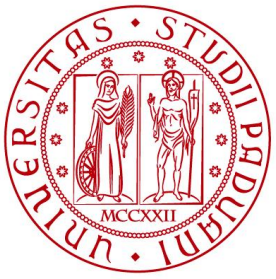
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea Magistrale in Molecular Biology



TESI DI LAUREA

A benchmark of tools for inferring Copy Number Alterations from single-cell RNA sequencing data

Relatore: Prof.ssa Chiara Romualdi
Dipartimento di Biologia

Correlatore: Prof.ssa Enrica Calura
Dipartimento di Biologia

Laureanda: Elena Cibola

ANNO ACCADEMICO 2023/2024

Contents

Abstract

High-grade serous ovarian cancer (HGSOC) is characterized by widespread genomic instability, with copy number alterations (CNAs) playing a key role in tumor progression and therapy resistance. Single-cell RNA sequencing (scRNA-seq) enables the study of genetic heterogeneity inside tumors at the single-cell level. In this thesis, CNAs inferred from scRNA-seq data with InferCNV, SCEVAN, and Numbat are compared to those derived from bulk whole-genome sequencing data, which serves as the ground truth, in order to evaluate their ability in CNA detection. Based on samples from patients with HGSOC, the analysis demonstrates that SCEVAN is the most accurate method. CNA profiles generated from scRNA-seq-based tools are further used to quantify CNA signature activities and predict platinum-based treatment response. However, the results indicate that these predictions are not consistent when compared with those derived from WGS data. In conclusion, this benchmark provides guidance for selecting the optimal tool for CNA inference when working with scRNA-seq data, being particularly relevant to future studies of chromosomal instability and tumor heterogeneity in HGSOC.

1. Introduction

1.1 High-Grade Serous Ovarian Cancer

Ovarian cancer (OC) is a critical global health concern. In 2022, it was the seventh most common cancer and the fifth leading cause of cancer-related deaths in women, with 275,335 new cases and 159,285 deaths reported globally (Cancer Today (IARC), 2022) (see Figure 1). High-grade serous ovarian cancer (HGSOC), the most common subtype of OC (see Figure 2), is responsible for the 70-80% of the disease's mortality (Bowtell et al., 2015; Kim et al., 2018), leading to an estimated 111,500 to 127,500 deaths in 2022.

Early detection of HGSOC can significantly improve treatment outcomes and survival rates. When diagnosed early, with tumors localized to the ovary, the 5-year survival rate is 92.3%; when the disease has spread to the pelvis, it is 74.5%. In advanced stages, when metastasis has occurred, the 5-year survival rate drops to 29.2% (Kim et al., 2018) and the 10-year survival rate is just 15% (Lisio et al., 2019). Unfortunately, over 75% of HGSOC cases are diagnosed at an advanced stage (Lisio et al., 2019; Kim et al., 2018; Forgó et al., 2024; Kurman et al., 2016), primarily due to the lack of effective early screening methods and the late presentation of non-specific symptoms.

Understanding the molecular mechanisms of HGSOC is crucial to improve early identification and patient outcomes. This includes chromosomal alterations, mutational processes, cell origin, early cancer progression, and metastatic transition.

Several biological challenges occur when dealing with HGSOC.

The first one regards the origin of HGSCOC. Initially, it was thought to arise from the ovarian surface epithelium, which is subsceptible to DNA damage during ovulation cycles. However, recent research suggests that pre-cancerous lesions, known as serous tubal intraepithelial carcinomas (STICs), generated by epithelial cells in the distal fimbriae of the fallopian tubes, may be the site of origin of HGSOC due to their shared genetic instability. Yet, not all HGSOC cases involve the fallopian tubes, other mechanisms may potentially contribute to its development (Lisio et al., 2019; Kim et al., 2018). This uncertainty complicates early detection and often leads to late-stage diagnoses.

HGSOC's metastatic nature is another significant problem. Unlike many other cancers, tumor cells do not primarily spread through the bloodstream or lymphatic system. Instead, cells detach from the primary tumor site, disseminate within the peritoneal cavity via peritoneal fluid, and implant on nearby organs such as the ovaries and the omentum. In later stages, HGSOC may extend to the liver, lungs, or pleural space. Furthermore, patients often develop malignant ascites, where clusters of tumor cells are floating in fluid, potentially promoting metastasis and chemotherapy resistance by preventing cell death in the absence of surface attachment (Lisio et al., 2019). By the time symptoms manifest, this metastatic process is often well advanced, contributing to the high mortality rate.

Finally, both the primary and the metastatic tumor are characterized by a high degree of chromosomal instability and nearly universal TP53 mutations (Kurman et al., 2016; Forgó et al., 2024). These factors contribute to HGSOC's complex mutational profile, which leads to immune evasion and resistance to conventional therapies, making it aggressive
(Vázquez-García et al., 2022). This leads to complicated treatment approaches, frequent recurrence, and a poor prognosis.

This thesis contributes to addressing the last challenge by benchmarking tools that infer copy number alterations, essential for understanding chromosomal instability in HGSOC.

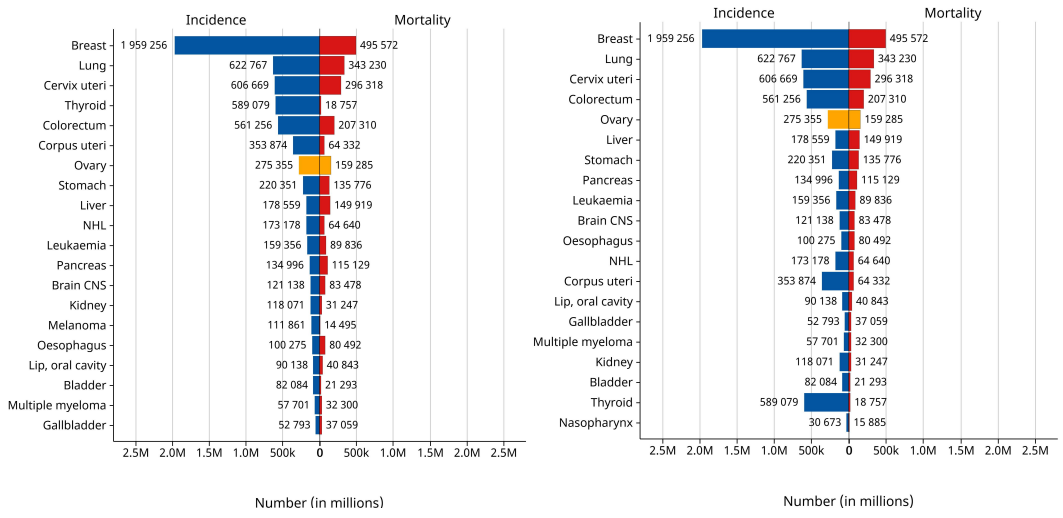


Figure 1: Absolute numbers of cancer incidence and mortality for females aged 0-74 worldwide in 2022 (Top 20 cancer sites). The left figure is ordered by incidence, while the right figure is ordered by mortality. It can be observed that there are 275,335 cases and 159,285 deaths attributed to ovarian cancer.
Source: Cancer TODAY | IARC - Globocan 2022 (version 1.1)

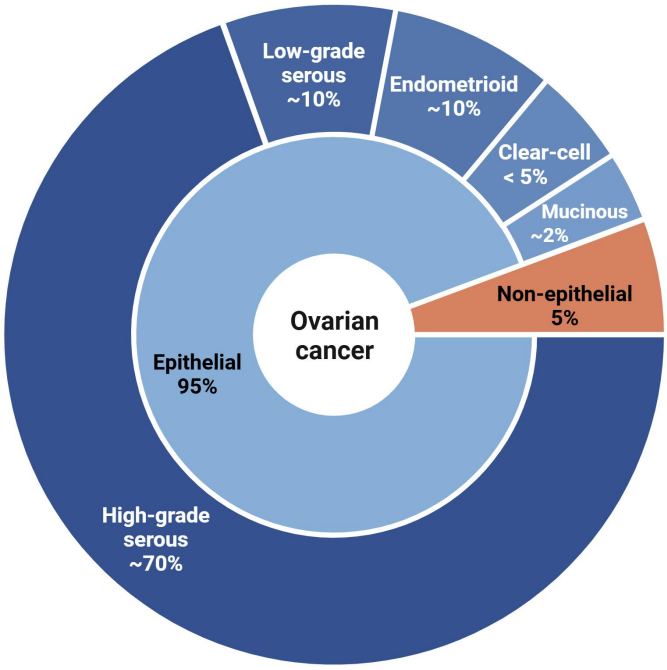


Figure 2: Ovarian cancers classification based on data in Arora T. et al. book *Epithelial Ovarian Cancer*. The percentages indicate the proportion of each ovarian cancer subtype out of the total ovarian cancer cases.

1.2 Chromosomal Instability and Copy Number Alterations

HGSOC is characterized by a high level of chromosomal instability (CIN), which contributes to its agressiveness and poor prognosis. CIN originates from various mechanisms, including frequent mis-segregation of chromosomes during mitosis, replication stress, homologous recombination deficiency, telomere crisis, and breakage-fusion-bridge cycles. These processes lead to the accumulation of structural variations in the genome, mostly copy number alterations (CNAs) (Lynch et al., 2024; Drews et al., 2022).

CNAs are changes in the number of copies of genomic regions larger than 10 kb, involving gains and losses ranging from partial chromosomal segments to entire arms or even whole chromosomes (Harbers et al., 2021).

These alterations frequently affect oncogenes and tumor suppressor genes, impacting critical cellular processes and promoting tumorigenesis. For instance, recurrent amplifications in MYC, TERT, CCNE1, and PIK3CA increase cell proliferation and survival, while deletions in BRCA1 and BRCA2 disrupt the homologous recombination DNA repair pathway, leading to further genomic instability and disease progression (Vázquez-García et al., 2022; Harber Martins et al., 2022)

The role of CNAs in cancer is not limited to initiating tumorigenesis. In fact, due to CIN, new CNAs continually emerge as the tumor evolves, leading to the coexistance of multiple genetically distinct subclones within a single tumor, each with its unique CNA profile that contributes to differential sensitivity to therapeutic agent (Lynch et al., 2024). This intra-tumor heterogeneity (ITH) poses a significant challenge in cancer treatment as it enables the tumor to adapt to selective pressures, such as chemotherapy, by allowing drug-resistant subclones to survive and expand, leading to recurrence and treatment resistance (Lynch et al., 2024; Harber Martins et al., 2022). CIN's dynamic nature ensures that new genetic alterations develop over time, creating new subclones and boosting tumor adaptability and aggressiveness (Lynch et al., 2024).

CIN, in addition to ITH, contributes to inter-tumor heterogeneity. Tumors from different HGSOC patients have different CNA profiles, reflecting diverse evolutionary trajectories (van Dijk et al., 2021). The variability of patients makes it difficult to develop standardized treatment strategies and predict therapy effects. As a result, personalized therapies based on the CNA profile of each tumor subclone of each patient are becoming increasingly important (Lynch et al., 2024; van Dijk et al., 2021).

CIN levels correlate with cancer progression, where moderate CIN enhances tumor adaptability and resistance to therapies, leading to poor prognosis, while extreme CIN leads to genomic catastrophe. (Lynch et al., 2024)(see Figure 3).

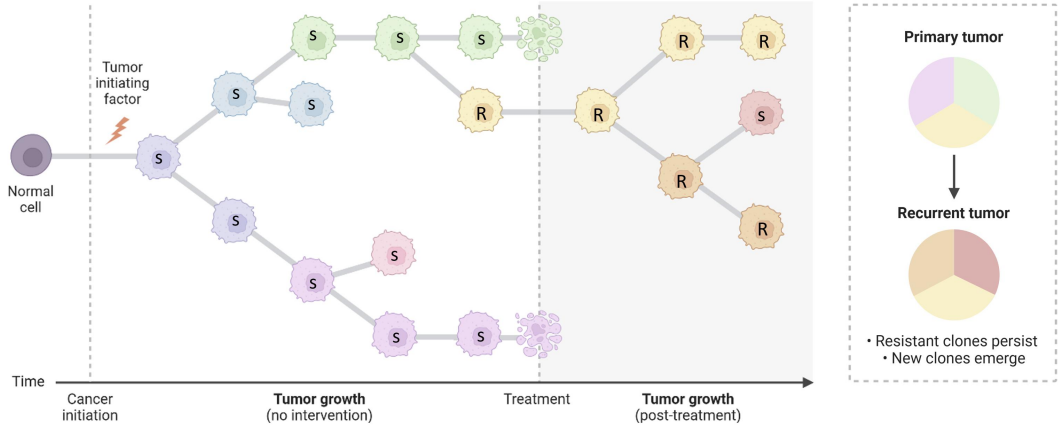


Figure 3: llustration of tumor growth and subclone selection during treatment:
A normal cell transforms into a tumor cell, which proliferates to form a clone. As the tumor grows, it accumulates mutations, giving rise to genetically distinct subclones, each represented by different colors. Treatment selectively eliminates sensitive subclones (S), but resistant ones (R) survive and continue to proliferate, leading to tumor recurrence and the formation of new subclones.

1.3 CNA signatures

To better understand tumor phenotypes and support diagnosis, prognosis, and treatment prediction, signatures derived from gene expression have been widely used (Nevins and Potti, 2007; Chibon, 2013). However, the tumor microenvironment affects gene expression, which may mask the tumor direct genetic signals (Macintyre et al., 2018).

Mutational signatures, which are genomic patterns that reflect cumulative genomic alterations caused by mutagenic processes over the course of a tumor cell's lifetime, can overcome this limitation (Macintyre et al., 2018). Unlike gene expression data, they provide a direct view of genomic alterations that are not impacted by the microenvironment (Drews et al., 2022).

Among mutational signatures, CNA signatures stand out as they capture the different patterns of CNAs that result from CIN, revealing the underlying mutational processes contributing to this genomic instability (Drews et al., 2022). This makes them particularly useful for studying complex tumors like HGSOC, where CIN plays a crucial role (Macintyre et al., 2018).

To derive meaningful insights, CNA patterns are condensed into CNA signatures by analyzing specific CNA features (Macintyre et al., 2018). In this context, Drews et al. (2022) discovered 17 CNA signatures across 33 cancer types, including ovarian cancer, by analyzing the CN difference between neighboring segments, the segment length, the number of breakpoints per 10 megabases, the number of breakpoints per chromosome arm, and the length of chains of contiguous CN segments that alternate between two CN states. Each signature reflects one or more processes that contribute to CIN, such as chromosome missegregation, telomere dysfunction, impaired homologous recombination, replication stress, impaired damage sensing, PI3K/AKT-mediated toleration, whole-genome duplication, and impaired non-homologous end joining. Understanding these CNA signatures reveals how different types of CIN contribute to cancer development, progression and resistance to treatment.

1.4 CNAs detection: from bulk DNA sequencing to single-cell RNA sequencing

Although CIN and CNAs play important roles in tumor growth and therapy resistance, their clinical application as biomarkers to inform patient prognosis or treatment selection is restricted due to challenges in measuring them accurately and efficiently (Lynch et al., 2024).

Traditional techniques to infer CNAs, such as cytogenetics and bulk sequencing of DNA or RNA, give a comprehensive picture but lack the resolution to capture subclonal variability within tumors (Lynch et al., 2024). Single-cell sequencing of DNA or RNA offers a more detailed view, revealing tumor subclones with distinct CNA patterns that bulk methods cannot detect (Kurt et al., 2024). Understanding this genomic heterogeneity is crucial for developing strategies to target therapy-resistant subclones, which frequently cause relapse (Mallory et al., 2020; Gao et al., 2023).

Although single-cell DNA sequencing (scDNA-seq) enables direct detection of CNAs at single-cell resolution, it has technical constraints, including the necessity for whole-genome amplification, which can introduce biases and uneven coverage, resuting in false positives (Mallory et al., 2020). Furthermore, scDNA-seq datasets are limited in availability and access (Kurt et al., 2024).

On the other hand, single-cell RNA sequencing (scRNA-seq) data are more widely available. CNAs cannot be detected directly from RNA, but can be inferred indirectly because they correlate with gene expression levels (Gao et al., 2023; Lynch et al., 2024). However, scRNA-seq presents its own challenges due to the complexity of the tumor microenvironment, such as the presence of immune cells and the cell-cycle state (Kurt et al., 2024), as well as sparse and noisy data (Gao et al., 2023). Despite limitations, this technology has the potential to improve clinical treatment of HGSOC and develop more targeted therapies. Therefore, several methods have been proposed to infer CNAs profiles from scRNA-seq data (Lynch et al., 2024).

1.5 Aim of the project

Several computational approaches have been developed to infer CNAs from single-cell RNA sequencing (scRNA-seq) data, however the predictions are not always consistent due to differences in algorithms and assumptions. In the context of HGSOC, inaccurate CNA profiles might lead to incorrect conclusions regarding tumor progression, therapy resistance and overall patient prognosis, interfering with clinical decision-making.

This thesis benchmarks three tools selected for their relevance in the field. InferCNV (Tickle et al., 2018) is one of the most widely used tools, while SCEVAN (De Falco et al., 2023) and Numbat (Gao et al., 2023) are more recent, actively maintained tools that in recent studies have outperformed other state-of-the-art methods, including CopyKAT and HoneyBADGER (De Falco et al., 2023; Gao et al., 2023).

By comparing the CNAs inferred from scRNA-seq data using these approaches with those derived from bulk whole-genome sequencing (WGS) data using ASCAT, this benchmark aims to determine the most accurate method for CNA detection. The reason the CNAs from bulk WGS are utilized as the ground truth is that ASCAT is considered a very reliable tool (Van Loo et al., 2010).

Some of these tools, including inferCNV, CopyKAT, CaSpER, HoneyBADGER, sciCNV, SCEVAN, Numbat, have already been benchmarked across several cancer types and cell lines in very recent studies not published yet but available as pre-print (Chen et al., 2024; Minfang et al., 2024). While these studies provide a broad benchmarking, this thesis specifically addresses the need for specific research on HGSOC, using patient-derived samples to reflect the complexity of the tumor microenvironment. By focusing on chromosomal instability and tumor heterogeneity in HGSOC, this work fills a gap in the literature and paves the way for future research in this aggressive cancer.

An additional goal of this thesis is to assess whether the CN profiles inferred from scRNA-seq-based tools are sufficiently consistent with those obtained by ASCAT from WGS data (considered the ground truth) to be used for quantifying the 17 CNA pan-cancer signatures discovered by Drews et al. (2022). This analysis also aims to identify which tool performs best for signature quantification. Demonstrating such consistency would support the use of scRNA-seq data for studying CIN in HGSOC.

2. Materials and Methods

2.1 Datasets overview

This thesis uses two datasets derived from the Vázquez-García et al. study (2022) conducted at Memorial Sloan Kettering Cancer Center (MSK), which investigates the influence of mutational processes and anatomical tumor sites on immune evasion mechanisms in HGSOC using multimodal approaches, including scRNA-seq and bulk WGS.

The original cohort comprised 42 patients. However, this analysis focused on 39 patients with paired scRNA-seq data and tumor-normal bulk WGS samples. Patient SPECTRUM-OV-004 was excluded due to the lack of scRNA-seq data, while SPECTRUM-OV-071 and SPECTRUM-OV-090 were excluded due to the lack of tumor bulk WGS samples. Among the 271 scRNA-seq samples collected from multiple sites in these 39 patients, only 128 CD45- and 6 Unknown-sorting samples were included, while the CD45+ samples were excluded from the analysis.

Samples collection

Tissue samples were obtained from various anatomical sites (adnexa, omentum, peritoneum, bowel, ascites, and other intra-peritoneal sites) of treatment-naive HGSOC patients during laparoscopic biopsies or debulking surgeries.

In addition to the tissue biopsies, blood samples were collected pre-surgery to isolate peripheral blood mononuclear cells (PBMCs), which were used for normal bulk WGS, ensuring a tumor-normal matched WGS dataset for each patient.

Samples processing

- scRNA-seq processing:
  - Fresh tumor tissue were dissociated into single cells immediately after collection.
  - Single-cell suspensions were sorted into CD45+ (immune cells) and CD45- (tumor cells) using fluorescence-activated cell sorting.

  - Sequencing libraries containing 1400-5000 cells per sample were prepared using the Chromium Single-Cell 3' Reagent kit v3 (10x Genomics) and sequenced on the Illumina HiSeq 2500 or NovaSeq 6000.

- Bulk WGS processing:
  - Fresh tumor tissue was snap-frozen.
  - DNA was extracted from microdissected tumor sections (for tumor WGS) and from PBMCs isolated from blood samples (for normal WGS).
  - Sequencing libraries were prepared using the KAPA Hyper Prep Kit and sequenced on the Illumina NovaSeq 6000.

These datasets were selected because they provided both genomic and transcriptomic profiles from the same cohort of HGSOC patients, allowing the use CNAs inferred from WGS as a ground truth for benchmarking. Additionally, the samples were treatment-naive, meaning no alterations in the tumor and microenvironment composition were caused by prior therapies.

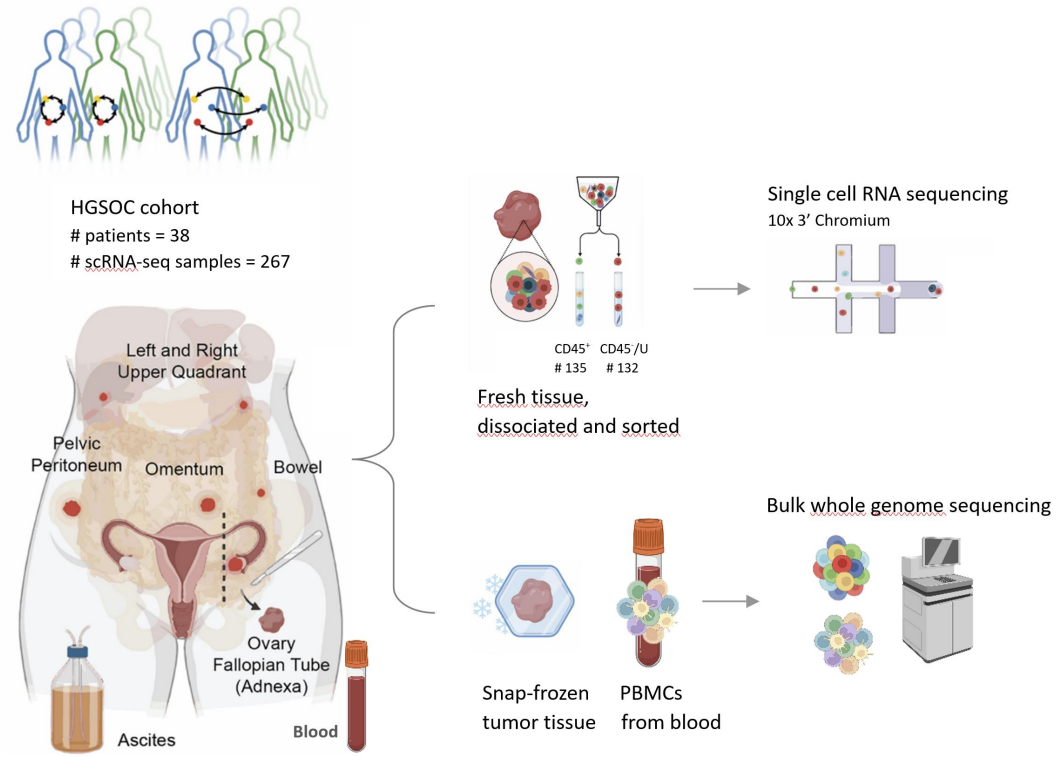


Figure 4: Data collection and processing.
Source: inspired by the work of Vázquez-García et al. (2022)

2.2 CNAs inference from scRNA-seq

In this section, InferCNV (Tickle et al., 2018), SCEVAN (De Falco et al., 2023), and Numbat (Gao et al., 2023) are presented with a focus on their inputs, workflows, outputs, and the parameter settings used in this benchmark. For simplificy, only the outputs relevant to the benchmark are included.

2.2.1 Main inputs

The raw count matrix is the main input for InferCNV, SCEVAN, and Numbat. Depending on the tool, additional information on cells types may also be required. These inputs are processed differently depending on the tool, with varying format requirements. Below is overview of both an explanation of how they were obtained.

scRNA-seq data from multiple anatomical sites of the same patients were merged to ensure comprehensive coverage of tumor heterogeneity and to increase the pool of cells. This merging step resulted in higher sequencing depth, which in turn improved the reliability of cell type classification and CNA inference.

Raw count matrix

The raw count matrix contains gene expression data, with rows corresponding to genes and columns to individual cell barcodes. The values within the matrix indicate the number of unique molecular identifiers (UMIs) mapped to each gene for a specific cell.

The matrix is generated using the Cell Ranger pipeline from 10x Genomics, which processes the raw FASTQ files to quantify gene expression.

As is common in scRNA-seq data, the matrix is sparse, meaning most values are zero, as shown in the example table below:

	GSM5467087_AAACCCAAGTGATAGT.1	GSM5467087_AAACCCACATTCACCC.1	GSM5467087_AAACCCAGTCCAAGAG.1
MIR1302-2HG	0	0	0
FAM138A	0	0	0
OR4F5	0	0	0
AL627309.1	0	0	0
AL627309.3	0	0	0
LINC01128	0	0	1
LINC00115	0	0	0
AL645608.6	0	1	0

Before analysis, this matrix undergoes preprocessing steps, such as filtering out low-quality cells and genes with low expression. The genes are then ordered by their genomic coordinates.

It's important to note that while Numbat requires `tumor_matrix`, a modified version of the matrix that excludes normal cells, InferCNV and SCEVAN use the full count matrix, which includes both normal and tumor cells.

Cell type information

To differentiate between normal and malignant cells, cell type annotations are required, though the specific format depends on the tool being used.

- **InferCNV**: Requires a header-free, tab-delimited `annotations_file` that maps cell barcodes to their cell type classification labels (malignant or specific normal cell type) and a `ref_group_names` vector containing the cell type classification labels of cell types to be used as reference.
  - Example of the `annotations_file`:

GSM5467087_AAACCCAAGTGATAGT.1	malignant_GSM5467087
GSM5467087_AAACCCACATTCACCC.1	malignant_GSM5467087
GSM5467087_AAACCCAGTCCAAGAG.1	malignant_GSM5467087
GSM5467087_AAACGAAGTCTTGAAC.1	malignant_GSM5467087
GSM5467087_AAACGAATCACGTCCT.1	Endothelial.cell
GSM5467087_AAAGAACTCCATAAGC.1	malignant_GSM5467087

  - Example of the `ref_group_names` vector:

```
c("Endothelial.cell", "Fibroblast", "other", "T.cell", "Myeloid.cell")
```

- **SCEVAN**: Requires a `norm_cell` vector containing barcodes of normal cells.
  Example of the `norm_cell` vector:

```
c("GSM5467087_AAACGAATCACGTCCT.1", "GSM5467087_AAAGAACTCGAGATGG.1",
"GSM5467087_AAAGGATCATTCATCT.1", "GSM5467087_AAAGTGAAGGCTCAAG.1", ...)
```

- **Numbat**: Requires a `lambdas_ref` reference matrix, with rows corresponding to genes and columns to normal cell types. The marix values represent gene expression values for each normal cell type, normalized such that the total expression within each cell type group sums to 1.
  Example of the `lambdas_ref` matrix:

	Endothelial.cell	Fibroblast	Myeloid.cell	other	T.cell
MIR1302-2HG	0	0.000000e+00	0	0	0
FAM138A	0	0.000000e+00	0	0	0
OR4F5	0	0.000000e+00	0	0	0
AL627309.1	0	1.768628e-07	0	0	0
AL627309.3	0	0.000000e+00	0	0	0

While InferCNV requires the annotation file to distinguish between malignant and normal cells, SCEVAN and Numbat can classify cells internally using their algorithms. However, for consistency in this benchmark, the same cell type information was provided a priori to all tools.

To classify cell types, the `CellAssign` tool (version 0.99.2) (Zhang et al., 2019) was applied to the merged scRNA-seq data for each patient. The marker genes used by `CellAssign` were derived from a previous study (Vázquez-García et al., 2022). The output is formatted the same way as the `annotations_file` input required by InferCNV.

2.2.2 InferCNV

Overview

InferCNV (Tickle et al., 2018) compares gene expression levels of tumor cells to those of reference normal cells to identify CNAs. The tool uses a six-state Hidden Markov Model (HMM) to detect CNAs and a Bayesian Network to refine the results and reduce false positives.

Inputs

- `raw_counts_matrix` (see Main inputs - raw counts matrix)
- `annotations_file` and `ref_group_names` (see Main inputs - cell type information)
- `gene_order_file`: provides the chromosomal locations for many genes, including all those present in the count matrix. For this benchmark, a pregenerated file was used, available at [TrinityCTAT]
  The file lists each gene's name along with its chromosome and start/end coordinates, as shown in the example below:

DDX11L1	chr1	11869	14412
WASH7P	chr1	14363	29806
FAM138A	chr1	34554	36081

Method workflow

1. COUNTS MATRIX PREPROCESSING: Genes expressed in fewer than a minimum number of cells (`min_cells_per_gene`) are removed from the counts matrix. Read counts are scaled to sum the median total read count across all cells to adjust for sequencing depth, and a log-transformation ($log_2(counts + 1)$) is applied to reduce the variance associated with higher mean values.

2. CENTERING BY NORMAL GENE EXPRESSION: For each gene, the mean expression value across reference (normal) cells is subtracted from the expression values of all cells in log space, yielding log-fold changes relative to normal expression. These values are thresholded so that log-fold changes exceeding a predefined limit (default=3) are capped.

3. SMOOTHING BY CHROMOSOME: Expression levels of genes ordered along each chromosome are smoothed using a weighted running average (default window size: 101 genes), reducing noise. The mean expression level of normal cells is again subtracted from tumor cells to correct discrepancies introduced during the smoothing.

4. CENTERING BY CELL: The median expression for each cell is centered at zero, under the assumption that most genes are not in CNA regions.

5. LOG TRANSFORMATION REVERSION: The log transformation applied earlier is reversed to make gains and losses more symmetrical around the mean.

6. CNAS CALLING: Application of a six-state HMM to model CNAs, with the following CNA states: complete loss (state 1), loss of one copy (state 2), neutral (state 3), addition of one copy (state 4), addition of two copies (state 5), and a placeholder for amplifications beyond two copies (state 6).

7. SIMULATED DATA: Generation of simulated scRNA-seq data, referred to as "hspike" data. This dataset is constructed based on characteristics of the input normal (reference) cells, matching both the mean/variance expression intensity trend and zero-inflation properties. Tumor cells are simulated to have chromosomal regions with CNA levels corresponding to each of the six states. The variance for each CNA state is estimated based on the sample size (or subclone size, depending on the analysis mode) and is standardized across all CNA states to the median variance of all hspike distributions. The hspike data goes through the same processing steps as the real data, and are used to calibrate the HMM's emission probabilities for each CNA state.

8. BAYESIAN NETWORK ADJUSTMENT: A Bayesian Network is applied to compute the posterior probability of each CNA being in a specific state, filtering out false positives by assigning low confidence to regions that are likely normal, despite HMM predictions.

9. FINAL FILTERING OF FALSE POSITIVES: CNAs with a high posterior probability of being normal (typically above 0.5) are filtered out to reduce false positives.

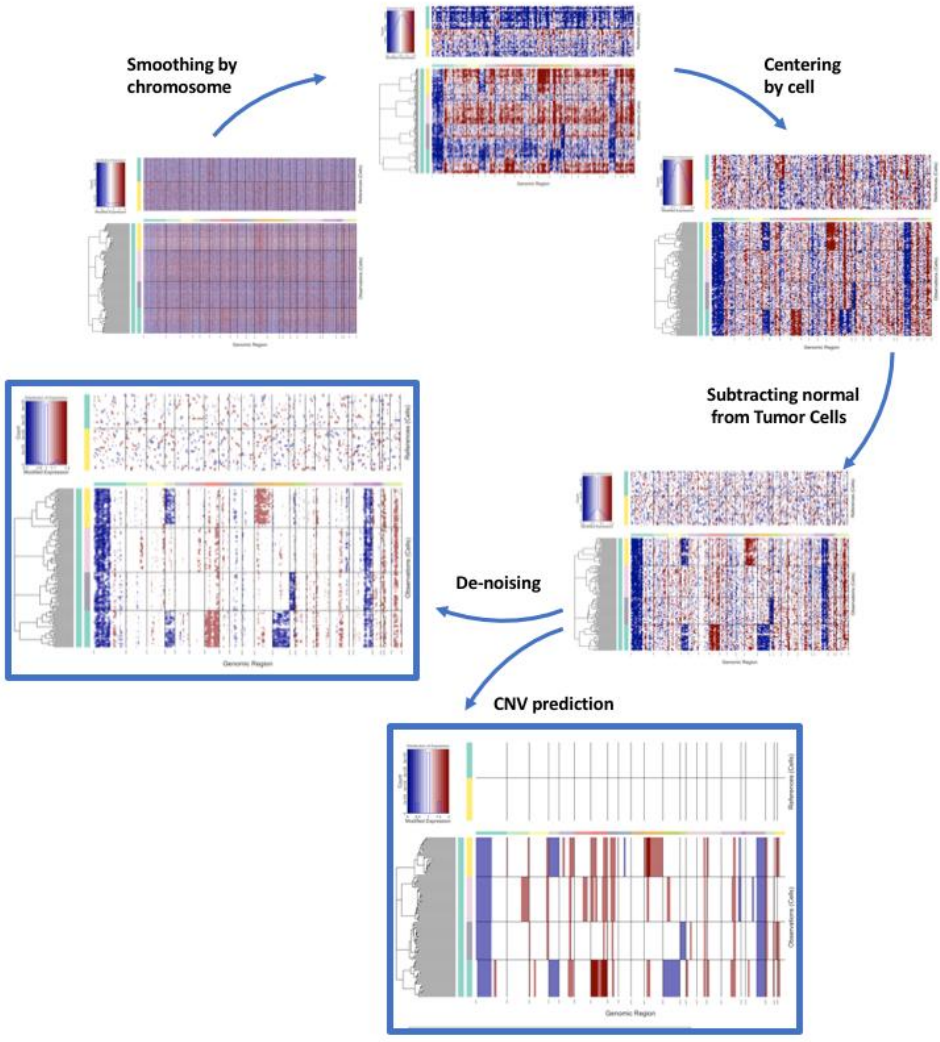


Figure 5: InferCNV workflow
Source: Tickle et al. (2019)

Output

- Segmentation file (`HMM_CNV_predictions.*.pred_cnv_regions.dat`) containing the group of cells affected by the CNA, CNA genomic region coordinates (chromosome, start and end posisions), and copy number state assignments. If the HMM is run on all cells for a given sample, the `cell_group_name` 'all_observations.all_observations_s1' corresponds to all tumor cells in the sample. However, if subcluster analysis is used, the sample is divided into subclusters.

  An example format with `analysis_mode = 'samples'` is shown below:

cell_group_name	cnv_name	state	chr	start	end
all_observations.all_observations_s1	chr1-region_2	4	chr1	6221193	7781432
all_observations.all_observations_s1	chr1-region_4	4	chr1	215311817	43991170
all_observations.all_observations_s1	chr2-region_7	4	chr2	6221193	217756593
all_observations.all_observations_s1	chr3-region_10	2	chr2	62318973	99799226

  An example format with `analysis_mode = 'subclusters'` is shown below:

cell_group_name	cnv_name	state	chr	start	end
all_observations.all_observations_s2	chr1-region_2	2	chr1	32247233	36464437
all_observations.all_observations_s2	chr1-region_4	3	chr1	53916574	70205620
all_observations.all_observations_s1	chr1-region_68	4	chr1	3772788	9196983
all_observations.all_observations_s1	chr4-region_88	2	chr4	55853616	93810157
all_observations.all_observations_s10	chr6-region_273	4	chr6	33299694	38703141

Parameters used in this benchmark

The `hg38_gencode_v27.txt` file was used as the `gene_order_file` to create the InferCNV object. InferCNV (version 1.20.0) was then run with the following key parameters: `cutoff = 0.1`, `min_cells_per_gene = 3`, `cluster_by_groups = FALSE`, `denoise = TRUE`, `HMM = TRUE`, `analysis_mode = 'samples'`. The `analysis_mode = 'samples'` option ensures that CNV prediction is performed by grouping all malignant cells from a single patient together, rather than analyzing smaller subclusters of malignant cells.

2.2.3 SCEVAN (Single-Cell Evolutionary Variational ANalysis)

Overview

SCEVAN (De Falco et al., 2023) segments the genome using a variational approach and performs joint segmentation across cells within the same clone, assuming that cells within a clone share similar CNAs breakpoints. Thus, the expression profile of every individual cell, seen as a function of the genomic coordinates, contributes to the evidence of CNAs in each subclone.

Inputs

- `count_mxt` (see Inputs - raw counts matrix)
- `norm_cells` (see Inputs - cell type information)

Method workflow

1. COUNTS MATRIX PREPROCESSING: cells with fewer than 200 detected genes, genes expressed in fewer than 1% of cells, and genes involved in the cell cycle are removed from the counts matrix. The remaining genes are annotated with genomic locations and sorted by genomic coordinates. The Freeman-Tukey transformation (Freeman et al., 1950) is applied to the read counts to reduce the variance associated with higher mean values. Then, each gene is scaled by subtracting its mean to center the data around zero.
2. CONFIDENT NORMAL CELLS IDENTIFICATION: Confident normal cells are identified using gene expression signatures of various types of normal cells from public collections. However, when the vector `norm_cells` is provided by the user, a priori information can be used.
3. CENTERING BY NORMAL GENE EXPRESSION: For each gene, the median expression value across reference (normal) cells is subtracted from the expression values of all cells, yielding a matrix of values relative to normal expression.
4. SMOOTHING: Application of edge-preserving smoothing to the gene expression data. This process reduces noise while preserving sharp transitions between different CNA states, which correspond to chromosomal breakpoints.
5. JOINT SEGMENTATION: A greedy multichannel segmentation algorithm identifies the boundaries of homogeneous copy number regions across all cells in a clone, resulting in a CNA matrix where expression values between breakpoints are averaged.

A Mumford-Shah energy model (Mumford et al., 1989) is applied to minimize the number of breakpoints, where adjacent regions are iteratively merged based on segment size and mean expression differences. The segmentation process assumes that cells in the same clone share similar copy number breakpoints. By segmenting across multiple cells at once, the tool ensures that CNAs are detected at the clone level, reducing the impact of single-cell noise.

6. CLASSIFICATION OF MALIGNANT AND NON-MALIGNANT CELLS: The CNA matrix obtained with segmentation is splitted into two groups using hierarchical clustering, with cells in the cluster having the highest number of confident normal cells classified as non-malignant. The final matrix is obtained by subtracting again the mean expression values of the non-malignant cells.

7. SUBCLONAL STRUCTURE CHARACTERIZATION: Malignant cells are clustered based on their CNA profiles. Joint segmentation is applied to each subclone, identifying alterations that are subclone-specific, shared between some subclones or clonal. Alterations in different subclones are considered the same if breakpoints are within 10 Mb and differ in size by less than 40%.

8. CNAS CALLING: A 5-states mixture model-based algorithm, with each component defined as a truncated normal distribution, is used to assign copy number states (0 = deletion, 1 = loss, 2 = neutral, 3 = gain, 4 = amplification) to each segmented region. Clonal CN profiles can be obtain from all tumour cells (`ClonalCN = TRUE`) or from each subclone.

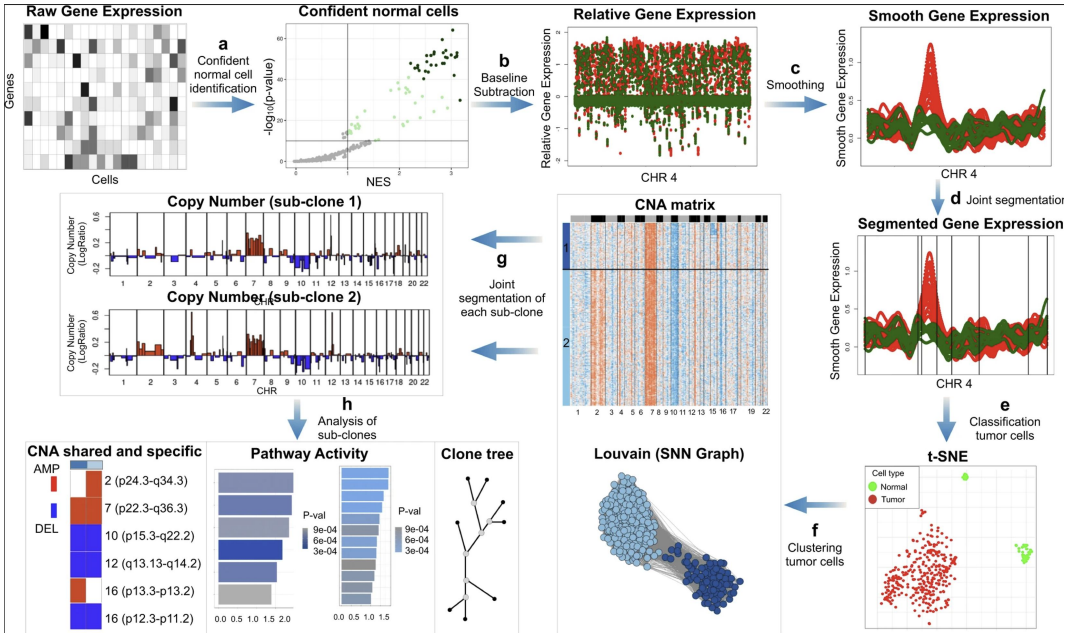


Figure 6: SCEVAN workflow
Source: (De Falco et al., 2023)

Output

- Segmentation file (`*_Clonal_CN_seg`) containing the identified CNAs and their corresponding genomic region coordinates (chromosome, start and end positions), copy number state assignments, and the segment mean.
  An example format is shown below:

Chr	Pos	End	CN	segm.mean
1	825138	151249536	2	0.020223
1	151281618	161964070	4	0.162726
1	161983192	248859144	2	0.02111
2	38814	69674349	2	0.004106
2	69644425	134918710	3	0.102261

Parameters used in this benchmark

The `pipelineCNA` function in SCEVAN (version 1.0.1) was run with the following key parameters: `ClonalCN = TRUE`, `norm_cell = norm_cell`. The `ClonalCN = TRUE` option ensures that clonal CNA prediction is performed by grouping all tumor cells from a single patient together, while `norm_cell = norm_cell` ensures that a vector of a priori identified normal cells is used as a reference.

2.2.4 Numbat

Overview

Numbat (Gao et al., 2023) uses an haplotype-aware Hidden Markov Model (HMM) that integrates scRNA-seq data, allelic ratios, and population-derived haplotype information to detect allele-specific CNAs. Differently from InferCNV and SCEVAN, it is able to distinguish allele-specific events, like loss of heterozygosity (LOH) and biallelic amplification (bAMP).

Inputs

- `count_mat` (see Main inputs - raw counts matrix)
- `lambdas_ref` (see Main inputs - cell type information)
- `allele_df`: dataframe containing phased allele counts per single nucleotide polymorphisms (SNP)

Method workflow

1. PILEUP AND PHASING: Allele counts are generated for known SNPs using `cellsnp-lite`, and heterozygous SNPs are phased into maternal and paternal haplotypes using `Eagle2` based on reference panels (e.g., 1000 Genomes). The result is a phased allele dataframe containing SNP positions, genotypes, and allele counts.

2. COUNTS MATRIX PREPROCESSING: cells with a total gene expression of zero and low-expressed genes are removed from the counts matrix. Each gene's raw counts are normalized by the total counts in each cell to account for differences in sequencing depth between cells. A log transformation is applied to the read counts to reduce the variance associated with higher mean values. The transformation is scaled by a factor of 10^6 to bring normalized counts to a readable scale.

$$\text{count}_{\text{gene, cell}} = \log\left(\frac{\text{raw\_count}_{\text{gene, cell}}}{\sum_{\text{gene}} \text{raw\_count}_{\text{gene, cell}}} \times 10^6 + 1\right)$$

   Reference expression profiles from matched normal cells or external datasets are used to build the `lambdas_ref` matrix.

3. CNAS CALLING: A 15-states haplotype-aware HMM is used to detect CNAs across the genome. This HMM integrates gene expression (to detect copy number gains or losses), ratio of reference to alternate alleles at heterozygous SNP sites (deviations from the expected 1:1 ratio in diploid cells indicate the presence of CNAs), and phased haplotype information (to differentiate between monoallelic and biallelic CNAs and thus to detect allele-specific events, such as LOH or bAMPs).
   The HMM models the likelihood of the observed data under 15 states, capturing a continuum of copy number variations and haplotype fractions. It assigns probabilities to different states (deletion, LOH, neutral, bAMP, amplification) and then computes the most likely state for each genomic region.

4. FILTERING OF FALSE POSITIVES: CNAs with a log-likelihood ratio > 5 or a posterior probability < 0.5 are filtered out, reducing false positives.

5. CLONAL PHYLOGENY RECONSTRUCTION: Malignant cells are clustered based on their CNA profiles, and a distance matrix is used to infer subclone relationships. Numbat generates a phylogenetic tree that shows the evolutionary relationships among subclones, with CNAs annotated at each evolutionary step.

6. PHYLOGENY RECONSTRUCTION AND SUBCLONES IDENTIFICATION: The inferred CNAs are used to build a maximum-likelihood phylogeny of the cells, capturing clonal and subclonal structure. Numbat identifies distinct subclones within the tumor by grouping cells based on their phylogenetic relationships.

7. FINAL CNA CALLS: After iterative optimization of the phylogeny and CNAs detection, Numbat outputs a consensus segmentation file and joint posterior probabilities of CNV states for each genomic segment in each cell.

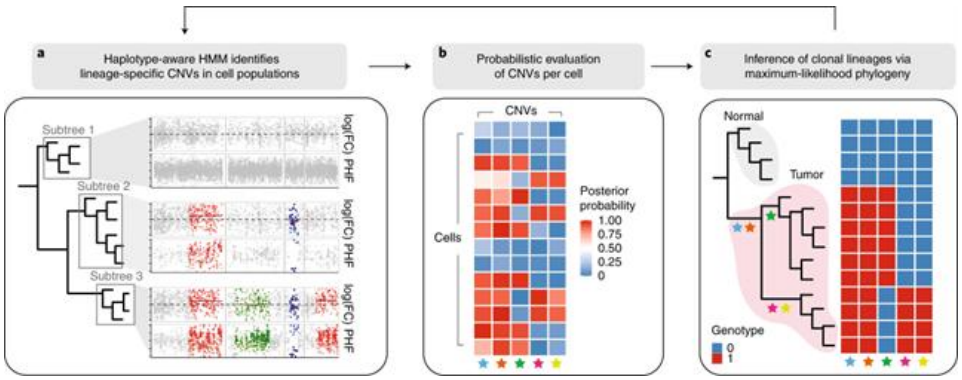


Figure 7: Numbat workflow
Source: Gao et al. (2023)

Output

- Segmentation file (`segs_consensus_i.tsv.gz`) containing the identified CNAs and their corresponding genomic region coordinates (chromosome, start and end positions), copy number state assignments, and additional information that is not relevant for this analysis.
  An example format displaying only the columns of interest is shown below:

CHROM	cnv_state_post	seg_start	seg_end
1	loh	779047	115691884
1	neu	115747427	248906235
2	neu	8814	241883646
3	loh	209229	186106855
3	loh	86139752	195529015
3	amp	195543189	197960200

Parameters used in this benchmark

To generate the phased allele dataframe `allele_df` using `pileup_and_phase.R`, the following files were downloaded: `gmap = genetic_map_hg38_withX.txt`, `snpvcf = genome1K.phase3.SNP_AF5e2.chr1toX.hg38.vcf`, `paneldir = 1000G_hg38`.

Numbat (version 1.4.2) was then run with the following key parameters: `genome = 'hg38'`, `t = 1e-5`, `lambdas_ref = lambdas_ref`. The `lambdas_ref = lambdas_ref` option ensures that a priori identified normal cells are used as a reference.

2.3 CNAs inference from WES - ASCAT (Allele - Specific Copy Number Analysis of Tumors)

Overview

ASCAT (Van Loo et al., 2010; Ross et al., 2021; Van Loo et al., 2012) is an algorithm designed to automate the discovery of tumor ploidy (the average number of DNA copies in tumor cells) and purity (fraction of tumor cells in the sample) to derive allele-specific CNAs from bulk WGS data.

ASCAT relies on LogR (log2 intensity ratio, a measure of total copy number) and BAF (B-allele frequency) values, which can be extracted from WGS BAM files. These data reveal imbalances that differentiate the tumor genome from a normal genome and help determine whether a sample is actually aneuploid (Figure 8).

In solid tumors, there is often a mixture of normal and tumor cells leading to mosaic CNAs profiles in LogR and BAF data. ASCAT addresses both aneuploidy and non-aberrant cell mixtures simultaneously.

In earlier versions, ASCAT required matched normal WGS samples to distinguish tumor-specific CNAs from germline variations, providing a baseline of normal genomic variation. Although later versions removed this requirement, this benchmark includes matched normal samples to improve the accuracy of CNA detection.

Inputs

- Tumor BAM file
- Normal BAM file

Note: A BAM file stores aligned sequences from WGS data in a binary format. It includes a header section with metadata about the sequencing run and an alignment section with details for each read aligned to the reference, such as position, mapping quality, and a CIGAR string, which describes the alignment pattern (e.g. matches, insertions, deletions). It is compressed for efficient storage and indexed (via `.bai` files) to enable fast retrieval of alignments within specific regions (Marshall, 2024).

Method workflow

1. LogR and BAF data extraction: LogR and BAF data are generated for each SNP locus in both tumor and normal cells from WGS BAM files.
2. LogR correction: LogR values are corrected for GC content and replication timing to ensure that the total copy number measurements are adjusted for potential biases in sequencing data.
3. Segmentation: The Allele-Specific Piecewise Constant Fitting (ASPCF) algorithm segments the genome into regions based on differences in copy number and allelic balance (using the LogR and BAF data).
4. Estimation of purity and ploidy: A sunrise plot is generated to determine the best estimates for tumor purity and ploidy by evaluating different possible values. The matched normal sample improves the accuracy of these estimates, ensuring that only tumor-specific CNAs are analyzed.
5. Allele-specific CNAs calling: The number of maternal and paternal copies in tumor cells is determined for each segment of the genome by combining the segmented LogR and BAF data with the purity and ploidy estimates.

Output

- Tumor purity and ploidy estimates.
- Segmentation file (`segments_raw.txt`) containing the identified CNAs, their corresponding genomic region coordinates (chromosome, start and end positions), and raw and rounded copy number state assignments for both minor and major alleles. An example format is shown below:

sample	chr	startpos	endpos	nMajor	nMinor	nAraw	nBraw
SPE-OV-002	1	809641	3964554	2	1	2.4933291	1.162646898
SPE-OV-002	1	3965681	16678804	2	1	2.3824587	1.092205206
SPE-OV-002	1	16678933	18180875	3	1	3.3821671	1.015665293
SPE-OV-002	1	18182422	30389876	2	1	2.1023286	1.055509309
SPE-OV-002	1	30389957	30394257	1	0	1.0154774	0.342194144
SPE-OV-002	1	30394908	50397494	2	1	2.0524650	1.018038938

Parameters used in this benchmark

The allele-specific copy number data (LogR and BAF) was generated from the WGS BAM files of both tumor and matched normal samples using the ASCAT (version 3.1.3) function `ascat.prepareHTS` with the following key parameters:

- `alleles.prefix = "path/G1000_alleles_hg38_chr"`
- `loci.prefix = "path/G1000_loci_hg38_chr"`
- `gender = "XX",`
- `genomeVersion = "hg38"`
- `chrom_names = seq(22)`
- `ref.fasta = "path/Homo_sapiens.GRCh38.dna.primary_assembly.fa".`

ASCAT algorithm was run using the function `ascat.runAscat` on an ASCAT object that had previously undergone GC content and replication timing correction and segmentation, with the following key parameters: `gamma = 1` and `write_segments = TRUE"`. A gamma value of one is standard for sequencing data, ensuring the correct interpretation of LogR shifts for a one-copy loss.

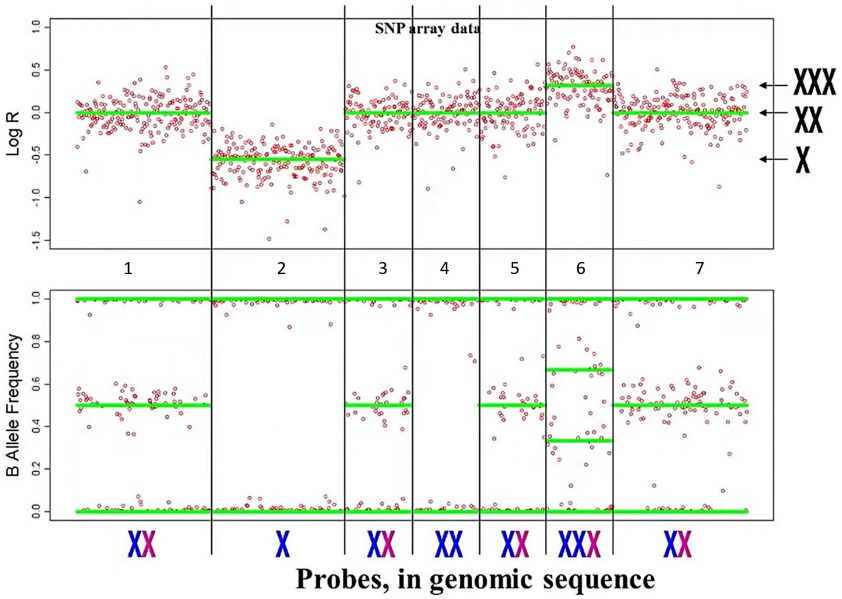


Figure 8:

- Panel 1, 3, 5, 7: In a normal diploid situation, denoted by the blue X for the paternal allele and the pink X for the maternal allele, the LogR is zero, and there are three distinct BAF bands.
- Panel 2: When an allele is lost (e.g., the maternal allele), the LogR decreases, and heterozygosity (middle BAF band) is lost, leaving only homozygous bands.
- Panel 4: In cases of copy-neutral LOH, one allele is lost, and the other is duplicated, leading to homozygosity with LogR = 0.
- Panel 6: Gains, such as trisomy, increase the LogR and result in four distinct BAF bands.

Source: van Loo video file (2013)

2.4 Tools benchmarking

To assess the ability of InferCNV, SCEVAN, and Numbat in detecting CNAs, a detailed comparison was conducted between the CN regions inferred by these tools, representing their predictions, and those identified using bulk WGS data analyzed with ASCAT, which served as the ground truth. The comparison was done at the base-pair level, allowing to capture even the smallest discrepancies between predicted and actual CNAs across the genome.

ASCAT's CNAs are obtained by summing the number of copies of the minor and major alleles, and are adjusted for ploidy to ensure comparability with outputs from other tools.

For all benchmarked tools and ASCAT, including ASCAT, regions with losses of one or two copies are labeled as "loss," while neutral regions or those with LOH are labeled as "neutral," and amplifications of one or more copies, whether balanced or unbalanced, are labeled as "gain."

Since InferCNV does not explicitly call regions it assigns a neutral copy number, uncalled regions from InferCNV that are called by ASCAT are labeled as "neutral." On the other hand, SCEVAN and Numbat do call neutral regions, and any regions uncalled by these tools but identified by ASCAT were labeled as "NA."

Regions called by scRNA-seq based tools but not by ASCAT were excluded from the analysis, as their ground truth state is unknown, making it impossible to assess the correctess of these predictions.

Classification metrics

For each CNA state (loss, neutral, or gain) classification metrics (True Positives TP, False Positives FP, True Negatives TN, False Negatives FP) were computed based on the length of the segments in base pairs (bp), rather than the number of segments. This approach better reflects the physical size of the alterations detected and ensures a more precise assessment of performance: longer regions have a greater impact on the metrics than smaller regions.

The classification metrics for each state are defined as follows (see Figure 9a and 10):

- For losses predicted by scRNA-seq based tools:
  - TP: Regions classified as loss using WGS data and as loss using scRNA-seq data.
  - FP: Regions classified as neutral or gain using WGS data and as loss using scRNA-seq data.
  - TN: Regions classified as neutral or gain using WGS data and as neutral, gain or no-call using scRNA-seq data.
  - FN: Regions classified as loss using WGS data and as neutral, gain or no-call using scRNA-seq data.
- For neutrals predicted by scRNA-seq based tool:
  - TP: Regions classified as neutral using WGS data and as neutral using scRNA-seq data.
  - FP: Regions classified as loss or gain using WGS data and as neutral using scRNA-seq data.
  - TN: Regions classified as loss or gain using WGS data and as loss, gain or no-call using scRNA-seq data.
  - FN: Regions classified as neutral using WGS data and as loss, gain or no-call using scRNA-seq data.
- For gains predicted by scRNA-seq based tool:
  - TP: Regions classified as gain using WGS data and as gain using scRNA-seq data.
  - FP: Regions classified as loss or neutral using WGS data and as gain using scRNA-seq data.
  - TN: Regions classified as loss or neutral using WGS data and as loss, neutral or no-call using scRNA-seq data.
  - FN: Regions classified as gain using WGS data and as loss, neutral or no-call using scRNA-seq data.

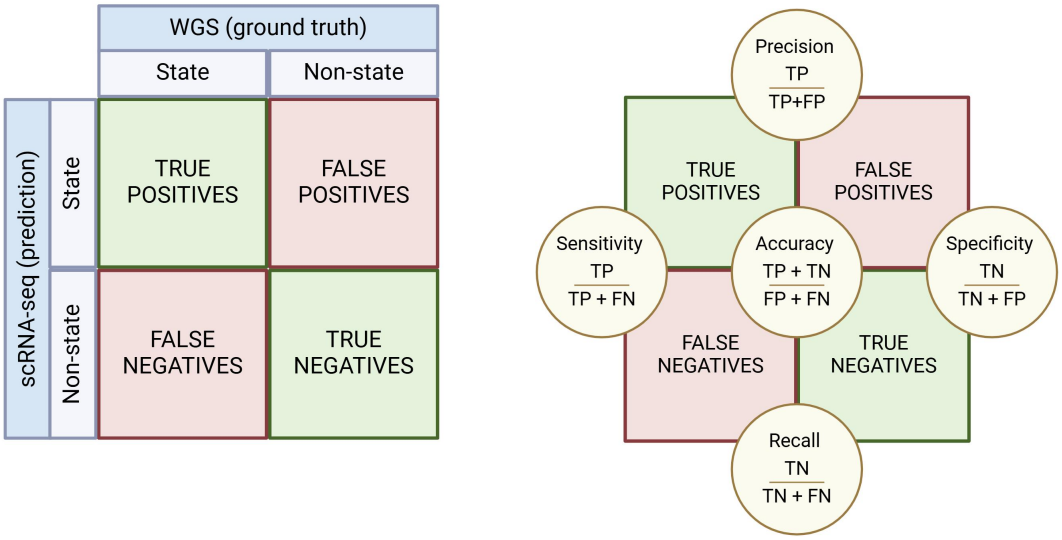
Performance metrics

Beyond the raw classification metrics (TP, FP, TN, FN), performance metrics were calculated to provide a more intuitive understanding of the tools' overall performance (see Figure 9b). These metrics were computed for each copy number state and include:

- Sensitivity: The proportion of actual positive regions that were correctly identified by the tool $\left(\frac{\text{TP}}{\text{TP+FN}}\right)$.
- Specificity: The proportion of actual negative regions that were correctly identified by the tool $\left(\frac{\text{TN}}{\text{TN+FP}}\right)$.
- Positive predictive value (PPV): The proportion of predicted positive regions that were truly positive $\left(\frac{\text{TP}}{\text{TP+FP}}\right)$.
- Negative predictive value (NPV): The proportion of predicted negative regions that were truly negative $\left(\frac{\text{TN}}{\text{TN+FN}}\right)$.
- Accuracy: The proportion of true results over the total number of regions $\left(\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}\right)$.

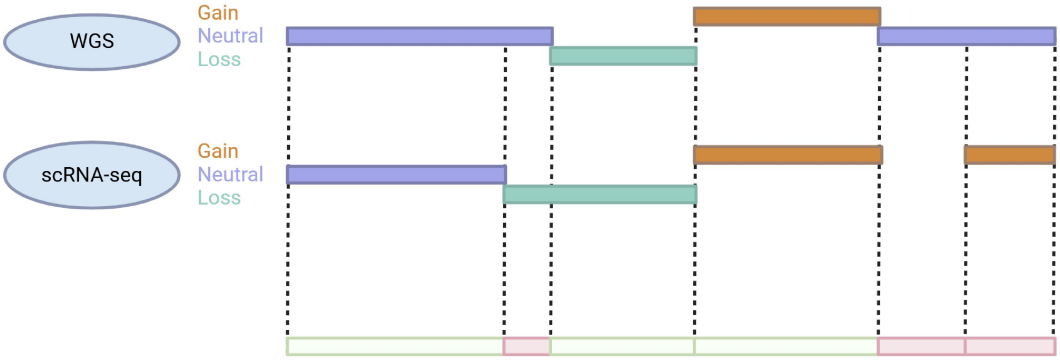
Additionally, an overall accuracy metric was calculated by considering the cumulative length of regions where predictions from WGS and scRNA-seq data were concordant. Specifically, it is computed by dividing the total length of regions where both ASCAT and tools using scRNA-seq data classified the regions as either gain, loss, or neutral (i.e., green regions in Figure 10) by the total length of all regions with a call in WGS (green and red regions combined).

In an ideal scenario, where the tool credictions perfectly agree with the ground truth, each performance metrics would be as close to 1 as possible, reflecting maximum TP and TN rates and minimal FP and FN rates.



(a) (b)

Figure 9:
(a) Classification metrics: TP, FP, TN, and FN. The "state" can represent loss, neutral, or gain, and "non-state" can represent non-loss, non-neutral, or non-gain. ASCAT serves as the ground truth, while the predictions come from tools using scRNA-seq data.
(b) Performance metrics: accuracy, sensitivity, specificity, PPV, and NPV.



Classification metrics	For losses	TN	FP	TP	TN	TN	TN
	For neutrals	TP	FN	TN	TN	FN	FN
	For gains	TN	TN	TN	TP	TN	FP

Figure 10: Visual representation of state concordance between regions inferred by ASCAT and by tools using scRNA-seq data. Green segments indicate regions where the predictions agree, while red segments indicate regions where the predictions disagree. The bottom table illustrates how each region contributes to the classification metrics.

2.5 CNA signatures by Drews et al.

Drews et al. (2022) used ASCAT to derive high-quality CN profiles from SNP array data across a large set of tumor samples with detectable CIN, spanning 33 cancer types. This dataset expanded their previously developed framework for identifying CNA signatures in ovarian cancer (Macintyre et al., 2018). Their CNA signatures discovery workflow included:

- PREPROCESSING: CN profiles were processed to extract values for five features from each CNA segment across samples. These features (segment size, change-point, lengths of chains of oscillating CN, breakpoints per 10MB, and breakpoints per chromosome arm) are known to capture distinct CNA patterns associated with CIN causes (see Figure 1).
- MIXTURE MODELING: Mixture modeling was applied to each feature's values, classifying variations within each feature's distribution into components representing core characteristics of CNA patterns associated with CIN. For each tumor sample, the probabilities of each CNA event belonging to each component were calculated, and then summed to produce a vector of posterior probabilities. These vectors were merged into a sample-by-component matrix.
- SIGNATURE IDENTIFICATION: Non-negative matrix factorization was applied to the full sample-by-component matrix, identifying 10 pan-cancer signatures, and to matrices from individual cancer types with at least 100 samples, leading to 7 additional cancer-type-specific signatures. The pan-cancer and cancer-type-specific signatures were then combined into a compendium of 17 CNA signatures. Following identification, potential causes for each of the signatures were explored (see Figure 1).
- ROBUSTNESS EVALUATION: The robustness of these signatures was assessed across various genomic technologies (SNP6, WGS, shallow WGS, and WES). Results indicated that they were stable, although WES showed limitations due to its lower resolution.

In this thesis, the `quantifyCNSignatures` function (with parameters `method = 'Drews'`, `build = 'hg38'`) and the `clinPredictionPlatinum` funtion from Drews et al.'s `CINSignatureQuantification` package were used to quantify the activity of the 17 CNA signatures based on CN profiles generated both by ASCAT (from WGS data) and scRNA-seq-based tools, and to predict platinum-based treatment response. A Pearson correlation was then computed for each patient to compare the consistency of signature activities across these tools.

Pattern	Seg. size	CNC	BP10	BPARM	OSC
Amplification	x	x	x		
Aneuploidy	x	x			
Breakage-fusion-bridge		x		x	
Chromothripsis		x	x	x	x
Complex genomic rearrangement	x	x	x	x	x
Deletions		x			
Extrachromosomal DNA (ecDNA)	x	x			
Homologous recombination deficiency (HRD)	x	x	x	x	x
Loss of heterozygosity (LOH)	x	x		x	
Micronuclei	x	x			
Tandem duplications	x	x		x	x
Whole-genome duplication (WGD)		x			

Figure 11: Copy number features and the CNA patterns they capture. The features are:

- Segment size (Seg. size): Length of CNA segments in base pairs.
- Changepoint (CNC): Absolute CN difference between a CNA segment and its left neighboring segment.
- Breakpoints per 10MB (BP10): Number of breaks within a 10MB sliding window across the genome.
- Breakpoints per chromosome arm (BPARM): Total number of breaks per chromosome arm.
- Lengths of chains of oscillating CN (OSC): Length of contiguous where CN oscillates between two states, often due to mutational processes like chromothripsis.

Source: Drews et al., Supplementary Methods (2022)

Signature	Pan-cancer prevalence (%)	Putative cause	Confidence	Pattern of change
CX1	80	Chromosome missegregation via defective mitosis and/or telomere dysfunction	★★★	Whole-arm/chromosome changes
CX2	69	IHR	★★★	Short-to-medium-sized, oscillating, single-copy changes
CX3	63	IHR with replication stress and impaired damage sensing	★★★	Long-sized, single-copy changes
CX4	58	PI3K–AKT-mediated toleration of whole-genome duplication	★★★	Medium-sized, clustered and oscillating, two-copy changes
CX5	56	IHR with replication stress	★★★	Medium-sized, two-to-three-copy changes
CX6	51	Chromosome missegregation via defective mitosis	★★★	Whole-arm/chromosome changes
CX7	50	Unknown	☆☆☆	Long-sized, less than one-copy change
CX8	38	Replication stress	★★★	Short-to-medium-sized, oscillating, low-level amplifications
CX9	32	Replication stress	★☆☆	Short-to-medium-sized, clustered, mid-level amplifications
CX10	29	Impaired NHEJ with replication stress	★★☆	Short-to-medium-sized, clustered and oscillating, two-copy changes
CX11	28	Replication stress	★★☆	Short-sized, clustered and oscillating, mid-level amplifications
CX12	27	Unknown	☆☆☆	Very short oscillating changes
CX13	26	Replication stress	★☆☆	Short-to-medium-sized, clustered, high-level amplifications
CX14	26	Chromosome missegregation via defective mitosis	★★☆	Whole-arm/chromosome, less than one-copy changes
CX15	25	Unknown	☆☆☆	Long-sized, less than one-copy changes
CX16	17	Unknown	★☆☆	Very short, clustered and oscillating, two-to-three-copy changes
CX17	16	Unknown	★★☆	Short-to-long-sized, one-to -three-copy changes

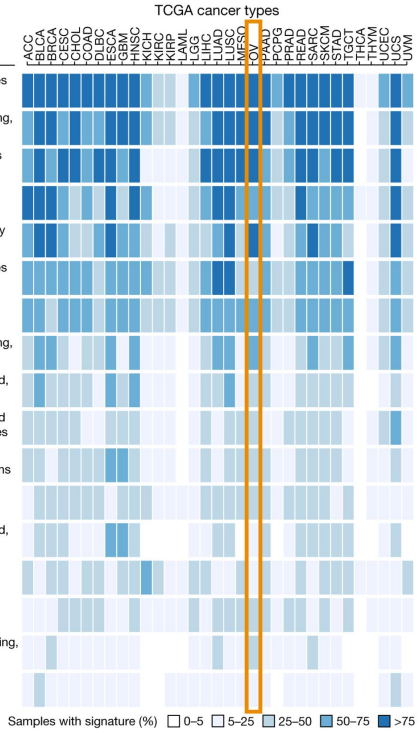


Figure 12: A summary of the pan-cancer frequency, proposed aetiology, aetiology confidence rating, pattern of copy number change, and distribution across cancer types for each signature. In HGSOC (highlighted in orange), the most frequent signatures are CX1, CX2, CX3, and CX5.
Source: Drews et al. (2022)

3. Results and discussion

3.1 Introduction to results

Table 2-8 (Appendix) provide patient-level information on classification and performance metrics, as well as the cumulative lengths of genomic regions classified as loss, neutral, or gain for each tool. Additionally, they include ploidy and purity values inferred by ASCAT, and cell counts.

Out of the initial 38 patients, 22 were excluded due to unresolved errors encountered in Numbat's execution, leaving 16 of them (listed in the rownames of Table 2 - Appendix). To ensure fair metric comparisons, these 22 patients were excluded across all tools, even though SCEVAN, InferCNV, and ASCAT generated valid results, potentially resulting in a loss of valuable data.

Visual representations, presented in the following sections, have been made to aid in the interpretation of this information.

3.2 Classification trends and visualization of CNAs across the genome

As observed in Figure 14, ASCAT and SCEVAN classify similar proportions of the genome as gain, loss, or neutral, suggesting a close agreement in their classification patterns. In contrast, InferCNV and Numbat classify a larger proportion of genome as neutral, suggesting a more conservative approach to detecting losses and gains.

The distribution along the genome of CNAs detected across each sample by ASCAT and scRNA-seq-based tools (Figure 13) can be compared with the CNA frequency data from Martins et al. (2022) and Punzón-Jiménez et al. (2022) in HGSOC.

Both studies identified frequent gains on chromosomes 3q and 8q, and losses on chromosomes 5q, 6q, 16, 17, 18q, 19q, and 22. These altered chromosomal regions contain oncogenes (e.g., PIK3CA and MECOM on 3q, and MYC on 8q) and tumor suppressor genes (e.g., NF1 on 17q). ASCAT detected all these alterations across most samples, confirming its validity as a ground truth for CNAs in HGSOC.

While scRNA-seq-based tools detected most of these frequently altered regions as well, they showed limitations specifically with losses on chromosome 19q, which ASCAT identified across many samples. This suggests that, while capturing the majority of critical CNAs, scRNA-seq-based tools may be less sensitive to some specific events.

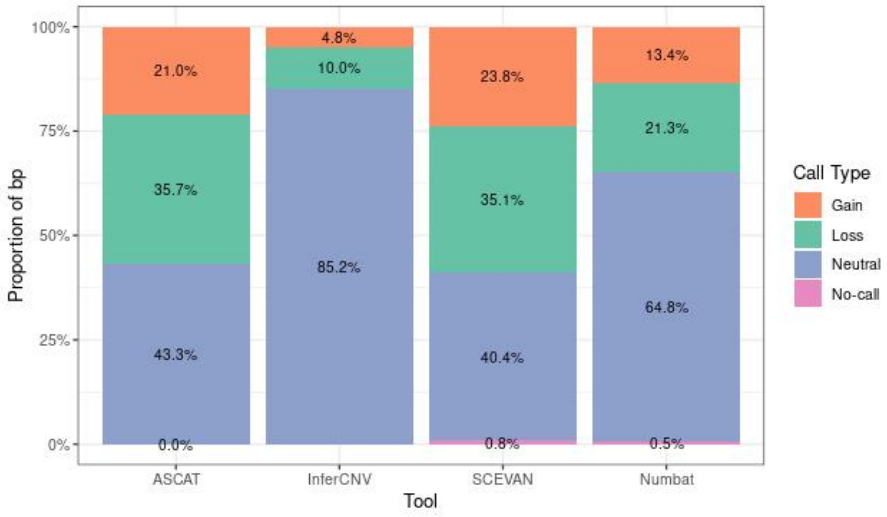


Figure 14: Proportions of genome classified as loss, gain, neutral, or no-call by each tool, calculated as the total base pairs for each call type summed across all samples.

Figure 13: Visualization of CNA segments across the genome for each sample, comparing ASCAT (top line for each sample) with InferCNV, SCEVAN, and Numbat. Colors represent different CN states (loss, neutral, gain, and no-call)

3.3 Classification metrics

Confusion matrices (Figure 13) show the number of bases classified as TP, FP, TN, and FN for loss, neutral, and gain calls using InferCNV, SCEVAN, and Numbat. Each confusion matrix shows the distribution of these classification metrics across the 16 samples.

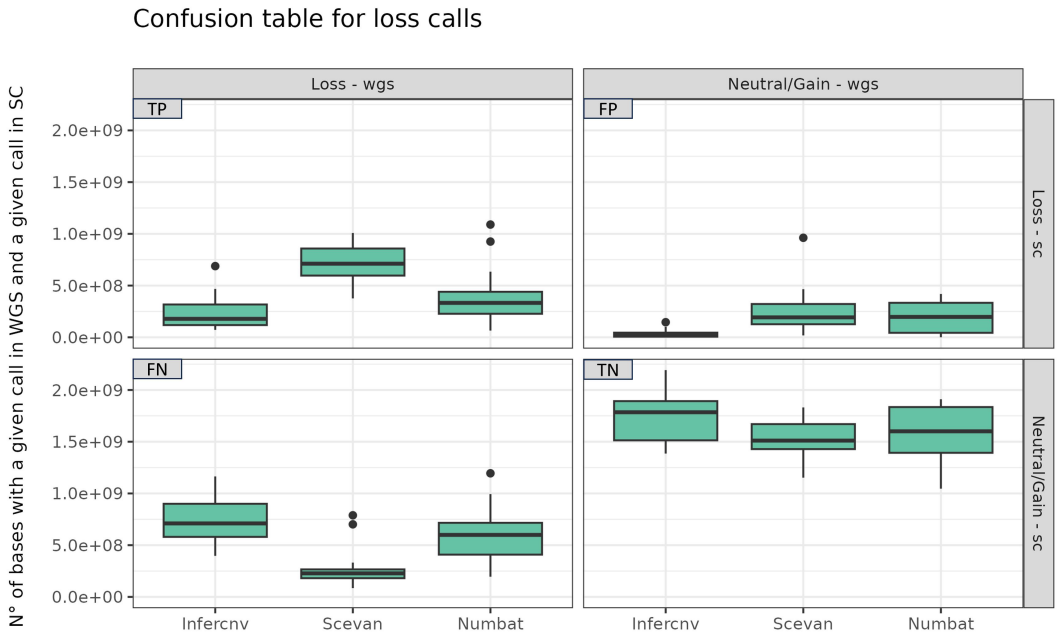
- Loss calls (Figure 13a): SCEVAN achieves the highest TP and lowest FN. InferCNV achieves less FP but has the lowest TP and highest FN.
- Neutral calls (Figure 13b): In terms of TP and FN, InferCNV outperforms, with the highest TP and fewest FN. SCEVAN outperforms in both FP and TN, with the lowest FP and highest TN. Numbat remains in the middle of both sets of metrics.
- Gain calls (Figure 13c): SCEVAN achieves the highest TP and lowest FN, indicating effective detection of gain regions, but also the highest FP and slightly lower TN compared to the other tools.

The classification metrics reveal distinct strengths and weaknesses across the tools.

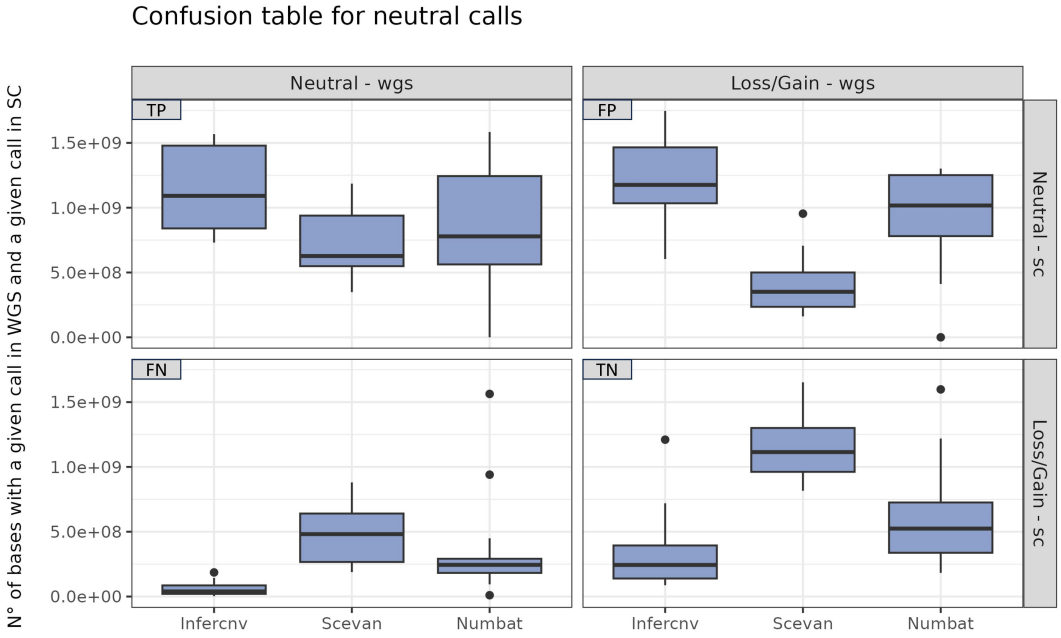
SCEVAN has the highest FP and lowest TN for loss and gain calls, but the lowest FP and the highest TN for neutral calls. This indicates that both InferCNV and Numbat tend to minimize FP at the expense of higher FN in loss and gain regions: they make few loss and gain calls at cost of missing some true alterations. This suggests that InferCNV and Numbat have a tendency to classify regions as neutral more often, aligning with the observations from Figure 14.

Numbat generally performs between SCEVAN and InferCNV without excelling in any single category.

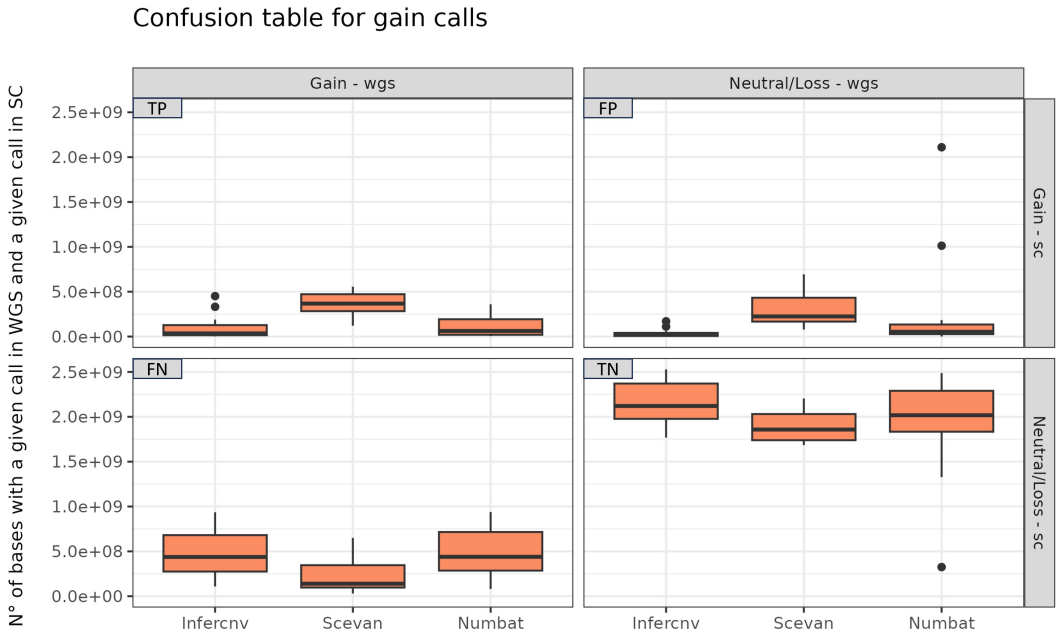
While classification metrics provide an overview, they rely on raw counts and may not fully reflect each tool's effectiveness. Detailed performance metrics, discussed in the next section, allow for a more immediate comparison.


Confusion table for loss calls
Loss - wgs
Neutral/Gain - wgs
TP
FP
FN
TN
Loss - sc
Neutral/Gain - sc
N° of bases with a given call in WGS and a given call in SC
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Infercnv
Scevan
Numbat

(a)


Confusion table for neutral calls
Neutral - wgs
Loss/Gain - wgs
TP
FP
FN
TN
Neutral - sc
Loss/Gain - sc
N° of bases with a given call in WGS and a given call in SC
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Infercnv
Scevan
Numbat

(b)



(c)

Figure 13: Each figure shows four panels illustrating TP, FP, TN, and FN for loss (a), neutral (b), and gain (c) calls, respectively. Each panel contains three boxplots (one for each of the scRNA-seq-based tools) that display the distribution of the number of bases with a specific call in ASCAT (WGS) and scRNA-seq-based tools (SC) across the 16 samples.

3.4 Performance metrics

Figure 16 and Figure 17 present performance metrics for InferCNV, SCEVAN, and Numbat.

Figure 16 shows overall, loss, neutral, and gain accuracy metrics. SCEVAN achieves the highest median overall accuracy (0.67), especially outperforming in neutral regions (median neutral accuracy = 0.68), above InferCNV (median neutral accuracy = 0.57) and Numbat (median neutral accuracy = 0.52).

Figure 17 shows sensitivity, specificity, PPV, and NPV metrics for each call type (loss, neutral, gain). SCEVAN consistently achieves median values above 0.50 across all metrics and call types. InferCNV and Numbat achieve lower performance, particularly with NPV values below 0.50 for loss and gain calls, and both PPV and specificity below 0.50 for neutral calls.

The performance metrics confirm insights from the classification metrics. The lower performance of InferCNV and Numbat reflects that many regions classified as neutral by these tools should not be neutral. In particular, the low NPV for loss and gain calls derives from high FN rates in these categories (i.e., many true loss

regions are incorrectly classified as neutral or gain, and many true gain regions are incorrectly classified as neutral or loss). Similarly, the low PPV and specificity for neutral calls are due to high FP for neutral regions (i.e., many regions classified as neutral are actually altered).

Given the prevalence of neutral regions (as visible in Figure 14 - Appendix), accuracy in neutral calls greatly influences overall accuracy. Therefore, even though InferCNV and Numbat achieve similar accuracy for loss and gain, SCEVAN's higher neutral accuracy has a positive impact on its median overall accuracy (0.67), making it the most accurate tool.

SCEVAN's higher overall accuracy reflects its greater ability to minimize errors across CNA types. Figure 21 (Appendix) shows that SCEVAN achieves the most balanced values across pairs of performance metrics for gain, loss, and neutral calls. In particular, Figure 18 highlights that, although SCEVAN exhibits a slightly lower median sensitivity for neutral calls, it compensates with a much higher specificity, achieving a superior balance between these two metrics.

It was tested whether performance, specifically overall accuracy, is influenced by factors such as tumor purity inferred by ASCAT from WGS data, the number of normal and malignant cells, the fraction of malignant cells, or coverage in scRNA-seq data (Figure 22, Figure 23 - Appendix). However, no correlation was observed.

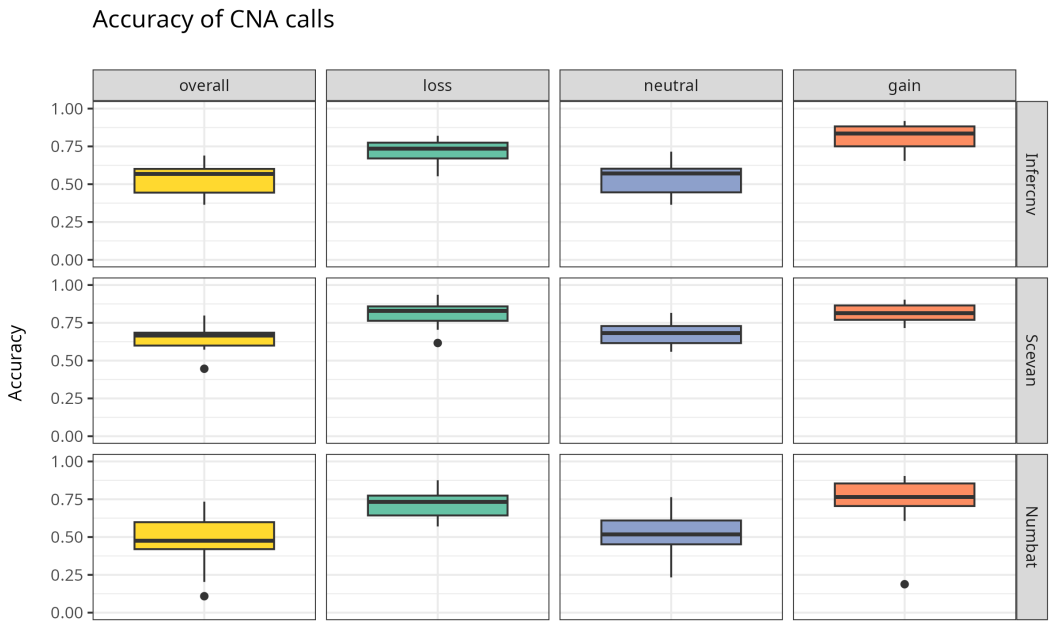


Figure 16: Accuracy metrics (overall, loss, neutral, gain) for each tool.

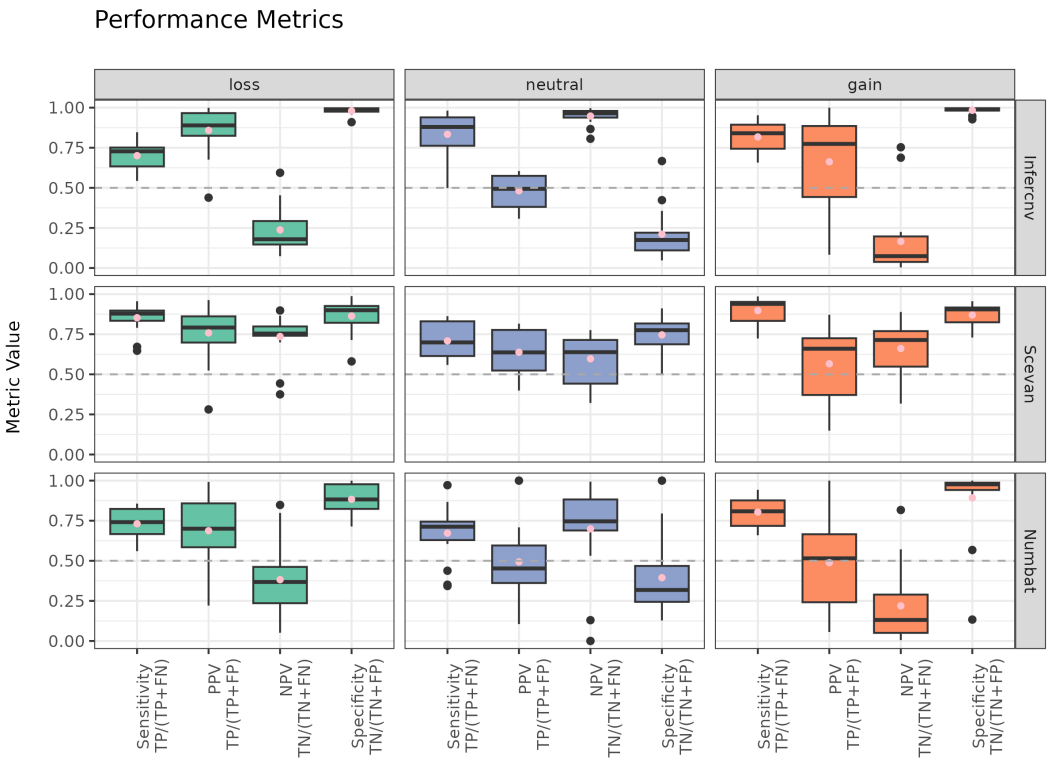


Figure 17: Performance metrics (sensitivity, specificity, PPV, and NPV) across each call type for each tool. The mean of each metric across all samples is represented by a pink dot. Ideally, all metrics should be close to 1, indicating high performance.

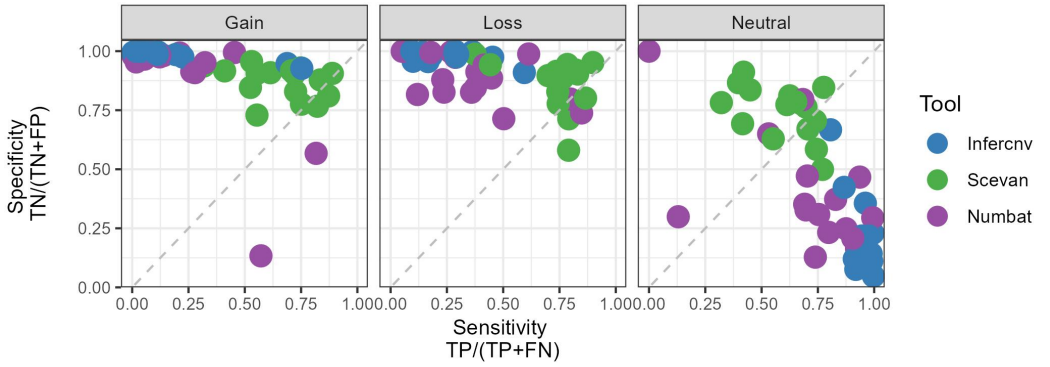


Figure 18: Scatter plot comparing sensitivity and specificity across CNA call types for each patient, colored by tool. Points near the diagonal indicate patients with balanced metrics, with those toward the top right along the diagonal reflecting optimal performance for both metrics.

3.5 Signatures

In Figure 19, CNA signatures were quantified across all samples starting from CNA segments detected by ASCAT and by scRNA-seq-based tools. Signature activity was not computed from scRNA-seq data in samples with fewer than 20 altered segments, affecting especially Numbat, which could only quantify signatures for 3 of 16 samples, limiting the comparison across methods.

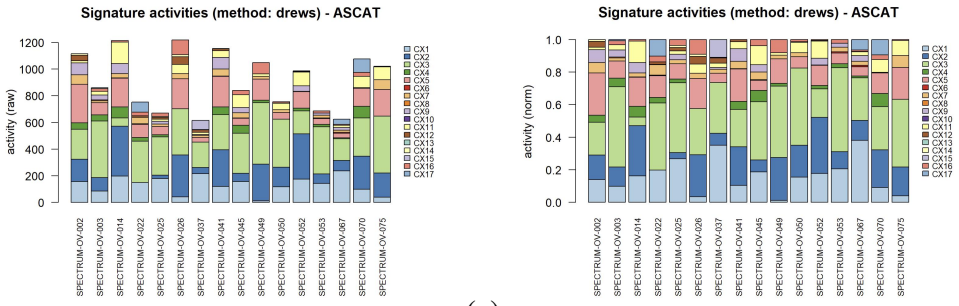
Using ASCAT, CX1, CX2, CX3, and CX5 were found active in the majority of samples. This supports the findings of Drews et al. (2022) that they were prevalent in HGSOC. However, while employing scRNA-seq-bases methods, CX5 was not found. This could be due to the nature of CX5, which has medium-sized, clustered, two-to-three-copy changes. In the benchmark analysis, all copy number changes were classified as gains or losses, but it would be helpful to compare the tools by distinguishing between types of gains and losses, as this could reveal whether the tools correctly identify gains or losses while misclassifying the actual number of copies changed. Also CX2 (short to medium-sized, clustered, single-copy changes) was difficult to detect, CX3 (long-sized, single-copy changes) was detected when using SCEVAN, and CX1 (whole arm or chromosome changes) was consistently detected when using all scRNA-seq tools.

The differences in WGS and scRNA-seq signature activities may be due to how each tool outputs copy number segments: ASCAT generates unrounded values, while scRNA-seq-based tools generate rounded values, making direct comparisons difficult. Furthermore, ASCAT does not correct for ploidy, while scRNA-seq tools algorithms integrate ploidy adjustments.

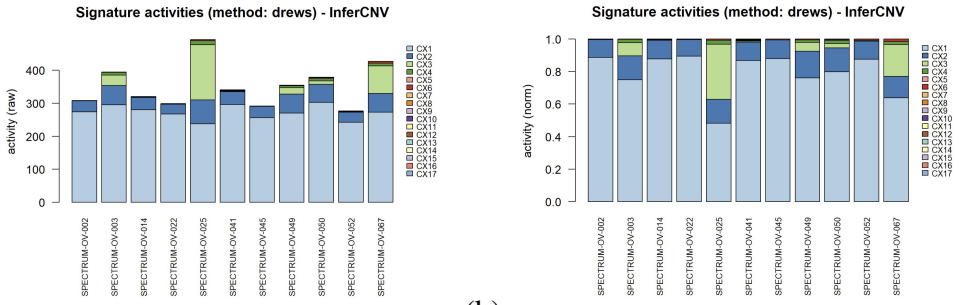
Pearson correlation analysis was used to assess the correlation between the signature activities quantified from WGS and scRNA-seq data. SCEVAN had the highest median correlation with ASCAT-derived signatures (about 0.45), although it is still quite low (Figure 20). This suggests that while SCEVAN captures some CIN characteristics, scRNA-seq–derived CNA profiles generally diverge from WGS-derived profiles, raising questions about whether the information captured is sufficient for clinical applications of these signatures.

Using the clinical classifier from Drews et al. (2022), no scRNA-seq tool identified platinum-sensitive patients as classified by ASCAT (Table 1). The classifier relies on higher CX3 activity over CX2 to predict sensitivity and high CX5 activity to predict resistance. Since CX2 activity is hard to detect in scRNA-seq data, the

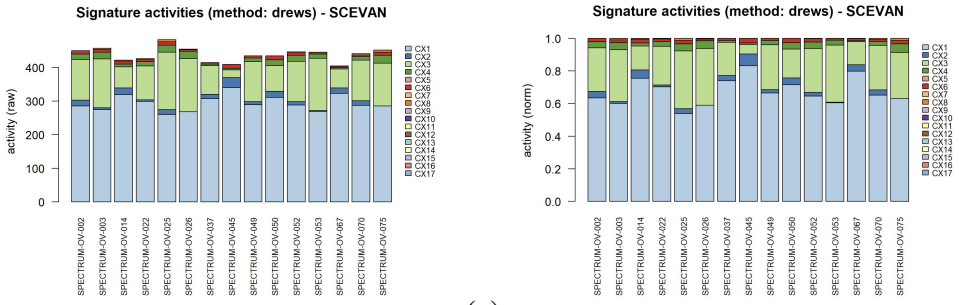
baseline is compromised, contributing to the classifier's limited performance with scRNA-seq data. These discrepancies between WGS and scRNA-seq signatures suggest a need for clinical classifiers specifically tailored to scRNA-seq data, which would require larger datasets to develop effectively.



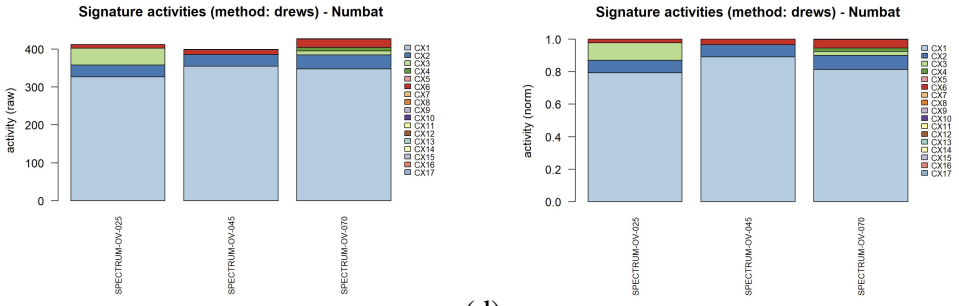
(a)



(b)



(c)



(d)

Figure 19: Raw (on left panel) and normalized (on right panel) signature activities computed from CNA segments inferred by ASCAT (a), considered the ground truth, InferCNV (b), SCEVAN (c) and Numbat (d).

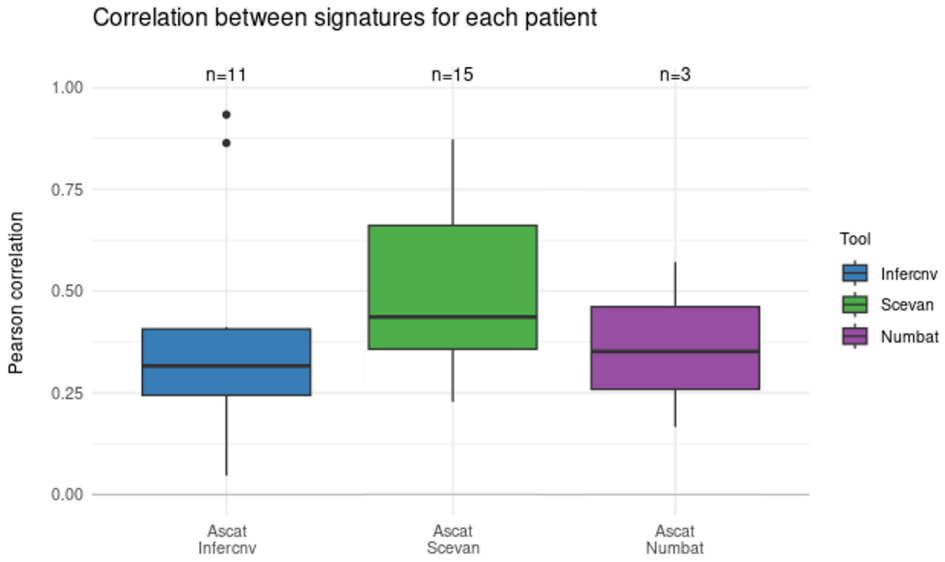


Figure 20: Boxplots showing the Pearson correlation between signature activities derived from CNA segments inferred by ASCAT and by scRNA-seq-based tools. Correlations are computed across all samples where both tools provided signature activity quantifications. The number of samples is indicated above each boxplot.

Patient	ASCAT	InferCNV	SCEVAN	Numbat
SPECTRUM-OV-002	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-003	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-014	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-022	Sensitive	Resistant	Resistant	<NA>
SPECTRUM-OV-025	Sensitive	Resistant	Resistant	Resistant
SPECTRUM-OV-026	Resistant	<NA>	Sensitive	<NA>
SPECTRUM-OV-037	Resistant	<NA>	Resistant	<NA>
SPECTRUM-OV-041	Resistant	Resistant	<NA>	<NA>
SPECTRUM-OV-045	Resistant	Resistant	Resistant	Resistant
SPECTRUM-OV-049	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-050	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-052	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-053	Sensitive	<NA>	Resistant	<NA>
SPECTRUM-OV-067	Resistant	Resistant	Resistant	<NA>
SPECTRUM-OV-070	Resistant	<NA>	Resistant	Resistant
SPECTRUM-OV-075	Resistant	<NA>	Resistant	<NA>

Table 1: Platinum-treatment resistance prediction. Green cells represent predictions concordant with the ASCAT ground truth, red cells represent discordant predictions, and gray cells represent cases where predictions are unavailable due to missing signature activity quantifications.

4. Conclusion and future perspectives

In this thesis, three tools (InferCNV, SCEVAN, and Numbat) for inferring CNA profiles from scRNA-seq data are benchmarked using CNAs inferred from WGS data as the ground truth, specifically within the context of HGSOC. It was shown that SCEVAN performed the best, even though all three tools can infer CNA profiles to some extent. The results also reveal a limitation in InferCNV and Numbat, which have a tendency to classify regions as neutral, thereby overlooking significant genomic gains and losses. SCEVAN is therefore the most reliable tool for studying HGSOC.

The observed differences between CNAs inferred from scRNA-seq and WGS data, such as in regions like chromosome 19q, highlight the need for algorithms refinement.

A larger and more diverse patient cohort, including additional cancer types and samples processed using different scRNA-seq platforms and protocols, could be analyzed to improve the reliability and broaden the applicability of this benchmarking. Furthermore, evaluating additional tools, such as CopyVAE (Kurt et al., 2024), would provide a more complete benchmarking.

Future research should explore the clinical potential of CNA signatures derived from scRNA-seq data, particularly for predicting responses to platinum-based therapies. The results demonstrate that current classifiers optimized for SNP6/WGS-derived signatures do not perform well on scRNA-seq-derived signatures, highlighting the need for classifiers specifically made for scRNA-seq data.

This thesis evaluated the ability of the tools to infer CNA profiles from pseudo-bulk RNA-seq produced from scRNA-seq data, however future studies are still needed to determine how accurately they can identify CNAs of individual tumor subclones.

In conclusion, this field of study has the potential to advance precision oncology. Tools that infer CNAs from scRNA-seq data provide detailed insights into the genomic and clonal composition of tumors, holding promise to predict treatment response and improve patient outcomes.

References

1. Cancer Today (IARC), 2022, https://gco.iarc.fr/today/en

2. Vázquez-García, I., Uhlitz, F., Ceglia, N. et al. Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* **612**, 778–786 (2022) https://doi.org/10.1038/s41586-022-05496-1

3. Forgó E., Longacre T.A. High grade serous carcinoma. PathologyOutlines.com website. https://www.pathologyoutlines.com/topic/ovarytumorserouscarcinomahg.html

4. Harbers L., Agostini F., Nicos M., Poddighe D., Bienko M., Crosetto N. Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas. *Front Oncol* **11** (2021) https://doi.org/10.3389/fonc.2021.700568

5. Harbers Martins, F. C., Couturier, D. L., de Santiago, I., Sauer, C. M. et al. Clonal somatic copy number altered driver events inform drug sensitivity in high-grade serous ovarian cancer. *Nat comm* **13**(1) (2022) https://doi.org/10.1038/s41467-022-33870-0

6. van Dijk, E., van den Bosch, T., Lenos, K.J. et al. Chromosomal copy number heterogeneity predicts survival rates across cancers. *Nat comm* **12** (2021) https://doi.org/10.1038/s41467-021-23384-6

7. Mallory, X.F., Edrisi, M., Navin, N. et al. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol* **208** (2020) https://doi.org/10.1186/s13059-020-02119-8

8. Tickle, T., Tirosh, I., Georgescu, C., Brown, M., Haas, B. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard (2019). https://github.com/broadinstitute/inferCNV

9. De Falco, A., Caruso, F., Su, XD. et al. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat Commun* **14**, 1074 (2023) https://doi.org/10.1038/s41467-023-36790-9

10. Kurt, S., Chen, M., Toosi, H., Chen, X., Engblom, C., Mold, J., Hartman, J., Lagergren, J. CopyVAE: a variational autoencoder-based approach for copy number variation inference using single-cell transcriptomics. *Bioinformatics* **40**(5), btae284 (2024) https://doi.org/10.1093/bioinformatics/btae284

11. Van Loo, P., Nordgard, S. H., Lingjærde, O. C. et al. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**(39), 16910–16915 (2010) https://doi.org/10.1073/pnas.1009843107

12. Ross, E. M., Haase, K., Van Loo, P. et al. Allele-specific multi-sample copy number segmentation in ASCAT. *Bioinformatics* **37**(13), 1909–1911 (2021) https://doi.org/10.1093/bioinformatics/btaa538

13. Van Loo, P., Nilsen, G., Nordgard, S.H. et al. Analyzing Cancer Samples with SNP Arrays. In: Wang, J., Tan, A., Tian, T. (eds) Next Generation Microarray Bioinformatics: Methods and Protocols. *Humana Press*. **802**, 57-72 (2012) https://doi.org/10.1007/978-1-61779-400-1_4

14. Van Loo, P. ASCAT Algorithm. *BioDiscovery, Inc.* (2013) [Video File]

15. Chen X., Fang L.T., Chen Z. et al. A benchmarking study of copy number variation inference methods using single-cell RNA-sequencing data. *bioRxiv* (2024) https://doi.org/10.1101/2024.09.09.612120

16. Minfang S., Shuai M., Zhenzhen Y. et al. Benchmarking copy number aberrations inference tools using single-cell multi-omics datasets. *bioRxiv* (2024) https://doi.org/10.1101/2024.09.26.615284

17. Adhikari L., Hassell LA. WHO classification. PathologyOutlines.com website. https://www.pathologyoutlines.com/topic/ovarytumorwhoclassif.html

18. Lynch, A., Bradford, S., Burkard, M. E. The reckoning of chromosomal instability: past, present, future. *Chromosome Res* **32**(2) (2024) https://doi.org/10.1007/s10577-024-09746-y

19. Kurman, R. J., Shih, I.-M. The Dualistic Model of Ovarian Carcinogenesis. *Am J Pathol* **186**(4), 733–747 (2016) https://doi.org/10.1016/j.ajpath.2015.11.011

20. Bowtell, D. D. et al. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat Rev Cancer* **15**, 668–679 (2015) https://doi.org/10.1038/nrc4019

21. Ovary cancer. Mortality DB. https://platform.who.int/mortality/themes/theme-details/topics/indicator-groups/indicator-group-details/MDB/ovary-cancer

22. Matulonis, U. A. et al. Ovarian cancer. *Nat Rev Dis Primers* **2**, 16061 (2016) https://doi.org/10.1038/nrdp.2016.61

23. Ovarian Cancer SPORE. Gynecologic Disease Laboratory. http://www.gynecologycancer.org/contact

24. Lisio, M.-A., Fu, L., Goyeneche, A., Gao, Z.-H., Telleria, C. High-Grade Serous Ovarian Cancer: Basic Sciences, Clinical and Therapeutic Standpoints. *Int J Mol Sci* **20**(4), 952 (2019) https://doi.org/10.3390/ijms20040952

25. Gao, T. et al. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol* **41**, 417–426 (2023) https://doi.org/10.1038/s41587-022-01468-y

26. Macintyre, G. et al. Copy-number signatures and mutational processes in ovarian carcinoma. *Nat Genet* **50**, 1262–1270 (2018) https://doi.org/10.1038/s41588-018-0179-8

27. Kim, J. et al. Cell Origins of High-Grade Serous Ovarian Cancer. *Cancers* **10**(11), 433 (2018) https://doi.org/10.3390/cancers10110433

28. Arigoni, M. et al. A single cell RNAseq benchmark experiment embedding "controlled" cancer heterogeneity. *Sci Data* **11**(159) (2024) https://doi.org/10.1038/s41597-024-03002-y

29. Drews, R.M., Hernando, B., Tarabichi, M. et al. A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022) https://doi.org/10.1038/s41586-022-04789-9

30. Hatano, Y. et al. A Comprehensive Review of Ovarian Serous Carcinoma. *Adv Anat Pathol* **26**(5), 329–339 (2019) https://doi.org/10.1097/PAP.0000000000000243

31. Zhang, A.W., O'Flanagan, C., Chavez, E.A. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* **16**, 1007–1015 (2019) https://doi.org/10.1038/s41592-019-0529-1

32. Mumford, D. B. & Shah, J. Optimal approximations by piecewise smooth functions and associated variational problems. *Commun. Pure Appl. Math.* **42**, 577–685 (1989) https://doi.org/10.1002/cpa.3160420503

33. Freeman, M.F., Tukey J.W. Transformations Related to the Angular and the Square Root *Ann. Math. Statist* **21**(4), 607-611 (1950) https://doi.org/10.1214/aoms/1177729756

34. Marshall J. Sequence Alignment/Map Optional Fields Specification (version 9 Sep 2024) https://github.com/samtools/hts-specs/blob/5a6f5e9ab18104b33110796604409782a612cb52/SAMtags.pdf

35. Nevins, J. R., Potti, A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature reviews. Genetics.* **8**(8), 601–609 (2007) https://doi.org/10.1038/nrg2137

36. Chibon F. Cancer gene expression signatures - the rise and fall? *European journal of cancer* **49**(8), 2000–2009 (2013) https://doi.org/10.1016/j.ejca.2013.02.021

37. Punzón-Jiménez, P. et al. Molecular Management of High-Grade Serous Ovarian Carcinoma. *Int J Mol Sci* **23**(22), 13777 (2022) https://doi.org/10.3390/ijms232213777

38. Kurt, S., Chen, M., Toosi, H. et al. CopyVAE: a variational autoencoder-based approach for copy number variation inference using single-cell transcriptomics. *Bioinformatics* **40**(5) (2024) https://doi.org/10.1093/bioinformatics/btae284

Appendix

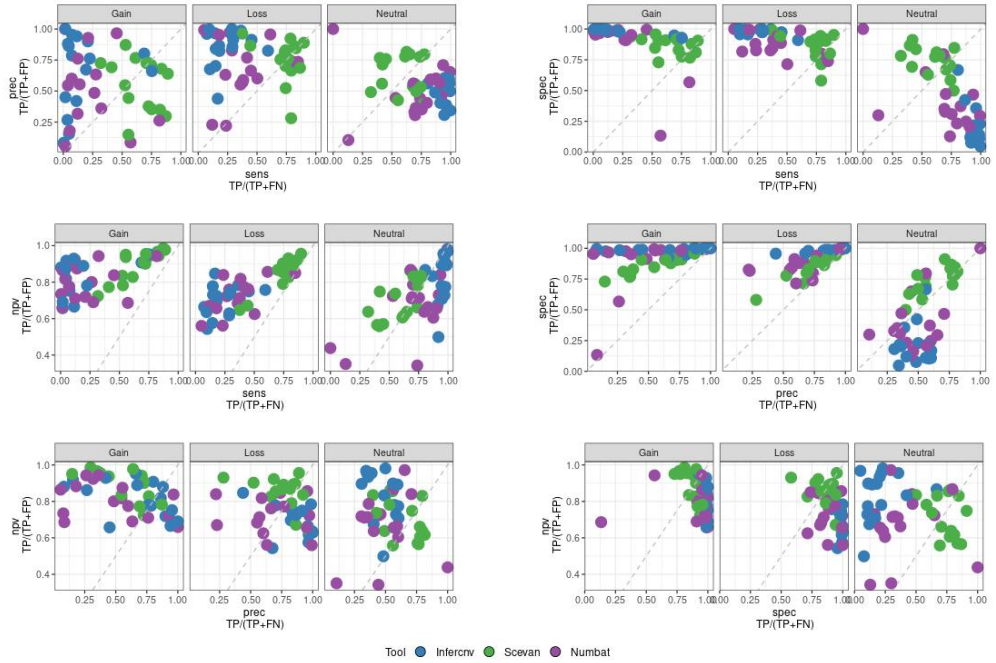


Figure 21: Scatter plots comparing pairs of performance metrics across CNA call types for each patient, colored by tool. Points near the diagonal indicate patients with balanced metrics, with those toward the top right along the diagonal reflecting optimal performance for both metrics.

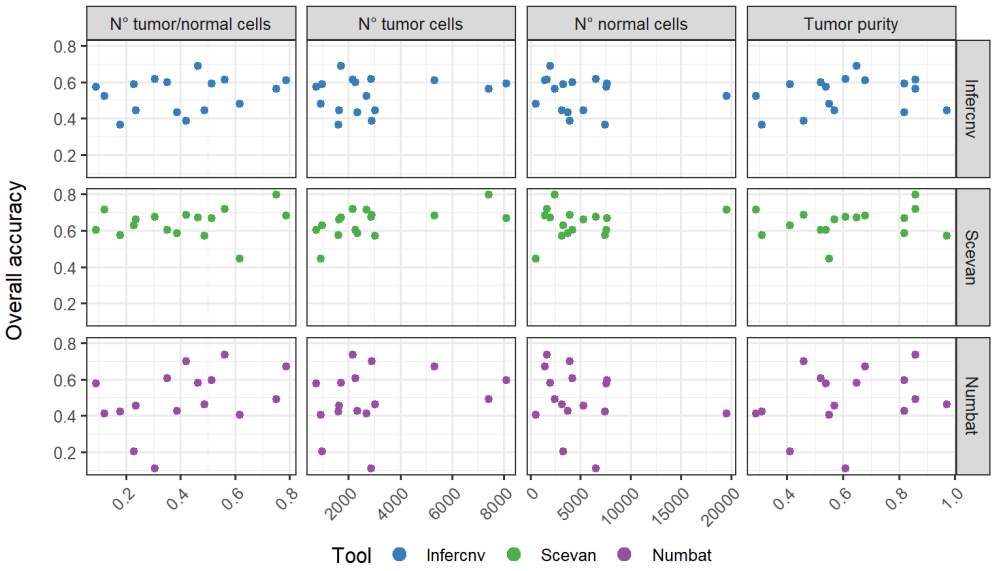


Figure 22: Scatter plots illustrating the relationship between overall accuracy in CNA detection and the tumor cell fraction, the number of tumor and normal cells in the scRNA-seq sample, and tumor purity inferred by ASCAT from WGS data, colored by tool.

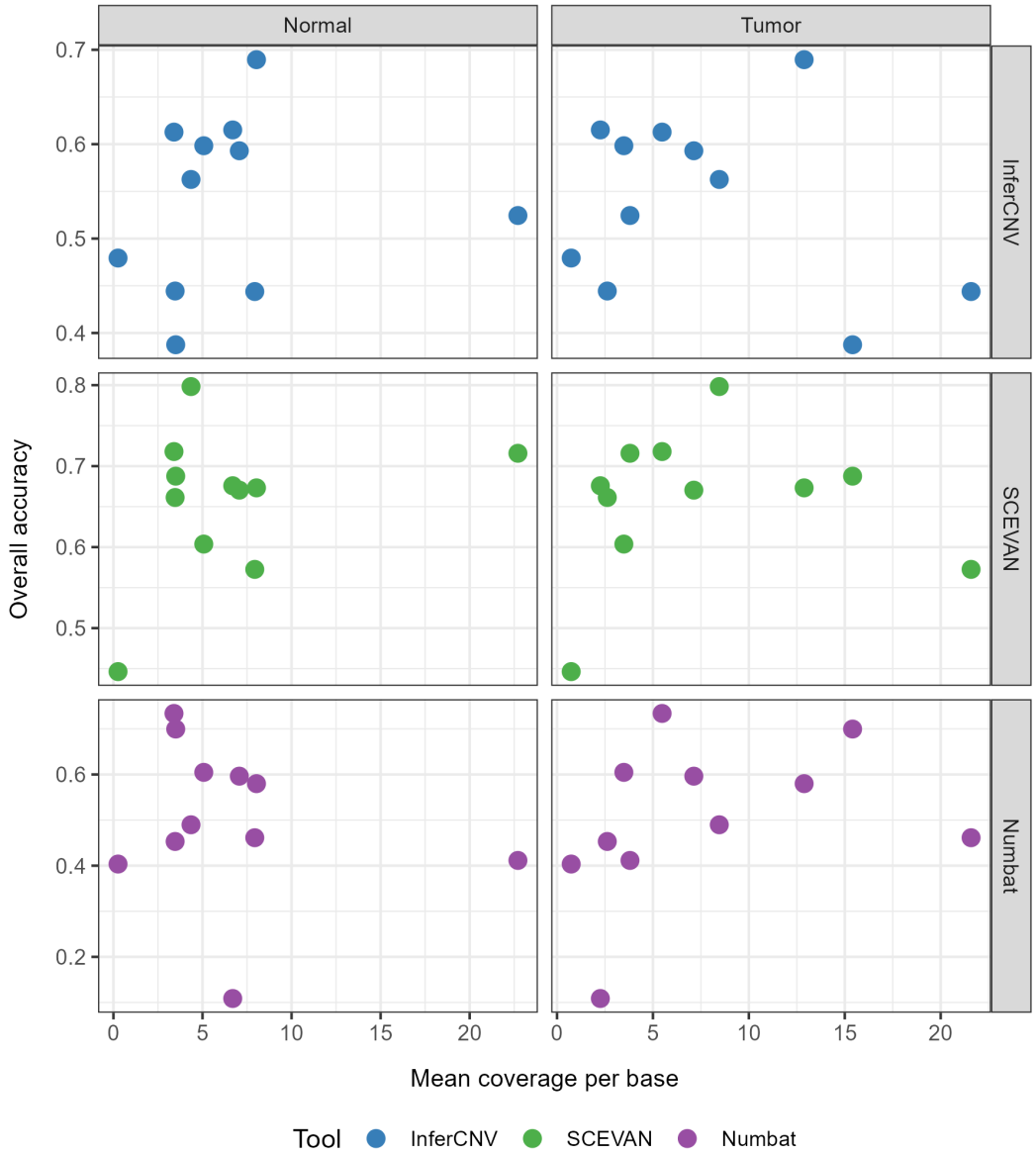


Figure 23: Scatter plots illustrating the relationship between overall accuracy in CNA detection and the mean coverage per base in pseudo-bulk samples derived from normal and tumor cells in each scRNA-seq sample, colored by tool.

patient	ploidy	purity
SPECTRUM-OV-002	3.31	0.55
SPECTRUM-OV-003	3.16	0.29
SPECTRUM-OV-014	3.06	0.86
SPECTRUM-OV-022	3.68	0.68
SPECTRUM-OV-025	3.02	0.65
SPECTRUM-OV-026	1.98	0.82
SPECTRUM-OV-037	1.83	0.86
SPECTRUM-OV-041	3.88	0.97
SPECTRUM-OV-045	2.92	0.46
SPECTRUM-OV-049	3.25	0.57
SPECTRUM-OV-050	1.83	0.52
SPECTRUM-OV-052	2.58	0.82
SPECTRUM-OV-053	3.1	0.31
SPECTRUM-OV-067	2.74	0.41
SPECTRUM-OV-070	3.67	0.61
SPECTRUM-OV-075	1.7	0.54

Table 2: Tumor ploidy and purity values estimated by ASCAT from WGS data.

patient	n_cells	n_mal	n_nor	frac_mal
SPECTRUM-OV-002	1512	933	579	0.6171
SPECTRUM-OV-003	22206	2690	19516	0.1211
SPECTRUM-OV-014	9880	7416	2464	0.7506
SPECTRUM-OV-022	6750	5313	1437	0.7871
SPECTRUM-OV-025	3674	1708	1966	0.4649
SPECTRUM-OV-026	15753	8096	7657	0.5139
SPECTRUM-OV-037	3853	2166	1687	0.5622
SPECTRUM-OV-041	6185	3019	3166	0.4881
SPECTRUM-OV-045	6852	2889	3963	0.4216
SPECTRUM-OV-049	6903	1637	5266	0.2371
SPECTRUM-OV-050	6452	2268	4184	0.3515
SPECTRUM-OV-052	6059	2349	3710	0.3877
SPECTRUM-OV-053	9051	1613	7438	0.1782
SPECTRUM-OV-067	4257	974	3283	0.2288
SPECTRUM-OV-070	9432	2877	6555	0.3050
SPECTRUM-OV-075	8322	746	7576	0.0896

Table 3: Total number of cells (n_cells), number of malignant (n_mal) and normal cells (n_nor), and fraction of malignant cells (frac_mal) in scRNA-seq data.

patient	cl_loss_A	cl_neu_A	cl_gain_A	cl_loss_I	cl_neu_I	cl_gain_I	cl_loss_S	cl_neu_S	cl_gain_S	cl_loss_N	cl_neu_N	cl_gain_N
SPECTRUM-OV-002	475666307	1340553739	952775125	179984954	2550786811	38223406	1337929957	995936704	417141482	510808694	2235838107	14526387
SPECTRUM-OV-003	1031019714	823153936	905433830	512063741	2037675135	209868604	983252093	1031997079	724958221	670917250	1875520308	205915634
SPECTRUM-OV-014	832288512	1200962249	723368518	306746786	2392815120	57057373	840415162	1170543400	716138200	289285324	2408553076	48273469
SPECTRUM-OV-022	953773176	1594649243	234585462	209218538	2500027509	73761834	858451270	1221225354	687358495	262340348	2422553954	90046756
SPECTRUM-OV-025	1158900816	959352303	655374251	834145101	1377395042	562087227	1379299928	564163819	805372272	1252992676	1145203601	357111946
SPECTRUM-OV-026	687076673	1568737371	522411700	144446438	2605578118	28201188	976413121	1234001185	551068197	446761143	2282615940	37552262
SPECTRUM-OV-037	827964134	1604242506	329803358	147069320	2600728063	14212615	802310041	1527168302	411823754	531024543	2118670476	99778445
SPECTRUM-OV-041	1262216803	821522647	643621185	373277113	2294292032	59791490	491689412	1586928508	628058818	1053329644	1585026818	76820216
SPECTRUM-OV-045	1284741638	772664379	725133231	191252618	2377224528	214062102	1092769468	973285771	696603487	1481430710	939472869	340986271
SPECTRUM-OV-049	895098172	880062364	1004305855	291479172	2364900194	123087025	899163333	1246954344	610931089	483312688	1839733553	440636730
SPECTRUM-OV-050	930891575	1461115345	390770499	316597284	2414894228	51285907	1174044939	756910872	830514963	627876108	2040098420	98131090
SPECTRUM-OV-052	1293445860	1084980589	362245036	190927043	2445796120	103948322	1160826748	709224101	852226841	474298823	2136486765	114886641
SPECTRUM-OV-053	971282219	917907840	862799634	74738548	2661822367	15428778	1027200191	1186906929	511710088	641049257	1960096086	139530384
SPECTRUM-OV-067	1259634023	1080548877	440546884	361635872	1917264260	501829652	645403747	1509832340	612191252	65367111	1332355511	1372457772
SPECTRUM-OV-070	871769949	1563306533	347059540	1795082238	2587302369	15325415	807066569	1279564437	667856666	454372385	107828	2308646301
SPECTRUM-OV-075	1077995415	1484158677	218567985	101654744	2620872873	58194460	1060192903	879246057	814555032	202090241	2369544062	195831048

Table 4: Cumulative lengths of regions for each type of CNA (loss, neu = neutral, gain) for each tool (A = ASCAT, I = InferCNV, S = SCEVAN, N = Numbat).

patient	sen_loss_I	spe_loss_I	ppv_loss_I	npv_loss_I	sen_loss_S	spe_loss_S	ppv_loss_S	npv_loss_S	sen_loss_N	spe_loss_N	ppv_loss_N	npv_loss_N
SPECTRUM-OV-002	0.1660	0.9559	0.4388	0.8467	0.7905	0.5805	0.2810	0.9304	0.2371	0.8264	0.2208	0.8393
SPECTRUM-OV-003	0.4533	0.9741	0.9128	0.7492	0.8204	0.9205	0.8602	0.8957	0.3593	0.8262	0.5522	0.6838
SPECTRUM-OV-014	0.3634	0.9978	0.9862	0.7837	0.8980	0.9516	0.8893	0.9557	0.2874	0.9740	0.8268	0.7596
SPECTRUM-OV-022	0.1805	0.9797	0.8232	0.6963	0.7358	0.9143	0.8175	0.8690	0.2673	0.9960	0.9719	0.7228
SPECTRUM-OV-025	0.5938	0.9096	0.8250	0.7572	0.7912	0.7136	0.6648	0.8264	0.7985	0.7971	0.7386	0.8465
SPECTRUM-OV-026	0.1475	0.9793	0.7016	0.7776	0.7441	0.7775	0.5236	0.9024	0.3830	0.9122	0.5891	0.8182
SPECTRUM-OV-037	0.1493	0.9878	0.8405	0.7306	0.7423	0.9029	0.7660	0.8911	0.6132	0.9879	0.9561	0.8564
SPECTRUM-OV-041	0.2862	0.9918	0.9678	0.6172	0.3750	0.9875	0.9627	0.6471	0.5026	0.7141	0.6023	0.6250
SPECTRUM-OV-045	0.1437	0.9955	0.9654	0.5754	0.7848	0.9436	0.9227	0.8364	0.8482	0.7385	0.7356	0.8501
SPECTRUM-OV-049	0.2980	0.9868	0.9151	0.7474	0.7644	0.8859	0.7609	0.8878	0.4153	0.9408	0.7692	0.7721
SPECTRUM-OV-050	0.2912	0.9754	0.8562	0.7324	0.8658	0.8012	0.6865	0.9223	0.4484	0.8863	0.6648	0.7617
SPECTRUM-OV-052	0.0997	0.9572	0.6758	0.5433	0.7436	0.8625	0.8286	0.7901	0.2314	0.8791	0.6311	0.5614
SPECTRUM-OV-053	0.0739	0.9983	0.9607	0.6640	0.7423	0.8280	0.7019	0.8548	0.3766	0.8454	0.5706	0.7132
SPECTRUM-OV-067	0.2869	0.9998	0.9993	0.6286	0.4430	0.9426	0.8647	0.6714	0.0515	0.9997	0.9919	0.5600
SPECTRUM-OV-070	0.8805	0.8794	0.9865	0.1426	0.6991	0.8965	0.7551	0.8672	0.1192	0.8166	0.2288	0.6701
SPECTRUM-OV-075	0.0940	0.9998	0.9967	0.63544	0.8305	0.9031	0.8445	0.8938	0.1785	0.9943	0.9520	0.6566

Table 5: Performance metrics (spe = specificity, ppv = positive predictive value, npv = negative predictive value, sen = sensitivity) for losses for each scRNA-seq based tool (I = InferCNV, S = SCEVAN, N = Numbat)

patient	sen_neu_I	spe_neu_I	ppv_neu_I	npv_neu_I	sen_neu_S	spe_neu_S	ppv_neu_S	npv_neu_S	sen_neu_N	spe_neu_N	ppv_neu_N	npv_neu_N
SPECTRUM-OV-002	0.9184	0.0761	0.4826	0.4988	0.4156	0.6928	0.5594	0.5581	0.0152	1.0000	1.0000	0.6594
SPECTRUM-OV-003	0.9607	0.3561	0.3881	0.9552	0.6970	0.7633	0.5559	0.8556	0.2111	0.9920	0.9284	0.7203
SPECTRUM-OV-014	0.9942	0.2294	0.4990	0.9810	0.7758	0.8464	0.7959	0.8302	0.0052	0.9781	0.0777	0.7343
SPECTRUM-OV-022	0.9482	0.1686	0.6048	0.7081	0.6247	0.8107	0.8157	0.6168	0.1219	0.9759	0.3175	0.9235
SPECTRUM-OV-025	0.8061	0.6670	0.5614	0.8667	0.4198	0.9110	0.7140	0.7481	0.2644	0.9132	0.4852	0.8005
SPECTRUM-OV-026	0.9701	0.1040	0.5840	0.7286	0.6122	0.7738	0.7782	0.6060	0.0392	0.9924	0.5457	0.8169
SPECTRUM-OV-037	0.9772	0.1077	0.6028	0.7737	0.7390	0.7049	0.7763	0.6609	0.0544	0.9664	0.1799	0.8829
SPECTRUM-OV-041	0.9768	0.2172	0.3497	0.9560	0.7703	0.4994	0.3988	0.8345	0.0716	0.9853	0.6002	0.7746
SPECTRUM-OV-045	0.9455	0.1807	0.3073	0.8962	0.6785	0.7765	0.5386	0.8627	0.4540	0.9943	0.9654	0.8378
SPECTRUM-OV-049	0.9689	0.2038	0.3605	0.9341	0.7059	0.6705	0.4982	0.8311	0.2780	0.9091	0.6337	0.6900
SPECTRUM-OV-050	0.9446	0.2172	0.5715	0.7803	0.3972	0.8664	0.7668	0.5652	0.1392	0.9817	0.5544	0.8747
SPECTRUM-OV-052	0.9116	0.1201	0.4044	0.6748	0.3212	0.7821	0.4914	0.6374	0.0177	0.9544	0.0557	0.8645
SPECTRUM-OV-053	0.9968	0.0475	0.3437	0.9674	0.5516	0.6289	0.4266	0.7370	0.1231	0.9824	0.7612	0.7104
SPECTRUM-OV-067	0.8668	0.4232	0.4885	0.8333	0.7433	0.5843	0.5319	0.7817	0.8169	0.5673	0.2622	0.9427
SPECTRUM-OV-070	0.9865	0.1426	0.5961	0.8924	0.6517	0.7860	0.7962	0.6376	0.5716	0.1334	0.0859	0.6860
SPECTRUM-OV-075	0.9890	0.1107	0.5601	0.8982	0.4491	0.8360	0.7581	0.5700	0.3231	0.9511	0.3606	0.9428

Table 6: Performance metrics (spe = specificity, ppv = positive predictive value, npv = negative predictive value, sen = sensitivity) for neutrals for each scRNA-seq based tool (I = InferCNV, S = SCEVAN, N = Numbat)

patient	sen_gain_I	spe_gain_I	ppv_gain_I	npv_gain_I	sen_gain_S	spe_gain_S	ppv_gain_S	npv_gain_S	sen_gain_N	spe_gain_N	ppv_gain_N	npv_gain_N
SPECTRUM-OV-002	0.0181	0.9884	0.4505	0.6574	0.3175	0.9369	0.7253	0.7235	0.7386	0.1279	0.4429	0.3428
SPECTRUM-OV-003	0.2087	0.9887	0.9004	0.7190	0.6138	0.9087	0.7667	0.8281	0.6972	0.3278	0.3060	0.7181
SPECTRUM-OV-014	0.0751	0.9986	0.9529	0.7521	0.7207	0.9041	0.7279	0.9009	0.9218	0.1633	0.4596	0.7300
SPECTRUM-OV-022	0.0488	0.9755	0.1555	0.9176	0.8722	0.8105	0.2976	0.9856	0.9936	0.2947	0.6540	0.9717
SPECTRUM-OV-025	0.6879	0.9475	0.8021	0.9075	0.8351	0.8781	0.6795	0.9450	0.5313	0.6497	0.4451	0.7239
SPECTRUM-OV-026	0.0465	0.9982	0.8627	0.8188	0.7477	0.9288	0.7089	0.9408	0.8752	0.2479	0.6015	0.6049
SPECTRUM-OV-037	0.0035	0.9946	0.0828	0.8804	0.5546	0.9058	0.4442	0.9375	0.9358	0.4667	0.7086	0.8400
SPECTRUM-OV-041	0.0729	0.9938	0.7857	0.7763	0.7072	0.9170	0.7247	0.9102	0.7031	0.4714	0.3644	0.7865
SPECTRUM-OV-045	0.2249	0.9752	0.7621	0.7811	0.5250	0.8464	0.5465	0.8349	0.6830	0.7952	0.5618	0.8671
SPECTRUM-OV-049	0.1153	0.9959	0.9410	0.6655	0.5302	0.9558	0.8716	0.7824	0.6913	0.3517	0.3307	0.7109
SPECTRUM-OV-050	0.0351	0.9842	0.2676	0.8619	0.7516	0.7755	0.3536	0.9502	0.8287	0.3726	0.5935	0.6630
SPECTRUM-OV-052	0.1931	0.9857	0.6730	0.8891	0.8222	0.7669	0.3495	0.9659	0.7980	0.2326	0.4053	0.6373
SPECTRUM-OV-053	0.0178	1.0000	1.0000	0.6903	0.4102	0.9165	0.6917	0.7728	0.7536	0.3084	0.3529	0.7143
SPECTRUM-OV-067	0.7530	0.9273	0.6611	0.9522	0.8887	0.9057	0.6395	0.9774	0.1299	0.2989	0.1053	0.3508
SPECTRUM-OV-070	0.5961	0.8924	0.7338	0.8794	0.7256	0.8291	0.3770	0.954	0.0001	1.0000	1.0000	0.4381
SPECTRUM-OV-075	0.1118	0.9868	0.4199	0.9287	0.5540	0.7293	0.1486	0.9504	0.9048	0.2081	0.5667	0.6563

Table 7: Performance metrics (spe = specificity, ppv = positive predictive value, npv = negative predictive value, sen = sensitivity) for gains for each scRNA-seq based tool (I = InferCNV, S = SCEVAN, N = Numbat)

patient	acc_ov_I	acc_loss_I	acc_neu_I	acc_gain_I	acc_ov_S	acc_loss_S	acc_neu_S	acc_gain_S	acc_ov_N	acc_loss_N	acc_neu_N	acc_gain_N
SPECTRUM-OV-002	0.4794	0.8203	0.4839	0.6545	0.4463	0.6166	0.5586	0.7238	0.4036	0.7252	0.4236	0.6612
SPECTRUM-OV-003	0.5244	0.7796	0.5365	0.7328	0.7159	0.8831	0.7436	0.8120	0.4115	0.6518	0.4380	0.7358
SPECTRUM-OV-014	0.5626	0.8063	0.5626	0.7563	0.7983	0.9355	0.8157	0.8560	0.4897	0.7667	0.4938	0.7228
SPECTRUM-OV-022	0.6093	0.7059	0.6153	0.8974	0.6837	0.8532	0.7042	0.8158	0.6712	0.7463	0.6952	0.9039
SPECTRUM-OV-025	0.6895	0.7777	0.7152	0.8862	0.6732	0.7461	0.7412	0.8680	0.5799	0.7977	0.6088	0.7599
SPECTRUM-OV-026	0.5930	0.7737	0.5931	0.8193	0.6703	0.7693	0.6826	0.8948	0.5963	0.7813	0.6021	0.8132
SPECTRUM-OV-037	0.6128	0.7365	0.6128	0.8763	0.7180	0.8548	0.7247	0.8640	0.7339	0.8756	0.7392	0.8575
SPECTRUM-OV-041	0.4439	0.6653	0.4460	0.7765	0.5725	0.7041	0.5810	0.8675	0.4613	0.6162	0.5412	0.7697
SPECTRUM-OV-045	0.3876	0.6023	0.3931	0.7797	0.6876	0.8703	0.7494	0.7627	0.6996	0.7891	0.7640	0.8535
SPECTRUM-OV-049	0.4445	0.7650	0.4461	0.6777	0.6613	0.8468	0.6818	0.8020	0.4531	0.7716	0.4592	0.6811
SPECTRUM-OV-050	0.5984	0.7465	0.5992	0.8510	0.6038	0.8229	0.6201	0.7722	0.6047	0.7398	0.6121	0.8634
SPECTRUM-OV-052	0.4335	0.5526	0.4335	0.8810	0.5868	0.8065	0.5997	0.7742	0.4275	0.5734	0.4564	0.8306
SPECTRUM-OV-053	0.3642	0.6721	0.3642	0.6921	0.5746	0.7978	0.6032	0.7578	0.4229	0.6800	0.4569	0.7130
SPECTRUM-OV-067	0.5861	0.6769	0.5956	0.8997	0.6304	0.7163	0.6462	0.9030	0.2032	0.5701	0.2332	0.6068
SPECTRUM-OV-070	0.6151	0.7339	0.6169	0.8794	0.6758	0.8347	0.7106	0.8162	0.1087	0.5981	0.4381	0.1881
SPECTRUM-OV-075	0.5731	0.6486	0.5795	0.9180	0.6053	0.8750	0.6295	0.7156	0.5775	0.6780	0.5800	0.9018

Table 8: Accuracy (acc_ov = overall accuracy, acc_loss = accuracy for losses, acc_neu = accuracy for neutrals, acc_gain = accuracy for gains) for each scRNA-seq based tool (I = InferCNV, S = SCEVAN, N = Numbat)

Acknowledgements

There are many people I would like to thank for helping me during these 5 years of university. Without your support, this thesis would not have been possible. Given my love for bullet points, here is a list of those I would like to acknowledge:

- My supervisor, Chiara Romualdi, first of all for inspiring my interest in computational biology right from her very first lecture on omics data, and, of course, for guiding me through the entire traineeship while making me feel valued and part of the team.
- My lab colleagues for the never-boring lunch breaks and their help with my research, with a special mention to Nicolò.
- All the professors who have shaped my education over the years, starting from middle school, with particular thanks to the most recent ones: Vezzi, for brightening the days at Vallisneri in the most unexpected ways, as well as Calura, Pagani, and Sales.
- Ermanno, Martina, and everyone in the Lavezzo-Toppo lab, for hosting me last summer, a crucial experience for developing my practical bioinformatics skills and reaffirming my passion for research.
- My university mates (Agus, Anja, Claudio, Cloey, Edo, Emam, Fra, Kate, Khaled, LordB, Mattia, Paòlo, Riky, Yasmine) for sharing notes, facing challenges together, creating endless memories (from aperitivi and outings to maté cosido sessions), and for the close friendships that have formed with some of you.
- My friends outside of university (it would take half a page to name you all, so please forgive me if I don't, but I must mention Chiara and Laura as I don't want them to kill me ♡) for always being there, encouraging and supporting me. This is not to be taken for granted.
- My volleyball teams for being a constant source of energy and a much-needed break from my studies.

- My parents for valuing and financially supporting my education, and for standing by me even when I wasn't the easiest to deal with.
- My sisters, Giorgia and Silvia, for supporting me during my whole life in countless ways, sometimes without even realizing it. Giorgia, your encouragement to explore teaching has been crucial in keeping me motivated and managing the stress of future uncertainties and financial needs over the past three years, helping me avoid burnout. Silvia, introducing me to Scratch as a child might have seemed small, but it was a seed that grew into my passion for computational thinking.
- My extended family, especially Nonna Lucia for our endless chats about my studies and for always telling me, "diventerai una scienziata." This thesis is a first step towards that dream.

Elena