Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

Niclas Björn
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

Niclas Björn

Linköping University Medical Dissertations No. 1700

Division of Drug Research
Department of Medical and Health Sciences
Linköping University, Sweden
Linköping 2019
The cover illustration was made by Esther Camín according to the author's wishes and was printed and published with the artist's permission.

During the course of the research underlying this thesis, Niclas Björn was enrolled in Forum Scientium, a multidisciplinary doctoral program at Linköping University, Sweden.

Published articles have been reprinted with the permission of the copyright holder.

Printed in Sweden by LiU-Tryck, Linköping, Sweden, 2019

ISBN 978-91-7685-004-6
ISSN 0345-0082
Your genetics is not your destiny
George M. Church
Supervisor
Henrik Gréen, Linköping University, Sweden

Co-supervisors
Svante Vikingsson, Linköping University, Sweden
Joakim Lundeberg, Royal Institute of Technology, Sweden

Faculty opponent
Johan Staaf, Lund University, Sweden
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>1</td>
</tr>
<tr>
<td>POPULÄRVETENSKAPLIG SAMMANFATTNING</td>
<td>3</td>
</tr>
<tr>
<td>LIST OF PAPERS</td>
<td>5</td>
</tr>
<tr>
<td>ABBREVIATIONS</td>
<td>7</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>9</td>
</tr>
<tr>
<td>Cancer</td>
<td>9</td>
</tr>
<tr>
<td>Non-small cell lung cancer</td>
<td>10</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>11</td>
</tr>
<tr>
<td>Toxicity</td>
<td>11</td>
</tr>
<tr>
<td>CTCAE</td>
<td>13</td>
</tr>
<tr>
<td>Myelosuppression</td>
<td>14</td>
</tr>
<tr>
<td>Pharmacogenomics</td>
<td>15</td>
</tr>
<tr>
<td>Source of DNA/RNA</td>
<td>16</td>
</tr>
<tr>
<td>Personalized medicine</td>
<td>16</td>
</tr>
<tr>
<td>Bioinformatics and next-generation sequencing</td>
<td>17</td>
</tr>
<tr>
<td>AIMS</td>
<td>19</td>
</tr>
<tr>
<td>MATERIALS AND METHODS</td>
<td>21</td>
</tr>
<tr>
<td>Patient cohorts</td>
<td>21</td>
</tr>
<tr>
<td>Ethical considerations</td>
<td>22</td>
</tr>
<tr>
<td>Measurements of myelosuppressive toxicity</td>
<td>22</td>
</tr>
<tr>
<td>Next-generation sequencing (NGS)</td>
<td>24</td>
</tr>
<tr>
<td>Sequencing library preparation</td>
<td>24</td>
</tr>
<tr>
<td>Sequencing by synthesis (SBS)</td>
<td>25</td>
</tr>
<tr>
<td>Applied bioinformatics and statistics</td>
<td>27</td>
</tr>
<tr>
<td>Sequencing data pre-processing</td>
<td>27</td>
</tr>
<tr>
<td>Variant calling</td>
<td>28</td>
</tr>
<tr>
<td>Gene expression quantification</td>
<td>30</td>
</tr>
<tr>
<td>SNV/INDEL analysis</td>
<td>31</td>
</tr>
<tr>
<td>Functional annotation</td>
<td>32</td>
</tr>
<tr>
<td>Prediction models</td>
<td>33</td>
</tr>
</tbody>
</table>
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

Dimensionality reduction and clustering .......................................................... 35

RESULTS AND DISCUSSION ........................................................................ 37

Paper I ........................................................................................................ 37
  ABCB1 and CYP2C8 genotyping ............................................................... 38
  Progression-free survival and overall survival ....................................... 38
  Toxicity ..................................................................................................... 39
  Inconclusive or informative? .................................................................... 41

Paper II ........................................................................................................ 42
  SNVs/INDELs, FDR, and gene-based tests ............................................. 42
  Hematopoiesis-related genes and pathways ........................................... 43
  Thrombocytopenia prediction models ..................................................... 44

Paper III ....................................................................................................... 46
  High and low coverage WES................................................................. 46
  Called variants, coverage, and genotype quality .................................... 47
  Discordant variant calls .......................................................................... 47

Paper IV ....................................................................................................... 49
  SNVs as a starting ground for GNMs ...................................................... 49
  GNMs enriched with affected functional elements .................................. 50
  Prediction model for maximal toxicity .................................................... 50

Paper V ........................................................................................................ 52
  Cultivation, treatment, and cell cycle analysis ....................................... 53
  Transcriptionally different cell clusters .................................................. 54
  Going forward with scRNA-seq............................................................... 56

CONCLUSIONS ........................................................................................... 57

FUTURE ASPECTS ...................................................................................... 59

ACKNOWLEDGEMENTS ............................................................................ 61

REFERENCES ............................................................................................. 63

APPENDIX .................................................................................................. 75
Cancer is a serious disease expected to be the world-leading cause of death in the 21st century. The use of harsh chemotherapies is motivated and accepted but, unfortunately, is often accompanied by severe toxicity and adverse drug reactions (ADRs). These occur because the classical chemotherapies’ common modes of action effectively kill and/or reduce the growth rate not only of tumour cells, but also of many other rapidly dividing healthy cells in the body. There are also considerable interindividual differences in ADRs, even between patients with similar cancers and disease stage treated with equal doses; some experience severe to life-threatening ADRs after one dose, leading to treatment delays, adjustments, or even discontinuation resulting in suboptimal treatment, while others remain unaffected through all treatment cycles. Being able to predict which patients are at high or low risk of ADRs, and to adjust doses accordingly before treatment, would probably decrease toxicity and patient suffering while also increasing treatment tolerability and effects. In this thesis, we have used next-generation sequencing (NGS) and bioinformatics for the prediction of myelosuppressive ADRs in lung and ovarian cancer patients treated with gemcitabine/carboplatin and paclitaxel/carboplatin.

Paper I shows that \textit{ABCB1} and \textit{CYP2C8} genotypes have small effects inadequate for stratification of paclitaxel/carboplatin toxicity. This supports the transition to whole-exome sequencing (WES) and whole-genome sequencing (WGS). Papers II and IV, respectively, use WES and WGS, and demonstrate that genetic variation in or around genes involved in blood cell regulation and proliferation, or genes differentially expressed at chemotherapy exposure, can be used in polygenic prediction models for stratification of gemcitabine/carboplatin-induced myelosuppression. Paper III reassuringly shows that WES and WGS are concordant and mostly yield comparable genotypes across the exome. Paper V proves that single-cell RNA sequencing of hematopoietic stem cells is a feasible method for elucidating differential transcriptional effects induced as a response to \textit{in vitro} chemotherapy treatment.

In conclusion, our results supports the transition to genome-wide approaches using WES, WGS, and RNA sequencing to establish polygenic models that combine effects of multiple pharmacogenetic biomarkers for predicting chemotherapy-induced ADRs. This approach could be applied to improve risk stratification and our understanding of toxicity and ADRs related to other drugs and diseases. We hope that our myelosuppression prediction models can be refined and validated to facilitate personalized treatments, leading to increased patient wellbeing and quality of life.
POPULÄRVETENSKAPLIG SAMMANFATTNING


I den här avhandlingen har nya sekvenseringsmetoder använts för att läsa hela eller delar av arvsmassan från patienter med lung- eller äggstockscancer som behandlats med cellgifterna gemcitabin, karboplatin och paklitaxel. Vi har sökt efter skillnader i patienternas arvsmassa och tagit reda på hur de hör ihop med cellgifternas allvarliga blodcellsbiverkningar (vilket under behandlingen ses som kraftigt minskat antal blodceller). Skillnaderna vi hittat i arvsmassan återfinns till stor del i biologiska signalsystem som är viktiga för blodcellernas normala produktion och funktion. Avhandlingen visar att man kan kombinera effekten av dessa skillnader i matematiska modeller. Modellerna kan kategorisera patienterna i olika riskgrupper och förutspå blodcellsbiverkningar. Metodiken kan framöver tillämpas för att förbättra förståelsen och riskkategoriseringen för andra läkemedelbiverkningar.

Förhoppningen är att de utvecklade modellerna för att förutspå risken för blodcellsbiverkningar till följd av cellgiftsbehandling kan optimeras och valideras i kommande studier. Detta med målet att i framtiden kunna vägleda individanpassade behandlingsbeslut i vården, vilket ger patienter bättre behandlingseffekt, mindre lidande och ökad livskvalitet.
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions
LIST OF PAPERS

The five papers listed below are included in this thesis and referred to in the text by their roman numerals (I-V).

I. **ABCB1** variation affects myelosuppression, progression-free survival and overall survival in paclitaxel/carboplatin treated ovarian cancer patients

II. Genes and variants in hematopoiesis-related pathways are associated with gemcitabine/carboplatin-induced thrombocytopenia

III. Comparison of variant calls from whole genome and whole exome sequencing data using matched samples

IV. Whole-genome sequencing and gene network modules predict gemcitabine/carboplatin-induced myelosuppression in non-small cell cancer patients

V. Single-cell RNA sequencing of hematopoietic stem cells treated with gemcitabine and carboplatin
    NICLAS BJÖRN, Ingrid Jakobsen, Kourosh Lotfi, and Henrik Gréen. Submitted - to BMC Genomics.
Other relevant co-authored papers, not included in the thesis:

**The effect of *JMJD1C* knockdown on myeloid cell lines proliferation, viability, and gemcitabine/carboplatin-sensitivity**

**Genetic association of gemcitabine/carboplatin-induced leukopenia and neutropenia in non-small cell lung cancer patients using whole-exome sequencing**
Anna Svedberg, Benjamín Sigurgeirsson, NICLAS BJÖRN, Sailendra Pradhananga, Eva Brandén, Hirsh Koyi, Rolf Lewensohn, Luigi De Petris, Maria Apellániz-Ruiz, Cristina Rodríguez-Antona, Joakim Lundeberg, and Henrik Gréen. Manuscript.

**Early changes in gene expression profiles in AML patients during induction chemotherapy**
Ingrid Jakobsen, Max Sundkvist, NICLAS BJÖRN, Kourosh Lotfi, and Henrik Gréen. Manuscript.
ABBREVIATIONS

ADR  adverse drug reaction
AE   adverse event
AUC  area under the curve
BQSR base quality score recalibration
BSA  body surface area
BWA  Burrows-Wheeler aligner
CADD combined annotation dependent depletion
CML  chronic myelogenous leukaemia
CTCAE common terminology criteria for adverse events
DNA  deoxyribonucleic acid
FDR  false discovery rate
HET  heterozygous
HOM  homozygous
HR   hazard ratio
HSC  hematopoietic stem cell
GATK genome analysis toolkit
GNM  gene network module
GO   gene ontology
GWAS genome-wide association study
IC_{50} the half-maximal inhibitory concentration
INDEL insertion or deletion
LASSO least absolute shrinkage and selection operator
LD   linkage disequilibrium
MAF  minor allele frequency
MCODE molecular complex detection
NGS  next-generation sequencing
NSCLC non-small cell lung cancer
OR   odds ratio
OS   overall survival
PCA  principal component analysis
PCR  polymerase chain reaction
PD   pharmacodynamics
PFS  progression-free survival
PK   pharmacokinetics
PPI  protein-protein interaction
REF  reference
RNA  ribonucleic acid
ROC  receiver operating characteristic
SBS  sequencing by synthesis
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>scRNA-seq</td>
<td>single-cell RNA sequencing</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SNP</td>
<td>single-nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>single-nucleotide variant</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-distributed stochastic neighbour embedding</td>
</tr>
<tr>
<td>UMAP</td>
<td>uniform manifold approximation and projection</td>
</tr>
<tr>
<td>VQSR</td>
<td>variant quality score recalibration</td>
</tr>
<tr>
<td>WES</td>
<td>whole-exome sequencing</td>
</tr>
<tr>
<td>wGRS</td>
<td>weighted genetic risk score</td>
</tr>
<tr>
<td>WGS</td>
<td>whole-genome sequencing</td>
</tr>
</tbody>
</table>
Cancer
Cancer is a common, gruelling, and debilitating disease that comes in many forms. It is expected to affect every third person and to be the world-leading cause of death in the 21st century [1, 2]. Cancer arises due to the acquisition of somatic mutations in normal cells leading to abnormal cell growth and the ability to spread to other tissues. These biological characteristics are acquired in a multistep developmental process leading to the formation of malignant tumours, summarized and overviewed by Hanahan and Weinberg in the Hallmarks of Cancer [3, 4]. The complexity of cancer is further increased by the effects of hereditary, environmental, and lifestyle factors. Cancers, even of the same type and stage, are heterogeneous between patients, and cancer cells within tumours are also heterogeneous. This complicates cancer treatment standardization and guidelines. Treatments include, but are not limited to, surgery, radiation, as well as newer, targeted, and older, classical chemotherapies.

Cancer therapies often induce considerable toxicity and adverse drug reactions (ADRs) in patients. We believe that selection of the right drugs and doses for each patient can be tailored in more personalized treatment schedules. This is dependent on selecting the most suitable drug for the tumour and the most tolerable dose for the patient. In this thesis, we focus on the latter by investigating patients’ constitutional genetics and its ability to predict patient response in terms of induced toxicity and ADRs. We have
done so using the DNA of patients with lung cancer and ovarian cancer treated with carboplatin, gemcitabine, and paclitaxel for predicting the commonly associated myelosuppressive toxicity induced by the drugs and then working towards finding pharmacogenetic biomarkers for predicting and understanding chemotherapy-induced ADRs. Although we have used a selected set of drugs, cancer types, and toxicities, our results and methods can hopefully be applied in future studies of pharmacogenetics and personalized medicine for chemotherapies in general.

**Non-small cell lung cancer**

In Sweden, lung cancer is the sixth most common cancer, yearly affecting around 3500 people, and is the most common cancer-related cause of death [1, 5]. It is common, lethal, with a similar pattern worldwide of high mortality and a relative five-year survival of only about 20% [1, 2, 5]. This is due to the tumour’s high growth rate, ability to metastasize early, and development of resistance to treatments, which makes relapse common. In addition, the symptoms of lung cancer are vague and nonspecific, which make the disease hard to detect in its earlier stages when the prognosis is often better. Lung cancer is divided into two main types of lung cancer, small cell lung cancer and non-small cell lung cancer (NSCLC). Patients with small cell lung cancer are usually treated with chemotherapy directly. NSCLC treatment today consists of surgery, radiation, and chemotherapy. Now, the first-line chemotherapy consists of targeted therapies or immune checkpoint inhibitors. But, depending on their success the older classical chemotherapies are still often used at later stages. The NSCLC patients in the presented studies were treated with gemcitabine and carboplatin combination chemotherapy, which was the standard treatment at the time, 2006-2008, and place, Karolinska University Hospital, Stockholm, Sweden, of the study.

Because of the bad prognosis for lung cancer, tough treatment regimens are both motivated and accepted. This is seen in gemcitabine/carboplatin treatment where myelosuppression is the most common ADR, expressed as anaemia, leukopenia, neutropenia, and thrombocytopenia [6-9]. Severe levels are observed in 50-70% of patients [8, 10-14]. These toxicities lead to increased patient suffering and hospitalization costs, and necessitate dose delays, adjustments, or discontinuation, none of which are optimal in the treatment of cancer.

The mechanisms of action of both drugs affect the synthesis and replication of DNA in cells resulting in reduced growth rate and increased cell death. Gemcitabine is a cytidine analogue with several properties responsible for its cytotoxic effects such as DNA polymerase inhibition, ribonucleoside reductase inhibition, and irreparable incorporation of the faulty cytidine
analogue in the DNA which leads to termination of DNA polymerization. All these effects intensify gemcitabine’s inhibition of DNA synthesis and induction of apoptosis [15]. Carboplatin binds guanine and adenine resulting in unwinding, bending, and cross-linking of DNA which prevents transcription and replication, and initiates apoptosis [9, 16, 17].

Any underlying genetic differences contributing to myelosuppression are today not understood, although, some biomarkers have been proposed in the literature [14, 18-23]. In Papers II and IV we have applied more agnostic exome and genome sequencing methodologies to predict this toxicity.

**Ovarian cancer**

Ovarian cancer in Sweden and world-wide represents about 1-2% of newly diagnosed cancer cases in women [1, 5, 24]. The five-year survival rate is around 50% but varies greatly depending on the stage of the cancer [5, 24]. First-line chemotherapy against ovarian cancer is normally a combination of paclitaxel and carboplatin (although other combinations are used) [25-27]. This treatment is also associated with considerable toxicity with a severe impact on the patient’s quality of life and practical aspects of the therapy, due in particular to neuropathy and neutropenia (but also myelosuppression in general). This frequently also results in dose delays, reductions, and cessation [25, 28-34]. The neuropathy is dose-dependent and about 70% of the patients suffer from neuropathy. In most cases, treatment is discontinued before the neuropathy develops too far, but in some cases the effects of neuropathy remain with the patients for the rest of their life.

Paclitaxel targets tubulin in the cytoskeleton, stabilizing it and preventing disassembly. This makes chromosome spindle configuration in the metaphase impossible and blocks mitosis.

Previous studies on paclitaxel have shown that genetic variants in *CYP2C8* and *ABCB1* affect the pharmacokinetics and treatment-associated toxicity [29, 30, 35-48]. These previous findings are contradictory, in particular for toxicity, and we hope, by using the candidate gene approach in Paper I, to shed some light on the reason for this.

**Toxicity**

The classical chemotherapies are chemical substances that kill and/or reduce the growth rate of tumour cells and other rapidly dividing cell types. These are exemplified by carboplatin, gemcitabine, and paclitaxel which are used in the treatment of NSCLC and ovarian cancer, as mentioned
previously. This general mechanism of action affects not only the patient’s cancer cells but also their healthy cells resulting in toxicity. These toxicities can take many forms from nausea, fatigue, pain, and alopecia, to myelosuppression and neuropathy [49]. Chemotherapies have narrow therapeutic windows, where even very small variations in the dosage can result in a lack of efficacy or toxic side effects. Furthermore, in the case of chemotherapy, the therapeutic and the toxic dose often overlap. Currently, most chemotherapy doses are based on the body surface area (BSA) of the patient, calculated from their height and weight. Some drug doses are also adjusted based on the patient’s renal function, for example, carboplatin [50]. However, the variation in, for example, BSA is far less than the variation in drug exposure, clinical effects, and the activity of drug-metabolizing enzymes and drug transporters [51]. This variation is largely due to genetic variation in genes encoding transporters and metabolizing enzymes. The patient’s genetic make-up is therefore of great importance for predicting toxicity and treatment response.

Knowledge on the relationship between chemotherapies and cancer biology has outpaced our understanding of their toxicity [52]. The ADRs cause substantial discomfort and distress to patients and their relatives, limit treatment tolerability and efficacy and in some cases, they may even remain with survivors after treatment. Newer and more targeted therapies are more specific and thought to have less toxicity than the classical therapies; however, even these are associated with considerable toxicity [53].

Toxicity is often viewed as a necessary evil. Doctors might comment that when the patient experiences toxicity we know that the treatment is working. In my opinion, this is an uncaring and unnecessary comment that should be used with caution. However, some studies on a variety of treatments and cancers do show that toxicity correlates with increased survival [54-57]. Toxicity and its accepted severity must be, and usually are, viewed in relation to the treatment and the patient. Palliative treatments should accept less toxicity because the patient’s wellbeing is more important, while curative treatments usually accept more toxicity. However, toxicity should not be viewed as a necessity. Higher levels of toxicity often require dose alterations resulting in suboptimal treatments. Being able to adjust the dosage at the start of treatment to a tolerable dose that can be maintained for the duration of the treatment is, potentially, a vast improvement. Another issue is that patients who do not experience toxicity could be undertreated. Being able to identify these patients and start them on higher doses could potentially lead to a substantial improvement in their overall cancer treatment.
Although the patients’ wellbeing and treatment should be a focus in relation to ADRs, the hospitalization costs must also be considered. ADRs and unmanaged toxicity lead to emergency hospital visits and admissions placing a financial burden on the healthcare system. The exact cost is difficult to estimate and depends on the drugs, their toxicity, the structure of the healthcare system, and the method used for evaluation [58]. Some estimates for chemotherapy-induced toxicity are in the ballpark of thousands of dollars per patient in the US [59, 60]. Although most drugs are safe, ADRs still constitute between 2.0% to 12.0% of all hospital admissions [61-67], even when administered and used appropriately. The incidence of deaths related to ADRs is estimated to a few percent (up to 5%) of all deaths [62, 63, 68, 69]; however, these numbers are hard to determine and interpret. Moreover, these ADR-related deaths are often thought to be preventable. Chemotherapy is harsh and has narrow therapeutic windows and it is easy to understand that ADRs and toxicity are of the utmost importance in chemotherapy-related patient suffering.

There is a substantial need for better use of these drugs, but it is a tough nut to crack. On the one hand, the treatment must kill the cancer cells and on the other, killing too many of the healthy cells is dangerous. Balancing on this fine line will be more easily achieved with better individualisation of therapies. In the next paragraph, I will introduce the most commonly used scoring and evaluation system for ADRs.

**CTCAE**

The evaluation and reporting of adverse events (AEs) has always been a critical part of clinical trials, in particular for cancer therapies which often have narrow therapeutic windows and are associated with induction of toxicity due to their harsh nature and broad mechanisms of action. Stringent criteria were needed for evaluating and facilitating decision-making concerning risks and benefits of treatments and their continuation. Therefore, the Common Toxicity Criteria were introduced by the National Cancer Institute in 1983 [70, 71]. This has since been updated several times and is now referred to as the Common Terminology Criteria for Adverse Events (CTCAE). The most recent version (5.0) was made available in 2017, but the studies presented in this thesis have been evaluated using version 4.03 from 2010. However, there are no differences between the two versions of the CTCAE for the specific myelosuppressive toxicities evaluated in this thesis.

The CTCAE was initially intended primarily for the evaluation of cancer therapies, but is now used in clinical trials for many different types of drugs. The CTCAE grades or similar measurements are now also used to some
extent in normal clinical settings to assist decision-making related to treatment schedules, doses, and toxicity.

In general, the CTCAE has five grades defining the severity of AEs. However, we also use grade zero to denote a normal state or no AE, as do many others. The grades have unique descriptions for all AEs based on the guidelines in Table I.

<table>
<thead>
<tr>
<th>CTCAE</th>
<th>Interpretation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>None</td>
<td>Normal or below mild measurable levels. No intervention indicated</td>
</tr>
<tr>
<td>1</td>
<td>Mild</td>
<td>Asymptomatic or mild symptoms. No intervention indicated</td>
</tr>
<tr>
<td>2</td>
<td>Moderate</td>
<td>Limiting instrumental activities of daily life. Minimal, local, or noninvasive intervention indicated</td>
</tr>
<tr>
<td>3</td>
<td>Severe</td>
<td>Medically significant, disabling, or limiting self-care activities of daily life. Hospitalization indicated</td>
</tr>
<tr>
<td>4</td>
<td>Life-threatening</td>
<td>Urgent intervention indicated</td>
</tr>
<tr>
<td>5</td>
<td>Death</td>
<td>Death related to the AE</td>
</tr>
</tbody>
</table>

**Myelosuppression**

Although there are many toxicities associated with chemotherapy, we have focused on myelosuppression (also known as bone marrow suppression or myelotoxicity) which is commonly induced by the drugs carboplatin, gemcitabine, and paclitaxel used in the presented studies. Myelosuppression is mainly the decreased production of neutrophils, leukocytes, platelets, and erythrocytes (not investigated in this thesis). The severity of myelosuppression is measured by taking a blood sample and counting the number of blood cells present. Many of the classic chemotherapies fiercely attack the production of blood cells in bone marrow because these cells proliferate at a fast rate. This can rapidly lead to a substantial reduction in the number of blood cells in patients. After only one cycle of treatment many patients exhibit a severe reduction, whereas others are basically unaffected during the entire course of treatment.

This deficit in blood cells affects patients in different ways. Loss of the leukocytes renders the patient vulnerable to infections; loss of platelets can lead to severe spontaneous bleeding; and loss of erythrocytes can lead to fatigue, weakness, and shortness of breath. Myelosuppression can be managed to a certain extent by administering G-CSF and blood transfusions, but often also necessitates treatment changes, delays, or cessations resulting in suboptimal cancer treatment.
The approach we have taken to investigate myelosuppressive toxicity using next-generation sequencing (NGS) should also be applicable for finding new means of predicting common and rare toxicities induced by many other types of drugs for numerous diseases.

**Pharmacogenomics**

It is well known that drug therapies have significant inter-individual variability in therapeutic effect, toxicity, and ADRs. Now, one of the most important factors responsible for these differences is recognized to be normal genetic variation, which has been estimated to account for roughly 20-95% of the variability seen in drug disposition and effects [72, 73]. The field of pharmacogenomics, which combines pharmacology and genomics, studies the role of the genome in relation to drug response. The initial focus of this field was the candidate genes in pathways involved in drug absorption, distribution, metabolism, and excretion, and usually involved many of the classical pharmacokinetic (PK) genes. To some extent, genes known to be important for drug pharmacodynamics (PD) have also been investigated. However, despite thousands of publications, this has shown limited usability for clinically used drugs in general [74-80]. Nonetheless, there are numerous examples of genetic sequence variations in genes coding for drug targets, metabolizing enzymes and transporters known to affect drug effects, and the list of genetic variants is continuously growing [73, 78, 80-82]. Two examples for chemotherapy are the importance of TPMT and DPYD mutations [75, 76, 83-85], where genotyping can determine whether a patient is at risk of severe ADRs; however, their clinical use is still limited.

Technological advancements in the last ten years have made NGS readily available to scientists, and have facilitated the move from candidate gene to genome-wide studies. This allows for more agnostic genome-wide studies without the biased decision to look only at genes that are potentially important. This is probably particularly useful for common phenotypes, where single rare genetic variants probably have limited effect, such as for many of the commonly induced toxicities from chemotherapy. This is also probably the reason why, for many drugs, the effect of single genetic variants shown in some studies has proven difficult to replicate. It works well for the rare phenotype rare genotype case, but the more common the phenotype, the more likely the genetic component will be affected by multiple genotypes, both common and rare, and with small and big effect sizes. Therefore, the field is currently moving from focussing on a single or a couple of genetic variants to integrating multiple factors in polygenic models for predicting phenotypes.
Pharmacogenomics has seen an increased interest from the public via the many companies now emerging to offer DNA sequencing consumer products. Using these products, people can acquire information not only about hereditary traits and the risk of certain diseases but also about their genetic make-up for some metabolizing enzymes, which can be important when prescribing standard drugs. In some instances, the patient now knows more about the importance of pharmacogenomics than their treating doctor.

One of the problematic areas remaining is that even if mutations are known to be highly related to a specific toxicity (or another phenotype for that matter), if these mutations are very rare the implementation of genetic analyses is not cost-effective. However, this will hopefully change in the future with the decreasing sequencing costs and the fact that some tests could be screened for at an earlier stage, or as people start to use consumer genetics profiling and share the results with healthcare providers. The good thing is that the germline genetics of a patient, once known, should remain the same. Thus, a broad genome-wide test should be applicable for other drugs at a later timepoint.

**Source of DNA/RNA**

In regard to cancer, we must remember that two genomes could be investigated, the germline genome and the cancer genome. The toxicity experienced by patients undergoing chemotherapy is a systemic response to the drugs affecting the patient’s normal cells. Therefore, we have focused on using the patients’ normal germline genetics to be able to predict patient response and toxicity in relation to the drugs. However, when selecting the most optimal drug, knowledge is needed on the somatic mutations in the cancer. Therefore true, personalized medicine in cancer chemotherapy will need to use both genomes. The same is true for gene expression analyses, for example, using RNA sequencing. Here, the expression in normal tissue is relevant in relation to toxicity, and the expression in tumour tissue is relevant in relation to whether or not a specific drug should be used.

**Personalized medicine**

The use of pharmacogenomics for personalized medicine (also known as precision medicine) is a long-sought goal for many researchers. Personalized medicine hopes to minimize toxicity while maximizing therapeutic effect through administering a suitable drug at the most appropriate dose to the right patient (Figure 1). Personalized medicine is a much broader term than pharmacogenetics and can include many additional non-genetic factors. However, pharmacogenetics is an important cornerstone of personalized medicine.
Personalized medicine will also require extensive replication and validation studies to prove that the doses and drugs used for the patients, when stratified using biomarkers, lead to a clear benefit measured in decreased toxicity and/or longer survival [74, 79, 80, 83, 86]. This is, and will be, a very important aspect of many association studies, albeit outside the scope of this thesis. We focus on finding models for risk stratification and understanding toxicity, while we hope the models can be used clinically in the future, they must first be validated in further studies.

![Figure 1. The visualized goal of personalized medicine in which each patient receives the most suitable drug and dose for them by integrating pharmacogenomic, or other, biomarkers in the clinic before the start of treatment.](image)

**Bioinformatics and next-generation sequencing**

In the beginning, pharmacogenetics did not include as many genes or as much data from as many sources as it does today. After the introduction of NGS, the amount of data has increased significantly, from genotyping of a few mutations to sequencing all genes, and from a couple of patients to hundreds or thousands. The data can also be merged with sequences of noncoding regions and put in relation to the expression of all genes in relevant tissues. This transition to using large-scale multilevel data has come with the need for computational data-driven analyses carried out on supercomputing clusters. Therefore, scientific research on pharmacogenomics and personalized medicine now relies on biologists, computer scientists, and bioinformaticians (with experience in both biology and computer science) working together to make use of computers to analyse biological data and to make sense of it all using bioinformatics.

Bioinformatics includes everything from pre-processing and getting the data ready for analysis to developing novel tools for analysing the data and applying existing tools and protocols to, finally, trying to draw conclusions from the results. In the present studies, I have mainly applied and integrated existing tools for analysing and understanding our data.
NGS generates vast amounts of data that must be pre-processed to ensure data quality and reliability before they can be used for genotyping or expression analysis. These data can then be used for various analyses for finding and evaluating correlations and patterns in the data associated with the phenotype of interest (in our case the myelosuppressive toxicity induced by chemotherapy). Because thousands of tests are carried out, correlation and association analysis will most probably produce results. The real bioinformatics’ challenge here lays in making sense of all the results and correlations, and their applicability for understanding or predicting the phenotype. This can be partly done by validation, where one investigates whether the same associations can be found in other materials. But another means for understanding the relevance of the results is overlap and enrichment in known biological systems (such as pathways or gene ontologies (GOs)), genes expressed in relevant tissues, and protein-protein interaction networks. Multiple tools are available for predicting severity or effect of genetic variants which can be used for understanding the relevance or consequence of genetic variants for the phenotype. All this can be assessed and help to understand the associations relevance and importance for the phenotype.

Lastly, the most important property of bioinformatics is to be able to generate information that is useful and adds value, not only for the research community but also for clinicians and patients. In this thesis this is done by implementing models for understanding, predicting, and stratifying the patients’ risk of chemotherapy-induced toxicities. For this purpose, sophisticated polygenic scoring methods, weighted genetic risk score (wGRS) [87-93], logistic regression, and random least absolute shrinkage and selection operator (LASSO) [94], have been used to combine the effect of many genetic variants into reduced models for personalized medicine. These models need to be simple, effective, sensitive, and sufficiently robust while adding value to the patients’ treatment; if not, even the best models will not be used by the treating doctor in the clinic. Not to be forgotten is that, before any model can or should be introduced in the clinic, they must undergo rigorous validation to see whether patients classified to be at low risk or high risk of toxicity benefit from increased or reduced doses. The low-risk patient with increased doses should achieve better progression-free survival (PFS) or overall survival (OS) without the induction of life-threatening toxicity. While high-risk patients treated with reduced doses should experience less toxicity without the cost of decreased PFS or OS.
The overall aim was to investigate and improve the effectiveness and safety of cancer chemotherapy by utilizing patients’ constitutional genotypes for predicting and stratifying the risk of chemotherapy-induced ADRs. The hope is that the results can be used as pharmacogenetic biomarkers to individualize drug therapies and to reduce the risk of ADRs in the future. This would lead to improved treatments in terms of both the anti-tumour effect of the treatment and patient wellbeing.

The specific aims of the thesis were to

- evaluate $ABCB1$ and $CYP2C8$ effects on myelosuppression, PFS, and OS in paclitaxel/carboplatin treatment.

- use WES to find pharmacogenetic biomarkers for gemcitabine/carboplatin-induced thrombocytopenia.

- compare WES and WGS genotyping quality across the exome.

- use WGS for constructing genome-wide risk prediction models for gemcitabine/carboplatin-induced myelosuppression.

- evaluate whether scRNA-seq of HSCs is a feasible approach for identifying differential transcriptional effects induced by chemotherapies in vitro.
MATERIALS AND METHODS

In this section, the methods used to obtain the main findings in Papers I-V are described and discussed. However, details of all the methods employed can be found in the respective papers.

Patient cohorts

Three patient cohorts have been used in the five studies included in this thesis. Table II presents an overview of the number of patients and the papers in which they were included. For Papers I, II, III, and IV, peripheral blood samples taken before treatment start were used for genotyping, i.e. for looking at the germline (constitutional) genetics of the patients, and do not include any genetic data reflecting somatic mutations of the patients’ cancers. In Paper V, we used leftover harvested hematopoietic stem cells (HSCs) from patients with chronic myelogenous leukaemia (CML) and, due to the nature of CML, the harvests include both normal and cancerous cells. Clinical data, including patient characteristics and registered toxicity/ADRs, were retrieved from medical records.

In Paper I we investigated the relationship between \textit{ABCB1} and \textit{CYP2C8} genotypes and progression-free survival (PFS), overall survival (OS), and myelosuppression in ovarian cancer patients treated with paclitaxel/carboplatin. In Paper II we used whole-exome sequencing (WES) to investigate gemcitabine/carboplatin-induced thrombocytopenia
in NSCLC patients. In Paper IV we used whole-genome sequencing (WGS) to investigate gemcitabine/carboplatin-induced myelosuppression in 96 of the patients from Paper II. In Paper III we performed a technical comparison of genotype calls from WES and WGS utilizing the 96 patients sequenced using both methods in Paper II and Paper IV, respectively. In Paper V we performed single-cell RNA sequencing (scRNA-seq) of HSCs treated with gemcitabine and carboplatin from one of the included patients. This was a pilot study to showcase the feasibility of this approach for finding differential transcriptional effects induced by chemotherapy across cells.

**Table II. Overview of the studies.**

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Number of included patients</th>
<th>Sequencing approach</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer</td>
<td>525</td>
<td>Pyrosequencing</td>
<td>I</td>
</tr>
<tr>
<td>Non-small cell lung cancer</td>
<td>215</td>
<td>WES</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>WES and WGS</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td>96</td>
<td>WGS</td>
<td>IV</td>
</tr>
<tr>
<td>Chronic myelogenous leukaemia</td>
<td>20</td>
<td>scRNA-seq</td>
<td>V</td>
</tr>
</tbody>
</table>

**Ethical considerations**

The studies were performed with approval from regional ethics committees, and all the included patients gave informed consent in accordance with the Declaration of Helsinki.

The data handled within the studies were coded and anonymized so that the patients were not linked to their data. This minimized any potential for breach of personal information. Genetic information is regarded as personal sensitive data, and the risk of a privacy breach cannot be completely ruled out. However, this risk was considered to be extremely low due to the careful management of the coded and anonymized data. This low risk should be viewed in relation to the potentially large benefits arising from the results of the individual studies.

**Measurements of myelosuppressive toxicity**

Throughout the studies, we have used different measurements of toxicity, both categorical and continuous and, depending on the specific toxicity investigated or interpreted, the availability of a measurable continuous variable might be limited. Sometimes the variables are based on self-assessment or expert judgment, but in our studies the continuous variable is the actual count of neutrophils, leukocytes, and platelets in blood samples taken throughout the treatment cycles. This limits any
interpretation bias associated with different patient/physician-based toxicity assessments.

Neutrophil, leukocyte, and platelet counts were determined in blood samples taken during the treatment cycles. The lowest measured value during the given timeframe for each of the blood cell types, referred to as the *nadir*, was then used as the measurement of the myelosuppressive toxicities neutropenia, leukopenia, and thrombocytopenia. Our studies also gave us access to the baseline blood status measured immediately before treatment start. This can be used as a continuous co-variable in the statistical tests. Another approach that we used was to adjust the nadir values with the baseline variation to give the *decrease* (Equation 1) and *relative decrease* (Equation 2) parameters. Both represent the magnitude of the decrease from the baseline value to the nadir value. From a dose-response perspective, the decrease parameters are highly relevant as this is where a genotype-phenotype association would be expected. Genotypes affecting sensitivity to drugs should lead to a substantial decrease in the number of blood cells irrespective of the starting baseline value. By looking only at the nadir, patients with high baseline values might still have an “okay” blood status after treatment even if they have a large proportional decrease, meaning that if the nadir value only was used, the actual effect of some genotypes could be masked.

\[
decrease = \frac{\text{nadir}}{\text{baseline}}
\]  
Equation 1

\[
relative\;decrease = \frac{\text{baseline} - \text{nadir}}{\text{baseline}}
\]  
Equation 2

The categorical variables used in our studies are the nadir values of neutrophils, leukocytes, and platelets graded/scored according to the National Cancer Institute’s CTCAE version 4.03 (there is now a fifth version released in 2017, however without any changes to the toxicities investigated in Papers I-IV). In some instances, we also combined CTCAE grades 3 and 4 in a high toxicity variable, CTCAE grades 0-1 or 0-2 in a low toxicity variable, and occasionally CTCAE grade 2 as an intermediate toxicity variable. Exactly how the levels for these should be set depends both on the number and availability of patients, but mainly on the CTCAE grades on which clinical decisions such as treatment modifications or cessations are made. For myelosuppressive toxicities, this is generally when patients experience CTCAE grade 3 or higher. In Paper IV we also look at the highest registered CTCAE grade of neutrophils, leukocytes, and platelets overall as a maximal toxicity variable.
From a clinical perspective, the nadir values and their CTCAE interpretations are probably the most valuable and easiest to implement in clinical practice due to their standardization and regular use in clinical assessments. However, after measuring the patient’s baseline value, the risk of a large decrease can easily be recalculated back to the risk of a given CTCAE grade.

Next-generation sequencing (NGS)

DNA sequencing studies endeavour to determine the exact sequence of the billions of adenine (A), guanine (G), cytosine (C), and thymine (T) nucleotides in the human diploid genome which is efficiently stored in our chromosomes (1-22, X, and Y), whereas RNA sequencing studies endeavour to determine which genes are expressed by sequencing the bases in the RNA available in our cells.

DNA sequencing dates back to 1977 when Sanger et al. [95] published their method. However, this method was not appropriate for sequencing vast amounts of genetic material; more high-throughput sequencing techniques were needed. The first popular genome-wide technologies for assessing DNA and RNA were microarrays [96]. The full human genome was finally sequenced during the early 2000s [97-99], although it was labour-intensive and required substantial monetary investments to complete the task. Not long after, high-throughput massively parallel NGS has become readily available to the broader research community. This has been achieved through the development and commercialisation of numerous sequencing approaches [100-105].

In our sequencing projects, we have used machines from Illumina, based on sequencing by synthesis (SBS). This was developed in the study by Bentley et al. [100], who sequenced short sequences in parallel on a glass surface using fluorescent reversible terminator deoxyribonucleotides. This will be briefly described in the coming sections.

Sequencing library preparation

Before sequencing, DNA or RNA must be prepared into a sequencing library. The library preparation depends on the type of sequencing as well as the product brand. It is briefly outlined below:

- The WGS PCR free library preparation starts by fragmentation of the DNA. This is followed by reparation of blunt ends before index adapters are ligated to the DNA fragments.
• The WES library preparation starts with tagmentation (fragmentation and adaptor addition in a single step). PCR is then used to amplify the tagmented DNA and add indexes. Capture probes targeting the exome regions enrich that region before another round of PCR amplification.

• For bulk RNA-seq, unwanted RNA is removed either by rRNA depletion or mRNA enrichment. Thereafter the remaining RNA is reverse-transcribed into cDNA. The cDNA is then fragmented, amplified, and adaptors and indexes are added.

• The scRNA-seq starts with extraction of single cells into droplets containing beads with unique indexes that are incorporated into the cDNA during reverse-transcription of mRNA. The cDNA is then tagmented and amplified.

These preparations are also interspaced by clean-ups. The quantity and quality of the final libraries are then determined using a 2100 Bioanalyzer (Agilent Technology). The fragments now include indexes, sequencing primer binding sites, and adapters, see Figure 2.

Sequencing by synthesis (SBS)
In the projects in this thesis, we have performed short-read sequencing using NextSeq 500, HiSeq 2500, and HiSeq X Ten sequencing machines from Illumina (all supporting single-end and paired-end sequencing). They all utilize the same approach for sequencing with bridge amplification for cluster generation followed by SBS, briefly described below.

The sequencing libraries are added to the sequencing flow cell, a glass slide covered with a lawn of two types of oligonucleotides, one complementary to each adaptor. When the single-stranded fragments flow over the glass slide they randomly hybridize to the flow cell oligos. The DNA polymerase then synthesizes the other strand of the fragment. The double-stranded fragment is denatured, and the loose strand is washed away. The strands bend over and the top hybridizes to the other type of flow cell oligo. The DNA polymerase elongates the fragment to a double-stranded bridge which is later denatured. Two complementary fragments are now attached to the
flow cell in each cluster. This bridge amplification PCR process is repeated until the millions of clusters on the flow cell contain thousands of identical DNA fragments. All strands selectively binding to the second type of flow cell oligo are then cleaved off and washed away.

Then the actual SBS can start. This process uses reversible dye-terminator-bound deoxyribonucleoside triphosphates (dNTPs) which enables detection of single bases as the DNA polymerase synthesizes the complementary fragment. These dNTPs have specific fluorescent tags for A, C, T, and G which emit a characteristic wavelength when they are excited by a laser light source. This is a four-colour, two-colour, or single-colour system (depending on the sequencing machine). The dNTPs are 3’ blocked which means the DNA polymerase cannot continue because polymerization is terminated after the addition of one base to the fragment. However, this dye-terminator structure of the dNTPs is reversible. During SBS all dNTPs are present, and natural competition minimizes the DNA polymerase incorporation bias. Between each step, remaining unincorporated dNTPs are washed away before detection of the incorporated fluorescent base. After detection, the fluorescent dye is cleaved away and the 3’ hydroxyl group is restored so that the cycle can repeat and add the next base, and so on.

The signal from a single fluorescently-labelled base is not strong enough to be detected by the sequencer. But, from the clusters, thousands of bases emit the same signal simultaneously making it strong enough to be detected by the sequencing camera.

The SBS is then carried out roughly following these steps:

1. Read 1 is sequenced using the first sequencing primer.
2. The read 1 product is washed away.
3. Index 1 is sequenced.
4. The index 1 product is washed away.
5. The strand folds over and hybridizes to the second flow cell oligo.
6. Index 2 is sequenced from the bridge.
7. The index 2 product is washed away.
8. The DNA polymerase synthesizes the complementary strand.
9. The double-stranded bridge is denatured.
10. The strand on the first flow cell oligo is cleaved off and washed away.
11. Read 2 is sequenced using the second sequencing primer.

This is done in a massively parallel way for all clusters simultaneously. The base calls are made from the signal intensity and the base calls along with
their corresponding quality scores are then saved for all reads in FASTQ files.

Applied bioinformatics and statistics
Most analyses, comparisons, and figures presented in the Papers have been carried out using the free software environment for statistical computing and graphics, R [106], and extended using several R packages referred to when used. The computationally intensive sequencing data pre-processing was carried out on the high-performance computing cluster UPPMAX.

Sequencing data pre-processing
Before performing the main analyses on the sequencing data, the vast amounts of sequencing data, usually in FASTQ files, must go through several pre-processing steps. These steps ensure read quality and alignment to the genome before variant calling for DNA sequencing, and gene expression quantification for RNA-seq. I will here introduce the main procedures.

As a rule, it is always a good idea to start by assessing the overall quality of the sequencing reads, base quality scores, nucleotide distribution, GC content, overrepresented sequences, and sequencing adaptor content. We have mainly used FastQC for this. We have also used another tool, QualiMap, for assessing the quality of the later aligned data [107]. For summarizing quality metrics from multiple samples, the MultiQC tool [108] can merge all quality reports into a single comprehensible report which is much easier to interpret.

Next, the first main step of pre-processing is removal of adapter sequences and trimming low-quality bases from the sequencing reads. For this, we have applied cutadapt [109], which finds and removes adapter sequences, and performs quality trimming by removing bases from both of the read ends until the quality threshold is met. Lastly, cutadapt can control the read length and discard too-short reads.

After these steps, the data can be aligned to the reference genome (in our case the human reference genome, GRCh37 in Papers II, III, and IV, and GRCh38 in Paper V). For the alignment of the DNA-sequencing data in Paper II we used Bowtie2 [110], but for Papers III and IV, we used the Burrows-Wheeler aligner (BWA) [111, 112]. The alignment process in which the aligners identify the exact position (chromosome and base pair) of the bases in the reads is both tricky and time-consuming. For the RNA sequencing in Paper IV and scRNA-seq in Paper V, we used the software STAR [113] for rapid alignment [114]. RNA alignments are slightly different
and not as straightforward as alignment of DNA, mainly due to the splicing of exons to form RNA. The aligned data are outputted in SAM format (usually converted to the binary counterpart BAM to save space) and easily analysed, interpreted, and manipulated using Samtools [115]. In our study, Samtools was used to extract only the primary aligned reads (in the case of reads with multiple alignments the best alignment is kept and the others discarded) that mapped in proper pairs (in the case of paired-end sequencing). Picard Tools can also be implemented here to discard duplicate reads (mainly introduced by PCR amplification during library preparation or bridge amplification during clustering). These are reads that are exact copies with the same start and end point. In DNA sequencing, duplicates can skew the genotyping in favour of one allele or lower the overall genotype quality.

In the case of scRNA-seq, some of the initial pre-processing also differs somewhat, mainly because the second read contains the barcode corresponding to the cell identity. We applied the ddSeeker [116] and DropSeq [117] protocols for pre-processing the scRNA-seq data before alignment to GRCh38 using STAR [113].

Up to this point, processing RNA and DNA sequencing data is similar, but from this point, having achieved high quality aligned data, the processes start to diverge. DNA sequencing proceeds with variant calling to determine the genotypes in the data while RNA sequencing proceeds with gene expression quantification, both of which are further explained in the coming sections.

**Variant calling**

For the variant calling of DNA sequencing data, we have used the GATK [118, 119] and applied their best practices, which are frequently updated and refined. The main steps include base quality score recalibration (BQSR), variant calling using HaplotypeCaller [120], joint genotyping with GenotypeGVCFs, and variant quality score recalibration (VQSR). These processes are briefly described in general below.

The bases in all reads have quality scores with estimates of how likely they are to be correctly determined by the sequencing machines (calculated by the sequencing manufacturer’s proprietary algorithms). However, they may not reflect the true error rates and are, for example, affected by desynchronization of DNA copies in the same cluster, leading to biased fluorescence intensities [121]. So even if sequencing base qualities can appear to be high, a 99% certainty will still yield many errors when sequencing billions of bases. BQSR applies machine learning to model errors and adjusts the quality scores accordingly depending on several
parameters including quality scores on surrounding bases, position in the read, cycle, and the read group.

Variants are then called using HaplotypeCaller; its advantage lies in reference-based local reassembly of sequencing reads containing non-reference variants. This improves the capture of genetic variation (especially INDELs). HaplotypeCaller uses de Bruijn-like graph construction of reference sequence k-mers (10 and 25 nucleotides in length) to form candidate haplotypes that are realigned using Smith-Waterman alignment over the regions with non-reference variants. Genotyping is then performed using a Bayesian model to calculate genotype likelihoods from the underlying likelihoods for each read against each haplotype from paired Hidden Markov Models. Here, the probability of a candidate genotype is the product across all the sequencing reads mean likelihood for all alleles in the specified genotype. The genotype probabilities are converted to Phred-scaled (log-transformed and multiplied by negative ten) likelihoods. This is then normalized by subtracting the lowest value (the value of the most probable genotype). The genotype quality (GQ) is then defined as the difference between the two lowest values, of which one is zero. The GQ cannot be higher than 99 even though some genotypes might be more probable. However, this is slightly simplified as we have used the HaplotypeCaller in gVCF mode, where some additional conditions come into play when calculating the genotype likelihoods because there are overall fewer candidate haplotypes available for any given sample. Therefore, base quality scores and likelihood estimations of alleles not seen in the sample are used to model the genotype likelihoods. In gVCF mode, the final GQ is later calculated (as explained above) using the estimated likelihoods by GenotypeGVCFs which combines the sample gVCFs into the final VCF file. This quick step can be re-run at any time if more samples are added to the cohort later, thereby solving the previous N+1 problem with long computing times. The VCF file includes all genetic positions with variation in the cohort, and we have then evaluated all bi-allelic genetic positions where a genetic position in a patient can be reference homozygous (REF), heterozygous (HET), or variant homozygous (HOM) as shown in Figure 3.

VQSR is then applied to filter away low confidence genotypes. This uses Gaussian mixture models for classifying genetic variants based on their auxiliary information (including coverage, GQ, base quality scores, surrounding bases calls, and variant calls). The models are trained by using and clustering high confidence genetic variants from dbSNP and HapMap within the data sets. VCFtools [122] was then used to filter out genetic variants not labelled as PASS by VQSR, of low coverage, or high missingness in the cohort from the final VCF file.
As a side note, there is much development of newer, more accurate variant callers with improved machine learning and deep learning models. For example, DeepVariant (from Google Inc. and Verily Life Sciences) [123] utilizing convolutional neural networks (CNNs), NeoMutate (for somatic mutations in cancer) [124], and GATK who are working to improve their variant calls with deep learning, with the new tool CNNvariant also based on CNNs (as described on the Broad Institutes GATK blog [125, 126]) to replace VQSR.

**Gene expression quantification**

For the gene expression quantification of RNA sequencing data, the number of reads spanning gene regions defined in an annotated gene reference file are calculated, as visualized in Figure 4. This means that the higher coverage a gene has, the more reads will be counted for that gene.

For the RNA-seq of the cell lines in Paper IV, all uniquely mapping reads spanning a single gene region were summarized with featureCounts [127]. These counts can then be used as an average measurement of the gene expression in the sample. To make the data between samples comparable, gene expression is usually also normalized to both the length of the gene and to the size of the sequencing library for each sample. After this, another common analysis is differential gene expression to compare differences in the relative gene expression between samples representing different cells, states, or treatments. For this, we used edgeR [128, 129] and followed their recommendations.

The gene expression for the scRNA-seq in Paper V was determined using the Drop-seq [117] tool DigitalExpression. Thereafter the R toolkit Seurat [130-132] was used following their recommendations. One point to remember when analysing scRNA-seq data is that genes specifically expressed by certain cell types or housekeeping genes might not be
observable and expressed in all individual cells at a given time. Gene expression is irregular, with active periods alternated with inactive periods, this is known as transcriptional bursting [133, 134]. Therefore, rather than comparing single cells with one another, clusters of similar cells within a sample should be compared with cells in another cluster.

**SNV/INDEL analysis**

For association analysis of individual SNVs/INDELSs to the toxicity phenotypes, we used the toolset PLINK [135]. This toolset is designed to handle large-scale genetic datasets in a computationally convenient and efficient way. PLINK can manage and rapidly analyse genome-wide data including hundreds of thousands of genetic variants in large cohorts while also including both standard and novel tests [135-137]. For the association analysis in our studies, PLINK was used to apply linear regression of continuous phenotype data and Fisher’s exact test of case/control phenotype data on bi-allelic SNVs/INDELs. These tests could be carried out assuming different penetrance (that is, different genotype/phenotype relationship) including multiplicative, additive, recessive, and dominant genetic models.

In Paper II, we used the linear regression model implemented in the linear function assuming an additive genetic model for the association of genotypes to the continuous variables. This test assumes that the minor allele within the given data is the risk allele, and the $\beta$-value (the slope of the linear regression line) represents the direction and magnitude of the effect of the minor allele ($-\beta$ associated with higher toxicity, $+\beta$ associated with lower toxicity).

In Paper IV, we used Fisher’s exact test for categorical case/control variables, CTCAE 3-4 and CTCAE 0-1, implemented in an allelic fashion for which the estimated OR applies to the minor allele.

The tests above are excellent for finding associations between individual genetic variants. However, genetic variants of low frequency are harder to
find. In the rare phenotype/rare genotype case they will stand out, because mainly (only) the patients with the phenotype will have the genotype; however, in the case of the common phenotype/common genotype case, the effects of rare variants are easily missed [135, 138]. One can anticipate that rare genetic variants could have an influential effect, but it would not be discovered because other patients without the rare genotypes would to a high degree also exhibit the phenotype. One way to overcome this is by looking at the combined effect of many genetic variants, for example, within the same gene or haplotype. There are numerous association tests for this, such as burden and variance-component tests [139], but some of these tend to underestimate the effect of the common variants while overestimating the effect of rare variants. The optimal sequence kernel association test available in the R package SKAT [140], circumvents this by initially assigning all genetic variants equal weight. We applied the optimal SKAT to evaluate the combined effect of all genetic variants (both common and rare genetic variants) within the start of the first exon (-6 base pairs) and the end of the last exon (+6 base pairs) as defined by RefSeq [141] GRCh37.

**Functional annotation**

Next, is the question of the genetic association’s relevance. Although p-values, OR, and $\beta$-values all say something about the penetrance and phenotypic association of the genetic variants, it is still difficult to postulate anything about any actual functional effect of the genetic variants. The use of more specific information on the genetic variants such as missense, synonymous, frameshift, conservation, and within which gene it lies can enhance our understanding of their effect. There are many tools for annotation and extraction of this type of information; however, many of these methods are restricted to the coding region of the genome. For a more unbiased and comparable way to categorize and rank individual genetic variants, we applied combined annotation dependent depletion (CADD) scores [142]. CADD scores estimate the deleteriousness (which is correlated with the molecular function and pathogenicity) of genetic variants. CADD integrates effect, conservation, expression, regulation, and transcription factor annotations from multiple sources (including [143-152]) into a single score for each variant. The CADD score is then easily interpreted relative to all possible genetic substitutions in the human genome. Substitutions with the highest 10%, 1%, and 0.1% deleteriousness in the human genome have a score $>10$, $>20$, and $>30$, respectively.

Another way of understanding the effect and relevance of genetic variants is by mapping them to known biological systems such as pathways and GO terms [153]. This also allows some interpretation of whether the genetic
variants are active within a biological system that could have some functionality related to the toxicity phenotypes. Because we investigate multiple genetic variants or genes from large-scale datasets, we implement enrichment analysis to evaluate whether we have an overrepresentation of genetic variants within biological systems than that found by chance given the same number of genetic variants or genes, and the size of the pathway/GO. In Paper II, we used the meta-database ConsensusPathDB [154, 155], combining predefined pathways from several heterogeneous resources including KEGG [156], PharmGKB [157], and Reactome [158]. In Paper IV, we performed enrichment analysis of module genes in KEGG pathways [156] and GO terms [153] using the R package clusterProfiler [159].

Furthermore, in Paper IV, we applied the graph-theoretic clustering algorithm MCODE [160] on the String PPI network [161] for inferring GNMs starting from a set of genes based on the SNVs/INDELs in the data as seeds. However, there are many other methods/algorithms for inference of GNMs. GNMs is yet another way of extracting potential functional patterns from the data based on protein interactions. It uses the seed genes as input to extract interacting proteins from the interactome for constructing the GNM. This means that the elements within the GNM are extracted and interpolated from the seeds to other potential genes that could be relevant for the investigated phenotype but were not included as input for the algorithm. The validity of the GNM can then be proved (or disproved) by layering the genes in the GNM with information of enriched pathways/GO terms (as above) or even with gene expression data. We performed the latter by utilizing gene expression data from cell lines, human bone marrow, human kidney, and rat bone marrow all treated with the same drugs (gemcitabine/carboplatin) as our patients.

Prediction models

Constructing models that use the genetic variation associated with toxicity is an important tool for guiding future clinical decisions regarding ADRs. When the ADRs are common, complex, and influenced by multiple genetic factors, it is particularly necessary to use polygenic scoring. In our studies, we have implemented two prediction models, wGRS and logistic regression. These models combine the effects of multiple genetic factors into polygenic risk scores that can be used to stratify patients based on their risk of toxicity.

wGRS

The integration of genetic risk factors into clinically relevant wGRS for patient stratification has previously been implemented for many
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

phenotypes (including chemotherapy toxicity) [87-93]. In wGRS, the effects of many genetic variants are weighed together into one combined risk score for the overall investigated phenotype. We implemented wGRS using the formula in Equation 3:

$$wGRS = \sum_{i=1}^{n} w_i A_i$$

Equation 3

where $i$ is the genetic variant (SNV/INDEL), $w_i$ is the weight for genetic variant $i$, $n$ is the number of genetic variants in the model, and $A_i$ is the number of minor alleles (0, 1, or 2) for genetic variant $i$. The wGRS is then calculated by multiplying the number of minor alleles for each genetic variant by its weight and then summarizing across all genetic variants. This assumes that the genetic variants have an independent additive effect. In Paper II we used wGRS models for the main associated pathway found, and the $\beta$ values from the association analysis in PLINK as weights. Based on the wGRS quantiles, the patients were stratified into four risk groups, low, medium-low, medium-high, and high, for thrombocytopenia CTCAE grades 0-2 and 3-4, and the magnitude of the platelet decrease from the baseline value.

**Logistic regression**

Binomial logistic regression is another method that can be used to combine sets of genetic variants into a tool for stratifying patients according to their risk of toxicity. Logistic regression models could be implemented using the genetic variants’ $\beta$ values, as in the case of the wGRS, or the coefficients can be calculated within the logistic regression to optimize the separation between the two binomial states in the data.

In Paper II, this was done using the R package glm for constructing logistic regression models stratifying the risk of CTCAE grades 3-4 compared to 0-2, and the magnitude of decrease (large against small). In Paper IV, the logistic regression was implemented using the R package glmnet as this includes the random LASSO [94] for selection and shrinkage of model variables and cross-validation for stratification, which reduce the likelihood of overfitting the model and increasing the model’s robustness to classify previously unseen data. Here the risk of maximal toxicity was modelled as CTCAE grades 3-4 against CTCAE grades 0-1.

The logistic regression models’ performance was then evaluated using receiver operating characteristics (ROC) and AUC determined with the R-package ROCR [162]. An AUC = 100% is a perfect classification and anything less than 50% would be worse than guessing.
Dimensionality reduction and clustering
When it comes to analysing large-scale genetics data from GWAS, WES, WGS, RNA-seq, or scRNA-seq, much can be done and much can be missed. Multifactorial responses and effects are hidden within the data. Here, the use of dimensionality reduction has a role. This is a process for reducing data that contains thousands of parameters (high-dimensional data) into fewer dimensions that capture and explain most of the variability in the data. Dimensionality reduction can be used for multiple purposes, for example for finding patterns and interactions in data, but it can also be used for quality control and outlier discovery.

The main dimensionality reduction technique we have used is principal component analysis (PCA). PCA is a linear technique for dimensionality reduction and works by constructing a correlation matrix from which the eigenvectors are calculated. The eigenvectors with the largest eigenvalues are the main principal components (PCs). The first PCs can be plotted against each other for visual determination of clustering and data interpretation. Although this removes layers of data, it should retain the most important variability within the first couple of PCs.

For scRNA-seq, dimensionality reduction is a key technique implemented in the R toolkit Seurat [130-132] used by us. For single-cell data, dimensionality reduction aids in the detection of common patterns and similarity between cells. In Seurat [130-132], PCA is coupled with other dimensionality reduction analyses, such as canonical correlation analysis for identification of gene sets that are highly correlated between samples or sets of cells. These reduced dimensions can then be used to cluster cells, both within and between samples to find populations of cells with similar or differing characteristics. While there are multiple algorithms for clustering, both supervised and unsupervised, we mainly used the default unsupervised clustering in Seurat [130-132]. This uses the k-nearest neighbour algorithm to construct a shared nearest neighbour graph from which the clusters are determined.

The clusters can then be visualized by further reducing the PCs that were used into just two dimensions. For this, we have used the two algorithms: t-distributed stochastic neighbour embedding (t-SNE) [163] and uniform manifold approximation and projection (UMAP) [164], both available in Seurat [130-132]. The plots can be colour-coded according to the clusters the cells were assigned in the previous step. The plots should probably not be overinterpreted as they are severely reduced from higher dimensions. But they are very useful for identifying and visualizing different cell types or expression of various genes. UMAP is said to better preserve the global structure and continuity of subsets of cells [164].
RESULTS AND DISCUSSION

This section is subdivided based on the five papers and their use of different methodologies and analytic approaches: candidate genes, WES, WGS, and scRNA-seq. This flow through the papers also reflects the transition of pharmacogenomics from candidate assumptions towards the appliance of the more unbiased and agnostic WES and WGS methods and, lastly, the addition and integration of high-resolution functional analysis using scRNA-seq.

Paper I

In Paper I, the association between paclitaxel/carboplatin-induced myelosuppression, PFS, and OS, and the four ABCB1 SNPs, 1199G>A, 1236C>T, 2677G>T/A, and 3435C>T, and the CYP2C8 SNP 1196A>G were investigated. This study was performed in collaboration with Oasmia Pharmaceuticals AB (Uppsala, Sweden) and included 525 (hereafter referred to as All) ovarian cancer patients treated with paclitaxel (in two different formulations; Arm A, 250 mg/m² Paclical administered as a 1-hour intravenous infusion, n=260, and Arm B, 175 mg/m² Taxol administered as a 3-hour intravenous infusion, n=265) together with carboplatin (AUC=5-6) originally included in the OAS-07OVA phase III trial (EudraCT: 2008-002668-32, ClinicalTrials.gov: NCT00989131).
**ABCB1 and CYP2C8 genotyping**

The five SNPs were genotyped using pyrosequencing as previously described [35, 39] and there were no differences in genotype frequencies between the treatment arms A and B, see Table III. The genetic variant 2677G>A was excluded from further analysis due to low frequency. Based on the genetic structure of ABCB1 [165], the three SNPs 1236C>T, 2677G>T, and 3435C>T were used to construct the haplotype referred to as CCGGCC, CTGTCT, or TTTTTT.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Genotype</th>
<th>All</th>
<th>Frequency</th>
<th>n</th>
<th>Arm A</th>
<th>Frequency</th>
<th>n</th>
<th>Arm B</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1199G&gt;A</td>
<td>G/G</td>
<td>495</td>
<td>94.3%</td>
<td>242</td>
<td>93.1%</td>
<td>253</td>
<td>95.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/A</td>
<td>28</td>
<td>5.3%</td>
<td>17</td>
<td>6.5%</td>
<td>11</td>
<td>4.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A/A</td>
<td>2</td>
<td>0.4%</td>
<td>1</td>
<td>0.4%</td>
<td>1</td>
<td>0.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1236C&gt;T</td>
<td>G/G</td>
<td>168</td>
<td>32.0%</td>
<td>91</td>
<td>35.0%</td>
<td>77</td>
<td>29.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C/T</td>
<td>251</td>
<td>47.8%</td>
<td>117</td>
<td>45.0%</td>
<td>134</td>
<td>50.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T/T</td>
<td>106</td>
<td>20.2%</td>
<td>52</td>
<td>20.0%</td>
<td>54</td>
<td>20.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2677G&gt;T/A</td>
<td>G/G</td>
<td>150</td>
<td>28.6%</td>
<td>87</td>
<td>33.5%</td>
<td>63</td>
<td>23.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/T</td>
<td>238</td>
<td>45.3%</td>
<td>102</td>
<td>39.2%</td>
<td>136</td>
<td>51.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T/T</td>
<td>106</td>
<td>20.2%</td>
<td>52</td>
<td>20.0%</td>
<td>54</td>
<td>20.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3435C&gt;T</td>
<td>C/C</td>
<td>98</td>
<td>18.7%</td>
<td>56</td>
<td>21.5%</td>
<td>42</td>
<td>15.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C/T</td>
<td>273</td>
<td>52.0%</td>
<td>127</td>
<td>48.8%</td>
<td>146</td>
<td>55.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T/T</td>
<td>154</td>
<td>29.3%</td>
<td>77</td>
<td>29.6%</td>
<td>77</td>
<td>29.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haplotype</td>
<td>CCGGCC</td>
<td>73</td>
<td>13.9%</td>
<td>41</td>
<td>15.8%</td>
<td>32</td>
<td>12.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTGTCT</td>
<td>169</td>
<td>32.2%</td>
<td>74</td>
<td>28.5%</td>
<td>95</td>
<td>35.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TTTTTT</td>
<td>89</td>
<td>17.0%</td>
<td>47</td>
<td>18.1%</td>
<td>42</td>
<td>15.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1196A&gt;G</td>
<td>A/A</td>
<td>445</td>
<td>84.8%</td>
<td>227</td>
<td>87.3%</td>
<td>218</td>
<td>82.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A/G</td>
<td>77</td>
<td>14.7%</td>
<td>33</td>
<td>12.7%</td>
<td>44</td>
<td>16.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G/G</td>
<td>3</td>
<td>0.6%</td>
<td>0</td>
<td>0.0%</td>
<td>3</td>
<td>1.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table III.** Genotype frequencies in All, Arm A, and Arm B. This table was adapted from Supplementary Table 2 in Paper I.

**Progression-free survival and overall survival**

There were no differences in PFS or OS when comparing the two treatment arms. Continuing the analysis, we could show increased PFS and OS in All and Arm A for the variant alleles, mainly 3435TT and the haplotype TTTTTT. Cox proportional hazard models showed increased PFS, 10.8 months for 3435TT compared to 9.6 months for 3435CC in All (HR=0.623, p=0.002) and 11.6 months for 3435TT compared to 9.8 months for 3435CC in Arm A (HR=0.590, p=0.013). The same pattern was seen for OS and
3435TT, see Figure 5 for Kaplan-Meier curves. For the TTTTTT haplotype compared to the CCGGCC haplotype, increased OS was seen in All, 435 days compared to 387 days (HR=0.467, p=0.018), and Arm A, 444 days compared to 366 days (HR= 0.336, p=0.018). The same benefit of the variant alleles was not evident in treatment Arm B. Other studies have also shown increased OS and/or PFS for the 1236C>T, 2677G>T/A variant alleles [41]. However, in contrast to our finding, the 3435C>T wildtype has previously been found to be beneficial [42], while other studies have been inconclusive [38, 43].

![Figure 5. Kaplan-Meier curves for OS, in days, for the 3435C>T genotypes in All (A), Arm A (B), and Arm B (C). This figure was adapted from Supplementary Figure 6 in Paper I.](image)

**Toxicity**

At the start, the patients had normal baseline values of neutrophils, leukocytes, and platelets. However, during treatment, the proportion of patients with high myelosuppressive CTCAE grades increased, as illustrated in Figure 6. There was also significantly more myelosuppressive toxicity in Arm A, which is explained by the higher dose of paclitaxel used.
The main findings included 3435C>T and the haplotype. The 3435C>T variant alleles were associated with increased neutropenia in Arm A (p=0.018, first cycle neutrophil nadir) and less neutropenia in All (p=0.048, first cycle relative neutrophil decrease) and in Arm B (p=0.049, first cycle relative neutrophil decrease). The haplotype variant alleles were associated with less neutropenia in Arm B (p=0.039 and p=0.021 for first cycle neutrophil nadir and first cycle relative neutrophil decrease, respectively), and in All (p=0.048, first cycle relative neutrophil decrease), and with less leukopenia in Arm B (p=0.029, whole treatment leukocyte nadir).

To summarize, the variant alleles of the *ABCB1* SNPs and the haplotype were mainly associated with less myelosuppressive toxicity overall in All and in Arm B. However, these associations were of small effect and varied between neutropenia, leukopenia, and thrombocytopenia, and changed over the course of the treatment. In addition, these associations did not show the same effect in Arm A.

There are studies showing lower *ABCB1* mRNA expression and stability [166, 167] and increased neutropenia [44] for the 3435C>T variant allele, as opposed to our findings. On the other hand, higher *ABCB1* expression has also been reported for the 3435C>T variant allele [168]. The general idea for
the role of ABCB1 is that lower expression results in slower clearance of paclitaxel and higher intracellular concentrations leading to more cell death. This has two sides: killing both more cancer cells (positive) and healthy cells resulting in toxicity (negative), meaning that low ABCB1 expression in tumour cells and high expression in normal cells is preferable. Our findings of decreasing myelosuppression for ABCB1 variant alleles support the increased expression reported by Baldissera et al. [168]. However, one should keep in mind that neither the expression of ABCB1 in various cell types nor the effects of genetic variation in ABCB1 are fully known.

Neuropathy is another common and debilitating ADR related to the use of paclitaxel [25, 33, 34] that one must mention in relation to this type of study. Our intention was to investigate neuropathy in this cohort as well. However, when looking at the clinical data, the frequency of registered neuropathy was much lower than anticipated. Therefore, we decided to not investigate neuropathy as we would not be able to accurately represent the frequency of neuropathy, nor the effect of different treatment regimens on neuropathy, nor the association of the SNPs with risk or grade of neuropathy.

Inconclusive or informative?

Comparing our results with previously published studies it could be said that we only add more confusion to the inconclusive body of literature on this topic. However, we believe that the fact that our study had, in contrast to many others, a large and clearly defined cohort using two slightly different regimens can explain why the literature consists of contradictory findings - because we can show that the effects of these genetic variants are small and that they change with the different treatment regimens used in Arm A and Arm B. Our results indicate that the variant alleles have small protective effects for toxicity. But when using higher doses, this effect is negligible in terms of protection against toxicity. However, the variant allele carriers tended to have increased PFS and OS following the increased dose in Arm A. This shows that variant allele carriers might be slightly undertreated in Arm B with the lower dose.

Another lesson learned from this study is that genetic variation in the major genes important for PK/PD cannot explain everything, even if there are specific cases when they do. But the more common and complex the phenotypes, the more likely they are to be affected by many contributing factors, including the combinatorial effect of multiple genetic variants, some, of course, also in the PK/PD genes. This idea is also, reassuringly, reflected in the transition towards the use of GWAS, WES, and WGS in the field of pharmacogenetics. Approaches that solve some problems while
Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions

adding much complexity to evaluation, interpretation, validation, and implementation. Therefore, in the following studies, we expand the analysis to more agnostic genome-wide sequencing approaches and do not focus solely on candidate genes.

Paper II

In Paper II, WES was used to investigate gemcitabine/carboplatin-induced myelosuppression. The study included WES data from 212 (after sequencing data quality control) NSCLC patients treated with gemcitabine (1250 mg/m² at day 1 and 8) and carboplatin (AUC=5 at day 1) for at least one cycle. The number of platelets was registered before treatment start and on days 8, 15, and 21 of the first treatment cycle. Already during the first cycle, roughly a third of the patients experienced severe to life-threatening thrombocytopenia, see Figure 7. This figure shows that roughly 35% of the patients experienced thrombocytopenia CTCAE grades 3-4, as early as during the first cycle. Comparing this with paclitaxel/carboplatin in Paper I (Figure 5A) were the thrombocytopenia CTCAE grades for most patients were 0-1 during the first cycle, reveals that gemcitabine/carboplatin effects on platelets are both faster and more severe.

SNVs/INDELs, FDR, and gene-based tests

When looking at the statistical associations between genotypes and the toxicity phenotype thrombocytopenia, there are some things that are important to remember. The WES yielded 71374 common (MAF≥0.01) SNVs/INDELs which would, just by chance, give many significant associations. This is a known problem, recurrent among similar studies [169], and is reflected in the literature with associations that to date have proven hard to validate. Moreover, we were reluctant to focus only on our top associations as we thought this would not be the most efficient way to utilize our dataset. To find a suitable p-value cut-off, we instead implemented permutations (Figure 8). This showed that lowering the p-value cut-off initially gives a better FDR, but the lowest p-value cut-off is not necessarily the p-value cut-off with the lowest FDR. So not only would focusing on the top associations give an increased proportion of false positives, true positives would be thrown away in the process. This could
explain why previous top associations have been hard to replicate. Applying
this method, we found a set of 130 SNVs associated with thrombocytopenia.

We also showed that some of the SNVs/INDELs are predicted to have
deleterious effects using CADD scores [142]; however, these are not
necessarily correlated with their p-values. Further, by checking LD with
genetic variants found in a GWAS dataset [170] we could validate 23 of our
SNVs/INDELs. However, it should be noted that the validated genetic
variants are probably mostly related to thrombocytopenia induced by
carboplatin, as the validation dataset was treated with
paclitaxel/carboplatin and not gemcitabine/carboplatin. This type of
validation does not prove the effect of the SNVs/INDELs, but it does
indicate the genetic variants’ importance, credibility, and prioritization for
validation in future studies.

We also investigated the combined effect of multiple genetic variants, both
rare and common, within whole gene-based regions. This yielded 25 genes
associated with thrombocytopenia, some of which also included significant
SNVs/INDELs from the previous analysis.

**Hematopoiesis-related genes and pathways**

By using the genes represented by the gene-based test, genetic variants
with high CADD scores, and validated SNVs as input for pathway
enrichment, the two top-enriched pathways were *Hemostasis* (p=1.27×10⁻³) and *Factors involved in megakaryocyte development and platelet
production \((p=3.34\times10^{-4})\) showcasing how this approach has extracted relevant genetic variation from biological systems understandably of importance for the investigated phenotype thrombocytopenia. Furthermore, some of the genes within these two pathways have previously been shown to be important: \textit{SERPINC1} deficiency causes thrombophilia [171, 172], \textit{DOCK8} correlates with platelet volume [173], and \textit{JMJD1C} correlates with platelet volume [173-176] and affects megakaryocyte proliferation in mice [177]. We have also shown that \textit{JMJD1C} knockdown affects the proliferation and gemcitabine/carboplatin-sensitivity of two myelogenous cell lines \textit{in vitro} (submitted manuscript Schimek \textit{et al.}). Taking this together it seems that genetic variation in these genes can be a plausible cause affecting proliferation rates of megakaryocytes, the production of platelets, and blood cell drug-sensitivity. This is also well in line with the drug’s mode of action affecting the DNA in highly proliferative cells. Thus, an underlying genetic variation in these pathways could play an important role in the varying degree of the drug-induced toxicity thrombocytopenia experienced by patients.

**Thrombocytopenia prediction models**

Up to this point, we have only shown that we find associations in relevant biological systems that we can partly validate with previous findings. But we have yet to show if these genetic variants can be used to stratify toxicity risk. To do this we used the 17 SNVs/INDELs covered by the genes \textit{KIF6}, \textit{CAPZA2}, \textit{DOCK8}, and \textit{JMJD1C} found within the main pathway, \textit{Factors involved in megakaryocyte development and platelet production}. We implemented two prediction models, based on wGRSs and logistic regression, using the 17 SNVs. Figure 9 shows how the wGRSs could stratify thrombocytopenia risk levels for CTCAE (A) and decrease (B). This figure was adapted from Figure 1 in Paper II.
our patients into the four risk levels, Low, Medium-low, Medium-high, and High, for CTCAE in Figure 9A and decrease in Figure 9B. This shows accurate predictions, at least in the Low and High levels.

Patients with high toxicity risk were more likely than patients with low toxicity risk to exhibit CTCAE grades 3-4 (OR=22.35, p=1.55×10⁻⁸) or a large decrease down to ≤ 10% of the baseline platelet value (OR=66.82, p=5.92×10⁻⁹). The risk levels, Medium-low and Medium-high, include a number of low and high toxicity patients making them harder to draw any definite conclusions from.

We also implemented logistic regression models for CTCAE and decrease which are shown together with their ROC curves in Figure 10. These models show great predictive power with ROC AUC = 79% and 86%, respectively, for CTCAE and decrease. Keep in mind that the logistic regression models are binomial and because the decrease has three levels (small, medium, and large), only the patients with small or large decrease were used in the decrease model, the patients with medium decrease were omitted.

**Figure 10.** Logistic regression model for stratification of patients CTCAE (A), CTCAE ROC (B), decrease (C), and decrease ROC (D). This figure was adapted from Figure 2 in Paper II.
Thus, going further and using WES for understanding and stratifying toxicity based on the genetic variation appears to add predictive value. Interestingly, the main pathway and the SNVs/INDELs used for prediction do not stem from genes important for metabolism or transport of the drugs, rather they stem from genes important for the proliferation and activity of the cell type that produces the platelets. This indicates that the most important factor for thrombocytopenia in patients is the underlying megakaryocyte activity and platelet production. This could increase or decrease the general production of platelets or perhaps even produce platelets with varying quality and function. Speculating further, perhaps these genetic differences do not affect platelets in a normal situation, but only when the system has been perturbed, for example by cytotoxic agents, is the effect of the genetic variants on platelet levels observed.

These results triggered our interest in whether genetic variation within promoters, enhancers, or other distal parts of the genome contribute to the variation in induced toxicity, perhaps by affecting the expression and regulation of genes within key pathways. Therefore, we transitioned from WES to WGS, which has become readily available in recent years. First, we compared the sequencing output of WES and WGS across the exome to facilitate a better understanding of sequencing quality, variant filtering, and method concordance in Paper III. Then, we continued by investigating toxicity using the WGS approach in Paper IV.

Paper III
For the study in Paper III, we leveraged the fact that we have 96 DNA samples from NSCLC patients that have been both WES, for Paper II, and WGS, for Paper IV. These matched samples were used to compare variant calls and quality metrics over the exome target region covered by both methods, as defined by the Nextera Rapid Capture Exome Targeted Regions Manifest version 1.2. While there are previous studies comparing WES and WGS [178-182], the number of matched biological samples used here has not previously been seen in a technical comparison of the two sequencing methods.

High and low coverage WES
When interpreting the data, we immediately found that the mean WES coverage was uneven, see Figure 1 in Paper III. Based on this we decided that comparing all the data straight off would not be the most informative use of the data. Instead, we split the data in quantiles based on the WES coverage and compared the two extreme quantile groups SC_low (the 24 samples with the lowest mean WES coverage) and SC_high (the 24 samples
with the highest mean WES coverage) independently. This enabled us to compare both low coverage and high coverage WES data with WGS data. Bi-allelic genetic variants were then categorized into specific types and groups. Variant types: reference homozygous (REF); heterozygous (HET); or variant homozygous (HOM). Variant groups: called in both, variants called with the same zygosity; discordant, variants called with differing zygosity; exome only, for variants called using WES only; or genome only, for variants called using WGS only.

**Called variants, coverage, and genotype quality**
Looking at the data briefly presented in Table IV, the number of variants called from the two sequencing methods is higher for WGS in both SC_low and SC_high. This reflects the power of WGS to sequence rare genetic variants. While there are some genetic variants that are only captured by the respective methods, the number of genetic variants only called by WGS decreases substantially with increasing WES coverage. WGS coverage is homogenous over both differing variant types and variant groups. On the other hand, WES coverage is also stable (at least in the SC_high setting) over the variant groups but varies over the different variant types, where HET variant calls generally have a higher coverage. Looking at quality, exome only, genome only, and discordant calls have lower quality than the variants called in both. For the variant types, the quality is mainly higher for WGS in SC_low and WES in SC_high. Further, HET calls tend to have higher quality than the other types.

| Table IV. Overview of sequencing parameters displaying average ±SD comparing WES and WGS in the subgroups SC_low and SC_high. This table was adapted from Table 2 in Paper III. |
|---------------------------------|-----------|-----------|
| **Average number of variants**  | WES       | WGS       |
| SC_low                          | 91811 ± 1182 | 99271 ± 1802 |
| SC_high                         | 97401 ± 432  | 98340 ± 2435 |
| **Average coverage of variants**| WES       | WGS       |
| SC_low                          | 37.25 ± 5.66 | 29.60 ± 3.02 |
| SC_high                         | 74.56 ± 11.91 | 30.25 ± 3.58 |
| **Average genotyping quality**  | WES       | WGS       |
| SC_low                          | 75.83 ± 5.19 | 80.84 ± 5.27 |
| SC_high                         | 92.54 ± 1.87 | 81.43 ± 5.42 |
| **Number of discordant variants**| WGS/WES  |
| SC_low                          | 2108 ± 523  |
| SC_high                         | 1007 ± 429  |

**Discordant variant calls**
Looking more closely at the discordant variant calls, which represent roughly 2% of the calls in SC_low and 1% in SC_high. The most common
discordance is HET in WGS and REF in WES; this discordance has, on average, much higher quality in WGS for both SC_high and SC_low, see Figure 11. Furthermore, it appears that for discordant calls (as in the other variant groups) the HET calls are generally of higher quality and more likely correct irrespective of the sequencing approach. Judging from the quality of the discordances, putting a quality threshold of 40 would remove the variant calls, which arguably are wrong, based on their lower average quality. However, this is a representation of the average quality of the discordant calls, and filtering should be done on individual genotype calls for each patient and not by discarding the whole genotype call from the cohort. This would increase the average quality of variant calls. Filters on genotyping rate can then be applied to remove genetic positions not genotyped across the cohort. The case when WGS is HOM and WES is HET, for which both have intermediate quality, would still in many cases surpass this cut-off. Luckily this uncertain case is not very common in either of the SC_low or SC_high settings.

Two factors contributing to the lower quality, uneven GC coverage, and fewer genetic variants found using WES are the capturing probes and the usage of PCR during the library preparation [183-186].

To conclude, the results from Paper III mainly, reassuringly, show that WES and WGS are concordant on most of the genotype calls across the exome. Moreover, higher WES coverage reduced the discrepancies seen. With that said, we still believe that as sequencing costs continue to decline, WGS will become the method of choice also for exome studies. Based on this, we could continue with the WGS study in Paper IV knowing that most of the variation earlier found in the WES will be captured using this method together with the remaining variation outside of the exome that was not previously known for these patients.
Paper IV

In Paper IV we extended the use of genetics data from the protein-coding exome to WGS 96 of the patients who previously underwent WES in Paper II. The 96 patients were selected based on extreme toxicity, either high or low neutropenia, leukopenia, and thrombocytopenia. The patients’ CTCAE grades were then used to identify correlations with the WGS data and, lastly, the study presents a prediction model for stratifying patients’ risk of high maximal myelosuppressive toxicity (that is CTCAE grades 3-4 of either neutropenia, or leukopenia, or thrombocytopenia).

SNVs as a starting ground for GNMs

Inherent to the WGS we found thousands of genetic variants surpassing normal p-value thresholds. However, as shown in Paper II, only using a select set of variants surpassing a very stringent exome, or genome-wide adjusted significance level, might yield a larger proportion of false positives and high FDR. At the same time, this would also force us to discard large amounts of our sequencing data already in the starting blocks of the study; this would not be an ideal way to unlock the full human genome’s potential for prediction of toxicity. Instead, we chose the arbitrary significance level p<0.001 for the initial screening of SNVs association with neutropenia, leukopenia, and thrombocytopenia using Fisher’s exact test and comparing CTCAE grades 0-1 with CTCAE grades 3-4. This gave a starting point of 4594, 5019 and 5066 autosomal genetic variants nominally associated with neutropenia, leukopenia, and thrombocytopenia, respectively. PCA showed initial signs of these genetic variants’ predictive capabilities because the first two principal components start to separate the low (CTCAE 0-1) and high (CTCAE 3-4) toxicity patients with the intermediates lying in between (CTCAE 2) (Figure 12).

![Figure 12. PCA using all nominally associated genetic variants for neutropenia (A), leukopenia (B), and thrombocytopenia (C). This figure was adapted from Figure 1 in Paper IV.](image-url)
After annotating the nominal associations, we found that most (>99%) reside outside the exome, which is reasonable because the human exome comprises roughly 1-2% of the full genome. It is also worth pointing out that compared to the background there was enrichment of distal intergenic SNVs/INDELs. This means that there is extensive genetic variation far from the regions coding for proteins present within this cohort and which can be used in prediction models.

**GNMs enriched with affected functional elements**

We then applied the graph-theoretic clustering algorithm MCODE [160] together with the String PPI network [161] for inferring gene network modules for neutropenia, leukopenia, and thrombocytopenia by using the genes closest to the respective genetic variants as seeds. The module validity was confirmed by the enrichment of differentially expressed genes in rat bone marrow treated with gemcitabine/carboplatin. The combination of genes in at least two of the modules for neutropenia, leukopenia, or thrombocytopenia were combined to construct a combined toxicity module consisting of 215 genes, including the most integral genes more generally important for myelosuppression. The toxicity module showed even more enrichment. Furthermore, significant overlap was also seen with gene expression meta-analysis data from gemcitabine and carboplatin-treated human bone marrow and kidney [187]. We could also show, using RNA sequencing of two human myelogenous cell lines in our lab, that 152 of the module genes are expressed in blood cells and that 18 of them are differentially expressed after 24-hour exposure to gemcitabine and carboplatin.

The genes in the toxicity module do not overlap with the four genes used for the predictions of thrombocytopenia in Paper II (JMJD1c, DOCK8, CAPZA2, and KIF6). However, looking at the enrichment of the genes, we found blood cell and cancer-related pathways and GO terms, important for cell proliferation and maturation, which is in line with the findings in Paper II and the known effects of the treatment. This shows how GNM genes inferred from WGS data can be used, and that functional genetic elements that are regulated as a response to the treatment can be found. The GNM genes are likely also important for the underlying myelosuppressive ADRs.

**Prediction model for maximal toxicity**

In summary, we showed that the module genes are enriched with genes expressed in relevant cells, and that the genes are differentially expressed after treatment with the chemotherapeutic drugs gemcitabine and carboplatin. Now we wanted to use the genetic variation underlying the gene module inference for stratifying the patients’ risk of maximal
myelosuppressive toxicity. We focussed on one model for predicting maximal myelosuppressive toxicity because this would be easiest to validate and implement in the clinic. As we have a limited number of patients and a greater number of genetic variants, the problem of overfitting is overhanging. To circumvent this, we used 20% of the patients for testing, and 80% for training, the prediction model. Then we implemented the use of the random LASSO [94] for the selection of a reduced number of genetic variants to include in the prediction model. To improve the model’s robustness and lower the chance of overfitting we used 10-fold cross-validation, permutations, and a randomized penalization factor for modelling the training data. This resulted in a prediction model combining the effects of 62 genetic variants (SNVs/INDELSs). As shown in Figure 13, when applying the prediction model, both the training set and the previously unused test set were accurately predicted with a ROC AUC = 100%. The intermediates, however, would be predicted to be of both high and low toxicity characteristics; but because the model is binomial the intermediates were not included when calculating the ROC, which should be kept in mind when evaluating the overall model predictions.

The final prediction model includes 42 (67.7%) intron variants, 16 (25.8%) distal intergenic variants, and 4 (6.5%) promoter variants. This highlights the usability and predictive capability of non-coding genetic variation for common toxicity phenotypes, and adds to the increasing body of literature that has long since refuted the early notion that non-coding DNA, commonly referred to as “junk DNA” after Ehret et al. [188] coined the term, is unusable and irrelevant. Another important feature of the
presented prediction model is that it does not utilize the patient characteristics, which would likely improve the model and its robustness.

Another way to exploit the whole genome data would be to infer GNMs from expression analysis of chemotherapy-treated blood cells, and then use the genetic variation mapping to the GNM genes for predicting myelosuppressive toxicity. Therefore, in Paper V we initiated the use of culturing and treating HSCs with chemotherapy and subsequently using scRNA-seq to elucidate gene expression changes induced by the treatment. We hope that this can be used in conjunction with WGS data in the future.

**Paper V**

In Paper V we have, to the best of our knowledge, conducted the first study using scRNA-seq of chemotherapy-treated HSCs. For the whole project, we included 20 patients with CML, all of whom have leftover harvested HSCs. These HSCs were not used in their treatment and would have been discarded if we had not had access to them. After ethical approval and informed consent, we were able and allowed to use these cells in the study. In Paper V, because scRNA-seq is a rather new method without standardized protocols for analysing and interpreting the data, and the HSCs are a rare and invaluable source of cells, we have only utilized cells from one sample as a pilot study to optimize our protocols and to demonstrate the feasibility of using the method for understanding toxicity.

ScRNA-seq has in recent years shown promise for better understanding transcriptional differences, even among cells thought to be homogenous, which with conventional RNA-seq would be masked by the average gene expression (Figure 14) [189-192]. In 2013, scRNA-seq was recognized as the method of the year [193], anticipated to transform many areas of biology and medicine in the years to come. ScRNA-seq will need specific cells depending on the toxicity investigated. Hopefully, the use of HSCs for myelosuppression in Paper V can serve as a good rule of thumb even for

![Figure 14. An example of a set of 13 cells with three types of distinctly different cell populations. Bulk RNA-seq would only give an estimate of the average gene expression of all cells in the set, as one population, while scRNA-seq would retain not only information of the different cell populations but also for all unique cells in the sample.](image)
studies using other cell types, toxicities, and drugs.

**Cultivation, treatment, and cell cycle analysis**

Primary cell cultures are harder to work with and more sensitive than cancerous immortalized cell lines. Therefore, we first tested media, supplements, culture vessels, culture times, and culture density. From this, we found that tissue-culture-treated (TC-treated) culture vessels worked best, even though the cells do not adhere to the surface the HSCs prefered TC-treated over non-treated surfaces (conventionally used for suspension cells). There are many options for culturing media and supplements, and we found the StemMACS HSC Expansion Media XF supplemented with StemMACS HSC Expansion Cocktail, containing human stem cell factor, Flt3-ligand, and thrombopoietin (both from Miltenyi Biotec) worked very well. We also found during initial tests that the use of low density (5-20×10^4 cells/ml) primed substantial cell death in the first 24 hours of culturing after thawing the cells. Increasing the density to 10^6 cells/ml circumvented this and allowing 48 hours of acclimatisation gave stable and reproducible results.

Using the optimized culturing conditions we could show the cells immediate sensitivity to both gemcitabine and carboplatin using 24-hour exposure times and the MTT assay as visualized in Figure 15A. Based on the MTT data, three treatment conditions were decided to be used for the scRNA-seq of the HSCs: Carboplatin High (150 µg/ml), Carboplatin Low (18.75 µg/ml), and Gemcitabine (25 ng/ml). Single cells from the treated samples and one untreated control sample were extracted before RNA sequencing.

![Figure 15](image_url)

**Figure 15.** Effects of gemcitabine and carboplatin on HSCs. (A) 24-hour dose-response curves of treated HSCs determined using the MTT assay. (B) The proportion of HSCs in G1 (Gap 1), S (Synthesis), and G2M (Gap 2 and Mitosis) phases of the cell cycle. This figure was adapted from Figures 1 and 3 in Paper V.
After pre-processing the sequencing data and quality-filtering cells according to the recommendations in ddSeeker [116], Drop-Seq [117], and Seurat [130-132], 1172 high-quality cells remained. The expression of the cell cycle markers suggested by Tirosh et al. [194, 195] shows that the treatments decrease the proportion of proliferating cells (G2M) substantially compared to the control sample (see Figure 15B). The higher carboplatin treatment concentration also had a more pronounced effect than the lower concentration. This confirms that the chemotherapeutic agents can be used to kill HSC *in vitro* and that they affect cell proliferation through their modes of action on the DNA synthesis.

**Transcriptionally different cell clusters**

We then went forward with dimensionality reduction using PCA followed by clustering and visualization of clusters in the separate samples. This showed (see Figure 4 in Paper V) that the treated samples have more clusters than the control sample. The control sample has two clusters and their differences, using KEGG pathways and GO terms, lay within mRNA, endoplasmic reticulum, and translation. We think that the two clusters could represent the healthy cells and cancerous cells in the harvested HSCs. However, this should be investigated further using known cancer markers. In addition, there were half as many cells in the control sample compared to the treated samples, which could pose a barrier to finding more rare cell populations among the control cells.

All treated samples and the control sample were then integrated to find if and how they cluster together. Most of the control cells end up in the first two clusters (0 and 1) while only a few are found in the remaining clusters (see Figure 5 and Table 3 in Paper V). However, this indicates that with more control cells, more clusters could possibly be found in the control sample. From this, we recommend aiming to get $\geq 300$ high-quality cells for each sample, as this seems better able to capture the cell heterogeneity. But it could also be that the treatment induces the spread into more clusters compared to the control sample, and that even with more control cells, only a small proportion of control cells would be found in these other clusters.

We can then use the integrated analysis of Carboplatin High and the Control as an example, visualized in Figure 16. We can then compare differential gene expression between the cells from the two samples that cluster together. Doing this for cluster 0 shows that the main differences between the treated and control cells lay in KEGG pathways related to cancer and proliferation. This could be due to clustering of the cancer cells, with the treated ones being pushed towards apoptosis and lower proliferation as a result of the treatment, which they try to circumvent by regulating p53 signalling and longevity, whereas in cluster 1 the differences...
between treated and control cells were attributable to leukocyte regulation. This could be the normal cells responding to the treatment with immune and proliferative regulation as an acute effect. Another interesting finding was that when we compared gemcitabine treated cells to the control cells, we also saw an immune cell response coupled with tumour necrosis factor and apoptosis signalling. The identification of several clusters also points towards that there is not just one major effect induced by the treatments but rather multiple underlying effects.

For the milder carboplatin treatment option, there were fewer and less distinguishable differences. This indicates that if we are to find acute effects the treatment should be close to the IC$_{50}$ or, in the case of a lower dose, a longer exposure time should perhaps be used. However, this would necessitate more optimization and could be more difficult because long-term cultivation (weeks) of HSCs is difficult and will likely involve differentiation and maturation of the cells.

Because we only used a single donor for this study, bald conclusions should be avoided. However, it is worth pointing out that the preliminary results in Paper V are coherent with the results in Papers II and IV. Showing that the most important genetic factors for understanding and predicting myelosuppression seem to lay within regulation, proliferation, and maintenance of the blood cells in question and not among genes involved in absorption, distribution, metabolism, and excretion of the drugs. Although we believe the latter genes also to be of importance, and possibly needed together with our findings, for implementing polygenic risk scores that eventually can be used to predict patients risk of toxicity in the clinic.

**Figure 16.** Integrated Carboplatin High and Control cells. (A) Coloured according to the sample. (B) Coloured according to the identified clusters. This figure was adapted from Figures 5a and 5b in Paper V.
Going forward with scRNA-seq

This initial study using HSCs from one patient shows how scRNA-seq can be used to determine and distinguish transcriptional effects induced by chemotherapeutic agents. However, more things can be done such as determining specific cell types using known hematopoietic and cancer markers [194, 196-198]. Another interesting step forward would be trajectory analysis over pseudotime to help understand the dynamic behaviour and the order of cells’ transcriptional responses induced by the treatments [197, 199, 200]. We will also move forward using both more cells and by expanding the studies to use cells from more of the included patients.

Lastly, Paper V enables us to conclude that using scRNA-seq and HSCs is a feasible approach which can be used for elucidating transcriptional responses induced by chemotherapeutic agents.
In conclusion, the results show how various sequencing techniques and bioinformatics strategies can be used to find pharmacogenetic biomarkers for chemotherapy-induced ADRs. We hope that the presented approaches can be applied and used by other researchers. In addition to their application with the drugs, diseases, or toxicities specifically investigated in this thesis, they should moreover have broad applicability for finding genetic biomarkers for ADRs in many other settings. We have also shown that, as a new way of investigating and understanding myelosuppression and ADRs, HSCs can be treated with chemotherapy in vitro and that scRNA-seq can subsequently be used to distinguish transcriptional effects attributable to the treatments.

The results culminate in polygenic prediction models for stratifying risk of toxicity, interpreted as thrombocytopenia (Paper II) or maximal myelosuppression (Paper IV). These models have been refined down from genome-wide studies to a limited set of genetic markers (17 and 62) that are small enough to be easily implemented clinically. However, the models need further validation and replication to ensure their predictive capability before they can be implemented in the clinic.

Hopefully, this thesis can be used in future projects studying how to improve risk stratification of patient toxicity and ADRs, and for guiding treatment decisions. This would result in personalized medicine with improved treatments, fewer ADRs, and increased patient wellbeing.
The main take-home messages are presented briefly below.

**Paper I**
- The effects of *ABCB1* and *CYP2C8* on paclitaxel/carboplatin-induced toxicity are small and vary with the treatment dose.
- The combination of many contributing genetic factors is needed for risk stratification of paclitaxel/carboplatin-induced ADRs.

**Paper II**
- Top genetic associations should not be the only focus, as these can be inflated with false positives.
- Genetic variation in hematopoiesis-related pathways is important for and can predict treatment-induced thrombocytopenia.

**Paper III**
- WES and WGS are mainly in concordance across the exome.
- We recommend an average genotype quality (GQ) threshold of ≥40.
- WGS has higher genotype quality on most discordant variant calls and, overall, captures more genetic variants than WES.

**Paper IV**
- GNM$s$ inferred from SNVs pick up functional genetic elements that are regulated as a response to the treatment.
- Distal intergenic and intron variation shows predictive power.
- Using the WGS data and the inferred GNM we propose a prediction model that can stratify patients’ risk of myelosuppressive toxicity.

**Paper V**
- Chemotherapy-treated HSC and scRNA-seq can be used as a functional approach for investigation of treatment effects.
- Several clusters of HSCs show that there is not one coherent effect, rather there are multiple underlying differential effects.
- We recommend ≥300 high-quality cells per sample and treatments close to IC$_{50}$ to ensure that effects can be captured with scRNA-seq.
Future Aspects

FUTURE ASPECTS

Going forward there is still much to be done within this field. The first obvious direction would be to expand the WGS in Paper IV and sequence the remaining samples to further improve the model. Another, direction would be to sequence the genetic variants included in the proposed models in the 525 patients in Paper I. Even if they have been treated with a different drug combination, we believe that the mechanisms embedded in the models (in Paper II and IV) are somewhat generic and relate to thrombocytopenia or myelosuppression in general. Therefore, they could be useful for risk classification in settings that do not necessarily include both gemcitabine and carboplatin. However, the extent to which (the number of patients) and the way (WES, WGS, or candidate pharmacogenetic biomarkers) this would be done is unfortunately also a monetary question. Expanding our studies would help to further refine the models before new studies are conducted by us or others to confirm the models’ clinical validity, and to prove whether their use actually results in, at least, less toxicity and hopefully also a better prognosis.

I also believe that our studies enable us to say that the more agnostic genome-wide sequencing approaches that we have employed for finding genetic components underlying toxicity are, as discussed by Kerr et al. [76], the way forward for cancer and personalized medicine. However, the early focus on genes, mainly involved in drug absorption, distribution, metabolism, and excretion, should not be forgotten as they will continue to
have a role, and are probably needed in many areas of personalized medicine.

There have also been opinions such as that of Letai [201] on the issue of whether precision medicine is equivalent to genomic medicine, mainly in regard to whether sequencing of patient and tumour samples can answer all our questions. I believe, as he does, that there is both room and a need for more functional studies for interpreting and understanding drug effects. Similarly for toxicity where, for example, expression profiling of genes in tissues relevant for toxicity could be a step forward in understanding drug effects. For this purpose, Paper V is well-timed as it demonstrates that scRNA-seq is a feasible model for interpreting chemotherapies’ effects on HSCs. Hopefully, both our group and others can continue working on functional models for toxicity.

Another characteristic of constitutional genetic information is that once determined it is the same (unlike the genotypes in cancer). There has been talk and ideas that at some point in the future everybody’s genomes will be sequenced, and stored accessible to healthcare providers. This is coming closer to being realised not only because many large hospitals get access to and implement NGS technologies in their clinical routine and the fact that the price of NGS is still decreasing, but also because companies are making sequencing available to the public [80, 202]. If (probably when) this becomes reality it will mean that, at the time that a drug is to be administered, potentially valuable genetic information that could be used for making a decision on a suitable dose could already be available. The era of personalized medicine has only just started. Together with the ever-growing pool of available genetic information, the application of machine learning [203, 204] will be expanded to further help us understand and use the data in terms of quality control, analysis, classification, and evaluation.

Exciting times for pharmacogenetics and personalized medicine lay ahead...
ACKNOWLEDGEMENTS

There are so many I would like to thank for small and big gestures, that in many ways have contributed and helped me. With your support, I have been able to get this far. Family, friends, and colleagues thank you! I would like to express special gratitude to the following.

To my main supervisor, Henrik Gréen thank you for believing in me and trusting your projects to me. Your expertise and optimism through my research studies have been of utmost importance. It has been a pleasure to work with you and I have grown and learned so much during this time, not only related to research but also personally.

My co-supervisors Svante Vikingsson, for indispensable guidance and constructive comments for my PhD-studies and projects, and Joakim Lundeberg, for all sequencing support and guidance. Thank you!

A big thank you also to all co-authors - Ignace Vergote, Benjamín Sigurgeirsson, Sailendra Pradhananga, Eva Brandén, Hirsh Koyi, Rolf Lewensohn, Luigi De Petris, Pelin Sahlén, Zelmina Lubovac-Pilav, Maria Apellániz-Ruiz, Cristina Rodríguez-Antona, Tejaswi Venkata Satya Badam, Rapolas Spalinskas, Mika Gustafsson, and Kourosh Lotfi – and to all patients and healthcare professionals.

Thanks, Margareta Reis for leading me into this path.
To my office and research colleagues Anna Svedberg, Ingrid Jakobsen, Anna Åstrand, and Lucia Pellè. Thanks for always helping me with my projects, supporting me during hard work, reading my documents, discussions, and an awesome working environment. I also appreciate your tolerance for my music and my restless legs, running in and out of the office all the time.

Thanks, Arjan Mofers and Karthik Selvaraju for being weird (in a good way) and helpful around the labs and offices, it has enlightened my days at work. Of course, also thanks to all other PhD-students, Postdocs, and master students at the department for fika and discussions over my mostly later than average lunchtimes.

To all other past and present colleagues at LÄFO, Klinfarm, and KVM thank you for support, collaborations, and discussions.

To all at some point involved or enrolled in Forum Scientium thanks for conferences, study visits, seminars, discussions, and many new friends!


Jag vill också passa på att tacka min löparklubb IK Akele där jag sedan hösten 2009, när jag flyttade till Linköping för att studera tränat hårt (med några uppehåll för skador) och haft roligt. Där jag också på senare år blev med i Tränargruppen, tack för förtroendet. Tack också till Stefan och Erik för bra sparring och träningspass både i och utanför klubbens regi.


Slutligen, så vill jag tacka dig Marina. Sense tu, això mai no hauria estat possible! För stöttning, support och allt roligt vi gör och har gjort. Men framförallt för all din energi till smått och stort, även om vi kanske inte hinner med allt, jämt.
REFERENCES


Pharmacogenetic biomarkers for chemotherapy-induced adverse drug reactions


192. Liu S, Trapnell C: Single-cell transcriptome sequencing: recent advances and remaining challenges. **F1000Res** 2016, **5**.


199. Campbell KR, Yau C: Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. **Nature Communications** 2018, **9**.


APPENDIX

Paper I-V
Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-162138