

Christine MOIK

A computational workflow supporting the evaluation of drug efficacy

Master Thesis



Institute of Genomics and Bioinformatics
Graz University of Technology
Petersgasse 14, 8010 Graz

Supervisors:
Dr. Paul Perco
Andreas Heinzl, MSc

Evaluator:
Dr. Gerhard Thallinger

Graz, May, 2014

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

.....
date

.....
(signature)

Acknowledgment

First of all, I would like to thank Dr. Paul Perco and Andreas Heinzl, MSc from emergentec for their supervision and constant guidance. I am grateful for the possibility to write my thesis at emergentec.

Many thanks also to Maria Martin's group at the European Bioinformatics Institute, the collaboration partner of this project, for fruitful discussions and helpful comments.

Furthermore, I would like to express my gratitude to Dr. Gerhard Thallinger from the Institute of Genomics and Bioinformatics, University of Technology Graz for his supervision and valuable advices for my thesis.

Thanks a lot to all my colleagues from emergentec for the encouraging working atmosphere and enjoyable time there.

Last but not least, I would like to thank my family for their love and support.

Abstract

English

Drug development suffers from high failure rates. Apart from adverse side effects the limited efficacy of a drug under study in the entire disease population is often a reason for stopping development. Specifically, single nucleotide polymorphisms (SNPs) occurring in the drug targets binding site region may hamper the efficacy of a drug.

In the course of this thesis a computational workflow for evaluating drug efficacy in the context of mutation status of drug targets was developed. The implemented workflow was exemplarily applied to drugs used in breast cancer therapy. Relevant drugs, drug targets, and SNPs were retrieved from the following publicly accessible biological databases: ClinicalTrials.gov, ChEMBL, DrugBank and COSMIC. A major part of the in-silico workflow deals with the integration of such heterogeneous data from different sources and making them accessible for common processing. The identified set of drugs and their targets were analysed in greater detail including a target relevance ranking.

A significant difference in the mutation ratios between targets of the identified breast cancer drug set and other human proteins was observed. Drug targets of the identified set reveal lower mutation ratios. In addition, they show a higher number of associated breast cancer publications.

Furthermore, SNPs occurring in drug-targets were evaluated. For some of them, evidence of the influence on the drug binding efficacy and consequently on the therapy outcome was found in literature.

Keywords: breast cancer, drug efficacy, data integration, in-silico workflow, clinical trials, drug-target interactions, SNPs

Deutsch

Die Entwicklung von Medikamenten ist durch eine hohe Ausfallrate gekennzeichnet. Abgesehen von unerwünschten Nebenwirkungen führt oft die unzureichende Wirkung eines getesteten Medikaments zum Abbruch der klinischen Studie. Insbesondere können Einzelnukleotid-Polymorphismen (SNPs), die sich in Targetgenen des Medikaments befinden, dessen Wirksamkeit beeinträchtigen.

Im Zuge dieser Arbeit wurde ein *in silico* Workflow zur Evaluierung der Wirksamkeit unter Berücksichtigung von SNPs in der Targetregion des Medikaments entwickelt. Dieser wurde am Beispiel von Brustkrebs erstellt und anhand von öffentlich zugänglichen Daten validiert. Dazu wurden die relevanten Wirkstoffe, Zielproteine (Targets) und SNPs aus den biologischen Datenbanken ClinicalTrials.gov, ChEMBL, DrugBank und COSMIC extrahiert.

Ein erheblicher Teil des *in silico* Workflows befasst sich mit der Integration heterogener Daten aus verschiedenen Quellen mit dem Ziel diese einer gemeinsamen Verarbeitung zugänglich zu machen. Die für die Anwendung bei Brustkrebs ermittelten Wirkstoffe sowie ihre Zielproteine wurden in weiterer Folge näher analysiert und eine Reihung hinsichtlich der Relevanz der Targets erstellt.

Dabei wurde ein signifikanter Unterschied zwischen Targets der ermittelten Brustkrebswirkstoffe und anderen humanen Proteinen beobachtet. Jene aus dem ermittelten Set zeigen eine deutlich niedrigere Mutationsrate. Zugleich sind sie mit einer größeren Anzahl an Brustkrebs Publikationen assoziiert.

Des Weiteren wurden SNPs, welche in Genen vorkommen, die für Zielproteine kodieren, sowie ihr Einfluss auf die Medikamentenwirksamkeit näher ausgewertet. Für einige der identifizierten SNPs konnte in der Literatur ein Nachweis der tatsächlichen Auswirkung auf die Bindungsfähigkeit der Medikamente an ihr Target gefunden und so der negative Einfluss auf die Wirksamkeit der Therapie aufgezeigt werden.

Stichworte: Brustkrebs, Medikamentenwirksamkeit, Datenintegration, *in silico* Workflow, klinische Studien, Medikament-Zielprotein-Interaktionen, SNPs

Table of contents

List of Figures	III
List of Tables	IV
Listings	V
Glossary	VI
1 Introduction	1
1.1 Objectives.....	5
2 Materials and Methods.....	7
2.1 Biological Data sources	7
2.1.1 ClinicalTrials.gov database	7
2.1.2 ChEMBL	7
2.1.3 DrugBank.....	8
2.1.4 National Cancer Institute.....	8
2.1.5 DailyMed.....	8
2.1.6 UniProt.....	8
2.1.7 COSMIC	8
2.1.8 Anatomical Therapeutic Chemical classification system.....	9
2.2 Software Tools	9
2.2.1 R	9
2.2.2 C	9
2.2.3 Auxiliary tools.....	10
2.3 Applied Methods.....	11
2.3.1 Breast cancer trials identification	11
2.3.2 Drug retrieval and mapping.....	13
2.3.3 Drug target identification	14
2.3.4 Target relevance ranking	17

2.3.5	SNP identification.....	18
3	Results.....	21
3.1	Workflow and implementation	21
3.2	Workflow results	31
3.2.1	Breast cancer trials	31
3.2.2	Breast cancer drug set and targets.....	35
3.2.3	Target relevance ranking	40
3.2.4	Drug-SNP evaluation	43
4	Discussion.....	46
4.1	Workflow construction	46
4.2	Main findings	47
4.3	Assessment of data sources	49
4.4	Outlook	52
	Literature	53
	Appendix.....	58

List of Figures

Figure 1.1: Trends in attrition rates.....	2
Figure 1.2: Possible influences of SNPs on efficacy of BC therapy.....	3
Figure 2.1: Workflow for evaluating drug-targets.....	11
Figure 2.2: Sample of the trials data extracted from ClinicalTrials.gov.....	12
Figure 2.3: Screen dump of a DrugBank DrugCard.....	16
Figure 2.4: Screenshot of the COSMIC Biomart BC sample query.....	19
Figure 3.1: ChEMBL breast cancer drug identification.....	22
Figure 3.2: ChEMBL breast cancer target identification.....	24
Figure 3.3: DrugBank breast cancer target identification.....	28
Figure 3.4: Target relevance ranking.....	29
Figure 3.5: SNP identification.....	30
Figure 3.6: Filter process of breast cancer trials.....	31
Figure 3.7: Tag cloud of condition terms of BC studies.....	32
Figure 3.8: Number of trials per phase.....	33
Figure 3.9: Distribution of drugs versus the highest development phase.....	34
Figure 3.10: Distribution of drugs versus number of trials.....	34
Figure 3.11: Distribution of drugs versus number of drug targets (ChEMBL).....	36
Figure 3.12: Distribution of drugs versus number of drug targets (DrugBank).....	37
Figure 3.13: DrugBank breast cancer drug set and its targets.....	38
Figure 3.14: Venn diagram of target coverage.....	39
Figure 3.15: Venn diagram of drug-target interactions.....	40
Figure 3.16: Boxplots of number of GeneRIFs (BC aggregated) of BC targets.....	41
Figure 3.17: Boxplots of number of BC papers associated with BC targets.....	41
Figure 3.18: Boxplots of disease sensitivity of BC targets.....	43
Figure 3.19: Boxplots of the mutation ratio of BC targets.....	44

List of Tables

Table 2.1: Simplified classification of clinical trial recruitment states	12
Table 2.2: Narrower MeSH terms of “Breast Neoplasms”	17
Table 2.3: Example data for mutation ratio calculation	20
Table 3.1: Distribution of ChEMBL drugs over ATC code groups	35
Table 3.2: Example of differences in target assignment	38
Table 3.3: Excerpt of target relevance ranking	42
Table 3.4: Example of mutated genes	45
Table 3.5: Most commonly annotated mutated genes	45

Listings

Listing 3.1: Code snippet of highest development phase calculation.....	20
Listing 3.2: Code snippet for determination of the highest recruitment state	21
Listing 3.3: Code snippet of combi_id calculation and C function calling.....	23
Listing 3.4: Code snippet of the C function calculating the pChEMBL ratio	24

Glossary

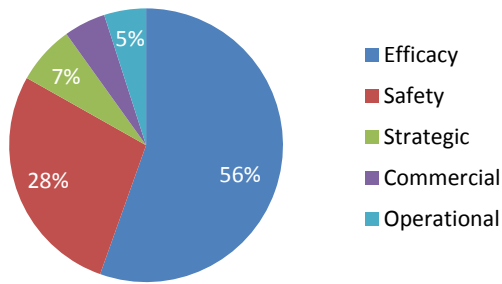
AA	Amino Acid
ATC	Anatomical Therapeutic Chemical
BC	Breast Cancer
CGP	Cancer Genome Project
COSMIC	Catalogue Of Somatic Mutations In Cancer
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENSG	Ensembl gene
ENST	Ensembl transcript
FDA	Food and Drug Administration (US)
GeneRIF	Gene Reference Into Function
MeSH	Medical Subject Heading
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NIH	National Institutes of Health
NLM	National Library of Medicine
PCR	Polymerase Chain Reaction
PIR	Protein Information Resource
SE	Sensitivity
SIB	Swiss Institute for Bioinformatics
SNP	Single Nucleotide Polymorphism
UniProt	Universal Protein Resource
URI	Uniform Resource Identifier
WHO	World Health Organisation
XML	Extensible Markup Language

1 Introduction

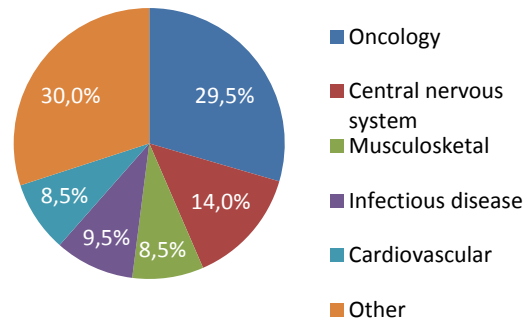
Breast cancer is the second most commonly diagnosed cancer type worldwide following lung cancer. Moreover, breast cancer is the most frequent cancer among women with an incidence rate of about 90 and a mortality rate of 21 per 100,000 women in Austria in 2012 [1]. Breast cancer medication regimes comprise chemotherapy, hormone blocking therapy and monoclonal antibodies (anti-HER2). With the identification of five intrinsic molecular subtypes of breast cancer, known as luminal A, luminal B, Her2-enriched, basal-like (triple-negative) and Claudin-low, as well as the advances in omics technologies, a strong focus on personalized therapy has evolved [2]. However, this also requires an assessment of drug binding efficacy taking into account the influence of inter-individual genomic variations.

Drug development is a long lasting process which has become more extensive and complex within the past few decades [3]. It comprises preclinical studies using animal models followed by clinical trials in humans including in general four phases during which the safety and efficacy is tested. Failure rates in clinical trials are still high [4]. Only about 10% of new agents tested from Clinical Phase II and 50% tested from Phase III [5] ultimately receive approval. Success rates of drug approvals vary considerably between different therapeutic areas with oncology still showing the lowest [6-8]. Only one of every 5,000-10,000 antineoplastic drugs under development reaches FDA approval and only 5% of oncology drugs entering Phase I clinical trials are finally approved [5, 9]. According to an analysis of Arrowsmith *et al.* in the years 2011 and 2012, 148 drugs failed in total between Phase II and approval. Among these, 105 had indicated reasons for attrition, e.g. safety issues (28%). However, the majority (56%) failed due to a lack of drug efficacy [10]. Figure 1.1 shows the trends in attrition rates resulting from the Arrowsmith analysis.

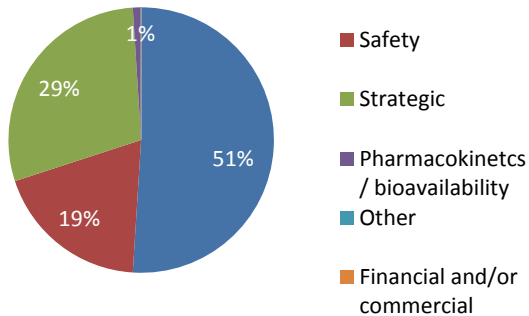
a Causes of failure



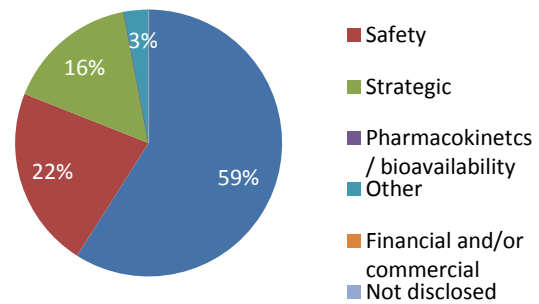
Failure by therapeutic area



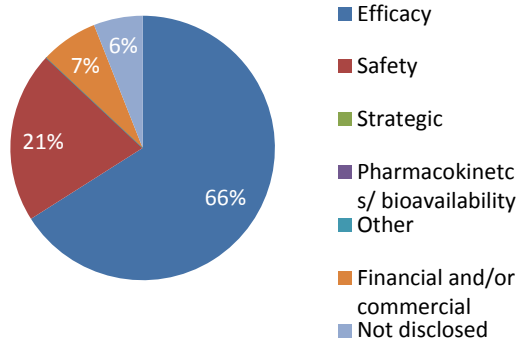
b Phase II failures 2008-2010



Phase II failures 2011-2012



Phase III and submission failures 2007-2010



Phase III and submission failures 2011-2012

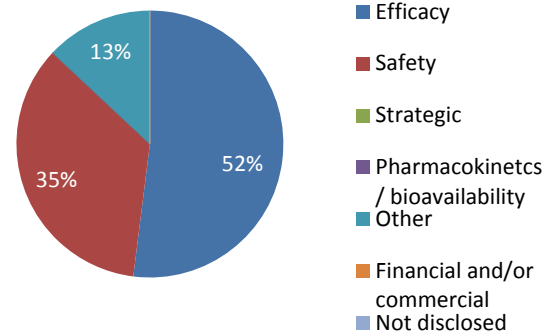


Figure 1.1: Trends in attrition rates. (a) On the left, the majority of failures were caused by limited efficacy. As shown on the right, oncology is the single therapeutic area with the highest failure rates. (b) Comparison of the reasons for drug failures in Phase II and Phase III trials in different periods of time (modified from [10]).

As shown in Figure 1.1, in Phase II as well as in Phase III and beyond, the limited drug efficacy represents the main reason for drug failure.

Drug efficacy is directly correlated to the proper functionality of the drug's binding mechanism to its intended target.

Drug targets can be genetic material such as DNA but mostly proteins such as enzymes, membrane receptors, transporters and ion channels [11]. The most important protein targets are the membrane receptors. Drugs can either stimulate them (i.e. they are agonists) or block them (i.e. they are antagonists) inhibiting a further interaction with an agonist. The activation of a receptor may directly lead to an opening or closing of ion channels or a transduction of the drug's signal (inhibition, activation) to enzymes and thus an influence of biological pathways. In the case of cancer, drugs may e.g. block the ligand binding domain of the epidermal growth factor receptor (EGFR) or inhibit the EGFR tyrosine kinase resulting in a prevention of EGFR activation and in this way interrupting its signalling pathway and tumour proliferation.

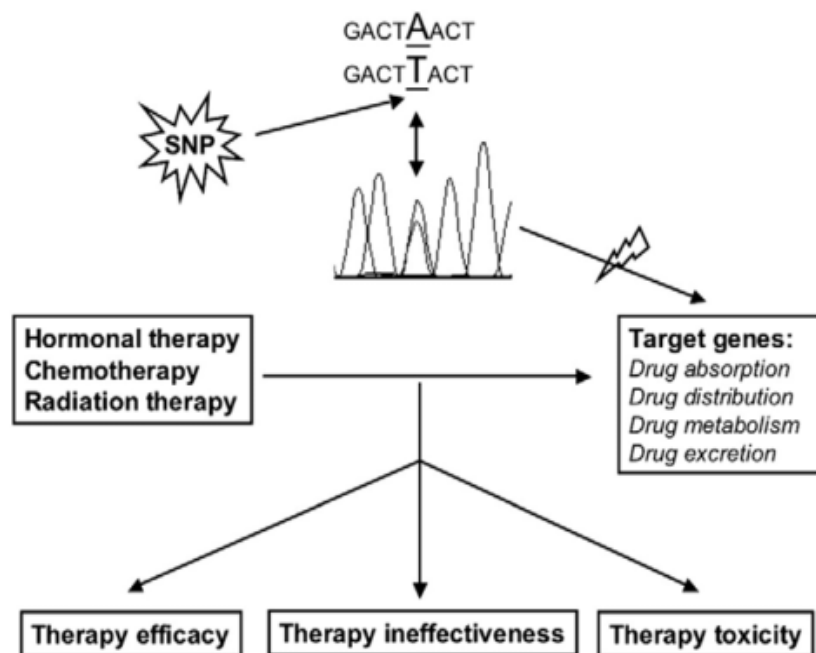


Figure 1.2: Possible influences of SNPs on efficacy of BC therapy. The drug activity may be altered by individual genetic variations in target genes involved in drug absorption, distribution, metabolism and excretion (figure from [15]).

However, the variations in the individual's genetic profile may influence the intended drug binding and lead to inter-patient variability of drug efficacy and toxicity. These differences occur partly due to mutations in genes encoding drug targets, e.g. receptors [12, 13]. Especially in the case of cancer, single nucleotide polymorphisms

(SNPs) located in the drug targets binding site region may hamper drug efficacy (see Figure 1.2). Consequently, the characterization of toxicity and efficacy in an early stage of drug development might reduce drug failure rates [14].

Genetic variability further plays a crucial role in choosing the appropriate therapy as well as predicting the clinical outcome [15, 16]. The availability of molecular criteria such as mutation status for predicting sensitivity to a specific group of drugs hence represents an important prerequisite for prescribing the correct medication [17].

Single nucleotide polymorphisms (SNPs) represent the most common polymorphisms in the human genome [18]. This form of genetic variation is found within every 1,000 nucleotides on average. SNPs may occur in non-coding as well as in coding regions of the genome. Those being located in a coding sequence may either change only a codon without any effect on the amino acid (synonymous SNPs) or result in amino acid substitutions (nonsynonymous SNPs) [19]. As a consequence, this change in the amino acid sequence might affect the function of the protein product [16].

With the completion of the Human Genome Project (HGP) in 2003, which aimed to sequence the whole human genome, followed by the International HapMap Project, millions of DNA sequence variants, mostly SNPs, have been detected. However, the majority of the identified SNPs has not yet been associated with drug response or disease susceptibility [20].

Nowadays, several technologies are available for SNP detection encompassing scanning of DNA sequences for previously unknown polymorphisms as well as genotyping individuals for known polymorphisms. These methods include DNA sequencing technologies such as pyrosequencing, hybridization-based methods (e.g. molecular beacons, SNP microarrays), enzyme-based methods (e.g. primer extension, PCR-based methods) as well as other methods based on physical properties of DNA (e.g. single strand conformation polymorphism) [21, 22]. With the emergence of high-throughput sequencing technologies, which enable large-scale characterization of SNPs, also vast amounts of sequencing data are produced and need to be stored. The most commonly used genomic repositories are the databases of the European Molecular Biology Laboratory (EMBL)/European Bioinformatics Institute (EBI) e.g. Ensembl [23], the National Center of Biotechnology (NCBI, GenBank [24]) database

and the DNA Database of Japan (DDBJ) [25]. One of the most popular SNP repositories maintained by the NCBI is the dbSNP database [26], a public-domain archive containing millions of genomic variants [27]. Further sources for genetic variants are OMIM [28], the Online Mendelian Inheritance in Man database, and COSMIC (Catalogue Of Somatic Mutations In Cancer) [29] which focuses on mutations in human cancer. This flood of individual genomic data leads to new bioinformatics challenges in several areas, one of them the interpretation of the functional effect and the impact of genomic variation [30]. With the growing number of public accessible biological databases also the demand on bioinformatics solutions to analyse these data is expanding. Data integration as a prerequisite plays a key role in this field. Next to the quantity in data, the complexity of the different datasets makes data integration efforts a challenge. This complexity arises for example from the diversity of sources, the variability of data quality and evidence levels and the diversity in types of data [31]. Similar data can be contained in several databases but their representation may be different. To overcome this problem of representational heterogeneity and to ease the entity identification across sources, several online databases provide cross references to other repositories but they are by far not complete.

1.1 Objectives

The overall aim of the present thesis was the development of a computational workflow for identifying a drug-target dataset for drugs in clinical development for a given indication in order to evaluate whether prevalent SNPs in drug targets hamper drug efficacy. The workflow should be exemplified on the disease term “breast cancer”. Specifically, it should include the following steps:

- Extraction of clinical trials on breast cancer
- Identification of investigated drugs in these trials
- Extraction of drug targets of these drugs
- Identification of prevalent SNPs within genes encoding these targets
- Prediction of drug binding domains
- Integration of the collected information for evaluating if SNPs located in the genomic regions of drug target binding sites have an influence on drug efficacy by evaluating information available from scientific publications

- Evaluation of functional association of the targets to breast cancer based on scientific literature

Drug-domain prediction is carried out by the collaboration partner of the project.

The workflow itself as well as the results on the breast cancer targets will be presented and discussed.

2 Materials and Methods

This chapter provides an overview of biological data resources and software tools used throughout this thesis. Furthermore, a description of the applied methods is given.

2.1 Biological Data sources

This work is based on data retrieved from the following data sources.

2.1.1 ClinicalTrials.gov database

ClinicalTrials.gov [32] is a publicly accessible database maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH). It lists interventional and observational studies and covers a wide range of diseases and conditions, including cancer. The registered studies are conducted in 185 countries and all 50 U.S. States. According to the FDA Amendments Act of 2007 “basic” results of clinical trials which involve testing of FDA-regulated and FDA-approved chemical or biologic agents or medical devices need to be reported in ClinicalTrials.gov.

The ClinicalTrials.gov results database contains information on study participants and a summary of study outcomes in a tabular format. On ClinicalTrials.gov, studies can be identified via “Basic Search”, “Advanced Search”, by topic or location. The study record data of the search result can then be downloaded in Extensible Markup Language (XML) and other data formats.

2.1.2 ChEMBL

ChEMBL [33] is an open access bioactivity database (established in 2009) that is maintained by the European Bioinformatics Institute (EBI). It contains drugs, drug-like small molecules and their bioactivity data.

ChEMBL data is regularly updated, with releases approximately every three to four months. In August 2013, EBI released the 17th version of the ChEMBL database with more than 1 million distinct compounds and 12 million activity measures for about 9,300 targets.

Access to ChEMBL is available via a web-based interface, data downloads, web services as well as via the EMBL-EBI RDF Platform [34].

2.1.3 DrugBank

DrugBank [35] is a publicly accessible database maintained by the University of Alberta. It serves as bioinformatics and chemoinformatics resource for drug and drug target information. DrugBank contains more than 7,600 drug entries of FDA-approved drugs, experimental drugs and nutraceuticals. The web page allows a direct search as well as a download in XML or CSV (Comma-separated values) format.

2.1.4 National Cancer Institute

The National Cancer Institute (NCI) belongs to the NIH (National Institutes of Health), which is part of the U.S. Department of Health and Human Services. It is the U.S. government's principal agency for cancer research and training. On the web site of the NCI, information on cancer, clinical trials and cancer statistics as well as a drug dictionary are provided.

2.1.5 DailyMed

The DailyMed [36] website provides information on FDA-approved drugs such as FDA labels, descriptions, usage, dosage and warnings. It is maintained by the U.S. National Library of Medicine (NLM) and updated daily by the U.S. Food and Drug Administration (FDA).

2.1.6 UniProt

UniProt [37], also known as “universal protein resource”, is a public resource for protein sequence and annotation data established by the European Bioinformatics Institute (EBI), the Swiss Institute for Bioinformatics (SIB) and the Protein Information Resource (PIR).

UniProt hosts several tools such as text search, sequence alignments and database identifier mappings.

2.1.7 COSMIC

The catalogue of somatic mutations in cancer (COSMIC) [29] is a public accessible online database hosted by the Sanger Institute, UK, which provides information on somatic mutations occurring in human cancers. Data are curated from scientific literature as well as from the Cancer Genome Project (CGP) at the Sanger Institute

(<http://www.sanger.ac.uk/research/projects/cancergenome/>). The COSMIC website offers a data search either by cancer type, gene- or sample name.

Via the COSMIC Biomart [38] somatic mutation data can be filtered, e.g. by primary histology, and easily downloaded.

2.1.8 Anatomical Therapeutic Chemical classification system

The Anatomical Therapeutic Chemical (ATC) classification system was established by the WHO (World Health Organisation) Collaborating Centre for Drug Statistics Methodology in 1990 [39].

This system can be used to compare data on drugs regarding the organ or anatomical system on which they act as well as their therapeutic, pharmacological and chemical characteristics. Drugs are classified into groups at five different hierarchy levels. The first three ATC levels indicate the anatomical main group as well as the therapeutic/pharmacological group. The fourth and the fifth level specify the drug's chemical structure.

2.2 Software Tools

2.2.1 R

R [40] is an open source programming language and environment for statistical computing and graphical representation of data. The data processing and analysis of the present master thesis was performed using R version 3.0.2 in combination with RStudio [41], version 0.97.551, which is an open source integrated development environment (IDE) for R.

2.2.2 C

For resource intensive data manipulations leading to long processing time, the programming language C [42] was used in addition. Programming was carried out by means of Code::Blocks 13.12 [43], a free integrated development environment, using the MinGW compiler. The functions coded in C were compiled into a shared library (dll) which can be called from inside the R script (see R manual "Writing R extensions", chapter 5.3: http://cran.r-project.org/doc/manuals/R-exts.html#dyn_002eload-and-dyn_002eunload).

2.2.3 Auxiliary tools

In addition to R, the following tools were used for visualisation:

- MS Visio (version 2007) for drawing the workflow charts and Venn diagrams
- Tagxedo (<http://www.tagxedo.com/>) for creating a tag cloud

The initial drug mapping step was performed by making use of the DiseaseLookup Excel Add-In developed as in-house software by emergentec. It is a managed plugin built with Visual Studio Tools for Office. Apache Lucene.Net (lucenenet.apache.org/) is used to construct a search index from a drug thesaurus. This search index is subsequently used for fuzzy matching of a user selected value against the drug thesaurus. The DiseaseLookup Excel Add-In provides a Task Pane that allows the user to query the index either with the value of the currently selected cell or with a manually entered value. Search results are presented in a *ListBox* from which the best hit can be selected and merged back into the active document.

2.3 Applied Methods

This section presents the components of the workflow as well as a detailed description of the applied methods. Figure 2.1 shows the procedure that was followed to evaluate drugs, their targets and prevalent SNPs.

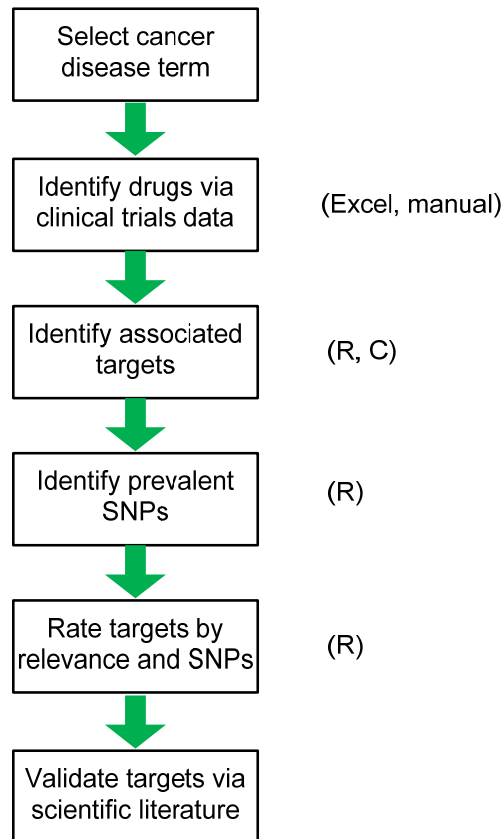


Figure 2.1: Workflow for evaluating drug-targets with respect to target relevance and prevalent SNPs. It consisted of the following steps: (i) selection of a cancer disease term, (ii) identification of drugs via clinical trials data by making use of a DiseaseLookup Excel Add-In and further manual mapping, (iii) identification of associated targets by integrating data extracted from biological databases using R and C, (iv) identification of prevalent SNPs using R for data manipulation, (v) target rating according to relevance and mutation frequency using R, (vi) validation of targets by means of scientific literature.

2.3.1 Breast cancer trials identification

An advanced search with the condition “breast cancer” was performed on 28th of October 2013 in ClinicalTrials.gov in order to identify all breast cancer trials available at that date. Figure 4 shows a sample of the search results returned.

NCT Number	NCT01547741
Title	Docetaxel and Cyclophosphamide Compared to Anthracycline-Based Chemotherapy in Treating Women With
Recruitment	Recruiting
Study Results	No Results Available
Conditions	Breast Cancer
Interventions	Drug: Doxorubicin Drug: Cyclophosphamide Drug: Docetaxel Drug: Paclitaxel
Sponsor/Collaborators	National Surgical Adjuvant Breast and Bowel Project (NSABP) National Cancer Institute (NCI)
Gender	Female
Age Groups	Adult Senior
Phases	Phase 3
Enrollment	1843
Funded Bys	Other NIH
Study Types	Interventional
Study Designs	
Other IDs	NSABP-B-49 NCI-2012-00701
First Received	6-Mar-12
Start Date	12-Apr
Completion Date	19-May
Last Updated	22-Oct-13
Last Verified	13-Oct
Results First Received	No Study Results Posted
Acronym	
Primary Completion Date	18-May
Outcome Measures	Invasive disease-free survival (IDFS) Disease-free survival (DFS-DCIS) Overall survival (OS) Recurrence-free i
URL	http://ClinicalTrials.gov/show/NCT01547741

Figure 2.2: Sample of the trials data extracted from ClinicalTrials.gov. Each registered study has a unique trial ID consisting of “NCT” followed by an 8-digit number. Furthermore, information on e.g. the recruitment state, interventions and study phase are usually indicated. The highlighted fields were used for further analysis.

From the resulting list of breast cancer trials only studies containing at least one intervention of type “Drug” were considered for further analysis, thus excluding all studies e.g. on medical devices, procedures and vaccines.

In the studies of ClinicalTrials.gov, eleven different recruitment states are used. These eleven states were grouped into the following five broader recruitment states to guarantee sufficient numbers in the individual groups in further analysis steps: active, completed, not active, terminated and withdrawn. The table below shows the classification used.

Table 2.1: Simplified classification of clinical trial recruitment states. The first column shows the introduced recruitment states combining all states listed in the second column.

Active	“Active, not recruiting”
Completed	“Approved for marketing”, “Completed”, “No longer available for expanded access”
Not active	“Available for expanded access”, “Enrolling by invitation”, “Not yet recruiting”, “Recruiting”, “Suspended”
Terminated	“Terminated”
Withdrawn	“Withdrawn”

In addition to the recruitment state, information on the phase of the trial (phase 0, phase 1, phase 2, phase 3, phase 4) was extracted. In case multiple phases were listed (e.g. phase 2/3) the higher phase was used for further analysis.

2.3.2 Drug retrieval and mapping

2.3.2.1 Extraction of drugs from clinical trials

From the set of eligible breast cancer trials described above, all tested drugs were extracted from the intervention attribute and their names were mapped to a common reference namespace. This process was carried out by making use of an already existing DiseaseLookup Excel Add-In that is accessing a list of all ChEMBL compounds including their synonyms and drug names as well as the URI of the Compound Report Card. As the Add-In was case sensitive and not tolerating additional spaces and hyphens, approximately two-thirds of the drugs could be mapped automatically with the remaining thirty percent requiring a manual step. This consisted of searching the drug in the ChEMBL database, extracting the ChEMBL ID and inserting it into the Excel sheet. Drugs not found in ChEMBL were searched in the NCI catalogue. Finally, the drug list was reviewed to exclude drugs which have not been specifically administered to tackle the tumour but merely given for the treatment of side effects (e.g. analgesics, anaesthetics, antibiotics) or compounds such as imaging agents (often recognisable by additional terms like “fluoro” or “PET scan”).

2.3.2.2 ChEMBL drugs to DrugBank

In order to verify the drugs retrieved via ChEMBL and to provide a curated repository of drug targets for the next steps, a second resource, DrugBank, containing detailed drug data with drug target information was used. For this purpose the identified ChEMBL breast cancer drug set was further mapped to DrugBank using (i) the cross reference mapping provided by ChEMBL and (ii) manual (human interference) identification of the respective entry in DrugBank. In general each ChEMBL drug ID referred to one unique DrugBank ID.

2.3.3 Drug target identification

On the one hand drug targets were extracted from ChEMBL providing information on drug mechanism of actions as well as bioactivity data, and on the other hand targets were retrieved via DrugBank as outlined in the next sections in more detail.

2.3.3.1 Drug target identification via ChEMBL

2.3.3.1.1 Curated

ChEMBL provides information on the mechanism of action of FDA-approved drugs. Targets are only included in the mechanism table if the drug interacts directly with the target and its interaction exerts the efficacy in the disease for which the drug is indicated. The mechanism table was extracted from the ChEMBL MySQL database dump (<ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/>) and used to identify the targets associated to the present breast cancer drug set.

2.3.3.1.2 Experimental

To identify further targets, specifically for not approved drugs, activity information available in the ChEMBL database was retrieved via the EMBL-EBI RDF platform [44]. This bioactivity data comprises several parameters such as confidence score and pChEMBL value. The confidence score value reflects on the one hand the type of a target tested in a specific assay and on the other hand the confidence in the correctness of the assigned target for a particular assay. It ranges from 0 (“target assignment has yet to be curated”) to 9 (“direct single protein target assigned”). The pChEMBL value (> 0) arises from several measures like affinity and potency (<https://www.ebi.ac.uk/chembl/faq#faq67>). In order to evaluate the drug-target interactions via bioactivity assays, a threshold for the confidence score as well as for the pChEMBL value was chosen. In this thesis only activity entries with pChEMBL values and assay data with assay confidence scores ≥ 8 (8 = homologous single protein target assigned, 9 = direct single protein target assigned) were selected for further analysis. Moreover all non-human targets were excluded and only reviewed human UniProt targets were considered. A pChEMBL threshold value of 7 was applied. Since multiple activity entries for the same drug-target interaction can exist having different pChEMBL values, the concept of a pChEMBL ratio was introduced for consolidation. The pChEMBL ratio was defined as the number of activity entries for a specific drug-target interaction with a pChEMBL value above the threshold divided by

the total number of activities for that drug-target interaction. It was calculated per drug-target interaction as follows:

$$pChEMBL_{ratio} = \frac{\text{Number of activity entries with } pChEMBL > 7}{\text{Total number of activity entries}} \quad (1)$$

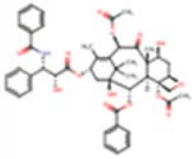
All drug-target interactions with pChEMBL ratios higher than 0.75 were considered as positive interactions.

The final ChEMBL breast cancer target list resulted from merging the directly identified ChEMBL breast cancer drug set with targets derived from the activity entries.

2.3.3.2 Drug target identification via DrugBank

DrugBank drug records hold drug target associations which are manually curated from scientific literature. Drug targets are characterized by their gene name, pharmacological action as well as references to PubMed citations.

Figure 2.3 shows clippings of a DrugBank drug entry, the “DrugCard”, containing specific information on the drug and its target(s).

Name	Paclitaxel
Accession Number	DB01229 (APRD00259, DB05261, DB05927)
Type	small molecule
Groups	approved
Description	Paclitaxel is a mitotic inhibitor used in cancer chemotherapy. It was discovered in a US National Cancer Research Triangle Institute in 1967 when Monroe E. Wall and Mansukh C. Wani isolated it from the bark <i>Taxus brevifolia</i> and named it taxol. Later it was discovered that endophytic fungi in the bark synthesize it. When it was developed commercially by Bristol-Myers Squibb (BMS), the generic name was changed to paclitaxel. In this formulation, paclitaxel is dissolved in Kolliphor EL, a delivery agent. A newer formulation, in which paclitaxel is bound to albumin, is sold under the trademark
Structure	 <p>MOL SDF PDB SMILES InChI View Structure</p>
Synonyms	10 records per page

1. **Tubulin beta-1 chain**
 Kind: protein
 Organism: Human
 Pharmacological action: **yes**
 Actions: inhibitor

Components

Name	UniProt ID	Details
Tubulin beta-1 chain	Q9H4B7	Details

References:

- Overington JP, Al-Lazikani B, Hopkins AL: How many drug targets are there? *Nat Rev Drug Discov* 2006 Dec;5(12):993-6. [Pubmed](#)
- Imming P, Sinning C, Meyer A: Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* 2006 Oct;5(10):821-34. [Pubmed](#)
- Cheung CH, Chen HH, Kuo CC, Chang CY, Coumar MS, Hsieh HP, Chang JY: Survivin counteracts the therapeutic effect of microtubule destabilizers by stabilizing tubulin polymers. *Mol Cancer* 2009 Jul 28;8:1

Figure 2.3: Screen dump of a DrugBank DrugCard showing the breast cancer drug Paclitaxel and one of its targets. The highlighted fields were used for further analysis.

For this work DrugBank drug target information was extracted from the drug target identifiers file provided in the download section of the DrugBank homepage. UniProt targets with the species category "human" were selected and used for the drug to target mapping.

2.3.4 Target relevance ranking

The targets of the final drug set (ChEMBL and DrugBank) were analysed with respect to breast cancer disease associations extracted from scientific publications.

Therefore, the number of breast cancer aggregated Gene References Into Function (GeneRIFs) for each gene was derived from the information provided in NCBI's gene2geneRIF file (ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/generifs_basic.gz).

GeneRIFs are short phrases which describe the function of a gene including the reference to a corresponding PubMed publication. They are manually added in NCBI's Gene database by scientists in order to enrich the functional annotation of a gene.

Articles in NCBI's PubMed are indexed by Medical Subject Headings (MeSH) [45] which is a controlled vocabulary that can also be used as a thesaurus for searching. Only GeneRIFs referencing a paper having the MeSH term "Breast Neoplasms" or one of its narrower terms (listed in Table 2.2) as major concept were considered for the score calculation, i.e. breast cancer "aggregated" GeneRIFs were counted.

Table 2.2: Narrower MeSH terms of "Breast Neoplasms"

Breast Neoplasms
Breast Neoplasms, Male
Carcinoma, Ductal, Breast
Hereditary Breast and Ovarian Cancer Syndrome
Inflammatory Breast Neoplasms
Triple Negative Breast Neoplasms

The number of breast cancer associated papers for each gene ("#disease BC agg papers") was derived by counting the number of breast cancer associated papers linked to the gene. The following PubMed query was used to identify a set of breast cancer associated papers:

```
"Breast Neoplasms/pathology[majr] OR Breast Neoplasms/physiopathology[majr] OR Breast Neoplasms/enzymology[majr] OR Breast Neoplasms/chemistry[majr] OR Breast Neoplasms/metabolism[majr]"
```

Gene paper links were used as provided in NCBI's gene2pubmed (available under <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz>).

The sensitivity of each gene for breast cancer (“SE disease BC”) was estimated based on the number of breast cancer papers linked to that gene. The genes having at least one associated paper (2,998 genes) were sorted according to this number. The sensitivity value represents the percentile ranking of the gene under investigation in this list.

$$SE_{disease\ BC} = \left(1 - \frac{\text{Number of genes with more associated papers}}{\text{Total number of genes with associated papers}}\right) * 100 [\%] \quad (2)$$

To compare the target relevance ranking of the different target sets, boxplots were generated. ChEMBL targets were grouped in the following subsets: positive assessed ChEMBL BC targets, negative assessed ChEMBL BC targets and ChEMBL targets being not part of the identified BC target set. DrugBank targets were classified into pharmacological active targets of BC approved drugs, pharmacological active targets of BC experimental drugs, pharmacological inactive targets of BC approved drugs, pharmacological inactive targets of BC experimental drugs and pharmacological inactive targets that are not part of the identified BC target set. In case a protein has been a target of an approved as well as an experimental drug, it was classified as approved drug target.

To assess the differences between the target and non-target sets as well as the pharmacological active and pharmacological inactive target set, Wilcoxon tests were calculated setting the significance level to < 0.05. The Benjamini-Hochberg correction was used to adjust the p-values for multiple comparisons. Statistical tests as well as generation of boxplots were done in R.

2.3.5 SNP identification

For identifying SNPs occurring in breast cancer tissue the COSMIC database was used. All COSMIC variants isolated from breast samples were extracted via COSMIC Biomart, COSMIC v68, (retrieval date: 2014-02-16). The figure below provides a screenshot of the COSMIC query illustrating the used dataset, filters and attributes.

The screenshot shows the COSMIC Mart interface. The dataset is COSMIC68. Filters applied are Primary Site: breast and Primary Histology: carcinoma. The results table is as follows:

Pubmed ID	Primary Histology	Primary Site	COSMIC Mutation ID	AA Mutation Syntax	Gene Name	Accession Number	COSMIC Sample ID	Sample Name	Sample Source	Tumour Source	Swissprot ID
12068308	carcinoma	breast			BRAF	ENST00000288602	687503	ZR-75-1	cell-line	NS	P15056
15908952	carcinoma	breast			FGFR1	ENST00000447712	749730	PD1236a	NS	primary	P11362
15908952	carcinoma	breast			ADCK4	ENST00000324464	749728	PD1234a	NS	primary	
15908952	carcinoma	breast			MASTL	ENST00000375946	749709	HCC1954	cell-line	primary	
15908952	carcinoma	breast			SRPK3	ENST00000393786	749732	PD1238a	NS	primary	Q9UPE1
15908952	carcinoma	breast			STK36	ENST00000295709	749717	HCC38	cell-line	NS	
15908952	carcinoma	breast			ACVR1C	ENST00000243349	749718	PD0025a	NS	primary	
15908952	carcinoma	breast			MTOR	ENST00000361445	749722	PD0119a	NS	primary	P42345
15908952	carcinoma	breast			GRK7	ENST00000264952	749722	PD0119a	NS	primary	Q8WTO7
15908952	carcinoma	breast			STK38	ENST00000229812	749717	HCC38	cell-line	NS	Q15208

Figure 2.4: Screenshot of the COSMIC Biomart BC sample query. As dataset COSMIC68 was chosen and the following filters were applied: Primary Site: breast, Primary Histology: carcinoma. In the right, an excerpt of the results table is illustrated.

The results table comprised 29 columns e.g. Pubmed ID, Gene Name, COSMIC Sample ID, Sample Name, Accession Number, Swissprot ID, AA Mutation Syntax. As the Swissprot ID was not indicated for every entry, the Accession Numbers, if consisting of an ENST accession, were mapped to UniProt ID by making use of the UniProt ID mapper (<http://www.uniprot.org/mapping/>). For entries which did not have ENST accession numbers but e.g. NCBI Refseq or GenBank ID, the Swissprot ID if provided by COSMIC was used.

In order to compare the frequency of diverse mutations detected per gene, a mutation ratio was calculated. The mutation ratio was defined as the number of mutated samples divided by the total amount of examined samples per gene. If there were several mutations found in the same sample, it was still counted only once as mutated sample. As example the mutation ratio for the data provided in Table 2.3 is calculated as follows. The data contains five different samples (unique sample IDs). Among these, three samples harbour mutations resulting in a mutation ratio of 0.6.

Table 2.3: Example data for mutation ratio calculation including UniProt ID, COSMIC mutation ID and COSMIC sample ID.

UniProt ID	COSMIC mutation ID	COSMIC sample ID
A1Z1Q3	NA	1184078
A1Z1Q3	NA	1230672
A1Z1Q3	162314	1331073
A1Z1Q3	162315	1331073
A1Z1Q3	NA	1779393
A1Z1Q3	443475	1779393
A1Z1Q3	443474	1899732

For further analysis only mutation ratios of genes with a tested sample size bigger than 20 were considered.

3 Results

In the first part of this chapter the design as well as the implementation of the developed workflow is described in greater detail. The main intention was to apply this workflow to breast cancer and to achieve a detailed overview of approved and experimental drugs, their targets and SNPs influencing their efficacy. The results of this analysis are covered in the second part of the results section.

3.1 Workflow and implementation

The first step consisted of identifying a breast cancer drug set. Figure 3.1 illustrates the input and output tables used in this step as well as the analyses executed. The initial table retrieved from ClinicalTrials.gov (Table A: Study) was remodelled and filtered for “Drug” interventions. By looking up the drug names in ChEMBL, the table was extended with a ChEMBL drug ID and a ChEMBL drug name. The mapping process finally leads to Table B (for details of the mapping process see chapter 2.3.1). This first procedure was executed by means of Microsoft Excel. All further steps were implemented as R scripts. This comprised the import of Table B as a data.frame, further simplifications leading to Table C and a common analysis of both tables.

Out of Table C four subsets were generated: ChEMBL drugs, NCI drugs, drugs that are not breast cancer relevant (“other”) and drugs that could neither be found in the ChEMBL drug catalogue nor in NCI’s drug dictionary. This information was stored in the attribute column “comment”. The column “ChEMBL ID” either contains ChEMBL drug links which include the ID or the ones of NCI. In subsequent steps of the workflow only drugs indexed in ChEMBL were considered thus forming the final drug set used for the following analysis.

One way to classify the ChEMBL drug set was via the Anatomical Therapeutic Chemical Classification System. The ATC codes of the ChEMBL drugs were retrieved via the EMBL-EBI RDF platform and merged into the unique ChEMBL drug set using the ChEMBL ID. A substring of the first three characters of the ATC code was generated in order to group the ChEMBL breast cancer drug set and to check the number of neoplastic agents.

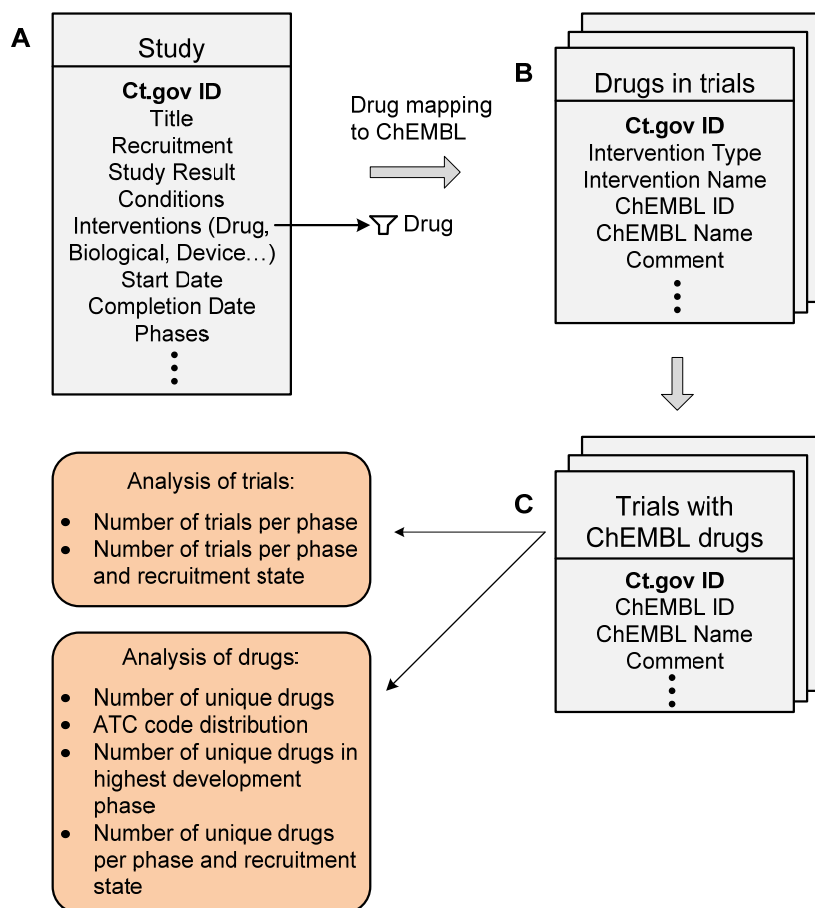


Figure 3.1: ChEMBL breast cancer drug identification: Input and output tables of the ChEMBL breast cancer drug identification step (grey boxes) and output of the analysis of trials and drugs (red boxes).

Furthermore, the identified drugs were analysed in greater detail pertaining to the phase and recruitment state of the trials they were tested in. To identify the number of unique drugs per phase and recruitment state Table A was merged with Table C by column “ClinicalTrials.gov ID”. However, this did not result in unique drug rows as the same drug might be used in several trials. Therefore, the highest development phase of each drug was determined. The following code snippet shows this procedure:

```
#data of drugs with no phase indicated are ignored:
data_phase <- subset(data_phase_drug,data_phase_drug$Phases.y != "")
#simplify phase labels so that only phase indices are left:
data_phase$Phases.y <- gsub("Phase","",data_phase$Phases.y)
#transform phase indices to type "integer":
data_phase$Phases.y <- as.integer(data_phase$Phases.y)
data_phase_s <- data_phase
data_phase_s$comment = NULL #delete the column "comment"
data_phase_s$ChEMBL_name = NULL #delete the column "ChEMBL_name"
data_max <- data.frame() #create new dataframe for max phases
for (i in 1: nrow(data_phase_s))
{
  #retrieve indices of all occurring phases of a specific drug:
```

```

g <- which(data_phase_s$ChEMBL_id == data_phase_s[i,1])
#return index of g of the highest phase of the selected drug:
k <- which.max(data_phase_s[g,2])
i_max <- g[k] #returns equivalent index of data_phase_s
#add the drug ID and its highest phase to data_max:
data_max <- rbind(data_max, data_phase_s[i_max,])
}
#unique ChEMBL and NCI drugs in highest development phase:
data_max <- unique(data_max)

```

Listing 3.1: Code snippet of highest development phase calculation.

As one drug (in its highest development phase) could also have multiple recruitment states, it was necessary to sort the recruitment states according to completeness state. To avoid time consuming loops with string comparisons, the recruitment state string was replaced by a completeness index which was derived in the following way: "Completed"= 5, "Active"= 4, "Not active"= 3, "Terminated"= 2, "Withdrawn"= 1.

This index was then used for sorting instead of the recruitment state as shown in the following code snippet:

```

data_max_state <- data.frame()#create new dataframe
for (i in 1: nrow(data_max))
{
g <- which(data_trials$ChEMBL_id == data_max[i,1] & data_trials$Phases ==
data_max[i,2]) #select drugs with certain ChEMBL ID and phase
#dataframe with drugs, their highest phase and recruitment states:
data_max_state <- rbind(data_max_state, data_trials[g,])
}
data_max_state <- unique(data_max_state) #remove duplicates

data_max_state_x <- data.frame()
data_max_state_num <- data_max_state
states_num <- c("5", "4", "3", "2", "1") #recruitment states as numeric
#recruitment states:
states <- c("Completed", "Active", "Not active", "Terminated", "Withdrawn")
#loop for substitution of recruitment states by the introduced numeric values:
for(k in 1:length(states))
{
data_max_state_num$Recruitment <-gsub(states[k],states_num[k],
data_max_state_num$Recruitment)
}
for (i in 1: nrow(data_max_state_num))
{
#dataframe subset containing all recruitment states of a certain drug (in
#highest development phase):
drugx_states <- subset(data_max_state_num, ChEMBL_id ==
data_max_state_num[i,1])
k <- which.max(drugx_states[,2]) #index of the most completed recruitment state
#dataframe of drugs with their highest development phase and most completed
#recruitment state
data_max_state_x <- rbind(data_max_state_x, drugx_states[k,])
}
}

```

```

data_max_state_x_unique <- unique(data_max_state_x) #remove duplicates
#indices of ChEMBL drugs:
i <- grep("http://rdf.ebi.ac.uk/resource/chembl/molecule/",
data_max_state_x_unique[,1], fixed = TRUE)
data_max_state_x_unique_chembl <- data_max_state_x_unique[i,]
#loop for substitution of numeric values by recruitment states:
for(k in 1:length(states))
{
  data_max_state_x_unique_chembl$Recruitment <- gsub(states_num[k], states[k],
data_max_state_x_unique_chembl$Recruitment)
}

```

Listing 3.2: Code snippet for determination of the highest recruitment state belonging to the respective highest development phase.

During the second step of the workflow targets associated to the drug set were identified either via information provided in ChEMBL or DrugBank. Figure 3.2 shows the input and output tables used in the ChEMBL breast cancer drug-target identification step.

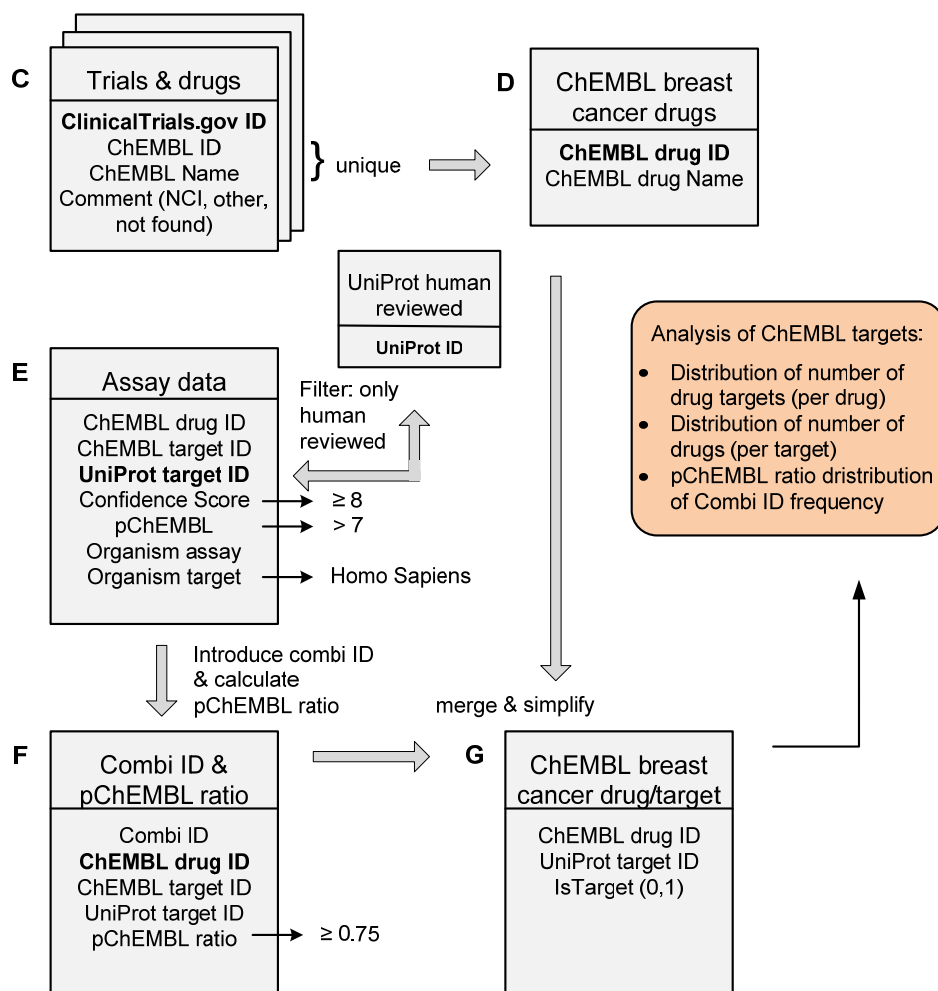


Figure 3.2: ChEMBL breast cancer target identification: Input and output tables of the ChEMBL breast cancer target identification step (grey boxes) and output of the analysis of ChEMBL targets (red box).

As described in chapter 2.3.3.1 the activity information extracted from the ChEMBL database was first pre-processed by defining a pChEMBL and confidence score threshold as well as by choosing the target organism “Homo Sapiens” and excluding non-human reviewed targets. In order to calculate the pChEMBL ratio (see chapter 2.3.3.1) all drug-target combinations were taken into account. To avoid time consuming string comparisons a unique numeric id for each drug-target combination was introduced (“combi_id”) according to the equation below.

$$combi_{id} = 10^{\text{ceil}(\log_{10}(\max(target_{id})))} * drug_{id} + target_{id} \quad (3)$$

This combi_id was then used to sort the total dataset.

As the calculation of the pChEMBL ratio requires nested loops the calculation time achieved in R was not satisfying due to the large amount of data containing many duplicates. Therefore, a C function was written to calculate the pChEMBL ratio. R can directly call C functions from a shared library (dynamic link library in Microsoft Windows).

The implemented function requires the following one-dimensional arrays as input: combi_id, pChEMBL, pChEMBL_threshold. Address references (pointers) to these elements are passed as arguments.

The function returns a reduced data set consisting of unique combi_ids together with the pChEMBL_ratio for each id.

The memory for these output arrays also has to be allocated by the caller and the address references are passed as further arguments.

R code calling C function:

```
df_int <- data.frame(drug_id = as.numeric(df$ChEMBL_drug_id), target_id =
as.numeric(df$ChEMBL_target_id), uniprot_id = df$uniprot_id, conf_score =
df$conf_score, pChEMBL = df$pChEMBL, combi_id = NA)
#identify the highest target_id:
mult_fact <- ceiling(log10(max(df_int$target_id)))
#create a unique id for each drug/target combination by multiplying the drug_id
#with mult_fact and adding the target_id:
df_int$combi_id <- (df_int$drug_id * 10^(mult_fact)) + df_int$target_id
#sort by combi_id so that same drug-target interactions are consecutive:
df_int_sort <- df_int[order(df_int$combi_id),]
```

```

#count of unique drug-target interactions (combi ids):
unique_combi <- length(unique(df_int_sort$combi_id))

cutoff <- 7 #choose pChEMBL cutoff
c_combi_id <- as.double(df_int_sort$combi_id)
c_pChEMBL <- as.double(df_int_sort$pChEMBL)
c_count_id <- as.integer(length(df_int_sort$combi_id))
#zero vector with the same length as unique combi ids:
c_ratio <- as.double(rep(0, unique_combi))

dyn.load("dll_path\\pchembl.dll") #loading dll
out <- .C("pChEMBL_ratio", c_combi_id, c_pChEMBL, c_count_id, as.double(cutoff),
c_ratio, as.integer(unique_combi)) #calling C function
str(out) #structure of C function output
#List of 6
# $ : num [1:930655] 1e+07 2e+07 2e+07 2e+07 2e+07 ...
# $ : num [1:930655] 6.38 9.49 9.29 9.49 9.49 5.8 9.07 5.39 4.9 5.67 ...
# $ : int 930655
# $ : num 7
# $ : num [1:852679] 0 1 1 1 0 1 0 0 0 1 ...
# $ : int 852679
result <- out[[5]] #calculated pChEMBL ratios

```

Listing 3.3: Code snippet of combi_id calculation and C function calling.

C function inside dll:

```

void pChEMBL_ratio (double* p_id, double* p_pCh, int* p_id_count, double*
p_cutoff, double* p_ratio, int* p_ratio_count)
{
    int count = 1;
    int count_cutoff = 0;
    int i;
    int j = 0;
    if (p_pCh[0] > *p_cutoff) {count_cutoff = count_cutoff + 1;}
    for( i = 1; i < *p_id_count; i++)
    {
        if (p_id[i] == p_id[i-1])
        {
            count = count + 1;
            if (p_pCh[i] > *p_cutoff) {count_cutoff = count_cutoff + 1;}
        }
        else
        {
            p_ratio[j] = (double)count_cutoff / (double)count;
            j = j+1;
            count = 1;
            count_cutoff = 0;
            if (p_pCh[i] > *p_cutoff) {count_cutoff = count_cutoff + 1;}
        }
    }
    p_ratio[j] = (double) count_cutoff / (double) count;
}

```

Listing 3.4: Code snippet of the C function calculating the pChEMBL ratio.

With the obtained pChEMBL ratio and the previously selected pChEMBL threshold (“cutoff”), ChEMBL targets were assessed and divided into “positive” and “negative” targets.

Apart from the ChEMBL database, also DrugBank was used for target retrieval. In Figure 3.3 the input and output tables used for the DrugBank drug-target identification are illustrated. In the first step the ChEMBL drug set was mapped to DrugBank (see the “Methods” section) resulting in Table H containing all ChEMBL breast cancer drug IDs and their respective DrugBank IDs. In the next step Table I containing all available targets in DrugBank, had to be remodelled as multiple drugs binding to the same target are recorded in the same field. A string split was therefore performed and entries of targets were duplicated. In addition, the species category was filtered by “human”. Finally the DrugBank targets could be identified by merging the resulting table with Table H by means of the DrugBank ID.

For further analysis Table K and Table L were created with data available in DrugBank. These tables contain information on drug status (approved, experimental) and pharmacological action of the target (data available from DrugBank).

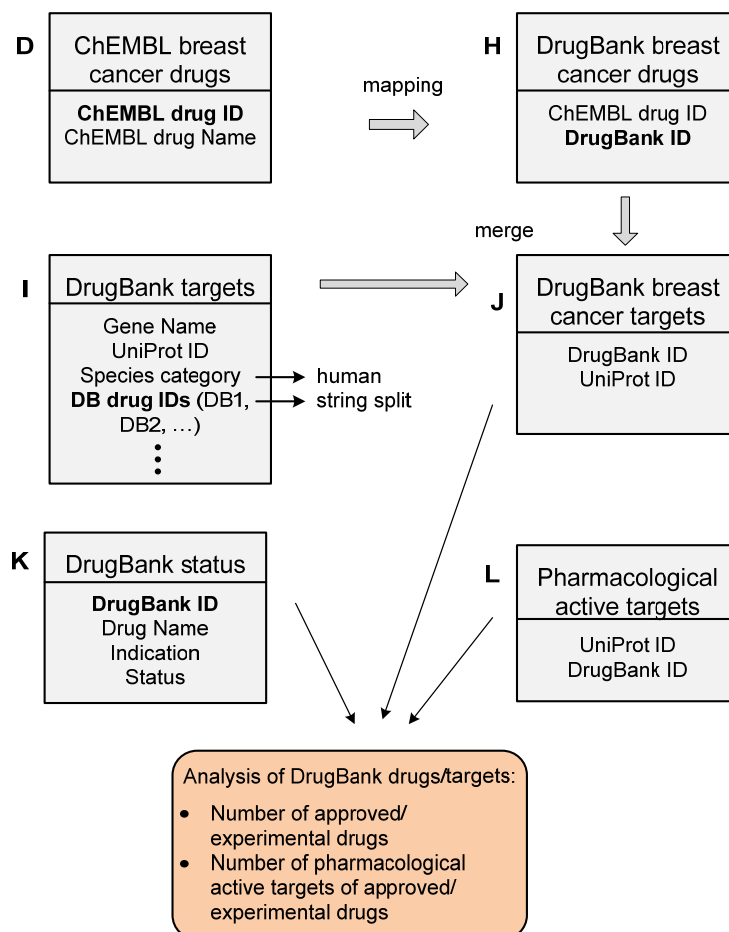


Figure 3.3: DrugBank breast cancer target identification: In- and output tables of the DrugBank breast cancer target identification step (grey boxes) and output of the analysis of DrugBank drugs/targets (red box).

In the next step the identified targets were analysed in greater detail by performing a relevance ranking (see Figure 3.4). As described earlier in the “Methods” section, information about the number of breast cancer relevant GeneRIFs, breast cancer associated papers and the sensitivity values for each gene already existed as extract from emergentec’s BIO software in the form of an Excel spreadsheet. The genes were identified by name and ENSG accession (Ensemble identifier for human genes). In order to make them compatible with the target set, the UniProt ID mapping service was used to convert ENSGs to UniProt IDs. Then the mapping result was merged with the target relevance ranking table (Table M) by means of the ENSG column. A further merging step with the identified target set (Table P) by UniProt ID was required to filter out only the relevant targets.

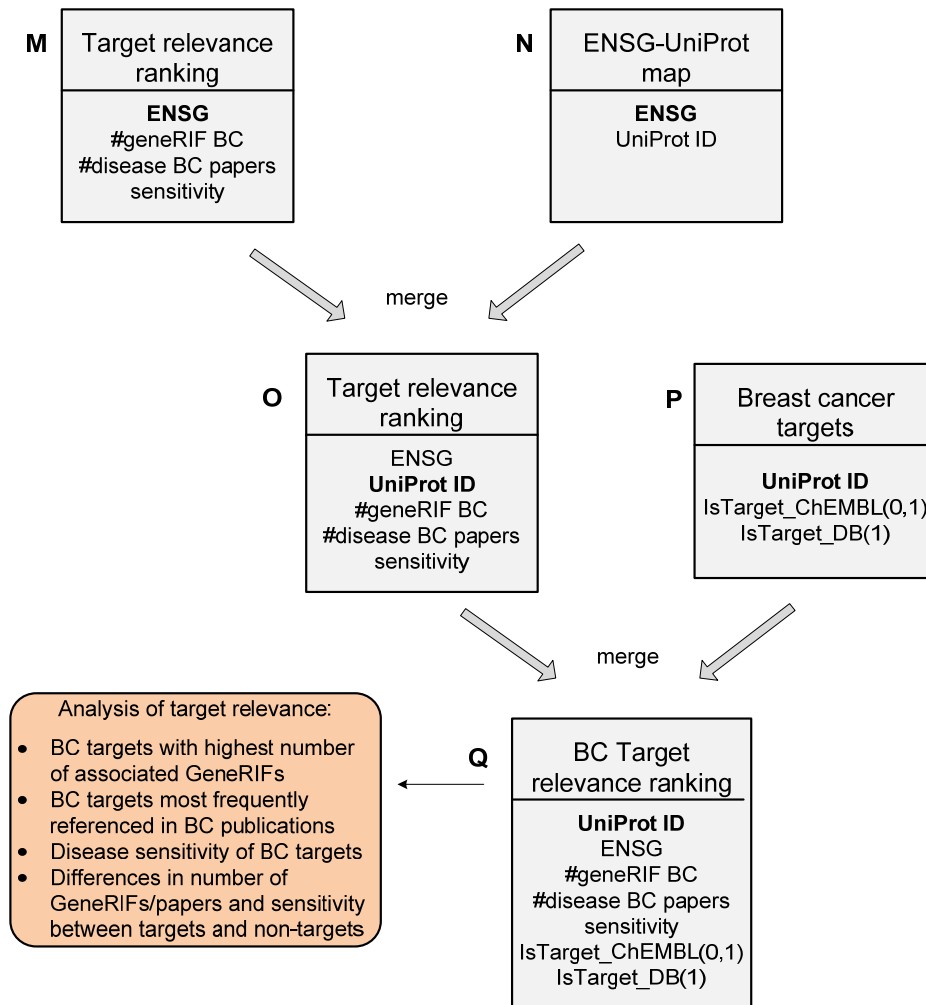


Figure 3.4: Target relevance ranking: In- and output tables of the target relevance ranking step (grey boxes) and output of the analysis of target relevance (red box).

The following step of the workflow comprised the SNP identification (see Figure 3.5). For this purpose, all breast cancer samples available in COSMIC database were extracted (for details see section 2.3.5). The indicated ENSTs were mapped to UniProt IDs making use of the UniProt ID mapper and further merged to the SNP data table (Table R).

The retrieved mutation data (Table T) was summarized by introducing the mutation ratio that was calculated per gene as follows:

$$Mutation_ratio = \frac{Number\ of\ mutated\ samples}{Total\ number\ of\ samples} \quad (4)$$

In order to remove sample data of genes which are not part of the identified target set, the COSMIC data (Table U) was further merged with the target set (Table P) by UniProt ID resulting in Table V.

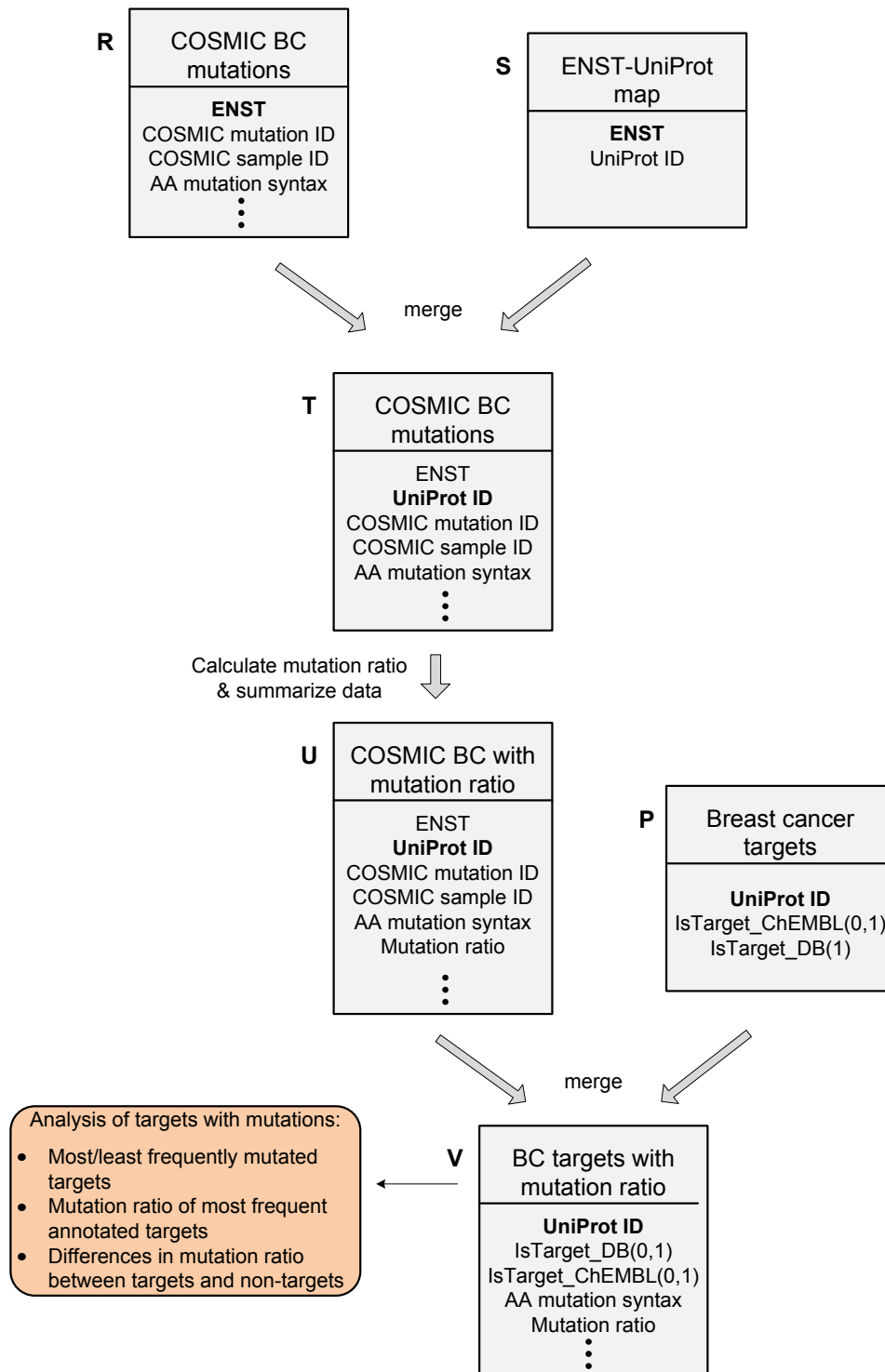


Figure 3.5: SNP identification: In- and output tables of the SNP identification step (grey boxes) and output of the analysis of targets with mutations (red box).

The final step of the workflow after the identification of SNPs located in the drug targets consisted of a plausibility check including an extensive literature search.

So far this chapter has focused on the implementation of the workflow. The following section illustrates the outcome of the previously described workflow on breast cancer as example disease term.

3.2 Workflow results

This section is divided into four main parts, each of which represents the results relating to a stage of the workflow.

3.2.1 Breast cancer trials

As described in chapter 2.3.1 all available breast cancer trials were extracted from ClinicalTrials.gov in order to generate an appropriate drug set. Therefore, pre-processing and filtering of the trial data was necessary at first, as shown in Figure 3.6.

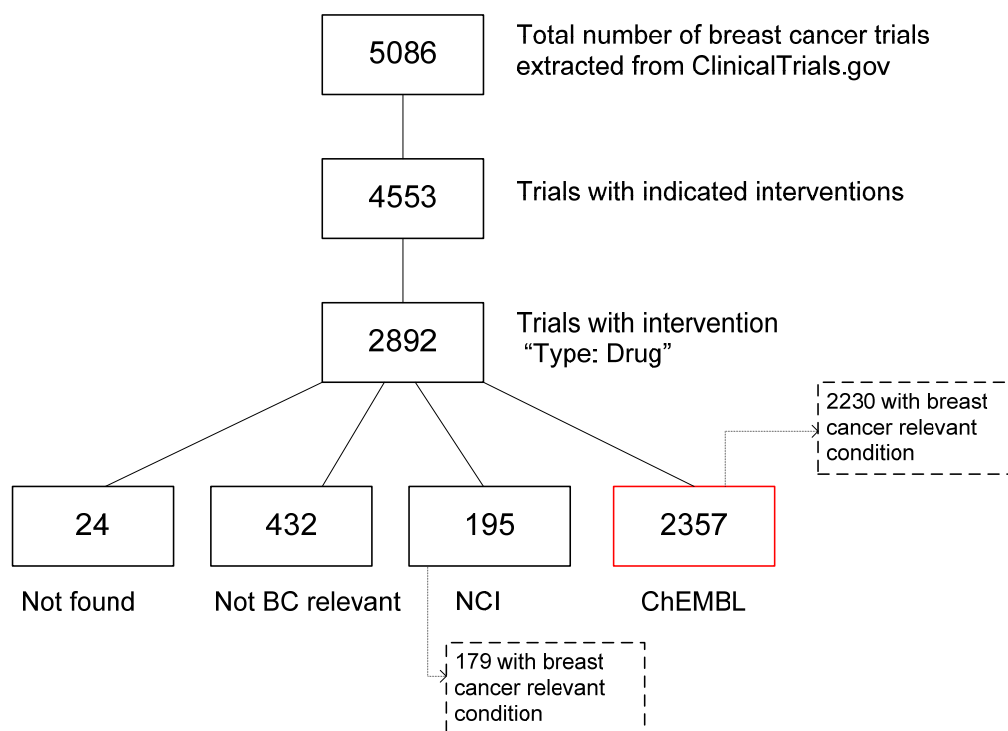


Figure 3.6: Filter process of breast cancer trials. The initial breast cancer trial set, retrieved from ClinicalTrials.gov, is reduced by selecting a distinct intervention type as well as by separating trials testing drugs not indexed in the ChEMBL database. Provided is the number of trials in the different filter steps.

Non breast cancer relevant terms and contradicting terms like “excluding breast cancer” could also be found in the condition attribute. 2230 trials of those investigating a drug indexed in ChEMBL had a breast cancer relevant condition indicated. The most common condition term is “breast cancer”, followed by “breast neoplasms” and “metastatic breast cancer”.

2,267 unique trials out of the ones with identified ChEMBL drugs showed an indicated phase and recruitment state. The following figure illustrates the number of trials per phase and recruitment state. The data set includes all breast cancer trials investigating an identified ChEMBL drug.

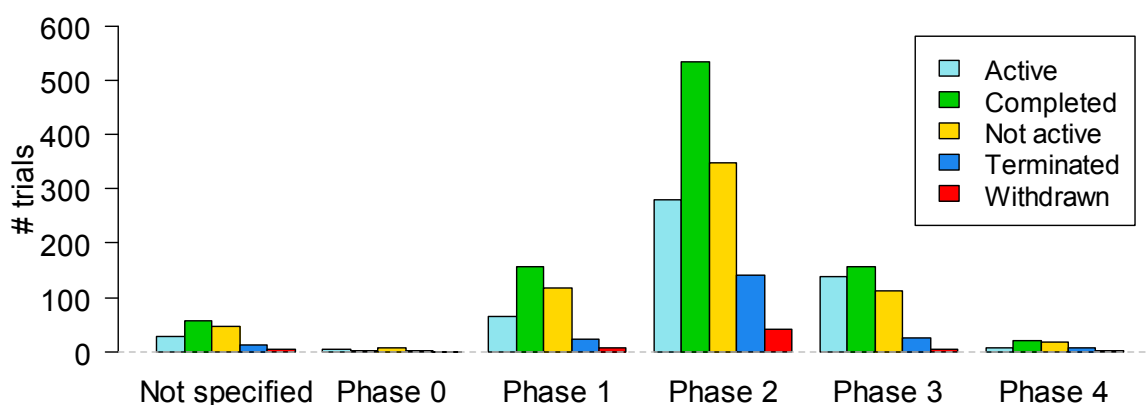


Figure 3.8: Number of trials per phase. Trials are grouped according to their clinical phase (see axis labels) and subdivided with respect to recruitment state (see legend). The group “Not specified” contains the trials without phase indication.

As can be seen in Figure 3.8 most of the trials are currently in clinical phase II of which the majority is of recruitment state “completed” (521).

Figure 3.9, illustrated below, demonstrates the distribution of drugs versus the highest development phase. If a drug was tested in multiple phases, the highest one was taken into consideration. The “total” set includes drugs referenced in ChEMBL or NCI and comprised 409 drugs. As shown in Figure 3.8 most breast cancer drugs are currently tested in clinical phase II. Furthermore it can be seen that drugs not referenced in ChEMBL but in NCI’s drug dictionary are almost exclusively in lower clinical phases (I, II).

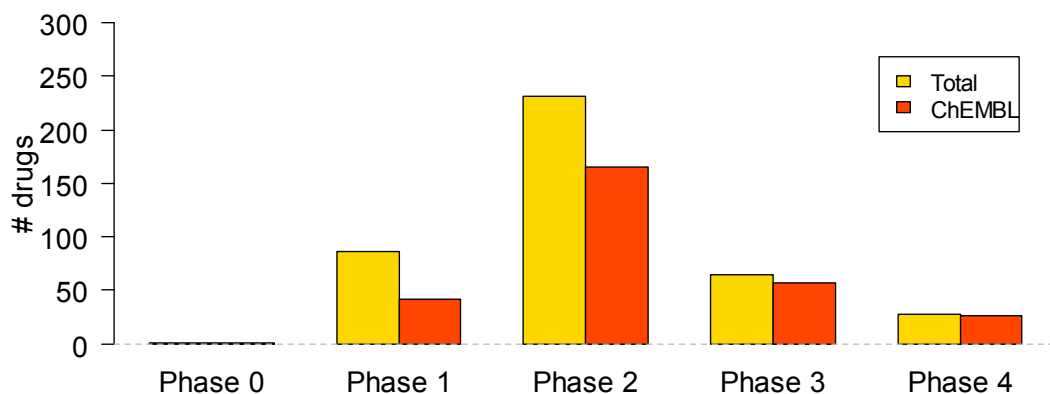


Figure 3.9: Distribution of drugs versus the highest development phase (total = NCI & ChEMBL, ChEMBL). If a drug was tested in multiple phases, the highest one was taken into consideration.

Figure 3.10 shows the distribution of the number of trials per drugs. The used data set comprised of all identified breast cancer drugs referenced in the ChEMBL database. The diagram shows that the majority of drugs are tested only in a small number of trials. Drugs that are frequently used in clinical breast cancer trials are e.g. Docetaxel, Cyclophosphamide and Paclitaxel.

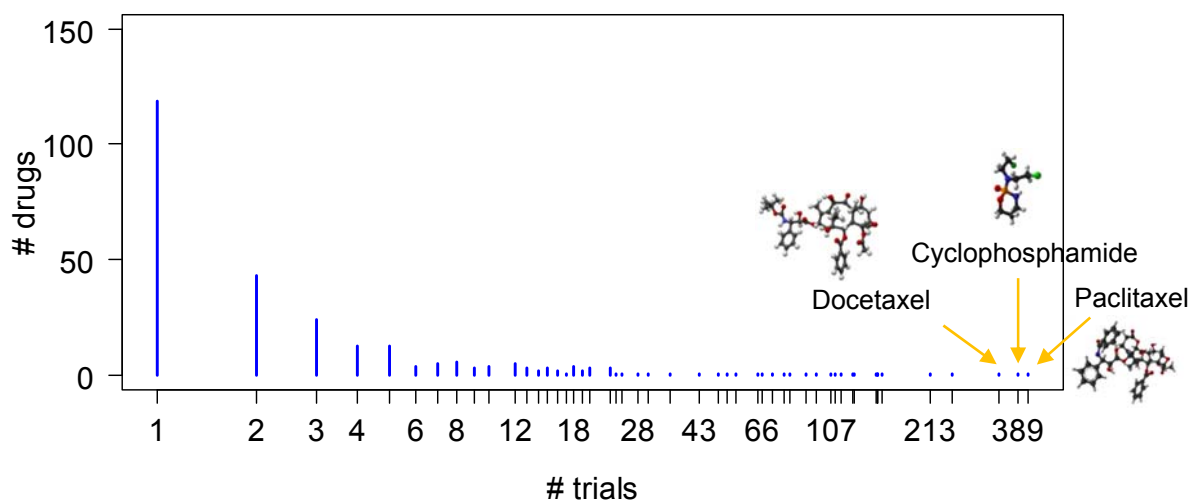


Figure 3.10: Distribution of drugs versus number of trials (used drug set: ChEMBL). The highest number of trials has been detected for Paclitaxel which is tested in 420 studies.

In summary 2,357 trials amendable for further analysis were identified leading to the generation of the drug set via ChEMBL described in the following section.

3.2.2 Breast cancer drug set and targets

This section describes the identification of a breast cancer drug set as well as a set of associated drug targets and the characteristics of these sets. The section is divided into two parts due to the databases used for drug and target retrieval, namely ChEMBL and DrugBank.

3.2.2.1 ChEMBL

The semi-automatic mapping of breast cancer drugs identified in clinical breast cancer trials to the ChEMBL database resulted in 292 unique ChEMBL drugs.

To check the reliability of the identified breast cancer drug set DailyMed (dailymed.nlm.nih.gov) was used on the one hand to determine FDA approved breast cancer drugs. On the other hand the ATC (Anatomical Therapeutic Chemical) code of the drugs was analysed and reviewed by a clinician. The table below shows the distribution of the ChEMBL drug set over ATC code groups (consisting of the first two levels only). It has to be noted that only approved drugs are classified by ATC codes.

Table 3.1: Distribution of ChEMBL drugs over ATC code groups (first two levels) of the identified ChEMBL breast cancer drug set. Drugs may occur in multiple groups.

ATC code (first two levels)	name of ATC label	number of drugs
A10	drugs used in diabetes	1
C01	cardiac therapy	1
C02	antihypertensives	1
C09	agents acting on the renin-angiotensin system	1
D11	other dermatological preparations	1
G02	other gynecologicals	1
G03	sex hormones and modulators of the genital system	7
H01	pituitary and hypothalamic hormones and analogues	3
L01	antineoplastic agents	75
L02	endocrine therapy	14
L03	immunostimulants	1
L04	immunosuppressants	6
M01	antiinflammatory and antirheumatic products	3
M02	topical products for joint and muscular pain	1
N03	antiepileptics	1
N04	anti-parkinson drugs	1
P01	antiprotozoals	1
S01	ophthalmologicals	2
V04	diagnostic agents	1

In total 19 different ATC codes are present in the used drug list which consists of 109 ChEMBL drugs. However, drugs may occur in more than one group as multiple ATC codes may be assigned to a single drug. About 90 drugs were classified into group L01 and L02 which are the categories “antineoplastic agents” and “endocrine therapy”.

In order to determine the targets of the identified ChEMBL breast cancer drug set the bioactivity data from ChEMBL was used. The initial set of bioactivity data (assay confidence score ≥ 8) comprised 1,404,385 entries. By excluding non-human targets, the dataset was reduced to 930,828 entries and, after removal of non-reviewed UniProt targets, 930,655 entries remained. These contained 852,679 unique drug-target pairs (852,799 on the level of UniProt protein accessions as target). Filtering these interactions by the ChEMBL breast cancer drug set resulted in 1805 drug-target interactions.

After evaluating the ChEMBL targets as described in the “Methods” section (see chapter 2.3.3.1) and after merging them with the selected cancer drug set 75 drugs and 177 distinct targets remained. In total 299 unique drug-target interactions were determined. The figure below shows the distribution of the number of drug targets associated to one drug. Most of the drugs affect one or two specific targets. One of the few exceptions is Dasatinib which is related to 45 targets.

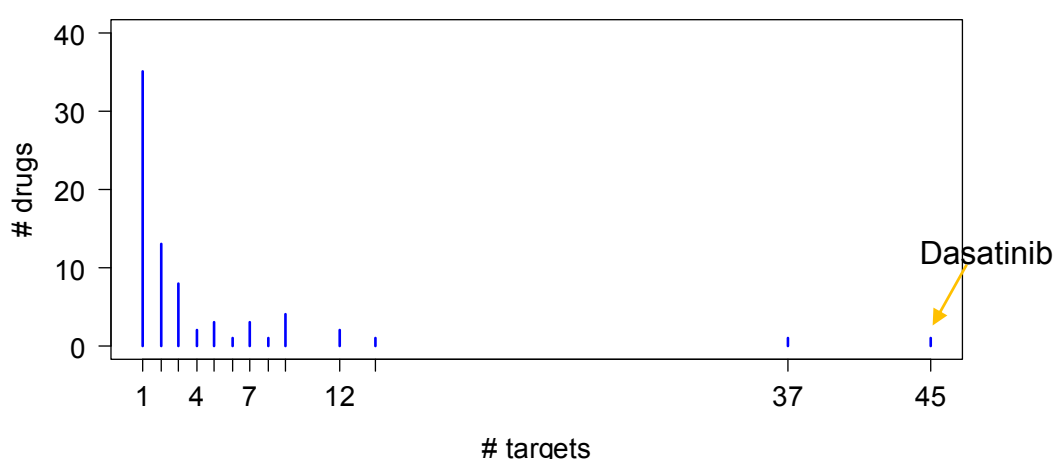


Figure 3.11: Distribution of drugs versus number of drug targets (ChEMBL). The majority of drugs are associated with a small number of targets.

In total 75 drugs and 177 targets from ChEMBL were considered for further analysis.

3.2.2.2 DrugBank

Mapping the identified ChEMBL drugs to DrugBank using the cross reference provided by ChEMBL resulted in 70 DrugBank breast cancer drugs and 169 distinct targets. After an additional manual identification a set of 141 DrugBank breast cancer drugs in total was established. 104 out of these drugs had at least one indicated target. In total 221 unique UniProt targets could be identified for the DrugBank drug set as well as 456 drug-target interactions. The number of targets associated with each drug ranges from 1 to 26. As can be seen in Figure 3.12 the majority of drugs has a small number of drug targets associated. Only a few drugs, e.g. Marimastat, affect more than ten targets.

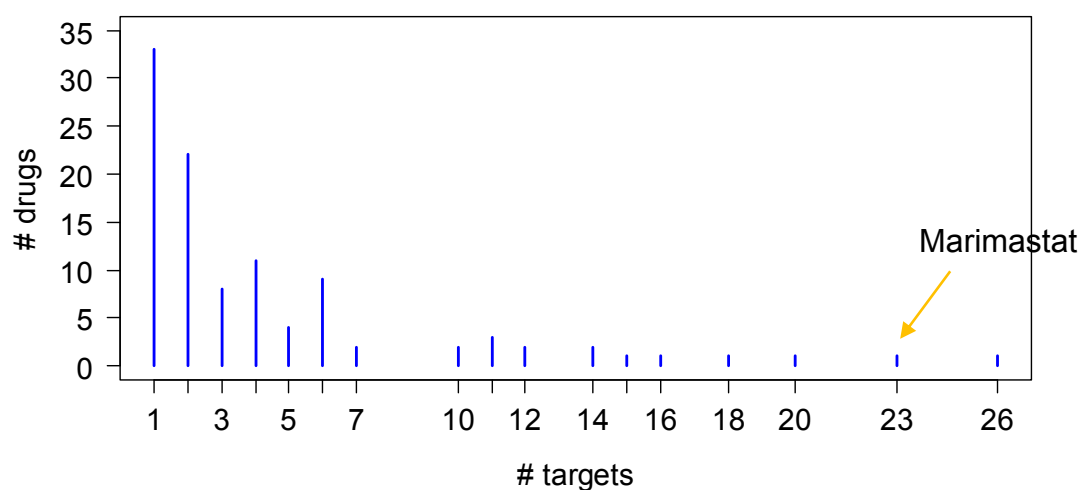


Figure 3.12: Distribution of drugs versus number of drug targets (DrugBank). Most drugs have a small amount of associated targets.

The DrugBank drug set could be further characterized by the status of the drug: approved, clinical stage and experimental (see Figure 3.13). This information as well as the pharmacological action of a drug target is indicated in DrugBank.

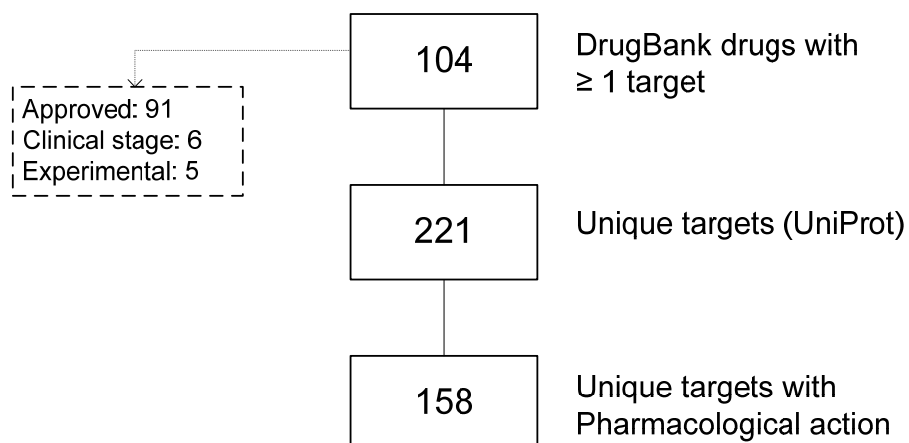


Figure 3.13: DrugBank breast cancer drug set and its targets. Provided is the number of identified DrugBank drugs (approved, in clinical stage, experimental) affecting at least one target, total number of unique targets (UniProt) and unique targets with pharmacological action.

The comparison of the ChEMBL and DrugBank final breast cancer drug-target set showed that several times the same drug was assigned to different targets in the two databases. A few examples are listed below in Table 3.2.

Table 3.2: Example of differences in target assignment via ChEMBL and DrugBank. Drugs are given as ChEMBL and DrugBank IDs, targets as UniProt IDs. Positive and negative assessed ChEMBL targets (via pChEMBL as described in the “Methods” section) are listed in separate columns.

ChEMBL ID	DrugBank ID	UniProt ID	IsTarget ChEMBL (positive)	IsTarget ChEMBL (negative)	IsTarget DrugBank
ChEMBL428647	DB01229	P27816	no	no	yes
ChEMBL428647	DB01229	O75469	no	no	yes
ChEMBL428647	DB01229	P11137	no	no	yes
ChEMBL428647	DB01229	P10636	no	no	yes
ChEMBL428647	DB01229	Q9H4B7	no	no	yes
ChEMBL428647	DB01229	P10415	no	no	yes
ChEMBL428647	DB01229	O15118	no	yes	no
ChEMBL1445	DB01185	P16471	no	no	yes
ChEMBL1445	DB01185	P04150	no	no	yes
ChEMBL1445	DB01185	P03372	no	no	yes
ChEMBL1445	DB01185	P10275	yes	no	yes
ChEMBL1445	DB01185	Q99714	no	yes	no
ChEMBL939	DB00317	P00533	yes	no	yes
ChEMBL939	DB00317	O14976	yes	no	no
ChEMBL24828	DB08764	P00374	no	no	yes
ChEMBL24828	DB08764	Q9UF33	yes	no	no
ChEMBL24828	DB08764	P06239	yes	no	no
ChEMBL24828	DB08764	P00533	yes	no	no

To check the overlap of the content of the ChEMBL and DrugBank database in general, the coverage of drugs, targets and drug-target interactions of these two databases was investigated. 3,411 of the 6,825 drugs indexed in DrugBank (version 3.0) are also present in the ChEMBL database. Compared to DrugBank, ChEMBL represents a vast repository of chemicals containing more than a million of compounds.

Figure 3.14 illustrates the Venn diagram of the target content comparison between ChEMBL and DrugBank at the UniProt identifier level. For this evaluation only targets belonging to the species category “human” were considered. All ChEMBL targets were retrieved via bioactivity data available in the ChEMBL database (retrieval date: 2013-12-23). As can be seen in the diagram below, DrugBank provides significantly more drug targets than ChEMBL. Approximately half of the targets detected in ChEMBL are annotated in DrugBank.

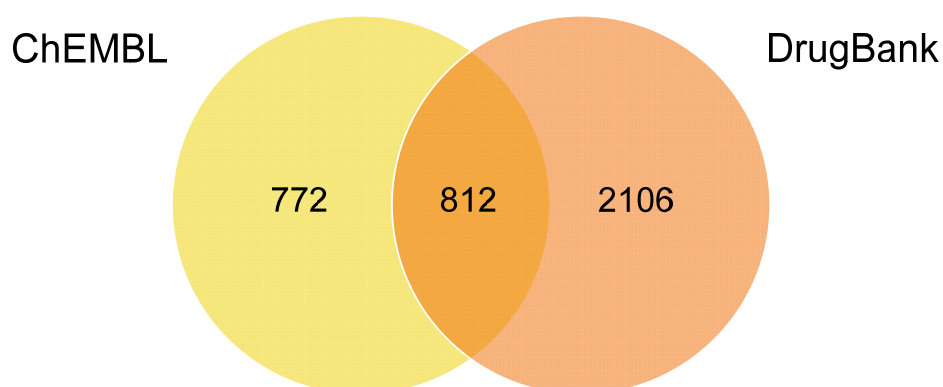


Figure 3.14: Venn diagram of target coverage. Overlap of targets available in ChEMBL (retrieval date: 2013-12-23) and DrugBank (version 3.0) for the species category “human” at the UniProt identifier level. ChEMBL targets were retrieved via ChEMBL bioactivity data.

Figure 3.15 illustrates the Venn diagram of all drug-target interactions available in ChEMBL and DrugBank. Only interactions of drugs contained in both databases were considered. Furthermore, ChEMBL targets that have been assessed as positive according to the assay evaluation procedure described in the “Methods” section were viewed separately. It has to be noted that this data solely covers targets belonging to the species category “human”. In total 1165 shared interactions between the two databases have been identified, among these 595 interactions being part of the positive ChEMBL set.

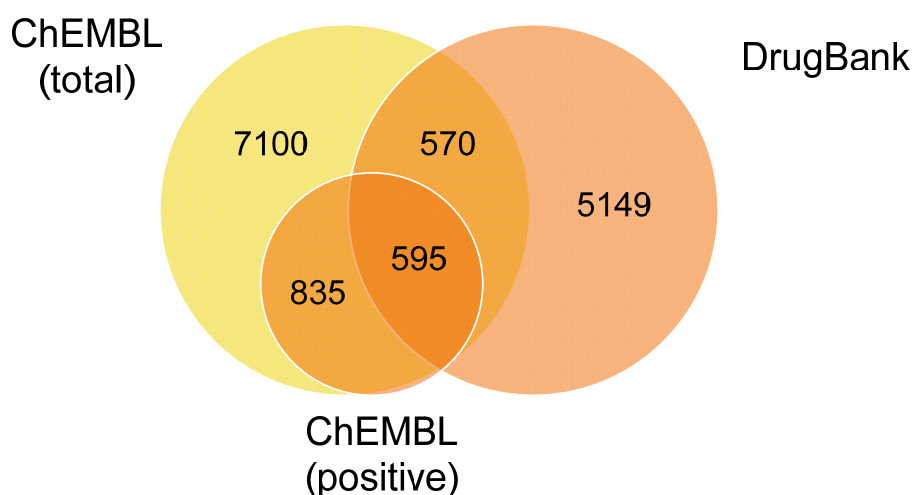


Figure 3.15: Venn diagram of drug-target interactions. Overlap of drug-target interactions in ChEMBL (total = positive + negative assessed targets, positive targets) and DrugBank for the species category “human”. Only interactions of the common ChEMBL and DrugBank drug set are included. ChEMBL (total), ChEMBL (positive) and DrugBank share 595 drug-target interactions.

The final DrugBank drug-target set comprised of 104 drugs associated to 221 unique UniProt targets. The ChEMBL drug–target set consisted of 75 drugs and 177 targets.

3.2.3 Target relevance ranking

In order to perform a target relevance ranking, the targets of the final ChEMBL and DrugBank drug set were analysed with respect to their functional disease context as described in the “Methods” section. The data used for the evaluation comprised the whole human protein coding genome, i.e. 20,288 genes. Among these, 3,033 genes were identified holding GeneRIFs extracted from publications on breast neoplasms or one of its narrower MeSH terms (aggregated mode). Figure 3.16 shows the distribution of the number of GeneRIFs (breast cancer, aggregated) in breast cancer targets and non-breast cancer targets. About thousand GeneRIFs for BRCA1 (breast cancer 1) are linked to publications on breast neoplasms. Furthermore, ERBB2 (human epidermal growth factor receptor 2), BRCA2 (breast cancer 2) and ESR1 (estrogen receptor 1) showed a high number of associations to breast cancer based on their GeneRIFs. These genes are not displayed in the boxplots below, as outliers have been excluded from the graphics.

Wilcoxon tests were applied using a p-value < 0.05 as significance level. A comparison of the target and non-target ChEMBL set revealed no significant difference. On contrary, the group of ChEMBL targets showed a significant difference compared to

the targets that are not part of the identified target set. However, in the DrugBank set higher variations between single target sets were detected. A significant difference was identified between pharmacological active and inactive targets of approved drugs as well as pharmacological inactive targets that are not part of the identified target set.

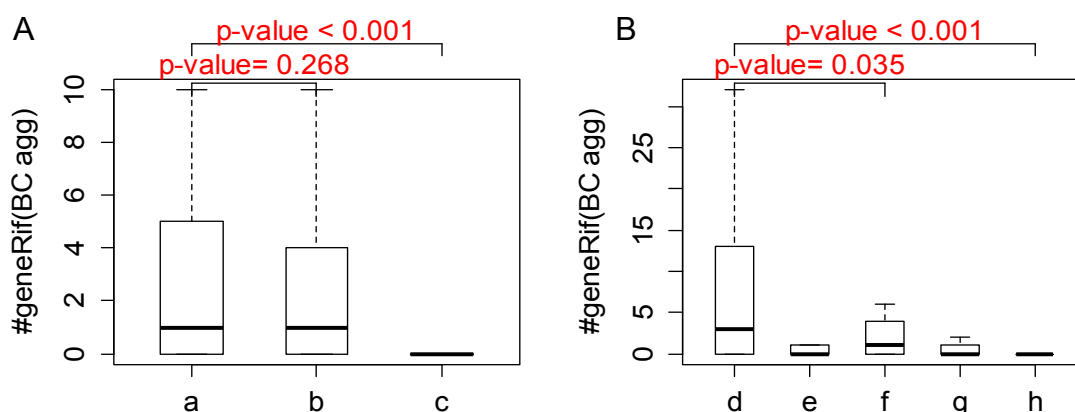


Figure 3.16: Boxplots of number of GeneRif(BC aggregated) of BC targets. (A) ChEMBL set: (a) targets (177), (b) no targets (278), (c) targets that are not part of the identified target set (18,520). (B) DrugBank set: (d) pharmacological active targets of approved drugs (153), (e) pharmacological active targets of experimental drugs (5), (f) pharmacological inactive targets of approved drugs (33), (g) pharmacological inactive targets of experimental drugs (29), (h) pharmacological inactive targets that are not part of the identified target set (18,755). Differences between the groups were assessed by Wilcoxon test using a significance level of < 0.05 .

Figure 3.17 represents the variability in the number of breast cancer papers associated with identified breast cancer targets (ChEMBL and DrugBank set) compared to non-targets. Outliers are not illustrated in the following boxplots.

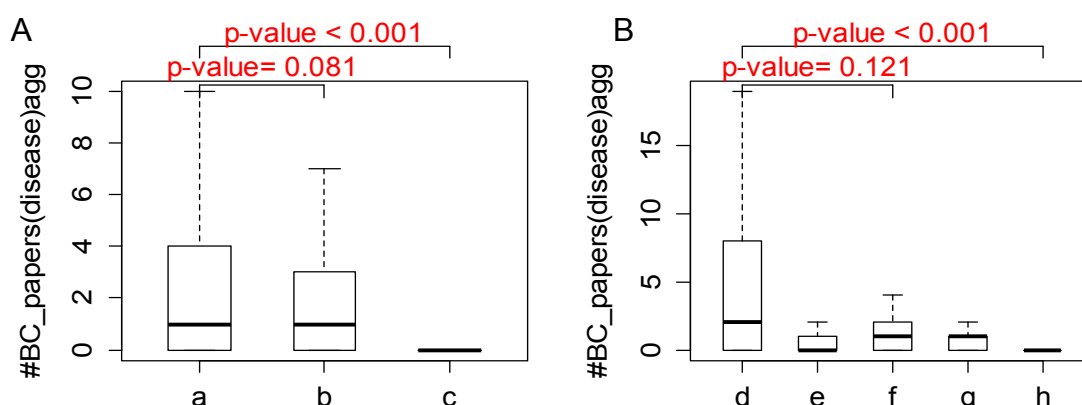


Figure 3.17: Boxplots of number of BC papers associated with BC targets (aggregated). (A) ChEMBL set: (a) targets (177), (b) no targets (278), (c) targets that are not part of the identified target set (18,520). (B) DrugBank set: (d) pharmacological active targets of approved drugs (153), (e) pharmacological active targets of experimental drugs (5), (f) pharmacological inactive targets of approved drugs (33), (g) pharmacological inactive targets of experimental drugs (29), (h) pharmacological inactive targets that are not part of the identified target set (18,755). Differences between the groups were assessed by Wilcoxon test using a significance level of < 0.05 .

In total, 2,998 genes were linked to publications which hold the selected disease term breast neoplasms or one of its child terms (aggregated mode) in the annotation. ERBB2 showed the highest number of associations (445), followed by ESR1 and BRCA1. Overall, the distribution looks similar to the boxplots based on the GeneRIF numbers. Again, a significant difference was detected between ChEMBL targets and targets not being in the identified target set as well as between pharmacological active DrugBank targets and DrugBank targets that do not belong to the identified target set.

The following table provides a list of the first ten genes with the highest number of GeneRIFs associated to breast neoplasms and the appropriate number of breast cancer papers and sensitivity.

Table 3.3: Excerpt of target relevance ranking. Listed are the first ten genes with number of GeneRIFs, number of breast cancer papers and sensitivity, ranked to the number of GeneRIFs.

UniProt ID	Gene name	#geneRif (BC agg)	#BC papers (disease agg)	sensitivity (disease)
P38398	BRCA1	1136	170	99.92
P04626	ERBB2	786	445	99.98
P51587	BRCA2	782	58	99.45
P03372	ESR1	529	331	99.95
P04637	TP53	454	155	99.88
P00533	EGFR	193	130	99.85
P15692	VEGFA	182	97	99.78
P06401	PGR	160	96	99.82
O96017	CHEK2	157	10	93.60
Q92731	ESR2	141	74	99.65

Figure 3.18 displays the disease sensitivity of the targets. In total 2,992 genes of the data set have a sensitivity value indicated. There are 287 genes that achieved a sensitivity rate higher than 90% for breast neoplasms. A significant difference in sensitivity was identified between the identified breast cancer targets and other targets. Pharmacological active and inactive targets of approved drugs from DrugBank did not show any significant difference.

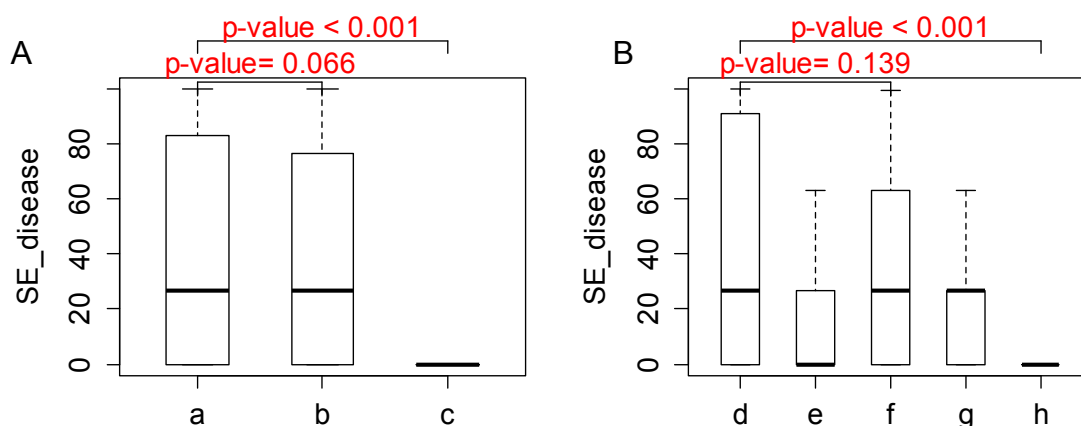


Figure 3.18: Boxplots of disease sensitivity of BC targets. (A) ChEMBL set: (a) targets (177), (b) no targets (278), (c) targets that are not part of the identified target set (18,520). (B) DrugBank set: (d) pharmacological active targets of approved drugs (153), (e) pharmacological active targets of experimental drugs (5), (f) pharmacological inactive targets of approved drugs (33), (g) pharmacological inactive targets of experimental drugs (29), (h) pharmacological inactive targets that are not part of the identified target set (18,755). Differences between the groups were assessed by Wilcoxon test using a significance level of < 0.05.

3.2.4 Drug-SNP evaluation

SNP information was retrieved via COSMIC Biomart as described in the “Methods” section. The extracted data comprised 702,531 unique entries with 34,692 unique sample IDs originating from 462 publications. 656,017 entries contained ENST accession numbers which were mapped to UniProt IDs making use of the UniProt ID mapper. For entries which did not contain ENST accession numbers and consequently could not be mapped to UniProt ID, the Swissprot ID - if indicated – was used instead. Thus, the data encompass 17,377 genes encoding human reviewed UniProt targets that have been tested for mutations. In 14,039 genes SNPs were detected. Overall 70,277 unique mutations (COSMIC mutation IDs) had been identified in breast carcinoma samples.

In the COSMIC data 170 out of the 177 identified ChEMBL targets as well as 207 out of the 221 DrugBank targets were represented. 157 ChEMBL and 181 DrugBank targets had mutated samples assigned in COSMIC respectively.

Figure 3.19 demonstrates the differences in the mutation ratio between targets and non-targets of the identified breast cancer drug set (ChEMBL, DrugBank).

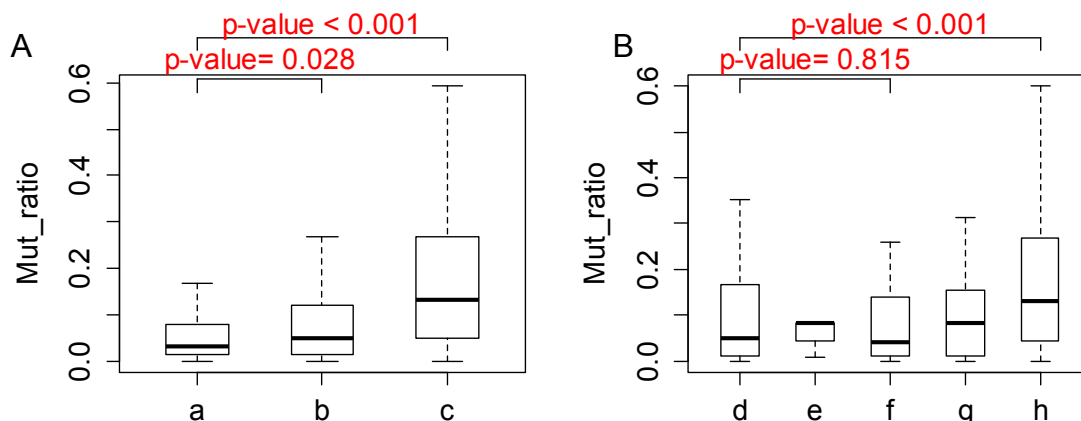


Figure 3.19: Boxplots of the mutation ratio of BC targets (sample size > 20). (A) ChEMBL set: (a) targets (total: 177, NA: 29), (b) no targets (total: 278, NA: 38), (c) targets that are not part of the identified target set (total: 18,520, NA: 5,293). (B) DrugBank set: (d) pharmacological active targets of approved drugs (total: 153, NA: 31), (e) pharmacological active targets of experimental drugs (total: 5), (f) pharmacological inactive targets of approved drugs (total: 33, NA: 2), (g) pharmacological inactive targets of experimental drugs (total: 29, NA: 1), (h) pharmacological inactive targets that are not part of the identified target set (total: 18,755, NA: 5,326). Differences between the groups were assessed by Wilcoxon test using a significance level of < 0.05.

Gene encoding protein targets belonging to the identified target set from ChEMBL as well as from DrugBank show a significantly lower mutation ratio than other targets. Furthermore, there is a significant difference in mutation ratio between ChEMBL positive and negative targets (according to assay data evaluation).

The most prevalent gene mutations in the COSMIC data are c.3140A>G and c.1633G>A occurring on gene PIK3CA. This gene shows an overall mutation ratio of 0.275. The maximal possible mutation ratio is 1 which means that mutations are found in all samples tested. Only a few protein targets included in the breast cancer tissue sample data extracted from COSMIC exhibit a mutation ratio of 1. Table 3.4 provides the mutation ratios of the five most and least mutated genes being part of the target set of both ChEMBL and DrugBank. The highest mutation ratio found is 0.33 occurring in HDAC6.

Table 3.4: Example of mutated genes. Tabular listing of the five most and least mutated genes included in both ChEMBL and DrugBank target set sorted by their mutation ratio in descending order. Only genes with a sample size > 20 are included.

UniProt ID	Gene name	Mutation ratio
Q9UBN7	HDAC6	0.333
P14780	MMP9	0.2
P45452	MMP13	0.167
Q9BY41	HDAC8	0.167
P22894	MMP8	0.136
P30872	SSTR1	0.005
P12931	SRC	0.004
P00374	DHFR	0
P30968	GNRHR	0
P50750	CDK9	0

The following table summarizes the mutation ratios of the most frequently annotated targets of the common ChEMBL and DrugBank breast cancer target set.

Table 3.5: Most commonly annotated mutated genes. Tabular listing of the mutation ratios of the most frequently referenced genes being part of the common BC target set of ChEMBL and DrugBank. Mutation ratios flagged with * are calculated on a basis of less than 21 different samples.

UniProt ID	Gene name	#BC papers (disease agg)	Mutation ratio
P04626	ERBB2	445	0.022
P03372	ESR1	331	0.019
P00533	EGFR	130	0.013
P31749	AKT1	81	0.031
Q92731	ESR2	74	*0.214

4 Discussion

The main goal of this thesis was to develop a computational workflow for evaluating drug efficacy in the context of mutation status of drug targets. The workflow was exemplarily applied to breast cancer drugs and their targets.

4.1 Workflow construction

The workflow consisted of the following steps, namely (i) the identification of clinical trials on breast cancer, (ii) the identification of investigated drugs in these trials, (iii) the extraction of drug targets of these drugs, (iv) the identification of prevalent SNPs in these targets, and (v) the evaluation of functional association of the targets to breast cancer based on scientific literature. Prediction of drug binding domains proved to be difficult and could therefore not be completed in time. Consequently, SNP integration was performed on the level of genes.

While major parts of the workflow could be fully automated making use of R scripting supported by a function implemented in C to speed up processing, certain parts of the workflow required manual interaction. Mapping of drug names extracted from clinical trials to common drug namespaces for example was optimized by manual inspection in order to reduce the rate of false positives.

Basically, it is thus possible to adapt the workflow to investigate other diseases besides breast cancer, however, requiring in addition to the modification of the filter conditions manual curation of certain mapping steps to fully exploit the potential of the whole workflow.

In addition, it has to be mentioned that the interface of the web data sources and their structure is undergoing a permanent change. ChEMBL for instance has changed from version 17 to version 18 since start of this thesis. To cope with such changing interfaces the import sections of the script needs to be modified accordingly.

Nevertheless, the required amount of work to adapt the existing workflow to changing interfaces and/or different disease terms is considered to be manageable.

4.2 Main findings

In this thesis the workflow was implemented using the example of breast cancer. As shown in the “Results” section, a breast cancer drug set could be identified by obtaining the required data from ClinicalTrials.gov as well as the ChEMBL database. Associated drug targets were detected by making use of information provided in ChEMBL and DrugBank. The performed drug-SNP evaluation was based on SNPs that occurred in breast cancer tissue samples extracted from COSMIC. Interestingly, the targets of the identified breast cancer drugs reveal a significant lower mutation ratio than other targets (p -value < 0.001). This is basically valid for targets derived from ChEMBL as well as for targets from DrugBank. Of note, ChEMBL breast cancer targets that have been assessed as positive according to the assay data evaluation also show significant lower mutation ratios in comparison to those assessed as negative (p -value = 0.028). Surprisingly, no statistical difference between pharmacological active and inactive targets of approved DrugBank breast cancer drugs can be determined. In addition, targets that are frequently referenced in breast cancer papers typically reveal rather small mutation ratios.

As expected, both of the breast cancer target sets from ChEMBL and DrugBank show a higher number of associated BC papers and GeneRIFs than those targets which are not in the identified sets. A significant difference can be seen between pharmacological active and inactive targets from DrugBank.

To assess the SNPs extracted from COSMIC with respect to therapy outcome of the treatment an extensive online search was performed. An example is L858R, a mutation of the epidermal growth factor receptor (EGFR). This target was ranked in the sixth place of the performed target relevance ranking with 193 breast cancer associated GeneRIFs and 130 linked publications. Its mutation L858R occurs in 12 different samples extracted from COSMIC and is reported in several studies. The study by Rajasekaran *et al.* [46] focused on in-silico identification of significant detrimental missense mutations of EGFR and their effect with 4-anilinoquinazoline-based drugs. In the study the three major inhibitors of this drug class, erlotinib, gefitinib and lapatinib were examined which are all used in the treatment of breast cancer. The atomic contact energy between EGFR and its inhibitors was computed in order to evaluate the changes in binding efficiency of the drugs related to mutations of EGFR. They were

able to show that lapatinib establishes maximum binding affinity with native EGFR as well as mutants except for the mutations R748P and L858R. Consequently they suggest that a combination of these drugs might improve the outcome of the cancer treatment. Further groups studied the efficacy of lapatinib against HER2 (epidermal growth factor receptor 2, ERBB2) variants, as lapatinib is a dual inhibitor of EGFR and HER2 tyrosine kinases. HER2 is one of most relevant targets in the performed relevance ranking with 445 associated breast cancer papers showing a mutation ratio of 0.022. The group of Kancha *et al.* and Trowe *et al.* identified several potential mutations in HER2 kinase domain linked to lapatinib resistance [47, 48].

However, SNPs in drug targets may also have benefits on therapeutic intervention. The group of Colomer *et al.* investigated three different aromatase CYP19 gene variants and their effect on antiaromatase therapy in postmenopausal women with breast cancer. They identified one SNP variant (rs4646) that was associated with improved treatment efficacy when treating patients with hormone receptor-positive advanced breast carcinoma with the aromatase inhibitor letrozole [49].

As shown in the “Results” section, the most common gene mutations found in COSMIC data are harboured in the oncogene PIK3CA. Its mutations may result in an activation of the PI3K (Phosphatidylinositol 3-kinase) pathway, which mediates cell proliferation and survival, and moreover confers increased resistance to some well-known cancer drugs, e.g. paclitaxel [50]. Several research groups demonstrated that somatic mutations in PIK3CA occur at high frequency in breast cancer tissue [51, 52] but with varying occurrence in different subtypes [53]. Furthermore, the vast majority of PIK3CA mutations are located at three hotspots thus representing ideal targets for therapeutic development [50, 54].

The target with the highest mutation ratio of 0.33 in the common ChEMBL and DrugBank breast cancer target set is HDAC6 (histone deacetylase 6). According to the relevance ranking HDAC6 does not belong to the top targets (ranked on place 434), but still reveals seven breast cancer associated GeneRIFs as well as ten linked publications. So far one approved drug inhibiting this protein and thus blocking the cell proliferation is on the market, called Vorinostat. This drug is used in the treatment of skin cancer, whereas several studies are currently ongoing to investigate the drug’s applicability in breast cancer therapy as well.

4.3 Assessment of data sources

- Breast cancer trials

Although FDA requires that “basic” results of clinical trials are reported in ClinicalTrials.gov, still not all of the studies conducted in the U.S. are listed. The number is increasing, though, as registration requirements are expanded.

The majority of investigated breast cancer trials are currently in clinical phase II mostly having the recruitment state “completed” (see Figure 3.8). However, only drugs of trials that are completed either in phase III or in phase IV can be assumed as approved. This leads to the conclusion that phase II constitutes a substantial obstacle for drugs under approval and the drop-out rate rises significantly during this phase. This issue was already discussed in several publications. The group of Gupta *et al.*, for example, points out that in 2006-2007 the success rates for Phase II studies was only at 28% whereas in 2008-2009 it even fell to 18% [5]. Arrowsmith *et al.* propose that these low Phase II success rates in combination with decreasing Phase III failure rates may be an indication that the pharmaceutical industry is redesigning their Phase II programmes in order to be able to decide earlier on termination and thereby potentially reduce the costs of Phase III failures [10].

Concerning the conditions indicated in each clinical trial entry, the analysis showed that apart from the selected disease term “breast cancer” also other cancer types showed up in the condition attribute e.g. lung cancer and colorectal cancer. This fact is due to the possibility of multiple conditions being linked to one clinical study, as certain cancer drugs are evaluated in the treatment of more than one specific cancer type.

As shown in the “Results” section (Figure 3.10), well-established drugs as e.g. Paclitaxel are used in a large number of clinical trials. This is understandable since they often serve as basis for comparison for novel drugs being still in development.

- Breast cancer drugs and targets

The two publicly accessible biological databases ChEMBL and DrugBank were used for retrieval of drug and drug target information. Both provide approved as well as experimental drugs. Nevertheless, ChEMBL encompasses a larger proportion of early

stage drugs. The ChEMBL drug set identified in this work comprised 292 breast cancer drugs in total, whereas only 142 of them have an equivalent in DrugBank. DrugBank provides drug-targets for all FDA-approved drugs. The ChEMBL database includes drug mechanism of action information for all FDA-approved drugs as well. In addition, to identify further targets, especially those for not yet approved drugs, bioactivity data can be retrieved from ChEMBL. However, the usage of the bioactivity data, which comprises pChEMBL values and confidence scores as key values, poses a challenge as it is difficult to decide on a proper separation algorithm between positive and negative targets. In this thesis drug-target interactions with a confidence score ≥ 8 and pChEMBL ≥ 7 were considered as positive. By making use of the bioactivity data from ChEMBL for determining the associated targets the drug set was reduced to less than half of the initial set. This was mainly due to the limited availability of bioactivity data in ChEMBL as well as the stringent pChEMBL cutoff and the pChEMBL ratio values used.

A comparison between the ChEMBL and DrugBank final breast cancer target set revealed that several times the same drug was assigned to different targets in the two databases (see Table 3.2).

Furthermore, to assess the conformity between drug target information provided in ChEMBL and DrugBank in general, the information of both data sources was contrasted with each other. Surprisingly, only a minor overlap between ChEMBL and DrugBank drug target interaction data was detected (see Figure 3.15 for further details). Among the 9,100 (positive and negative) interactions from ChEMBL and 6,314 interactions from DrugBank only 1,165 shared drug-target interactions have been identified. Restricting the ChEMBL drug-target interaction set to only potentially positive interactions (pChEMBL value > 7 and pChEMBL-ratio ≥ 0.75) reduced the number of common interactions further down to 595. This small overlap in drug-target interactions results from the already low agreement between drugs available in ChEMBL and DrugBank. The difference in the chemical content of these two databases is probably due to diverging data capture strategies [55]. Furthermore, also the target coverage between ChEMBL and DrugBank seems quite low (see Figure 3.14). In DrugBank, far more targets are represented than really identified via ChEMBL assay data. Consequently, the low overlap in drug-target interactions is not as surprising as it seems at first glance. The group of Tiikkainen *et al.* analysed

commercial and publicly available bioactivity databases, including the largest data supplier ChEMBL, and estimated their error rates [56]. Typically, these resources are manually curated from scientific publications and patents. Thus, apart from errors in the source document, the human factor during the extraction process may lead to data inconsistency between databases. Tiikkainen *et al.* reported common data discrepancies between the vendors when comparing data derived from the same article. Furthermore, they showed a minor overlap between bioactivity databases as well as a great number of unique data contained in each resource [56, 57].

- SNP evaluation

Data available in the COSMIC database have to be interpreted with caution. One issue is the variability in the resolution for mutations when screening genes [58]. Some genes are characterized by so called mutation hot spots. These are gene regions where mutations are most likely to occur. As a consequence such genes are especially screened in this genomic area and mutations located in other regions of the gene might not be detected. This strategy might therefore lead to an underestimation of the frequency of mutations and moreover to a distorted view of the distribution of genetic mutations. Another possible curational bias in COSMIC data originates from sample screenings that are reported twice in two different publications. Although these samples might have the same name, it is impossible to confirm equality of the data. Thus, multiple entries with different COSMIC sample IDs might exist for an identical sample. To account for this fact, multiple occurrences of the same sample name being assigned to different COSMIC IDs had to be checked. The result showed that sometimes identical sample names existed with different sample IDs. As the frequency of duplicated sample names was low, this potential bias was neglected.

Another issue concerning COSMIC data are missing UniProt IDs. It is possible to download COSMIC samples with additional Swissprot IDs but actually they are not indicated for every entry. Even well-known breast cancer related genes such as BRCA1, BRCA2 and ESR1 are not assigned to UniProt IDs in the data extracted from COSMIC Biomart. Consequently, mapping of accession numbers via UniProt ID mapper is required as well as further manual data integration work.

Furthermore, it has to be noted that targets with a sample size smaller than 20 were not considered in the analysis although among these are well-known breast cancer targets as e.g. ESR2.

It should also be mentioned that the calculated mutation ratio could only be based on genes but not on individual SNPs. As a consequence a statement of the prevalence of SNPs could not be derived.

4.4 Outlook

In this work approved and experimental drugs for breast cancer treatment have been investigated irrespectively of the breast cancer subtype. For instance, one drug of the identified set, the experimental drug Olaparib, also known as AZD-2281, was designed particularly for patients with BRCA1/2 mutations as it acts as a PARP-1 (Poly ADP ribose polymerase 1) inhibitor [59]. Several breast cancer studies showed that molecularly distinct subtypes of breast cancer may respond differently to pathway-targeted therapies. Docetaxel and Cisplatin for example revealed preferential activity in basal or claudin-low subtypes [60]. Thus, a possible further refinement of the present work could be to focus on a specific type of breast cancer.

Within the scope of this thesis all SNPs occurring on the target gene affected by one of the identified breast cancer drugs were taken into consideration. To determine only the SNPs located exactly in the binding domain of the drug, further work would be required which includes drug-domain prediction as well as SNP-domain integration.

In summary, personalized/stratified medication in cancer therapy has gained more and more importance over the years as the inter-individual variation on the genomic and molecular level influences drug efficacy. Knowledge on the mutation status of drug targets is an important piece of information in the selection of appropriate drug targets as well as in the design of clinical trials.

Literature

1. Ferlay J, Steliarova-Foucher E, *et al.*: **Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012.** *Eur J Cancer* 2013, **49**:1374-1403.
2. De Abreu FB, Schwartz GN, *et al.*: **Personalized Therapy for Breast Cancer.** *Clin Genet* 2014, in press, doi: 10.1111/cge.12381.
3. DiMasi JA, Feldman L, *et al.*: **Trends in risks associated with new drug development: success rates for investigational drugs.** *Clin Pharmacol Ther* 2010, **87**:272-277.
4. Pammolli F, Magazzini L, *et al.*: **The productivity crisis in pharmaceutical R&D.** *Nat Rev Drug Discov* 2011, **10**:428-438.
5. Gupta SC, Sung B, *et al.*: **Cancer drug discovery by repurposing: teaching new tricks to old dogs.** *Trends Pharmacol Sci* 2013, **34**:508-517.
6. Hay M, Thomas DW, *et al.*: **Clinical development success rates for investigational drugs.** *Nat Biotechnol* 2014, **32**:40-51.
7. Kamb A, Wee S, *et al.*: **Why is cancer drug discovery so difficult?** *Nat Rev Drug Discov* 2007, **6**:115-120.
8. Kola I, Landis J: **Can the pharmaceutical industry reduce attrition rates?** *Nat Rev Drug Discov* 2004, **3**:711-715.
9. Hutchinson L, Kirk R: **High drug attrition rates--where are we going wrong?** *Nat Rev Clin Oncol* 2011, **8**:189-190.
10. Arrowsmith J, Miller P: **Trial watch: phase II and phase III attrition rates 2011-2012.** *Nat Rev Drug Discov* 2013, **12**:569.
11. Kenakin TP: *Pharmacology in Drug Discovery: Understanding Drug Response*, 1 edn. Waltham, MA, USA: Academic Press; 2012.
12. Evans WE, Johnson JA: **Pharmacogenomics: the inherited basis for interindividual differences in drug response.** *Annu Rev Genomics Hum Genet* 2001, **2**:9-39.
13. McLeod HL, Yu J: **Cancer pharmacogenomics: SNPs, chips, and the individual patient.** *Cancer Invest* 2003, **21**:630-640.
14. Semizarov D, Blomme E: *Genomics in Drug Discovery and Development*, 1 edn. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2009.

15. Wiechec E, Hansen LL: **The effect of genetic variability on drug response in conventional breast cancer treatment.** *Eur J Pharmacol* 2009, **625**:122-130.
16. Mah JTL, Low ESH, *et al.*: **In silico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery.** *Drug Discov Today* 2011, **16**:800-809.
17. Ferrer-Soler L, Vazquez-Martin A, *et al.*: **An update of the mechanisms of resistance to EGFR-tyrosine kinase inhibitors in breast cancer: Gefitinib (Iressa) -induced changes in the expression and nucleo-cytoplasmic trafficking of HER-ligands (Review).** *Int J Mol Med* 2007, **20**:3-10.
18. Dua S, Chowriappa P: *Data Mining for Bioinformatics*, 1 edn. Boca Raton, FL, USA: CRC Press; 2012.
19. Jazwinska EC: **Exploiting human genetic variation in drug discovery and development.** *Drug Discov Today* 2001, **6**:198-205.
20. Voisey J, Morris CP: **SNP Technologies for Drug Discovery: A Current Review.** *Curr Drug Discov Technol* 2008, **5**:230-235.
21. Kwok P-Y: **Methods for genotyping single nucleotide polymorphisms.** *Annu Rev Genomics Hum Genet* 2001, **2**:235-258.
22. Kwok P-Y, Chen X: **Detection of single nucleotide polymorphisms.** *Curr Issues Mol Biol* 2003, **5**:43-60.
23. Flicek P, Amode MR, *et al.*: **Ensembl 2014.** *Nucleic Acids Res* 2014, **42**:D749-755.
24. Benson DA, Cavanaugh M, *et al.*: **GenBank.** *Nucleic Acids Res* 2013, **41**:D36-42.
25. Teufel A, Krupp M, *et al.*: **Current bioinformatics tools in genomic biomedical research (Review).** *Int J Mol Med* 2006, **17**:967-973.
26. Sherry ST, Ward MH, *et al.*: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**:308-311.
27. Perkel J: **SNP genotyping: six technologies that keyed a revolution.** *Nat Methods* 2008, **5**:447-453.
28. Hamosh A, Scott AF, *et al.*: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**:D514-517.

29. Forbes SA, Tang G, *et al.*: **COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer.** *Nucleic Acids Res* 2010, **38**:D652-657.
30. Fernald GH, Capriotti E, *et al.*: **Bioinformatics challenges for personalized medicine.** *Bioinformatics* 2011, **27**:1741-1748.
31. Goble C, Stevens R: **State of the nation in data integration for bioinformatics.** *J Biomed Inform* 2008, **41**:687-693.
32. **ClinicalTrials.gov.** [<http://clinicaltrials.gov>]. 2013.
33. Gaulton A, Bellis LJ, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40**:D1100-1107.
34. Bento AP, Gaulton A, *et al.*: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res* 2014, **42**:D1083-1090.
35. Knox C, Law V, *et al.*: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39**:D1035-1041.
36. **DailyMed.** [<http://dailymed.nlm.nih.gov>]. 2014.
37. The UniProt Consortium: **Activities at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2014, **42**:D191-198.
38. Shepherd R, Forbes SA, *et al.*: **Data mining using the Catalogue of Somatic Mutations in Cancer BioMart.** *Database* 2011, **2011**:bar018.
39. WHO Collaborating Centre for Drug Statistics Methodology: *Guidelines for ATC classification and DDD assignment, 2014*, 17 edn. Oslo, Norway: Norwegian Institute of Public Health; 2013.
40. R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2013, [<http://www.R-project.org/>]. [ISBN 3-900051-07-0].
41. Allaire JJ: *RStudio: Integrated development environment for R.* RStudio, Boston, MA 2013, [<https://www.rstudio.com/>].
42. Kernighan BW, Ritchie DM: *The C programming Language*, 1 edn. Englewood Cliffs, NJ, USA: Prentice-Hall; 1978.
43. **Code::Blocks.** [<http://www.codeblocks.org/>]. 2013.
44. Jupp S, Malone J, *et al.*: **The EBI RDF platform: linked open data for the life sciences.** *Bioinformatics* 2014, in press, doi: 10.1093/bioinformatics/btt765.
45. Rogers FB: **Medical subject headings.** *Bull Med Libr Assoc* 1963, **51**:114-116.

46. Rajasekaran R, Sethumadhavan R: **In silico identification of significant detrimental missense mutations of EGFR and their effect with 4-anilinoquinazoline-based drugs.** *Appl Biochem Biotechnol* 2010, **160**:1723-1733.
47. Kancha RK, Bubnoff N, *et al.*: **Differential sensitivity of ERBB2 kinase domain mutations towards lapatinib.** *PLoS One* 2011, **6**:e26760.
48. Trowe T, Boukouvala S, *et al.*: **EXEL-7647 inhibits mutant forms of ErbB2 associated with lapatinib resistance and neoplastic transformation.** *Clin Cancer Res* 2008, **14**:2465-2475.
49. Colomer R, Monzo M, *et al.*: **A single-nucleotide polymorphism in the aromatase gene is associated with the efficacy of the aromatase inhibitor letrozole in advanced breast carcinoma.** *Clin Cancer Res* 2008, **14**:811-816.
50. Isakoff SJ, Engelman JA, *et al.*: **Breast cancer-associated PIK3CA mutations are oncogenic in mammary epithelial cells.** *Cancer Res* 2005, **65**:10992-11000.
51. Bachman KE, Argani P, *et al.*: **The PIK3CA gene is mutated with high frequency in human breast cancers.** *Cancer Biol Ther* 2004, **3**:772-775.
52. Cizkova M, Susini A, *et al.*: **PIK3CA mutation impact on survival in breast cancer patients and in ER α , PR and ERBB2-based subgroups.** *Breast Cancer Res* 2012, **14**:R28.
53. Stemke-Hale K, Gonzalez-Angulo AM, *et al.*: **An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer.** *Cancer Res* 2008, **68**:6084-6091.
54. Gustin JP, Cosgrove DP, *et al.*: **The PIK3CA gene as a mutated target for cancer therapy.** *Curr Cancer Drug Targets* 2008, **8**:733-740.
55. Muresan S, Sitzmann M, *et al.*: **Mapping Between Databases of Compounds and Protein Targets.** In *Bioinformatics and Drug Discovery*. Edited by Larson RS. Totowa, NJ, USA: Humana Press; 2012: 145-164.
56. Tiikkainen P, Bellis L, *et al.*: **Estimating error rates in bioactivity databases.** *J Chem Inf Model* 2013, **53**:2499-2505.
57. Tiikkainen P, Franke L: **Analysis of commercial and public bioactivity databases.** *J Chem Inf Model* 2012, **52**:319-326.

58. Greshock J: **Bioinformatics and cancer genetics**. In *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data*. Edited by Barnes MR, 2 edn. Chichester, WSX, UK: John Wiley & Sons; 2007: 413-439.
59. Deenen MJ, Cats A, *et al.*: **Part 4: pharmacogenetic variability in anticancer pharmacodynamic drug effects**. *Oncologist* 2011, **16**:1006-1020.
60. Heiser LM, Sadanandam A, *et al.*: **Subtype and pathway specific responses to anticancer compounds in breast cancer**. *Proc Natl Acad Sci USA* 2012, **109**:2724-2729.

Appendix

Drugs, trademarks and trademark owners

Drug	Trademark	Trademark Owner
Cisplatin	Platinol	Bristol-Myers Squibb
Cyclophosphamide	Cytoxan	Bristol-Myers Squibb
Dasatinib	Sprycel	Bristol-Myers Squibb
Docetaxel	Taxotere	Sanofi-Aventis
Erlotinib	Tarceva	Astellas Pharma, Genentech
Gefitinib	Iressa	AstraZeneca
Lapatinib	Tykerb	GalaxoSmithKline
Marimastat	n/a (development terminated)	British Biotech
Olaparib	n/a (investigational)	AstraZeneca
Paclitaxel	Taxol	Bristol-Myers Squibb
Vorinostat	Zolinza	Merck