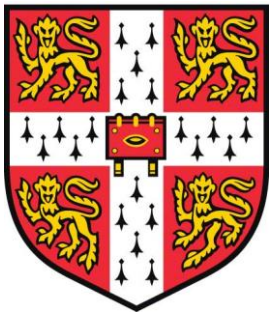
Identification and functional characterisation of gene fusions in human cancer cell lines



Elisabeth Daisy Chen

Wellcome Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Christ's College

September 2018

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

This dissertation contains fewer than 60,000 words, exclusive of tables, footnotes, bibliography and appendices.

Acknowledgements

This dissertation would not have been possible for the support of many people and institutions over the past four years. Thank you to the Wellcome Sanger Institute and the Medical Research Council for providing me the funds and a space to carry out this work. I would like to express my gratitude particularly to my supervisor, Mathew Garnett, who taught me by example not only how to be an impactful thinker and scientist, but also a thoughtful and patient mentor to a student.

Thank you also to all my colleagues in Team 215 for providing a scientifically engaging, stimulating and immensely enjoyable environment to work in the past few years. Special thanks to Gabriele Picco for being the ideal co-first author, and for complementing my computational interest with his strengths in experimental biology. I would also like to specifically thank (in alphabetical order): Fiona Behan, Thomas Cokelaer, Hayley Francies, Emanuel Gonçalves, and Francesco Iorio, for playing an active role in teaching me the hard and soft skills that I needed to complete my work.

There are not enough pages to express my appreciations of my friends who shared with me their time and laughter. All the summer BBQ's, breakfasts and lunches, celebrations, climbing sessions, extracurriculars, game nights, grad halls, morning runs, pub evenings, tea sessions, week(end) trips, work clubs, etc. etc., filled my every-day with happiness and special memories. Without naming names, I feel lucky to have all of them in my life.

Finally, I would like to lovingly acknowledge my family: my mother 陈燕青, my father 钱发强, my brother David, as well as Jake, my family away from my family.

Abstract

Advances in next-generation sequencing have accelerated the rate at which novel gene fusions are discovered. The discovery of gene fusions such as *EML4-ALK* in lung cancer and *BCR-ABL1* in chronic myeloid leukaemia have already led to changes in clinical care. However, important questions remain about the role of gene fusions in promoting oncogenic phenotypes and their relevance in drug response.

In this study, I combine RNA sequencing, CRISPR/Cas9 screens and high-throughput drug sensitivity data in a panel of 1,011 human cancer cell lines across 42 tissue types to examine the occurrence and functional relevance of gene fusions in cancer.

Fusions were called using three algorithms and filtered to reveal 8,354 fusion events with a validation rate of 70%. Cell lines exhibit known fusions in their corresponding tissue types as well as a large number of putative passenger events and fusion recurrence across tissue types correlates with that found in patient samples.

The panel of 1,011 cell lines has previously undergone high-throughput drug screening of 409 drug compounds. I implemented a systematic analysis to identify associations between fusion occurrence and drug response. It reliably recapitulates known associations (e.g. *BCR-ABL1* and sensitivity to ABL inhibitors). However, the number of novel findings is low, likely due to the low numbers of recurrent fusions, a lack of prior knowledge of novel gene fusions as well as a narrow range of drug targets.

Next, I developed a computational approach using whole-genome CRISPR/Cas9 screening data for 339 cell lines. It utilises CRISPR/Cas9 data on a guide-level to systematically evaluate essentiality of novel gene fusions. My analysis predicts essentiality of known gene fusions with high accuracy and provides evidence for the oncogenic relevance of novel gene fusions. A gene fusions in *YAP1-MAML2* represents a particularly interesting finding showing functionality across multiple distinct cancer types.

Altogether, in my thesis, I demonstrate that innovative computational approaches leveraging new datasets can enable us to elucidate the functionality of rare gene fusions in human cancer. These types of discoveries may aid in the development of targeted therapies and supports the use of clinical basket trials to capture cancer events across multiple tissue types.

Table of Contents

1 Introduction & Literature review

1.1 General introduction and thesis outline

1.1.1 Context and role of genomics in cancer treatment

The advent of next-generation sequencing technologies has allowed the scientific community an unprecedented insight into cancer genomes. For a disease that is estimated to cause 28% of deaths in the United Kingdom each year (Cancer Research UK, 2015), insights into the biology of cancer initiation, progression, maintenance and treatment are crucial in helping to create a better standard of care for patients.

Cancer has been studied in some form or another since several hundred years BC, when the term "carcinoma" was termed by Hippocrates. The vast majority of our understanding into the underlying biological causes has been developed over the last 70 or so years, enabled by the elucidation of the DNA structure and sequencing technology. Now, international consortia such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), have already sequenced tens of thousands cancer genomes. An on-going effort by Genomics England aims to sequence 100,000 genomes of National Health Service patients with cancer and rare diseases in the United Kingdom (Genomics England). As of July 2018, they have sequenced their 70,000th genome.

With the explosion of genomic data available, significant improvements are being made in the diagnostic, prognostic and treatment opportunities made available to patients today and in the future.

At its' essence, cancer describes the transformation of normal cells into a malignant disease that can be lethal to the host. It encompasses over 100 distinct diseases, all of which follow essentially the same pattern of excessive proliferation and death avoidance that underpin the hallmarks of cancer defined by Hanahan and Weinberg in 2000 and refined in 2011 (Hanahan and Weinberg, 2000, 2011).

The genetic and genomic revolution has enabled the research that brought about and cemented one of the now fundamental principles of cancer research – that cancer arises from changes in the genome. Genomic events, such as point mutations, changes in copy number and others, can deregulate the function and expression levels of proteins. They imbalance the intricate protein signalling networks that regulate a range of basic

cellular functions, from response to growth factors, apoptotic blockade, cell adhesion and migration, immune evasion and many more.

As genetic alterations were understood to be drivers of cancer, the resulting oncogenic proteins became appealing targets for chemical inhibition. The standard of care for most cancers diagnosed today is a combination of surgery, chemotherapy and radiotherapy. While they reduce overall mortality on a population level, not all patients respond and they often carry harsh side effects and a risk of treatment-related mortality.

Targeted therapy is the name for a treatment that is tailored to a molecular alteration that underpins a specific cancer subtype. The first and best-cited example of a successful targeted therapy is that of imatinib, which targets the disease-defining *BCR-ABL1* fusion in chronic myeloid leukaemias (CML) (Druker et al., 2001). Imatinib became the first targeted treatment to be approved by the US Food and Drug Administration (FDA) in 2001 and five years after initial administration showed an exceptional 95% overall survival for CML patients (Druker et al., 2006).

Since then, many more targeted treatments have been approved by the FDA across a range of tumours with varying genomic biomarkers. For example, these include *BRAF*-mutated melanomas being treated with BRAF and MEK inhibitors (Flaherty et al., 2012; Hauschild et al., 2012), or EGFR inhibitors being used in *KRAS* and *NRAS* wild-type colorectal cancers and *EGFR*-mutated non-small-cell lung cancers (Douillard et al., 2013; Mok et al., 2009). Similarly, the immunotherapy agent pembrolizumab has recently been approved for the treatment of microsatellite instable solid tumours (Le et al., 2015).

Unfortunately, in advanced tumours many of the targeted treatments merely slows down disease progression and patients often relapse and exhibit new genomic alterations that bypass the inhibited mechanism. Similarly, many patients simply do not have any actionable biomarkers for tailored treatments (McDermott, 2015).

Thus, there is still a pressing need to further advance our understanding of the cellular mechanisms of different subtypes of cancer. Discoveries of new cancer drivers and markers can greatly benefit future patients by allowing early diagnosis, better prognoses and the development of new targets for cancer treatment.

1.1.2 Gene fusions in cancer research

This PhD thesis is focused on the functional role of fusion transcripts in cancers. Gene fusions have formed some of the best-cited success stories in targeted cancer treatment, starting with the treatment of *BCR-ABL1*-positive tumours with imatinib. In 2011 followed the approval of crizotinib for the treatment of *ALK*-fused non-small-cell lung cancer (Kwak et al., 2010). Aside from treatment targets, gene fusions also play major roles in driving tumourigenesis in various tumour subtypes.

Since the 1970's, sequential improvements in technology allowed the study of gene fusions in greater and greater detail. More recently, the invention of paired-end RNA-sequencing (RNA-Seq) technology opened up new opportunities to study gene fusions at even greater scale. A paper published in March 2018 represents a further milestone, where fusions were called from RNA-Seq data from almost 10,000 TCGA patient tumours (Gao et al., 2018). With large-scale fusion detection made accessible in this age of high-throughput sequencing, the next challenge is to sift through biological passengers and technical artefacts in order to identify functional drivers of oncogenesis and markers for treatment response.

To address this question, I used a panel of 1,011 genomically annotate human cancer cell lines. The cell lines also have been screened across over 400 drug compounds and a subset of cell lines also has essentiality profiles of over 18,000 genes that were generated through CRISPR/Cas9 whole genome drop-out screens. Our group has now generated RNA-Seq data for these cell lines, and analysed it with fusion-calling algorithms.

The aim of my dissertation has been to 1) analyse and apply a set of filters to remove technical artefacts and arrive at a final set of fusions and 2) conduct a set of analyses that would attribute a functional role to fusions detected by large-scale sequencing approaches. This work could help us discover and inform the functional implications of gene fusions in cancer.

1.1.3 PhD chapter outline

Here in **chapter 1**, I will introduce the i) concept, successes and challenges of precision cancer medicine, ii) state-of-art of the technologies and findings on gene fusions in cancer and iii) the important role of cancer cell lines as experimental models of tumour biology and treatment response, especially in the high-throughput setting.

Following that, in **chapter 2** I describe the development of a set of filtering criteria that created a working list of gene fusions with a ~72% validation rate in our cancer cell lines. In **chapter 3**, I briefly describe the landscape of the gene fusions across our panel of cell lines. In **chapter 4,** I query the association of fusion occurrence with sensitivity to over 400 drug compounds. In **chapter 5**, I develop a computational method to assess functionality of gene fusions using the CRISPR/Cas9 whole-genome data and in **chapter 6** I present the results and provide examples of fusions as novel candidates of cancer maintenance and potential treatment targets. Throughout the thesis, I attempted to highlight not just the biological significance of the results, but also points of learnings and potential improvements that emerged while working with the data. Finally, **chapter 7** summarises the previous chapters and link it back to the state-of-art.

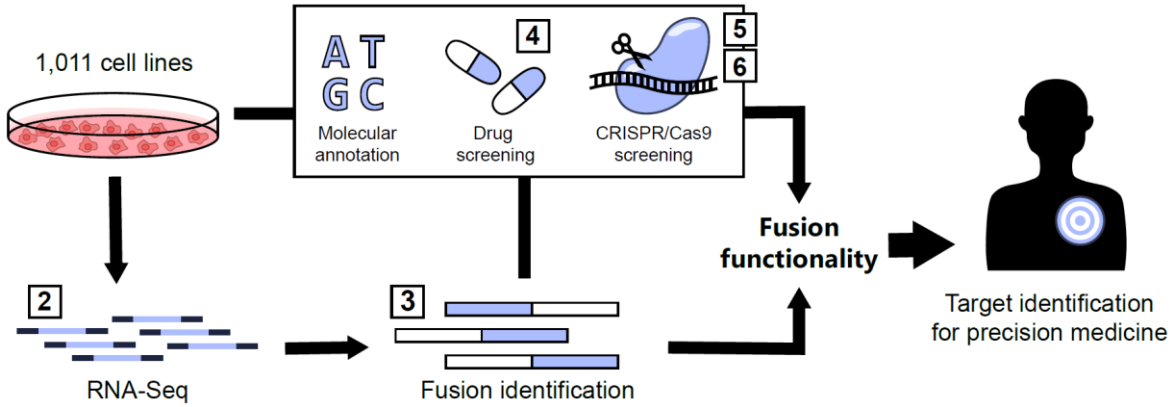


Figure 1.1: PhD thesis outline of identifying gene fusions in cancer cell lines. Using the extensive genomic and phenotypic data that is available for the cell lines, my analysis attempts to find functionally relevant gene fusions that may be used as targets of precision medicine. Numbers in boxes refer to the chapters in which I focus on certain elements of this process map.

1.2 Precision cancer medicine

A central goal of cancer research is to improve the clinical outlook for cancer patients and to increase their survival chances while maintaining an adequate quality of life. The two biggest opportunities for medical intervention are 1) by providing good treatment options available at the time of diagnosis that will eradicate or delay disease progression while minimising harmful side effects and 2) achieving earlier diagnosis of cancer, as treatment of early-stage cancers typically has much better outcomes than at advanced-stages.

Precision cancer medicine falls into the first category of improving treatment options. As indicated in the previous section, the standard-of-care often produces mixed results depending on the exact cancer diagnosis and carries the risk of harmful side-effects. Precision cancer medicine aims to select the treatment that produces the best survival outcomes for each patient according to the tumour's molecular make-up. An important principle here is that of "oncogenic addiction", whereby a cancer cell becomes dependent on constitutive activity of an oncogene for tumour maintenance and survival. Many targeted medicines aim to inhibit the oncogene and thus cause tumour regression.

The basic concept is that when a patient presents with cancer in the clinic, a sample is taken and sequenced either for a panel of cancer-associated molecular alterations and based on the genetic changes and the current state-of-the art of cancer knowledge, a suitable treatment can then be selected.

1.2.1 Case study: Non-small cell lung cancer

A prominent example illustrating precision cancer medicine is that of non-small cell lung cancer (NSCLC). Lung cancer, of which 85% are classified as NSCLC, is one of the most common and deadly cancers world-wide, partially due to the wide-spread production and consumption of its main carcinogens, tobacco smoke and air pollution (Reck and Rabe, 2017). When a patient presents at the clinic and is diagnosed with NSCLC, a biopsy is taken and tested for the presence of 1) activating *EGFR*, *BRAF* and *KRAS* mutations using sequencing, 2) *ALK* or *ROS1* rearrangements using fluorescence in situ hybridisation (FISH) and 3) *PDL-1* expression levels using immunohistochemistry. Treatment is then chosen accordingly (Reck and Rabe, 2017).

Firstly, mutations in epidermal growth factor receptor (*EGFR*) occur in about 10-20% of Caucasians and 48% of East Asians, and are a good indicator of response to EGFR inhibitors, such as gefitinib, afatinib and erlotinib. Compared to standard-of-care chemotherapy, these targeted EGFR inhibitors effectively doubled progression-free survival time from 5.1-6.1 months to 10.0-11.8 months (range depended on the exact nature of the *EGFR* mutation) (Lee et al., 2015a). Moreover, *EGFR* inhibitors tend to show lower toxicity than chemotherapy, with a lower percentage of patients with grade 4 side effects (Maemondo et al., 2010).

Rearrangements in ALK receptor tyrosine kinase (*ALK)* and ROS proto-oncogene 1 (*ROS1)* occur in 1-7% of tumours and qualify a patient for treatment with crizotinib, an inhibitor of multiple tyrosine kinases: MET proto-oncogene (MET), ROS1 and ALK. Treatment outcomes in terms of response rate and progression-free survival were better in all clinical trials using crizotinib for *ALK*-rearrangements for both first-line and advanced-stage treatments (response rates of 65%-74% vs. 20%-45% using chemotherapy and median progression-free survival: 7.7-10.9 months vs. 3.0-7.0 months) (Shaw et al., 2013; Solomon et al., 2014). Advanced NSCLC with *ROS1*-rearrangements performed even better with crizotinib, with a 72% response rate and median progression-free survival of 19.2 months (Shaw et al., 2014).

Another 2% of patients carry a B-Raf proto-oncogene (*BRAF)* mutations, which is treated with a combination of BRAF inhibitor dabrafenib and MEK inhibitor trametinib and has showed a response rate of 54% in a clinical trial of advanced-stage NSCLC (Planchard et al., 2016).

Conversely, a negative indicator for response is a mutation in KRAS proto-oncogene (*KRAS*). *KRAS* mutations occur in about 30% of Caucasians (with a higher incidence in smokers) and mainly cause the constitutive activation of the RAF-MEK-ERK pathway of proliferative signalling and the PI3K-AKT pathway of anti-apoptotic signalling (Suda et al., 2010). Unfortunately, NSCLC patients with *KRAS* mutations generally have very poor prognosis with standard-of-care (Slebos et al., 1990) and show no response to *EGFR* inhibitors (Pao et al., 2005). KRAS itself has proven itself undruggable so far and a clinical trial found no added benefit of targeting downstream MEK protein in combination with chemotherapy compared to standard chemotherapy (Jänne et al., 2016).

1.2.2 Case study: *BRAF* mutations

While the aim of precision medicine is to tailor the therapy to the genetic background of cancers, an important factor to consider is also the tissue type of origin. A particularly potent example of that is the example of the tyrosine kinase *BRAF*, an oncogene of the RAS-RAF-MEK signalling pathway that is mutated in various different tissue types, including melanoma, colorectal cancers, and ovarian carcinomas, and the majority of mutations are a *BRAF*(V600E) mutation in the activation loop (Davies et al., 2002).

The observation of the wide-spread mutation of *BRAF* across several cancer types lead to the development of *BRAF* inhibitors and vemurafanib was the first one to be approved after it showed a 48% response rate in *BRAF*-mutated advanced-stage melanomas, compared with 5% with chemotherapy (Chapman et al., 2011).

However, in colorectal cancers, where about 10% of patients showed a *BRAF*(V600E) mutation, the response rate was only 5% with median progression-free survival at only 3.7 months (Kopetz et al., 2010). Two years later, an RNA interference screen revealed that the mechanism of BRAF inhibitor resistance in colorectal cancer was a rapid feedback activation of EGFR by phosphorylation, which activates other RAF family kinases (Prahallad et al., 2012). The authors further showed that melanomas had lower levels of *EGFR*, which is typically expressed in epithelial tissues, and explaining their relative sensitivity to BRAF inhibitors. A recently published clinical trial treated 91 advanced-stage metastatic patients with a triple-combination of EGFR, BRAF and MEK inhibitors (Corcoran et al., 2018). They reported promising response rates of 21% in a disease subtype which outcomes are usually poor despite chemotherapy treatment (Tran et al., 2011).

Overall, this case demonstrates how the treatment of the same genetic subtype requires different considerations in different cancer types. While BRAF inhibitors work well in melanomas with *BRAF*(V600E) mutations, in colorectal cancer combinations with EGFR and MEK inhibitors are required to suppress primary resistance. At the same time, this case also illustrates that in instances where monotherapy does not show a desired effect, combinations may be used to target resistance mechanisms.

1.2.3 Challenges in precision cancer medicine

Despite promising case studies that show that an understanding of cancer biology can lead to better treatment options for patients, the successful deployment of precision cancer medicine still faces many challenges.

Firstly, the number of genetic biomarkers used in the clinic is still small. A large proportion of the patient population do not present with actionable mutations and will be treated with the standard of care.

Secondly, as illustrated in the examples above, even in a carefully pre-selected patient population, a proportion of patients do not respond to the treatment. While the development of better drugs can lead to improved response rates compared to initial trials, they are still rarely even close to 100%.

At the same time, only a few patients show a stable disease over a prolonged time period and many patients eventually develop resistance to targeted medicines. For instance, the mitogen-activated protein kinase (MAPK) pathway forms a central node in proliferative signalling which involves BRAF, MEK and ERK. This pathway is commonly targeted using inhibitors, however resistance often develops through a variety of genetic alterations that lead to increased reactivation of the pathway. These include *KRAS* and NRAS proto-oncogene (*NRAS*) activation by mutation or amplification, *BRAF*(V600E) amplification and other mutations that either lead to mutation in the drug-inhibitor binding sites, hyper-activation of key pathway components or by-passing the blocked signalling pathways (Ahronian et al., 2015).

In the context of acquired resistance to targeted therapies, intra-tumoural heterogeneity also represents a challenge as resistant subclones may already be present in tumours at the time of treatment (Diaz Jr et al., 2012).

In order to combat the issue of treatment resistance, it will be important to clarify the exact mechanisms of resistance. In particular, cheaper, more readily available and reliable next-generation sequencing technologies allows the detection of genome-wide alterations after treatment. Improvements of liquid biopsy technologies may facilitate the collection and analysis of cancer DNA and cells from the blood. As illustrated with the BRAF/MEK/EGFR triple combination, drug combinations may delay the emergence of resistance by targeting multiple components of signalling pathways at once.

Precision cancer medicine may also benefit from combining or stratifying patients for treatment with immunotherapy in addition to targeted therapy. In melanoma and other cancer with a high mutation burden, antibodies that target negatively regulating immune receptors such as PDL-1 and CTL-4 have shown great promise in recent years. For instance, ipilimumab, an anti-CTL-4 antibody achieved a sustained 3-year progression-free survival rate of 21% in advanced-stage melanomas (Schadendorf et al., 2015). Thus the ability to predict which patients would perform better with immunotherapy vs. targeted therapy, or which would benefit from combining both treatment types, could make a significant difference in future treatment outcomes.

Ultimately, advances in the discovery of novel cellular processes, new biomarkers and drug targets will rely heavily on the availability of pre-clinical models which can model the disease and be used as proxies for anti-tumour compound testing. Later, I will introduce cell lines as experimental models of cancer in section 1.4. The next section will first describe gene fusions, their role in cancer precision medicine to date and open opportunities in the future.

1.3 Gene fusions in precision cancer medicine

In order to understand how we arrived at our current understanding of gene fusions in human cancers, it is useful to first understand the experimental conditions that provided the opportunities and limitations in gene fusion analyses since the 1960's. Thus, I will first provide a short overview of the seminal experimental techniques that paved the way from the most basic cytogenetics to next-generation sequencing. Then, I will describe the experimental and computational processing of samples for fusion-calling from RNA-Seq to highlight currently available tools and their limitations.

Finally, in this chapter I will provide context on the current state of knowledge of the types of fusions known to contribute to cancer progression and give an overview of recent efforts to provide landscapes of fusions across large cancer cohorts.

1.3.1 Brief history of technological advances in gene fusion analysis

In the 1960, only five years after Peter Harper discovered that the number of human chromosomes is 46, Peter Nowell and David Hungerford characterised a recurrent abnormally small chromosome in patients with chronic myeloid leukaemia (Nowell and Hungerford, 1960). This "minute" chromosome was soon dubbed the Philadelphia chromosome, named after the city in which it was discovered. It was the first time that a specific genetic alteration was found to be so strongly recurrent in patients of a given tumour subtype. Fast forwarding through the years with the development of new cytogenetic techniques, the Philadelphia chromosome was of course later identified as a translocation between chromosomes 9 and 22, which resulted in the fusion of breakpoint cluster region (*BCR*) to ABL proto-oncogene 1 (*ABL1*) (Groffen et al., 1984; Rowley, 1973).

At a time when the community believed that viruses were one of the key causes of cancer, Nowell and Hungerford's finding provided early evidence to support the hypothesis that a critical single genetic change could give rise to a tumour (Nowell, 2007).

The analysis of chromosomal translocations and gene fusions in cancer evolved significantly throughout the years, from 1960 to today. In the 1960's, Peter Nowell discovered by serendipity that rinsing cells with tap water made the chromosomes expand (Nowell and Hungerford, 1960). In the following two decades, chromosomal banding techniques enabled scientists to distinguish different chromosomal regions based on their

banding patterns. The study of these disrupted and rearranged banding patterns led to further discovery, such as t(8;14)(q24;q32), i.e. *IGH-MYC*, in Burkitt lymphoma (Zech et al., 1976), or t(11;22)(q24;q12), i.e. *EWSR1-FLI1* in Ewing's sarcoma (Aurias et al., 1983). From the 1980's onwards, developments in DNA cloning, capillary sequencing and fluorescence *in situ* hybridisation (FISH) provided further tools to study chromosomal translocation at ever higher resolution.

The first step into the direction of high-throughput fusion analysis came with the invention of array-based genetic tools, in particular gene expression and copy number arrays. For instance, in 2005, Tomlins and colleagues used microarray gene expression data to identify the now widely studied transmembrane serine protease 2 (*TMPRSS2*) fusions in prostate cancers (Tomlins et al., 2005). They did this by first mathematically identifying gene expression outliers in individual samples in publicly available data. Among the top results were recurrent outliers in the ETS transcription factor ERG (*ERG*) and ETS variant 1 (*ETV1*) genes, which are members of the ETS transcription factor family to which also Fli-1 proto-oncogene (*FLI1*) (of the *EWSR1-FLI1* fusion) belongs. Further sequencing in cell lines with the same expression patterns then revealed the mechanism of high expression to be a fusion of the 5' untranslated region (UTR) to that of the androgen-regulated *TMPRSS2* gene. The discovery of the *TMPRSS2-ERG* was further significant, as it was the first fusion found in an epithelial-derived tumour, while previously oncogenic fusions were thought to be exclusive to blood cancers and sarcomas.

Now, next generation sequencing has opened up a new frontier in high-throughput analyses of gene fusions. For the first time an unguided discovery of gene fusions at large-scale and with nucleotide-level resolution was possible in a single experiment.

First, whole genome sequencing (WGS) allowed researchers to look at structural rearrangements on a genomic level. In a seminal study, Campbell and colleagues used paired-end DNA sequencing data from two lung cancer cell lines (Campbell et al., 2008). For paired-end sequencing, sequencing adapters are fitted to both ends of a DNA fragment of a defined length (often ~200-400 bp). When paired nucleotide sequences consistently map to genomic regions that are further apart than the expected fragment length, there is strong evidence for structural rearrangements. Applied on a whole genome level, these types of computational analyses allowed researchers to create a map of structural rearrangements across the entire genome of a single sample.

In 2009, Maher and colleagues published the first paper that identifies gene fusions from whole transcriptome sequencing, i.e. RNA-Seq, from both single long reads and paired-end reads (Maher et al., 2009a, 2009b). Unlike WGS, where there is no information on whether a structural rearrangement results in an expressed transcript or not, analysis using RNA-Seq data only detects expressed transcripts. At the same time, RNA-Seq data can report instances where gene splicing events leads to alternative transcripts.

While the first studies focused on small panels of a handful of cell-lines at a time, since then the number of tumours that have been sequenced by RNA-Seq increased dramatically. A review written in 2015 by Mertens and colleagues highlighted that of the ~10,000 gene fusions known in tumours that year, over 90% were discovered by deep-sequencing approaches (Mertens et al., 2015). More recently, a paper published in March 2018 found almost 24,000 distinct gene fusions in over 9,500 TCGA patient samples (Gao et al., 2018).

In order to facilitate the interpretation of the huge amount of RNA-Seq data that has been generated so-far, numerous computational tools have been developed. A particular challenge is that RNA-Seq data interpretation is error-prone and thus new algorithms aiming to provide robust fusion-calling function are continuously being developed, with varying degrees of success. In the next section, I will review the difficulties of calling fusions in RNA-Seq data and some of the methods available to solve them.

1.3.2 RNA-Seq data generation and analysis

1.3.2.1 A note on terminology

In order to establish a common vocabulary for the comparison of multiple RNA-Seq processing algorithms, in my thesis, these terms should hereby be defined as following. Also see Figure 1.2 for an illustration of selected definitions.

Transcript: a full length RNA strand that represents a uniquely expressed unit in the cell.

Fragment: generated by breaking up transcripts, often into a few hundred base-pair long strands. Fragments are physically captured in their entirety as part of the sequencing reaction.

(Sequencing) read: a part of a fragment that is actually sequenced. Reads can be sequenced from both ends of a fragment (paired-end RNA-Seq), or from a single-end. The

length of the sequenced read depends on the machine and approach used for sequencing, but is typically 30-100 bp long.

Alignment: computationally generated by comparing the read sequence to a reference genome to identify the genomic location from which a read is predicted to originate. Due to sequence homology in the genome, a single read can have multiple alignments that fit perfectly or with a low number of base mismatches.

Fragment insert: For paired-end sequencing, the sequence between the read pairs that is not sequenced on the sequencing platform. Where read pairs map on a genomic location that is approximately the size of the fragment length this sequence is inferred.

Spanning reads: a pair of reads that map onto genomic regions that are further apart than the expected fragment length. This suggests that a fusion has occurred, and that the breakpoint is inferred to be within the fragment insert sequence.

Splitting reads: a pair of reads, for which one read maps onto a standard genomic region. The other read maps directly across the fusion breakpoint, thus revealing the exact breakpoint sequence.

Cluster (of reads): multiple splitting and spanning reads that are predicted to be derived from the same fusion transcript and are thus grouped together.

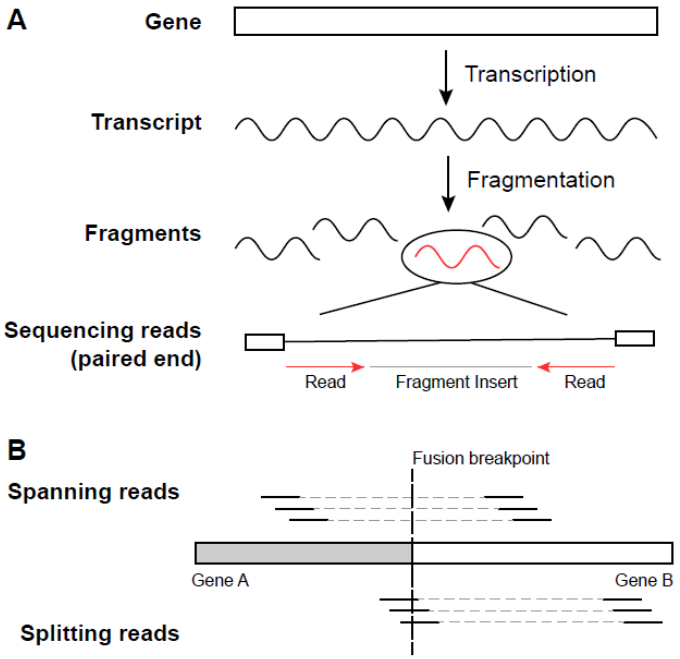


Figure 1.2: (A) Basic concept of paired-end sequencing. (B) Spanning reads and splitting reads are defined by whether or not the fusion breakpoint is captured in the read sequence.

1.3.2.2 *Brief overview of RNA-Seq data generation from samples*

In order to collect the RNA used for sequencing, cell are first grown for several days in order to achieve adequate cell volume. To extract the RNA, centrifugation generates a concentrated cell pellets which are then lysed and homogenised to release the cell contents. Proteins are separated from nucleic acids by treating the lysate using a standard phenol:chloroform technique (Sambrook and Russell, 2006). Then, typically using commercially available kits, the nucleic acid mixture is passed through a spin column that contains a membrane that selectively binds and retains RNA while DNA is washed out. DNA can be further removed using DNAse enzymes that selectively digest DNA inside the spin columns. To further increase purity of RNA mixture and to remove any other potential forms of contaminants, an extra step of RNA-clean-up is provided by kits such as Agencourt's RNAClean XP kit, which uses magnetic beads that bind to RNA molecules and removes excess nucleotides, salts and enzymes from the solution. Next, RNA is converted to cDNA prior to sequencing, amplified and fragmented into fragments of several hundred base pairs. Finally, in order for the fragment to be recognised by the sequencing platform, adapters are fitted on both ends of each fragment and the fragment is then processed for next-generation sequencing (typically sequencing-by-synthesis).

Sequencing identifies the nature and order of nucleotides on the captured fragments. Due to a higher likelihood of sequencing errors at greater distances from the sequencing adaptors, reads are typically 30-100 bp long. In order to identify the location of the genome from which a read originates, computational tools called "aligners" compare the read sequence to a reference genome and output "alignments" that indicate the most likely area from which a read originates (see definitions in section 1.3.2.1).

The computational methodology and tools that predict the presence of fusions from RNA-Seq alignments will be discussed in the next section.

1.3.2.3 *Computational fusion-calling tools*

Since the invention of RNA-Seq in 2008, over 20 different computational algorithms for fusion-calling have been produced (Carrara et al., 2013; Haas et al., 2017; Kumar et al., 2016). The methodology of fusion-calling can be distilled into a simple principle: to identify sequenced reads (whether single or paired) which partially map onto two different regions of the genome. In general, computational fusion-calling tools all follow three basic steps:

1) **Chimeric transcript identification**: this is done by finding reads that map onto two distinct regions in the genome. Some algorithms (e.g. deFuse and Maher *et al*'s method) do this by mapping all RNA-Seq normally and then finding spanning reads, i.e. paired-end reads that map onto two different regions. Other algorithms (e.g. TopHat-Fusion and STAR-Fusion) first identify splitting-reads, i.e. reads that contain a potential fusion-breakpoint and thus do not map perfectly onto any genomic region, which they then split into several smaller segments which can then be aligned to unique regions. While conceptually both approaches are logical ways of identifying the same fusion transcripts, in practice they use different read pools to flag potentially chimeric reads, and may thus lead to different final results.

Additionally, fusion algorithms use different methods of resolving ambiguous alignments. For instance, while Maher's script and STAR-Fusion use only the best-matching alignment for each read, TopHat-Fusion removes any read with more than two alignments, and deFuse uses a "maximum parsimony" solution which considers all alignments and chooses the one matching the most likely fusion transcript. Again, these differences in read processing can lead to different final results.

2) **Grouping of reads to resolve final fusion breakpoint:** In order to retrieve the full sequence around the putative fusion breakpoint, reads are grouped, typically based on genomic mapping location either before or after identification of the chimeric reads. Here, some algorithms (e.g. deFuse) take advantage of the paired-end system, by searching for reads that map the predicted fragment length away from the putative fusion boundary region – the paired read is likely to reveal the exact fusion boundary.
3) **Further filters to exclude likely false positives:** This step is perhaps one of the most differentiating points between algorithms. Filtering the initial results can reduce the number of candidate fusion events by almost 99% (Kim and Salzberg, 2011; McPherson et al., 2011). Many algorithms devise extensive filters, but some considerations are commonly observed among most algorithms, e.g.: i) the minimum number of reads that provide evidence for the breakpoint, ii) homology in nearby regions that may cause mapping errors, iii)

differential alignment coverage and gaps in the genomic area immediately surrounding the putative breakpoint.

Notably, while on the perfect data-set, fusion-calling algorithms should yield similar results, in practice the output is often vastly different. For instance, different algorithms can yield from 10 to over 5,000 fusion calls on the exact same sets of RNA-Seq data (Carrara et al., 2013). Applied by different groups on different samples, the same algorithm can also output anything ranging from a handful to several 100,000. (For a more extensive comparison and benchmarking of fusion-calling algorithms, see Chapter 2). Aside from intrinsic differences in the methodology, these differences likely also reflect that small changes in the algorithm parameters can have large effects on the effectiveness of the filters.

1.3.2.4 Challenges and limitations

Since fusion-calling by definition is the unguided detection of non-standard fragments, standard methods that recognise and remove abnormal nucleotide sequences in WGS are not useful in RNA-Seq. Possible artefacts can arise from all steps from sample preparation to the alignment.

For instance, during sequencing-by-synthesis, usually a single fluorescent nucleotide is integrated per each sequencing cycle. If either none or multiple bases integrate in one cycle, this can lead to so-called "phasing"-errors. Similarly, an effect called "cross-talk" can arise when fluorescent emission spectra are insufficiently resolved by the sequencing machine (Ledergerber and Dessimoz, 2011). In the short-reads that are generated from NGS, small mismatches can thus lead to misalignments in later steps.

Another source of error comes from the fact that in order to be sequenced on NGS platforms, RNA is first converted to complementary DNA (cDNA) using reverse transcriptase enzymes. Previous analyses have shown that the enzymes are relatively error-prone, for example adding artefactual single-base substitutions and occasional frameshifts (Roberts et al., 1989). Spontaneous template switching of the reverse transcriptases can lead to chimeric transcripts (Cocquet et al., 2006) introduced after sample collection that could be misinterpreted as real gene fusions. Furthermore, contamination of either DNA or other organisms can add artefacts as algorithms permissively align multiple regions.

While higher sequencing depth should add more support to real gene fusions, the larger number of RNA particles may also increase the number of false positives. Unfortunately many of the fusion-calling algorithms have been developed with and benchmarked against test data-sets with similar ranges of read number (around 20-30 M reads per sample), and to the best of my knowledge no formal analysis has been performed to show the influence of increased read depth on the number of fusions detected.

In summary, issues with error-prone RNA-Seq sample processing and sequencing differences in fusion-calling algorithm design can lead to vastly different outputs for the same tested sample. Further, there is a lack of clear, empirically-determined guidelines for parameter and tool choice for RNA-Seq data-sets of different read depths and fragment sizes, which is exemplified and perhaps aggravated by the observation that algorithms can show different accuracies whether benchmarking is performed on simulated or real RNA-Seq data (Haas et al., 2017).

A trend is recently emerging, where instead of using a single algorithm on its default setting, researchers now use multiple fusion-calling algorithms for the same study and tailor filtering criteria based on their own data. Indeed, in our study we use three different algorithms, the processing steps for which I will describe in chapter 2. In our data, we do find that fusions that are picked up by multiple fusion-calling algorithms had very high validation rates by polymerase chain reaction (PCR) and most known oncogenic fusions that we find in our data sets are supported by multiple algorithms. However, when using only the overlap of multiple different algorithms, one runs the risk of discarding real putative fusions that perhaps due to low expression are only picked up by single algorithms.

Overall, fusion-calling from RNA-Seq data still represents a highly promising and perhaps the only way to call fusions in a high-throughput setting with nucleotide-level resolution. Improvements in technology, e.g. direct sequencing of RNA without the conversion into cDNA and longer read lengths, will likely allow improved sensitivity and specificity for computational algorithms. Until then, one should be careful not to over-interpret fusions called from RNA-Seq data and validate promising candidates using traditional techniques such as FISH, PCR and capillary sequencing.

1.3.3 The structural and functional context of gene fusions

This section is dedicated to understanding how gene fusions can occur and how they are able to modulate gene function in driving oncogenic processes.

1.3.3.1 Structural and non-structural changes can lead to the formation of gene fusions

The best-understood examples of oncogenic fusions arise from structural rearrangements in the genome, and often from balanced rearrangements. These changes refer to rearrangements such as translocations, insertions and inversions, in which a genomic region is transferred to another location in the genome. Unlike in unbalanced rearrangements such as deletions, duplications and aneuploidies, in balanced rearrangements the overall copy number does not change.

While specific initiation factors for gene fusions are largely unknown, previous research has indicated that DNA double-strand breaks (DSB) are required for gene fusions (Aplan, 2006). Inherited syndromes that are associated with genetic instability and the inability to efficiently repair DSBs can lead to a predisposition of chromosomal translocations and hematologic cancers, e.g. in Nijmegen's breakage syndrome patients that show nibrin (*NBN*) mutations. *NBN* is normally involved in non-homologous end-joining (NHEJ), one of the main mechanisms of DSB repair (Digweed and Sperling, 2004). Similarly, knocking out other NHEJ-related genes in mice induced high rates of chromosomal translocations and cancer development (Aplan, 2006). Chemotherapy agents that inhibit DNA topoisomerase II can also cause increased occurrences of therapy-related chromosomal translocations. DNA topoisomerase II normally play a role in untangling DNA loops and supercoils by creating and later re-ligating DSBs. *In vitro* treatment of cells with the chemotherapy agent etoposide has shown that stabilisation of the cleaved complex created by DNA topoisomerase II could lead to chromosomal translocations between different stalled cleaved complexes (Zhou et al., 1997).

Furthermore, DNA-region specific factors are also believed to play a role in causing recurrent chromosomal translocations. These include proximity of different genomic regions in the nucleosome during interphase. But also sequence-specific features can play a role – for example alternating purine and pyrimidine tracts create a special helical structure that tends to be found in inter-nucleosomal regions and may thus be especially prone to DSBs (Garner and Felsenfeld, 1987; Thandla et al., 1999). Similarly, the lysine methyltransferase 2A (*KMT2A*) and super elongation complex subunit (*MLLT3*) loci that are

commonly fused in leukaemias and lymphomas were found to be nuclear scaffold-associated regions that are thought to be attachment sites for chromosomal loops, and thus especially sensitive to DSBs in the context of topoisomerase II malfunction (Stanulla et al., 2001).

In terms of unbalanced structural rearrangements, fusion transcripts may commonly result from deletions and amplifications. Deletions of regions within tumour suppressors can lead to the expression of aberrant transcripts that are most likely to have lost the original function (e.g. in our panel, we find several fusions that involve the phosphatase and tensin homologue (*PTEN*) tumour suppressor). Amplifications are similarly one of the common mechanisms by which oncogenes are overexpressed; well-known examples include erb-b2 receptor tyrosine kinase 2 (*ERBB2*), *EGFR*, MYC proto-oncogene (*MYC*) and fibroblast growth factor receptor (*FGFR*). The repeated breaking and re-joining of DNA that pre-date amplifications is likely to facilitate the formation of fusion transcripts. Indeed a survey of 14 breast cancer cell lines showed that a large proportion of detected gene fusions map to amplicons, although it is possible that the majority of these types of events are passenger events without cancer-related functionality (Kalyana-Sundaram et al., 2012).

Other than structural changes that hard-code gene fusions into the genome, fusions have also been detected that have no detectable underlying structural rearrangements. So-called transcription-induced gene fusions (TIGF) can occur through either *cis*-splicing or *trans*-splicing events. *Cis*-splicing, also known as gene read-throughs, occurs when neighbouring genes are transcribed and then spliced together to produce a chimeric transcript. An example of a *cis*-spliced TIGF is a fusion transcript found in prostate cancer samples between solute carrier family 45 member 3 (*SLC45A3*) and ETS transcription factor (*ELK4*), which is another member of the *ETS* family of transcription factors that *ERG* belongs to (e.g. in *TMPRSS2-ERG* and *EWSR1-ERG*) (Rickman et al., 2009). The two genes are located approximately 50 kb apart from each other and the fusion transcript was not accompanied any detectable structural rearrangements. Further research showed that *SLC45A3* and chimera expression similar to the *TMPRSS2-ERG* fusion is androgen-regulated, that *SLC45A3-ELK4* expression correlated with more advanced stages of prostate cancer and that siRNA knock-down of *SLC45A3-ELK4* significantly reduced cell

line growth (Zhang et al., 2012). In 2011, a paper that surveyed TIGF events using deep sequencing techniques showed that TIGF's are common events in both tumour and normal samples and that they are associated with high expression of the 5′ fused gene partner (Nacu et al., 2011).

In contrast to *cis*-splicing events between neighbouring genes, *trans*-splicing events can occur between genes that are further apart and even on separate chromosomes. These result when mRNA are expressed separately for two genes and subsequently spliced together before translated into a chimeric protein (Finta and Zaphiropoulos, 2002). One example in cancer is the fusion transcript of exons 1-3 of JAZF zinc finger 1 (*JAZF1*) and exons 2-16 of polycomb repressive complex 2 subunit (*SUZ12*), which are found on chromosomes 7 and 17 respectively (Li et al., 2008). *JAZF1-SUZ12* had previously been found in endometrial stromal sarcomas that presented a chromosomal translocation, and expression of the transcript in HEK cells had shown suppression of hypoxia-induced apoptosis (Li et al., 2007). Intriguingly, an expressed *JAZF1-SUZ12* transcript was also found in non-neoplastic endometrial cell lines, however extensive analysis of the cellular DNA, FISH and karyotyping showed no evidence underlying translocation or rearrangements (Li et al., 2008). It is important to note that an independent study was unable to verify the presence of a *JAZF1-SUZ12* transcript in the endometrial cell line, which suggests that if true, these events may be transiently expressed only in specific circumstances (Panagopoulos, 2010).

While the presence of TIGFs is well validated, overall there are only few replicated examples of transcription-induced gene fusions with a cancer-related role. Nonetheless, they are likely to make up a significant proportion of fusion transcripts detected from RNA-Seq data and their potentially transient nature may explain discrepancies between fusions identified in the same cell lines during different experiments.

1.3.3.2 *Gene fusions in deregulating normal protein function*

Known oncogenic gene fusion events perform their function using two main mechanisms: they either create a translated chimeric protein, or deregulate mRNA expression through the fusion of regulatory untranslated regions (UTR's) (Figure 1.3).

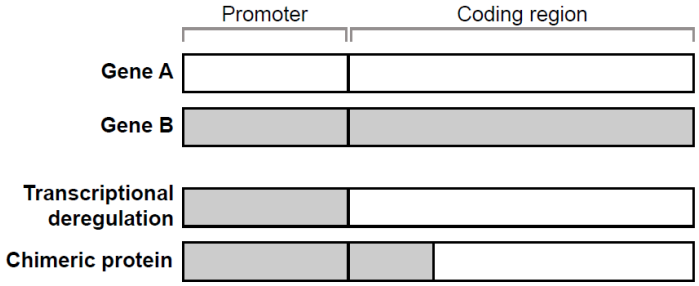


Figure 1.3: Two main types of oncogenic gene fusion.

Chimeric proteins make up the most common and best understood examples. Most commonly, one of the gene partners has normal activity as a kinase or transcriptional co-activators, and the fusion either removes a self-inhibitory domain and/or adds a domain that facilitates activation of the catalytic domains. Next, I describe two examples to illustrate some of the mechanisms of oncogenic activity.

In the case of the *BCR-ABL1* fusion, ABL1 contains three retained SRC homology (SH) domains that facilitate protein-protein interaction and tyrosine kinase activity (Figure 1.4). Also retained in the chimeric protein is a nuclear localisation domain and actin binding domain. Conversely, the fusion eliminates an N-terminal inhibitory cap and a myristoylation site, both of which stabilise an inactive conformation of ABL1 (Colicelli, 2010). On the other hand, the N-terminal fused portion of BCR contains a coiled-coil domain, which allows the formation and kinase activity of BCR-ABL1 homotetramers (McWhirter et al., 1993). Aside from BCR, ABL1 is also found to be fused to multiple other

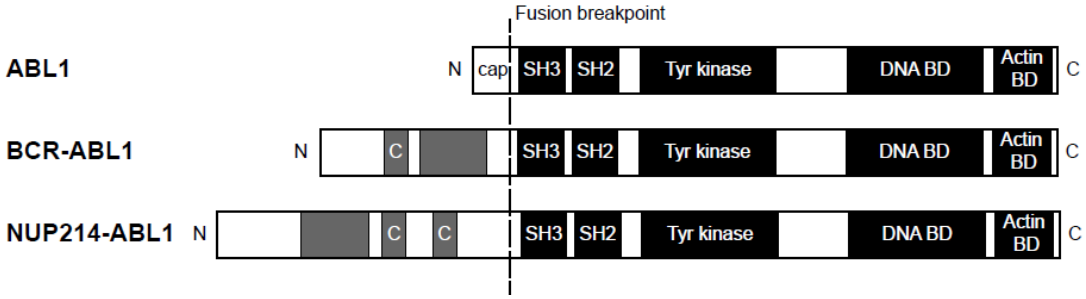


Figure 1.4: Known breakpoints of selected ABL1 fusions. BD = binding domain. C = coiled-coil domain. Adapted from De Braekeleer et al. (2011)

proteins in rarer subtypes. Many of the other fusion partners, for instance nucleoporin 214 (NUP214), echinoderm microtubule associated protein like 1 (EML1) and RCSD domain containing 1 (RCSD1), also carry similar coiled-coil domains or helix-loop-helix domains that allow oligodimerisation (Braekeleer et al., 2011). Similarly to this mechanism, the oncogenic activity of *ALK* fusions (e.g. *NPM1-ALK*, *EML4-ALK*) results from constitutive activation of the ALK tyrosine kinase domain by fused partner genes which encourage dimerisation (Mossé et al., 2009).

An example where the fusion involves a transcriptional co-activator is the fusion between CREB regulated transcription coactivator 1 (*CRTC1*) and mastermind like transcriptional coactivator 2 (*MAML2*), which is found in about 50% of mucoepidermoid head and neck carcinomas, with transforming activity in cell lines (Chen et al., 2014). MAML2 normally acts as a transcriptional co-activator with NOTCH receptors, with an N-terminal domain that enables it to bind to Notch receptors and a C-terminal transcriptional activation domain which facilitates transcription of down-stream Notch pathway genes (Wu et al., 2002). Conversely, CRTC1 is a co-activator of the CREB pathway which plays multiple roles in cell proliferation, differentiation and apoptosis (Sakamoto and Frank, 2009). In this chimeric protein, the N-terminal CREB-binding domain of CRTC1 and the C-terminal transcriptional activation domain of MAML2 are brought together and lead to transforming activity by constitutive CREB pathway activation (Wu et al., 2005).

Another type of oncogenic transformation is enabled by **transcriptional deregulation**, which can occur at various regulatory regions. For instance, the fusion of *TMPRSS2* to ETS family transcription factors in prostate cancer occurs at the 5' untranslated region (UTR) (Tomlins et al., 2005). Expression levels of *TMPRSS2* are regulated by androgen, and thus in prostate tissue with high androgen levels this type of "promoter switching" leads to overexpression of the ETS genes. Consequently, this has downstream effects on differentiation, and cellular invasion (Tomlins et al., 2008).

In Burkitt lymphomas and many other B-cell related cancers, a translocation between *MYC* and various immunoglobulin loci (*IGH/IGK/IGL*) is common. This puts *MYC* under the regulatory control of immunoglobulin enhancers and leads to stark overexpression of the *MYC* transcript (Dalla-Favera et al., 1983). Similar scenarios are found

in T-cell related cancers, where oncogenes are commonly fused to regulatory elements of T-cell receptor genes (Vlierberghe and Ferrando, 2012).

A haematological oncogene, TAL bHLH transcription factor 1 (*TAL1*), can be fused at both the 5′ and 3′ UTRs. Indeed, the first discovered case of a *TAL1* fusion in patients was a translocation that lead to a fusion at the 3′ UTR of *TAL1* with regulatory elements of the TCR locus in a T-cell leukaemia (Begley and Green, 1999). A second mechanism of *TAL1* activation is through a 90 kb deletion that fuses the coding regions of *TAL1* to the 5′ regulatory regions of the ubiquitously expressed *STIL* gene (Brown et al., 1990).

Notably, structural rearrangements that lead to transcriptional deregulation of oncogenic genes do not necessarily lead to chimeric transcripts. Chimeric transcripts can be detected for fusions that involve transcribed untranslated regions such as *TMPRSS2-ERG*, however immunoglobulin and TCR translocations are often only detectable on a structural level, since only the surrounding regulatory regions and not transcribed regions are fused.

Finally, structural rearrangements may perform cancer-driving function by silencing tumour suppressors effectively by **gene truncation**. For example, individual cases have been reported whereby a translocation caused the complete inactivation and loss of expression of tumour suppressors cyclin dependent kinase inhibitor 2A (*CDKN2A*) in a B-cell leukaemia (Duro et al., 1996) and a biallelic inactivation of *NF1* in sporadic neurofibromas (Storlazzi et al., 2005). Similarly, a survey of 180 TCGA acute myeloid leukaemia samples found several truncating fusions across various samples, including some of the tumour suppressor runt related transcription factor 1 (*RUNX1*) (Cancer Genome Atlas Research Network et al., 2013).

In summary, in the above section I described the structural and non-structural mechanisms by which functional fusions transcripts can be formed in cancer cells. Further, functional fusions generally exert their functionality either by producing an aberrant and typically constitutively activated chimeric protein, or by transcriptional deregulation that leads to overexpression.

1.3.4 Gene fusions in cancer

Throughout the previous sections, I have already given various examples of gene fusions that throughout the last three decades have been found to play a role in cancer. In this section, I would like to take a more detailed look at currently relevant developments of gene fusion research, in particular 1) current approaches and efforts on using RNA-Seq data to identify cancer-relevant gene fusions and 2) the role of gene fusions in precision cancer medicine.

1.3.4.1 New learnings from large-scale RNA-Seq fusion landscaping studies

As outlines in section 1.3.2, fusion-calling from RNA-Seq data is a relatively new field, with the first algorithms being developed and refined only about 10 years ago. Since then, this technology has enabled researchers to look for gene fusions in an unbiased manner without pre-requisite knowledge and in ever larger number of samples.

Many of the early analyses were conducted in a tissue-specific manner and perhaps one of the first important learnings was that expressed fusion transcripts were common without necessary playing an obvious role in carcinogenesis. For instance, in 2010 Berger and colleagues gave the first account of gene fusions in melanoma, with a small panel of 8 patient samples and 2 cell lines (Berger et al., 2010). The authors found 11 fusion transcripts which all validated by RT-PCR of which none had strong evidence for oncogenic functionality.

Focussing on known cancer driver genes did help to identify previously unknown oncogenic fusions. In 2010, Palanisamy and colleagues analysed paired-end RNA-Seq data of prostate cancers without the classic *ETS* gene fusions (Palanisamy et al., 2010). The authors found rearrangements that involved *BRAF* and Raf-1 proto-oncogene (*RAF1*), which deleted the proteins' auto-inhibitory domain, and were able to induce tumour formation in mice and were also sensitive to RAF inhibitors. Targeted sequencing of larger tumour panels confirmed that *BRAF* and *RAF1* fusions could be found in about 1-2% of prostate carcinomas (Palanisamy et al., 2010). Similarly, RNA-Seq analyses were also able to identify *FGFR3-TACC3* fusions as a rare subtype that occur in about 3% of glioblastomas (Singh et al., 2012).

Later on, RNA-Seq panels were further expanded to include several hundred, and then even thousands of samples within a single study. In 2014 Stransky and colleagues published a paper that focussed on recurrent kinase fusions in over 6,800 TCGA patient

samples (Stransky et al., 2014). This resulted in the discovery of novel fusions involving known fused oncogenes (e.g. *BRAF* and MET proto-oncogene (*MET*)) and also novel fusions with kinase genes which were previously not implicated in cancer. In 2015, Yoshihara and colleagues sequenced 4,366 patient samples in 13 tumour types and found over 7,800 fusion events (Yoshihara et al., 2015). While largely descriptive, the authors noted that only few fusions are recurrent (n = 263) and the relatively enrichment of kinase fusions in thyroid cancer and an enrichment of chromatin modifier fusions in acute myeloid leukaemias.

The fast rate of fusion gene discovery that came with next-generation sequencing technologies has also led to the establishment of databases to record cancer-associated fusion transcripts. The currently largest data base is likely ChimerDB, which was created in 2006 and regularly updated since then. At the time of the latest publication in January 2017, ChimerDB listed over 33,000 fusions, of which 1,066 were mined from other databases, 2,760 retrieved through text mining from PubMed and over 30,000 from TCGA RNA-Seq data from ~4,500 patients detected by combining the results of 2 different fusion-calling algorithms (Lee et al., 2017). It is important to point out that similar to most other large-scale efforts of annotating gene fusions, the fusions listed in ChimerDB should also be approached with a degree of caution, as the authors themselves acknowledge that the false positive rate of the fusions called from RNA-Seq data may be around half of those predicted.

Recently, in March 2018, Gao and colleagues published the analysis that encompasses the so-far largest number of samples with ~9,600 TCGA patient tumours in 33 cancer types (Gao et al., 2018). Notably, they are also one of the first large-scale papers to analyse their data using three different fusion-calling algorithms. The authors also integrated gene expression data and showed that oncogene and kinase fusions tended to be overexpressed, while tumour suppressor fusions tend to be underexpressed.

Notably, with ever easier method that facilitate fusion detection, it has become clear that fusion transcripts themselves are common and may result as passenger events of structural rearrangements. With larger studies and tens of thousands fusions detected, the functional assessment of fusions becomes more difficult. Many of the above cited papers are mainly descriptive, but often it remains unresolved whether fusions involving known cancer-related genes have occurred by chance or are indeed functional. A paper in 2017 by Lu and colleages describes a protocol with which they engineered 20 novel fusion

genes into *in vitro* cell models to functionally characterise them (Lu et al., 2017). However, engineering fusion genes into cells to observe their impact on growth, transformative ability and sensitivity to drug inhibition is a time-intensive process, which is unrealistic to conduct on the tens of thousands fusions observed in each large-scale study.

As existing studies tend to focus on fusions involving known cancer-related genes, true oncogenic fusions may have remained hidden needles in the haystack. This thus reveals a need for a method to study the functionality of gene fusions in cancer in an unbiased manner on a large-scale, which is one of the main aims of my PhD thesis.

1.3.4.2 Clinical relevance of gene fusions

The functional assessment of gene fusions is critical for the understanding of cancer biology as well as clinical opportunities. Already, gene fusions have played a major role in the development of precision medicine (Table 1-1).

As mentioned before, the treatment of *BCR-ABL1* fused cancers with Imatinib has become one of the best-cited success stories for treatment based on genetic biomarkers as patient survival chances improved vastly with the new therapy. At the same time, the treatment of *ALK*-fused non-small cell lung cancers with ALK inhibitor crizotinib was FDA-approved after a phase 3 trial showed a 65% response rate (vs. 20% with chemotherapy), an increase in median progression-free survival from 7.7 months vs. 3.0 months and better quality of life (Shaw et al., 2013).

Kinase-fused tumours often show signs of "oncogene addiction", i.e. they require the continuous activity of the kinase for cellular survival. Thus, many other fusion genes make appealing targets for inhibition with compounds that are either existing or currently in development. For example, many ALK inhibitors are also potent inhibitors of ROS1, a closely related tyrosine kinase (Shaw et al., 2013). *ROS1*-rearranged cancers have shown promising initial response rates to crizotinib in an on-going Phase I clinical trial (Ou et al., 2013). In early 2018, the treatment of *ROS1*-positive advanced-stage non-small cell lung cancer has been recommended as per the official guidelines of NICE, the National Institute for Health and Care Excellence which submits guidance to the UK National Health Service (NICE, 2018). Similarly, *BRAF* and *RAF1* fusions have been shown to respond well to inhibition by RAF inhibitors, of which there are several already approved by the FDA for treatment of *BRAF*-mutated melanoma (Palanisamy et al., 2010). Other promising targets

that are currently under clinical investigation include neurotrophic receptor tyrosine kinase (*NTRK*), ret proto-oncogene (*RET*), and *FGFR* fusions (Shaw et al., 2013).

Other than as pharmacological targets, fusion genes may also aid in the diagnosis and prognosis of cancers. For instance, the *KMT2A*-rearranged acute myeloid leukaemias with different fusion partners had vastly different progression-free survival curves ranging from very favourable 92% at 5-years for some subtypes to 16% for others (Balgobind et al., 2009). At the same type, advances in liquid biopsies mean that it may soon become possible to diagnose patients and monitor disease progression from circulating tumour DNA and cells in the blood (Crowley et al., 2013; Krebs et al., 2014). Due to their non-invasive nature, these methods would be well-suited to be used at regular time-points to test for disease recurrence and the emergence of resistance tumour subclones. Already, research has shown that *ALK*-fusions in non-small cell lung cancer could be successfully identified in patient populations using circulating tumour cell technology (Pailler et al., 2013).

Overall, improvements in the clinical care of cancer patients, be it in the diagnosis, prognosis, monitoring or treatment, relies on the continuous discovery and improved understanding of cancer genetics. Therefore, being able to separate true cancer-related functional gene fusions from non-functional fusion transcript may reveal currently unknown opportunities in this area.

Table 1-1: Summary table of selected clinically relevant fusions. CML = chronic myelogenous leukaemia. NSCLC = non-small cell lung cancer.

Fusion	**Predominant cancer types**	**Clinical characteristics**
BCR-ABL1	CML	Treated with imatinib and other ABL1-targeting drugs
EML4-ALK	NSCLC	Treated with crizotinib and other ALK-targeting drugs
KMT2A fusions	Haematological malignancies	Favourable/unfavourable prognosis depending on fusion partner
NPM1-ALK	Anaplastic large cell lymphoma	In clinical trials for ALK-targeting drugs
ROS1 fusions	NSCLC	In clinical trials for ALK/ROS1-targeting drugs
RAF1/BRAF fusions	Various solid carcinomas	In clinical trials for RAF inhibitors.
NTRK fusions	Various solid carcinomas and sarcomas	In clinical trials for larotrectinib (TRK inhibitor)
RET fusions	Papillary thyroid cancer, NSCLC	In clinical trials for cabozantinib (RET inhibitor)
FGFR1/2/3 fusions	Various solid carcinomas	In clinical trials for various *FGFR* inhibitors

1.4 Cell lines as experimental models of cancer

1.4.1 Overview of different types of cancer cell models

Cancer research relies heavily on experimental models, as the ability to study cancers in patients is limited to biopsies and treatment outcome statistics. In general, cancer models are divided between *in vivo*, i.e. animal models, and *in vitro*, i.e. cells cultivated in plastic dishes.

In vivo models are most often mice. Tumours can generally either be induced, e.g. with the use of carcinogens, irradiation or genetic manipulation, or are transplanted, which involves the injection of tumour-derived tissue or cell lines into the animal. Mouse models are widely regarded to most accurately represent the human disease in the lab, since they provide a complex microenvironment and extra-cellular matrix interactions. An exception is the lack of immune capabilities in immunosuppressed mice. Genetically engineered mouse models are also a powerful method of modelling the consequences of genetic modifications on a whole living system. However, *in vivo* models are time consuming, expensive, technically challenging to maintain and there are ethical issues around using live animals. This means that *in vivo* experiments are mainly recommended for relatively advanced stage of research or where there are no other viable alternatives.

Conversely, *in vitro* models, in particular cancer cell lines, are widely accessible and have fewer ethical implications[1]. Immortalised cancer cell lines give an essentially unlimited supply of cells with a highly similar genotype and phenotype.

There are some concerns that genomic instability, culture conditions and contamination of cell lines can limit how representative cancer cell lines are of the original disease. However, most of these issues can be avoided with proper quality control. For example, regular mycoplasma testing and DNA barcoding of cell lines prevent contamination and mix-ups. Keeping the cells in passage for no longer than 3 months also minimises genetic drift.

Human cancer cell lines transplanted into immunodeficient mice have the ability to generate tumours, and the histopathology has been shown to faithfully correlate with that of the tumour of origin (Fogh et al., 1977). Gene expression patterns of cell lines also cluster

[1] Note: in this thesis, by "cancer cell lines" I always refer to human cancer cell lines, as opposed to cell lines derived from other model animals.

according to their tissue of origin, which indicates that culture conditions are not sufficient to overwrite the tissue-specific gene expression program (Ross et al., 2000). The genotypes of cancer cell line panels are also largely representative of those of the original cancer type. The alteration frequencies of the most common gene alterations in tumours correlates significantly with that of cell lines. Similarly, there is a high concordance in terms of the classification by alteration subclass, i.e. whether alterations in the tumour type of origin are predominantly mutations, copy number alterations, or hypermethylated regions (Iorio et al., 2016). Altogether, the evidence suggests that with proper quality control, cancer cell lines are a useful and representative tool to study cancer tissues and specific cancer-related genetic alterations.

There are certain caveats to using cancer cell lines. For one, immortal cancer cell lines are difficult to derive and advanced-stage cancers have higher success rates of immortalisation (Masters, 2000). This means that early-stage cancers are likely to be underrepresented in cancer cell line panels. Due to a lack of a complex microenvironment, cancer cell lines also only represent a simplified view of cancer biology.

Better models of cell-cell interaction and microenvironments may be cell line cultures grown in 3D. They better reproduce the differential exposure to oxygen, nutrients and growth factor within tumour masses, which can influence proliferation rate, gene expression and drug uptake (Lin and Chang, 2008).

Even more recently, we have seen the development of cancer organoids, which are derived from tumour cells that expand into a 3D structure that closely resembles the tissue architecture, cell type composition and self-renewal dynamics (Sato et al., 2009). They have been found to recapitulate the genetic background of tissue donors and to be amenable to high-throughput drug screening, which opens up future possibilities of ever-closer personalisation of cancer treatments (van de Wetering et al., 2015).

While the use of 3D *in vitro* cultures as experimental models of cancer is likely to become increasingly important in future pre-clinical studies, for now standard protocols for the derivation, maintenance, genetic manipulation and high-throughput screening are still being written. Thus, 2D cancer cell lines are still favoured by many as pliable models for novel techniques and discoveries, such as high-throughput drug screening, CRISPR/Cas9 whole genome drop-out screening and more.

In the next section, I will provide a brief overview of human cancer cell line panels that have featured prominently in modern cancer research, their origin and large-scale data sources that are available and being generated today.

1.4.2 NCI60: The first drug screening panel

The first large cancer cell line panel was established in the 1980's by the National Cancer Institute in the USA, in the form of the NCI60. It consisted of 60 cancer cell lines of 9 distinct tissue types that were characterised by karyotype banding to exclude cell lines that were derivatives of HeLa or one another (Shoemaker, 2006). The purpose of this panel was to act as the first *in vitro* screening platform for cancer compounds. Before that, anti-cancer compounds were typically screened against leukaemia mouse xenografts, which were unreliable at predicting growth response in solid tumours (Weinstein et al., 1997). The NCI60 on the other hand were faster to screen and could act as the first stage of a drug discovery pipeline to identify the activity of potential anti-cancer compounds. From its establishment in the 1980's to 2000, more than 300,000 compounds were screened across this panel (Shoemaker, 2006). The NCI60 panel has since been deployed to identify the mechanism of action of potent unknown natural compounds and discovery of novel anti-cancer compounds. For instance, bortezomib is now a FDA-approved chemotherapy compound and the first proteasome-inhibitor to be used in humans (Adams et al., 1999). Later, molecular characterisation of the cell lines was integrated, which enabled further important biological discoveries, for example that overexpression of ABC family membrane transporter could confer resistance to specific compounds (Szakács et al., 2004). The sensitivity of *BRAF*(V600E)-positive cell lines to MEK inhibitors was also discovered using the NCI60 cell lines panel in 2006 (Solit et al., 2006).

1.4.3 Expanding the number of cell lines and genomic biomarker analyses

In March 2012, two papers were published in the same issue of Nature, both describing drug screening pipelines in even larger panels of cancer cell lines. The first paper described the Genomics of Drug Sensitivity in Cancer (GDSC) cell lines panel at the Wellcome Sanger Institute in the UK and the other the Cancer Cell Line Encyclopedia (CCLE) at the Broad Institute in the USA (Barretina et al., 2012; Garnett et al., 2012). Unlike the NCI60, these two studies drug screened an excess of 600 and 400 cell lines respectively. With an increased number of cell lines, the newer panels are able to represent a larger

number of tissue types and genomic variety, overall and within tissue types. By annotating the cell lines for mutations in cancer-related genes, copy number alterations (CNAs), and gene expression, the studies provided the statistical power for the unguided identification of genomic biomarkers of drug sensitivity. Both cell line panels were able to replicate many known biomarker-drug associations, e.g. *BRAF*-mutations with BRAF/MEK inhibitors, *ERBB2*-mutations with EGFR/ERBB2 mutations and others. Additional, the GDSC panel was able to identify a novel sensitivity of Ewing's sarcoma cell lines with *EWSR1-FLI1* gene translocations to PARP inhibitors, while the CCLE panel found that an aryl hydrocarbon receptor (*AHR*) gene expression signature in *NRAS*-mutated cell lines correlated with sensitivity to MEK inhibitors.

To-date, both the GDSC and CCLE have expanded the number of models further, with now over 1,000 overlapping cancer cell lines screened and annotated by both panels. With improvements in technology, it has also become easier, cheaper and faster to sequence cell lines for more detailed genomic annotations, perform high-throughput drug screening at larger scale and to genetically engineer cell lines for a richer data set.

Next, I will provide a brief overview on technologies that produced these data-sets. At the Wellcome Sanger Institute, my work is based on the GDSC cell line panel, and therefore they will be my main focus in the next few paragraphs.

1.4.4 Genomic characterisation of cell line panels

As sequencing technologies have become cheaper, the genomic characterisation of cell lines has expanded accordingly. For the original paper in 2012, the authors sequenced the full coding exons of 64 commonly mutated cancer genes in about 600 cell lines. Now, the GDSC cell line panel has whole exome sequencing data and single-nucleotide polymorphism (SNP) called for over 1,000 cell lines (Iorio et al., 2016). This also includes targeted PCR sequencing or FISH analysis for three selected translocations (*BCR-ABL1*, *EWSR1-FLI1* and *EWSR1-ERG*). The molecular annotation was expanded to include hypermethylated regions, as called using Infinium HumanMethylation450 BeadChip arrays. RNA-sequencing data now also provides gene expression data for the entire transcriptome, rather than selected genes from microarray expression data. An overview of the data types available for each cell line is given in Figure 1.5.

Analysing too many potential genomic alterations dilutes the statistical power for biomarker analysis and also likely contains alterations that are only found in cancer cell

lines but not patient tumours. To identify patient-relevant genomic alterations, Iorio and colleagues processed genomic annotation available from over 6,000 human tumours and statistically identified alterations that are likely to be cancer-related. This involved the use of algorithms that evaluated the i) recurrence and functional impact of SNPs (MutSigCV, OncodriveFM and Oncodrive-CLUST) ii) recurrence and size of CNA regions (ADMIRE), and iii) methylated regions that show multi-model distribution in at least one cancer type (Iorio et al., 2016). This produced a reference set of clinically-relevant cancer functional events (CFEs) and only alterations that matched this list were considered for further downstream analyses.

The authors then showed that panels of cell lines captured the genomic variety of tumours on a tissue type level, since 1) the percentage of cell lines altered for specific cancer-associated pathways and 2) the class of the dominant genomic alterations within tissue types both correlated well. Additionally, out of 1,273 CFEs identified in the tumour samples, 86% (n = 1,063) occurred in at least one cell line. There are the caveats that this is lower for CFEs that are present in < 5% of patient samples and only around 79% of CFEs are found in at least three cell lines (Iorio et al., 2016). Biomarker analyses often require at least 2-3 samples with a specific genotype to have adequate statistical power to separate signal from noise. This thus suggests that expanding the *in vitro* models collection beyond the existing 1,011 models has the potential to reveal further trends in rare cancer subtypes.

1.4.5 High-throughput drug screening and identification of biomarkers of drug response

The original NCI60 platform screened 100,000's of compounds across 60 human cancer cell lines, which enabled researchers to query the general anti-cancer activity of compounds (Shoemaker, 2006). Conversely, the GDSC and CCLE studies screen fewer compounds but across many more cell lines, which provided the resolution to identify rarer cancer subtypes that are particularly sensitive to certain compounds. Screening fewer compounds does mean that they are typically selected more carefully, e.g. only testing those that have already shown anti-tumour activity in smaller sets of cell lines. Nevertheless, technological advances have enabled a rapid growth in the number of compounds and cell lines screened is rapidly growing. From 2012 to 2016, the number of compounds/cell lines screened increased from 130/639 to 265/990 (Garnett et al., 2012). My thesis will include drug sensitivity data for > 400 drug compounds.

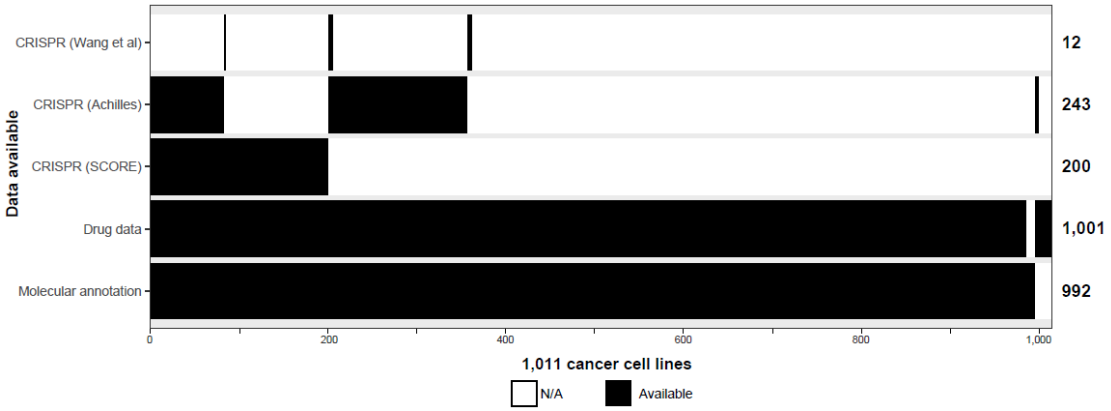


Figure 1.5: Summary of the data types available for the GDSC panel of 1,011 cancer cell lines used for my thesis. RNA-Seq data is available for all cell lines. CRISPR/Cas9 whole genome drop-out screening data is available from multiple sources (see chapter 5 for details and references). Number at the right represents the number of cell lines a given data type is available for.

To briefly describe the protocol for cell line screening: cell lines are grown in multi-well plates for 4 days. 24 hours after plating, cell lines are exposed to a certain concentration of the drug compound, titrated with a dilution series. Viability is tested 72 hours after exposure using fluorescent chemical assays that measure the nucleic acid content (Syto60), or in later studies, metabolic activity (CellTitre Glo). Viability is normalised to negative control (DMSO) and positive control wells (media only). The percentage viability from each experimental well is then fitted onto a viability curve from which an inhibitory concentration at which viability is 50% (IC_{50}) and an area under the curve (AUC) can be calculated. Lower values for both IC_{50} and AUC mean that a lower concentration of a drug can elicit the same loss of viability and therefore represents a sensitivity to a given drug.

Across the years, developments in technology have increased the pace of screening. Notably, while 96-well screening plates were used for the NCI60 and the 2012 GDSC screen, the 2016 GDSC screen used 384-well plates. In 2018, the 1536-well plate is now standard in our group, effectively increasing the screening throughput 16-fold. Robotic platforms have also been invaluable in reducing the amount of manual handling needed. Tasks for which a person requires several hours can now be carried out in the span of minutes, e.g. evenly dispensing cell-line containing media evenly across screening plates, drug titration, and pipetting nano litres of drug directly into screening wells. To-date, the platform is powerful enough to screen 900 new compounds across a core-set of 750 cell lines per year.

For the identification of biomarkers of drug sensitivity, analyses statistically test whether the IC_{50}'s, i.e. drug sensitivity, for cell lines of a given genotype are significantly different from all other cell lines. For instance, cell lines with a *BCR-ABL1* rearrangement on average have a much lower IC_{50} when exposed to Imatinib than the bulk of cell lines (see chapter 4.3.1.2). The analysis of biomarkers in the GDSC panel is typically performed using a multivariate analysis of variance model (MANOVA), which tests the association of each CFE to each drug and also controls for tissue type and MSI status of cell lines as co-factors (Garnett et al., 2012; Iorio et al., 2016). Other computational algorithms have been developed that use variants of machine learning to deconvolute associations of combined features (e.g. *RAS* OR *RAF* mutations both lead to MEK inhibitor sensitivity). However, the significant associations returned by these models are often combinations of the gene expression of large sets of genes, which can be difficult to interpret for biological and clinical relevance.

So far, the GDSC high-throughput drug-screens have shown that *in vitro* systems can recapitulate many known associations (e.g. with genetic alterations involving *BRAF*, *EGFR*, *ABL*, tumor protein p53 (*TP53*), fms related tyrosine kinase 3 (*FLT3*), etc.). In the 2016 publication, Iorio and colleagues also report > 100 novel associations, though the meaning and biological significance of many of these is yet to be understood (Iorio et al., 2016).

1.4.6 Current developments and future outlook

As the above paragraphs have hopefully shown, high-throughput drug screens in genomically annotated cancer cell line panels are a valuable tool for the unguided exploration of cancer biology and treatment opportunities. Implementing novel techniques in the existing pipeline can further broaden the opportunities for discovery.

For one, our group and others have deployed CRISPR/Cas9 whole genome drop-out screens in cancer cell line panels (Behan et al., 2018; Meyers et al., 2017; Wang et al., 2017). The CRISPR/Cas9 system was originally discovered in bacteria and is now a widely-used technique for creating gene knock-outs in mammalian cells (Jinek et al., 2012; Mali et al., 2013). In the drop-out screens, a short guide RNA (sgRNA) library targeting about 17,000-20,000 different genes is transduced into Cas9-expressing cell lines. Those sgRNA that induce a loss-of-fitness phenotype will be statistically depleted after several weeks in culture. Compared to drug compounds, these screens can reveal a much broader spectrum of genetic dependencies of cancer cells, as they have two major advantages: 1) the sgRNA

can be designed to have low off-target activity and 2) they can essentially target any, even undruggable, gene in the genome.

CRISPR/Cas9 technologies are also evolving beyond just knock-outs, as more complex phenotypes such as double knock-outs and transcriptional activation are being achieved. This further opens up possibilities in querying the biology of cancer in *in vitro* systems. In chapters 5 and 6, I will show how CRISPR/Cas9 whole genome drop-out screens data can be leveraged to examine the functionality of gene fusions in the context of my thesis.

Secondly, finding potent drug combinations is another area of interest for cell line screening. As illustrated in the example of *BRAF*(V600E) mutated colorectal cancer (see section 1.2.2), well selected drug combinations can overcome innate and acquired drug resistances (Al-Lazikani et al., 2012). The sheer number of combinations that can be created using currently available drugs is staggering, especially when considering triple and even quadruple combinations. Nonetheless, high-throughput cell line panels can be a first step in the search for drug combinations. In our group, we have so far screened 1,800 unique combinations in a subset of the cancer cell line panel. Cancer cell line co-dependencies may also be identified in knock-out screens. For example an RNA interference screen found that alterations of *PI3K*/*PTEN* may confer resistance to anti-HER2 antibody trastuzumab in breast cancers (Berns et al., 2007).

Finally, the dominant type of *in vitro* model used in large-scale screens may also see some changes in the near future. High-throughput drug screening protocols are currently being developed for cancer organoids in our lab and others. Similarly, improvements in technology have made it possible to derive experimental models such as patient derived tumour organoids in the span of several weeks. This may provide future opportunities for direct feedback from the bench back into the clinic (Pauli et al., 2017).

Altogether, while high-throughput drug screening has historically been a tool for the discovery of new drugs and biomarkers, new developments may allow it to soon become part of a more integrated personalised medicine programme for patients in real-time.

1.5 Final introductory words

In the past chapter, I have reviewed the concepts of precision cancer medicine, the role of gene fusions as cancer oncogenes and large cell line panels as experimental models for cancer research.

In summary, it is clear that precision medicine has had an overall positive impact on the treatment opportunities open to eligible patients today. At the same time, not all cancer patients are diagnosed an actionable biomarker, and many treated tumours develop secondary resistances, which underscores a persisting need for novel research.

Gene fusions have so far been shown to be important cancer oncogenes as well as excellent markers of diagnosis, prognosis and drug sensitivity. Thanks to new technologies such as RNA-Seq that have enabled the unguided large-scale discovery of gene fusions across thousands of samples, the number of known gene fusions has been growing exponentially. Now, while the identification of gene fusions was once a limiting factor, it is the functional annotation that presents a challenge.

Large cell line panels in the past have been valuable models for anti-cancer compound discovery and genetic biomarker analysis. Over the next few chapters in my thesis, I would like to examine the hypothesis that they can be used to assist in identifying the functional relevance of oncogenic fusions.

My thesis has the following structure to address the question:

1) Create a filtering pipeline to annotate fusions in 1,011 human cancer cell lines.
2) Brief description of the landscape of fusions as we find them in the cell lines.
3) Using extensive drug-screening data, I examine the use of gene fusions as biomarkers of drug response.
4) Set-up and explore the results of an unguided computational approach to examine the functionality of individual gene fusions using CRISPR/Cas9 whole-genome drop-out screening data.

With that, I hope that the following pages represent a small contribution to the grand challenge that is the continuous improvement in the well-being of future patients of a terrible disease.

2 Filtering gene fusions in 1,011 cancer cell lines

The first step required in the study of gene fusions in cancer cell lines is to create a reliable catalogue of all fusion transcripts found in the same cell lines.

As discussed in the introduction chapter, for an unguided fusion transcript discovery in large number of samples, RNA-Seq analysis followed by the computational devolution of fusion transcripts is the best option. At the same time, fusion calling algorithms are highly error-prone, both due to sequencing artefacts and variable methods used by fusion-calling algorithms.

For this reason, our fusion-calling pipeline uses three different fusion-calling algorithms to analyse our RNA-Seq data. These algorithms are DeFuse, Tophat Fusion and STAR. The reasoning behind the decision was that the fusion-calling algorithms could either validate or complement each other.

Thus, the first aim of my thesis is to examine the output produced by the Wellcome Sanger Institute's Cancer, Aging and Somatic Mutations (CASM) program's fusion-calling pipeline and to put together a set of filtering criteria that would produce a catalogue of high-confidence gene fusions in our 1,011 cancer cell lines. In this chapter, I first introduce the data available, the pre-processing steps and the three filtering algorithms which were used in our analysis. I then show two ways of benchmarking the fusion transcripts called by our fusion-calling pipeline. Based on these, I evaluate how effectively proposed filters retain true positive fusion transcripts while reducing false positives.

2.1 Initial data processing and overview of fusion algorithms

2.1.1 Note on terminology

Throughout this thesis, the term "fusion" refers to any joining of two specific partner genes, with the first-named gene being the 5' and the second-named gene being the 3' partner. The term "fusion event" refers to the presence of a given fusion in a specific cell line. "Fusion transcript" refers to a fusion event with a specific breakpoint. Thus, multiple fusion transcripts of a single fusion event may occur in a given cell line.

2.1.2 1,011 cancer cell lines

The panel of 1,011 human cancer cell lines have been collected from publicly available repositories and private collections and are curated by the Genomics of Drug Sensitivity in Cancer (GDSC) project team at the Wellcome Sanger Institute. They span 42

different cancer types in 29 different tissue classifications (Figure 2.1; Supplementary Table 1).

The cells lines have previously been extensively characterised via whole exome sequencing (EGAD00001001039) and Affymetrix SNP6-based copy number analysis and genotyping (EGAD00010000644). To identify the cancer-relevant mutations, copy number-altered and hypermethylated regions, each cell line is also annotated for a set of cancer functional events (CFE's), which represent genomic alterations that are recurrently altered in TCGA tumour samples (Iorio et al., 2016).

Further, high-throughput drug-screening data for 997 cell lines is available for more than 400 drug compounds. CRISPR/Cas9 screening data is available through several in-house and publically available resources for 371 cell lines.

These data resources will be leverages in chapters 3 and 4 to query the functional relevance of the identified fusion transcripts.

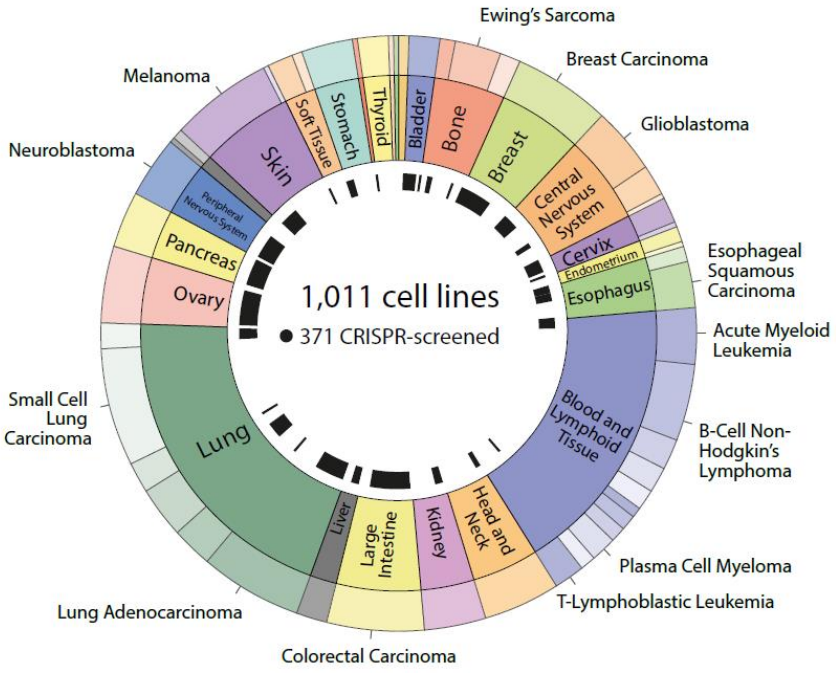


Figure 2.1: Overview of tissue types of 1,011 cell lines. Inner ring shows the coverage of CRISPR-screening data.

2.1.3 Data pre-processing

Prior to my joining the project, RNA-Seq data was generated in-house for 447 cell lines and downloaded from CGHub for 587 cell lines. This data set includes 1,011 unique cell lines. For 23 of those, we obtained RNA-Seq data from both sources. The cell lines had

a median of 77.6M read pairs (range: 40M-175M) in the CGHub data (median: 36X; range: 7X-83X) and 227M read pairs (range: 79M-543M) in the in-house data (median: 89X; range: 16X-206X).

Fusion calling was performed using the cgpRna pipeline developed by the Cancer Genome Project at the Wellcome Trust Sanger Institute (Cancer IT, 2018a). RNA-Seq reads were mapped onto the GRCh38 human reference genome (Ensembl version 82 gene model) via three independently developed pipelines using their default options: STAR, TopHat-Fusion and deFuse (Dobin et al., 2013; Kim and Salzberg, 2011; McPherson et al., 2011). The outputs of all three fusion calling algorithms are compiled into a single list and the pipeline then employs the Gene Rearrangement AnalySiS (GRASS) algorithm to predict consequences of gene fusions, including whether fusions are in coding or non-coding regions and their frames (Cancer IT, 2018b).

The cgpRna pipeline automatically combines fusions identified by multiple pipelines into a single entry. After splitting the entries into individual fusion calls, we found that the cgpRna pipeline yielded 1,934,711 putative fusions transcripts across the 1,011 cell lines (Table 2-1). Overall, fusion calls by STAR made up the bulk of putative transcripts (N = 1,576,062), with TopHat-Fusion calling the smallest number of putative transcripts identified by a single algorithm (N = 58,397). Only 14,118 (0.9%) of putative transcripts were called by more than one algorithm.

2.1.4 Fusion-calling algorithms

The basic principles of fusion-calling algorithms is already described in the introductory chapter in section 1.3.2. Here, I would like to briefly outline algorithm-specific characteristics and the results from previous attempts at benchmarking the algorithms.

Table 2-1: Number of putative fusions called by individual or multiple algorithms. (D = DeFuse, T = TopHat-Fusion, S = STAR-Fusion.)

Algorithm (unique)	Number of Fusions	Algorithm (shared)	Number of Fusions
D	286,134	DTS	5,959
T	58,397	DT	1,770
S	1,576,062	DS	2,856
		TS	3,533

2.1.4.1 DeFuse

This fusion-caller was created in 2009 by McPherson and colleagues at a time when few alternatives existed (McPherson et al., 2011). It was the first aligner to take into account multiple alignments for a single read pair, while other aligners typically either discard any read pair with multiple alignments, or choose the alignment with the highest alignment score. DeFuse does this by applying a so-called "maximum parsimony" approach, which picks the alignment which supports a putative fusion transcript which has the highest support from other read alignments.

The deFuse algorithm was tested in 44 samples (40 ovarian cancer patient samples, 3 sarcoma patient samples and 1 ovarian cancer cell line), which mainly contained 17-45 M read pairs. The algorithm detected 20,237 candidate fusion transcripts, i.e. a mean of 460 per sample.

The authors then designed 11 additional filters which characterise the fusions and perform statistical tests to evaluate the distribution of reads across a predicted breakpoint, the most likely distribution of fragment lengths and potential confounding factors, e.g. homologous nucleotides at the fusion boundaries. In total, 268 fusion transcripts (1.3%) passed these additional filters.

In order to validate the efficacy of these filters, the authors then performed PCR validation for selected fusion transcripts. Of 46 fusions that pass all additional filters, 42 validated. 14 that failed at least one of the additional filters also failed the PCR validations. Of 40 randomly selected fusion transcripts, only one validated using PCR.

While the additional filters appear to perform well at selecting true positive fusions, some known positive fusions did not pass. The authors also acknowledge that the lack of reliable positive and negative controls still represent a challenge when designing fusion-calling algorithms. Similarly, as the filters were trained on this dataset, it is unclear how they will perform in independent datasets.

2.1.4.2 TopHat-Fusion

This algorithm was developed by Kim and Salzberg in 2011 as an algorithm that only uses gene annotations to identify known intron boundaries at a later step, which speeds up the analysis significantly (Kim and Salzberg, 2011).

To do this, TopHat first identifies any reads that do not map entirely within exons and then splits them into 25 bp segments which are aligned independently. Fusions are

separated from introns by whether or not two segments align at a genomic distance of more than 100,000 bp apart. TopHat then stitches reads together to reconstruct the full 600 bp window around the putative fusion breakpoint and filters out those that show any gaps in this window.

An advantage over algorithms that only recognises fusions from reads mapping to two different annotated genes, TopHat-Fusion can derive fusion products from known genes, unknown genes and unannotated splice variants.

The authors tested their algorithm on four breast cancer cell lines and one prostate cancer cell line, which contained 7-21 M read pairs. Compared with previous analyses of the same data, TopHat-Fusion found 44 of 47 known fusion events in the same cell lines and an additional 61 novel fusion events. It is important to note that individual splitting and spanning read thresholds were implemented for each sample based on the *a priori* known fusions, which could enrich for true positives while providing little guidance for setting thresholds for samples that were not previously tested.

The authors estimated their false positive rates by hypothesising that non-neoplastic tissue should have no fusion transcripts. In two normal tissue samples, they found 10 fusion events, albeit with more stringent supporting read thresholds that the authors devised due to a higher number of total read pairs in those samples.

The authors also compared the output of their algorithm to that of deFuse for two cell lines. There, they found that both programs found seven out of nine previously reported fusion events, but with TopHat-Fusion predicting 42 and deFuse 1,670 fusions events in total.

Thus, TopHat-Fusion's algorithm predicts many of the expected fusions in the cell lines that they examined, and a lower number of unexpected events compared to deFuse. However, the authors still faced the same challenge as the authors of deFuse faced: namely that the lack of gold-standard true-positives and true-negatives makes it difficult to predict the likelihood of the unexpected fusion-calls to represent true fusion events.

2.1.4.3 STAR

This fusion-caller is an in-built function of STAR aligner, an ultra-fast aligner for RNA-Seq data on reference transcriptomes (Dobin et al., 2013; Haas et al., 2017). The chimeric read detection algorithm in STAR follows the following basic steps of 1) creating "local genomic windows" by stitching together alignments that align perfectly to the

reference transcriptome, 2) identify reads that align partially to these local genomic windows and 3) aligning the overhang of partially aligning reads to other local genomic windows. These partially aligning reads are thus splitting reads that support the presence of fusion transcripts.

At the time of release, the built-in chimeric read detection algorithm of STAR was only a small component of overall tool and was minimally benchmarked. In 2017 the authors of STAR released a bioRxiv pre-print for STAR-Fusion, an additional stand-alone tool that uses the output of the chimeric sequence detection programme in STAR and adds an additional set of criteria to further dissect the results. These include filters that discard fusion transcripts that do not have sense-sense orientation of the genes, are not supported by at least 1 splitting read and 2 total reads, do not have a fully aligned genomic window of 23 bp either side of the breakpoint, isoforms with low read evidence, fused genes with high sequence homology and genes that are commonly fused across the same sample.

Importantly, with the publication of STAR-Fusion, the authors also perform a comprehensive comparison of the performance of 15 other fusion-calling algorithms, including deFuse and TopHat-Fusion, which highlight some of the variations between fusion-calling algorithms.

2.1.4.4 Variations in accuracy of fusion-calling algorithms

In the previous section I have outlined three different fusion-calling tools that are available for RNA-Seq data. The exact methodology, as well as the filters applied have a large impact on the final list of fusions. For instance TopHat-Fusion's filters cut down the fusions called in 6 data sets from > 10,000 to several hundred. Similarly, only about 1% of fusions passed deFuse's additional filters.

Several papers have now compared the performances of over 20 fusion-calling algorithms, the most notable by Carrara *et al* in 2013, Kumar *et al* in 2016 and Haas *et al* in 2017, examining 8, 12 and 16 different algorithms respectively (with some overlap) (Carrara et al., 2013; Haas et al., 2017; Kumar et al., 2016).

In general, all three papers used a combination of 1) simulated RNA-Seq data in which fusion transcripts were planted across different expression levels and 2) real RNA-Seq data, for which the benchmark was either based on previously detected fusions in these samples (Carrara and Kumar), or the concordance across different algorithms (Haas).

All three reviews found large variations both in the sensitivity and specificity of the algorithms in the simulated data, as well as the total number of fusions in the real RNA-Seq data sets. In the simulated data, fusion-calling algorithms have varying levels of success in extracting true positives (sensitivities ranging from 4% to 88% in assays by Kumar and Carrara) but with generally low false positives rates (10-20%).

Real data sets are much messier to interpret. For instance, in the same sample, Kumar *et al.* found that different fusion calling algorithms identified anywhere from 0-10 to over 5,000 fusion transcripts. Carrara *et al.* extracted 11-73,000 different fusion transcripts in the same sample using different algorithms. Moreover the analysis sometimes had highly variable findings, while deFuse and TopHat-Fusion found on average 34 and 3 fusions per sample under Kumar, the number was much higher at 454 and 43,301 under Carrara. These numbers stand in further contrast to the study by Kim and Salzberg, where deFuse identified 32x more fusions than TopHat (Kim and Salzberg, 2011). Furthermore, while EricScript had good sensitivity and specificity under Kumar *et al.*, but had underwhelming performance in Haas *et al.*'s analysis. Haas *et al.* also highlighted discrepancies in the performances of the same fusion-calling algorithms in the simulated and the real data sets. For instance, nFuse and ChimeraScan perform well in the simulated sets, but had much lower accuracy than other calling algorithms in the real data set.

Overall, the unpredictability of RNA-Seq fusion-calling algorithms is important to consider when assessing the choice of method. No one tool emerges as being consistently reliable at calling expected fusions without introducing large numbers of false positives. Similarly, different interpretations of reliability of the same algorithms across different papers is likely a result of a lack of consensus guidelines on how tools should be best applied across samples with different characteristics (e.g. number and length of reads, etc). In general, these findings supports the use of multiple fusion-calling algorithms in order to minimise unpredictable results from individual algorithms.

2.2 Filtering approach

As outlined in the previous section, algorithms for fusion calling from RNA-seq data suffer from noisy results with a large number of false positives due to sequencing artefacts.

In order to create a catalogue of fusion transcripts in our cancer cell lines, I examined the expected true positive and false positive rates of our putative fusion transcripts. To do this, I utilised three sets of benchmarking data that were generated before I started the work on this thesis.

2.2.1 Benchmarks

We developed three benchmarks that were subsequently used to gauge the effectiveness of individual filters, as well as the quality of our final list of fusions. These consist of: 1) 945 putative transcripts that were validated by PCR, 2) comparisons of fusion events called from cell lines with RNA-Seq data obtained both in-house and from TCGA and 3) fusions described in the same cell lines in previous literature. This following section describes these benchmark datasets

2.2.1.1 PCR validations

945 putative fusions transcripts (798 fusion events) in 23 cell lines were selected for validation through PCR of cDNA libraries using two different sets of primers per transcript[2]. PCR reactions confirmed the presence of 51% fusion transcripts while 49% failed. Of putative fusions predicted by single algorithms, those predicted by deFuse had the highest rate of successful validations (56.2%), while those predicted by STAR had the lowest validation rates (14.4%). Putative fusions predicted by all three fusion calling algorithms

Table 2-2: Validation success across 945 putative fusion transcripts. D = Defuse, T = TopHat Fusion, S = STAR. Combinations of letters denote a combination of multiple algorithms.

Algorithm	Total # fusions (%)	Total PCR reactions	% passed (#)	Adj. pass rate
D	286,134 (14.8)	192	56.2% (108)	8.3%
T	58,397 (3.0)	107	33.6% (36)	1.0%
S	1,576,062 (81.5)	111	14.4% (16)	11.7%
DT	1,770 (0.1)	90	61.1% (55)	0.06%
DS	2,856 (0.2)	107	58.9% (63)	0.09%
ST	3,533 (0.2)	112	48.2% (54)	0.09%
DST	5,959 (0.3)	226	78.8% (178)	0.24%
Total	**1,934,711**	**945**		**21.5%**

[2] PCR validations were planned by Graham Bignell and carried out by Elizabeth Anderson. Both were core members at CASM facilities. For methods, see Picco* & Chen* *et al.* (manuscript under review). See supplementary table 2.

had the highest rates of validations overall (Table 2-2). The set of PCR validations served as guide lines when deciding on thresholds for fusion filters.

Note, that fusion transcripts that were selected for validation are not evenly distributed across different fusion-calling algorithms and combinations thereof. For instance, there is a clear overrepresentation in fusion transcripts that are called by multiple fusion-calling algorithms, which only represent less than 1% of all fusion calls. Similarly, both median and mean splitting reads is higher in the subset of fusion events that are PCR validated, compared to the overall set of fusion calls (median: 11 vs. 1, mean: 60.2 vs. 8.4). This is partially due to fusion transcripts called by multiple algorithms generally having higher numbers of splitting reads.

Since validation rates vary considerably between the calling methods, when approximating the overall PCR validation rate of a set of samples, it is advisable to adjust it to the proportion of fusion transcripts called by a given calling method (e.g. if the proportion of fusions called are: 70% by STAR, 20% by DeFuse, 10% by TopHat, then calculate validation rate as: 0.7*14.4% + 0.2*56.2% + 0.1*33.6%, etc.). Plugging in the actual numbers for the validation rates given in Table 2-2, I find an adjusted PCR validation rate of 21.2% across the pool of 1.9 M fusions. Note also, that this number is likely to still be an overestimate, due to the overrepresentation of fusion transcripts with higher number of splitting reads.

At the same time, it may be important to keep in mind that there may be technical reasons for the failure to validate fusions by PCR, in particular since most fusions do not have positive control samples that can confirm the efficacy of a pair of PCR primers. In fact, out of the 945 fusion transcripts tested, only for 791 (84%) were the two different primer sets in agreement. For the other 154 transcripts, one primer set failed while the other passed. Since negative controls test that a PCR product is not a false positive, it is likely that in these 16% of cases, the failed primer set was in fact a false negative. Thus, while the PCR validation is a useful benchmark for the downstream analysis, this caveat should be born in mind.

2.2.1.2 Comparison of TCGA and in-house RNA-Seq data

In order to measure biological replicability of our pipeline, for 23 cell lines we performed both in-house RNA-Seq and used data from CGHub. Both sequencing data sets were then processed by the same downstream pipeline using cgpRna.

In total, 76,358 fusion transcripts were found in the 46 datasets. Although, the in-house RNA-Seq data was associated with significantly more read pairs in sequencing data (paired t-test: t: -11.1, $p < 1 \times 10^{-10}$; Figure 2.2A), there was no significant difference between the number of putative fusions identified (paired t-test: t = -2.0, p = 0.06; Figure 2.2B). With the unfiltered output, the mean proportion of shared fusion events was 6.7% (SD = 2.2%; median across cell lines = 6.8%) (Figure 2.2B).

The relatively small proportion of overlapping fusion events indicates a highly noisy dataset and suggests that only a minority of the unfiltered putative fusion events are reproducible. Ideally, our filtering process should enrich our list of putative fusion events for those that are shared.

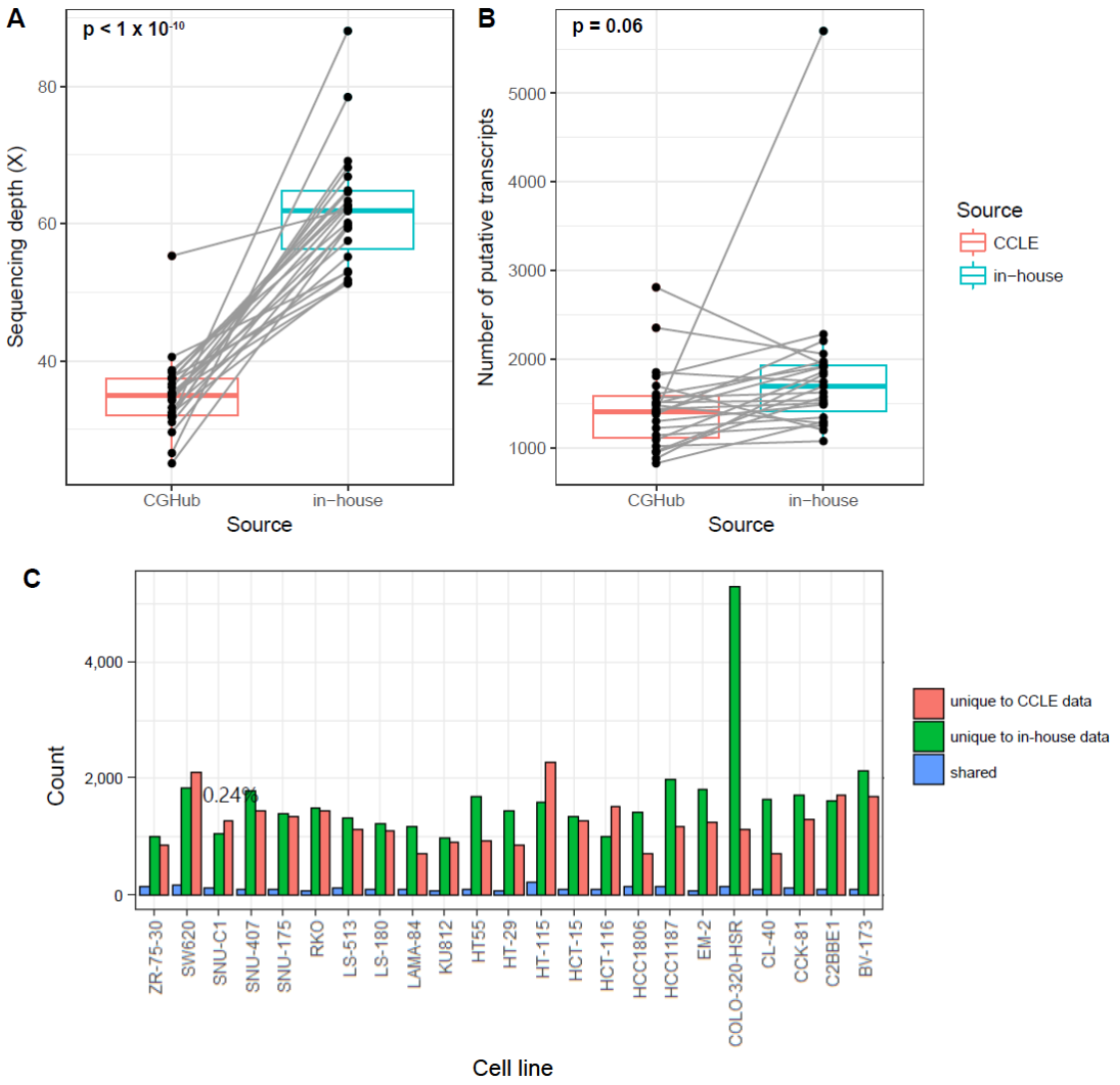


Figure 2.2: Relationship of (A) read pairs for each data set and (B) number of fusions identified. Lines connect the pair of datasets (CCLE or in-house sequencing) for the same cell lines. (C) Overlap of fusions called from two separate RNA-Seq data sources by cell line.

2.2.1.3 *Fusions described in literature*

Finally, a list of previously reported fusions events in the same cell lines that are used in our analysis was manually compiled by Graham Bignell through ChimeraDB, Mitelmann DB, the Atlas of Genetics and Cytogenetics in Oncology and Haematology and TICdb. It includes 1,916 previously reported fusion events across 417 samples from 91 sources.

While some of these fusion events were a result of high-throughput, little-filtered landscape studies, the list also includes 253 annotated fusion events that were validated in either a second study, or through high-confidence methods such as FISH and capillary sequencing. These annotated fusion events include well known oncogenic drivers, for example 20 instances of *EWSR1-FLI1*, 7 instances of *BCR-ABL1* and 1 instance of *TMPRSS2-ERG*.

Due to the inherently high error rate in calling true fusion transcripts as well as incomplete coverage, this list is not necessarily comprehensive or free of false-positives. However, it serves as a helpful tool to compare how many previously reported fusion events are captured using our pipeline. Similarly, whether these previously reported fusion events are preferentially retained by our filtering process can give an indication of the filters' effectiveness in retaining true positives.

Overall, it is important to recognise that all benchmarks cover a subset of the fusion transcripts/cell lines that are examined in my thesis. The PCR validations are conducted on 945 fusions transcripts, which represents less than 0.5 percent of all fusions called within our pipeline. Meanwhile, the analysis of shared fusions between replicates covers 23 cell lines, i.e. 2.3% of our entire panel. Similarly, the fusions described in literature also cover only 1,916 fusion events in 417 samples. While the curation of this benchmarking data is nevertheless a valuable resource, the representation should be born in mind when interpreting the downstream analyses.

2.2.2 Evaluation of suggested set of filters

In the next section, I describe how I evaluated the set of filters that were suggested prior to my participation in this project. For that, I benchmarked the impact of filters using the benchmarks described above, mainly with the aim of removing fusion transcripts and

events that do not pass PCR validations and that are not shared in the independently sequenced datasets.

2.2.3 Split read filter

The split read filter sets the requirement that a putative fusion transcript must be supported by a certain number of sequencing reads that align directly onto the breakpoint and is standard for fusion calling algorithms. In the cgpRna pipeline, only TopHat-Fusion and DeFuse apply such a filter by default, while STAR reports putative fusion transcripts even when supported only by a single read or read pair. This likely explains the large number of fusion transcripts found by STAR (81.5% of total), opposed to other fusion calling algorithms. In fact, out of the 1,588,410 putative fusion transcripts found by STAR, only 102,958 (6.5%) had at least 3 splitting reads.

The assumption underlying the split read threshold is that putative fusion transcripts with little supporting evidence from single sequencing reads may be false-positives resulting from technical artefacts.

First, I examine the total putative fusions retained at a given splitting read threshold, and the number and proportion of passing and failing PCR validation. By increasing the splitting read threshold, the proportion of PCR validating fusions increases substantially (Figure 2.3). Even a basic threshold of 3, which removes 80% of all fusion transcripts, we remove 38.2% of false-positives tested by PCR, while retaining 92.9% of those that do

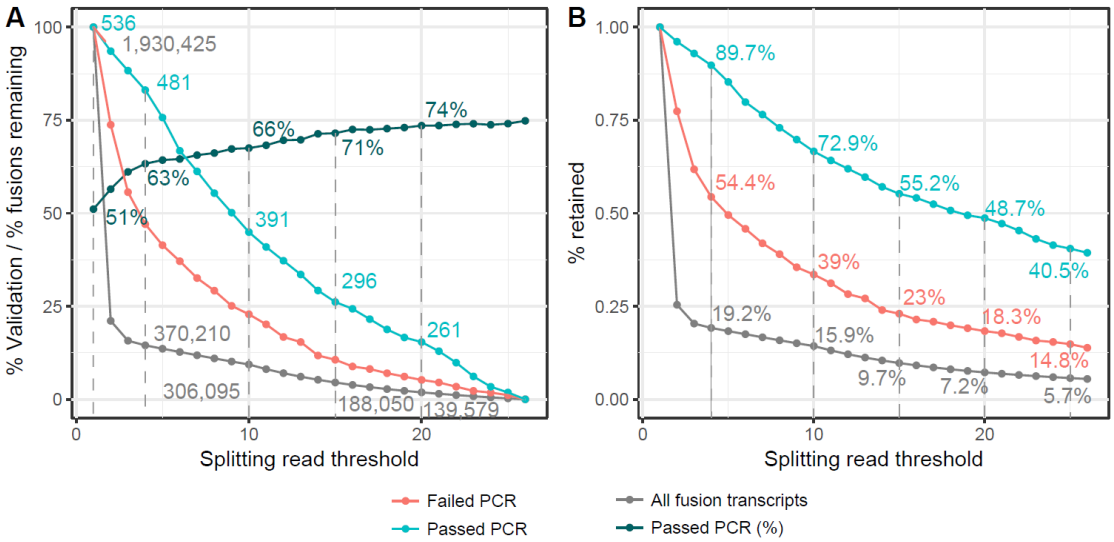


Figure 2.3: Impact of a splitting read filter. At any threshold (x-axis), the plot shows (A) the number and percentage of PCR validated fusion transcripts, as well as the total number of fusion transcripts that pass that threshold. (B) The percentage of fusion transcripts that are retained from a given category (validating, non-validating and total) at each threshold.

validate (Figure 2.3B). On the other end of the spectrum, at very stringent threshold of 25 splitting reads, 74% of putative fusion transcripts validate, but only 40% of validating transcripts are retained.

Even true fusion transcripts can have varying numbers of supporting splitting reads, which can be explained by varying transcript expression levels, i.e. true fusions that are expressed lowly may only have limited number of sequencing reads supporting it.

Since no clear division exist between true positive and false positive fusions, the ideal threshold would likely depend on the study and in particular the read depth of samples. Because my project aims to discover novel fusions, which will be validated in later experiments, I decided to implement a relatively permissive threshold of 4 splitting reads, as it strikes a compromise by removing approximately 45.6% of false-positives while retaining 89.7% of validated putative fusion transcripts in a single filtering step. Overall, this retains 370,210 (19%) of 1,930,425 fusion transcripts.

2.2.4 Multi-algorithm filter

At the time that the cgpRna script was developed, most research articles used only a single fusion-calling algorithm. Like more recent papers that tend to use multiple algorithms to call fusion transcripts (Gao et al., 2018), for this project the decision was made to implement three separate algorithms, with the reasoning that: 1) multiple algorithms may capture fusions that are missed by single algorithms 2) fusions called by multiple algorithms have more streams of supporting evidence and therefore are more likely to be true positives rather than technical artefacts.

To examine if these assumed benefits hold true, I again examined to our PCR validation data, as well as the data on 23 cell lines obtained from different sources.

First, I broke down the 945 PCR validation by which algorithm called a given putative transcript (Figure 2.4A). Notably, none of the algorithms or combinations showed 100% validation rates by PCR. The data for the most part supports the assumption that fusions called by multiple algorithms are more likely to be true positives, as they validate in higher proportion by PCR. Notably, deFuse outperforms STAR and TopHat in terms of true positives. Nonetheless, deFuse's validation rate at 56.2% still suggests that more than two in five fusion calls may represent a false positive.

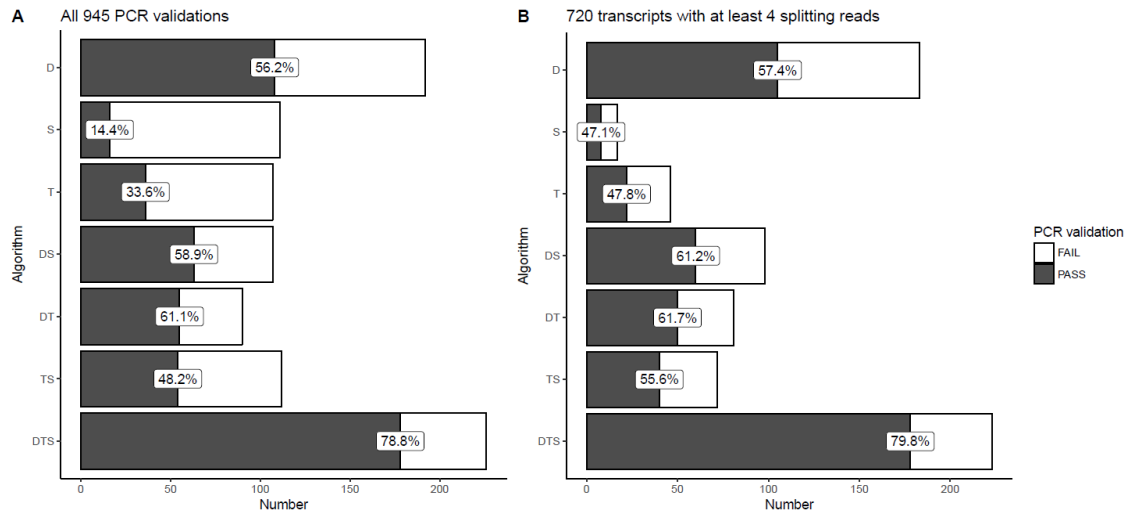


Figure 2.4: Breakdown of 945 PCR validations by algorithm either on the unfiltered list (left), or after applying a basic splitting read threshold of 4 (right).

We hypothesised that the low threshold for splitting reads may have caused the low validation rates particularly for STAR and TopHat fusions. After applying our suggested filter of 4 splitting reads (Figure 2.4B), the proportion of validating fusions per algorithm is increased across almost the entire dataset, but with the most striking improvements in STAR and TopHat-Fusion, where the validation rates rose to 47.1% and 47.8% from 14.4% and 33.6% respectively. Overall, for fusion transcripts called by single algorithms, our validation rate is 39.0% before filtering and 54.9% after filtering by splitting reads. Validation rates across fusion transcripts with multiple supporting algorithms also rose substantially, with the average validation rate for those fusions that are found by more than two algorithms increasing from 65.4% to 69.2%.

Next, I examined the fusion calls obtained by analysing RNA-Seq data from two separate sources in 23 cell lines. In section 2.2.1.2, I showed that when unfiltered, few fusion transcripts overlap between the two sources, which indicates that the data is highly noisy. Figure 2.5 shows that after applying the filters which were identified as effective in the previous sections, we can increase the proportion of shared fusion calls. Unfiltered, the proportion of shared transcripts is merely 6.7%. Implementing the split read filter with a minimum threshold of 4 splitting reads already increases the proportion to 33.7%. When considering only fusions called by multiple algorithms, the proportion of shared transcripts is 64.6%. Finally, when I combine both the split-read filter and the multiple algorithm filter

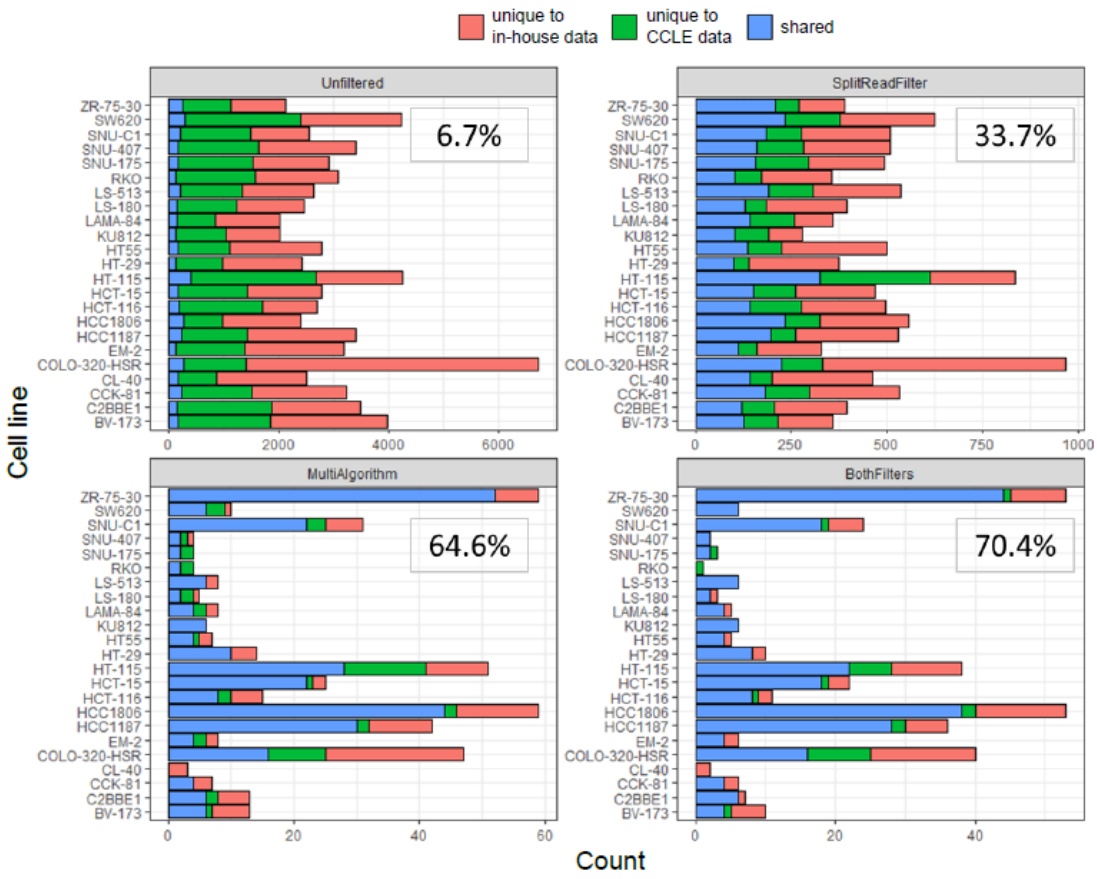


Figure 2.5: Breakdown of fusion calls from the analysis from two separate sources in 23 cell lines. Colour of the bar shows whether a specific putative fusion is called from both sources (Shared: blue), or if it is uniquely called from either the in-house data (red) or the CGHub data (green). Each panel shows the data after a different filtering method was applied. The number in the panel represents the percentage of shared fusions across the data.

together, I observe a much improved proportion of 70.4% shared fusion transcripts. On a per cell line level, the median cell line has an even higher proportion of 77.8% shared fusion transcripts. The overall mean is likely lower due to a few outlier cell lines that have particularly low overlap.

Thus, by limiting our list of putative fusion transcripts to only those that are called by two or more algorithms, I see substantial improvements in proportion of validating fusions from PCR analysis. Perhaps more importantly, the proportion of shared fusions called from biological replicates is more than doubled to 70.4%. In the interest of ensuring that I am working with a biologically replicable dataset, I therefore decided to implement the multi-algorithm filter to obtain our final dataset. It is nonetheless important to keep in mind, that a large amount of fusions that do validate by PCR are removed in the process, which is a caveat that is unfortunately difficult to avoid when analysing fusion calls from RNA-Seq data.

2.2.5 Fusions discovered in normal tissues

Next I examined fusions called in normal tissues, since we believe that true oncogenic fusions are unlikely to be present in non-neoplastic tissues. RNA-Seq data for 245 non-neoplastic tissues was downloaded from the Genotype-Tissue Expression project (GTEx) (GTEx Consortium, 2013) and fed into the cgpRna pipeline. The raw output contained 126,897 putative fusion transcripts. After filtering by number of splitting reads (≥ 4) and removing any putative fusions that were called only be a single algorithms, our dataset retained 78 fusion transcripts. No sample showed more than one fusion transcript and approximately 1 fusion transcript was found per every 3.1 GTEx samples (Figure 2.6A).

Of the 78 fusions found in GTEx samples, 6 fusions were also found in our cancer cell line panel, including the three most common GTEx fusions. Interestingly, one pancreatic sample carries a fusion in *TMPRSS2-ERG*, a well-known pancreatic cancer driver fusion (Figure 2.6B). The presence of this fusion may indicate that the sample originated from a pre-malignant lesion, although recent literature shows that non-neoplastic tissue can show surprisingly high frequencies of cancer driver alterations without leading to obvious oncogenic transformation (Martincorena et al., 2015).

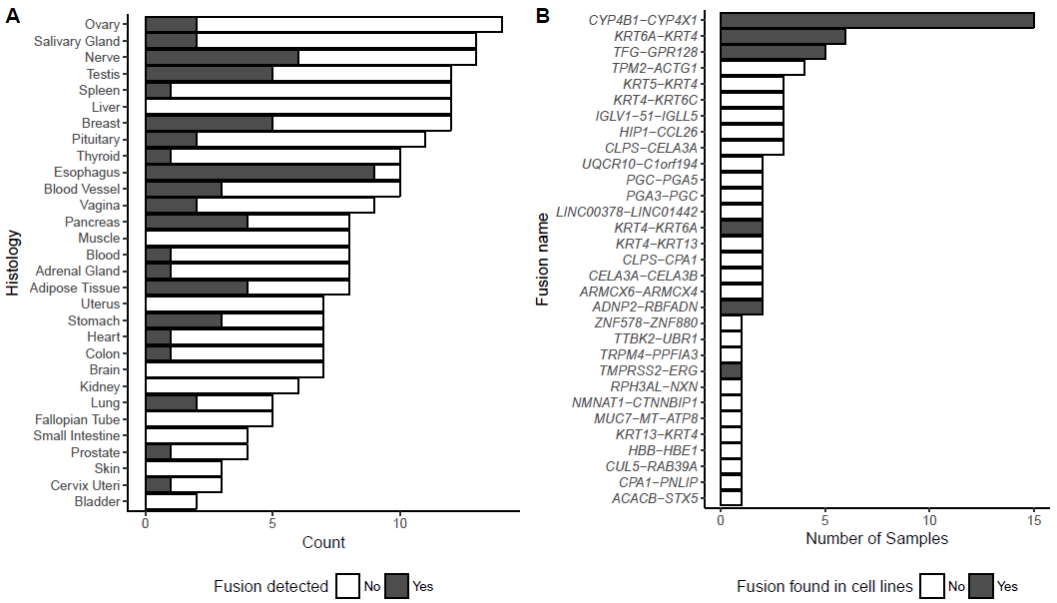


Figure 2.6: Histology and fusions in 245 GTEx samples. (A) We detected fusions in 58 of 245 GTEx samples. (B) 31 unique fusion pairs were detected in GTEx samples. Of those, 6 overlap with fusions found in our cancer cell lines.

With the exception of *TMPRSS2-ERG*, I did not observe any other potentially oncogenic fusion events (i.e. those matching the COSMIC fusion curation) within the GTEx sample results. As fusions that occur recurrently in GTEx samples are unlikely to play significant functional roles, we removed the small number of recurrent GTEx fusions from our list of putative fusions in cancer cell lines (n = 29).

2.2.6 Summary and overview of a framework for fusion filtering

In the above paragraphs, I examined the effectiveness of a set of filters in removing technical artefacts from our list of putative fusion transcripts. Overall, I found that using a splitting read threshold of at least 4 splitting reads per transcript increased the projected true positive rate from 51% to 63% (Figure 2.3). Also, only using fusion calls supported by two or more algorithms significantly increased the biological replicability. Finally, I also removed a small number of fusions that were recurrently found in a set of non-neoplastic GTEx samples.

Note, that during the course of my PhD, I also examined the efficacy of a further 4 filters, that examine false positives filtered by 1) regions of high repeats, 2) ratios of splitting reads to spanning reads, 3) the deFuse additional filters and 4) highly recurrent fusions. No efficacy was found for those filters (keeping in mind the limitations of my benchmarking approach) and for the sake of brevity, the data is not shown.

Thus, for our final filtering pipeline summarised in Figure 2.7, I decided to implement only the splitting read and multi-algorithm filters, as they showed the greatest efficacy in increasing the proportion of true positives.

Implemented across our 1,011 cell lines, of the 1,934,711 putative fusion transcripts outputted by cgpRna, 10,514 (0.5%) pass my filtering process (Supplementary Table 3).

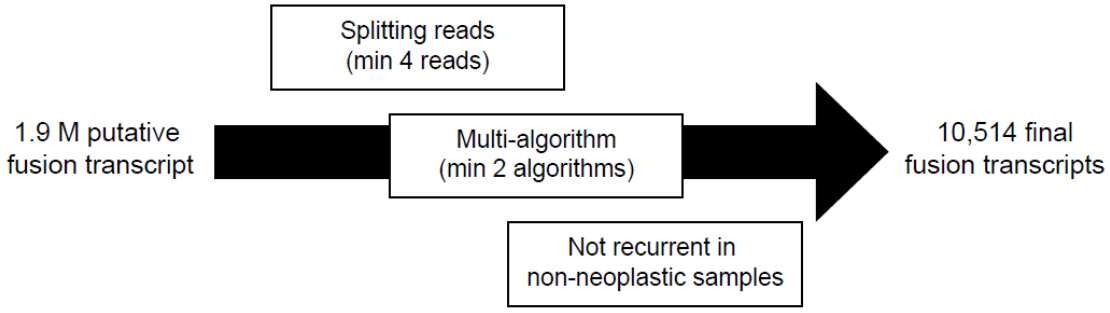


Figure 2.7: Overview of the filtering and annotation pipeline to identify our final catalogue of gene fusions in 1,011 cancer cell lines.

This represents 8,354 fusion events, i.e. unique cell line/fusion combinations. PCR validations performed on a subset of these fusion transcripts (n = 474) yields a projected 69.2% pass rate. In terms of the 23 cell lines that were RNA-Seq'd at two independent sources, applying the two filters substantially improved the proportion of shared fusions from 6.7% to 70.4%.

Finally, I compared our filtered list of fusions with the list of fusion events described in literature (see section 2.2.1.3). Our raw output contained 84.13% of 'validated' and 52.6% of all remaining fusion events described in literature. After filtering, we still matched 74.21% of validated and 39.74% of the remaining fusion events. Although we lost a small number of fusion events that had previously been validated, overall considering that 99.95% of putative fusion events were removed, our filtering strongly enriches our set with previously observed fusion events, i.e. likely true positives.

Although my set of filters provides a vast improvement over the validation rate of our set of fusion transcripts, it is clear that separating true positive from false positive fusions with high reliability remains a challenge. A high validation rate suggests a high specificity of a calling and filtering approach in excluding false negatives. However, in many instances implementing our filtering pipeline also reduces the sensitivity, as fusion transcripts that validate by PCR are filtered out.

At the same time, a limitation of my approach are the benchmarking tools that are available to test filters against. While performing the PCR validations for 945 was without a doubt a big effort, these only represented a fraction of the fusion transcripts called. Similarly, they under-represent certain types of fusion calls which may have created a skew in the analysis. For instance, the vast majority of PCR validations were performed on fusion transcripts that were called with more than 4 splitting reads and there was an overrepresentation of fusion transcripts called by multiple algorithms.

Similarly, the analysis of shared fusions across the same 23 samples from different RNA-Seq sources was extremely valuable. It highlighted how single fusion-calling algorithms often called fusion transcripts that were not replicated in another sequencing source. Some variability observed between different cell lines is currently unexplained but may be related to data quality from either RNA-Seq data source.

The difficulty in obtaining a list of fusions that comprehensive and free from false positives is not unique to our data-set. As mentioned in the introduction, multiple publications that examined RNA-Seq fusion calling algorithms agreed that there is a room for improvement in the sensitivity and specificity of the algorithms (Carrara et al., 2013; Kumar et al., 2016). It is possible that a certain degree of noise in the data is unavoidable with the current tools available, especially as fusion-calling relies on the detection of non-standard read alignments that may be particularly error-prone. Notably, very few papers that conduct fusion-calling from RNA-Seq data report any validation rates beyond the validation of a selected few high-confidence hits. The recent TCGA analysis reports a validation rate of 63.3%, based on a subset of samples with available whole-genome sequencing data (Gao et al., 2018). Their method of validation differs from my method of benchmarking fusion transcripts based on 945 PCR validations and the overlap of shared fusions in 23 different samples. Nonetheless, the ~70% validation rate from our methods suggest that our fusion-calling and filtering approach performs relatively well. At the same time, this analysis shows that working with data produced by single fusion-calling algorithms can be extremely unreliable and caution should be used when multiple algorithms are not used.

Future research may yield more insight into the nature and causes of sequencing and fusion-calling artefacts and could lead to advances in the sensitivity and specificity of computational fusion-calling algorithms and associated filters. Thus, in the future, it may be useful to reanalyse the RNA-Seq data for fusion calls, as any improvements in sensitivity and specificity can lead to substantial noise reduction in my downstream analyses.

3 Landscape of fusions in cancer cell lines

Our final list of fusions comprises 10,514 unique fusion transcripts identified in 1,011 cancer cell lines. This corresponds to 8,354 unique fusion events, i.e. fusion/cell line combinations, as the merged transcripts are likely to represent alternative splicing events.

Of the 1,011 cell lines examined, 52 cell lines showed no fusions. These cell lines do not appear to be particularly enriched within specific tissue types (Figure 3.1A). For all other cell lines, the number of fusions detected per cell line ranged from 1 to 96, with a median of 8. Notably, there was a significant correlation between sequencing depth and number of fusions detected in the cell line (spearman correlation, rho = 0.15, $p < 2.7*10^{-6}$) and also a small but significant difference in the mean of fusions found in CGHub data (median =

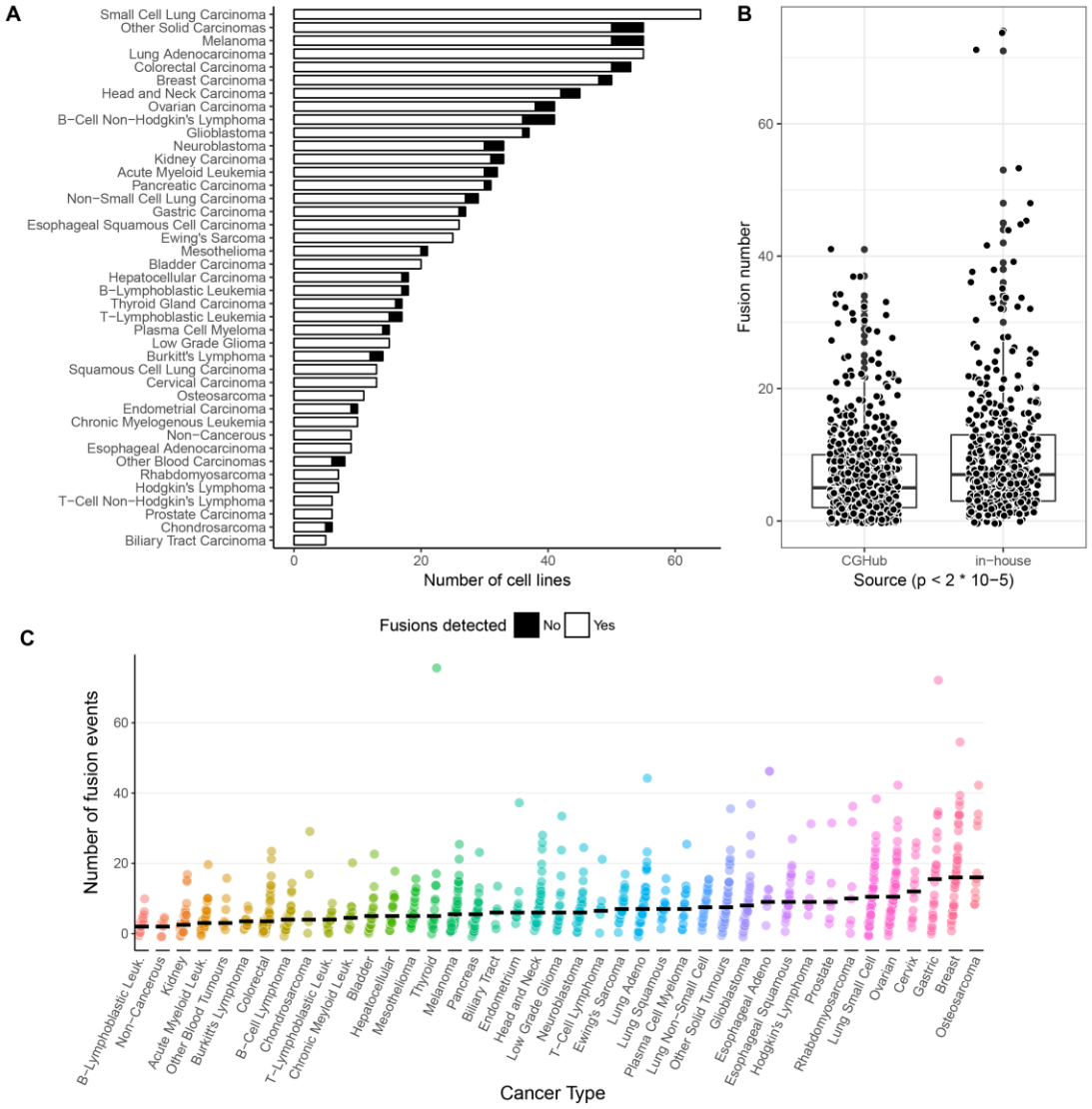


Figure 3.1: (A) Number of cell lines examined by cancer type. (B) Mean number of fusions by RNA-Seq data source. (C) Number of identified fusions for each cell line, split by tissue type.

7.2) compared to in-house sequenced samples (median = 9.6) (t-test: t = -4.3, p < 1.8 * 10^{-5}). This suggests that sequencing depth can influence the number of fusions detected in a sample, potentially due to improved sensitivity to detect transcripts that are expressed at a low levels.

3.1 Recurrence of fusions by cancer type

First, I examined the distribution of gene fusions across different cancer types (Figure 3.1C). Osteosarcoma and breast cancer showed the highest prevalence, with a median of 21 and 18.5 fusions per cell line respectively. They are followed by stomach and cervical cancers that present a median of 16 fusions per cell line. At the other end of the spectrum, a handful of non-cancerous cell lines showed a median of 2 fusions per cell line. Burkitt lymphomas, kidney cancers and B-lymphoblastic leukaemias all showed a median of 3 fusions per cell line.

In line with this, genomic landscape studies often find that breast cancers show one of the highest number of structural variation compared to other cancer types (Yang et al., 2016). This tendency may be due to a high frequency of *BRCA* mutations that are associated with genomic instability (Nik-Zainal et al., 2016; Sedic et al., 2015). Interestingly, in our data we indeed find that *BRCA1* mutations are significantly associated with a higher number of fusions in cell lines (p < 0.01, fdr = 13.4%. Controlled for tissue type and MSI status. Corrected for testing against 724 cancer events) (Figure 3.2).

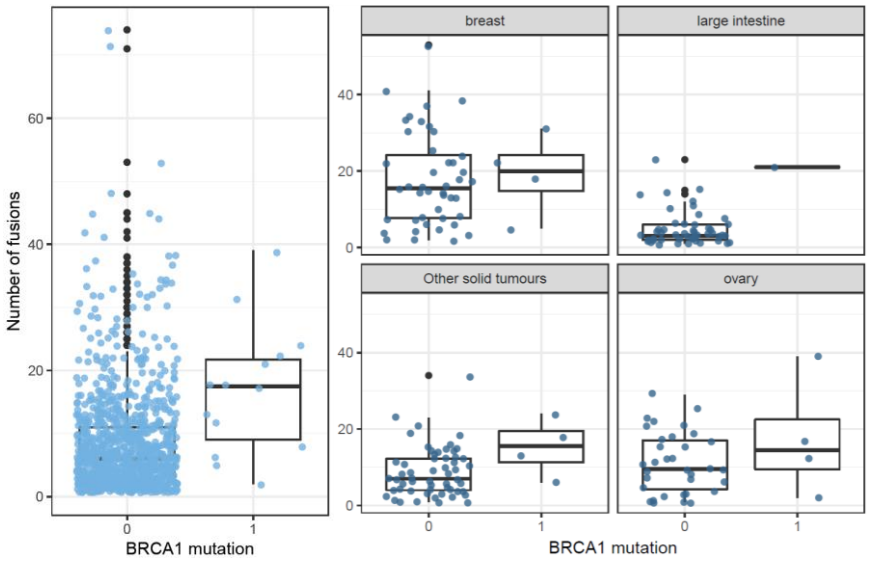


Figure 3.2: BRCA1 mutation status is significantly correlated with number of fusions detected in cell lines (p < 0.0086, fdr = 13.4%. (Left) number of fusions in a given cell line by BRCA1 mutational status across 952 cell lines. (Right) Same plots broken down by tissue categories with frequent BRCA1 mutations.

Osteosarcomas are rarely part of multi-tumour landscape papers, and their overall genomic instability is therefore difficult to directly compare against other cancer types. Nonetheless literature suggests that osteosarcomas distinguish themselves from other sarcomas due to large cytogenetic alterations and chromosomal instability (Cleton-Jansen et al., 2005; Lorenz et al., 2015). This may therefore explain a markedly above average occurrence of fusion genes.

Recently, Lorenz *et al.* characterised gene fusions in 11 osteosarcoma cell lines by RNA-Seq and whole genome sequencing (Lorenz et al., 2015). They validated 17 of 502 putative fusion transcripts by Sanger sequencing, with a validation rate of 76%. Interestingly, only 7 of the validating fusions detected by RNA-Seq were supported by structural changes predicted by whole genome sequencing, suggesting that the remaining gene fusions may result from non-structural changes (e.g. transcription-induced gene fusions). One of their examined cell lines, MG-63, is also part of our cell line panel. Interestingly, while Lorenz *et al.*'s pipeline returns 101 fusion transcripts in this cell line, our own analysis lists only 13. Both pipelines recapitulate all six fusions that are predicted by structural variation, however no fusions overlap that are not supported by structural changes. While the sample size is in this case small, and the data processing approaches are different (their pipeline constitutes of two fusion-mapping algorithms for which they combined any transcripts that have more than 3 split and 8 spanning reads), this finding puts into question whether the fusion transcripts not resulting from structural changes are stably expressed in time and across different experimental conditions, and their relevance in an oncogenic context.

3.1.1 Fusion recurrence in cell lines vs. patient tumours

Next, I wanted to compare the median number of fusions found in cell lines to those found in patient samples. For that, I mapped the tissue types reported by Gao *et al* for their ~10,000 TCGA samples to those we assign to our cell lines. TCGA tissue types were split into 33 different categories with letter codes and I manually retrieved the definitions from their website (NCI Genomic Data Commons, 2018). On the other hand, the cell line panel is split into 41 cancer types. I found corresponding TCGA tissue types for 26 different cancer types. Cancer types that were not represented included a number of haematological cancers (e.g. chronic myeloid leukaemia, Hogkin's lymphoma and lymphoblastic leukaemia) and some solid cancers (e.g. small cell lung cancers, Ewing's sarcoma or neuroblastomas).

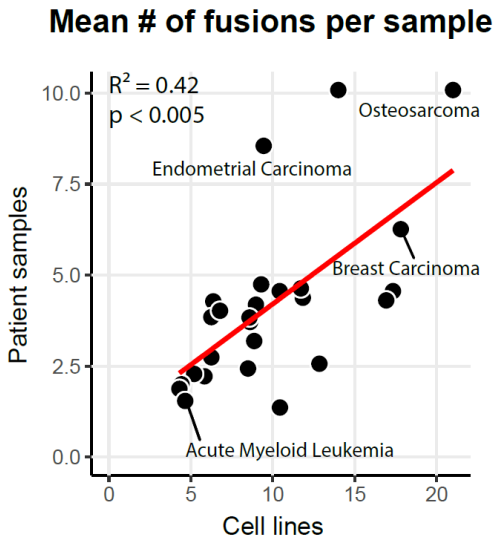


Figure 3.3: The mean number of fusions found in patient samples and cell lines correlate significantly.

For 22 TCGA tissue types, I found a direct match (e.g. BRCA – breast carcinoma, OV – ovarian carcinoma, LUSC – squamous lung cell carcinoma). For 5 TCGA tissue types, I had to find approximate matches, i.e. I combined KIRC (kidney renal clear cell carcinoma) and KIRP (kidney renal cell papillary cell carcinoma) to match the "Kidney Carcinoma" cancer type, which includes cell lines matching both of the TCGA categories. I also mapped ESCA (Esophageal carcinoma) to both Esophageal Squamous Cell Carcinoma and Esophageal Adenocarcinoma in our cancer type annotation, and the SARC (sarcoma) category was mapped to both Osteosarcoma and Rhabdomyosarcomas cell lines. Within these 26 mapped categories, the mean number of fusions per cell line to patient samples correlates significantly (R-square = 0.4 and p-value < 0.005) (Figure 3.3).

It is important to note that the scale differs – where patient samples show a range of 1.4-10 mean fusions per TCGA tissue, our cell lines have a range of 5.2-27.2. This difference is also observed on a data-set level, as our cell lines have a mean of 8.7 (range: 2-92) fusions per sample, while TCGA samples have a mean of 4.2 (range: 1-96) fusions per sample (t-test, $p < 2*10^{-16}$).

The higher range could potentially be due to cell line-specific biases. However, as in the previous chapter I show that higher sequencing depth is associated with a higher number of fusions, this may also be due to the exceptionally high sequencing depth at which RNA-Seq was performed for our samples. Unfortunately, the paper published by Gao and colleagues does not report sequencing depth of the tumours. Further, the real

sequencing depth in cell lines can be expected to be even higher than that of primary tissues, as there would be no contamination from non-tumour cells.

3.1.2 Tissue type specific fusions

Generally, I also find known oncogenic fusions in the tissue types in which they were previously reported in patients. For instance, *BCR-ABL1* and *NUP214-ABL1* fusions are found exclusively in blood cancers, e.g. chronic myelogenous leukaemia (n = 9), T-lymphoblastic leukaemias (n = 2) and B-lymphoblastic leukaemia (n = 1). Similarly, *EML4-ALK* fusions are only in lung adenocarcinomas (n = 2) while *NPM1-ALK* is found in T-cell non-Hodgkin's lymphoma (n = 4). *KMT2A* fusions are also mainly found in blood cancer cell lines, e.g. *KMT2A-MLLT1* (n = 1), *KMT2A-MLLT3* (n = 4), *KMT2A-MLLT4* (n = 1) and *KMT2A-AFF9* (n = 1) are all in leukaemia cell lines. *EWSR1-FLI1* fusions are mainly found in Ewing's sarcoma cell lines (n = 22), with the exception of one fusion event in the SK-NEP-1 cell line, which is annotated as a kidney cancer. However, a recent paper shows that this cell line was in fact misclassified and originates from Ewing's sarcoma (Smith et al., 2008). *TMPRSS2-ERG* (n = 2) are also found exclusively in prostate carcinoma cell lines.

Overall, the above data shows that cell lines represent tissue types reasonably well, at least in the relative recurrence of fusions events, as well as the tissue type-specificity of well-known oncogenic fusions. The latter observation in particular suggests that cancer-related fusions are faithfully represented in cell line model systems, in terms of their occurrence and functionality.

3.2 Fusion recurrence

To understand the recurrence of fusion events, I examined the number of samples in which a given fusion is present. Of 7,430 fusions, I found that the vast majority (n = 7,028; 94.6%) occur in only a single cell line. Moreover, only 1.3% of fusions are present in at least three cell lines or more (Figure 3.4A).

I found the same trend for fusions detected in the 9,624 TCGA patient tumours (Gao et al., 2018) (Figure 3.4B). There, of 23,999 fusions, 96% (n = 23,056) are called in only a single tumour and 1% (n = 237) fusions are recurrent in three samples or more.

Overall, the observation that only a small proportion of fusions are recurrent in both cell lines and patient fusions suggests that most fusion events are passenger events

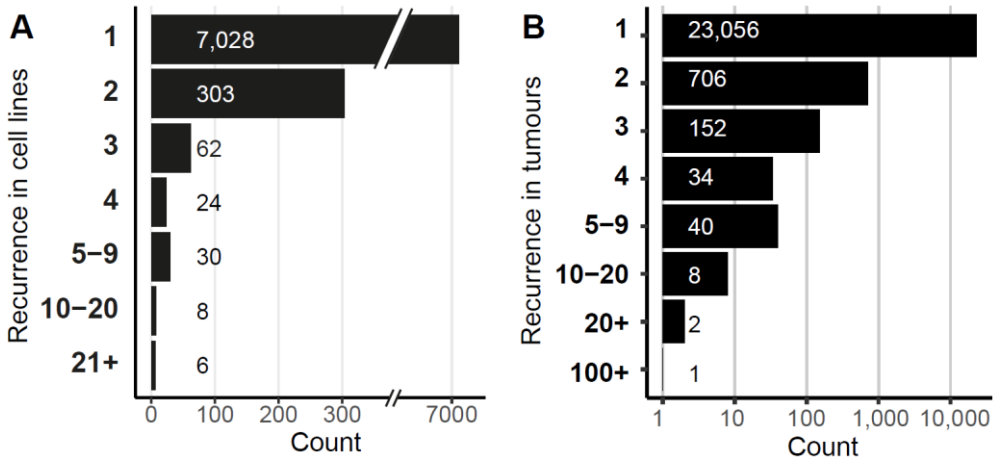


Figure 3.4: Recurrence of fusions in (A) cell lines and (B) patient samples (data from Gao et al, 2018).

that result from wide-spread structural rearrangement, and that only a minority are driver events selected for during oncogenic processes.

3.3 Cancer-related fusions and recurrence

Here, I annotated cancer-related fusions that belong into one of 3 categories: 1) known oncogenic fusions that have previously been causally implicated in cancer initiation and development, 2) fusions that have previously been observed in a subset of patient samples, and 3) fusions that involve known cancer-driver genes.

Fusions that have previously been causally implicated in cancer development and progression are important and useful tools in my study of functional gene fusions. As their oncogenic function is pre-established, they serve as valuable positive controls and benchmarks in my downstream analyses. To annotate known oncogenic fusions, I used the manually and expert-curated COSMIC gene fusion database (https://cancer.sanger.ac.uk/cosmic/fusion). The version downloaded in May 2018 contains 295 fusions.

Similarly, annotating fusions that have been observed in patient samples can help to put fusions that are found in cancer cell lines into context, as only fusions that actually occur in the patient population will be relevant for therapeutic applications. It also gives us an indication on the proportion of fusions observed in patients that we can functionally assess in our cancer cell lines. For this purpose, I use the recent paper published by Gao and colleagues, which analysed 9,624 TCGA tumour samples and found 25,664 gene

fusions (Gao et al., 2018). Fusions are considered as "observed in patients", if the same gene partners are fused in the same position (5'/3') in any of the samples analysed there.

Finally, fusions that involve known cancer driver genes may activate oncogenic functionality. For example, the fusion of a looped-coil domain to the kinase domain of *ALK* leads to constitutive kinase activation. Similarly, fusion genes involving known tumour suppressors can indicate knock-out of tumour-suppressor function. To annotate known cancer genes, I used the COSMIC Cancer Gene Census database, an expert-curated database that documents genes in which alterations have been causally implicated in cancer (https://cancer.sanger.ac.uk/census). My version of the database was downloaded in August 2017 and contains 609 genes. Of those, 91 were further annotated to be tumour suppressor genes, 116 were annotated as oncogenes and 37 genes have dual roles.

Of the 8,354 fusion events, 14.2% (n = 1,184) contained a cancer gene, 1.2% (n = 98) are known oncogenic fusions and 5.6% (n = 471) were previously detected in patients.

Known oncogenic fusions listed in the COSMIC fusion census show a tendency to have a higher recurrence than the rest, with the three top-most recurrent fusion in the

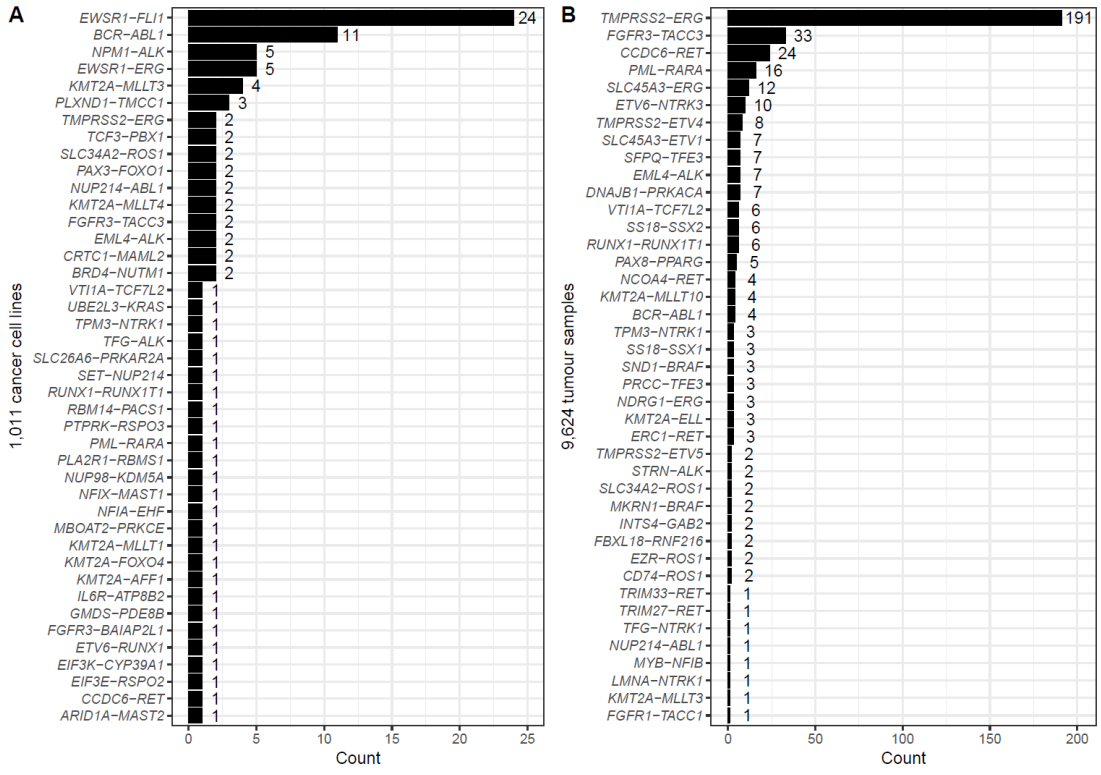


Figure 3.5: Recurrence of known oncogenic fusions in (A) cell lines and (B) patient samples (data from Gao et al, 2018).

TCGA data set being *TMPRSS2-ERG* (n = 191), *FGFR3-TACC3* (n = 33) and *CCDC6-RET* (n = 24). Similarly, *EWSR1-FLI1* is with 24 cell lines one of the most commonly fused fusions (Figure 3.5A-B).

When looking at the top 40 most recurrent fusions overall, I find that within cell lines, there is less of an enrichment for COSMIC fusions and fusions that involve COSMIC cancer genes than in the patient samples (Figure 3.6A-B). This is likely because the authors removed any fusion that recurs in 10 or more samples unless they were reported in previous TCGA studies (Gao et al., 2018). Although the number of fusions that were removed with this filter is not reported in the manuscript, considering that only 40 fusions recurred in 5-9 different samples (Figure 3.4B), it is likely to be small.

In cell lines, two of the top three recurrent fusions, *TSHZ2-SLC35A1* and *NCOR2-UBC*, have been described before, but never characterised (Obholzer et al., 2015; Roberts et al., 2012). Some of the top recurrent fusions contain entirely uncharacterised genes (e.g. *CTD-2334D19.1* and *RP11-120D5.1*) and long intergenic non-protein coding RNA (e.g. *LINC01340*). These fusions tend not to be tissue-type specific and are often not in-frame, which suggests that they are less likely to perform oncogenic functionality.

It is notable that of the COSMIC fusions found in our cell lines, only a subset (28%) are observed in Gao's patient samples. Well known oncogenic fusions that are not observed in Gao's patient samples include *EWSR1-FLI1*, *NPM1-ALK* and *BRD4-NUTM1*. The

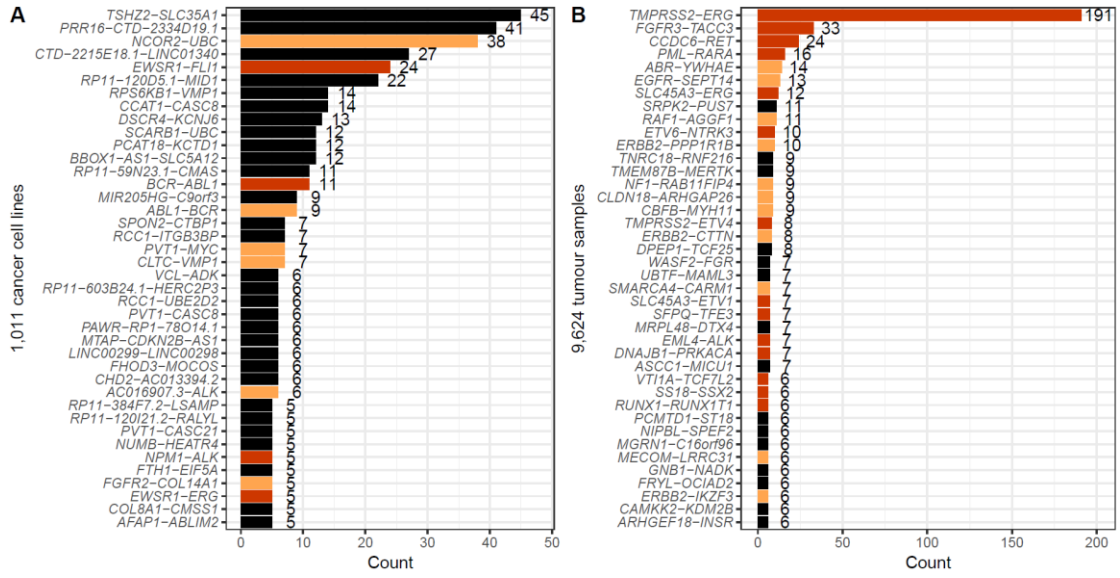


Figure 3.6: Top 40 most recurrent fusions in (A) cell lines and (B) patient samples (data from Gao et al, 2018). Red bars indicate that the fusion is listed in the COSMIC fusion data base, dark yellow bars indicate that one of the genes is listed in the COSMIC cancer gene census.

opposite is also true, e.g. no cell lines show *ETV6-NTRK3* fusions, even though those are among the most recurrent in the patient samples. Furthermore, a large number of well-understood driver events also occur at only very low frequencies (e.g. *KMT2A-MLLT3* only being in a single patient sample and *PML-RARA* only occurring in a single cell line). This likely reflects the biases of tissue selection within both data sets, e.g. there being no Ewing's sarcoma samples analysed in the TCGA data set. It also supports the argument that fusions are exceedingly rare events, and that even with almost 10,000 samples sequenced, we are still far off from being able to provide a comprehensive landscape of oncogenic fusion transcripts. Similarly, it demonstrates that while recurrence is now commonly used as a predictor for the oncogenic functionality of single-nucleotide variations, with the current resources available, fusions recurrence alone is currently insufficient in assigning functionality.

3.4 Predicted frame of fusion transcript

GRASS (Gene Rearrangement AnalySiS) is an algorithm developed by CASM IT at the Wellcome Trust Sanger Institute that predicts the impact that a breakpoint has on the formation of potential fusion constructs based on the ensemble gene annotation (Cancer IT, 2018b). It is automatically run with the cgpRna pipeline and outputs a number code ("flag") that annotates fusion transcripts with relevant features, including frame (e.g. in-frame, out-of-frame, stop codon) and breakpoint region (i.e. intronic, extronic or UTR).

In total, there are 49 different grass flags, which represents different combinations of frame, breakpoint region etc (e.g. an in-frame transcript where the breakpoint is in an intron has a code of 910, while an in-frame transcript with an exonic breakpoint has a code of 900). The original 49 grass flags were designed for an algorithm to predict breakpoints in whole genome sequencing (Cancer IT, 2018c), and are not applicable to transcriptome data. 26 codes appear for the fusion transcripts called by our pipeline, and these can be summarised in 7 categories (Table 3-1). Four categories affected coding regions: In-frame, ambiguous frame, out-of-frame, stop-codon introduced. A fifth category annotates any transcript that involves a fusion at an untranslated region (UTR). Two further categories annotate, where a fusion involves "something else" or "no fusion". These two categories, together with the "ambiguous frame" mentioned above, are usually assigned in situations, where the ensemble annotation on coding regions or gene boundaries is ambiguous.

Table 3-1: 26 GRASS codes and their descriptions. The first column ("category") is the custom categories I divided the codes into to simplify interpretation of the GRASS flags.

Category	GRASS flag	Type	Breakpoint region	Gene orientation
In-frame	910	in-frame	exon	same
	900	in-frame	intron	same
Ambiguous frame	740	ambiguous	intron	same
	520	ambiguous	exon	different
	540	ambiguous	intron	different
	500	ambiguous	coding	different
Out-of-frame	710	out-of-frame	exon	same
	700	out-of-frame	intron	same
Stop codon	715	stop-codon		same
UTR fusion	860	5' UTR to 5' UTR		same
	820	5' UTR to coding		same
	660	5' UTR to 3' UTR		same
	620/600	3' UTR to coding		same
	440	3' UTR to 5' UTR		same
	400	3' UTR to 3' UTR		same
	360	5' UTR to 5' UTR		different
	580/570	5' UTR to coding		different
	320	5' UTR to 3' UTR		different
	560	3' UTR to coding		different
	280	3' UTR to 5' UTR		different
No fusion	0	No Fusion		NA
Something else	200	intron to something else		NA
	180	something else to intron		NA
	100	something else to something else		NA

Unfortunately, the documentation for GRASS does not provide further detail on the specific circumstances in which a fusion is assigned "ambiguous frame", "something else" or "no fusion", and an attempt to reach out to the original authors of the code was unsuccessful.

I broke down all fusion events by the 7 GRASS flag categories – where several fusion transcripts exist for the same fusion event, I assigned the GRASS flag with the highest value (i.e. the outcome that is most likely to have functional significance) to the event. I found that 26% of all transcripts are annotated as in-frame (Figure 3.7A). Fusions events in coding regions make up 50.8% of all events, another 14.6% are fused in UTR's and the remainder are fusions that fall into the "something else" or "no fusion category".

When looking specifically at fusion events that involve COSMIC fusions, I find that the vast majority (91%) are in-frame events (Figure 3.7B). Only 9 fusion events do not fall

into this category. A *TMPRSS2-ERG* fusion in a prostate carcinoma cell line shows the standard phenotype, where a fusion to the 5′ UTR region of androgen-regulated *TMPRSS2* drives overexpression of *ERG* (Tomlins et al., 2008). A further 5 of not in-frame fusions (*PLXND1-TMCC1* in two cell lines, *IL6R-ATP8B2*, *NFIX-MAST1*, *RBM14-PACS1*) were originally identified in high-throughput landscape papers for which no functional validation has been performed. In fact, in a later chapter, I show evidence to argue that these fusions should not be included in the COSMIC fusion census (see chapter 6.3).

Two more fusion events, *FGFR3-TACC3* in RT4 and *EIF3E-RSPO2* in ESO51 are both annotated as "ambiguous frame". In later chapters, I show that the *FGFR3-TACC3* is likely to be functional, as RT4 is sensitive to FGFR3 inhibitor PD173074 (see chapter 4.3.1.5). On the other hand, *EIF3E-RSPO2* fusions have previously been found to be sensitive to porcupine inhibitors in colorectal cancer, and to which ESO51, an oesophageal cancer cell line showed no response, suggesting that this fusion event is non-functional.

The final not in-frame fusion event is an *UBE2L3-KRAS* fusion detected in DU145, a prostate adenocarcinoma cell line that is annotated as "something else". In fact, this fusion in the same cell line was already previously functionally characterised by another group and showed attenuated xenograft growth and cell invasion when knocked-down (Wang et al., 2011). I found the break-point reported by our fusion-calling pipeline to be identical to the one described by the authors, which suggests that in this case, the GRASS algorithm failed to correctly annotate the fusion.

Thus overall known oncogenic fusions listed in the COSMIC database where overwhelmingly in-frame, except in the case of the 5′ UTR fusion of *TMPRSS2-ERG*.

Next, in the subset of fusion events that involve a COSMIC cancer gene, I see a much wider range of predicted outcomes (Figure 3.7C). In general, the pattern follows that of all other fusion events, though with a slightly higher proportion of in-frame fusion events (33% vs. 26%). This proportion of in-frame events is even higher when looking at just the subset of COSMIC genes that are annotated as "onocogene" (41%), while in COSMIC genes annotated as "tumour suppressor genes" (TSG), the proportion of in-frame events is just 29% (Figure 3.7D-E). Although not all in-frame fusion events will necessarily be functional, an enrichment of them in fusion events involving cancer driver oncogenes could suggest positive selection. For instance, these events include two in-frame *RAF1* fusions, *MBNL1-RAF1* and *ATG7-RAF1*, which will be discussed in chapter 6.5, where a

fusion most likely acts by removing an N-terminal auto-inhibitory domain of *RAF1*. At the same time, the nonetheless large proportion of COSMIC cancer gene-carrying fusion events that are out-of-frame or with ambiguous mapping suggests that one should be careful not to over-interpret the occurrence of a cancer gene in a fusion transcript.

Notably, the subset of fusion events that are also seen in patient samples has an even higher proportion of in-frame fusion events (41.5%) (Figure 3.7F). Furthermore compared to all fusion events, an even smaller proportion of fusion events fall into the

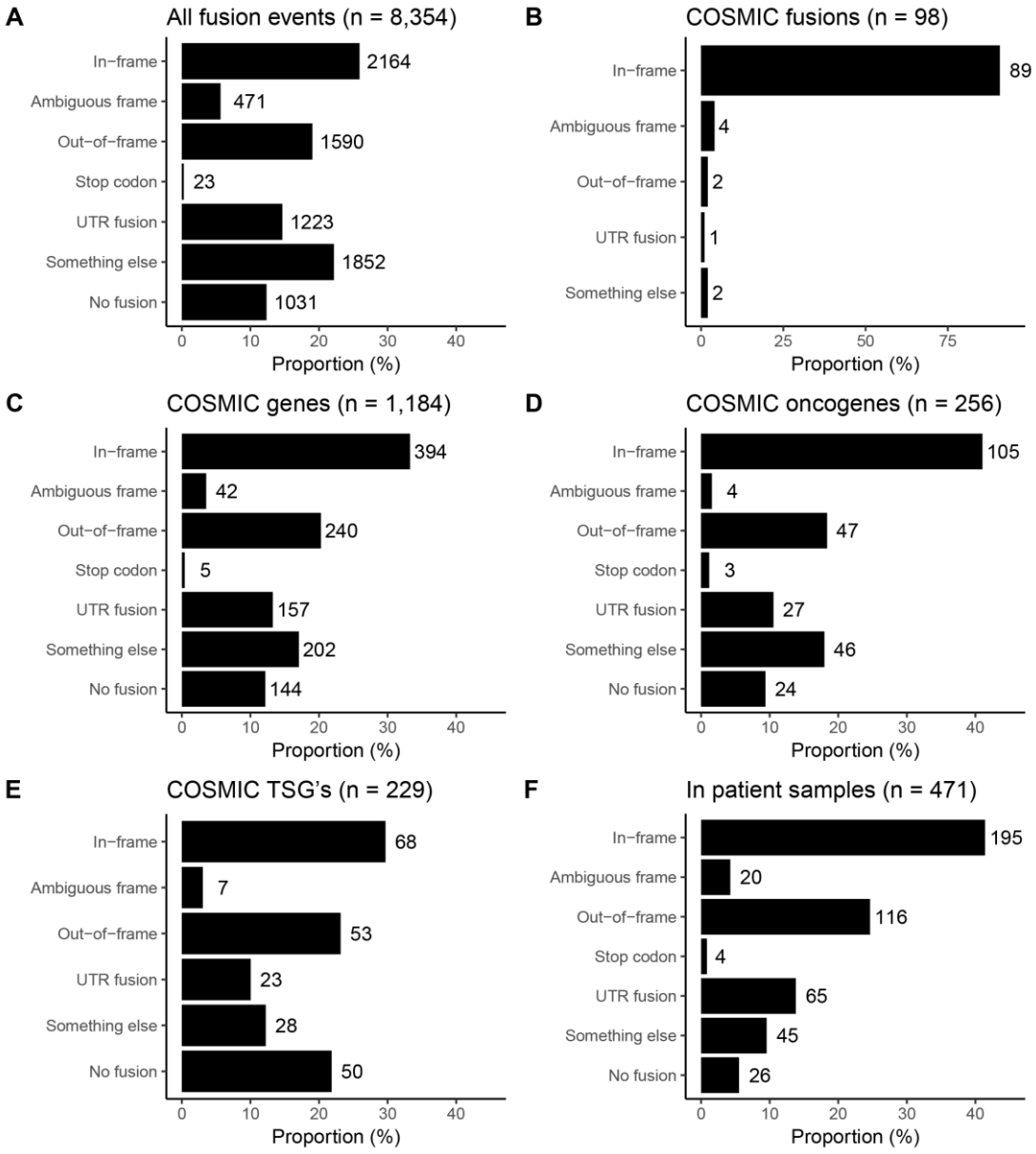


Figure 3.7: Break-down of fusion transcripts/events by GRASS flag category for (A) all fusions transcripts, (B) fusion events matching COSMIC fusions, (C) fusion transcripts involving COSMIC cancer genes and (D) fusion events that were observed in patient samples. Numbers next to bars represent the number of fusion events that fall into a given category.

"something else" (9.6% vs. 22.2%) and "no fusion" (5.5% vs. 12.3%) categories. In the future, it could be interesting to examine the predicted impact across the entire list of fusion transcripts called in patient fusions, to see whether this difference is representative of fusion transcripts found in patients, or due to biased subsetting. As of now, it is unclear whether ambiguous fusions are artefacts that arise from cell culture, differences in sequencing and fusion-calling techniques, or simple an effect of subsetting. Either way, for future iterations, a deeper understanding of the mechanisms that cause ambiguous fusion events could potentially help to filter out artefacts from the fusion calling pipeline.

3.5 Chapter summary

In this chapter, I attempted to briefly characterise the list of fusions obtained. I find that 1) median number of fusions varies by tissue type of origin in a manner that correlates significantly with that found in patient samples. 2) Fusions are exceedingly rare events and the majority of fusions (~95%) occur only in a single cell line. I also found that recurrence alone is a poor predictor of functionality. 3) Although known oncogenic fusions are almost exclusively in-frame events, in-frame events make up a relatively low proportion of fusions, suggesting that the proportion of true functional fusions may be low.

Known oncogenic fusions serves as a useful positive control, although they only make up only around 1% of all detected fusion events. They are largely found in cell lines with the expected cancer type of origin (e.g. *BCR-ABL1* found exclusively in blood cancers). This gives us confidence that our cell lines present a good model system for truly cancer-relevant fusions.

Thus the next challenge lies in finding a method of separating functional fusion events from passengers of structural rearrangements. In the up-coming chapters, I will describe two analysis that are aimed to identify novel functionality and therapeutic potential for previously uncharacterised fusions. Firstly, I use the high-throughput drug-screening data of > 400 drug compounds to examine the use of fusions as biomarkers of drug response. Secondly, I developed a system to leverage CRISPR/Cas9 whole genome drop-out screening data to construct fusion-specific essentiality profiles.

4 Identification of fusion genes as biomarkers for drug response

Having identified 8,388 gene fusions in our panel of 1,011 cancer cell lines, my next aim was to identify the potential of using fusion genes as biomarkers of drug response. Fusion genes are already playing a major role in helping to provide the best treatment to given patients, e.g. patient with *BCR-ABL1* have the best treatment outcomes when treated with ABL1 inhibitors and *EML4-ALK* fusions are best treated with ALK inhibitors. Novel fusion-drug associations may lead us to discover more therapeutic opportunities. At the same time, a fusion-drug association can indicate functionality in novel fusions.

The panel of 1,011 cancer cell lines has previously undergone extensive high-throughput drug screening across hundreds of drug compounds (Garnett et al., 2012; Iorio et al., 2016). I implemented this with an ANOVA framework and considered well-understood cancer functional events (CFEs) as confounding factors.

Finally, I ran three distinct analyses: 1) An ANOVA including only CFEs that would provide a basis for 2) a fusion-level and 3) a gene-level ANOVA.

4.1 Set-up and data-sets

4.1.1 ANOVA model

I tested high-throughput drug sensitivity measurements for 409 drug screening data sets against any gene fusion that occurred in 2 or more cell lines (n = 412) under the following ANOVA model:

It was tested for all combinations of a drug screening data set ($IC50_D$) and relevant gene fusions ($Fusion_F$), for a total of 168,508 combinations.

$$IC50_D \sim Tissue\ Type + MSI\ Status + CE_D + Fusion_F$$

Equation 1: Fusion ANOVA model

The 42 tissue types and MSI status were used as covariates in the model, as a large amount of variance from drug response can be attributed to these features[6,7]. Further, I implemented a co-variate (CE_D), which controls for other non-fusion genetic alterations that have a significant association with drug response (see section 4.2).

4.1.2 409 drug IDs

For this analysis, I utilised 409 drug screening data sets, each with a distinct drug ID. These data sets included data for 379 unique chemical compounds, of which some were

screened multiple times. Since duplicated measurements were taken several years apart on different screening formats (e.g. in 384-well plates vs. 1,536-well plates), I decided to keep them separate for my analyses. To maximise cell line coverage, multiple drug screening data sets for the same drug compounds were included where they covered different sets of cell lines. For the sake of simplicity, unique drug screening data sets are referred to "drug IDs" within this thesis.

All drug IDs were screened as part of the GDSC1000 high-throughput drug screening pipeline at the Wellcome Sanger Institute (Cancerrxgene.org; Garnett et al., 2012; Iorio et al., 2016). They span 279 unique drug targets and are categorised into 24 effector pathways. A subset of 77 compounds are FDA-approved, while further subsets are currently in clinical trials, or still experimental (Figure 4.1; Supplementary Table 4). The 409 drug IDs had a median coverage of 805 cell lines (range: 42-947).

4.2 Cancer Functional Events ANOVA

Another covariate included in the model is CFE_D. This denotes the occurrence of other cancer functional events (CFEs) that significantly correlate with sensitivity to a given drug. With this approach, I aim to minimise potential false positives associations that can

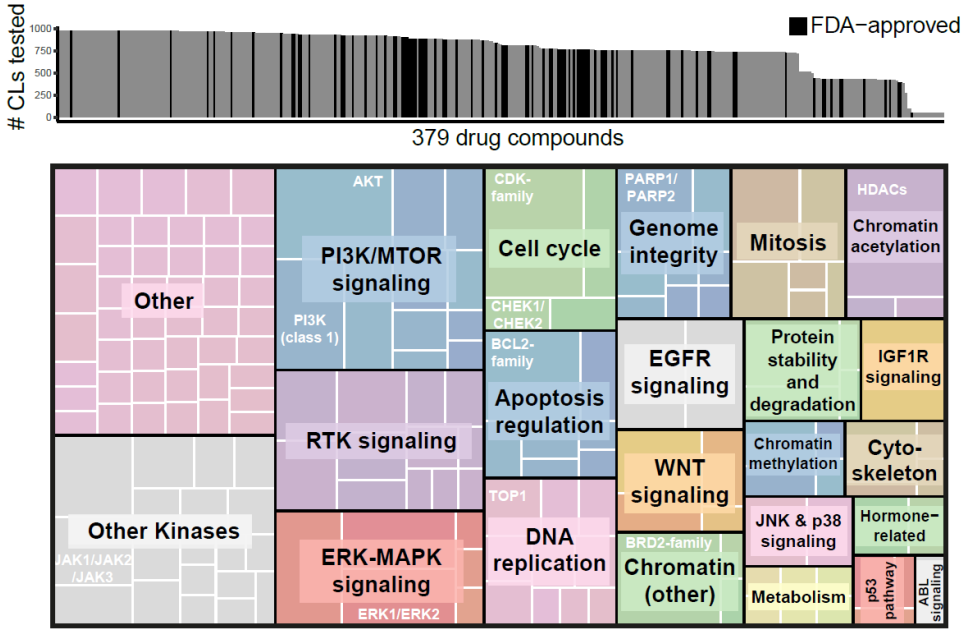


Figure 4.1: Overview of 379 drug compounds that make up the drug-screening data. (Top) The total number of cell lines tested for each of the 379 drug compounds, with black bar colour indicating FDA-approved drugs. (Bottom) Tree-map of 24 major effector pathways that are targeted by the panel of drug compounds. Size of the field is proportional to number of drugs with a given target.

occur when gene fusions by chance co-occur with CFEs that confer sensitivity to a given drug.

To define the set of known associations between CFEs and drugs, I ran a separate ANOVA under the model described in the following equation across the same 409 drugs ($IC50_D$) against a set of 717 recurrent cancer functional events (CFE_D):

$$IC50_D \sim Tissue\ Type + MSI\ Status + CFE_D$$

Equation 2: Model for CFE ANOVA

The ANOVA model was run across all combinations of drugs and CFEs. I adjusted p-values for significance of the CFE using the Benjamini-Hochberg method to calculate a False Discovery Rate (FDR) (Benjamini and Hochberg, 1995).

4.2.1 Cancer Functional Events

CFEs were defined on our cell line panel by Iorio *et al.* (Iorio et al., 2016), by overlaying recurrent gene alterations found in a large set of patient samples with those found in our cell lines. The binary annotation was downloaded from cancerrxgene.org/gdsc1000. Methylation data was filtered to include only those segments that were annotated to contain a known cancer driver gene compared against the Cancer Gene Census. The final input data includes mutation status for 281 genes, copy number alteration for 424 segments and methylation status for 12 segments.

4.2.2 Incorporating significant results in the gene fusion analysis

As in the paper published by Iorio *et al.* (2016), I considered any result with an FDR < 25% and both positive and negative Glass Delta > 1 as large-effect association. The Glass Delta is a measure of effect size and is calculated by dividing the difference in mean log IC_{50} of the two groups by the standard deviation of one of the groups (hence each association has two Glass Delta's, one genetic-alteration-positive and one negative). A Glass Delta of > 1 shows that the difference in mean log IC_{50}s is larger than the standard deviation. In total, 101 combinations of CFE and drug sensitivity passed these criteria (Supplementary Table 5). For the gene fusion ANOVA, I implemented for any drug involved in a significant association, the status of associated CFEs in cell lines as a covariate as shown in Equation 1.

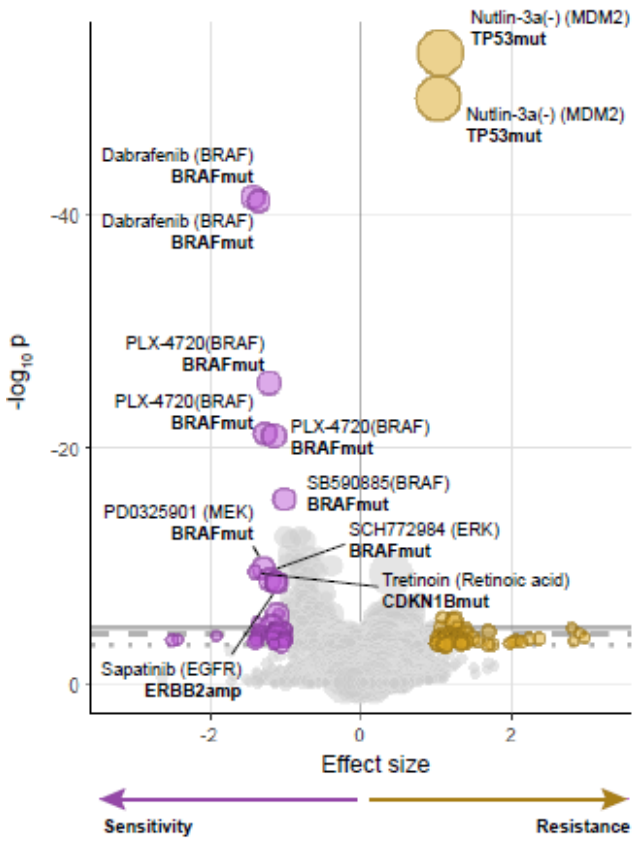


Figure 4.2: Volcano plot of all significant results from the CFE ANOVA. Results that do not meet our criteria for FDR and Glass Delta are represented in grey. Effect size denotes the minimum Glass Delta for a given association. Size of circle is relative to number of cell lines with a given CFE.

4.2.3 Characterisation of significant findings

Our results show 62 resistant and 39 sensitising associations. Results with the highest p-values include prominent well-known associations. *TP53*-mutations, for example, are associated with resistance to Nutlin-3a, a compound that induces tumour suppression by inhibiting the MDM2-interaction with p53 (Vassilev et al., 2004). Similarly, *BRAF*-mutations typically lead to enhanced kinase activity and activation of the down-stream MEK pathway, and are associated with sensitivity to several BRAF-inhibitors, including Dabrafenib and PLX-4720 (Hauschild et al., 2012; Kopetz et al., 2010).

The results recapitulate 88% of those shown in Iorio *et al.* (2016). Small differences are to be expected, as an updated annotation of cancer tissue types was used in my analysis that annotated over 300 cell lines that were previously unannotated by Iorio *et al*.

Novel associations not reported in Iorio *et al.* (2016) include 16 sensitivities and 40 resistances. Of the sensitivities, 4 represented new drugs targeting the same targets as known associations (e.g. *BRAF*-mutation associating with MEK/ERK-inhibitors and an *ERBB2*-amplification with an EGFR-inhibitor).

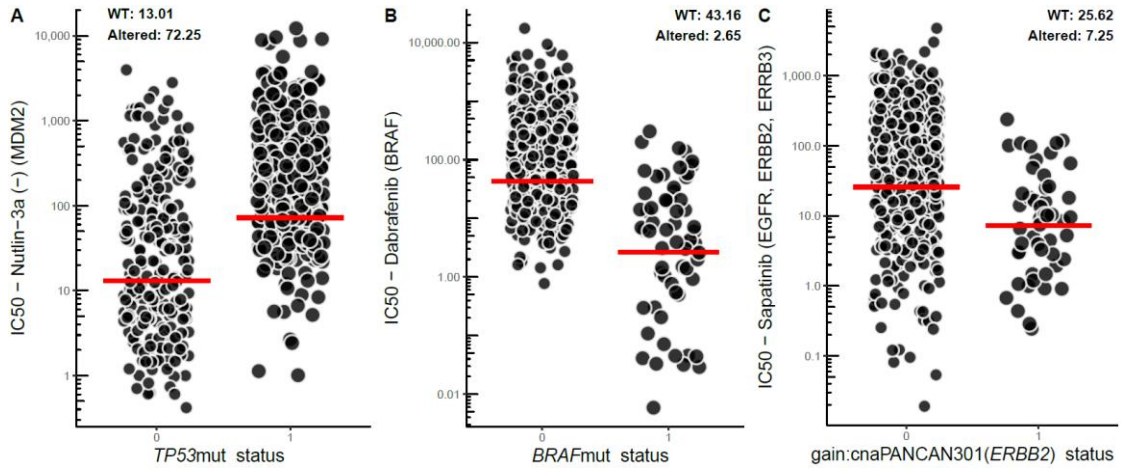


Figure 4.3: Known associations between CFEs and drug compounds. (A) TP53-mutation and MDM2 inhibition (B) BRAF-mutation and BRAF-inhibition (C) Amplification of ERBB2 and EGFR/ERBB2-inhibition.

One interesting novel finding is the association of EPH receptor A2 (*EPHA2*) with sensitivity to a PI3K-alpha inhibitor, GNE-317 (Figure 4.4A-B). EPHA2 is suggested to act with PI3K in tumour vascularisation (Wang et al., 2016). Further, EPHA2-signalling in a PI3K/AKT dependent manner can be upregulated as a potential mechanism to evade BRAF-inhibition in melanoma (Paraiso et al., 2015). In our analysis, *EPHA2*-mutations are found in a variety of cancer types of origin, suggesting that *EPHA2*-altered tumours may benefit from PI3K-pathway inhibitors across several cancer types. Although not all significant, when ranking all associations involving the *EPHA2*-mutation by FDR, there is a significant enrichment of inhibitors targeting PI3K or AKT (ROC AUC: 0.654, p = 0.0075 with 10,000 permutations; Figure 4.4C).

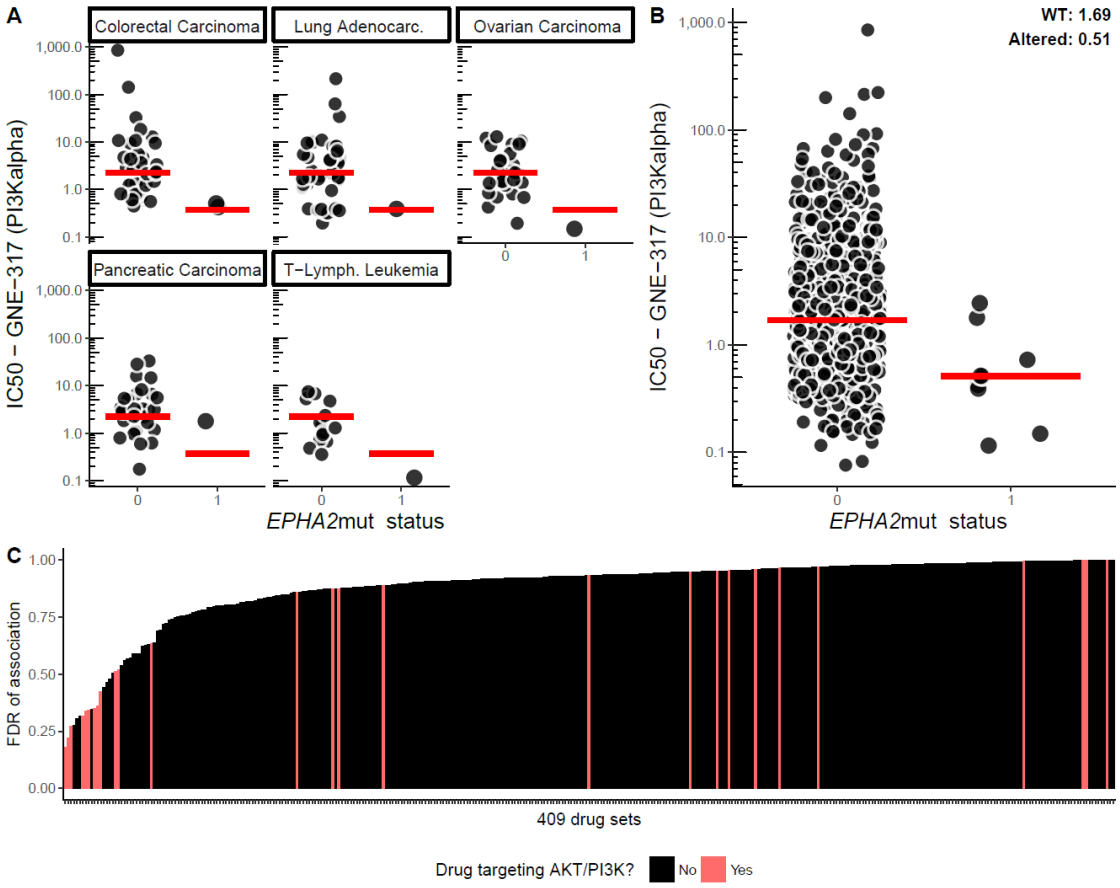
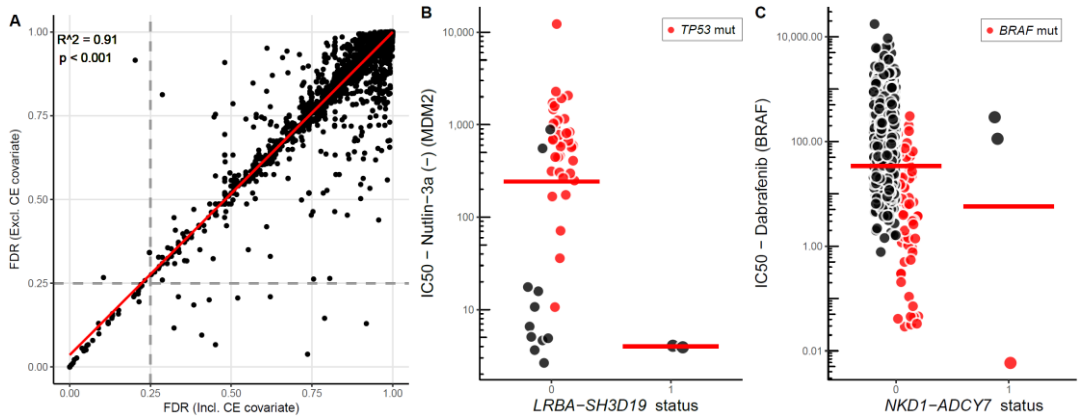


Figure 4.4: EPHA2-mutational status is significantly associated with sensitivity to PI3K-alpha inhibitor GNE-317. LogIC50 by mutation status (A) by tissue and (B) across all cell lines. (C) When ranked by FDR, top drug IDs tested for association with EPHA2-mutational status are enriched for drugs targeting either PI3K or AKT.

4.2.4 Using significant CFEs as covariate in the fusion ANOVA

Implementing the CFEs as covariates, where the associations fit the previously described filters (FDR < 25% and Glass Delta's > 1), I assigned co-variates to 73 drug IDs.



FEATURE	Drug Name	Drug Target	CFE covariate(s)	FDR	Uncorr. FDR
GPHN-MPP5	NSC-87877	SHP-1, SHP-2	MAP3K4-mut, SRGAP1-mut	32.3	11.6
KIF26B-EFCAB2	Sepantronium bromide	BIRC5	HGF-mut, RPGR-mut	73.7	3.8
KMT2A-MLLT3	Sorafenib	PDGFR, KIT, VEGFR	U2AF1_mut	62.1	21.0
LRBA-SH3D19	Nutlin-3a (-)	MDM2	TP53-mut	45.1	6.7
NKD1-ADCY7	Dabrafenib	BRAF	BRAF-mut	40.9	9.5
NKD1-ADCY7	SCH772984	ERK1, ERK2	BRAF-mut	43.2	20.0
NPM1-ALK	Alisertib	AURKA	loss:cnaPANCAN425	38.4	18.5
RCC1-SNX5	Idelalisib	PI3Kdelta	MYC-mut	33.7	21.0
UBA2-GPI	NSC-87877	SHP-1, SHP-2	MAP3K4-mut, SRGAP1-mut	52.0	20.6
YAP1-MAML2	Dabrafenib	BRAF	BRAF-mut	91.9	13.0
ZAK-RAPGEF4	AR-42	HDAC1	TBX3-mut, CCND1-mut	78.9	14.5

Figure 4.5: (A) False discovery rates of fusion-drug associations with and without including the CFE covariate. (B) Dabrafenib sensitivity by NKD1-ADCY7 fusion (C) Nutlin-3a sensitivity by LRBA-SH3D19 fusion for colorectal carcinomas. (D) Eleven associations for which the covariate corrects the significance of the association.

After running the fusion ANOVA model (Equation 1) with and without the covariate CE_D, the significance of the majority of associations is largely unchanged (Figure 4.5A), although eleven associations are significant at FDR < 25% are returned as not-significant when including the covariate.

Covariates that cause a shift in significance fall into one of two patterns. In some cases, fused cell lines are a subset of those with the covariate. For example, in the association of the *NKD1-ADCY7* fusion with BRAF-inhibitor Dabrafenib, the cell line that

shows the highest sensitivity also harbours a *BRAF*-mutation, which likely drives the association (Figure 4.5C).

In other cases, a covariate may have a large effect on the variance of a drug response. After having mean-centred for the covariate, the variance for the remaining cell lines could be greatly reduced. For example in the case of *LRBA-SH3D19* and Nutlin-3a, *TP53*-mutated cell lines are significantly less sensitive than non-mutated cell lines. As the two *LRBA-SH3D19* fused cell lines are non-TP53-mutated, their drug sensitivity is expected to be low (Figure 4.5B). A further two associations that became more significant by including the CFE covariate also fall into this latter category.

The number of cases affected by including the CFE covariate is low. Nonetheless, after manual inspection of all 11 cases, I concluded that the quality of the results is improved by its inclusion by removing false-positives such as the ones outlined above.

4.3 Fusion ANOVA

To run the fusion ANOVA, I use Equation 1 to query the associations between 409 drug IDs and 431 fusions. As before, I adjust p-values for multiple-hypothesis correction by calculating an FDR. From 176,279 potential combinations, 324 had a FDR < 25%, of which 227 were large-effect associations that also had Glass Deltas of > 1 (Supplementary Table 6). Of those, 172 were sensitising and 55 resistant associations.

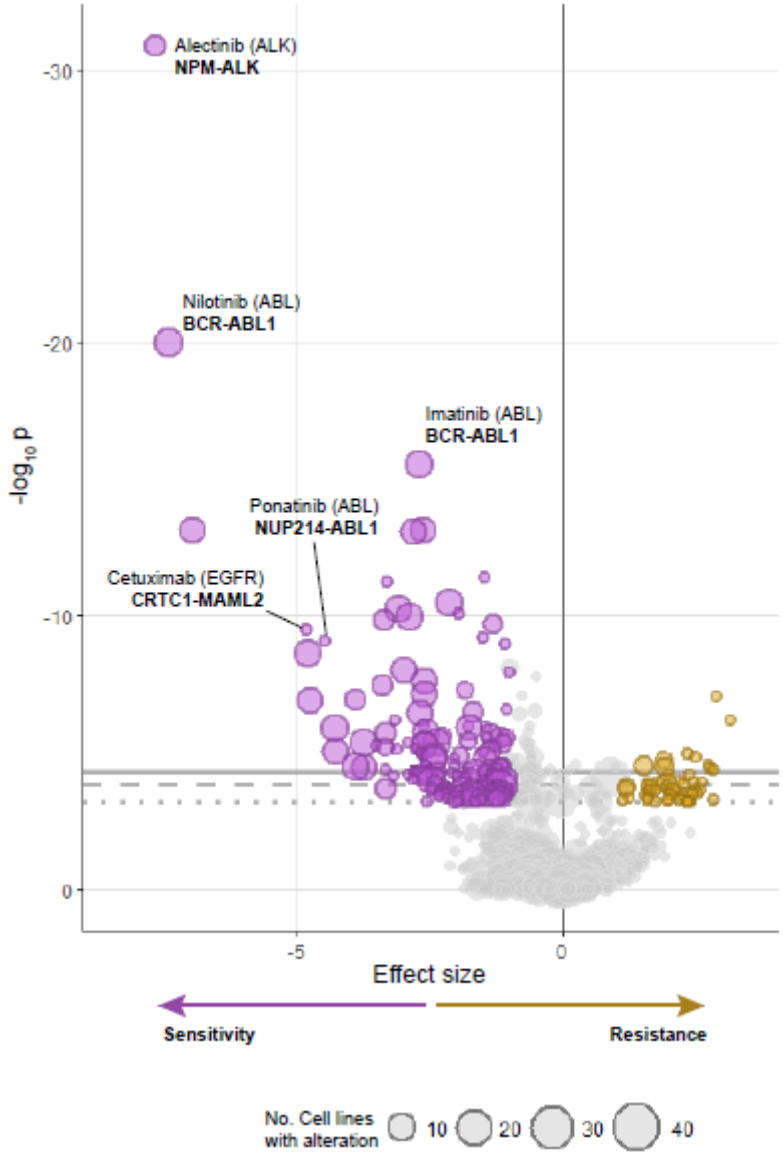


Figure 4.6: Volcano plot of all significant results from the fusion ANOVA. Results that do not meet our criteria for FDR and Glass Delta are represented in grey. Effect size denotes the minimum Glass Delta for a given association. Size of circle is relative to number of cell lines with a given fusion.

4.3.1 Known fusion-drug associations

Known fusion-drug associations make up a large portion of large-effect associations. Below follows a brief review of the most commonly studied fusions and their representation in my analysis.

4.3.1.1 ALK fusions

ALK is a tyrosine receptor kinase that is commonly found altered in many haematological and solid cancers. While sometimes mutated, e.g. in a subset of neuroblastomas (Chen et al., 2008), more often the tyrosine kinase domain of ALK is fused to the C'-terminus of another protein. Partner genes in the *ALK* fusion encourage dimerization in absence of a ligand and therefore lead to activation of the ALK kinase domain (Mossé et al., 2009). In normal development, ALK is specifically expressed in the neural system, mainly in neonates (Iwahara et al., 1997), with a suggested role in neuronal differentiation. In cancer, the ALK fusions encourage cell cycle progression, survival, cell shaping and has transforming ability.

In anaplastic lymphomas and some T-cell non-Hodgkin's lymphomas, *ALK* is commonly fused to nucleophosmin 1 (*NPM1)*, which contributes a nuclear localisation domain to the chimeric protein. *ALK* fusions are also commonly found fused in 3-6% of non-small-cell lung cancer with echinoderm microtubule associated protein like 4 (*EML4*) (Koivunen et al., 2008; Perner et al., 2008).

Small molecule inhibitors targeting the tyrosine kinase of ALK show great response rates in *in vivo* experiments, and clinical trials for several ALK-inhibitors are underway (Mossé et al., 2009).

My analysis recapitulates these previous findings robustly. The 5 T-cell non-Hodgkin's lymphoma cell lines that carrying an *NPM1-ALK* fusion, and the two lung adenocarcinoma cell lines with an *EML4-ALK* fusion, significantly associate with sensitivity to numerous ALK inhibitors, including crizotinib, alectinib and XMD14-99 (all at $p < 0.001$ and FDR < 0.01%; Figure 4.7).

4.3.1.2 ABL1 fusions

The fusion between BCR ABL has been known since the 1960s with the discovery of the Philadelphia chromosome, a minute chromosome resulting from a translocation of chromosomes 9 and 22 t(9;22)(q34;q11), in chronic myeloid leukaemia patients (Braekeleer et al., 2011). The fused, C-terminal domain of ABL1 contains the SH1 tyrosine kinase

domain as well as a nuclear localisation domain (Colicelli, 2010). The N-terminal partner gene, BCR, facilitates dimerization through a coiled-coil domain and thus leads to constitutive activation of the tyrosine kinase of ABL1 through reciprocal phosphorylation (McWhirter et al., 1993).

The *BCR-ABL1* fusion is a hallmark of chronic myeloid leukaemia, but is also found in other haematological malignancies, such as a small subset of B-lymphoblastic leukaemias. A rare subtype in about 5% of T-lymphoblastic leukaemias carries a variation on the fusion (Braekeleer et al., 2011). There, ABL1 is fused to the N-terminal NUP214, resulting in the NUP214-ABL1 fusion. Similar to BCR, NUP214 carries a coiled-coil domain which would encourage dimerization of the fusion protein.

Small molecule inhibitors targeting the ATP-binding domain of ABL1 show great efficacy in patients and to date, three ABL1 inhibitors, imatinib, nilotinib and dasatinib, have been FDA-approved for clinical use for the treatment of *BCR-ABL1*-fused cancers (McDermott, 2015). Pre-clinical studies suggest that *NUP214-ABL1*-fused cancers would similarly benefit from treatment with ABL1-inhibitors (Deenik et al., 2009; Quintás-Cardama et al., 2008).

Again, our data recapitulates previous literature. The 11 cell lines, 9 of which are chronic myeloid leukaemias, that were identified to carry the BCR-ABL1 gene fusion are also significantly sensitive to multiple ABL-inhibitors including imatinib, ponatinib and dasatinib (all at $p < 0.001$ and FDR < 0.01%; Figure 4.7). Moreover, the same drug compounds are significantly associated to the same thresholds with the *NUP214-ABL1* fusion detected in two T-lymphoblastic leukaemia cell lines, despite the small sample size (Figure 4.7).

4.3.1.3 EWSR1-FLI1 fusions

The fusion of Ewing's sarcoma breakpoint region 1 (*EWSR1*/*EWS*) to Friend leukaemia virus integration site 1 (*FLI1*) is a characteristic fusion of Ewing's sarcoma, a childhood soft tissue tumour with poor prognosis (Janknecht, 2005). In this fusion, the C-terminal DNA-binding domain of FLI1 is fused to the N-terminal transcriptional activation domain of EWSR1, which constitutively activates a transcriptional program that leads to transforming and tumorigenic activity *in vivo* and *in vitro* (Janknecht, 2005). While the *EWSR1-FLI1* fusion is prevalent in about 80% of Ewing's sarcomas, alternative fusions of *EWSR1* to other members of the ETS gene family to which *FLI1* belongs are also observed,

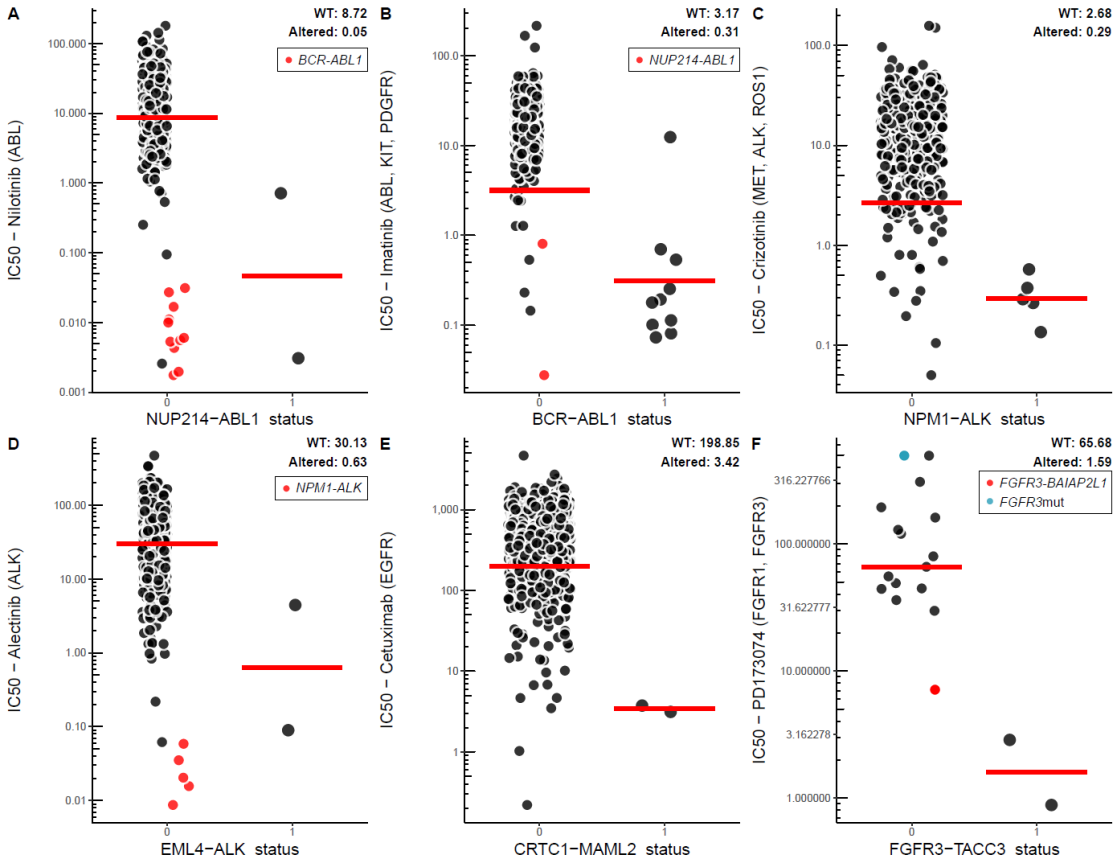


Figure 4.7: Known associations of fusions and drugs are recapitulated well, such as in the cases of (A) NUP214-ABL1 and (B) BCR-ABL1 with ABL1 inhibitors nilotinib and imatinib, and (C) NPM1-ALK and (D) EML4-ALK with ALK-inhibitors crizotinib and alectinib. Similarly, associations with small sample sizes are also identified, e.g. for CRTC1-MAML2 and cetuximab and FGFR3-TACC3 and PD173074 (latter only showing bladder carcinoma cell lines).

and found to activate transcriptional activity through the conserved DNA-binding domain in a similar fashion (Janknecht, 2005).

Previous high-throughput drug screens from our lab showed that cell lines carrying a *EWSR1-FLI1* fusion are highly sensitive to inhibition of poly(ADP-ribose) polymerases (PARP), a component of the DNA-damage response pathway to repair single-strand DNA breaks (Garnett et al., 2012). Although *in vivo* experiments show efficacy of the drug in mouse xenografts (Brenner et al., 2012), clinical trials testing PARP inhibitors in Ewing's sarcoma patients have so far failed to show efficacy (Choy et al., 2013). Since then, follow-up studies showed that using the DNA alkylating agent temozolomide in combination with PARP-inhibitors may increase PARP-trapping and enhance cytotoxicity *in vitro* and *in vivo* (Gill et al., 2015; Murai et al., 2014; Stewart et al., 2014). Several clinical trials are currently underway to test the combination in Ewing's sarcoma patients (NCT02044120, NCT01858168 and NCT02116777).

Our cell line panel exhibit 24 detected cases of *EWSR1-FLI1* and 4 cases of *EWSR1-ERG* fusions. Of those, the majority is found in Ewing's sarcoma cell lines, however I observe isolated cases with the same conserved breakpoints also in SK-NEP-1, a kidney carcinoma (*EWSR1-FLI1*), TASK1, a cell line of peripheral nervous system origin (*EWSR1-FLI1*) and EB-3, a Burkitt's lymphoma (*EWSR1-ERG*) cell line. SK-NEP-1 and TASK1 were unfortunately not screened as part of the high-throughput screening process, but EB-3 showed no enhanced sensitivity to PARP-inhibition (Figure 4.8A), which suggests that the sensitivity may be specific to Ewing's sarcomas carrying the fusion.

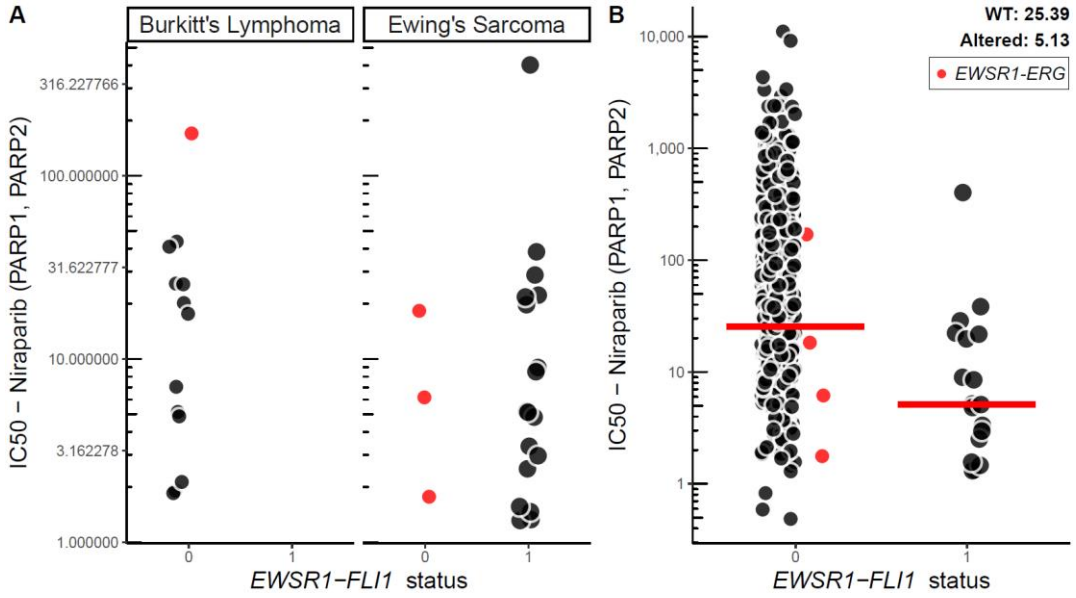


Figure 4.8: (A) The Burkitt's lymphoma cell line EB-3 does not respond to PARP-inhibition despite harbouring the EWSR1-ERG fusion. (B) The association of EWSR1-FLI1 fusion status is significant in a t-test ($F = 37.8$, $p < 0.001$), but not when tested in our ANOVA model that uses tissue type as a covariate (Tissue type: $F = 13.1$, $p < 0.001$; EWSR1-FLI1: $F = 0.08$, $p = 0.77$).

It is worth noting that unlike in the original screen, where *EWSR1-FLI1* and PARP-inhibition showed a highly significant association, I do not reproduce the statistical significance in my analysis. This is due to a slight change in the tissue type input into the model. Ewing's sarcoma cell lines in the previous study were part of a set of 300 cell lines whose tissue type was uncurated. For my analysis, I compiled an updated annotation that extended the number of tissue types from 28 to 42, and created a tissue type specifically for Ewing's sarcoma. Using this input, although a t-test of *EWSR1-FLI1* with Olaparib outputs a highly significant result ($F = 37.8$, $p < 0.001$), the ANOVA model attributes most of the variance due to tissue type (Tissue type: $F = 13.1$, $p < 0.001$; *EWSR1-FLI1*: $F = 0.08$, $p = 0.77$) (Figure 4.8B).

4.3.1.4 *CRTC1-MAML2*

Two of our cell lines also carry a *CRTC1-MAML2* fusion, which has a known transforming activity and was originally described in mucoepidermoid carcinomas (Enlund et al., 2004). The fusion brings together a CREB-binding domain in *CRTC1* and a transcriptional activation domain of *MAML2* and activates constitutive transcription of CREB downstream genes (Wu et al., 2005). One of the downstream targets of CREB-transcription is amphiregulin (AREG), a ligand for EGFR. Overexpression therefore leads to constitutive EGFR activation, and thus aberrant growth and survival signalling with decreased apoptosis and increased proliferation and metastasis signalling (Chen et al., 2014).

As has been demonstrated in previous studies (Chen et al., 2014), both cell lines with *CRTC1-MAML2* fusions show significant sensitivity to EGFR-inhibition in my analysis (Figure 4.7).

4.3.1.5 *FGFR3-TACC3*

We call three fusions that involve an N-terminal part of FGF receptor 3 (FGF). All three are bladder carcinomas; SW780 harbours a fusion of *FGFR3* to BAI1 associated protein 2 like 1 (*BAIAP2L1*), while in RT-112 and RT4 *FGFR3* is fused to transforming acidic coiled-coil containing protein 3 (*TACC3*). The fusions in the exact same three cell lines were previously characterised to demonstrate FGFR3 fusions as an alternate means of activating FGFR3 in bladder cancers, where overexpression and activating mutations are common (Williams et al., 2013). Since then, evidence for *FGFR3-TACC3* fusions have since been found in patient populations (Nassar et al., 2018) as well as other cancer types, such as glioblastoma and head and neck carcinomas (Singh et al., 2012).

Our analysis shows a highly significant association of between the *FGFR3-TACC3* fusion and sensitivity to the FGFR3-inhibitor PD173074, which is in line with results from previous studies (Singh et al., 2012). Aside from the *FGFR3-TACC3*-fused cell lines, SW780 shows the third-highest sensitivity to PD173074 (Figure 4.7).

In my analysis, six fusions (*NPM1-ALK, EML4-ALK, BCR-ABL1, NUP214-ABL1, CRTC1-MAML2* and *FGFR3-TACC3*) involving the known fusion oncogenes described above represent 67 of the 227 significant large-effect associations (30%). This is largely due to the large number of ALK (n = 8) and ABL-inhibitors (n = 10) screened in our high-

throughput screening set and the variations of partner genes with the same central fusion oncogene. The results therefore demonstrate that my analysis is able to identify clinically relevant fusion-drug associations.

4.3.2 Novel associations

Disregarding the 67 associations known *a priori*, I find a strongly depleted landscape of significant interactions. Further removing from considerations fusions that have never been identified in a patient sample, those that are not in-frame, excluded fusions that are unlikely to be causative and those that are unlikely to be relevant for clinical use. This retains 35 novel associations, of which 24 are sensitising and 11 are resistant (Figure 4.9). The remaining fusion-drug associations are listed in (Table 4-1).

Upon closer inspection, I find that 7 fusion-drug associations show potentially confounding factors. For instance, the two cell lines, MFM-223 and KATOIII, which carry a *FGFR2-ATE* also carry duplication of *FGFR2* (amplification > 14) (Figure 4.9B) (Forbes et al.,

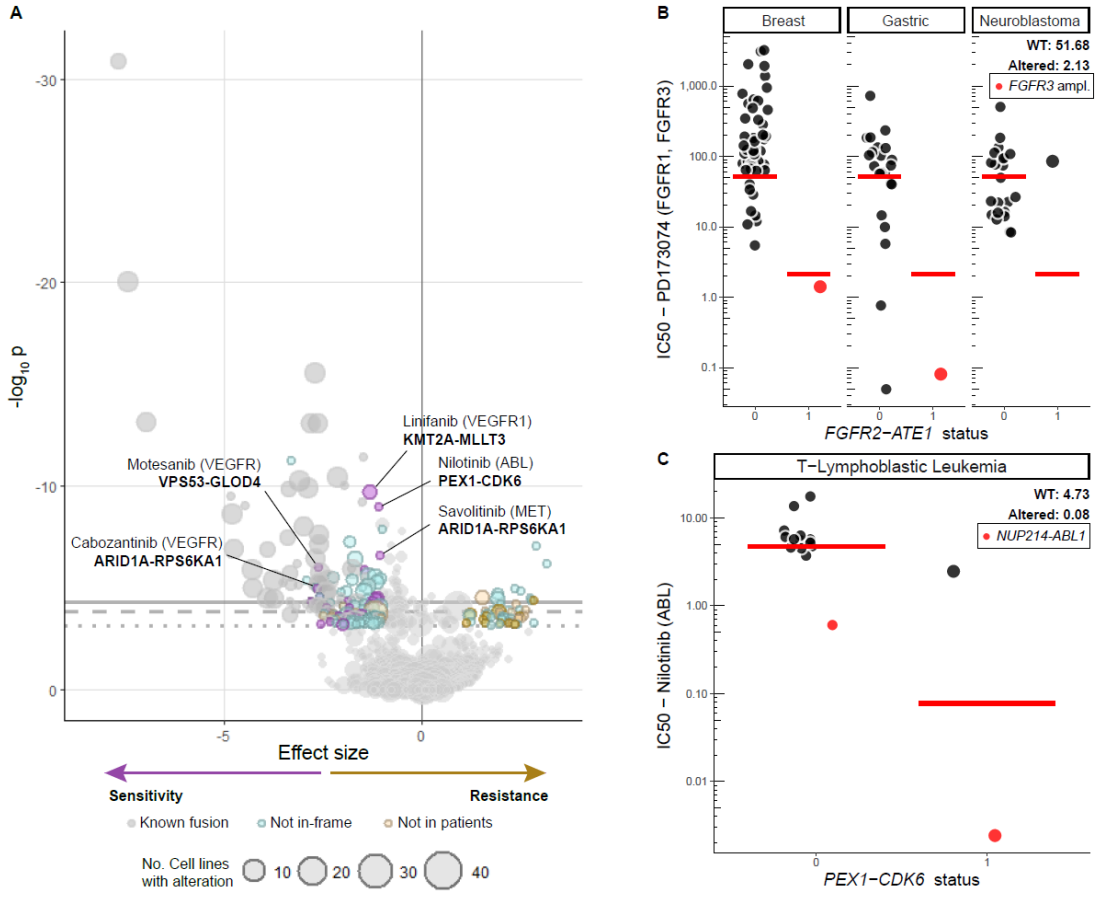


Figure 4.9: (A) Volcano plot of significant fusion ANOVA interactions after annotating known (grey), not-in-frame (blue) and non-patient fusions (beige). Often, associations have confounding factors, as in the case of (B) FGFR2-ATE with FGFR2-inhibitor and (C) PEX1-CDK6 with ABL inhibitor.

2017), which is more likely to explain their sensitivity to FGFR2-inhibitors. In the association of *PEX1-CDK6* and the ABL-inhibitor nilotinib, one of the two cell lines with the fusion, ALL-SIL, carries a *NUP214-ABL1* fusion (Figure 4.9C). In a similar vein, the association of *STK24-DOCK9* and EGFR inhibitors Gefitinib and Afatinib is confounded by an EGFR-amplification in 1 of the two cell lines. One of the cell lines harbouring the *ARID1A-RPS6KA1* fusion sensitive to various MET inhibitors also carries a *MET* amplification.

KMT2A-MLLT3

Of the remaining novel associations, those involving *KMT2A-MLLT3* were significant across four distinct drugs, which target a range of tyrosine kinases, but all share VEGFR inhibition (Figure 4.10). Translocations of *KMT2A,* also known as *MLL*, with over 60 different fusion partners occur in about 5%-10% of acute leukaemias (Meyer et al., 2009). *KMT2A-MLLT3* (also known as *MLL-AF9*) fusions make up about 8.5% of *MLL* fusions (Meyer et al., 2009). Unlike most tumours with *KMT2A*-translocations that are associated with aggressive

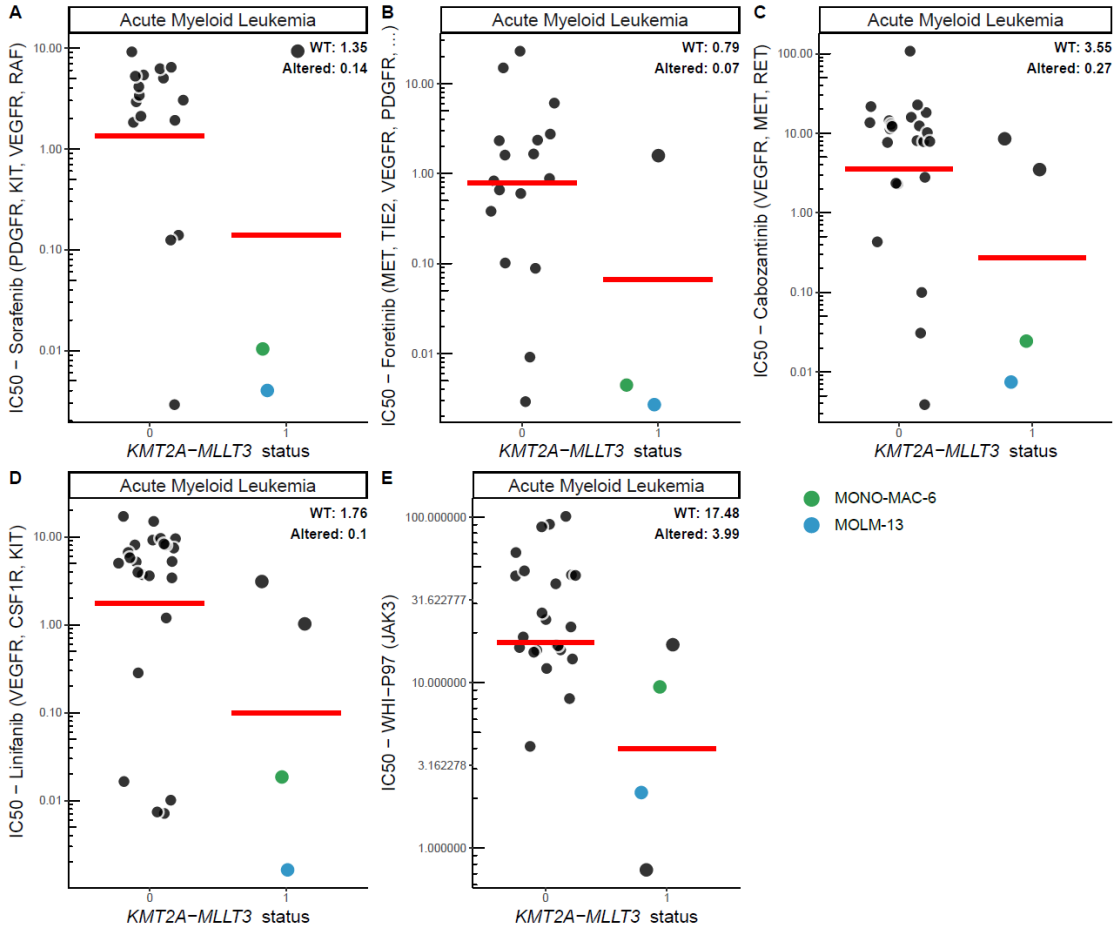


Figure 4.10: Significant associations between the KMT2A-MLLT3 fusion and various VEGFR-inhibitors.

disease and poor prognosis, tumours with *KMT2A-MLLT3* fusion have a favourable outcome (Grimwade et al., 2010).

Upon closer inspection of the associations of *KMT2A-MLLT3*, it appears that of the 4 acute myeloid leukaemia cell lines harbouring the fusion, only two are highly sensitive across all drugs: MONO-MAC-6 and MOLM-13. In the association of *KMT2A-MLLT3* and JAK3-inhibitor WHI-P97, the two most sensitive cell lines are MOLM-13 and THP-1.

Vascular endothelial growth factor receptors (VEGFR) are typically targeted for its role in endothelial cell proliferation and angiogenesis (Ferrara et al., 2003). Although VEGFR expression has shown to be a factor in leukaemia development (Song et al., 2012), my literature search did not reveal any known associations of KMT2A activity with VEGFR. Since only a subset of *KMT2A-MLLT3* carrying fusions exhibit heightened sensitivity, the sensitivity may present itself due to other shared factors within the cell lines that may be unrelated to the *KMT2A-MLLT3* fusion.

Upon further inspection, I found no significant upregulation of the VEGF receptor fms related tyrosine kinase 1 (*FLT1*), fms related tyrosine kinase 1 (*FLT4*) and kinase insert domain receptor (*KDR*) in any of the 4 fusion-containing cell lines using voom-normalised gene expression data. I then looked for any fusions or CFEs which coincide only the $VEGFR_i$ responsive cell lines, but not of the non-responsive cell lines, but found none matching the criteria. Finally, I analysed the gene essentiality profiles derived from CRISPR/Cas9 data[3], but found that none of the guides targeting the three VEGFR genes were significantly depleted in MOLM-13 (no data available for MONO-MAC-6). The drug compounds may still lead to loss of viability by inhibiting several targets at once, however this finding strongly suggests that the sensitivity to $VEGFR_i$ does not result from the inhibition of an individual VEGFR protein.

Thus, I conclude that the interest of the association is inconclusive and that further research is needed to determine whether sensitivity to VEGFR inhibitors is truly enriched in a subset of *KMT2A-MLLT3* fused cell lines, or is caused by other non-determined confounding factors.

[3] I derived the binarised gene essentiality data using BAGELR from CRISPR/Cas9 whole genome drop-out screening data from Wang et al (Wang *et al.*, 2017). Further information on data processing to follow in chapter 0.

On another topic relating to *KMT2A-MLLT3*, previous studies show that DOT1 like histone lysine methyltransferase (*DOT1L*) is required in the tumorigenicity and cell proliferation of *KMT2A-MLLT3* fusions *in vivo* and *in vitro*, and sensitive to DOT1L inhibition using EPZ-5676 in mice and cell lines (AML cell lines NOMO-1 and MOLM-13) (Daigle et al., 2013; Nguyen et al., 2011). In contrast, my analysis finds no statistically significant association of *KMT2A-MLLT3* and any of the DOT1L inhibitors (EPZ5676, SGC0946 and EPZ004777), as most of the variation is attributed to the differential sensitivity of tissue types. As our screen also screens NOMO-1 and MOLM-13 for this interaction, a possible explanation to these opposing results may be related to differences in screen design. In particular, previous studies exposed cell lines to EPZ5676 for 14 days, while our high-throughput screen uses a 72-hour assay (Daigle et al., 2013).

Once interactions with confounding factors and those involving *KMT2A-MLLT3* are removed, 12 sensitising and 11 resistant hits remain. The fusions involved in the remaining associations have so far not been characterised and the respective partnering genes are largely unknown. At the same time, no particular patterns emerge in the drugs that they associate with, making calls of their clinical importance difficult at this stage.

With the information to be published into the public domain, with improvements in the knowledge of the role of the fusions proteins or the respective cancer pathways, the importance of the associations may be revealed at a future time point. As it stands so far, although our fusion ANOVA recapitulates prominent known fusion-drug sensitivities very well, novel findings are largely confounded by other factors or unable to be interpreted.

Table 4-1: Novel associations in fusion ANOVA. Conf: confounded. Descr: described in text.

Notes	FEATURE	DRUG_NAME	DRUG_TARGET	CL #	FDR	Type
	AFAP1--ABLIM2	GSK269962A	ROCK1, ROCK2	4	9.5	resistant
	AFAP1--ABLIM2	Vorinostat	HDAC inhibitor Class I, IIa, IIb, IV	3	11.9	resistant
	LDLR--SMARCA4	ZG-10	JNK1	2	13.2	resistant
	PLXND1--TMCC1	Epirubicin	Anthracycline	2	13.2	resistant
	PLXND1--TMCC1	Luminespib	HSP90	2	21.9	resistant
	RCOR1--TECPR2	Tanespimycin	HSP90	2	16	resistant
	SLC12A7--TERT	BX796	TBK1, PDK1 (PDPK1), IKK, AURKB, AURKC	2	24.7	resistant
	SWAP70--WEE1	Piperlongumine	Induces reactive oxygen species	2	4.8	resistant
	SWAP70--WEE1	PFI-1	BRD4	2	21.7	resistant
	UBQLN1--GKAP1	Bortezomib	Proteasome	2	21.9	resistant
	UBQLN1--GKAP1	Sabutoclax	BCL2, BCL-XL, BFL1, MCL1	2	22	resistant
(Conf.)	ARID1A--RPS6KA1	Savolitinib	MET	2	0.1	sensitive
(Conf.)	ARID1A--RPS6KA1	Cabozantinib	VEGFR, MET, RET, KIT, FLT1, FLT3, FLT4, TIE2,AXL	2	1.8	sensitive
	ARID1A--RPS6KA1	Bicalutamide	AR	2	3.7	sensitive
	ARID1A--RPS6KA1	BMS-345541	IKK1, IKK2	2	7	sensitive
	ARID1A--RPS6KA1	WIKI4	TNKS1, TNKS2	2	11.9	sensitive
(Conf.)	FGFR2--ATE1	PD173074	FGFR1, FGFR3	3	4	sensitive
(Descr)	KMT2A--MLLT3	Linifanib	VEGFR1, VEGFR2, VEGFR3, CSF1R, FLT3, KIT	4	0	sensitive
(Descr)	KMT2A--MLLT3	Cabozantinib	VEGFR, MET, RET, KIT, FLT1, FLT3, FLT4, TIE2,AXL	4	4	sensitive
(Descr)	KMT2A--MLLT3	Sorafenib	PDGFR, KIT, VEGFR, RAF	3	5.1	sensitive
(Descr)	KMT2A--MLLT3	Foretinib	MET, KDR, TIE2, VEGFR3/FLT4, RON, PDGFR, FGFR1, EGFR	3	6.5	sensitive
(Descr)	KMT2A--MLLT3	WHI-P97	JAK3	4	14.1	sensitive
	NKD1--ADCY7	KIN001-270	CDK9	3	12.2	sensitive
	NPLOC4--PDE6G	Cediranib	VEGFR, FLT1, FLT2, FLT3, FLT4, KIT, PDGFRB	2	5.1	sensitive
	NPLOC4--PDE6G	Ibrutinib	BTK	2	13.2	sensitive
	NPLOC4--PDE6G	BIX02189	MEK5, ERK5	2	15.1	sensitive
(Conf.)	PEX1--CDK6	Nilotinib	ABL	2	0	sensitive
(Conf.)	PEX1--CDK6	Nilotinib	ABL	2	6.5	sensitive
	QKI—PACRG	RO-3306	CDK1	3	22.7	sensitive
	SLC12A7--TERT	IOX2	EGLN1	2	4.9	sensitive
(Conf.)	STK24--DOCK9	Gefitinib	EGFR	2	21.9	sensitive
(Conf.)	STK24--DOCK9	Afatinib	ERBB2, EGFR	2	24.6	sensitive
	VPS53--GLOD4	Motesanib	VEGFR, RET, KIT, PDGFR	2	0.3	sensitive
	VPS53--GLOD4	Axitinib	PDGFR, KIT, VEGFR	2	0.3	sensitive
	VPS53--GLOD4	AZD7762	CHEK1, CHEK2	2	19.8	sensitive

4.4 Gene-centric ANOVA

Next, I implemented a gene-centric version of the ANOVA model. Here, I consider any gene which is fused at the same position (5′ or 3′) multiple times across the cell line panel. I am specifically looking for cases where oncogenes can be fused to multiple partner genes which can all lead to the same constitutive activation of the oncogene (e.g. through removal of auto-inhibitory domains or the fusion of any dimerization domain). An example of that would be how constitutive activation of ABL1 can be triggered by either a *BCR-ABL1* or a *NUP214-ABL1* fusion, or an activation of EWSR1 by either *EWSR1-FLI1* or *EWSR1-ERG* as described above (section 4.3.1).

To achieve this, I use a similar model as for the fusion ANOVA, where the final term is substituted to $Gene_G$:

$$IC50_D \sim Tissue\ Type + MSI\ Status + CE_D + Gene_G$$

Equation 3: Model for gene-centric ANOVA.

4.4.1 Input data

Note, that I take into account the direction of the gene fusion. For example, I find 9 different fusion constructs involving the *ALK* gene (brackets represent number of cell lines): *ALK-AC006486.9* (1), *ALK-ERF* (1), *NPM1-ALK* (5), *AC016907.3-ALK* (6), *TERT-ALK* (1), *ALK-AC016907.3* (1), *EML4-ALK* (2), *ALK-PTPN3* (1) and *TFG-ALK* (1).

Thus, the model would run two variations of *ALK*: 1) 3′-ALK would contain *NPM1-ALK*, *AC016907.3-ALK*, *TERT-ALK*, *EML4-ALK* and *TFG-ALK*, and 2) 5′-ALK would contain the remaining constructs.

In total, I used data for 941 genes, of which 98 are tested in both 5′ and 3′ positions, for a total of 1,039 features to be tested against 409 drug IDs. There were two criteria for inclusion in the gene-centric ANOVA: 1) the gene must be fused in at least two distinct fusion constructs and 2) fusions containing the gene must be found in at least 3 different cell lines.

The majority of gene inputs consist of 3 distinct fusions and are present in 3 different cell lines (Figure 4.11). Genes represented in more than 4 distinct fusions disproportionally tend to be fused at the 5′. The three genes involved in the highest number of distinct fusions transcripts are Pvt1 oncogene (*PVT1*), regulator of chromosome

condensation 1 (*RCC1*) and neuroblastoma amplified sequence (*NBAS*), with 44, 22 and 21 unique fusion constructs, respectively.

4.4.2 Overview of results

Of 424,951 potential combinations, 563 passed our previous criteria for large-effect significant associations (FDR < 25% and Glass Delta > 1), with 281 sensitising and 282 resistant associations. Due to the larger set of input features, I decided to implement a stricter FDR threshold, raising it from 25% to 5%. Implementing this, we maintain 159 associations, with 110 sensitising and 49 resistant (Figure 4.11) (Supplementary Table 7).

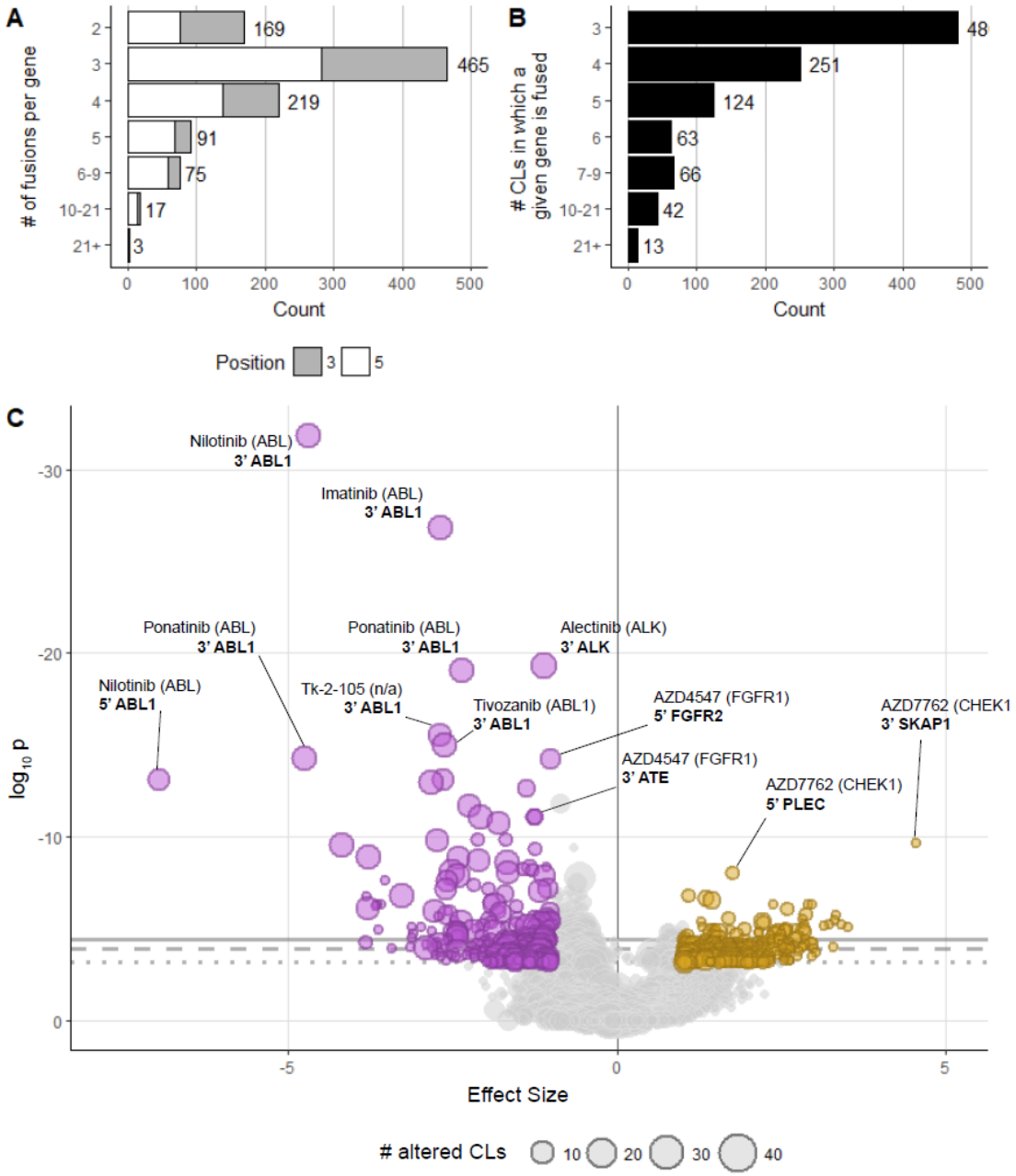


Figure 4.11: Volcano plots illustrating the results from the gene-centric ANOVA.

4.4.3 Known gene-drug associations

The associations with the lowest p-values are dominated by known associations, as the ones described in the fusion ANOVA (Figure 4.12).

4.4.3.1 ABL1

Our cell line panel contains only two types of *ABL1*-fusions with *ABL1* at the 3': the well-understood *BCR-ABL1* and *NUP214-ABL1*. Unsurprisingly, grouping the two fusions into one variable resulted in highly significant associations ($p < 0.001$, fdr < 1%) across 8 different ABL inhibitors, including FDA-approved nilotinib (Figure 4.12A).

4.4.3.2 BCR

The most prevalent partner gene in the *BCR-ABL1* fusion is fused multiple times at the 5' end. Associations of 5'-*BCR* and ABL1 inhibitors are highly significant, however when I manually inspect the drug sensitivities, it becomes apparent that the association is mainly driven by the *BCR-ABL1* fusion. Other *BCR*-fusions, e.g. *BCR-ABHD2* and *BCR-PI4KA*, show IC_{50}'s similar to the median of non-fused samples. None of the other *BCR* fusions have previously been reported in patient samples. Running the gene-centric ANOVA model with a covariate for *BCR-ABL1* status, the association of the remaining *BCR*-fused genes becomes non-significant ($p = 0.74$).

4.4.3.3 NUP214

Another common partner gene to *ABL1*, *NUP214* is fused to several other 3' partners, such as N-terminal Xaa-Pro-Lys N-methyltransferase 1 (*NTMT1*), PBX homeobox 3 (*PBX3*), proline rich coiled-coil 2B (*PRRC2B*) and XK related 3 (*XKR3*). 5'-*NUP214* is significantly associated with nilotinib, however as in the previous case, this significance is mainly driven by the *NUP214-ABL1* fusion. Only three cell lines of the six cell lines with 5'-*NUP214* fusions were screened under this drug ID. Of those, as expected, the two *NUP214-ABL1* cell lines are sensitive to drug inhibition. While K-562, a chronic myeloid leukaemia cell line with a *NUP214-XKR3* fusion appears to be sensitive, it also harbours a *BCR-ABL1* fusion, which is more likely to drive sensitivity to nilotinib.

Further, 5'-*NUP214* tested for sensitivity to ABL inhibitors across a larger set of fusion-containing cell lines do not show significant associations. For example, imatinib was screened also in a *NUP214-PBX3*-containing Melanoma cell line, where it shows no difference in IC_{50} from non-fused cell lines. In conclusion, 5'-*NUP214* fusions are highly unlikely to cause a sensitivity to ABL inhibitors outside of the context of the *NUP214-ABL1*

fusion. Neither *NUP214-XKR3* nor *NUP214-PBX3* have previously been reported in patient samples.

4.4.3.4 ALK

As described above (section 4.3.1), our cell line panel includes a two distinct recurrent fusions involving the tyrosine receptor kinase ALK. While the majority of blood cancers with *ALK*-fusions carry the *NPM1-ALK* fusion, a smaller subset is fused to other partner genes that encode coiled-coil domains that promote dimerization and activation of ALK's tyrosine receptor kinase domains. One example is that of a fusion with Tyrosine receptor kinase-fused gene (TFG), a protein with preserved N-terminal coiled-coil domains (Greco et al., 1998; Hernández et al., 1999).

Our cell line panel includes one such instance of a *TFG-ALK* fusion in SCC-3, a B-cell lymphoma. Other *ALK*-fused constructs include *AC016907.3-ALK*, and *TERT-ALK*, neither of which were previously described in tumours. In our drug screen, SCC-3, but not the other two fused-cell lines, shows a sensitivity to alectinib that is comparable to *NPM1-ALK* and *EML4-ALK* fused cell lines (Figure 4.12C).

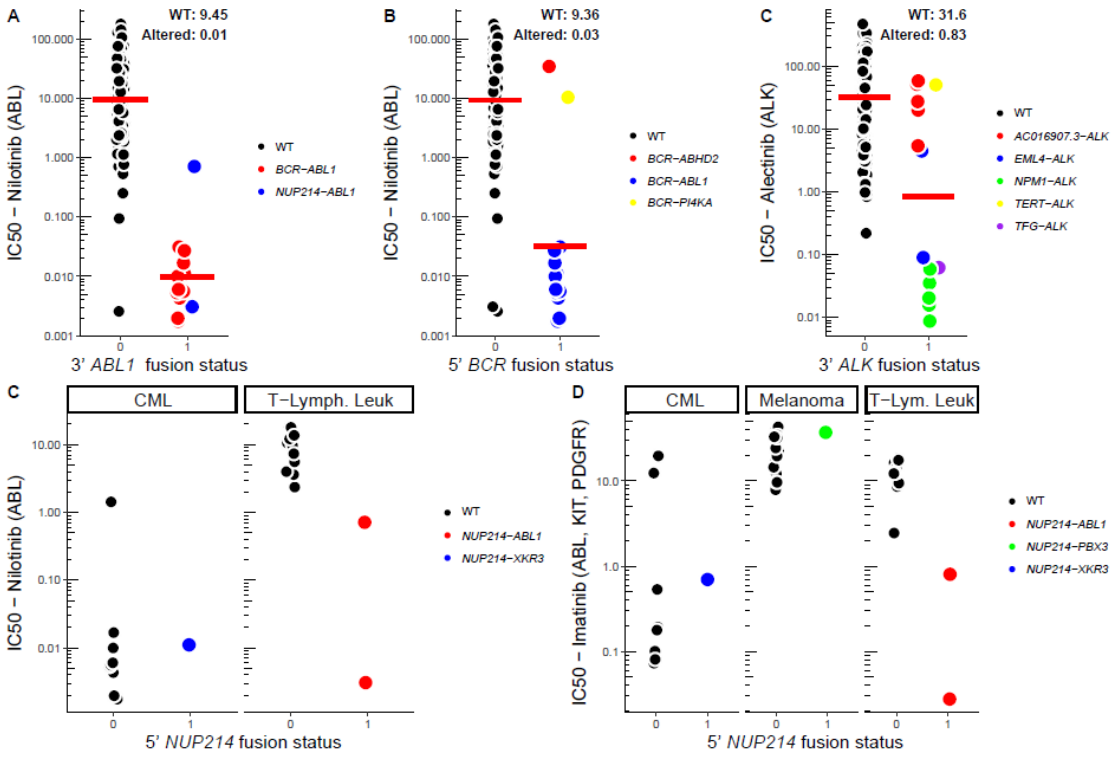


Figure 4.12: (A-D) Known examples in the gene-centric ANOVA.

4.4.3.5 *FGFR3*

Sensitivity of *FGFR3*-fused cell lines to FGFR-inhibitors is described in section 4.3.1. As expected, the association of 5'-*FGFR3* fusions with FGFR-inhibitor PD173074 is highly significant in this analysis.

4.4.4 Confounding factors

4.4.4.1 *FGFR2*

Similarly as in the fusion ANOVA, *FGFR2* amplifications are a confounding factor in the gene-centric ANOVA. Ten cell lines show FGFR2 fusions, of which 5 cell lines also have a copy number gain (PICNIC score = 14) at the segment containing FGFR2. When comparing the activity of the FGFR-inhibitor AZD457, I find that the amplification neatly divides responding verses non-responding cell lines, regardless of the fusion status (Figure 4.12).

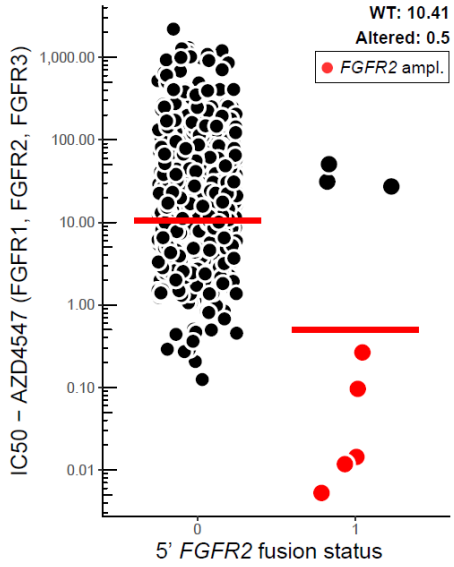


Figure 4.13: FGFR-amplification separates cell lines that respond to FGFR-inhibitors from those that do not, regardless of fusion status.

4.4.4.2 *ERBB2*

Another common confounding factor is that cell lines with amplification of ERBB2 are associated with ERBB2-inhibitors such as afatinib. In our data-set, associations with afatinib sensitivity are found with cell lines with ERBB2-fusion, which like in FGFR2 are likely to be a side-effect of ERBB2-amplifciations causing passenger rearrangements and translocations.

Aside from those, afatinib sensitivity is also associated with 3′ IKAROS family zinc finger 3 (*IKZF3*) and 3′ titin-cap (*TCAP*) fusions. Both genes are less than 50,000 bp down or up-stream of ERBB2. As cell lines with the fusions all carry an ERBB2 amplification, it is likely a passenger event from the amplification.

Altogether, the above events demonstrate how fusions in commonly amplified oncogenes can be correlated with a drug sensitivity but not causative, due to co-occurrence with an oncogenic event. Further, these passenger fusions occur commonly not only at the oncogene whose amplification is beneficial to tumour growth, but also in the surrounding genomic regions.

Events such as these pose more questions on the frequency of passenger fusion events compared to cancer-driving fusion events and how these may be distinguished from high-throughput screening.

4.4.5 "Novel" findings

4.4.5.1 PTEN

5′-*PTEN* fusions, found in breast, prostate and biliary tract cell lines, are associated with resistance to tankyrase inhibitor AZ6102 (Figure 4.14). Here, *PTEN*-rearrangement is likely to be a mechanism for disruption of the tumour suppressor gene. A previous study show that tankyrases ADP-ribosylate PTEN to be promoted for degradation and thus tumour growth. Knockdown of tankyrases stabilises PTEN and resulted in downregulation of the PI3K/AKT pathway and thus tumour suppression (Li et al., 2015). In cells where PTEN is disrupted through a rearrangement, tankyrase inhibition is thus less likely to lead to cell death. However, cell lines with *PTEN* mutations or deletions appear to have similar sensitivity to the tankyrase inhibitor as wild type cells (Figure 4.14). This suggests that simple deletion of *PTEN* through fusion events may not be the sole mechanism of tankyrase inhibitor resistance in fused cell lines.

4.4.5.2 FBXW7

5′ F-box and WD repeat domain containing 7 (*FBXW7*) fusions are another feature associated with resistance to another tankyrase inhibitor, WIKI4. The fusions are present in

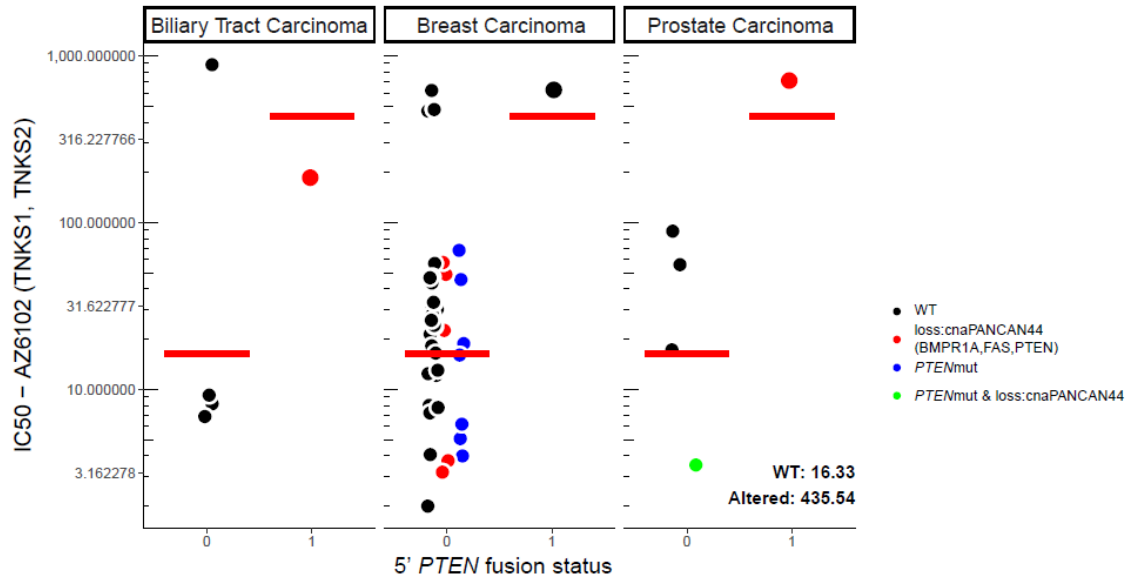


Figure 4.14: PTEN-fusion and sensitivity to the tankyrase inhibitor AZ6102.

breast, cervical and head and neck cell lines and do not co-occur with *PTEN*-fusions. In general, *FBXW7* targets several known cancer-drivers for ubiquitination and subsequent degradation, such as Yes associated protein 1 (YAP1) and Aurora kinases (Tu et al., 2014). Interestingly, another study showed the cooperation of *FBXW7* and *PTEN* in tagging Aurora kinases for ubiquitination (Kwon et al., 2012). It would be interesting to note in the future whether PTEN stabilisation following tankyrase inhibition provides survival benefits related to *FBXW7* activity.

Altogether, after removing the associations that are either known, confounded or described above, 78 significant large-effect associations remain in the list (Supplementary Table 7). Further curation efforts are complicated by the lack of prior knowledge of the function of the associated genes that could help to establish any causative links with the associated drug.

As I showed in this analysis as the one before, confounding variables unknown to the model can further complicate the task of differentiating true fusion-drug associations from false-positives, making a manual curation effort based on literature research difficult at this point in time.

4.5 Chapter discussion & conclusion

In this chapter, I sought to discover novel fusion biomarkers of drug response using our high-throughput screening data of 409 drug IDs. Any results from such an analysis uncover essential pathways in cancer cell lines that may be regulated by gene fusions. Moreover, these results may provide a basis for further research into the opportunities of screening for novel gene fusions in the clinic to provide improved treatment opportunities, along the same veins as the successful previous case studies of *ABL1* and *ALK*-fusions.

While my analyses reliably recapitulated known and expected findings, after curation of novel associations, no strong candidates emerge that warrant a recommendation for further study at the current point. Most potential associations that are ruled out fall into one of two categories: 1) the presence of confounding factors may explain drug sensitivity through fusion-unrelated mechanisms. 2) A lack of prior knowledge in low-recurrence fusions puts into question the robustness of the association, as well as their relevance in patient populations.

Despite a lack of emerging candidates, the analyses do provide learnings that may become useful when setting up future analyses and drug-screens.

Low recurrence give low statistical power

Truly recurrent fusions are rare in our dataset. For comparison, our group's previously published analysis of high-throughput drug-screening data used 717 cancer mutations and copy number alterations, of which the vast majority was present in at least 10+ cell lines (Figure 4.15A) (Iorio et al., 2016). However, in my analysis even with over 8,000 fusions detected in the 1,011 cancer cell lines, only a minority of fusions are recurrent, and even less in more than two cell lines (Figure 4.15B). Moreover, the list of 717 CFEs in Iorio *et al.* was curated to include only patient-relevant events, while our fusion analysis was not curated to that extent due to a lack of comprehensive databases for the annotation of fusion genes described in patients.

As recurrent fusions are not only rare, but also uncurated, the low numbers of novel findings are perhaps unsurprising. Due to the rarity of recurrent fusions, it perhaps becomes even more important to be able to separate potentially functional gene fusions from passenger events. For example, *ALK*-fusions are sensitive to ALK-inhibitors only in a

subset of detected constructs. For gene fusions of unknown function that may be less recurrent, being able to pick out the functional constructs may be the difference between a significant association and a false negative for any drug-associations tested in a gene-centric ANOVA.

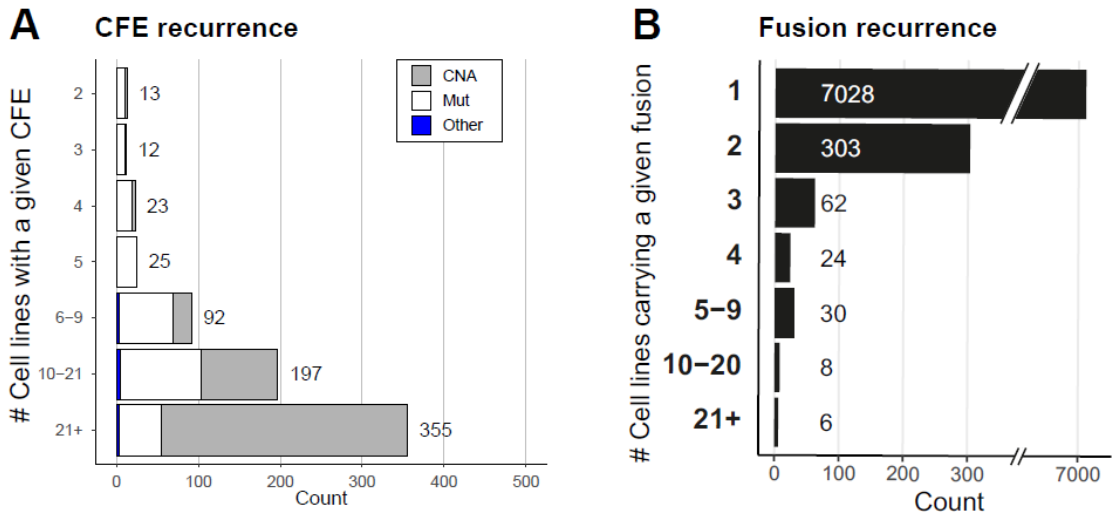


Figure 4.15: (A) Recurrence of 717 curated CFEs and (B) 8,388 fusions in our panel of 1,011 cell lines.

Confounding CFEs complicate interpretation of associations

What further complicates interpretation is the frequency of confounding events whose occurrences are linked to those of fusions. For example, *ERBB2* amplified cell lines often carry fusions of genes at the same genomic region, which are in turn significantly associated with ERBB2-inhibitors. This is a problem that the inclusion of significant hits from the CFEs ANOVA (4.2) was designed to solve. Unfortunately, in the event of *ERBB2* amplification, the CFE failed the criteria for large-effect interaction due to the Glass delta of *ERBB2*-amplified cell lines being 0.8. The threshold was originally set following the example in Iorio, *et al* (2016), and is necessary as the inclusion of too many covariates can significantly decrease significance of even true associations. However, the frequency of *ERBB2*-amplification being a confounding variable shows that perhaps future iterations of this model should re-analyse thresholds for a broader inclusion of potential confounders.

Similarly, *FGFR2*-amplifications were another common confounder in my analysis. Unfortunately, *FGFR2*-amplifications were not a feature included in the curated binary CFE annotation upon which the CFE ANOVA was based, and thus not flagged to be included as a covariate in the fusion ANOVA. Thus, further identifying and annotating this and other

unknown confounding factors could contribute significantly to the robustness of future iterations of this analysis.

Lack of prior knowledge in fusion selection

As implied in the previous paragraph, lack of prior knowledge of fusion function can complicate model input decision and interpretation of novel significant associations.

Firstly due to a community-wide lack of understanding on how to distinguish functional from non-functional fusions resulting from high-throughput transcriptome sequencing data, the data-set still contains a lot of noise. Unlike point mutations for which nowadays there are meticulous filters to determine the likelihood of causing functional disruption, there are no comprehensive resources available for gene fusions at this point. Though breakpoint predictors such as GRASS are available to predict the genomic location of fusions (e.g. intron vs exonic, UTR vs coding) and their likelihood of result in in-frame gene products, there are still shortcomings. For one, if we assume that passenger fusions occur at random frequencies across the genome, a third of fusions in coding regions would be expected to be in-frame. Further, gene fusions may exert functionality beyond the production of chimeric proteins, for instance a fusion of a UTR to a coding region may lead to differential regulation due to post-transcriptional regulatory events that affect splicing or mRNA stability. In other cases, a gene fusion may be functional not through the chimeric protein produced but due to a disruption of a tumour suppressor gene, as the case of 5'-*PTEN* fusions suggest. Another draw-back of the GRASS algorithm is that it depends on the quality of genomic annotations that is used as an input file. For instance, the two well-studied oncogenic fusions *EIF3E-RSPO* and *UBE2L3-KRAS*, are flagged with a GRASS score of 100, defined as a "something else to something else" fusion. However when I manually aligned breakpoints and reviewed literature, I found strong evidence supporting that the fusion should be functionally in-frame.

Due to the examples described above, I conclude that while the GRASS algorithm is a useful tool for annotation of fusions, I would not recommend it as a filtering method to exclude any fusion events.

Available drug compounds are biased towards well-studied oncogenic targets

Most targeted compounds that are included in our high-throughput screen were developed as a response to known oncogenic events. For instance, the development of ABL and ALK-inhibitors was sparked by the discovery of the oncogenic activity of the ABL

and ALK-fusions. As fusion-discovery using high-throughput transcriptome analyses is still a recent endeavour, and due to the rarity of recurrent fusion events, currently unknown rare oncogenic fusions may yet to be discovered in the future.

At the same time, it is important to keep in mind that due to a lack of hydrophobic pockets or their subcellular localisation, many appealing molecular targets are currently considered undruggable (Verdine and Walensky, 2007).

Thus, the lack of targeted drugs for potential unknown fusions with oncogenic activity may present a blind spot in this type of analysis of high-throughput drug screens.

In conclusion, my analysis recapitulates current knowledge on fusion-drug associations. That no strong candidates emerge among novel associations is somewhat surprising considering the successes of previous high-throughput drug-screening efforts in identifying biomarkers from oncogenic mutations and copy number alterations. I speculate that the lack of novel fusion biomarkers is mostly due to 1) the relative lack of novel recurrent fusions leading to reduced statistical power and 2) limited prior knowledge on fusion genes complicating interpretation of hits.

Future developments may yield a better understanding of the characteristics of functional fusions and broaden the currently limited landscape of gene fusions in patient samples. This could then be used to filter out noise in the form of non-functional passenger fusions and increase statistical power and confidence for the remaining fusions. Similarly, advances in drug development may improve the breadth of pathways and targets available. Due to the low recurrence of fusions, expanding the number of models available can also increase the statistical power of the biomarker analysis.

At that point, a similar statistical model querying high-throughput drug-screening data may indeed yield valuable insight into therapeutic sensitivities of fusion genes and aid the targeted treatment of cancer patients in the future.

5 Functional analysis using CRISPR/Cas9 whole-genome screening data

5.1 Introduction and concept

The topic of utilising CRISPR/Cas9 whole genome screening data has been split across two different chapters for ease of reading. This chapter will provide an introduction to the concept, the methodology of the computational approach and some quality control analyses. After that, the next chapter will discuss the results of the approach in more detail, putting the focus onto discovering functional gene fusions.

5.1.1 State-of-art and current challenges

One of the biggest challenges that high-throughput transcriptome analyses face is how to separate the noise from the signal. In the age of genomics, cancer genes and specific oncogenic mutations are commonly identified because they are mutated at a higher rate than the background. With now tens of thousands tumour genomes sequenced, statistical models identifying such mutation hotspots have helped us build comprehensive landscapes and databases of cancer driver mutations (Barretina et al., 2012; Forbes et al., 2017).

However, as I discussed in the previous chapter, recurrent fusions are strikingly rare. Even though more than 10,000 patient samples have already undergone fusion calling, analytical methods can still only identify the handful of known recurrent fusions. Similarly, even recurrent fusions may represent passenger events of common genetic alterations and need not necessarily be functional. Known examples of this are the *FGFR2* and *ERBB2* fusions described in the previous chapter, but unknown examples may further confound our goal to identify truly oncogenic fusions.

Innovative solutions that step away from using recurrence to identify cancer driver fusions are necessary in order to overcome these challenges. Here, I developed a systematic method to query the functionality of unknown fusions using existing CRISPR/Cas9 whole-genome screening data.

5.1.2 Review of CRISPR/Cas9 whole-genome drop-out screening

CRISPR/Cas9 whole genome drop-out assays are designed to identify genes that contribute to cellular fitness, i.e. genes that are necessary for the normal cell function and

survival. The cell line is first modified to express functional CRISPR associated protein 9 (Cas9), which has the ability to introduce double-strand breaks in a genomic region that is defined by a sequence given on a short guide RNA (sgRNA). Then, a whole-genome guide library is virally transduced into a population of Cas9-expressing cells, so that on average one cell will carry one sgRNA. Upon encountering a sgRNA, Cas9 will introduce a double-strand break at the target site that is complementary to the sgRNA sequence. The double-stranded break is repaired using non-specific non-homologous repair mechanisms that are extremely prone to introducing insertion/deletions which effectively knock-out the target gene (Behan et al., 2018; Meyers et al., 2017; Wang et al., 2017). The guide library will include several sgRNAs targeting each of a vast majority of known genes in the human genome. Typically, whole-genome guide libraries target ~17,000-20,000 genes with 4-10 sgRNA for each gene. After keeping the cell lines in culture for several days, the sgRNA in the population is sequenced to reveal the proportion of remaining sgRNAs. sgRNA that target genes required for cellular fitness should be depleted from the population due to the deleterious effect of the CRISPR system. Depleted genes can be identified using several statistical models, with the most prevalent including MAGeCK and BAGEL (Hart and Moffat, 2016; Li et al., 2014).

5.1.3 Applying CRISPR/Cas9 screening data to identify functional gene fusions

Due to sometimes unpredictable sgRNA efficacy and off-target effects depending on the target and PAM sequences (Doench et al., 2014; Munoz et al., 2016), each gene will usually have 4-10 sgRNA spread across the target gene. The drop-out values are typically then averaged across the sgRNA to produce final essentiality scores for each gene.

My method takes advantage of the multiple sgRNAs per gene to create a fusion-specific CRISPR scoring system. For each fusion gene, I map the sgRNAs of both fusion partners onto the predicted breakpoint of the gene fusions. I then categorise sgRNAs that map onto the predicted fusion-gene as "mapping" and those that fall outside the fused-region as "non-mapping". The key concept is that for essential functional fusions, CRISPR/Cas9 generated double strand breaks should be lethal at the mapping regions, but have no effect at non-mapping regions. Conversely, if the fusion is just a passenger event, then disrupting the gene fusion should not yield any essentiality. Similarly, where the normal function of a gene is essential rather than the fusion, I expect to see depletion of not only mapping- sgRNA but also of non-mapping sgRNA (Figure 5.1).

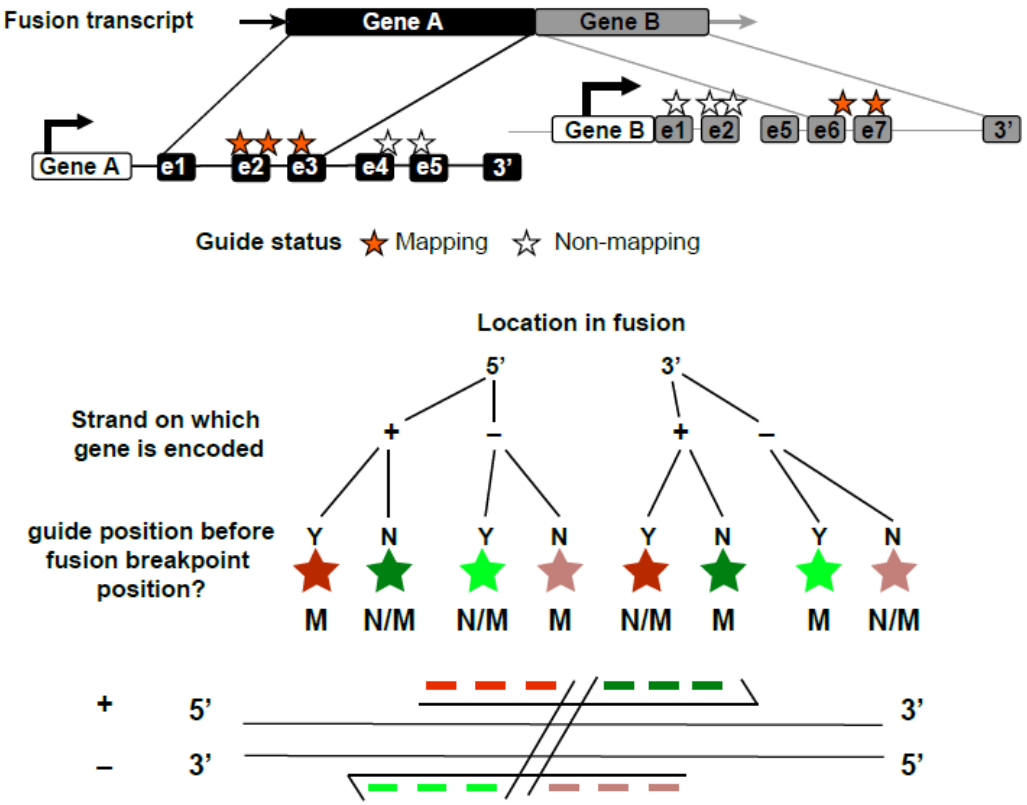


Figure 5.1: Visualisation to illustrate the classification of sgRNA into mapping and non-mapping for each fusion transcript.

Therefore, in the ideal scenario where both mapping and non-mapping sgRNA are present for a gene fusion, the difference between mapping and non-mapping sgRNA will be an indication of the essentiality specific to the gene-fusion.

In this chapter, I describe the development of a computational approach that uses CRISPR/Cas9 whole genome drop-out data to query the essentiality of fusion genes. First I will outline the datasets that I used for the analysis, then explain the pre-processing and the individual steps applied in my computational approach. I then performed a brief quality control of the data that looks at the coverage and quality of sgRNA. Finally, I describe the results of the analysis and give a few examples of my findings.

5.2 Data-sets and data processing

5.2.1 Description

For this analysis, I used three sources of whole-genome CRISPR/Cas9 drop-out screening in human cancer cell lines (Table 5-1).

An in-house effort to generate CRISPR/Cas9 whole-genome drop-out screening data for our panel of cell lines has to date produced 200 Cas9-expressing cell lines and screened them with the Sanger V1 guide library (Tzelepis et al., 2016), with more being released in the coming few months. This data will be publishes under the name Project Score and covers 206 cell lines which have undergone fusion-calling.

Next, Project Achilles performed at the Broad Institute has so far screened another 342 cancer cell lines using the Avana library, of which 243 overlap with my dataset (Meyers et al., 2017). I also include a smaller study by Wang *et al.* on 14 acute myeloid leukaemia cell lines (12 overlapping) that were screened with the Wang *et al.* library.

The three data-sets used library that covered a similar number of genes (~17,000-18,500), although the number of sgRNAs per gene varies, with Wang *et al.* using a median of 10 sgRNAs per gene, compared to 4 and 5 used by Project Achilles and Project Score respectively. Due to this difference, the data-set produced by Wang *et al.* contains mapping sgRNAs for a higher percentage of fusions called within the screened cell lines and may produce higher-resolution data than the other two resources. Conversely, the

Table 5-1: Comparison of three resources for CRISPR-Cas9 drop-out screening in human cancer cell lines

	Project Score	**Project Achilles**	**Sabatini**
Paper	Behan *et al.* (2018)	Meyers *et al.* (2017)	Wang *et al.* (2017)
Guide library	Human CRISPR library v1.0/v1.1	Avana	Wang *et al.* (2015)
# Genes	~18,000	~17,500	~18,500
Median sgRNA / Gene	5	4	10
Cell lines in our panel (total screened)	206	243 (342)	11 (14)
# Fusions assessed	2,148	2,229	69
% Fusions with mapping sgRNA	77%	74%	88%
Reps/Cell line	3	3	1
Data download	(Pre-publication data, personal communication)	[https://depmap.org/ceres/]	[http://sabatinilab.wi.mit.edu/wang/2017/]

Wang *et al.* dataset contains only one biological replicate for each cell line screened, while both Project Achilles and Project Score generated three replicates per cell line.

With some exception in the details, the three resources otherwise followed the same principles in their screening protocols, transducing Cas9-expressing cell lines with a puromycin-selectible lentiviral sgRNA library and leaving the cells in culture for 2-3 weeks.

Finally, cells were harvested, sgRNA PCR amplified and sequenced. The sequences were aligned to the reference genome and raw counts in each cell line for each sgRNA RNA were reported.

5.2.2 Processing flow chart

For my analysis, I downloaded the raw counts reported for each cell line and processed the data for all three resources in a standardised manner (Figure 5.2).

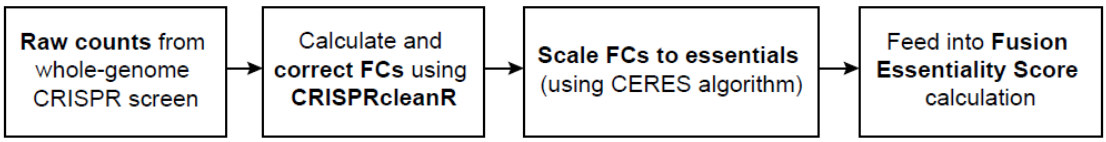


Figure 5.2: Standardised process for data processing

5.2.3 Data download

I downloaded raw counts for each guide in each cell line and guide annotations from the links provided in each of the original publications (Table 5-1). Guide annotations for all three resources were based on GRCh37 and to convert them to GRCh38, to which our fusion data was aligned, I used the NCBI Genome Remapping Service online tool (https://www.ncbi.nlm.nih.gov/genome/tools/remap).

5.2.4 Log-fold change calculation

First, using the CRISPRcleanR R package that was also used to process the original Project Score dataset (Iorio et al., 2018), I calculated log2-fold changes (logFCs) from raw counts for each sgRNA. logFCs simply denote the ratio of sgRNA counts in the final population over that of the initial population, at log base 2. Note that for the Project Score and the Project Achilles resource, the final raw counts were quantified against library plasmid-read counts. For the Wang *et al.* resource, no library plasmid-read counts were available. Instead, I followed the methods described in their paper and averaged the read

counts obtained from 12 of the 14 cell lines at 72 hours after library transduction and before puromycin selection.

5.2.5 CRISPR-bias correction

The depletion of guides in CRISPR-screens may be subjected to biases that are unrelated to the essentiality of a particular gene. Previous studies have shown that the wide-spread DNA damage created by guides that cut at copy number amplified regions tend to lead to high depletion levels even when targeting non-expressed genes (Aguirre et al., 2016; Munoz et al., 2016). To correct for copy-number biases, the Project Score data uses the CRISPRcleanR algorithm, the Project Achilles data used the CERES algorithm, while Wang *et al.* calculate a sliding window score (SWS).

Of the three, the SWS method is the most simplistic. Each gene targeted by a sgRNA is analysed in the context of a "sliding window" of 20 upstream and 20 downstream neighbouring genes (after discarding pan-essential genes). Any neighbouring gene that is in the top 3% of depleted genes in a given cell line adds a value of 1 to the SWS. Genes that have a SWS of > 12 are considered to be in a copy-number amplified region and are removed from further analysis. For my analysis, I excluded the SWS from consideration, as it does not correct logFCs but simply removes those which are in clusters of highly depleted genes.

The CRISPRcleanR method uses a binary segmentation algorithm to identify clusters of neighbouring genes that have fold-changes that are on average significantly different from the background (Iorio et al., 2018). The logFCs in each segment are then corrected by mean-centering, which therefore allows them to be used in further down-stream analyses. An advantage of this method is also that this unsupervised algorithm can identify and correct unknown region-specific biases other than the copy number-bias.

The CERES algorithm for copy-number correction was published together with the release of the Project Achilles data (Meyers et al., 2017). CERES requires input of *a priori* copy-number information for each cell line. It then uses a computational model to estimate the impact of copy-number effects, as well as sgRNA-specific effects shared between all cell lines in a screen. CERES then corrects the logFC for each gene so that the mean copy number effect is shifted to zero.

To decide on which algorithm to use for my analysis, I examined the differences between the supervised CERES and unsupervised CRISPRcleanR methods. The main

difference is that the CERES correction of logFCs depends on *a priori* data on copy number changes in the cell line, while CRISPRcleanR uses a flexible algorithm that corrects logFCs based on detected differences between genomic segments.

A key consideration is that while the number of cuts a guide evokes is strongly correlated with depletion of non-expressed genes, variation still exists with some highly amplified non-expressed genes showing little lethality and single-cut non-targeting guides showing high depletion (Aguirre et al., 2016; Meyers et al., 2017; Munoz et al., 2016). This strongly suggests that other factors other than just copy-number can influence CRISPR-biases.

In fact, a recent study from our group shows that the cell fitness effect associated with copy number amplifications is mainly driven by tandem duplications, rather than chromosomal duplications (Goncalves et al., 2018). The authors show that regions that the CRISPR-bias is stronger in amplified regions in cell lines with low ploidy, and severely reduced in cell lines with high ploidy. This suggests that in fact a large number of DNA-breaks spread out across the nucleus are less lethal than the same number of cuts within a single genomic region.

As copy number amplifications can results from both tandem duplications and chromosomal duplications, correcting them equally, as the CERES algorithm does, may introduce further biases into the data, i.e. over-correcting genes at chromosomal duplications and under-correcting genes in tandem duplications.

Further, as CRISPRcleanR makes no assumptions about the cause of CRISPR-biases, the algorithm has the ability to take into account any potentially unknown region-specific biases. Indeed, the authors of the CRISPRcleanR algorithm identify genomic segments that exhibit biased logFCs without any apparent copy number changes (Iorio et al., 2018).

For the above reasons, I implemented the CRISPRcleanR algorithm for CRISPR-bias correction of the logFCs for all three datasets.

5.2.6 Normalising fold-changes

Slightly different experimental conditions between different cell lines and experimental variation can lead to different ranges of raw counts and the resulting logFCs. This means that if no normalisation is performed, the logFCs for the same guides are not directly comparable between cell lines. As my down-stream analysis requires a direct

comparison of logFCs across different cell lines, I sought to implement a normalisation step.

Normalisation of logFCs was performed by Wang *et al.* using quantile normalisation, while Project Achilles scaled the distribution by *a priori* essential and non-essential genes. The Project Score data was further processed into essentiality scores using BAGEL and MAgECK which did not require normalisation.

Quantile normalisation was first introduced for the analysis of microarrays (Bolstad et al., 2003). Essentially, it works by first calculating an average distribution of logFCs across all samples and then substituting the sorted logFCs with the equivalent position in the average distribution. The result is that the distributions of logFCs of all samples will be identical.

The method developed for CERES to scale logFCs to essential takes as input a set of *a priori* essential and non-essential genes. For each cell line, it then scales a distribution of gene effects so that the median score of essential genes is at -1 and the median score of non-essential genes is at 0.

The advantage of scaling the logFCs to essential is that the distribution of guides are preserved for each cell line. Additionally, scaling the logFCs to the essential genes provides a simple and useful anchor against which the essentiality of unknown guides can be benchmarked.

For these reasons, I decided to scale the CRISPRcleanR-corrected logFCs of all datasets using the CERES scale-to-essential R function provided in the CERES github repository (https://github.com/cancerdatasci/ceres/blob/master/R/scale_to_essentials.R). For the *a priori* essential and non-essential genes, I used the lists supplied within the CERES R package and that were previously defined by Hart *et al.* (Hart et al., 2014). Scaled and CRISPRcleanR-corrected logFCs are abbreviated to "scaled logFCs" in all of the following text.

5.2.7 Mapping guides to fusion transcripts

For the calculation of a fusion essentiality score, I first needed to classify guides into either mapping or non-mapping guides (Figure 5.1). In order to automate this, I devised a set of rules that take into consideration (1) the fusion breakpoint and (2) the mapping location of the guides.

Guides map onto the 5′ gene if the gene is on the positive strand and the guides maps before the position of the fusion breakpoint, or if the gene is on the negative strand and the guide maps after fusion breakpoint (red stars in figure Figure 5.1). Guide map onto the 3′ gene if the opposite is true (green stars in Figure 5.1).

5.2.8 Calculation of fusion fitness score

In order to calculate the fusion essentiality, I made the following considerations:

i. At the most basic level, in true essential fusion transcripts, mapping guides should be highly depleted, while non-mapping guides should have near zero fold-changes.

ii. Where a fusion transcript has only mapping guides, a direct comparison of mapping and non-mapping scores is not possible. However, as only 54.7% of the fusion transcripts tested have differentially-mapping guides, I aimed to develop a score that could also be applied to fusion transcripts with only mapping guides. It is important to note that these types of fusions may suffer from potential false positives, as it cannot distinguish between gene-specific vs. true fusion-specific effects.

iii. Mapping guides for true fitness fusion transcripts should be highly depleted, but this depletion may be caused by other confounding factors and doesn't take into account the average essentiality and range of essentialities of the guide across all cell lines.

Thus in order to calculate the fusion essentiality score, I take the following steps:

1) Z-normalise the scaled logFC across all cell lines screened for a given guide. The Z-score is calculated by taking a scaled logFC, subtracting the mean of all scaled logFCs produced by the guide across all cell lines, and then dividing by their standard deviation. The Z-score therefore represents the magnitude of deviation from the average. Compared to a mean-centred scaled logFC, the Z-score gives added weight to guides that deviate from a narrow distribution than for guides that deviate from a broad distribution.
2) Calculate the difference in Z-score between the mean of mapping and the mean of non-mapping guides for each gene involved in a fusion transcript (Equation 4).

Where a gene has only mapping guides the difference is taken from 0. Where a gene only has mapping guides, a high Z-score indicates cell-line specific essentiality of a gene, which we expect to see in true essential fusions.

3) Finally, to calculate the Fusion Essentiality Score (FES), the difference scores calculated for both fusion genes is averaged. In fusion transcripts where only one gene has mapping guides, the FES = GeneDiff.

$$GeneDiff = mean(sgRNA_{mapping}) - mean(sgRNA_{non-mapping})$$
$$FES = mean(GeneDiff_{5'} + GeneDiff_{3'})$$

Equation 4: Gene difference score and fusion essentiality score calculation.

5.2.9 Calculating a significance score for fusion essentiality scores

In order to estimate the statistical likelihood of a given FES occurring by chance, I performed 10,000 randomisations of the scaled logFCs. Randomisations were performed on a per cell line basis in order to preserve the distribution of guide effects.

After each randomisation, a "randomised FES" is calculated using the same methods as described in the previous section. For each fusion transcript, I calculated a p-value as the number of "randomised FES" that were larger than the non-randomised FES, divided by the number of randomisations. Finally, p-values were adjusted for multiple hypothesis testing using the false discovery rate method.

Figure 5.3A shows the distribution of FES scores across the entire dataset. The scores follow an approximately normal distribution with a mean of mean = 0.21 (sd = 1.01). There is a long tail at the positive end of the distribution and a small secondary peak around +4, which signifies a slight enrichment in positive scores. More positive scores represent fusion transcripts that are more highly depleted.

Fusion essentiality scores have a similar distribution in the Project Score and Achilles data set, while the Wang *et al.* data set appears to have a narrower range and more defined secondary peaks (Figure 5.3B). Means and medians are comparable across all three datasets and whether or not fusion transcripts have differentially mapping or non-mixed guides.

A more thorough analysis of the distribution of the FES using functionally relevant annotation will follow in chapter 6.2.

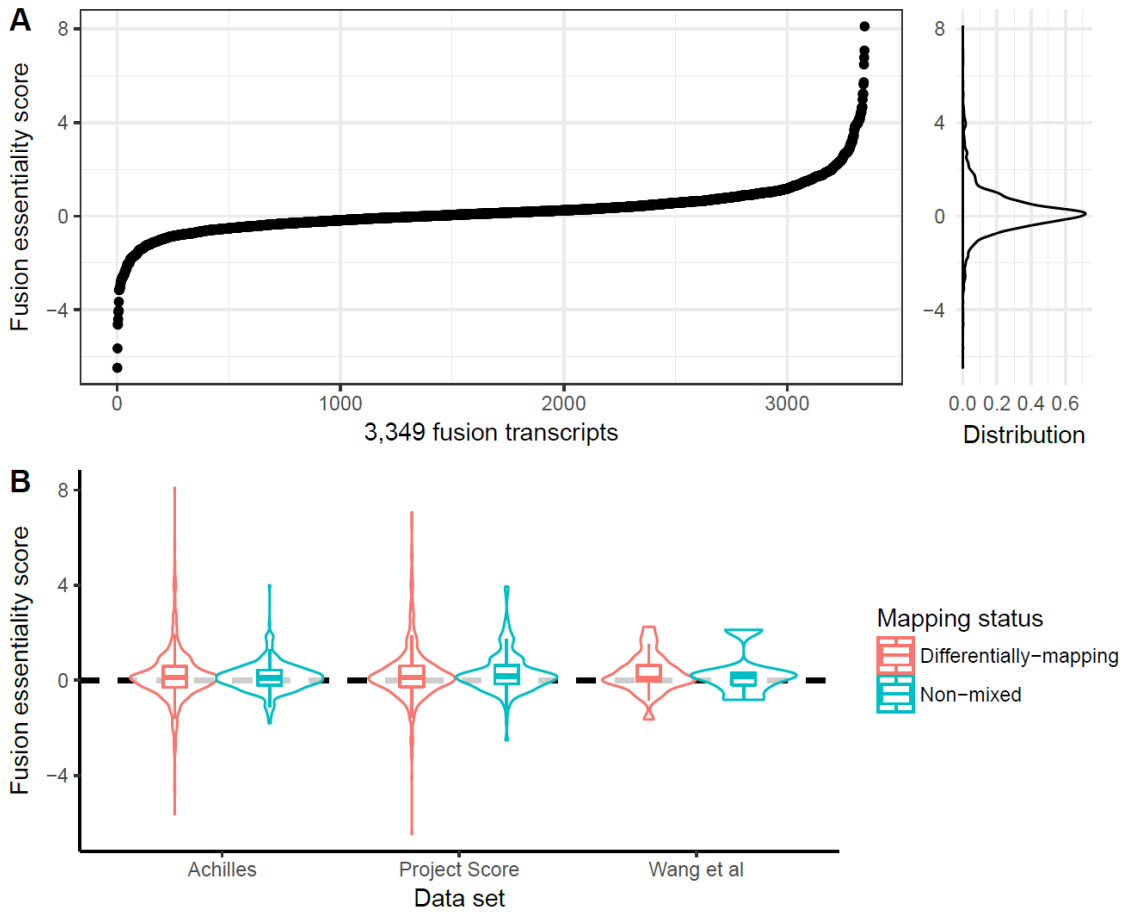


Figure 5.3: Distribution of fusion essentiality scores (A) across all three data sets and (B) by data set and sgRNA mapping status.

5.2.10 Essentiality prediction with BAGEL

Apart from calculating the FES, in some of my downstream analysis I require a binary annotation on whether a given gene is significantly depleted in a cell line. In order to calculate a fitness score for each gene and in each cell line, I implemented the BAGEL framework developed by Trevor Hart and colleagues, which identifies essential genes from pooled library screens (Hart and Moffat, 2016; Hart et al., 2015). BAGEL is a supervised computational algorithm that uses a Bayesian framework to estimate the essentiality of a guide RNA. In short, it uses *a priori* lists of essential and non-essential genes and first estimates the distribution of fold changes of guides targeting these genes within each sample. Then, using a bootstrap process of 1,000 iterations, it calculates the likelihood of the logFC of a given guide to have come from either the distribution of essential or non-essential genes. The likelihood to come from the essential distribution is captured in a Bayes Factor score (BF).

The set of *a priori* essential and non-essential genes can then be used to generate a False discovery rate (FDR), the precision and the recall for each of the computed BFs (Equation 5).

Since the distribution of essential and non-essential genes will differ from sample to sample, the FDR of BFs also accordingly vary by sample. In order to create an easily comparable score, Iorio and colleagues propose a scaled BF (Behan et al., 2018; Iorio, 2018). To calculate this score, for each sample, the BF at which FDR = 5% is subtracted from all BFs calculated for the sample. Thus, across all samples, a scaled BF of > 0 means that a guide is likely to be essential at 5% FDR, a threshold that was originally adopted by Hart *et al.* (Hart and Moffat, 2016).

$$FDR = \frac{FP}{FP + TP}$$
$$Precision = 1 - FDR$$
$$Recall = \frac{TP}{TP + FN}$$

Equation 5: Equations for FDR calculation. FP = False positives; TP = True positives; FN = False negatives.

I implemented the method described above using the R implementation of BAGEL written by Francesco Iorio (Iorio, 2018) to calculate BFs from the CRISPRcleanR-corrected logFCs. To calculate the BF at FDR 5%, I used the *ccr.PrRc_Curve* function within the CRISPRcleanR package.

5.3 Data characteristics & quality control

5.3.1 Data set coverage

In total, the screening data from the three data sets encompasses 371 cell lines in 33 cancer types that were part of our panel (Figure 5.4). Of those, 90 cell lines were screened in two separate data sets. No cell line was screened in all three data sets.

The cell lines harbour 3,598 fusion transcripts (35% of the total). 2,821 transcripts (76%) have at least one guide mapping to the predicted protein. Of those, 528 transcripts (19%) are in cell lines screened by two resources.

Fusion essentiality scores were calculated in 339 of the 371 cell lines, because in 24 cell lines no fusion transcripts were detected and in 8 cell lines the detected fusion transcripts had no mapping guides.

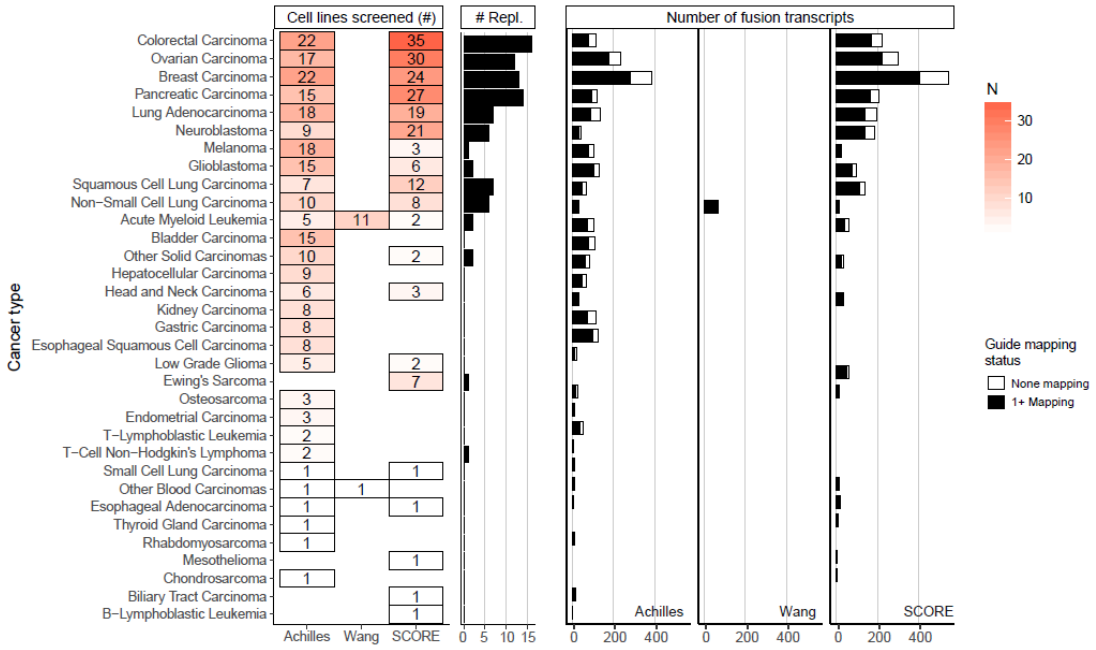


Figure 5.4: Overview of CRISPR/Cas9 screening data available by cancer type. Number of different cell lines screened per resource, number of cell lines replicated between data-sets and number of fusion transcripts in the screened cell lines. White fill in bars indicate that no sgRNA were mapped to a fusion transcript.

5.3.2 Data quality: precision-recall scores across data sets

One measure of data quality is to calculate the precision and recall of expected essential and non-essential genes for each sample in the CRISPR-screen.

This is calculated by ordering the values (here, I used the scaled logFC's) for the sgRNA targeting *a priori* essential/non-essential genes from smallest to largest. At each point, one can calculate the precision, which refers to the proportion of true positives among all positive values and the recall, which refers to the number of true positives from all known essential elements.

In the ideal sample, all the essential genes will have low scaled logFC's (i.e. be highly depleted), and all non-essential genes will have high scaled logFC's. In that scenario, plotting the precision vs. the recall for that sample at each data point will yield a curve with a very high AUC. Conversely, if scaled logFC's were distributed randomly within a sample, the AUC is expected to be at around 0.5, as for each highly depleted essential guide one would also draw out a highly depleted non-essential guide.

Thus, the AUC of precision-recall curves can give an indication of the quality and the noise levels for an individual sample, and is indeed a common quality control step when analysing CRISPR/Cas9 whole genome drop-out screening data. For this analysis, I calculated the precision-recall AUC's for each sample at both the gene and the guide-level across all three data-sets, using the same sets of *a priori* known essential and non-essential genes as described in section 5.2.10.

For both the gene-level and the guide-level precision-recall AUC's, Project Score data outperforms both of the other data-sets (Figure 5.5A-B). For data from Wang *et al.*, the guide-level AUC's are lower than average, while the gene-level AUC's are narrowly higher than that of the Project Score data (mean AUC = 0.902 vs. 0.900). This likely reflects that the screen quality is lower than in the other two screens, possibly due to only having performed one biological replicate for each cell line. That gene-level AUC's are nevertheless high might be due to the use of 10 guides per gene, which is 2x as many as the next data-set, and which would likely yield more accurate gene-level values on average.

Further, I found that the guide-level AUC's for the same cell lines between Project Score and Achilles (n = 83) are significantly correlated ($p < 0.002$, Figure 5.5). This suggests that certain cell lines are inherently more difficult to screen than others, but even on a per-

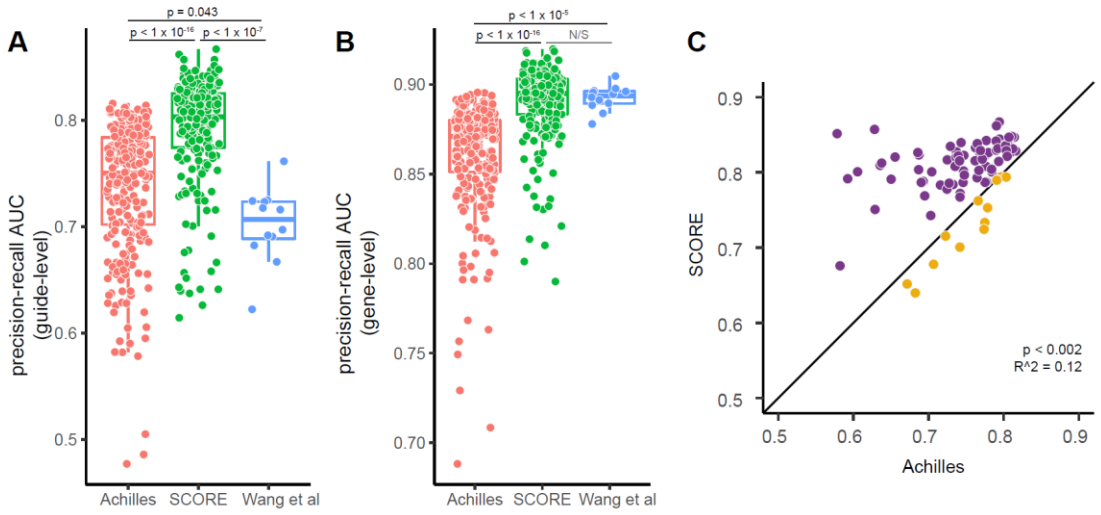


Figure 5.5: Precision-recall AUC's for each sample screened by CRISPR/Cas9, calculated by (A) guide-level and (B) gene-level. (C) Correlation of guide-level precision-recall AUC's for 83 cell lines that were screened by both Project Score and Achilles. Purple dots represent cell lines where Project Score has higher AUC than Achilles and yellow dots represent the reverse.

cell line level, more cell line showed higher AUC's when screened in the Project Score data set over the Achilles data set.

Overall, with some individual exceptions, the majority of precision-recall AUC's are relatively comparable across all samples on both the guide-level (IQR: 0.725-0.805) and the gene-level (IQR: 0.862-0.894). Instances for which the AUC's are low should however be kept in mind and should be considered in down-stream analyses during the interpretation of findings. This analysis shows that all data sets are of comparable quality. Also considering that the sgRNA distribution across fusion genes may vary by data set, I included data from all three datasets in my future analyses. This also means that for fusions in cell lines tested by multiple datasets I also calculated multiple FES.

5.3.3 sgRNA coverage

FES calculation can be performed for any fusion transcript that has at least one sgRNA mapping to a fused portion of either partner gene. However, in general the calculation will yield a more informative result 1) the higher the total number of mapping sgRNAs is and 2) if the number of mapping and non-mapping sgRNA are approximately balanced.

The calculation benefits from a higher number of mapping sgRNA in order to reduce noise from known and unknown CRISPR sgRNA biases. Whole genome CRISPR/Cas9 screens typically rely on 4-10 sgRNA per gene to give an accurate

representation of gene essentiality, but since not all sgRNA in the whole genome sgRNA libraries will be transcript-mapping, some FES may be calculated from a low number of sgRNA. The Wang *et al.* CRISPR library has the highest mean number of targeting sgRNA per fusion transcript, followed by the Project Score and then Achilles libraries (9.95, 5.0 and 3.96 respectively, Figure 5.6A). However, the variation is quite large, with 11.1% of tests (n = 372) being performed based on only a single mapping sgRNA, and a further 9.5% of tests (n = 319) based on two mapping sgRNAs. FES calculations using the lower numbers may be more prone to noise and when evaluating the results of the analysis, it is important to be aware of these potential caveats.

Secondly, a FES can be calculated for a fusion transcript where for both partner genes only one type of sgRNA is available (i.e. mapping OR non-mapping, but not both). In total, 54.7% of the tests conducted involve differentially mapped sgRNA to either the 5′ or the 3′ fused gene, though as might be expected, this proportion is highest in the library by Wang *et al*, which has 10 sgRNA per genes (Figure 5.6B). Of the remaining, there is a slightly higher percentage of transcripts for which the 5′ gene has only mapping sgRNA, compared to the 3′ gene (19.6% vs. 17.2%), although this seems to be driven mainly by the trend in the Achilles data (Figure 5.6B). Also interesting to observe is that around a third of the fused genes are targeted by not a single mapping sgRNA, this number being slightly

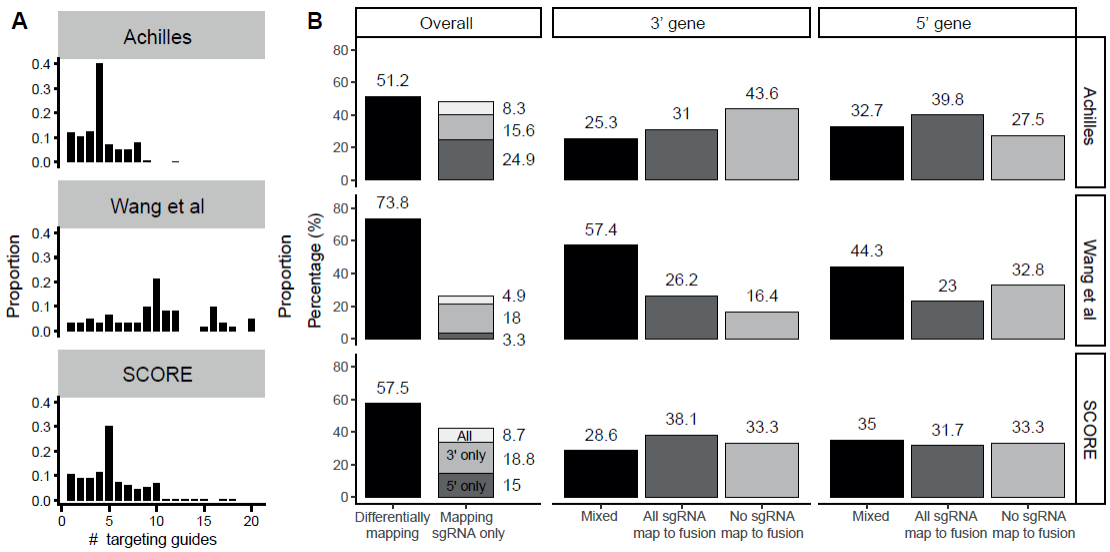


Figure 5.6: Overview of sgRNA coverage across fusion transcripts. (A) Distribution of total number of fusion-targeting sgRNA per test performed for each data set. (B) Percentage of transcripts with differentially mapping sgRNA for either gene (overall), and specifically for the 3′ or the 5′ gene.

lower for the 3′ gene in the Wang *et al.* data set (16.4%) and higher for the 3′ gene in the Achilles data set (43.6%) (Figure 5.6B).

It is important to keep in mind that FES calculations for transcripts with differentially mapping sgRNA for neither gene may be more prone to false positives, as fusion-specific and gene-specific effects may be difficult to untangle. In those cases, further evidence should be considered, for example whether any confounding events are present (e.g. amplification events that may lead to local wide-spread structural rearrangements). These cases would also benefit from validation experiments to confirm the fusion-specific effects, e.g. using CRISPR/Cas9 system with tailored sgRNA.

5.3.4 sgRNA efficacy

A further metric to estimate the effectiveness of a CRISPR/Cas9 screen is to examine the ability of individual sgRNA to have a deleterious effect in any cell line. One approach is to examine the number of cell lines within which a given sgRNA has a low scaled logFC, i.e. a high deleterious effect. I used a threshold of -1 to define sgRNAs with high effect in a given cell line, since the sgRNA depletion scores were scaled so that on average, essential genes would have a scaled logFC of -1. Then counting the number of sgRNA which are depleted at a scaled logFC of at least -1, I find that overall, 45.5% of sgRNA are depleted in at least one cell line. The proportions are much higher in the Achilles and Project Score data sets (66.1% and 60.6%; Figure 5.7), presumably due to the larger number of samples tested compared to Wang *et al*.

When targeting genes of unknown essentiality, a lack of deleterious effect in any cell line can be due to several reasons: 1) the gene is not an essential gene, 2) gene is essential in some instances, however those are not captured in the samples tested and 3) the gene is essential, but the sgRNA was not effective.

To focus on cases where the gene should be essential, but the sgRNA was not effective, I specifically examined sgRNA that are *a priori* known essential and non-essential (the same set as used to scale logFCs in section 5.2.6). Among 6,465 sgRNA that target known essential genes, 83.4% where depleted in at least one cell line, with some variation between different data sets (Figure 5.7B). The number of depleted cell lines per sgRNA varied greatly, e.g. sgRNA targeting *C11orf24* were depleted to that threshold in only 5 instances, while at the other end *RPS3A* and *RPS8* targeting sgRNA were on average depleted in about 90% of samples. However out of the 360 *a priori* known essential genes,

all exhibited a high depletion value with at least one sgRNA, which suggests that the 16.6% of sgRNA that are not highly depleted in any cell line may in fact have low efficacy in creating gene knock-out.

On the other hand, the depletion pattern of non-essential guides broadly resembles that of all other guides (Figure 5.7C). Of those targeting non-essential genes, 37.1% of sgRNA still show high depletion in at least one cell line. These observations may be due to noisiness of the screen or off-target effects. Similarly, in the original publication of this set of reference non-essential genes, the authors selected protein-coding genes with expression levels below 0.1 FPKM in 15 of 16 BodyMap tissues and 16 of 17 ENCODE cell lines (Hart et al., 2014). As our data set of over 300 CRISPR-screened cell lines is much larger than that, it is indeed a possibility that a portion of the *a priori* non-essential may in fact be essential in a low number of cell lines.

In conclusion, the analysis above gives an indication that a proportion of sgRNA are indeed expected to have low efficacy in the screen. This result may impact FES calculations performed using very few targeting sgRNA. At the same time, when the FES calculation yields non-significant results, it may be useful to check the general efficacy of the involved sgRNA across other samples to avoid misinterpreting a false negative.

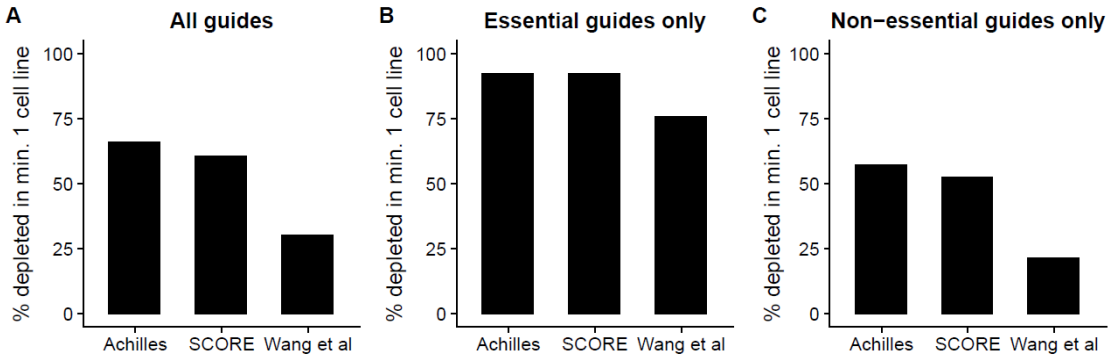


Figure 5.7: Number of guides that show high depletion (scaled logFC < -1) in at least one cell line, for (A) all sgRNA, (B) essential sgRNA only and (c) non-essential sgRNA only.

5.4 Chapter conclusion

This chapter outlines my method as well as quality control elements in my approach of using CRISPR/Cas9 screening data to generate scores of fusion-specific essentiality in cancer cell lines. I showed how I calculated a fusion essentiality score (FES) for 339 cancer cell lines for which I had data available, covering 27% (n = 2,821) of the fusion transcripts identified overall. Further, my data QC showed that precision-recall AUCs are relatively comparable across all samples at the guide-level as well as the gene-level. Importantly, 11% of tests are based on only a single fusion-mapping sgRNA and only 54.7% of transcripts have differentially mapping sgRNA to a single gene. Either of these may contribute to noisiness in the data, which should be kept in mind for future data interpretation.

Overall, the above chapter provides the basis for the next chapter, which will give a general overview of the results as well as some specific examples that illustrate the power of the approach. A comprehensive discussion of the important findings and learnings in creating and using this computational approach will follow at the end of the next chapter.

6 Identifying functional gene fusions using CRISPR/Cas9 whole genome screening data

This chapter is based on the concept and methodology described in chapter 0. Here, I first provide a general overview of the results and a brief introduction into how to interpret the data and graphs in this chapter. I then set out to understand the distribution of significantly essential fusion essentiality scores across the data set, show the validation of known oncogenic fusions and finally end by highlighting a few specific interesting novel findings.

6.1 Fusion Essentiality Scores

Using the methods described in chapter 5.2, I calculated 3,349 fusion essentiality scores (FES) for 2,821 unique fusion transcripts in 339 cell lines (Supplementary Table 8).

Across all fusion transcripts in all three data sets, only 4.6% of transcripts had a FES with a FDR < 5%, 6.8% at FDR < 10% and 13.7% at FDR < 25%. Within the data-sets, some differences could be observed (Figure 6.1).

The data set by Wang *et al* shows a particularly high proportion of fusion transcripts with highly significant FES. This is likely due to the fact that the 11 acute myeloid leukaemia cell lines contain a high proportion of known oncogenic driver fusions. In fact, 8% of fusion transcripts tested in the Wang *et al* dataset are listed in the COSMIC fusion census, compared to 2% of the fusion transcripts in the other two datasets.

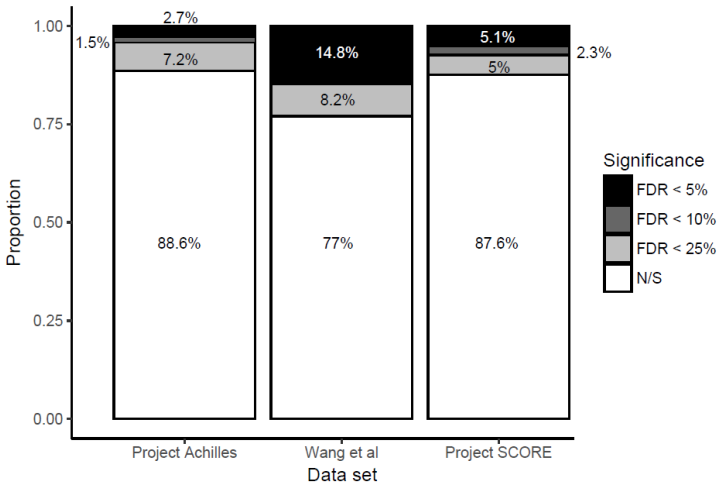


Figure 6.1: Proportion of significant results in each of the three CRISPR/Cas9 data sets.

6.1.1 Two specific examples of the fusion essentiality analysis

To help interpret the FES analysis, I next describe two examples, of which one is a known essential fusion and another a known confounder.

KMT2A-MLLT3

As described in the previous chapter, *KMT2A-MLLT3* fusions are well studied for inducing tumour initiation in acute leukaemias. Previous studied have shown that withdrawal of *KMT2A-MLLT3* in an inducible *in vivo* system leads to disease regression (Zuber et al., 2011). The MOLM-13 cell line is an acute myeloid leukaemia cell line that carries the fusion protein. The fusion transcript brings together exon 8 of *KMT2A* and exon 6 of *MLLT3*.

MOLM-13 was previously screened in a whole-genome CRISPR drop-out screen by Wang *et al.* (Wang et al., 2017). Their whole-genome CRISPR guide library (Wang et al., 2014) contains 10 guides that map to *KMT2A* and another 10 guides that map to *MLLT3*. Of those, 8 guides map to the fused region of *KMT2A* and 3 to that of *MLLT3*, while the rest map to the unfused portions of the genes (Figure 6.2A). By mapping scaled fold-changes on a guide-by-guide-basis, as expected from an essential fusion, I find that mapping guides have a lower fold-change than non-mapping guides (Figure 6.2B), indicating they are preferentially deleted. The FES for this fusion transcript is 1.85 and highly significant at $p < 0.001$ and FDR < 0.001.

ERBB2-CDK12

The gastric carcinoma cell line NCI-N87 harbours an *ERBB2-CDK12* fusion transcript that is likely a passenger fusion due to a copy number gain of *ERBB2*. NCI-N87 has been screened as part of the Project Achilles (Meyers et al., 2017) using a library that contains three mapping and one non-mapping sgRNA on *ERBB2* and four non-mapping guides on cyclin dependent kinase 12 (*CDK12*) (Figure 6.2C).

After mapping scaled fold-changes on a guide-level basis, as expected, all sgRNA mapping onto ERBB2 are depleted, regardless of whether they target the fused portion or not (Figure 6.2D). Thus, as the depletion of sgRNA is not specific to fusion-mapping regions, the FES is low at 0.05 and not significant ($p > 0.5$ & FDR > 85%).

Here it is important to highlight a potential caveat of the method: since the CRISPR sgRNA library was not specifically designed for this usage, for many fusion transcripts, the sgRNA mapping locations may be unbalanced. For instance, in NCI-N87, aside from *ERBB2-*

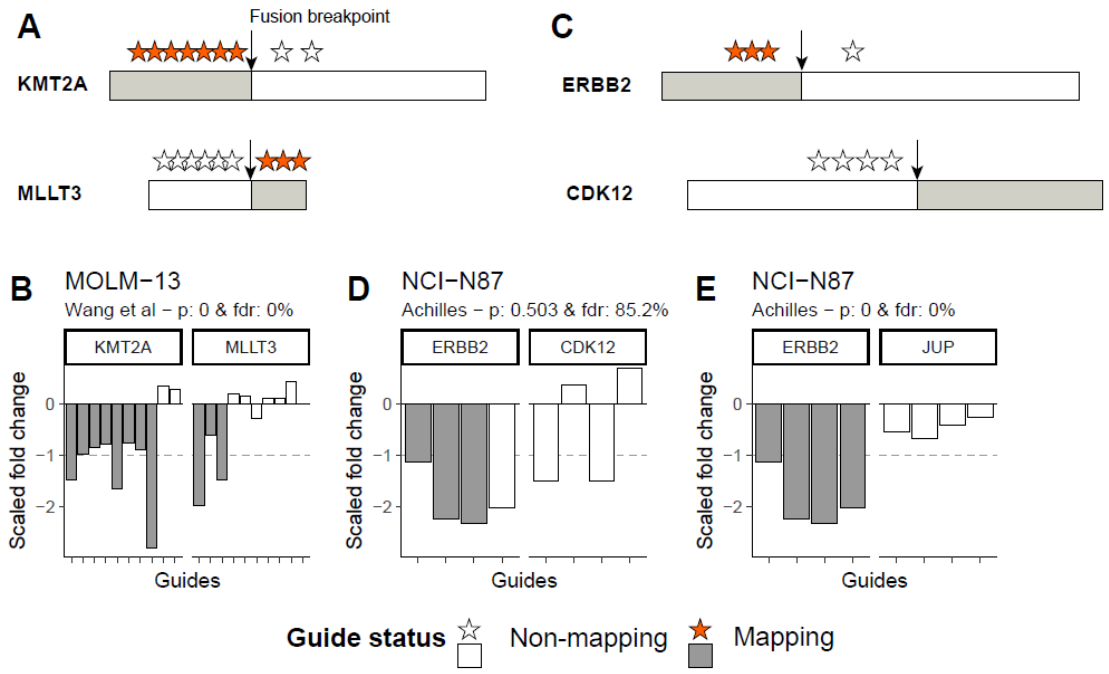


Figure 6.2: Examples for fusion essentiality score calculation for (A,C) a known driver fusion transcript, KMT2A-MLLT3 and (B,D,E) for a known confounded fusion transcript. Data source for each set of sgRNA is indicated above the plot (Wang et al, 2017; Project Achilles = Meyers et al, 2018; Score = Behan et al, 2018.)

CDK12, we also detect another *ERBB2*-fused transcript, *ERBB2-JUP*. In this transcript however, *ERBB2* is fused in a way so that all sgRNA target locations are preserved. When calculating the FES, thus all fusion-mapping guides will be depleted due to the essentiality of *ERBB2*, and since for transcripts with no differential mapping guides the difference is taken from zero, the FES is high at 4.0 and significant at $p < 0.001$ and FDR < 0.001 (Figure 6.2E).

As we are aware of the oncogenic driver effects and essentiality of amplified *ERBB2*, we can declare this transcript a false positive with high confidence. However in cases where the cell-line specific essentialities of genes are not understood, it is important to be aware that this type of sgRNA mapping pattern may lead to misleadingly high significance of passenger events.

6.2 Gene set enrichment analysis of fusion essentiality score significance

In order to find out whether certain categories of gene fusions are enriched for significant FES, I next annotated all tested fusion transcripts according to 10 distinct categories, as described below, and conducted a gene-set enrichment analysis using the piano R package with its default settings (Väremo et al., 2013). I used the p-values of FES

as gene level statistics. Only 2 out of 10 categories were significantly enriched after p-value adjustment: 1) fusion transcripts that were listed in the COSMIC fusion census and 2) fusion transcripts that had significant associations with any drug in our fusion-drug association analysis (Table 6-1, Figure 6.3).

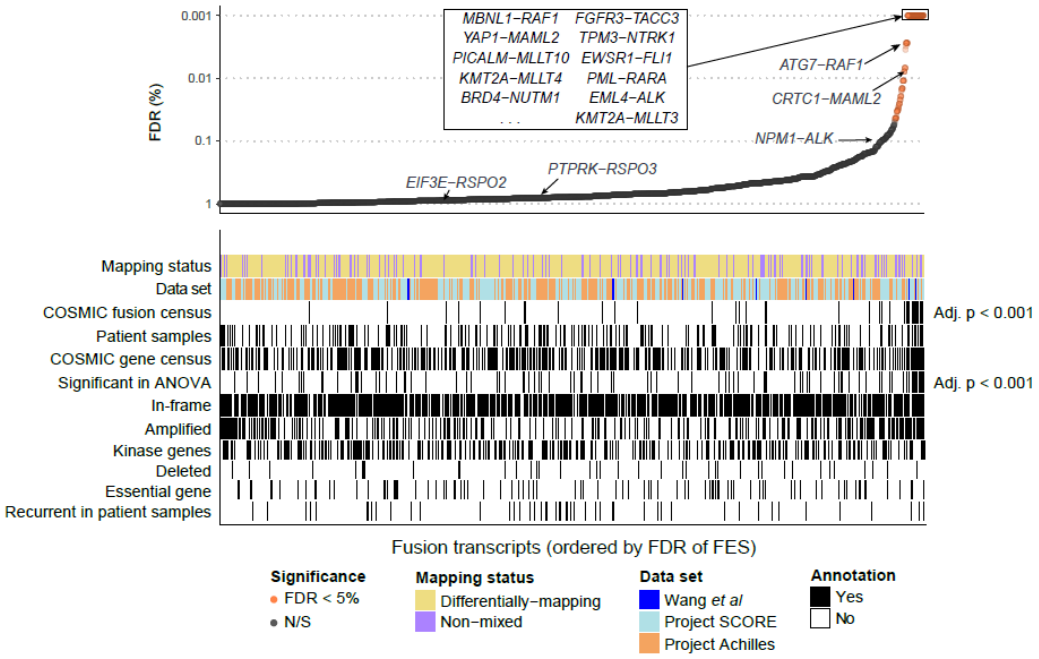


Figure 6.3: Overview of 3,349 fusion essentiality scores. Annotation shows fusion essentiality scores where the fusion transcripts belong to a certain category (e.g. fusions includes genes in COSMIC gene census, kinase genes, essential genes, etc.).

Table 6-1: Gene-set enrichment test scores, ordered by significance of adjusted p-value.

Name	Genes	Stat	p	p adj.
COSMIC fusion census	48	0.17736	1.00E-04	0.00035
Significant in ANOVA	106	0.29554	1.00E-04	0.00035
In-frame	750	0.4057	0.039796	0.11143
Patient samples	187	0.40543	0.21268	0.33083
Deleted	50	0.38332	0.18128	0.33083
COSMIC gene census	443	0.41159	0.20428	0.33083
Recurrent in patient samples	63	0.40188	0.29697	0.41576
Amplified	299	0.41573	0.34647	0.44096
Essential gene	105	0.42031	0.46345	0.5407
Kinase genes	363	0.44485	0.91481	0.91481

In the next paragraphs, I describe the definition and context of annotation sets, and the results from this analysis.

Amplified/deleted

For this annotation, I used the gene-level PICNIC (predict integral copy number in cancer) scores, published by COSMIC for each cell line (Greenman et al., 2010), for both of the partner genes involved in a fusion event. PICNIC scores represent the total number of copies found for a gene in the genome of a given cell line. A fusion event is annotated as deleted or amplified, if the PICNIC score of either one of the two partner genes is either -1 (deleted) or higher than 8 (amplified). The threshold of 8 to define amplified genes was used to follow the classification used for the COSMIC cell line data (Forbes et al., 2017).

The annotation mainly served as a negative control. As amplified genes are known to cause CRISPR-screen biases (as discussed in chapter 5.2.5), although I already performed a correction of CNV regions using CRISPRcleanR (see section 5.2.5), I still wanted to test whether fusion events containing amplified or deleted genes may showed biased FES FDR's. P-values were not significant for either (adj. p = 0.44 and 0.33 respectively), which showed that the FES calculation was not biased by copy number status.

Essential genes

Here, I annotated fusion transcripts where either partner gene is part of the list of high-confidence pan-essential genes published by Hart and colleagues (Hart et al., 2014). An enrichment of essential genes at higher significance level might have indicated that genes that are universally highly essential could cause biased FES for reasons unrelated to the fusion transcript. After performing the gene set enrichment analysis, I observed no such enrichment (adj. p = 0.54).

Kinase genes

Kinase genes, which include *ALK* and *ABL1*, are commonly fused oncogenes, and previous publications specifically analyse the landscape of kinase fusions in cancer genomes (Gao et al., 2018; Stransky et al., 2014). A list of kinase genes was published by Gao and colleagues as part of their work to landscape fusion events in patient tumours (Gao et al., 2018). I annotated any fusion event that includes a kinase gene in either the 5′ or 3′ partner gene, but there was no significant enrichment (adj. p = 0.92). This suggests that although known oncogenic fusions commonly involve kinase genes, simply detecting a fusion involving a kinase gene gives no indication of functional relevance. In those cases,

the identity of the partner genes, the predicted breakpoint and the tissue context should be considered and functionality needs to be confirmed in further analyses.

In-frame

The frame of fusion transcripts is predicted using the GRASS score (see section 3.4). Any fusion transcript that is called with a GRASS flag of 880 and above is defined as in-frame (Cancer IT, 2018b) and thus annotated. The gene set enrichment analysis shows that in-frame fusions are not significantly enriched across the FES FDR (adj. p = 0.11). This is perhaps not surprising, since if we assume that non-oncogenic breakpoints are randomly distributed across the genome, a third of breakpoints within coding regions would be in-frame simply by chance. Since only about 4.5% of fusion transcripts have significant essentiality using our analyses, the majority of in-frame fusion transcripts are likely to be passenger events.

COSMIC fusion census

Here, I annotated any fusion transcript between two specific partner genes that are listed in the COSMIC fusion census. The census is a manually curated database for gene fusions for which there is published evidence of their involvement in cancer processes (https://cancer.sanger.ac.uk/cosmic/fusion). There is a clear and significant enrichment of fusion transcripts that are listed within the COSMIC fusion census ($p < 0.0004$). This enrichment suggests that my computational approach is able to detect functional cancer driver fusion with good sensitivity. Fusion transcripts that match the COSMIC fusion census are discussed in further detail in section 6.3 below.

COSMIC gene census

This category includes fusion events that involve a gene that is listed in the COSMIC gene census, a data-base of manually curated cancer driver genes (http://cancer.sanger.ac.uk/census). Known cancer driver genes that are fused may have cancer-related functionality. For example, fusions of oncogenes can lead to constitutive activation of a kinase, e.g. in the case of *BCR-ABL1* and *NPM1-ALK*, or of a transcription factor binding site, e.g. in *CRTC1-MAML2*. Nonetheless, I observe no enrichment of fusions involving cancer driver genes in the FES FDR's ($p = 0.33$). This suggests that gene fusions in cancer driver genes might be common and researchers that observe such fusions in individual samples should be careful not to over-interpret those findings without additional evidence of functionality.

(Recurrent in) patient samples

Here, I annotated any fusion that was reported by Gao *et al.* (Gao et al., 2018) in their high-throughput RNA-seq analysis of almost 10,000 TCGA patient samples. A recurrent patient sample is any fusion that is reported in at least two or more different samples. That neither category was significantly enriched for low FES FDR's ($p = 0.33$ and 0.42 respectively) underscores the fact that mere observation of a fusion, even if recurrent, is not a good predictor of potential functionality of a fusion.

Significant in ANOVA

This category contains any fusion that is significantly (FDR < 25% and Glass Deltas > 1) associated with any drug as part of the fusion-drug-association analysis (see chapter 3). That this category is significantly enriched is perhaps not surprising, as it contains several known oncogenic driver fusions.

In total, 57 fusions fall under this category, making up 130 fusion transcripts. 10 of those fusions have at least one transcript that is significant at FDR < 5%. Five of the ten are well-known oncogenic drivers (*EML4-ALK*, *EWSR1-FLI1*, *KMT2A-MLLT3*, *CRTC1-MAML2, FGFR3-TACC3*). Two of the ten (*FGFR2-ATE1, ERBB2-JUP*) are known confounded events, as described in chapter 3.

A further 3 fusions belong to neither category (*APIP-SLC1A2, SLC12A7-TERT, LRP5-RP11-554A11.9*). However, it is important to note that mapping sgRNA were unequally distributed across the fusions. In *APIP-SLC1A2 and SLC12A7-TERT*, mapping sgRNA were concentrated in a single gene that also had no non-mapping sgRNA (Figure 6.4A). Unlike fusions with differentially mapping sgRNA within the same gene, this type of mapping status makes false positives more likely when the gene with mapping sgRNA is essential in the cell line regardless of the fusion status, as in Figure 6.4B. Similarly, the *LRP5-RP11-554A11.9* fusion only has sgRNA mapping onto the *LRP5* gene, as the 3' portion aligns to an uncharacterised mRNA *RP11-554A11.9*. The significance of non-protein forming mRNA is currently unknown.

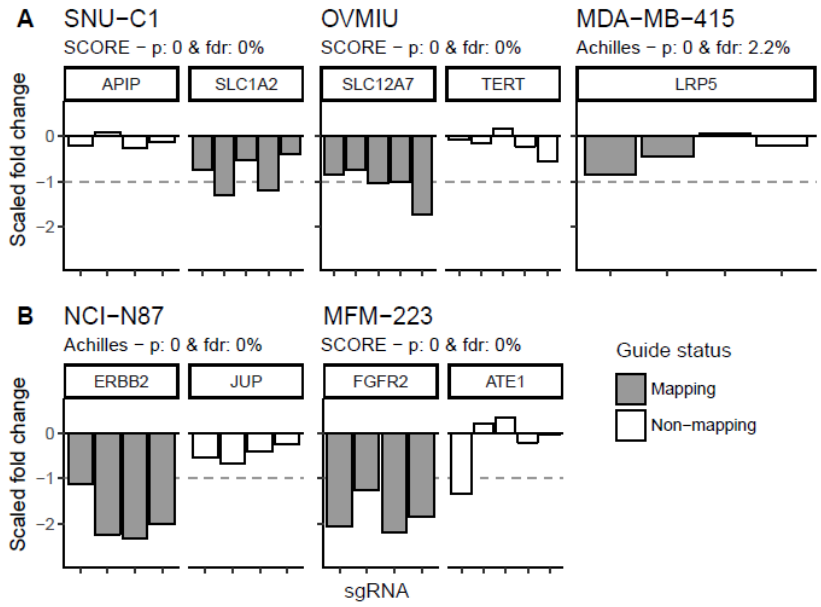


Figure 6.4: Visualisation of selected fusions with high significance in both the drug-association analysis and the fusion essentiality analysis. (A) uncharacterised fusions and (B) known confounded fusions.

While these fusions have been found in multiple cell lines, CRISPR-screening data was only available in a single cell line. Future CRISPR-screening data in the untested cell lines may reveal either opposing or supporting evidence for the functionality of these fusions.

6.3 Detection of known oncogenic gene fusions

Across all tests performed, 61 transcripts involved genes in the COSMIC fusion census. This involves 30 unique fusion events with 20 unique fusions. The remaining 31 are fusion transcripts with alternative breakpoints (n = 18) and the same fusion transcripts tested with data from another data set (n = 13).

Of the 30 unique fusion events, 16 (53%) have a significant FDR of < 5% in at least one transcript in one data set. These include extensively studied oncogenic fusions, including *EWSR1-FLI1*, *KMT2A-MLLT3*, *EML4-ALK* and others (Figure 6.5A). Two more fusion events, *KMT2A-MLLT3* in NOMO-1 and *NPM1-ALK* in KARPAS-299, narrowly miss the FDR threshold at FDR = 11.8% and 13.5% respectively (Figure 6.5B).

A further twelve (40%) fusion events that match the COSMIC fusion census have an insignificant FDR at > 25%. However when I manually curated these fusions, they were linked to single publications (Figure 6.6). Six of the seven (*PLXND1-TMCC1*, *EIF3K-CYP39A1*, *PLA2R1-RBMS1*, *NFIA-EHF*, *MBOAT2-PRKCE* and *SLC26A6-PRKAR2A*) were detected in the

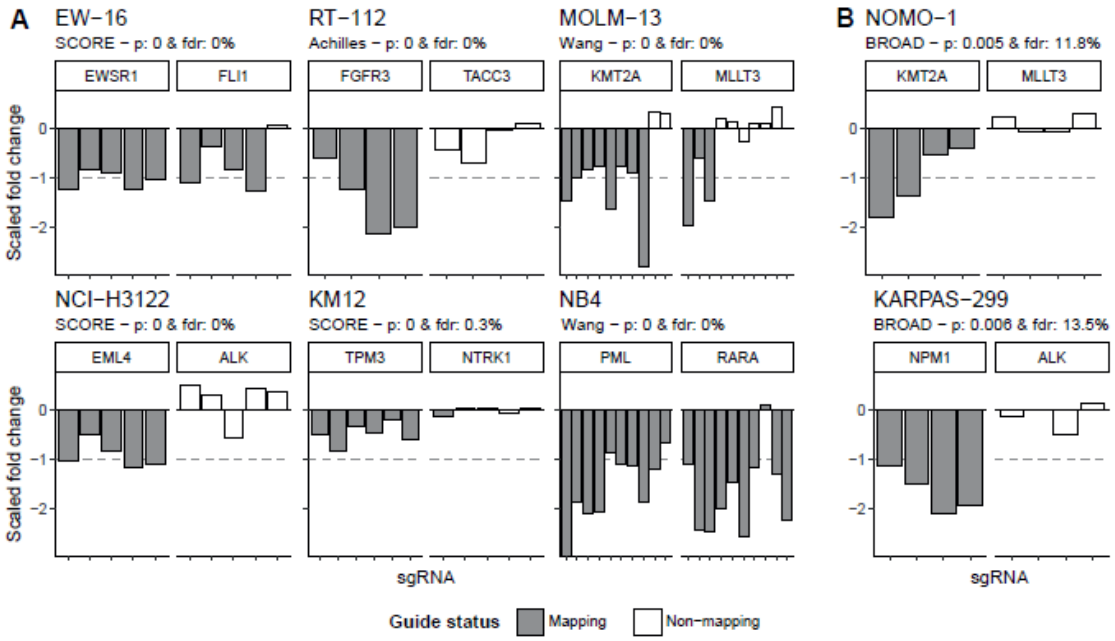


Figure 6.5: Visualisation of scaled log fold-changes across well-studied oncogenic gene fusions for both (A) highly significant examples at 5% FDR and (B) less significant examples at FDR < 25%.

same paper which described somatic rearrangements in 24 human breast cancer genomes (Stephens et al., 2009). Notably, none of the six fusion transcripts were described or even named in the main text, so the inclusion into the COSMIC fusion census is likely to have been due simply to detection. Further, three of the above fusions were tested in two different data-sets and the lack of significance is in agreement in all tests conducted (Figure 6.6B-D)

A seventh fusion, an *ARID1A-MAST2* fusion in the breast cancer cell line MDA-MB-468, was described in a publication by Robinson and colleagues (Robinson et al., 2011) in the exact same cell line. In their publication, siRNA knock-down of microtubule associated serine/threonine kinase 2 (*MAST2*) mRNA led to reduced cell growth of MDA-MB-468, and introduction of *MAST2*-fused gene vectors into a control cell line produced accelerated cell growth. The authors did not perform experiments that tested whether this effect was caused by the specific fusion transcript, or whether these cell lines are simply more reliant on MAST2. By chance, MDA-MB-468 was screened across two different data-sets, Project Score and Achilles and the FES for the *ARID1A-MAST2* fusion is significant in neither (Figure 6.6A). Notably, several MAST2-targeting sgRNA in both the Achilles (n = 2), as well as the Project Score (n = 3) data show high depletion levels (i.e. scaled logFCs < -1) in at least 2 other samples, which suggests that the sgRNA are functional and the lack of depletion is

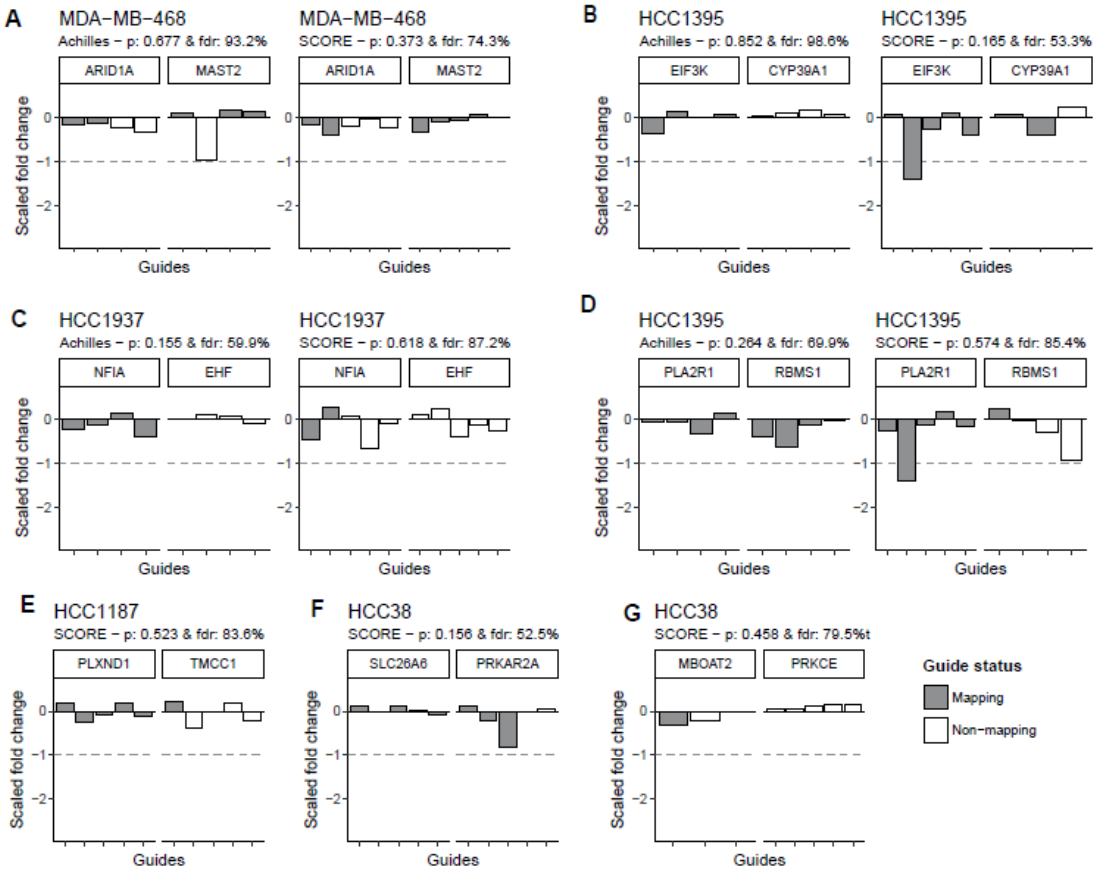


Figure 6.6: Visualisation of scaled log fold changes for seven fusion transcripts that match the COSMIC fusion census. A-D were tested by multiple data-sets (Project Achilles and Project Score).

not a false negative. Considering the negative finding and that aside from the original publication in 2011, no other publication has since been published on *ARID1A-MAST2*, I conclude that the evidence for functionality of the *ARID1A-MAST2* gene fusion in breast cancers is unclear and the fusion may constitute a passenger event rather than an oncogenic driver.

To sum up the above two paragraphs, I argue that for the seven fusions described above, despite their inclusion in the COSMIC fusion census, evidence for their functionality is weak and that they should thus be removed when considering the sensitivity of my computational approach to detect true positive fusions.

Thus considering only the remaining 23 fusion transcripts, my approach detects 16 (70%) with a stringent 5% FDR threshold and 18 (78%) using a less stringent 25% FDR threshold.

Table 6-2: All 30 fusion events that match the COSMIC fusion census, ordered by FDR. Thick black borders denote stringent threshold (FDR < 5%), permissive threshold (FDR < 25%), not significant fusion events and not significant fusion events excluded for reasons described in section 6.2. In patient samples refers to fusions found in TCGA samples by Gao and colleagues (Gao et al*, 2018). Excl refers to fusions which are excluded due to weak evidence of functionality.*

Cell line	Fusion name	FES	p-value	FDR	In patient tumours?	Excl?
ES4	*EWSR1-FLI1*	1.82	0	0	No	
ES6	*EWSR1-FLI1*	2.43	0	0	No	
ES8	*EWSR1-FLI1*	1.97	0	0	No	
EW-16	*EWSR1-FLI1*	2.18	0	0	No	
EW-7	*EWSR1-FLI1*	1.96	0	0	No	
KM12	*TPM3-NTRK1*	1.92	0	0	Yes	
MC-IXC	*EWSR1-FLI1*	2.32	0	0	No	
MHH-ES-1	*EWSR1-FLI1*	2.69	0	0	No	
MOLM-13	*KMT2A-MLLT3*	2.31	0	0	Yes	
NB4	*PML-RARA*	2.12	0	0	Yes	
NCI-H3122	*EML4-ALK*	4.47	0	0	Yes	
OCI-AML2	*KMT2A-MLLT4*	3.05	0	0	No	
RT-112	*FGFR3-TACC3*	4.18	0	0	Yes	
SBC-3	*BRD4-NUTM1*	3.84	0	0	No	
THP-1	*KMT2A-MLLT3*	1.41	0	0.31	Yes	
H3118	*CRTC1-MAML2*	4.4	0	0.78	No	
NOMO-1	*KMT2A-MLLT3*	1.89	0.005	11.77	Yes	
KARPAS-299	*NPM1-ALK*	0.76	0.006	13.48	No	
RT4	*FGFR3-TACC3*	2.62	0.102	46.82	Yes	
HCC-78	*SLC34A2-ROS1*	2.35	0.146	51.12	Yes	
EGI-1	*PTPRK-RSPO3*	0.64	0.363	73.38	No	
SUP-M2	*NPM1-ALK*	0.08	0.436	81.64	No	
ESO51	*EIF3E-RSPO2*	-0.9	0.644	88.68	No	
HCC38	*SLC26A6-PRKAR2A*	0.29	0.156	52.52	No	Y
HCC1395	*EIF3K-CYP39A1*	0.77	0.165	53.26	No	Y
HCC1937	*NFIA-EHF*	0.29	0.155	59.92	No	Y
HCC1395	*PLA2R1-RBMS1*	0.24	0.264	69.85	No	Y
MDA-MB-468	*ARID1A-MAST2*	0.16	0.373	74.32	No	Y
HCC38	*MBOAT2-PRKCE*	0.05	0.458	79.48	No	Y
HCC1187	*PLXND1-TMCC1*	-0.07	0.523	83.58	No	Y

6.4 Reproducibility of significant FES across data sets

In order to understand the reproducibility of significant FES in different testing conditions and with different sgRNA libraries, I next examined the significantly essential fusion events tested in different data sets. In total, 27 significantly essential (FDR < 5%) fusion events were tested in two different data sets. None was tested by all three data sources. Of those, only 10 (37%) were significant at a less stringent FDR < 25% threshold in both data sets (Figure 6.7A). Among those that replicate across multiple data sources, I observe a large number of known cancer drivers, e.g. *KMT2A-MLLT3*, *PML-RARA*, *TPM3-NTRK1*.

In previous chapters I showed that non-differentially mapping sgRNA may be more prone to false positive errors. Thus, I was interested to examine whether there would be differences in reproducibility from FES calculated from differentially mapping and non-differentially mapping sets of sgRNA. Indeed, I found that significantly essential fusion events called from non-differentially mapping sgRNA are less likely to reproduce in a different data set (one-tailed t-test: $p = 0.033$, $t = -2.1$; Figure 6.7B). This corresponds to 71% of differentially mapping transcripts validating, vs. 30% for the remaining transcripts.

An example of that is the *FGFR2-MRPL13* fusion in NCI-H716, a colorectal cancer cell line with a known *FGFR2* copy number gain (Forbes et al., 2017). In the Project Score data, all 8 sgRNA targeting *FGFR2* and mitochondrial ribosomal protein L13 (*MRPL13*) are fusion-mapping and on average highly depleted ($p < 0.001$ & FDR < 0.001, Figure 6.7C). In the Achilles data, only a single sgRNA targeting *MRPL13* maps to the fusion, while the non-fusion-targeting sgRNA for *FGFR2* are still highly depleted, thus giving an insignificant result ($p = 0.089$; FDR > 44%).

On the other hand, a non-reproducing result may also be caused by screening noise. For instance, a fusion between phosphatidylinositol binding clathrin assembly protein (*PICALM*) and *MLLT10* is detected in about 7.5% of patients with T-cell acute lymphomas and leukaemias (Nigro et al., 2013). A previous study showed that RNAi-based knock-down of *PICALM-MLLT10* leads to reduced viability *in vitro* and decreased invasive ability when transplanted *in vivo* (Okada et al., 2006). A *PICALM-MLLT10* fusion in P31-FUJ, an acute myeloid leukaemia cell line, is significantly essential according to the Achilles data, but not the Wang *et al* data. Although in the latter, the majority of targeting guides show

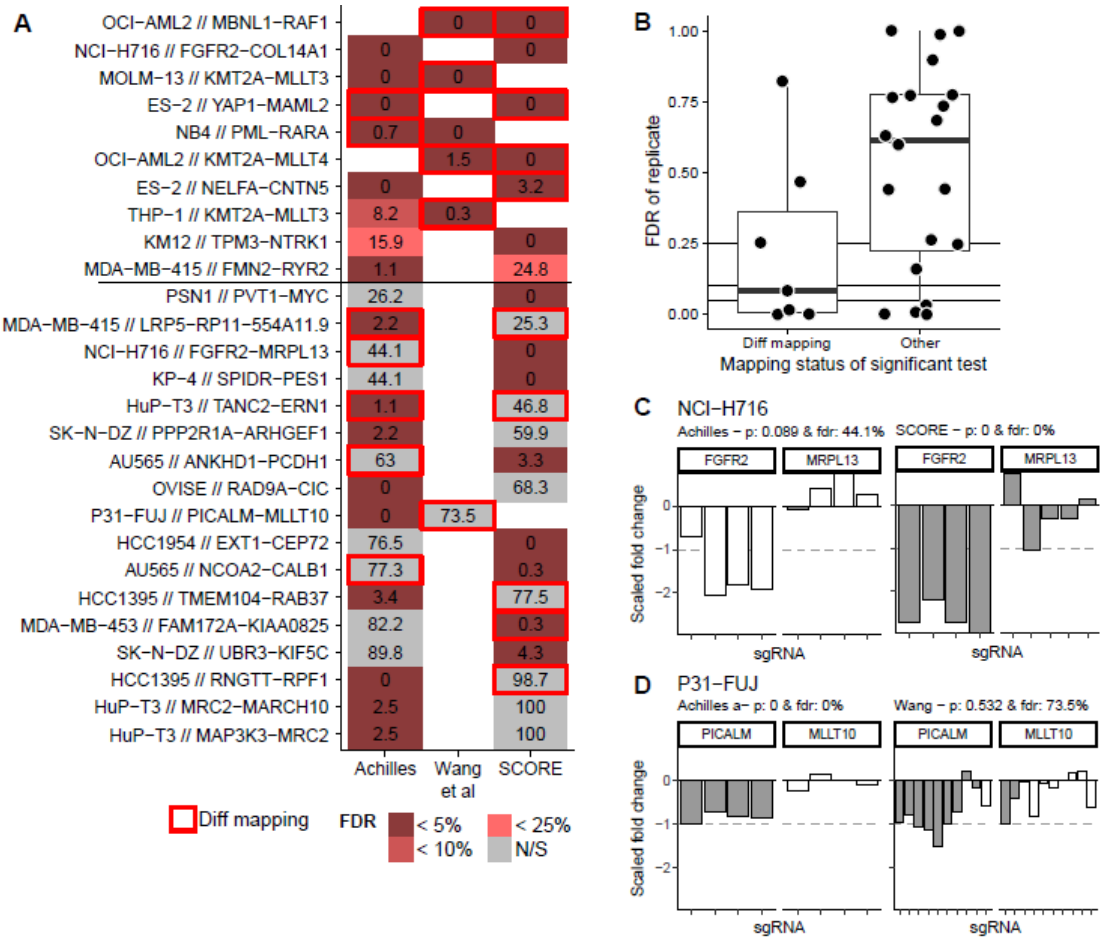


Figure 6.7: Examination of reproducibility of significantly essential fusion transcripts across multiple data sets. (A) 27 fusion events that are significantly essential in at least one dataset. Numbers represent the FDR (%). The black line separates fusion events for which the two data sets are in concordance (top) and those in opposition (bottom). (B) Where two FES are available for the same fusion transcript, the FES with the higher FDR was assigned as "replicate" (y-axis), while the FES was assigned as either differentially mapping, or other, based on its mapping status. Significantly essential fusion events are more likely to be reproducible if they were based on differentially-mapping sgRNA (one-tailed t-test; p = 0.033; t = -2.1). (C) Non-differentially mapping sgRNA lead to the false positive calling of the FGFR2-MRPL13 fusion in the Project Score data. (D) Screening noise leads to the likely false negative result in the Wang et al. data for PICALM-MLLT10.

relatively consistent high depletion, some noisiness in the data on MLLT10 non-fusion-mapping guides yields a lower scaled logFC (Figure 6.7D).

A third source of noisiness could arise from the Z-normalisation that I apply on scaled logFCs that seeks to consider how much the effect of a sgRNA differs in a given cell line compared to all other samples. The Z-normalisation identifies and gives weight to true outliers from a population. Most scaled logFCs, depending on the genes in question, are distributed approximately normal, with significantly depleted sgRNA often producing a secondary peak. However, in a noisy screen, highly exceptional depletion effects may arise through one-off technical artefacts, especially when a sgRNA overall shows a narrow

distribution of low depletion effects. In such cases, a Z-normalisation may distort data and yield false positives. In the short-term, this issue can be circumvented by manually inspecting the scaled logFCs to confirm that an assumed high depletion is also found in the non-normalised data. In the long-term, it would be advisable to conduct an analysis to compare the impact of the Z-normalisation on data distortion, and whether there may be better methods to solve the question of outlier-identification.

In conclusion, while previous sections showed that my computational approach is able to detect known oncogenic fusions with high sensitivity, this section shows that noise is nonetheless expected to exist and may add false positives as well as false negatives. Non-differentially mapping sgRNA are especially prone to calling false positives. However, as due to the pre-determined sgRNA library design and as they are still effective in calling known true positives (e.g. *TPM3-NTRK1*, *FGFR3-TACC3* etc; see section 6.3), I decided to retain these tests. However, when evaluating those results, it is important to be aware of any confounding factors, and consider stringent validation experiments to confirm that findings are indeed true positives.

6.5 Overview of less studied significant fusion transcripts

After establishing some aspects of the sensitivity and specificity of my computational approach in identifying true known oncogenes, I return to the aim of discovering novel functionality of fusion events.

In total, 138 of the 3,349 tests returned a significant FES at FDR < 5%. Next, because I am specifically interested in fusions that are highly essential, but a high FES can be caused by non-mapping guides showing positive selection, I also filter out fusion transcripts for which the mean scaled logFCs of mapping guides is larger than -0.45. This leaves 93 high effect, significantly essential tests. The threshold of -0.45 was chosen to retain known oncogenic fusions.

Out of those, 14 tests confirmed the same fusion transcripts across two data-sets. A further 49 of them tested the same fusion events across an alternative breakpoint. Thus, the number of fusion events that have at least one significant transcript from any data set at FDR < 5% is 65.

The 63 unique fusion events constitute 56 unique fusions. In total, and nine of the 56 (16%) are COSMIC fusions (Figure 6.8). 37 (66%) fusion events involve a gene in the COSMIC gene census. 12 fusion events (21%) were previously observed in the ~10,000 TCGA patient samples (Gao et al., 2018). However, here it is worth noting that some known oncogenic fusions are also not observed in that set of patient samples, e.g. *EWSR1-FLI1*, *NPM1-ALK*. This indicates that although that set of patient samples brings valuable insight into the landscape of fusions, it is by all means not comprehensive in representing all true oncogenic fusions. 4 fusion events (7%) involve a gene that is amplified and a known oncogenic driver and are thus likely to be false positives (*ERBB2-JUP*, *ERBB2-LTBP4*, *PVT-MYC and FGFR2-COL14A1*).

Notably, only 3 fusions are detected as significantly essential in more than a single cell line (Figure 6.8). These are *EWSR1-FLI1* (n = 7), *YAP1-MAML2* (n = 3) and *KMT2A-MLLT3* (n = 2). *EWSR1-FLI1* and *KMT2A-MLLT3* are both well-studied oncogenic fusions, and their essentiality is present in cell lines from cancer types consistent with the clinical profile, i.e. *EWSR1-FLI1* fusions appear only in Ewing's sarcoma and *KMT2A-MLLT3* only appear in Acute myeloid leukaemia cell lines. The third recurrent fusion, the previously

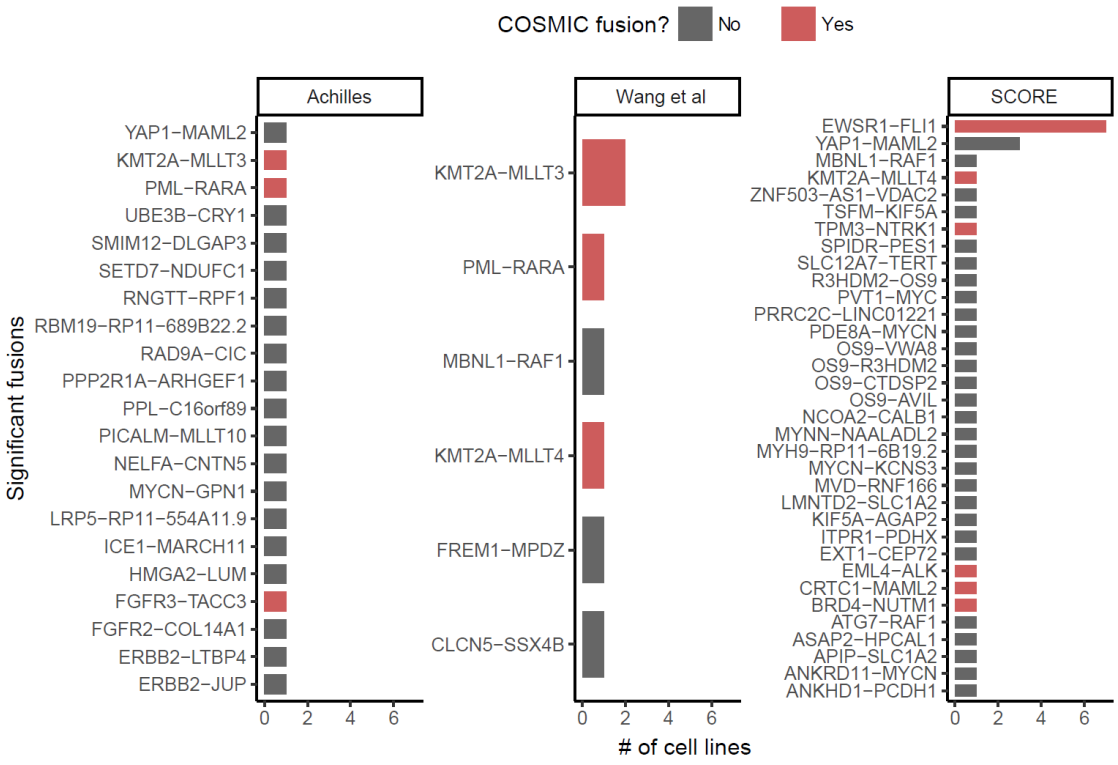


Figure 6.8: Overview of high effect significant (FDR < 5%; mean scaled logFC < -0.45) fusion transcripts from CRISPR/Cas9 analysis for fusion essentiality.

uncharacterised *YAP1-MAML2* is an interesting case, which, I examine in further detail in the next paragraph.

6.5.1 YAP1-MAML2

Unlike any other fusion of unknown background, I found significant essentiality in all three cell lines in which *YAP1-MAML2* was detected, AM-38 (glioblastoma), ES-2 (ovarian carcinoma) and SAS (head and neck carcinoma). Further, one of the samples, ES-2, was tested and *YAP1-MAML2* was found as significantly essential in both the Achilles and the Project Score data, thus adding robustness and confidence in this being a true positive observation. Further, depletion of fusion-mapping sgRNA targeting *MAML2* is specific to the three cell lines with the *YAP1-MAML2* fusion (Figure 6.9B). One notable exception is the head and neck cell line H3118, which harbours the known oncogenic *CRTC1-MAML2* fusion.

While multiple transcripts are found in AM-38 (n = 2) and SAS (n = 3), one of the breakpoints is conserved across all three cell lines. This breakpoint of position 102,206,074 and 96,093,517, both on chromosome 1, produces an in-frame transcript that brings together exons 1-5 of *YAP1* and exons 2-5 of *MAML2* (Figure 6.9C). In order to confirm the presence of the fusion transcript, we performed PCR validation using cDNA purified from the cell lines, interphase FISH and fibre-FISH (Figure 6.9D-E)[4].

Interestingly, although *YAP1-MAML2* has never been functionally characterised to date, it was previously reported in a nasopharyngeal carcinoma, a sub-type of head and neck carcinomas, where the authors showed that the fusion is likely to result from a coupled inversion (Valouev et al., 2014). A *YAP1-MAML2* fusion was also detected in a skin cancer sample analysed as part of the ~10,000 TCGA patient samples (Gao et al., 2018). Strikingly, in both of those patient samples, the breakpoint of the fusion is conversed and identical to that found in our cell lines.

[4] The follow-up work on the *YAP1-MAML2* fusion was performed in close collaboration with Gabriele Picco, a post-doctoral researcher at the Wellcome Sanger Institute. Gene set enrichment analysis, PCR validation and FISH were planned and performed by him. Methods are described in Picco* and Chen*, *et al* (manuscript under review).

The evidence above strongly suggest that it is indeed the *YAP1-MAML2* fusion and not the individual genes themselves that are essential for cell line survival. YAP1 overexpression has previously been linked to oncogenic effects in multiple solid tumours. For instance, an expression signature related to YAP1 overactivation is associated with poor prognosis in colorectal cancer (Lee et al., 2015b). In pancreatic carcinomas YAP1 activation

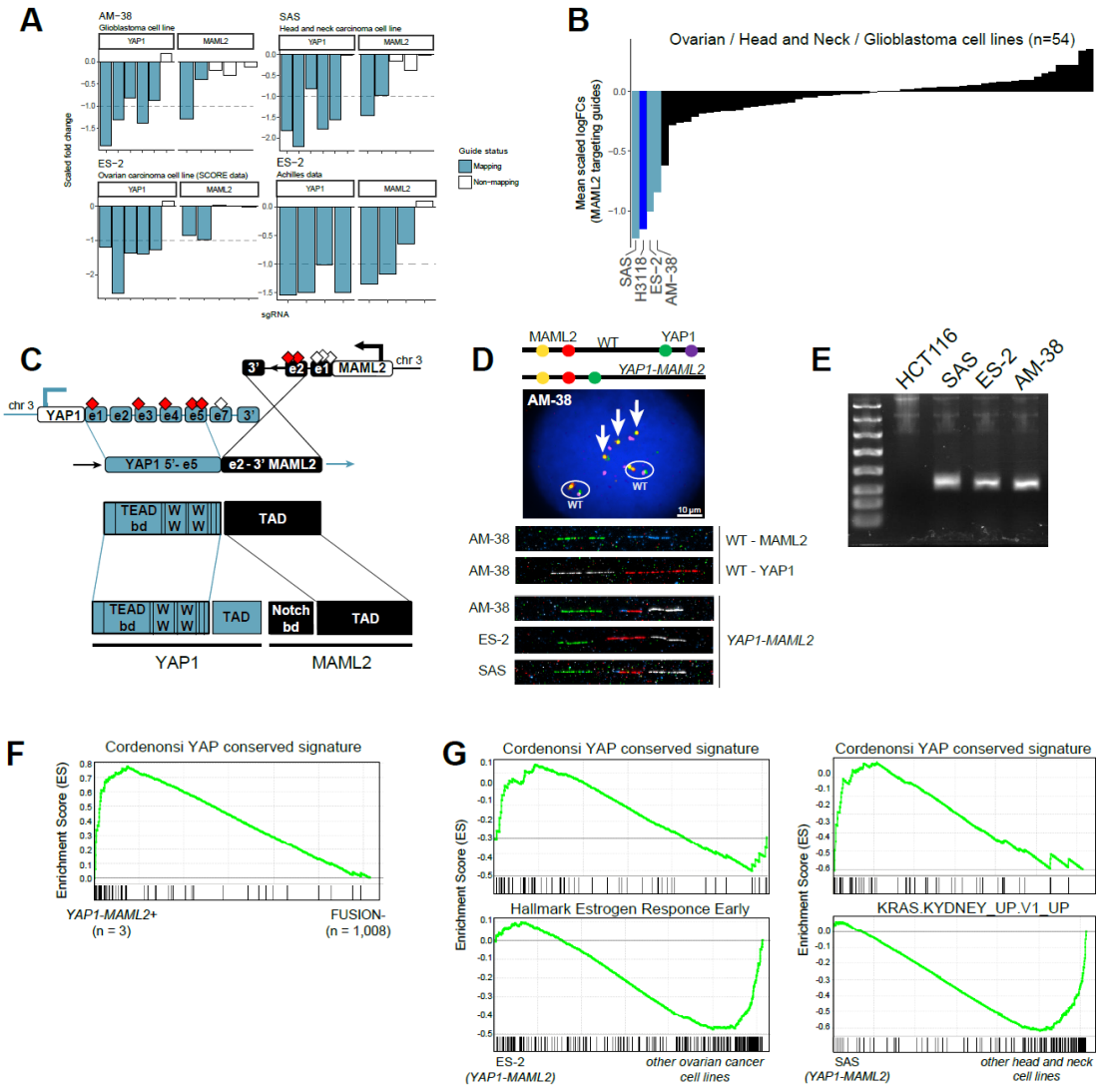


Figure 6.9: Novel functionality for YAP1-MAML2 fusion found in three cell lines. (A) CRISPR profiles in all three cell lines. ES-2 was tested in both Project Score and Achilles guide libraries. All p < 0.001 & FDR < 0.001. (B) Fusion-targeting sgRNA for MAML2 are uniquely depleted in MAML2-fused cell lines. (C) Exons 1-5 of YAP1, which contain the TEAD binding domain, are fused to exons 2-5 of MAML2, which contains a transcription activation domain (TAD). (D) Interphase FISH in AM-38 and Fibre-FISH in AM-38, ES-2 and SAS confirm the presence of the YAP1-MAML2 fusion in all three cell lines. (E) PCR validation of the fusion in all three cell lines. HCT116 is a negative control without the fusion. (F) Gene set enrichment analysis shows YAP-conserved signature as the top most enriched gene set in cell lines with YAP1-MAML2 fusion (adj. p < 0.001). (G) YAP-conserved signature is enriched when comparing ES-2 against all other ovarian cancer cell lines and SAS against all other head and neck cell lines. Conversely, the tissue prototypic signatures (i.e. oestrogen response and KRAS signature respectively) are negatively enriched in the fused cell lines.

bypassed KRAS addiction (Kapoor et al., 2014). In oesophageal cancer, YAP1 upregulation led to EGFR expression and resistance to chemotherapy (Song et al., 2015).

The oncogenic function of YAP1 is typically attributed to its role as a transcriptional co-activator that activates the Hippo pathway. Previous work shows that Hippo pathway activation is driven in particular with binding to the TEA domain transcription factor 1 and 2 (TEAD1/2) (Harvey et al., 2013). On the other hand, MAML2 is a transcriptional co-activator normally involved in NOTCH signalling (Kitagawa, 2016). MAML2 is also known in the context of the well characterised CRTC1-MAML2 fusion. There, the CREB binding domain encoded in the exon 1 of *CRTC1* is fused to the transcriptional activation domain on exons 2-5 of *MAML2*, leading to constitutive activation of the CREB signalling pathway. As a direct result of that, a ligand to EGF receptors, AREG, is upregulated, which thus induces oncogenic activity through the EGFR signalling pathway (Chen et al., 2014).

As the protein resulting from the *YAP1-MAML2* fusion contains both the transcriptional activation domain of *MAML2* and the TEAD-binding domain of *YAP1*, I hypothesised that the fusion may lead to constitutive activation of the YAP1/TEAD signalling pathway. In accordance with that, we performed a gene-set enrichment analysis comparing the three YAP1-MAML2 fusion positive cell lines against all others. Of 189 pathways tested, the YAP1-conserved signature was the top most significantly enriched (adj. $p < 0.001$) (Figure 6.9F). YAP1-conserved signature was also enriched when ES-2 was compared against all other ovarian cancer cell lines and SAS against all other head and neck cell lines, while expression of prototypic tissue-specific oncogenic signatures, such as oestrogen receptor signalling in ovary, were depleted (Figure 6.9G).

In summary, my data suggests strongly that recurrent *YAP1-MAML2* fusions are associated with increased YAP1 signalling and essential for cell fitness. In the future, it would be interesting to follow-up the question of whether *YAP1-MAML2* fused tumours respond to inhibitors of YAP1 and other members of the Hippo-signalling cascade.

6.5.2 RAF1 fusions

Another interesting vignette revolves around *RAF1* fusions, of which two instances are significantly essential in the CRISPR/Cas9 analysis. Rare RAF1 fusions have been reported in several patient tumours across multiple tissue types, including a *SRGAP3-RAF1* fusion in pilocytic astrocytoma, a *HACL-RAF1* fusion in pancreatic acinar cell carcinoma and

an *ESRP1-RAF1* fusion in prostate adenocarcinoma (Chmielecki et al., 2014; Jones et al., 2009; Palanisamy et al., 2010) (Figure 6.10A). The fusions all preserve the kinase domain in *RAF1* while removing an N-terminal auto-inhibitory domain, which likely leads to MAPK pathway activation. In particular, fibroblast cells transduced with the *SRGAP3-RAF1* showed increased phosphorylation of MEK1/2 and exhibited anchorage-independent growth (Jones et al., 2009). Similarly, *ESRP1-RAF1* transduced cells showed increased cell proliferation and were sensitive to RAF and MEK inhibitors (Palanisamy et al., 2010).

In our cell line panel, we detect a *MBNL1-RAF1* fusion in the acute myeloid leukaemia cell line OCI-AML2, which is consistent with two previous high-throughput analyses (Klijn et al., 2015; Wang et al., 2017). We also identify a previously unreported in-frame *ATG7-RAF1* fusion in a pancreatic adenocarcinoma cell line, PL-5. Both fusions are significantly essential at $p < 0.001$ and FDR < 0.5% (Figure 6.10B).

An *ATG7-RAF1* fusion was previously also found in another KRAS wild type pancreatic cancer cell line, PL5, which is not in our cell line panel (Giacomini et al., 2013). There, the authors already showed that cell growth of PL5 was inhibited upon siRNA knockdown of *ATG7*.

To characterise *ATG7-RAF1* in PL-18, we confirmed the presence of the fusion of autophagy related 7 (*ATG7*) to *RAF1* by Sanger sequencing across the breakpoint and interphase FISH (Figure 6.10C) [5]. Like the RAF1-fusions found in patients and in PL5, the breakpoint of *ATG7-RAF1* leads to deletion of the N-terminal regulatory domain while preserving the kinase domain (Figure 6.10D). Unlike most pancreatic cell lines, PL-18 also retained wild-type KRAS (28 of 32 cell lines are mutated in KRAS). In our high-throughput drug screening data, PL-18 nonetheless showed high sensitivity to MEK-inhibitors trametinib and PD0325901 (Figure 6.10E), which suggests that MAPK pathway signalling is still essential. The result from the high-throughput drug screening was validated by exposing both PL-18, OCI-AML2 (*MBNL1-RAF1* fused) and a negative control cell line, SU8686 to trametinib over 72 hours and observing viability compared to the same cell lines exposed to DMSO. Both PL-18 and OCI-AML2 showed similarly strong loss of viability

[5] Validation experiments for *RAF1*-fused cell lines were performed in collaboration with Gabriele Picco. He designed and performed/delegated validation drug screens, FISH and Sanger sequencing. Methods are described in Picco & Chen, *et al.* (manuscript under review).

(Figure 6.10F). Furthermore, PL-18 was the only pancreatic cell line in our panel for which sgRNA targeting the fused portion of *ATG7* were significantly depleted (Figure 6.10G).

The above experiments provide strong evidence that the *ATG7-RAF1* fusion is essential for the viability of the PL-18, likely through the upregulation of the MAPK pathway. By mining sequencing data for 126 pancreatic adenocarcinoma PDX models, we identified one *KRAS* wild-type tumour with a *PDZRN3-RAF1* fusion (PA2409; supplied by CrownBio), which suggests that *RAF1* fusions are likely to also be found in clinical pancreatic adenocarcinoma.

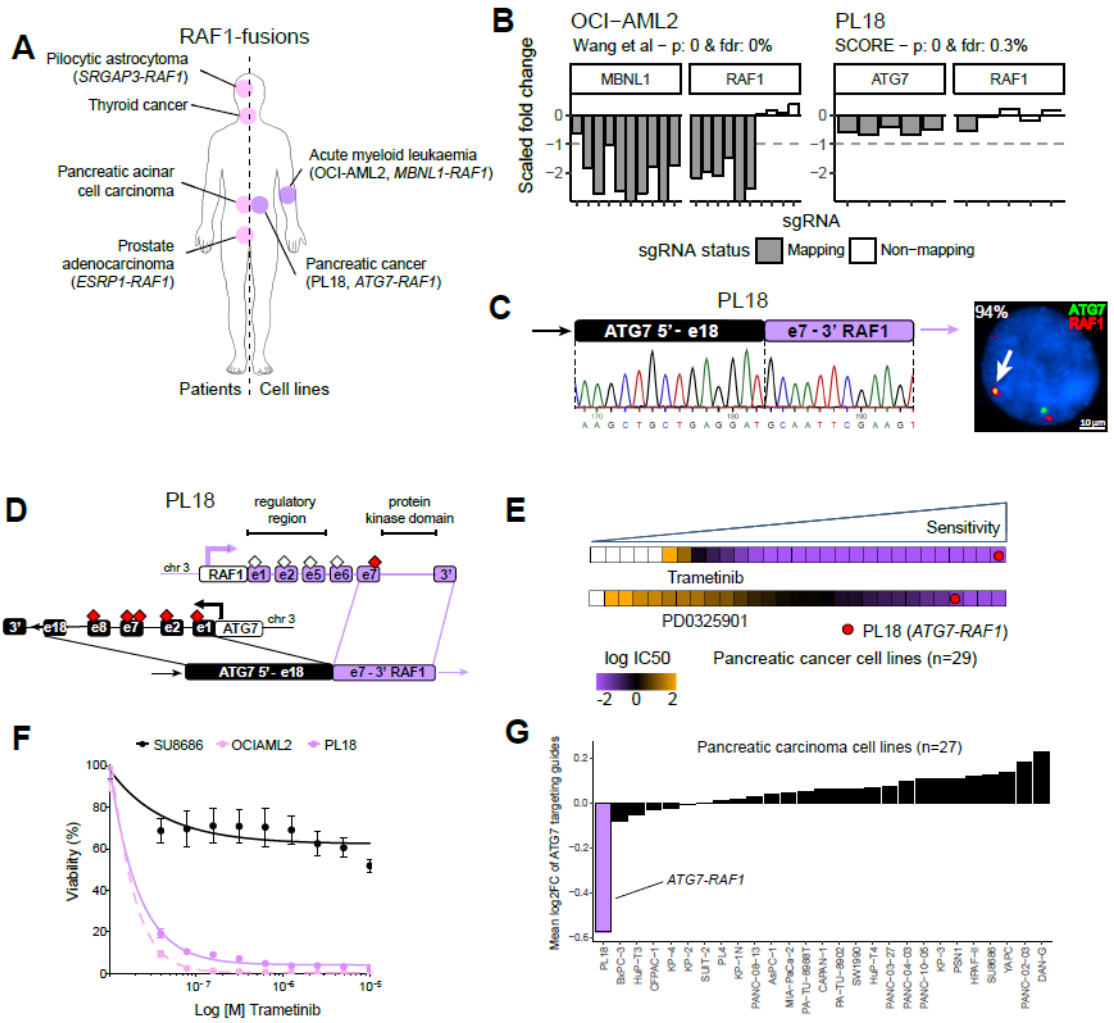


Figure 6.10: Essential ATG7-RAF1 fusion identified in pancreatic cancer cell line PL18. (A) Tissue types where RAF1 fusions have previously been identified. (B) RAF1 fusions are significantly depleted in the CRISPR/Cas9 analysis. (C) Capillary sequencing and FISH confirm the presence of a fusion breakpoint. (D) Exon 18 of ATG7 is fused to exon 7 if RAF1. (E) Across 29 pancreatic cancer cell lines, PL18 is among the most sensitive to MEK inhibitors trametinib and PD0325901. (F) Drug response curves of OCI-AML2, PL18 and a negative control pancreatic carcinoma cell line SU8686. (G) Among pancreatic carcinoma cell lines, PL18 is the only one for which sgRNA targeting sgRNA are depleted.

Thus, our data provides evidence for the second instance of a functional *ATG7-RAF1* fusion in pancreatic cancer cell lines. This indicates that *RAF1* fusions in pancreatic cancer cell lines may have a higher incidence than previously believed. Further, we show for the first time that cell lines carrying the fusion are sensitive to MEK inhibition, highlighting that this rare subtype of *KRAS* wild-type pancreatic carcinomas is potentially clinically actionable.

6.5.3 BRD4-NUTM1

The *BRD4-NUTM1* fusion typically characterises a rare and aggressive disease termed NUT midline carcinoma, a subtype of head and neck carcinomas (French, 2010). It involves the fusion of the N'-terminal bromodomains of bromodomain containing 4 (*BRD4*) at exon 9 with virtually the entire NUT midline carcinoma family member 1 (*NUTM1*) gene at exon 2. Although usually only expressed in testis, the fusion puts *NUTM1* under the control of the *BRD4* promoter, leading to ubiquitous expression (French et al., 2003). Through siRNA knock-down, the fusion has been shown to prevent differentiation and proliferation, although the exact mechanism of action is currently poorly understood (French et al., 2008). Importantly, in an initial study four fusion-positive patients responded well to the BET inhibitor OTX015/MK-8628, with resulting disease stabilisation and tumour regression (Stathis et al., 2016).

Curiously, in our cell line panel, we find two cell lines with *BRD4-NUTM1* fusions. One, RPMI-2650 is a known cell line of NUT midline carcinoma that is sensitive to BET inhibitors (Stirnweiss et al., 2017). Another, SBC-3 is a small cell lung cancer cell line, which carries a *BRD4-NUTM1* fusion with the canonical break point (Figure 6.11A). Intriguingly, *BRD4-NUTM1* has a significant fusion essentiality score in SBC-3 (Figure 6.11B), which suggest that it is necessary for tumour maintenance. Follow-up experiments by collaborator Gabriele Picco confirmed that *NUTM1* is highly expressed in the cell line and that SBC-3 is just as sensitive to BET inhibition as RPMI-2650 (Figure 6.11C-D).

This observation is in line with a previous study that found *NUTM1* fusions in small cell lung cancer patients (Taniyama et al., 2014). Together, our findings suggest that rare *BRD4-NUTM1* fusions may play a functional role in small cell lung cancers and may indicate a sensitivity to treatment with BET inhibitors in some small cell lung cancer patients.

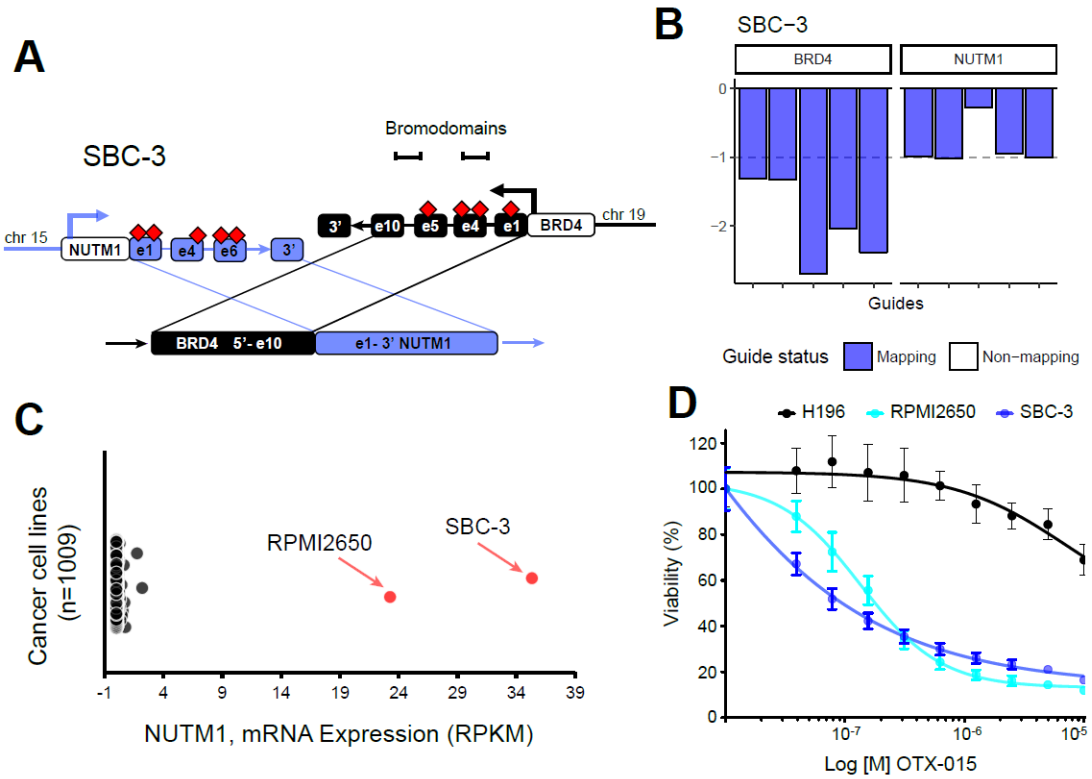


Figure 6.11: Essential BRD4-NUTM1 fusion identified in small cell lung cancer cell line SBC-3. (A) Illustrated fusion breakpoint. (B) BRD4-NUTM1 fusions are significantly depleted in the CRISPR/Cas9 analysis. (C) NUTM1 is highly expressed in SBC-3 and RPMI2650 compared to all other cancer cell lines. (D) Drug response curves of RPMI-2650, SBC-3 and a negative control small cell lung carcinoma cell line NCI-H196.

6.6 Chapter summary and discussion

6.6.1 Summary and implications of findings

In this chapter, I described the development and implementation of a novel computational approach to identify functional gene fusions using existing CRISPR/Cas9 whole genome screening data. I applied the approach on 371 cell lines for which screening data was available and tested the functional essentiality for the 2,821 fusion transcripts for which at least one sgRNA was available.

My method of mapping sgRNA effects on a per-transcript basis was able to identify 78% of known oncogenes as essential at FDR < 25%. I also highlighted several cases of recurrent fusions in the COSMIC fusion census with weak evidence of oncogenic activity (section 6.3).

Across all transcripts tested, the method identified 59 fusions in 63 events as being high-effect significantly essential (FDR < 5% and mean scaled logFC < -0.45). This represents just 2.7% of all fusions tested. I also found that within fusion essentiality scores with highly significant FDR, there was an enrichment only for known oncogenic fusions and fusions with significant drug associations. No such enrichment was found for other categories, e.g. in-frame fusions, kinase genes or fusions found in patients. Together, the evidence strongly suggests that in tumours, out of all detectable fusion transcripts, the vast majority are likely to be passenger events resulting from wide-spread structural rearrangements.

In this chapter, I described two instances where my approach identified new oncogenic roles for 1) a previously uncharacterised fusion and 2) a fused oncogene in a novel tissue context.

I showed that the fusion of the TEAD-binding domain of *YAP1* and the transcription activation domain of *MAML2* is likely to lead to constitutive activation of YAP signalling pathways in a fashion that is essential to tumour viability. Unlike most oncogenic fusions that are tissue-specific (e.g. *EWSR1-FLI1* in Ewing's sarcoma, *BCR-ABL1* in chronic myeloid leukaemia etc), *YAP1-MAML2* is detected and predicted to be essential in multiple tissue types including head and neck carcinomas, glioblastoma and ovarian carcinomas. This is perhaps less surprising considering that YAP1 overexpression has been linked to

oncogenic function ranging from poor prognosis, chemoresistance and bypass of KRAS addiction in multiple types of solid tumours. Several YAP1 inhibitors are currently in development, and in the future it would be interesting to test whether they would cause decreased viability cell lines, xenografts and tumours harbouring *YAP1-MAML2* fusions. Positive results in these types of experiments would provide evidence to support testing and treating patients for *YAP1-MAML2* positive tumours.

In a second case, the *ATG7-RAF1* fusion in a KRAS wild-type pancreatic adenocarcinoma cell lines describes a *RAF1* fusion with a novel fusion partner in a novel tissue context. Like previously described *RAF1* fusions, the upregulation of the MAPK pathway are likely driven by the deletion of N-terminal auto-inhibitory domain on *RAF1*. Cell lines with *RAF1* fusions were indeed extremely sensitive to MEK1/2 inhibition by trametinib. The *ATG7-RAF1* fusion demonstrates an alternate and potentially therapeutically targetable mechanism of MAPK pathway activation in a tissue type for which the majority of cancers are driven by *KRAS* mutations.

6.6.2 Limitations of approach

My computational approach of calculating fusion essentiality has a number of caveats. The most obvious caveats stem from the fact that the CRISPR/Cas9 whole genome guide libraries were of course originally designed with different objectives in mind.

Skewed sgRNA distribution across fused genes

In the ideal experiment, sgRNAs should be distributed equally across both fused and un-fused portion of the involved genes. However, in 45% of tests conducted (chapter 5.3.2), I observe skewed distributions where one of the fused partners has sometimes exclusively mapping guides, while the other gene partner has exclusively non-mapping guides.

In situations like those, without additional knowledge, it is difficult to call fusion-specific essentiality. False positives are more likely to occur, when a gene is essential regardless of fusion status, which is illustrated in the example of the *ERBB2-JUP* fusion which has a significant FES despite being a passenger fusion event (section 6.1.1). Additionally, in section 6.4 I show that reproducibility for unknown fusion transcripts is lower when the sgRNA are not differentially mapping.

A more suitable sgRNA library design would potentially include equal numbers of sgRNA that target the 5′ and 3′ regions of the genes. In the present situation, it is important

to bear in mind the distribution of sgRNA when analysing the results of this analysis. For fusions of potential clinical interest, it would be advisable to use follow-up validation experiments to confirm the fusion-specific essentiality before drawing strong conclusions.

Noise and unpredictability of CRISPR/Cas9 whole-genome screens

In addition to the distribution and number of sgRNA on fusions, the effectiveness of sgRNA is also a source of potential biases. Effectiveness of sgRNA can be biased by screening noise, as well as by sgRNA design effected by both known and unpredictable biases. Different sgRNA targeting the same genes can exhibit large variation in term of depletion effect size, with biases emerging for example from GC content of the guide target sequence, position of the target sequence along the gene and even the nucleotide identity at specific guide RNA positions (e.g. a guanine at position 20 is strongly preferable and adenines at positions 9-19 are favourable) (Doench et al., 2014; Wang et al., 2014).

For that reason, whole genome CRISPR/Cas9 screens prefer to deploy multiple sgRNA's per gene, ranging from 4-10 across the three data-bases. In chapter 5.3.2 I showed that screens using 5 sgRNA per gene can give good retrieval of essential genes (possibly improved by performing multiple biological replicates). In this analysis, only 76% of fusion transcripts in the screened cell lines have any mapping sgRNA. Where tests were conducted, although the median number of sgRNA are very similar (~4-5 sgRNA per fusion), the distribution is much broader and about 20% of fusion transcripts have only one or two mapping sgRNA (see chapter 5.3.2).

Improving computational techniques

Inefficiencies in sgRNA library design could be addressed through a redesign of the library. However, this is perhaps unrealistic, considering the time, resources and specialist knowledge necessary to undertake that task and to subsequently carry out the CRISPR/Cas9 drop-out screens. Instead, the computational approach could be improved to consider and highlight any potential short-comings.

In relationship to the noisiness around screening data addressed above, it would for example be interesting to be able to approximate the effectiveness of a single sgRNA across the entire gene. This may mean calculating how much it varies compared to the depletion effect on the gene as a mean of all sgRNA, or more simply by setting a threshold on the number of samples an sgRNA has a strong deleterious effect on (for instance, this could be achieved by calculating a sgRNA-level BAGEL score). sgRNAs with low

effectiveness could then be removed from the down-stream calculations, which should lead to lower noise and more accurate fusion essentiality calling.

Further, part of the analytical process involves calculating a Z-score of the scaled logFC's across all cell lines for a given sgRNA. This Z-score may distort data when there are screening artefacts leading to extreme outliers (see chapter 5.3.4). To reduce noise from distortion, it may be useful to remove cell lines with particularly bad data quality, for instance by implementing a threshold on the AUC of the precision-recall curves (see chapter 5.3.2). Also, one could investigate alternative methods to scaling corrected log fold-changes and/or to Z-normalisation. In particular, it would be interesting to see whether mean-centring instead of Z-normalising, or whether Z-normalisation after taking the difference of mapping and non-mapping sgRNA would yield better results. Good benchmarks to use would be the recovery of known oncogenic fusions and any skew towards potential biases such as amplified/deleted/essential genes (see section 6.2).

6.6.3 Future directions & conclusion

Implementing solutions to the potential caveats above may reduce noise and improve signal for any potential current false negatives. Moreover, CRISPR/Cas9 whole genome drop-out screening projects are still on-going, both at the Broad Institute (Achilles data), and here at the Sanger Institute (Project Score data). Thus, in the nearby future, we may fill out further gaps in the 640 cell lines for which there is currently no screening data available. Similarly, a joint project is currently underway that looks at the comparability and possible integration of the Achilles and Project Score data (F. Iorio, personal communications). sgRNA libraries share less than 2% of target locations across their sgRNA, and thus may complement each other where one of the libraries lack differentially mapping sgRNA. Similarly, the ability of being integrated in the future can lead to higher numbers of sgRNA per test, thus increasing robustness and reliability.

Concluding, in this chapter I have demonstrated that my computational approach is able to 1) detect known oncogenic fusions with high accuracy and 2) provide leads for novel oncogenic functionality of previously unknown fusions. My findings also suggest that truly functional, oncogenic fusions are exceedingly rare (< 2.7% in my data), and thus that the vast majority of detected fusions are likely to be passenger events. As my approach leverages high-throughput data to examine the essentiality of fusions in a large-scale

manner to examine even rare (n = 1) events, it opens up new possibilities of oncogene detection. Future advances in data collection will hopefully allow us to discover more functional oncogenic fusions as targets of personalised medicine.

7 Discussion and conclusion

At the beginning of this dissertation, I stated the aim of identifying functional gene fusions in cancer. To achieve the goal, I created a catalogue of gene fusions and developed computational approaches that integrated data from several high-throughput screens.

In this final chapter, I will first briefly summarise and discuss the important learnings and implications of the previous chapters. Finally, I conclude by reflecting on the overall contribution that this work has made and discussing the opportunities that are opening up in the future.

7.1 Interpreting fusion-calling algorithms to create a catalogue of gene fusions (Chapter 2)

I filtered and annotated processed RNA-Seq data and created a catalogue of 8,354 gene fusions in 1,011 cancer cell lines (chapter 2). My benchmarks of 945 PCR validations and two independent sets of RNA-Seq data in 23 cell lines showed that fusion transcripts called by individual algorithms can be unreliable and contain a large proportion of false positive fusions. Using multiple different fusion-calling algorithms and only considering the overlap of different fusions returned the most reliable set of fusions based on my analyses.

These conclusions are in line with other studies that examined fusion transcripts identified from RNA-Seq data. Meta-analyses of multiple fusion-calling algorithms found large discrepancies in performance, not only between fusion-calling algorithms, but even within the same algorithm depending on the type of sample (Carrara et al., 2013; Haas et al., 2017; Kumar et al., 2016). Similarly, recently published papers are now also starting to utilise multiple fusion-calling algorithms for an individual study (Gao et al., 2018). Altogether, the reliability of the algorithms is currently a weak point in the interpretation of RNA-Seq fusion calling data, and the community should be cautious when interpreting otherwise unvalidated lists of gene fusions called using single algorithms.

As of now, different analyses use different ways of benchmarking the validity of gene fusions. Some choose to validate a small subset of gene fusions by PCR, some consider breakpoints confirmed by whole genome sequencing as true positives and yet others only validate a small number of fusions that are of particular interest due to a pre-existing connection to cancer. Each approach has its own weaknesses and validation rates

are often incomparable across different studies. The community has yet to construct a gold-standard set of techniques, quality controls or benchmarks. This is further complicated by the fact that different samples can have vastly different characteristics, be it read length, read depth, sequencing type (single vs. paired-end), which may cause further discrepancies between samples called even by the same algorithm.

Altogether, fusion-calling algorithms still have a way to go before being able to reliably minimise false negatives and false positives. However, the first algorithms were only developed in 2009, and at the speed at which technologies are being developed, we may soon see rapid improvements. Most algorithms were developed with samples sequenced with ~30 M read pairs. The average RNA-Seq data used in my dissertation contains 70-230 M read pairs (corresponding to ~36-89X) that were made possible by the large-scale sequencing pipelines that are developed both by CCLE, as well as here at the Wellcome Sanger Institute. Algorithms trained on samples like this, once the correct parameters and filtering criteria are identified, may be able to utilise the high read-depth to return more accurate results as well as picking up gene fusion that are lowly expressed.

For now, the validation rate of the fusions used in my analysis is approximated at ~70%. While it is important to be mindful of potential false positives that will be found in the data, it is nonetheless possible to extract meaningful biological information on cancer-related fusions, as I showed in my subsequent analyses.

7.2 Landscape of the occurrence and recurrence of gene fusions in the cancer cell line panel (Chapter 3)

Next, I examined the landscape of this catalogue of gene fusions. I found that the occurrence of gene fusions in cancer cell lines of different tissue types of origin correlated well with those found in patient samples. Similarly, the fusions that are known to be cancer-related (e.g. *BCR-ABL1*, *ALK* fusions, *EWSR1-FLI1*), are found almost exclusively in cell lines of the expected tissue type of origin. Overall, this indicates that the cell line panel can be a useful model for gene fusions in cancer.

In this chapter I also describe how recurrent fusions are incredibly rare, with 96% of the gene fusions in our catalogue only occurring in a single cell line. In fact, only 237 fusions (1%) were recurrent in three or more samples. I found a similar pattern of recurrence in the gene fusions identified in almost 10,000 patient samples, which shows

that it is not a specific characteristic of the genomic instability of cell lines. Importantly, I found that some commonly observed known oncogenic fusions are found in our cell line collection, but not in the almost 10,000 patient samples, e.g. *EWSR1-FLI1, NPM1-ALK* and *BRD4-NUTM1*. This is likely to be due to a bias in tissue representation, illustrating that even with large overall sample sizes and with the low recurrence of gene fusions, current resources are unlikely to be comprehensive. Increasing our number of sequenced cancer types and models will likely provide opportunities to increase representation of missing cancer subtypes.

Finally, I examined the representation of gene fusions according to the predicted frame of the breakpoints. Overall, known functional fusions were overwhelmingly in-frame events, which is to be expected considering that most chimeric proteins must retain intact catalytic domains. On the other hand, the remaining gene fusions had a much wider spread of predicted fusion outcomes, including out-of-frame fusions, UTR fusions and others. As most not in-frame events are likely to be passenger events, this further underscores that the many of the fusions found in this analysis are likely to be passenger events.

Overall, this chapter was a mainly descriptive look into the occurrence and recurrence of gene fusions across our panel of 1,011 cancer cell lines. In order to delve deeper into the analysis of functional fusions, it is necessary to involve other data sets that can link the occurrence of fusions to meaningful phenotypic indicators, e.g. of cell viability.

7.3 Biomarker analysis with high-throughput screening data (Chapter 4)

In this chapter, I used high-throughput drug screening data of over 400 drug IDs to find fusions that are biomarkers of drug response. My analysis reliably identified known cancer-related drug biomarkers, with the top hits being a sensitivity of *BCR-ABl1* fused cell lines to ABL1 inhibitors such as Imatinib and the sensitivity of *NPM1-ALK* fused cell lines to *ALK* inhibitors. The detection of true positives suggest that the approach is correct, but unfortunately, no novel candidates showed strong enough evidence to warrant intensive further investigation.

A major challenge in this analysis is probably the relatively low number of recurrent gene fusions reducing the statistical power of the analysis. This low sample size also means that even statistically significant hits are difficult to interpret, as the associations may be significant due to other confounding factors or simply by chance despite having performed

multiple-hypothesis correction. At the same time, the screened drug compounds are limited largely to those that were selected for showing promise in different cancer-related settings and targets that are currently druggable. These characteristics would have aided the rapid translations of any novel findings, but also mean I only test a relatively narrow set of potential molecular targets of cell line dependencies.

These types of high-throughput drug screens have previously been successful at identifying molecular biomarkers of drug sensitivity in cancer cell lines (Barretina et al., 2012; Garnett et al., 2012), which is also demonstrated by the efficient identification of known fusion-drug biomarkers. However, in the context of identifying unknown rare gene fusions that could play a functional role in cancer, perhaps a different approach is needed.

7.4 Identifying essential fusions using CRISPR/Cas9 screening data (Chapter 5/6)

In the final chapter, I developed a novel computational approach to identify functional gene fusions based on CRISPR/Cas9 whole genome drop-out screening data for 339 cell lines. By mapping the coordinates of sgRNA of the CRISPR/Cas9 guide library to the coordinates of the gene fusions I was able to create fusion-specific depletion scores for 2,821 fusion transcripts (27% of total).

My approach was able to correctly identify many known oncogenic fusions in this data set, including *EWSR1-FLI1*, *PML-RARA, KMT2A-MLLT3* and *EML4-ALK*. Importantly, this method also highlighted novel fusions which produce a loss-of-fitness phenotype when knocked out in cell lines.

The most interesting is perhaps a fusion in *YAP1-MAML2*, which has previously been described in patient samples of head and neck carcinoma origin, but have never before been functionally validated. *YAP1-MAML2* was significantly essential in all three cell lines in which it was identified in our cell lines. *YAP1* has previously been linked to cancer in multiple solid tumours, and I hypothesise that the functionality of the fusion is driven by the activation of the TEAD binding domain of *YAP* through the transcriptional activation domain of *MAML2*. Indeed, the YAP conserved gene expression signature is significantly enriched in the three cell lines compared to all other cell lines.

That the fusion has been previously identified in the clinic gives hope that this finding may help real patients in the future. What is further interesting is that the fusion is found in three cell lines of different tissue types (head and neck, glioblastoma and ovarian

carcinoma), and appears to be significantly essential in all. This suggests that the functionality of the fusion may be tissue-type independent and argues for a broadly scoped approach to diagnosis and treatment. It would be interesting to test next whether fusion-carrying cell lines and tumours are sensitive to small molecule inhibition of YAP1 as well as transforming activity of the fusion. Further, understanding the frequency of this fusion in cancer patient populations will be important for the realisation of any treatment plans.

My analysis also identified two further clinically interesting functional fusions. The first was an *ATG7-RAF1* fusion in pancreatic cancer cell line PL-18. While this fusion has been identified previously in another pancreatic cancer cell line, PL5 (Giacomini et al., 2013), we provided evidence for the first time of the sensitivity to MEK inhibition in fused cell lines. Together, this indicates that *RAF1* fusions in pancreatic cancer cell lines are potentially clinically actionable and clinically actionable in pancreatic cancer patients, supporting further research into the clinical use of this fusion as diagnostic marker for personalised treatment.

Finally, I identified a functional *BRD4-NUTM1* fusion in small cell lung cancer cell line SBC-3. Similarly to nut midline carcinomas in which the fusion is usually found, we showed that SBC-3 is sensitive to BET inhibitors. The data thus supports that using *BRD4-NUTM1* as a biomarker may provide treatment benefit in a subset of small cell lung cancer cell patients.

My computational approach of identifying essential fusions from CRISPR/Cas9 data was able to facilitate some novel findings. Overall however, it is striking how few fusions (~3%) have evidence of significant essentiality in the cell lines. This strongly suggests that the vast majority of fusions identified in cancer models and samples are passenger events that are unlikely to contribute to cancer maintenance. This likely complicates the interpretation of RNA-Seq fusion-calling data in research and the clinic, as potentially important impactful fusions may be metaphorical needles in the haystack. Taken together with the low recurrence of even some known oncogenic fusions (e.g. *BRAF* and *RAF1* fusions), this means that identifying novel oncogenic fusions will need novel approaches beyond simply looking for highly recurrent fusions (as is currently the standard of identifying cancer-related mutations).

Using CRISPR/Cas9 data in cell lines represents one way of approaching this. There are still caveats that can be ironed out that may make the approach more impactful (for instance using larger sgRNA libraries, developing thresholds and indicators of potential false positive fusions, etc.). As more cell lines are being screened across large institutions such as the Wellcome Sanger Institute and the Broad Institute, the discoveries to be made with this computational approach may yet grow. The coverage of this analysis will also depend on the availability of models. As RNA-Seq is collected and CRISPR/Cas9 screening data is generated in cancer organoids, this approach can easily be transferrable into different model systems as well.

7.5 Final impressions

With the aim of the thesis in mind, as finding functional gene fusions in cancer, my PhD work makes the following specific contributions to the community:

1. A catalogue of gene fusions in a panel of 1,011 cancer cell lines that will be released into the public domain. In the future, this resource can be integrated with further molecular and phenotypic annotations to make scientific discoveries, and may also be used by scientists who are interested in finding cell lines with a particular gene fusions.
2. A computational approach to determine the specific essentiality of gene fusions. This approach can be replicated in future studies that generate RNA-Seq and CRISPR/Cas9 screening data, potentially leading to the discovery of further novel functional gene fusions.
3. Evidence of the functionality of the *YAP1-MAML2* fusion in three different cancer cell lines. My results suggest that knock-out of the fusion may lead to loss-of-viability in tumour, which may have direct implications for any future patients that present with this fusions in the clinic.
4. The observation that only a small proportion of gene fusions have evidence for functionality. Perhaps an expected result, my computational approach provides hard evidence that most gene fusions do not contribute significantly to the viability of cancer cell lines. This suggests that the interpretation of gene fusions identified in cancer cells should be

treated with caution, and that developing better approaches of identifying functional cancer-related gene fusions could be valuable and worthwhile.

Overall, these contributions are only a small step towards a larger goal of providing a better diagnosis, treatment and care environment for cancer patients. My findings have been built on the work of a lot of literature before me and it is my hope that they can provide a stepping stone for important scientific discoveries that are yet to come.

8 Supplementary tables

Here follows an index of the supplementary tables and titles. All supplementary tables are included as a supplement to the thesis on a USB stick.

Table 8-1: Index of supplementary tables in this dissertation.

Supp Table #	Title
1	Annotation of 1,034 human cancer cell lines used in our study and the source of RNA-seq data. All cell lines are part of the GDSC cancer cell line project (see COSMIC IDs).
2	List of primers sequences used to validate gene fusions.
3	List and annotation of 10,514 fusion transcripts found in 1,011 cell lines.
4	High-throughput cell line drug sensitivity data (IC50's) used in this study together with compound annotation. Column names are COSMIC id's of cell lines.
5	Significant results from the drug-association analysis using cancer functional events.
6	Significant results from the drug-association analysis using gene fusions.
7	Significant results from the gene-centric drug-association analysis.
8	Fusion essentiality score and significance calculation for all 2,821 fusion transcripts with mapping guides. For 525 fusion transcripts where multiple data sets contained mapping guides, both are reported.

9 Bibliography

Adams, J., Palombella, V.J., Sausville, E.A., Johnson, J., Destree, A., Lazarus, D.D., Maas, J., Pien, C.S., Prakash, S., and Elliott, P.J. (1999). Proteasome Inhibitors: A Novel Class of Potent and Effective Antitumor Agents. Cancer Res. *59*, 2615–2622.

Aguirre, A.J., Meyers, R.M., Weir, B.A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W.F., Doshi, M.B., et al. (2016). Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. *6*, 914–929.

Ahronian, L.G., Sennott, E.M., Allen, E.M.V., Wagle, N., Kwak, E.L., Faris, J.E., Godfrey, J.T., Nishimura, K., Lynch, K.D., Mermel, C.H., et al. (2015). Clinical Acquired Resistance to RAF Inhibitor Combinations in BRAF-Mutant Colorectal Cancer through MAPK Pathway Alterations. Cancer Discov. *5*, 358–367.

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. Nat. Biotechnol. *30*, 679–692.

Aplan, P.D. (2006). Causes of oncogenic chromosomal translocation. Trends Genet. *22*, 46–55.

Aurias, A., Rimbaut, C., Buffe, D., Dubousset, J., and Mazabraud, A. (1983). [Translocation of chromosome 22 in Ewing's sarcoma]. Comptes Rendus Seances Acad. Sci. Ser. III Sci. Vie *296*, 1105–1107.

Balgobind, B.V., Raimondi, S.C., Harbott, J., Zimmermann, M., Alonzo, T.A., Auvrignon, A., Beverloo, H.B., Chang, M., Creutzig, U., Dworzak, M.N., et al. (2009). Novel prognostic subgroups in childhood 11q23/MLL-rearranged acute myeloid leukemia: results of an international retrospective study. Blood *114*, 2489–2496.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Begley, C.G., and Green, A.R. (1999). The SCL Gene: From Case Report to Critical Hematopoietic Regulator. Blood *93*, 2760–2770.

Behan, F.M., Iorio, F., Gonçalves, E., Picco, G., Beaver, C., Santos, R., Rao, Y., Ansari, R., Harper, S., Jackson, D.A., et al. (2018). Prioritisation of oncology therapeutic targets using CRISPR-Cas9 screening. Manuscr. Rev.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C., et al. (2010). Integrative analysis of the melanoma transcriptome. Genome Res. *20*, 413–427.

Berns, K., Horlings, H.M., Hennessy, B.T., Madiredjo, M., Hijmans, E.M., Beelen, K., Linn, S.C., Gonzalez-Angulo, A.M., Stemke-Hale, K., Hauptmann, M., et al. (2007). A Functional Genetic Approach Identifies the PI3K Pathway as a Major Determinant of Trastuzumab Resistance in Breast Cancer. Cancer Cell *12*, 395–402.

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

Braekeleer, E.D., Douet-Guilbert, N., Rowe, D., Bown, N., Morel, F., Berthou, C., Férec, C., and Braekeleer, M.D. (2011). ABL1 fusion genes in hematological malignancies: a review. Eur. J. Haematol. *86*, 361–371.

Brenner, J.C., Feng, F.Y., Han, S., Patel, S., Goyal, S.V., Bou-Maroun, L.M., Liu, M., Lonigro, R., Prensner, J.R., Tomlins, S.A., et al. (2012). PARP-1 inhibition as a targeted strategy to treat Ewing's sarcoma. Cancer Res. *72*, 1608–1613.

Brown, L., Cheng, J.T., Chen, Q., Siciliano, M.J., Crist, W., Buchanan, G., and Baer, R. (1990). Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia. EMBO J. *9*, 3343–3351.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. *40*, 722–729.

Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., Laird, P.W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

Cancer IT (2018a). Contribute to cancerit/cgpRna development by creating an account on GitHub (Cancer IT).

Cancer IT (2018b). Gene Rearrangement AnalySiS. Contribute to cancerit/grass development by creating an account on GitHub (Cancer IT).

Cancer IT (2018c). Breakpoints via assembly: Identifies breaks and attempts to assemble rearrangements. - cancerit/BRASS (Cancer IT).

Cancer Research UK (2015). Cancer mortality statistics.

Cancerrxgene.org Home page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer.

Carrara, M., Beccuti, M., Lazzarato, F., Cavallo, F., Cordero, F., Donatelli, S., and Calogero, R.A. (2013). State-of-the-Art Fusion-Finder Algorithms Sensitivity and Specificity. BioMed Res. Int.

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., et al. (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. N. Engl. J. Med. *364*, 2507–2516.

Chen, Y., Takita, J., Choi, Y.L., Kato, M., Ohira, M., Sanada, M., Wang, L., Soda, M., Kikuchi, A., Igarashi, T., et al. (2008). Oncogenic mutations of ALK kinase in neuroblastoma. Nature *455*, 971–974.

Chen, Z., Chen, J., Gu, Y., Hu, C., Li, J.-L., Lin, S., Shen, H., Cao, C., Gao, R., Li, J., et al. (2014). Aberrantly activated AREG–EGFR signaling is required for the growth and survival of CRTC1–MAML2 fusion-positive mucoepidermoid carcinoma cells. Oncogene *33*, 3869–3877.

Chmielecki, J., Hutchinson, K.E., Frampton, G.M., Chalmers, Z.R., Johnson, A., Shi, C., Elvin, J., Ali, S.M., Ross, J.S., Basturk, O., et al. (2014). Comprehensive Genomic Profiling of Pancreatic Acinar Cell Carcinomas Identifies Recurrent RAF Fusions and Frequent Inactivation of DNA Repair Genes. Cancer Discov. *4*, 1398–1405.

Choy, E., Butrynski, J., Harmon, D., Morgan, J., George, S., Wagner, A., D'Adamo, D., Cote, G., Rubinstein, Y., Benes, C., et al. (2013). Abstract LB-174: Translation of preclinical predictive sensitivity of Ewing sarcoma to PARP inhibition: Phase II study of olaparib in adult patients with recurrent/metastatic Ewing sarcoma following failure of prior chemotherapy. Cancer Res. *73*, LB-174-LB-174.

Cleton-Jansen, A.-M., Buerger, H., and Hogendoorn, P.C.W. (2005). Central high-grade osteosarcoma of bone: Diagnostic and genetic considerations. Curr. Diagn. Pathol. *11*, 390–399.

Cocquet, J., Chong, A., Zhang, G., and Veitia, R.A. (2006). Reverse transcriptase template switching and false alternative transcripts. Genomics *88*, 127–131.

Colicelli, J. (2010). ABL Tyrosine Kinases: Evolution of Function, Regulation, and Specificity. Sci. Signal. *3*, re6–re6.

Corcoran, R.B., André, T., Atreya, C.E., Schellens, J.H.M., Yoshino, T., Bendell, J.C., Hollebecque, A., McRee, A.J., Siena, S., Middleton, G., et al. (2018). Combined BRAF, EGFR, and MEK Inhibition in Patients with *BRAF* ^{V600E} -Mutant Colorectal Cancer. Cancer Discov. *8*, 428–443.

Crowley, E., Nicolantonio, F.D., Loupakis, F., and Bardelli, A. (2013). Liquid biopsy: monitoring cancer-genetics in the blood. Nat. Rev. Clin. Oncol. *10*, 472–484.

Daigle, S.R., Olhava, E.J., Therkelsen, C.A., Basavapathruni, A., Jin, L., Boriack-Sjodin, P.A., Allain, C.J., Klaus, C.R., Raimondi, A., Scott, M.P., et al. (2013). Potent inhibition of DOT1L as treatment of MLL-fusion leukemia. Blood *122*, 1017–1025.

Dalla-Favera, R., Martinotti, S., Gallo, R.C., Erikson, J., and Croce, C.M. (1983). Translocation and rearrangements of the c-myc oncogene locus in human undifferentiated B-cell lymphomas. Science *219*, 963–967.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. Nature *417*, 949–954.

Deenik, W., Beverloo, H.B., Luytgaarde, S.C.P.A.M. van der P. de, Wattel, M.M., Esser, J.W.J. van, Valk, P.J.M., and Cornelissen, J.J. (2009). Rapid complete cytogenetic remission after upfront dasatinib monotherapy in a patient with a NUP214-ABL1-positive T-cell acute lymphoblastic leukemia. Leukemia *23*, 627–629.

Diaz Jr, L.A., Williams, R.T., Wu, J., Kinde, I., Hecht, J.R., Berlin, J., Allen, B., Bozic, I., Reiter, J.G., Nowak, M.A., et al. (2012). The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature *486*, 537–540.

Digweed, M., and Sperling, K. (2004). Nijmegen breakage syndrome: clinical manifestation of defective response to DNA double-strand breaks. DNA Repair *3*, 1207–1217.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. *29*, 15–21.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. Nat. Biotechnol. *32*, 1262–1267.

Druker, B.J., Talpaz, M., Resta, D.J., Peng, B., Buchdunger, E., Ford, J.M., Lydon, N.B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., et al. (2001). Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. N. Engl. J. Med. *344*, 1031–1037.

Druker, B.J., Guilhot, F., O'Brien, S.G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M.W.N., Silver, R.T., Goldman, J.M., Stone, R.M., et al. (2006). Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. N. Engl. J. Med. *355*, 2408–2417.

Duro, D., Bernard, O., Valle, V.D., Leblanc, T., Berger, R., and Larsen, C.-J. (1996). Inactivation of the P16INK4/MTS1 Gene by a Chromosome Translocation t(9;14)(p21–22;q11) in an Acute Lymphoblastic Leukemia of B-Cell Type. Cancer Res. *56*, 848–854.

Enlund, F., Behboudi, A., Andrén, Y., Öberg, C., Lendahl, U., Mark, J., and Stenman, G. (2004). Altered Notch signaling resulting from expression of a WAMTP1-MAML2 gene fusion in mucoepidermoid carcinomas and benign Warthin's tumors. Exp. Cell Res. *292*, 21–28.

Ferrara, N., Gerber, H.-P., and LeCouter, J. (2003). The biology of VEGF and its receptors. Nat. Med. *9*, 669.

Finta, C., and Zaphiropoulos, P.G. (2002). Intergenic mRNA Molecules Resulting from *trans* -Splicing. J. Biol. Chem. *277*, 5882–5890.

Flaherty, K.T., Robert, C., Hersey, P., Nathan, P., Garbe, C., Milhem, M., Demidov, L.V., Hassel, J.C., Rutkowski, P., Mohr, P., et al. (2012). Improved Survival with MEK Inhibition in BRAF-Mutated Melanoma. N. Engl. J. Med. *367*, 107–114.

Fogh, J., Fogh, J.M., and Orfeo, T. (1977). One Hundred and Twenty-Seven Cultured Human Tumor Cell Lines Producing Tumors in Nude Mice. JNCI J. Natl. Cancer Inst. *59*, 221–226.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

French, C.A. (2010). NUT midline carcinoma. Cancer Genet. Cytogenet. *203*, 16–20.

French, C.A., Miyoshi, I., Kubonishi, I., Grier, H.E., Perez-Atayde, A.R., and Fletcher, J.A. (2003). BRD4-NUT Fusion Oncogene: A Novel Mechanism in Aggressive Carcinoma. Cancer Res. *63*, 304–307.

French, C.A., Ramirez, C.L., Kolmakova, J., Hickman, T.T., Cameron, M.J., Thyne, M.E., Kutok, J.L., Toretsky, J.A., Tadavarthy, A.K., Kees, U.R., et al. (2008). BRD–NUT oncoproteins: a family of closely related nuclear proteins that block epithelial differentiation and maintain the growth of carcinoma cells. Oncogene *27*, 2237–2242.

Gao, Q., Liang, W.-W., Foltz, S.M., Mutharasu, G., Jayasinghe, R.G., Cao, S., Liao, W.-W., Reynolds, S.M., Wyczalkowski, M.A., Yao, L., et al. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. *23*, 227-238.e3.

Garner, M.M., and Felsenfeld, G. (1987). Effect of Z-DNA on nucleosome placement. J. Mol. Biol. *196*, 581–590.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Genomics England Genomics England.

Giacomini, C.P., Sun, S., Varma, S., Shain, A.H., Giacomini, M.M., Balagtas, J., Sweeney, R.T., Lai, E., Vecchio, C.A.D., Forster, A.D., et al. (2013). Breakpoint Analysis of Transcriptional and Genomic Profiles Uncovers Novel Gene Fusions Spanning Multiple Human Cancer Types. PLOS Genet. *9*, e1003464.

Gill, S.J., Travers, J., Pshenichnaya, I., Kogera, F.A., Barthorpe, S., Mironenko, T., Richardson, L., Benes, C.H., Stratton, M.R., McDermott, U., et al. (2015). Combinations of PARP Inhibitors with Temozolomide Drive PARP1 Trapping and Apoptosis in Ewing's Sarcoma. PLOS ONE *10*, e0140988.

Goncalves, E., Behan, F.M., Louzada, S., Arnol, D., Stronach, E., Yang, F., Yusa, K., Stegle, O., Iorio, F., and Garnett, M.J. (2018). Tandem duplications lead to loss of fitness effects in CRISPR-Cas9 data. BioRxiv 325076.

Greco, A., Fusetti, L., Miranda, C., Villa, R., Zanotti, S., Pagliardini, S., and Pierotti, M.A. (1998). Role of the TFG N-terminus and coiled-coil domain in the transforming activity of the thyroid TRK-T3 oncogene. Oncogene *16*, 809–816.

Greenman, C.D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., et al. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics *11*, 164–175.

Grimwade, D., Hills, R.K., Moorman, A.V., Walker, H., Chatters, S., Goldstone, A.H., Wheatley, K., Harrison, C.J., Burnett, A.K., and Group, on behalf of the N.C.R.I.A.L.W. (2010). Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. Blood *116*, 354–365.

Groffen, J., Stephenson, J.R., Heisterkamp, N., de Klein, A., Bartram, C.R., and Grosveld, G. (1984). Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. Cell *36*, 93–99.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Haas, B., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T., Pochet, N., et al. (2017). STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. BioRxiv 120295.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hart, T., and Moffat, J. (2016). BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics *17*.

Hart, T., Brown, K.R., Sircoulomb, F., Rottapel, R., and Moffat, J. (2014). Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol. Syst. Biol. *10*, 733.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell *163*, 1515–1526.

Harvey, K.F., Zhang, X., and Thomas, D.M. (2013). The Hippo pathway and human cancer. Nat. Rev. Cancer *13*, 246–257.

Hauschild, A., Grob, J.-J., Demidov, L.V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C.U., Miller, W.H., Kaempgen, E., et al. (2012). Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. The Lancet *380*, 358–365.

Hernández, L., Pinyol, M., Hernández, S., Beà, S., Pulford, K., Rosenwald, A., Lamant, L., Falini, B., Ott, G., Mason, D.Y., et al. (1999). TRK-Fused Gene (TFG) Is a New Partner of ALK in Anaplastic Large Cell Lymphoma Producing Two Structurally DifferentTFG-ALK Translocations. Blood *94*, 3265–3268.

Iorio, F. (2018). R implementation of the BAGEL method to call for gene essentiality significance: francescojm/BAGELR.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell *166*, 740–754.

Iorio, F., Behan, F.M., Gonçalves, E., Bhosle, S.G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., et al. (2018). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. BMC Genomics *19*, 604.

Iwahara, T., Fujimoto, J., Wen, D., Cupples, R., Bucay, N., Arakawa, T., Mori, S., Ratzkin, B., and Yamamoto, T. (1997). Molecular characterization of ALK, a receptor tyrosine kinase expressed specifically in the nervous system. Oncogene *14*, 439.

Janknecht, R. (2005). EWS–ETS oncoproteins: The linchpins of Ewing tumors. Gene *363*, 1–14.

Jänne, P.A., van den Heuvel, M., Barlesi, F., Cobo, M., Mazieres, J., Crinò, L., Orlov, S., Blackhall, F., Wolf, J., Garrido, P., et al. (2016). Selumetinib in combination with docetaxel as second-line treatment for patients with KRAS-mutant advanced NSCLC: Results from the phase III SELECT-1 trial. Ann. Oncol. *27*.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Jones, D.T.W., Kocialkowski, S., Liu, L., Pearson, D.M., Ichimura, K., and Collins, V.P. (2009). Oncogenic RAF1 rearrangement and a novel BRAF mutation as alternatives to KIAA1549:BRAF fusion in activating the MAPK pathway in pilocytic astrocytoma. Oncogene *28*, 2119–2123.

Kalyana-Sundaram, S., Shankar, S., DeRoo, S., Iyer, M.K., Palanisamy, N., Chinnaiyan, A.M., and Kumar-Sinha, C. (2012). Gene Fusions Associated with Recurrent Amplicons Represent a Class of Passenger Aberrations in Breast Cancer. Neoplasia N. Y. N *14*, 702–708.

Kapoor, A., Yao, W., Ying, H., Hua, S., Liewen, A., Wang, Q., Zhong, Y., Wu, C.-J., Sadanandam, A., Hu, B., et al. (2014). Yap1 Activation Enables Bypass of Oncogenic Kras Addiction in Pancreatic Cancer. Cell *158*, 185–197.

Kim, D., and Salzberg, S.L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol. *12*, R72.

Kitagawa, M. (2016). Notch signalling in the nucleus: roles of Mastermind-like (MAML) transcriptional coactivators. J. Biochem. (Tokyo) *159*, 287–294.

Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., et al. (2015). A comprehensive transcriptional portrait of human cancer cell lines. Nat. Biotechnol. *33*, 306–312.

Koivunen, J.P., Mermel, C., Zejnullahu, K., Murphy, C., Lifshits, E., Holmes, A.J., Choi, H.G., Kim, J., Chiang, D., Thomas, R., et al. (2008). EML4-ALK Fusion Gene and Efficacy of an ALK Kinase Inhibitor in Lung Cancer. Clin. Cancer Res. *14*, 4275–4283.

Kopetz, S., Desai, J., Chan, E., Hecht, J.R., O'Dwyer, P.J., Lee, R.J., Nolop, K.B., and Saltz, L. (2010). PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. J. Clin. Oncol. *28*, 3534–3534.

Krebs, M.G., Metcalf, R.L., Carter, L., Brady, G., Blackhall, F.H., and Dive, C. (2014). Molecular analysis of circulating tumour cells—biology and biomarkers. Nat. Rev. Clin. Oncol. *11*, 129–144.

Kumar, S., Vo, A.D., Qin, F., and Li, H. (2016). Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Sci. Rep. *6*, 21597.

Kwak, E.L., Bang, Y.-J., Camidge, D.R., Shaw, A.T., Solomon, B., Maki, R.G., Ou, S.-H.I., Dezube, B.J., Jänne, P.A., Costa, D.B., et al. (2010). Anaplastic Lymphoma Kinase Inhibition in Non–Small-Cell Lung Cancer. N. Engl. J. Med. *363*, 1693–1703.

Kwon, Y.-W., Kim, I.-J., Wu, D., Lu, J., Stock, W.A., Liu, Y., Huang, Y., Kang, H.C., DelRosario, R., Jen, K.-Y., et al. (2012). Pten Regulates Aurora-A and Cooperates with Fbxw7 in Modulating Radiation-Induced Tumor Development. Mol. Cancer Res. *10*, 834–844.

Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N. Engl. J. Med. *372*, 2509–2520.

Ledergerber, C., and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. Brief. Bioinform. *12*, 489–497.

Lee, C.K., Wu, Y.-L., Ding, P.N., Lord, S.J., Inoue, A., Zhou, C., Mitsudomi, T., Rosell, R., Pavlakis, N., Links, M., et al. (2015a). Impact of Specific Epidermal Growth Factor Receptor (EGFR) Mutations and Clinical Characteristics on Outcomes After Treatment With EGFR Tyrosine Kinase Inhibitors Versus Chemotherapy in EGFR-Mutant Lung Cancer: A Meta-Analysis. J. Clin. Oncol. *33*, 1958–1965.

Lee, K.-W., Lee, S.S., Kim, S.-B., Sohn, B.H., Lee, H.-S., Jang, H.-J., Park, Y.-Y., Kopetz, S., Kim, S.S., Oh, S.C., et al. (2015b). Significant Association of Oncogene YAP1 with Poor Prognosis and Cetuximab Resistance in Colorectal Cancer Patients. Clin. Cancer Res. *21*, 357–364.

Lee, M., Lee, K., Yu, N., Jang, I., Choi, I., Kim, P., Jang, Y.E., Kim, B., Kim, S., Lee, B., et al. (2017). ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. Nucleic Acids Res. *45*, D784–D789.

Li, H., Ma, X., Wang, J., Koontz, J., Nucci, M., and Sklar, J. (2007). Effects of rearrangement and allelic exclusion of JJAZ1/SUZ12 on cell proliferation and survival. Proc. Natl. Acad. Sci. *104*, 20001–20006.

Li, H., Wang, J., Mor, G., and Sklar, J. (2008). A Neoplastic Gene Fusion Mimics Trans-Splicing of RNAs in Normal Human Cells. Science *321*, 1357–1361.

Li, N., Zhang, Y., Han, X., Liang, K., Wang, J., Feng, L., Wang, W., Songyang, Z., Lin, C., Yang, L., et al. (2015). Poly-ADP ribosylation of PTEN by tankyrases promotes PTEN degradation and tumor growth. Genes Dev. *29*, 157–170.

Li, W., Xu, H., Xiao, T., Cong, L., Love, M.I., Zhang, F., Irizarry, R.A., Liu, J.S., Brown, M., and Liu, X.S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. *15*, 554.

Lin, R.-Z., and Chang, H.-Y. (2008). Recent advances in three-dimensional multicellular spheroid culture for biomedical research. Biotechnol. J. *3*, 1172–1184.

Lorenz, S., Barøy, T., Sun, J., Nome, T., Vodák, D., Bryne, J.-C., Håkelien, A.-M., Fernandez-Cuesta, L., Möhlendick, B., Rieder, H., et al. (2015). Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. Oncotarget *7*, 5273–5288.

Lu, H., Villafane, N., Dogruluk, T., Grzeskowiak, C.L., Kong, K., Tsang, Y.H., Zagorodna, O., Pantazi, A., Yang, L., Neill, N.J., et al. (2017). Engineering and Functional Characterization of Fusion Genes Identifies Novel Oncogenic Drivers of Cancer. Cancer Res. *77*, 3502–3512.

Maemondo, M., Inoue, A., Kobayashi, K., Sugawara, S., Oizumi, S., Isobe, H., Gemma, A., Harada, M., Yoshizawa, H., Kinoshita, I., et al. (2010). Gefitinib or Chemotherapy for Non–Small-Cell Lung Cancer with Mutated EGFR. N. Engl. J. Med. *362*, 2380–2388.

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C., Yu, J., et al. (2009a). Chimeric transcript discovery by paired-end transcriptome sequencing. Proc. Natl. Acad. Sci. *106*, 12353–12358.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., and Chinnaiyan, A.M. (2009b). Transcriptome sequencing to detect gene fusions in cancer. Nature *458*, 97–101.

Mali, P., Esvelt, K.M., and Church, G.M. (2013). Cas9 as a versatile tool for engineering biology. Nat. Methods *10*, 957–963.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and

pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886.

Masters, J.R.W. (2000). Human cancer cell lines: fact and fantasy. Nat. Rev. Mol. Cell Biol. *1*, 233–236.

McDermott, U. (2015). Next-generation sequencing and empowering personalised cancer medicine. Drug Discov. Today *20*, 1470–1475.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G.F., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., et al. (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput. Biol. *7*, e1001138.

McWhirter, J.R., Galasso, D.L., and Wang, J.Y. (1993). A coiled-coil oligomerization domain of Bcr is essential for the transforming function of Bcr-Abl oncoproteins. Mol. Cell. Biol. *13*, 7587–7595.

Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. Nat. Rev. Cancer *15*, 371–381.

Meyer, C., Kowarz, E., Hofmann, J., Renneville, A., Zuna, J., Trka, J., Abdelali, R.B., Macintyre, E., Braekeleer, E.D., Braekeleer, M.D., et al. (2009). New insights to the *MLL* recombinome of acute leukemias. Leukemia *23*, 1490–1499.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Mossé, Y.P., Wood, A., and Maris, J.M. (2009). Inhibition of ALK Signaling for Cancer Therapy. Clin. Cancer Res. *15*, 5609–5614.

Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G., et al. (2016). CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov. *6*, 900–913.

Murai, J., Zhang, Y., Morris, J., Ji, J., Takeda, S., Doroshow, J.H., and Pommier, Y. (2014). Rationale for Poly(ADP-ribose) Polymerase (PARP) Inhibitors in Combination Therapy with Camptothecins or Temozolomide Based on PARP Trapping versus Catalytic Inhibition. J. Pharmacol. Exp. Ther. *349*, 408–416.

Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C.S., Stinson, J., Peters, B.A., Modrusan, Z., Jung, K., Seshagiri, S., et al. (2011). Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med. Genomics *4*.

Nassar, A., Lundgren, K., Pomerantz, M., Van Allen, E.M., Choudhury, A.D., Harshman, L.C., Preston, M.A., Mouw, K.W., Wei, X.X., McGregor, B.A., et al. (2018). FGFR3-TACC3 fusion in

bladder cancer: Enrichment in the young, never-smokers, and Asians. J. Clin. Oncol. *36*, 465–465.

NCI Genomic Data Commons (2018). TCGA Study Abbreviations | NCI Genomic Data Commons.

Nguyen, A.T., Taranova, O., He, J., and Zhang, Y. (2011). DOT1L, the H3K79 methyltransferase, is required for MLL-AF9–mediated leukemogenesis. Blood *117*, 6912–6922.

NICE (2018). Crizotinib for treating ROS1-positive advanced non-small-cell lung cancer | Guidance and guidelines | NICE.

Nigro, L.L., Mirabile, E., Tumino, M., Caserta, C., Cazzaniga, G., Rizzari, C., Silvestri, D., Buldini, B., Barisone, E., Casale, F., et al. (2013). Detection of *PICALM-MLLT10* (*CALM-AF10*) and outcome in children with T-lineage acute lymphoblastic leukemia. Leukemia *27*, 2419–2421.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Nowell, P.C. (2007). Discovery of the Philadelphia chromosome: a personal perspective. J. Clin. Invest. *117*, 2033–2035.

Nowell, P., and Hungerford, D. (1960). A minute chromosome in chronic granulocytic leukemia. Science 1497.

Obholzer, N.D., Haas, B.J., Landau, D.-A., Pochet, N., Regev, A., and Wu, C. (2015). Abstract 4859: Development of a cancer transcriptome analysis toolkit: identification of gene fusions in chronic lymphocytic leukemia. Cancer Res. *75*, 4859–4859.

Okada, Y., Jiang, Q., Lemieux, M., Jeannotte, L., Su, L., and Zhang, Y. (2006). Leukaemic transformation by CALM-AF10 involves upregulation of Hoxa5 by hDOT1L. Nat. Cell Biol. *8*, 1017–1024.

Ou, S.-H.I., Bang, Y.-J., Camidge, D.R., Riely, G.J., Salgia, R., Shapiro, G., Solomon, B.J., Engelman, J.A., Kwak, E.L., Clark, J.W., et al. (2013). Efficacy and safety of crizotinib in patients with advanced ROS1-rearranged non-small cell lung cancer (NSCLC). J. Clin. Oncol. *31*, 8032–8032.

Pailler, E., Adam, J., Barthélémy, A., Oulhen, M., Auger, N., Valent, A., Borget, I., Planchard, D., Taylor, M., André, F., et al. (2013). Detection of Circulating Tumor Cells Harboring a Unique ALK Rearrangement in ALK-Positive Non–Small-Cell Lung Cancer. J. Clin. Oncol. *31*, 2273–2281.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K., et al. (2010). Rearrangements of the *RAF* kinase pathway in prostate cancer, gastric cancer and melanoma. Nat. Med. *16*, 793–798.

Panagopoulos, I. (2010). Absence of the JAZF1/SUZ12 chimeric transcript in the immortalized non-neoplastic endometrial stromal cell line T HESCs. Oncol. Lett. *1*, 947–950.

Pao, W., Wang, T.Y., Riely, G.J., Miller, V.A., Pan, Q., Ladanyi, M., Zakowski, M.F., Heelan, R.T., Kris, M.G., and Varmus, H.E. (2005). KRAS Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib. PLOS Med. *2*, e17.

Paraiso, K.H.T., Thakur, M.D., Fang, B., Koomen, J.M., Fedorenko, I.V., John, J.K., Tsao, H., Flaherty, K.T., Sondak, V.K., Messina, J.L., et al. (2015). Ligand-Independent EPHA2 Signaling Drives the Adoption of a Targeted Therapy–Mediated Metastatic Melanoma Phenotype. Cancer Discov. *5*, 264–273.

Pauli, C., Hopkins, B.D., Prandi, D., Shaw, R., Fedrizzi, T., Sboner, A., Sailer, V., Augello, M., Puca, L., Rosati, R., et al. (2017). Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. Cancer Discov. *7*, 462–477.

Perner, S., Wagner, P.L., Demichelis, F., Mehra, R., Lafargue, C.J., Moss, B.J., Arbogast, S., Soltermann, A., Weder, W., Giordano, T.J., et al. (2008). EML4-ALK fusion lung cancer: a rare acquired event. Neoplasia N. Y. N *10*, 298–302.

Planchard, D., Besse, B., Groen, H.J.M., Souquet, P.-J., Quoix, E., Baik, C.S., Barlesi, F., Kim, T.M., Mazieres, J., Novello, S., et al. (2016). Dabrafenib plus trametinib in patients with previously treated BRAFV600E-mutant metastatic non-small cell lung cancer: an open-label, multicentre phase 2 trial. Lancet Oncol. *17*, 984–993.

Prahallad, A., Sun, C., Huang, S., Nicolantonio, F.D., Salazar, R., Zecchin, D., Beijersbergen, R.L., Bardelli, A., and Bernards, R. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature *483*, 100–103.

Quintás-Cardama, A., Tong, W., Manshouri, T., Vega, F., Lennon, P.A., Cools, J., Gilliland, D.G., Lee, F., Cortes, J., Kantarjian, H., et al. (2008). Activity of tyrosine kinase inhibitors against human *NUP214-ABL1*-positive T cell malignancies. Leukemia *22*, 1117–1124.

Reck, M., and Rabe, K.F. (2017). Precision Diagnosis and Treatment for Advanced Non–Small-Cell Lung Cancer. N. Engl. J. Med. *377*, 849–861.

Rickman, D.S., Pflueger, D., Moss, B., VanDoren, V.E., Chen, C.X., de la Taille, A., Kuefer, R., Tewari, A.K., Setlur, S.R., Demichelis, F., et al. (2009). SLC45A3-ELK4 Is a Novel and Frequent Erythroblast Transformation-Specific Fusion Transcript in Prostate Cancer. Cancer Res. *69*, 2734–2738.

Roberts, J.D., Preston, B.D., Johnston, L.A., Soni, A., Loeb, L.A., and Kunkel, T.A. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. Mol. Cell. Biol. *9*, 469–476.

Roberts, K.G., Morin, R.D., Zhang, J., Hirst, M., Zhao, Y., Su, X., Chen, S.C., Payne-Turner, D., Churchman, M.L., Harvey, R.C., et al. (2012). Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia., Genetic alterations

activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell Cancer Cell *22*, *22*, 153, 153–166.

Robinson, D.R., Kalyana-Sundaram, S., Wu, Y.-M., Shankar, S., Cao, X., Ateeq, B., Asangani, I.A., Iyer, M., Maher, C.A., Grasso, C.S., et al. (2011). Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. Nat. Med. *17*, 1646–1651.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. *24*, 227–235.

Rowley, J.D. (1973). A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. Nature *243*, 290–293.

Sakamoto, K.M., and Frank, D.A. (2009). CREB in the Pathophysiology of Cancer: Implications for Targeting Transcription Factors for Cancer Therapy. Clin. Cancer Res. *15*, 2583–2587.

Sambrook, J., and Russell, D.W. (2006). Purification of Nucleic Acids by Extraction with Phenol:Chloroform. Cold Spring Harb. Protoc. *2006*, pdb.prot4455.

Sato, T., Vries, R.G., Snippert, H.J., Wetering, M. van de, Barker, N., Stange, D.E., Es, J.H. van, Abo, A., Kujala, P., Peters, P.J., et al. (2009). Single Lgr5 stem cells build crypt–villus structures *in vitro* without a mesenchymal niche. Nature *459*, 262–265.

Schadendorf, D., Hodi, F.S., Robert, C., Weber, J.S., Margolin, K., Hamid, O., Patt, D., Chen, T.-T., Berman, D.M., and Wolchok, J.D. (2015). Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma. J. Clin. Oncol. *33*, 1889–1894.

Sedic, M., Skibinski, A., Brown, N., Gallardo, M., Mulligan, P., Martinez, P., Keller, P.J., Glover, E., Richardson, A.L., Cowan, J., et al. (2015). Haploinsufficiency for BRCA1 leads to cell-type-specific genomic instability and premature senescence., Haploinsufficiency for BRCA1 leads to cell-type-specific genomic instability and premature senescence. Nat. Commun. Nat. Commun. *6*, *6*, 7505–7505.

Shaw, A.T., Kim, D.-W., Nakagawa, K., Seto, T., Crinó, L., Ahn, M.-J., De Pas, T., Besse, B., Solomon, B.J., Blackhall, F., et al. (2013). Crizotinib versus Chemotherapy in Advanced *ALK* -Positive Lung Cancer. N. Engl. J. Med. *368*, 2385–2394.

Shaw, A.T., Ou, S.-H.I., Bang, Y.-J., Camidge, D.R., Solomon, B.J., Salgia, R., Riely, G.J., Varella-Garcia, M., Shapiro, G.I., Costa, D.B., et al. (2014). Crizotinib in *ROS1* -Rearranged Non–Small-Cell Lung Cancer. N. Engl. J. Med. *371*, 1963–1971.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Singh, D., Chan, J.M., Zoppoli, P., Niola, F., Sullivan, R., Castano, A., Liu, E.M., Reichel, J., Porrati, P., Pellegatta, S., et al. (2012). Transforming Fusions of FGFR and TACC Genes in Human Glioblastoma. Science *337*, 1231–1235.

Slebos, R.J.C., Kibbelaar, R.E., Dalesio, O., Kooistra, A., Stam, J., Meijer, C.J.L.M., Wagenaar, S.S., Vanderschueren, R.G.J.R.A., van Zandwijk, N., Mooi, W.J., et al. (1990). K-ras Oncogene Activation as a Prognostic Marker in Adenocarcinoma of the Lung. N. Engl. J. Med. *323*, 561–565.

Smith, M.A., Morton, C.L., Phelps, D., Girtman, K., Neale, G., and Houghton, P.J. (2008). SK-NEP-1 and Rh1 are Ewing family tumor lines. Pediatr. Blood Cancer *50*, 703–706.

Solit, D.B., Garraway, L.A., Pratilas, C.A., Sawai, A., Getz, G., Basso, A., Ye, Q., Lobo, J.M., She, Y., Osman, I., et al. (2006). BRAF mutation predicts sensitivity to MEK inhibition. Nature *439*, 358–362.

Solomon, B.J., Mok, T., Kim, D.-W., Wu, Y.-L., Nakagawa, K., Mekhail, T., Felip, E., Cappuzzo, F., Paolini, J., Usari, T., et al. (2014). First-Line Crizotinib versus Chemotherapy in *ALK* -Positive Lung Cancer. N. Engl. J. Med. *371*, 2167–2177.

Song, G., Li, Y., and Jiang, G. (2012). Role of VEGF/VEGFR in the pathogenesis of leukemias and as treatment targets (Review). Oncol. Rep. *28*, 1935–1944.

Song, S., Honjo, S., Jin, J., Chang, S.-S., Scott, A.W., Chen, Q., Kalhor, N., Correa, A.M., Hofstetter, W.L., Albarracin, C.T., et al. (2015). The Hippo Coactivator YAP1 Mediates EGFR Overexpression and Confers Chemoresistance in Esophageal Cancer. Clin. Cancer Res. *21*, 2580–2590.

Stanulla, M., Chhalliyil, P., Wang, J., Jani-Sait, S.N., and Aplan, P.D. (2001). Mechanisms of MLL gene rearrangement: site-specific DNA cleavage within the breakpoint cluster region is independent of chromosomal context. Hum. Mol. Genet. *10*, 2481–2491.

Stathis, A., Zucca, E., Bekradda, M., Gomez-Roca, C., Delord, J.-P., Rouge, T. de L.M., Uro-Coste, E., Braud, F. de, Pelosi, G., and French, C.A. (2016). Clinical Response of Carcinomas Harboring the BRD4–NUT Oncoprotein to the Targeted Bromodomain Inhibitor OTX015/MK-8628. Cancer Discov. *6*, 492–500.

Stephens, P.J., McBride, D.J., Lin, M.-L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., et al. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature *462*, 1005–1010.

Stewart, E., Goshorn, R., Bradley, C., Griffiths, L.M., Benavente, C., Twarog, N.R., Miller, G.M., Caufield, W., Freeman, B.B., Bahrami, A., et al. (2014). Targeting the DNA repair pathway in Ewing sarcoma. Cell Rep. *9*, 829–841.

Stirnweiss, A., Oommen, J., Kotecha, R.S., Kees, U.R., and Beesley, A.H. (2017). Molecular-genetic profiling and high-throughput in vitro drug screening in NUT midline carcinoma—an aggressive and fatal disease. Oncotarget *8*, 112313–112329.

Storlazzi, C.T., Steyern, F.V.V., Domanski, H.A., Mandahl, N., and Mertens, F. (2005). Biallelic somatic inactivation of the NF1 gene through chromosomal translocations in a sporadic neurofibroma. Int. J. Cancer *117*, 1055–1057.

Stransky, N., Cerami, E., Schalm, S., Kim, J.L., and Lengauer, C. (2014). The landscape of kinase fusions in cancer. Nat. Commun. *5*, 4846.

Suda, K., Tomizawa, K., and Mitsudomi, T. (2010). Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation. Cancer Metastasis Rev. *29*, 49–60.

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K.J., Reinhold, W., Guo, Y., Kruh, G.D., Reimers, M., et al. (2004). Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells. Cancer Cell *6*, 129–137.

Taniyama, T.K., Nokihara, H., Tsuta, K., Horinouchi, H., Kanda, S., Fujiwara, Y., Yamamoto, N., Koizumi, F., Yunokawa, M., and Tamura, T. (2014). Clinicopathological Features in Young Patients Treated for Small-Cell Lung Cancer: Significance of Immunohistological and Molecular Analyses. Clin. Lung Cancer *15*, 244–247.

Thandla, S.P., Ploski, J.E., Raza-Egilmez, S.Z., Chhalliyil, P.P., Block, A.W., Jong, P.J. de, and Aplan, P.D. (1999). ETV6-AML1 Translocation Breakpoints Cluster Near a Purine/Pyrimidine Repeat Region in the ETV6 Gene. Blood *93*, 293–299.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. Science *310*, 644–648.

Tomlins, S.A., Laxman, B., Varambally, S., Cao, X., Yu, J., Helgeson, B.E., Cao, Q., Prensner, J.R., Rubin, M.A., Shah, R.B., et al. (2008). Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. Neoplasia N. Y. N *10*, 177–188.

Tran, B., Kopetz, S., Tie, J., Gibbs, P., Jiang, Z.-Q., Lieu, C.H., Agarwal, A., Maru, D.M., Sieber, O., and Desai, J. (2011). Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. Cancer *117*, 4623–4632.

Tu, K., Yang, W., Li, C., Zheng, X., Lu, Z., Guo, C., Yao, Y., and Liu, Q. (2014). Fbxw7 is an independent prognostic marker and induces apoptosis and growth arrest by regulating YAP abundance in hepatocellular carcinoma. Mol. Cancer *13*, 110.

Tzelepis, K., Koike-Yusa, H., De Braekeleer, E., Li, Y., Metzakopian, E., Dovey, O.M., Mupo, A., Grinkevich, V., Li, M., Mazan, M., et al. (2016). A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. *17*, 1193–1205.

Valouev, A., Weng, Z., Sweeney, R.T., Varma, S., Le, Q.-T., Kong, C., Sidow, A., and West, R.B. (2014). Discovery of recurrent structural variants in nasopharyngeal carcinoma. Genome Res. *24*, 300–309.

van de Wetering, M., Francies, H.E., Francis, J.M., Bounova, G., Iorio, F., Pronk, A., van Houdt, W., van Gorp, J., Taylor-Weiner, A., Kester, L., et al. (2015). Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients. Cell *161*, 933–945.

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391.

Vassilev, L.T., Vu, B.T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., et al. (2004). In Vivo Activation of the p53 Pathway by Small-Molecule Antagonists of MDM2. Science *303*, 844–848.

Verdine, G.L., and Walensky, L.D. (2007). The Challenge of Drugging Undruggable Targets in Cancer: Lessons Learned from Targeting BCL-2 Family Members. Clin. Cancer Res. *13*, 7264–7270.

Vlierberghe, P.V., and Ferrando, A. (2012). The molecular basis of T cell acute lymphoblastic leukemia. J. Clin. Invest. *122*, 3398–3406.

Wang, H., Lin, H., Pan, J., Mo, C., Zhang, F., Huang, B., Wang, Z., Chen, X., Zhuang, J., Wang, D., et al. (2016). Vasculogenic Mimicry in Prostate Cancer: The Roles of EphA2 and PI3K. J. Cancer *7*, 1114–1124.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science *343*, 80–84.

Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell *168*, 890-903.e15.

Wang, X.-S., Shankar, S., Dhanasekaran, S.M., Ateeq, B., Sasaki, A.T., Jing, X., Robinson, D., Cao, Q., Prensner, J.R., Yocum, A.K., et al. (2011). Characterization of KRAS Rearrangements in Metastatic Prostate Cancer. Cancer Discov. *1*, 35–43.

Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., et al. (1997). An Information-Intensive Approach to the Molecular Pharmacology of Cancer. Science *275*, 343–349.

Williams, S.V., Hurst, C.D., and Knowles, M.A. (2013). Oncogenic FGFR3 gene fusions in bladder cancer. Hum. Mol. Genet. *22*, 795–803.

Wu, L., Sun, T., Kobayashi, K., Gao, P., and Griffin, J.D. (2002). Identification of a Family of Mastermind-Like Transcriptional Coactivators for Mammalian Notch Receptors. Mol. Cell. Biol. *22*, 7688–7700.

Wu, L., Liu, J., Gao, P., Nakamura, M., Cao, Y., Shen, H., and Griffin, J.D. (2005). Transforming activity of MECT1-MAML2 fusion oncoprotein is mediated by constitutive CREB activation. EMBO J. *24*, 2391–2402.

Yang, L., Lee, M.-S., Lu, H., Oh, D.-Y., Kim, Y.J., Park, D., Park, G., Ren, X., Bristow, C.A., Haseley, P.S., et al. (2016). Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. Am. J. Hum. Genet. *98*, 843–856.

Yoshihara, K., Wang, Q., Torres-Garcia, W., Zheng, S., Vegesna, R., Kim, H., and Verhaak, R.G. (2015). The landscape and therapeutic relevance of cancer-associated transcript fusions., The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene Oncogene *34*, *34*, 4845, 4845–4854.

Zech, L., Haglund, U., Nilsson, K., and Klein, G. (1976). Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with burkitt and non-burkitt lymphomas. Int. J. Cancer *17*, 47–56.

Zhang, Y., Gong, M., Yuan, H., Park, H.G., Frierson, H.F., and Li, H. (2012). Chimeric Transcript Generated by cis-Splicing of Adjacent Genes Regulates Prostate Cancer Cell Proliferation. Cancer Discov. *2*, 598–607.

Zhou, R.-H., Wang, P., Zou, Y., Jackson-Cook, C.K., and Povirk, L.F. (1997). Complex Formation by Topoisomerase II in Amsacrine-treated Chinese Hamster. 5.

Zuber, J., Rappaport, A.R., Luo, W., Wang, E., Chen, C., Vaseva, A.V., Shi, J., Weissmueller, S., Fellman, C., Taylor, M.J., et al. (2011). An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. Genes Dev. *25*, 1628–1640.