Università
di Genova

Ph.D. in Mathematics and Applications

A multiscale journey through omic sciences for cancer: from somatic copy number aberrations to cellular dynamics via mathematical modelling

Candidate:
Giorgia Biddau

Supervisors:
Prof. Michele Piana
Dr. Sara Sommariva

UniGe
DIMA

XXXVIII Cycle

Research Outputs and Dissemination Activities

Papers

Biddau G., Razzetta C., Ferrando L., Zoppoli G., Sommariva S., Piana M. A Multi-Resolution Strategy for Automatic Parameter Selection in ASCAT. In preparation.

Biddau G. & De Falco A., Piana M., Ceccarelli M. Extending SCEVAN to an Allele-Specific Copy Number Inference Framework. In preparation.

Servetti, M., Caramia, M., Parodi, G., Loiacono, F., Nano, E., Biddau, G., ... & Sterlini, B. (2025). Optimization of Transcription Factor-Driven Neuronal Differentiation from Human Induced Pluripotent Stem Cells for Disease Modelling and Drug Screening. Stem Cell Reviews and Reports, 21(3), 816-833.

Parodi, D., Dighero, E., Biddau, G., ... & Sambuceti, G. (2024). Localized FDG loss in lung cancer lesions. EJNMMI research, 14(1), 102.

Brunetti, N., Campi, C., Biddau, G., ... & Tagliafico, A. S. (2024). Radiomic and clinical model for predicting atypical ductal hyperplasia upgrades and potentially reduce unnecessary surgical treatments. European Journal of Radiology, 181, 111799.

Biddau, G., Caviglia, G., Piana, M., & Sommariva, S. (2024). PCA-based synthetic sensitivity coefficients for chemical reaction network in cancer. Scientific Reports, 14(1), 17706.

Sommariva, S., Berra, S., Biddau, G., Caviglia, G., Benvenuto, F., & Piana, M. (2023). In-silico modelling of the mitogen-activated kinase (MAPK) pathway in colorectal cancer: mutations and targeted therapy. Frontiers in systems biology, 3, 1207898.

Talks and Poster at international meetings and Conferences

(2025) Poster presentation: "A PCA-based synthetic tool for local sensitivity analysis of chemical reaction networks". Annual international conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB). Liverpool, UK.

(2025) Flash Talk presentation: "Integration of a multi-resolution approach for automatic parameter selection in ASCAT."21st Annual Meeting of the Bioinformatics Italian Society (BITS), Napoli, Italy.

(2024) Talk: "A local sensitivity analysis tool for chemical reaction networks and application to G1-S phase of colorectal cells". Joint annual meeting of the Korean Society for Mathematical Biology and the Society for Mathematical Biology (KSMB-SMB), Seoul, Republic of Korea.

(2023) Poster presentation: A computational tool for sensitivity analysis in chemical reaction networks modeling cancer cell signaling. 14th conference of Dynamical Systems Applied to Biology and Natural Sciences (DSANBS), Bilbao, Spain.

Participation in Workshops and Schools

(2025) Poster presentation: "A multi-scale approach for automatic parameter selection in copy number analysis with ASCAT". 1st Data Science 4 Health and Biology (DS4HB) workshop, Politecnico di Milano, Milano, Italy.

(2023) Participant presentation: "Sensitivity analysis of chemical reaction networks modelling G1-S phase of physiological and mutated colorectal cells". Come School on Cancer Evolution (CSCE), Como, Italy.

Visiting period

(1st May- 30th June 2025) Visiting Period at Biogem, Ariano Irpino (AV), Italy.

Abstract

The present thesis focuses on the development of robust statistical and computational methods for cancer modelling, with a particular emphasis on the analysis of genomic, transcriptomic, and proteomic data.
In the field of genomics, the analysis performed had the objective of detecting and characterizing mutations, with a particular focus on somatic copy number aberrations (SCNAs), defined as the gain or loss of DNA regions. Specifically, this thesis developed a multi-resolution approach for the segmentation of LogR Ratio and B Allele Frequency data, with a focus on the analysis of data with non-homogeneous noise and region-specific signal to noise ratio.
With regard to transcriptomic data, this thesis proposes an extension of a widely used SCNA detection pipeline for single-cell data, adapting it to resolve allele-specific configurations and enabling more reliable identification of cancer cell subpopulations. The proposed multi-step method addresses the inherent sparsity and noise of single-cell data, and incorporates novel strategies tailored to omics datasets, including graph-based data integration, state-of-the-art clustering techniques, and probabilistic models.
Finally, in the context of proteomics, this thesis proposes a framework for local sensitivity analysis to fine-tune parameters of a Chemical Reaction Network (CRN) modeling key signaling pathways in colorectal cells. By appropriately modifying the network model, it could be possible to simulate mutations induced by SCNAs, and the proposed sensitivity analysis enables the assessment of their effects on cellular dynamics and comparison with the physiological cell state.

This work is for you.
I know you would be proud of me.

Contents

List of Figures

List of Tables

Introduction

One of the central challenges in modern medicine is to disentangle the molecular and cellular processes that govern cancer initiation and progression. Carcinogenesis is indeed a multistep process driven by genetic and cellular alterations, which disrupt the tightly regulated balance governing normal cellular function. At the same time, tumour development and progression vary substantially depending on the tissue of origin, individual genomic background, and environmental or lifestyle factors, precluding the existence of a single, universal explanation.

In response to this complexity, cancer research has increasingly adopted a multidisciplinary perspective, supported by the generation of large-scale and high-dimensional datasets. The integration of diverse fields, such as medicine, biology, mathematics, and statistics, is now essential for investigating carcinogenesis across multiple biological scales, as such complexity cannot be addressed without formal quantitative frameworks.

In this context, mathematical and statistical modelling provide the tools required to represent, analyse, and interpret high-dimensional biological systems, while modern statistical approaches have become indispensable for data processing and interpretation.

Among the research domains most profoundly transformed by these advances are the omics sciences, which aim to characterise biological processes at the molecular level from complementary perspectives. While traditional bulk omics technologies provided averaged measurements across populations of cells, recent developments in single-cell technologies have further expanded

the capabilities of omics approaches, allowing the exploration of cellular heterogeneity, clonal diversity and, in some cases, spatial organisation within tumours. Within this context, genomics investigates alterations in the deoxyribonucleic acid (DNA) sequence and structure of the genome, transcriptomics studies ribonucleic acid (RNA) transcripts and their expression levels, reflecting gene regulation and activity, while proteomics examines the abundance, interactions, and dynamics of proteins, which constitute the fundamental components of cellular functions.

The high-resolution capabilities of modern omics technologies have revealed multiple classes of genomic alterations, among which somatic copy number aberrations (SCNAs) play a central role. These events, involving gains or losses of DNA regions, are widely studied in cancer research because they serve as important molecular biomarkers and can influence tumour behaviour across many cancer types. Importantly, SCNAs affect cellular function by perturbing signalling pathways, which orchestrate key cellular decisions, such as proliferation, survival, and differentiation. These pathways can be described as large-scale dynamical systems governed by complex networks of molecular interactions.

Many computational pipelines have been developed to detect SCNAs, providing powerful frameworks to tackle complex inference problems by integrating statistical modelling, signal processing, and algorithmic optimisation. Moreover, mathematical models of signalling pathways, such as Chemical Reaction Networks (CRNs), formalise biochemical reactions through systems of equations and kinetic parameters, enabling the systematic investigation of how genomic perturbations influence cellular dynamics.

However, both inference from omics data and dynamical modelling of signalling networks present significant methodological challenges. In SCNAs analysis, many pipelines rely on user-defined parameter choices that can strongly influence the outcomes of genomic profiles, and single-cell data introduce additional challenges due to noise, sparsity, and technical variability, motivating the development of dedicated methodological frameworks. In

the context of signalling pathway modelling, the high dimensionality of the resulting networks complicates interpretation, parameter identifiability, and robustness assessment. Systematic sensitivity analysis becomes therefore essential to identify the parameters and mechanisms that most strongly influence system behaviour, supporting model simplification and dimensionality reduction.

Within this integrated perspective, the objective of this thesis is to develop mathematical and computational approaches for the quantitative analysis and improvements of existing pipeline for the detection of genomic aberrations and their downstream effects on cellular signalling. These methods aim to provide a structured framework for interpreting complex biological data, improving our understanding of cancer development and progression, and leveraging the opportunities offered by emerging high-throughput technologies.

The thesis is organised as follows.

- Chapter 1 introduces the biological background necessary to understand the main mechanisms underlying cancer. It provides an overview of DNA and other key biomolecules, together with the fundamental processes underlying gene regulation and cellular function. The chapter also presents a concise review of sequencing technologies and the main types of genomic data, together with basic concepts of signalling pathway modelling.

- Chapter 2 introduces an optimization-based framework for the automatic selection of parameters in copy number analysis. Specifically, the identification of SCNA occurrences across the genome can be formulated as a change-point detection problem, which can be addressed through regularization-based methods. Within this framework, the choice of the regularization parameter plays a crucial role, as different values may lead to solutions that vary in smoothness and noise levels.

  The proposed methodology formulates parameter tuning as a model selection problem, based on the maximization of a criterion derived from Bayesian principles. By formally defining this selection strategy, the ap-

proach is further extended to ASCAT [1], a widely adopted tool for bulk DNA copy number analysis, thereby reducing user-dependent variability in somatic copy number inference.

- Chapter 3 presents a multi-step algorithmic framework for allele-specific somatic copy number alteration inference in single-cell RNA sequencing data. While existing approaches, including SCEVAN [2], primarily focus on total copy number estimation, they do not explicitly model allele-specific aberrations, which are crucial for accurately characterizing clonal architecture and tumour evolution. The proposed method extends SCEVAN by jointly integrating gene expression signals and allele-level information within a unified computational framework. By combining graph-based data integration, clustering strategies, and probabilistic modelling, the approach enables robust high-resolution inference of allele-specific copy number states, while explicitly accounting for sparsity, technical noise, and intra-tumour heterogeneity inherent to single-cell sequencing data.

- Chapter 4 develops a methodological framework for local sensitivity analysis in large-scale CRNs. The approach introduces a set of indices, termed Statistical Sensitivity Indices (SSIs), which quantify the response of equilibrium states to parameter perturbations, providing a structured assessment of parameter influence in high-dimensional nonlinear systems. This framework is then applied to cellular signalling pathway models, where the large number of interacting components makes sensitivity analysis a crucial tool for dimensionality reduction and for the biological interpretation of the network.

Chapter 1

A Gentle Introduction to Molecular and Systems Biology

1.1 A Molecular Biology Overview for Cancer Genomics

In recent years, the integration of mathematical and computational frameworks into biological research has enabled a quantitative understanding of complex cellular and molecular processes. Advances in biotechnology have yielded rich, high-dimensional datasets that allow unprecedented quantitative investigation of biological systems. At the same time, mathematical frameworks have been developed to analyse these datasets and derive predictive models of cellular behaviour. To make the most of this joint effort, mathematical approaches must first accurately model biological systems and, secondly, rely on a solid understanding of the nature of the data, the underlying mechanisms, as well as awareness of possible sources of variability and error. Therefore, mathematical approaches to cancer research must account for the structure and function of pivotal biomolecules, as well as the biochemical structure of cellular processes, and a solid understanding of basic biology is essential to determine which mathematical tools are most appropriate for their analysis and modelling.
To this end, this chapter provides an overview of the key molecular compo-

nents involved in carcinogenesis and introduces a specific class of aberrations associated with tumour initiation and progression. The concepts presented here form the biological background required for understanding the mathematical models discussed in the following chapters. The sources consulted are listed below and are gratefully acknowledged: "Overview of deoxyribonucleic acid (DNA)", "Signal transduction overview", "Somatic copy number alterations in human cancers: an analysis of publicly available data from the cancer genome atlas", "The human genome: an introduction", "The essence of SNPs", "Guide to the draft human genome" [3, 4, 5, 6, 7, 8].

Among the possible mechanisms, the accumulation of unrepaired alterations in deoxyribonucleic acid (DNA) is one of the most common events leading to the transformation of a normal cell into a cancer cell. DNA alterations have indeed a significant impact on cellular processes, such as protein production and other cellular activities, since one of DNA's primary functions is to store genetic information that regulates essential cellular processes, including cell growth and division.

At the molecular level, DNA is located within the nucleus of most human cells and consist of two strands that coil around each other to form a double helix. Each strand is composed of simpler units called nucleotides, formed of one of four fundamental nucleobases, i.e. cytosine [C], guanine [G], adenine [A] and thymine [T], a sugar called deoxyribose, and a phosphate group. In a single DNA strand, two consequent nucleotides are linked by covalent phosphodiester bonds, which are formed between the sugar of one nucleotide and the phosphate group of the other. Conversely, the two strands are held together by specific rules stabilized by hydrogen bonds, forming the so-called base pairs (bp): A pairs with T, and G pairs with C. The resultant bonds are responsible for the characteristic aforementioned double-stranded structure of DNA, with strands running in opposite directions (antiparallel) and carrying complementary genetic information.

In healthy human somatic cells, that is, all non-reproductive cells of the body, DNA is organized into 46 chromosomes arranged in 23 homologous pairs, a

condition referred to as diploidy.
Out of these chromosomes in somatic cells, 44 are known as autosomes, while the remaining 2 are defined as sex chromosomes, since they are responsible for an individual's biological sex. Autosomes are organized into 22 homologous pairs and are numbered from 1 to 22 primarily according to size, with chromosome 1 being the largest and chromosome 22 the smallest. Each pair of homologous autosomes consists of one chromosome inherited from the biological mother and one from the biological father. Sex chromosomes are commonly referred to as X and Y due to their shape and are also inherited in pairs, but with the difference that the mother always contributes an X chromosome, while the father contributes either an X or a Y chromosome.
Each chromosome is divided into two arms (or bands), the short arm (p) and the long arm (q), which are separated by a specialized chromosomal region known as the centromere.
As mentioned above, DNA contains the instructions for a living cell to function correctly, and the regions that operate as a carrier of this type of genetic information are designated as genes. Genes consist of segments of DNA comprising a certain sequence of nucleotides, and are located in precise positions on specific chromosomes, called loci, and their arrangement along each chromosome reflects an organized and evolutionarily conserved genomic structure. Within this context, genes are embedded in chromatin, a complex structure composed of DNA and associated proteins, mainly histones, which contributes to the spatial organization and accessibility of genetic information. The number of genes in the human genome has been estimated to be on the order of tens of thousands, while their size can vary considerably, typically ranging from 10.000 to 150.000 base pairs.
Genes exist in multiple forms, called alleles, which contribute to genetic diversity, and the combination of alleles on homologous chromosomes determines an individual's genotype. When both copies of a gene carry the same allele, the genotype is said to be homozygous, when they differ, it is called heterozygous. This genetic diversity is often caused by changes in DNA, including single-

nucleotide polymorphisms (SNPs). SNPs are defined as variants that occur at specific loci within a population, where the less frequent allele has a minimum frequency of 1%. Although most SNPs do not alter gene function, some can influence gene products and their functioning, contributing to the previously mentioned genetic differences observed among individuals.
Even SNPs may present at different allelic forms, and when two alleles are observed, they are conventionally denoted as A and B. The B allele often refers to the alternative allele, that is, the allele less frequent in a population, while A denotes the reference allele, i.e. the most common one. Beyond individual loci, the joint configuration of SNPs along a chromosome is often inherited and is referred to as a haplotype, defined as a specific combination of alleles at multiple SNP loci transmitted together on the same chromosome.
It is estimated that approximately ten million SNPs are present in the human genome [9], and a substantial fraction of known SNPs has been identified and curated within dedicated genomic databases, which provide standardized information on genomic coordinates, population allele frequencies, and potential functional impact [10, 11, 12].
To provide a visual overview of the genomic structures just discussed, Figure 1.1 presents a simplified scheme of the organization of DNA, SNPs, and alleles, while Figure 1.2 shows a representation of the complete set of chromosomes, or karyotype, under normal conditions.
From a functional viewpoint, one of the primary roles of genes is to store the genetic information necessary for a cell to perform its essential activities. Different alleles of the same gene can influence how these activities are carried out, leading to variations in cellular behavior and function. One of the main cellular activities directed by genes is protein synthesis, which translates genetic information into functional molecules that sustain cell life.

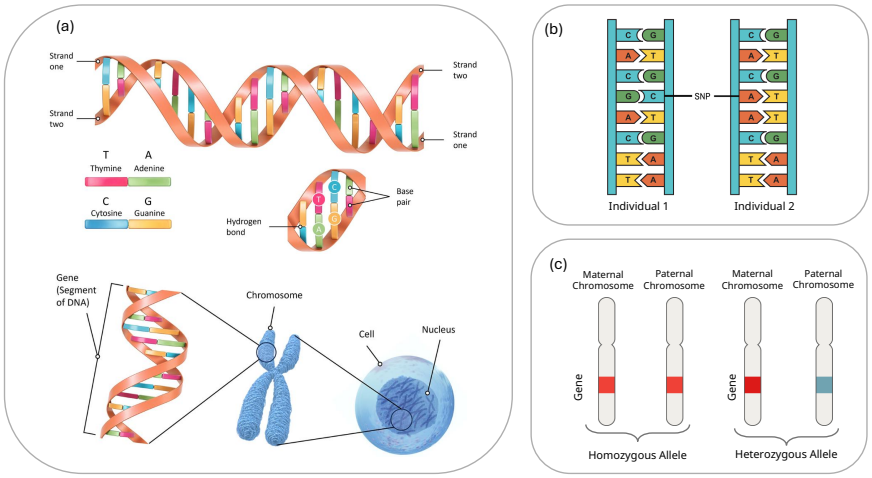

Figure 1.1: (a) Simple schematic illustration of genome organization, showing the hierarchical relationship between the cell nucleus, chromosomes, and the DNA double helix. Nucleotide base pairs (A–T and C–G) are shown in the magnified DNA structure. (b) Illustrative example of a SNP in two individuals (1 and 2). (c) Illustrative example of heterozygous and homozygous alleles of a gene. (Original figure adapted from MicrobeNotes science-resources.co.uk and NIH BioArt)

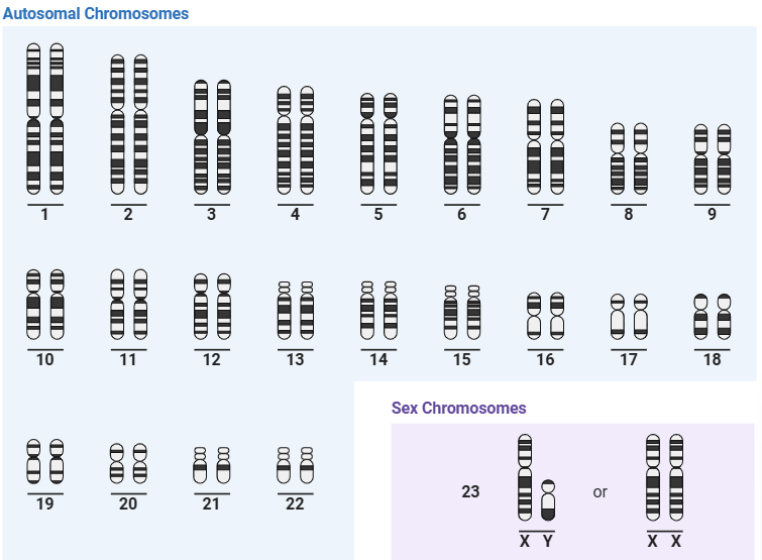

Figure 1.2: Schematic representation of a normal human karyotype. Each chromosome is present in two copies, resulting in a diploid ploidy. Original figure adapted from MicrobeNotes (https://microbenotes.com/chromosomes/).

More precisely, the process of protein synthesis is mediated by the ribonucleic acid (RNA) molecules, which are typically constituted of single-stranded sequences of nucleotides, with uracil (U) replacing thymine (T) as the base pair. Among the various types of RNA, messenger RNA (mRNA) and transfer RNA (tRNA) are involved in these this processes, characterizing two sequential phases of this cellular process, transcription and translation.
During the transcription phase, the nucleotide sequence of a specific gene is copied from DNA into a molecule of mRNA, which carries the genetic information from the nucleus to the cellular environment where protein synthesis occurs. During translation, the mRNA sequence is read in groups of three nucleotides, each specifying a particular amino acid. Amino acids, which are the basic building blocks of proteins, are delivered by tRNA molecules and assembled in the correct order to form a functional protein. It is important to note that different triplet combinations of nucleotides can encode the same aminoacid, a property known as the degeneracy of the genetic code.
Once synthesized, proteins act as the primary functional units of the cell, carrying out a wide range of biological fundamental activities, including helping cells receive and responding to signals from the extracellular environment and from their own internal state. These instructions are transmitted through protein-mediated biochemical interactions, termed signalling pathways, which regulate a wide range of cellular responses, including proliferation, differentiation, and programmed cell death.
Of these processes, the cell cycle is surely one of particular significance, since it ensures a proper cell growth and proliferation. The cell cycle is classically divided into four fundamental phases. During the G1 phase, the cell grows and synthesizes the proteins required for DNA replication. During the subsequent S phase, DNA replication occurs, while the following G2 phase serves as a preparation for cell division. Finally, during the M phase, the cell divides into two daughter cells through mitosis.
Potential genetic errors introduced during DNA replication are normally counteracted by DNA repair mechanisms and by cell cycle checkpoints, which act

as regulatory control points at the transitions between successive phases of the cell cycle. However, in instances where these repair mechanisms are compromised, genomic instability may arise, resulting in cells containing aberrant genetic information. Abnormal variations and modifications carried by the DNA of these cells have the capacity to alter the processes of protein synthesis, thus affecting the functionality of proteins, and modifying the effective communication through signalling pathways.

Among the genetic aberrations that most significantly contribute to the development of cancer, there are those affecting specific types of genes, known as oncogenes and tumour suppressor genes. Under physiological status, oncogenes are responsible of promoting cell growth and proliferation, while tumour suppressor genes inhibit proliferation or induce apoptosis. Alterations in these genes can therefore disrupt normal cellular regulatory mechanisms, contributing to uncontrolled cell growth and progression.

Among the most common alterations that modify the structure of these pivotal genes, somatic copy number aberrations (SCNAs) represent a particularly relevant class. These alterations arise in somatic cells and involve gains or losses of genomic material, affecting individual genes, groups of genes, or larger chromosomal segments. As a consequence, SCNAs disrupt the normal diploid state of the cell, in which each genomic region is typically present in two copies, one on each homologous chromosome. When the number of DNA copies of a given region deviates from this two-copy configuration, the cell reaches a condition known as aneuploidy arises, characterised by an unequal distribution of genomic material across homologous chromosomes.

SCNAs are commonly classified according to whether DNA regions are gained or lost. Amplifications are typically defined as substantial increases in copy number, often involving multiple additional copies of a DNA region, whereas gains generally refer to more moderate increases, such as the acquisition of one or a few extra copies. Conversely, losses are usually defined as partial reductions in copy number, for example when one copy of a diploid region is lost. Deletions, sometimes referred to as deep deletions, are often used to

describe the complete loss of a specific DNA region.
In addition to changes in total copy number, a particular class of SCNAs is represented by copy-neutral loss of heterozygosity (CNLoH), which occurs when heterozygosity is lost in a chromosomal region without a change in the total number of DNA copies. From a schematic viewpoint, a heterozygous genotype (AB) is converted into a homozygous state (AA or BB), while maintaining an overall copy number of two. A final class of aberrations involves changes affecting the entire chromosomal complement of a cell. In these cases, all chromosomes are present in additional copies, leading to polyploid states such as tetraploidy or higher-order ploidy levels. When this genome-wide increase in copy number results from a doubling of the complete genomic content, the condition is referred to as whole-genome duplication (WGD).
Table (1.1) summarizes the common classes of SCNAs.

SCNA class	Total copy number	Allelic configuration (examples)
Neutral	2	AB
Gain	≥ 3	AAB, ABB
Amplification	≥ 4	AAAB, ABBB
Loss	1	A, B
Deletion	0	–
CNLoH	2	AA, BB
WGD	4	AAAA, BBBB, AABB

Table 1.1: Schematic representation of common SCNA classes based on total copy number and allelic configuration. Definitions may vary across studies and biological contexts.

By altering the number of copies of genes, SCNAs can modify the abundance of the corresponding encoded proteins, with direct consequences on normal cellular function. For instance, signalling pathways rely on tightly regulated protein concentrations and interactions, and disruptions in these levels may result in aberrant pathway activity. In this way, SCNAs provide a direct mechanistic link between genomic alterations and the dysregulation of cellular signalling networks observed in cancer.

Accordingly, SCNAs are frequently observed across a wide range of tumour types and contribute to cancer initiation and progression by affecting key oncogenes and tumour suppressor genes. Through gains or losses of genomic material, these alterations can lead to sustained proliferative signalling, impaired cell-cycle control, and altered protein production, ultimately promoting uncontrolled cell division.

Distinct patterns of SCNAs are observed across different cancers, including tumour subtypes arising within the same tissue. As a result, SCNAs serve as important molecular biomarkers with direct implications for diagnosis, prognosis, and therapeutic decision-making, since cancers affecting the same organ may require markedly different treatment strategies depending on their underlying copy number profiles.

Finally, SCNAs contribute to tumour heterogeneity and evolution over time. Due to the uncontrolled proliferation of aberrant cells, cancer cell populations expand and progressively accumulate selective advantages genomic alterations. In particular, copy-neutral loss of heterozygosity (CNLoH) can render cells homozygous for deleterious alleles, such as inactivated tumour suppressor genes, thereby promoting the expansion of specific clones. The accumulation and distribution of such events provide valuable insights into clonal evolution and the genetic mechanisms underlying cancer progression. This evolutionary process leads to the coexistence of multiple genetically distinct subclones competing within the same tumour, a phenomenon referred to as tumour heterogeneity.

Understanding the SCNA profiles of individual subclones, together with the reconstruction of their evolutionary relationships, tehrefore represents a central aspect of cancer research, as it enables the study of tumour progression and cellular dynamics, the identification of potential molecular biomarkers, and the assessment of treatment response.

A comprehensive characterization of SCNAs, and of their impact on signalling pathways, is therefore essential not only for understanding tumour biology, but also for linking genomic alterations to functional changes at the cellular level.

By inducing gains or losses in genes encoding key signalling components, SCNAs can alter protein abundance and pathway activity, thereby contributing to the dysregulation of cellular processes commonly observed in cancer.
For these reasons, the study of SCNAs provides a unifying perspective connecting genomic aberrations and altered signalling behaviour. This biological framework motivates the need for experimental and computational approaches capable of detecting copy number alterations and investigating their downstream functional consequences.

1.2 Detecting Genomic Aberrations: Sequencing Technologies and Copy Number Signals

The detection and characterization of somatic copy number aberrations are essential for the study of tumour genomes and their evolutionary dynamics. From a practical perspective, this requires the identification of genomic regions affected by copy number changes, together with the determination of their genomic boundaries, referred to as breakpoints, and their associated copy number states.
Different levels of copy number analysis can be considered, each providing complementary information relevant to cancer research. Total copy number analysis focuses on identifying genomic segments involved in copy number alterations and estimating their copy number values across the genome. Allele-specific copy number analysis refines this information by quantifying the number of copies of each allele at a given locus, enabling the study of heterozygosity, allele-specific gains or losses, and their functional consequences. A further level of resolution is provided by haplotype-specific analysis, which determines whether copy number alterations affect the maternal or paternal chromosome.
The application of these analytical approaches relies on genome-wide data generated by experimental technologies that have progressively evolved over time and thus provided advances in measurement accuracy, genomic resolution,

and data availability. These technologies yield characteristic genomic signals, which form the basis for SCNA detection and analysis.
This section introduces the main experimental technologies used to detect copy number alterations and describes the genomic signals they generate, which will be used throughout the remainder of this thesis.

1.2.1 The Arise of Sequencing Techniques

The ability to detect SCNAs and other genomic aberrations has undergone a remarkable transformation over the past two decades. The methodologies for discovering and analysing SCNAs have evolved significantly, driven both by advances in biotechnologies and by the development of increasingly sophisticated mathematical models, that have adapted in parallel with the growing diversity, complexity, and scale of genomic data.
Systematic genome-wide analysis of SCNAs became feasible in the late 1990s with the advent of array-based technologies, most notably array Comparative Genomic Hybridization (aCGH). This approach enabled the direct comparison of tumour DNA with matched normal DNA (tissue or blood-derived) by hybridizing both samples to arrays containing probes that target specific genomic loci [13]. Each probe consists of a short synthetic DNA sequence designed to hybridize to a complementary region of the genome, where hybridization refers to the process by which single-stranded nucleic acid sequences bind to complementary sequences through base pairing. Tumour and matched normal DNA samples, differentially labelled with fluorescent dyes, competitively bind to these probes based on sequence complementarity, and the relative fluorescence intensity measured at each probe reflects the balance between tumour and normal DNA, thereby allowing the detection of genomic copy number gains and losses.
The resulting signal is typically expressed as the LogR Ratio, or LogR, defined as the logarithmic ratio of tumour to reference signal intensity, and provides a quantitative measure of relative DNA copy number across the genome. However, early aCGH platforms were largely limited to detecting copy number

alterations over broad genomic regions and did not provide nucleotide-level resolution or information on allelic imbalance. [14, 15, 16].

A major advancement in genome-wide copy number analysis was achieved with the introduction of single nucleotide polymorphism arrays (SNP arrays), which employ probes specifically designed to target known SNPs distributed across the genome. By measuring hybridization intensities at these loci, SNP arrays enable both genotype determination and the indirect inference of copy number variation [17]. Similarly to array-based comparative genomic hybridization (aCGH), SNP arrays provide a LogR signal that reflects total copy number variation across the genome. In addition, SNP arrays offer an allele-specific measure through the B-Allele Frequency (BAF), defined as the proportion of the alternative allele observed at a given SNP.

BAF provides information complementary to LogR, as it captures the relative contribution of the two alleles at heterozygous loci, thereby enabling the detection of allelic imbalance. By jointly analysing LogR and BAF signals, it is possible to distinguish between different classes of copy number alterations that may appear similar when considering total copy number alone [16].

Compared to aCGH, SNP arrays therefore increase the informational content of copy number data by jointly capturing total and allele-specific variation. However, these technologies remain constrained by the uneven genomic distribution of SNPs and by their reliance on selected variants, which limit resolution in certain regions and hinder the detection of novel or very small copy number aberrations.

The next major progression in genome-wide analysis was the emergence of Next-Generation Sequencing (NGS) technologies in the mid-2000s. Unlike SNP arrays, which target predefined loci, NGS provides base-by-base sequencing, allowing the detection of both known and novel variants. Furthermore, NGS can be applied to RNA, enabling the study of gene expression alongside DNA copy number, and thus providing a more comprehensive view of genomic and functional variation.

In this context, gene expression refers to the process by which the information

encoded in a gene is transcribed into RNA molecules, which may act as functional products themselves or serve as templates for protein synthesis. From a data-driven perspective, gene expression is commonly quantified through the abundance of messenger RNA (mRNA) molecules, which serves as a proxy for the activity of a gene within a cell. While mRNA levels do not necessarily correspond directly to protein abundance, they provide a measurable and informative link between genomic alterations and downstream functional effects. The fundamental concept underlying NGS technologies is sequencing, namely the determination of the exact nucleotide sequence of DNA or RNA molecules. NGS performs sequencing through a stepwise approach: the molecules are first fragmented into DNA or RNA fragments, which are then read by the sequencing machine, producing short or long reads. These reads are subsequently aligned to a reference genome, which acts as a coordinate system, allowing each read to be mapped to a specific genomic locus and enabling the reconstruction of the original sequences [18, 19]. Reads are the fundamental units of NGS. Short reads are typically 100–300 base pairs long and are characteristic of widely used sequencing platforms based on sequencing-by-synthesis technologies, such as those developed within the Illumina ecosystem [20]. These reads provide high accuracy but can be challenging to map to repetitive regions of the genome. In contrast, long reads, generated by more recent sequencing technologies, enable more precise mapping across complex genomic regions and allow the direct observation of larger structural variants [21].

Through standard bioinformatic pipelines [22, 23, 24, 25], reads are converted into numerical signals that reflect two different biological aspects. For DNA sequencing (DNA-seq) and RNA sequencing (RNA-seq), reads can be interpreted as numerical signals that reflect different biological aspects. In RNA-seq, gene-level read counts quantify gene expression, reflecting transcriptional activity. In DNA-seq, read counts provide information on copy number and allelic composition. These signals can be further processed into read depth (DP), which indicates the total number of reads covering a specific genomic

position. Bioinformatic pipelines also allow the quantification of allele-specific counts (AD) at particular positions, such as known SNPs. Together, DP and AD represent complementary types of numerical data derived from sequencing. In NGS, another important quantity is coverage, that can be considered at different scales. Local coverage is DP essentialy, while global coverage refers to the average depth across the regions of interest. Depending on whether nucleic acids are extracted from a population of cells or from individual cells, NGS experiments are commonly classified as bulk or single-cell sequencing. In bulk sequencing, DNA and RNA are extracted from a population of cells. In tumour analysis, this population is typically obtained from biopsies and may contain a mixture of normal and cancerous cells. In this setting, the resulting data represent an aggregate signal across all cells, capturing overall genomic or transcriptomic patterns while masking cellular heterogeneity. Consequently, bulk sequencing measurements reflect the average genomic or transcriptomic state of the sample rather than cell-specific variations. NGS bulk sequencing applications are commonly divided into three main classes: Targeted Sequencing (TS), Whole Exome Sequencing (WES), and Whole Genome Sequencing (WGS), depending on the genomic regions of interest. [19, 20]. TS focuses on a selected group of genes, known as panels, which can be designed for specific purposes, such as cardiological or oncological applications. TS offers high coverage at relatively low cost, but it cannot detect aberrations outside the selected gene set.

WES targets the exome, i.e., the protein-coding regions of the genome, which represent approximately 1–2% of the entire DNA. WES provides good sequencing depth for coding regions but may miss larger structural variants or copy number alterations outside exons.

WGS sequences the entire genome, enabling comprehensive detection of all variant types, including non-coding regions and structural variants. However, WGS is more expensive and generates very large datasets.

To overcome bulk sequencing limitations, the focus in genomics and transcriptomics has shifted towards single-cell technologies, commonly referred

to as scRNA-seq and scDNA-seq. Technically, single cell sequencing extends standard NGS workflows by incorporating a cellular isolation step, which enables reads to be assigned to individual cells. From a data perspective, this results in cell-resolved, read-based signals.
While bulk NGS provides an average signal over millions of cells, single-cell approaches reveal cellular heterogeneity, enabling the identification of cell types, cell states, and dynamic processes that are otherwise masked in bulk analyses. This methodology provides a reconstruction of single-cell genomic landscapes, identification of mutations, copy number alterations, and clonal populations. Moreover, these approaches are enabling to resolve SCNA patterns and transcriptional profiles at single cell resolution [26, 27], despite the technical challenges posed by low coverage. Altogether, the field has progressed from hybridization-based, bulk-level measurements to high-resolution, allele-specific, and single-cell SCNA inference. Each technological advance has brought improvements in resolution, interpretability, and biological relevance, while also introducing new analytical and computational challenges. Today, the integration of multi-omic single-cell data and long-read sequencing represents the frontier of SCNA research, promising deeper insights into the complexity of cancer and genomic disorders.

1.2.2 A Class of Genomic Signals: LogR, BAF and gene expression

A more detailed description of data that can be recorded through the techonologies presented in the previous sectiona and that serve as a starting point for somatic copy number analysis is now presented.
The study of SCNAs typically relies on three derived quantities: LogR, BAF, and gene expression. These measures are related to underlying biological phenomena, consisting in total copy number, allelic imbalance, and transcriptional activity, respectively.
While these concepts are well established in the literature, their quantitative

definitions are inherently technology-dependent. Indeed, LogR, BAF, and gene expression are not directly observed biological variables, but are instead inferred from platform-specific signals, such as probe intensities in SNP arrays or read counts in next-generation sequencing data. As a consequence, their computation varies across technologies and even across analytical pipelines within the same technology, reflecting differences in data acquisition, normalization, and modelling assumptions.
Each of these quantities is therefore defined operationally within a given experimental and computational framework; consequently, there is no single, technology-agnostic mathematical definition.
The following paragraph provides a concise overview of commonly adopted computation strategies across the technologies described above.

Log R Ratio and B Allele Frequency: Although SNP array methodologies are based on the same conceptual framework, the resulting signal output may differ due to variability in experimental protocols and normalisation methods. However, all SNP array platforms provide only two quantities that are fundamental for SCNA analysis: the LogR and BAF [28, 1, 29, 30]. LogR measures the total tumor intensity signal relative to the diploid state. Consequently, samples from a healthy population or a matched normal pair are often analyzed together with the tumor sample to facilitate interpretation. In detail, if R^i_A and R^i_B are the reconstructed signal for allele A and allele B for the i-th SNP, $R^i_{obs} = R^i_A + R^i_B$ represents the total signal observed and R^i_{exp} is the expected signal from the normal population, it can be concluded that for each SNP locus i, the LogR is

$$r_i = \log_2 \left(\frac{R^i_{obs}}{R^i_{exp}} \right) \tag{1.1}$$

The LogR signal is expressed on a logarithmic scale, which facilitates the detection and interpretation of copy number alterations by symmetrizing gains

and losses around a common reference. Since LogR values are centered around zero when the observed signal matches the reference, indicating a copy-neutral state, copy number increases and decreases become comparable in magnitude and easier to identify.

Positive LogR values correspond to copy number gains, whereas negative values indicate copy number losses. The magnitude of these deviations depends on several factors, including the underlying copy number, tumour purity, ploidy, and technical normalization, and therefore LogR values are not confined to a fixed numerical range.

While LogR captures the total copy number, it provides no information on allele-specific changes and, consequently, the same copy number event affecting either allele A or allele B produces an identical LogR signal. To capture allele-specific information, the key signal is the B allele frequency, known as BAF. BAF consists in a quantitative metric that indicates the proportion of the allele B in the sample locus. For each SNP position i, the signal is computed as the ratio of the observed signal for allele B to the total signal at that locus, i.e.

$$b_i = \frac{R_B^i}{R_A^i + R_B^i} \tag{1.2}$$

In a normal sample, SNPs can be either homozygous or heterozygous. Homozygous SNPs have two copies of the same allele (either A or B), resulting in a BAF of 0 or 1, respectively. Conversely, heterozygous SNPs, carrying one copy of each allele, exhibit a BAF of 0.5.

In tumor samples, copy number gains or losses at a locus will alter the relative proportion of alleles, leading to deviations in the BAF signal. In the presence of a copy number gain, the BAF of homozygous SNPs remains at 0 or 1, but heterozygous SNPs can exhibit intermediate values between 0.5 and 1 or between 0 and 0.5, reflecting the allelic imbalance introduced by the gain. When a copy number loss occurs, part of a chromosome is lost, leaving most SNPs with fewer copies of one allele. In such cases, BAF values for heterozygous SNPs shift toward the extremes of the allowed range, which may

differ from exactly 0 or 1 in samples with ploidy strongly deviating from 2 (e.g.: whole-genome duplication). In all samples, tumoral and normal, BAF values remain symmetric around 0.5 up to allele labeling.

While SNP arrays generate allele-specific intensity signals from which LogR and BAF are derived, bulk NGS produces short DNA fragments known as reads (see Section 1.2.1). In particular, at each nucleotide locus within the genomic region of interest, the sequencing output provides a DP measure, that is, the number of reads aligned to that position. A related quantity is the read count (RC), which aggregates the number of reads falling within a predefined genomic region in the reference genome [31].

A common working assumption in NGS-based copy number analysis is that sequencing reads provide an approximately uniform representation of the targeted genomic regions. Under this assumption, the mean or median RC observed over a genomic segment is expected to be proportional to its total copy number. However, this proportionality is affected by several sources of technical and biological bias, including mappability issues in repetitive regions, variability in sequencing depth, technological artifacts, including GC content, defined as the proportion of G and C nucleobases in a DNA region, which can influence sequencing uniformity and read coverage. GC content is particularly relevant because genomic regions with high G and C composition are often sequenced less uniformly due to amplification and sequencing biases, which in turn affect read coverage estimates. For this reason, read count signals typically require normalization and bias correction before being used for reliable copy number inference.

Moreover, for known variant sites such as SNPs, it is also possible to quantify the number supporting each specific allele. As previously mentioned, this information is reported as the allele depth (AD), which indicates the number of reads aligned to each possible allele. For semplicity, AD will be referred to the alternate allele depth, that is the total read counts for the B allele [23, 32, 33, 31].

Depending on the type of tool and analysis strategy, different approaches

can be used to calculate LogR. Specifically, when using genomic bins rather than individual nucleotides, the genome is divided into bins of similar size, and the base-2 logarithm of the aggregated read counts within each bin is computed. Conversely, if one is interested in single-nucleotide resolution, the base-2 logarithm of the DP at each position can be used. When a set of normal samples or a paired normal is available, the LogR ratio is commonly defined as the base-2 logarithm of the ratio between the reads in the sample of interest and the reference ones, providing a relative measure of copy number changes [31, 34, 35, 36].

Regarding allele-specific SCNAs, different computational methods implement different strategies. Generally, the BAF signal at specific SNP loci is defined as the ratio between the allele depth and the total read depth. The definition provided here is consistent with that one reported in equation (1.2), with the substitution of signal intensities with aligned read counts. Other methods employ alternative measures of allelic imbalance, such as log-odds ratios (logOR) [36]. This specific measure is employed for heterozygous sites and is described as

$$\text{logOR} = \log_2 \left(\frac{AD}{DP - AD} \right) \tag{1.3}$$

Regardless the computation strategy, SNP arrays or bulk sequencing techniques, the resultant LogR and BAF can be considered as two vectors of different dimensions $\boldsymbol{r} \in \mathbb{R}^N$ and $\boldsymbol{b} \in [0,1]^M$, with $N, M \in \mathbb{N}^*$. For SNP arrays, $N = M$, and they represent the total number of SNPs of the array, while for NGS-based strategies, N is the number of considered nucleotide positions or genomic bins, while M is the number of SNPs under investigation.

Conversely, in scDNA-seq, information analogous to LogR and BAF can be derived using read counts per cell aggregated over genomic windows or informative SNPs, as single-nucleotide coverage is insufficient to reliably estimate these metrics. These metrics can also be calculated using the pseudobulk technique [37]. In this scenario, the LogR or analogous metrics can be represented as matrices

$$\boldsymbol{R} \in \mathbb{R}^{C \times N} \tag{1.4}$$

where C is the total number of cells, while N is the cardinality of the set of genomic bins. As for BAF signal, it can be formulated as a matrix

$$\boldsymbol{B} \in \mathbb{R}^{C \times N} \tag{1.5}$$

whose exact structure may vary depending on the specific analytical technique.

Gene Expression: Gene expression is quantified through RNA-seq or scRNA-seq analysis. Gene expression data is derived from the number of reads that are aligned and mapped to various genes of interest. For each gene, the number of reads mapped to it is counted to obtain the raw counts. However, raw data is rarely used directly in analyses. In most cases, normalisation and scaling steps are performed to correct for factors such as sequencing depth and gene length, as well as other sources of technical variability [38, 39]. While in bulk RNA-seq the expression signal can be considered homogeneous at the sample level, as it represents an average of the expression across all cells present, in the case of scRNA-seq this measure refers to the mRNA detected in each individual cell. In this scenario, gene expression from RNA-seq for an individual can be represented as a vector

$$\boldsymbol{x} = (x_1, ..., x_G) \in \mathbb{R}^G \tag{1.6}$$

where $G \in \mathbb{N}$ represents the total number of genes under analysis, while gene expression from scRNA-seq can be represented as a matrix

$$\mathbf{X} \in \mathbb{R}^{C \times G} \tag{1.7}$$

where C is the total number of cells and G is the total number of genes. The entry $X_{c,g}$ is the gene expression of cell c for gene g. The quantities x_g and $X_{c,g}$ typically represent normalized and transformed expression measures rather than raw read counts.

In bulk RNA-seq and scRNA-seq data, a coherent BAF definition can also be estimated at heterozygous SNP positions, as reads mapping to expressed transcripts provide allele-specific counts that can be used to quantify reference

and alternative alleles [40, 41].

Bulk sequencing vs Single Cell sequencing: In the context of bulk approaches, the measured signal originates from the aggregation of numerous cells, thereby constituting an average of the cell population present within the sample. From a mathematical perspective, bulk measurements can be interpreted as population-level averages of the underlying single-cell signals. This configuration provides the advantage of greater signal stability and high coverage, given that the readings are spread over a large number of cells. However, it should be noted that information on cell heterogeneity is thereby lost. Furthermore, in tumour samples, bulk data generally represent a mixture of malignant and non-malignant cells, necessitating the inference of the tumour fraction, that is the so-called purity, and the reconstruction of the subclonal structure using deconvolution methods. Purity can be defined as the proportion of malignant cells contributing to the observed bulk signal.
In contrast, single-cell technologies preserve information at the single-cell level, allowing direct identification of populations, subpopulations, and clonal structures. However, this granularity comes at the cost of increased technical and sampling noise. As a result, single-cell measurements exhibit higher variance and lower coverage compared to bulk data. In the case of scRNA-seq, the frequent phenomenon of dropout causes many genes to go undetected in single cells, even when biologically expressed, due to low mRNA quantities and the capture limitations of the technology. This results in zero-inflated and overdispersed gene expression distributions. Similarly, in scDNA-seq, coverage is typically low affected by amplification bias and allelic dropout. This necessitates the use of dedicated probabilistic models for copy number inference and clonal reconstruction [41].
To increase the robustness of single-cell data, it is sometimes useful to aggregate similar cells by generating so-called pseudobulks, in which the signal is summed to obtain a more stable representation while maintaining the information at the level of defined cell populations [25, 42, 27]. From a modelling

perspective, pseudobulk representations reduce variance by averaging signals over predefined cell groups while preserving population-level structure.
In conclusion, bulk and single-cell sequencing technologies provide distinct data on the same underlying cellular population. Bulk measurements correspond to population-level averages with reduced variance, while single-cell measurements provide direct but noisy realizations of cell-specific signals. These differences in data structure and noise characteristics naturally motivate the use of different mathematical and statistical models, which will be introduced in the following chapters.

Key bioinformatics concepts and file formats In next-generation sequencing analyses, quantitative signals such as LogR and BAF (for DNA) and gene expression (for RNA) are derived from aligned sequencing reads, typically stored in BAM (Binary Alignment Map) files. BAM files contain reads mapped to a reference genome, along with quality scores and alignment information. The most frequently referenced human genome assemblies are GRCh37 and GRCh38, also known as hg19 and hg38, respectively. The main differences between these assemblies lie in sequence corrections, updated coordinates, and annotations, which allow identification of known genes, SNPs, and other genomic features.
Another important file type is the BED file. BED files define genomic regions of interest in a pre-defined assembly, such as exons, allowing aggregation of read-level signals over these regions. Variant information for each SNP can be stored in VCF (Variant Call Format) files, which record the alleles observed at each position along with additional metadata such as genotype quality and read depth.
Two key bioinformatic concepts that will be useful for understanding the following chapters are pileup and phasing. Reads can be summarized in a pileup, which counts the number of reads supporting each allele at each SNP position, enabling the detection of allelic imbalance. Phasing assigns variants to parental haplotypes, which is important for allele-specific copy number and

expression analyses.

1.3 Introduction to Chemical Reaction Networks: A Proteomic View of Cellular Dynamics

The aberrations occurring within a cell are not static events, but actively influence cellular behavior and contribute to the continuous evolution of cancer. Somatic copy number alterations affect the number of copies of a gene, thereby altering protein production and triggering cascades of downstream molecular events that can profoundly modify cellular function and the response to external stimuli. While SCNAs are measured as static quantities, the cellular processes they affect such as signalling pathways are inherently dynamic. This motivates the use of mathematical models that capture temporal evolution, such as Chemical Reaction Networks (CRNs). CRNs consist in mathematical models that describe the reactions among chemical species, providing a natural framework for modelling signalling pathways as systems of interacting molecular species evolving over time. This section is devote to provide the keys ideas and definitions for analyzing these chemical processed from a biochemical point of view and to present the foundamental hypotheses necessary for the consequent mathematical modelling of CRNs that will be presented in Chapter 4. For futher details the interested reader may refer to the following works: "Mathematical modeling in systems biology: an introduction" and “Modeling and analysis of mass-action kinetics” [44, 43].
Signalling pathways consist in the interactions between specific chemical species, such as proteins, ions, molecules, macromolecules or complex molecules. These interactions can be categorised into diverse types, representing a wide range of processes, including their creation, dissolution or regulation of cellular activities that lead to the production, transport or consumption of the species involved. All the aforementioned interactions are merely chemical reactions, which can be elucidated through the utilisation of diagrams. In these diagrams, a chemical reaction is denoted by an arrow, with the reactants indicated on

the left and the products on the right. As a purely illustrative example, three chemical species will be considered: A, B, and C. In the reaction

$$A \rightarrow B + C,$$

A represents the reactant and the product of the reaction is $B + C$. The '+' symbol between the species is a standard notation to indicate that both species B and C are present in the environment. The main types of biochemical reactions can be summarised as follows:

- conversion, $A \rightarrow B$, where species A is transformed into species B;
- association, $A + B \rightarrow C$, where species A associates with B, forming a new species C with a bond;
- dissociation, $C \rightarrow A + B$, where species C dissociates into species A and B;
- production, $\emptyset \rightarrow A$, where species A is increased at a constant rate and the reactants do not belong to the environment;
- consumption, $A \rightarrow \emptyset$, where species A decomposes at a constant rate and the products do not belong to the environment.

It can be demonstrated that every chemical reaction can be described as a combination of the three elementary processes just defined [45]. For example, a reversible reaction of the type $A \leftrightarrow B$ can be broken down into two conversions: $A \rightarrow B$ and $B \rightarrow A$. The topology of a network describes how species interact across reactions, and networks can be classified as closed or open, with most biochemical CRNs being open. The term 'closed' refers to a set of reactions in which all chemical species participate as both reactants and products, while the term 'open' stands for those where there is an exchange of material with the surrounding environment [46]. To translate CRNs into mathematical models, two fundamental assumptions are commonly made:

i. 'spatial homogeneity': the volume in which the reactions take place is well-mixed. Consequently, the reactants are distributed uniformly throughout the volume, ensuring that the speed of each reaction is independent of its spatial position.

ii. 'continuity': the number of molecules is large enough that their concentrations can be treated as continuous variables, varying smoothly over time. This quantity is called molar concentration and the unit of measurement for this physical quantity is nanomoles (nM).

The 'spatial homogeneity' results in the determination of the reaction rate in the specified volume with unambiguity. Assuming these conditions hold, the 'law of mass action' can be applied, which states that the rate of a chemical reaction is directly proportional to the product of the concentrations of its reactants. Consider again the reaction $A \to B + C$. Let $x_A(t) \in \mathbb{R}_+$ denote the concentration of species A at time $t \in \mathbb{R}_+$ and $k \in \mathbb{R}_+$ the rate constant, which is usually determined experimentally. The law of mass action states that the rate $v(t)$ at which the reaction occurs is

$$v(t) = k \cdot x_A(t) \tag{1.8}$$

and the evolution of species A, B and C concentrations over time is described by the following differential equations

$$\frac{d}{dt}x_A(t) = -kx_A(t) \qquad \frac{d}{dt}x_B(t) = kx_A(t) \qquad \frac{d}{dt}x_C(t) = kx_A(t).$$

for any $t \in \mathbb{R}_+$ The negative sign for x_A reflects its consumption, while the positive signs for x_B and x_C indicate their production. By denoting with $[T]$ and $[C]$ the unit of measurement for time and for concentrations, respecively, the units of k can be consequently defined. This unit depends also on the number of reactants: for a single reactant, k has units of $[T]^{-1}$. When two reactants are present, the reaction constant assumes the dimensions $[C]^{-1}[T]^{-1}$. Lastly, the unit of measurement of k in the reaction $\emptyset \to A$ is designated as $[C][T]^{-1}$ [44].

By providing a quantitative description of how molecular species change over time, CRNs offer a rigorous framework for modelling dynamic cellular processes, also influenced by alternations such as SCNAs. They form the foundation for mathematical models that capture signalling dynamics, and pathway interactions. In summary, this biochemical overview provides the basis for a rigorous description of chemical reaction systems using dynamical systems concepts, which will be addressed in subsequent chapters.

Chapter 2

Towards a Multi-Resolution Optimization Approach for Bulk DNA Somatic Copy Number Aberration Analysis

Somatic copy number aberrations are a hallmark of a wide range of tumour types and have been largely studied using genomic data generated by technologies such as SNP arrays and bulk sequencing. More generally, copy number analysis represents a fundamental source of genomic heterogeneity and has also been investigated in other biological contexts beyond cancer, including cell reprogramming processes, in which cells are derived or induced from a common cell of origin and genomic stability is a critical requirement.
Despite methodological differences, computational approaches to copy number analysis are based on common fundamental steps, among which genome segmentation plays a central role. Given the measured LogR data as input, often together with BAF or LogOR data, the segmentation algorithm aims to identify genomic regions characterised by homogeneous copy number states. Over the years, several segmentation methods have been proposed, each with specific strengths and limitations. A clear understanding of these approaches

is therefore essential for both the correct interpretation of copy number profiles and the development of improved inference methods.
The present work focuses on the segmentation algorithm implemented within ASCAT [1], which is a reference method in the field of allele-specific somatic copy number analysis. This approach relies on the optimization of a penalized loss function, which requires the corresponding regularization parameter to be accurately selected by balancing sensitivity to true copy number changes against robustness to noise. Here a novel extension of the original algorithm to introduce a data-driven parameter selection is pursued.
In this Chapter, Section 1 provides an overview of the main segmentation strategies employed in copy number analysis. Section 2 introduces ASCAT in detail as a reference framework and an application of the method in a cell reprogramming context is then presented. Section 3 describes an extension of the original framework aimed at improving the segmentation step through the introduction of an automatic multi-resolution approach for penalty parameter selection. Finally, the performance of the proposed approach is evaluated on both real and synthetic datasets, and the results are discussed, highlighting their implications and limitations.

2.1 A Change-Point Detection Problem for Segmentation within SCNA analysis

Standard pipelines for detecting somatic copy number aberrations from bulk DNA data typically involve two main steps. The first is the segmentation step, which aims to partition the input data into segments with homogeneous underlying copy number, thereby producing a smoother representation of the original signal affected by noise. The second, commonly referred to as copy number calling, focuses on inferring absolute or allele-specific copy number states along the genome.
Among these two stages, segmentation plays a central role, as it not only focuses on signal denoising but also determines how the genome is partitioned

into regions with homogeneous copy number. Errors introduced at this stage may propagate to downstream analyses, ultimately leading to inaccurate final SCNA calls. For this reason, segmentation represents one of the most critical components of SCNA detection pipelines.

In this section, an overview of the most widely used segmentation methodologies is provided, with the aim of establishing a common mathematical framework for SCNA analysis. Particular attention is devoted to a class of approaches that form the methodological basis for the techniques developed and discussed in the following sections and chapters. To this end, the relationship between the underlying biological problem and its mathematical formulation is first introduced, followed by a description of representative algorithms that offer practical solutions to the segmentation task.

All segmentation approaches rely on the assumption that, in a noise-free genomic signal, regions affected by the same type of somatic copy number aberration exhibit similar signal levels, whereas different copy number states produce distinct signal intensities. In real samples, where genomic signals are typically noisy, the goal of segmentation is therefore to recover the latent piecewise constant signal by reducing noise, while accurately identifying the genomic positions at which abrupt changes occur.

From a mathematical modeling perspective, this problem can be formulated as a change-point detection task on piecewise constant signals. The objective is to partition the observed signal into contiguous intervals such that the underlying signal remains constant within each interval, while simultaneously estimating the locations of the change points, also referred to as breakpoints.

Formally, given a sequence of observations

$$\mathbf{y} = (y_1, ..., y_N) \in \mathbb{R}^N \tag{2.1}$$

the change-point detection problem aims at partitioning the index set $\{1, ..., N\}$ into an unknown number Q of contiguous sets, $I = \{I_1, ..., I_Q\}$, called segments, such that within each I_k the underlying signal remains constant. By defining

the breakpoints as a $(Q+1)$-tuple of increasing locations

$$\{(\tau_0, ..., \tau_Q) \in \{1, ..., N+1\}^{Q+1} | \tau_0 = 1, \tau_Q = N+1, \tau_0 < \tau_1 < \tau_2 < ... < \tau_Q\} \tag{2.2}$$

then each I_k is defined as

$$I_k = \{i \in \mathbb{N} : \tau_{k-1} \leq i \leq \tau_k - 1\}, \quad k = 1, ..., Q \tag{2.3}$$

The observed signal is then modeled as

$$y_i = \sum_{k=1}^{Q} z_k \mathbb{1}_{\{i \in I_k\}} + \epsilon_i, \quad i = 1, \ldots, N, \tag{2.4}$$

where $z_k \in \mathbb{R}$ represents the underlying piece-wise constant signal within the segment I_k. The objective of the change-point detection is then to determine the number of breakpoints Q, their locations $(\tau_0, ..., \tau_{Q+1})$ across the genome, and the segment-level constants $\{z_k\}_{k=1}^{Q}$ across each segment $\{I_k\}_{k=1}^{Q}$.
Typically, the noise vector $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_N)$ is assumed to have independent entries, while its distribution depends on the type of sequencing data [47]. For read counts (RC), a Poisson model is commonly adopted due to the discrete nature of sequencing data. In contrast, LogR and BAF signals are often modeled using a Gaussian distribution, i.e.

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i = 1, ..., N \tag{2.5}$$

This choice is supported by empirical observations of LogR and BAF measurements, since they tend to have approximately symmetric distribution. This assumption is also convenient from a methodological perspective, as it leads to simple and well-established estimation procedures in segmentation algorithms [48]. In particular, under a Gaussian noise model, once a segmentation is fixed, the optimal estimate of the segment-level signal $\hat{z}_k$ corresponds to the average

of the observations within each segment:

$$\hat{z}_k = \frac{\sum_{i \in I_k} y_i}{|I_k|} \tag{2.6}$$

where $|\cdot|$ indicates the length of the segment.

Before proceeding with the overview of segmentation methods in genomics, it is important to clarify the distinction between univariate and multivariate segmentation in the context of SCNA analysis. In total copy number analysis, segmentation is typically performed on a single genomic signal, most commonly the LogR profile. In this setting, the problem is referred to as univariate segmentation, as each genomic position is associated with a single observed values.

In contrast, allele-specific SCNA analysis relies on multiple genomic signals measured at the same loci, most notably the LogR signal and the BAF, or related transformations such as the LogOR, defined in equation (1.3). In this case, segmentation is performed jointly across signals, leading to a bivariate segmentation problem. More generally, several methods extend this framework to a multichannel setting, in which more than two signals are segmented simultaneously.

From a mathematical perspective, the underlying change-point model remains unchanged: the key assumption is that the latent genomic signals are piecewise constant along the genome. However, instead of observing a scalar value at each genomic position, a vector of measurements is observed. Joint segmentation enforces a common set of breakpoint positions across LogR and mBAF, while allowing the two channels to contribute with different strengths to the detection of a change. A breakpoint is retained when the shift is consistently supported by both channels, or when one channel exhibits a clear and well-defined deviation. This selective behaviour enhances robustness and substantially reduces channel-specific false positives.

Formally, each genomic position i is associated with a vector of p observations

$\mathbf{y}_i \in \mathbb{R}^p$ modeled as

$$\mathbf{y}_i = \sum_{k=1}^{Q} \mathbf{z}_k \mathbb{1}_{i \in I_k} + \boldsymbol{\epsilon}_i, \tag{2.7}$$

where $\mathbf{z}_i$ is piecewise constant across genomic positions and $\boldsymbol{\epsilon}_i$ denotes a multivariate noise term. The segmentation task therefore consists in identifying a common set of breakpoints, and thus of intervals $\{I_k\}_{k=1}^{Q}$, shared across all p signals. Consequently, segmentation reduces to the identification of the number, Q, and locations of the breakpoints $\{\tau_0, \ldots, \tau_Q + 1\}$, together with the estimation of the corresponding piecewise constant signal values $\{\mathbf{z}_k\}_{k=1}^{Q}$. The following discussion focuses on common segmentation approaches in the univariate setting, ranging from classical statistical methods to regularization-based techniques. It is worth noting, however, that several extensions have been developed for multivariate frameworks. Moreover, a variety of methodologies for copy number analysis integrate segmentation and copy number calling within a unified model, for example through hidden Markov models [29].

Statistical-based approaches The most widely used statistical approaches for segmentation are based on the original or modified versions of Circular Binary Segmentation (CBS) [49]. CBS has proven highly effective in denoising genomic profiles and accurately delineating regions of amplification or deletion, establishing it as a standard tool in computational genomics. CBS was originally developed for the analysis of aCGH-data and was later extended to other technologies, including SNP arrays and NGS.
In particular, CBS aims to detect statistically significant changes in the mean signal between adjacent genomic regions, under the assumption that the observed data are affected by additive Gaussian noise with zero mean and unit variance. Conceptually, CBS can be viewed as a modification of the binary segmentation procedure introduced by Sen and Srivastava [50], which is based on a likelihood ratio test for detecting a single change-point at an unknown

location. In binary segmentation, this test is applied recursively to the resulting subsegments until no further significant changes are detected. However, a well-known limitation of this approach is its reduced ability to detect short altered segments embedded within much larger regions, since only one change-point is tested at each step.

To overcome this limitation, CBS introduces a circular representation of the data. Specifically, the genomic signal is conceptually spliced at its ends to form a circle, allowing the simultaneous detection of two change-points. Within this circular framework, any pair of genomic positions (i, j) defines an arc, corresponding to the contiguous interval from position $i + 1$ to j along the circle. The CBS then evaluates whether the mean signal on the arc differs significantly from that of its complementary segment using the following test statistic:

$$Z_{ij} = \frac{1}{\sqrt{\frac{1}{j-i} + \frac{1}{n-j+i}}} \left(\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i} \right). \tag{2.8}$$

where $S_i = y_1 + \cdots + y_i$, for $1 \leq i \leq n$, denotes the partial sum of the signal. For each pair (i, j), then Z_{ij} measures the deviation between the mean signals of the two segments. The global test statistic

$$Z_C = max_{1 \leq i < j < n} |Z_{ij}| \tag{2.9}$$

represents the largest observed difference across the genome. A breakpoint is declared if the test statistic Z_C exceeds an appropriate threshold level based on the null distribution i.e., the distribution it would have under the hypothesis of no change. When the y_i can be reasonably modeled as Gaussian, the corresponding critical values can be obtained from the reference distribution used in classical binary segmentation, as originally derived via Monte Carlo simulations [50] or through the tail probability approximation [51]. In the case of non-normal distributed data, the procedure can be generalized by generating a reference distribution using a permutation approach [49]. The procedure is applied recursively to identify all significant genomic changes.

An example of an adaptation of CBS for NGS data is FACETS, which performs

allelic analysis of SCNAs [36]. FACETS implements a bivariate version of CBS to jointly segment LogR and LogOR signals. Specifically, the method combines the CBS framework with a linear combination of two different Z_{ij}, each representing Mann-Whitney test statistic, one for each genomic signal. By relying on Mann-Whitney test statistics, the method no longer requires the data to follow a normal distribution. Despite the modification of the test statistic, the fundamental principle of circular recursive segmentation remains unaltered. While CBS methods are robust and widely used, they may result in over-segmentation or reduced sensitivity to subtle genomic alterations.
To address the limitations of purely statistical segmentation approaches, a broad class of regularization-based methods has been introduced. These methods control model complexity by explicitly penalizing the number of detected changes, thereby discouraging overfitting and improving robustness, particularly in noisy genomic settings.

Regularization-based approaches A wide range of regularization-based approaches for segmentation have been developed, depending on the employed regularization parameter. In this section, a widely used formulation is presented, which provides a clear mathematical interpretation of the segmentation problem and serves as a foundation for the approaches discussed throughout this thesis.
Winkler and Liebscher introduced the Potts Model as a regularization-based approach for solving one-dimensional piecewise constant segmentation problems [52]. In this framework, the Potts Model induces a penalty on the number of discontinuities in the underlying signal, that is, on the locations where abrupt changes occur. Within this setting, such discontinuities can be formally characterized as the elements of the set

$$\{i \sim j : z_i \neq z_j\}, \tag{2.10}$$

where the notation $i \sim j$ denotes consecutive index positions.
Combined with a Gaussian assumption on the observation noise, this modeling choice leads to a regularized optimization problem with an L_2–L_0 structure, where the L_2 term accounts for data fidelity, while the L_0 term reflects the penalty on the number of discontinuities.
This penalty can be expressed through a matrix formulation. In particular, by introducing the first-order difference matrix $\boldsymbol{D} \in \mathbb{Z}^{N-1\times N}$, defined as

$$\boldsymbol{D}_{ij} = \begin{cases} 1 & j = i, \\ -1 & j = i+1, \quad i = 1, \dots, N-1, \\ 0 & \text{otherwise}, \end{cases} \tag{2.11}$$

the cardinality of the set $\{i \sim j : z_i \neq z_j\}$ can be represented by the L_0 norm of the vector $\boldsymbol{Dz}$. Since $\boldsymbol{z}$ is assumed to be piecewise constant, the product $\boldsymbol{Dz}$ encodes differences between consecutive positions, and its non-zero entries correspond exactly to breakpoint locations. In particular, the i-th component of $\boldsymbol{Dz}$ is given by

$$(\boldsymbol{Dz})_i = z_i - z_{i+1}, \quad i = 1, \dots, N-1. \tag{2.12}$$

Within each segment, the signal remains constant and therefore $(\boldsymbol{Dz})_i = 0$, while non-zero values occur only at segment boundaries.
The resulting regularization-based formulation of the Potts Model leads to the optimization problem

$$\arg\min_{z \in \mathbb{R}^N} \ \|\boldsymbol{y} - \boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{Dz}\|_0, \tag{2.13}$$

where $\lambda > 0$ is a regularization parameter controlling the trade-off between data fidelity and segmentation complexity. The L_0 term $\|\boldsymbol{Dz}\|_0$ counts the number of non-zero first-order differences in $\boldsymbol{z}$, corresponding to the number of change points in the signal. Consequently, estimating a piecewise constant signal $\boldsymbol{z}$ is equivalent to identifying both the number of segments and the

partition of the index set into contiguous regions.
Despite its conceptual simplicity, the presence of a non-differentiable and non-convex penalty term makes the optimization problem computationally challenging. In particular, the resulting L_2-L_0 formulation does not admit a closed-form solution, and identifying the global optimum is non-trivial.
Nevertheless, several practical algorithms have been proposed to directly address problem (2.13), relying on dynamic programming strategies or heuristic procedures. These approaches aim at solving the original non-convex formulation without resorting to convex relaxations.
Among these methods, the Piecewise Constant Fitting (PCF) algorithm introduced by Nilsen et al. to avoid the computational intractability of a direct analytical solution and has been widely adopted and extended within commonly used copy number analysis tools, including ASCAT, and ascatNGS [1, 34]. An alternative approach is VEGA, proposed by Morganella et al., that solves additional practical challenges, including the selection of the regularization parameter, by relying on data-driven heuristics [54]. Although VEGA was originally proposed for aCGH data, its formulation can easily be extended to more complex scenarios, such as the multivariate approach in the context of scRNA-seq data.
Within the PCF framework, the optimal signal value within each segment is given by the mean of the observations in that segment, as reported in equation (2.6). As a result, the Potts Model of equation (2.10) can be reformulated as a criterion where the minimization is performed over the total number of segments Q and their corresponding partition $I = \{I_1, \ldots, I_Q\}$:

$$\mathcal{L}(Q, I \mid \boldsymbol{y}, \lambda) = \sum_{k=1}^{Q} \sum_{j \in I_k} (y_j - \bar{y}_{I_k})^2 + \lambda Q, \tag{2.14}$$

where $\bar{y}_{I_k}$ denotes the mean of the observations within segment I_k.
The key idea underlying PCF is that the global segmentation problem can be decomposed into a collection of nested subproblems, which can be efficiently solved using dynamic programming. This decomposition relies on the funda-

mental assumption that, once a breakpoint is fixed, the optimal segmentations on either side of it are mutually independent.

More precisely, let k denote a candidate breakpoint position. If the optimal segmentation up to position $k-1$ is known, the optimal segmentation up to any subsequent position can be constructed by evaluating the cost of introducing a new segment starting at position k. At each step, the algorithm computes the total cost associated with all possible previous breakpoints and selects the configuration by minimizing a derived cost function. The algorithm stops when the last breakpoint is detected as the last possible index N. A more detailed description of the PCF algorithm and its extensions is provided in Appendix A.1.

Although original PCF algorithm was proposed for univariate segmentation of LogR signals, it was subsequently extended to the bivariate setting to enable allele-specific SCNA detection. Specifically, this method, called ASPCF, jointly segments LogR and BAF profiles, enforcing shared breakpoint locations across both channels. This bivariate formulation constitutes the basis of the ASCAT framework [1], and will be discussed in the following section.

Further extensions to the multi-sample setting have also been proposed, such as the algorithm called asmultiPCF, coupling segmentations across samples through shared breakpoint penalties to improve robustness and statistical power [55].

The original PCF, together with its extensions, require the specification of a penalty parameter, which is tipically fixed prior to the execution of the algorithm and strongly influences the resulting segmentation. Although the authors recommend a default parameter, the noise in each sample varies, so higher or lower values can lead to more reliable solutions. Grid searching for the parameter can therefore make segmentation a slow process. It is often necessary for a specialist in the field to visually select the best solution.

VEGA replaces the problem of selecting a single regularization parameter with the exploration of a multi-scheduled segmentation path and an automatic, data-driven stopping rule.

The method is based on the minimization of the functional (2.14) through a region-growing procedure, in which smaller regions are iteratively merged to form larger ones whenever this operation leads to a decrease in a well-defined energy-based functional. More precisely, the variation in energy associated with merging two neighboring segments I_k and I_{k+1} is

$$\mathcal{L}(Q-1, I^{(k)}|\boldsymbol{y}, \lambda) - \mathcal{L}(Q, I|\boldsymbol{y}, \lambda) = \frac{|I_k|\,|I_{k+1}|}{|I_k| + |I_{k+1}|}(\bar{y}_{I_k} - \bar{y}_{I_{k+1}})^2 - \lambda, \qquad (2.15)$$

where $I^{(k)}$ represents the partition where segments I_k and I_{k+1} are merged, i.e. where the set of breakpoints is $\{\tau_0, \ldots, \tau_Q\} \setminus \{\tau_k\}$.
At each step the algorithm choose as the pair of regions to be merged those producing the maximum decrease of the energy functional (2.15). If no further merge is possible for the selected value of λ, the regularization parameter is increased to allow coarser merges, and the procedure is repeated.
In this way, the penalty parameter is progressively updated according to a data-driven schedule, generating a sequence of segmentations at increasing levels of regularization. To determine the highest possible value for λ, a stopping criterion is implemented, based on the overall variability of the data ν, given by the chromosome-wise standard deviation of the signal, together with a tuning parameter $\beta > 0$, which controls the sensitivity of the stopping rule. Heuristically, Morganella et al. suggest $\beta = 0.5$ as a reasonable default choice based on empirical performance. When the updated penalty parameter exceeds the product $\nu \cdot \beta$, the algorithm stops and the determined solution is the final sementation.
For a more thorough exposition of the VEGA algorithm and its extensions, please refer to the Appendix A.2.
To overcome the non-convex formulation of (2.13), a plethora of relaxations have been proposed, including Total Variation, Fused Lasso, or combinations with other metrics, as well as continuous approximations of the L_0 penalty [56, 57, 58, 59, 60]. These approaches typically replace the L_0 penalty on the number of discontinuities with convex surrogates, most commonly an L_1 norm

applied to the first-order differences of the underlying signal.
These strategies have been adopted by several widely used copy number analysis tools, such as CNVkit [35] and their efficacy has been demonstrated in the literature.

In conclusion, statistical-based methods grounded in the CBS context are characterised by their robustness to noise and robust validation. This enables them to discern both global and local alterations without necessitating intricate parameter calibration. However, these methods can be computationally expensive, and may lead to over-segmented solutions.
On the other hand, regularization-based methods for copy number segmentation are powerful and robust, particularly in noisy or complex datasets. They enable automatic detection of breakpoints, and can be adapted using different norms for penalty terms to capture sparsity or smoothness. Regularization also helps prevent overfitting by penalizing overly short segments or excessive breakpoints. However, these methods have a common limitation, which is the choice of the regularization parameter. Its selection strongly affects results, since inappropriate tuning may lead to overfitting. These approaches consequently need the fine tuning of the regularization parameter that aims to realized an optimal trade-off between high specificity, which consist in few false aberrations, and high sensitivity, that leads to few missed aberrations.
The conventional techniques employed for parameter selection, based on predictive risk or more prevalent methods such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [61, 62], frequently prove to be challenging to implement in this context. The primary issue is evidently the non-differentiability of numerous method's objective function and this limitation motivates the development of alternative parameter selection strategies.
In this context, ASCAT is considered as a case study for allele-specific SCNA analysis. Despite having been introduced several years ago, ASCAT remains one of the most widely used tools in this setting and incorporates many of the

methodological challenges discussed in this chapter, including regularization-based segmentation and user-dependent parameter tuning.
The following chapters are therefore devoted to a detailed description of the ASCAT framework and an application in an unconventional field. Subsequently, an integrated methodology for estimating the regularization parameter within ASCAT is introduced, with the aim of reducing user-dependent tuning and improving robustness in allele-specific copy number inference.

2.2 ASCAT as a Baseline Framework for SCNA Analysis

In the field of allele-specific SCNA detection, ASCAT (allele-specific copy number aberrations of tumours) is a well known and widely utilised tool [1]. ASCAT was originally implemented to accurately detect the exact allele-specific copy number of solid tumors in the context of SNP arrays. Its original version aimed to simultaneously estimate tumor ploidy and purity, as the percentage of aberrant cell admixture. Over the past decade, the authors have refined the software to analyze NGS DNA bulk data, thereby expanding its functionality [34]. ASCAT relies on the following assumptions. First, it assumes that all aberrant cells in the tumour share the same copy number alterations, and therefore subclonal events are not explicitly modelled. Second, the method requires data from both tumour and matched germline samples, which are jointly analysed by ASCAT.

Modeling of LogR and BAF Signals In an ideal setting of a pure tumour sample, LogR and BAF signals can be described using the formulations introduced in Chapter 1.2.2. In practice, however, bulk DNA data measurements are typically affected by contamination from normal cells, and ASCAT addresses this issue by explicitly modelling the contribution of normal and tumour cells, while accounting for tumour aneuploidy.

Let $\rho \in (0, 1]$ denote the aberrant cell fraction and ψ_t the tumour ploidy, which can be interpreted as the genome-wide average copy number across all tumoural cells. ASCAT defines the total ploidy as

$$\psi = 2(1 - \rho) + \psi_t \rho, \tag{2.16}$$

which corresponds to the average total number of copies in the sample after accounting for the contamination of normal cells. Let i denote a genomic locus, and let $n_{A,i}$ and $n_{B,i}$ represent the copy numbers of alleles A and B in tumour cells at that locus. The expected LogR signal at locus i, r_i, is then modelled as

$$r_i = \gamma log_2 \left(\frac{\rho(n_{A,i} + n_{B,i}) + 2(1 - \rho)}{\psi} \right) \tag{2.17}$$

with $\gamma > 0$ a scaling factor, typically fixed to $\gamma = 0.5$ in the literature. This formulation reflects the fact that, at each locus, the observed signal arises from a mixture of normal cells contributing two copies and tumour cells contributing $n_{A,i} + n_{B,i}$ copies, weighted by their respective proportions.
Similarly, the expected BAF at locus i is modelled as

$$b_i = \frac{1 - \rho + \rho n_{B,i}}{2(1 - \rho) + \rho(n_{A,i} + n_{B,i})} \tag{2.18}$$

since B allele is expected with a frequence of ρ in tumour cells. It is important to aknowledge that this expression is valid only for loci that are heterozygous in the germline: within a SCNA event, loci that are germline homozygous remain homozygous in the tumour sample, resulting in $b_i = 0$ or $b_i = 1$. Such loci do not carry information about allele-specific copy numbers and are therefore excluded from subsequent analyses.

Joint segmentation of LogR and BAF via ASPCF As previously discussed, ASCAT performs a bivariate extension of the PCF algorithm, referred to as ASPCF. Within this framework, the LogR and BAF signals are segmented jointly, enforcing shared breakpoint locations across both signals. This con-

straint reflects the biological assumption that copy number changes affect both measurements simultaneously and contributes to reducing the occurence of false positive SCNAs.
Prior to segmentation, ASCAT applies a preprocessing step to the input signals. Homozygous germline SNPs are removed, as they do not carry allele-specific information, as discussed in the previous paragraph. Subsequently, heterozygous BAF values are mirrored around 0.5. This transformation yields a symmetric signal and improves numerical stability during segmentation.
ASPCF then seeks for the optimal partitioning independently on predefined genomic regions, corresponding to chromosomes or chromosome arms, by minimizing the following loss function

$$\mathcal{L}(Q, I|\mathbf{r}, \mathbf{b}, \lambda) = \sum_{k=1}^{Q} \sum_{i \in I_k} (r_i - \bar{r}_{I_k})^2 + (b_i - \bar{b}_{I_k})^2 + \lambda Q \tag{2.19}$$

where $\bar{r}_{I_j}$ and $\bar{b}_{I_j}$ are the LogR and BAF means over the segment I_j, respectively. For a fixed value of the penalty parameter λ, the optimization yields a partition of the signals into $Q(\lambda)$ segments, $\hat{I}(\lambda) = \{\hat{I}_1^\lambda, \ldots, \hat{I}_{Q^\lambda}^\lambda\}$, which is shared across both LogR and BAF signals.

Copy Number Calling Step After segmentation, ASCAT performs the copy number call step. The mean of LogR and BAF over the segments, $(\bar{r}_{I_1^\lambda}, ..., \bar{r}_{I_Q^\lambda})$ and $(\bar{b}_{I_1^\lambda}, ..., \bar{b}_{I_Q^\lambda})$, are used to estimate the parameters ρ and ψ_t.
These parameters are estimated via a grid search over biologically plausible ranges, $\{0.10 + q \cdot 10^{-2} : q = 1, 2, \ldots, 95\}$ for ρ and $\{1 + 5q \cdot 10^{-2} : q = 1, 2, \ldots, 440\}$ for ψ_t. For each combination of parameters on the defined grid, allele-specific copy numbers at each genomic locus i are inferred by inverting the LogR and BAF models (Equations (2.17) and (2.18)), yielding

$$\hat{n}_{A,i} = \frac{\rho - 1 + \psi 2^{\frac{r_i}{\gamma}}(1 - b_i)}{\rho}$$

$$\hat{n}_{B,i} = \frac{\rho - 1 + \psi 2^{\frac{r_i}{\gamma}} b_i}{\rho}$$

The optimal solution is selected by minimizing the discrepancy between observed and expected LogR and BAF signals, expressed by

$$d(\rho, \psi_t) = \sum_{i=1}^{N} ((\hat{n}_{A,i} - [\hat{n}_{A,i}])^2 + (\hat{n}_{B,i} - [\hat{n}_{B,i}])^2) \tag{2.20}$$

where $[\cdot]$ is the rounding function. Finally, ASCAT selects the optimal local minimum and rounds copy number estimates to the nearest integer to obtain the final allele-specific copy number profile.

ASCAT has played a pivotal role in the field of allele-specific SCNA detection. Its widespread use, together with successive refinements, has established ASCAT as a reference tool that is still routinely adopted in contemporary research pipelines [63, 64].
Nevertheless, ASCAT also presents well-known limitations. In the context of NGS data, the assumption of clonal copy number alterations has been addressed by more recent algorithms, such as FACETS [36], which explicitly model subclonal populations. Moreover, ASCAT relies on SNP-level measurements, which may provide a less informative signal compared to read-depth binning strategies adopted by tools such as CNVkit [35].
A further critical limitation concerns the calibration of the segmentation parameter, which has a strong impact on the inferred copy number profile and may lead to overfitting if not properly tuned. Addressing this latter issue is the focus of Section 2.3, where a multi-resolution segmentation strategy is introduced with the aim of assessing the robustness of breakpoint detection and copy number inference across different genomic scales.
Prior to the presentation of this novel improvement of ASCAT, an illustrative example of its utilisation in a modern practical application is provided in the subsequent section.

2.2.1 Assessment of Genomic Stability in Reprogrammed Cell Lines Using ASCAT

Copy number analysis is a well-established approach in cancer research, where it has been extensively used to characterise genomic alterations and tumour heterogeneity. Nevertheless, its application has recently been extended to additional biological contexts in which the assessment of genomic stability is of central importance.

One such context is the generation of human cells in vitro through cellular reprogramming approaches, such as induced pluripotent stem cell (iPSC) technologies [65]. These approaches are increasingly investigated as alternatives to traditional animal models, which may fail to fully recapitulate human-specific traits due to interspecies differences. As these technologies enable the generation of new cell populations derived from a single cell line, genomic alterations, such as gains or deletions of chromosomal regions, may be introduced during reprogramming, making the evaluation of copy number changes a relevant data quality control step.

In the work "Optimization of Transcription Factor-Driven Neuronal Differentiation from Human Induced Pluripotent Stem Cells for Disease Modelling and Drug Screening", an established cellular reprogramming protocol was optimised and standardised to induce human fibroblasts, which are terminally differentiated somatic cells that are easily accessible and commonly used as starting material for reprogramming, to generate neuronal populations in vitro [66]. This approach creates new opportunities to study brain function and disease while emphasising the importance of ensuring genomic integrity in reprogrammed cells.

As part of the study, genomic stability was assessed through allele-specific copy number analysis using the ASCAT framework to compare reprogrammed clones with their parental fibroblast lines. Although ASCAT was originally developed for oncological applications, where tumour samples are analysed relative to matched normal controls, the framework can be naturally adapted to the cellular reprogramming setting. In this study, iPSC clones were treated

analogously to tumour samples, while the corresponding fibroblasts were used as normal references. This strategy enabled the inference of allele-specific copy number profiles, together with estimates of ploidy and cellularity for each sample.

Copy number assessment was performed using a high-density SNP array, with around 560000 probes, providing LogR and BAF measurements at each probe. The choice of this high-level resolution SNP array prevented the selection of genetic alterations not detectable by commonly used methods such as karyotyping.

Following standard quality control procedures, only SNPs mapping to autosomal chromosomes were retained to ensure consistency and interpretability of the allele-specific analysis. SNPs with missing LogR or BAF values were excluded.

Since for each parental fibroblast line, multiple iPSC clone-derived cell lines were available, segmentation was performed using the asmultiPCF algorithm by jointly analysing all samples derived from the same fibroblast line. This approach facilitates the identification of shared breakpoints, which become more robustly detectable within a multi-sample framework.

Penalty values in the range of 70 to 150 were explored, and the optimal segmentation was selected by visual inspection performed by a medical doctor.

The copy number analysis revealed no newly acquired genomic aberrations in the iPSC lines relative to their parental fibroblasts, and copy number profiles showed a high concordance across clones, indicating that the reprogramming protocol preserved genomic stability. These results demonstrate that the derived iPSCs satisfy a critical quality criterion for their use in further studies and research.

2.3 An automatic multi-resolution approach for ASCAT segmentation step

As introduced in the previous section, ASPCF solves the segmentation step using a regularized-based method, whose loss function is defined in Equation (2.19). Within this context, the regularization parameter has the role of penalizing the introduction of additional breakpoints, and its selection involves certain inherent difficulties. First, the choice of this parameter strongly depends on the noise level and data characteristics. Therefore, the usage of a recommended default parameter, that in the ASCAT GitHub repository [67] is empirically fixed equal to 70, may not be appropriate across different datasets and experimental conditions. In addition, since genomic regions may differ in the number of heterozygous SNPs, a single global parameter could lead to suboptimal partitions of the genome. Indeed, the usage of a unique value for the regularization parameter implies homogeneity in signal quality and SNP density, a hypothesis that is rarely satisfied in practice, and regions with sparse SNP coverage may require stronger regularization than regions with dense coverage. These limitations motivate the introduction of an improved version of ASCAT, based on the following main two novelties:

(i) a different level of regularization is selected for predefined regions, for example one for each chromosome, or each chrosome bands;

(ii) a data-driven criterion is applied for determining the optimal regularization parameters.

By allowing the degree of regularization to vary across regions, the proposed strategy aims to enhance segmentation performance, particularly in high-noise settings, and to improve the resolution at which SCNAs can be detected. Moreover, since less experienced users often favour segmentations characterized by a limited number of inferred segments, which may result in simplified solutions with loss of biologically relevant information [53], this approach can potentially guide them in the more detailed characterization of SCNAs.

The modification proposed in (i) is straightforward to implement. First, recall that $\{1, ..., N\}$ denotes the index set of genomic loci. Let $\{C_i\}_{i=1}^{C}$ denote the predefined finite set of pairwise disjoint genomic regions, corresponding to chromosomes or chromosome bands, with $C_i \subset \{1, ..., N\}$. The ASPCF algorithm is then applied separately to each region C_i by restricting the input signals $\mathbf{r}$ and $\mathbf{b}$ to the set of indices in C_i, for each candidate penalty parameter in a predefined set $\Lambda \subset \mathbb{R}_+$. For each $\lambda \in \Lambda$, the number of segments $Q(\lambda)$ and genome partition $\hat{I}(\lambda)$ are then determined and stored.

The modification proposed in point (ii) aims to select the optimal $\lambda \in \Lambda$ for each $C_i \in \{C_1, ..., C_C\}$ with a data-driven approach. From a mathematical perspective, this can be viewed as model selection problem, where different values of the regularization parameter correspond to diverse segmentation models, characterized by varying numbers of segments and breakpoints. The goal is therefore to select the one achieving an optimal trade-off between goodness of fit and model complexity, balancing sensitivity and specificity in breakpoint detection.

A number of criteria for model selection exists in the literature, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [61, 62]. However, it has been shown that these standard criteria are generally not well suited for change-point detection and copy number segmentation problems [68, 53]. Indeed, AIC and BIC rely on regular asymptotic assumptions that fail in change-point models, where breakpoint locations are discrete parameters and model dimensionality increases with the number of segments. Moreover, BIC requires the differentiability of the likelihood function, a hypothesis that does not hold in piecewise change-point detection problems. Finally, empirical studies have shown that both AIC and BIC tend to select overly complex models in this context, resulting in an overestimation of the number of segments [53].

To address these limitations, a modified version of the Bayesian Information Criterion (mBIC) developed for change-point detection models has been proposed by Zhang and Siegmund [69]. The authors introduced mBIC in the

context of univariate change-point detection for aCGH data, but since the criterion is formulated under the assumption of Gaussian-distributed observations, it can be generalized to SNP array data and other sequencing platforms data. The mBIC is actually an asymptotic approximation of the Bayes Factor, which quantifies the evidence provided by the data in favour of a genome segmentation with Q segments relative to simpler alternatives, while penalizing model complexity. A technical definition of the mBIC criterion is provided in Appendix A.4, while its application and adaptation to the ASPCF framework is discussed in the following.

In the original framework proposed by Zhang and Siegmund, a set of admissible numbers of segments $\{1, \dots, Q_{\max}\} \subset \mathbb{N}$ needs to be defined a priori, and the corresponding partitions computed through a faster and more accurate version of the CBS algorithm [49]. Subsequently, the mBIC term is computed for each expreimental design, the optimal number of segments $\hat{Q}$ is defined as the one that maximizes the mBIC criterion, i.e.:

$$\hat{Q} = \text{argmax}_{Q \in \{1,\dots,Q_{\max}\}} \text{mBIC}(Q). \tag{2.21}$$

The direct application of this framework is not straightforward in the context of ASPCF, as the number of segments is not an explicit parameter of the model and the segmentation setting is bivariate rather than univariate. The adaptation of the mBIC to ASPCF thus requires two adjustments.

First, as the segmentation is obtained as the solution of a regularization-based approach, the number of segments implicitly depends on the regularization parameter. As a result, the mBIC criterion can be viewed as a function of this parameter, and the optimal model is defined as the one corresponding to the penalty value that induces the segmentation maximizing the mBIC criterion. Specifically, the selection criterion of the optimal model becomes

$$\hat{\lambda} = \text{argmax}_{\lambda \in \Lambda} \text{mBIC}(Q(\lambda)). \tag{2.22}$$

Second, when applied within the ASPCF framework, the mBIC approach needs to be extended to account for multiple genomic signals within the same sample, namely LogR and BAF. Following the strategy proposed by Picard et al., who extended the mBIC criterion to jointly segment multiple LogR signals across samples [70], a tailored version of mBIC is introduced here.
First of all, LogR and mBAF are concatenated into a single vector of observations $\boldsymbol{y} \in \mathbb{R}^{2N}$, i.e

$$\boldsymbol{y} = (y_1, ..., y_{2N}) = (r_1, ..., r_N, b_1, ..., b_N) \tag{2.23}$$

Since LogR and mBAF are modeled as Gaussian variables and are assumed to be independent, the concatenated vector $\boldsymbol{y}$ is also modeled component-wise as Gaussian. Moreover, following the ASPCF formulation, a common noise variance is assumed for the two channels within each segment. Under these assumptions, each element of the concatenated vector satisfies

$$y_i \sim \mathcal{N}(\mu_j, \sigma^2), \quad j = \theta_{k-1}, ..., \theta_k, \quad k = 1, ..., 2Q \quad \forall i = 1, ..., N \tag{2.24}$$

where μ_j denotes the mean level of the segment to which locus i belongs. Within this formulation, once the joint segmentation of LogR and mBAF is obtained, the corresponding segmentation of the concatenated vector follows automatically by translating the breakpoints of the two channels into the indexing of $\boldsymbol{y}$ into a new set of breakpoints $\{\theta_k : k = 1, ..., 2Q\}$. The only removable change-point is the artificial junction at index N, where the two signals are concatenated. This point does not correspond to a biological breakpoint and is therefore not counted as part of the segmentation.
The original definition of the mBIC, reported in Equation A.71, can therefore be applied directly to y. A more detailed discussion of the independence assumption and of the equal-variance hypothesis is provided in the Discussion section.
Incorporating the two proposed adjustments into the original definition of the mBIC, reported in Equation A.71, the criterion adapted to the ASPCF

framework takes the form

$$\begin{aligned}\text{mBIC}(\lambda) =& \left(\frac{2N - 2Q(\lambda) + 1}{2}\right) \log\left[1 + \frac{SS_{bg}(\mathbf{I}(\lambda))}{SS_{wg}(\mathbf{I}(\lambda))}\right] + \log\left[\frac{\Gamma\left(\frac{2N-2Q(\lambda)+1}{2}\right)}{\Gamma\left(\frac{2N+1}{2}\right)}\right] \\ &+ Q(\lambda)\log(SS_{all}) - \frac{1}{2}\sum_{j=1}^{2Q(\lambda)} \log|I_j| + \left(\frac{1}{2} - 2(Q(\lambda) - 1)\right)\log(2N)\end{aligned} \tag{2.25}$$

where N is the total number of SNP positions, $Q(\lambda)$ denotes the number of inferred segments on each individual signal, and $I(\lambda) = I_1, \ldots, I_{2Q(\lambda)}$ is the resulting partition of the concatenated signal. The quantities SS_{all}, SS_{bg} and SS_{wg} denote the total, between-group and within-group sums of squares, respectively, and are defined as follows. Letting $\mathbf{y} = (\mathbf{r}, \mathbf{b})$ denote the concatenation of the LogR and BAF signals,

$$\bar{y} = \frac{1}{2N}\sum_{i=1}^{2N} y_i \tag{2.26}$$

the global mean and

$$\bar{y}_j = \frac{1}{|I_j|}\sum_{i \in I_j} y_i, \quad j = 1, \ldots, 2Q(\lambda), \tag{2.27}$$

the segment-wise mean, then the aforementioned quantities quantities are computed as

$$SS_{all} = \sum_{i=1}^{2N}(y_i - \bar{y})^2 \tag{2.28}$$

$$SS_{bg} = \sum_{i=1}^{2Q(\lambda)-1}(\bar{y}_i - \bar{y})^2 \tag{2.29}$$

$$SS_{wg} = SS_{all} - SS_{bg} \tag{2.30}$$

The two proposed modifications, consisting in allowing region-specific regu-

larization and selecting the penalty parameter via the adapted mBIC criterion, naturally lead to the following extension of the original ASPCF segmentation procedure in ASCAT, presented in Algorithm 1.

For each genomic region $C_i \in \{C_1, ..., C_C\}$, a set Λ of candidate penalty parameters is considered. For each $\lambda \in \Lambda$, ASPCF is performed to segment LogR and BAF within the region under consideration, producing a number of segments $Q(\lambda)$ and a partition $I(\lambda)$.

The corresponding mBIC score is then computed, and the penalty value λ that maximizes the mBIC is selected.

The resulting segmentations from all regions are combined to obtain the final genome-wide segmentation. This procedure ensures that the degree of regularization is adapted to local signal characteristics.

Algorithm 1 Region-wise mBIC-based selection of ASPCF penalty

Require: LogR signal $\boldsymbol{r}$, BAF signal $\boldsymbol{b}$, regions $\{C_i\}_{i=1}^{C}$, penalty grid Λ
Ensure: Segmented signals $\boldsymbol{r}_{\text{optimal}}, \boldsymbol{b}_{\text{optimal}}$

for each region C_i **do**
  Restrict **r** and **b** to region C_i
  for each $\lambda \in \Lambda$ **do**
    Compute ASPCF segmentation with penalty λ
    Compute $Q(\lambda)$ and mBIC(λ)
  end for
  Select $\hat{\lambda} = \arg\max_{\lambda} \text{mBIC}(\lambda)$
  Store segmentation corresponding to $\hat{\lambda}$ in $\mathbf{r}_{\text{optimal}} = []$, $\mathbf{b}_{\text{optimal}} = []$
end for

The grid of candidate penalty parameters Λ is inherited from the ASCAT implementation, where the minimum value is 35 and the maximum 140 (GitHub Repository: `https://github.com/VanLoo-lab/ascat/blob/master/ASCAT/`, lines 38-39, code *"ascat.aspcf.R"*, v3.2 as a fixed reference). In this work, the grid $\Lambda = \{35, 40, \ldots, 140\}$ is recommended, although it is acknowledged that future users may wish to modify these values depending on their dataset. This range is thus consistent with values commonly used in ASCAT-based analyses and provides a suitable trade-off between computational efficiency and

resolution in model selection. In practice, ASPCF segmentations change only at a finite number of critical values of the penalty parameter, in order to capture all possible relevant segmentation regimes. This two step algorithm aims thus at selecting the optimal penalty parameter from a finite and predefined grid, and the resulting solution is fully deterministic and reproducible given the same input data and parameter grid. From a computational perspective, the proposed procedure requires solving the ASPCF segmentation problem for each candidate value of the penalty parameter and for each genomic region, resulting into a computationally feasible approach for genome-wide analyses, as it will be reported in the following section.

2.3.1 Preliminary results

In this section some preliminary results are shown, aimed to evaluate whether the proposed workflow achieves performances comparable to ASCAT in low-noise scenarios, while providing improved segmentation in challenging settings characterized by high noise levels. The comparison was carried out using the default ASPCF parameter setting, as this represents the standard operating configuration and provides a fair assessment of performance. The results presented here provide initial evidence supporting the potential of the suggested strategy and help identify directions for further investigation. Specifically, two complementary experiments were considered: one based on real data by high noise levels to assess practical performance, and one based on synthetic data with known ground-truth segmentation, enabling a quantitative evaluation of breakpoint position and and copy number estimation accuracy.

Illustrative comparison between the multi-resolution approach and ASCAT on real data The proposed multi-resolution approach was applied to real sequencing data from the TCGA Research Network (`https://www.cancer.gov/tcga`), which represents a large-scale consortium providing multi-omics data across a wide range of cancer types. Specifically, the analysis focused on a single sample of the Uterine Corpus Endometrial Carcinoma (UCEC) cohort,

due to the high prevalence of complex somatic copy number alterations in this tumor type.

Three complementary analyses were performed to evaluate the proposed method. First, WGS data from this sample were analyzed to verify whether the multi-resolution approach achieves comparable performance to the standard ASCAT implementation in high-resolution sequencing data. Second, the corresponding WES data were analyzed to assess in a setting that is more commonly used in practice, providing an additional perspective on robustness. Finally, the WES-derived segmentation profiles were compared to WGS profiles obtained using the standard ASPCF procedure to qualitatively assess segmentation performance in challenging, high-noise regions. WGS data were used as an approximate reference, as they cannot be considered ground truth for WES due to inherent differences in SNP positions and measured signals between the two technologies.

To evaluate this analysis, BAM files corresponding to tumor and matched normal tissue were downloaded from the Genomic Data Commons Data Portal [71]. The files were then preprocessed to extract LogR and BAF values at predefined SNP positions, provided by the 1000 Genomes Project [12]. Preprocessing was performed using functions available in the ASCAT R package, including allele counting at the specified loci and GC-content correction with the ASCAT reference files. For WES data, the SNP positions were further restricted to the target regions defined in the corresponding BED file.

Following preprocessing, both the standard ASPCF and the proposed multi-resolution approach were applied to the WGS and WES datasets to obtain segmentations, followed by the corresponding copy number calls. The resulting segmentations and allele-specific copy number profiles are shown in Fig. 2.1 and Fig. 2.2. Overall, the multi-resolution approach produced results highly consistent with the standard method, with segmentations and copy number calls showing strong agreement across both datasets. For the multi-resolution analysis, regional specific penalty parameters are reported in Table 2.1 and Table 2.2.

A more detailed comparison shown in Fig. 2.3 reveals that the multi-resolution approach generates segmentations of both LogR and BAF signals that are more closely aligned with the reference WGS segmentation than those obtained using a fixed global penalty parameter. The differences between the two approaches are most pronounced in genomic regions with multiple, closely spaced copy number alterations, where finer segmentation is required. An example of such a region is chromosome 11, where the multi-resolution method identifies a larger number of biologically plausible segments in the LogR signal, while maintaining consistency with the corresponding WGS profile. By allowing the regularization parameter to vary locally, the method captures subtle, local copy number changes while avoiding excessive fragmentation in low-noise regions. In order to assess data quality, two metrics relevant to SCNA analysis were computed from LogR obtained from WES and WGS data, prior to the analysis. The Median Absolute Pairwise Difference (MAPD), which measures the technical noise in the signal between adjacent genomic regions, was approximately 0.2, indicating high-quality data, as values below 0.5 are generally related to acceptable data [72]. The median autocorrelation scanning profile (ASP), which evaluates the dependence of measurement error between contiguous probes, was equal to 0.3. Since false positives in segmentation increase when the median ASP exceeds 0.4, this value suggests that the sample quality is acceptable for reliable SCNA detection [73].

Chromosomal region	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Chromosome	chr1	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr16	chr17	chr18	chr18	chr19	chr20	chr21	chr22	chrX
Penalty	100	35	110	40	35	75	35	70	90	35	45	35	35	35	35	65	35	35	35	100	35	40	65	35	35	35	45

Table 2.1: Penalty values per chromosomal region inferred by the proposed procedure for the WGS real sample

Chromosomal region	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Chromosome	chr1	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr7	chr8	chr9	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr16	chr17	chr18	chr18	chr19	chr20	chr21	chr22	chrX	chrX
Penalty	130	35	55	80	130	75	85	35	35	35	35	40	65	50	70	35	65	35	45	35	105	35	35	135	60	35	75	80	35

Table 2.2: Penalty values per chromosomal region inferred by the proposed procedure for the WES real sample

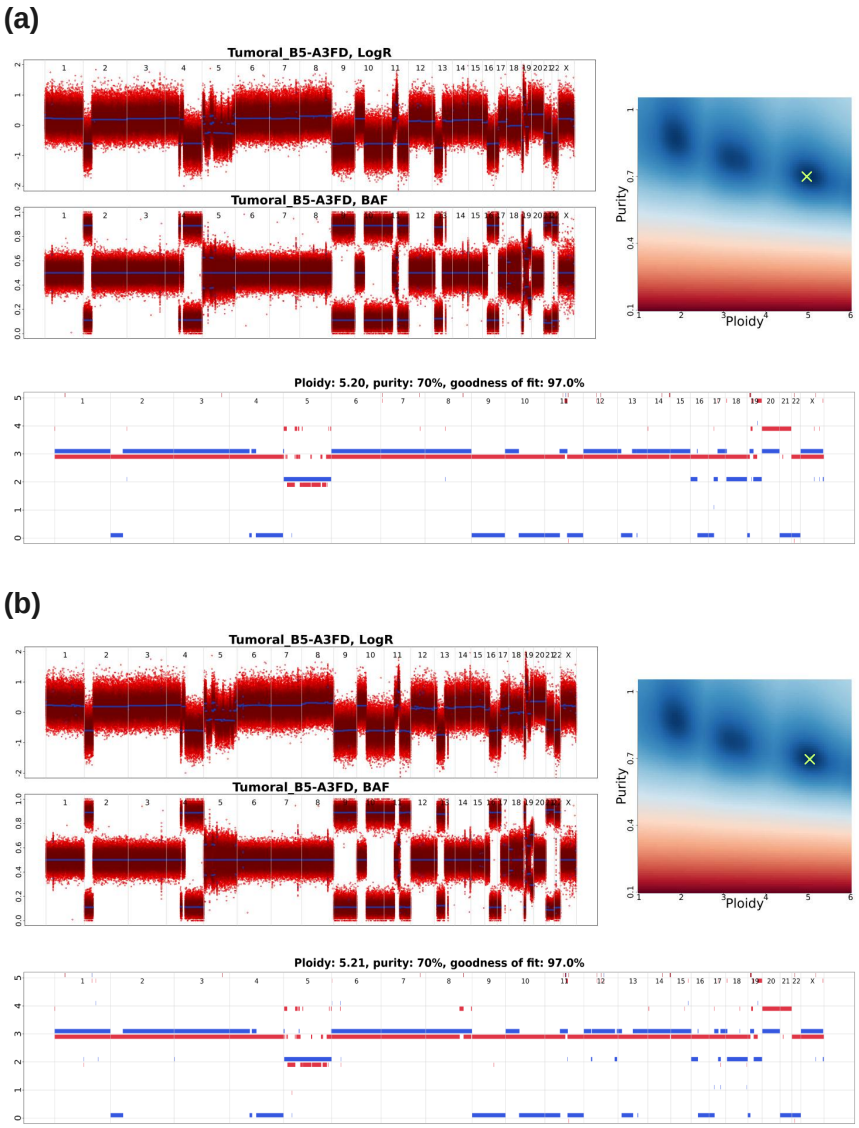

Figure 2.1: Comparison between the results obtained by ASCAT with penalty parameter equal to 70 (a) and the proposed pipeline (b) on WGS data. For both panel (a) and (b) the upper left plot shows in blue the segmentation of Log R (top) and BAF (bottom), on top of the recorded data (red points). The upper right heatmap shows the best combination of purity and ploidy (green cross); the lower plot show the allele-specific copy number computed using the tumor purity and ploidy combination. In these plots, blue lines represent the minor allele (i.e. lowest copy numer) and red lines represent the major allele (i.e. highest copy number).

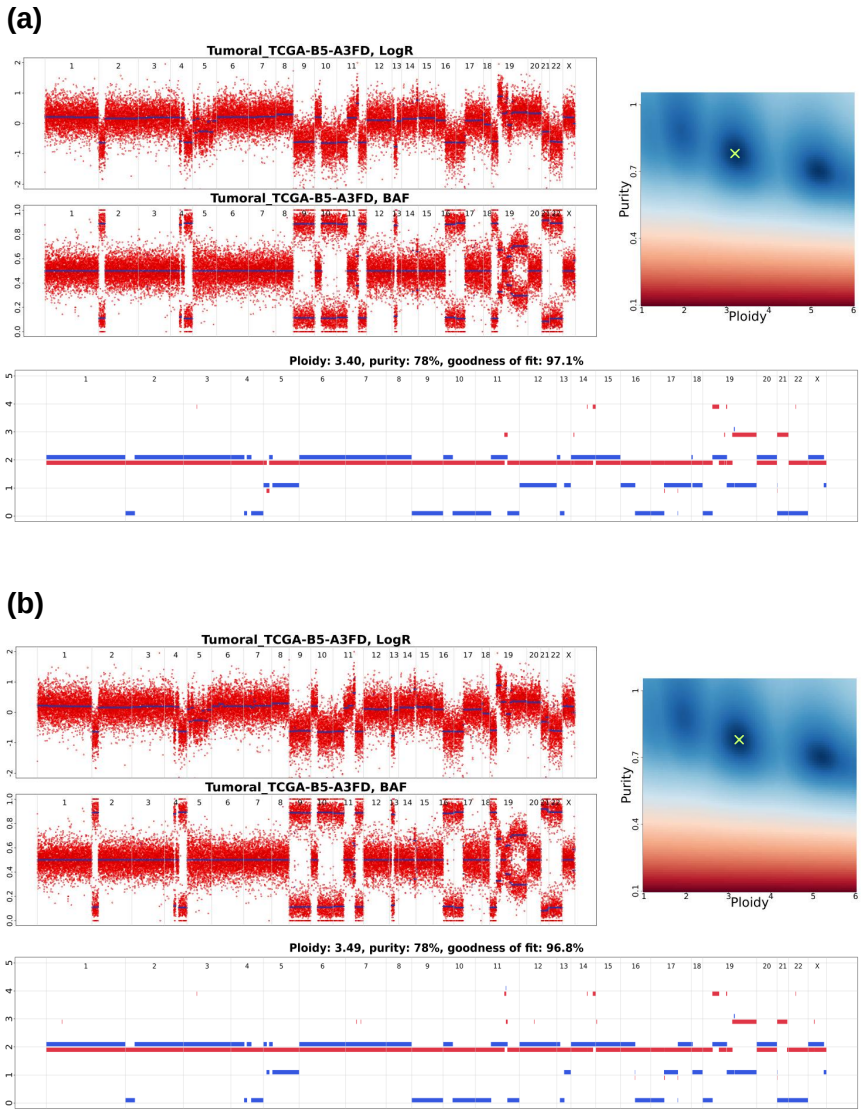

Figure 2.2: Comparison between the results obtained by ASCAT with penalty parameter equal to 70 (a) and the proposed pipeline (b) on WES data. For both panel (a) and (b) the upper left plot shows in blue the segmentation of Log R (top) and BAF (bottom), on top of the recorded data (red points). The upper right heatmap shows the best combination of purity and ploidy (green cross); the lower plot show the allele-specific copy number computed using the tumor purity and ploidy combination. In these plots, blue lines represent the minor allele (i.e. lowest copy numer) and red lines represent the major allele (i.e. highest copy number).

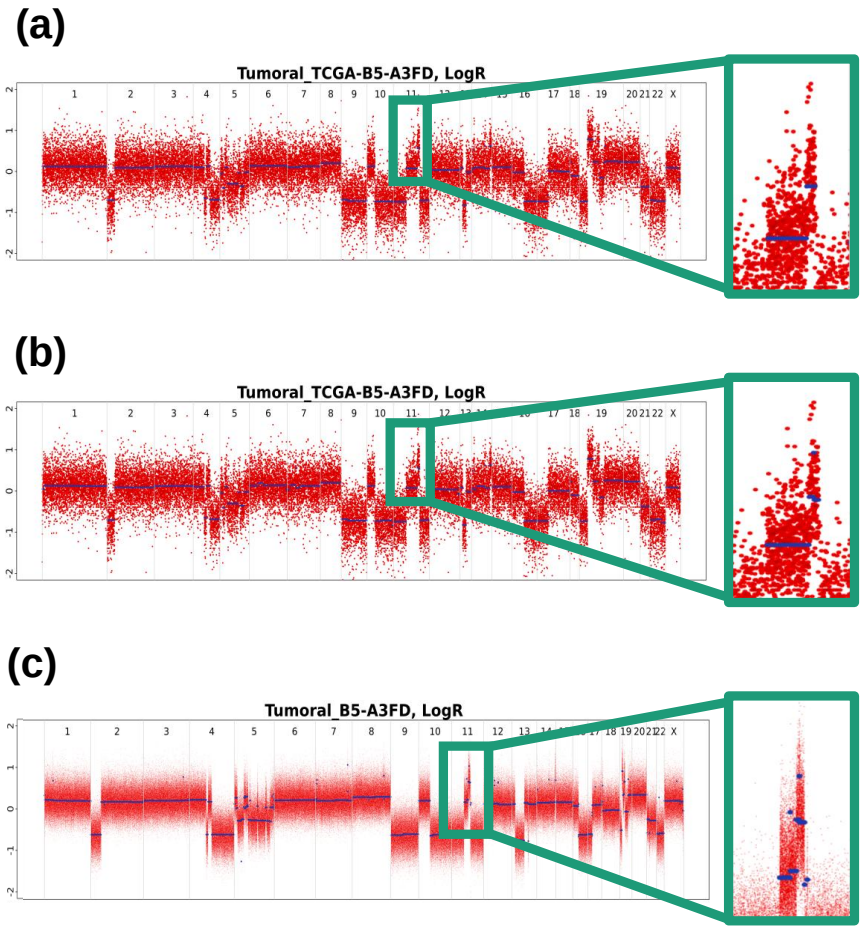

Figure 2.3: Comparison of segmentation obtained by ASCAT with penalty parameter equal to 70 (a) and our pipeline (b) on WES data and the one obtained by ASCAT with penalty parameter equal to 70 (c) on the corresponding WGS data. As illustrated in the green panel, the segmented version of the chromosome 11 data has been magnified.

Synthetic data generation and performance of the proposed method To enable a controlled and quantitative evaluation of the method, a synthetic dataset in which the ground truth is known by construction, with predefined SCNA events and fully specified allele-specific copy numbers, has been generated. Indeed, synthetic data provides a fully known reference against which the accuracy and robustness of the proposed method could be assessed. However, generating synthetic sequencing data with these characteristics is non-trivial, as no standard approach currently exists in the literature.

In recent years, several tools have been developed to facilitate the benchmarking of computational methods for the accurate interpretation of somatic genomic profiles, but only a few are specifically tailored for copy number analysis. For example, Bamgineer [74] provides a framework for introducing user-defined haplotype-phased allele specific copy number events into an existing BAM, but the software has not been updated to recent standards. Other tools, such as SECNVs [75], can generate newer BAM files with novel copy number aberrations, but is not suitable for benchmarking methods such as ASCAT or its extensions, as it does not incorporate SNP-level data. These limitations motivated the choice of Synggen [76], a novel computational tool designed to generate large-scale realistic and heterogeneous synthetic NGS datasets, simulating both normal and tumoral samples. In particular, Synggen was developed for the simulation of cancer sequencing data characterized by aberrations such as point mutations and allele-specific copy number aberrations in a realistic and flexible manner.

From a computational viewpoint, Synggen is divided into two modules. In the first module, starting from a panel of normal sequencing data (BAM format), the tool constructs statistical models that capture the distinctive characteristics of the data inherent in the panel, such as coverage distribution, base quality profiles, and sequencing error frequencies. Within the second module, these models are combined with user-specified genomic features, to generate normal or tumoral synthetic sequencing data with a specified total number of sequencing reads. During this module, somatic allele-specific SCNA profiles can be

introduced, by defining the chromosomes, the breakpoint positions and the clonality of each aberration at which they occur. The tool generates data that is compatible with ASCAT, as it allows for the direct incorporation of a list of germline-phased SNPs, which can be obtained with a pileup of normal input BAMs. In addition, the simulation framework allows for the specification of tumor purity, thereby enabling the generation of different datasets characterized by the same genomic features. Synggen outputs can then be processed through dedicated pipelines to generate associated BAM files, which can be easily used by various downstream analysis tools, including ASCAT and the proposed extension, to generate LogR and BAF data. Moreover, the generated normal and matched tumor BAMs are associated to the reference files of the input data, such as the BED files.

A BAM file associated to a normal WES sample from the TCGA Breast Invasive Carcinoma (BRCA) cohort was selected as the reference input for the generation of a WES synthetic data. This choice was motivated by the high sequencing quality of the available normal sample, with a coverage of $\sim$ 100x, and the associated synthetic normal and tumor sample were generated by simulating 100 million reads. The introduced germline SNPs corresponded to the one of the original sample. A total of 11 SCNAs of different type and width were selected to be introduced in the tumoral synthetic sample. These aberrations fall in the original WES target regions defined by the BED focal copy number aberrations with a span of less than 50000 bp were also taken into account. To avoid subclonality events that are not appropriate for ASCAT analysis, all simulated aberrations were assumed to be fully clonal, i.e. the fraction of tumor cells harboring the aberration is equal to 1. These aberrations are reported in Table 2.3.

The resulting tumoral sample quality was high, given a MAPD associated to the LogR signal equal to approximately 0.04, denoting a profile with low-noise. The LogR and BAF measurements of the tumoral synthetic sample were generated following the ASCAT preprocessing pipeline described in the previous paragraph, and the resulting profiles are illustrated in Figure 2.4.

These measurements demonstrate a low noise level, thus clearly highlighting all clonal copy number events. The implementation of both the proposed approach and standard ASCAT was carried out, and the resulting plots are presented in Figure 2.5. The table with the selected penalty parameters are reported in Table 2.5

Chromosome	Start	End	Major CN	Minor CN	Clonality
chr1	4774360	50086909	3	1	1
chr4	73146745	108622482	4	1	1
chr5	3596404	66736867	7	3	1
chr5	96662536	96979783	11	1	1
chr6	33086365	33112548	2	2	1
chr7	63124646	64219715	2	1	1
chr8	213432	73793009	4	1	1
chr12	9253595	9256647	1	0	1
chr14	23953586	23965812	2	2	1
chr17	156162	41123669	9	6	1
chr20	95994	275844	2	1	1

Table 2.3: Simulated allele-specific copy number aberrations introduced in the synthetic dataset. Chromosome and genomic coordinates (bp) define the segment boundaries. Major CN and Minor CN denote allele-specific copy numbers. Clonality indicates the fraction of tumor cells harboring the aberration.

As illustrated, both approaches yielded consistent estimates of the optimal purity and ploidy, with purity equal to 79% and ploidy equal to 2.56. The correct inference of tumor purity, conditional on the segmentation results, suggests that the underlying smoothed signal was of sufficient quality to support reliable inference. A qualitative assessment of the copy number profiles reveals strong agreement between the inferred and the true copy number states for both approaches, indicating accurate identification of segment boundaries and their allele-specific copy number. The inferred genomic coordinates and allele-specific copy number by the proposed method and ASCAT, respectively shown in Tables 2.6 and 2.7, report that the true segments are fully recovered. To quantitatively assess the agreement between the reconstructed allele-specific

copy number aberrations and the ground truth, the mean squared error (MSE) across all SNPs has been computed for total copy number, major allele copy number (Major CN), and minor allele copy number (Minor CN). Both ASCAT and the proposed approach exhibited highly comparable error profiles, with only marginal differences across all metrics. The MSE values for total copy number were approximately 5.19 for ASCAT and 5.20 for the proposed method, the absolute difference between the two methods was only 0.006, corresponding to a relative deviation of approximately 0.12% from ASCAT. Similar behaviour was observed for major and minor copy number. These values are reported in Table 2.4. These results indicate that the proposed method achieves SNP-level accuracy that is similar from the original method, demonstrating that the methodological modifications introduced in the proposed approach do not compromise the fidelity of allele-specific copy-number reconstruction. Nevertheless, the proposed method exhibits a tendency towards excessive

Metric	ASCAT	Proposed Approach
Total CN	5.19237	5.198571
Major CN	2.665232	2.669517
Minor CN	0.6871737	0.6868647

Table 2.4: Mean squared error (MSE) between ground truth and ASCAT and the proposed approach for total copy number (Total CN), Major copy number (Major CN), and Minor copy number (Minor CN)

segmentation, in the sense that it divides some regions into smaller adjacent segments. However, it can be concluded that both methods are comparable in terms of their ability to accurately determine the copy number status in low-noise case scenarios.

Chromosomal region	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Chromosome	chr1	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr7	chr8	chr9	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr16	chr17	chr18	chr18	chr19	chr20	chr21	chr22	chrX	chrX
Penalty	110	70	55	55	100	50	60	35	50	100	60	115	130	50	90	35	125	55	50	60	40	35	35	70	40	35	40	55	35

Table 2.5: Penalty values per chromosomal region selected by the bivariate version of mBIC

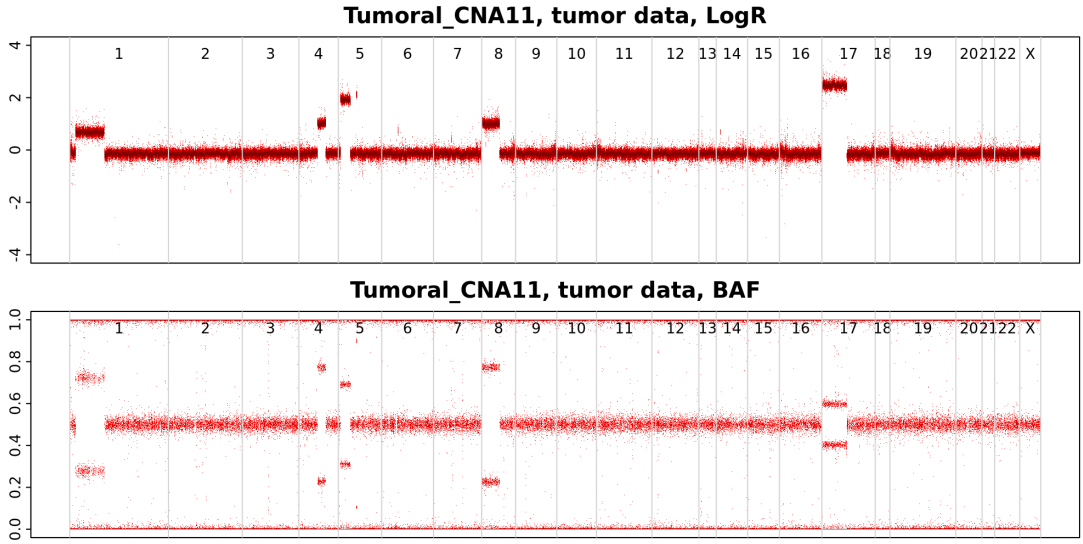

Figure 2.4: LogR (top) and B-allele frequency (BAF, bottom) measurements generated from synthetic data showing copy number alteration patterns and allelic imbalance.

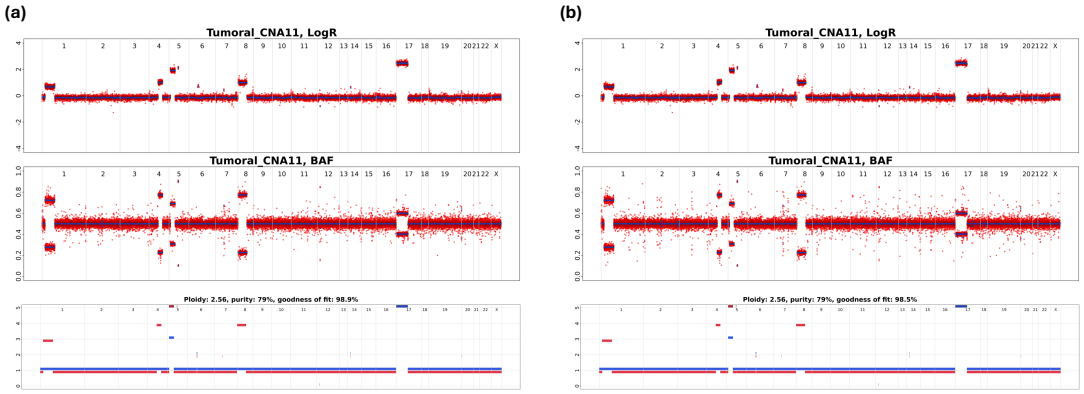

Figure 2.5: Comparison between the results obtained by ASCAT (a) with penalty parameter fixed equal to 70 and the proposed approach (b) on the synthetic data. For both panel (a) and (b), the upper plot shows in clue the segmentation of LogR (top) and BAF (bottom), on top of SNP measurements (red points). The lower plot show the reconstructed allele-specific copy number profile, and the optimal combination of purity and ploidy. In these plots, blue lines represents the minor allele (lowest copy number), while red lines represent the major allele (highest copy number).

Chromosome	Start	End	Major CN	Minor CN
chr1	4774360	50086909	3	1
chr4	73146745	108622482	4	1
chr5	3596184	66736867	7	3
chr5	96662536	96979783	11	1
chr6	33084946	33116921	2	2
chr7	63124646	64219755	2	1
chr8	213432	73805505	4	1
chr12	9253595	9256739	1	0
chr14	23953524	23988921	2	2
chr17	156162	41124087	9	6
chr17	4530819	4545672	10	6
chr17	4545798	16924431	9	6
chr17	16925244	16925824	10	6
chr17	16925845	21304604	9	6
chr17	21304980	21535247	10	6
chr17	21535268	41124087	9	6
chr20	95994	276122	2	1

Table 2.6: Inferred copy number segments with the proposed approach. Chromosome and genomic coordinates (bp) define the segment boundaries. Major CN and Minor CN denote allele-specific copy numbers.

Chromosome	Start	End	Major CN	Minor CN
chr1	4774360	50086909	3	1
chr4	73146745	108622482	4	1
chr5	3596184	66736867	7	3
chr5	96662536	96979783	11	1
chr6	33084946	33116921	2	2
chr7	63124646	64219755	2	1
chr8	213432	73805505	4	1
chr12	9253595	9256739	1	0
chr14	23953524	23988921	2	2
chr17	156162	41124087	9	6
chr20	95994	276122	2	1

Table 2.7: Inferred copy number segments with ASCAT. Chromosome and genomic coordinates (bp) define the segment boundaries. Major CN and Minor CN denote allele-specific copy numbers.

2.4 Discussion and future improvements

The detection of SCNAs from DNA bulk sequencing data has been investigated over the past few decades, leading to the development of numerous computational pipelines aimed at ensuring the reliability and robustness of the resulting genomic profiles. These methodologies have progressively evolved in parallel with advances in biotechnologies, which have enabled the generation of data with increasingly higher resolution and throughput. Among these tools, ASCAT is one of the most widely adopted approaches for SCNA analysis. Originally developed for SNP array data and later extended to NGS applications, it enables allele-specific copy number estimation, and tumor purity and ploidy inference.

Central to its implementation is the segmentation step, which aims to identify genomic breakpoints and partition the genome into regions with homogeneous underlying copy number states . This step is performed in a bivariate setting, incorporating both LogR and BAF measurements, to ensure a more reliable detection considering allelic imbalances. More specifically, ASCAT relies on a regularized approach called ASPCF, which is performed region-wise and whose solution depends on the choice of the penalty parameter: lower values increase breakpoint detection at the risk of over-segmentation, whereas higher penalties favor smoother profiles but might omit focal alterations. The selection of the penalty parameter is therefore pivotal, as it also depends on data characteristics and experimental conditions.

The method proposed in this section aims to overcome two key limitations of the existing ASPCF framework. First, ASPCF assumes homogeneous noise variance across signals and genomic regions, an assumption that is rarely satisfied in real genomic data, where variability may differ substantially across chromosomes. To address this issue, a multi-resolution strategy is hereby introduced. This strategy allows the penalty parameter to vary across genomic regions, thereby enabling different levels of regularization depending on the characteristics of the region under consideration. Second, the selection of the penalty parameter should be data-driven, and not left to the user. To this end,

a multi-resolution approach in which segmentation is performed at different scales, and the optimal level of regularization is selected using the modified Bayesian Information Criterion (mBIC) is proposed. This model selection criterion was originally developed specifically for change-point detection problems. To evaluate its effectiveness, the proposed multi-resolution framework was applied to both real and synthetic datasets, allowing for a comprehensive assessment of its performance in realistic scenarios as well as under controlled experimental settings. The synthetic dataset under consideration was generated with a recently developed tool, Synggen, which facilitates the creation of realistic synthetic cancer data harbouring different types of aberrations, including SCNAs. Preliminary results showed a high concordance with the original ASCAT method with a fixed penalty equal to the default value, in detecting copy number aberrations, while achieving higher resolution in real datasets.

To qualify the computational feasibility of the proposed multi-resolution framework for genome-wide analysis, here is a report indicative estimates of runtime and memory usage for typical WGS and WES samples considering the whole pipeline, with pre-processing of data and the proposed pipeline for segmentation. On average, whole-genome sequencing (WGS) data require approximately 400 MB of storage and 200 minutes of computation, whereas whole-exome sequencing (WES) data require 250 MB and 10 minutes, respectively. These values refer to single-core execution and include the full multi-resolution segmentation pipeline and the evaluation of the penalty grid.

From a computational perspective, the multi-resolution segmentation strategy is computationally efficient by design. First the genome is decomposed into separated regions, each of which is segmented independently. This decomposition reduces the effective problem size, for a full-chromosome profile. For each region, segmentation is performed over a fixed grid of penalty values, and the optimal solution is selected via the mBIC criterion. The total computational cost depends on the total number of genomic regions and number of penalty values explored. Despite this, the runtime remains manageable. Since genomic

regions are processed independently and each parameter value defines an independent segmentation task, the algorithm is naturally amenable to parallelization across both dimensions. A parallel implementation would therefore reduce runtime substantially, particularly for WGS-level data or large synthetic cohorts.

Although the preliminary results are encouraging, several limitations still need to be addressed. A first priority concerns computational scalability. The current implementation processes each genomic region independently for every penalty value, which results in a bilinear cost. However, the algorithm naturally decomposes across both genomic regions and penalty parameters, making it highly amenable to parallelization. Future work will therefore focus on exploiting this structure to develop a fully parallel implementation, substantially reducing runtime and improving scalability for large-scale studies.

From a methodolical viewpoint, a first consideration concerns the independence assumption. In the proposed method, LogR and mBAF were treated as independent variables. This choice was motivated both by simplicity and by the different sources of variability affecting the two signals. Although both derive from sequencing data, LogR is mainly influenced by GC bias and local depth fluctuations, whereas BAF is strongly affected by genotyping uncertainty and sequencing errors. These distinct noise sources justify, at least as a first approximation, the use of independent Gaussian components.

Relaxing the independence assumption would substantially modify the AS-PCF formulation and, consequently, the derivation of the mBIC. Indeed, if LogR and mBAF were modelled as non-independent, the concatenated vector $\boldsymbol{y} = (\boldsymbol{r}, \boldsymbol{b})$ would follow a Gaussian distribution with a full covariance matrix. This change would directly affect the likelihood function, the structure of the Bayes Factor, and the associated reparametrization, ultimately requiring a different model selection criterion. Exploring this direction represents a promising avenue for future work, with potential implications for a broader class of multichannel segmentation problems.

Secondly, the current framework assumes that the two signals share the same

variance. While this assumption is consistent with the original ASPCF formulation, assuming a different variance would allow the model to better capture signal-specific noise characteristics. Future work may therefore focus on incorporating channel-specific variance estimates or on introducing adaptive rescaling strategies to account for local variability. The mBIC formulation itself could be extended to explicitly model vectors with heterogeneous noise distributions, by adapting the Bayes Factor and its asymptotic approximations to allow channel-specific variance terms.

It is also important to note that the proposed approach does not provide a formal guarantee of global convergence due to the dynamic heuristic procedure of ASCPF, which addresses a non-convex optimization problem without theoretical guarantees of reaching the global minimum. In this sense, the proposed approach inherits the heuristic nature of the underlying segmentation framework, while remaining robust and effective for the detection of somatic copy-number alterations.

Finally, future work will focus on a more extensive evaluation of the proposed multi-resolution framework, taking into account the methodological extensions discussed above, such as channel-specific noise modeling or adaptive rescaling strategies. A first direction involves the construction of a comprehensive synthetic cohort using synggen. This cohort will include SCNA profiles with diverse architectures, purity and ploidy levels, and controlled noise intensities, in order to mimic realistic scenarios such as Formalin-Fixed, Paraffin-Embedded (FFPE) derived samples. Additional parameters, such as variable sequencing coverage, will be incorporated to assess the robustness of the method under increasingly challenging conditions.

A complementary validation strategy will rely on real data. In particular, WES samples paired with their corresponding WGS profiles may be used to evaluate consistency across technologies, while accounting for their intrinsic differences. Moreover, WES data could be systematically downsampled or selectively masked to emulate TS designs, enabling an assessment of how segmentation accuracy degrades as coverage and genomic resolution decrease.

Together, these analyses would provide a comprehensive characterization of the strengths and limitations of the proposed framework across a wide spectrum of experimental settings.
Depending on the type of source data, different tools will be used for benchmarking. For WES synthetic and real datasets, tools such as CNVkit, FACETS, and EXCAVATOR [35, 36, 77] are appropriate comparators, given their widespread use and their distinct segmentation strategies, including the bivariate formulation implemented in FACETS. For WGS-like data, additional methods such as Battenberg or HATCHet [78, 60] can be included to broaden the comparison. The evaluation should consider both breakpoint accuracy, i.e., the concordance between inferred and true change-point, and the correctness of the inferred copy-number states. This dual assessment will allow a rigorous quantification of segmentation performance across heterogeneous scenarios.

Chapter 3

A Multi-Step Computational Framework for Allele-Specific Somatic Copy Number Aberration Detection from Single-Cell RNA Sequencing Data

Single-cell sequencing technologies have revolutionised cancer research by enabling the analysis of tumours at single-cell resolution. Among these, single-cell RNA sequencing has become increasingly widespread, and several computational methods have been developed to infer somatic copy number aberrations from the type of data generated by these platforms. These computational approaches differ in their underlying assumptions, analytical strategies, and statistical models, and have shown promising results across a variety of application contexts. However, only a limited number of methods are able to perform allele-specific analyses, mainly due to the high level of noise and sparsity that characterises scRNA-seq data.

In this Chapter an extension of SCEVAN [2] for the detection of allele-specific SCNAs from scRNA-seq data is presented. SCEVAN has been shown to be a

fast method for identifying SCNAs, while also providing strong accuracy in distinguishing subclonal populations within tumour samples [79].
The Chapter is organized as follow. Section 1 discusses the technical limitations of scRNA-seq data and reviews the main existing approaches, together with the reference method. Section 2 describes the proposed extension in detail, while Section 3 reports preliminary results obtained on real and synthetic datasets. Finally, the results are discussed in Section 4, highlighting the potentiality and limitations of the method, as well as possible future developments.

3.1 A Review on Computational Tools for SCNA Analysis from Single-Cell RNA Sequencing Data

In recent decades, there has been a significant increase in the generation and analysis of single-cell data, as it has proven to be valuable in studying tumour heterogeneity and structural complexity [80, 81]. As previously mentioned, the advent of single-cell sequencers has also facilitated a more profound and detailed comprehension of SCNAs, and their clonal evolution. In particular, recent advances in single-cell DNA sequencing and analysis aim to become a gold standard for copy number analysis, and have already improved the analysis of heterogeneity and its potential correlation with clonal fitness, i.e. cell growth dynamics. [82].
However, single-cell DNA sequencing is often less widely used due to its high cost and experimental complexity, whereas single-cell RNA sequencing is more accessible, making it increasingly attractive for bioinformatic analyses, including the detection of SCNAs at single-cell resolution [41]. Despite the scalability and cost advantages of scRNA-seq, several factors must be considered when developing methods for SCNA detection from this type of data. A key challenge is the non-uniform genome coverage: highly expressed genes produce more reads, while lowly expressed genes are underrepresented, potentially leading to false-positive SCNA calls unrelated to actual genomic changes. Bio-

logical factors such as transcriptional regulation and diverse cellular states can also significantly affect physiological gene expression. For example, cellular heterogeneity within a sample can lead to variation in expression patterns, even among cells that exhibit the same genomic alterations [83, 2].
Many computational methods for inferring SCNAs from scRNA-seq data have been developed, reflecting differences in data characteristics across scRNA-seq platforms [84]. Most of these tools are expression-based, meaning they infer copy number directly from gene expression, and primarily focus on total copy number analysis. The most used and known tools include inferCNV [85], CopyKAT [84], and SCEVAN [2]: inferCNV utilizes a corrected moving average over gene windows, CopyKAT employs an integrative Bayesian segmentation approach, while SCEVAN uses a multi-channel segmentation algorithm to obtain segment-level copy number states [79].
In recent years, only a limited number of methods have been developed to integrate allelic information for the detection of allele-specific SCNAs. This scarcity reflects the technical challenges inherent to scRNA-seq data, such as dropout events and cellular heterogeneity.
Dropout occurs when a gene fails to be detected in a cell despite active transcription, resulting in missing or zero-valued measurements. This phenomenon is especially prevalent in single-cell RNA sequencing, where the amount of RNA captured from each cell is inherently limited. Beyond gene expression, dropout events also affect the estimation of allelic imbalance, commonly quantified by the B-allele frequency. As a result, scRNA-seq data are characterized by a high proportion of zeros or missing values that reflect both technical limitations and low expression levels. Such sparsity poses significant challenges for reliable inference and requires specialized computational approaches for modeling and imputation, as standard bulk RNA-seq methods are not directly applicable [86].
Cellular heterogeneity within tumor samples adds an additional layer of complexity. Differences in gene expression across clonal populations can lead to substantial variability in BAF estimates across cells, even in the presence

of shared genomic alterations. Overall, the inference of allelic or haplotype-specific SCNAs from scRNA-seq data remains an open problem and an active area of research. Addressing this challenge requires sophisticated statistical models capable of accounting for data sparsity, technical noise, and allelic variation, enabling more accurate allele-specific SCNA inference at single-cell resolution.

Numbat [41] and XClone [82] are two recent methods specifically designed to leverage allelic information while accounting for sparsity and technical noise. These tools demonstrate the feasibility of allele-specific SCNA inference from scRNA-seq data, however they focus on modeling frameworks that may not fully exploit the robustness and scalability of segment-level approaches developed for expression-based inference.

Motivated by the strengths and robustness of SCEVAN, this work aims to explore its extension to allele-specific SCNA detection. By integrating allele information into the existing implementation of SCEVAN, the goal is to infer allele-specific copy number states while preserving the scalability, robustness, and suitability of the method for single-cell datasets.

3.1.1 Existing methods for allele-specific SCNA detection

Over the past five years, algorithms such as Numbat and XClone have been developed to detect copy number aberrations at the level of individual alleles. Both methods exploit proxies for BAF when phasing SNPs along the two homologous chromosomes. These proxies, derived from parental allele frequencies (PFs), allow the reconstruction of maternal and paternal haplotypes despite the sparsity of scRNA-seq data.

In particular, Numbat is a computational framework designed not only to detect allele-specific SCNAs, but also distinguish tumor from normal cells, and infer the subclonal architecture and evolutionary phylogeny of malignant cell populations. To achieve these goals, Numbat iteratively performs three main steps:

- **Pseudobulk SCNA inference.** Single-cell data are aggregated into pseu-

dobulk profiles corresponding to major clades identified from a single-cell phylogeny. In the first iteration, cells are grouped into three clades based on gene expression similarity, whereas in subsequent iterations the clade structure depends on the implementation of successive refinement steps. Numbat then applies a haplotype-aware hidden Markov model (HMM) to these pseudobulk profiles to reconstruct lineage-specific SCNAs at the pseudobulk level. The HMM comprises 15 hidden copy-number states, including copy-neutral loss of heterozygosity (CN-LOH). Each state emits observations at both the gene and SNP levels through state-specific probability mass functions. Gene expression is modeled using a Poisson-lognormal distribution, while paternal allele counts at each SNP are modeled using a beta-binomial distribution. The parameters of these distributions are either specified via hyperparameters or estimated directly from the data. Transition probabilities encode both copy-number state changes and haplotype switches, with the latter modeling phase-switch events between major and minor haplotypes. At this stage, information is aggregated across all cells within each pseudobulk.

- **Single-cell SCNA assignment.** Single-cell SCNAs are inferred by jointly modeling gene expression and allele count information within a Bayesian framework. For each genomic region and each cell, posterior probabilities over a discrete set of copy-number genotypes are computed by combining expression- and allele-based likelihoods with prior information derived from the pseudobulk analysis. In particular, posterior probabilities of alteration types inferred at the pseudobulk level are propagated as priors for single-cell genotype inference, while the likelihood functions follow the model described above.

- **Phylogenetic reconstruction and iteration.** The inferred SCNA states across cells and genomic regions are subsequently used to reconstruct subclonal relationships. Numbat builds a maximum-likelihood phylogenetic tree that captures the clonal and subclonal architecture of the tumor,

with copy-number alterations mapped along evolutionary branches. Cells are assigned to distinct subclones based on their phylogenetic relationships. The resulting phylogeny can then be used to redefine pseudobulk groupings, enabling iterative refinement of both SCNA inference and evolutionary reconstruction.

On the other hand, XClone is a statistical model to infer single-cell allele-specific and haplotype-specific SCNAs from low-coverage and sparse scRNA-seq. XClone operates through four direct fundamental steps.

- **Preprocessing: extract count matrices** A multi-level allele phasing procedure is perfomed to estimate haplotype-specific allele frequencies. This step requires a partition of the genome into genomic bins, to increase statistical robustness in sparse single-cell data. These bins involve 100 genes by default.

- **Pre-analysis: horizontal and vertical smoothing** Raw gene expression and phased allele frequency signals are smoothed along the genome and across cells. This smoothing procedure is achieved via a weighted moving average applied chromosome-wise, while cell-to-cell relationships are incorporated through a k-nearest neighbor (KNN) graph, enabling information sharing across transcriptionally similar cells.

- **Mixture models for expression and allele-specific data** Copy number states are inferred by using separate Bayesian probabilistic models in which smoothed gene expression is modeled by a negative binomial distribution, and haplotype-specific allele frequencies are modeled using a beta-binomial distribution. A negative binomial distribution is used to model gene expression counts, while a beta-binomial distribution to model the observed allele counts. Then, the SCNA state assignment probability matrices are computed.

- **Combination modules** Posterior probabilities of copy number states obtained from the expression and allele-specific components are integrated

within a unified module to produce haplotype-specific SCNA calls for each cell.

Results reported in recent benchmarking studies suggest that, while Numbat performs well in detecting clonal allele-specific SCNAs and major copy gains, it may suffer from reduced sensitivity in detecting copy losses and focal events, particularly in datasets characterized by low tumor purity or weak allelic signal. Moreover, Numbat initially relies on pseudobulk inference, refined to single-cell via hierarchical Bayes, and its performance and robustness remain to be fully evaluated, particularly on samples with extreme clonal SCNA compositions. Additionally, it is a computationally intensive method [82].
To date, XClone has limited dedicated benchmarking studies focusing on its behavior across different SCNA scales. Nevertheless, some potential limitations can be hypothesized based on its methodological design. In particular, XClone applies a relatively wide smoothing window (on the order of 100 genes) to allelic signals. While this strategy is effective in stabilizing noisy scRNA-seq measurements and in detecting large-scale or arm-level events, it may reduce sensitivity to smaller, non-focal SCNAs that span genomic regions shorter than the smoothing window but are larger than single-gene events. As a consequence, intermediate-scale SCNAs may be partially smoothed out and thus harder to detect.

3.1.2 Expression-based SCNA inference: SCEVAN

SCEVAN is a fast variational algorithm that was originally developed with the intention to infer SCNAs from single-cell RNA-seq gene expression raw counts data. This algorithm is reported to be extremely accurate in the automatic discrimination between malignant and non-malignant cells and suitable for detecting the clonal copy number substructure of tumors as well [87]. In particular, SCEVAN is capable of identifying total copy number alterations, including both single-copy losses and double-copy losses (complete deletions), as well as gains and amplifications.

Due to the complexity of single-cell expression data, SCEVAN is organized into multiple stages. These include data preprocessing, identification of non-malignant cells, and detection of tumor subclones. Subsequently, copy-number states are inferred for each cell within each subclone, enabling the identification of genomic regions affected by gains, amplifications, and deletions. The detailed description of the SCEVAN pipeline provided in this chapter is intended to highlight the assumptions and algorithmic components that will be extended in the following section, while a schematic representation is illustrated in figure 3.1.

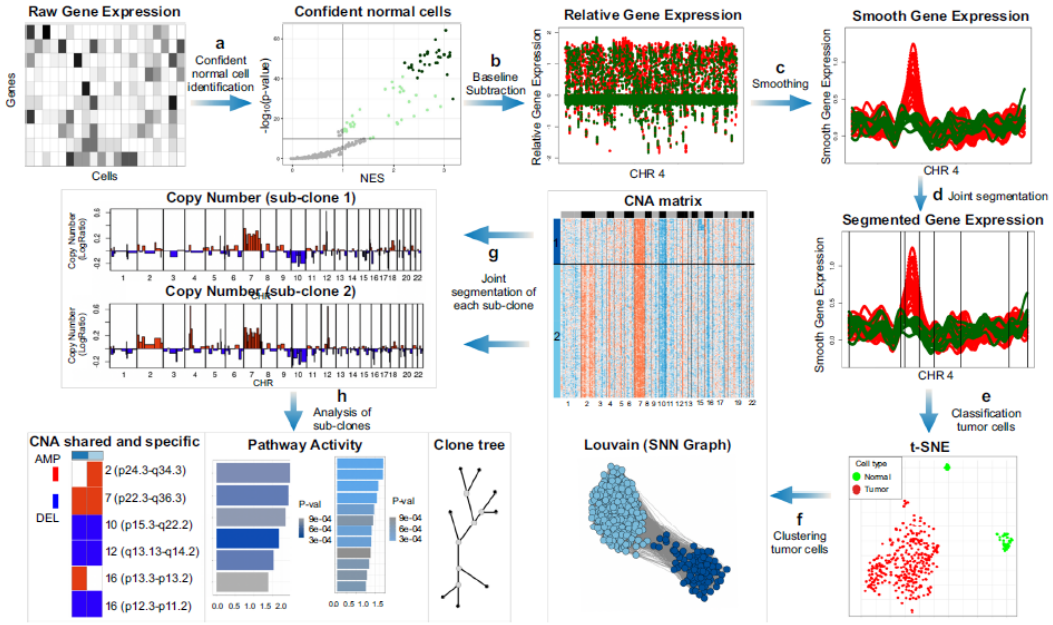

Figure 3.1: SCEVAN workflow. Source De Falco et al. [2]

Preprocessing of scRNA-seq data SCEVAN starts with a preprocessing of the raw count matrix $\mathbf{X} \in \mathbb{N}^{C \times G}$, with C representing the total number of cells, and G the amount of genes being examined. Specifically, cells with less than 200 detected genes and the genes expressed in less than 1% of cells are filtered out, as well as those associated with the cell cycle pathway, in order to reduce the artificial signal caused by the cell cycle itself. The resulting matrix

is $\tilde{\mathbf{X}} \in \mathbb{N}^{\tilde{C} \times \tilde{G}}$, with $\tilde{C} < C$ and $\tilde{G} < G$.

Identification of non-malignant cells This procedure consists of three main stages: (i) identification of a high-confidence reference set of normal cells, (ii) detection of shared clonal breakpoints, and (iii) identification of normal and malignant cells across the entire dataset. The rationale underlying this approach is that identifying a reference set of normal cells allows the removal of a normal gene expression baseline, thereby making expression deviations of tumoral cells more evident. These deviations can then be detected by a segmentation algorithm, while normal cells are identified as those exhibiting no significant variation across genomic segments. The following sections describe the algorithmic details of this procedure.

(i) The first step comprises the computation of the relative expression matrix $\tilde{\mathbf{X}}_r$, by subtracting a baseline vector $\mathbf{b} \in \mathbb{R}^G$ from each row of $\tilde{\mathbf{X}}$. The baseline is estimated from the median expression of high-confidence normal cells and aims at removing biological noise, isolating tumour-specific alterations, and obtaining more accurate and interpretable expression profiles.

This high-confidence set of normal cells is identified using a single-sample Mann-Whitney-Wilcoxon test to assess enrichment of predefined tumor microenvironment gene signatures in each cell's expression profile. According to the original method, cells with a normalized enrichment score greater than 1 and a p-value below 10^{-10} are classified as normal, and at most 30 such cells are retained to define a reference baseline. The baseline vector $\mathbf{b}$ is estimated from the median expression of the selected normal cells. If no normal cells are detected, a synthetic baseline is constructed via gene-wise Gaussian perturbation within clusters identified by hierarchical clustering, with the number of clusters selected using the Calinski-Harabasz criterion [88]. The latter evaluates clustering quality by comparing the variance between clusters to the variance within clusters.

(ii) Before segmentation, the relative expression matrix $\tilde{\mathbf{X}}_r$ is smoothed to preserve edges, i.e., potential clonal breakpoints of tumour cells, and reduce

outliers. This is achieved via nonlinear smoothing based on total variation minimization, producing a new matrix $\mathbf{U}_r \in \mathbb{R}^{\tilde{C}\times\tilde{G}}$.
The rows of $\mathbf{U}_r$ are then jointly segmented using VegaMC [89], a multichannel extension of the VEGA algorithm that was explained in Section 1 whose details can be found in Appendix A.2, where each cell is treated as a separate channel. A full overview of the mothod is available in Appendix A.2. Shared genomic breakpoints $\{\tau_0, \ldots, \tau_Q\}$ are identified across cells, and a segmented matrix $\mathbf{C} \in \mathbb{R}^{\tilde{C}\times\tilde{G}}$ is constructed. Each entry $\mathbf{C}_{ck}$ represents the average expression of cell c over the segment containing gene k:

$$\mathbf{C}_{ck} = \frac{1}{\tau_{k+1} - \tau_k} \sum_{i=\tau_k}^{\tau_{k+1}-1} \tilde{\mathbf{X}}_{ci}. \tag{3.1}$$

(iii) The identification of all non-malignant cells is then performed, through a hierarchical clustering of $\mathbf{C}$ into two groups. Cells within the cluster containing the largest number of confident normal cells are classified as normal and removed from further analysis. Denoting by C_M the total number of malignant cells, SCEVAN then proceeds with the analysis on $\tilde{\mathbf{X}}_r^M \in \mathbb{R}^{C_M\times\tilde{G}}$, the matrix restricted to rows corresponding to the identified tumor cells.

Subclones identification To infer the subclonal architecture, the malignant cell matrix $\tilde{\mathbf{X}}_r^M$ is first clustered using Louvain clustering [90] applied to a shared nearest-neighbor graph [91]. Louvain clustering identifies groups of transcriptionally similar cells by detecting densely connected communities in the graph, making it well-suited to partition malignant cells into candidate subclones. Denoting the resulting clusters as $S_1, \ldots, S_K$, where $S_i \subset \{1, \ldots, C_M\}$, each submatrix $\tilde{\mathbf{X}}_r^{S_i}$, corresponding to the cells in S_i, is then segmented using the multichannel VegaMC procedure to identify subclone-specific genomic breakpoints.

Identification of subclones and single-cell SCNA calling Copy number states of each genomic segment are estimated using a mixture model applied

to the mean expression levels of cells within each subclone and segment, following Magi et al. [77]. Segment means are modeled as a mixture of five truncated normal distributions, corresponding to canonical copy number states: deletion, loss, neutral, gain, and amplification. Formally, the mixture is defined as

$$g_l^u(m_i; \mathbf{p}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \, g_l^u(m_i \mid \mu_k, \sigma_k), \quad \sum_{k=1}^{K} p_k = 1, \tag{3.2}$$

where $g_l^u(m_i \mid \mu_k, \sigma_k)$ denotes a Gaussian density truncated to the interval $[l, u]$, m_i is the mean expression of segment i, $\mathbf{p} = \{p_1, \ldots, p_K\}$ are the mixture weights, and $\boldsymbol{\theta} = \{\mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K\}$ are the component means and standard deviations. In this precise case, $K = 5$, representing each possible SCNA event, while truncation bounds are introduced to restrict the support of the model to biologically plausible expression values. Fixed variables, such as l, u are described in the suppementary material of Magi et al. [77].
Model parameters are estimated using the classical Expectation-Maximization (EM) algorithm. EM iteratively computes the posterior probabilities of each segment belonging to the five copy number states (E-step) and updates the mixture proportions, means, and variances of the components (M-step) until convergence. Each segment is then assigned to a discrete copy number state by selecting the state with the highest posterior probability: double loss/deletion (0 copies), loss (1 copy), neutral state (2 copies), gain (3 copies), or amplification (>3 copies).

Conclusions In this section, the main computational steps underlying SCEVAN, a scalable and widely adopted method for SCNA detection from scRNA-seq data, have been reviews. By combining expression smoothing, joint segmentation, and mixture model-based copy number classification, SCEVAN provides a robust framework for identifying total copy number alterations and subclonal structures in large single-cell datasets. Despite its strengths, the method relies exclusively on total gene expression and does not explicitly model allelic imbalance. As a result, copy-neutral loss of heterozygosity and

other allele-specific SCNAs cannot be directly detected. These limitations motivate an extension of SCEVAN that incorporates allele-specific information from scRNA-seq. In the following section, a novel allele-aware formulation that integrates BAF information into the inference pipeline will be presented, enabling the detection of allelic SCNAs while preserving the scalability and robustness of the original method.

3.2 Algorithmic Implementation of the Allele-Specific SCEVAN Extension

As stated in the previous section, integrating allelic information into scRNA-seq-based copy number inference is challenging due to the extreme sparsity and noise of allele-specific read counts at single-SNP and single-cell resolution. Existing methods typically address this issue either by aggregating cells into pseudobulks, which can smooth subclonal variation, or by smoothing allelic signals within genomic regions of individual cells, creating a trade-off between noise reduction and spatial resolution in the presence of dropout events.
To overcome these limitations, an extension of SCEVAN is proposed that integrates gene expression and allelic imbalance for robust segment-level copy number inference. The approach proceeds in multiple steps. First, allelic signals are stabilized through chromosome-arm level aggregation (p/q) and incorporated within a multi-modal framework to capture shared structures across cells. These shared structures are then used to estimate segment-level allelic information. Subsequently, expression and allelic imbalance are modeled separately and combined probabilistically to infer discrete copy number states at both single-cell and subclone levels.
This strategy balances robustness and resolution, enabling the detection of subclonal copy number alterations while preserving the scalability and interpretability of the original SCEVAN framework. The model and inference procedure are described in detail in the following sections, and an overview of the method is illustrated in Fig.3.2.

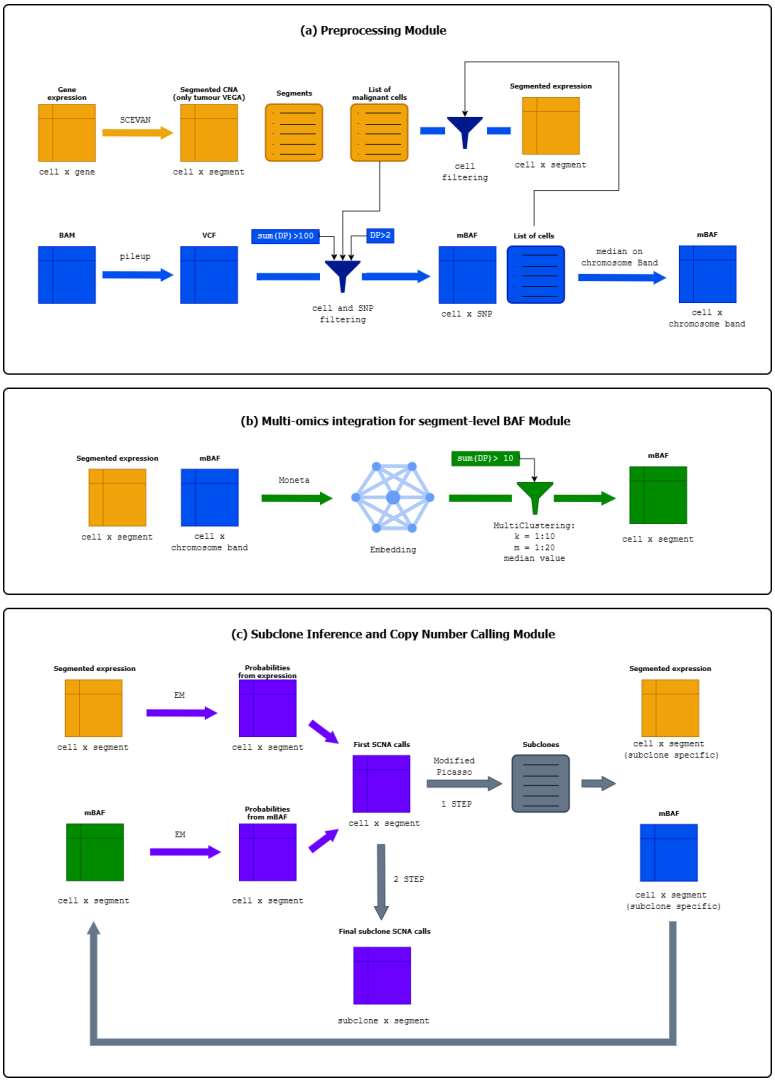

Figure 3.2: Overview of the allele-specific framework for SCEVAN pipeline. (a) Preprocessing of gene expression and allelic data. Low-quality cells, genes, and SNPs are removed, resulting in a cell $\times$ segment expression matrix and a chromosome-band mBAF profile. (b) Data integration of expression and mBAF through MoNETA produces an embedding space that is used to compute an iterative clustering based on induced cell-cell similarity. This step results in a cell $\times$ segment mBAF profile, obtained by iteratively aggregating allelic information across similar cells. (c) Primary copy number calls are inferred through a combination module that integrates cell $\times$ segment expression and mBAF posterior probabilities, estimated via an Expectation-Maximization procedure under a probabilistic model. A modified version of PICASSO is then employed to infer the subclonal composition, followed by a second copy number calling step. Both single-cell and subclone-level copy number profiles are obtained.

3.2.1 Preprocessing of Expression Matrix and Allelic Data

The preprocessing pipeline handles two main types of input data: the raw gene expression count matrix $\mathbf{X} \in \mathbb{N}^{C \times G}$, with C and G representing the total number of cells and genes respectively, and a VCF file containing allele-specific read counts at the SNP level for each cell. In particular, this file encodes allele depth (AD) and total depth (DP) information for each detected SNP in each cell, where individual SNPs are observed in subsets of cells.

First, the gene-expression raw count matrix $\mathbf{X}$ undergoes the SCEVAN preprocessing pipeline to obtain the relative expression matrix. The identified non-malignant cells and cells with a total DP across all SNPs lower than 100 are also excluded from the analysis, to further reduce technical noise. Subsequently, the segmentation step by means of VegaMC is applied to the obtained relative expression matrix restricted to malignant cells, with the aim of detecting common breakpoints across all malignant profiles. This procedure yields a segmented matrix

$$\mathbf{C} \in \mathbb{R}^{\tilde{C}_M \times \tilde{Q}}, \tag{3.3}$$

where Q denotes the total number of obtained segments. This segmentation leads to a set of breakpoints $\{\tau_0, ..., \tau_Q\}$ and each segment I_s is defined by its bounding breakpoints τ_s and τ_{s+1}, consisting of all positions in between.

Having processed the expression data, the focus now turns to the allele-specific information in the VCF to construct a complementary BAF matrix. First of all, homozygous SNPs, which are non-informative, and SNPs with low coverage (DP < 3) are removed from the VCF.

Working directly with single-SNP BAF is problematic because DP is highly variable both across SNPs and across cells, and there is no a priori knowledge of cell similarity that would allow the construction of reliable pseudobulks. Moreover, aggregating information at segmentation level, using expression information, could yield very fine regions, which may contain few or no SNPs in a given cell, or would require selecting a fixed number of genes per segment. To overcome these limitations, allelic imbalance is summarized at

the chromosome-band level, providing broader regions that are biologically meaningful and sufficiently populated with SNPs across cells.
Since simple aggregation of AD and total depth DP is not appropriate due to the natural oscillation of BAF around 0.5, the mirrored BAF (mBAF) is computed for each band. For each SNP, the BAF is computed as

$$\text{BAF} = \frac{\text{AD}}{\text{DP}} \tag{3.4}$$

and the mirrored around 0.5 version is determined as follows

$$\text{mBAF} = \begin{cases} 1 - \text{BAF}, & \text{if BAF} < 0.5, \\ \text{BAF}, & \text{otherwise}, \end{cases} \tag{3.5}$$

Then, the median of these values within the band is taken to obtain a robust measure of allelic imbalance. The use of the median is particularly advantageous because it smooths out potential phasing errors that might incorrectly classify a homozygous SNP as heterozygous, providing stable estimates even in the presence of sparse coverage.
The resulting values lie in the interval $[0.5, 1]$ and are stored in a matrix

$$\mathbf{M} \in \mathbb{R}^{\tilde{C}_M \times C_b}, \tag{3.6}$$

where $\tilde{C}_M$ denotes the number of filtered malignant cells and C_b the number of chromosome bands considered. Here, C_b corresponds to twice the number of chromosomes in the experiment, reflecting the subdivision into bands.
Due to both filtering and the inherent sparsity of allelic signals, some entries in $\mathbf{M}$ may be missing, i.e. no SNPs are present in a given cell within a particular chromosome band. To address this, the chromosome-band-level mBAF matrix is subjected to a filtering and imputation procedure. First, bands with insufficient allelic information are removed: only bands for which the proportion of non-missing BAF values across cells exceeds a predefined threshold are retained. This filtering threshold ensures that each retained band has sufficient

allelic information for reliable downstream inference. Next, missing mBAF values are imputed using a k-nearest neighbors (kNN) approach, whereby the allelic profiles of similar cells are used to estimate unobserved values. Imputation is performed in the cell space using 10 nearest neighbors, yielding the imputed mBAF matrix

$$\tilde{\mathbf{M}} \in \mathbb{R}^{\tilde{C}_M \times \tilde{C}_b}, \tag{3.7}$$

where $\tilde{C}_b$ denotes the number of chromosome bands retained after filtering. In this work, the threshold is set to 30% and was chosen empirically, as a conservative criterion to remove chromosome bands with limited information and could bias downstream inference, while still preserving a sufficient number of bands for robust analysis.

3.2.2 Cell-Wise BAF Signal Estimation via Graph-Based Data Integration and Iterative Clustering

Following the initial preprocessing step, a two two-step strategy is adopted to infer a segment-level allelic information. This enables analysis within a shared genomic space, with the aim of conducting subsequent joint analysis with segment-level expression data. First, a multi-modal data integration framework is used to infer a cell-cell similarity structure by jointly considering gene expression information, represented by the matrix $\mathbf{C}$, and chromosome-band-level allelic matrix $\tilde{\mathbf{M}}$. Second, the resulting similarity structure is exploited to aggregate cells through an iterative clustering approach, yielding more stable estimates of segment-level BAF.

Graph-based multi-modal integration with MoNETA Cell-cell similarities are inferred using MoNETA, a fast and scalable framework for graph-based multi-modal integration at single-cell resolution [92]. Since a detailed description of MoNETA is provided in Appendix A.5, only a brief overview of its application is given here. MoNETA aims to derive a low-dimensional latent representation of the multi-modal molecular profiles, encoded in a matrix

$\mathbf{E} \in \mathbb{R}^{\tilde{C}_M \times D}$, with $D < \tilde{C}_M$. To this end, omic-specific similarity networks are first constructed, each capturing cell-cell relationships within a single data modality. These networks are then integrated into a unified multi-modal graph, and a Random Walk with Restart procedure is applied to the integrated network that results in a low-dimensional embedding, summarizing the shared and complementary structure across omics layers.
In this work, MoNETA is used to integrate segment-level gene expression information, represented by the matrix $\mathbf{C}$, together with chromosome-band-level allelic information, $\tilde{\mathbf{M}}$, thereby capturing complementary sources of biological variability across cells.

An iterative clustering approach to estimate cell-wise BAF The embedding $\mathbf{E}$ obtained from the multi-modal integration step provides a low-dimensional representation of cells, capturing their similarity across the two modalities, and it serves as an input to the following step. An iterative clustering strategy inspired by DrImpute [86] is now adopted to compute the segment-level mBAF matrix.
Let $K_{\max} \in \mathbb{N}$ denote the maximum number of clusters considered. At each iteration $k = 1, ..., K_{\max}$, a k-means clustering is applied to the embedding matrix $\mathbf{E}$, to partition cells into clusters. To account for variability due to different k-means initializations, a total of $M_{\max} = 20$ clustering iterations are performed to capture alternative plausible partitions. Note that for $k = 1$, all cells are assigned to a single cluster and one iteration is performed.
The cluster-level mBAF profile is computed as follows:

1. Once fixed a partition into k clusters at iteration $m = 1, \ldots, M_{\max}$, let

$$C_j = C_j(k, m) \subseteq \{1, \ldots, \tilde{C}_M\}$$

   denote the j-th cluster, with $j = 1, \ldots, k$.

2. Restrict the AD and DP information within the filtered VCF to the cells

in C_j. For simplicity in notation, let AD_n^i and DP_n^i be the AD and DP of the $i-$th SNP in cell n, respectively. If no SNP is detected, then $AD_n^i = 0$ and $DP_n^i = 0$.

3. For each SNP at position i, aggregate allele and total depths across all cells in the cluster:

$$\mathrm{BAF}_{C_j,k,m,i} = \frac{\sum_{n \in C_j} \mathrm{AD}_n^i}{\sum_{n \in C_j} \mathrm{DP}_n^i}. \tag{3.8}$$

4. Compute the mirrored BAF ($\mathrm{mBAF}_{C_j,k,m,i}$) to ensure symmetry around allelic balance:

$$\mathrm{mBAF}_{C_j,k,m,i} = \begin{cases} 1 - \mathrm{BAF}_{C_j,k,m,i} & \text{if } \mathrm{BAF}_{C_j,k,m,i} < 0.5, \\ \mathrm{BAF}_{C_j,k,m,i} & \text{otherwise.} \end{cases} \tag{3.9}$$

5. For each previously computed segment I_s with a total DP aggregated across cells and SNPs greater than 10, the algorithm computes the median of mBAF values across SNPs to obtain a cluster-level segment mBAF:

$$M_{C_j,I_s}^{(k,m)} = \mathrm{median}_{i \in I_s} \mathrm{mBAF}_{C_j,k,m,i}. \tag{3.10}$$

The total DP threshold ensures sufficient coverage, filtering poorly covered segments. This resulting segment-level mBAF is assigned to all cells in C_j within columns for the I_s segment, and a cell-level matrix with cluster-shared values for the k-th clustering at iteration m is

$$\mathbf{M}^{(k,m)} \in \mathbb{R}^{\tilde{C}_M \times Q}, \tag{3.11}$$

where $\tilde{C}_M$ is the number of malignant cells and Q is the total number of segments.

6. Finally, all cluster-induced segment-level matrices obtained across the different values of k and clustering initializations are aggregated element-wise using the median to produce a robust final estimate of the segment-

level mBAF:

$$\mathbf{M}^{\text{final}} = \text{median}_{(k,m)}\mathbf{M}^{(k,m)}. \tag{3.12}$$

For each cell and genomic segment, the final mBAF value therefore reflects the median across all cluster-based estimates obtained from the different partitions in which the cell is included.

The strategy reduces single-cell noise by borrowing information across groups of cells with similar expression and allelic profiles, as captured by the multimodal embedding. By aggregating allele counts within locally shared genomic segments across these groups, sparse and noisy SNP-level observations are stabilized, while preserving biologically coherent segment-level allelic states.

3.2.3 Initial EM-Based Allele-Specific Copy Number Inference

At this stage, for each cell and each genomic segment, two complementary sources of information are available: relative expression levels from $\mathbf{C}$ and allelic imbalance from $\mathbf{M}^{\text{final}}$. To infer segment-level copy number states, an Expectation-Maximization (EM) framework is employed to estimate the posterior probability that a given cell belongs to a specific copy number category.

For expression data, the EM model follows the truncated Gaussian mixture framework implemented in SCEVAN, providing posterior probabilities over five discrete copy number states (loss, deletion, neutral, gain, amplification). Please, refer to section 3.1.2 for details.

Independently, mBAF values in $\mathbf{M}^{\text{final}}$ are modeled using a three-component truncated Gaussian mixture corresponding to canonical allelic states (loss, neutral, and amplification), following the framework defined in equation 3.2, with $K = 3$. Then, the EM algorithm estimates mixture parameters and posterior probabilities for each cell and segment, using initial component means fixed at 0.9 (loss), 0.6 (neutral), and 0.75 (amplification). Since mBAF values are mirrored around 0.5 and constrained to [0.5, 1], these values reflect empirically observed allelic imbalance patterns. Initial priors are set according

to the expected biological prevalence of each state (0.8 for neutral, 0.1 for deviations).
Let

$$\mathbf{P}_{c,s}^{\text{expr}} = \left(P_{c,s,\text{amp}}^{\text{expr}}, P_{c,s,\text{gain}}^{\text{expr}}, P_{c,s,\text{neu}}^{\text{expr}}, P_{c,s,\text{del}}^{\text{expr}}, P_{c,s,\text{loss}}^{\text{expr}}\right) \tag{3.13}$$

denote the posterior distribution over expression-based copy number states for cell c and segment s, and

$$\mathbf{P}_{c,s}^{\text{mBAF}} = \left(P_{c,s,\text{loss}}^{\text{mBAF}}, P_{c,s,\text{neu}}^{\text{mBAF}}, P_{c,s,\text{amp}}^{\text{mBAF}}\right) \tag{3.14}$$

the posterior distribution over allelic imbalance states.
Since the two posteriors have different numbers of states, the joint posterior is computed over the Cartesian product of expression and mBAF states. Each entry corresponds to a unique combination of states, as defined in Table 3.1. The joint posterior is obtained by element-wise multiplication:

$$\mathbf{P}_{c,s}^{\text{joint}} = \mathbf{P}_{c,s}^{\text{expr}} \odot \mathbf{P}_{c,s}^{\text{mBAF}}. \tag{3.15}$$

The combined copy number call for cell c and segment s is the combination with the highest joint posterior probability:

$$\arg\max_{state} \mathbf{P}_{c,s}^{\text{joint}}. \tag{3.16}$$

For segments lacking mBAF data, the most probable copy number state is imputed from adjacent segments. If multiple consecutive segments are missing, imputation proceeds iteratively from the nearest available segments. Entire chromosomes, or chromosome bands, without mBAF information rely solely on expression-based posterior probabilities.
The joint posterior defines a Cartesian product of expression and mBAF states, with biologically interpretable allele-specific copy number states: copy-neutral (Neu), gain (Gain), amplification (Amp), deletion (Del), loss (Loss), and copy-neutral loss of heterozygosity (CNLoH). Whole-genome duplication (WGD), which could occur when expression indicates copy loss or deletion while

mBAF remains neutral, is not currently modeled. Its exclusion does not affect inference of other states and will be implemented in future work.
The final copy number states matrix is

$$\mathbf{CN} \in \mathbb{N}^{\tilde{C}_M \times Q}, \tag{3.17}$$

where states are encoded as ordered categorical variables: loss, deletion, neutral, gain, amplification, and CNLoH mapped to $-2, -1, 0, 1, 2, 3$, respectively.

	Expression state				
mBAF state	Loss	Del	Neu	Gain	Amp
Loss	Loss	Del	CNLoH	Gain	Amp
Neu	Loss	Del	Gain	Amp	Amp
Del	Loss	Del	Neu	Gain	Amp

Table 3.1: Deterministic mapping between expression-based and mBAF-based states to allele-specific copy number configurations.

3.2.4 Subclone identification from copy number states

To increase confidence in the inferred copy number profiles, the successive step is to infer the subclonal structure of the tumour. Since cells within the same subclone are expected to share consistent SCNA patterns, inference is computed from the matrix of copy number calls **CN**. For this purpose, an approach inspired by the recent method PICASSO [93] is adopted, which infers subclones from single-cell copy number calls. PICASSO is particularly suitable for noisy single-cell copy number calls such as those in **CN**, where uncertainty arises from technical variability, dropouts, and indirect inference of SCNA. The algorithm also infers hierarchical relationships among subclones, producing a phylogenetic tree representing their evolutionary history. A detailed description of the original algorithm and its extension to the proposed multi-step apprach is provided in Appendix A.6. However, a brief presentation of the method and its application are now presented.

PICASSO operates recursively on the copy number call matrix within a tree-based decision algorithm, where each leaf node represents a candidate clone, i.e. a group of cells with similar copy number states. At each iteration, the algorithm evaluates whether a leaf should be subdivided into two descendant subclones. This decision is based on an EM step that estimates the posterior probability of each cell belonging to one of the two potential subclones, and a Bayesian Information Criterion (BIC) which balances goodness of fit and model complexity.
If a split is supported, cells are probabilistically assigned to one of the two subclones, otherwise, the leaf is declared terminal. This recursive process continues until no further subdivisions are warranted. Terminal leaves define the inferred subclones, while the recursion itself induces a phylogenetic tree describing their evolutionary relationships.
In this work, PICASSO is extended to explicitly model additional copy number states, including amplifications, deletions, and CNLoH, allowing integration of allele-aware information. The BIC penalty term is set to 2.5 to favor parsimonious models and reduce overfitting. Applying the extended framework to **CN** yields a partition of malignant cells into subclones $S_1, \ldots, S_N$.

3.2.5 Final Copy Number Calls

The final step of the algorithm concerns copy number calls at both single-cell and subclone levels. This procedure involves three processing steps.
First, consistent with the original SCEVAN framework, the expression matrix is re-segmented using VegaMC, restricting the analysis to each subclone. The β parameter is set to 0.5 to obtain a fine segmentation, enabling the detection of subclone-specific breakpoints.
Second, segment-level mBAF matrix is recomputed at the subclone level, where SNP-level allelic counts are aggregated across all cells within the same subclone. This follows the iterative clustering strategy described above, but using a single subclone-defined cluster, i.e. all cells in the subclone are treated as a pseudobulk, to generate subclone-specific mBAF profiles.

Finally, segment-level copy number inference is performed independently for each subclone using the joint expression and mBAF framework described in Section 3.2.3. This results in a refined copy number call matrix at single-cell resolution. Subclone-level SCNA profiles are then derived via majority vote among the constituent cells, as in the original SCEVAN pipeline previously described in 3.1.2.

3.3 Preliminary Results

This Section presents the results obtained using the proposed approach, with the aim of assessing its feasibility and practical performance. Both real and generated synthetic datasets were considered for the analysis.
First, a description of the real datasets under analysis is reported, together with results obtained using the original SCEVAN framework as a baseline.
Subsequently, the outputs of the main steps of the proposed pipeline are illustrated, highlighting how each process contributes to the final inference.
Finally, results are compared with established state-of-the-art methods, namely Numbat and XClone, to assess consistency and differences in the inferred copy number profiles, since an explicit ground truth is not available. In particular, Numbat was run directly on the same datasets using the recommended parameter settings, while results for XClone were taken from the original published studies. This choice is motivated by the fact that the published results represent the reference performance reported by the authors under optimized conditions, and therefore provide a meaningful benchmark for comparison, without introducing additional sources of variability related parameter tuning.
Finally, results obtained on a synthetic dataset generated through controlled modifications of one of the real samples are presented. In this case, ground-truth copy number states are available, and this enables a direct assessment of the method's performance under controlled conditions. These results are organized by first presenting empirical evidence, followed by comparison with the Numbat outcomes. Together, these analyses highlight the strengths

and limitations of the proposed framework, and shape directions for future methodological development.

3.3.1 Performance on real datasets

Dataset description

Two samples were selected to assess the performance of the proposed method on real single-cell RNA sequencing data. Specifically, a triple-negative breast cancer (TNBC) sample and an anaplastic thyroid cancer (ATC) sample were considered. For both samples, data were retrieved from the Gene Expression Omnibus (GEO, accession numbers GSM4476486 and GSM4476491) [94]. These data were obtained using 3' scRNA-seq on the 10x Genomics platform, which represents a high-throughput method for profiling gene expression across thousands of individual cells simultaneously [95]. In particular, both raw gene-by-cell count matrices and aligned reads in BAM format were directly retrieved from the GEO repository. BAM files were subsequently employed only for allele-specific pileup and phasing, using the *pileup_and_phase.R* script from the Numbat repository (Github: `https://github.com/kharchenkolab/numbat`). In the original dataset, the TNBC sample included 1097 cells, while the ATC sample included 2203 cells.

SCEVAN results

In this section, the results obtained with the original implementation of SCEVAN are presented.
For the TNBC sample, SCEVAN detects 797 malignant cells across a total of 1097 cells. A total of five tumoral subclones are identified, each carrying different copy number aberrations. SCEVAN detects copy number amplifications on chromosomes 1q, 6p, 8, 16p, and 18q in all cells, together with deletions on chromosomes 4q, 8p, 14, and 15p. The most prominent differences are detected on chromosome 7, 3p, and 11, where distinct clones display heterogeneous copy number states. Figure 3.3 shows the final subclone segmentation together

with the copy number calls. The bluest colors denote copy number losses, orange shades denote amplifications, and white segments indicate neutral copy number states.

Concerning the ATC sample, SCEVAN identifies a total of five subclones, comprising 945 malignant cells out of 2,203. Amplifications on chromosomes 11 and 20 and losses on chromosome 13 are shared across subclones. Major differences are observed on chromosomes 7, 17, and 9: three clones share a loss on chromosome 17, while one clone exhibits an amplification. On chromosome 7, one of these clones also carries this amplification, whereas the remaining clones display a neutral state. Results are shown in Figure 3.4. The color scheme is identical to that used previously.

The number of clones inferred by SCEVAN may be overestimated in both TNBC and ATC samples, as multiple clones share similar aberrations. Importantly, the number of malignant cells identified is consistent with results reported in literature by previous analyses using alternative tools, such as CopyCat and XClone. An allelic-specific framework could improve subclonal detection by enabling more precise copy number calls.

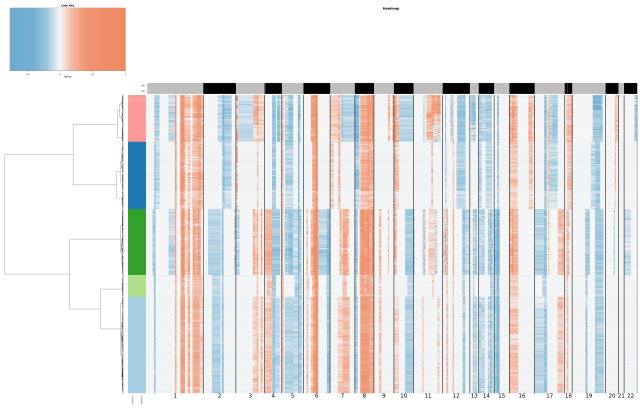

Figure 3.3: Subclonal somatic copy number aberrations identified by SCEVAN in the TNBC sample. Rows correspond to single cells and columns to genomic segments ordered along the chromosomes. Colors represent inferred copy number states, with blue indicating losses, orange indicating gains, and white denoting copy-number neutral regions. Five subclones are identified. Shared aberrations include amplifications on chromosomes 1q, 6p, 8, 16p, and 18q, and deletions on chromosomes 4q, 8p, 14, and 15p.subclone-specific differences are mainly observed on chromosomes 7, 3p, and 11.

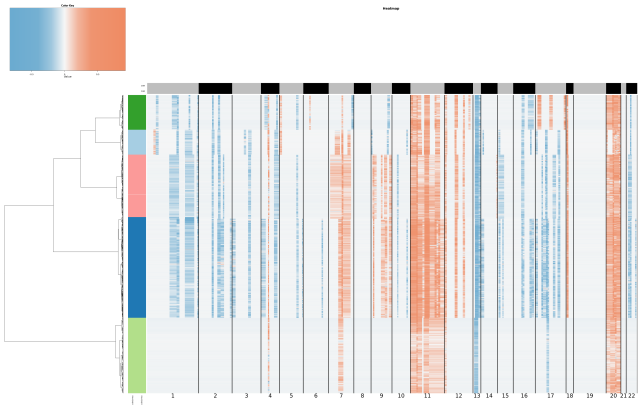

Figure 3.4: Subclonal somatic copy number aberrations identified by SCEVAN in the ATC sample. Rows correspond to single cells and columns to genomic segments ordered along the chromosomes. Colors represent inferred copy number states, with blue indicating losses, orange indicating gains, and white denoting copy-number neutral regions. Five subclones are identified. Shared alterations include amplifications on chromosomes 11 and 20 and losses on chromosome 13, while subclone-specific differences are mainly observed on chromosomes 7, 9, and 17.

Analysis results using the proposed framework

Preprocessing results Prior to the major steps of the pipeline, the preprocessing step involves the removal of SNPs with low sequencing depth (DP). This procedure represents a double-edged sword: on the one hand, it increases the reliability of the results, while on the other hand, it inevitably leads to the loss of information.

Figure 3.5 shows the distribution of SNPs across chromosome bands before and after applying a DP > 2 filter for the TNBC sample (upper panel) and ATC sample (lower panel) to identified malignant cells only, as described in the preprocessing step in Section 3.2.1. As illustrated, no SNP is shared across all cells, and the majority are present in a limited number of cells, typically fewer than 50. After filtering, this number is substantially reduced to a range of 0-10 in both samples, indicating that no more than 10 cells carry a given SNP.

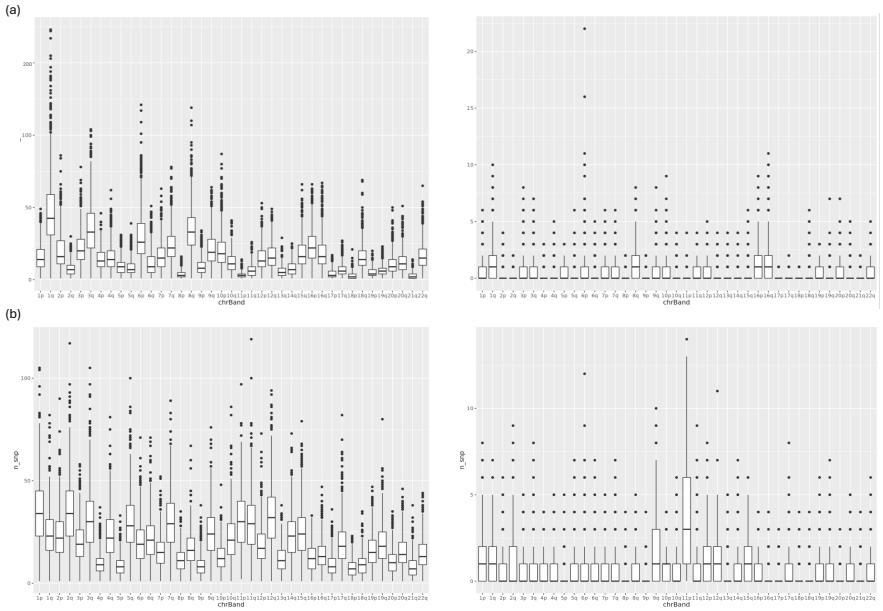

Figure 3.5: (a) Upper Panel: SNP distribution across chromosome bands in TNBC sample. The left subpanel displays all detected SNPs, while the right subpanel shows only SNPs with a read depth (DP) greater than 2. (b) Lower Panel: SNP distribution across chromosome bands in ATC sample. The left subpanel displays all detected SNPs, while the right subpanel shows only SNPs with a read depth (DP) greater than 2.

After cell filtering, 157 cells were retained for the TNBC sample and 436 cells for the ATC sample. The information carried by the filtered SNPs was then used to construct the chromosome-band-level mBAF matrix **M**, shown in Figure 3.6 (left panels (a) and (c)). In these heatmaps, grey shades indicate missing data, corresponding to chromosome bands with no detected SNPs, and have no quantitative meaning.
Chromosome bands excluded due to low SNP coverage involved chromosomes 2, 5, 8, 9, 11, 13, 14, 15, 17, 18, and 21 in the TNBC sample (upper panel), whereas in the ATC sample (lower panel) bands on chromosomes 5, 8, 9, 13, 15, 16, 17, 18, 20, 21, and 22 were filtered out. The right panels display the preprocessed matrices, $\tilde{\mathbf{M}}$, obtained after filtering out these low-coverage chromosome bands. The resulting mBAF heatmaps reveal a clear primary pattern, with copy number losses readily detectable. In the TNBC sample, deletions are visible on chromosomes 1, 10q, 11, and 19, whereas in the ATC sample a deletion is evident on chromosome 17. Moreover, a distinct cluster of cells in the ATC sample sharing a deletion on chromosome 7 can also be identified.
Panels (b) and (d) of Figure 3.6 display the segmented expression matrix **C** for the two samples, with a color scheme consistent with the previous figures described within the SCEVAN pipeline. In the TNBC sample, segment-level expression values clearly indicate the presence of two potential subclones, with major differences on chromosomes 3, 7, and 11. In the ATC sample, the most prominent differences are observed on chromosomes 7, 9, and 17, suggesting the presence of three distinct clusters of cells.
Based on segment-level expression patterns, two distinct clusters are anticipated in the TNBC sample, whereas three clusters are expected in the ATC sample. These matrices constitute the input for the MoNETA embedding.

Multi-modal integration and clustering based mBAF The preprocessed matrices **C** and $\tilde{\mathbf{M}}$ shown in the previous section were converted into graph-based networks, which serve as the foundation for the embedding space obtained by

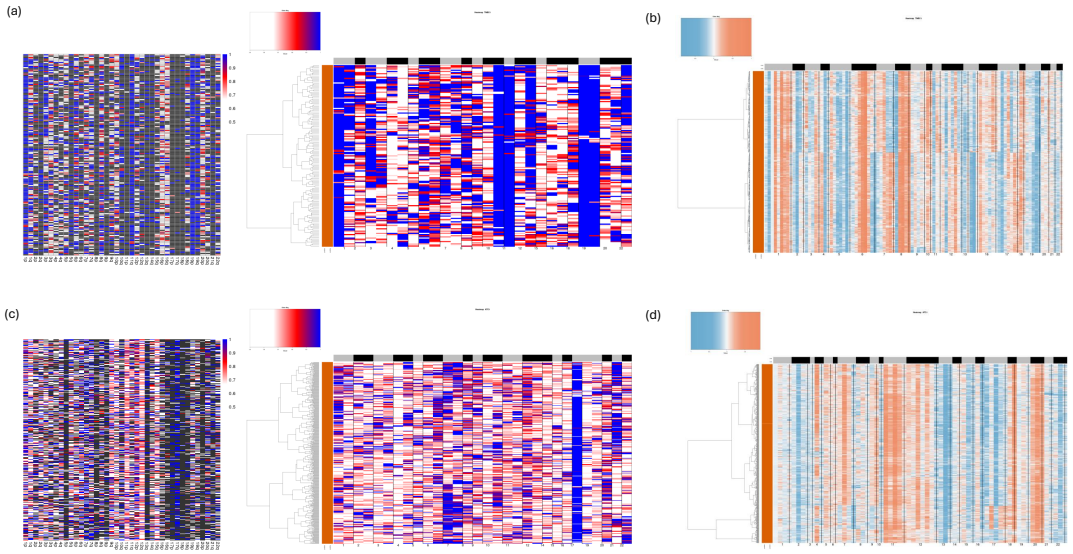

Figure 3.6: Panels (a) and (c) illustrate heatmap of mBAF values for the TNBC and ATC samples across chromosome bands, respectively. Rows represents individual cells and columns corresponds to chromosome bands. Left figures represent pre-filtering mBAF matrix where grey areas indicate bands where no detected SNPs. Right panel shows values after applying a K-nearest neighbors smoothing step with $k = 10$ to bands covered by at least the 30% of cells. Bluer color are related to loss or deletion, while oranger are correlated to gains and amplification. Panels (b) and (d) show expression segmentation for the TNBC and ATC sample across malignant cells, respectively

MoNETA application. As shown in Figure 3.7, the segment-level expression graph of the TNBC sample (left (a) panel) reveals two well-defined communities, whereas the ATC sample exhibits more than two distinct communities (left (b) panel). In contrast, the mBAF graph representations do not show a clear separation of clusters (right panels).
This figure indicates that mBAF contributes to a reduced extent to the separation of cell subpopulations in comparison with gene expression. This phenomenon can be explained by the fact that the mBAF signal is more subject to noise than the expression profile, which exhibits more distinct similarity patterns, as shown in Figure 3.6. Consequently, the final pattern between the two modalities is determined primarily by expression, while mBAF provides additional but less decisive information.
Integration of omics data produces, for each sample, a cell–cell similarity matrix computed from genome-wide signals. A single, global UMAP projection of this matrix is then used to obtain a two-dimensional representation of the cellular landscape, as shown in Figures 3.8 and 3.9. In this embedding, each point corresponds to a cell, while the color scale varies across panels to highlight either mean expression (upper panels) or median mBAF for selected chromosome bands (lower panels). The chromosome bands displayed were chosen based on major SCNAs identified by SCEVAN (chromosomes 3, 7 and 11 for the TNBC sample, and chromosomes 7 and 17 for the ATC sample).
In the TNBC sample, the UMAP embedding clearly identifies two subclones with distinct expression and mBAF profiles. For instance, at chromosome band 3p, all cells display a neutral mean expression profile, whereas a subset of cells exhibits a loss state in the mBAF signal (highlighted in red). On chromosome band 7p, both neutral and loss expression states are observed across the two subclones, while the mBAF signal does not show clear discriminative patterns between them. Finally, chromosome band 11q is characterized by neutral expression levels combined with a loss in mBAF, suggesting a potential copy-neutral loss of heterozygosity (CNLoH) event.
In the ATC sample, expression profiles reveal visually distinguishable subpop-

ulations. The majority of cells exhibit amplifications on chromosome 7 and either neutral or loss states on chromosome 17 in their expression profiles, whereas a minority subpopulation displays the opposite pattern. Consistently, mBAF profiles reflect these differences, with deletions observed in regions where expression remains neutral. By jointly considering both expression and mBAF signals, three putative subclones can be inferred.

Finally, the multi-modal integration leads to the iterative clustering step, whose result is the segment-level mBAF matrix, $\mathbf{M}^{\text{final}}$, which is derived from detected cell-to-cell similarities. Figure 3.10 illustrates the resulting matrix for both samples. In the TNBC sample (left panel), losses are observed across chromosome bands including 1p, 11, and 19p, shared by all cells, while aberrations on chromosomes 3, 6, 12, and 22 indicate subclonal structure. Moreover, a possible clonal amplification stated is detected across chromosome 8. In the ATC sample (right panel), a cluster of cells shows deletions on chromosomes 7, 8, 9, and 17. Notably, many chromosomes and cells are inferred to be in a neutral state, while a possible amplification stated is detected across chromosome 11 and 20. These differences enable the preliminary identification of distinct subclonal populations. These segment-level mBAF results display patterns consistent with the SCEVAN results, particularly shared amplifications and specific losses.

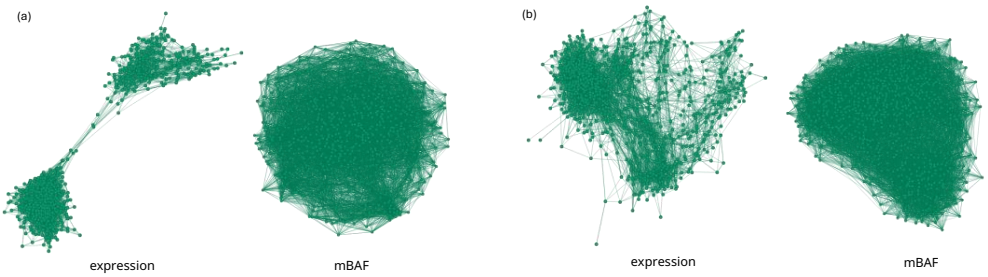

Figure 3.7: Single-network representation plots for each omic layer (expression and mBAF). Each node represents a cell, and each edge represents a connection between cells. Cells positioned farther apart in the network share fewer similarities, whereas cells located closer together have more features in common. Panels (a) and (b) show the single-network representations of the TNBC and ATC samples, respectively, for expression (left) and mBAF (right). Cell-cell similarities are more clearly detectable in the expression layer than in the mBAF layer. In particular, two communities are clearly visible in the TNBC expression graph, whereas more than three communities can be observed in the ATC expression graph.

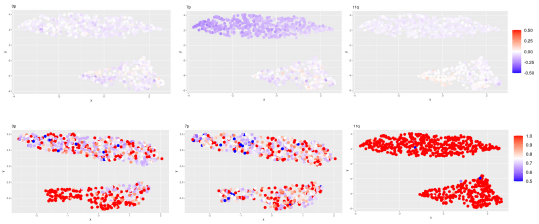

Figure 3.8: UMAP projection of the MONETA similarity matrix for the ATC sample. Each point represents a single cell in a global embedding computed once for the entire genome-wide similarity matrix. Colors vary across panels to highlight either mean expression (upper panels) or median mBAF for selected chromosome bands (lower panels). Two distinct subclones are clearly identified, showing different expression and mBAF profiles. At chromosome band 3p, one subclone displays neutral states, while chromosome 7 shows neutral and amplified states across the two subclones. Color coding follows the same palette described in the previous figures.

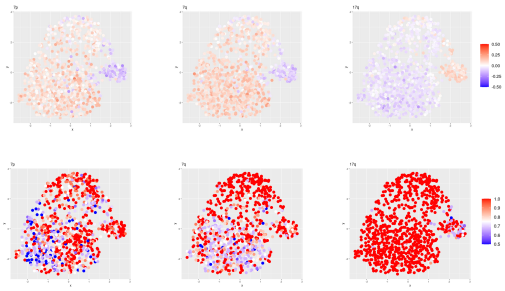

Figure 3.9: UMAP projection of the MONETA similarity matrix for the ATC sample. Each point represents a single cell in a global embedding computed once for the entire genome-wide similarity matrix. Colors vary across panels to highlight either mean expression (upper panels) or median mBAF for selected chromosome bands (lower panels). Expression profiles reveal two distinct subpopulations: a major cluster showing amplification on chromosome 7 and neutral or loss states on chromosome 17, and a minor cluster exhibiting the opposite pattern. mBAF profiles consistently reflect these differences, highlighting deletions in regions that appear neutral at the expression level. Color coding follows the same palette used in previous figures.

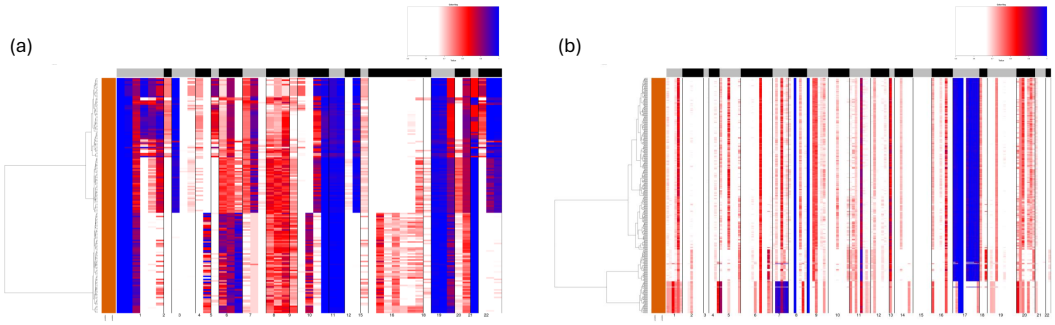

Figure 3.10: Heatmap representing the segment-level allelic information for the TNBC sample (a) and the ATC sample (b). Each row represents a cell, while each column represent a chromsome. Cells order is determined by expression-level similarity. Color coding follows the same palette described in the previous figures.The dendrogram on the left represents the hierarchical clustering of cells based on their gene expression profiles.

Subclones Identification Starting from the preprocessed expression matrix $\tilde{\mathbf{X}}$ and the derived final segment-level mBAF matrix $\mathbf{M}^{\text{final}}$, two-state posterior probabilities were computed to obtain the joint posterior matrix $\mathbf{P}^{\text{final}}$. For

each cell and genomic segment, the most probable copy number state was then selected, yielding final copy number call matrices for the TNBC and ATC samples (Figure 3.11, left panels). Notably, this joint posterior analysis reveals CN-LOH events that were not detected by SCEVAN, highlighting its ability to capture additional copy number alterations.

Although cell similarity patterns are already apparent at this stage, the resulting signal remains noisy, making direct interpretation challenging. This motivates the application of PICASSO (Section 3.2.4), which exploits these emerging similarity patterns to robustly infer subclonal structure. The chosen penalty term equal to 2.5 for the BIC value is reasonable in light of the high noise level in the data. Notably, this step differs from the MoNETA-based analysis, which focuses on inferring cell-cell similarities, whereas here the objective is to determine the number of subclones and their cellular composition.

As shown in the right panels of Figure 3.11, PICASSO effectively separates cells according to their copy number profiles, identifying two subclones in the TNBC sample and four in the ATC sample. Each subclone is characterized by internally consistent copy number profiles that are not shared by the others. In the TNBC sample, the two subclones mainly differ at segments 21, 51, and 146, whereas in the ATC sample distinct alterations are observed across segments 46, 61-66, and 126-131.

Overall, the subclonal structure inferred by PICASSO is fully consistent with observations from the previous analysis steps, supporting the reliability and coherence of the proposed inference framework.

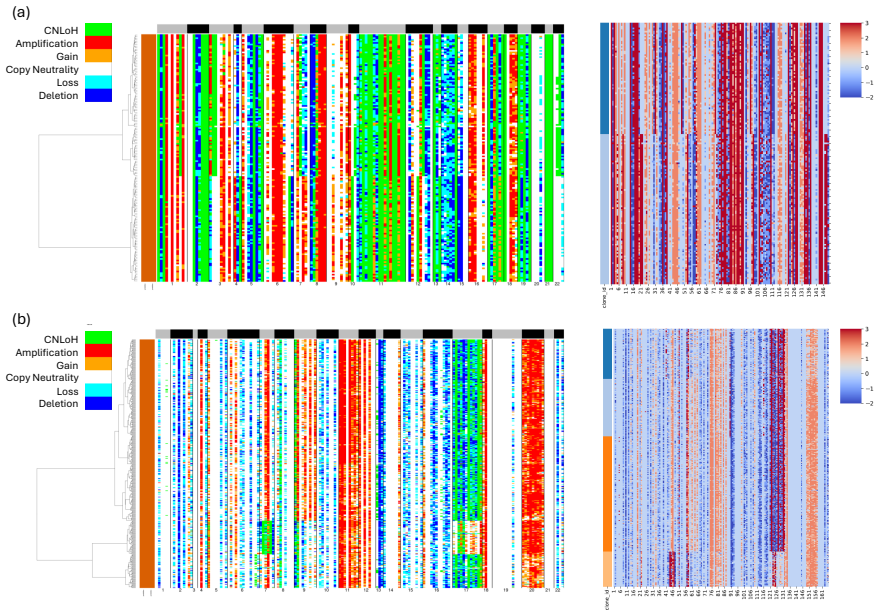

Figure 3.11: Subclone identification. Panels (a) and (b) show results for the TNBC and ATC samples, respectively. The left panels display the initial copy number calls, where each row represents a single cell and each column corresponds to a genomic segment across all chromosomes. Colors indicate copy number states: blue for loss, light blue for deletion, white for neutral, orange for gain, red for amplification, and green for CN-LOH. The right panels show subclone assignments obtained from PICASSO. Rows represent individual cells and columns correspond to genomic segments. Cells are ordered according to copy number similarity, and the color-coded bar on the left indicates subclone membership. Colors represent discrete copy number states as follows: -2 for Loss, -1 for Deletion, 0 for Neutral, 1 for Gain, 2 for Amplification, and 3 for CN-LOH. Two subclones are identified in the TNBC sample and four in the ATC sample.

Final allele-specific SCNA calls The final step of the analysis consists of single-cell and subclone-level copy number calling. First, a finer segmentation parameter was employed ($\beta = 0.5$) to capture subtle variations within each subclone from expression data. This choice of β provides a fine resolution, while higher values could result in overly large segments, potentially masking subclonal differences. Afterwards, single-cell and subclone-leve with majority vote final copy number calls are obtained, and displayed in Figure 3.12. These

results reveal substantial differences compared to the original SCEVAN calls. In particular, the refined analysis allows the detection of CN-LOH events that were previously undetected, providing a more detailed and accurate representation of subclonal architecture.
Additionally, fewer subclones are inferred, suggesting that some of the subpopulations detected in the initial analysis were likely artifacts arising from coarser segmentation or noise in the data.
The most prominent differences in the TNBC sample detected by SCEVAN were located on chromosome 7, 3p, and 11, where distinct subclones exhibited heterogeneous copy number states, and are preserved within these results. Additional CN-LOH events, previously undetected, are now observed on chromosomes 1 and 19.
For the ATC sample, SCEVAN originally identified major alterations on chromosomes 7, 17, and 9. In the final subclonal-level analysis, these chromosomes continue to display differences among subclones, now including both CN-LOH events and deletions.
This provides a more comprehensive view of subclonal heterogeneity and highlights the added resolution obtained through the joint expression-mBAF framework. This confirms the persistence of previously detected alterations while providing a more refined characterization of their subclonal distribution. Overall, this analysis provides a more comprehensive view of subclonal heterogeneity, confirming previously detected alterations while offering a more refined characterization of their subclonal distribution. These results highlight the increased resolution afforded by the allele-specific framework compared to the original approach, enabling a more accurate reconstruction of tumor heterogeneity.

Comparison with state-of-the-art tools

This section compares the results obtained with the proposed allele-specific version of SCEVAN to those produced by two widely used allele-specific scRNA-seq analysis tools, Numbat and XClone. In the absence of ground-truth

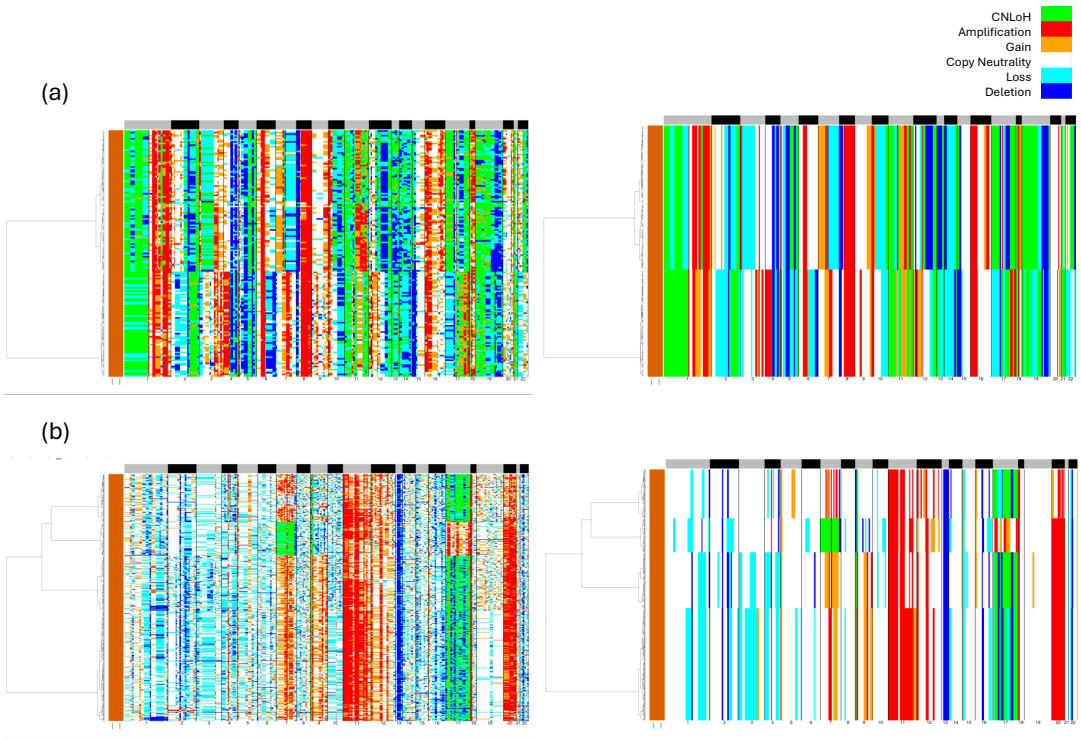

Figure 3.12: Final copy number call at the single cell (left) and subclonal (right) level. Panel (a) shows the TNBC sample, and panel (b) shows the ATC sample. The color scheme is as follows: blue for loss, light blue for deletion, white for neutral, orange for gain, red for amplification, and green for CN-LOH. These final calls, obtained using a finer segmentation parameter ($\beta = 0.5$), reveal CN-LOH events undetected by SCEVAN and provide a more accurate representation of the subclonal architecture.

SCNA profiles, as well as bulk or single-cell DNA sequencing data for direct validation, this section offers an appropriate framework for benchmarking the proposed method against existing approaches.

Numbat was run using its default parameters, as specified in the official GitHub repository and user guide, to ensure a fair and reproducible comparison reflecting standard usage. As anticipated in the introduction of this section, results reported in the literature were used for XClone.

The comparison is carried out both qualitatively and quantitatively. Qualitative assessment is based on the inspection of graphical representations of

copy number profiles, whereas quantitative evaluation relies on cell-subclone assignments assessed using two widely adopted clustering metrics: the Jaccard Index and the Adjusted Rand Index (ARI) [96]. The Jaccard Index measures the proportion of pairs of observations assigned to the same cluster in both partitions, while the ARI evaluates the agreement between two clusterings corrected for chance. Both metrics were computed to assess the concordance between the proposed framework and the results obtained with Numbat .
To enable a fair comparison of subclonal structures, only malignant cells identified by the proposed framework were considered, and cells classified as normal by Numbat were excluded. In addition, clusters containing fewer than 20 cells were removed, as they were deemed non-informative.

Results on the TNBC sample Numbat and XClone infer different subclonal structures and SCNA patterns in the TNBC sample. Numbat identifies a total of 12 subclones, whereas XClone reports only two major subclones. While a high number of inferred subclones may suggest increased sensitivity, several of the subclones defined by Numbat consist of a very limited number of cells, as will be discussed. In the literature, subclones are generally considered biologically meaningful only when containing a sufficiently large number of cells: for example, PICASSO requires subclones to include more than 60 cells to ensure robustness [93].
Figure 3.13 shows the allele-specific SCNA profiles inferred by Numbat for each subclone. Here, no explicit normal cell population is identified, since the largest subclone, comprising 303 cells, displays exclusively normal states and CN-LOH events. However, these events are not supported by the parental haplotype frequency profiles, as SNP values remain centered around 0.5 rather than approaching 0 or 1. This suggests that this subclone likely corresponds to a population of normal cells rather than a tumour subclone. Other inferred subclones consist of very few cells, such as subclones 4 and 7, which include only 17 and 33 cells, further questioning their biological relevance. In addition, several subclones (from subclone 9 to subclone 12 and subclones 2 together

with 5, 6, 7, 8) exhibit highly similar SCNA patterns and may represent the same underlying population.

Moreover, some inferred SCNAs appear to be driven primarily by gene expression rather than allelic evidence. In several regions classified as neutral or amplified, parental haplotype frequencies do not display the characteristic patterns expected under allelic imbalance. In particular, SNP values fail to cluster around canonical states (e.g., 0, 0.5) and instead appear distributed across the interval. For instance, chromosome 16 in cluster 10 is inferred as neutral despite sparse parental haplotype signals.

In other regions, a very limited number of informative SNPs are used to infer the underlying copy number state (e.g., as few as three), which prevents a reliable estimation of haplotype imbalance (e.g., subclone 12 on chromosomes 11, 13, and 14). The limited availability of informative SNPs, together with potential phasing errors leading to incorrect heterozygous calls, likely contributes to inaccurate SCNA assignments and may ultimately affect the reliability of subclonal inference.

In contrast, XClone applied to the TNBC sample reconstructs a much simpler subclonal architecture, with several regions exhibiting CNLoH, that are supported by differential expression analysis, as shown in figure 3.14. One subclone shows CNLoH on chromosomes 1p and 21, as well as strong allelic bias on chromosomes 11p, 13, 14, 17, and 19, while the second subclone exhibits CNLoH on chromosomes 3p and 22. These allele-specific alterations are consistent with literature reports and highlight the advantage of approaches that explicitly integrate allelic information. Figure 3.14 illustrates the strong qualitative agreement between the proposed framework and XClone results. In particular, CN-LOH events are concordant at the clonal level on chromosomes 1, 19, and 21, and at the subclonal level on chromosome 22.

However, this comparison reflects fundamental differences between the approaches. The proposed framework performs inference on a more restricted set of high-quality cells and informative SNPs, prioritizing confidence over coverage. In contrast, Numbat and XClone employ all available SNPs and cells,

potentially increasing sensitivity at the cost of precision.
From a quantitative point of view, the two subclones identified and described by the proposed allele-specific framework partially overlap with the ones defined by Numbat. For simplicity of notation, the identified subclones using the poposed approach are now called Subclone 1 and Subclone 2. Table 3.2 summarize the distribution of cells across the 12 subclones identified by Numbat for the two subclones. The fourth subclone identified by Numbat is removed due to the limitated number of cells. Most cells of Subclone 1 are

Table 3.2: Distribution of the two TNBC subclones identified by the proposed framework across Numbat-defined clones.

TNBC Numbat Subclones	1	2	3	5	6	7	8	9	10	11	12	TOT
Subclone 1	0	0	0	2	0	0	0	8	8	47	2	67
Subclone 2	1	1	16	36	19	6	9	0	0	0	0	80

assigned to Numbat subclone 11, while the remaining cells are distributed across clones 5, 9, 10, and 12. Subclone 2 overlaps primarily with clone 5, with smaller contributions to other clusters. The presence of a small number of cells (two cells in Numbat clone 5 for Subclone 1, one cell in clone 1 and 2 for Subclone 2) for Subclone 1, should not be considered a major discrepancy, since these represent only a minor fraction of the total cells correctly assigned. When comparing the clustering obtained with the proposed framework to Numbat's results, Numbat's first and fourth subclones are removed, due to low quality. A Jaccard index of approximately 0.35 and an ARI of approximately 0.34 were observed, indicating a moderate to low level of concordance between the two clustering solutions.

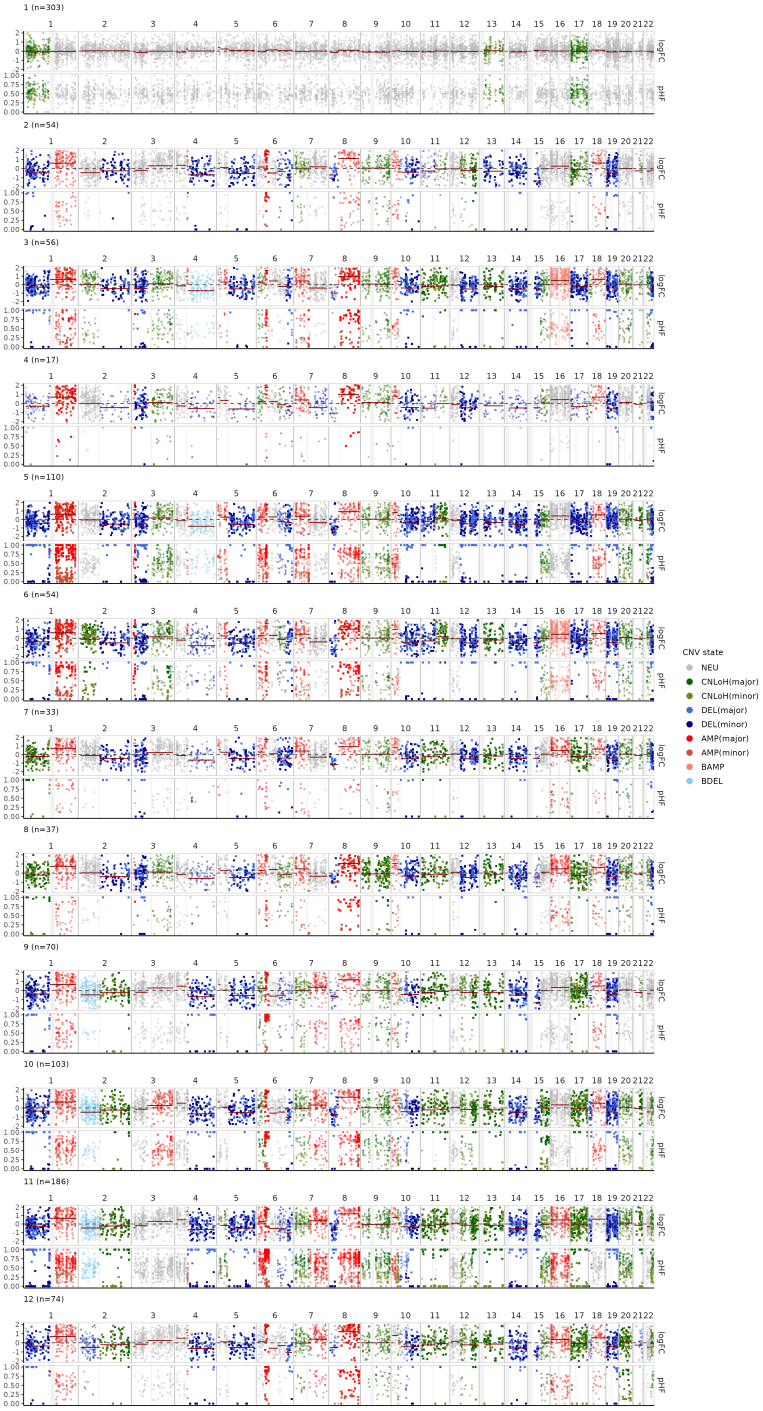

Figure 3.13: Pseudobulk SCNA profile of the detected sublclones. Each dot represents a different gene (upper figure) or SNP (lower figure), whose log expression fold-change (logFC) and paternal haplotype frequency (pHF) are represented in upper and lower graphs, repspectively. Gray vertical bars represent centromeres and gap regions.

However, after merging Numbat subclones based on prior observations of clonal similarity, the agreement between the two approaches increased substantially, with an ARI of approximately 0.93 and a Jaccard index of approximately 0.94, indicating a very strong concordance and an almost identical subclonal structure.

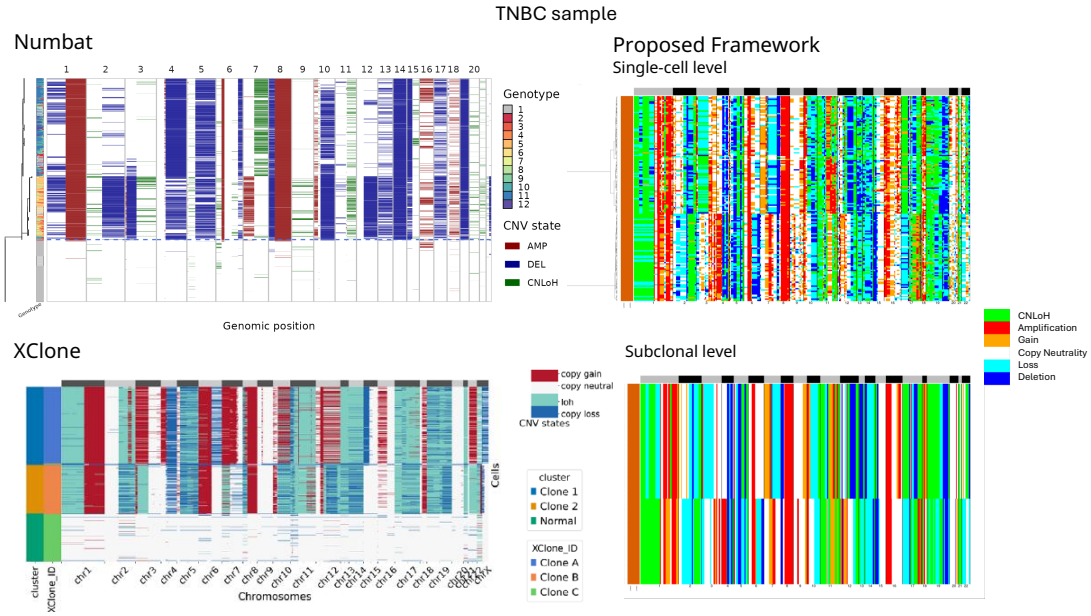

Figure 3.14: The heatmaps provide a visual representation of the SCNAs generated by Numbat, XClone and the proposed framework. XClone results are taken from Huang et al.[82].

Results on the ATC sample For the ATC sample, the comparison is limited to Numbat, as no XClone results are available in literature.
Numbat identifies a total of four subclones, whose SCNA patterns are displayed in Figure 3.15. The first subclone comprises only normal cells, since no SCNA have been detected. Moreover, this subclone is represented by the majority of cells in the sample. The second subclone consists of 28 cells and, since this is a considerably low number, it is removed frome the analysis. The remaining two subclones show different and mirrored SCNA patterns. Specifically, one subclone shows a CNLoH on chromosome 7 and an amplification on chromosome 17, while the other presents CNLoH on chromosome 17 and

an amplification on chromosome 7. Additionally, the q-arm of chromosome 9 reflects the difference among the subclones, showing an amplification in one subclone and a neutral state in the other.
The results obtained with the proposed framework and presented in the previous section are consistent with those identified by Numbat, as depicted in Figure 3.16. However, small differences are detectable and might be attributed to the finer-grained segmentation adopted by VegaMC within the proposed approach. Specifically, an additional subclonal population is detected, characterized by a copy number profile not identified by Numbat, comprising copy number gains on chromosome 7, copy-neutral status on chromosome 9, and a combination of deletions and CNLoH affecting chromosome 17.
With regard to cell distribution, Table 3.3 summarizes the allocation of the four subclones across the clones defined by Numbat. As in the case of the TNBC sample, here Subclone 1-4 represent the four tumour subclones identified with the allele-specific version of SCEVAN. Subclones 3 and 4 identified by the

Table 3.3: Distribution of the first set of ATC subclones across Numbat-defined clones.

Subclone / Numbat clone	1	2	3	4	TOT
1	1	1	9	74	86
2	1	1	57	1	62
3	0	0	0	97	100
4	0	0	0	195	199

proposed framework show a perfect correspondence with Numbat subclone 4, suggesting a potential overestimation of the number of subclones by the proposed method. In contrast, subclone 2 predominantly overlaps with Numbat subclone 3. Only a limited number of cells are shared between Numbat clones 1 and 2, supporting the interpretation that these clusters likely represent normal cell populations.
For clustering comparison, only the third and the fourth clone identified by Numbat were considered, since they represent the most reliable results.
When comparing the clustering obtained with the proposed framework to

Numbat's results, a Jaccard index of approximately 0.40 and an ARI of approximately 0.24 were observed, indicating a moderate level of concordance between the two approaches. However, after merging Subclone 3 and Subclone 4, the agreement increased substantially, with an ARI of approximately 0.50, suggesting a markedly improved alignment between the two subclones detections.

In conclusion, the results obtained are highly comparable to those generated by Numbat and may provide improved resolution with respect to the estimated number of subclones and cell-to-subclone assignment.

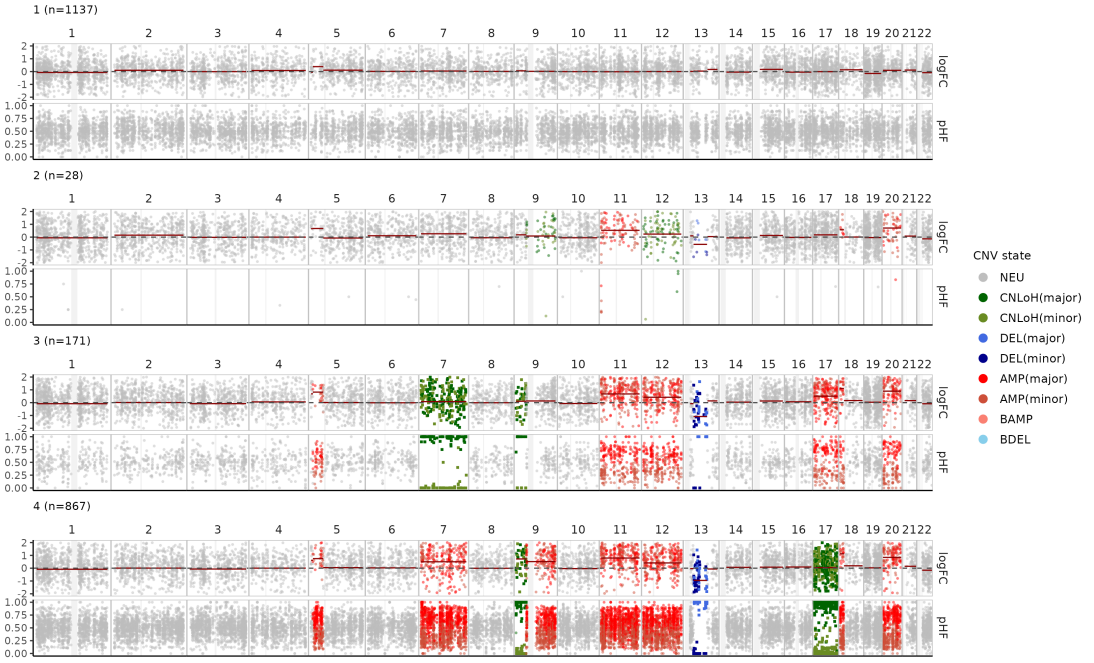

Figure 3.15: Pseudobulk SCNA profile of the detected sublclones. Each dot represents a different gene (upper figure) or SNP (lower figure), whose log expression fold-change (logFC) and paternal haplotype frequency (pHF) are represented in upper and lower graphs, repspectively. Gray vertical bars represent centromeres and gap regions.

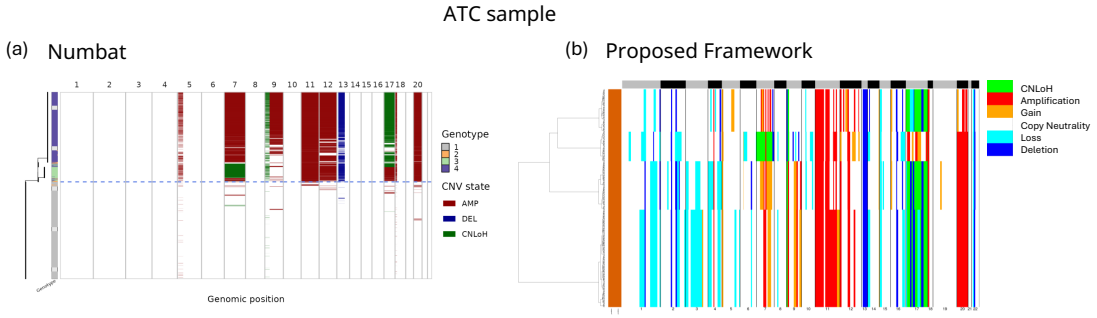

Figure 3.16: The heatmaps provide a visual representation of the SCNAs generated by Numbat (a) and the proposed framework (b).

3.3.2 Results on synthetic dataset

The benchmark of the results obtained on real data are strictly dependent on the inferences produced by other tools. For this reason, experiments on synthetic data are proposed in order to assess the strengths of the method and to identify potential directions for future improvements of the algorithm.
Specifically, a synthetic dataset generated from the ATC sample has been created. This choice is motivated by the relatively simple SCNA profile of the sample, which appears neutral across several genomic regions, in contrast to the more complex SCNA landscape observed in the TNBC sample. In particular a an artificial CNLoH event at chromosomal level has been introduced, since only allele-aware methods are expected to reliably detect such events. As both Numbat and the proposed framework consistently identified chromosome 19 as neutral, 100 cells were randomly sampled from Subclone 4 and retained for downstream analysis. For these cells, all SNPs mapping to chromosome 19 were assigned a loss configuration, with allelic depths set to either 0 or 10 and a fixed total read depth of 10. This DP value was selected to ensure a reliable sequencing coverage.
Figure 3.17 illustrates the results of the key steps of the proposed algorithm. As expected, the mBAF profile (Panel a) shows a clear loss of allelic balance, confirming that the synthetic CNLoH event was correctly introduced. Furthermore, this pattern differs markedly from the mBAF profile of the original data,

shown in Panel (b) of Figure 3.6.
Subsequently, the MoNETA approach is applied to jointly analyze the two omics layers. The similarity graphs derived from mBAF profiles (panel b, upper left) reveal that cells harboring the synthetic CNLoH event cluster closely together, indicating a clear separation of the altered population from the remaining cells. Consistently, the UMAP representation of the resulting embedding space on chromosome 19 (panel b, lower left) shows that cells carrying the CNLoH event occupy nearby regions of the embedding, reflecting their signal similarity.
At the segment-level mBAF resolution (panel b, right), an overestimated number of cells exhibit mBAF values close to 1, and cells harboring the true CNLoH event are not clearly separable from the remaining population. However, the modified version of PICASSO (panel c, right) is able to exploit the emerging similarity patterns to correctly infer the subclonal structure. Consequently, the final copy-number calls (panel d) successfully recover the synthetic CNLoH event on chromosome 19.
However, three out of the six inferred subclones are classified as harboring the CNLoH event on chromosome 19, suggesting that the algorithm may overestimate the number of subclones and fragment the underlying altered population. To better characterize this behavior, the distribution of cells belonging to the synthetic CNLoH subclone across the inferred subclones is analyzed.
Table 3.4 reports the total number of cells assigned to each subclone, showing that several subclones contain a limited number of cells and could potentially be merged. This behavior is consistent with previous observations in the literature, where biologically meaningful subclones are typically required to comprise a sufficiently large number of cells
As shown in Table 3.5, cells carrying the synthetic CNLoH event are mostly assigned to subclones 2, and 3, accounting for approximately more than a half of the cells in each of these subclones. This pattern indicates that the introduced alteration is consistently detected, but fragmented across multiple closely related subclones, suggesting an over-partitioning of the altered

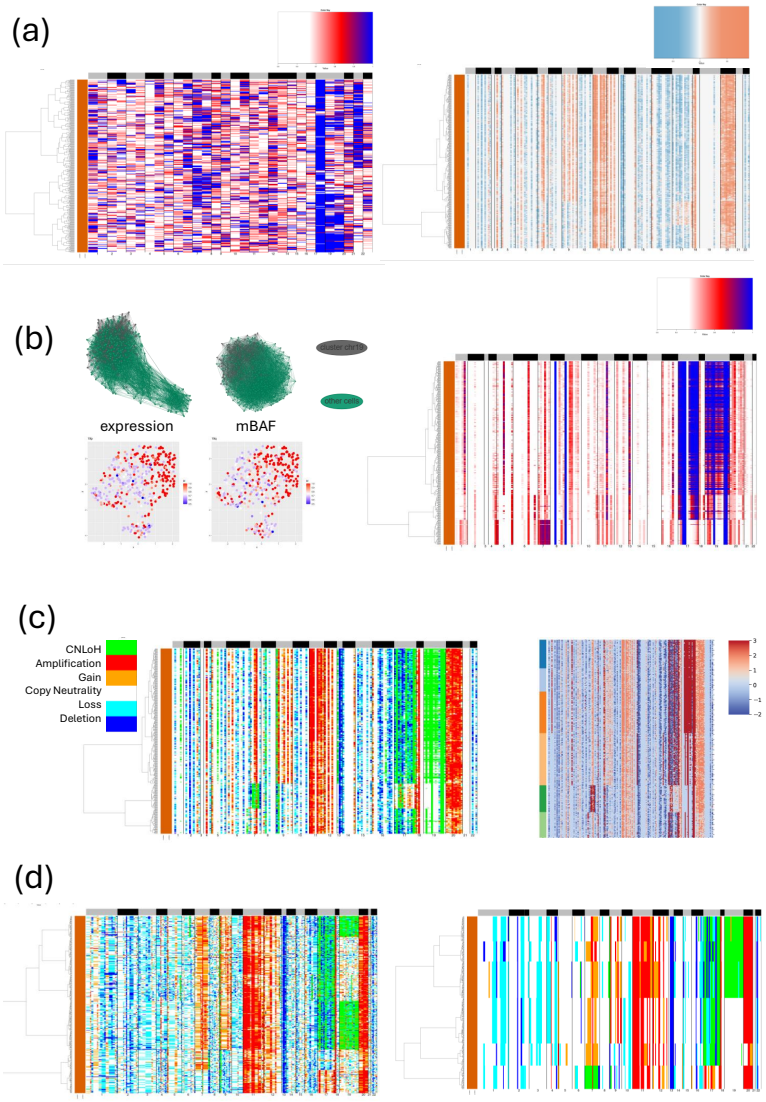

Figure 3.17: Overview of the proposed framework for allele-specific copy-number inference on synthetic data. (a) Heatmaps showing chromosome band-level mBAF (left) and segment-level gene expression (right). (b) Left panel: single-cell similarity graphs constructed from expression and mBAF profiles at the single-omics level. Cells belonging to the clone harboring alterations on chromosome 19 are shown in grey, while the remaining cells are highlighted in green. The lower-left panel displays the corresponding cell distribution in the UMAP embedding. The right panel shows segment-level mBAF profiles obtained after the iterative clustering procedure. (c) Left: initial copy-number calling. Right: subclonal structure inferred using the modified version of PICASSO. (d) Final copy-number call at single-cell (left) and subclone-level (right), highlighting the detection of a CNLOH event on chromosome 19. Colour scheme follows panel (c)

population rather than a failure to recover the signal.

Table 3.4: Number of cells in each subclone inferred from the CNLoH-positive population on chromosome 19

Subclone	1	2	3	4	5	6
Number of cells	63	51	92	115	59	56

Table 3.5: Distribution of CNLoH-positive cells across subclones

Subclone	1	2	3	4	5	6
Number of CNLoH-positive cells	11	28	59	2	0	0

Overall, these results show that the proposed framework can reliably detect allele-specific alterations and recover the underlying altered cell population, , even when it only affects the allelic signal. However, the number of subclones detected is higher than expected, and cells with the synthetic CNLoH do not belong exclusively to one or more subclones, but rather to subclones in which cells without this aberration are also present. This phenomenon can be explained by referring to the topology of the input network for MoNETA, particularly with regard to that of mBAF.

As presented in Figure 3.19, panel (b), the graph-based mBAF representation clearly indicates that the cells harbouring the synthetic alteration are located in close proximity to each other, yet do not form a separate cluster. The mBAF signal therefore shows a high level of similarity between cells across all chromosomes, with a tendency to bring them closer together rather than separate them. Consequently, in the iterative clustering phase, cells that are positioned close together in the resulting embedding will mix their signal with that of cells with CNLoH aberration, leading to false positives.

For comparison purposes, Numbat was applied to the synthetic dataset, and Figure 3.18 illustrates the results. These results are consistent with those presented in the previous section, showing an additional subclone harbouring the introduced CNLoH alteration. This subclone contains all cells harbouring the introduced SCNA, but also includes 30 more cells. As illustrated in panel

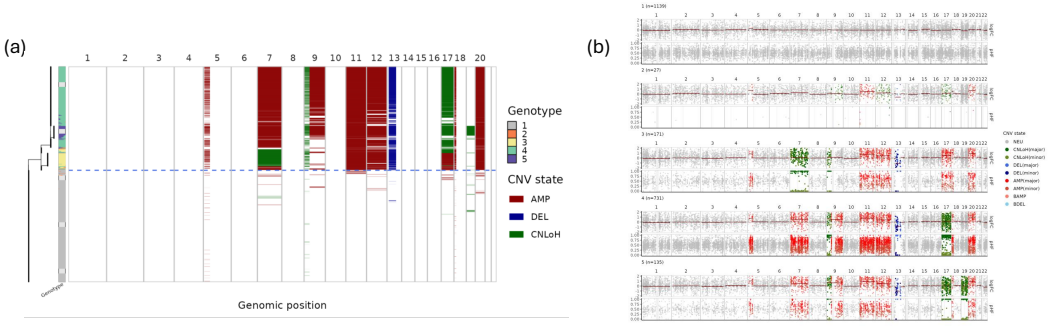

Figure 3.18: Numbat SCNA landscape and reconstructed phylogeny of synthetic sample. (a) Branch lengths correspond to the number of CNVs. Blue dashed line separates predicated tumor and normal cells. Genotype refers to subclone belonging. Three states are identified: amplification, deletion and CNLOH. profile. CNLOH is clearly detected on chromosome 19. (b) Pseudobulk SCNA profile of the detected sublclones. Each dot represents a different gene (upper figure) or SNP (lower figure), whose log expression (logFC) and paternal haplotype frequency (pHF) are represented in upper and lower graphs, repspectively. Gray vertical bars represent centromeres and gap regions. Subclone harbouring CNLOH over chromosome 19 comprises 135 cells.

(b), this is likely due to the fact that these cells had a parental allelic fraction of 0 or 1. This error may be due to two factors: first, these cells may have been mislabelled during the phasing step, and second, Numbat considers all SNPs and cells, including those with lower reliability.

3.4 Dicussion and future perspective

The proposed extension of SCEVAN shows promising potential in the field of omics sciences, enabling allele-specific SCNA detection within scRNA-seq frameworks. The proposed pipeline integrates seamlessly with the original SCEVAN implementation for total copy number analysis, following a precise strategy aimed at inferring SCNAs at an appropriate resolution.
This methodology integrates single-cell allele-specific information with gene expression data using an approach based on graphical multi-modal integration. This technique allows similarities between cells to be identified, providing a

more comprehensive understanding of the cellular composition and heterogeneity of tumours. This similarity is then used to identify feasible subclones and, through this implementation, it is possible to deduce the copy number status within single-cell and subclonal resolution.
This strategy employs and integrates cutting-edge methods in the field of scRNA-seq data analysis (MoNETA, PICASSO and DrImpute), specifically tailored to address the strengths and limitations of this technology, such as high dimensionality, sparsity, and technical noise.
While still under development, preliminary results on real datasets presented in Section 3.3 indicate that the method can detect allele-specific events with meaningful resolution, demonstrating performance consistent with expectations for current approaches. Comparisons with previously published tools, such as Numbat and XClone, further support the consistency of the proposed framework, particularly in the detection of tumors with low purity and complex SCNA profiles. In this analysis, the primary comparison was performed against Numbat, one of the most widely used state-of-the-art methods for allele-specific SCNA inference from scRNA-seq data. Numbat is known to provide highly accurate detection of SCNA patterns, although it tends to overestimate subclonality in certain settings. Notably, in genomic regions where the proposed approach diverged from Numbat, the resulting SCNA profiles showed strong agreement with those reported by XClone, indicating that these discrepancies likely reflect methodological differences rather than inconsistencies in the proposed framework.
From a computational perspective, the scalability of the proposed framework depends primarily on the number of cells, the number of SNPs used for allele-specific inference, and the dimensionality of the gene expression matrix. Increasing any of these components affects both the construction of the multi-modal similarity graph and the downstream subclone inference. Future work will focus on characterising the runtime behaviour of the method and on implementing computational optimisations to further improve its scalability.
Moreover, further research is necessary to overcome the current limitations of

the present methodology.
First, the performance of the method benefits from datasets with a high number of cells and sufficient coverage, and its application to low-quality samples may compromise subclonal inference. This limitation is partly driven by the filtering strategies applied during the preprocessing step. As a mitigation approach, users may relax filtering criteria, rely on unfiltered data when appropriate, or integrate multiple datasets obtained from the same patient and time point, as is common in multifocal studies, to increase coverage and robustness.
Secondly, as illustrated in Section 3.3.2, experiments on the synthetic dataset show an overestimation of the number of subclones and an inaccurate definition of their composition. As previously hypothesized, this behavior appears to be largely driven by cell-cell similarity inferred from expression data rather than from allelic signals. For this reason, the MoNETA input could be revised to construct an mBAF-based graph that better emphasizes allelic differences between cells and promotes stronger clustering of altered populations.
As a first step in this direction, the selective filtering of non-informative chromosome bands at this stage of the analysis is currently being explored: the underlying idea is that removing highly uninformative regions may improve the quality of the embedding and, consequently, increase the robustness of subclone detection.
Within this framework, an initial experiment consists of identifying the most variable chromosome bands, defined as those showing deviations from the neutral state and higher heterogeneity across cells. For each band, the median deviation from 0.5 and its variance are used as indicators of systematic allelic imbalance and clonal heterogeneity. Bands with both median and variance values above the 75th percentile are then selected for downstream analysis.
Figure 3.19 shows the results obtained by applying this additional filtering to the synthetic dataset. The selected chromosome bands are 6q, 7p, 7q, 19p, 19q, and 21q. This selection is consistent with the experimental setting, as chromosome 19 contains the synthetic CNLoH event, while chromosome 7 was identified by all methods as exhibiting heterogeneous SCNA patterns across

cells. By restricting the allelic information used for the embedding to these regions, cells harboring the CNLoH event on chromosome 19 become more clearly separated in the network (Figure 3.19, panel (a), upper right), leading to the identification of two subclones enriched for the altered population. These two subclones contain 88 and 61 cells, respectively, of which approximately 70% carries the synthetic CNLoH, as shown in panel (b). Preliminary evidence suggests that pre-selecting informative chromosome bands can reduce over-clustering. However, further work is required to refine this filtering strategy while carefully avoiding overfitting.

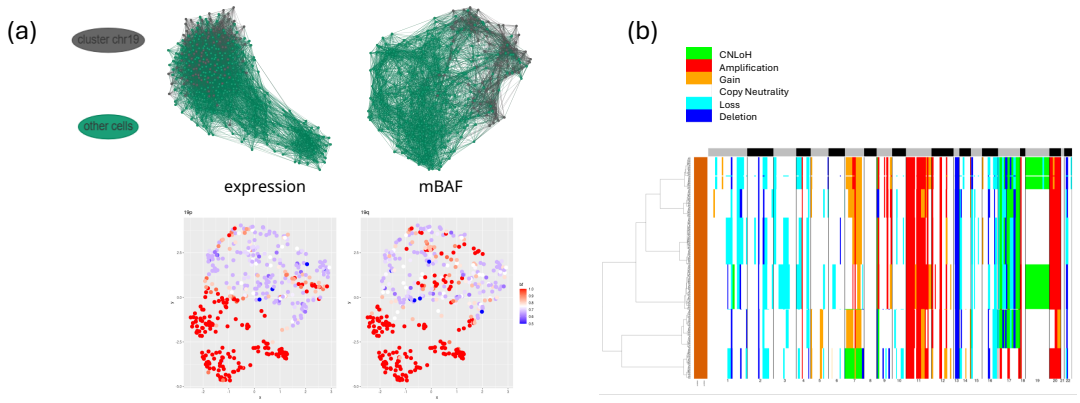

Figure 3.19: Preliminary findings of the extension of the algorithm.(a) Graph representation of input to the MoNETA algorithm. In the upper right figure, mBAF clearly shows possible clusters. The UMAP representation on chromosome 19 demonstrates that SNPs exhibiting analogous mBAF are in close proximity. (b) Final SCNA calls demonstrates the presence of two subclones with CNLOH on chromosome 19.

Alternative approaches to identify the most informative chromosome bands could include treating all cells as a single pseudobulk and detecting altered regions as those deviating from 0.5 through an mBAF segmentation step, as proposed in "CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data" [97]. This strategy could also be extended to incorporate expression data within a bivariate framework, in order to identify genomic regions showing consistent alterations across

modalities.
Third, the algorithm currently appears to perform effectively when CNLoH events span large genomic regions. This observation suggests that future refinements should place greater emphasis on detecting alterations affecting smaller genomic segments, potentially through tailored strategies aimed at increasing sensitivity and resolution in such regions. For example, final copy number calls could be refined through a bivariate segmentation approach that jointly integrates expression and allelic information. Such integration will require careful methodological design, as the two data modalities may not naturally reside in the same genomic space.
Once the core algorithmic framework is fully established, a systematic benchmarking on larger synthetic and real datasets will be necessary to assess the robustness and general applicability of these strategies. The validation of the method would benefit from a three-level evaluation. First, real datasets for which both scDNA-seq and scRNA-seq data are available [98] would provide an essential validation step at the single-cell level, enabling the assessment of subclonal structure and allele-specific events with high resolution. Second, datasets combining bulk DNA-seq, or RNA-seq, with scRNA-seq would offer a complementary validation: although bulk data lack the granularity of single-cell measurements, they typically exhibit consistent and less noisy SCNA profiles. Comparing the proposed framework with state-of-the-art bulk methods, such as ASCAT, would therefore help verify whether major SCNA patterns are consistently recovered. Finally, the generation of synthetic datasets would strengthen the credibility of the approach and provide a controlled environment for further methodological refinements. These datasets should span a wide range of tumor purities, numbers of subpopulations, and cell counts per subpopulation. Moreover, the synthetic SCNAs should vary in scale and genomic length, allowing the evaluation of the method's ability to detect focal alterations, which remain particularly challenging. Synthetic profiles should also reflect complex tumor heterogeneity, including scenarios where subclones are detectable exclusively through CNLOH events, in which only

the BAF signal is altered. Benchmarking on such datasets would enable a comprehensive assessment of breakpoint accuracy, SCNA pattern reconstruction, and subclonality inference, thereby strengthening the reliability of the approach.

Finally, it is worth discussing the extensibility of the proposed framework, highlighting two straightforward and promising directions for future developments. Future work may include the explicit modeling of whole-genome duplication (WGD) events, which could further improve subclonal copy number inference. This extension would be relatively simple to implement, as it mainly affects the copy number calling step of the algorithm.

Another interesting direction for extending the proposed pipeline is the integration of single-cell Assay for Transposase-Accessible Chromatin sequencing (scATAC-seq) data. Unlike scRNA-seq, which captures the transcriptional activity of the cell, scATAC-seq sequences genomic DNA fragments originating from regions of accessible chromatin. As such, scATAC-seq provides a direct DNA-based signal, although characterized by a highly non-uniform genome-wide coverage [99]. In this context, scATAC-seq can be viewed as a proxy for single-cell DNA sequencing, offering a trade-off between cost and scalability. Moreover, recent single-cell technologies increasingly enable the joint measurement of chromatin accessibility and gene expression within the same cell through multi-ome assays, making scATAC-seq a natural complementary layer to scRNA-seq. Integrating scATAC-seq with scRNA-seq therefore has the potential to yield a more stable and robust allele-specific framework for somatic copy number alteration (SCNA) analysis, by jointly modeling genomic and transcriptional output. To date, only one method has been explicitly designed to infer SCNAs from joint scRNA-seq and scATAC-seq data [100]. NumbatMulti-OME represents a natural extension of the original Numbat framework, in which scATAC-seq data are aggregated over genomic windows and jointly modeled with gene expression signals within a hidden Markov model to infer pseudobulk SCNA profiles.

Within the proposed framework, scATAC-seq datasets could be incorporated

alongside scRNA-seq during the MoNETA integration step, to strengthen cell-cell similarity estimates and improve the robustness of downstream SCNA inference. Future efforts will be devoted to developing coherent strategies for an effective extension of the method, leading to more robust allele-specific copy number calls.

Chapter 4

Local Sensitivity Analyisis for Chemical Reaction Networks

4.1 Chemical Reaction Networks as a Founding Model in Cancer Research

Genomic aberrations, such as SCNAs, are widely recognized as distinctive hallmarks of carcinogenesis and tumour progression. While the identification of these aberrations is often associated with the analysis of the structural organization of the genome, their impact extends beyond a purely genomic level, affecting the functional behaviour of the cell.

In particular, by influencing gene expression and protein synthesis, these alterations can significantly modify intracellular signalling pathways, which comprises sets of biochemical reactions that enable cells to perceive external signals, such as hormones, growth factors, stress and nutrients, and transmit this information to the nucleus or other compartments to induce a specific biological response [101]. Variations in protein concentrations perturb the underlying cascade of biochemical reactions, ultimately leading to a disequilibrium in the cellular environment and to altered outcomes.

In this context, mathematical modelling plays a central role. By explicitly modelling both physiological and mutational states, it is possible to capture

the dynamical consequences of genomic alterations and to investigate mechanisms that are difficult to isolate through wet-lab experiments alone. Such models provide a systematic framework to explore the propagation of genomic perturbations through signalling pathways and to analyse their impact on cellular behaviour. Moreover, these tools could serve as a guide for the design of new therapeutic strategies [102].

Among the various models, Chemical Reaction Networks (CRNs) provide a well-established and flexible mathematical framework for describing intracellular processes, including the molecular interactions underlying cancer-related signalling pathways. CRNs represent biochemical systems as sets of interacting species and reactions, enabling the simulation and analysis of complex cellular dynamics. Unlike approaches that focus exclusively on reactions directly affected by cancer-related alterations, CRNs allow for the representation of more complex mechanisms, capturing the downstream effects of cancer-induced mutations on the overall cellular behaviour [103].

Moreover, CRNs can be regarded as a foundational component for the development of digital twins, whose objective is to reproduce virtual models of cancer patients and to enable the simulation of disease progression and treatment response. In oncology, the implementation of digital twins holds the potential to support the prediction of tumour growth at the individual level and to simulate different therapeutic strategies, such as chemotherapy or immunotherapy, in order to identify the most effective interventions while minimising adverse effects.

Nevertheless, further methodological refinements are required to enhance the descriptive and predictive power of CRN-based models. Addressing these limitations represents a crucial step towards their effective application in precision oncology and provides a clear roadmap for future developments.

CRNs are typically characterised by a large number of parameters describing molecular abundances and reaction rates, enabling a quantitative and mechanistic representation of cellular dynamics. Furthermore, the inherent structural properties of CRNs render them particularly well-suited for rigorous

mathematical formalization, thereby enhancing their applicability in analytical and computational investigations.
A plethora of studies have been conducted on CRNs, with Sommariva et al. [103] presenting a formal mathematical framework capable of simulating the physiological interactions of proteins and the chemical reactions occurring within a healthy cell. In addition, the authors developed an efficient computational tool for the in silico analysis of the global effects induced on a CRN by specific classes of mutations leading to loss or gain of function of individual proteins. This framework has been further extended to incorporate the targeted action of inhibitory drugs, thereby reinforcing the role of CRNs in the development of novel targeted therapeutic strategies [104, 105, 106, 107].
It is well established that the outcomes of CRN simulations are highly sensitive to the values assigned to model parameters and initial conditions. This observation raises a central question concerning the quantitative assessment of the impact of parameter uncertainty on the behaviour of a CRN. Closely related to this issue is the challenge of designing effective numerical simulations and of correctly interpreting their results, both of which strongly depend on the choice of parameters. A further difficulty arises from the intrinsic complexity of signalling networks, which are typically characterised by a large number of interacting components and kinetic parameters. In this context, the application of dimensionality reduction techniques represents a valuable strategy for simplifying the system while preserving its essential dynamic features.
The present work addresses these challenges by proposing an approach grounded in sensitivity analysis, with the aim of systematically quantifying the influence of parameters on the model behaviour. In particular, the focus is placed on local sensitivity analysis, which examines how the equilibrium state of the system responds to small variations in its parameters. Although this choice entails certain limitations, as it neglects the full dynamics of the network, it is well justified by the timescale of the biological processes under investigation.
Indeed, it can be reasonably assumed that the mechanisms governing cell

signalling operate on an almost instantaneous timescale when compared with other cellular processes, such as the synthesis of large proteins or cell duplication. Consequently, the concentrations of the proteins involved in these signalling networks can be regarded as being at, or very close to, equilibrium [108].

Several methods have been proposed to compute the equilibrium state of CRN models under this assumption. Among them, the approach developed by Berra et al. [109] is particularly noteworthy. Their method is based on a Newton–gradient descent scheme specifically tailored to CRN models of signalling pathways, and is both computationally efficient and scalable.

This theoretical framework avoids the computational burden associated with global dynamical sensitivity approaches [110, 111], which are often difficult to adapt to large-scale signalling CRNs, and builds upon ideas introduced in earlier studies [112].

The chapter is organized as follows. First, a brief introduction to the construction of chemical reaction network models and their related mathematical framework is provided. Subsequently, the implementation of the modifications induced by two specific classes of mutations is discussed. A description of CRN involving both mutations and target drugs is finally provided.

The core methodology is then presented, which relies on a two-step procedure: first, the construction of the sensitivity matrix is undertaken, followed by the computation of the statistical sensitivity indices (SSIs). These indices quantify the relevance of individual parameters and reactions within the network.

Finally, the proposed method is applied to a CRN modelling the G1–S phase of the cell cycle of colorectal cells. This application will demonstrate that SSIs provide a reliable quantitative measure of the impact of mutations on protein expression levels. Moreover, they enable the identification of reactions that are most affected by these mutations. Importantly, SSIs also have the potential to guide the identification of novel therapeutic targets and to inform the design of in silico drug dosage simulations.

4.2 CRNs: A Mathematical Description

Following the mathematical framework introduced in Section 1.3, this section is devoted to the formalisation of CRNs and to the definition of the associated mathematical model. Proofs of Theorems 1, 2, and Propositions 1, 2, 3 and 4 can be found in "Computational quantification of global effects induced by mutations and drugs in signaling networks of colorectal cancer cells" [103].

4.2.1 From reactions to the CRN model

A CRN representing a signalling pathway typically comprises a large number of reactions involving several chemical species. According to the framework introduced in Section 1.3, the full set of reactions can be translated into a system of ordinary differential equations (ODEs), whose dimension equals the number of chemical species involved in the network. Before presenting this mathematical formulation, the fundamental mathematical objects of the model are introduced.
More precisely, let

- r denote the total number of reactions, labelled as R_j, $j = 1, \dots, r$;
- n denote the total number of chemical species, each denoted by A_i, $i = 1, \dots, n$.

Under the continuity assumption, concentrations are treated as continuous variables. The vector of species concentrations at time t is therefore defined as

$$\mathbf{x}(t) = (x_1(t), \dots, x_n(t))^T \in \mathbb{R}^n_+, \tag{4.1}$$

where T denotes transposition and $x_i(t)$ represents the concentration of species A_i at time t. In the following, time dependence will be omitted whenever this does not cause ambiguity. Let

$$\mathbf{k} = (k_1, \dots, k_r) \in \mathbb{R}^r_+ \tag{4.2}$$

be the vector of reaction rate constants, where k_j denotes the rate constant associated with reaction R_j, $j = 1, \ldots, r$. According to the law of mass action, the reaction rates depend on species concentrations through proportionality constants. The reaction flux vector

$$\mathbf{v} = (v_1, \ldots, v_r) \in \mathbb{R}_+^r \tag{4.3}$$

is therefore expressed as a function of the concentration vector and the rate constants,

$$\mathbf{v} = \mathbf{v}(\mathbf{x}, \mathbf{k}). \tag{4.4}$$

CRNs as systems of ODEs Having introduced chemical reactions at the level of individual processes in 1.3, the focus now shifts to a network-level description. In this setting, each chemical species may contribute differently to each reaction, or may not be involved at all. This observation motivates the introduction of a structured representation capable of encoding, in a compact form, the effect of all reactions on all species.

If involved, it may appear either as a reactant or as a product. For each pair A_i and R_j, the contribution of reaction R_j to the time evolution of the concentration of species A_i is therefore either proportional to the reaction flux v_j or identically zero.

This contribution is conveniently described by means of stoichiometric coefficients. In particular, the stoichiometric coefficient of species A_i in reaction R_j is negative if A_i is a reactant, positive if it is a product, and equal to zero if the species does not participate in the reaction.

Let $\mathbf{S} \in \mathbb{Z}^{n \times r}$ denote the stoichiometric matrix of the network, whose entry S_{ij} corresponds to the stoichiometric coefficient of species A_i in reaction R_j. The dynamics of the CRN can then be compactly written as the following system of ordinary differential equations:

$$\dot{\mathbf{x}} = \mathbf{S}\,\mathbf{v}(\mathbf{x}, \mathbf{k}), \tag{4.5}$$

where $\dot{\mathbf{x}}$ denotes the time derivative of the concentration vector $\mathbf{x}(t)$.
Assuming that the initial concentrations of the chemical species are known, and denoted by $\mathbf{x}_0 \in \mathbb{R}^n_+$, the dynamics of the CRN are described by the following Cauchy problem:

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{S}\,\mathbf{v}(\mathbf{x}, \mathbf{k}), \\ \mathbf{x}(0) = \mathbf{x}_0. \end{cases} \tag{4.6}$$

The solution of this system represents the temporal evolution of the concentrations of the chemical species over the time interval of interest.
On the basis of this model, together with a set of biological observations, it is possible to investigate structural and dynamical properties of the underlying CRN. This modelling framework provides a direct link between theoretical predictions and experimental data, offering insights into the mechanisms governing the system.

Conservation laws and dynamical constraints In this paragraph, a fundamental concept in the analysis of CRNs is introduced, namely conservation laws. Conservation laws play a crucial role in characterising the invariants and constraints of the network dynamics, providing insight into the behaviour of the system.
From a biochemical perspective, a conservation law states that the total amount of molecules associated with a given combination of chemical species remains constant over time. As discussed by Shinar, Alon, and Feinberg [108], this reflects the preservation of conserved moieties throughout the evolution of the network. This notion can be formalised mathematically as follows [46, 113].

Definition 1. *Given the initial concentrations and the rate constants, let* $\mathbf{x}(t)$ *be a solution of the ODE system* (4.6)*. A constant vector* $\boldsymbol{\gamma} \in \mathbb{N}^n \setminus \mathbf{0}$ *is defined as a non-negative conservation vector if there exists a constant* $c \in \mathbb{R}_+$ *such that*

$$\boldsymbol{\gamma}^T \mathbf{x}(t) = c \quad \forall t \geq 0. \tag{4.7}$$

The relation (4.7) *defines a conservation law.*

The quantity $\gamma^T \mathbf{x}(t)$ represents a linear combination of the concentrations of the involved species and corresponds to the total number of molecules associated with the conservation vector γ, which remains constant over time. This constant total amount is uniquely determined by γ and the initial condition $\mathbf{x}_0$ through

$$c = \gamma^T \mathbf{x}(t) = \gamma^T \mathbf{x}_0. \tag{4.8}$$

The following proposition provides a sufficient condition for the identification of non-negative conservation vectors.

Proposition 1. *If* $\gamma \in \ker(\mathbf{S}^T) \cap \mathbb{N}^n \setminus \{\mathbf{0}\}$, *then* γ *is a non-negative conservation vector.*

All vectors belonging to $\ker(\mathbf{S}^T)$ define conservation laws. In the remainder of this work, attention is restricted to conservation vectors satisfying

$$\gamma \in \ker(\mathbf{S}^T).$$

Under this assumption, the previous proposition ensures that any such vector defines a non-negative conservation law in the sense of Definition 4.7.
As described in [46], the set of all non-negative conservation vectors associated with a CRN forms a convex cone, and a set of linearly independent generators of this cone can be computed using the algorithm proposed by Schuster and Höfer [114].
Let $\{\gamma_1, \ldots, \gamma_p\}$ denote a set of generators of the cone of non-negative conservation vectors, and let $\mathbf{N} \in \mathbb{R}^{p \times n}$ be the matrix whose rows are given by these vectors, namely

$$\mathbf{N} = \begin{bmatrix} \gamma_1^T \\ \vdots \\ \gamma_p^T \end{bmatrix}. \tag{4.9}$$

By further assuming that

$$p = n - \operatorname{rank}(\mathbf{S}), \tag{4.10}$$

the following statements hold:

(i) The set $\{\gamma_1, \ldots, \gamma_p\}$ forms a basis of $\ker(\mathbf{S}^T)$. Indeed, the vectors $\gamma_1, \ldots, \gamma_p$ are linearly independent by construction and their number equals

$$\dim\big(\ker(\mathbf{S}^T)\big) = n - \operatorname{rank}(\mathbf{S}) = p.$$

(ii) The conservation laws associated with the CRN are completely characterised. Recalling the notation $\mathbf{x}_0 := \mathbf{x}(0)$ for the vector of initial species concentrations, it follows from (4.6) that p independent conservation laws are given by

$$\mathbf{N}\mathbf{x}(t) = \mathbf{N}\mathbf{x}_0 =: \mathbf{c} \qquad \forall t \geq 0. \tag{4.11}$$

Equivalently, relation (4.11) implies that

$$\mathbf{N}\dot{\mathbf{x}} = \mathbf{0},$$

which is consistent with the observations made in the test example at the end of the previous section.

(iii) The matrix $\mathbf{N}$ satisfies the algebraic relation

$$\mathbf{N}\mathbf{S} = \mathbf{0}. \tag{4.12}$$

This follows directly from the definition of $\mathbf{N}$ and from statement (i), since the rows of $\mathbf{N}$ belong to $\ker(\mathbf{S}^T)$.

(iv) All solutions $\mathbf{x}(t)$ of the Cauchy problem (4.6) evolve within the same affine subspace of $\mathbb{R}^n$, determined by the conservation relations (4.11).

Weakly elemented CRNs Within this thesis, a specific class of CRNs will be considered, which now will be presented.

Definition 2. *A CRN is defined as elemented if:*

(i) it admits a set of generators $\gamma_1, \ldots, \gamma_p$ *such that the matrix* $\mathbf{N}$ *contains at least one minor equal to the* $p \times p$ *identity matrix, denoted* $\mathbf{I}_p$*;*

(ii) each chemical species belongs to at least one conservation law, that is,

$$\forall i = 1, ..., n, \exists k \in \{1, ..., p\} \text{ such that } \gamma_{ki} \neq 0.$$

If only condition (i) is fulfilled, the CRN is weakly elemented. Species associated with the identity matrix are referred to as elementary or basic species.

It follows that, within weakly elemented CRNs, an elementary species corresponding to the k-th column of the identity matrix appears only in the k-th conservation law. Moreover, under the weakly elemented assumption, the columns of the matrix $\mathbf{N}$ can be reordered so that $\mathbf{N}$ admits the following block decomposition:

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_p & \mathbf{N}_2 \end{bmatrix}, \tag{4.13}$$

where $\mathbf{N}_2 \in \mathbb{R}^{p \times (n-p)}$. As a consequence, a corresponding reordering of the concentration vector is possible. In particular, by appropriately permuting the components of $\mathbf{x}$, one obtains

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{4.14}$$

where $\mathbf{x}_1 \in \mathbb{R}^p$ and $\mathbf{x}_2 \in \mathbb{R}^{n-p}$. The components of $\mathbf{x}_1$ represent the concentrations of the elementary species. Recalling the conservation relation (4.11), it follows that

$$\mathbf{c} = \mathbf{N}\mathbf{x} = \begin{bmatrix} \mathbf{I}_p & \mathbf{N}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathbf{x}_1 + \mathbf{N}_2\mathbf{x}_2,$$

which yields

$$\mathbf{x}_1 = \mathbf{c} - \mathbf{N}_2\mathbf{x}_2. \tag{4.15}$$

Within a weakly elemented CRN, multiple species may be associated with the same conservation law. In such cases, the matrix $\mathbf{N}$ may contain several columns that can be selected to form a $p \times p$ minor equal to the identity

matrix $\mathbf{I}_p$. This situation does not affect the present analysis. Indeed, once a representative elementary species is selected for each conservation law, the remaining species are treated as non-elementary with respect to the same law. Therefore, without loss of generality, it is assumed that exactly one elementary species is associated with each conservation law. Moreover, as will be shown later, the relation (4.15) enables the analysis of a reduced version of the original ODE system.
Relation (4.15) also motivates the following definition:

Definition 3. *Given a weakly elemented CRN, a vector* $\mathbf{x} \in \mathbb{R}^n$ *is said to be an ideal state of the network if and only if all elementary species have nonzero concentrations. In other words, in the present notation, an ideal state is such that* $\mathbf{x}_2 = \mathbf{0}$.

The equilibrium state of a weakly elemented CRN The analysis now focuses on the study of equilibrium points, denoted by $\mathbf{x}^*$, which represent the concentrations of the chemical species at steady state. The components of this vector are typically defined as

$$x_i^* = \lim_{t \to \infty} x_i(t), \quad i = 1, \ldots, n. \tag{4.16}$$

By introducing the affine spaces in which the solution trajectories evolve, namely the stoichiometric compatibility classes, it is possible to formulate general conditions guaranteeing the existence and uniqueness of an equilibrium point within this class.
Since the solution of the Cauchy problem (4.6) satisfies

$$\mathbf{x}(t) - \mathbf{x}_0 = \mathbf{S} \int_0^t \mathbf{v}(\mathbf{x}(\tau), \mathbf{k})\, d\tau, \quad t \in [0, T], \tag{4.17}$$

it follows that the vector $\mathbf{x}(t) - \mathbf{x}_0$ can be expressed as a linear combination of $\text{rank}(\mathbf{S})$ columns of $\mathbf{S}$, with time-dependent coefficients. Therefore, $\mathbf{x}(t) - \mathbf{x}_0$ belongs to a subspace of dimension $\text{rank}(\mathbf{S})$ [115, 116].

Definition 4. *Given a state variable* $\mathbf{x} \in \mathbb{R}^n_+$ *for the system* (4.6)*, the* stoichiometric

compatibility class *(SCC) of* $\mathbf{x}$ *is defined as the set*

$$\mathcal{SC}(\mathbf{x}) = (\mathbf{x} + \text{span}(\mathbf{S})) \cap \mathbb{R}^n_+. \tag{4.18}$$

Accordingly, if $\mathbf{x}_0$ denotes the initial condition at time $t = 0$, then all solutions of (4.6) remain in the same stoichiometric compatibility class $\mathcal{SC}(\mathbf{x}_0)$. In the case of weakly elemented CRNs, the definition of the SCC admits a simpler characterization.

Proposition 2. *Let* $\gamma_1, ..., \gamma_p$ *be a set of generators of the convex cone defined by the conservation laws, with* $\mathbf{N}$ *as in* (4.13)*. If* $p = n - \text{rank}(\mathbf{S})$*, then for any* $\mathbf{x} \in \mathbb{R}^n+$ *it holds that*

$$\mathcal{SC}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n_+ : \mathbf{N}\mathbf{y} = \mathbf{N}\mathbf{x}\}. \tag{4.19}$$

Equality (4.19) shows that, for weakly elemented CRNs, stoichiometric compatibility classes can be characterized directly in terms of the matrix $\mathbf{N}$, since they consist of all concentration vectors lying in the same affine subspace defined by the conservation relations (4.11).
Moreover, solutions of (4.6) remain within the same stoichiometric compatibility class for all $t \geq 0$. Consequently, once the initial condition $\mathbf{x}_0$ is specified, not only are the conservation laws determined, but also the stoichiometric compatibility class to which the solution trajectory belongs.
Finally, attention can be turned to the uniqueness of the equilibrium point. In general, the system of ODEs may admit multiple equilibria; therefore, an additional condition is introduced to ensure both uniqueness and global stability of the equilibrium within a given stoichiometric compatibility class.

Definition 5. *A CRN satisfies the global stability condition if, for each stoichiometric compatibility class, there exists a unique asymptotic point* $\mathbf{x}^*$*.*

Under the global stability assumption, all trajectories originating from a given stoichiometric compatibility class asymptotically converge to the same equilibrium point $\mathbf{x}^*$. Further details on the properties of equilibrium points in CRNs can be found in [116, 115, 43].

The following prepositions demonstrate the relation between the equilibrium point and the solution of a linear system defined starting from the mathematical objects characterizing the CRN. This relation will serve in the characterization of the SSIs.

Proposition 3. *If $\mathbf{x}^*$ is an equilibrium point, then it always satisfies*

$$\begin{cases} \mathbf{S}\mathbf{v}(\mathbf{x}, \mathbf{k}) = 0 \\ \mathbf{N}\mathbf{x} - \mathbf{c} = 0 \end{cases} \tag{4.20}$$

If the CRN satisfies the global stability condition, then the solution is unique.

Proposition 4. *For a weakly elemented CRN that satisfies the global stability condition, the system* (4.20) *is equivalent to the square system*

$$\begin{cases} \mathbf{S_2}\mathbf{v}(\mathbf{x}, \mathbf{k}) = 0 \\ \mathbf{N}\mathbf{x} - \mathbf{c} = 0 \end{cases} \tag{4.21}$$

where $\mathbf{S_2}$ is a $(n - p) \times r$ matrix formed by the last $n - p$ rows of $\mathbf{S}$, corresponding to the non-elementary species of the CRN.

Proposition 5. *Given a weakly elemented CRN satisfying the global stability condition, and such that the matrix $\mathbf{N}$ defines multiple elementary species for the same conservation law, then the systems* (4.21) *constructed from these choices are equivalent and share the same solution.*

Since there is no ambiguity in the choice of the elementary species, these results are fundamental, as they enable an efficient characterization of the unique equilibrium point $\mathbf{x}^*$. This, in turn, allows the application of standard numerical methods for its computation, such as the one developed by Berra et al [109], and provides the theoretical foundation for the subsequent developments presented in this thesis.

4.2.2 Modelization of mutations and drugs

When modelling genetic mutations and drug interventions, the procedure follows a systematic framework. Starting from the original CRN, which represents the physiological state of the cell, the model is modified through the addition or removal of reactions and, when necessary, molecular species. Once the altered network has been defined, all assumptions introduced in the previous section, including those related to the weakly elemented structure, must be re-evaluated. Consequently, the associated conservation laws and structural properties of the system are recomputed. The specific reactions or species to be added or removed depend on the nature of the genetic or pharmacological perturbation under consideration.

In this thesis, two main classes of genetic alterations are considered: Loss of Function (LoF) and Gain of Function (GoF) mutations. LoF mutations result in a partial or complete loss of activity of the affected protein. Two subclasses can be distinguished: null LoF mutations, in which the protein activity is entirely abolished and the corresponding concentration is assumed to vanish, and leaky LoF mutations, where residual activity remains and the concentration of the affected species is only partially reduced. In this thesis, only null LoF are taken under consideration. Conversely, GoF mutations increase the activity of a specific protein, thereby amplifying its downstream effects. Such alterations may arise from gene-level events, including copy number aberrations, leading to increased or decreased synthesis or prolonged activation of the corresponding protein.

As demonstrated by Sommariva, Caviglia, and Piana [46], such modifications enable the construction of CRNs that accurately capture the altered biochemical dynamics induced by genetic mutations and pharmacological perturbations.

From a mathematical viewpoint, each class of biological or pharmacological alteration corresponds to a specific modification of the underlying CRN.

Null LoF mutations can be modelled by introducing an operator that sets to zero the directly affected elementary species, together with all chemical species involved in the associated conservation laws.

Definition 6. *Consider a weakly elemented CRN and let A_i denote the i-th elementary species of the network. A null LoF mutation affecting A_i is modelled by the operator*

$$P_{L_i} : \mathbb{R}^n \to \mathbb{R}^n$$

which maps a concentration vector $\mathbf{x}$ *into a modified state*

$$P_{L_i}(\mathbf{x}) = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \tilde{\mathbf{x}}_2 \end{bmatrix},$$

where the components of $\tilde{\mathbf{x}}_2$ *are defined as*

$$\tilde{x}_{2,j} = \begin{cases} 0 & \text{if } \gamma_{j,i} \neq 0 \\ x_{2,j} & \text{otherwise} \end{cases} \qquad j = p+1, ..., n. \tag{4.22}$$

The corresponding elementary concentrations $\tilde{\mathbf{x}}_1$ *are determined through the conservation relations*

$$\tilde{\mathbf{x}}_1 = \tilde{\mathbf{c}} - \mathbf{N}_2 \tilde{\mathbf{x}}_2,$$

where $\tilde{\mathbf{c}}$ *is obtained from* $\mathbf{c} = \mathbf{N}\mathbf{x}$ *by setting its i-th component to zero.*

Theorem 1. *Consider a weakly elemented CRN and let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$ *be two states such that*

$$\mathcal{SC}(\mathbf{x}) = \mathcal{SC}(\mathbf{y}).$$

Then, for a null LoF mutation affecting the elementary species A_i, it holds that

$$\mathcal{SC}(P_{L_i}(\mathbf{x})) = \mathcal{SC}(P_{L_i}(\mathbf{y})).$$

Moreover, if the CRN satisfies the global stability condition, the solution trajectories originating from $P_{L_i}(\mathbf{x})$ and $P_{L_i}(\mathbf{y})$ converge to the same globally asymptotically stable state.

Therefore, it is ensured that trajectories starting from compatible initial conditions evolve on the same stoichiometric surface.

A GoF mutation of a given protein is implemented by removing from the network the reactions involved in the deactivation of such a protein, i.e., by setting the corresponding reaction rate constants to zero. Equivalently, the same modification can be implemented at the stoichiometric level by nullifying the stoichiometric coefficients associated with these reactions.

Definition 7. *Consider a CRN and let* $\mathbf{S}$ *be the corresponding stoichiometric matrix. Given a set of reactions identified by the indices* $H \subseteq \{1, ..., r\}$ *a GoF mutation results in the operator*

$$G_H : \mathbb{R}^{n \times r} \to R^{n \times r}$$

that projects $\mathbf{S}$ *into a novel stoichiometric matrix* $\tilde{\mathbf{S}} = G_H(\mathbf{S})$*, with the following property*

$$\tilde{S}_{i,h} = \begin{cases} 0 & \text{if } h \in H \\ S_{i,h} & \text{otherwise} \end{cases} \quad i = 1, ..., n. \tag{4.23}$$

Theorem 2. *Consider a CRN having stoichiometric matrix* $\mathbf{S}$*, and satisfying the global stability condition and consider* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$ *such that*

$$SC(x) = SC(y).$$

Given $H \subseteq \{1, ..., r\}$ *such that equation* (4.23) *holds, consider the CRN having stoichiometric matrix* $G_H(\mathbf{S})$*, i.e. described by the system of ODEs*

$$\dot{\mathbf{x}} = G_H(\mathbf{S})\mathbf{v}(\mathbf{x}, \mathbf{k}). \tag{4.24}$$

If also this projected CRN satisfies the global stability condition, then the trajectories obtained by solving the system of ODEs (4.24) *with initial points* $\mathbf{x}$ *and* $\mathbf{y}$ *tend asymptotically to the same steady state.*

Theorems (1) and (2) are necessary for the following theorem, which shows that the concatenation of GoF and LoF mutations leads to the formulation of a new CRN.

Theorem 3. *Consider an elemented CRN with stoichiometric matrix* $\mathbf{S}$*. Consider a set of* l *mutations, resulting in the LoF of elemental species* A_{j1}*, ...,* A_{jl} *and a set*

of q GoF mutations, resulting in the supporession of sets of reactions $H_1, ..., H_q \in \{1, ..., r\}$. *The combined effect of the considered mutations is quantified by computing the asimptotically stable state of modified systems of ODEs* $\dot{\mathbf{x}} = \tilde{\mathbf{S}}\boldsymbol{v}(\boldsymbol{x}, \mathbf{k})$, *with*

$$\tilde{\mathbf{S}} = G_{H_q} \circ ... \circ G_{H_1}(\boldsymbol{S}) \tag{4.25}$$

and

$$\boldsymbol{x}(0) = P_{L_{jl}} \circ ... \circ P_{L_{j1}}(\boldsymbol{x}^*) \tag{4.26}$$

where $\circ$ *denotes function composition and* $\boldsymbol{x}^*$ *are the values of species concentration reached by the cell in physiological condition. Then the steady state computed through the procedure does not depend on the order of the composition in the projectors and the projected system of ODEs satisfy the global stability condition*

The CRN modelling the effect of a targeted drug is constructed starting from the network that already incorporates the considered mutations. Additional chemical species are introduced to represent the drug itself and the molecular complexes formed through its interaction with the target protein. As a consequence, the stoichiometric matrix is extended by adding the rows and columns associated with the new reactions describing drug binding and its downstream effects. The reaction rate vector is likewise updated in order to capture the dynamics of these additional processes.
Each modification of the network may alter its structural properties. Therefore, the validity of the assumptions introduced in the previous sections must be reassessed whenever a new network configuration is obtained.

4.3 The Methodological Framework for Local Sensitivity Analysis of CRNs

4.3.1 Computation of Sensitivity Matrices

In order to define the sensitivity criterion, it is first necessary to introduce the so-called sensitivity matrices. Consider a weakly elemented CRN consisting

of r reactions and n chemical species, with stoichiometric matrix $\mathbf{S}$, assumed to satisfy the global stability condition. Under these assumptions, for a given initial concentration vector $\mathbf{x}_0$, all trajectories belonging to the associated stoichiometric compatibility class $\mathcal{SC}(\mathbf{x}_0)$ asymptotically converge to the same equilibrium point, denoted by $\mathbf{x}^*$, which is unique.
Moreover, since the network is weakly elemented, there exist p elemental species among the n species of the system, where $p = n - \text{rank}(\mathbf{S})$. As a consequence, once the set of species, the reaction structure, and the conservation relations defining the stoichiometric subspace are specified, the network dynamics are fully characterized by the reaction rate constants and by the values of the conservation constants. Variations in these parameters induce changes in the equilibrium point along the corresponding stoichiometric compatibility class.
It follows that the local sensitivity of the equilibrium state can be naturally expressed with respect to these quantities

$$\frac{\partial x_i^{eq}}{\partial k_j} \text{ and } \frac{\partial x_i^{eq}}{\partial c_k}, \quad \forall i, j, k. \tag{4.27}$$

Sensitivity matrices are defined as the matrices collecting the sensitivities of the equilibrium concentrations with respect to the model parameters, namely the reaction rate constants and the conservation constants.

Definition 8. *The matrices*

$$\Delta_{\mathbf{k}}\mathbf{x}^* = \begin{bmatrix} \frac{\partial x_1^*}{\partial k_1} & \cdots & \frac{\partial x_1^*}{\partial k_r} \\ \vdots & & \vdots \\ \frac{\partial x_n^*}{\partial k_1} & \cdots & \frac{\partial x_n^*}{\partial k_r} \end{bmatrix} \tag{4.28}$$

and

$$\Delta_{\mathbf{c}}\mathbf{x}^* = \begin{bmatrix} \frac{\partial x_1^*}{\partial c_1} & \cdots & \frac{\partial x_1^*}{\partial c_p} \\ \vdots & & \vdots \\ \frac{\partial x_n^*}{\partial c_1} & \cdots & \frac{\partial x_n^*}{\partial c_p} \end{bmatrix}, \tag{4.29}$$

are referred to as the sensitivity matrices of the equilibrium point $\mathbf{x}^*$ *with respect to the parameters* $\mathbf{k}$ *and* $\mathbf{c}$*, respectively.*

If an explicit analytical expression of the equilibrium concentrations as functions of the model parameters is available, the computation of the sensitivity matrices is straightforward. However, solving the system (4.21) is not always simple, especially for networks involving a large number of reactions and chemical species. In such cases, equilibrium points are typically computed numerically, yielding solutions in numerical rather than analytical form, which prevents the direct evaluation of derivatives.
The method proposed herein enables the computation of the Jacobian matrices introduced in Definition (8) without requiring an explicit analytical expression of the equilibrium point $\mathbf{x}^*$ as a function of the parameters $\mathbf{k}$ and $\mathbf{c}$.
The core idea of the method is to implicitly characterize the equilibrium point by defining a function whose zeros correspond to $\mathbf{x}^*$. This formulation allows the Jacobians with respect to the selected parameters to be computed directly, relying solely on the information provided by system (4.21). To this end, the Implicit Function Theorem is employed. [117].

Theorem 4 (Implicit Function Theorem)**.** *Let* $f : A \subseteq \mathbb{R}^n \times \mathbb{R}^r \longrightarrow \mathbb{R}^n$ *be of class* C^1 *on an open set* A*, and let* $(\mathbf{x}_0, \mathbf{y}_0)$ *satisfy* $f(\mathbf{x}_0, \mathbf{y}_0) = 0$*. Consider the Jacobian matrix of* f *at* $(\mathbf{x}_0, \mathbf{y}_0)$ *with respect to the first n variables:*

$$\nabla_{\mathbf{x}} f(\mathbf{x}_0, \mathbf{y}_0) = \begin{bmatrix} \frac{\partial f_1(x_0, y_0)}{\partial x_1} & \cdots & \frac{\partial f_1(x_0, y_0)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(x_0, y_0)}{\partial x_1} & \cdots & \frac{\partial f_n(x_0, y_0)}{\partial x_n} \end{bmatrix}$$

and assume it is invertible. Then there exist neighborhoods $U \subset \mathbb{R}^{n+r}$ *and* $V \subset \mathbb{R}^r$ *such that* $(\mathbf{x}_0, \mathbf{y}_0) \in U$*,* $\mathbf{y}_0 \in V$*, and for all* $\mathbf{y} \in V$*, there exists a unique* $\mathbf{x}$ *satisfying* $(\mathbf{x}, \mathbf{y}) \in U$ *and* $f(\mathbf{x}, \mathbf{y}) = 0$*. Furthermore, defining the function* $g : V \longrightarrow \mathbb{R}^n$ *by* $g(\mathbf{y}) = \mathbf{x}$*,* g *is of class* C^1 *and*

(i) $g(\mathbf{y}_0) = \mathbf{x}_0$*;*

(ii) $\nabla_{\mathbf{y}} g(\mathbf{x}_0, \mathbf{y}_0) = -\nabla_{\mathbf{x}} f(\mathbf{x}_0, \mathbf{y}_0)^{-1} \nabla_{\mathbf{y}} f(\mathbf{x}_0, \mathbf{y}_0)$.

The theorem states that, if there exist two functions $f^{(\mathbf{c}^*)}$ and $f^{(\mathbf{k}^*)}$ satisfying the hypotheses of the Implicit Function Theorem, then the corresponding implicit functions $g^{(\mathbf{c}^*)}$ and $g^{(\mathbf{k}^*)}$ are differentiable and describe the equilibrium point $\mathbf{x}^*$ as a function of the parameters $\mathbf{k}$ and $\mathbf{c}$, respectively. Moreover, the Jacobian matrices of these implicit functions coincide with the sensitivity matrices defined in (4.28) and (4.29).
As a consequence, explicit expressions for the computation of the sensitivity matrices are available and depend solely on quantities that are known or computable from the model structure and the equilibrium conditions.
An explicit choice for the functions $f^{(\mathbf{c}^*)}$ and $f^{(\mathbf{k}^*)}$ can be constructed as follows. Let $\mathbf{k}^*$ and $\mathbf{c}^*$ denote fixed and measured parameter vectors for the CRN under consideration, and assume that p elemental species have been selected.
Two implicit functions are now introduced, corresponding to the two classes of parameters with respect to which sensitivities are computed. First, by fixing the conservation constants $\mathbf{c}^*$, define the function

$$f^{(\mathbf{c}^*)} : \mathbb{R}^n \times \mathbb{R}^r \longrightarrow \mathbb{R}^n \tag{4.30}$$

as

$$f^{(\mathbf{c}^*)}(\mathbf{x}, \mathbf{k}) = \begin{bmatrix} \mathbf{S_2 v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N x} - \mathbf{c}^* \end{bmatrix}.$$

Similarly, by fixing the reaction rate constants $\mathbf{k}^*$, define

$$f^{(\mathbf{k}^*)} : \mathbb{R}^n \times \mathbb{R}^p \longrightarrow \mathbb{R}^n \tag{4.31}$$

as

$$f^{(\mathbf{k}^*)}(\mathbf{x}, \mathbf{c}) = \begin{bmatrix} \mathbf{S_2 v}(\mathbf{x}, \mathbf{k}^*) \\ \mathbf{N x} - \mathbf{c} \end{bmatrix}.$$

The definitions of $f^{(\mathbf{c}^*)}$ and $f^{(\mathbf{k}^*)}$ follow directly from the equilibrium system (4.21). In particular, the following properties hold:

1. $f^{(\mathbf{c}^*)} \in C^1(\mathbb{R}^{n+r})$ and $(\mathbf{x}^*, \mathbf{k}^*)$ satisfies

$$f^{(\mathbf{c}^*)}(\mathbf{x}^*, \mathbf{k}^*) = \mathbf{0};$$

2. $f^{(\mathbf{k}^*)} \in C^1(\mathbb{R}^{n+p})$ and $(\mathbf{x}^*, \mathbf{c}^*)$ satisfies

$$f^{(\mathbf{k}^*)}(\mathbf{x}^*, \mathbf{c}^*) = \mathbf{0}.$$

Therefore, equilibrium points of the system (4.21), together with the fixed parameters of the network, correspond to zeros of the implicit functions defined above.
Assuming that the Jacobian matrices $\nabla_{\mathbf{x}} f^{(\mathbf{c}^*)}(\mathbf{x}^*, \mathbf{k}^*)$ and $\nabla_{\mathbf{x}} f^{(\mathbf{k}^*)}(\mathbf{x}^*, \mathbf{c}^*)$, the Implicit Function Theorem can be applied. As a consequence, there exist open neighborhoods $V^{(\mathbf{c}^*)} \subset \mathbb{R}^r$ and $V^{(\mathbf{k}^*)} \subset \mathbb{R}^p$, and differentiable functions

$$g^{(\mathbf{c}^*)} : V^{(\mathbf{c}^*)} \longrightarrow \mathbb{R}^n, \qquad g^{(\mathbf{k}^*)} : V^{(\mathbf{k}^*)} \longrightarrow \mathbb{R}^n,$$

such that the following properties hold:

(i) $g^{(\mathbf{c}^*)}(\mathbf{k}^*) = \mathbf{x}^*$ and $g^{(\mathbf{k}^*)}(\mathbf{c}^*) = \mathbf{x}^*$;

(ii) the Jacobians of the implicit functions at the equilibrium point are given by

$$\nabla_{\mathbf{k}} g^{(\mathbf{c}^*)}(\mathbf{k}^*) = -\nabla_{\mathbf{x}} f^{(\mathbf{c}^*)}(\mathbf{x}^*, \mathbf{k}^*)^{-1} \nabla_{\mathbf{k}} f^{(\mathbf{c}^*)}(\mathbf{x}^*, \mathbf{k}^*),$$
$$\nabla_{\mathbf{c}} g^{(\mathbf{k}^*)}(\mathbf{c}^*) = -\nabla_{\mathbf{x}} f^{(\mathbf{k}^*)}(\mathbf{x}^*, \mathbf{c}^*)^{-1} \nabla_{\mathbf{c}} f^{(\mathbf{k}^*)}(\mathbf{x}^*, \mathbf{c}^*).$$

Therefore, the Jacobian matrices of $g^{(\mathbf{c}^*)}$ and $g^{(\mathbf{k}^*)}$ coincide with the sensitivity matrices defined in (4.28) and (4.29), respectively.
In particular,

$$\nabla_{\mathbf{k}} g^{(\mathbf{c}^*)}(\mathbf{k}^*) = \begin{bmatrix} \frac{\partial x_1^*}{\partial k_1} & \cdots & \frac{\partial x_1^*}{\partial k_r} \\ \vdots & & \vdots \\ \frac{\partial x_n^*}{\partial k_1} & \cdots & \frac{\partial x_n^*}{\partial k_r} \end{bmatrix}, \tag{4.32}$$

and

$$\nabla_{\mathbf{c}} g^{(\mathbf{k}^*)}(\mathbf{c}^*) = \begin{bmatrix} \frac{\partial x_1^*}{\partial c_1} & \cdots & \frac{\partial x_1^*}{\partial c_p} \\ \vdots & & \vdots \\ \frac{\partial x_n^*}{\partial c_1} & \cdots & \frac{\partial x_n^*}{\partial c_p} \end{bmatrix}. \tag{4.33}$$

Hence, the sensitivity matrices can be computed explicitly from known quantities of the CRN, without requiring an explicit analytical expression of the equilibrium point.

Verification of Well-Posedness of Sensitivity Matrices

In order to compute the sensitivity matrices, the functions $f^{(\mathbf{c}^*)}$ and $f^{(\mathbf{k}^*)}$ are defined by selecting a specific set of p elementary species. The aim of this paragraph is to show that the resulting sensitivity formulation is independent of the particular admissible choice of elementary species. In particular, it will be shown that different selections of elementary species lead to function formulations related through an invertible linear transformation. As a consequence, the corresponding sensitivity matrices are invariant with respect to the choice of the elementary species, ensuring both consistency and generality of the proposed sensitivity definition.

In the following, the demonstration over $f^{(\mathbf{k}^*)}$ is provided, since the case corresponding to $f^{(\mathbf{c}^*)}$ is analogous. To this end, consider a CRN admitting multiple elementary species associated with the same conservation laws. Without loss of generality, assume that a subset of conservation laws involves exactly two elementary species. This assumption is not restrictive: if a conservation law involves more than two elementary species, say A, B, and C, the argument can be applied pairwise (e.g. (A, B) or (A, C)) without any modification to the proof.

A couple of observations can be formulated over the most important quantities that describe the functions, i.e $\mathbf{N}$ and $\mathbf{S}_2$.

By letting i denote the number of such conservation laws, the matrix $\mathbf{N}$ contains $p + i$ columns associated with the identity matrix $\mathbf{I}_p$, due to the definition

of weakly elemented CRN. Recalling the decomposition of $\mathbf{N}$ given in (4.13),

$$\mathbf{N} = \left[\mathbf{I}_p \quad | \quad \mathbf{N}_2\right],$$

the matrix $\mathbf{N}_2$ can be thus partitioned, up to a permutation of its columns consistent with the ordering of the elementary species, as

$$\mathbf{N}_2 = \left[\tilde{\mathbf{I}}_i \quad | \quad \mathbf{N}_3\right], \tag{4.34}$$

with

$$\tilde{\mathbf{I}}_i = \begin{bmatrix} \mathbf{I}_i \\ \mathbf{0}_{(p-i)\times i} \end{bmatrix},$$

where $\mathbf{I}_i$ is the $i \times i$ identity matrix, while $\mathbf{0}_{(p-i)\times i}$ is the null matrix of dimesion $(p-i) \times i$. The matrix $\mathbf{N}_3 \in \mathbb{R}^{p\times(n-p-i)}$ denotes the submatrix associated with the remaining non-elementary species.
By suitably reordering the vector of chemical species, the concentration vector can be partitioned as

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_E \\ \boldsymbol{x}_{E'} \\ \boldsymbol{x}_N \end{bmatrix}, \tag{4.35}$$

where

$$\boldsymbol{x}_E = (x_1, \ldots, x_p)^T, \qquad \boldsymbol{x}_{E'} = (x_{p+1}, \ldots, x_{p+i})^T, \qquad \boldsymbol{x}_N = (x_{p+i+1}, \ldots, x_n)^T.$$

Here, $\boldsymbol{x}_E$ contains the p elementary species initially selected, $\boldsymbol{x}_{E'}$ represents the additional i elementary species associated with the same conservation laws, and $\boldsymbol{x}_N$ collects the remaining non-elementary species.
Finally, two different reduced stoichiometric matrices, denoted by $\mathbf{S}_2$ and $\tilde{\mathbf{S}}_2$, can be constructed from the original stoichiometric matrix $\mathbf{S}$: each corresponds to different selections of the blocks $\boldsymbol{x}_E$ and $\boldsymbol{x}_{E'}$ from the partition above.
The reaction flux vector $\mathbf{v}(\mathbf{x}, \mathbf{k})$ and the conservation laws are unchanged: the reaction rates do not depend on the selection of elementary species, and the

conservation relations originate from the same underlying network.
These observations provide the basis for the proof of the following theorem.

Theorem 5 (Well-defined sensitivity with respect to **k**). *Consider a weakly elemented chemical reaction network satisfying the global stability condition, with r reactions, n chemical species and $p < n$ conservation laws. Assume that i of these conservation laws involve more than one elemental species. Fix a conservation vector $\mathbf{c}^*$ and let f and $\tilde{f}$ be the functions associated with two different choices of elementary species, defined as*

$$f : \mathbb{R}^n \times \mathbb{R}^r \longrightarrow \mathbb{R}^n \qquad \tilde{f} : \mathbb{R}^n \times \mathbb{R}^r \longrightarrow \mathbb{R}^n$$

$$f(\mathbf{x}, \mathbf{k}) = \begin{bmatrix} \mathbf{S_2 v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N x} - \mathbf{c}^* \end{bmatrix} \qquad \tilde{f}(\mathbf{x}, \mathbf{k}) = \begin{bmatrix} \tilde{\mathbf{S}}_2 \mathbf{v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N x} - \mathbf{c}^* \end{bmatrix}$$

Then, the two functions are related by a linear change of coordinates. In particular, there exists an invertible matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$, independent of $\mathbf{x}$ and $\mathbf{k}$, such that

$$\tilde{f}(\mathbf{x}, \mathbf{k}) = \mathbf{T} f(\mathbf{x}, \mathbf{k}), \qquad \forall \mathbf{x} \in \mathbb{R}^n, \ \forall \mathbf{k} \in \mathbb{R}^r.$$

Moreover, $\mathbf{T}$ is involutory, i.e. $\mathbf{T}^{-1} = \mathbf{T}$. With the ordering of species introduced above, the matrix $\mathbf{T}$ takes the form

$$\mathbf{T} = \begin{bmatrix} -\mathbf{I}_i & -(\mathbf{N_3})_{1:i} & \mathbf{0}_{i,p} \\ \mathbf{0}_{n-p-i,i} & \mathbf{I}_{n-p-i} & \mathbf{0}_{n-i-p,p} \\ \mathbf{0}_{p,i} & \mathbf{0}_{p,n-p-i} & \mathbf{I}_p \end{bmatrix}.$$

where $\mathbf{I}_m$ denotes the identity matrix of dimension $m \times m$, and $\mathbf{0}_{s \times t}$ denotes the zero matrix of dimension $s \times t$.

Proof. Since the term $\mathbf{Nx} - \mathbf{c}$ appears in the definition of both f and $\tilde{f}$, it suffices to analyze the relation between the reduced stoichiometric matrices $\mathbf{S_2}$ and $\tilde{\mathbf{S}}_\mathbf{2}$. These two matrices are naturally defined by $\mathbf{S}$, which can be

decomposed in

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{1:i} \\ \mathbf{S}_{i+1:p} \\ \mathbf{S}_{p+1:p+i} \\ \mathbf{S_3} \end{bmatrix},$$

where $\mathbf{S}_{1:i}$ and $\mathbf{S}_{p+1:p+i}$ contain the rows associated with the doubled elementary species, $\mathbf{S}_{i+1:p}$ corresponds to the unique elemental species, and $\mathbf{S}_3$ corresponds to the non-elementary species. Thanks to this decomposition, $\mathbf{S_2}$ and $\tilde{\mathbf{S}}_\mathbf{2}$ are

$$\tilde{\mathbf{S}}_\mathbf{2} = \begin{bmatrix} \mathbf{S}_{1:i} \\ \mathbf{S_3} \end{bmatrix} \quad \text{and} \quad \mathbf{S_2} = \begin{bmatrix} \mathbf{S}_{p+1:p+i} \\ \mathbf{S_3} \end{bmatrix}$$

The relation among these matrices is given from the following observation. By recalling that conservation laws imply

$$\mathbf{NS} = \mathbf{0},$$

and the block decomposition

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_p & | & \tilde{\mathbf{I}}_i & | & \mathbf{N_3} \end{bmatrix},$$

it follows that

$$\mathbf{S}_{1:i} = -\mathbf{S}_{p+1:p+i} - (\mathbf{N_3})\mathbf{S_3}$$

As a consequence,

$$\tilde{\mathbf{S}}_\mathbf{2} = \begin{bmatrix} -\mathbf{I}_i & -(\mathbf{N_3})_{1:i} \\ \mathbf{0}_{n-p-i\times i} & \mathbf{I}_{n-p-i} \end{bmatrix} \mathbf{S_2}.$$

One can conclude that there exists a block matrix

$$\mathbf{T} = \begin{bmatrix} -\mathbf{I}_i & -(\mathbf{N_3})_{1:i} & \mathbf{0}_{i,p} \\ \mathbf{0}_{n-p-i,i} & \mathbf{I}_{n-p-i} & \mathbf{0}_{n-i-p,p} \\ \mathbf{0}_{p,i} & \mathbf{0}_{p,n-p-i} & \mathbf{I}_p \end{bmatrix},$$

such that

$$\tilde{f}(\mathbf{x}, \mathbf{k}) = \mathbf{T} f(\mathbf{x}, \mathbf{k}).$$

Moreover, the matrix $\mathbf{T}$ is upper triangular, and therefore invertible.
By the same reasoning, it follows that $\mathbf{T^{-1}} = \mathbf{T}$, hence

$$\tilde{f} = \mathbf{T} f \text{ e } f = \mathbf{T}\tilde{f}.$$

□

This theorem ensures the existence of an invertible linear trasformation among f and $\tilde{f}$.
According to Proposition (4), all representations of the same underlying CRN share the same equilibrium point $\mathbf{x}^*$. By applying the Implicit Function Theorem to f and $\tilde{f}$, provided that $\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{k}^*)$ and $\nabla_{\mathbf{x}} \tilde{f}(\mathbf{x}^*, \mathbf{k}^*)$ are invertible, it is possible to define the following functions

$$g : V \longrightarrow \mathbb{R}^n \text{ and } \tilde{g} : \tilde{V} \longrightarrow \mathbb{R}^n$$

where V and $\tilde{V}$ are open set of $\mathbf{k}^*$. Moreover,

1. $\nabla_{\mathbf{k}} g = -(\nabla_{\mathbf{x}} f)^{-1} \nabla_{\mathbf{k}} f$ and $\nabla_{\mathbf{k}} \tilde{g} = -(\nabla_{\mathbf{x}} \tilde{f})^{-1} \nabla_{\mathbf{k}} f$.
2. $\nabla_{\mathbf{x}} \tilde{f} = \mathbf{T} \nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{k}} \tilde{f} = \mathbf{T} \nabla_{\mathbf{k}} f$. Indeed, for every $l, m = 1, ..., n$

$$(\nabla_{\mathbf{x}} \tilde{f})_{lm} = \frac{\partial \tilde{f}_l}{\partial x_m} = \frac{\partial (\mathbf{T} f)_l}{\partial x_m} = \sum_{j=1}^{n} \frac{\partial T_{l,j} f_j}{\partial x_m} = \sum_{j=1}^{n} T_{l,j} \frac{\partial f_j}{\partial x_m} = (\mathbf{T} \nabla_{\mathbf{x}} f)_{lm}.$$

The two sensitivity matrices coincides:

$$\begin{aligned} \nabla_{\mathbf{k}} \tilde{g} = -(\nabla_{\mathbf{x}} \tilde{f})^{-1} \nabla_{\mathbf{k}} f = -(\mathbf{T} \nabla_{\mathbf{x}} f)^{-1} \mathbf{T} \nabla_{\mathbf{k}} f = \\ = -(\nabla_{\mathbf{x}})^{-1} \mathbf{T}^{-1} \mathbf{T} \nabla_{\mathbf{k}} f = -(\nabla_{\mathbf{x}} f)^{-1} \nabla_{\mathbf{k}} f = \nabla_{\mathbf{k}} g \end{aligned}$$

Before proceeding with the computation of the Statistical Sensitivity Indices, a unified formulation for the sensitivity matrices is provided. This will simplify their description.

A Unified Formulation for the Sensitivity Matrices

In the previous sections, the functions $f^{(c^*)}$ and $f^{(k^*)}$ were introduced in order to characterize the equilibrium of the network with respect to variations in the conservation laws constants $\mathbf{c}$ and in the kinetic parameters $\mathbf{k}$, respectively. Although treated separately for clarity, both functions describe the same equilibrium state $\boldsymbol{x}^*$ of the CRN. More precisely, the equilibrium depends simultaneously on the kinetic parameters and on the conserved quantities. This motivates considering a unified formulation in which the equilibrium is described as an implicit function of the combined parameter vector $\mathbf{h} = (\mathbf{k}, \mathbf{c})$. To this end, one may define a single function

$$\begin{aligned} F &: \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^p \longrightarrow \mathbb{R}^n \\ F(\mathbf{x}, \mathbf{k}, \mathbf{c}) &= \begin{bmatrix} \mathbf{S_2 v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N x} - \mathbf{c} \end{bmatrix} \end{aligned} \tag{4.36}$$

whose zeros characterize the equilibrium of the network.
Under the assumptions of the Implicit Function Theorem, the equilibrium $\boldsymbol{x}^*$ can then be expressed locally as a differentiable function of both $\mathbf{k}$ and $\mathbf{c}$. As a consequence, a total sensitivity matrix can be defined, collecting the sensitivities with respect to kinetic parameters and conserved quantities in a single object:

$$\Delta_{\mathbf{h}} \mathbf{x}^* = \left[\Delta_{\mathbf{k}} \mathbf{x}^* \quad | \quad \Delta_{\mathbf{c}} \mathbf{x}^* \right]. \tag{4.37}$$

This formulation provides a unified and comprehensive description of equilibrium sensitivity. For readability, the theorem regarding the well-posedness, together with a sketch of its implication, is deferred to Appendix A.7.

4.3.2 SSIs: Statistical Sensitivity Indices

In this section, the procedure for defining and computing the sensitivity indices (SSIs) is introduced. The proposed methodology is inspired by the approach of Liu, Swihart, and Neelamegham. [112], but differs from it in two key aspects.

1. By focusing on the equilibrium state of the network, the sensitivity matrices can be computed analytically, as shown in the previous section. This substantially reduces the computational cost while preserving a high level of accuracy.

2. The proposed approach leads to a more statistically robust definition of the sensitivity indices. In particular, it avoids the ambiguity associated with the arbitrary sign of eigenvectors obtained through principal component analysis.

Sensitivity indices are introduced to assess how strongly the equilibrium concentrations respond to small perturbations of the parameters defining the network. In particular, they are designed to rank the parameters according to their impact on the equilibrium state. To this end, the effect of an infinitesimal relative perturbation $\hat{\Delta}\mathbf{h}$ on the corresponding relative variation of the equilibrium concentrations $\hat{\Delta}\mathbf{x}$ is studied, with

$$\hat{\Delta}\boldsymbol{x} = \big[\frac{\Delta x_1}{x_1^*}, \dots, \frac{\Delta x_n}{x_n^*}\big]^T, \qquad \hat{\Delta}\mathbf{h} = \big[\frac{\Delta h_1}{h_1^*}, \dots, \frac{\Delta h_r}{h_{r+p}^*}\big]^T, \tag{4.38}$$

More specifically, the overall effect of a parameter perturbation on the equilibrium state is quantified by measuring the magnitude of the relative concentration variation

$$Q = ||\hat{\Delta}\mathbf{x}||_2^2 \tag{4.39}$$

. The Euclidean norm provides a scalar measure of the total deviation of the equilibrium concentrations, aggregating the contributions of all species into a single quantity. Moreover, the squared Euclidean norm yields a quadratic form, which allows for a direct spectral analysis and enables a clear interpretation of the resulting sensitivity indices in terms of dominant directions in the parameter space.
Under the assumption of small perturbations of the parameters $\mathbf{h}$, the equilibrium concentrations $\mathbf{x}(\mathbf{h})$ can be approximated using a first-order Taylor

expansion around the nominal value $\mathbf{h}^*$:

$$\Delta \boldsymbol{x} = \nabla_{\mathbf{h}} \boldsymbol{x}(\mathbf{h}^*)\ \Delta \mathbf{h} + o(|\Delta \mathbf{h}|)\ \ , \tag{4.40}$$

where

$$\Delta \boldsymbol{x} := \boldsymbol{x}(\mathbf{h}) - \boldsymbol{x}^* \tag{4.41}$$

and

$$\Delta \mathbf{h} := \mathbf{h} - \mathbf{h}^*\ \ . \tag{4.42}$$

Here, $\nabla_{\mathbf{h}} \mathbf{x}(\mathbf{h}^*)$ is the sensitivity matrix.
To express relative changes, the diagonal scaling matrices are introduced

$$\mathbf{D}_{\mathbf{x}}^{eq} = \mathrm{diag}[x_1^*, \ldots, x_n^*], \qquad \mathbf{D}_{\mathbf{h}}^{*} = \mathrm{diag}[h_1^*, \ldots, h_{r+p}^*]. \tag{4.43}$$

Using these, the normalized sensitivity matrix is defined as

$$\boldsymbol{\mathcal{S}} = (\mathbf{D}_{\mathbf{x}}^{eq})^{-1}\, \nabla_{\mathbf{h}} \mathbf{x}(\mathbf{h}^*)\, \mathbf{D}_{\mathbf{h}}^{*}. \tag{4.44}$$

The relative variations of the concentrations and parameters can be written as

$$\hat{\Delta}\mathbf{x} = (\mathbf{D}_{\mathbf{x}}^{eq})^{-1}\Delta \mathbf{x}, \qquad \hat{\Delta}\mathbf{h} = (\mathbf{D}_{\mathbf{h}}^{*})^{-1}\Delta \mathbf{h}, \tag{4.45}$$

so that the linear relation (4.40) becomes

$$\hat{\Delta}\mathbf{x} = \boldsymbol{\mathcal{S}}\, \hat{\Delta}\mathbf{h}. \tag{4.46}$$

This relation allows to compute the quadratic measure

$$Q(\hat{\Delta}\mathbf{h}) = \|\hat{\Delta}\mathbf{x}\|_2^2 = \hat{\Delta}\mathbf{h}^T\, \boldsymbol{\mathcal{S}}^T \boldsymbol{\mathcal{S}}\, \hat{\Delta}\mathbf{h} \tag{4.47}$$

which represents a measure of the relative deviation of concentrations resulting from a relative change in $\mathbf{h}$ at equilibrium.
Since Q is a positive semi-definite quadratic form, it can be expressed in terms

of the principal components of the sensitivity matrix. By performing a PCA on $\boldsymbol{\mathcal{S}}^T\boldsymbol{\mathcal{S}}$, the quadratic form can be written as

$$Q(\hat{\Delta}\mathbf{h}) = \hat{\Delta}\mathbf{h}^T\, \boldsymbol{U}\, \boldsymbol{\Lambda}\, \boldsymbol{U}^T\, \hat{\Delta}\mathbf{h}\,, \tag{4.48}$$

The matrices $\boldsymbol{\Lambda} = \text{diag}[\lambda_1, ..., \lambda_{r+p}]$, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{r+p} \geq 0$, and $\boldsymbol{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_{r+p}]$ are formed by the eigenvalues and the corresponding normalized eigenvectors of $\boldsymbol{\mathcal{S}}^T\boldsymbol{\mathcal{S}}$, respectively.
This representation highlights how the relative change in the equilibrium concentrations is decomposed along the directions of the principal components, each weighted by its corresponding eigenvalue.
Equation (4.48) allows one to quantify the relevance of the j-th parameter, corresponding to the fractional change $\hat{\Delta}h_j$. By introducing a single variation to the j-th parameter, then $\hat{\Delta}\mathbf{h}$ possesses only one non-zero component and equation (4.48) reduces to

$$Q(\hat{\Delta}h_j) = \hat{\Delta}h_j^2 \sum_{a=1}^{r+p} \lambda_a\, u_{ja}^2, \tag{4.49}$$

where $(\lambda_a, \mathbf{u}_a)$ denote the eigenvalues and eigenvectors of the matrix $\boldsymbol{\mathcal{S}}^T\boldsymbol{\mathcal{S}}$. Based on this observation, the Statistical Sensitivity iIndex (SSI) of the j-th parameter is defined as

$$e_j^{\mathbf{h}} = \frac{\displaystyle\sum_{a=1}^{r+p} \lambda_a\, u_{ja}^2}{\displaystyle\sum_{a=1}^{r+p} \lambda_a}, \tag{4.50}$$

which represents the average contribution of the j-th parameter to the total variance of the system, weighted by the eigenvalues. By construction, $e_j^{\mathbf{h}}$ satisfies

$$0 \leq e_j^{\mathbf{h}} \leq 1.$$

When applied to the original decomposition of the parameter vector $\mathbf{h} = (\mathbf{k}, \mathbf{c})$, this definition yields

- $e_j^{\mathbf{k}}$, which provides a compact measure of the overall influence of the j-th reaction rate constant on the equilibrium state of the CRN;
- $e_j^{\mathbf{c}}$, which quantifies the sensitivity of the network to variations in the j-th conservation law.

Therefore, the SSIs provide a statistically robust and interpretable measure of parameter importance, capturing the impact of each parameter on the equilibrium concentrations across all directions in the parameter space.

Sensitivity analysis for a subnetwork

One of the strengths of the SSIs lies in their applicability to the sensitivity analysis of subnetworks within a CRN.
Let $\mathcal{A} = \{A_{i_1}, \dots, A_{i_t}\}$, with $t < n$, denote the set of chemical species involved in a given subnetwork. The corresponding submatrix $\widetilde{\boldsymbol{\mathcal{S}}}$ of $\boldsymbol{\mathcal{S}}$, obtained by selecting the t rows associated with $\mathcal{A}$, represents the desired sensitivity matrix. The diagonalization of $\widetilde{\boldsymbol{\mathcal{S}}}^T\widetilde{\boldsymbol{\mathcal{S}}}$ then yields the SSI values associated with $\mathcal{A}$. These SSI values make it possible to identify the network parameters whose variations most significantly influence the equilibrium concentrations of the species in the subset $\mathcal{A}$.
In the simplest case where $t = 1$, i.e. $A = \{A_i\}$, by letting $\mathbf{s}$ be the i-th row of $\boldsymbol{\mathcal{S}}$ and given the diagonalization $\mathbf{s}^T\mathbf{s} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, it is easy to demonstrate that

$$\mathbf{s}_j^2 = (\mathbf{s}^T\mathbf{s})_{jj} = (\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T)_{jj} = \sum_{a=1}^{r+p} \lambda_a u_{ja}^2 \,. \tag{4.51}$$

Hence the statistical sensitivity indices can be computed as

$$e_j^{\mathbf{h}} = \frac{\mathbf{s}_j^2}{\sum_{a=1}^{r+p} \lambda_a} \,. \tag{4.52}$$

It is worth noting that, since the denominator in equation (4.52) is independent of j, the parameter rankings obtained from the SSIs coincide with those derived

from the squared components of **s**. This consistency reflects the fact that the components of **s** correspond to the entries of the rescaled sensitivity matrix $\mathcal{S}$, which quantifies the relative sensitivity of the equilibrium concentrations of the species in $\mathcal{A}$ with respect to the network parameters. Figure 4.1 illustrates a complete schematic representation of the algorithm.

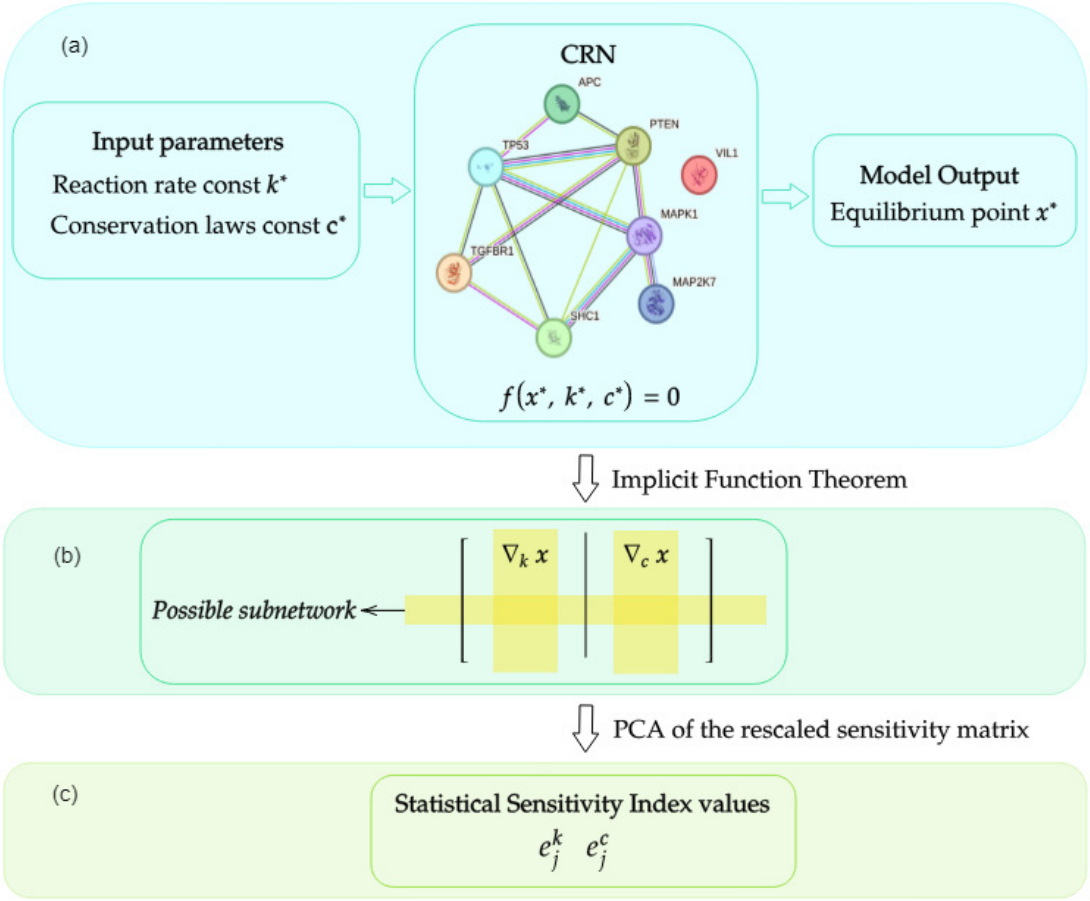

Figure 4.1: Schematic representation of the workflow for computing the SSIs. (a) Sensitivity analysis allows the quantification of how small changes in the input parameters, namely the rate constants and the conservation laws' constants, impact the protein concentrations at equilibrium. (b) By applying the implicit function theorem, the sensitivity matrix can be analytically computed. (c) SSI values are then determined by performing a PCA of the whole rescaled sensitivity matrix or of a proper submatrix (yellow bands) representing a given subnetwork.

4.4 Results on a CRN modelling a colorectal cell

4.4.1 The CRN for the colorectal cancer

Colorectal cancer (CRC) is one of the most deadly and commonly diagnosed tumors worldwide. According to data reported by the Global Cancer Observatory (https://gco.iarc.fr/), CRC ranks as the second most common cancer among women and the third among men. In order to comprehend and capture the underlying hallmarks associated with this disease, a CRN for CRC has been developed by Tortolina et al. [102], and a related in-silico model is available for computational analysis [103].

This CRC-CRN has been designed with particular reference to the checkpoint of the G1-S phase of the cell cycle: the rationale behind this choice is due to the importance of this phase in cellular life, as previously discussed in Chapter 1. The transition between these phases is pivotal in ensuring that the cell is prepared to progress through the cycle, and significant decisions are made at this point, including whether to differentiate, repair DNA or proliferate. Moreover, this transition phase has the capacity to suspend subsequent stages in response to improperly or partially replicated DNA. The failure of this critical regulatory checkpoint can result in cellular alterations and the development of pathological conditions, including cancer.

The CRN modelling the physiological state of the cell encompasses the most common signalling pathways, including MAPK, WNT, SMAD, JAK/STAT, TGFβ, and TP53 [118, 58, 119, 120]. This comprehensive network accounts for 419 chemical species, 850 reactions, and 81 conservation laws.

Such mutations include those affecting the TP53, APC, SMAD4, and KRAS genes, and thus the encoded proteins. A considerable amount of effort has been dedicated to the field of drug modelling, which has established the modelling of two target drugs, addressing the KRAS GoF mutation [102, 121, 122, 107].

In previous studies, both the physiological CRC-CRN and its variants incorporating mutations and targeted drug actions have been shown to share

several structural properties: they are open networks, they satisfy the condition $p = n - \text{rank}(\mathbf{S}^T)$, and they are weakly elemented.
Despite the extensive numerical evidence supporting global stability, a general theoretical proof valid for all stoichiometric compatibility classes is still lacking. Therefore, global stability is assumed throughout this work. For this reason, the global stability of the CRC-CRN is postulated, consistently with the numerical evidence previously reported [103, 107].

4.4.2 Sensitivity analysis for the physiological colon-rectal CRN

The first experiment aims at validating the proposed SSIs on the physiological CRN modelling the G1–S transition in a healthy colorectal cell. Specifically, the effectiveness of the SSI definition is assessed by quantifying the perturbation induced on the network equilibrium when modifying, one at a time, the components of the rate constant vector $\mathbf{k}$ and of the conservation-law vector $\mathbf{c}$ associated with the maximum, median, and minimum SSI values.
To compute the sensitivity indices, the algorithm outlined in Fig. 4.1 was applied twice. In the first case, the parameter vector $\mathbf{k}$ was considered, while keeping the conservation laws fixed; in the second case, the roles of $\mathbf{k}$ and $\mathbf{c}$ were exchanged. In both instances, the full set of chemical species was retained.
Figure 4.2 reports the resulting SSI distributions. Panel (a) shows that only 131 out of the 850 rate constants are associated with an SSI larger than 10^{-3}. A further 462 rate constants exhibit SSI values in the range $[10^{-4}, 10^{-3}]$, while all remaining parameters have SSI values below 10^{-4}.
Conversely, the conservation law constants displayed in Panel (b) illustrate a different behaviour. Among the 81 components of $\mathbf{c}$, 24 have an SSI greater than 10^{-2}, while 43 lie in the interval $[10^{-3}, 10^{-2}]$. The minimum SSI observed for the conservation laws is 5.6×10^{-4}.
The impact of variations in the rate constants and conservation laws corre-

sponding to the minimum, median, and maximum SSIs is shown in the lower panels (c) and (d) of Figure 4.2, respectively. These panels illustrate that the selected parameters indeed correspond to the lowest, median, and highest sensitivities within the network.
Sensitivity is assessed through the Euclidean norm of the relative deviation between the equilibrium of the original network and that reached after perturbing a single parameter, as defined in equation (4.39). This quantity provides a global measure of the change in equilibrium protein concentrations induced by parameter variations.
The results confirm that parameters associated with higher SSIs produce larger deviations in the equilibrium state, whereas parameters with low SSIs induce only marginal changes.

4.4.3 Sensitivity analysis for the CRC-CRN

In this sub-section, SSIs are employed to address three complementary questions of biological and pharmacological relevance in the context of CRC.
First, SSIs are employed to investigate how variations in the concentrations of different chemical species affect the equilibrium concentration of a selected target protein. To this end, conservation laws are exploited as the natural parameters through which the impact of protein-level perturbations on the equilibrium state is quantified. This analysis allows the identification of those chemical species whose modulation induces the largest changes in the equilibrium concentration of the target species, which could support the development of targeted therapies. Second, the variation of SSIs across different network configurations is analysed in order to investigate how the sensitivity structure of the system is reshaped by mutations and drug-induced modifications. This analysis allows for a comparative assessment of the robustness and reorganization of the network dynamics under different biological conditions. Finally, SSIs are exploited to evaluate the sensitivity of the network to drug-related parameters, providing insight into which drug concentrations have the greatest impact on the equilibrium state, and supporting the identification of dosage

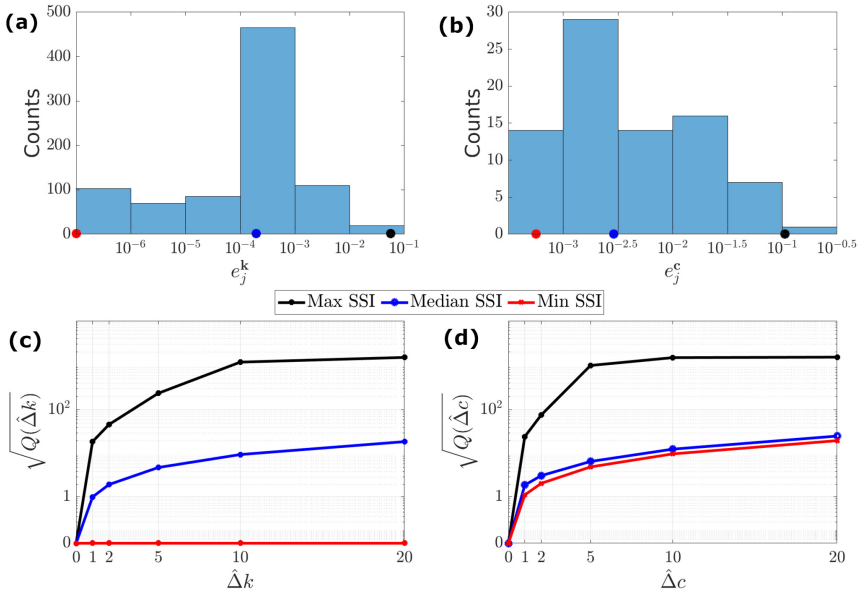

Figure 4.2: Sensitivity analysis for the CRN modelling a healthy colorectal cell. (a-b): distribution of the SSIs associated to the rate constants (left) and to the conservation laws' constants (right). In both panels, colored dots represent the values of the minimum (red), median (blue), and maximum (black) SSI. (c): $\sqrt{Q(\hat{\Delta}k)}$ as a function of the relative variation of the rate constants with minimum, median, and maximum SSI values. (d): $\sqrt{Q(\hat{\Delta}c)}$ as a function of the relative variation of the conservation laws' constants with minimum, median, and maximum SSI values.

ranges.

SSIs for identifying therapeutic targets One of the aims of this work is to illustrate the feasibility of the proposed approach in supporting the development of targeted therapies. In this regard, the analysis of the tumour suppressor p53 is presented, since mutations of the corresponding gene TP53, and in particular LoF mutations, are among the most frequent genetic events observed in colorectal cancer.
Specifically, the analysis investigates the sensitivity of the equilibrium concentration of p53 with respect to the conservation-law constants, and candidate therapeutic targets are identified as those associated with the highest SSI val-

ues. To this end, the SSI computation is performed by running the algorithm sketched in Fig. 4.1, restricting the analysis to a subnetwork consisting of a single chemical species, $\mathcal{A} = \{\text{p53}\}$.
As reported in Table 4.1, the largest SSI corresponds to the conservation law involving MDM2, which is a protein involved in tumour progression by targeting tumour suppressors, such as p53 [123, 103, 124].
From a network perspective, the conservation law associated with MDM2 involves the chemical species MDM2P, which promotes p53 degradation via the following reaction scheme:

$$\text{MDM2P} + \text{p53} \leftrightarrow \text{MDM2P\_TP53} \rightarrow \text{MDM2P} + \text{pP53U}$$

$$\text{p53U} \rightarrow \emptyset.$$

The interaction between MDM2P and pP53 leads to the formation of the complex MDM2P_p53, which can either reversibly dissociate into the original species or undergo an irreversible reaction producing MDM2P and p53U. The latter species is subsequently degraded through a consumption reaction.
An increase in the constant associated with this conservation law would raise MDM2P availability, thereby enhancing p53 degradation and reducing its equilibrium concentration. Consequently, inhibitory drugs targeting MDM2 or proteins involved in its production could restore p53 equilibrium levels.
A large body of literature on this topic provides a solid biological background for the interpretation of the results [125, 126, 127].

$e_j^{\mathbf{c}}$	$j-$th elemental species
0.455	MDM2
0.351	ARF
0.024	TP53_generator
0.022	Pho13
0.017	Pase2
0.016	AKT
0.016	Pase1

Table 4.1: SSI for the conservation laws' constants when only the concentration of TP53 is considered. For each of the SSIs in the first column, the second column shows one representative protein (elemental species) of the corresponding conservation law. Only SSI higher than 0.015 have been displayed.

SSIs for highlighting the impact of mutations A key advantage of the implemented workflow for local sensitivity analysis lies in its application to CRNs modelling GoF and LoF mutations. In this work, four networks were considered, incorporating into the CRN the most common colorectal cancer mutations: LoF of APC, SMAD4, and TP53, and GoF of KRAS. For each, two sets of SSIs, one for rate constants $\mathbf{k}$ and one for the conservation law constants $\mathbf{c}$, were computed using the previously described algorithm.
In this context, SSIs serve two purposes: they enable a rigorous evaluation of each mutation's overall effects by comparison with the physiological CRN's SSIs, and they identify the specific reactions most affected by each mutation.
Figure 4.3 reports the two computed sets of SSI values for the colorectal cancer networks under study, highlighting that the KRAS GoF mutation induces the most substantial changes. This finding is consistent with previous observations, where the KRAS GoF mutation produced larger shifts in equilibrium concentrations compared to other mutations.
The second role of SSIs is highlighted in Table 4.2, which lists reactions ranked according to the relative differences between SSI values in the physiological and KRAS associated mutated network. All these reactions are associated with Ras signalling pathways, which play a central role in regulating cell growth, proliferation, and survival across many cancer types [128, 129].

Notably, KRAS encodes the Ras protein, so the GoF mutation in KRAS directly leads to increased Ras activity, which explains the substantial changes observed in the SSIs for reactions involving Ras.
Specifically, R41 converts free Ras into its inactive form Ras_GDP, while R52 reduces the active form Ras_GTP, which is expected to accumulate due to the mutation. The same compound of R52 is also generated in reaction R49. Among reactions R294, R297, and R302, only the last directly involves Ras through Ras_GDP, whereas the first two involve $ERBP\_ShP\_G\_S$ but are crucial for signal transduction from receptor activation at the membrane to Ras activation. Similar considerations apply to reactions R412, R415, and R420 (see the Supplementary Material in "Advances in dynamic modeling of colorectal cancer signaling-network regions, a path toward targeted therapies" [102]). [103].

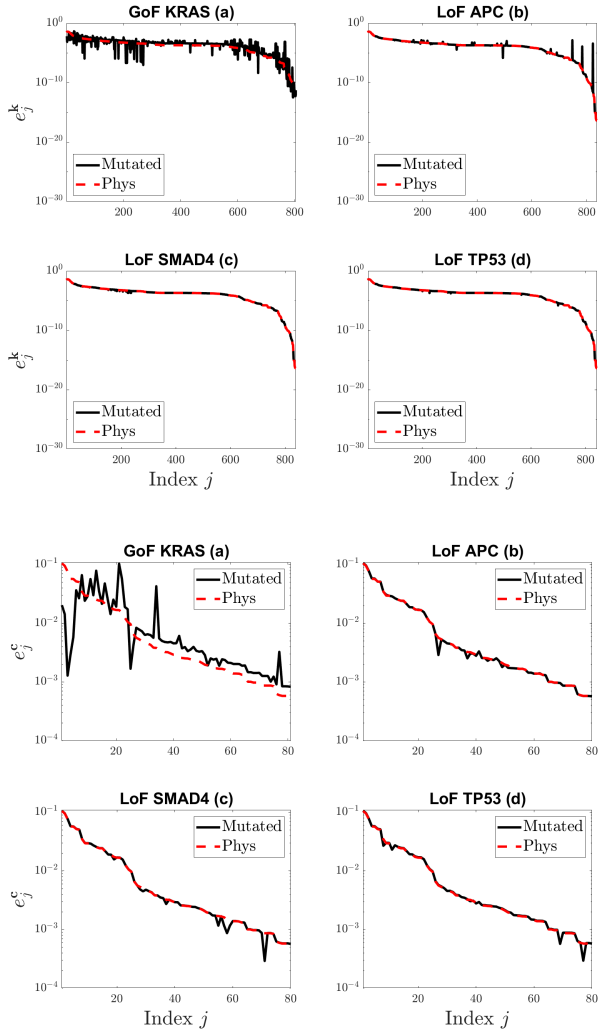

Figure 4.3: SSI values for all rate constants (left panel) conservation laws (right panel) in the case of four mutated networks (black solid line) with respect to the SSI values of the physiological CRN (red dashed line).

SSIs for in silico-based drug dosage The last important role of the SSIs regards the identification of those parameters that require a fine calibration, for example when devising novel networks or when an existing network is employed for the in-silico comparison of different therapies.

In order to provide support for the second objective, a CRC-CRN incorporating

Reference #	Reaction	$\delta e_j^{\mathbf{k}}$
R41	Ras + GDP → Ras_GDP	1.91×10^7
R294	ERBP_ShP_G_S → ERBP_ShP_G + SOS	2.91×10^4
R49	RP_ShP_G_S_Ras + GTP → RP_ShP_G_S_Ras_GTP	2.88×10^4
R412	ERB3P_Sh → ERB3P + Shc	2.86×10^4
R415	ShP + ERB3P → ERB3P_ShP	1.23×10^4
R297	ERBP_ShP + G_S → ERBP_ShP_G_S	8.64×10^3
R302	ERBP_ShP_G_S_Ras + GDP → ERBP_ShP_G_S_Ras_GDP	5.78×10^2
R57	RP_G_S_Ras_GDP → RP_G_S_Ras + GDP	5.78×10^2
R420	ERB3P_ShP_G + SOS → ERB3P_ShP_G_S	5.78×10^2
R52	RP_ShP_G_S + Ras_GTP → RP_ShP_G_S_Ras_GTP	4.95×10^2

Table 4.2: Reactions whose sensitivity indexes are mostly affected by the GoF mutation of KRAS. For each reaction listed in the second column, the first column shows the reference number attributed to the reaction in the supplementary material of Tortolina and colleagues[102], while the third column contains the relative differences $\delta e_j^{\mathbf{k}}$ between the value of the SSI $e_j^{\mathbf{k}}$ in the original and in the mutated network. Only the first 10 reactions with highest absolute value of $\delta e_j^{\mathbf{k}}$ have been displayed.

a both a GoF mutation of KRAS and the effect of two target drugs has been established. This model follows the proposal set out in a previous work [107], and comprises two groups of reactions mimicking the effect of two drugs: Dabrafenib (DBF) and Trametinib (TMT), targeting B-Raf and MEK, respectively.

To illustrate this application, the action of 40 nM of DBF and 240 nM of TMT was considered, since these concentrations were found to minimize the difference between the equilibrium of the modified and the original healthy network [107].

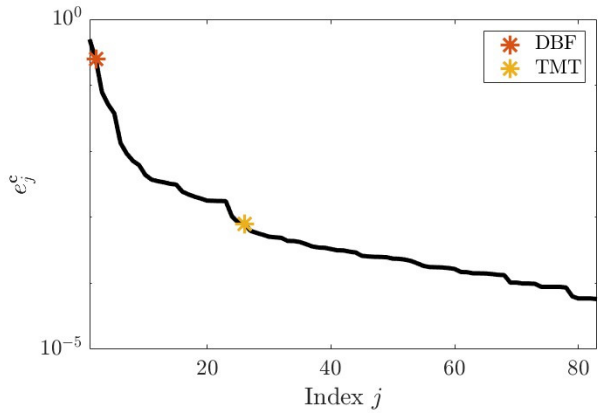

Figure 4.4: SSI values for all conservation laws' constants for the network incorporating a GoF mutation of KRAS and the combined effect of two targeted drugs, namely DBF and TMT. The red and yellow stars highlight the value of the SSI associated to the conservation law involving DBF and TMT, respectively

As illustrated in Figure 4.4, the SSI values associated with the conservation law involving DBF is equal to 0.247, a value which is considerably higher than the SSI associated with the conservation law involving TMT, which is equal to 7.98×10^{-4}.
This finding indicates that, in the context of in silico analysis employing network modelling to assess pharmaceutical dosages, it is recommended to utilise a more refined set of DBF values. This is due to the fact that even minor fluctuations in DBF concentration can exert a substantial influence on the network's equilibrium.

4.5 Discussion

CRNs that describe signalling pathways are usually characterised by a large number of kinetic parameters that must be calibrated to ensure reliable results. Sensitivity analysis can be used to determine the parameters that most influence the network behaviour, providing a basis for dimensionality reduction and simpler model descriptions. In this regard, a local sensitivity analysis provides a first step towards this goal, through the definition of indices that

rank the kinetic parameters according to the impact of their variation on the network behaviour. To this end, a set of indices, called Statistical Sensitivity Indices, have been proposed, with a particular focus on the analysis of the effects on the network equilibrium state. The utilisation of these indices offers a number of advantages. Primarily, they are adapted to the static condition, and their calculation is straightforward. Furthermore, they are not affected by the intrinsic ambiguity of the sign of the eigenvectors, as demonstrated by previous studies [112]. These indices provide the desired ranking and have been validated on a CRN describing the G1–S phase of the cell cycle in healthy colorectal cells.

The role of SSIs in the analysis of CRNs, however, extends beyond their use for parameter ranking. The application of SSIs to the CRN modelling the colorectal cell cycle, while considering different parameter configurations corresponding to physiological and mutated conditions, highlights three main uses of these indices. First, SSIs enable the identification of those chemical species whose concentration variations have the strongest impact on specific proteins, thereby providing valuable guidance for experimental investigations and the design of targeted therapies. Second, by analysing CRN models incorporating gain of function and loss of function mutations, SSIs make it possible to assess which kinetic parameters and, consequently, which reactions are most affected by a given mutation. This offers insight into how mutations influence the overall network behaviour, rather than merely altering its equilibrium state. Finally, SSIs can be exploited to identify drug target concentrations that require more accurate calibration in order to recover physiological behaviour in networks modelling pathological conditions.

In the context of modelling reduction, further developments would consist of the integration of these SSIs with more advanced techniques, such as those based on information geometry or sloppy modelling [130], which could be exploited to identify and isolate the most sensitive pathways, thereby further reducing the effective dimensionality of the system. Within this reduced framework, computational algorithms grounded in inverse problems theory [131]

could then be employed to calibrate the parameters associated with the highest SSI values by using experimental measurements of protein concentrations at equilibrium. Conversely, the remaining parameters, characterised by low sensitivity, could be more coarsely approximated using values available in the literature, without significantly compromising the predictive capability of the model.

It is important to acknowledge that obtaining precise estimates for all network parameters is practically unattainable, as any CRN model is inherently incomplete, with numerous proteins and reactions either missing or poorly characterised [132]. Nevertheless, the primary objective is the construction of a reliable quantitative framework capable of accurately simulating signalling pathways.

Conversely, it is important to acknowledge the limitations of the present work. The first one concerns the adoption of a local sensitivity analysis framework, which requires the specification of a reference set of network parameters, typically inferred from literature data. This was primarily motivated by the fact that most global sensitivity analysis The choice of a local sensitivity approach over global alternatives is justified by the limited scalability of the latter. Indeed, global methods, including Monte Carlo sampling and polynomial chaos expansions, are presently computationally feasible only for relatively small-scale problems, which precludes their direct application to large biochemical networks such as the one investigated here [133, 134]. A second limitation lies in the focus on sensitivity analysis at the steady state of the network, under the assumption that a unique equilibrium exists once all parameter values are fixed. While the definition of SSIs based on the PCA of the scaled local sensitivity matrix can, in principle, be extended to the dynamic setting, the analytical evaluation of the local sensitivity matrix at different time points becomes significantly more challenging. In contrast to the steady-state setting, where the implicit function theorem provides a direct and explicit expression for the sensitivities, the dynamic case would require the development of dedicated analytical or numerical tools.

These limitations certainly open up different perspectives within the domain of CRN research. Firstly, future efforts should be dedicated to the implementation of a global sensitivity approach that is not susceptible to scalability issues. In this context, sensitivity analysis methods based on network topology represent an appealing alternative, as they exploit structural properties of the interaction graph and do not require detailed kinetic information. However, the large size and density of biochemical CRN, involving hundreds of species and reactions, currently prevent a meaningful application of purely topology-based approaches. The development of new strategies encompassing limitations would represent a promising research direction [135].
Secondly, it is evident that the development of an algorithm capable of efficiently computing the local sensitivity matrix at each transition state of a solution to the ODE system is also necessary. This approach has the potential to facilitate the analysis of complex networks involving hundreds of proteins and reactions, such as the CRC-CRN considered in this work. Linked to this future perspective, it is important to note that a recent study was conducted with the objective of utilising Gröbner bases for the identification of potential values of the kinetic parameters for which the equilibrium of the network is not unique [136]. In the event that the identified parameter values correspond to biologically plausible scenarios, a sensitivity analysis could be performed at each equilibrium reached by the network. In this context, since the computed SSIs associated with distinct equilibrium states are expected to differ, a dedicated algorithm should be developed for their comparison and integration. For instance, a statistical comparison of the PCA of the local sensitivity matrices evaluated at different equilibria could be employed [137].

Conclusions

Cancer is a complex disease whose investigation requires the integration of multiple disciplines in order to achieve a multiscale understanding of the underlying biology. Among the research areas contributing to a comprehensive view of the problem, omics sciences play a central role by providing insight at the molecular scale. These disciplines, including genomics, transcriptomics, and proteomics, enable the study of cancer at increasingly fine resolution. By directly interrogating fundamental molecular components such as DNA and RNA, omics approaches allow the precise characterization of tumour-associated genomic alterations, thereby capturing the complexity of cancer at the genomic level. On the other hand, proteomics enables the study of fundamental cellular processes by describing interconnected cascades of biochemical reactions, such as signalling pathways, which transmit and process information within the cell. In cancer research, these signalling pathways can be explored along two complementary directions: the characterization of physiological cellular behaviour and the study of cancer-associated perturbations that lead to aberrant signalling activity.

In this thesis, two complementary and interconnected research directions have been pursued. The first concerns the characterization of a widespread class of genomic aberrations, called somatic copy number aberrations (SCNAs), through the development and application of statistical and computational approaches. From a mathematical viewpoint, SCNA inference constitutes a multi-step inverse problem, where copy number states must be inferred from noisy, indirect genomic measurements.

The second regards the modelling of signalling pathways, with a particular focus on sensitivity analysis, which provides a principled framework for pathway exploration and model simplification. Signalling networks can be formalised as high-dimensional dynamical systems, whose analysis requires principled strategies to reduce complexity while preserving essential system behaviour. The main scientific contributions of this thesis can be summarized as follows:

- the extension of a widely utilized pipeline for SCNA detection in the context of bulk-sequencing data, called ASCAT [1], to tackle the analysis of datasets characterized by non-homogeneous noise and high signal-to-noise ratios. In particular, a multi-resolution segmentation strategy has been developed, incorporating automated parameter selection, representing a trade-off among sensitivity and specificity of the results;
- the design of a novel pipeline for detecting allele-specific SCNAs from single-cell RNA sequencing data. Specifically, it extends the widely used and benchmarked pipeline SCEVAN [2] for total SCNA analysis, addressing the challenges posed by allelic data sparsity and technical noise. The proposed strategy integrates advanced methodologies from omics data analysis, including graph-based data integration and state-of-the-art clustering techniques.
- the introduction of a local sensitivity analysis framework for Chemical Reaction Networks (CRNs), which are widely used to model complex signaling pathways characterized by a large number of kinetic parameters. The proposed tool provides a systematic strategy for identifying key parameters, guiding model reduction and enabling the comparison between mutated and physiological network states.

Despite the rapid development of single-cell technologies, bulk sequencing data remain widely used in both research and clinical settings, due to their robustness, lower cost, and availability across large cohorts. Consequently, reliable and interpretable pipelines for SCNA detection in bulk data continue to play a central role in cancer genomics.

In this context, the first constribution over relies on ASCAT as a reference framework, and this choice is well motivated. Although originally introduced over a decade ago, ASCAT remains one of the most widely adopted methods for SCNA inference, including in NGS-based analyses. Moreover, its continued development by the original authors, with extensions ranging from multi-sample analyses to single-cell applications, testifies to its enduring relevance and flexibility [34, 138].

The contribution presented in this thesis aims to enhance the usability and robustness of ASCAT by reducing the dependence on manual, user-defined parameter selection. By introducing an automated strategy for parameter choice, this work mitigates the risk of arbitrary or default-driven decisions, making the pipeline more accessible to users without advanced mathematical or statistical expertise, while preserving analytical accuracy and interpretability.

At the same time, single-cell sequencing technologies are increasingly generated and adopted in cancer research. The growing availability of single-cell data opens new opportunities for investigating tumour heterogeneity and clonal evolution at an unprecedented resolution. As a consequence, modern analytical pipelines must adapt to this emerging framework, and it is within this context that the second contribution of this thesis is positioned. In particular, this work proposes an extension of the SCEVAN algorithm to an allele-specific setting. This choice is motivated by the fact that only a limited number of methods are currently available for allele-specific SCNA inference from single-cell RNA sequencing data, many of which still present methodological limitations. Moreover, SCEVAN represents a robust and well-established framework in the literature, making it a solid foundation.

Although this represents an early stage in the development of the proposed method, preliminary results already demonstrate its reliability and robustness across different experimental settings. Importantly, the methods provides a finer resolution in the characterisation of SCNAs compared to existing approaches, thereby providing a more detailed view of tumour heterogeneity and clonal structure.

These results suggest that the proposed framework constitutes a solid foundation for further methodological refinement and extension, with the potential to become a valuable tool for allele-specific SCNA analysis in single-cell transcriptomic studies.

The third contribution of this thesis addresses the modelling and analysis of cellular signalling pathways, which represent a fundamental component of cancer biology. Models such as Chemical Reaction Networks are often characterised by high structural complexity and a large number of kinetic parameters, which poses significant challenges in terms of parameter identifiability, interpretability, and computational tractability.

To tackle these issues, this thesis introduces a sensitivity analysis framework tailored to signalling network models, with the aim of identifying the parameters that most strongly influence system behaviour. The proposed approach is computationally efficient, easy to apply, and designed to be adapted to specific network structures and biological contexts. Importantly, it builds upon modelling assumptions commonly adopted in the literature, ensuring compatibility with a wide range of existing signalling models.

Despite the encouraging results presented in this thesis, several limitations should be acknowledged. Although the proposed extension of ASCAT for automatic parameter selection shows promising preliminary results, further experimental validation is required to fully assess its robustness and general applicability. In particular, benchmarking against a broader set of simulated data and real datasets would allow a more comprehensive evaluation of the method's accuracy and limitations. These additional analyses represent a natural next step towards the full consolidation of the proposed approach.

With regard to the single-cell RNA sequencing, data are affected by high levels of technical noise, dropout events, and uneven coverage, which may limit the resolution and robustness of allele-specific copy number inference in certain scenarios. While the proposed approach mitigates some of these issues, further improvements in data preprocessing remain an open challenge.

Third, the modelling of signalling pathways necessarily relies on simplifying

assumptions regarding network structure, parameter values, and kinetic interactions. Although sensitivity analysis provides an effective strategy to identify key parameters and reduce model complexity, the accuracy of the results ultimately depends on the completeness and correctness of the underlying biological models.

However, several future research directions naturally arise from the methodological developments presented in this thesis.

For the bulk sequencing framework, the proposed strategy for automated parameter selection could be adapted to other widely used SCNA inference methods and analytical pipelines. Extending this approach to additional models and datasets may improve robustness and reproducibility in bulk copy number analysis, particularly in contexts where user-defined parameters have a strong impact on downstream results.

In the single-cell RNA sequencing setting, the proposed allele-specific SCNA framework could be further enhanced through integration with complementary single-cell data modalities, such as single-cell ATAC sequencing. Such multi-omics integration would provide orthogonal genomic and transcriptomic information, potentially increasing inference accuracy and robustness while enabling a more comprehensive characterization of tumour heterogeneity and clonal architecture.

Finally, in the context of signalling pathway modelling, future work may focus on extending the proposed sensitivity analysis from a local to a global framework. Global sensitivity analyses would enable a more exhaustive exploration of the parameter space and a deeper understanding of pathway dynamics and robustness. These developments would further strengthen the role of mathematical modelling as a quantitative tool for interpreting complex cellular signalling mechanisms in cancer.

Appendices

A.1 The PCF algorithm and its extensions

In this appendix, the main methodological aspects of the PCF algorithm are presented, with the aim of providing a deeper understanding of the method. The input of standard PCF algorithm is the vector of LogR signals at each probe, denoted by $\boldsymbol{y} \in \mathbb{R}^N$, while the output are the segment start indices $\tau_0, \ldots, \tau_Q$ and segment averages $\bar{y}_1, ..., \bar{y}_Q$.
As a first step, the objective functional is simplified by exploiting some technical properties of the model. In particular, the likelihood term of $\mathcal{L}(Q, I|\mathbf{y}, \lambda)$ can be rewritten as

$$\sum_{k=1}^{Q} \sum_{j \in I_k} (y_j - \bar{y}_{I_k})^2 = \sum_{k=1}^{Q} \sum_{j \in I_k} (y_j)^2 + \sum_{k=1}^{Q} \sum_{j \in I_k} (\bar{y}_{I_k})^2 - \sum_{k=1}^{Q} \sum_{j \in I_k} 2 y_j \bar{y}_{I_k} \tag{A.53}$$

Since the sets I form a partition of $\{1, ..., N\}$, the first term reduces to

$$\sum_{k=1}^{Q} \sum_{j \in I_k} y_j^2 = \sum_{i=1}^{N} y_i^2. \tag{A.54}$$

Moreover, by simple algebraic manipulation, the remaining terms can be combined to yield

$$\sum_{i=1}^{N} y_i^2 + \sum_{k=1}^{Q} |I_k| (\bar{y}_{I_k})^2 - 2 \sum_{k=1}^{Q} \bar{y}_{I_k} \sum_{j \in I_k} y_j = - \sum_{k=1}^{Q} \frac{(\sum_{j \in I_k} y_j)^2}{|I_k|}$$

Since the minimization is performed with respect to Q and I, the constant term $\sum_{i=1}^{N} y_i^2$ can be omitted. Therefore, the optimization problem reduces to minimizing the following functional

$$\mathcal{L}'(Q, I|\boldsymbol{y}, \lambda) = -\sum_{k=1}^{Q} \frac{(\sum_{j \in I_k} y_j)^2}{|I_k|}$$

Thanks to this reformulation, the PCF algorithm can be efficiently implemented using a dynamic programming approach. For each position $k = 1, ..., N$, the algorithm evaluates all possible starting points of the last segment by minimizing a total error cost e_k, which corresponds to the value of the objective function (A.1) restricted to the first k observations. For each candidate starting position $j \leq k$, the cost of approximating the observations from j to k with a constant segment is computed. This cost, denoted by

$$d_j^k = \frac{1}{j-k-1} \left(\sum_{r=j}^{k} y_r \right)^2 \tag{A.55}$$

measures the within-segment fitting error. Let e_{j-1} denote the optimal segmentation cost up to position $j-1$. The total error at position k is then obtained by selecting the breakpoint position j that minimizes the sum of these contributions, together with a penalty term

$$e_k = \min_{j=1,\dots,k} \left(d_j^k + e_{j-1} + \lambda \right). \tag{A.56}$$

at which the minimum is attained, the optimal segmentation can be recovered by backtracking. Specifically, the segment start indices are obtained recursively as $\tau_1 = t_p, \tau_2 = t_{\tau_1} - 1, \dots, \tau_M = 1, M \geq 1$. The segment averages are then computed over the resulting segments.

This procedure is valid for PCF, and is extendible to ASPCF, as well as to the multivariate version of the algorithm. As to ASPCF, the total functional

$\mathrm{L}(Q, I|\mathbf{r}, \mathbf{b}, \lambda)$ can be decomposed as

$$\mathrm{L}(Q, I|\mathbf{r}, \mathbf{b}, \lambda) = \mathrm{L}(Q, I|\mathbf{r}, \frac{\lambda}{2}) + \mathrm{L}(Q, I|\mathbf{b}, \frac{\lambda}{2}). \tag{A.57}$$

In the bivariate setting, the PCF algorithm retains the same dynamic programming structure as in the single-sample case. The only modification concerns the computation of the segment cost: for each candidate segment, the sums and sums of squares are accumulated separately for each signal (LogR and BAF), and the overall segment cost is obtained by summing the corresponding contributions.
The same approach naturally extends to the multi-sample PCF framework, where the segment cost is computed by aggregating the contributions across all samples.

A.2 The VEGA algorithm and the multi-channel version

In this appendix, the VEGA algorithm is described.
The objective of the VEGA algorithm is the minimization of

$$\mathcal{L}(Q, I \mid \mathbf{y}, \lambda) = \sum_{k=1}^{Q} \sum_{j \in I_k} (y_j - \bar{y}_{I_k})^2 + \lambda Q, \tag{A.58}$$

which is achieved through a region-growing process. More specifically, VEGA starts from a fine partition of the signal and progressively merges adjacent regions to form larger segments, resulting in a hierarchical (pyramidal) segmentation structure.
Given two adjacent regions I_k and I_{k+1}, the change in the objective function resulting from their merging is given by

$$\mathcal{L}(Q-1, I^{(k)} \mid \boldsymbol{y}, \lambda) - \mathcal{L}(Q, I \mid \boldsymbol{y}, \lambda) = \frac{|I_k|\,|I_{k+1}|}{|I_k| + |I_{k+1}|}(\bar{y}_{I_k} - \bar{y}_{I_{k+1}})^2 - \lambda, \tag{A.59}$$

where $I^{(k)}$ denotes the partition obtained by merging regions I_k and I_{k+1}, that is, by removing the breakpoint τ_k from the set $\{\tau_0, \ldots, \tau_Q\}$.
VEGA initializes the segmentation by assigning each observation to a distinct region. For a fixed value of λ, the algorithm iteratively merges the pair of adjacent regions that yields the largest decrease of the objective function in (A.59). This procedure is repeated until no further merges are possible for the current value of λ.
Once no additional merges can be performed, the value of the penalty parameter λ is increased and the region-growing procedure is repeated. The algorithm terminates when a predefined maximum value of λ is reached.
The sequence of penalty values explored by the algorithm is referred to as the λ-schedule and is constructed as follows. When no further merges are possible for the current value λ_s, the next value is defined as

$$\lambda_{s+1} = \min_{i \in \{2,\ldots,Q\}} \hat{\lambda}_i + \epsilon, \tag{A.60}$$

where $\epsilon > 0$ is a small constant. The quantities $\hat{\lambda}_i$ are obtained by setting the energy variation in (A.59) equal to zero, yielding

$$\hat{\lambda}_i = \frac{|I_i|\,|I_{i+1}|}{|I_i| + |I_{i+1}|}(\bar{y}_{I_i} - \bar{y}_{I_{i+1}})^2. \tag{A.61}$$

Since $\hat{\lambda}_i$ quantifies the degree of compatibility between two adjacent regions, it provides an estimate of the variability between neighboring segments. Consequently, the difference between two consecutive values in the λ-schedule reflects the stability of the corresponding segmentation: small differences indicate the merging of highly compatible regions, whereas larger differences suggest more substantial structural changes.
The overall variability of the data is characterized by the standard deviation σ, computed chromosome by chromosome. Based on this quantity, a stopping

criterion for the algorithm is defined as

$$\lambda_{l+1} - \lambda_l < \beta\sigma, \tag{A.62}$$

where the parameter β is set to 0.5.
The VEGA framework can be naturally extended to the multi-channel setting in order to jointly segment multiple samples sharing a common genomic structure. Letting M denote the number of observed samples, the energy variation defined in (A.59) is then extended by incorporating information from all channels, leading to

$$\sum_{m=1}^{M} \frac{|I_k|\,|I_{k+1}|}{|I_k| + |I_{k+1}|} \left(\bar{y}_{I_k}^{(m)} - \bar{y}_{I_{k+1}}^{(m)}\right)^2 - \lambda, \tag{A.63}$$

where $\bar{y}_{I_k}^{(m)}$ denotes the mean value of segment I_k for sample m. Accordingly, the λ-schedule is adapted by applying the same VEGA procedure based on the multi-channel energy variation in (A.63). As in the PCF framework, this multi-channel formulation can be further extended to allele-specific segmentation, where multiple signals (LogR and BAF) are jointly analyzed for each sample.

A.3 Common statistical-based criteria for model selection

In this appendix, we briefly recall the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), two widely used criteria for model selection [61, 62]. Let $\mathcal{M}$ denote a statistical model with parameter vector $\boldsymbol{\theta}$ of dimension k, and let $\ell(\boldsymbol{\theta} \mid \boldsymbol{y})$ be the corresponding log-likelihood function evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$.
If the primary goal is to select a model that best predicts unobserved data, model parsimony becomes less critical and prediction error plays a central role. In this context, the Akaike Information Criterion (AIC) is commonly employed

as a model selection criterion. The AIC is defined as

$$\text{AIC} = -2\,\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) + 2k. \tag{A.64}$$

On the other hand, if the objective is to extract structural information from the observed data, model simplicity and interpretability become primary concerns. In this setting, model selection is often performed by choosing the model with the largest posterior probability given the data. Among commonly used criteria, the Bayesian Information Criterion (BIC) provides an asymptotic approximation of the logarithm of the posterior model probability under suitable regularity conditions. The BIC is defined as

$$\text{BIC} = -2\,\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) + k\log(n), \tag{A.65}$$

where n denotes the number of observations.

A.4 The modified version of BIC

In this appendix, the modified Bayesian Information Criterion (mBIC) is introduced. The mBIC provides a principled criterion for model selection in segmentation problems, allowing the selection of the optimal number of segments among a set of candidate solutions. Its formulation is grounded in a Bayesian framework and is motivated by the use of the Bayes Factor for comparing competing segmentation models.

Let $\boldsymbol{y} = (y_1, \ldots, y_T)$ be a sequence of observations, and let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two competing models for its description. The Bayes Factor BF in favor of $\mathcal{M}_1$ over $\mathcal{M}_2$ is defined as

$$\text{BF} = \frac{P(\boldsymbol{y} \mid \mathcal{M}_1)}{P(\boldsymbol{y} \mid \mathcal{M}_2)}, \tag{A.66}$$

where $P(\boldsymbol{y} \mid \mathcal{M})$ denotes the marginal likelihood of $\boldsymbol{y}$ under model $\mathcal{M}$.

Model comparison can equivalently be performed by considering the posterior

probabilities:

$$\frac{P(\mathcal{M}_1 \mid \boldsymbol{y})}{P(\mathcal{M}_2 \mid \boldsymbol{y})} = \mathrm{BF} \cdot \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}, \tag{A.67}$$

where $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ denote the prior probabilities assigned to the models.

Zhang and Siegmund derived an explicit approximation of this latter quantity in the case of multiple change-point detection problem with gaussian distributed data with unknown variance [69]. In this context, $\mathcal{M}_1$ represents a model with m change-points, whereas $\mathcal{M}_2$ corresponds to a null model with no breakpoints. This result is summarized in the following findings.

Let $\boldsymbol{y}$ be a sequence of observations that are normally distributed with constant unkown variance and piecewise constant means

$$y_i \sim \mathcal{N}(\mu_j, \sigma^2), \quad \text{for } i = \tau_j + 1, ..., \tau_{j+1}, \quad j = 0, ..., m. \tag{A.68}$$

This formulation represents a model $\mathcal{M}_m$ where the set $\{\tau_0, ..., \tau_{m+1}\}$ contains all the breakpoints locations, and m is the total number of segments. After the reparametrization

$$\mu_j = \mu_0 + \sum_{l=1}^{j} \delta_l, \tag{A.69}$$

the parameters of the model M_m are

$$\Theta = \{(\mu_0, \tau_j, \delta_j, \sigma) : j = 0, ...mm\} \tag{A.70}$$

and the following theorem can be established.

Teorema 1. *Under uniform priors on $\{\mu_0, \tau_j, \delta_j : j = j, ..., m\}$, and non informative prior $\frac{1}{\sigma}$ for σ, an approximation of logarithm transformation of the following form of the Bayes Factor*

$$\log \frac{P(\mathcal{M}_m|\boldsymbol{y})}{P(\mathcal{M}_0|\boldsymbol{y})}$$

is given by

$$\begin{aligned} mBIC(m) = &\left(\frac{T-m+1}{2}\right)\log\left[1+\frac{SS_{bg}(\hat{\mathbf{t}})}{SS_{wg}(\hat{\mathbf{t}})}\right]+\log\left[\frac{\Gamma\left(\frac{T-m+1}{2}\right)}{\Gamma\left(\frac{T+1}{2}\right)}\right] \\ &+\frac{m}{2}\log(SS_{all})-\frac{1}{2}\sum_{j=1}^{m}\log n_j(\hat{\mathbf{t}})+\left(\frac{1}{2}-m\right)\log(T) \end{aligned} \tag{A.71}$$

where Γ *is the Gamma function, while* $SS_{bg}(\hat{\mathbf{t}})$, $SS_{wg}(\hat{\mathbf{t}})$ *and* SS_{all} *are the between-group, within-group, and total sum of squares respectively. In particular, by defining*

- $n_i(\hat{\mathbf{t}}) = \hat{t}_i - \hat{t}_{i-1}$, $i = 1, ...m$
- $\bar{y}_i = \frac{1}{n_i(\hat{\mathbf{t}})}\sum_{j=\hat{t}_{i-1}}^{\hat{t}_i} y_j$, $i = 1, ..., m$
- $\bar{y} = \frac{1}{T}\sum_{j=1}^{T} y_j$

the computation of these measures becomes

- $SS_{bg}(\hat{\mathbf{t}}) = \sum_{j=1}^{m} \hat{n}_j(\bar{y}_j - \bar{y})^2$,
- $SS_{all} = \sum_{i=1}^{T}(y_i - \bar{y})^2$,
- $SS_{wg}(\hat{\mathbf{t}}) = SS_{all} - SS_{bg}$

The demonstration of this theorem is provided in [139]. In practice, the mBIC is evaluated for a set of candidate segmentation results, and the configuration maximizing the criterion is selected.

A.5 An overview of the MoNETA algorithm

MoNETA is a multi-omics integration framework designed to identify relevant relationships between biological entities by jointly analyzing heterogeneous omics data, at both bulk and single-cell resolution. In this thesis, MoNETA is employed to integrate segment-level gene expression data and chromosome-band-wise allelic information in order to construct a latent space representation

of cellular similarity, as described in Section 3.2.2.
Here, a brief overview of the MoNETA algorithm is provided. For a detailed description, its applicability and performance, please refer to "MoNETA: MultiOmics Network Embedding for SubType Analysis" [92].
The MoNETA workflow starts from a collection of L omics matrices, denoted by $M = \{M^\alpha : \alpha = 1, ..., L\}$. Each matrix $M^\alpha \in \times$ contains measurements of f_α molecular features for a subset of n_α entities, drawn from a total cohort of n entities. The objective of MoNETA is to derive a low-dimensional embedding matrix $\mathbf{E} \in \mathbb{R}^{n \times d}$, with $d << n$, which captures a latent representation of the multi-omics molecular profiles underlying the input data.
To achieve this goal, MoNETA follows a structured workflow composed of four main steps.

Data preprocessing Input data are collected and subjected to basic preprocessing procedures, including filtering of low-quality features and scaling, performed independently for each omics layer.

Construction of omics-specific networks For each omics matrix M^α, MoNETA constructs an Entity Similarity Network $ESN_\alpha = (V_\alpha, E_\alpha)$, where nodes represent entities and edges encode pairwise similarities. Two strategies are available for defining network connectivity: a static k-nearest neighbors (kNN) approach, in which each entity is connected to its k closest neighbors, and a dynamic neighborhood k^*-NN approach, which adapts the neighborhood size to local data density.

Multiplex network integration Omics-specific networks are then merged into a multiplex graph $G_M = (V_M, E_M)$. The node set contains all possible nodes, i.e. $V_M = \cup_{\alpha=1,...,L} V_\alpha$. The edge set E_M is composed of three types of connections:

(a) intra-layer edges corresponding to the omics-specific similarity networks;

(b) inter-layer edges linking nodes representing the same entity across different omics layers;

(c) cross-layer neighborhood edges connecting an entity in one omics layer to the neighbors of the same entity in another layer.

Random Walk with Restart embedding A modified Random Walk with Restart (RWR) procedure is applied to the multiplex graph in order to estimate the probability of connectivity between nodes. At each iteration, the random walker may:

(a) move to a neighboring node within the same layer;

(b) transition to the corresponding node in a different omics layer, according to an omics transition probability matrix;

(c) restart from the seed node with probability r, selecting the omics layer according to a user-defined probability vector $\tau = [\tau_1, ..., \tau_L]$.

This procedure yields a probability matrix $RW \in \mathbb{R}^{n \times n}$. This matrix contains the probability of reaching other nodes in the multi-omics network through a random walk process originating from the entity node.

Dimensionality Reduction RW is subsequently used as input for a dimensionality reduction algorithm to obtain the final embedding matrix, representing the latent multi-omics similarity structure among entities. This dimensionality reduction can be implemented trhough a pletora of specialized mwthods, including Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), t -distributed Stochastic Neighbor Embedding (t-SNE) or the multiVERSE algorithm.

A.6 The original PICASSO framework and its extension to the allele-specific context

PICASSO (Phylogenetic Inference from Copy Number Alterations in Single-cell Sequencing Observations) is a probabilistic framework aimed at inferring cellular subclones and their phylogenetic relationships from single-cell copy number alteration profiles derived from expression data.
In this thesis, PICASSO is employed as a downstream step to characterize subclonal structure introduced in Section 3.2.4.
Here, a detailed overview of the algorithm and of the proposed extension is provided. Interested readers are referred to the original literature for a more comprehensive description [93].
Before describing the algorithmic details, it is important to state the fundamental assumption underlying PICASSO. The method assumes that the observed single-cell copy number profiles are noisy measurements of latent clonal profiles, such that cells belonging to the same clone share similar SCNA patterns.
The input to PICASSO is a matrix containing copy number calls. Let $B \in \mathbb{Z}^{n \times A^{20+}}$ denote this matrix, where n represents the total number of cells and A^{20+} the total number of altered regions. The entry $B_{c,r}$ represents the copy number state of cell c in altered region r.
The algorithm starts by encoding B into a matrix

$$M \in \mathbb{Z}^{n \times \sum_{r=1}^{A^{20+}} p_r},$$

where p_r denotes the maximum absolute copy call, i.e. $p_r = \max_c |B_{c,r}|$. The encoding is determined through the following steps

1. for each cell c in region r, let $k = B_{c,r}$;
2. if $k \geq 0$, set $M_{c,1:p_r} = [1, \ldots, 1, 0, \ldots, 0]$ with k ones and $p_r - k$ zeros;
3. if $k < 0$, set $M_{c,1:p_r} = [-1, \ldots, -1, 0, \ldots, 0]$ with $|k|$ minus ones and $p_r - |k|$ zeros.

The matrix M is then obtained by concatenating the expanded representations across all regions. By defining $d = \sum_{r=1}^{A^{20+}} p_r$ and $M_c = [M_{c,1:p_1}, \ldots, M_{c,1:p_{A^{20+}}}]$, then

$$M = \begin{bmatrix} M_1 & \cdots & M_n \end{bmatrix}^T.$$

Once the encoded matrix M is constructed, PICASSO applies an Expectation-Maximization (EM) procedure to split cells into two candidate evolutionary subclones, based on the similarity structure of their SCNA profiles.

PICASSO models the data using a finite mixture model with two components, each corresponding to a hypothetical subclone. Letting $k \in \{1, 2\}$ denote the subclone index, then each subclone k is characterized by a set of parameters $\{\pi_k, \phi_k\}$, where π_k is the mixture proportion, representing the prior probability that a cell belongs to subclone k, and ϕ_k encodes the clonal copy-number profile. Specifically, $\phi_k \in \mathbb{R}^{3\times d}$ parametrizes a collection of categorical distributions, one for each of the d expanded genomic regions, assigning probabilities to the possible copy-number states $\{-1, 0, 1\}$. Let $z_{ck} \in \{0, 1\}$ be a latent indicator variable such that $z_{ck} = 1$ if cell c belongs to subclone k and $z_{ck} = 0$ otherwise, with $\sum_k z_{ck} = 1$. Under this model, the complete-data log-likelihood is given by

$$\log p(M, Z \mid \pi, \phi) = \sum_{c=1}^{n} \sum_{k=1}^{2} z_{ck} \left(\log \pi_k + \sum_{j=1}^{d} \log \phi_{k, m_{c,j}, j} \right),$$

where M is the encoded copy-number matrix and Z denotes the latent subclone assignments.

The EM algorithm iteratively alternates between the Expectation (E) step and the Maximization (M) step. The procedure is initialized by randomly assigning cells to the two subclones, resulting in initial values for the posterior membership probabilities.

E-step Given the current parameter estimates, the E-step computes the posterior probability that each cell belongs to each subclone. These probabilities,

also referred to as responsibilities, are defined as

$$\gamma_{ck} = \mathbb{E}[z_{ck} \mid M_c] = \frac{\pi_k \prod_{j=1}^{d} \phi_{k,m_{cj},j}}{\sum_{l=1}^{2} \pi_l \prod_{j=1}^{d} \phi_{l,m_{cj},j}},$$

where M_c denotes the encoded SCNA profile of cell c.

M-step. In the M-step, the model parameters are updated to maximize the expected log-likelihood. The new mixture proportions are computed as

$$\pi_k = \frac{1}{n} \sum_{c=1}^{n} \gamma_{ck},$$

while the categorical distribution parameters becomes

$$\phi_k(z) = \frac{\sum_{c=1}^{n} \sum_{j=1}^{d} \gamma_{ck}\, \delta(m_{cj}, z)}{\sum_{c=1}^{n} \sum_{j=1}^{d} \gamma_{ck}}, \quad z \in \{-1, 0, 1\},$$

where $\delta(\cdot, \cdot)$ denotes the Kronecker delta. The update of $\phi_k(z)$ does not preserve region specific information. This modeling choice reflects the goal of assessing whether the observed single cell SCNA patterns support a binary evolutionary split, while maintaining statistical stability under high noise and sparsity.
These updates ensure that the parameters reflect the expected sufficient statistics of the data under the current soft assignment of cells to subclones.

In order to determine whether a subclone should be further subdivided, PICASSO adopts two alternative stopping criteria.
The first criterion is based on the BIC, which is used to compare two resulting models:

(i) the current model with a fixed number of inferred subclones;

(ii) an alternative model in which one of the terminal subclones is further split into two descendant subclones.

The BIC term is computed according to the formulation reported in Equation (A.65) (Appendix A.3), where n denotes the total number of cells. The number of model parameters k is updated at each split to account for the additional mixing proportions and clonal copy-number profiles, resulting in an increased complexity penalty.
This splitting procedure is continued as long as the BIC score improves. The process is terminated when no improvement in the BIC is observed for two consecutive splitting steps. The authors recommend the use of the BIC-based criterion in settings with a large number of cells.
As an alternative, PICASSO implements a stopping criterion based on assignment confidence. Using the responsibilities values γ_{ck} obtained from the EM algorithm, the proportion of confidently assigned cells is evaluated. A cell is considered confidently assigned if this value exceeds a predefined threshold (e.g. 75%). If the proportion of confidently assigned cells falls below another user-defined threshold (typically $60 - 80\%$), the splitting process is stepped.
In addition, PICASSO applies two minimal constraints on the inferred subclones: each subclone must contain at least 75 cells and exhibit at least one copy number alteration occurring at high frequency. These constraints are introduced to reduce the likelihood of identifying spurious subclones arising from noise in the copy number inference process.

A.6.1 Extended Picasso framework to incorporate additional copy number states

The original version of PICASSO models single-cell copy number profiles using a three-state categorical alphabet corresponding to losses, neutral states, and gains. In this thesis, we extend the PICASSO framework to accommodate a richer set of allele-specific copy number alterations, including amplifications, deep deletions, and copy-neutral loss of heterozygosity (CNLoH).
First, the encoding matrix M is extended to take values in the discrete state space $S = \{-2, -1, 0, 1, 2, 3\}$, where each element represents a distinct copy

number state. Amplifications and deletions are encoded following the original PICASSO expansion procedure, while regions classified as CNLoH are encoded by assigning

$$M_{c,1:p_r} = [-3, ..., -3]$$

Second, the clonal emission parameters are generalized accordingly. In the extended model, each clone k is associated with a collection of categorical distributions

$$\phi_k \in \mathbb{R}^{|S| \times d},$$

where $\phi_{k,s,j}$ represents the probability of observing copy number state $s \in S$ at expanded region j for subclone k. The EM algorithm proceeds unchanged, with parameter updates naturally extending to the enlarged state space.

A.7 A Unified Formulation for Sensitivity Matrix

In this appendix, an overview of the unified formulation for sensitivity matrix is provided. The hypothesis that $(\mathbf{x}^*, \mathbf{k}^*, \mathbf{c}^*)$ satisfies $F(\mathbf{x}^*, \mathbf{k}^*, \mathbf{c}^*) = 0$ follows by the definition of the equilibrium of the considered CRN. Moreover, $F \in C^1(\mathbb{R}^{n+r+p})$, since both $f^{(c^*)}$ and $f^{(k^*)}$ were differentiable over their domains.

When the Jacobian with respect to $\mathbf{x}$ evaluated at $(\mathbf{x}^*, \mathbf{k}^*, \mathbf{c}^*)$, i.e. $\nabla_{\mathbf{x}} F(\mathbf{x}^*, \mathbf{k}^*, \mathbf{c}^*)$, is invertible, the Implicit function theorem can be applied. As a result, there exists a function

$$g : V \longrightarrow \mathbb{R}^n$$

of class $C^1(V)$ such that

$$\nabla_{\mathbf{h}} g(\mathbf{h}^*) = -(\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{h}^*))^{-1} \nabla_{\mathbf{h}} f(\mathbf{x}^*, \mathbf{h}^*).$$

where $\mathbf{h} = (\mathbf{k}, \mathbf{c})$. Following the definition of Jacobians $\nabla_{\mathbf{h}} g(\mathbf{c}^*)$ and $\nabla_{\mathbf{h}} g(\mathbf{c}^*)$ in Section 4.3.1 (equations (4.32) and (4.33)), $\nabla_{\mathbf{h}} g(\mathbf{h}^*)$ represents the total sensitivity matrix, i.e.

$$\nabla_{\mathbf{h}} g(\mathbf{h}^*) = \nabla_{\mathbf{h}} F(\mathbf{h}^*). \tag{A.72}$$

By direct differentiation of the augmented function $F(\mathbf{x}, \mathbf{h})$ with respect to the parameter vector $\mathbf{h} = (\mathbf{k}, \mathbf{c})$, it follows that

$$\nabla_{\mathbf{h}} F(\mathbf{x}, \mathbf{h}) = \begin{bmatrix} \mathbf{S_2} \nabla_{\mathbf{k}} \mathbf{v}(\mathbf{x}, \mathbf{k}) & \mathbf{0}_{(n-p)\times p} \\ \mathbf{0}_{p\times r} & \mathbf{N} \end{bmatrix} = \left[\nabla_{\mathbf{k}} f^{(\mathbf{c}^*)} \quad \middle| \quad \nabla_{\mathbf{c}} f^{(\mathbf{k}^*)} \right]. \tag{A.73}$$

Therefore, the Jacobian $\nabla_{\mathbf{h}} g(\mathbf{h}^*)$ coincides with the augmented sensitivity matrix, collecting the sensitivities of the equilibrium with respect to both reaction rate constants and conservation parameters.
The block structure of the Jacobian $\nabla_{\mathbf{h}} F(\mathbf{x}, \mathbf{h})$ follows directly from the definition of the augmented function F. Indeed, the first $n - p$ components of f correspond to the equilibrium condition $\mathbf{S_2 v}(\mathbf{x}, \mathbf{k}) = \mathbf{0}$, which does not depend on the conservation parameters $\mathbf{c}$. Consequently, the partial derivatives of these components with respect to $\mathbf{c}$ vanish. Conversely, the last p components are given by the conservation laws $\mathbf{Nx} - \mathbf{c} = \mathbf{0}$, which are independent of the reaction rate constants $\mathbf{k}$, and therefore have zero partial derivatives with respect to $\mathbf{k}$. As a consequence, the Jacobian $\nabla_{\mathbf{h}} F(\mathbf{x}, \mathbf{h})$ admits a block decomposition that directly reflects the separation between kinetic and conservation parameters.
For simplicity of notation, the obtained sensitivity matrix will be defined as

$$\nabla_{\mathbf{h}} \mathbf{x}^* \tag{A.74}$$

This observation allows the extension of the well-posedness results for the sensitivity matrices with respect to the choice of elementary species to the joint sensitivity framework, as stated in the following corollary.

Corollary 1. *Consider a chemical reaction network that is weakly elemented, satisfying the global stability condition. Suppose that this network consists of r reactions, n chemical species and $p < n$ conservation laws. Assume that among the p conservation laws, there exist i laws involving at least two elemental species. Let f and $\tilde{f}$ denote the two functions*

$$F : \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^p \longrightarrow \mathbb{R}^n$$

$$F(\mathbf{x}, \mathbf{k}, \mathbf{c}) = \begin{bmatrix} \mathbf{S_2 v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N}\mathbf{x} - \mathbf{c} \end{bmatrix}$$

and

$$\tilde{F} : \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^p \longrightarrow \mathbb{R}^n$$

$$\tilde{F}(\mathbf{x}, \mathbf{k}, \mathbf{c}) = \begin{bmatrix} \tilde{\mathbf{S}}_\mathbf{2} \mathbf{v}(\mathbf{x}, \mathbf{k}) \\ \mathbf{N}\mathbf{x} - \mathbf{c} \end{bmatrix}$$

associated with the reduced systems, as previously described.
Then, there exists an invertible matrix $\mathbf{T}$ *such that* $\tilde{f} = \mathbf{T}f$ *and* $\mathbf{T}^{-1} = \mathbf{T}$.
In particular,

$$\mathbf{T} = \begin{bmatrix} -\mathbf{I}_i & -(\mathbf{N_3})_{1:i} & \mathbf{0}_{i,p} \\ \mathbf{0}_{n-p-i,i} & \mathbf{I}_{n-p-i} & \mathbf{0}_{n-i-p,p} \\ \mathbf{0}_{p,i} & \mathbf{0}_{p,n-p-i} & \mathbf{I}_p \end{bmatrix}.$$

Bibliography

[1] Peter Van Loo et al. "Allele-specific copy number analysis of tumors". In: *Proceedings of the National Academy of Sciences* 107.39 (2010), pp. 16910–16915. DOI: 10.1073/pnas.1009843107.

[2] Antonio De Falco et al. "A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data". In: *Nature Communications* 14.1 (2023), p. 1074.

[3] Havey M James. "Overview of deoxyribonucleic acid (DNA)". In: *INOSR Scientific Research* 2 (2016), pp. 1–6.

[4] Timothy T Cornell and Thomas P Shanley. "Signal transduction overview". In: *Critical care medicine* 33.12 (2005), S410–S413.

[5] Luuk Harbers et al. "Somatic copy number alterations in human cancers: an analysis of publicly available data from the cancer genome atlas". In: *Frontiers in oncology* 11 (2021), p. 700568.

[6] Jeroen Aerssens et al. "The human genome: an introduction". In: *The Oncologist* 6.1 (2001), pp. 100–109.

[7] Anthony J Brookes. "The essence of SNPs". In: *Gene* 234.2 (1999), pp. 177–186.

[8] Tyra G Wolfsberg, Johanna McEntyre, and Gregory D Schuler. "Guide to the draft human genome". In: *Nature* 409.6822 (2001), pp. 824–826.

[9] Leonid Kruglyak and Deborah A Nickerson. "Variation is the spice of life". In: *Nature genetics* 27.3 (2001), pp. 234–236.

[10] Stephen T Sherry et al. "dbSNP: the NCBI database of genetic variation". In: *Nucleic acids research* 29.1 (2001), pp. 308–311.

[11] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, et al. "The mutational constraint spectrum quantified from variation in 141,456 humans". In: *Nature* 581 (2020), pp. 434–443.

[12] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), p. 68.

[13] Johji Inazawa, Jun Inoue, and Issei Imoto. "Comparative genomic hybridization (CGH)-arrays pave the way for identification of novel cancer-related genes". In: *Cancer science* 95.7 (2004), pp. 559–563.

[14] Veronika Gordeeva, Elena Sharova, and Georgij Arapidi. "Progress in methods for copy number variation profiling". In: *International journal of molecular sciences* 23.4 (2022), p. 2143.

[15] Daniel Pinkel et al. "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays". In: *Nature genetics* 20.2 (1998), pp. 207–211.

[16] JH Lee and JT Jeon. "Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels". In: *Cytogenetic and genome research* 123.1-4 (2008), pp. 333–342.

[17] Peter Van Loo et al. *Analyzing cancer samples with SNP arrays*. 2011.

[18] David R Bentley. "The human genome project—an overview". In: *Medicinal research reviews* 20.3 (2000), pp. 189–196.

[19] Elaine R Mardis. "Next-generation sequencing platforms". In: *Annual review of analytical chemistry* 6.1 (2013), pp. 287–303.

[20] Sara Goodwin, John D McPherson, and W Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies". In: *Nature reviews genetics* 17.6 (2016), pp. 333–351.

[21] David Deamer, Mark Akeson, and Daniel Branton. "Three decades of nanopore sequencing". In: *Nature biotechnology* 34.5 (2016), pp. 518–524.

[22] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pp. 1754–1760.

[23] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *bioinformatics* 25.16 (2009), pp. 2078–2079.

[24] Ravi K Patel and Mukesh Jain. "NGS QC Toolkit: a toolkit for quality control of next generation sequencing data". In: *PloS one* 7.2 (2012), e30619.

[25] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis". In: *Genome biology* 17.1 (2016), p. 13.

[26] Sten Linnarsson and Sarah A Teichmann. "Single-cell genomics: coming of age". In: *Genome biology* 17.1 (2016), p. 97.

[27] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". In: *Nature Reviews Genetics* 17.3 (2016), pp. 175–188.

[28] C Yau and CC Holmes. "CNV discovery using SNP genotyping arrays". In: *Cytogenetic and genome research* 123.1-4 (2008), pp. 307–312.

[29] Kai Wang et al. "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data". In: *Genome research* 17.11 (2007), pp. 1665–1674.

[30] Laura Balagué-Dobón, Alejandro Cáceres, and Juan R González. "Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure". In: *Briefings in bioinformatics* 23.2 (2022), bbac043.

[31] Biao Liu et al. "Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges". In: *Oncotarget* 4.11 (2013), p. 1868.

[32] Petr Danecek et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (2011), pp. 2156–2158.

[33] Alberto Magi et al. "Read count approach for DNA copy number variants detection". In: *Bioinformatics* 28.4 (2012), pp. 470–478.

[34] Keiran M Raine et al. "ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data". In: *Current protocols in bioinformatics* 56.1 (2016), pp. 15–9.

[35] Eric Talevich et al. "CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing". In: *PLoS computational biology* 12.4 (2016), e1004873.

[36] Ronglai Shen and Venkatraman E Seshan. "FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing". In: *Nucleic acids research* 44.16 (2016), e131–e131.

[37] Simone Zaccaria and Benjamin J Raphael. "Characterizing allele-and haplotype-specific copy numbers in single cells with CHISEL". In: *Nature biotechnology* 39.2 (2021), pp. 207–214.

[38] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[39] Rory Stark, Marta Grzelak, and James Hadfield. "RNA sequencing: the teenage years". In: *Nature reviews genetics* 20.11 (2019), pp. 631–656.

[40] Joel Rozowsky et al. "AlleleSeq: analysis of allele-specific expression and binding in a network framework". In: *Molecular systems biology* 7.1 (2011), p. 522.

[41] Teng Gao et al. "Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes". In: *Nature Biotechnology* 41.3 (2023), pp. 417–426.

[42] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & molecular medicine* 50.8 (2018), pp. 1–14.

[43] Vijaysekhar Chellaboina et al. "Modeling and analysis of mass-action kinetics". In: *IEEE Control Systems Magazine* 29.4 (2009), pp. 60–78.

[44] Brian P Ingalls. *Mathematical modeling in systems biology: an introduction*. MIT press, 2013.

[45] Scussolini Mara. "Inverse Problems in data-driven multi-scale Systems Medicine: application to cancer physiology". PhD thesis. Università degli studi si Genova, 2018.

[46] Sara Sommariva, Giacomo Caviglia, and Michele Piana. "Gain and Loss of Function mutations in biological chemical reaction networks: a mathematical model with application to colorectal cancer cells". In: *Journal of Mathematical Biology* 82.6 (2021), pp. 1–25.

[47] Yibo Zhang, Wenyu Liu, and Junbo Duan. "On the core segmentation algorithms of copy number variation detection tools". In: *Briefings in Bioinformatics* 25.2 (2024), bbae022.

[48] Jane Fridlyand et al. "Hidden Markov models approach to the analysis of array CGH data". In: *Journal of multivariate analysis* 90.1 (2004), pp. 132–153.

[49] Adam B Olshen et al. "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5.4 (2004), pp. 557–572.

[50] Ashish Sen and Muni S Srivastava. "On tests for detecting change in mean". In: *The Annals of statistics* (1975), pp. 98–108.

[51] David Siegmund. "Boundary crossing probabilities and statistical applications". In: *The Annals of Statistics* (1986), pp. 361–404.

[52] Gerhard Winkler and Volkmar Liebscher. "Smoothers for discontinuous signals". In: *Journal of Nonparametric Statistics* 14.1-2 (2002), pp. 203–222.

[53] Gro Nilsen et al. "Copynumber: efficient algorithms for single-and multi-track copy number segmentation". In: *BMC genomics* 13.1 (2012), p. 591.

[54] Sandro Morganella et al. "VEGA: variational segmentation for copy number detection". In: *Bioinformatics* 26.24 (2010), pp. 3020–3027.

[55] Edith M Ross et al. "Allele-specific multi-sample copy number segmentation in ASCAT". In: *Bioinformatics* 37.13 (2021), pp. 1909–1911.

[56] Robert Tibshirani and Pei Wang. "Spatial smoothing and hot spot detection for CGH data using the fused lasso". In: *Biostatistics* 9.1 (2008), pp. 18–29.

[57] Fatima Zare, Abdelrahman Hosny, and Sheida Nabavi. "Noise cancellation using total variation for copy number variation detection". In: *BMC bioinformatics* 19.Suppl 11 (2018), p. 361.

[58] Ying E Zhang. "Mechanistic insight into contextual TGF-β signaling". In: *Current opinion in cell biology* 51 (2018), pp. 1–7.

[59] Ralph CA Rippe, Jacqueline J Meulman, and Paul HC Eilers. "Visualization of genomic changes by segmented smoothing using an L 0 penalty". In: *PloS one* 7.6 (2012), e38230.

[60] Simone Zaccaria and Benjamin J Raphael. "Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data". In: *Nature communications* 11.1 (2020), p. 4301.

[61] Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6 (2003), pp. 716–723.

[62] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.

[63] Hassan Foroughi-Asl et al. *BALSAMIC: Bioinformatic Analysis pipeLine for SomAtic MutatIons in Cancer*. Version v18.0.1. URL: `https://github.com/Clinical-Genomics/BALSAMIC`.

[64] Solrun Kolbeinsdottir et al. "Absolute copy number aware CNV calling of sub-megabase segments in ultra-low coverage single-cell DNA sequencing data". In: *Nucleic Acids Research* 53.17 (2025), gkaf919.

[65] Justin K Ichida and Evangelos Kiskinis. "Probing disorders of the nervous system using reprogramming approaches". In: *The EMBO journal* 34.11 (2015), pp. 1456–1477.

[66] Martina Servetti et al. "Optimization of Transcription Factor-Driven Neuronal Differentiation from Human Induced Pluripotent Stem Cells for Disease Modelling and Drug Screening". In: *Stem Cell Reviews and Reports* 21.3 (2025), pp. 816–833.

[67] Van Loo Lab. *ASCAT: Allele-Specific Copy number Analysis of Tumors*. `https://github.com/VanLoo-lab/ascat`. GitHub repository. Accessed: 20 February 2026. 2026.

[68] Franck Picard et al. "A statistical approach for array CGH data analysis". In: *BMC bioinformatics* 6.1 (2005), p. 27.

[69] Nancy R Zhang and David O Siegmund. "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data". In: *Biometrics* 63.1 (2007), pp. 22–32.

[70] Franck Picard et al. "Joint segmentation, calling, and normalization of multiple CGH profiles". In: *Biostatistics* 12.3 (2011), pp. 413–428.

[71] Allison P Heath et al. "The NCI genomic data commons". In: *Nature genetics* 53.3 (2021), pp. 257–262.

[72] David Brown et al. "Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations". In: *Nature communications* 8.1 (2017), p. 14944.

[73] Liangcai Zhang and Li Zhang. "Use of autocorrelation scanning in DNA copy number analysis". In: *Bioinformatics* 29.21 (2013), pp. 2678–2682.

[74] Soroush Samadian, Jeff P Bruce, and Trevor J Pugh. "Bamgineer: introduction of simulated allele-specific copy number variants into exome and targeted sequence data sets". In: *PLoS Computational Biology* 14.3 (2018), e1006080.

[75] Yue Xing et al. "SECNVs: a simulator of copy number variants and whole-exome sequences from reference genomes". In: *Frontiers in Genetics* 11 (2020), p. 82.

[76] Riccardo Scandino, Federico Calabrese, and Alessandro Romanel. "Synggen: fast and data-driven generation of synthetic heterogeneous NGS cancer data". In: *Bioinformatics* 39.1 (2023), btac792.

[77] Alberto Magi et al. "EXCAVATOR: detecting copy number variants from whole-exome sequencing data". In: *Genome biology* 14.10 (2013), R120.

[78] Serena Nik-Zainal et al. "The life history of 21 breast cancers". In: *Cell* 149.5 (2012), pp. 994–1007.

[79] Minfang Song et al. "Benchmarking copy number aberrations inference tools using single-cell multi-omics datasets". In: *Briefings in Bioinformatics* 26.2 (2025), bbaf076.

[80] Mihriban Karaayvaz et al. "Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq". In: *Nature communications* 9.1 (2018), p. 3588.

[81] Liwen Zhang et al. "Interrogating subclonal heterogeneity of liver cancer with single-cell multi-omics analysis". In: *Scientific Reports* 15.1 (2025), p. 39021.

[82] Rongting Huang et al. "Robust analysis of allele-specific copy number alterations from scRNA-seq data with XClone". In: *Nature Communications* 15.1 (2024), p. 6684.

[83] Ana Ortega-Batista et al. "Single-Cell Sequencing: Genomic and Transcriptomic Approaches in Cancer Cell Biology". In: *International Journal of Molecular Sciences* 26.5 (2025), p. 2074.

[84] Ruli Gao et al. "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes". In: *Nature biotechnology* 39.5 (2021), pp. 599–608.

[85] Anoop P Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–1401.

[86] Wuming Gong et al. "DrImpute: imputing dropout events in single cell RNA sequencing data". In: *BMC bioinformatics* 19.1 (2018), p. 220.

[87] Katharina T Schmid et al. "Benchmarking scRNA-seq copy number variation callers". In: *Nature Communications* 16.1 (2025), p. 8777.

[88] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[89] Sandro Morganella and Michele Ceccarelli. "VegaMC: a R/bioconductor package for fast downstream analysis of large array comparative genomic hybridization datasets". In: *Bioinformatics* 28.19 (2012), pp. 2512–2514.

[90] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[91] Chen Xu and Zhengchang Su. "Identification of cell types from single-cell transcriptomes using a novel clustering method". In: *Bioinformatics* 31.12 (2015), pp. 1974–1980.

[92] Giovanni Scala et al. "MoNETA: MultiOmics Network Embedding for SubType Analysis". In: *NAR Genomics and Bioinformatics* 6.4 (2024), lqae141.

[93] Alejandro Jiménez-Sánchez et al. "Transcriptomic plasticity is a hallmark of metastatic pancreatic cancer". In: *bioRxiv* (2025), pp. 2025–02.

[94] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic acids research* 30.1 (2002), pp. 207–210.

[95] Jeanette Baran-Gale, Tamir Chandra, and Kristina Kirschner. "Experimental design for single-cell RNA sequencing". In: *Briefings in functional genomics* 17.4 (2018), pp. 233–239.

[96] Navid Ahmadinejad, Yunro Chung, and Li Liu. "J-score: A robust measure of clustering accuracy". In: *PeerJ Computer Science* 9 (2023), e1545.

[97] Akdes Serin Harmanci, Arif O Harmanci, and Xiaobo Zhou. "CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data". In: *Nature communications* 11.1 (2020), p. 89.

[98] Ignacio Vázquez-García et al. "Ovarian cancer mutational processes drive site-specific immune evasion". In: *Nature* 612.7941 (2022), pp. 778–786.

[99] Akshaya Ramakrishnan et al. "epiAneufinder identifies copy number alterations from single-cell ATAC-seq data". In: *Nature Communications* 14.1 (2023), p. 5846.

[100] Ruitong Li et al. "Numbat-multiome: inferring copy number variations by combining RNA and chromatin accessibility information from single-cell data". In: *Briefings in Bioinformatics* 26.5 (2025), bbaf516.

[101] Bruce Alberts Alexander Johnson Julian Lewis, David Morgan Martin Raff Keith Roberts, and Peter Walter. *ALBERTS 6, Molecular Biology of the Cell*.

[102] Lorenzo Tortolina et al. "Advances in dynamic modeling of colorectal cancer signaling-network regions, a path toward targeted therapies". In: *Oncotarget* 6.7 (2015), p. 5041.

[103] Sara Sommariva et al. "Computational quantification of global effects induced by mutations and drugs in signaling networks of colorectal cancer cells". In: *Scientific reports* 11.1 (2021), pp. 1–13.

[104] Tapesh Santra. "Fitting mathematical models of biochemical pathways to steady state perturbation response data without simulating perturbation experiments". In: *Scientific Reports* 8.1 (2018), pp. 1–14.

[105] Bertram Klinger et al. "Network quantification of EGFR signaling unveils potential for targeted combination therapy". In: *Molecular systems biology* 9.1 (2013), p. 673.

[106] Özgür Sahin et al. "Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance". In: *BMC systems biology* 3 (2009), pp. 1–20.

[107] Sara Sommariva et al. "In-silico modelling of the mitogen-activated protein kinase (MAPK) pathway in colorectal cancer: mutations and targeted therapy". In: *Frontiers in Systems Biology* (2023).

[108] Guy Shinar, Uri Alon, and Martin Feinberg. "Sensitivity and robustness in chemical reaction networks". In: *SIAM Journal on Applied Mathematics* 69.4 (2009), pp. 977–998.

[109] Silvia Berra et al. "Combined Newton-Gradient Method for Constrained Root-Finding in Chemical Reaction Networks". In: *Journal of Optimization Theory and Applications* 200.1 (2024), pp. 404–427.

[110] Zhike Zi. "Sensitivity analysis approaches applied to systems biology models". In: *IET systems biology* 5.6 (2011), pp. 336–346.

[111] Andrea Saltelli et al. "Sensitivity analysis for chemical models". In: *Chemical reviews* 105.7 (2005), pp. 2811–2828.

[112] Gang Liu, Mark T Swihart, and Sriram Neelamegham. "Sensitivity, principal component and flux analysis applied to signal transduction: the case of epidermal growth factor mediated signaling". In: *Bioinformatics* 21.7 (2005), pp. 1194–1202.

[113] Jens Reich and Evgeni Selkov. *Energy Metabolism of the Cell.Academic Press.London*. Jan. 1981.

[114] Stefan Schuster and Thomas Höfer. "Determining all extreme semi-positive conservation relations in chemical reaction systems: a test criterion for conservativity". In: *Journal of the Chemical Society, Faraday Transactions* 87.16 (1991), pp. 2561–2566.

[115] Polly Y Yu and Gheorghe Craciun. "Mathematical analysis of chemical reaction systems". In: *Israel Journal of Chemistry* 58.6-7 (2018), pp. 733–741.

[116] Martin Feinberg. "Chemical reaction network structure and the stability of complex isothermal reactors—I. The deficiency zero and deficiency one theorems". In: *Chemical engineering science* 42.10 (1987), pp. 2229–2268.

[117] Steven George Krantz and Harold R Parks. *The implicit function theorem: history theory and applications*. Springer Science & Business Media, 2002.

[118] Markus Morkel et al. "Similar but different: distinct roles for KRAS and BRAF oncogenes in colorectal cancer development and therapy resistance". In: *Oncotarget* 6.25 (2015), p. 20785.

[119] Sung-Young Shin et al. "Positive-and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway". In: *Journal of cell science* 122.3 (2009), pp. 425–435.

[120] Sung-Young Shin and Lan K Nguyen. "Dissecting cell-fate determination through integrated mathematical modeling of the ERK/MAPK signaling pathway". In: *ERK Signaling: Methods and Protocols* (2016), pp. 409–432.

[121] Kanwal Tariq and Kulsoom Ghias. "Colorectal cancer carcinogenesis: a review of mechanisms". In: *Cancer biology & medicine* 13.1 (2016), p. 120.

[122] Matthew W Anderson et al. "Mathematical modeling highlights the complex role of AKT in TRAIL-induced apoptosis of colorectal carcinoma cells". In: *IScience* 12 (2019), pp. 182–193.

[123] DA Freedman, L Wu, and AJ Levine. "Functions of the MDM2 oncoprotein". In: *Cellular and Molecular Life Sciences CMLS* 55 (1999), pp. 96–107.

[124] Jiaqi Li and Manabu Kurokawa. "Regulation of MDM2 stability after DNA damage". In: *Journal of cellular physiology* 230.10 (2015), pp. 2318–2327.

[125] Wendy De Roock et al. "KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer". In: *The lancet oncology* 12.6 (2011), pp. 594–603.

[126] Noa Rivlin et al. "Mutations in the p53 tumor suppressor gene: important milestones at the various steps of tumorigenesis". In: *Genes & cancer* 2.4 (2011), pp. 466–474.

[127] Diamantis I Tsilimigras et al. "Clinical significance and prognostic relevance of KRAS, BRAF, PI3K and TP53 genetic mutation analysis for resectable and unresectable colorectal liver metastases: A systematic review of the current evidence". In: *Surgical oncology* 27.2 (2018), pp. 280–288.

[128] Richard J Orton et al. "Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway". In: *Biochemical Journal* 392.2 (2005), pp. 249–261.

[129] Arnold J Levine, Nancy A Jenkins, and Neal G Copeland. "The roles of initiating truncal mutations in human cancers: the order of mutations and tumor cell type matters". In: *Cancer cell* 35.1 (2019), pp. 10–15.

[130] Mark K Transtrum et al. "Perspective: Sloppiness and emergent theories in physics, biology, and beyond". In: *The Journal of chemical physics* 143.1 (2015).

[131] M Bertero and M Piana. "Inverse problems in biomedical imaging: modeling and methods of solution". In: *Complex systems in biomedicine*. Springer, 2006, pp. 1–33.

[132] Zachary G Nicolaou and Adilson E Motter. "Missing links as a source of seemingly variable constants in complex reaction networks". In: *Physical Review Research* 2.4 (2020), p. 043135.

[133] Alejandro Fernandez Villaverde et al. "Assessment of prediction uncertainty quantification methods in systems Biology". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics: 15455963* (2023).

[134] Hans-Michael Kaltenbach, Sotiris Dimopoulos, and Joerg Stelling. "Systems analysis of cellular networks under uncertainty". In: *FEBS letters* 583.24 (2009), pp. 3923–3930.

[135] Takashi Okada and Atsushi Mochizuki. "Sensitivity and network topology in chemical reaction systems". In: *Physical Review E* 96.2 (2017), p. 022322.

[136] Paola Ferrari et al. "When algebra twinks system biology: a conjecture on the structure of Gröbner bases in complex chemical reaction networks". In: *Bollettino dell'Unione Matematica Italiana* (2025), pp. 1–16.

[137] James R Schott. "Common principal component subspaces in two groups". In: *Biometrika* 75.2 (1988), pp. 229–236.

[138] Van Loo Lab. *ASCAT.sc*. `https://github.com/VanLoo-lab/ASCAT.sc`. 2026.

[139] Nancy R Zhang. *Change-point detection and sequence alignment: statistical problems of genomics*. Stanford University, 2005.