Automated Identification of Targeted Therapy Strategies in Precision Oncology

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

M.Sc. Bilge Sürün

aus Istanbul, Türkei

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	20.10.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Prof. Dr. Nico Pfeifer

"A person’s life purpose is nothing more than to rediscover, through the detours of art or love or passionate work, those one or two images in the presence of which his heart first opened."

Albert Camus

Erklärung

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel:

Automated Identification of Targeted Therapy Strategies in Precision Oncology

selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

____________________ ____________________

Ort, Datum Unterschrift

Abstract

Precision in cancer treatment builds upon targeted strategies tailored to the genomic traits of patients instigating pathological abnormalities. Extrapolating phenotype to genotype translations to oncology clinics has led to a less costly and more efficient cancer care model. However, its implementation remains challenging due to the complex analysis trajectory requiring various bioinformatics tools and databases. It relies on the individual expertise of MTBs executing a non-standard framework with a limited number of pharmacogenomics sources. The disadvantages of existing tools emanate from requiring programmatic skills, not addressing data privacy concerns, the large number of clinical evidence databases, and the lack of GUI tailored to MTB's workflow.

We created ClinVAP, a cohesive framework for clinical annotation of genomic variants which automates the process of generating patient-specific diagnostic reports by translating the long list of mutations to clinical implications. We enriched it with the gene-gene interactions that also reveal the content of disrupted pathways. We provided the combined results in an interactive GUI which isolates backend operations from the users and allows them to operate through the results. We measured the adaptability of ClinVAP using retrospective cases to compare their contentwise equality to the MTB's implementation. The differences were mainly based on expert opinion. The content and the structure of the automated patient reporting tools form a comprehensive foundation to be used in decision making.

The future of precision oncology depends on the accessibility of the accumulated molecular knowledge of the disease-contributing factors. The number of bioinformatics tools and the sheer size of genome data is a barrier to making this information available in hospitals. Our solutions not only increase their clinical applicability, but also demonstrate the field's readiness to generate automated solutions. Moreover, standardization and archiving will facilitate population studies, allowing molecular analyses to be archived and returned to the system as information.

Zusammenfassung

Individualisierung in der Krebsbehandlung beruht auf gezielten Strategien, die auf die genomischen Merkmale der Patienten zugeschnitten sind, die pathologische Anomalien verursachen. Die Extrapolation von Phänotyp-Genotyp-Beziehungen auf onkologische Kliniken hat zu einem weniger kostspieligen und effizienteren Krebsbehandlungsmodell geführt. Die Umsetzung ist jedoch nach wie vor schwierig, da für die komplexe Analyse verschiedene Bioinformatik-Tools und Datenbanken erforderlich sind. Sie beruht auf dem individuellen Fachwissen der MTBs, die einen nicht standardisierten Rahmen mit einer begrenzten Anzahl von Quellen ausführen. Die Nachteile bestehender Tools bestehen darin, dass sie Programmierkenntnisse erfordern, den Datenschutz nicht berücksichtigen, eine Vielzahl von Datenbanken mit klinischer Evidenz enthalten und keine auf die Arbeitsabläufe von Molekularen Tumorboards (MTBs) zugeschnittene Benutzeroberfläche haben.

Wir haben ClinVAP entwickelt, ein kohärentes Framework für die klinische Annotation von Genomvarianten, das den Prozess der Erstellung patientenspezifischer Diagnoseberichte automatisiert, indem es die lange Liste von Mutationen in klinische Implikationen übersetzt. Wir haben es mit den Gen-Gen-Interaktionen angereichert, die auch den Inhalt der gestörten Signalwege aufzeigen. Wir haben die kombinierten Ergebnisse in einer interaktiven grafischen Benutzeroberfläche (GUI) bereitgestellt, die die Backend-Operationen von den Nutzern isoliert und es ihnen ermöglicht, die Ergebnisse zu bearbeiten. Wir haben die Anpassungsfähigkeit von ClinVAP anhand von retrospektiven Fällen gemessen, um ihre inhaltliche Gleichheit mit der manuellen Implementierung im MTB zu vergleichen. Die Unterschiede beruhten hauptsächlich auf Expertenmeinungen. Der Inhalt und die Struktur der automatisierten Patientenberichts-Tools sind eine umfassende Grundlage für die Entscheidungsfindung.

Die Zukunft der Präzisionsonkologie hängt von der Zugänglichkeit des gesammelten molekularen Wissens über die krankheitsverursachenden Faktoren ab. Die Vielzahl der Bioinformatik-Tools und die schiere Größe der Genomdaten stellen ein Hindernis für

die Bereitstellung dieser Informationen in Krankenhäusern dar. Unsere Lösungen erhöhen nicht nur ihre klinische Anwendbarkeit, sondern zeigen auch, dass das Feld bereit ist, automatisierte Lösungen zu entwickeln. Darüber hinaus werden Standardisierung und Archivierung Populationsstudien erleichtern, da molekulare Analysen archiviert und als Informationen an das System zurückgegeben werden können.

Acknowledgments

I would like to express my gratitude to my supervisor Oliver Kohlbacher whose mentorship has had a profound impact on my personal and professional development.

I would like to thank Kohlbacher Lab members for the academic insights we shared over interesting conversations, and for the incredible support they provided every time I needed assistance with my projects. I would like to thank my colleagues Thorsten Tiede and Sam Wein for their efforts in proofreading my thesis and providing invaluable comments, to Andras Szolek for being an amazing office companion and a cherished friend, to Koray Kirli and Leon Kuchenbecker for their motivational support and understanding during the writing phase of my dissertation.

I would also like to acknowledge the administrative and technical teams in Tübingen for ensuring the smooth operation of our activities, particularly to Claudia Walter, for her invaluable assistance and patience in navigating the bureaucratic challenges I encountered.

A special acknowledgement goes to my flatmates, who have not only provided support throughout my PhD journey but have also become my second family in Germany. Furthermore, I wish to express my appreciation to my "Kletter Crew". Their presence has transformed my PhD experience into an adventurous chapter of my life, filled with unforgettable memories.

Last but not least, I would like to thank my parents, whose love and encouragement have been a source of strength, and to my siblings for the joy they bring into my life.

General Remarks

- In accordance with the standard scientific protocol, I will use the personal pronoun ***we*** to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Chapter 1

Introduction

Motivation

The advancements in the sequencing methods commercially enabled Next-Generation Sequencing (NGS) technologies, which have revolutionized genomic research by producing an unprecedented amount of data. With such advancements, 30 years after the Human Genome Project, the sequencing cost is under 1,000 dollars per genome and the rate of data growth is doubling every seven months [1,2]. Our ability to generate genomic data at reduced costs with speed makes sequencing the entire genome and the whole exome possible and produces approximately one terabase of data per run with today's instruments[3].

Rapid developments in NGS technologies have revealed genetic variability among humans and other species. It leveraged the study of evolution and natural selection, genome editing, and synthetic biology. One of the most compelling applications is discovering novel molecular disease mechanisms to link genotypes with the complex traits in the population with large-scale genome-wide association studies (GWAS). GWAS has reached more than one million individuals and resulted in the risk loci mappings for a vast amount of diseases with the identification of more than 50,000 unique single nucleotide variant (SNV)-trait associations[4–6]. Although such large-scale population studies revealed the molecular rationale of monogenic diseases and led to successful clinical applications, the lack of causal gene findings in polygenic complex diseases among many associated traits has hampered its precise clinical translation. Even though the promise of solving complex genetic diseases at the molecular level was not fully achieved, NGS data provided information on inter-individual variations.

The discovery of the effect of individuals' genetic mark-up on their healthcare pathway from prevention to treatment emerged in precision medicine, which tailors the healthcare regimen based on the molecular patient profile to deliver the optimal treatment with increased efficacy and reduced costs. It created an opportunity for translating the molecular findings contributing to disease heterogeneity into the clinical routine. It enabled patient stratification based on the existence of biomarkers contributing to disease progression and defining the response to the therapy. Precision medicine has especially shown promising applications in oncology, as cancer is a complex genomic disease that resulted in the accumulation of mutations and the alteration of molecular paths[7].

The conventional methods of clinical oncology heavily rely on morphological and histopathological diagnostic methods as well as on surgical removal of tumor tissue and/or chemo- or radiation therapy as a treatment. However, this first line of treatment methods frequently lead to complications such as the increased risk of surgery-related micro-metastases, therapy resistance, toxicity, short relapse time, and increased risk of secondary tumors[8–10]. The problems in the standard therapies have compelled patient care to shift from conventional clinical workflow to patient-specific biomarker-driven targeted therapies, so-called precision oncology.

Target identification has been accelerated with the advancements in genomics by linking aberrations to therapies. Data from big consortia such as ICGC and TCGA have been analyzed individuals systematically and created a large catalog of the driver genes which provide a selective advantage to the cells to initiate uncontrolled proliferation. The shift from population-based to biomarker-driven clinical trials yields a growing arsenal of drugs. As we know more about molecular cancer mechanisms, biological reasons, and the pathways involved, translation of this information to clinics is getting more and more promising. However, target identification is complex and follows a long data analysis trajectory from sequencing of the genetic material to concluding on a responsive therapy option based on a set of informative and actionable variants. One of the main bottlenecks in closing the gap between clinics and the genomics field is the complexity of the analysis platforms and the labor-intensive work of clinical annotations. Thereby refinement of in silico methods is needed to integrate genomics into the clinical routine to improve healthcare by increasing the usability of these platforms for healthcare professionals. Corroboration is needed in selecting optimal analytical platforms, standardization of the sequencing assays, reporting metrics, and assessing the efficacy of the selected therapies [11].

In this thesis, we provide several approaches to translate genomic data into patient-specific cancer profiling. Our contributions cover a wide range of steps in the clinical decision-making workflow varying from the development of a fully automated pipeline for molecular profiling of cancer patients to combining different levels of clinical information and provide a visual analytic decision as support tool.

Part I: ClinVAP

Targeted therapy is proven as a gateway to treatment with an increased overall survival rate in various cancer types such as metastatic renal cell carcinoma, non-small-cell lung cancer, and ovarian cancer, due to data-driven decision(s) which are made based on the patient's unique disease profile[12–14]. The complexity of target identification emerges from the need of translating large and unstructured data into clinical information using bioinformatics tools. The amount of data to be processed, the variety of the annotation tools and data sources together with the complex programmatic interfaces provide oncology clinics a disadvantage[15]. Hence, there is a need for automated pipelines to respond to the data analysis needs of clinics such as ensuring data security, reproducibility, scalability, and interoperability.

In the first part of the thesis, we implemented a fully automated pipeline to convert genomic data into concise clinical information. Our method includes the functional variant annotation that prioritizes the long list of mutations to identify molecular mechanisms that drive carcinogenesis. For therapeutic target and response identification, we developed a clinical variant annotation method that also integrates clinical evidence from a wide range of publicly available information sources. The results give evidence-based therapeutic suggestions based on the level of gene disruption in a patient's cancer profile.

Part II: Case profiling with reports and networks via visual analytic tool

Cancer is a complex system disease[16]. Not only driver genes but also driver events are important in target identification due to the additive effects of the mutations on oncogenic signaling pathways[17]. Additionally, a big part of the human cancer genome is not yet directly druggable with FDA-approved drugs[18] [19,20]. This makes the off-label use and re-purposing of the available drugs crucial in optimizing treatments using the clinical evidence based on upstream of driver genes. Therefore a patient's molecular profile together with networks showing the molecular interactions of genes would give a complete case overview, present opportunities for identifying re-purposing drug candidates as well as pointing out the secondary drug resistance mechanisms, and reveal pathways to take action on. Such clinomics approaches are arduous due to the complex nature of data and the lack of comprehensive tools which conduct variant and network annotation and visualize the results with an effective graphical user interface.

In the second part of the thesis, we developed a visual analytics tool that combines evidence-based case reports with the networks showing the other molecular players in the proximity of the disrupted genes. It provides the results in a compendious graphical user interface. Our tool has an added value in the precision oncology clinical setup as it provides multi-level information at once which is necessary in clinical decision making and enables users to conduct a more comprehensive case analysis.

Part III: Systematic MTB data analysis and comparison

As targeted therapies were introduced to the standard line of care, oncology clinics established molecular tumor boards (MTBs), a multi-disciplinary committee that is assigned to make treatment suggestions based on the patient's genomics information. Although precision oncology has proven to be beneficial in various studies[21–24], lack of reproducibility and reliability remain its major hurdles[25]. The low agreement rate between the MTBs is associated with the absence of the standard workflows[26,27]. Assigning significance to the aberrations mostly rely on the expertise of individual MTBs, which is often an error-prone manual task. All these obstacles assert the need for standardized and automated workflows to increase user acceptance in clinics. In addition to the software requirements of decision support tools, it is of crucial importance to assess the reliability of the produced results for clinical implementation.

In the third part of the thesis, we conducted a stratified case-cohort analysis to demonstrate the reliability of ClinVAP which was developed to offer a solution for MTBs in their need of a standardized, automated, and reproducible analysis tool. We assessed the adaptability of our annotation tool over the current clinical practice in therapeutic decision-making. With the retrospective genomics data analysis, we have shown the efficiency of our *in silico* treatment stratification by proving the completeness of its content in comparison to manually prepared case reports.

Chapter 2

Background

2.1 The Human Genome

The genetic material of all living species is composed of deoxyribonucleic acid (DNA) which encloses all the information to define the molecular composition of an organism[28]. The genetic information is decoded via transcription, forming messenger ribonucleic acid (RNA) that governs protein synthesis. This process that is termed as *central dogma of molecular biology*[29] assembles the proteins, the building blocks of organisms carrying out the essential functions including the catalyzing the reactions and generating the defense units of the immune system.

The genetic material is densely compacted as chromosomes in the cells' nucleus. Human cells contain 23 pairs of chromosomes[28] that approximately consists of 3 billion DNA bases[30]. The DNA regions corresponding to the total of the protein-coding units (genes) makes up approximately 1% of the overall base count[30]. The remaining portion of DNA which was perceived as junk two decades ago is found to have essential regulatory roles[30] in orchestrating many processes varying from cell differentiation to coding for untranslated regulatory RNAs (e.g. microRNAs).

Each genome contains fragmented chunks of exon sequences, which are the protein-coding domains interspersed with DNA's non-coding fragments, introns. RNA polymerase (RNAP) initiates transcription by binding to the promoter region located in the 5' terminus of the DNA. This process is orchestrated by transcription factors (TFs) to regulate the transcription of the specific regions at a specific time. RNAP moves along the sequence from the 5' to 3' end and creates an immature RNA (pre-mRNA) copy of the DNA. pre-mRNA is transduced into mature RNA (mRNA) by removing introns it contains via the splicing process. Then mRNA moves to the cytoplasm and synthesizes proteins.

It is estimated that the human genome contains 24,000 protein-coding genes which encode more than 100,000 different proteins. The mechanism leading to large protein diversity is found to be the alternative splicing through exon arrangements during the maturation of the pre-mRNA[31]. It provides evolutionary benefits to the organisms by efficiently forming new functionalities with the inclusion or exclusion of the exons without changing the source of information[32].

The developments in sequencing technologies paved the way to assemble the human reference genome. It has been continually improved over the years to increase its coverage which reached 90% in the latest releases, GRCh37 and GRCh38[33]. Although the field is turning toward generating population-specific consensus genomes[34], the available versions have been a valuable source in understanding the genetic structure of humankind and contributed to improving health care.

2.1.1 Genetic Variation

Genetic variation is introduced by mutations and sexual reproduction during meiosis through the independent assortment of the chromosomes and the cross-over. Mutations affecting an organism manifest themselves in two major settings depending on the cell type they occurred in. Germline variants take place in reproductive cells (sperm or eggs) and are passed to the offspring with the zygote formation of either the mutated sperm or the oocyte. Somatic variants are the post-zygotic mutations that occur in any type of the cells other than the germ cells; thus, they are non-heritable.

Human genome analysis revealed that 99.9% of the genetic material of the humans are identical and the individuality manifests itself in the 0.1% fraction[35]. Through comparative genomic studies focusing on the small fraction of the genome, the link between the genomic variation and the phenotype has started to be established. Genetic variation is expressed in a wide range of alteration classes varying from single nucleotide variants (SNVs) to large chromosomal rearrangements. SNVs are the result of single nucleotide substitution where the amino acid length is preserved and they are found to be the most abundant variant type in the human genome[36]. Small INDELs are the insertions and deletions between 1 bp to 9,989 bp in length[37], the second most prevalent in the population after SNVs[38]. Structural variant (SV) is an umbrella term covering large chromosomal aberrations including copy number variants (CNVs), inversions, translocations, and segmental uniparental disomies. CNVs denote the duplications, insertions, and deletions causing changes in the copy number compared to the reference genome[35]. The aberrations such as translocations, inversions, and insertions might emerge gene fusions[39] which may form chimeric RNA translating novel proteins[40] or deregulate transcription by affecting the regulatory DNA regions[41].

Genetic variation is the main source of the phenotypic diversity determining the unique traits of a person with polymorphisms shared within the populations with the line of ancestry. Genetic diversity does not only control the phenotypic distinctions but also governs a person's susceptibility to diseases and their treatment responses. The pathogenic effects of the acquired mutations through the change of protein structure and function are found to be the underlying causes of many diseases emphasizing the importance of the population studies to assign significance to the observed variation.

2.1.2 Genetic Variations in Diseases

The advancements in sequencing technologies facilitated to generate genomics data for millions of individuals. Although this substantial influx of data has been translating into immense progress in the clinical understanding of the molecular disease mechanisms, it remains challenging to identify the associated loci with the complex traits.

The disease phenotype association of the genetic variations requires large screenings among the different populations before predicting their impact to catalog them based on their pathogenesis. Genotyping is one of the SNV discovery and screening methods that are used to identify variants based on genotyping assays for detection. The method uses the regions harboring the variants as a backbone and imputes the space between the fragments with haplotypes. It is heavily used in genome-wide association studies (GWAS) to map complex traits and the loci by comparing the population exhibiting a certain disease phenotype with the healthy individuals. One problem with the genotyping assays is that the statistical significance might not be the main cause of the disease traits. The assays may fail to capture the rare alleles which are the main contributors to the anomalies[42]. Moreover, the variants found significant in GWAS studies need further investigation since due to the linkage disequilibrium, multiple variants are found associated and the process requires more examination of this overestimation to pinpoint the causative variants[43].

Another method to identify the variants is through sequencing technologies. Panel sequencing is used to test the patients for specific exonic regions that are known to be associated with a disease trait. It requires a pre-defined set of targets, it is cost and time-efficient; hence, it has the properties of a diagnostic tool that is routinely used in clinics rather than the function of assigning significance to the novel genes. Whole exome sequencing (WES) has the advantage over panel sequencing since it profiles all the DNA coding regions, provides information also on the CNVs and provides the set of aberrations that are readily interpretable. It has more sensitivity in identifying the regions that fall outside of the main target regions compared to its equivalent genotyping methods. However, it fails to detect large chromosomal changes such as the SVs[44]. Moreover, even though its interpretation is still challenging, the mutations in the non-coding regions of the DNA are also known as resembling the crucial part of the disease relevancy due to their function in regulatory mechanisms such as histone modification and transcription[45]. Thus, whole-genome sequencing (WGS) is another method for obtaining the entire genome including both coding and non-coding regions. The ability of WES and WGS to identify the potentially disruptive genes without having a prior candidate gene promotes their usage in both research and clinical diagnostics. However, the knowledge about disease-causing mutation has a modest translation to clinical diagnosis with the average molecular diagnosis rate being approximately 30%, differing based on the disease[46,47].

The challenges of increasing the known variant associations with diseases lie within the various aspects such as the elusive process of capturing SVs with short-read sequencing technologies and the high cost of alternative methods based on long reads[36], the under-representation of the large chromosomal variances in the reference genome and the need for more diverse, multiple reference genomes[36], the difficulties in elucidating the effect of non-coding variants and accounting for incomplete penetrance[48]. Regardless of these bottlenecks, the field is progressing rapidly through large-scale national sequencing consortiums and global collaborations of data sharing. In efforts to distinguish the disease-related variants among the ones that are attributing unique non-benign traits to humans, there are immense efforts to collect complete genomics data not only from the unhealthy individuals but also from the healthy individuals. Consortia such as All of Us launched in the USA aiming to sequence millions of individuals[49]. Another initiative 1,000 Genomes are collecting variants with 1 % or higher frequency among the population from the healthy individuals[50]. The large consortium created collaborative efforts to collect genomic data from large populations with diverse ethnicity and medical backgrounds such as the UK Biobank and FinnGen project, each genotyping 500,000 individuals[51]. Large databases are hosting the available data from the established associations closing the gap between the research and clinical practice laying the stepping stones for mitigating the complex process of establishing the molecular diagnosis and discovering more associations.

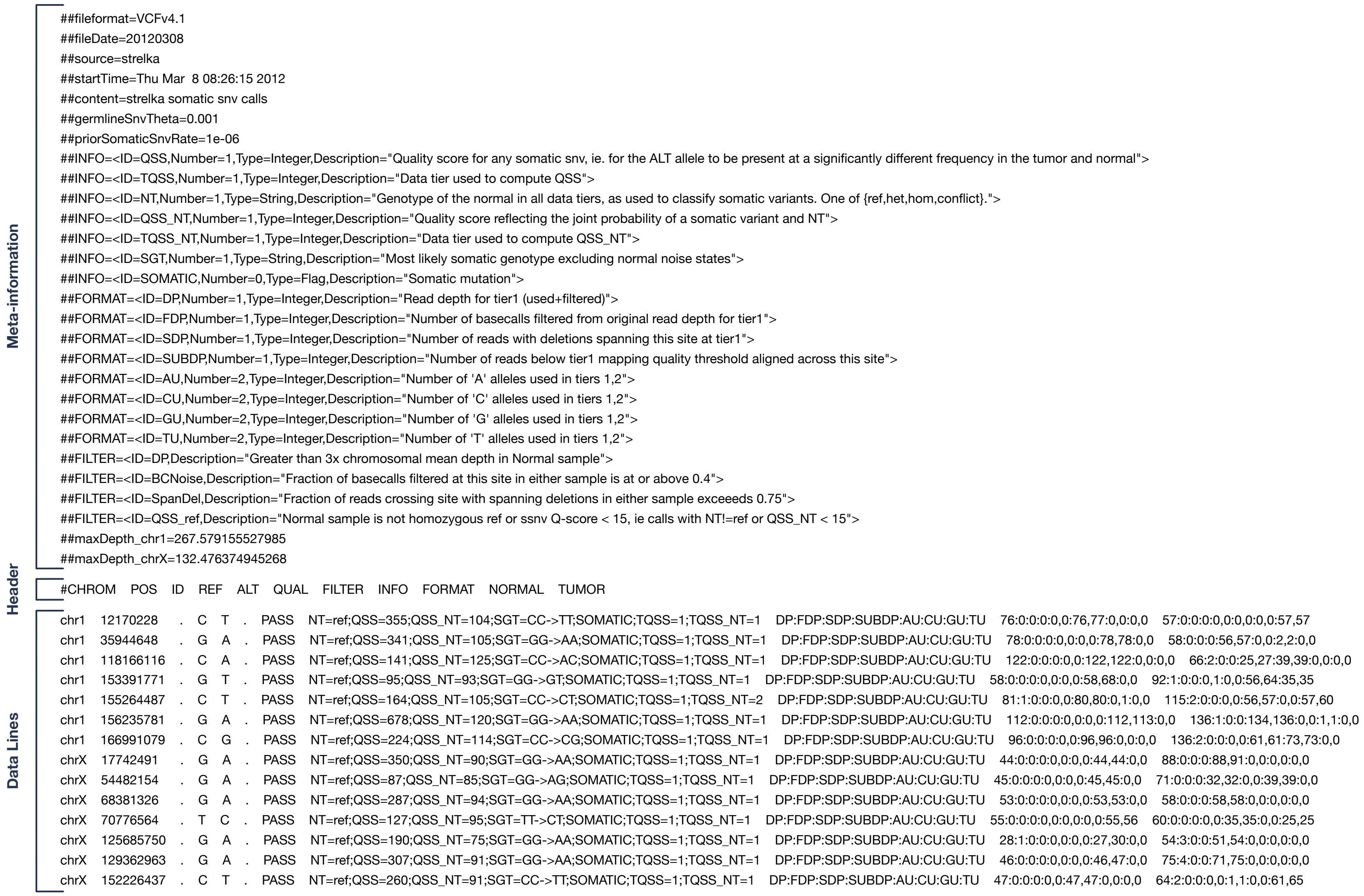
2.1.3 From Raw Reads to Variants

Making biological interpretations of the data produced by WES, WGS, and panel sequencing technologies lays the foundation for understanding human pathology. However, the process of extracting biological meaning from the copious amount of raw sequence reads is a challenging task. The process of identifying not yet alone SNVs but also the insertions, deletions, structural and copy number variations requires sophisticated and well-established tools to make inferences from the sequencing data. Calling the variants is a major step of NGS data analysis which is conducted with the pipelines developed for the specific type of analysis that the data required. While SNV calling focuses on finding the short deviations from the reference genome, the algorithms to call large SVs differ from the ones pinpointing SNVs and INDELs since they focus on recognizing the breakpoints where an unexpected change is observed on sequencing depth or misalignment of paired read ends[52].

A standard pipeline to call somatic SNVs and small INDELs from DNA sequencing data usually encompasses the main steps of pre-processing the raw sequences, mapping them to the reference genome, and annotating the variants, using a wide range of refined tools tackling the individual steps of the analysis. Pre-processing of the raw sequence reads provided as FASTQ file format, includes the quality control of the reads which leads to trimming the low-quality read ends to increase the success of the mapping with tools such as FastQC[53], Sickle[54] or Cutadapt[55]. The processed reads are then mapped to the reference genome to find the genomic locations of the short reads. There are many tools available for read mapping such as Bowtie2[56] or BWA[57]. The benchmarking studies showed that there is no one tool outperforming the others in each test and the method should be chosen based on the needs of users such as the tolerated amount of the false positives or the run time[58]. The read mapping process generates Binary Alignment Map (BAM), or its uncompressed version, Sequence Alignment Map (SAM) files from the raw sequencing reads FASTQ files. After conducting the post-processing steps such as filtering the mappings based on the mapping quality or eliminating the duplicate fragments, variant calling algorithms are implemented to pinpoint the variants observed in the data in comparison to the reference genome. Among various variant calling algorithms, Genome Analysis Toolkit (GATK)[59], SAMtools[60], VarScan[61] and Strelka[62] are the most popular variant calling tools each employing a different algorithm. The variant calling algorithms are specialized based on the analysis type mainly clustered around single sample variant calling, matched tumor-normal variant calling, and unique molecular identifiers (UMI) based variant calling. Over 40 publicly available somatic variant calling tools each specialized in one of the aforementioned types of analysis and their base algorithms are reviewed by [52]. The result of the variant calling step is the variant call format (VCF) which represents the genomic variation. VCF file is also generalized in the way of representing large chromosomal changes. The next step in extracting biological inference is to predict the potential effects of the observed variants which are discussed in the following Section.

A VCF file is a generic text file storing the SNVs, INDELs, and large SVs with standard specifications. It includes the meta-information, header, and the data lines showing the indexed variants (Figure 2.1). The meta-information section introduces the tags and the annotation used by the NGS analysis tools which enable users to tailor the format based on the information produced by the tools. It also incorporates the details of the tool that generated the data such as the name and version of it, the date that the data was created, the reference genome used in mapping, and the version of the VCF format. The header line introduces the category names of the columns. The mandatory columns are the chromosome number (CHROM), the starting position of the variant (POS), the ID of the observed variant (ID), reference and the alteration bases (REF, ALT), the quality score (QUAL), the filtering column (FILTER), and the info column (INFO) showing the annotation tags and their corresponding values. The mandatory columns are followed by the FORMAT column which indicates the genotype-related fields such as the read depth and the genotype quality. If the experiment contained more than one sample, their corresponding columns follow the FORMAT column (Figure 2.1). Having the variants in a generic standard format improves the interoperability of the tools producing the data and the results[63].

There are various pipelines combining the aforementioned tools to process the NGS data from raw reads to called variants[64–67]. However, the benchmarking studies demonstrated that there is no high concordance between the results of different pipelines due to the lack of standards for handling the experimental artifacts reflected in the raw data. Thus, it is difficult to suggest one size fits all solution and the user needs to optimize the combinations of available tools to increase the accuracy of the called variants[68].



Meta-information
Header
Data Lines

```
##fileformat=VCFv4.1
##fileDate=20120308
##source=strelka
##startTime=Thu Mar  8 08:26:15 2012
##content=strelka somatic snv calls
##germlineSnvTheta=0.001
##priorSomaticSnvRate=1e-06
##INFO=<ID=QSS,Number=1,Type=Integer,Description="Quality score for any somatic snv, ie. for the ALT allele to be present at a significantly different frequency in the tumor and normal">
##INFO=<ID=TQSS,Number=1,Type=Integer,Description="Data tier used to compute QSS">
##INFO=<ID=NT,Number=1,Type=String,Description="Genotype of the normal in all data tiers, as used to classify somatic variants. One of {ref,het,hom,conflict}.">
##INFO=<ID=QSS_NT,Number=1,Type=Integer,Description="Quality score reflecting the joint probability of a somatic variant and NT">
##INFO=<ID=TQSS_NT,Number=1,Type=Integer,Description="Data tier used to compute QSS_NT">
##INFO=<ID=SGT,Number=1,Type=String,Description="Most likely somatic genotype excluding normal noise states">
##INFO=<ID=SOMATIC,Number=0,Type=Flag,Description="Somatic mutation">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth for tier1 (used+filtered)">
##FORMAT=<ID=FDP,Number=1,Type=Integer,Description="Number of basecalls filtered from original read depth for tier1">
##FORMAT=<ID=SDP,Number=1,Type=Integer,Description="Number of reads with deletions spanning this site at tier1">
##FORMAT=<ID=SUBDP,Number=1,Type=Integer,Description="Number of reads below tier1 mapping quality threshold aligned across this site">
##FORMAT=<ID=AU,Number=2,Type=Integer,Description="Number of 'A' alleles used in tiers 1,2">
##FORMAT=<ID=CU,Number=2,Type=Integer,Description="Number of 'C' alleles used in tiers 1,2">
##FORMAT=<ID=GU,Number=2,Type=Integer,Description="Number of 'G' alleles used in tiers 1,2">
##FORMAT=<ID=TU,Number=2,Type=Integer,Description="Number of 'T' alleles used in tiers 1,2">
##FILTER=<ID=DP,Description="Greater than 3x chromosomal mean depth in Normal sample">
##FILTER=<ID=BCNoise,Description="Fraction of basecalls filtered at this site in either sample is at or above 0.4">
##FILTER=<ID=SpanDel,Description="Fraction of reads crossing site with spanning deletions in either sample exceeeds 0.75">
##FILTER=<ID=QSS_ref,Description="Normal sample is not homozygous ref or ssnv Q-score < 15, ie calls with NT!=ref or QSS_NT < 15">
##maxDepth_chr1=267.579155527985
##maxDepth_chrX=132.476374945268

#CHROM   POS   ID   REF   ALT   QUAL   FILTER   INFO   FORMAT   NORMAL   TUMOR

chr1   12170228    .   C   T   .   PASS   NT=ref;QSS=355;QSS_NT=104;SGT=CC->TT;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   76:0:0:0:0,0:76,77:0,0:0,0   57:0:0:0:0,0:0,0:0,0:57,57
chr1   35944648    .   G   A   .   PASS   NT=ref;QSS=341;QSS_NT=105;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   78:0:0:0:0,0:0,0:78,78:0,0   58:0:0:0:56,57:0,0:2,2:0,0
chr1   118166116   .   C   A   .   PASS   NT=ref;QSS=141;QSS_NT=125;SGT=CC->AC;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   122:0:0:0:0,0:122,122:0,0:0,0   66:2:0:0:25,27:39,39:0,0:0,0
chr1   153391771   .   G   T   .   PASS   NT=ref;QSS=95;QSS_NT=93;SGT=GG->GT;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   58:0:0:0:0,0:0,0:58,68:0,0   92:1:0:0:0,1:0,0:56,64:35,35
chr1   155264487   .   C   T   .   PASS   NT=ref;QSS=164;QSS_NT=105;SGT=CC->CT;SOMATIC;TQSS=1;TQSS_NT=2   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   81:1:0:0:0,0:80,80:0,1:0,0   115:2:0:0:0,0:56,57:0,0:57,60
chr1   156235781   .   G   A   .   PASS   NT=ref;QSS=678;QSS_NT=120;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   112:0:0:0:0,0:0,0:112,113:0,0   136:1:0:0:134,136:0,0:1,1:0,0
chr1   166991079   .   C   G   .   PASS   NT=ref;QSS=224;QSS_NT=114;SGT=CC->CG;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   96:0:0:0:0,0:96,96:0,0:0,0   136:2:0:0:0,0:61,61:73,73:0,0
chrX   17742491    .   G   A   .   PASS   NT=ref;QSS=350;QSS_NT=90;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   44:0:0:0:0,0:0,0:44,44:0,0   88:0:0:0:88,91:0,0:0,0:0,0
chrX   54482154    .   G   A   .   PASS   NT=ref;QSS=87;QSS_NT=85;SGT=GG->AG;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   45:0:0:0:0,0:0,0:45,45:0,0   71:0:0:0:32,32:0,0:39,39:0,0
chrX   68381326    .   G   A   .   PASS   NT=ref;QSS=287;QSS_NT=94;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   53:0:0:0:0,0:0,0:53,53:0,0   58:0:0:0:58,58:0,0:0,0:0,0
chrX   70776564    .   T   C   .   PASS   NT=ref;QSS=127;QSS_NT=95;SGT=TT->CT;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   55:0:0:0:0,0:0,0:0,0:55,56   60:0:0:0:0,0:35,35:0,0:25,25
chrX   125685750   .   G   A   .   PASS   NT=ref;QSS=190;QSS_NT=75;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   28:1:0:0:0,0:0,0:27,30:0,0   54:3:0:0:51,54:0,0:0,0:0,0
chrX   129362963   .   G   A   .   PASS   NT=ref;QSS=307;QSS_NT=91;SGT=GG->AA;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   46:0:0:0:0,0:0,0:46,47:0,0   75:4:0:0:71,75:0,0:0,0:0,0
chrX   152226437   .   C   T   .   PASS   NT=ref;QSS=260;QSS_NT=91;SGT=CC->TT;SOMATIC;TQSS=1;TQSS_NT=1   DP:FDP:SDP:SUBDP:AU:CU:GU:TU   47:0:0:0:0,0:47,47:0,0:0,0   64:2:0:0:0,0:1,1:0,0:61,65
```

Figure 2.1: *Meta-information* section encapsulates i) the file format version and the details on file generation including the tool and the date, time ii) ##INFO lines providing the name, description and the data type of the annotations, iii) ##FORMAT lines giving the description and the data types of the genotypical tags, iv) ##FILTER lines stating the filtering constraints applied to the data set followed by additional annotations, which is ##maxDepth_chr in this example provided as a demonstration of user-tailored annotation fields. *Header* denotes the column names of the data fields which contain two samples in this example which are NORMAL and TUMOR whose genotype information is provided following the FORMAT column tag order. The data is a subset of the publicly available VCF file provided by Strelka[62]

2.1.4 Variant Effect Prediction

Variant effect prediction is an essential step in NGS analysis forming the concrete foundation for interpreting the consequences of the observed variations on the molecular level. Predicted consequences of the variants enable the scientists to develop a broad understanding of the molecular rationale for the non-infectious diseases that are very common in the population such as cardiovascular diseases, cancer, and diabetes, which then translated into clinical usage.

There are various methods in the literature such as SnpEff[69], ANNOVAR[70] and Ensembl Variant Effect Predictor (VEP)[71], with different algorithms factoring into the information such as the genomic location, evolutionary conservation, searching databases for known variants and the reference-query sequence similarity to predict the functional effects of the variants. We will give the details about the Ensembl VEP in the following section, since it was selected for our projects due to advantages on the VCF file support as both input and output, being a command line tool that enables its integration into workflows and its offline option ensuring the data security.

2.1.4.1 Ensembl Variant Effect Predictor

Ensembl VEP is one of the most commonly used tools to assess the severity of the variants over the transcripts, proteins, and regulatory functions. It provides annotations on SNVs, small INDELs, and the large structural variants exceeding 50 base pairs in length. It incorporates the phenotype information and the allele frequency into its annotations for the known variants. It is an open-source tool that is well maintained with stable releases ensuring reproducibility of the results. It is developed in Perl programming language and supported with the C programming language for the components where the runtime was a constraint.

The algorithm reads the input in blocks and converts them into a variant object with the genomic location and the set of alleles information. For the VCF inputs, the variants undergo pre-processing to create Ensembl compatible variant coordinates for the unbalanced variants. Forking functionality spawns sub-processes of the aforementioned steps to ensure fast data processing whose results are combined upon the completion of the analysis. The quality control step checks for the potential errors in the data such as a mismatch between the allele length and the coordinates, or a mismatch among the reference alleles of the same position between the data and the reference genome.

Enseml VEP has multiple files stored as VEP cache, containing DNA fragments larger than 1,000 kb, which are referred to as "regions". In the analysis, the regions that are overlapping the variants are converted into objects carrying the information of the region's corresponding transcripts, regulatory regions e.g. promoters and known variants. Those regions are cached in the memory to provide a time advantage to prevent loading the same region from the disk for another overlapping variant. Each variant object is annotated with the reference and alteration bases if the algorithm detects an overlap with the transcript. The variant consequences calculated by the API through a set of functions are added to the variant annotation in standard Sequence Ontology (SO) terms. Configurations supplied by the user such as implementing plugins are applied after the consequence prediction. After the HGVS enrichment of the results, the variant objects are converted into the output.

Transcript annotation. Ensembl VEP provides a comprehensive annotation on transcript information using GENCODE or RefSeq as the prediction source of transcript isoforms[72,73]. Due to alternative splicing, one alteration allele may overlap multiple regions resulting in transcript isoforms. VEP reports all the annotations for those multiple isoforms in a single data line for a given allele. It uses multiple internal and external data sources to pinpoint the dominantly expressed transcripts[73–75] and tools to predict and prioritize their consequences[76,77].

Protein annotation. Ensembl VEP annotates the effect of amino acid changes on the biophysical features of the proteins which provides an advantage in evaluating the severity of the change, even when the alteration does not have a prior phenotype association.

VEP predicts the impact of missense variants as "moderate" which is an umbrella term implying that the change might cause alterations in the expression levels of the protein. For a more scrupulous assessment of their impact, it employs Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping (PolyPhen) as plugins to the main application. SIFT utilizes a conservation-based approach based on the assumption that the mutations observed in highly conserved regions are more likely to have a detrimental effect than the ones not highly conservative and already exhibiting variation[78,79]. Additionally, high chemical similarities between the amino acid conversions are considered less disruptive, such as the change from valine to leucine where the hydrophobic feature is kept[80]. SIFT labels the mutations as "tolerated" or "deleterious" based on the normalized position-specific probability estimations of an amino acid to be observed at a certain position[78]. PolyPhen which is another commonly used tool to assess the pathogenicity of non-synonymous SNVs incorporates both the sequence and the protein structure features into its naive Bayes classifier model. It returns qualitative labels for the variants such as "benign", "possibly damaging", "probably damaging" or "unknown"[81]. VEP also allows the incorporation of other tools such as LOFTEE to identify the alterations leading to the loss of function in nonsense, splice site, and frameshift variants[82].

Non-coding annotation. VEP provides annotation of the variants harbored in the regulatory parts of the non-coding regions. The annotations are based on the Ensembl Regulatory Build[83] which is constructed as a human regulatory region catalog using data associated with epigenomic and transcription factor regulatory elements from publicly available databases[84–86]. It also provides further prioritization of the non-coding variants through additional plugins (e.g., Combined Annotation Dependent Depletion (CADD)[87], GWAVA[88]).

In addition to the functional annotation of the variants, it incorporates information related with the known phenotypic indications[89–92], provides cross-references for the known variants[93–95], and variant allele frequencies[50,96,97] from various publicly available databases. it supports standard inputoutput data formats as well as using standardized terms such as SO and HGVS nomenclature in its annotations. Next to the web interface and application programming interface (API), it supports running the tool through VEP scripts which is powerful since it gives users to do customized configurations and it can be used as an annotation module in larger pipelines and workflows. Its offline mode provides additional data security and enables it to be a module as a part of larger applications and workflows[71]. Input and output formats include the VCF file which is the standard data format - another advantage.

2.2 Molecular Mechanisms of Cancer

Cancer is a complex genetic disease in which the acquired series of mutations derive the normal cells into malignant cells which gain selective growth advantage and undergoes uncontrollable proliferation. Large cohort studies have revealed a set of genes, known as driver genes, conferring the selective growth advantage to the cells, differentiated from passenger mutations based on their mutation frequency and predicted effects on the protein functions. For example in solid tumors from brain, breast, pancreas or colon, 33 to 66 genes are found to have mutations altering their protein product, whereas in melanomas and lung tumors, this number increases to approximately 200 missense mutations. Tumor in the tissues without self renewing property such as leukemia manifest fewer mutations than the other types[7]. It is estimated that there are up to 10 driver genes dictating tumorigenesis in lung and colorectal cancer[7,98,99]. Although the mutational load varies between different cancer types, when compared to the overall mutation rate, it is clear to see the distinction between the driver events and the accompanying passenger mutations which are forming the 99% of the mutations, tolerated and not subject to negative selection[99].

Driver genes have two distinct features. They work through oncogenes, which gains function through mutations, shows high expression profile and takes role in governing abnormal proliferation. Other mechanism works though tumor suppressor genes (TSG) acquiring loss of function, leaving the checkpoints of the cell cycle unregulated. Driver genes affect the core cellular processes which are related with the cell fate, the survival and the genome stability through repair, checkpoint and division control events[7]. The molecular mechanisms involved in core cellular processes forms the eight hallmarks of cancer[100] which are briefly explained in the following sections.

2.2.1 Sustaining Cell Proliferation

Cell proliferation is a process where the cell enters into the division cycle which involves cell growth (G1), synthesis (S) where the DNA is replicated, mitosis preparation (G2) and the mitosis (M) where two daughter cells are created[101]. This cell cycle is stimulated by the growth-promoting signals which are thoroughly controlled and regulated in the healthy tissues[100].

The most distinctive feature of cancer is the recurring and uncontrolled cell division due to the alterations in growth factor signaling leading unbalanced tissue homeostasis[100]. The chronic state of cell proliferation is acquired through distinct mechanisms such as i) the cells gaining autonomy through the autocrine secretion of growth factors[102], ii) the cells prompting the surrounding normal cells to supply growth factors[100], iii) overexpression or mutation in the surface receptor proteins resulting in hypersensitivity to the available growth factors[17]. The cells also sustain proliferation through the deregulation downstream of the growth signaling pathways via constitutive mutations, such as the hyperactivation of the Ras-ERK pathway due to the mutations in Ras protein[103], or responding oncogenes by discarding the cell senescence or apoptotic signalling cascades[100].

2.2.2 Escape from Antigrowth Signals

The extracellular space is equipped with the agents monitoring the external signals which prevents cells to advance through G1 state when necessary. While normal cells have mechanisms to revert the cycle back to the quiescence state (G0), the cycle escapes the checkpoint controllers by evading antigrowth signals. This process mostly advances due to tumor suppressor genes undergoing a loss-of-function alteration[100].

RB protein processes the signals from extracellular space. It functions as a switch for G1-S checkpoint (determines whether the cell proceeds to S-phase) by forming a complex with E2F gene in the beginning of G1[104]. Based on its activity of either releasing E2F later in the phase or not, it dictates the prospect of the cycle such as prohibiting the transitioning to S phase as a response to antigrowth signals (Figure 2.2). Its inactivation through mutations or as an after effect of disturbance in upstream regulators are highly correlated with causing neoplasms[105].

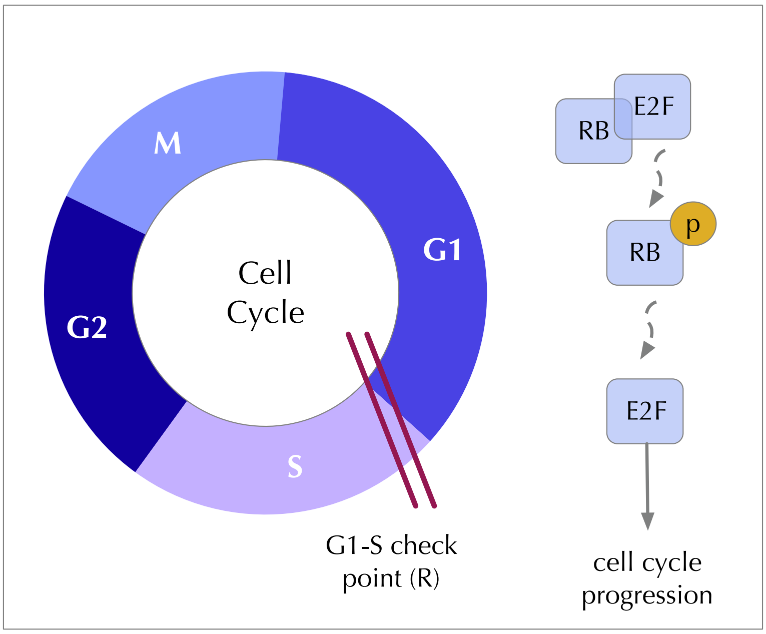


Figure 2.2: RB cell cycle regulation. The cell cycle phases are i) growth (G1), ii) synthesis (S) in which DNA replication occurs, iii) mitosis preparation (G2), iv) mitosis (M) where the active cell division occurs. RB protein forms a complex with E2F and blocks its transcriptional activity. When RB releases E2F through its phosphorylation, E2F binds to the promotors of proteins required for cell division. Deregulation of E2F is frequently seen in most cancers.

TP53 gets the signals within the cell to maintain the genomic stability. A wide range of stressor events such as DNA damage or growth stimulating signals trigger TP53 to intervene the cell cycle progression until the conditions reach an optimum state, or impose apoptosis[100,106]. The disturbances in TP53 pathway due to the inactivation of TP53 oncogene is one of the most frequently observed mechanisms in various cancer types[106], since it leaves the cells vulnerable to the stressor events by disturbing the repair mechanisms.

The cells also exhibit insensitivity to antigrowth stimulus by evading the differentiation machinery, which results in a permanent intermitotic state[17]. One of such mechanisms occur through *c-myc* oncogene. When it is overexpressed, it causes the cell to skip the differentiation and to go through the growth process[107].

2.2.3 Avoiding of Apoptosis

Apoptosis is the programmed cell death mechanisms that exists in every cell across the tissues in a quiescent form. When triggered, it eliminates the unwanted cells in response to numerous stress signals including an increased oncogenic activity and DNA damage. In the healthy cells, this process functions through transmitting the signal via ligands binding to cell surface death receptors to activate the apoptosis effectors cascade[108]. Mithocondrial pathway orchestrated by Blc-2 family proteins adjusting cytochrome-c release[109] and endoplasmic reticulum pathway activating the apoptotic program due to non-repairable damages of the organelle[110] have a major role in dictating the fate of the cell.

Apoptosis is one of the most pre-eminent program that cancer cells circumvents through various strategies. Disrupted expression levels of Bcl-2 family of proteins and inactivation of TP53 tumor suppressor gene have pivotal role in enhancing tumor cell survival[108]. Downregulation of the main effectors initiating the apoptotic pathway or breaking down the cellular structures during the apoptosis is another major mechanism of cell survival[108,111]. Impairments in apoptosis pathways due to downregulation of the death receptors or abnormal levels of decoy receptors, also result in the inhibition of the process[108].

2.2.4 Unlimited Proliferative Potential

Healthy cells go through a limited number of growth-division cycle and the majority of them reach the senescence state in which cells are stable without being able to proliferate. In vivo studies showed that the cells that defer the senescence state are observed to enter a crisis phase after additional division, which is characterized as the massive cell death of the most senescence-surpassing cells. There is a third state which characterizes the cells that escaped the senescence and crisis states and acquired the trait of limitless proliferation, which is referred as cell immortality. Both *in vivo* and *in vitro* studies showed that cancer cell populations acquired immortality[112].

Telomeres are the hexanucleotide repeats located at the end of the chromosomes functioning as protective caps preventing the chromosome end-to-end fusions and providing genomic stability[113]. In normal cells these regions get gradually shortened after every division which is a phenomenon known as "end replication problem"[114]. When the length reaches a critical point, cell death mechanism is triggered as a DNA damage response[115]. In cancer cells, telomerase (the enzyme elongating the chromosome by adding repeated tandem regions) is often found to be abnormally upregulated. While the shortening of telomeres together with the lack of telomerase function as tumor suppressor mechanism, cancer overcomes it by abnormally upregulating telomerase which elongates the telomeres and provides the cells immortality[116].

2.2.5 Aberrant Angiogenesis

Both healthy and tumor tissues provide their need of oxygen and nutrients and dispose of cellular waste products using the vascular system. In adults, the formation of new blood vessels and sprouting branches from the existing vessels (angiogenesis) are in a dormant state, except the temporary activation of angiogenesis due to physiologic processes (e.g., wound healing). However, in tumor tissues, angiogenesis is transduced from quiescent to active state creating new capillaries from the existing vasculature to provide sustenance[117].

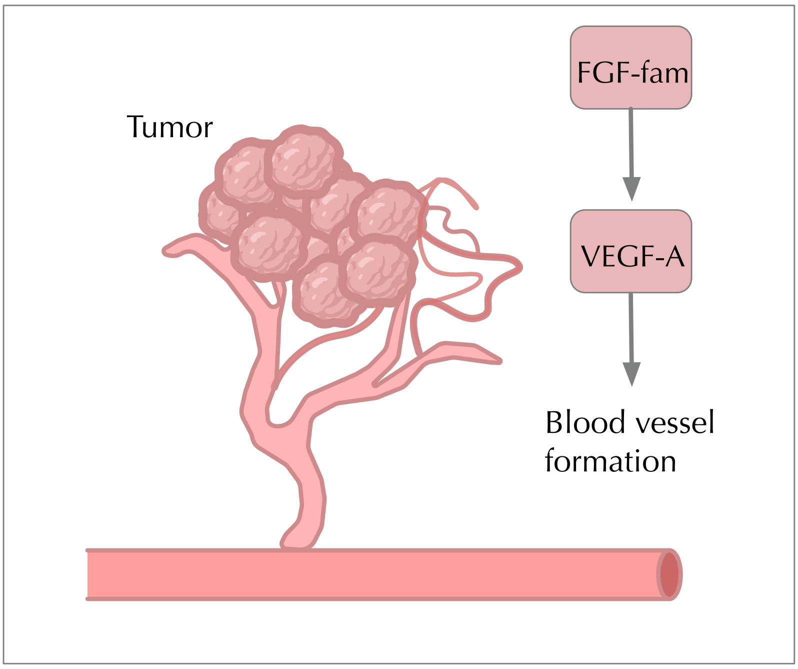


Figure 2.3: Angiogenic switch. VEGF-A is a major pro-angiogenic gene that is found to be up-regulated in various cancer types. FGF family genes indirectly activates new blood vessel formation through increasing the VEGF expression levels.

Coupling of signaling proteins with the surface receptors of the vascular endothelial cells regulates the process of neovascularization through the balance between the pro- and anti-angiogenic factors. "Angiogenic switch" shifts this balance in favor of the pro-angiogenic factors[118]. Vascular endothelial growth factor-A (VEGF-A) is a major pro-angiogenic gene that encodes for ligands regulating the neovasculature formation during the embryonic and postnatal development. Fibroblast growth factor (FGF) gene family also promotes angiogenesis through indirectly inducing VEGF expression[119,120]. Both of them are found to be up-regulated in various cancer types as a response to oncogenic signalling that enable tumors to grow through blood vessel formation[120,121].

2.2.6 Adaptations for Invasion and Metastasis

Metastasis is the process of cancer cells spreading to the distant tissues to form new colonies where the resources is less limited, which is known as a primary cause of deaths[122]. It encompasses a multistep cascade of invading the local tissues, intravasation to near by vessels, extravasation from blood stream to adjacent tissues, forming micrometastatic colonie and formation of macroscopic tumor masses[100].

Metastasis mechanisms are interrelated to the aforementioned characteristics of cancer cells with additional molecular adaptations. The known molecular mechanisms for determining invasion and metastasis phenotypes involve the genes encoding for cell-to-extracellular matrix (ECM) and cell-to-cell adhesion molecules such as integrins, cadherin and immunoglobin protein families[17]. Those molecules have various functions in maintaining tissue architecture, forming epithelial sheets, regulating cell migration and the crosstalk between cells. Decreased expression levels in cadherins found to be inducing the malignant cell migration phenotype. Integrins are also observed to be up-regulated in various cancer types[123,124].

2.3 Precision Oncology

With NGS technologies becoming widely available due to the decrease in cost and the short turn-over time, the amount of genomics data has burgeoned. Large population studies enabled researchers to link underlying genotypes to disease phenotypes and to create molecular atlases encompassing predictive and prognostic biomarkers. Rapid growth in uncovering complex molecular traits of various cancer types extrapolated to oncology clinics as an emerging medical care model termed precision oncology. It has introduced a therapeutic paradigm shift towards targeted therapies which are new drugs affecting a specific molecular target that is considered as the main factor in the disease initiation and progression. Targeted therapy strategies factor in the unique genetic make-up of the individuals to create a molecular map of the patient's mutational landscape which is used to identify the driver mechanisms and predictive biomarkers. The implementation of precision oncology follows the steps of obtaining patients' genomic data from sequencing technologies, identification of variants of known significance, coupling the genetic aberrations and the therapeutics based on clinical evidence, and providing the treating physician with the overall annotations.

2.3.1 Targeted Therapies

Malignant cell colonies emanate from the disregulations in biochemical cascades responsible for the proliferation, stress response, cell migration and extracellular communication. Functional abnormalities in tumor suppressor genes and oncogenes are found to be the fundamental cause of the biochemical disturbances, which are either directly involved in the corresponding pathways or serve as an upstream regulator. However, standard treatment strategies with antineoplastics, radiotherapy and surgery are dissonant with the complex genetic nature of the disease. The empirical treatment approaches tackles uncontrolled proliferation by introducing toxicity to the cells, killing it with radiation or removing it with surgery, as a standalone method or in combination. While providing the benefits on the shrinkage of the tumor mass, they often entail high toxicity for healthy tissue as well, lead to micrometasis or secondary tumors, contribute to developing treatment resistance and short relapse free survival times[8–10]. Targeted therapies aim to provide more efficient care by impeding the tumor growth and metastasis with therapeutic agents specialized for certain molecules, in stead of targeting a wide range of cells with rapid growth. The mechanism of action of this new generation of precision drugs works through small inhibitory molecules and monoclonal antibodies[125,126].

Kinases are enzymes that catalyze the phosphorylation of intracellular proteins which play an important role in many essential signaling cascades such as growth, proliferation and apoptosis whose deregulation is strongly found associated with developing carcinogenesis[127,128]. Thereby, they are successful target candidates for tumors whose growth depend on this specific kinase activity. Imatinib is the first example for this drug class which was approved by FDA in 2001[129]. It was proven to create a very effective treatment response for the chronic myeloid leukemia with the patients harboring a mutant kinase fusion protein, BCR-ABL. Imatinib inhibits the fusion protein's ABL domain to prevent constant ABL kinase activation[130]. The effective use of kinase inhibitors in treatment was followed by many example such as sorafenib targeting VEGFR in hepatocellular carcinoma, or gefitinib inhibiting EGFR activation[131]. Inhibition of small molecules strategy is extended to targeting the activated oncogenic pathways in the absence of its negative tumor suppressor regulator. One example is the effect of loss-of-function mutations in the PTEN tumor suppressor gene which results in the upregulation of (PI3K)AktmTOR pathway regulated by mTOR inhibitors[132].

Monoclonal antibodies are the agents engineered in the laboratory to recognize specific proteins that are involved in malignancy. They utilize the immune system by various mechanisms, such as enhancing the immune system by blocking its inhibitors, working as a tumor cell flag for detection and mimicking the immune system by triggering antibody dependent cytotoxicity[131]. This class of drugs are also effective to elude the cell crosstalk within the tumor stroma which harbors pro-oncogenic features[132], such as the effect of bevacuzimab on VEGF-A in reducing angiogenesis[126,133].

Molecularly defined treatment strategies have created a momentum in changing the histology-based clinical trial study design to the biomarker-driven trials. Similar to the targeted therapy strategies, biomarker-based oncology trials require the stratification of the population based on the shared molecular profile which involves well-defined biomarkers whose validity is already proven. Single targets are studied in basket trials, which clusters patients from various disease types harboring a specific biomarker. Umbrella trials assess a set of biomarkers from a single disease histology. Both methods provide the opportunity of investigating the molecular heterogeneity and revealing new drug-target associations which move towards closing the gap between the undruggable oncological targets and the available pharmaceuticals[134].

2.3.1.1 Network-based Approach

Pathways represent specific biological activities through the molecular interactions of their components. Directly connected nodes imply higher functional relation then the ones in the distant neighborhoods. An aberration in the upstream regulator of a specific pathway might disregulate other mechanisms due to the high complexity and connectivity of biochemical networks. In the literature, module detection algorithms are implemented to examine the systematic effect of cancer genes to identify novel driver genes, synergistic drug combinations and driver pathways which are thoroughly reviewed in Ozturk et al.,2018.[135].

The therapeutic potential of targeting driver genes has revolutionized the cancer healthcare. However, the alternative tumor mechanisms were found in the unresponsive patients receiving this new line of treatment. For instance, while in melanoma BRAF V600E inhibition has a high response rate, patients with colorectal cancer were resistant to the therapy since they exhibited EGFR activation[136]. Similar studies pointed out the fact that cancer mechanisms are highly heterogeneous and the functional impact of somatic mutations are transmitted to the collateral gene products[137]. Additionally, the focus on the accumulated effects rather than the single causative gene enables new treatment strategies for patients who lack actionable targets, via targeting essential pathways such as the cell cycle regulation cascades[138,139].The molecular interplay of network members is used to estimate the therapy efficiency, to identify driver modules, to assess therapeutic response and to select combination therapies[135]. The hypothesis that the genes that are outside of the core disease pathways also contribute to the pathogenicity in polygenic traits (termed as omnigenic model) is utilized to project the mutational profile to interaction networks (i.e. protein-protein interaction, pathway-pathway interactions) and implement a scoring schema to extend the affected nodes based on the parameters such as proximity to candidate genes, mutational burden, similar biochemical properties in making inferences.

Network-based methods have also been adopted in population studies to accurately stratify patients into cancer subtypes. It enables further clustering of patients who manifest a similar phenotype with molecular differences such as having the same core pathway activated through different mutant genes[135,140]. Identifying the sub-networks (aka modules) relies on the assumption that the genes causing similar phenotypes are likely to interact with each other through the involvement of the similar biochemical mechanisms or harboring variants linked with the same disease. Network-based stratification (NBS) method implements the similar phenotype-connected node assumption by projecting the candidate genes to the interaction network and iteratively propagating the network to the near by nodes based on their gene proximity score until convergence[141].

2.3.2 Pharmacogenomics

Pharmacogenomics (PGx) is the study of the correlation between an individual's genomic make-up and their response to a drug. Its sub-field pharmacodynamics (PD) concerns with the effect of mutant targets on the response, whereas pharmacokinetics (PF) relates how the variation in the absorption, distribution, metabolism, or elimination (ADME) genes influences the effect of a drug[142].

The underlying molecular complexity of cancer renders it crucial to depend its therapeutic strategies to PGx. Molecular differences of the same organ and histology malignancies define the disease subtypes requiring the precision in target-drug selections. The prognostic estimation of the selected treatment depends on the genetic variation assessment methods similar to GWAS studies, where molecular case control analysis is conducted to find significant variants on the drug response, dose adjustment or adverse effects. The known associations from medical publications are publicly available and accumulated in many databases (e.g. the Pharmacogenomics Knowledge Base (PharmGKB)[143], Clinical Interpretation of Variants, in Cancer (CIViC)[144], Cancer Genome Interpreter (CGI)[145], OncoKB[146] and MyCancerGenome[147]) which provide the data integration service, standard ontological representation of variants, curated datasets with the association significance level score.

The PGx effect does not only depend on the monogenic cause for most cases, but rather is manifests as a cumulative effect of the disrupted genes on the oncological pathways (e.g. the combinatorial effect of EGFR and KRAS mutations EGFR inhibitor cetuximab response which is rendered unresponsive due to the KRAS involvement in the EGFR pathway[148]). The identification of such mechanisms requires mapping the mutational profile to associated pathways to infer the level of pathway disruption.

In clinical trials, it is not always possible to factor in the effect of rare variants since they are observed in less than 1% of the population. However, it is known that the rare variants in the proximity of the targets involved in mechanism of action (MoA) alter can the drug response. The assessment method in such cases builds upon the *in silico* tools explained in Section 2.1.4 to predict the variant impact and match their profile with the list of MoA genes. Some publicly available databases such as DrugBank provide the list off molecules involved in the drug mechanism[149].

The molecular knowledge from PGx studies continually increases with additional methods of non-coding region variants effect prediction and the 3D space simulations to pinpoint the broken mechanism with protein docking and chemical molecular dynamics. However, due to the complexity of these methods, they are rather research questions than the direct methods involved in clinical routine. But the results might be used in the clinical decision making process through the publicly available databases curating the corresponding publication, which again emphasis the overall clinical workflow PGx studies which is the sequential step of impact prediction, variant mapping to existing clinical evidence through databases and interpreting the prognostics.

2.3.3 Clinical Applications

Molecular tumor boards (MTBs) are the interdisciplinary advisory bodies for precision oncology in larger clinics. Their clinical workflow revolves around two major phases of case preparation and discussion. Case preparation involves collecting and interpreting the molecular data to create patient-specific reports which are presented to the committee in the discussion phase. After a patient is referred to the personalized oncology program by the treating physician, a clinical coordinator manages the patient consent and collects all necessary clinical documentation. Typically the pathology unit examines the biopsy sample to diagnose the tumor type and evaluate tumor content. The tissue block with high tumor content is sent to a sequencing facility for sequencing and bioinformatics analysis. The hospital obtains the results containing the patient's genomics sequence together with the mutational signatures. The case preparation team implements clinical annotation which is the process of interpreting the reported aberrations in terms of the clinical actionability and druggability. A variant is selected as actionable if it i) predicts the prognosis of a particular drug, ii) regulates a cancer gene which can be targeted directly or indirectly, iii) is an enrollment pre-condition for open clinical trials, iv) is known for causing adverse effects, v) increases the cancer susceptibility and subject to preventive therapy[150]. The team refers to the literature and the publicly available data sources briefly mentioned in the previous section to couple the relevant variants with their targeting drugs and prioritize based on the significance level of their variant association. As a result, a case-specific molecular report is prepared together with the clinical data (e.g. treatment history, diagnosis, patient demographics). As a last step, the MTB, consists of specialists from various fields, discusses the options and suggests the optimal therapy strategy based on the information reported[11].

In 2012, Moores Cancer Center identified a small cohort of 12 patients who exhausted multiple conventional therapies and found eligible for targeted therapy. They reported partial response for three patients who were resistant to their prior therapies. The remaining patients could not be administered with the therapy due to logistic, insurance, or lack of actioable problems[151]. This modest number of clinical success was scaled up to larger cohorts. The same center applied targeted therapy to 87 patients among a cohort of 347 and reported longer progression free survival and overall survival rates[152]. The Danish Renal Cancer Group reported increased overall survival times of 744 metastatic cancer patients[12]. Many other retrospective cohort studies utilizing molecular patient stratification reported successful clinical implementation[22,24,153–156].

Two decades of precision oncology practices pinpointed the challenges of the field. While demonstrated utility was criticized of selecting the most suitable patients, the requirement of exhausting conventional therapies was claimed to lower the efficiency of the method. Additionally, the lengthy turnaround time to retrieve the mutational signatures[151], not having a standardized system of analysis leading to non-reproducibility of the results[25], and the labor intensive, mostly manual and thus error-prone process of actionability assessment challenged physicians to integrate the robust precision oncology workflows to the clinical routine.

2.3.4 Data Requirements of Precision Oncology

Efficient implementation of precision oncology necessitates a framework for standardizing the vocabulary and minimum information required to report the somatic variants. Clinically relevant data being scattered over many publicly available databases each with a different database model exacerbate the data integration efforts. Additionally, the lack of ontology use creates discrepancies between the vocabulary of similar

Allele Descriptive Fields	
Information field	**Details**
Assembly version	GRCh37 or GRCh38
Chromosome number	Number or letter representation of the chromosome of the variant
Variant position	DNA position in HGVS format
Transcripts	RefSeq identifiers of all possible transcripts
Proteins	RefSeq identifiers of proteins

Table 2.1: Complete list of required allele descriptive fields.

data fields from different sources which hinder standardizing the information to be reported.

Major cancer data stakeholders such as Clinical Genome Resource (ClinGen), ClinVAR, American Society of Clinical Oncology (ASCO), Global Alliance For Genomics and Health (GA4GH) created a consensus on reporting clinical relevance of somatic variants[157–159] to improve interoperability. Standardized representation is organized on descriptive and interpretative categories based on the information represented. *Allele descriptive fields* mostly includes allele-specific coordinate and identifier properties (e.g., the chromosome number and variant start position on DNA). *Allele interpretive fields* represent the functional classification of the variants such as the variant class, the predicted consequence and the sequence alterations. *Cancer interpretive fields* contains the clinically relevant information of the variants such as the therapy relevance and drug response[160]. The entire list of requirements are given in Table 2.1-Table 2.3.

Allele Interpretive Fields	
Information field	**Details**
Variant category	somatic, germline, unknown
DNA substitution and position	HGVS format
Protein substitution and position	HGVS format
Variant type	SNV, INS, DEL, multinucleotide variant (MNV)
Variant consequences	Sequence Ontology (SO) terms, i.e. Nonsense, missense, frameshift
Supportive publications	PubMed identifier

Table 2.2: Complete list of required allele interpretive fields.

Cancer Interpretive Fields	
Information field	**Details**
Diagnosis	International Classification of Diseases (ICD) cod
Biomarker type	Prognostic, diagnostic, predictive
Drug association	FDA approved drugs, National Comprehensive Cancer Network (NCCN), DrugBank
Drug response	Resistant, responsive, not-responsive, sensitive, reduced sensitivity
Evidence level	There is not a consensus on the level stratification.
Supportive publications	PubMed identifier

Table 2.3: Complete list of required cancer interpretive fields.

Chapter 3

Targeted Therapy Identification in Precision Oncology

The content of this chapter is an extended version of the article:

Sürün, B., Schärfe, C. P., Divine, M. R., Heinrich, J., Toussaint, N. C., Zimmermann, L., ... & Kohlbacher, O. (2020). ClinVAP: a reporting strategy from variants to therapeutic options. Bioinformatics, 36(7), 2316-2317[161].

3.1 Introduction

Understanding a patient's genetic-molecular profile to assess clinical actionability is a key to establishing a working model of precision oncology where the ultimate goal is the selection of patient-specific molecular target(s) with the evidence of treatment response to a cancer drug. In clinics, this assessment is conducted by MTBs who make case-specific decisions based on a set of therapy informing biomarkers obtained from the patient's genomic data. In this respect, extracting the list of genes that have a major contribution to the disease progression is essential for therapeutic decision-making as well as the identification of variants that are amenable to drug treatment or possibly conferring treatment resistance.

The decreasing price of NGS technologies combined with its critical role in precision oncology has resulted in large amounts of genomics data. While elucidating links between phenotypes and underlying molecular causes, it came with an additional cost of data analysis complexity. The arduous search for clinical significance requires annotation methods to extract major molecular players which would form a basis for MTBs for therapeutic decision making.

The efforts in understanding tumorigenesis led to the discovery of cancer driver genes which provide a selective advantage to the cells resulting in malignancy[162]. Various strategies have been applied to big cancer cohorts to capture positive selection signals of genes and resulted in driver gene catalogs[7,94,163–167]. Additionally, the advancements in the PGx field revealed the underlying causes of inter-individual variation in drug response which utilizes the genomic data as an indicator to estimate treatment success. Unveiling the promoting effect of driver genes and the clinical evidence on drug response leveraged the potential of targeted therapies. However, the major drawback in its application lies within the same system that equipped us with the knowledge in the first place: information is scattered over many different sources. It is very time-consuming to query those sources manually. Moreover, sending patient-related information to the web services for such annotations creates data privacy issues and thus hinders the use of services such as PharmGKB[168] and Cancer Genome Interpreter (CGI)[145]. On the other hand, variants have to be examined within the context of the severity of the observed mutation, which requires employing tools[70,71] to predict the potential effect of the variants on the cell functioning. The local instances of such functional annotation tools can be difficult to use due to often complex command-line interfaces. All these complexities impede the use of annotation tools in the clinical routine and require fast and robust data analysis pipelines that automatize the arduous task of variant annotation.

In 2018, Perera et al. developed a local tool that generates evidence-driven reports of treatment options from somatic variants. They assembled a PGx dataset from GDKB, CIViC, and TARGET as the main source of clinical evidence on drug-gene associations[169]. Their tool eases the therapeutic search by automatically matching genes to this dataset and provides a ranked list of suggestions based on the strength of drug-variant associations and their relatedness to a given tumor type. Nevertheless, it skips the assessment of gene disruption levels and lacks standard rules to filter them based on their predicted consequences. Hence, it requires input pre-processing prior to the analysis. Moreover, it has shortcomings in its implementation due to the difficulties in installation caused by package dependencies and reproducibility due to lacking consistent versioning. In overall, it fails to encompass the entire analysis framework and software requirements.

CGI is another source to annotate the cancer genome. It has a rich database involving a catalog of known and predicted driver genes compiled via OncodriveMUT, and molecular actionability gathered from publicly available sources[145]. Although it provides a complete clinical annotation, it does not support a standard VCF input file and thus requires additional input preparation. Not having a publicly available local instance is another disadvantage since it requires transferring sensitive patient data to their web servers and stores the analysis results for six months, which violates many clinical data protection guidelines. And lastly, it gets variant impacts by matching their coordinates to the database entries, which could result in overlooking rare variants in the dataset. oncoPDSs is another web-based tool similar to CGI, with the focus on emphasizing the expected outcome of variant-drug associations[170]. Even though it has a variant effect prediction mechanism, the aforementioned data security and lack of reproducibility concerns are also pertinent for it.

Swiss Molecular Tumor Board (SwissMTB) created a comprehensive workflow covering each step of annotation from variant calling to report generation by systematic variant prioritization. Although their workflow fulfills the needs of MTBs, it is a standard operating procedure (SOP) rather than an automated pipeline since report generation solely relies on manual work for obtaining gene-drug associations and assessing their therapy relevance[171]. Other disadvantages of their system are not systematically reporting driver genes, and leaving the relevance of genes to the clinicians' judgment on a mutated gene frequency chart. Since it relies on manual work, consistent versioning is not ensured which creates a lack of reproducibility because of the dynamic content of annotation databases due to updates. They also lack the generalizability of their evidence levels since it is specialized in Swissmedic-approved drugs.

This chapter introduces Clinical Variant Annotation Pipeline (ClinVAP) which is an automated pipeline to create patient-specific reports based on their mutational profile. It processes SNVs to extract functionally significant variants and augments the variants with a user-provided list of CNVs which are then annotated by driver gene status and druggability. ClinVAP is available as a fully containerized, self-contained pipeline maximizing reproducibility and scalability allowing the analysis of large-scale data. It works with standard data files and eliminates an additional input preparation step. The resulting JSON-based report is suited for automated downstream processing, but ClinVAP can also automatically render the information into a user-defined template to yield a human-readable report.

3.2 Design and Implementation

ClinVAP is intended as a self-contained pipeline with one-way data flow. Its target audience are clinical practitioners with an MTB and this audience also shaped the design choices. Custom scripts are written in Python 3.0 and the pipeline is implemented in Nextflow.

3.2.1 Design Concepts and Principles

The main factor defining users acceptance of a decision support tool in health care is the ease of use since the target users generally are not trained in bioinformatics. We built our pipeline around this principle by mainly focusing on two important aspects. First aspect was to ensure simple and smooth installation. Second aspect was to make it a simple command tool, so that running it would not be complex and prone to user mistakes. Another decision that contributes to this purpose was to limit the number of parameters that are required from the user. To limit the parameter space, we used widely accepted standards on functional and clinical annotation of the variants. We also created extensive documentation and a clear step by step guide to run the pipeline.

Discussions with MTB members revealed that the key requirement was to automate the analysis at the highest level and to produce a complete case report that is broad enough to not to leave any significant therapy-related information out and yet be as concise as it can be. Although the latter is mostly related to the data analysis content, it indeed pointed to the key functionality which is to produce one complete case report via automated processing. To achieve maximum level of automation we limited input types to standard data formats (VCF for SNVs, TSV for CNVs, ICD10 code for diagnosis), which enables users to directly channel the data from sequencing centers into the pipeline. As an addition to machine-readable case reports, we produce the output in a human readable form (PDF and Microsoft Word DOCX). Working with the standard data formats also increases the pipeline's level of interoperability.

Another main design principle was reproducibility. It represents the software's ability of producing same results in different operating systems with the same source code and data. It is especially important for clinics to produce same case report in different analysis runs, and backtrack the source of the information.

Bioinformatics tools have two main aspects that hamper reproducibility. The first issue is mainly applicable for the tools which rely on publicly available databases in their enrichment and annotation steps like ClinVAP. Due to the immense acceleration in the field, those data sources are frequently updated in a way that they either include more entries or have more curated results. This emerges the need of storing a snapshot of the background sources in order to get the same results in a future execution. We solve this problem via an integrated background knowledge base, in which we fixed versions of different sources, and provided knowledge base versioning that is maintained as GitHub releases.

The second issue is related with the effect of configuration, dependencies and the operating system on software installation. Not providing a fixed environment for the software with all the dependencies cause problems from installation of it to producing identical results. A general solution is the usage of container technologies such as Docker and Singularity which enables the shipment of the code in its suitable operating system with configured environment and installed dependencies with fixed versions. To ensure reprodicibility, our first strategy was to create Docker and Singularity versions of our pipeline. High-performance computing (HPC) clusters do not support Docker due to its security vulnerability of having root rights from the Docker daemon. Although Singularity does not have this specific feature, it did not support image orchestration at that time. Our efforts to create one version of the core pipeline with both Docker and Singularity resulted in two different architectures and raised maintainability issues since both version required their own configuration and standards. The issue became more problematic with the developments of the second version of the pipeline, since it became more complex and required the separation the main code into different processes based on their functionality. The solution which will ensure reproducibility, supporting the process based implementation and enabling reproducibility and maintainability, was found in NextFlow framework.

3.2.2 Architecture

ClinVAP is implemented as a NextFlow framework (Figure 3.1) which forms a cohesive pipeline from the individual tasks communicating through input/output channels[172]. Reproducibility is ensured via NextFlow's compatibility with Docker, an image containing the conda environment with pre-installed dependencies in which the NextFlow script is executed. Automated integration between the GitHub code repository and Docker Hub is set via Docker Hub's continuous integration feature which ensures the automated build of the Docker image subsequent to GitHub code commits. The scripts of the processes except for the ones employing the Ensembl VEP tool are written in Python3. The clinical annotation process encapsulates three modules responsible for

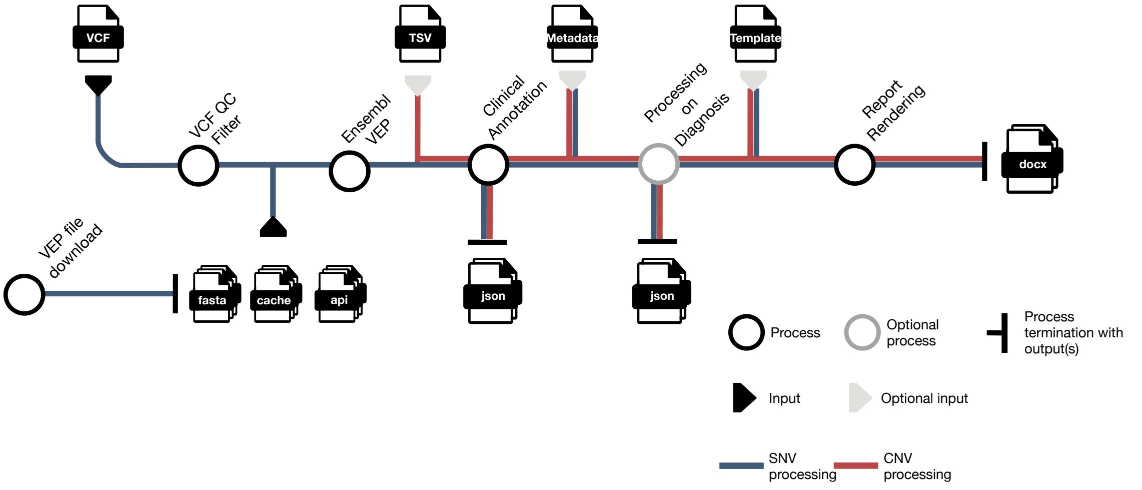


Figure 3.1: ClinVAP implementation in Nextflow. The processes are channeled linearly through their input/output files. Ensembl VEP depends on the cache and api files. If they are not provided in the working directory, "VEP file download" initiates the downloading process. The main input of the workflow is the VCF file containing SNVs which is passed to "Clinical Annotation" after being pre-processed. The processes "Processing on Diagnosis" is triggered only when the ICD code of the diagnosis is given in the metadata file (in JSON format). The human readable case report rendering relies on the default report template if a customized template is not provided by the users. In the case of the presence of the CNVs (as TSV file), the clinical annotation applied to those separately and two reports one for each generated as a result.

processing the annotated input types, querying the knowledge base, and creating the table contents. The input and output types of every process are shown in the Figure 3.1 pointing out that the data flow is mostly linear and the pipeline builds up the results based on the process dependencies. Inputs and the outputs are managed in a way to support the highest level of generalization which is the reason for choosing the widely used standard data formats. The entire pipeline is available at https://github.com/KohlbacherLab/nextflow-clinvap under MIT license.

3.3 Materials and Methods

3.3.1 Clinical Annotation Knowledge Base

We created a clinical annotation knowledge base (KB) an annotation source of the pipeline which is queried by the reporting application for each variant and disrupted gene observed in the sample. It is designed to integrate various publicly available databases and cover several annotation categories (e.g., mechanistic drugs, PGx, and adverse effects) each contributing to a different layer of the informed decision-making process with various specificity, as summarized in Table 3.1.

3.3.1.1 Data integration

The KB was constructed using the complete set of 42,596 genes sourced from HGNC and UniProt databases[164,173]. This number surpasses the approximately 21,000 genes reported in the neXtProt release[174], as the genes from HGNC and UniProt do not consist of only protein-coding genes.

The genes were annotated with driver genes, PGx information, mechanistic drug targets, and adverse effects. Each annotation is enriched with its source name and source id to keep track of the origin of the information. While integrating data from different background sources, gene names were normalized by converting their corresponding identifiers to HGNC id and HGNC gene symbol.

Data structures and the annotation content show differences from source to source. To unify the data model with the least amount of missing data, we implemented integration strategies including information extraction with cross-referencing among different sources and processing text fields for keywords and notations.

Table 3.1: Knowledge Base Sources

(a) Driver Gene Source

Data Source	Version
Literature	Vogelstein et al.
Intogen	2014.12
UniProtKB	2020.02
COSMIC	v90
TSGene	v2.0

(b) Mechanistic Drugs

Data Source	Version
Literature	Santos et al.
TTD	7.1.01
IUPHAR	2017.5
DrugBank	5.1.4

(c) Pharmacogenomics Effects

Data Source	Version
CGI	2018.01
CIViC	2019.11
DrugBank	5.1.4

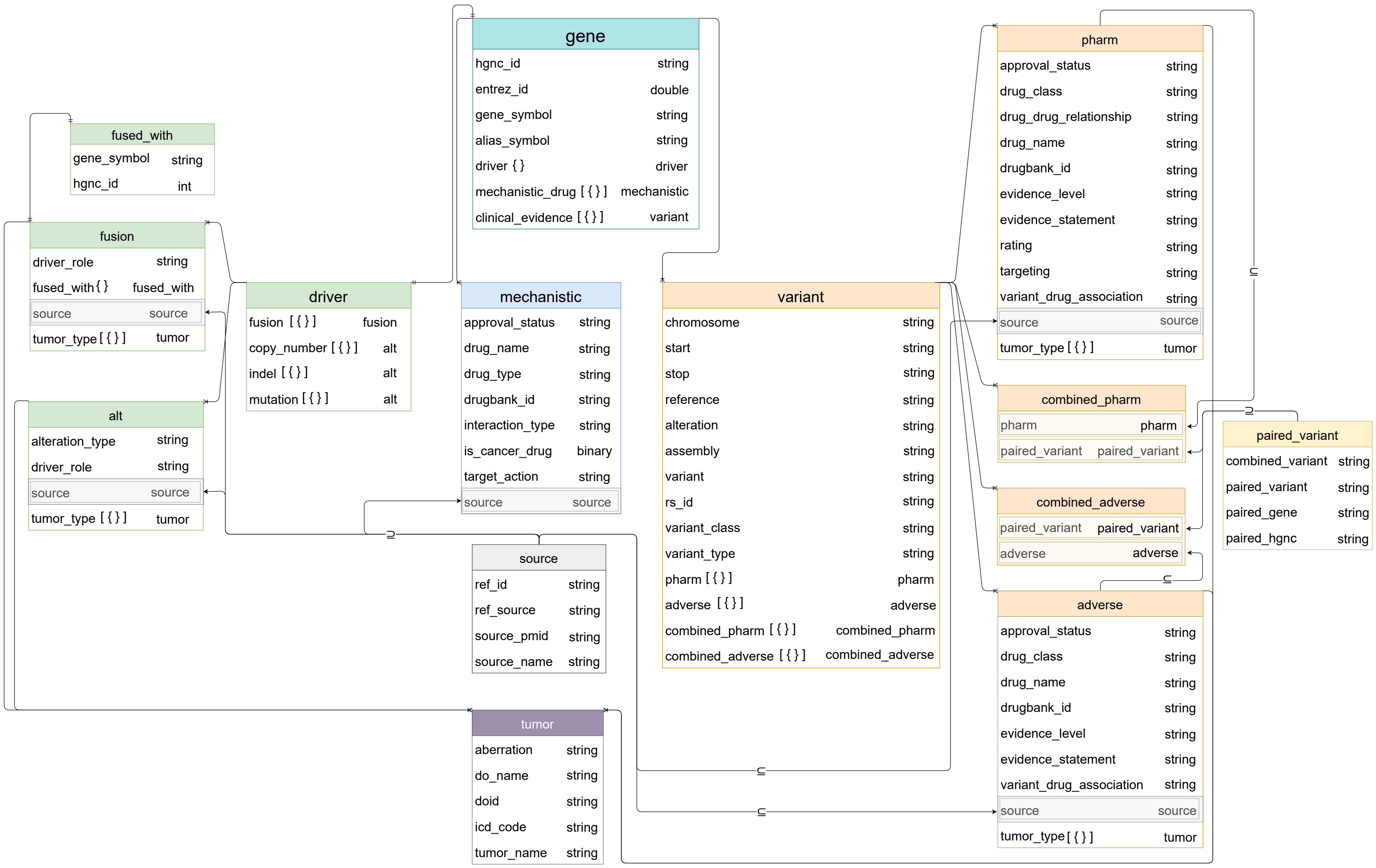


Figure 3.2: Knowledge base schema. It is designed as one collection with embedded documents. For readability, the common data fields were not repeated but presented as a subset of embedded or array of embedded documents. Embedded documents have one-to-one relationships and the array of embedded documents is represented with one too many relationships to their superset document.

3.3.1.2 Driver gene catalog

A driver gene catalog was assembled from literature and publicly available databases, TSGene2.0, COSMIC v90, UniProtKB, IntoGen (Table 3.11) [7,94,163–165].

To obtain the driver gene catalog from UniProt, we queried UniProtKB database using the keywords "tumor suppressor" and "proto-oncogene". We only selected the results if they were curated and marked as reviewed. The corresponding disease information was also included in the results. We obtained the mutation and cancer types by parsing the disease involvement column using a set of keywords. Then we finalized the disease information by manually curating the initial list.

The information contained in the catalog denotes the driver gene type i.e., oncogene, tumor suppressor gene (TSG), or Oncogene/TSG. To ensure precise querying on the driver gene catalog, we further processed corresponding data columns to label the mutation categories as SNV, CNV, and FUSION. The same strategy is applied to retrieve the cancer types in which the genes were identified as drivers. As a final step, the information from these sources was merged after gene name normalization using HGNC id conversion via HGNC gene symbols, Entrez, and UniProt ids.

3.3.1.3 Clinical evidence catalog

PGx information is collected from Cancer Biomarkers Database (CGI, 2018.01), CIViC (2019.11), and DrugBank (5.1.4) (Table 3.1c). Data integration required data specific pre-processing before merging information from different sources, to standardize the content and data structure.

Evidence score mapping: For variants with documented effects, we adopted a scoring schema to provide a metric that shows the confidence of the provided association. We used CIViC's evidence level schema as a base and extended it to cover the associations coming from clinical guidelines[144]. We conducted evidence-level mappings for CGI based on the schema given in Table 3.2.

Genome assembly mapping: Even though the most recent human genome assembly is GRCh38, both GRCh37 and GRCh38 assembly versions are still in use. Thus, not all publicly available databases are up to date with the most recent version of the human genome assembly. In order to provide clinical annotations for the sequencing data coming from both assemblies, we converted the variant coordinates to GRCh38, if it was not available.

Evidence Level Mapping		
Level	Name	Definition
A	Validated association	Proven association or association with medical consensus
B	Clinical evidence	Association proven by clinical trials or with other primary patient data
A/B	Clinical guidelines	Associations given by the clinical guidelines such as FDA, NCCN, CPIC etc.
C	Case study	Associations given by the individual case reports of clinical journals
D	Preclinical evidence	Associations supported by in vitro or in vivo experiments
E	Inferential association	Association is inferred but not directly measured

Table 3.2: Evidence level schema used in cross reference evidence mapping.

For CGI and CIViC, we used NCBI ReMap to convert the genomic coordinates from GRCh37 to GRCh38[175]. We applied a two-step control of the conversions to ensure its robustness. First, we obtained the reference base of the newly converted coordinates using samtools faidx and compare those with that of GRCh37. Second we checked whether the new coordinates were mapped to the same genes as in GRCh37. Since the variant coordinates were not given in DrugBank database, we mined them from RefSeq database for both assemblies using the corresponding RefSeq identifiers. For the RefSeq ids representing more than one type of base change in the same region, extraction of the coordinates was made by matching the base changes reported in DrugBank with the HGVS notation of interest.

Further data standardization: We made the drug response vocabulary consistent between the resources. For better categorization of the data, we assigned the variant class to each entry. This information was either obtained from RefSeq database or assigned by linking variant consequences to variant classes using the SO hierarchies[176]. We separated PGx effects from adverse effects. We re-structured the data entries pointing to a combined effect of more than one variant. We pre-filtered CIViC data based on variant origin (somatic), evidence status (accepted), and evidence direction (supports). As a final step, we merged the processed data to create separate data files for pharmacogenomic effects and adverse effects.

3.3.1.4 Drug mechanism catalog

We further annotated genes with drug target data compiled from DrugBank 5.1.4, Therapeutic Target Database (TTD), IUPHAR Guide to Pharmacology 2017.5, and the manually curated dataset by Santos et al. (Table 3.1b)[177–180].

Data-specific pre-processing steps were required to put the data into a standard form. We extracted a cancer drug catalog from the sources by processing their drug type information fields. Cancer drugs were identified from 1) DrugBank by processing ATC codes, then by searching the keywords ’immunotherapy’, ’antineoplastic agents’ or ’Lutetium’ (a radioligand therapeutic agent that is not listed under the other keywords) in the drug category if an ATC code is not available; 2) TTD by processing ATC codes, then by searching drug-disease links given in the database, for a defined set of cancer vocabulary to identify drugs associated with cancer even when their ATC code is absent; 3) IUPHAR by searching drug-disease links retrieved via IUPHAR API calls, for a pre-defined cancer vocabulary to associate drugs with cancer; 4) SANTOS by cross-referencing drugs with the ones contained by DrugBank to retrieve cancer drug status since none of the information to extract such association was given in the data file.

We standardized the drug approval status among the sources. DrugBank’s approval status did not need any pre-processing and its vocabulary was taken as standard. TTD provides more than one approval status for a drug due to its different indications. For such cases, the highest approval status was used unless the drug has a withdrawn or preclinical label for any of its indications. In this strategy, a false negative is preferred over a false positive in case of ambiguity of the approval labels. IUPHAR’s and TTD’s approval status vocabulary is standardized to have the same notation among the data sources. SANTOS did not provide approval status; hence, we completed this field by cross-referencing between sources using DrugBank ids, drug names, or drug synonyms.

After bringing the data into the same format from different sources by completing information fields, assigning DrugBank ids to other sources if possible, and gene identifier normalization, we merged them and created a mechanistic drug targets table.

3.3.2 Variant Annotation

We annotate the variants to reveal their functional effects and the clinical evidence on their contribution to treatment response.

3.3.2.1 Functional variant annotation

We implemented a functional variant annotation step to predict the potentially damaging effects of somatic SNVs. Prior to functional annotation, the pipeline filters the variants which did not pass the quality control measures of its NGS pipeline. Prefiltering provides an advantage in reducing the size of the data; thus, decreasing the time required to annotate the file. We use Ensembl VEP v95[71] to annotate the remaining variants and calibrated the tool to run offline to ensure data security by restricting input data from being sent outside of our local application.

In addition to obtaining the severity of the variant effects from Ensembl VEP, we predict the functional effects of variants on the canonical transcripts using SIFT and Polyphen[181,182]. We then pass the results to the clinical annotation step for further processing.

3.3.2.2 Clinical variant annotation

In this step, the descriptive and interpretive information on variants such as DNA position, variant consequence, etc., are retrieved after parsing the VEP-annotated VCF file for SNVs[160]. Among the variants annotated with more than one functional effect due to alternative splicing possibilities in each region, we choose the one that is most damaging. We excluded the variants that are predicted as non-coding and low effect (harmless) by VEP. Furthermore, we set additional filters to remove the variants predicted as "tolerated" or "tolerated low confidence" by SIFT and "benign" by PolyPhen. We also included a step to calculate variant allele frequencies (VAF) of the SNVs if they are from Strelka variant caller, as described in version 2.9 user guide[183]. Data from other variant callers is not subject to the same calculation due to the non-standard VCF fields and the lack of documentation. We query the KB, with the remaining list of SNVs and the user-provided list of CNV, at variant and gene levels using base coordinates and HGNC identifiers to reveal driver genes and gather clinical evidence on the actionability of those variations.

As a part of clinical annotation, we convert "evidence level" to a "match level" indicator, which is an extension to the evidence schema with the predicted variant consequence match. It ranks the level of matches between the observed variant and the KB results. When there is not an exact match in the KB for the given variant, we search for other variants of the same gene and assign priorities to evidence levels based on the functional effect of that variant. While level 1 represents an exact match for the variant, level 2 depicts different variants, same gene, same consequence, and level 3 points to the different variants, same gene, different consequence. For CNVs, we create the same ranking through the CNV type, i.e., amplification or deletion. In this way, we rank the KB results based on their compatibility with the observed variants and the directness of their potential impact on the treatment response.

We equipped the pipeline with an optional step to further process the data based on diagnosis, which provides an advantage in particularization of the KB results for a given cancer type. Since disease names used in our KB data sources do not follow a standardized ontology, any data processing attempts based on diagnosis require either standardization or a name matching strategy on the disease keywords. We created a hybrid approach, which searches for exact matches and if there is no exact match, calculates a similarity score between nonstandard disease names and standard ICD10 codes based on their common features (e.g., organ type, system, and histology). To achieve this, we created a KB diagnosis look-up table 1) by mapping ICD10 codes to disease names when it is possible, 2) by dissecting the disease terms into main disease features. We created an ICD10 code look-up table by applying the same dissection to its vocabulary. If the KB disease term equivalent of the ICD10 code of interest is not found in the KB look-up table, the similarity score between the ICD10 code and the KB disease terms is calculated by dividing the number of common features between them by the number of total features.

Particularization of the results can be done in three ways. The first option is to filter clinical annotation results that do not belong to the diagnosis of interest. The second option is to sort those results based on the diagnosis similarity score. The third option is to show all the results for evidence levels A, B, and C, and only show the results that belong to the same diagnosis for evidence groups D and E.

3.3.3 Report Generation

We structure the results into categories based on their context, e.g. PGx and mechanistic drugs, and the specificity, e.g. drugs directly targeting the observed variant or targeting another variant of the mutated gene. We use JSON structure to store the reports in machine-readable form. Since human-readable formats are preferred in clinics, we render the *.JSON* into a *.DOCX* template using python-docx-template library[184] that creates the document through Jinja2 package tags and filters in the provided word template. The pipeline is configurated to accept the template as one of the optional parameters which enable users to customize the report based on the preferred structure and corporate design elements. For the simplicity of the generated reports, we create separate reports for SNVs and CNVs through the same process.

3.4 Results

3.4.1 ClinVAP Overview

We devised a fully automated pipeline which takes SNVs and CNVs of a patient as input and creates evidence based patient specific reports as output (Figure 3.3). The multistep process builds on Ensembl VEP for functional SNV annotation. It is followed by a clinical annotation step constructed on processing the query results obtained from a knowledge base using disrupted genes and user provided list of CNVs, to reveal existing clinical evidence for therapeutic strategies. Final results are rendered into a reporting template. The whole pipeline is available as a NextFlow workflow on GitHub https://github.com/KohlbacherLab/nextflow-clinvap

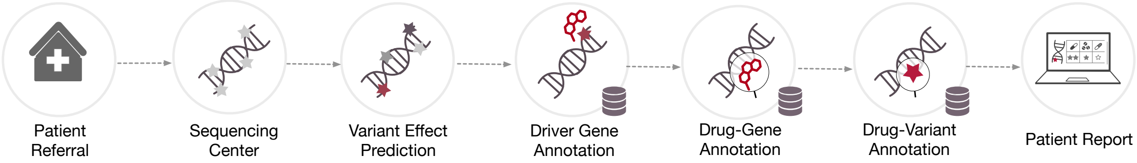


Figure 3.3: ClinVAP overview in the clinical setting. The targeted therapy cycle starts with the patient's referral to the MTB program. ClinVAP processes the sequencing data through variant effect prediction, driver gene annotation, actionability/druggability of disrupted genes and the observed variants, structuring and rendering the results into a final case report, consecutively. The resulting report serves as the pre-MTB case preparation which then is used as a major resource in the MTB discussion.

3.4.2 Clinical Annotation Knowledge Base

We integrated in total 11 publicly available data sources and created a clinical annotation knowledge base. The resulting knowledge base consists of three main information levels for each gene which are driver gene annotation, mechanistic drugs and PGx information including adverse effects. The knowledge base created as one document encapsulating embedded documents of each information category in JSON structure which is compatible with MongoDB.

Driver gene catalog:

We integrate the sources Vogelstein et al., IntOGen, UniProt, COSMIC, TSGene and created a driver gene catalog which includes the details of the driver genes such as driver role, mutational class, tumor type and the source of information. Figure 3.5 represents contribution of each sources to the catalog. It points out that each of our background sources has an essential contribution to the KB. The sources contain different amount of information for the same variant class. In addition, some sources have specialized in specific classes, for instance TSGene covers mostly the content for SNV driver genes while COSMIC contributes mainly to the structural variant class of driver genes. Moreover, the sources do not have a large intersection apart from the content of Vogelstein which is due to the common use of it in performance assessment of the new classification tools.

Integrating those 5 complementary sources resulted in a comprehensive catalog of 1,727 SNV driver genes, 102 CNV driver genes and 429 fusion driver gene pairs. The total number of genes that are classified as driver for at least one of these variant classes is 1,883.

Drug mechanisms: We integrated the data from Santos et al.[180], TTD[178], IUPHAR/BPS Guide to Pharmacology[179], DrugBank[149] and created a mechanistic cancer drugs catalog. It includes the details of the mechanistic drugs such as approval status, drug type, target action and the source of information. Our catalog includes 2,626 cancer drugs/compound in total, that are known to target at least one gene mechanistically. Among those cancer therapeutics, there are 324 drugs approved by a regulatory agency for at least one drug indication. The amount of drugs which are in clinical and preclinical trial phase are 1,868 and 454, respectively. Figure 3.4 shows the contribution of each source to the mechanistic cancer drugs catalog. The diagrams indicate that the main contribution to the catalog is rooted in DrugBank and TTD, whereas Santos and IUPHAR have rather a limited number of approved drugs which provides more evidence for the corresponding association subsets of DrugBank and TTD.

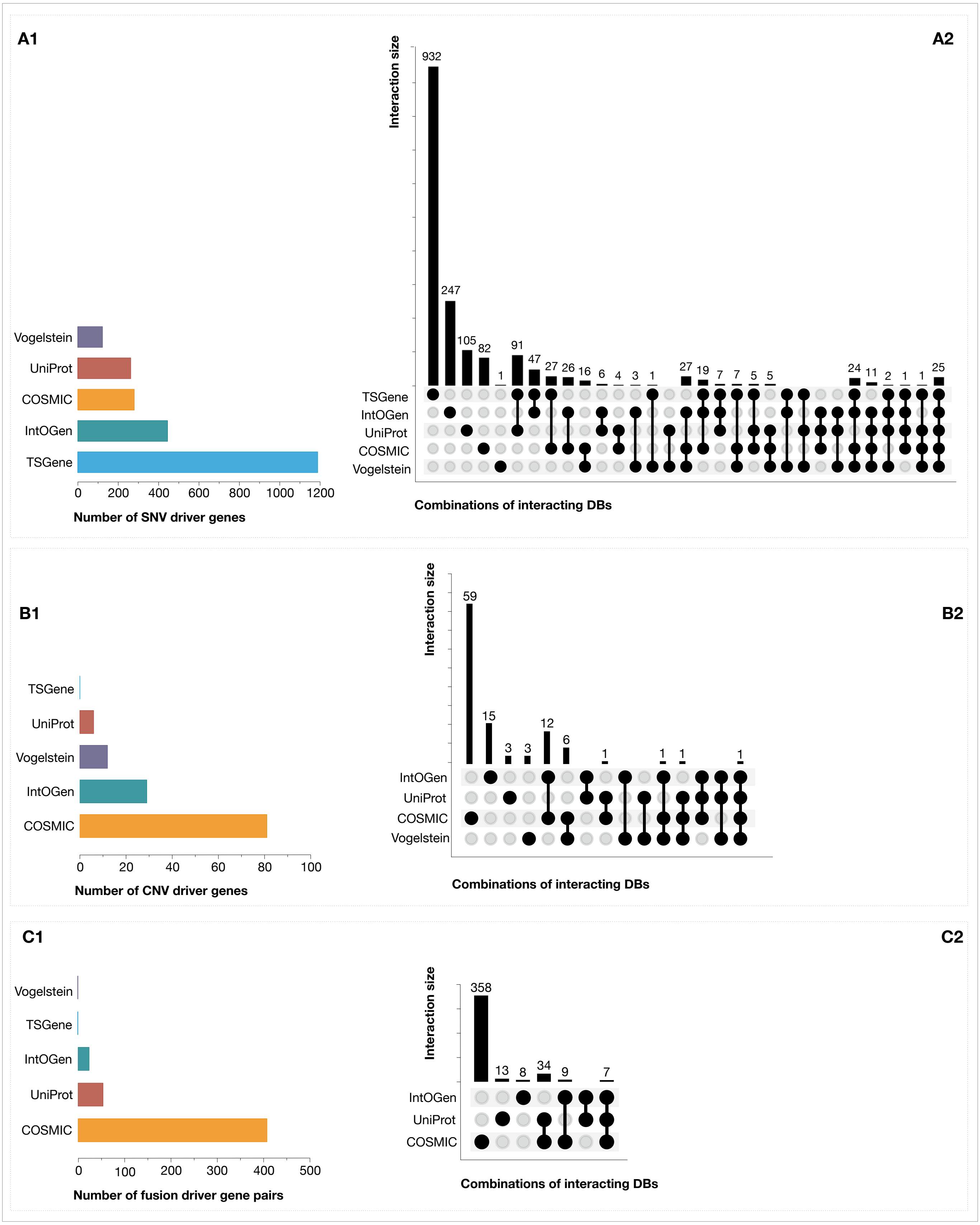


Figure 3.5: Source specific driver gene numbers per variant class. (A) shows the distribution of SNV driver genes. (B) shows the distribution of CNV driver genes. (C) shows the distribution of fusion driver gene pairs. In each class, diagram (1) shows the total contribution of individual sources to the KB and diagram (2) shows the unique contributions of the sources and the combinations.

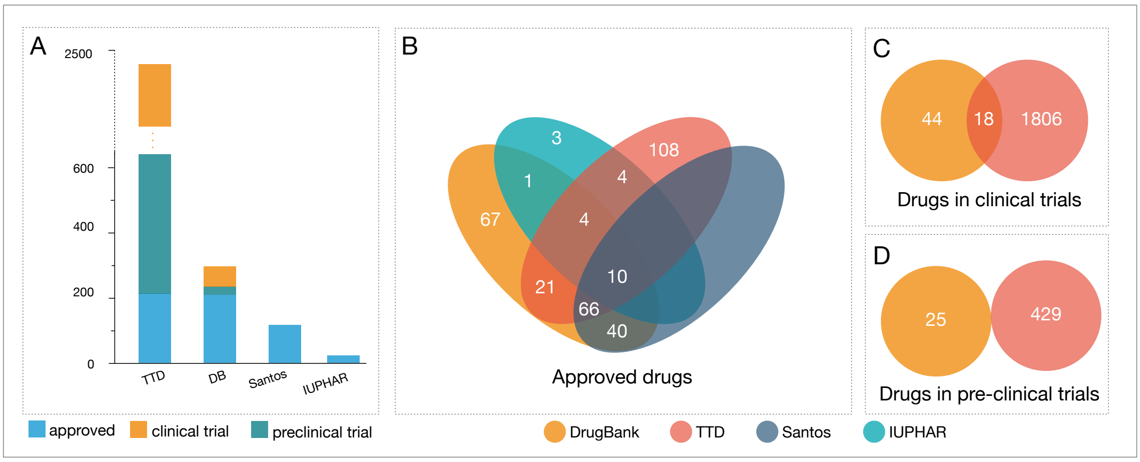


Figure 3.4: Source specific mechanistic drug numbers per approval status. (A) shows the overall drug numbers coming from each sources. (B) shows the source intersections for approved drugs. (C) and (D) show the intersection between data sources for drug in clinical trials and in pre-clinical trials, respectively.

Pharmacogenomics data: We integrated the sources Cancer Genome Interpreter, CIViC, and DrugBank and created a catalog of the drugs which have a known documented clinical response on the observed variants and disrupted genes. The catalog includes descriptive properties of variants such as their location and the base changes, and interpretative details on the variant-drug associations such as predicted variant consequences and observed drug response. It includes information on 943 gene-drug pairs from a total of 243 genes. CGI, CIViC, and DrugBank contribute with 430, 592, and 11 gene-drug pairs, respectively. CGI and CIViC only have 100 common gene-drug pairs.

Adverse effect data: We integrated the sources DrugBank and CGI and created a catalog showing the documented adverse effects of 52 drug-gene pairs from a total of 31 genes, in the presence of specific variants. 42 of the gene-drug associations were provided by DrugBank. We obtained only 4 associations from CGI that were not included in DrugBank.

3.4.3 Patient-Specific Case Reports

The main result of the pipeline is the clinical annotations of the SNVs and CNVs that 1) occur in a known cancer driver gene, 2) have been observed previously in the context of altered treatment response, or 3) fall in the coding region of the mechanistic target gene of cancer therapeutic, 4) cause adverse effects for treatment. They are classified into six categories and saved as a JSON file (the structure is given in APPENDIX).

The report in DOCX format is then created by rendering the JSON file into the report template by Python mail merging tool[184] using the default template. Given a case, the reports for SNVs and CNVs are generated separately. We prioritized the results by selecting the most damaging effect/impact for SNVs, identifying driver genes to reveal the difference between the contribution of damaging variants to cancer initiation and progression, and providing a match level rank to indicate the strength of the database entry's association with the observed variant.

We used a three-tier system in drug content categorization. The complete case overview is presented starting from variant-specific information to gene-specific information with the option of further narrowing down the results based on the cancer type. The table views are shown in Figure 3.6 and 3.7.

Patient Data table holds patient information and general statistics of the molecular profile. Since the *name* and the *birthdate* are personal identifiers, the pipeline is agnostic to such input. Those together with the *additional information* cells are inputted by the clinicians. Another field requiring a post-report registry is the tumor mutation load. It is the total number of nonsynonymous mutations per coding area[185]. Since its calculation factors in the length of the sequenced regions that vary on experiment design which is not a part of inputs, the pipeline does not determine it.

Somatic Mutations in Known Driver Genes table represents the results of driver gene annotation. The columns *driver type*, *tumor type* and *references* are returned from the KB by gene level queries. The remaining fields are obtained as a result of functional annotation. Variant and gene column represent the observed mutations and the gene that the variant is mapped on, respectively.

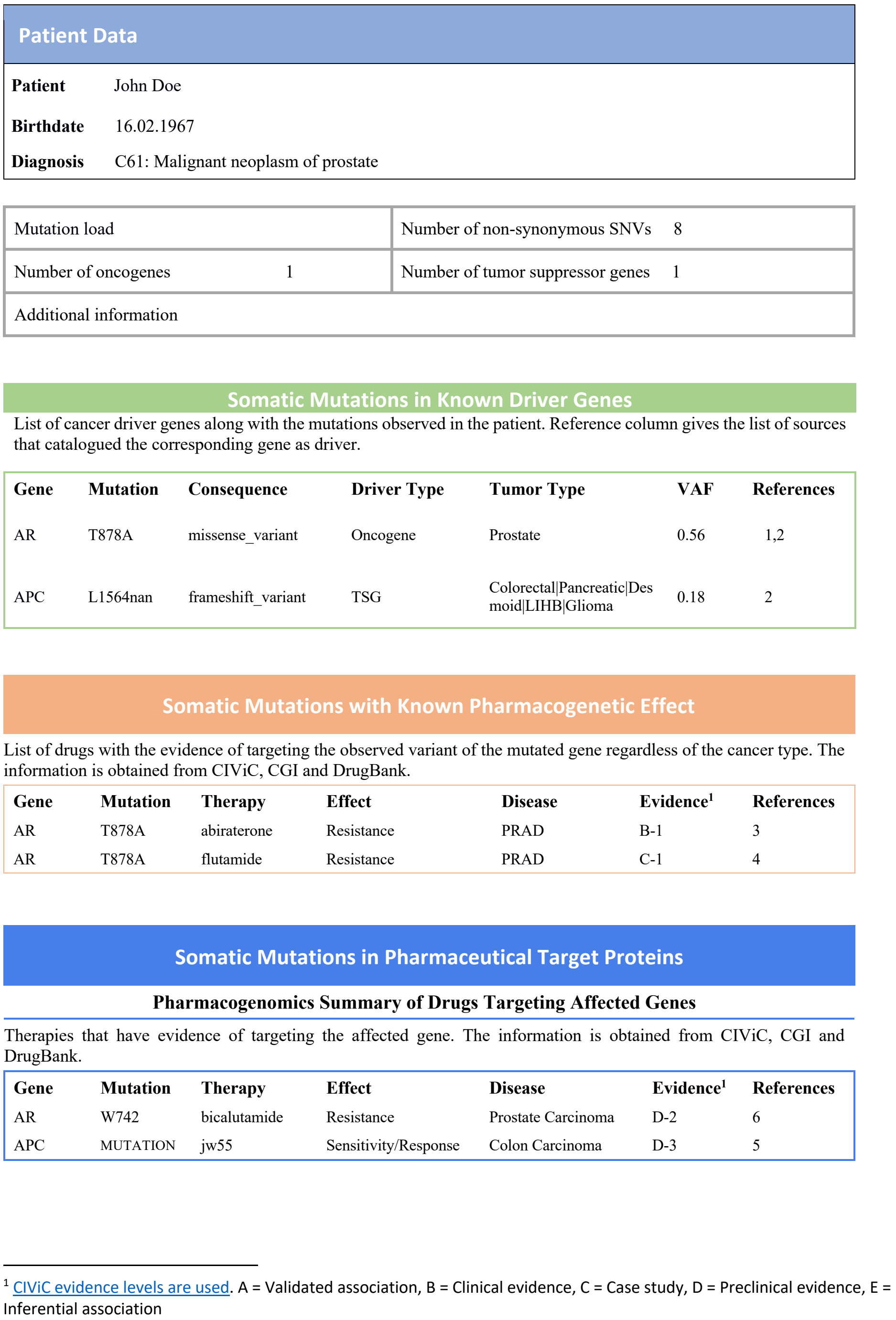
Somatic Mutations with Known Pharmacogenetic Effects is dedicated to therapeutics that have a clinical evidence of targeting the observed variants of the mutated genes regardless of the cancer type, aka list of direct drug-gene association. For example, the corresponding table in Figure 3.6 shows the example of T878A variant of gene AR is resistant to abiraterone in prostate adenocarcinoma cases and the association is supported by clinical trials or other primary patient data.

Somatic Mutations in Pharmaceutical Target Proteins has two sub-tables. *Pharmacogenomics Summary of Drugs Targeting Affected Genes* lists the therapeutics with the evidence of targeting another variant in the affected gene. The corresponding table of Figure 3.6 suggests that W742 variant of AR gene causes resistance when treated with bicalutamide. The evidence level D suggests that there is a preclinical evidence. The match level indication '2' given as a part of evidence level represents that variant

W742 has a same consequence of the observed variant in gene AR. *Summary of Cancer Drugs Targeting Affected Genes* gives the list of drugs that are known to mechanistically target the genes. In this table, the strength of the clinical association is replaced by the approval status of the drugs 3.7.

References and ***Appendix*** tables have the supplementary content of the main tables. The Reference table provides the publications supporting the annotations given in the main tables. Providing the source of information strengthens the clinical adaptability of the pipeline and allows users to expand their investigations with further literature. Additionally, we provide an entire list of non-synonymous mutations of the case in the *Appendix* table together with the mapped genes, VAFs. We link the known variants to their dbSNP or COSMIC id. This table is the main source to identify the rare variants that do not have clinical evidence of a therapeutic association.

As a final remark, providing the reports as a JSON file paves the way for integrating results into electronic health records and the Word document version allows users to store custom information such as the sensitive fields and the measures from sequencing experiments such as chromosomal instability or mutational load.



Patient Data

Patient	John Doe
Birthdate	16.02.1967
Diagnosis	C61: Malignant neoplasm of prostate

Mutation load		Number of non-synonymous SNVs	8
Number of oncogenes	1	Number of tumor suppressor genes	1
Additional information			

Somatic Mutations in Known Driver Genes

List of cancer driver genes along with the mutations observed in the patient. Reference column gives the list of sources that catalogued the corresponding gene as driver.

Gene	Mutation	Consequence	Driver Type	Tumor Type	VAF	References
AR	T878A	missense_variant	Oncogene	Prostate	0.56	1,2
APC	L1564nan	frameshift_variant	TSG	Colorectal|Pancreatic|Desmoid|LIHB|Glioma	0.18	2

Somatic Mutations with Known Pharmacogenetic Effect

List of drugs with the evidence of targeting the observed variant of the mutated gene regardless of the cancer type. The information is obtained from CIViC, CGI and DrugBank.

Gene	Mutation	Therapy	Effect	Disease	Evidence[1]	References
AR	T878A	abiraterone	Resistance	PRAD	B-1	3
AR	T878A	flutamide	Resistance	PRAD	C-1	4

Somatic Mutations in Pharmaceutical Target Proteins

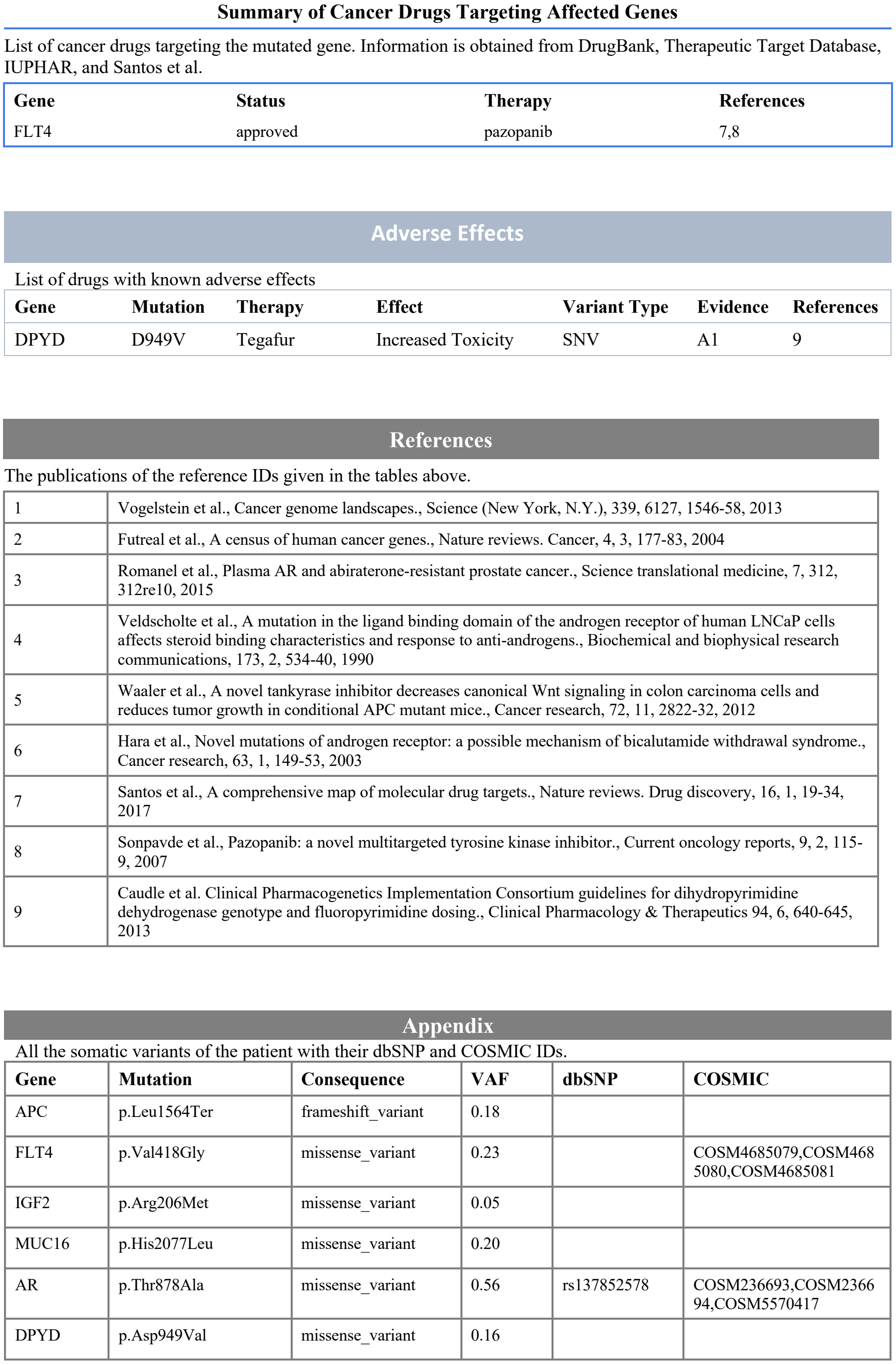
Pharmacogenomics Summary of Drugs Targeting Affected Genes

Therapies that have evidence of targeting the affected gene. The information is obtained from CIViC, CGI and DrugBank.

Gene	Mutation	Therapy	Effect	Disease	Evidence[1]	References
AR	W742	bicalutamide	Resistance	Prostate Carcinoma	D-2	6
APC	MUTATION	jw55	Sensitivity/Response	Colon Carcinoma	D-3	5

[1] CIViC evidence levels are used. A = Validated association, B = Clinical evidence, C = Case study, D = Preclinical evidence, E = Inferential association

Figure 3.6: Case report of SNVs of a mock dataset. The report consists of tables for i) patient data and molecular overview, ii) mutations in driver genes, iii) the list of drug associations with a direct effect on the observed variants, and iv) the list of drug associations with the effect on the disrupted gene.



Summary of Cancer Drugs Targeting Affected Genes

List of cancer drugs targeting the mutated gene. Information is obtained from DrugBank, Therapeutic Target Database, IUPHAR, and Santos et al.

Gene	Status	Therapy	References
FLT4	approved	pazopanib	7,8

Adverse Effects

List of drugs with known adverse effects

Gene	Mutation	Therapy	Effect	Variant Type	Evidence	References
DPYD	D949V	Tegafur	Increased Toxicity	SNV	A1	9

References

The publications of the reference IDs given in the tables above.

1	Vogelstein et al., Cancer genome landscapes., Science (New York, N.Y.), 339, 6127, 1546-58, 2013
2	Futreal et al., A census of human cancer genes., Nature reviews. Cancer, 4, 3, 177-83, 2004
3	Romanel et al., Plasma AR and abiraterone-resistant prostate cancer., Science translational medicine, 7, 312, 312re10, 2015
4	Veldscholte et al., A mutation in the ligand binding domain of the androgen receptor of human LNCaP cells affects steroid binding characteristics and response to anti-androgens., Biochemical and biophysical research communications, 173, 2, 534-40, 1990
5	Waaler et al., A novel tankyrase inhibitor decreases canonical Wnt signaling in colon carcinoma cells and reduces tumor growth in conditional APC mutant mice., Cancer research, 72, 11, 2822-32, 2012
6	Hara et al., Novel mutations of androgen receptor: a possible mechanism of bicalutamide withdrawal syndrome., Cancer research, 63, 1, 149-53, 2003
7	Santos et al., A comprehensive map of molecular drug targets., Nature reviews. Drug discovery, 16, 1, 19-34, 2017
8	Sonpavde et al., Pazopanib: a novel multitargeted tyrosine kinase inhibitor., Current oncology reports, 9, 2, 115-9, 2007
9	Caudle et al. Clinical Pharmacogenetics Implementation Consortium guidelines for dihydropyrimidine dehydrogenase genotype and fluoropyrimidine dosing., Clinical Pharmacology & Therapeutics 94, 6, 640-645, 2013

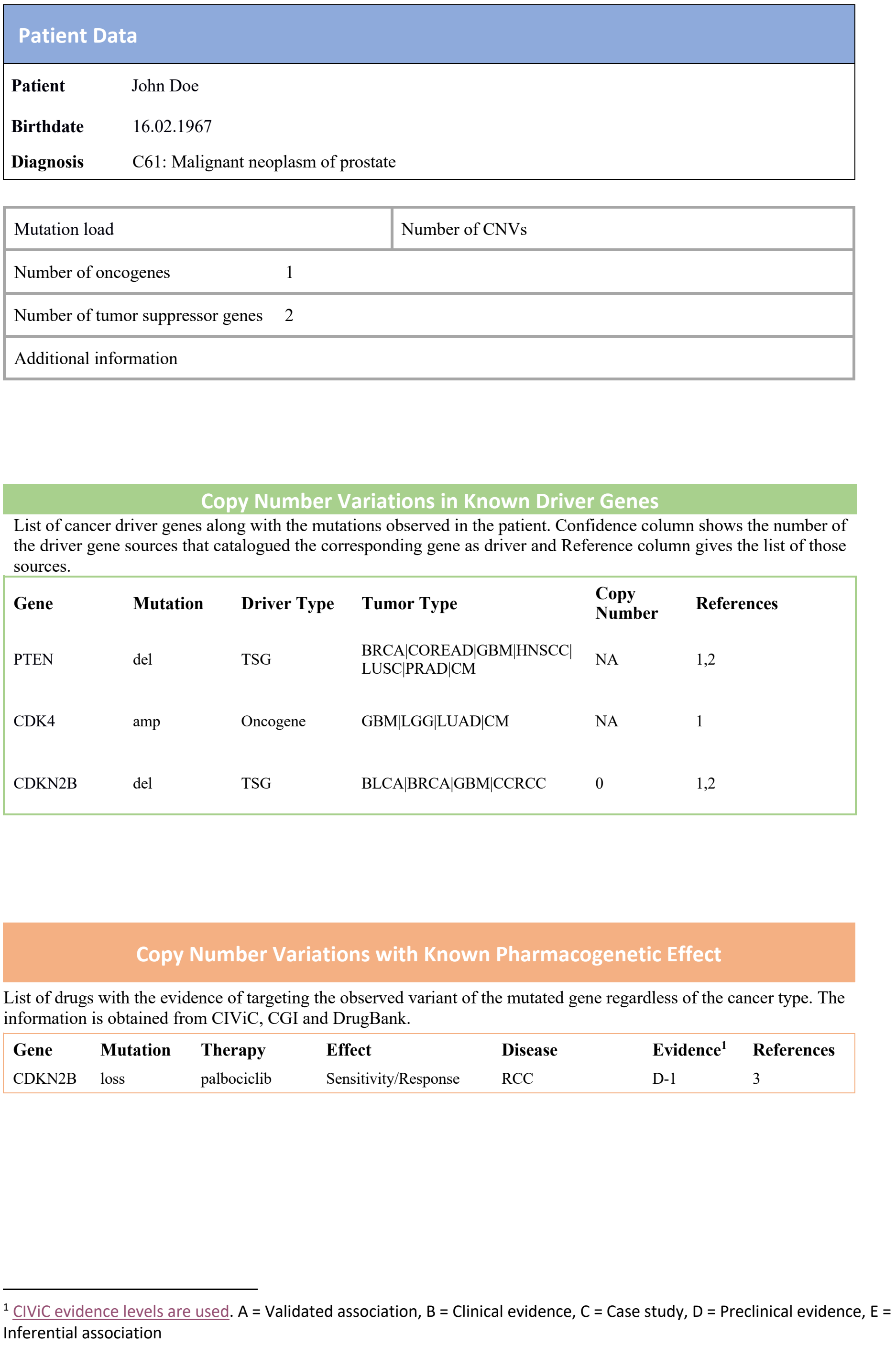
Appendix

All the somatic variants of the patient with their dbSNP and COSMIC IDs.

Gene	Mutation	Consequence	VAF	dbSNP	COSMIC
APC	p.Leu1564Ter	frameshift_variant	0.18		
FLT4	p.Val418Gly	missense_variant	0.23		COSM4685079,COSM4685080,COSM4685081
IGF2	p.Arg206Met	missense_variant	0.05		
MUC16	p.His2077Leu	missense_variant	0.20		
AR	p.Thr878Ala	missense_variant	0.56	rs137852578	COSM236693,COSM236694,COSM5570417
DPYD	p.Asp949Val	missense_variant	0.16		

Figure 3.7: Case report of SNVs continued. The report consists of tables for i) the list of therapeutics with the evidence of targeting the disrupted genes mechanistically, ii) the list of variant drug associations coupled with a known adverse effect, iii) the list of publications that are referenced in the tables, iv) the list of all the non-synonymous mutations observed in the patient.

The report structure has slight differences for the clinical annotations of CNVs. The *consequence* column is removed from the known driver genes table and the effect of the change in the copy number e.g. deletion or amplification is represented with the *mutation* column (Figure 3.8). In the same table, VAF is replaced with copy number (Figure 3.8). One major difference is the removal of the adverse effects table (Figure 3.9), since the KB does not have any adverse effect records associated with the copy number changes. The appendix table is modified to represent the type and the amount of the change in the copy number (Figure 3.9).



Patient Data

Patient	John Doe
Birthdate	16.02.1967
Diagnosis	C61: Malignant neoplasm of prostate

Mutation load		Number of CNVs
Number of oncogenes	1	
Number of tumor suppressor genes	2	
Additional information		

Copy Number Variations in Known Driver Genes

List of cancer driver genes along with the mutations observed in the patient. Confidence column shows the number of the driver gene sources that catalogued the corresponding gene as driver and Reference column gives the list of those sources.

Gene	Mutation	Driver Type	Tumor Type	Copy Number	References
PTEN	del	TSG	BRCA|COREAD|GBM|HNSCC|LUSC|PRAD|CM	NA	1,2
CDK4	amp	Oncogene	GBM|LGG|LUAD|CM	NA	1
CDKN2B	del	TSG	BLCA|BRCA|GBM|CCRCC	0	1,2

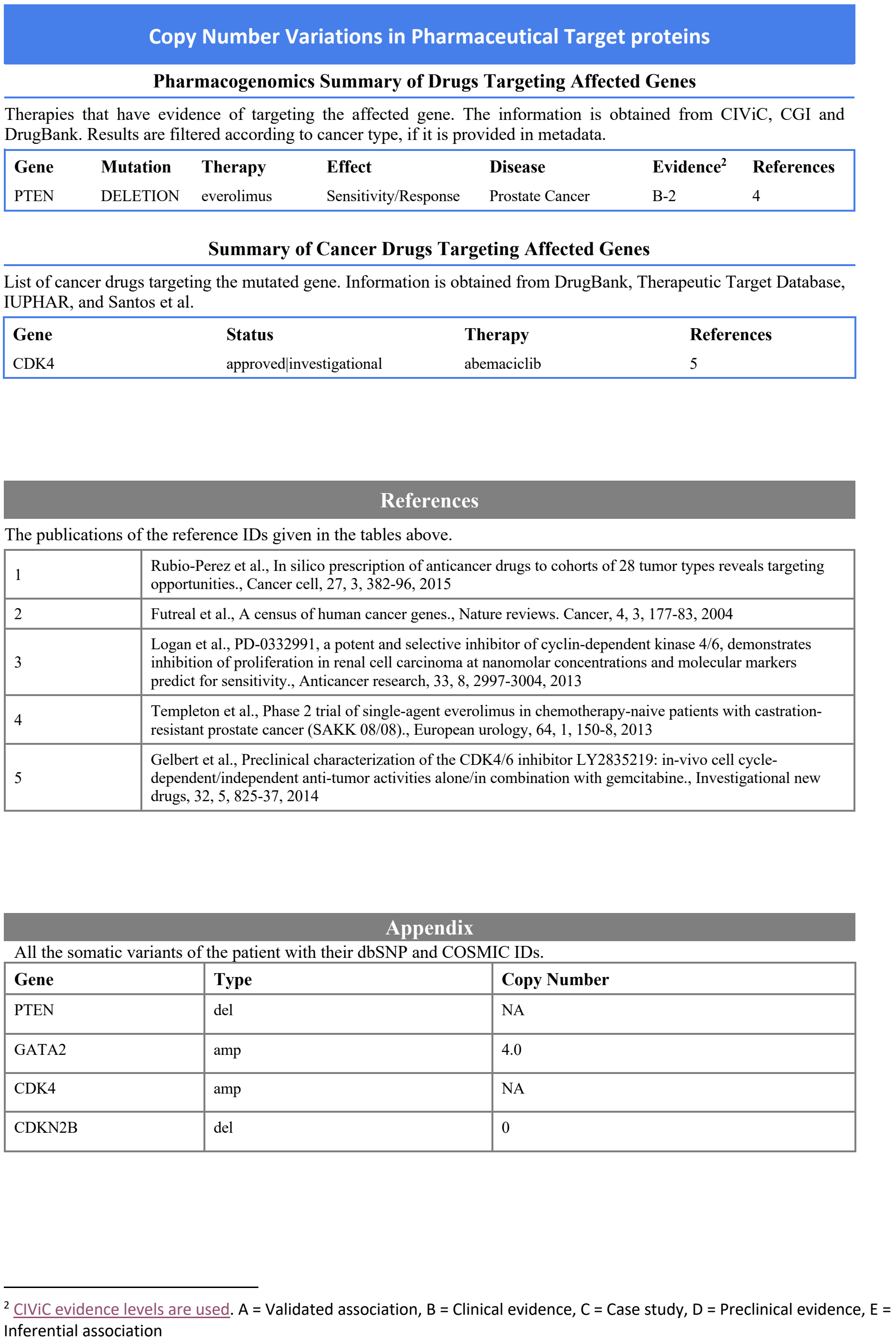
Copy Number Variations with Known Pharmacogenetic Effect

List of drugs with the evidence of targeting the observed variant of the mutated gene regardless of the cancer type. The information is obtained from CIViC, CGI and DrugBank.

Gene	Mutation	Therapy	Effect	Disease	Evidence[1]	References
CDKN2B	loss	palbociclib	Sensitivity/Response	RCC	D-1	3

[1] CIViC evidence levels are used. A = Validated association, B = Clinical evidence, C = Case study, D = Preclinical evidence, E = Inferential association

Figure 3.8: Case report of CNVs of a mock data. The report consists of tables for i) patient data and molecular overview, ii) mutations in driver genes, iii) the list of drug associations with a direct effect on the observed variants, and iv) the list of drug associations with the effect on the disrupted gene.



Copy Number Variations in Pharmaceutical Target proteins

Pharmacogenomics Summary of Drugs Targeting Affected Genes

Therapies that have evidence of targeting the affected gene. The information is obtained from CIViC, CGI and DrugBank. Results are filtered according to cancer type, if it is provided in metadata.

Gene	Mutation	Therapy	Effect	Disease	Evidence[2]	References
PTEN	DELETION	everolimus	Sensitivity/Response	Prostate Cancer	B-2	4

Summary of Cancer Drugs Targeting Affected Genes

List of cancer drugs targeting the mutated gene. Information is obtained from DrugBank, Therapeutic Target Database, IUPHAR, and Santos et al.

Gene	Status	Therapy	References
CDK4	approved|investigational	abemaciclib	5

References

The publications of the reference IDs given in the tables above.

1	Rubio-Perez et al., In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities., Cancer cell, 27, 3, 382-96, 2015
2	Futreal et al., A census of human cancer genes., Nature reviews. Cancer, 4, 3, 177-83, 2004
3	Logan et al., PD-0332991, a potent and selective inhibitor of cyclin-dependent kinase 4/6, demonstrates inhibition of proliferation in renal cell carcinoma at nanomolar concentrations and molecular markers predict for sensitivity., Anticancer research, 33, 8, 2997-3004, 2013
4	Templeton et al., Phase 2 trial of single-agent everolimus in chemotherapy-naive patients with castration-resistant prostate cancer (SAKK 08/08)., European urology, 64, 1, 150-8, 2013
5	Gelbert et al., Preclinical characterization of the CDK4/6 inhibitor LY2835219: in-vivo cell cycle-dependent/independent anti-tumor activities alone/in combination with gemcitabine., Investigational new drugs, 32, 5, 825-37, 2014

Appendix

All the somatic variants of the patient with their dbSNP and COSMIC IDs.

Gene	Type	Copy Number
PTEN	del	NA
GATA2	amp	4.0
CDK4	amp	NA
CDKN2B	del	0

[2] CIViC evidence levels are used. A = Validated association, B = Clinical evidence, C = Case study, D = Preclinical evidence, E = Inferential association

Figure 3.9: Case report of CNVs continued. The report consists of tables for i) the list of therapeutics with the evidence of targeting the disrupted genes mechanistically, ii) the list of publications that are referenced in the tables, iii) the list of all CNVs observed in the patient.

3.4.4 Large-scale Data Processing

3.4.4.1 Deployment and Performance

ClinVAP is available as a self-contained Nextflow workflow configured to use Docker or Singularity image environments[186,187]. Through the containerized execution environment of the pipeline, we ensured easier versioning, full reproducibility of results, and convenient execution on large-scale datasets.

In order to test the robustness and performance of ClinVAP, we processed 500 VCF files from 430 donors containing simple somatic mutations from ICGC cancer projects[188]. The pipeline is executed over the processes VCF QC filter (filter_vcf), Ensemble VEP (vep_on_input_file), clinical annotation (snv_report_generation) and report rendering (render_report_snv).

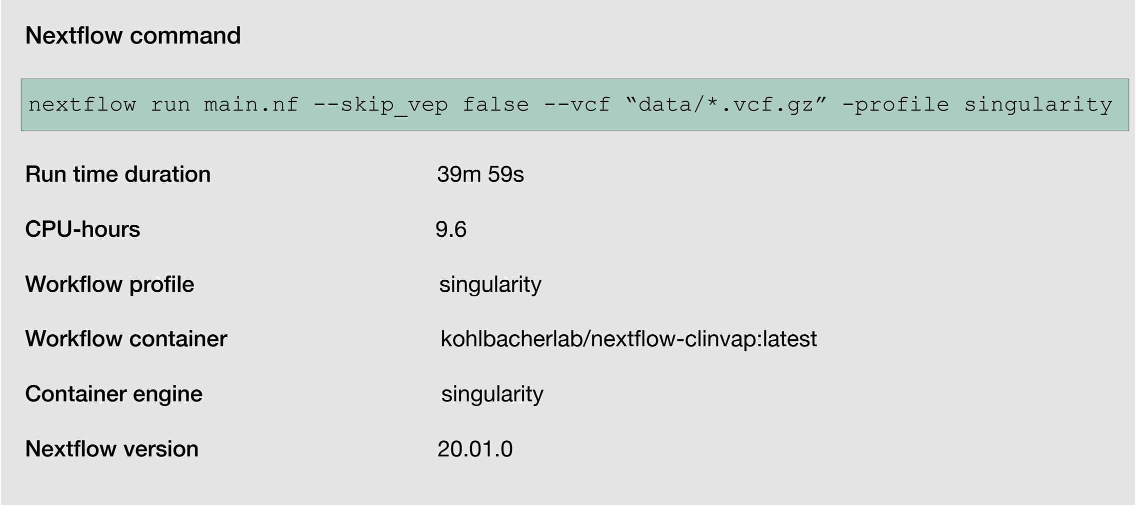
We executed the pipeline in an interactive compute node with 80 cores and 256 GB RAM. We utilized Nextflow's built-in support for container technologies that provides containers as execution environment with the workflow dependencies. We configured ClinVAP execution to use Singularity container engine that converted the existing ClinVAP Docker image (kohlbacherlab/nextflow-clinvap) registered in Docker Hub into Singularity image. The overall run time was 39m 59s corresponding to 9.6 CPU hours (Table 3.3).

The resource usage metrics of ClinVAP large-scale run is generated by Nextflow's tracing and visualisation feature. Figure 3.10A shows the amount of CPU resources used by each process. CPU usage is calculated based on the weighted average of amount of CPUs that the tasks are distributed over and the total duration of each tasks[189]. The most demanding process was the functional annotation with Ensemble VEP (vep_on_input_file) of which the tasks were distributed over 4 CPUs where as the mean of the CPU usage for this process is slightly over 2 CPUs (Table D.1). Figure 3.10B shows the amount of RAM that is used by a process[190]. The most memory demanding process is again vep_on_input_file due to the fact that it works with the large scale data that is not filtered due to the requirements of the clinical annotation step yet. Figure 3.10C shows the execution time for each process[191]. As consistent with previous metrics, the most time consuming step is vep_on_input_file.

3.4.4.2 Case Content Statistics

Among 500 VCF files, 59 of them did not return any variant which passed the ClinVAP's functional annotation filters. We obtained 441 case reports with content. The number

Table 3.3: ClinVAP execution summary.



Nextflow command

```
nextflow run main.nf --skip_vep false --vcf "data/*.vcf.gz" -profile singularity
```

Run time duration	39m 59s
CPU-hours	9.6
Workflow profile	singularity
Workflow container	kohlbacherlab/nextflow-clinvap:latest
Container engine	singularity
Nextflow version	20.01.0

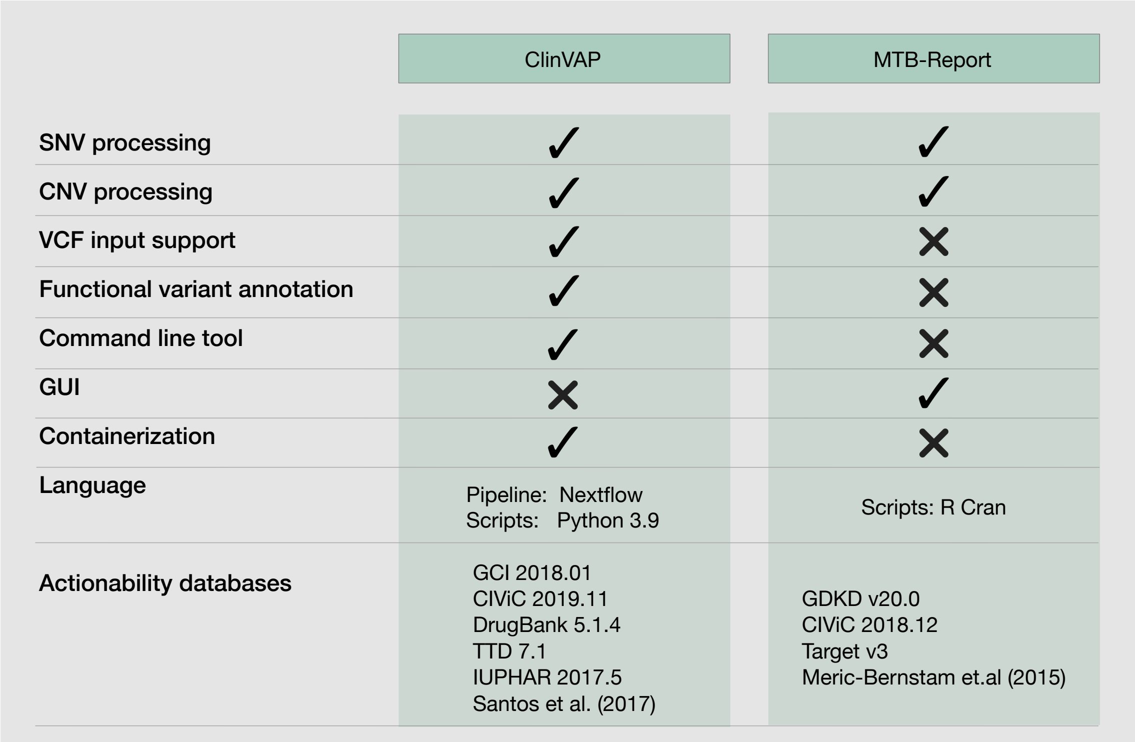
of driver genes identified per case is shown in Figure 3.11A. The majority of the cases had zero to 20 driver genes. Among 441 case reports, 93% of them had non-empty driver gene list. The median and the mean of the driver genes per report were four and 7.3 respectively, with individual donors having up to 162 driver genes. The number of suggested drugs showed more variation between the cases (Figure 3.11B). The median and the mean of drugs per case were found 12 and 49, respectively. 79% of the non-empty case reports had that least one therapeutics suggestion. Only 6% of the non-empty case reports returned neither driver genes nor suggested therapeutics

ClinVAP reported 26,068 genes (11601 unique) with non-synonymous mutations in total. All the case reports had at least one variant with a predicted functional impact. The median of the genes containing non-synonymous mutations was 29, whereas the average was 59 (Figure 3.11C). The majority of the cases had zero to 100 variants, with few cases having over a thousand variants.

3.4.4.3 Benchmarking

We compared ClinVAP with MTB-Report[169] which is another tool for generating evidence-driven reports from somatic mutations for MTBs (Table 3.4). Even though, MTB-Report is developed to generate case reports of the druggabble targets via searching its background databases for mutations, it does not conduct functional variant annotation and variant prioritization steps. It relies on a tab-separated input format for SNVs containing gene name, variant classification and protein change. It uses fewer sources to annotate actionable genes.

Table 3.4: ClinVAP's comparison with MTB-Report



	ClinVAP	MTB-Report
SNV processing	✓	✓
CNV processing	✓	✓
VCF input support	✓	✗
Functional variant annotation	✓	✗
Command line tool	✓	✗
GUI	✗	✓
Containerization	✓	✗
Language	Pipeline: Nextflow Scripts: Python 3.9	Scripts: R Cran
Actionability databases	GCI 2018.01 CIViC 2019.11 DrugBank 5.1.4 TTD 7.1 IUPHAR 2017.5 Santos et al. (2017)	GDKD v20.0 CIViC 2018.12 Target v3 Meric-Bernstam et.al (2015)

Since MTB-Report is not available as a command line tool, we used its GUI to obtain the case reports and compare the contents with that of ClinVAP's. Unfortunately, none of the cases returned a result from MTB-Report since the tool kept crashing in its attempts to search GDKD knowledge base.

3.5 Discussion

The precision oncology era requires translational workflows from bench to bedside to identify the treatments that are optimal given a molecular profile. Another aspect of improving cancer care involves efficient archiving and re-usage of the patient data which have the potential to contribute to decisions on similar future cases and population stratification. Even though targeted therapies are found to improve cancer care, the clinical implementations of phenotype-genotype associations are precluded due to the complexity of the data, the abundance of publicly available databases with complementary content, and the variety of bioinformatics tools required for a complete annotation workflow, of which many necessitates programming background. The available methods fail to comply with the requirements of a complete pipeline facilitating various steps from functional variant assessment to clinical annotations.

We addressed these issues with ClinVAP, a fully automated, fast, and robust annotation pipeline to equip clinicians with evidence-based patient reports which reveal the molecular driving forces in cancer formation and actionable therapeutic targets among the patients' somatic variants. Nextflow implementation of the pipeline provides reproducibility and scalability. Besides Nextflow's feature of running the pipeline in a pre-installed Docker environment, we fix the data source versions of the KB, so that the report content does not deviate due to database updates. We proved the robustness of the pipeline with stress tests using 500 VCF files including simple somatic mutations from the ICGC cancer projects[188]. In addition to the therapeutic content, we provide a complete list of non-synonymous variants to notify the users of the variants of unknown significance. The local installation of ClinVAP with offline functional annotation and knowledge base access prevents sending the patient data to remote servers, ensures patient confidentiality, and abides by the data security regulations. The pipeline provides clinicians the flexibility to add additional annotations/notes via outputting the report directly to a Microsoft Word Document. The idea here is that the physicians have reports that they can file for insurance and administrative purposes. The JSON version of the report comes with the advantage of interoperability with the electronic health record systems or digital medical data archiving solutions.

ClinVAP is specialized on somatic variants. Germline variants are out of its scope due to the strict legal regulations on their usage. Additionally, germline testing is not a standard procedure for patients who do not satisfy certain eligibility criteria[192]. However, germline alterations are associated with increased susceptibility to cancer and influence therapy choice and clinical trial eligibility[193] (e.g., the heritable alterations on FANC genes increasing the risk of cancer). The mutations in FANC genes disrupt the FA pathway causing genomic instability[194]. They drive Falconi anemia which is not cancer but induces pre-disposition for various tumor types[195]. Evaluating somatic and germline mutations together with automated decision support tools can increase the sensitivity of treatment suggestions. However, the prerequisite of implementing such a feature would be germline testing becoming prevalent in the clinical decision-making workflow.

Throughout the chapter, we referenced the publicly available databases as a source of actionability and tumorigenesis impact information of genomic variants of our KB[94,144,145,163,177–179]. Although they do not follow the same update cycle, some of them have more dynamic content change than others for example, the CIViC database continuously curates its none-approved entries. This suggests an overwhelmingly frequent necessity of updating KB. However, the content of the sources is mutually inclusive, and we do not expect the minor updates would result in significant content loss. The manual effort of MTBs also includes a PubMed search for recent developments. However, this nonsystematic and time-consuming approach brings disadvantage to reproducibility and increase fallibility. Implementing a systematic search module over PubMed for the recent relevant publications can be considered as a solution that comes with the cost of introducing noise due to the lack of manual curation and explodes the content of the case reports. Therefore, we suggest a yearly update cycle for the KB and consider MTBs as the responsible body for conducting a rigorous PubMed search when they see the need.

In conclusion, offering automated, concise, and robust solutions to the daunting challenge of translating genomics data into clinical information increases the efficiency of cancer care. Although the precision oncology field is still in constant evolution and the current requirements will need enrichment, ClinVAP responds to the crucial aspects of the current MTB workflow.

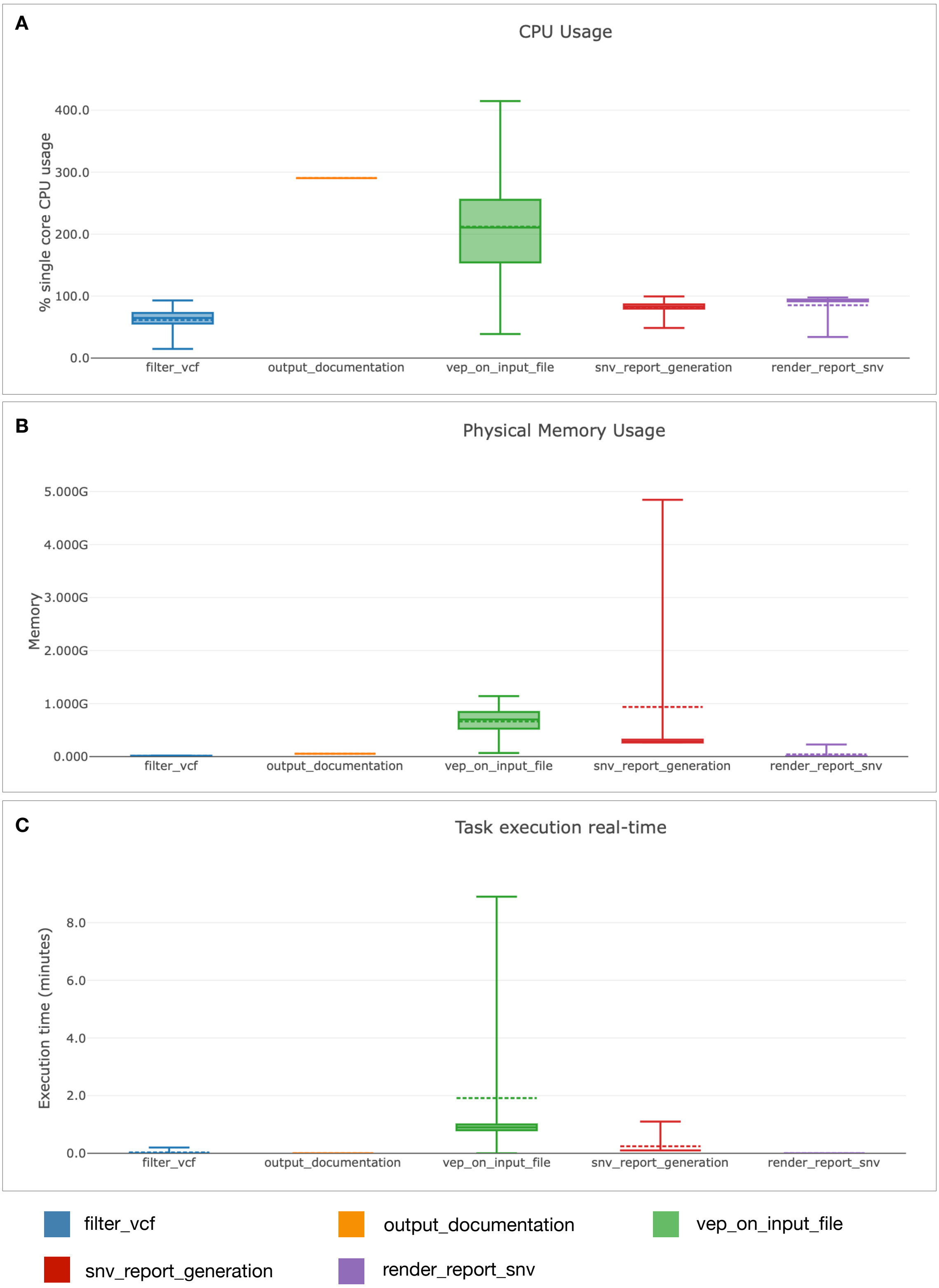


Figure 3.10: ClinVAP's resource usage. A) CPU usage. The most CPU demanding process is vep_on_input_file which is the step of functional annotation on 500 files. B) Memory usage. The most memory demanding process is vep_on_input_file since it works with the large scale unprocessed data. Although its overall memory usage is less than vep_on_input_file, snv_report_generation requires up to 5GB of memory for the tasks that work with large processed-vcf-inputs C) Execution time. As consistent with other metrics, vep_on_input_file process executes the most time-consuming tasks.

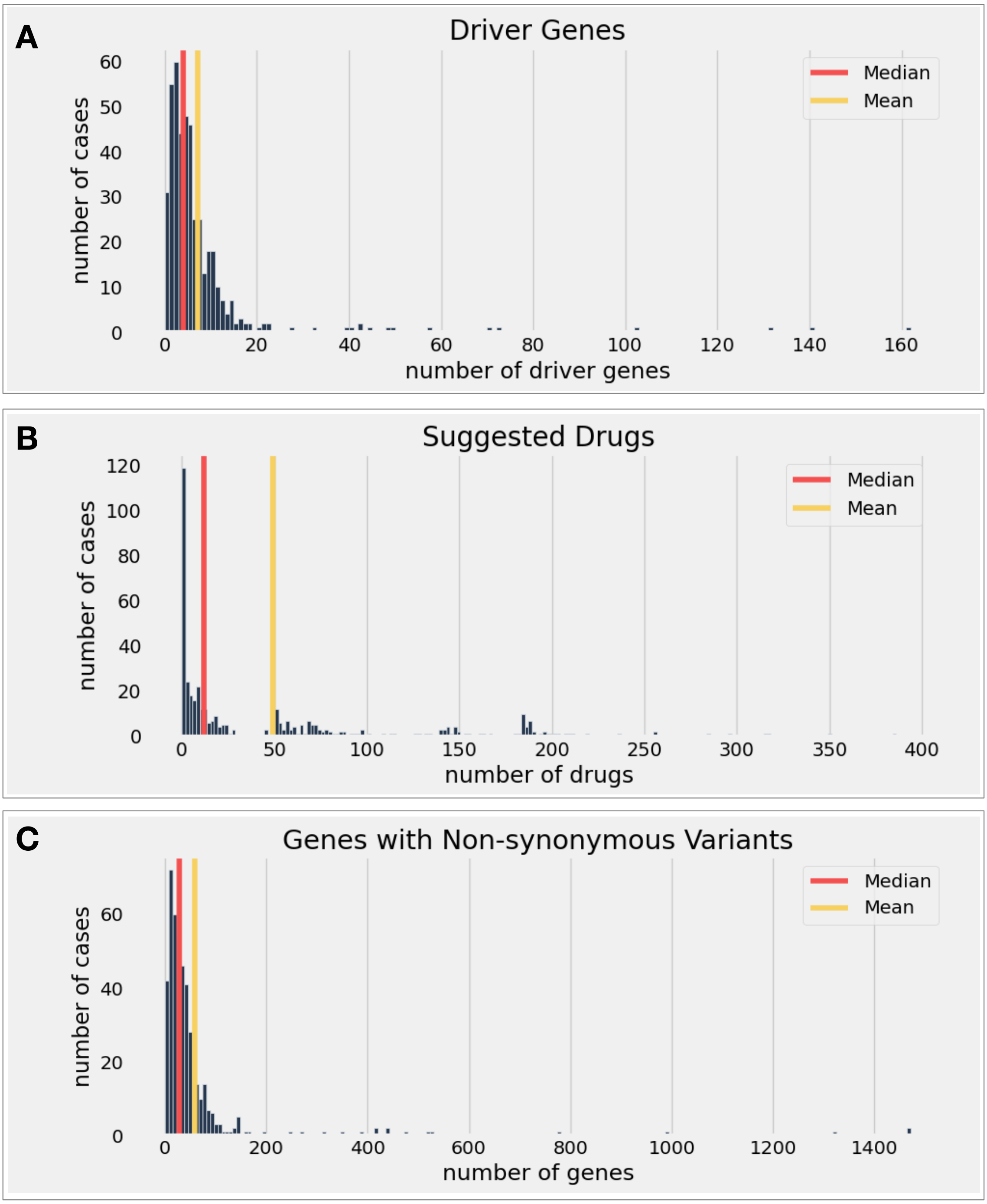


Figure 3.11: The distribution of driver genes, drugs and non-synonymous mutations over the cases. A) Driver genes. The majority of the cases had zero to 20 driver genes. The average driver gene number per report was 7.3 and the median is four. B) Drugs. The median number of drugs was 12 and the average number of drugs per case was 49. The number of suggested drugs showed higher variation between the cases. C) Genes with non-synonymous mutations. All cases returned at least one such gene. The average number of mutated genes per case was 59 and the median was 29.

Chapter 4

Interactive Case Exploration with PeCaX

The content of this chapter is an extended version of the article:

Figaschewski, M., Sürün, B., Tiede, T., & Kohlbacher, O. (2023). The personalized cancer network explorer (PeCaX) as a visual analytics tool to support molecular tumor boards. BMC bioinformatics, 24(1), -11[196].

4.1 Introduction

NGS data becoming widely available with reduced costs and short turn-over time enabled a deeper understanding of disease mechanisms. Oncology clinics adopted precision oncology operations to guide clinical decisions based on the patients' unique molecular profiles. Precision oncology has evolved around the identification of genomic biomarkers to predict the likely drug response and the possible drug resistance mechanisms[197] to create personalized therapeutic interventions which would increase the treatment efficiency with reduced costs.

Biomarker profiling starts with referring the biopsy sample from patients selected for targeted treatment, usually after exhausting conventional therapies, to sequencing facilities. Clinics obtain a list of pre-processed variants that are likely to have a damaging effect and significant contribution to the disease progression. Personalized medicine units investigate the long list of variants to prioritize them based on the therapy relevancy. Institutional MTBs then hold a discussion based on the processed list of variants to identify targeted therapy strategies. It is an arduous, mostly manual, and thus error-prone task to identify targeted therapies tailored to the patient. Moreover, additional challenges are introduced to genome-based cancer care such as acquired drug resistance mechanisms and "undruggable" cancer targets[18,198].

The most common acquired resistance mechanism seen in oncogene-based therapies is the re-activation of the proliferation pathway which circumvents the drug action through alternate routes affecting downstream signal transduction pathways[199]. The compensatory mechanisms are formed by either the crosstalk with another pathway, byproducts substituting each other within the same pathway, or by a parallel pathway performing a similar cell function[200]. Another factor contributing is the co-occurring loss of function variant(s) in a tumor suppressor gene downstream of the proliferation pathway resulting in cell death inhibition[200]. Sequence-based comparison methods[201] in combination with pathway annotations is a common strategy to demonstrate both acquired and intrinsic resistance mechanisms through gene-gene and drug-gene interactions in close proximity to disrupted genes. Revealing affected pathways can also result in the identification of the independent parallel pathways that contribute to resistance, which then creates the rationale for combination therapies to circumvent drug resistance[202].

Another bottleneck for precision oncology implementations is the fact that only 10-20 % of the cancer genome is directly targetable[18]. This limits the therapy options for patients who do not have any actionable targets. One potential mitigation strategy is to exploit gene interactions to trigger metabolic inhibition creating a shift from targeting driver genes to defining driver events. Understanding the interactions of undruggable genes would contribute to reveal driver events that could be targeted by taking advantage of their protein-protein interaction (PPI) network to disrupt their oncogenic function. The network-based methods also have a notable potential to stratify patients based on their molecular profile to factor in the genomic heterogeneity of cancer into the treatment plan. All that emphasizes the importance of assessing not only the observed aberrations but also their network interactions.

To overcome those challenges, assessing the patient's susceptibility to drug resistance should be streamlined to a clinical cancer care routine in addition to actionable target evaluation. Since the data is intrinsically complex, it is difficult to combine clinical annotations with a network layer of information manually. Thus, pipelines to determine actionability/druggability (elaborated in Chapter 3) require an extension to automate network generation and pathway annotation for providing an overall view of the patient's mutational landscape with the least amount of programmatic efforts. The clinical acceptance of such pipelines is strongly tied to the user-friendliness of the graphical user interface (GUI) which allows users to operate through results such as filtering, sorting, note-taking, and cross-referencing to other sources to enable thorough investigation.

There are available tools designed for specific steps of the multiplex process of precision oncology workflow with the least amount of required programming knowledge. VCF visualization tools such as VCF-Miner and VCF-Explorer[203,204] provide users from non-bioinformatics backgrounds an interface to conduct VCF operations such as presenting the file content in a human-readable format, allowing users to conduct operations such as filtering with customized options while favoring the memory efficiency of the operations. However, they mainly rely on the variant annotation conducted before the VCF operations. Thus, they do not reduce the complexity of analyzing genomic data and are neither advantageous nor preferable in the clinical routine where a quick turnover time for the genomic analysis is prioritized. Any additional steps those tools require, such as conducting the annotation, re-shaping data based on the required input format, and selection of quality control filters based on experimental measurements, increase the workload in clinics making them inapplicable in the health care routine.

Other available tools do not solely focus on variant operations but specialize in the clinical annotation to create a concise list of therapy-relevant variants reviewed in Chapter 3.1[169,170]. They do not integrate gene-gene interaction networks to their results; therefore, they fail to enable users to assess cumulative resistance mechanisms and alternative therapeutic options required for undruggable targets.

Another major obstacle hindering the use of these tools is their lack of an interactive GUI supporting MTB actions. Overall, the field lacks one central tool which is capable of conducting necessary operations to explore genomic data with a clinomics approach of integrating interaction networks with the pathway information in a user-friendly GUI allowing clinicians to examine a case in depth.

This chapter introduces Personalized Cancer Network Explorer (PeCaX), a clinical decision support tool that performs clinical annotation on the genomics data, provides gene-drug interaction networks of the identified aberrations, and ensures communication with clients with its interactive GUI. PeCaX is compliant with microservice architecture with each component containerized with Docker attributing the ease of maintenance, deployment, and reproducibility to the software. It is a local application ensuring data security by performing all the analysis on the local infrastructure and removing the input data once the analysis is completed. It is supported by all modern web browsers across platforms. Its usage requires no programming knowledge since all the backend operations are hidden from the users. It is easily integrable into diagnostic and MTB workflows to investigate the relevance of variants from individual cases or patient cohorts.

4.2 Design and Implementation

PeCaX is intended as a microservice-based tool consisting of individual and independently deployable components that communicate via the REST API. Its web-based frontend is implemented using the NuxtJS framework. The main target users for the tool are MTB members preparing the cases and selected representatives of these users were consulted during the design phase to increase its adoption in clinical routine.

4.2.1 Design Concepts and Principles

Since the main objective of PeCaX is to provide a decision support system for MTBs to assist the task of case preparation and discussion, we determined user acceptability as the most important design principle. One of the most important requirements was providing a tool that does not require any programming knowledge and is free of complex programming interfaces. Hence, we isolated all the backend operations from the user and limited users' interaction with our application through the GUI.

We designed our tool to work with standard data formats to eliminate the manual input preparation step. Additionally, we focused on boosting the user-friendliness of the GUI to receive the input, provide the results as concisely as possible while allowing the users to interact with it, and provide convenient archiving options such as downloading the results in a human and machine-readable format. The frontend is a web service that works on any modern browser independent of the operating system. It supports concurrent use allowing access from browsers not running on the same machine but within the same network.

Another important aspect that shaped our design was the data privacy concerns as genomics data is intrinsically highly confidential. To avoid sending data to servers for analysis, we constructed PeCaX as a local application and ensured that sensitive data is only processed on the machine PeCaX is installed on, not the machines that access it via the GUI. We prevent unauthorized data access by deleting the main input genomics data as soon as the data analysis is completed.

We highly prioritized reproducibility of the results to provide standard and coherent analysis through consistent versioning and containerization of our application. We fixed all the versions of third-party analysis tools and annotation sources to avoid the uncontrolled effects of software and database updates on the results. We used containerization technologies to ship our application with all the package and platform dependencies.

PeCaX is easy to maintain due to its microservice architecture design. We separated the main components based on their functionality such as clinical annotation, network generation, and GUI in a way that each component is a stand-alone service independent of the other. We containerized each microservice separately and orchestrated them through Docker compose. We established microservice communication via REST APIs and consistent data sharing among the individual containers through Docker data volumes. Microservice architecture together with each service working with standard data formats also provides added advantage on healthcare interoperability, since it enables the results to be integrated into electronic health records, it also allows employing the services in a larger application ensuring the communication with REST APIs.

4.2.2 Implementation

4.2.2.1 Architecture

We implemented a microservice-based architecture with three main service layers to provide clinical annotation, network generation, and the GUI to visualize the results and enable users to interact with the tool (Figure 4.1). The clinical annotation service employs ClinVAP[161] to create patient-specific case reports from the patient's genomic data. Network generation service utilizes SBML4j[205] to create gene-gene and gene-drug interaction networks of the disrupted genes observed in a patient. Visualization service is a GUI including two main modules, one providing the results from the clinical annotation in a tab-separated table format and the second module, BioGraphVisart to provide network visualization[206]. We also implemented an additional data management module as an ArangoDB database. It creates a collection to store unique job ids, selected parameters, and the analysis results which enable users to retrieve a previous session or start a new session from the previously downloaded results. We containerized every service layer individually using Docker and orchestrated via docker-compose. We used Docker volumes to store and share the data within the individual services. We set up service communications via a client/server model mediated by RESTful APIs handling HTTP requests.

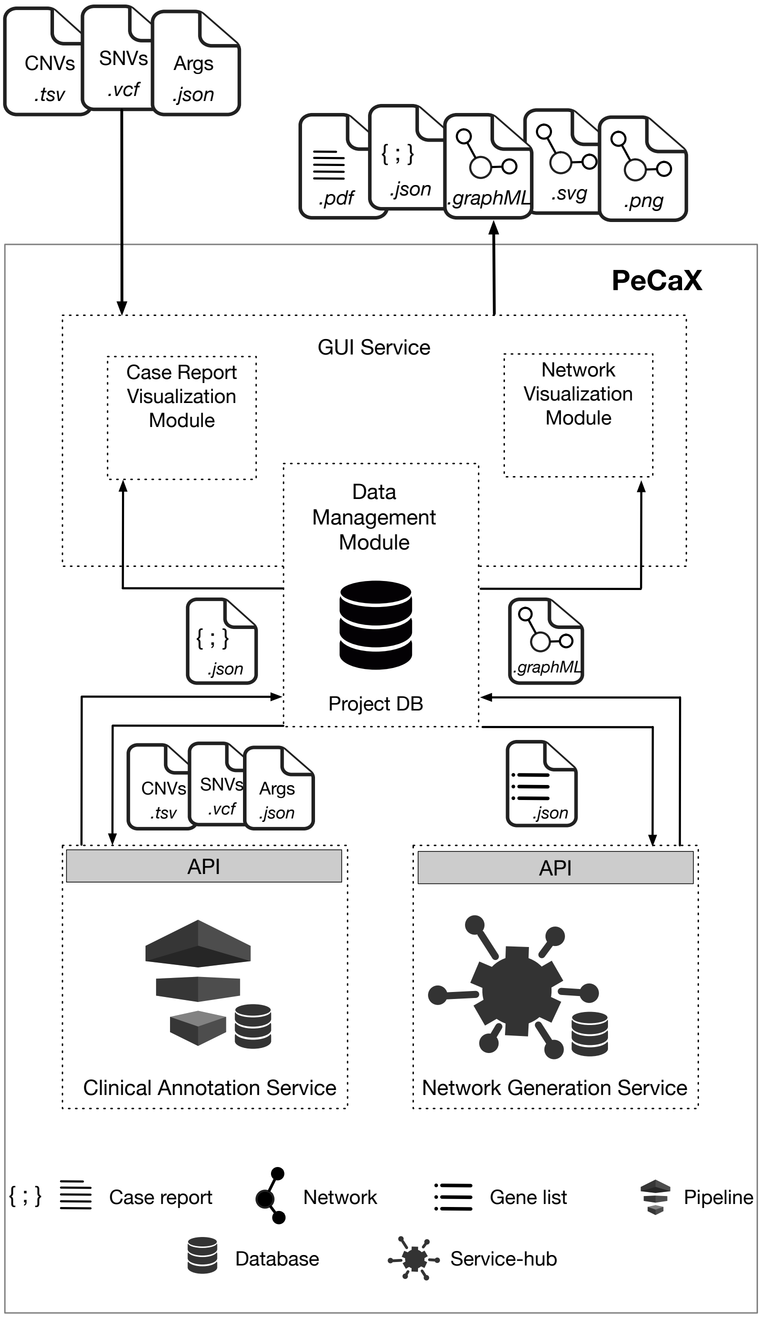


Figure 4.1: Overall PeCaX Architecture. The communication between the backend services and the client is established via the GUI. User inputs are SNVs in .VCF format, CNVs in .TSV format and additional arguments such as the human genome version and the diagnosis passed to the application as JSON. Data sharing between the services are ensured via the data management module implemented in ArangoDB. SNVs, CNVs and the arguments are passed to the clinical annotation module. The resulting case report is sent to the report visualization module. A list of genes from the case report is passed to the network generation module. The resulting networks (GraphML) are sent to the network visualization module.

4.2.2.2 Clinical Annotation Service

We converted ClinVAP (Chapter 3) into a PeCaX service component to generate the case reports from functional and clinical annotations of a mutational profile. We equipped ClinVAP with an Nginx webserver to handle HTTP protocols. We used Flask, a WSGI-compliant web application Python framework, to provide access from the server to the backend application. ClinVAP Nextflow pipeline (Chapter 3.2.2) is used as the backend application providing the clinical annotation functionality.

Clinical annotation service is consistent with microservice design principles. It incorporates Nginx, Flask, Nextflow, and VEP file deployment services, each is an independent component of the overall service. Each microservice is containerized separately with Docker containerization technology (Figure 4.2). We implemented minor changes to the ClinVAP Nextflow pipeline to encapsulate the whole pipeline in a Docker image. Instead of providing pipeline dependencies as a Docker image and running the script within this environment, we replaced the dependency image with a Conda environment management system, to avoid the unrecommended use of running a Docker container within a Docker container.

Once the application is up, the Nginx container starts the web server and publishes the port to receive and process the requests and return the response. Flask container monitors the published port, accepts the users' requests, and calls the application method. ClinVAP Nextflow container constantly observes the data volume through a Python `FileSystemEventHandler` script for newly created VCF files to automatically start the pipeline. When the process finishes, the same event handler deletes the inputted SNVs, CNVs, and additional arguments file. The data sharing among the microservices is ensured via local Docker volumes. Flask and ClinVAP Nextflow containers have common volumes to pass the input and output files between themselves. Additionally, to establish fast file transfer for Ensembl VEP cache and FASTA files required for offline variant effect prediction, we created a VEP file deployment image that contains all the required files and copies them to one of the shared Docker volumes on ClinVAP Nextflow container, if they are absent.

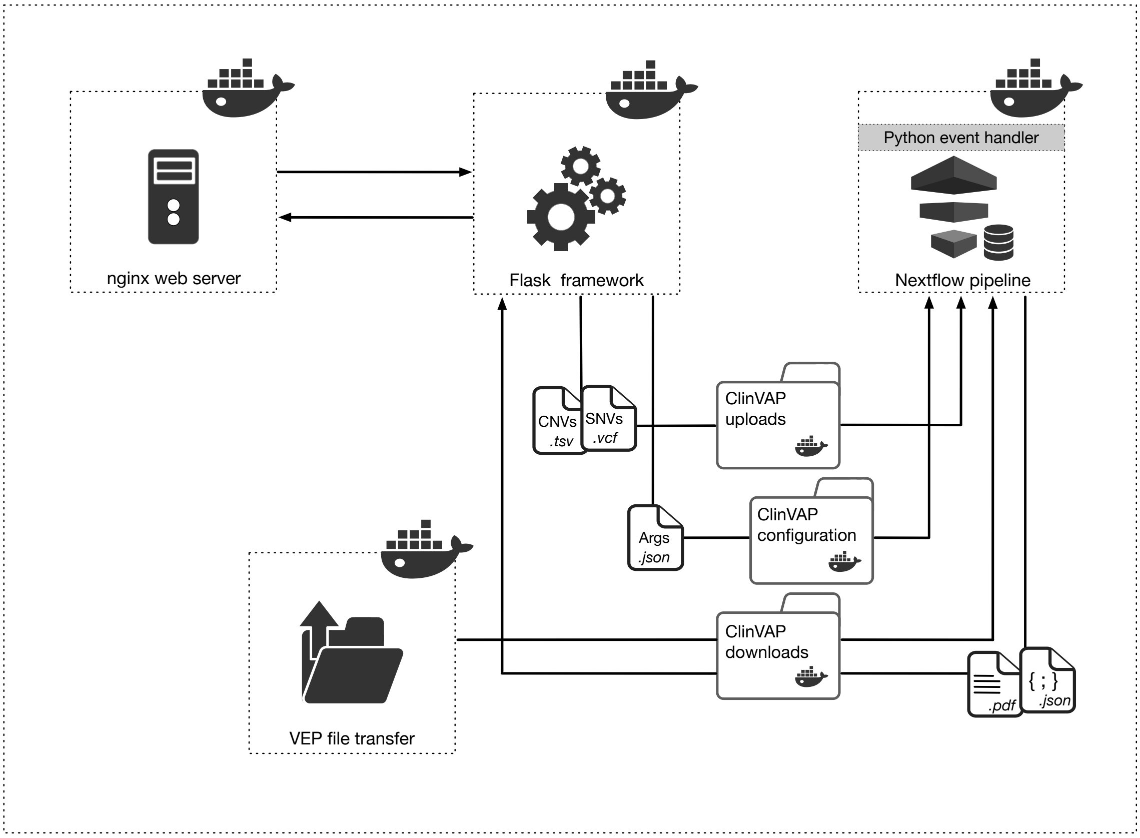


Figure 4.2: Clinical annotation service architecture. Each component is containerized with Docker and orchestrated with docker-compose. ClinVAP pipeline image depends on Flask and VEP file transfer services. Flask framework transfers the user inputs to *ClinVAP uploads* and *configuration* volumes. Python event handler watches the *uploads* volume for newly created VCF files to trigger the Nextflow pipeline. VEP file transfer image carries the cache and FASTA files necessary for offline variant effect prediction. It delivers those files to *ClinVAP downloads* volume. The resulting case reports are shared with the Flask framework through the *downloads* volume and exposed to the client upon request. Nginx web server and Flask framework services work together to ensure communication with the client.

We exposed endpoints to receive input parameters and data files, publish the driver gene list and the entire case report in JSON format, and signal PeCaX GUI when the analysis is done or interrupted due to an error.

The REST endpoint `/upload-input` is modeled as `POST` method and dedicated to send input files and parameters to the server. It is realized with the content-type `multipart/form-data` since the aim is to transfer different types of data in one request. It transports two input files: SNVs as *VCF* and CNVs in *TSV* format. The validity of files to be uploaded is checked by the file name extension. The files are rejected if they have a different extension than the allowed ones. Next to the input files, it requests the parameters `human genome assembly version`, `filtering type` and `ICD10 code` from the user.

The only required argument is the *VCF* file. Uploading a CNV file and the remaining parameters are optional. The parameters `human genome assembly version` and `filtering type` have a default value, if not provided. The transferred files are saved to Docker volume `clinvap_uploads`. Additional arguments are bundled in a JSON file and saved to `clinvap_conf` volume as the Nextflow pipeline's configuration file.

```
requestBody:
 description:
 content:
  multipart/form-data:
   schema:
    type: object
     properties:
      filename:
       type: array
       properties:
        snv:
         type: string
         format: binary
        cnv:
         type: string
         format: binary
        arguments:
         type: object
         properties:
          assembly:
           type: string
          diagnosis:
           type: string
          filter:
           type: string
```

Listing 4.1: Definition of the POST request, uploading inputs

The endpoint `/results/<filename>/status` is modeled as a `GET` method to handle the request of the pipeline's run status. It accesses the pipeline's log file on the data volume. It parses the log file and reports the pipeline's status at the moment of the request as a *JSON* object with response code 200. The responses signaling the status of the pipeline are "finished with errors", "finished with success" and "running". It also handles the `FileNotFoundError` as an exception and returns a 404 response if the log file is not in the data volume indicating that the pipeline has not been started and the request is invalid.

`/results/<path:filename>` endpoint is modeled as a `GET` method to send the case report as a *JSON* file attachment for users to download. The route takes the `filename` as a variable to create a path under the static folder "results" which is the base directory in the local Docker volume in which the pipeline saves the resulting report. The users' access to the folder is controlled with the `filename` variable which is the name of the inputted VCF file with *JSON* extension. The endpoint also handles `FileNotFoundError` as an exception and returns a 404 code if the report does not exist in the path.

`/results/<filename>/tables/driver-genes` is modeled as a `GET` method to handle users' request of obtaining the driver gene list. It parses the case report to extract the driver gene list and returns it as a *JSON* response. Similar to the previous endpoint, it takes the `filename` as a variable and searches for the case report in the created path under the static results folder. It returns a 404 file not found error code if the case report does not exist in the path.

4.2.2.3 Network Generation Service

The indirect or accumulated effect of disrupted genes on the biological regulatory mechanisms harbors the potential to be informative in clinical decision-making. The interplay of such genes within their proximity has the potential to reveal possible drug resistance mechanisms. Moreover, examining disrupted genes in their network creates the possibility of finding alternative paths for drug intervention indirectly affecting an undruggable target. Hence, precision oncology requires examining the interactions between the observed biomarkers in their network neighborhood.

PeCaX infers the genes which are up- and downstream of a candidate target concerning gene regulatory and signaling pathways together with the drugs associated with the genes included in the networks. It enables the network-level examination of a patient's molecular aberrations by employing SMBL4j[205] to generate the gene-gene and gene-drug interaction networks. SBML4j is a service hub that provides biological network management and querying the graph data. It provides a standardized REST API which enables its integration into bioinformatics pipelines and workflows. It persists biological models in the Systems Biology Markup Language (SBML) format, a standard representation of biological networks, in a graph database. It is coded in the Java spring framework which is able to communicate with its Neo4j backend database.

SBML4j is a broad application with diverse functionality in network operations. Its usage in PeCaX depends on the knowledge graph generated from KEGG pathways and the drug-gene annotations obtained from DrugBank[177,207]. Given a set of genes with HGNC symbols as identifiers, SBML4j creates a sub-network of the paths connecting them and extends the network with their first neighbors.

PeCaX communicates with SBML4j through its REST APIs. Using the POST request `/overview`, PeCaX publishes the list of gene symbols as JSON object to trigger the sub-network generation. The universally unique identifier (UUID) of the generated network is provided by SBML4j as a response and the network is stored in its local database. To access the network's GraphML content, PeCaX sends a GET request to SBML4j's `/networks/UUID` endpoint. The resulting networks are annotated with the pathway information through a POST request, `/mapping/UUID`. PeCaX conducts this operation for the list of genes given in every table category separately to present the corresponding networks of every table.

Additionally, SBML4j obtains the list of driver genes from the case reports and integrate the driver gene information into the mapped network. SBML4j also provides links to the additional publicly available sources i.e., ENCODE, Ensembl, HGNC, MD Anderson, KEGG, and UniProt as node annotations, which are processed by PeCaX and included in the user interface for enabling users to expand their investigation.

SBML4j is deployed with Docker as a part of PeCaX's micro service-oriented application. It is based on a Neo4j database image mounted to a local Docker volume for storing the data needed for the network database such as the configurations, data sources, and plugins. Another image it depends on contains the SBML4j service itself with a data volume assigned to keep its log files.

4.2.2.4 GUI Service

Since PeCaX's main user group is MTBs, the most important design principle was achieving high user acceptance. It required minimizing the programmatic efforts to use the tool and presenting the results in the most concise way to not overwhelm users with the deluge of data attributing vital importance to the design of the graphical user interface and the biological data visualization methods.

Our GUI's first main function is to ensure the user's initial communication with the tool in providing inputs and receiving the results of the analysis. Another crucial function is to enable users to conduct operations on the results, by providing them in a concise, structured, and interactive manner for users to understand the data and gain insights from the results. For this purpose, we implemented two main modules, one is to visualize the clinically annotated genomics data and the other one is for the networks generated from the disrupted genes.

The case report visualization module, implemented in the Nuxt.js framework, displays the results generated by ClinVAP in an interactive tabular form with the same table structure as described in Chapter 3.4.3. The network visualization module employs BioGraphVisart[206], which is an automated tool designed for visualizing biological networks based on Gene Regulatory Networks (GRNs). BioGraphVisart is a web-based tool written in the Node.js JavaScript library that can handle many connections concurrently. Additionally, it inherits the functionality of Cytoscape via Cytoscape.js. It surpasses Cytoscape's functionality with its features to include extra node annotations such as grouping the genes based on their pathway involvement and allows to incorporate additional node types such as drugs with the targeting relationship to the genes. BioGraphVisart uses GraphML files as input and creates a Cytoscape core network object from the nodes and edges obtained from the input file. It initializes the network object from the nodes and automates i) the layout of the network graph, ii) the labeling of nodes (genes, drugs) and edges (interactions), iii) the edge style for different interaction types, iv) the node coloring according to easily modifiable node attributes, and v) the generation of legends. GUI service and the BioGraphVisart are deployed separately with Docker containerization technology with GUI services dependent on BioGraphVisart.

4.2.2.5 Data Management

Data management between the microservices of GUI, network generation, and clinical annotation is ensured with a local ArangoDB database. Upon the start of the analysis, an empty ArangoDB database collection is created with a unique job ID. When the user uploads the variant files and enters the parameters, the empty JSON collection is appended by the arguments of the *ICD10 code*, *assembly version*, and the *filtering option*. The arguments are then posted to ClinVAP via its REST API. Once the clinical annotation of the uploaded variants is finished, the resulting case report is added to the database collection and the original variant files are deleted from PeCaX. PeCaX sends the list of genes and their annotation labels to SBML4j to generate the networks. The

UUID of the generated network is sent back to the database and stored in the project collection. The corresponding networks are saved to the Neo4j database and sent to the BioGraphVisart for visualization. The custom notes made by the users through the GUI are also appended to the analysis collection.

The main advantage of the data management module is to store the analysis results and enable the user to access the previous analysis results with the job id and the project name. It also offers the advantage of performing multiple analyses gathered in one collection (e.g., holding the analysis results of different patients presented in one MTB session).

4.2.2.6 Deployment and Availability

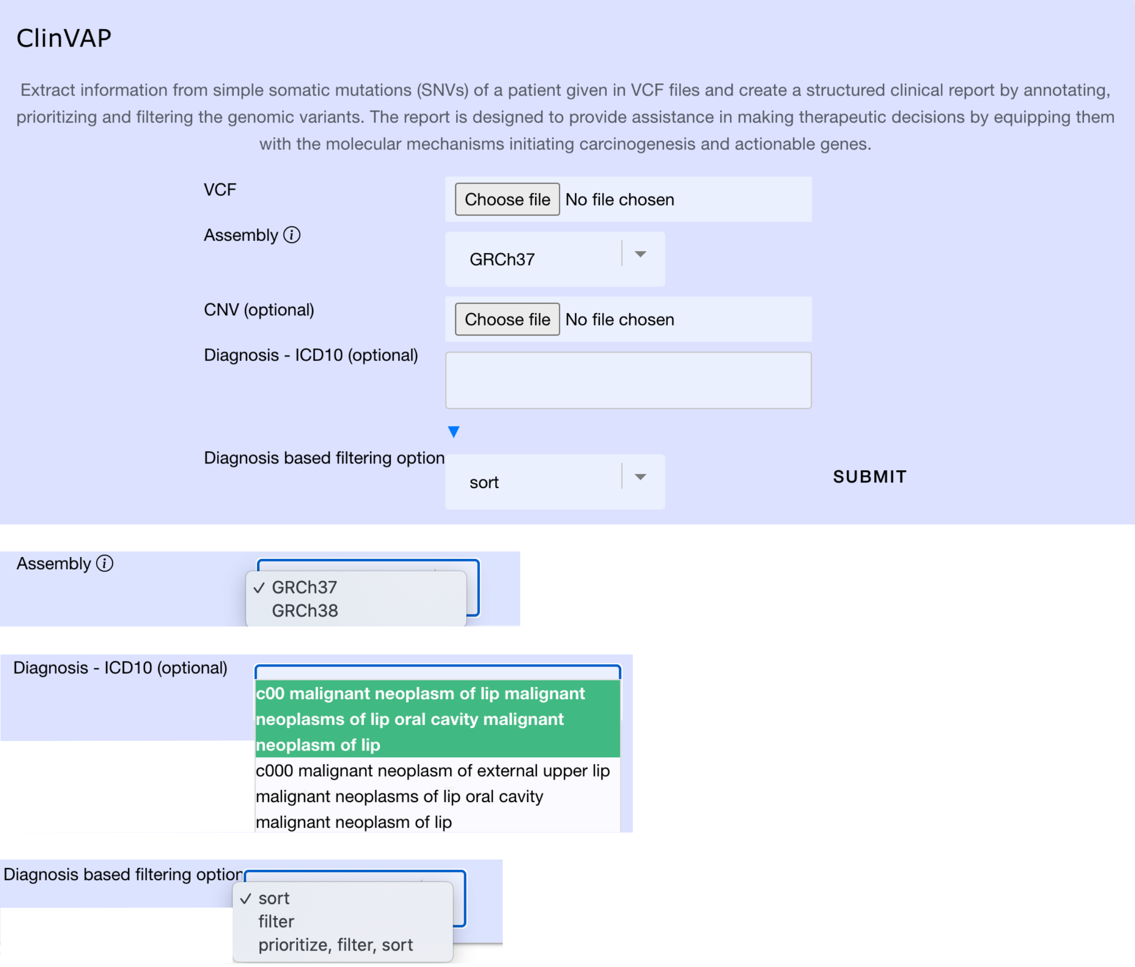
PeCaX can be installed locally on a personal computer or for groups of users in an access-controlled intranet. It is deployed with pre-built docker containers and orchestrated with docker-compose. Complete deployment is enabled with all the dependencies required for the software installation and the configuration via Docker. The reproducibility of the results is also achieved with containerization since all the dependencies and the source code itself are built on fixed package versions, tools, and data sources. Local Docker data volumes are used to store and share the data between PeCaX's service components.

PeCaX is an open-source tool under the MIT license. The source code is maintained in GitHub repository `https://github.com/KohlbacherLab/PeCaX-docker`. All the Docker images are publicly available in the Docker hub PeCaX repository, `https://hub.docker.com/repositories/pecax`. Due to its microservice architecture, the changes in the microservices are automatically reflected in the PeCaX software due to the continuous integration functionality of the Docker hub based on the new commits on GitHub.

4.3 Results

4.3.1 Initialization and Data upload

PeCaX requires a local submission of the data to initialize the analysis. To trigger the automated process the main requirement is the VCF file upload. If a TSV file containing the list of CNVs is provided, PeCaX factors it into the clinical annotation process. Other optional parameters are 1) the human genome assembly version which was used in the variant calling step of the NGS pipeline with a default value GRCh37, 2) the diagnosis given as ICD10 code, 3) the filtering option only factored into the analysis as described in Chapter 3.3.2.2 if the diagnosis is provided (Figure 4.3). When the data upload and parameter selection is completed, PeCaX groups the data sets and the results and attributes them to a specific project which provides the advantage of saving the session information which might include more than one patient entity.



ClinVAP

Extract information from simple somatic mutations (SNVs) of a patient given in VCF files and create a structured clinical report by annotating, prioritizing and filtering the genomic variants. The report is designed to provide assistance in making therapeutic decisions by equipping them with the molecular mechanisms initiating carcinogenesis and actionable genes.

VCF: Choose file No file chosen

Assembly ⓘ: GRCh37

CNV (optional): Choose file No file chosen

Diagnosis - ICD10 (optional):

Diagnosis based filtering option: sort

SUBMIT

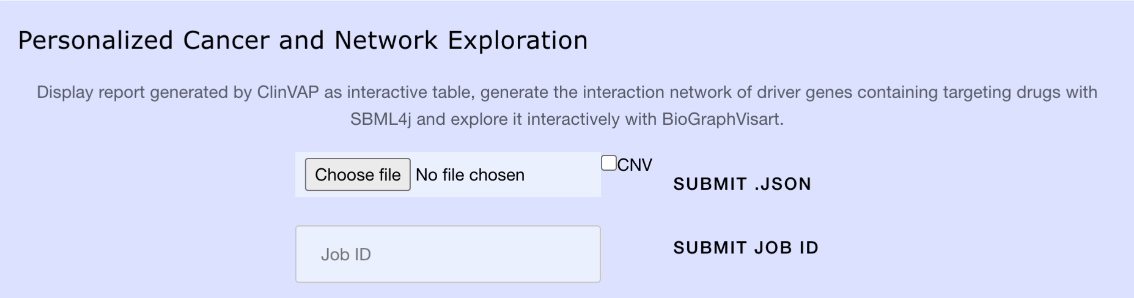
Assembly ⓘ
✓ GRCh37
GRCh38

Diagnosis - ICD10 (optional)
c00 malignant neoplasm of lip malignant neoplasms of lip oral cavity malignant neoplasm of lip
c000 malignant neoplasm of external upper lip malignant neoplasms of lip oral cavity malignant neoplasm of lip

Diagnosis based filtering option
✓ sort
filter
prioritize, filter, sort

Figure 4.3: PeCaX data upload and parameter initialization interface. The values of *assembly*, *diagnosis*, and *diagnosis based filtering* is selected from the drop-down menu. If the optional parameter *diagnosis* is not provided, the annotation skips the diagnosis based filtering.

Another option to start the PeCaX session is to upload a previously downloaded JSON file or to enter the corresponding job ID which will skip the clinical annotation and directly load the results to GUI. All job IDs of a given project are listed on a subpage where the user can select and delete them individually. Deletion of a job ID removes all information stored for this ID in the project database as well as the generated networks from the network database to ensure data privacy. In the same manner, it is also possible to delete the entire project (Figure 4.4).



Personalized Cancer and Network Exploration

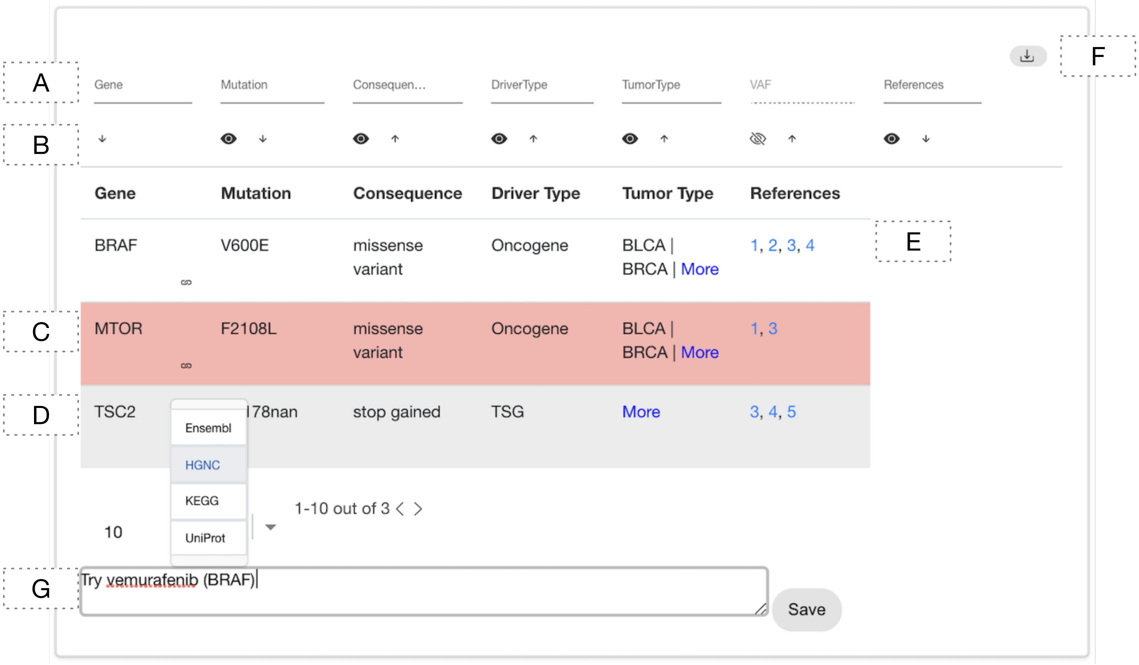
Display report generated by ClinVAP as interactive table, generate the interaction network of driver genes containing targeting drugs with SBML4j and explore it interactively with BioGraphVisart.

Figure 4.4: PeCaX initialization from previous sessions. The results can be retrieved with the job id, or the previously downloaded JSON file.

4.3.2 Interactive visualizations

The results are rendered as interactive, responsive tables. The structure of the case report follows the same representation as described in Chapter 3.4.3 where the tables are separated based on 1) the list of known driver genes observed in the patient, 2) the list of drugs with the evidence of targeting a specific variant of the gene and their documented drug response, 3) the therapies that have evidence of targeting the affected gene, 4) the list of cancer drugs targeting the mutated gene mechanistically, 5) the list of variant-drug pairs known as causing adverse effects, 6) the list of scientific publications supporting the associations found for the mutational profile and the therapy options, 7) the complete list of non-synonymous variants observed with their dbSNP and COSMIC IDs. Each table has an information text displayed upon selection to provide the table explanations which is useful for users who are not familiar with the report structure.

The GUI enables users to query, sort, and filter the content of the columns of each table (Figure 4.5A,B). The table view supports a wide range of table operations to simplify navigation of the data such as hiding/showing columns (Figure 4.5B), highlighting rows across sections (Figure 4.5C), collapsing/expanding tables, searching the tables based on a column value even with a partial text, and choosing the number of displayed entries which is ten by default. For each gene listed, the tables contain cross-links to the external data sources (Figure 4.5D) Uniprot[208], KEGG[207], Ensembl[209] and HGNC[210] which are accessible through the drop-down menu next to the gene symbol. Similarly, the references of the extracted associations are directly linked to the web page of the related publications on PubMed (Figure 4.5E). The content of all external sources is shown in a separate browser tab upon selection. The GUI also allows users to download each table along with customized notes (Figure 4.5F). At the end of each section, the tables contain a text field dedicated to entering custom notes which are stored along with the annotation data(Figure 4.5 G).



Gene	Mutation	Consequence	Driver Type	Tumor Type	References
BRAF	V600E	missense variant	Oncogene	BLCA | BRCA | More	1, 2, 3, 4
MTOR	F2108L	missense variant	Oncogene	BLCA | BRCA | More	1, 3
TSC2	78nan	stop gained	TSG	More	3, 4, 5

Try vemurafenib (BRAF)

Figure 4.5: PeCaX table interactive view. A and B) show the individual column operations (e.g., filtering, sorting). C) demonstrates the row highlight function. D) is the drop down menu for gene cross referencing. E) provides the links to the evidence publications. F) points to the table download button. G) shows the field for customized note taking.

The networks are generated separately for each table except the appendix tables if at least one gene symbol is associated with an entry in the SBML4j database. The networks are displayed side by side with their corresponding table (Figure 4.7). In the networks, the nodes are either the genes or the drugs. The edges represent the interactions between them such as signaling, regulation (Figure 4.6 A). Genes that are sourced from the tables are specified with the color red and labeled with their HGNC symbol (Figure 4.6 B). The drugs associated with any of the genes are represented with diamond-shaped nodes and labeled with their drug name (Figure 4.6 B). Multiple drug nodes targeting the same genes are merged into one expandable hub, providing the advantage of a more concise network representation. Different interaction types are depicted by different edge styles. If two nodes have multiple interactions, their edges are merged into one by default. Since drug and gene names may become very long, they are shortened, and moving the mouse over a node reveals the full node name. The same principle is applied to the edge types.

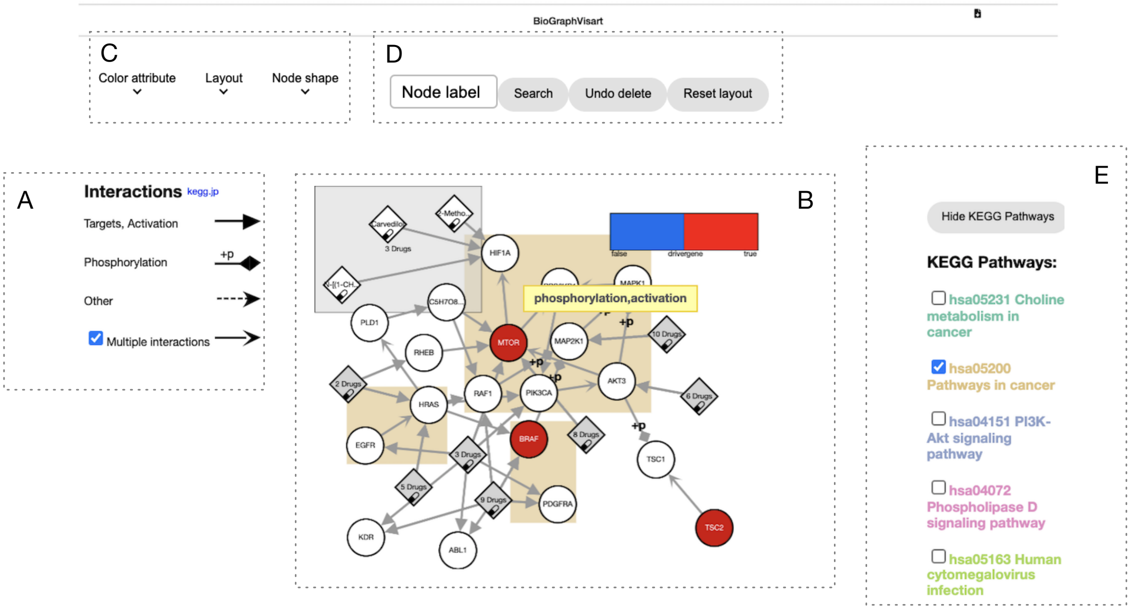
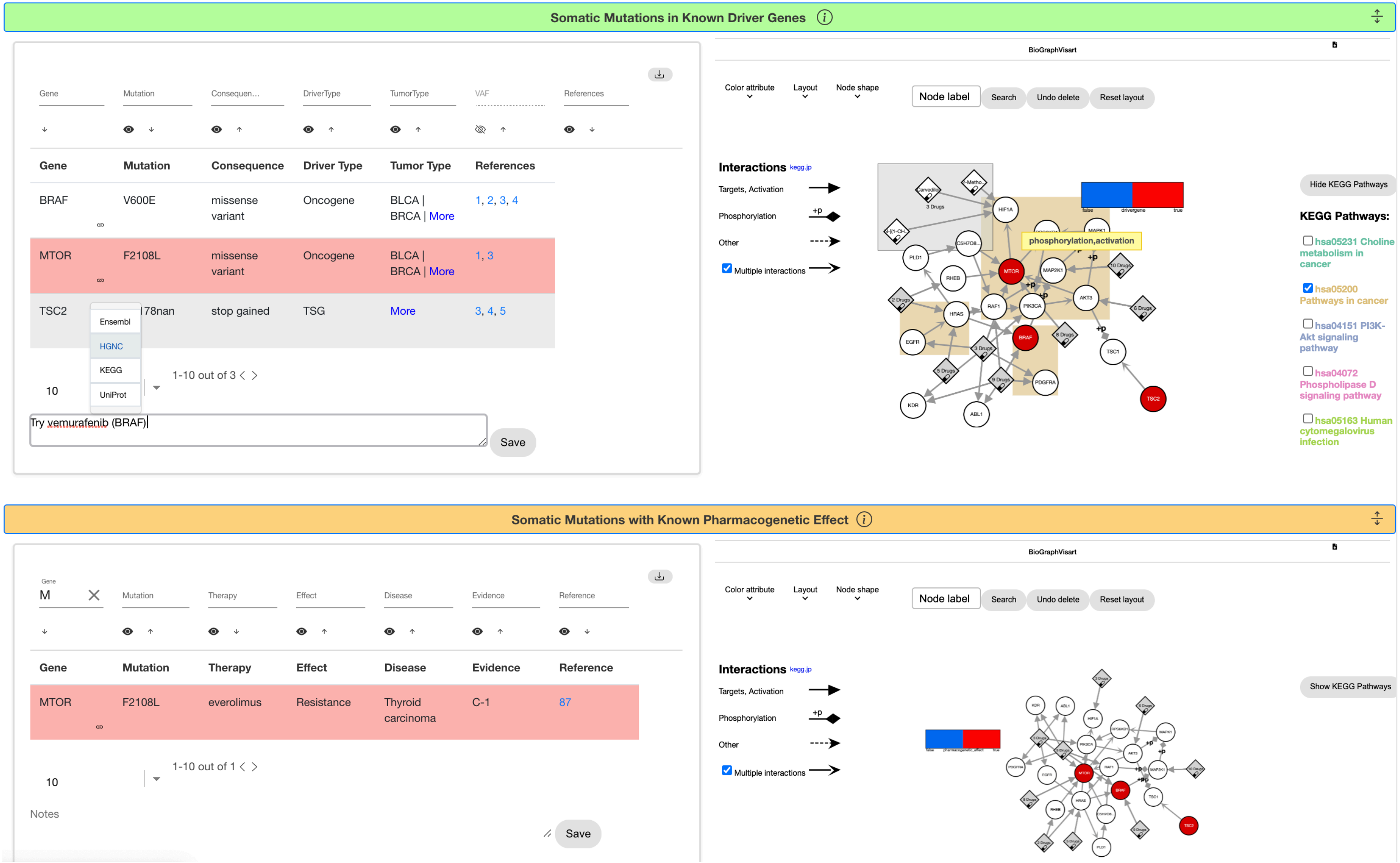


Figure 4.6: PeCaX network interactive view. A) shows the different interaction types and their symbols. B) represents displays the network properties such as the difference in the node shapes for drugs and genes, the highlighted regions based on their pathway involvement and collapsing the drug nodes and edges to provide a simpler view. C) shows the menu to alter the layout, node shape and colors. D) is the menu for node search, undo the layout changes and retrieve the deleted nodes. E) shows the pathway menu. Upon the selection of a pathway, corresponding regions are highlighted.

The default network layout is selected as Compound Spring Embedder[211] which can be replaced with additional four layouts from the drop-down menu (Figure 4.6 C). It also supports manual alteration of the network's layout by dragging the nodes, arranging the network at users' convenience, and modifying the content by deleting certain edges and the nodes. The networks have the search function by node label (Figure 4.6 D). The pathway information is displayed on the networks by grouping the genes and highlighting the associations extracted from KEGG pathways (Figure 4.6 E). Drug nodes harbor the link to a drug overview page that includes cross-references to external drug databases, i.e., Drugbank[149], HGNC[212], and PDB[213].



Somatic Mutations in Known Driver Genes

Gene	Mutation	Consequence	Driver Type	Tumor Type	References
BRAF	V600E	missense variant	Oncogene	BLCA | BRCA | More	1, 2, 3, 4
MTOR	F2108L	missense variant	Oncogene	BLCA | BRCA | More	1, 3
TSC2	178nan	stop gained	TSG	More	3, 4, 5

10 1-10 out of 3 < >

Try vemurafenib (BRAF)

Somatic Mutations with Known Pharmacogenetic Effect

Gene: M

Gene	Mutation	Therapy	Effect	Disease	Evidence	Reference
MTOR	F2108L	everolimus	Resistance	Thyroid carcinoma	C-1	87

10 1-10 out of 1 < >

Notes

Figure 4.7: PeCaX side by side view of tables and corresponding networks.

4.3.3 Exporting tables and networks

The entire case report including the custom user notes can be exported as a PDF file as well as the individual table sections. It is possible to download entire case report in JSON format which can be incorporated into other programmatic workflows or uploaded back to PeCaX to re-create the visualization. The gene-drug interaction networks are available for download individually in the formats PNG, SVG, and GraphML.

4.3.4 Performance

Since genome data is highly sensitive, its usage and sharing are regulated by laws. We created synthetic data to evaluate the performance of PeCaX. We downloaded the list of cancer biomarkers from the CGI database. We partitioned it into subsets each specific to a cancer type. For SNVs, we converted those files into VCF format by extracting the coordinates and other standard VCF fields. For CNVs, we prepared TSV files that are compatible with our tools. We chose nine cancer types as our test set, which vary in terms of the number of variants they contain (Table D.4) and thus provide a good measure of PeCaX's behavior with changing data sizes. All the example datasets are available at `https://github.com/KohlbacherLab/PeCaX-docker/tree/main/test_files`.

We measured the time from 1) job submission to case report display, 2) gene list submission to SBML4j to network display and 3) the job submission to display of the entire results. For each test file, we run the analysis 3 times. The performance in terms of average time is provided in Table 4.1 and the detailed performance evaluation is given in Table D.5. The performance was evaluated on a MacBook Pro with a 3.1 GHz Dual-Core Intel Core i5 processor and 16 GB 2133 MHz LPDDR3 memory with a local installation of PeCaX.

PeCaX needs about 92 s on average to analyze the VCF data and display the results. When the CNV file is included, the time required increases to 352 s. The main reason for the increase is the genenames.org REST API calls to retrieve the gene symbols of the provided gene list. The average time PeCaX requires to create and display the networks is 58.19 s which is more than the time required for report generation instinctively due to the computationally expensive network operations such as tree traversing. Overall, PeCaX needs 205 s on average for the analysis of the data until the results are displayed.

Input file	Clinical Annotation	Network Generation	Overall
SNV	45.5	58.2	91.8
SNV & CNV	177.4	203.4	352.4
Overall	103.2	121.7	205.8

Table 4.1: Average processing time [s].

4.4 Discussion

Since NGS data has become widely accessible with the drop in sequencing costs, it has entered the routine practice in oncology clinics as a foundation of targeted therapy applications. Even though the targeted therapy strategies increase the treatment efficiency, translating genomics data into clinical implications remains challenging due to the complexity and the multiplicity of the bioinformatics tools. Requiring programmatic knowledge decreases the clinical adaptability of such tools and instigates personalized medicine units to conduct time-consuming and error-prone manual annotations. Besides the strenuous efforts to analyze the large influx of data, the clinicians face other difficulties in finding strategies for undruggable targets and assessing the cumulative effect of the variants leading to the drug resistance. Resolving these issues requires incorporating the interaction networks in the analysis as an additional layer of information which further complicates the entire process.

We addressed these issues with PeCaX, which is a novel tool developed as a decision support system for oncology to analyze the genomics data for extracting gene biomarkers and exploring their systematic effect on the interaction networks in developing targeted therapy strategies tailored to the unique molecular profile of the patients. It automates the entire workflow by eliminating the need for manual annotations and data preparation. It contains an interactive GUI that enables users to initialize the workflow and obtain the results without requiring any programmatic knowledge. It is deployed as Docker containers with all the package and data dependencies which not only attribute complete reproducibility to PeCaX but also provide easy installation since it is independent of the operating system. It ensures data security by keeping the data in the local system and deleting the input files as soon as the analysis is finished. Besides the convenience it provides due to its software design and architecture, the combination of clinical annotation, interaction networks visualizing variants in their pathway context, and interactive web-based visualizations make PeCaX unique in the precision oncology practice.

PeCaX is designed for MTBs with a high focus on user adaptability which implies that its powerful back-end operations are based on the MTB case preparation requirements and separated from the presentation layer. While its MTB-specific design increases its adaptability in oncological clinics, it may prompt limitations to some user groups such as the researchers aiming to customize the VCF operations, investigate the variants of unknown significance, conduct cohort studies, and make inferences based on patient similarity.

To reduce the complexity, the clinical annotation step uses a standard filter on the quality control metrics of the called variants. It accepts the variants labeled as *PASS* in the VCF file, which means all the quality control measures for those reads were over the cut-off threshold. This might be a limitation for users who require to customize the quality control metrics to filter variants prior to the variant effect prediction. Such functionality would require a GUI for VCF file content visualizations which do not fall within the focus of our tool since it contradicts our aim to eliminate manual processing and automatize the analysis pipeline. Thus, the limitations on VCF operations would require users to implement these operations prior to the import of the VCF into PeCaX.

In a similar manner to variant filtering, clinical annotation services apply pre-defined *de facto* filters to the predicted consequences of the remaining variants. Next to filtering the low-impact variants, it also does not report the ones falling into the non-coding regions and labeled as modifier impact by its annotation tool, which implies that either the prediction is difficult or there is no evidence of the impact. Even though ClinVAP provides a complete list of the non-synonymous variants as an appendix regardless of their actionability to the case reports, the SNVs contained in non-coding regions with the modifier impact are not reported. Although these filters are found to be robust for the MTB use cases, filtering modifier impact variants might be a limitation for users who aim to investigate the variants of unknown significance.

PeCaX manages the data on a case-by-case basis and treats every case individually due to its MTB-specific design. However, it has considerable potential to serve as a data analysis medium in large cohort analysis which requires additional features of batch data upload and patient similarity assessment to provide stratified clusters based on molecular profiles. In future work, we plan to allow the upload of multiple

VCF files at once and provide patient similarity based on their network overlap which has already been implemented in the stand-alone version of the visualization module, BioGraphVisart[206].

In conclusion, PeCaX offers a user-friendly platform for MTBs to perform case preparation and detailed case investigation. The combination of variant annotations with interaction networks holds the potential to unravel complex phenotypes. PeCaX empowers clinicians and researchers to navigate large-scale datasets, foster evidence-based decision-making and improve healthcare efficiency in oncology clinics.

Chapter 5

Clinical Assessment of Evidence-based Reporting Strategy in the Precision Oncology Workflow

5.1 Introduction

The advancements in NGS technologies have increased the knowledge of underlying molecular disease mechanisms, which revealed the effect of inter-individual variation on disease progression and therapeutic response. These information have been extrapolated to clinics for establishing biomarker-driven targeted therapies. Clinical translation of the high-throughput data created a paradigm shift towards precision oncology due to cancer being a complex and individualized genetic disease[7]. Using NGS has became mainstream clinical practice arising from the need for consistent frameworks to evaluate the diagnostic and predictive biomarkers. In oncology clinics, developing such workflows is delegated to institutional MTBs which are multidisciplinary committees involving experts from related disciplines such as oncology, biology, bioinformatics, pathology, and genetics. Although it is a non-standardized procedure varying between institutes, the underpinning of MTB operations is constructed by patient selection, tumor profiling, clinical annotation, and treatment strategy assessment[214]. Patients for whom the conventional therapy options are exhausted are enrolled in the MTB system. The discussion content emanates from diagnostics reports containing an initial set of pre-processed and -filtered variants produced by NGS laboratories. The identification of actionable variants from those reports mainly relies on labor-intensive and error-prone manual processing. Discussion content is used by the committee to evaluate the therapy options with possible outcomes (e.g., enrollment of patients in clinical studies or suggesting off-label drugs with the evidence of treatment response). Even though it is viewed as an unproven hypothesis[215], precision oncology has been demonstrated to increase the healthcare efficiency and/or overall survival times for various cancer types such as ovarian cancer, colorectal cancer, non-small-cell lung cancer, metastatic renal cell carcinoma[12–14,216]. However, the major bottleneck of the MTB process is that the entire process lacks reliability and reproducibility[25]. Due to the absence of standard workflows, different MTBs have been found to exhibit low agreement on the same patients[26]. The variability in decision-making heavily relies on the differences in the interpretation of the clinical assessment of the molecular profiles.

Diagnostic reports provided by NGS laboratories cannot be directly translated into clinical action due to liability issues and lack of information prioritization[217], requiring specialists to evaluate the reported aberrations[218]. Moreover, these reports introduce discrepancies to MTB procedures due to non-standardized annotations made by a non-transparent set of operations. Since there is no standard scale of assigning clinical significance to reported alterations[219], the efforts rely heavily on MTB workflows utilizing a small set of technologies and biomarker databases, manually. The complexity of assessing actionability/druggability is precluding the effective clinical use of genomics data and thus pointing to the need for standardized systems utilizing a large pool of data sources and bioinformatics tools. However, the standardization solely relies on the expertise of individual MTBs, mostly with the efforts of creating an in-house biomarker database and a standard operating procedure[25,171]. Available decision support tools do not match the extensive needs of MTBs. The major issues are not supporting standard data formats and the need for input preparation, the absence of the gene disruption level assessment, and relying on the prioritized list of variants provided by users[145,169]. Moreover, many of those tools hinder patient privacy due to sending sensitive data to servers[145,170,217].

Retrospective comparisons as a performance measure of the available tools focused on a small set of data. The coverage was limited to the final MTB recommendations without extending the analysis to the overall reported molecular diagnostics[169,171]. The assessments based on follow-up patient data on survival intervals do not focus on the discussion content overlap but on the utilities of the entire MTB workflow[25,220]. Another performance measure is to calculate the percentage of the cases for which the tools were able to make a targeted therapy suggestion[171] without comparing the results to the MTBs' decisions.

We previously published ClinVAP as a fully automated and self-contained pipeline to generate case reports, ensuring data privacy, covering all aspects of therapy relevant molecular players, working with standard data formats, and ensuring reproducibility[161]. In this work, we implemented a comparison analysis to demonstrate the utility of ClinVAP in precision oncology workflow using retrospective neuro-oncology cases from University Hospital Tübingen (UKT). We showed its content-wise equivalence with the current MTB case preparation step and evaluated its advantages in identifying actionable variants.

5.2 Materials and Methods

5.2.1 Retrospective Patient Data Pre-processing

The target population is selected among neuro-oncology cases that were enrolled in the personalized medicine program between January 2016 - April 2020. The study design and the data access request were approved by the responsible internal review board (IRB, 192/2020BO2) for the patients who are older than 18 years and signed a broad consent for the scientific use of their data. We obtained genomics data in VCF file for SNVs and variant exports as *.TXT* files for CNVs. Additionally, we received case discussion content including the therapy-relevant variants and the MTB protocol excerpts stating the clinical action(s) suggested by the committee. We digitized the case presentations from *PDF* to *JSON* format concordant to the categories included in the slides. We extracted the selected targets and therapeutic options from the protocols with the suggestions' priority rank for a given case and included them in the case evaluation JSON file (evaluation JSON object 5.4). Case digitization revealed ambiguities such as the cases that were sequenced twice or recommended for a second discussion based on additional immunohistochemical staining (IHC) test results. For such cases, the evaluation files were compiled from the newest data when the complete case information content was available. Otherwise, the evaluation content was created from the early data since case completeness was prioritized. The same strategy was applied to the cases for which a second discussion was suggested but did not take place. The cases were excluded if 1) the case was not presented to the MTB committee, 2) the case data set were not complete i.e., missing variant export or VCF files, not being presented to MTB, missing MTB case presentation, or the resulting protocol, or 3) the case was not subject to targeted therapy or found eligible for immunotherapy.

```
{
    "point_mutation_schema": {
        "type": "dict",
        "properties": {
            "cell_type": {"type": "str"},
            "gene_symbol": {"type": "str"},
            "variant": {"type": "str"},
            "hgvsc": {"type": "str"},
            "hgvsp": {"type": "str"},
            "predicted_consequence": {"type": "str"},
            "functional_change": {"type": "str"},
            "priority": {"type": "str"}
        }
    }
}

defined vocabularies:
cell_type: somatic, germline
priority: high, null, low
```

JSON object 5.1: Point mutation schema for SNVs and small INDELs. The information fields are used to represent the content of SNVs, insertions, deletions and INDELs provided in MTB case presentations.

```
{
    "rearrangement_schema": {
        "type": "dict",
        "properties": {
            "gene_symbol": {"type": "str"},
            "location": {"type": "str"},
            "functional_change": {"type": "str"},
            "priority": {"type": "str"}
        }
    }
}
defined vocabulary:
priority: high, null, low
```

JSON object 5.2: Chromosomal re-arrangement schema. The information fields are used to represent the content of fusions and inversions provided in MTB case presentations.

```

{
    "drug_schema": {
        "type": "dict",
        "properties": {
            "drug_name": {"type": "str"},
            "drugbank_id": {"type": "str"},
            "drug_drug_relationship": {"type": "str"},
            "drug_class": {"type": "str"},
            "evidence_level": {"type": "str"},
            "target": {"type": "str"}
        }
    }
}
defined vocabulary:
drug_drug_relationship: combination, substitution
```

JSON object 5.3: Drug schema. The information fields are used to represent the content of the therapeutics that was either suggested or selected through the MTB process.

```
{
    "type": "object",
    "properties": {
        "mtb_id": {"type": "str"},
        "patient_id": {"type": "str"},
        "sample_id": {"type": "str"},
        "target": {"type": "str"},
        "suggestion": {"type": "boolean"},
        "icd_10": {"type": "str"},
        "driver_genes": {
            "type": "array",
            "properties": {
                "gene_symbol": {"type": "str"},
                "variant": {"type": "str"},
                "hgvsc": {"type": "str"},
                "hgvsp": {"type": "str"},
                "predicted_consequence": {"type": "str"},
                "variant_type": {"type": "str"},
                "driver_type": {"type": "str"},
                "priority": {"type": "str"}
            }
        },
        "snv": {
            "type": "array",
            "properties": "point_mutation_schema"
        },
        "insertion": {
            "type": "array",
            "properties": "point_mutation_schema"
        },
        "deletion": {
            "type": "array",
            "properties": "point_mutation_schema"
        },
        "indel": {
            "type": "array",
            "properties": "point_mutation_schema"
        },
        "cnv": {
            "type": "array",
            "properties": {
                "cell_type": {"type": "str"},
                "gene_symbol": {"type": "str"},
                "cnv_type": {"type": "str"},
                "functional_change": {"type": "str"},
                "allelic_change": {"type": "str"},
                "priority": {"type": "str"}
            }
        },
        "fusion": {
            "type": "array",
            "properties": "rearrangement_schema"
        },
        "inversion": {
            "type": "array",
            "properties": "rearrangement_schema"
        },
        "suggested_targets": {
            "type": "array",
            "properties": {
                "gene_symbol": {"type": "str"},
                "variant_type": {"type": "str"}
            }
        },
        "suggested_drugs": {
            "type": "array",
            "properties": "drug_schema"
        },
        "selected_targets": {
            "type": "array",
            "properties": {
                "gene_symbol": {"type": "str"},
                "variant_type": {"type": "str"},
                "recommendation_type": {"type": "str"}
            }
        },
        "selected_drug": {
            "type": "array",
            "properties": "drug_schema"
        }
    }
}
defined vocabularies:
cnv_type: amplification, deletion
variant_type: snv, insertion, deletion, indel, cnv, fusion,
    inversion
recommendation_type: main recommendation, alternative
   recommendation
priority: high, null, low
```

JSON object 5.4: Evaluation schema. It represents the information of an entire case.

5.2.2 Annotation Knowledge Base Coverage on MTB Case Contents

Driver genes exhibit the properties of potential therapeutic targets[145,162], which attributes importance to their categorization on diagnostic reports. Thus, broad coverage of our annotation KB (Section 3.4.2) over the genes reported as drivers in MTB case presentations is of critical importance. To measure the KB's driver gene representation, we calculated the percentage of the genes that are labelled as driver in both KB and the MTB case presentations over the total number of MTB driver genes.

We investigated MTB driver genes without coverage in the KB through a manual PubMed search to find evidence of the established driver labeling for them to identify the reasons for the misclassifications.

Target identification is an arduous task requiring a total mutational profile investigation. Thus, not only driver genes contribute to target prioritization but also the variants with damaging and non- tolerated impact on the protein function. Consequently, it is of crucial importance to have a broad KB representation of the overall molecular profile utilized in MTB. We used all the mutated genes that were reported on the MTB case presentations at least once, to assess the KB's coverage over the patients' molecular profile favored by MTB. We expanded this strategy to different variant priority levels to see the coverage distribution over the therapeutically related information. We distributed MTB molecular data into different priority levels based on their MTB labeling. All the genes that were mentioned at least once in the MTB case slides were categorized as "reported genes". "Priority genes" assigned to MTB genes given as "potential therapy targets" in the MTB cases. "Target candidates" was used for the genes that were suggested as targets by the case preparation committee. Genes that were recommended as targets and those selected as the main target as a result of the committee's discussion were labeled as "targets" and "main targets", respectively. We then calculated the coverage individually for every priority level as a percentage using the amount of overlap with the KB over the total number of genes of a given level.

5.2.3 Content Comparison on an Individual Case Level

We measured the robustness of ClinVAP's variant filtering constraints by showing the content equality of its case reports to the MTB case presentations, to exhibit its utility at the case level. We compared the list of MTB genes with the ones predicted as non-synonymous variants by ClinVAP on a case-specific level. We then calculated the case overlap ratio of the number of genes that were shared between two instruments over the total number of MTB genes. We only considered SNVs in the analysis since ClinVAP does not apply impact-based filters on the reported list of CNVs.

We investigated the underlying reasons for the non-overlapping MTB genes (MTB-specific genes) by using unfiltered ClinVAP functional variant annotations. We obtained variant impacts and SO terms of the consequences predicted by Ensembl VEP for all mutations, and by SIFT and PolyPhen for the mutations affecting the gene coding regions. We gathered similar information for MTB-specific genes to investigate the annotation differences, the severity of them and their contribution to MTB's decisions.

5.2.4 ClinVAP Case Report Coverage on MTB Recommendations

We demonstrated ClinVAP's utility in providing a complete framework of druggability by testing the reports' coverage on drug-gene associations that were recommended by MTB as a therapeutic action plan. We first conducted a binary search on covering the recommended gene-drug pairs in ClinVAP reports. Then, we investigated the cases where a full match was not observed. We excluded the cases in which 1) the target was a germline variant, 2) the suggested drug-target association is not well established in the literature, but was made based on expert opinion, 3) the suggestion was not made entirely on a molecular rationale but based on a known increase in the efficiency of the suggested drug by a variant, 4) the genes that are not given in the somatic CNV files, but falling -entirely or partially- to the reported CNV coordinates.

We assigned cases to "zero coverage" if the recommendation was not included in the ClinVAP report. We used "partial coverage" if, among multiple selected targets, at least one drug-gene pair was included in the ClinVAP report. We used the same term for the cases where ClinVAP offers another drug from the same class of the main recommendation. We assigned the cases to "full coverage" if their recommendations were covered entirely.

5.3 Results

5.3.1 Patient Cohort

We obtained 131 MTB identifiers of the patients enrolled in the targeted cancer therapy program. 40 of the total cases were not referred to the MTB committee due to either decision of continuing conventional therapies, patient withdrawal, or death before the start of the treatment. Among 71 patients whose case was discussed by the MTB committee, only 44 of them had an archived *VCF* and CNV variant export files. We filtered the data based on the existence of actionable targets with therapeutic suggestions which led to 33 eligible cases.

5.3.2 KB's Coverage on Overall MTB Case Content

Driver genes. Driver genes exhibit a high potential for actionability due to their function in tumorigenesis. Thus, pinpointing the driver genes of a case together with their druggability has a profound impact on unveiling promising therapy options. To demonstrate our annotation source's comprehensiveness, we calculated its coverage over the MTB driver genes. We revealed that our KB covers 99% of the driver genes reported in MTB case presentations. Driver gene annotation is underrepresented in MTB case presentations due to the different reporting formats and the content used by the sequencing labs. It suggests that MTB's decision-making is agnostic to the driver gene annotation and the selection of a set of candidate genes depends on expert knowledge. The comprehensiveness of our KB in driver annotation would enable MTBs to obtain a complete driver landscape and further expand their expert-knowledge-based list of significant genes.

The initial coverage on MTB driver genes was found as 89% with 11 misclassifications. MTB diagnostic reports did not include the information source for their driver annotation which decreases the reliability of the provided information. Our literature review revealed that even though most of the misclassified genes are known as increasing cancer susceptibility, driver classification is not yet established for nine of them[195,221–230]. We could only find the driver associations for FRS2 and PAK1 (Figure 5.1A)[231–234]. After the literature review, our KB's driver gene coverage increased to 99%.

Next, we uncovered the impact of not classifying FRS2 and PAK1 as drivers. PAK1 was seen in two cases, identified as driver only for one of them, and neither suggested nor selected as a therapy target. FRS2 was observed in three cases, categorized as a driver in two of them, and selected as a target for one of those cases. However, the case that it was selected did not include driver gene information which points to inconsistent annotations between the cases harboring similar aberrations. We concluded that the effect of FRS2 was well-known to MTB and their decision was agnostic to its driver gene annotation. To prove this misclassification does not result in missing target-drug content, we searched our KB for therapeutics targeting FRS2 which revealed that KB links it to therapeutic options which would still provide a good content coverage for MTB.

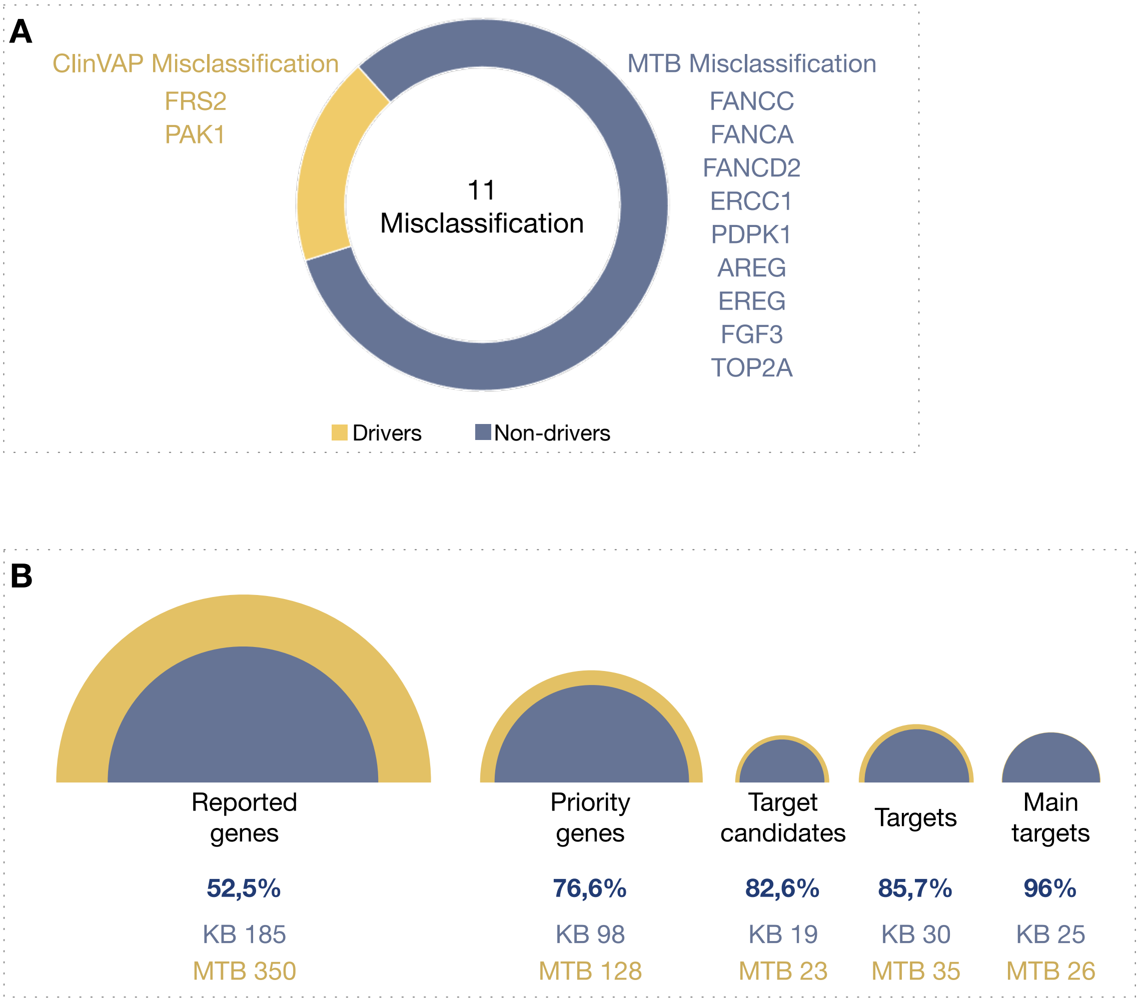


Figure 5.1: Annotation knowledge bases (KB) coverage on MTB case content. A) KB's coverage on the driver genes reported in MTB case discussion contents. Among MTB driver genes, 11 of them do not have the same labeling in our KB. In the literature, no established driver annotation was found for nine of them which are shown as MTB misclassifications. Our KB has only two driver misclassifications for the genes FRS2 and PAK1. B) KB's coverage on the genes that are reported in MTB case discussion contents. Its coverage is calculated separately for every gene category provided in the case contents. Initial coverage on the genes reported at least once in MTB case presentations was 52.5%. The KB's coverage increased with the therapeutic relevance of the reported genes.

MTB's content of altered genes. MTB case presentations provide a complete mutational profile of patients categorized as "potential therapy targets", "strong therapy candidates", and "other observed mutations". However, regardless of the completeness of a case's mutational profile, MTB case presentations do not concentrate on the actionability/druggability of the aberrations and report non-therapy informant variations. Our annotation KB is specialized in therapy relevance exhibiting high coverage on the "strong target candidates", pointing to its ability to precise MTB content towards actionability. The differences found between the strong therapy targets and the KB were mainly for the connections that are not well established but depend on expert opinion.

We measured the KB's coverage over the genes reported in MTB case content separately for every priority level that we assigned based on their labeling on the case presentations and protocol excerpts (5.1B). The coverage over the genes that were mentioned in the MTB slides at least once was 52.8%. This increased to 96% on the MTB main selected targets and 87.5% on MTB selected targets.

We investigated the differences between the KB's content and the MTB's selected targets. For the main selected targets, the only difference was ERG-TMPRSS2 fusion whose case received the combination of PARP inhibitors with radiotherapy as a treatment suggestion. Even though our KB annotates ERG-TMPRSS2 as a driver aberration and lists PARP inhibitors for TMPRSS2 fusions, we could not find established targeted therapy options with clinical evidence for ERG-TMPRSS. Evaluating the biomarker together with its treatment suggestion revealed that PARP inhibitors increase the effect of radiotherapy[235]. Thus, this case was not subject to targeted therapy and was out of ClinVAP's scope.

Another difference was found for NBN, PMS2, POLQ, and C11orf30. C11orf30 was suggested for one case together with germline BRCA2 mutation as a target for PARP inhibitors. C11orf30 is frequently over-expressed in cancer and its amplification is known as impairing the DNA damage repair[236]. However, its association with increased sensitivity to PARP inhibitors is not yet fully established[237]. The lack of evidence both in literature and the MTB case protocol suggests that the therapy option was identified based on expert opinion.

We identified that NBN, PMS2, and POLQ were observed in the same case and reported with BRCA1/2 which was a clustered set of mutations on the DNA repair genes. After visiting MTB's evidence publications, the difference was reduced to only NBN because two out of three pieces of evidence belong to clinical trials with no focus on genomic biomarkers[238,239] and the third one only provides evidence on NBN[240]. PMS2 and POLQ have supporting roles identified based on expert knowledge due to their function in DNA repair mechanisms. Our KB provides the same evidence as MTB on the effect of PARP inhibitors for BRCA1/2 deletions which implies that the absence of NBN does not have any negative effect on the suggestions overall.

5.3.3 Case Report Content Analysis

We investigated the robustness of ClinVAP's variant filtering constraints by showing the content equality of its case reports to the corresponding MTB case presentations. ClinVAP provided the same annotations for the set of genes that were only reported in MTB case contents (MTB-specific genes) and filtered them out due to their predicted low impact. We observed that most MTB-specific genes did not contribute to MTB's decisions. While MTB content included many aberrations regardless of the severity of their potential impact, ClinVAP focused on reporting the high-impact variants with a higher potential to be selected as targets.

Case-by-case comparisons lead to the identification of 55 ClinVAP-specific variants from 14 cases and 59 MTB-specific variants from 23 cases. We could not investigate the reason that ClinVAP-specific variants were not reported by the sequencing centers, since we did not know their analysis configurations. We investigated the reporting differences by clustering MTB-specific genes based on their functional annotation.

17 of the MTB-specific genes were predicted as synonymous mutations where the impact was low implying that the change was harmless. ClinVAP had the same annotation for 16 of them (Figure 5.2A). One gene's coordinates were reported differently in the VCF file than the one in variant export which led ClinVAP to predict it as an intron variant.

23 MTB-specific genes were clustered as missense mutations which correspond to the amino acid changes on the protein-coding region with a moderate impact implying that it might change the protein's effectiveness. ClinVAP returned the same annotation for 20 of them (Figure 5.2B). Additionally, ClinVAP predicted 20 of them as "tolerated and benign" through SIFT and PolyPhen which was the main reason of their absence in ClinVAP reports. The impact for those variants in MTB diagnostic reports was either ambiguous or not available.

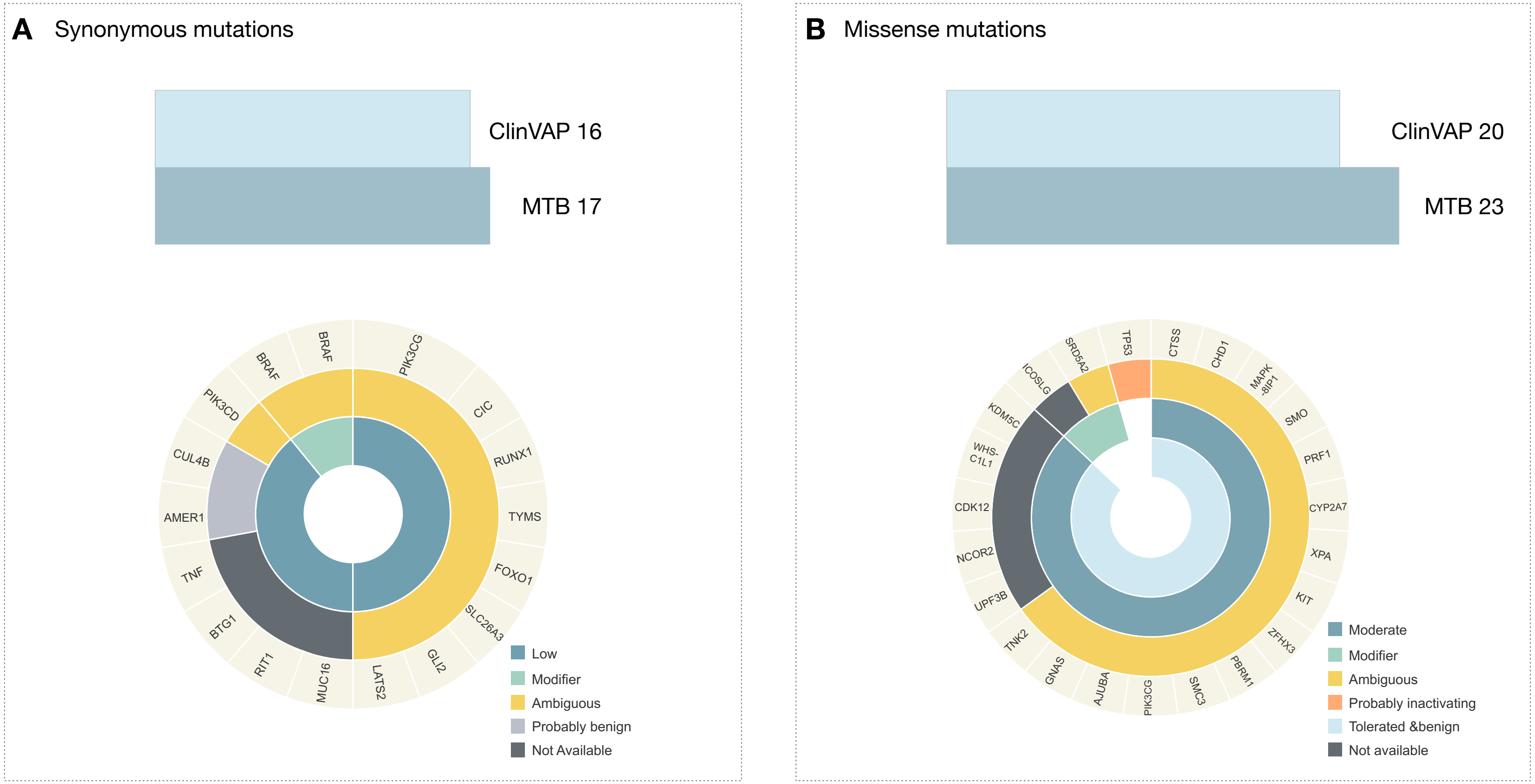


Figure 5.2: Annotation comparison for MTB-specific genes. A) Synonymous. 16 have the same ClinVAP annotation. Ensembl VEP and MTB predictions are in the inner and outer ring, respectively. They are not therapy informative, thus filtered by ClinVAP. B) Missense. The inner and middle layers are for SIFT-PolyPhen and Ensembl VEP impacts, respectively. The third layer shows the MTB impacts. 20 have the same ClinVAP annotation which states therapy insignificance.

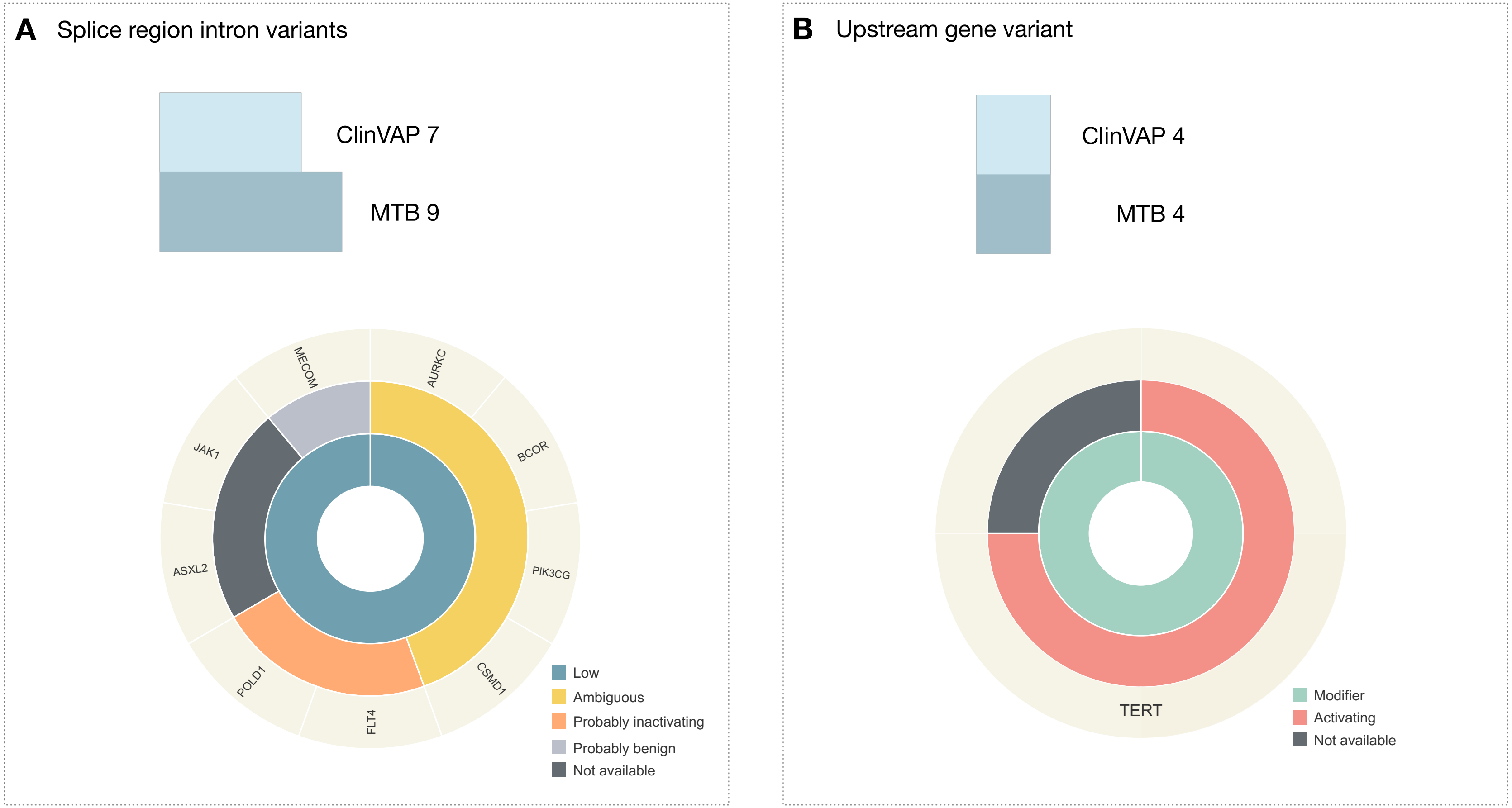


Figure 5.3: Annotation comparison for MTB-specific genes cont. A) Splice region intron. Inner and outer rings are Ensembl VEP and MTB predictions, respectively. Seven genes have the same ClinVAP annotation. They are not therapy informative. B) Upstream gene. The impact is strenuous to predict. MTB listed them as activating mutations. ClinVAP filters modifier impact variants.

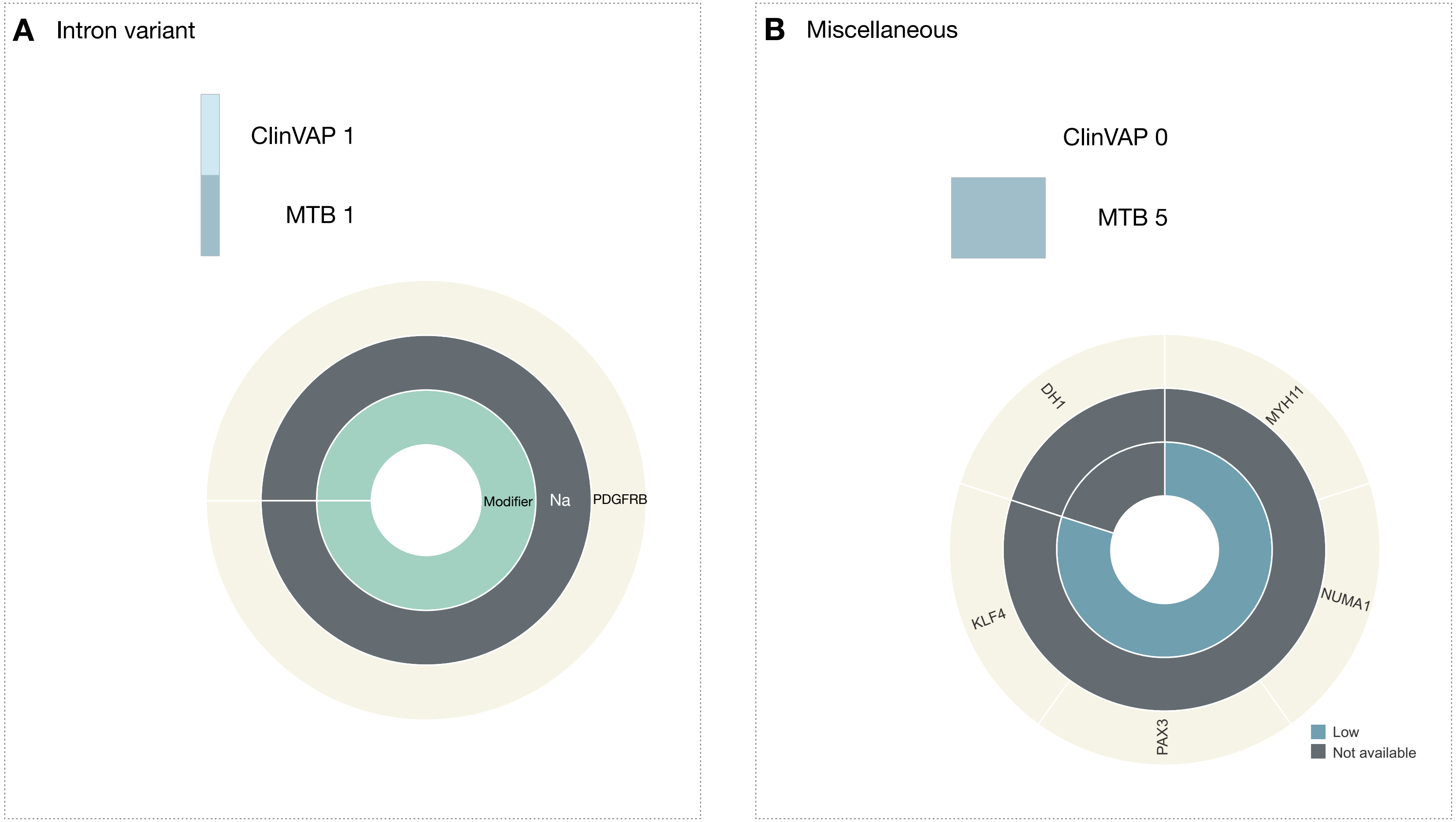


Figure 5.4: Annotation comparison for MTB-specific genes cont. A) Intron. ClinVAP filtered PDGFRB variant due to its modifier impact. MTB did not receive an annotation for it. B) Miscellaneous. It shows MTB-specific genes without an annotation. Four are synonymous mutations in ClinVAP. One gene did not have a ClinVAP annotation due to a typo in MTB content.

Another predominant category was the splice region intron variant of which nine of the MTB-specific variants were labeled as low predicted impact. For seven of them, ClinVAP provided the same annotation (Figure 5.3A). The remaining two were categorized as synonymous splice region variants by ClinVAP suggesting that these were silent mutations observed in the exon region. We used the NCBI genome data viewer to reveal the region that the reported coordinates fall into, which confirmed that these variants were indeed in the exon region and the ClinVAP's annotations were robust. None of those nine splice region intron variants were considered in target identification which supports the reasoning behind their filtration.

Four of the MTB-specific variants were the upstream gene variants which are the changes in the non-coding region or affecting non-coding genes. It has modifier impact which means that there is no evidence of an impact or the impact assessment is tenuous, which causes difficulty in assigning priority to such variants. All four variants observed in this category belong to the TERT gene and ClinVAP filtered them due to their modifier effect (Figure 5.3B). However, the same variants have an activating effect in MTB diagnostic reports which suggests unjustified filtering, since activating mutation in TERT is known for its role in cancer initiation[241].

One MTB-specific variant had intron variant annotation which has a similar consequence to the upstream gene variants and was filtered out by the ClinVAP due to its non-assessable impact (Figure 5.4A). MTB diagnostics reports did not appoint any significance to this variant. Its predicted impact was stated as ambiguous in MTB diagnostic report, and it was not considered a candidate target.

We assigned five genes to the miscellaneous category (Figure 5.4B). Those are the genes whose consequences were not given in MTB diagnostic reports. Four of them were predicted as synonymous annotations by ClinVAP, that increase the redundancy in the MTB content without contributing to the therapeutic decision-making. The remaining one was not in ClinVAP due to a typo made in case presentations. Our conclusion remains the same with the synonymous mutations that it is necessary to filter them to not increase the amount of insignificant information in diagnostic reports.

5.3.4 Coverage on final MTB recommendations

We demonstrated ClinVAP's ability to identify MTB recommended treatment strategies by calculating its coverage on drug-target pairs at the case level. ClinVAP's case reports included an exact coverage for 81.5% and provided another drug from the same class for 7.5% of the 51 final recommendations. We determined that the non-covered recommendations were identified based on expert knowledge of the recruiting clinical trials.

We extracted 41 main and 10 alternative suggestions from MTB protocol excerpts of 31 cases that were subject to analysis. For the main suggestions, the full coverage was 73% and the partial coverage was 15%. Only 5% of the recommendations had zero coverage which corresponds to two cases (Figure 5.5). Both cases had EGFR as the target and the same "Depatuxizumab mafodotin/Temozolomide" combination therapy recommendation which is absent in our annotation KB. Even though it includes depatuximab mafodotin in mechanistic cancer drugs, it was not returned in the results due to its investigational drug tag. ClinVAP results did not list Temozolomide for EGFR either. Besides the zero coverage cases, we identified three cases that only had partially covered suggestions (Figure 5.6). For one case, the main target was FRS2 CNV coupled with Regorafenib or Lenvatinib. ClinVAP listed FGFR inhibitors for FRS2 CNV, but it did not specify a drug name that is classified by partial coverage on drug class. The second case had EGFR CNV suggested as a target for AfatinibTemozolomide combination therapy. ClinVAP included Afatinib for EGFR CNV but not the Temozolomide. The third case had the target IDH1 SNV suggested for BAY1436032, which was a clinical trial. ClinVAP listed PARP inhibitors considered as coverage on drug class. For the alternative suggestions, ClinVAP had an exact coverage of the 90% of the suggestions. 10% of the alternative and 7% of the main suggestions were excluded from the comparison due to not following the eligibility criteria.

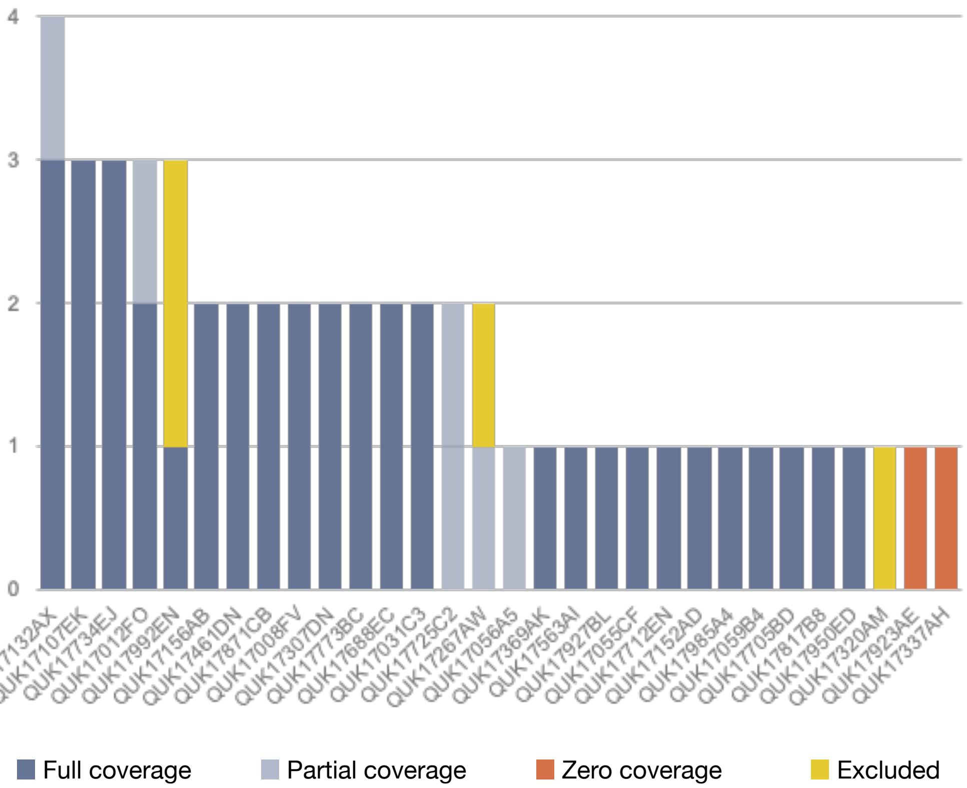


Figure 5.5: MTB therapy suggestion coverage per case. It represents the number of MTB recommendations per case with their coverage status in ClinVAP case reports.

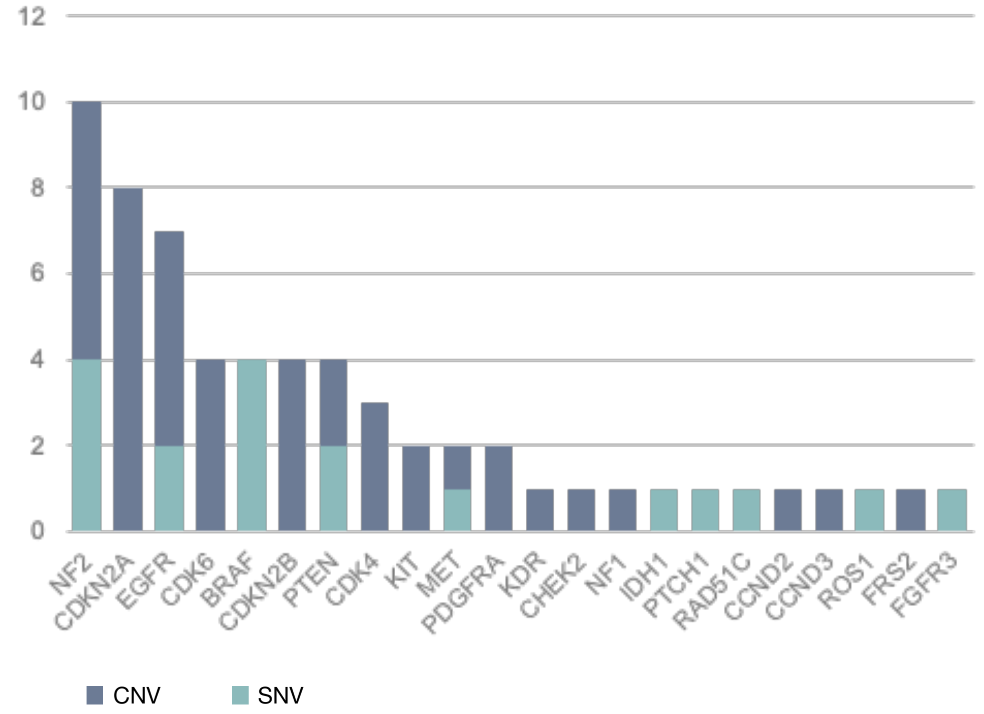


Figure 5.6: MTB suggested target distribution.

We then investigated the likelihood of overestimating the final decision coverage due to the genes that were frequently used as a target by the MTB. As shown in Figure 4B, NF2, CDKN2A, and EGFR were used as a target more frequently than other genes. However, there were also so to say non-mainstream target selections such as FRS2 which is altered in 0.78% of all cancer types. ClinVAP made at least one suggestion for all these targets. Its coverage discrepancy started at the suggested therapeutics for these targets. Among 22 selected targets, 13 of them were observed in less than 6% of the patients in the literature. Those 13 targets were still identified by ClinVAP which is an indication that it does not have a bias in favor of the well-known highly mutated genes.

We then investigated the likelihood of overestimating the final decision coverage that could have been originated from the genes that were frequently used as targets by the MTB. As shown in Figure 3B, NF2, CDKN2A, EGFR were used as a target more frequently than other genes. However, there were also so to say non-mainstream target selections such as FRS2 which is altered in 0.78% of all cancer types. ClinVAP made at least one suggestion for all these targets. Its coverage discrepancy started at the suggested therapeutics for these targets. Among 22 selected targets, 13 of them were observed in less than 6% of the patients in the literature. Those 13 targets were still identified by ClinVAP which is an indication that it does not have bias over the well-known highly mutated genes.

Overall, ClinVAP exhibits complete coverage over the selected targets and high coverage on the suggested drugs. The discrepancies in the drug coverage mostly resulted from the decisions based on expert opinion which cannot and not aimed to be replaced by decision support mechanisms.

5.4 Discussion

The increasing amount of genomics data that needs to be processed to extract its clinical implications necessitates automated, reproducible, and robust data processing pipelines to support MTBs in their decision-making. Besides the software requirements which increase the adaptability of such tools in the clinical setup, it is of crucial importance to demonstrate the content-wise equivalence of the results compared to current clinical practice. Previously, we developed ClinVAP which processes genomics data to create patient-specific case reports representing molecular actionability/druggability. To evaluate its clinical utility, we measured its agreement with the MTB content of retrospective patient data.

We measured the completeness of our annotation KB on the MTB content over the driver and the altered gene lists. The KB exhibited extensive coverage on driver genes and therapeutically relevant variants demonstrating a high focus on actionability. Moreover, we observed that the driver genes were underrepresented in MTB diagnostic reports due to non-standardized reports varying between sequencing labs.

Case-by-case comparisons revealed that ClinVAP has stricter filters on predicted variant impacts to classify them as relevant. Its filtering strategy does not leave out important information, since most of the MTB-specific genes were proven insignificant for therapy decisions. While MTB's content covered complete mutational profiles including low or non-impact variants, ClinVAP prioritized therapeutically informative ones with higher impacts.

Finally, we revealed the high coverage of ClinVAP reports over the MTB suggested therapeutic options. The discrepancies between MTB and ClinVAP contents were predominantly due to the expert knowledge, fundamentally related to open clinical trials, which is not aimed to be replaced by the evidence-based decision management systems. Although not listing the enrolling clinical trials could be perceived as a disadvantage at the first glance, in the regular MTB set-up, this information is heavily dependent on the current knowledge of the attending clinicians since the enrollment criteria depend on many different parameters including the location. Therefore, expert opinion/knowledge will remain an important parameter.

In our analysis, we also demonstrated the disadvantages of non-transparent data processing and annotation procedures in creating MTB diagnostic reports. It was challenging to investigate the discrepancies between MTB and ClinVAP mainly because of the unavailability of the information sources and the non-reproducibility of the filtering constraints from sequencing centers. For instance, an unavailable source of information that was used to label genes as drivers required us to conduct an extensive literature review to distinguish true misclassifications. We had a similar bottleneck when we attempted to investigate the reason for ClinVAP-specific genes being filtered out in the diagnostic reports. Unfortunately, reverse engineering the filtering constraints set by sequencing labs was not possible. These non-transparency problems together with the under-represented information categories point to the need for standardizing the case report structure and transparent and interoperable data management systems.

Another disadvantage of the current MTB system is the compatibility issues of the molecular diagnostic reports with the electronic health records. MTBs receive the reports in PDF format and they use the same format in case discussion content they created. Processed genomics data do not become a part of patient stratification, and no future conclusions are being drawn from it based on patient similarity. Lack of data integration hinders the efficiency of clinical implications of precision oncology as well as oncology research. The labor-intensive manual work we spent on digitizing the MTB case presentations is a strong indicator of the essentiality of the machine-readable molecular diagnostic reports. Next to a human-readable medium, ClinVAP produces patient reports in *JSON* format integrable to data portals which enables clinicians and researchers to re-use the previous data in drawing conclusions from patient similarity and stratifying cohorts.

In conclusion, we assessed the comprehensiveness of our annotation KB and the content-wise sufficiency of ClinVAP reports compared to the current clinical process. We demonstrated the importance of standardized, reproducible, and transparent pipelines to ensure delivering reliable and reproducible translation of molecular aberrations. Even though the expert knowledge will stay an important component, ClinVAP's content is shown as comprehensive enough to be a basis for MTB case discussions which increases the clinical efficiency by eliminating manual case preparation steps. Additionally, it has the potential of benefiting the population-level studies by enhancing genomic data integration for patient stratification and research on patient similarity combined with expected treatment outcomes.

Chapter 6

Conclusion and Outlook

In this thesis, we presented our contributions to the challenges of the clinical adaptation of precision oncology. We developed a pipeline that automates each step of clinical annotation from deciphering the impact of the observed alterations to determining the therapeutic associations of the actionable variants. We assembled a knowledge base as the core source of the clinical implications of molecular aberrations by integrating multiple publicly available databases. We composed an evidence-level index to depict the strength of the clinical associations found, based on the study type of the initial causation (e.g., pre-clinical study, clinical study) and the similarity of the reported consequences to the observed variants. The commonality of the information among the different sources is also depicted as an indication of the association's strength. We aggregated the results in a structured patient report as the final product of our pipeline which can be integrated into the MTB workflow without any modification.

Moreover, we developed a clinomics approach that goes beyond the observed aberrations by extending the case investigation to their neighboring gene interactions. With our network approach, we cluster the main and adjacent genes into their corresponding pathways to signal the user about the potentially affected cascades. We interpolated the direct drug nodes to the neighboring extensions to indicate off-label therapeutic candidates and possible drug combinations to bypass potential therapy resistance. We implemented an interactive GUI to present the results in an organized form that contains brief but comprehensive information. It also allows users to conduct MTB operations for both case preparation and board discussion.

We started the development of the aforementioned decision support tools intending to bridge the bench and the bedside. We demonstrated the benefits of our standardized and automated pipeline approach to clinical routine through a stratified case-cohort study. We compared the case discussion contents and the therapeutic suggestions of the neuro-oncology cases, retrospectively. Our analysis proved that the discrepancy found in the automated report content is mostly introduced by expert opinionsknowledge. We verified the content equivalence of ClinVAP reports indicating that it is suitable to be integrated into the MTB workflow.

At the beginning of the 20th century, Sir Archibald Edward Garrod construed the phenotypical differences of manifesting alkaptonuria between the family members as "chemical individuality"[242]. He reasoned that "...the thought naturally presents itself that these [conditions] are merely extreme examples of variation of chemical behavior which are probably everywhere present in minor degrees and that just as no two individuals of a species are absolutely identical in bodily structure neither are their chemical processes carried out on exactly the same lines" which stands out as an early definition of interpersonal variation in the pre-genomic era[243]. A century later, factoring in the effect of a person's genomic markup on clinical decisions has soared and paved the way for precision oncology. Observations such as the increased survival rates among the patients treated with imatinib in the existence of bcr-abl fusion mutation[244] increased the expectations from precision oncology. However, the premises have not met the expectations. Clinical implementations estimated the number of patients who are eligible for targeted therapy as 8.3% and among those patients, the ratio of responding to the therapy is estimated as 4.9% for 2018 and 7% in 2020[245,246] pointing that a very small part of the cancer population benefit from it.

Increasing the clinical efficiency of precision oncology heavily depends on an all-encompassing translation of genome data into clinical implications varying from actionability/druggability assessment to forming drug repositioning strategies. That translation relies heavily on existing knowledge. As the main pillar is to uncover more molecular mechanisms, genome characterization through cohort studies will continue to be at the center of the cancer research field. There are ongoing efforts to accelerate research and provide clinical benefit from large initiatives such as 1+ Million Genomes (1+MG) and Beyond One Million Genomes (B1MG) through "FAIR"ification of the genome data[247,248]. The largest cancer consortiums ICGCTCGA have entered a new phase that focuses on the efficient use of the information generated over the last decade in the genomic characterization of over 50 cancer types. With the Accelerating Research in Genomic Oncology (ARGO) initiative, the community aims to factor in the effect of regional diversity, uncover the new synergistic therapy strategies and evaluate the commonalities and discrepancies in the therapy outcomes[249]. Most of the hurdles in the application of precision oncology are expected to be overcome by closing the gap in our knowledge of genotype-phenotype links through such initiations that will make the data available to research communities and enables its analysis together with digitized healthcare data. This also requires a change in the phenotype-based clinical trial design for the new therapeutics. It has been shown that the biomarker-based trials have a significantly reduced timeline which creates a profound advantage for decreasing the amount of undruggable genome[250].

Concurrent with the aforementioned strategies, there are affords to make the knowledge available to the community through publicly available databases, mostly implemented by academia for thorough variant classification. One of the major obstacles is that there is no single source that would serve as a "ground truth". Due to the abundance of annotation sources, every MTB relies on a different subset which creates the discrepancy between and within MTBs. The same standardization problem is valid for evidence-based clinical reporting tools. Querying every source through their API services is not favorable due to the re-assembly risk of the sensitive genomics data. Data integration is extremely challenging due to the number of different data models which is equal to the number of sources. Among the sea of those databases, the knowledge contribution to integration efforts ratio is hard to estimate resulting in positive selection towards the well-known sources such as CIViC. To standardize the clinical workflow, it is of crucial importance to have a knowledge base that is a complete representation of the union of the available databases. Unfortunately, having such a source is almost a utopia for the near future.

Apart from the scientific concerns, the availability and the processing of the data require clear legislation. Patients and medical doctors need to be trained in the usage of such data. Patients need to be informed and consented to the scientific use of their data. Society's trust has to be established in the secure storage of the genomics data. From a financial aspect, democratizing genetic testing relies upon the reimbursement of genetic testing by insurance companies. A tangible impact of precision oncology can only be achieved when the regulations are aligned with scientific interests.

Bibliography

[1] Richard A Gibbs. The human genome project changed everything. *Nature Reviews Genetics*, 21(10):575–576, 2020. 1

[2] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015. 1

[3] Stefania Morganti, Paolo Tarantino, Emanuela Ferraro, Paolo D'Amico, Bruno Achutti Duso, and Giuseppe Curigliano. Next generation sequencing (ngs): a revolutionary technology in pharmacogenomics and personalized medicine in cancer. *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics*, pages 9–30, 2019. 1

[4] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019. 1

[5] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[6] Navigating 2020 and beyond. *Nature Genetics*, 52(1):1–1, Jan 2020. 1

[7] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013. 2, 14, 15, 30, 38, 83

[8] Samer Tohme, Richard L Simmons, and Allan Tsung. Surgery for cancer: a trigger for metastases. *Cancer research*, 77(7):1548–1552, 2017. 2, 21

[9] Ryan Morrison, Stephen M Schleicher, Yunguang Sun, Kenneth J Niermann, Sungjune Kim, Daniel E Spratt, Christine H Chung, and Bo Lu. Targeting the mechanisms of resistance to chemotherapy and radiotherapy with the cancer stem cell hypothesis. *Journal of oncology*, 2011, 2011.

[10] Chinna Babu Dracham, Abhash Shankar, and Renu Madan. Radiation induced secondary malignancies: a review article. *Radiation oncology journal*, 36(2):85, 2018. 2, 21

[11] Chandan Kumar-Sinha and Arul M Chinnaiyan. Precision oncology in the age of integrative genomics. *Nature biotechnology*, 36(1):46–60, 2018. 2, 25

[12] Anne V Soerensen, Frede Donskov, Gregers G Hermann, Niels V Jensen, Astrid Petersen, Henrik Spliid, Rickard Sandin, Kirsten Fode, and Poul F Geertsen. Improved overall survival after implementation of targeted therapy for patients with metastatic renal cell carcinoma: results from the danish renal cancer group (darenca) study-2. *European Journal of Cancer*, 50(3):553–562, 2014. 3, 25, 84

[13] Cesare Gridelli, Filippo De Marinis, Federico Cappuzzo, Massimo Di Maio, Fred R Hirsch, Tony Mok, Floriana Morgillo, Rafael Rosell, David R Spigel, James Chih-Hsin Yang, et al. Treatment of advanced non–small-cell lung cancer with epidermal growth factor receptor (egfr) mutation or alk gene rearrangement: results of an international expert panel meeting of the italian association of thoracic oncology. *Clinical lung cancer*, 15(3):173–181, 2014.

[14] Manasi K Mayekar and Trever G Bivona. Current landscape of targeted therapy in lung cancer. *Clinical Pharmacology & Therapeutics*, 102(5):757–764, 2017. 3, 84

[15] Katherine C Kurnit, Ann M Bailey, Jia Zeng, Amber M Johnson, Md Abu Shufean, Lauren Brusco, Beate C Litzenburger, Nora S Sánchez, Yekaterina B Khotskaya, Vijaykumar Holla, et al. "personalized cancer therapy": a publicly available precision oncology resource. *Cancer research*, 77(21):e123–e126, 2017. 3

[16] Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: a systems biology disease. *Biosystems*, 83(2-3):81–90, 2006. 3

[17] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000. 3, 15, 17, 20

[18] Chi V Dang, E Premkumar Reddy, Kevan M Shokat, and Laura Soucek. Drugging the'undruggable'cancer targets. *Nature Reviews Cancer*, 17(8):502, 2017. 3, 62

[19] Junfei Zhao, Feixiong Cheng, Yuanyuan Wang, Carlos L Arteaga, and Zhongming Zhao. Systematic prioritization of druggable mutations in 5000 genomes across 16 cancer types using a structural genomics-based approach. *Molecular & cellular proteomics*, 15(2):642–656, 2016. 3

[20] Sohini Sengupta, Sam Q Sun, Kuan-lin Huang, Clara Oh, Matthew H Bailey, Rajees Varghese, Matthew A Wyczalkowski, Jie Ning, Piyush Tripathi, Joshua F McMichael, et al. Integrative omics analyses broaden treatment targets in human cancer. *Genome medicine*, 10(1):1–20, 2018. 3

[21] Daniel D Von Hoff, Joseph J Stephenson Jr, Peter Rosen, David M Loesch, Mitesh J Borad, Stephen Anthony, Gayle Jameson, Susan Brown, Nina Cantafio, Donald A Richards, et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. *Journal of clinical oncology*, 28(33):4877–4883, 2010. 4

[22] Apostolia-Maria Tsimberidou, Nancy G Iskander, David S Hong, Jennifer J Wheler, Gerald S Falchook, Siqing Fu, Sarina Piha-Paul, Aung Naing, Filip Janku, Rajyalakshmi Luthra, et al. Personalized medicine in a phase i clinical trials program: the md anderson cancer center initiative. *Clinical cancer research*, 18(22):6373–6383, 2012. 25

[23] Siân Jones, Valsamo Anagnostou, Karli Lytle, Sonya Parpart-Li, Monica Nesselbush, David R Riley, Manish Shukla, Bryan Chesnick, Maura Kadan, Eniko Papp, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Science translational medicine*, 7(283):283ra53–283ra53, 2015.

[24] Mark E Burkard, Dustin A Deming, Benjamin M Parsons, Paraic A Kenny, Marissa R Schuh, Ticiana Leal, Nataliya Uboha, Joshua M Lang, Michael A Thompson, Ruth Warren, et al. Implementation and clinical utility of an integrated academic-community regional molecular tumor board. *JCO Precision Oncology*, 1:1–10, 2017. 4, 25

[25] Mario Lamping, Manuela Benary, Serge Leyvraz, Clemens Messerschmidt, Eric Blanc, Thomas Kessler, Moritz Schütte, Dido Lenze, Korinna Jöhrens, Susen Burock, et al. Support of a molecular tumour board by an evidence-based decision management system for precision oncology. *European Journal of Cancer*, 127:41–51, 2020. 4, 25, 84

[26] Damian T Rieke, Mario Lamping, Marissa Schuh, Christophe Le Tourneau, Neus Basté, Mark E Burkard, Klaus H Metzeler, Serge Leyvraz, and Ulrich Keilholz. Comparison of treatment recommendations by molecular tumor boards worldwide. *JCO Precision Oncology*, 2:1–14, 2018. 4, 84

[27] Bart Koopman, Harry JM Groen, Marjolijn JL Ligtenberg, Katrien Grünberg, Kim Monkhorst, Adrianus J de Langen, Mirjam C Boelens, Marthe S Paats, Jan H Von der Thüsen, Winand NM Dinjens, et al. Multicenter comparison of molecular tumor boards in the netherlands: Definition, composition, methods, and targeted therapy recommendations. *The Oncologist*, 26(8):e1347–e1358, 2021. 4

[28] The human genome project completion: Frequently asked questions. `https://www.genome.gov/11006943/`, 2010. Accessed: June 2022. 5

[29] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958. 5

[30] Elizabeth Pennisi. Encode project writes eulogy for junk dna, 2012. 5

[31] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature genetics*, 30(1):13–19, 2002. 6

[32] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11(5):345–355, 2010. 6

[33] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338–345, 2018. 6

[34] Sara Ballouz, Alexander Dobin, and Jesse A Gillis. Is it time to change the reference genome? *Genome biology*, 20(1):1–9, 2019. 6

[35] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006. 6, 7

[36] Evan E Eichler. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1):64–74, 2019. 6, 8

[37] Ryan E Mills, Christopher T Luttig, Christine E Larkins, Adam Beauchamp, Circe Tsui, W Stephen Pittard, and Scott E Devine. An initial map of insertion and deletion (indel) variation in the human genome. *Genome research*, 16(9):1182–1190, 2006. 6

[38] Julienne M Mullaney, Ryan E Mills, W Stephen Pittard, and Scott E Devine. Small insertions and deletions (indels) in human genomes. *Human molecular genetics*, 19(R2):R131–R136, 2010. 6

[39] Fredrik Mertens, Bertil Johansson, Thoas Fioretos, and Felix Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, 15(6):371–381, 2015. 7

[40] Milana Frenkel-Morgenstern, Vincent Lacroix, Iakes Ezkurdia, Yishai Levin, Alexandra Gabashvili, Jaime Prilusky, Angela Del Pozo, Michael Tress, Rory Johnson, Roderic Guigo, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric rna transcripts. *Genome research*, 22(7):1231–1242, 2012. 7

[41] Musaffe Tuna, Christopher I Amos, and Gordon B Mills. Molecular mechanisms and pathobiology of oncogenic fusion transcripts in epithelial tumors. *Oncotarget*, 10(21):2095, 2019. 7

[42] Aaron K Wong, Rachel SG Sealfon, Chandra L Theesfeld, and Olga G Troyanskaya. Decoding disease: From genomes to networks to phenotypes. *Nature Reviews Genetics*, 22(12):774–790, 2021. 7

[43] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018. 7

[44] Zuoheng Wang, Xiangtao Liu, Bao-Zhu Yang, and Joel Gelernter. The role and challenges of exome sequencing in studies of human diseases. *Frontiers in genetics*, 4:160, 2013. 8

[45] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012. 8

[46] David R Adams and Christine M Eng. Next-generation sequencing to diagnose suspected genetic disorders. *New England Journal of Medicine*, 379(14):1353–1362, 2018. 8

[47] Kyle Retterer, Jane Juusola, Megan T Cho, Patrik Vitazka, Francisca Millan, Federica Gibellini, Annette Vertino-Bell, Nizar Smaoui, Julie Neidich, Kristin G Monaghan, et al. Clinical application of whole-exome sequencing across clinical indications. *Genetics in Medicine*, 18(7):696–704, 2016. 8

[48] Eleanor G Seaby and Sarah Ennis. Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Briefings in Functional Genomics*, 19(4):243–258, 2020. 8

[49] All of Us Research Program Investigators. The “all of us” research program. *New England Journal of Medicine*, 381(7):668–676, 2019. 8

[50] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015. 8, 14

[51] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O’Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. 8

[52] Chang Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal*, 16:15–24, 2018. 9, 10

[53] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010. 9

[54] Fass JN Joshi NA. ickle: A sliding-window, adaptive, quality-based trimming tool for fastq files, 2011. Accessed: April 2022. 9

[55] M Martin. Cutadapt removes adapter sequences from highthroughput sequencing reads. embnet j 17: 10–12, 2011. 9

[56] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012. 9

[57] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009. 9

[58] Ayat Hatem, Doruk Bozdağ, Amanda E Toland, and Ümit V Çatalyürek. Benchmarking short sequence mapping tools. *BMC bioinformatics*, 14(1):1–25, 2013. 9

[59] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010. 9

[60] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009. 9

[61] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, 2012. 9

[62] Christopher T Saunders, Wendy SW Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 2012. 9, 11

[63] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011. 10

[64] Yunfei Guo, Xiaolei Ding, Yufeng Shen, Gholson J Lyon, and Kai Wang. Seqmule: automated pipeline for analysis of human exome/genome sequencing data. *Scientific reports*, 5(1):1–10, 2015. 10

[65] Jason L Causey, Cody Ashby, Karl Walker, Zhiping Paul Wang, Mary Yang, Yuanfang Guan, Jason H Moore, and Xiuzhen Huang. Dnap: a pipeline for dna-seq data analysis. *Scientific reports*, 8(1):1–9, 2018.

[66] Maxime Garcia, Szilveszter Juhos, Malin Larsson, Pall I Olason, Marcel Martin, Jesper Eisfeldt, Sebastian DiLorenzo, Johanna Sandgren, Teresita Díaz De Ståhl, Philip Ewels, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research*, 9, 2020.

[67] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W Richard McCombie, and Peter P Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):1–10, 2014. 10

[68] Tyler S Alioto, Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D Eldridge, Eivind Hovig, Lawrence E Heisler, Timothy A Beck, Jared T Simpson, Laurie Tonon, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature communications*, 6(1):1–13, 2015. 10

[69] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012. 12

[70] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010. 12, 30

[71] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):122, 2016. 12, 14, 30, 41

[72] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012. 13

[73] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. Refseq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763, 2014. 13

[74] Adam Frankish, Barbara Uszczynska, Graham RS Ritchie, Jose M Gonzalez, Dmitri Pervouchine, Robert Petryszak, Jonathan M Mudge, Nuno Fonseca, Alvis Brazma, Roderic Guigo, et al. Comparison of gencode and refseq gene annotation and the impact of reference geneset on variant effect prediction. *BMC genomics*, 16(8):1–11, 2015.

[75] Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. Ensembl 2022. *Nucleic acids research*, 50(D1):D988–D995, 2022. 13

[76] Ensembl UCSC. Transcript supporting level (tsl). `http://www.ensembl.org`. Accessed: April 2022. 13

[77] Jose Manuel Rodriguez, Paolo Maietta, Iakes Ezkurdia, Alessandro Pietrelli, Jan-Jaap Wesselink, Gonzalo Lopez, Alfonso Valencia, and Michael L Tress. Appris: annotation of principal and alternative splice isoforms. *Nucleic acids research*, 41(D1):D110–D117, 2013. 13

[78] Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001. 13

[79] Robert Vaser, Swarnaseetha Adusumalli, Sim Ngak Leng, Mile Sikic, and Pauline C Ng. Sift missense predictions for genomes. *Nature protocols*, 11(1):1–9, 2016. 13

[80] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 13

[81] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010. 14

[82] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020. 14

[83] Daniel R Zerbino, Steven P Wilder, Nathan Johnson, Thomas Juettemann, and Paul R Flicek. The ensembl regulatory build. *Genome biology*, 16(1):1–8, 2015. 14

[84] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012. 14

[85] David Adams, Lucia Altucci, Stylianos E Antonarakis, Juan Ballesteros, Stephan Beck, Adrian Bird, Christoph Bock, Bernhard Boehm, Elias Campo, Andrea Caricasole, et al. Blueprint to decode the epigenetic signature written in blood. *Nature biotechnology*, 30(3):224–226, 2012.

[86] Casey E Romanoski, Christopher K Glass, Hendrik G Stunnenberg, Laurence Wilson, and Genevieve Almouzni. Roadmap for regulation. *Nature*, 518(7539):314–316, 2015. 14

[87] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014. 14

[88] Graham RS Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature methods*, 11(3):294–296, 2014. 14

[89] MD) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore. Online mendelian inheritance in man, omim®. `https://omim.org/`. Accessed: May 2022. 14

[90] INSERM 1997. Orphanet: an online database of rare diseases and orphan drugs. `http://www.orpha.net`. Accessed: May 2022.

[91] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2014.

[92] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014. 14

[93] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001. 14

[94] Simon A Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, et al. Cosmic: somatic cancer genetics at high-resolution. *Nucleic acids research*, 45(D1):D777–D783, 2016. 30, 38, 57

[95] Peter D Stenson, Edward V Ball, Matthew Mort, Andrew D Phillips, Katy Shaw, and David N Cooper. The human gene mutation database (hgmd) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics*, 39(1):1–13, 2012. 14

[96] Nhlbi exome sequencing. `http://evs.gs.washington.edu/EVS/`. Accessed: May 2022. 14

[97] European network staff exchange for integrating precision health in the health care systems, the exact project. `http://www.exactproject.net`. Accessed: May 2022. 14

[98] Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, 2015. 15

[99] Iñigo Martincorena, Keiran M Raine, Moritz Gerstung, Kevin J Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041, 2017. 15

[100] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011. 15, 16, 19

[101] N Yang, SD Ray, and K Krafts. Cell proliferation. In *Encyclopedia of Toxicology: Third Edition*, pages 761–765. Elsevier, 2014. 15

[102] Michael B Sporn and Anita B Roberts. Autocrine growth factors and cancer. *Nature*, 313(6005):745–747, 1985. 15

[103] Richard Sever and Joan S Brugge. Signal transduction in cancer. *Cold Spring Harbor perspectives in medicine*, 5(4):a006098, 2015. 15

[104] AS Lundberg and RA Weinberg. Control of the cell cycle and apoptosis1. *European journal of cancer*, 35(14):1886–1894, 1999. 16

[105] C Giacinti and Antonion Giordano. Rb and cell cycle progression. *Oncogene*, 25(38):5220–5227, 2006. 16

[106] Brandon J Aubrey, Andreas Strasser, and Gemma L Kelly. Tumor-suppressor functions of the tp53 pathway. *Cold Spring Harbor perspectives in medicine*, 6(5):a026062, 2016. 16, 17

[107] Kevin P Foley and Robert N Eisenman. Two mad tails: what the recent knockouts of mad1 and mxi1 tell us about the myc/max/mad network. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1423(3):M37–M47, 1999. 17

[108] Rebecca SY Wong. Apoptosis in cancer: from pathogenesis to treatment. *Journal of experimental & clinical cancer research*, 30(1):1–14, 2011. 17

[109] John C Reed. Bcl-2 family proteins: regulators of apoptosis and chemoresistance in hematologic malignancies. In *Seminars in hematology*, volume 34, pages 9–19, 1997. 17

[110] Eva Szegezdi, Susan E Logue, Adrienne M Gorman, and Afshin Samali. Mediators of endoplasmic reticulum stress-induced apoptosis. *EMBO reports*, 7(9):880–885, 2006. 17

[111] Susan L Fink and Brad T Cookson. Apoptosis, pyroptosis, and necrosis: mechanistic description of dead and dying eukaryotic cells. *Infection and immunity*, 73(4):1907–1916, 2005. 17

[112] Leonard Hayflick. Mortality and immortality at the cellular level. a review. *Biochemistry-New York-English Translation of Biokhimiya*, 62(11):1180–1190, 1997. 17

[113] Carol W Greider. Telomere length regulation. *Annual review of biochemistry*, 65(1):337–365, 1996. 18

[114] Michael Z Levy, Richard C Allsopp, A Bruce Futcher, Carol W Greider, and Calvin B Harley. Telomere end-replication problem and cell aging. *Journal of molecular biology*, 225(4):951–960, 1992. 18

[115] Maria A Blasco. Telomeres and human disease: ageing, cancer and beyond. *Nature Reviews Genetics*, 6(8):611–622, 2005. 18

[116] Shiaw-Yih Lin and Stephen J Elledge. Multiple tumor suppressor pathways negatively regulate telomerase. *Cell*, 113(7):881–889, 2003. 18

[117] Douglas Hanahan and Judah Folkman. Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *cell*, 86(3):353–364, 1996. 18

[118] Vanessa Baeriswyl and Gerhard Christofori. The angiogenic switch in carcinogenesis. In *Seminars in cancer biology*, volume 19, pages 329–337. Elsevier, 2009. 19

[119] Masahiro Murakami and Michael Simons. Fibroblast growth factor regulation of neovascularization. *Current opinion in hematology*, 15(3):215, 2008. 19

[120] Marco Presta, Patrizia Dell'Era, Stefania Mitola, Emanuela Moroni, Roberto Ronca, and Marco Rusnati. Fibroblast growth factor/fibroblast growth factor receptor system in angiogenesis. *Cytokine & growth factor reviews*, 16(2):159–178, 2005. 19

[121] Peter Carmeliet. Vegf as a key mediator of angiogenesis in cancer. *Oncology*, 69(Suppl. 3):4–10, 2005. 19

[122] Christine L Chaffer and Robert A Weinberg. A perspective on cancer cell metastasis. *science*, 331(6024):1559–1564, 2011. 19

[123] Ilona Kaszak, Olga Witkowska-Piłaszewicz, Zuzanna Niewiadomska, Bożena Dworecka-Kaszak, Felix Ngosa Toka, and Piotr Jurka. Role of cadherins in cancer—a review. *International Journal of Molecular Sciences*, 21(20):7624, 2020. 20

[124] Hellyeh Hamidi and Johanna Ivaska. Every step of the way: integrins in cancer progression and metastasis. *Nature Reviews Cancer*, 18(9):533–548, 2018. 20

[125] Jianming Zhang, Priscilla L Yang, and Nathanael S Gray. Targeting cancer with small molecule kinase inhibitors. *Nature reviews cancer*, 9(1):28–39, 2009. 21

[126] Gregory P Adams and Louis M Weiner. Monoclonal antibody therapy of cancer. *Nature biotechnology*, 23(9):1147–1157, 2005. 21

[127] Tony Hunter and Jonathan A Cooper. Protein-tyrosine kinases. *Annual review of biochemistry*, 54(1):897–930, 1985. 21

[128] Jonas Cicenas, Egle Zalyte, Amos Bairoch, and Pascale Gaudet. Kinases and cancer. *Cancers*, 10(3), 2018. 21

[129] Philip Cohen. Protein kinases—the major drug targets of the twenty-first century? *Nature reviews Drug discovery*, 1(4):309–315, 2002. 21

[130] Charles Sawyers. Targeted cancer therapy. *Nature*, 432(7015):294–297, 2004. 21

[131] Troy A Baudino. Targeted cancer therapy: the next generation of cancer treatment. *Current drug discovery technologies*, 12(1):3–20, 2015. 21

[132] Charles L Sawyers. Will mtor inhibitors make it as cancer drugs? *Cancer cell*, 4(5):343–348, 2003. 21

[133] Herbert Hurwitz, Louis Fehrenbacher, William Novotny, Thomas Cartwright, John Hainsworth, William Heim, Jordan Berlin, Ari Baron, Susan Griffing, Eric Holmgren, et al. Bevacizumab plus

irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *New England journal of medicine*, 350(23):2335–2342, 2004. 21

[134] Chen Hu and James J Dignam. Biomarker-driven oncology clinical trials: Key design elements, types, features, and practical considerations. *JCO Precision Oncology*, 1:1–12, 2019. 22

[135] Kivilcim Ozturk, Michelle Dow, Daniel E Carlin, Rafael Bejar, and Hannah Carter. The emerging potential for network analysis to inform precision cancer medicine. *Journal of molecular biology*, 430(18):2875–2899, 2018. 22, 23

[136] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr. *Nature*, 483(7387):100–103, 2012. 22

[137] Theodoros I Roumeliotis, Steven P Williams, Emanuel Gonçalves, Clara Alsinet, Martin Del Castillo Velasco-Herrera, Nanne Aben, Fatemeh Zamanzad Ghavidel, Magali Michaut, Michael Schubert, Stacey Price, et al. Genomic determinants of protein abundance variation in colorectal cancer cells. *Cell reports*, 20(9):2201–2214, 2017. 22

[138] Erik S Knudsen and Jean YJ Wang. Targeting the rb-pathway in cancer therapy. *Clinical Cancer Research*, 16(4):1094–1099, 2010. 22

[139] Laura Spring, Aditya Bardia, and Shanu Modi. Targeting the cyclin d–cyclin-dependent kinase (cdk) 4/6–retinoblastoma pathway with selective cdk 4/6 inhibitors in hormone receptor-positive breast cancer: rationale, current status, and future directions. *Discovery medicine*, 21(113):65, 2016. 22

[140] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. In *Biocomputing 2012*, pages 55–66. World Scientific, 2012. 23

[141] Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551–562, 2017. 23

[142] Joel T Dudley and Konrad J Karczewski. *Exploring personal genomics*. Oxford University Press, 2013. 23

[143] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165, 2002. 23

[144] Malachi Griffith, Nicholas C Spies, Kilannin Krysiak, Joshua F McMichael, Adam C Coffman, Arpad M Danos, Benjamin J Ainscough, Cody A Ramirez, Damian T Rieke, Lynzey Kujan, et al. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics*, 49(2):170, 2017. 23, 38, 57

[145] David Tamborero, Carlota Rubio-Perez, Jordi Deu-Pons, Michael P Schroeder, Ana Vivancos, Ana Rovira, Ignasi Tusquets, Joan Albanell, Jordi Rodon, Josep Tabernero, et al. Cancer genome

interpreter annotates the biological and clinical relevance of tumor alterations. *Genome medicine*, 10(1):1–8, 2018. 23, 30, 31, 57, 84, 90

[146] Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E Rudolph, Rona Yaeger, Tara Soumerai, Moriah H Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO precision oncology*, 1:1–16, 2017. 23

[147] My cancer genome: Genetically informed cancer medicine. `https://www.mycancergenome.org`. Accessed: June 2022. 23

[148] Sandra Misale, Rona Yaeger, Sebastijan Hobor, Elisa Scala, Manickam Janakiraman, David Liska, Emanuele Valtorta, Roberta Schiavo, Michela Buscarino, Giulia Siravegna, et al. Emergence of kras mutations and acquired resistance to anti-egfr therapy in colorectal cancer. *Nature*, 486(7404):532–536, 2012. 24

[149] David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl_1):D901–D906, 2008. 24, 44, 77

[150] Funda Meric-Bernstam, Amber Johnson, Vijaykumar Holla, Ann Marie Bailey, Lauren Brusco, Ken Chen, Mark Routbort, Keyur P Patel, Jia Zeng, Scott Kopetz, et al. A decision support framework for genomically informed investigational cancer therapy. *JNCI: Journal of the National Cancer Institute*, 107(7), 2015. 25

[151] Maria Schwaederle, Barbara A Parker, Richard B Schwab, Paul T Fanta, Sarah G Boles, Gregory A Daniels, Lyudmila A Bazhenova, Rupa Subramanian, Alice C Coutinho, Haydee Ojeda-Fournier, et al. Molecular tumor board: the university of california san diego moores cancer center experience. *The oncologist*, 19(6):631–636, 2014. 25

[152] Maria Schwaederle, Barbara A Parker, Richard B Schwab, Gregory A Daniels, David E Piccioni, Santosh Kesari, Teresa L Helsten, Lyudmila A Bazhenova, Julio Romero, Paul T Fanta, et al. Precision oncology: The uc san diego moores cancer center predict experienceprecision oncology: Moores cancer center experience. *Molecular cancer therapeutics*, 15(4):743–752, 2016. 25

[153] Laura J Tafe, Ivan P Gorlov, Francine B De Abreu, Joel A Lefferts, Xiaoying Liu, Jason R Pettus, Jonathan D Marotti, Kasia J Bloch, Vincent A Memoli, Arief A Suriawinata, et al. Implementation of a molecular tumor board: the impact on treatment decisions for 35 patients evaluated at dartmouth-hitchcock medical center. *The oncologist*, 20(9):1011–1018, 2015. 25

[154] Benjamin Besse, Ludovic Lacroix, Laura Faivre, Virginie Kahn-Charpy, Maud Ngo Camus, Nathalie Auger, Valerie Koubi-Pick, Julien Adam, Vincent Thomas De Montpreville, Peter Dorfmuller, et al. Molecular multidisciplinary tumor board (mmtb) for lung cancer patients: 2-year experience report. In *JOURNAL OF THORACIC ONCOLOGY*, volume 8, pages S245–S246. LIPPINCOTT WILLIAMS & WILKINS 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA, 2013.

[155] Milan Radovich, Patrick J Kiel, Stacy M Nance, Erin E Niland, Megan E Parsley, Meagan E Ferguson, Guanglong Jiang, Natraj R Ammakkanavar, Lawrence H Einhorn, Liang Cheng, et al. Clinical

benefit of a precision medicine based approach for guiding treatment of refractory cancers. *Oncotarget*, 7(35):56491, 2016.

[156] Kim M Hirshfield, Denis Tolkunov, Hua Zhong, Siraj M Ali, Mark N Stein, Susan Murphy, Hetal Vig, Alexei Vazquez, John Glod, Rebecca A Moss, et al. Clinical actionability of comprehensive genomic profiling for management of rare or refractory cancers. *The oncologist*, 21(11):1315–1325, 2016. 25

[157] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016. 26

[158] Heidi L Rehm, Jonathan S Berg, Lisa D Brooks, Carlos D Bustamante, James P Evans, Melissa J Landrum, David H Ledbetter, Donna R Maglott, Christa Lese Martin, Robert L Nussbaum, et al. Clingen—the clinical genome resource. *New England Journal of Medicine*, 372(23):2235–2242, 2015.

[159] Sharon F Terry. The global alliance for genomics & health. *Genet Test Mol Biomarkers*, 18(6):375–6, 2014. 26

[160] Deborah I Ritter, Sameek Roychowdhury, Angshumoy Roy, Shruti Rao, Melissa J Landrum, Dmitriy Sonkin, Mamatha Shekar, Caleb F Davis, Reece K Hart, Christine Micheel, et al. Somatic cancer variant curation and harmonization through consensus minimum variant level data. *Genome medicine*, 8(1):117, 2016. 26, 41

[161] Bilge Sürün, Charlotta PI Schärfe, Mathew R Divine, Julian Heinrich, Nora C Toussaint, Lukas Zimmermann, Janina Beha, and Oliver Kohlbacher. Clinvap: a reporting strategy from variants to therapeutic options. *Bioinformatics*, 36(7):2316–2317, 2020. 29, 65, 85

[162] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009. 30, 90

[163] Min Zhao, Jingchun Sun, and Zhongming Zhao. Tsgene: a web resource for tumor suppressor genes. *Nucleic acids research*, 41(D1):D970–D976, 2012. 30, 38, 57

[164] UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2016. 35

[165] Carlota Rubio-Perez, David Tamborero, Michael P Schroeder, Albert A Antolín, Jordi Deu-Pons, Christian Perez-Llamas, Jordi Mestres, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer cell*, 27(3):382–396, 2015. 38

[166] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3(1):1–10, 2013.

[167] Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016. 30

[168] Michelle Whirl-Carrillo, Ellen M McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, Russ B Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012. 30

[169] Júlia Perera-Bel, Barbara Hutter, Christoph Heining, Annalen Bleckmann, Martina Fröhlich, Stefan Fröhling, Hanno Glimm, Benedikt Brors, and Tim Beißbarth. From somatic variants towards precision oncology: evidence-driven reporting of treatment options in molecular tumor boards. *Genome medicine*, 10(1):1–15, 2018. 30, 55, 63, 84

[170] Quan Xu, Jin-Cheng Zhai, Cai-Qin Huo, Yang Li, Xue-Jiao Dong, Dong-Fang Li, Ru-Dan Huang, Chuang Shen, Yu-Jun Chang, Xi-Ling Zeng, et al. Oncopdss: an evidence-based clinical decision support system for oncology pharmacotherapy at the individual level. *BMC cancer*, 20(1):1–10, 2020. 31, 63, 84

[171] Franziska Singer, Anja Irmisch, Nora C Toussaint, Linda Grob, Jochen Singer, Thomas Thurnherr, Niko Beerenwinkel, Mitchell P Levesque, Reinhard Dummer, Luca Quagliata, et al. Swissmtb: establishing comprehensive molecular cancer diagnostics in swiss clinics. *BMC medical informatics and decision making*, 18(1):1–18, 2018. 31, 84

[172] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319, 2017. 33

[173] Kristian A Gray, Bethan Yates, Ruth L Seal, Mathew W Wright, and Elspeth A Bruford. Genenames. org: the hgnc resources in 2015. *Nucleic acids research*, 43(D1):D1079–D1085, 2014. 35

[174] Monique Zahn-Zabal, Pierre-André Michel, Alain Gateau, Frédéric Nikitin, Mathieu Schaeffer, Estelle Audot, Pascale Gaudet, Paula D Duek, Daniel Teixeira, Valentine Rech de Laval, et al. The nextprot knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research*, 48(D1):D328–D334, 2020. 35

[175] Ncbi genome remapping service. `https://www.ncbi.nlm.nih.gov/genome/tools/remap`. Accessed: August 2022. 39

[176] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):1–12, 2005. 39

[177] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006. 40, 57, 71

[178] Ying Hong Li, Chun Yan Yu, Xiao Xu Li, Peng Zhang, Jing Tang, Qingxia Yang, Tingting Fu, Xiaoyu Zhang, Xuejiao Cui, Gao Tu, et al. Therapeutic target database update 2018: enriched

resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic acids research*, 46(D1):D1121–D1127, 2017. 44

[179] Simon D Harding, Joanna L Sharman, Elena Faccenda, Chris Southan, Adam J Pawson, Sam Ireland, Alasdair JG Gray, Liam Bruce, Stephen PH Alexander, Stephen Anderton, et al. The iuphar/bps guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology. *Nucleic acids research*, 46(D1):D1091–D1106, 2017. 44, 57

[180] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19, 2017. 40, 44

[181] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009. 41

[182] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010. 41

[183] Strelka user guide. `https://github.com/Illumina/strelka/blob/v2.9.x/docs/userGuide/README.md#variant-prediction`. Accessed: May 2023. 41

[184] My cancer genome: Genetically informed cancer medicine. `https://docxtpl.readthedocs.io/en/latest/`. Accessed: August 2022. 43, 47

[185] Bárbara Meléndez, Claude Van Campenhout, Sandrine Rorive, Myriam Remmelink, Isabelle Salmon, and Nicky D'Haene. Methods of measurement for tumor mutational burden in tumor tissue. *Translational lung cancer research*, 7(6):661, 2018. 47

[186] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459, 2017. 54

[187] Dirk Merkel et al. Docker: lightweight linux containers for consistent development and deployment. *Linux j*, 239(2):2, 2014. 54

[188] International Cancer Genome Consortium et al. International network of cancer genome projects. *Nature*, 464(7291):993, 2010. 54, 57

[189] Metrics, cpu usage. `https://www.nextflow.io/docs/latest/metrics.html#cpu-usage`. Accessed: May 2023. 54

[190] Metrics, memory. `https://www.nextflow.io/docs/latest/metrics.html#cpu-usage`. Accessed: May 2023. 54

[191] Metrics, job duration. `https://www.nextflow.io/docs/latest/metrics.html#cpu-usage`. Accessed: May 2023. 54

[192] Kim DeLeonardis, Lauren Hogan, Stephen A Cannistra, Deepa Rangachari, and Nadine Tung. When should tumor genomic profiling prompt consideration of germline testing? *Journal of Oncology Practice*, 15(9):465–473, 2019. 57

[193] D Mandelker, M Donoghue, S Talukdar, C Bandlamudi, P Srinivasan, M Vivek, S Jezdic, H Hanson, K Snape, A Kulkarni, et al. Germline-focussed analysis of tumour-only sequencing: recommendations from the esmo precision medicine working group. *Annals of Oncology*, 30(8):1221–1231, 2019. 57

[194] Hongbin Dong, Daniel W Nebert, Elspeth A Bruford, David C Thompson, Hans Joenje, and Vasilis Vasiliou. Update of the human and mouse fanconi anemia genes. *Human genomics*, 9(1):1–10, 2015. 57

[195] Joshi Niraj, Anniina Färkkilä, and Alan D D'Andrea. The fanconi anemia pathway in cancer. *Annual review of cancer biology*, 3:457–478, 2019. 57, 92

[196] Mirjam Figaschewski, Bilge Sürün, Thorsten Tiede, and Oliver Kohlbacher. The personalized cancer network explorer (pecax) as a visual analytics tool to support molecular tumor boards. *BMC bioinformatics*, 24(1):1–11, 2023. 61

[197] David M Hyman, Barry S Taylor, and José Baselga. Implementing genome-driven oncology. *Cell*, 168(4):584–599, 2017. 61

[198] Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013. 62

[199] David B Solit and Neal Rosen. Resistance to braf inhibition in melanomas. *New England Journal of Medicine*, 364(8):772–774, 2011. 62

[200] Ruth Nussinov, Chung-Jung Tsai, and Hyunbum Jang. A new view of pathway-driven drug resistance in tumor proliferation. *Trends in pharmacological sciences*, 38(5):427–437, 2017. 62

[201] Michael F Berger and Elaine R Mardis. The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, 15(6):353–365, 2018. 62

[202] Ian A Cree and Peter Charlton. Molecular chess? hallmarks of anti-cancer drug resistance. *BMC cancer*, 17(1):1–8, 2017. 62

[203] S N Hart, P Duffy, D J Quest, A Hossain, M A Meiners, and JP Kocher. Vcf-miner: Gui-based application for mining variants and annotations stored in vcf files. *Briefings in bioinformatics*, 17(2):346–351, 2016. 63

[204] M Akgün and H Demirci. Vcf-explorer: filtering and analysing whole genome vcf files. *Bioinformatics*, 33(21):3468–3470, 2017. 63

[205] Thorsten Tiede. Sbml4j, 2019. 65, 70

[206] Mirjam Figaschewski. Biograohvisart, 2019. 65, 72, 82

[207] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000. 71, 75

[208] UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015. 75

[209] Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002. 75

[210] Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames. org: the hgnc and vgnc resources in 2021. *Nucleic acids research*, 49(D1):D939–D946, 2021. 75

[211] Ugur Dogrusoz, Erhan Giral, Ahmet Cetintas, Ali Civril, and Emek Demir. A layout algorithm for undirected compound graphs. *Information Sciences*, 179(7):980–994, 2009. 77

[212] Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, Michael Lush, and Hester Wain. The hugo gene nomenclature committee (hgnc). *Human genetics*, 109(6):678–680, 2001. 77

[213] Andrei Kouranov, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E Bourne, and Helen M Berman. The rcsb pdb information portal for structural genomics. *Nucleic acids research*, 34(suppl_1):D302–D305, 2006. 77

[214] R Danesi, S Fogli, S Indraccolo, M Del Re, AP Dei Tos, L Leoncini, L Antonuzzo, L Bonanno, V Guarneri, A Pierini, et al. Druggable targets meet oncogenic drivers: opportunities and limitations of target-based classification of tumors and the role of molecular tumor boards. *ESMO open*, 6(2):100040, 2021. 83

[215] Vinay Prasad, Tito Fojo, and Michael Brada. Precision oncology: origins, optimism, and potential. *The Lancet Oncology*, 17(2):e81–e86, 2016. 84

[216] Mariangela Russo, Sandra Misale, Ge Wei, Giulia Siravegna, Giovanni Crisafulli, Luca Lazzari, Giorgio Corti, Giuseppe Rospo, Luca Novara, Benedetta Mussolin, et al. Acquired resistance to the trk inhibitor entrectinib in colorectal cancer. *Cancer discovery*, 6(1):36–44, 2016. 84

[217] David Tamborero, Rodrigo Dienstmann, Maan Haj Rachid, Jorrit Boekel, Richard Baird, Irene Braña, Luigi De Petris, Jeffrey Yachnin, Christophe Massard, Frans L Opdam, et al. Support systems to guide clinical decision-making in precision oncology: The cancer core europe molecular tumor board portal. *Nature Medicine*, 26(7):992–994, 2020. 84

[218] Todd C Knepper, Gillian C Bell, J Kevin Hicks, Eric Padron, Jamie K Teer, Teresa T Vo, Nancy K Gillis, Neil T Mason, Howard L McLeod, and Christine M Walko. Key lessons learned from moffitt's molecular tumor board: The clinical genomics action committee experience. *The oncologist*, 22(2):144, 2017. 84

[219] Claudio Luchini, Rita T Lawlor, Michele Milella, and Aldo Scarpa. Molecular tumor boards in clinical practice. *Trends in Cancer*, 6(9):738–744, 2020. 84

[220] Shumei Kato, Ki Hwan Kim, Hyo Jeong Lim, Amelie Boichard, Mina Nikanjam, Elizabeth Weihe, Dennis J Kuo, Ramez N Eskander, Aaron Goodman, Natalie Galanina, et al. Real-world data from

a molecular tumor board demonstrates improved outcomes with a precision n-of-one strategy. *Nature communications*, 11(1):1–9, 2020. 84

[221] Hiroshi Kobayashi, Sumire Ohno, Yoshikazu Sasaki, and Miyuki Matsuura. Hereditary breast and ovarian cancer susceptibility genes. *Oncology reports*, 30(3):1019–1029, 2013. 92

[222] Bertrand C Liang, Donald A Ross, and Eddie Reed. Genomic copy number changes of dna repair genes ercc1 and ercc2 in human gliomas. *Journal of neuro-oncology*, 26(1):17–23, 1995.

[223] C Raimondi and Marco Falasca. Targeting pdk1 in cancer. *Current medicinal chemistry*, 18(18):2763–2769, 2011.

[224] Benoit Busser, Lucie Sancey, Elisabeth Brambilla, Jean-Luc Coll, and Amandine Hurbin. The multiple roles of amphiregulin in human cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1816(2):119–131, 2011.

[225] Shuli Liu, Yang Wang, Yong Han, Weiya Xia, Ling Zhang, Shengming Xu, Houyu Ju, Xiangkai Zhang, Guoxin Ren, Liu Liu, et al. Ereg-driven oncogenesis of head and neck squamous cell carcinoma exhibits higher sensitivity to erlotinib therapy. *Theranostics*, 10(23):10589, 2020.

[226] David J Riese II and Richard L Cullum. Epiregulin: roles in normal physiology and cancer. In *Seminars in cell & developmental biology*, volume 28, pages 49–56. Elsevier, 2014.

[227] Noriaki Sunaga and Kyoichi Kaira. Epiregulin as a therapeutic target in non-small-cell lung cancer. *Lung Cancer*, 6:91, 2015.

[228] Boriana M Zaharieva, Ronald Simon, Pierre-Andre Diener, Daniel Ackermann, Robert Maurer, Göran Alund, Hartmut Knönagel, Marcus Rist, Kim Wilber, Franz Hering, et al. High-throughput tissue microarray analysis of 11q13 gene amplification (ccnd1, fgf3, fgf4, ems1) in urinary bladder cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 201(4):603–608, 2003.

[229] V Fantl, R Smith, S Brookes, C Dickson, and G Peters. Chromosome 11q13 abnormalities in human breast cancer. *Cancer surveys*, 18:77–94, 1993.

[230] Qinfei Zhao, Huaying Li, Longyu Zhu, Suping Hu, Xuxiang Xi, Yanmei Liu, Jianfeng Liu, and Tianyu Zhong. Bioinformatics analysis shows that top2a functions as a key candidate gene in the progression of cervical cancer. *Biomedical Reports*, 13(4):1–1, 2020. 92

[231] Leo Y Luo, Eejung Kim, Hiu Wing Cheung, Barbara A Weir, Gavin P Dunn, Rhine R Shen, and William C Hahn. The tyrosine kinase adaptor protein frs2 is oncogenic and amplified in high-grade serous ovarian cancer. *Molecular Cancer Research*, 13(3):502–509, 2015. 92

[232] Leo Y Luo and William C Hahn. Oncogenic signaling adaptor proteins. *Journal of Genetics and Genomics*, 42(10):521–529, 2015.

[233] Si Wang, Si-Yao Wang, Feng Du, Qiang Han, En-Hua Wang, En-Jie Luo, and Yang Liu. Knockdown of pak1 inhibits the proliferation and invasion of non–small cell lung cancer cells through the erk pathway. *Applied Immunohistochemistry & Molecular Morphology*, 28(8):602–610, 2020.

[234] Christy C Ong, Adrian M Jubb, Peter M Haverty, Wei Zhou, Victoria Tran, Tom Truong, Helen Turley, Tom O'Brien, Domagoj Vucic, Adrian L Harris, et al. Targeting p21-activated kinase 1 (pak1) to induce apoptosis of tumor cells. *Proceedings of the National Academy of Sciences*, 108(17):7177–7182, 2011. 92

[235] C Powell, C Mikropoulos, SB Kaye, CM Nutting, SA Bhide, K Newbold, and KJ Harrington. Pre-clinical and clinical evaluation of parp inhibitors as tumour-specific radiosensitisers. *Cancer treatment reviews*, 36(7):566–575, 2010. 94

[236] Petar Jelinic, Laura A Eccles, Jill Tseng, Paulina Cybulska, Monicka Wielgos, Simon N Powell, and Douglas A Levine. The emsy threonine 207 phospho-site is required for emsy-driven suppression of dna damage repair. *Oncotarget*, 8(8):13792, 2017. 94

[237] Olga Kondrashova and Clare L Scott. Clarifying the role of emsy in dna repair in ovarian cancer. *Cancer*, 125(16):2720–2724, 2019. 94

[238] Suresh S Ramalingam, Normand Blais, Julien Mazieres, Martin Reck, C Michael Jones, Erzsebet Juhasz, Laszlo Urban, Sergey Orlov, Fabrice Barlesi, Ebenezer Kio, et al. Randomized, placebo-controlled, phase ii study of veliparib in combination with carboplatin and paclitaxel for advanced/metastatic non–small cell lung cancer. *Clinical Cancer Research*, 23(8):1937–1944, 2017. 94

[239] Minesh P Mehta, Ding Wang, Fen Wang, Lawrence Kleinberg, Anthony Brade, H Ian Robins, Aruna Turaka, Terri Leahy, Diane Medina, Hao Xiong, et al. Veliparib in combination with whole brain radiation therapy in patients with brain metastases: results of a phase 1 study. *Journal of neuro-oncology*, 122(2):409–417, 2015. 94

[240] Joaquin Mateo, Suzanne Carreira, Shahneen Sandhu, Susana Miranda, Helen Mossop, Raquel Perez-Lopez, Daniel Nava Rodrigues, Dan Robinson, Aurelius Omlin, Nina Tunariu, et al. Dna-repair defects and olaparib in metastatic prostate cancer. *New England Journal of Medicine*, 373(18):1697–1708, 2015. 94

[241] Xiaotian Yuan, Catharina Larsson, and Dawei Xu. Mechanisms underlying the activation of tert transcription and telomerase activity in human cancer: old actors and new players. *Oncogene*, 38(34):6172–6183, 2019. 99

[242] ArchibaldE Garrod. The incidence of alkaptonuria: a study in chemical individuality. *The Lancet*, 160(4137):1616–1620, 1902. 108

[243] Laura H Goetz and Nicholas J Schork. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963, 2018. 108

[244] Michael E O'Dwyer and Brian J Druker. Status of bcr-abl tyrosine kinase inhibitors in chronic myelogenous leukemia. *Current opinion in oncology*, 12(6):594–597, 2000. 108

[245] John Marquart, Emerson Y Chen, and Vinay Prasad. Estimation of the percentage of us patients with cancer who benefit from genome-driven oncology. *JAMA oncology*, 4(8):1093–1098, 2018. 108

[246] A Haslam, MS Kim, and V Prasad. Updated estimates of eligibility for and response to genome-targeted oncology drugs among us cancer patients, 2006-2020. *Annals of Oncology*, 32(7):926–932, 2021. 108

[247] European 1+ million genomes initiative. https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes. Accessed: November 2022. 108

[248] Beyond 1 million genomes. https://b1mg-project.eu. Accessed: November 2022. 108

[249] Acceleration research in genomic oncology. https://www.icgc-argo.org/page/64/about-icgc-argo. Accessed: November 2022. 108

[250] Denis Leonardo Jardim, Maria Schwaederle, David S Hong, and Razelle Kurzrock. An appraisal of drug development timelines in the era of precision oncology. *Oncotarget*, 7(33):53037, 2016. 109

Appendix A

Abbreviations

A

API	Application Programming Interface
ARGO	Accelerating Research in Genomic Oncology
ASCO	American Society of Clinical Oncology
ATC	Anatomical Therapeutic Chemical

B

BAM	Binary alignment map
B1MG	Beyond 1 Million Genomes

C

CGI	Cancer Genome Interpreter
CIViC	Clinical Interpretation of Variants in Cancer

ClinVAP	Clinical variant annotation pipeline
ClinGen	Clinical Genome Resource
CNV	Copy number variant
COSMIC	Catalogue of Somatic Mutations in Cancer

D

DNA	Deoxyribonucleic acid

E

ECM	Extra cellular matrix

F

FGF	Fibroblast growth factor

G

GA4GH	Global Alliance For Genomics and Health
GATK	Genome Analysis Toolkit
GDKB	Gene Drug Knowledge Base
GRN	Gene Regulatory Network
GUI	Graphical User Interface

GWAS	Genome-wide association study

H

HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HPC	High Performance Computing

I

ICGC	International Cancer Genome Consortium
ICD-10	International Classification of Diseases,Tenth Revision
IHC	Immunohistochemical staining
INDEL	Insertion/deletion
IRB	Internal review board

J

JSON	JavaScript Object Notation

K

KB	Knowledge base
KEGG	Kyoto Encyclopedia of Genes and Genomes

L	
LOFTEE	Loss-Of-Function Transcript Effect Estimator

M	
MoA	Mechanism of action
MTB	Molecular Tumor Board

N	
NBS	Network-based stratification
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NSCLC	NonSmall Cell Lung Cancer

P	
PD	Pharmacodynamics
PGx	Pharmacogenomics
PF	Pharmacokinetics
PeCaX	Personalized Cancer Network Explorer
PharmGKB	Pharmacogenomics Knowledge Base
PolyPhen	Polymorphism Phenotyping

PPI	Protein-protein interaction

R

REST	Representational State Transfer
RNA	Ribonucleic acid
RNAP	RNA polymerase

S

SAM	Sequence alignment map
SBML	Systems Biology Markup Language
SNV	Single nucleotide variant
SO	Sequence ontology
SOP	Standard Operating Procedure
SIFT	Sorting Intolerant From Tolerant
SV	Structural variant
SwissMTB	Swiss Molecular Tumor Board

T

TARGET	Tumor Alterations Relevant for Genomics-driven Therapy
TCGA	The Cancer Genome Atlas

TF	Transcription factor
TSG	Tumor suppressor gene
TSGene	Tumor suppressor gene database
TTD	Therapeutic Target Database

U

UKT	University Hospital Tübingen
UMI	Unique molecular identifier
UUID	Universally unique identifier

V

VAF	Variant allele frequency
VCF	Variant call format
VEP	Variant effect predictor
VEGF-A	Vascular endothelial growth factor-A

W

WES	Whole-exome sequencing
WGS	Whole-genome sequencing

Appendix B

Contributions

All ideas, approaches and results presented in this work were developed and discussed with my supervisors Prof. Dr. Oliver Kohlbacher (OK). The collaborators who contributed to the different projects are:

- Bryant Joseph Gilot (BJG)
- Charlotta Scharfe(CH)
- Ghazaleh Tabatabai (GT)
- Julian Heinrich (JH)
- Lukas Zimmermann (LZ)
- Matthew Divine (MD)
- Mirjam Figaschewski (MF)
- Thorsten Tiede (TT)

Chapter 3: Targeted Therapy Identification in Precision Oncology

OK, CS, JH, MD and myself: Project concept and the study design, drafting or revising the manuscript; LZ: Insights on Singularity containers; BJG: Input on diagnosis match score calculation via ICD10 term partitions. Implementation of ClinVAP, data acquisition, stress test and interpretations of the results are performed by me.

Chapter 4: Interactive Case Exploration with PeCaX

OK, MS, TT and myself: Project concept and the study design, drafting and revising the manuscript; myself: Implementation of the PeCaX compatible version of the variant annotation component; MF: Implementation of PeCaX GUI; TT: Implementation of the network component SBML4j; MF, TT and myself: Implementation the communication between PeCaX microservices and their deployment.

Chapter 5: Clinical Assessment of Evidence-based Reporting Strategy in the Precision Oncology Workflow

OK, GT, BJG and myself: Project concept and the study design, drafting or revising the manuscript; myself: data acquisition and data analysis; OK, GT and myself: interpretations of the results.

Appendix C

Publications

2023

Figaschewski M., **Sürün B.**, Tiede T., Kohlbacher O. (2023) "The personalized cancer network explorer (PeCaX) as a visual analytics tool to support molecular tumor boards" *BMC Bioinformatics* **24**, 88.

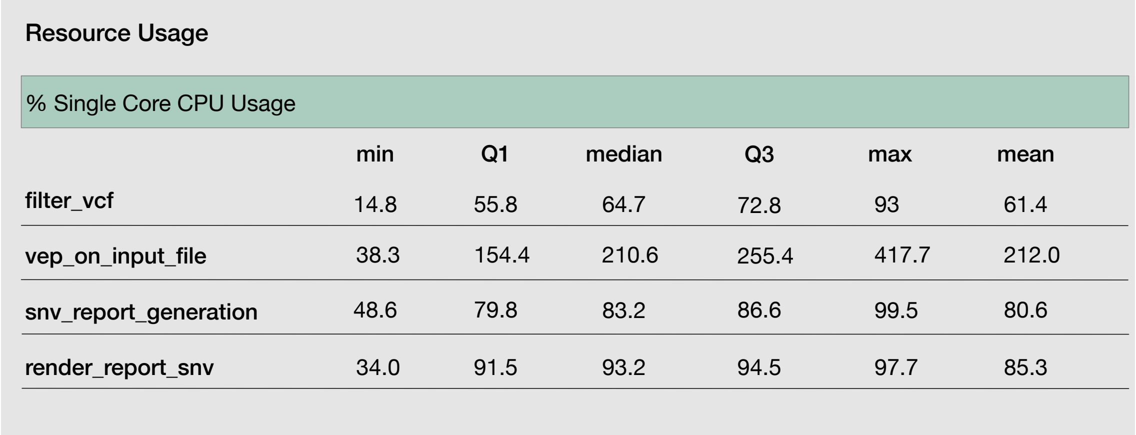
2020

Sürün B., Schärfe C.P.I., Divine M.R., Heinrich J., Toussaint N.C., Zimmermann L., Beha J., Kohlbacher O. (2020) "ClinVAP: a reporting strategy from variants to therapeutic options" *Bioinformatics* **36**, 2316-2317.

Appendix D

Supporting Tables

Table D.1: CPU usage

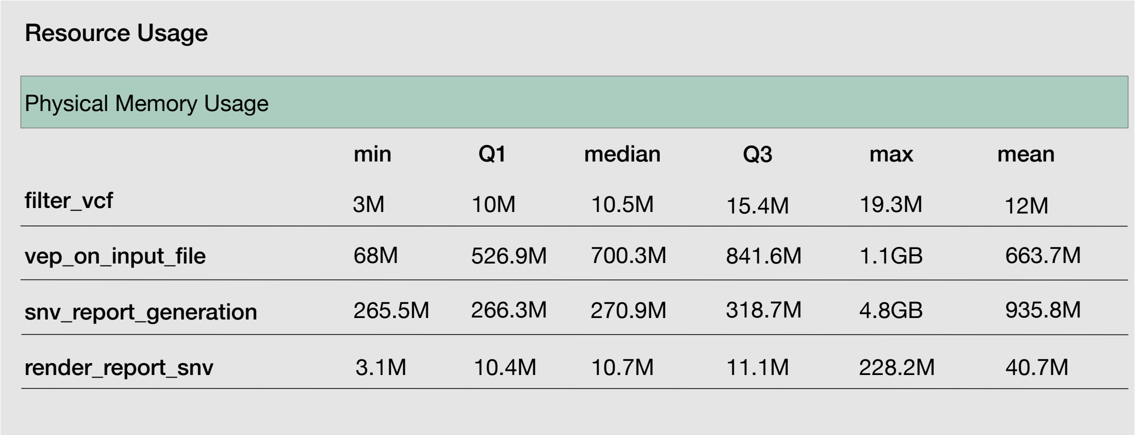


Resource Usage

% Single Core CPU Usage

	min	Q1	median	Q3	max	mean
filter_vcf	14.8	55.8	64.7	72.8	93	61.4
vep_on_input_file	38.3	154.4	210.6	255.4	417.7	212.0
snv_report_generation	48.6	79.8	83.2	86.6	99.5	80.6
render_report_snv	34.0	91.5	93.2	94.5	97.7	85.3

Table D.2: Memory usage

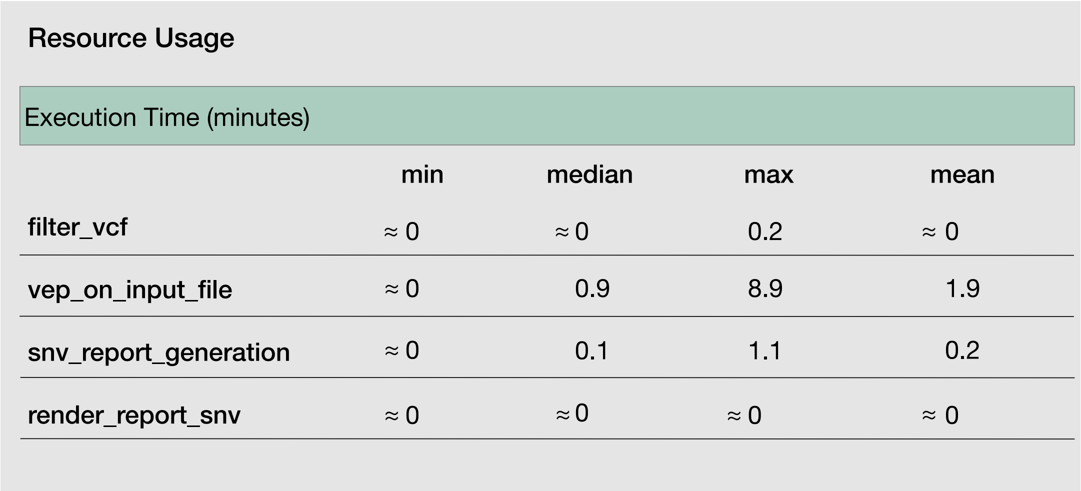


Resource Usage

Physical Memory Usage

	min	Q1	median	Q3	max	mean
filter_vcf	3M	10M	10.5M	15.4M	19.3M	12M
vep_on_input_file	68M	526.9M	700.3M	841.6M	1.1GB	663.7M
snv_report_generation	265.5M	266.3M	270.9M	318.7M	4.8GB	935.8M
render_report_snv	3.1M	10.4M	10.7M	11.1M	228.2M	40.7M

Table D.3: Execution time



Resource Usage

Execution Time (minutes)

	min	median	max	mean
filter_vcf	≈ 0	≈ 0	0.2	≈ 0
vep_on_input_file	≈ 0	0.9	8.9	1.9
snv_report_generation	≈ 0	0.1	1.1	0.2
render_report_snv	≈ 0	≈ 0	≈ 0	≈ 0

Table D.4: Number of variants included in the PeCaX pseudo test data

Cancer type	number of SNVs	number of CNVs
Chronic myeloid leukemia	131	0
Any cancer type	82	30
Cutaneous melanoma	70	12
Lung cancer	54	10
Non-small cell lung cancer	53	3
Lung adenocarcinoma	43	12
Breast adenocarcinoma	31	22
Colorectal adenocarcinoma	18	19
Thyroid carcinoma	3	0

Data		Clinical Annotation			Network generation			Overall		
		1	2	3	1	2	3	1	2	3
Chronic myeloid leukemia	SNV	42	50	53	10	9	9	52	59	62
Any cancer type	SNV	36	42	43	122	140	130	158	182	173
	SNV & CNV	381	374	382	248	291	349	629	665	731
Cutaneous melanoma	SNV	42	42	53	63	48	55	105	90	108
	SNV & CNV	194	166	183	183	221	208	377	387	391
Lung cancer	SNV	42	52	42	22	28	30	64	80	72
	SNV & CNV	208	163	172	86	133	136	294	296	308
NSCLC	SNV	42	42	52	16	20	25	58	62	77
	SNV & CNV	85	112	82	21	38	33	106	150	115
Lung adenocarcinoma	SNV	43	42	42	65	69	73	108	111	115
	SNV & CNV	182	183	180	64	110	95	246	293	275
Breast adenocarcinoma	SNV	43	46	43	23	20	20	66	66	63
	SNV & CNV	283	289	296	63	82	88	346	371	384
Colorectal adenocarcinoma	SNV	42	42	42	48	48	50	90	90	92
	SNV & CNV	254	268	137	73	81	81	327	349	218
Thyroid carcinoma	SNV	42	52	62	17	17	20	59	69	82

Table D.5: PeCaX's processing time[s] on pseudo case data. It is measure from the start of the processes until the results are displayed. Data was analyzed three times.