Guiding Cancer Therapy: Evidence-driven Reporting of Genomic Data

Dissertation
for the award of the degree

Doctor of Philosophy

Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen
within the doctoral program *Molecular Biology of Cells*
of the Georg-August-University School of Science (GAUSS)

submitted by
Júlia Perera Bel
from Barcelona

Göttingen, 2018

Thesis Committee:

Prof. Dr. Tim Beißbarth
Department of Medical Statistics
University Medical Center Göttingen

Prof. Dr. Ulrich Sax
Department of Medical Informatics
University Medical Center Göttingen

Prof. Dr. Edgar Wingender
Department of Bioinformatics
University Medical Center Göttingen

Members of the Examination Board:

1st Referee: Prof. Dr. Tim Beißbarth
Department of Medical Statistics
University Medical Center Göttingen

2nd Referee: Prof. Dr. Ulrich Sax
Department of Medical Informatics
University Medical Center Göttingen

Further members of the Examination Board:

Prof. Dr. Edgar Wingender
Department of Bioinformatics
University Medical Center Göttingen

Prof. Dr. Burkhard Morgenstern
Department of Bioinformatics
Georg August University Göttingen

Prof. Dr. Gregor Bucher
Department of Developmental Biology
Georg August University Göttingen

Prof. Dr. Dieter Kube
Department of Haematology and Oncology
University Medical Center Göttingen

Date of oral examination: 19th of November 2018

Declaration

I herby declare that this doctoral thesis entitled "Guiding Cancer Therapy: Evidence-driven Reporting of Genomic Data" has been written independently with no other sources and aids than those quoted.

Göttingen, October 5th, 2018

Júlia Perera Bel

Acknowledgements

I want to thank in the very first place my supervisor Prof. Tim Beißbarth for giving me the opportunity to learn so much in your group, for guiding me along this work and for envisioning beyond my ideas.

I am also very grateful to Dr. med. Annalen Bleckmann for your very valuable clinical insights and feedback. I would like to thank the thesis committee members, Prof. Ulrich Sax and Prof. Edgar Wingender, for very constructive discussions.

I would like to acknowledge all collaborators of this work: Dr. Barbara Hutter, Dr. Christoph Heining, Dr. Martina Fröhlich, Prof. Stefan Fröhling, Prof. Hanno Glimm and Prof. Benedikt Brors. Their contributions and expert insights, as well as the data provided, are crucial to the work presented in this thesis.

I am very grateful to current and former colleagues for the great environment, without which the quality of my (working) time wouldn't have been the same. Special thanks to Astrid, Mish and Xenia for your friendship. Also, I am deeply grateful to Mish and Astrid for reading this thesis.

Many thanks to my friends, for never doubting that I would return driving a wagon.

Special thanks to my family, Aurora, Joan and Enric. Because your support shortens distances.

The warmest thanks goes to Igor, because we shape each other personally, politically and scientifically.

Júlia Perera Bel
Göttingen
October 5th, 2018

Abstract

Next Generation Sequencing (NGS) has been crucial for the breakthrough experienced by cancer genomics during the last decade. In turn, the knowledge gathered has fostered the development of targeted drugs and genomics-driven cancer treatment. Some university hospitals have built the infrastructure and invested in human resources for the implementation of NGS within precision medicine initiatives. However, the expertise required to integrate the data with available knowledge spans several disciplines; the information to decipher the clinical implications codified in the genome of a tumor is scattered across many resources; and the complexity of the data demands of computational support.

In the anticipation of a widespread use of clinical sequencing, this thesis describes an evidence-based workflow, the *Molecular Tumor Board* (MTB) Report, aimed at paving the way for genomics-driven oncology. Deciding whether or not a molecular alteration entails clinical action (i.e. if the variant is actionable) involves a wide-range of expertise and the need to keep up with the pace of new discoveries (e.g. clinical trials, conferences, preclinical studies). The workflow presented here uses public resources to narrow somatic variants from a tumor's genomic profile down to actionable variants. Furthermore, actionable variants are classified into a six-level system based on the evidence that supports the actionability. The variables considered are cancer type in which the evidence exists and grade of predictive association between a variant and a drug. The classified variants and the evidence that supports their actionability are detailed in a concise report to support clinical discussions. To increase the usability and availability of the workflow, it has been implemented as a web-based application. The user can provide custom data as well as explore a public dataset. Actionable variants can be visualized in an interactive setting and downloaded in the aforementioned report format or in a tabular data file.

The MTB Report workflow was tested over two different large public datasets to evaluate its scope and strengths, *The Cancer Genome Atlas* (TCGA) and *Genomics Evidence Neoplasia Information Exchange* (GENIE). The results concerning variants currently used to guide treatment were in line with published numbers of patients receiving genomics-driven therapies. The results also suggested that these numbers could be increased to a large extent if low-evidence (clinical and preclinical evidence) and predictive associations that have not been established for the cancer type in the patient being tested

(i.e. off-label) were considered.

A retrospective comparison study was performed for a subset of patients from the *Molecularly Aided Stratification for Tumor Eradication Research* (MASTER) precision medicine program. The variants identified by the MTB Report were compared to the variants suggested by the experts of the MASTER program. The results showed high concordance between both approaches, as the majority of expert suggestions were identified by the workflow. The workflow identified a plethora of other variants, that, though not yet actionable, depicted a comprehensive landscape of the actionability of the patient.

In all, this thesis work established a computational workflow aimed at enabling a widespread use of NGS for guiding clinical decisions. We envision that such efforts on standardizing genomic data interpretation and reporting will become useful resources in the field of precision medicine.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AACR	American Association for Cancer Research
ALL	Acute Lymphocytic Leukemia
AML	Acute Myeloid Leukemia
bp	Base Pair
CIViC	Clinical Interpretation of Variants in Cancer
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CNV	Copy Number Variations
ctDNA	Circulating Tumor DNA
DGIdb	Drug Gene Interaction Database
DNA-Seq	DNA Sequencing
EMA	European Medicines Agency
FDA	Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
FISH	Fluorescence *in situ* Hybridization
GATK	Genome Analysis Toolkit
Gb	Giga Base Pairs
GIST	Gastrointestinal Stromal Tumor
GoF	Gain-of-Function
GDKD	Gene Drug Knowledge Database
GENIE	Genomics Evidence Neoplasia Information Exchange
HGNC	*Human Genome Organization* (HUGO) Gene Nomenclature Committee
HGVS	Human Genome Variation Society

HUGO	Human Genome Organization
IHC	Immunohistochemistry
IQR	Interquartile Range
indel	Small Insertions or Deletions
LoF	Loss-of-Function
mAb	Monoclonal Antibodies
MAF	Mutation Annotation Format
MASTER	Molecularly Aided Stratification for Tumor Eradication Research
Mb	Mega Base Pairs
MTB	Molecular Tumor Board
NCCN	National Comprehensive Cancer Network
NCT	National Center for Tumor Diseases
NGS	Next Generation Sequencing
NSCLC	Non-Small Cell Lung Cancer
OS	Overall Survival
PCR	Polymerase Chain Reaction
PDX	Patient-derived Xenografts
PMKB	Precision Medicine Knowledgebase
PFS	Progression-free Survival
RNA-Seq	RNA Sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural variations
TARGET	Tumor Alterations Relevant for Genomics-driven Therapy
TCGA	The Cancer Genome Atlas
VCF	Variant Call Format
VUS	Variants of Unknown Significance
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

PART I

INTRODUCTION

Chapter 1

Cancer Genomics

– a Breakthrough in Cancer Therapy –

Cancer is a heterogeneous disease characterized by an uncontrolled cell growth and the ability to spread to distant body parts (i.e. metastasize). Though cancer survival rates are higher now than in the past, the number of new cases per year keeps increasing and cancer is among the leading causes of death worldwide. For many decades, cancer treatment has relied almost entirely on cytotoxic agents (i.e chemotherapy) and *one-size-fits-all* approach. This type of treatment essentially ignores the underlying biology of the disease by attacking all dividing cells and consequently provoking highly impairing side effects. Despite this inability to target only cancer cells, cytotoxic agents are largely used and prove effective for many patients. However, the ability of tumor cells to overcome cytotoxicity and the difficulty in finding the right dosage show the need for new treatment strategies more in accordance with new molecular discoveries. Indeed, conventions for diagnosis and treatment decisions are still based, to a large extent, on morphological aspects (anatomic site, staging, histology) rather than molecular aspects (expressed proteins, DNA mutations, gene signatures) (Levy et al., 2012).

Prior to the rise of *Next Generation Sequencing* (NGS) techniques, efforts on identifying cancer-causing genes led to the development of first targeted agents in the late 90s[1]. However, the availability of NGS allowed a better understanding of the molecular mechanisms of cancer development, which, in turn, has translated into a faster and significant development of targeted

(1) FDA approves trastuzumab (*Herceptin*) in 1998 for *HER2* positive breast cancer and imatinib (*Gleevec*) in 2001 for Philadelphia chromosome-positive (*BCR-ABL* fusion gene) chronic myeloid leukemia

agents. NGS has also the potential to characterize patient genomes at the point of care to better profile the disease and guide treatment. As NGS gains sensitivity, speed and throughput in detecting molecular alterations, the main challenge is filtering and contextualizing the alterations that are relevant to make clinical decisions (Good et al., 2014). These alterations are here referred to as *actionable*.

Deciding whether or not a molecular alteration is actionable requires expertise in several areas (oncology, molecular pathology, bioinformatics, genetics). Hence, the best setting to incorporate genomic data into treatment planning is the *Molecular Tumor Board* (MTB). MTBs are multidisciplinary meetings in which complex cancer cases are discussed among a team of experts. However, expertise to interpret genomic findings requires keeping up with the pace of new discoveries (e.g. preclinical studies, conferences, clinical trials). For that, computational algorithms, data integration methods and informatics infrastructures are crucial to provide rigorous clinical interpretation of comprehensive genomic data (Garraway et al., 2013). Therefore, the aim of this thesis is reducing the complexity and workload that represents including genomic data in clinical decisions. To address this aim, a tool that reports clinically relevant genomic findings (i.e. actionable variants) was developed.

1.1 High-throughput Genomic Measurement Techniques

First discoveries in cancer genomics where achieved with the use of cytogenetic and molecular biology techniques, such as chromosome banding, *Fluorescence in situ Hybridization* (FISH) or *Polymerase Chain Reaction* (PCR). In 2003, The Human Genome Project completed the first draft of the human genome (International Human Genome Sequencing Consortium, 2004). Performed entirely with Sanger-based sequencing technologies, it was a tipping point in the genomics field, as it provided the community with a reference genome. Having a reference genome is crucial to identify alternative variants in a given sample: from single-base substitutions, to *Small Insertions or Deletions* (indel), to *Copy Number Variations* (CNV) and to *Structural variations* (SV) (rearrangements and inversions). As regards to single-base substitutions, a single *Base Pair* (bp) position in which alternative alleles exist in more than 1% of the population is known as *Single Nucleotide Polymorphism* (SNP); otherwise, the term is *Single Nucleotide Variant* (SNV).

The reference genome was widely used for the design of probes for array-based high-throughput methods. SNP arrays were initially designed to interrogate allele frequencies of thousands of SNPs for genome-wide association

studies, but have also been used to quantify copy number events. The allele-specific intensity measures can be used to identify copy number imbalances across the genome and have also the potential to detect copy-neutral loss-of-heterozygosity (Cooper et al., 2008). Affymetrix SNP Array 6.0 is designed to interrogate around 2 million probes, half of which are SNP and the other half are copy number probes. In contrast, array comparative genomic hybridization (array-CGH) measures the fluorescence intensity ratio between two labeled samples hybridized to the array, that, in turn, is proportional to the ratio of DNA copy numbers in the two samples. On the downside, this method is only able to detect unbalanced chromosomal abnormalities (i.e. those that affect copy number, such as reciprocal translocations and inversions) (Shinawi and Cheung, 2008). A combination of both methods (SNP-CGH) has also been used in cancer to detect both kinds of events (Peiffer et al., 2006).

In 2008, a second generation of sequencing technologies –also known as NGS, *high-throughput* or *massively parallel* sequencing– revolutionized again the field: throughput and accuracy increased, and costs per base decreased beyond any expectations (see curve of costs per base over time in `www.genome.gov/sequencingcostsdata`). The applications of NGS have from then onwards spread to a large number of research fields, creating the terminology (*-omics*) for the study of particular molecular layers: genomics (DNA), transcriptomics (RNA), proteomics, epigenomics (DNA-protein interactions), and similar. In turn, the high-dimensional data generated with high-throughput techniques (e.g. NGS or mass spectrometry), are commonly referred to as *omics*.

1.1.1 DNA Sequencing Approaches

There are mainly two approaches for *DNA Sequencing* (DNA-Seq): *Whole-Genome Sequencing* (WGS) and targeted sequencing. On the one hand, WGS offers the most comprehensive characterization of a genome as it has the power to detect alterations from single-base substitutions (SNV) to chromosomal rearrangements (SV). Yet, the large amount of sequencing required to have a standard coverage (that is, the average number of times (x) each bp is covered) is very costly. For instance, a human genome has 3 thousand million bp, or *Giga Base Pairs* (Gb); so 30x coverage requires the generation of 90 Gb per sample. On the other hand, targeted sequencing is often preferred as a most cost-effective approach. *Whole-Exome Sequencing* (WES) is a type of targeted sequencing, in which protein-coding regions (i.e. exons, which account for 1% of the genome) are captured during the library preparation. This technique

generates higher coverage at the regions of interest compared to WGS, as 75x coverage of a human exome (30 million bases, or *Mega Base Pairs* (Mb)) requires the generation of around 3 Gb per sample[2]. Hence, it provides higher throughput at lower cost and, in turn, requires less input DNA.

DNA-Seq consist of three main steps: library preparation, sequencing and data analysis. The specific combination of platforms and protocols determines the type and quality of data obtained. Library preparation protocols convert the isolated DNA into standard libraries suitable for the sequencing machine. DNA is cleaved into short fragments (final read length depends on the specific technology, ranging from 50 to 700 bp) and the DNA fragments are ligated to adapters (they will bind to primers attached to the surface in which sequencing will take place). Assessing quality and quantity of the DNA before and after library preparation is recommended to identify degraded, fragmented, and low-purity samples (Illumina, 2017). For many applications, DNA needs to be amplified to have enough bulk input amount. Sequencing generally consists of a cluster amplification step (PCR-based) that is required to generate enough signal to be detected by the measurement instruments. Next, the actual measurement happens in parallel for all reads. Different methods are used depending on the technology: sequencing-by-synthesis approach (Illumina platforms), pyrosequencing (454 platforms), or sequencing by ligation (SOLiD platforms). Despite the important advances in high-throughput and costs, most protocols still require a non-negligible amount of input DNA and an amplification step which introduces technical biases (e.g. PCR duplicates). Also, the short read length poses a challenge for downstream computational processing. All these limitations are overcome by third-generation sequencing –Nanopore and PacBio– platforms; however, error rates of these technologies are not yet acceptable for most applications. For a comprehensive review of the different technologies, the reader is referred to Metzker (2010).

1.1.2 DNA Sequencing Data Analysis

For some organisms (mainly model organisms) we have good representative assemblies of the species genome (i.e reference genome). In such cases, the bioinformatic pipeline to analyze DNA-Seq data consists of i) quality check of the raw reads; ii) map reads to reference genome; iii) mark (or remove) duplicates arisen from PCR; and iv) actual detection of variants (i.e. variant calling). The large amount of generated data and the short lengths of reads

(2) Exon capture efficiency rates are around 70%. For more information, visit: `https://genohub.com/exome-sequencing-library-preparation/#inefficiency`

have forced the development of new bioinformatic algorithms, open source software, pipelines and data formats for sequencing data analysis. The reader is referred to Bao et al. (2014) for a review of software for WES and to the *Genome Analysis Toolkit* (GATK) (McKenna et al., 2010) best practices for a detailed recommendation of pipelines[3].

The most challenging aspects are being able to reconstruct the genome using short reads, detecting duplicates from the PCR step and differentiating sequencing errors from actual nucleotide variations in the data in the variant calling step (e.g. SNV and indel) (Bao et al., 2014). The latter is especially important, as the majority of genomic studies focus on variant discovery. In cancer genomics, variant discovery is aimed at identifying germline and somatic variants. Germline variants are genomic alterations that are present in germ cells and are therefore passed to all cells of the offspring (i.e. inherited). In contrast, somatic variants are genetic alterations acquired during life and are not transmitted to the offspring (e.g. variants present only in tumor cells). Somatic variant calling is an application of particular interest for cancer research, and ideally uses matched tumor and normal samples. The simplest approach consists of separately calling variants in matched tumor and normal samples and subtracting calls found in the normal (germline) from variants found in the tumor (somatic+germline). This approach is used by VarScan2 (Koboldt et al., 2012). In contrast, other algorithms, such as MuTect (Cibulskis et al., 2013) and Strelka (Saunders et al., 2012), simultaneously call variants using information from both matched samples and do not assume diploidy (Xu, 2018).

Though SNV calling is the most popular application, NGS has the power to detect CNVs and other SVs overcoming many of the limitations of array-based approaches (e.g. hybridization noise, limited coverage and resolution). CNVs from sequencing data are detected using combined information from allele frequencies, depth of coverage and read level information. In contrast to array based methods, the specific breakpoints of CNVs can be inferred from soft clipped reads (a read that spans over two separate regions of the chromosome and therefore only one end of the read was mapped) (Tattini et al., 2015). Currently, low-coverage WGS is recommended for CNVs detection as it gives a genome-wide picture, offers a reliable and even sequence coverage and is usually PCR-free (Sims et al., 2014). WES can also be used to detect CNVs, but with certain limitations. Different efficiency of probe hybridization yields an uneven coverage distribution that affects CNV calling. Besides, the

(3) https://software.broadinstitute.org/gatk/best-practices/

probability to find breakpoints within the exome is very low (Zhao et al., 2013). The same applies to the detection of fusion genes, as both WGS or WES can potentially detect chimeric transcripts (with the difference that WES is restricted to coding regions whereas WGS can detect, for example, promoter-gene fusions). Yet, the functional impact of the chimeric transcript can only be asserted by gene expression data (for instance, *RNA Sequencing* (RNA-Seq)). Therefore, a combination of shallow WGS and RNA-Seq rises as an optimal solution for many applications (Mertens et al., 2015).

1.1.3 Public Data Repositories for Omics Data

The idea of sharing NGS data started with The Human Genome Project (International Human Genome Sequencing Consortium, 2004). Without a draft of the human reference genome, bioinformatics mapping algorithms would need to be more complex (as it happens in *de novo* sequencing for non-model organisms). Other projects followed, such as HapMap (International HapMap Consortium, 2003), ENCODE (The ENCODE Project Consortium, 2007) and 1000 Genomes (The 1000 Genomes Project Consortium, 2010). In the field of cancer genomics, *The Cancer Genome Atlas* (TCGA) project is one of the largest, most well-known effort of multi-omics cancer data generation. It consists of matched tumor-normal samples from over 11,000 patients across 33 cancer types, covering 7 data types measured with 15 different techniques. The International Cancer Genome Consortium (ICGC) currently coordinates 55 cancer research projects (including TCGA) that generate *omics* data of cancer patients. The last of such sequencing efforts is the *American Association for Cancer Research* (AACR)'s project *Genomics Evidence Neoplasia Information Exchange* (GENIE), which focuses on advanced cancer patients and has the aim of standardizing the aggregation, registering and sharing of NGS and clinical data.

Data generated by two of these projects (TCGA and GENIE) are analyzed within the scope of this thesis. Such resources of *omics* measurements linked to clinical data are crucial for a successful identification of molecular traits linked to the disease (i.e. biomarkers). Moreover, multi-center studies allow gathering of high sample sizes that would otherwise be impossible and are necessary to establish any kind of statistical inference. Thus, the growing availability of high-throughput technologies and international efforts to generate, gather, analyze and share cancer *omics* data have fostered the perfect setting for promoting cancer genomic research.

1.2 Cancer is an Acquired Genetic Disease

Cancer is the manifestation of sequential alterations accumulated in the genome and epigenome of cells during lifetime. Some of these alterations can have an inherited origin (i.e. germline) and cause susceptibility to develop cancer, such as *BRCA1/2* mutations in breast and ovarian cancer (Miki et al., 1994; Wooster et al., 1994; Foulkes, 2014). However, the majority of mutations are acquired (i.e. somatic) and therefore only present in a subset of cells in the organism. Cells with cancer-driving alterations gain growth advantages in contrast to normal cells by the deregulation of pathways from three crucial molecular processes: cell survival, genome maintenance and cell fate (Vogelstein et al., 2013). These advantages have been classified into 10 core biological capabilities of tumor cells, which are known as the *hallmarks of cancer* (Hanahan and Weinberg, 2000, 2011).

Pan-cancer genomic studies have shown that, as opposed to the traditional taxonomy of cancer merely based on the site/tissue of origin, there are cross-tissue patterns at a molecular level. For example, basal breast cancer and endometrial serous-like carcinoma share *ATM* mutations, *BRCA1* and *BRCA2* inactivation, *RB1* loss and *CCNE1* amplification (The Cancer Genome Atlas Network, 2012). Squamous malignancies also present common molecular patterns (Yan et al., 2010). These observations have led to suggest that *cell-of-origin* might be the main factor underlying molecular patterns in cancer (Hoadley et al., 2014, 2018).

It has been observed that mutation load and mutational signatures correlate with cancer type (e.g. childhood malignancies and leukemia have few mutation events, whereas lung cancer or melanomas high mutation load due to external mutagens). However, cancer type only explains about half of the genomic variability between tumor samples (Lawrence et al., 2013). Underlying biology and driver mechanisms differ from patient to patient, even within the same cancer type. As a matter of fact, though some genes are highly tumor type specific (e.g. *APC* and *KRAS* in colorectal cancer, *VHL* in kidney malignancies), mutations in these genes are also observed in other cancer types at lower frequencies (Kandoth et al., 2013). Hence, no gene is always mutated nor exclusively altered in just one tumor type. This genomic heterogeneity is also observed within the same patient: between metastases, within a metastasis and within one tumor. Indeed, heterogeneity is as crucial to cancer development as genetic diversity is to evolution.

1.2.1 Mutational Processes drive Tumor Progression

Large-scale pan-cancer sequencing of human tumors has made it possible to understand the extent to which specific alterations contribute to cancer development. As it has been said, cancer is the consequence of sequentially acquired alterations. Thus, a cell –or a group of cells– evolves from benign to malign with the acquisition of alterations, each of which confers a selective growth advantage (i.e. increased *fitness* in terms of population genetics) over surrounding cells. However, this is an iterative process: a cell with a selective growth advantage is selected, followed by an expansion of this clone. In turn, the progeny of this clone will acquire new alterations (genetic diversification) and a new clone will be selected, and so on (Greaves and Maley, 2012). The combination of genetic heterogeneity, selective pressure from the environment and competition with surrounding cells with other genomic landscapes (subclones) gives rise to simultaneous linear and branching evolution processes. As a consequence, and again using concepts from population genetics, some genomic alterations are shared by all tumor cells (clonal alterations) whereas others are only present in subclones. Furthermore, as this is a dynamic process, the genomic architecture might change with time and/or with other selective pressures, such as treatment, giving rise to resistances (Yates and Campbell, 2012).

There is a difference between those genomic alterations generated during the genetic diversification step (but without any effect on growth) and those that are actually selected. Or, in other words, *passenger* (no effect on cell's fitness) and *driver* (direct or indirect positive effect on cell's fitness) events. As a matter of fact, solid tumors have between tens and hundreds of mutations that affect the protein primary structure, yet only from 2 to 10 mutation are within driver genes (Tamborero et al., 2013; Kandoth et al., 2013). A proper identification of driver events (and genes) has only been possible with the advent of NGS and international initiatives like TCGA, which have provided enough sample size and resolution to identify frequent alterations within and across tumor types. However, Chang et al. (2018a) observed that driver genes present different rates of mutation discovery when increasing sample size, and showed that just a few have reached saturation. Recent studies set the number of cancer driver genes around 800 (Tamborero et al., 2018a). As more and more cancer-causing genes are identified it becomes clear that the catalog of driver mutations is not yet complete.

Genomic alterations can affect the protein products by either enhancing or diminishing their function. The former are known as *Gain-of-Function* (GoF)

alterations; the latter, as *Loss-of-Function* (LoF). Following this line of thought, a gene that, when affected by a GoF alteration, increases the cell's fitness, is known as *oncogene*; likewise, a gene that, when affected by a LoF alteration, increases the cell's fitness, is known as *tumor-suppressor gene*. Biologically, genes that favor proliferation, growth, survival and migration are potential oncogenes. On the contrary, tumor-suppressor genes are responsible for ensuring genome stability, acting as cell growth checkpoints and promoting apoptosis.

In general terms, the aim of new therapeutic strategies is to design drugs that i) inhibit oncogenes or ii) restore or compensate the function of tumor-suppressor genes. Whereas the former is achievable with small molecules or *Monoclonal Antibodies* (mAb) that block their enzymatic function or avoid ligand binding, the latter would require an introduction of a functional protein. For more details, see §1.3.

1.2.2 Effects of Genomic Variants on Protein Function

Back in the 60s, Phildelphia chromosome was the first recurrent chromosomal abnormality discovered by cytogenetic analyses in *Chronic Myeloid Leukemia* (CML). Yet, the mechanism through which this chromosomal rearrangement contributed to cancer development was not known. The establishment of FISH allowed the characterization of fusion oncogones *BCR-ABL* in CML (Shtivelman et al., 1985) and *MYC* fusions in Burkitt lymphoma (Leder et al., 1983). Subsequent improvements in molecular techniques –first and second generation sequencing platforms – allowed the discovery of a large variety of genomic events causing GoF and LoF of *oncogenes* and *tumor-suppressor genes*, respectively. Table 1.1 summarizes the main genomic events that occur in cancer cells.

SNVs and indels located within gene-coding or promoter regions can change the protein function by modifying the coded aminoacid (missense, frameshift) or by truncating the protein (nonsense), among others. However, predicting the functional impact and thus the effect of the mutation (namely GoF or LoF) is a difficult task, as illustrated by the controversy of both tumor-suppressor and oncogenic mutations in *TP53* (Deppert, 2007). As far as CNVs are concerned, the main challenge is the correct identification of the target gene (i.e. gene that, when deleted or amplified, is responsible for the selective growth advantage). This is especially difficult for broad CNVs, which are large in size (may affect hundreds of genes) and shallow in amplitude. As for focal CNVs, the effect on target genes might be more visible, since they are small in

Table 1.1: **Types of genomic alterations.** Table adapted from Good et al. (2014).

Event type	Description	Types	Example
Single nucleotide variants (SNV)	Single-base substitutions	Missense, nonsense, nonstop, silent	*BRAF* c.1799T>A (V600E)
Small insertions or deletions (indel)	Small numbers of nucleotides deleted or inserted	Frameshift, inframe	*PTEN* c.800delA (K267fs*9)
Structural variations (SV)	Large-scale genomic rearrangements (can yield fusion genes if breakpoints fall close to or within gene coding regions)	Translocation, inversion, and CNV	*FLT3* tandem duplication; *BCR-ABL* fusion
Copy number variations (CNV)	Broad (large-scale) or focal (1Kb-3Mb) gains or losses of DNA. They are types of SV.	Amplification, deletion	*ERBB2* amplification

size but deep in amplitude (peak amplifications of oncogenes, homozygous deletions of tumor-suppressor genes). Most cancer cells present both SNVs and CNVs, but, interestingly, Ciriello et al. (2013) observed a pattern of mutual exclusivity between high numbers of SNVs and CNVs, which they called the *cancer genome hyperbola*. In other words, it means that some tumors are characterized by many SNVs, others by many CNVs, but never by both. Finally, with regard to SV, when breakpoints of a translocation, inversion or CNV fall within gene-coding regions, portions of these two genes can fuse and create a fusion (chimeric) gene. Most common products are oncogenes by means of i) transcriptional deregulation (the promoter of one gene affects the expression of the other, as in *TRA-MYC*) or ii) hybrid proteins with abnormal enzymatic activity (enhanced tyrosine kinase activity in *BCR-ABL* and *PML-RARA*). Nonetheless, gene truncations can create tumor-suppressor fusion genes (e.g. *CDKN2A-NF1* prevents normal activity of *CDKN2A*) (Mertens et al., 2015).

1.3 Targeted Therapies

Understanding the commonalities among tumor genomes has allowed a rational development of drugs. The so-called *targeted drugs* target the aberrant protein products of driver genes, which are not only necessary for tumor evolution, but also for its survival. Thanks to phenomena like oncogene addiction or synthetic lethality (see *Targeting strategy*), targeted drugs can selectively kill tumor cells while leaving normal cells intact and, thus, maximizing clinical responses and minimizing side effects. This rationale was already used in the development of the first tyrosine kinase inhibitors (imatinib) and mAb

(trastuzumab). However, cancer genomics has increased the pace of clinical translation over the last years (Chin et al., 2011). Several aspects of targeted therapies are important for understanding their mechanisms of action:

- **Targeting strategy**. The easiest and most common approach is directly blocking an oncogene to which tumor cells are addicted (e.g. *EGFR*, *MET* inhibitors). Yet, some oncogenes are difficult to target and *indirect targeting* of downstream effectors is the only possibility to efficiently block their oncogenic function (e.g. *MEK*, *AKT* inhibitors). Targeting tumor-suppressor genes is even more complex since they are usually not expressed due to mutations, deletions or silencing. Restoring their function is theoretically possible through gene therapy (introducing a functional copy of the gene), but most attempts have failed so far due to the low or uncontrolled expression of transfered genes (Guo et al., 2014). A more effective approach consists of restoring a tumor-suppressor activity by targeting its negative regulators (e.g. inhibition of *MDM2*, a degrader of p53). Synthetic lethal interactions can also be exploited for drug design. Synthetic lethality is the condition in which inhibition of two independent molecules has little effect on a cell, but a simultaneous inhibition yields cell death (Lord et al., 2015). PARP inhibitors exploit this phenomenon: the loss of BRCA1/2 activity makes tumor cells dependent on PARP function to repair double strand DNA breaks. Thus, inhibiting PARP dooms BRCA1/2 deficient cells to cell death because they are not able to repair DNA breaks [4] (Farmer et al., 2005; Fong et al., 2010; Ledermann et al., 2014; Lord and Ashworth, 2017) .

- **Type of molecule**. Targeted drugs are divided into two main groups according to their molecular nature: i) small molecules, that due to their small molecular weight (<900 Da) can easily permeate the cell membrane and block intracellular signaling (e.g. kinase and proteasome inhibitors); and ii) biological substances, prepared or derived from living organisms, including vaccines, hormones, cytokines, gene therapy, growth factors, viruses and mAb. Kinase inhibitors and mAbs account for most targeted drugs (Table A.1).

(4) PARP inhibitor olaparib was approved by FDA in 2014 for germline *BRCA*-mutated ovarian tumors, expanded later to breast tumors (U.S. Food and Drug Administration, 2014, 2018).

- **Target pathway/process/function**. Cancer is caused by the deregulation of molecular processes that provide tumor cells with the *hallmarks of cancer*. Targeted drugs aim at restoring the original function of the deregulated pathways by targeting specific proteins. Table A.1 shows targeted drugs grouped by the pathways they target.

- **Target gene(s)**. A targeted drug is characterized by blocking specific proteins essential for tumor cells. Some drugs selectively target one protein (e.g. most mAb), but more commonly, they target a group of proteins of the same family (e.g. ponatinib inhibits ABL, FLT3 and FGFR1/2/3; afatinib inhibits EGFR and HER2; pan-PI3K inhibitor copanlisib, inhibits predominantly alpha and delta isoforms). Information retrieved from `http://dgidb.org/`.

- **Target mutation**. Targeted drugs have mutation specificity. Different genomic alterations in the same gene can trigger diverse phenotypes and, in turn, pharmacological responses. For example, imatinib successfully inhibits *ABL*, *PDGFB* and *PDGFRA* fusion genes, as well as mutated *KIT*. However, imatinib fails to generate response against *ABL* mutations and D816 mutation in *KIT* (Heinrich et al., 2003; Jabbour et al., 2006). In a similar fashion, *EGFR* inhibitors erlotinib and gefitinib accomplish high response rates in lung tumors with mutations in exons 19 and 21 (Lynch et al., 2004; Pao et al., 2004), but fail against T790M resistance mutation in exon 20 or when *EFGR* is amplified or overexpressed (Pao et al., 2005; Sone et al., 2015). Therefore, considering down to mutation resolution is crucial for drug prescription as well as for drug prioritization and repurposing.

- **Cancer type**. Genomic alterations conferring sensitivity to a targeted drug are context specific. For instance, while inhibition with vemurafenib is very efficient in *BRAF* V600E mutant melanoma (Chapman et al., 2011), colorectal patients with the same mutation show little response to vemurafenib (Kopetz et al., 2010). In this context, the tumor microenvironment bypasses the drug effect by increasing *EGF*, and, instead, a combination of *BRAF* plus *EGFR* inhibitor is recommended (Prahallad et al., 2012).

1.3.1 Predictive Biomarkers and Companion Diagnostics

A biomarker is a measurable indicator that is associated with some clinical feature (e.g. *Progression-free Survival* (PFS), *Overall Survival* (OS), diagnosis, treatment response, etc.). Any genomic alteration can be a biomarker

(SNV, CNV, fusions, etc.), but also RNA, proteins or metabolites. Common classifications divide biomarkers into *prognostic*, *diagnostic* and *predictive* biomarkers. Prognostic biomarkers provide information on clinical outcome (e.g. future onset of a disease, chance of survival). Diagnostic biomarkers inform about presence of a disease or disease subtype. Finally, a biomarker that dichotomizes the clinical outcome upon treatment is known as predictive biomarker. In other words, a predictive biomarker stratifies patients into responders and non-responders (Figure 1.1).

In the context of targeted drugs, the biological rationale supports the idea that a tumor should be sensitive to a drug the target of which is mutated in the tumor. For instance, erlotinib was first approved for pretreated unselected *Non-Small Cell Lung Cancer* (NSCLC) patients after showing improved PFS and OS compared to standard therapy. Studies showed that *EGFR* mutations were common among responsive patients, and later phase III trials showed that patient selection based on *EGFR* mutations could better stratify patients benefiting from *EGFR* inhibitors: patients with *EGFR* mutations showed longer PFS with *EGFR*-inhibitors, whereas patients with wild-type *EGFR* showed longer PFS with chemotherapy (Kobayashi and Hagiwara, 2013). Since 2013, medical guidelines recommend testing for *EGFR* mutations for first-line treatment of advanced NSCLC[5] (Lindeman et al., 2013).

This logical approach of checking the mutational status of drug targets works in some cases, such as erlotinib, but it does not necessarily have to. For instance, cetuximab or panitumumab are anti-*EGFR* mAb whose prescriptions are not bound to *EGFR* status –their target– but to *KRAS* status –mutations in which predict drug resistance (Lièvre et al., 2006). The same is true for drugs that indirectly target an oncogene or tumor-suppressor gene: *BRCA1/2* mutations or copy-number loss predict response to *PARP* inhibitors. Therefore, to differentiate between a drug target and alterations that make cells sensitive (or resistant) to a certain targeted agent, the latter are referred to as predictive biomarkers.

A list of *predictive* biomarkers included as part of a drug label by *Food and Drug Administration* (FDA) can be found in Table A.2. Both FDA and *European Medicines Agency* (EMA) use the terminology *companion diagnostic* to specify that a biomarker test is required for a drug's prescription, and, as such, it is

(5) in 2013 FDA included EGFR status in erlotinib's and afatinib's indications; in 2015, gefitinib's (Cancer Network, 2013; U.S. Food and Drug Administration, 2013; Kazandjian et al., 2016)

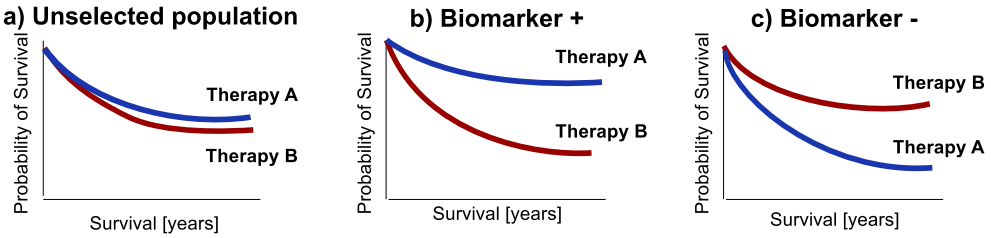


FIGURE 1.1: **Concept of predictive biomarkers depicted in Kaplan-Meier survival curves**. a) Probability of survival of patients randomized to treatment A (e.g. targeted drug) and treatment B (e.g. chemotherapy) shows no substantial benefit of one treatment over the other. However, a subset analysis shows that b) biomarker-positive patients are more likely to respond to therapy A, whereas c) biomarker negative patients are more likely to survive under therapy B.

considered a medical device with all its regulatory implications. However, for a biomarker to be included in the drug indication, a randomized clinical trial has to show that a drug performs better than the standard treatment in the biomarker-positive arm, but not in the biomarker negative arm (or the other way around for biomarkers predicting resistance). Yet, achieving the required sample size appears to be very challenging for some tumor entities due to the fact that most genes are altered at low frequencies across cancer types (Kandoth et al., 2013; Lawrence et al., 2014). To address this issue, tumor agnostic strategies enriching for patients with the biomarker are commonly used as a proof-of-concept in early phase trials, which, in the best case, will lead to cancer type specific clinical studies.

Chapter 2

Genomics-driven Oncology

– Sequencing Data and its Interpretation Challenges –

Cancer research is currently inconceivable without genomics. This technology has provided a comprehensive picture of the molecular mechanisms of cancer and has largely guided the development of new therapeutic strategies. Furthermore, the instruments that make genomic studies possible (i.e. high-throughput sequencing technologies) are being developed at an incredibly fast pace, exceeding all expectations. Today, sequencing a human genome costs around a thousand dollars and takes a few hours; it is possible to sequence tumors, circulating tumor cells and even sequence at a single-cell level. Therefore, these technologies are of high interest not only for research but also for advanced clinical diagnostics.

Testing thousands of molecular biomarkers at the same time allows an easier and more elaborate diagnosis of complex and heterogeneous diseases such as cancer. It opens the possibility to stratify patients into different subtypes and thus choose an optimal "personalized" treatment option for each patient. This concept opposes to "one-size-fits-all" medicine and is embraced by all the terms ambiguously used by scientific literature and press: *personalized*, *precision*, *stratified*, *biomarker-driven*, *genomics-driven* medicine. These terms stress different aspects of the same idea: tailoring clinical management (e.g. prevention, diagnosis and treatment) to the molecular characteristics of the disease.

Over the last years, we can find growing number of examples in which high-dimensional molecular data –*omics* data– have been used to redefine disease classifications. For instance, gene signatures are measurements of a combination of genes that have been extensively used as biomarkers. PAM50

is one of the most widely known gene expression signature, which defines five breast cancer molecular subtypes that are not only largely concordant with *Immunohistochemistry* (IHC) subtypes but have added prognostic and predictive information (Parker et al., 2009). Similar approaches have been used to define molecular subtypes in colorectal (Guinney et al., 2015), gastric (Lei et al., 2013) and pancreatic cancer (Bailey et al., 2016), just to mention some examples. As illustrated by these studies, molecular subtypes can be defined by a group of biomarkers – signature – but also by single biomarkers. For instance, NSCLC subtypes are defined by the genomic status of individual genes: *EGFR*, *KRAS*, *ALK*, *MET*, *BRAF*, *HER2*, *ROS1* – of which *EGFR*, *ALK*, *BRAF* and *ROS1* subtypes have already a companion diagnostic (see Table A.2) (Vargas and Harris, 2016). Also, we can talk about pan-cancer subtypes, such as those with *NTRK* fusions which account only for 0.5% of all solid and hematologic tumors but have shown durable responses to NTRK inhibitors (Vaishnavi et al., 2015; Drilon et al., 2018). These new subtypes are known as molecular subtypes (in contrast to classical histological or clinical subtypes).

Omics technologies have revolutionized the field of cancer research by capturing information of all molecular layers of cells which play an important role in the disease, namely the transcriptome, the epigenome, the genome, and so on. Interestingly, the main translational and clinical achievements in precision oncology have been in the field of genomics. Whereas transcriptomics has been widely used for stratification purposes, the approval of genomic companion diagnostics for targeted drugs has been a major shift in cancer treatment.

To deal with the increasing complexity of cancer diagnosis and treatment, many clinical institutions have established so-called tumor boards. These disease-oriented multidisciplinary teams address the clinical management of patients by integrating findings from various medical specialties. New disciplines like genomics are slowly being introduced in such teams due to the need to perform genome-wide biomarker tests; these teams are often called MTB. MTBs have the potential to perform genomics-driven (genetically informed) medicine, but face common obstacles as genomic data is complex to interpret, there are multiple platforms, tools and workflows to chose from, and there is limited drug access. No major attempts to establish quality requirements, guidelines or tools have been done to date (Velden et al., 2017). Besides, decision making in clinical practice requires stronger evidence than a mechanistic explanation; any genomics-driven treatment has to be better than placebo and that *one-size-fits-all* approach.

The remaining part of this chapter will describe the different aspects of the implementation of genomics-driven oncology. Namely, practical and technical challenges of NGS clinical applications (§2.1) and the conceptual and practical difficulties for the identification of actionable alterations (§2.2).

2.1 Next Generation Sequencing in Clinical Applications

Growing numbers of biomarkers are being discovered and slowly translated into companion diagnostics. However, traditional detection methods are limited to one or few genomic alterations (e.g. *ALK* fusion using FISH; exon 19 mutations in *EGFR* using PCR). NGS opens the door to a cost-effective way to increase the number of genes and alterations tested in one single experiment.

Obtaining and processing patient samples entails important limitations with regard to the quality and quantity of tumor material. Clinical biospecimens are usually fixed as *Formalin-Fixed Paraffin-Embedded* (FFPE) material and stored in biobanks for a long time, which deteriorates the quality of the DNA. Also, most of the resected sample is used for initial diagnosis, reducing the quantity of available biomaterial for NGS. Furthermore, the resected tumor often contains stromal cells and other non-tumor tissue. This yields low tumor content (also known as purity of the sample) which affects subsequent variant calling as the allele frequencies are modified. Other intrinsic aspects of tumors, such as clonality and ploidy, render the process of variant calling more challenging. Alioto et al. (2015) showed that low tumor content limits the maximum number of called mutations and that 100x coverage of both matched normal and tumor samples is necessary to accurately detect clonal and subclonal mutations.

Many clinical institutions have overcome most of these challenges and tumor samples are routinely subjected to NGS in many hospitals. There are three types of NGS technologies for clinical applications: targeted gene panels, WES and WGS. Targeted panels can focus only on frequently mutated *hotspot* positions[1] or sequence entire coding regions of a selected panel of genes[2]. Panels are an affordable and customizable approach that require little input DNA, can detect a combination of genomic events (SNVs, CNVs, rearrange-

(1) *TrueSeq Amplicon Cancer Panel* from Illumina Inc. covers 212 *hotspots*; *Ion AmpliSeq Cancer Hotspot panel v2* from ThermoFisher covers 1974 *hotspots*.

(2) *FoundationOne* and *FoundationOne Heme* from Foundation Medicine cover SNVs, CNVs and fusions in over 300 genes; *Oncoplex* from University of Washington covers SNVs, CNVs and fusions in 234 genes.

ments) and can reach high coverage (to overcome low purity, for example). As a matter of fact, all contributing centers to the GENIE project have decided for panel sequencing, which allowed to collect the largest number of patients to date in this kind of study (The AACR Project GENIE Consortium, 2017). On the downside, panels are restricted to genes known to be implicated in cancer and, therefore, leave no room for new discoveries. In contrast, WES allows the identification of new or rare mutations, which is the reason why large clinical research institutions have decided for this approach (Van Allen et al., 2014; Beltran et al., 2015). Finally, WGS allows a genome-wide interrogation of SNVs, indels, SVs and CNVs. However, increased costs and complexity to interpret the findings make WGS less suitable for clinical applications. All in all, the choice has to be a balance between cost and the desired depth/breadth of sequencing.

As for any other drug or medical device, clinical trials must be carried out to test: i) whether NGS is a feasible approach according to clinical standards (turnaround time, clinical validity), and ii) if NGS is of any benefit for cancer patients (clinical utility). Many world-wide institutions are already trying to answer these questions. For example, NCI-MPACT and MATCH (Lih et al., 2016, 2017), MOSCATO (Massard et al., 2017) or SHIVA (Le Tourneau et al., 2015a) trials have established bioinformatic infrastructures for NGS and used treatment algorithms to stratify the patients into predefined arms (or create new arms when necessary). Other institutions, such as MD Anderson Cancer Center (Tsimberidou et al., 2012), New York Presbyterian Hospital–Weill Cornell Medical College (Beltran et al., 2015), Memorial Sloan Kettering Cancer Center (Zehir et al., 2017) and Heidelberg *National Center for Tumor Diseases* (NCT) (Horak et al., 2017), opted for observational studies. These studies also consist of standardizing techniques and workflows, however, the treatment decision does not take place within the clinical trial; the MTB decides, based on the NGS data generated within the study, whether to enroll the patient into an interventional clinical trial with matched targeted therapy, or to undergo some other treatment strategy. On the one hand, observational studies are more flexible and they leave room for discussion of rare events and *Variants of Unknown Significance* (VUS). On the other hand, no conclusions can be drawn about NGS as a treatment strategy over conventional (non-genomic) strategies.

Overall, these studies have set the infrastructure and standards for sequencing clinical specimens (both FFPE and fresh frozen tissue) with turnaround times from written consent to report generation below 6 weeks, thus demonstrating the feasibility of implementing NGS in clinical routine

(Van Allen et al., 2014; Rennert et al., 2016; Horak et al., 2017). The number of patients with reported actionable variants varies from 40-80% across studies (this variation depends on the definition of actionable, as will be discussed in next section). However, actual numbers of patients treated with genomics-driven therapies were substantially lower (5-35%), in general due to complications such as access to drugs under development, access to clinical trials and clinical deterioration of patients (André et al., 2014; Beltran et al., 2015; Roychowdhury et al., 2011; Sohal et al., 2016; Horak et al., 2017). The implications on patient outcome are yet a controversial aspect. Whereas some trials and meta-analysis have been able to observe improved outcomes using genomics-driven strategies to assign treatment (Schwaederle et al., 2015b; Wheler et al., 2016; Massard et al., 2017), others have failed to show the clinical utility of such an approach (Le Tourneau et al., 2015a; Tannock and Hickman, 2016; Marquart et al., 2018).

2.2 The Challenge of Identifying Actionable Variants

As more and more clinical institutions are technically ready for daily use of NGS, the challenge lies in identifying patient-specific genomic alterations that are relevant for guiding treatment. In the best case scenario, the treating clinician will identify a well-known mutation among all called genomic alterations. However, most mutations are infrequent at a cohort scale and their implications are yet to be discovered. Thus, the variants called in the bioinformatic analysis will be to a large extent VUS.

To date, cancer genomics studies have focused mainly on the identification of driver and biologically relevant events. In a clinical context, though, the focus of cancer genomics shifts from cancer-causing variants towards variants that influence drug response, course and severity of the disease. A genomic variant that informs about treatment action is usually referred to as *actionable*. This definition can be restricted to biomarkers for approved drugs following the indications of cancer type. Yet, it can be broadened to cancer types not included in the drug indication (*off-label* use), or to variants being studied in clinical trials. In this work *actionable* is used in this broad definition. Similarly, *targetable* usually refers to a variant in a gene which is the target of a drug. Again, this definition can refer only to direct targets (to which the drug binds), or include also indirect targets of a drug. The term *druggable* is also used in the literature in a similar fashion. The terminology targetable is reserved for drug targets, which do not necessarily need to inform about drug response – in contrast to a predictive biomarker, which can or cannot be a drug target, but by definition informs about drug response.

Actionability is more and more understood as a dynamic concept defined by several variables: gene, variant, drug and cancer type. To keep the most inclusive definition but differentiate among strength of evidence, tiers, classes or levels are commonly used. Wagle et al. (2012) introduced a 3-tier system to differentiate actionable variants from prognostic and from VUS. Since then, such systems have been adopted by genomics-driven medical community as a simple and clinically relevant classification approach (Sukhai et al., 2016; Van Allen et al., 2014; Dienstmann et al., 2015a; Griffith et al., 2017; Chakravarty et al., 2017; Meric-Bernstam et al., 2015b; Tamborero et al., 2018a; Hintzsche et al., 2018; Horak et al., 2017; Beltran et al., 2015).

Determining the actionability of a variant can be very complex, for targeted therapies have complex mechanisms of action (discussed in §1.3). For instance, the expected relation between GoF mutations in oncogenes and LoF mutations in tumor suppressors is not always fulfilled: in *FGFR2*, N549K is a GoF missense mutation which in endometrial cancer predicts response to FGFR inhibitors (Nakanishi et al., 2014), whereas R251Q in the same gene is a LoF missense mutation and does not predict FGFR inhibitors response in melanoma (Gartside et al., 2009). The ability of a drug to bind to its target can be an important factor to predict the effect of a variant: first-generation ALK inhibitor crizotinib is effective against *ALK* rearrangements but not against acquired resistance mutations (L1196M, C1156Y) or *ALK* amplification. Yet, new generation inhibitors –ceritinib, alectinib, lorlatinib, brigatininb– bind more strongly to the target and are effective also against the latter (Sullivan and Planchard, 2016). So, when one has to interpret a VUS or a known variant in a new context, several questions arise: is this mutation a LoF or GoF? Does it confer sensitivity to any targeted drug? Is there some evidence in any other cancer type? If so, can this evidence be projected to a new cancer type? In the case of multiple actionable alterations, how should they be prioritized? If there are no alterations in actionable genes, can indirect targeting be considered?

2.2.1 Overview of Data and Knowledge Resources

There are many databases, integrative efforts and curated resources that can assist in determining the actionability of a variant. However, sometimes little overlap is found among databases which in principle have the same aim (Ahmed et al., 2011; Griffith et al., 2017). Also, information is irregular and spread across many resources. For instance, some databases focus on drug targets (also known as drug-gene interactions), while others focus on biomarkers (variants associated to clinical outcome). Yet, biomarkers are in many

cases also drug targets. Curated medical guidelines include recommendations for genomic tests, information on drug targets and biomarkers. Finally, lists of biologically relevant cancer genes are often used to provide mechanistic explanations and justify the use of a drug. Table 2.1 summarizes the main resources grouped according to the data type they focus on.

The physician responsible for deciding on a treatment with the support of genomics data faces an overwhelming range of resources to query and decide on (Table 2.1). For that, the genomics medicine community has claimed the need for a comprehensive knowledge database as well as computational algorithms for matching patient's variants to drugs. Last but not least, the information needs to be reported to the treating clinician following standards, providing a minimal set of information, linking statements to evidence source and prioritizing results. Yet, the report has to be a compromise between comprehensiveness and the right level of compression (Johnson et al., 2015; Welch and Kawamoto, 2013; Good et al., 2014; Garraway et al., 2013).

Table 2.1: Data and knowledge resources for tumor genome interpretation. Abbreviations: dgiDB, drug-gene interaction database; TTD, Therapeutic Target Database; PMKB, Precision Medicine Knowledgebase; GDKD, Gene Drug Knowledge Database; JAX-CKB, Jackson Laboratory Clinical Knowledgebase; CanDL, Cancer Driver Log; WHO-ICTRP, World Health Organization-International Clinical Trials Registry Platform; EU-CTR, European Union Clinical Trials Register; DRKS, German Clinical Trials Register; COSMIC, Catalogue of Somatic Mutations in Cancer; DoCM, Database of Curated Mutations; intOgen, interactive Onco Genomics; FDA, Food and Drug Administration; EMA, European Medical Agency; NCCN, National Comprehensive Cancer Network; CAP, College of American Pathologists; AMP, Association of Molecular Pathology; ACMG, American College of Medical Genetics and Genomics; ASCO, American Society of Clinical Oncology ; NIH, National Institutes of Health

Type of data	Description	Databases	Comment
Drug targets	Interactions of drugs or chemical compounds with their target molecules	ChEMBL (Gaulton et al., 2012)	Bioactivity of compounds against drug targets
		Drugbank (Wishart et al., 2006)	Detailed drug-target information
		dgiDB (Griffith et al., 2013)	Drug-gene interactions from the integration of 10 databases
		PharmGKB (PharmGKB, 2018)	Pharmacogenomics (implications of germline variants in drug response)
		TTD (Chen et al., 2002)	Drug targets, disease annotations
		CancerResource (Gohlke et al., 2016)	Over 3000 cancer-related drug targets
		The drug repurposing hub (Corsello et al., 2017)	Drug compounds and their targets
Clinically actionable variants (biomarkers)	Diagnostic, prognostic and predictive biomarkers	CIViC (Griffith et al., 2017)	Knowledgebase of clinical interpretations of somatic/germline variants
		OncoKB (Chakravarty et al., 2017)	Biological effects and treatment implications of somatic variants
		PMKB (Huang et al., 2017)	Free text clinical interpretations of cancer variants (no drug annotation)
		GDKD (Dienstmann et al., 2015a)	Predictive biomarkers for cancer drugs
		Cancer Biomarkers (Tamborero et al., 2018a)	Cancer predictive biomarkers, part of the Cancer Genome Interpreter
		TARGET (Meric-Bernstam et al., 2015b)	Therapeutic implications of cancer genes
		JAX-CKB (Patterson et al., 2016)	Gene/variant annotations, therapy knowledge, diagnostic/prognostic information, and clinical trials related to oncology. No bulk-download
		CanDL (Damodaran et al., 2015)	Cancer driver genes with therapeutic associations (no drug annotation)
		ClinVar (Landrum et al., 2018)	Somatic and germline variations and associated phenotypes (including drug response). Not cancer focused.
Clinical trials	Registries for clinical trials	Clinicaltrials.gov	Database of clinical trials world-wide (though enriched in US trials)
		WHO-ICTRP (WHO-ICTRP, 2018)	World-wide clinical trials registry (integrates 17 regional registries)
		EU-CTR (EU-CTR, 2018)	European clinical trials register
Cancer gene lists	Driver genes, tumor supressors and oncogenes, somatic and germline variants	Cancer Gene Census (Futreal et al., 2004)	Curated list of genes with cancer-causing mutations
		COSMIC (Forbes et al., 2017)	Largest database of somatic mutations
		DoCM (Ainscough et al., 2016)	Database of pathogenic mutations in cancer
		IntOgen (Gonzalez-Perez et al., 2013)	Predicted driver genes
		Catalog of Validated Oncogenic Mutations; Cancer genes	Integrates several databases, homogenizes cancer type taxonomy and variant annotations. Maintained by Cancer Genome Interpreter
Organizations	Drug approval and guidelines organizations	FDA/EMA	Companion diagnostics
		NCCN/CAP/AMP/ACMG	Guidelines for molecular biomarker testing – specific for cancer types
		ASCO	Guidelines and abstracts from the annual ASCO meeting
Comprehensive resources	Curated educational resources	`https://www.mycancergenome.org/`	GenomOncology
		`https://www.cancer.gov/`	National Cancer Institute (NIH)

2.3 Aims and Organization of the Thesis

Interpretation of genomic data is a cumbersome task for a clinician. In order to fulfill the needs of the medical community, genomics expertise in MTBs is inevitably required. Computational algorithms and visualization tools are crucial to support analysis and interpretation of genomic data. The overall goal of this thesis is to pave the way for the use of genomic technologies in clinical decisions. More precisely, this is a four-fold aim:

Propose a workflow for genomics-driven clinical decisions — Fulfill the interpretation and reporting gap of genomics-driven oncology: assemble existing public knowledge into a workflow that filters genomic variants of a patient and reports the actionable variants. The main concept is depicted in Figure 2.1.

Implement the workflow for a broad audience — Implement the workflow as a web tool with the R Shiny framework to contribute to the standardization of reporting genomic findings. The i(nteractive)MTB-Report app is available as a web page at `http://www.ams.med.uni-goettingen.de:3838/iMTB-Report/app/` and distributed in GitHub (`https://github.com/jperera-bel/iMTB-Report`).

Evaluate the scope of the workflow on large public datasets. — Assess the feasibility and scope of the workflow by applying it onto two datasets: TCGA and the AACR's project GENIE.

Provide a proof-of-concept of the workflow — Demonstrate the workflow's strengths in terms of clinical utility by analyzing patients sequenced within the NCT *Molecularly Aided Stratification for Tumor Eradication Research* (MASTER) program.

This thesis is organized as follows: *Materials and Methods* part is divided into two chapters. Chapter 3 describes specific concepts for knowledge organization of clinically actionable variants and the selection of databases used in this thesis. Chapter 4 details the datasets analyzed within this thesis and the tools used. The following part, *Results*, starts with the description of the workflow that was developed, the MTB Report workflow, and its implementation as an R Shiny application in Chapter 5. Next, the results of applying our workflow to large public genomic datasets are presented in Chapter 6. Chapter 7 consists of a retrospective analysis to provide a proof-of-concept of the clinical utility of the workflow.

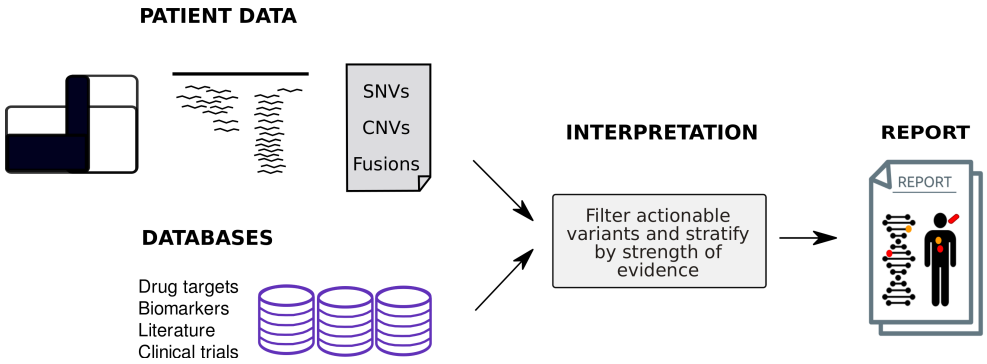


FIGURE 2.1: **Concept of computationally assisted genomic-data interpretation**. Assemble public knowledge into a workflow that filters patient's genomic variants and reports actionable variants.

Finally, *Discussion and Conclusions* part includes a discussion in chapter 8 in which the main additions of this thesis to the genomic-driven oncology toolbox are discussed and put into context with new developments. The thesis is wrapped up in the conclusions in Chapter 9.

PART II

MATERIALS AND METHODS

Chapter 3

Knowledge of Actionable Variants

To introduce the knowledge required for understanding the findings of this thesis, this chapter describes the concepts for the organization of knowledge of clinically actionable variants (§3.1). Next, the databases used in the workflow presented in results are described in detail (§3.2).

3.1 Organization of Actionable Variants Knowledge

In order to suggest drugs to patients based on their genomic profile (i.e. interpret a tumor's genome), the term actionable variant has to be defined. In this work, the term actionable variant is used as a synonym for predictive biomarker. The actionability considered ranges from approved biomarkers (i.e. companion diagnostics) to investigational biomarkers (i.e. assessed in preclinical studies). Here, *off-label* cases are taken into account (e.g. cancer type indication, VUS). More details will be given in §5.4.

Along with the development of genomics-driven medicine, there has been a parallel attempt to set standards on how to organize the knowledge required to enable this new type of medicine. A catalog of actionable variants must comprise, at least, four layers of information: i) gene, ii) genomic variant, iii) drug, and iv) cancer type. Good et al. (2014) and Dienstmann et al. (2014) suggested a minimum of data layers needed for curation of actionable variants:

- **Gene annotation**. The gene that has an actionable genomic mutation. The recommended nomenclature for gene annotation is *HUGO Gene Nomenclature Committee* (HGNC) gene symbols, as it is human readable and it is related to biological concepts (e.g. *BRAF*). Yet, other gene nomenclatures can be used in addition (Entrez *–673*, Ensembl *–ENSG00000157764*).

- **Variant annotation**. The genomic alteration that is considered actionable. A list of possible alterations types can be found in Table 1.1. To ensure good practice and reproducibility, *Human Genome Variation Society* (HGVS) recommendations and standards for variant annotations should be followed (`https://varnomen.hgvs.org/`). In principle, genomic-level annotation should be used (e.g. chr12:g.25398284C>T) along with coding-annotation (e.g. c.35G>A) to avoid transcript conflicts (e.g. KRAS:NM_004985:exon2:c.G35A:p.G12D vs. KRAS:NM_033360:exon2:c.G35A:p.G12D). Yet, in publications and medical community, protein-level annotations are widely used (e.g. p.G12D) and, as such, should also be specified.

- **Tumor type annotation**. Cancer subtype (histological subtype, site subtype, or even molecular subtype) in which the variant is considered actionable. Ideally, disease ontologies should be used: Disease Ontology (`http://www.disease-ontology.org/`), ICD-O (`http://codes.iarc.fr/`), MeSH terms (`https://meshb.nlm.nih.gov/`), OncoTree ontology (`http://oncotree.mskcc.org/`).

- **Drug annotation**. Treatment for which the genomic variant indicates response or resistance. Besides generic names, drugs should be annotated using identifiers. Also, providing the type of inhibitor is useful when the drug is in preclinical settings (e.g. BTK inhibitors). Remarkably, drug annotation is the field in which the lower level of standardization has been accomplished to date.

- **Effect or direction of association**. The direction of the predictive association between the variant and the drug (e.g. response, resistance, no response). In other words, whether the annotated variant or the lack of the variant is the predictor (e.g. *EGFR* mutations in lung cancer vs. *KRAS* wild-type in colorectal cancer).

- **Type of association**. Prognostic, diagnostic or predictive association. In this thesis, only predictive associations are considered as actionable variants. However, databases curating variants for clinical interpretation can include other associations between a variant and clinical outcome (e.g.high risk of relapse, better outcome, molecular subtype). Eventually, biologically relevant variants can also be included in these kind of knowledge bases.

- **Level of evidence**. Classification system to rate the clinical utility of the association, or, in other words, the level of actionability. Common systems differentiate between consensus (i.e. standard) and emerging

(i.e. investigational) evidence. The former includes biomarkers that are recognized by either FDA and EMA or recommended by the *National Comprehensive Cancer Network* (NCCN) or similar expert-panels. The latter refers to biomarkers for which compelling clinical evidence exists (clinical trials, case reports) but the predictive association between the variant and the drug has not been approved in any scenario (here, preclinical evidence can also be considered). Yet, *off-label* scenarios fall within a gray-zone.

3.2 Databases of Actionable Variants

As depicted in Table 2.1, several databases collect information on actionable variants. Yet, some of these databases do not have enough annotation details for clinical implementation and follow different curation and accessibility methods. Within the scope of this thesis, the following selection criteria were used: the depth of annotation (layers explained in the previous section), bulk-download option, up-to-date, clinically focused, cancer focused and somatic variants focused. As a result, two databases and two lists of actionable genes were used within the scope of this thesis: *Clinical Interpretation of Variants in Cancer* (CIViC) (Griffith et al., 2017), *Gene Drug Knowledge Database* (GDKD) (Dienstmann et al., 2015a), *Tumor Alterations Relevant for Genomics-driven Therapy* (TARGET) (Van Allen et al., 2014) and gene list from Meric-Bernstam et al. (2015b)[1].

3.2.1 Clinical Interpretation of Variants in Cancer (CIViC)

The CIViC database is a community-driven effort born from the proposed knowledge base system by Good et al. (2014). Both open access and open source, it is built upon a collaborative process in which anyone can be a curator. Yet, user-roles and limited powers ensure the reliability and revision of all interpretations. The scope of the database includes all types of genomic biomarkers in cancer: prognostic, diagnostic, predisposing and predictive. It also contains both germline and somatic variants. It is the largest database compiling curated information of this kind. The database acknowledges a bias towards *Acute Myeloid Leukemia* (AML), breast and lung cancer, since Washington University is especially focused on clinical research in these entities.

(1) Databases such as Cancer Biomarkers (Tamborero et al., 2018a) and OncoKB (Chakravarty et al., 2017) would fulfill this thesis' criteria, however, they are not included as they were in their first stages at the time the work for this thesis was performed.

A detailed example of the structure of this database is shown in Table 3.2. Moreover, comprehensive annotation for gene, variant, disease and sequence are provided by an automated import of data from other databases and ontologies. Level of evidence annotation consists of 5 tiers: A (validated), B (clinical), C (case study), D (preclinical) and E (inferential). Also, a free-text summary is stated for each association.

For the analyses of TCGA, GENIE and MASTER presented in Chapter 6 and Chapter 7, CIViC version from 1st of June 2017 was used. In this version, the database contained 1931 variant-drug evidences on 290 genes (of which 213 had predictive associations with a drug) across 177 cancer types.

3.2.2 Gene Drug Knowledge Database (GDKD)

Similarly, GDKD was built upon the concepts proposed in Dienstmann et al. (2014). As shown in Table 3.3, annotation levels and structure are very similar to CIViC. However, this database follows an expert-only curation model and data bulks are made available through the Synapse repository periodically (`https://www.synapse.org/#!Synapse:syn2370773`). It focuses exclusively on somatic variants which predict response to anti-cancer drugs (genomic predictive biomarkers). Level of evidence is annotated in a similar fashion as CIViC, thus, a simple relation can be drawn between both databases (Table 3.1). In contrast to CIViC, which has plentiful of actionable variants with evidence supported by preclinical studies, GDKD has a more filtered list of preclinical actionable variants based on their scientific soundness and translational power.

For the analyses of TCGA, GENIE and MASTER presented in Chapter 6 and Chapter 7, version 19.0 of GDKD was used. This version contained 618 variant-drug evidences on 170 genes across 65 cancer types.

3.2.3 Tumor Alterations Relevant for Genomics-driven Therapy (TARGET) and Meric-Bernstam et al. (2015b)

TARGET was published in 2014 as part of a WES clinical pipeline and three versions have been released. The latest version (TARGET_db_v3_02142015.xlsx) available at `http://software.broadinstitute.org/cancer/cga/target` consists of a list of 135 genes manually curated by experts from the Dana-Farber Cancer Institute with predictive, prognostic and diagnostic implications in cancer. Although the list has few annotation layers (see Table 3.4), it comprises several genes not included in CIViC nor GDKD.

TABLE 3.1: **Relation of levels of evidence between CIViC and GDKD.**

CIViC	**GDKD**	
Level of evidence	**Evidence**	**Status**
A (validated)	Consensus	FDA-approved
		NCCN-guidelines
B (clinical)	Emerging	Late trials
		Early trials
C (case study)		Case report
D (preclinical)		Preclinical
E (inferential)		-

Meric-Bernstam et al. (2015b) published a list of therapeutically actionable genes (i.e. predictive biomarkers) with a focus on genes included as selection criteria in clinical trials.

Both lists were published with the aim to engage the community into curation of clinical implications of genomic alterations before CIViC and GDKD were released. Although there is a high overlap between them, TARGET spans a broader range of clinical implications (e.g. prognostic, diagnostic) ,whereas Meric-Bernstam et al. (2015b) focuses only on actionable variants as defined in this thesis.

TABLE 3.2: **CIViC database structure.** Five selected variants are reproduced, including diagnostic, prognostic and predictive associations.

Gene	Variant	Disease	Drugs	Evidence type	Evidence direction	Evidence level	Clinical significance	Pubmed id
NRAS	Q61	Melanoma		Diagnostic	Supports	B	Positive	23861977
MAP2K1	Q56P	Melanoma	Selumetinib (AZD6244)	Predictive	Supports	D	Resistance or Non-Response	19915144
DNMT3A	R882	Acute Myeloid Leukemia		Prognostic	Supports	B	Poor Outcome	21067377
ERBB2	amplification	Gastric Adenocarcinoma	Trastuzumab	Predictive	Supports	A	Sensitivity	20728210
ARAF	S214C	Non-small Cell Lung Carcinoma	Sorafenib	Predictive	Supports	C	Sensitivity	24569458

TABLE 3.3: **GDKD database structure.** Four selected actionable variants are reproduced

Disease	Gene	Variant	Description	Effect	Association	Therapeutic context	Status	Evidence	PMID
breast	BRCA2	any	mutation	loss-of-function	response	PARP inhibitors	early trials	emerging	20609467
gastric	ERBB2	amplification	copy number gain	gain-of-function	response	trastuzumab	FDA-approved	consensus	FDA
colorectal	BRAF	V600	missense mutation	gain-of-function	resistance	cetuximab, panitumumab	late trials	emerging	20619739
lung adeno	ALK	amplification	copy number gain	gain-of-function	resistance	crizotinib	case report	emerging	22277784

TABLE 3.4: **TARGET database structure.** Four selected genes are reproduced: the first three have predictive associations to drugs, the last one is a diagnostic biomarker.

Gene	Rationale	Types of recurrent alterations	Examples of Therapeutic Agents
ABL1	Translocations predict sensitivity to tyrosine kinase inhibitors such as imatinib, dasatinib, and nilotinib. Secondary mutations can cause resistance to these agents.	Rearrangement; Mutation	Imatinib, Dasatinib, Nilotinib, ABL1 inhibitors
AKT2	Mutations may predict sensitivity to AKT/MTOR inhibitors	Mutation; Amplification	AKT/MTOR inhibitors
ALK	Translocations predict sensitivity to ALK-inhibitors such as crizotinib. Secondary mutations can cause resistance. Amplification and activating mutations may also be sensitive to these agents.	Rearrangement; Mutation; Amplification	Crizotinib, ALK inhibitors
CDH1	Diagnostic in lobular breast carcinoma. In gastric cancer, may signal the presence of a germline mutation.	Mutation	

Chapter 4

Data and Resources

This chapter summarizes the datasets, resources and tools used within this thesis. §4.1 details the two public patient datasets used to evaluate the scope and feasibility of the workflow (results presented in Chapter 6). §4.2 describes the *Molecularly Aided Stratification for Tumor Eradication Research* (MASTER) dataset used as a proof-of-concept for the clinical utility of the described workflow (results presented in Chapter 7). Finally, section §4.3 details the software and packages used for the implementation of workflow.

4.1 Public Datasets

Two large multi-center datasets of cancer patient samples profiled with high-throughput genomic techniques were analyzed in this thesis: *The Cancer Genome Atlas* (Weinstein et al., 2013) and *Genomics Evidence Neoplasia Information Exchange* (The AACR Project GENIE Consortium, 2017).

4.1.1 The Cancer Genome Atlas (TCGA)

The Pan-Cancer 12 dataset comprises the first 12 cancer types profiled by TCGA. It was chosen for being the most well-established TCGA dataset, which has been studied in many publications and for which a data *freeze* is provided (Weinstein et al., 2013). Synapse repository assembles high-level analyses into a robust and consistent data *freeze* (latest version V4.7). For this work, ready-to-use data was downloaded from the Synapse repository. See Table 4.1 for a dataset description. Next, a short description of the provenance of the data is provided.

- **SNVs**. Exome-sequencing was performed in different institutions, and, thus, exome-capture and sequencing platforms of tumor and matched

TABLE 4.1: **TCGA Pan-Cancer 12 dataset.**

Acronyms	Tumor Type	SNV	CNV	Clinical	Overlap
BLCA	Bladder cancer	99	126	153	97
BRCA	Breast carcinoma	772	887	929	756
COAD/READ	Colorectal cancer	224	586	592	224
GBM	Glioblastoma	291	578	598	287
HNSC	Head and Neck Cancer	306	310	343	306
KIRC	Kidney cancer	417	457	459	417
LAML	Acute myeloid leukemia	200	198	202	190
LUAD	Lung adenocarcinoma	230	357	508	172
LUSC	Lung squamous carcinoma	178	345	389	178
OV	Ovarian cancer	316	577	592	313
UCEC	Uterine corpus endometrial cancer	248	511	512	244
Total		3281	4932	5277	3184

normal samples differed among centers. For that, a standardization process was performed by the analysis working groups of the TCGA Research Network: recurrent false positives (blacklist) were removed, germline variants and single nucleotide polymorphisms (allele frequency > 0.1 in population studies) were also removed variants present in dbSNP database (Sherry et al., 2001) were removed, and all variant coordinates were transferred to GRCh37 and re-annotated using the Gencode human transcript annotation imported from Ensembl release 69. The downloaded data *freeze* is a tab-delimited *Mutation Annotation Format* (MAF) file[1] which contains strict filters to ensure high quality mutation calls.: e.g. recurrent false positives (blacklist) were removed, germline variants and single nucleotide polymorphisms (allele frequency > 0.1 in population studies) were also removed. Data was downloaded at this stage from `https://www.synapse.org/Synapse:syn1729383`. For a more detailed explanation, see Kandoth et al. (2013) and Synapse repository.

- **CNVs**. Affymetrix SNP Array 6.0 were used to measure CNVs. GISTIC (Mermel et al., 2011) was used to identify recurrent regions with CNVs (noise threshold of 0.3, a broad length cutoff of 0.5 chromosome arms, a confidence level of 95% and a copy-ratio cap of 1.5). The downloaded file contains gene-wise CNV calls in which two thresholds are applied: i) CNVs that passed the noise thresholds are given a value of +1 (amplification) or -1 (deletion); and ii) focal high-level CNVs with >4.4 copies are given +2 (focal amplifications) and with <1 copies are

(1) description here: `https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/`

assigned -2 (focal deletions). Data was downloaded at this stage from https://www.synapse.org/Synapse:syn1711454. For a detailed information on the processing of CNV data, see Zack et al. (2013); The Cancer Genome Atlas Network (2012).

- **Clinical**. Clinical data was downloaded from https://www.synapse.org/Synapse:syn2325436. The annotated metadata may differ between the different cancer types.

4.1.2 Genomics Evidence Neoplasia Information Exchange (GENIE)

GENIE's main goal is to set standards for multi-center data aggregation in a clinical routine basis. As a result, a harmonized dataset from eight contributing clinical centers has compiled, by the time of this thesis, over 18.000 patient samples from 32 tissues. The dataset is enriched in late-stage samples. Samples were profiled using different DNA targeted panels (covering from 50 to over 400 genes), which in some cases included also CNV and SV profiling (Table 4.2). For detailed information on the methods, see GENIE data guide[2].

Data was also downloaded from synapse repository. In this case, clinical data was downloaded from syn7851246, SNVs from syn7851250, CNVs from syn7851245 and fusion data from syn7851249. Both SNVs and CNVs file formats were the same as for TCGA (see §4.1.1). Fusions are detailed in a tab-delimited file containing the gene fusion in the format "*BCR-ABL1* fusion" or "*AKT2* fusion" (HGNC gene symbols).

4.2 The Molecularly Aided Stratification for Tumor Eradication Research (MASTER) Dataset

Molecularly Aided Stratification for Tumor Eradication Research (MASTER) is a clinical sequencing program within the *National Center for Tumor Diseases* (NCT) with an institutional review board-approved protocol in Heidelberg. In this program, a sequencing platform has been developed in order to perform prospective stratification of advanced cancer patients in clinical context (Horak et al., 2017). Somatic variants (SNV, CNVs and SVs) derived from WES of tumor and normal samples of 11 patients from this program were used within this thesis. Next, the bioinformatics analysis performed within the MASTER protocol are detailed.

(2) https://www.aacr.org/Research/Research/Documents/GENIE%20Data%20Guide.pdf

TABLE 4.2: **GENIE dataset by center.** # stands for number; X means that data is available for the patients from that center. Fusions are only available for a subset of patients: *1911; **409. Column *Samples* refers to the strategy to call somatic variants: T stands for tumor-only, T-N stands for matched tumor-normal samples.

Center	Center Name	#Patients	SNVs	CNVs	Fusions	Samples
DFCI	Dana-Farber Cancer Institute, USA	6137	X	X		T
GRCC	Institut Gustave Roussy, France	529	X			T
JHU	Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, USA	1203	X			T
MDA	MD Anderson Cancer Center, USA	961	X			T
MSK	Memorial Sloan Kettering Cancer Center, USA	7341	X	X	X*	T-N
NKI	Netherlands Cancer Institute, The Netherlands	505	X			T
UHN	Princess Margaret Cancer Centre, University Health Network, Canada	1296	X			T-N
VICC	Vanderbilt-Ingram Cancer Center, USA	832	X	X	X**	T

TABLE 4.3: **MASTER dataset.**

MASTER ID	Tumor	Gender	Tumor content	Coverage tumor	Coverage normal
01	Breast cancer metastasis	Male	35%	138x	114x
02	Pancreatic adenocarcinoma	Male	90%	128x	108x
03	Leiomyosarcoma of the retroperitoneum	Female		155x	113x
04	Ovarian carcinoma	Female	20%	109x	131x
05	Myxoid liposarcoma	Female	100%	112x	125x
06	Neuroendocrine tumor	Male		154x	143x
07	Neuroendocrine tumor	Male	70%	150x	160x
08	Cholangiocarcinoma	Male	60%	159x	123x
09	Clear cell sarcoma	Female		135x	152x
10	Histiocytic sarcoma	Male	70%	98x	96x
11	Pulmonary Adenocarcinoma	Female	80%	131x	121x

Average coverage for tumor and normal samples was 133X and 126X, respectively (details of each sample are listed in Table 4.3). Read alignment was performed using BWA (version 0.6.2, Li and Durbin (2009)) against reference human genome NCBI build 37.1. Default parameters and maximum insert size set of 1000 bp were used. Next, SAMtools (version 0.1.19) was used for sorting BAM files and Picard tools (version 1.90) for marking duplicates (Li et al., 2009).

SNV calling was done using standard mpileup and bcftools (SAMtools). Next, heuristic rules were applied to filter the variants, as previously done in Jones et al. (2012, 2013); Yaktapour et al. (2014). Namely, a minimum tumor allele frequency (10%), a minimum number of tumor reads at the position (5), a minimum number of normal reads at the position (12) and a minimum tumor content (20%). Indels were called with Platypus (version 0.5.2, Rimmer et al. (2014)). Functional annotation was done with RefSeq model in ANNOVAR (version of September 2013, Wang et al. (2011)). Finally, non-silent coding variants (nonsynonymous, stop-gain, stop-loss and indels) were selected.

Read-depth plots and an in-house pipeline which uses VarScan2 copy number and copyCaller modules were used for CNV calling (Koboldt et al., 2012). Regions with unmappable genomic stretches were filtered and merged. Regions were annotated with RefSeq genes using BEDTools (Quinlan and Hall, 2010). Regions with a tumor/normal coverage log ratio > 0.55 or < -0.55 were called as amplifications and deletions, respectively. Finally, SVs that can lead to gene fusions – inversions, deletions, duplications– were called with CREST at DNA level (Wang et al., 2011).

The high-confidence list of somatic variants was published as part of the supplementary material in Perera-Bel et al. (2018).

4.3 Tools and Implementation

The workflow presented in this thesis was implemented using the scripting language and environment for statistical computing R (Team, 2014). Core functions were written to match input genomic data to databases of actionable variants, to classify the filtered matched data into levels of evidence, and to generate an output report with the results (method detailed in §5). For the output report, a .tex file is automatically generated from a .Rnw file using *knitr* and *xtbale* R packages, which is then converted to a .pdf using LaTeX (`texi2pdf` R function). The core functions were published as Additional File 6 in Perera-Bel et al. (2018) and is being maintained on the online version control website GitHub[3].

The implementation as a web-based tool of the workflow was done using the *Shiny* package (Chang et al., 2018b), a framework for building interactive web applications using R. A user interface was build around the aforemen-

(3) Repository on GitHub: `https://github.com/jperera-bel/MTB-Report`

tioned core functions of the workflow. The user interface uses the *shinydashboard* and *DT* packages. The user provides the required inputs (SNVs, CNVs, gene fusions and cancer type). Alternatively, the user can explore data from the TCGA project. This feature uses the R client for Broad Institute's Firehose web API *FirebrowseR* (Deng et al., 2017) to retrieve genomic and clinical data of the selected TCGA sample. On the server side, these inputs (user defined or from TCGA) are reactive values connected to a main reactive expression that calls the core functions to match, filter and classify actionable variants. Several endpoints are connected to the reactive expression: a figure with a visual summary of the number of findings per level identified; a table with the actionable variants matched to database information; and file download. The table can be filtered by selecting one or several levels of evidence, sorted according to columns or searched for specific patters. The download option offers to save the table as a .csv file, or to generate the full .pdf report with LaTeX.

The *Shiny* based web application is available as open source under MIT license on GitHub[4] and hosted for use at the University web page[5]. As of September 2018, version v0.1.1 has been released. By default, it supports GDKD's version v20.0 and CIViC's version from 01 May 2018, but other versions can be uploaded by the user.

(4) Repository on GitHub: `https://github.com/jperera-bel/iMTB-Report`
(5) University web page: `http://www.ams.med.uni-goettingen.de:3838/iMTB-Report/app`

PART III

RESULTS

Chapter 5

Molecular Tumor Board Report Workflow

The *Results* part follows the same order as the aims (§2.3). Hence, in the first place, Chapter 5 presents a workflow to enable genomics-driven oncology and describes its implementation as a web application. Next chapters will present the results of applying the workflow on two public datasets (Chapter 6) and the results of a comparison analysis using patient data from a German precision medicine initiative, the MASTER program (Chapter 7).

The workflow described in this chapter is referred to as the *Molecular Tumor Board* (MTB) Report. Briefly, the MTB Report takes as input all somatic variants of a sample and finds, with the use of public databases, which variants have predictive evidence on drug response (i.e. actionable variants). The output collects a filtered list of actionable variants and provides a detailed information on the clinical evidence with the goal to allow a genomics-guided, evidence-based clinical discussion.

The first section (§5.1) documents the modifications done to the databases upon which the workflow relies. Next, the different steps of the workflow (input data, filtering of variants, classification of variants and output report) are detailed in four separate sections (§5.2, §5.3, §5.4 and §5.5). Finally, the implementation as a web application is detailed (§5.6). A research article describing the workflow has been published in Genome Medicine (Perera-Bel et al., 2018). Furthermore, the source code is available in the GitHub repository (`https://github.com/jperera-bel/MTB-Report`).

5.1 Parsing of Databases

Curated databases are an essential part of the proposed workflow, as the output fully relies on them. Some modifications to the databases described in methods (§3.2) are required to ensure a standardization and a correct

matching in later steps.

- **GDKD**. This database contains redundant rows in cases in which multiple variants of one gene have the same clinical implications. To reduce both search time and output length, variants in the same gene are merged (comma separated) if they share annotations at the disease, drug, evidence and association levels. Further minor modifications to ensure a correct functioning of the workflow consist of the removal of i) blank spaces in gene names (manual *errors* introduced in the curation process) and ii) special characters (^ and _) in columns containing PubMed and abstract IDs ("PMID") (these characters can generate errors in the steps that use *LaTeX*).

- **CIViC**. This database contains three types of clinical associations a genomic variant can be annotated with (e.g. prognostic, diagnostic and predictive). For the purpose of this thesis, only rows containing "Predictive" evidence (column "evidence_type") are selected. Finally, for the same reason as in GDKD, variants in the same gene are merged (comma separated) if they share annotations at the disease, drug, evidence and association levels.

- **TARGET**. This database, as CIViC, also contains various types af associations. Hence, rows containing "Predictive" evidence (those with annotations in "Drug" column) are selected. Taking advantage of the similar structure in terms of annotation levels (layers) of this database compared to the supplementary table 2 from Meric-Bernstam et al. (2015b), and because neither of them are regularly updated, 20 gene annotations from the latter are manually added to the TARGET database.

The databases are stored as CSV files and are used later on in the filtering step (§5.3). The full list of genes covered by each of the databases is shown in Table A.4.

5.2 Genomic Data and Cancer Type handling

The two inputs required by the workflow are genomic data and cancer type. However, as will be presented in §5.5, other clinically relevant data might be provided by the user and will be displayed in the final report. For instance, an ID of the patient, previous therapies received and sample information (biospecimen type, sequencing technology, tumor content).

Genomic data refers to fully-annotated and quality-filtered somatic variants. This is an important remark, as it is assumed that only high-quality variants are provided and, thus, no quality filtering is performed. Nevertheless, quality information (e.g. allele frequencies, position depth, Phred scores, functional predictions) can be provided and will be displayed, but merely for informative purposes. Three kinds of somatic variants can be used as input: SNVs (and indels), CNVs, and fusion genes. SNVs must be a table with at least three columns, containing i) HGNC gene symbols, ii) variant type (MAF format, see Table A.3) and iii) aminoacid change (e.g. V600E). MAF format can easily be adapted from *Variant Call Format* (VCF) files[1]. CNVs must be a table with at least two columns: the first column contains HGNC gene symbols, the second column specifies the type of copy number alteration (amplification or deletion). Gene fusions must be provided in a table with two columns, each containing one of the two fused genes using HGNC gene symbols.

Cancer type refers to the cancer diagnosis and has to be provided by the user. A cancer type is generally defined by the organ or tissue in which the cancer originated and is often combined with the type of cell that formed the tumor (e.g. lung adenocarcinoma). Cancer type information is used by the MTB Report workflow to classify actionable variants into levels of evidence (detailed in §5.4). Since the databases use different cancer type annotations, a table with equivalences between the different databases was manually built. TCGA and GENIE cancer type annotations were also included in this mapping file, and consensus cancer types were created (to be selected by the user as input and to be displayed in the output report). The table was constructed in the way that entries were duplicated if they could map to more than one disease type to account for hierarchical relations between subtypes. For instance, the consensus *lung (adeno)* maps to TCGA's type LUAD, GDKD's *lung adeno* and *lung*, CIViC's *Bronchogenic Lung Adenocarcinoma*, *Lung Acinar Adenocarcinoma*, *Lung Adenocarcinoma*, *Lung Carcinoma*, *Lung Cancer*, *Non-small Cell Lung Carcinoma* and to GENIE's NSCLC. Likewise, *lung* consensus maps to all the above plus TCGA's type LUSC, GDKD's *lung squamous*, CIViC's *Lung Squamous Cell Carcinoma* and GENIE's *Small Cell Lung Cancer*. The complete table can be found as supplementary material in Perera-Bel et al. (2018) or in GitHub (`https://github.com/jperera-bel/MTB-Report/blob/master/data/cancer_types.csv`).

This approach that requires manual curation was developed since only

(1) VCF to MAF: `https://github.com/mskcc/vcf2maf`

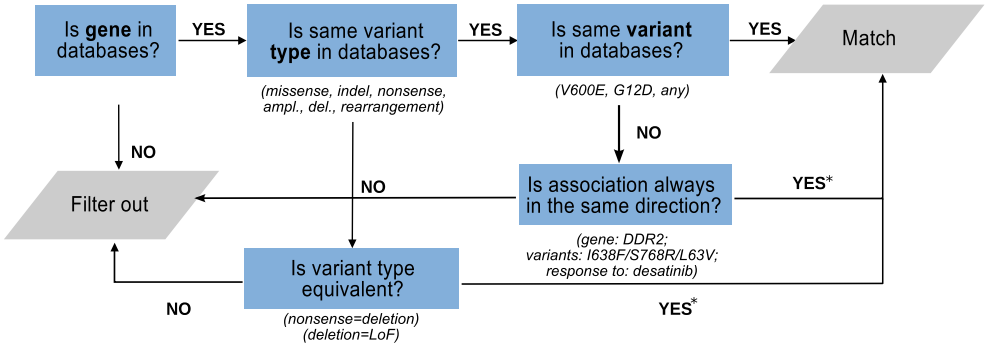


FIGURE 5.1: **Flowchart of steps to filter actionable variants.** The flowchart depicts the rules applied to genomic variants in order to determine their actionability. Asterisks depict variants marked with a *repurposing* flag.

CIViC provided disease-ontology annotations among the databases used. Ontologies are valuable as long as they can be used to match different resources to each other. As the field evolves, new releases and new databases are incorporating the use of ontologies. Thus, further releases of the MTB Report workflow will incorporated disease ontologies.

5.3 Filtering of Actionable Variants

The filtering consists of matching somatic variants from the patient. The abstraction of the rules applied to filter actionable variants are outlined in Figure 5.1. For practical reasons, the steps shown in the figure are implemented in a database- and variant type-basis.

To understand the filtering procedure, it is important to mention the two manners in which a SNV can be annotated in a database (see Table 3.3). The first manner is for very well described variants (*hotspots*): the specific aminoacid changes are given (e.g. *BRAF* V600). The second manner is common in clinical and preclinical studies that study the predictive value of all variants in a given gene: these cases are annotated in the databases as *any* or LoF/GoF SNVs. In order to handle both annotations, the algorithm narrows down from gene level to variant level and filters out SNVs at every step.

The general workflow depicted in Figure 5.1 consists of three main filtering steps and two side rules to recover (i.e. *repurpose*) variants. In the first place, filtering is done at gene level: genes not present in the databases are filtered out. Second, variant types are matched between the patient and the database (i.e. missense mutations, indels, nonsense mutations, amplifications,

deletions and rearrangements). Nonsense mutations, in which a stop codon is introduced in the middle of the coding sequence causing LoF of the protein product, are matched to deletions or LoF with a *repurposing* flag. Third, SNVs from the patient are matched to database annotations. If the same aminoacid change from the patient is annotated in the database (first manner), the predictive evidence in the database is matched to the patient's variant. If the gene in which the considered SNV is annotated as *any* (second manner), the patient's variant (regardless the aminoacid change) is automatically matched to that predictive evidence. Finally, if the patient variant is not annotated in the database, but other aminoacid changes are, the variant will be matched with a *repurposing* flag if the entries in the database of that gene support always the same effect (i.e. response vs. resistance) towards the drug.

Repurposing flags are used to recover VUS that are likely to be actionable. Flagged variants are treated as normal matches for downstream classification (§5.4), but the flag will be shown in the final report to acknowledge its VUS nature.

In case a gene has two different mutations, or several annotations of the same variant are provided (e.g. different transcripts), the algorithm checks all of them.

5.4 Classification into Levels of Evidence

Patient's variants matched by the filtering algorithm to database entries are considered to be actionable. A variant is actionable when it informs about treatment action; in other words, the variant is a biomarker of a given drug. For clarification, an actionable variant can be matched to multiple database entries (e.g. if the variant is actionable in distinct tumor types, towards distinct drugs, proven at distinct clinical or preclinical studies). Each predictive association (i.e. each finding of the workflow) is unique for a given variant, with a given drug in a given context.

Actionability of a variant can be supported by different levels of evidence. For that, it is common to use classification schemes or tiers, in order to inform about the strength of the evidence. Four main aspects are important to determine the actionability of a variant: the affected gene, the genomic variant, the cancer type and the drug. The first two are accounted for during the filtering step (see previous section and Figure 5.1). Hence, at this stage of the workflow, the last two remain to be accounted for.

In this work, a six-level system is proposed, defined by two variables: can-

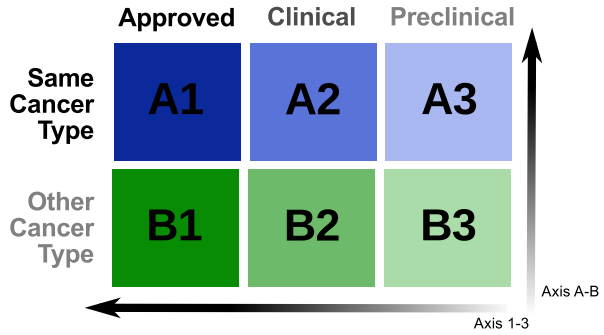


FIGURE 5.2: **Levels of evidence.** Six-levels system that classifies actionable variants along two axes: strength of clinical evidence (axis 1-3) and cancer type (axis A-B). On the 1-3 axis, 1 stands for biomarkers approved by e.g. FDA, EMA, NCCN. Level 2 regards clinical trials and case reports. Finally, level 3 consists of preclinical evidence including animal models and cell lines.

cer type and strength of clinical evidence. Cancer type is dichotomized into the *same* cancer type –as the patient in question– and *other* cancer types. The strength of clinical evidence refers to the validation stage of the drug–biomarker predictive association and is divided into three categories: *approved*, *clinical* and *preclinical*. Hence, the combination of these two variables yields six levels of evidence (Figure 5.2).

This classification scheme is flexible in terms of repurposing between cancer types and VUS. However, it does not account for biologically relevant variants (unless they are actionable) nor variants in drug targets that are not known to be predictive biomarkers. These aspects are important when compared to other classification systems, as classification systems are to a large extent interchangeable but each of them has a specific focus. As for our six-level system, the focus is on the validation stage of the biomarker-drug predictive association and the *repurposing* between cancer types. The reason behind it is the fact that genomic events are present across cancer types and most of the genes recurrently altered in cancer are already being studied in at least one context. Thus, by making these cross-cancer cross-clinical evidence knowledge available, the patient's actionable landscape can be expanded.

5.5 Design of the Molecular Tumor Board Report

The format of the output is one of the main features of this workflow. It is named after MTBs, as it is designed having in mind such a medical setting (though, as stated in the disclaimer, the report is intended for research use only and should not be used for medical or professional advice). Hence, the report contains some technical terminology that happens to be appropriate for

clinicians, geneticists, oncologists, pathologists and scientists, but that could not be used in direct-to-consumer products. A report that targets clinicians, who are often short in time, has to be concise. Yet, conciseness should not be at the expense of complexity and completeness. The MTB Report uses a tabular format, as tables allow a neat, structured and visual presentation of results. Along these lines, the design of the report follows recommendations, standards and guidelines on reporting genomic variants in clinical settings (Richards et al., 2008, 2015; Matthijs et al., 2016; Li et al., 2017).

The MTB Report is structured into two main blocks (a sample report is shown in Figure 5.3). Under "Patient information" the user can find anonymous patient details, clinical and specimen information –given by the user– along with a summary of the genomic data provided as input. The second block, entitled "Gene-drug predictive associations" details the variants found by the algorithm as actionable. A summary of the variants identified as actionable and their quality (if provided by the user) is shown in genomic type specific tables (that is, one for SNVs, one for CNVs and one for fusion genes). The filtering method, the databases, and the classification system are briefly explained. Finally, the table of results provides an interpretative design that informs about the context of actionability. Each row represents a unique predictive association between a variant and a drug in a given context (e.g. response of drug X in a cancer type Y in a clinical trial). Each predictive association is considered as one *finding*. The number of findings stratified by cancer types is provided in the figure on the right side of the report. Patient's variants are listed on the left side and the matched predictive evidences are on the right side of the table.

5.6 Implementation and Visualization with R Shiny

In order to increase the usability and visibility of the MTB Report workflow, a Shiny-based web application was developed: the interactive Molecular Tumor Board Report (iMTB-Report).

The iMTB-Report is a web application that allows an interactive visualization of actionable variants of an individual tumor genome (Figure 5.4). The user can uploead genomic data in common tabular data files (.dat, .csv, .xlsx). The specific formats of genomic data have been detailed in §5.2.

The user can select which databases should be used in the analysis, and, in case previous versions are desired, the user can upload older versions after downloading them from the databases' websites. After data uploading and

MTB Report - From somatic variants to treatment options

Department of Medical Statistics, University Medical Center Goettingen, September 27, 2017

Disclaimer: This report is intended for research use only and should not be used for medical or professional advice. We make no guarantee of the comprehensiveness, reliability or accuracy of the information on this report. You assume full responsibility for all risks associated with using this report.

PATIENT INFORMATION

Patient ID	MASTER-04	Tissue Type	?
Gender	Female	Tumor Content (%)	20
Disease	Ovarian carcinoma metastasis	Number SNVs	98
Previous Therapies	-	Number CNVs	3555
Tumor Board Decision	TSC2 stopgain : mTOR-Inhibitor (Everolimus)	Number Fusions	16

GENE-DRUG PREDICTIVE ASSOCIATIONS

Method: Somatic variants of the patient (mutations, amplifications, deletions, rearrangments) are searched in curated databases of predictive biomarkers (GKDB[1], CIViC[2]) and reported according to their clinical evidence (see Levels of Evidence). In the following two tables (SNVs and CNVs), basic information of the somatic variants with relevant clinical implications can be found:

Gene	Patient's Variant	Level of Evidence	Variant Freq.	Zygosity	Quality (Phred)
BCOR	S1024T	B3	0.18	het	88
TP53	Y236C	A2 ,B2 ,B3	0.79	het	225
TSC2	R505X	B2 ,B3	0.42	het	221

Gene	Patient's Variant	Level of Evidenc	Segment Mean	Size (Mb)
AURKA	ampl.	B3	0.58	39.2
BIRC7	ampl.	B3	0.58	39.2
BRCA1	del.	A2 ,B2 ,A3 ,B3	-0.84	2.6
CCNE1	ampl.	A3 ,B3	0.73	15.2
CDK12	del.	A3	-0.87	1.9
FANCA	del.	B2	-0.71	59.8
FRS2	ampl.	B3	0.73	6.5
MDM2	ampl.	B2	0.73	6.5
NF1	del.	B2 ,B3	-0.85	2.6
PALB2	del.	B2 ,B3	-0.70	17.5
RICTOR	ampl.	B2	1.56	12.5
SUZ12	del.	B3	-0.87	4.5
TOP1	ampl.	B2	0.58	39.2
TSC2	R505X,del.	B2 ,B3	-0.78	11.2

Segment mean: tumor/normal coverage log 2 ratio of a genomic segment (e.g. diploid regions have segment mean of zero, positve values are ampifications and negative values, deletions)

Levels of Evidence: Findings are classified into 6 levels of evidence combining the axis A-B and the axis 1-2-3. Level A means evidence in the same cancer type. Level B means evidence in any other cancer type. On the 1-2-3 axis, level 1 means evidence supported by drug approval organizations or clinical guidelines, level 2 contains clinical evidence (clinical trials, case reports) and level 3 consists of preclinical evidence. The distribution of findings into levels is summarized in the right figure

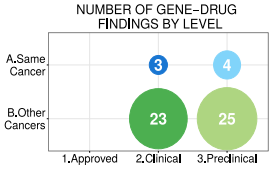


Table of Results: All the predictive associations are detailed in this table. The results are sorted by 1) drug frequency, 2) levels of evidence (A1-B1-A2-B2-A3-B3). To allow a quick interpretation, the type of association (response, resistance) is colored (green, red) and new variants are gray .

Patient				Gene-Drug	Associations			
Gene	Variant	Disease	Known Variant	Association	Drugs	Evidence	PMID	Level
FANCA	del.	prostate	any (LoF)	response	PARP inhibitors	early trials	26510020	B2
PALB2	del.	prostate	any (LoF)	response	PARP inhibitors	early trials	AACR 2015 (abstr CT322), 26510020	B2
BRCA1	del.	ovarian	del. (LoF)	sensitivity	PARP inhibitors	preclinical	22392482	A3
CDK12	del.	ovarian	any (LoF)	sensitivity	PARP inhibitors	preclinical	24240700, 24554720	A3
PALB2	del.	pancreatic	any (LoF)	sensitivity	PARP inhibitors	preclinical	25263539, NCT01585805	B3
TSC2	R505X	angiomyolipoma	any (LoF)	response	mTOR inhibitors	early trials	23312829, 21525172, 20048174	B2
RICTOR	ampl.	lung	ampl. (GoF)	response	mTORC1/2 inhibitors	case report	26370156	B2
NF1	del.	neurosarcoma	any (LoF)	sensitivity	mTOR inhibitors	preclinical	18483311, 20505189, 24509877	B3

[1] Dienstmann et al., Cancer Discov (2015), v19
[2] Griff th et al., Nat Genet (2017), version 01 June 2017

1

FIGURE 5.3: **MTB Sample Report.** First page of the report of patient MASTER-04 from the NCT MASTER dataset is shown. This figure has been modified from Perera-Bel et al. (2018) with the addition of the disclaimer at the top of the page.

database selection, input variants are queried against the databases and a list of actionable variants is compiled. The user can explore the subset of filtered actionable variants stratified by levels of evidence, and download a static .pdf report or a .csv file with the results (Figure 5.4). Links to the source of the information (links to PubMed IDs, google scholar searches and FDA website) and to other relevant databases (*Precision Medicine Knowledgebase* (PMKB) (Huang et al., 2017) and *Drug Gene Interaction Database* (DGIdb) (Griffith et al., 2013)) are provided. Moreover, the user can browse up to 34 cohorts of The Cancer Genome Atlas. The application is written in R using the *Shiny* framework (Chang et al., 2018b), which allows an easy access to users not familiar with R environment through a web interface (for more details, see §4.3).

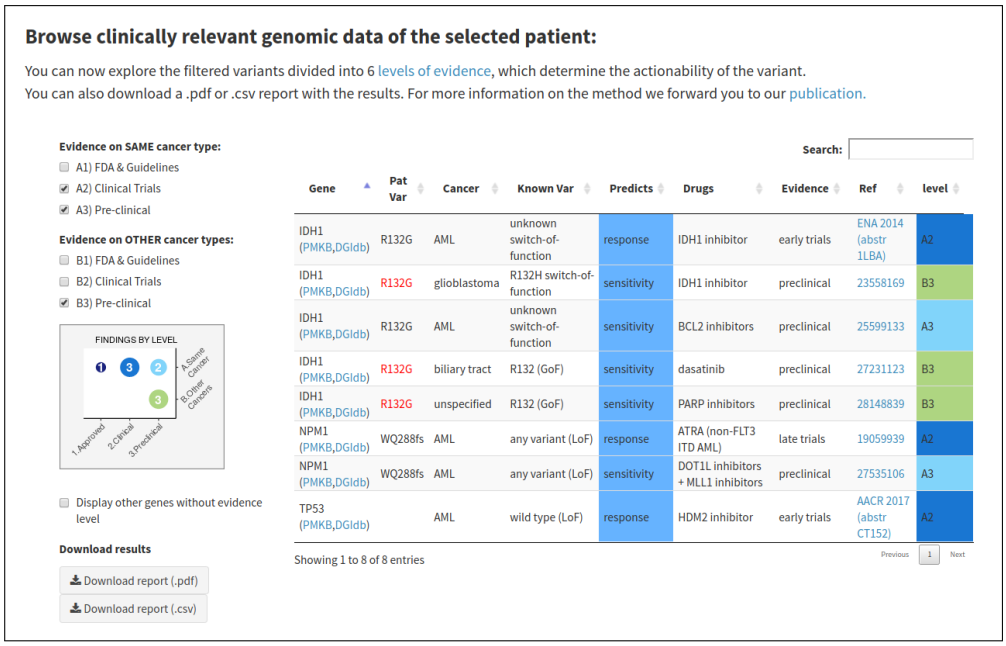


FIGURE 5.4: **Interactive MTB Report.** Interactive visualization of actionable variants of TCGA patient *TCGA-AB-2990* with AML. In the left, a figure summarizes the results by levels of evidence. On the right, the table displays the variant-drug associations, providing links to PMKB, DGIdb, Pubmed and Google Scholar. The table can be sorted according to a specific column by clicking the table headers. The variants with a *repurposing* flag are highlighted in red. Two buttons allow to download the table of results in csv format or in the MTB Report format (pdf).

Chapter 6

Scope of the Molecular Tumor Board Report Workflow

To determine the feasibility and scope of the proposed workflow (the third aim of this thesis), MTB Reports for 3184 samples from TCGA and 18804 samples from GENIE datasets were generated. In brief, data was downloaded from Synapse repository (see §4.1), actionable variants of each patient were filtered (§5.3) and then classified into levels of evidence (§5.4). The results presented in this chapter describe the actionability landscape of two cancer datasets from two projects according to the MTB Report workflow with emphasis on particular aspects: levels of evidence, cancer types, genes and pathways. Figure 6.1 depicts the results obtained in the two datasets at each level of evidence across cancer types.

It is important to highlight that the two datasets are not directly comparable, as several aspects directly influence the results. GENIE project used targeted gene panels which included fusions, whereas TCGA used WES. In A1 level we can clearly see the importance of fusions for lung cancer treatment: 14% of LUAD and 3.4% of LUSC patients in TCGA vs. 40% of NSCLC patients in GENIE (Figure 6.1c). Also, GENIE consists of patients with more advanced disease stages than TCGA. The impact of the advanced nature of the diseases can be observed by the fact that the MTB Report identifies both more actionable variants and more patients with actionable variants in GENIE dataset than in TCGA as regards to high levels (A1, B1) (Figure 6.1a,b,c). Finally, the use of selected panels in GENIE is reflected in preclinical levels, in which TCGA has higher number of patients with actionable variants than GENIE as genes investigated in preclinical studies are usually not considered actionable, and, thus, not included in gene panels designed for clinical use (Figure 6.2).

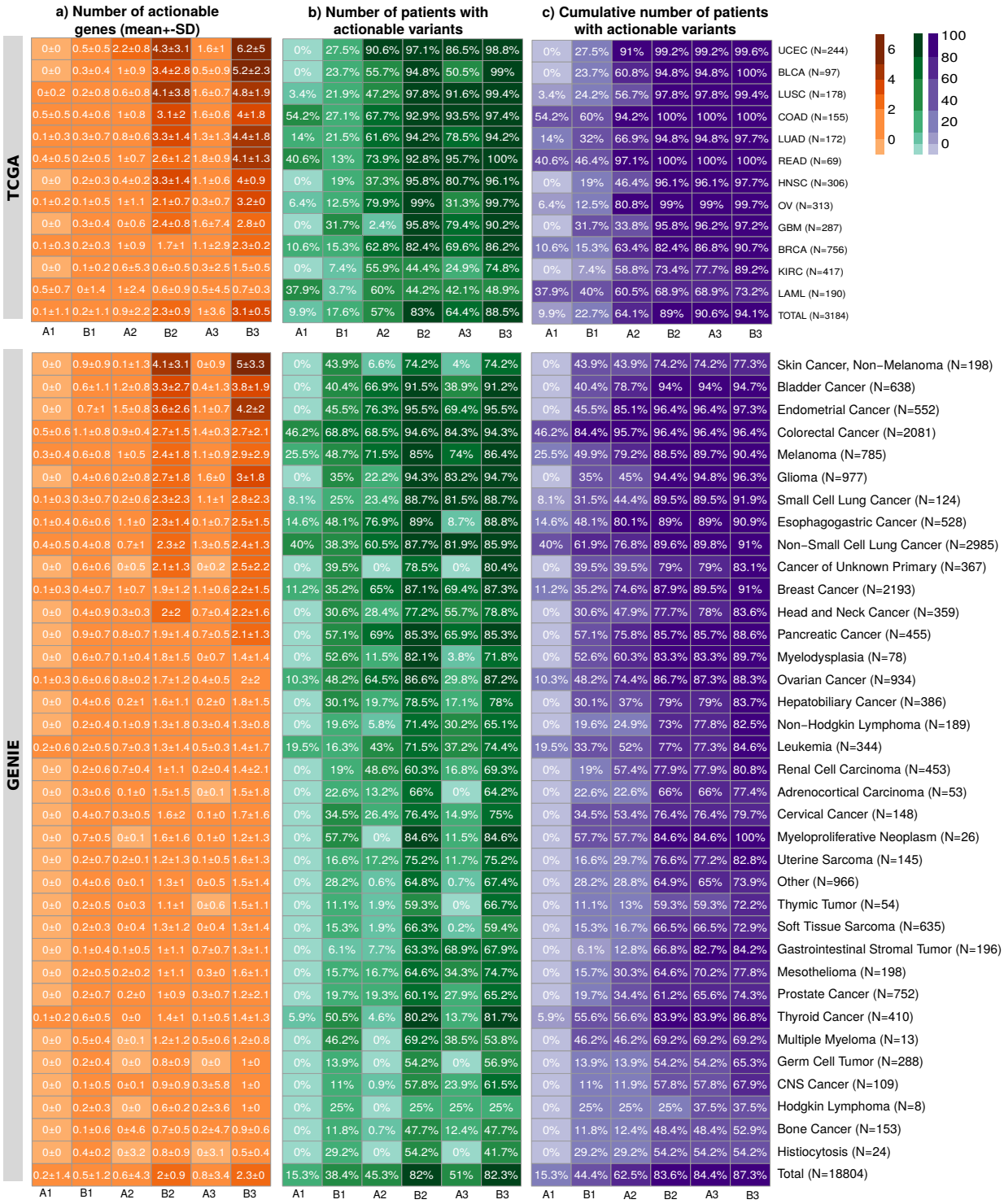


FIGURE 6.1: **TCGA (top) and GENIE (bottom) actionability landscape.** Heatmap representation of the number of actionable genes and the percentage of patients with actionable variants stratified by cancer type and level of evidence. Findings associated to resistance/no response are not included in this representation. Regarding wild-type variants, only findings in level A1 are considered, e.g., *NRAS*, *KRAS* wild type in colorectal cancer. This figure has been modified from Perera-Bel et al. (2018).

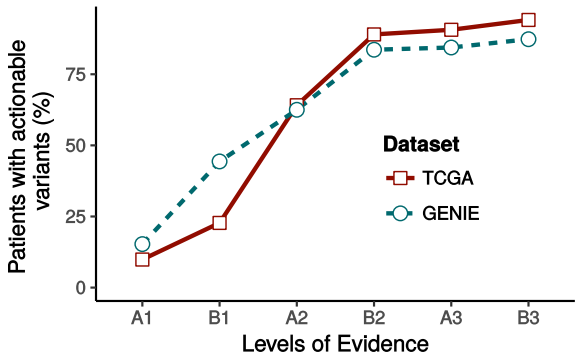


FIGURE 6.2: **Comparison of TCGA and GENIE datasets**. Cumulative number (percentage) of patients with actionable variants at each level of evidence.

6.1 Role of Levels of Evidence

A1 is the most intuitive level, as it is a synonym for companion diagnostics. In other words, A1 includes genes that are routinely checked in certain cancer entities because they are part of a drug label or are included in treatment guidelines. The workflow identified actionable variants in 9.9% and 15.3% of TCGA and GENIE patients, respectively. Notably, most cancer types do not have actionable variant at this level (in other words, just few cancer types have approved companion diagnostics, see Table A.2). The differences between the two datasets can be explained by the inclusion of fusions (important in NSCLC) and more cancer types (e.g. melanoma) in GENIE.

The simplest type of *repurposing*, in which the predictive value of a variant towards a drug is extrapolated to another cancer entity (B1 level), involves at least twice as many patients. This rise is due to the fact that the same genes that are companion diagnostics in certain entities (e.g. *EGFR-lung cancer*, *BRAF-lung cancer -melanoma -thyroid cancer*, *KRAS-lung cancer -colorectal cancer*, *NRAS-colorectal cancer*, *ERBB2-breast cancer -gastric cancer*, *BRCA1-ovarian cancer*, *BRCA2-ovarian cancer*, *RET-lung cancer*, *ALK-lung cancer*, *ROS1-lung cancer*) are also altered in other cancer types at lower frequencies (Figure 6.3 and Figure A.1). Indeed, in a dataset enriched in advanced cancer patients -more severe diseases and more acquired mutations– such as the GENIE, around 40% of patients were found to have actionable variants at B2 level (precisely 38.4%; 44.4% taking into account A1 and B1).

Clinical trials levels (2) include around 60% of patients as regards to the same cancer type (A2) and above 80% pooling clinical trials on any cancer type (B2). On the other hand, preclinical levels (A3, B3) do not have a large impact

on the aforementioned numbers, increasing only 1-5 points the percentage of covered patients. This supports the idea that clinical trials are already studying the majority of cancer genes, or at least a subset of genes altered in the majority of patients. Another issue is whether clinical trials are able to prove any predictive value of the variants studied (e.g. many trials study *TP53*, though none has yet been successful).

6.2 Cancer Type Particularities

Main differences in the percentage of patients with actionable variants between cancer types are observed at A levels (Figure 6.1b). Whereas some cancer types have high percentages of patients with actionable variants (breast cancer has 62.8% and 65% for A2 level, and 69.6% and 69.4% for A3 in TCGA and GENIE datasets, respectively), other cancer types do not present actionable variants at A2 level (germ cell tumor, multiple myeloma) and A3 level (adrenocortical carcinoma, germ cell tumor). In contrast, the same cancer types that do not include any patients at A levels, present similar percentages to the rest of cancer types at the corresponding B levels (germ cell presents 50% of patients with actionable variants at B2 and B3 levels; multiple myeloma 69.2% at B2 level; adrenocortical carcinoma 64.2% at B3 level). Overall, percentages at B levels are higher and more uniform than A levels.

Variations in A levels between cancer types reflect the relevance of certain pathways in shaping drug response in specific cancer entities: *BRAF* mutations in melanoma, *DNMT3A* and *NPM1* in AML, *KRAS* in colorectal cancer, *TP53* in ovarian cancers, *PIK3CA* in breast cancers, *EGFR* in head and neck malignancies and *PTEN* and *PIK3CA* in uterine cancer (Figure A.1). Unfortunately, TCGA does not properly reflect current status of lung cancer therapy (which is to a large extent fusion-based) because the data *freeze* from TCGA used in this thesis did not include fusions.

Kidney cancer presents a large range of mutated genes but at really low frequencies which complicates reaching half of the patients. 50% barrier is only reached at A2 level with late clinical trials studying *VHL* as a biomarker, and at preclinical studies of *BAP1*, *VHL*, *PBRM1 and SETD2*. AML also shows a distinct pattern, in which 37.9% of patients present actionable variants (*DNMT3A* and *NPM1* chemotherapy biomarkers) but then it has the flattest slope with the lowest cumulative percentage (73.2%).

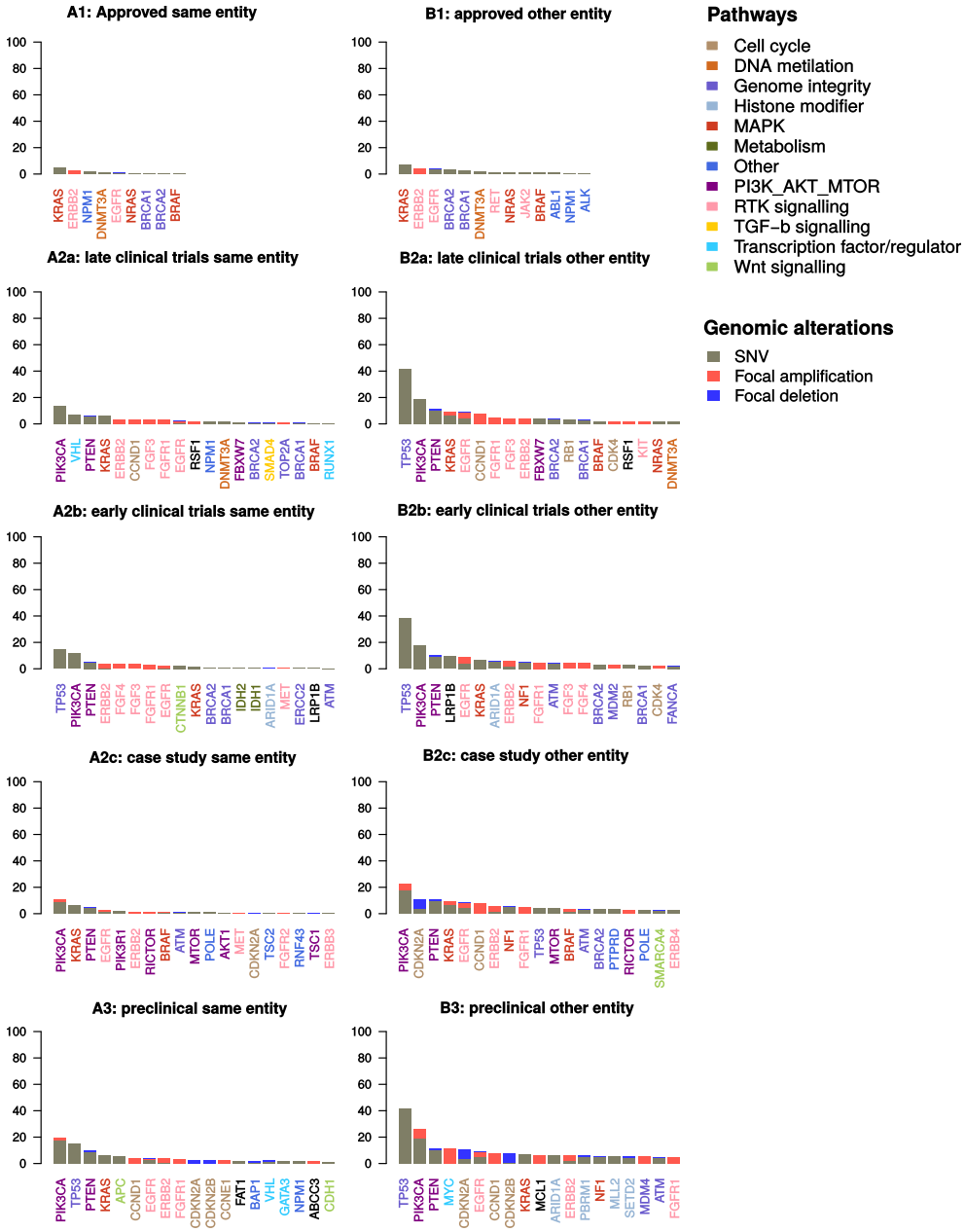


FIGURE 6.3: **Actionable genes by levels of evidence in the TCGA dataset**. Barplots depicting the percentage of patients (including all TCGA cancer types) of the top twenty biomarkers of each level of evidence. Level 2 is divided into three groups: 2a (late clinical trials), 2b (early clinical trials), and 2c (case reports). Colors in the bars denote the type of genomic alteration. Genes are colored according to manually curated pathway annotations.

6.3 Common Actionable Genes and Pathways

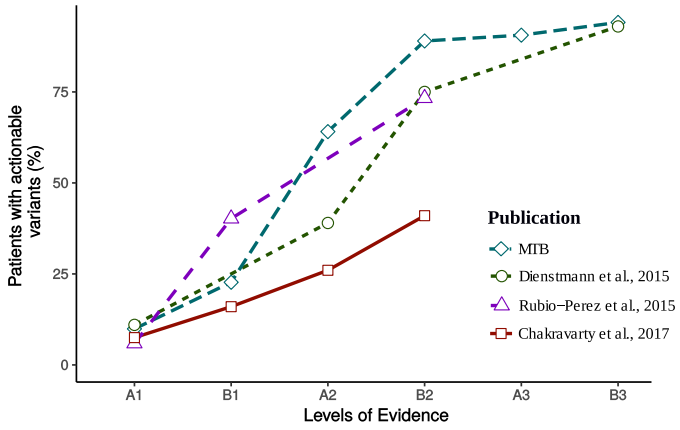
Exploring the more prevalent actionable genes can give an idea of the main targeting strategies (Figure 6.3). Drugs classified in high-evidence levels target mainly RTK-signalling, MAPK and Genome Integrity genes. Targets from the PI3K/AKT/MTOR and cell cycle signaling pathways may soon enter clinical routine (late trials). New strategies (early trials, case studies and preclinical) include genes involved in metabolism, WNT signaling and epigenomic pathways (histone modification).

The aforementioned lack of fusions in TCGA dataset is also illustrated in Figure 6.3. For instance, the lack of *ROS1*, *ALK* and *ABL* fusions in A1 level. Interestingly, this figure also depicts, by the colors of the bars, the fact that LoF alterations (deletions) are more difficult to target than GoF (SNVs and amplifications).

6.4 Comparison to other Publications

Taking advantage of the fact that TCGA has been studied in many contexts, we could find three studies with a scope similar to this thesis (i.e. identification of actionable genes). The results reported in these studies were compared to the results of the MTB Report workflow. The comparison is shown in Figure 6.4. Although there is high concordance between the studies at A1 level, as soon as *repurposing* is considered, the reported numbers start to diverge. OncoKB results are the most different, presumably by the use of a more stringent definition of actionability, that, moreover, does not consider preclinical evidence. The other three studies, including ours, present similar results. This indicates that the databases used are an important factor in the process of identifying actionable variants, as MTB Report uses GDKD, the database from Dienstmann et al. (2015a). It is therefore reasonable that, at most levels, the MTB Report shows the highest number of patients with actionable variants as it takes advantage of several databases.

In all, factors that can explain the divergence between studies are: databases used, definitions of actionability and characteristics of the TCGA dataset included in the analysis (e.g. cancer types, genomics data types). The general increasing trend along decreasing strength of evidence shows that low levels of evidence certainly increase the actionable landscape of cancer patients. Whereas an approach like OncoKB might be more accurately reflecting current clinical implementation of NGS, the other three approaches highlight the potential of NGS in guiding clinical trials and preclinical studies.



Publication	Standard therapy		Clinical trials		Preclinical		Total	# samples	Databases
	Label	Off-label	Label	Off-label	Label	Off-label			
MTB	9.9 (A1)	22.7 (B1)	64.1 (A2)	89 (B2)	90.6 (A3)	94.1 (B3)	94	3184	GDKD, CIViC,TARGET
Dienstmann et al., 2015	11 (5)	-	39 (4)	75 (3)	93 (1-2)		93	4392	GDKD
Rubio−Perez et al., 2015	5.9	40.2	73.3		-	-	73.3	4068	Rubio-Perez et al., 2015
Chakravarty et al., 2017	7.5 (1-2A)	16 (2B)	26 (3A)	41 (3B)	-	-	41	5983	OncoKb

FIGURE 6.4: **Comparison of TCGA actionability in different *in-silico* studies**. Cumulative number (percentage) of patients with actionable variants at each level of evidence from our workflow (MTB) and three other publications. Equivalences between levels of evidence is not always one-to-one: Dienstmann et al. (2015a) does not report B1 level, and A3 and B3 levels are reported together; Rubio-Perez et al. (2015) reports together the equivalents of A2 and B2 levels.

Chapter 7

Proof-of-Concept Application

This last chapter of *Results* part details the analysis addressing the last of the aims of this thesis, that is, to provide a proof-of-concept for the clinical utility of the MTB Report workflow. The MTB Report was used to identify actionable variants in WES of matched tumor and normal samples of eleven patients from the MASTER study. The clinical aspects of the samples, the sequencing details, the bioinformatic processing and the identification of somatic variants are described in §4.2.

7.1 Genomic Landscape of Patients from the MASTER program

A highly varying number of SNVs and CNVs can be observed from patient to patient (Table 7.1). Patients MASTER-05, -09 and -10 had as little as 10-11 SNVs, whereas MASTER-06 had 2703 SNVs. MASTER-06 was a clear case of an hyper-mutated tumor and also presented many broad CNVs, which, rather than indicating driver events, are most likely a consequence of undergoing too many chemotherapy cycles. The three patients with the lowest number of SNVs had also the lowest number of genes affected by CNVs (107, 1, and 5, respectively). However, most patients presented broad CNVs, which translated to thousands of genes affected by CNVs (e.g. MASTER-01-04, -06 and -08). By definition, broad CNVs alter the copy number status of large numbers of genes as they span at least 3Mb of the genome (some affect up to whole chromosomal arms).

The complete list of actionable variants and their therapeutic implications identified in each MASTER patient by the MTB Report workflow can be found in the eleven reports published as Additional File 5 from Perera-Bel et al. (2018). Nonetheless, a summary of the findings is also provided in Table 7.1. The MTB Report identified from 0 to 21 actionable SNVs per patient (median of 2, *Interquartile Range* (IQR) of 2.25), and from 0 to 14

CNVs per patient (median of 5, IQR of 4.75). The number of findings (i.e. drugs matched in a given context) associated to the actionable variants ranged from 1 to 92 (median of 43, IQR of 55.75). Overall, 10 unique high-level (A1, B1, A2), potentially clinically actionable findings were identified: two *BRCA* deletions, one *EGFR* exon 19 indel, two *TP53*, two *ABL1* and one *MTOR* missense mutations. Besides, one missense mutation in *TP53* and one in *KRAS* predicted both response and resistance. Due to the advanced nature of MASTER patients (inclusion criteria of this study are young adults with advanced diseases), low-evidence actionable variants were of special interest.

7.2 Comparison to Actionable Variants Identified by the MASTER program

A comparison study was performed to assess the relevance of the variants filtered by the MTB Report workflow. Table 7.1 shows the comparison of the actionable variants identified by MTB Report with the actionable variants selected by the experts' panel of the MASTER study. The experts' comprised bioinformaticians and translational oncologists responsible to interpret and prioritize the results produced by WES. The general process for the interpretation of variants consisted of a manual revision of the annotated list of high-quality variants, visualization of the alignments, and prioritization of actionable variants (described in Horak et al. (2017)). Main considerations were i) quality of the variant (e.g. allele frequency, segment mean, gene overlap and length of copy number events) and ii) clinical, functional and biological implications (by means of annotations to databases). The most relevant candidates (one to three) were discussed in the MASTER's program molecular tumor board and are detailed in Table 7.1.

With regard to this comparison study, platin-based chemotherapies and PARP inhibitors were considered as equivalents. Hence, a *match* status is shown in Table 7.1 when, upon the disruption of DNA repair pathway genes (*BRCA1*, *BRCA2*, *RAD51*, *PALB2*, *CDK12*), either drug was suggested (as in case MASTER-03). This decision is based on the evidence that PARP inhibitors have antitumor activity in *BRCA1/2* mutation ovarian cancer, which is in turn associated with platin-based agents response (Fong et al., 2010).

Out of 20 variant-drug associations manually identified by the experts' panel, 15 were also filtered by the MTB Report. These results show a high concordance of MTB Report with experts judgments, though perfect matching was not achieved. The mismatches can be explained by mainly two reasons: information lacking in the MTB Report workflow-associated databases (*PTPRJ*,

PTPN12, *LCK*), and MTB Report filtering rules do not allow extrapolation of fusion evidence to missense mutations (e.g. MASTER-09 *NTRK3* missense mutation and *NTRK3* it is present in GDKD only as a gene fusion biomarker).

The MTB Report reported many more findings per patient than the ones discussed in MASTER's program molecular tumor board. Each actionable variant was matched to an average of 4.9 unique findings (ranging from 1 to 15) (calculated dividing the number of findings by the number of actionable variants in Table 7.1). Necessarily, many gene–drug associations are repeated among the findings; however, each finding is unique in terms of cancer type or clinical evidence. MASTER-02 and -06 showed the highest number of findings: the first, because it carried a *KRAS* G12D mutation, which is a very prevalent and highly controversial mutation in terms of drug response; the latter, because many cancer-related genes were mutated thus yielding many actionable mutations.

Patient MASTER-04 (Figure 5.3 presented *BRCA1* deletion (sensitive for PARP inhibitors in ovarian cancer and level of evidence A2) and *TSC2* nonsense mutation (sensitive for MTOR inhibitors, level B2). The MTB Report reports more and higher-ranked findings for PARP inhibitors than for MTOR inhibitors; yet, the latter was selected by the experts' panel. This apparent inconsistency is explained by the fact that this patient case is from 2014 whereas the MTB Report was generated after 2016, when olaparib was already approved (December 2014). So back in 2014, it was not possible to prescribe this drug and, hence, it was not considered. Furthermore, a heterozygous *BRCA1* loss would not have been considered a rationale for PARP inhibition. This view only changed after the paper by Mateo et al. (2015) who postulated that heterozygous alterations in homologous recombination DNA repair genes confer sensitivity to olaparib. This case illustrates the rapidly evolving landscape of targeted cancer drugs and highlights that NGS can be critical as the catalog of actionable genetic lessons is constantly expanding.

The hypermutated genotype (2703 SNVs) of patient MASTER-06 led to the second largest number of findings reported (89) and the largest number of actionable SNVs (21). Among them, *ATM* missense mutation could be suggested as one of the causes of the hypermutated genotype, as it is a DNA repair protein. According to the experts, although four actionable genes were discussed within the MASTER program, the current common practice for patients with more than 400 missense mutations (or more than 100 in patients with colorectal cancer) would be to check for PD-L1/PD1 expression as a rationale for checkpoint inhibitors. Conversely, the other case with as many

findings reported was MASTER-02 (92), in which 60% of them belong to just one genomic variant: *KRAS* G12D. Whereas *KRAS* is a negative biomarker for some cancer entities, targeting *KRAS*-mutant tumors remains to be one of the main challenges in oncology.

MTB Report did not reveal any actionable variant matching to expert's panel suggestions for patient MASTER-09. The bad quality of this sample and the few mutations to be assessed (11 SNVs and 1 fusion), yielded the consideration of extreme repurposing from the experts' side, suggesting *NTRK3* missense mutation as potentially actionable. Hence, it is acceptable that the MTB Report workflow was not able to detect this actionable variant.

It is worth mentioning that two of the matches were achieved by the use of TARGET and Meric-Bernstam et al. (2015b) list, namely *FGF1* in MASTER-01 and *ERRFI1* in MASTER-08 (no level of evidence specified in Table 7.1).

TABLE 7.1: **Retrospective study of MASTER actionability**. Comparative evaluation of eleven patients in which the actionable variants found by an experts panel (MASTER Report) were compared to the actionable variants found by the MTB Report workflow (MTB Report). # stands for number. Here, CNVs refers to the genes within copy number regions. The column Match and Match Level indicate whether the expert suggestion was found in the MTB Report. Drug Support refers to the number of times the drug suggested by the experts appeared in the MTB Report. #Findings refers to the number of unique predictive associations in the report. This table has been modified from Perera-Bel et al. (2018).

ID	Cancer	#SNV	#CNV	Fusions	MASTER Report		MTB Report					
					Gene	Drug	Match	Match Level	Drug Support	#Actionable SNVs	#Actionable CNVs	#Findings
01	Breast cancer met.	104	3045	5	BRCA1/2 deletions RAF1, PDGFRA amplifications FGF1 (T8N)	PARP inhibitors sorafenib FGFR inhibitor	YES YES YES	B2 B3 -	11.60% 4.60% 2.30%	2	13	43
02	Pancreatic adenocar.	49	1433	1	KRAS (G12D)	-	-	-	-	1	9	92
03	Leiomyosarcoma	31	3964	6	PTPRJ deletion CDK12+BRCA2 deletions	pazopanib cisplatin	NO YES	- B2	- 50.00%	0	6	12
04	Ovarian carcinoma	98	3555	16	TSC2 (R505X)	mTOR inhibitor	YES	B2	12.50%	4	14	56
05	Myxoid liposarcoma	11	107	2	PIK3CA (C420R) + PTEN (R130G)	mTOR inhibitor AKT inhibitor PI3K inhibitor	YES YES YES	B2 B2 B2	9.30% 2.40% 32.00%	2	3	75
06	Neuroendocrine	2703	2114	5	MTOR (P2490L, G332R) PTPN12 (S509N, G523S) KIT (A837T) LCK (P74L)	mTOR inhibitor lapatinib, erlotinib, imatinib, desatinib imatinib, desatinib desatinib	YES NO YES NO	B2 - B3 -	16.80% - 1.10% -	21	7	89
07	Neuroendocrine	657	425	5	ERBB3 (V104M), RAF1 (S259P), MTOR (E1485G)	mTOR inhibitor	YES	A2	12.50%	10	2	40
08	Cholangiocarcinoma	28	1001	1	ERRFI1 (R199X)	erlotinib	YES	-	0.58%	2	2	17
09	Clear cell sarcoma	11	1	20	NTRK3 (R116W)	lestaurtinib, midostaurin	NO	-	-	1		1
10	Histiocytic sarcoma	7	5	1	BRAF (F595L) + HRAS (Q61R)	MEK inhibitor sorafenib (multi TKi)	YES YES	B2 B3	33.30% 5.50%	3		18
11	Pulmonary adenocar.	70	133	1	EGFR (p.745-750del)	erlotinib	YES	B2	22.20%	3	5	54

PART IV

DISCUSSION AND CONCLUSIONS

Chapter 8

Discussion

NGS is increasingly being used in clinical settings to characterize advanced cancer patients with the aim of informing about diagnosis, prognosis and treatment, in particular within precision oncology trials. NGS platforms (e.g. *hotspot* panels, WES, WGS) are able to identify large amounts of genomic variants, many of which still have unknown clinical implications. Hence, assigning clinical meaning to the genome of a patient is an overwhelming task. The MTB Report is a workflow that automates a number of cumbersome and time-consuming steps which are usually carried out manually in preparation of MTBs. It offers a pre-filtered list of actionable variants that may indicate vulnerabilities of the tumor thus facilitating the clinicians' work in deciding for a treatment.

Through the analysis of large public datasets we have shown that the MTB Report results are largely concordant to other studies (Marquart et al., 2018), with regard to variants with high-evidence predictive associations: only around 10% of cancer patients in TCGA and 15% of cancer patients in GENIE to date are eligible for genomically-guided therapies – the chances are completely determined by the cancer type, as companion diagnostics are approved for specific cancer entities. Furthermore, the MTB Report is able to identify actionable variants in over 90% of patients when low-evidence levels are considered. The main strength of the workflow is the automatic aggregation of information which makes the whole process scalable for clinical practice. Nevertheless, as noted by experts of the MASTER trial, all MTB Report findings have to be rated and reevaluated for their therapeutic impact with regard the clinical course of each patient (e.g. response to prior therapies, side effects), patient characteristics or relevant pathway interactions in case of multiple findings.

The implementation of the workflow as a web-based application increased

the usability of the MTB Report, allowing user-defined data, providing interactivity and visualization of the data, and multiple download options. The web-page distribution makes the workflow accessible to users without programming skills (researchers, clinicians). Instead, the distribution as a stand-alone application through GitHub promotes the incorporation of the tool as part of bioinformatic pipelines.

8.1 Aspects of the Molecular Tumor Board Report

The strengths and weaknesses of the MTB Report are discussed in the following pages and will hopefully be addressed in future publications of this fast-growing field of genomics-driven oncology.

8.1.1 Definition of Actionability

Actionability is a concept that has been used in the context of precision medicine with slightly distinct connotations. Different definitions of actionability have been shown to influence the results of *in silico* prescription publications, as we have seen in the comparison between four studies that analyzed the actionability landscape of the TCGA dataset (Figure 6.4). Differences between Rubio-Perez et al. (2015) and the MTB Report can be explained by a combination of three factors: 1) their dataset comprises more cancer types than ours, 2) there is a lapse of time between both publications and the field evolves rapidly, and most importantly, 3) the gene-drug associations considered are different. Whereas some clinical interpretation approaches include genes that are drug targets even if the drug prescription is independent of the mutational status of the target (Rubio-Perez et al., 2015; Hintzsche et al., 2016), our method focuses only on variants that have been shown to have a predictive value on drug response (i.e. predictive biomarkers). For instance, an *EGFR* mutation in head and neck cancer is not considered as A1 level in our study because cetuximab prescription is independent of *EGFR* status (Network, 2006). Differences with Chakravarty et al. (2017) are mostly due to the database they used, OncoKB. This database differentiates between oncogenic (i.e. driver) and actionable variants. A variant is considered actionable only if there is compelling clinical evidence of the biomarker as being predictive of response to a drug (early phase trials, preclinical and VUS are not included). As a result, OncoKB contains 38 actionable genes[1], which translates into 41% of samples with actionable variants (equivalent to our

(1) Genes with annotations in levels 1 to 3, downloaded 25 September 2018

A1+B1+A2+B2 but excluding case reports and variants with a *repurposing* flag) (Chakravarty et al., 2017). In contrast, the databases used by the MTB Report contain a larger number of actionable genes: 170 in GDKD, 213 in CIViC and 111 in TARGET[2].

Even though these *in silico* studies differ in the percentages, the overall message remains the same: reporting low-evidence biomarkers (i.e. *off-label* use and on substances in clinical trials) undeniably increases treatment recommendations for cancer patients.

With respect to clinical studies, in which mostly high-evidence actionable variants are considered, recent prospective trials using NGS to guide treatment decisions have reported informative variants in a wide range of patients, reflecting the definition of actionability used in each study. Studies comprising high-evidence findings only (A1, B1, and open clinical trials within A2) have reported actionable variants in 49% (Sohal et al., 2016), 48% (Massard et al., 2017) and 36% of patients (Zehir et al., 2017). In contrast, studies that include also low-levels of evidence (therapies in clinical or preclinical development) reported actionable variants in the majority of patients: 94% (Beltran et al., 2015), 82% (Rennert et al., 2016) and 75% (Horak et al., 2017). Our analyses of TCGA and MASTER datasets are comparable to the studies that use a broad actionability definition, therefore, above 90% of patients were reported to have actionable variants. Interestingly, in cases in which no evidence was found, MASTER experts considered extreme repurposing between variant types; in this case, from evidence on fusions to missense mutation in *NTRK3*.

Actionability rises as a dynamic concept in which variant-drug associations are supported by a range of evidence strengths. Besides, the rate of actionability varies by cancer type and over time as more evidence is generated. The MTB Report expands the treatment options delivered by applying a broad definition of actionability to associations in which neither the drug nor the variant have been approved. Such strategy can be acceptable for patients whose therapeutic options have been exhausted beyond the standard of care (Meric-Bernstam et al., 2015b), but always informing about the actionability strength through the classification into levels of evidence (see §8.1.4).

(2) Genes with predictive associations, versions specified in §3.2

8.1.2 Off-label Prescription

Off-label prescription of drugs has been successfully done for many patients for which no further treatment lines nor clinical trials were available (Conti et al., 2013; Kordes et al., 2016; Chau et al., 2016). It is also true, though, that there is explicit clinical evidence against off-label prescription of a drug in some cases. *BRAF* V600E mutation is a predictive biomarker in melanoma and NSCLC; however, clinical studies already showed that colorectal patients with this same mutation do not respond to RAF inhibitors (Kopetz et al., 2015; Hyman et al., 2015). The same applies to *ERBB2* amplification, which is a predictive biomarker for breast and gastric cancers (Table A.2), but has failed to show predictive value upon treatment with trastuzumab in lung cancer patients with *ERBB2* amplification (Lara et al., 2004). Indeed, all these examples are compiled in the databases used by the MTB Report, so both the repurposing options and the resistances will be identified by the workflow. These cases of conflicting evidences render interpretation of MTB Reports more complex, such as, for instance, patient MASTER-02 in which *KRAS* mutation yields 92 findings.

8.1.3 Standardization of Annotations

At the time this thesis' work started, the knowledge bases were mainly lists/tables curated by oncologists collecting the most important pieces of evidence, and, as such, did not use systematic annotations (e.g. ontologies, databases). Next databases, like CIViC, PMKB, OncoKB, Cancer Biomarkers, established a new concept and adopted the use of standardized annotations which allowed a semi-automatic curation process and integration with other databases.

Currently, the use of reference annotations for genes and variants is widely embraced in the community. All the above-mentioned databases standardize genomic annotations by using gene IDs, reference genome builds, genomic coordinates, transcript identifiers and HGVS format. The MTB Report will incorporate such standards in next releases.

The use of public ontologies to annotate cancer types has been already adopted by some databases, such as CIViC, JAX-CKB (Disease Ontology), OncoKB (OncoTree ontology), clinicaltrials.gov (MeSH terms) and COSMIC (National Cancer Institute thesaurus and Experimental Factor Ontology). However, many other databases still use *in-house* taxonomies for tumor types and tissues (GDKD, Cancer Biomarkers, PMKB, TARGET). Databases (and tools such as the MTB Report) that want to be widely used will have to adopt

the use of ontologies as the only way to ensure reproducibility.

Cancer type annotations are especially important, as the classification system used (levels of evidence) depends on the definition of cancer type. For instance, a patient with lung adenocarcinoma will have A-level actionable variants of lung adenocarcinoma and NSCLC. However, if there are findings on particularly squamous lung cancer, they will be considered as B together with the findings on AML, for instance. Hence, the more generic the tumor type supplied as input is, the less specific the results at A levels of the MTB Report will be.

It will be likewise important to include molecular (sub)classification of cancer types, as they will complement current histological classifications (Hoadley et al., 2014, 2018). The MTB Report is able to match genomic alterations to molecular subtypes that respond better to a therapy (e.g. breast cancers with *HER2* amplification are matched to trastuzumab; lung cancers with *EGFR* mutations are matched to EGFR inhibitors, and with *MET* amplifications to crizotinib). However, this is true as long as the molecular subtype (HER2 positive) and the biomarker are measured at the same molecular level (that is, genomic). Transcriptome profiling has been used to identify molecular subtypes in breast cancer (Perou et al., 2000), lymphoma (Alizadeh et al., 2000; Barton et al., 2012) and colorectal cancer (Guinney et al., 2015). The subsequent step of such studies is the identification of subtype-specific biomarkers to stratify treatment (Barton et al., 2012; Bramsen et al., 2017). For that, the MTB Report should allow as *cancer type* input not only tumor types and tissues but also molecular subtypes. Yet, this will require a change of paradigm, as ontologies do not include molecular/genetic subtypes (besides some exceptions like breast cancer subtypes based on hormone receptor status, or breast and ovarian hereditary cancers).

Regarding drugs, annotations are far less standardized. Databases such as DrugBank, Chembl, PubChem or Target-based Classification of Drugs from KEGG collect some standardized names and descriptions of drugs and their targets. Yet, no ontology exists that integrates drug categories (e.g. *FGFR* inhibitors), generic and commercial names of drugs, and compounds under investigation. Still, publications seeking some kind of standardization invest large efforts on *in-house* solutions (Griffith et al., 2013; Iorio et al., 2016; Tamborero et al., 2018a; Piñeiro-Yáñez et al., 2018).

8.1.4 Levels of Evidence

Levels of evidence serve for highlighting the strength of evidence supporting the actionability of a variant and, hence, make the MTB Report easier to interpret in terms of the significance of the clinical impact. Many other classification systems have been proposed, and our 6-levels system is not a substitute for any of those. It is rather an orientation tool from which clear relations can be established to the other classification systems (as it was done in Figure 6.4 to compare results of four studies analyzing the TCGA dataset).

The Association for Molecular Pathology, the American Society of Clinical Oncology and the College of American Pathologists have recently proposed a classification system for actionable variants in a joint effort to establish recommendations for reporting somatic variants (Li et al., 2017). Li et al.'s classification comprises six levels: *A1*, *A2*, *B*, *C1*, *C2* and *D*. Compared to our classification, *A* and *B* segregate more precisely our A1 level in terms of approval-status (*A1*), professional guidelines (*A2*) and well-powered studies with consensus from experts (*B*). *C1* equals our B1, whereas *C2* in our case is further divided into A2, B2 and includes both clinical trials and case studies. Finally, level *D* is equivalent to both A3 and B3. We can find in the literature a large range of complexity regarding classification systems, from three simple categories (targets of approved drugs, drugs under development, and VUS) in Beltran et al. (2015), to ten categories in Van Allen et al. (2014). Between these extremes, Meric-Bernstam et al. (2015b) and Horak et al. (2017) have proposed similar six-level systems stressing the statistical soundness and power of clinical trials to be able to justify the prescription of the drug.

Our six-level classification scheme emphasizes i) the strength of the clinical evidence of the predictive association between the biomarker and the drug (axis 1–3) and on ii) the activity in cancer type (axis A–B). The first, because drug approval is crucial for drug prescription; the second, to better inform about *off-label* use and expand the *repurposing* of drugs.

8.1.5 Report Design

The MTB Report design is one of the main features of the workflow. The report is designed as a supporting tool for researchers, oncologists, pathologists, bioinformaticians, who have to interpret results from NGS technologies in search for actionable variants. Therefore, it includes the maximum knowledge relevant for assessing the actionability of variants but keeping the presentation as simple and compressed as possible. The report includes a disclaimer of

liability withdrawal, as it not a clinical test and is, hence, intended for research use only.

To allow an intuitive reading of the report, a first table summarizes the genes found to be actionable. In this table, information on the allelic frequency of the variant, allelic status and the quality of the variant is provided. For CNVs, we report the segment mean and the size of the segment. However, in the iMTB Report the information provided can be customized according to user's preferences.

The main table contains the predictive associations between the identified actionable variants and drugs. The table is sorted according to drug frequency because it was shown to be more convenient for interpreting the therapeutic options. Sorting by levels of evidence scatters the drugs along the table, which complicates the interpretation. However, genes might be repeated along the table. In this respect, the iMTB-Report implementation allows a more flexible and user-oriented experience. The user can decide to sort the findings in the interactive table according any variable or to filter the results to only one level of evidence. Nevertheless, the static version of the report with the date of issue is a general practice for medical records and that is the reason why the MTB Report (and not a CSV or the interactive web-based tool) is the main output of the workflow (Li et al., 2017).

As the use of NGS has entered clinical practice, guidelines and recommendations for reporting this kind of data have been published. The MTB Report adopted a great number of such recommendations. Although mainly addressed at germline variants, some of the reporting recommendations of Richards et al. (2008, 2015) are universal. Tables are recommended for large numbers of results (e.g. NGS tests) as they are able to convey any information in a simple format. Essential components that should be included are: variant, gene, disease, allele frequency, tumor content and variant classification. The report should keep record of the method (sequencing technique, bioinformatics software, interpretation) and the version of any databases used to provide the possibility to reanalyze in the future and to reproduce past analyses. Any variant classification (levels of evidence in the MTB Report) should be supported by literature. Recommendations for the design of molecular genetic reports (Suthers, 2009) were incorporated in the MTB Report design: provide consistent and informative headings, limit the information under each heading, provide visual clues to the structure of the report, set the context, meet the needs of other readers and keep the report length to the minimum. Matthijs et al. (2016), on behalf of EuroGentest and the Eu-

ropean Society of Human Genetics, coincide in this point, and also in the minimum content of the report: laboratory, patient identification, sample type, context (e.g. diagnosis), test description and results. Furthermore, they state that laboratories need to define the list of variants being tested and need to have an automated system to match patients and variants to allow a re-classification if necessary (Bowdin et al., 2016), both statements being fulfilled by the MTB Report. Finally, the discussion of whether VUS should or should not be reported is left to the laboratory's choice. Nevertheless, the choice should be stated beforehand and it is, in general, recommended to do so under a separate category (Matthijs et al., 2016; Li et al., 2017).

One of the main shortcomings of the MTB Report is the lack of genomic-level annotations of the variants. If provided as input, it is possible to include genomic and coding annotations in the first table containing the summary of the actionable variants. However, the incorporation of this information to unambiguously match patient's variants requires the use of HGVS standards by the databases used by the MTB Report, which only CIViC does. Future releases of the workflow should include new databases (e.g. Cancer Biomarkers, OncoKB) in which standard annotations are provided; more will be discussed in §8.1.6.

8.1.6 Databases for the Interpretation of Actionable Variants

Interpretation of genomic data requires to keep up with a combination of knowledge on different areas such as molecular biology, oncology, pathology, bioinformatics and genomics. Since 2011, My Cancer Genome (`https://www.mycancergenome.org/`) has offered a synthesis of the most relevant literature on all these areas. Though it is still one of the most used resources to guide genomics-driven therapies, the data is not structured and informatics tools can not systematically parse the information. Since approximately 2014, there has been an important development of genomics-driven oncology, which is reflected by the number of databases with a very similar scope being published within the last couple of years (notice in Table 2.1 the year of publication of the databases under the category *Clinically actionable variants (biomarkers)*). Unfortunately, little overlap has been observed among databases of actionable variants (Griffith et al., 2017).

Between 2015 and 2016, GDKD and CIViC set the precedent for the curation of actionable somatic variants in cancer. The MTB Report was to a large extent developed around these two knowledge resources. Since then, many other databases have emerged that use similar annotation layers: JAX-CKB,

OncoKB, PMKB, Cancer Biomarkers, myvariant.info (Xin et al., 2016), DEPO (Sun et al., 2018), among others. The Cancer Genome Interpreter has partially overtaken the principle of GDKD and recently published a web-tool that hosts the Cancer Biomarkers database, which is an improved version of GDKD in terms of annotations (it uses complete genomic annotations, in-house drug and cancer type taxonomies) and accessibility (Tamborero et al., 2018a). Institutions behind several of these databases have gathered in the Variant Interpretation for Cancer Consortium (http://cancervariants.org/). Created as a working group under the umbrella the Global Alliance for Genomics Health, their aim is to integrate individual databases (CIViC, OncoKB, PMKB, MolecularMatch, JAX-CLKB, Cancer Biomarkers and BCCancer) into a cross-knowledgebase platform [3] and set standards for the community to share, organize and collect this kind of data. The natural course of the MTB Report will be to incorporate those databases that are regularly updated and maintained. Though time will show which databases will be able to get the institutional support required, such cross-institution efforts will improve sustainability, standardization and accessibility of cancer variants' actionability knowledge.

8.1.7 Tools for the Interpretation of Actionable Variants

Recently, several computational algorithms have been added to the toolbox of matching genomic alterations to drugs. These tools are comparable to the MTB Report, as they are oncology-focused, allow a multi-query of genomic variants against selected knowledge bases and apply certain heuristic rules or prediction algorithms to prioritize drugs. With regard to approved drugs, all tools rely on the same resources. However, to expand the therapeutic landscape, each tool follows a particular approach. PanDrugs uses multiple resources and puts emphasis on pathway repurposing (Piñeiro-Yáñez et al., 2018). The Cancer Genome Interpreter identifies driver variants and uses the Catalog of Validated Oncogenic Mutations and the Cancer Biomarkers database, which they also maintain, to identify actionable driver variants (Tamborero et al., 2018a). The Personal Cancer Genome Reporter (Nakken et al., 2017) integrates several databases and performs annotations for which VCF files are required as input. The workflow is distributed with Docker technology, a solution that might be very useful for the inclusion in in-house pipelines but requires of programming expertise. Therefore, this solution does not encourage other potential target users such as clinicians or researchers

(3) alpha version available under `https://search.cancervariants.org`

without computational support. The IMPACT pipeline (Hintzsche et al., 2016) and web portal (Hintzsche et al., 2018) have a strong focus on pharmacogenomic (drug-target) interactions. In contrast, the MTB Report focuses on evidence-based actionable variants and provides an expanded catalog by reporting cancer type *repurposing* and low evidence levels (case studies and preclinical evidence). The iMTB-Report has the benefits of an online web page distribution through the institute web page, but can also be run or installed locally with the Shiny app distributed in GitHub.

Interestingly, from those actionable variants missed by the MTB Report compared to the experts' suggestions, *LCK* would have been matched to dasatinib by IMPACT (as well as to other 15 approved drugs and 7 investigational) and by PanDrugs (as well as to other 4 approved drugs, 40 in clinical trials and 73 experimental), as they include target-drug databases. *NTRK3* R116W would have been matched to larotrectinib (resistance) and to novel receptor tyrosine kinase inhibitors by the Cancer Genome Interpreter upon activating aminoacid change repurposing (allowed in MTB Report), and to *IGF1R* inhibitors, *PI3K* inhibitors and midostaurin upon variant type repurposing (not allowed in MTB Report). Yet, these comparisons are time-biased, as all MTB Reports presented here were performed with database versions from 2014 to June 2017, and, for instance, the sources supporting *NTRK3* actionability in SNVs date from the end of 2017 (Drilon et al., 2017). Indeed, this evidence is now included in CIViC and thus, would be now identified by the MTB Report. As for *PTPRJ* and *PTPN12*, which were suggested based on indirect targeting rationale, PanDrugs was not able to identify drugs that indirectly target these genes. In light of these findings, we believe that including drug-target interactions and a pathway visualization could be highly informative for direct and indirect targeting.

8.2 Challenges of Genomics-driven Oncology

Genomics-driven oncology needs the expertise on how to combine the data, knowledge, technical and biological tools that are available. We regard the MTB Report presented here as a small but crucial piece in a larger precision medicine workflow. Such a workflow has to ensure efficient sequencing of patients' samples, accurate bioinformatic processing of the data, clinically meaningful interpretation of the results, decision-making based on the integration of all available data of the patient, pursuing of follow-up and, finally, using this information for interpreting future patients (Hyman et al., 2017). In this section, some relevant challenges for the implementation of genomics-driven oncology are discussed.

8.2.1 Clinical Trials Accessibility

Our initial results demonstrate the potential of WGS/WES to perform genomics-driven cancer treatment and justify the systematic evaluation of the clinical utility of such reporting workflows in larger cohorts of cancer patients. Other studies have forecasted similar numbers of patients with actionable variants (Schwaederle et al., 2015a; Beltran et al., 2015; Van Allen et al., 2014; Rennert et al., 2016; Schuh et al., 2018). Conversely, the actual number of patients treated with genomics-driven therapies is known to be highly depending on the availability of genomically-matched clinical trials. For instance, large clinical trials, such as SHIVA, MD Anderson Cancer Center or NCT-MASTER have been able to assign around 40% of tested patients to matched clinical trials (Le Tourneau et al., 2015a; Meric-Bernstam et al., 2015a; Tsimberidou et al., 2014; Horak et al., 2017). However, other studies reported numbers limited to 5-20%, mainly due to lack of clinical trial access, individual preferences and patient deterioration (Beltran et al., 2015; Sohal et al., 2016; Zehir et al., 2017; Massard et al., 2017).

Several aspects render clinical trial enrollment difficult. Most actionable variants are present in a small proportion of patients, hindering the achievement of large sample sizes. In contrast, for a patient with a rare mutation, it may be geographically difficult to participate in a genomically-matched clinical trial. The first step towards a solution is to accommodate the design of clinical trials to the new precision medicine paradigm. New early phase clinical trials are designed to minimize sample sizes to provide proof-of-concept for larger randomized designs (Dienstmann et al., 2015b). Next, robust computational methods are needed to match patients with rare mutations to genomically-matched clinical trials (Eubank et al., 2016) and harmonize genomic and clinical trial data registries (Siu et al., 2016). Finally, multi-center trials and data sharing across institutions could increase patient recruitment. However, incompatible electronic clinical information systems and acquisition and shipment of biomaterial pose major logistic barriers (Rubin, 2015; Horak et al., 2017).

8.2.2 Precision Medicine Infrastructures and Treatment Algorithms

Institutional bioinformatic pipelines have been developed in the context of precision medicine clinical trials to deal with genomic data analysis, integration with clinical information, interpretation to guide treatment and matching to clinical trial arms. Some of these tests, infrastructures, platforms or algorithms have been published, as EXaCT-1 (Rennert et al., 2016), MSK-

IMPACT (Cheng et al., 2015) and GeneMed (Zhao et al., 2015). GeneMed was developed in order to manage the NCI trials MPACT and MATCH (Coyne et al., 2017) and has also been implemented as a distributable system, OpenGeneMed, which allows for customization of study-specific rules to define actionability (Palmisano et al., 2017). Knowledge and Data Integration (KDI) was developed for SHIVA and RAIDS trials as a data analysis and integration platform (Servant et al., 2014). Any precision medicine infrastructure uses databases for variant annotation and includes rules to match genomic variants to therapies. In this context, tools like the MTB Report could prove useful to standardize this step of data analysis.

An interesting reflection arose from this precision medicine effort, claiming that what genomics-driven trials actually test is the efficiency of the treatment algorithm designed to assign therapies based on molecular data (Le Tourneau et al., 2015b). There is a thin line to decide whether or not a bioinformatic pipeline (or data infrastructure) is a treatment algorithm. A treatment algorithm incorporates expert rules into the NGS bioinformatic pipeline with the final aim of assigning a treatment to a patient. At the same time, it ensures standardization and reproducibility by regulating technical aspects such as minimum coverage, allele frequency, fold change, size of amplicons, prediction scores, etc., making them more suitable for clinical trials. On the other hand, reporting tools such as the MTB Report display all the treatment options based on NGS results and leave the decision to the liable person. There is room for variability between reporting tools with regard to the databases used, the prioritization rules, and the visualization approaches.

8.2.3 Ethical Concerns on Secondary Genomic Findings

As germline DNA is required for a proper identification of somatic variants, this poses an ethical dilemma on whether or not the consequences of germline variants (e.g. predisposition to hereditary diseases) should be tested for and returned to the clinician and the patient (Lolkema et al., 2013). Issues such as privacy, potential benefit to the patient and informed consent should be carefully balanced. The American College of Medical Genetics and Genomics seems to have reached a consensus with regard to a minimum list of genes in which secondary genomic findings should be returned in a clinical setting. The list of 59 medically actionable genes has recently been published (see Table 1 in Kalia et al. (2017)). The list focuses on variants with high penetrance and includes few genes predisposing to hereditary cancers (e.g. *BRCA1/2*, *RET*, *TP53*).

As noticed by Biesecker (2018), many laboratories offering gene panel tests (e.g. gene panels for autism, cardiovascular and cancer diseases) actually perform WES but disclose only a subset of genes relevant to one disease. The author refers to these practices as *exome slices* or *virtual panels* and suggests that, as long as the 59 genes are tested for in the sequencing (as it happens in WES and WGS), the recommendations for secondary findings should apply and mutations in the 59 genes should be disclosed. Likewise, though the focus of the MTB Report is on cancer-related genes, as long as the input would comprise any of the 59 genes, they should be returned. Such a list of genes could easily be included as an appended table in the report.

8.2.4 Further Approaches to Tumor Treatment

Sequencing of cell-free *Circulating Tumor DNA* (ctDNA) –also known as liquid biopsy– is emerging as a less invasive method suitable for diagnostic purposes, for following the course of the disease or the treatment response. It is still under discussion whether this technique is also appropriate for diagnostic purposes and selection of therapy. Though high concordance between plasma-based and tissue-based NGS has been shown for genomic predictive biomarkers (Lebofsky et al., 2015; Kim et al., 2017; Zill et al., 2018), other publications have shown poor concordance (Barata et al., 2017; Kuderer et al., 2017). Most likely, the complexity risen due to the dynamic evolutionary nature of tumors makes very difficult to capture in a snapshot (biopsy) both clonal and subclonal mutations, leading to the observed poor concordance. In a similar fashion, DNA, RNA, proteins and lipids extracted from tumor-derived exosomes or microvesicles can be used to profile tumors in a less-invasive multi-*omics* way (Hoshino et al., 2015; Menck et al., 2017).

Organoids (3D cell cultures, in this case, derived from tumor cells of a patient) emerge as a feasible technique to perform screening of drugs selected based on NGS data in clinical applications. The results of such screening can, in turn, be tested *in vivo* in *Patient-derived Xenografts* (PDX) models. PDX have been used in preclinical studies and are now starting to be employed in observational and some interventional trials as avatars to test drug response *in vivo* (NCT03134456, NCT02720796). However, high technical complexity, low engrafting rates, and high costs will most likely prevent PDX from becoming common in clinical practice. However, neither PDX nor organoids provide an accurate model of immune and vascular microenvironment (Pauli et al., 2017; Dienstmann and Tabernero, 2017).

The interaction of tumor cells with other cell types, such as those present

in the microenvironment and immune infiltrates, have been shown to have important effects on shaping tumor response to therapies (Hanahan and Weinberg, 2011; Prahallad et al., 2012; Blank et al., 2016; Tamborero et al., 2018b). These interactions play an important role in immunotherapies, which have led to durable responses and long-term remissions across tumor types (Le et al., 2015; Sharma and Allison, 2015a). Efforts are focused now on identifying biomarkers for immunotherapies (e.g. mutational load, PD-L1 expression, immune gene signatures, immune infiltrates (Gibney et al., 2016)), elucidation of mechanisms of resistance to immunotherapies and finding the right combination regiments (e.g. blocking two immune checkpoints with PD-1 and CTL4 inhibitors or blocking an immune checkpoint and one genomic vulnerability with CTL4 and BRAF inhibitors) (Sharma et al., 2017). Indeed, combinations of targeted therapies and immune checkpoint therapies are envisioned to have a synergistic effect that would i) improve median OS and ii) increase the number of patients with durable responses (Sharma and Allison, 2015b).

8.3 Perspectives of the Molecular Tumor Board Report and of Genomics-driven Oncology

The current applications of the MTB Report are restricted to genomic alterations that predict response to mono-therapies. As more and more patients are developing drug resistance (Greaves, 2015, 2018), the databases are starting to include genomic alterations that predict response to combination therapies, drug response given a secondary alteration, thus complicating the interpretability of the reports and drug prioritization. Furthermore, databases also include genomic alterations that predict response –and resistance– to immunotherapies. Unfortunately, the degree to which low response rates, development of resistance, interaction with tumor microenvironment and high toxicity of drug combinations will impact the benefit achieved by genomics-driven medicine remains unclear. An important issue, especially with drug combinations, is to ascertain the best combination at the right dose that leads to higher response rates. However, to our best knowledge there is no resource that records drug response rates in patients cohorts matched to genomic profiles.

Although this thesis' focus has been on somatic variants, germline variants can further expand the actionability landscape of cancer patients (besides informing about syndrome predispositions and pharmacodynamics) (Mandelker et al., 2017). As shown by Pritchard et al. (2016), metastatic prostate cancers present actionable germline mutations in *BRCA1*, *BRCA2*

and *ATM* genes; and tumors with mismatch repair deficiencies may benefit from anti-checkpoint blockade immunotherapies (Le et al., 2015). Besides genomic alterations, cancer is also driven by epigenomic regulatory alterations, RNA editing, alternative splicing, post-translational alterations, etc. Thus, multi-*omics* integration will be able to provide more information on a tumor's actionability landscape than a single level (e.g. DNA). Such integration of data originated from different molecular levels requires *systems biology* approaches that take into account prior knowledge to construct complex networks of molecular interactions using mathematical models. However, to begin with, RNA-Seq could prove useful to define pathway activation and target expression. Indeed, mutational status is often not enough to support the use of a drug, and functional data is required to measure the transcription of mutant and wild-type alleles. Furthermore, RNA-Seq is better suited to detect fusion genes than DNA sequencing. A combination of both DNA- and RNA-Seq can provide a comprehensive view suitable for clinical translation (Roychowdhury et al., 2011; Roychowdhury and Chinnaiyan, 2016).

Furthermore, the distribution of tumor mutations presents a long tail of rare mutations that are still incompletely characterized (Chang et al., 2018a). Sharing genomic data will accelerate the identification of rare mutations, thereby expanding the reach of precision oncology in patients with cancer. However, many other factors influence the course of a disease and treatment response. Coupling genomic data to follow-up, treatment, diagnosis, family history, and other kinds of data (diet, environment, etc.) increase the potential of precision medicine (Rubin, 2015; Hyman et al., 2017). The use of medical informatics infrastructure, algorithms and machine learning, will help to find similar patients and design a personalized treatment strategy (Eubank et al., 2016; Shameer et al., 2017). In turn, these predictions can be tested *in vitro* and *in vivo* to select the optimal combinations that will tackle all vulnerabilities (Hyman et al., 2017; Dienstmann and Tabernero, 2017). We envision the MTB Report as one piece of such precision medicine framework that takes advantage of all available information and links it to scientific evidence.

Chapter 9

Conclusions

As NGS technologies are becoming more sensitive and affordable, the knowledge gathered on how specific mutations shape tumor development and treatment response has grown to a large extent. Clinical implications of genomic variants had thus become unmanageable in practical terms. One important branch of precision medicine deals with the use of genomic data to guide cancer treatment. The main contribution of this thesis towards this field has been the development of a workflow aimed at reducing the workload that represents finding the clinical implications of genomic data for a bioinformatician, oncologist or researcher. The MTB Report workflow interrogates somatic variants using public databases of actionable variants. The MTB Report delivers a filtered list of somatic actionable variants in a report with the supporting evidence that indicates the actionability strength and context.

The MTB Report showed high concordance with experts' manual interpretations and identified a comprehensive landscape of actionable variants of a patient's tumor. The proposed approach expands the number of actionable variants by including investigational drugs and applying relaxed repurposing rules. Hence, as it was shown in the analysis of two large public datasets (TCGA and GENIE), a significant fraction of patients can get treatment recommendations guided by genomic data. Of course, all MTB Report findings have to be rated and reevaluated for their therapeutic impact with regard to the clinical history of each patient, as the efficacy of off-label use of targeted drugs is still under discussion.

In the anticipation of a widespread use of clinical sequencing, this study provides a proof of concept for the utility of the MTB Report for the clinical interpretation of genomic data. Despite its limitations, this is one of the first workflows fully available to the research community for the identification of actionable variants. The implementation of the workflow as a web tool

increased end-user feedback and will help in evaluating the extent to which the tool facilitates interpretation of NGS studies.

For a valid interpretation of genomic data with regard to the actionability of mutations, it is necessary to accommodate public databases into the same organizational scheme. The dependence of the workflow on up-to-date knowledge databases is an obvious caveat and appears as a prospective challenge. Therefore, integrative efforts are needed to standardize and maintain the curation of actionable variants' knowledge. Also, further effort should be dedicated to include enrollment suggestions for open genome-guided clinical trials, interpretation of germline variants, prioritization in cases in which many actionable variants co-occur and integration with other *omics* data.

In conclusion, this work demonstrates the potential of combining public resources with a bioinformatic workflow to translate complex genomic profiles into a format suitable for clinical interpretation. Availability of source code and open access databases enables reproducibility, transparency, customization, knowledge sharing.

PART V

APPENDIX

TABLE A.1: **Pathways involved in cancer and drugs which target them.** Information adapted from mycancergenome.org and Iorio et al. (2016).

Pathway	Drug types	Kinase inhibitors	mAbs	Others
Apoptosis	BCL-2 inhibitors			Venetoclax, oblimerson
B-catenin/Wnt signaling	FZD inhibitors		Vantictumab	
Cell cycle control	CDK inhibitors	Alvocidib, roniciclib, dinaciclib, seliciclib		
	CDK4/6 inhibitors	Ribociclib, abemaciclib, palbociclib		
	WEE1 inhibitors	AZD1775, MK1775		
Cellular architecture and microenvironment	PTK2 (FAK)inhibitors	Masitinib, defactinib		
	SRC inhibitors	Saracatinib, ilorasertib, dasatinib		
Chromatin remodeling/DNA methylation	DNMT inhibitors			Guadecitabine
	Histone deactylase			Romidepsin, tucidinostat, panobinostat
DNA damage/repair	PARP inhibitors			Rucaparib, olaparib, talazoparib
Hedgehog signaling	SMO inhibitors			Vismodegib, sonidegib
Hormone signaling	Aromatase inhibitors			Letrozole, exemestane
	HSP90 inhibitors			Retaspimycin, ganetespib
	Hormone (AR, ER) receptor antagonists			Tamoxifen (ER), nilutamide (AR)
Immune checkpoints	Anti-CTLA4 antibodies		Tremelimumab, ipilimumab	
	Anti-PD-1 antibodies		Nivolumab, pidilizumab, pembrolizumab	
	Anti-PD-L1 antibodies		Durvalumab, atezolizumab, avelumab	
	Immunotherapies			
JAK/STAT signaling	JAK inhibitors	Ruxolitinib, lestaurtinib		
Kinase fusions*	ABL inhibitors	Nilotinib, ponatinib, dasatinib, bosutinib, imatinib		
	ROS1 inhibitors	Crizotinib, lorlatinib, entrectinib		
MAP kinase signaling	BRAF inhibitors	Dabrafenib, vemurafenib		
	ERK inhibitors	Ralimetinib		
	MEK inhibitors	Selumetinib, cobimetinib, trametinib		
Metabolic signaling	IDH1 inhibitors			Ivosidenib, enasidenib
PI3K/AKT1/MTOR	mTOR inhibitors	Sirolimus, temsirolimus, everolimus		
	(pan) PI3K inhibitors	Buparlisib, gedatolisib		
	AKT inhibitors	Cenisertib, ipatasertib		
Protein degradation/ubiquitination	Proteaseome inhibitors			Bortezomib, ixazomib, carfilzomib
Receptor tyrosine kinase/growth factor signaling	EGFR inhibitors	Osimertinib, afatinib, erlotinib, gefitinib	Panitumumab, necitumumab, cetuximab	
	ERBB2 inhibitors	Afatinib, lapatinib, neratinib	Pertuzumab, trastuzumab	
	PDGFR inhibitors	Pazopanib, imatinib, sorafenib, regorafenib, sunitinib	Olaratumab, tovetumab	
	KIT inhibitors	Regorafenib, dasatinib, pazopanib, imatinib, sunitinib		
	FLT3 inhibitors	Ponatinib, sorafenib, lestaurtinib, sunitinib, cabozantinib		
	MET inhibitors	Crizotinib, cabozantinib	Telisotuzumab, emibetuzumab, onartuzumab	
	ALK inhibitors	Brigatinib, alectinib, ceritinib, crizotinib		
	VEGFR inhibitors	Axitinib, regorafenib, nintedanib, pazopanib, ponatinib	Ramucirumab	
	RET inhibitors	Vandetanib, sorafenib, regorafenib, sunitinib, cabozantinib		
	NTRK inhibitors	Sitravatinib, entrectinib, lestaurtinib		
	FGFR inhibitors	Ponatinib, nintedanib		
	IGFR inhibitors	Linsitinib	Ganitumab, figitumumab, cixutumumab	
TGF signaling	TGFBR1 inhibitors	Galunisertib		

TABLE A.2: **Companion diagnostics in oncology approved by FDA, July 2018.** Table modified from https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm. Abbreviations: expr. (=expression) indicates that the biomarker is not measured at DNA level; (-) indicates that biomarker negative status is used for that indication; HR, hormone receptor including ESR (estrogen receptor) and PGR (progesterone receptor); APL, acute prolymphocytic leukemia; HES, Hypereosinophilic syndrome; CEL, chronic eosinophilic leukemia; MDS, myelodysplastic syndrome; MPD, myeloproliferative disorders.

Cancer type	Biomarker	Biomarker type	Drugs
ALL	BCR-ABL1 (-)	biomarker	Blinatumomab
	BCR-ABL1	target	Dasatinib; Imatinib; Ponatinib
AML	FLT3	target	Midostaurin
	IDH2	target	Enasidenib
APL	PML-RARA	biomarker	Arsenic Trioxide; Tretinoin
Breast	BRCA	biomarker	Olaparib
	ERBB2 (-)	biomarker	Abemaciclib; Everolimus; Fulvestrant; Palbociclib; Ribociclib
	ESR (expr.)	biomarker	Abemaciclib; Everolimus; Palbociclib
	HR (expr.)	biomarker	Lapatinib; Ribociclib; Tamoxifen
	HR (expr.)	(indirect) target	Anastrozole; Exemestane; Letrozole
	ERBB2	target	Ado-Trastuzumab Emtansine; Lapatinib; Neratinib; Pertuzumab; Trastuzumab
	ESR, PGR (expr.)	target	Fulvestrant
Cancer	MSI, MMR	biomarker	Nivolumab; Pembrolizumab
Cervical	PD-L1 (expr.)	(indirect) target	Pembrolizumab
CLL	Chr. 17p	biomarker	Ibrutinib
	BCR-ABL1	target	Ponatinib
	CD20 (expr.)	target	Rituximab
CML	BCR-ABL1	target	Bosutinib; Dasatinib; Imatinib; Nilotinib
Colorectal	RAS (-)	biomarker	Cetuximab; Panitumumab
	EGFR (expr.)	target	Cetuximab
Cutaneous T - cell lymphoma	IL2RA (expr.)	target	Denileukin Diftitox
Gastric adenoc.	ERBB2	target	Trastuzumab
GIST	KIT (expr.)	target	Imatinib
HES/CEL	FIP1L1-PDGFRA	target	Imatinib
MDS/MPD	PDGFRB	target	Imatinib
Melanoma	BRAF (-)	biomarker	Nivolumab
	BRAF	target	Cobimetinib; Trametinib; Vemurafenib; encorafenib; binimetinib; dabrafenib
Non-Hodgkin lymphoma	MS4A1 (CD20) (expr.)	target	Rituximab
NSCLC	PD-L1 (expr.)	(indirect) target	Pembrolizumab
	ALK	target	Alectinib; Brigatinib; Ceritinib; Crizotinib
	BRAF	target	Dabrafenib; Trametinib
	EGFR	target	Afatinib; Erlotinib; Gefitinib; Osimertinib
	ROS1	target	Crizotinib
Ovarian	BRCA	biomarker	Olaparib; Rucaparib
Thyroid cancer	BRAF	target	Dabrafenib; Trametinib
Urothelial carc.	PD-L1 (expr.)	(indirect) target	Pembrolizumab

TABLE A.3: **MAF variant types.** List of possible variant categories in MAF file format and their description. Information taken from `https://gatkforums.broadinstitute.org/gatk/discussion/comment/35514`

MAF Variant type	Description
Intron	variant lies between exons within the bounds of the chosen transcript
5'UTR	variant is on the 5'UTR for the chosen transcript
3'UTR	variant is on the 3'UTR for the chosen transcript
IGR	intergenic region. Does not overlap any transcript
5'Flank	the variant is upstream of the chosen transcript (within 3kb)
3'Flank	the variant is downstream of the chosen transcript (within 3kb)
Missense_Mutation	the point mutation alters the protein structure by one amino acid
Nonsense_Mutation	a premature stop codon is created by the variant
Nonstop_Mutation	variant removes stop codon
Silent	variant is in coding region of the chosen transcript, but protein structure is identical. I.e. a synonymous mutation
Splice_Site	the variant is within two bases of a splice site. See the secondary classification to determine if it lies on the exon or intron side.
In_Frame_Del	deletion that keeps the sequence in frame
In_Frame_Ins	insertion that keeps the sequence in frame
Frame_Shift_Ins	insertion that moves the coding sequence out of frame
Frame_Shift_Del	deletion that moves the coding sequence out of frame
Start_Codon_SNP	point mutation that overlaps the start codon.
Start_Codon_Ins	insertion that overlaps the start codon.
Start_Codon_Del	seletion that overlaps the start codon.
De_novo_Start_In Frame	New start codon is created by the given variant using the chosen transcript. However, it is in frame relative to the coded protein.
De_novo_Start_Out OfFrame	New start codon is created by the given variant using the chosen transcript. However, it is out of frame relative to the coded protein.
RNA	variant lies on one of the RNA transcripts.
lincRNA	variant lies on one of the lincRNAs.
Translation_Start_Site	initiator_codon_variant, start_lost

TABLE A.4: **List of actionable genes included in each database**. Acronyms are used to indicate each database: G, GDKD; C, CIViC; T, TARGET; M-B, Meric-Bernstam et al. (2015b). Genes with * were added upon expert suggestion.

Gene	G	C	T	M-B.
ABCB1		X		
ABCC10		X		
ABCC3		X		
ABL1	X	X	X	X
AGR2		X		
AKT1	X	X	X	X
AKT2	X	X	X	X
AKT3	X	X	X	X
ALCAM		X		
ALDH1A2		X		
ALK	X	X	X	X
APC	X	X	X	
AR	X	X	X	X
AR-V7	X			
ARAF	X	X	X	X
AREG	X	X		
ARID1A	X			
ASNS		X		
ATM	X	X	X	X
ATR	X	X	X	X
ATRX		X		
AURKA	X	X	X	X
B2M	X	X		
BAP1	X	X	X	X
BCL2	X	X	X	X
BCOR	X			
BCR	X	X	X	X
BIRC5		X		
BIRC7		X		
BRAF	X	X	X	X
BRCA1	X	X	X	X
BRCA2	X	X	X	X
BRD2			X	
BRD3			X	
BRD4	X	X	X	
BTK	X	X		
c15orf55			X	
CALR		X		
CASP8		X		
CBL	X	X		
CBLC		X		
CCND1	X	X	X	X
CCND2	X		X	X
CCND3	X	X	X	X
CCNE1	X	X	X	X
CD274	X	X		
CD44		X		
CDH1	X			
CDK12	X	X	X	
CDK4	X	X	X	X
CDK6	X	X	X	X
CDKN1A	X	X	X	
CDKN1B	X	X	X	X
CDKN2A	X	X	X	X
CDKN2B	X	X	X	X
CDKN2C	X			X
CEBPA		X	X	
CFLAR		X		
CHEK2	X			X
COL1A1	X			
CRKL			X	
CRLF2	X			
CSF1R	X	X		X
CSF3R	X			
CTLA4		X		
CTNNB1	X	X	X	
CXCR4		X		
DDR2	X	X	X	X
DDX43		X		
DEFA1		X		
DKK1		X		
DNMT1		X		
DNMT3A	X	X	X	X
DPYD		X		
DUSP6		X		
ECSCR		X		
EGF	X	X	X	
EGFR	X	X	X	X
EIF4EBP1		X		
EPAS1		X		
EPHA2	X			
EPHA3	X		X	X

Continued...

Continued...

Gene	G	C	T	M-B.
EPHB4		X		
ERBB2	X	X	X	X
ERBB3	X	X	X	X
ERBB4	X	X	X	X
ERCC1	X	X		
ERCC2	X	X	X	
ERCC4	X			
ERCC6	X			
EREG	X	X		
ERG			X	
ERRFI1		X	X	
ESR1	X	X	X	X
ETS2		X		
EZH2	X	X	X	X
FANCA	X			
FANCC	X	X		
FAT1	X			
FBXW7	X	X	X	
FCGR3A		X		
FGF2		X		X
FGF3	X	X		X
FGF4	X			X
FGFR1	X	X	X	X
FGFR2	X	X	X	X
FGFR3	X	X	X	X
FGFR4	X			X
FLCN	X		X	
FLT3	X	X	X	X
FNTB		X		
FOS		X		
FOXA1	X			
FOXP3		X		
FRS2	X			
GAS6		X		
GATA2		X		
GATA3	X			
GNA11	X	X	X	X
GNAQ	X	X	X	X
GNAS	X	X	X	
GSTP1		X		
HAVCR2		X		
HDAC2	X			
HGF	X	X		X
HIF1A		X		

Continued...

Gene	G	C	T	M-B.
HLA-C		X		
HLA-DRA		X		
HMOX1		X		
HRAS	X	X	X	X
HSPA5		X		
HSPB1		X		
HSPH1		X		
IDH1	X	X	X	X
IDH2	X		X	X
IGF1R	X	X	X	X
IGF2	X	X		X
IL7R	X			
INI1	X			
INPP4B	X			
JAK1	X	X		X
JAK2	X	X	X	X
JAK3	X		X	X
JUN		X		
KDR	X	X	X	X
KIAA1524		X		
KIT	X	X	X	X
KRAS	X	X	X	X
LRP1B	X	X		
MAGEH1		X		
MAP2K1	X	X	X	X
MAP2K2	X		X	X
MAPK1		X	X	X
MAPK3			X	
MCL1	X		X	
MDM2	X	X	X	X
MDM4	X		X	
MED12			X	
MERTK		X		
MET	X	X	X	X
MGMT	X	X		
MITF	X		X	
MLH1		X		
MLL	X		X	X
MLL2	X			
MMP2		X		
MMP9		X		
MPL	X		X	X
MRE11		X		
MSH2		X		

Continued...

Gene	G	C	T	M-B.
MSH3	X			
MSH6		X		
MTAP		X		
MTHFR		X		
MTOR	X	X	X	X
MYC	X	X		
MYCL		X		
MYCN		X		X
MYD88	X	X	X	
NAPRT		X		
NF1	X	X	X	X
NF2	X	X	X	X
NOTCH1	X	X	X	X
NOTCH2	X		X	X
NPM1	X	X	X	X
NQO1		X		
NRAS	X	X	X	X
NRG1	X	X		
NT5C2		X		
NT5E		X		
NTRK1	X	X		X
NTRK3	X	X	X	X
PAK1	X			
PALB2	X	X		X
PBK		X		
PBRM1	X			
PDCD4		X		
PDGFRA	X	X	X	X
PDGFRB	X	X	X	X
PDPK1	X			
PGR		X		
PIK3CA	X	X	X	X
PIK3CB	X		X	X
PIK3R1	X	X	X	X
PIK3R2	X			X
PLCG2	X			
PML	X	X		X
PMS2		X		
POLE	X	X		
POU5F1		X		
PPP1R15A		X		
PRKAA2		X		
PRKCH	X			
PROM1		X		

Continued...

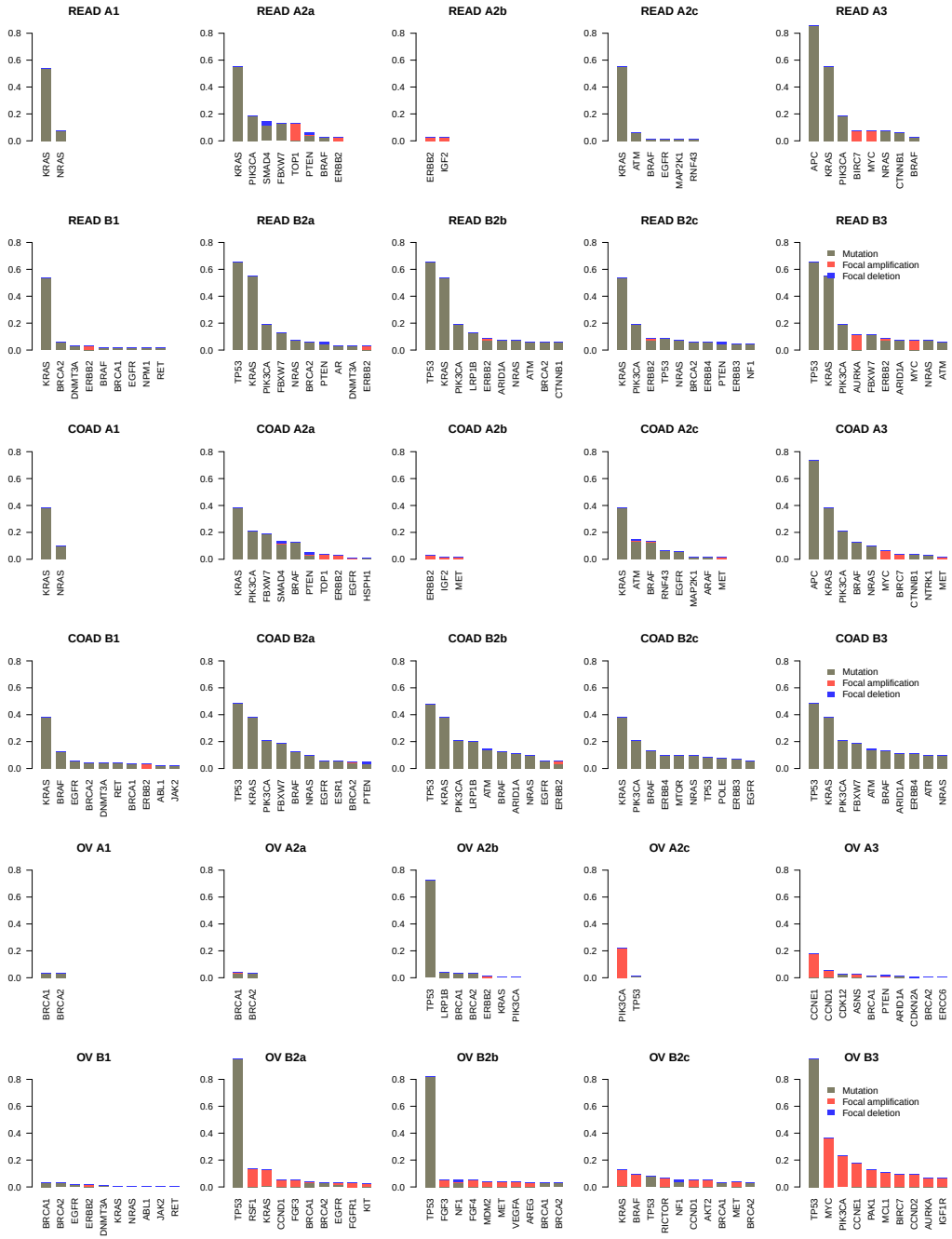
Gene	G	C	T	M-B.
PTC1	X			
PTCH1	X	X	X	X
PTEN	X	X	X	X
PTGS2		X		
PTP4A3		X		
PTPRB		X		
PTPRD	X	X		
PTPRT		X		
RAB35			X	
RAC1	X	X		
RAD23B		X		
RAD50	X			X
RAD51	X			
RAD51C	X			
RAF1	X	X	X	X
RARA	X		X	X
RB1	X	X	X	
RET	X	X	X	X
RHEB			X	
RICTOR	X	X		X
RIT1		X		
RNF43	X		X	
ROBO4		X		
ROS1	X	X	X	X
RPS6		X		
RRM1		X		
RRM2		X		
RSF1		X		
RSPO2			X	
RUNX1		X		X
SERPINB3	X			
SETD2	X			
SF3B1	X	X		
SGK1		X		
SH2B3	X			
SIRT1		X		
SLCO1B1		X		
SLFN11		X		
SMAD4		X		
SMARCA1	X			
SMARCA4	X	X	X	
SMARCB1	X	X	X	
SMO	X	X	X	X
SOCS1	X			

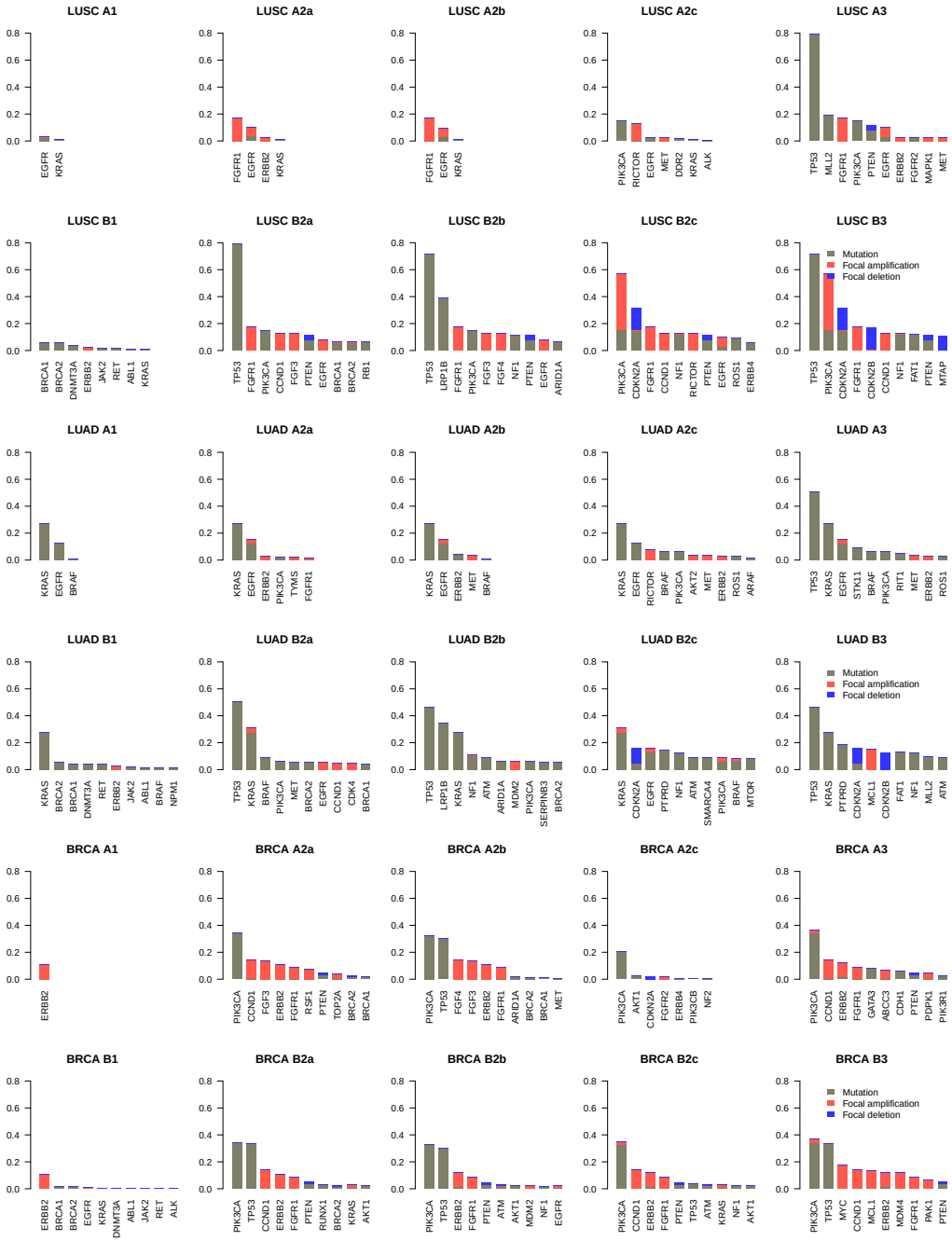
Continued...

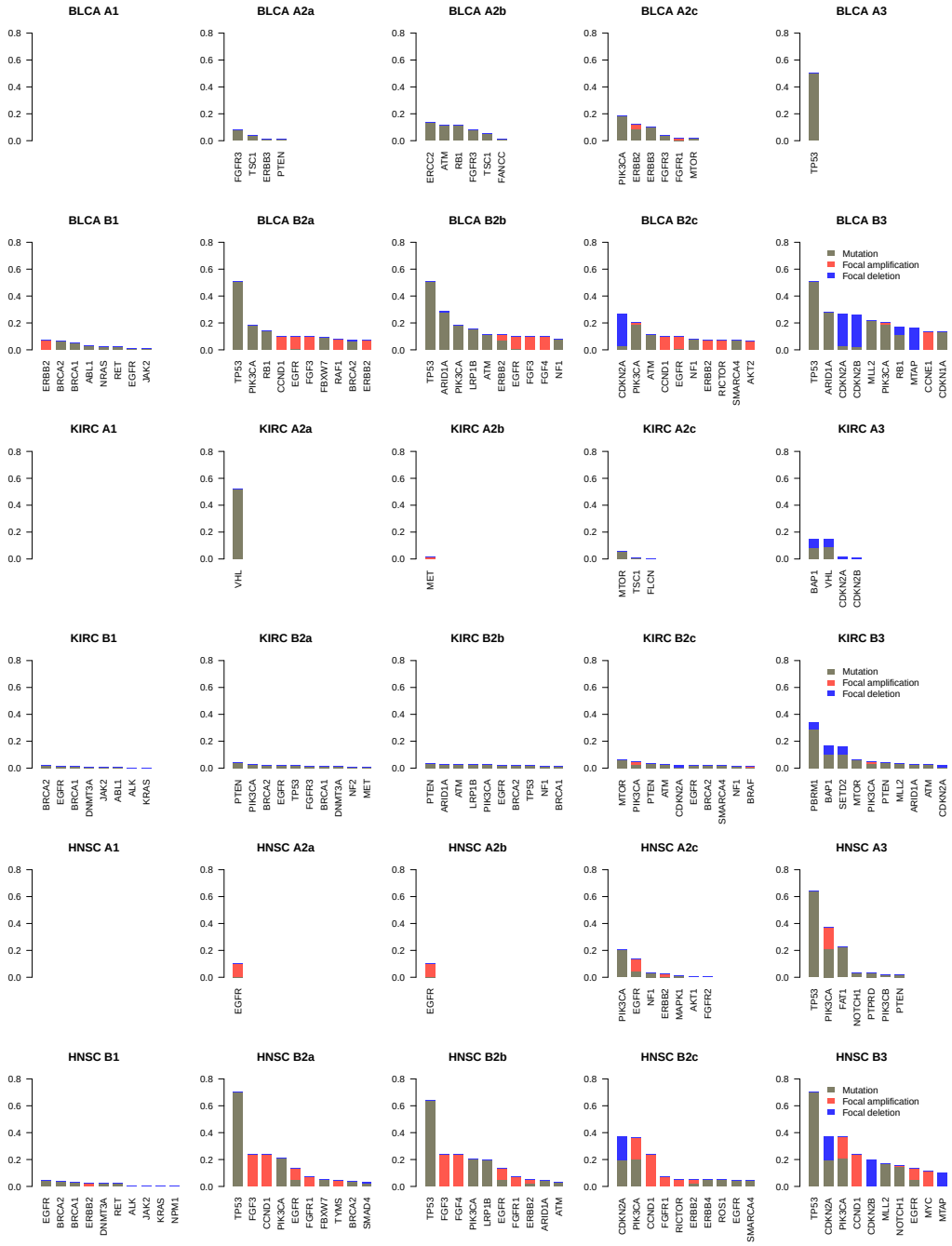
Gene	G	C	T	M-B.
SOX10		X		
SRSF2	X			
STAG2	X	X		
STAG3		X		
STK11	X	X	X	X
STMN1		X		
SUZ12	X			
SYK	X	X	X	X
TBK1		X		
TET2			X	X
TFF3		X		
TIMP1		X		
TMPRSS2	X		X	
TOP1		X		
TOP2A	X	X		X
TP53	X	X	X	
TSC1	X	X	X	X
TSC2	X	X	X	X
TUBB3		X		
TYMS		X		
U2AF1	X			
UGT1A		X		
UGT1A1		X		
VEGFA	X	X		
VHL		X		
WEE1		X		
WT1		X		
XPO1			X	
XRCC1		X		
ZEB1		X		
ZNRF3	X		X	
ABL2				X
AURKB				X
AURKC				X
CBFB				X
DDR1				X
DOT1L				X
FLT1				X
FLT4				X
HDAC9				X
KMT2A				X
MAP2K4				X
MAP3K1				X
MAP3K4				X
MAPK8				X
NOTCH3				X
NOTCH4				X
PDGFB				X
PIK3CD				X
PTPN11				X
SRC				X
FGF1				X*
EGF1				X*

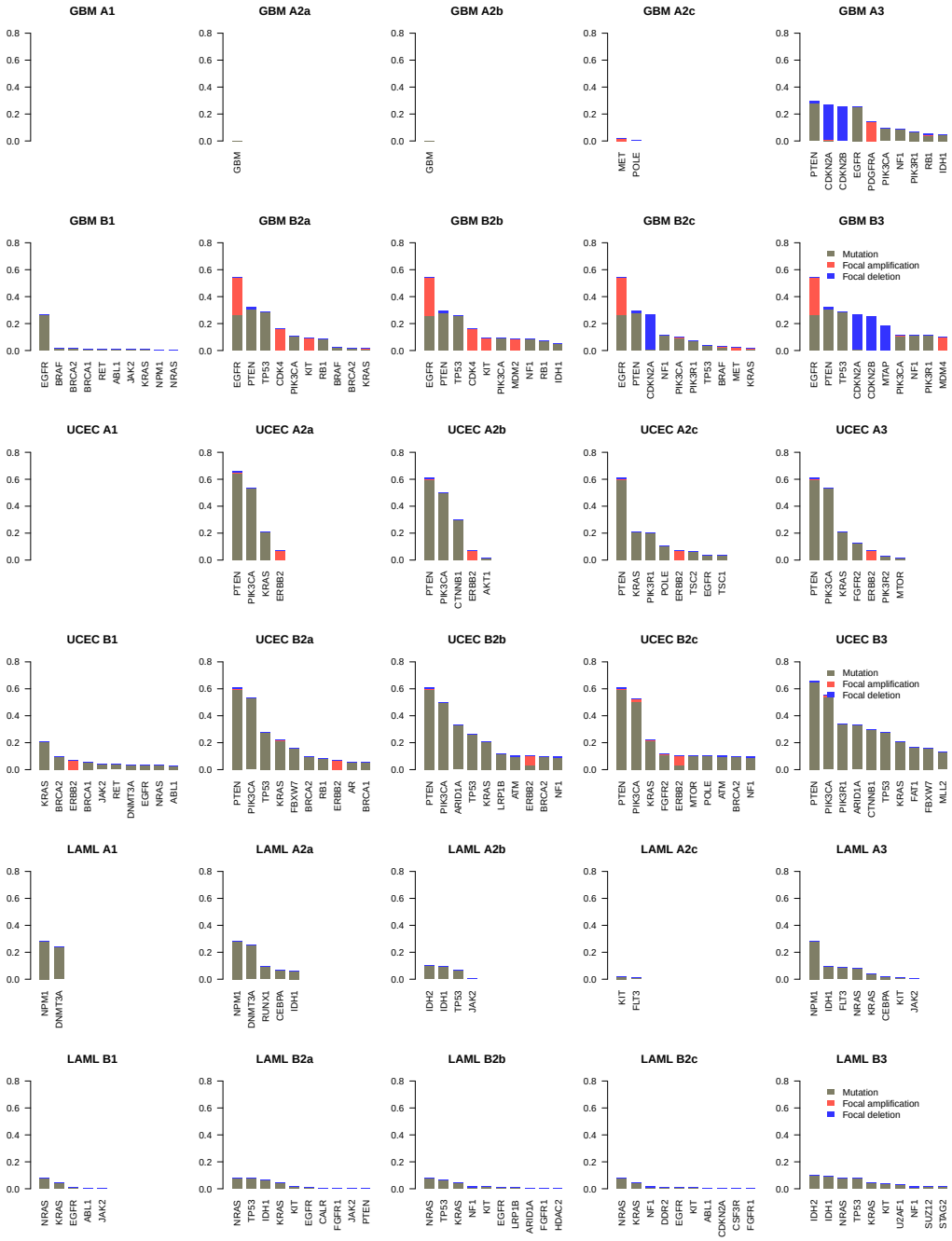
Continued...

FIGURE A.1: Percentage of patients per gene showing top ten genes at each level of evidence in each tumor type. Level 2 is divided into three further groups: 2a (late clinical trials), 2b (early clinical trials), and 2c (case reports).




LUSC A1
LUSC A2a
LUSC A2b
LUSC A2c
LUSC A3
LUSC B1
LUSC B2a
LUSC B2b
LUSC B2c
LUSC B3
LUAD A1
LUAD A2a
LUAD A2b
LUAD A2c
LUAD A3
LUAD B1
LUAD B2a
LUAD B2b
LUAD B2c
LUAD B3
BRCA A1
BRCA A2a
BRCA A2b
BRCA A2c
BRCA A3
BRCA B1
BRCA B2a
BRCA B2b
BRCA B2c
BRCA B3
Mutation
Focal amplification
Focal deletion


BLCA A1
BLCA A2a
FGFR3 TSC1 ERBB3 PTEN
BLCA A2b
ERCC2 ATM RB1 FGFR3 TSC1 PANCC
BLCA A2c
PIK3CA ERBB2 ERBB3 FGFR3 FGFR1 MTOR
BLCA A3
TP53
BLCA B1
ERBB2 BRCA2 BRCA1 ABL1 NRAS RET EGFR JAK2
BLCA B2a
TP53 PIK3CA RB1 CCND1 EGFR FGF3 FBXW7 RAF1 BRCA2 ERBB2
BLCA B2b
TP53 ARID1A PIK3CA LRP1B ATM ERBB2 EGFR FGF3 FGF4 NF1
BLCA B2c
CDKN2A PIK3CA ATM CCND1 EGFR NF1 ERBB2 RICTOR SMARCA4 AKT2
BLCA B3
Mutation
Focal amplification
Focal deletion
TP53 ARID1A CDKN2A CDKN2B MLL2 PIK3CA RB1 MTAP CCNE1 CDKN1A
KIRC A1
KIRC A2a
VHL
KIRC A2b
MET
KIRC A2c
MTOR TSC1 FLCN
KIRC A3
BAP1 VHL CDKN2A CDKN2B
KIRC B1
BRCA2 EGFR BRCA1 DNMT3A JAK2 ABL1 ALK KRAS
KIRC B2a
PTEN PIK3CA BRCA2 EGFR TP53 FGFR3 BRCA1 DNMT3A NF2 MET
KIRC B2b
PTEN ARID1A ATM LRP1B PIK3CA EGFR BRCA2 TP53 NF1 BRCA1
KIRC B2c
MTOR PIK3CA PTEN ATM CDKN2A EGFR BRCA2 SMARCA4 NF1 BRAF
KIRC B3
Mutation
Focal amplification
Focal deletion
PBRM1 BAP1 SETD2 MTOR PIK3CA PTEN MLL2 ARID1A ATM CDKN2A
HNSC A1
HNSC A2a
EGFR
HNSC A2b
EGFR
HNSC A2c
PIK3CA EGFR NF1 ERBB2 MAPK1 AKT1 FGFR2
HNSC A3
TP53 PIK3CA FAT1 NOTCH1 PTPRD PIK3CB PTEN
HNSC B1
EGFR BRCA2 BRCA1 ERBB2 DNMT3A RET ALK JAK2 KRAS NPM1
HNSC B2a
TP53 FGF3 CCND1 PIK3CA EGFR FGFR1 FBXW7 TYM5 BRCA2 SMAD4
HNSC B2b
TP53 FGF3 FGF4 PIK3CA LRP1B EGFR FGFR1 ERBB2 ARID1A ATM
HNSC B2c
CDKN2A PIK3CA CCND1 FGFR1 RICTOR ERBB2 ERBB4 ROS1 EGFR SMARCA4
HNSC B3
Mutation
Focal amplification
Focal deletion
TP53 CDKN2A PIK3CA CCND1 CDKN2B MLL2 NOTCH1 EGFR MYC MTAP


GBM A1
GBM A2a
GBM A2b
GBM A2c
GBM A3
GBM B1
GBM B2a
GBM B2b
GBM B2c
GBM B3
UCEC A1
UCEC A2a
UCEC A2b
UCEC A2c
UCEC A3
UCEC B1
UCEC B2a
UCEC B2b
UCEC B2c
UCEC B3
LAML A1
LAML A2a
LAML A2b
LAML A2c
LAML A3
LAML B1
LAML B2a
LAML B2b
LAML B2c
LAML B3
Mutation
Focal amplification
Focal deletion

References

Ahmed, J., Meinel, T., Dunkel, M. et al (2011). CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res*, 39(Database issue):D960–D967.

Ainscough, B.J., Griffith, M., Coffman, A.C. et al (2016). DoCM: a database of curated mutations in cancer. *Nat Methods*, 13(10):806–807.

Alioto, T.S., Buchhalter, I., Derdak, S. et al (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, 6:10001.

Alizadeh, A.A., Eisen, M.B., Davis, R.E. et al (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503.

André, F., Bachelot, T., Commo, F. et al (2014). Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). *The Lancet Oncology*, 15(3):267–274.

Bailey, P., Chang, D.K., Nones, K. et al (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

Bao, R., Huang, L., Andrade, J. et al (2014). Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Cancer Inform*, 13(Suppl 2):67–82.

Barata, P.C., Koshkin, V.S., Funchain, P. et al (2017). Next-generation sequencing (NGS) of cell-free circulating tumor DNA and tumor tissue in patients with advanced urothelial cancer: a pilot assessment of concordance. *Ann Oncol*, 28(10):2458–2463.

Barton, S., Hawkes, E.A., Wotherspoon, A. et al (2012). Are we ready to stratify treatment for diffuse large B-cell lymphoma using molecular hallmarks? *The oncologist*, pages theoncologist–2012.

Beltran, H., Eng, K., Mosquera, J.M. et al (2015). Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncol*, 1(4):466–474.

Biesecker, L.G. (2018). Secondary findings in exome slices, virtual panels, and anticipatory sequencing.

Blank, C.U., Haanen, J.B., Ribas, A. et al (2016). The "cancer immunogram". *Science*, 352(6286):658–660.

Bowdin, S., Gilbert, A., Bedoukian, E. et al (2016). Recommendations for the integration of genomics into clinical practice. *Genet Med*, 18(11):1075–1084.

Bramsen, J.B., Rasmussen, M.H., Ongen, H. et al (2017). Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell Reports*, 19(6):1268–1280.

Cancer Network (2013). FDA Approves Erlotinib (Tarceva) as First-Line Lung Cancer Therapy for Certain Patients: http://www.cancernetwork.com/lung-cancer/fda-approves-erlotinib-tarceva-first-line-lung-cancer-therapy-certain-patients.

Chakravarty, D., Gao, J., Phillips, S. et al (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16.

Chang, M.T., Bhattarai, T.S., Schram, A.M. et al (2018a). Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov*, 8(2):174–183.

Chang, W., Cheng, J., Allaire, J.J. et al (2018b). *shiny: Web Application Framework for R*.

Chapman, P.B., Hauschild, A., Robert, C. et al (2011). Improved Survival with Vemurafenib in Melanoma with BRAF V600e Mutation. *New England Journal of Medicine*, 364(26):2507–2516.

Chau, N.G., Li, Y.Y., Jo, V.Y. et al (2016). Incorporation of Next-Generation Sequencing into Routine Clinical Care to Direct Treatment of Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.*, 22(12):2939–2949.

Chen, X., Ji, Z.L. and Chen, Y.Z. (2002). TTD: Therapeutic Target Database. *Nucleic Acids Res*, 30(1):412–415.

Cheng, D.T., Mitchell, T.N., Zehir, A. et al (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *J Mol Diagn*, 17(3):251–264.

Chin, L., Andersen, J.N. and Futreal, P.A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 17(3):297–303.

Cibulskis, K., Lawrence, M.S., Carter, S.L. et al (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213–219.

Ciriello, G., Miller, M.L., Aksoy, B.A. et al (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*, 45(10):1127–1133.

Conti, R.M., Bernstein, A.C., Villaflor, V.M. et al (2013). Prevalence of Off-Label Use and Spending in 2010 Among Patent-Protected Chemotherapies in a Population-Based Cohort of Medical Oncologists. *J Clin Oncol*, 31(9):1134–1139.

Cooper, G.M., Zerr, T., Kidd, J.M. et al (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics*, 40(10):1199–1203.

Corsello, S.M., Bittker, J.A., Liu, Z. et al (2017). The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med*, 23(4):405–408.

Coyne, G.O., Takebe, N. and Chen, A.P. (2017). Defining precision: The precision medicine initiative trials NCI-MPACT and NCI-MATCH. *Curr Probl Cancer*, 41(3):182–193.

Damodaran, S., Miya, J., Kautto, E. et al (2015). Cancer Driver Log (CanDL). *J Mol Diagn*, 17(5):554–559.

Deng, M., Brägelmann, J., Kryukov, I. et al (2017). FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database (Oxford)*, 2017.

Deppert, W. (2007). Mutant p53: from guardian to fallen angel?

Dienstmann, R., Dong, F., Borger, D. et al (2014). Standardized decision support in next generation sequencing reports of somatic cancer variants. *Mol. Oncol.*, 8(5):859–73.

Dienstmann, R., Jang, I.S., Bot, B. et al (2015a). Database of Genomic Biomarkers for Cancer Drugs and Clinical Targetability in Solid Tumors. *Cancer Discov.*, 5(2):118–123.

Dienstmann, R., Rodon, J. and Tabernero, J. (2015b). Optimal design of trials to demonstrate the utility of genomically-guided therapy: Putting Precision Cancer Medicine to the test. *Molecular oncology*, 9(5):940–950.

Dienstmann, R. and Tabernero, J. (2017). Cancer: A precision approach to tumour treatment. *Nature*, 548(7665):40–41.

Drilon, A., Laetsch, T.W., Kummar, S. et al (2018). Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children. *N. Engl. J. Med.*, 378(8):731–739.

Drilon, A., Nagasubramanian, R., Blake, J.F. et al (2017). A Next-Generation TRK Kinase Inhibitor Overcomes Acquired Resistance to Prior TRK Kinase Inhibition in Patients with TRK Fusion-Positive Solid Tumors. *Cancer Discov*, 7(9):963–972.

EU-CTR (2018). Clinical Trials Register: https://www.clinicaltrialsregister.eu/ctr-search/search.

Eubank, M.H., Hyman, D.M., Kanakamedala, A.D. et al (2016). Automated eligibility screening and monitoring for genotype-driven precision oncology trials. *J Am Med Inform Assoc*, 23(4):777–781.

Farmer, H., McCabe, N., Lord, C.J. et al (2005). Targeting the DNA repair defect in *BRCA* mutant cells as a therapeutic strategy. *Nature*, 434(7035):917–921.

Fong, P.C., Yap, T.A., Boss, D.S. et al (2010). Poly(ADP)-Ribose Polymerase Inhibition: Frequent Durable Responses in BRCA Carrier Ovarian Cancer Correlating With Platinum-Free Interval. *JCO*, 28(15):2512–2519.

Forbes, S.A., Beare, D., Boutselakis, H. et al (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*, 45(Database issue):D777–D783.

Foulkes, W.D. (2014). BRCA1 and BRCA2 – update and implications on the genetics of breast cancer: a clinical perspective. *Clinical Genetics*, 85(1):1–4.

Futreal, P.A., Coin, L., Marshall, M. et al (2004). A CENSUS OF HUMAN CANCER GENES. *Nat Rev Cancer*, 4(3):177–183.

Garraway, L.A., Verweij, J. and Ballman, K.V. (2013). Precision Oncology: An Overview. *JCO*, 31(15):1803–1805.

Gartside, M.G., Chen, H., Ibrahimi, O.A. et al (2009). Loss-of-Function Fibroblast Growth Factor Receptor-2 Mutations in Melanoma. *Mol Cancer Res*, 7(1):41–54.

Gaulton, A., Bellis, L.J., Bento, A.P. et al (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 40(Database issue):D1100–D1107.

Gibney, G.T., Weiner, L.M. and Atkins, M.B. (2016). Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *The Lancet Oncology*, 17(12):e542–e551.

Gohlke, B.O., Nickel, J., Otto, R. et al (2016). CancerResource—updated database of cancer-relevant proteins, mutations and interacting drugs. *Nucleic Acids Res*, 44(D1):D932–D937.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J. et al (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082.

Good, B.M., Ainscough, B.J., McMichael, J.F. et al (2014). Organizing knowledge to enable personalization of medicine in cancer. *Genome Biol.*, 15(8):438.

Greaves, M. (2015). Evolutionary Determinants of Cancer. *Cancer Discov*, 5(8):806–820.

Greaves, M. (2018). Nothing in cancer makes sense except.... *BMC Biology*, 16:22.

Greaves, M. and Maley, C.C. (2012). CLONAL EVOLUTION IN CANCER. *Nature*, 481(7381):306–313.

Griffith, M., Griffith, O.L., Coffman, A.C. et al (2013). DGIdb: mining the druggable genome. *Nature Methods*, 10(12):1209–1210.

Griffith, M., Spies, N.C., Krysiak, K. et al (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*, 49(2):170–174.

Guinney, J., Dienstmann, R., Wang, X. et al (2015). The Consensus Molecular Subtypes of Colorectal Cancer. *Nat Med*, 21(11):1350–1356.

Guo, X.E., Ngo, B., Modrek, A.S. et al (2014). Targeting Tumor Suppressor Networks for Cancer Therapeutics. *Curr Drug Targets*, 15(1):2–16.

Hanahan, D. and Weinberg, R. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

Hanahan, D. and Weinberg, R.A. (2000). The Hallmarks of Cancer. *Cell*, 100(1):57–70.

Heinrich, M.C., Corless, C.L., Demetri, G.D. et al (2003). Kinase Mutations and Imatinib Response in Patients With Metastatic Gastrointestinal Stromal Tumor. *JCO*, 21(23):4342–4349.

Hintzsche, J., Kim, J., Yadav, V. et al (2016). IMPACT: a whole-exome sequencing analysis pipeline for integrating molecular profiles with actionable therapeutics in clinical samples. *Journal of the American Medical Informatics Association*, page ocw022.

Hintzsche, J.D., Yoo, M., Kim, J. et al (2018). IMPACT web portal: oncology database integrating molecular profiles with actionable therapeutics. *BMC Medical Genomics*, 11(2):26.

Hoadley, K.A., Yau, C., Hinoue, T. et al (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6.

Hoadley, K.A., Yau, C., Wolf, D.M. et al (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, 158(4):929–944.

Horak, P., Klink, B., Heining, C. et al (2017). Precision oncology based on omics data: The NCT Heidelberg experience. *Int. J. Cancer*, 141(5):877–886.

Hoshino, A., Costa-Silva, B., Shen, T.L. et al (2015). Tumour exosome integrins determine organotropic metastasis. *Nature*, 527(7578):329–335.

Huang, L., Fernandes, H., Zia, H. et al (2017). The cancer precision medicine knowledge base for structured clinical-grade mutations and interpretations. *J Am Med Inform Assoc*, 24(3):513–519.

Hyman, D.M., Puzanov, I., Subbiah, V. et al (2015). Vemurafenib in Multiple Non-melanoma Cancers with BRAF V600 Mutations. *N Engl J Med*, 373(8):726–736.

Hyman, D.M., Taylor, B.S. and Baselga, J. (2017). Implementing Genome-Driven Oncology. *Cell*, 168(4):584–599.

Illumina (2017). Scalable Nucleic Acid Quality Assessments for Illumina Next-Generation Sequencing Library Prep (accessed 12.9.2018). page 4.

International HapMap Consortium (2003). The international HapMap project. *Nature*, 426(6968):789.

International Human Genome Sequencing Consortium, . (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Iorio, F., Knijnenburg, T.A., Vis, D.J. et al (2016). A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 166(3):740–754.

Jabbour, E., Kantarjian, H., Jones, D. et al (2006). Frequency and clinical significance of BCR-ABL mutations in patients with chronic myeloid leukemia treated with imatinib mesylate. *Leukemia*, 20(10):1767–1773.

Johnson, A., Zeng, J., Bailey, A.M. et al (2015). The right drugs at the right time for the right patient: the MD Anderson precision oncology decision support platform. *Drug Discovery Today*, 20(12):1433–1438.

Jones, D.T.W., Hutter, B., Jäger, N. et al (2013). Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma. *Nat Genet*, 45(8):927–932.

Jones, D.T.W., Jäger, N., Kool, M. et al (2012). Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–105.

Kalia, S.S., Adelman, K., Bale, S.J. et al (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.*, 19(2):249–255.

Kandoth, C., McLellan, M.D., Vandin, F. et al (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339.

Kazandjian, D., Blumenthal, G.M., Yuan, W. et al (2016). FDA Approval of Gefitinib for the Treatment of Patients with Metastatic EGFR Mutation–Positive Non–Small Cell Lung Cancer. *Clin Cancer Res*, 22(6):1307–1312.

Kim, S.B., Dent, R., Wongchenko, M.J. et al (2017). Concordance between plasma-based and tissue-based next-generation sequencing in LOTUS. *The Lancet Oncology*, 18(11):e638.

Kobayashi, K. and Hagiwara, K. (2013). Epidermal growth factor receptor (EGFR) mutation and personalized therapy in advanced nonsmall cell lung cancer (NSCLC). *Target Oncol*, 8(1):27–33.

Koboldt, D.C., Zhang, Q., Larson, D.E. et al (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576.

Kopetz, S., Desai, J., Chan, E. et al (2010). PLX4032 in metastatic colorectal cancer patients with mutant BRAF tumors. *JCO*, 28(15_suppl):3534–3534.

Kopetz, S., Desai, J., Chan, E. et al (2015). Phase II Pilot Study of Vemurafenib in Patients With Metastatic BRAF-Mutated Colorectal Cancer. *J Clin Oncol*, 33(34):4032–4038.

Kordes, M., Röring, M., Heining, C. et al (2016). Cooperation of BRAFF595l and mutant HRAS in histiocytic sarcoma provides new insights into oncogenic BRAF signaling. *Leukemia*, 30(4):937–946.

Kuderer, N.M., Burton, K.A., Blau, S. et al (2017). Comparison of 2 Commercially Available Next-Generation Sequencing Platforms in Oncology. *JAMA Oncol*, 3(7):996–998.

Landrum, M.J., Lee, J.M., Benson, M. et al (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, 46(Database issue):D1062–D1067.

Lara, P.N., Laptalo, L., Longmate, J. et al (2004). Trastuzumab plus docetaxel in HER2/neu-positive non-small-cell lung cancer: a California Cancer Consortium screening and phase II trial. *Clin Lung Cancer*, 5(4):231–236.

Lawrence, M.S., Stojanov, P., Mermel, C.H. et al (2014). Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature*, 505(7484):495–501.

Lawrence, M.S., Stojanov, P., Polak, P. et al (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Le, D.T., Uram, J.N., Wang, H. et al (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *http://dx.doi.org/10.1056/NEJMoa1500596*.

Le Tourneau, C., Delord, J.P., Gonçalves, A. et al (2015a). Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The Lancet Oncology*, 16(13):1324–1334.

Le Tourneau, C., Kamal, M., Tsimberidou, A.M. et al (2015b). Treatment Algorithms Based on Tumor Molecular Profiling: The Essence of Precision Medicine Trials. *J Natl Cancer Inst*, 108(4).

Lebofsky, R., Decraene, C., Bernard, V. et al (2015). Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. *Molecular Oncology*, 9(4):783–790.

Leder, P., Battey, J., Lenoir, G. et al (1983). Translocations among antibody genes in human cancer. *Science*, 222(4625):765–771.

Ledermann, J., Harter, P., Gourley, C. et al (2014). Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. *The Lancet Oncology*, 15(8):852–861.

Lei, Z., Tan, I.B., Das, K. et al (2013). Identification of Molecular Subtypes of Gastric Cancer With Different Responses to PI3-Kinase Inhibitors and 5-Fluorouracil. *Gastroenterology*, 145(3):554–565.

Levy, M.A., Lovly, C.M. and Pao, W. (2012). Translating genomic information into clinical medicine: Lung cancer as a paradigm. *Genome Res.*, 22(11):2101–2108.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A. et al (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, M.M., Datto, M., Duncavage, E.J. et al (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of Molecular Diagnostics*, 19(1):4–23.

Lih, C.J., Harrington, R.D., Sims, D.J. et al (2017). Analytical Validation of the Next-Generation Sequencing Assay for a Nationwide Signal-Finding Clinical Trial:Molecular Analysis for Therapy Choice Clinical Trial. *The Journal of Molecular Diagnostics*, 19(2):313–327.

Lih, C.J., Sims, D.J., Harrington, R.D. et al (2016). Analytical Validation and Application of a Targeted Next-Generation Sequencing Mutation-Detection Assay for Use in Treatment Assignment in the NCI-MPACT Trial. *J Mol Diagn*, 18(1):51–67.

Lindeman, N.I., Cagle, P.T., Beasley, M.B. et al (2013). Molecular Testing Guideline for Selection of Lung Cancer Patients for EGFR and ALK Tyrosine Kinase Inhibitors: Guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *Journal of Thoracic Oncology*, 8(7):823–859.

Lièvre, A., Bachet, J.B., Corre, D.L. et al (2006). KRAS Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer. *Cancer Res*, 66(8):3992–3995.

Lolkema, M.P., Gadellaa-van Hooijdonk, C.G., Bredenoord, A.L. et al (2013). Ethical, Legal, and Counseling Challenges Surrounding the Return of Genetic Results in Oncology. *JCO*, 31(15):1842–1848.

Lord, C.J. and Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. *Science*, 355(6330):1152–1158.

Lord, C.J., Tutt, A.N. and Ashworth, A. (2015). Synthetic Lethality and Cancer Therapy: Lessons Learned from the Development of PARP Inhibitors. *Annu. Rev. Med.*, 66(1):455–470.

Lynch, T.J., Bell, D.W., Sordella, R. et al (2004). Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *New England Journal of Medicine*, 350(21):2129–2139.

Mandelker, D., Zhang, L., Kemel, Y. et al (2017). Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline Testing. *JAMA*, 318(9):825–835.

Marquart, J., Chen, E.Y. and Prasad, V. (2018). Estimation of The Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. *JAMA Oncol*.

Massard, C., Michiels, S., Ferté, C. et al (2017). High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. *Cancer Discov*.

Mateo, J., Carreira, S., Sandhu, S. et al (2015). DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. *New England Journal of Medicine*, 373(18):1697–1708.

Matthijs, G., Souche, E., Alders, M. et al (2016). Guidelines for diagnostic next-generation sequencing. *Eur J Hum Genet*, 24(1):2–5.

McKenna, A., Hanna, M., Banks, E. et al (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*

Menck, K., Bleckmann, A., Wachter, A. et al (2017). Characterisation of tumour-derived microvesicles in cancer patients' blood and correlation with clinical outcome. *J Extracell Vesicles*, 6(1).

Meric-Bernstam, F., Brusco, L., Shaw, K. et al (2015a). Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. *JCO*, 33(25):2753–2762.

Meric-Bernstam, F., Johnson, A., Holla, V. et al (2015b). A Decision Support Framework for Genomically Informed Investigational Cancer Therapy. *JNCI J Natl Cancer Inst*, 107(7):djv098.

Mermel, C.H., Schumacher, S.E., Hill, B. et al (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12:R41.

Mertens, F., Johansson, B., Fioretos, T. et al (2015). The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, 15(6):371–381.

Metzker, M.L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Miki, Y., Swensen, J., Shattuck-Eidens, D. et al (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71.

Nakanishi, Y., Akiyama, N., Tsukaguchi, T. et al (2014). The Fibroblast Growth Factor Receptor Genetic Status as a Potential Predictor of the Sensitivity to CH5183284/Debio 1347, a Novel Selective FGFR Inhibitor. *Mol Cancer Ther*, 13(11):2547–2558.

Nakken, S., Fournous, G., Vodák, D. et al (2017). Personal Cancer Genome Reporter: variant interpretation report for precision oncology. *Bioinformatics*.

Network, C. (2006). FDA Approves Cetuximab to Treat Head and Neck Cancer: http://www.cancernetwork.com/head-neck-cancer/fda-approves-cetuximab-treat-head-and-neck-cancer.

Palmisano, A., Zhao, Y., Li, M.C. et al (2017). OpenGeneMed: a portable, flexible and customizable informatics hub for the coordination of next-generation sequencing studies in support of precision medicine trials. *Brief Bioinform*, 18(5):723–734.

Pao, W., Miller, V., Zakowski, M. et al (2004). EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *PNAS*, 101(36):13306–13311.

Pao, W., Miller, V.A., Politi, K.A. et al (2005). Acquired Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib Is Associated with a Second Mutation in the EGFR Kinase Domain. *PLOS Medicine*, 2(3):e73.

Parker, J.S., Mullins, M., Cheang, M.C. et al (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *JCO*, 27(8):1160–1167.

Patterson, S.E., Liu, R., Statz, C.M. et al (2016). The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. *Human Genomics*, 10:4.

Pauli, C., Hopkins, B.D., Prandi, D. et al (2017). Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. *Cancer Discov*, 7(5):462–477.

Peiffer, D.A., Le, J.M., Steemers, F.J. et al (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, 16(9):1136–1148.

Perera-Bel, J., Hutter, B., Heining, C. et al (2018). From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. *Genome Medicine*, 10:18.

Perou, C.M., Sørlie, T., Eisen, M.B. et al (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

PharmGKB (2018). PharmGKB: https://www.pharmgkb.org.

Piñeiro-Yáñez, E., Reboiro-Jato, M., Gómez-López, G. et al (2018). PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data. *Genome Medicine*, 10:41.

Prahallad, A., Sun, C., Huang, S. et al (2012). Unresponsiveness of colon cancer to BRAF(V600e) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–103.

Pritchard, C.C., Mateo, J., Walsh, M.F. et al (2016). Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *https://doi.org/10.1056/NEJMoa1603144*.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Rennert, H., Eng, K., Zhang, T. et al (2016). Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision cancer care. *npj Genomic Medicine*, 1:16019.

Richards, C.S., Bale, S., Bellissimo, D.B. et al (2008). ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.*, 10(4):294–300.

Richards, S., Aziz, N., Bale, S. et al (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423.

Rimmer, A., Phan, H., Mathieson, I. et al (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*, 46(8):912–918.

Roychowdhury, S. and Chinnaiyan, A.M. (2016). Translating cancer genomes and transcriptomes for precision oncology. *CA: A Cancer Journal for Clinicians*, 66(1):75–88.

Roychowdhury, S., Iyer, M.K., Robinson, D.R. et al (2011). Personalized Oncology Through Integrative High-Throughput Sequencing: A Pilot Study. *Sci Transl Med*, 3(111):111ra121.

Rubin, M.A. (2015). Health: Make precision medicine work for cancer care. *Nature*, 520(7547):290–291.

Rubio-Perez, C., Tamborero, D., Schroeder, M. et al (2015). In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities. *Cancer Cell*, 27(3):382–396.

Saunders, C.T., Wong, W.S.W., Swamy, S. et al (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

Schuh, A., Dreau, H., Knight, S.J.L. et al (2018). Clinically actionable mutation profiles in patients with cancer identified by whole-genome sequencing. *Cold Spring Harb Mol Case Stud*, 4(2):a002279.

Schwaederle, M., Daniels, G.A., Piccioni, D.E. et al (2015a). On the Road to Precision Cancer Medicine: Analysis of Genomic Biomarker Actionability in 439 Patients. *Mol Cancer Ther*, page molcanther.1061.2014.

Schwaederle, M., Zhao, M., Lee, J.J. et al (2015b). Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *JCO*, 33(32):3817–3825.

Servant, N., Roméjon, J., Gestraud, P. et al (2014). Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial. *Front. Genet.*, 5:152.

Shameer, K., Badgeley, M.A., Miotto, R. et al (2017). Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Brief Bioinform*, 18(1):105–124.

Sharma, P. and Allison, J.P. (2015a). The future of immune checkpoint therapy. *Science*, 348(6230):56–61.

Sharma, P. and Allison, J.P. (2015b). Immune Checkpoint Targeting in Cancer Therapy: Toward Combination Strategies with Curative Potential. *Cell*, 161(2):205–214.

Sharma, P., Hu-Lieskovan, S., Wargo, J.A. et al (2017). Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell*, 168(4):707–723.

Sherry, S.T., Ward, M.H., Kholodov, M. et al (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.

Shinawi, M. and Cheung, S.W. (2008). The array CGH and its clinical applications. *Drug Discovery Today*, 13(17):760–770.

Shtivelman, E., Lifshitz, B., Gale, R.P. et al (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature*, 315(6020):550–554.

Sims, D., Sudbery, I., Ilott, N.E. et al (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–32.

Siu, L.L., Lawler, M., Haussler, D. et al (2016). Facilitating a culture of responsible and effective sharing of cancer genome data. *Nature Medicine*, 22(5):464–471.

Sohal, D.P.S., Rini, B.I., Khorana, A.A. et al (2016). Prospective Clinical Study of Precision Oncology in Solid Tumors. *JNCI J Natl Cancer Inst*, 108(3):djv332.

Sone, T., Araya, T., Tambo, Y. et al (2015). A Phase II study to evaluate the efficacy of erlotinib in advanced NSCLC patients who have wild-type EGFR and EGFR gene amplification. *JCO*, 33(15_suppl):e19028–e19028.

Sukhai, M.A., Craddock, K.J., Thomas, M. et al (2016). A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. *Genet Med*, 18(2):128–136.

Sullivan, I. and Planchard, D. (2016). ALK inhibitors in non-small cell lung cancer: the latest evidence and developments. *Ther Adv Med Oncol*, 8(1):32–47.

Sun, S.Q., Mashl, R.J., Sengupta, S. et al (2018). Database of evidence for precision oncology portal. *Bioinformatics*.

Suthers, G. (2009). *Guidelines for reporting molecular genetic tests to medical practitioners*.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C. et al (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports*, 3:2650.

Tamborero, D., Rubio-Perez, C., Deu-Pons, J. et al (2018a). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10:25.

Tamborero, D., Rubio-Perez, C., Muiños, F. et al (2018b). A Pan-cancer Landscape of Interactions between Solid Tumors and Infiltrating Immune Cell Populations. *Clin Cancer Res*.

Tannock, I.F. and Hickman, J.A. (2016). Limits to Personalized Cancer Medicine. *New England Journal of Medicine*, 375(13):1289–1294.

Tattini, L., D'Aurizio, R. and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol*, 3.

Team, R.C. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

The AACR Project GENIE Consortium, . (2017). AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov*, 7(8):818–831.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Tsimberidou, A.M., Iskander, N.G., Hong, D.S. et al (2012). Personalized Medicine in a Phase I Clinical Trials Program: The MD Anderson Cancer Center Initiative. *Clin Cancer Res*, 18(22):6373–6383.

Tsimberidou, A.M., Wen, S., Hong, D.S. et al (2014). Personalized Medicine for Patients with Advanced Cancer in the Phase I Program at MD Anderson: Validation and Landmark Analyses. *Clin Cancer Res*, 20(18):4827–4836.

U.S. Food and Drug Administration (2013). FDA Approves Afatinib: http://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm360574.htm.

U.S. Food and Drug Administration (2014). FDA Approves Olaparib for the Treatment of Advanced Ovarian Cancer: http://fdaguidance.net/2014/12/fda-approves-olaparib-for-the-treatment-of-advanced-ovarian-cancer/.

U.S. Food and Drug Administration (2018). Approved Drugs - FDA approves olaparib for germline BRCA-mutated metastatic breast cancer: https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm592357.htm.

Vaishnavi, A., Le, A.T. and Doebele, R.C. (2015). TRKing Down an Old Oncogene in a New Era of Targeted Therapy. *Cancer Discov*, 5(1):25–34.

Van Allen, E.M., Wagle, N., Stojanov, P. et al (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.*, 20(6):682–8.

Vargas, A.J. and Harris, C.C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. *Nat Rev Cancer*, 16(8):525–537.

Velden, V.D., L, D., Herpen, V. et al (2017). Molecular Tumor Boards: current practice and future needs. *Ann Oncol*, 28(12):3070–3075.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E. et al (2013). Cancer Genome Landscapes. *Science*, 339(6127):1546–1558.

Wagle, N., Berger, M.F., Davis, M.J. et al (2012). High-Throughput Detection of Actionable Genomic Alterations in Clinical Tumor Samples by Targeted, Massively Parallel Sequencing. *Cancer Discov*, 2(1):82–93.

Wang, J., Mullighan, C.G., Easton, J. et al (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Meth*, 8(8):652–654.

Weinstein, J.N., Collisson, E.A., Mills, G.B. et al (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet*, 45(10):1113–1120.

Welch, B.M. and Kawamoto, K. (2013). The Need for Clinical Decision Support Integrated with the Electronic Health Record for the Clinical Application of Whole Genome Sequencing Information. *Journal of Personalized Medicine*, 3(4):306–325.

Wheler, J.J., Janku, F., Naing, A. et al (2016). Cancer Therapy Directed by Comprehensive Genomic Profiling: A Single Center Study. *Cancer Res*, 76(13):3690–3701.

WHO-ICTRP (2018). ICTRP Search Portal: http://apps.who.int/trialsearch/.

Wishart, D.S., Knox, C., Guo, A.C. et al (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue):D668–D672.

Wooster, R., Neuhausen, S.L., Mangion, J. et al (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, 265(5181):2088–2090.

Xin, J., Mark, A., Afrasiabi, C. et al (2016). High-performance web services for querying gene and variant annotation. *Genome Biology*, 17(1):91.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24.

Yaktapour, N., Meiss, F., Mastroianni, J. et al (2014). BRAF inhibitor–associated ERK activation drives development of chronic lymphocytic leukemia. *J Clin Invest*, 124(11):5074–5084.

Yan, W., Wistuba, I.I., Emmert-Buck, M.R. et al (2010). Squamous cell carcinoma – similarities and differences among anatomical sites. *Am J Cancer Res*, 1(3):275–300.

Yates, L.R. and Campbell, P.J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11):795–806.

Zack, T.I., Schumacher, S.E., Carter, S.L. et al (2013). Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, 45(10):1134–1140.

Zehir, A., Benayed, R., Shah, R.H. et al (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, 23(6):703–713.

Zhao, M., Wang, Q., Wang, Q. et al (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14(11):S1.

Zhao, Y., Polley, E.C., Li, M.C. et al (2015). GeneMed: An Informatics Hub for the Coordination of Next-Generation Sequencing Studies that Support Precision Oncology Clinical Trials. *Cancer Inform*, 14(Suppl 2):45–55.

Zill, O.A., Banks, K.C., Fairclough, S.R. et al (2018). The landscape of actionable genomic alterations in cell-free circulating tumor DNA from 21,807 advanced cancer patients. *Clin Cancer Res*, page clincanres.3837.2017.

Curriculum Vitae

Júlia Perera Bel

Personal Information

Ewaldstr. 25
37085 Göttingen, Germany
E-Mail: julia.perera@med.uni-goettingen.de

Date of Birth: 21/09/1990
Place of Birth: Barcelona

Education and Research

02/2015 – present **PhD student at Department of Medical Statistics, Göttingen, Germany**
PhD thesis 'Guiding Cancer Therapy: Evidence-driven Reporting of Genomic Data'

09/2012 – 07/2014 **Master in Bioinformatics for Health Sciences, University Pompeu Fabra, Barcelona**
Master thesis 'Unveiling deep-ocean protists through metagenomics' at the Marine Science Institute, CSIC, Barcelona

09/2008 – 07/2012 **Bachelor in Human Biology, University Pompeu Fabra, Barcelona**
Bachelor thesis 'Effect of DREF and Ken proteins in drosophila telomeres regulation' at the Institute of Evolutionary Biology, CSIC-UPF, Barcelona

Professional Experience

10/2014 – present **Research Associate, Statistical Bioinformatics Group**
Department of Medical Statistics, University Medical Center Göttingen, Germany

07/2013 – 07/2014 **Internship, Marine Science Institute, CSIC, Barcelona**

Publications

Wolff, A., **Perera-Bel, J.**, Schildhaus, H. U., Homayounfar, K., Schatlo, B., Bleckmann, A., and Beißbarth, T. (2018). Using RNA-Seq Data for the Detection of a Panel of Clinically Relevant Mutations. Studies in health technology and informatics, 253, 217-221. doi: 10.3233/978-1-61499-896-9-217

Perera-Bel, J., Hutter, B., Heining, C., Bleckmann, A., Fröhlich, M., Fröhling, S., Glimm H., Brors B., and Beißbarth, T. (2018). From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. Genome medicine, 10(1), 18. doi: 10.1186/s13073-018-0529-2.

Pernice, M.C., Giner, C.R., Logares, R., **Perera-Bel, J**., Acinas, S.G., Duarte, C.M., Gasol J.M., and Massana, R. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. The ISME journal, 10(4), 945. doi: 10.1038/ismej.2015.170