Bioinformatic tools for analysis, mining and modelling large-scale gene expression and drug testing datasets

John Patrick Mpindi

Institute for Molecular Medicine Finland
University of Helsinki
and
Doctoral Programme in Biomedicine (DPBM)
Faculty of Medicine
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Lecture Hall 3, Biomedicum Helsinki on Friday, March 11th, 2016, at 12 noon.

Helsinki 2016
Supervised by
Professor Olli Kallioniemi MD, PHD
Director,
Institute for Molecular Medicine Finland
University of Helsinki, Finland

Thesis Committee
Professor Samuel Kaski, PHD
Director
Finnish Centre of Excellence in Computational Inference Research COIN
Aalto University and University of Helsinki

And

Juha Pekka Turunen, MD
Head of Education
The Finnish Medical Society Duodecim

Reviewed by
Adjunct Professor Laura Elo-Uhlgren
Research Director in Bioinformatics
Turku Centre for Biotechnology
University of Turku and ABO AKADEMI

And

Dr Silje Nord
Oslo University Hospital
Department of Cancer Genetics
Institute for Cancer Research

Official Opponent
Dr Bissan Al-Lazikani
Director Computational Biology and Chemogenomics
Cancer Therapeutics Unit
The Institute of Cancer Research UK

http://ethesis.helsinki.fi
Unigrafía
Helsinki 2016
“Never calculate without first knowing the answer.”

John Archibald Wheeler

To Piia, Petra and Susan,

Family is a gift that lasts forever.
Table of Contents

List of Original Publications 1

Abbreviations 2

Abstract 4

1. Introduction 6

2. Review of the literature 12
   2.1 Personalized medicine for cancer 12
   2.2 Personalized medicine datasets 13
   2.3 Bioinformatics methods in personalized medicine 18
   2.4 Oncogene expression profiling 20
      2.4.1 What are oncogenes? 20
      2.4.2 Oncogenes as therapeutic targets 20
      2.4.3 Oncogene expression profiling 20
      2.4.4 Methods for analyzing oncogene outlier expression 23
   2.5 Stepwise validation of oncogene outlier hits using HT-RNAi screening 25
      2.5.1 Limitations of HT-RNAi technology as a validation tool 26
   2.6 Meta-analysis of gene expression and drug testing data 28
      2.6.1 Inconsistency and reproducibility challenges in drug testing studies 29

3. Aims of the study 31

4. Materials and methods 32
   4.1 Datasets 32
      4.1.1 Gene expression datasets 32
      4.1.2 Drug testing datasets 32
   4.2 Data preprocessing and quality control 33
      4.2.1 Raw gene expression data processing 33
      4.2.2 Processing and quality control of raw drug testing data 34
   4.3 Outlier gene expression analysis 35
      4.3.1 The t-statistic 36
      4.3.2 Cancer outlier profile analysis (COPA) 36
      4.3.3 Outlier sums (OS) 37
      4.3.4 Outlier robust t-statistic (ORT) 37
      4.3.5 Gene tissue index (GTI) 38
   4.4 High-throughput screening normalization methods 39
   4.5 Drug testing analysis methods 40
      4.5.1 Percent inhibition 40
      4.5.2 Drug response scoring 40
   4.6 Data visualizations 41

5. Results 44
   5.1 Comparison of outlier analysis methods for the analysis of gene expression data 45
      5.1.1 Comparisons using a simulated gene expression dataset 45
      5.1.2 Comparison using a large-scale glioma microarray integrated dataset 45
   5.2 Other applications of the GTI algorithm in cancer drug targets and biomarker search 49
      5.2.1 The GTI identifies cancer biomarkers and drug targets 49
      5.2.2 The GTI and differential correlation map (corDiff-map) identify cancer-specific co-expression networks 52
   5.3 Impact of normalization methods on high-throughput screening data 55
   5.4 Analysis of consistency of large-scale drug testing and pharmacogenomics studies 57
5.4.1 Consistent bioinformatic meta-analysis gives consistent results 59
5.4.2 Correlations made when there is little response lead to a low concordance score 62

6. Discussion 65
7. Conclusion and future prospects 74
8. Acknowledgements 76
9. References 80
List of Original Publications

This thesis is based on the following publications (referred to in the text by their Roman numerals I–IV).


Publications included in another thesis

Publication II was included in the Ph.D. thesis of Paula Vainio with the title “High throughput screening for novel prostate cancer drug targets”, University of Turku 2011.

The results section includes previously unpublished data, such as visualizations of bioinformatics analyses.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIM1</td>
<td>Absent in melanoma 1</td>
</tr>
<tr>
<td>AMACR</td>
<td>Alpha-methylacyl-CoA racemase</td>
</tr>
<tr>
<td>CCLE</td>
<td>Cancer cell line encyclopedia</td>
</tr>
<tr>
<td>CDF</td>
<td>Chip definition file</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>Cyclin-dependent kinase inhibitor 2A</td>
</tr>
<tr>
<td>CGP</td>
<td>Cancer Genome Project</td>
</tr>
<tr>
<td>CLDN3</td>
<td>Claudin 3</td>
</tr>
<tr>
<td>COPA</td>
<td>Cancer outlier profile analysis</td>
</tr>
<tr>
<td>corDiff-map</td>
<td>Differential correlation map</td>
</tr>
<tr>
<td>cormap</td>
<td>Correlation map (correlation heatmap)</td>
</tr>
<tr>
<td>CYP4F8</td>
<td>Cytochrome P450, family 4, subfamily F, polypeptide 8</td>
</tr>
<tr>
<td>DHFR</td>
<td>Dihydrofolate reductase</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSS</td>
<td>Drug sensitivity score</td>
</tr>
<tr>
<td>EC50</td>
<td>Half maximal effective concentration</td>
</tr>
<tr>
<td>EPHX2</td>
<td>Epoxide hydrolase 2, cytoplasmic</td>
</tr>
<tr>
<td>ERBB2</td>
<td>ErbB2 receptor tyrosine kinase 2</td>
</tr>
<tr>
<td>ERGIC1</td>
<td>Endoplasmic reticulum-golgi intermediate compartment (ERGIC) 1</td>
</tr>
<tr>
<td>EST</td>
<td>Expressed sequence tags</td>
</tr>
<tr>
<td>FAAH</td>
<td>Fatty acid amide hydrolase</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and drug administration</td>
</tr>
<tr>
<td>FOXA1</td>
<td>Forkhead box A1</td>
</tr>
<tr>
<td>GBM</td>
<td>Glioblastoma</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene expression omnibus</td>
</tr>
<tr>
<td>GFAP</td>
<td>Glial fibrillary acidic protein</td>
</tr>
<tr>
<td>GTI</td>
<td>Gene tissue outlier index</td>
</tr>
<tr>
<td>HPA</td>
<td>Human protein atlas</td>
</tr>
<tr>
<td>HTS</td>
<td>High throughput screening</td>
</tr>
<tr>
<td>IC50</td>
<td>Half-maximal inhibitory concentration</td>
</tr>
<tr>
<td>MM</td>
<td>Mismatch (probe)</td>
</tr>
<tr>
<td>MoA</td>
<td>Mechanism of action</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger ribonucleic acid</td>
</tr>
<tr>
<td>MTDH</td>
<td>Metadherin</td>
</tr>
<tr>
<td>NCI</td>
<td>National Cancer Institute</td>
</tr>
<tr>
<td>ODC1</td>
<td>Ornithine decarboxylase 1</td>
</tr>
<tr>
<td>ORT</td>
<td>Outlier robust t-statistic</td>
</tr>
<tr>
<td>OS</td>
<td>Outlier sums</td>
</tr>
<tr>
<td>PM</td>
<td>Perfect match (probe)</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>RMA</td>
<td>Robust multichip average</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>Ribonucleic acid sequencing</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>RNAi</td>
<td>Ribonucleic acid interference</td>
</tr>
<tr>
<td>RRM2</td>
<td>Ribonucleoside-diphosphate reductase M2 subunit</td>
</tr>
<tr>
<td>SIM2</td>
<td>Single-minded homolog 2</td>
</tr>
<tr>
<td>siRNAi</td>
<td>Small interfering ribonucleic acid</td>
</tr>
<tr>
<td>SSMD</td>
<td>Strictly standardized mean difference</td>
</tr>
<tr>
<td>TCGA</td>
<td>The Cancer Genome Atlas</td>
</tr>
<tr>
<td>TMED3</td>
<td>Transmembrane emp24 protein transport domain containing 3</td>
</tr>
<tr>
<td>TPX2</td>
<td>TPX2, microtubule-associated, homolog (Xenopus laevis)</td>
</tr>
<tr>
<td>TYMS</td>
<td>Thymidylate synthetase</td>
</tr>
<tr>
<td>UBE2C</td>
<td>Ubiquitin-conjugating enzyme E2C</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>MRC</td>
<td>Median rank correlation</td>
</tr>
</tbody>
</table>
Abstract

Bioinformatic tools applied to large-scale genomic and gene expression datasets have helped in developing our understanding of the molecular basis of cancer. They have also become an important component of the drug discovery and development process, and potentially of personalized medicine for the future. Bioinformatic studies are now benefiting from the wealth of large datasets generated in laboratories through the use of new high-throughput technologies and their massive public repositories of data. As of October 2015, the GEO database (www.ncbi.nih.gov/geo/) alone comprised 1,597,783 samples across 15,040 platforms, and it is being updated on a daily basis. It is becoming evident that biological data are accumulating faster than the capacity of the scientific community to analyse, integrate and mine the data, as well as to create knowledge, understanding, and insights from the data. Thus, there is a growing need for better bioinformatic tools for analysing, mining and modelling both local and global datasets.

Many data analysis projects have called for the assembly of specialized data analysis tasks and pipelines. In the future, bioinformaticians need to be involved in both methods development and in applied bioinformatics. Methods development refers to developing new algorithms, while applied bioinformatics involves putting together existing tools/pipelines in a creative way to perform an analysis task. Bioinformatics is complicated due to the heterogeneous nature of the data, varying experimental settings, small sample sizes with little replication and the existence of many distributions in the data. There is also no uniformly accepted method for large-scale integrated data analysis. The aim of this study was to develop bioinformatic and statistical tools to perform an integrated analysis of large-scale microarray gene expression, high-throughput RNAi screening and drug testing data, as well as to demonstrate the applicability of these approaches in cancer research, drug target discovery and drug testing.

First, the gene tissue index (GTI) outlier analysis method was developed to identify cancer outlier genes from large-scale microarray datasets. The need to identify genes (‘outlier genes’) highly expressed in a subgroup of samples rendered some of the traditional differential expression analysis methods inadequate. The GTI method enabled the analysis and mining of outlier expression profiles from heterogeneous large-scale microarray datasets that usually contain a variable number of samples for each gene being compared. Using real and simulation study datasets, the performance of the GTI method was evaluated. We observed that the GTI performed equally well in single study settings compared to existing outlier analysis methods. Furthermore, the performance of the GTI method based on discovery studies in glioblastoma and prostate cancer was notable based on the biology of the top genes identified by the GTI. This analysis revealed many genes with outlier expression patterns, and the approach is directly applicable to the identification of drug targets and cancer biomarkers, and for cancer subtype classification studies.

Secondly, there have been significant concerns over the reproducibility of high-throughput screening data in the microplate format for both RNAi screening and for drug testing data in cancer cell lines. Some of this variability may be related to the
study design and statistical methods, which could be further controlled. Here, we carried out a systematic study to assess the impact of normalization methods on the reproducibility and quality of high-throughput screening data with high hit rates and drug testing with dose–response data. This study revealed that the hit rate and the plate layout significantly affect the performance of normalizations, and hence the quality of high-throughput screening data.

Finally, high-throughput drug testing data were analysed for consistency across three large-scale pharmacogenomic datasets, which were systematically processed using standardized bioinformatic analysis methods while controlling assumptions for statistical inference on large-scale data matrices. We standardized data processing and analysis methods for generating dose–response curves and drug response scoring across the three datasets. For example, the concentration of one drug screened at all the three sites was merged in one standard window, and the meta-analysis was performed either between cell lines or between measurements, such as genes and drugs. The results based on standardized bioinformatic analysis of drug testing and gene expression datasets demonstrated a high correlation between two of the sites tested, and moderate agreement between the others.

In conclusion, broad standardization of the methods both for laboratory measurements as well as for applied bioinformatics will be necessary to ensure greater reproducibility of biological findings in cancer research and therapeutic/biomarker discovery. I envisage that improved methods for the analysis and interpretation of large-scale datasets might accelerate our ability to advance personalised medicine.
1. Introduction

Personalized medicine is an approach based on the tailoring of treatment options to an individual by identifying the molecular disease signature from a patient and then matching it with the most appropriate treatment[1-3]. For decades, clinicians have offered the same kind of treatment for the same type of cancer, despite being aware that drug treatments only work on a subset of patients based on retrospective evidence from clinical trials[4]. New studies have led a shift in cancer treatment from relatively nonspecific cytotoxic agents to targeted therapies and cancer immunotherapies. Targeted therapies are based on inhibiting oncogenic pathways and mechanisms behind the progression of cancer[4,5]. However, the successful transition from cytotoxic agents to individualized targeted therapies hinges on our ability to effectively interpret, analyse and integrate large-scale genomic, transcriptomic, proteomic and drug testing datasets into actionable and tailored treatment regimens for individual patients or small patient subgroups.

Cancer is a heterogeneous disease with variable clinical, pathological and molecular features. The heterogeneous nature of cancer has made it an important human health problem worldwide, accounting for 1,323,600 deaths annually in Europe alone[6]. The identification of subgroups of patients with distinct genomic, molecular and clinical features is crucial for the effective development and delivery of treatments to the clinic. Many molecular techniques, particularly genome sequencing and gene expression profiling, have increasingly been used to help identify cancer subtypes and driver oncogenes, and to assess treatment outcomes. The identification of driver oncogenes behind the progression of cancer is applicable for the diagnosis of cancer and the development of new targeted therapies[7-10]. In these approaches, it is assumed that tumour driving processes turn on certain genes that are not usually expressed in normal cells (oncogenes) and down-regulate genes that are needed for critical regulatory functions in normal cells (tumour suppressors)[7] or alter them to enhance their effect.

Microarrays and, more recently, RNA sequencing approaches can reveal global changes in gene expression in response to genetic and environmental changes. They
are thus well suited to constructing hypotheses concerning the differences between normal and cancer cells and how this would be useful for biomarker search and therapeutic discovery[11,12]. Likewise, functional analyses, such as high-throughput RNAi screening and drug testing, are increasingly being applied, for instance, to established cancer cell lines, drug-resistant cancer cell models, primary cancer cells, and progenitor/stem cell models of disease[8,13-19]. This facilitates investigation of the molecular mechanisms of disease and the functional effect of molecular and chemical inhibitors on these processes. Researchers now have the ability to identify key driver oncogenes and signatures enriched in cancer pathways that drive tumour progression, which will increase our understanding of disease phenotypes and will help the practice of individualized medicine in the future.

The wealth of high-throughput molecular profiling data currently available in public repositories such as GEO, Array Express, Oncomine, TCGA and the Connectivity Map[11,20-23] provides an opportunity to systematically analyse these data for oncogene expression profiles (‘outlier genes’). While there has recently been a shift towards using RNA sequencing as a means for quantifying gene expression in cancer, the majority of the publicly available datasets are still based on microarray gene expression profiles[24-26]. Furthermore, the growth of large-scale pharmacogenomics datasets from public repositories and our increasing ability to screen thousands of compounds across a large number of, for example, established cancer cell lines, drug-resistant cancer cell models, primary cancer cells, iPS and other stem cell models is driving the promise of personalised medicine[13,14].

These large-scale biomedical datasets are creating tremendous challenges in terms of their storage and the application of bioinformatics methods to analyse them. As noted by Manish Parashar, a computer scientist and head of the Rutgers Discovery Informatics Institute in Piscataway, New Jersey, “We can collect data faster than we can analyze them”[26-29]. Bioinformaticians are facing complex choices. Many are opting for applied bioinformatics ‘methods repurposing’ rather than pure bioinformatics ‘methods development’. Here, pure bioinformatics refers to the development of new algorithms that are described in new publications, while methods repurposing involves putting together tools/pipelines in a creative way to perform an analysis task. Some service cores are spending less than 5% of the time on pure
bioinformatics, as reported by Chang[30]. Efforts to develop new pure bioinformatics methods have been hindered by the growing volume of customized data analysis tasks with few routines that warrant automation. The situation is even more complicated for methods repurposing approaches due to the heterogeneous nature of the data and the variety of data platforms. For example, according to the OmicsMaps report on all known sequencing platforms in the world[24], we have more than 2,500 high throughput instruments manufactured by several companies located in about 1000 sequencing centres in universities, hospitals, and research laboratories of 55 countries. The biomedical research community is constantly developing pure and applied bioinformatics methods to analyse data from each sequencing platform, but systematic large-scale platform integration approaches are still needed. It might be possible that methods designed for one technology, such as microarray gene expression, could be repurposed in a creative way to address personalised medicine questions using data from other unrelated high-throughput technologies.

For example, most microarray studies have focused on the identification of differentially expressed genes, using a panel of perturbed and control samples collected at the same time and analysed on a single microarray platform. These study settings represent datasets that are relatively homogeneous, containing a small number of samples for all genes. When results from such individual studies are compared with each other, the overlap of the differentially expressed gene sets is sometimes disappointing and insignificant[31]. Consequently, current approaches to analysing gene expression data have focused on integrating expression data from various microarray platforms from the same manufacturer[22,32,33], e.g. Affymetrix, into a single large-scale integrated dataset with thousands of samples acquired from different cancer types. Most large-scale integrated microarray datasets are typically diverse collections of small studies that feature different experimental conditions, varying sample preparation methods and labelling methods or scanner settings, and different microarrays or microarray platforms. These multiple layers of variability introduce numerous challenges to the statistical methods applied in meta-analyses. The enormous complexity of patterns of gene expression observed in large-scale integrated microarray datasets renders traditional statistical and bioinformatic methods that focus on identifying uniform overexpression such as the t-statistic (see Methods section of study I) inadequate and bound to fail to adequately identify
oncogene signatures. Due to the generic heterogeneity, any given oncogene is also typically aberrant in a small fraction (e.g. 1–20%) of cancers. According to a study on prostate cancer by Tomlins et al. [34], searching for oncogene outlier expression patterns in a subgroup of samples uncovers new oncogenes that cannot be captured by traditional differential expression analysis. In general, outlier analysis facilitates a powerful way to identify critical pathogenic genes involved in a subset of disease samples. Important discoveries of new oncogenes based on outlier profile analysis among different tumour subtypes have already been described in the literature, such as the *TMPRSS2-ERG* [34] fusion gene, the best-known example in prostate cancer. Thus, with the development of new bioinformatic methods, it is possible to identify new oncogenic expression patterns from large-scale integrated microarray expression datasets.

Large-scale pharmacogenomic [35-38] studies are generating data on drug sensitivity profiles, and the volume of such data is also soaring, as they can be created using efficient high-throughput screening methods. Integrated analysis of genomic/gene expression and drug testing data will facilitate drug discovery and repurposing based on matching data from gene mutations or expression levels with the drug sensitivity and resistance patterns seen with matching oncoprotein inhibitors. The US National Cancer Institute (NCI) pioneered these efforts by assembling the NCI60 tumour cell line panel, which as of January 2016 has been cited in over 336 PubMed articles. In 1992, Weinstein et al. performed studies using the NCI60 panel of cell lines to identify the mechanism of action (MoA) of drugs [39]. In 2000, Weinstein et al. published a study [40] on linking bioinformatics and chemoinformatics by correlating gene expression and drug response patterns in the NCI60 panel of cell lines. Recent large-scale pharmacogenomic studies [41,42] have demonstrated that drug sensitivity and resistance testing could be linked to molecular profiles across a large panel of cell lines.

These recent pharmacogenomic studies have observed differences in the way cell lines cluster based on the different types of datasets, but have not systematically assessed whether the differences were driven by the data processing and scoring methods, which were repurposed from traditional compound screen studies. In a recent study, Haibe-Kains et al. [43] highlighted the lack of consistency in large-scale
studies, and the follow-up articles and comments[44-46] discussed many factors that could be behind the observed inconsistencies. Thus, the development and systematic evaluation of bioinformatic methods is highly necessary for the advancement of personalised medicine.

With the four studies presented in this thesis, I have demonstrated a way of identifying new oncogenes from large-scale integrated microarray datasets and systematically shown the importance of integrated analysis of gene expression data, high-throughput RNAi screening and drug testing data. The laboratory protocols for drug testing bear close resemblance to high-throughput RNA interference (HT-RNAi) screening, but the drug testing approach has some fundamental differences. The hit rate is very high for drug testing studies compared to the traditional drug discovery application or RNAi screening.

In the first publication, I explain the development of a new statistical method (the gene tissue index, GTI) based on modifying and adapting algorithms originally developed for statistical problems in economics. I explain how I performed a comparative study of methods used to identify genes up-regulated in subsets of samples of a given tumour subtype (‘outlier genes’), and continue to demonstrate the performance of the GTI in both single and combined study settings (large-scale integrated datasets). The second study shows the utility of the GTI outlier method for identifying biologically relevant genes in prostate cancer.

In the third study, I demonstrated the impact of normalization methods on high-throughput screening data with high hit rates. Most high-throughput screening methods were developed for screens with low hit rates, and these methods do exist in most readily available commercial and open source software. The most common approach for biomedical researchers is to use the method repurposing strategy by adapting already existing tools to these new biological questions. This study demonstrated potential problems that could arise from using the method repurposing approach on datasets with different experiment setups and suggested an assessment criterion. The small overlap between the top hits from two high-throughput RNAi (HT-RNAi) experiments performed in the second study highlights the need for improving data reproducibility based on RNAi screening data. The low
reproducibility of HT-RNAi data analysis results across different cancer types or experimental settings has been discussed previously[16], and three potential reasons were identified: false positives, a lack of standards for data normalization or processing methods and the varying transfection protocols. The first and second reasons are statistical and computationally based problems, and they could be solved by careful planning of experiments to enable robust utilization of method repurposing approaches based on existing data processing methods or the development of new pure bioinformatic methods. The third reason is more complicated to solve due to the lack of standards across different laboratories and suppliers of laboratory reagents.

In the fourth study, I explored and evaluated an applied bioinformatic “method repurposing” approach based on performing standardized drug response scoring across three drug testing studies while controlling assumptions for statistical inference on data matrices[47]. The standardization and the systematic assessment of the reproducibility of drug testing data demonstrates stronger consistency than what had been reported by Haibe-Kains et al. [43]. The analysis was complemented with data visualization tools to enable the systematic evaluation of our drug testing analysis standardization approach.
2. Review of the literature

2.1 Personalized medicine for cancer

The central goal of personalized medicine is to enable clinicians to match a patient’s molecular profile to a tailored treatment regimen. Determining actionable treatment regimens from large-scale complex and heterogeneous genetic and transcriptomics datasets requires specialized tests and bioinformatics tools that are compatible with the clinical workflow and are economically cost optimal for smooth implementation in the current healthcare system [48].

Clinical oncologists have for long known that cancers of the same type manifest in different ways for each individual patient, and drug responses are therefore difficult to predict [49,50]. Thus, the ability to predict how patients will respond to treatment is an important goal of molecular medicine and bioinformatics. With the advance of technologies enabling the systematic analysis of the cancer genome, transcriptome and proteome using bioinformatics tools, scientists and clinicians have been able to better understand the genomic and molecular heterogeneity of human cancers. Molecular profiling techniques, particularly gene expression, have been used to predict patient outcomes, classify patients into subgroups, discover new oncogenes and predict the sensitivity of patients to drugs [23,51-53]. Until the late 1990s, almost all drugs used for cancer treatment were cytotoxic agents (except hormone treatments), and these worked by killing dividing cells [54]. These cytotoxic agents are unspecific and do not spare dividing normal cells, although cancer cells, given their higher proliferation rate, are more strongly affected.

The ultimate strategy of targeted therapies is to target biological processes that are more active in cancer cells than in the normal cells. These processes could include those that control growth, division, cell death, proliferation, replication, immortality and the migration of cancer cells or stroma interactions [55]. It has been determined that abnormal types of growth factors or abnormally high levels of growth factors contribute to the proliferation of cancer cells [55]. In 1982, de Klein et al. [56] discovered that in some cases of chronic myeloid leukemia (CML), the *ABL proto-
Review of the literature

oncogene is translocated to the *BCR* gene located on chromosome 22, and this initiated many studies on genomic arrangements leading to the identification of oncogenes and subsequent efforts to develop anti-cancer therapeutics for patients with these genetic alterations. In a study by Slamon et al., up to 20% of breast cancers were found to amplify and overexpress the *ERBB2/HER-2* gene[57,58].

Even though most of the early studies to characterize genetic lesions in cancers were performed in a low-throughput way, the emerging concepts of cancer genomics paved the way for the current high-throughput technologies that enable the querying of virtually all genes in a given specimen for their oncogenic potential. Integrated analysis of genetics, molecular profiling and drug testing is today being used to tailor individualized treatment regimens[13,35,37,42,59]. However, it has been shown that the quality of these interpretations is highly dependent on the data on which they are based[53]. For this reason, high-quality reference databases provide the foundation for personalized medicine.

### 2.2 Personalized medicine datasets

The growth of datasets in public repositories arising from high-throughput technologies, such as molecular profiling, has surpassed our capacity to analyse them[25,26,60-63]. For example, the use of DNA sequencing data has been growing at an astronomical pace and offering large datasets, such as the 1000 Genomes Project, aggregating hundreds of human genomes by 2012[64], the Cancer Genome Atlas (TCGA) [65] providing several normal/tumour genome pairs, the Exome Aggregation Consortium (ExAC) [66] aggregating over 60,000 human exomes, and the International Cancer Genome Consortium (ICGC), supporting and accelerating the translation of genomic discoveries and also managing the Expression Project for Oncology (expO)[67]. Even though the majority of these efforts have focused on exome rather than whole genome sequencing, this is expected to change in the future[28]. Likewise, large-scale pharmacogenomics datasets such as the Genomics of Drug Sensitivity in Cancer project (GDSC) [41] the Cancer Cell Line Encyclopedia (CCLE) [36] and the NCI60 cell line panel[68] as well as molecular profiling datasets such as the Gene Expression Omnibus (GEO) [20], ArrayExpress [21] and the Connectivity Map (Cmap) [23], have been published and are maintained as public resources. Recent work by Jiao *et al.* summarized a list of data repositories supporting
computational drug repositioning approaches (Table 1). These datasets are huge and currently stored in separate formats and databases, making their integration difficult. Hence, the scientific community is not making optimal use of them. On the other hand, the pace of developing new bioinformatics tools for mining these data has been slow. While the use of applied bioinformatics approaches has worked for some tasks, many methods still produce inconsistent and less reproducible results. Our ability to integrate large-scale pharmacogenomic and molecular profiling datasets is hindered by the heterogeneous nature of these data [26]. There are numerous distributions in these datasets that cannot be captured by traditional methods assuming a normal distribution.

The growth of large-scale datasets is fuelled by continued advances in the development of high-throughput technologies. At the same time, the costs of generating new data are decreasing, making it possible to produce data on a large number of samples across many features. For example, the current price of approximately US$1,000 to sequence one human genome has led to sequencing projects covering from thousands up to hundreds of thousands of specimens[25,62,69].

A large number of technologies are used to produce biomedical datasets. For example, gene expression can currently be acquired from a variety of technology platforms, each with a different output design. It is virtually impossible to query the expression of one gene using a single gene ID across a multitude of gene expression platforms without writing a customized bioinformatic script and performing normalization across all the technology platforms. Despite full guidelines being issued in 1979 to adopt gene nomenclature standards[70], a variety of names do exist in large-scale datasets. The data formats range from genomic, proteomic and transcriptomic to drug responses and clinical data. Without standards, we would need to develop customized data processing methods for each format, which is beyond the capacity of the scientific community. For example, in genomics, we currently have no standard data format that technology producers or analysts use, but rather, a set of commonly used formats (e.g. FASTA/Q, SAM, VCF, GFF/GTF), meaning that a lot of time is spent on converting multiple data formats to achieve comparability[71].
Table 1: Data resources supporting computational repositioning strategies. Adapted from Jiao et al., Briefings in Bioinformatics, 2015[69]

<table>
<thead>
<tr>
<th>Large-scale datasets</th>
<th>Database</th>
<th>URL</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArrayExpress</td>
<td></td>
<td><a href="http://www.ebi.ac.uk/arrayexpress">http://www.ebi.ac.uk/arrayexpress</a></td>
<td>Public repositories of functional genomics data.</td>
</tr>
<tr>
<td>Gene Expression Atlas</td>
<td><a href="http://www.ebi.ac.uk/geo">http://www.ebi.ac.uk/geo</a></td>
<td>Genomic data (e.g. Exome, SNP, Methylation, mRNA, Clinical) of &gt;10,000 patient tissue samples across &gt;30 common cancers.</td>
<td></td>
</tr>
<tr>
<td>The Cancer Genome Atlas (TCGA)</td>
<td><a href="http://cancergenome.nih.gov">http://cancergenome.nih.gov</a></td>
<td>Provides the Expression Project for Oncology (expO) and the TCGA (e.g. Exome, SNP, Methylation, mRNA, Clinical) of &gt;1000 cancer cell lines.</td>
<td></td>
</tr>
<tr>
<td>International Genomics Consortium</td>
<td><a href="http://www.intgen.org/about-igc/">http://www.intgen.org/about-igc/</a></td>
<td>Genomic data (e.g. DNA copy number, mRNA expression, mutation data) of &gt;1000 cancer cell lines.</td>
<td></td>
</tr>
<tr>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td><a href="http://www.broadinstitute.org/ccle">http://www.broadinstitute.org/ccle</a></td>
<td>A comprehensive description of genomic, transcriptomic and epigenomic changes.</td>
<td></td>
</tr>
<tr>
<td>International Cancer Genome Consortium</td>
<td><a href="https://icgc.org">https://icgc.org</a></td>
<td>Gene expression profiles of &gt;1000 drugs across three primary cell lines.</td>
<td></td>
</tr>
<tr>
<td>The Connectivity Map (CMap)</td>
<td><a href="http://www.broadinstitute.org/cmap">http://www.broadinstitute.org/cmap</a></td>
<td>Aim to produce &gt;1 million gene expression profiles of drugs and genetic perturbagens across &gt;15 cell lines.</td>
<td></td>
</tr>
<tr>
<td>Molecular Signature Database (MsigDB)</td>
<td><a href="http://www.broadinstitute.org/gsea/msigdb">http://www.broadinstitute.org/gsea/msigdb</a></td>
<td>Functional annotation tools.</td>
<td></td>
</tr>
<tr>
<td>Gene Signature Database (GeneSigDB)</td>
<td><a href="http://compbio.dfci.harvard.edu/genesigdb">http://compbio.dfci.harvard.edu/genesigdb</a></td>
<td>Resource for understanding high-level functions and utilities of the biological system from molecular-level information.</td>
<td></td>
</tr>
<tr>
<td>Database for Annotation, Visualization and Integrated Discovery (DAVID)</td>
<td><a href="http://david.abcc.ncifcrf.gov">http://david.abcc.ncifcrf.gov</a></td>
<td>Tool to determine whether an a priori defined gene signature shows statistically significant, concordant differences between two biological states.</td>
<td></td>
</tr>
<tr>
<td>Gene Set Enrichment Analysis (GSEA)</td>
<td><a href="http://www.broadinstitute.org/gsea">http://www.broadinstitute.org/gsea</a></td>
<td>Genetic variation in drug response.</td>
<td></td>
</tr>
<tr>
<td>Drug versus Disease (DvD)</td>
<td><a href="http://www.ebi.ac.uk/saezrodriguez/dvd">www.ebi.ac.uk/saezrodriguez/dvd</a></td>
<td>The relationship between genes and genetic phenotypes, particularly disorders.</td>
<td></td>
</tr>
<tr>
<td>The Pharmacogenomics and Pharmacogenomics Knowledge Base (PharmGKB)</td>
<td><a href="http://www.pharmgkb.org">http://www.pharmgkb.org</a></td>
<td>Adverse drug reactions to &gt;900 drugs.</td>
<td></td>
</tr>
<tr>
<td>Online Mendelian Inheritance in Man (OMIM)</td>
<td><a href="http://www.omim.org">http://www.omim.org</a></td>
<td>A registry and result database of publicly and privately supported clinical studies.</td>
<td></td>
</tr>
<tr>
<td>Side Effect Resource (SIDER)</td>
<td><a href="http://sieffects.embl.de">http://sieffects.embl.de</a></td>
<td>Information about FDA-approved drugs, such as brand name, therapeutic products.</td>
<td></td>
</tr>
<tr>
<td>Phenome/Drug</td>
<td>Drugs@FDA Database</td>
<td><a href="http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm">http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm</a></td>
<td>A comprehensive, publicly accessible collection of approved and investigational drugs.</td>
</tr>
<tr>
<td>Drug Combination Database (DCDB)</td>
<td><a href="http://www.cls.zju.edu.cn/dcdb">http://www.cls.zju.edu.cn/dcdb</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Review of the literature

<table>
<thead>
<tr>
<th>Large-scale datasets</th>
<th>Database</th>
<th>URL</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Data Bank (PDB)</td>
<td><a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a></td>
<td>3D structure of proteins, nucleic acids.</td>
<td></td>
</tr>
<tr>
<td>SWEETLEAD</td>
<td><a href="https://simtk.org/home/sweetlead">https://simtk.org/home/sweetlead</a></td>
<td>A database containing chemical structures representing approved drugs, chemical isolates from traditional medicinal herbs and regulated chemicals.</td>
<td></td>
</tr>
<tr>
<td>DrugBank</td>
<td><a href="http://www.drugbank.ca/">http://www.drugbank.ca/</a></td>
<td>Detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.</td>
<td></td>
</tr>
<tr>
<td>canSAR</td>
<td><a href="http://cansar.icr.ac.uk">http://cansar.icr.ac.uk</a></td>
<td>canSAR integrates genomic, protein, pharmacological, drug and chemical data with structural biology, protein networks and druggability data.</td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td><a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a></td>
<td>A large literature-derived database of molecule structures and molecule-protein interactions. This includes a catalogue of approved drugs.</td>
<td></td>
</tr>
</tbody>
</table>
Another important characteristic of large biomedical datasets is that each public data repository generates unique identifiers. As an example, the drug Dasatinib has been assigned over ten different identifiers, depending on the database (Table 2)[72,73].

Table 2: Identifiers for Dasatinib listed in drug bank databases.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG Drug</td>
<td>D03658</td>
</tr>
<tr>
<td>PubChem Compound</td>
<td>3062316</td>
</tr>
<tr>
<td>PubChem Substance</td>
<td>46505143</td>
</tr>
<tr>
<td>ChemSpider</td>
<td>2323020</td>
</tr>
<tr>
<td>BindingDB</td>
<td>13216</td>
</tr>
<tr>
<td>ChEBI</td>
<td>49375</td>
</tr>
<tr>
<td>ChEMBL</td>
<td>CHEMBL1421</td>
</tr>
<tr>
<td>Therapeutic Targets Database</td>
<td>DAP000004</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>PA162372878</td>
</tr>
<tr>
<td>Drug Product Database</td>
<td>20122</td>
</tr>
<tr>
<td>RxList</td>
<td><a href="http://www.rxlist.com/cgi/generic/sprycel.htm">http://www.rxlist.com/cgi/generic/sprycel.htm</a></td>
</tr>
<tr>
<td>Drugs.com</td>
<td><a href="http://www.drugs.com/cdi/dasatinib.html">http://www.drugs.com/cdi/dasatinib.html</a></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Dasatinib</td>
</tr>
</tbody>
</table>

Minimum information standards and guidelines have been popular and set for some datasets to ensure that these data are easily interpreted and integrated. Most of these guidelines have not focused on the raw data formats and structure, but rather on high-level features that are a result of data processing techniques.

Large-scale datasets tend to exist in several terabytes, even in small labs. Those researchers without high-throughput instruments can also quickly become big data users by querying large data repositories such as EBI, canSAR[74] or the US National Center for Biotechnology Information. It is estimated that each day in 2012, the EBI received over 9 million requests to query its large datasets [29]. Another example, canSAR 3.0[74], as of January 2016 was holding data on over one million bioactive, small molecule drugs and compounds, and over 140,000 users were utilizing this resource to generate new hypotheses.
2.3 Bioinformatics methods in personalized medicine

Bioinformatic analysis and mining of large-scale datasets is a complex process that requires specialized skills in programming, statistics, and mathematics, as well as biological understanding. Alyass et al. have suggested that personalized medicine needs hybrid education [48]. Fields such as meteorology, finance, astronomy and companies such as Google, Yahoo, Microsoft, Apple, and many others have for long handled large datasets, but these datasets are often fairly homogeneous [28,29]. In contrast, the heterogeneous nature of personalized medicine datasets requires sufficient understanding of the underlying biological concepts and analysis algorithms. Very often, no appropriate algorithm exists with which these kinds of datasets can readily be analysed. Moreover, bioinformaticians trained to analyse these types of datasets are in short supply. There has been a decline in the number of new analysis methods being published [30], and this could be attributed to the lack of uniformity in the biological analysis tasks assigned to bioinformaticians and the constant need to investigate many biological questions using applied bioinformatics ‘method repurposing’ approaches. An article by Chang[30] notes that many bioinformaticians at core facilities spend less than 5% of their time on pure bioinformatics. Furthermore, the complex nature of large-scale datasets makes it complicated to perform method-repurposing approaches.

Recently, Chang introduced the terms ‘pure’ and ‘applied’ bioinformatics, and defined them as: a) pure bioinformatics “method development” if a bioinformatician is developing new algorithms that fit into a publication, and b) applied bioinformatics “method repurposing” if one is putting together tools/pipelines in a creative way to perform an analysis task. Both of these approaches are complicated due to the complex nature of large-scale personalized medicine datasets. The challenges faced by bioinformaticians applying these strategies arise from the fact that they need to understand a) the biological question, b) the analysis methods and c) the pros and cons of different algorithms and programming language with the most efficient analysis algorithm, and appropriate statistical methods. Some of the straightforward method-repurposing approaches do lead to low reproducibility of the results and only identify the most obvious hits.
Recently, studies by Haibe-Kains et al. [43] and Yadav et al. [75] on integrated analysis of pharmacogenomic datasets have demonstrated that the use of the most common drug response scoring metric, the IC50 (half maximal inhibitory concentration), is not robust. A prostate cancer study by Tomlins et al. revealed that the traditional t-statistic commonly used for outlier statistics could not identify oncogenes showing an outlier expression profile in a subgroup of patient samples [76]. In addition, the B-score algorithm (see Methods section of study III), which is one of the most used normalization methods in high-throughput screening [77-79], has been assessed in several recent publications [80,81]. For example, Murie et al. [80] showed that Spatial Polish and Well Normalization (SPAWN), with an iterative polish technique similar to that for the B-score, performs well based on using the trimmed mean, rather than a median. The trimmed-mean approach has been shown to have good robustness [82] and leads to fewer false positives generated by the B-score.

The skill set required to perform integrated personalized medicine method development and repurposing is multidisciplinary. Pure method development could occur in independent bioinformatics laboratories and statistics research groups, but applied bioinformatics (method repurposing) involves constant interaction with biologists in order to understand the methodological design of experiments and high-throughput experiment readouts from non-conventional technologies. Most of the applied bioinformatic “method repurposing” tasks do not lead to their own research projects and cannot easily be listed as research projects [30]. The growing size of personalized medicine datasets and the heterogeneity of the data makes the development of pure bioinformatic methods complicated. Performing simulations based on large-scale personalized medicine datasets can take days and could delay the deployment of new methods. Those involved in the development of bioinformatic methods usually come from fields such as physics, mathematics, statistics and biology, with little experience of best practices in software development such as code optimization, versioning, issue tracking and task parallelization [83]. The creation of interdisciplinary research institutes and networks under the European biosciences research initiative (the European Molecular Biology Laboratory, EMBL), which aim to translate research findings back to clinics, could be an active driver for more efficient personalized medicine studies in the future. New career track schemes optimized for those carrying out applied bioinformatic method development in
personalized medicine projects have been suggested[84]. Informatics tools for analysis, mining, and modelling large-scale personalized medicine datasets are needed to enable efficient delivery of the promise of personalised medicine.

2.4 Oncogene expression profiling

2.4.1 What are oncogenes?
Oncogenes are normal/regular cellular genes, so-called proto-oncogenes[85], which are usually involved in cellular signalling and growth regulation and may undergo aberrant activation in cancer, for instance through gene amplification, mutation or gene fusions. Genetic alterations of oncogenes lead to a higher concentration or activity of the encoded protein in cancer cells or overexpression of the gene in a peculiar way in the wrong cell type, leading to deregulated growth. Oncogenes also participate in biological processes such as differentiation, senescence and apoptosis. Oncogenes can be divided into three broad groups based on functional and biochemical properties of the protein products of their normal counterparts[10]. These groups are: (1) growth factors and their receptors, (2) signal transducers and (3) nuclear proto-oncogenes, including transcription factors.

2.4.2 Oncogenes as therapeutic targets
Studying the biological functions of oncogenes has many practical applications, including the development of new therapies, molecular diagnosis and the monitoring of cancer progression. Most importantly, oncogenes represent potential targets for new types of cancer therapies[86]. Mutations, fusions and deregulated expression of oncogenes can be used in the diagnostic assessment of cancer. Targeting oncogene products that play a crucial role in cancer development provides an opportunity to kill cancer cells without targeting normal cells[87,88]. This concept can be powerfully exploited in the treatment of cancer, such as the BCR–ABL oncogene in leukaemia through the inhibition of c-ABL kinase with Gleevec (imatinib) [89], and the inhibition of ErbB2 amplification in breast cancers using Herceptin[58].

2.4.3 Oncogene expression profiling
It is now possible to systematically evaluate genome-wide expression profiles due to the increased availability of new high-throughput gene expression profiling technologies. The methods available for gene expression profiling have evolved from basic quantitative PCR to microarrays to next-generation sequencing. Over the years,
the research community has employed a variety of techniques for large-scale gene expression profiling, such as microarrays (single or dual channel) and SAGE (Serial Analysis of Gene Expression). RNA sequencing (RNA-seq) is now replacing most of the existing gene expression profiling technologies. RNA-seq will probably soon be the dominant technology used for the analysis of expression, but at present, the biggest datasets in gene expression analysis are still based on microarray technology[65,69]. A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it originated[90,91]. Thousands of gene sequences are placed in known locations on a glass slide called a gene chip. A sample containing cDNA developed from cellular mRNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces a signal that is measured as a gene expression intensity score. The amount of mRNA bound to each site on the gene chip indicates the expression level of the various genes.

Currently, the scientific community is mainly using RNA-seq technology to research topics such as new transcript discovery, gene structure or allele-specific expression, while microarrays are to some extent still the tool of choice for studies that require larger numbers of publicly available datasets for statistical significance testing. The number of samples included in a typical RNA-seq experiment is still small due to several factors, including data handling and the cost[92]. A recent study observing the trend in the number of articles in PubMed including one of the two keywords “microarray” or “next-generation sequencing” in their article showed that microarray-based articles were overall the most dominant until 2012[93]. Figure 1 displays the new trends when searching with the keyword “genome sequencing”. These statistics mainly depend on keywords, since many next-generation sequencing (NGS) studies no longer mention this term. Furthermore, another study estimated that only 25% of all published microarray studies between 2000 and 2009 deposited their data in a public archive[94]. After 2009, the corresponding number was approximately above 45% [94] because the scientific community became aware of the value of making data public and the majority of scientific journals made this a standard requirement for all new publications based on microarray data. Currently, I would estimate the number to have stabilised at approximately 80% for the period after 2013, when publications based on microarray data stagnated.
The focus in this thesis research was mainly on microarray gene expression data, because the study started before NGS data were widely available. Generating primary data is expensive, and to reach the statistical power needed to identify significantly expressed oncogene outliers in a subset of samples, one needs to resort to a meta-analysis of studies with small sample numbers. To this end gene expression profiling studies are performed in a robust manner using meta-datasets such as Oncomine [22] and GeneSapiens (Medisapiens) [95]. The Oncomine database contains data on a variety of microarray platforms, while the GeneSapiens database contains data on different versions of Affymetrix platforms. Several other meta-analyses have been reported in the literature that were based on integrating publicly available data[33,96-98].

Affymetrix arrays are single channel arrays, and the platform employs a set of probes to measure the expression of each gene[99-102]. Probe sets contain two types of probes consisting of 25-nucleotide-long perfect match probes (PM) and mismatch probes (MM), together making a probe pair. Perfect match probes are designed to match the target sequence of a particular gene. Even though each probe is unique, some cross-reaction between homologous sequences may occur. Mismatch probes are the same as perfect match probes, except that the 13th position of the sequence is altered with a base mismatch. The hybridization process is very sensitive to

![Figure 1. Line graph indicating the number of articles identified in PubMed with a search including one of the two keywords “microarray” or “genome sequencing” in their articles.](image)
mismatches, and a single base mismatch is enough to also disrupt other results. The mismatch results inform about the background signal level that must be subtracted from the actual biologically meaningful signal measured. Data processing in the GeneSapiens database is based on Microarray Suite 5.0 (MAS5), which utilizes mismatch probes. On Affymetrix chips, the probes are randomly scattered so as to avoid positional hybridization artefacts. To determine the expression level of a single gene, probe level readings are summarized for each transcript. One can generate custom Chip Description Files (CDF) by dropping probes mapping to more than one transcript, leading to a new CDF file that generally improves the quality, reliability and reproducibility of the data, and thus the results of related gene expression studies[100]. Although Affymetrix provides a standard method for summarizing the probe level intensity readouts in a single gene transcript score, many approaches are available, such as the brainarray CDF file, to mention one of the most popular [91,101,103], that offer a more robust way of mapping probe sets to a single gene transcript. A recent study examining inconsistencies in large pharmacogenomic studies concluded that gene expression data are currently fairly reproducible across two separate laboratories [43].

2.4.4 Methods for analysing oncogene outlier expression

One of the first studies in prostate cancer to use microarray gene expression profiling data to identify oncogenes was performed by Tomlins et al. [34]. They hypothesized that both gene rearrangements and amplifications result in the overexpression of an oncogene, and that these should be evident in DNA microarray data as outlier profiles. This means a subgroup of cases with high levels of expression, which would not necessarily be detectable by traditional applied bioinformatic analysis approaches searching for uniform differential expression based on means and medians. In the majority of cancer types, heterogeneous patterns of gene expression are observed, often within a small subgroup of patient samples showing elevated expression. This renders traditional method repurposing approaches such as the t-test unsuitable for this type of analysis. Tomlins et al. developed the cancer outlier profile analysis (COPA) method and applied it to the Oncomine database [22], comprised of 132 gene expression datasets representing 10,486 microarray experiments [22]. Since then, some pure bioinformatic oncogene outlier analysis methods have been developed to identify oncogenes from both single study settings and large-scale integrated datasets,
Review of the literature

namely outlier sums (OS), the outlier robust t-test (ORT), and the maximum ordered subset t-statistic (MOST) [104-106]. These oncogene outlier analysis methods confirmed earlier findings that high-level gene expression in small subgroups of patients may highlight potential cancer-causing oncogenes.

Gene expression analysis studies with a small number of samples in relation to the number of genes suffer from the “curse of dimensionality” [107]. This is an issue of “large \( p \), small \( n \)”, where the number of experimental measurements \( p \) is far greater than the number of independent samples \( n \). When the number of samples in, for example, a repeated experiment is small, statistical estimates of metrics such as the mean from such draws involving small sample numbers is likely to be biased. The estimated values are very likely to deviate far from the true distribution parameters. The likelihood of such deviations increases with the number of different measurements in a single independent sample. This makes it difficult to reliably estimate parameters when multiple hypotheses are tested simultaneously, and raises the issue of the proportion of false positives in any of the results based on a given statistical test involving a small \( n \). For example, researchers performing differential gene expression analysis perform multiple testing correction, because a statistical test is calculated across thousands of genes and a small number of samples, making statistical confidence scores such as the \( P \)-value inadequate. One issue that is sometimes ignored is that multiple testing correction methods were developed for repeated measures occurring in industrial production. However, their application to several high-throughput data analysis approaches in medical research has been a point of discussion in the literature[107,108]. Several oncogene outlier analysis methods are non-parametric, and thus require some permutation-based approaches to generate \( P \)-values. Most often, the power of these methods has been assessed by simulation studies. Another common approach is to evaluate whether the method can identify previously known oncogenes and enriched pathways among the candidate outlier genes. The capabilities and experimental setup for some of the outlier analysis methods indicate that there is room for the development of more robust methods.
2.5 Stepwise validation of oncogene outlier hits using HT-RNAi screening

Elbashir et al. defined RNA interference (RNAi) as a biological phenomenon in which small double-stranded RNA (dsRNA) molecules present in the cytoplasm of a cell lead to the destruction of cognate mRNAs[109]. Especially if the oncogene blocked using siRNAs is a known drug target, it is possible to repurpose an already approved drug to treat a new disease. The phenotypic results from high-throughput RNA interference (HT-RNAi) can be monitored by assaying for specific endpoints, such as promoter activation, cell proliferation, survival and death. As with most high-throughput genomic screening technologies, data processing and analysis methods are required to handle the unique aspects of data normalization and statistical processing[80,110].

HT-RNAi screens have shown considerable utility in oncogene validation studies and have been utilized for validations based on several “omics” oncogene discovery approaches (Figure 2) [87]. Traditionally, HT-RNAi libraries have been generated by both academic and commercial suppliers and tend to contain thousands of genes without any pre-filtering. The results based on un-filtered large-scale RNAi libraries are usually dominated by a high number of false positives due to off-target effects. The schematic pipeline for systematic selection of hits based on several “omics” analyses, as described in Figure 2[87], has become popular, since it helps to eliminate the majority of the false positives that would otherwise occur in HT-RNAi screens using large-scale un-filtered libraries.
2.5.1 Limitations of HT-RNAi technology as a validation tool

The key challenge faced by scientists using HT-RNAi technology is to determine whether the target gene identified from an RNAi screen is responsible for the phenotype[111]. In recent years, the application of HT-RNAi technology to functional genomic studies has led to an increase in the number of “hit lists (signatures)”, or in other words, genes that elicit a positive response in various functional assays. As with most large-scale “omics” studies, these studies tend to contain a high number of false positives and negatives that could be caused by the quality of the reagents, the experimental assay setup[112], or data processing and analysis issues. Recent studies have shown that it is possible to eliminate off-target effects by reagent design alone[111]. The actual process of demonstrating that the knockdown of a hit gene is directly responsible for the phenotype involves performing a so-called “rescue experiment” [15]. Rescue experiments often involve reversing the phenotype induced by the siRNA by re-expressing an siRNA-resistant cDNA that has mismatches with

Figure 2. Stepwise selection of anti-tumoural therapies based on oncogene targets identified from several “omic” analyses. Adapted from Alberto et al., Mol Cancer, 2010[87]
the siRNA and cannot be targeted for degradation. With a large number of hits, performing rescue experiments for all hit genes is virtually impossible, and in many cases, overexpression of cDNA has aberrant effects, which makes this approach difficult to implement.

When designing an HT-RNAi experiment, one needs to take into account several factors, including the following:

a) **Choice of the assay system**

The selection of the model system is crucial for the successful implementation of an RNAi experiment. The assay system must be robust, reproducible and affordable. Most laboratories use human tumour cell lines for RNAi screens, but, depending on the phenotype examined, the use of other model systems such as conditionally reprogramed cell lines (CRCs) might be considered.

b) **Choice of the RNAi library**

The choice of the library has been a point of debate in many studies to date due to the many formats that exist on the market[8,111]. Library selection is usually driven by a single question: to pool, or not to pool? If a pooling strategy is chosen, screens have mostly been limited to vector-based libraries [8]. Another issue to consider is the scope of the experiment, i.e. genome-wide screens or specific subsets of the genome, typically based on functional categories or oncogene outliers, which are less expensive.

c) **Choice of the screening approach**

Even though the choice of the screening approach tends to mainly be determined by the library format, some additional issues need to be considered, such as non-pooling approaches requiring the library to be administered in 96-well or 384-well plates. To optimally exploit the benefits of this arrangement of the assay, some level of automation is needed. Conversely, a pooling approach allows the entire library to be used as single, or a small number of pool(s), but requires considerable effort in data processing.

d) **Choice of controls**

Efficient utilization of HT-RNAi data greatly depends on the positive and negative controls placed on each plate. They are useful in the quality control assessment of each screen. The choice of controls on each screen is usually pre-determined by the library chosen, which for some model organisms tends
to be ineffective. This makes controls not behave as expected, thus narrowing
the dynamic range between the positives and the negatives. Controls are
essential for the calculation of quality control metrics such as the z-factor[113]
and the strictly standardized mean (SSMD)[114] (see Methods section).

2.6 Meta-analysis of gene expression and drug testing data

The ability to grow established (ATCC) cancer cell lines, drug-resistant cancer cell
models and ex vivo patient cancer cells in primary cultures has made a tremendous
contribution to cancer research and has also helped to improve our capacity to
understand the oncogene addiction mechanisms that could be targeted with a
matching drug[13,37,59]. Advances in high-throughput drug testing have helped to
uncover novel ways of defining cancer patient subtypes and have accelerated the
delivery of individualized therapies to patients[13,59]. High-throughput drug testing
has become a common practice in many research centres around the
world[13,37,42,59]. The drug testing approach has been successful in the treatment of
cancer where a single drug that shows favourable response is associated with a known
cancer oncogene, such as the drug Imatinib in BCR-ABL gene fusion-positive
cells[115-118]. Most drug screening approaches involving oncogene hits have been
based on kinase pathways constantly implicated in cancer pathogenesis. One example
in leukaemia is the tyrosine kinases family of genes, consisting of oncogenes such as
FLT3, ABL and c-KIT. These are abnormally expressed in several leukaemia
subtypes, including chronic myeloid leukaemia (CML)[119], chronic
myelomonocytic leukaemia (CMML) [120-123], other myeloproliferative neoplasms
(MPN)[124-127], acute myeloid leukaemia (AML)[121,128-132], acute
lymphoblastic leukaemia (ALL) [133-138], and chronic lymphocytic leukaemia
(ALL)[139-142].

The current paradigm of treatment options based on matching kinase inhibitors to a
patient’s disease is only helpful for a handful of patients due to the lack of knowledge
of the specific kinase pathways activated in the majority of cancers. Competing
strategies based on deep sequencing have been tested in many studies, leading to
important discoveries of activated pathways. However, most of these studies have not
been able to pin down specific mutations in key kinase genes[143-146]. Already in
2007, Greenman et al.[143] showed that most somatic mutations are likely to be
"passengers" and do not contribute to oncogenesis. They found some evidence of mutation-driver association links for approximately 120 genes that warranted further characterization to ascertain the functional mechanisms of these mutated genes. These results continue to shed light on the complexity of predicting therapeutically relevant kinase targets and matching them with the relevant inhibitors.

The drug testing approach is emerging as an important tool based on identifying inhibitors capable of inhibiting the growth of cancer cells in a set of screened cancer samples. These drug-testing studies are then followed by bioinformatic prediction approaches to help predict the gene targets underlying a set of inhibitor sensitivity profiles. These approaches utilize existing knowledge of known drug targets from publicly available drug annotation databases (Table 1) [73,147,148]. Using the drug targets for drugs that are identified as being sensitive in a cohort of cancer samples enables a direct match of key drug targets and signalling pathways commonly activated in a set of cancer samples. Performing large-scale integrated gene expression analysis and gene-drug functional association studies enables the utilization of drug testing data in accelerating our understanding of the molecular aetiologies of cancer. Recent articles presenting pharmacogenomics studies have demonstrated that it is possible to apply some methods formerly used for gene expression to the new drug testing datasets[37,149]. Even though some successful gene expression method repurposing approaches have already been used in drug testing studies [35,37], careful assessment of the performance of new methods adapted from gene expression studies is needed.

2.6.1 Inconsistency and reproducibility challenges in drug testing studies
The reproducibility of scientific work has recently become a hot topic in many scientific articles, and many suggestions have been presented to address the problem. Large-scale pharmacogenomics studies have been cited in some articles exposing the lack of consistency and reproducibility[44,45,149]. For example, Haibe-Kains et al.[149] highlighted inconsistencies among two large-scale pharmacogenomics studies, which they attributed to the lack of drug response measurement standards and differences in pharmacological assays. As a standard practice, most pharmacogenomics studies use metrics such as IC$_{50}$ or EC$_{50}$ (half-maximal inhibitory/effective concentration), as well as the area under the curve (AUC)[36,37].
Recently, it has been shown that multi-response parameters could offer improved standardized drug response scoring metrics, as demonstrated by Yadav et al.[75] and Fallahi-Sichani et al. [150]. However, Weinstein et al.[44] pointed out that the pharmacological assay used by the Cancer Genome Project (CGP) (the CellTiter 96 AQueous One Solution Cell Proliferation Assay from Promega) measures metabolic activity in terms of a reductase-enzyme product after a 72-hour incubation of cells with a drug. The assay of the Cancer Cell Line Encyclopedia (CCLE) (the CellTiter-Glo assay from Promega) measures metabolic activity by assessing levels of the energy-transfer molecule ATP after 72 hours of incubation. Both experiments lead to some drug response measurements, but they are not expected to mirror each other across all cell lines and drug types, even if all other factors remain constant. Furthermore, they observed that many variables could affect drug response scoring from such assays. For example, effects based on batches of foetal bovine serum, the storage conditions of drugs, the relative decay of drugs due to a long storage time, cell culture conditions, and the coating on the plastic culture wells could affect drug response scoring. However, they raised an interesting question regarding the usefulness of drug testing data by asking whether the drugs would cluster into two separate groups on the basis of response data from two different projects. It remains to be seen how useful drug testing data will be in advancing the efficient delivery of individualized treatments to patients.
3. **Aims of the study**

The overall objective of this study was to develop pure and applied bioinformatics and statistical tools for the analysis, mining and modelling of large-scale gene-expression and drug testing datasets.

The specific aims were to develop:

1. Statistical models for identifying oncogenic expression profiles from integrated large-scale microarray datasets to provide information on the molecular biology of cancer, biomarker discovery and facilitate the identification of therapeutic targets;

2. Methods and statistical approaches to identify cancer subtypes from large-scale molecular profiling and drug testing datasets;

3. Tools and analysis pipelines to improve the quality and reproducibility of results from HT-RNAi and drug testing studies based on high-throughput screening technologies;

4. Tools and analysis pipelines to improve the consistency (agreement) of results from pharmacogenomic studies.
4. Materials and methods

4.1 Datasets

4.1.1 Gene expression datasets

The dataset (Table 3) used in publications I and II was extracted from the publicly available GeneSapiens (MediSapiens) database[95]. The GeneSapiens database contains microarray gene expression data from several Affymetrix platforms. The majority of studies in the database have been downloaded from publicly available microarray gene expression data repositories such as GEO[20] and ArrayExpress[21]. The data processing methods and normalization tools are extensively described in the original papers of Kilpinen et al. [91,95]. Specifically, all CEL files were preprocessed with the Microarray Suite 5.0 (MAS5) algorithm, but not with the Robust Multichip Average (RMA) method [151]. The database contained over 9783 samples covering approximately 175 different types of healthy and cancerous tissues. The data were meticulously annotated by a team of several researchers to obtain detailed clinical histological annotations for each sample.

The study reported in publication IV examined the reproducibility of gene expression data by comparing Affymetrix microarray data from the Cancer Genome Project (CGP) [37] and the Cancer Cell Line Encyclopedia (CCLE) [36]. The CGP dataset contained 727 cell lines compared to 917 cell lines available in the CCLE dataset. Altogether, 471 cell lines were tested in both studies, but on different Affymetrix microarray platforms (Affymetrix GeneChip HG-U133A for CGP and HG-U133Plus2 for CCLE). The raw CEL files for the CCLE dataset were downloaded from the GEO database (GSE36133) and that for the CGP dataset from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/). These datasets were then processed using the aroma.affymetrix Bioconductor package[152].

4.1.2 Drug testing datasets

The drug testing datasets (Table 3) used in publications III and IV were derived from experiments performed in-house under the personalised medicine grand challenge projects carried out by researchers at the Institute for Molecular Medicine Finland (FIMM), haematologists at the Helsinki University Central Hospital (HUCH)
Comprehensive Cancer Center and Helsinki Urological Biobank Finland. The drug testing data used in publication III were screened at the FIMM High Throughput Biomedicine Unit (FIMM-HTB). This dataset contained two replicate measurements per drug tested on VCaP and LAPC4 prostate cancer cell lines that were tested with 306 Food and Drug Administration (FDA) approved and investigational drugs. Importantly, each drug was screened at five different concentrations to enable the generation of cell viability dose–response curves. Additionally, in publication IV, three drug testing datasets (Table 3) were used, which were acquired from the Cancer Genome Project (CGP), the Cancer Cell Line Encyclopedia (CCLE) and data from FIMM. The CGP dataset contained drug response profiles for 138 drugs, the CCLE contained response profiles for 24 drugs, and the FIMM drug testing dataset contained 308 drug response profiles covering 106 cancer cell lines. The FIMM dataset had 45 compounds in common with CGP and 14 with the CCLE, whereas 15 compounds were in common between the CCLE and CGP datasets.

Table 3: Datasets used for gene expression and drug testing analysis

<table>
<thead>
<tr>
<th>Tumor Type</th>
<th>Data type</th>
<th>Source</th>
<th>Link</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 tumor types</td>
<td>Affymetrix microarray gene expression</td>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td><a href="http://www.broadinstitute.org/ccle">www.broadinstitute.org/ccle</a></td>
<td>Barretina et al.</td>
</tr>
<tr>
<td>36 tumor types</td>
<td>Drug Sensitivity testing data (24 drugs)</td>
<td>Cancer Cell Line Encyclopedia (CCLE)</td>
<td><a href="http://www.broadinstitute.org/ccle">www.broadinstitute.org/ccle</a></td>
<td>Barretina et al.</td>
</tr>
</tbody>
</table>

4.2 Data preprocessing and quality control

4.2.1 Raw gene expression data processing

Raw Affymetrix gene expression data acquired from public repositories were processed using the Bioconductor suite of packages[152] in the R statistical programming language[153]. Specifically, the aroma.affymetrix[154] Bioconductor package was used to process the raw CEL files. The raw data that we processed later outside of the GeneSapiens database were normalized using the Robust Multichip Average method (RMA)[151]. Custom Chip Definition Files (CDF) from brainarray[100] were used to filter out uninformative probesets. Results for several transcripts mapping to a single gene were summarized using Tukey’s Biweight method[155].
4.2.2 Processing and quality control of raw drug testing data

Raw drug testing data used in publication III were preprocessed using R statistical software. The raw Cell Titer-Glo (CTG) viability assay data outputs from the Pherastar FS plate reader (Ortenberg, Germany) were converted to matrices and visualized as heatmaps and well scatter plots as a normal quality control step. Each screen contained five 384-well plates seeded with cells and incubated with 306 drugs and controls. The individual assay plates included 16 negative controls with dimethyl sulfoxide (DMSO) only and 8 positive control wells with 100 μM benzethonium chloride. In addition, 19 of the remaining wells were left blank and 35 wells contained cells only. The drugs were plated at five different concentrations in 10-fold dilutions covering a 10,000-fold concentration range.

Quality control (QC) analysis was performed for each plate before proceeding to downstream analysis. The calculation of quality control metrics was regarded as an essential step in the processing of drug testing data, because the percent inhibition metrics largely depend on the average signal from the controls. I used the Z’-factor (Z’)[113] and the strictly standardized mean difference (SSMD) [114]. Acceptable values of SSMD and the Z’-factor depend on the strength of the controls on the plate. Each plate included two types of controls placed in different wells, namely negative control (DMSO only) and positive control wells (containing 100 benzethonium chloride), applied at a toxic dose to achieve the maximum viability decrease in cells. The fundamental hypothesis of the quality control metrics is that the positive and negative controls come from two independent distributions. The Z’-factor can be represented as:

\[
Z' = 1 - \left\{ 3 \left( \sigma_{h.c} + \sigma_{l.c} \right) / |\mu_{h.c} - \mu_{l.c}| \right\},
\]

where \( \mu = \text{mean} \), \( \sigma = \text{standard deviations} \), \( h.c = \text{high.control} \), which represents the signal detected from control wells (DMSO) leading to no cell death, while \( l.c = \text{low.control} \) (positive control wells containing 100 μM benzethonium chloride). We also calculated QC scores using the SSMD quality control metric. The SSMD offers a robust assessment of the quality of the screen, and recent work by Birmingham et al. demonstrated that this metric is a less conservative indicator of quality than the Z’-factor[156]. The SSMD formula can be represented as:
Materials and methods

\[ SSMD = \frac{(\mu_{h.c} - \mu_{l.c})}{\sqrt{(\sigma^2_{h.c} + \sigma^2_{l.c})}}, \]  

where \( h.c = high\_control \) and \( l.c = low\_control \), used as described above.

### 4.3 Outlier gene expression analysis

In publication I, outlier gene expression analysis was performed using the GeneSapiens gene expression database developed by Kilpinen et al. [95]. The outlier analysis covered a total of 16,868 human genes, with each gene represented by a different number of normal and cancer samples in the database. As the compositions of microarrays are continuously updated, as well as the custom CDF file mappings to enable the incorporation of new genes with improved target sequences, it is evident that combining data from different microarray generations of the same microarray technology will result in largely varying samples per gene. Our initial attempt at finding outlier genes started by applying the traditional t-statistic and three other outlier expression analysis methods conceived in previous studies, including cancer outlier profile analysis (COPA) [34] the outlier sum (OS) statistic[105] and the outlier robust t-statistic (ORT)[106].

However, the problem at hand resembled income inequality problems in economics, and there are well-established formulas that could be efficiently adapted and modified to suit cancer research questions of outlier analysis in large-scale gene expression datasets. One formula developed by Nobel prize winner Amartya Sen[157] for calculating an income index for each country by quantifying the number of people living below the poverty line was adapted and applied to outlier expression analysis. Similar methods are very popular in development economics, and they have been applied to address several problems in economics[158]. To adapt the algorithm developed by Amartya Sen for gene expression analysis, we inverted the original question by asking “in how many samples from the same body part is a gene X expressed above a fixed cut-off threshold?” This index is determined as a robust proportion of outlying samples given that every gene is represented by an adequate number of samples.

In publication I, the gene tissue index (GTI) method was introduced and systematically compared with existing methods, i.e. the t-statistic, COPA, OS and ORT. Here, we let \( x_{ij} \) be the expression values for genes \( j = 1,2, \ldots, p \) and samples \( i \).
Materials and methods

We assumed that the samples were obtained from two groups \((k = 1\) and \(k = 2\)), where \(n = n^{(1)} + n^{(2)}\). In our case, \(n^{(1)}\) represents the number of samples from normal group and \(n^{(2)}\) represents the number of samples from the cancer group. Let \(C_k\) be the set of indices of the observations in group \(k\), for \(k = 1\) and 2.

### 4.3.1 The t-statistic

The standard unpaired sample t-statistic for gene \(j\) is defined as:

\[
T_j = \frac{x_j^{(2)} - x_j^{(1)}}{s_j} \sqrt{\frac{n^{(1)} n^{(2)}}{n}},
\]

where \(x_j^{(k)}\) is the mean expression of samples for gene \(j\) in group \(k\),

\[
x_j^{(1)} = \frac{\sum_{i \in C_1} x_{ij}}{n^{(1)}}, \quad x_j^{(2)} = \frac{\sum_{i \in C_2} x_{ij}}{n^{(2)}},
\]

\(s_j\) is the pooled within-group standard deviation of gene \(j\),

\[
s_j^2 = \frac{\sum_{i \in C_1} (x_{ij} - x_j^{(1)})^2 + \sum_{i \in C_2} (x_{ij} - x_j^{(2)})^2}{n-2}.
\]

The t-statistic assumes that all samples in the disease group are uniformly overexpressed, which is not the case for outlier gene expression studies.

### 4.3.2 Cancer outlier profile analysis (COPA)

The second method compared was the COPA statistic. This is one of the most-cited gene expression outlier statistics and is based on the \(r^{th}\) percentile of the standardized expression values \(\bar{x}_{ij}\) for disease samples, defined as:

\[
q_r (\bar{x}_{ij}; i \in C_2) = \frac{q_r (x_{ij}; i \in C_2) - \text{med}_j}{\text{mad}_j},
\]

where, using \(r = 75, 90\) or \(95\), the \(r^{th}\) percentile of the disease samples is \(q_r (x_{ij}; i \in C_2)\).

Observations for gene \(j\) are standardized by subtracting the median \(\text{med}_j\) from each expression value \((x_{ij})\) divided by the median absolute deviation \(\text{mad}_j\),

\[
x_{ij} = \frac{x_{ij} - \text{med}_j}{\text{mad}_j}, i = 1, \ldots, n, j = 1, \ldots, p,
\]

where \(\text{med}_j\) is the median and \(\text{mad}_j\) is the median absolute deviation of gene \(j\)’s expression values.

\[
\text{med}_j = \text{median}_{i=1,\ldots,n} (x_{ij}), \quad \text{mad}_j = 1.4826 \times \text{median}_{i=1,\ldots,n} (|x_{ij} - \text{med}_j|),
\]
where the product of \( mad_j \) and the constant 1.4826 is approximately equal to the standard error for normally distributed random variables.

Compared to the t-statistic, COPA intuitively replaces the normal sample average with the all-sample median \( med_j \), the sample standard error \( s_j \) with the median absolute deviation \( mad_j \), and the disease sample average with the \( r^{th} \) percentile \( q_r(x_{ij} : i \in C_2) \).

### 4.3.3 Outlier sums (OS)

The outlier sums metric was introduced as an improvement to the COPA statistic. Here, the OS statistic\[104-106\] was proposed to replace the \( r^{th} \) percentile with a sum over the outlier samples from the disease group above a given cut-off. OS was designed to lower the false discovery rate (FDR) of COPA, as noted by Wu et al. OS standardizes each expression value of gene \( j \) \( (x_{ij}) \) by dividing the result of \( (x_{ij} – med_j) \) by \( mad_j \). However, only expression values above a given cut-off are used to obtain the final score.

\[
OS \, score_j = \frac{\sum_{i\in O_j}(x_{ij} – med_j)}{mad_j},
\]

where \( O_j \) is the set of outlier samples from the disease group defined by the following heuristic criterion:

\[
O_j = \{i : i \in C_2, x_{ij} > q_{75}(x_{mj} : m = 1, \ldots, n) + IQR(x_{mj} : m = 1, \ldots, n)\},
\]

where \( m \) refers to samples \( 1, 2, \ldots, n, n_1 + 1, \ldots, n \), \( q_{75} = \text{75}^{th} \text{ quantile} \) \((q)\) and the interquartile range \((\text{IQR})\): \( IQR(x_{ij} : i = 1, \ldots, n) = q_{75}(x_{ij} : i = 1, \ldots, n) – q_{25}(x_{ij} : i = 1, \ldots, n) \).

### 4.3.4 Outlier robust t-statistic (ORT)

The outlier robust t-statistic\[104,106,159\] is a direct robust generalization of the two-sample t-statistic. With ORT, the sample mean is replaced with the median and the squared difference with the absolute difference. The overall median used as a common estimate for the two group medians was suggested to be inefficient, since we know that the normal and disease samples are different. The ORT statistic was proposed to replace the overall median estimate used in calculating the COPA and OS score with a median calculated from the group median-centred expression values.

\[
\left| x_{ij} – med_j^{(1)} \right|, i = 1, \ldots, n; \left| x_{ij} – med_j^{(2)} \right|, i = n_1 + 1, \ldots, n,
\]
Materials and methods

where \( \text{med}_j^{(1)} \) and \( \text{med}_j^{(2)} \) are the sample medians for normal and disease groups. ORT is then described as

\[
ORT = \frac{\sum_{i \in R_j} (x_{ij} - \text{med}_j^{(1)})}{\text{median}\{x_{ij} - \text{med}_j^{(1)} | i \in C_1, x_{ij} - \text{med}_j^{(2)} | i \in C_2\}}, \quad j = 1, \ldots, p ,
\]

where \( R \) is the set of outlier disease samples for gene \( j \) defined by

\[
R_j = \{i : i \in C_2, x_{ij} > q75(x_{mj}; m = 1, \ldots, n_1) + IQR(x_{mj}; m = 1, \ldots, n_1)\}, \quad (10)
\]

where \( m \) refers to samples \( 1, 2, \ldots, n_i \).

It should be noted that only samples in the normal group were used to estimate outliers when calculating an ORT score.

4.3.5 Gene tissue index (GTI)

In publication I of this thesis, the GTI method is introduced. This method was designed to produce an index quantifying the proportion of outlying samples beyond a given cut-off. A similar approach has been applied in economics, and the cut-offs are assigned based on global income estimates. For outlier gene expression studies, there was no globally defined standard cut-off per gene for each normal tissue sample. For this, a cut-off \( (B) \) based on the expression of gene \( j \) among all samples \( (n) \) was defined. These samples were obtained from one anatomical part, such as the breast of individuals with a normal \((k = 1)\) and cancer \((k = 2)\) state. We then asked whether the proportion of samples above the cut-off was larger than it should be. Our choice of \( B \) was the standard statistical outlier cut-off \((q75 + IQR)\). We proposed the following score, which weighs the proportion of outliers by a robust measure of how outlying the outliers are in a single group:

\[
\text{GTI}_j^{(k)} = \frac{T_j^{(k)}}{n_j^{(k)}} \times \frac{(A_j^{(k)} - B_j)}{A_j^{(k)}}, \quad (11)
\]

where \( T_j^{(k)} \) is the number of samples with expression values above the cut-off (number of elements in set \( O_j^{(k)} \)), \( n_j^{(k)} \) is the total number of samples in group \( k \) and \( A_j^{(k)} \) is the average expression of the samples above the cut-off for gene \( j \). For the interquartile range \((IQR)\), we write

\[
IQR(x_{ij}; i = 1, \ldots, n) = q75(x_{ij}; i = 1, \ldots, n) - q25(x_{ij}; i = 1, \ldots, n).
\]

Expanding the definition of GTI and substituting our choice of \( B \), we get
Materials and methods

\[
GTI_{j}^{(k)} = \frac{x_{j}^{(k)} - \left( q_{75}(x_{ij}:i=1,\ldots,n) + IQR(x_{ij}:i=1,\ldots,n) \right)}{\bar{x}_{j}^{(k)}},
\]  

where \(x_{j}^{(k)}\) is the mean of “outlier samples” in group \((k = 1 \text{ or } k = 2)\) for genes \(j = 1, \ldots, p\),

\[
\bar{x}_{j}^{(k)} = \frac{\sum_{i \in O_{j}^{(k)}} x_{ij}}{n},
\]

where the set \(O_{j}^{(k)}\) consists of the outliers in group \(k\). The set \(O_{j}^{(k)}\) is defined using the following criterion:

\[
O_{j}^{(k)} = \{i : i \in C_{k}, x_{ij} > q_{75}(x_{mj}:m = 1, \ldots, n) + IQR(x_{mj}:m = 1, \ldots, n)\},
\]

where \(m\) refers to samples \(1, 2, \ldots, n^{(i)}, n^{(i)} + 1, \ldots, n\).

We calculated the actual GTI scores for each group \(k\) and gene \(j\) multiplied by one hundred, as this made them more readable.

Finally, the index per gene was determined as a direct association between two groups defined by \(GTI_{j} = GTI_{j}^{(2)} - GTI_{j}^{(1)}\), where 2 and 1 represent the grouping. The index \(GTI_{j}\) can be a large positive number if there are outliers in group 2 or a large negative number if there are outliers in group 1. All samples (cancer and normal combined) were used to determine the cut-off point for each gene. As in the existing methods, we used permutations to estimate the null distribution for GTI and the \(P\)-values.

4.4 High-throughput screening normalization methods

High-throughput screening (HTS) data might contain unwanted variation in the raw data across plates or within a single plate. In drug testing experiments, normalization across plates is achieved by calculating percent inhibition values for all data points in one screen. Although this step is necessary for performing downstream drug response scoring, it cannot correct for within-plate systematic effects. Systematic errors due to row, column and edge effects do occur in high-throughput screening experiments. In publication III, I examined the B-score HTS normalization method, together with an adaptation of the Loess method. The B-score method is simply a ratio of the residual values calculated from the median polish to the median absolute deviation (MAD) of the sample. Brideau et al. have described the implementation of the B-score approach[77], which I also describe in publication III.
Materials and methods

Publication III introduces a local regression fit approach based on fitting a smooth surface to each plate. This was followed by well correction by subtracting from or adding to the original value depending on the sign of the deviation, as shown below. A span value was applied to determine the number of wells in one location that could be used to estimate Loess fitted values. Low values of span enable the identification of small areas with an uneven concentration of hits or non-hits. Setting high span values leads to over-fitting of the data and less accurate detection of small areas with systematic within-plate effects.

Given plate $p$ with $n$ rows and $m$ columns, which I represent as an $n \times m$ data matrix, where $x_{ij}$ is the raw signal readout for row $i$ and column $j$, we calculate the Loess-fit result $\hat{x}_{ij}$ as follows:

$$\hat{x}_{ij} = x_{ij} - (\text{loess.fit}_{ij} - \text{median(loess.fit}_p)), \quad (14),$$

where $\text{loess.fit}_i$ is the value from the loess-smoothed data matrix for row $i=1,\ldots,n$ and column $j=1,\ldots,m$, $\text{median(loess.fit}_p) = \text{median}_{i=1,\ldots,n,j=1,\ldots,m}(\text{loess.fit}_ij)$. The $\text{loess.fit}$ values were calculated using the Loess function in the stats package of R software.

### 4.5 Drug testing analysis methods

#### 4.5.1 Percent inhibition

In publications III and IV, percent inhibition values were calculated for each drug as a scaling step to enable the direct comparison of data from different plates. Given plate $p$ and plate raw signal readout $x_{ij}$, the percent inhibition ($I_{ijp}$) for the value at row $i$ and column $j$ is as follows:

$$I_{ijp} = \left(\frac{\mu_{h.c} - x_{ij}}{\mu_{h.c} - \mu_{l.c}}\right) * 100, \quad (15)$$

where $\mu_{h.c}$ is the mean of high.control sample value on plate $p$ and $\mu_{l.c}$ is the mean of the low.control sample values on plate $p$.

#### 4.5.2 Drug response scoring

In publications III and IV, drug response scoring was performed using standard curve fitting algorithms and a modified area under the curve metric. For each drug, we obtained a vector of drug concentration values and percent inhibition values. The values were then sorted by the concentration values in descending order. A four-parameter sigmoid dose–response curve was fitted to this data set using the
Levenberg-Marquardt[160] optimization algorithm to obtain IC50, SLOPE, and top and lower asymptotes. The formula for fitting the four-parameter sigmoid curve-fit model is as follows:

\[
y = d + \frac{(a-d)}{1+(10^{(b(c-x)))}},
\]

where \(a, b, c\) and \(d\) are the parameters to be fitted and \(x\) is the concentration. Given that:

- \(a\) = maximum response
- \(b\) = slope
- \(c\) = dose IC50 (half maximal inhibitory concentration)
- \(d\) = minimum asymptote

I used the nonlinear least squares (nls) model-fitting[161-163] function implemented with the \textit{minpack.lm} R package. It should be noted that the nonlinear least squares model fitting approach is based on an iterative process in order to identify the optimal parameter estimates. When the doses of a drug are relatively sparse and few percent inhibition points fall within a designated response window, such as between 0% and 100%, the model overfits the data, leading to false dose–response curve-fit parameters. The nonlinear least squares model fitting approach requires the setting of proper model starting parameters for all the four parameters to be estimated. Here, we set the model starting parameters as \(b = 1, d = 0\), while \(a\) and \(c\) were estimated using the linear regression model described in the \textit{drc} R package (self-starter function) [164]. The model parameters estimated from the curve-fitting step were then used to calculate the drug sensitivity score (DSS) of Yadav \textit{et al.} [165]. The DSS provides a more standardized way of comparing drug response profiles obtained from several large-scale drug-testing studies.

4.6 Data visualizations

Our outlier analysis method developed in publication I produces a long list of genes, including genes that are over-expressed together with those lost in a subgroup of samples. Here, the differential correlation map “corDiff-map” statistical visualization tool was developed to compare hundreds of correlations in the disease state versus those in the normal state. Previous studies on differential correlation analysis of time-resolved transcriptomic data had been based on implementing group average metrics,
and the tools developed in these studies such as weighted gene co-expression network analysis (WGCNA) [166,167] and differential co-expression profile analysis (DCp) [168], ignored heterogeneity between sample series. Recently, a newly published tool called Dynamically Co-expressed Neighborhoods (DCeN) [169] has provided a more advanced approach to differential correlation based on the individual correlations approach to time-series data. The group averaging approach used by WGCNA and DCp methods loses information on the strong pairwise correlation within each individual group. The corDiff-map tool employs an approach similar to that used by the DCeN tool. However, DCeN and corDiff-map tools differ in the metrics used to score the differences between two correlations, and the corDiff-map approach has not yet been tested on time-series data. The hypothesis behind the corDiff-map approach is that among the genes identified in outlier gene expression studies, some are driver genes while the majority are passengers. We assumed that genes activated in a given pathway might be co-regulated and thus show strong co-expression. To construct a corDiff-map, we calculate correlations in one disease state, for example cancer, and in another disease state, such as among normal samples, all from one tissue. The order of the genes is selected from the first correlation map (corMap) and retained in all the other correlation maps. We then test whether there is a statistically significant difference between the correlations of the first correlation map (disease state 1) and those of the second correlation map (disease state 2).

In usual practice, when we assess correlation, we compare our correlation value against the hypothesis that there is no correlation $R = 0$. However, we can always test the hypothesis that the correlation value is equal to some other assigned value. This is the idea behind assessing the significance between two correlation values. Therefore, given correlation $R^{(k)}$ where $k = 1$ or 2, with $k$ representing disease states 1 and 2, we test the null hypothesis:

$H_0: R^{(1)} = R^{(2)}$

where

$H_0 = $ the null hypothesis,

$R^{(1)} = $ the correlation of a feature (gene or drug) among samples from disease state 1,

$R^{(2)} = $ the correlation of a feature (gene or drug) among samples from disease state 2.
Materials and methods

We need to estimate whether the difference between the two correlations is statistically significant. However, the sampling distribution of Pearson’s correlation coefficient $r$ does not follow a normal distribution. Fisher developed a transformation, “Fisher’s z transformation”[170], used to convert Pearson’s $r$ values to the normally distributed variable $z_{fk}$, which can be calculated using the formula below:

$$z_{fk} = \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right).$$  \hspace{1cm} (17)

After acquiring $z_{f1}$ and $z_{f2}$, we estimate the $z$ value of the difference as follows:

$$z = (z_{f1} - z_{f2})/\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)},$$  \hspace{1cm} (18)

which is normally distributed with a known standard error ($\sigma_{z_f}$) and where $n_1$ and $n_2$ are the sample sizes for the two sample groups representing different disease states. It is now straightforward to use the $z$ value to determine the level of significance ($P$-values). This visualizing tool requires more than 10 samples in each group to enable a robust estimation of the significance. Large sample sizes increase the power of the test to be able to reject the null hypothesis that the two correlations are not equal.
5. Results

In this thesis study, I developed a novel method for outlier gene expression analysis adapted from the analysis of economic data. This algorithm was compared with four existing outlier analysis methods and was then adapted to discovery studies of gliomas in publication I and prostate cancer in publication II. Next, I examined the impact of normalization methods on the reproducibility of high-throughput screening data. Finally, I presented applied bioinformatic analysis approaches and visualization tools for mining large-scale pharmacogenomic datasets.

The integration of large-scale pharmacogenomic datasets is both a challenge and an opportunity in bioinformatics with the increasing number of large biomedical datasets. The ability of biomedical researchers to produce large-scale high-throughput biomedical data is increasing beyond the capacity of those who analyse them. The need for pure and applied bioinformatic methods to deliver the promise of personalized medicine is becoming a reality at many research centres in the world[48]. As a first step in performing integrated biomedical data analysis, bioinformatic methods for the analysis and mining of molecular profiling and high-throughput screening (HTS) data have been developed in this thesis research (Figure 3).

![Figure 3: Schematic overview of the data integration model and the analytical tools developed in this thesis research.](image)
5.1 Comparison of outlier analysis methods for the analysis of gene expression data

5.1.1 Comparisons using a simulated gene expression dataset
To identify key cancer driver genes and drug targets from large-scale gene expression studies, we developed the Gene Tissue Index (GTI) method (see methods section 4.3.5 for details). Next, we performed a systematic comparison of the GTI with existing outlier expression analysis methods. First, for this comparison, simulation studies using the outlier robust t-statistic (ORT), outlier sums (OS) and cancer outlier profile analysis (COPA) were performed using a fixed outlier cut-off (q75 + IQR). This analysis also included the t-statistic, which requires no cut-off threshold in its calculation. The simulation study involved 1000 genes present in an equal number of normal and cancer samples. For the cancer samples, a constant value was added to the expression value to make the sample an outlier. We generated three datasets by varying the number of outlier samples (k = 1, 10, 20 and 30), and these were taken as the true positives (TP). The actual true positives (TP) and false positives (FP) were generated by performing 50 simulations. A P-value representing the probability of an identified gene being a false positive was determined as the proportion of genes with a score greater than that of the true positive. This process generated 50 P-values, and the true positive rate corresponding to a given false positive rate threshold was determined as the proportion of simulations identifying the true positive gene using the false positive threshold. The simulations were repeated for each of the methods compared in this study. The results were summarized as receiver operating characteristic (ROC) curves, and the size of the area under the curve was used to evaluate the performance of the methods. The results of this study demonstrated that most of the existing methods performed equally well, especially when the sub-population of outlier samples was very large. The ORT method showed poor performance for cases when the majority of the samples were detected as outliers. This type of outlier expression profile is usually recorded for biomarkers, while the most common outlier profile for oncogenes is close to the present study, with 10 outliers out of 30 cancer samples[34].

5.1.2 Comparison using a large-scale glioma microarray integrated dataset
Tumours of the central nervous system (CNS) consist of a diverse group of neoplasms that are derived from various different cell lineages. Gliomas refer to a brain cancer
Results

subtype that has its origin traced from the glial tissue. They represent approximately 80% of brain cancer cases and there is an enormous number of large-scale publicly available gene expression microarray datasets in GEO[20] and TCGA[11], as well as in the GeneSapiens database[95]. This study focused on two major subgroups of gliomas, anaplastic astrocytoma (WHO grade III) (74 samples) and glioblastoma multiforme (GBM, WHO grade IV) (353 samples). A reference dataset containing healthy CNS tissue samples profiled using the Affymetrix microarray platform was used in this study. The overlap of the top outlier genes identified by the GTI, COPA and the OS statistic were visualized using a Venn diagram (Figure 4).

![Figure 4: Venn diagram showing the overlap of the top 100 genes identified by each method. Only 49 genes were identified by all three methods, thus indicating the existence of many false positives or false negatives among the other genes not identified by all the methods.](image)

To assess the findings uncovered by the GTI method, a comprehensive study of the 29 GTI unique genes and a merged gene set containing the top 100 genes identified by each method was performed. A literature search to identify genes already associated with gliomas (see publication I, Tables 1 and 2) and protein expression mining based on the 29 unique GTI genes (publication I, Table 4) was carried out to assess the association with the biology of gliomas. The literature search revealed that COPA identified seven genes already known to be associated with the biology of gliomas, while the GTI identified six genes (see publication I, Table 1). GFAP was the only gene on the hit list for COPA but not for the GTI and, interestingly, GFAP is a known differentiation marker for normal cells of astroglial origin as well as a glioma marker, but not an outlier gene. According to a literature search to identify the commonly deregulated genes in gliomas out of 16,868 analysed genes, the GTI was able to
correctly rank genes such as CDKN2A and CDKN2B (with a high GTI score; publication I, Table 2) that were previously known to have a biological role in gliomas. Genes that are commonly lost in gliomas, such as PTEN, acquired a low GTI score. We further followed up the unique 29 GTI genes by analysing the protein expression profile using data available in the Human Protein Atlas database (HPA) [171]. We discovered that 19 out of the 29 genes contained immunohistochemical staining images deposited in the HPA database. Interestingly, 17 out of the 19 genes (90%) showed an outlier protein staining pattern in the HPA database. Most of the specific hits for COPA and OS were genes such as GFAP (see publication I, Figure 4), which were present in many samples but not with a strong outlier expression profile.

A very strong outlier protein expression profile was observed for thymidylate synthase (TYMS) in the HPA database. In the literature, there was no publication linking TYMS to glioma. In fact, only 7 out of the 29 (24%) genes had been previously linked to glioma in two or more publications. TYMS, which was uniquely identified by the GTI (rank 35), COPA (rank 105), OS (rank 1864) and ORT (rank 1646), showed an outlier expression pattern in the glioma cancer samples and with barely any expression in the normal samples. Remarkably, TYMS is a known target of many antifolate drugs such as 5’-fluorouracil (5-FU) and gemcitabine. These drugs had been given to glioma patients but offered little advantage over other treatment options[172-174]. Indeed, this is in line with the challenges in the general treatment of cancers, whereby all patients are treated the same way. Our study demonstrated that only a small subpopulation of glioma patients may benefit from the use of antifolate drugs.

A drug testing study was carried out using the above two antifolate drugs together with a new inhibitor, raltitrexed, and using four established glioblastoma cell lines (A172, U87-MG, LN-405, and U373-MG) and an immortalized foetal astrocyte cell line, SVGp12. The TYMS-specific inhibitor raltitrexed and gemcitabine showed favourable response values on the nM scale against the GBM cell lines expressing high levels of TYMS, while the normal foetal astrocyte SVGp12 was not affected. Furthermore, we used GTI outlier genes to define new glioma subtypes beyond the commonly known subtypes (Figure 5a). The hypothesis behind our approach was that
cancer subtypes defined by hierarchical clustering of expression data are usually illustrated using heatmap plots by performing un-supervised clustering of samples placed on the columns and as well as clustering of genes placed on rows. The correlation metric is one of the metrics used to generate the cluster distances. In our study, we displayed the correlations used to generate gene clusters that define the cancer subtypes as reported in several heatmap figures found in earlier publications[51].

**Figure 5: The correlation map (corMap) from the corDiff-map tool better defines cancer subtypes based on GTI outlier genes.** a) Three major subtypes of gliomas were identified based on the top GTI outlier genes. b) This plot illustrates a similar study performed using breast cancer samples and a GTI filtered list of the 534 breast cancer gene signature previously defined by Sørlie et al. [51,175]. Our approach identified six major breast cancer subtypes but not five as previously described.

We also examined whether our approach identified biologically meaningful subtypes by performing a similar analysis using the 534 intrinsic breast cancer gene signature published by Sørlie et al. [175]. Breast cancer subtypes have been verified in several publications[176-179] and they include luminal A, luminal B, HER2-enriched, basal-like, and normal-like. The 534 gene signature that best defines the five major breast cancer subtypes was assessed for outlier expression profiles, and only 114 genes exceeding the GTI outlier cut-off were used in making the breast cancer subtype correlation map (corMap). This approach identified all five major breast cancer subtypes previously described based on the 534 intrinsic gene signature, but importantly, a sixth subtype of breast cancer was identified with a clearly different set of genes driving the subtype. The rediscovery of breast cancer subtypes in our proof of concept study demonstrated that our subtypes identified in glioblastoma were likely to be clinically relevant, and these findings suggest that the GTI may indeed uncover
important genes behind the progression of different cancer subtypes, e.g. ErbB2 for the HER2-enriched breast cancer subtype. Strikingly, these results show a novel means of identifying subgroups of patients derived from systematic analysis of large cohorts of molecularly profiled cancer samples. To assign a new sample to a subtype identified using the correlation map tool, we generated cluster expression centroids for the gene signature acquired from the correlation map. The expression profile of a new sample is correlated with the centroid score and the sample is finally assigned to the group yielding the highest correlation score.

5.2 Other applications of the GTI algorithm in cancer drug targets and biomarker search

The GTI method could be used in a number of ways for performing outlier analysis in large-scale personalized medicine projects. It is very common that large-scale high-throughput biomedical datasets are characterized by varying numbers of samples per gene or drug, and the GTI method is well suited to working on datasets with such settings. The computation of the GTI is very efficient, since it does not involve any iterative process.

5.2.1 The GTI identifies cancer biomarkers and drug targets

The early detection of cancer using specialized assays is providing clues to early curative treatment. However, most cancer biomarkers lack specificity and often lead to some level of incorrect diagnosis. This challenge is often attributed to the lack of tissue specificity of the biomarkers being evaluated. The GTI approach is computationally efficient and it can be also efficiently utilized for identifying novel cancer biomarkers based on mining large-scale integrated gene expression datasets. The original use of the GTI described in study I was to identify genes showing an outlier expression profile, but because the GTI was not sensitive to variation in the number of samples per gene, its use was extended to biomarker search. The GTI biomarker search approach is performed in a similar way to the ROKU method approach previously described by Kadota et al. [180]. It should be noted that ROKU was designed for biomarker search studies, but not for outlier expression analysis studies. Using the GeneSapiens database[95], the GTI method was used to mine for tissue-specific expression profiles across 175 healthy and pathological tissues. This analysis was performed on 349 prostate cancer samples, 147 healthy prostate tissue
samples and a set of normal samples (n = 1476) as elaborated in publication II. Altogether, 295 outlier and prostate cancer-specific genes were identified from this large-scale in silico analysis. Among the genes were prostate cancer specific genes such as *AMACR* (alpha-methylacyl-CoA racemase), outlier genes such as *TPX2* and prostate-specific genes such as PSA (*KLK3*). Further analysis of the 295 genes using functional studies in two cultured prostate cancer cell lines (VCaP and LNCaP) revealed 112 hit genes from the proliferation assay. The high-throughput screening results were processed and normalized as described in publication II before hits were selected.

These results showed very minimal overlap between the hit genes of the two prostate cancer cell lines (17 out of 112). There was no obvious reason why the overlap between the two screens was minimal compared to what has been observed using drug testing assays[59]. In addition, *AMACR* was one of the genes on the list of the 295 GTI hits that did not appear on the list of 112 proliferation hit genes. This was a surprise due to its strong association with prostate cancer in the literature. It is widely recognized as a promising drug target[181-183], since it shows high protein levels and enzyme activity in prostate cancer[184], a subset of colon cancers[185] and some other cancer subtypes. Furthermore, several articles report that knock-down of *AMACR* decreases the proliferation of cultured cancer cells and rescues androgen-dependent growth in some prostate cancer cell lines[186,187]. To further demonstrate that *AMACR* is high in prostate cancer, a publicly available large-scale gene expression database, Medisapiens[95], was used to generate a body-wide expression profile boxplot for *AMACR* (Figure 6). These results demonstrate a powerful utility of the GTI method for identifying cancer-specific expression profiles for use in biomarker discovery studies and for drug repurposing approaches by efficiently prioritizing hits from whole genome expression studies to small biologically relevant gene signatures. However, the low reproducibility of findings from the two functional studies using siRNA technology and missed potential hits such as *AMACR* warrant further investigation. Some of the variability may be related to the study design and statistical methods, which could be further controlled.
Figure 6: Tissue-specific expression profile for AMACR. Extracted from MediSapiens. This plot is split into two parts with the boxplots for normal tissue types coloured green, while those for cancer subtypes are colored red. There is high expression of AMACR among prostate adenocarcinoma samples (185). This is an example of a cancer-specific gene showing high expression in a small number of cancer subtypes.
5.2.2 The GTI and differential correlation map (corDiff-map) identify cancer-specific co-expression networks

Systematic bioinformatic data analysis is crucial for the identification of potential driver oncogenes. In this context, the benefit in identifying statistically significant differential co-expression for a network of genes correlating well in the disease state and showing no correlation in the normal state is very important for identifying cancer-specific co-expression networks. In publication II, a systematic in silico analysis of large-scale gene expression data from the GeneSapiens database[95] was performed, focusing on prostate cancer as described in the previous section. The siRNA validation screen based on the identified 295 GTI hits revealed a 112 proliferation hit gene signature in prostate cancer. The whole proliferation hit gene signature of 112 genes was followed up further with co-expression analysis. Using the differential correlation map (corDiff-map) tool, we identified three prostate cancer subtypes from the correlation map generated using prostate cancer samples. The largest group was rich in genes involved in the endoplasmic reticulum (ER) and Golgi apparatus, prostate gland development, and oxidation-reduction, while the two smaller groups were enriched for genes involved in muscle contraction, actin cytoskeleton and mitosis based on gene ontology analysis. These tools allowed us to determine whether a cancer sample belongs to a given cancer subtype in a robust way, compared to using hierarchical clustering tree diagrams alone. There are multiple previously published prostate cancer drug targets, such as CLDN3, SIM2, CYP4F8, UBE2C, EPHX2, FAAH, ODC1, FOXA1 and MTDH[188-196], belonging to different groups identified using the corDiff-map tool. Further studies focused on four novel candidate drug targets not previously associated with prostate cancer, namely AIM1, ERGIC1, TMED3 and TPX2. Most importantly, these are among the 17 genes commonly identified from the two screens. The first three, AIM1, ERGIC1 and TMED3, were all co-expressed and functionally annotated to ER and the Golgi apparatus, while TPX2 was functionally annotated to the group of mitosis co-expressed genes. As a proof of concept, the corDiff-map tool was tested on a breast cancer gene signature for genes correlating with ERBB2, a target for trastuzumab. Indeed, the tool determined that the ERBB2 network of genes is co-expressed in cancer (Figure 7). Other differential co-expression analysis tools such as WGCNA[166,167], DCp[168] and DCeN[169] have been mentioned in the literature,
but only the individual correlations approach employed by DCeN comes close to the corDiff-map approach.
**Results**

Figure 7: *ERBB2* correlation network of genes displayed using the corDiff-map tool. Correlation heatmaps were created using a) correlations in breast cancer samples, b) correlations in normal samples and c) the significance of the difference of correlations of two maps expressed as *P*-values (-log10(*P*-value)). The black box shows a network of genes only correlating in the cancer state (a) but not in the normal state (b).
Results

This tool identifies a network of cancer driver genes in a more robust way based on differential co-expression between cancer and normal states assessed by significance scores (P-values >0.05 in red (Figure 7c)) compared to the traditionally used tree diagrams. The P-value correlation map (Figure 7c) effectively identifies networks of genes co-expressed in the cancer state and not in the normal state, and could enable the delivery of therapies targeting cancer-specific networks of genes and sparing normal cells in which the genes are not co-activated. Thus, this newly developed bioinformatic tool could be potentially used to facilitate our understanding of many personalized medicine questions.

5.3 Impact of normalization methods on high-throughput screening data

In publication II, we explored the biological relevance of outlier genes identified using the GTI method in a prostate cancer study using siRNAs. This study revealed minimal overlap of the hit genes identified from the VCaP and LNCaP (17 genes) screens. The findings from publication II prompted us to further examine the impact of normalization methods on the quality of high-throughput screening (HTS) data.

In publication III, we compared the B-score together with the Loess-fit approach as described in the materials and methods (see section 4.4). Here, we aimed to determine whether the plate layout or the normalization methods could significantly affect the quality of HTS data. A layout based on placing control samples on the edge is commonly used in HTS experiments. However, HTS assays exhibit systematic edge effects, which could affect the quality of the data from control wells. The data from control wells are essential for QC assessment and for the calculation of percent inhibition scores needed for curve fitting in drug testing studies.

First, we performed a high-throughput drug testing experiment with the controls scattered across the plate. Then, the processed raw data were normalized using the B-score and Loess-fit approach in order to remove within-plate effects. According to the results, the plates containing drugs applied at a high concentration showed poor post-normalization QC scores, especially for the B-score method. The VCaP replicate screens with plates that showed poor QC before normalization could not be corrected
using either of the two normalization methods. When we visualized the B-score normalized data using scatter plots (Figure 8), a gradual shift in signal intensities following the hit rates on the plates was observed.

![Figure 8: B-score-normalized data displayed according to concentration levels of the drugs. The B-score normalization effect could be seen in the data from the plate with the highest concentration of drugs, D5.](image)

The scatter plots are arranged according to increasing dose of each drug (D1 to D5), and we would therefore expect to see more hits on the fifth plate (D5), but the B-score compressed the hits. The Loess-fit approach in general generated data of high quality and close to or even higher than the recommended QC thresholds, as shown in publication III.

A simulation experiment was then designed mimicking the commonly used plate layouts under increasing hit rate scenarios to enable the identification of the most optimal hit rate and plate layout combination for normalizations to perform well above the recommended QC thresholds. The simulation data involved picking non-hits from the distribution of negative controls (DMSO) and hits from the distribution of positive controls, as described in detailed in publication III. We started with a hit rate of 20 wells out of 384 wells (5%) and increased this number iteratively until a hit rate of 160 wells (42%) was obtained. A dataset of 142 plates was generated and QC scores were analysed across the different simulation datasets. The critical hit rate cutoff was estimated as the point at which the quality control scores were continuously reported below the recommended thresholds ($Z'$-factor < 0.5 and SSMD < 6) with increasing hit rates. The maximum tolerable hit rate for normalizations to perform well was determined to be 20% or 77/384 wells (Figure 9).
Results

Figure 9: Simulation data results showing increasing hit rate versus post-normalization quality control scores for a) Z’-factor and b) SSMD images. In summary, a scattered layout of the controls and a hit rate of less than 20% were identified to be the most optimal. The dashed line shows the recommended thresholds for each method.

These results indicate that normalization methods, as well as the plate layout, indeed impact on the quality of high-throughput screening experiments with high hit rates above 20%.

5.4 Analysis of consistency of large-scale drug testing and pharmacogenomics studies

Recently, Haibe-Kains et al. [43] published an article stressing the importance of standardization in large-scale drug testing studies. They suggested the need for the implementation of better standards for drug testing protocols and potentially also the standardization of drug scoring methods and analysis approaches. For example, here the consistency of gene expression data for 471 overlapping cell lines from CGP and CCLE was assessed together with three drug testing datasets from CGP, CCLE and FIMM using the same bioinformatic analysis approach. Each dataset was organized as a matrix of the high throughput measurements (genes and drugs) placed on the rows
while all the tested samples were placed on the columns. This step was consistent
with the way in which gene expression is organized using Bioconductor
ExpressionSets[152]. Calculations from HTS measurements that were placed on the
rows “measured features, e.g. genes and drugs” cannot be directly used to infer
interpretations from calculations made between the samples “tested items, e.g. cell
lines”, which were placed on the columns of our matrices due to varying underlying
distributions and heterogeneity. Measurements could share the same experimental
protocols and inherent properties, while samples could share a similar biology.

Here, we showed how large-scale bioinformatic data analysis tasks need to be
supported by a vast number of visualizations to enable improved understanding of the
underlying data. For example, when we examined the CGP dataset (release 2, July
2012) containing 138 oncology drugs screened across a range of 329–668 cell lines,
we discovered that the majority of the drugs show effective sensitivity (drug
sensitivity score (DSS) ≥ 15) in less than 25% of all the 647 tested cell lines (Figure
10). Mathematical calculations produced using drugs that show almost no response in
the majority of the cell lines are likely to be biased due to the lack of a measured
signal above the technical noise. Most high-throughput screening technologies do not
always provide robust measurements for low signal readouts below technical
noise[197,198].
Results

When we compared the consistency of drug testing data from CGP, CCLE and FIMM organized in matrices as illustrated in Figure 11, we discovered that correlations performed across cell lines, as was done for gene expression in the article by Haibe-Kains et al. [43], lead to very high consistency (Figure 12a). The gene expression data had been reported to show a high level of consistency (median rank correlation, MRC = 0.85) when compared between hundreds of cancer cell lines. The key difference between how we were able to achieve high consistency for drug testing data far beyond what had been reported was the use of a consistent data analysis scheme (Figure 11) and the standardization of data processing procedures (see methods section 4.5 for details). This study demonstrates the complexity of mining

Figure 10: Circular stacked barplot showing the proportion of cell lines in which a drug is categorized as being effective (DSS ≥ 15), intermediate (DSS between 5 and 15) and not effective (DSS < 5)

5.4.1 Consistent bioinformatic meta-analysis gives consistent results

When we compared the consistency of drug testing data from CGP, CCLE and FIMM organized in matrices as illustrated in Figure 11, we discovered that correlations performed across cell lines, as was done for gene expression in the article by Haibe-Kains et al. [43], lead to very high consistency (Figure 12a). The gene expression data had been reported to show a high level of consistency (median rank correlation, MRC = 0.85) when compared between hundreds of cancer cell lines. The key difference between how we were able to achieve high consistency for drug testing data far beyond what had been reported was the use of a consistent data analysis scheme (Figure 11) and the standardization of data processing procedures (see methods section 4.5 for details). This study demonstrates the complexity of mining
Results

large-scale biomedical datasets with many dimensions, several distributions and varying sample numbers. Here, performing correlations based on a row-wise or column-wise scheme for genes and at the same time for analysis based on drug responses enabled us to achieve differing levels of consistency (Table 4).

Table 4: Median Spearman’s rank correlation coefficients for drug testing data across paired sites.

<table>
<thead>
<tr>
<th></th>
<th>CGP and CCLE</th>
<th>CGP and FIMM</th>
<th>CCLE and FIMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between cell lines</td>
<td>0.63</td>
<td>0.54</td>
<td>0.74</td>
</tr>
<tr>
<td>(column-wise)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Across cell lines</td>
<td>0.37</td>
<td>0.29</td>
<td>0.49</td>
</tr>
<tr>
<td>(row-wise)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The row-wise correlations across all site pairs were very poor based on data generated using standardized drug response scoring (see Table 4, second row). The conclusions of Haibe-Kains regarding the lack of consistency among drug testing experiments were based on comparing the row-wise drug testing correlation score for CGP and CCLE (MRC = 0.37, see Table 4) against the CGP versus CCLE cell line gene expression correlation score (MRC = 0.85). Moreover, the drug testing correlation score between cell lines for comparisons between CGP and CCLE (MRC = 0.63, see Table 4) should have been used, as we have shown here. Furthermore, comparisons with a new dataset profiled using the FIMM drug testing assay[59], showed a high positive correlation between CCLE and FIMM (MRC = 0.74) compared to that between CGP and FIMM (MRC = 0.54). This suggests that other factors were contributing to the strong consistency observed between CCLE and FIMM. Comparing the experimental procedures and the drug testing assays from the three sites to some extent explained the observed findings. First, similar experimental protocols were applied at CCLE and FIMM, including the use of the same controls (the vehicle as a negative control and positive controls consisting of toxic compounds 100 μM benzethonium chloride or 1 μM MG132) and the same readout (CellTiter-Glo, Promega). There were also differences, such as the plate format (1536 vs 384 wells) and no effort to standardize the cell numbers used, the passage number and cell culture media, or the drug source, supplier and handling. Thus, there was still a high likelihood for substantial improvement of this observed level of agreement between CCLE and FIMM if extra effort was made towards standardizing experimental protocols. Second, the CGP experimental protocol differed from the other two in
Results

terms of the readout (fluorescent nucleic acid stain Syto 60, Life Technologies), choice of controls (drug-free cells as negative and no cells as positive controls), and plate format (96-or 384-well plates). Unlike drug testing datasets, measurements at the gene expression level based on microarray technology for one sample are performed across thousands of genes. In drug testing, by comparison, the same sample could be screened with tens to hundreds of drugs, such as the CCLE study, which covered only 24 drugs. This makes statistical calculations using all readouts from microarray gene expression data very robust when performed at the sample level (column-wise), but not necessarily robust at the level of each gene (row-wise) due to varying degrees of freedom. Statistical calculations at the gene level could be affected by small sample sizes. Surprisingly, there was no study in the literature describing the consistency and reproducibility of gene expression data when compared between gene measurements (row-wise) across cell lines.

![Diagram](image)

**Figure 11:** Schematic diagram illustrating the approach we used to achieve consistency.

The mean of the expression data for each gene was used to categorize genes into five groups based on the mean expression level. The results of this study demonstrated
that comparisons using gene expression data across cell lines (row-wise) similarly produced discordant results (MRC = 0.58 between CGP and CCLE) (Figure 12b). Interestingly, the majority of poor correlations at the expression were observed for genes with Affymetrix expression signal intensities approximately below 100. The drug testing correlation comparisons between cell lines in Figure 12a were on average higher than the median correlation comparisons at the gene expression level in Figure 12b. This illustrates how inconsistent bioinformatic analysis approaches could distort important findings in favour of an incorrect hypothesis.

Figure 12: The above boxplots show a) correlations between cell lines using drug testing data and b) correlations across 485 cell lines using gene expression data. The results indicate that correlations at the sample level (between cell line) are not comparable to correlations at the measurement level (gene or drug measurements across cell lines). The mean drug testing correlation at the sample level was slightly higher than the mean correlation at the gene level across cell lines.

5.4.2 Correlations made when there is little response lead to a low concordance score

Since the majority of the large-scale drug testing experiments included targeted and cytotoxic drugs, it is expected that some targeted drugs are likely to show no response across all cell lines where there is no matching target. However, due to the challenges of quantifying signal readouts from high-throughput drug testing experiments, a certain amount of background noise is measured as a real signal. Basically, for these cases, drugs tend to show low DSS scores that could be interpreted as a low response. However, when calculating hundreds of correlations, these drugs with extremely low DSS scores will also be quantified with a correlation score, leading to similar inconsistency observed among genes with low expression levels (Figure 12b). A meta-analysis of sensitivity profiles of CGP data revealed that several targeted drugs
showed a good response (DSS > 5) in less than 5 cell lines and no connecting node was drawn for such drugs (Figure 13). If two drugs were sensitive in more than five overlapping cell lines, a connecting line was drawn between the two nodes displayed using TVNViewer tool[199].

Figure 13: Illustration of pairs of different drugs with enough data to calculate correlations (n > 5 cell lines). Cytotoxic drugs show considerable overlap among several cell lines, and could thus better be used for calculating correlations than many targeted drugs.

These results demonstrate that several targeted drugs show a response in only a few cell lines as outliers. Their drug response profiles may not be suitable for making consistent correlations due to the strong influence of data points with low response scores close to measuring technical noise.

In summary, comparisons at the sample level (between cell lines) demonstrated that drug testing data are highly consistent between different laboratories, especially when
applied bioinformatic analysis approaches are kept consistent and when experimental conditions and assays are standardized between laboratories.
6. Discussion

The increasing application of high-throughput technologies is fuelling the ongoing explosion in biomedical datasets. This has led to serious capacity issues related to the ability of the scientific community to handle such data [26,29]. The low availability of specialized bioinformatic tools limits our ability to store, mine, analyse and to make interpretations from the data. Integrated analysis of large-scale molecular profiling and drug testing datasets requires in-depth knowledge of several biological concepts, analysis algorithms, and programming expertise. It is also very clear that there is no uniformly accepted standard way to perform bioinformatic data mining and integration of large-scale biomedical datasets. The unique nature of most data analysis projects using large-scale biomedical data requires customized assembly of specialized data analysis modules [25,29,48]. While these significant challenges in translating large-scale molecular profiling and drug testing data do exist, it is evident that they will severely impact on our ability to understand cancer, develop standardized drug discovery approaches and to tailor treatment options to individual patients[48,87]. The development of applied and pure bioinformatic methods for the analysis, mining and modelling of large-scale molecular profiling and drug testing datasets should, therefore, accelerate not only cancer research but also drug discovery and personalised medicine.

In this thesis research, I utilized both pure and applied bioinformatic approaches to mine, analyse and to draw new interpretations from large-scale molecular profiling and drug testing datasets. The bioinformatic methods developed in this thesis work enabled us to identify a) genes with outlier expression profiles, b) genes with cancer-specific expression profiles (cancer biomarkers), c) cancer subtypes using the correlation map tool and d) cancer-specific co-expression gene networks using corDiff-map tool, as well as to e) improve the normalization of HTS data and f) generate consistent results across pharmacogenomic studies. The findings presented in this thesis demonstrate the use of bioinformatic and statistical tools for the analysis, mining and integrating large-scale molecular profiling of drug testing datasets.
In publication I, we adapted a statistical method originally designed for solving outlier analysis problems in economics to cancer research. The hypothesis that a normal gene abnormally deregulated in a subgroup of cancer samples leads to oncogenesis forms the basis for outlier analysis studies in cancer research [34]. Some microarray gene expression studies have successfully applied a variety of methods such as COPA, OS, ORT and the t-statistic for the identification of genes overexpressed in a subgroup of cancer patients [76,104-106]. The design of the existing methods addresses outlier analysis tasks based on datasets of single study settings with a relatively equal number of samples per gene, whereas in large-scale integrated datasets [22,95], their performance remains unknown. We developed a new outlier expression profiling metric, the GTI, with a normalization factor in the formula catering for varying sample numbers found in large-scale microarray datasets, enabling the systematic identification of outlier expression profiles in comparison to four other existing outlier analysis methods.

The varying sample numbers did not affect the results based on the t-statistic; however, the t-statistic mainly searches for uniformly overexpressed genes. Tomlins et al. demonstrated in their recent study that COPA was more powerful than the t-statistic for detecting cancer outliers [34]. The results from the other outlier analysis methods included in this study other than the GTI indicated that their performance was affected by varying sample numbers per gene, since the formulas do not include sample number normalization. Furthermore, the use of data from both normal and cancer samples to define the outlier cut-off point significantly improved our GTI results compared to the approach of using cancer samples alone utilized in both COPA and OS metrics or using normal samples alone in the ORT statistic. Wu et al. researched this topic and showed that the COPA and OS methods employ only cancer samples to define the outlier cut-off point, while the ORT method uses only normal samples [106]. Determining the outlier cut-off based on samples from one group could be less efficient, because data from one group do not adequately represent the true biological variation of the gene expression profiles between normal and cancer samples. The task of defining the most optimal cut-off in outlier gene expression studies is difficult, since each gene is expressed differently in a multitude of normal and cancer tissues.
The unique findings for both COPA and OS were genes showing high overall expression signal intensities in the majority of cancer samples, as well as in a non-malignant normal reference group such as GFAP\cite{200,201}. A systematic exploration of the GTI unique genes revealed highly interesting results associated with GBM cancer biology. For example, \textit{TYMS} showed outlier protein expression in its staining pattern in glioblastoma immunohistochemical protein staining images from the HPA database\cite{171}, and overall, 62% of the other 29 genes showed an outlier expression pattern at the protein level. Although the levels of protein expression have a direct link to biological processes, mRNA gene expression measurements are very often used as a proxy to infer functional differences occurring at the protein level. This is because studying protein expression is more challenging due to the more costly and laborious technologies, e.g. tissue microarray technologies (TMAs). Researchers have reported a moderate or weak positive correlation between mRNA and protein expression with correlation values ranging from 0.2 to 0.6\cite{202-205}, whereas comparing mRNA gene expression findings to protein level expression profiles allows us to establish whether the observed outlier expression profiles have any functional biological consequences.

Interestingly, some GTI unique hits such as \textit{TYMS}, a drug target for several antifolate drugs such as 5'-fluorouracil (5-FU) and gemcitabine, represent biologically attractive opportunities for drug repurposing. Previous studies on utilizing these antifolate drugs for the treatment of glioblastoma patients have demonstrated little success, whereas the results from this study revealed a favourable response on the nM scale of TYMS inhibitors tested against four GBM cell lines expressing high levels of \textit{TYMS}. We postulate that the selection process of study participants could have led to the low success observed in previous studies, because \textit{TYMS} shows an outlier expression in a small subgroup of GBM patients. Furthermore, 5-FU and gemcitabine are known to inhibit RRM2 (ribonucleoside-diphosphate reductase M2 subunit) and DHFR (dihydrofolate reductase)\cite{73}, while the new drug raltitrexed may be more specific to TYMS. However, the use of raltitrexed for treating glioblastoma remains to be tested. The observations from this study indicate that outlier analysis studies will enable the design and adaptation of new cancer subtype classifications to translate the potential therapeutic gain in clinical trials. We performed validation studies \textit{in vitro} based on these findings in a proof-of-concept experiment.
In addition, we further validated the application of the GTI method for the identification of biomarkers and drug targets through functional studies based on high-throughput screening of siRNAs in two prostate cancer cell lines. The use of siRNA screens to probe for genetic bias for cancer-related processes, including cell proliferation, migration and apoptosis, indicates that it is a powerful tool for basic research [15-17]. More recently, Ramalingam et al. described the CRISPR (clustered regularly interspaced short palindromic repeats) and CRISPR-associated (Cas)9-mediated gene editing technology as a new cost-effective method that combines targeted genome editing with a simple, less time-consuming assay set-up and with potential utility in high-throughput screening approaches[206,207]. Carrying out genome-wide siRNA studies is very expensive, and bioinformatics data processing could be laborious, making hit selection challenging. The GTI tool focusing on identifying cancer-specific and outlier genes enabled an efficient way of selecting genes for functional studies in prostate cancer. Using hit prioritization statistics and the B-score from the HT-RNAi normalization method[77], we generated a signature of 112 proliferation hit genes and performed further bioinformatics analysis using the corDiff-map tool.

Overall, we observed low overlap between the hits of the VCaP screen and those for the LNCaP screen (17 genes in common). In particular, the results for the VCaP screens showed poor Pearson correlation (r) between replicates (as low as r = 0.36), which is not surprising given that researchers have previously reported that cell lines differ in their transfection efficiency, which could affect the reproducibility of the results[112,208]. Furthermore, a literature search revealed that multiple teams have discussed the low overlap of results from siRNA studies[17,208-210]. In part, some researchers postulate that data processing and statistical methods may contribute to the lack of agreement among siRNA studies due to the absence of standard methods for different types of experiments. For example, we used the B-score normalization method in study II while assuming a low hit rate from our experiments, because the B-score method was designed for the analysis of datasets with relatively low hit rates[211-213]. Surprisingly, our assumption of a low hit rate might not have been met, because the analysis missed some well-known genes such as AMACR, a gene previously characterized as a prostate cancer proliferation hit gene [186].
reasoned that the results from siRNAi screening could be improved with an extra assessment of the data analysis procedures as well as the screening protocols.

The corDiff-map enabled us to identify three major prostate cancer subtypes exhibiting different mechanisms for cell growth regulation based on our 112 proliferation hit gene signature from functional studies. Importantly, these subtypes showed distinct co-expression patterns mostly enriched in the ER and Golgi apparatus, prostate gland and oxidation-reduction as functional annotations. The other two small subtypes were enriched in muscle contraction associated genes and mitosis. This confirmed previous reports that prostate cancer is a heterogeneous disease, which could manifest in different forms[194]. A proof-of-concept experiment based on novel drug targets selected from the 112 hits not previously linked to prostate cancer, namely AIM1, TMED3, ERGIC1 and TPX2, revealed that, for example, ERGIC1 silencing regulated the proliferation of ERG oncogene-positive prostate cancers and inhibited ERG mRNA expression in VCAP cells, and TPX2 silencing reduced KLK3 expression, implying that TPX2 regulates androgen receptor-mediated signalling.

The use of the GTI and the corDiff-map tool demonstrated an efficient way of performing genome-wide hit prioritization to identify cancer subtypes and gene co-expression networks, and we demonstrated the utility of the approach by rediscovering the known breast cancer subtypes using a published breast cancer subtype gene signature from Sørlie et al. [51]. Interestingly, the corDiff-map tool identified all six major breast cancer subtypes that had previously been defined. A recent study by Curtis et al. [214] based on a larger number of breast cancer samples (2000) revealed ten breast cancer subtypes. Furthermore, the application of the corDiff-map tool to the ERBB2 co-expression network confirmed the cancer specific co-regulation of the ERBB2 genes, including GRB7, PSMD3 and STARD3. The identified cancer-specific network included the target for trastuzumab, thereby illustrating the clinical utility of the approach. Many cancer studies have identified genes involved in oncogenesis, such as EZH2[215-219], with no defined clinically actionable inhibitors, but drugging some other genes in the cluster (drug targets) in the cancer-specific gene network could facilitate the delivery of new therapies or repositioning of existing drugs to new diseases. The corDiff-Map tool could also be used in personalized drug testing studies involving samples from normal patients. The
tool could easily identify cancer-specific drug response networks that are not present in normal cells.

In publication III, we assessed the impact of normalization methods on the reproducibility and quality of HTS data based on simulated and real high-throughput drug testing data. Normalization is performed to minimize within-plate effects and across-plate effects so that data are comparable. Despite careful design of HTS experiments, some level of row, column and edge effects within a plate is expected to be observed[79]. The need for assessing the impact of HTS normalization methods became apparent when we observed low overlap between two HT-RNAi studies for VCaP and LNCaP in study II. Moreover, the use of high-throughput drug screening/testing approaches similar to those previously applied for HT-RNAi studies has recently become popular at many research centres[36,37,59]. Most of these studies have used applied bioinformatic approaches adapted from HT-RNAi protocols to process, mine and analyse these large-scale drug testing datasets.

We tested 306 drugs over a dose range of five dilutions for each drug testing experiment. The setup of our drug testing experiment implied that the fifth plate contained a high hit rate, whereas existing HTS normalization methods that use data on the whole plate to serve as the negative control ‘no response’ assume that most of the samples are inactive[77,81,212]. This assumption does not hold for drug testing data as well as HT-RNAi screens with a high hit rate due to a customized library selection criterion that tends to result in screens with high hit rates. Besides, as many reports have shown, several data processing workflows and analytical methods are used to generate drug response scoring metrics such as IC50 or AUC as the final drug response measurement in drug testing experiments[36,37,165]. Comparing the reproducibility of drug testing data based on results from two separate data processing pipelines could be complicated due to differences in the methods used. For example, the cellHTS[211,220] Bioconductor package for HTS data analysis allows normalization to be performed based on the median (default), mean, percent of control (POC), normalized percent inhibition (NPI), negatives, the B-score and robust local fit regression (locfit), all showing different outputs. These methods enable a variety of ways to analyse HTS data, and our literature search revealed that there are few, if any, agreed standards or best practices for data processing and analysis. For
example, we found that a typical curve-fitting process could be easily extrapolated beyond what is biologically meaningful, and AUC measurements could be produced based on boundaries that are not comparable. Furthermore, the data for the same drug screened at different unmatched concentrations cannot be compared, and the aggregated scores do not show the concentration of the drug. It is evident that achieving quality and reproducibility of drug testing data based on analysing data acquired from non-standardized laboratory assays and bioinformatic analysis pipelines across many studies remains a challenge [45]. However, our findings demonstrate that we can achieve good reproducibility and quality if the laboratory assays are similar and robust quality control assessment is performed.

Assessing the quality of the data before and after normalization facilitates systematic assessment of the impact of different normalization methods, thus enabling an efficient way to assess the effect of different methods on the quality of HTS data. Study III clearly demonstrated that normalization impacts the quality of HTS data generated with perturbations, leading to high hit rates. Furthermore, a plate layout based on placing controls on the edge of the plate was found to be less effective compared to a layout based on scattering controls across the entire plate. Murie et al. introduced an approach based on the control plate regression method, which depends on adjusting signal intensities for the treatment plates by scaling the data based on bias estimates for the control plate[80]. The challenge with this approach is that it requires an extra plate, whereas it is impossible to calculate QC metrics without placing controls on each plate. Adapting methods designed for normalizing microarray data such as Generalized Procrustes Analysis[221] and modified Loess (LoessM)[222] would be an alternative solution, but the design of the experiments is essential if one is to adapt any of these methods. For example, Generalized Procrustes Analysis requires replicate experiments, while drug testing often involves expensive drugs and the number of primary cells in personalized medicine projects is limited.

Our analysis based on comparing two HTS normalization methods, the B-score[77] and a local polynomial fit[81] method called Loess, revealed that low-quality screens show low reproducibility. We observed that the B-score normalization algorithm negatively impacted on the quality of the data for the plate with the highest concentration of the drugs, leading to QC scores below the recommended thresholds
Discussion

among all screens in the VCaP and LAPC4 experiments. Poor QC results indicate that
the controls do not represent two independent distributions for positive and negative
controls. Moreover, the calculation of the percent inhibition scores strongly depends
on the quality of the data from the control wells. These findings confirmed the
original hypothesis that the B-score was designed for normalizing data with low hit
rates, and the results showed that methods with similar assumptions might impact on
the quality of drug testing data. The simulation study revealed that normalizations
perform poorly beyond a hit rate of 20%, especially for the B-score method compared
to the Loess fit approach combined with a scattered layout. It was also observed from
the real drug testing study that no normalization could correct for strong within-plate
effects, and new methods have to be designed to effectively correct for such effects.

Haibe-Kains et al. recently reported poor agreement of correlation analysis results
based on drug testing data compared between two large-scale pharmacogenomic
datasets from CGP and CCLE, and thus raised fundamental questions in the field of
personalized medicine [43]. Many aspects of data normalization and analytical
methods used for drug testing were adapted from HT-RNAi studies, and we assumed
that the lack of agreement between the two large-scale drug testing datasets was due
to bioinformatic data processing and statistical analysis methods. For example,
statistical inferences are often made during the design of statistical models to ease
computations or to adapt existing methods (applied bioinformatics). However, when
conducting inference on matrix data, there is a general assumption that the variables
along one dimension, such as columns, are independent and that they enable us to
pool these observations to make inferences on the measurements across the rows, as
described by Genevera et al. [47]. We examined the research of Genevera et al. on the
topic of modelling the effects of row and column correlations based on transposable
data in a matrix format, such as gene expression microarray data. This research
guided our statistical approach to comparing two large-scale pharmacogenomic
datasets. We observed that several groups of statisticians (Dudoit et al., Efron, Owen
et al., Qui et al. and Schwartzman et al.)[223-227] have previously studied
correlations among the genes placed on rows and determined their effect on standard
statistical methods used for large-scale inference. Given this background, it was
important to take into account this large body of literature before performing
statistical inferences on columns or rows.
In study IV, we merged datasets from both gene expression and drug testing experiments according to overlapping cell lines and the HTS measurements (genes and drugs). The HTS measurements were placed on the rows while the samples (cell lines) were all on the columns, as shown in Figure 11. In all comparisons, we calculated correlations either row-wise or column-wise to achieve comparable correlations across cell lines (row-wise) or between cell lines (column-wise). The comparisons of CGP and CCLE demonstrated that correlations should either be compared column-wise between samples or row-wise across samples. Overall, the correlations performed column-wise (between cell lines) were higher ($r > 0.5$) compared to correlations performed row-wise (across cell lines) ($r < 0.5$). Further validation of the approach using an extended third dataset from FIMM revealed similar results that correlations between cell lines (column-wise) were much higher than correlations across cell lines (row-wise). Interestingly, we observed marked similarities between FIMM[59] and CCLE[36] laboratory screening protocols, and the correlation analysis yielded a very high correlation (MRC = 0.74) for the comparisons made between FIMM and CCLE compared to correlations between FIMM and CGP (MRC = 0.54). These findings highlighted the need for standardization of laboratory protocols for drug testing assays among different laboratories. A joint effort by CGP and CCLE teams to identify the reasons for the inconsistencies has demonstrated that there is an agreement between the two large-scale pharmacogenomic datasets [46]. The new study by CCLE and CGP [46] did not examine the approach of exploring correlations row-wise or column-wise, as presented in our study IV.

Hence, the results from the meta-analysis in study IV implied that measurements from high-throughput data such as gene, drug, and siRNA measurements present highly discontinuous distributions across large panels of cell lines compared to distributions explored between cell lines across a large panel of measurements. In summary, study IV revealed that correlations between samples are more consistent and higher than correlations among measurements (genes or drugs) across samples.
7. Conclusion and future prospects

The GTI method adapted from economics and modified in this thesis was used to identify key cancer outlier genes, and these signatures were found to be highly associated with biological processes of tumours. Drawing on the successful examples, we demonstrated the ability of the GTI to rank genes according to overall outlier expression and to detect genes whose expression pattern was strongly linked to cancer proliferation and biological processes. The analysis based on GTI-ranked genes revealed many genes with outlier protein expression patterns, and the approach is directly applicable to identifying drug targets, cancer biomarkers and subtype classification of cancers.

The systematic assessment of normalization methods for HTS datasets with high hit rates and different plate layouts revealed that normalizations impact on the quality and reproducibility of HTS data. In this study, we discovered that no normalization could correct for strong within-plate effects and that datasets with poor quality control metrics showed low reproducibility. The results based on using Loess combined with a scattered layout showed more consistent pre- and post-normalization quality control results given low hit rate scenarios below 20%, and suggested Loess as an improved way of normalizing some types HTS datasets.

Furthermore, drug testing data are highly consistent between large-scale independent studies when bioinformatic data assembly, processing, and analysis are performed systematically while checking assumptions regarding statistical inference on large-scale transposable data matrices. Correlations between cell lines should not be compared directly with those across measurements, such as genes and drugs across cell lines.

Despite these interesting findings, there are still several challenges in analysing large-scale drug testing data. Focusing on eliminating these challenges and limitations, the scientific community will need to develop standards for drug response scoring, drug nomenclature, assay protocols and experiment designs. Novel statistical methods will still be needed to improve the integration of multi-layered large-scale
Conclusion and future prospects

pharmacogenomic datasets. An improved ability to store, analyse, normalize and mine datasets will be needed for personalized medicine to become a reality.


8. Acknowledgements

The studies presented in this thesis were carried out at VTT Medical Biotechnology Turku, Finland and at the Institute for Molecular Medicine Finland, FIMM, between 2009 and 2016. I would like to thank the administration at VTT, especially Heidi Sid, who helped with all the practical matters in joining the Finnish working environment and giving me valuable tips on networking with colleagues at VTT. I would like to thank Docents Harri Siitari and Kirsi-Marja for providing an excellent working atmosphere at VTT.

The last part of the research was performed at FIMM. My work at FIMM started way back in 2009 as a commuting student, and my sincere appreciation goes to Susanna Rosas for helping me with all the practical matters at FIMM, especially the SAP system, and arranging thesis committee meetings. Susanna, thank you very much for the lovely seasonal cards to my family. Since moving to FIMM in 2011, I have enjoyed interacting with the human resource and administration team at FIMM led by Reetta Niemelä and the research training coordinator Gretchen Repasky. At FIMM, I would also like to thank all my collaborators who have helped me to work on a diverse number of projects. I am also grateful for the financial support I received from the Cancer Society of Finland, the European Union’s Seventh Framework Programme (FP7/2007–2013) under grant agreement n°258068; EU-FP7-Systems Microscopy NoE, the Sigrid Juselius Foundation, the Cancer Society of Finland, the Academy of Finland Centre of Excellence and the Magnus Ehrnrooth Foundation. I also thank the Doctoral Programme in Biomedicine (DPBM) for organizing scientific seminars and funding part of my Ph.D. studies. I am very grateful for the financial support for my Master’s degree that I received from the United States National Institutes of Health (NIH) Fogarty International Center (FIC) provided through the Makerere University and University of California, San Francisco (MU-UCSF) research collaboration.

The official reviewers, Adjunct Professor Laura Elo-Uhlgren and Dr Silje Nord, are thanked for the timely, helpful and constructive comments on my thesis. I thank Dr Roy Siddall for reviewing the language of my thesis and for the useful English
language courses. I thank Dr. Bissan Al-Lazikani for accepting to be the Opponent in my thesis defense.

This Ph.D. work would not have been possible without the support, care and inspiration from my supervisor, Professor Olli Kallioniemi. He is so kind, welcoming, extremely accommodating, very professional, very motivating, very trustful, respectful and with impeccable integrity. My interactions with him have always been very motivating, and I do look at him as my mentor, and a role model for outstanding leadership and scientific excellence. He always focuses on my strengths and helps me overcome my weaknesses. For example, he has trusted me with so many big projects without fear that I might fail or embarrass him. He takes care of me financially, academically and has helped me grow as a scientist, future leader, and manager.

I thank the past and present members of the Kallioniemi group for being very friendly colleagues, and give special thanks to members who are not co-authors on my papers in this thesis: Ville Härmä, Juha Rantala, Magnus Marcin, Arho Virkki, Ilmari Ahonen, Johannes Virtanen, Santosh Gupta, Kirsi Ketola, Jeroen Pouwels, Pekka Tiikainen, Rolf Skotheim, Sirkku Pollari, Teijo Pellinen, Elina Mattila, Laasola Petra, Petri Saviranta, Rami Mäkelä, Vuoriluoto Karoliina, Maija Wolf, Malani Dishia, Poojitha Kota, Akira Hirasawa, He Liye, P Niettäinen Vilja M, Mikkonen Pia M, Mariliina arjama, Sami Blom, Rahkama Vesa, Fiere Clement, Lassi Paavolainen, Katja valinaki, Antonio ribeiro, Tanja Ruokoranta, Malyutina Alina, Kovanen Ruusu-Maaria, Maija Puhka and Taija af Hällström. I have liked the good meeting discussions in our group and the comments you gave when I presented my work. It showed you care. Most especially, I thank Ville Härmä, Magnus Marcin, Santosh Gupta, Akira Hirasawa, Clement Fiere, Bychkov Dmitrii, Maria Anastasina and Astrid Murumägi for the nice lunch discussions, and Teijo Pellinen for teaching me how to ski.

Special thanks to all my co-authors: Sara Henri, Haapa-Paananen Saija, Kilpinen Sami, Pisto Tommi, Bucher Elmar, Ojala Kale, Iljin Kristiina, Vainio Paula, Björkman Mari, Gupta Santosh, Kohonen Pekka, Nees Matthias, Fey Vidal, Mirtti Tuomas, Alanen KA, Perälä Merja, Swapnil Potdar, Dmitrii Bychkov, Jani Saarel, Saeed Khalid, Wennerberg Krister, Aittokallio Tero, Östling Päivi, Yadav Bhagwan,
Acknowledgements

Murumägi Astrid, Hirasawa Akira, Kangaspeska Sara and Gautam Prson. I thank Dr Matthias Nees for inviting me to my first interview at VTT and for all the help with understanding cancer biology and the guidance to enable me to design biologically interesting bioinformatic analysis questions. Dr Matthias Nees, Dr Kristiina Iljin and Dr Päivi Östling, as senior researchers in Prof. Olli Kallioniemi’s group and who co-supervised most of my studies, have helped me so much with understanding cancer research, editing my presentations, papers, guiding and discussing with me on a daily basis, as well as helping me with running multi-disciplinary collaborative projects. To all my workmates at FIMM for creating a real multidisciplinary working environment. This Ph.D. thesis work benefited from the large Personalised Cancer Medicine project involving several research groups at FIMM, the Hematology Research Unit Helsinki and the Helsinki Urological Biobank, Finland. Many tools developed in this thesis research were tested on the data from this large collaborative project. I would like to thank the principal investigators (PIs) in the project, Tero Aittokalio, Caroline Heckman, Jonathan Knowles, Olli Kallioniemi, KRister Wennerber, Satu Mustjoki and Kimmo Porkka, for providing an excellent environment and resources for research. The supportive attitude and timely response of Tero Aittokalio to many questions concerning this study have been invaluable. I would also wish to thank all former and current fellow Ph.D. students in the project for their collaboration and feedback on the tools I developed in the project.

Henri Sara, my Master’s thesis supervisor, is especially thanked for teaching me R, microarray analysis and helping with optimizing my R code when writing algorithms.

I thank Dr Päivi Östling for always initiating interesting bioinformatic projects to work on and for supporting, promoting and beta testing my bioinformatic tools. It has been great sharing an office with you, and during these many years you have taken care of so many things that sometimes I wonder how you manage. I specifically want to thank you for the great teamwork, drive for excellence and support towards building Breeze.

I specifically thank the FIMM Breeze team, Östling Päivi, Jani Saarela, Dmitrii Byckov, Swapnil Potdar, Fiere Clement and He Liye, for bearing with my inexperienced way of running things. I am very grateful for the skills I have gained
while working with you. In addition, at the FIMM Technology Center, Timo Miettinen, Olle Hansson, Kari Tuomainen, Kyösti Sutinen, and Teemu Perheentupa have always been very helpful with all our IT needs.

I want to thank my friends Sam Mukoza, William Bukyanagandi, Aschwin van der Woude, Kaisa Seppäla, Charles Wasswa, Rebekka Muyanja, Micheal Muyanja, Peter Kalanzi, Bernard Kawere and Akinrinade-Oyediran O for all their support and encouragement.

I sincerely thank my parents (John Baptist Nsanja and Margret Nakiganda), brothers and sisters for their encouragement and love. Lastly, I thank my wife Susan Nabukeera for taking care of our lovely daughters Piia Mpindi and Petra Mpindi during the whole journey of my Ph.D. and for tolerating my long hours of work.

Helsinki, March 2016
John Patrick Mpindi
9. References

References

32. Kilpinen et al In silico transcriptomics (IST) database.
36. Kilpinen et al In silico transcriptomics (IST) database.
References


84. [https://www.researchgate.net/]. ResearchGate: A network dedicated to science and research..
References


86. [http://www.nature.com/onc ]. Oncogene.


References


References

References


