PhD Program in Biomolecular Sciences
Department of Cellular, Computational and Integrative Biology – CIBIO
37th Cycle

Non-Invasive cancer detection: computational applications in liquid biopsy and radiomics

Tutor
Prof. Alessandro Romanel
University of Trento

Ph.D. Thesis of
Riccardo Scandino
University of Trento

Academic Year 2023-2024

Declaration

I, Riccardo Scandino, confirm that this is my own work and the use of all material from other sources have been properly and fully acknowledged.

Riccardo Scandino

Table of Contents

Abstract

This thesis explores the field of non-invasive cancer detection, in particular, it focuses on the development and application of computational approaches for analyzing liquid biopsy and radiomic data. The research presented herein addresses the limitations of current methods in comprehensively analyzing circulating tumor DNA (ctDNA), particularly at low ctDNA fractions, and introduces novel tools and strategies to enhance the sensitivity and specificity of ctDNA analysis. The first part of the thesis focuses on the development of Synggen, a tool for generating synthetic cancer data. Synggen addresses the critical need for standardized datasets to benchmark cancer analysis tools by creating synthetic data that closely mimics real-world samples. This tool allows researchers to evaluate algorithms in a controlled environment, overcoming the limitations associated with relying solely on real patient data, such as privacy concerns and inherent heterogeneity. The second part of the thesis introduces eSENSES, a novel NGS panel designed for sensitive and specific ctDNA characterization in breast cancer through the analysis of somatic copy number alterations (SCNAs). The panel's unique combination of elements, including genome-wide and focal single nucleotide polymorphisms (SNPs), exonic regions of key cancer-related genes, and a bespoke computational strategy, enhances the detection and quantification of SCNAs and ctDNA fraction. This approach aims to improve disease monitoring and inform therapeutic decisions in breast cancer patients. The final part of the thesis investigates the potential of radiomics in predicting treatment response in lung cancer patients. This exploration aims to complement liquid biopsy approaches by incorporating quantitative features extracted from medical images.

List of Abbreviations

AF - Allelic Fraction

CBS - Circular Binary Segmentation

cfDNA – Circulating-Free DNA

cfRNA – Circulating-Free RNA

CNA - Copy Number Aberration

CTCs – Circulating Tumor Cells

ctDNA - Circulating Tumor DNA

dNLR – derived Neutrophil-to-Lymphocyte Ratio

ECOG - Eastern Cooperative Oncology Group

EVs - Extracellular Vesicles

FDA – Food and Drug Administration

FISH - Fluorescence in situ hybridization

gMAF - global minor allele frequency

ICIs - Immune Checkpoint Inhibitors

LDH – Lactate dehydrogenase

LOO - Leave-One-Out

lpWGBS – low-pass Whole-Genome Bisulfite Sequencing

mAF - mirrored allelic fraction

Mb - Megabase

MRD - Minimal Residual Disease

NGS - Next-Generation Sequencing

NSCLC - Non-Small Cell Lung Cancer

OS - Overall Survival

PCA - Principal Component Analysis

PBE - Position-Based Error

PCR - Polymerase Chain Reaction

PFS - Progression-Free Survival

PMs - Point Mutations

QM - Quality Model

RD - Read Depth

RDM - Read Depth Model

RMB - Reference Mapping Bias

SCNA - Somatic Copy Number Alteration

SNP - Single Nucleotide Polymorphism

SNV - Single Nucleotide Variant

TC - Tumor Content

TCGA - The Cancer Genome Atlas

TEPs - Tumor-Educated Platelets

TS - Targeted Sequencing

WES - Whole-Exome Sequencing

WGBS – Whole-Genome Bisulfite Sequencing

WGS – Whole-Genome Sequencing

WHO - World Health Organization

List of Figures

List of Supplementary Figures

List of Tables

Chapter 1 - Introduction

Cancer remains a formidable health challenge, characterized by its intricate and dynamic nature (1,2). This disease is marked by a series of extensive somatic alterations accumulated during its evolution, making it a complex adversary in the quest for effective treatments and cures (1,3,4). Research efforts are heavily invested in deciphering the mechanisms of cancer development and progression, with a particular focus on the role of genetic alterations (2). Advances in genomic techniques have provided invaluable insights into the intricate genetic landscape of malignant tumors, paving the way for the identification of potential therapeutic targets and the development of personalized treatment strategies (5,6).

The direct application of this concepts is cancer detection, the identification of cancerous cells or tissues at various stages of malignancy, plays a pivotal role in improving survival rates (7,8). Ideally, detection at an early stage, when treatment outcomes are most favorable, is crucial for effective management of the disease (9). In research and clinical settings, cancer detection encompasses a range of diagnostic methodologies, each designed to identify specific biomarkers indicative of malignancy. These biomarkers can include genetic mutations, abnormal protein expression, or metabolic changes, and can be derived from various sources such as tissue biopsies, blood, or medical images (10–12). The development of accurate and cancer detection methods is essential for timely therapeutic intervention, precise staging, and effective monitoring of disease progression or response to treatment.

While traditional cancer detection methods, such as tissue biopsies, have proven valuable, they often come with limitations, including invasiveness and potential sampling bias (13). Non-invasive cancer detection methodologies have revolutionized the landscape of oncology, offering a means to detect and monitor cancer without the need for traditional invasive biopsies or surgeries (10,12,14,15). These techniques enable clinicians to obtain critical insights into the molecular and functional characteristics of tumors. In this thesis work we focus on liquid biopsy and radiomics, both of which rely on innovative technological advances and big data analytics.

1.1 Liquid Biopsy

The field of oncology has been significantly impacted by the emergence of non-invasive cancer detection methodologies, offering alternatives to traditional invasive procedures. Among these, liquid biopsy stands out as a promising approach, allowing for the detection and monitoring of cancer without the need for surgical biopsies (12). Liquid biopsy involves the collection and analysis of circulating biomarkers from bodily fluids, primarily blood, providing a minimally invasive window into the molecular characteristics of tumors. This technique offers several advantages, including the ability to track tumor dynamics in real-time, assess genetic mutations, and monitor treatment response without the inherent risks and limitations of tissue biopsies.

Several molecular markers are detectable through liquid biopsy, such as circulating tumor cells (CTCs), circulating tumor DNA (ctDNA), tumor-derived extracellular vesicles (EVs), tumor-educated platelets (TEPs), and circulating free RNA (cfRNA) (16).

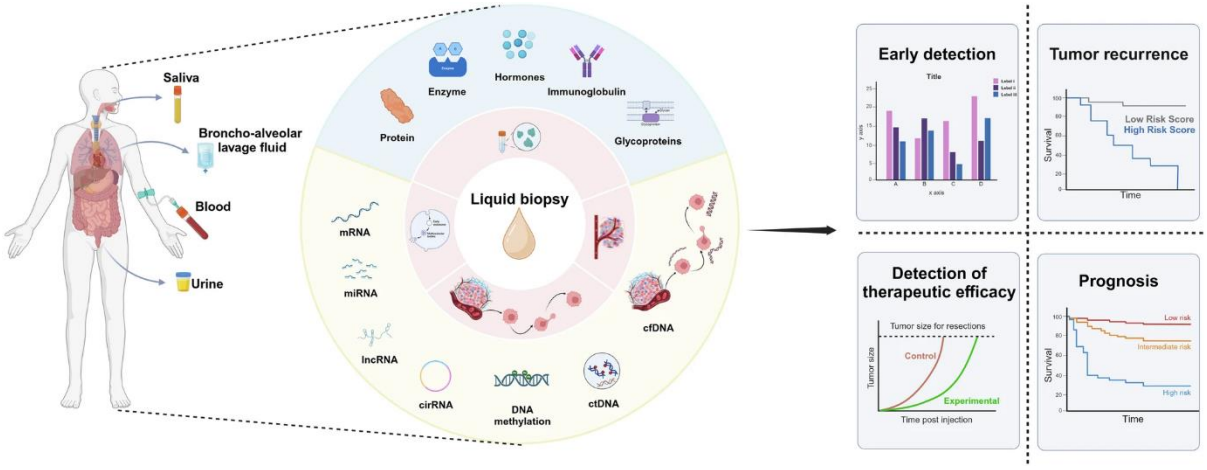

Figure 1.1 Flowchart of applying liquid biopsy in cancers. Applications of liquid biopsies and types of biomarkers for liquid biopsies. Adapted from (16) with the creative common license (https://creativecommons.org/licenses/by/4.0/).

Despite ongoing research into various approaches, only one method based on Circulating Tumor Cells (CTCs) (17) and a limited number of tests based on circulating free DNA (cfDNA) have received approval from the U.S. Food and Drug Administration (FDA) for clinical use, primarily as companion diagnostic tests (18).

The relatively low rate of FDA approvals can be attributed to the generally low reproducibility of preclinical studies, aided by the intrinsically low concentration of these analytes, especially in specific cases, which further diminishes reproducibility (19).

Despite these challenges, liquid biopsy application and clinical acceptance are steadily increasing, particularly in areas such as cancer biomarker identification, real-time monitoring of therapeutic efficacy and response (20), disease progression assessment (21), and early cancer detection (22,23).

1.1.1 Circulating free DNA as a molecular marker

One of the most studied analyte in liquid biopsies is circulating cell-free DNA (cfDNA), which refers to small fragments of DNA found in the bloodstream, originating from various cell types undergoing processes such as apoptosis, necrosis, or active secretion (12,24). These fragments, typically 150-200 base pairs in length, can be detected in both healthy and diseased individuals, offering a non-invasive snapshot of the body's physiological and pathological state (25,26). In oncology, cfDNA, particularly the fraction derived from tumor cells (ctDNA), has gained significant attention as a non-invasive biomarker for cancer detection and monitoring (9,10,24,26). ctDNA carries the genetic fingerprints of the tumor, providing valuable information about its molecular characteristics and potential vulnerabilities (27), these include point mutations (PMs), methylation modifications, DNA fragment size, and somatic copy number aberrations (SCNAs). However, the analysis of cfDNA presents several challenges, including the low abundance of ctDNA in the bloodstream, especially in early-stage cancers, and the inherent heterogeneity of tumors (12,28).

The ability to accurately detect and quantify ctDNA is crucial for understanding the evolution of cancer and developing effective treatment strategies.

A significant portion of ctDNA assays in oncology focuses on detecting specific mutations in plasma (18,29). Point mutations, which are single-base somatic genetic alterations in the DNA sequence, are known to contribute to cancer development (5,30–32). Their identification holds potential for enabling cancer detection, diagnosis, and tissue-of-origin identification. Many clinical assays have been developed around point mutations utilizing different approaches (18,33). Highly sensitive, targeted techniques, such as Polymerase Chain Reaction (PCR)-based methods, can detect these mutations at very low concentrations with high specificity (34,35). On the other hand, broader sequencing approaches using Next Generation

Sequencing (NGS) allow for the analysis of entire regions of the DNA rather than targeting specific mutations. Although NGS-based techniques are generally less cost-effective compared to targeted assays, they provide more comprehensive genotyping (36). One of the main challenges in point mutation detection is the risk of false positives, particularly with NGS-based methods. Library preparation and sequencing steps can introduce errors, necessitating careful bioinformatic analysis to differentiate true mutations from background noise. A common practice in tissue biopsy to compensate for these errors is comparing paired tumor and healthy tissues for accurately evaluating the presence of mutations. However, obtaining matched germline samples with same molecular context as the plasma sample is particularly difficult in liquid biopsy due to the heterogeneous origin of cfDNA (37).

NGS-based methods such as targeted sequencing (TS), whole-exome sequencing (WES), whole-genome sequencing (WGS), and whole-genome bisulfite sequencing (WGBS), enable the detection and analysis of various characteristics of ctDNA.

ctDNA detection methods based on NGS sequencing tend to focus either entirely on focal approaches for specific somatic events identification, such as TS assays (38), or entirely on broader, less specific, ones, such as WGS, and WGBS (39).

WGBS, and bisulfite sequencing in general, allows for the distinction of ctDNA from the remaining cfDNA based on epigenetic dysregulation (40). Epigenetic dysregulation, which occurs early in tumorigenesis, involves widespread alterations in DNA methylation and chromatin organization within cancer cells and the tumor microenvironment (41,42), thereby facilitating cancer detection and tissue characterization. However, conventional bisulfite-based methylation sequencing can cause significant DNA damage, making it unsuitable for profiling other important cfDNA features (43).

In recent years, WGS and other sequencing approaches have been used to analyze the properties of the circulating fragments. This field, also known as fragmentomics, focuses on studying size, ends and strandness of cfDNA fragments (36). Changes in these features allow the detection and characterization of ctDNA (44). Moreover, recent studies have shown how fragmentomics has the potential to estimate the epigenomic and transcriptomic landscape of a sample (26,45,46).

NGS applications allow the identification of SCNAs from cfDNA which are another useful feature for the characterization of tumors (47,48). These events, considered a hallmark of cancer (2,6), can drive tumor development and progression, making their detection and

characterization essential for understanding cancer biology and identifying potential therapeutic targets (49–52). In addition to SCNAs, single nucleotide polymorphisms (SNPs) represent another important source of genetic variation with implications for cancer development and progression. SNPs, the most common type of genetic variation in the human genome, arise from single base-pair changes in the DNA sequence (53). These variations, occurring approximately once every 300 base pairs, contribute significantly to the genetic diversity observed within and between populations (54,55). SNPs can occur in both coding and non-coding regions of the genome, with those located within coding regions potentially altering protein structure and function. While many SNPs have no discernible phenotypic effect, some have been linked to various diseases and traits, including cancer susceptibility, drug response, and predisposition to complex diseases (56).

In the context of cancer genomics, SNPs can be exploited to detect and characterize allelic imbalance, a phenomenon where the normal equal representation of two alleles at a heterozygous locus is disrupted. This imbalance can arise from somatic copy number alterations (SCNAs), where gains or losses of genomic regions lead to deviations from the expected allelic ratios (57). By analyzing allelic imbalance patterns using SNPs as markers, we can gain valuable insights into the genomic alterations driving tumor development and progression (58,59). For instance, the relative abundance of different SNP alleles in circulating tumor DNA (ctDNA) can be used to infer the presence of SCNAs and estimate the fraction of ctDNA in a sample, aiding in cancer detection and monitoring (60).

1.2 Radiomics

Radiomics is an emerging field that leverages advanced computational techniques to extract a vast array of quantitative features from medical images, providing a comprehensive characterization of tumor phenotypes beyond what is visually apparent (14,15,61). These features, encompassing tumor shape, size, texture, and intensity, can capture subtle intratumoral heterogeneity and provide insights into the underlying biological processes driving tumor behavior (61,62). By applying sophisticated image processing and machine learning algorithms, radiomics transforms standard medical images into mineable high-dimensional data, enabling the identification of patterns that correlate with specific clinical

outcomes (62,63). This non-invasive approach offers a complementary tool to liquid biopsy, enhancing our understanding of tumor biology and potentially guiding personalized treatment strategies (61).

However, the field of radiomics faces several challenges. The lack of standardized imaging protocols and feature extraction methods can introduce variability in results across different institutions and studies, hindering reproducibility and generalizability (64). Additionally, the high dimensionality of radiomic data requires robust feature selection methods and sophisticated computational tools to analyze and interpret the data effectively (14,64,65). Building robust radiomic models relies on large, well-annotated datasets, which are not always readily available. Finally, while radiomics can generate hypotheses about tumor biology based on image-derived features, these findings must be validated through rigorous biological studies to ensure their clinical relevance and applicability (64).

1.3 Aim of the thesis

This thesis aims to develop and apply innovative computational approaches for analyzing cancer data obtained through non-invasive methodologies. The research focuses on enhancing the detection and characterization of circulating tumor DNA (ctDNA) in cancer patients, facilitating the development and benchmarking of computational tools in this field, and exploring the integration of complementary non-invasive approaches, such as radiomics, with liquid biopsy for a more comprehensive understanding of cancer biology and improved patient care.

Assessing the performance of computational approaches in the context of liquid biopsy is complex due to the accessibility of the data and the accurate estimation of the tumor signal. Therefore, as a first step in building a computational approach, a simulator of genomic data, Synggen, was created (Chapter 2) to test and fine-tune the computational approach developed in Chapter 3.

Chapter 3 focuses on the development of eSENSES, a targeted NGS panel for analyzing cfDNA in breast cancer patients. This panel combines broad genome-wide and focal SNPs, exonic regions of key cancer-related genes, and a tailored computational strategy to enhance the detection and quantification of SCNAs and ctDNA fraction.

While eSENSES demonstrated promising results, its integration with other clinical data, such as imaging, could further enhance its prognostic and predictive value. This multimodal approach, combining diverse biological signals through machine learning models, offers the potential for a more comprehensive and personalized understanding of cancer progression and treatment response. However, due to the nature of the data used in this project, direct exploration of this integration was not feasible.

Therefore, Chapter 4 explores another non-invasive cancer detection approach, radiomics, to evaluate the potential of a different approach. Although the context of application for this part of the work is different from the previous chapters (NSCLC versus breast cancer, immunotherapy versus cfDNA), it serves as a proof of concept to evaluate whether this approach could potentially provide a different angle of characterization compared to the use of a liquid biopsy approach, leading to a future prospect of integration of these different approaches for a more complete patient characterization.

Chapter 2 - Synggen: fast and data-driven generation of synthetic heterogeneous NGS cancer data

Authors: Riccardo Scandino, Federico Calabrese, Alessandro Romanel

Journal: Bioinformatics, December 2022

DOI: https://doi.org/10.1093/bioinformatics/btac792

This chapter explores the development of synggen, a tool for generating synthetic cancer data, recently published in *Bioinformatics*. My contributions to this project encompassed testing and refining the tool, as well as participating in the design and implementation of some of its key features.

Synggen addresses the critical need for standardized datasets to benchmark cancer analysis tools. By creating synthetic data that closely mirror real-world samples, synggen allows researchers to evaluate algorithms in a controlled environment, overcoming the limitations associated with relying solely on real patient data, such as privacy concerns and inherent heterogeneity.

Synggen leverages real control samples as a foundation for creating synthetic datasets. It permits the incorporation of essential genomic features, including germline and somatic mutations and copy number variations, while also allowing for user-defined parameters like sequencing coverage and tumor content. This approach ensures that the generated data are both biologically realistic and adaptable to specific research requirements.

To evaluate synggen's performance, I generated synthetic whole-exome sequencing (WES) and targeted sequencing (TS) samples. The results demonstrated efficient scaling with computational resources and accurate representation of complex somatic events. Moreover, I employed synggen to simulate liquid biopsy cell-free DNA (cfDNA) scenarios, highlighting its proficiency in incorporating somatic alterations and preserving platform-specific genomic characteristics.

2.1 Introduction

The interrogation of next-generation sequencing (NGS), principally whole-exome (WES) and targeted sequencing (TS), is rapidly becoming a preferred approach for the exploration of large tissue and liquid biopsy-based cohorts, especially in the context of precision medicine. In this setting, a precise benchmarking of the large collection of computational tools that are continuously developed is fundamental for the correct interpretation of somatic genomic profiling results. Although many simulators of synthetic cancer genomes and NGS data have been developed in the last years(66), only a few of them focus on WES and TS data and most of them either require the generation and merge of intermediate large FASTA (67), incorporate random genomic events (68) or do not allow to incorporate complex somatic cancer-specific patterns (69). The generation of large-scale benchmarking datasets that are able to capture cancer-specific heterogeneous scenarios together with NGS platform-specific characteristics is hence still hindered, limiting the broad applicability of synthetic NGS-based computational tool benchmarking. To overcome these limitations, we developed synggen, a tool that enables a fast and scalable generation of platform-specific cancer and matched control WES and TS synthetic data, incorporating phased germline polymorphisms together with complex and allele-specific cancer genomic events.

2.2 Approach

Synggen is a tool written in C programming language to generate synthetic NGS files, in the form of WES or TS experiments, representing heterogeneous cancer genomes and matched controls. The tool provides two execution modes which allow to (i) exploit a set of control (non-cancer) NGS sequencing files (BAM format) to generate three *reference models* capturing a collection of data summary statistics; and (ii) combine these reference models and a set of user-specified germline and somatic genomic profiles to create synthetic sequencing files (FASTQ format).

Reference models are built across a single or a collection of WES or TS sequencing control samples, profiled with the same platform, using a fast and efficient multi-threaded cumulative pileup strategy based on (70). Specifically, synggen generates: (i) a *Read Depth Model* (*RDM*), which measures the average intra-sample depth of coverage variability by calculating the probability, across all input files, of observing sequencing reads at any captured genomic

region and position; only reads aligning at the lowest genomic coordinate are considered when paired-end data is used; (ii) a *Quality Model* (*QM*), which measures the distribution of base qualities across sequencing read positions; (iii) *Position-Based Error* (*PBE*), which measures for each captured genomic position [not representing a common single nucleotide polymorphism (SNP)] the probability of observing platform-specific systematic errors supported by high quality reads and bases (71). When paired-end data are used to generate the reference models, insert size statistics are also computed and embedded in the RDM model.

Exploiting the created reference models, synggen allows to directly generate platform-specific cancer and matched control NGS files by incorporating in the simulated genomes a user-specified list of germline-phased SNPs and user-specified lists of somatic *allele-specific* copy number alterations (CNAs) and point mutations (PMs). Phased SNPs, which are specified by the user for both their presence and their phasing, are used to realistically represent the genetic background of specific individuals and to generate both cancer and matched control samples. It is important to consider the introduction of phased SNPs when simulating a specific biological scenario to preserve real and ancestry specific haplotype structure of individuals. Somatic CNAs are defined as allele-specific copy numbers (pairs of values specifying the number of copies for each allele) and are incorporated across all affected WES/TS captured regions to modulate the RDM distribution and to adjust SNP allelic fractions. Of note, both overlapping and nested somatic copy number alterations can be represented (see Materials and Methods for details). Point mutations are incorporated by specifying the affected allele and indicating the number of allele copies (in case copy number alterations are present) carrying the specific point mutation. When generating cancer genomes NGS files, CNAs and PMs are defined specifying also alteration-specific clonality values, which are used to adjust the fraction at which they are incorporated in the data. A global tumor content value is also specified by the user to adjust the fraction at which all somatic events are incorporated. Provided all these information, synggen uses an efficient multi-threaded approach to generate a specified number of synthetic sequencing reads implementing for each of them the following steps: (i) sample a genomic region and position from the RDM distribution; (ii) Generate the corresponding read sequence provided a specific read length; (iii) randomly choose one of the two alleles, sampling from a probability distribution that reflects the number of region allele-specific copies; (iv) incorporate all SNPs spanning the read (if any); (v)

Generate base qualities and errors sampling first from the PBE (for all positions there annotated) and then from the QM for all the remaining read bases; (vi) In case of cancer samples, incorporate PMs if any; and (vii) Choose the strand considering the strand bias (default 0.5). When paired-end data is considered, two reads are generated in Step 2 using an insert size sampled from a normal distribution with mean and standard deviation as specified in the RDM model; Steps from 3 to 7 are executed for both reads. Details are available in the Materials and Methods section.

2.3 Results

We developed synggen to allow efficient and scalable generation of cancer and matched control targeted NGS synthetic data. Compared to previous methods (**Table 2.1**), synggen is the only simulator providing built-in generation of platform-specific WES- or TS-based NGS reads, reproducing hence sequencing characteristics of real data and incorporating germline and somatic variants without the need of intermediate FASTA files. Differently from most available simulators, synggen allows to input specific lists of germline variants and somatic genomic events, including phased germline SNPs and somatic allele-specific CNAs and PMs, together with local and global parameters including the clonality of somatic events and the overall sample tumor content, allowing for the emulation of varied and realistic cancer- and patient-specific data across the different multi-subclones composition, tumor purity, aneuploidy and tumor evolution scenarios. Our approach allows for the generation of large-scale bulk WES and TS datasets, including multi-regional and longitudinal tissue sequencing datasets and liquid biopsy cfDNA sequencing datasets.

Table 2-1. Comparison table of synggen features against a list of relevant available simulators. Features reported in the original publications were used as comparison. Specific genomic variants refer to the capability of incorporating a specific variant in a specific genomic position.

		Simulators				
		Synggen	HeteroGenesis	BAMSurgeon	XomeBlender	GemSim
Formats	Input	BAM	FASTA	BAM	BAM	SAM
	Output	FASTQ	FASTA	BAM	BAM	FASTQ
	Generation of intermediate FASTA	No	Yes	No	No	No
Sequencing params	Learn error distribution	Yes	No	Yes (no model creation)	Yes (no model creation)	Yes
	Learn quality model	Yes	No	Yes (no model creation)	Yes (no model creation)	Yes
	Learn coverage distribution	Yes	No	Yes (no model creation)	Yes (no model creation)	?
Somatic variants parameters	Allele-specificity	Yes	Not directly	No	No	No
	Clonality of somatic events	Yes	Not directly	Yes	Yes	No
	Global tumor content	Yes	Not directly	No	Yes	No
Specific genomic variants	SNP	Yes	No	Yes	No	Yes
	PM	Yes	No	Yes	No	Yes
	CNA	Yes	No	No	Yes	No
NGS capture strategy	TS	Yes	?	?	Yes	Yes
	WES	Yes	Yes	Yes	Yes	Yes

To test the performances of synggen, we generated synthetic WES and TS samples and measured the computational time required for reference models' construction and sequencing reads generation. The construction of reference models was tested for an increasing number of input samples selected among breast cancer control (non-tumor) samples (BAM files) available from The Cancer Genome Atlas dataset, focusing on patients having high (>80%) tumor content WES (Sure Select All Exome v3) cancer tissue samples. Generation of synthetic NGS cancer data was instead tested for increasing number of sequencing reads produced. A cancer sample was generated with tumor content at 80% and incorporating patient-specific germline SNPs and patient-specific somatic CNAs and PMs (details in the Materials and Methods). The overall reference models' construction time using one input sample on a HPE Proliant DL560 server and exploiting 16 cores took approximately 2.5 min for the WES sample and 1 second for the TS sample (**Figure 2.1**). Using the same running configuration, 56 s and 1 min were instead required to generate a FASTQ file representing, respectively, a cancer WES sample with 100x average coverage (∼36 million reads) and a cancer TS sample with 1000× average coverage (∼57 million reads) (**Figure 2.1**). Overall, results show that reference models' creation and FASTQ files generation scales well with the number of threads used (**Figures 2.1 and 2.2**).

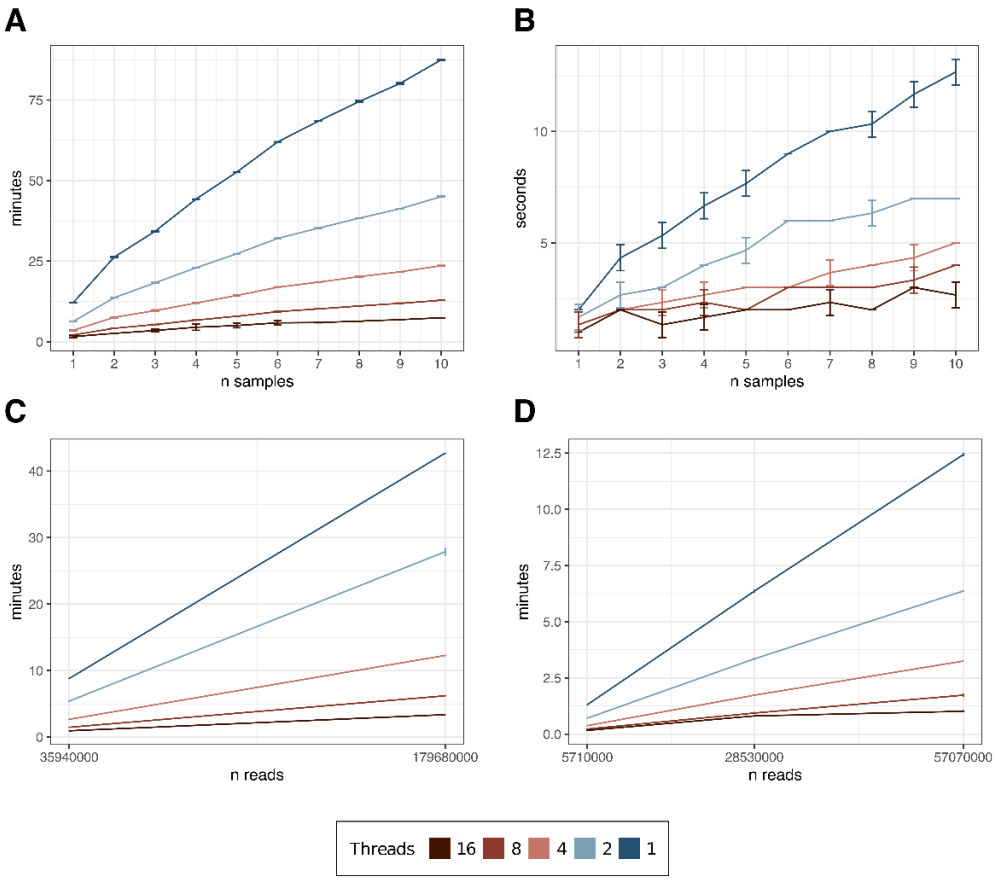

Figure 2.1 Benchmark of time required by synggen.
Mean time required for model generation at increasing samples using increasing threads in WES (A) and TS (B) scenarios. Time required for reads generation is tested for increasing number of threads and increasing number of reads, simulating a generation of a WES sample with average 100x and 500x coverage (C) and a TS sample with average 100x, 500x and 1000x coverage (D).

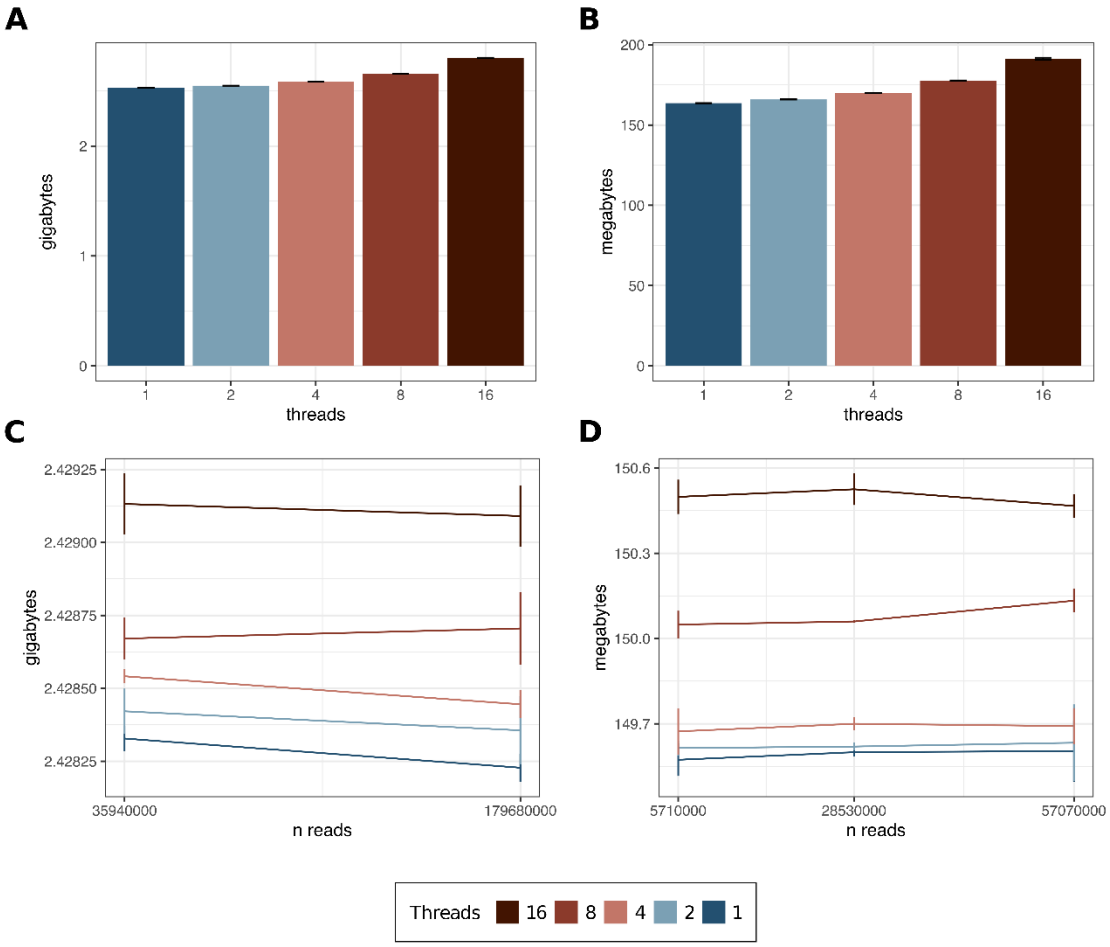

Figure 2.2 Benchmark of memory required by synggen.
Mean memory required for model generation using 5 samples at increasing number of threads in WES (A) and TS (B) scenarios. Memory required for reads generation is tested for increasing number of threads and increasing number of reads, simulating a generation of a WES sample with average 100x and 500x coverage (C) and a TS sample with average 100x, 500x and 1000x coverage (D).

In addition, complex and patient's specific combinations of somatic events were generated (**Figures 2.3**, **2.4 and 2.5**) keeping genomic characteristics like depth of coverage distributions and error rates consistent with the original WES sample (**Figures 2.6 and 2.7**).

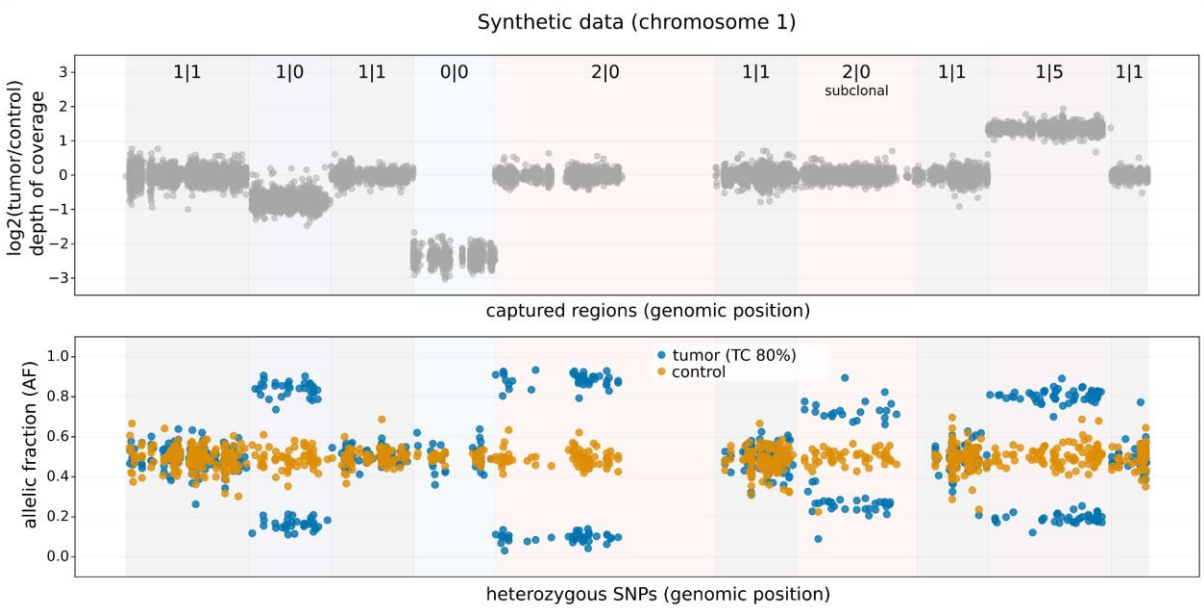

Figure 2.3 Examples of somatic allele-specific CNAs.

(Top) log2(tumor/control) of regions' average depth of coverage. (Bottom) AF of heterozygous SNPs in tumor and matched control synthetic samples. TC = tumor content.

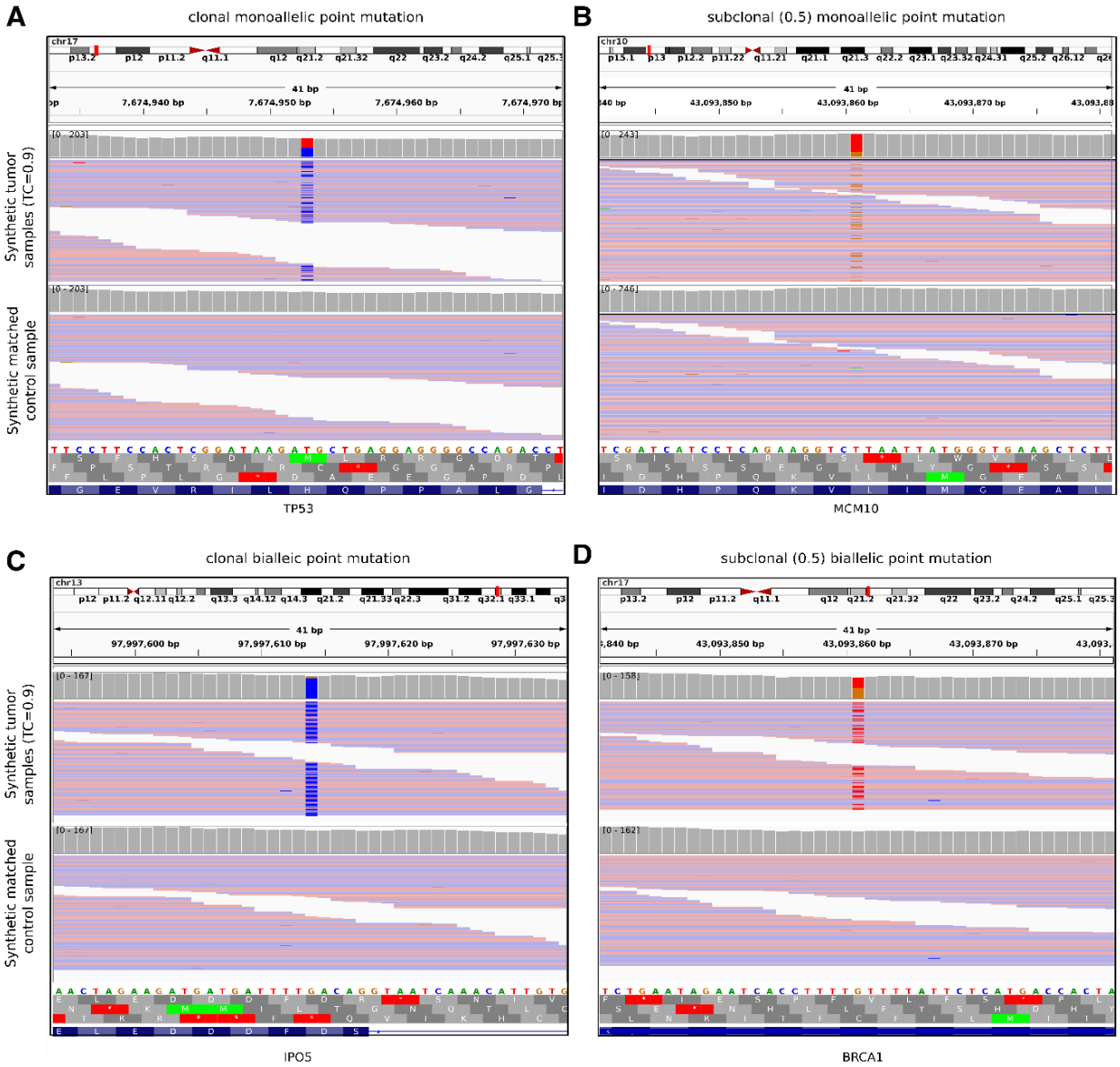

Figure 2.4 Examples of point mutations incorporated in a synthetic tumor sample with tumor content at 90%.

A) Monoallelic point mutation with 100% clonality; B) Subclonal (50% clonality) monoallelic point mutation; C) Biallelic point mutation with 100% clonality; D) Subclonal (50% clonality) biallelic point mutation.

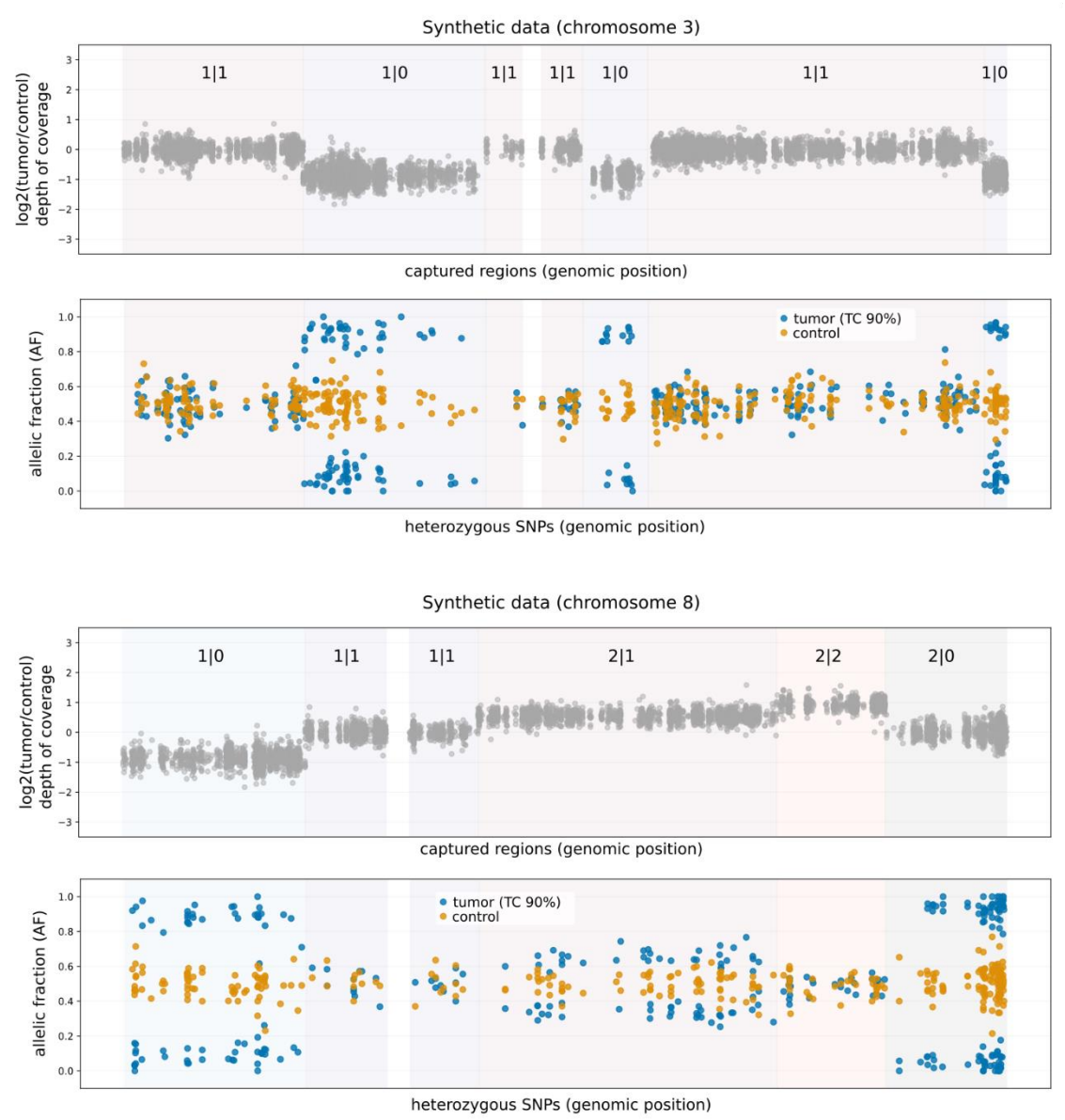

Figure 2.5 Examples of copy number aberrations incorporated in a synthetic tumor WES sample.

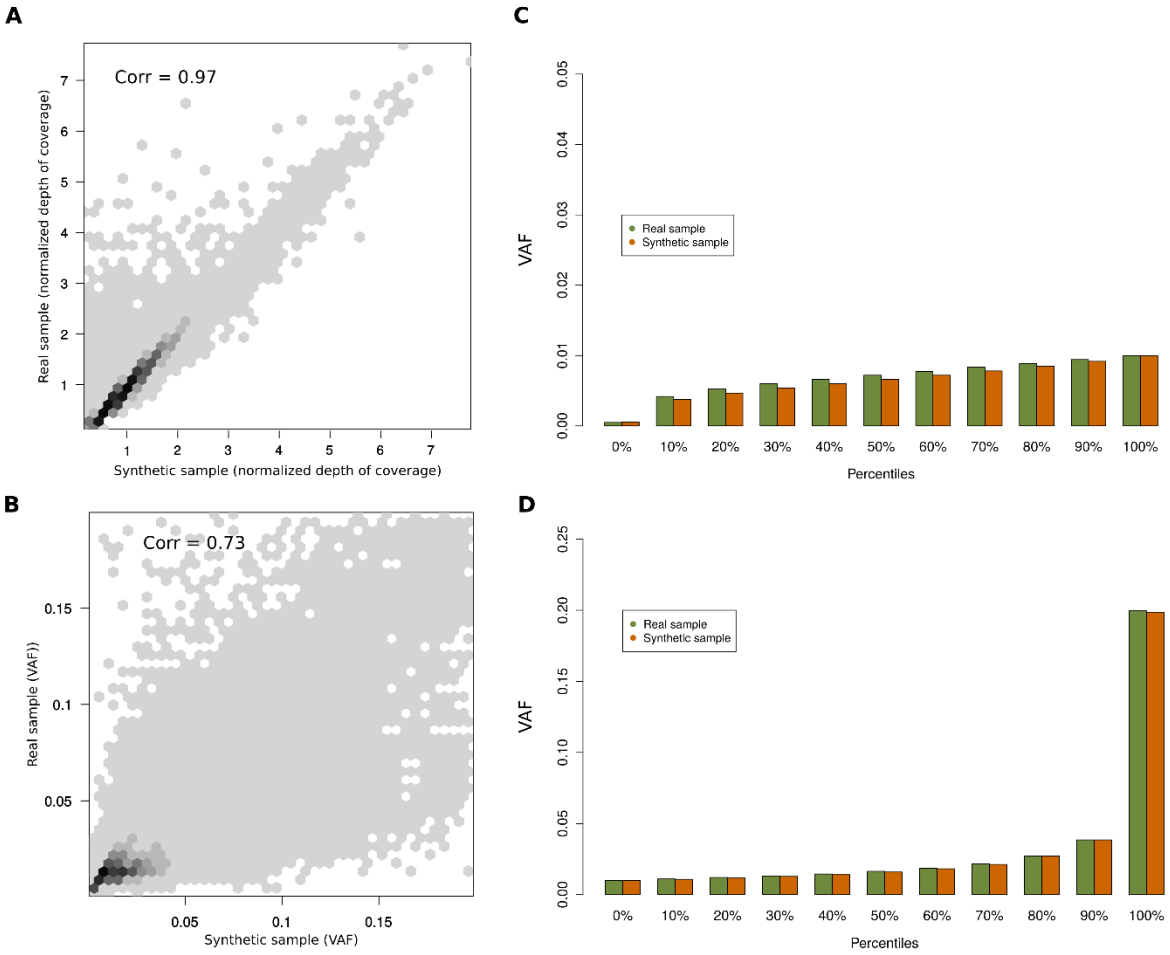

Figure 2.6 Synthetic versus real samples concordance of captured regions depth of coverage and genomic positions VAFs.

The synthetic WES sample is generated with reference models created from a real WES sample. A) Concordance of captured regions depth of coverage. B) Concordance of VAF values in the range (0,0.2], considering only positions with VAF>0 in both samples and with coverage >10x. C) Distributions of VAFs in the range (0,0.01] for both synthetic and real samples. D) Distributions of VAFs in the range (0.01,0.2] for both synthetic and real samples.

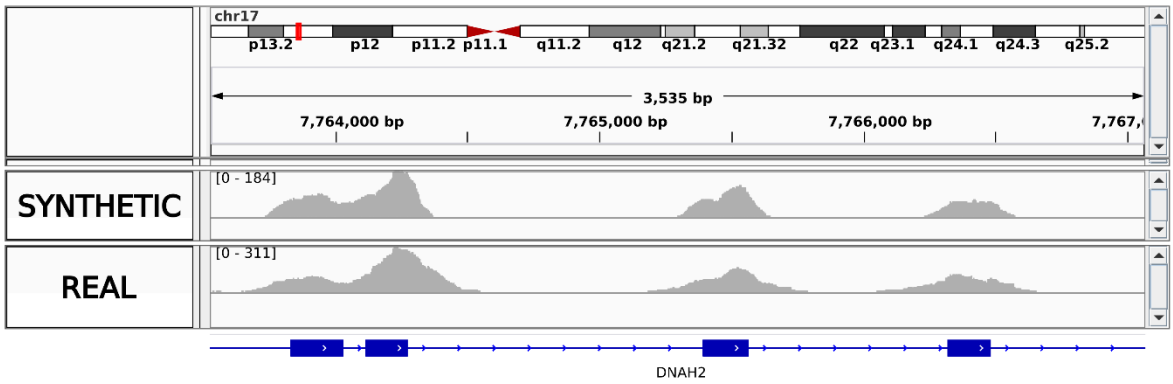

Figure 2.7 Example of depth of coverage distribution in synthetic and real NGS data across four genomic captured regions.

To demonstrate the effectiveness of synggen in generating benchmarking datasets, we simulated two liquid biopsy cfDNA scenarios considering data produced in (72,73). In those studies, different commercial TS panels that exploit different technologies were used. Considering genomic regions covered by both panels, we first generated synthetic NGS cancer data at decreasing tumor content, simulating a patient with both clonal and sub-clonal somatic copy number alterations and point mutations. Then, we generated synthetic NGS cancer data simulating a temporal sampling from a patient with regressing and emerging independent cancer sub-clones, at fixed tumor content. **Figures 2.8 and 2.9** demonstrate that log2 ratios of incorporated somatic copy number alterations and allelic fractions of incorporated somatic point mutations are well represented in both simulated scenarios and across all generated cancer samples' data but also demonstrate that the generated data consistently preserve the different and platform-specific genomic characteristics of the two considered TS panels. See Materials and Methods for additional details.

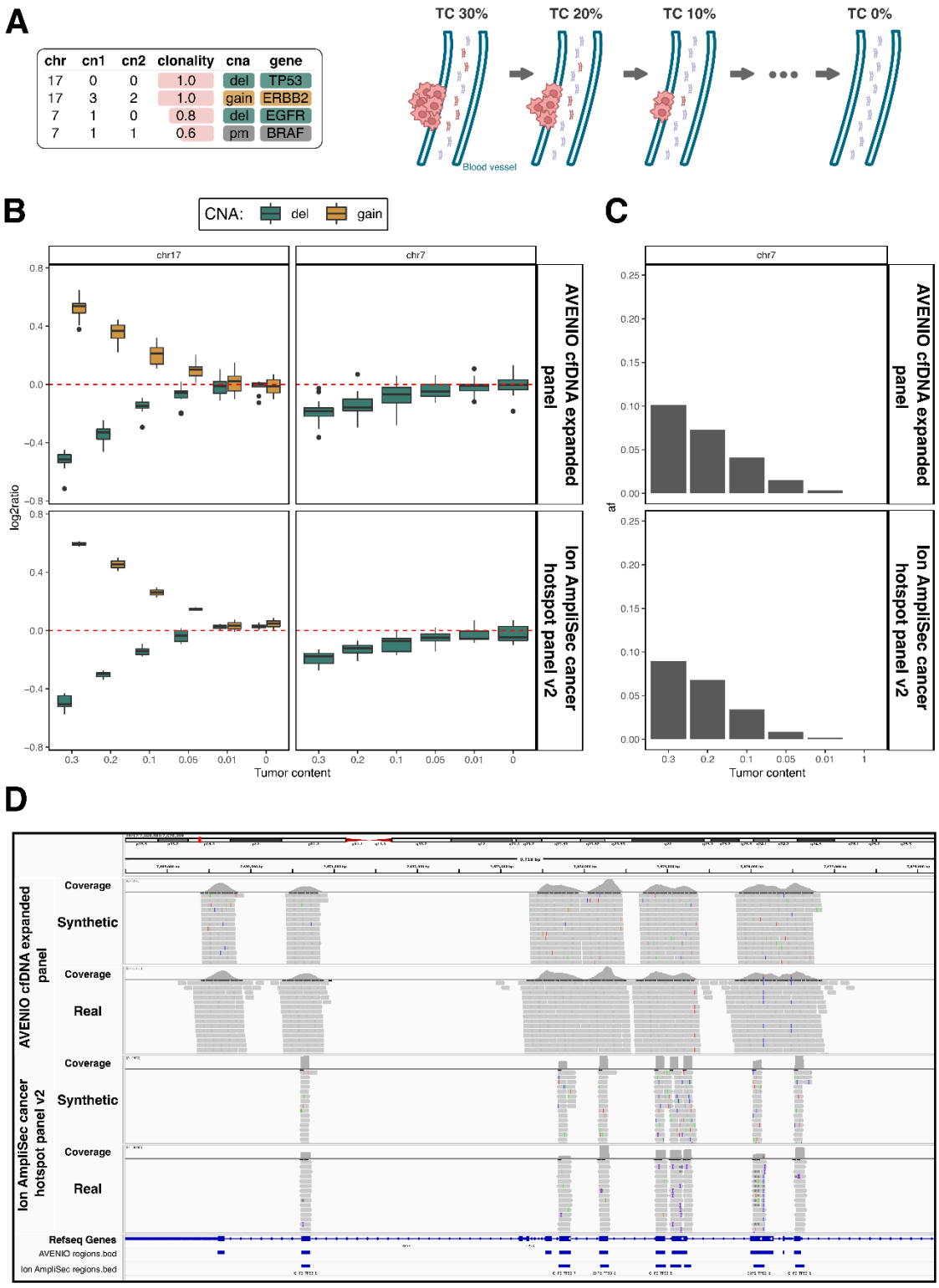

Figure 2.8 Tumor diluted samples simulation exploiting cfDNA data produced in (Qvick et al., 2021) and (Kaisaki et al., 2016).

A) Experiment design, generation of different tumor content samples from 30% to 0%. Table representing introduced copy number aberrations and point mutations at different clonalities. B) Copy number changes in tumor content dilution displayed through log2 ratio of tumor sample coverage over control sample coverage for both panels. C) Point mutation allelic fraction changes across different tumor contents for both panels. D) IGV visualization of synthetic sample depth of coverage distribution versus a real sample for both panels.

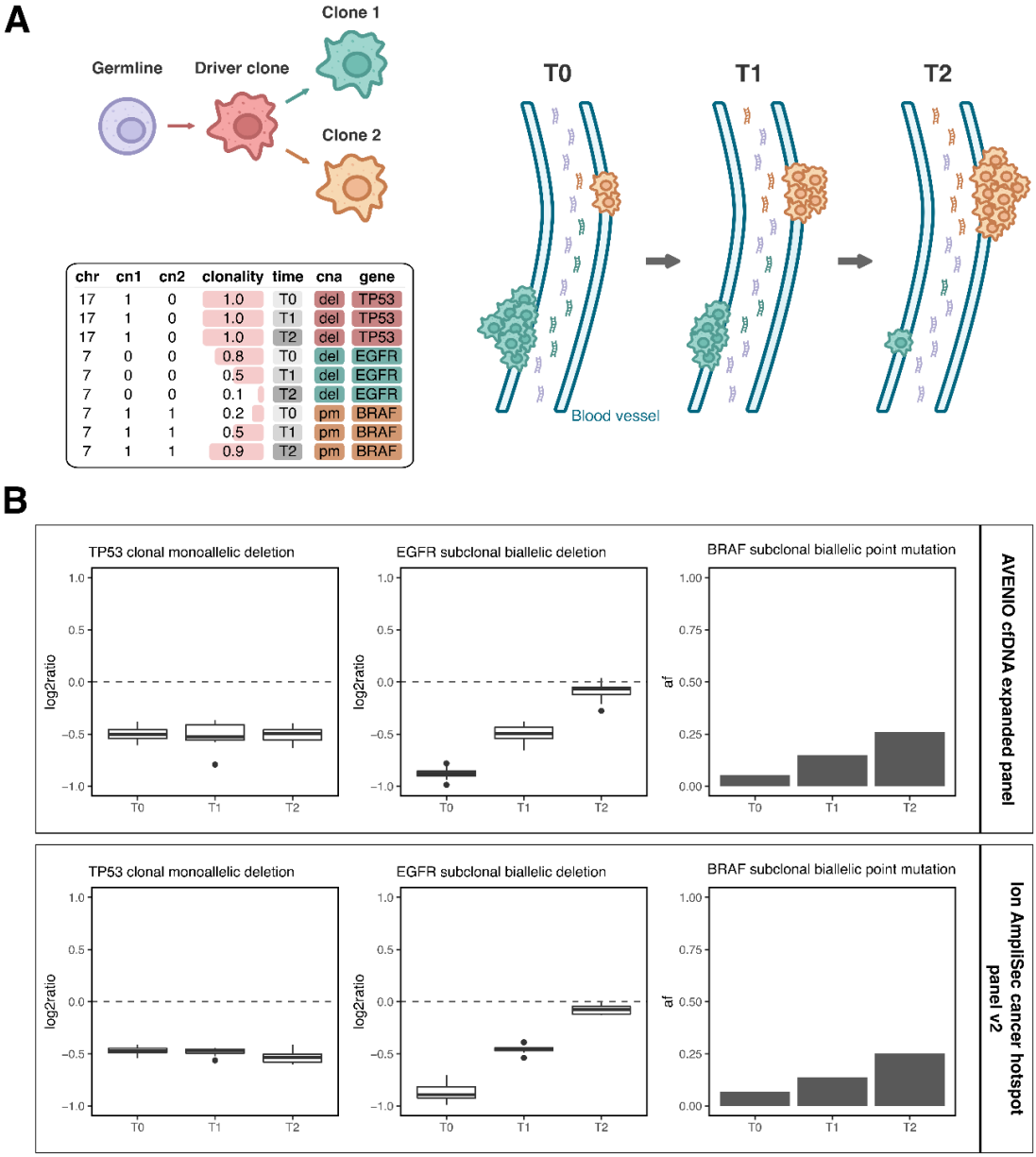

Figure 2.9 Temporal sampling from a patient with changing tumor sub-clones' populations simulation exploiting cfDNA data produced in (Qvick et al., 2021) and (Kaisaki et al., 2016).

A) Experiment design, generation of samples at 60% tumor content with a shared clonal deletion and private sub-clonal deletions and point mutations dynamically changing through three time points. Table representing introduced copy number aberrations and point mutations at different clonalities and time points. B) Changes over the time points of the three incorporated alterations in both datasets: fixed clonal deletion and sub-clonal deletion represented as coverage log2 ratio, and sub-clonal mutation represented as allelic fraction.

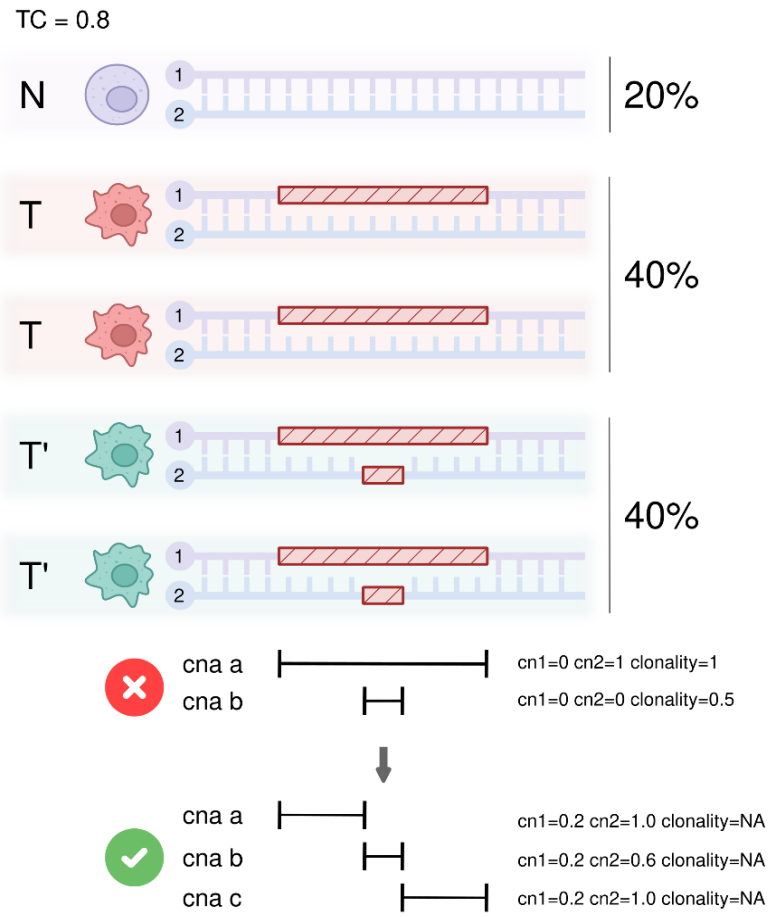

Figure 2.10 Example of complex nested somatic copy number aberrations definition in synggen.

TC represents the tumor content. N represents the fraction of normal contribution, while T and T' represent fractions of different tumor clones' contributions. The bottom part of the figure shows how a nested copy number definition should be defined in synggen.

2.4 Discussion

We developed synggen to perform fast and scalable generation of cancer and matched control WES and TS synthetic data without the need for intermediate FASTA files generation and exploiting multi-threaded computation. Synggen allows to emulate platform-specific NGS data characteristics and allows to incorporate germline-phased SNPs and complex somatic allele-specific copy number alterations and point mutations. In addition, it allows to create clonal and sub-clonal events across different tumor content scenarios enabling the generation of large-scale complex benchmarking datasets. Synggen is easy to use and is provided with a collection of additional scripts to further simplify its usability.

2.5 Materials and Methods

2.5.1 Read Depth Model (RDM)

To emulate the variability of the depth of coverage across platform-specific whole-exome sequencing (WES) and targeted sequencing (TS) captured regions, synggen synthetic reads are generated using a model RDM which measures the frequency of observing sequencing reads at any captured genomic region and position. Although an RDM can be synthetically designed, synggen provides a specific execution mode to generate an RDM starting from a set of control platform-specific BAM files (non-tumor samples). Exploiting an adapted version of the multi-threaded computational pileup engine we previously proposed in (Valentini et al., 2019) (70), multiple BAM files are processed and a single-base pileup is performed across all captured genomic regions and positions to count, for any genomic position p, the number of high quality reads (mapping quality over a certain threshold tunable by the user, default 10) having start alignment position p, and finally aggregating these counts at the level of captured genomic regions. These counts are saved in a gzip compressed *.rdm.gz file with each line reporting counts in the following format:

…

34;0;1;2;3;4;0;0;4;…

36;2;3;0;1;4;2;1;3;…

…

Each line represents a genomic captured region with the first number representing the count at the level of the region, followed by the counts across all positions in the region. Of note, when a paired-end protocol is used, only the read in the pair with the lowest alignment coordinate is used in the RDM computation. In addition, when paired-end protocol is used, insert-size statistics (average and standard deviation) are computed and saved at the beginning of the RDM.

2.5.2 Quality Model (QM)

To emulate platform-specific base qualities, sequencing errors are introduced by synggen in synthetic reads following a quality model QM, which measures the distribution of base qualities across read positions. Also in this case, although a QM can be synthetically designed, synggen quality models are created in parallel with RDM starting from a set of control platform-specific BAM files. More specifically, while fetching reads from control BAM files during the pileup computation, synggen inspects the reads base qualities across all positions to create a global quality model. The counts are saved in an gzip compressed *.qm.gz file having the following format:

0	1	0	0	…	1201	1340	…	200
0	1	2	0	…	1102	1621	…	112
…								
1	3	2	1	…	1471	1050	…	78

where columns (41 columns) represent phred quality scores from 0 to 40, rows represent read positions (specified by the user, default 100 positions) and values at a specific row and

column represent the number of time a quality has been observed at that read position. To eliminate potential outliers, quality distributions across read positions are smoothed using a slicing window approach:

$$q'_{ij} = \begin{cases} q_{i-1j} + q_{i+1j}, & q_{i-1j} * 3 < q_{ij} \; and \; q_{i+1j} * 3 < q_{ij} \; and \; (q_{i-1j} > 0 \; or \; q_{i+1j} > 0) \\ q_{ij}, & otherwise \end{cases}$$

$$with \; i \in \{1, \dots, 39\} \; and \; j \in \{1, \dots, read\_length\}$$

with q'_{ij} representing the smoothed quality at column i and row j. Also in this case, no distinction is made between single-end and paired-end protocols at QM creation.

2.5.3 Position-Based Error (PBE) model

Platform-specific errors enriched at specific genomic positions are emulated by synggen using a position-based error model $PBEM$. The model measures, for each genomic captured position (not representing a common SNP), the probability of observing an error that is supported by high quality reads and bases (read and base qualities above a user-specific threshold, default 10) and with a VAF above a specific threshold (default 1%). Also in this case, although a $PBEM$ can be synthetically designed, synggen position-based error models are created in parallel with models RDM and QM starting from a set of control platform-specific BAM files. In detail, when the pileup of all input control BAM files is performed across all captured genomic regions and positions, for each captured genomic position p the number of reads supporting A, C, G and T bases at that position is computed and alternative base-specific errors are calculated and stored in a compressed *.pbe.gz file having the following format:

…

85|11|0|0|0|1;160|14|0|0|1|1; … ;596|17|0|1|1|1

411|54|0|1|2|2;513|65|1|1|2|3; … ;789|32|0|0|1|1

…

where each row represents a captured genomic region and each string 411|54|0|1|2|2 represents a base in the region with the first number (e.g. 411) representing the relative

position of the base in the region, the second number (e.g. 54) representing the total depth of coverage, and the remaining numbers (e.g. 0,1,2 and 2) representing the cumulative number of errors found across the bases A, C, G and T (reference bases are not considered in the count). In the example we have one error with base C and one error with base G. To avoid capturing private SNPs when few BAM files are used in the construction of the model, positions with overall VAF>20% are not included in the $PBEM$. Also in this case, no distinction is made between single-end and paired-end protocols at PBE creation. It is important to consider that synggen does not embed duplication rate in its model, a commonly caused Polymerase Chain Reaction (PCR) artifact during the sequencing approach, therefore deduplication step is recommended before providing inputs to synggen.

2.5.4 Single Nucleotide Polymorphisms (SNPs)

SNPs are provided to synggen using different file formats depending on its execution mode. For reference models creation, a standard VCF file with common SNPs is provided, while for synthetic reads generation, sample specific SNPs are provided through a file with a specific format listing genomic coordinates of SNP positions, alternative bases and phased genotypes (allele 1 and allele 2, respectively):

…

chr1	957418	A	0	1
chr1	961678	T	0	0
chr1	965180	C	1	1
chr1	999792	A	1	0

…

2.5.5 Somatic Copy Number Alterations (CNAs)

CNAs are provided to synggen as allele specific copy numbers through a file with a specific format, listing genomic CNA regions and characteristics:

…

chr8	100529524	105561481	4	2	0.8

```
chr3    200765464    200798862    0.6    1.0    NA
…
```

Each line of the file represents a somatic copy number alteration and specifies the chromosome, the starting and end genomic positions, the allele-specific copy number (allele 1 and allele 2, respectively) and the clonality of the somatic alteration. There are two ways to specify allele-specific copy numbers:

- Define integer allele-specific copy numbers for both alleles and a value for the clonality of the copy number. This mode is simpler but allows to specify only one copy number per genomic captured region. Hence, nested/overlapping copy numbers cannot be specified in this way.
- Define allele-specific copy numbers using real values. This mode is more general and allows to specify complex scenarios were the observed copy number of a genomic region could be the result of different clones carrying different copy numbers for that region. When this mode is selected, the real values used to specify the allele-specific copy number should already consider the contribution of the global tumor content and the CNA clonality; the clonality value should hence be set to “NA” (if a value is specified, it is ignored by synggen).

2.5.6 Somatic point mutations (PMs)

Somatic PMs are provided to synggen as allele specific SNVs through a file with a specific format listing the SNVs’ coordinates and information:

```
…
chr8    100529524    A    0.5    1    2
…
```

Each line of the file represents a point mutation and specifies the chromosome, the genomic position, the alternative base, the clonality of the mutation, the allele on which the mutation is incorporated (allele 1 or allele 2) and the number of allele copies carrying the mutation; the number of allele copies should be consistent with the somatic allele-specific copy number specified for that genomic locus. Of note, biallelic point mutations are specified using two

different lines in the file; they indeed may have different clonality values and different number of alleles affected. In the example, a point mutation is incorporated with clonality 0.5 at position chr8:100529524 on two copies of the first allele, hence assuming that a somatic allele-specific copy number for the allele >=2 is specified in the CNAs file; a control is implemented by synggen to check the consistency between PM and CNAs definitions.

2.5.7 Implementation of WES/TS captured regions

Synggen implementation represents captured regions with an ad-hoc data structure that allows scalable multi-threaded analysis both at the level of reference models creation and at the level of synthetic reads generation. Regions are provided to synggen in input with a BED file. When loaded, the genomic coordinates in the BED file are extended by a size equal to N*read length (N=1 for the single-end protocol and N=2 for the paired-end protocol) to realistically model the distribution of reads in WES and TS captured regions. Of note, read length is specified by the user (default 100bp) during the creation of reference models and the information is saved in the QM; loading a QM when synthetic reads are generated will hence automatically set the read length.

Each captured region loaded from the BED file is associated with data structures storing characteristics and information of the (extended) region, together with characteristics and information of all single base positions defining the region. In particular, each region and its positions are associated with RDM corresponding counts, each region has associated an allelic fraction (AF) with default value 0.5 and each region has data structures listing germline SNPs present in the region and, when cancer data is generated, somatic CNAs and PMs affecting the region. All germline and somatic variants are provided to synggen using the file formats we previously described. Of note, when cancer data is generated and allele-specific CNAs are incorporated, RDM counts of affected regions are adjusted by one of the following formulas to correctly represent the region-specific depth of coverage. Specifically, when the CNA is specified using integer values, the used formula is:

$$RDM(r)' = RDM(r) * (1 - (tc * CN^{r}_{clon})) + RDM(r) * \left(\frac{CN^{r}_{1} + CN^{r}_{2}}{2}\right) * (tc * CN^{r}_{clon})$$

In the formula, r is the region, CN^r is the CNA with CN_1^r and CN_2^r being the copy numbers associated to the two alleles, CN_{clon}^r is the clonality of CN^r and tc is the sample tumor content. When the CNA is instead specified using real values, the formula used is:

$$RDM(r)' = RDM(r) * \left(\frac{CN_1^r + CN_2^r}{2}\right)$$

As previously highlighted, real valued representation of CNAs should already consider the contribution of global tumor content and local CNA clonality, hence these two values are not used in the formula. Similarly, RDM values associated to single base positions are updated by the formulas:

$$RDM(p|r)' = RDM(p|r) * (1 - (tc * CN_{clon}^r)) + RDM(p|r) * \left(\frac{CN_1^r + CN_2^r}{2}\right) * (tc * CN_{clon}^r)$$

or

$$RDM(p|r)' = RDM(p|r) * \left(\frac{CN_1^r + CN_2^r}{2}\right)$$

were p is a single base position of the region r. In addition, to correctly represent phased SNP allelic fractions in CNAs affected genomic regions, AF values associated to CNA affected regions are updated by the following formulas:

$$AF_r' = \frac{CN_1^r * (tc * CN_{clon}^r) + \left(1 - (tc * CN_{clon}^r)\right)}{(CN_1^r + CN_2^r) * (tc * CN_{clon}^r) + 2 * \left(1 - (tc * CN_{clon}^r)\right)}$$

$$= \frac{(CN_1^r - 1) * tc * CN_{clon}^r + 1}{(CN_1^r + CN_2^r - 2) * tc * CN_{clon}^r + 2}$$

or

$$AF_r' = \frac{CN_1^r}{CN_1^r + CN_2^r}$$

After RDM counts adjustment, region-specific and position-specific RDM cumulative counts are calculated and used to determine the probabilities used to generate reads starting at any specific position across all captured genomic regions.

2.5.8 Definition of complex nested CNAs

As previously described, to define complex nested or overlapping CNAs in synggen, real values should be used to define allele-specific copy numbers. An example is reported in **Figure 2.10**

where we depict the generation of a sample with tumor content at 80% incorporating a clonal (100%) monoallelic deletion and a nested sub-clonal (50%) biallelic deletion. In this scenario, the description of the two nested copy numbers using integer allele-specific values with the specification of CNAs clonalities (first description in the bottom) would not work because the integer-based formulas used for RDM and AF adjustments assume that the portion (1-tc*clonality) is copy number neutral. To represent this scenario, real allele-specific values can be used instead. In detail, the description of the nested CNAs should be unrolled into a description of three non-overlapping CNAs, each of them described using real values and indicating "NA" for the clonality value. Specifically, cna a and cna c in the bottom definition will have a value of 0.2 for CN_1, indicating that 20% of the sample has that genomic portion in allele 1, and value of 1 for CN_2, indicating that 100% of the sample has that genomic portion in allele 2. Description of cna b should instead indicate a value of 0.2 for CN_1, indicating that 20% of the sample has that genomic portion in allele 1, and value of 0.6 for CN_2, indicating that 60% of the sample has that genomic portion in allele 2. Of note, synggen CNAs' input file could be designed exploiting both integer and real valued allele-specific copy number specifications to define different CNAs in the same input file. Synggen will inspect each CNA definition and apply the right formula for the adjustment of RDM and AF values.

2.5.9 Generation of sequencing reads

Synggen generates sequencing reads using an efficient multi-threaded algorithm that implements the following steps:

1) A genomic position p part of an extended captured region r is sampled from the probability distribution described by the RDM. Selection is performed in two steps. First a region is sampled and then a position in the region is sampled. Once sampled, information of the selected region and position are rapidly retrieved using a binary search in our ad-hoc data structures.
2) The sequence representing the read is built accessing to the reference genome FASTA. Specifically, a sequence from p to $p + read\_length$ is retrieved. If paired-end generation is active, an additional read with sequence from $p + read\_length + insert\_size$ to $p + read\_length + insert\_size + read\_length$ is generated. Insert

size value is sampled from a normal distribution with mean and standard deviation as specified in the RDM.

3) One of the two alleles is selected based on a probability that is proportional to the region AF.
4) SNPs are incorporated in the read(s). Specifically, if the genomic region represented by the generated read(s) contains one or more SNPs on the selected allele, then the read(s) sequence is modified to incorporate the SNPs alternative bases.
5) Errors are incorporated in the read(s) and base qualities are accordingly generated. Specifically, for each genomic position represented by the read(s) bases, we first check if the position is annotated in our PBE model. If that is the case, an error is incorporated in the read by changing the base to another base as specified by the PBE model for that genomic position. In this case the quality of the base is extracted from the QM considering the specific position of the base in the read and considering only the distribution of qualities > mbq, the parameter the user can set (by default mbq is equal to 10) to determine the PBE using only high quality bases. If the genomic position is not in the PBE, then we extract a quality from QM, considering the position of the base in the read and sampling from the qualities distribution at the corresponding QM entry. The selected quality is then used to determine the base error probability:

$$error = 10^{\left(-\frac{quality}{10}\right)}$$

The calculated error probability is then used to eventually incorporate the error in the read, changing the current base with another base that is randomly selected. Of note, at the end of these steps, the base qualities for all read bases are available.

6) PMs are incorporated in the read(s). Specifically, for each genomic position represented by the read(s) bases, if a PM is specified and is present in the selected allele N of corresponding region r, then the mutation is incorporated with a probability:

$$p(PM) = tc * PM_{clon} * \frac{PM_{alleles}}{CN_N^r}$$

Value tc is the sample tumor content, PM_{clon} is the clonality of the point mutation, $PM_{alleles}$ is the specified number of alleles carrying the mutation and CN_N^r is the number of copies for the selected allele associated to the region (1 in case of no copy number or as specified by the user in the CNA file). As previously highlighted, consistency of $PM_{alleles}$ is checked by synggen during PMs loading.

7) A strand is selected based on a probability proportional to the region strand bias (default 0.5) and read(s) sequences are complemented and reversed depending on the selected strand.
8) Generated read(s) is(are) written to the output file in FASTQ format.

As described, the generation of sequencing reads is multi-threaded and the final number of FASTQ files will be proportional to the number of cores used in the parallel computation.

2.5.10 Generation of benchmarking data for performance analyses

To test the performances of synggen, we generated synthetic samples and measured the computational time and memory required for reference models' construction and sequencing reads generation. These information are available from synggen standard output. The benchmark analyses were performed on a HPE Proliant DL560 server.

Two scenarios were tested, WES and TS, for both model and synthetic sequencing reads generations. Each benchmark condition in each scenario was repeated three times and averaged results with standard deviation were evaluated and visualized.

Ten samples were selected among breast cancer control (non-tumor) samples (BAM files) available from The Cancer Genome Atlas (TCGA) dataset having matched high (>80%) tumor content WES (Sure Select All Exome v2) cancer tissue sample. BED files used consist of: i) WES kit's captured/target regions BED file (35,937,293 base pairs); ii) a BED file generated selecting 100 genes randomly from the larger WES BED (5,811,848 base pairs). Samples used in the TS scenario were generated subsampling each WES sample given the smaller TS BED file. The construction of reference models was tested for an increasing number of input samples (1 to 10) and increasing number of threads (1, 2, 4, 8, 16). Generation of synthetic NGS cancer data was instead tested for increasing number of sequencing reads produced (100x and 500x

calculated based on bed regions dimension and read size of 100bp) and increasing threads (1, 2, 4, 8, 16). A cancer sample was generated with tumor content at 80% and incorporating patient's specific germline SNPs and somatic CNAs retrieved from cBioPortal (www.cbioportal.org).

2.5.11 Generation of benchmarking data for liquid biopsy cfDNA scenarios

To demonstrate the effectiveness of synggen in generating benchmarking datasets, we focused on liquid biopsies cfDNA data produced in (72) and (73), and simulated two benchmarking scenarios: generation of NGS cancer data at decreasing tumor content, and generation of NGS cancer data simulating temporal sampling from a patient with dynamic tumor sub-clones' populations.

The two datasets are based on the use of different commercial targeted sequencing panels which exploit different technologies. In (72) the AVENIO cfDNA Expanded Panel with paired-end sequencing protocol was used, while in (73) the Ion AmpliSeq Cancer Hotspot Panel v2 was used (single-end sequencing protocol).

For both datasets we first generated reference models, then, synthetic NGS cancer data was generated for both benchmark scenarios. Models and reads generated were based on paired-end protocol for AVENIO data, while were based on single-end protocol for Ion AmpliSeq data.

Considering the first benchmarking scenario, we generated synthetic NGS cancer data at decreasing tumor content (30%, 20%, 10%, 5%, 1%, 0.5%, 0%) simulating a patient with both clonal and sub-clonal CNAs and PMs (**Figure 2.8A**). A different sample at tumor content 0% was generated as control. Generated data was analyzed showing log2 ratios (tumor samples over control sample) of incorporated somatic CNAs (**Figure 2.8B**) and allelic fractions of incorporated somatic PMs (**Figure 2.8C**).

Considering the temporal sampling scenario, we generated synthetic NGS cancer data simulating a patient with regressing and emerging independent tumor sub-clones at fixed tumor content 60%. In particular, as shown in **Figure 2.9A**, we generated three time points samples, incorporating a shared clonal monoallelic deletion, a private biallelic sub-clonal deletion in clone 1, and a private biallelic point mutation in clone 2. Presence of tumor signal representing clone 1 in circulation was increasing (emerging clone) across the three time

points while presence of tumor signal representing clone 2 was decreasing (regressing clone) across the three time points. The generated data was analyzed showing for each incorporated somatic alteration its dynamic change through the three time points, using log2 ratios for CNAs and allelic fractions for PMs (**Figure 2.9B**).

Chapter 3 - Enabling sensitive and precise detection of tumor signals through somatic copy number aberrations in cfDNA from breast cancer patients

Authors: Riccardo Scandino, Agostina Nardone, Nicola Casiraghi, Francesca Galardi, Mattia Genovese, Dario Romagnoli, Marta Paoli, Chiara Biagioni, Andrea Tonina, Ilenia Migliaccio, Marta Pestrin, Erica Moretti, Luca Malorni, Laura Biganzoli, Matteo Benelli, Alessandro Romanel

Journal: npj Breast Cancer, March 2025

DOI: https://doi.org/10.1038/s41523-025-00739-6

This chapter focuses on the development and evaluation of eSENSES, a novel targeted next-generation sequencing (NGS) panel designed specifically for comprehensive characterization of ctDNA in breast cancer through the analysis of SCNAs. eSENSES distinguishes itself through a unique combination of elements: genome-wide and focal single nucleotide polymorphisms (SNPs), exonic regions of key cancer-related genes, and a bespoke computational strategy. This multifaceted approach enhances the detection and quantification of somatic copy number alterations (SCNAs) and circulating tumor DNA (ctDNA) fraction.

This project was carried out in collaboration with the Department of Oncology, Hospital of Prato, Azienda USL Toscana Centro, Italy.

My principal contribution to this project centers on the implementation, development, and evaluation of the computational framework underpinning eSENSES. This involved the design and optimization of algorithms that leverage SNP allelic fraction information for precise copy number calling, thus enabling accurate quantification of ctDNA and identification of SCNAs.

This chapter will provide a detailed exposition of the eSENSES panel's design and development, with particular emphasis on the computational methodology and its validation using both synthetic and clinical samples. The overarching aim is to present a robust and cost-effective tool for ctDNA analysis, thereby contributing to improved monitoring of disease progression and informing therapeutic decisions in breast cancer patients.

3.1 Introduction

Breast cancer is a multifaceted disease affecting millions of women worldwide. It is characterized by diverse histological characteristics, metastatic potential and treatment responses. In both early and advanced stages, tissue sampling is essential for pathological assessments and selecting biomarker-based therapies (52). However, these procedures can delay diagnosis and are often complicated by tumor heterogeneity, sampling limitations, and their invasive nature.

Circulating biomarkers, such as cell-free DNA (cfDNA), are gaining popularity as non-invasive new biomarker source in the field of oncology (12,74). While early detection of ctDNA signal promises to improve cancer diagnosis (75,76), the quantification of tumor-derived content in plasma cfDNA (referred to as ctDNA) from cancer patients can be used to monitor the response to therapeutic interventions (77–80). Indeed, plasma cfDNA carries the genomic characteristics of tumor cell material shed into the bloodstream and in the presence of metastatic disease and/or of multifocal tumors, where single tissue biopsies would fall short in allowing heterogeneity assessment, ctDNA represents an ideal alternative to capture the disease genomic features. Although ctDNA fraction in plasma cfDNA samples ranges typically from 0% to 25%, several studies demonstrated the prognostic value of ctDNA and the ability to track tumor dynamics through the analysis of genomic lesions detected in the circulation of cancer patients (81,82).

However, biological and technical issues can influence the ability of cfDNA next-generation sequencing (NGS) assays to accurately stratify patients, especially at low ctDNA fractions (83). Indeed, independently from the NGS approach that is used, low ctDNA fraction strongly influences and limits the detection sensitivity and specificity of different types of genomic events. In particular, somatic copy-number alterations (SCNAs) have been demonstrated extremely challenging to detect in samples with ctDNA below 20%.

SCNAs are changes in the number of copies of specific genomic regions and their detection and characterization provides information about the progression and aggressiveness of a cancer that are fundamental to guide treatment decisions. Breast cancer is a heterogeneous disease characterized by the presence of both large and focal SCNAs (84) and different breast cancer subtypes typically exhibit distinct SCNA patterns (85). Large SCNAs, involving extensive chromosome segments, contribute to genomic instability and resistance to conventional

therapies (86). Focal SCNAs, which affect specific genomic regions, can lead to the overexpression of key oncogenes like ERBB2 and CCND1 or the loss of tumor suppressor genes like TP53 and PTEN.

Various methods, including whole-genome sequencing (87), whole-exome sequencing (87,88), and targeted sequencing assays (60,89,90), have been developed to detect SCNAs from cfDNA and estimate ctDNA levels across different cancers. However, characterizing samples with ctDNA fractions below 10%-15% remains highly challenging, regardless of the method's advantages and disadvantages. Furthermore, none of the existing approaches can comprehensively characterize both large-scale and focal SCNAs. Hence, although SCNAs hold clinical relevance with important prognostic implications for localized and metastatic breast cancer patients, their detection and characterization through cfDNA NGS non-invasive assays is still hindered and innovative approaches are needed.

In this work, we propose an innovative and cost-effective breast-cancer specific platform that combines a custom NGS cfDNA panel with tailored computational approaches. This panel integrates genome-wide and focal common SNP loci to enhance the sensitivity and specificity of SCNA detection and ctDNA level. Additionally, it includes exonic regions of genes frequently mutated in breast cancer, aiding in the interpretation of complex genomic profiles. Ultimately, this technology aims to improve the implementation of liquid biopsies in clinical practice for breast cancer patients.

3.2 Results

3.2.1 Design and development of a breast cancer panel for the sensitive detection of SCNAs, SNVs and tumor signal from cfDNA

To enhance the detection and quantification of ctDNA and to enable precise characterization of genome-wide and focal somatic copy number aberrations SCNAs in breast cancer patients, we developed a custom targeted sequencing panel, named eSENSES (*e*nabling *S*ensitive and precis*E* detection of ctD*N*A through *S*omatic copy number ab*E*rrations in brea*S*t cancer). The panel leverages patients' genetic backgrounds to improve estimates of allelic imbalances. Because allelic imbalance can only be measured using inherited heterozygous SNPs, our panel

design is enriched for genome-wide and gene-specific SNPs with high global minor allele frequency (gMAF). We hypothesized that this approach would maximize the sensitivity of SCNA detection. Our goal was to achieve a genome-wide SNP density of 5 SNPs per megabase (Mb) and, based on previous work (60), an average of 150 SNPs per focal genomic region containing genes frequently deleted or amplified in breast cancer, including CCND1, ERBB2, PTEN, TP53, and ESR1.

Additionally, to facilitate parallel and effective profiling of Single Nucleotide Variants (SNVs), we incorporated insights from large-scale breast cancer genomic studies (5,31,32,49,91–93). This allowed us to include in our panel the exonic regions of genes that are recurrently aberrant in both localized and metastatic breast cancer (Supplementary Data 1). Overall, our panel comprised approximately 15,000 SNPs and over 2,000 exons from 81 genes, spanning a total of around 2.3 Mbp.

The panel is matched with a tailored computational approach that, exploiting the panel design, enables sensitive and specific detection of somatic copy number alterations and ctDNA detection and estimation also in patients with low ctDNA level. Briefly, we developed a novel SCNA detection algorithm (**Figure 3.1**) that integrates: (i) a read-depth estimation module which models the local and global coverage noise observed across all captured genomic regions and (ii) a SNP-based estimation module that is based on the detection of SNPs' allelic imbalance. These steps are then followed by estimation of tumor ploidy with data correction, copy number states computation and ctDNA level detection and estimation. The approach embeds the use of a pre-computed panel of controls that provides local and global read-depth and allelic fraction statistics for all captured genomic regions and SNPs and that is generated only once using a set of pre-defined control samples.

For an initial evaluation of eSENSES performance, we conducted plasma sequencing on samples from 20 healthy individuals. We examined the depth of coverage distribution across all captured genomic regions and explored the distribution of allelic fractions for SNPs. Overall, we observed both reasonable read duplication rates (on average 21.6%) and off-target reads (on average 18.1%). Considering processed data, we observed on average a sample's mean depth of coverage of about 750x (**Figure 3.2**) with roughly 80% of the captured genomic regions exhibiting stable relative coverage across the samples (CV<20%). In line with the panel design, each individual displayed approximately 40% heterozygous SNPs out of all those captured, ranging from a minimum of 38% to a maximum of 44%. Two samples

exhibited suboptimal distribution of SNPs' allelic fraction (**Figure 3.3**) and were excluded from further analyses.

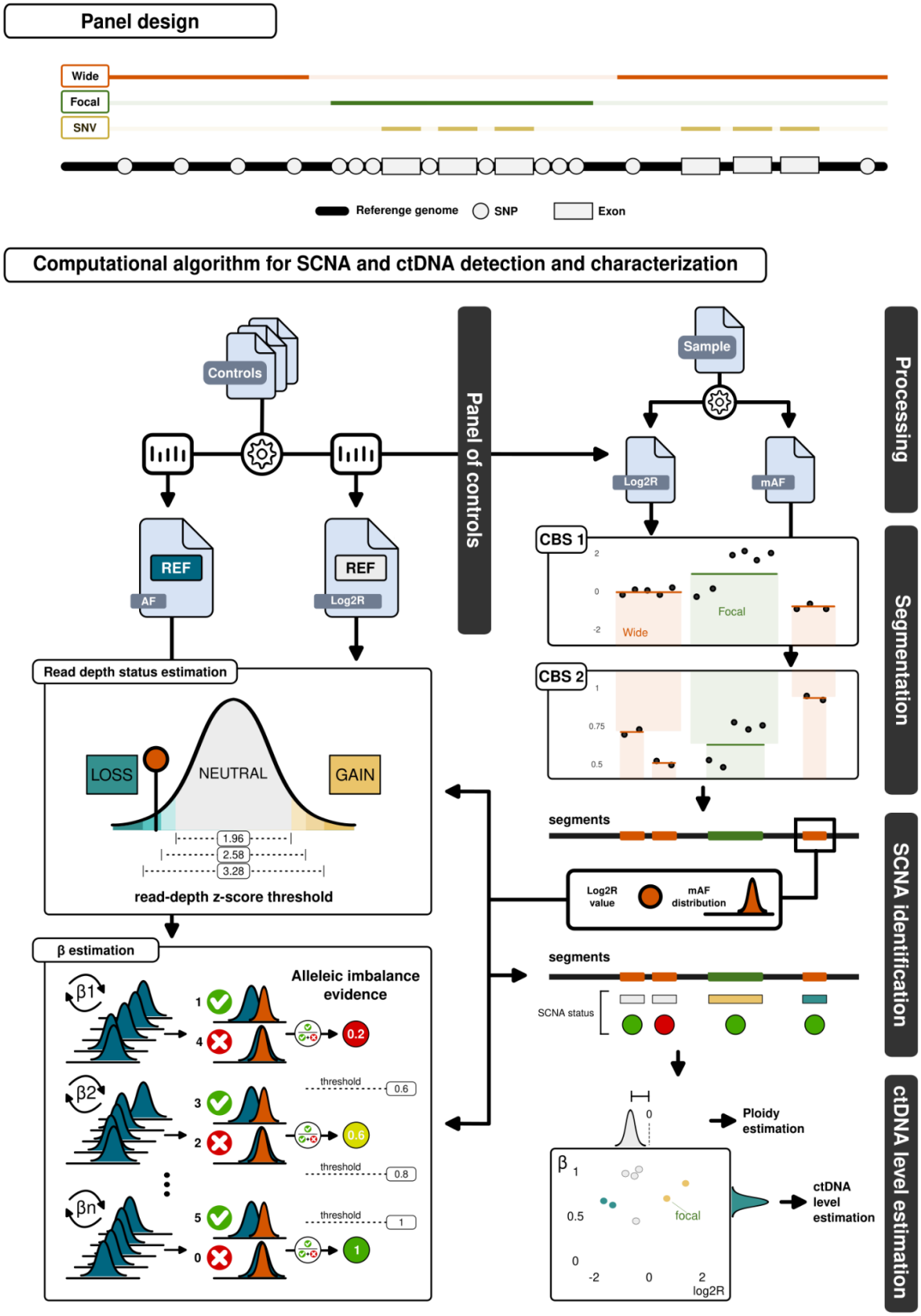

Figure 3.1 Panel design and eSENSES workflow.
The panel design is illustrated at the top of the figure. It includes targeted exons of common breast cancer genes and heterozygous SNPs evenly distributed across the genome ("wide"), as well as SNP-enriched regions specific to certain breast cancer driver genes ("focal"). Additionally, the exons within the wide regions are specifically designed to capture potential SNV events. The panel is paired with a customized computational workflow designed for the sensitive and specific detection of SCNAs and the estimation of ctDNA levels.

Control cfDNA samples are processed to generate statistics on SNP allelic fraction (AF) and log2 ratio (Log2R) reference distributions, which are then used to create a panel of controls. For a given patient's cfDNA sample, this panel computes Log2R values while mirrored allelic fractions (mAF) are processed and filtered. Circular binary segmentation (CBS) is applied to Log2R values, and a second CBS is performed on mAF values for each identified segment region. Focal regions are considered in their entirety and are not segmented. SCNA identification relies heavily on the panel of controls, which is used to estimate the read depth copy number status and its β for each obtained segment—a value proportional to the allelic imbalance of the SNPs in the segment. The sensitivity of these estimations for read depth status and β is adjusted through the parameters of read-depth z-score and allelic imbalance evidence. The assigned SCNA, along with its respective copy number status, β, and Log2R values, are then used to estimate the ploidy of the sample and the ctDNA level. These values can also be combined to explore the allele-specific (Log2R vs β) SCNA space.

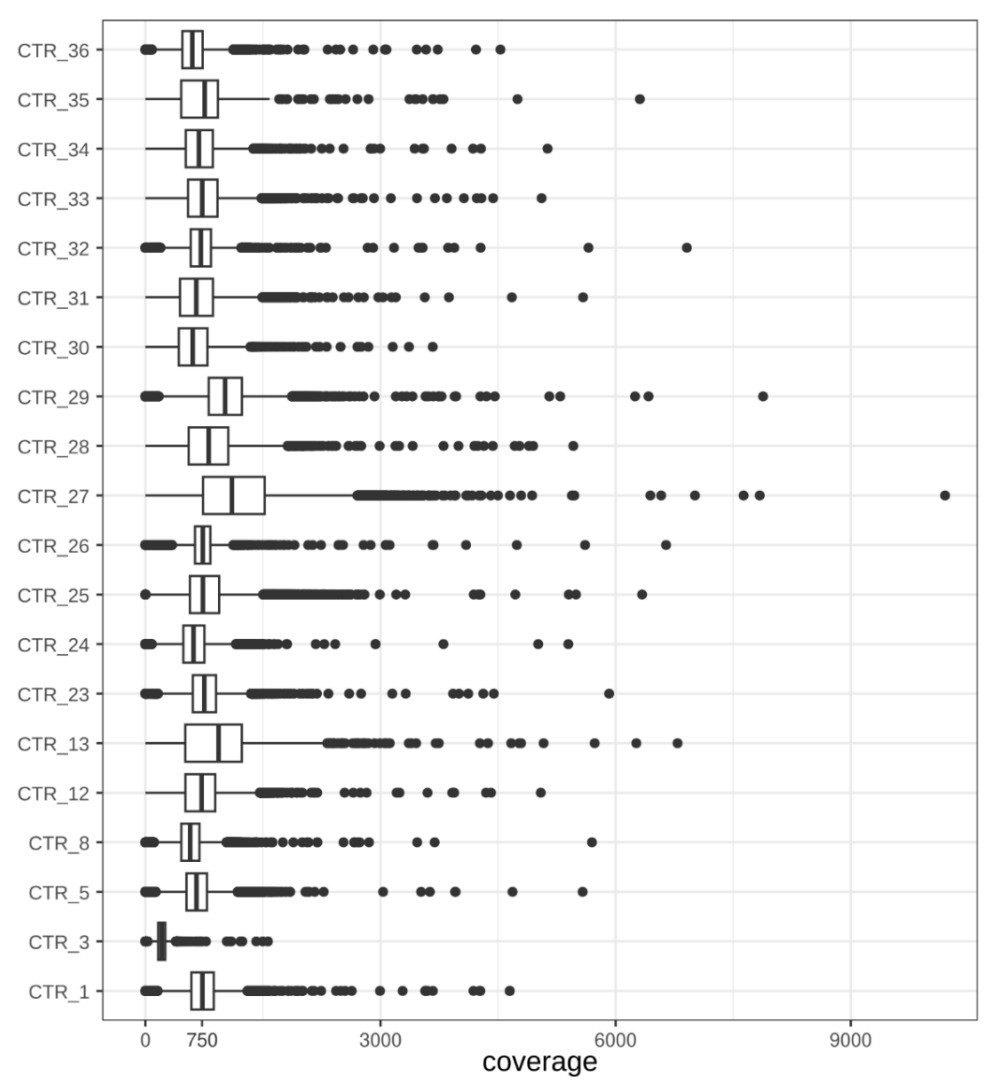

Figure 3.2 Processed control samples' distributions of read depth coverage.

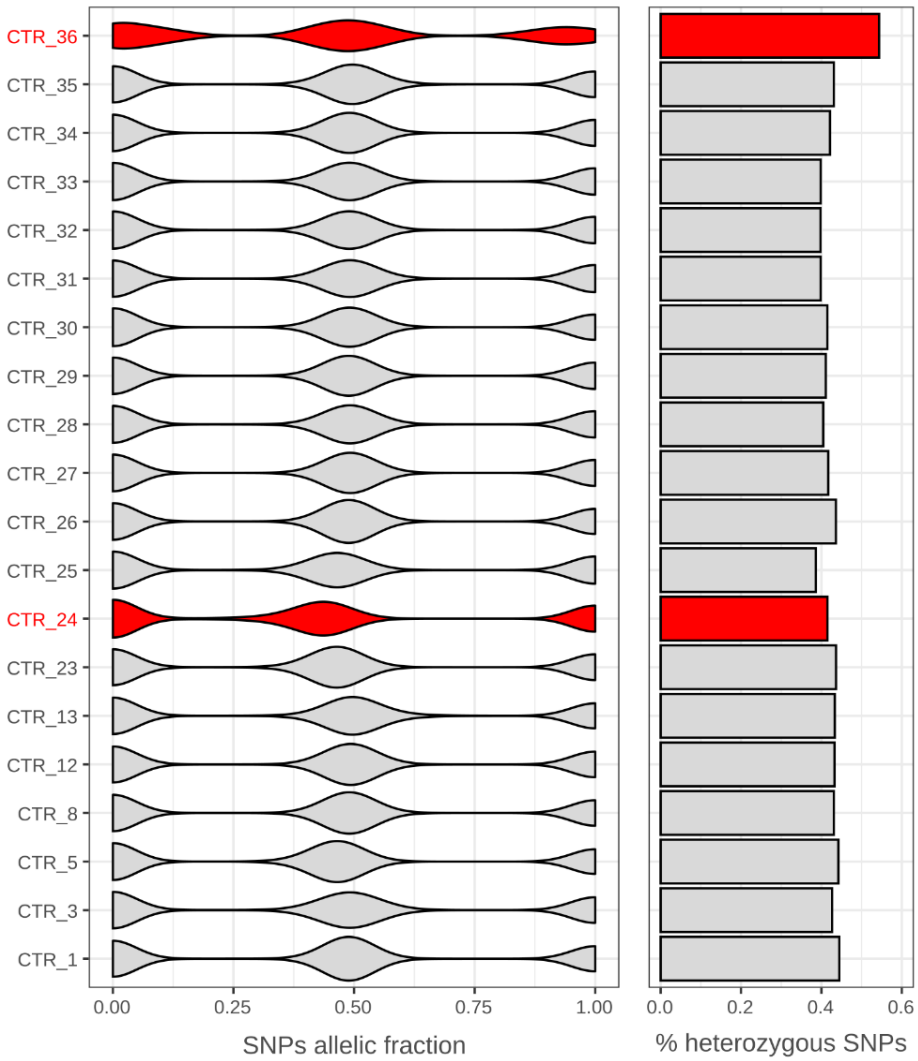

Figure 3.3 Processed control samples' distributions of allelic fractions with reported percentage of heterozygous SNPs (allelic fraction between 0.1 and 0.9) among total SNPs per patient. Patients highlighted in red were removed due to the not standard distribution of allelic fraction.

3.2.2 Performance of ctDNA detection and estimation

To assess the detection and estimation limits of our approach and to fine tune the parameters of our algorithm, we employed an in-silico approach and built a large benchmarking dataset exploiting the synggen computational tool (94). Synggen is a software we recently developed for the scalable generation of large-scale realistic and heterogeneous cancer sequencing synthetic datasets, including cfDNA synthetic datasets. Specifically, we utilized sequencing data derived from the plasma samples of the 18 healthy individuals (henceforth referred to as control samples) to produce a series of statistical models via synggen. These models encapsulate the distinctive data characteristics inherent to the panel, including coverage distribution across the captured genomic regions and sequencing error frequencies. Subsequently, these models were integrated with germline SNPs and somatic allele-specific SCNA profiles, respectively sourced from 1,000 Genomes Project and TCGA datasets, to produce realistic synthetic control and cfDNA samples. These synthetic samples were designed to encompass various levels of ctDNA and average depth of coverage. In particular, considering that the variability of allelic fraction estimation increase by decreasing the coverage we aimed at testing high coverage design of the panel and further explore whether performances could be improved on this aspect, an average coverage of synthetic samples spanned from ~800x to 2500x and 20 representative breast cancer samples from TCGA were used to generate synthetic samples with ctDNA levels spanning from 80% to 0%. Overall, we generated 940 cases/profiles among synthetic cfDNA and control samples.

Our approach's performance was assessed using the generated benchmarking dataset and various algorithm parameter combinations. The evaluation aimed to gauge and fine-tune eSENSES's ability to detect tumor signals within synthetic cfDNA samples, known as detection performance, as well as to estimate the ctDNA level. In detail, we started our evaluation by generating cfDNA synthetic samples with an average depth of coverage of 800x, representing the average coverage of the data we obtained by sequencing our control samples. Using realistic germline SNP haplotypes and somatic allele-specific SCNA profiles, we generated 260 cfDNA synthetic samples with decreasing ctDNA level, from 80% to 0.1%. Additional 40 synthetic samples with no embedded tumor signal were generated; half (N=20) representing control samples to create the panel of controls and half representing cfDNA samples with ctDNA level at 0% to measure the false positive rate (FPR) of our approach. The performance

of our approach was tested for different combinations of allelic imbalance evidence (from 0.2 to 1) and read-depth z-score (from 1.96 to 3.89) thresholds, representing the key parameters of our SCNA detection algorithm.

As shown in **Figure 3.4A**, no major detection performance deviations were observed across the different parameters' combinations. As expected, increasing the read-depth z-score might slightly reduce sensitivity, while decreasing allelic imbalance evidence might increase the False Positive Rate (FPR). Overall, we noted a sensitivity nearing 100% for ctDNA levels as low as 4%-5%, with a slight decline to 80%-90% when ctDNA levels reach 3%, and a rapid decrease for levels below 3%. Although supported by low sensitivity, our data suggest that our ctDNA detection limit is as low as 0.5%. When assessing ctDNA estimation performance (**Figure 3.4B** and **Figure 3.5**), we consistently observed a strong correlation with expected ctDNA levels (on average >99%). Of note, for synthetic cfDNA samples with expected ctDNA levels below or equal to 5%, less than 50% of the samples with positive ctDNA detection had also estimations available (**Figure 3.4C**); in addition, no ctDNA level estimations were available for samples with expected ctDNA level below 2%.

Next, we tested to what extent an increase in average depth of coverage would increase the performance of our approach. As illustrated in **Figure 3.4D** and **Figure 3.6**, the in-silico benchmarking indicates that increasing the coverage could potentially improve sensitivity of ctDNA detection to over 80% even for samples with an expected ctDNA level of 2%. However, achieving this would theoretically necessitate tripling the coverage. Similarly, as shown in **Figure 3.4E-F** and **Figure 3.6**, increasing the coverage would also increase the fraction of samples with ctDNA level estimation available. Nonetheless, it would still not enable estimations in samples with ctDNA levels below 2%.

Taking into account the outcomes of our benchmarking analysis, we opted for setting the read-depth z-score to 2.58 and the allelic imbalance evidence to 0.8. This configuration strikes a balanced compromise between sensitivity and specificity, optimizing performance in both ctDNA detection and ctDNA level estimation.

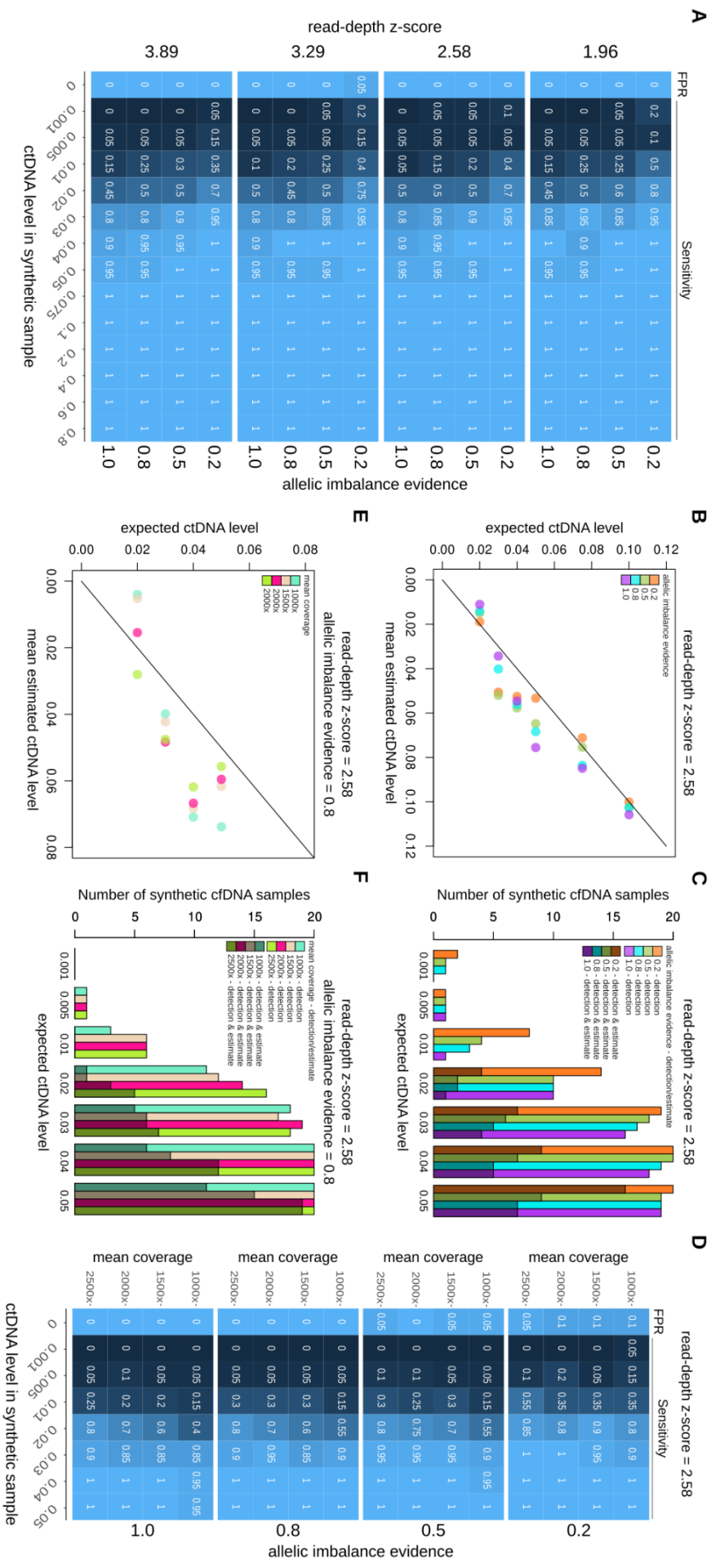
A
read-depth z-score
3.89
3.29
2.58
1.96
FPR
Sensitivity
ctDNA level in synthetic sample
allelic imbalance evidence
B
expected ctDNA level
mean estimated ctDNA level
read-depth z-score = 2.58
allelic imbalance evidence
C
Number of synthetic cfDNA samples
expected ctDNA level
read-depth z-score = 2.58
allelic imbalance evidence - detection/estimate
D
mean coverage
read-depth z-score = 2.58
E
expected ctDNA level
mean estimated ctDNA level
read-depth z-score = 2.58
allelic imbalance evidence = 0.8
mean coverage
F
Number of synthetic cfDNA samples
expected ctDNA level
read-depth z-score = 2.58
allelic imbalance evidence = 0.8
mean coverage - detection/estimate

Figure 3.4 eSENSES performances from the in-silico benchmarking.
A) *Sensitivity and FPR results of ctDNA detection for increasing simulated ctDNA level across different read-depth z-score and allelic imbalance evidence parameters' values.* ***B)*** *Concordance between the mean estimated ctDNA level, calculated across the results obtained for the simulated cfDNA samples, and the expected ctDNA level. Results are shown for read-depth z-score parameter equal to 2.58 and increasing allelic imbalance evidence parameter' values.* ***C)*** *Fraction of samples with ctDNA detection and ctDNA detection and ctDNA level estimation for expected ctDNA levels up to 5%. Results are shown for read-depth z-score parameter equal to 2.58 and increasing allelic imbalance evidence parameter' values.* ***D)*** *Sensitivity and FPR results of ctDNA detection for increasing simulated depth of coverage. Results are shown for read-depth z-score parameter equal to 2.58 and increasing allelic imbalance evidence parameter' values.* ***E)*** *Concordance between the mean estimated ctDNA level, calculated across the results obtained for the simulated cfDNA samples, and the expected ctDNA level. Results are shown for read-depth z-score parameter equal to 2.58 and allelic imbalance evidence equal to 0.8.* ***F)*** *Fraction of samples with ctDNA detection and ctDNA detection and ctDNA level estimation for expected ctDNA levels up to 5%. Results are shown for read-depth z-score parameter equal to 2.58 and allelic imbalance evidence equal to 0.8.*

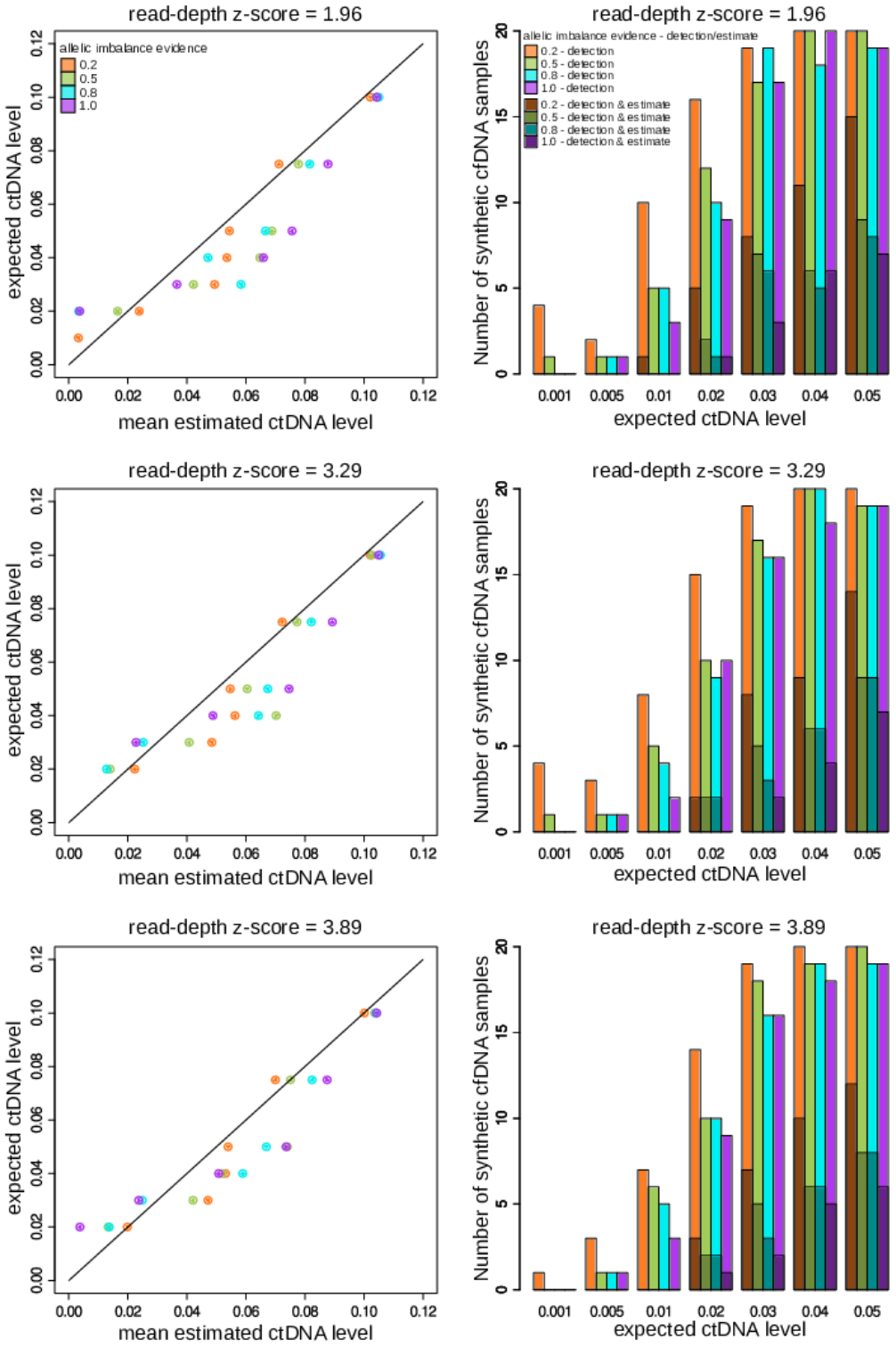

Figure 3.5 (Left) Concordance between the mean estimated ctDNA level, calculated across the results obtained for the simulated cfDNA samples, and the expected ctDNA level. Results are shown for increasing read-depth z-score parameter and increasing allelic imbalance evidence parameter' values. (Right) Fraction of samples with ctDNA detection and ctDNA dection and ctDNA level estimation for expected ctDNA levels up to 5%. Results are shown for increasing read-depth z-score parameter and increasing allelic imbalance evidence parameter' values

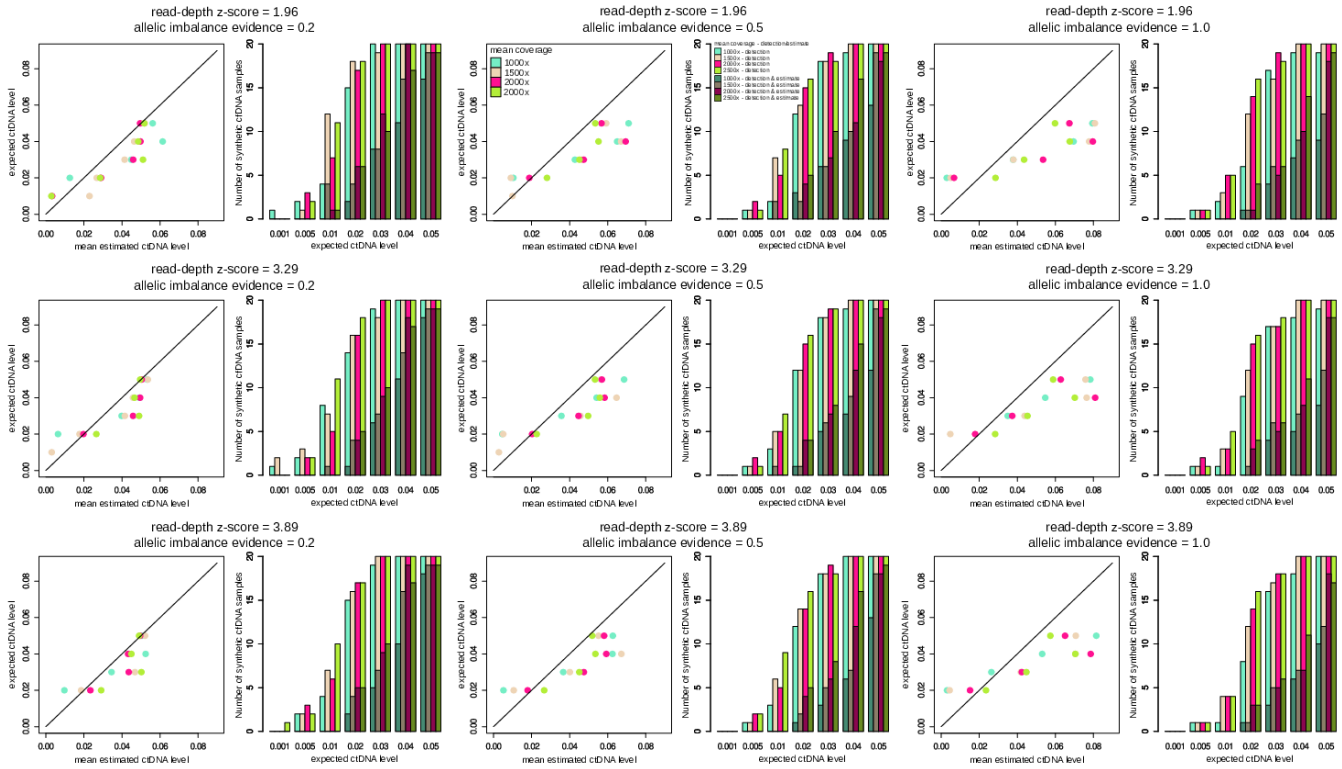

Figure 3.6 Scatterplots show the concordance between the mean estimated ctDNA level, calculated across the results obtained for the simulated cfDNA samples, and the expected ctDNA level.

Results are shown for increasing read-depth z-score parameter and increasing allelic imbalance evidence. Barplots show the fraction of samples with ctDNA detection and ctDNA dection and ctDNA level estimation for expected ctDNA levels up to 5%. Results are shown for increasing read-depth z-score parameter and increasing allelic imbalance evidence.

3.2.3 Clinical application of the panel in metastatic breast cancer serial samples

To assess the clinical applicability of eSENSES, we profiled 44 cfDNA samples from a cohort of 15 patients with metastatic breast cancer, collected in the context of a larger collaborative project named MIMESIS (40,95). For all but one of the patients, we had available plasma samples before the start of the therapy (T0), at the second cycle of therapy (T1) and at progression disease (T2). Two patients had ER+/HER2+ disease, 11 ER+/HER2-, one ER-/HER+ and another one ER-/HER2-.

Sequencing was performed resembling the characteristics of our control samples. In particular, in the patients' cohort, we consistently observed reasonable read duplication rates

(on average 20.3%), reasonable off-target reads (on average 19.1%) and an average sample mean depth of coverage of approximately 800x (**Figure 3.7**).

For 14 samples across 8 patients, eSENSES was unable to detect any tumor signal. Among the remaining 30 samples with detectable ctDNA, 24 had ctDNA estimations available, with values ranging from 5% to 60% (mean 20.4%) (**Figure 3.8A** and Supplementary Data 4). These values align with the aggressive features of the analyzed patients, showing no significant distribution differences between samples from ER+/HER2+ and ER+/HER- patients (P = 0.18). Although not statistically significant, we observed a general decrease in median ctDNA levels from 21% at T0 to 14% at T1. In addition, an increased number of involved sites was observed in patients with detectable ctDNA at T0 compared to those with non-detectable ctDNA (Supplementary Data 5).

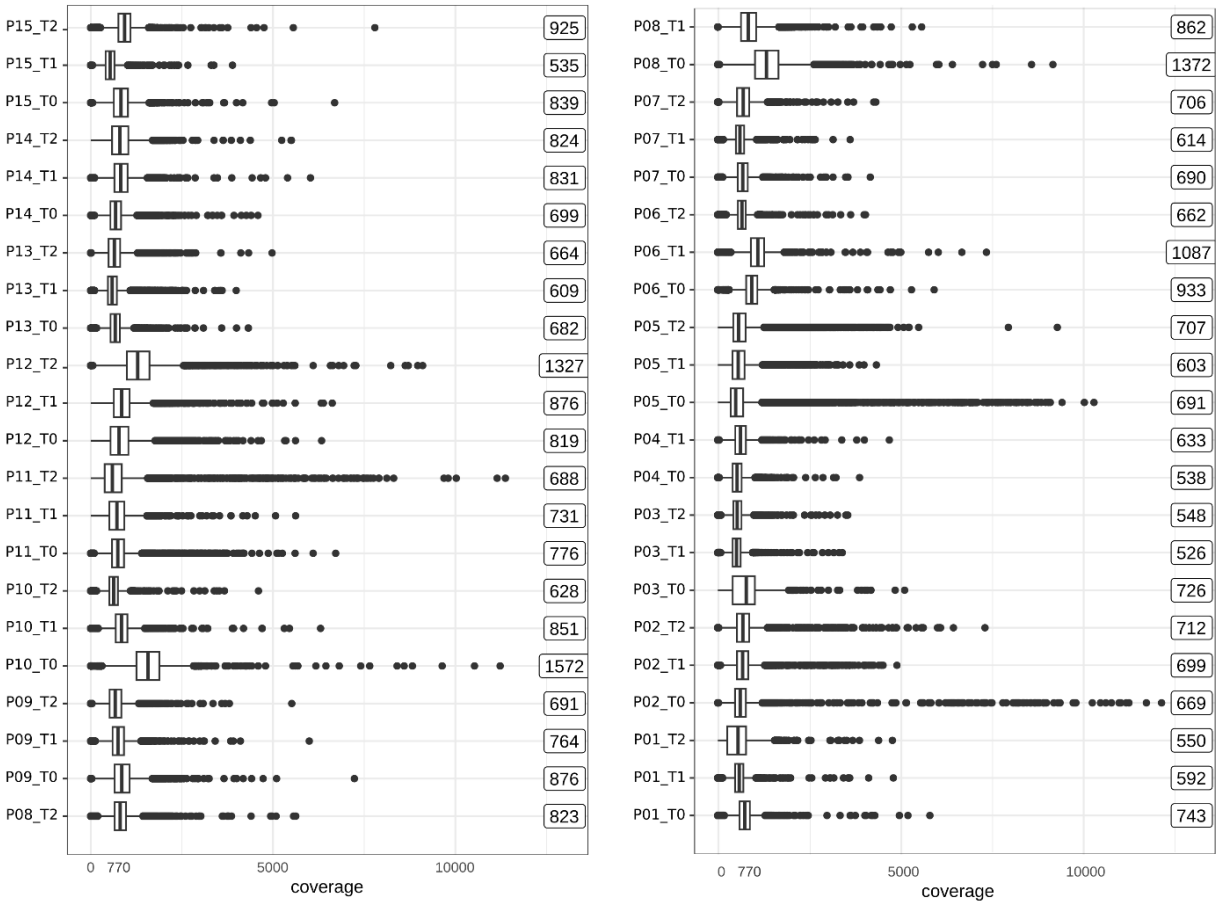

Figure 3.7 Processed patients' cfDNA samples' distributions of read depth coverage.

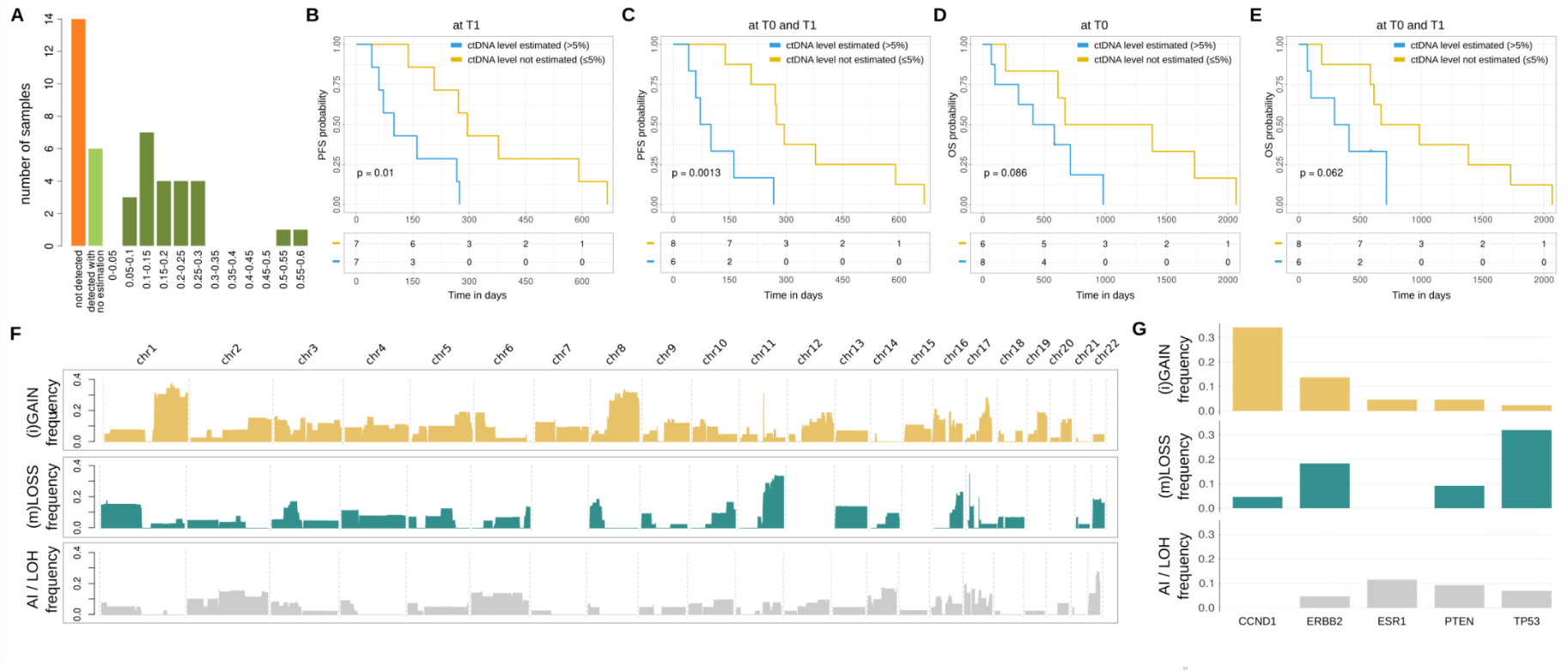
A
number of samples
not detected
detected with no estimation
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
0.2-0.25
0.25-0.3
0.3-0.35
0.35-0.4
0.4-0.45
0.45-0.5
0.5-0.55
0.55-0.6
B
at T1
ctDNA level estimated (>5%)
ctDNA level not estimated (≤5%)
PFS probability
p = 0.01
Time in days
C
at T0 and T1
p = 0.0013
D
at T0
OS probability
p = 0.086
E
at T0 and T1
p = 0.062
F
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
(i)GAIN frequency
(m)LOSS frequency
AI / LOH frequency
G
CCND1
ERBB2
ESR1
PTEN
TP53

Figure 3.8 Profiling of the clinical cohort.

A) *Distribution of ctDNA detection and ctDNA level estimation across all cohort' samples.* ***B)*** *PFS results stratifying samples based on availability of ctDNA level estimation at T1.* ***C)*** *PFS results stratifying samples based on availability of ctDNA level estimation at T0 or T1.* ***D)*** *OS results stratifying samples based on availability of ctDNA level estimation at T0.* ***E)*** *OS results stratifying samples based on availability of ctDNA level estimation at T0 or T1.* ***F)*** *Landscape of genome-wide SCNAs detected across all cohort's samples.*
G) *Landscape of focal SCNAs detected across all cohort's samples.*

Despite the cohort being limited to 15 patients, which reduced its statistical power for clinical analyses, we found that ctDNA estimation served as a prognostic indicator. Specifically, patients with estimated ctDNA levels at T1 had inferior progression-free survival (PFS) (P = 0.01, **Figure 3.8B**). The prognostic signal was even stronger when considering available ctDNA level estimations at both T0 and T1 (P = 0.0013, **Figure 3.8C**). Additionally, while not statistically significant, patients with estimated ctDNA levels at T0 or both at T0 and T1 exhibited a clear trend toward inferior overall survival (OS) (**Figure 3.8D-E**).

Although assigning precise copy number status to genomic segments at low ctDNA levels is challenging, the somatic copy number aberration profiles (Supplementary Data 6) reconstructed by our approach were consistent with previously published data (**Figure 3.9**), showing a clear enrichment of large gains/amplifications in the 1q, 8q, 16p and 17q arms, and significant large deletions in the 8p, 11q, 16q, 17p, and 22q arms (**Figure 3.8F**). Additionally, focal somatic copy number aberration analysis detected, as expected, amplification signal for CCDN1 gene and deletion signal for TP53 gene large fractions of patients (about 35% and 30%, respectively) and a clear amplification signal for ERBB2 gene in both HER2+ patients for which we had detectable ctDNA (**Figure 3.8G** and **Figure 3.10**).

Regarding the analysis of single nucleotide variants, performed exploiting our tool ABEMUS (71) shown to be a valuable choice for high coverage sequencing setup with no healthy matched samples' cohort (96), along with a custom post-processing pipeline, we were able to detect 80 non-synonymous SNVs across 32 genomic positions, with an average of 6 SNVs (from 1 to 20) per patient (across T0, T1 and T2) and an average of 3 SNVs (from 1 to 7) per sample. Median observed allelic fraction of non-synonymous SNVs was 3.7% (from 0.2% to 56%). The complete list of the non-synonymous SNVs detected in all samples is reported in Supplementary Data 7. The most frequently mutated genes at T0 or T2 were ESR1 (5/15, 33%), PIK3CA (5/15, 33%) and TP53 (4/15, 27%), aligning with published large genomic (Supplementary Data 8). ESR1 mutations were often observed at T0 but not at T2 (**Figure 3.11**) and were primarily subclonal (**Figure 3.12**), whereas PIK3CA mutations were predominantly

clonal and detected at both T0 and T2. In some samples where ctDNA was undetectable via SCNA, we identified damaging non-synonymous SNVs in key driver genes, aiding in the interpretation of circulating tumor signals. For example, although patient P03 had a positive ctDNA estimation only for sample T1 (at 7.5%), the PIK3CA p.E545K hotspot gain-of-function mutation was detected as heterozygous mutation in all three samples, with allelic fractions of 2.3%, 4.6%, and 1.6% in samples T0, T1, and T2, respectively.

Overall, the dynamics of arm-level SCNAs, along with focal SCNAs and non-synonymous SNVs in breast cancer driver genes across longitudinal samples, revealed a high genomic similarity of tumor clones between T0 and T2 for the majority of patients (**Figure 3.13A**). In particular, for matched samples with ctDNA estimations above 10%, these similarities can be further explored using a detailed allele-specific SCNA analysis, which leverages the allelic imbalance information available from our approach. For example, the analysis of patient P05 samples shows genomic profiles that are identical across all longitudinal samples, with mono-allelic loss of PTEN and TP53 genes, together with broad mono-allelic losses in p-arms of chromosomes 1 and 8 and q-arms of chromosomes 11 and 22, gain of q-arm of chromosome 1, loss of heterozygosity (LOH) of BRCA1 gene and amplifications of ERBB2 and CCND1 genes (**Figure 3.13B**). Another example is depicted in **Figure 3.13C**, where the analysis of patient P12's samples consistently shows mono-allelic losses in chr8.p and chr11.q, as well as gains in chr1.q and chr8.q, with unaltered ctDNA levels across all time points.

Overall, eSENSES demonstrated prognostic ctDNA level estimations and the capacity to offer sensitive genomic profiling across various abstraction levels. It fully leveraged allelic-imbalance information, while also illustrating the specific temporal evolution of each patient's cancer.

Figure 3.9 Arm-level SCNA frequencies retrieved across multiple large scale genomic studies obtained using BroadBand (bcglab.cibio.unitn.it/broadband).

BroadBand is a web-based resource we developed for exploring genomic data selected from a comprehensive array of breast cancer studies containing information on SNVs and SCNAs.

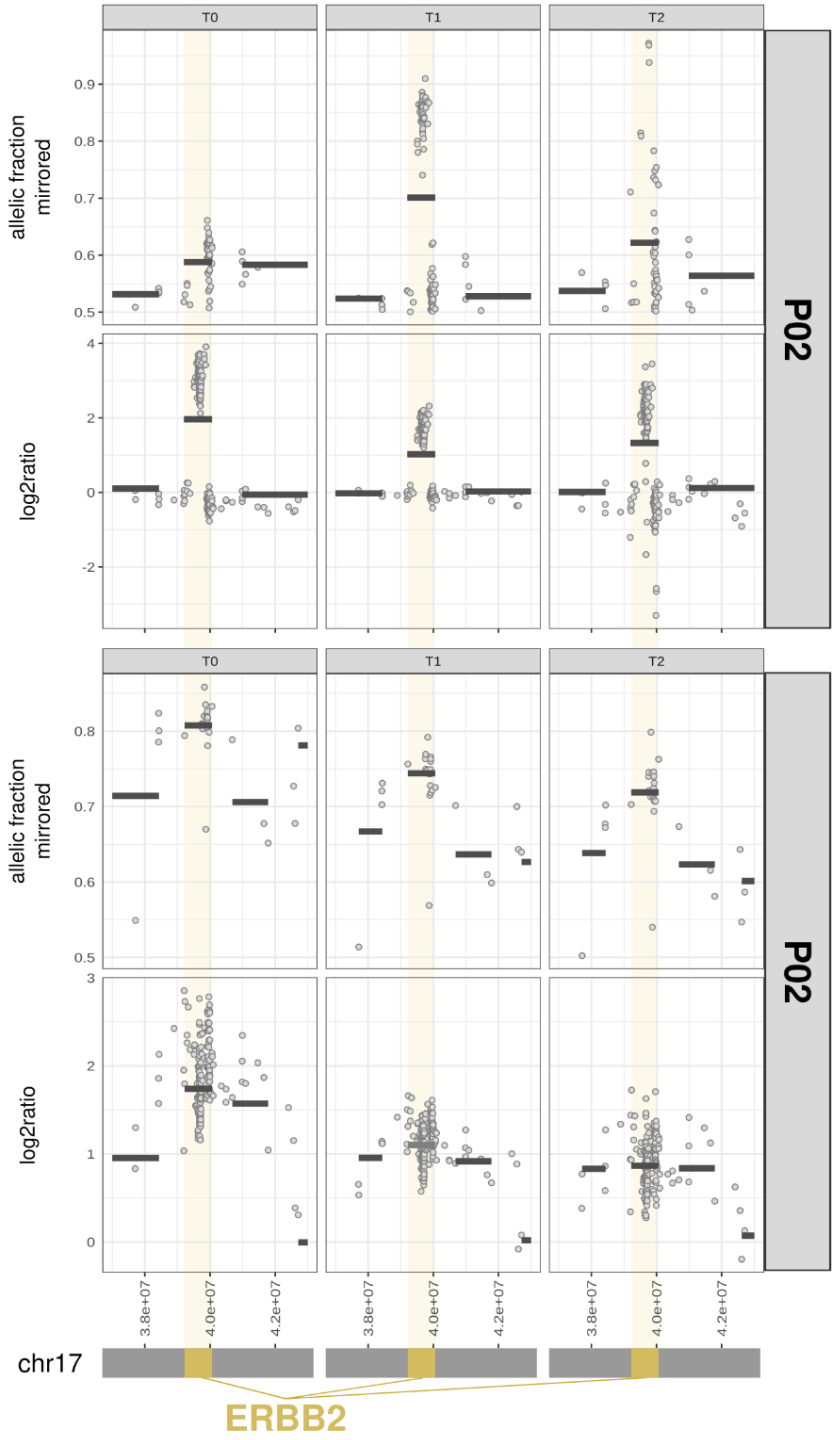

Figure 3.10 Log2 ratio and mirrored allelic fraction signals at each time point of the two HER2+ patients showing detected focal somatic copy number segments corresponding to the genomic region of ERBB2 gene.

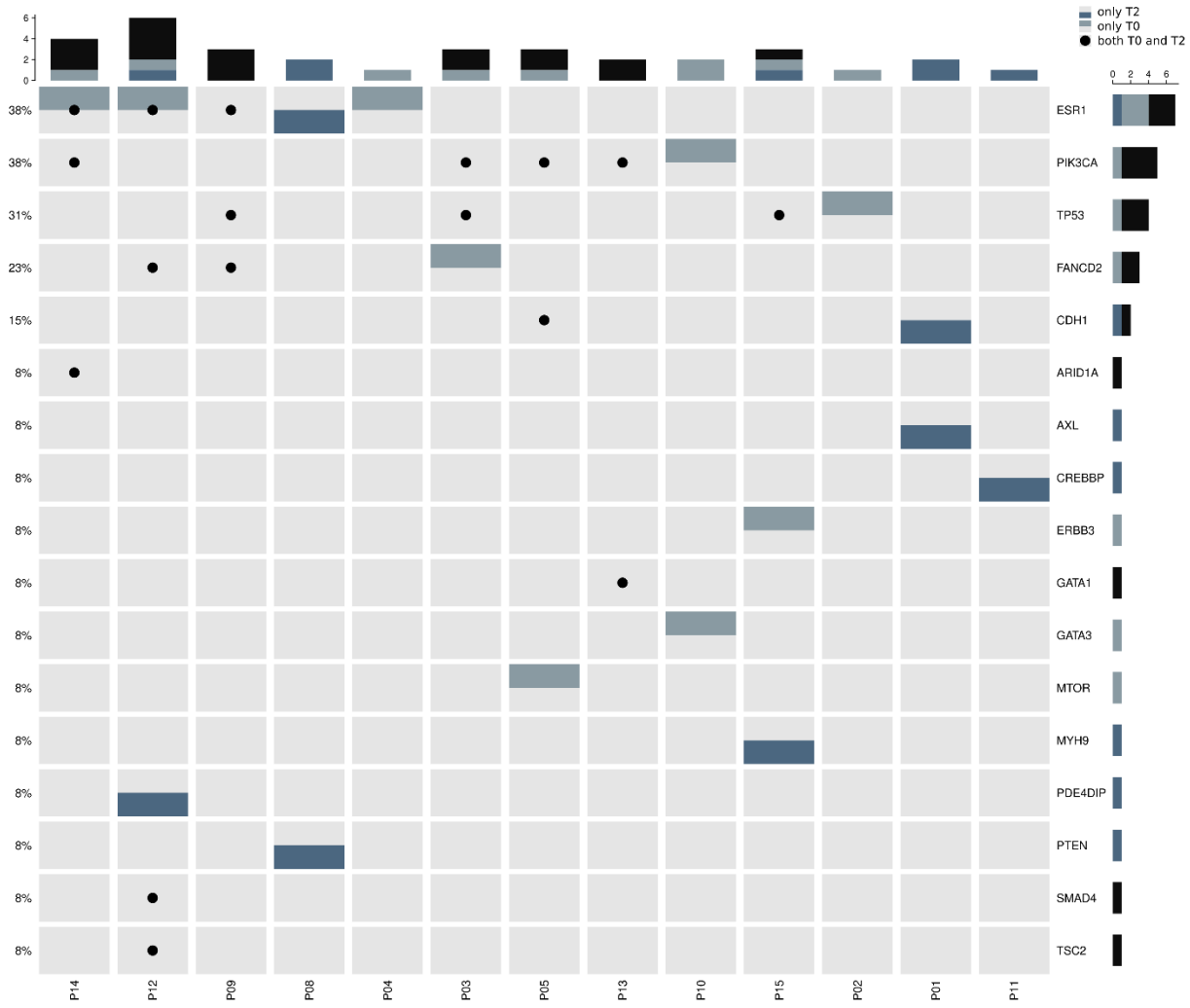

Figure 3.11 Oncoplot of non-synonymous SNV detected across patients' cfDNA samples at T0 and T2.

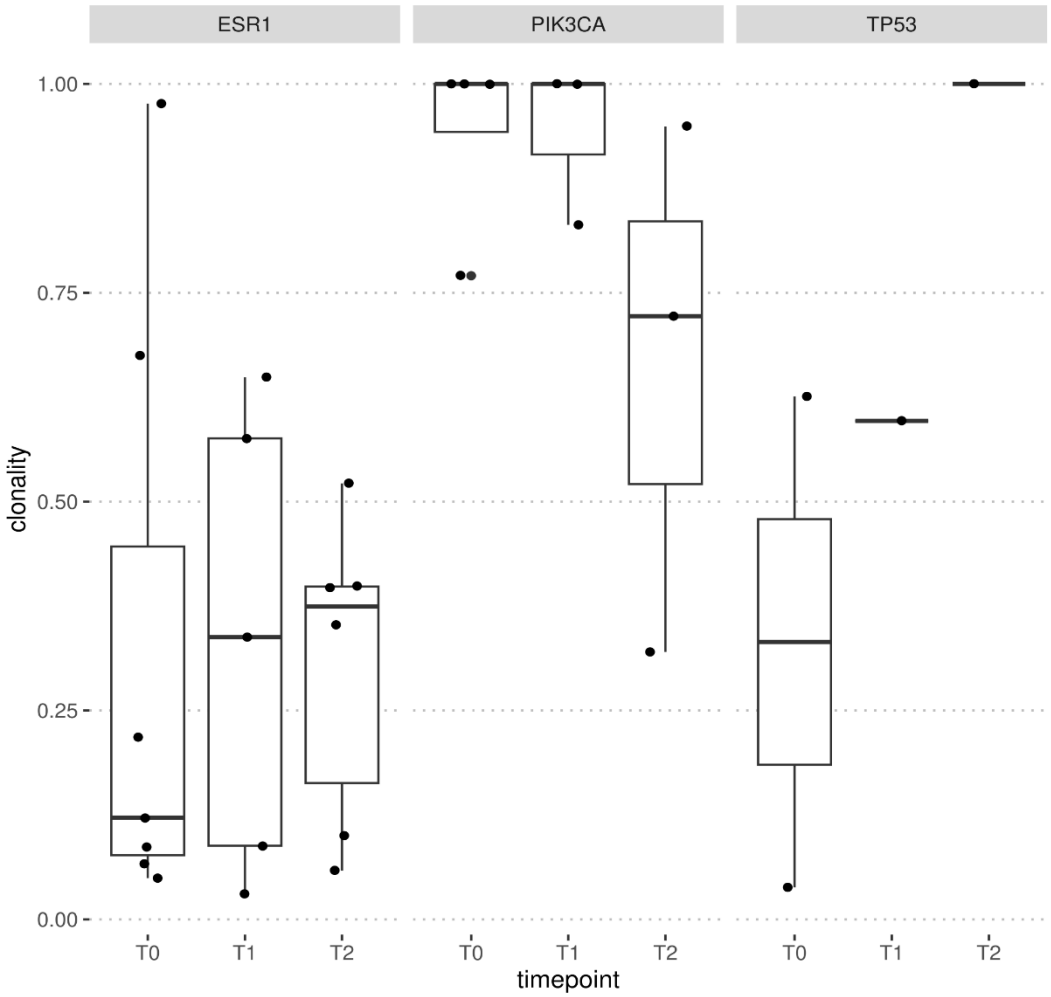

Figure 3.12 Clonality of damaging/deleterious SNVs across the most frequently mutated genes. Values above 0.75 are considered associated with clonal SNVs.

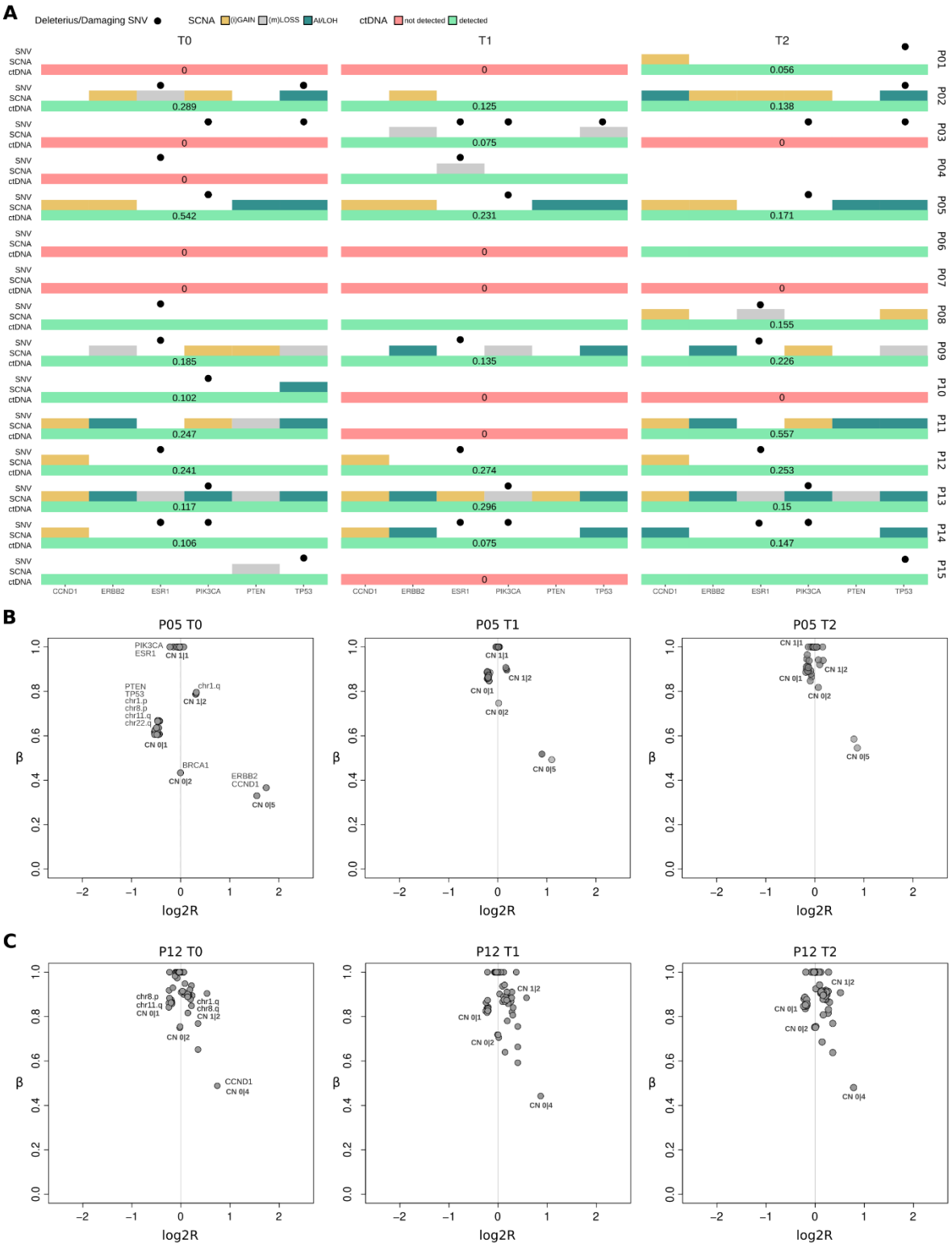

Figure 3.13 Longitudinal analysis of the clinical cohort.

A) *Distribution of SCNAs, SNVs and ctDNA detection/estimation for breast cancer driver genes across all patients at the different time points.* ***B)*** *Allele-specific analysis of SCNAs across patient's P05 T0, T1 and T2 samples.* ***C)*** *Allele-specific analysis of SCNAs across patient's P12 T0, T1 and T2 samples. Points indicate SCNA segments in the sample and labels indicate genomic position and (in bold) the allele-specific copy number. Points indicate SCNA segments in the sample and labels indicate genomic position and (in bold) the allele-specific copy number.*

3.2.4 Validation of eSENSES performances using independent and complementary approaches

To further validate our approach, we exploited additional profiling data available for our cohort's patients across the same time points in the context of the collaborative MIMESIS project (40). More specifically, the same cfDNA samples were also characterized generating low-pass methylation whole-genome bisulfite sequencing (lpWGBS) and analyzed with ichorCNA (87,97), a state-of-the-art tool for cfDNA WGS data analysis. Given our findings in (98), which showed that bisulfite pretreatment for WGS does not introduce significant biases in SCNA characterization or ctDNA detection, we were able to directly compare the performance of our approach with a complementary methylation-based assay.

IchorCNA was executed on lpWGBS data using two different configurations suggested by the authors: a default run with parameters optimized for estimating ctDNA levels in the context of moderate to high expected ctDNA, and a sensitive run designed for estimating ctDNA levels when low expected ctDNA is anticipated.

A high concordance was observed when comparing ctDNA level estimations obtained using the two complementary strategies. Specifically, there was an 80% correlation with the default run, which had a clear detection limit of 10% ctDNA (**Figure 3.14**). This correlation increased to 93% with the sensitive run (**Figure 3.15A**). In this mode, ichorCNA did not provide any negative calls, detecting positive ctDNA levels below 3% in 16 samples, which is the theoretical limit of detection declared by the authors of ichorCNA. Interestingly, considering that the majority of these 16 calls are likely to be false positives, for 11 of them (70%) our approach detected no tumor signal and for the 5 remaining ones our approach was able to detect the presence of ctDNA without however providing a ctDNA level estimation.

To further validate ctDNA of eSENSES and considering that ichorCNA was developed for ultra-low-pass whole-genome sequencing (ulpWGS) data, we utilized off-target reads from our cfDNA samples to achieve an average whole-genome depth of coverage of around 0.2x per sample, thereby emulating ulpWGS data.On these samples, we first ran ichorCNA in *sensitive* mode, focusing on patients' data and using control samples to create an ichorCNA panel of normals (PoN). As shown in **Figure 3.15B**, ctDNA level estimation on patients' cfDNA data exhibited strong concordance between our approach and ichorCNA, which was run on our panel of off-target reads. This concordance closely resembled the results obtained with

lpWGBS data. Also in this case, ichorCNA did not provide any negative calls, detecting positive ctDNA levels below 3% in 11 samples. Among these, 9 had no call by our approach, and the remaining 2 had a positive call with no estimation available, overall suggesting that our approach may offer improved precision in detecting ctDNA signals below the 3% threshold.

To support this observation and further validate the specificity of our approach, we finally conducted a leave-one-out (LOO) cross-validation analysis. For each control sample, we tested the eSENSES estimation using the remaining control samples to construct the panel of controls. Notably, the LOO analysis yielded negative ctDNA results for all control samples, confirming the high specificity of our method.

Overall, these results clearly demonstrate that eSENSES matches the sensitivity of state-of-the-art approaches while maintaining high specificity, thereby preventing the generation of false positive ctDNA calls.

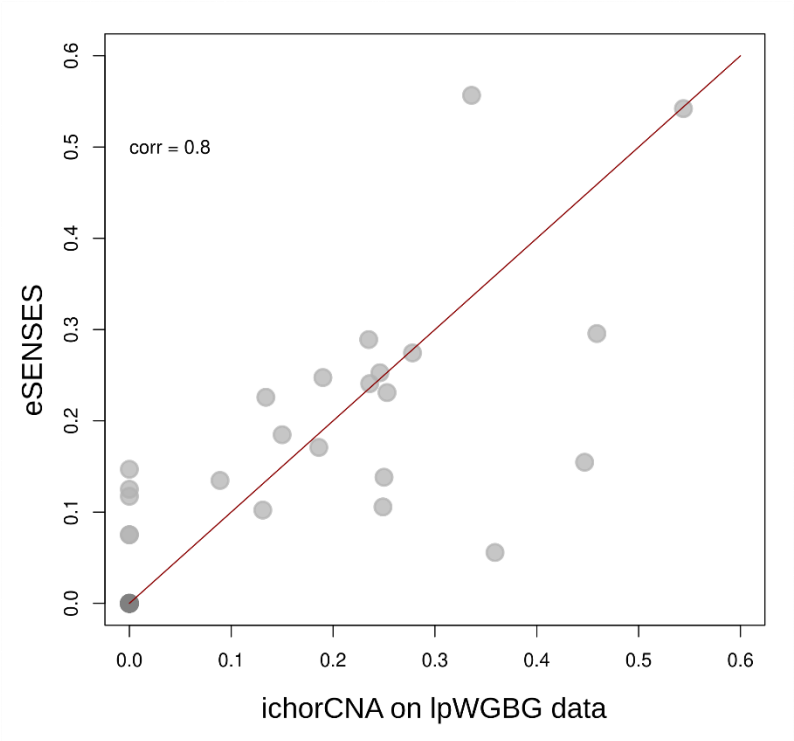

Figure 3.14 Concordance of eSENSES ctDNA level estimations with estimations obtained with ichorCNA (default parameters) from lpWGBS samples available for the same patients at the same time points

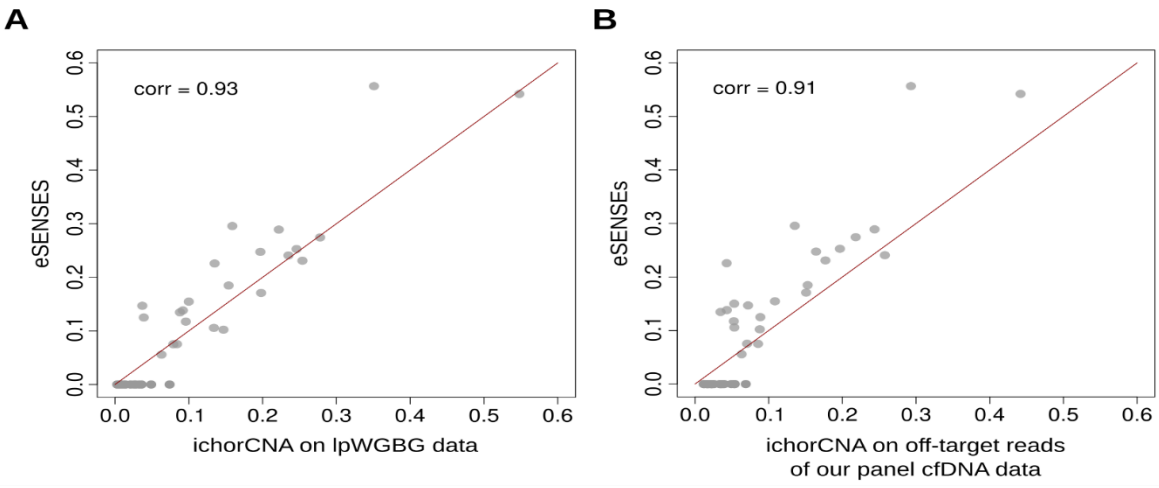

Figure 3.15 Validation analysis.
A) *Concordance of eSENSES ctDNA level estimations with estimations obtained with ichorCNA from lpWGBS samples available for the same patients at the same time points.* ***B)*** *Concordance of eSENSES ctDNA level estimations with estimations obtained from ulpWGS built from eSENSES off-target sequencing reads.*

3.3 Discussion

In this work, we introduced eSENSES, a custom targeted sequencing cfDNA panel combined with a tailored computational approach for a sensitive and specific detection and quantification of ctDNA and somatic copy number aberrations in patients with metastatic breast cancer.

By leveraging a comprehensive set of focal and genome-wide SNPs, eSENSES enhances the characterization of SCNAs and their allelic imbalances, providing an advanced framework for analyzing allele-specific SCNAs in low ctDNA breast cancer samples. Additionally, by including exonic regions of recurrently mutated breast cancer driver genes, eSENSES enables the detection and characterization of driver breast cancer SNVs, further improving the understanding and interpretation of complex allele-specific SCNA patterns.

Current NGS methods for estimating the tumor fraction in cfDNA based on SCNA profiles typically utilize either a horizontal approach, such as whole-genome (83,97) or whole-exome sequencing (87,88), or a vertical approach with targeted sequencing (89,90). While those approaches optimize either sequencing breadth or sequencing depth, eSENSES has the advantage of maximizing both, offering an innovative method that combines the strengths of horizontal and vertical sequencing. An approach in this direction was proposed for prostate cancer in (40), where only a thousand genome-wide SNPs and the exonic regions of only 13 genes were captured. Although the authors reported a ctDNA detection limit of 5%, which is one-fold higher than our in-silico estimate for eSENSES, they did not thoroughly explore the sensitivity and specificity of their detection limit.

To enhance the sensitivity of SCNA detection, other methods fully exploiting SNP allelic imbalance have been proposed (60,99). Specifically, in (99), an approach was developed to improve SCNA detection in pancreatic cancer cfDNA samples with low ctDNA. Although detection at ctDNA levels <5% was shown to be feasible, the sensitivity remained limited, only large-scale SCNAs were potentially detectable, and deep WES (>200x average depth of coverage) was required to achieve the enhanced performance levels.

Another advantage of eSENSES is that it does not require matched control samples. Instead, it utilizes a pooled set of control samples to build a reference model a single time, which is then applied by our computational approach to analyze all patients' cfDNA samples.

Our in-silico benchmarking analysis demonstrated that eSENSES, when applied to samples with an average depth of coverage of 800X—resulting in a sequencing cost comparable to that of low-pass whole genome sequencing (approximately 30-40 million reads)—achieves high sensitivity and specificity in detecting and estimating circulating tumor DNA (ctDNA) levels. This performance is particularly notable in samples with low ctDNA content. Our approach demonstrated the potential to achieve a detection sensitivity of 90% for ctDNA levels as low as 3% with zero FPR and a detection limit at 0.5% ctDNA. Increasing the depth of coverage to approximately 2500x allowed us to maintain a detection sensitivity of over 80% for ctDNA levels as low as 2%, while still ensuring a zero FPR.

The effectiveness of eSENSES was demonstrated on a small cohort of patients with metastatic breast cancer. In particular, the analysis of 44 cfDNA samples from 15 patients demonstrated that ctDNA levels were consistent with aggressive disease features and were associated with both overall survival and progression-free survival. Although our cohort predominantly included ER+ breast cancer patients, with approximately 60% evaluated at the onset of metastatic treatment, the distribution of ctDNA estimates closely mirrored those reported in ref. (100) . This study, focusing on metastatic TNBC patients, found that 70% of analyzed samples had confidently detectable ctDNA, with 63% of patients presenting at least one sample where ctDNA levels exceeded 10%. Notably, this threshold, which aligns with the one used in our study, was also deemed by the authors as adequate for high-confidence SCNA characterization.

The SCNA profiles of the cohort were consistent with established breast cancer profiles described by large genomic studies (5,31,32,49,91–93), identifying significant large-scale and focal amplifications and deletions in key genes and genomic regions. Of note, eSENSES was able to precisely delineate the amplification status of ERBB2 gene's genomic region in all HER2+ patients and to provide a confident allele-specific characterization of all detected SCNAs when ctDNA levels exceeded 10%. This capability facilitated the precise characterization of intricate SCNA patterns and LOH events.

Furthermore, the SNV analysis accurately identified non-synonymous mutations in well-known breast cancer driver genes like ESR1, PIK3CA, and TP53. The patterns in SNV allelic fractions also contributed to the analysis of longitudinal samples from patients, revealing, in conjunction with SCNA profiles, a prevalent high genomic similarity among tumor clones circulating across various disease time points within our cohort.

To further validate eSENSES's performance, we demonstrated that ctDNA detection and level estimations from our cohort samples closely matched those provided by the state-of-the-art tool ichorCNA. We applied ichorCNA to low-pass WGBS data from the same patients and utilized off-target reads from our targeted approach to simulate low-pass WGS samples. Importantly, eSENSES exhibited comparable ctDNA detection sensitivity to ichorCNA while maintaining high specificity.

To conclude, we introduced eSENSES, an approach that exploits a combination of focal and genome-wide SNPs with a tailored computational approach, providing a sensitive and specific cost-effective technology for detecting ctDNA and characterizing SCNAs in breast cancer patients. The eSENSES panel demonstrates significant clinical utility in monitoring disease progression and offers a reliable alternative to existing methods, maintaining high specificity and sensitivity.

3.4 Materials and Methods

3.4.1 Samples and patients' characteristics

Plasma samples (n=44) were collected from 15 patients with metastatic breast cancer. Clinical and pathological characteristics for each of the patients are reported in Supplementary Data 3. In particular, samples were drowned at baseline (T0, n=15), after one cycle of treatment (T1, n=15) and at disease progression (T2, n=14). Plasma samples were prepared within an hour from blood collection and stored at -80°C until cfDNA extraction. Normal K2EDTA plasma samples from 20 female healthy donors were purchased from Precision for Medicine (Precision for Medicine, MA).

3.4.2 Selection of SNPs to include in the panel

Two sets of high gMAF SNPs were included in the panel; a first set, including loci distributed uniformly along the human genome to ensure the detection of SCNAs involving extensive genomic regions, and a second set, consisting of SNPs located across genes frequently and focally altered by SCNAs.

For the first set, to ensure greater experimental efficiency, priority in selection was assigned to SNPs included in the Infinium Omni2.5-8 Kit microarray and characterized by a MAF ≥ 0.45,

as reported in the dbSNP version v151 catalog (54) considering all populations included in the 1,000 Genomes Project (101). More specifically, for each chromosomal band, a number of SNPs was selected to ensure a total density of 5 SNPs/Mb, a minimum number of 20 loci for the least populated chromosomal arm (chr21p) and maintaining a spacing of at least 200 bases between two consecutive SNPs. A total of 14,170 loci were selected, 5% of which located in exonic regions.

For the second set, SNPs with gMAF ≥ 0.35 were selected in genomic windows of approximately 500 kbp both upstream and downstream of 5 selected genes (CCND1, ERBB2, PTEN, TP53, ESR1) that are frequently affected by focal SCNAs. In total, 856 loci were selected, with an average of 171 SNPs per gene.

3.4.3 Selection of genes to include in the panel for SNV analysis

Publicly available data from the cBioPortal for Cancer Genomics (102,103) and the Integrative Onco Genomics (IntOgen) database (104) were utilized to define a list of genes of interest in the clinical context of breast carcinoma. Specifically, mutational data from the METABRIC studies (32,49,91), TCGA PanCancer Atlas (www.cancer.gov/tcga) INSERM (31), MSK-IMPACT (5), and The Metastatic Breast Cancer Project (www.mbcproject.org) were analyzed to associate a mutation frequency with each gene. A gene was considered mutated if at least one of its coding regions was found altered by the presence of at least one somatic point mutation causing a change of a base to a different and "non-synonymous" one.

To ensure the selection is as accurate and informative as possible, the data and their mutational frequencies were stratified based on the molecular subtype of the tumor (HER2+, HR+, Triple Negative) and its stage of evolution (primary or metastatic). This stratification allowed the selection of the most frequently mutated genes (N = 25) for each of the 6 classes. The unique set of genes identified from this initial selection (N = 68) was validated and further extended by analyzing tumor-specific aberration frequencies reported in the IntOgen database. The combined results of these two analyses identified 81 genes of interest, whose coding regions (exons, excluding UTRs) were selected for sequencing. The final selection included 2,119 exons of 81 genes.

3.4.4 Panel design

All selected SNPs and coding regions of genes of interest were included in a single BED file. This file served as input for the Illumina HyperDesign software (hyperdesign.com), with stringency parameters (a value defining the specificity of a probe for the target region; min=1, max=20, optimal interval for unique probes from 1 to 4) and overhang (a value defining the number of bases by which the probe deviates from the target region; min=0, max=120, optimal interval from 0 to 30) set to 4 and 30, respectively. The output produced by HyperDesign defined the optimal genomic coordinates for 17,040 sequencing probes, covering a total of approximately 2.3 Mbp.

3.4.5 cfDNA extraction, library preparation and sequencing

Plasma cfDNA was isolated by QIAmp Circulating Nucleic Acid Kit (Qiagen) and quantified by Qubit (Qubit dsDNA High Sensitivity Kit, Life Technologies). Libraries were prepared starting from 20 ng of DNA using KAPA HyperPrep Kit and Kapa Universal UMI adapter (Roche), according to the manufacturer's instructions. Libraries size distribution and concentration were analyzed respectively by Bioanalyzer DNA High Sensitivity Kit (Agilent) and Qubit (Qubit dsDNA High Sensitivity Kit, Life Technologies) before target enrichment with the 17,040 customized Kapa hyper Cap Target Enrichment Probes (Roche). For the enrichment step four samples were combined using 500 ng of each amplified indexed library. The enriched DNA samples were amplified according to the manufacturer's instructions with 16 cycles of Post-Capture PCR. The amplified enriched DNA samples libraries were quantified by Qubit dsDNA High Sensitivity Kit and libraries size distribution were evaluated by Bioanalyzer DNA High Sensitivity Kit. Sequencing was performed on Illumina Novaseq 6000 generating 150bp paired-end reads.

3.4.6 Sequencing data pre-processing

Reads were trimmed to remove adapters and UMI extraction, consensus and sorting were performed using UMI_tools (version 1.1.2) (105), fgbio (version 2.0.2), and GATK (version 4.2) (106). Alignment to the human GRCh38 reference genome was performed using BWA-MEM. Realignment and recalibration were performed using GATK. MD tags were calculated using

samtools calmd (version 1.7) (107) and overlapping read pairs were clipped using bamUtil (1.0.14). PaCBAM was used to generate pileup and depth of coverage statistics' files (70). In particular, we used rc files, reporting the average depth of coverage of all captured genomic regions along with their GC content, and snps files, reporting pileup statistics for all considered SNPs such as total coverage, coverage supporting the reference allele, coverage supporting alternative allele, and the variant allelic fraction (AF).

3.4.7 Reference Mapping Bias correction

Reference Mapping Bias (RMB), namely the presence of a main AF peak value different from the expected 0.5, is addressed to ensure proper downstream analysis of SNPs' AF data and comparison of AF distributions from independent samples, similarly to what previously described in (108). The peak correction was applied separately for all cfDNA samples, applying a Kernel Density Estimation on the heterozygous SNPs AF distribution and extracting peaks by computing the local maxima of the smoothed distribution (60); the central peak was extracted and data centered to the 0.5 theoretical value.

3.4.8 Panel of controls

Sequencing data of control samples was used to generate a panel of control, which characterizes the captured genomic regions of all genes of interest and all SNPs.

First, read depth of coverage of all captured regions across all control samples was normalized for both GC content and sample's mean coverage. Normalized read depth (RD) was then used to compute, for each region across all control samples, mean and standard deviation values. Then, we performed a leave-one-out procedure in which, for each control sample i and each captured region c, a log2 ratio (log2R) value was calculated using the following formula:

$$log2R_{ci} = log2R\left(\frac{cov_{ci}}{mean(cov_{cn-i})}\right)$$

with cov_{ci} being the RD of the control sample i in region c divided by the mean of the RD of region c across all control samples excluding sample i (cov_{cn-i}). In this way a table of log2R

values for each captured region across each control sample was obtained. While performing this operation, a mask ratio parameter (set at 0.5) was used. This value determines the minimum percentage of samples having finite values for a considered captured region in order to perform the log2R calculation. If the mask ratio was not reached that region was not included in the panel of controls. Overall, read depth of coverage statistics are organized two tables: the first, referred to as rc table, having the RD mean and standard deviation for each captured region across all controls; the second table, referred to as log2R table, having for each captured region all the controls' log2R values.

Then, for SNPs' AF data two main data structures were built: (1) a collection of SNPs summary statistics; (2) AF dispersion stratified by changes in SNPs local coverages. Briefly, for each control sample, captured SNPs that have a heterozygous genotype ($0.2 < AF < 0.8$) were kept. Summary statistics for each heterozygous SNP across all control samples were computed, including, AF distribution mean, coefficient of variation, proportion of samples out of N harboring the heterozygous genotype, and mean coverage. AF dispersion was instead modeled collecting AF standard deviations stratified by local coverage quantiles Q (min 0%, max 100%, step 10%).

3.4.9 SCNA identification

For a given cfDNA plasma sample, raw read depth data was processed as mentioned above. Upon calculation of log2R for all captured regions, Circular Binary Segmentation (CBS) (109) was performed, using the R package DNAcopy (version 1.68) (110), to identify putative read depth distribution change points representing copy number variations. The segmentation analysis was performed considering single arms of each chromosome separately. Focal genes' regions, enriched for high MAF SNPs in our panel, were smoothed in order to prevent over-segmentation, in other words, segmentation was not performed on these genes and their SNPs enriched regions were considered in their entirety for gaining a better signal.

Then, for each identified segment, a second run of CBS was performed. This time upon mirrored AF values (mAF), calculated as:

$$mAF = \begin{cases} 1 - SNP_{af}, & SNP_{af} > 0.5 \\ SNP_{af}, & otherwise \end{cases}$$

mAF values were used to better identify possible SCNA events not detectable by the log2R signal. As in this case, focal regions were analyzed in their entirety without segmentation.

3.4.10 Computation of allelic imbalance per segmented region

For each of the identified segments, an allelic imbalance value was computed using a methodology that extends our work in (60,111). In detail, given a cfDNA sample and an identified SCNA segment S, the set of SNPs that are heterozygous in the cfDNA sample, contained in the segment and also present in the panel of controls were selected and used to compute a value representing the evidence of allelic imbalance for the segment S and another value representing an estimate of the β_S value, which represents the proportion of local read depth signal that is not imputable to ctDNA (112). The evidence of allelic imbalance was computed with the formula:

$$E(S) = \frac{\sum_1^k W(d > D)}{k}$$

were k = 5 (by default), d is the observed mAF distribution in the cfDNA sample, D is a simulated mAF distribution generated sampling one time for each SNP in S from a normal distribution with mean and standard deviation obtained from the panel of controls, and W is a function returning 1 if the difference between d_T and D applying a Wilcoxon signed-rank test with significance cutoff of 1% is statistically significant, 0 otherwise.

The β_S estimate was instead computed by comparing d with simulated distributions mimicking different levels of β (representing different tumor content levels) and searching for the most similar one. Formally:

$$\beta_S = min\{\beta \vee W(d > D_\beta)\} - (min\{\beta \vee W(d > D_\beta)\} - max\{\beta \vee W(d < D_\beta)\}) * P$$

with

$$P = \frac{median(d - min(d))}{max(d) - min(d)} and\ \beta \in \{0.01, 0.02, \dots, 0.99, 1\}$$

and where $W(d > D_\beta)$ is the Wilcoxon signed-rank statistics (significance cutoff of 1%) comparing d and D_β.

3.4.11 Assignment of copy number state

To assign a copy number state to each SCNA segment identified in a cfDNA sample, a computational approach combining allelic imbalance evidence and read-depth z-score thresholds is used. In detail, given a SCNA segment S identified in a cfDNA sample, the panel of controls log2R table was queried to obtain the set of captured genomic regions that are contained in the segment, denoted as R_S. Then, for each control sample, the median log2R of all R_S genomic regions was computed resulting in a vector of reference log2R values for the segment S, denoted as $log2R_S$. The z-score for the segment S was then calculated as follows:

$$zscore_S = \frac{(log2R_S - mean(log2R_S))}{std(log2R_S)}$$

A statistical read-depth z-score threshold Z_{thr} (by default 2.58) was finally combined with an allelic imbalance evidence threshold E_{thr} (by default 0.8) to assign a copy number state to each segment S in the following way:

$$SCNA_S = n\begin{cases} imbalanced\ GAIN\ (iGAIN), & zscore_S > Z_{thr} \wedge E(S) > E_{thr} \\ monoallelic\ LOSS\ (mLOSS), & zscore_S < -Z_{thr} \wedge E(S) > E_{thr} \\ allelic\ imbalance\ (AI), & zscore_S \in [-Z_{thr}, Z_{thr}] \wedge E(S) > E_{thr} \\ GAIN, & zscore_S > Z_{thr} \wedge E(S) \leq E_{thr} \\ LOSS, & zscore_S < -Z_{thr} \wedge E(S) \leq E_{thr} \end{cases}$$

The state AI is used to identify all segments that have evidence of allelic imbalance but for which there is no statistical evidence of SCNA from the read depth analysis. Of note, when the tumor content estimation (see below) for the sample is above 15% we can confidently assume that AI events are LOH events.

3.4.12 Tumor content and ploidy estimation

Tumor content detection in cfDNA plasma samples was assessed considering, for each sample, the presence of at least one segment having an allelic imbalance evidence value greater than E_{thr}. Tumor content for a cfDNA sample was instead estimated considering the β values of all SCNA segments identified as $mLOSS$ (112), which represent the set of mono-allelic deletions identified in the cfDNA sample. In detail, an integrated β value was calculated as a weighted mean of all peaks identified in the β segment values distribution (β was weighted by considering the magnitude of the peak). Then, as described in (112), the tumor content was estimated as:

$$TC = 1 - \frac{\beta}{(2 - \beta)}$$

For each cfDNA sample having evidence of tumor signal, a ploidy estimation was computed. In detail, samples segments were filtered for those with β equal to 1, indicating no presence of allelic imbalance. Clustering using dbscan (113) of log2R values for these segments was then performed and the left most cluster was identified as the one representing the balanced copy number 2 (one copy per allele) and used as shift value to adjust the overall sample' log2R distribution. More specifically, the adjustment was applied to all segments in order to center putative copy number neutral segments to zero:

$$log2r.corrected = log2r - shift$$

Of note, since tumor content is typically low in cfDNA samples and ploidy values are extremely challenging to calculate at low tumor content level, the ploidy adjustment was applied only when the tumor content estimation was greater than 15%. When ploidy adjustment was applied, a new tumor content estimation was calculated after the adjustment.

3.4.13 In-silico benchmarking

To assess the theoretical detection/estimation limits of our approach, an in-silico cohort of synthetic samples was generated. To this end, synggen, a computational tool we recently

developed for the fast generation of large-scale realistic and heterogeneous cancer sequencing synthetic datasets, was used (94). To generate the in-silico cohort, profiles of germline SNPs and somatic allele-specific SCNA were collected and retrieved, respectively, from CEU individuals in the 1,000 Genomes Project collection and from the TCGA dataset (cbioportal.org).

In detail, sequencing data (BAM files) of the control samples from the available cohort were provided in input to synggen using a specific execution mode that, from those files, extracts a series of statistical models that summarize platform specific data characteristics, such as the distribution of the read depth of coverage, the distributions of read and base qualities, and base-specific systematic errors. These models were then used, in conjunction with the collected SNPs and allele-specific SCNA profiles, to generate synthetic control and cfDNA samples at different levels of tumor content and average depth of coverage.

More precisely, for a read depth of coverage of ~800x (representing the average coverage of the data we generated), 20 representative breast cancer samples for each of the following decreasing tumor content were generated [80%, 60%, 40%, 20%, 10%, 7.5%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.1%]. Then, for increasing read depth of coverage scenarios [1000x, 1500x, 2000x, 2500x], 20 representative breast cancer samples for each of the following decreasing tumor content were generated [5%, 4%, 3%, 2%, 1%, 0.5%, 0.1%]. In addition, 40 control samples were generated for each simulated depth of coverage scenario, half representing cfDNA samples with no tumor signal and half to generate the panels of controls. All allele-specific SCNA were included in the synthetic cfDNA data as clonal events.

The simulated cohort was used to assess the performance of our assay. In particular, we looked into the ability of our computational approach to identify presence of tumor signal in a synthetic cfDNA sample, namely a detection performance, and, if possible, into the estimation of the tumor content present in the sample analyzed. Detection performances were tested looking into the accuracy of identifying samples with positive tumor content at each different simulated condition (coverage / tumor content) independently (N=20).

3.4.14 SNV calls

To detect somatic single nucleotide variants (SNVs) we applied ABEMUS (71), a method we previously implemented and that is specifically designed for SNVs detection in cfDNA samples. ABEMUS was run using our control samples to create the ABEMUS reference model, which captures platform specific characteristics that are used by the tool to improve precision of SNVs calling. ABEMUS analyses were then performed using the standard computational workflow across all cfDNA samples considering only the exonic regions that are captured by the panel (i.e. we excluded all regions capturing assay SNPs). Considering that we had no matched control samples for our cfDNA samples, we then implemented an ad-hoc post-processing strategy to reduce the number of false positives. First, we annotated all identified position using the SNP Nexus web server. Exploiting the sequential samples, we then excluded all identified SNVs that were annotated in the dbSNP database, had a MAF>0.01 in gnomAD and had in 2/3 of the cfDNA sequential samples an AF>0.2. From the remaining calls, we then excluded the ones that were annotated in the dbSNP database, had a MAF>0 in gnomAD and had in all the sequential cfDNA samples an AF>0.4. Finally, we retained only the SNV calls supported by a number of alternative reads >1 and annotated for the presence in the COSMIC database (114) or in breast cancer datasets available from the cBioPortal. The clonality of single nucleotide variants (SNVs) was determined by calculating the ratio of the SNV's allele frequency to the sample’s ctDNA level. This calculation assumes the presence of a mono-allelic mutation and includes a correction for mLOSS. If the resulting ratio exceeds 1, it is normalized to 1. Values above 0.75 are considered associated with clonal SNVs.

Chapter 4 - Radiomics as a predictive biomarker for immunotherapy response in advanced NSCLC

This chapter investigates the potential of integrating radiomic features derived from computed tomography (CT) images with blood-based markers of systemic inflammation to predict treatment response in patients with advanced non-small cell lung cancer (NSCLC) receiving immune checkpoint inhibitors (ICIs). This innovative approach seeks to enhance the prediction of ICI response by combining the rich phenotypic information captured by radiomics with the systemic immune landscape reflected in blood-based markers. This research was conducted in collaboration with the Medical Oncology Unit of the University Hospital of Parma (Italy), which provided data regarding the patient cohort.

My central contribution to this research lies in the development and implementation of a comprehensive computational framework for analyzing and integrating radiomic and blood-based data. This involved the design and optimization of a feature selection pipeline to identify the most informative radiomic features, the training and evaluation of machine learning models for predicting treatment response, and the exploration of strategies for integrating these features with blood-based markers.

This chapter provides a detailed account of the computational methodology employed, encompassing feature selection techniques, machine learning models, and model evaluation strategies. Furthermore, it presents the results of applying this framework to a cohort of NSCLC patients undergoing ICI therapy, demonstrating the potential of this integrated approach for improving patient stratification and informing personalized treatment decisions.

4.1 Introduction

Non-small cell lung cancer (NSCLC) stands as a significant global health concern, representing a heterogeneous group of lung cancers with varying molecular profiles and clinical behaviors (115,116). Despite advances in treatment, including surgery, chemotherapy, and radiation therapy, the prognosis for advanced NSCLC remains poor, underscoring the urgent need for novel therapeutic strategies and predictive biomarkers (117,118). The emergence of immune

checkpoint inhibitors (ICIs) has brought renewed hope for patients with advanced NSCLC, offering the potential for durable responses and improved survival (119).

ICIs, a revolutionary class of immunotherapy drugs, work by blocking immune checkpoints, such as programmed cell death protein 1 (PD-1) or its ligand (PD-L1), which are often exploited by cancer cells to evade immune surveillance (120,121). By inhibiting these checkpoints, ICIs unleash the power of the immune system to recognize and destroy cancer cells. While ICIs have demonstrated remarkable efficacy in a subset of NSCLC patients, their response rates remain variable, highlighting the need for robust predictive biomarkers to identify patients most likely to benefit (122–124).

PD-1, a key immune checkpoint receptor expressed on T cells, plays a critical role in regulating immune responses. Its interaction with PD-L1, often overexpressed on tumor cells, suppresses T cell activation and promotes immune evasion. Blocking the PD-1/PD-L1 axis with ICIs reinvigorates anti-tumor immunity, leading to tumor regression and improved survival in some patients. However, the complex interplay of factors influencing ICI response, including tumor microenvironment characteristics, PD-L1 expression levels, and the presence of other immune escape mechanisms, necessitates a multifaceted approach to patient selection (125,126).

Radiomics, an emerging field in medical imaging analysis, leverages advanced computational techniques to extract a vast array of quantitative features from medical images, providing a comprehensive characterization of tumor phenotypes (62,127). These features, encompassing tumor shape, size, texture, and intensity, can capture subtle intratumoral heterogeneity and provide insights into the underlying biological processes driving tumor behavior. Radiomics has shown promise in predicting treatment response and prognosis in various cancers, including NSCLC, and may offer a non-invasive means of identifying patients who are most likely to benefit from ICI therapy (128–130).

Integrating radiomic features with blood-based markers of systemic inflammation, such as basal lactate dehydrogenase (LDH) and derived neutrophil-to-lymphocyte ratio (dNLR), represents a promising avenue for enhancing the prediction of ICI response in NSCLC (131–133). While radiomics provides a detailed portrait of the tumor microenvironment, blood-based markers offer a glimpse into the broader immune landscape and its interaction with the tumor. Combining these two non-invasive modalities may provide a more comprehensive and accurate assessment of tumor biology and immune response, leading to improved

patient stratification and personalized treatment decisions. This study aims to explore the synergistic potential of radiomics and blood-based markers in predicting ICI response, ultimately contributing to the advancement of precision oncology in NSCLC.

4.2 Results

4.2.1 Cohort dataset construction

This study included 117 stage IV non-small cell lung cancer (NSCLC) patients who received immune checkpoint inhibitors (ICIs) at the University Hospital of Parma between March 2017 and July 2019 (11% as first line, 89% as second or more line of treatment). Median overall survival (OS) and progression free survival (PFS) were 7.8 (95%CI, 5.4-12.8) and 2.5 months (95% CI, 1.1-3.9), respectively. According to RECIST criteria (134), 76 patients (65%) belonged to non-responders, while the remaining 41 (35%) were responders. All patients underwent evaluations including CT imaging, clinical laboratory investigations (ECOG performance status, clinical stage), and peripheral blood sampling for LDH and dNLR immunological markers.

CT examinations were acquired within one week of clinical laboratory assessments using various multidetector CT scanners. All acquisitions were performed with contrast medium administration, with slice thicknesses ranging from 1.5 to 2.5 mm, tube voltage ranging from 100 to 120 kVp, and smooth reconstruction kernels.

For radiomic analysis, a 5-year experienced thoracic radiologist segmented the primary tumor on post-contrast CT images using 3D Slicer software. A 3-year experienced thoracic radiologist segmented 25 randomly selected tumors to evaluate feature robustness. They then proceeded extracting radiomic features using pyRadiomics (62), obtaining 847 radiomic features.

4.2.2 Clinical variables performance baseline

To establish a baseline for predicting immunotherapy response, we assessed the performance of blood-derived pro-inflammatory markers, specifically basal lactate dehydrogenase (LDH) and derived neutrophil-to-lymphocyte ratio (dNLR). Several machine learning models were trained and evaluated using stratified Monte Carlo cross-validation. Logistic regression and

support vector classifier (SVC) models achieved an average area under the receiver operating characteristic curve (AUC) of 0.74, while the random forest classifier returned a slightly higher AUC of 0.78 (**Figure 4.1**). Interestingly, incorporating both immunological markers and clinical variables yielded similar predictive performance, whereas models based solely on clinical variables demonstrated lower accuracy (**Figure 4.1**). These findings support LDH and dNLR clinical potential as valuable predictors of immunotherapy response (131–133), potentially capturing distinct biological information not fully reflected in standard clinical variables.

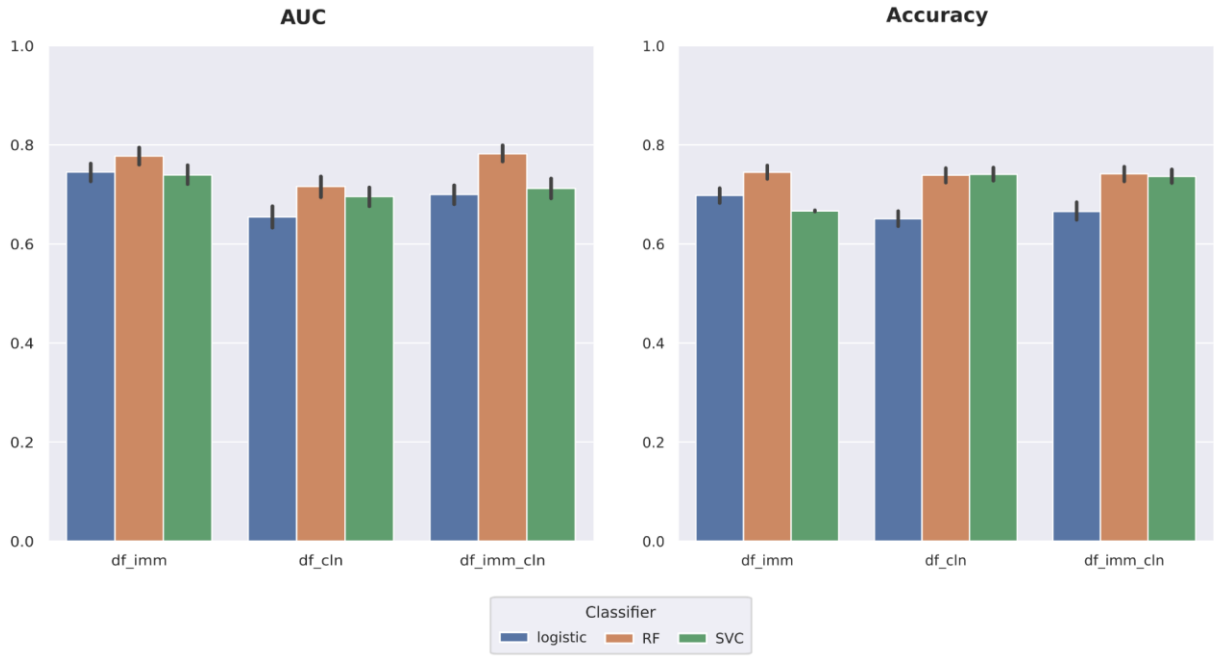

Figure 4.1: Predictive performance of different models and feature sets.

The figure displays the AUC (left) and accuracy (right) of logistic regression, random forest, and SVC models evaluated through stratified Monte Carlo cross-validation. Performance is shown for three different feature sets: immunological blood markers, clinical variables, and the combination of both.

4.2.3 Exploring radiomics data heterogeneity

Radiomic studies are susceptible to batch effects, which can arise from variations in imaging protocols, scanners, and reconstruction algorithms. These technical variations can introduce non-biological variability into the extracted features, potentially compromising the reliability and generalizability of radiomic models. To mitigate the impact of batch effects, rigorous

preprocessing steps are essential. Batch effect identification techniques, in particular principal component analysis (PCA), were used to visualize these variations, finding no evident contribution from factors such as the protocol used, the type of scanner, and the type of reconstruction algorithm (**Figure 4.2**). Given this, no correction methods were required to adjust for batch-specific biases.

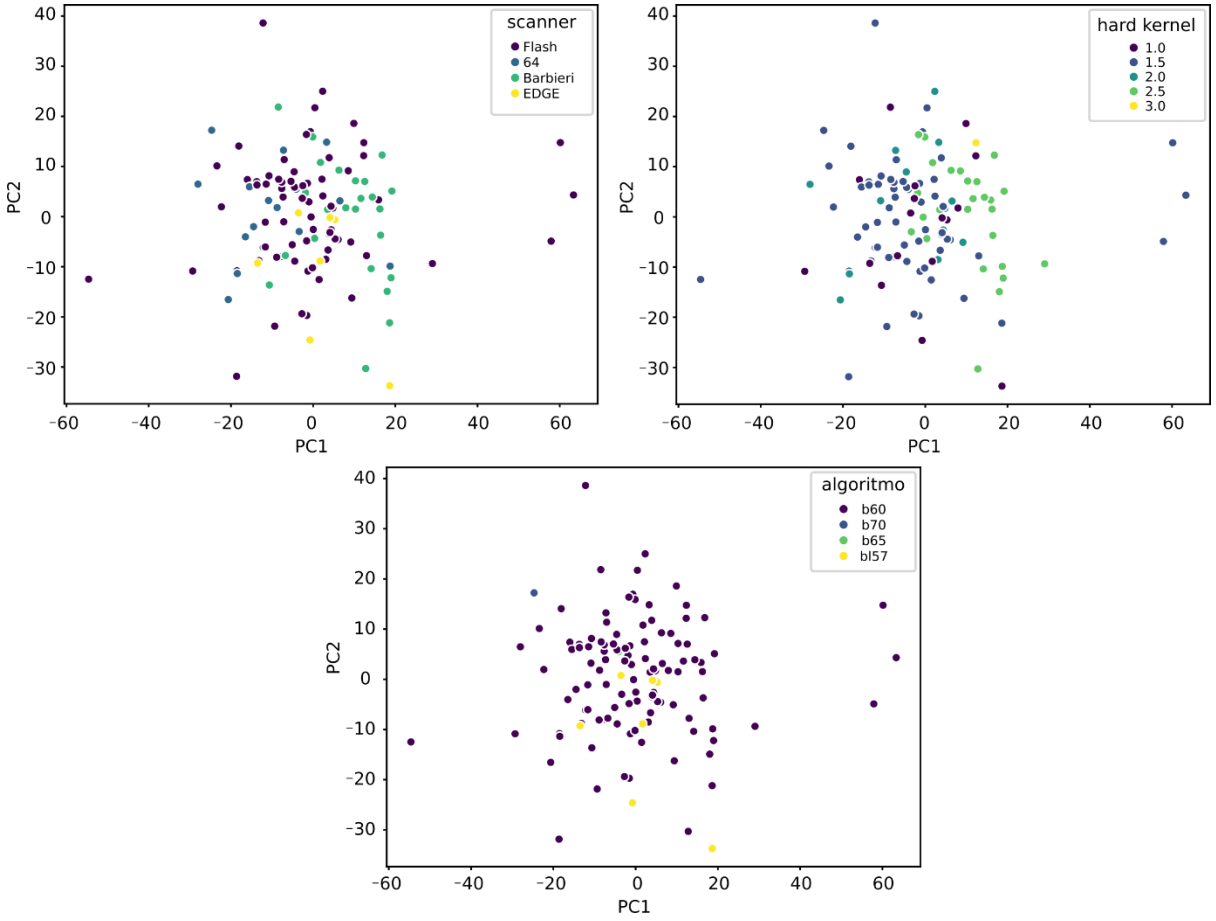

Figure 4.2 Assessment of batch effects using principal component analysis (PCA).
The figure displays PCA plots examining the influence of four potential confounding factors on radiomic features: (clockwise from top left) scanner used, slice thickness (hard kernel) in mm, and reconstruction algorithm. Each point represents a patient sample, and the distribution of points across the principal components illustrates the degree to which each factor contributes to variability in the radiomic data.

Moreover, in radiomics, the extraction of a large number of features can lead to high-dimensional datasets, often exceeding the number of samples. In our case, the number of features available is 851 compared to 117 total samples. This high-dimensional space can pose challenges for model training, depending on the model used (e.g., logistic regression), increasing the risk of overfitting and hindering model generalization. Additionally, the

presence of redundant, highly correlated features can further exacerbate these issues (e.g., in the construction of models such as random forest, SVC). Therefore, feature selection is crucial in radiomics to identify and retain the most informative and independent features, improving model performance, reducing overfitting, and enhancing model interpretability.
To address this problem, we implemented a feature selection pipeline, testing different filtering approaches to better leverage the performance of our dataset.

4.2.4 Radiomics feature selection pipeline

To identify the most informative radiomic features, we employed a feature selection pipeline and evaluated its effectiveness using stratified Monte Carlo cross-validation. This robust approach helps assess model performance and mitigate the risk of overfitting. At each fold of the cross-validation, a specific filter was applied to the radiomic dataset. These filters included correlation analysis, hierarchical correlation analysis, L1-regularized SVC penalty, and combinations thereof (see Materials and Methods for details). Subsequently, different machine learning models, including logistic regression, SVC, and random forest, were trained and tested on the corresponding training and test folds.
Applying the various filters resulted in distinct sets of selected features, yet the overall predictive performance remained consistent across different filters and machine learning models, with an average AUC of approximately 0.7 (**Figure 4.3B**). This performance represents a substantial improvement over the baseline AUC of 0.6, observed when using all available features without filtering (**Figure 4.3A**). Notably, precision and recall metrics were also significantly enhanced by feature selection. These results clearly indicate that reducing the number of features through filtering leads to improved predictive performance. Interestingly, the choice of machine learning model did not substantially influence the results, suggesting that a simpler, more interpretable model like logistic regression may be preferred in this context, given its ease of clinical application and understanding.
To further investigate feature importance, the selected features were ranked by their frequency of selection, categorized into groups of 1, 5, 10, 20, 30, and all features available. The performance of these ranked feature subsets was assessed in a similar manner exploiting the logistic regression model to determine whether more frequently selected features were indeed more informative.

As a result, more frequently selected features appear to be more informative than complete sets of features, likely due to a combination of factors: the inherent presence of more informative features and the influence of dimensionality on model performance. These reduced feature sets achieved performance levels of ~0.8 AUC (**Figure 4.3C**). This observation suggests that further feature selection aimed at reducing the dimensionality of the feature sets and maximizing their informative power could potentially yield improved results when coupled with an easily interpretable model such as logistic regression.

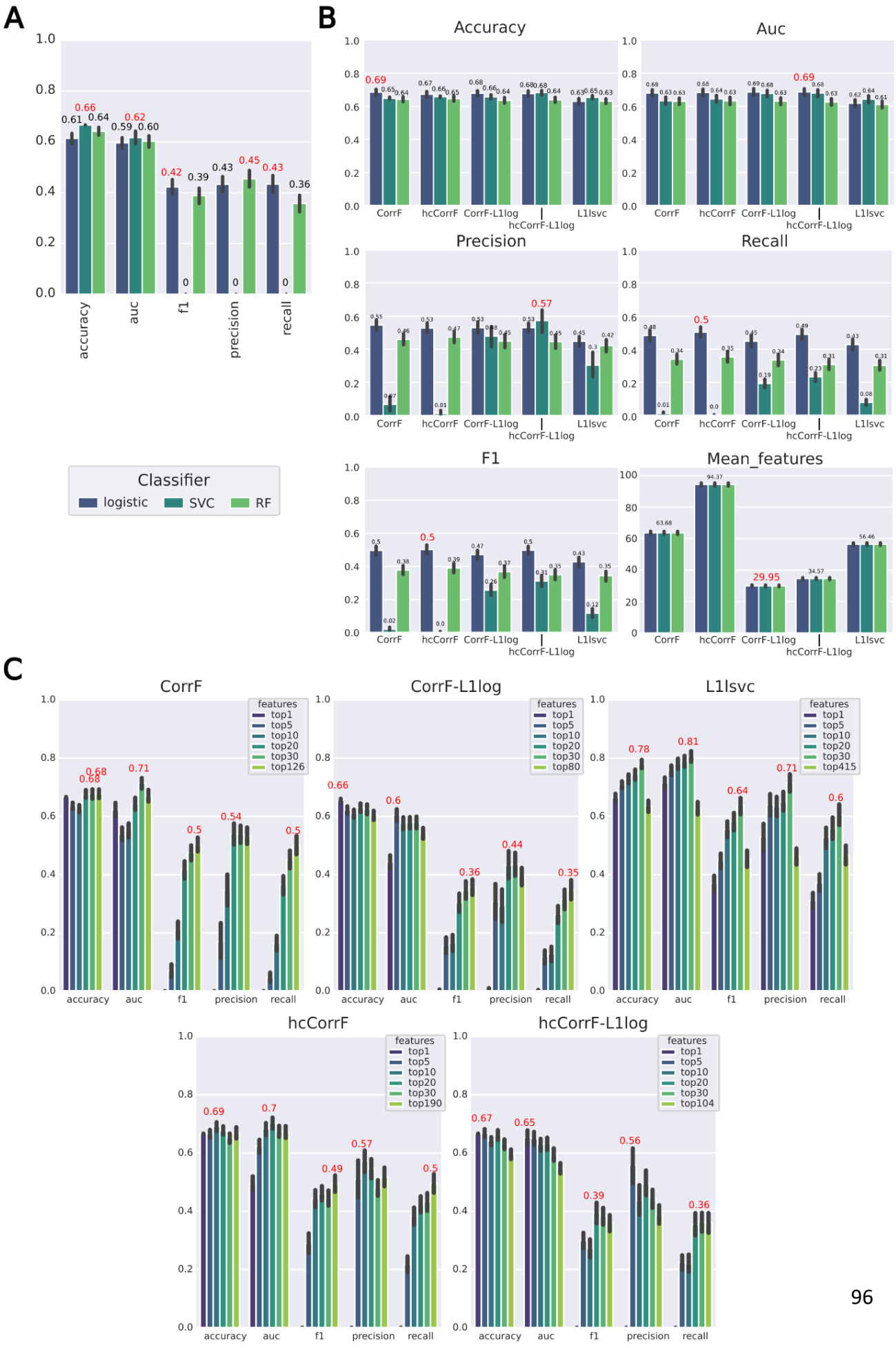
A
B
C
Accuracy
Auc
Precision
Recall
F1
Mean_features
Classifier
logistic
SVC
RF
CorrF
hcCorrF
CorrF-L1log
hcCorrF-L1log
L1lsvc
accuracy
auc
f1
precision
recall
features
top1
top5
top10
top20
top30
top126
top80
top415
top190
top104

Figure 4.3 Performance evaluation of the radiomics feature selection pipeline.
(A) *Baseline performance of the unfiltered radiomic feature set. The plot displays the performance of logistic regression, SVC, and random forest models trained on all available radiomic features, evaluated through stratified Monte Carlo cross-validation. Performance metrics include AUC, accuracy, F1-score, precision, and recall.* ***(B)*** *Performance of different feature selection filters. Each plot shows the performance of the three models (logistic regression, SVC, and random forest) using features selected by different filters (indicated on the x-axis). Performance is evaluated using various metrics (one per plot).* ***(C)*** *Predictive performance of top-ranked features. For each feature selection filter, the most frequently selected features across cross-validation folds were identified and used to train a logistic regression model. Each plot represents a different filter, and performance is evaluated using the metrics shown on the x-axis.*

To further refine feature selection, we employed a meta-heuristic optimization approach known as a genetic algorithm (GA). Genetic algorithms are inspired by natural selection, starting with a random population of solutions (sets of features) and iteratively selecting and recombining the fittest solutions to generate improved offspring. This process continues until a satisfactory solution is found.

In our implementation, the population comprised a fixed number of randomly selected features from the pools generated by the filtering pipeline. The fitness function was defined as a logistic regression model evaluated within a stratified Monte Carlo cross-validation framework. The choice of logistic regression was motivated by its simplicity and interpretability, allowing for a clearer understanding of feature contributions. Applying a similar approach to more complex models like random forests, which possess inherent feature importance discrimination mechanisms, could result in less effective fitness functions. This is because such models are capable of mitigating the impact of feature redundancy and noise, potentially obscuring the true contribution of individual features within the genetic algorithm optimization process.

This genetic algorithm process was repeated for each filtered feature set at various solution sizes, ranging from 10 to 35 features, to prevent overfitting the logistic regression model. This approach identified three optimal feature sets from different initial sets, demonstrating an increase in performance, reaching average ~0.9 AUC in two filter sets (**Figure 4.4A**).

Interestingly, an analysis of the composition of the three top-performing feature sets (correlation filter with 15 features, hierarchical correlation filter with 20, and 15 features) revealed that the selected features were largely unique to each set (**Figure 4.4B**). This observation suggests that the performance improvement is not solely driven by individual features but rather by the complex interplay among them.

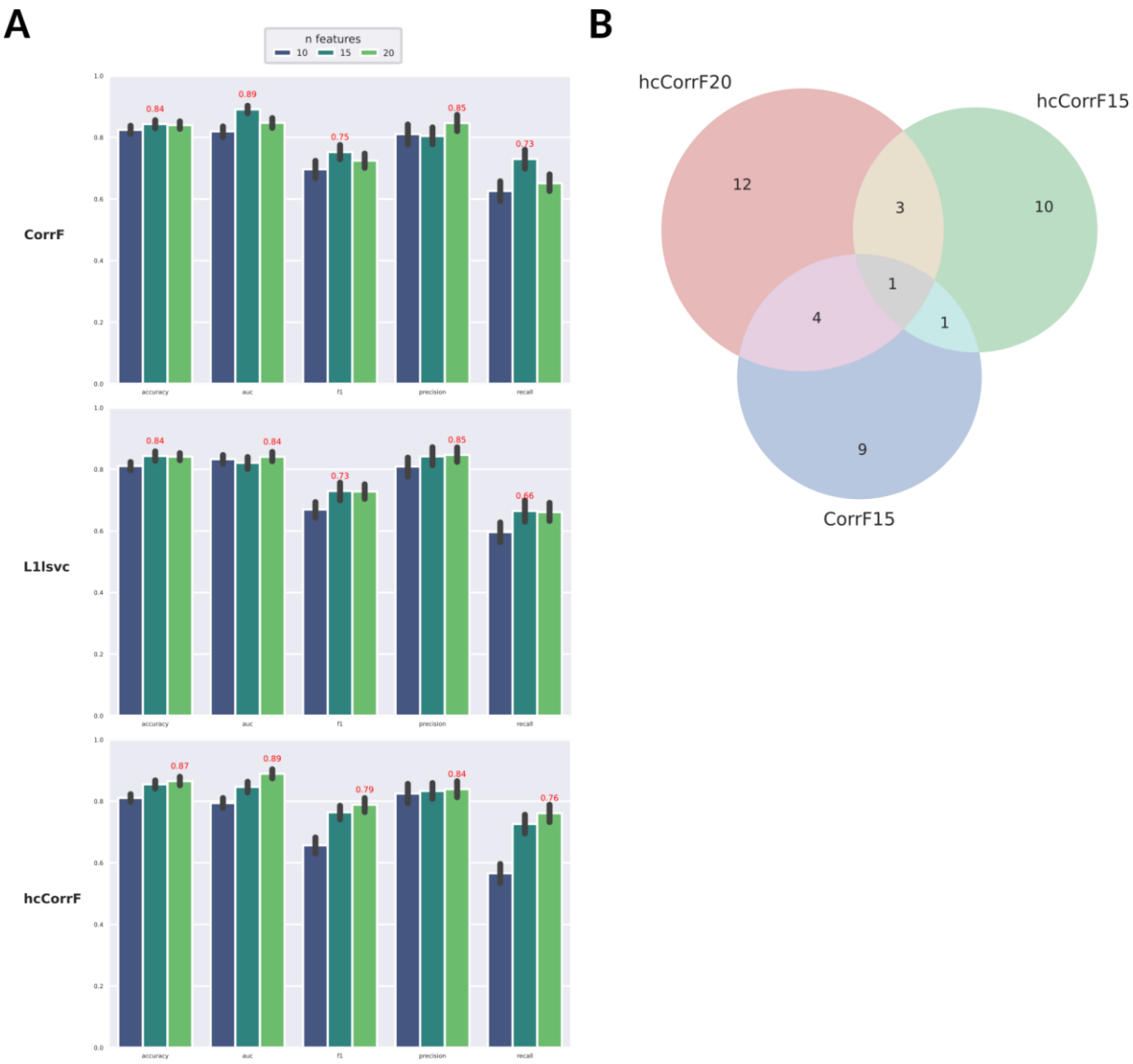

Figure 4.4 Optimization of radiomic feature selection using a genetic algorithm.

(A) *Performance of logistic regression models trained on GA optimized feature sets. Each plot displays the performance of logistic regression models trained on feature sets of varying sizes (10, 15, and 20) selected by a genetic algorithm from an initial pool of filtered features. Performance is evaluated using the metrics shown on the x-axis.* ***(B)*** *Overlap of features in top-performing sets. The Venn diagram illustrates the shared and unique features among the three top-performing feature sets identified by the genetic algorithm.*

4.2.5 Integration of blood derived markers with radiomics features

Although integrating blood-based markers with radiomic features did not significantly improve predictive performance (**Figure 4.5**), it is crucial to note that the performance did not deteriorate. This observation suggests that both sets of features contribute valuable, albeit potentially overlapping, information regarding the underlying biological processes associated with treatment response. Interestingly, while logistic regression models achieved the highest performance, likely due to the optimization process being tailored to this model, the other models (SVC and random forest) also demonstrated improved performance when applied to the selected feature sets. This finding indicates that the selected features, while optimized for logistic regression, possess inherent informative power for discriminating patient response, independent of the specific model employed.

Further supporting this notion, an additional genetic algorithm selection process was conducted. This process, initiated with the same filtered radiomic sets but augmented with immunological markers (LDH and dNLR) and clinical variables (smoking status, bone lesions, hepatic lesions, number of disease sites, ECOG PS), resulted in the selection of a feature set combining radiomic features with both immunological markers. This augmented set demonstrated a slight increase in performance (**Figure 4.5**), suggesting that a judicious combination of features derived from these two non-invasive approaches could potentially enhance the prediction of therapy response. This observation highlights the potential for synergistic interactions between radiomic and blood-based markers, warranting further investigation into optimal strategies for integrating these modalities.

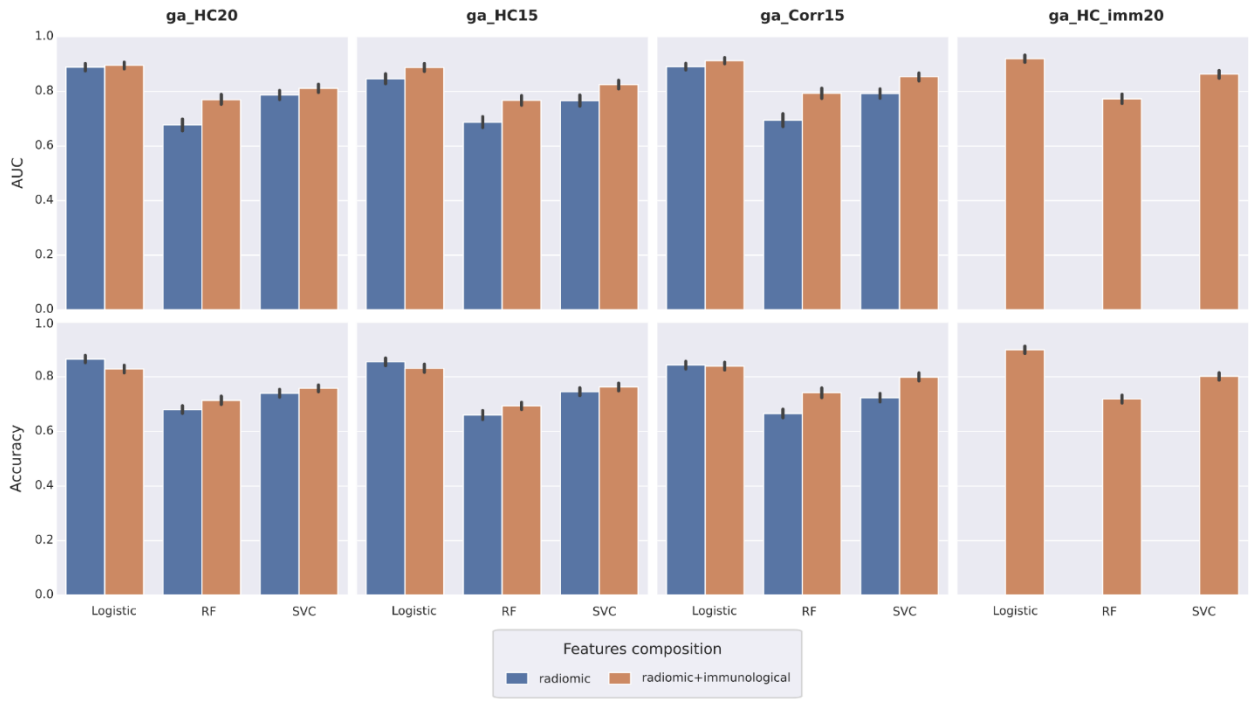

Figure 4.5 Predictive performance of integrated radiomic and blood-based features.

This figure illustrates the predictive performance of three machine learning models (logistic regression, random forest, and SVC) trained on distinct feature sets optimized through a genetic algorithm. Each panel showcases the results for a different feature set, including the top three performing sets composed solely of radiomic features and a set derived from hierarchically clustered radiomic features combined with immunological and clinical variables. Model performance is evaluated using AUC (top row) and accuracy (bottom row). Blue bars denote models trained on radiomic features only, while orange bars indicate models incorporating both radiomic and immunological features. Notably, the rightmost panel highlights a feature set where the genetic algorithm selected both radiomic and immunological features, underscoring the potential for synergistic interactions between these modalities to enhance predictive accuracy.

4.2.6 Selected feature's ability to predict patients' survival

To further evaluate the optimized feature sets, we assessed their ability to predict patient outcomes. Predictions derived from these sets were used to generate survival curves, which were then compared to the survival curve obtained using the actual therapy response variable. Median overall survival (OS) times were calculated and used to perform survival analysis, stratifying patients based on the predicted response from each feature set. The results, illustrated in **Figure 4.6**, demonstrate that the selected features effectively capture the prognostic information related to therapy response and can accurately mimic the observed survival patterns in the patient cohort.

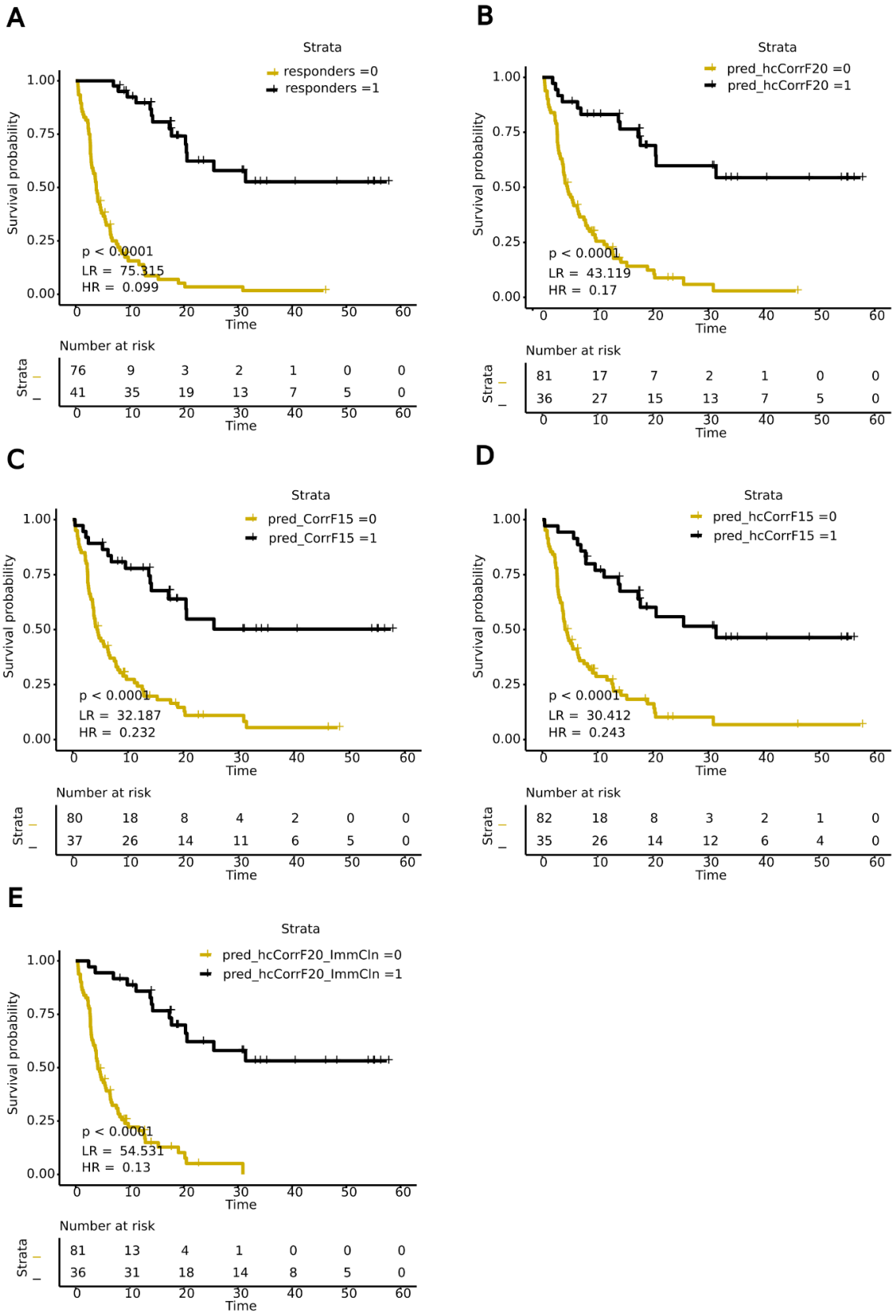

Figure 4.6 Survival analysis based on predicted treatment response.

(A) *Survival analysis of patients stratified by actual treatment response. This Kaplan-Meier curve displays the survival probability of patients based on their actual response to immunotherapy.* ***(B-E)*** *Survival analysis of patients stratified by predicted treatment response. These Kaplan-Meier curves illustrate the survival probability of patients stratified by their predicted response to immunotherapy, as determined by the top-performing feature sets identified in this study (one set per plot).*

4.3 Discussion

This study investigated the potential of integrating radiomic features with blood-based markers of systemic inflammation to predict treatment response in patients with advanced NSCLC receiving immune checkpoint inhibitors (ICIs). Our findings demonstrate the feasibility of constructing a robust predictive model by combining these two modalities, offering a non-invasive approach to patient stratification and personalized treatment strategies.

The analysis of blood-derived pro-inflammatory markers, namely basal LDH and dNLR, revealed their potential as indicators of immunotherapy response. While their individual predictive performance, as assessed by the AUC of 0.7, highlights their clinical relevance, it also underscores the need for complementary approaches to enhance predictive accuracy. This observation aligns with previous research emphasizing the complex interplay of factors influencing ICI response, including tumor microenvironment characteristics, PD-L1 expression, and systemic inflammatory responses.

Radiomic analysis, with its ability to capture intratumoral heterogeneity and provide a comprehensive quantitative assessment of tumor phenotypes, offers a promising avenue for improving patient stratification. However, the high dimensionality of radiomic data necessitates careful feature selection to mitigate the risks of overfitting and ensure model generalizability. In this regard, the results obtained from our radiomics feature selection method (AUC ~0.8) and further genetic algorithm selection (AUC ~0.9) appear promising, there remains a significant concern regarding the potential for overfitting, despite the use of cross-validation techniques. While cross-validation is a widely accepted approach for model evaluation, it does not entirely eliminate the risk of overfitting, especially when working with relatively small datasets or when there is a high number of features in relation to the sample size. This issue becomes particularly evident in our context where the high dimensionality of the data can lead to models that are overly tailored to the specific characteristics of the training set.

To accurately assess the true performance and generalizability of the model, it is crucial to validate the findings on an external cohort or a larger independent dataset. Such external validation would help determine whether the model's performance holds up across different patient populations, and imaging acquisition protocols. A more robust train-test separation, using larger and more diverse cohorts, is necessary to ensure that the model is not simply

capturing noise or idiosyncratic patterns specific to the training data. Only through such validation could we firmly confirm the model's performaces, reproducibility, and clinical utility. Therefore, additional efforts towards obtaining larger, external cohorts and performing rigorous cross-validation with independent datasets are essential to strengthen the reliability and robustness of the radiomics-based feature selection method.

Interestingly, the integration of literature's assessed blood-based markers LDH and dNLR with radiomic features did not further enhance predictive performance, suggesting that both modalities capture similar aspects of the underlying biological processes related to treatment response. However, a subsequent analysis incorporating immunological markers and clinical variables alongside radiomic features demonstrated a slight performance improvement. This finding indicates the potential for combining these non-invasive approaches to achieve a more comprehensive and accurate prediction of treatment response.

The selected radiomic features, when used to predict patient outcomes, demonstrated the ability to mimic the survival patterns observed in the cohort. This finding supports the notion that radiomic features can capture essential prognostic information related to treatment response, potentially complementing or even surpassing traditional biomarkers.

The integration of radiomic features with blood-based inflammatory markers represents a significant step towards a more comprehensive and personalized approach to managing advanced NSCLC in the era of ICIs. This strategy could potentially guide treatment decisions, identify patients most likely to benefit from immunotherapy, and ultimately improve clinical outcomes.

This study has certain limitations that warrant discussion. Firstly, as already discussed, the reproducibility of radiomics studies is influenced by the inherent heterogeneity of the data, stemming from various technical and biological factors. Although validation strategies were employed to mitigate the risk of overfitting, future studies on independent cohorts are crucial to confirm the robustness of our findings.

Furthermore, exploring more sophisticated approaches for integrating blood-based biomarkers with radiomic information is necessary. In this study, we prioritized simple and interpretable statistical models to facilitate clinical understanding and application. However, more complex approaches, such as applying deep learning techniques directly to images, could potentially mitigate the heterogeneity and intrinsic noise of radiomics approaches, leading to improved patient characterization.

In conclusion, the integration of radiomic features with blood-based inflammatory markers represents a significant step towards a more comprehensive and personalized approach to managing advanced NSCLC in the era of ICIs. This strategy could potentially guide treatment decisions, identify patients most likely to benefit from immunotherapy, and ultimately improve clinical outcomes. Further research is warranted to validate these findings in larger, independent cohorts and to explore more sophisticated integration strategies.

4.4 Materials and Methods

4.4.1 Cohort of study

All the non–oncogene-addicted stage IV NSCLC patients who initiated ICIs from March 2017 to July 2019 at the Medical Oncology Unit of the University Hospital of Parma (Italy) were considered for inclusion in the present study. In particular, 117 NSCLC patients were selected. All these patients underwent evaluations comprehensive of CT imaging, clinical laboratory investigations, including Eastern Cooperative Oncology Group (ECOG) performance status, clinical stage, and peripheral blood sampling for LDH and dLNR immunological markers. Primary endpoint was the response to ICIs per RECIST (version 1.1) (134). Patients with complete/partial response and stable disease lasting at least 6 months were defined as responders and those with stable disease lasting less than 6 months or progression as non-responders.

4.4.2 Computed Tomography

CT examinations were acquired within 1 week from the respective clinical laboratory assessment using various multidetector CT scanners (SOMATOM Definition Flash, SOMATOM Definition Edge, SOMATOM Sensation 64 Cardiac, SOMATOM Emotion 6; Siemens Healthineers, Erlangen, Germany). All the acquisitions were performed following contrast medium administration. Parameters of the imaging series retrieved for the analyses were as follows: slice thickness, 1.5 to 2.5 mm; tube voltage, 100 to 120 kVp; automated exposure

control for tube current; contiguous or overlap reconstruction interval; smooth kernels. The tumor burden (TB) computed by consultant radiologists in charge following the RECIST 1.1 criteria was retrospectively collected from the electronic medical records for the purposes of the current analyses.

4.4.3 Radiomic dataset construction

The postcontrast CT images were uploaded to a dedicated workstation for processing and feature extraction. For each examination, a volumetric segmentation of the primary tumor was performed by a 5-year experienced thoracic radiologist blinded to clinical data other than the tumor location. All the segmentations were performed semiautomatically using open-source software (3D Slicer, Version 4.11), adjusting the window width and level to properly annotate tumor borders, excluding bronchi, blood vessels, vacuoles, normal lung tissue, chest wall, and mediastinal structures. A total of 25 randomly selected tumors were also segmented by a 3-year experienced thoracic radiologist to evaluate the feature robustness. The volumes of interest were resampled to 1x1x1-mm voxel size using linear interpolation to counter different slice thicknesses. Moreover, signal intensity values were discretized to a bin width of 25 to further reduce variability. RFs were extracted using the open-source tool pyRadiomics (62) integrated as a plugin into 3D Slicer software.

4.4.4 PCA for batch inspection

Dimensionality reduction technique called Principal Component Analysis (PCA) was used to address the presence of batched effects. PCA identifies patterns of variance within a dataset by projecting it onto a lower-dimensional subspace formed by principal components (PCs), which represent directions of maximum variance in the data. In the context of batch identification and exploration, PCA is utilized to identify clusters or distinct batches of samples with similar patterns of variation. Visualization of these components allows to discriminate co-factors' role in the similarity composition of the dataset. Python (3.9.20) package scikit-learn (v0.24.2) (135) was used for scaling the data and performing PCA decomposition. Results were plotted using python package matplotlib (v3.3.4) with inspected variables data highlighted by different colors.

4.4.5 Logistic regression

Logistic regression analyzes the relationship between a binary categorical dependent variable and multiple independent continuous variables, also called explanatory variables. The logistic model estimates the probability of occurrence of an event by fitting data to a logistic curve. In this work, logistic regression is used both in the context of feature filter selection and as a model to discriminate between patients who do and don't respond to therapy. Logistic regression is implemented through the python package scikit-learn (v0.24.2) with default parameters.

4.4.6 Support vector machine

Support Vector Machines analyze the relationship between a binary categorical dependent variable and multiple independent continuous variables, also called explanatory variables. The SVM model estimates the decision boundary by maximizing the margin between classes, ensuring optimal classification accuracy. In this work, SVM is used both in the context of feature selection and as a model to classify patients into different response categories based on their characteristics. SVM is applied using the python package scikit-learn (v0.24.2) with internal parameters C=1 and gamma=0.1.

4.4.7 Random Forest Classifier

Random Forest is an ensemble learning method that leverages the power of multiple decision trees to perform classification tasks. Each tree in the forest is constructed using a random subset of features and data points, promoting diversity and robustness. The final classification is determined by aggregating the predictions of individual trees, often through a majority voting scheme. In this study, the Random Forest classifier is employed to predict patient response to therapy based on a given feature set. The implementation utilizes the class from the python package scikit-learn (v0.24.2) with parameters number of trees=100, criterion=Gini impurity, max_depth=4, max_features=None (all features considered).

4.4.8 Correlation based filter

To mitigate the impact of redundant features in the radiomic dataset, a correlation-based filtering approach was employed. This method identifies and removes highly correlated

features, which can introduce noise and instability into machine learning models. Pairwise Pearson's correlations were calculated for all radiomic features, generating a correlation matrix. A threshold of 0.75 was applied to the absolute correlation values in the upper triangle of this matrix. Features exceeding this threshold were considered highly correlated and thus redundant in terms of informative value. One feature from each highly correlated pair was removed from the dataset to reduce dimensionality and improve model performance. This filtering process was implemented using custom functions compatible with the scikit-learn (v0.24.2) Python library.

4.4.9 Hierarchical correlation based filter

This method groups features into clusters based on their correlation patterns, allowing for a more nuanced understanding of feature relationships. Specifically, hierarchical clustering with a correlation-based affinity metric and complete linkage was applied to the radiomic feature set (with a default cluster number of k=10). This resulted in the formation of distinct clusters, each containing features exhibiting similar correlation profiles.

Following cluster identification, each cluster was subjected to the correlation-based filtering method described previously. This within-cluster filtering ensured that only a representative subset of features from each cluster was retained, reducing redundancy while preserving the diversity of information captured across different clusters. This combined approach leverages both the global correlation structure across all features and the local correlation patterns within individual clusters to achieve a more robust and informative feature selection.

4.4.10 L1 based filter

The L1 penalty filter leverages regularization techniques inherent to logistic regression and support vector classification (SVC) to identify and eliminate uninformative features. By incorporating an L1 penalty term into the loss function, the model is encouraged to assign small or zero coefficients to features with weak or negligible contributions to the prediction task. This effectively shrinks the coefficients of irrelevant features, driving them towards zero.

In this study, the L1 penalty filter is applied to both logistic regression and SVC models trained on the radiomic dataset. Features with coefficients below a model internal threshold (|β| < threshold) are deemed uninformative and removed. Conversely, features with coefficients exceeding the threshold (|β| > threshold) are retained as relevant for subsequent analyses. This approach facilitates feature selection by identifying and prioritizing features that exert a substantial influence on the model's predictive performance.

4.4.11 Stratified Monte Carlo cross-validation

Monte Carlo Cross-Validation is employed to evaluate the performance of machine learning models and selected pools of features. One fold consists in the creation of a train and test partitions from the dataset, respectively 80% and 20% of its sample size. To ensure that each fold represents similar characteristics to the entire dataset, stratified random sampling is used to maintain the same proportion of classes within each fold. The cross-validation process involves evaluating the model's performance on each fold using metrics such as accuracy, precision, and recall. By repeating this process 100 times, Monte Carlo Cross-Validation provides an estimate of the model's expected performance on unseen data.

4.4.12 Feature selection pipeline

Stratified Monte Carlo cross-validation exploiting feature selection is performed. In particular, at each cross-validation iteration, a pre-processing step of features scaling, followed by a feature selection step and logistic regression model training are performed on the training fold which is then evaluated on the test fold. The cross validation process is performed 100 times and metrics such as accuracy, AUC, precision, and recall are collected. This is repeated for different filters, in particular: Correlation, Hierarchical correlation, Correlation followed by L1 through logistic, Hierarchical correlation followed by L1 through logistic, and L1 through SVC. For each applied filter, frequent features appearing in the various folds of the cross validation are collected.

4.4.13 Genetic Algorithm for feature selection

Genetic Algorithms (GAs) are a type of optimization technique part of the population-based metaheuristic algorithms inspired by the principles of natural selection and genetics (136).

GAs consist of three primary components: an individual population of candidate solutions, a fitness function to evaluate the quality of each solution, and a set of operators (such as mutation, crossover, and selection) (**Supplementary Figures S4.1, S4.2**) that manipulate the population over generations. As shown in **Supplementary Fig. S4.1**, the algorithm begins with an initial population of random or predefined solutions, which are then evaluated using the fitness function to determine their relative "fitness" or quality. At each generation, the fittest individuals are selected for reproduction, individuals undergo mutation and crossover operations to generate new offspring. Through this iterative process of selection, mutation, and crossover, GAs iteratively converge towards a set of optimal solutions that satisfy the problem's constraints. In the context of application for feature selection, the population of candidate solutions consists of 50 individuals with initial random selection of a fixed number of features from a provided set. The fitness function consists of a monte carlo validation approach using the dataset provided for evaluating the performance of the selected population. In particular, pools of features are selected from the dataset and monte carlo validation approach coupled to a Logistic Regression Model is performed. Average accuracy performance throughout the validation is then used as fitness score. The selection step is responsible for choosing two individuals from the current population to be included in the new generation. Each individual is chosen through tournament selection, a random subset of 3 individuals is chosen from the current population. The individual with the highest fitness value within this subset is then selected as a parent.

The two selected parents are then subjected to crossover with a crossover probability rate. If crossover condition is met, two children are generated and moved to the mutation step, otherwise, the parents are. The mutation step has a specified probability rate, in which a mutation, in the form of a random substitution of an introduced filter, is performed. The newly generated individuals are then assigned to the new population. These steps are repeated until the new population reaches the specified dimension. At this point, elitism is performed, in which the best candidate so far is reintroduced into the population. One iteration of these steps is considered as one generation. The algorithm runs for a fixed number of generations i_{total}. During this generation, crossover and mutation probability's rates are modified in order to better explore the space of optimal solutions, with crossover reducing at each iteration i_{cycle} as

$$i_{cycle} \,/\, i_{total}$$

and mutation increasing as

$$1 - \frac{i_{cycle}}{i_{total}}$$

At its conclusion, the algorithm returns the best individual obtained throughout the various generations.

4.4.14 Survival analysis

Survival analysis is an area of statistics designed for modeling data of time to disease remission, progression or death for cohorts of patients or to compare different treatments within a clinical trial.

Data regarding Median OS of the 117 advanced NSCLC analyzed is used to perform survival analysis. To perform the survival analysis the R package survival was used. This package allows to apply Kaplan-Meier *survivor* function, Log Rank Test, and the Cox proportional hazard ratio. The survivor function represents the probability that a subject will be yet to experience death (or the considered endpoint) by a given time. The Log Rank Test statistically evaluates generated survival curves. The hazard ratio represents the instantaneous risk of a subject experiencing death (or the considered endpoint) at a given time and is reported alongside its corresponding p-value.

Chapter 5 - Conclusions

This thesis aimed to develop and apply innovative computational approaches for analyzing cancer data obtained through non-invasive methodologies. The research focused on enhancing the detection and characterization of circulating tumor DNA (ctDNA) in cancer patients, and exploring the integration of complementary non-invasive approaches, such as radiomics, with liquid biopsy for a more comprehensive understanding of cancer biology and improved patient care.

First, we developed Synggen, a tool that generates realistic synthetic cancer data crucial for developing and benchmarking new ctDNA analysis methods. This addresses the limitations of using real patient data, such as privacy concerns and data heterogeneity, by providing a standardized and flexible resource for researchers.

Second, we created eSENSES, a custom NGS panel combined with a tailored computational approach that enhances the sensitivity and specificity of ctDNA detection and quantification, even at low ctDNA fractions. This enables more accurate identification of SCNAs, providing a valuable tool for monitoring and understanding tumor evolution.

Finally, we investigated the potential of radiomic features extracted from medical images to predict treatment response in lung cancer patients. This exploration aimed to complement liquid biopsy approaches by providing additional non-invasive information for patient stratification and personalized treatment decisions.

This research is not without limitations. The clinical evaluations of eSENSES and the radiomics analysis were conducted on relatively small cohorts, potentially limiting the generalizability of the findings. Larger-scale studies are needed to validate these results. Additionally, while eSENSES demonstrated promising results in detecting ctDNA at low fractions, the challenge of capturing the full spectrum of tumor heterogeneity remains. Further research is needed to optimize the approach to address the complexities of clonal evolution. Despite the potential of radiomics, the lack of standardized imaging protocols and feature extraction methods poses a challenge for reproducibility and widespread clinical implementation.

Looking ahead, we see several promising future directions for this work. Incorporating additional genomic features and data modalities into synggen will further enhance its utility for benchmarking and developing more sophisticated ctDNA analysis tools. Further validation

and refinement of eSENSES will pave the way for its integration into clinical practice, potentially improving disease monitoring and treatment decisions for breast cancer patients. In the future, eSENSES could be extended to incorporate fragmentomics data. Fragmentomics is the study of the size distribution of cfDNA fragments. This distribution can provide valuable information about the origin and nature of cfDNA, as different biological processes can lead to distinct fragmentation patterns. By incorporating fragmentomics data, eSENSES could potentially improve the accuracy of ctDNA characterization and quantification.

Combining liquid biopsy, radiomics or more in general image analysis, and other clinical data through advanced machine learning models holds the promise of achieving a more comprehensive understanding of cancer biology and enabling truly personalized medicine.

This thesis wants to contribute to precision oncology by providing valuable tools and insights for non-invasive cancer detection and characterization. The development of Synggen and eSENSES, along with the exploration of radiomics, represents a step towards more accurate disease monitoring, personalized treatment strategies, and improved patient outcomes.

Chapter 6 – Supplementary Material

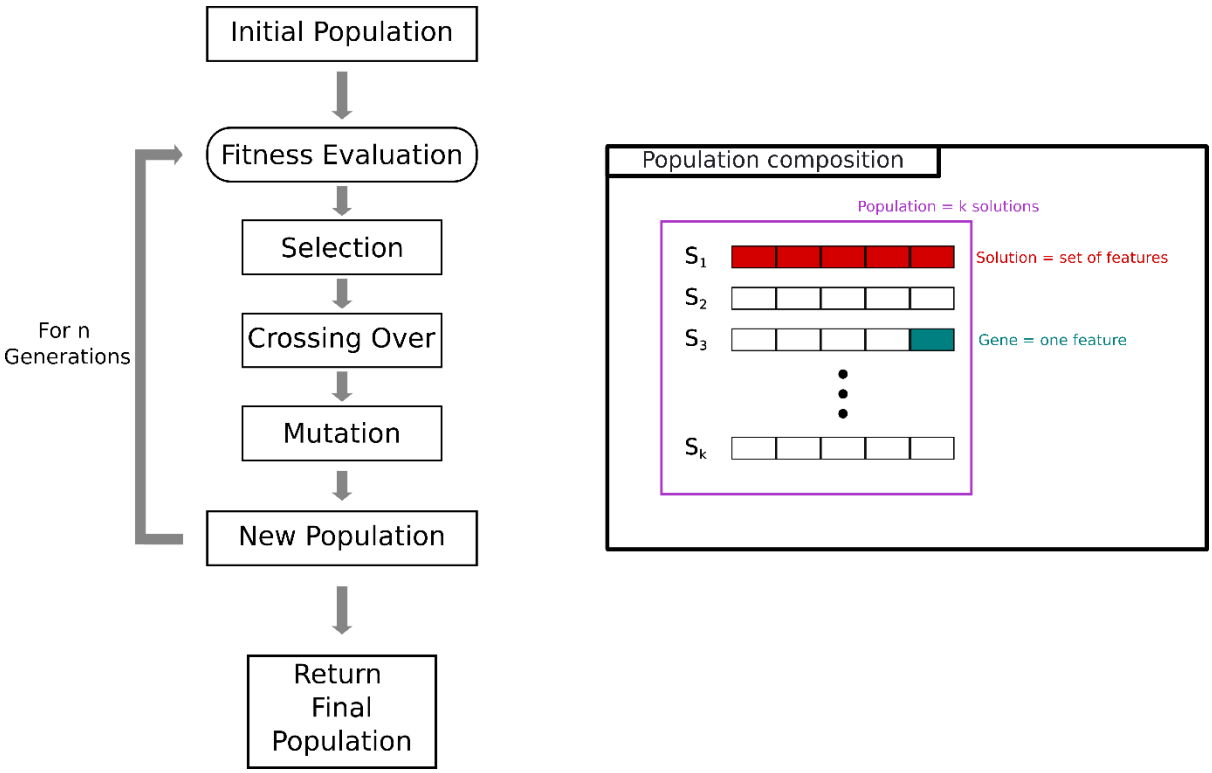

Figure S 6.1 Genetic Algorithm workflow and population composition explanation.

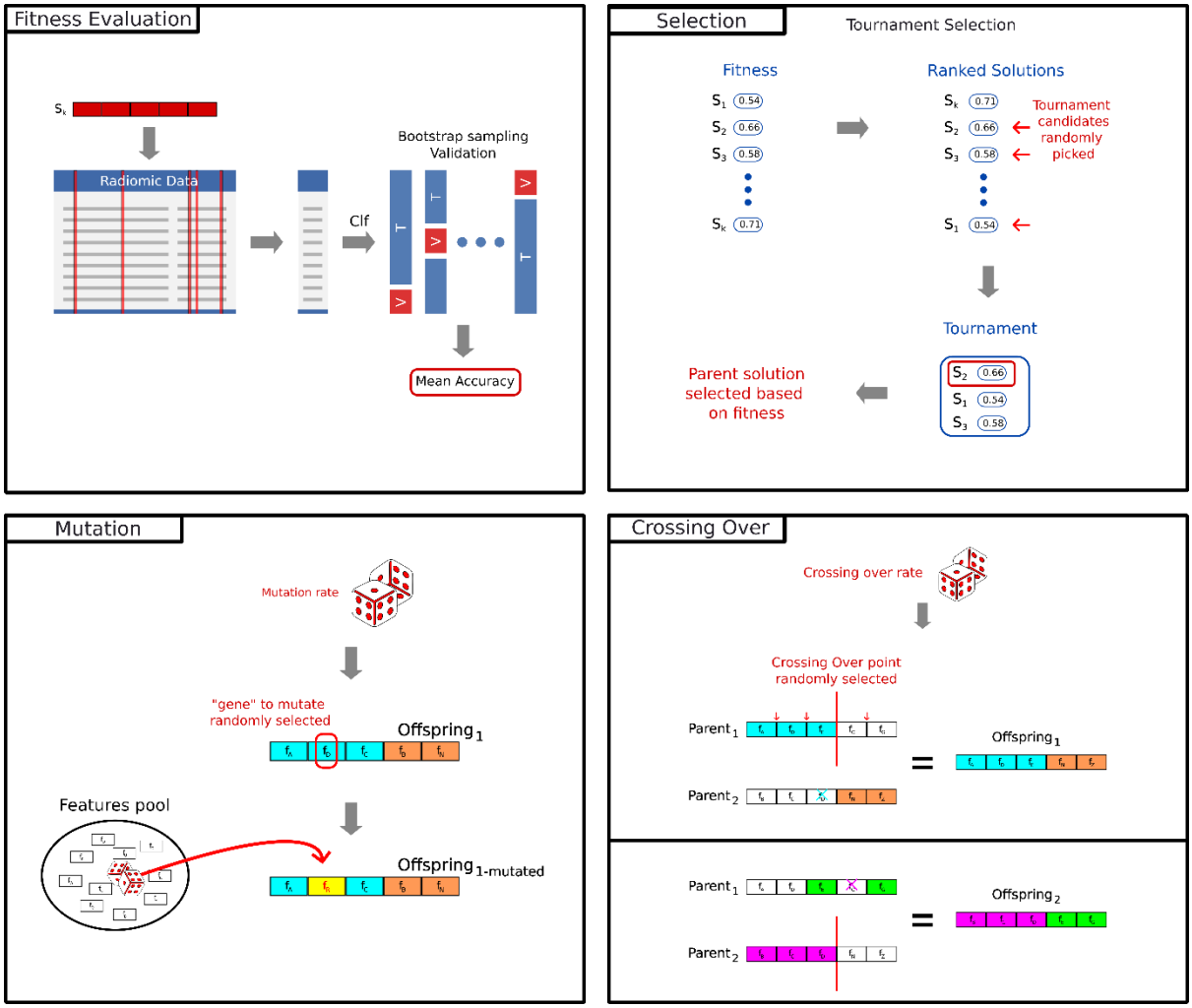

Figure S 6.2 Genetic Algorithm operators infographics.

Chapter 7 – Bibliography

1. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018 Feb;15(2):81–94.

2. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000 Jan;100(1):57–70.

3. Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. Translational Implications of Tumor Heterogeneity. Clin Cancer Res. 2015 Mar 15;21(6):1258–66.

4. Sansregret L, Vanhaesebroeck B, Swanton C. Determinants and clinical implications of chromosomal instability in cancer. Nat Rev Clin Oncol. 2018 Mar;15(3):139–50.

5. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017 Jun;23(6):703–13.

6. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018 Apr;173(2):321-337.e10.

7. International Early Lung Cancer Action Program Investigators, Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, et al. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med. 2006 Oct 26;355(17):1763–71.

8. Neal RD, Tharmanathan P, France B, Din NU, Cotton S, Fallon-Ferguson J, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. Br J Cancer. 2015 Mar 31;112(S1):S92–107.

9. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Sci Transl Med [Internet]. 2016 Jul 6 [cited 2024 Dec 9];8(346). Available from: https://www.science.org/doi/10.1126/scitranslmed.aaf6219

10. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol. 2017 Sep;14(9):531–48.

11. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019 Nov;16(11):703–15.

12. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet. 2019 Feb;20(2):71–88.

13. Blanco BA, Wolfgang CL. Liquid biopsy for the detection and management of surgically resectable tumors. Langenbecks Arch Surg. 2019 Aug;404(5):517–25.

14. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018 Dec;2(1):36.

15. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014 Jun 3;5(1):4006.

16. Caputo V, Ciardiello F, Corte CMD, Martini G, Troiani T, Napolitano S. Diagnostic value of liquid biopsy in the era of precision medicine: 10 years of clinical evidence in cancer. Explor Target Anti-Tumor Ther. 2023;4(1):102–38.

17. Cristofanilli M, Broglio KR, Guarneri V, Jackson S, Fritsche HA, Islam R, et al. Circulating Tumor Cells in Metastatic Breast Cancer: Biologic Staging Beyond Tumor Burden. Clin Breast Cancer. 2007 Feb 1;7(6):34–42.

18. Health C for D and R. List of Cleared or Approved Companion Diagnostic Devices (In Vitro and Imaging Tools). FDA [Internet]. 2025 Mar 5 [cited 2025 Mar 10]; Available from: https://www.fda.gov/medical-devices/in-vitro-diagnostics/list-cleared-or-approved-companion-diagnostic-devices-in-vitro-and-imaging-tools

19. Song J, Ye X, Xiao H. Liquid biopsy entering clinical practice: Past discoveries, current insights, and future innovations. Crit Rev Oncol Hematol. 2025 Mar 1;207:104613.

20. Ulz P, Heitzer E, Geigl JB, Speicher MR. Patient monitoring through liquid biopsies using circulating tumor DNA. Int J Cancer. 2017;141(5):887–96.

21. Pasini L, Ulivi P. Liquid Biopsy for the Detection of Resistance Mechanisms in NSCLC: Comparison of Different Blood Biomarkers. J Clin Med. 2019 Jul;8(7):998.

22. Lamb YN, Dhillon S. Epi proColon® 2.0 CE: A Blood-Based Screening Test for Colorectal Cancer. Mol Diagn Ther. 2017 Apr 1;21(2):225–32.

23. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol Off J Eur Soc Med Oncol. 2021 Sep;32(9):1167–77.

24. Alix-Panabières C, Pantel K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. Cancer Discov. 2016 May 1;6(5):479–91.

25. Diaz LA, Bardelli A. Liquid Biopsies: Genotyping Circulating Tumor DNA. J Clin Oncol. 2014 Feb 20;32(6):579–86.

26. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018 Nov 7;10(466):eaat4921.

27. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem. 2015 Jan;61(1):112–23.

28. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic - implementation issues and future challenges. Nat Rev Clin Oncol. 2021 May;18(5):297–312.

29. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nat Rev Genet. 2019 Feb;20(2):71–88.

30. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555–72.

31. Lefebvre C, Bachelot T, Filleron T, Pedrero M, Campone M, Soria JC, et al. Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. PLoS Med. 2016 Dec;13(12):e1002201.

32. Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016 May 10;7:11479.

33. Keller L, Belloum Y, Wikman H, Pantel K. Clinical relevance of blood-based ctDNA analysis: mutation detection and beyond. Br J Cancer. 2021 Jan;124(2):345–58.

34. Elazezy M, Joosse SA. Techniques of using circulating tumor DNA as a liquid biopsy component in cancer management. Comput Struct Biotechnol J. 2018 Jan 1;16:370–8.

35. Ou CY, Vu T, Grunwald JT, Toledano M, Zimak J, Toosky M, et al. An ultrasensitive test for profiling circulating tumor DNA using integrated comprehensive droplet digital detection. Lab Chip. 2019 Mar 13;19(6):993–1005.

36. Bruhm DC, Vulpescu NA, Foda ZH, Phallen J, Scharpf RB, Velculescu VE. Genomic and fragmentomic landscapes of cell-free DNA for early cancer detection. Nat Rev Cancer. 2025 Mar 4;1–18.

37. Stetson D, Ahmed A, Xu X, Nuttall BRB, Lubinski TJ, Johnson JH, et al. Orthogonal Comparison of Four Plasma NGS Tests With Tumor Suggests Technical Factors are a Major Source of Assay Discordance. JCO Precis Oncol. 2019 Dec;3:1–9.

38. Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. Comput Struct Biotechnol J. 2019 Jan 1;17:1348–59.

39. Bagger FO, Borgwardt L, Jespersen AS, Hansen AR, Bertelsen B, Kodama M, et al. Whole genome sequencing in clinical practice. BMC Med Genomics. 2024 Jan 29;17(1):39.

40. Paoli M, Galardi F, Nardone A, Biagioni C, Romagnoli D, Donato SD, et al. Sensitive tumor detection, accurate quantification, and cancer subtype classification using low-pass whole methylome sequencing of plasma DNA [Internet]. 2024 [cited 2024 Dec 4]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.06.10.598204

41. Dawson MA. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. Science. 2017 Mar 17;355(6330):1147–52.

42. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. The Lancet. 2018 Sep 1;392(10149):777–86.

43. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. Bioorg Med Chem Lett. 2007 Apr 1;17(7):1912–5.

44. Qi T, Pan M, Shi H, Wang L, Bai Y, Ge Q. Cell-Free DNA Fragmentomics: The Novel Promising Biomarker. Int J Mol Sci. 2023 Jan 12;24(2):1503.

45. Sirajee AS, Kabiraj D, De S. Cell-free nucleic acid fragmentomics: A non-invasive window into cellular epigenomes. Transl Oncol. 2024 Nov 1;49:102085.

46. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA | Nature Communications [Internet]. [cited 2025 Mar 12]. Available from: https://www.nature.com/articles/s41467-022-35076-w

47. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009 Aug;10(8):551–64.

48. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006 Nov;444(7118):444–54.

49. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012 Apr 18;486(7403):346–52.

50. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb;463(7283):899–905.

51. Yang L, Shi P, Zhao G, Xu J, Peng W, Zhang J, et al. Targeting cancer stem cell pathways for cancer therapy. Signal Transduct Target Ther. 2020 Feb 7;5(1):8.

52. Speirs V. Quality Considerations When Using Tissue Samples for Biomarker Studies in Cancer Research. Biomark Insights. 2021;16:11772719211009513.

53. Sunyaev S. Prediction of deleterious human alleles. Hum Mol Genet. 2001 Mar 1;10(6):591–7.

54. Sherry ST. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–11.

55. The International SNP Map Working Group, Cold Spring Harbor Laboratories:, Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928–33.

56. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep;467(7311):52–8.

57. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 2005 Jul 15;65(14):6071–9.

58. Van Loo P, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010 Sep 28;107(39):16910–5.

59. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016 Sep 19;44(16):e131.

60. Orlando F, Romanel A, Trujillo B, Sigouros M, Wetterskog D, Quaini O, et al. Allele-informed copy number evaluation of plasma DNA samples from metastatic prostate cancer patients: the PCF_SELECT consortium assay. NAR Cancer. 2022 Jun;4(2):zcac016.

61. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012 Mar;48(4):441–6.

62. Van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017 Nov 1;77(21):e104–7.

63. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep. 2015 Aug 17;5(1):13087.

64. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol. 2018 Nov;102(4):1143–58.

65. Van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—“how-to” guide and critical reflection. Insights Imaging. 2020 Dec;11(1):91.

66. Peng B, Leong MC, Chen HS, Rotunno M, Brignole KR, Clarke J, et al. Genetic Simulation Resources and the GSR Certification Program. Bioinformatics. 2019 Feb 15;35(4):709–10.

67. Tanner G, Westhead DR, Droop A, Stead LF. Simulation of heterogeneous tumour genomes with HeteroGenesis and in silico whole exome sequencing. Bioinforma Oxf Engl. 2019 Aug 15;35(16):2850–2.

68. Semeraro R, Orlandini V, Magi A. Xome-Blender: A novel cancer genome simulator. PLOS ONE. 2018 Apr 5;13(4):e0194472.

69. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. PLOS ONE. 2016 Nov 28;11(11):e0167047.

70. Valentini S, Fedrizzi T, Demichelis F, Romanel A. PaCBAM: fast and scalable processing of whole exome and targeted sequencing data. BMC Genomics. 2019 Dec;20(1):1–5.

71. Casiraghi N, Orlando F, Ciani Y, Xiang J, Sboner A, Elemento O, et al. ABEMUS: platform-specific and data-informed detection of somatic SNVs in cfDNA. Bioinforma Oxf Engl. 2020 May 1;36(9):2665–74.

72. Qvick A, Stenmark B, Carlsson J, Isaksson J, Karlsson C, Helenius G. Liquid biopsy as an option for predictive testing and prognosis in patients with lung cancer. Mol Med Camb Mass. 2021 Jul 3;27(1):68.

73. Kaisaki PJ, Cutts A, Popitsch N, Camps C, Pentony MM, Wilson G, et al. Targeted Next-Generation Sequencing of Plasma DNA from Cancer Patients: Factors Influencing Consistency with Tumour DNA and Prospective Investigation of Its Utility for Diagnosis. PLOS ONE. 2016 Sep 14;11(9):e0162809.

74. Freitas AJA de, Causin RL, Varuzza MB, Calfa S, Hidalgo Filho CMT, Komoto TT, et al. Liquid Biopsy as a Tool for the Diagnosis, Treatment, and Monitoring of Breast Cancer. Int J Mol Sci. 2022 Sep 1;23(17):9952.

75. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 2020 Jul 21;11(1):3475.

76. Reinert T, Henriksen TV, Christensen E, Sharma S, Salari R, Sethi H, et al. Analysis of Plasma Cell-Free DNA by Ultradeep Sequencing in Patients With Stages I to III Colorectal Cancer. JAMA Oncol. 2019 Aug 1;5(8):1124–31.

77. Mazzitelli C, Santini D, Corradini AG, Zamagni C, Trerè D, Montanaro L, et al. Liquid Biopsy in the Management of Breast Cancer Patients: Where Are We Now and Where Are We Going. Diagnostics. 2023 Mar 25;13(7):1241.

78. Garcia-Murillas I, Chopra N, Comino-Méndez I, Beaney M, Tovey H, Cutts RJ, et al. Assessment of Molecular Relapse Detection in Early-Stage Breast Cancer. JAMA Oncol. 2019 Oct 1;5(10):1473–8.

79. Hrebien S, Citi V, Garcia-Murillas I, Cutts R, Fenwick K, Kozarewa I, et al. Early ctDNA dynamics as a surrogate for progression-free survival in advanced breast cancer in the BEECH trial. Ann Oncol Off J Eur Soc Med Oncol. 2019 Jun 1;30(6):945–52.

80. Chabon JJ, Hamilton EG, Kurtz DM, Esfahani MS, Moding EJ, Stehr H, et al. Integrating genomic features for non-invasive early lung cancer detection. Nature. 2020 Apr;580(7802):245–51.

81. Reichert ZR, Morgan TM, Li G, Castellanos E, Snow T, Dall'Olio FG, et al. Prognostic value of plasma circulating tumor DNA fraction across four common cancer types: a real-world outcomes study. Ann Oncol Off J Eur Soc Med Oncol. 2023 Jan;34(1):111–20.

82. Cisneros-Villanueva M, Hidalgo-Pérez L, Rios-Romero M, Cedro-Tanda A, Ruiz-Villavicencio CA, Page K, et al. Cell-free DNA analysis in current cancer clinical trials: a review. Br J Cancer. 2022 Feb;126(3):391–400.

83. Herberts C, Annala M, Sipola J, Ng SWS, Chen XE, Nurminen A, et al. Deep whole-genome ctDNA chronology of treatment-resistant prostate cancer. Nature. 2022 Aug;608(7921):199–208.

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

85. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015 Oct 8;163(2):506–19.

86. Shahrouzi P, Forouz F, Mathelier A, Kristensen VN, Duijf PHG. Copy number alterations: a catastrophic orchestration of the breast cancer genome. Trends Mol Med. 2024 Aug;30(8):750–64.

87. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017 Nov 6;8(1):1324.

88. Beltran H, Romanel A, Conteduca V, Casiraghi N, Sigouros M, Franceschini GM, et al. Circulating tumor DNA profile recognizes transformation to castration-resistant neuroendocrine prostate cancer. J Clin Invest. 2020 Apr 1;130(4):1653–68.

89. Yu L, Lopez G, Rassa J, Wang Y, Basavanhally T, Browne A, et al. Direct comparison of circulating tumor DNA sequencing assays with targeted large gene panels. Brody JP, editor. PLOS ONE. 2022 Apr 28;17(4):e0266889.

90. Li W, Huang X, Patel R, Schleifman E, Fu S, Shames DS, et al. Analytical evaluation of circulating tumor DNA sequencing assays. Sci Rep. 2024 Feb 29;14(1):4973.

91. Rueda OM, Sammut SJ, Seoane JA, Chin SF, Caswell-Jin JL, Callari M, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019 Mar;567(7748):399–404.

92. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.

93. Wagle N, Painter C, Anastasio E, Dunphy M, McGillicuddy M, Kim D, et al. The Metastatic Breast Cancer (MBC) project: Accelerating translational research through direct patient engagement. J Clin Oncol [Internet]. 2017 May 20 [cited 2024 Dec 4]; Available from: https://ascopubs.org/doi/10.1200/JCO.2017.35.15_suppl.1076

94. Scandino R, Calabrese F, Romanel A. Synggen: fast and data-driven generation of synthetic heterogeneous NGS cancer data. Bioinforma Oxf Engl. 2023 Jan 1;39(1):btac792.

95. Romagnoli D, Nardone A, Galardi F, Paoli M, De Luca F, Biagioni C, et al. MIMESIS: minimal DNA-methylation signatures to quantify and classify tumor signals in tissue and cell-free DNA samples. Brief Bioinform. 2023 Mar 19;24(2):bbad015.

96. Carrié H, Sim NL, Wong PM, Gan A, Lau YT, Poon P, et al. Comprehensive benchmarking of methods for mutation calling in circulating tumor DNA [Internet]. bioRxiv; 2025 [cited 2025 Mar 10]. p. 2025.02.16.638482. Available from: https://www.biorxiv.org/content/10.1101/2025.02.16.638482v1

97. Rickles-Young M, Tinoco G, Tsuji J, Pollock S, Haynam M, Lefebvre H, et al. Assay Validation of Cell-Free DNA Shallow Whole-Genome Sequencing to Determine Tumor Fraction in Advanced Cancers. J Mol Diagn JMD. 2024 May;26(5):413–22.

98. Paoli M, Galardi F, Nardone A, Biagioni C, Romagnoli D, Donato SD, et al. Sensitive tumor detection, accurate quantification, and cancer subtype classification using low-pass whole methylome sequencing of plasma DNA [Internet]. bioRxiv; 2024 [cited 2025 Mar 7]. p. 2024.06.10.598204. Available from: https://www.biorxiv.org/content/10.1101/2024.06.10.598204v1

99. Huebner A, Black JRM, Sarno F, Pazo R, Juez I, Medina L, et al. ACT-Discover: identifying karyotype heterogeneity in pancreatic cancer evolution using ctDNA. Genome Med. 2023 Apr 20;15(1):27.

100. Stover DG, Parsons HA, Ha G, Freeman SS, Barry WT, Guo H, et al. Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. J Clin Oncol. 2018 Feb 20;36(6):543–53.

101. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.

102. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012 May;2(5):401–4.

103. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1.

104. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020 Oct;20(10):555–72.

105. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017 Mar;27(3):491–9.

106. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

107. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

108. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinforma Oxf Engl. 2009 Dec 15;25(24):3207–12.

109. Hsu FH, Chen HIH, Tsai MH, Lai LC, Huang CC, Tu SH, et al. A model-based circular binary segmentation algorithm for the analysis of array CGH data. BMC Res Notes. 2011 Dec;4(1):394.

110. Venkatraman E. Seshan AO. DNAcopy [Internet]. Bioconductor; 2017 [cited 2024 Dec 6]. Available from: https://bioconductor.org/packages/DNAcopy

111. Romanel A, Gasi Tandefelt D, Conteduca V, Jayaram A, Casiraghi N, Wetterskog D, et al. Plasma AR and abiraterone-resistant prostate cancer. Sci Transl Med. 2015 Nov 4;7(312):312re10.

112. Prandi D, Baca SC, Romanel A, Barbieri CE, Mosquera JM, Fontugne J, et al. Unraveling the clonal hierarchy of somatic genomic aberrations. Genome Biol. 2014 Aug 26;15(8):439.

113. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast Density-Based Clustering with R. J Stat Softw. 2019 Oct 31;91:1–30.

114. Sondka Z, Dhir NB, Carvalho-Silva D, Jupe S, Madhumita null, McLaren K, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic Acids Res. 2024 Jan 5;52(D1):D1210–7.

115. Barlesi F, Mazieres J, Merlio JP, Debieuvre D, Mosser J, Lena H, et al. Routine molecular profiling of patients with advanced non-small-cell lung cancer: results of a 1-year nationwide programme of the French Cooperative Thoracic Intergroup (IFCT). Lancet Lond Engl. 2016 Apr 2;387(10026):1415–26.

116. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018 Jan 24;553(7689):446–54.

117. Rotow J, Bivona TG. Understanding and targeting resistance mechanisms in NSCLC. Nat Rev Cancer. 2017 Oct 25;17(11):637–58.

118. Duma N, Santana-Davila R, Molina JR. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. Mayo Clin Proc. 2019 Aug;94(8):1623–40.

119. Osmani L, Askin F, Gabrielson E, Li QK. Current WHO guidelines and the critical role of immunohistochemical markers in the subclassification of non-small cell lung carcinoma (NSCLC): Moving from targeted therapy to immunotherapy. Semin Cancer Biol. 2018 Oct;52(Pt 1):103–9.

120. Alsaab HO, Sau S, Alzhrani R, Tatiparti K, Bhise K, Kashaw SK, et al. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. Front Pharmacol. 2017;8:561.

121. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012 Mar 22;12(4):252–64.

122. Bagchi S, Yuan R, Engleman EG. Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. Annu Rev Pathol. 2021 Jan 24;16:223–49.

123. Hargadon KM, Johnson CE, Williams CJ. Immune checkpoint blockade therapy for cancer: An overview of FDA-approved immune checkpoint inhibitors. Int Immunopharmacol. 2018 Sep;62:29–39.

124. Rizvi H, Sanchez-Vega F, La K, Chatila W, Jonsson P, Halpenny D, et al. Molecular Determinants of Response to Anti–Programmed Cell Death (PD)-1 and Anti–Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With Non–Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. J Clin Oncol. 2018 Mar 1;36(7):633–41.

125. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer. 2019 Mar;19(3):133–50.

126. Pitt JM, Vétizou M, Daillère R, Roberti MP, Yamazaki T, Routy B, et al. Resistance Mechanisms to Immune-Checkpoint Blockade in Cancer: Tumor-Intrinsic and -Extrinsic Factors. Immunity. 2016 Jun 21;44(6):1255–69.

127. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014 Jun 3;5:4006.

128. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine Learning methods for Quantitative Radiomic Biomarkers. Sci Rep. 2015 Aug 17;5:13087.

129. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018 Sep;19(9):1180–91.

130. Huang Y, Liu Z, He L, Chen X, Pan D, Ma Z, et al. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non-Small Cell Lung Cancer. Radiology. 2016 Dec;281(3):947–57.

131. Guthrie GJK, Charles KA, Roxburgh CSD, Horgan PG, McMillan DC, Clarke SJ. The systemic inflammation-based neutrophil-lymphocyte ratio: experience in patients with cancer. Crit Rev Oncol Hematol. 2013 Oct;88(1):218–30.

132. Mezquita L, Auclin E, Ferrara R, Charrier M, Remon J, Planchard D, et al. Association of the Lung Immune Prognostic Index With Immune Checkpoint Inhibitor Outcomes in Patients With Advanced Non-Small Cell Lung Cancer. JAMA Oncol. 2018 Mar 1;4(3):351–7.

133. Diem S, Schmid S, Krapf M, Flatz L, Born D, Jochum W, et al. Neutrophil-to-Lymphocyte ratio (NLR) and Platelet-to-Lymphocyte ratio (PLR) as prognostic markers in patients with non-small cell lung cancer (NSCLC) treated with nivolumab. Lung Cancer Amst Neth. 2017 Sep;111:176–81.

134. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer Oxf Engl 1990. 2009 Jan;45(2):228–47.

135. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825–30.

136. Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl. 2021 Feb 1;80(5):8091–126.