In silico models of drug response in cancer cell lines based on various molecular descriptors

This dissertation is submitted for the degree of Doctor of Philosophy in Biological Science.

Michael Patrick Menden
University of Cambridge
Darwin College

Submission date: 12th of August 2015

Total word count: 40,620

Declaration of contribution

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text. This declaration should also make clear any parts of the dissertation that have been submitted for any other qualification

_____ _____

Michael Patrick Menden date

Publications specific to my PhD

- Eduati F, Mangravite LM, Wang T, Tang H, Bare JC, Huang R, Norman T, Kellen M, **Menden MP**, Yang J, Zhan X, Zhong R, Xiao G, Xia M, Abdo N, Kosyk O, the NIEHS-NATS-UNC DREAM Toxicogenetics Collaboration, Friend S, Dearry A, Simeonov A, Tice R, Rusyn I, Wright FA, Stolovitzky G, Xie Y, Saez-Rodriguez J.
  "Prediction of human population responses to toxic compounds by a collaborative competition." Nature Biotechnology. 2015

- Bansal M, Yang J, Karan C, **Menden MP**, Costello JC, Tang H, Xiao G, Li Y, Allen J, Zhong R, Chen B, Kim M, Wang T, Heiser LM, Realubit R, Mattioli M, Alvarez MJ, Shen Y; NCI-DREAM Community, Gallahan D, Singer D, Saez-Rodriguez J, Xie Y, Stolovitzky G, Califano A; NCI-DREAM Community.
  "A community computational challenge to predict the activity of pairs of compounds." Nature Biotechnology. 2014

- Costello JC, Heiser LM, Georgii E, Gönen M, **Menden MP**, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, Mpindi JP, Kallioniemi O, Honkela A, Aittokallio T, Wennerberg K; NCI DREAM Community, Collins JJ, Gallahan D, Singer D, Saez-Rodriguez J, Kaski S, Gray JW, Stolovitzky G.
  "A community effort to assess and improve drug sensitivity prediction algorithms." Nature Biotechnology. 2014

- **Menden MP**, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J.
  "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties." PLoS One. 2013

- Iorio F, Rittman T, Ge H, **Menden M**, Saez-Rodriguez J. "Transcriptional data: a new gateway to drug repositioning?" Drug Discovery Today. 2013

Acknowledgments

It is with immense gratitude that I acknowledge the mentoring and help of my supervisor Prof. Dr. Julio Saez-Rodriguez. Dr. Saez-Rodriguez helped me to identify scientific questions, solve them and make the solution available to the whole scientific community. Besides scientific support during the whole duration of my PhD, he also guided me in the 'emotional rollercoaster' that is an invariable part of doing a PhD. I value Dr. Saez-Rodriguez not only as my mentor, but also as my friend.

I would like to highlight Dr. Francesco Iorio, a colleague and mentor. He not only shared his knowledge and expertise, he actually inspired many of the presented analyses. I consider it an honour to have worked with Dr. Iorio, which goes beyond just being colleagues.

I am indebted to Dr. Gos Micklem, Dr. John Overington, Dr. Christoph Merten and Dr. Pedro Ballester, who together formed my thesis advisory committee.

This dissertation would have not been possible without the collaboration with the GDSC project lead by Dr. Cyril Benes, Dr. Mathew Garnett and Dr. Ultan McDermott. They kindly shared their screen data prior-to-publication, but furthermore contributed insightful discussions to analyses and method development. Moreover, I would to show my gratitude to the Lodewyk Wessels Lab from NKI, who also contributed to discussion.

I share the credit of my work with Dr. Oliver Stegle and his research group, regarding GWAS and impact of germline mutations. Within this project, I am delighted to highlight Francesco Paolo Casale and Johannes Stephan, who I mainly collaborated with.

Furthermore, I am indebted to many other colleagues who supported me and contributed to discussion: Dr. Graham Bignell, Michael Schubert, Christoph Wiedermann, Dr. Camille Terfve, Emanuel Gonçalves, Dr. Thomas Cokelaer, Dr. Federica Eduati, Dr. Luz Garcia Alonso, Dr. Marti Bernardo Faura, Dr. Denes Turei, Damien Arnol, Dr. Fatemeh Zamanzad, Dr. Marc Brehme, Konrad Rudolph, Aleksandra Kolodziejczyk and Jakob Wirbel.

This thesis would also not have been possible unless of the enormous support of my parents, Hartmut and Vera Menden and my siblings Nicole, Sven and Marc and last but not least, my fiancé Emma Carter.

Preface

“Time is shortening. But every day that I challenge this cancer and survive is a victory for me.”

- Ingrid Bergman

“Sunshine all the time makes a desert.”

- Arab proverb

“You never know how strong you are until being strong is the only choice you have.”

- Cayla Mills

“Attitude is a little thing that makes the big difference.”

- Winston Churchill

Summary

In the UK approximately one-third of the mortality rate is due to cancer. We are continuously undergoing cell renewal with ~$120x10^3$ novel somatic point mutations per cell division, and each has the potential to become a tumour driver and cause various types of cancer. Understanding that cancer is a disease with many subtypes and various drivers, implicates the need for personalised treatments tailored to individuals.

In this study I predicted drug responses in cancer cell lines from various molecular descriptors. Although cell lines are not perfect reflections of *in vivo* tumours, various leaps forward in modern cancer treatment have been achieved through such cell line models. In order to identify potential biomarkers of drug response, I analysed the two largest pan-cancer high-throughput screens available, namely, the Genomics of Drug Sensitivity in Cancer (GDSC) project and the Cancer Cell Line Encyclopaedia (CCLE).

At first, I predicted the drug responses of different cell lines and analysed driving factors for drug sensitivities based on genomic, epigenomic and transcriptional features separately as well as in combination. I built models based on as few features as possible, but as many as needed for a good prediction so that those models may be translatable to clinics.

Then I explored the increased predictive power by adding chemistry. I combined genomic features of the cell lines and chemical properties of the compounds, and showed their predictive improvement over using genomics alone. This model reproduced known biomarkers and showed superior predictability over models based solely on genomics.

In addition to using somatic alterations for patient stratification (i.e. responders vs. non-responders), I leveraged the impact of germline variances. Germline traits are usually applied as toxicity markers due to the fact that germline traits are identical in every cell in an individual. For leveraging germline mutations as sensitivity markers, I explored interactions with somatic mutations.

In summary, I developed and applied methods to predict drug responses and studied biomarkers of drug sensitivity based on various descriptors. Furthermore, I analysed lead drug candidates regarding their mode-of-action (MoA) and hypothesised future clinical applications.

Table of Contents

Table of Figures

Chapter 1: Introduction

“Cancer is a word, not a sentence.”

- John Diamond

1.1: Declaration of contribution

This chapter is based exclusively on my own perspective of cancer biology, drug discovery and prediction methods towards personalised medicine.

It must be understood that genomic alterations are the prime causality of cancer, i.e. "*Cancer is a genetic disease of somatic cells.*" (Knudson, 2002). Cancer is predominantly driven by somatic mutations that accumulate during the life span of an organism (Vogelstein & Kinzler, 1993). It has to be noted that a cell originates from another cell (in Latin "*omnis cellula e cellula*"), which was described by Rudolf Virchow in 1855, except for meiosis (diploid cell produces 4 haploid cells, i.e. gametes). Nowadays, this is common knowledge, but in the past this was the key to understanding that cancer originates from healthy cells in an individual (Ghoshal, 2012). This accumulation of somatic mutations is comparable to a mini-evolution of cells in our body, which start to grow uncontrolled, divide in large numbers and finally become hostile for the multicellular organism and cause cancer.

1.2: Cancer biology

1.2.1: Somatic mutations

Cancer is a genetic disease that is driven by somatic mutations, which are mutations that are accumulated through the lifespan of an individual. Most of these mutations occur during DNA replication, which is carried out by DNA polymerase enzymes. Although they can copy the DNA template with high accuracy, their error rate is estimated to be 1 in 100×10^3 replicated nucleotides (Pray, 2008). The human nuclear DNA consists of ~6×10^9 base pairs. In a diploid cell, during each DNA replication $2\times6\times10^9$ / 100×10^3 = 120×10^3 errors are made (Pray, 2008). Furthermore, an average human body is estimated to consist of ~3.72×10^{13} cells (Bianconi et al., 2013), which are constantly renewed. The average age of a cell in an adult human body is estimated to be 7-10 years (Spalding, Bhardwaj, Buchholz, Druid, & Frisén, 2005). However, the renewal rate is strongly dependent on tissue type. For example, in extremely hostile environments, such as in the small intestine with low pH values, epithelial turnover occurs at least once a week (Cliffe et al., 2005). Assuming an average cell renewal of ~8.5 years, this would lead to 3.72×10^{13} / $8.5\times365\times24\times60 \approx 8.3\times10^6$ cells renewed per minute in an adult human body. In total, this means that on average approximately

$8.3\times10^6\times120\times10^3 = 996\times10^9$ synthesis errors per minute are occurring in an adult human body. However, the human body has remarkable intracellular repair mechanisms that are discussed below and further fail-safe options such as a controlled cell death (apoptosis) or cell senescence. On the extracellular level the immune system is another 'guardian' against cancer (Hanahan & Weinberg, 2011). Nevertheless, in my thesis I study cancer cell lines and therefore focus solely on the intracellular defence mechanisms against cancer.

1.2.2: DNA damage repair

Despite having a relatively accurate DNA-replication mechanism, the cell has various DNA damage repair mechanisms in place. Those DNA damage repair mechanisms could be classified depending on the type of DNA damage into: (i) mismatch repair, (ii) excision repair, (iii) single strand break repair and (iv) double strand break repair.

Mismatches of bases are mainly a direct result of replication errors (Lodish et al., 2000a). Excision errors are induced through external influences, e.g. UV light might cause two neighbouring thymine residues to covalently bind each other (Lodish et al., 2000a). Single strand breaks predominantly occur through disintegration of deoxyribose or unsuccessful TOP1 activity and might collapse into replication forks or block transcription (Caldecott, 2008).

X-rays and oxygen radicals predominantly cause double strand breaks (Negritto, 2010). Double strand breaks may cause translocations across chromosomes, which furthermore may fuse proto-oncogenes together and stimulate consistent cell proliferation, e.g. *BCR-ABL* (see 1.2.7: Fusion genes). Translocations may also alter gene expression by rearranging promoters to cause oncogenic overexpression, e.g. an additional inserted promoter in front of the *BCL6* gene causes *BCL6* overexpression in B cell lymphoma (B. H. Ye et al., 1995).

In cancer therapy, DNA alkylating agents have been developed to induce DNA double strand breaks, which at first seems counter-intuitive when bearing in mind that exactly those DNA damages are causal for certain cancer types. However, tumour cells proliferate faster and are more vulnerable to DNA damaging agents compared to healthy cells, making them a valuable therapeutic option in late state cancer treatment while being carcinogenic at the same time (Jackson, 2009).

The DNA damage response is a complex signal transduction process, which depends on the type of DNA damage. Here I focus on single and double strand break repair pathways, since those are potentially caused by chemotherapy, as previously mentioned. Proteins involved in this repair machinery are usually categorised into protein cascades consisting of sensors, transducers, mediators, downstream kinases and effectors (Polo & Jackson, 2011; Sulli, Di Micco, & d'Adda di Fagagna, 2012) (Figure 1.1).

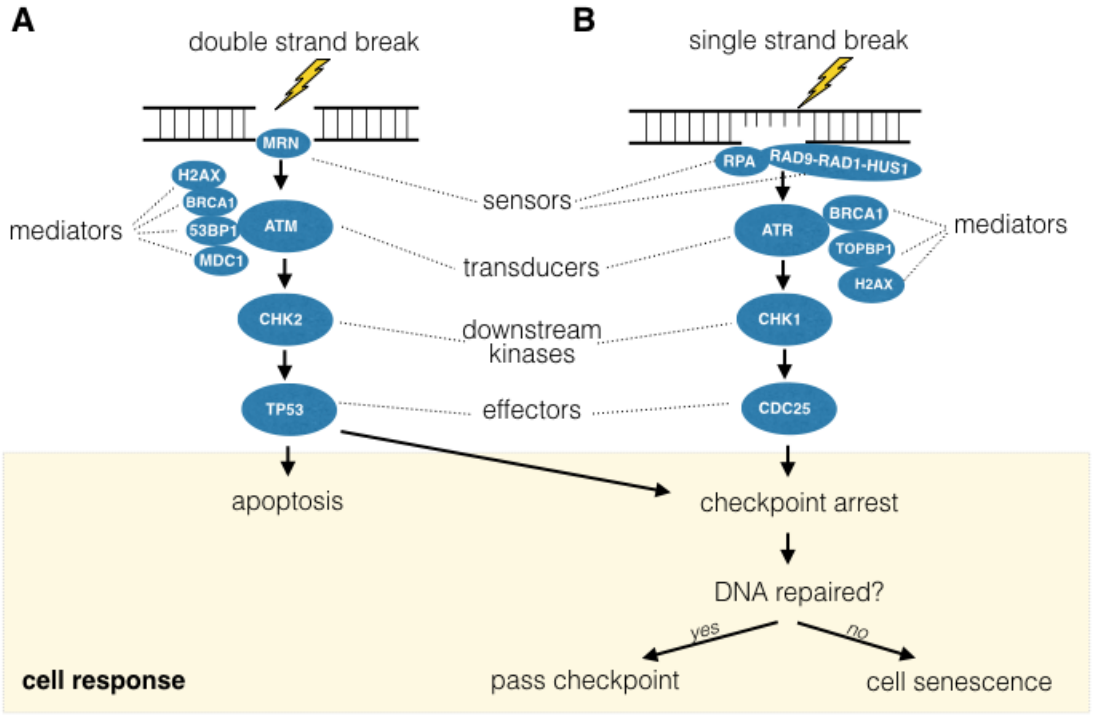

Figure 1.1: DNA damage response.

Shows the DNA damage response to (A) double strand breaks and (B) single strand breaks. Double strand breaks may cause severe damage to the genetic material, which might lead the cell to programmed cell death (apoptosis) or cell cycle arrest. A cell cycle arrest is reversible if the DNA is repairable, otherwise it may transition to irreversible cell senescence.

For example, the MRN complex (consisting out of Mre11, Rad50 and Nbs1) senses DNA double strand breaks (J.-H. Lee & Paull, 2004, 2005), while single-strand DNA breaks are detected by the replication protein A (RPA) and RAD9-RAD1-HUS1 complex (R. Chen & Wold, 2014). MRN activates through the transducer ATM, several DNA damage mediators such as 53BP1, MDC1, etc. (Shiloh, 2006), while the RPA and RAD9-RAD1-HUS1 complex rather activates the transducer ATR, which further phosphorylates TOPBP1. ATM and ATR signalling pathways share common mediators such as BRCA1 and H2AX (Sulli et al., 2012). Downstream of ATM is CHK2 and the effector TP53 (Smith, Tho, Xu, & Gillespie, 2010), the activation of which leads to either apoptosis or checkpoint arrest (Norbury & Zhivotovsky, 2004). ATR mediates its signal through CHK1 and activates the effector CDC25 (Xiao et al., 2003), which also forces the cell to checkpoint arrest. After checkpoint arrest, two outcomes are possible; either a successful DNA repair leading to further cell proliferation, or alternatively, a shift from temporary cell cycle arrest to permanent cell senescence (Sancar, Lindsey-Boltz, Unsal-Kaçmaz, & Linn, 2004; Sulli et al., 2012)

1.2.3: Passenger and driver mutations

Besides relatively high DNA-replication accuracy and various DNA damage repair mechanisms, a cell can tolerate most somatic mutations. Somatic mutations with no impact on cell fitness are called ‘passengers’, while mutations that are beneficial for the cancer cell are called ‘driver’ mutations (Figure 1.2).

Cancer somatic mutations are rare. About ~98% of the human genome consists of non-coding sequences (i.e. introns and intergenic regions), while only ~2% are exons that contain gene coding sequences (Elgar & Vavouri, 2008). It is worth noting that not only exon regions are functionally important (Bernstein et al., 2012); mutations in regulatory elements such as promoters or enhancers can have a pronounced effect on gene expression and cause functional defects.

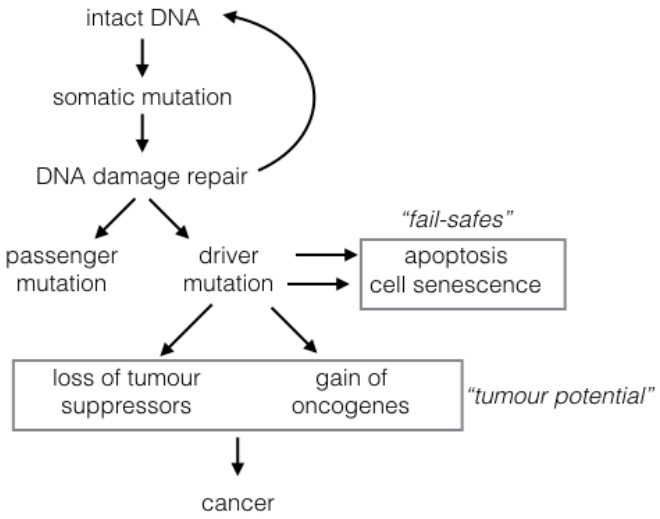

Figure 1.2: Cell fate after somatic mutation.

A somatic mutation might be repaired by DNA damage mechanisms, but if those fail it can result in a passenger or driver mutation. 'Fail-safes' might lead cells with driver mutations in cell senescence or controlled cell death (apoptosis). Driver mutations that bypass those 'fail-safes' might correspond to a loss or gain of tumour suppressors or oncogenes, respectively. Accumulation of such drivers ultimately leads to cancer (Vogelstein & Kinzler, 1993).

Mutations within a coding sequence can have a different impact. For example, mutation of the DNA sequence may not result in a change of protein sequence, because more than one nucleotide triplet codes each amino acid. Such mutations are called synonymous mutations or silent SNPs (single nucleotide polymorphisms) and do not alter the protein product, although silent SNPs might change the splicing pattern (Supek, Miñana, Valcárcel, Gabaldón, & Lehner, 2014), alter the gene expression profile (Sauna & Kimchi-Sarfaty, 2011) and modify tRNA abundance, leading to bottlenecks in protein synthesis (Sauna & Kimchi-Sarfaty, 2011).

A mutation causing an amino acid substitution may be tolerated, if the substituting and substituted amino acids have similar chemical properties. In cases when both amino acids are chemically very different, a substitution may still be tolerated, if it is located in a protein region that is not crucial for the protein's function. For example, functionally important regions are phosphorylation sites, catalytic pockets of an enzyme, DNA or RNA binding sites. The importance of protein regions can be estimated based on their

evolutionary conservation level across species, since important regions tend to be conserved across species (D. Lee, Redfern, & Orengo, 2007).

Evolution is not an optimised process that finds the best solution, rather a 'trial-and-error' selection of a good solution (Rudel & Sommer, 2003). This explains frequent redundancies in biological processes, providing high resilience against somatic mutations and ultimately against cancer.

1.2.4: Tumour suppressors and oncogenes

Genes related to cancer can be classified into either tumour suppressors or oncogenes (E. Lee & Muller, 2010). One objective of tumours is to bypass or disable key tumour suppressors, e.g. via loss-of-function mutations or homozygous deletions. The first discovered tumour suppressor was the gene *RB*, which is deleted in most retinal cancers (Murphree & Benedict, 1984). RB is a major regulator in the cell cycle (see next section for details), which prevents the G1- to S-phase cell cycle progression by inhibiting transcription factors from the *E2F* gene family. E2F together with the dimerisation protein (DP) forms a protein complex E2F-DP, which initiates the S-phase transition. Without a functional E2F-DP complex a cell remains in G1-phase (Harbour, 2000).

Besides bypassing tumour suppressors such as *RB*, cancer also actively drives tumorigenesis through oncogenic mutations, e.g. mutating *KRAS*. Most common mutations in KRAS are at position G12, G13 and Q61. These mutations are enriched in colorectal, lung and pancreatic cancer (Janakiraman et al., 2010). KRAS is a GTPase that is active when bound to GTP. It is part of the ERK signalling cascade and activates the PI3K-AKT pathway (Figure 1.3) (Dhillon, Hagan, Rath, & Kolch, 2007). In cancer, KRAS mutations predominantly cause activation of ERK signalling but also trigger PI3K-AKT signalling cascades. This leads to uncontrolled cell division and cell survival (Hanahan & Weinberg, 2011), which ultimately becomes independent from ligand induced activation, e.g. growth factors (Figure 1.3). This dependency on oncogenes is often referred to as 'oncogene addiction'.

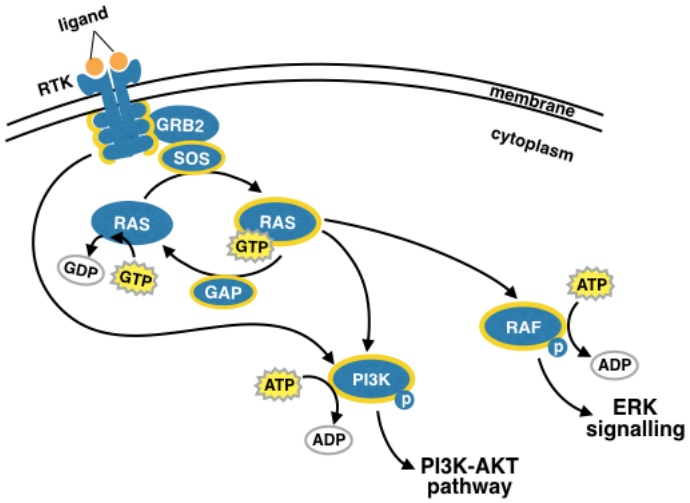

Figure 1.3: RAS activation.

Ras is activated through RTKs such as EGFR. EGFR is stimulated through the ligand EGF, which causes a conformational change in EGFR and consequently its autophosphorylation. Activated EGFR attracts the adaptor protein GRB2, which upon a conformational change binds SOS. SOS releases GDP, allowing GTP to bind RAS, which therefore activates RAS. RAS is deactivated by GAP, which hydrolyses GTP to GDP in RAS. Activated RAS binds to the RAS binding domain in RAF and PI3K, resulting in conformational changes in those kinases and promoting their phosphorylation. PI3K can also be directly phosphorylated by EGFR. Activation of RAF and PI3K ultimately cause cell proliferation and cell survival. RAS is one of the most frequently mutated oncogenes in cancer and actively causes hallmarks of cancer. Activation is highlighted with a yellow border.

Understanding the function of tumour drivers is the first step towards personalised treatments. In particular, oncogenes are a logical choice as drug targets, although they are often not directly targetable, e.g. KRAS wild type is not directly targetable (Zimmermann et al., 2013), mainly because it has no deep hydrophobic pockets where small molecules could bind (Cox, Fesik, Kimmelman, Luo, & Der, 2014). Yet, the $KRAS^{G12C}$ mutant is allosterically targetable due to the presence of a novel pocket, which is non-competitive to the GTP binding site (Ostrem, Peters, Sos, Wells, & Shokat, 2013).

1.2.5: Cell cycle

DNA damage and chromosomal abnormalities may occur at different stages in the cell cycle, which is the regulating process of cell division (Collins, Jacks, & Pavletich, 1997; Morgan, 2007). The cell cycle is separated into the following 2 distinct stages: (i) interphase and (ii) M-phase (Figure 1.4) (Campbell & Reece, 2006a).

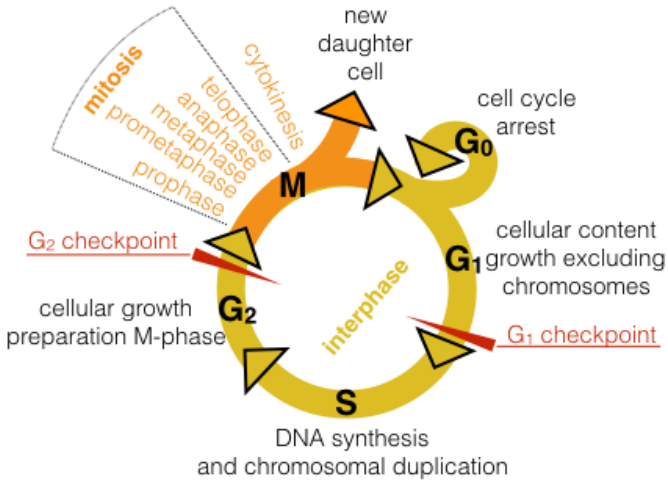

Figure 1.4: Cell cycle.

The cell cycle is separated in interphase (yellow) and M-phase (orange). The interphase is based on 4 distinct phases, G0, G1, S and G2 phase. The M-phase further separates in mitosis and cytokinesis. Mitosis is further split into prophase, prometaphase, metaphase, anaphase and telophase.

The interphase starts with the G1-phase, where the cell has 46 chromosomes and 46 chromatids. In the G1-phase the cellular content (organelles) grows, excluding the chromosomes. Before completing G1-phase, nutrition needs to be sufficient, growth factors have to be present and no DNA-damage occurred. This check at the end of the G1 phase is called the G1-checkpoint, which may or may not allow entering into the next part of the interphase (Morgan, 2007).

The next stage of interphase is the synthesis phase (S-phase), where the DNA is replicated and chromosomes are duplicated, leaving us with 46 chromosomes and 92 chromatids. The final interphase stage is G2, in which the cell further grows and prepares for mitosis. This checkpoint is the so-

called G2-checkpoint, which ensures that chromosome replication is completed and no DNA-damage has occurred (Morgan, 2007).

After successful finalisation of the interphase, the cell enters the M-phase, which is further separated into two stages: (i) mitosis and (ii) cytokinesis. Mitosis is the separation of the chromosomes within the original cell, and cytokinesis is the division of the cell into two presumably identical daughter cells (Morgan, 2007).

Mitosis is further separated into (i) prophase, (ii) prometaphase, (iii) metaphase, (iv) anaphase and (v) telophase. In prophase chromosomal material is condensed, the nucleolus envelope is dispersed, the cytoskeleton is dissembled and the mitotic spindle is assembled. In prometaphase the chromosomes are associated with the mitotic spindle, meaning the microtubules connect from the polar oriented centrosome to the chromosome kinetochore of each chromosome. In anaphase, sister chromatids are separated by the centrosomes, and move to the opposite poles. In the final telophase, the chromosomes cluster at opposite spindle poles, chromosomes disperse again, the nuclear envelope reassembles and organelles reform. Finally the cell is ready to engage cytokinesis, which separates both daughter cells. Those daughter cells might restart in the G1 phase or stay in a reversible cell cycle arrest called G0 (Campbell & Reece, 2006a; Morgan, 2007).

The cell cycle is regulated by transcriptional waves, which are initiated by transcription factors of the *E2F* family (Bertoli, Skotheim, & de Bruin, 2013). Those E2F transcription factors are inactive when bound to members of the pocket protein family such as RB, p130 and p107. The transcription waves are activated by cyclin-dependent kinases (CDKs), which are able to phosphorylate pocket protein family members. Hence, CDKs free up E2F transcription factors and thereby activate them. Notably, CDK activation further triggers a positive feedback loop for increased CDK activation. After passing from one cell cycle phase to the next, negative feedback loops decrease CDK activity again, e.g. $p27^{CDKN1B}$ after passing from the G1 to

S-phase. Those positive and negative feedback loops explain the observed waves of E2F regulated transcription (Malumbres & Barbacid, 2009; Vermeulen, Van Bockstaele, & Berneman, 2003). First, I exemplify the G0 to G1 transition, and secondly the G1 to S-phase transition (Figure 1.5).

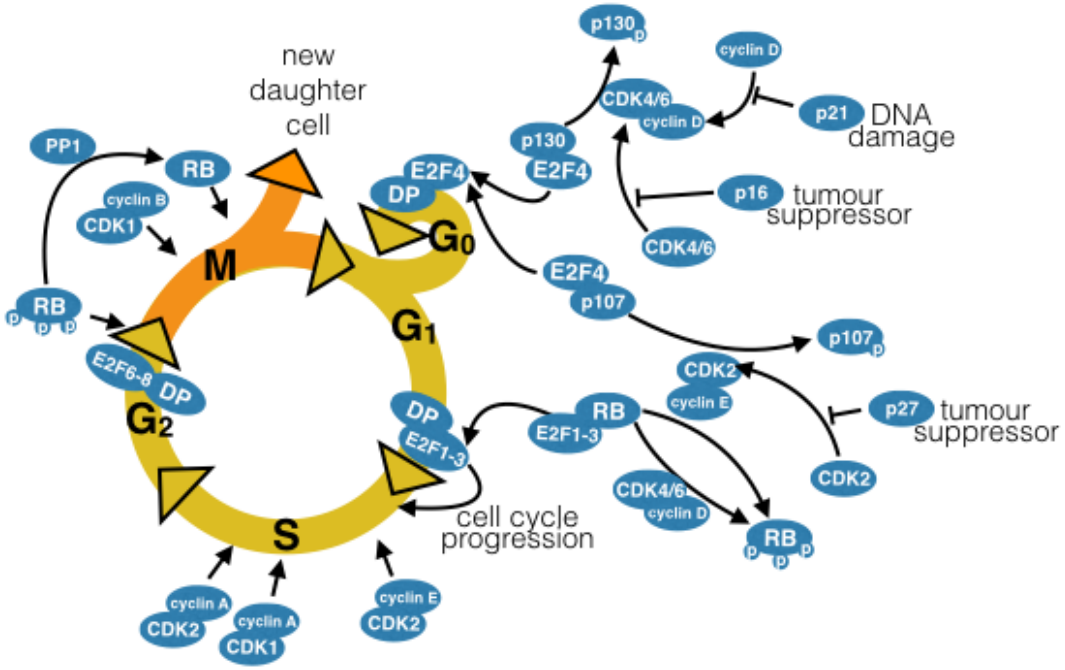

Figure 1.5: Cell cycle G0-G1 and G1-S phase transition.

This figure shows a simplification of the G0 to G1 phase transition as well as the transition from the G1 to S-phase. The E2F protein family are transcription factors, which upon binding with DP initiate a transcription wave. RB, p130 and p107 prevent those transcription waves, by inhibiting the E2F transcription factors upon binding. E2F can be released from binding to RB, p130 and p107 by the means of phosphorylating their RB, p130 or p107. CDK and cyclin complexes are phosphorylating RB, p130 and p107. There are tumour suppressors such as p16^{CDKN2A} and p27^{CDKN1B}, which prevent the formation of CDK and cyclin complexes to pause the cell cycle. p21^{CDKN1A} is downstream from the DNA damage pathway, which also enables a halt function of the cell cycle.

E2F4 activation is required to reverse cells from the G0 to G1 phase (Gaubatz et al., 2000). Inactive E2F4 binds predominantly p130 and p107, but also with lower affinity RB. E2F4 can be released from p130 by CDK4 or CDK6 in complex with cyclin D. p130 is phosphorylated in this process, which disables the binding to E2F4. The CDK2 and cyclin E complex are able to

phosphorylate p107 and also free up E2F4, which ultimately enables the transition from the G0 to G1 phase (Bertoli et al., 2013).

Tumour suppressors such as $p16^{CDKN2A}$ or $p27^{CDKN1B}$ can prevent the assembly of the CDK and cyclin complex, which consequently halts the cell cycle. For example, $p27^{CDKN1B}$ binds to CDK2 with higher affinity than to cyclin E, and thereby prevents the building of the CDK2 and cyclin E complex. In addition, $p16^{CDKN2A}$ binds to CDK4 and CDK6 with higher affinity than to cyclin D, and therefore prevents the complex assembling with cyclin D. Both, $p16^{CDKN2A}$ and $p27^{CDKN1B}$ are CDK inhibiting proteins (Bertoli et al., 2013).

Another CDK inhibitor is RB (Dowdy et al., 1993; Henley & Dick, 2012; Weinberg, 1995). RB was the first tumour suppressor to be discovered and loss of RB function is a driver mutation for retinoblastoma, which aggressively proliferates due to deregulated E2F1/2/3 (Weinberg, 1995). RB is particularly important in transition from the G1 to S-phase and passing the G1-checkpoint. RB binds with high affinity to the transcription factors E2F1, E2F2 and E2F3. Those transcription factors are required to be freed from RB to pass the G1 checkpoint. RB is phosphorylated by the CDK2 and cyclin E complex as well as CDK4/6 and cyclin D, which thereby frees up E2F1/2/3 and ultimately enters the DNA synthesis and duplication phase. PP1 again dephosphorylates RB in the last state of M phase, to regain RB's inhibitory activity towards E2F1/2/3 in the G1 phase (Vietri, Bianchi, Ludlow, Mittnacht, & Villa-Moruzzi, 2006).

Cells may also halt in G0 (cell cycle arrest) due to DNA damage response regulated by TP53. High expression of *TP53* are associated with apoptosis while lower expressions cause cell arrest and expression of $p21^{CDKN1A}$ (X Chen, Ko, Jayaraman, & Prives, 1996; Kracikova, Akiri, George, Sachidanandam, & Aaronson, 2013). $p21^{CDKN1A}$ binds to cyclin D and prevents the formation of CDK4 or CDK6 with cyclin D complex.

The cell cycle arrest (G0 phase) is reversible, which is not the case in cell senescence. Cell senescence is a permanent and irreversible arrest, but the

expression profile and phenotype of cell cycle arrest and senescence are similar (Pérez-Mancera, Young, & Narita, 2014). Cell senescence associated secretion is reported in some instances such as tumour suppressing and other instances such as tumour promoting and may be oncogene induced (Courtois-Cox, Jones, & Cichowski, 2008). Therefore, cell senescence is a 'double-edged sword' that might be beneficial or disadvantageous for cancer.

Loss of cell cycle control is a hallmark of cancer, i.e. increased cell proliferation without controlling mechanisms and ultimately avoiding 'fail-safes' such as cell senescence or apoptosis (Hanahan & Weinberg, 2011; Lodish et al., 2000c). Therefore, many cancers evolutionarily favour the loss-of-function of key regulators such as RB, p16^{CDKN2A} and p27^{CDKN1B}. The functionality of TP53 and p21^{CDKN1A} are also frequently lost to circumvent the DNA damage response, cell cycle arrest and apoptosis. In contrast, CDKs and cyclins are upregulated in many cancers to increase the proliferative potential and remove the 'breaks' of the cell cycle.

Understanding the function of CDKs and their protein inhibitors has given rise to various chemicals developed to target specific CDKs (Besson, Dowdy, & Roberts, 2008; Sherr & Roberts, 1999), e.g. RO-3306 (CDK1i), CGP-082996 (CDK4i), THZ-2-102-1 (CDK7i), AT-7519 (CDK9i), KIN001-270 (CDK9i) or THZ-2-49 (CDK9i). There are also inhibitors targeting multiple CDKs e.g. palbociclib (CDK4/6i) or pan-CDK inhibitors e.g. CGP-60474 (CDK1/2/5/7/9i), roscovitine (CDK1/2/4/5/6i), PHA-793887 (CDK1/2/4/5/7/9i). All those compounds aim to reduce cell cycle activity.

An alternative therapeutic approach to slowing the cell cycle might be to further accelerate it, allowing cells with damaged DNA to proceed past cell cycle checkpoints. At first, this may sound controversial since cancer aims for uncontrolled proliferation; however, their accumulated mutational burden might make them more vulnerable than normal cells (Huber et al., 2014).

Experimental compounds such as 681640, which target WEE1 and CHK1, are capable of removing the 'cell cycle brake'. WEE1 and CHK1 inhibition leads to

high cyclin-dependent kinases (CDK) activity, premature replication initiation and aberrant fork structures that ultimately cause susceptibility to DNA damage (Sørensen & Syljuåsen, 2012). CHK1 and WEE1 inhibition are rendering cells sensitive to alkylating agents (Ma, Janetka, & Piwnica-Worms, 2011), the latter particularly in acute myeloid leukaemia (AML) (Chaudhuri et al., 2014).

1.2.6: Copy number variations

Throughout the cell cycle, specifically during mitosis, nuclear DNA is error-prone. In mitosis the DNA is separated into two distinct cell nuclei. First the centriole divides and then moves to the opposite sites in the nucleus. At the same time, the DNA condenses into chromosomes. Afterwards, starting from both centrioles the spindle fibres connect with centromeres of each chromosome and finally pull them apart. A healthy cell does not separate the sister chromatid until all chromosomes are correctly connected (Lodish et al., 2000b). However, in cancer this is often not the case and abnormal chromosome numbers, such as the presence of only one chromosome of a pair (aneuploidy) and/or multiple additional chromosomes (polyploidy) are common (Nigg, 2002). Cancer cells also exploit inaccuracies specifically in mitosis. As a consequence of evolutionary pressure, copy number variations (CNVs) and fusion genes are naturally selected, favouring uncontrolled cell proliferation, bypassing apoptosis and setting the stage for various other hallmarks of cancer (Hanahan & Weinberg, 2011).

For example, the chromosomal region 9p21 is deleted in many non-small-cell lung cancers (NSCLC), gliomas, leukaemias and melanomas (Cairns et al., 1995). This region contains the gene *CDK2NA* (cyclin-dependent kinase inhibitor 2A), which encodes the protein p16. As previously described, this protein is a well-known tumour suppressor, which regulates cell progression from the G1-phase to S-phase. Deletion of the *CDK2NA* is evolutionary favourable for cancer cells to achieve uncontrolled cell proliferation (see page 11, Figure 1.5).

Another example is the amplification of the chromosomal region 12q13-q14 in sarcomas (Oliner, Kinzler, Meltzer, George, & Vogelstein, 1992). This region contains the gene *MDM2*, whose protein product suppresses TP53 functionality, the so-called "*guardian of the genome and policeman of the oncogenes*" (Efeyan & Serrano, 2007). TP53 is an initiator of controlled cell death (i.e. apoptosis) and cell senescence.

1.2.7: Fusion genes

In addition to copy number variations, cancer stochastically explores translocations of chromosomal regions to create novel fusion genes with oncogenic properties. For instance, in 1960 Nowell and Hungerford discovered the Philadelphia chromosome, which is a chromosomal abnormality in chronic myeloid leukaemia (CML) and acute lymphocytic leukaemia (ALL) (Nowell & Hungerford, 1960). Later in 1973, Janet Davison Rowley elaborated that this abnormality is a reciprocal translocation of chromosome 9q34 and 22q11, which lengthens chromosome 9 and shortens chromosome 22 at the same time (Rowley, 1973). A decade later this translocation has been identified to produce the fusion gene *BCR-ABL* (Bartram et al., 1983), which is a tyrosine kinase with increased activity compared to the proto-oncogene ABL (Lugo, Pendergast, Muller, & Witte, 1990) and ultimately leads to enhanced cell growth in ALL and CML.

1.2.8: Chaperones

The correct folding of proteins is of prime importance to execute their potential native function and is enabled by chaperones. Chaperones are key molecules in the response to physical assaults such as heat shock, which was first studied in *Drosophila melanogaster* in the 60's (Ritossa, 1962) and later expanded to be a global phenomena within all organisms in the late 80's (Lindquist, 1986). During heat shock several proteins may denature to catalytically inactive states, folding intermediates, be degraded, or form toxic aggregates, while some may autonomously refold or depend on the assistance of chaperones to re-gain their native fold. The main controller of heat shock response is the transcription factor HSF1, which triggers the

expression of chaperones. The overexpression of chaperones is the main manifestation of heat shock response giving this protein class their name, heat shock proteins (HSPs) (Feder & Hofmann, 1999).

As mentioned in previous sections, cancer mutations accumulate during one's lifetime and thereby produce unstable or metastable proteins (metastable proteins are stable in presence of chaperones), which are comparable to the physical assaults of a heat shock. Cancer exploits chaperones to maintain cancer deregulation, which produces such unstable and metastable proteins. The abundance of heat shock proteins (HSPs) in unstressed cells is estimated to be ~1-2% of the total protein mass, while this may double under stress (Wegele, Müller, & Buchner, 2004; Luke Whitesell & Lindquist, 2005). Recently this number has been re-estimated upwards to even 5% HSPs in some types of breast cancer (Jarosz, 2016).

Chaperones are so-called housekeeping genes that take important roles in healthy cells, but are also an 'Achilles Heel' of cancer that can be therapeutically exploited (Luke Whitesell & Lindquist, 2005). HSP90, particularly, has been identified as a key supporter of tumorigenesis and capacitor of evolution (Rutherford & Lindquist, 1998). HSP90 is a chaperone that stabilises more than 200 client proteins and enables their functionality (Shiro Soga, Akinaga, & Shiotsu, 2013; Taipale, Jarosz, & Lindquist, 2010). Among those client proteins are key mediators in cell signalling, cell cycle regulators and transcription factors; however, HSP90 also serves various oncoproteins e.g. ERBB2, EGFR, FLT3, BCR-ABL, SCR, EML4-ALK, ER, etc.

1.2.9: HSP90 machinery

The ATPase cycle of the HSP90 complex is called HSP90 machinery (Wegele et al., 2004; Luke Whitesell & Lindquist, 2005), exemplified by the stabilisation of the intracellular oestrogen receptor (ER) (Figure 1.6 A), which thereby enables its binding to oestrogen. Consecutively, the ligand-bound ER interacts with the oestrogen response elements (ERE) and ultimately causes the transcriptional response of ER target genes (Klinge, 2001). The ER binds

an early chaperone complex composed of the chaperone HSP70 (Pratt & Toft, 2003) which is activated through the co-chaperones HSP40 (Fan, Lee, & Cyr, 2003) and HIP (Höhfeld, Minami, & Hartl, 1995). The intermediate complex binds an HSP90 dimer and replaces HSP40 with HOP (Odunuga, Longshaw, & Blatch, 2004). This intermediate complex matures by releasing HSP70, HIP and HOP, while at the same time the HSP90 dimer binds two p23 co-chaperones in an ATP consuming process (Young & Hartl, 2000). Additionally, and specific for ER activation, the binding of the co-chaperone CYP40 is necessary (Ratajczak & Carrello, 1996). This enables activation of the ER, which allows the ligand oestrogen to bind and activate ER (Knoblauch & Garabedian, 1999). After activation, the HSP90 complex disassembles under dephosphorylation of ATP to ADP (Hartl, Bracher, & Hayer-Hartl, 2011).

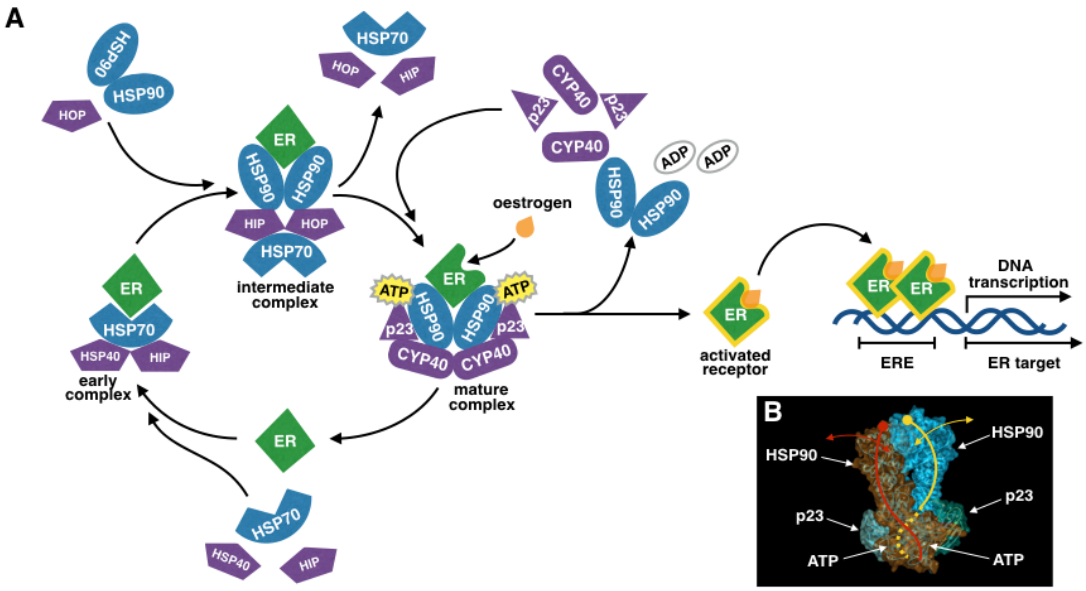

Figure 1.6: HSP90 machinery.

(A) shows the HSP90 complex cycle exemplified by the stabilisation of ER (green). Activation of ER is highlighted in yellow, which upon activation binds to oestrogen response elements (ERE) and causes ER targeted transcription. Adapted from Whitesell et al. (Luke Whitesell & Lindquist, 2005). Proteins in blue and purple are chaperones and co-chaperones, respectively. (B) shows the activated HSP90 complex binding ATP and p23. In yellow and red the spines of the HSP90 dimer are sketched. Generated with ProteinWorkshop 4.2.0 (J. L. Moreland, Gramada, Buzko, Zhang, & Bourne, 2005), RCSB PDB (Berman, 2000) identifier: 2CG9, (Ali et al., 2006).

The mature HSP90 complex is comparable to a pair of closed scissors, which opens up after the activation of its substrate (Figure 1.6 B) (Ali et al., 2006).

1.2.10: Summary

Somatic mutations stochastically occur at different stages in the cell cycle. The mutation rate is influenced by external factors, e.g. UV light exposure or smoking. Nevertheless, through evolution cells have developed mechanisms to repair DNA damage and acquired 'fail-safe' mechanisms such as apoptosis and cell senescence. Somatic mutations may still be passed on to viable daughter cells. Most of these mutations are passenger mutations that do not affect cellular homeostasis. However, some somatic mutations are oncogenic as they have a functional impact and may cause cancer. Common cancer somatic alterations are copy number variations (CNVs), fusion genes and non-synonymous SNPs potentially causing loss- or gain-of-function. Loss-of-function events in tumour suppressors and gain-of-function mutations in oncogenes are subject to natural selection. Ultimately, those oncoproteins alter cell signalling, which reprogram cells to become tumorigenic.

1.3: Cancer signalling

In order to understand tumour biology, it is important to bear in mind that a complex multicellular organism such as humans can only function as long as a secure communication between cells is assured. This concept of cell interaction is based on the pioneering work of Earl Wilbur Sutherland (Raju, 1999), who discovered the impact of the hormone adrenaline in the 60's. Adrenaline is secreted by the endocrine gland into the bloodstream and causes the release of glycogen in the liver and skeletal muscle cells. Glycogen is a polysaccharide that can be quickly transformed to energy. Sutherland found that adrenaline activates the glycogen phosphorylase within muscle and liver cells *in vivo*, but not *in vitro* when only combining glycogen phosphorylase, glycogen and adrenaline. This result implies that within muscle and liver cells other mediators of signal transduction must exist. Sutherland's finding launched the field cell signalling, which is based on 3 steps: (i) signal reception, (ii) transduction and (iii) response (Campbell & Reece, 2006b).

Lipophilic hormones are able to penetrate the cell membrane and enter the cell (e.g. oestrogen which binds intracellular to the oestrogen receptor), but most are too polar and/or too large. Rather they bind to receptor proteins that are anchored in the cell membrane. Here I focus on two receptor classes: (i) G-protein-coupled receptors (GPCRs) and (ii) receptor tyrosine kinases (RTKs).

GPCRs are embedded in the cell membrane with 7 alpha helices. Part of the protein is exposed to the extracellular space including a binding site for ligands, while the intracellular part interacts heterotrimeric G-protein subunits that in turn engage or regulate intracellular signalling proteins.

A heterotrimeric G-protein is inactive when it's Gα subunit is bound to GDP. When a specific ligand binds to the GPCR binding site, this enforces a structural change that results in GDP being replaced with GTP within the Gα subunit, which thereby becomes active. The activated Gα.GTP subunit

dissociates from the Gβγ dimer and both Gα.GTP and Gβγ activate downstream signalling effector proteins thereby transmitting a signal, which ultimately causes the cellular response (Figure 1.7) (Oldham & Hamm, 2008).

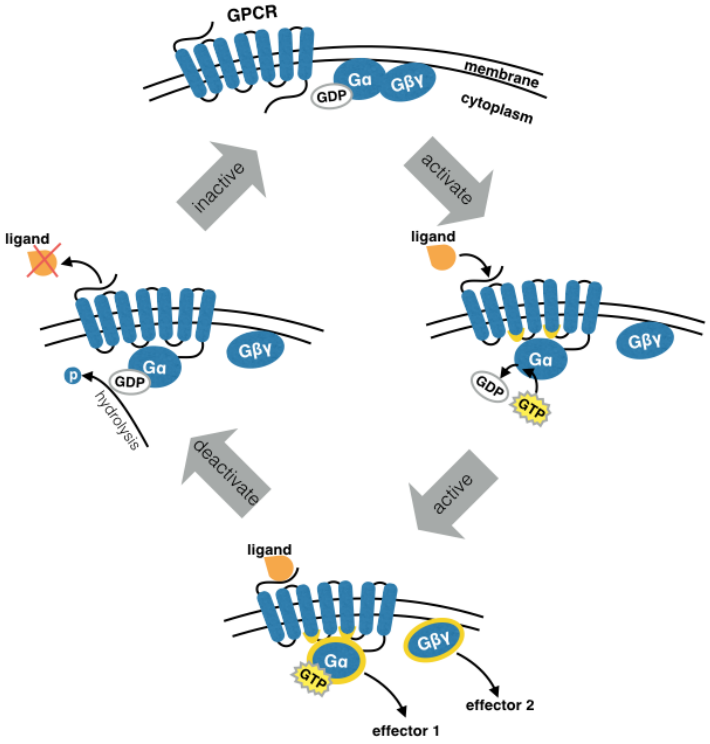

Figure 1.7: G-protein-coupled receptor signalling.

Figure shows the activation of GPCR upon ligand binding followed by a heterotrimeric GTPase cycle. Ligand binding leads to disassociation of Gα and Gβγ, resulting in replacement of GDP with GTP in Gα. Both subunits, Gα.GTP and Gβγ activate their own effectors until the ligand is removed, which consequently leads to hydrolysis of GTP to GDP and therefore inactivates the GPCR response.

Signalling is terminated by the intrinsic GTPase activity of the Gα subunit that hydrolyses the GTP to a GDP thereby inactivating itself.

Receptor tyrosine kinases (RTKs) are a different class of receptors containing 58 known RTKs with 20 subfamilies (Lemmon et al., 2010). In a cell, these mostly consist of two identical and unconnected monomers in the inactive state (Figure 1.8 A). The extracellular face of all RTKs has a ligand-binding site, while the cytoplasmic face of each monomer has a tyrosine kinase domain with an activation loop and a αC helix required for phosphorylating the substrates. Most RTKs dimerise upon ligand binding (Figure 1.8 B), although

there are also atypical RTKs requiring larger oligomers such as the Tie2 receptor (Barton et al., 2006) or Eph receptor (Himanen & Nikolov, 2003).

Dimerisation or assembling of larger oligomer allows RTKs to transition from cis-autoinhibtion (self inhibiton) to trans-autophosphorylation of the intracellular tyrosine kinase domain, leading to self-activation. The trans-autophosphorylated RTKs are able to recruit specific proteins containing phosphotyrosine-recognition domains, which are either SRC homology 2 (SH2) domains or phosphotyrosin biding (PTB) domains (Hubbard, 2004). SH2 and PTB domains are abundant in adapter proteins, which are capable of recruiting other substrates that are able to further transmit signal, e.g. GRB2 has an SH2 domain and recruits SOS and promotes its activation by the means of conformational changes (Schlessinger, 1994). Notably, adaptor proteins themselves are not capable to phosphorylate their interaction partners, but are able to change their partner's conformation, enabling them to be phosphorylated and thereby transmit a signal.

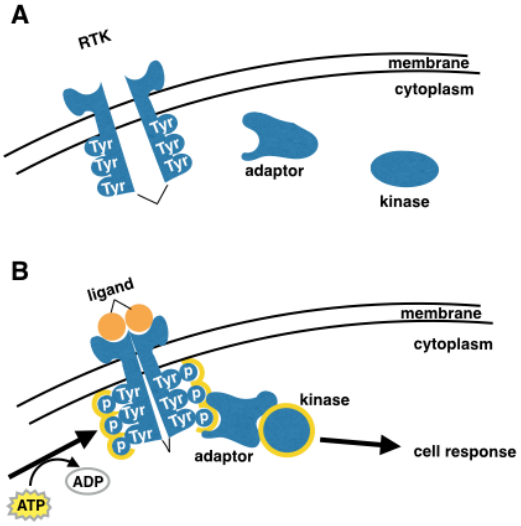

Figure 1.8: Receptor tyrosine kinase.

(A) shows an inactive receptor tyrosine kinase (RTK), while (B) illustrates the stimulated case, which recruits adaptors and cause a conformational change. This enables the adaptor to bind specific kinases, which changes the confirmation of those recruited kinases and promotes their phosphorylation. Phosphorylation is highlighted in yellow. Adapted from Campbell Biology (Campbell & Reece, 2006b).

By definition, all kinases are proteins that can phosphorylate their substrates by hydrolysing ATP to ADP and transferring the phosphate group onto their substrates phosphorylation sites. The substrate of a kinase could be activated (or deactivated) by this phosphorylation. Kinases themselves can be activated (or deactivated) through phosphorylation or other mechanisms. Although kinases are the most frequently mutated protein class in cancer and phosphorylation is the major mechanism for signal transduction, other post-translational modifications are used to transmit information, e.g. glycosylation, ubiquitination, acylation or methylation.

Signal transduction can follow a cascade of such activation or inhibition events, which ultimately culminate in a cellular response. Ligand-bindings to receptors may initiate such response cascades, although such 'cascades' have to be understood as models simplifying complex signalling networks. Nevertheless, the cascade concept is the basis for cellular signalling pathways, which remains a useful model to understand and conceptualise complex biological networks (see 1.3.2: Cancer pathways, page 26).

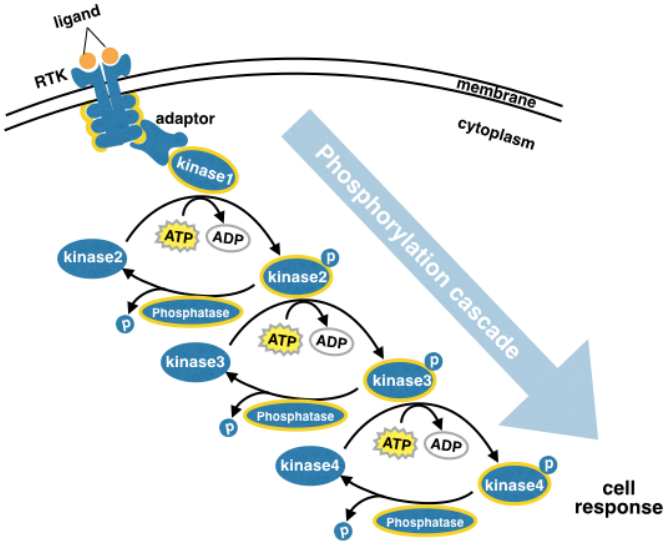

Figure 1.9: Phosphorylation cascade.

Figure shows a phosphorylation cascade involving 4 kinases. The stimulated RTK engages adaptors, which recruit kinase1 and enable its activation. The activated kinase1 phosphorylates kinase2, hydrolysing ATP to ADP and attaching the phosphate to its substrate kinase2. Activated kinase2 consequently phosphorylates kinase3, etc. The phosphorylation and dephosphorylation are in equilibrium with the incoming stimuli and phosphatase activity, which catalyse dephosphorylation. The activated kinases are highlighted in yellow. Adapted from Campbell Biology (Campbell & Reece, 2006b).

1.3.1: Kinase classes

Kinases are separated into two major classes: (i) tyrosine kinases and (ii) serine / threonine kinases, which either phosphorylate tyrosine or alternatively phosphorylate the amino acids serine or threonine(Manning, Whyte, Martinez, Hunter, & Sudarsanam, 2002). Tyrosine kinases can be further split into receptor and non-receptor tyrosine kinases (non-RTKs). Previously RTKs were introduced (Figure 1.8), an example of which is EGFR (Figure 1.10).

The non-RTKs are exemplified by SRC. Other non-RTKs not shown here are e.g. ABL, CHK or JAK2. The class of serine / threonine kinases is exemplified by PKA, PKC, PKG, CDK2.

At first, it was assumed that those classes of kinases are mutually exclusive, meaning a kinase would only be able to phosphorylate tyrosine or serine/threonine. This was disproved by the discovery of dual-specificity of the MAP kinase motif in the 80's (Dhanasekaran & Premkumar Reddy, 1998; Lindberg, Quinn, & Hunter, 1992), which separates MEK1/2 from other kinases (Figure 1.10).

All kinases consist of an N-tail, a catalytic core and a C-tail. The catalytic core is highly conserved across all kinases, while the N-tail contains conserved regulatory domains that depend on the kinase subfamily. Those domains are conserved building blocks, which determine the kinase's functionality and can be found in other kinases of the same family.

For example, the SH2 and SH3 domains found in SRC are also found in adapter proteins such as GRB2 (Schlessinger, 1994), which is important to mediate the signal from RTK to RAS (Figure 1.3). PKA is involved in the adrenaline response that was originally identified by Sutherland. The second messenger cAMP regulates PKA (Su et al., 1995). This regulatory subunit of PKA is also conserved in other multiple isoforms (Brandon, Idzerda, & McKnight, 1997). C1 and C2 domains enable PKC regulation through the second messenger calcium (Ca^{2+}) (Medkova & Cho, 1999) with also various isoforms (Steinberg, 2008). CDKs require binding to cyclin to regulate

progression in the cell cycle (John, Mews, & Moore, 2001), but no additional regulatory subunits are needed (see 1.2.5: Cell cycle). Docking domains are important for substrate specificity of mitogen-activated protein (MAP) kinases (Sharrocks, Yang, & Galanis, 2000), and are found in MEK1, MEK2 and ERK1/2.

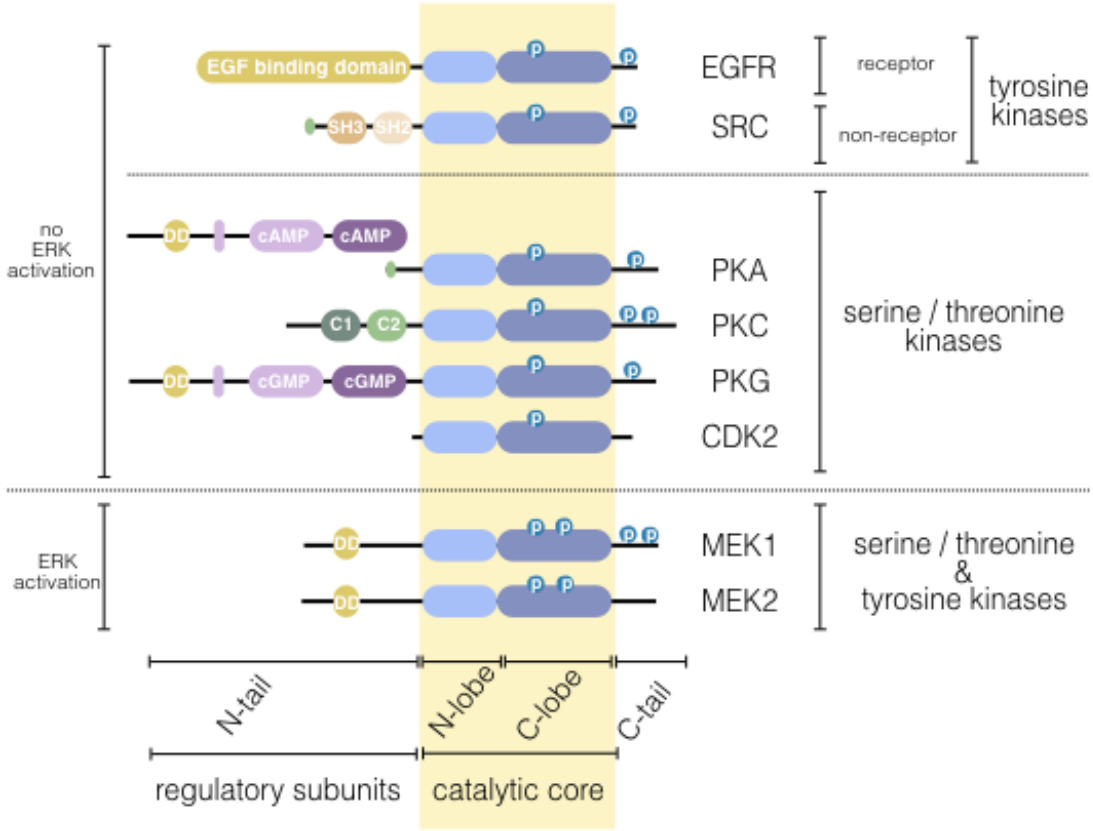

Figure 1.10: Kinases.

Illustrates the main classes of kinases: (i) tyrosine and (ii) serine/threonine kinases. Highlights the difference of MEK1/2 to all other kinases, which has 2 phosphorylation sites in the activation loop and is the only activator of ERK signalling. Adapted from Caunt et al. and Taylor et al. (Caunt, Sale, Smith, & Cook, 2015; Taylor, Ilouz, Zhang, & Kornev, 2012).

The N-lobe (part of the catalytic core) contains the so-called C helix, which is a highly conserved region, which contains a catalytic lysine residue (Carrera, Alexandrov, & Roberts, 1993). A conformational change of the C helix is required to activate and deactivate kinases such as c-ABL, BCR-ABL and MEK1/2 (Hantschel & Superti-Furga, 2004; Palmieri & Rastelli, 2013).

Allosteric drugs can exploit a displacement of the C helix, e.g. MEK inhibitors, which therefore do not compete with ATP (Figure 1.11).

MEK1 and MEK2 have two phosphorylation sites in their activation loop. Both phosphorylated amino acids are serine in MEK1 at position 218 and 222 (Zheng & Guan, 1994) (Figure 1.11 C), while in MEK2 at position 222 and 226 (Bromberg-White, Andersen, & Duesbery, 2012).

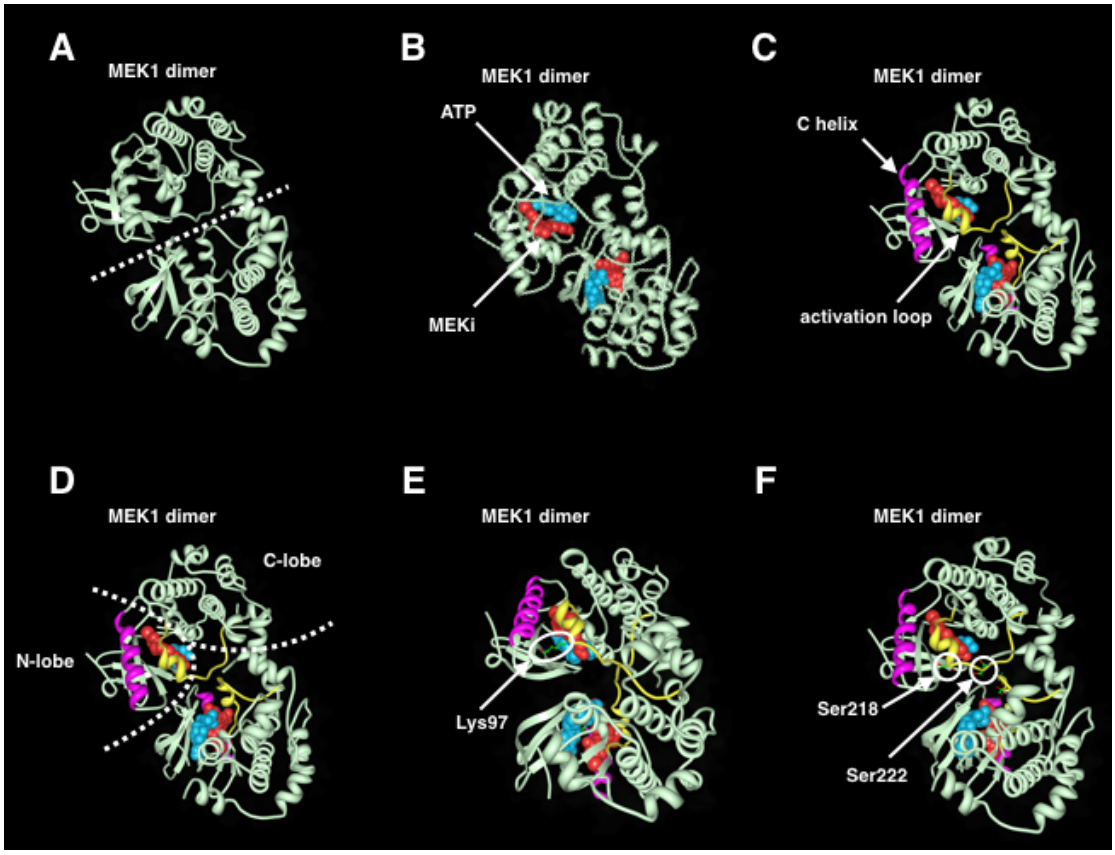

Figure 1.11: Allosteric MEK inhibitor and ATP binding in MEK1 dimer.

(A) shows the crystal structure of the MEK1 dimer, which consists of two identical MEK1 amino acid sequences. (B) labels in blue is ATP, while in red is an allosteric MEK inhibitor. (C) tags the C helix and activation loop. (D) highlights the N-lobe and C-lobe region. (E) shows the lysine at position 97, which separates ATP and the allosteric MEK inhibitor. (F) labels the serine amino acids of the activation loop. Generated with ProteinWorkshop 4.2.0 (J. L. Moreland et al., 2005), RCSB PDB (Berman, 2000) identifier: 3MBL, (Wallace et al., 2010).

Most kinase inhibitors are ATP competitive (Garuti, Roberti, & Bottegoni, 2010), and therefore are in best case, concentration-dependent, but not

allosteric. The unique activation loop of MEK1/2 can be exploited to generate truly allosteric drugs, which are not ATP competitive and only target MEK1/2 (Ishii et al., 2013; Iverson et al., 2009). MEK1/2 is inactivated by the drug-induced conformational displacement of the C helix and the activation loop of MEK1/2 (Ohren et al., 2004).

In particular, MEK1/2 is an interesting drug target due to its substrates. ERK1 and ERK2, which are exclusively activated by MEK1/2, making MEK1/2 the so called 'gatekeeper' of ERK1/2 (Caunt et al., 2015). ERK1/2 is reported to regulate hundreds of other client proteins involved in cell proliferation and cell survival (von Kriegsheim et al., 2009), which therefore increases the interest in MEK1/2 as a drug target.

Another example of an allosteric drug is rapamycin (Benjamin, Colombi, Moroni, & Hall, 2011; Choi, Chen, Schreiber, & Clardy, 1996), which inhibits mTORC1 in an ATP non-competitive manner (H. Yang et al., 2013). Rapamycin is historically interesting given the fact that it was extracted from the bacterium *Streptomyces hygroscopicus* found on the island of Rapa Nui (Seto, 2012), and afterwards led to the scientific discovery of mTOR, which is the abbreviation for the mammalian target of rapamycin (Cheng, Huang, Qiang, Lin, & Demain, 2001). Rapamycin is a showcase of the hypothesised allosteric transition of Monod (Monod, Wyman, & Changeux, 1965).

Kinases account for ~2% of the whole human proteome (Manning et al., 2002). In cancer, the mutational frequency within kinase is 4 fold larger than randomly expected (Futreal et al., 2004). This increased frequency of mutation highlights their importance in cancer, making kinases an interesting drug target.

1.3.2: Cancer pathways

Cells are capable of reacting to their direct microenvironment. Various stimuli are detected via receptors, which are embedded in the cell membrane, and transmit their stimuli through signalling cascades (pathways) to achieve an

appropriate response, e.g. cell growth or apoptosis. Signals in those cascades are transmitted through GTPases (e.g. RAS superfamily) and mostly conveyed by subsequent phosphorylation of kinases (e.g. BRAF, MEK1/2, ERK, etc.), but can also be transmitted through cleavage (e.g. PARP1 cleavage through CASP3), protein-protein binding (e.g. MDM2 and TP53 binding), or other actions.

There are several pathways involved in tumours including the ERK signalling (Dhillon et al., 2007), PI3K-AKT pathway (Liu, Cheng, Roberts, & Zhao, 2009), TGF-β pathway (Massagué, 2008), WNT-β-catenin pathway (Anastas & Moon, 2013), JAK-STAT pathway (Quintás-Cardama & Verstovsek, 2013) and many more.

1.3.3: ERK signalling

One of the key pathways to accelerate cell growth and promote cell survival is ERK signalling. This pathway is activated through ligands binding to RTKs. After ligand binding to RTKs, those receptors recruit the adaptor protein such as GRB2, which consecutively activates the guanine nucleotide exchange factor SOS. SOS displaces GDP from RAS allowing binding of GTP, which thereby activates RAS. GAP reverses the activation of RAS by hydrolysing GTP to GDP. The GTP bound RAS activates RAF isoforms, namely ARAF, BRAF and CRAF. All three isoforms are capable of activating MEK1/2, which can also be triggered by other MAP3Ks, e.g. MEKK1/2/3, MAP3K8, or MLK1/2/3/4. MEK1/2 activates ERK1/2, which further has hundreds of substrates (Figure 1.12) (Caunt et al., 2015).

ERK signalling is often simplified as a cascade of RAS-RAF-MEK-ERK, which causes the impression of a one-way directed signal. However, ERK signalling also has various negative feedback loops, e.g. ERK inhibits MEK, RAF and RAS (Fritsche-Guenther et al., 2011; Sturm et al., 2010). These negative feedback loops increase the robustness of ERK activation (Fritsche-Guenther et al., 2011; Shin et al., 2009). For instance, a decreased stimulus of the ERK pathway also decreases at the same time the negative feedback, which consequently leads to the reactivation of ERK.

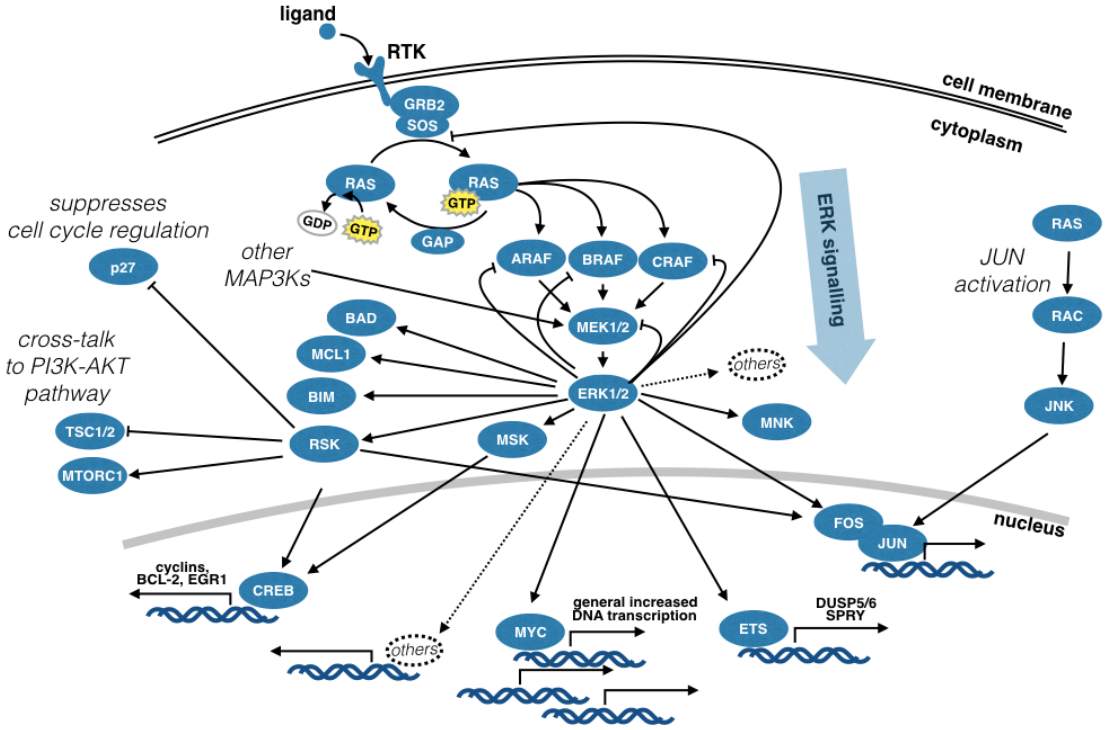

Figure 1.12: ERK signalling.

The above figure shows the ERK signalling, which follows the path of RAS-RAF-MEK-ERK. This cascade is triggered by activation through a ligand binding to RTK, which thereby recruits the adaptor protein GRB2 and activates SOS. Activated SOS is capable of activating RAS by displacing GDP, allowing GTP binding. RAS is deactivated by GAP. There are 3 known isoforms of RAF activating MEK1/2, i.e. ARAF, BRAF, CRAF. MEK1/2 might be also activated by other MAP3Ks, but solely MEK1/2 is capable of activating ERK1/2, making it a 'gatekeeper' of ERK. ERK is a master regulator of hundreds of proteins and here only a few are exemplified. The negative feedback loops of ERK signalling increase the robustness of this pathway.

Not shown in Figure 1.12 is the complex negative regulation of ERK signalling from ERK substrates. For example, ETS is a transcription factor, which causes the expression of *DUSP5*, *DUSP6* and *SPRY*. DUSP5/6 negatively regulate ERK1/2 activity itself (Z. Zhang et al., 2010), while SPRY suppresses RAS activity (Hanafusa, Torii, Yasunaga, Matsumoto, & Nishida, 2004), which is upstream of ERK signalling.

Another substrate of ERK is MSK, which further activates the transcription factor CREB (Arthur, 2008). CREB activation causes expression of *cyclin D*,

which is required for CDK4/6 activity. The cyclin D and CDK4/6 complex is required to escape cell cycle arrest and overcome the G1 checkpoint through phosphorylation of RB (see 1.2.5: Cell cycle, page 9). Thereby, ERK activity drives uncontrolled cell proliferation. The cyclin expression is also driven by ERK-RSK-CREB. RSK also inhibits the tumour suppressor p27 (Fujita, Sato, & Tsuruo, 2003), which thereby promotes uncontrolled cell proliferation. RSK is a substrate of ERK, which also activates PI3K-AKT signalling (Anjum & Blenis, 2008).

MYC is another substrate of ERK, which is a transcription factor that universally enhances DNA transcription (Lin et al., 2012; Nie et al., 2012). MYC is over-activated in many cancers and generally increases the current transcription profile and is a pan-cancer hallmark.

ERK also activates FOS, a component of the FOS-JUN complex also known as AP-1 transcription complex. In parallel, JUN is activated through RAS-RAC-JNK-JUN (Hilfiker-Kleiner et al., 2006). The AP-1 transcription complex is a 'master regulator' of cell proliferation, transformation and apoptosis (Shaulian & Karin, 2002).

Cancer exploits ERK signalling to increase cell proliferation, avoid cell death as well as other hallmarks of cancer. Hyperactivation of this pathway may be achieved through amplifications of *RTK*s, e.g. *ERBB2* amplification in breast cancer (Garnock-Jones, Keating, & Scott, 2010), or amplification of downstream kinases, e.g. *KRAS* or *BRAF* in colorectal cancer (Little et al., 2011).

As an alternative to amplifications of *RTK*s and *ERK* pathway members, other mutations within this pathway aim to decouple from a ligand-induced stimulus and become fully independent from the presence or absence of growth factors. For example, the mutation $EGFR^{T790M}$ causes intrinsic activation of this RTK in non-small-cell lung cancer (Gazdar, 2009). RAS mutations cause intrinsic activation in various cancer types (Prior, Lewis, & Mattos, 2012). The $BRAF^{V600E}$ mutation causes intrinsic activation in melanomas (Chapman et al.,

2011). These oncogenic mutations give an evolutionary advantage and thus are naturally selected. Those alterations lead to 'oncogenic addiction', allowing the development of targeted therapies (Sharma & Settleman, 2007; Torti & Trusolino, 2011; I. B. Weinstein & Joe, 2008).

It is important to note that MEK1/2 is the only activator of ERK1/2, making it a master regulator of ERK1/2. ERK1/2 is a kinase with many substrates such as BAD, MCL1, BIM, MNK, etc. that are involved in cell proliferation, apoptosis, differentiation, transformation and tumorigenesis. This pathway gained further attention due to the availability of allosteric inhibitors for MEK1/2 (e.g. CH5126766) (Ishii et al., 2013), which are not ATP-competitive and therefore highly target specific. As an alternative to the allosteric inhibition of MEK1/2, specific inhibitors against $BRAF^{V600E}$ mutants (e.g. vemurafenib) also became available (Chapman et al., 2011).

1.3.4: PI3K-AKT pathway

In addition to the previously mentioned ERK signalling, here I focus on the PI3K-AKT pathway (Figure 1.13), which is a rapidly emerging target for cancer therapies (Engelman, 2009). There are several isoforms of PI3K and their functionalities are not fully understood (Thorpe, Yuzugullu, & Zhao, 2014), but it is well-known that *PI3K* is a key oncogene in tumour development and maintenance, e.g. cell growth, cell survival, cell migration, angiogenesis, autophagy, endocytosis, etc. A particularly well-studied example is the activation of PI3K through GPCRs (Figure 1.7) or RTKs (Figure 1.8). After these stimuli, PI3K phosphorylates the membrane anchored PIP_2 to PIP_3. PTEN is a tumour suppressor, which reverses the PIP_3 to PIP_2 phosphorylation. PIP_3 recruits AKT (i.e. PKB) and PDK1 at the cell membrane, which leads to AKT activating MTOR driven cell proliferation and promotes cell survival (Mendoza, Er, & Blenis, 2011).

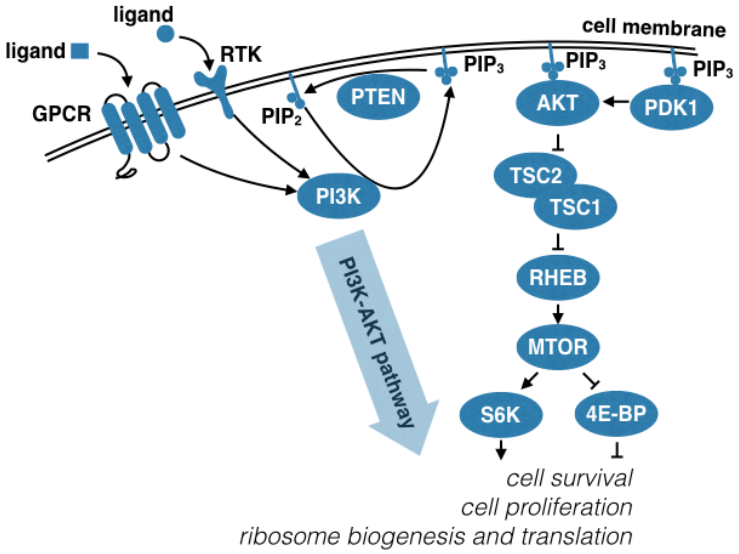

Figure 1.13: PI3K-AKT pathway.

This figure shows the simplification of the PI3K-AKT pathway, which may be activated through receptor tyrosine kinases (RTKs) or the G-protein coupled receptor (GPCR). PI3K phosphorylates PIP_2 to PIP_3, while PTEN reverses the phosphorylation. PIP_3 recruits AKT and PDK1 to the cell membrane, which then inhibits the TSC1 and TSC2 complex, which further inhibits RHEB. RHEB activates MTOR, which then activates S6K and inhibits 4E-BP, which ultimately leads to cell survival and proliferation as well as ribosome biogenesis and translation. Notably, $A \dashv B \dashv C \rightarrow D$ could also be simplified as $A \rightarrow D$, for example $AKT \dashv TSC1/TSC2\ complex \dashv RHEB$ could be modelled as $AKT \rightarrow mTOR$.

As previously mentioned, the PI3K-AKT pathway can also be activated by RAS (Figure 1.3) or crosstalk with ERK signalling (Figure 1.12).

1.3.5: Summary

To summarise, cancer exploits cell signalling to reprogram the cell for tumorigenic behaviour. Cancer modifies cell signalling at the level of signal reception and transduction to achieve the cancer phenotype. Kinases are prominently responsible for the signal transmission and therefore, with higher likelihood, modified in cancer (4 fold larger than randomly expected) (Futreal et al., 2004). In this section, I showed examples of tyrosine as well as serine/threonine kinases. I exemplified cancer signalling with ERK signalling and the PI3K-AKT pathway, which cause accelerated cell proliferation and increased cell survival. Finally, I explained the concept of allosteric drugs in

the context of MEK1/2 inhibitors and put oncogenic addiction in the context of pathways.

1.4: Cancer drug discovery

The first compound that was approved for cancer treatment was nitrogen mustard in 1949. Nitrogen mustard is the main component of mustard gas, which was extensively used in World War II as a chemical weapon. Today, it is still used as chemotherapy to treat Hodgkin lymphoma. Nitrogen mustard is a DNA alkylating agent that causes DNA double strand breaks (Cornell & Blauw, 1949). Several other chemotherapeutic drug classes followed such as microtubule inhibitors (e.g. paclitaxel), antimetabolites (e.g. gemcitabine) or cytotoxic antibiotics (e.g. doxorubicin). The big challenge and disadvantage of these chemotherapies is that they do not distinguish between cancer and healthy tissue, effectively being toxic across all cells in an individual. Chemotherapies are sometimes even carcinogenic (e.g. alkylating agents), however, cancer cells are often lacking various repair mechanisms and dividing faster than healthy cells, which makes them more fragile to chemotherapeutic assaults.

Modern cancer treatment is commonly based on the exploitation of oncogenic addiction, a strong dependency of a tumour on a particular oncogene. For example, melanoma is often addicted to the $BRAF^{V600E}$ mutation, which is heavily reliant on this alteration (Sharma & Settleman, 2007; I. B. Weinstein & Joe, 2008). Oncogenic addictions are the basis of targeted therapies and are nowadays commonly used in clinics, such as ERBB2 (i.e. HER2) inhibitors (e.g. trastuzumab), BCR-ABL inhibitors (e.g. imatinib), PI3K inhibitors (e.g. AZD6482) etc.

Targeted therapies are usually classified as small or large molecules. Small molecules are chemicals that need to enter the cell, while large molecules are antibodies (biological products) that block particular cell receptors on the cell surface. For example, cetuximab and panitumumab are two monoclonal antibodies targeting the EGF receptor (EGFR) on the cell surface, while lapatinib and afatinib are small molecules targeting EGFR and ERBB2 from inside the cell (Modjtahedi, Cho, Michel, & Solca, 2014) and are therefore required to penetrate the cell membrane.

Orthosteric are drugs targeting the active site of a protein and have many undesirable off-target effects. The off-target effects are caused by drug interactions with other proteins with the same conserved active site (Nussinov & Tsai, 2012). The opposite of orthosteric therapies are allosteric treatments, which are not competitive at the active site. Identifying allosteric drugs is challenging since non-active sites within proteins are also less evolutionarily conserved than active sites (De Smet, Christopoulos, & Carmeliet, 2014). Although allosteric drugs are desirable due to fewer off-target effects, most of the current targeted therapies are orthosteric drugs, which are ATP-competitive and thereby in best-case, concentration-dependent.

1.4.1: History of HSP90 inhibitors

Many cancers pursue unsustainable deregulation, which are buffered by chaperones such as HSP90 (see 1.2.8: Chaperones, page 15). Due to the cancer addiction to chaperones, investigations in the mid 90's led to the discovery of geldanamycin, which was the first compound targeting HSP90 and proved high potency *in vitro*, but large toxicities *in vivo* (Uehara, 2003; L Whitesell, Mimnaugh, De Costa, Myers, & Neckers, 1994). Shortly after the ground-breaking work with geldanamycin, an alternative structure called radicicol was discovered to target HSP90 (Schulte et al., 1998). This newer structure showed *in vitro* good potency comparable to geldanamycin as well as less toxicity, but was discontinued due to its chemical stability *in vivo* (S Soga et al., 1999). Geldanamycin and radicicol are the scaffolds for many HSP90 inhibitors currently under consideration in clinical trials.

The first class of HSP90 inhibitors was predominantly derived from the toxic compound geldanamycin. 17-AAG was the first chemical modification of geldanamycin reducing the toxicity and taken into clinical trials, which was discontinued due to small potency in patients or remaining toxicity (Gartner et al., 2012; Heath et al., 2008; Kim et al., 2009). There were some attempts to improve 17-AAG such as 17-DMAG, which increases the water solubility for

oral administration, but ultimately this failed again due to high toxicities (Glaze et al., 2005).

Radicicol was discontinued due to chemical stability. The 2nd generation of HSP90 inhibitors was dominated by resorcinol derivatives, which improved the chemical stability and solved the previous bottleneck of radicicol (e.g. AUY-922, KW-2478, ganetespib, AT-13387). The 2nd generation of HSP90 inhibitors also contained structures, which are neither derivatives of radicicol nor geldanamycin, for example purine related structures (e.g. BIIB-021, MPC-3100, PU-H71) and others (e.g. PF-04929113, XL-888, DS-2248) (Shiro Soga et al., 2013).

The inhibition of HSP90 might not be the cure to cancer due to its 'housekeeping nature', but in combination with other target therapies may overcome drug resistances (Workman, Clarke, & Al-Lazikani, 2016). For example, HSP90 inhibition in combination with hormonal therapy in ER positive breast cancer delayed resistance (Luke Whitesell et al., 2014). A similar result is seen in skin cancer cell lines that developed delayed resistances against the allosteric inhibitor vemurafenib (Smyth et al., 2014).

1.4.2: Personalised medicine and biomarkers

Personalised medicine, also known as precision medicine, is an idea of tailoring and administering drugs based each patient's unique tumour profile. Understanding the concept of oncogenic addictions led to targeted treatments and furthermore to identification of biomarkers, which distinguish responders from non-responding tumours (Torti & Trusolino, 2011).

An example of a resistance biomarker is *TP53* mutations and the MDM2 inhibitor nutlin-3a (Figure 1.14). TP53 regulates initiation of apoptosis, the controlled cell death, and is also involved in the arrest of defective cells in the senescence. This key tumour suppressor gene is mutated in ~50% of all cancers (Toledo & Wahl, 2006). Drugs targeting proteins upstream in the signalling of TP53 are not effective when the function of TP53 is already lost

due to a mutation. For instance, nutlin-3a targets the oncoprotein MDM2, which is known to inhibit TP53 functionality (Vassilev et al., 2004). Nutlin-3a treatment is only useful, as long as the function of TP53 is suppressed by MDM2 and not already lost due to a somatic mutation in *TP53*. Therefore, mutations in the tumour suppressor *TP53* are indirect biomarkers of drug resistance against nutlin-3a (Garnett et al., 2012)

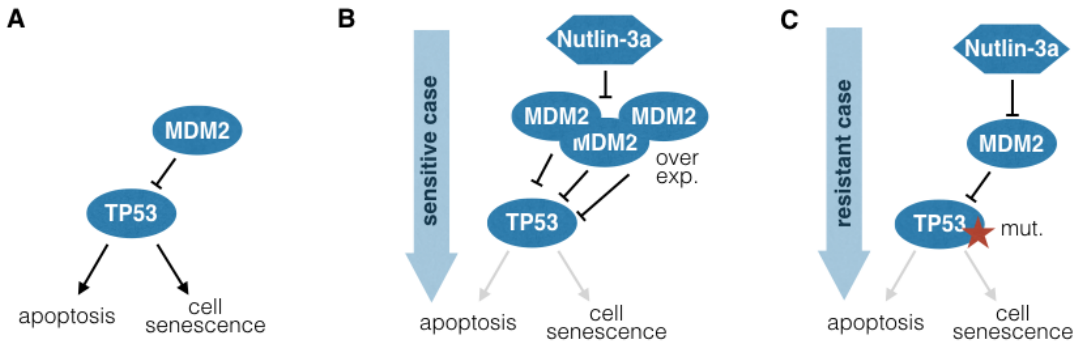

Figure 1.14: Nutlin-3a drug response.

(A) *TP53*s protein product activates apoptosis and cell senescence. MDM2 is inhibiting TP53 functionality. (B) Drug sensitivity to MDM2 inhibition is given when MDM2 is overexpressed and TP53 is a wild type. (C) Nutlin-3a treatment is ineffective if a loss-of-function mutation occurred in TP53. Notably, TP53 mutation and MDM2 overexpression are mutually exclusive, which makes TP53 wild type to an indirect marker of nutlin-3a drug sensitivity.

An example of a sensitivity biomarker is the $BRAF^{V600E}$ and BRAF or MEK1/2 inhibition. The $BRAF^{V600E}$ mutation in melanoma causes a continuous signal of cell growth as previously described. Realising that this particular mutation causes uncontrolled rapid growth, specific BRAF inhibitors such as PLX4720 (i.e. progenitor of vemurafenib) and MEK1/2 inhibitors (targeting downstream of BRAF) have been developed in the past decades. For those drugs the tumour driver $mutation^{V600E}$ in BRAF is a biomarker of drug sensitivity (Chapman et al., 2011; Garnett et al., 2012).

Most approved biomarkers in oncology are enriched for genomic markers (FDA approved drugs, 2015). For example,

- Crizotinib response (MET and ALK inhibitor) is linked to *ALK* rearrangement (Ou, Bartlett, Mino-Kenudson, Cui, & Iafrate, 2012; Soda et al., 2007).
- Dabrafenib sensitivity depends on $BRAF^{V600E}$ mutation (Chapman et al., 2011).
- Nilotinib response is associated with *BCR-ABL* fusion (Druker et al., 2006).
- Lapatinib (small molecule targeting EGFR and ERBB2) sensitivity is associated with *ERBB2* (i.e. *HER2*) amplification (Konecny et al., 2006).
- Afatinib (small molecule targeting EGFR and ERBB2) and erlotinib (EGFR inhibitor) are associated with sensitivity for the activating mutation L858R (i.e. leucine substitutes arginine at codon 858) in EGFR (Gazdar, 2009).
- Cetuximab or panitumumab (both monoclonal antibody targeting EGFR) are associated with resistance to KRAS codon 12 and 13 mutations (Peeters et al., 2013; Yen et al., 2010), which may cause an intrinsic activation of KRAS and by-passes EGFR inhibition. Notably, this strong resistance in KRAS mutant cells to EGFR inhibition is not given in the previously mentioned small molecules inhibiting EGFR (e.g. lapatinib, afatinib and erlotinib), suggesting the presence of some off-target effects of those compounds.

Most biomarkers are genetic mutations, but may also be observed at other molecular levels (i.e. gene expression or methylation), e.g. *ERBB2* has higher expression levels due to genomic amplification of this gene, and therefore responds well to the ERBB2 inhibition (Hanawa et al., 2006).

1.4.3: Drug resistance

Although targeted treatments are very promising tools in the battle against cancer, this disease has a repertoire of mechanisms to avoid drug response. There are known multidrug resistance ATP binding cassette (ABC)

transporters, which actively pump small molecules out of the cancer cell and thereby cause drug resistance (Fletcher, Haber, Henderson, & Norris, 2010).

Cell signalling is not a process that is optimised by an engineer for the best solution, rather a process that is constantly evolving to reach a good solution (Chosed, Mukherjee, Lois, & Orth, 2006). Therefore, in cell signalling there is a lot of redundancy, i.e. different routes to activate the same pathway. In case a chemical compound remains within the cytoplasm (not pumped out), a common mechanism of resistance is pathway rewiring, e.g. the paradoxical activation of CRAF when inhibiting BRAF can activate MEK1/2 in cells with wild type BRAF (Heidorn et al., 2010).

Alternative activation of a pathway can happen through cross talk between different pathways (Gilbert, 2000). For example the cross-activation between the PI3K-AKT pathway and ERK signalling causes resistance to PI3K inhibitors in murine lung cancer with a *KRAS* mutation (Engelman et al., 2008). However, tumour cells with those *KRAS* mutants respond synergistically to the combination of PI3K and MEK1/2 inhibition, which targets both pathways and stops cross activation (Engelman et al., 2008).

Another primary mechanism of drug resistance is a negative feedback signal within a pathway. For example, negative feedback loops from ERK to RAF grant robustness of ERK signalling (Fritsche-Guenther et al., 2011). Targeting multiple members in the RAS-RAF-MEK-ERK signalling cascade decreases the effect of negative feedback loops, e.g. targeting BRAF and MEK improved prognosis in melanoma (Long et al., 2014). Also colon cancer with $BRAF^{V600E}$ is known to be robust to BRAF inhibition through a negative feedback loop to EGFR, while the inhibition of BRAF and EGFR is highly synergistic (Corcoran et al., 2012; Prahallad et al., 2012).

Secondary resistance is an evolutionary response to treatment and may involve additional secondary mutations. For example, cetuximab is an effective antibody for treating EGFR wild type colorectal cancer, but after the first successful treatment KRAS mutations and amplifications cause

secondary resistance (Diaz et al., 2012; Misale et al., 2012). Another example of secondary resistance is the MEK1/2 inhibition with AZD6422 in colorectal cancer cell lines, which trigger resistance through the amplification of $KRAS^{13D}$ or $BRAF^{600E}$ (Little et al., 2011).

Drug resistance may also be achieved through dynamically changing the gene expression, methylation or metabolic profile (Zahreddine & Borden, 2013). For example, the HSP90 inhibitor 17-AAG requires functional expression of *NQO1*, an enzyme that metabolises 17-AAG into its pharmacological active form (Kelland, Sharp, Rogers, Myers, & Workman, 1999). Primary resistance might be achieved by dynamically lowering the expression of *NQO1*, but also through secondary resistances by *de novo* loss-of-function mutations within *NQO1* (Gaspar et al., 2009).

1.4.4: Clonal diversity

The most commonly accepted cause for secondary resistance in tumours are not *de novo* mutations but clonal diversity, i.e. the heterogeneity of cell types and their mutational status in tumours (Hanahan & Weinberg, 2011). Within a tumour there may be a population of cells already carrying resistance mechanisms against the administered drug. Those resistant cells have an evolutionary advantage and ultimately outgrow the sensitive cancer cells. Therefore, what appears to be the most promising targeted treatment initially, might give rise to even more aggressive tumours. Current opinion in cancer research is that a solution is to understand the heterogeneity of the tumour and to use drug combinations to address all types of cells within the tumour at the same time (Al-Lazikani, Banerji, & Workman, 2012).

1.4.5: Biological models

Cancer cell lines are models of real *in vivo* tumours and are controversially discussed in literature. On one hand, they are acknowledged to be the key to modern cell biology, but on the other hand they are different from *in vivo* tumours (Domcke, Sinha, Levine, Sander, & Schultz, 2013; van Staveren et al., 2009).

For generating a novel cell line, first primary cells have to change their molecular machinery to become immortal and evade apoptosis and cell senescence, e.g. through expression of *TERT* and deletion of key regulators such as *TP53* or *RB*, (Maqsood, Matin, Bahrami, & Ghasroldasht, 2013). This is the first step, which makes cell line models different from primary tumours. Notably, a real tumour consists of many different cell types that fulfil various tasks, while a cell line is only a single cell type from this heterogeneous cell population (Meacham & Morrison, 2013).

In 1952, the first cell line in culture was derived from a very aggressive cervical tumour and was named HeLa from the first letters of the name of the patient, Henrietta Lacks (Gey, Coffman, & Kubicek, 1952). Kept in culture for over 60 years, cell lines such as HeLa go through many cell cycles and acquire many additional mutations, which often makes them questionable as suitable human cancer models. Some clones of HeLa express different surface proteins and consequently change their cohesion and adhesion properties, e.g. HeLa Kyoto line clone. As a notable side effect, HeLa Kyoto cell lines have an improved cell adhesion, which particularly increases the attachment to petri dishes. This makes HeLa Kyoto cells a suitable model for confocal microscopy (Yu, Strohmeyer, Wang, Müller, & Helenius, 2015), but at the same time the HeLa Kyoto clones are undoubtedly more different to their *in vivo* origin (Landry et al., 2013).

Primary tumour tissue models are certainly closer to the real tumour biology than immortalised and cultured cell lines (Cree, Glaysher, & Harvey, 2010), however, for technical purposes cell lines are more manageable for large-scale high throughput screens such as from the Genomics of Drug Sensitivity in Cancer (GDSC) project (Garnett et al., 2012) or Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012).

Other alternative models to cell lines for studying cancer are organoids (Matano et al., 2015), which have the advantage of consisting of multiple

cells. However, the limitation of organoids is that they are only available for a limited number of tissues e.g. colorectal or prostate.

Another alternative to cancer cell lines are *in vivo* mice models. For example, studies can be performed with mouse knockouts or xenografts. In a knockout mouse a particular gene or group of genes is genetically modified for studying differences, while xenografts are mice with reduced immune systems and implanted human tumours. Mice models are in some aspects much closer to a real human tumour, but more expensive and cannot be used in high-throughput.

To summarise, cancer cell lines resemble to a certain degree behaviours of real tumours, however, any association found in cell lines needs to be validated in primary tumours. Moving from putative biomarkers in cell lines to markers in patients is a long and complicated process including tests for precision, trueness, limitation of quantification in clinics, sensitivity and clinical implication (Drucker & Krapfenbauer, 2013; Füzéry, Levin, Chan, & Chan, 2013). Notably, the closer the cancer model is to a real human tumour, the more challenging and laborious it is in high-throughput.

1.4.6: Summary

To summarise, cancer drugs can be classified as chemotherapies or targeted therapies. Targeted therapies are based on large or small molecules, which target specific oncogenic addictions, and therefore, are the basis of personalised medicine and biomarkers. However, secondary resistances remain a limitation in these approaches. It is a general observation that the more targeted a therapy, the quicker secondary resistances arise. Therefore targeted therapies are mostly administered in combination with chemotherapies, and the future might be to combine targeted therapies to achieve synergistic response in tumours. Nevertheless, the aim of this thesis is to predict drug responses to initial treatments and identify biomarkers in cancer cell lines, which is the first step towards personalised medicine.

1.5: Pharmacological screens

In the early 90's the first pharmacological high-throughput screen including various cancer types was established. This screen was set up at the National Cancer Institute (NCI) originally including 60 cell lines, which gave it its name NCI-60 (Shoemaker, 2006). One of those 60 cell lines turned out to be a clone of another included in the NCI-60, therefore, it is also often referred to as NCI-59. Nevertheless, this pioneering NCI-60 screen was the first large-scale screen across different cancers (pan-cancer), which is different to cancer-type specific screens, which only focus on a single organ, e.g. breast (Neve et al., 2006) or lung (Sos et al., 2009).

The difference between pan-cancer and cancer-type specific is that a pan-cancer screen allows studying pan-cancer biomarkers, which is not possible with a cancer-type specific one alone. Alternatively to pan-cancer screens, leveraging multiple different cancer type specific screens is challenging due to different experimental setups and batch effects.

1.5.1: GDSC & CCLE datasets

Following the pioneering work of NCI-60, the Genomics of Drug Sensitivity in Cancer (GDSC) project (Garnett et al., 2012) and the Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012) extended their screen efforts to find pan-cancer associations. GDSC screened 714 cells and 138 drugs, while the CCLE contained 504 cell lines and 24 compounds. Both pharmacological screens captured 30 different cancer types and are the core datasets of my thesis.

In addition to the pharmacological screens, both projects characterised the cell lines at a molecular level in the absence of treatment. Both measured copy number variations with Affymetrix SNP6.0 arrays. GDSC detected gene expression with Affymetrix HT-HGU122A arrays, while CCLE used the more advanced Affymetrix U133 plus 2.0 platform. GDSC reported capillary sequencing of 67 tumour drivers and leveraged breakpoint-specific sequence primers for 3 cancer related fusion genes. In comparison, CCLE evaluated the

mutational status of more than 1,600 genes with next-generation sequencing (i.e. targeted massive parallel sequencing), but removed germline variants and confirmed 392 high frequently observed mutated genes with mass spectrometric genotyping including 33 known tumour drivers.

Current extensions of the pharmacological and molecular data of the GDSC project are in preparation (Iorio et al, unpublished). The pharmacological data will increase from 714 cells x 138 drugs to 1,001 cells x 265 drugs. The capillary sequencing data will be replaced by whole exome sequencing from the Illumina HiSeq 2000 platform (European Genome-phenome Archive identifier: EGAS00001000978) (Lappalainen, Almeida-King, Kumanduri, Senf, Spalding, Ur-Rehman, et al., 2015). The gene expression will be updated to the Affymetrix U219 platform (ArrayExpress identifier: E-MTAB-3610) (Brazma et al., 2003). The fusion gene list will be extended by 7 additional cancer related translocations. Furthermore, methylation profiles measured with Infinium HumanMethylation450 v1.2 BeadChip will be added (Gene Expression Omnibus identifier: GSE68379) (Edgar, Domrachev, & Lash, 2002). This expansion of pharmacological data and large extension of deep molecular characterisation is not only an incremental increase in volume of the data, but allows studying associations which may be not have been observable before.

1.5.2: GDSC & CCLE primary goals

The main goal of CCLE and GDSC was to identify biomarkers of drug response, which was enabled through the deep molecular characterisation of cell lines and their large pharmacological screen. The aim was to understand the molecular profile of a cell and based on its profile to predict the drug response. Ultimately, CCLE and GDSC strived to identify the basis of a truly personalised treatment by exploring the landscape of cancer.

CCLE and GDSC were both focusing on targeted therapies that are either in clinical use, clinical trials or an advanced experimental state to increase their potential of immediate clinical application. Both screens also contain

state-of-the-art chemotherapies (cytotoxic compounds), e.g. paclitaxel (J. Zhou & Giannakakou, 2005), which inflict reproductive stress on normal and cancer cells, but ultimately cancer cells are more vulnerable due to their increased growth and mutational burden (Corrie, 2008). Target therapies aim to inhibit certain molecules within a cell and may inhibit oncogenic drivers that do not exist (or of less relevant) in normal cells. The concept of oncogenic addiction drove the development of target therapies (Torti & Trusolino, 2011), which often comes at the price of secondary resistances due to *de novo* alterations or clonal diversity (Greaves & Maley, 2012a). Nevertheless, targeted therapies and their combinations are the basis of modern personalised medicine.

For the task of personalised treatment, a large panel of diverse cell lines is required to capture the heterogeneity of cancer; therefore the number of testable compounds is restricted due to labour and costs. In my thesis I will explore various statistical and machine learning methods to explore the prioritisation of compounds as well as the importance of different features to reduce this dimensionality.

Another goal from both pharmacological screens is to understand drug response in cancer. In my thesis, I will predict drug response, explore good predictive models to reveal the reason for their predictability and ultimately relate it to their respective Mode of Action (MoA).

1.5.3: Limitations of pharmacological screens

The pharmacological data for CCLE and GDSC has been criticised for its poor correlation (Haibe-Kains et al., 2013), which may sound concerning, but is also somehow expected when considering the limitations of CCLE and GDSC. The main reasons for the pharmacological data being poorly rank-correlated are due to (i) the definition of drug sensitivity in cells, (ii) misinterpretation of resistant cell lines and (iii) different experimental setups.

Drug response in cell lines is a relative measurement. For example, a cell line is sensitive if we observe a drug response lower than the maximal tested drug concentration; however, this cell line is only truly sensitive if the remaining majority of cell lines in the screen do not respond. In case all cell lines respond this is unlikely a cure for cancer, it rather indicates a strong cell toxicity, which would most likely also kill the patient: bleach, for example, is not a cure for cancer. Therefore, sensitivity is relative defined to the other cell lines in the drug screen.

Another major reason for poor ranked correlations is the chosen drug concentrations. For instance, GDSC adjusts the maximal concentration for each drug individually, while CCLE choses a fixed concentration. GDSC optimises their screen effort to identify 10-20% of the sensitive cell lines and avoid global toxicities. This highlights that 80-90% are non-responding cell lines which is not necessarily the same as resistance to drug treatment. Cell lines might just not be treated with a high enough concentration to respond. Therefore, we can only make a statement of sensitive and non-responding cell lines at the observed maximal concentration. A ranked correlation would try to correlate 80-90% of noise, which obviously would lead to poor concordance.

A limitation of GDSC is that their screen is optimised for the sensitive area, which potentially increases the direct implication in clinics. For instance, a biomarker matching drug sensitivity identifies the correct patient cohort for successful treatment, while the resistance marker only states ‘do not treat’. The identification of successful treatment is arguably clinically more relevant than excluding treatments. For the later task the GDSC screen is not optimised, which explains poor correlations in the resistant area.

Another challenge is that drug sensitivity is a rather rare event. The overlap of CCLE and GDSC is only 15 drugs and 291 cell lines, which decreases the likelihood of observing overlapping sensitivity. Again, correlating non-responders against each other would be similar to correlate noise, which is a misleading comparison of both screens.

1.5.4: Summary

GDSC and CCLE were praised as the new gateway to battle cancer (J. N. Weinstein, 2012), but unfortunately failed to live up to their expectations (Haibe-Kains et al., 2013). This mismatch of expectation and delivered biological insight is partially the product of not openly sharing the screen limitations. The misinterpretation of the data led to a loss of trust in such large pan-cancer high-throughput screens. A response to the mismatched expectation was to clarify that biological insight such as observed drug sensitivity markers were reproducible in both screens (Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium., 2015). I anticipate that the novel molecular characterisation and extension of pharmacological data will be an important resource for the community and will reveal novel drug associations guiding future drug discovery despite the previous outlined limitations.

1.6: Computational biology of pharmacogenomics

Computational biology is a broad field including various topics. Methods applied to pharmacogenomics could be divided into 'top-down' and 'bottom-up' approaches (Pillay, Hofmeyr, Mashamaite, & Rohwer, 2013; Shahzad & Loor, 2012). These two mindsets describe the directionality of how to explore a complex dataset and derive hypotheses. Top-down approaches explore the drug response data in a systematic manner to identify hypotheses to follow-up. In contrast, bottom-up driven analyses start from an already well-defined model and further integrate observations to extend the current knowledge.

For instance, a bottom-up approach could leverage specific prior knowledge of a pathway to derive a mechanistic model of drug response, while a top-down approach would rather explore the data in a systematic manner with either statistical methods or machine learning. In my thesis, I leverage from both approaches: I start with applying top-down approaches for identifying good predictive models and subsequently explore those good models with a bottom-up approach to understand their MoA.

1.6.1: Statistical framework to identify biomarkers

A statistical approach towards exploring biomarkers is to identify different cohorts of cell lines and test if those different cohorts respond differently to a given treatment. Those cohorts could be defined by the mutational status of oncogenes and ultimately mapped to the drug response of various compounds. Such an oncogene and drug response association could be simply tested with a t-test (Witt & McGrain, 1985) in a cancer-type specific setting.

It is known that tissue label is a good predictor of drug response in a pan-cancer setting (B.-J. Chen, Litvin, Ungar, & Pe'er, 2015), e.g. blood cancer cell lines are generally more sensitive than solid tumours. Any molecular biomarker distinguishing the blood cell lines from solid tumours would perform well, but would be trivial and not clinically applicable.

Therefore, a correction of the tissue type is necessary in a pan-cancer analysis. This is achieved with a generalisation of the t-test, which is commonly known as analysis of variance (ANOVA) (Chambers, Freeny, & Heiberger, 1992). The t-test is able to test the mean difference in drug response based on a categorical variable, i.e. 'mutation' or 'wild type' gene. The ANOVA is generalised to also fit an additional term, the categorical tissue label.

In order to apply an ANOVA model, I make the following 3 assumptions:

- The drug responses are normally distributed.
- The observed drug responses are independent.
- Variances of the different cell line cohorts are equal.

For calculating the ANOVA p-values, I (i) need to define the necessary sum of squares, (ii) afterwards the degree of freedom and (iii) the total sum of squares (TSS) as defined as,

$$TSS = \sum_{i=1}^{N} (\bar{y} - y_i)^2$$

where, N is the number of experimental measured drug responses y and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the average of all drug responses.

The sum of squares for y under consideration of c_0 (i.e. tissue covariate) is defined as,

$$SS_{y.c_0} = TSS - \sum_{t=1}^{T}\sum_{j=1}^{J} (y_{t,j} - \bar{y}_t)^2$$

where, T is the number of tissues I correct for, and J is the number of cell lines within a cancer type context. $\bar{y}_t = \frac{1}{J}\sum_{j=1}^{J} y_{t,j}$ is the average drug response within a tissue context.

The sum of squares of the drug response y is calculated as the following,

$$SS_{y \cdot c_0, x} = \sum_{m=1}^{M} \sum_{j=1}^{J} (y_{m,j} - \bar{y}_m)^2,$$

where, M is stratified based on tissue covariate c_0 and mutational status of the oncogene x.

For example, $x = \{0, 0, 0, 1, 1, 1, 1\}$ and $c_0 = \{a, a, b, b, b, c, c\}$ would lead to the following ordered pairs $\{(a, 0), (a, 0), (b, 0), (b, 1), (b, 1), (c, 1), (c, 1)\} \epsilon M$, where M would be to the following set of unique tuples $\{(a, 0), (b, 0), (b, 1), (c, 1)\}$. M is explored to calculate SS_y.

The sum of squares of mutation x under the consideration of c_0 is defined as

$$SS_{x \cdot c_0} = TSS - SS_{y \cdot c_0} - SS_{y \cdot c_0, x}.$$

After defining the sum of squares $SS_{y \cdot c_0}$, $SS_{y \cdot c_0, x}$ and $SS_{x \cdot c_0}$, I need to identify the degrees of freedom for c_0 (dgf_{c_0}), x (dgf_x) and y (dgf_y).

dgf_x and dgf_{c_0} are the number of unique elements of x and c_0 minus one, respectively. Recapping the previously mentioned example, I explored two mutational states $\{0,1\} \epsilon x$ and three different tissues $\{a, b, c\} \epsilon c_0$, which correspond to $dgf_x = 1$ and $dgf_{c_0} = 2$.

dgf_y is the total number of elements in M minus one, dgf_{c_0} and dgf_x, which would be in our example $dgf_y = 7 - 1 - 2 - 1 = 3$.

After obtaining the $SS_{x \cdot c_0}$ and dgf_x, the mean sum of squares of x corrected for c_0 can be calculated as the following,

$$MS_{x \cdot c_0} = \frac{SS_{x \cdot c_0}}{dgf_x}.$$

The corresponding mean sum of squares is defined as

$$MS_{y \cdot c_0, x} = \frac{SS_{y \cdot c_0, x}}{dgf_y}.$$

The p-value of the ANOVA is calculated with an F-test, which is based on the F-distribution. The F_{value} is calculated as the following,

$$F_{value} = \frac{MS_{x \cdot c_0}}{MS_{y \cdot c_0, x}}$$

The outlined ANOVA is a sequential two-way ANOVA, which first corrects for the tissue covariate and afterwards estimates how much remaining variance can be explained with the mutational feature.

A p-value might be inflated due to the sample size of the analysed groups and therefore is complemented with an effect size, which is independent of sample size, e.g. the p-value can be low in large groups, although the effect size might be relatively small. In a statistical model you ultimately search for small p-values and a large effect size. Here, in this analysis, I used the Cohen's D effect size (Lakens, 2013), which is the mean difference divided by the pooled standard deviation.

As an alternative to the two-way ANOVA, I also applied linear mixed models to identify germline variations associated with drug response (Fusi, Lippert, Lawrence, & Stegle, 2014) (for details see 'Germline model (equation 5-2)'). For this task, I needed to additionally incorporate fixed effects such as population structures, which are known confounding factors in genome-wide association studies (GWAS) (Kang et al., 2008; Stephan, Stegle, & Beyer, 2015). Furthermore, mixed models are capable of incorporating random effects, e.g. noise, which is not separately captured in the ANOVA model.

1.6.2: Machine learning to systematically predict drug response

Machine learning methods have similar applications to statistical methods, but are heuristics validated by their predictability. They could be divided into two major classes: (i) unsupervised and (ii) supervised algorithms (Bishop, 2006; Mitchell, 1997).

An unsupervised algorithm searches patterns in a dataset independently of a phenotype, e.g. clustering algorithms. Usually, after clustering, a phenotype is mapped on the different groups to understand the data driven clustering. Common applications of unsupervised machine learning algorithms are quality controlled. For example, the clustering of cell lines based on their gene expression profile are expected to group by their respective tissue-of-origin (Domcke et al., 2013). If this is not the case a cell line might be wrongly annotated and presumably fail the quality control.

Supervised algorithms aim to learn a phenotype. In my thesis I predicted the drug response of cell lines with supervised training. Supervised algorithms aim to predict the phenotype and prove through their predictability in an independent test set. Such an independent test set could be an external data set or alternatively achieved with cross-validation or bootstrapping (see 1.6.4: Cross-validation and bootstrapping).

1.6.3: Overfitting

Before explaining the key concepts of cross-validation and bootstrapping, their necessity will be exemplified. They are techniques to prevent under- and overfitting. The term overfitting may not sound as negative as underfitting, but both substantially reduce the performance of a supervised algorithm (Figure 1.15).

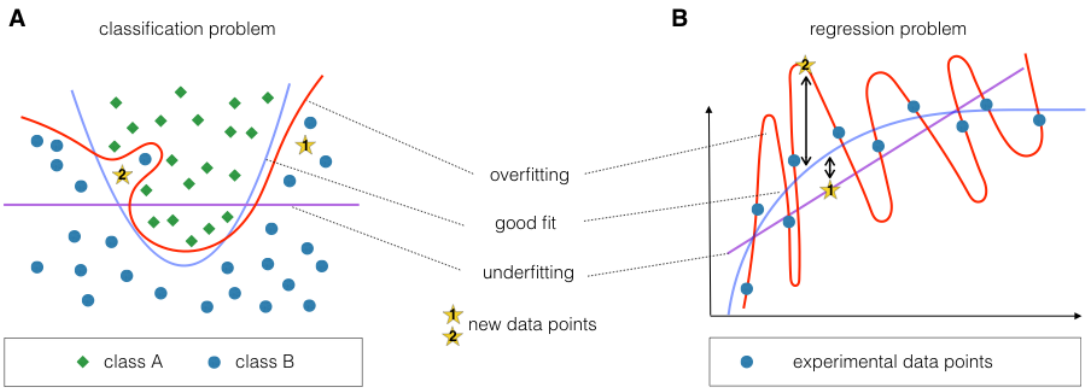

Figure 1.15: Fitting problem for classification and regression.

(A) visualises a 2-class classification problem, where green diamonds and blue dots symbolise the different classes. (B) shows a regression problem, where the aim is to fit the experimental points in blue. The well, under and overfitted models are shown in blue, purple and red, respectively. The stars highlight new data points tested after the initial training.

Machine learning algorithms either aim to classify groups (Figure 1.15 A) or solve regressions (Figure 1.15 B). Both problems may be under- or overfitted, meaning the model parameters are optimised too little or too much. Neither produces a good model and in machine learning the most common mistake is overfitting on the train set, which ultimately shows weak reproducibility of novel predicted data points.

For example, in Figure 1.15 A, the novel predicted point $star_1$ would be wrongly classified as a green diamond by the underfitted model (purple line). This point may be correctly identified by the overfitted model (red line), but $star_2$ would not. However, the well-fitted model (blue line) would identify both novel points correctly.

A similar observation is made in regression problems (Figure 1.15 B). A good fit would decrease the observed error in the predicted versus observed data point, which presumably behaves like the well-fitted model.

1.6.4: Cross-validation and bootstrapping

Cross-validation is a technique to overcome the risk of under- or overfitting (Bishop, 2006; Mitchell, 1997). In cross-validation the original data set is split into equally sized bins, which is called n-fold cross-validation (Figure 1.16). One bin is exclusively dedicated to an independent test and never used in any kind of training, while the remaining n-1 bins are used to train the algorithm. For avoiding an overfitting on the train set, commonly one of those training-bins is exclusively used for cross-training. The cross-training bin is not involved in direct training of the model, but explores the performance of various tested parameters of the model. Usually the best performing parameters on the cross-training set are chosen, and then evaluated on the test set. Therefore, in an n-fold cross-validation $^1/_n$ of the data is used for testing predictability, another $^1/_n$ of the data for cross-training parameters and the remaining $n - {}^2/_n$ data is used for training.

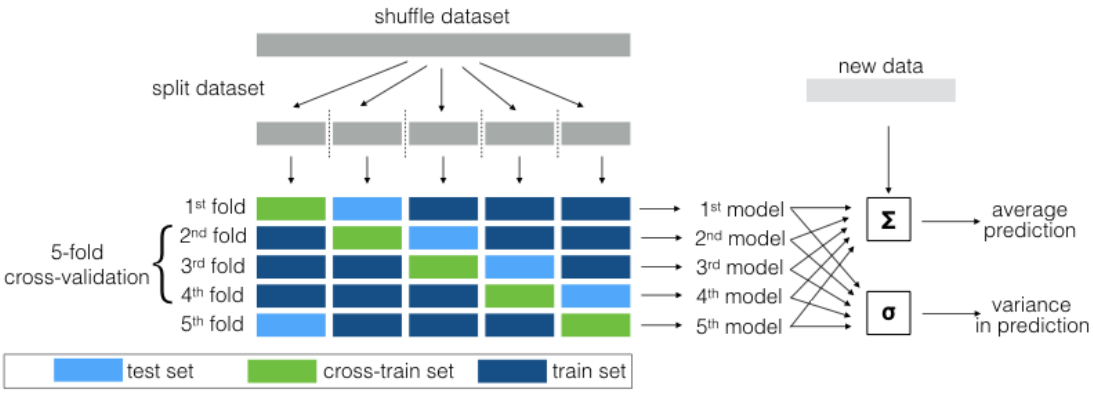

Figure 1.16: 5 fold cross-validation.

Shows a 5-fold cross-validation and how the data is split into different bins. This would result in 5 presumably identical models, but through noise in the data and a finite number of training samples, small variations are expected. Novel data points could be predicted as the average of them, while the variance could also be used to assess confidence in the prediction. The noise in the data would amplify for non-confident predictions, which would be reflected in larger differences between the models.

In an n-fold cross-validation all configurations of test set and cross-train set are tested in a 'round-robin' manner. This results in n different models, which may or may not choose different parameters and lead to slightly different

performances on their respective test sets. In an ideal scenario, all those performances should be identical. However, due to having a finite number of training samples and noise in the data, those models might be slightly different. The variance in cross-trained models could be used to estimate the noise in novel predicted data points and leveraged as reliability / confidence in predictions.

There are no guidelines for how many folds are optimal in machine learning and those most commonly applied are 5-fold, 8-fold or 10-fold. By increasing the fold number, the train set increases, but at the same time the test set decreases. The train set could excessively be increased to the point where only a single data point is left in the test set, which is called leave-one-out cross-validation. Such a leave-one-out validation maximises the available training data, but at the same time performances based on this test set might be overestimated. Therefore, the leave-one-out validation is not recommended.

The split of the bins has an impact on the achieved performances. Bootstrapping could reduce this bias, which selects randomly different bins multiple times. Some samples in the bootstrapping are more often selected than others, but ultimately the reported performance is more robust. Bootstrapping is the superior validation method in terms of estimating the performance and confidence in predictions, but at the same time is computationally more exhaustive than n-fold cross-validation.

The training of machine learning algorithms is comparable to a ‘chicken-egg’ problem, which questions if the egg or the chicken was there first. In this analogy, the machine learning method needs to be trained to make a prediction, but the prediction is required to train the model. This is systematically addressed with cross-validation and bootstrapping.

1.6.5: Machine learning algorithms

Machine learning algorithms could be used for (i) classification or (ii) regression. For example, the binary prediction of sensitive versus non-responding cell lines is a classification, while a regression model would aim to predict the continuous drug response. In my thesis, I predominantly use the latter approach.

Typical machine learning algorithms are either linear or non-linear predictors. Linear predictors are based on linear combinations, which are terms adding up to the expected prediction. A term is defined by a weight and an input (i.e. feature). The weights of this linear combination are adjusted during the training.

Sparse linear regression models are trained with penalisation norms, e.g. LASSO, ridge or elastic net. When case features are correlated with each other, it might not be clear which is the driver gene for explaining drug response. The LASSO regression would arbitrarily pick a representative gene, while the ridge regression would leverage all co-correlated genes with lower weights. Therefore, the ridge regression generally contains more terms with smaller weights comparable to LASSO. *A priori*, it is not known which approach is more suitable for a problem. Therefore, the state-of-the-art approach became the elastic net penalisation, which is a combination of LASSO and ridge regression. Elastic net balances both approaches with an additional mixing parameter.

Likewise sparse linear regression, the previously described ANOVA is a linear model, but only contains two terms: One for the tissue covariate, which is discarded, and the mutational status of an oncogene. Therefore, ANOVA is a univariate model that tests a single gene at a time (tissue is discarded), while the sparse linear regression is a multivariate model including many genes. The advantage of any linear model is its interpretability. By adding multiple terms, the complexity and potential predictive power increases.

Predictive power may even be further increased by applying non-linear methods, which are capable of solving non-linear problems, but are also more prone to overfitting and less trivial to interpret. Comparable to not knowing which penalisation norm might be the best, it can only be evaluated if a non-linear algorithm is necessary to grasp the problem by applying it and comparing the achieved performances with other linear models. In my thesis I explored random forests and neural networks, which were capable of learning non-linear problems; however, no significant improvement to elastic nets was observed.

Other non-linear methods commonly applied are support vector machines (unless a linear kernel is chosen). A kernel transforms the data into another hyperdimensional space, which becomes linearly separable again. Depending on the chosen kernel, a support vector machine might be linear or not. The support vector machine aims to fit a hyperdimensional plane into this hyperdimensional space and maximises the distances to different groups within this space.

Generally, the applied algorithm is not the most important determining factor of performance, rather it is the features that are chosen and how they have been prepared. A common machine learning tenet is ‘garbage in – garbage out’, meaning only informative features will lead to informative models.

1.6.6: Summary

The era of high-throughput biology and big data promotes the importance of computational biology. Top-down methods such as statistical methods and machine learning have been robustly applied to other problems outside the scope of biology and evolved to useful tools in cancer biology. However, the key to derive biological insights is to learn from bottom-up approaches and also systematically combine them with top-down approaches. Particularly, the used features determine achieved performances. My thesis aims to minimise the gap between bottom-up and top-down approaches, by exploring the MoA

of good predictive drug models and identifying the right features for each model.

1.7: Thesis outlook

The poor correlation of CCLE's and GDSC's pharmacological screens split the scientific community about the value of high-throughout screens (Haibe-Kains et al., 2013). I saw the need for investigating this unexpected inconsistency to offer explanations as well as solutions for resolving this discrepancy. I believed this poor correlation was due to experimental noise as well as batch effects. My hypothesis was that this poor correlation should neither affect the biological results nor the clinical implications. To prove both screens' consistencies, I reproduced their biological and clinical implications within their overlap. Particularly, I showed that biomarkers of drug response hold true independently of which screen is explored.

After proving the consistency within pharmacological screens, my aim was to derive guidelines for clinics, which shed light on how the combination of different molecular layers may improve drug response predictions. For this task, I used machine learning algorithms to study molecular layers including genomics, epigenomics (i.e. methylation) and transcriptomics (i.e. gene expression). My hypothesis was that cancer is a genetic disease and therefore drug responses are predominantly driven by genomic alterations, but may be improved by additional information. I systematically analysed those molecular layers within a pan-cancer and cancer-specific setting.

The translation of predictive models to clinical indication of treatment was limited due to the lack of highly predictive models. To bridge this gap and improve the predictive power of existing genomic models, I additionally explored the integration of drug properties. I anticipated that this integration of chemistry and genomics would improve predictive power and enable *in silico* screens for previously untested drugs.

All systematic genomic explorations have predominantly focused on somatic mutations, which accumulate during the lifetime of a patient and ultimately are the drivers of cancer. However, the inherited differences are often neglected in the exploration of drug sensitivity. Germline variances are predominantly studied as a toxicity marker, since each cell of a patient carries the same inherited variants. My hypothesis was that germline variances are not only responsible for a baseline response, but furthermore could be leveraged as an indication of drug sensitivity when studied in interaction with somatic mutations. In the next chapters, I will explore those hypotheses.

Chapter 2: Consistencies in large pharmacogenomic studies

"Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work."
- Daniel B. Wright

2.1: Declaration of contribution

All data for the comparison of the Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012) and the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012) project were published and accordingly referenced. I was not involved in any data generation. I developed the ANOVA source code, implemented it, and carried out all presented analysis. The analysis and results were discussed with Francesco Iorio, Mathew Garnett and Julio Saez-Rodriguez, particularly in the context of a submitted rebuttal from Levi Garraway's and Cyril Benes's groups to "*Inconsistencies in large pharmacogenomic studies*" (Haibe-Kains et al., 2013).

2.2: Introduction

Pharmacological high-throughput screens are used in cancer research, firstly for the identification of potential new treatments, but secondly also for exploring biomarkers of drug sensitivity (i.e. molecular features that are predictive for drug response).

For the first task, usually thousands of chemicals or biological agents are screened against cancer cell lines for identifying so-called hits that either cause cell growth inhibition or reduce cell viability. Afterwards, it has to be determined that the lead compound is non toxic and additionally effective in patients (Conway, Carragher, & Timpson, 2014; Paul et al., 2010).

For the second task of biomarker discovery, pharmacological screens have been set up for various different cancer types, e.g. breast (Neve et al., 2006) or lung cancer (Blanco et al., 2009). Alternatively, screens were also built across various cancer types at the same time (i.e. pan-cancer).

Following the pioneering work of NCI-60 from the late 1980's (Shoemaker, 2006), which screened thousands of compounds against 59 pan-cancer cell lines, the Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012) and Genomics of Drug Sensitivity in Cancer (GDSC) project (Garnett et al., 2012) expanded their screen effort to hundreds of cell lines, while thereby decreasing the number of tested compounds.

The molecular profiles extracted from cancer cell lines display a high degree of heterogeneity. Cancer driver mutations may be rare and hidden in a background of passenger events and noise. This poses a major challenge for the identification of biomarkers of drug sensitivity and resistance, which demands a large number of cell lines for statistical power. The labour and costs of drugs to screen is multiplied with the cell line panel size, which therefore decreases the possible number of compounds to explore with the same budget. For these reasons, GDSC and CCLE focused on clinically approved drugs, in order to maximise their direct clinical impact.

CCLE and GDSC were praised as the new gateway to battle cancer. For instance, John Weinstein said:

"*Both data sets will be made publicly available and will undoubtedly be used by numerous investigators to generate or test their own hypotheses about particular genes, proteins, pathways, cell lineages or drugs*" (J. N. Weinstein, 2012).

Those expectations were lowered by Haibe-Kains study "*Inconsistencies in large pharmacogenomic studies*" (Haibe-Kains et al., 2013), which concluded that the pharmacological data from CCLE and GDSC was not well correlated. Haibe-Kains pointed out that the ranked Spearman correlation coefficient (R_s) of observed drug responses of CCLE and GDSC are mostly weak ($R_s \approx 0.3$) and rarely moderately ($R_s \approx 0.5$) correlated. This observation from Haibe-Kains is clearly contradictory to previous expectations towards CCLE and GDSC.

Besides poor correlation in the pharmacological data, Haibe-Kains observed good concordance in gene expression profiles between both projects. Based on this observation he suggested that a standardisation of pharmacological screen methods might be a solution to increase concordance between CCLE and GDSC, which might be comparable to the previous successful efforts in standardising gene expression (Haibe-Kains et al., 2013).

Notably, pharmacological screens are more complex than measuring gene expression. For example, to measure the response of drug X in cell line Y, drug X needs to be given at various concentrations to estimate the drug response curve. Different target therapies might require different drug concentration ranges to be on target. Different drug classes might require longer treatment periods to show activity, e.g. PARP inhibitors that reduce DNA damage repair require longer treatment periods than drugs targeting ERK-signalling to show effects. The efficacy can be assessed using different experimental assays. A standardisation of pharmacological screens is

desirable, but at this stage might rather limit the potential of new discoveries, since there is no expertise yet which protocols and methodologies are clearly the best practice.

Weinstein also noticed the difference in GDSC and CCLE, and rather suggested considering this as a test of robustness rather than a challenge in normalisation (J. N. Weinstein, 2012). By that he meant that strong biological associations should be observed independently in both screens.

Following up on Weinstein's robustness test and aiming to resolve Haibe-Kain's strong concerns, my hypothesis was that drug sensitivity and resistance markers observed in GDSC should also be observed in CCLE. In this chapter I focused on the comparison of 15 compounds and 292 cell lines screened in CCLE and GDSC, which was the same overlap used in Haibe-Kain's study. My aim was to prove that the pharmacological data is highly concordant when exploring biomarkers of drug response, despite differences in experimental setups.

2.3: Methods

2.3.1: Overlap GDSC and CCLE pharmacological screens

The GDSC dataset contained 714 cell lines and 138 drugs (Garnett et al., 2012); the CCLE database held 504 cell lines and 24 components (Barretina et al., 2012). These two datasets shared 292 cell lines and 15 drugs in common (Figure 2.1 A). The drug response was defined by the IC_{50}, which is the drug concentration needed to reduce cell viability by half (see 2.3.4: Drug response curve fitting and metrics; Figure 2.1 B).

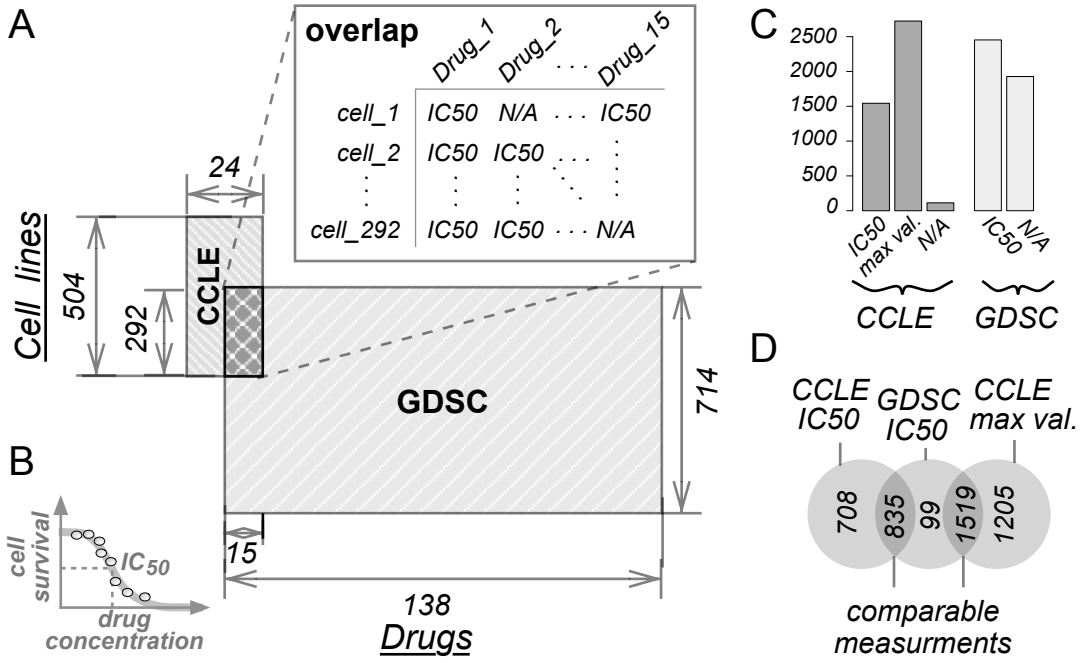

Figure 2.1: Comparison of CCLE and GDSC.

(A) Overlap of CCLE and GDSC. (B) Drug response measured in IC_{50}, which is the drug concentration needed to decrease cell viability by half. (C) Number of cell lines measured in CCLE and GDSC. In CCLE IC_{50}s were capped at a maximal concentration of greater equal $8\mu\mathrm{Mol}$, while GDSC extrapolated IC_{50}s outside of the tested drug concentration range. In both datasets NAs indicated unmeasured or failed quality control. (D) Shows the overlap between comparable IC_{50}s of CCLE and GDSC.

In the overlapping area, CCLE and GDSC held 1,543 and 2,453 IC_{50} values, respectively (Figure 2.1 C). CCLE contained only 113 missing values compared to the 1,927 unavailable ones in GDSC; However, CCLE also

included 2,724 maximum IC_{50} values of 8 µMol, which was the capped maximum concentration for cell lines with no IC_{50} observed in concentration range. In contrast, the GDSC project extrapolated values outside of the measured drug range rather than setting them to a default maximum value which enabled the study of semi-responders, i.e. the drug decreases cell viability within a tested concentration range, but not under 50%. The extrapolation of semi-responders might be useful, but extrapolating completely non-responding cell lines leads to very high IC_{50}values, which are meaningless and potentially a source of misinterpretation.

In the final analysis, the overlap of experimentally measured IC_{50}s of CCLE and GDSC was 835 cell-to-drug combinations. Furthermore, 1,519 max values of CCLE were mapped to measured or extrapolated GDSC IC_{50}s, summing up to 2,354 potentially comparable drug response values.

2.3.2: Cell viability assays

For measuring the drug response in the GDSC screen, cells were seeded in 384-well microtiter plates. For each cell line the seed number was individually optimised to ensure cells entering growth phase until the fixation point of the assay. The assay contained 5% supplements of fetal bovine serum (FBS) and penicillin/streptomycin. FBS was extracted from calf embryos and contains necessary growth factors for cell culture, which are not all identified yet, but essential for cell culture. Penicillin/streptomycin was added to reduce chances of microbial contamination. Cell lines were treated with 9 different concentrations in 2-fold dilution. Notably, also the maximal drug concentration was optimised for each compound, with the aim to be on target while not being generally toxic for the whole cell line panel. The assay was incubated for 72 hours and fixated at the end of this period. After fixation, dead cells were washed off the plate and remaining viable cells at fixation points were DNA stained with SYTO-60. Afterwards, fluorescence intensity of different conditions was measured.

To correct systematic errors from the assay as well as correcting for the background, on each microtiter plate, multiple blank and control wells were measured. Blank and control wells were empty and seeded with cells, respectively. Both types of wells were not treated with any compound.

Cell viability at concentration i was calculated as following:

$$\mathrm{cv}_i = \frac{\boldsymbol{I}_{i,treated} - mean(\boldsymbol{I}_{blank})}{mean(\boldsymbol{I}_{control}) - mean(\boldsymbol{I}_{blank})} \quad \text{(Eq. 2-1)}$$

$\boldsymbol{I}_{i,treated}$ was the measured intensity of well i after 72 hours treatment. $\boldsymbol{I}_{control}$ and $\boldsymbol{I}_{blank}$ were all controls and blanks measured on that particular microtiter plate. The cell viability, $\boldsymbol{cv}_i$, should be in a range from 0 to 1, however, due to noise or the unlikely case that cells benefit from drug treatment, $\boldsymbol{cv}_i$ might be larger than 1.

2.3.3: Experimental pipeline differences in CCLE and GDSC

There were various differences in GDSC and CCLE concerning how the cell viability is estimated, starting with minor differences in the assay. GDSC used 5% FBS and penicillin/streptomycin while CCLE used 10% FBS and 1% penicillin/streptomycin. Furthermore, GDSC operated on 384-well microtiter plates, while CCLE was screened on 1,536-well plates.

Another difference was that GDSC used 9 concentrations at 2-fold dilution, i.e. a concentration range of $\left[C_{max}, {C_{max}}/{2^9}\right]$ where C_{max} was maximal concentration. In contrast, CCLE quantified cell viability with 8 concentrations at 3.16-fold, i.e. a concentration range of $\left[C_{max}, {C_{max}}/{3.16^8}\right]$. The ratio of the CCLE and GDSC concentration range was ${3.16^8}/{2^9} \approx 19$. Therefore, CCLE explored a concentration range that was 19 times larger than GDSC, while at the same time sampling one concentration value less.

Not all parameters of both pipelines could be compared due to lack of documentation (Haibe-Kains et al., 2013), e.g. both screens used different unpublished curve fitting algorithms (see 2.3.4: Drug response curve fitting and metrics).

Haibe-Kains suggested that the main difference was the method to measure the cell viability. As mentioned above, GDSC assessed cell viability through SYTO-60, which releases fluorescence when binding to the DNA of living cells. CCLE measures cell viability with bioluminescent quantitation of intracellular ATP content. In other words, GDSC approximated cell count through DNA while CCLE approximated it through the ATP level. A disadvantage of ATP depending assays is that they underestimate potencies and efficacy of DNA synthesising agents (i.e. antimetabolites) such as gemcitabine or cytarabine (Hatzis et al., 2014)(Chan, Kleinheinz, Peterson, & Moffat, 2013). DNA synthesising agents imitate purine (adenine and guanine) or pyrimidine (cytosine and thymine) structures, which are the building blocks of the DNA, and ultimately stop the normal synthesis in the S-phase of the cell cycle. This had less impact on the ATP level than on the DNA staining, and therefore ATP depending assays tended to underestimate the potency of DNA synthesising agents. ATP is an approximation of metabolic activity, which was previously also linked to chemoresistance in colon cancer (Y. Zhou et al., 2012). Such metabolic dependency would affect more the bioluminescent quantitation of intracellular ATP content than the DNA staining, thereby potentially confounding results.

Despite those differences, another main discrepancy that Haibe-Kains did not consider is that GDSC optimised their maximum concentration (C_{max},) for each compound, while CCLE explored a fixed range from 8 µMol to 2.5 nMol across all agents (Figure 2.2).

For example, in extreme cases such as paclitaxel, which was highly cytotoxic, GDSC set the concentration range from 0.1024 µMol to 0.4 nMol, while CCLE remained at the fixed larger range of 8 µMol to 2.5 nMol. Therefore, in CCLE

more cell lines would be more likely reported as responders compared to GDSC, since CCLE also experimentally explored higher concentrations.

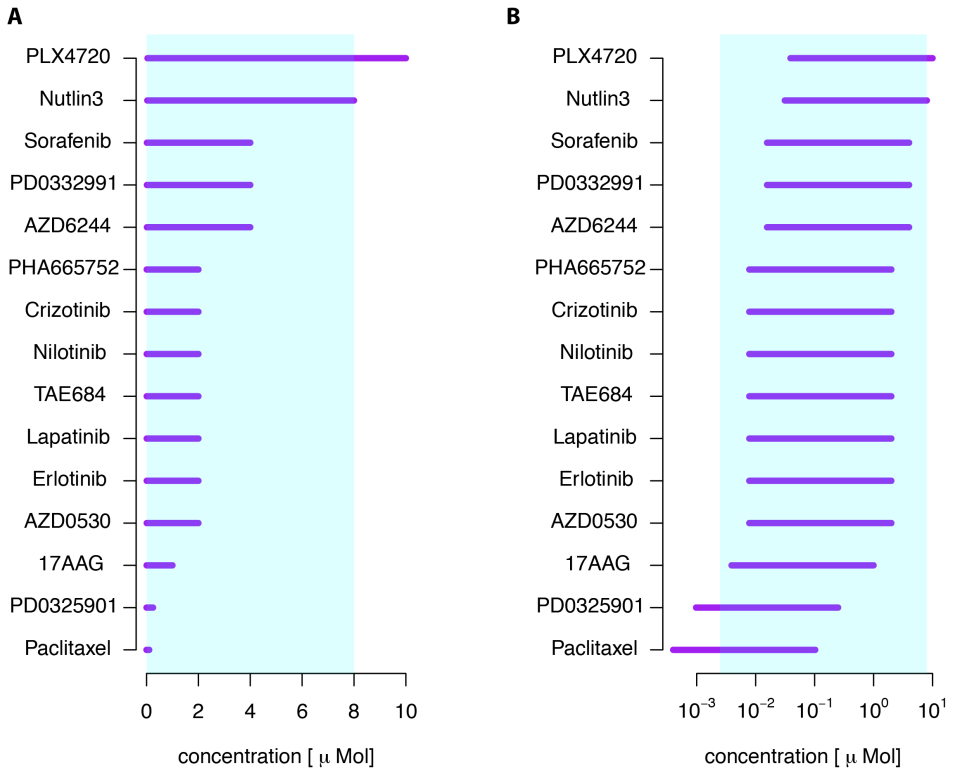

Figure 2.2: Concentration ranges CCLE and GDSC.

CCLE and GDSC had different tested drug concentration ranges. Drugs in CCLE were always treated in a range from 2.5 nMol to 8 µMol (in light blue), while drug ranges for compounds in GDSC were dynamically adjusted suited to each agent (purple lines). (A) and (B) were drug response ranges in normal and log10 scale, respectively.

2.3.4: Drug response curve fitting and metrics

In a pharmacological screen a robust drug response is usually fitted with a sigmoid function (see equation 2-2).

$$\boldsymbol{y} = res(\boldsymbol{x}) = \frac{o-\kappa}{1+e^{-\tau\,(x\,-\varpi)}} \qquad \text{(Eq. 2-2)}$$

$\boldsymbol{x}$ was a vector that contains the treated drug concentrations, while $\boldsymbol{y} = res(\boldsymbol{x})$ was the corresponding vector with estimated cell viability in percentage. All other parameters were scalar values to be estimated during curve fitting. o

was the origin of the drug response curve, which should be 100%, but could vary depending on noise, errors in the control measurement or the unlikely case that a cancer cell benefited from the treatment and therefore had a more rapid cell growth. κ was the absolute decrease in cell viability. The drug response saturates at $1 - \kappa$ percentage and these remaining cells became resistant to further treatment. This saturation point was called E_{max} (Figure 2.3 A), although this point was rather theoretical since at a high enough concentration every treatment would be lethal, citing Paracelsus from the 15th century: "*Sola dosis facit venenum – The dose makes the poison*". $\boldsymbol{\tau}$ was the slope of the sigmoid function (Figure 2.3 B). ϖ was the inflection point of the sigmoid function, i.e. the point with a gradient equal to zero. If $\kappa = 1$ than the inflection point would be equal to the IC_{50} value, which was the required concentration to reduce cell viability by half (Figure 2.3 C).

The area under the drug response curve (AUC) was the integral from the minimum treated concentration (C_{min}) to maximal concentration (C_{max}), which was normalised into a range from 0 to 1 (Figure 2.3 D). This was achieved by equation 2-3.

$$AUC = \frac{\int_{C_{min}}^{C_{max}} res(\boldsymbol{x})}{C_{max} - C_{min}} \qquad \text{(Eq. 2-3)}$$

Commonly reported is the area above the curve, which was $1 - AUC$ (Figure 2.3 E). In contrast to AUC, a large $1 - AUC$ value corresponded to a stronger drug response.

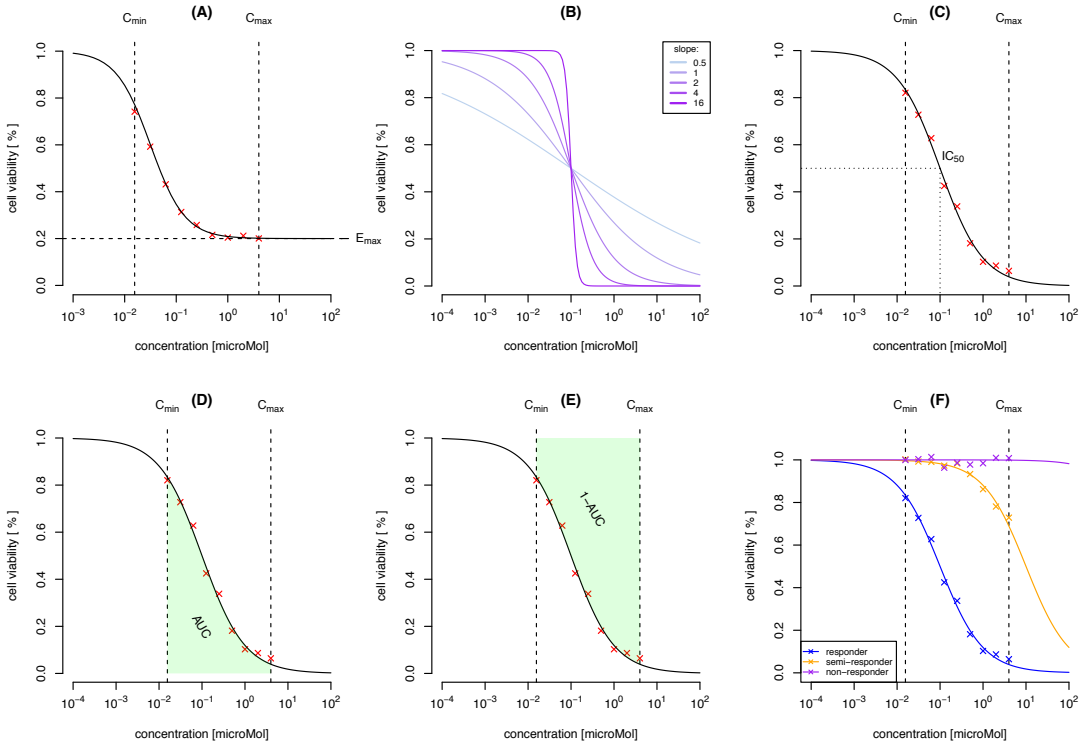

Figure 2.3: Drug response curve fitting.

Shows the sigmoid drug response curves. On the x-axis are the concentrations in µMol. On the y-axis are cell viability in percentage. Crosses mark the 9-measured concentrations with 2-fold dilution. $\mathrm{C_{min}}$ and $\mathrm{C_{max}}$ are the minimum and maximum concentration, respectively. (A) Example of $\mathrm{E_{max}} = 1 - \kappa = 0.2$. (B) Shows slopes τ equal 0.5, 1, 2, 4 and 16. (C) Example of $\mathrm{IC_{50}}$, which is the concentration to reduce cell viability by half. (D) and (E) is the area under the curve (AUC) and the area above the curve ($1 - \mathrm{AUC}$), respectively. (F) Drug response of three different cell lines, while one cell line responds, one semi-responds and another does not respond in blue, orange and purple, respectively.

2.3.5: Definition of resistance and sensitivity in cell lines

Measured drug response metrics as shown above were AUC, $1 - AUC$, $\mathrm{E_{max}}$, slope or $\mathrm{IC_{50}}$. By measuring those metrics, it was challenging to estimate if a single cell line was sensitive or resistant without prior knowledge. However, measuring multiple cell lines and comparing their responses in relation to each other could solve this problem. For instance, a responding cell line required less drug concentration than a non-responding one to reduce cell viability (Figure 2.3 F), and therefore was more sensitive in relation to the other cell lines.

As previously mentioned, in the GDSC screen the range of drug concentrations was adjusted for each compound so that approximately 15 - 20% of the whole cell line panel responded while the remaining ones were non-responders. This adjustment ensures that the drug is on target, while not being toxic for all cell lines. Cell lines that did and did not respond were called sensitive and resistant, respectively. However, the use of the term 'resistant' was inaccurate, since it was only confirmed that those 'resistant' cell lines were not responding at the given concentration.

IC_{50} values carried information about the concentration and were absolute measurements. If an IC_{50} was an 'absolute measurement' of given concentration, why is this information not directly translatable to sensitivity or resistance?

This was only possible for well-studied compounds where prior screens defined the concentration range including response level, but this was not given for all experimental agents. For experimental drugs without known effective concentrations, the mapping to resistance or sensitivity was not possible.

It must also be noted that there was no direct link between concentrations in cell lines and patients. For example, the required drug concentration in a patient depended on its ADME, i.e. adsorption, distribution, metabolism and excretion. Furthermore, the necessary concentration might be influenced by factors such as tumour density (Kobayashi, Takemura, & Ohnuma, 1992). Those influences were not observable in cell lines and therefore cannot be modelled. This implied that for many compounds no conclusion could be drawn towards sensitivity in clinics.

2.3.6: Comparing IC_{50} and 1-AUC

The most commonly used metrics of drug response are IC_{50} and $1 - AUC$ values. IC_{50}'s had the advantage of retaining its relationship to the original given drug concentration, while this was lost when using $1 - AUC$ values.

However, IC_{50} values faced challenges when they were outside of the concentration range of a particular compound. Such out-of-range IC_{50} values either need to be capped or extrapolated. Both options might cause artefacts, which further could cause biases in the analysis.

The advantage of $1 - AUC$ was that those values seemed to be very robust when modelling drug sensitivity. Furthermore, $1 - AUC$ values neither required extrapolation nor capping values, but as previously mentioned the relationship to original drug concentration was lost.

2.3.7: CCLE and GDSC curve fitting differences

In the GDSC screen the drug response curve was fitted with a Markov Chain Monte Carlo (MCMC) algorithm, with the assumption that $o = 1$ and $\kappa = 0$ (see equation 2-2). In essence, the GDSC curve fitting ignored E_{max} values and forced the sigmoid function to start at 100% cell viability and ultimately converged to 0% cell viability. The MATLAB implementation written by Greenman assumed two additional data points to the experimental measured concentrations for the fitting. The first data point was the concentration at 1,000 fold smaller than C_{min}, which was set to 100% cell viability, and the second point was the concentration at 1,000 fold larger than C_{max} was forced to be 0% cell viability. Also as previously stated, this assumption was only partially correct since cells might converge at a theoretically E_{max} value, although ultimately the whole cell population would be erased with enough drug treatment.

The CCLE curve fitting was very different. Only in 5% of the cases a sigmoid function was fitted and CCLE considered E_{max}. In case the cell lines were complete non-responders they fitted a constant model of $o = 1$, alternatively they used a non-parametric spline interpolation. A chi-squared test determined which of the three fitted models was chosen.

Another major difference between the CCLE and GDSC curve fitting was that GDSC extrapolated values, while CCLE capped outliers. CCLE reported only values measured within the observed drug concentration or alternatively the minimum or maximum value.

2.3.8: Biomarker definition and GDSC genomic dataset

The main application of CCLE and GDSC was identifying biomarkers of sensitivity. A biomarker was based on basal features of cell lines and their response to different treatments. For example, a biomarker might be a mutation of gene X in cell line Y that was causing sensitivity to drug Z, and therefore Y should be treated with Z.

A biomarker translatable to clinics required basal information of the cell lines and not before and after treatment differences. Before and after treatment ratios were often applied in other projects to derive drug response signatures, e.g. the Connectivity-Map approach (Lamb et al., 2006). Furthermore, such before and after treatment features were not feasible to measure across large cell line panels and many treatment conditions such as those provided from CCLE and GDSC. Nevertheless, the scope of CCLE and GDSC was the exploration of basal biomarkers and not drug signatures.

For the purpose of biomarker detection, I extracted the mutational information from the GDSC project (Garnett et al., 2012), and mapped them to the 292 overlapping cell lines with CCLE. The genomic basal features were generated with capillary sequencing and SNP 6.0 Affymetrix arrays to determine mutations and copy number variations, respectively for 68 oncogenes. Additionally, 3 fusion genes *BCR-ABL*, *EWSR1-FLI1*, and *MLL-AFF1* were measured with breakpoint-specific sequence primers and added to the oncogene list, making 71 oncogenes in total.

A requirement for running the statistical analysis of biomarkers was to binarise the oncogenes into mutant (=1) or wild type (=0). For fusion genes, this binary encoding was straightforward: 1 or 0 for fused gene or wild type, respectively. For mutations, I only considered genes as mutant if they were predicted to not be a passenger (i.e. non-synonymous SNP that were part of the list of non-passengers provided by COSMIC). This binary encoding of oncogene mutants ignored the fact that different mutations within an oncogene might have different impacts on functionality and hence on the drug response;

however, due to sparseness of mutational events across those 68 oncogenes I rather grouped them together in order to decrease the number of statistical tests and thereby improve the statistical power. Copy number variations were binary encoded as wild type if a gene had between 1 or 7 copies, and as mutant if the gene had 0 copies (deletion) or more than 7 copies (amplification). Deletions and amplifications could be grouped together as the 'mutant' group, since they were mutually exclusive, meaning a segment could be either deleted or amplified, but not both at the same time. Defining more than 7 copy numbers as 'amplified' was chosen based on previous recommendations of the GDSC project (Garnett et al., 2012). The impact of which copy number was considered as amplified might also depend on each oncogene independently. However, for the same reason as for the mutations (number of tests and statistical power) I used the suggested copy number threshold of >7 for amplifications.

2.3.9: Analysis of variance (ANOVA)

For identifying biomarkers I performed an analysis of variance (ANOVA) in pan-cancer, as it was proposed in "*Systematic identification of genomic markers of drug sensitivity in cancer cells*" (Garnett et al., 2012). Briefly, I searched for genomic descriptors that stratified cell lines for differential drug response with previously removing the impact of tissue specificity.

In the ANOVA, I considered the mutational status as input and tissue-of-origin as covariate. Tissue type on its own was a powerful predictor of drug response, which is a trivial feature in clinics since the primary tumour location is usually known. Moreover, several oncogenic mutations were enriched in different tissue types, e.g. in the cell line panel KRAS was enriched in pancreas and gastrointestinal (GI) tract to 92% and 43%, respectively. Hypothetically, if we would ignore the tissue as a covariate, false biomarkers might be identified, which would trivially reflect the tissue-of-origin. For this purpose, it was crucial to first remove all variance explained by tissue type and afterwards explore the remaining explanatory power of mutations.

I used the 'aov' function from R package 'stats' version 3.1.2 with the following formula (see equation 2-4).

$$\mathbf{res} \sim \mathbf{tissue} + \mathrm{gene} \qquad \text{(Eq. 2-4)}$$

res was the drug response, while gene was a binary value for mutant or wild type. **tissue** was a covariate, e.g. breast or lung. 'aov' performed a sequential ANOVA and firstly fitted tissue, and then the mutational status towards explaining the drug response.

2.3.10: Correlations metrics

Correlations were measured in various alternatives. Two commonly applied metrics are Pearson ($\mathrm{R_p}$, see equation 2-5) and the Spearman correlation ($\mathrm{R_s}$, see equation 2-6).

$\mathrm{R_p}$ captured linear relationships of two vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, containing continuous values, while $\mathrm{R_s}$ was a rank based metric depicting monotonic relationships. $\mathrm{R_p}$ was strongly influenced by strong outliers, while $\mathrm{R_s}$ debilitates the impact of outliers.

$$R_p(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{x})(\boldsymbol{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{y})}} \qquad \text{(Eq. 2-5)}$$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ denoted the mean observed value for vector $\boldsymbol{x}$, while $\bar{y}$ was the corresponding mean value for vector $\boldsymbol{y}$.

The Spearman correlation is a variation of the Pearson correlation, which is defined as the following:

$$R_s(\boldsymbol{x}, \boldsymbol{y}) = R_p(rank(\boldsymbol{x}), rank(\boldsymbol{y})) \qquad \text{(Eq. 2-6)}$$

For $\mathrm{R_s}$, the continuous elements of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ were transformed into rankings, e.g. $rank([0.2,\ 7.25,\ 5.1,\ 3.1,\ 0.1]) = [2,\ 5,\ 4,\ 3,\ 1]$. Thus, Spearman correlations allowed testing for monotonic relationships between the variables.

2.4: Results

2.4.1: Spearman versus Pearson & IC_{50} versus 1-AUC

Haibe-Kains and Quackenbush suggested in their paper to apply the Spearman correlation (R_s, see equation 2-6) of overlapping drug responses in CCLE and GDSC (Haibe-Kains et al., 2013). R_s captured monotonic behaviours of the drug responses, which naturally would be expected when comparing replicates. However, those drug responses generated from CCLE and GDSC cannot be considered as replicates due to different screen pipelines, experimental set-ups, drug concentration ranges, curve fittings, etc.

Another important fact was that GDSC optimised their screen so that ~15-20% of the cell lines were responders, while the remaining cell lines were resistant or non-responders at a given dosage. Assuming those ~80-85% of cell lines behaved monotonically was a false assumption and therefore sorting those resistant cell lines was meaningless. However, a more relevant measurement was the linearity of GDSC versus CCLE drug response. In other words, could we capture outliers? Outliers, particularly highly sensitive cell lines, were truly positive responders, which might lead to sensitivity biomarkers. Linearity and outliers could be captured with a Pearson correlation (R_p, see equation 2-5; Figure 2.4 A).

Furthermore, Haibe-Kains and Quackenbush proposed in their analysis to focus on IC_{50} values rather than $1-\mathrm{AUC}$ values. $1-\mathrm{AUC}$ values would have the advantage that non-responding and resistant cell lines would converge towards 0, while semi responders and truly sensitive cell lines would achieve $1-\mathrm{AUC}$ values towards 1. By using $1-\mathrm{AUC}$ values, many of the cell lines that could not be clearly classified as resistant or non-responders would be lumped together as resistant, i.e. resistant outliers would receive less weight. Notably, resistant outliers were most likely the product of artefacts and should be discarded. R_p depicted the linear behaviour of outlying sensitive cell lines in $1-\mathrm{AUC}$ while there were no resistant outliers. The improvement of

consistency between CCLE and GDSC through $1 - \mathrm{AUC}$ compared to IC_{50} could be seen in Figure 2.4 A versus Figure 2.4 B.

Although $1 - \mathrm{AUC}$ values lost the relation to the administered concentration compared to IC_{50} values, the $1 - \mathrm{AUC}$ metric focused on the scientifically interesting sensitive area, since those cell lines could hint at successful treatments.

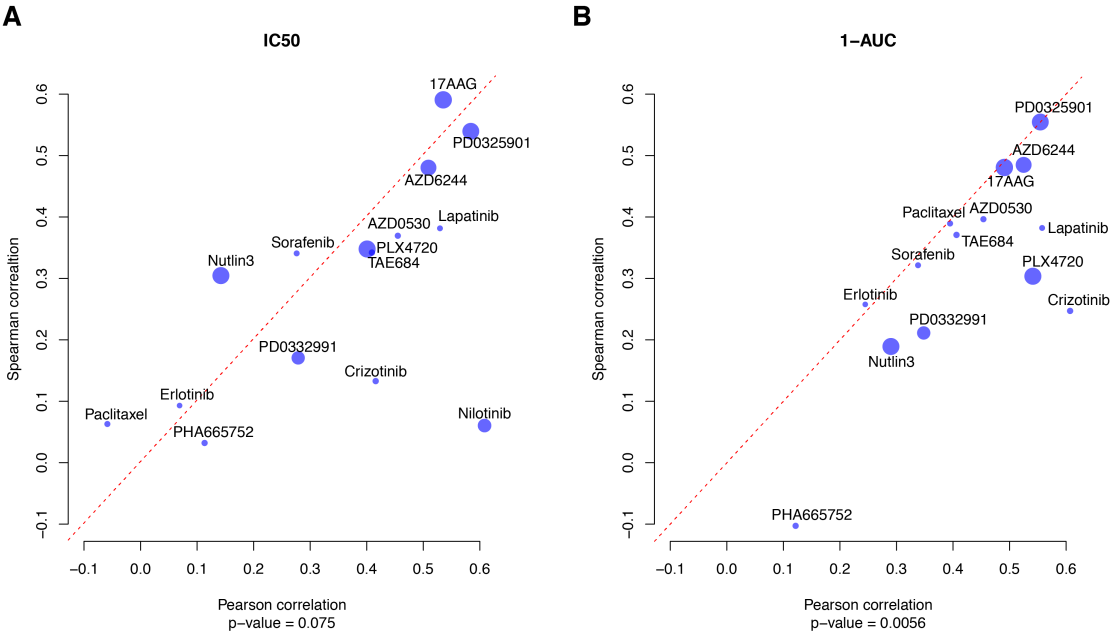

Figure 2.4: Spearman versus Pearson & IC_{50} versus 1-AUC.

Comparison of overlapping compounds in CCLE and GDSC. On the x-axis is the achieved Pearson correlation ($\mathrm{R_p}$), while on the y-axis is the corresponding Spearman correlation ($\mathrm{R_s}$). (A) and (B) are based on IC_{50} and $1 - \mathrm{AUC}$ values, respectively. The p-value under each panel is the paired t-test between achieved performances.

To summarise, there are several limitations in CCLE and GDSC (e.g. 'resistant' cell lines were not distinguishable from non-responders), which were not adequately discussed in their original publication. However, with the right metric for capturing drug response (1-AUC) and the appropriate correlation metric (R_p) the consistency in both screens could be increased.

2.4.2: Biomarker identification of CCLE and GDSC

The aim of both screens was to identify biomarkers of drug response. Mathew Garnett compared the conclusions of both original studies for the 15 overlapping drugs. Many of the biomarkers were independently identified across both studies (see Table 1).

Drug	Therapeutic Target	Clinical indication	Clinical / Pre-clinical Biomarker	GDSC study	CCLE study
Lapatinib	HER2 (ERBB2)	breast	HER2 amp.	✔	✔
PLX4720	BRAF	melanoma	BRAF mut.	✔	✔
Crizotinib	ALK MET	NSCLC	EML4-ALK MET amp. HGF exp.	✔ ALK exp.	✔ HGF exp.
Nilotinib	ABL	CML AML	BCR-ABL	✔	✔ CML lineage
Erlotinib	EGFR	NSCLC	EGFR mut.		✔
AZD0530	SRC-family ABL		BCR-ABL	✔	✔ CML lineage
PD0332991	CDK4 CDK6		CDKN2A del. RB del.	✔	✔
PD0325901	MEK1 MEK2		BRAF mut. NRAS mut.	✔	✔
AZD6244	MEK1 MEK2		BRAF mut. NRAS mut.	✔	✔
Nutlin-3a	MDM2		TP53 mut.	✔	✔
17-AAG	HSP90		NQO1 exp.	✔	✔
TAE684	ALK		EML4-ALK	✔ ALK exp.	✗
Sorafenib	PDGFRA PDGFRB KDR KIT FLT3	kidney liver	FLT3 mut. KDR mut.	✔	✗
PHA665752	MET		MET amp. HGF exp.	✗	✗
Paclitaxel	Microtubules	ovarian breast NSCLC	none	NA	NA

Table 1: Identified biomarkers in the original CCLE and GDSC study.

Extracted biomarkers from Barretina et al 2012 and Garnett et al. 2012 for the set of 15 overlapping drugs. This table and comparison of the original publications was produced by Mathew Garnett.

Those concordances were independently observed in the CCLE and GDSC study by leveraging all cell lines, i.e. for CCLE and GDSC contained 504 and 714 cell lines, respectively. As a complementary analysis, I focused only on the 292 overlapping cell lines for the purpose of showing concordances even within this much smaller subset. Therefore, I reproduced the ANOVA analysis of the original GDSC study, to identify genomic biomarkers in CCLE and GDSC, while only considering the overlapping cell lines of both screens.

Strong associations of drug sensitivity or resistance with a genomic marker were often reproduced in both screens, even with the reduced set of 292 overlapping cell lines (Figure 2.5).

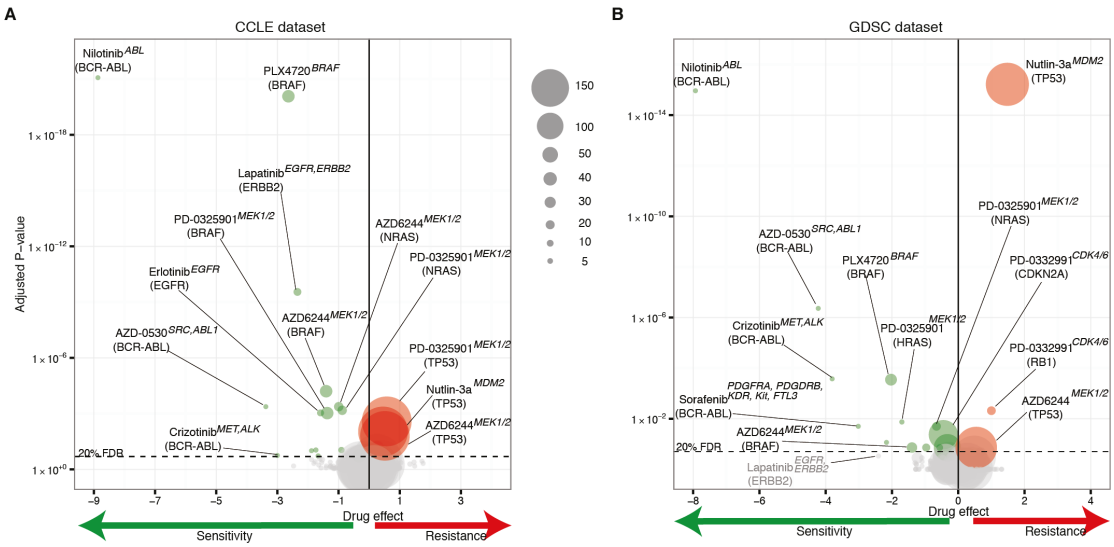

Figure 2.5: ANOVA results CCLE and GDSC.

Volcano plots show biomarkers of drug response estimated with the analysis of variance (ANOVA). For this analysis only the overlapping cell lines in CCLE and GDSC were considered. (A) and (B) show the results obtained with CCLE and GDSC drug response, respectively. On the y-axis is the significance of an association, while on the x-axis is the effect size in Cohen's D. Negative effect sizes point at sensitivity while positive ones indicate resistance. The circle size corresponds to the number of cell lines harbouring the mutation. The biomarker label is the drug name followed by the putative target as superscript and in brackets the associated oncogene.

For example, the clinical biomarker in CML (chronic myeloid leukaemia) and AML (acute myeloid leukaemia) of drug response to the ABL inhibitor nilotinib

was *BCR-ABL* fusion (Druker et al., 2006). This association was highly significant and with a strong effect size in both CCLE and GDSC.
Also AZD-5030 (another ABL inhibitor) was significantly associated with the *BCR-ABL* fusion in both screens.

Another clinical association of drug sensitivity was PLX4720 (BRAF inhibitor), which was the progenitor of vemurafenib (PLX4032), with BRAF mutations in melanoma (Chapman et al., 2011). This clinical biomarker was also independently recapitulated in CCLE and GDSC.

As previously reported, the MEK1/2 inhibitor AZD6244 was consistently associated with BRAF mutations, which was not surprising since MEK1/2 is downstream of BRAF (see page 8, Figure 1.3).

Moreover, in both screens concordance was observed in identifying the association of NRAS mutants with PD-0325901, which was another MEK1/2 inhibitor. Notably, in GDSC the drug PD-0325901 was also associated with mutations of HRAS, which is another RAS isoform with the potential to activate the same downstream kinases (Parikh, Subrahmanyam, & Ren, 2007).

In addition, not yet clinically approved associations were consistently identified. For instance, crizotinib was a MET and ALK inhibitor, where the clinically approved biomarker was the EML4-ALK fusion gene in NSCLC (Kwak et al., 2010). The EML4-ALK fusion gene was not part of the GDSC genomic dataset and therefore cannot be identified in this study, but interestingly in both screens BCR-ABL mutant cell lines were significantly (20% FDR) sensitive to crizotinib.

Consistently identified resistance markers were TP53 mutant cell lines treated with either the MEK1/2 inhibitor AZD6244 or MDM2 inhibitor nutlin-3a. Interestingly, the association of AZD6244 and TP53 was consistent in CCLE and GDSC, but not an approved clinical marker yet.

Furthermore, there were clinically approved associations that were at least identified in one of both screens. For example, *ERBB2* (*HER2*) amplification in breast cancer was a marker of sensitivity to lapatinib (Konecny et al., 2006), which was a ERBB2 inhibitor. This association was significantly identified (20% FDR) in CCLE, while just being under the 20% FDR threshold in GDSC.

Another biomarker that was significantly identified in CCLE but not in GDDC was the EGFR inhibitor erlotinib, which was associated with the sensitivity of EGFR mutant cell lines in NSCLC (Pérez-Soler et al., 2004; Y. Wang, Schmid-Bindert, & Zhou, 2012).

GDSC but not CCLE recapitulated the association of RB mutations inflicting resistance to PD0332991, which was a CDK4/6 inhibitor. As discussed earlier, RB is key regulator of the cell cycle, which is a tumour suppressor. CDK4/6 promotes the phosphorylation of RB, and thereby blocks the tumour supressing function of RB (Konecny et al., 2011). However, if RB functionality was already sufficiently neutralised through a loss-of-function mutation, a CDK4/6 inhibition would have no positive treatment effect. Therefore, RB was a marker of resistance to PD0332991.

Paclitaxel was a chemotherapy, which was a not targeted therapy. Although applied in various cancer types such as NSCLC, ovarian and breast cancer, no clinical biomarker exists. This was also concordantly observed in both screens where no biomarker for paclitaxel was assigned.

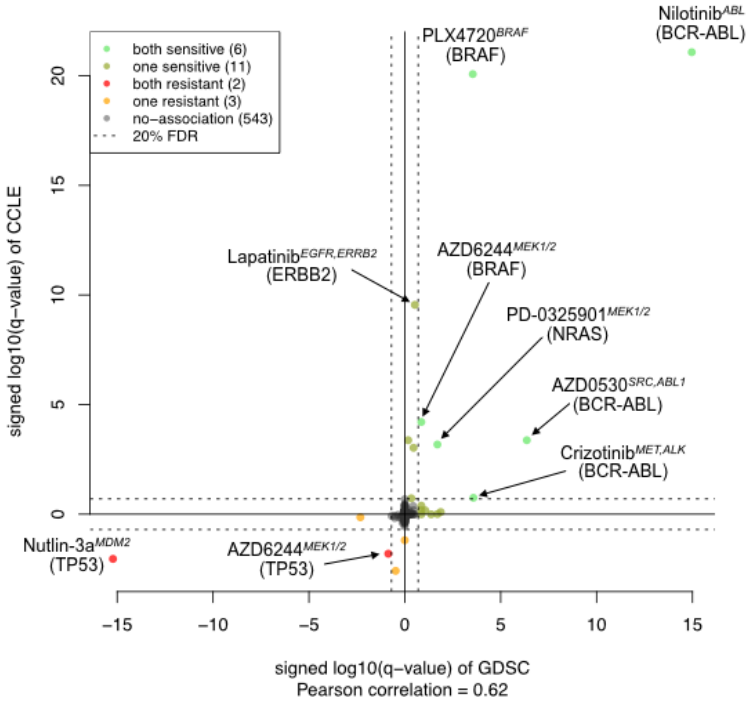

Figure 2.6: Concordances of CCLE and GDSC in biomarker space.

Comparison of signed log10 (q-values) of CCLE and GDSC, where the sign was derived from the effect size, e.g. negative log10 (q-values), corresponded to resistant markers, while positive signed log10 (q-values) were sensitivity markers. Q-values were derived from the ANOVA test described above. Those signed q-values achieved a Pearson correlation of 0.62.

To summarise, while only analysing the 291 overlapping cell lines from CCLE and GDSC, large concordance between the genomic biomarkers were found with an ANOVA analysis. I significantly identified 6 sensitivity and 2 resistance markers (Figure 2.6; 20% FDR). Not concordantly, I found 11 sensitivity and 3 resistance markers in at least one of the screens, however, in all those cases the effect size pointed in the same direction. Also concordantly, 543 associations were significantly discarded in both screens. Besides sorafenib, PHA665752 and drugs relying on *EML4-ALK* fusion or NQO1 expression (which were both excluded in the mutational data), at least one screen identified the clinical genomic biomarker. Furthermore, potentially novel associations were concordantly reported in both screens, e.g. *BCR-ABL* fusion as a sensitivity marker of crizotinib.

2.5: Discussion

2.5.1: Exploratory power and standardisation

The conclusion of the Haibe-Kains and Quackenbush study was that a standardisation would be necessary between different pharmacological screens for increasing the correlation of drug responses (Haibe-Kains et al., 2013). This conclusion was chiefly driven by the fact that they found the gene expression data of CCLE and GDSC highly correlated, and reasoned that this was due to previous efforts for developing robust gene expression platforms and standardisation towards reproducibility. Therefore, Haibe-Kains and Quackenbush also suggested a similar standardisation of pharmacological screens, although they did not offer a solution of how to achieve this.

Pharmacological response metrics, e.g. IC_{50} or $1 - \mathrm{AUC}$ values, were based on various data points, i.e. differently treated concentrations, controls and blanks. Expression of a particular gene might be estimated through a single probe on a chip, which simplified the experimental setup of gene expression compared to drug response metrics. Therefore, the standardisation of gene expression was simpler than of drug screens, although remaining a desirable goal.

CCLE and GDSC were the first pharmacological screens that incorporate hundreds of cell lines. They made various different methodological decisions regarding experimental set-up, chosen drug concentrations and computational assessment. Differences in methodology could be seen as a test of robustness in biological findings (J. N. Weinstein, 2012). Some of those methodological steps might be better in one or the other screen, however, without exploring those steps we would limit ourselves to potentially suboptimal ways of performing pharmacogenomic screens. Therefore, premature standardisation of not yet fully explored choices might limit advancement in technology and scientific discoveries. In the long run, a standardisation is desirable, but with current knowledge this task is not feasible.

2.5.2: Limitations in CCLE and GDSC not enough stressed

Independently, CCLE and GDSC reproduced biological biomarkers of drug sensitivity as previously shown in this study (see 2.4.2:Biomarker identification of CCLE and GDSC). However, both original studies failed to explicitly point out their limitations. As a result, Haibe-Kains and Quackenbush's study assessed both screens in an unsuitable manner, since they disregarded key limitations and differences.

For example, the GDSC screen was optimised for 15-20% of the cell lines responding to a treatment to avoid widespread toxicities. This means that the remaining cell lines were non-responders or resistant in a rather undefined state. This behaviour would be appropriately evaluated with an outlier approach (see 2.4.1:Spearman versus Pearson & IC_{50} versus 1-AUC). However, Haibe-Kains and Quackenbush chose to use a ranked based metric, which also leveraged 80-85% noise in their benchmark.

It had to be understood that both screens produced potential leads of biomarkers, which had to be further validated. To cite John N. Weinstein: "*… robust observations will often be reflected in data from both. Therefore, the differences in methodology can be considered either a disadvantage or a test of robustness.*" (J. N. Weinstein, 2012).

2.5.3: Science communication of reproducibility

This chapter was not only scientifically motivated, but also influenced by poor communication in science. All three studies were published in the high prestige journal Nature, the original CCLE and GDSC papers back-to-back in 2012 as well as the Haibe-Kains and Quackenbush's study in 2013. It was somewhat controversial to first publish two screens that reflect similar biological insights and afterwards report them as inconsistent.

This could only be hypothesised and was not based on measurable facts, but there was a movement at Haibe-Kains and Quackenbush's publication time, which tried to improve reproducibility in science. The Economist published the

following statement, "*Is science wrong? Human after all*" on the 17th of October 2013 and "*Unreliable research – Trouble at the lab*" shortly after on the 19th of October 2013. Those two publications in The Economist started a snowball effect that damaged the perception of sciences in public and the trust of funding agencies. They also had further implications in the scientific community, e.g. delays of future data releases.

All in all, academic science is not perfect and not solely measurable in the resulting publications. A lack of communication and openness about limitations are challenges to advance as a scientific community, particularly, since prestigious journals are more inclined to publish exciting breakthroughs.

2.6: Conclusion

Despite the small overlap as well as different experimental set-ups, pipelines, and drug concentration ranges, etc., large consistencies were found in the sensitive area when considering $1 - \mathrm{AUC}$ and the Pearson correlation.

Particularly, the Biomarker analysis highlighted the large concordance in significantly identified biomarkers. 6 sensitivity and 2 resistance markers were significantly reproduced in both screens. In 14 instances, one of the screens identified a marker, but the other screen at least identified the same directionality in effect size, even if not significant. Large concordance was also seen in significantly rejecting biomarker associations. The Pearson correlation in signed log10 (q-value) space was 0.62, reflecting the strong concordance in the biomarker space.

In summary, strong drug biomarker associations (such as *BCR-ABL* fusions causing sensitivity to nilotinib treatment) were reproduced across both screens, which ultimately supported the robustness of CCLE and GDSC pharmacological data.

Chapter 3: Investigating the contributions of different molecular features to predict drug response

"*An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*"

- John Tukey

3.1: Declaration of contribution

Under the lead of Ultan McDermott and Mathew Garnett within the scope of the Genomics of Drug Sensitivity in Cancer (GDSC) project, all wet lab experiments were independently carried out by them and I share no contribution. Additionally, the gene expression data was produced from GDSC in collaboration with Petra Ross-MacDonald, Heshani Desilva and Aiquing He at Bristol-Myers Squibb (BMS). The Manel Esteller group from Bellvitge Biomedical Research Institute (IDIBELL) generated the methylation dataset. Mainly, Francesco Iorio, Emanuel Gonçalves, Daniel J Vis and Graham Bingell curated and prepared the data for this study. Particularly, Graham Bingell pre-processed the raw sequencing data and performed the variant calling for mutations and copy number variations. Francesco Iorio, Emanuel Gonçalves processed the methylation data set, while Francesco prepared the gene expression. Daniel J Vis performed the drug response curve fitting. The Nuria Lopez group from ICIQ shared an extended oncogene list for this publication. All the machine learning models in this chapter were solely my contribution. Furthermore, my work involved presenting the prediction results, visualisation, biological interpretation and the gained clinical insights from the machine learning. Those analyses were part of a GDSC resource update, which currently is in preparation for submission.

3.2: Introduction

Predicting the drug response from basal molecular features has several applications. The main scientific goal of such drug response predictions aligns with the general goal of the CCLE (Barretina et al., 2012) and GDSC (Garnett et al., 2012; W. Yang et al., 2013) project, which was to stratify patients into responders versus non-responders depending on their cancer profile.

CCLE and GDSC are resources for pharmacological high-throughput screens, as well as repositories of basal molecular data of the cell lines including genomics (i.e. copy number variations and mutations) and transcriptomics (i.e. gene expression). This enabled the development of various prediction methods leveraging the molecular features. For example, methods that utilise genomics (i.e. copy number variations and mutations) (Garnett et al., 2012; Menden et al., 2013), or were based on gene expression (Dong et al., 2015; Geeleher, Cox, & Huang, 2014; Khan et al., 2014) or integrated both molecular layers (Barretina et al., 2012; Fang et al., 2015; Garnett et al., 2012; Gönen & Margolin, 2014; Jang, Neto, Guinney, Friend, & Margolin, 2014). There is a trend towards integrating both layers and additional prior knowledge (e.g. pathway information) to obtain the best possible prediction.

With the aim of obtaining the best possible prediction from various molecular layers, in the scope of the National Cancer Institute – Dialogue on Reverse Engineering Assessment and Methods (NCI-DREAM) challenge, 44 research teams predicted drug sensitivity of 31 anonymous drugs in 53 breast cancer cell lines (Costello et al., 2014). All cell lines from the challenge were fully molecularly characterised with the following basal layers: whole exome sequencing, copy number variations (CNVs), reverse phase protein arrays (RPPAs), methylation, gene expression arrays and RNA-seq. The best performing methods leveraged all feature layers (Costello et al., 2014).

In contrast to the NCI-DREAM goal, my aim was not to obtain the best possible prediction, but a good one that is clinically applicable. In clinics it is not feasible to screen a patient with all imaginable technologies available with

the hope of obtaining a minimally improved prediction. The aim of this chapter was to produce predictions that are applicable for clinics through using as few data layers as possible, but as many as necessary.

By the same token to achieve a more clinically relevant outcome, in this chapter the analysis was limited to three different molecular layers:

i. Genomics (i.e. copy number variation and mutations)
ii. Epigenomics (i.e. methylation)
iii. Transcriptomic (i.e. gene expression)

Genomics was the union of copy number variations (CNVs) and mutations, since both alterations could be derived from the same technology, i.e. whole exome sequencing (Magi et al., 2013).

Following the pioneering work of the breast cancer NCI-DREAM challenge and in order to understand the contribution of each layer, here I analysed 1,001 cell lines derived from 30 cancer types treated with 265 compounds.

The aim of this chapter was to identify which molecular features were most important for determining drug response in cell lines. Cancer is a genetic disease and therefore my hypothesis was that the drug response was predominately driven by genomic features. However, the degree of contribution from epigenetics and transcriptomic had to be determined. Was it necessary to measure all layers, or were they redundant or might they complete each other and could lead to better understanding drug sensitivity? The aim of this analysis was to explore the contribution of different molecular features for predicting drug response and ultimately deriving guidelines that would be translatable to clinics.

3.3: Methods

3.3.1: Drug response dataset

The drug response dataset was an update of the original GDSC dataset (Garnett et al., 2012), which was produced by the Ultan McDermott and Mathew Garnett lab in Sanger. The enlarged dataset contained 265 compounds and 1,001 cell lines across 30 different tissue types. In total, 214,490 drug response values had been measured, which corresponded to ~81% of all drug to cell line combinations experimentally determined. For this analysis, I chose the $1 - \mathrm{AUC}$ as metric to capture drug sensitivity (see 2.3.4: Drug response curve fitting and metrics).

3.3.2: Genomic dataset

The genomic dataset contained mutations and copy number variations (CNVs) of the 1,001 cell lines. The McDermott and Garnett groups generated mutational and CNV information with whole exome sequencing (Agilent SureSelect / Illumina) and SNP6.0 affymetrix microarrays, respectively. The raw BAM and CEL files wee available as a download from the European Genome-phenome Archive (Lappalainen, Almeida-King, Kumanduri, Senf, Spalding, ur-Rehman, et al., 2015) (https://www.ebi.ac.uk/ega/studies/EGAS00001000978).

Briefly, Graham Bingell called missense mutations with cancer variants through expectation maximisation (CaVEMan) (Stephens et al., 2012; Varela et al., 2011) and nucleotide insertions/deletions with the Pindel algorithm (K. Ye, Schulz, Long, Apweiler, & Ning, 2009). In order to exclude germline and passenger mutations, he considered two filters: (i) Graham removed any SNPs from dbSNP (Sherry et al., 2001) and (ii) discarded common variants with minor allele frequency (MAF) larger than 40% in the 1,000 human genome project (Abecasis et al., 2012). Those filtered somatic variants of the whole exome sequencing were made available on the Cancer Genome Project webpage (http://cancer.sanger.ac.uk/cosmic) (Forbes et al., 2014).

Additionally, to focus on drivers and further discard passenger mutations or germline variants, I used a list of high confidence tumour drivers guided by the IntOGene pipeline (Gonzalez-Perez et al., 2013; Gundem et al., 2010), which defined oncogenes based on prior knowledge and mutational frequencies (e.g. tumour drivers have a higher likelihood to be mutated). Those oncogene lists were generated for pan-cancer as well as cancer-specific settings by the Nuria Lopez lab in collaboration with the GDSC team (Rubio-Perez et al., 2015).

Copy number alterations were called with Analytical Multi-scale Identification of Recurring Events (ADMIRE) (van Dyk, Reinders, & Wessels, 2013) and filtered by Daniel Vis for recurrently altered chromosomal segments. Daniel called and derived those segments in a cancer-specific as well as pan-cancer setting.

Mutations and copy number variations were complemented with an additional set of known fusion-genes related to cancer, e.g. *BCR-ABL* fusion. The GDSC group used break-point specific sequence primers to detect chromosomal rearrangements for identifying fusion genes.

In the presented analysis a gene is either considered as mutant or wild type.

$$\mathrm{gene}_{\mathbf{i}} = \begin{cases} 1 & , \text{if del} \vee \text{ampl} \vee \text{mut} \vee \text{fused} \\ 0 & , \text{if wild type} \vee \text{filtered e.g. passenger} \end{cases} \qquad \text{(Eq. 3-1)}$$

At first it might be unintuitive to combine different mutations as one input feature, but different mutations within the same tumour driver had mostly the same effect (tumour suppressor or oncogene), except passenger mutations that had no functional impact. This was a simplification and assumption of our models, which ignored the impact of cell state.

In the case of a gene being a tumour suppressor, the objective of cancer would be to disable the functionality of this gene. Various mutations might achieve this goal, e.g. homozygous deletions or loss-of-function missense

mutations. Such different mutations were merged to one feature, since their biological objective was the loss-of-function of the tumour suppressor.

For oncogenes, only certain mutations might increase the fitness of the tumour, which would be naturally selected, e.g. $BRAF^{V600E}$ in melanoma intrinsically activates the ERK signalling, which ultimately leads to increased cell proliferation (Chapman et al., 2011) provided that key ERK-responsive tumour suppressors have first been inactivated. In the case of an oncogene, the kind of mutation matters. Due to selective pressure, cancer favours the tumour beneficial mutations, which would be reflected in their increased frequency across different patients (Stratton, Campbell, & Futreal, 2009). Therefore, also in the case of oncogenes, different mutations in a gene could be united as one feature.

Tumour drivers had mostly either tumour suppressing or enhancing capabilities, which justified the above described encoding. However, there were various exceptions where a driver behaved as either an oncogene or tumour suppressor depending on the cell state and specific alterations.

For example, as previously described, the melanoma $BRAF^{V600E}$ usually has an oncogenic functionality, while within melanocytic nevi based on interaction with $p16^{CDKN2A}$ the functionality of $BRAF^{V600E}$ is tumour suppressing. In the melanoma the $BRAF^{V600E}$ mutation causes cell proliferation, while in the melanocytic nevi the consequence is cell senescence (Gray-Schopfer et al., 2006; Michaloglou et al., 2005; Romagosa et al., 2011).

Another example would be TP53, which is generally perceived as a tumour suppressor. TP53 is mutated in 50% of all human tumours and is a key regulator of apoptosis and cell senescence. TP53, however, was shown to also act as an "*oncogenic transcription factor*" based on specific mutations (Strano et al., 2007), e.g. $TP53^{R273H}$ has tumour promoting capabilities (W. Wang, Cheng, Miao, Mei, & Wu, 2013). $TP53^{R273H}$ suppresses the expression of miR27a by blocking its promoter region. miR27a has been reported to supress EGFR activation, which subsequently reduces the ERK-signalling.

Therefore, $TP53^{R273H}$ indirectly promotes cell proliferation by means of reducing the regulation of EGFR and ERK-signalling.

Another example of a gene being either a tumour suppressor or an oncogene within different cell states would be RB. As previously shown, RB is a tumour suppressor due to stopping the progression from the G1 to S-phase; however, RB has anti-apoptotic functionalities when caspase-cleavage resistant through mutations in the C-terminus in the intestine (Borges et al., 2005; Goodrich, 2006). This anti-apoptotic behaviour is tumour promoting in the intestine. Therefore, RB can be either a tumour suppressor or oncogene depending on the cell state.

The exceptions mentioned above are not modelled by my encoding and the biological signal might be diluted in specific cases; however, it has to be noted that those are exceptions. Separating all mutations as different features would also increase the sparseness in the genomic dataset and ultimately increase the challenge to build predictive models. It further has to be noted that by including the gene expression profiles in the models, the cell state is captured (described below), which overcame the lack of cell state in the genomic dataset and thereby enabled the modelling of those exceptions by a combination of genomics with gene expression profiles.

3.3.3: Methylation dataset

The Esteller group from the IDIBELL institute generated the methylation dataset with Illumina Infinium HumanMethylation450 v1.2 BeadChip. Emanuel Gonçalves and Francesco Iorio in collaboration with the Esteller group carried out the data processing of the methylation data.

In DNA a Cytosine nucleotide could be followed by a Guanine nucleotide, which forms a Cytosin-phosphate-Guanine (CpG) site. A CpG site could be either methylated or unmethylated. The methylation status $\boldsymbol{beta}_i$ of a CpG site i was captured with the ratio of methylated probe intensity and overall intensity (Du et al., 2010):

$$\boldsymbol{beta}_i = \frac{max(\boldsymbol{y}_{i,methylated},0)}{max(\boldsymbol{y}_{i,methylated},0) + max(\boldsymbol{y}_{i,unmethylated},0) + \alpha} \quad \text{(Eq. 3-2)}$$

The measured intensity of the methylated and unmethylated CpG site i was $\boldsymbol{y}_{i,methylated}$ and $\boldsymbol{y}_{i,unmethylated}$, respectively. The $\max(\boldsymbol{y}, 0)$ function set negative values to 0, which were artefacts from the background adjustments. α was an Illumina recommended offset with default at 100.

Several CpG sites might reside in close proximity and formed a so-called CpG island. I used the UCSC (University of California, Santa Cruz) genome browser (Kent et al., 2002) definition of CpG Islands and focused on informative islands. An CpG island was considered as informative if across all measured cell lines the island showed bimodal behaviour, or put differently, if it was a mixture model of two Gaussian distributions. Emanuel Goncalves identified informative islands with the R package 'mixtools' version 1.0.3 (Benaglia, Chauveau, Hunter, & Young, 2009).

For those informative CpG islands two states were considered for each individual cell line. Either the Island was hyper-methylated or not, corresponding to one or zero:

$$\boldsymbol{CpG}_{i,island} = \begin{cases} 1 & ,if \ \frac{\sum_i^{n_{island}} \boldsymbol{beta}_i}{n_{island}} > 80\% \\ 0 & ,if \ \frac{\sum_i^{n_{island}} \boldsymbol{beta}_i}{n_{island}} \leq 80\% \end{cases} \quad \text{(Eq. 3-3)}$$

3.3.4: Gene expression

The GDSC team in collaboration with BMS generated the gene expression with the Affymetrix Human Genome U219 Array Plates. Briefly, Francesco Iorio read the chip description file with the R-package 'makecdfenv' (Irizarry, Gautier, Huber, & Ben, 2006) and afterwards normalised the gene expression with Robust Multi-array Average (RMA) normalisation by using the R-package 'affy' (Gautier, Cope, Bolstad, & Irizarry, 2004).

The raw transcriptomic data was made available in ArrayExpress (Parkinson et al., 2007) (E-MTAB-3610).

3.3.5: Elastic net

Along with my goal to obtain a reliable prediction, but not necessarily the best, the NCI-DREAM challenge proved that the chosen method had marginal impact on the predictive performance. For example, in NCI-DREAM ~2/3 of the participants would either use some sort of non-linear or sparse linear regression which lead to equivalent predictive power than more complex models (Costello et al., 2014). Therefore, I chose as a representative model the elastic net (see 3.3.5: Elastic net), a state-of-the-art linear regression.

Elastic net (Hui Zou, 2005) is a linear regression method, which balances LASSO (least absolute shrinkage and selection operator) and the ridge (i.e. Tikhonov) regularisation term with the mixing parameter α. The mixing parameter is in the range from 0 to 1, where 1 corresponds to a LASSO regression and 0 to ridge regression. With cross validation α was determined in 0.1 steps in [0, 1]. The mixing parameter had to be set before solving the following problem:

$$min_{\beta_0,\boldsymbol{\beta}}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{f}\left(\boldsymbol{x}_{i,j}\boldsymbol{\beta}_j\right)\right)^2 + \lambda\left((1-\alpha)\,\frac{1}{2}\,\|\boldsymbol{\beta}\|_{l_2}^2 + \alpha\|\boldsymbol{\beta}\|_{l_1}\right)\right) \qquad \text{(Eq. 3-4)}$$

After fixing α, the tuning parameter λ was estimated during cross-training of the elastic net. N was the number of samples and f was the number of features. $\boldsymbol{y}$ were the drug response values to fit. $\boldsymbol{x}$ were the input features, where $\boldsymbol{\beta}$ were the corresponding weights and β_0 was the intercept of the linear regressed model.

The l_1 penalisation term from LASSO favours non-correlated features and sparsely selects from them. This means that if input features are correlated, then l_1 penalisation would arbitrarily select a representative feature of them. Accordingly, an arbitrary representative feature would receive a large weight

(i.e. large β), while other features that would be correlated with the chosen representative are set to 0.

The l_2 penalisation term from the ridge compensates the l_1 effect by sharing weights among all correlated features. Therefore, ridge regression leverages more features than LASSO, but with lower weights for each individual feature, i.e. the l_2 penalisation allows group wise selection from correlated features.

I used the elastic net implementation from the 'glmnet' R package version 2.0-2 (Friedman, Hastie, & Tibshirani, 2010) and fitted α and λ during cross-training.

3.3.6: Training and evaluation of models

I used bootstrapping to train the models, where 80% of data was allocated for pure training, 10% for identifying the α and λ parameter of the elastic net (see above, section 3.3.5: Elastic net), and the remaining 10% for an unbiased test set.

I explored with the 10% cross-training data in a grid search the local optima of α and λ. The α parameter was tested in a range from 0 to 1 in steps of 0.1, while for λ I used the default settings of the 'glmnet' R-package. The λ parameter was tested in 100 steps with a $\lambda_{min}/\lambda_{max}$ of ~0.01, where the R-package linear interpolates λ values in log scale from λ_{min} to λ_{max}.

The splitting process of 80% training, 10% cross-training and 10% testing was performed 1,000 times, resulting in 1,000 slightly different models. This might be surprising at first, but the variation in the models was due to having finite training data and noise. For robustly estimating the performance, I exclusively used the test sets of 1,000 different splits to evaluate the predictive power (see below, section 3.3.7: Predictive power metric). Reported performance was done on predicted mean values across 1,000 bootstrapped test sets.

3.3.7: Predictive power metric, threshold & confidence

To evaluate the predictive power, I reported the Pearson correlation (R_p; equation 2-5) of observed versus predicted $1 - AUC$ value of the drug response curve. R_p was naturally in a range from -1 to 1, while positive values inferred good predictions, values around zero indicated random predictions and negative values pointed towards reverse predictions.

Negative R_p corresponded to a model that systematically predicted drug responses wrongly. This was undesirable but might happen through overfitting on noise. Overfitting was a problem particularly when focusing on small tissue types with few training samples, e.g. Lung Squamous Cell Carcinoma (LUSC) contains 15 cell lines.

In a pan-cancer analysis the expected random mean of R_p might be systematically shifted towards positive R_p through leveraging the tissue type, i.e. the tissue was a good predictor and several features were correlated with tissue-of-origin (see 3.5.3: Tissue effect in pan-cancer study).

For identifying a threshold of 'predictive' models, which also considers biases through overfitting and tissue dependencies, I built models for all 265 compounds based on all molecular layers and their combinations in a pan-cancer setting. One observation was that based on some features I could build a predictive (i.e. informative) model, but based on others not (see 3.4.1:Pan-cancer analysis of molecular features). Therefore, I assumed bimodality: (i) a distribution predictive/informative models (in green; D_{inform}) and (ii) another distribution reflecting random noise including the tissue bias (in red; D_{rand}). I fitted two Gaussian distributions (bimodal distributions) with the R-package 'mixtools' version 1.0.3 (see Figure 3.1).

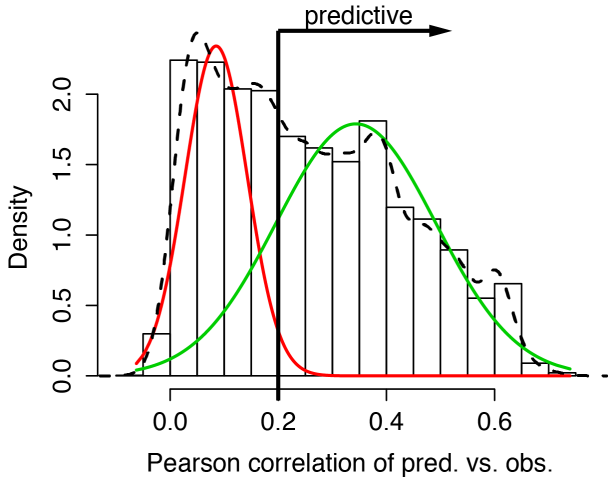

Figure 3.1: Predictive power threshold.

Shows the threshold for distinguishing predictive performance into 'predictive' or 'informative' (in green; D_{inform}) and 'random' distribution (in red; D_{rand}), which is at ~0.2 Pearson correlation ($\mathrm{R_p}$). On the x-axis is $\mathrm{R_p}$, while on the y-axis is the density of all pan-cancer models using all molecular layers and their combination. The performance of models is split into a bimodal distribution: 'random' and 'predictive' model fits are in red and green, respectively. The dashed line is the density fit of all models.

In pan-cancer the expected distribution for random was a normal distribution of $N_{rand}(\mu_{rand} \approx 0.085, \sigma_{rand} \approx 0.058)$, where μ_{rand} and σ_{rand} were expected mean and standard deviation, respectively. As informative predictions, I assumed $R_p > \mu_{rand} + 2\,\sigma_{rand} \approx 0.2$, which removed random models to ~97.6%.

In addition to the cut-off of a ~0.2 Pearson correlation, which I considered as predictive, I also provide a continuous estimate for the likelihood to be from the informative (D_{inform}) versus random distribution (D_{rand}) (Figure 3.2).

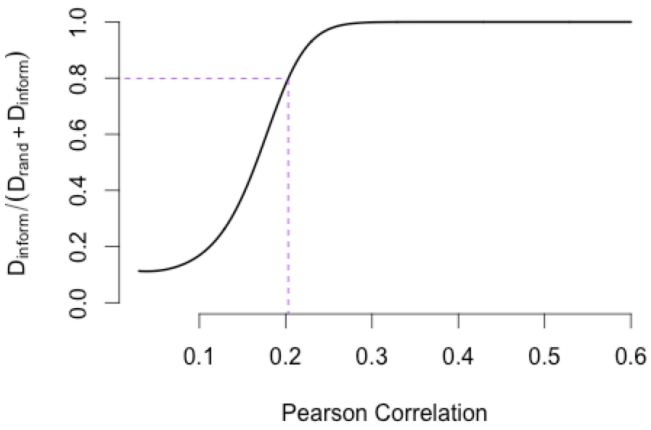

Figure 3.2: Pearson correlation and its relationship to be informative.

Shows the achieved Pearson correlation and its corresponding likelihood to be drawn from the informative distribution (D_{inform}) versus non-informative (D_{rand}). The purple line is the chosen threshold for a predictive model. At this given threshold the likelihood is ~80% to be derived from D_{inform}, while the remaining 20% are likely from D_{rand}. Models with Pearson correlation > 0.3 are guaranteed to carry information.

There are limitations to translating an informative drug response model in cell lines to patients. As previously described, cell lines are very different to primary tumours (see 1.4.5: Biological models). Cell line artefacts might explain drug response, but those artefacts might not be clinically relevant. It has to be noted that cell line screens were built to identify potential hits, which have to be further investigated in animal models before making any assumption for human beings and clinical applications. Nevertheless, the advantage of cell line screens is their high-throughput, which would not be feasible in any animal or human model. Furthermore, such an exploratory study in human beings would be unethical. Therefore, I considered the threshold of $R_p \approx 0.2$ to avoid the early filtering out of potential hits.

In a cancer-type specific analysis no tissue bias existed. At the same time the potential bias through overfitting would be larger because of a smaller sample size, which further decreased expected performance. However, I 'conservatively' applied the same threshold as in pan-cancer of $R_p > 0.2$ for identifying 'predictive' models in the cancer-type specific setting.

3.4: Results

To assess the contribution of different molecular layers in explaining the variation in drug response, I built linear machine learning models (see 3.3.5: Elastic net) and compared their performances. Linear regression models leverage multiple genes at the same time and enable therefore the comparison of feature layers, while the ANOVA from the previous chapter is limited to single genes.

Here, I benchmark three molecular layers: the genomic layer containing copy number variations and mutations, the methylation layer of informative CpG isles and the gene expression layer of ~17k genes.

For each of the 265 compounds I built models across all tissue-types at the same time (i.e. pan-cancer) to identify global trends. Subsequently, I also built models individually for each tissue type (i.e. cancer specific), which leads to a limited sample size through sub-setting the cell line panel, but at the same time avoids any dilution of signal from other tissues.

3.4.1: Pan-cancer analysis of molecular features

In the pan-cancer analysis, the most predictive feature was gene expression, interestingly, closely followed by the tissue-of-origin (i.e. cancer type) of cell lines as an input feature. In contrast, the genomic layer performed worst overall (Figure 3.3 A). The predictive power of gene expression was strongly correlated with the predictive power of tissue, while genomics was less correlated with the tissue predictions (Figure 3.3 B). This was expected as gene expression is known to be strongly affected by the tissue of origin (Ross et al., 2000), and the tissue on its own is a known good predictor of drug response (Majumder et al., 2015). Another observation was that methylation was more similar to genomics than gene expression, which was the result of filtering for informative CpG islands.

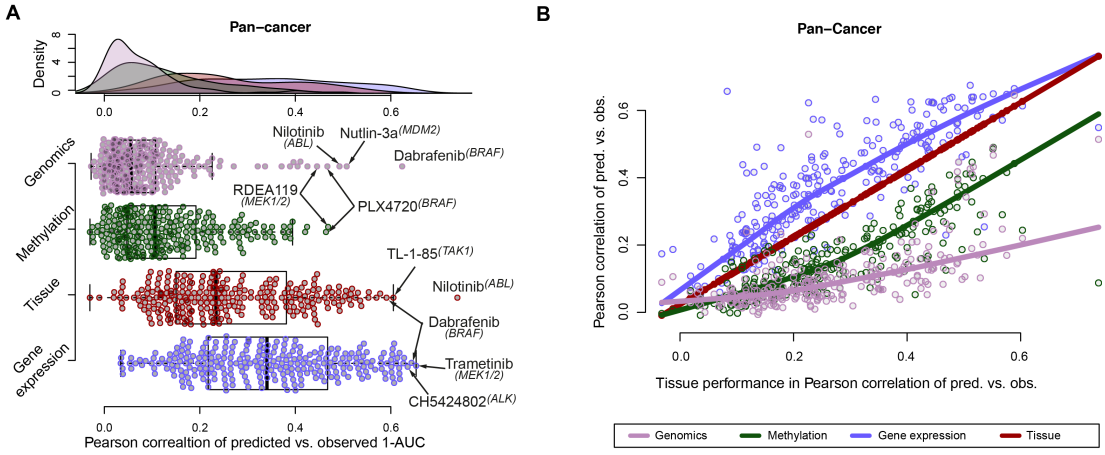

Figure 3.3: Pan-cancer predictions from various molecular layers.

(A) Predictive performance of each drug model in a pan-cancer setting for each individual molecular layer: genomics, methylation, tissue and gene expression. Notably, 'tissue' is not a molecular layer, rather the binary encoding of the primary tumour location of each given cell line. Predictive performance is measured as the Pearson correlation of predicted versus observed drug response using $1 - \mathrm{AUC}$. (B) Scatter plot of achieved performance from tissue versus the other molecular layers.

Notably, in pan-cancer MEK1/2 inhibitors and BRAF inhibitors are among the top performing models across all molecular layers. In contrast, ABL inhibitors such as nilotinib are exclusively well predicted from the tissue label and genomic layer, while neither the gene expression nor the methylation layer captures this association well. The tissue label and genomic layer perform well because of the fusion gene *BCR-ABL* in AML and CML lineage, which infers sensitivity when chromosomally rearranged.

3.4.2: Cancer-Specific analysis of molecular features

Evaluating each molecular layer (genomics, methylation and gene expression) and their combinations in a cancer-type specific manner is arguably more clinically relevant, since the primary tumour location is usually known in a patient. Here, I evaluated each drug in the different tissue contexts and report the used molecular layer of the best performing models.

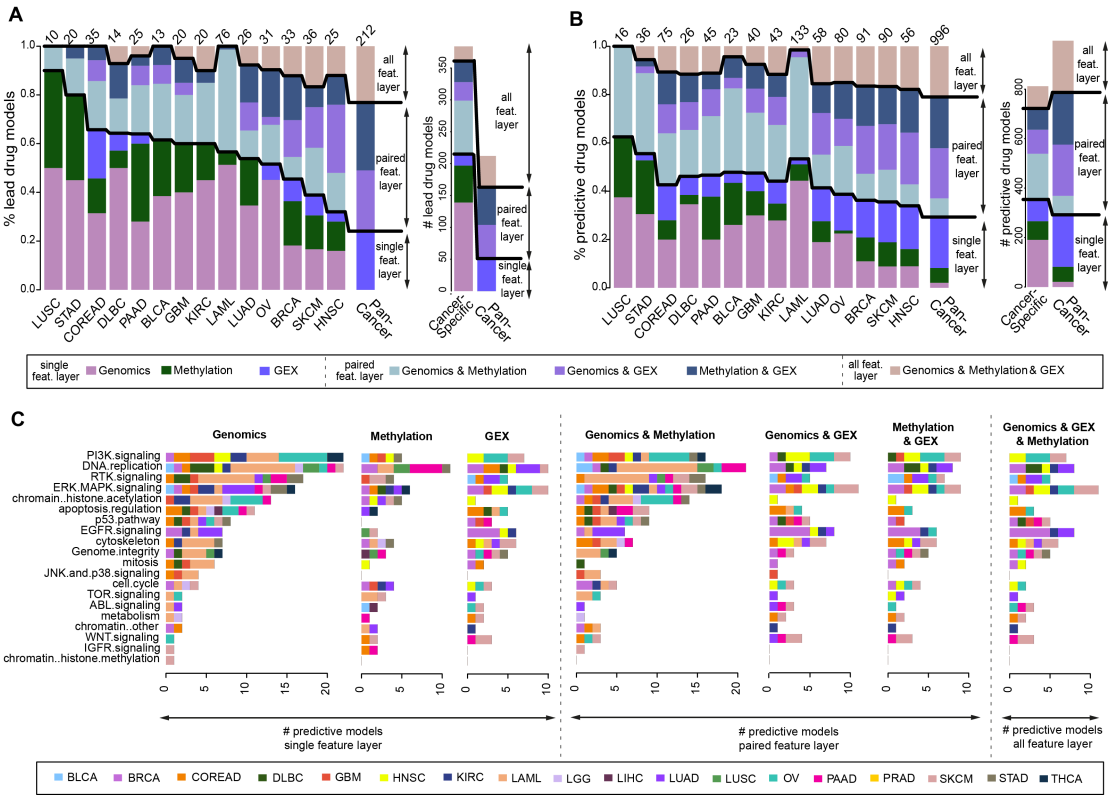

Figure 3.4: Cancer-specific predictions from various molecular layers.

(A) Shows Lead predictive model (R_p > 0.2 & 'best') for each drug within a different tissue context. Lead drug model was defined as the best performing model within a given context, i.e. tissue type. The black lines separate single from paired layer, and paired from all feature layers. On top of the bar plot is the absolute number of drug models per cancer type displayed. Pan-cancer was the union of all cell lines and added for reference. Tissue types where less than 10 predictive models could be built were discarded, leaving us with 14 cancer types. The figure further indicates as a bar plot the absolute numbers of all drug models in cancer-specific and pan-cancer settings. (B) has the same layout as (A), but counts all models which at least perform better than 0.2 R_p. This included redundant models, for example, a drug might be predictable ($R_p > 0.2$) by different molecular layers and in (B) all of those models would be counted, while (A) only counted the best performing of them, i.e. only so called 'lead' models. (C) Shows the number of all models R_p > 0.2 in a tissue specific and drug pathway centric view.

Towards the goal of evaluating feature layer importance, I first reported the used molecular layer for the 'predictive' ($R_p > 0.2$) and best performing so called 'lead' models within different tissue types and compared those to the pan-cancer setting for reference (Figure 3.4 A).

In pan-cancer all the best performing models incorporate gene expression, although 111 drugs (~52%) benefitted from including either methylation or genomics, and 49 drugs (23%) gained performance by leveraging all three molecular layers (Figure 3.4 A). On the contrary, the dominance of the gene expression layer observed in pan-cancer disappeared in a cancer-specific analysis.

Focusing on a single cancer-type at a time, for 146 drugs (~36%) the 'lead' model (i.e. the most predictive model among the $\mathrm{R_p} > 0.2$) was solely based on genomics. Furthermore, genomics in combination with the methylation layer resulted in additional 90 lead drug models (~22%), in combination with gene expression in another 29 lead models (~7%) and genomics integrated with both other layers in 23 lead models (~6%). This summed up to 288 lead drug models (~71%) including genomics in a cancer-specific setting, while only 116 lead models (~29%) were independent from genomics.

Subsequently after analysing the 'lead' models (i.e. $R_p > 0.2$ & 'best' models; Figure 3.4 A), I produced results for all 'predictive' models (i.e. R_p > 0.2; Figure 3.4 B). The aim of this analysis was to evaluate redundancies across different feature layers and to investigate their general information contents. In a cancer-specific setting the total number of lead compared to all predictive models increased from 404 to 842, while in pan-cancer the number boosted from 212 to 996 (Figure 3.4 A & B). Therefore, the ratio of predictive to lead models in cancer-specific and pan-cancer was ~2 and ~5, respectively. In conclusion, more pan-cancer models had alternatively predictive models than cancer-specific models.

Another observation was that the fraction differed from models based on single versus multiple layers, when comparing 'lead' (Figure 3.4 A) against all predictive models (Figure 3.4 B). In a cancer-specific setting the 'lead' models based on single and multiple molecular layers were 229 (~57%) and 175 (~43%), respectively (Figure 3.4 A). When focusing on all predictive models ($R_p > 0.2$) the fraction of single and multiple features changed to 373

(~44%) and 469 (~56%), respectively (Figure 3.4 B). This illustrated a trend that by adding an additional molecular layer the performance did generally not improve, although models based on multiple layers were valid alternatives.

Notably, in the pan-cancer setting the 'lead' models relied always on gene expression (Figure 3.4 A), however, 159 alternative and predictive models (~16%) could be independently built without gene expression (Figure 3.4 B).

After comparing lead (Figure 3.4 A) and all predictive models (Figure 3.4 B), I categorised the screened compounds based on their ability to target 19 specific biological processes (Figure 3.4 C). Concordantly with previous observations in a cancer-specific setting the genomics (and in combination with methylation) layer was mostly able to predict drug response.

To summarise, gene expression was the strongest predictor of drug response in pan-cancer screens, likely due to its high tissue specificity. However, in arguably the more clinically relevant cancer-specific analysis, genomics (either alone or in combination with methylation) was the most predictive data layer for the majority of the screened compounds. Nevertheless, this would need to be evaluated for each compound individually in clinical settings.

3.5: Discussion

3.5.1: FLT3, NRAS and TP53 in AML

For acute myeloid leukaemia (AML, or alternatively the TCGA abbreviation LAML) a total of 76 drug models could be built, which was comparably large to other cancer-types (Figure 3.4 A). 71 (~94%) of those best performing models incorporated genomic features, particularly based on the FLT3, NRAS and TP53 mutations, which were among the most common alterations in AML (Fidler et al., 2004; Stirewalt et al., 2001). In AML I observed sensitivity of FLT3 mutant lines to FLT3 inhibitors (Wander, Levis, & Fathi, 2014) and conversely resistance to DNA damaging agents in TP53 and NRAS mutant cell lines (Dunna et al., 2014; Wong et al., 2014).

3.5.2: Drug class enrichment in tissue centric analysis

The enrichment analysis of predictive models in a pathway-centric view with a cancer type revealed a few insights (Figure 3.4 C). Particularly in AML (i.e. LAML) good models were enriched for 'DNA replication' and 'receptor tyrosine kinases (RTK) signalling'. The enrichment for RTK signalling targeting drugs in AML was a result of FLT3 mutations (i.e. FLT3 is a RTK), which is a well-known sensitivity biomarker of FLT3 inhibitors in AML (Wander et al., 2014) (see 3.5.1:FLT3, NRAS and TP53 in AML). The increased number of predictive AML models in DNA replication targeting drugs was due to TP53 and NRAS mutations, and their association to drug resistance (Dunna et al., 2014; Wong et al., 2014).

Ovarian cancer (OV) was enriched in genomics and PI3K signalling, which was frequently altered in OV (Cheaib, Auguste, & Leary, 2015). ERK signalling was predictable in various cancer-types, while for drugs targeting 'chromatin and histone methylation' almost no predictive model could be built. Unexpected, 'chromatin and histone methylation' targeting agents were also not well predicted by the methylation layer, however, this might be due to the fact that I focused on informative CpG islands (see 3.3.3: Methylation dataset).

3.5.3: Tissue effect in pan-cancer study

This analysis was carried out in a pan-cancer and cancer type specific setting. Strength of pan-cancer approaches was that the sample size of 1,001 cell lines was large compared to the cancer specific setting, which improved statistical power. However, at the same time the weakness of the pan-cancer analysis was that drug response is very tissue specific. For instance, solid tumours were generally more resistant than blood cancers (Minchinton & Tannock, 2006), meaning blood cancer cell lines generally tended to respond more sensitively at lower drug concentrations than solid tumours. Furthermore, tissue specific signals might be diluted.

In a pan-cancer setting a model would perform well if it was capable of classifying the tissue type, since the tissue on its own is a powerful predictor.

This would be a transitive ($X \rightarrow Y \rightarrow Z$) and trivial association. For example, the molecular layer X explained tissue type Y ($X \rightarrow Y$), tissue type Y explained drug response Z ($Y \rightarrow Z$); therefore the molecular layer X seemed to be a good biomarker for drug Z ($X \rightarrow Z$), but the trivial dependency was that tissue Y explains drug Z ($Y \rightarrow Z$). In such a scenario, no molecular biomarker X would be necessary and the drug response Z could be fully explained by the tissue-of-origin Y.

Particularly the gene expression layer was strongly correlated with the tissue-of-origin (Ross et al., 2000), and therefore predictive power in pan-cancer from gene expression might trivially reflect the tissue type. The genomics feature layer was comparably less correlated with tissue-of-origin and therefore clinically more relevant.

3.6: Conclusion

In this chapter, I studied which 'omics' features were the most important for predicting drug response. In this analysis, I built machine learning models to predict the drug response in cancer cell lines and found genomic features (combining copy number variations and mutations) as the most predictive feature within a cancer type specific context. This supports my original hypothesis that drug response in cancer cell lines would be predominantly driven by genetic alterations, based on the observation that cancer is a genetic disease. Furthermore, this study revealed that genomic features in combination with methylation improved the predictive power. Within a pan-cancer context (across different cancer types), transcriptomic was by far the most informative feature due to reflecting the tissue type. This was less clinically relevant since the cancer tissue would be usually known upon treatment. Therefore, prioritising copy number variations, mutations and methylation profiles might be the most cost effective way to deliver effective treatments with biomarkers to stratify patient cohorts in clinics.

Chapter 4: Predicting drug response based on genomics and chemistry

"*Essentially, all models are wrong, but some are useful.*"
-George E. P. Box

4.1: Declaration of contribution

The planning for this project was cooperatively performed with Pedro Ballester, Julio Saez-Rodriguez and myself. Biological data was exclusively retrieved from published data of the Genomics of Drug Sensitivity in Cancer (GDSC) project. I jointly designed the software pipeline with Pedro Ballester and Julio Saez-Rodriguez, while I alone implemented the shown software solution in this chapter and all presented results were solely my contribution. I jointly discussed the experimental setup, results and biological interpretation with Pedro Ballester, Julio Saez-Rodriguez, Francesco Iorio, Cyril Benes, Mathew Garnett and Ultan McDermott, which led to the publication "*Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomics and Chemical Properties*" (Menden et al., 2013).

4.2: Introduction

In the context of personalised medicine, the ultimate goal of methods to predict drug sensitivity is to provide a treatment that is tailored to the patient's need based on their tumour profile. However, there were several challenges to overcome. For example, building highly accurate models within cell lines was already a challenging task on its own (Costello et al., 2014), and further obstacles such as translation from cell lines to real patients had to be solved (Cree et al., 2010; Domcke et al., 2013). The goal of this chapter was to improve predictions of drug sensitivity in cell lines in comparison to results of the previous chapter.

Aligned with the aim of this chapter to improve the predictive power I chose a pan-cancer setting. As pointed out in the previous chapter, a pan-cancer analysis would have more statistical power through its larger sample size than a cancer-specific setting, while at the same time introducing a trivial bias through the tissue type.

In order to minimise the tissue bias in this pan-cancer analysis, I exclusively selected the genomic layer as input for the models. As shown in the previous chapter, the genomic layer was less correlated with the tissue than the epigenetic or transcriptomic layer.

In this chapter, I explored the integration of the genomic layer of cancer cell lines and chemical properties of the drugs for improving predictive power compared to genomics alone (Figure 4.1). Through the integration of chemical features of the drugs, the similarity between different compounds was modelled. Weak effects observed in one compound might be more obvious in another drug, and identified through a 'guilt-by-association' approach (Keiser et al., 2009), i.e. 'guilt-by-similarity'. For instance, off-target effects in one drug might mildly contribute to the drug response, which could be learned through other compounds that exactly target this off-target and show a stronger drug response. Therefore, a machine learning algorithm trained across different drugs was more likely to pick up weak but global biological signals.

Furthermore, by integrating those two orthogonal but complementary data streams of genomics and chemistry the sample number increased, which now became the number of cell lines times the number of drugs (more training data available).

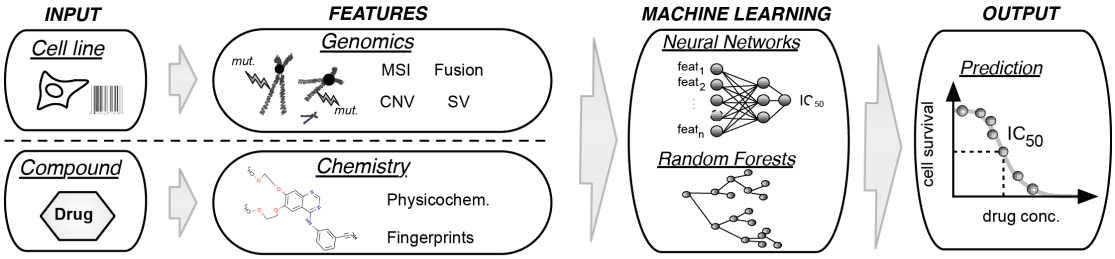

Figure 4.1: Genomic & chemistry model for predicting drug response.

Shows the integration of the two different data streams to predict the drug response. Inputs were the drug and cell line features, while the drug features were physiochemical properties and fingerprints, and the cell features were microsatellites (MSI), fusion genes, sequence variations (SV) and copy number variations (CNV). As a representative machine learning model I trained a neural network, while Pedro Ballester built a random forests for reference *(Menden et al., 2013)*. Finally, based on both data streams the drug response was predicted in IC_{50} values (concentration to reduce cell viability by half). Figure was adapted from Menden et al. PLoS One 2013.

An application of this prediction method was the identification of genomic biomarkers (i.e. measured molecules indicating drug sensitivity or resistance). Notably, approximately 90% of all cancer drugs entering clinical trials failed the final clinical approval (Paul et al., 2010), which was mainly due to not enough efficacy. Efficacy could be increased by identifying a suitable patient cohort, therefore, a cancer drug that had a validated biomarker was more likely to be clinically approved (Kelloff & Sigman, 2012). Therefore, in this chapter I explored the reproducibility of genomic biomarkers by leveraging the drug response predictions from genomics and chemistry.

4.2.1: State of the art in modelling drug response

In a community effort to predict drug response of breast cancer cell lines, 44 different research teams applied various machine learning algorithms (Costello et al., 2014). ~1/3 applied sparse linear regression methods such as elastic net, lasso or ridge regression. Another ~1/3 used non-linear prediction

methods such as random forests or neural networks. The remaining ~1/3 used kernel methods (e.g. support vector machine), PC or PLSR regression, ensemble models (i.e. leveraging multiple models) or others. However, most of the methods performed similarly, i.e. the method had modest impact on predicting drug sensitivity. This was also observed in other studies comparing multiple methods (Jang et al., 2014).

Another outcome of the NCI-DREAM drug sensitivity challenge (Costello et al., 2014) was that most predictions did not perform much better than random. For instance, focusing on a single drug at the time, the predicted cell line sensitivity ranking was within the distribution of random. However, 34 out of the 44 teams managed to predict the ranking better than randomly expected (two-sided, t-test, FDR < 5%), but not a single drug prediction on its own was significantly distinguishable from random. Furthermore, all teams (beside one) managed to predict more correct than wrong. This implied that almost all teams captured biological signals across all drugs, but it also highlighted the difficulty in accurately predicting drug response in cell lines.

The best performing team in the drug sensitivity challenge leveraged to their advantage multiple drugs at the same time when building the predictive model, rather than building a model for each drug individually. This technique is called multi-task learning and aimed to train communalities amongst the drugs without neglecting individual drug responses. In essence, the winning method benefitted from lumping together similar drugs, which resulted in more training data. The idea of multi-task learning was an already established and successfully applied technology in other fields such as HIV research (Qi, Tastan, Carbonell, Klein-Seetharaman, & Weston, 2010) and Alzheimer disease prediction (D. Zhang & Shen, 2012).

The approach I used is similar to the idea of a multi-task learning model, but I additionally include chemical information of the drug (Menden et al., 2013). A multi task model as well as the approach to add chemical information of drugs as an additional feature enabled leveraging more training data and was also capable of learning similarities across drugs, i.e. off target effects that were

not prominent in one drug might be more visible in another compound and improved predictive power. However, by including the chemical feature space in the model, as shown in this chapter, it generalised the problem and became extendable to novel and previously untested compounds, which previously was not possible by a pure multi-task learned model.

Several other models followed up the idea to leverage multiple drug responses to predict mono-therapy and showed superior predictive power; for example, with either using multi-task learning (Gönen & Margolin, 2014) or, following the idea of this chapter, by including features of the drugs (Khan et al., 2014; Kumar, Chaudhary, Singla, Gautam, & Raghava, 2014).

Following up on the results of the previous chapter, which highlighted that genomics features were the most predictive ones in a cancer-specific context, I tried to further improve the performances within a pan-cancer setting. For improving the performance, my hypothesis was that this might be achieved by additionally using features of the different compounds. I applied established Quantitative Structure-Activity Relationship (QSAR) methods to derive those drug features. In this analysis, I combined the genomic features and drug properties, which I hypothesised to be superior to genomic models alone, and ultimately enabled drug repositioning of unscreened drugs.

4.3: Methods

4.3.1: Drug response data

Drug response data version 1.0 was downloaded from the GDSC project website (http://www.cancerrxgene.org/downloads/). It contained drug responses for 638 cell lines, 131 drugs and 65,614 IC_{50} values. In this chapter the focus was on predicting the IC_{50} (concentration values to reduce cell viability by half) rather than the area under the drug response curve (AUC) with the aim of also predicting the concentration dosage of a drug.

4.3.2: Genomics data

Genomic information version 1.0 was publicly available from the GDSC project website (http://www.cancerrxgene.org/downloads/). It contained the mutational status of 67 oncogenes determined with capillary sequencing, while I focused exclusively on sequence variance (i.e. SV) that changed the protein sequence.

$$\mathrm{SV} = \begin{cases} 1, & if\ mutation \\ 0, & if\ wild\ type \end{cases} \quad \text{(Eq. 4-1)}$$

For those 67 oncogenes the copy number variation (i.e. CNV) was generated with Affymetrix SNP6.0 array data.

$$\mathrm{CNV} = \begin{cases} 1, & if\ amplified \\ 0, & if\ wild\ type \\ -1, & if\ deleted \end{cases} \quad \text{(Eq. 4-2)}$$

Furthermore, the GDSC project also provided the fusion gene status of *BCR-ABL*, *MLL-AFF1* and *EWSR1-FLI1*, which was produced with breakpoint-specific sequence primers and followed capillary sequencing to detect chromosomal rearrangements.

$$\mathrm{fusion} = \begin{cases} 1, & if\ rearranged \\ 0, & if\ wild\ type \end{cases} \quad \text{(Eq. 4-3)}$$

As a side note, the features SV, CNV and $fusion$ were mostly mutually exclusive for the 70 oncogenes, and therefore could alternatively be encoded as in Chapter 3.3.2: Genomic dataset.

Additionally, the GDSC project supported information of microsatellite stability (MSI), which was encoded as following:

$$\mathrm{MSI} = \begin{cases} 1, & if\ stable \\ 0, & if\ instable \end{cases} \quad \text{(Eq. 4-4)}$$

4.3.3: Chemical information

There was a large community which focused on the generation of chemical features of compounds, namely, Quantitative Structure Activity Relationship (QSAR) models (Devillers, 2013; Jaworska, Nikolova-Jeliazkova, & Aldenberg, 2005).

In this study I used the PaDEL-descriptors (Yap, 2011), which generated chemical properties and fingerprints of each compound from simplified molecular-input line entry system (i.e. SMILES) structures. PaDEL generated 722 chemical properties and comprehended features such as molecular weight, atom count, bond count, lipophilicity or rule of five. The Lipinski's rule of five is an approximation for identifying drug-like compounds, where the molecular weight < 500 daltons, H-bond acceptors <= 10, H-bond donors <= 5 and calculated octanol-water partition coefficient <= 5 (note: naming for the rule of five derives from all estimated numbers being multiples of 5). H-bond acceptors were oxygen (O) and nitrogen (N) atoms, while H-bond donors were -NH and -OH groups (Lipinski, Lombardo, Dominy, & Feeney, 2001).

Furthermore, the PaDEL software generated 10 different kinds of fingerprints, which led to 881 binary input variables. Fingerprints are bitmaps that were vectors of a specific size containing either 0's or 1's. Most algorithms to calculate fingerprints were based on the absence or presence of a molecule, or alternatively encoded the count of certain molecules (Swamidass & Baldi,

2007). Different algorithms might lead to different distances between compounds, which could be calculated with the Tanimoto distance (Rogers & Tanimoto, 1960). The underlying algorithm used to define a fingerprint might produce technical biases and furthermore could bias the selection of compound libraries (Hert, Irwin, Laggner, Keiser, & Shoichet, 2009).

The set of chemical properties and fingerprints summed up to 1,603 chemical features. However, I discarded features that could not be calculated or have identical values across all drugs, i.e. features with identical values had no informative content for prediction. This resulted in a final set of 689 chemical features.

4.3.4: Neural network

The neural network implementation was in Java from the Encog 3.0.1 framework (http://www.heatonresearch.com/encog) (Heaton, 2008, 2011). The neural network was a feed-forward multi-layer implementation based on 3 layers: input, hidden and output. The input layer size corresponded to the number of chemical and genomic features. The hidden layer was between 1 to 30 units, which were determined during cross-training (usually 21 or 26 units selected). Each unit of a layer was connected to all units in the following layer. Furthermore, each of the units in the input and hidden layer had a bias, meaning a constant activation input of 1. There was only one output unit for the regression problem, for predicting the IC_{50} value.

The neural network used as an activation function, a logistic function in the range from 0 to 1. Therefore the IC_{50} values required a normalisation in the range from 0 to 1 as well, which was achieved with the following function (equation 4-5):

$$norm(\boldsymbol{y}) = \frac{1}{1+\boldsymbol{y}^{-0.1}} \qquad \text{(Eq. 4-5)}$$

$\boldsymbol{y}$ was a vector where the elements were IC_{50} values. A $\boldsymbol{y}$ element of 0 would lead to infinity (i.e. $\boldsymbol{y}_i = 0 \rightarrow +\infty$), however, by definition the IC_{50} had to be larger than 0 (i.e. $0 \ngeq IC_{50} > 0$).

The neural network was trained with resilient error backpropagation implementation from Encog with default parameters (Riedmiller & Braun, 1993) and maximally trained for 400 iterations, while mostly converted to an optimum in performance of 300 iterations.

4.3.5: 8-fold cross-validation

The models were validated with an 8-fold cross-validation, meaning that 75% of the data was used for training, 12.5% for fitting hyperparameters, and the remaining 12.5% for testing the predictive performance, which was not used at any stage of training. Hyperparameters that had to be cross-trained for the neural network were the number of iterations to perform back propagation and the number of units in the hidden layer. In cross-validation, trainset, cross-trainset and testset were iteratively rotated so that each data point was used once for testing. In other words, 8 models were generated based on different splits of the data.

As a side note, in an ideal scenario all those 8 models should have identical predictive power, but small variations in performance were expected due to noise in the data. Therefore, the prediction of novel data points would be recommended with an ensemble (i.e. average) of all 8 models.

4.4: Results

4.4.1: Genomic & chemistry versus genomic model in pan-cancer

In contrast to the NCI-DREAM challenge that solely focused on breast cancer, this chapter investigated in a pan-cancer setting the relative improvement by integrating chemistry and genomics over models leveraging genomics alone. A neural network that was based on the genomics of cells and additionally chemical properties of the drugs (i.e. used all drugs for training), performed moderately but significantly better (p-value < 0.01, paired t-test) than building for each drug an individual model using genomic features exclusively (Figure 4.2).

This result was concordant with other literature, which showed the superiority of multi-task learned models (Gönen & Margolin, 2014; Khan et al., 2014; Kumar et al., 2014; Menden et al., 2013).

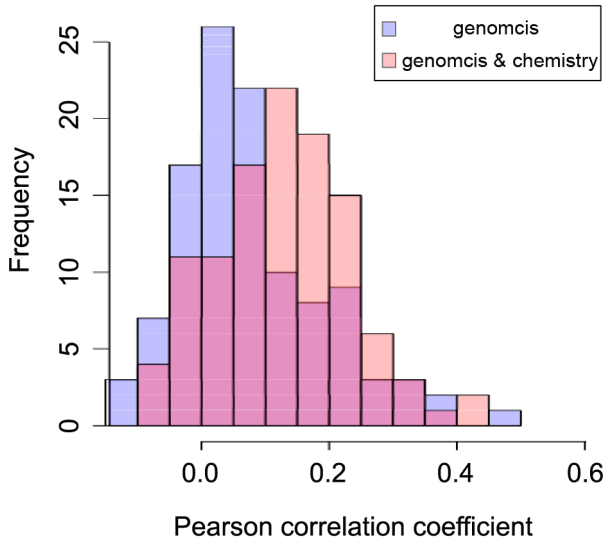

Figure 4.2: Genomic models versus genomic & chemistry model.

Benchmark of genomic models versus the model that additionally integrated the chemical information of the drugs. Performance was measured in Pearson correlation of predicted versus observed IC_{50} for each drug. Models that were exclusively based on genomics are shown in blue, while models combining genomics and chemistry of the drug are in red.

4.4.2: Genomic & chemistry model reproduces biomarkers

In the GDSC project the genomic biomarkers were identified with an analysis of variance (ANOVA) (Garnett et al., 2012). An oncogene was considered as a biomarker of drug response, if its mutational status separated the cell lines into two populations: one with higher IC_{50} values versus the rest with lower IC_{50} values. A biomarker of resistance was when the cell population carried the mutant oncogene that responded significantly less than the wild type cell population, i.e. the IC_{50} values of the mutant cell lines were much higher than the wild type. Conversely, if the mutant cell lines had lower IC_{50} values than the wild type, this was considered as a biomarker of sensitivity.

The reported volcano plot shows an ANOVA on the experimentally observed IC_{50} values (Figure 4.3 A). Any biomarker effect that was correctly predicted with the genomics and chemistry neural network is shown as a blue dot, while wrongly predicted tendencies are in red. For 168 out of 213 (~79%) experimentally and statistically identified biomarkers, the drug effect tendency was correctly predicted (positive or negative drug sensitivity). When focusing on the associations significantly predicted (>20% FDR), 59 out of the 213 (~28%) experimentally verified biomarkers were estimated with the neural network. As expected, the genomics and chemistry model lacked the capability to identify tendencies when the effect size was small (separation of mean values by mutant versus wild type population), or when the experimental identified biomarkers were not highly significant.

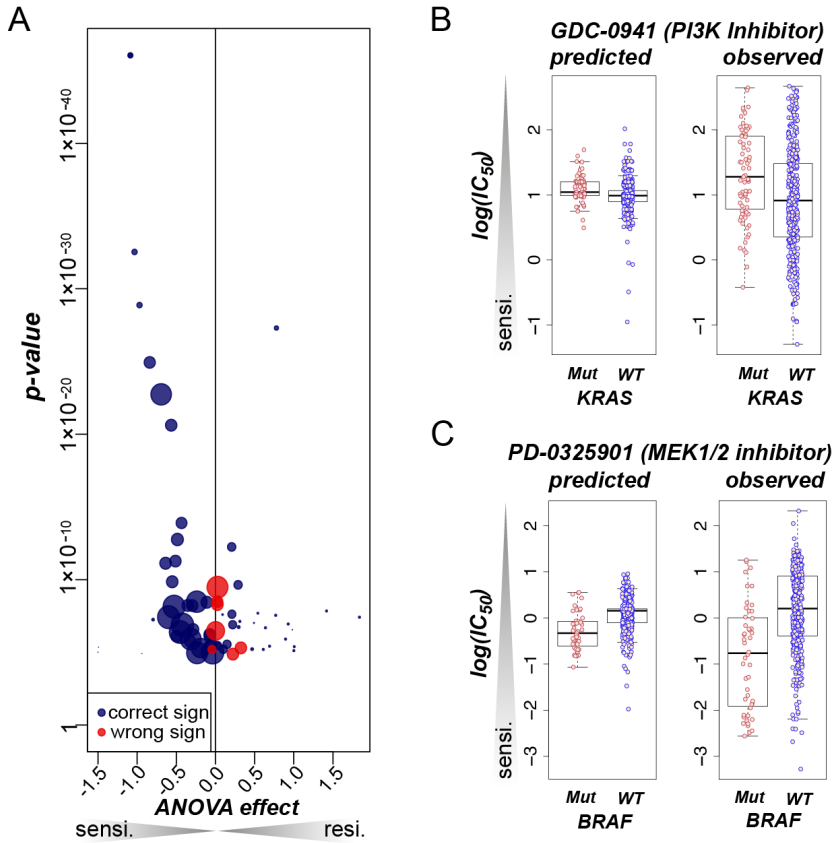

Figure 4.3 Biomarkers reproduced with genomic & chemical model.

(A) shows ANOVA volcano plot where each dot corresponds to a drug-oncogene association. Size of each circle corresponds to the cell population size with mutated oncogene. ANOVA effect size is the relative difference in mean IC_{50} of the two cell line populations that carry oncogene mutant or wild type. The tissue label was used as a covariate of the model to compensate for the tissue effects. In case the prediction reflected the correct ANOVA effect sign the dot is coloured blue, otherwise it is coloured red. Only associations are shown which performed better than 20% FDR. (B) shows the predicted versus observed IC_{50} values and separation by KRAS mutation and wild type. (C) is in the same format as (B), but separates the cells by BRAF mutation and wild type. Figure was adapted from Menden et al. PLoS One 2013

As an example of resistance biomarkers, here I showed KRAS mutations and the PI3K inhibitor GDC-0941 (Figure 4.3 B). KRAS mutations were experimentally observed to cause resistance as well as being predicted by my method.

This was supported by the finding that the combination of MEK1/2 and PI3K inhibition was synergistic for KRAS mutant cell lines (Engelman et al., 2008),

which targets the ERK signalling and PI3K-AKT pathway simultaneously (see 1.3.2: Cancer pathways).

As an alternative to the PI3K-AKT pathway activation through RTKs and GPCRs, this pathway could also be triggered through KRAS mutations (Castellano & Downward, 2011; Downward, 2008). However, KRAS mutations also intrinsically activate the ERK signalling, which may be the cause of resistance in KRAS mutants to PI3K inhibition alone (Downward, 2003, 2008), since between those two key pathways were several cross-talks. For example, mTOR is downstream of the PI3K-AKT pathway, but mTOR could also be activated through a strong signal of the ERK signalling (Mendoza et al., 2011).

As an example of drug sensitivity markers, I chose the well-known example of BRAF mutations and MEK1/2 inhibition (Figure 4.3 C). The sensitivity of $BRAF^{V600E}$ positive cell lines to MEK1/2 inhibition was confirmed in clinics (Chapman et al., 2011). BRAF mutations, particularly V600E, strongly activate ERK signalling, and therefore promote ERK-dependent cell proliferation (Davies et al., 2002; Wan et al., 2004). This oncogenic addiction to BRAF mutations could be therapeutically exploited by targeting MEK1/2, which is downstream of BRAF and the only activator of ERK.

4.5: Discussion

In summary, I demonstrated that the combination of chemistry and genomics data stream moderately, but significantly, improved the predictive power compared to models based on genomics alone. As a highlighted application, I showed the capability of those genomics and chemistry models to predict biomarker tendencies.

4.5.1: Limitations of genomic & chemistry model

The predictive power of this approach was still the limitation. Most likely, this could be improved by including more genetic features. Here I focused on 67 tumour drivers and 3 fusion genes that were well-known tumour suppressors or oncogenes. However, including more genes also increased the problem of multiple hypotheses testing of biomarkers, while performing more tests the false positive rate (FDR) correction method became more stringent in accepting p-values as significant.

Particular limitations regarding multiple hypothesis testing could be improved by including knowledge of pathways. For instance, statistical performance towards identifying biomarkers might be boosted by considering mutual exclusivities (Ciriello, Cerami, Sander, & Schultz, 2012). Such an approach would be based on the concept that mutations in the same pathway responsible for the same phenotype would most likely only naturally select one of those mutations. For instance, in case one of the mutations already occurred there would be no evolutionary pressure anymore for selecting the other mutation with the same effect, resulting in one or the other mutation occurrence (i.e. mutual exclusivity) in the cancer patient. This biological phenomenon could be exploited to boost statistical power by reducing the number of performed tests to the number of mutually exclusive modules with high abundance.

Another limitation was that the dynamic range of the predictions was much smaller than the experimentally generated IC_{50} values. This was partially due to the fact that the IC_{50} values were extrapolated and several of the

concentrations were unrealistically high or beyond the measurable minimal concentration (see 2.3.4: Drug response curve fitting and metrics).

4.5.2: Chemotype versus MoA

In this analysis, I assumed that the Mode-of-Action (MoA) of drug was related to its chemotype. This might be true for several compounds, although was undeniably a false assumption for various exceptions.

An extreme example would be large or small molecules which could both inhibit receptor tyrosine kinases (RTKs). Large molecules are antibodies, while small molecules are chemically synthesised compounds. Large Y-shaped proteins are undeniably different to small chemical structures, however, their MoA might be the same.

For example, cetuximab and panitumumab are both monoclonal antibodies targeting EGF receptors (EGFR) from outside the cell. This might also be achieved with the small molecules lapatinib, afatinib or erlotinib, which need to penetrate the cell membrane to target EGFR from inside the cell. Although their structure and mechanism to inhibit EGFR is very different, ultimately they share the same MoA, which stops EGFR from further stimulating the ERK signalling.

To systematically investigate the difference of chemotye and MoA, I hierarchically clustered the compounds based on their fingerprints and chemical properties. This was done with Manhattan distance. Alternatively, the chemotype may be described by exclusively using fingerprints, which would then allow the use of Tanimoto distance (Rogers & Tanimoto, 1960). However, Tanimoto distance is restricted to compare binary vectors of the same length, which is true for fingerprints, but false for chemical properties, e.g. molecular weight is a continuous float value. The hierarchical clustering of fingerprints and chemical properties is shown in the following illustration (Figure 4.4).

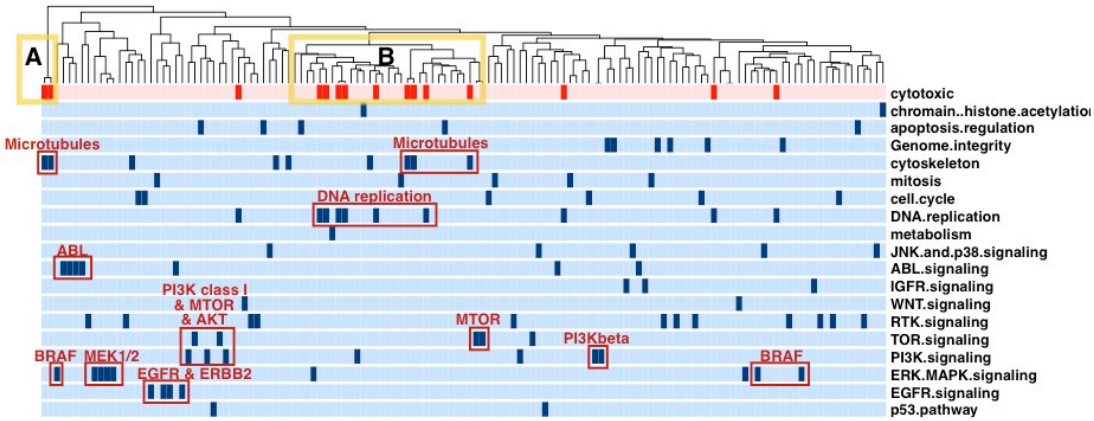

Figure 4.4 Compounds clustered by chemotype and labelled by pathway.

Shows the hierarchical clustering of fingerprints and chemical properties, which was done with Manhattan distance. (A) is a distinct cluster containing vinblastine and vinorelbine, which are both microtubule inhibitors. This is an example of where the MoA is closely related to the chemotype. (B) shows a mixed cluster of predominantly cytotoxic compounds. This cluster contains the remaining 3 microtubule inhibitors of the GDSC compound panel, which are fairly different in chemoptype to those in cluster (A). Those 3 microtubule inhibitors are docetaxel, epothilone B and paclitaxel. This shows that the chemotype might relate to the MoA, but does not necessarily have to.

A good example of the chemical similarity indicating similar MoAs in the GDSC screen were the tested ABL inhibitors, MEK1/2 inhibitors, EGFR & ERBB2 inhibitors (notably only small molecules were tested), while there was less conservation observed in the drugs targeting BRAF or PI3K.

As an improvement and future development, it would be interesting to integrate more information about the MoA. For example, this could be achieved by additionally integrating bioactivity profiles of the compounds.

4.5.3: Goals of genomic & chemistry model

At the same time as improving the predictive power by combining those two complementary input streams, the interpretability became more challenging. However, achieving the best possible prediction had clear technical and clinical applications:

i. Imputation in pharmacological screens
ii. *In silico* screen of compounds

A practical goal of an accurate prediction could be to experimentally run sparse screens of mono-therapies on various cell lines and complete the remaining missing data points with predictions. This could substantially reduce screen and labour costs by imputing missing drug response values rather than experimentally measuring all. In essence, more compounds could be tested with the same research budget

Through incorporating the dimension of chemical properties, drugs that have not yet been screened could be tested in an *in silico* screen. Such an *in silico* screen could be solely simulated in the computer, and therefore a cost efficient solution. Compounds could be associated by similarity, or as stated before: 'guilt-by-association' (Keiser et al., 2009). This approach could explore not yet experimentally screened cancer drugs, but would also be potentially extendable to drug repositioning, i.e. drugs that were not yet associated with being a treatment for cancer, but might be useful assets against cancer in the future. Drug repositioning is particularly valuable for clinically approved drugs, since those compounds could skip various steps in drug development, e.g. toxicity in patients was already ruled out through previous clinical trials. Therefore, such a repositioned compound could take a short cut in clinical development and only requires further confirmation of efficacy in cancer.

Another application of this approach would be to prioritise a list of novel compounds, which should be added in an experimental screen for increasing chances to find potential hits. However, the GDSC drug screen was already

biased towards clinically approved targeted therapies. Therefore, *in silico* screens of novel compounds might be biased by the initial set of drugs.

4.6: Conclusion

In this chapter, I compared models based on genomics alone against models additionally leveraging the chemotype. As a result, I found that models trained with both data streams moderately performed better when predicting mono-therapies than models solely based on genomics. I believe this moderate improvement is the result of leveraging more training data to build the model, but is also due to the fact that the chemotype is related to the MoA in many instances. For further improving these models and better capturing the MoAs, the integration of bioactivity profiles might be the next leap forward. Also the integration of detailed pathway knowledge may improve the predictive power. In essence, improvement may be achieved by adding novel informative features, but also by improving the current existing set, e.g. including other key tumour drivers from prior knowledge.

In addition to my original hypothesis that chemistry might improve predictive power, here I showed a potential application of genomics and chemistry models. These new models were capable of predicting biomarker tendencies (effect size signs), although predicting exact drug response remained a challenge.

The most promising future application of genomic and chemistry models might be a systematic drug repositioning, or in other words, an *in silico* screen of potential new cancer treatments and potential biomarkers. Notably, this was technically not possible with previous models based on either genomics or chemistry alone, but this remains to be explored in follow up studies.

Chapter 5: Leveraging germline variations for drug response

"*Prediction is very difficult, especially if it's about the future*"

- Niels Bohr

5.1: Declaration of contribution

Under the lead of Ultan McDermott and Mathew Garnett within the scope of Genomics of Drug Sensitivity in Cancer (GDSC) project, all wet lab experiments were independently carried out by them and I share no contribution. I curated and prepared the data for this particular study, unless stated differently. Dry lab experiments for this association study were jointly planned with Julio Saez-Rodriguez, Oliver Stegle, Francesco Paolo Casale, Johannes Stephan and myself. Francesco Paolo Casale and Johannes Stephan ran the existing LIMIX association pipeline and eQTL scan. I parsed the data in a LIMIX suitable format, while Francesco Paolo Casale additionally performed the LD filtering on germline mutations. Furthermore, my independent contribution was presenting the results, visualisation, biological interpretation and outlining the gained medical insights. This chapter is currently in preparation for submission.

5.2: Introduction

Germline variations are inherited from parents to offspring, while somatic mutations are accumulated during the life span of an individual (Vogelstein & Kinzler, 1993) (Figure 5.1). The occurrence of somatic mutations is a stochastic process whose rate may depend on germline variations (Stratton et al., 2009). Additionally, somatic mutation rates can be increased by environmental factors such as smoking (Gibbons, Byers, & Kurie, 2014), UV light exposure (Saladi & Persaud, 2005), alcohol consumption (Ahrendt et al., 2000) and unhealthy nutrition (Donaldson, 2004), etc.

Over the last 20 years life expectancy increased but the healthy life expectancy in proportion only slightly grew (Salomon et al., 2012). Therefore, the question should be whether we live long enough to develop cancer, which is inevitable but also depends on our inherited germline variations, stochastic processes, age, environmental factors and personal habits.

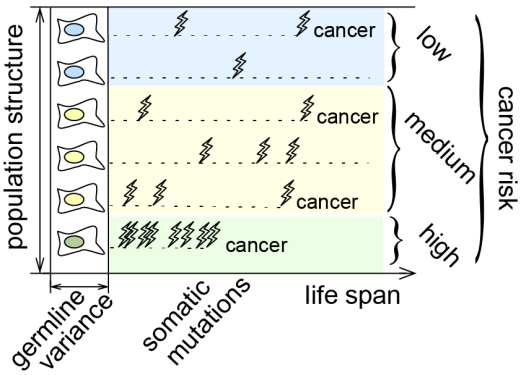

Figure 5.1: Nature of germline variations and somatic mutations.

Germline variations are inherited genetic differences, while somatic mutations occur during the life span of an organism. Germline variations are differences between individuals, but are identical within an individual. Somatic mutations locally occur in a single cell and are passed on to daughter cells after cell division. The somatic mutation rate may be altered by the germline background. Most somatic mutations are tolerated (i.e. passengers), however, some might be cancer somatic (i.e. oncogenic) and may increase cell proliferation, avoiding the controlled cell death (apoptosis), and ultimately causing cancer.

5.2.1: Germline variants for estimating cancer risk

Germline variations and their phenotypic traits are used in clinics to estimate the risk of developing certain cancer types. For example, mutations in BRCA1 and BRCA2 increases the likelihood of developing breast and ovarian cancer (Antoniou et al., 2003; Pal et al., 2005; Walsh et al., 2006). Both genes are involved in DNA damage repair. A loss-of-function event in BRCA1 or BRCA2 increases the risk to progress cell division with damaged DNA. This leads to a faster accumulation of somatic mutations and finally increases chances of producing tumour potential cells compared to human beings with functional BRCA1 and BRCA2. Implications are that individuals carrying risk alleles should be more often screened for certain cancer types, which would ultimately enable early diagnosis and initiation of appropriate treatment in time (Moyer, 2014).

5.2.2: Treatment decisions based on somatic and germline mutations

As previously discussed, somatic mutations occur stochastically during the lifespan of an individual (Stratton et al., 2009). Most somatic mutations have no impact on human health (i.e. passenger mutations), but some may lead to cancer (i.e. oncogenic mutations). Based on the nature of somatic mutations, which are the difference between healthy and cancer cells, their clinical application is straightforward and involves their use as biomarkers for treatment decisions (Figure 5.2 A).

Every single cell in a human individual contains the exact same genetic material, except somatic mutations. Germline variations are identically abundant in every single cell within an individual. Therefore, the broad application of germline mutations is their association with drug toxicity and predicting pharmacokinetics (Coate et al., 2010) (Figure 5.2 B). Germline variations can be used to separate patients into weak or strong responders, and are used to estimate toxicity and side effect strengths.

For example, the metabolism from SN-38 (i.e. the active component of camptothecin) to the pharmacologically active glucuronidated SN-38 (i.e. SN-38G) is decreased in patients with polymorphisms in chromosome 2 at

location 2q37 in the gene *UGT1A1* (Xiaohong Chen et al., 2012). Particularly, patients treated with camptothecin and harbouring an additional TA (i.e. $[TA]^7TAA$ instead of $[TA]^6TAA$ wild type) in the promoter TATA box of UGT1A1 lack efficient metabolism from SN-38 to SN-38G, which increases the risk of diarrhoea and leukopenia (Gagné et al., 2002). In leukopenia (i.e. neutropenia) the white blood cell count is reduced which further leads to a decrease in immune system functionality (Ing, 1984).

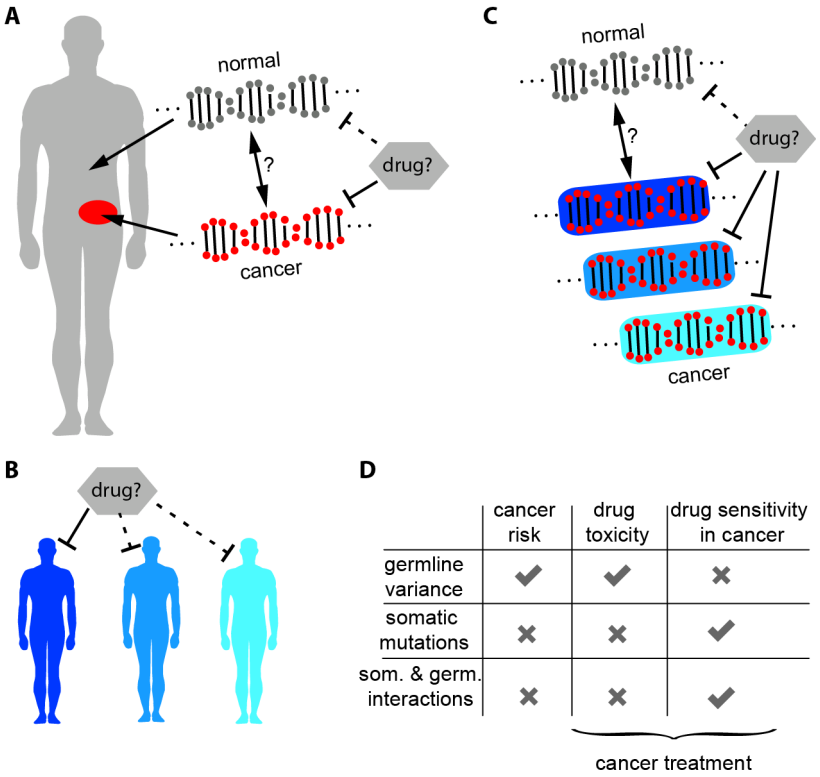

Figure 5.2: Clinical applications of germline and somatic mutations.

(A) Shows differences between cancer and normal tissue that may be potential therapeutic targets. A drug is desirable if it targets cancer but not healthy tissue. (B) Germline mutations in identifying toxicities. Patients with different inherited germline backgrounds might respond differently to treatment, e.g. some patients might be able to metabolise a compound and some others not. (C) Somatic mutations might respond differently under certain germline conditions, in essence, somatic mutations might interact with germline variations and translate to novel sensitivity biomarkers. (D) Germline variations are used to predict cancer risk and drug toxicity. In contrast, somatic mutations are applicable as biomarkers of drug sensitivity. However, somatic mutations and their interaction with germline variations could be used as markers for cancer drug sensitivity. Human silhouette adapted from *("Human Free Vector," 2015)*.

In the past, germline and somatic mutations were mostly studied separately due to their different nature and clinical applications. Germline variations were mainly used as toxicity markers, while somatic mutations served as biomarkers of drug sensitivity. Germline traits are shared in every single cell within an individual and those are not suitable drug sensitivity markers (i.e. occur in healthy and cancer). However, in my study I additionally investigated the interactions of somatic and germline background (Figure 5.2 C), which would be again specific for cancer response. I tested if somatic mutations within different germline traits gave novel insights for patient stratification. Such interaction of germline and somatic mutations occur locally in tumours and not in the whole patient, therefore, making them an additional marker of drug sensitivity.

In summary, my hypothesis was that cancer drug sensitivity would be determined by somatic mutations and their interaction with germline variations. Germline variations alone set a baseline response to a drug, i.e. general toxicity. In this chapter I explored drug efficacy based on germline variants, somatic mutations and their interactions.

5.3: Methods

5.3.1: Drug response screen

I used the drug response from Genomics of Drug Sensitivity in Cancer (GDSC) project release-5.0 (http://www.cancerrxgene.org/downloads) (Garnett et al., 2012; W. Yang et al., 2013), which contained 140 drugs and 707 cell lines. This data set was extended by the drug response of 269 unpublished cell lines, which were additionally provided by GDSC summing up to 976 cell lines in total. This extended data will be part of the next GDSC web release. Drug response was measured in 1-AUC (see 2.3.4: Drug response curve fitting and metrics).

5.3.2: Cancer somatic dataset

The cancer somatic dataset was the same as described on page 90, Chapter 3, section 3.3.2: Genomic dataset. Furthermore, the pre-processing is the same as previously outlined previously.

5.3.3: Germline dataset

Germline variants were extracted from Affymetrix SNP6.0 microarrays, which captured the simple genotype (i.e. disregard copy number information) of common single nucleotide polymorphisms (SNPs). For example, a SNP could either be homozygous AA, heterozygous AT or homozygous TT. Now, assuming AA would be the minor allele, the SNP would be encoded as following: $AA \mapsto 0$, $AT \mapsto 1$ and $TT \mapsto 2$.

SNPs with low minor allele frequency (MAF) led to inflated statistics, i.e. often associations with strong effect sizes were observed in SNPs with low MAF due to limitations of the association test. Therefore, SNPs with a minor allele frequency lower than 2% were discarded from analysis, which shrank the putative germline set from 884,145 to 650,130 SNPs.

To guarantee that each SNP was from a germline origin rather than a somatic mutation, Francesco Paolo Casale applied a linkage disequilibrium (LD) filter.

The linkage disequilibrium (Golding, 1984) was based on the observation that germline variations were inherited from parents to offspring and gradually changed over various generations, therefore, germline variants in close proximity of the genome remained conserved. In contrast, somatic mutations stochastically occured during the live span, making them punctual events unrelated to surrounding mutations. This LD filter calculated the r^2 from the 50 closest neighbouring SNPs and removed all SNPs where the r^2 was lower than 0.5. This left me with a final set of 489,285 germline SNPs.

5.3.4: eQTL

An expression Quantitative Trait Loci (eQTL) is a SNP associated with the expression of a particular gene. The GDSC team measured gene expression with Affymetrix Human Genome U219 Array Plates (see section 3.3.4: Gene expression) while Francesco Paolo Casale calculated the eQTLs. Motivation for identifying the eQTLs was to filter for interpretable SNPs. Particularly, exploring somatic and germline interactions of all combinations would lead to a larger number of statistical tests and multiple hypothesis testing would 'kill' all signal. Therefore, I filtered interactions for germline variants that are eQTLs.

A cis-eQTL is a SNP in close proximity to the gene, which expression it affects. I assumed a window of 1Mb upstream and downstream of each of the 16,481 genes and for identifying the lead eQTL SNP (i.e. most significant SNP for each gene). Francesco Paolo Casale used a linear mixed model where the phenotype is gene expression, and generated p-values with a log-likelihood ratio test.

I identified 1,856 SNPs as significant eQTLs at a false discovery rate (i.e. FDR) of 20% (Benjamini & Hochberg, 1995). This set of 1,856 SNPs was used for interaction models between germline and the 315 somatic mutations (see section 5.3.5: Interactions between somatic and germline).

5.3.5: Somatic, germline and interaction linear mixed-effect model

The analysis was based on linear mixed-effect models, which were jointly designed and developed with Francesco Paolo Casale, and implemented and ran by him using the existing LIMIX pipeline (Lippert, Casale, Rakitsch, & Stegle, 2014). The linear mixed model combined fixed and random effects to explain a particular phenotype across samples, where random effects were a distribution with a fitted covariance matrix (e.g. noise in the data) and all remaining terms were defined as fixed effects.

I built models for 3 different use cases:

i. Somatic model (equation 5-1)
ii. Germline model (equation 5-2)
iii. Interactions between somatic and germline (equation 5-3)

For all models I defined a null model (i.e. $H0$) and alternative model (i.e. $H1$), which I tested for 'goodness-of-fit' with a log-likelihood ratio test (Wilks, 1938). The effect size between $H0$ and $H1$ was assumed to be zero, and the log-likelihood ratio test estimated p-values for the improvement of $H1$ over $H0$. For making the p-value estimates robust against outliers, the drug response was quantile normalised to a Gaussian distribution (Bolstad, Irizarry, Astrand, & Speed, 2003), while for the fitting of reported effect size, the raw drug response was used for reflecting the dynamic range of drug response. P-values were multiple hypotheses corrected with a false discovery rate (FDR) (Benjamini & Hochberg, 1995).

Notably, all models included noise (i.e. random effect) in $H0$ and $H1$, but also the tissue type as a covariate of the model (i.e. fixed effect). This was necessary since tissue type in a pan-cancer setting effectively explained drug response, however, tissue type on its own would be a fairly trivial biomarker regarding clinical decisions, e.g. treat a breast cancer patient with a breast cancer drug (see 3.5.3: Tissue effect in pan-cancer study).

For technical purposes, I also included in all our linear mixed models a random interception. This allowed me to fit for each drug a different intercept, for improving the linear fit overall, i.e. the linear regression was not forced to intersect at 0 but could be randomly shifted for improving the fit.

Somatic model

My general aim was to fit $\boldsymbol{y}$, which was the drug response in $1 - AUC$ of a given compound. To test for somatic associations, I included a binary matrix $\boldsymbol{X}_s$ of 315 cancer somatic mutations (i.e. columns) measured in 976 cell lines (i.e. rows). The effect size of each somatic mutation was measured in the vector $\boldsymbol{\beta}_s$.

$$\begin{aligned} H0&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ H1&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ &where\ \boldsymbol{\psi} \sim N\,(0,\ \sigma_e^2\boldsymbol{I}) \end{aligned} \qquad \text{(Eq. 5-1)}$$

As outlined above, all models had a noise term $\boldsymbol{\psi}$ (i.e. random effect), which was assumed to be normally distributed with mean 0 and an unknown covariance matrix $\sigma_e^2\boldsymbol{I}$ (i.e. variance-covariance or dispersion matrix), which was fitted during training.

Also mentioned previously, to discard trivial molecular associations as a result of correlations with the tissue label, I added the tissue term, $\boldsymbol{T\alpha}$, as a fixed effect in my $H0$ and $H1$ model. $\boldsymbol{T}$ was a matrix binary encoding the tissue-of-origin of all cell lines and $\boldsymbol{\alpha}$ was a vector denoting the effect size of tissue effect. In essence, this was how I corrected for the tissue effects.

Notably, $\mathbf{1}\vartheta$ was the random interception, where ϑ was a scalar and $\mathbf{1}$ a vector exclusively containing ones.

Germline model

For evaluating germline associations, I exchanged the term $\boldsymbol{X}_s\boldsymbol{\beta}_s$ (equation 5-1) for $\boldsymbol{X}_g\boldsymbol{\beta}_g$ (equation 5-2), where $\boldsymbol{X}_g$ was a matrix containing the

simple genotype of the filtered 489,285 germline SNPs (i.e. columns) measured in 976 cell lines (i.e. rows), and $\boldsymbol{\beta}_g$ was its effect size vector of each SNP, respectively.

$$\begin{aligned} H0&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{g} + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ H1&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{g} + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ &where\ \boldsymbol{g} \sim N\,(0, \sigma_g^2\boldsymbol{K}) \\ &and\ \boldsymbol{\psi} \sim N\,(0, \sigma_e^2\boldsymbol{I}) \end{aligned} \qquad \text{(Eq. 5-2)}$$

Furthermore, to build the germline model, I included another random effect, $\boldsymbol{g}$, for capturing the population structure (equation 5-2), which reflects the shared ancestry of the individual cell lines (Weir, Anderson, & Hepler, 2006). $\boldsymbol{g}$ was assumed to be normally distributed with mean 0, and $\sigma_g^2\boldsymbol{K}$ was the covariance matrix measuring the genetic relatedness.

Interactions between somatic and germline

The interaction model combined all terms from the somatic (equation 5-1) and germline model (equation 5-2).

$$\begin{aligned} H0&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{g} + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ H1&: \boldsymbol{y} = \boldsymbol{T\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + \boldsymbol{X}_g\boldsymbol{\beta}_g + \boldsymbol{X}_s\odot\boldsymbol{X}_g\boldsymbol{\beta}_{gs} + \boldsymbol{g} + \boldsymbol{\psi} + \mathbf{1}\vartheta, \\ &where\ \boldsymbol{g} \sim N\,(0, \sigma_g^2\boldsymbol{K}) \\ &and\ \boldsymbol{\psi} \sim N\,(0, \sigma_e^2\boldsymbol{I}) \end{aligned} \qquad \text{(Eq. 5-3)}$$

The interaction model included an interaction term, $\boldsymbol{X}_s\odot\boldsymbol{X}_g\boldsymbol{\beta}_{gs}$, where $\odot$ denoted the Cartesian product of $\boldsymbol{X}_s$ and $\boldsymbol{X}_g$. In other words, each element of $\boldsymbol{X}_s$ and $\boldsymbol{X}_g$ was tested, i.e. each somatic and germline pair. $\boldsymbol{X}_s$ was exactly the same matrix as described in the somatic model (equation 5-1). $\boldsymbol{X}_g$ contained germline variations but was filtered for eQTLs (see section 5.3.4: eQTL) for reducing the number of tests and focusing on interpretable germline variations. $\boldsymbol{\beta}_{gs}$ was the effect size of the interaction model.

Interactions were statistically more difficult to identify than additive effects (e.g. somatic or germline on their own), since interactions focused on cell

lines that carried a particular somatic mutation and germline variant, which decreased the original training set size. This results in many statistical tests that were underpowered. Therefore, only interactions with at least 10 non-zero elements in $\boldsymbol{X}_s \odot \boldsymbol{X}_g$ were considered.

5.4: Results

5.4.1: Comparing lead somatic and lead germline

For the purpose of comparing somatic versus germline models, in a first attempt I compared the 'lead' somatic mutations versus 'lead' germline variations across all drugs (Figure 5.3). The so-called 'leads' were the most significant somatic and germline association for a particular drug.

There was a tendency that the drug response of a compound was either significantly associated with a somatic mutation or germline variation (Figure 5.3 A).

Somatic mutations were more often significantly associated than germline variants, which was driven by biology but also notably influenced by incorporating prior knowledge in the selection process of oncogenes (i.e. somatic mutations). For somatic mutations I focused on 315 oncogenes (see 5.3.2: Cancer somatic dataset), while for germline variation I inspected 489,285 SNPs (see 5.3.3: Germline dataset). Therefore, the number of tests for somatic and germline was different, and consequently a more stringent false discovery (i.e. FDR) threshold (Benjamini & Hochberg, 1995) was chosen for germline variants. For somatic association the 20% FDR threshold was p-values of ~$10^{-3.5}$, while for germline association the threshold was more stringent at ~10^{-7}.

The effect sizes of somatic mutations was larger compared to the germline variations (Figure 5.3 B). This was biologically driven, but also due to the fact that germline variants were filtered with minor allele frequency smaller than 2% but somatic mutations not. Therefore, somatic mutations might occur with lower frequency than germline mutations, which might inflate p-values of somatic mutations.

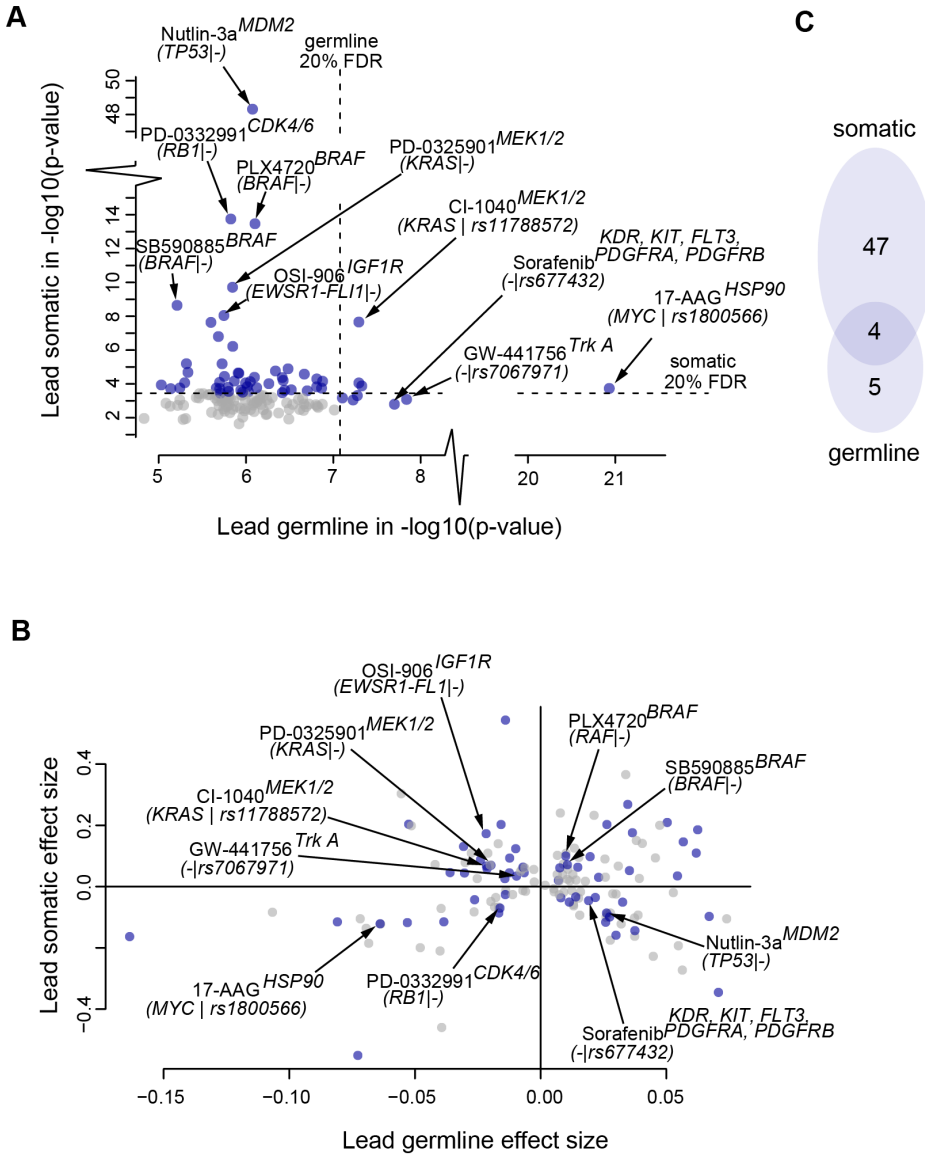

Figure 5.3: Comparison of lead germline and lead somatic mutations.

(A) Shows the lead germline and lead somatic mutations on the x-axis and y-axis, respectively, where 'lead' was the most significant association for each drug. Axis are in -log10 scale and blue dots are significant (20% FDR), while grey dots are not. The horizontal stripped line indicates the 20% FDR threshold for somatic mutations, while the vertical stripped line is the threshold for germline variations. (B) Effect sizes of the lead somatic and lead germline models. All labelled models in (A) were also tagged in (B). (C) Count of significant associations with either germline or somatic, or with both associated.

Out of 140 compounds, 9 drugs were significantly associated with one SNP, while 50 drugs are significantly associated with somatic mutations (Figure

5.3 C). The significant overlap between somatic and germline association was 3 compounds.

Lead somatic associations

The most significant identified somatic association was the drug response of nutlin-3a and TP53 mutations (Figure 5.3 A). Notably, the protein product TP53 of TP53 gene could be inhibited by E3 ubiquitin-protein ligase Mdm2 (i.e. MDM2 gene). Nutlin-3a targets MDM2, and therefore neglects the inhibition of TP53, i.e. TP53 functionality can be suppressed by increased MDM2 expression and activity (Vassilev et al., 2004). As previously discussed, TP53 is a key tumour suppressor which is mutated in ~50% of all tumours, which generally leads to a loss-of-function event in TP53. In the case that TP53 is mutated, the MDM2 inhibition is redundant and has no further effect on cell viability. Therefore, the TP53 mutation was identified as a resistance marker of nutlin-3a.

PLX4720 (i.e. progenitor of vemurafenib) and SB590885 are BRAF inhibitors that both were associated with BRAF mutations (Figure 5.3 A). Particularly, the $BRAF^{V600E}$ mutation caused a constant activation of the ERK signalling pathway in melanoma, and ultimately increased cell proliferation (Chapman et al., 2011).

PD-0332991 is a CDK4/6 inhibitor that I associated with RB mutations. RB is a tumour suppressor, where active RB inhibits the cell cycle and particularly causes arrest in G1 to S-phase transition. CDK4/6 promotes phosphorylation of RB and thereby deactivates RB tumour supressing function (Konecny et al., 2011). A loss-of-function mutation within RB would be redundant to the treatment with CDK4/6 inhibitors and therefore have no positive treatment effect, making RB mutations a marker of resistance to CDK4/6 inhibition.

I also identified the *EWSR1* and *FLI1* fusion gene (i.e. *EWSR1-FLI1*), which was associated with the drug response to OSI-906. This drug targets the IGF1 receptor. Approximately 85% of ewing sarcoma are driven by *EWSR1-FLI1*

fusion (Owen, Kowalewski, & Lessnick, 2008) and the association between this fusion gene and IGF1R was also concordant with the literature (Prieur, Tirode, Cohen, & Delattre, 2004).

Lead germline and lead somatic associations

As previously outlined, there was a trend that a drug was either linked to somatic or germline associations. However, there were also a few instances where both associations were observed at a significant level.

For example, CI-1040 is a MEK1/2 inhibitor that was associated with the lead oncogene KRAS and the germline SNP rs11788572 (Figure 5.3 A). The association with KRAS was concordant with the literature, which is the lead biomarker for MEK1/2 inhibition in colorectal cancer cells (Little et al., 2011; Yeh et al., 2009). The germline SNP was located in the intron region of *NTNG2*, which is an axonal membrane adhesion protein (Matsukawa et al., 2014; Yaguchi, Nishimura-Akiyoshi, Kuroki, Onodera, & Itohara, 2014) and is involved in axon guidance (Seiradake et al., 2011).

Notably, NTNG1 (i.e. same protein family as NTNG2) had previously been associated with ERK signalling activation (Forcet et al., 2002), and this might be the missing link between NTNG2 and the drug response to CI-1040.

The KRAS mutation was the lead association for CI-1040 (i.e. above discussed MEK1/2 inhibitor), but CI-1040 was also significantly associated with other oncogenes such as BRAF (Figure 5.4). Surprisingly, KRAS was the lead oncogene and not BRAF, which is immediately upstream of the drug target (i.e. MEK1/2). Notably, BRAF mutations were the main driver in melanoma, and ~75% of all melanoma cell lines of the GDSC panel were BRAF mutants. The predominant occurrence of the BRAF mutation in melanoma and correction for tissue type down weighed the importance of BRAF associations in this pan-cancer setting, leaving KRAS as the lead somatic association upon CI-1040 treatment.

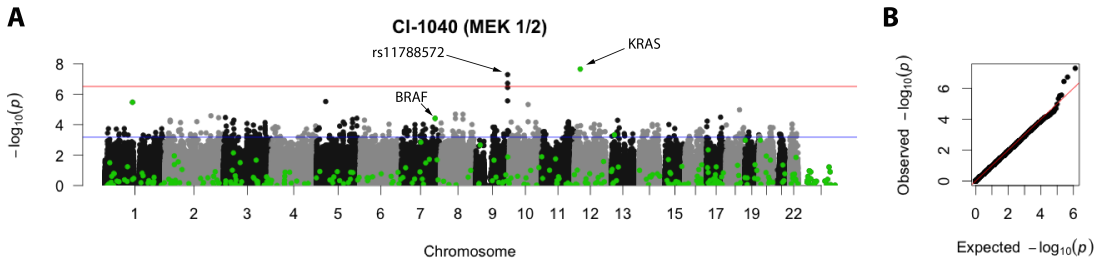

Figure 5.4: Manhattan plot of CI-1040.

(A) Manhattan plot of CI-1040 (i.e. MEK1/2 inhibitor), where on the x-axis are the chromosomal positions and on the y-axis the corresponding p-value. Germline variants are in black and grey (odd and even chromosome number, respectively). Somatic mutations are dots in green. 20% FDR threshold are horizontal lines for somatic and germline mutations in blue and red, respectively. (B) Expected versus observed p-values (i.e. QQ-plot).

It has to be noted that KRAS mutations occurred in different tissue types with high frequency hence it was a pan-cancer event. For example, ~96% pancreatic (PAAD) cancer cell lines were KRAS mutant, ~54% colon and rectal (COREAD) cancer lines, ~44% thyroid (THCA) cancer lines and ~33% lung adenocarcinoma (LUAD) cell lines are KRAS mutant. Notably, KRAS mutant frequencies in cell lines aligned fairly with the frequencies observed in primary tumours, i.e. in PAAD ~95% KRAS mutant, in COREAD ~35%, in THCA ~55% and in LUAD ~35% were KRAS mutants (Kranenburg, 2005).

Germline variants associated with drug response appeared as a peak in the Manhattan plot, in contrast to somatic mutations, which had no co-correlated mutations in close proximity (Figure 5.4). This was due to the linkage disequilibrium (Golding, 1984) (see also 5.3.3: Germline dataset).

Germline associations

I found 9 significant 'lead' germline associations (i.e. best hit per drug) with top hits being most strikingly 17-AAG (rs1800566), followed by GW-441756 (rs7067971) and sorafenib (rs677432) (see page 141, Figure 5.3 A).

Sorafenib was a FDA-approved drug for thyroid, liver and kidney cancer. This compound putatively targeted PDFGRA, PDGFRB, KDR, KIT and FLT3. I

found the ‘lead’ germline SNP rs677432 to be associated with the sorafenib drug response. rs677432 was located in the intronic region of FRMD4A (i.e. FERM domain containing 4A), a gene that had been associated with the regulation of epithelial polarity (Ikenouchi & Umeda, 2010).

GW-441756 is a TrkA inhibitor and I found it to be associated with the germline SNP rs7067971, which was on chromosome 10 in close proximity to the ABC transporter *ABCC2*. ABC transporters are involved in a well-known mechanism of multi-drug resistance, where in particular, cancer cells exploit ABC transporters as efflux pumps for chemotherapies (Evers et al., 1998). Despite the proximity of rs7067971 to *ABCC2*, and additional to the vicinity of a CTCF binding site and enhancer regions, I found no evidence that this germline variant affects the expression of ABCC2 directly (Figure 5.5). This result was consistent with previous findings reporting that rs7067971 was not an eQTL for ABCC2 (Nguyen et al., 2013). The underlying mechanism remained unexplained, however, my analysis revealed that rs7067971 was associated with drug response to GW-441756.

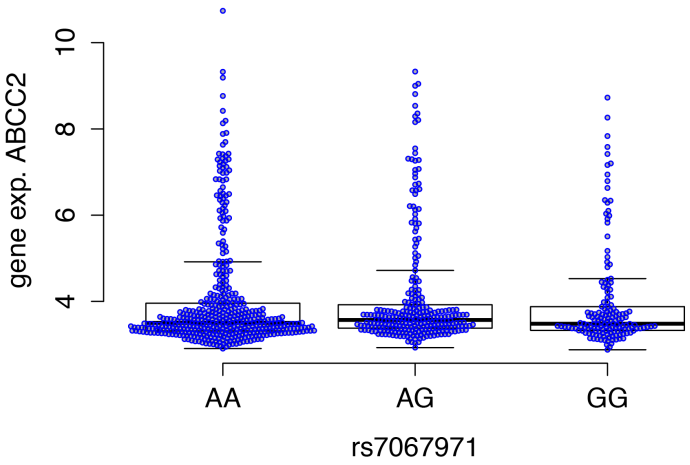

Figure 5.5: ABCC2 expression is independent from rs7067971.

Gene expression of ABBC2 stratified by the germline SNP rs7067971 into AA, AG and GG allele.

17-AAG and its strong dependency on germline variation

The most striking germline association was the SNP rs1800566 with 17-AAG (Figure 5.6 A). The associated SNP rs1800566 from C>T causes a structural change from Proline>Serine at position 186 in NQO1 (i.e. $NQO1^{C609T}$, Figure 5.6 B). *NQO1*'s protein product is often also referred to as DT-diaphorase. NQO1 is an enzyme that reduces quinones to hydroquinones. This substitution is a change from a hydrophobic to a nucleophile amino acid in a disordered region close to the active pocket. PolyPhen predicts $NQO1^{C609T}$ as being a 'probably damaging function' of the enzyme (Adzhubei et al., 2010).

17-AAG was the first HSP90 inhibitor taken into clinical trails, derived from the cytotoxic geldanamycin. The targeted HSP90 is a chaperone that stabilises several kinases such as AKT, ERK signalling members and CDK proteins while folding. HSP90 is a housekeeping gene expressed in healthy and cancer cells, and is believed to be vital in many tumours. Unlike novel HSP90 inhibitors such as AUY922, 17-AAG relied on functionally expressed NQO1 (Figure 5.6 C). 17-AAG was a quinone-like compound that needed to be metabolised by NQO1 into the pharmacological active version 17-AAGH2. This metabolised compound became a more potent HSP90 inhibitor than the unmodified 17-AAG, which required functionally expressed NQO1 to be metabolised (Kelland et al., 1999).

To rule out a direct effect of rs1800566 on the expression of *NQO1* I explored the gene expression of *NQO1* and the dependency of rs1800566 (Figure 5.6 D). The gene expression of the CC & CT genotypes was not significantly different from TT (t-test p-value=0.29). This confirmed that the SNP exerts its functional effect through the structural change in NQO1, and is independent from *NQO1* expression.

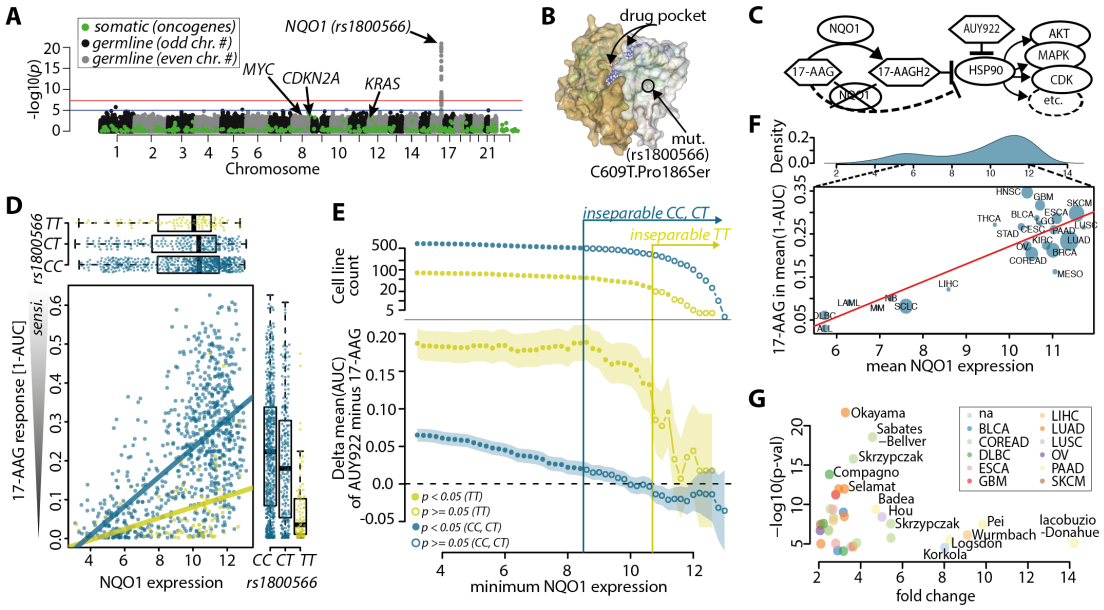

Figure 5.6: 17-AAG in detail.

(A) Manhattan plot of 17-AAG drug response. On the x-axis are the chromosomes while on the y-axis are the p-values of each SNP. In black and grey are the germline SNP, while in green are the somatic mutations. (B) 3D structure of NQO1 *(Asher, Dym, Tsvetkov, Adler, & Shaul, 2006)*, which is visualised with Protein Workshop (J. Moreland, Gramada, Buzko, Zhang, & Bourne, 2005) (PDB ID: 2F1O). (C) Mode-of-action of 17-AAG and AUY922, notably both are HSP90 inhibitors. 17-AAG needs to be metabolised into 17-AAGH2 to become a potent HSP90 inhibitor. (D) Gene expression versus drug response of 17-AAG. In blue are the function alleles of NQO1 while in yellow are the non-functional TT genotype. (E) Difference between AUY922 versus 17-AAG. I plot as a function of NQO1 expression, the resulting difference in drug response between AUY922 and 17-AAG. Colour coding is the same as in (D), while solid filled points are significantly separable, while empty points correspond to not distinguishable between AUY922 and 17-AAG responses. (F) 17-AAG drug response as well as the gene expression of NQO1 are tissue dependent. (G) Differential gene expression analysis of NQO1 between cancer patients versus normal. This analysis was performed with ONCOMINE (Rhodes et al., 2004).

Concordantly, the correlation of NQO1 and drug response to 17-AAG was strongly influenced by separating cell lines into CC against TT genotype and CT against TT genotype, corresponding to 0.59 (Kendall's tau p-value<2.2×10^{-16}) and 0.38 (Kendall's tau p-value=0.17×10^{-3}), respectively. Since there was no direct effect of the SNP on the expression of NQO1, this implied that their interaction determined the drug response to 17-AAG in our pan-cancer analysis.

In contrast, in 2005 a phase I clinical trial of 17-AAG of 21 patients with advanced cancer led to the conclusion that the heterozygous CT and homozygous TT genotype of the SNP (rs1800566) in NQO1 did not affect the clearance or toxicity of 17-AAG (Goetz et al., 2005). In this study, 15 patients with homozygous CC alleles for the active form were compared against six individuals carrying either the homozygous TT allele or the heterozygous CT genotype, although I showed previously that CT produces a rather functional NQO1 (Figure 5.6 D). However, in this clinical trial both TT and CT were considered as non-functional. This initial phase I trial shaped several other 17-AAG phase I & phase II trials in which rs1800566 was often discarded (Bagatell et al., 2007; Heath et al., 2008; D. B. Solit et al., 2008; David B Solit et al., 2007; Weigel et al., 2007).

Having 890 cell lines available, the comparison of CT and TT against CC still led to a significant separation in our dataset (t-test p-value= 1.3×10^{-9}). However, I found a much stronger difference when binning CC and CT and comparing it against the non-functional form TT (t-test p-value<$2.2\text{x}10^{-16}$). This suggests that the heterozygous form CT was more similar to the functional homozygous CC than the non-functional TT. In agreement with this, I found a strong difference between TT and CT, but a minor one from CC to CT (t-test p-value=0.018). Hence, my results suggest that the clinical trial above might have derived an inaccurate conclusion about the importance of rs1800566 in drug response to 17-AAG.

Besides the binning of CT with functional CC rather than with non-functional TT, an important factor that allowed me to obtain a strong statistical signal was the much larger sample size, with 890 samples (575 CC, 221 CT and 94 TT) instead of 21 patients. It must be noted that the minor allele frequency was 0.28 (TT) as observed in the 1,000 genome project (Abecasis et al., 2012), an important fraction of patients might not benefit from 17-AAG treatment compared to the other HSP90 inhibitors which do not require NQO1.

Modern HSP90 inhibitors such as AUY922 are considered superior to 17-AAG, due to the fact that they are independent from NQO1. Notably, *NQO1* was expressed in a few tissue types, making it a suitable HSP90 inhibitor for a subset of cancer types (Figure 5.6 F). Interestingly, *NQO1* was differentially expressed in cancer versus normal for many tissue types (Figure 5.6 G), turning the 'disadvantage' of 17-AAG relying on NQO1 into a potential strength in certain circumstances. Such a hypothesis had been already proposed in the context of non-small lung cancer (D Siegel, Franklin, & Ross, 1998; David Siegel & Ross, 2000), and potentially could be extended to multiple tissue types such as pancreatic cancer, glioblastoma, etc. 17-AAG might reduce toxicities compared to other HSP90 inhibitors in cancers with *NQO1* overexpression, although secondary resistances induced by down regulation or mutation of NQO1 remain an obstacle (Gaspar et al., 2009).

5.4.2: Interactions between germline and somatic mutations

Identification of drug sensitivity biomarkers is the basis for personalised medicine. Usually such biomarkers were cancer somatic mutations (i.e. oncogenes), while here I investigated the interaction of oncogenes with germline mutations. For reducing the number of tests to a manageable size and interpretability of results, I focused on 1,856 cis-eQTLs (see section 5.3.4: eQTL). Furthermore, I filtered for a minimal number of 10 cell lines carrying the oncogene in a minor and heterozygous allele. For my 140 tested drugs, the above outlined filters were satisfied in 9,445 instances. This resulted in 12 associations that were having a q-value lower than a 20% false discovery rate (Figure 5.7 A).

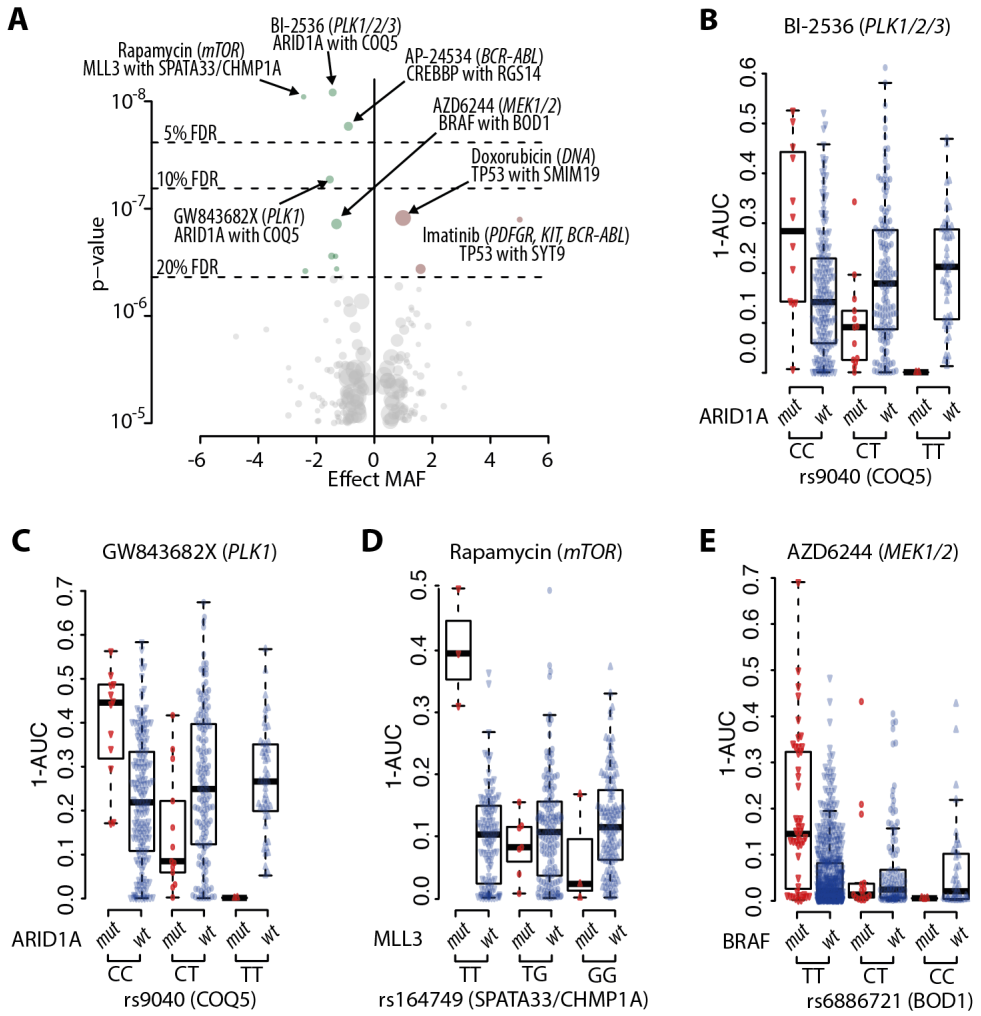

Figure 5.7: Interactions of germline and somatic mutations.

(A) Top interactions between germline and somatic mutations regarding drug response of a compound. On the x-axis is the effect size of the minor allele frequency, and on the y-axis the p-value. Annotations of interactions are the drug name, in brackets the putative target, below is written the somatic mutation and interacting eQTL. (B) and (C) illustrates boxplots of drug response from PLK inhibitors. In red are the somatic mutant cell lines while blue are wild type. The populations were split by germline background, e.g. CC, CT or CT. (D) Shows rapamycin association with MLL3 and SPATA33/CHMP1A. (E) Illustrates a MEK1/2 inhibitor associated with the interaction of BRAF and BOD1.

Reassuringly, the drug response of two structural different PLK inhibitors, BI-2536 (Figure 5.7 B) and GW843682X (Figure 5.7 C) were associated with the same interacting oncogene and eQTL, ARID1A and COQ5, respectively. PLK1 inhibitors regulate the progression of the cell division, however, their mechanistically function is not fully understood yet (Plyte & Musacchio, 2007).

ARID1A is part of the SWI/SNF chromatin-remodelling complex and is potentially an epigenetic tumour suppressor (Wu & Roberts, 2013). COQ5 is a methyltransferase that is involved in converting DDMQH2 to DMQH2 (Dai et al., 2014). There might be a biological connection between COQ5 and ARID1A that influenced drug sensitivity of PLK1 inhibitors.

Another weak association was the MTOR inhibitor rapamycin and the interaction of MLL3 and the SNP rs164749 (Figure 5.7 D). This SNP was an eQTL that influences the expression of *SPATA33* and *CHMP1A*. SPATA33 is the spermatogenesis-associated protein 33 while CHMP1A is the charged multivesicular body protein 1a that is hypothesised to be involved in chromosome condensation (Stauffer, Howard, Nyun, & Hollenberg, 2001). MLL3 is a histone methyltransferase. CHMP1A and the interaction with MLL3 might play a role in the drug response to rapamycin, while none of them alone explained the drug sensitivity.

Likewise, the association between AZD6244 and the interaction between BRAF and BOD1 is biologically and clinically interesting (Figure 5.7 E). MEK1/2 inhibitors such as AZD6244 were commonly applied in clinics for suppressing the overregulated ERK signalling, which phenotypic outcome otherwise would be uncontrolled constant cell proliferation among other hallmarks of cancer. BOD1 is responsible for the bi-orientation of chromosomes in cell division protein 1, which mislocates MCAK and therefore disturbs the efficient phosphorylation of Aurora B (Porter et al., 2007). Aurora B is the checkpoint kinases for mitosis and is overexpressed in many cancer types (Gully et al., 2012). Furthermore, Aurora B is downstream of the ERK signalling (Bonet et al., 2012). As previously described, the $BRAF^{V600E}$ mutation caused a constant activation of ERK signalling, and drug response of AZD6244 for BRAF mutant lines was increased when within the eQTL rs6886721 TT allele. This germline association might be used to further stratify BRAF mutant patients into responders and non-responders.

To summarise, the reported interactions have to be interpreted carefully due to relatively high p-values (low confidence), however, they revealed new

insights into drug sensitivity. Germline mutations on their own were mainly considered in risk assessment or explaining toxicity, while when considering their interaction with somatic mutations, which distinguished healthy from cancerous tissue, they became markers of drug sensitivity.

5.5: Discussion

In the results section, it was shown that drug response in cell lines depended on germline variations, somatic mutations and their interaction. Through the demand of DNA sequencing in the last decades, a technological leap forward had been established through next generating sequencing, which made DNA sequencing affordable in clinics estimated to be $1,000 per sample (Grada & Weinbrecht, 2013). This trend enabled large-scale explorations of germline and somatic mutations in primary tumours and their impact on treatment response.

5.5.1: Ethical concerns with germline

Genome wide associations (GWAS) and germline variations in general cause ethical concerns regarding clear identification of individuals, parental information and disease risks (Kaye, Boddington, de Vries, Hawkins, & Melham, 2010). Using such data with human specimens, who are clearly identifiable, does not ensure anonymity of participants in GWAS studies. Unwilling parental information of participants might be revealed to third parties. Publically, participants might knowingly or unknowingly reveal their disease risk to friends, family, colleagues or employers, which can have social implications for their lives. The risk allele might not even be established yet, but may be associated with future studies.

Germline data is a powerful tool in clinics and risk estimation, however, at the same time it causes ethical concerns regarding incidental genomic findings (Clayton et al., 2013; Townsend et al., 2012; Wolf, Annas, & Elias, 2013). For instance, obligations of a clinician, to tell a patient risks unrelated to the currently treated disease, are not established. Furthermore, what if those unrelated risk alleles might be inherited to the patient's children, who are currently healthy and showing no symptoms but are in a high risk group?

5.5.2: Limitations of association studies

A key limitation of association studies was the number of performed tests and the consequently required multiple test correction resulting in more stringent false discovery rate (FDR) thresholds.

For somatic associations, the number of performed tests was reduced by leveraging prior knowledge, e.g. in this study the focus was on 315 well-known oncogenes and tumour suppressors. This list might be refined in the future.

I filtered with the linkage disequilibrium potential somatic mutations from the germline analysis set. This filter reduced the number of performed tests for germline association. Trust in germline associations might be further increased through linkage disequilibrium, since multiple germline SNPs in close proximity were correlated with ancestry, but also with the drug response of a particular compound. This causes a redundancy in germline signal (e.g. spikes in Manhattan plot) and ultimately increased trust in such germline associations.

The main bottleneck in the interaction analysis was the number of possible combinations between germline variations and somatic mutations, and the resulting number of tests. Therefore, I reduced the set of interaction tests to germline SNPs that had been identified as a cis-eQTL. This was done to perform a manageable size of tests and made results interpretable regarding expressed genes. This list of informative SNPs might be extended by further prior knowledge, e.g. focusing on germline mutations with functional impact.

An additional bottleneck of interaction associations was that they were rather rare events. Somatic mutations were already fairly rare events, and inspecting them within different germline backgrounds was decreasing the sample size of that particular somatic mutation even further. This reduced the statistical power of identifying interactions.

5.5.3: Faith in geldanamycin derivatives

At first, the scientific community sceptically perceived HSP90 as a cancer target, since HSP90 is a so-called 'housekeeping' gene that takes important roles in healthy cells. However, the demand of cancer to compensate for their deregulation made HSP90 an interesting drug target (see 1.2.8: Chaperones, page 15), and geldanamycin was the first compound proving this concept.

I believe that the discontinuation of the geldanamycin derivatives (particularly 17-AAG) might have been too early, and some of the toxicities might be due to solely the dependencies on NQO1 germline variations and expression. I showed that NQO1 had to be considered for identifying the responding patient cohort to 17-AAG, which ultimately would reduce the toxicity in responding patients.

The 2nd generation of HSP90 inhibitors does not rely on NQO1 expression or germline variations, which is often considered as a 'weakness' of geldanamycin derivatives. However, I showed that this weakness might be explored as a strength within certain conditions, i.e. NQO1 is functionally overexpressed in various cancer types.

Nevertheless, a full resurrection of geldanamycin derivatives (although still tested in drug combinations) would be unlikely due to various reasons. Firstly, the radicicol derivatives were independent from NQO1 metabolism (although missing a therapeutic opportunity) and further developed in the meantime. Secondly, the patent life of 17-AAG is expiring soon which makes 17-AAG an unattractive candidate for pharmaceutical companies to pursue.

5.6: Conclusion

The hypothesis that drug response was driven by somatic mutations, germline variations and their interaction was shown in various instances in this chapter.

For example, I reproduced the leading somatic associations found in Garnett at al. 2012. Concordantly, I identified the MEK1/2 inhibitors associated with BRAF mutations, the MDM2 inhibitor associated with TP53, the IGF1R inhibitor associated with *EWSR1-FLI1*, and other somatic associations. This reflected the expected relation of somatic mutations and drug response, which was driven by the causality of cancer (i.e. somatic mutations).

Additionally, I showed a systematic approach to consider germline variations as a baseline for treatment decisions, for choosing the appropriate patient cohort and ultimately minimising toxicities and identifying suitable dosages. This was exemplified with the striking association of 17-AAG and rs1800566, a germline variation that influences NQO1 activity, which was necessary to metabolise 17-AAG to the pharmacological active state. This mode-of-action is well-known, but controversially discussed in literature, particularly in the context of clinical trials. Therefore, 17-AAG was abandoned in several clinical studies and replaced by 'modern' HSP90 inhibitors, which are independent from NQO1. However, my results suggested revisiting some of those 17-AAG clinical trials. This exemplified the usage of NQO1 as a baseline for identifying toxicities.

Surprisingly, I found an interaction with the gene expression of *NQO1* and the germline variation in NQO1, which extended the application of identifying drug toxicities to another therapeutic application. For instance, this interaction could be exploited to boost the therapeutic index. I highlighted this potentially new application of 17-AAG, particularly in tissues where 17-AAG was differentially expressed for minimising overall toxicities and optimising drug response within cancer.

Based on my original hypothesis, that drug response was also based on the interaction of somatic and germline mutations, here I presented a novel systematic approach to investigating those interactions, although the presented results have to be taken with caution due to low sample sizes and low p-values. Interactions, as shown with 17-AAG, were scientifically interesting due to their nature of being different in healthy versus normal tissue, making them biomarkers of drug sensitivity.

In regards to the future of personalised medicine, I believe that a treatment decision could be based on a baseline response due to germline variation, which reflects toxicities and determines appropriate dosages of drugs. Afterwards, the drug sensitivity (so called therapeutic index) could be estimated by somatic mutations and additionally the interaction with germline variations.

Chapter 6: Discussion and future outlook

“Cancer is a word, not a sentence.”

- John Diamond

6.1: Declaration of contribution

This chapter is based exclusively on my own perspective. It summarises my thesis and outlines future development.

In previous decades the pharmaceutical industry had less interest in understanding how drugs may function, as long as they work. This mindset has changed with the demand for personalised cancer treatment, which is a profitable business. Although by subsetting the patient into smaller cohorts, this is still a lucrative business for the following reasons: Firstly, our current population ages more, which increases the risk of developing cancer (ultimately meaning more costumers) and secondly, the last 'real luxury' in our society may be prolonging life which we are willing to pay for.

Tailoring drugs to the patient's need is a desirable medical achievement, but one which will also trigger ethical questions such as: Who can afford those treatments? The National Institute for Health and Care Excellence (NICE) already made calculations concerning which treatments might be 'cost-effective', meaning they prolong the patient's life enough to be worth administering. For example, trastuzumab emtansine, a successor drug of lapatinib, significantly decreased toxicities, prolonged progression-free survival as well as overall survival in HER2-positive advanced breast cancer patients (Verma et al., 2012), but was rejected as funded treatment by the National Health Service (NICE evaluation tastuzumab estansine, 2015). Such cost-effective calculations might sound cruel, but our society needs to debate these conflicts between access to medicine and affordability. As a first response, the NHS increased the cancer treatment budget from £280 million in 2014/15 to an expected £340 million (NHS cancer fund update, 2015), but this might be a drop in the ocean.

Besides the costs and potential ethical issues, personalised treatments would be beneficial for the patient's survival and life quality, since it is tailored to the their individual need. A treatment tailored to the patient would increase the treatment success. This 'tailoring' leverages molecular profiles of patients and decides the most suitable treatment based on the molecular characterisation. It has to be noted that all recent FDA approved drugs were approved with a specific biomarker indication for a responding patient cohort. I anticipate that this trend will continue further and a next step might be biomarkers for

targeted drug combinations and even targeted cocktails tailored to each patient individually, which highlights the importance of my study.

In my thesis, I studied pharmacological cell line screens and leveraged various baseline molecular descriptors to predominantly (i) predict the drug response, (ii) identify biomarkers of sensitivity and resistance and (iii) explore potential MoA's of leading associations. These three topics were not independent and had large commonalities. For example, good predictions were based on biological associations, which could be translated to biomarkers. Those associations hinted at the underlying mechanism of the drugs and their response, which could be personalised for each patient (i.e. subsetting by biomarkers). Therefore, the prediction of drug response, the identification of biomarkers and the understanding of MoA's are very closely related topics.

All models I presented in this thesis were based on cancer cell lines. Those cancer cell lines are just 'models' of real tumours (Domcke et al., 2013). A tumour is comparable to a small novel organ growing in a patient, e.g. they supply themselves with nutrition by creating blood vessels (Folkman, 1971), interact with the immune system (Anderson & LaBaer, 2005) and consist of a heterogenic population of cells with different tasks (Greaves & Maley, 2012b), reshaping their tumour microenvironment (Hanahan & Coussens, 2012), etc. In comparison, cancer cell lines are a single cell type kept in culture under laboratory conditions. When realising the large difference between cell lines and tumours, it is obvious that any biomarker observed in cell lines needs to be validated *in vivo*. In the context of my thesis, this implies that all identified associations exclusively indicate potential leads, which require experimental follow-ups to make any reliable statement in patients.

6.1.1: Systematic analysis of biomarkers

For systematically analysing the drug responses and identifying biomarkers in cell lines, I used an ANOVA model. An ANOVA is a univariate and statistical model, meaning a single feature was tested for association with the

dependent variable. Here the dependent variable was the drug response of a compound, while the mutational status of a particular oncogene was tested across all cell lines. This statistical framework was powerful enough to derive biomarkers when the tested gene pool, as well as compound panel, was relatively small, but failed when the number of performed tests became too large. In essence, multiple-hypothesis testing was the analysis bottleneck that decreased the confidence in achieved p-values, which ultimately leads to less or even no significant results.

A future improvement to reduce the number of performed tests and decrease the sparseness of oncogenic events, might be to explore mutual exclusivities within a signalling pathway (Ciriello et al., 2012). The underlying assumption is that only one of these mutations within a pathway is necessary to cause the desired phenotype in cancer. By means of evolutionary pressure and natural selection, only one of those mutations would be abundant in a cancer cell. Through uniting different mutations within the same signalling pathway, the problem of multiple hypothesis testing could be minimised, although there would be challenges such as defining the pathways and the rare occurrences of oncogenic mutations that generally leads to being mutual exclusive in a large dataset.

6.1.2: Predictive models to understand drug response

To overcome the limitation of statistical models, I also built multivariate machine learning models, which included multiple features at the same time. Machine learning models do not require a statistical test (p-value independent method), but do require an internal cross-validation or bootstrapping to prove their predictability. Machine learning models were evaluated based on their achieved performance within the testset or alternatively, the external dataset. The advantage of machine learning models was their improved predictive power compared to a univariate model, but at the same time interpretability was decreased.

In this thesis, I explored different machine learning algorithms that were linear e.g. elastic nets, lasso regression and ridge regression, and also non-linear algorithms such as neural networks and random forests. Notably, marginal improvements might be achieved by applying different algorithms, but the largest improvement could be achieved by improving the used feature set. A common machine learning tenet is "*garbage in, garbage out*", meaning the key to improve predictive power would be to improve the feature set.

In my thesis I combined mutations, copy number variations, gene expressions and methylation profiles to predict drug responses. The molecular data was processed in various steps starting with the variant calling algorithms, filtering for passenger mutations and mutations with no functional impact, etc. All those preprocessing steps might have a larger impact on the achieved performances than other fine-tunings of the machine learning algorithms. I anticipate a boost in performance by leveraging additional features such as pathway knowledge, protein-protein interaction networks, metabolites, bioactivity profiles of the drugs and various other features.

Another outcome of my thesis was that genomic alterations were not only causal for cancer, but also predominantly drove drug responses. Those drug response predictions were often improved by including methylation profiles, although this statement must be individually reconsidered for each drug.

For most drugs, I was not able to build highly predictive models for which I confidently believe might translate to patient care. This lack of performance might be improved by tailoring the molecular input for each compound based on its MoA as outlined above, but at the same time this decreases the computational high-throughput and applicability to novel compounds. My thesis aimed for the systematic analysis of 265 drugs and their biomarkers as well as being extendable to novel untested compounds.

The current application of my prediction methods could be the systematic prioritisation of novel compounds to be added to the current drug screen. In the future with improved genomic and chemical features, I anticipate

applications such as drug repositioning, where we *in silico* identify potential drugs for a given cancer sub type, which is characterised by its molecular profile.

For those models with exceptionally good performance, I explored their MoA, since a good predictive model should be based on biological mechanisms (except for technical / biological artefacts). By inspecting those highly predictive models, I derived further hypotheses, which might contribute to our functional knowledge of good performing drug models and might indicate potential applications in clinics. For example, the interaction of NQO1's expression and NQO1's germline variant explained potential failures of early clinical trials of 17-AAG. A systematic approach to include more tailored information of each compound would be a desirable extension of my method.

6.1.3: The impact of cancer type

In my work, I observed that the tissue label was a powerful predictor for drug response. This association was expected and might be trivial, e.g. treat a melanoma with a melanoma drug. It was counterintuitive to the trend of tailoring a drug to each patient's molecular profile, e.g. "Can histology already tell us all we need to know for a cancer treatment?" The answer would be "no", but the histology was an important covariate to reveal the underlying biological association.

For example, when considering all cell lines at the same time, regardless of their tissue-of-origin, there was no association between the PDK1 inhibitor OSU-03012 and TP53 mutations, but when exploring TP53's mutations within different tissue contexts an association was observed in the urogenital system (p-value$\approx$0.006).

The stratification and exploration of associations within each cancer type reduced the sample sizes and at the same time increased the number of tests. This subsetting, therefore, decreased the statistical power and required multiple hypothesis corrections. For instance, when exploring TP53's

mutational association with OSU-03012 in 30 different cancer contexts, I needed to correct the p-values that might be inflated. The Bonferroni correction for a significance level of $\alpha = 0.05$ (false discovery rate of 5%) would be $^{0.05}/_{30} \approx 0.001$, implying that the association of OSU-03012 and TP53 within the urogenital system still significantly held true (p-value≈0.006 < 0.001), although the p-value was too optimistic.

As an alternative to the exploration within a cancer type, it was also possible to correct for a covariate such as the cancer type (pan-cancer analysis). This reduced the number of performed tests, but explored another biological rational. An association independent of the cancer type would be of high value, since it would hold true for all cancer types (pan-cancer association), but would be generally more challenging to observe in biological systems. Notably, the previously highlighted association of the OSU-03012 and TP53 mutation held true in a pan-cancer setting, but only if the tissue type was used as a covariate (p-value≈0.04), which without correction was not observed. This highlighted the fact that covariates could hide potential statistical and biological signal.

Pan-cancer associations were arguably more scientifically impactful since they would ultimately hold true for all cancer, but many associations were exclusively cancer type dependent and therefore not observable on a pan-cancer level. In my thesis, I offered systematic approaches to tackle both: (i) inspecting different tissues separately, and (ii) alternatively correcting for the tissue covariate and ultimately identifying pan-cancer associations.

6.1.4: The power of covariates in pharmacological screens

As previously described, the cancer type was an important factor of understanding drug response. Another example of covariates hiding biological signal were the germline variations. A cancer somatic mutation might be only associated with a certain drug within a germline background. This hidden association could be revealed by either inspecting the somatic mutation in subsets of germline backgrounds or correcting for the germline mutations. As

shown in the previous chapter, I identified 35 novel interactions between somatic mutations and germline variants, which were otherwise neglected by only inspecting somatic mutations on their own. I anticipate that this analysis could be expanded in cancer specific settings in the near future, when more samples per tissue type will be available.

This idea of covariates hiding biological signal in pharmacological screens could be further expanded to various other information. This information might be technical artefacts of the screen, which were previously discarded in cell line screens, since they were presumed to be non-translatable to patients. However, my thesis showed that all covariates were highly important for understanding the drug response within cell lines and might otherwise reveal hidden biological associations. For example, it would be interesting to also integrate information about the media type, doubling time, adherent vs. suspension growth, etc. A systematic integration of covariates in the analysis of pharmacological screens would be desirable and give more insight into the underlying cancer biology. This might even include features such as images of the untreated cancer cell lines, which were undeniably different to the real tumour histology, e.g. cell lines consist only of a single cell type and have no stroma.

6.1.5: Drug resistance markers

Another outcome of my thesis was that with the current drug screens, a direct observation of resistance markers would be very challenging or if not impossible. CCLE and GDSC were suitable for identifying markers of drug sensitivity, but not for drug resistance. For example, GDSC optimised their screen effort to have ~15-20% responders (sensitive cell lines) and ~80-85% non-responders (presumably resistant). The responders were defined as sensitive in comparison to the non-responders. If all cell lines would respond, this would unlikely be the cure for cancer and rather indicate cell toxicity. Too few cell lines responding would either hint at an ineffective compound or more likely indicate a small given drug concentration. My thesis highlighted this

bottleneck, which was also reflected in less reproducibility of resistance markers comparable to sensitivity markers in CCLE and GDSC.

A follow up study could be to systematically and indirectly investigate the drug resistances. By the term ‘indirectly’, I mean that a focus should be on supposedly sensitive cell lines, but which are not. For example, the $BRAF^{V600E}$ mutation caused drug sensitivity to PLX4720 in most melanoma cell lines (Chapman et al., 2011), but not in all of them. Those non-responding cell lines were expected to respond, but did not, which might be explained by another mechanism of resistance. It would be interesting to indirectly ask for drug resistance, by investigating the supposedly responding cell lines.

A direct exploration of drug resistance was doomed to fail due to the nature of the data, i.e. ‘resistant’ cell lines were non-responders for various reasons, e.g. low administered drug concentration. However, a systematic and indirect exploration of known resistant markers such as multi-drug resistance efflux pumps and channels, might refine current sensitivity biomarkers and uncover hidden associations.

6.1.6: Final words

My thesis highlighted bottlenecks and strengths of pharmacological screens, which could be further explored. It underscored the impact of tissue covariates and that mutations are not only causal of cancer, but also predictive. Furthermore, my thesis showed the meaning of pan-cancer versus cancer specific associations and offered a framework to systematically analyse pharmacological screens. I also highlighted future applications such as prioritisation of compound selection and drug repositioning. Challenges such as systematically relating predictions to the MoA and applications in clinics remained unsolved, but in this thesis I suggested approaches to tackle those problems.

List of abbreviations

Proteins	Names
4E-BP	eukaryotic translation initiation factor 4E-binding protein 1
53BP1	tumour protein p53 binding protein 1
ABCC2	ATP-binding cassette, sub-family C [CFTR/MRP], member 2
ABL	Abelson murine leukaemia viral oncogene homolog 1
AFF1	AF4 / fragile mental retardation 2 (FMR2) family member 1
AKT	v-akt murine thymoma viral oncogene (Protein kinase B) 2
ALK	anaplastic lymphoma receptor tyrosine kinase
ARAF	A-Raf proto-oncogene, serine/threonine kinase
ARID1A	AT-rich interaction domain 1A
ATM	ataxia telangiectasia mutated serine/threonine kinase
ATR	ataxia telangiectasia and RAD3 related serine/threonine kinase
Aurora B	aurora kinase B
BAD	BCL2 associated agonist of cell death
BCL2	B-cell lymphoma 2 protein
BCL6	B-cell lymphoma 6 protein
BCR	breakpoint cluster region
BCR-ABL	fusion gene of BCR and ABL
BIM	BCL2 like protein 11
BOD1	biorientation of chromosomes in cell division 1
BRAF	v-Raf murine sarcoma viral oncogene homolog B
BRCA1	breast cancer 1
BRCA2	breast cancer 2
CASP3	caspase 3
CDK	cyclin-dependent kinase
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)
CDKN2A	cyclin-dependent kinase inhibitor 2A
CDKN2B	cyclin-dependent kinase inhibitor 2B
CHK1	checkpoint kinase 1

CHK2	checkpoint kinase 2
CHMP1A	charged multivesicular body protein 1A
COQ5	coenzyme Q5, methyltransferase
CRAF	Raf-1 proto-oncogene, serine/threonine kinase
CREB	cAMP response element-binding protein
CYP40	peptidylprolyl isomerase D
DDMQH2	2-polyprenyl-6-methoxy-1,4-benzoquinol
DMQH2	2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinol
DP	dimerisation partners
DUSP5	dual specificity phosphatase 5
DUSP6	dual specificity phosphatase 6
E2F	E2F transcription factor
EGFR	epidermal growth factor receptor
EML4	echinoderm microtubule associated protein like 4
EML4-ALK	fusion gene of EML4 and ALK
ER	oestrogen receptor
ERBB2	erb-b2 receptor tyrosine kinase 2
ERK	extracellular-signal regulated kinase
ETS	external transcribed spacer (ETS) transcription factor
EWSR1	Ewing sarcoma breakpoint region 1 (RNA binding protein EWS)
EWSR1-FLI1	fusion gene of EWSR1 and FLI1
FLI1	Friend leukaemia integration 1 transcription factor
FLT3	FMS related tyrosine kinase 3
FOS	Finkel-Biskis-Jinkins (FBJ) murine osteosarcoma viral oncogene homolog
GAP	GTPase activating protein
GPCR	G protein–coupled receptor
GRB2	growth factor receptor bound protein 2
H2AX	H2A histone family member X
HER2	see ERBB2
HIP	huntingtin-interacting protein
HOP	HSP70/HSP90-organising protein

HRAS	Harvey rat sarcoma viral oncogene homolog
HSP40	heat shock 40kD protein
HSP70	heat shock 70kD protein
HSP90	heat shock 90kD protein
IGF1R	insulin-like growth factor 1 receptor
JAK2	janus kinase 2
JNK	c-Jun N-terminale kinase
JUN	jun proto-oncogene
KDR	kinase insert domain receptor
KIT	KIT proto-oncogene receptor tyrosine kinase
KRAS	Kirsten rat sarcoma viral oncogene homolog
MAP3K	mitogen-activated protein kinase kinase kinases
MCAK	mitotic centromere-associated kinesin
MCL1	myeloid cell leukaemia 1
MDC1	mediator of DNA damage checkpoint 1
MDM2	mouse double minute 2, human homolog of (TP53 binding protein)
MEK1	mitogen-activated protein kinase kinase 1
MEK2	mitogen-activated protein kinase kinase 2
MET	MET proto-oncogene, receptor tyrosine kinase
MLL-AFF1	fusion gene of MLL and AFF1
MLL3	lysine methyltransferase 2C
MNK	Menkes' protein
MRN	complex of MRE11, RAD50 and NBS1
MSK	salt inducible kinase 1
MTOR	mechanistic target of rapamycin
MTORC1	mammalian target of rapamycin complex 1
MTORC2	mammalian target of rapamycin complex 2
MYC	v-myc avian myelocytomatosis viral oncogene homolog
NQO1	NAD(P)H dehydrogenase, quinone 1
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog
NTNG1	netrin G1
NTNG2	netrin G2

p107	retinoblastoma-like protein 1
p130	retinoblastoma-like protein 2
p16	see CDKN2A
p21	see CDKN1A
p23	HSP90 co-chaperone
p27	see CDKN1B
PARP1	poly(ADP-ribose) polymerase 1
PDGFRA	platelet derived growth factor receptor alpha
PDGFRB	platelet derived growth factor receptor beta
PDK1	pyruvate dehydrogenase kinase 1
PI3K	phosphatidylinositol-4,5-bisphosphate 3-kinase
PIP_2	phosphatidylinositol 4,5-bisphosphate
PIP_3	phosphatidylinositol (3,4,5)-trisphosphate
PKA	protein kinase A
PKC	protein kinase C
PKG	protein kinase, CGMP-dependent, type I
PP1	protein phosphatase 1
PTEN	phosphatase and tensin homolog
RAC	ras-related C3 botulinum toxin substrate 1
RAD9-RAD1-HUS1	complex of RAD9, RAD1 and HUS1
RAF	raf proto-oncogene, serine/threonine kinase
RAS	rat sarcoma viral oncogene
RB	retinoblastoma protein 1
RHEB	ras homolog enriched in brain
RPA	Replication protein A
RSG14	regulator of G-protein signalling 14
RSK	ribosomal protein S6 kinase
RTK	receptor tyrosine kinases
S6K	ribosomal protein S6 kinase B1
SMIM19	small integral membrane protein 19
SOS	son of sevenless
SPATA33	spermatogenesis associated 33

SPRY	sprouty
SRC	tyrosine-protein kinase src
STAT	signal transducer and activator of transcription
SYT9	synaptotagmin 9
TERT	telomerase reverse transcriptase
TGF-β	transforming growth factor, beta 1
TOP1	topoisomerase (DNA) I
TOPB1	topoisomerase (DNA) II binding protein 1
TP53	tumour protein p53
TSC1	tuberous sclerosis 1
TSC2	tuberous sclerosis 2
UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1
WEE1	WEE1 G2 checkpoint kinase
WNT	wingless-related integration site

Tissues	**Names**
ACC	adrenocortical carcinoma
ALL	acute lymphoblastic leukemia
AML	acute myelogenous leukemia
BLCA	bladder urothelial carcinoma
BRCA	breast invasive carcinoma
CESC	cervical squamous cell carcinoma and endocervical adenocarcinoma
CML	chronic myelogenous leukemia
COREAD	colon and rectum adenocarcinoma
DLBC	lymphoid neoplasm diffuse large B-cell lymphoma
ESCA	esophageal carcinoma
GBM	glioblastoma multiforme
HNSC	head and neck squamous cell carcinoma
KIRC	kidney renal clear cell carcinoma
LAML	acute myeloid leukaemia
LGG	lower grade glioma
LIHC	liver hepatocellular carcinoma

LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
MB	medulloblastoma
MESO	mesothelioma
MM	multiple myeloma
NB	neuroblastoma
NSCLC	non-small cell lung cancer
OV	ovarian serous cystadenocarcinoma
PAAD	pancreatic adenocarcinoma
PRAD	prostate adenocarcinoma
SCLC	small cell lung carcinoma
SKCM	skin cutaneous melanoma
STAD	stomach adenocarcinoma
THCA	thyroid carcinoma
UCEC	uterine corpus endometrial carcinoma

Others	**Names**
ADP	adenosine diphosphate
ATP	adenosine triphosphate
ANOVA	analysis of variance
AUC	area under the curve
CCLE	cancer cell line encyclopaedia
COSMIC	catalogue of somatic mutations in cancer
CNV	copy number variation
DREAM	dialogue for reverse engineering assessments and methods
EGF	epidermal growth factor
eQTL	expression quantitative trait loci
FDR	false discovery rate
GEX	gene expression
GDSC	genomics of drug sensitivity in cancer
GDP	guanosine diphosphate
GTP	guanosine triphosphate

IC50	half maximal inhibitory concentration
HGF	hepatocyte growth factor
ICGC	international cancer genome consortium
LASSO	least absolute shrinkage and selection operator
MAF	minor allele frequency
MoA	mode of action
QSAR	quantitative structure-activity relationship
R_p	Pearson correlation
R_s	Spearman correlation
SNP	single nucleotide polymorphism
tRNA	Transfer RNA
TCGA	the cancer genome atlas

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., … McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. doi:10.1038/nature11632

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. doi:10.1038/nmeth0410-248

Ahrendt, S. A., Chow, J. T., Yang, S. C., Wu, L., Zhang, M.-J., Jen, J., & Sidransky, D. (2000). Alcohol Consumption and Cigarette Smoking Increase the Frequency of p53 Mutations in Non-Small Cell Lung Cancer. *Cancer Res.*, *60*(12), 3155–3159.

Al-Lazikani, B., Banerji, U., & Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, *30*(7), 679–92. doi:10.1038/nbt.2284

Ali, M. M. U., Roe, S. M., Vaughan, C. K., Meyer, P., Panaretou, B., Piper, P. W., … Pearl, L. H. (2006). Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*, *440*(7087), 1013–7. doi:10.1038/nature04716

Anastas, J. N., & Moon, R. T. (2013). WNT signalling pathways as therapeutic targets in cancer. *Nature Reviews. Cancer*, *13*(1), 11–26. doi:10.1038/nrc3419

Anderson, K. S., & LaBaer, J. (2005). The sentinel within: exploiting the immune system for cancer biomarkers. *Journal of Proteome Research*, *4*(4), 1123–33. doi:10.1021/pr0500814

Anjum, R., & Blenis, J. (2008). The RSK family of kinases: emerging roles in cellular signalling. *Nature Reviews. Molecular Cell Biology*, *9*(10), 747–58. doi:10.1038/nrm2509

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., … Easton, D. F. (2003). Average risks of breast and

ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *American Journal of Human Genetics*, *72*(5), 1117–30. doi:10.1086/375033

Arthur, J. S. C. (2008). MSK activation and physiological roles. *Frontiers in Bioscience : A Journal and Virtual Library*, *13*, 5866–79.

Asher, G., Dym, O., Tsvetkov, P., Adler, J., & Shaul, Y. (2006). The crystal structure of NAD(P)H quinone oxidoreductase 1 in complex with its potent inhibitor dicoumarol. *Biochemistry*, *45*(20), 6372–8. doi:10.1021/bi0600087

Bagatell, R., Gore, L., Egorin, M. J., Ho, R., Heller, G., Boucher, N., … Trippett, T. M. (2007). Phase I Pharmacokinetic and Pharmacodynamic Study of 17-N-Allylamino-17-Demethoxygeldanamycin in Pediatric Patients with Recurrent or Refractory Solid Tumors: A Pediatric Oncology Experimental Therapeutics Investigators Consortium Study. *Clinical Cancer Research*, *13*(6), 1783–1788. doi:10.1158/1078-0432.CCR-06-1892

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–7. doi:10.1038/nature11003

Barton, W. A., Tzvetkova-Robev, D., Miranda, E. P., Kolev, M. V, Rajashankar, K. R., Himanen, J. P., & Nikolov, D. B. (2006). Crystal structures of the Tie2 receptor ectodomain and the angiopoietin-2-Tie2 complex. *Nature Structural & Molecular Biology*, *13*(6), 524–32. doi:10.1038/nsmb1101

Bartram, C. R., de Klein, A., Hagemeijer, A., van Agthoven, T., Geurts van Kessel, A., Bootsma, D., … Stone, M. (1983). Translocation of c-ab1 oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. *Nature*, *306*(5940), 277–80.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical*

Software, *32*(6), 1–29.

Benjamin, D., Colombi, M., Moroni, C., & Hall, M. N. (2011). Rapamycin passes the torch: a new generation of mTOR inhibitors. *Nature Reviews. Drug Discovery*, *10*(11), 868–80. doi:10.1038/nrd3531

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289 – 300. doi:10.2307/2346101

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. doi:10.1093/nar/28.1.235

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Bertoli, C., Skotheim, J. M., & de Bruin, R. A. M. (2013). Control of cell cycle transcription during G1 and S phases. *Nature Reviews. Molecular Cell Biology*, *14*(8), 518–28. doi:10.1038/nrm3629

Besson, A., Dowdy, S. F., & Roberts, J. M. (2008). CDK Inhibitors: Cell Cycle Regulators and Beyond. *Developmental Cell*, *14*(2), 159–169. doi:10.1016/j.devcel.2008.01.013

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., … Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, *40*(6), 463–71. doi:10.3109/03014460.2013.807878

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.

Blanco, R., Iwakawa, R., Tang, M., Kohno, T., Angulo, B., Pio, R., … Sanchez-Cespedes, M. (2009). A gene-alteration profile of human lung cancer cell lines. *Human Mutation*, *30*(8), 1199–206. doi:10.1002/humu.21028

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*,

19(2), 185–93.

Bonet, C., Giuliano, S., Ohanna, M., Bille, K., Allegra, M., Lacour, J.-P., … Bertolotto, C. (2012). Aurora B is regulated by the mitogen-activated protein kinase/extracellular signal-regulated kinase (MAPK/ERK) signaling pathway and is a valuable potential target in melanoma cells. *The Journal of Biological Chemistry*, *287*(35), 29887–98. doi:10.1074/jbc.M112.371682

Borges, H. L., Bird, J., Wasson, K., Cardiff, R. D., Varki, N., Eckmann, L., & Wang, J. Y. J. (2005). Tumor promotion by caspase-resistant retinoblastoma protein. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15587–92. doi:10.1073/pnas.0503925102

Brandon, E. P., Idzerda, R. L., & McKnight, G. S. (1997). PKA isoforms, neural pathways, and behaviour: making the connection. *Current Opinion in Neurobiology*, *7*(3), 397–403. doi:10.1016/S0959-4388(97)80069-4

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., … Sansone, S.-A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *31*(1), 68–71.

Bromberg-White, J. L., Andersen, N. J., & Duesbery, N. S. (2012). MEK genomics in development and disease. *Briefings in Functional Genomics*, *11*(4), 300–10. doi:10.1093/bfgp/els022

Cairns, P., Polascik, T. J., Eby, Y., Tokino, K., Califano, J., Merlo, A., … Westra, W. (1995). Frequency of homozygous deletion at p16/CDKN2 in primary human tumours. *Nature Genetics*, *11*(2), 210–2. doi:10.1038/ng1095-210

Caldecott, K. W. (2008). Single-strand break repair and genetic disease. *Nature Reviews. Genetics*, *9*(8), 619–31. doi:10.1038/nrg2380

Campbell, N. A., & Reece, J. B. (2006a). *Biology*. *Pearson Studium*. Pearson Studium.

Campbell, N. A., & Reece, J. B. (2006b). *Biology*. *Pearson Studium* (6th

Editio). Pearson Studium.

Cancer Cell Line Encyclopedia Consortium; Genomics of Drug Sensitivity in Cancer Consortium. (2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, *528*(7580), 84–7. doi:10.1038/nature15736

Carrera, A. C., Alexandrov, K., & Roberts, T. M. (1993). The conserved lysine of the catalytic domain of protein kinases is actively involved in the phosphotransfer reaction and not required for anchoring ATP. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(2), 442–6.

Castellano, E., & Downward, J. (2011). RAS Interaction with PI3K: More Than Just Another Effector Pathway. *Genes & Cancer*, *2*(3), 261–74. doi:10.1177/1947601911408079

Caunt, C. J., Sale, M. J., Smith, P. D., & Cook, S. J. (2015). MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nature Reviews. Cancer*, *15*(10), 577–92. doi:10.1038/nrc4000

Chambers, J. M., Freeny, A., & Heiberger, R. M. (1992). *Analysis of variance; designed experiments. Chapter 5 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.* Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/stats/html/aov.html

Chan, G. K. Y., Kleinheinz, T. L., Peterson, D., & Moffat, J. G. (2013). A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. *PloS One*, *8*(5), e63583. doi:10.1371/journal.pone.0063583

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., … McArthur, G. A. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine*, *364*(26), 2507–16. doi:10.1056/NEJMoa1103782

Chaudhuri, L., Vincelette, N. D., Koh, B. D., Naylor, R. M., Flatten, K. S., Peterson, K. L., … Tibes, R. (2014). CHK1 and WEE1 inhibition combine synergistically to enhance therapeutic efficacy in acute myeloid leukemia ex vivo. *Haematologica*, *99*(4), 688–96.

doi:10.3324/haematol.2013.093187

Cheaib, B., Auguste, A., & Leary, A. (2015). The PI3K/Akt/mTOR pathway in ovarian cancer: therapeutic opportunities and challenges. *Chinese Journal of Cancer*, *34*(1), 4–16. doi:10.5732/cjc.014.10289

Chen, B.-J., Litvin, O., Ungar, L., & Pe'er, D. (2015). Context Sensitive Modeling of Cancer Drug Sensitivity. *PloS One*, *10*(8), e0133850. doi:10.1371/journal.pone.0133850

Chen, R., & Wold, M. S. (2014). Replication protein A: single-stranded DNA's first responder: dynamic DNA-interactions allow replication protein A to direct single-strand DNA intermediates into different pathways for synthesis or repair. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *36*(12), 1156–61. doi:10.1002/bies.201400107

Chen, X., Ko, L. J., Jayaraman, L., & Prives, C. (1996). p53 levels, functional domains, and DNA damage determine the extent of the apoptotic response of tumor cells. *Genes & Development*, *10*(19), 2438–51.

Chen, X., Peer, C. J., Alfaro, R., Tian, T., Spencer, S. D., & Figg, W. D. (2012). Quantification of irinotecan, SN38, and SN38G in human and porcine plasma by ultra high-performance liquid chromatography-tandem mass spectrometry and its application to hepatic chemoembolization. *Journal of Pharmaceutical and Biomedical Analysis*, *62*, 140–8. doi:10.1016/j.jpba.2012.01.008

Cheng, Y. R., Huang, J., Qiang, H., Lin, W. L., & Demain, A. L. (2001). Mutagenesis of the rapamycin producer Streptomyces hygroscopicus FC904. *The Journal of Antibiotics*, *54*(11), 967–72.

Choi, J., Chen, J., Schreiber, S. L., & Clardy, J. (1996). Structure of the FKBP12-rapamycin complex interacting with the binding domain of human FRAP. *Science (New York, N.Y.)*, *273*(5272), 239–42.

Chosed, R., Mukherjee, S., Lois, L. M., & Orth, K. (2006). Evolution of a signalling system that incorporates both redundancy and diversity: Arabidopsis SUMOylation. *Biochemical Journal*, *398*(3), 521. doi:10.1042/BJ20060426

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, *22*(2), 398–406. doi:10.1101/gr.125567.111

Clayton, E. W., Haga, S., Kuszler, P., Bane, E., Shutske, K., & Burke, W. (2013). Managing incidental genomic findings: legal obligations of clinicians. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *15*(8), 624–9. doi:10.1038/gim.2013.7

Cliffe, L. J., Humphreys, N. E., Lane, T. E., Potten, C. S., Booth, C., & Grencis, R. K. (2005). Accelerated intestinal epithelial cell turnover: a new mechanism of parasite expulsion. *Science (New York, N.Y.)*, *308*(5727), 1463–5. doi:10.1126/science.1108661

Coate, L., Cuffe, S., Horgan, A., Hung, R. J., Christiani, D., & Liu, G. (2010). Germline genetic variation, cancer outcome, and pharmacogenetics. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *28*(26), 4029–37. doi:10.1200/JCO.2009.27.2336

Collins, K., Jacks, T., & Pavletich, N. P. (1997). The cell cycle and cancer. *Proceedings of the National Academy of Sciences*, *94*(7), 2776–2778. doi:10.1073/pnas.94.7.2776

Conway, J. R. W., Carragher, N. O., & Timpson, P. (2014). Developments in preclinical cancer imaging: innovating the discovery of therapeutics. *Nature Reviews. Cancer*, *14*(5), 314–28. doi:10.1038/nrc3724

Corcoran, R. B., Ebi, H., Turke, A. B., Coffee, E. M., Nishino, M., Cogdill, A. P., … Engelman, J. A. (2012). EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. *Cancer Discovery*, *2*(3), 227–35. doi:10.1158/2159-8290.CD-11-0341

Cornell, V. H., & Blauw, A. S. (1949). Histopathologic observations in cases of Hodgkin's disease treated with nitrogen mustard. *The American Journal of Pathology*, *25*(2), 233–7.

Corrie, P. G. (2008). Cytotoxic chemotherapy: clinical aspects. *Medicine*, *36*(1), 24–28. doi:10.1016/j.mpmed.2007.10.012

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, *32*(12), 1202–1212. doi:10.1038/nbt.2877

Courtois-Cox, S., Jones, S. L., & Cichowski, K. (2008). Many roads lead to oncogene-induced senescence. *Oncogene*, *27*(20), 2801–9. doi:10.1038/sj.onc.1210950

Cox, A. D., Fesik, S. W., Kimmelman, A. C., Luo, J., & Der, C. J. (2014). Drugging the undruggable RAS: Mission Possible? *Nature Reviews. Drug Discovery*, *13*(11), 828–851. doi:10.1038/nrd4389

Cree, I. A., Glaysher, S., & Harvey, A. L. (2010). Efficacy of anti-cancer agents in cell lines versus human primary tumour tissue. *Current Opinion in Pharmacology*, *10*(4), 375–9. doi:10.1016/j.coph.2010.05.001

Dai, Y.-N., Zhou, K., Cao, D.-D., Jiang, Y.-L., Meng, F., Chi, C.-B., … Zhou, C.-Z. (2014). Crystal structures and catalytic mechanism of the C-methyltransferase Coq5 provide insights into a key step of the yeast coenzyme Q synthesis pathway. *Acta Crystallographica. Section D, Biological Crystallography*, *70*(Pt 8), 2085–92. doi:10.1107/S1399004714011559

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., … Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, *417*(6892), 949–54. doi:10.1038/nature00766

De Smet, F., Christopoulos, A., & Carmeliet, P. (2014). Allosteric targeting of receptor tyrosine kinases. *Nature Biotechnology*, *32*(11), 1113–20. doi:10.1038/nbt.3028

Devillers, J. (2013). Methods for building QSARs. *Methods in Molecular Biology (Clifton, N.J.)*, *930*, 3–27. doi:10.1007/978-1-62703-059-5_1

Dhanasekaran, N., & Premkumar Reddy, E. (1998). Signaling by dual specificity kinases. *Oncogene*, *17*(11 Reviews), 1447–55. doi:10.1038/sj.onc.1202251

Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling

pathways in cancer. *Oncogene*, *26*(22), 3279–90. doi:10.1038/sj.onc.1210421

Diaz, L. A., Williams, R. T., Wu, J., Kinde, I., Hecht, J. R., Berlin, J., … Vogelstein, B. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, *486*(7404), 537–40. doi:10.1038/nature11219

Domcke, S., Sinha, R., Levine, D. A., Sander, C., & Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*, *4*, 2126. doi:10.1038/ncomms3126

Donaldson, M. S. (2004). Nutrition and cancer: a review of the evidence for an anti-cancer diet. *Nutrition Journal*, *3*(1), 19. doi:10.1186/1475-2891-3-19

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., & Zheng, X. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*, *15*, 489. doi:10.1186/s12885-015-1492-6

Dowdy, S. F., Hinds, P. W., Louie, K., Reed, S. I., Arnold, A., & Weinberg, R. A. (1993). Physical interaction of the retinoblastoma protein with human D cyclins. *Cell*, *73*(3), 499–511.

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. *Nature Reviews. Cancer*, *3*(1), 11–22. doi:10.1038/nrc969

Downward, J. (2008). Targeting RAS and PI3K in lung cancer. *Nature Medicine*, *14*(12), 1315–6. doi:10.1038/nm1208-1315

Drucker, E., & Krapfenbauer, K. (2013). Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. *The EPMA Journal*, *4*(1), 7. doi:10.1186/1878-5085-4-7

Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., … Larson, R. A. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *The New England Journal of Medicine*, *355*(23), 2408–17. doi:10.1056/NEJMoa062867

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying

methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1), 587. doi:10.1186/1471-2105-11-587

Dunna, N. R., Vuree, S., Anuradha, C., Sailaja, K., Surekha, D., Digumarti, R. R., … Vishnupriya, S. (2014). NRAS mutations in de novo acute leukemia: prevalence and clinical significance. *Indian Journal of Biochemistry & Biophysics*, *51*(3), 207–10.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–10.

Efeyan, A., & Serrano, M. (2007). p53: guardian of the genome and policeman of the oncogenes. *Cell Cycle (Georgetown, Tex.)*, *6*(9), 1006–10.

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics : TIG*, *24*(7), 344–52. doi:10.1016/j.tig.2008.04.005

Engelman, J. A. (2009). Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nature Reviews. Cancer*, *9*(8), 550–62. doi:10.1038/nrc2664

Engelman, J. A., Chen, L., Tan, X., Crosby, K., Guimaraes, A. R., Upadhyay, R., … Wong, K.-K. (2008). Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. *Nature Medicine*, *14*(12), 1351–6. doi:10.1038/nm.1890

Evers, R., Kool, M., van Deemter, L., Janssen, H., Calafat, J., Oomen, L. C., … Borst, P. (1998). Drug export activity of the human canalicular multispecific organic anion transporter in polarized kidney MDCK cells expressing cMOAT (MRP2) cDNA. *The Journal of Clinical Investigation*, *101*(7), 1310–9. doi:10.1172/JCI119886

Fan, C.-Y., Lee, S., & Cyr, D. M. (2003). Mechanisms for regulation of Hsp70 function by Hsp40. *Cell Stress & Chaperones*, *8*(4), 309–16.

Fang, Y., Qin, Y., Zhang, N., Wang, J., Wang, H., & Zheng, X. (2015). DISIS: prediction of drug response through an iterative sure independence

screening. *PloS One*, *10*(3), e0120408. doi:10.1371/journal.pone.0120408

FDA approved drugs. (2015). Genomics - Table of Pharmacogenomic Biomarkers in Drug Labeling. Center for Drug Evaluation and Research. Retrieved from http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm

Feder, M. E., & Hofmann, G. E. (1999). Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual Review of Physiology*, *61*, 243–82. doi:10.1146/annurev.physiol.61.1.243

Fidler, C., Watkins, F., Bowen, D. T., Littlewood, T. J., Wainscoat, J. S., & Boultwood, J. (2004). NRAS, FLT3 and TP53 mutations in patients with myelodysplastic syndrome and a del(5q). *Haematologica*, *89*(7), 865–6.

Fletcher, J. I., Haber, M., Henderson, M. J., & Norris, M. D. (2010). ABC transporters in cancer: more than just drug efflux pumps. *Nature Reviews. Cancer*, *10*(2), 147–56. doi:10.1038/nrc2789

Folkman, J. (1971). Tumor angiogenesis: therapeutic implications. *The New England Journal of Medicine*, *285*(21), 1182–6. doi:10.1056/NEJM197111182852108

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., … Campbell, P. J. (2014). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, gku1075–. doi:10.1093/nar/gku1075

Forcet, C., Stein, E., Pays, L., Corset, V., Llambi, F., Tessier-Lavigne, M., & Mehlen, P. (2002). Netrin-1-mediated axon outgrowth requires deleted in colorectal cancer-dependent MAPK activation. *Nature*, *417*(6887), 443–7. doi:10.1038/nature748

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Fritsche-Guenther, R., Witzel, F., Sieber, A., Herr, R., Schmidt, N., Braun, S., … Blüthgen, N. (2011). Strong negative feedback from Erk to Raf confers robustness to MAPK signalling. *Molecular Systems Biology*, *7*, 489. doi:10.1038/msb.2011.27

Fujita, N., Sato, S., & Tsuruo, T. (2003). Phosphorylation of p27Kip1 at threonine 198 by p90 ribosomal protein S6 kinases promotes its binding to 14-3-3 and cytoplasmic localization. *The Journal of Biological Chemistry*, *278*(49), 49254–60. doi:10.1074/jbc.M306614200

Fusi, N., Lippert, C., Lawrence, N. D., & Stegle, O. (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications*, *5*, 4890. doi:10.1038/ncomms5890

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, *4*(3), 177–83. doi:10.1038/nrc1299

Füzéry, A. K., Levin, J., Chan, M. M., & Chan, D. W. (2013). Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. *Clinical Proteomics*, *10*(1), 13. doi:10.1186/1559-0275-10-13

Gagné, J.-F., Montminy, V., Belanger, P., Journault, K., Gaucher, G., & Guillemette, C. (2002). Common human UGT1A polymorphisms and the altered metabolism of irinotecan active metabolite 7-ethyl-10-hydroxycamptothecin (SN-38). *Molecular Pharmacology*, *62*(3), 608–17.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., … Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, *483*(7391), 570–5. doi:10.1038/nature11005

Garnock-Jones, K. P., Keating, G. M., & Scott, L. J. (2010). Trastuzumab: A review of its use as adjuvant treatment in human epidermal growth factor receptor 2 (HER2)-positive early breast cancer. *Drugs*, *70*(2), 215–39. doi:10.2165/11203700-000000000-00000

Gartner, E. M., Silverman, P., Simon, M., Flaherty, L., Abrams, J., Ivy, P., & Lorusso, P. M. (2012). A phase II study of 17-allylamino-17-demethoxygeldanamycin in metastatic or locally advanced, unresectable

breast cancer. *Breast Cancer Research and Treatment*, *131*(3), 933–7. doi:10.1007/s10549-011-1866-7

Garuti, L., Roberti, M., & Bottegoni, G. (2010). Non-ATP competitive protein kinase inhibitors. *Current Medicinal Chemistry*, *17*(25), 2804–21.

Gaspar, N., Sharp, S. Y., Pacey, S., Jones, C., Walton, M., Vassal, G., … Workman, P. (2009). Acquired resistance to 17-allylamino-17-demethoxygeldanamycin (17-AAG, tanespimycin) in glioblastoma cells. *Cancer Research*, *69*(5), 1966–75. doi:10.1158/0008-5472.CAN-08-3131

Gaubatz, S., Lindeman, G. J., Ishida, S., Jakoi, L., Nevins, J. R., Livingston, D. M., & Rempel, R. E. (2000). E2F4 and E2F5 play an essential role in pocket protein-mediated G1 control. *Molecular Cell*, *6*(3), 729–35.

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, *20*(3), 307–15. doi:10.1093/bioinformatics/btg405

Gazdar, A. F. (2009). Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors. *Oncogene*, *28 Suppl 1*, S24–31. doi:10.1038/onc.2009.198

Geeleher, P., Cox, N. J., & Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biology*, *15*(3), R47. doi:10.1186/gb-2014-15-3-r47

Gey, G., Coffman, W., & Kubicek, M. (1952). Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Res*, *12*, 264–265.

Ghoshal, S. (2012). The Emperor of All Maladies: A Biography of Cancer. *Journal of Postgraduate Medicine Education and Research*. doi:10.5005/jp-journals-10028-1025

Gibbons, D. L., Byers, L. A., & Kurie, J. M. (2014). Smoking, p53 mutation, and lung cancer. *Molecular Cancer Research : MCR*, *12*(1), 3–13. doi:10.1158/1541-7786.MCR-13-0539

Gilbert, S. F. (2000). Cross-Talk between Pathways. Sinauer Associates.

Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK10087/

Glaze, E. R., Lambert, A. L., Smith, A. C., Page, J. G., Johnson, W. D., McCormick, D. L., … Tomaszewski, J. E. (2005). Preclinical toxicity of a geldanamycin analog, 17-(dimethylaminoethylamino)-17-demethoxygeldanamycin (17-DMAG), in rats and dogs: potential clinical relevance. *Cancer Chemotherapy and Pharmacology*, *56*(6), 637–47. doi:10.1007/s00280-005-1000-9

Goetz, M. P., Toft, D., Reid, J., Ames, M., Stensgard, B., Safgren, S., … Erlichman, C. (2005). Phase I trial of 17-allylamino-17-demethoxygeldanamycin in patients with advanced cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *23*(6), 1078–87. doi:10.1200/JCO.2005.09.119

Golding, G. B. (1984). The sampling distribution of linkage disequilibrium. *Genetics*, *108*(1), 257–74.

Gönen, M., & Margolin, A. A. (2014). Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics (Oxford, England)*, *30*(17), i556–63. doi:10.1093/bioinformatics/btu464

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., … Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, *10*(11), 1081–2. doi:10.1038/nmeth.2642

Goodrich, D. W. (2006). The retinoblastoma tumor-suppressor gene, the exception that proves the rule. *Oncogene*, *25*(38), 5233–43. doi:10.1038/sj.onc.1209616

Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *The Journal of Investigative Dermatology*, *133*(8), e11. doi:10.1038/jid.2013.248

Gray-Schopfer, V. C., Cheong, S. C., Chong, H., Chow, J., Moss, T., Abdel-Malek, Z. A., … Bennett, D. C. (2006). Cellular senescence in naevi and immortalisation in melanoma: a role for p16? *British Journal of Cancer*, *95*(4), 496–505. doi:10.1038/sj.bjc.6603283

Greaves, M., & Maley, C. C. (2012a). Clonal evolution in cancer. *Nature*, *481*(7381), 306–13. doi:10.1038/nature10762

Greaves, M., & Maley, C. C. (2012b). Clonal evolution in cancer. *Nature*, *481*(7381), 306–13. doi:10.1038/nature10762

Gully, C. P., Velazquez-Torres, G., Shin, J.-H., Fuentes-Mattei, E., Wang, E., Carlock, C., … Lee, M.-H. (2012). Aurora B kinase phosphorylates and instigates degradation of p53. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24), E1513–22. doi:10.1073/pnas.1110287109

Gundem, G., Perez-Llamas, C., Jene-Sanz, A., Kedzierska, A., Islam, A., Deu-Pons, J., … Lopez-Bigas, N. (2010). IntOGen: integration and data mining of multidimensional oncogenomic data. *Nature Methods*, *7*(2), 92–3. doi:10.1038/nmeth0210-92

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. W. L., & Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, *504*(7480), 389–93. doi:10.1038/nature12831

Hanafusa, H., Torii, S., Yasunaga, T., Matsumoto, K., & Nishida, E. (2004). Shp2, an SH2-containing protein-tyrosine phosphatase, positively regulates receptor tyrosine kinase signaling by dephosphorylating and inactivating the inhibitor Sprouty. *The Journal of Biological Chemistry*, *279*(22), 22992–5. doi:10.1074/jbc.M312498200

Hanahan, D., & Coussens, L. M. (2012). Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell*, *21*(3), 309–22. doi:10.1016/j.ccr.2012.02.022

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–74. doi:10.1016/j.cell.2011.02.013

Hanawa, M., Suzuki, S., Dobashi, Y., Yamane, T., Kono, K., Enomoto, N., & Ooi, A. (2006). EGFR protein overexpression and gene amplification in squamous cell carcinomas of the esophagus. *International Journal of Cancer. Journal International Du Cancer*, *118*(5), 1173–80. doi:10.1002/ijc.21454

Hantschel, O., & Superti-Furga, G. (2004). Regulation of the c-Abl and Bcr-Abl tyrosine kinases. *Nature Reviews. Molecular Cell Biology*, *5*(1), 33–44. doi:10.1038/nrm1280

Harbour, J. W. (2000). The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes & Development*, *14*(19), 2393–2409. doi:10.1101/gad.813200

Hartl, F. U., Bracher, A., & Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. *Nature*, *475*(7356), 324–332. doi:10.1038/nature10317

Hatzis, C., Bedard, P. L., Birkbak, N. J., Beck, A. H., Aerts, H. J. W. L., Stem, D. F., … Haibe-Kains, B. (2014). Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Research*, *74*(15), 4016–23. doi:10.1158/0008-5472.CAN-14-0725

Heath, E. I., Hillman, D. W., Vaishampayan, U., Sheng, S., Sarkar, F., Harper, F., … Liu, G. (2008). A phase II trial of 17-allylamino-17-demethoxygeldanamycin in patients with hormone-refractory metastatic prostate cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *14*(23), 7940–6. doi:10.1158/1078-0432.CCR-08-0221

Heaton, J. (2008). *Programming Neural Networks with Encog3 in Java* (2nd Editio). Heaton Research. Retrieved from http://www.heatonresearch.com/book/programming-neural-networks-encog3-java.html

Heaton, J. (2011). *Introduction to Neural Networks for Java* (2nd Editio). Heaton Research. Retrieved from http://www.heatonresearch.com/book/programming-neural-networks-java-2.html

Heidorn, S. J., Milagre, C., Whittaker, S., Nourry, A., Niculescu-Duvas, I., Dhomen, N., … Marais, R. (2010). Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF. *Cell*, *140*(2), 209–21. doi:10.1016/j.cell.2009.12.040

Henley, S. A., & Dick, F. A. (2012). The retinoblastoma family of proteins and

their regulatory functions in the mammalian cell division cycle. *Cell Division*, *7*(1), 10. doi:10.1186/1747-1028-7-10

Hert, J., Irwin, J. J., Laggner, C., Keiser, M. J., & Shoichet, B. K. (2009). Quantifying biogenic bias in screening libraries. *Nature Chemical Biology*, *5*(7), 479–83. doi:10.1038/nchembio.180

Hilfiker-Kleiner, D., Hilfiker, A., Castellazzi, M., Wollert, K. C., Trautwein, C., Schunkert, H., & Drexler, H. (2006). JunD attenuates phenylephrine-mediated cardiomyocyte hypertrophy by negatively regulating AP-1 transcriptional activity. *Cardiovascular Research*, *71*(1), 108–17. doi:10.1016/j.cardiores.2006.02.032

Himanen, J.-P., & Nikolov, D. B. (2003). Eph signaling: a structural view. *Trends in Neurosciences*, *26*(1), 46–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12495863

Höhfeld, J., Minami, Y., & Hartl, F. U. (1995). Hip, a novel cochaperone involved in the eukaryotic Hsc70/Hsp40 reaction cycle. *Cell*, *83*(4), 589–98.

Hubbard, S. R. (2004). Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nature Reviews Molecular Cell Biology*, *5*(6), 464–471. doi:10.1038/nrm1399

Huber, K. V. M., Salah, E., Radic, B., Gridling, M., Elkins, J. M., Stukalov, A., … Superti-Furga, G. (2014). Stereospecific targeting of MTH1 by (S)-crizotinib as an anticancer strategy. *Nature*, *508*(7495), 222–227. doi:10.1038/nature13194

Hui Zou, T. H. (2005). Regularization and variable selection via the Elastic Net. *Royal Statistical Society, Series B*.

Human Free Vector. (2015). Retrieved from http://www.freevectors.net/details/Free+Vector+Human+Silhouette+

Ikenouchi, J., & Umeda, M. (2010). FRMD4A regulates epithelial polarity by connecting Arf6 activation with the PAR complex. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(2), 748–53. doi:10.1073/pnas.0908423107

Ing, V. W. (1984). The etiology and management of leukopenia. *Canadian Family Physician Médecin de Famille Canadien*, *30*, 1835–9.

Irizarry, R. A., Gautier, L., Huber, W., & Ben, B. (2006). makecdfenv: CDF Environment Maker.

Ishii, N., Harada, N., Joseph, E. W., Ohara, K., Miura, T., Sakamoto, H., … Sakai, T. (2013). Enhanced inhibition of ERK signaling by a novel allosteric MEK inhibitor, CH5126766, that suppresses feedback reactivation of RAF activity. *Cancer Research*, *73*(13), 4050–60. doi:10.1158/0008-5472.CAN-12-3937

Iverson, C., Larson, G., Lai, C., Yeh, L.-T., Dadson, C., Weingarten, P., … Quart, B. (2009). RDEA119/BAY 869766: a potent, selective, allosteric inhibitor of MEK1/2 for the treatment of cancer. *Cancer Research*, *69*(17), 6839–47. doi:10.1158/0008-5472.CAN-09-0679

Jackson, S. P. (2009). The DNA-damage response: new molecular insights and new approaches to cancer therapy. *Biochemical Society Transactions*, *37*(Pt 3), 483–94. doi:10.1042/BST0370483

Janakiraman, M., Vakiani, E., Zeng, Z., Pratilas, C. A., Taylor, B. S., Chitale, D., … Solit, D. B. (2010). Genomic and Biological Characterization of Exon 4 KRAS Mutations in Human Cancer. *Cancer Research*, *70*(14), 5901–5911. doi:10.1158/0008-5472.CAN-10-0192

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. A. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 63–74.

Jarosz, D. (2016). *Hsp90 in Cancer: Beyond the Usual Suspects*. *Advances in cancer research* (Vol. 129). Elsevier. doi:10.1016/bs.acr.2015.11.001

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Alternatives to Laboratory Animals : ATLA*, *33*(5), 445–59.

John, P. C., Mews, M., & Moore, R. (2001). Cyclin/Cdk complexes: their

involvement in cell cycle progression and mitotic division. *Protoplasma*, *216*(3-4), 119–42.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–23. doi:10.1534/genetics.107.080101

Kaye, J., Boddington, P., de Vries, J., Hawkins, N., & Melham, K. (2010). Ethical implications of the use of whole genome methods in medical research. *European Journal of Human Genetics : EJHG*, *18*(4), 398–403. doi:10.1038/ejhg.2009.191

Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., … Roth, B. L. (2009). Predicting new molecular targets for known drugs. *Nature*, *462*(7270), 175–81. doi:10.1038/nature08506

Kelland, L. R., Sharp, S. Y., Rogers, P. M., Myers, T. G., & Workman, P. (1999). DT-Diaphorase expression and tumor cell sensitivity to 17-allylamino, 17-demethoxygeldanamycin, an inhibitor of heat shock protein 90. *Journal of the National Cancer Institute*, *91*(22), 1940–9.

Kelloff, G. J., & Sigman, C. C. (2012). Cancer biomarkers: selecting the right drug for the right patient. *Nature Reviews. Drug Discovery*, *11*(3), 201–14. doi:10.1038/nrd3651

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, a. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi:10.1101/gr.229102

Khan, S. A., Virtanen, S., Kallioniemi, O. P., Wennerberg, K., Poso, A., & Kaski, S. (2014). Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics (Oxford, England)*, *30*(17), i497–504. doi:10.1093/bioinformatics/btu456

Kim, Y. S., Alarcon, S. V, Lee, S., Lee, M.-J., Giaccone, G., Neckers, L., & Trepel, J. B. (2009). Update on Hsp90 inhibitors in clinical trial. *Current Topics in Medicinal Chemistry*, *9*(15), 1479–92.

Klinge, C. M. (2001). Estrogen receptor interaction with estrogen response elements. *Nucleic Acids Research*, *29*(14), 2905–19.

Knoblauch, R., & Garabedian, M. J. (1999). Role for Hsp90-associated cochaperone p23 in estrogen receptor signal transduction. *Molecular and Cellular Biology*, *19*(5), 3748–59.

Knudson, A. G. (2002). Cancer genetics. *American Journal of Medical Genetics*, *111*(1), 96–102. doi:10.1002/ajmg.10320

Kobayashi, H., Takemura, Y., & Ohnuma, T. (1992). Relationship between tumor cell density and drug concentration and the cytotoxic effects of doxorubicin or vincristine: mechanism of inoculum effects. *Cancer Chemotherapy and Pharmacology*, *31*(1), 6–10.

Konecny, G. E., Pegram, M. D., Venkatesan, N., Finn, R., Yang, G., Rahmeh, M., … Slamon, D. J. (2006). Activity of the dual kinase inhibitor lapatinib (GW572016) against HER-2-overexpressing and trastuzumab-treated breast cancer cells. *Cancer Research*, *66*(3), 1630–9. doi:10.1158/0008-5472.CAN-05-1182

Konecny, G. E., Winterhoff, B., Kolarova, T., Qi, J., Manivong, K., Dering, J., … Slamon, D. J. (2011). Expression of p16 and retinoblastoma determines response to CDK4/6 inhibition in ovarian cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *17*(6), 1591–602. doi:10.1158/1078-0432.CCR-10-2307

Kracikova, M., Akiri, G., George, A., Sachidanandam, R., & Aaronson, S. A. (2013). A threshold mechanism mediates p53 cell fate decision between growth arrest and apoptosis. *Cell Death and Differentiation*, *20*(4), 576–88. doi:10.1038/cdd.2012.155

Kranenburg, O. (2005). The KRAS oncogene: past, present, and future. *Biochimica et Biophysica Acta*, *1756*(2), 81–2. doi:10.1016/j.bbcan.2005.10.001

Kumar, R., Chaudhary, K., Singla, D., Gautam, A., & Raghava, G. P. S. (2014). Designing of promiscuous inhibitors against pancreatic cancer cell lines. *Scientific Reports*, *4*, 4668. doi:10.1038/srep04668

Kwak, E. L., Bang, Y.-J., Camidge, D. R., Shaw, A. T., Solomon, B., Maki, R. G., … Iafrate, A. J. (2010). Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *The New England Journal of Medicine*, *363*(18), 1693–703. doi:10.1056/NEJMoa1006448

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. doi:10.3389/fpsyg.2013.00863

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, *313*(5795), 1929–35. doi:10.1126/science.1132939

Landry, J. J. M., Pyl, P. T., Rausch, T., Zichner, T., Tekkedil, M. M., Stütz, A. M., … Steinmetz, L. M. (2013). The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda, Md.)*, *3*(8), 1213–24. doi:10.1534/g3.113.005777

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., … Flicek, P. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, *47*(7), 692–695. doi:10.1038/ng.3312

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., Ur-Rehman, S., … Flicek, P. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics*, *47*(7), 692–5. doi:10.1038/ng.3312

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature Reviews. Molecular Cell Biology*, *8*(12), 995–1005. doi:10.1038/nrm2281

Lee, E., & Muller, W. (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harbor Perspectives in Biology*, *2*(10), a003236. doi:10.1101/cshperspect.a003236

Lee, J.-H., & Paull, T. T. (2004). Direct activation of the ATM protein kinase by the Mre11/Rad50/Nbs1 complex. *Science (New York, N.Y.)*, *304*(5667), 93–6. doi:10.1126/science.1091496

Lee, J.-H., & Paull, T. T. (2005). ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science (New York, N.Y.)*, *308*(5721), 551–4. doi:10.1126/science.1108297

Lemmon, M. A., Schlessinger, J., Amit, I., Citri, A., Shay, T., Lu, Y., … Yarden, Y. (2010). Cell Signaling by Receptor Tyrosine Kinases. *Cell*, *141*(7), 1117–1134. doi:10.1016/j.cell.2010.06.011

Lin, C. Y., Lovén, J., Rahl, P. B., Paranal, R. M., Burge, C. B., Bradner, J. E., … Young, R. A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*, *151*(1), 56–67. doi:10.1016/j.cell.2012.08.026

Lindberg, R. A., Quinn, A. M., & Hunter, T. (1992). Dual-specificity protein kinases: will any hydroxyl do? *Trends in Biochemical Sciences*, *17*(3), 114–119. doi:10.1016/0968-0004(92)90248-8

Lindquist, S. (1986). The Heat-Shock Response. *Annual Review of Biochemistry*, *55*(1), 1151–1191. doi:10.1146/annurev.bi.55.070186.005443

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *46*(1-3), 3–26.

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014). *LIMIX: genetic analysis of multiple traits*. *bioRxiv*. Cold Spring Harbor Labs Journals. doi:10.1101/003905

Little, A. S., Balmanno, K., Sale, M. J., Newman, S., Dry, J. R., Hampson, M., … Cook, S. J. (2011). Amplification of the driving oncogene, KRAS or BRAF, underpins acquired resistance to MEK1/2 inhibitors in colorectal cancer cells. *Science Signaling*, *4*(166), ra17. doi:10.1126/scisignal.2001752

Liu, P., Cheng, H., Roberts, T. M., & Zhao, J. J. (2009). Targeting the phosphoinositide 3-kinase pathway in cancer. *Nature Reviews. Drug Discovery*, *8*(8), 627–44. doi:10.1038/nrd2926

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell,

J. (2000a). DNA Damage and Repair and Their Role in Carcinogenesis. In *Molecular Cell Biology* (4th ed.). W. H. Freeman. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21554/

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000b). Microtubule Dynamics and Motor Protein. In *Molecular Cell Biology* (4th ed.). W. H. Freeman. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21537/

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000c). Mutations Causing Loss of Cell-Cycle Control. W. H. Freeman. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21526/

Long, G. V, Stroyakovskiy, D., Gogas, H., Levchenko, E., de Braud, F., Larkin, J., … Flaherty, K. (2014). Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *The New England Journal of Medicine*, *371*(20), 1877–88. doi:10.1056/NEJMoa1406037

Lugo, T. G., Pendergast, A. M., Muller, A. J., & Witte, O. N. (1990). Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science (New York, N.Y.)*, *247*(4946), 1079–82.

Ma, C. X., Janetka, J. W., & Piwnica-Worms, H. (2011). Death by releasing the breaks: CHK1 inhibitors as cancer therapeutics. *Trends in Molecular Medicine*, *17*(2), 88–96. doi:10.1016/j.molmed.2010.10.009

Magi, A., Tattini, L., Cifola, I., D'Aurizio, R., Benelli, M., Mangano, E., … Gensini, G. F. (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biology*, *14*(10), R120. doi:10.1186/gb-2013-14-10-r120

Majumder, B., Baraneedharan, U., Thiyagarajan, S., Radhakrishnan, P., Narasimhan, H., Dhandapani, M., … Majumder, P. K. (2015). Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. *Nature Communications*, *6*, 6169. doi:10.1038/ncomms7169

Malumbres, M., & Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nature Reviews. Cancer*, *9*(3), 153–66. doi:10.1038/nrc2602

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science (New York, N.Y.)*, *298*(5600), 1912–34. doi:10.1126/science.1075762

Maqsood, M. I., Matin, M. M., Bahrami, A. R., & Ghasroldasht, M. M. (2013). Immortality of cell lines: challenges and advantages of establishment. *Cell Biology International*, *37*(10), 1038–45. doi:10.1002/cbin.10137

Massagué, J. (2008). TGFbeta in Cancer. *Cell*, *134*(2), 215–30. doi:10.1016/j.cell.2008.07.001

Matano, M., Date, S., Shimokawa, M., Takano, A., Fujii, M., Ohta, Y., … Sato, T. (2015). Modeling colorectal cancer using CRISPR-Cas9–mediated engineering of human intestinal organoids. *Nature Medicine*, *21*(3), 256–262. doi:10.1038/nm.3802

Matsukawa, H., Akiyoshi-Nishimura, S., Zhang, Q., Luján, R., Yamaguchi, K., Goto, H., … Itohara, S. (2014). Netrin-G/NGL complexes encode functional synaptic diversification. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *34*(47), 15779–92. doi:10.1523/JNEUROSCI.1141-14.2014

Meacham, C. E., & Morrison, S. J. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, *501*(7467), 328–37. doi:10.1038/nature12624

Medkova, M., & Cho, W. (1999). Interplay of C1 and C2 Domains of Protein Kinase C-alpha in Its Membrane Binding and Activation. *Journal of Biological Chemistry*, *274*(28), 19852–19861. doi:10.1074/jbc.274.28.19852

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PloS One*, *8*(4), e61318. doi:10.1371/journal.pone.0061318

Mendoza, M. C., Er, E. E., & Blenis, J. (2011). The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends in Biochemical Sciences*, *36*(6), 320–8. doi:10.1016/j.tibs.2011.03.006

Michaloglou, C., Vredeveld, L. C. W., Soengas, M. S., Denoyelle, C., Kuilman,

T., van der Horst, C. M. A. M., … Peeper, D. S. (2005). BRAFE600-associated senescence-like cell cycle arrest of human naevi. *Nature*, *436*(7051), 720–724. doi:10.1038/nature03890

Minchinton, A. I., & Tannock, I. F. (2006). Drug penetration in solid tumours. *Nature Reviews. Cancer*, *6*(8), 583–92. doi:10.1038/nrc1893

Misale, S., Yaeger, R., Hobor, S., Scala, E., Janakiraman, M., Liska, D., … Bardelli, A. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, *486*(7404), 532–6. doi:10.1038/nature11156

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math. Retrieved from http://personal.disco.unimib.it/Vanneschi/McGrawHill_-_Machine_Learning_-Tom_Mitchell.pdf

Modjtahedi, H., Cho, B. C., Michel, M. C., & Solca, F. (2014). A comprehensive review of the preclinical efficacy profile of the ErbB family blocker afatinib in cancer. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *387*(6), 505–521. doi:10.1007/s00210-014-0967-3

Monod, J., Wyman, J., & Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, *12*(1), 88–118. doi:10.1016/S0022-2836(65)80285-6

Moreland, J., Gramada, A., Buzko, O., Zhang, Q., & Bourne, P. (2005). The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*, *6*(1), 21. doi:10.1186/1471-2105-6-21

Moreland, J. L., Gramada, A., Buzko, O. V, Zhang, Q., & Bourne, P. E. (2005). The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*, *6*, 21. doi:10.1186/1471-2105-6-21

Morgan, D. O. (2007). The Cell Cycle: Principles of Control (Primers in Biology) (Primers in Biology): David O. Morgan: 9780878935086: Amazon.com: Books.

Moyer, V. A. (2014). Risk assessment, genetic counseling, and genetic testing for BRCA-related cancer in women: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, *160*(4), 271–81. doi:10.7326/M13-2747

Murphree, A. L., & Benedict, W. F. (1984). Retinoblastoma: clues to human oncogenesis. *Science (New York, N.Y.)*, *223*(4640), 1028–33.

Negritto, C. (2010). Repairing Double-Strand DNA Breaks. *Nature Education*, *3*(9), 23.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., … Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, *10*(6), 515–27. doi:10.1016/j.ccr.2006.10.008

Nguyen, T. D., Markova, S., Liu, W., Gow, J. M., Baldwin, R. M., Habashian, M., … Kroetz, D. L. (2013). Functional characterization of ABCC2 promoter polymorphisms and allele-specific expression. *The Pharmacogenomics Journal*, *13*(5), 396–402. doi:10.1038/tpj.2012.20

NHS cancer fund update. (2015). NHS England » NHS increases budget for cancer drugs fund from £280 million in 2014/15 to an expected £340 million in 2015/16. Retrieved from https://www.england.nhs.uk/2015/01/cancer-drug-budget/

NICE evaluation tastuzumab estansine. (2015). Breast cancer (HER2 positive, unresectable) - trastuzumab emtansine (after trastuzumab & taxane) [ID603] | Guidance and guidelines | NICE. Retrieved from http://www.nice.org.uk/guidance/indevelopment/gid-tag350

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., … Levens, D. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, *151*(1), 68–79. doi:10.1016/j.cell.2012.08.033

Nigg, E. A. (2002). Centrosome aberrations: cause or consequence of cancer progression? *Nature Reviews. Cancer*, *2*(11), 815–25. doi:10.1038/nrc924

Norbury, C. J., & Zhivotovsky, B. (2004). DNA damage-induced apoptosis. *Oncogene*, *23*(16), 2797–808. doi:10.1038/sj.onc.1207532

Nowell, C. P., & Hungerford, D. (1960). A minute chromosome in human chronic granulocytic leukemia. *Science*, *132*, 1497.

Nussinov, R., & Tsai, C.-J. (2012). The different ways through which specificity works in orthosteric and allosteric drugs. *Current Pharmaceutical Design*, *18*(9), 1311–6.

Odunuga, O. O., Longshaw, V. M., & Blatch, G. L. (2004). Hop: more than an Hsp70/Hsp90 adaptor protein. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *26*(10), 1058–68. doi:10.1002/bies.20107

Ohren, J. F., Chen, H., Pavlovsky, A., Whitehead, C., Zhang, E., Kuffa, P., … Hasemann, C. A. (2004). Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nature Structural & Molecular Biology*, *11*(12), 1192–7. doi:10.1038/nsmb859

Oldham, W. M., & Hamm, H. E. (2008). Heterotrimeric G protein activation by G-protein-coupled receptors. *Nature Reviews Molecular Cell Biology*, *9*(1), 60–71. doi:10.1038/nrm2299

Oliner, J. D., Kinzler, K. W., Meltzer, P. S., George, D. L., & Vogelstein, B. (1992). Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature*, *358*(6381), 80–3. doi:10.1038/358080a0

Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A., & Shokat, K. M. (2013). K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, *503*(7477), 548–51. doi:10.1038/nature12796

Ou, S.-H. I., Bartlett, C. H., Mino-Kenudson, M., Cui, J., & Iafrate, A. J. (2012). Crizotinib for the treatment of ALK-rearranged non-small cell lung cancer: a success story to usher in the second decade of molecular targeted therapy in oncology. *The Oncologist*, *17*(11), 1351–75. doi:10.1634/theoncologist.2012-0311

Owen, L. A., Kowalewski, A. A., & Lessnick, S. L. (2008). EWS/FLI mediates

transcriptional repression via NKX2.2 during oncogenic transformation in Ewing's sarcoma. *PloS One*, *3*(4), e1965. doi:10.1371/journal.pone.0001965

Pal, T., Permuth-Wey, J., Betts, J. A., Krischer, J. P., Fiorica, J., Arango, H., … Sutphen, R. (2005). BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer*, *104*(12), 2807–16. doi:10.1002/cncr.21536

Palmieri, L., & Rastelli, G. (2013). αC helix displacement as a general approach for allosteric modulation of protein kinases. *Drug Discovery Today*, *18*(7-8), 407–14. doi:10.1016/j.drudis.2012.11.009

Parikh, C., Subrahmanyam, R., & Ren, R. (2007). Oncogenic NRAS, KRAS, and HRAS exhibit different leukemogenic potentials in mice. *Cancer Research*, *67*(15), 7139–46. doi:10.1158/0008-5472.CAN-07-0778

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., … Brazma, A. (2007). ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, *35*(Database issue), D747–50. doi:10.1093/nar/gkl995

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews. Drug Discovery*, *9*(3), 203–14. doi:10.1038/nrd3078

Peeters, M., Douillard, J.-Y., Van Cutsem, E., Siena, S., Zhang, K., Williams, R., & Wiezorek, J. (2013). Mutant KRAS codon 12 and 13 alleles in patients with metastatic colorectal cancer: assessment as prognostic and predictive biomarkers of response to panitumumab. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *31*(6), 759–65. doi:10.1200/JCO.2012.45.1492

Pérez-Mancera, P. A., Young, A. R. J., & Narita, M. (2014). Inside and out: the activities of senescence in cancer. *Nature Reviews. Cancer*, *14*(8), 547–58. doi:10.1038/nrc3773

Pérez-Soler, R., Chachoua, A., Hammond, L. A., Rowinsky, E. K., Huberman,

M., Karp, D., … Bonomi, P. (2004). Determinants of tumor response and survival with erlotinib in patients with non--small-cell lung cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *22*(16), 3238–47. doi:10.1200/JCO.2004.11.057

Pillay, C. S., Hofmeyr, J.-H., Mashamaite, L. N., & Rohwer, J. M. (2013). From top-down to bottom-up: computational modeling approaches for cellular redoxin networks. *Antioxidants & Redox Signaling*, *18*(16), 2075–86. doi:10.1089/ars.2012.4771

Plyte, S., & Musacchio, A. (2007). PLK1 inhibitors: setting the mitotic death trap. *Current Biology : CB*, *17*(8), R280–3. doi:10.1016/j.cub.2007.02.018

Polo, S. E., & Jackson, S. P. (2011). Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes & Development*, *25*(5), 409–33. doi:10.1101/gad.2021311

Porter, I. M., McClelland, S. E., Khoudoli, G. A., Hunter, C. J., Andersen, J. S., McAinsh, A. D., … Swedlow, J. R. (2007). Bod1, a novel kinetochore protein required for chromosome biorientation. *The Journal of Cell Biology*, *179*(2), 187–97. doi:10.1083/jcb.200704098

Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., … Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, *483*(7387), 100–3. doi:10.1038/nature10868

Pratt, W. B., & Toft, D. O. (2003). Regulation of signaling protein function and trafficking by the hsp90/hsp70-based chaperone machinery. *Experimental Biology and Medicine (Maywood, N.J.)*, *228*(2), 111–33.

Pray, L. (2008). DNA replication and causes of mutation. *Nature Education*, *1*(1), 214.

Prieur, A., Tirode, F., Cohen, P., & Delattre, O. (2004). EWS/FLI-1 silencing and gene profiling of Ewing cells reveal downstream oncogenic pathways and a crucial role for repression of insulin-like growth factor binding protein 3. *Molecular and Cellular Biology*, *24*(16), 7275–83. doi:10.1128/MCB.24.16.7275-7283.2004

Prior, I. A., Lewis, P. D., & Mattos, C. (2012). A comprehensive survey of Ras mutations in cancer. *Cancer Research*, *72*(10), 2457–67. doi:10.1158/0008-5472.CAN-11-2612

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., & Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics (Oxford, England)*, *26*(18), i645–52. doi:10.1093/bioinformatics/btq394

Quintás-Cardama, A., & Verstovsek, S. (2013). Molecular pathways: Jak/STAT pathway: mutations, inhibitors, and resistance. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *19*(8), 1933–40. doi:10.1158/1078-0432.CCR-12-0284

Raju, T. N. (1999). The Nobel chronicles. 1971: Earl Wilbur Sutherland, Jr. (1915-74). *Lancet (London, England)*, *354*(9182), 961. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10489991

Ratajczak, T., & Carrello, A. (1996). Cyclophilin 40 (CyP-40), mapping of its hsp90 binding domain and evidence that FKBP52 competes with CyP-40 for hsp90 binding. *The Journal of Biological Chemistry*, *271*(6), 2961–5.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., … Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia (New York, N.Y.)*, *6*(1), 1–6.

Riedmiller, M., & Braun, H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *IEEE INTERNATIONAL CONFERENCE ON NEURAL NETWORKS*.

Ritossa, F. (1962). A new puffing pattern induced by temperature shock and DNP in drosophila. *Experientia*, *18*(12), 571–573. doi:10.1007/BF02172188

Rogers, D. J., & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science (New York, N.Y.)*, *132*(3434), 1115–8. doi:10.1126/science.132.3434.1115

Romagosa, C., Simonetti, S., López-Vicente, L., Mazo, A., Lleonart, M. E.,

Castellvi, J., & Ramon y Cajal, S. (2011). p16Ink4a overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. *Oncogene*, *30*(18), 2087–2097. doi:10.1038/onc.2010.614

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., … Brown, P. O. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, *24*(3), 227–35. doi:10.1038/73432

Rowley, J. D. (1973). Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Annales de Génétique*, *16*(2), 109–12.

Rubio-Perez, C., Tamborero, D., Schroeder, M. P., Antolín, A. A., Deu-Pons, J., Perez-Llamas, C., … Lopez-Bigas, N. (2015). In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, *27*(3), 382–396. doi:10.1016/j.ccell.2015.02.007

Rudel, D., & Sommer, R. J. (2003). The evolution of developmental mechanisms. *Developmental Biology*, *264*(1), 15–37. doi:10.1016/S0012-1606(03)00353-1

Rutherford, S. L., & Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, *396*(6709), 336–342. doi:10.1038/24550

Saladi, R. N., & Persaud, A. N. (2005). The causes of skin cancer: a comprehensive review. *Drugs of Today (Barcelona, Spain : 1998)*, *41*(1), 37–53. doi:10.1358/dot.2005.41.1.875777

Salomon, J. A., Wang, H., Freeman, M. K., Vos, T., Flaxman, A. D., Lopez, A. D., & Murray, C. J. L. (2012). Healthy life expectancy for 187 countries, 1990-2010: a systematic analysis for the Global Burden Disease Study 2010. *Lancet*, *380*(9859), 2144–62. doi:10.1016/S0140-6736(12)61690-0

Sancar, A., Lindsey-Boltz, L. A., Unsal-Kaçmaz, K., & Linn, S. (2004). Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual Review of Biochemistry*, *73*, 39–85. doi:10.1146/annurev.biochem.73.011303.073723

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews. Genetics*, *12*(10), 683–91. doi:10.1038/nrg3051

Schlessinger, J. (1994). SH2/SH3 signaling proteins. *Current Opinion in Genetics & Development*, *4*(1), 25–30.

Schulte, T. W., Akinaga, S., Soga, S., Sullivan, W., Stensgard, B., Toft, D., & Neckers, L. M. (1998). Antibiotic radicicol binds to the N-terminal domain of Hsp90 and shares important biologic activities with geldanamycin. *Cell Stress & Chaperones*, *3*(2), 100–8.

Seiradake, E., Coles, C. H., Perestenko, P. V, Harlos, K., McIlhinney, R. A. J., Aricescu, A. R., & Jones, E. Y. (2011). Structural basis for cell surface patterning through NetrinG-NGL interactions. *The EMBO Journal*, *30*(21), 4479–88. doi:10.1038/emboj.2011.346

Seto, B. (2012). Rapamycin and mTOR: a serendipitous discovery and implications for breast cancer. *Clinical and Translational Medicine*, *1*(1), 29. doi:10.1186/2001-1326-1-29

Shahzad, K., & Loor, J. J. (2012). Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism. *Current Genomics*, *13*(5), 379–94. doi:10.2174/138920212801619269

Sharma, S. V, & Settleman, J. (2007). Oncogene addiction: setting the stage for molecularly targeted cancer therapy. *Genes & Development*, *21*(24), 3214–31. doi:10.1101/gad.1609907

Sharrocks, A. D., Yang, S. H., & Galanis, A. (2000). Docking domains and substrate-specificity determination for MAP kinases. *Trends in Biochemical Sciences*, *25*(9), 448–53.

Shaulian, E., & Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nature Cell Biology*, *4*(5), E131–E136. doi:10.1038/ncb0502-e131

Sherr, C. J., & Roberts, J. M. (1999). CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes & Development*, *13*(12), 1501–12.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.

M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–11.

Shiloh, Y. (2006). The ATM-mediated DNA-damage response: taking shape. *Trends in Biochemical Sciences*, *31*(7), 402–10. doi:10.1016/j.tibs.2006.05.004

Shin, S.-Y., Rath, O., Choo, S.-M., Fee, F., McFerran, B., Kolch, W., & Cho, K.-H. (2009). Positive- and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway. *Journal of Cell Science*, *122*(Pt 3), 425–35. doi:10.1242/jcs.036319

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews. Cancer*, *6*(10), 813–23. doi:10.1038/nrc1951

Siegel, D., Franklin, W. A., & Ross, D. (1998). Immunohistochemical detection of NAD(P)H:quinone oxidoreductase in human lung and lung tumors. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *4*(9), 2065–70.

Siegel, D., & Ross, D. (2000). Immunodetection of NAD(P)H:quinone oxidoreductase 1 (NQO1) in human tissues11This work is dedicated to the memory of Professor Lars Ernster, who provided us with enthusiastic support, scientific insight, and constant encouragement in our many interactions. *Free Radical Biology and Medicine*, *29*(3-4), 246–253. doi:10.1016/S0891-5849(00)00310-5

Smith, J., Tho, L. M., Xu, N., & Gillespie, D. A. (2010). The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Advances in Cancer Research*, *108*, 73–112. doi:10.1016/B978-0-12-380888-2.00003-0

Smyth, T., Paraiso, K. H. T., Hearn, K., Rodriguez-Lopez, A. M., Munck, J. M., Haarberg, H. E., … Wallis, N. G. (2014). Inhibition of HSP90 by AT13387 delays the emergence of resistance to BRAF inhibitors and overcomes resistance to dual BRAF and MEK inhibition in melanoma models. *Molecular Cancer Therapeutics*, *13*(12), 2793–804. doi:10.1158/1535-7163.MCT-14-0452

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S.,

… Mano, H. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*, *448*(7153), 561–6. doi:10.1038/nature05945

Soga, S., Akinaga, S., & Shiotsu, Y. (2013). Hsp90 inhibitors as anti-cancer agents, from basic discoveries to clinical development. *Current Pharmaceutical Design*, *19*(3), 366–76.

Soga, S., Neckers, L. M., Schulte, T. W., Shiotsu, Y., Akasaka, K., Narumi, H., … Akinaga, S. (1999). KF25706, a novel oxime derivative of radicicol, exhibits in vivo antitumor activity via selective depletion of Hsp90 binding signaling molecules. *Cancer Research*, *59*(12), 2931–8.

Solit, D. B., Ivy, S. P., Kopil, C., Sikorski, R., Morris, M. J., Slovin, S. F., … Scher, H. I. (2007). Phase I trial of 17-allylamino-17-demethoxygeldanamycin in patients with advanced cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *13*(6), 1775–82. doi:10.1158/1078-0432.CCR-06-1863

Solit, D. B., Osman, I., Polsky, D., Panageas, K. S., Daud, A., Goydos, J. S., … Chapman, P. B. (2008). Phase II Trial of 17-Allylamino-17-Demethoxygeldanamycin in Patients with Metastatic Melanoma. *Clinical Cancer Research*, *14*(24), 8302–8307. doi:10.1158/1078-0432.CCR-08-1002

Sørensen, C. S., & Syljuåsen, R. G. (2012). Safeguarding genome integrity: the checkpoint kinases ATR, CHK1 and WEE1 restrain CDK activity during normal DNA replication. *Nucleic Acids Research*, *40*(2), 477–86. doi:10.1093/nar/gkr697

Sos, M. L., Michel, K., Zander, T., Weiss, J., Frommolt, P., Peifer, M., … Thomas, R. K. (2009). Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions. *The Journal of Clinical Investigation*, *119*(6), 1727–40. doi:10.1172/JCI37127

Spalding, K. L., Bhardwaj, R. D., Buchholz, B. A., Druid, H., & Frisén, J. (2005). Retrospective birth dating of cells in humans. *Cell*, *122*(1), 133–43. doi:10.1016/j.cell.2005.04.028

Stauffer, D. R., Howard, T. L., Nyun, T., & Hollenberg, S. M. (2001). CHMP1

is a novel nuclear matrix protein affecting chromatin structure and cell-cycle progression. *Journal of Cell Science*, *114*(Pt 13), 2383–93.

Steinberg, S. F. (2008). Structural basis of protein kinase C isoform function. *Physiological Reviews*, *88*(4), 1341–78. doi:10.1152/physrev.00034.2007

Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, *6*, 7432. doi:10.1038/ncomms8432

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., … Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature*, *486*(7403), 400–4. doi:10.1038/nature11017

Stirewalt, D. L., Kopecky, K. J., Meshinchi, S., Appelbaum, F. R., Slovak, M. L., Willman, C. L., & Radich, J. P. (2001). FLT3, RAS, and TP53 mutations in elderly patients with acute myeloid leukemia. *Blood*, *97*(11), 3589–95.

Strano, S., Dell'Orso, S., Di Agostino, S., Fontemaggi, G., Sacchi, A., & Blandino, G. (2007). Mutant p53: an oncogenic transcription factor. *Oncogene*, *26*(15), 2212–9. doi:10.1038/sj.onc.1210296

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–24. doi:10.1038/nature07943

Sturm, O. E., Orton, R., Grindlay, J., Birtwistle, M., Vyshemirsky, V., Gilbert, D., … Kolch, W. (2010). The mammalian MAPK/ERK pathway exhibits properties of a negative feedback amplifier. *Science Signaling*, *3*(153), ra90. doi:10.1126/scisignal.2001212

Su, Y., Dostmann, W. R., Herberg, F. W., Durick, K., Xuong, N. H., Ten Eyck, L., … Varughese, K. I. (1995). Regulatory subunit of protein kinase A: structure of deletion mutant with cAMP binding domains. *Science (New York, N.Y.)*, *269*(5225), 807–13.

Sulli, G., Di Micco, R., & d'Adda di Fagagna, F. (2012). Crosstalk between chromatin state and DNA damage response in cellular senescence and cancer. *Nature Reviews. Cancer*, *12*(10), 709–20. doi:10.1038/nrc3344

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, *156*(6), 1324–35. doi:10.1016/j.cell.2014.01.051

Swamidass, S. J., & Baldi, P. (2007). Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. *Journal of Chemical Information and Modeling*, *47*(2), 302–17. doi:10.1021/ci600358f

Taipale, M., Jarosz, D. F., & Lindquist, S. (2010). HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular Cell Biology*, *11*(7), 515–528. doi:10.1038/nrm2918

Taylor, S. S., Ilouz, R., Zhang, P., & Kornev, A. P. (2012). Assembly of allosteric macromolecular switches: lessons from PKA. *Nature Reviews. Molecular Cell Biology*, *13*(10), 646–58. doi:10.1038/nrm3432

Thorpe, L. M., Yuzugullu, H., & Zhao, J. J. (2014). PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nature Reviews Cancer*, *15*(1), 7–24. doi:10.1038/nrc3860

Toledo, F., & Wahl, G. M. (2006). Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews. Cancer*, *6*(12), 909–23. doi:10.1038/nrc2012

Torti, D., & Trusolino, L. (2011). Oncogene addiction as a foundational rationale for targeted anti-cancer therapy: promises and perils. *EMBO Molecular Medicine*, *3*(11), 623–36. doi:10.1002/emmm.201100176

Townsend, A., Adam, S., Birch, P. H., Lohn, Z., Rousseau, F., & Friedman, J. M. (2012). “I want to know what’s in Pandora's Box”: comparing stakeholder perspectives on incidental findings in clinical whole genomic sequencing. *American Journal of Medical Genetics. Part A*, *158A*(10), 2519–25. doi:10.1002/ajmg.a.35554

Uehara, Y. (2003). Natural product origins of Hsp90 inhibitors. *Current Cancer Drug Targets*, *3*(5), 325–30.

van Dyk, E., Reinders, M. J. T., & Wessels, L. F. A. (2013). A scale-space method for detecting recurrent DNA copy number changes with analytical

false discovery rate control. *Nucleic Acids Research*, *41*(9), e100. doi:10.1093/nar/gkt155

van Staveren, W. C. G., Solís, D. Y. W., Hébrant, A., Detours, V., Dumont, J. E., & Maenhaut, C. (2009). Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells? *Biochimica et Biophysica Acta*, *1795*(2), 92–103. doi:10.1016/j.bbcan.2008.12.004

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., … Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, *469*(7331), 539–42. doi:10.1038/nature09639

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., … Liu, E. A. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science (New York, N.Y.)*, *303*(5659), 844–8. doi:10.1126/science.1092472

Verma, S., Miles, D., Gianni, L., Krop, I. E., Welslau, M., Baselga, J., … Blackwell, K. (2012). Trastuzumab emtansine for HER2-positive advanced breast cancer. *The New England Journal of Medicine*, *367*(19), 1783–91. doi:10.1056/NEJMoa1209124

Vermeulen, K., Van Bockstaele, D. R., & Berneman, Z. N. (2003). The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Proliferation*, *36*(3), 131–49.

Vietri, M., Bianchi, M., Ludlow, J. W., Mittnacht, S., & Villa-Moruzzi, E. (2006). Direct interaction between the catalytic subunit of Protein Phosphatase 1 and pRb. *Cancer Cell International*, *6*, 3. doi:10.1186/1475-2867-6-3

Vogelstein, B., & Kinzler, K. W. (1993). The multistep nature of cancer. *Trends in Genetics : TIG*, *9*(4), 138–41.

von Kriegsheim, A., Baiocchi, D., Birtwistle, M., Sumpton, D., Bienvenut, W., Morrice, N., … Kolch, W. (2009). Cell fate decisions are specified by the dynamic ERK interactome. *Nature Cell Biology*, *11*(12), 1458–64. doi:10.1038/ncb1994

Wallace, M. ., Adams, M. ., Kanouni, T., Mol, C. ., Dougan, D. ., Feher, V. ., …

Dong, Q. (2010). Structure-based design and synthesis of pyrrole derivatives as MEK inhibitors. *Bioorg.Med.Chem.Lett.*, *20*, 4156–4158.

Walsh, T., Casadei, S., Coats, K. H., Swisher, E., Stray, S. M., Higgins, J., … King, M.-C. (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA*, *295*(12), 1379–88. doi:10.1001/jama.295.12.1379

Wan, P. T. C., Garnett, M. J., Roe, S. M., Lee, S., Niculescu-Duvaz, D., Good, V. M., … Marais, R. (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, *116*(6), 855–67.

Wander, S. A., Levis, M. J., & Fathi, A. T. (2014). The evolving role of FLT3 inhibitors in acute myeloid leukemia: quizartinib and beyond. *Therapeutic Advances in Hematology*, *5*(3), 65–77. doi:10.1177/2040620714532123

Wang, W., Cheng, B., Miao, L., Mei, Y., & Wu, M. (2013). Mutant p53-R273H gains new function in sustained activation of EGFR signaling via suppressing miR-27a expression. *Cell Death & Disease*, *4*, e574. doi:10.1038/cddis.2013.97

Wang, Y., Schmid-Bindert, G., & Zhou, C. (2012). Erlotinib in the treatment of advanced non-small cell lung cancer: an update for clinicians. *Therapeutic Advances in Medical Oncology*, *4*(1), 19–29. doi:10.1177/1758834011427927

Wegele, H., Müller, L., & Buchner, J. (2004). Hsp70 and Hsp90--a relay team for protein folding. *Reviews of Physiology, Biochemistry and Pharmacology*, *151*, 1–44. doi:10.1007/s10254-003-0021-1

Weigel, B. J., Blaney, S. M., Reid, J. M., Safgren, S. L., Bagatell, R., Kersey, J., … Adamson, P. C. (2007). A phase I study of 17-allylaminogeldanamycin in relapsed/refractory pediatric patients with solid tumors: a Children's Oncology Group study. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *13*(6), 1789–93. doi:10.1158/1078-0432.CCR-06-2270

Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell*, *81*(3), 323–30.

Weinstein, I. B., & Joe, A. (2008). Oncogene addiction. *Cancer Research*, *68*(9), 3077–80; discussion 3080. doi:10.1158/0008-5472.CAN-07-3293

Weinstein, J. N. (2012). Drug discovery: Cell lines battle cancer. *Nature*, *483*(7391), 544–5. doi:10.1038/483544a

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews. Genetics*, *7*(10), 771–80. doi:10.1038/nrg1960

Whitesell, L., & Lindquist, S. L. (2005). HSP90 and the chaperoning of cancer. *Nature Reviews. Cancer*, *5*(10), 761–72. doi:10.1038/nrc1716

Whitesell, L., Mimnaugh, E. G., De Costa, B., Myers, C. E., & Neckers, L. M. (1994). Inhibition of heat shock protein HSP90-pp60v-src heteroprotein complex formation by benzoquinone ansamycins: essential role for stress proteins in oncogenic transformation. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(18), 8324–8.

Whitesell, L., Santagata, S., Mendillo, M. L., Lin, N. U., Proia, D. A., & Lindquist, S. (2014). HSP90 empowers evolution of resistance to hormonal therapy in human breast cancer models. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(51), 18297–302. doi:10.1073/pnas.1421323111

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, *9*(1), pp. 60–62.

Witt, P. L., & McGrain, P. (1985). Comparing two sample means t tests. *Physical Therapy*, *65*(11), 1730–3.

Wolf, S. M., Annas, G. J., & Elias, S. (2013). Point-counterpoint. Patient autonomy and incidental findings in clinical genomics. *Science (New York, N.Y.)*, *340*(6136), 1049–50. doi:10.1126/science.1239119

Wong, T. N., Ramsingh, G., Young, A. L., Miller, C. A., Touma, W., Welch, J. S., … Wilson, R. K. (2014). Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature*, *518*(7540), 552–5. doi:10.1038/nature13968

Workman, P., Clarke, P. A., & Al-Lazikani, B. (2016). Blocking the survival of the nastiest by HSP90 inhibition. *Oncotarget*, *7*(4), 3658–61. doi:10.18632/oncotarget.6971

Wu, J. N., & Roberts, C. W. M. (2013). ARID1A mutations in cancer: another epigenetic tumor suppressor? *Cancer Discovery*, *3*(1), 35–43. doi:10.1158/2159-8290.CD-12-0361

Xiao, Z., Chen, Z., Gunasekera, A. H., Sowin, T. J., Rosenberg, S. H., Fesik, S., & Zhang, H. (2003). Chk1 mediates S and G2 arrests through Cdc25A degradation in response to DNA-damaging agents. *The Journal of Biological Chemistry*, *278*(24), 21767–73. doi:10.1074/jbc.M300229200

Yaguchi, K., Nishimura-Akiyoshi, S., Kuroki, S., Onodera, T., & Itohara, S. (2014). Identification of transcriptional regulatory elements for Ntng1 and Ntng2 genes in mice. *Molecular Brain*, *7*, 19. doi:10.1186/1756-6606-7-19

Yang, H., Rudge, D. G., Koos, J. D., Vaidialingam, B., Yang, H. J., & Pavletich, N. P. (2013). mTOR kinase structure, mechanism and regulation. *Nature*, *497*(7448), 217–23. doi:10.1038/nature12122

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., … Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, *41*(Database issue), D955–61. doi:10.1093/nar/gks1111

Yap, C. W. (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, *32*(7), 1466–74. doi:10.1002/jcc.21707

Ye, B. H., Chaganti, S., Chang, C. C., Niu, H., Corradini, P., Chaganti, R. S., & Dalla-Favera, R. (1995). Chromosomal translocations cause deregulated BCL6 expression by promoter substitution in B cell lymphoma. *The EMBO Journal*, *14*(24), 6209–17.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–71. doi:10.1093/bioinformatics/btp394

Yeh, J. J., Routh, E. D., Rubinas, T., Peacock, J., Martin, T. D., Shen, X. J., … Der, C. J. (2009). KRAS/BRAF mutation status and ERK1/2 activation as biomarkers for MEK1/2 inhibitor therapy in colorectal cancer. *Molecular Cancer Therapeutics*, *8*(4), 834–43. doi:10.1158/1535-7163.MCT-08-0972

Yen, L.-C., Uen, Y.-H., Wu, D.-C., Lu, C.-Y., Yu, F.-J., Wu, I.-C., … Wang, J.-Y. (2010). Activating KRAS mutations and overexpression of epidermal growth factor receptor as independent predictors in metastatic colorectal cancer patients treated with cetuximab. *Annals of Surgery*, *251*(2), 254–60. doi:10.1097/SLA.0b013e3181bc9d96

Young, J. C., & Hartl, F. U. (2000). Polypeptide release by Hsp90 involves ATP hydrolysis and is enhanced by the co-chaperone p23. *The EMBO Journal*, *19*(21), 5930–40. doi:10.1093/emboj/19.21.5930

Yu, M., Strohmeyer, N., Wang, J., Müller, D. J., & Helenius, J. (2015). Increasing throughput of AFM-based single cell adhesion measurements through multisubstrate surfaces. *Beilstein Journal of Nanotechnology*, *6*(1), 157–166. doi:10.3762/bjnano.6.15

Zahreddine, H., & Borden, K. L. B. (2013). Mechanisms and insights into drug resistance in cancer. *Frontiers in Pharmacology*, *4*, 28. doi:10.3389/fphar.2013.00028

Zhang, D., & Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895–907. doi:10.1016/j.neuroimage.2011.09.069

Zhang, Z., Kobayashi, S., Borczuk, A. C., Leidner, R. S., Laframboise, T., Levine, A. D., & Halmos, B. (2010). Dual specificity phosphatase 6 (DUSP6) is an ETS-regulated negative feedback mediator of oncogenic ERK signaling in lung cancer cells. *Carcinogenesis*, *31*(4), 577–86. doi:10.1093/carcin/bgq020

Zheng, C. F., & Guan, K. L. (1994). Activation of MEK family kinases requires phosphorylation of two conserved Ser/Thr residues. *The EMBO Journal*, *13*(5), 1123–31.

Zhou, J., & Giannakakou, P. (2005). Targeting microtubules for cancer chemotherapy. *Current Medicinal Chemistry. Anti-Cancer Agents*, *5*(1), 65–71.

Zhou, Y., Tozzi, F., Chen, J., Fan, F., Xia, L., Wang, J., … Weihua, Z. (2012). Intracellular ATP levels are a pivotal determinant of chemoresistance in colon cancer cells. *Cancer Research*, *72*(1), 304–14. doi:10.1158/0008-5472.CAN-11-1674

Zimmermann, G., Papke, B., Ismail, S., Vartak, N., Chandra, A., Hoffmann, M., … Waldmann, H. (2013). Small molecule inhibition of the KRAS-PDEδ interaction impairs oncogenic KRAS signalling. *Nature*, *497*(7451), 638–42. doi:10.1038/nature12205