UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA E BIOINFORMATICA
XXXVIII CICLO - 2025

Design and clinical deployment of OncoVI - an oncogenicity variant interpretation tool

PhD Thesis by

Maria Giulia Carta

Advisors:
Prof. Paolo Magni
PD Dr. Fulvia Ferrazzi

PhD Program Chair:
Prof. Riccardo Bellazzi

La vita ha senso solo se si vive «contro».

Il conformismo uccide la creatività e finisce per annientare l'uomo.

Oliviero Toscani

Abstract (English)

In precision oncology, next-generation sequencing (NGS) has played a crucial role in identifying biomarkers that guide accurate diagnosis and inform treatment decisions. Uncovering these biomarkers typically requires the analysis of multiple molecular data layers, such as genomics, transcriptomics, and proteomics, to capture the complexity of tumour cells.

While the technical and financial barriers to NGS data generation have been constantly diminishing, an important bottleneck now lies in the data analysis and clinical interpretation of results. Interdisciplinary molecular tumour boards (MTB)s have been formed to align molecular findings with available treatment options. Beyond establishing a correct diagnosis, central to this process is the consistent and accurate interpretation of genomic variants, aimed at determining their oncogenic role in tumour development and progression. This involves three key steps: (i) variant functional annotation, which predicts the potential effect of variants on gene transcripts, and protein sequence, (ii) collection of biomedical evidence supportive of this effect in the context of the neoplastic disease, and ultimately (iii) classification of the variant as benign or oncogenic. However, there is still no consensus on how these steps should be carried out. On the one hand, the first two steps lack standardised procedures and well-defined criteria, leaving institutions to adopt heterogeneous approaches. On the other hand, although internationally-recognised guidelines have been published to harmonise the classification of variant oncogenicity, there are very few publicly available tools that automate their assessment, which hinder consistent variant classification across clinical institutions.

This PhD thesis work focused on streamlining and harmonising the interpretation of genomic variants in the context of tumour molecular profiling, with the ultimate goal of advancing precision oncology.

Chapter 1 outlines the motivations driving this PhD work and provides an overview of the clinical applications of the strategies developed throughout the thesis.

Chapter 2 presents the current NGS technologies for DNA analysis employed for tu-

mour molecular characterisation in precision oncology, namely targeted, whole exome, and whole genome sequencing. In this Chapter, the bioinformatics analysis of genomic data is described and the current challenges in genomic variant interpretation are introduced.
Chapter 3 describes the development and implementation of a bioinformatics framework designed to streamline the first two steps of the variant interpretation process, i.e., variant functional annotation and interrogation of biomedical knowledgebases. In addition to pipeline design, the Chapter presents the results obtained from testing the pipeline on a cohort of advanced urothelial carcinoma patients.
Chapter 4 introduces a novel oncogenicity variant interpreter tool, OncoVI. In addition to incorporate the bioinformatics framework described in Chapter 3, OncoVI provides the oncogenicity classification of cancer variants in accordance with internationally-recognised guidelines. Thus, OncoVI represents a unique tool performing all the three aforementioned steps. The Chapter describes the design of OncoVI, its implementation, and data sources utilised. Furthermore, results obtained from its application on gold-standard and real-world data sets of genomic variants are presented.
Ultimately, Chapter 5 presents the deployment of OncoVI in the clinical practice of two pathology departments: the pathology unit in Erlangen (Germany) and a fully digitized pathology department in Caltagirone (Italy). OncoVI was integrated into a broader, fully-automated, integrative framework developed to streamline the entire clinical workflow, from the download of NGS data to the execution of all the required downstream bioinformatics analyses.
All research activities described in this PhD work were developed and tested at the Institute of Pathology, University Hospital Erlangen, Germany (UKER), where I worked as a bioinformatician supporting the Molecular Diagnostics unit.

Abstract (Italian)

Nell'oncologia di precisione, il sequenziamento di nuova generazione (NGS) ha giocato un ruolo cruciale nell'identificazione di biomarcatori che guidino una diagnosi accurata e supportino le decisioni terapeutiche. L'individuazione di questi biomarcatori richiede tipicamente l'analisi di molteplici livelli di dato molecolare, quali genomica, trascrittomica e proteomica, al fine di catturare la complessità delle cellule tumorali.
Sebbene le barriere tecniche ed economiche alla generazione di dati NGS si stiano costantemente riducendo, oggi un'importante criticità risiede nell'analisi di questo tipo di dati e nell'interpretazione clinica dei risultati. A tal proposito, sono stati istituiti i Molecular Tumour Boards (MTBs), ovvero dei gruppi interdisciplinari che hanno l'obiettivo di allineare le risultanze molecolari con le opzioni terapeutiche disponibili. Oltre a stabilire una corretta diagnosi, l'elemento centrale di questo processo è l'interpretazione coerente e accurata delle varianti genomiche, con lo scopo di determinarne il loro ruolo nello sviluppo e nella progressione tumorale. Ciò comporta tre passaggi chiave: (i) l'annotazione funzionale delle varianti, che ha lo scopo di predirre il potenziale effetto delle varianti sui trascritti genici e le sequenze proteiche, (ii) la raccolta di evidenze biomediche a supporto di tale effetto nel contesto delle malattie neoplastiche, e infine (iii) la classificazione della variante intesa come benigna o oncogenica. Tuttavia, per la comunità scientifica non è ancora chiaro come questi tre passaggi debbano essere affrontati. Da un lato, per i primi due passaggi mancano delle procedure standardizzate e criteri ben definiti, lasciando così agli istituti la libertà di adottare approcci eterogenei. Dall'altro lato, sebbene siano state pubblicate delle linee guida riconosciute a livello internazionale al fine di armonizzare la classificazione di oncogenicità delle varianti, esistono pochi strumenti disponibili pubblicamente che effettivamente automatizzino la valutazione delle linee guida, rappresentando un ostacolo alla classificazione coerente delle varianti tra cliniche diverse.
Questo lavoro di tesi di dottorato si è focalizzato sull'ottimizzazione ed armonizzazione dell' interpretazione delle varianti genomiche nel contesto della profilazione molecolare dei tumori, con lo scopo principale di favorire il progresso dell'oncologia di precisione.

Il Capitolo 1 illustra le motivazioni alla base di questo lavoro di dottorato e fornisce una panoramica delle applicazioni cliniche delle strategie sviluppate nel corso della tesi.
Il Capitolo 2 presenta le attuali tecnologie NGS per l'analisi del DNA, quali il sequenziamento mirato, quello dell'esoma e del genoma completo, impiegate nell'ambito della caratterizzazione molecolare dei tumori nell'oncologia di precisione. Inoltre, in questo Capitolo viene descritta l'analisi bioinformatica dei dati genomici e vengono introdotte le sfide attuali che caratterizzano l'interpretazione delle varianti genomiche.
Il Capitolo 3 descrive lo sviluppo e l'implementazione di un framework bioinformatico progettato per ottimizzare i primi due passaggi del processo di interpretazione delle varianti, ovvero: l'annotazione funzionale e l'interrogazione di risorse biomediche. Oltre alla progettazione della pipeline, questo Capitolo presenta i risultati ottenuti dalla sua valutazione su una coorte di pazienti con carcinoma uroteliale avanzato.
Il Capitolo 4 introduce un nuovo tool, un interprete dell'oncogenicità delle varianti, chiamato OncoVI. Oltre ad incorporare il framework bioinformatico sviluppato nel Capitolo 3, OncoVI fornisce la classificazione di oncogenicità delle varianti tumorali, in accordo con le linee guida riconosciute a livello internazionale. Quindi, OncoVI rappresenta un unico strumento che esegue tutti e tre gli steps sopracitati. Il Capitolo descrive il design di OncoVI, la sua implementazione e le fonti di dati utilizzate a tal fine. Inoltre, il Capitolo presenta i risultati ottenuti dall'applicazione dello strumento su gruppi di varianti gold standard e provenienti dal mondo reale.
Infine, il Capitolo 5 descrive l'integrazione di OncoVI nella pratica clinica di due dipartimenti di anatomia patologica: un'unità di patologia a Erlangen (Germania) e un dipartimento di patologia completamente digitalizzato a Caltagirone (Italia). OncoVI è stato integrato in un più ampio flusso di lavoro completamente automatizzato, volto a ottimizzare l'intero framework clinico, a partire dal download dei dati NGS fino all'esecuzione di tutte le analisi bioinformatiche a valle.
Tutte le attività di ricerca descritte in questa tesi di dottorato sono state sviluppate e valutate presso l'Istituto di Patologia dell'Ospedale Universitario di Erlangen in Germania (UKER), dove ho lavorato come bioinformatico a supporto dell'unità di Diagnostica Molecolare.

List of Abbreviations

ACK acknowledgment

AF allele frequency

API application programming interface

AP-LIS anatomic-pathology laboratory information system

BAF B-allele frequency

BAM binary alignment map

B Benign

BED browser extensible data

B/LB Benign/Likely Benign

BQSR base quality score recalibration

BCL binary base call

BQ base quality

BWA-mem Burrows-Wheeler aligner

CGC Cancer Gene Census

cBioPortal cBio Cancer Genomics Portal

CGI Cancer Genome Interpreter

COSMIC Catalogue Of Somatic Mutations In Cancer

CI confidence interval

CIViC Clinical Interpretation of Variants in Cancer

ClinGen/CGC/VICC Clinical Genome Resource/Cancer Genomics Consortium/Variant Interpretation for Cancer Consortium

CNA copy number alteration

CNV copy number variation

CUP cancer of unknown primary

dbNSFP database of non-synonymous functional predictions

DDR DNA Damage Response

DKTK Deutsches Konsortium für Translationale Krebsforschung

DNA deoxyribonucleic acid

ExAC Exome Aggregation Consortium

FDA Food and Drug Administration

FFPE formalin-fixed paraffin-embedded

gDNA genomic DNA

GIS genomic instability score

gnomAD Genome Aggregation Database

H&E hematoxylin and eosin

HL7 Health Level Seven

HRD homologous recombination deficiency

ICA Illumina Connected Analytics

indels insertion/deletions

LB Likely Benign

LO Likely Oncogenic

LOH loss of heterozygosity

LOVD Leiden Open Variation Database

LST large-scale transitions

KBs knowledgebases

MAF minor allele frequency

MAF file mutation annotation format file

MANE Matched Annotation NCBI EBI

MQ mapping quality

MIBC muscle-invasive bladder cancer

MIER cohort Muscle-Invasive ERlangen cohort

MLLP Minimal Lower Layer Protocol

MMR mismatch repair

mRNA messenger RNA

mRNA-Seq mRNA sequencing

MSI microsatellite instability

MTBP Molecular Tumor Board Portal

MTB cohort Molecular Tumour Board cohort

MTB molecular tumour board

mUC metastatic urothelial carcinoma

NAS network-attached storage

NCT Nationales Centrum für Tumorerkrankungen

NCBI National Center for Biotechnology Information

NGS next-generation sequencing

NMD nonsense-mediated mRNA decay

OM1 Oncogenic Moderate-1

OM2 Oncogenic Moderate-2

OM3 Oncogenic Moderate-3

OM4 Oncogenic Moderate-4

OMLˆO33 Laboratory Order Messages

OncoKB Oncology Knowledge Base

O Oncogenic

OG oncogene

O/LO Oncogenic/Likely Oncogenic

OncoVI Oncogenicity Variant Interpreter

OP1 Oncogenic Supporting-1

OP2 Oncogenic Supporting-2

OP3 Oncogenic Supporting-3

OP4 Oncogenic Supporting-4

OS1 Oncogenic Strong-1

OS2 Oncogenic Strong-2

OS3 Oncogenic Strong-3

OULˆR21 Unsolicited Laboratory Observation Messages

OVS1 Oncogenic Very Strong-1

PARPi PARP inhibitors

PCGR Personal Cancer Genome Reporter

phastCons phastCons100way_vertebrate_rankscore

phyloP phyloP100way_vertebrate_rankscore

P/LP Pathogenic/Likely Pathogenic

PON panel of normal

PCR polymerase chain reaction

RefSeq Reference Sequence Database

RNA ribonucleic acid

SAM sequence alignment map

SBP1 Somatic Benign Supporting-1

SBP2 Somatic Benign Supporting-2

SBS1 Somatic Benign Strong-1

SBS2 Somatic Benign Strong-2

SBVS1 Somatic Benign Very Strong-1

SNV single nucleotide variant

SOP Standard Operating Procedure

SV structural variant

TAI telomeric allelic imbalance

TCGA The Cancer Genome Atlas

TCP/IP Transmission Control Protocol/Internet Protocol

TSO500 TruSight Oncology 500

TSO500+HRD TruSight Oncology 500 HRD

TST170 TruSight Tumor 170

TMB tumour mutational burden

TSG tumour suppressor gene

UC urothelial cancer

UCSC University of California Santa Cruz

UKER University Hospital Erlangen, Germany

VAF variant allele frequency

VCF variant call format

VEP Variant Effect Predictor

VUS Variant of Unknown Significance

WES whole exome sequencing

WGS whole genome sequencing

WT wildtype

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Next-generation sequencing (NGS) refers to comprehensive and highly effective high-throughput methods that allow sequencing different molecules such as deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) in a short time, thus enabling large-scale analyses at genome-wide level [1]. These technologies work by fragmenting the genetic material into millions to billions of small pieces, allowing for example to sequence an entire human genome simultaneously, which was not feasible with earlier low-throughput methods such as Sanger sequencing [2, 3].

Although at the beginning these technologies have mainly been used in the research field, in recent years they have been increasingly adopted in the clinical practice, such as in the context of hereditary diseases, cancer, and infectious diseases [4–8]. In particular, the oncology field has greatly benefited from the adoption of NGS techniques. Indeed, in the last century cancer was classified as a genomic disease [9] and it became clear that tumour cells carry particular modifications in the genetic information, which can be identified on a large-scale only with NGS assays. These changes in the genetic material of tumour cells are defined "somatic" when they arise during a person's lifetime, often due to the exposure to carcinogenic agents or random errors in DNA replication. Somatic changes can be identified in tumour cells only. In contrast, hereditary tumours are caused by germline alterations, which are typically inherited by children from their parents and are present in all cells of the human body [10].

In cancer research, genomic NGS approaches have been adopted earlier to uncover the molecular foundations underlying the biological and clinical diversity of human cancer [11, 12]. Whole transcriptome profiling offered the first comprehensive overview of tumour molecular heterogeneity [13]. In particular, mRNA sequencing (mRNA-Seq) has been exploited to quantify RNA transcripts within a given tumour entity and identify molecular subgroups sharing similar expression profiles [14]. In breast and bladder can-

cer, for example, transcriptome analysis allowed identifying clinically relevant molecular subtypes, which can influence prognosis and treatment decisions [15, 16]. Transcriptome analysis is also able to support the identification of underlying biological pathways, immune microenvironment signatures, and potential therapeutic targets [17]. In clinical oncology, targeted mRNA-Seq is commonly used for the identification of therapeutically-relevant gene fusion events at the RNA level, resulting from chromosomal rearrangements in the DNA sequence [18, 19]. Focusing on a limited set of genes, this approach sequences the RNA transcripts expressed in the tumour, allowing the detection of fusion transcripts that typically arise from the abnormal joining of two genes in the DNA sequence. In addition to bulk RNA-sequencing (RNA-Seq), in cancer research more recent techniques like single-cell RNA-Seq and spatial transcriptomics have been leveraged to reveal the composition and spatial organisation of cell subpopulations in the tumour microenvironment, enhancing our understandings of intra-tumoral heterogeneity, cellular interactions, and treatment resistance [12, 20–22].

While RNA-based methods provide functional and expression-based insights, DNA sequencing remains central to cancer diagnostics, enabling the identification of key genomic alterations that guide prognosis and therapeutic decisions. Large genomics studies have been conducted to analyse the DNA of tumour cells using different approaches, from targeted sequencing to whole exome sequencing (WES) and whole genome sequencing (WGS) [23–26]. All techniques, with their strength and limitations, detect somatic alterations at single or few nucleotides level as well as variations involving gene copy number, structural events, and other biomarkers that might be predictive of drug response and inform cancer care. Targeted sequencing focuses on predefined genomic regions of interest, allowing for high-depth analysis of tens to few hundreds of genes commonly mutated in cancer, which may have diagnostic, prognostic or therapeutic significance. WES expands this approach to all protein-coding regions of the entire human genome, commonly known as the "exome", and allows for comprehensive detection of mutations in coding regions, including those with potential clinical significance. In contrast, WGS further extends to the non-coding elements and regulatory regions of the genome, thus offering a broader coverage by capturing more complex genomic events that involve the entire gene length or large chromosomal regions [27]. Together, RNA- and DNA-based technologies contribute to a multidimensional understanding of tumour biology, facilitating more precise tumour classification and informing targeted therapeutic strategies.

While targeted sequencing approaches remain the most widely used molecular tests in clinical oncology, they may miss less common or unexpected clinically relevant mutations. Instead, more comprehensive strategies such as WES and WGS can capture the full molecular landscape of tumours, providing additional insights that may guide pa-

tient stratification and therapeutic decisions. As the costs of WES and WGS continue to decrease, their clinical application is increasingly becoming physically and economically feasible [23]. Together, it has become more concrete to physicians and healthcare stakeholders the strong potential of integrating comprehensive NGS approaches into clinical oncology practice, thus opening up the possibility of applying precision oncology also to rare tumour types. Indeed, at present, it is difficult to imagine applying precision oncology, which identifies personalised therapies based on the molecular characteristics of each tumour, without genomic profiling using NGS [28].

Although many of the economic and technical barriers to NGS data acquisition and sequencing have been largely addressed [19], significant challenges persist in the downstream stages, particularly in the data analysis and interpretation. The key steps central to unleash the full potential of precision oncology, but still complex and time-consuming, include: (i) variant functional annotation, (ii) biomedical knowledgebases interrogation, and (iii) final classification of variant oncogenicity.

Variant functional annotation, which predicts the potential effects of variants on gene transcripts, is a labor-intensive task requiring the integration of information from multiple resources to gather all existing, publicly available biological evidence supporting the variant potential deleterious impact on protein function. However, the data collected from these knowledgebases are often limited, particularly for rare variants, or even contradictory. Thus, variant interpretation, i.e., the process of evaluating and synthesizing these evidence, is both time-consuming and heavily reliant on manual curation [29]. In addition, due to the nature of this process, different institutions often perform variant interpretation independently relying on local procedures, which hinders the consistent and standardised translation of molecular findings into therapeutics decisions. This relevant issue has been already recognised and to address the downstream step of variant classification, i.e., the standardised assignment of variants into oncogenicity categories, internationally-recognised guidelines have been proposed, defining classes of oncogenicity for somatic variants [30–32]. Nevertheless, there are few tools automating this evaluation, thus missing the opportunity to streamline and standardise variant classification across different institutions.

Consequently, as the volume of NGS data continues to grow, molecular biologists and clinicians face increasing challenges in bridging the gap between the alterations detected in patients' cancer genomes and the number of these that can be interpreted and explained [33]. Indeed, the analysis of a cancer patient's exome or genome can produce hundreds to thousands of potentially harmful genomic variants. However, in routine diagnostics settings, it is not feasible to interpret each detected variant and establish its final oncogenicity classification, i.e., defining its potential association with the neoplastic disease.

A key limitation is the turnaround time, which is the period from when the tumour tissue arrives at the pathology department to the generation of a clinical report. Consequently, robust bioinformatics frameworks that efficiently and accurately support genomic data analysis, and implement structured guidelines for variant classification are essential, thus enabling the identification of clinically relevant alterations, ultimately delivering reproducible and accurate diagnoses [34].

This PhD thesis work stemmed from real-world needs that are broadly shared among institutions seeking to implement precision oncology in the clinical practice. My involvement as bioinformatician supporting the Molecular Diagnostics unit at the Institute of Pathology UKER, provided me with a clear and practical understanding of these challenges and of potential strategies to address them . Here, in 2016, a Molecular Tumour Board (MTB) was established, bringing together a multidisciplinary team to advance precision oncology for cancer patients who had exhausted standard-of-care treatments with curative intent. Initially relying on targeted sequencing approaches covering few hundred genes, the centre has since expanded its capabilities, ultimately leveraging WES and WGS profiling as part of routine molecular diagnostics. As outlined above, while WGS and WES provide unprecedented opportunities to detect rare or unexpected clinically relevant mutations beyond the scope of targeted sequencing, their analysis and interpretation remain technically challenging. In practice, these steps emerged as the major bottleneck in the clinical workflow, limiting experts' ability to fully understand patient's tumour biology and translate clinically relevant alterations into available treatment decisions.

To address the challenges of time-consuming and inconsistent genomic data interpretation, my PhD work focused on streamlining and standardising this process in the context of precision oncology. For this purpose, first a bioinformatics framework was developed to: (i) filter variants, (ii) perform variant functional annotation, and (iii) biomedical knowledgebases interrogation, thereby assisting biologists in the reproducible collection of the required evidence for variant interpretation. This work required selecting appropriate filters to remove sequencing artefacts, as well as recurrent mutations in the population, unlikely to be cancer-associated. Then, an appropriate tool to perform variant functional annotation was chosen and application programming interfaces (API)s were integrated in the bioinformatics workflow to query their respective knowledgebases. This bioinformatics framework was tested on tumour profiling data of a cohort of advanced urothelial carcinoma patients, demonstrating its potential to accelerate and standardise variant annotation and knowledgebases interrogation in a real-world clinical setting.

The assessment and synthesis of the collected evidence is often influenced by the evaluator subjectivity if not supported by structured guidelines, leading to inconsistent variant classifications across institutions. To overcome this limitation, a novel tool was

developed to automate the application of internationally-recognised guidelines for the oncogenicity classification of somatic variants (ClinGen/CGC/VICC) [30]. This led to the development of OncoVI, an oncogenicity variant interpreter, that incorporating the bioinformatics framework previously developed: (i) performs variant functional annotation, (ii) interrogates knowledgebases to collect biological and clinical evidence, and (iii) classifies variant oncogenicity according to structured guidelines. To develop OncoVI, first the text-based rules proposed by the ClinGen/CGC/VICC guidelines were translated into Python IF-ELSE rules, then publicly available resources were selected to gather supporting evidence for each rule, and the point-based scoring system proposed by the guidelines authors was implemented [30]. The performance of OncoVI was evaluated using both gold-standard and real-world variants, showing its effectiveness in reducing subjectivity and promoting harmonised classification across institutions. The results of this evaluation are presented and discussed throughout this PhD thesis.

Ultimately, the final goal of this PhD work was to enable the deployment of the developed tools into clinical practice, thereby addressing the real-world needs that originally motivated this research. To this end, OncoVI was integrated in the clinical practice as part of a broader fully-automated framework designed to streamline the entire NGS workflow within MTBs. Indeed, this integrative framework automatically performs: (i) the transfer of NGS data from a remote storage to a local environment and (ii) the initiation of the downstream bioinformatics analyses, including variant functional annotation and oncogenicity classification through OncoVI. The effectiveness and flexibility of this integrative framework were tested through its implementation not only in the traditional (analogue) Institute of Pathology UKER but also in a fully digitized Pathology Department in Caltagirone (Italy). By supporting the harmonised and efficient use of NGS data in diverse clinical settings, this framework represents a further step towards translating genomic findings into patient care management.

Chapter 2

Genomic profiling for precision oncology

Precision oncology leverages the genomic information of a tumour to identify personalised therapeutic strategies tailored to individual patients [35]. In clinical oncology practice, patients who do not respond, or develop resistance to standard-of-care treatments with curative intent may be eligible for genomic testing, which can facilitate access to targeted therapies offering greater efficacy and reduced toxicity. In addition, tumour molecular profiling can identify candidates for molecularly guided clinical trial scenarios and provide access to experimental therapies that are not yet approved by medicines agencies [36, 37]. The more NGS technologies advance and new therapeutic biomarkers gain approval, the more likely genomic testing will be integrated in routine oncology diagnostics. Although precision oncology relies on a wide range of omics technologies, this discipline is often used to refer to DNA analyses since it represents the most commonly employed approach for diagnosis and therapy response assessment across many cancer types [19]. The rest of this Chapter will present the NGS-based technologies for DNA analysis currently employed in precision oncology for tumour molecular profiling, and the associated bioinformatics analysis.

2.1 NGS-based technologies for DNA analysis

In oncology, NGS-based technologies for DNA analysis span from targeted sequencing, which covers tens to a few hundreds of genes, to WES and WGS, targeting the full set of coding-genes and the entire genome, respectively. Each technology offers distinct advantages in terms of resolution and clinical utility (**Table 1**) [38, 39].

Table 1: **Comparison of the three major DNA sequencing approaches for clinical oncology applications.** All criteria are reported on a per-sample basis.

Criterion	**Targeted sequencing**	**Whole exome sequencing**	**Whole genome sequencing**
Region Covered	From 10 to hundreds of genes or specific regions of interest	All protein-coding genes (˜1–2% of human genome)	Entire human genome
Typical Read Coverage (Depth)	High (>500X)	Moderate (>100X)	Low (<100X)
Data Output	Small (<500 MB)	Moderate (<200 GB)	Large (>200 GB)
Clinical Utility	High for well-known cancer-associated mutations	High for genes coding functional proteins. Moderate for copy number changes of coding genes	Best for discovery of genome-wide variants, gene copy number changes, and structural rearrangements
Variant Detection	Single nucleotide variants (SNV)s and small insertion/deletions (indels) in cancer-associated genes	SNVs, indels and copy number variations (CNV)s in protein-coding genes	SNVs, indels, CNVs, structural variants, and non-coding variants
Limitations	Require a priori knowledge to select target regions	Moderate discovery potential	High cost, interpretation complexity, fresh tissue required
Best Use Case	Unavailability of fresh tissue	Rare cancers with unavailability of fresh tissue	Comprehensive cancer genomics, rare cancers

2.1.1 Targeted sequencing

Targeted sequencing focuses on a set of cancer-associated genes, typically ranging from tens to hundreds (up to 500), often referred to as "gene panel". By focusing on clinically relevant regions, this approach achieves very deep coverage, enabling the sensitive detection of single nucleotide variants (SNV)s, i.e., changes at single nucleotide level in the genomic sequence and small insertion or deletions (indels) of up to 50-100 nucleotides, even when present in a small fraction of tumour cells (**Figure 1**).

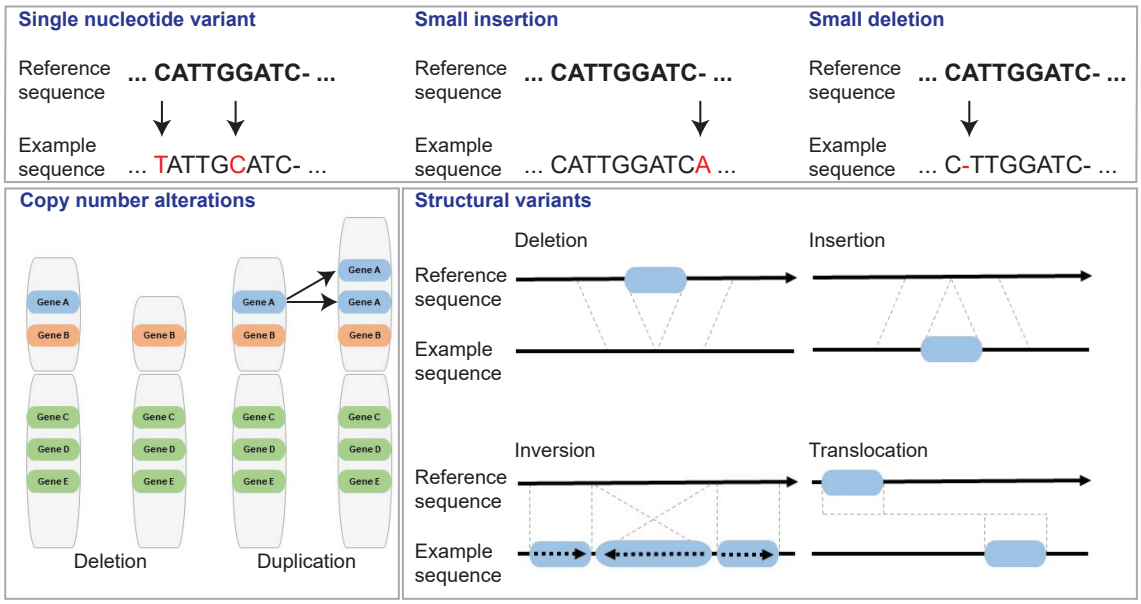

Figure 1: Different types of genomic alterations. Small nucleotide variants such as single nucleotide variants, small deletions and small insertions are changes involving from one to up to 50-100 nucleotides in the example sequence compared to the reference sequence. Copy number alterations and structural variants fall in the category of structural rearrangements, involving the entire gene length or large portions of chromosomes.

Targeted sequencing has become the standard in routine molecular diagnostics thanks to its cost-effectiveness, rapid turnaround time, and compatibility with formalin-fixed paraffin-embedded (FFPE) tissue. Indeed, in clinical practice, solid tumours are commonly archived as FFPE blocks and sliced into one or more thinner tumour tissue slices for two main purposes: 1) histopathological evaluation, i.e., the microscopic examination of tissue morphology, and 2) immunohistochemistry assessments, which detect proteins using antibody-based staining. For such FFPE samples, library preparation commonly relies on hybrid capture-based enrichment, where oligonucleotides probes target coding regions [40]. However, this method requires at least 50 nanograms of input DNA, sometimes unachievable from small tumour biopsies [19]. To overcome this limitation, amplicon-based protocols have been developed, using specific primers and polymerase chain reaction (PCR) amplification to cover target regions from smaller DNA quantities [41].

Since panels typically cover well-characterised genes, variant annotation and interpretation are more straight forward compared to more comprehensive approaches such as WES and WGS. This streamlined analysis makes targeted sequencing particularly amenable for laboratories aiming to extract clinically relevant information without facing the complexity and data volume of WES and WGS methods.

2.1.2 Whole exome sequencing

WES is an intermediate-scale sequencing approach focusing on the protein-coding regions of the genome. Although the protein-coding regions represent approximately ˜1-2% of the whole human genome, they harbour variants in genes coding for functional proteins. Compared to targeted sequencing, WES extends the scope of the analysis beyond a predefined set of genes, enabling a more comprehensive identification of potentially clinically relevant coding alterations. Beyond detecting SNVs and small indels, WES can also provide estimates of copy number alterations (CNA)s within protein-coding regions (**Figure 1**). CNAs refer to changes in DNA segments ranging from approximately 1,000 base pairs to 5 megabases in length. These alterations can manifest as duplications, where extra copies of a DNA segment are generated, or as deletions, where existing copies of a DNA segment are lost [42].

WES is a cost-effective alternative to WGS, especially in scenarios where fresh tumour material, typically required for WGS, is not available. Although it generates more data than targeted sequencing, WES still involves intermediate requirements in terms of computation and variant interpretation complexity, making it more accessible than WGS. Thus, WES represents a balanced compromise: it broadens the discovery potential beyond gene panels but without the full cost, data volume, and analytical challenges of whole genome approaches.

2.1.3 Whole genome sequencing

WGS is the most comprehensive approach for capturing the complete set of DNA molecules in the human genome. Unlike WES and targeted sequencing that focus on specific regions, WGS evaluates the entire genomic landscape, including non-coding elements such as regulatory sequences, introns, and intergenic regions [43]. Compared to WES and targeted sequencing, the broad coverage of WGS enables the identification of SNVs and indels at genome-wide level. In addition, WGS offers superior resolution for the detection of CNAs and structural variants (SV)s, which can involve large portions of chromosomes. SVs are alterations in the chromosome structure compared to the reference genome [44, 45] larger that indels and often spanning regions that cannot be captured within a single sequencing

read (**Figure 1**). SVs encompass nucleotides that have been deleted, inserted, inverted or relocated to a different region of the genome, i.e., translocated.

2.2 Bioinformatics analysis of genomics data

The analysis of NGS data from DNA sequencing experiments relies on robust computational platforms and bioinformatics analyses that can manage the high volume of sequencing output. The following sections will describe the steps of a standard bioinformatics workflow for DNA sequencing data by focusing on the detection of SNVs, which represent the most frequent type of alteration in the human genome.

According to the central dogma of molecular biology, DNA is transcribed into mRNA, whose nucleotide sequence is read in triplets (codons); each codon specifies an amino acid, which is incorporated into a growing polypeptide chain that subsequently folds into a functional protein. When a small change occurs within the DNA sequence, this can be classified as a silent, nonsense, or missense variant based on its effect on the encoded amino acid (**Figure 2**).

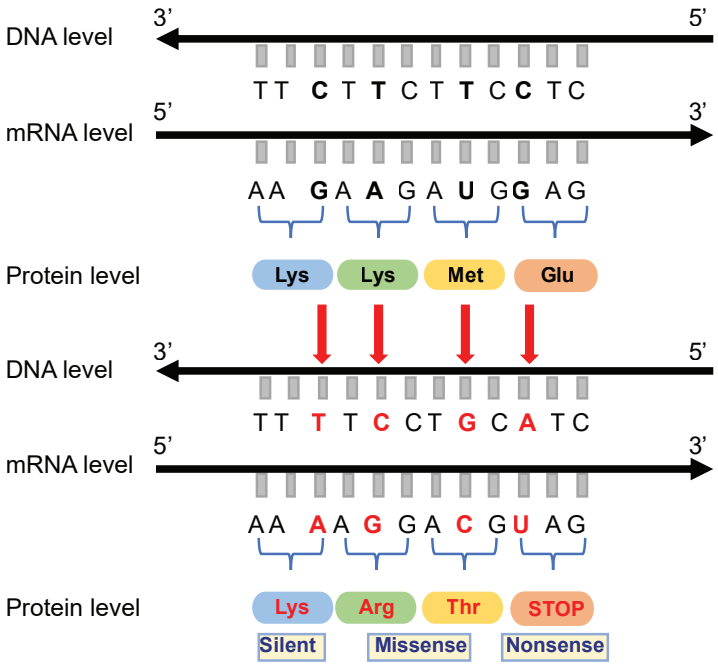

Figure 2: Classification of single nucleotide variants. Silent variant: a nucleotide substitution in the DNA sequence (C → T) corresponds to a change in the mRNA sequence (G → A), but the encoded amino acid remains unchanged (Lys → Lys). Missense variant: a nucleotide substitution in the DNA sequence (T → C or T → G) corresponds to a change in the mRNA sequence (A → G or U → C), resulting in an amino acid substitution at the protein level (Lys → Arg or Met → Thr). Nonsense variant: a nucleotide substitution in the DNA sequence (C → A) corresponds to a change in the mRNA sequence (G → U), which introduces a premature stop codon during translation into protein (Glu → STOP).

Silent variants are single-nucleotide changes within a coding sequence that alter the DNA sequence but do not change the encoded amino acid. As a result, the protein sequence remains unaltered and such variants are generally considered to have low clinical relevance. Missense variants are single-nucleotide substitutions that replace one amino acid with another. They are among the most frequent types of variants in the human genome, with effects that are often benign or of uncertain significance; yet some are associated with diseases. Because their impact on protein function is context-dependent, interpreting missense variants can be challenging. Ultimately, nonsense variants introduce a premature stop codon into the coding sequence, leading to an early termination of the translated sequence. This typically produces a truncated protein with reduced or absent function. Such variants are often associated with diseases, including cancer.

The bioinformatics detection and analysis of SNVs involves multiple steps, usually structured into a standardised workflow, commonly referred to as "pipeline" (**Figure 3**). The following sections will discuss the bioinformatics steps performed in the identification of SNVs, ranging from quality control, to alignment, variant calling, and variant functional annotation, for which various bioinformatics tools have been developed (**Table 2**) [46, 47].

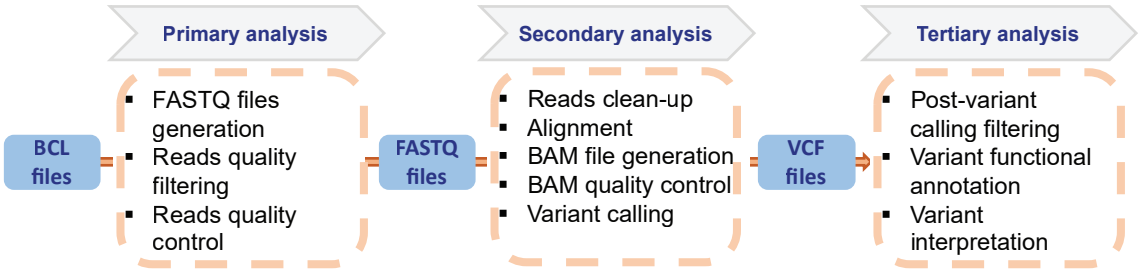

Figure 3: Standard workflow for the analysis of DNA sequencing data for the identification of single nucleotide variants. The bioinformatics pipeline for the analysis of DNA sequencing data includes three main steps: primary analysis, secondary analysis, and tertiary analysis.

Table 2: **Main steps performed in a standard bioinformatics workflow for the analysis of genomic data, their purpose, and the most commonly used bioinformatics tools for each step.**

Analysis Step	**Purpose**	**Commonly Used Tools**
Quality Control	Assess read quality and identify sequencing artefacts	FastQC, MultiQC
Adapter/Quality Trimming	Remove adapters and low-quality bases	Trimmomatic, Cutadapt, fastp
Read Alignment	Map reads to a reference genome	BWA-MEM, Bowtie2
Post-Alignment Processing	Mark duplicates, base quality recalibration	mosdepth, Picard
Somatic Variant Calling	Identify variants in cancer samples	MuTect2, Strelka2
Variant Filtering	Filter out low-confidence variants	BCFtools
Variant Functional Annotation	Predict the effect of the variant on gene transcripts and sequence protein	ANNOVAR, Ensembl Variant Effect Predictor
Interrogation of Knowledgebases	Collect biological and clinical information for variants	COSMIC, OncoKB, ClinVar, CIViC
Variant Interpretation	Evaluate and synthesize the resources collected for the variants	MolTB Portal, Cancer Genome Interpreter
Variant Classification	Assign the variant to one class (e.g., benign or oncogenic), which defines variant role in tumour development and progression	Personal Cancer Genome Reporter

2.2.1 Primary analysis

After DNA sequencing, which produces a large volume of raw data, the very first steps in the bioinformatics pipeline are collectively called “primary analysis”. Primary analysis comprises the generation of FASTQ files and their quality assessment. First, the raw binary base call (BCL) files, produced by the sequencing instrument, are converted into

FASTQ files. A FASTQ file is a text-based file containing the nucleotide sequence of one end of the sequenced DNA fragments , technically referred to as a "read", together with a quality score assigned to each base call (**Figure 4**). In paired-end sequencing, both ends of each DNA fragment are sequenced, resulting in two reads per fragment. Consequently, two FASTQ files are generated for each sample, one for the so called forward reads (R1) and one for the reverse reads (R2).

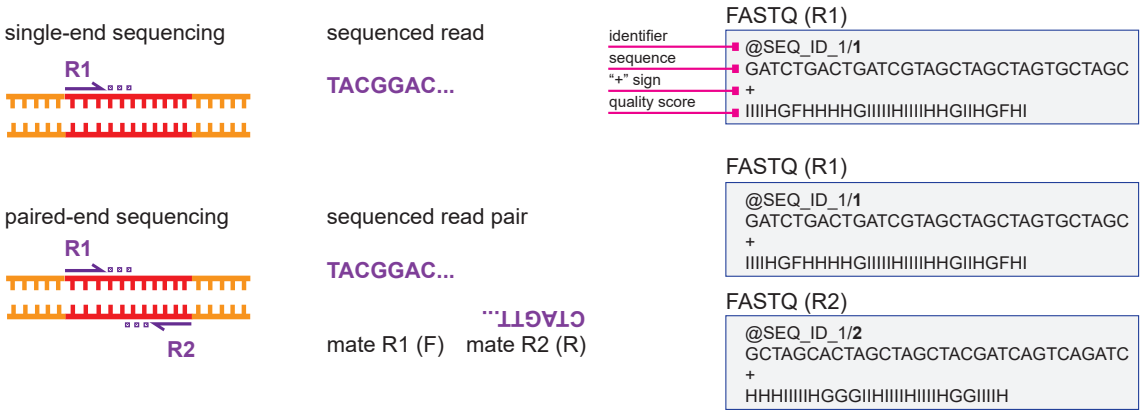

Figure 4: Comparison of single-end and paired-end sequencing. In single-end sequencing, DNA fragments are sequenced from one end only, producing a single FASTQ file containing all reads. In paired-end sequencing, both ends of each DNA fragment are sequenced, resulting in two FASTQ files: one containing the forward (F) reads (R1) and another containing the reverse (R) reads (R2).

FASTQ files are the standard format to store read sequences. More specifically, each read in a FASTQ file is represented by a block of four lines:

1. Line 1 always begins with the "@" character, followed by a unique sequence identifier. This identifier typically includes metadata about the sequencing run, such as the instrument name, flow cell ID, lane, and read number, allowing each read to be uniquely tracked.

2. Line 2 always contains the raw nucleotide sequence, which is a string of letters (A, T, C, G, or N for ambiguous bases) as determined by the sequencer during base calling.

3. Line 3 always starts with the "+" symbol and serves as a separator between line 2 and line 4.

4. Line 4 always provides the quality scores corresponding to each base in line 2. These scores reflect the confidence of the base calls, with one character per base. The quality score, called base quality (BQ) score or Phred score, indicates the probability that a base has been incorrectly identified by the sequencer.

Read pre-processing and quality control are performed during the conversion from BCL to FASTQ files and include quality filtering with the aim of discarding poor quality reads, typically shorter than 35 bases. After quality filtering, quality control is carried out using tools such as FastQC (`https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). This tool provides a comprehensive overview of key quality metrics, including the total number of reads per FASTQ file, BQ scores distribution, sequences length, duplication level, and detection of adapter contamination. FastQC enables bioinformaticians to assess sequencing quality and determine whether processing steps such as base trimming, or read filtering are required before alignment [48].

2.2.2 Secondary analysis

Once the FASTQ files are generated, the "secondary analysis" starts. This, includes read clean-up, alignment, post-alignment controls, and variant calling. Read clean-up typically starts with the trimming of adapters using tools such as fastp [49], Trimmomatic [50] or Cutadapt [51]. Adapter sequences are artificial sequences added to DNA fragments to facilitate their attachment to the flow cell during sequencing and need to be removed.

Subsequently, alignment of the sequenced reads to the human reference genome is performed. Alignment aims at determining the most likely genomic location from which a read originated by identifying the reference sequence portion that best matches the read. This process is complicated by several factors: (i) natural genomic variation (e.g., SNVs, indels, chromosomal rearrangements and CNAs), (ii) sequencing errors, and (iii) the presence of repetitive, or (iv) highly similar genomic regions. A widely used tool for read alignment is the Burrows-Wheeler aligner (BWA-mem) [52], which supports both single-end and paired-end sequencing data. Paired-end sequencing offers significant advantages over single-end sequencing, as the two reads from each fragment provide complementary information, improving alignment accuracy.

The output of the alignment process is stored in a binary alignment map (BAM) file, which is a compressed binary version of the sequence alignment map (SAM) format. BAM files are made up of:

1. A header section that contains metadata about the sample, reference genome, and command used for the alignment process.

2. An alignment section that lists each aligned read, with one record per line organised into 11 mandatory columns including the read identifiers, alignment position, mapping quality, and other attributes.

Alignment cannot be perfect because a sequencing read may align equally well to multiple locations, due to repetitive genomic elements, low-complexity sequences, or gaps or errors in the reference genome. To assess the reliability of each alignment, a mapping quality (MQ) score is assigned to each read. The MQ score, expressed as a Phred-scaled value, reflects the probability that the read is correctly mapped to its reported location. In the case of paired-end sequencing, mapping quality is calculated considering both reads in the pair, i.e., even if one read maps ambiguously, the proper placement of its mate can help to resolve its location.

Following alignment, post-alignment controls must be performed to evaluate the accuracy of the alignment process. Tools such as mosdepth [53] or picard (`https://broadinstitute.github.io/picard/`) are commonly used to assess a variety of metrics, including the percentage of reads mapped to the reference genome, the proportion of uniquely aligned reads, and the percentage of duplicated reads. Genome- or exome-wide coverage of mapped reads are also examined. These QC metrics are often compiled in a html report using tools like MultiQC [54], enabling a streamlined overview of sequencing and alignment quality.

Two additional steps are considered best practice for post-alignment controls: the base quality score recalibration (BQSR) and duplicate marking or removal. BQSR is implemented by tools such as the GATK BaseRecalibrator (`https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-BaseRecalibrato`) that adjusts the BQ scores originally assigned by the sequencing instrument. Indeed, BQ scores can be systematically biased by factors such as: (i) the position of the base within a read, (ii) the local sequence context, (iii) technical artefacts from the sequencing chemistry, or (iv) hardware. Since variant callers heavily rely on accurate BQ scores, systematic over- or under-estimation can reduce the reliability of variant detection [55]. Notably, the GATK procedure applies a machine learning model to detect and correct such biases, using known variant databases, like dbSNP [56], to distinguish true SNVs from sequencing errors. BQSR does not change the base calls themselves but recalibrates the associated quality score up or down according to the modeled error patterns.

Duplicate marking addresses a different source of bias, i.e., the overrepresentation of identical DNA fragments introduced by PCR amplification during library preparation. Such PCR duplicates artificially inflate coverage and can lead to false-positive variant calls. Duplicate reads, identified as having the same start and end positions, are flagged in the BAM file using a bitwise flag field but not removed. Duplicate rates vary by experiment type, with <5% duplicates produced by WGS and up to ~30% for targeted approaches. Since positional information is needed to identify reads potentially originating from the same DNA fragment, duplicate marking is always performed after alignment. Once the

data is cleaned and processed after alignment, variant calling is performed to identify genomic mutations such as SNVs and small indels with respect to the reference genome. This process considers multiple factors, including (i) the BQ scores of supporting reads, (ii) their mapping qualities, and (iii) the sequencing depth. The standard format for storing variant calls is the variant call format (VCF) file, which allows the representation of variants across multiple samples as well as matched tumour-normal pairs. A VCF file consists of two sections (**Figure 5**):

1. Header: one or more lines beginning with the character "##", followed by a mandatory column-definition line beginning with the character "#";
2. Body: one line per site, including mandatory columns: CHROM, POS, ID, REF, ALT, QUAL, FILTER, and INFO. The INFO field can store rich annotations and genotype-specific details are provided in the sample-specific columns.

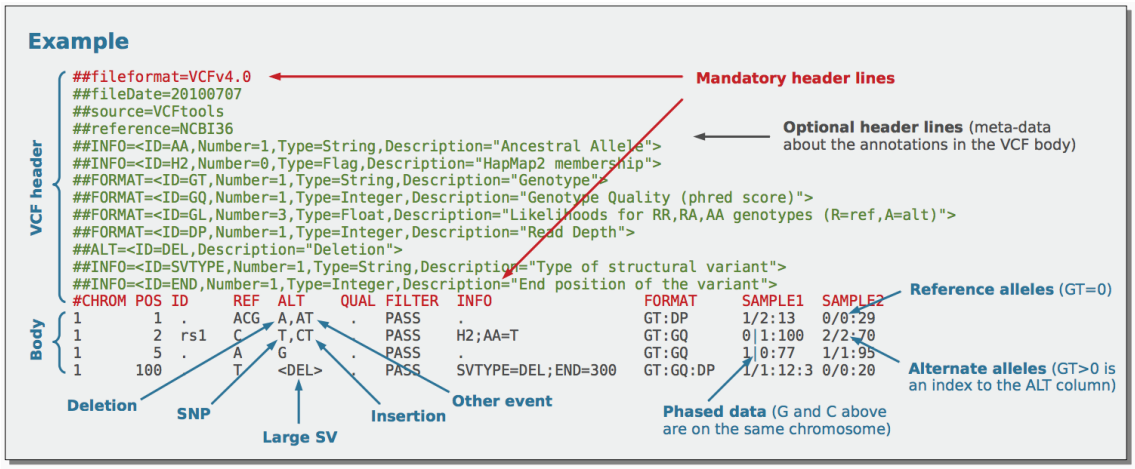

Figure 5: Structure of an exemplary VCF file. The VCF file is organised into two main sections: (i) the header, composed by one or more lines starting with the character "##", and (ii) the body. Data columns are identified by the character "#" and the body section contains as many lines as the variants identified in the sample. The figure is taken from `https://davetang.github.io/learning_vcf_file/`.

In case of matched tumour-normal pairs, the final goal of variant calling is to identify somatic variants present in the tumour but absent in the matched normal sample, thereby excluding germline variants and the majority of sequencing artefacts. The variant allele frequency (VAF), which represents the fraction of reads supporting the variant, is a key parameter for distinguishing somatic from potential germline events. Indeed, germline variants have a VAF close to 0.5 in case of heterozygosity and close to 1 in case of homozygosity. Although somatic variants often have a VAF below 5%, this value is strongly influenced by several factors including: (i) tumour purity, i.e., the fraction of tumour cells

in the sequenced sample, (ii) CNAs, and (iii) tumour heterogeneity, which refers to the coexistence of genetically distinct group of cells (subclones) within the same tumour. In low purity tumours, i.e., with a fraction of tumour cells below 20%, a higher sequencing depth can improve the detection of mutations at low VAF [27, 43]. On the other hand, distinguishing true variants from sequencing artefacts is facilitated by including a matched normal sample alongside the tumour, which serves as critical control against false-positive calls.

Common algorithms for somatic variant calling include Mutect2 (`https://gatk.broadinstitute.org/hc/en-us/articles/9570422171291-Mutect2`) and Strelka2 [57]. The choice of the variant caller should be driven by the type of experiment. For example, Strelka2 performs reliably with WGS but is more sensitive to artefacts in WES, whereas Mutect2 is reliable for both WES and WGS but has been shown to perform poorly with high coverage WGS data [27]. For WES or targeted sequencing experiments, variant calling is typically restricted to genomic intervals defined in a browser extensible data (BED) file, which is a tab-delimited text file specifying the genomic intervals of the sequences targeted by the experiment.

2.2.3 Tertiary analysis

Following variant calling, the "tertiary analysis" aims at identifying variants with potential clinical relevance. This process typically involves steps from post-variant calling filtering to variant functional annotation, and variant interpretation, with the final aim of classifying variant oncogenicity (**Table 2**). Currently, there are no standard guidelines on how variant filtering should be performed [34]. Therefore, institutions use different criteria: some use soft filters, which add information tags to the VCF without excluding variants, while others apply hard filters that remove variants failing specific criteria [34].

In tumour sequencing, the primary goal of hard filtering is to remove false positives and germline variants. Common strategies include excluding variants with low BQ scores (e.g., BQ <20), insufficient depth, or low mapping quality (e.g., MQ <50-60). However, thresholds should be selected according to the project design and sequencing strategy. A widely accepted practice is to retain only high-confidence variants that pass all applied variant calling filters, i.e., those marked with the "PASS" flag in the VCF "FILTER" field.

When a matched normal sample is unavailable, which is often the case in clinical oncology, distinguishing between somatic variants, germline variants, and artefacts becomes more challenging. In such scenarios, common filtering strategies involve removing known variants relying on the dbSNP database and filtering variants based on population allele

frequency data from databases like the Genome Aggregation Database (gnomAD) [58], Exome Aggregation Consortium (ExAC) [59], or the 1000 Genomes Project [60]. Additionally, a panel of normal (PON) samples can be used to exclude recurrent sequencing artefacts and false positives. To ensure the effectiveness of this approach, the normal samples in the PON must be prepared using the same library preparation and sequencing protocol used for the tumour samples [43].

After post-variant calling filtering, variant functional annotation starts with the prediction of the functional consequence of the variant on gene transcripts and protein sequence (e.g., silent, missense, or nonsense). Therefore, the raw variant calls in the VCF file are enriched with the information provided by annotation tools like ANNOVAR [61] and the Ensembl Variant Effect Predictor (VEP) [62]. Predicting the functional consequence of somatic variants on protein function is challenging and heavily depends on the transcript database used [e.g., National Center for Biotechnology Information (NCBI) Reference Sequence Database (RefSeq), Ensembl, University of California Santa Cruz (UCSC) or GENCODE]. In addition, a single variant may have different effects depending on the transcript isoform it affects. Since the experimental validation of the consequences computed by these annotation tools is not feasible for all the detected variants, variant annotation also relies on computational predictions from in-silico algorithms. Tools such as SIFT [63], PolyPhen-2 [64], MutationTaster [65] and CADD [66], integrate evolutionary conservation, biochemical properties, and structural data to estimate the deleterious potential of coding variants. In addition to these computational predictions, variant annotation also incorporates the collection of biomedical information from databases like the Catalogue Of Somatic Mutations In Cancer (COSMIC) [67], ClinVar [68], the Oncology Knowledge Base (OncoKB) [69], and the Clinical Interpretation of Variants in Cancer (CIViC) [70]. These resources provide information on previously reported cancer-associated or benign variants and often refer to other knowledgebases or literature, thus favouring the extrapolation of potential oncogenic mechanisms, therapeutic targets, and clinical outcomes.

Similarly to post-variant calling filtering, there is no consensus on how to perform variant annotation; some laboratories treat variant annotation as a prioritisation task, ranking variants by their likely clinical relevance, whereas others rely on curated gene lists to restrict interpretation to predefined variants [34]. The retrieval and integration of up-to-date biomedical knowledge from such diverse sources, including scientific publications, clinical trial registries, and real-world data, is a complex task, as it requires continuous maintenance to keep annotation data sources relevant and up-to-date, given the rapid pace of genomic research [71]. Due to these limitations, variant annotation remains heterogeneous across evaluators from different locations and largely manually performed [72, 73], representing a critical step in the clinical interpretation of results

coming from NGS data.

The final steps of tertiary analysis are variant interpretation and classification, where all the collected information about the biological and clinical relevance of each alteration are evaluated, with the final goal of assigning the somatic variant to a level of oncogenicity. Oncogenicity classification defines the potential of the variant to drive tumour initiation and progression [74]. To standardise the oncogenicity classification of somatic variants, three internationally-recognised consortia jointly published a set of guidelines [30]. These guidelines provide structured criteria for categorising somatic variants into five classes, ranging from class "Oncogenic", which denotes variants of strongest oncogenic potential, to class "Benign", which includes benign variants. Within this framework, each criterion is supported by the collected evidence and contributes to a point-based scoring system that determines the variant final classification into one of the five classes. Since each criterion must be carefully evaluated for the individual variant, the manual application of these guidelines requires the integration of different resources for each piece of evidence, thus making the adoption of the proposed framework time-consuming and non-trivial. To address these challenges, one of the major contributions of this PhD work consisted in the implementation of the automated evaluation of these guidelines in a novel tool called OncoVI. The oncogenicity guidelines together with the tool will be extensively described in Chapter 4.

2.3 Tumour molecular characterisation in Molecular Tumour Boards

All the results produced by the bioinformatics pipeline described above must ultimately be interpreted within a clinical context, to identify possible effective therapeutic options for patients. In this translational process, MTBs play a pivotal role, serving as the backbone where complex genomic data are integrated with clinical, pathological, and therapeutic considerations [75].

MTBs have emerged as interdisciplinary clinical frameworks for the interpretation of molecular data generated by NGS and other high-throughput assays. MTBs typically include oncologists, molecular pathologists, clinical bioinformaticians, and genetic counsellors who collaboratively review each patient's molecular profile considering the patient's clinical history, histopathology, and available treatment options. The primary objective of MTBs is to translate molecular findings into actionable therapeutic strategies, including targeted therapies, immunotherapies, and enrolment in genomically-guided clinical trials. In some cases, profiling may also suggest the use of off-label treatments, i.e., therapies

approved for a different tumour entity from the one under evaluation. The access to such treatment is subject to regional and national guidelines, which vary in flexibility and scope [19].

2.3.1 The Erlangen Molecular Tumour Board

At the Institute of Pathology UKER, the MTB takes place weekly in person. Requests for genomic testing can be submitted either by oncologists within the Erlangen MTB or by external oncologists. Patients are considered eligible for tumour molecular profiling if they meet one of the following criteria: (i) advanced tumour disease with expected exhaustion of standard therapy within six months; (ii) cancer of unknown primary (CUP) syndrome; or (iii) a tumour with an atypical course or the presence of multiple malignancies [76].

For tumour molecular profiling using targeted NGS, the required material is the most recent FFPE tissue sample. Tumour tissue slices are cut from the FFPE block for histopathological evaluation and NGS analysis. Those selected for histopathological evaluation undergo haematoxylin and eosin (H&E) staining, a routine histological technique that uses haematoxylin to stain cell nuclei blue, and eosin to stain the cytoplasm and extracellular matrix pink. This step allows pathologists to evaluate tumour morphology and estimate tumour cell content before proceeding with the NGS analysis. From the remaining unstained slide, DNA is extracted and processed with the Illumina TruSight Oncology 500 HRD (TSO500+HRD, Illumina, Inc., San Diego, CA, USA) cancer gene panel, following the manufacturer's instructions (`https://support.illumina.com.cn/content/dam/illumina-marketing/apac/china/documents/tso500-hrd-data-sheet-m-gl-00748.PDF`). Library sequencing is performed in-house on an Illumina NextSeq500/550 sequencer. FASTQ file generation, alignment to hg19 (UCSC) human reference genome, and SNVs calling are performed on the Illumina Connected Analytics (ICA) cloud environment.

Subsequently, variant annotation and classification are currently carried out using the tools developed within this PhD work and described in detail in Chapter 3 and Chapter 4. The generation of the clinical report is performed by experts specialised in cancer biology, prior to patient's discussion in the interdisciplinary MTB. In this context, the variants identified in the patient's tumour are stratified for their actionability, i.e., their potential to guide clinical decision-making, according to national evidence levels defined by the Nationales Centrum für Tumorerkrankungen (NCT) and the Deutsches Konsortium für Translationale Krebsforschung (DKTK). Accordingly, variants are assigned to one of four categories: M1 (biomarker validated for the same tumour entity), M2 (biomarker validated in another tumour entity), M3 (biomarker supported by preclinical in vitro or in vivo data), and M4 (biomarker supported by biological rationale).

Chapter 3

A bioinformatics framework to support variant interpretation in clinical oncology

This Chapter is based on a manuscript recently submitted to a journal, with the following authors and title:

Maria Giulia Carta, Lars Tögel, Miriam Angeloni, Marie Sieger, Christian Fiebig, Sven Wach, Helge Taubert, Norbert Meidenbauer, Silvia Spoerl, Arndt Hartmann, Bernd Wullich, Florian Haller, Markus Eckstein & Fulvia Ferrazzi. *Landscape of actionable genetic alterations in advanced urothelial carcinoma: high prevalence but limited clinical use.*

In the previous Chapter (section 2.2.3), the importance of post-variant calling filtering and variant functional annotation for accurate variant interpretation was introduced, along with their main challenges. These processes are traditionally labour-intensive, due to the multitude and heterogeneity of the resources to be collected. In addition, as they are predominantly manually curated, they lack reproducibility. These limitations highlight the need for consistent bioinformatics approaches to effectively support the interpretation of somatic variants in cancer. To address this challenge, in the context of this PhD work, a Python-based bioinformatics framework was developed within the MTB of the Institute of Pathology UKER. The framework was designed to assist biologists in post-variant calling filtering and variant functional annotation of genomic data from tumour molecular profiling. It is now routinely applied within the MTB, contributing to evidence-based interpretation of somatic variants in cancer.

This Chapter presents the design of the developed bioinformatics framework and the results of its comprehensive evaluation on a cohort of advanced urothelial carcinoma patients, which underwent tumour molecular profiling retrospectively.

3.1 Methods and patient cohorts

3.1.1 Implementation of the bioinformatics framework

To assist biologists in post-variant calling filtering and variant functional annotation, a dedicated bioinformatics framework was developed in Python (version 3.8.8). To systematically process the VCF files from tumour molecular profiling data, the pipeline was executed within a dedicated conda environment (conda version 24.11.1), on a remote server based on Ubuntu 20.04.6 LTS operating system, using as input file the "MergedSmallVariants.genome.vcf", which contains genomic coordinates of SNVs and indels identified in each patient's sample (**Figure 6**).

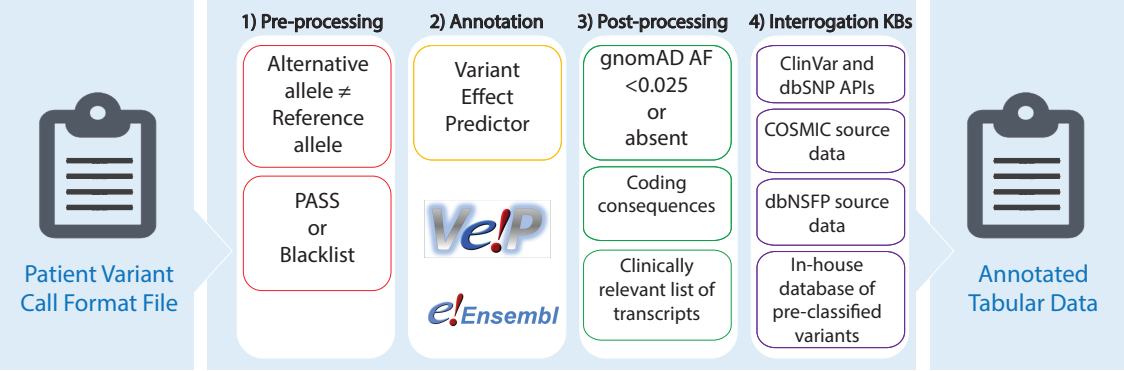

Figure 6: Main steps of the developed bioinformatics framework. The bioinformatics framework is based on the Variant Effect Predictor (VEP) and takes in input a single variant call format (VCF) file to perform variant annotation and knowledgebases (KBs) interrogation, in a multi-step workflow: 1) VCF pre-processing, 2) VEP-based functional annotation of the variants, 3) post-processing based on VEP-annotated functional consequences and gnomAD_AF population allele frequency (AF), 4) interrogation of multiple databases. The pipeline cross-references variants with an in-house database of pre-classified variants according to their effect on protein function and provides as output all the annotated variants in a tabular format.

Starting from a single VCF file containing the detected genomic variants in the patient, the framework integrates multiple steps: 1) VCF pre-processing, 2) variant functional annotation with the VEP tool, 3) VCF post-processing, and 4) automated knowledgebases interrogation.

In the first step, i.e., VCF pre-processing, variants are retained only if the alternate allele differs from the reference sequence and the "FILTER" status is either "PASS" or "Blacklist". This step is implemented using the filter_vcf function from the VcfFilterPy package (`https://github.com/superDross/VcfFilterPy`). The second step, i.e., variant functional annotation, is performed with the VEP of Ensembl (version 104.3) [62] to predict the functional consequences of nucleotide changes at the protein level. GRCh37 is used as reference genome and RefSeq [77] serves as transcript annotation cache. Selected VEP plugins include: 1) the nonsense-mediated mRNA decay (NMD) plugin to flag vari-

ants escaping the NMD, and 2) the database of non-synonymous functional predictions (dbNSFP) [78] plugin (version 4.3a) to integrate in silico functional prediction scores.

Following annotation, variants are filtered using the filter_vep utility from the VEP package. Retained variants meet at least one of the following criteria: (i) annotated consequence by VEP, or (ii) minor allele frequency (MAF) <0.025 reported in gnomAD [79]. In addition, only the variants predicted to negatively affect protein function are further kept, based on a curated list of predicted functional consequences. In the case of variants overlapping multiple transcript isoforms, the variant associated with the clinically relevant transcript, based on a curated transcripts list, is selected if available. Otherwise, the first transcript reported by VEP is used.

3.1.2 Knowledgebase interrogation

In the final step, the pipeline retrieves biological and clinical information for each variant by interrogating multiple curated knowledgebases. Data from ClinVar [68] and dbSNP [56] databases are accessed through their respective APIs. The information from ClinVar include the aggregated pathogenicity classification, review status, and number of submissions, while the corresponding "rs" identifier is obtained from dbSNP. Information from COSMIC [80] (CosmicMutantExport.tsv dataset, v. 95) include the FATHMM scores [81], the number of reported samples, and associated tissue types.

In addition, selected predictive scores are extracted from dbNSFP and the pipeline calculates an average score across these predictions. The final output consists of a tabular file integrating all collected evidence for each variant filtered at the previous step. Variants are then cross-referenced against a curated in-house database of previously characterised variants to determine, where applicable, their established classification according to the variant effect on protein function.

3.1.3 Variant assessment

The output generated by the developed bioinformatics framework is utilised by a molecular biologist to evaluate each variant with respect to its potential impact on wildtype (WT) protein function, allowing the categorisation into one of six classes: "Pathogenic", "Likely Pathogenic", "Variant of Unknown Significance (VUS)", "Likely Benign", "Benign", and "Artefact". This classification process is supplemented by manual interrogation of three external databases: the Molecular Tumor Board Portal (MTBP) database [82], varSEAK (`https://varseak.bio/`), and the Leiden Open Variation Database (LOVD) [83], in order to achieve a consensus on the classification.

The VCF file produced by filter_vep prior to knowledgebase interrogation is uploaded to MTBP and the cancer type must be specified in order to assess the predictive relevance of each variant. MTBP aggregates evidence from multiple sources, including ClinVar, BRCA-Exchange [84], OncoKB [69], and CIViC [70], to classify variants as "putative functionally relevant variants", "unclassified", or "putative functionally neutral variants". A supplementary text file listing variants identified as SNPs is also generated. In the final report, variants present in this list were labelled as "Polymorphism", whereas those with evidence of functional relevance were tagged as "MOL-TB Portal"; all others remained untagged.

Splicing-related variants, i.e., variants predicted by VEP to affect pre-mRNA splicing and annotated as "splice_region_variant", "splice_donor_variant", or "splice_acceptor_variant", are evaluated using varSEAK. For each variant, the gene symbol, cDNA change, and transcript are submitted to the varSEAK website to obtain a predicted splicing effect, reported as "no splicing effect", "likely no splicing effect", "unknown splicing effect", "likely splicing effect", or "splicing effect". "splice_donor_variant" and "splice_acceptor_variant" variants not processed through varSEAK are classified as "Likely Pathogenic" based on their potential to disrupt protein function, despite the absence of corroborating evidence. LOVD links generated by the bioinformatics framework are used to retrieve clinical classification and variant entry count for integration into the patient report.

Final classifications of variant effect on protein function are assigned using our in-house database, when available; otherwise, results from the resources extracted both automatically and manually are used.

- Variants annotated by VEP as "stop_gained" or "frameshift_variant" that met any of the following conditions:

  1. "FILTER" equal to "Blacklist" and/or;
  2. occurrence in the final two exons of the protein and/or;
  3. ClinVar or MTBP classification as "Likely Benign" or "Benign";

  are classified as "VUS".

- Variants with the same annotation by VEP but that did not meet any of the three abovementioned conditions are classified as "Likely Pathogenic".

- For variants annotated differently and with a ClinVar classification (without contradictory evidence from other databases), the ClinVar classification is adopted.

- In absence of ClinVar data, the classification is based on the average dbNSFP score:

1. Values < 0.3 are deemed "Likely Benign";
2. Values ≥ 0.3 are deemed "VUS" when no conflicting information is present.

- In cases where evidence from multiple sources is deemed contradictory, the classification is "VUS".

3.1.4 Patient cohorts

Two distinct cohorts of patients with muscle-invasive bladder cancer (MIBC) and metastatic urothelial carcinoma (mUC) disease were utilised to systematically assess the effectiveness of the developed bioinformatics framework in supporting biologists in variant interpretation:

- The "Muscle-Invasive ERlangen (MIER) cohort" (MIER cohort): the first cohort consists of 242 consecutive patients with MIBC/mUC who underwent radical cystectomy with lymphadenectomy followed by adjuvant chemotherapy with curative intent. None of the patients received neoadjuvant chemotherapy. Detailed clinical and pathological characteristics of this cohort have been previously reported [85–87].

- The "Molecular Tumour Board cohort (MTB cohort)": the second cohort comprises 41 patients with MIBC/mUC who were referred to the MTB of the Institute of Pathology UKER between 2019 and 2023. These patients underwent tumour genetic counselling with the aim of implementing personalised off-label treatment recommendations.

3.1.5 Tumour molecular profiling via targeted DNA-sequencing

For both cohorts, DNA was isolated from FFPE tumour tissue at the Institute of Pathology UKER. For the MIER cohort: DNA library preparation and sequencing were performed by IMGM Laboratories GmbH using the TruSight Oncology 500 (TSO500, Illumina, Inc., San Diego, CA, USA) assay, a targeted sequencing panel covering 523 cancer-associated genes. In brief, genomic DNA (gDNA) was extracted and purified via hybrid-capture chemistry with streptavidin-coated magnetic beads. Sequencing was carried out on the NovaSeq 6000 NGS sequencing platform (Illumina). For the MTB cohort: DNA library preparation and sequencing were conducted at the Institute of Pathology UKER on the Illumina NextSeq500/550 sequencing platform (Illumina). Of the 41 analysed tumour samples, 22 were processed with the TruSight Tumor 170 (TST170, Illumina) assay, targeting 148 cancer-related genes, while the remaining 19 were analysed with the Illumina TSO500 assay.

For all samples, BCL files were processed with the Illumina TruSight Oncology 500 v2.2. Local App, which performs demultiplexing, FASTQ file generation, and alignment of DNA and RNA sequences to the human reference genome (UCSC, hg19). Variant calling was conducted for SNVs, CNAs, gene fusions, and splice variants. Genomic coordinates of the identified variants for each patient were exported in a VCF file named "MergedSmallVariants.genome.vcf".

For the samples analysed with the Illumina TSO500 assay, additional metrics were determined:

- Microsatellite instability (MSI): it reflects length alterations in short tandem repeats ("microsatellites"), typically consisting of 1–6 base pair motifs repeated consecutively (e.g., "CACACACACA..."). MSI arises from defects in the DNA mismatch repair (MMR) pathway and serves as a predictive biomarker for response to immunotherapy [88–90].

- Tumour mutational burden (TMB): it quantifies the number of somatic non-silent mutations per megabase of interrogated genomic sequence and is a predictive biomarker for response to immunotherapy [91].

3.1.6 Curated gene list of potentially actionable off-label genomic alterations

To identify potentially actionable off-label genomic alterations in the MIER cohort, a curated list of 24 genes with approved therapies from the Food and Drug Administration (FDA) (`https://www.fda.gov/`) was compiled. To this aim, only patients harbouring Pathogenic or Likely Pathogenic variants in at least one of this genes list were considered. The same list was utilised also for the comparison of the potentially actionable off-label genomic alterations between MIER and MTB cohorts. The druggable genes were further categorised into functional or mechanistic groups, namely "DNA repair", "MAPK Signalling", "PI3K/mTOR Signalling", "Oestrogen Signalling", and "Receptor Tyrosine Kinases".

Oncoprints representing the mutational spectrums of the two analysed cohorts were generated with the R package ComplexHeatmap (version 2.13.1), whereas bar plots showing immune biomarkers distribution stratified by *FGFR3* alteration status (altered vs. WT) were created using the R packages ggplot2 (version 3.5.1) and ggpubr (version 0.6.0).

3.2 Results

3.2.1 Assessment of hands-on time for variant interpretation

In addition to its adoption within the MTB of the Institute of Pathology UKER for selected routine molecular diagnostics patients, the developed bioinformatics framework was systematically evaluated on an in-house curated cohort of 242 patients diagnosed with advanced urothelial carcinoma (MIER cohort). Tumour samples underwent retrospective tumour molecular profiling with the TSO500 gene panel (Illumina) and were processed as described in the previous sections. NGS data obtained from tumour profiling were complemented by the collection of clinical and pathological characteristics of the cohort. First, the VCFs named "MergedSmallVariants.genome.vcf" of the MIER patients were processed through the developed bioinformatics framework. Then, the pipeline output, containing all the variants of the VCF file that passed the applied filters, were interpreted by molecular biologists and pathologists according to the criteria described in the previous sections. An output was produced by the bioinformatics framework for 97% (n=235/242) of the patients, since the VCFs of seven patients did not contain any PASS variants.

The hard filters applied by the pipeline in the pre- and post-processing steps (steps 1 and 3, **Figure 6**) decreased the number of variants to be further interpreted by the expert biologist from a median value of 1,273 annotated variants per patient to a median value of 50 coding variants per patient (**Figure 7a**).

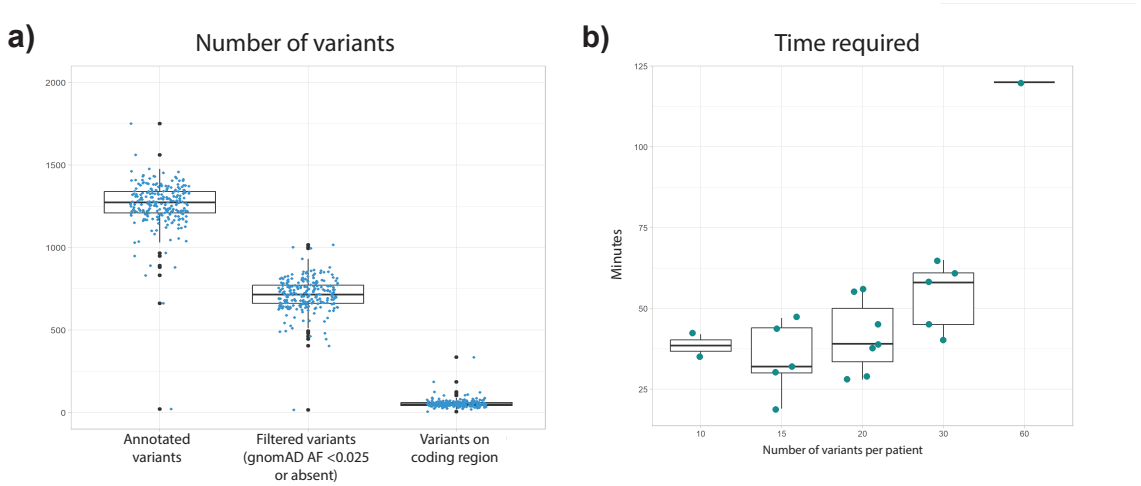

Figure 7: Assessment of the bioinformatics framework effectiveness on the reduction of biologist hands-on time. a) Boxplots show the distribution of the number of variants after variant functional annotation (step 2, left), variants after gnomAD_AF filter in post-processing (step 3, middle), and variants after filters on the VEP-annotated coding functional consequences in post-processing (step 3, right). **b)** Time per patient required by an expert biologist to interpret the pipeline output (n = 20 patients). Boxplots show the distribution of the time (minutes) required to interpret data from patients with different numbers of variants.

In addition, when focusing on a subset of cases (n=20/242), we observed that the time required by the molecular biologist for variant interpretation per patient was less than 60 minutes when the patient had less than 30 variants (**Figure 7b**). The higher the number of variants to be interpreted, the higher the time consumption per patient. However, for patients with 60 variants to be interpreted the required time was about 120 minutes. Taken together, the use of the bioinformatics framework dramatically reduced the number of variants to be further interpreted by the biologist and the biological evidence retrieved by the pipeline substantially lowered the time dedicated by the biologist to the process of variant interpretation, which according to the previous assessment without pipeline use was in the range of 1 to 2 hours per case.

3.2.2 Actionable in-label genomic alterations in the MIER cohort

To identify patients that could potentially benefit from therapies targeting validated markers in urothelial bladder cancer, we first evaluated the mutational spectrum of patients harbouring Pathogenic/Likely Pathogenic variants in the top 15 frequently mutated genes stratified by the *FGFR3* alteration status (**Figure 8a**).

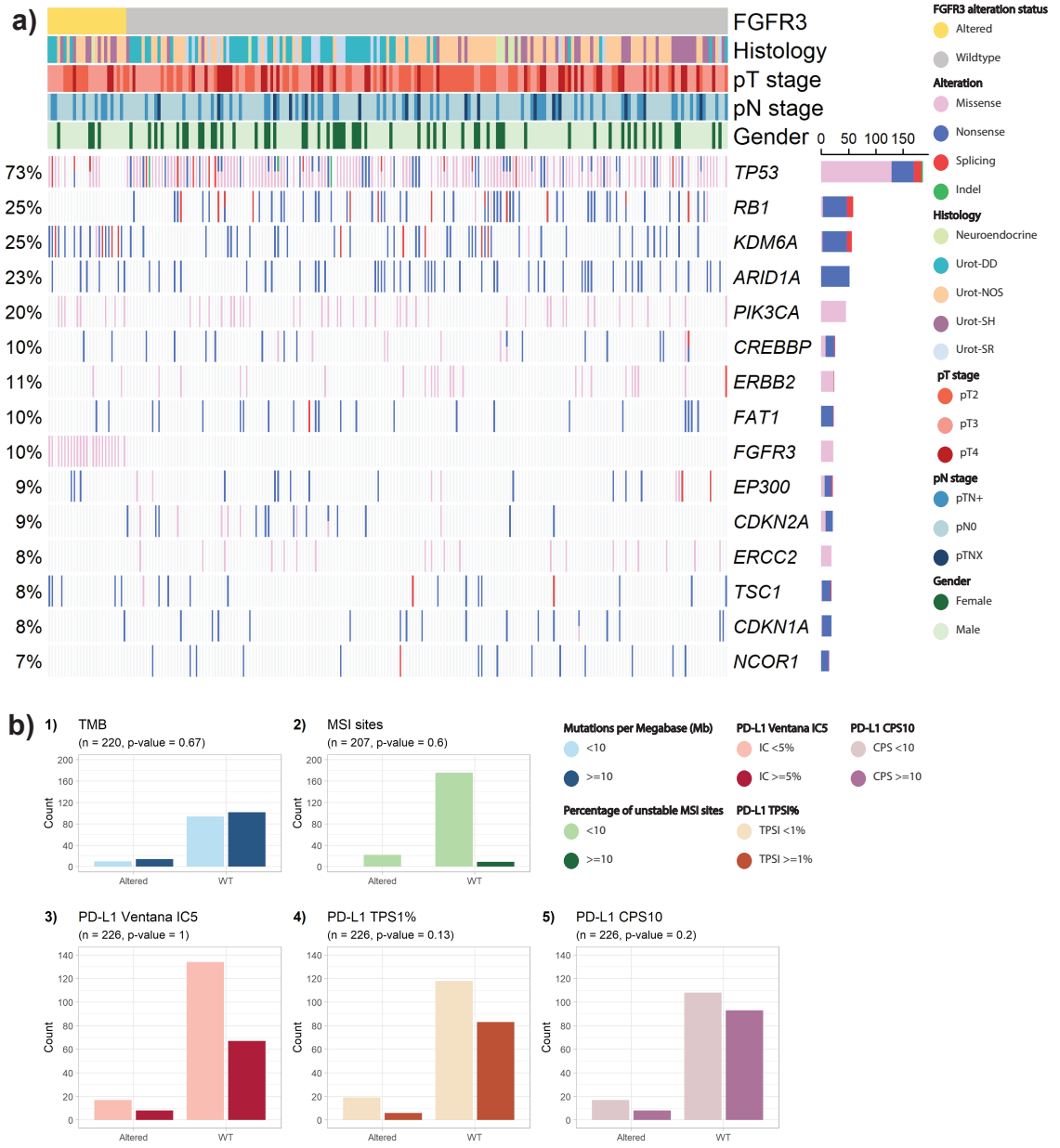

Figure 8: Spectrum of potentially actionable variants in the MIER cohort (n=226). a) Oncoprint showing the VEP-annotated consequences of the Pathogenic/Likely Pathogenic variants identified. Rows represent genes ordered according to decreasing frequency. Columns represent patients with a *FGFR3* alteration status equal to altered or wildtype (WT). **b)** Distribution of immune biomarkers stratified by *FGFR3* alteration status (altered vs. wildtype). TMB: Tumour mutational burden, MSI: microsatellite instability, IC score: immune cells score, CPS: combined positive score.

A subset of n=26/226 (12%) patients harboured at least one actionable *FGFR3* gene alteration indicating potential responsivity to erdafitinib, an approved therapy for *FGFR3* altered urothelial carcinomas. All *FGFR3* mutations identified in the MIER cohort by NGS were confirmed with two different PCR-based methods, i.e., the TheraScreen (mutations, fusions) and the SnapShot multiplex PCR (mutations), with a concordance rate

of 100%. The majority of the *FGFR3*-mutated cases (n=22/26) showed variants in well-known hotspot regions (S249, R248) of the protein structure [92]. Notably, only one of the 22 *FGFR3*-altered cases also harboured the *FGFR3::TACC3* gene fusion in addition to the variant. The remaining three *FGFR3*-altered samples exhibited the *FGFR3::TACC3* gene fusion as the sole alteration.

In second instance, we investigated the association between *FGFR3* mutations and variants of clinical relevance in other frequently mutated genes. We observed mutual exclusivity between the Pathogenic/Likely Pathogenic mutations in *FGFR3* and all the 15 frequently mutated genes, except for *FAT1*. In addition, we observed that 50% (n=13/26) of the *FGFR3*-altered patients had variants in *KDM6A*. The association of the *FGFR3* alteration status with the immune biomarkers qualifying for treatment with immunotherapy was evaluated (**Figure 8b**). No significant association was identified between *FGFR3*-alteration status (altered vs. WT) and TMB-high, MSI-high, or PD-L1 positive tumour statuses.

3.2.3 Potentially actionable off-label genomic alterations in the MIER cohort

To explore the potential responsiveness of MIER cohort patients to novel therapeutic targets, we focused on 24 selected genes implicated in key pathways commonly dysregulated in bladder cancer, including DNA repair, MAPK, PIK3/mTOR, estrogen signalling pathways, and receptor tyrosine kinases (**Table 3**).

Table 3: **Genes per pathway investigated for potentially off-label genomic alterations in the MIER cohort.**

DNA Damage Response	ATM, CHEK2, FANCA, BRCA1/2, BRIP1, CDK12, FANCI, PALB2, BARD1, CHEK1, NBN
MAPK signalling	KRAS, NRAS, BRAF, MAP2K1
PI3K/mTOR signalling	PIK3CA, MTOR
Estrogen signalling	ESR1
Receptor tyrosine kinases	EGFR, MET, KIT, FLT3, ALK

Within this gene subset, potential therapeutic targets were identified in 41% of the patients (n=92/226), with the most frequently mutated gene being the kinase *PIK3CA* (19.9%) (**Figure 9**).

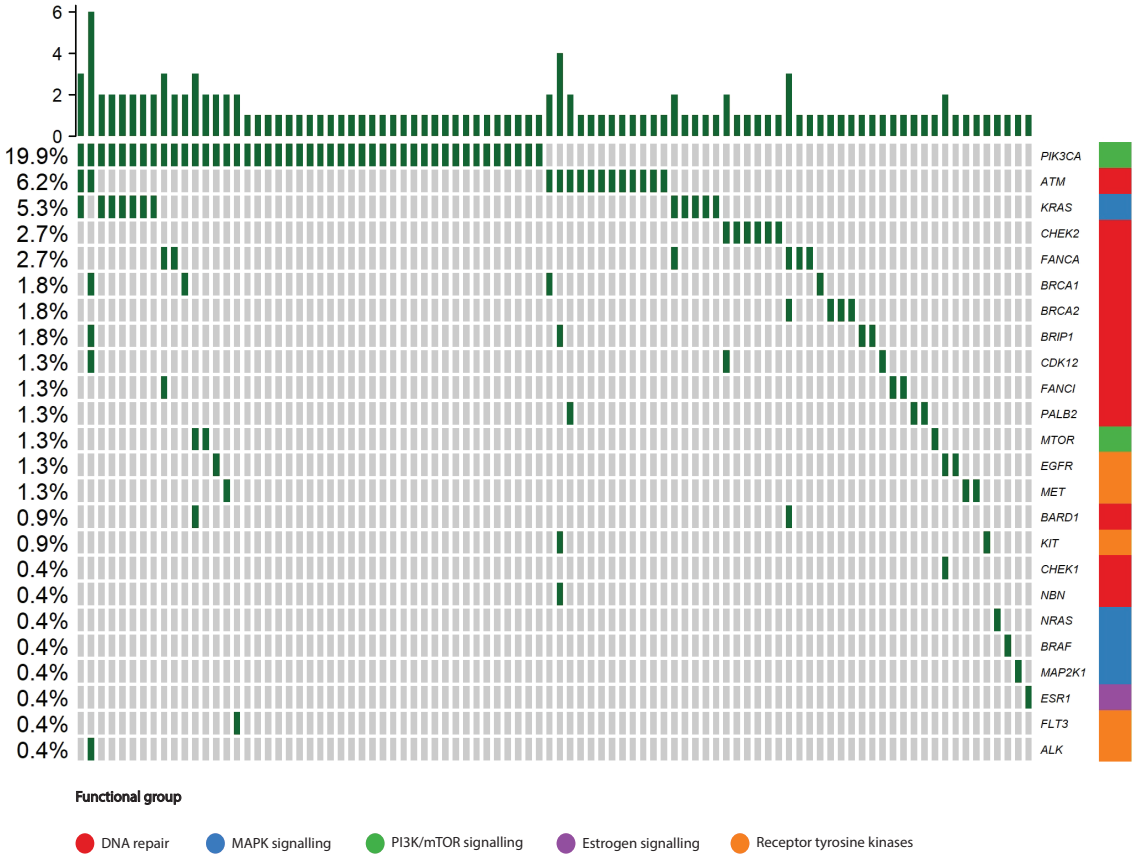

Figure 9: Mutational spectrum of MIER patients in potentially actionable off-label genes. Oncoprint showing the Pathogenic/Likely Pathogenic variants identified in patients of the Muscle-Invasive Erlangen (MIER) cohort (n=226) in potentially actionable off-label genes. Rows represent genes ordered according to decreasing frequency. Columns represent patients. MAPK: mitogen-activated protein kinases. PI3K/mTOR: Phosphatidylinositol 3-kinase /mammalian target of rapamycin.

Alterations in the DNA Damage Response (DDR) pathway were observed in 40 patients (18%, n=40/226) of the MIER cohort. The most frequently mutated gene was *ATM* (6%, n=14/226), with nonsense mutations accounting for the majority of the mutated cases (57%). *BRCA1* (2%, n = 4/226), *BRCA2* (2%, n = 4/226) and *BRIP1* (2%, n = 4/226) were mutated at similar frequencies to each other and showed predominantly nonsense mutations. Such mutations may confer increased sensitivity to PARP inhibitors (PARPi), which exploit synthetic lethality by blocking single-strand DNA repair in homologous recombination-deficient tumours.

Within the MAPK signalling pathway, 15 patients carried mutations (7%, n=15/226), most commonly in *KRAS* (5%, n=12/226). Patients with *KRAS* mutations may represent a subgroup potentially responsive to targeted inhibitors of this pathway.

Mutations in the PI3K/mTOR pathway were found in 46 patients (20%, n=46/226). *PIK3CA* was the most frequently altered gene overall (19.9%, 45/226), with exclusively missense variants. *MTOR* mutations were less common (1%, n=3/226), including both

missense and nonsense variants. Patients harbouring mutations in these genes may be potentially responsive to PI(3)K inhibitors (*PIK3CA*) or mTOR inhibitors (*MTOR*), respectively.

Additional, less frequent mutations were detected in other genes driving tumour development. One patient harboured a splicing mutation in *ESR1*. Finally, mutations in receptor tyrosine kinases were identified in 10 patients (4%, n=10/226), with *EGFR* and *MET* altered each in three cases. *EGFR* acted predominantly via missense mutations, while *MET* via spicing mechanism.

3.2.4 Mutational agreement with the real-world MTB cohort

To assess the clinical relevance of the potentially actionable variants in the MIER cohort, we compared its mutational spectrum with that of 41 mUC patients analysed in the context of Erlangen MTB (MTB cohort). Nineteen of these samples were analysed using the same targeted sequencing approach as the MIER cohort, while 22 were profiled with a smaller gene panel as part of the routine diagnostics. Patients were discussed in the MTB of the Institute of Pathology UKER between 2019 and 2023, with recommended targeted therapies and available follow-up data. Of the 15 most frequently mutated genes in the MIER cohort, nine (*TP53, RB1, PIK3CA, KDM6A, ARID1A, EP300, TSC1, FGFR3,* and *CREBBP*) were also found in the MTB cohort (**Figure 10**). The genes *FBXW7, JAK2, KRAS, MSH3, PTEN, STAG2* were among the top 15 frequently mutated genes exclusively in the MTB cohort.

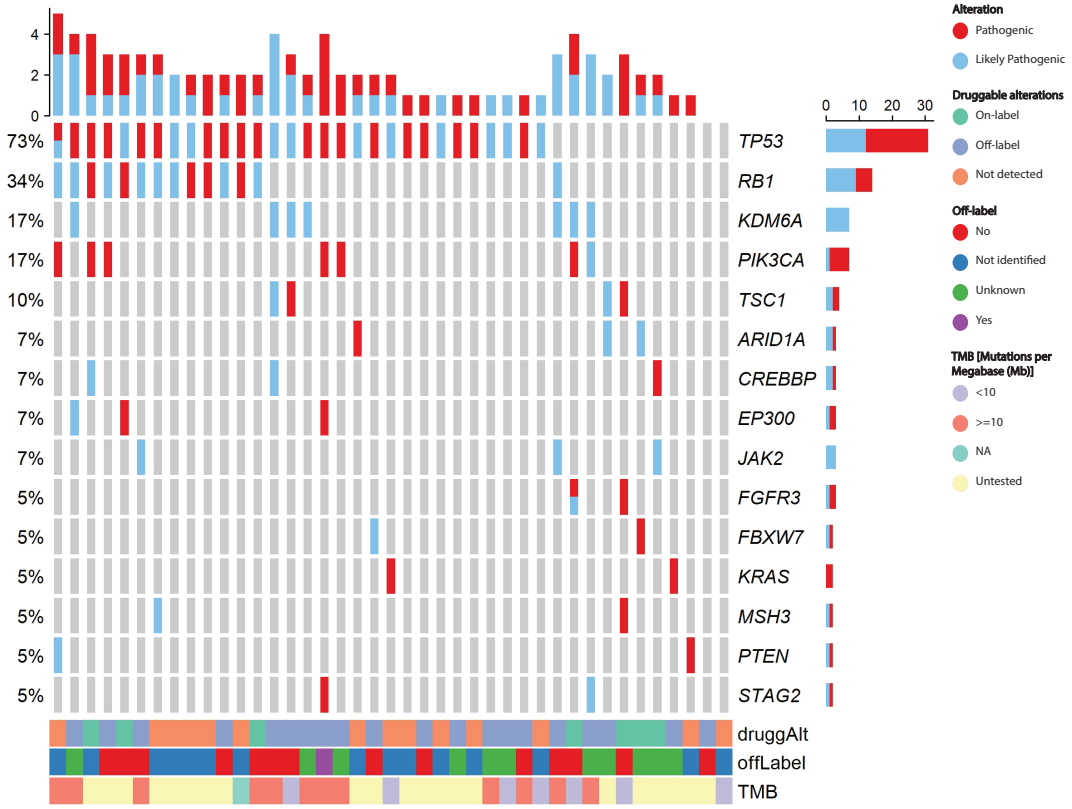

Figure 10: Mutational spectrum of the MTB cohort. Oncoprint showing the Pathogenic/Likely Pathogenic variants of the patients in the Molecular Tumour Board (MTB) cohort (n=41). Rows represent genes ordered according to decreasing frequency. Columns represent patients. druggAlt: druggable alterations, offLabel: off label therapy implemented for the patient, TMB: tumour mutational burden.

Mutational frequency of the top frequently mutated genes was comparable in the two cohorts, with *TP53, RB1, KDM6A, PIK3CA*, and *TSC1* mutated in 73%, 34%, 17%, 17%, and 10% of the MTB cohort patients, respectively. In the MIER cohort *ARID1A, ERBB2,* and *FGFR3*, had higher mutation frequencies compared to the MTB cohort. The assessment of the immune biomarkers in the MTB cohort was limited to 46% of the patients (n=19/41) and revealed a higher prevalence of TMB-high patients (63%, n=12/19) compared to the MIER cohort.

Actionable molecular alterations were identified in 66% (n=27/41) of MTB patients: 17% (n=7/41) with in-label and 49% (n=20/41) with off-label indications (**Table 4**).

Table 4: **Comparison of actionable alterations in MIER versus MTB cohorts.** HRD: homologous recombination deficiency.

Gene	**Type of alteration**	**MTB (n=41)**	**MIER (n=226)**	**Clinical Relevance**	**Treatment Indication**
FGFR3	Mutation	2 (5%)	22 (10%)	In-label	erdafitinib
ERBB2	Gene amplification	5 (12%)	35 (15%)	Off-label	HER2 inhibitors
FGFR1	Gene amplification	3 (7%)	11 (5%)	Off-label	erdafitinib / FGFR inhibitors
BRCA2	Mutation	1 (2%)	4 (2%)	Off-label	PARP inhibitors
RAD51D	Mutation	1 (2%)	0 (0%)	Off-label	(HRD) PARP inhibitors
BRIP1	Mutation	1 (2%)	4 (2%)	Off-label	(HRD) PARP inhibitors
PIK3CA	Mutation	3 (7%)	45 (20%)	Off-label	PI3K inhibitors
NF1	Mutation	1 (2%)	0 (0%)	Off-label	MEK inhibitors
TSC1	Mutation	4 (10%)	0 (0%)	Off-label	mTOR inhibitors
EGFR	Amplification	2 (5%)	10 (4%)	Off-label	EGFR inhibitors
KRAS	Mutation	2 (5%)	12 (5%)	Off-label	KRAS inhibitors

Two patients with *FGFR3* alterations had an in-label indication for erdafitinib, while three (7% n=3/41) with *FGFR1* gene amplifications received off-label recommendations for erdafitinib and FGFR-inhibitors; yet none of these recommendations were implemented. Similarly to the MIER cohort, *BRCA2* mutations were rare (2%, n=1/41) in MTB patients, occurring in only one patient, which was given a recommendation for PARPi therapy; though the outcome remains unknown. Additional HRD-related alterations (i.e., *RAD51D, BRIP1* mutation and *ATM* gene deletion) were identified in three patients (7%, n=3/41), and all of them received off-label indications for PARPi. However, analogously to the *BRCA2*-mutated MTB patient, it remains unknown whether the

therapy has been implemented by the treating oncologist.

ERBB2 gene amplifications in MTB patients (12%, n=5/41) had a frequency comparable to that in MIER patients (15%, n=35/226) and in some cases co-occurred with *RAF1* fusion or *PIK3CA* mutation, leading to combined recommendation for PIK3CA and HER2-inhibitors. One patient who received alpelisib for a *PIK3CA/ERBB2* co-alteration achieved an eight-month response, highlighting the potential benefit of MTB-guided therapy. Alterations in the PI3K/mTOR signalling pathway (**Table 3**) were at a comparable frequency in the MTB patients (17%, n=7/41) compared to the MIER cohort. Three of these seven patients with alterations in *PIK3CA* had co-occurrent *FGFR3* or *ERBB2* alterations, and as described above, received off-label indications for the respective findings. Of the remaining four patients, two received off-label recommendations for PIK3CA inhibitors, the other two did not receive any recommendation. Four additional patients with mutations in distinct genes, i.e., *NF1, TSC1, EGFR* and *KRAS*, received off-label recommendations for MEK-, mTOR-, EGFR-, and KRAS-inhibitors, respectively. However, similarly to the majority of the MTB patients, the implementation of these indications in the clinical practice remains unknown or impracticable.

Overall, while a substantial proportion of MTB patients presented actionable alterations, the actual implementation of off-label therapies was limited, highlighting a persistent gap between molecular findings and clinical application. Nevertheless, the case where off-label recommendation was applied demonstrates that MTB-guided off-label therapy can yield tangible clinical benefit.

3.3 Final considerations

This Chapter presented a bioinformatics framework to support biologists in the challenges associated with the process of variant interpretation in the context of tumour molecular profiling. The systematic evaluation of this framework on our in-house cohort of urothelial carcinoma patients demonstrated its effectiveness in streamlining critical steps such as variant functional annotation and knowledgebases interrogation.

By integrating robust filtering steps and annotation through VEP, the bioinformatics framework reduced the complexity of variant interpretation from many annotated variants per patient to only a few tens of coding variants. This reduction not only streamlined variant functional annotation and knowledgebases interrogation, but also made expert interpretation substantially faster, while allowing for a more standardised and reproducible process. Biologists and clinicians were thus provided with curated, biologically meaningful variant lists that could be readily assessed for therapeutic relevance, thereby bridging the gap between results from DNA sequencing data and clinical decision-making.

Beyond the evaluation in terms of time efficiency, the application of the framework to a homogeneous curated cohort of advanced urothelial carcinoma also generated relevant biological insights. In particular, results showed that tumour molecular profiling can uncover potential actionable alterations in advanced urothelial carcinoma, where the full implementation of precision oncology remains hindered by several factors. Indeed, despite the existence of publicly-available comprehensive molecular landscapes of MIBC/mUC, including The Cancer Genome Atlas (TCGA), these resources often lack a detailed annotation and expert clinical interpretation necessary for: 1) the precise identification of actionable alterations, and 2) patient-specific therapeutic recommendations beyond standard-of-care protocols [93–98]. Furthermore, few studies have addressed the actionable landscape of MIBC/mUC stemming from real-world data and cohorts [99]. In this context, the application of the developed bioinformatics framework to the MIER cohort enabled the identification of actionable alterations, most notably *FGFR3* and *ERBB2*, for which in-label therapies are available. Indeed, patients of the MTB cohort with analogous alterations received recommendations for the associated in-label treatments. In addition, several other potentially actionable alterations were detected, particularly in the RTK/RAS/PI3K pathway, and one patient of the MTB cohort with a PIK3CA/ERBB2 co-alteration achieved a meaningful response to alpelisib, underscoring the clinical potential of such findings.

However, the results also highlighted persistent barriers to clinical translation. Despite the ability of the developed framework to identify actionable alterations, the majority of MTB recommendations were not implemented, particularly when they involved off-label therapies. This limited adherence is not unique to our study. Indeed, other MTBs reported a low implementation rate due to systemic challenges such as regulatory, financial, and organizational constraints [100, 101].

The take-home message of this work is therefore twofold: first, bioinformatics frameworks can enable the efficient and reliable detection of actionable alterations, ensuring that clinically relevant information is made available; and second, the potential of MTBs to realise their full clinical value depends on consistent implementation of their recommendations. Without this, the benefits of precision oncology remain largely theoretical. Future research should thus aim not only to refine variant interpretation pipelines, but also focus on overcoming the translational gap between MTB discussions and real-world treatment decisions.

Chapter 4

OncoVI: a novel tool for oncogenicity variant interpretation

This Chapter is based on a manuscript currently under review in a journal, with an earlier version available as a preprint on medRxiv [102]:

Maria Giulia Carta, Lars Tögel, Annett Hölsken, Christoph Schubart, Heinrich Sticht, Robert Stöhr, Silvia Spoerl, Norbert Meidenbauer, Arndt Hartmann, Paolo Magni, Florian Haller, Fulvia Ferrazzi. *Oncogenicity Variant Interpreter (OncoVI): oncogenicity guidelines implementation to support somatic variants interpretation in precision oncology.* medRxiv. 2024.10.10.24315072. doi: `https://doi.org/10.1101/2024.10.10.24315072`.

In the introduction (section 2.2.3 and section 2.3), the importance of providing MTBs with accurate oncogenicity classification of somatic variants to guide therapy-decision making was discussed. Internationally-recognised guidelines, such as the ClinGen/CGC /VICC framework, have been established for this purpose [30]. However, their manual application is labour-intensive and prone to variability, leading to different interpretations and implementations across institutions. To promote a standardised and reproducible oncogenicity classification process, automated implementations are required. To address this need, during this PhD work the Oncogenicity Variant Interpreter (OncoVI), an open-source Python-based tool, was developed for the automated evaluation of the oncogenicity guidelines.

OncoVI was established within the MTB of the Institute of Pathology UKER and first validated using a gold-standard dat aset of variants, whose oncogenicity classifications were manually curated by the authors of the ClinGen/CGC/VICC guidelines. Subsequently, OncoVI was tested on a data set of more than 7,802 variants derived from

over 500 patients who underwent tumour molecular profiling in the MTB. This Chapter presents the implementation of OncoVI to support the harmonised oncogenicity classification of somatic variants in precision oncology and the results obtained from its application on both gold-standard and real-world variant data sets. OncoVI is currently maintained on GitHub at the following link: `https://github.com/MGCarta/oncovi`.

4.1 Background and motivation

In Chapter 2, we already highlighted that establishing precision oncology as routine practice requires two prerequisites: (i) mutation identification based on comprehensive molecular profiling of cancer cell DNA, and (ii) precise, reproducible classification of the oncogenicity of somatic variants in the context of the disease. Although identifying mutations is a widely consolidated practice in clinical settings [4, 5, 24–26, 103, 104], the classification remains an open challenge.

The interpretation of pathogenicity of a somatic variant with respect to neoplastic diseases (oncogenicity) defines its role in tumour initiation and progression [30]. Guidelines have been proposed to support the assessment of the variant clinical relevance in the somatic context [31, 32]. However, the harmonisation between laboratories remains challenging, since the guidelines are applied differently, resulting in conflicting classifications of variants [72, 73]. In section 2.2.3, we described how the interpretation of variant oncogenicity requires three fundamental steps: (i) functional annotation, (ii) collection of biomedical evidence, and (iii) final classification. The first step, usually relies on the use of tools such as Annovar [61], VEP [62], and SnpEff [105], which are able to predict the potential effect of variants on genes, transcripts and protein sequences. The collection of biomedical evidence requires the identification and integration of different resources, e.g., population data, biomedical literature, and in-silico predictions. Tools that integrate and display different data types in an intuitive way have been published, such as VarSome and cBio Cancer Genomics Portal (cBioPortal) [106, 107]. Yet, they leave the responsibility for the final interpretation to the user. On the one hand, to support the laborious task of interpreting the aggregated evidence for genomic variants in cancer, databases containing curated classifications, such as CIViC [70], OncoKB [69], and the Database of Curated Mutations [108] are available. However, these databases contain clinical interpretations predominantly limited to well-known and characterised variants [29]. On the other hand, to support the interpretation of somatic variants in precision oncology, tools that generate interactive reports have been developed, leveraging computational methods and publicly available knowledge, such as MTBP and the Cancer Genome Interpreter (CGI) [82, 109]. However, the variant assessment provided by these tools is not based on standardised

guidelines.

To increase the consistency of variant interpretation between different institutions as well as between different assessors, in 2022 three relevant consortia (i.e., ClinGen/CGC/-VICC) have published the so far only internationally-recognised guidelines for the classification of oncogenicity of somatic variants in cancer [30]. This Standard Operating Procedure (SOP) comprises 12 and five criteria for oncogenic and benign effect, respectively. For each somatic variant, all criteria are evaluated and, according to a point-based evaluation, a variant-specific score is calculated. According to this score the variant is classified into one of five possible classes, i.e., "Oncogenic", "Likely Oncogenic", "VUS", "Likely Benign", and "Benign". In the SOP all criteria are provided as textual indications only and for some criteria a careful interpretation by the user is essential, such as for criteria that refer to well-established *in vitro or in vivo* functional studies, supportive of an oncogenic or benign effect of the variant. Furthermore, the authors provide recommendations for the use of resources for some but not all criteria. To the best of our knowledge, the Personal Cancer Genome Reporter (PCGR) [110] is currently the only publicly-available tool that explicitly aims at automating the oncogenicity classification of somatic variants, based on these guidelines. However, PCGR does not implement all criteria, possibly also due to practical limitations related to data availability for criteria requiring direct experimental evidence of oncogenicity (e.g., OS2 based on *in vitro or in vivo* functional studies). As acknowledged by the PCGR developers, these limitations may lead to oncogenic variants being classified as "VUS". Furthermore, PCGR adopts modified scoring thresholds for classification compared to those originally proposed by Horak et al. It appears that the impact of these deviations on the final oncogenicity classification has not been systematically evaluated.

OncoVI takes a further step towards the harmonisation of the interpretation of somatic variants, providing a comprehensive automated evaluation of the oncogenicity guidelines based on the point-based system proposed by Horak et al. Indeed, OncoVI's implementation includes criteria that are missing in PCGR by adopting, for example, clinical significance annotations from ClinVar to support the OS2 criterion. OncoVI is a unique tool that incorporates the bioinformatics framework described in the previous Chapter (Chapter 2) to perform: (i) functional annotation of genomic alterations, and (ii) collection of the available evidence from the interrogated resources, by additionally (iii) evaluating each criterion and providing a final classification of oncogenicity. The implementation of the oncogenicity guidelines in OncoVI has been evaluated using both on gold-standard and real-world data sets of somatic variants.

4.2 Methods and variant data sets

4.2.1 Implementation of the oncogenicity criteria in OncoVI

All the criteria for oncogenic and benign effect as well as the point-based system for the classification of oncogenicity of somatic variants from the SOP proposed by Horak et al., [30] were implemented in OncoVI, except for the "Oncogenic Supporting-2" (OP2, 1 point, *"Somatic variant in a gene in a malignancy with a single genetic etiology"*) criterion, for which an appropriate resource could not be identified (**Figure 11**).

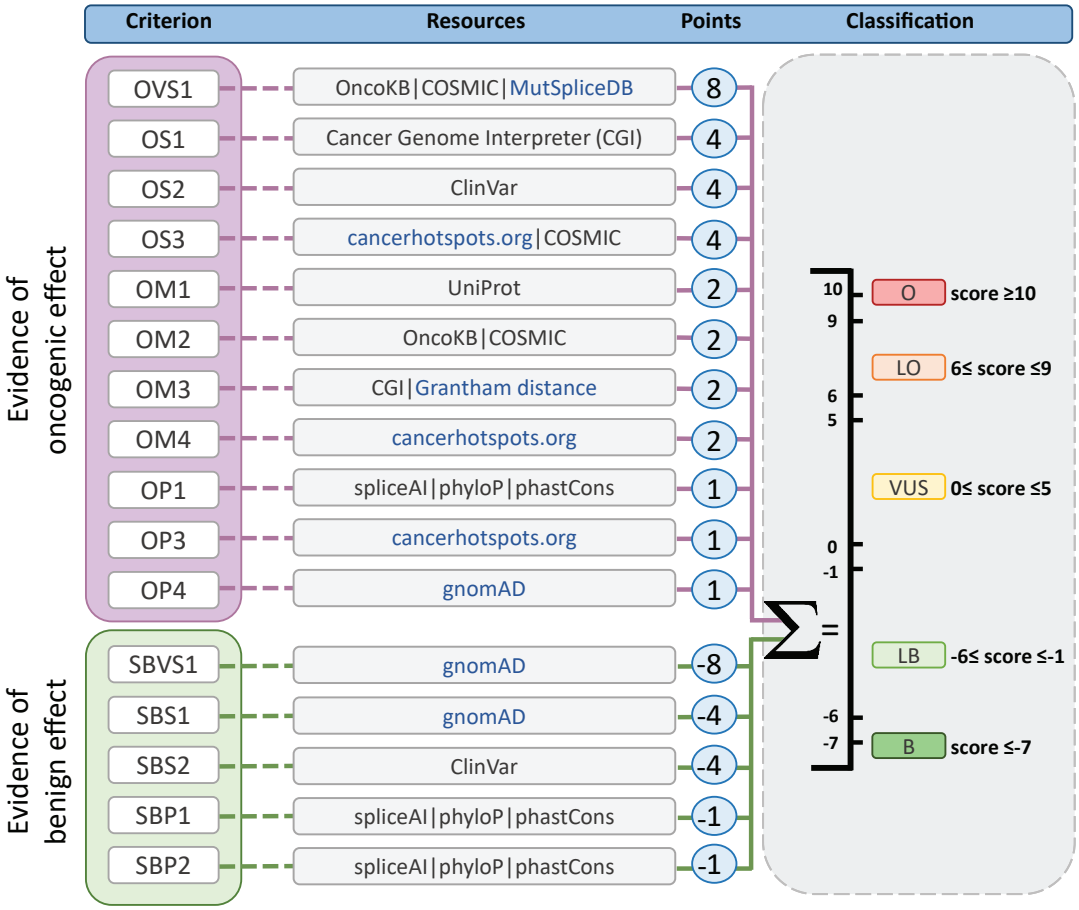

Figure 11: Workflow of OncoVI. The figure shows the implemented criteria in OncoVI (11 and five criteria for evidence of oncogenic and benign effect respectively), the public resources utilised to assess each criterion and the points associated with, and the classification of oncogenicity into one of five classes on the basis of the variant-specific score. This score is obtained as the sum of the points associated with the criteria triggered by OncoVI for the variant and the correspondence with the classification is: score$\geq$ 10: Oncogenic (O), 6 $\leq$score$\leq$ 9: Likely Oncogenic (LO), 0 $\leq$score$\leq$ 5: Variant of uncertain significance (VUS), -6 $\leq$score$\leq$ -1: Likely Benign (LB), score$\leq$ -7: Benign (B). Resources in blue text are suggested by the Standard Operating Procedure, resources in black text are identified by the authors of this study. Figure from [102].

To enable a correct implementation of the criteria, we first needed to interpret each criterion, whenever this resulted unclear from the textual indication provided by the SOP.

For example, the criteria "Somatic Benign Very Strong-1" (SBVS1, -8 points, *"Minor allele frequency is >5% in gnomAD in any 5 general continental populations"*) and "Somatic Benign Strong-1" (SBS1, -4 points, *"Minor allele frequency is >1% in gnomAD in any 5 general continental populations"*) for evidence of benign effect did not require further interpretation. In contrast, "Oncogenic Supporting-4" (OP4, 1 point, *"Absent from controls (or at an extremely low frequency) in gnomAD"*) required choosing a threshold for the gnomAD population frequency, which was set to 1% in our implementation in OncoVI. Furthermore, for criteria like "Oncogenic Moderate-1" (OM1, 2 points, *"Located in a critical and well-established part of a functional domain"*) an extended investigation into the biological characteristic addressed by the criterion was needed. For example, our interpretation of OM1 addresses variants located in a functional domain of the protein. Indeed, a functional domain by definition folds and functions independently from the rest of the polypeptide chain, and we therefore hypothesised that variants affecting protein domains are more likely to disrupt the protein function.

The oncogenicity assessment of somatic variants is articulated in OncoVI in two steps. First, information is collected from the publicly available resources identified as reference for each criterion. Then, OncoVI evaluates the obtained information and decides, based on an IF-ELSE rule, whether the information supports the criterion, thus assigning the associated points. The sum of the points aggregated from the criteria triggered by one variant, the "variant-specific score", allows the classification of oncogenicity.

The resources to use as reference for each criterion were either adopted from those suggested in the SOP or identified from publicly available databases, when no suggestion by the SOP was provided. Our implementation of OM1, for example, employs human protein domains in UniProt [111] as reference resource. Based on an IF-ELSE rule, OncoVI verifies whether UniProt includes a protein encoded by the gene affected by the variant under evaluation. If the nucleotide triplet affected by the variant belongs to a functional domain, then OM1 is triggered, otherwise not.

"Oncogenic Very Strong-1" (OVS1, 8 points, *"Null variant in a bona fide tumour suppressor gene"*) is triggered for nonsense variants located in bona fide tumour suppressor genes (TSG)s, defined as the union of the genes reported in OncoKB and the Cancer Gene Census (CGC) Tier 1 of COSMIC. Alternatively, OVS1 is triggered if the variant is present in MutSpliceDB [112], as suggested by the SOP. OncoKB and the CGC were also employed in "Oncogenic Moderate-2" (OM2, 2 points, *"Protein length changes as a result of in-frame deletions/insertions in a known oncogene or tumor suppressor gene or stop-loss variants in a known tumor suppressor gene"*) to define known oncogenes (OG)s and TSGs, yet this time regardless of the Tier.

ClinVar was used for "Oncogenic Strong-2" (OS2, 4 points, *"Well-established in vitro*

or in vivo functional studies, supportive of an oncogenic effect of the variant") and "Somatic Benign Strong-2" (SBS2, Strong, -2 points, *"Well-established in vitro or in vivo functional studies show no oncogenic effects"*), which consider variants supported by well-established *in vitro or in vivo* functional studies showing no oncogenic or benign effects, respectively (variant_summary.txt.gz, accessed 28 February 2024). For SBS2, only variants classified as either Benign or Likely Benign were evaluated, while for OS2 a manually curated list of different classifications in ClinVar was considered. These classifications include not only *"in vivo and in vitro* studies" as submission methods, but also others such as "clinical testing". This choice was mainly dictated by the fact that ClinVar variants from *"in vivo and in vitro* studies" represent only the 0.2% (6,891 of 3,462,730; submission_summary.txt.gz, accessed 12.04.2025).

For the implementation of "Oncogenic Strong-3" (OS3, 4 points, *"Located in one of the hotspots in cancerhotspots.org with at least 50 samples with a somatic variant at the same amino acid position, and the same amino acid change count in cancerhotspots.org in at least 10 samples"*), "Oncogenic Moderate-4" (OM4, 2 points, *"Located in one of the hotspots in cancerhotspots.org with <50 samples with a somatic variant at the same amino acid position, and the same amino acid change count in cancerhotspots.org is at least 10"*), and "Oncogenic Supporting-3" (OP3, 1 point, *"Located in one of the hotspots in ccancerhotspots.org and the particular amino acid change count in cancerhotspots.org is below 10"*) we used the resource suggested by the SOP, i.e., cancerhotspots.org [113, 114] that contains well-known hotspot variants. Additionally, to ensure a comprehensive assessment of hotspot variants, we included COSMIC data, considering variants found in at least 50 samples, with at least ten showing the exact nucleotide change.

The criteria "Oncogenic Strong-1" (OS1, Strong, 4 points, *"Same amino acid change as a previously established oncogenic variant (using this standard) regardless of nucleotide change"*) and "Oncogenic Moderate-3" (OM3, Moderate, 2 points, *"Missense variant at an amino acid residue where a different missense variant determined to be oncogenic (using this standard) has been documented"*) required variants previously classified according to the SOP. However, since this set of variants is not available when applying the guidelines for the first time, in OncoVI it was chosen to rely on the CGI Catalog of Validated Oncogenic Mutations (Accessed 01 February 2024). Moreover, in the implementation of criterion OM3, the amino acid difference is calculated based on the Grantham's distance [115], a measure of dissimilarity between amino acids, as recommended in the SOP.

The computational algorithms consulted by OncoVI to trigger "Oncogenic Supporting-1" (OP1, 1 point, *"All used lines of computational evidence support an oncogenic effect of a variant"*), "Somatic Benign Supporting-1" (SBP1, -1 point, *"All used lines of computational evidence suggest no effect of a variant"*), and "Somatic Benign Supporting-2"

(SBP2, -2 points, *"A synonymous (silent) variant for which splicing prediction algorithms predict no effect on the splice consensus sequence nor the creation of a new splice site and the nucleotide is not highly conserved"*) are those cited by the SOP to predict conserved elements across species and splicing effect, i.e., phyloP100way_vertebrate_rankscore (phyloP) [116], phastCons100way_vertebrate_rankscore (phastCons) [117], and spliceAI. Variants with a phyloP OR phastCons score ≥ 0.5 OR a spliceAI score equal to "PASS" were considered to trigger OP1, otherwise SBP1 and SBP2 were evaluated.

4.2.2 SOP data set

The SOP data set contains 93 variants and was obtained as the manually curated union of Supplementary Tables 1 and 3 provided by Horak et al [30]. The 93 variants are located in ten well-known genes: i.e., six OGs, three TSGs, and one gene with a context-dependent role. OG are genes that normally regulate cell growth and proliferation, but that, when mutated, acquire a gain of function that drives uncontrolled cell division. Conversely, TSGs act as safeguards by inhibiting proliferation, maintaining genomic stability, or triggering apoptosis. Activating mutations often involve OGs; loss-of-function mutations in TSGs, instead, disrupt these safeguard mechanisms, thereby promoting cancer progression [74]. The 70 missense variants in Supplementary Table 3 were provided with several information for both genome reference builds 37 and 38, in addition to DNA change, i.e., chromosome, genomic position, strand orientation, reference base, and alternate base. This additional information was missing for the remaining 23 variants in Supplementary Table 1. Thus, they were derived manually from the DNA change. Furthermore, for all variants the triggered criteria, the variant-specific scores, and the oncogenicity classification were available in the Supplementary Tables.

4.2.3 MTB data set

The real-world MTB data set is composed of 7,802 unique somatic variants identified in 557 patients with various tumour diagnoses (**Figure 12**), enrolled in the setting of the MTB of the Institute of Pathology UKER from February 2022 to February 2024.

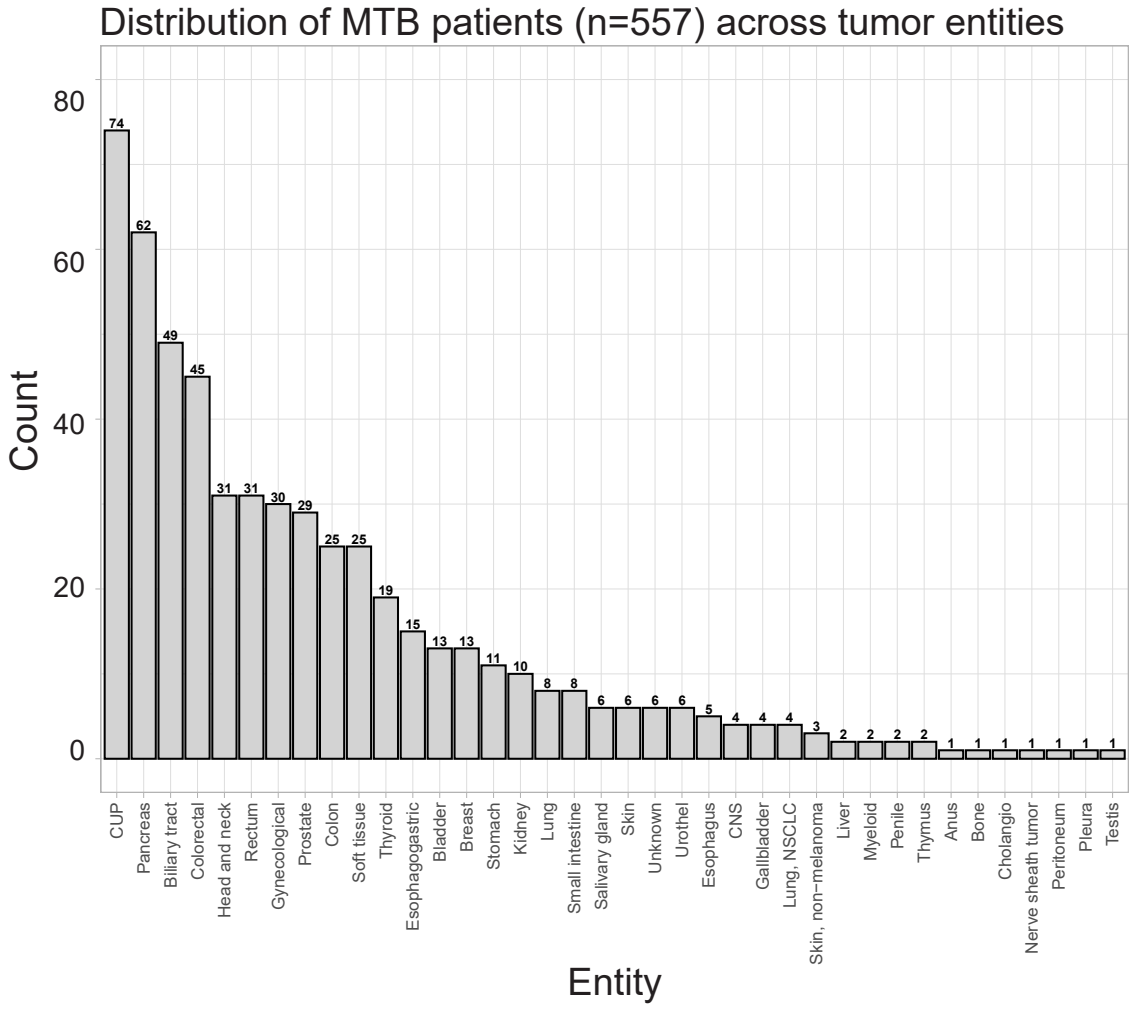

Figure 12: Distribution of MTB patients across tumour entities. Distribution of the tumour entities associated with the 557 patients of the Molecular Tumour Board (MTB) cohort. CUP: cancer of unknown primary. CNS: central nervous system. Lung, NSCLC: Non-small-cell lung cancer. Figure panel modified from [102].

DNA sequencing libraries were prepared and sequenced following the protocol detailed for the MTB cohort in section 3.1.5 of this PhD work. Three expert cancer biologists assessed variants in terms of their effect on the WT protein function and assigned them to one of five classes ("Pathogenic", "Likely Pathogenic", "VUS", "Likely Benign", and "Benign"). Later, the multidisciplinary team of the MTB evaluated variants for their actionability.

4.2.4 Validation data set

The validation data set of 135 somatic variants was created from the MTB data set to validate the implementation of the oncogenicity criteria in OncoVI. The 135 variants were chosen *a posteriori*, on the basis of results obtained on the entire MTB data set

(see section 4.3.2). The same three expert cancer biologists who had manually curated the classification of the MTB data set were engaged to classify the subset of 135 somatic variants, this time according to the oncogenicity criteria. To this aim, the validation data set was randomly divided into three separate sets and the list of the resources implemented in OncoVI as reference for each criterion was provided to the experts.

4.2.5 ClinVar data set

The complete set of 691 somatic variants for which the oncogenicity classification was available in ClinVar was retrieved (variant_summary.txt.gz, accessed 12.04.2025). For these variants, oncogenicity was calculated by NCBI based on data from submitters using the oncogenicity guidelines [30]. Chromosome, PositionVCF, ReferenceAlleleVCF, and AlternateAlleleVCF in the hg38 reference genome build were used to create the input VCF file for annotation with the bioinformatics framework described in Chapter 3 and incorporated in OncoVI. Oncogenicity was considered as ground truth to evaluate OncoVI performance on the ClinVar data set.

4.2.6 Variant annotation via custom bioinformatics framework

Functional annotation of the variants in the SOP, MTB, and ClinVar data sets was performed relying on the custom Python-based bioinformatics framework described in Chapter 3, which is incorporated in OncoVI. The pipeline runs in a dedicated conda environment (version 24.11.1) on a remote server based on Ubuntu 20.04.6 LTS operating system and takes the genomic positions of the variants as input. As described in the previous sections, for the SOP and ClinVar data sets, the coordinates in the hg38 reference genome build were used. For the variants of the MTB data set, genomic coordinates were obtained by uplifting of the original hg19-based VCFs (i.e., the "MergedSmallVariants.genome.vcf" files) to the genome reference build hg38 via the LiftOverVcf function of the picard package (version 3.0.0).

This time the variants were functionally annotated with the Ensembl VEP (VEP version 111.0, January 2024) utilising GRCh38 as reference genome build. An additional VEP plugin, spliceAI, was employed to retrieve scores from predictive algorithms in addition to dbNSFP (version 4.5a). For variants for which multiple transcripts were annotated, the canonical transcript according to the Matched Annotation NCBI EBI (MANE) project [118] was favoured. Eventually, the pipeline interrogates several knowledgebases to retrieve existing biomedical information of the somatic variants, as described in section 3.1.2. The final output of the pipeline is the list of annotated variants in the form of tabular data (csv and Microsoft Excel).

All statistical analyses were performed within R (version 4.1.2)/ Bioconductor (version 3.13) environment [119]. Spearman correlations were computed with the function cor.test() of the stats package. Contingency tables were generated using the R package caret (version 6.0.94). Odds ratios were calculated to test the association between the criteria triggered by OncoVI and the correct classification of the Oncogenic or Likely Oncogenic variants of the SOP data set. Statistical significance of this relationship was tested via a two-sided Fisher's exact test relying on the R stats package. p-values ≤ 0.05 were considered statistically significant. All plots were created with the R packages ggplot2 (version 3.5.1) and ggalt (version 0.4.0). OncoPrints were produced with the R package ComplexHeatmap (version 2.13.1). Alluvial plots were created using the R package ggalluvial (version 0.12.5).

4.3 Results

4.3.1 Performance of OncoVI on the SOP data set

To assess the validity of the resources allocated to each criterion and of the automated interpretation of the oncogenicity criteria, OncoVI was tested on the SOP data set of 93 somatic variants. Variant-associated triggered criteria, variant-specific scores, and oncogenicity classifications were provided by the SOP. Variants of the SOP data set were distributed across the five oncogenicity classes as follows: Oncogenic: 14, Likely Oncogenic: 29, VUS: 38, Likely Benign: 6, and Benign: 6. Variant functional annotation revealed that the most prevalent types of variants were missense (n=71), truncating (n=9), and variants affecting the upstream gene region (n=5).

To evaluate OncoVI performance on the SOP data set, we grouped Likely Oncogenic (LO) and Oncogenic (O) variants into the class Oncogenic/Likely Oncogenic (O/LO) and we grouped Likely Benign (LB) and Benign (B) variants into the class Benign/Likely Benign (B/LB). When considering the SOP oncogenicity classification based on three oncogenicity classes (i.e., O/LO, VUS and B/LB) as ground truth, the overall accuracy of OncoVI was 81% (n=75/93 correctly classified variants), with a sensitivity of 88% for the O/LO class (n=38/43) and a sensitivity of 83% for the B/LB class (n=10/12) (**Figure 13a**). Taken together, OncoVI showed very good agreement with the variant oncogenicity classification of the SOP.

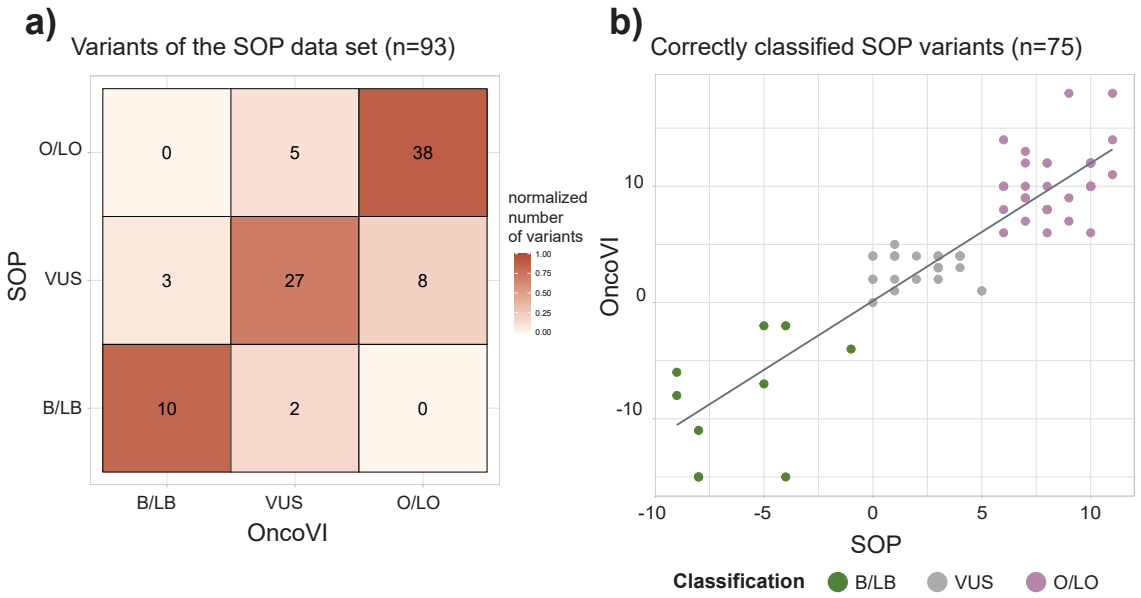

Figure 13: Results of OncoVI on the SOP data set. a) Confusion matrix of the agreement between the Standard Operating Procedure (SOP) and OncoVI in classifying the variants of the SOP data set. The colour scale indicates the normalised number of variants, i.e., the ratio (calculated by row) between the number of variants of each cell and the total number of variants. **b)** Scatterplot between the points assigned by the SOP and the points assigned by OncoVI for the 75 correctly classified variants. B/LB: Benign/Likely Benign. VUS: Variant of uncertain significance. O/LO: Oncogenic/Likely Oncogenic. Figure panel modified from [102].

Next, the agreement between the SOP and OncoVI in terms of variant-specific scores was evaluated. When considering the 75 correctly classified variants of the SOP data set, a strong positive correlation (0.87, p-value <2.2e-16) between the points assigned by the SOP and by OncoVI was observed (**Figure 13b**). Furthermore, in case of divergent scores, OncoVI usually assigned higher scores than the SOP both in correctly classified O/LO (**Figure 14a**) and variants classified as VUS (**Figure 14c**), and lower scores in correctly classified B/LB variants (**Figure 14b**). Overall, OncoVI assigned higher points than the SOP in correctly classified variants, which aligns with our aim to have a high sensitivity for the O/LO class.

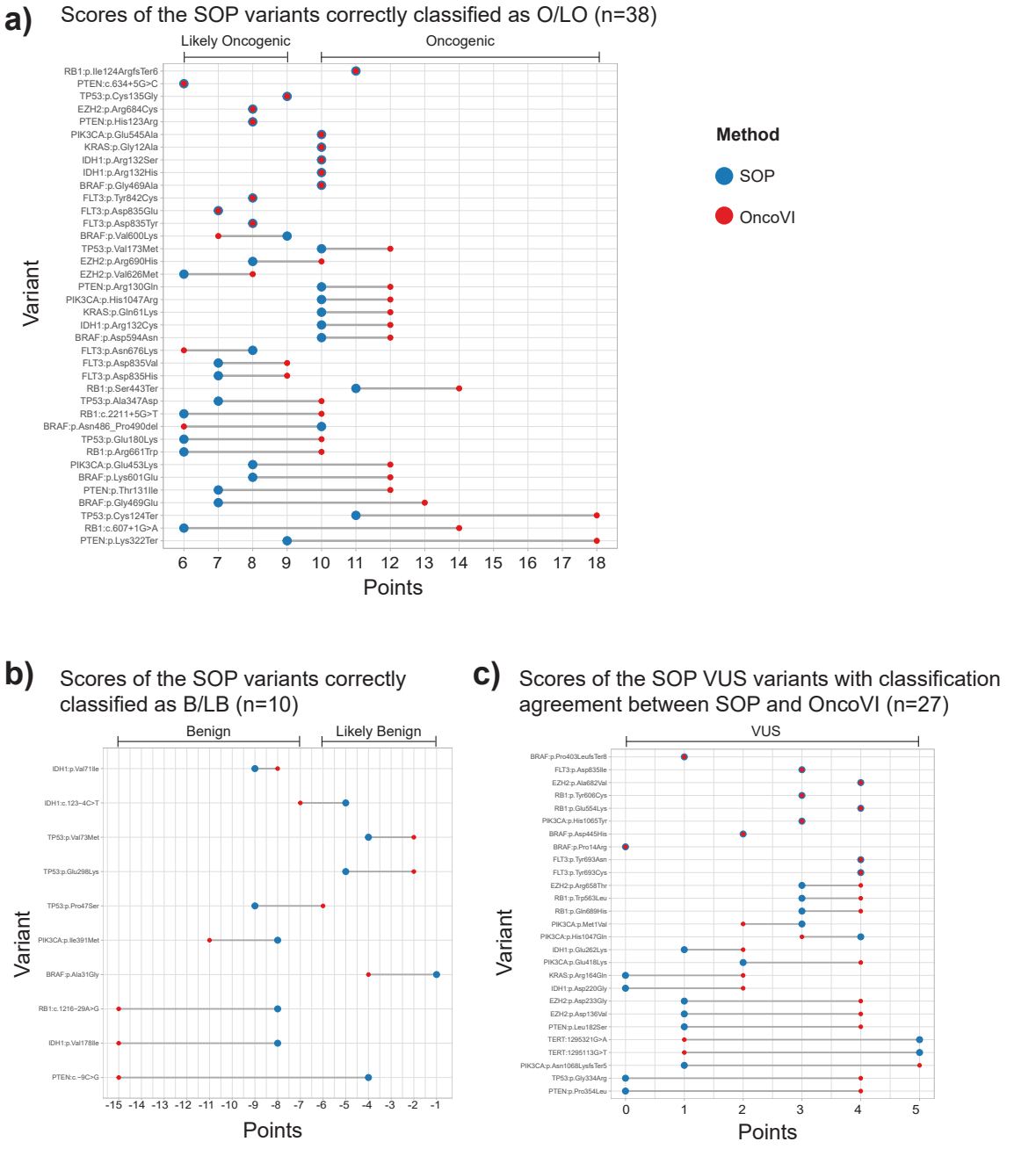

Figure 14: Variant-specific scores of the SOP data set. a) Dumbell plot of the 38 variants correctly classified as Oncogenic/Likely Oncogenic (O/LO). **b)** Dumbell plot of the ten variants correctly classified as Benign/Likely Benign (B/LB). **c)** Dumbell plot of the 27 SOP variants of uncertain significance (VUS) with classification agreement between the SOP and OncoVI. Horizontal bars indicate the classification of the variants according to the Standard Operating Procedure (SOP) point-based system (i.e., score$\geq$ 10: Oncogenic, 6 $\leq$score$\leq$ 9: Likely Oncogenic, -6 $\leq$score$\leq$ -1: Likely Benign, score-7 $\leq$: Benign, 0 $\leq$score$\leq$ 5: VUS). Figure panel modified from [102].

To further compare OncoVI and the SOP, the most frequently criteria triggered by OncoVI in the 75 correctly classified variants were investigated. In the correctly classified O/LO variants, OncoVI triggered Very Strong (8 points) and Strong (4 points) criteria

for oncogenic effect, which are associated with the highest points, and thus contribute the most to the variant-specific score (**Figure 15a**). Analogously, in correctly classifying B/LB variants OncoVI assigned Very Strong (-8 points) and Strong (-4 points) criteria for benign effect, which represent the criteria with the highest negative score, thus leading to a low overall variant-specific score for correct benign classification (**Figure 15b**).

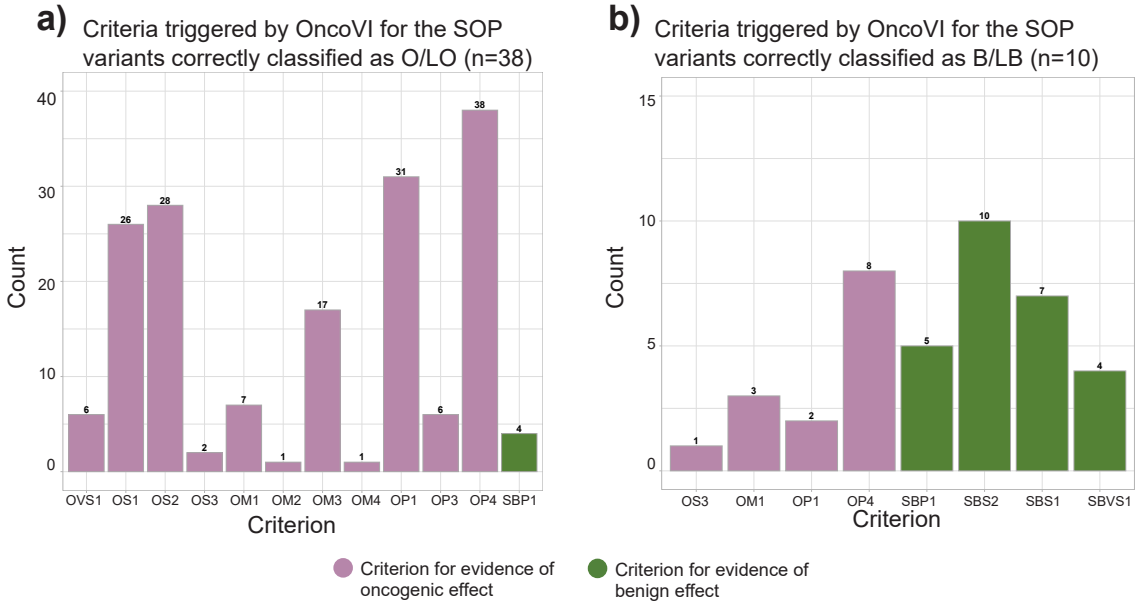

Figure 15: Criteria triggered by OncoVI on the SOP data set. a) Barplot of the criteria triggered by OncoVI in the 38 variants correctly classified as Oncogenic/Likely Oncogenic (O/LO) of the Standard Operating Procedure (SOP) data set. **b)** Barplot of the criteria triggered by OncoVI in the ten variants correctly classified as Benign/Likely Benign (B/LB) of the SOP data set. Criteria are sorted according to decreasing corresponding points. Figure panel modified from [102].

To further assess the choice of the supporting resources, in addition to investigating how often OncoVI activated each criterion, we evaluated the association between the triggering of each criterion by OncoVI and the correct classification of O/LO variants of the SOP data set. The oncogenic criteria, whose triggering was most associated with a correct classification, were OS1 (Odds ratio: 11.8, p-value =0.013), and OS2 (Odds ratio: 6.29, p-value =0.036), and no benign criterion showed association (**Table 5**). Taken together, the assessment of the triggered criteria supported the appropriateness of the chosen resources and showed that the classification of both oncogenic and benign variants mainly depends on the triggering of Strong criteria.

Table 5: **Odds ratio of association between triggering of criteria and the correct classification of Oncogenic or Likely Oncogenic variants of the SOP data set.** OR: Odds ratio. CI: Confidence interval. Bold italic values indicate statistically significant p-values (p-value <0.05).

Criterion	**OR**	**CI**	**p-value**
OVS1	1.27	0.12-66.78	1
OS1	11.8	1.24-591.71	***0.013***
OS2	6.29	0.87-75.17	***0.036***
OS3	0.5	0.03-29.75	0.488
OM1	0.2	0.024-1.41	0.06
OM2	0.33	0.015-21.71	0.39
OM3	4.77	0.51-237.82	0.22
OM4	0.33	0.01-21.71	0.39
OP1	2.912	0.36-21.45	0.33
OP3	1.27	0.12-66.78	1
OP4	6.09	0.07-520.21	0.28
SBVS1	0.16	0.002-14.07	0.28
SBS1	0.16	0.002-14.07	0.28
SBS2	0.16	0.002-14.07	0.28
SBP1	0.86	0.07-47.21	1
SBP2	0.16	0.002-14.07	0.28

To identify potential reasons for the higher variant-specific scores assigned by OncoVI, we investigated the criteria triggered by the SOP and OncoVI individually, focusing on each of the 38 variants correctly classified as O/LO (**Figure 16**).

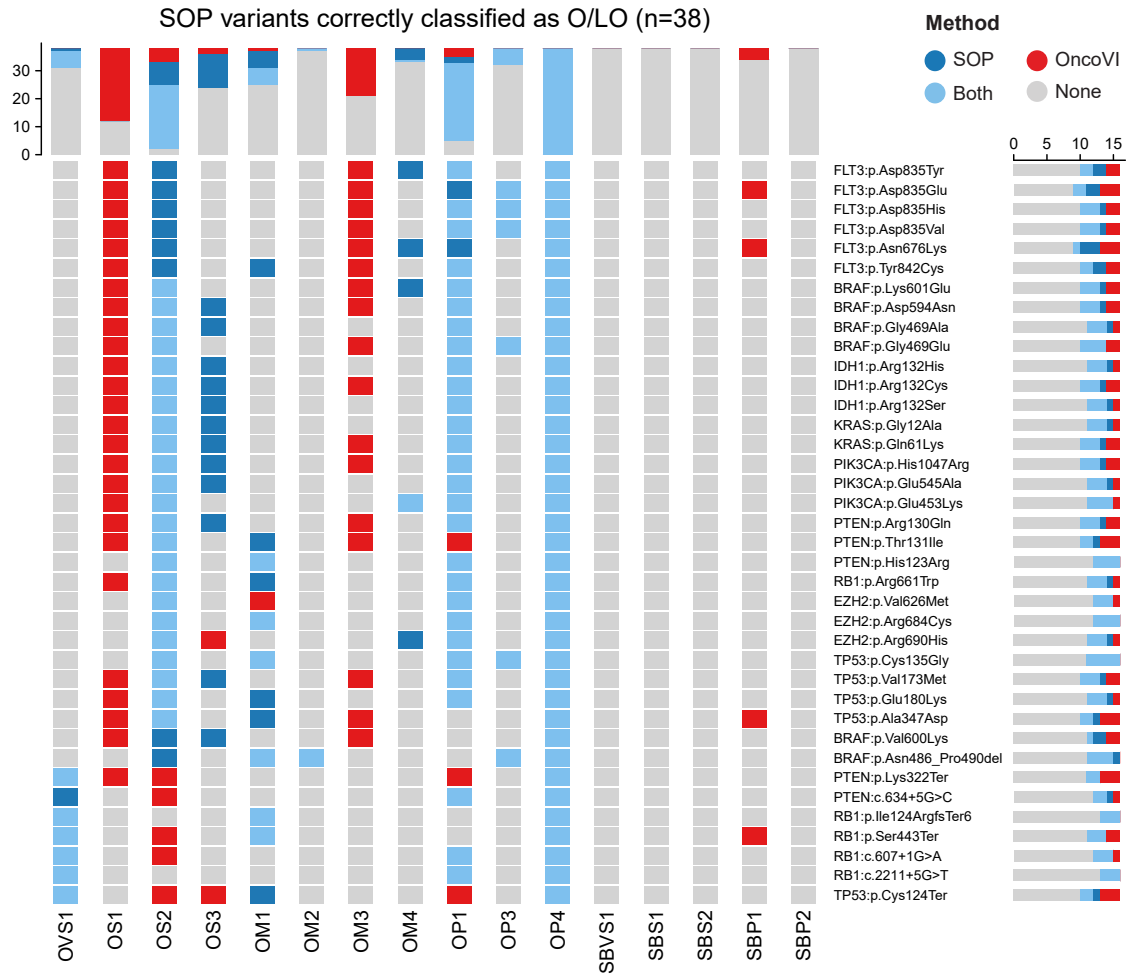

Figure 16: Criteria triggered by OncoVI and SOP on correctly classified O/LO variants of the SOP data set. Oncoprint of the 38 variants of the Standard Operating Procedure (SOP) data set correctly classified as Oncogenic/Likely Oncogenic (O/LO). Each row corresponds to one variant, each column corresponds to an assessed criterion. The colour of the cell indicates whether a criterion was triggered by the SOP only (dark blue), OncoVI only (red), both (light blue), or none of the two (grey). The top barplot shows the sum of cells of each colour calculated across all variants, the barplot on the right shows the sum across all criteria. Figure panel modified from [102].

The strongest agreement between the two methods in the triggered criteria was observed for OP3 (n=6/6 variants), OP4 (n=38/38 variants), OP1 (n=28/33 variants), OVS1 (n=6/7 variants), OM1 (n=6/12 variants), and OS2 (n=23/36 variants). Instead, OS3 was triggered almost exclusively by the SOP (n=12/14 variants), while OS1 (Strong, 4 points, *"Same amino acid change as a previously established oncogenic variant (using this standard) regardless of nucleotide change"*) and OM3 (Moderate, 2 points, *"Missense variant at an amino acid residue where a different missense variant determined to be oncogenic (using this standard) has been documented"*) were exclusively triggered by OncoVI. Indeed, for OS1 and OM3 OncoVI used an external data set of validated oncogenic mutations to define the "variant determined to be oncogenic (using this standard)". In

addition, according to the guidelines, OS3 is not applicable when OS1 is activated. Taken together, the comparison between OncoVI and the SOP confirmed the validity of the automated implementation offered by OncoVI and showed a good agreement in triggered criteria for the variants correctly classified as O/LO.

4.3.2 Performance of OncoVI on the MTB data set

To evaluate OncoVI performance in a real-world setting, we applied it to a data set of 7,802 unique somatic variants previously classified by three expert cancer biologists in the context of our Erlangen MTB. The 7,802 variants ("MTB data set") had been detected in 489 different genes in a cohort of 557 patients with different tumour entities. In contrast to the ten genes of the SOP data set, not for all genes of the MTB data set a clear indication about their role as either OG or TSG was available. Thus, the MTB data set was not only larger but also more heterogeneous. Another key difference is that variants of the MTB data set had been assessed for their effect on the protein function and not for oncogenicity. Indeed, the 7,802 variants had been assigned to one of five classes: Pathogenic, Likely Pathogenic, VUS, Likely Benign, and Benign (**Figure 17**).

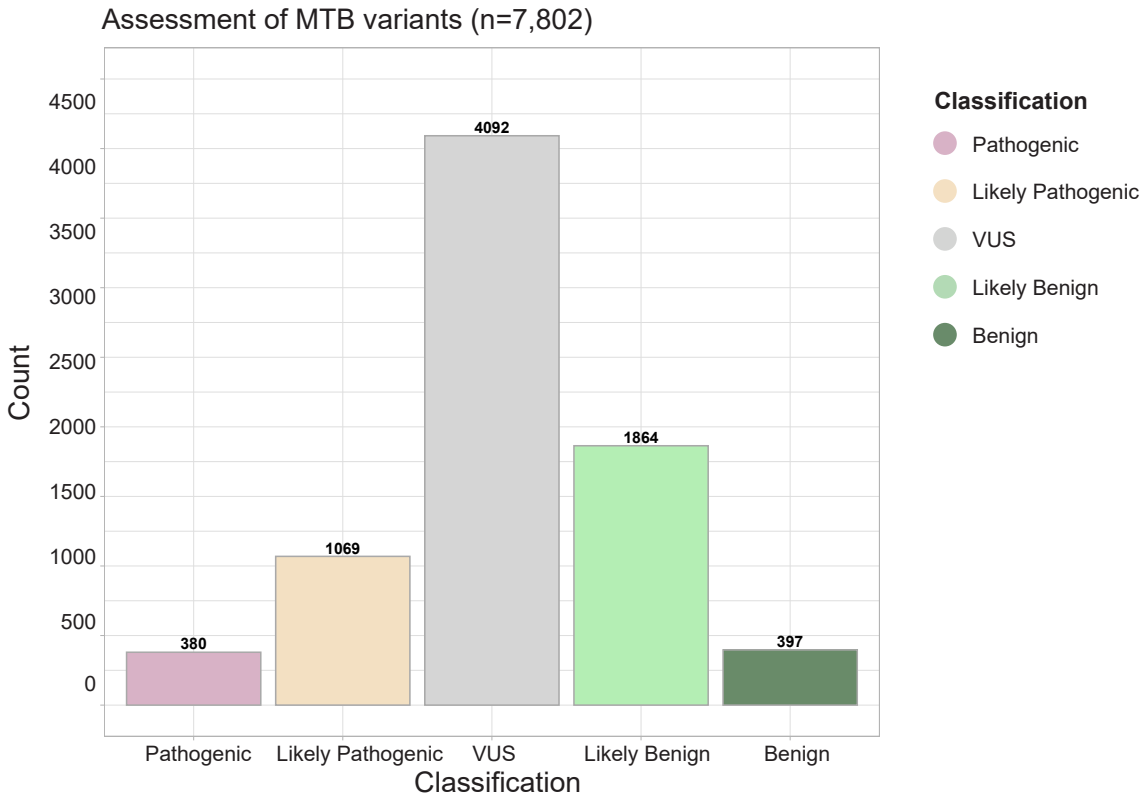

Figure 17: MTB data set of real-world routine diagnostic variants. Distribution of the 7,802 variants of the Molecular Tumour Board (MTB) data set across the MTB assessment classes. VUS: Variant of uncertain significance. Figure panel modified from [102].

To compare OncoVI with the MTB assessment, the variants belonging to Pathogenic and Likely Pathogenic classes were grouped into the "Pathogenic/Likely Pathogenic" (P/LP) class. In addition, the variants classified as Benign or Likely Benign were grouped together into the "Benign/Likely Benign" (B/LB) class.

Variant functional annotation revealed that the three most frequent types of variants were missense (n=6,363), frameshift (n=540), and stop gained (n=467). The performance of OncoVI was evaluated by assessing the agreement of its oncogenicity classification into O/LO, VUS, and B/LB with the MTB classification into P/LP, VUS, and B/LB. The observed agreement was 78% (n=6,060/7,802 variants), in line with the accuracy observed for the SOP data set (81%, n=75/93 variants). In particular, VUS and B/LB classes had an agreement of 97% (n=3,955/4,092 variants) and 43% (n=961/2,261 variants), respectively, while the concordance between OncoVI O/LO and MTB P/LP was 79% (n=1,144/1,449 variants) (**Figure 18**).

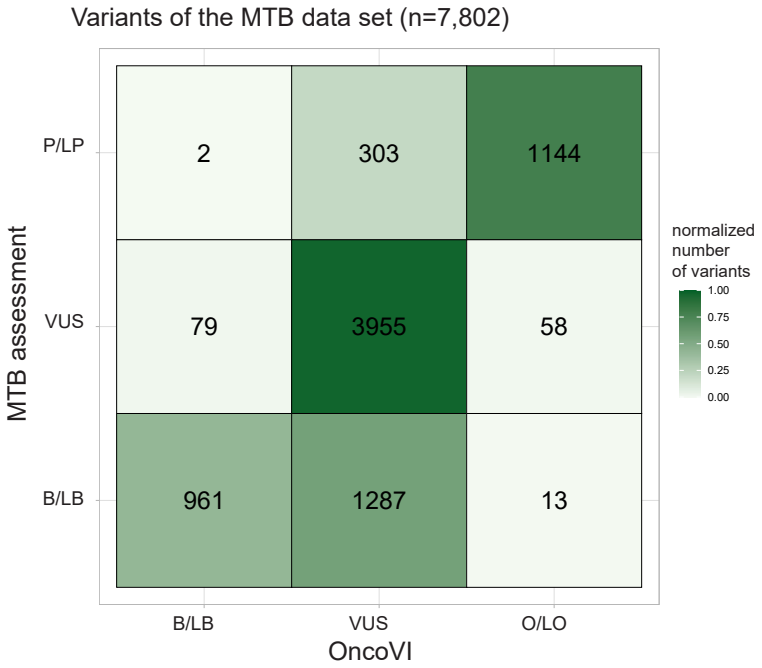

Figure 18: Results of OncoVI on the MTB data set. Confusion matrix of the agreement between the Molecular Tumour Board (MTB) assessment and OncoVI oncogenicity classification on the 7,802 variants of the MTB data set. B/LB: Benign/Likely Benign. VUS: Variant of uncertain significance. O/LO: Oncogenic/Likely Oncogenic. P/LP: Pathogenic/Likely Pathogenic. The colour scale indicates the normalised number of variants, i.e., the ratio (calculated by row) between the number of variants of each cell and the total number of variants. Figure panel modified from [102].

The most critical issue in using OncoVI for oncogenicity classification are false negative oncogenic variants. Further examination of the 303 P/LP variants classified as VUS by OncoVI showed that the majority (64%, n=196/303) were truncating (i.e., frameshift, stop gained) or splice site mutations, according to our bioinformatics framework.

To determine whether OncoVI behaved differently on real-world variants compared to the SOP data set, we analysed the most frequently triggered criteria. For 1,144 O/LO variants with agreement between MTB assessment and OncoVI oncogenicity classification, the top three most frequently triggered criteria were: OP4, OVS1, and OP1 for evidence of oncogenic effect (**Figure 19**).

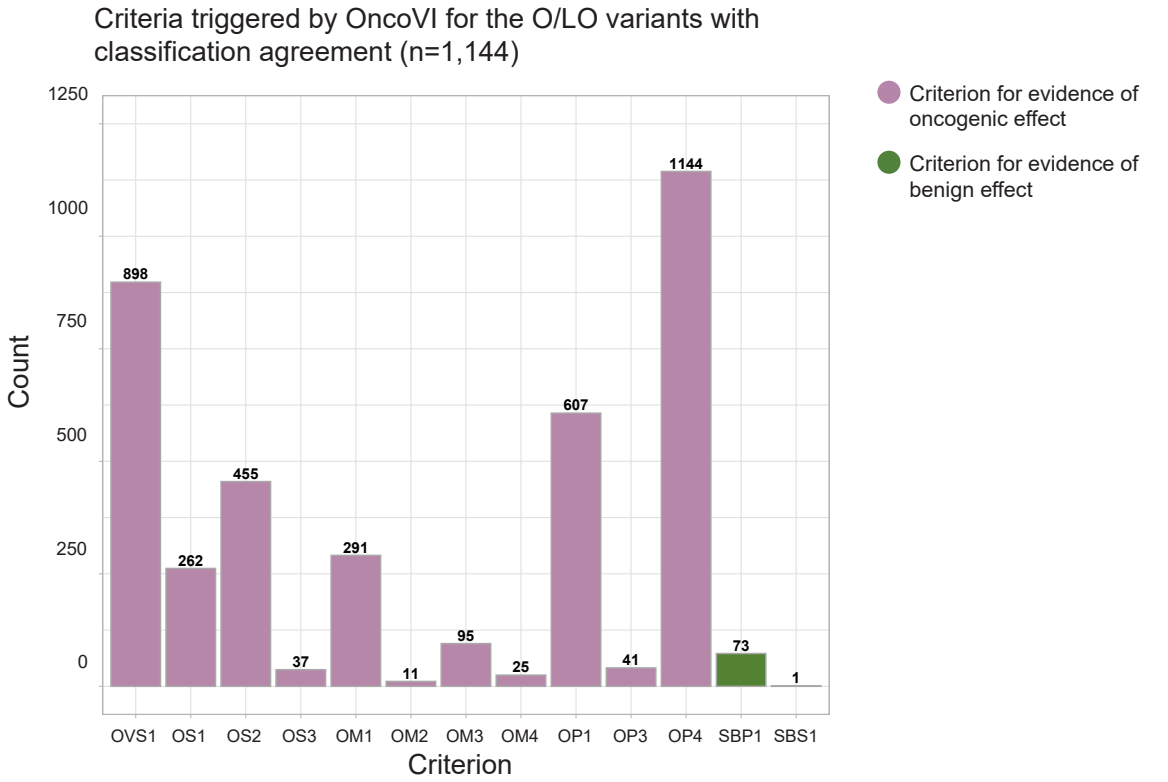

Figure 19: Criteria triggered by OncoVI on the variants of the MTB data set. Barplot of the criteria triggered by OncoVI in the 1,144 O/LO variants with classification agreement between Molecular Tumour Board (MTB) and OncoVI assessments. Criteria are sorted according to decreasing corresponding points. Figure panel modified from [102].

As for the SOP data set, OP4 (Supporting, 1 point, *"Absent from controls (or at an extremely low frequency) in gnomAD"*) was the top triggered criterion. OVS1 (Very Strong, 8 points, *"Null variants in a bona fide TSG"*) was triggered in 79% (n=898/1,144) of the variants, thus revealing a higher percentage of TSGs in the real-world MTB data set than in the SOP data set (16%, n=6/38). Among the 961 B/LB variants with agreement between MTB and OncoVI oncogenicity classification, SBS2 (Strong, -4 points, *"Well-established in vitro or in vivo functional studies showing no oncogenic effects"*) was the most frequently triggered criterion (n=821/961), confirming ClinVar as suitable reference for SBS2 (**Figure 20a**). SBP1 (n=549), SBS1 (n=496), and SBVS1 (n=149) were also frequently triggered as benign evidence.

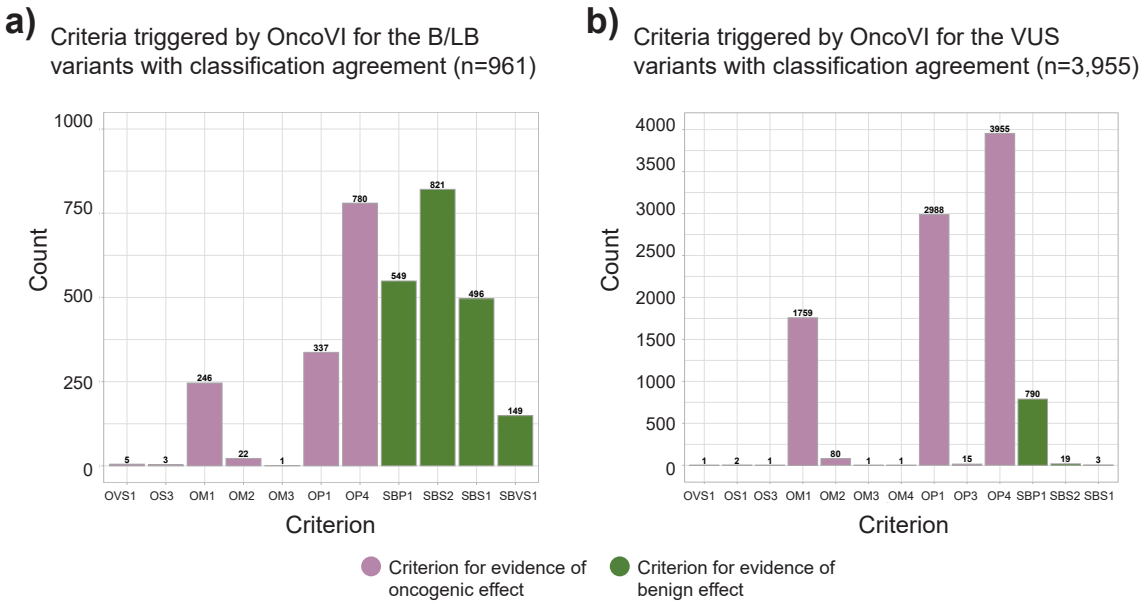

Figure 20: Criteria of the B/LB and VUS variants of the MTB data set with classification agreement. a) Barplot of the criteria triggered by OncoVI in the 961 variants classified as Benign/Likely Benign (B/LB) by both Molecular Tumour Board (MTB) and OncoVI. **b)** Barplot of the criteria triggered by OncoVI in the 3,955 variants classified as variant of uncertain significance (VUS) by both MTB and OncoVI. Criteria are sorted according to decreasing corresponding points. Figure panel modified from [102].

In the 3,955 variants classified as VUS by both OncoVI and the MTB, OP4, OP1 and OM1 were the most frequently triggered criteria as oncogenic evidence (OP4: n=3,955, OP1: n=2,988, OM1: n=1,759). SBP1 was triggered 790 times because supported by computational algorithms that reported a benign effect (**Figure 20b**).

Taken together, although the MTB assessment evaluated the variant effect on protein function and OncoVI focused on oncogenicity, a satisfactory overall agreement was observed. In addition, OncoVI specifically triggered Strong oncogenic and benign criteria in O/LO and B/LB variants, respectively, in agreement with the MTB classification.

4.3.3 Performance of OncoVI on the validation data set

The expert cancer biologists had not assessed MTB variants for oncogenicity, thus the comparison with OncoVI classification was limited to the assessment of the variant effect on protein function. To evaluate OncoVI performance in classifying the oncogenicity of real-world somatic variants, the expert cancer biologists re-assessed a subset of the MTB variants using the oncogenicity guidelines and the same resources used by OncoVI. Yet, the experts were blinded to the IF-ELSE translation of the criteria implemented in OncoVI. The subset of 135 re-assessed variants ("validation data set") was selected based on the results obtained in the comparison between OncoVI and the MTB assessment (**Figure 18**) in order to: (i) identify potential reasons for the observed discrepancies in

the classification, and (ii) assess the consistency between OncoVI criteria implementation, and expert interpretation of the resources. Specifically, the validation data set included the two variants classified within the MTB as P/LP and classified by OncoVI as B/LB, the 13 variants classified within the MTB as B/LB and misclassified by OncoVI as O/LO, and randomly chosen variants classified as O/LO or VUS by OncoVI, both with agreement and with discrepancies with the MTB assessment (**Figure 21a**).

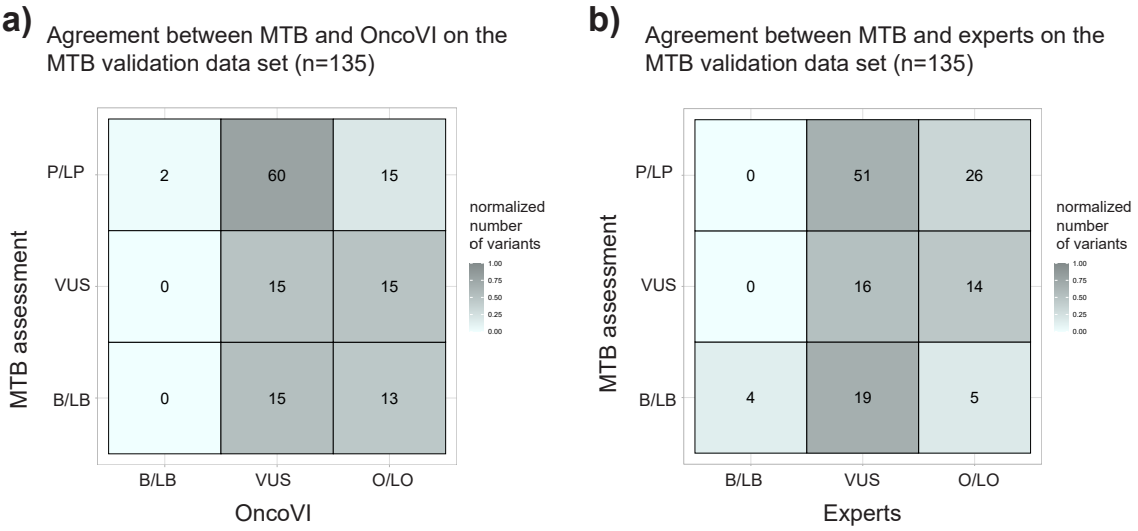

Figure 21: MTB variants of the validation data set. a) Confusion matrix of the agreement between OncoVI oncogenicity and Molecular Tumour Board (MTB) assessment in classifying the 135 variants of the validation data set. **b)** Confusion matrix of the agreement between MTB and expert re-assessment, based on the oncogenicity guidelines, in classifying the 135 variants of the validation data set. The colour scale indicates the normalised number of variants, i.e., the ratio (calculated by row) between the number of variants of each cell and the total number of variants. B/LB: Benign/Likely Benign. VUS: Variant of uncertain significance. O/LO: Oncogenic/Likely Oncogenic. P/LP: Pathogenic/Likely Pathogenic. Figure panel modified from [102].

Variants of the validation data set were located in 104 different genes, of which 67 OGs, 61 TSGs, and seven with unknown role. Similarly to the SOP data set, the most prevalent types of variants were truncating (n=63), missense (n=50), and variants affecting the splice site. Expert classification based on the oncogenicity guidelines achieved an agreement of 34% (n=46/135 variants) with their earlier assessment in the MTB (**Figure 21b**, **Figure 22**).

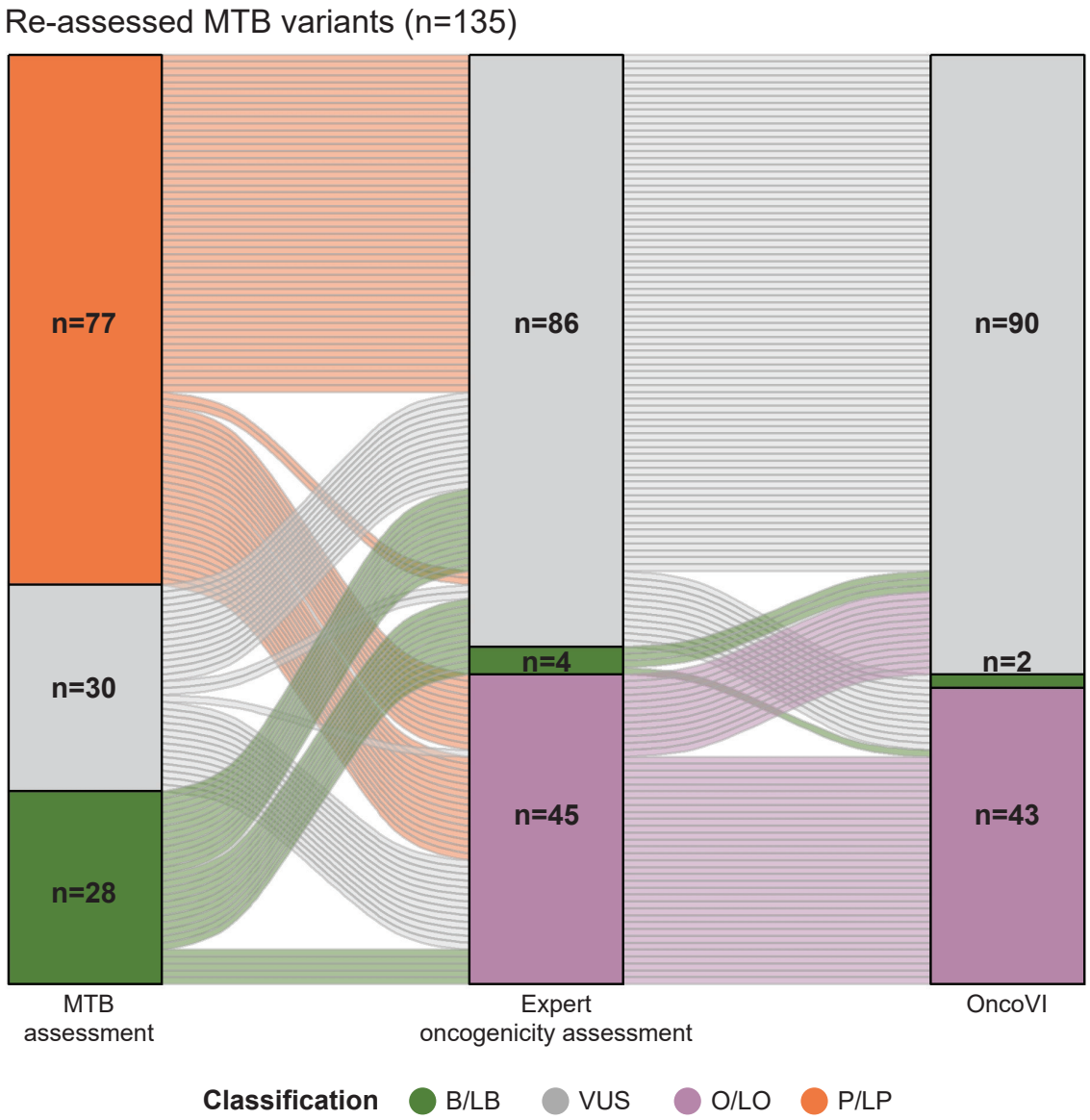

Figure 22: Assessment of the variants of the validation data set by MTB, experts and OncoVI. Alluvial plot of the 135 variants of the validation data set re-assessed by experts based on oncogenicity. Horizontally-distributed columns (axes) represent variants classified by the Molecular Tumour Board (MTB) (left), by the re-assessment of the experts based on the oncogenicity guidelines (middle), and by OncoVI (right). Alluvial flows between axes show the correspondence between variants' classifications. B/LB: Benign/Likely Benign. VUS: Variant of uncertain significance. O/LO: Oncogenic/Likely Oncogenic. P/LP: Pathogenic/Likely Pathogenic. Figure panel modified from [102].

The agreement of OncoVI with the earlier assessment of the MTB based on the variant effect on protein function was 22% (n=30/135 variants) (**Figure 21a**) and the agreement with expert re-assessment based on the oncogenicity guidelines was 80% (n=108/135 variants) (**Figure 23a**). No variants assessed by experts as O/LO were classified as B/LB by OncoVI. In addition to the classification comparison, the concordance of variant-specific scores between OncoVI and the expert re-assessment was also evaluated, as for

the SOP data set. When considering the 108 correctly classified variants, a positive correlation of 0.7 (p-value <2.2e-16) between the scores was observed (**Figure 23b**).

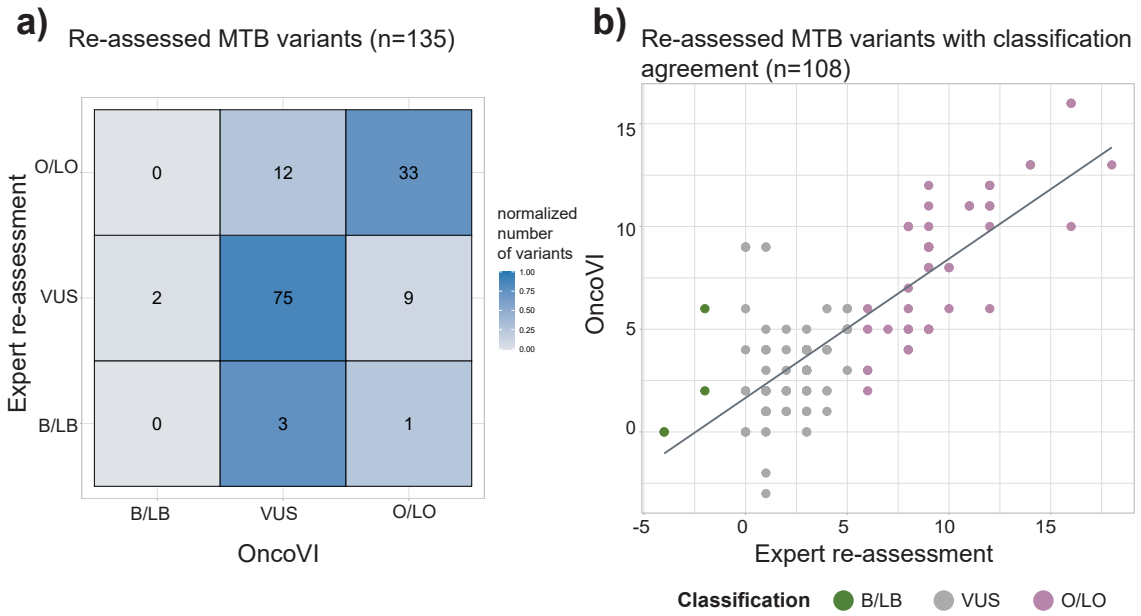

Figure 23: Results of OncoVI on the subset of MTB variants re-assessed by experts based on oncogenicity. a) Confusion matrix of the agreement between OncoVI and the expert re-assessment, based on the oncogenicity guidelines, in classifying the variants of the validation data set. The colour scale indicates the normalised number of variants, i.e., the ratio (calculated by row) between the number of variants of each cell and the total number of variants. **b)** Scatterplot between the points assigned by the experts and the points assigned by OncoVI for the 108 variants with classification agreement. B/LB: Benign/Likely Benign. VUS: Variant of uncertain significance. O/LO: Oncogenic/Likely Oncogenic. Figure panel modified from [102].

For the 33 O/LO variants with agreement between expert and OncoVI oncogenicity classification the variant-specific scores by OncoVI were the same or lower (**Figure 24a**), while for the 75 VUS variants with classification agreement the scores by OncoVI were the same or higher (**Figure 24b**). In summary, the concordance between OncoVI and the experts' re-assessment was also reflected by the assignment of comparable variant-specific scores.

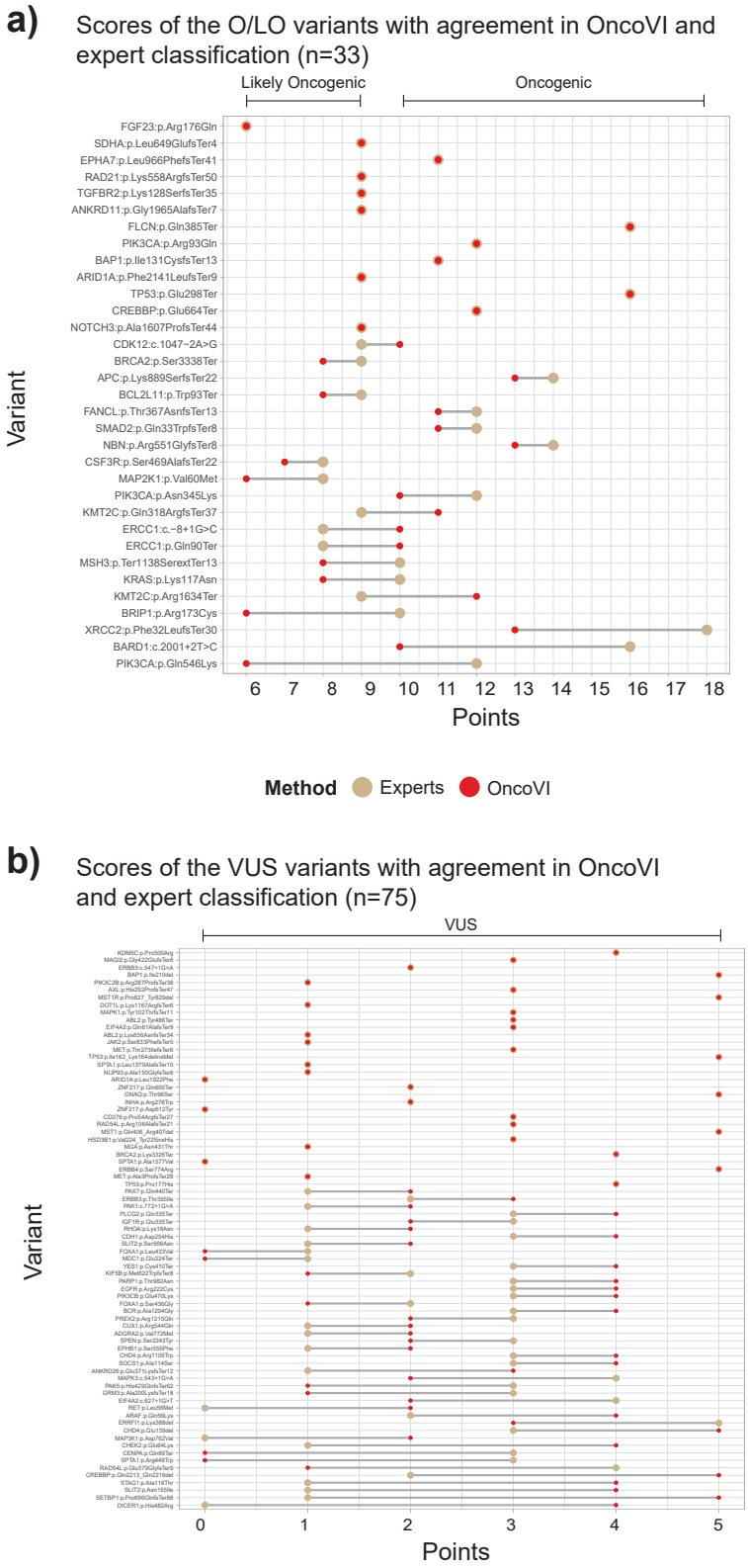

Figure 24: Variant-specific scores of the validation data set. a) Dumbell plot of the 33 variants classified as Oncogenic/Likely Oncogenic (O/LO) by both experts and OncoVI. **b)** Dumbell plot of the 75 variants classified as variant of uncertain significance (VUS) by both experts and OncoVI. Horizontal bars indicate the classification of the variants according to the SOP point-based system (i.e., score$\geq$ 10: Oncogenic, 6 $\leq$score$\leq$ 9: Likely Oncogenic, 0 $\leq$score$\leq$ 5: VUS). Figure panel modified from [102].

In addition, as in the SOP data set, the evaluation of the most frequently triggered criteria in the validation data set confirmed OncoVI ability to activate Strong oncogenic criteria, which play a key role in accurately classifying O/LO variants (**Figure 25a**, **Figure 25b**).

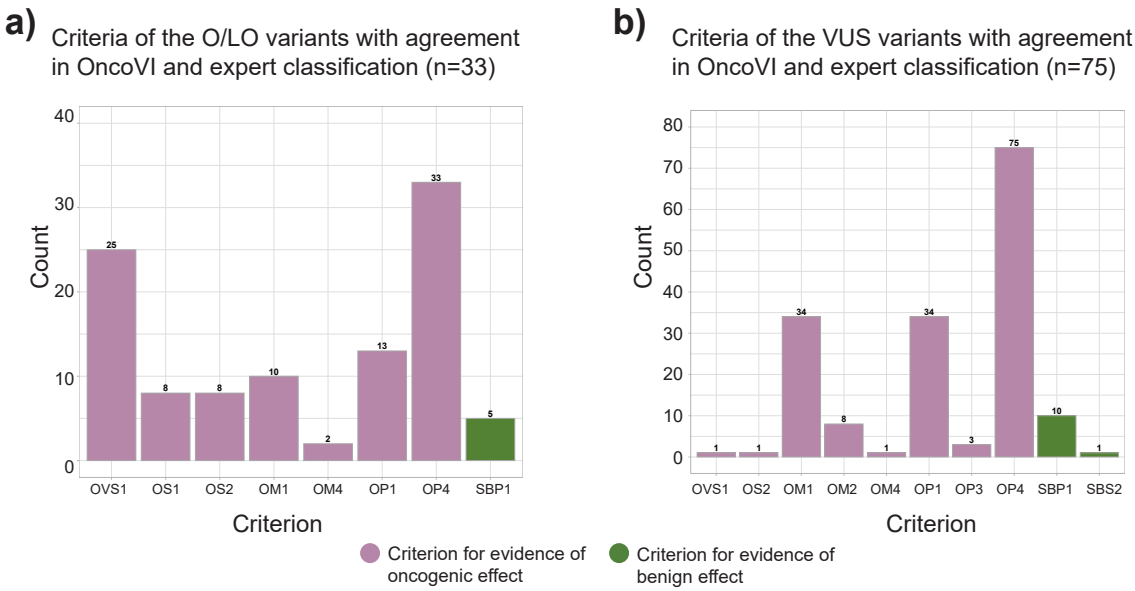

Figure 25: Criteria triggered by OncoVI on the variants of the validation data set. a) Barplot of the criteria triggered by OncoVI in the 33 variants classified as Oncogenic/Likely Oncogenic (O/LO) by both experts and OncoVI. **b)** Barplot of the criteria triggered by OncoVI in the 75 variants classified as variant of uncertain significance (VUS) by both experts and OncoVI. Criteria are sorted according to decreasing corresponding points. Figure panel modified from [102].

To assess how OncoVI criteria reflected expert interpretation, we examined the 108 variants with concordant oncogenicity classification in terms of criteria triggered by the experts and OncoVI (**Figure 26**). High concordance was observed for OP4 (triggered by both in 107 variants), OVS1 (26 variants), OS1 (7 variants), and OM2 (8 variants). Notably, in nine variants originally classified as VUS by the MTB but reclassified as O/LO by both experts and OncoVI, the criterion OVS1 (Very Strong, 8 points, *"Null variants in a bona fide TSG"*) was frequently triggered.

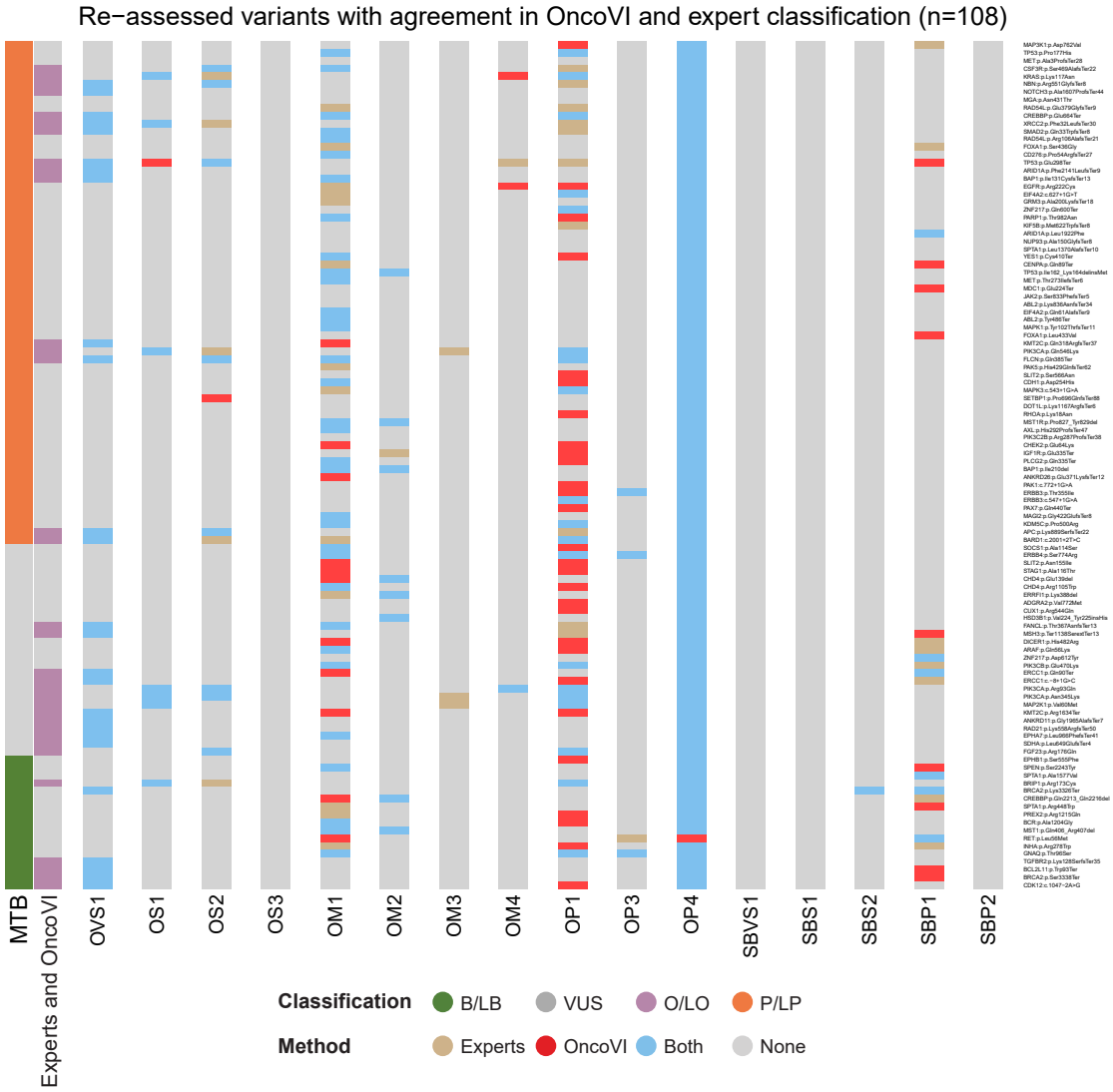

Figure 26: Re-assessed MTB variants with agreement between expert and OncoVI classifications. Each row corresponds to one of the 108 variants with agreement between expert and OncoVI classification, each column corresponds to an assessed criterion. The colour of the cell indicates whether a criterion was triggered by the experts only (dark gold), OncoVI only (red), both (light blue) or none of the two (grey). The barplot on the left indicates the classification of each variant according to the Molecular Tumour Board (MTB), OncoVI, and expert assessment. B/LB: Benign/Likely Benign, O/LO: Oncogenic/Likely Oncogenic, VUS: Variant of uncertain significance, P/LP: Pathogenic/Likely Pathogenic. Figure from [102].

In contrast, criteria such as OS2 (Strong, 4 points, *"Well-established in vivo or in vitro functional studies showing oncogenic effect of the variant"*) and SBS2 (Strong, -4 points, *"Well-established in vivo or in vitro functional studies showing no oncogenic effect of the variant"*) were more frequently applied by experts alone, reflecting their ability to resolve conflicting evidence in databases like ClinVar.

The strongest disagreement occurred for OM1 (Moderate, 2 points, *"Variant located in a well-established part of a functional domain"*), OP1 (Supporting, 1 point), and SBP1

(Supporting, -1 point) criteria, which rely on functional domain annotation or computational predictions (i.e., phyloP, phastCons, spliceAI). Collectively, these findings indicate that OncoVI and experts showed strong agreement for criteria not requiring interpretation of the resources (e.g., OP4, OVS1, OS1, OM2), while discrepancies arose mainly for criteria dependent on experts' interpretation of available resources.

4.3.4 Performance of OncoVI on the ClinVar data set

The ClinVar database has recently started providing oncogenicity classifications, based on manual curation by NCBI. In particular in April 2025, oncogenicity classification was available for 691 variants (11%, n=691/6,113; accessed 12.04.2025) involving 227 genes, classified as Benign (n=1), Likely Benign (n=1), Likely Oncogenic (n=373), Oncogenic (n=58), Uncertain significance (n=258). To further assess OncoVI performance, OncoVI was applied to classify these variants ("ClinVar data set") and the obtained oncogenicity classifications were compared with ClinVar assessment, considered as ground-truth.

OncoVI correctly classified 573 variants (n=573/691, 83%), with a sensitivity of 84% (n=362/431) for the O/LO class. No ClinVar O/LO variants were classified as B/LB by OncoVI or vice versa. Further investigations in the 69 O/LO variants misclassified by OncoVI as VUS revealed that the most triggered criteria by OncoVI were of category Supporting (1 point) or Moderate (2 points) (OP4: n=69/69, OP1: n=48/69, OM1: n=35/69). In addition, 55% (n=38/69) of the variants had variant-specific scores of 4 and 5, i.e., close to the upper limit of the score range corresponding to the class VUS. In addition, the majority (n=32/48) of the 48 variants classified by ClinVar as VUS and reclassified by OncoVI as O/LO had variant-specific scores of nine and ten, i.e., near the cut-off between the LO and O classes. Furthermore, in n=41/48 variants OncoVI triggered OVS1 (8 points) for null variants in a bona fide TSG due to VEP-annotated consequences typically associated with loss of protein function.

Taken together, OncoVI showed very good agreement with the oncogenicity classification in ClinVar.

4.4 Final considerations

This Chapter has demonstrated the potential of OncoVI as a reliable and reproducible tool to broaden the clinical adoption of the oncogenicity guidelines, further contributing to the standardisation of the oncogenicity classification. To implement the oncogenicity guidelines in OncoVI, efforts were required to define the valid biological meaning of each criterion, supported by consensus among expert biologists. In parallel, an exten-

sive screening of publicly available resources was conducted to identify the most suitable resource for each criterion.

OncoVI was validated on both gold-standard and real-world data sets of variants, namely, the SOP and MTB data sets, respectively. The validation revealed high agreement with manual classifications, as provided by the authors of the guidelines and the experts' evaluation based on the oncogenicity guidelines and variant impact on protein function, respectively. In addition, the analysis highlighted areas where OncoVI offered additional insights or exposed limitations in terms of chosen resources and guideline applicability.

The results presented in this Chapter underscore two key contributions. First, results showed that the interpretation of the biological meaning addressed by each criterion was accurate and the resources selected for the implementation of each criterion were appropriate. Indeed, this was confirmed by: (i) the strong agreement between SOP and OncoVI in the variants specific scores assigned to concordantly classified variants (although in general higher in OncoVI), and (ii) the assessment of the criteria triggered by OncoVI.

Second, OncoVI offers a scalable framework that can efficiently process large data sets compared to the manual assessment of the guidelines, while still leaving space for improvements. This emerged in the analysis on the validation data set, where the experts' assessment using a manual application of the guidelines showed a high agreement with OncoVI, for instance in classifying as O/LO variants previously assessed as VUS within the MTB, based on the variant effect on protein function. Such concordance between experts and OncoVI was largely driven by the frequent activation by both of the highest-scoring criterion (OVS1, 8 points) for nonsense variants. Follow-up investigations confirmed the original MTB classification originally based on the variant effect on protein function. Indeed, these nonsense variants were predominantly affecting the far terminal part of the protein, thus possibly leading to a functional expression. This result exposed the limitation of an automated application of this criterion for nonsense variants. In parallel, this result highlighted how misclassifications by OncoVI often reflected intrinsic differences in evaluating oncogenicity versus variant effect on protein function, as carried out in the MTB assessment. On the one hand, the comparison of the criteria triggered by OncoVI and the experts revealed strong alignment in those not requiring specific interpretation of the resources. On the other hand, assessment of some specific subcategory of variants revealed limitations in the automatic implementation of OncoVI with respect to expert assessment.

Together, these results suggest that OncoVI represents a reliable implementation of the oncogenicity guidelines, while reducing subjectivity in interpretation and supporting consistency in variant classification. In addition, OncoVI is available as open-source, Python-based tool, ensuring adaptability and integration into existing pipelines, and fos-

tering transparency and reproducibility in somatic variant classification. Furthermore, the systematic use of OncoVI may inform future revisions of the oncogenicity guidelines, ultimately contributing to the refinement of oncogenicity assessment frameworks and the advancement of precision oncology.

Chapter 5

Clinical deployment of OncoVI

The content of this Chapter is based on a manuscript available as preprint on medRxiv [120]:

Maria Giulia Carta*, Miriam Angeloni*, Lars Tögel, Christoph Schubart, Annett Hölsken, Robert Stoehr, Simona Vatrano, Davide Rizzi, Paolo Magni, Filippo Fraggetta, Arndt Hartmann, Florian Haller, Fulvia Ferrazzi. *A fully-automated integrative workflow to streamline NGS-based analyses within Molecular Tumour Boards.* medRxiv. 2025.12.12.25341897. doi: `https://doi.org/10.64898/2025.12.12.25341897`. *:shared first authors.

As outlined in the background (section 2.2.3), a major barrier to the implementation of precision oncology in clinical practice is the need for bioinformatics approaches that automatically interrogate existing knowledgebases for variant annotation, harmonised tools for oncogenicity classification, as well as for complex biomarkers evaluation. We aimed at overcoming this barrier and supporting variant interpretation in the clinical practice through the main contributions described in Chapter 3 and Chapter 4, which ultimately cumulated in the development of OncoVI.

Nevertheless, in precision oncology the typical bioinformatics analysis framework to process raw sequencing data to clinically meaningful variants is multi-steps and inherently complex. Also, it often requires the use of different tools. As a result, human intervention is necessary to interconnect these tools, e.g., to reformat the output of one tool to ensure compatibility with downstream tools. Thus, the overall bioinformatics framework remains fragmented, labour-intensive, and requires constant manual input.

To address the challenges of tools interoperability in NGS-based clinical decision-making within MTBs, this Chapter presents an end-to-end fully-automated workflow that

integrates variant interpretation, oncogenicity classification via OncoVI, and estimation of clinically relevant biomarkers, thus eliminating the need for human intervention. Using the TSO500+HRD gene panel as case study, the integrative solution was first developed and established at the Institute of Pathology UKER (Germany), and later adapted to the fully digitized Pathology Department at the Gravina Hospital in Caltagirone (Italy). The design and development of this fully-automated solution were carried out in close collaboration with Miriam Angeloni.

5.1 Background and motivation

In precision oncology, the curation and interpretation of clinically relevant alterations requires the application of a variety of tools and the consultation of diverse knowledgebases [121]. This process often involves experts running different software programs, querying heterogeneous sources, and reconciling outputs across the different steps before variant interpretation and clinical reporting. Because the interoperability between the differently tools is rarely guaranteed, outputs from tools often need to be manually reformatted to meet the input requirements of the downstream tools. As a result, human intervention cannot be avoided, leading to a fragmented and labour-intensive workflow. These limitations not only increase the workload of the healthcare team, but also prolong MTB preparation, potentially delaying treatment decisions, and compromising optimal patient care management [122].

The growing volume of NGS data, combined with associated clinical information, further underscores the need for digital solutions and visualisation strategies that can support data integration, reuse, and efficient interpretation [28, 121, 123]. Tools such as cBioPortal already facilitate genomic data integration and visualisation for treatment recommendations in patients with similar mutational profiles [124]; yet they require structured and standardised input. Without harmonised data formats, the exploitation and interoperability of such platforms remain limited, as manual transformations are often needed to align outputs with platform requirements. While automation could mitigate these issues, MTBs that rely on locally implemented and poorly integrated implementations remain heavily dependent on manual intervention, limiting scalability and efficiency [122].

Together these limitations highlight the need for automated and standardised solutions that can improve reproducibility, reduce manual workload, and accelerate MTB preparation. In addition, automation allows free up experts from repetitive data-handling tasks, such as reformatting, copying and pasting commands for tool execution, enabling them to focus on higher-level activities like variant interpretation and clinical decision-making.

To address these challenges, we developed a fully-automated, integrative workflow to

support NGS-based clinical decision-making within MTBs. The framework streamlines the entire process, from raw sequencing data through quality control, variant annotation, and oncogenicity classification. In addition, it supports the generation of outputs in specific data formats, suitable for the integration into visualisation platforms such as cBioPortal.

5.2 Methods

5.2.1 Erlangen workflow from sample pre-processing to MTB discussion

The Institute of Pathology UKER has established a robust infrastructure for NGS-based analyses, specifically utilised for cancer patients discussed in the MTB, dating back to 2016. Starting from February 2022, all MTB patients have been systematically tested using the TSO500+HRD gene panel, with an average of approximately 250 cases analysed per year. Starting from 2025 onwards, WES and WGS have been implemented for tumour molecular profiling as part of routine molecular diagnostics for patients who fulfil specified criteria by national guidelines.

The standard NGS-analysis workflow at the Institute of Pathology UKER begins with DNA extraction from a slide of FFPE tumour tissue using a dedicated kit according to the manufacturer's instructions (**Figure 27**).

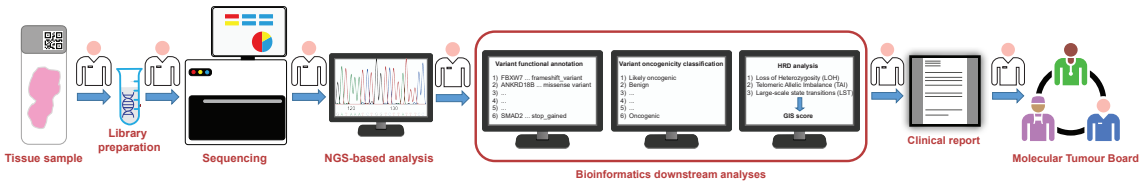

Figure 27: Standard workflow for NGS DNA-analysis at the Institute of Pathology UKER. DNA is extracted from FFPE tumour tissue, followed by in-house library preparation and sequencing on Illumina NextSeq500/550 platform. FASTQ generation, alignment to hg19, and SNVs calling are performed on the Illumina Connected Analytics cloud, while bioinformatics downstream analyses are run on a dedicated remote server. Automated pipelines for variant annotation, oncogenicity classification, and HRD scoring enable streamlined reporting for clinical interpretation and discussion in the Molecular Tumour Board.

Library preparation and sequencing are performed in-house, with sequencing carried out on the Illumina NextSeq500/550 sequencing platform. Once DNA libraries are prepared, the sequencing run is initiated by the technicians of the molecular diagnostics laboratory. Following sequencing, NGS data analysis is automatically performed on the Illumina Connected Analytics (ICA) cloud environment. On the ICA cloud environment, the following steps take place: (i) FASTQ files generation, (ii) reads alignment to the

human reference genome (hg19, UCSC), and (iii) SNVs calling. Instead, all the downstream bioinformatics analyses such as variant functional annotation and oncogenicity classification through OncoVI are conducted on a dedicated remote server. In addition, a pipeline developed for HRD scores estimation, an approved biomarker for PARP inhibitors treatment in HR-deficient tumours, is available for running.

5.2.2 Design and implementation of the integrative workflow

The fully-automated integrative workflow was developed in Python (version 3.10) to streamline the steps that require human intervention outlined in the previous section, i.e., from NGS data transfer to a remote server and to variant interpretation.

The workflow consists of four main steps: 1) following sequencing, a trigger event initiates the download of NGS data without manual intervention, 2) a sequencing quality report is automatically generated from the sequencing quality metrics, 3) downstream bioinformatics analyses are run automatically, and 3) analyses outputs including plots, text files, and tabular data are generated for interpretation by molecular biologists. At the Institute of Pathology UKER, the trigger event is implemented via a Python script that monitors the presence of new NGS data on the ICA cloud platform every 30 minutes (**Figure 28a**).

Upon detection of new NGS data in the Illumina cloud, the download of new NGS data is automatically performed relying on icav2 (version 2.34.0), the command line interface for the ICA cloud. Afterwards, the workflow is configured to automatically initiates all the downstream bioinformatics analyses specified in the previous section. All the analyses steps performed in the developed integrative workflow are run on a remote server with Ubuntu 20.04.6 LTS operating system and executed within dedicated conda environments (version 24.11.1) to manage dependencies and ensure reproducibility.

In Caltagirone Pathology Department, primary and secondary analyses of NGS data are performed on the Illumina DRAGEN remote server. The outputs generated by such analyses are manually downloaded by molecular biologists on a network-attached storage (NAS). The NAS is, in turn, mapped on a remote server where all the downstream bioinformatics analyses take place (**Figure 28b**). In this context, the trigger event for the integrative workflow is a request that comes from the molecular biologists through the anatomic pathology laboratory information system (AP-LIS) for an individual sample. Once the analysis request for a specific sample has been placed, all the samples analysed along with this sample are processed through the workflow, and analysis outputs are provided for the entire batch of samples.

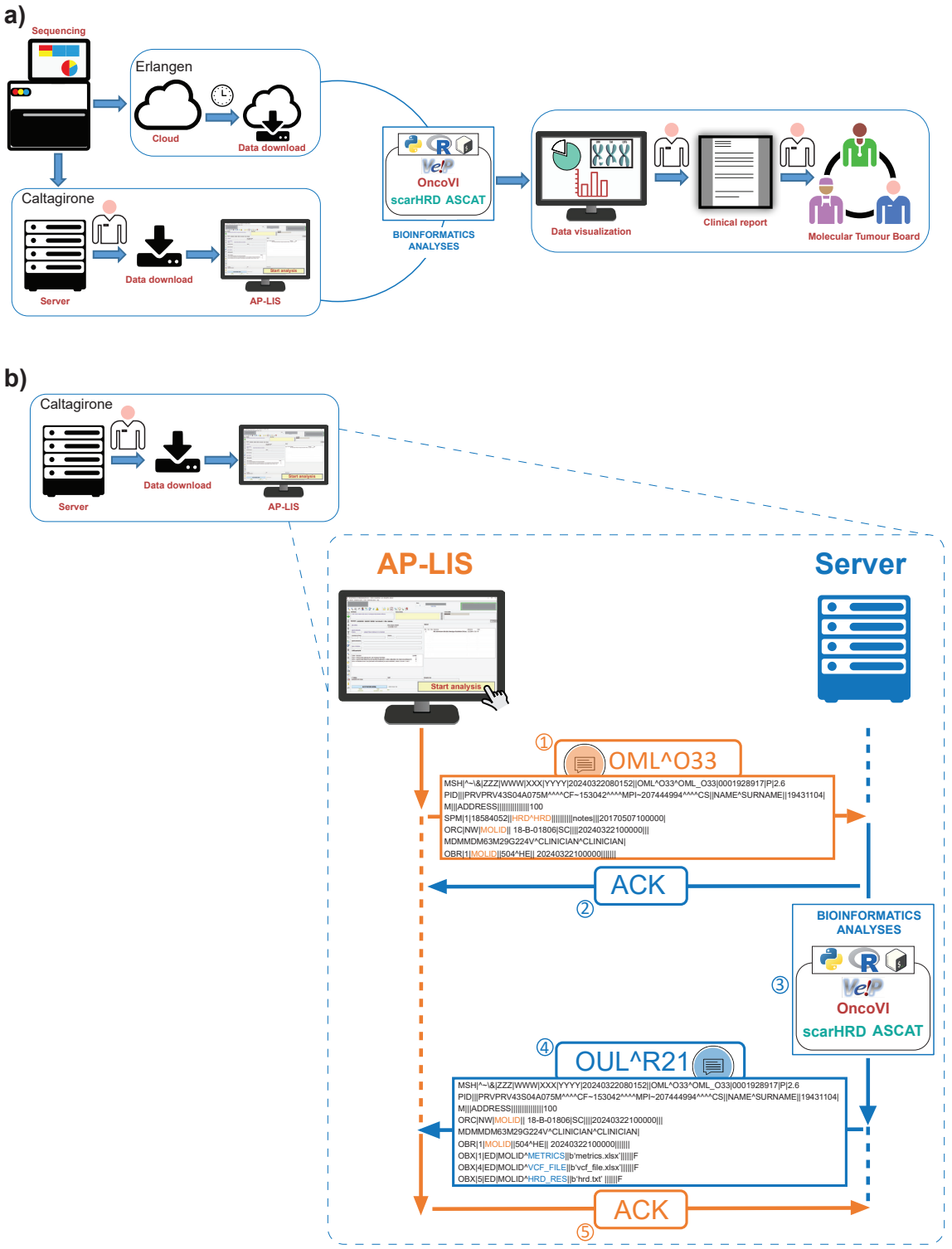

Figure 28: Configuration of the fully-automated integrative workflow. a) At the Institute of Pathology UKER (Erlangen), a python script continuously monitors for newly available NGS data on the ICA cloud every 30 minutes. Upon detection, the script automatically downloads NGS-data via icav2 and executes all downstream bioinformatics analyses without manual intervention. **b)** At the fully digitized Pathology Department of Caltagirone (Italy), NGS data are manually retrieved from the Illumina DRAGEN server. Subsequent bioinformatics analyses for a given sample of a sequencing run can be requested by molecular biologists through the AP-LIS via Health Level 7 (HL7) messaging, which triggers the following sequence of events: (1) a laboratory order message (OMLˆO33) is sent from the AP-LIS to the remote server storing NGS data; (2) an acknowledgment (ACK) message is sent from the remote server to the AP-LIS upon reception of the OMLˆO33 message; (3) the remote server processes the OMLˆO33 message. First, the molecular identifier of the patient (bold orange) is retrieved and all the downstream bioinformatics analyses are performed; (4) a laboratory observation message (OULˆR21) is sent from the remote server to the AP-LIS. Here, analysis results are transmitted to the AP-LIS as OBX segments (bold blue); (5) an ACK message is sent from the AP-LIS to the remote server upon reception of the OULˆR21 message. The figure panel depicting the HL7 connection between AP-LIS and server was adapted from [125].

5.2.3 Pipelines for variant interpretation integrated within the workflow

The developed integrative workflow consists of two main components: 1) OncoVI for variant functional annotation and oncogenicity classification as described in Chapter 3, and Chapter 4, 2) a custom pipeline for HRD scoring estimation.

Before steps 1) and 2) are performed, quality control sequencing metrics are collected for the samples analysed within the same batch. By default, pre-processing of NGS data from TSO500+HRD generates, for each sample, a "MetricsOutput.tsv" file containing quality indicators (e.g., the percentage of bases with BQ$\geq$ 30, the percentage of aligned reads, the mean target coverage). In the developed integrative workflow, to facilitate run-level evaluation, where a run represents a batch of eight samples prepared and sequenced simultaneously with the TSO500+HRD assay, sample-level quality control sequencing metrics are aggregated into a comprehensive table and provided as an Excel sheet.

Once the collection of the quality metrics has been completed, the workflow proceeds with variant functional annotation performed with the bioinformatics framework incorporated into OncoVI and detailed in Chapter 3. The input for this step is the "samplename_hard-filtered.vcf" file generated by DRAGEN, which contains the genomic positions of the detected variants for a given patient. The output is a human readable file containing the variants that passed all the filters applied by the pipeline and enriched with the information collected from the interrogated knowledgebases.

The variants retained from the previous step and enriched for additional information, undergo oncogenicity classification through OncoVI, as described in detail in Chapter 4. More specifically, OncoVI assigns each variant to one of five oncogenicity classes (i.e., "Oncogenic", "Likely Oncogenic", "VUS", "Likely Benign", "Benign") along with variant-specific score and triggered criteria. A final table is produced, containing both variant annotation and oncogenicity classification, to provide molecular biologists with a structured and interpretable data set, which serves as the basis for the following clinical interpretation and reporting.

Finally, to enable the integration of the most critical variants into our instance of cBioPortal, where variants previously classified as well as clinical data of our MTB patients are collected, the annotated VCF is converted to a mutation annotation format file (MAF file). The MAF file, i.e., a tab-delimited text file that associates each variant in the annotated VCF file to its most clinically relevant gene transcript (typically the one most severely affected), is obtained using the tool vcf2maf (conda package, version 1.6.21). The tool takes in input: 1) the VCF file annotated by VEP and filtered with the filter_vep utility (as described in section 3.1.1), and 2) the reference genome used for the alignment.

The resulting MAF file is then ready for integration into cBioPortal.

5.2.4 Pipeline for HRD scoring estimation

While the previous section describes the first part of the workflow that focuses on supporting SNVs interpretation, the second part assesses a clinically relevant biomarker, namely the HRD score.

The assessment of the HRD score is performed in R (conda package v. 4.2.2) [119] and consists of two steps: (i) inference of allele-specific copy number profiles of tumour only using ASCAT (version 3.1.1) [126] with a segmentation penalty set to 70, and (ii) estimation of HRD-associated genomic features using scarHRD (version 0.1.1) [127]. For each analysed sample, ASCAT takes in input the LogR ratios and B-allele frequency (BAF) values generated by DRAGEN. The first represents a normalised measure of DNA copy number, while the latter is used to infer allelic imbalance. To predict sample germline genotypes, the ASCAT function predictGermlineGenotypes() is run providing in input customised parameters specific for the TSO500+HRD gene panel and not available in the original version of the tool. These parameters were derived through the evaluation of the VAF distribution of an exemplary case without CNAs and set to the following values: proportionHetero= 0.29, proportionOpen= 0.03, and proportionHomo= 0.68 so that their sum is nearly 1. Additional parameters were set as follows: maxHomozygous= 0.05 and segmentLength= 100. The allele-specific copy number profiles generated by ASCAT are then used in input to scarHRD to estimate the HRD-associated genomic features, i.e., telomeric allelic imbalance (TAI) [128], loss of heterozygosity (LOH) [129], number of large-scale transitions (LST) [130]. The sum of the values associated with these three genomic features gives the total genomic instability score (GIS), which is provided in a tabular format by the pipeline, together with genomic instability plots produced by ASCAT.

5.2.5 Adaptation of the framework to the Caltagirone Pathology Department

Gravina Hospital Pathology Department in Caltagirone has recently adopted large gene panels for the molecular profiling of patients discussed in the MTB. Starting from November 2024, all cases have been systematically tested either using the TSO500+HRD or the TSO500 gene panel.

The workflow in Caltagirone partially differs from the workflow in Erlangen (section 5.2.1). Analogously to the Institute of Pathology UKER, in Caltagirone DNA extraction,

library preparation, and sequencing are performed in-house, with sequencing carried out on the Illumina NextSeq500/550 sequencing platform. Differently from Erlangen, Caltagirone relies on the Illumina DRAGEN remote server for primary and secondary analyses of NGS data, requiring an additional manual step for the download of NGS data to the NAS, where all the bioinformatics downstream analyses take place.

The developed integrative workflow was adapted to the fully digitized Italian Pathology Department by interconnecting the remote server to the AP-LIS via Health Level 7 (HL7) messaging through a Python-based server-client architecture as previously described [125]. Briefly, the communication between the AP-LIS and the remote server starts with an HL7 Laboratory Order Messages (OMLˆO33) request transmitted from the AP-LIS to the server. Conversely, the results of the fully-automated workflow are transmitted to the AP-LIS via HL7 Unsolicited Laboratory Observation Messages (OULˆR21) as OBX segments. The trigger event for the OMLˆO33 HL7 message is an analysis request placed by molecular biologists directly from the AP-LIS for the patient under evaluation. Communication between the AP-LIS and the remote server is set-up via intranet connection using socket programming, and HL7 messaging exchange takes place via a Transmission Control Protocol/Internet Protocol (TCP/IP) connection using the Minimal Lower Layer Protocol (MLLP) [125].

5.2.6 Evaluation of hands-on time

The efficacy of the fully-automated integrative workflow in reducing the overall time required for downloading NGS data and performing downstream bioinformatics analyses in MTB setting was evaluated at the Institute of Pathology UKER. Three key steps of the standard molecular diagnostics workflow were identified and the three cancer biologists of the Institute of Pathology UKER recorded the hands-on time required for each step. The steps were as follows:

1. Download of NGS data from the ICA cloud to the remote server used for downstream computational analyses;

2. Execution of OncoVI and HRD scoring estimation pipeline;

3. Generation of MAF files starting from VCF files.

The hands-on time recorded by each biologist for the aforementioned three steps was assessed across five routine TSO500+HRD runs, each comprising data from eight samples. For comparison, the corresponding times required by the integrative workflow to perform the same steps were also obtained.

The Friedman test was conducted to assess differences in timing among the three molecular biologists and the implemented integrative workflow. To this aim, the rstatix (version 0.7.2) package of R (version 4.5.1) was used. Boxplots showing timing distribution across raters for each step were obtained relying on the R package ggplot2 (version 3.5.2).

5.2.7 Survey design

To assess user satisfaction with the newly implemented integrative workflow, a structured survey was conducted among the three biologists of the Institute of Pathology UKER. The survey comprised a total of 12 statements aimed at evaluating molecular biologists' perception of the workflow in terms of: (i) automation impact and results accessibility, (ii) efficiency, (iii) impact on work, and (iv) overall satisfaction. Biologists were asked to rate each statement using a Likert scale (ranging from "Strongly Disagree" to "Strongly Agree"), addressing factors such as reduced manual intervention, clarity of results availability, elimination of redundant steps, improved turnaround time for variant interpretation, decreased troubleshooting effort, and overall productivity gains. Results from the survey were visualized through a heatmap representation. For each question and response category, the number of raters selecting that category (ranging from 0 to 3) was computed. The resulting 12 (questions) $\times$ 5 (Likert categories) matrix was visualized as a heatmap using the pheatmap R package (version 1.0.13).

5.3 Results

5.3.1 Establishment of the framework at the Institute of Pathology UKER

Since September 2021, at the Institute of Pathology UKER bioinformatics has gained increasing importance in the support of MTB. Notably, the implementation of OncoVI, described in Chapter 4, significantly improved the analysis workflow in the MTB setting. Although it was developed to standardise variant interpretation, until July 2025 molecular biologists still had to manually download NGS output from the ICA cloud environment and launch OncoVI on the remote server. When needed, the pipeline for HRD scoring estimation, as well as the pipeline for VCF to MAF conversion, had to be manually run.

Starting from July 2025, with the development of the fully-automated integrative workflow the MTB analysis framework now eliminates human intervention. Indeed, molecular biologists are no longer required to perform manual runs of downstream bioinformatics analyses, which further streamlines NGS-based clinical decision-making. Instead, molec-

ular biologists interpret the results and generate a comprehensive molecular report highlighting clinically relevant variants. These findings are subsequently discussed in the MTB, where their potential clinical implications are collectively evaluated for implementation into patient care.

5.3.2 Analysis outputs produced by the integrative workflow

The fully-automated integrative workflow designed to support and streamline NGS-based clinical decision-making within MTBs was developed and established at the Institute of Pathology UKER (Germany), and later adapted to the fully digitized Pathology Department of Caltagirone (Italy). In the German Pathology Institute, a custom Python-based script runs persistently in a detached session in the background on a remote server to monitor the ICA cloud every 30 minutes and detect the presence of new NGS data. Once new NGS data are found, their download as well as the initiation of the downstream bioinformatics pipelines are directly executed. To provide biologists with an overview of the sequencing metrics quality control, the metrics of all the samples belonging to the same sequencing run are collected into a unique tabular file, called "Metrics_Summary.xlsx" (**Figure 29**). In this table, rows correspond to the metrics calculated by DRAGEN and columns represent the analysed samples. A cell will be assigned to a "PASS" flag if the threshold set by DRAGEN for the specific metric in a given sample is reached (with the calculated value in brackets), otherwise a "FAILED" flag will be assigned to the cell and automatically marked in red to better highlight samples that failed that specific quality metric.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8
PCT_PF_READS (%) [80.0-nan]	PASS (93.2)	-	-	-	-	-	-	-
PCT_Q30_R1 (%) [80.0-nan]	PASS (94.7)	-	-	-	-	-	-	-
PCT_Q30_R2 (%) [80.0-nan]	PASS (91.9)	-	-	-	-	-	-	-
CONTAMINATION_SCORE (NA) [0-1457]	PASS (166)	PASS (152)	PASS (664)	PASS (445)	PASS (131)	PASS (231)	PASS (282)	PASS (417)
MEDIAN_INSERT_SIZE (bp) [70-nan]	PASS (114)	PASS (101)	PASS (97)	PASS (96)	PASS (111)	PASS (95)	PASS (102)	PASS (103)
MEDIAN_EXON_COVERAGE (Count) [150-nan]	PASS (389)	PASS (435)	PASS (291)	FAILED (19)	PASS (434)	PASS (284)	PASS (398)	PASS (335)
PCT_EXON_50X (%) [90.0-nan]	PASS (97.4)	PASS (97.9)	PASS (97.4)	FAILED (2.3)	PASS (97.5)	PASS (97.6)	PASS (97.5)	PASS (97.0)
USABLE_MSI_SITES (Count) [40-nan]	PASS (120)	PASS (119)	PASS (112)	FAILED (3)	PASS (120)	PASS (118)	PASS (118)	PASS (117)
GENE_SCALED_MAD (Count) [0.000-0.134]	PASS (0.088)	PASS (0.074)	PASS (0.079)	PASS (0.133)	PASS (0.085)	PASS (0.081)	PASS (0.084)	PASS (0.085)
MEDIAN_BIN_COUNT_CNV_TARGET (Count) [1.0-nan]	PASS (4.6)	PASS (5.4)	PASS (3.9)	FAILED (0.3)	PASS (5.1)	PASS (3.9)	PASS (5.1)	PASS (4.3)
PCT_TARGET_HRD_50X (%) [50.0-nan]	PASS (98.3)	PASS (98.6)	PASS (98.3)	FAILED (0.6)	PASS (98.8)	PASS (98.9)	PASS (97.8)	PASS (96.9)
EXCESSIVE_TF (NA) [nan-0]	PASS (0)	PASS (0)	PASS (0)	PASS (0)	PASS (0)	PASS (0)	PASS (0)	PASS (0)
MEDIAN_CV_GENE_500X (NA) [0.00-0.93]	PASS (0.62)	PASS (0.60)	PASS (0.71)	PASS (0.84)	PASS (0.62)	PASS (0.70)	PASS (0.64)	PASS (0.80)
TOTAL_ON_TARGET_READS (Count) [9000000-nan]	PASS (18026304)	PASS (18077758)	PASS (17454201)	PASS (18263180)	PASS (17627597)	PASS (16951841)	PASS (18425256)	PASS (16554704)
MEDIAN_INSERT_SIZE (Count) [80-nan]	PASS (108)	PASS (108)	PASS (98)	FAILED (70)	PASS (109)	PASS (99)	PASS (103)	PASS (91)
TOTAL_PF_READS (Count) [nan-nan]	PASS (20595204)	PASS (20658932)	PASS (20289684)	PASS (20528220)	PASS (20001510)	PASS (20211174)	PASS (21014802)	PASS (19957860)
MEAN_FAMILY_SIZE (Count) [nan-nan]	PASS (2.2)	PASS (1.9)	PASS (2.7)	PASS (38.8)	PASS (2.0)	PASS (2.7)	PASS (2.0)	PASS (2.4)
MEDIAN_TARGET_COVERAGE (Count) [nan-nan]	PASS (372.0)	PASS (409.0)	PASS (275.0)	PASS (18.0)	PASS (413.0)	PASS (273.0)	PASS (387.0)	PASS (328.0)
PCT_CHIMERIC_READS (%) [nan-nan]	PASS (0.42)	PASS (0.43)	PASS (0.50)	PASS (0.66)	PASS (0.39)	PASS (0.50)	PASS (0.48)	PASS (0.56)
PCT_EXON_100X (%) [nan-nan]	PASS (95.4)	PASS (96.6)	PASS (94.9)	PASS (0.1)	PASS (95.9)	PASS (95.7)	PASS (95.2)	PASS (94.0)
PCT_READ_ENRICHMENT (%) [nan-nan]	PASS (65.9)	PASS (62.8)	PASS (63.7)	PASS (59.7)	PASS (65.7)	PASS (63.8)	PASS (62.0)	PASS (63.8)
PCT_USABLE_UMI_READS (%) [nan-nan]	PASS (99.9)	PASS (99.9)	PASS (99.9)	PASS (99.9)	PASS (99.9)	PASS (99.9)	PASS (99.9)	PASS (99.9)
MEAN_TARGET_COVERAGE (Count) [nan-nan]	PASS (381.7)	PASS (412.0)	PASS (280.1)	PASS (19.6)	PASS (415.7)	PASS (280.7)	PASS (391.8)	PASS (329.5)
PCT_ALIGNED_READS (%) [nan-nan]	PASS (96.9)	PASS (95.3)	PASS (96.1)	PASS (96.9)	PASS (98.2)	PASS (96.5)	PASS (97.7)	PASS (96.6)
PCT_CONTAMINATION_EST (%) [nan-nan]	PASS (5.1)	PASS (1.5)	PASS (7.2)	PASS (22.9)	PASS (1.6)	PASS (1.6)	PASS (3.2)	PASS (6.6)
PCT_TARGET_0.4X_MEAN (%) [nan-nan]	PASS (91.1)	PASS (92.2)	PASS (91.6)	PASS (91.1)	PASS (91.4)	PASS (93.3)	PASS (89.1)	PASS (89.4)
PCT_TARGET_50X (%) [nan-nan]	PASS (96.8)	PASS (97.2)	PASS (96.5)	PASS (2.5)	PASS (96.9)	PASS (97.0)	PASS (96.8)	PASS (96.5)
PCT_TARGET_100X (%) [nan-nan]	PASS (94.5)	PASS (95.3)	PASS (93.0)	PASS (0.2)	PASS (94.9)	PASS (94.4)	PASS (94.1)	PASS (93.0)
PCT_TARGET_250X (%) [nan-nan]	PASS (77.0)	PASS (83.2)	PASS (55.1)	PASS (0.0)	PASS (82.7)	PASS (57.2)	PASS (75.4)	PASS (65.3)
PCT_SOFT_CLIPPED_BASES (%) [nan-nan]	PASS (3.58)	PASS (4.05)	PASS (4.20)	PASS (2.62)	PASS (2.38)	PASS (4.45)	PASS (4.09)	PASS (4.39)
PCT_Q30_BASES (%) [nan-nan]	PASS (94.01)	PASS (93.23)	PASS (93.82)	PASS (94.24)	PASS (94.74)	PASS (93.75)	PASS (94.54)	PASS (94.05)
ALLELE_DOSAGE_RATIO (NA) [nan-nan]	PASS (0.495)	PASS (0.647)	PASS (0.354)	PASS (nan)	PASS (0.528)	PASS (0.374)	PASS (0.267)	PASS (0.383)
MEDIAN_TARGET_HRD_COVERAGE (Count) [nan-nan]	PASS (141.0)	PASS (141.0)	PASS (135.0)	PASS (14.0)	PASS (156.0)	PASS (134.0)	PASS (154.0)	PASS (141.0)
PCT_CHIMERIC_READS (%) [nan-nan]	PASS (0.42)	PASS (0.43)	PASS (0.50)	PASS (0.66)	PASS (0.39)	PASS (0.50)	PASS (0.48)	PASS (0.56)
PCT_ON_TARGET_READS (%) [nan-nan]	PASS (87.5)	PASS (87.5)	PASS (86.0)	PASS (89.0)	PASS (88.1)	PASS (83.9)	PASS (87.7)	PASS (82.9)
SCALED_MEDIAN_GENE_COVERAGE (Count) [nan-nan]	PASS (3706.9)	PASS (3729.4)	PASS (3314.1)	PASS (2998.7)	PASS (3596.6)	PASS (3205.4)	PASS (3607.1)	PASS (2803.9)
TOTAL_PF_READS (Count) [nan-nan]	PASS (20595204)	PASS (20658932)	PASS (20289684)	PASS (20528220)	PASS (20001510)	PASS (20211174)	PASS (21014802)	PASS (19957860)
GENE_MEDIAN_COVERAGE (Count) [nan-nan]	PASS (285.35)	PASS (229.21)	PASS (142.58)	PASS (28.13)	PASS (226.00)	PASS (122.71)	PASS (194.11)	PASS (100.49)
GENE_ABOVE_MEDIAN_CUTOFF (Count) [nan-nan]	PASS (46)	PASS (46)	PASS (36)	PASS (22)	PASS (45)	PASS (36)	PASS (44)	PASS (29)
PCT_SOFT_CLIPPED_BASES (%) [nan-nan]	PASS (3.58)	PASS (4.05)	PASS (4.20)	PASS (2.62)	PASS (2.38)	PASS (4.45)	PASS (4.09)	PASS (4.39)
RNA_PCT_Q30_BASES (%) [nan-nan]	PASS (91.41)	PASS (90.57)	PASS (89.89)	PASS (92.33)	PASS (92.86)	PASS (88.44)	PASS (90.16)	PASS (88.29)

Figure 29: Overview of the metrics summary file. Example of a metrics summary file ("Metrics_Summary.xlsx") storing the sequencing metrics, as calculated by DRAGEN, for all the samples belonging to the same sequencing run. Metrics that failed quality thresholds imposed by DRAGEN are highlighted in red.

After quality control, OncoVI and the pipeline for HRD scoring estimation are executed, and the results are stored locally on a predefined folder path directly accessible to the biologists. The bioinformatics workflow incorporated into OncoVI (described in Chapter 3) for variant functional annotation and knowledgebases interrogation produces an annotation table with the extension "_prediction_vep.xlsx". In this file, each row corresponds to a variant that passed the filters applied by the pipeline, while the columns provide curated information from the interrogated knowledgebases. Afterwards, the oncogenicity evaluation of the variants performed by OncoVI is stored in a tabular file with extension "_OncoVI_prediction.xlsx". This file reports for each variant in the "_prediction_vep.xlsx" file the variant-specific score, the oncogenicity classification, and the triggered criteria contributing to the variant-specific score. Upon OncoVI assessment completion, the generation of the MAF files takes place using as input for each sample the "filter_vep_samplename_dna_filtered.vcf" file, namely the VCF file annotated by VEP before the conversion to the human readable format.

Once the MAF files are generated, the pipeline for HRD scoring estimation is executed and several plots provided by ASCAT are made available to molecular biologists to support the evaluation of the patient's tumour genomic instability. These plots include: (i) the LogR and BAF tracks, (ii) the allele-specific copy number profile of the tumour, and (iii) the sunrise plot (**Figure 30**). Specifically, the sunrise plot shows tumour purity

and ploidy estimates generated by ASCAT. The green cross represents the best purity-ploidy combination identified by ASCAT that fits the observed LogR and BAF data. Furthermore, the total HRD score as well as TAI, LOH, and LST values are made available in the text file named "HRDresults.txt".

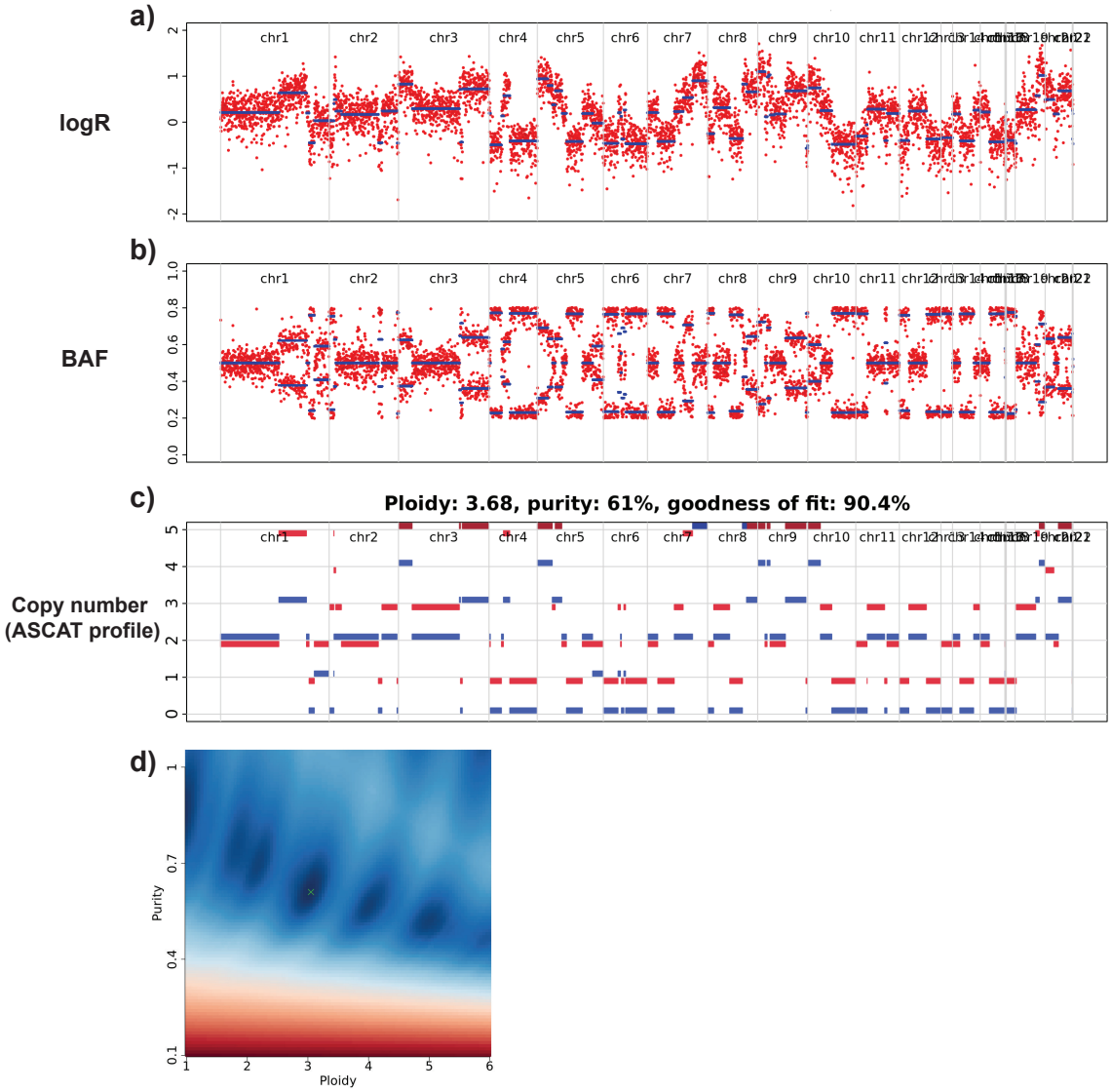

Figure 30: Plots generated by the HRD scoring estimation pipeline for a given sample. a, b) The LogR and BAF tracks provide cleaned genome-wide signals of a sample's copy number variation. **c)** From these tracks, the allele-specific copy number profile is generated by ASCAT and the major and minor copy number across the genome are summarised in two different colours (blue and red). **d)** In addition, the estimation of the tumour purity and ploidy made by ASCAT is plotted in the sunrise plot.

The framework was additionally implemented in the fully digitized Italian Pathology Department of Caltagirone (Italy). Here, an HL7-based connection was used to make analysis requests from the AP-LIS for the processing of new NGS runs. Analysis results

are directly accessible from the AP-LIS as shown below (**Figure 31**).

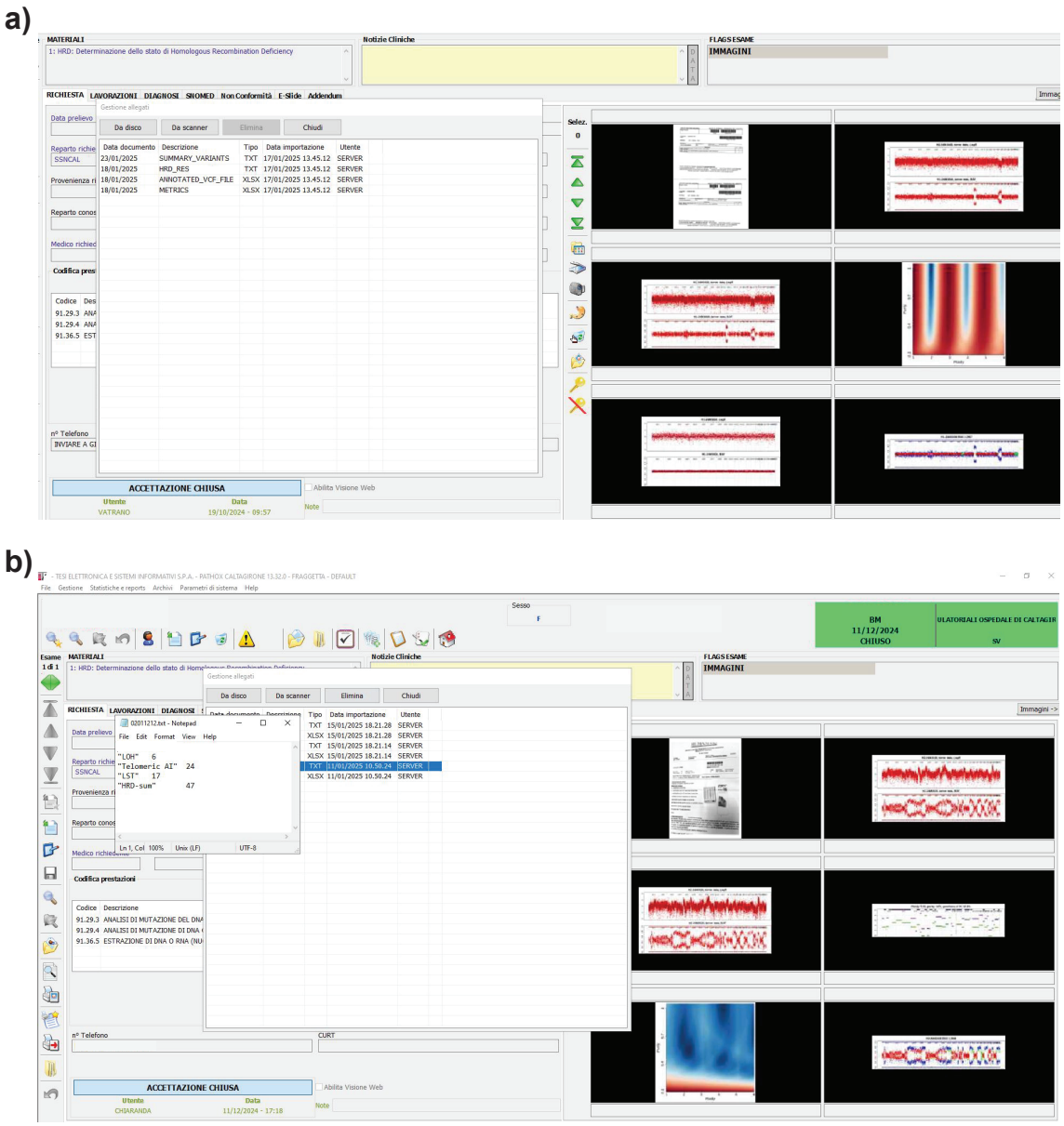

Figure 31: Visualisation of the output produced by the integrative workflow in the AP-LIS. Example of analyses outputs visualisation in the anatomic pathology laboratory information system (AP-LIS) of the Caltagirone Pathology Department. The text and tabular files generated by the workflow are available as attachments and can be opened with a double click (left), whereas all the plots are made available in the patient's gallery (right).

Taken together, the developed integrative framework allows a fully automated execution of all the downstream bioinformatics analyses performed by molecular biologists in the context of MTBs. This allows the complete removal of any required human intervention, thus making the results directly available for clinical interpretation and generation of the final report. In addition, the integrative workflow shows flexibility to different

pathology environments with distinct informatics set-ups such as in the case of HL7-based communication setting.

5.3.3 Hands-on time comparison between the automated workflow and the three molecular biologists

To systematically evaluate the efficiency of the proposed fully-automated integrative workflow, the time taken by the three molecular biologists of the Institute of Pathology UKER to manually perform three pre-defined analysis steps automated by the workflow (i.e., (1) download of NGS data from the ICA cloud to a remote server, (2) run of OncoVI and HRD scoring estimation pipeline, and (3) MAF files generation) was compared to the execution time of the workflow itself. This assessment was conducted and repeated for five independent TSO500+HRD routine sequencing runs.

The first step refers to the download of raw NGS data from the ICA cloud to the remote server used for subsequent analyses. The manual download of raw NGS data from the ICA cloud to the local server used for downstream bioinformatics analyses required a median 12.73, 10.33, and 14.55 minutes, respectively, for the three biologists performing the procedure (**Figure 32a**). In contrast, the integrative workflow exhibited a markedly lower execution time, with a median of 9.69 minutes across the five runs. Statistical evaluation confirmed that this difference was significant (Friedman test, $p = 0.026$), thereby demonstrating the superior efficiency of the automated approach in the data acquisition phase.

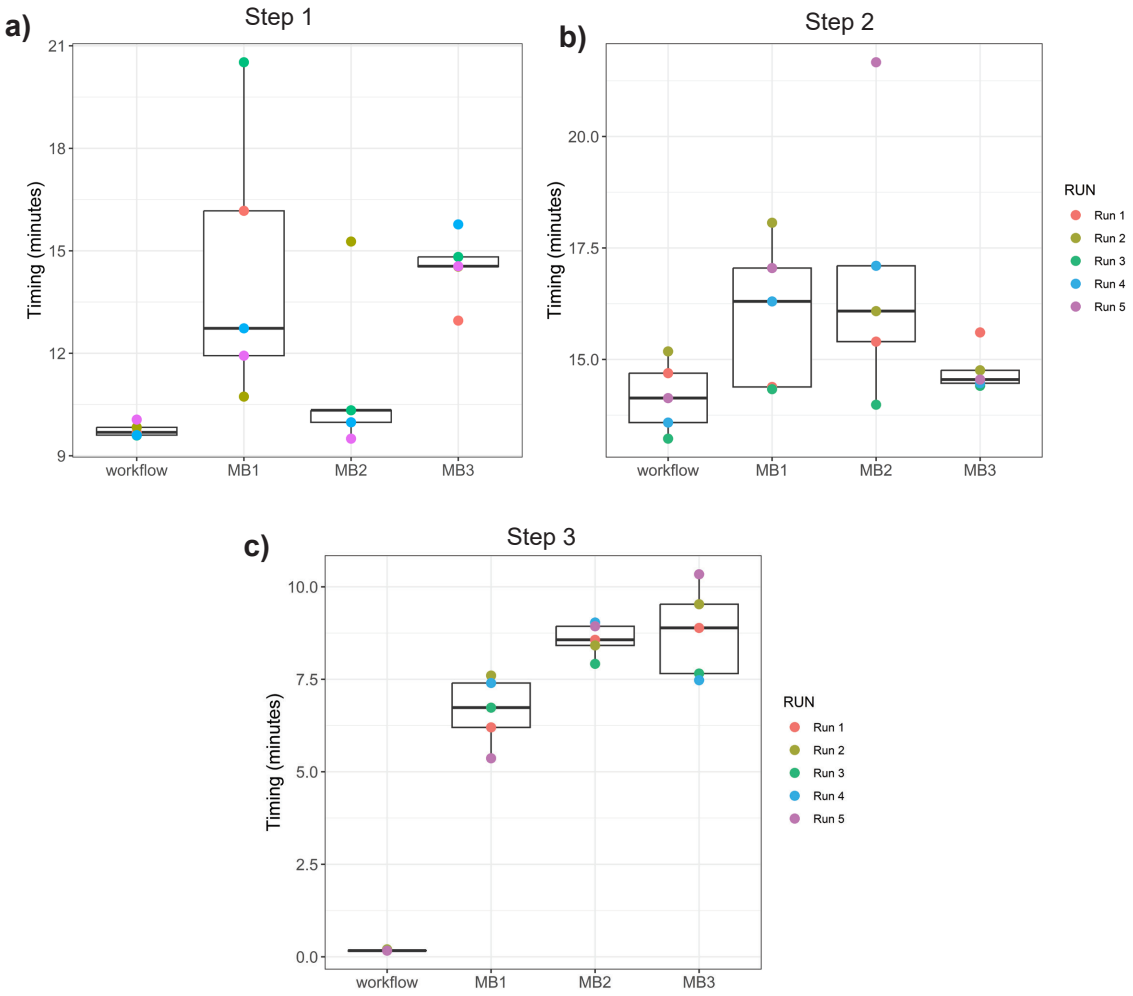

Figure 32: Hands-on time comparison between the automated workflow and the three molecular biologists at the Institute of Pathology UKER. Boxplot distribution of the time registered across five different sequencing runs to: **a)** download NGS data from the ICA cloud to the remote server, **b)** run the downstream bioinformatics analyses in the remote server; and **c)** generate the MAF files starting from the VCF file. MB = molecular biologist.

In the execution of OncoVI and HRD scoring estimation pipeline, the integrative workflow achieved a median execution time of 14.13 minutes, while the biologists took a median time of 16.30, 16.08, and 14.55 minutes, respectively. Although the automated framework demonstrated a slight advantage, statistical analysis did not reveal a significant difference between manual and automated execution of this step (Friedman test, $p = 0.14$) (**Figure 32b**).

The third step, involving the transformation of VCF files into MAF files, showed a marked difference between manual and automated execution. Indeed, the biologists spent a median of 6.73, 8.57, and 8.89 minutes, respectively, for the conversion of all samples of a given run. In contrast, the integrative workflow completed the task in a median of 0.17 minutes. The Friedman test confirmed the statistical significance (Friedman test, p =

0.0036), underscoring the substantial efficiency gain provided by the automated approach (**Figure 32c**).

Collectively, the comparison shows how the automation implemented by the integrative workflow conferred a tangible improvement in terms of time efficiency compared to the manual analysis steps.

5.3.4 Survey results

To evaluate the user satisfaction with the adoption of the fully-automated integrative workflow in the clinical practice, the three biologists of the Institute of Pathology UKER were asked to participate into a survey aimed at evaluating the developed integrative solution in terms of: (i) automation impact and results accessibility, (ii) efficiency, (iii) impact on work, (iv) overall satisfaction. When evaluating the impact of workflow automation and the accessibility of the results (**Figure 33**), all participants agreed that the automated framework requires no human intervention and delivers results in an intuitive way.

With respect to workflow efficiency, three participants strongly agreed that the automated workflow reduces repetitive tasks required in the non-automated solution and agreed that it saves time during variant interpretation. Regarding the impact on their daily work, three participants agreed in reporting that no troubleshooting was longer needed compared to the previous experience and strongly agreed that continuous monitoring of the analysis was no longer necessary, leading to a positive effect on their overall productivity. Finally, in terms of overall satisfaction, all responders indicated that they were highly satisfied with the benefits gained with the integrative workflow and consider it as an improvement over the previous solution. Moreover, all participants recognized its potential usefulness for colleagues. Collectively, these findings demonstrate that the implementation of the integrative framework in the clinical practice was well-accepted by the users, enhanced overall efficiency, and had a positive impact on their workload.

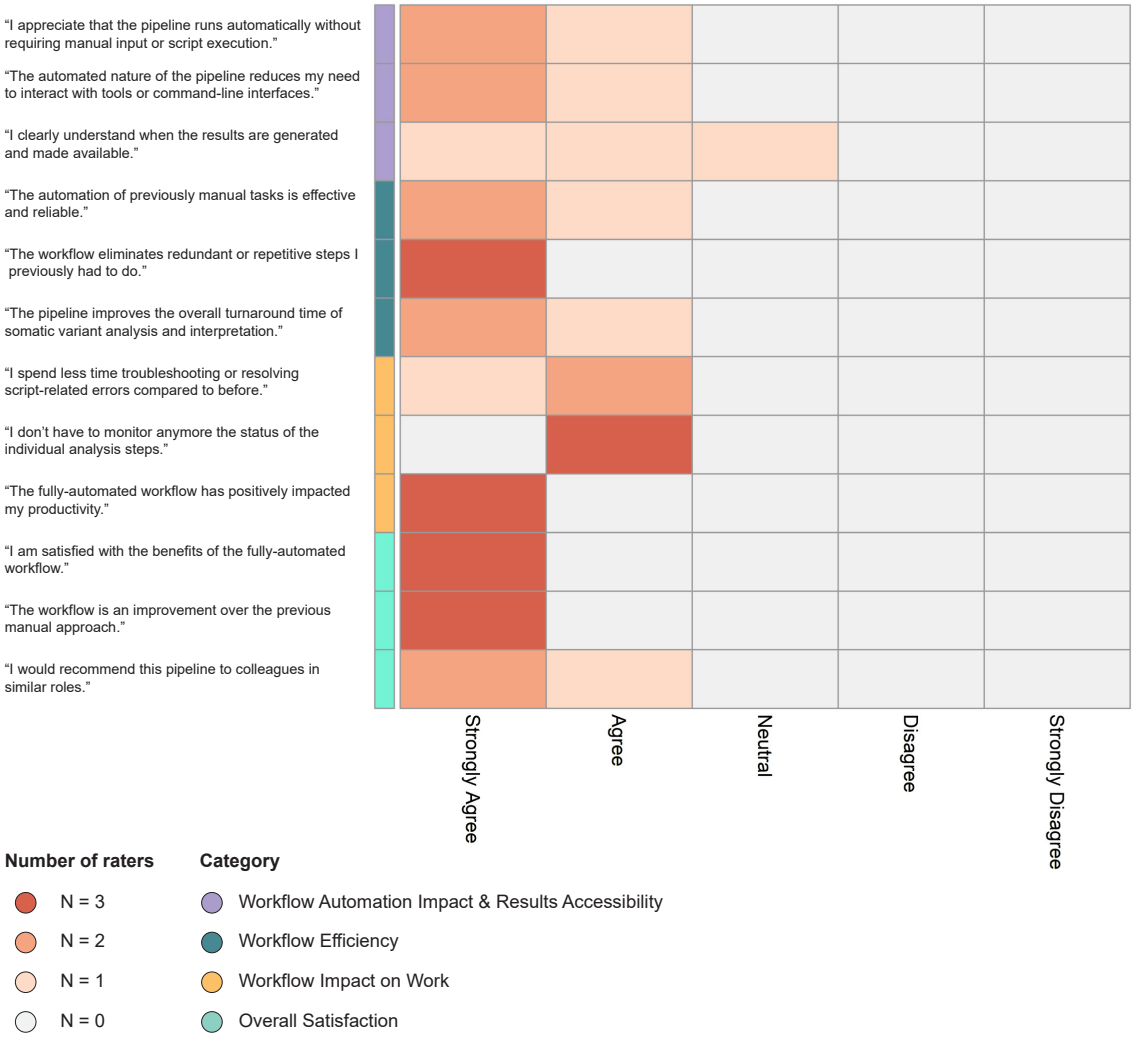

Figure 33: Heatmap representation of survey responses. For each question (rows) and response category (Strongly Agree-Strongly Disagree) (columns), the colour intensity reflects the number of raters (0–3) who selected a given Likert response for each question, with light grey indicating 0 raters and dark red indicating 3 raters.

5.4 Final considerations

This Chapter presented the implementation and evaluation of a fully-automated integrative workflow for the harmonised analysis of NGS data in two different clinical settings: the analogue Institute of Pathology UKER (Germany) and the fully digitized Pathology Department of Caltagirone (Italy). By automating and streamlining critical steps of the data management and bioinformatics analyses in MTB settings, the workflow removed the need for manual intervention on: (i) monitoring and downloading sequencing data, (ii) executing downstream analyses, such as OncoVI and the developed pipeline for HRD scoring estimation. In addition, when adapted to the fully digitized Italian Pathology Department, the developed integrative strategy allowed the direct visualization of outputs

within the AP-LIS, thus ensuring a convenient and rapid access to well-organised results for quality control, variant interpretation, and HRD scoring.

A systematic comparison conducted on five analysis runs between the developed framework and the three molecular biologists of the Institute of Pathology UKER in terms of hands-on time highlighted a strong reduction in time with the adoption of the automated solution. Indeed, compared to the manual execution, the workflow significantly reduced the time required for data retrieval from the cloud and MAF file generation. Although not statistically significant, a reduced time could also be observed for step 2, i.e., the execution of OncoVI and the HRD scoring estimation pipeline. However, these analyses had already been integrated in a larger automated standalone process, which might explain the non-significant results. Nevertheless, the overall reduction in terms of manual effort compared to the conventional approach was substantial. Indeed, benefits on the developed framework were supported by the insights emerged from the user feedback. Biologists found the framework intuitive, efficient, and beneficial for reducing workload.

Previous studies on digital strategies in MTBs have shown how automation reduces preparation and discussion times, standardises documentation, and facilitates case presentation [122]. In line with these findings, our integrative framework demonstrates how automated solutions can shift the focus of experts from repetitive and error-prone tasks, such as manually executing commands, toward interpretation and decision-making. For example, the collection of the individual sequencing metrics in one single output per run and the automated oncogenicity classification provided by OncoVI, represent key advantages over conventional methods that rely on fragmented and non-standardised processes.

The adoption of the developed framework in the Caltagirone fully digitized Pathology Department underscores its flexibility and real-world applicability. Nonetheless, some limitations must be acknowledged. The workflow was tested on a limited number of sequencing runs derived from an assay common to both institutions. This choice was intentional, as it enabled the development and validation of the workflow in a realistic clinical setting using a widely adopted gene panel. Applicability to diagnostics frameworks including WES and WGS analyses is also possible, although performance metrics were not assessed in this context. In addition, the framework's efficiency was not evaluated under varying hardware configurations or across heterogeneous computing infrastructures, which could influence runtime. To facilitate broader adoption by other pathology laboratories, future work should therefore focus on further generalising these components. Moreover, the Caltagirone implementation benefited from the full interoperability with the AP-LIS, while Erlangen setup lacked this integration. Future efforts should be focused on investigating solutions that enhance interoperability also in the Erlangen Pathology Department, allowing access to the framework results directly from patient records. Furthermore, the

connection of this workflow with additional digital solutions increasingly used in MTB settings such as cBioPortal should be explored.

Overall, the strategy presented in this Chapter reduced manual effort and automated molecular analyses in the clinical practice while standardising outputs, also thanks to the integration of OncoVI. We believe that the integrative workflow will offer a valuable solution to support decision-making within MTBs, with the potential to be extended to more diverse type of NGS data analyses and hospital settings.

Chapter 6

Conclusions

This PhD thesis was motivated by the real-world needs of MTBs to achieve reproducible and standardised interpretation of somatic variants identified from NGS-based tumour molecular profiling. While NGS-based tumour molecular profiling provides technical foundation for precision oncology, its routine applicability still faces significant limitations [121]. Indeed, downstream analyses and interpretation of these data remain time-consuming, inconsistent across institutions, and dependent on expert manual curation, thereby limiting actionable findings from being translated into patient care.

To this aim, this PhD work explored the opportunities of developing bioinformatics approaches to support harmonised annotation and oncogenicity classification of somatic variants as presented in Chapter 3 and Chapter 4, ultimately aiming at their integration in the clinical practice (Chapter 5).

First, a reproducible bioinformatics framework to systematically apply robust filtering and annotate variants according to their impact on protein function was developed. When applied to a retrospective (MIER) cohort of advanced urothelial carcinoma, the pipeline reduced the initial number of variants to a manageable set of biologically meaningful coding variants, thus streamlining expert review and improving reproducibility, compared to manual curation. In addition, the developed bioinformatics framework enabled the identification of potentially clinically relevant alterations, associated with existing therapeutic indications in our real-world MTB cohort of urothelial carcinoma. However, the practical implementation of these findings within the MTB cohort was limited, underscoring how regulatory, financial, and organisational factors continue to shape the impact of precision oncology beyond the purely technical challenges [76, 100, 101]. Nevertheless, the framework represented a tangible contribution to bridging the gap between the generation of genomic data and clinical decision-making in MTBs.

Building on this foundation, OncoVI was developed as an open-source Python-based

tool to automate the oncogenicity classification based on internationally-recognised guidelines (ClinGen/CGC/VICC) [30]. By translating guideline criteria into IF-ELSE rules and integrating external publicly-available resources, OncoVI reduced subjectivity, harmonised classifications across institutions, thus representing a concrete step toward more reproducible variant interpretation. When evaluated on both gold-standard and real-world data sets of somatic variants, OncoVI achieved strong concordance with classifications both by guideline authors' and our expert biologists' evaluations, respectively, thereby confirming its correct implementation and use of resources. Yet, limitations emerged where guideline application required human judgement, such as the evaluation of evidence from resources. These findings illustrated both the utility and the boundaries of computational approaches.

Finally, to evaluate the impact of the developed tools in the clinical practice, this PhD thesis work focused on deploying OncoVI as well as an additional pipeline for complex biomarker estimation into two distinct pathology environments: a conventional analogue Pathology Department at the Institute of Pathology UKER (Germany) and a fully digitized Pathology Department in Caltagirone (Italy). In these contexts, OncoVI was integrated into a broader framework aiming at automating the entire process, from NGS data download to downstream bioinformatics analyses. By connecting together NGS data transfer, variant annotation, oncogenicity classification, HRD scoring, and VCF to MAF conversion into a single framework, the fully-automated integrative solution significantly reduced biologists' manual effort, accelerated turnaround times, and saved time potentially usable for clinical reporting. The deployment in both analogue and fully digitised pathology departments demonstrated the adaptability and real-world relevance of the framework. At the same time, limitations were identified pointing to key areas for further refinement. To this aim, future work should follow the direction of the improvement of interoperability and scalability, ensuring automated workflows can be adopted across heterogeneous clinical environments, including resource-limited settings.

In conclusion, the bioinformatics tools developed during this PhD work represent concrete steps toward reducing subjectivity, improving reproducibility, and embedding automated workflows within routine clinical frameworks, thereby moving precision oncology closer to routine clinical practice.

Bibliography

[1] T. Hu, N. Chitnis, D. Monos, & A. Dinh. Next-generation sequencing technologies: An overview. *Hum Immunol*, 82(11):801–811, 2021.

[2] S. Goodwin, J. D. McPherson, & W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–51, 2016.

[3] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, & T. Shafee. Transcriptomics technologies. *PLoS Comput Biol*, 13(5):e1005457, 2017.

[4] M. Aldea, D. Vasseur, A. Italiano, & S. I. Nikolaev. WGS/WES-RNAseq compared to targeted NGS in oncology: is there something to unlock? *Ann Oncol*, 34(12): 1090–1093, 2023.

[5] E. R. Malone, M. Oliva, P. J. B. Sabatini, T. L. Stockley, & L. L. Siu. Molecular profiling for precision cancer therapies. *Genome Med*, 12(1):8, 2020.

[6] M. M. Clark, Z. Stark, L. Farnaes, T. Y. Tan, S. M. White, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med*, 3:16, 2018.

[7] Marta Gwinn, Duncan MacCannell, & Gregory L. Armstrong. Next-Generation Sequencing of Infectious Pathogens. *JAMA*, 321(9):893–894, 2019.

[8] H. Satam, K. Joshi, U. Mangrolia, S. Waghoo, G. Zaidi, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)*, 12 (7), 2023.

[9] I. B. Weinstein & K. Case. The history of Cancer Research: introducing an AACR Centennial series. *Cancer Res*, 68(17):6861–2, 2008.

[10] Z. Yu, T. H. H. Coorens, M. M. Uddin, K. G. Ardlie, N. Lennon, & P. Natarajan. Genetic variation across and within individuals. *Nat Rev Genet*, 25(8):548–562, 2024.

[11] T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, et al. International network of cancer genome projects. *Nature*, 464(7291):993–8, 2010.

[12] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–4, 2011.

[13] C. C. Guo, S. Lee, J. G. Lee, H. Chen, M. Zaleski, et al. Molecular profile of bladder cancer progression to clinically aggressive subtypes. *Nat Rev Urol*, 21(7):391–405, 2024.

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

[15] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871): 530–6, 2002.

[16] A. Kamoun, A. de Reyniès, Y. Allory, G. Sjödahl, A. G. Robertson, et al. A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. *Eur Urol*, 77(4):420–433, 2020.

[17] J. Griffin, J. Down, L. A. Quayle, P. R. Heath, E. A. Gibb, et al. Verification of molecular subtyping of bladder cancer in the GUSTO clinical trial. *J Pathol Clin Res*, 10(2):e12363, 2024.

[18] E. Cuppen, O. Elemento, R. Rosenquist, S. Nikic, I. Jzerman M, et al. Implementation of Whole-Genome and Transcriptome Sequencing Into Clinical Cancer Care. *JCO Precis Oncol*, 6:e2200245, 2022.

[19] W. Walter, N. Pfarr, M. Meggendorfer, P. Jost, T. Haferlach, & W. Weichert. Next-generation diagnostics for precision oncology: Preanalytical considerations, technical challenges, and available technologies. *Semin Cancer Biol*, 84:3–15, 2022.

[20] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*, 6(5):377–82, 2009.

[21] Y. Wang, J. Waters, M. L. Leung, A. Unruh, W. Roh, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513): 155–60, 2014.

[22] C. G. Williams, H. J. Lee, T. Asatsuma, R. Vento-Tormo, & A. Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Med*, 14(1):68, 2022.

[23] B. Kinnersley, A. Sud, A. Everall, A. J. Cornish, D. Chubb, et al. Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. *Nat Genet*, 56(9):1868–1877, 2024.

[24] A. Sosinsky, J. Ambrose, W. Cross, C. Turnbull, S. Henderson, et al. Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med*, 30(1):279–289, 2024.

[25] Michael Menzel, Mihaela Martis-Thiele, Hannah Goldschmid, Alexander Ott, Eva Romanovsky, et al. Benchmarking whole exome sequencing in the German network for personalized medicine. *European Journal of Cancer*, 211, 2024.

[26] M. Menzel, S. Ossowski, S. Kral, P. Metzger, P. Horak, et al. Multicentric pilot study to standardize clinical whole exome sequencing (WES) for cancer patients. *NPJ Precis. Oncol.*, 7(1):106, 2023.

[27] W. Xiao, L. Ren, Z. Chen, L. T. Fang, Y. Zhao, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat Biotechnol*, 39(9):1141–1150, 2021.

[28] F. Meric-Bernstam. The need for molecular tumor boards. *Nat Med*, 2025.

[29] A. H. Wagner, B. Walsh, G. Mayfield, D. Tamborero, D. Sonkin, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat Genet*, 52(4):448–457, 2020.

[30] P. Horak, M. Griffith, A. M. Danos, B. A. Pitel, S. Madhavan, et al. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): Joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for Cancer Consortium (VICC). *Genet Med*, 24(9):1991, 2022.

[31] J. Leichsenring, P. Horak, S. Kreutzfeldt, C. Heining, P. Christopoulos, et al. Variant classification in precision oncology. *Int J Cancer*, 145(11):2996–3010, 2019.

[32] M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn*, 19(1):4–23, 2017.

[33] S. J. Chanock. Harnessing cancer genomes for precision oncology. *Nat Genet*, 56 (9):1768–1769, 2024.

[34] Ksenia Lavrichenko, Emilie Sofie Engdal, Rasmus L. Marvig, Anders Jemt, Jone Marius Vignes, et al. Recommendations for Bioinformatics in Clinical Practice. *bioRxiv*, page 2024.11.23.624993, 2024.

[35] L. A. Garraway, J. Verweij, & K. V. Ballman. Precision oncology: an overview. *J Clin Oncol*, 31(15):1803–5, 2013.

[36] J. Rodon, J. C. Soria, R. Berger, W. H. Miller, E. Rubin, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat Med*, 25(5):751–758, 2019.

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

[38] P. Ramarao-Milne, O. Kondrashova, A. M. Patch, K. Nones, L. T. Koufariotis, et al. Comparison of actionable events detected in cancer genomes by whole-genome sequencing, in silico whole-exome and mutation panels. *ESMO Open*, 7(4):100540, 2022.

[39] P. Horak, S. Fröhling, & H. Glimm. Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. *ESMO Open*, 1(5):e000094, 2016.

[40] A. H. Mah, X. Qi, J. Zhao, K. Wiseman, L. Edoli, et al. A simplified hybrid capture approach retains high specificity and enables PCR-free workflow. *BMC Genomics*, 26(1):799, 2025.

[41] A. P. So, A. Vilborg, Y. Bouhlal, R. T. Koehler, S. M. Grimes, et al. A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. *NPJ Genom Med*, 3:2, 2018.

[42] S. H. Yim, S. H. Jung, B. Chung, & Y. J. Chung. Clinical implications of copy number variations in autoimmune disorders. *Korean J Intern Med*, 30(3):294–304, 2015.

[43] I. Cortés-Ciriano, D. C. Gulhan, J. J. Lee, G. E. M. Melloni, & P. J. Park. Computational analysis of cancer genome sequencing data. *Nat Rev Genet*, 23(5):298–314, 2022.

[44] Iaem van Belzen, A. Schönhuth, P. Kemmeren, & J. Y. Hehir-Kwa. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol*, 5(1):15, 2021.

[45] C. Alkan, B. P. Coe, & E. E. Eichler. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–76, 2011.

[46] S. Roy, C. Coldren, A. Karunamurthy, N. S. Kip, E. W. Klee, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn*, 20(1):4–27, 2018.

[47] D. C. Koboldt. Best practices for variant calling in clinical sequencing. *Genome Med*, 12(1):91, 2020.

[48] R. Bao, L. Huang, J. Andrade, W. Tan, W. A. Kibbe, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform*, 13(Suppl 2):67–82, 2014.

[49] S. Chen, Y. Zhou, Y. Chen, & J. Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.

[50] A. M. Bolger, M. Lohse, & B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–20, 2014.

[51] A. Kechin, U. Boyarskikh, A. Kel, & M. Filipenko. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *J Comput Biol*, 24(11):1138–1143, 2017.

[52] H. Li & R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[53] B. S. Pedersen & A. R. Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 2018.

[54] P. Ewels, M. Magnusson, S. Lundin, & M. Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19): 3047–8, 2016.

[55] Sunhee Kim, Young-suk Lee, & Chang-Yong Lee. Effect of Database Size in the Genetic Variants Calling. In *Bioinformatics*, 2019.

[56] L. Phan, H. Zhang, Q. Wang, R. Villamarin, T. Hefferon, et al. The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Res*, 53(D1): D925–d931, 2025.

[57] S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, 15(8):591–594, 2018.

[58] S. Gudmundsson, M. Singer-Berk, N. A. Watts, W. Phu, J. K. Goodrich, et al. Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat*, 43(8):1012–1030, 2022.

[59] K. J. Karczewski, B. Weisburd, B. Thomas, M. Solomonson, D. M. Ruderfer, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*, 45(D1):D840–d845, 2017.

[60] A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[61] K. Wang, M. Li, & H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, 2010.

[62] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, et al. The Ensembl Variant Effect Predictor. *Genome Biol*, 17(1):122, 2016.

[63] N. L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, & P. C. Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40(Web Server issue):W452–7, 2012.

[64] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, et al. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4): 248–9, 2010.

[65] J. M. Schwarz, D. N. Cooper, M. Schuelke, & D. Seelow. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11(4):361–2, 2014.

[66] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, & M. Kircher. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, 47(D1):D886–d894, 2019.

[67] Z. Sondka, N. B. Dhir, D. Carvalho-Silva, S. Jupe, Madhumita, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res*, 52(D1):D1210–D1217, 2024.

[68] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, 46(D1):D1062–d1067, 2018.

[69] D. Chakravarty, J. Gao, S. M. Phillips, R. Kundra, H. Zhang, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*, 2017, 2017.

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

[71] J. Corominas, S. P. Smeekens, M. R. Nelen, H. G. Yntema, E. J. Kamsteeg, et al. Clinical exome sequencing-Mistakes and caveats. *Hum Mutat*, 43(8):1041–1055, 2022.

[72] C. Berrios, E. A. Hurley, L. Willig, I. Thiffault, C. Saunders, et al. Challenges in genetic testing: clinician variant interpretation processes and the impact on clinical care. *Genet Med*, 23(12):2289–2299, 2021.

[73] E. Zukin, J. O. Culver, Y. Liu, Y. Yang, C. N. Ricker, et al. Clinical implications of conflicting variant interpretations in the cancer genetics clinic. *Genet Med*, 25 (7):100837, 2023.

[74] M. R. Stratton, P. J. Campbell, & P. A. Futreal. The cancer genome. *Nature*, 458 (7239):719–24, 2009.

[75] C. Luchini, R. T. Lawlor, M. Milella, & A. Scarpa. Molecular Tumor Boards in Clinical Practice. *Trends Cancer*, 6(9):738–744, 2020.

[76] L. Tögel, C. Schubart, S. Lettmaier, C. Neufert, J. Hoyer, et al. Determinants Affecting the Clinical Implementation of a Molecularly Informed Molecular Tumor Board Recommendation: Experience from a Tertiary Cancer Center. *Cancers (Basel)*, 15 (24), 2023.

[77] T. Goldfarb, V. K. Kodali, S. Pujar, V. Brover, B. Robbertse, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res*, 53(D1):D243–d257, 2025.

[78] X. Liu, C. Li, C. Mou, Y. Dong, & Y. Tu. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*, 12(1):103, 2020.

[79] K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[80] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*, 47(D1):D941–d947, 2019.

[81] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, 31(10):1536–43, 2015.

[82] D. Tamborero, R. Dienstmann, M. H. Rachid, J. Boekel, R. Baird, et al. Support systems to guide clinical decision-making in precision oncology: The Cancer Core Europe Molecular Tumor Board Portal. *Nat Med*, 26(7):992–994, 2020.

[83] Ifac Fokkema, M. Kroon, J. A. López Hernández, D. Asscheman, I. Lugtenburg, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. *Eur J Hum Genet*, 29(12):1796–1803, 2021.

[84] M. S. Cline, R. G. Liao, M. T. Parsons, B. Paten, F. Alquaddoomi, et al. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet*, 14(12):e1007752, 2018.

[85] M. Angeloni, S. Wach, H. Taubert, D. Sikic, B. Wullich, et al. Robust Consensus Molecular Subtyping of Muscle-Invasive Bladder Cancer Via 3' RNA Sequencing of Formalin-Fixed Paraffin-Embedded Tissues: Potential Impact for Clinical and Trial Settings. *Lab Invest*, 105(9):104191, 2025.

[86] M. Eckstein, P. Strissel, R. Strick, V. Weyerer, R. Wirtz, et al. Cytotoxic T-cell-related gene expression signature predicts improved survival in muscle-invasive urothelial bladder cancer patients after radical cystectomy and adjuvant chemotherapy. *J Immunother Cancer*, 8(1), 2020.

[87] C. Pfannstiel, P. L. Strissel, K. B. Chiappinelli, D. Sikic, S. Wach, et al. The Tumor Immune Microenvironment Drives a Prognostic Relevance That Correlates with Bladder Cancer Subtypes. *Cancer Immunol Res*, 7(6):923–938, 2019.

[88] D. F. Bajorin, J. A. Witjes, J. E. Gschwend, M. Schenker, B. P. Valderrama, et al. Adjuvant Nivolumab versus Placebo in Muscle-Invasive Urothelial Carcinoma. *N Engl J Med*, 384(22):2102–2114, 2021.

[89] M. D. Galsky, A. Arija JÁ, A. Bamias, I. D. Davis, M. De Santis, et al. Atezolizumab with or without chemotherapy in metastatic urothelial cancer (IMvigor130): a multicentre, randomised, placebo-controlled phase 3 trial. *Lancet*, 395(10236):1547–1557, 2020.

[90] A. Necchi, A. Anichini, D. Raggi, A. Briganti, S. Massa, et al. Pembrolizumab as Neoadjuvant Therapy Before Radical Cystectomy in Patients With Muscle-Invasive Urothelial Bladder Carcinoma (PURE-01): An Open-Label, Single-Arm, Phase II Study. *J Clin Oncol*, 36(34):3353–3360, 2018.

[91] M. Scimeca, J. Bischof, R. Bonfiglio, E. Nale, V. Iacovelli, et al. Molecular profiling of a bladder cancer with very high tumour mutational burden. *Cell Death Discov*, 10(1):202, 2024.

[92] I. T. Nakamura, S. Kohsaka, M. Ikegami, H. Ikeuchi, T. Ueno, et al. Comprehensive functional evaluation of variants of fibroblast growth factor receptor genes in cancer. *NPJ Precis Oncol*, 5(1):66, 2021.

[93] K. Yuen, M. Meagher, J. Mercer, B. Yilma, M. Stoppler, et al. Comprehensive Comparison of Somatic, Germline, and Immune Cell Profiles in Upper Tract and Bladder Urothelial Carcinoma. *JCO Precis Oncol*, 9:e2500289, 2025.

[94] The Cancer Genome Atlas Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–22, 2014.

[95] A. G. Robertson, J. Kim, H. Al-Ahmadie, J. Bellmunt, G. Guo, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*, 174(4):1033, 2018.

[96] Y. Loriot, N. Matsubara, S. H. Park, R. A. Huddart, E. F. Burgess, et al. Erdafitinib or Chemotherapy in Advanced or Metastatic Urothelial Carcinoma. *N Engl J Med*, 389(21):1961–1971, 2023.

[97] J. A. Nakauma-González, M. Rijnders, J. van Riet, M. S. van der Heijden, J. Voortman, et al. Comprehensive Molecular Characterization Reveals Genomic and Transcriptomic Subtypes of Metastatic Urothelial Carcinoma. *Eur Urol*, 81(4):331–336, 2022.

[98] A. Necchi, R. Madison, S. K. Pal, J. S. Ross, N. Agarwal, et al. Comprehensive Genomic Profiling of Upper-tract and Bladder Urothelial Carcinoma. *Eur Urol Focus*, 7(6):1339–1346, 2021.

[99] Q. Tang, W. Zuo, C. Wan, S. Xiong, C. Xu, et al. Comprehensive genomic profiling of upper tract urothelial carcinoma and urothelial carcinoma of the bladder identifies distinct molecular characterizations with potential implications for targeted therapy & immunotherapy. *Front Immunol*, 13:1097730, 2022.

[100] F. Nichetti, M. Brambilla, M. Duca, A. Piccolo, D. Miliziano, et al. Real-World Outcomes of Molecular Tumor Board Treatment Recommendations. *JCO Precis Oncol*, 9:e2400387, 2025.

[101] A. Scheiter, F. Hierl, F. Lüke, F. Keil, D. Heudobler, et al. Critical evaluation of molecular tumour board outcomes following 2 years of clinical practice in a Comprehensive Cancer Centre. *Br J Cancer*, 128(6):1134–1147, 2023.

[102] Maria Giulia Carta, Lars Tögel, Annett Hölsken, Christoph Schubart, Heinrich Sticht, et al. Oncogenicity Variant Interpreter (OncoVI): oncogenicity guidelines implementation to support somatic variants interpretation in precision oncology. *medRxiv*, page 2024.10.10.24315072, 2024.

[103] F. O. Bagger, L. Borgwardt, A. S. Jespersen, A. R. Hansen, B. Bertelsen, et al. Whole genome sequencing in clinical practice. *BMC Med Genomics*, 17(1):39, 2024.

[104] M. Meggendorfer, V. Jobanputra, K. O. Wrzeszczynski, P. Roepman, E. de Bruijn, et al. Analytical demands to use whole-genome sequencing in precision oncology. *Semin Cancer Biol*, 84:16–22, 2022.

[105] P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, et al. 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 (Austin)*, 6(2):80–92, 2012.

[106] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5):401–4, 2012.

[107] C. Kopanos, V. Tsiolkas, A. Kouris, C. E. Chapple, M. Albarca Aguilera, et al. VarSome: the human genomic variant search engine. *Bioinformatics*, 35(11):1978–1980, 2019.

[108] B. J. Ainscough, M. Griffith, A. C. Coffman, A. H. Wagner, J. Kunisaki, et al. DoCM: a database of curated mutations in cancer. *Nat Methods*, 13(10), 2016.

[109] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*, 10(1):25, 2018.

[110] S. Nakken, G. Fournous, D. Vodák, L. B. Aasheim, O. Myklebost, & E. Hovig. Personal Cancer Genome Reporter: variant interpretation report for precision oncology. *Bioinformatics*, 34(10):1778–1780, 2018.

[111] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*, 51(D1):D523–d531, 2023.

[112] A. Palmisano, S. Vural, Y. Zhao, & D. Sonkin. MutSpliceDB: A database of splice sites variants with RNA-seq based evidence on effects on splicing. *Hum Mutat*, 42 (4):342–345, 2021.

[113] M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol*, 34(2):155–63, 2016.

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

[115] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185(4154):862–4, 1974.

[116] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, & A. Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–21, 2010.

[117] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 2005.

[118] J. Morales, S. Pujar, J. E. Loveland, A. Astashyn, R. Bennett, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905): 310–315, 2022.

[119] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *Available at: http://www.R-project.org/*, 2022.

[120] Maria Giulia Carta, Miriam Angeloni, Lars Tögel, Christoph Schubart, Annett Hölsken, et al. A fully-automated integrative workflow to streamline NGS-based analyses within Molecular Tumour Boards. *medRxiv*, page 2025.12.12.25341897, 2025.

[121] Daniel Hübschmann, Simon Kreutzfeldt, Benjamin Roth, Katrin Glocker, Janine Schoop, et al. Knowledge Connector: Decision support system for multiomics-based precision oncology. *medRxiv*, page 2025.02.23.25322403, 2025.

[122] L. C. Chang, H. C. Kuo, H. M. Wang, Y. C. Kuo, C. T. Wang, et al. The Use of an Integrated Digital Tool to Improve the Efficiency of Multidisciplinary Tumor Boards-A Prospective Trial in Taiwan. *Cancers (Basel)*, 17(3), 2025.

[123] C. Strantz, D. Böhm, T. Ganslandt, M. Börries, P. Metzger, et al. Empowering personalized oncology: evolution of digital support and visualization tools for molecular tumor boards. *BMC Med Inform Decis Mak*, 25(1):29, 2025.

[124] J. Bossenz, I. Manuilova, A. B. Weise, S. Schulze, S. Hiemer, et al. Prototypical Visualization of Patient Similarities in cBioPortal to Enhance Decision-Making in Molecular Tumor Boards. *Stud Health Technol Inform*, 327:487–491, 2025.

[125] M. Angeloni, D. Rizzi, S. Schoen, A. Caputo, F. Merolla, et al. Closing the gap in the clinical adoption of computational pathology: a standardized, open-source framework to integrate deep-learning models into the laboratory information system. *Genome Med*, 17(1):60, 2025.

[126] P. Van Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*, 107(39): 16910–5, 2010.

[127] Z. Sztupinszki, M. Diossy, M. Krzystanek, L. Reiniger, I. Csabai, et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer*, 4:16, 2018.

[128] N. J. Birkbak, Z. C. Wang, J. Y. Kim, A. C. Eklund, Q. Li, et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov*, 2(4):366–375, 2012.

[129] V. Abkevich, K. M. Timms, B. T. Hennessy, J. Potter, M. S. Carey, et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br J Cancer*, 107(10):1776–82, 2012.

[130] T. Popova, E. Manié, G. Rieunier, V. Caux-Moncoutier, C. Tirapo, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res*, 72(21):5454–62, 2012.

Acknowledgements

First of all, I would like to express my gratitude to Prof. P. Magni and PD Dr. F. Ferrazzi for their mentorship and guidance throughout the course of this PhD work.

My sincere thanks also go to Prof. Dr. med. A. Hartmann and Prof. Dr. med. F. Haller for their support and contribution to my scientific growth.

I am grateful to the reviewers who generously dedicated their time and expertise to evaluate this work.

I would like to thank my colleagues from the Molecular Diagnostics group — Dr. L. Tögel, Dr. C. Schubart, Dr. A. Hölsken, PD Dr. E.A. Moskalev, and all collaborators at the Institute of Pathology — for their cooperation, technical support, and pleasant working atmosphere.

I thank also my colleagues from the Bioinformatics and Computational Pathology group — Dr. N. Feldker, Dr. R. Liguori, M. Sieger, and Dr. Vi Dang — for their collaboration and stimulating discussions. A very special mention goes to M. Angeloni, who has been not only a fundamental point of reference within the group but also a precious friend.

Finally, from the bottom of my heart, I wish to thank my family and friends for their endless love, encouragement, and warm closeness, even from afar.