CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Copy Number Profiling of ctDNA via Shallow WGS in NSCLC

Evaluation of an Integrated Workflow for the Detection of Somatic Alterations

Master's Thesis in Biotechnology

JOSEFIN MILLER

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2026
www.chalmers.se

MASTER'S THESIS 2026

Copy Number Profiling of ctDNA via Shallow WGS in NSCLC

Evaluation of an Integrated Workflow for the Detection of Somatic Alterations

JOSEFIN MILLER

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Copy Number Profiling of ctDNA via Shallow WGS in NSCLC
Evaluation of an Integrated Workflow for the Detection of Somatic Alterations

JOSEFIN MILLER

Supervisor: Anna Rohlin, Department of Clinical Genetics and Genomics,
Sahlgrenska University Hospital

Examiner: Erik Kristiansson, Division of Applied Mathematics and Statistics,
Departments of Mathematical Sciences, Chalmers University of Technology

Master's Thesis 2026
Department of Mathematical Sciences
Division of *Applied Mathematics and Statistics*
Chalmers University of Technology
SE-412 96 Göteborg
Telephone: + 46 (0)31-772 1000

Printed by Chalmers Reproservice
Göteborg, Sweden 2026

Copy Number Profiling of ctDNA via Shallow WGS in NSCLC
Evaluation of an Integrated Workflow for the Detection of Somatic Alterations
JOSEFIN MILLER
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Non-small cell lung cancer (NSCLC) remains a leading cause of cancer-related mortality worldwide, necessitating robust and minimally invasive methods for longitudinal monitoring of therapeutic response and tumor evolution. Liquid biopsies utilizing cell-free DNA (cfDNA) offer a promising approach, however, capturing somatic copy number alterations (CNAs) from circulating tumor DNA (ctDNA) is highly challenging due to low tumor fractions and signal dilution from healthy background cfDNA. This thesis evaluates and optimizes an integrated workflow for copy number profiling via shallow whole genome sequencing (sWGS) at ultra-low sequencing depths (~0.0044x to 0.01x). The bioinformatic pipelines Ion Reporter, ichorCNA, and WisecondorX were systematically benchmarked using healthy controls, an undiluted A549 cell line positive control, and clinical plasma samples. To increase sensitivity for low-fraction ctDNA signals, a bisection strategy was used to adjust the hidden Markov model (HMM) parameters of an existing clinical Ion Reporter workflow. All three pipelines were evaluated across different bin sizes (0.5 Mb, 1.0 Mb, and 2.0 Mb), with ichorCNA and WisecondorX processed under default configurations. The benchmarking framework revealed distinct architectural trade-offs across the pipelines at this ultra-low depth. While the optimized Ion Reporter workflow acted conservatively to minimize false-positive calls in validation controls, WisecondorX achieved the highest overall call accuracy (Macro F1 = 0.548) on the positive control data, and ichorCNA provided the highest sensitivity in clinical samples at the expense of vulnerability to background technical noise. Ultimately, this study demonstrates that at this ultra-low sequencing depth, pipeline selection should align with the clinical objective: a conservative architecture like Ion Reporter is well-suited to reduce false positives, whereas tools like WisecondorX or ichorCNA are better suited when maximizing the detection of low-amplitude variants is the priority.

Keywords: ctDNA, copy number alterations, sWGS, NSCLC, biomarker, Ion Reporter, ichorCNA, WisecondorX

Acknowledgements

First, I would like to thank my supervisor Anna Rohlin for giving me the opportunity to work with this project, and for believing in my ideas. I would also like to thank Johanna Svensson for supporting me throughout the project, and everyone at Clinical Genetics and Genomics at Sahlgrenska University Hospital for welcoming me to the department. I am also deeply grateful to Maria Yhr for her exceptional guidance during the practical laboratory work, teaching me to perform library preparation and next-generation sequencing procedures. Lastly, I would like to thank all the patients of the BioLung study, whose participation helps drive scientific research forward.

Josefin Miller, Gothenburg, June 2026

List of Acronyms

Below are relevant acronyms used throughout this thesis in alphabetical order.

CBS	Circular Binary Segmentation
cfDNA	Cell-Free DNA
CN	Copy Number
CNA	Copy Number Alteration
ctDNA	Circulating Tumor DNA
DOC	Depth of Coverage
ENPB	Expected Normal Ploidy Buffer
FFPE	Formalin-Fixed Paraffin-Embedded
gDNA	Genomic DNA
HMM	Hidden Markov Model
ICB	Immune Checkpoint Blockade
LOY	Loss of Y
MAPD	Median Absolute Pairwise Difference
NGS	Next-Generation Sequencing
NIPT	Noninvasive Prenatal Testing
NSCLC	Non-Small Cell Lung Cancer
PD-1	Programmed Death Protein 1
PD-L1	Programmed Death-Ligand 1
PGT	Preimplantation Genetic Testing
PoN	Panel of Normals
RMSE	Root Mean Square Error
SNV	Single Nucleotide Variant
sWGS	Shallow Whole Genome Sequencing
TFx	Tumor Fraction
TMAP	Torrent Mapping Alignment Program
TMB	Tumor Mutational Burden
TP	Transition Penalty
WGS	Whole Genome Sequencing

Contents

List of Figures

List of Tables

1 Introduction

Non-Small Cell Lung Cancer (NSCLC) remains a leading cause of cancer-related mortality worldwide [1], [2], [3]. While immunotherapy has significantly improved survival [2], [3], [4], the lack of precise, non-invasive biomarkers for monitoring treatment response remains a critical clinical barrier [2], [3], [4]. Liquid biopsies of cell-free DNA (cfDNA) offer a minimally invasive alternative to traditional tissue biopsies, possibly overcoming the spatial and temporal limitations of tissue sampling [5], [6], [7]. To improve genomic profiling in precision oncology, this thesis evaluates the optimization of shallow whole genome sequencing (sWGS) for detecting copy number alterations (CNAs) from cfDNA [7], [8]. Specifically, it investigates the adaptation of an existing clinical workflow, originally designed for preimplantation genetic testing (PGT), to meet the more complex analytical requirements of precision oncology.

1.1 Aim

The aim of this study is to establish an optimal bioinformatic workflow for reliable copy number detection in liquid biopsy. The specific objectives are to:

- Optimize the sequencing protocol to increase total data yield and per-sample sequencing coverage.
- Adapt the clinical Ion Reporter PGT Workflow to ctDNA analysis.
- Evaluate and compare it against two open-source ctDNA analysis pipelines.
- Benchmark the three pipelines using a positive control.
- Evaluate the analytical performance on clinical NSCLC patient samples.

2 Theory

This chapter established the theoretical foundation necessary to understand the biological principles and computational methodologies evaluated in this study. The sections cover the genomic landscape of cancer and immunotherapy, the molecular characteristics of cell-free DNA, and the principles of whole genome sequencing. Finally, the computational depth-of-coverage workflows used for copy number profiling are detailed.

2.1 Genomic Alterations in Cancer

Genomic alterations can be broadly categorized into germline and somatic variants based on their cellular origin. Germline variants are inherited from parents, present at conception, and found in every cell of an individual, meaning they can be passed on to future offspring [9]. In contrast, somatic variants are non-inherited genetic alterations that emerge in non-reproductive cells after conception. Both germline and somatic alterations can range in scale from Single Nucleotide Variants, where a single nucleotide base is substituted for another, to Copy Number Alterations (CNAs), which include large-scale genomic gains and losses spanning megabase (Mb)-sized regions or entire chromosomes.

When these alterations occur somatically, they accumulate naturally over an individual's lifetime due to spontaneous DNA replication inaccuracies or exposure to environmental factors, such as tobacco smoke or ultraviolet radiation. Because a somatic variant is transmitted exclusively to the descendants or the original altered cell, the affected tissue becomes a cellular mixture of altered and normal DNA, which is called genomic mosaicism.

While the vast majority of somatic variants are benign and represent a normal part of cellular aging, specific alterations that disrupt the cell cycle or genome stability can lead to the development of cancer. Somatic variants that drive tumor growth and metastatic spread are termed "drivers", while those that are biologically inert are called "passengers" [10], [11]. In precision oncology, identifying specific driver variants is of particular clinical utility, as they may serve as predictive biomarkers for treatment response [6]. However, effective clinical translation requires distinguishing these truly actionable drivers from passenger events. Misinterpreting these individual variants risks exposing patients to ineffective, potentially toxic therapies or overlooking intrinsic drug resistance mechanisms

2.2 Non-Small Cell Lung Cancer (NSCLC)

Globally, lung cancer remains the primary cause of both cancer incidence and mortality [2], [3]. The disease is broadly divided into two subtypes: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). NSCLC accounts for 80-85% of lung cancer cases, presenting primarily as either adenocarcinoma (LUAD) or squamous cell carcinoma (LUSC). The high mortality associated with the disease is driven largely by a high frequency of late-stage diagnoses. Swedish clinical data from 2020-2024 shows that 64% of all NSCLC patients are diagnosed at Stage III-IV [12], when the disease has already metastasized. In contrast, only

approximately 28% of NSCLC cases are diagnosed at Stage I. The clinical impact of this late-stage detection is evident in the survival outcomes for Swedish patients. According to data from 2018-2024, the five-year survival rate for patients diagnosed with NSCLC at Stage I is approximately 68%, while it drops down to around 22% for those diagnosed at Stage III, and 9% for Stage IV [12].

Historically, the clinical treatment of metastatic (Stage IV) NSCLC relied primarily on platinum-based chemotherapy [2], [3]. Today, the therapeutic landscape is instead defined by molecular stratification. Patients with specific actionable oncogenic drivers (such as EGFR, ALK, ROS1) are treated with targeted therapies as a first-line standard of care. Conversely, for patients lacking these specific driver alterations, the therapeutic landscape has shifted significantly toward immunotherapy, which is often administered either as a monotherapy or in combination with traditional platinum-based chemotherapy.

2.3 Immunotherapy and Immune Checkpoint Blockade (ICB)

Unlike traditional cytotoxic drugs, immunotherapy functions by modulating the host's immune system to recognize and eliminate malignant cells [3], [13]. This is primarily achieved through immune checkpoint blockade (ICB), which targets regulatory pathways that tumors exploit to evade immune surveillance. In a healthy immune system, these checkpoints act as a protective mechanism to prevent immune cells from attacking the body's own cells.

The core of this process involves a highly specific receptor-ligand interaction: checkpoints engage when proteins on the surface of T cells bind to partner proteins on tumor cells [13]. When this binding occurs, a suppressive "off" signal is sent to the T cell, deactivating its ability to attack. Many tumors exploit this by producing high levels of the partner protein programmed death-ligand 1 (PD-L1), which binds to the programmed cell death protein 1 (PD-1) receptor on T cells to turn off the immune response (Figure 2.1).

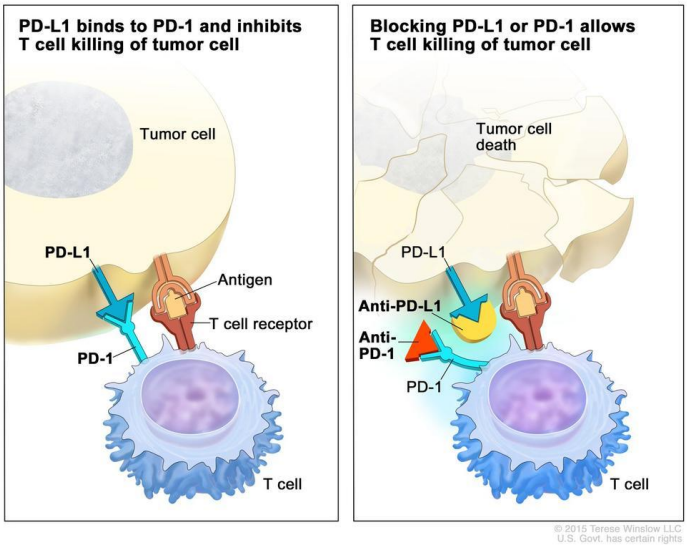

Figure 2.1: Mechanism of PD-1/PD-L1 Immune Checkpoint Inhibition.
The left panel demonstrates tumor immune evasion, where the binding of programmed death-ligand 1 (PD-L1) expressed on the surface of the tumor cell to programmed death 1 (PD-1) receptors on the T cell delivers an inhibitory signal, suppressing T-cell activation and preventing tumor cell lysis. The right panel illustrates the therapeutic intervention of immune checkpoint inhibitors; monoclonal antibodies targeting either PD-1 (anti-PD-1) or PD-L1 (anti-PD-L1) disrupt this inhibitory interaction, releasing the immunological brake and restoring T-cell-mediated cytotoxicity against the tumor cell. *Image credit: © 2015 Terese Winslow LLC, U.S. Govt. has certain rights. "Immune Checkpoint Inhibitors" was originally published by the National Cancer Institute* [13] *(https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors).*

At present, ICB in NSCLC mainly involves therapeutic agents designed to disrupt this PD-1/PD-L1 axis [1], [2], [3]. By inhibiting either PD-1 or PD-L1, ICB prevents the "off" signal from being transmitted, thereby reversing tumor-mediated immune resistance and restoring the capacity of T-cells to identify and destroy the tumor. Despite the success of ICB in improving overall survival, therapeutic response is not uniform across all patient cohorts. A significant proportion of patients do not respond to treatment, and some may experience severe immunotoxicities. To optimize patient selection, clinical research is currently directed toward the identification of biomarkers capable of predicting treatment response.

2.3.1 Current Predictive Biomarkers

Currently, the most widely utilized biomarker for predicting ICB response in clinical practice is the immunohistochemical assessment of PD-L1 expression on tumor cells [1], [2], [4]. Additionally, Tumor Mutational Burden (TMB) has emerged as a potential genomic indicator of response, as a higher abundance of somatic mutations is associated with increased immune recognition. However, these markers have significant limitations. For instance, PD-L1 has limited predictive power as treatment responses have been observed in PD-L1-negative patients, and TMB lacks standardized threshold across laboratories and sequencing platforms. While TMB has received regulatory approval for clinical use by the US Food and Drug Administration [14], it is currently not approved by the European Medicines Agency [15].

Furthermore, both markers are typically derived from tissue biopsies, which might be subject to sampling bias and fail to capture the spatial and temporal heterogeneity of metastatic disease.

These challenges highlight the need for dynamic, minimally invasive biomarkers. Previous work within this research group has focused on tracking targeted somatic point mutations and small insertion/deletions in patient plasma, establishing specific driver mutations, such as co-occurring *KRAS* and *LRP1B* variants, as predictive indicators of immunotherapy outcomes [4].

However, tumor evolution is a highly complex and heterogeneous process, and biomarkers based solely on individual gene variants may not fully capture the broader genomic instability that influences clinical outcomes. Therefore, additional markers reflecting large-scale genomic instability may be required to improve patient stratification. In particular, CNAs represent a form of this genomic instability, where wide chromosomal disruptions as well as regional copy number gains and losses can directly increase or decrease the expression levels of hundreds of genes simultaneously. Capturing these systemic genomic shifts requires high-throughput screening methods optimized to detect structural variations across various sample types, establishing the foundations for the shallow whole genome sequencing (sWGS) framework evaluated in this study.

2.4 Liquid Biopsy and Cell-Free DNA (cfDNA)

Traditionally, genetic biomarkers are identified via tissue biopsies. However, these are invasive, carry procedural risks, and can be difficult to obtain depending on the location of the tumor [5], [6], [16]. Furthermore, a single-site biopsy can fail to capture the spatial and temporal heterogeneity of evolving tumors. Liquid biopsies on the other hand, rely on tumor-derived DNA present in bodily fluids, primarily blood plasma [5], [6], [17]. This includes the analysis of Circulating Tumor Cells or the more widely adopted cell-free DNA (cfDNA) [2], [18]. Since liquid biopsies can be acquired through a simple blood draw, it allows for longitudinal monitoring of disease progression and treatment response [5], [6], [17].

2.4.1 Biology and Molecular Characteristics of cfDNA

Cell-free DNA (cfDNA) refers to the degraded DNA fragments released into the extracellular environment, primarily the bloodstream [19], [20], [21]. In healthy individuals, the vast majority of cfDNA originates from the natural turnover of hematopoietic cells [19], [21]. A typical cfDNA molecule consists of approximately 147 bp of DNA associated with the nucleosomal core, plus an approximately 20 bp unbound linker segment interspersed between nucleosomes (Figure 2.2) [20], [21]. This structural arrangement results in a highly characteristic fragment length distribution with a modal peak at approximately 167 bp. This non-random fragmentation pattern is widely attributed to apoptotic processes, where cellular DNA is enzymatically cleaved at accessible linker regions.

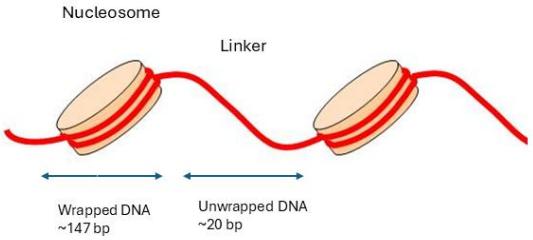

Figure 2.2: Nucleosomal Structure of Chromatin and the Basis for cfDNA fragmentation.
The schematic illustrates approximately 147 bp of DNA wrapped tightly around the histone octamer core of a nucleosome, bound to a single ~20 bp free linker DNA segment. Cleavage at these accessible linker sites during apoptotic degradation produces the characteristic ~167 bp modal fragment length observed in blood plasma. *Modified from original work by Pcauchy* [22]*, CC BY-SA 4.0.*

2.4.2 ctDNA and Tumor Fraction

In patients with cancer, a fraction of the total cfDNA might be derived from tumor cells and is termed circulating tumor DNA (ctDNA) [19], [21], [21]. Emerging studies suggest that tumor-derived fragments may exhibit a slightly shorter length profile compared to non-tumor cfDNA [20], [21]. Because ctDNA carries the same genetic alterations as the tumor cells it came from, analyzing blood samples may allow researchers to map out the genetic profile of a patient's cancer without needing a physical tissue biopsy [19]. The relationship between the ctDNA and the total cfDNA is defined by the Tumor Fraction (TFx), which represents the proportion of tumor-derived DNA relative to the total cfDNA in a sample [7]. A primary limitation of ctDNA analysis is that the tumor fraction is often very low [6]. Consequently, identifying the subtle signal of tumor-specific genetic alterations against the dominant "noise" of healthy DNA represents a significant challenge for bioinformatic detection.

2.5 Next-Generation Sequencing (NGS)

The establishment of sequencing technologies capable of massively parallel analysis enabled high-throughput sequencing and significantly reduced costs [23], [24], [25]. These technologies, known as Next-Generation Sequencing (NGS), can sequence an entire human genome in a single run. Ion Torrent, the NGS platform used in this project, utilizes a semiconductor chip to track sequencing-by-synthesis (Figure 2.3) [23], [26]. During this process, the incorporation of a nucleotide by DNA polymerase releases a hydrogen ion (H^+) as a byproduct. This localized pH shift is detected by an integrated sensor at the bottom of each microwell and converted into a proportional voltage signal. By sequentially flooding the chip with individual nucleotide species, the system maps these electrical spikes back to the corresponding bases.

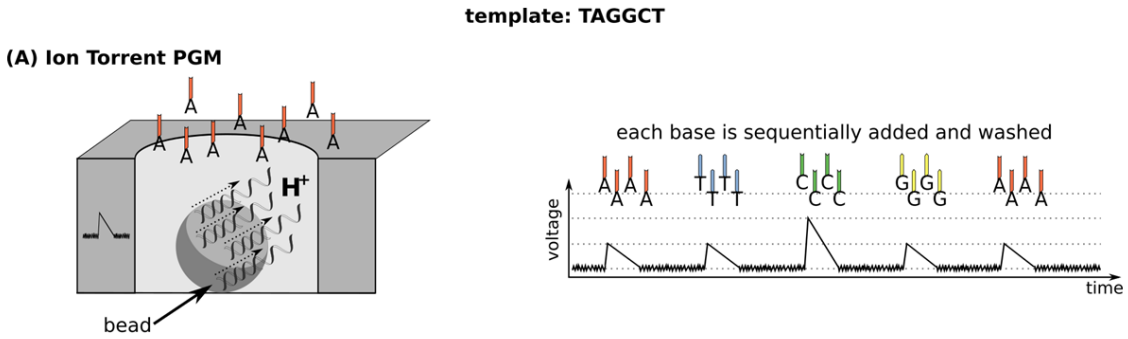

Figure 2.3: Schematic Cross-Section and Signal Registration of an Ion Semiconductor Sequencing Microwell.
(Left) Individual clonally amplified DNA templates on an ISP bead isolated within a microwell cavity. Nucleotide incorporation releases a hydrogen ion (H^+) byproduct, altering localized pH. (Right) The pH shift is registered as a voltage pulse proportional to base incorporation. *Adapted from Hupé (2012)* [27], *CC BY-SA 3.0.*

2.5.1 Whole Genome Sequencing (WGS) and Applications

One of the most common applications of NGS is Whole Genome Sequencing (WGS), which allows for comprehensive, non-targeted assessment of the whole genome. Unlike targeted approaches that require prior selection of specific genes or mutations, WGS enables the detection of both well-known and novel alterations across both coding and non-coding regions [24], [25]. However, WGS remains more expensive than targeted methods. This is primarily because targeted approaches, such as gene panels, sequence significantly smaller genomic regions, allowing them to achieve higher coverage at a lower overall cost per sample. Furthermore, while NGS technologies have made genomic data more accessible, the subsequent challenge of analyzing and interpreting the massive volume of raw data has become the new bottleneck in genomic analysis.

Traditional WGS typically aims for a coverage of 30x, which is sufficient for identifying germline variants but often lacks the resolution to capture rare, low-frequency somatic mutations present in cancer genomes [25]. However, the detection of CNAs does not require the same high-resolution depth as point mutation discovery [17], [28], [29]. Shallow WGS (sWGS), also referred to as low-pass WGS, is a cost-effective method that frequently utilizes a coverage of 0.1x to 1x for accurate CNA calling and tumor fraction estimation in liquid biopsies.

2.6 Depth of Coverage (DOC) Methods

The identification of CNAs in this study relies on Depth of Coverage (DOC) analysis. This approach utilizes the number of sequenced reads within specific genomic regions to determine their relative copy number states [28], [30], [31]. DOC analysis generally consists of three stages: binning and normalization, segmentation, and alteration calling [7], [28]. In this study, the Ion Reporter (Thermo Fisher Scientific) platform, and the ichorCNA [7] and WisecondorX [28] pipelines, were utilized to perform these analyses.

2.6.1 Genomic Binning and Normalization

Since sWGS provides sparse coverage that does not span every nucleotide position, the genome is first partitioned into "bins", which are contiguous windows of a fixed size [7], [28]. By aggregating the reads within these windows, whole-genome coverage is achieved at a bin-level resolution. The selection of bin size involves a critical trade-off between genomic resolution and statistical noise. Larger bins contain more reads and thus exhibit less stochastic variation. However, this reduction in noise comes at the expense of genomic resolution, thereby reducing the sensitivity for detecting smaller, focal alterations that span only a limited number of bins. Because a copy number calling algorithm typically requires a consecutive run of at least two to three bins to confidently identify an event, the chosen bin size directly determines the physical limit of detection.

Following binning, normalization is required to correct for technical artifacts, such as GC-content and mappability biases [28]. By adjusting the raw read counts against a reference baseline, the data is converted into read-depth ratios representing the relationship between observed counts in the sample and those expected in a diploid reference. These ratios provide the necessary input for downstream analysis to identify and characterize true CNAs.

2.6.2 Segmentation and Copy Number Calling

Following normalization, read-depth ratios undergo segmentation to group neighboring bins with similar ratios into larger segments [28]. This process identifies breakpoints where ratios shift significantly, simultaneously allowing the software to assign a discrete biological state (Gain, Loss, or Neutral) to each segment. In this study, these analyses were performed using the Ion Reporter Aneuploidy, ichorCNA, and WisecondorX pipelines.

The Ion Reporter Aneuploidy and ichorCNA pipelines utilize a Hidden Markov Model (HMM) for segmentation and calling. An HMM estimates the most likely sequence of "hidden" biological states (Gain, Loss, or Neutral) based on the "observed" read-depth ratios [32]. This is achieved by balancing emission probabilities, which describe the likelihood that an observed read-depth ratio belongs to a specific copy-number state, with transition probabilities, which describe the likelihood of moving from one state to another between adjacent bins. A ploidy buffer is often used to ensure only significant deviations from the diploid baseline are reported as alterations. Aside from alteration calling, the ichorCNA pipeline estimates global ploidy, the tumor fraction, and the subclonal prevalence within the ctDNA population.

For WisecondorX, the underlying logic is based on Circular Binary Segmentation (CBS), a change-point detection method that recursively splits the genome into regions where the copy-number level is statistically constant [33]. The algorithm identifies positions where the statistical properties of the read-depth ratios shift significantly, suggestion a transition in the underlying genomic state rather than stochastic noise. The pipeline then separates significant segments from normal regions using statistical thresholds: alpha determines the significance of new breakpoints, while the Z-score determines how many standard deviations a segment must deviate from a reference baseline to be officially "called" as an alteration.

2.6.3 Pipeline-Specific Inputs and Outputs

Beyond the core steps of normalization and segmentation, the specific software platforms utilized in this study rely on distinct operational parameters to control calling sensitivity and filter background noise.

Ion Reporter [34]:

- **Tile Size:** The size of the genomic bins, requiring an input value between 1 and 10,000,000 base pairs.
- **Mosaicism Detection:** A setting to enable or disable the identification of fractional copy number (CN) states, allowing the detection of mosaic events. When disabled, the calling algorithm constraints all reported alterations to discrete, integer copy numbers.
- **CNV Transition Penalty (TP):** The parameter for tuning the transition possibility within the HMM. Less negative values (lower penalties) increase sensitivity by making state changes easier to trigger. This parameter can be configured via sensitivity presets (low, medium, and high) or custom numerical values, where -1.05 is the highest sensitivity when mosaicism is disabled, and -2.31 when mosaicism is enabled.
- **Expected Normal Ploidy Buffer (ENPB):** A threshold value that defines a linear, absolute copy number safety margin (e.g. CN 1.8-2.2) around the diploid baseline (CN = 2) to filter out minor fluctuations and prevent them from being called as alterations.
- **Smoothing:** A setting to enable or disable the graphical averaging of the scattered raw data points into a continuous trendline, without altering the underlying algorithmic calls. However, because this design was high-purity, high-quality DNA samples, its utility remains uncertain for low-fraction ctDNA where CN signals are inherently low-amplitude and might be obscured by visual averaging.
- **Median Absolute Pairwise Difference (MAPD**): A global metric that quantifies the noise level of the data. It is calculated by determining the sample-to-reference read ratio for every genomic bin, and then measuring how much this value varies between all adjacent bins. A lower MAPD indicates more stable and uniform data, increasing the reliability of the CNA calling.
- **Analytical Output:** A list of detected genomic alterations, reported as absolute CN states, and their genomic start and end positions. This data is accompanied by a genome-wide CN plot scaled by absolute CN, displaying the raw read-depth data points overlaid with segmented CN lines.

ichorCNA [35]:

- **Bin Size:** The size of the genomic bins. The pipeline natively supports bin sizes of 10 kilobases (kb), 50 kb, 500 kb, and 1000 kb due to pre-compiled companion files for GC-content and mappability corrections.
- **Normal Fraction**: Defines the initial parameter value for the expected dilution of the sample by non-tumor cfDNA. The parameter accepts a list of multiple starting

estimates between 0 and 1.0, prompting the algorithm to run parallel optimization paths to prevent premature convergence on local maxima.

- **Subclonal Detection**: A setting to enable or disable the statistical estimation of subclonal CN events within the ctDNA population.
- **Transition Probability**: The parameter that sets the transition probability within the HMM, given as a value between 0 and 1.0. Higher values (closer to 1) increase the model's resistance to state changes, thereby lowering sensitivity.
- **Analytical Output:** A set of data and visualization files comprising:
    - **Alteration Segment File**: A list of detected genomic alterations and their genomic start and end positions, providing the discrete status (gain or loss) alongside segment-averaged log2 ratios and absolute CN states adjusted for the estimated tumor fraction.
    - **Genome-Wide Bin File**: A comprehensive data file containing the calculated log2 ratio for every individual genomic bin across the entire genome.
    - **Model Estimation**: A report detailing alternative optimized solutions, including combinations of estimated tumor fraction, ploidy, and subclonal fraction alongside the log-likelihood score for each model.
    - **Visualization plots:** Both genome-wide and chromosome-level CN plots scaled by log2 ratios, displaying the raw read-depth data points overlaid with segmented CN lines.

WisecondorX [36]:

- **Bin Size**: The size of the genomic bins. The pipeline is optimized for bin sizes ranging from 50 to 500 kb for sWGS (0.1x-1x coverage).
- **Significance Threshold (alpha):** The p-value cutoff used during CBS to determine the statistical validity of genomic breakpoints (default = 1×10^{-4}).
- **Z-score Cutoff:** Sets the standard deviation threshold used post-segmentation to evaluate read-depth deviations against the reference baseline (default = 5).
- **Analytical Output:** A set of data and visualization files comprising:
    - **Alteration Segment File:** A list of detected genomic alterations and their genomic start and end positions, providing the discrete status (gain or loss) alongside segment-averaged log2 ratios and z-scores.
    - **Genome-Wide Bin File:** A comprehensive data file containing the calculated log2 ratio and z-score for every individual genomic bin across the entire genome.
    - **Visualization plots:** Both genome-wide and chromosome-level CN plots scaled by log2 ratios, displaying the raw read-depth data points overlaid with segmented CN lines.

3 Method

This chapter outlines the experimental and computational methodologies implemented in this study. It describes the overarching study design and the patient cohort, as well as the experimental sequencing procedure. Furthermore, it details the computational pipelines, optimization strategies, and benchmarking metrics used to analyze the data.

3.1 Study Design and Cohort

3.1.1 The BioLung Cohort

The BioLung initiative started 2019, and is an ongoing prospective observational cohort study aimed at identifying biomarkers for stratifying NSCLC patients. The cohort includes patients with stage III-IV NSCLC undergoing immunotherapy targeting the PD-1/PD-L1 pathway.

Longitudinal blood samples are collected in conjunction with the patient's treatment session during the first five therapeutic cycles, designated A through E, as illustrated in the sampling timeline in Figure 3.1. This is followed by a 1-year follow-up sample or a disease progression sample (F) which can be drawn at any point during therapy. Additionally, treatment may be discontinued prematurely due to other clinical factors such as toxic reactions.

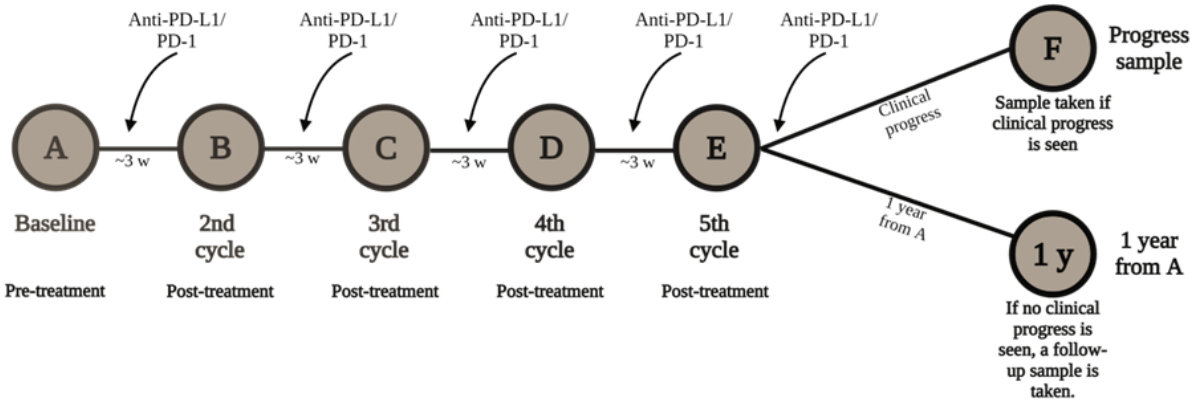

Figure 3.1: BioLung Longitudinal Plasma Sampling Protocol.
Blood collection windows are aligned with routine therapeutic administration cycles (A through E), with an extended clinical follow-up collection point at 1 year. The protocol features adaptive sampling pathways, allowing for an emergency collection point (F) upon verified radiological or clinical disease progression, or early discontinuation due to treatment-related toxicity. *Image courtesy of the BioLung Study Consortium.*

3.1.2 Patient Demographics and Associated Metadata

For this study, a subset of 24 patients was selected from the larger BioLung cohort. This specific patient cohort was pre-determined and provided by the parent project team. The relevant clinical metadata from this 24-patient subset are summarized in Table 3.1.

Table 3.1: Baseline Clinical Characteristics and Demographics of the Patient Cohort (N = 24).

Age at ICB Start	≥ 70	14 (58.3%)
	< 70	10 (41.7%)
	Median	71
	Range	58-80
Sex	Female	11 (45.8%)
	Male	13 (54.2%)
Histodiagnosis	LUAD	19 (79.2%)
	LUSC	4 (16.7%)
	NSCLC NOS	1 (4.2%)
Stage	III	3 (12.5%)
	IV	21 (87.5%)
PD-L1 Expression	High (≥ 50)	12 (50%)
	Low (< 50)	11 (45.8%)
	Not Determined	1 (4.2%)
Smoking History	Smoker (Current or Former)	23 (95.8%)
	Never Smoker	1 (4.2%)

Note: Percentages are rounded to one decimal place, and the row totals may deviate slightly from 100.0%.

3.1.3 Experimental Samples Sets

The genetic material analyzed in this study comprises specimens from NSCLC patients alongside healthy blood donor controls used for bioinformatic normalization and baseline generation. The analytical dataset is divided into two primary sub-cohorts based on historical processing timelines, as schematically outlined in Figure 3.2:

- **Initial Sub-cohort**: Features matching formalin-fixed paraffin-embedded (FFPE) tumor tissue DNA and plasma cfDNA pairs obtained prospectively from 16 NSCLC patients, accompanied by 12 healthy gDNA control samples.
- **New Sub-cohort**: Features matching FFPE tumor tissue DNA and plasma cfDNA pairs from 8 patients, accompanied by 7 healthy gDNA control samples. Additionally, a single positive control sample utilizing the A549 NSCLC cell line was integrated for analytical truth-set validation against public DepMap record [37].

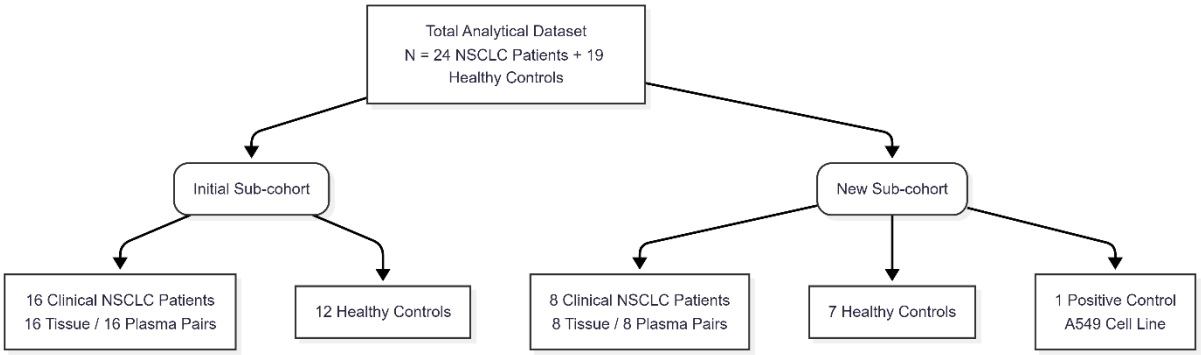

Figure 3.2: Flowchart of the Experimental Clinical Dataset Structure.
Schematic partitioning of the total analytical population (N = 24 NSCLC patients and 19 healthy controls) into Initial and New sub-cohorts. The diagram explicitly tracks sample types, separating matching clinical patient pairs (solid tissue tumor DNA and liquid biopsy plasma cfDNA) from the healthy blood donor reference controls. *Flowchart designed by the author.*

3.1.4 Overview of Experimental Workflow

This study utilized a multi-stage workflow encompassing wet-lab processing, high-throughput sequencing, and parallel computational benchmarking to evaluate CNA detection performance, as mapped out in Figure 3.3:

1. Wet-Lab Processing and Sequencing: High-throughput Ion Torrent sequencing was performed exclusive on the samples from the New sub-cohort (n = 8 patients, 7 healthy controls, and 1 positive control), following raw DNA extraction and library preparation.
2. Pipeline Optimization and Benchmarking: Execution parameters for Ion Reporter were systematically optimized, and outputs across all three analytical frameworks were evaluated using the computational benchmarking data from the A549 positive control.
3. Computational CN Profiling: Raw digital sequencing data from both the Initial sub-cohort and the newly sequenced datasets were processed uniformly across all three pipelines to generate and compare the final genomic CNA profiles.

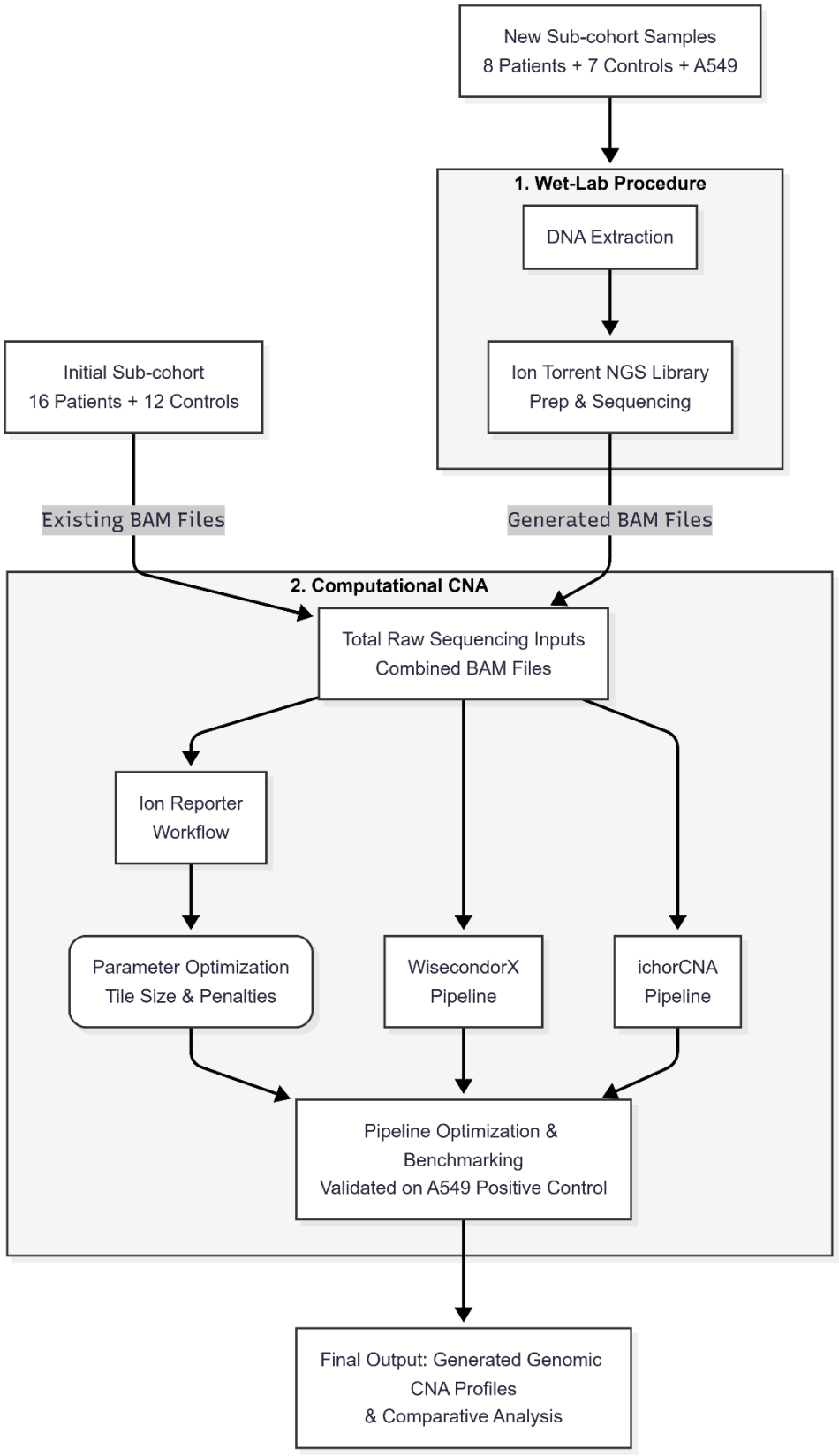

Figure 3.3: Flowchart of the Multi-Stage Experimental Workflow.
Visual tracking of sample progression, highlighting the distinct entry points for the prospective New sub-cohort (subjected to physical wet-lab extraction and Ion Torrent sequencing) and the historical Initial sub-cohort (integrated directly as pre-existing digital BAM sequencing files) prior to parallel bioinformatic profiling and benchmarking. *Flowchart designed by the author.*

3.2 Experimental Procedure

Blood samples were collected in Streck-tubes for cfDNA preservation. Plasma was subsequently isolated using a double-centrifugation protocol to minimize genomic DNA (gDNA) contamination from leukocyte lysis [38]. Total cfDNA was extracted using the EZ1&2 ccfDNA Kit (Qiagen) [39] and quantified using the Qubit® 1x dsDNA HS Assay [40]. Prior to library preparation, samples were diluted to a target concentration of 300 pg/µL.

Library preparation was performed using the Ion ReproSeq™ PGS Kit (Thermo Fisher Scientific). The initial cell extraction and lysis steps were omitted as the protocol was initiated directly with the purified cfDNA and FFPE tumor DNA inputs. The remaining steps were performed according to the manufacturer's instructions [41]. Following construction, post-amplification library integrity was qualitatively verified via agarose gel electrophoresis. Successful library generation was confirmed by the presence of a characteristic smear.

The template loading concentrations and pooling strategies were iteratively optimized across four distinct execution phases with the primary objective of increasing sequencing coverage, as shown in Figure 3.4:

- **Phase 1 & 2 (Prior to this study):** Initial sequencing runs utilized 16 samples (Phase 1) and 8 samples (Phase 2) of mixed types (tumor, cfDNA, and gDNA). These runs followed the institutional clinical PGT protocol at the time, which employed equivolume pooling and a loading concentration of 80 pM.
- **Phase 3 (Pilot Optimization):** An experimental run of 8 samples (mixed types) was conducted at a reduced loading concentration of 30 pM, mirroring an updated institutional loading concentration standard. In this phase, equimolar pooling was evaluated as a strategy to normalize coverage across the diverse sample types.
- **Phase 4 (Final Protocol):** The final workflow utilized 8 samples segregated by sample type and returned to equivolume pooling. To account for the distinct loading dynamics of different sample types, the target concentration was split: 30 pM was maintained for healthy gDNA controls, while 60 pM was used for cfDNA and tumor DNA libraries.

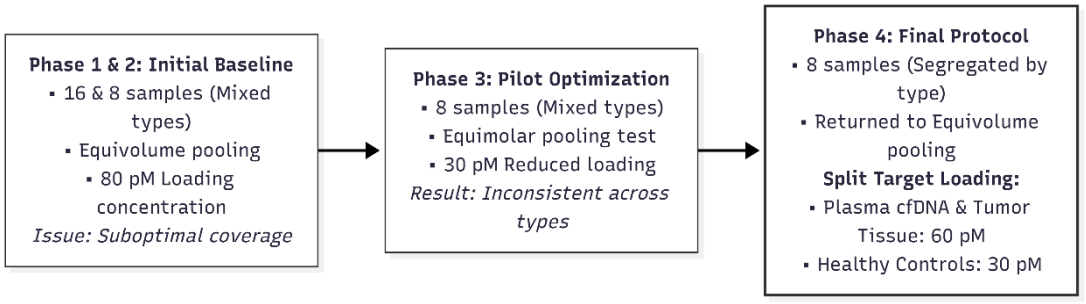

Figure 3.4: Target Loading Concentration and Pooling Adjustments across Execution Phases. Summary of the iterative changes made to the sequencing setup to improve cross-sample coverage. *Flowchart designed by the author.*

Following pooling, the libraries underwent a final magnetic bead-based purification step to remove residual reagents and adapter dimers. The purified pools were then diluted to their respective target loading concentrations. Automated template preparation and chip loading were conducted on the Ion Chef™ Instrument, followed by sequencing on the Ion GeneStudio™ S5 System using an Ion 510™ Chip.

3.3 Bioinformatic Analysis

3.3.1 Data Processing and Alignment

Following sequencing, raw data was processed by the Ion Torrent™ Suite Software (v. 5.20; Thermo Fisher Scientific). This included demultiplexing and automatic alignment to the hg19 (GRCh37) human reference genome using the Torrent Mapping Alignment Program (TMAP). The raw sequence data was subsequently remapped to the newer hg38 (GRCh38) human reference using the native TMAP. The resulting BAM files served as the primary input for the bioinformatic analysis by the Ion Reporter™ Aneuploidy, ichorCNA, and WisecondorX pipelines. The complete alignment data progression and subsequent parameters optimization framework are schematically structured in Figure 3.5.

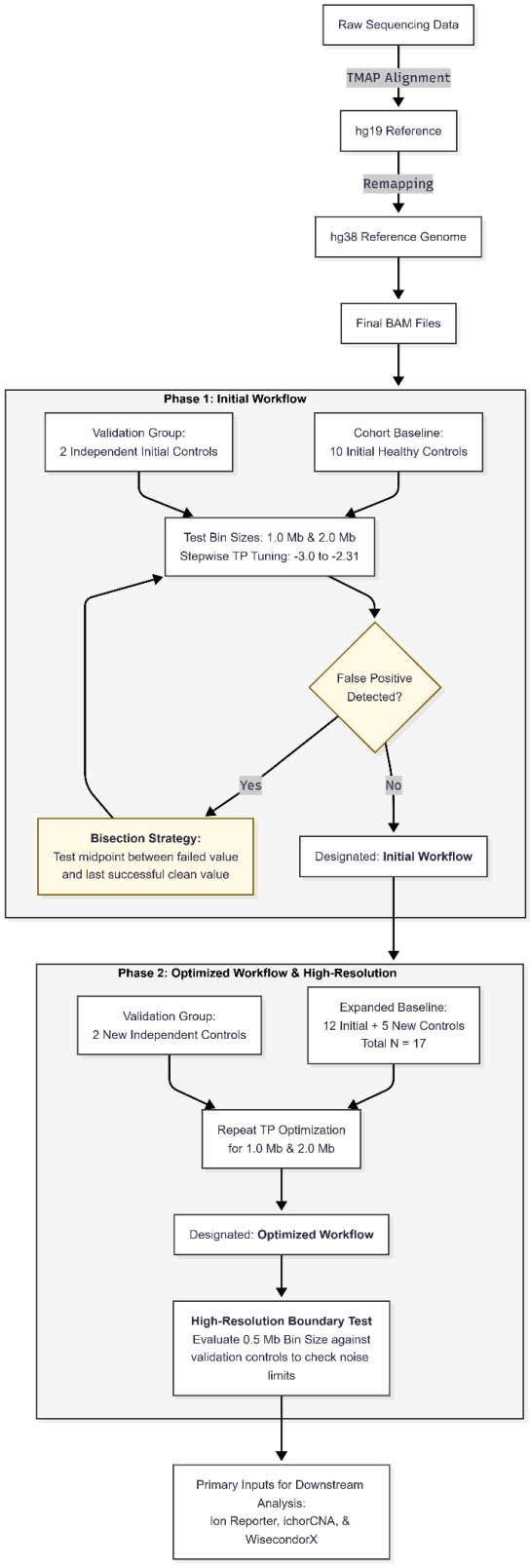

Figure 3.5: Bioinformatic Alignment Pathway and Iterative Ion Reporter Parameter Optimization Framework.

Schematic mapping of raw data alignment transitions from hg19 to hg38 alongside the two-phase workflow optimization matrix. The diagram outlines the baseline expansion strategy (from 10 to 17 healthy controls), the mathematical bisection loop used to counter false-positive calls, and the final high-resolution 0.5 Mb boundary evaluation. *Flowchart designed by the author.*

3.3.2 Baseline Framework

The institutional clinical PGT workflow at Sahlgrenska Clinical Genetics and Genomics served as the comparative baseline for this study. The pipeline operates under the following standard parameters:

- **Tile Size:** 2.0 Mb
- **TP:** -3.0
- **Mosaicism Detection**: Enabled
- **Smoothing:** Enabled
- **ENPB:** CN 1.7-2.3

3.3.3 Ion Reporter Optimization

To align with the biological reality of liquid biopsies, mosaicism detection remained enabled for all subsequent optimization phases to preserve sensitivity for low-fraction events. Furthermore, an ENPB of CN 1.8-2.2 was implemented. This adjustment was made to increase the likelihood of detecting these lower-magnitude alterations, accepting an increased risk for false positives compared to the more conservative 15% buffer used in the clinical PGT workflow. To ensure the highest mapping precision for the low-fraction ctDNA, the optimization was transitioned to the hg38 reference genome. Using this framework, a cohort-specific baseline was constructed, and the TP parameter was optimized across two phases to maximize sensitivity while strictly controlling for false positives.

In the first phase, the baseline comprised 10 healthy controls from the Initial sub-cohort (averaging 123,212 reads/sample). The TP parameter was evaluated for both 1.0 Mb and 2.0 Mb bin sizes, testing the values against two independent healthy controls (Initial sub-cohort) used strictly for validation.

- **The Process**: Starting from the clinical baseline of TP -3.0, the penalty was incrementally lowered toward a maximum sensitivity threshold of TP -2.31.
- **The Bisection Strategy**: Whenever a tested TP value generated a false positive in the validation samples, a bisection approach was applied: testing the exact midpoint between that failed value and the last successful “clean” value.

The optimized configuration resulting from this first phase was designated as the Initial Workflow.

In the second phase, the custom baseline was expanded to 17 samples by combining all 12 healthy controls from the Initial sub-cohort with 5 new, higher-depth healthy controls (averaging 227,998 reads/sample). The optimization process was then repeated across both 1.0 Mb and 2.0 Mb, evaluating the TP using the expanded baseline on two new independent healthy controls for final validation. The final configuration established at the end of this second phase was designated as the Optimized Workflow.

To explore the detection limits of optimized workflow, a high-resolution 0.5 Mb bin size was subsequently evaluated for the expanded 17-sample baseline. This evaluation was conducted to test the boundaries of the finalized settings, specifically applying the optimized TP value to a

higher resolution. Because smaller bin sizes are inherently prone to increased background noise, the workflow was assessed against the independent validation controls to determine if this tighter resolution could be utilized without compromising specificity or introducing false-positive calls.

3.3.4 Rationale for Pipeline Selection

The open-source copy number analysis pipelines, ichorCNA and WisecondorX, were selected based on their distinct analytical capabilities across varying tumor fractions. In addition to copy number calling, ichorCNA provides comprehensive estimation of global metrics, including: tumor fraction, baseline ploidy, and subclonal metrics. It operates under specific developer-defined thresholds: a minimum of 3% tumor fraction is required to detect the presence of tumor content in a sample; reliable, automated copy number calling requires 10% tumor fraction; and the remaining 3-10% span requires manual inspection. Conversely, WisecondorX was originally designed for Non-Invasive Prenatal Testing (NIPT), a clinical application where samples typically present with a lower biological baseline of 4% to 15% fetal fraction. WisecondorX was therefore selected to evaluate its performance at the lower tumor fractions where ichorCNA requires manual inspection or fails to detect tumor content.

3.3.5 Benchmarking

To evaluate the consistency between the three copy number platforms, a benchmarking analysis was performed using custom R scripts (v. 4.5.2; see Appendix A for full source code). The A549 positive control served as the primary benchmark, since it has a known, complex copy number profile. It includes 360 reported CNAs (109 gains and 251 losses) with a median length of 212,000 bp (range: 3,000-85,187,000 bp) [37]. This cell line is hypotriploid, meaning its neutral genomic baseline is shifted toward a copy number of three rather than the typical human diploid standard of two. The performance of the pipelines was assessed through two methodologies: signal correlation and bin-level concordance, using the DepMap public database as the ground truth reference dataset. The execution architecture of this two-tier benchmarking framework, including the grid standardization and downstream statistical assessments, is mapped out in Figure 3.6.

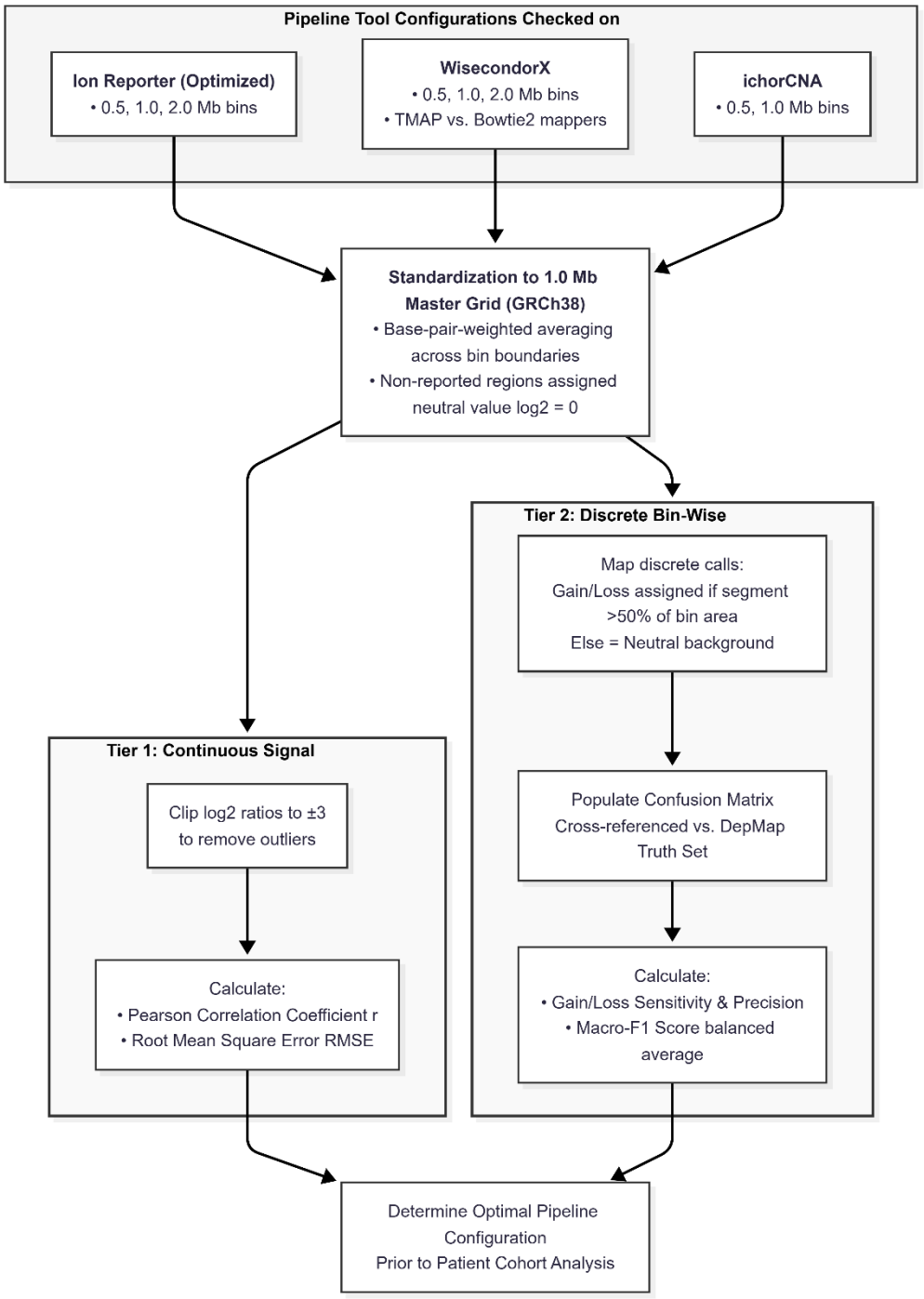

Figure 3.6: Flowchart of the Two-Tier Bioinformatic Benchmarking Architecture.
Structural overview of data harmonization across varying resolution bin sizes (0.5, 1.0, and 2.0 Mb) onto a unified 1.0 Mb master grid. Tier 1 handles continuous signal evaluation (Pearson *r*, RMSE) following extreme outlier clipping, while Tier 2 executes categorical classification tracking (Sensitivity, Precision, Macro F1-score) using spatial area thresholds cross-referenced against the public DepMap truth set. *Flowchart designed by the author.*

The benchmarking framework was used to evaluate how different configurations affected pipeline performance prior to patient cohort analysis. For ichorCNA, the analysis examined the impact of subclonality detection and initial non-tumor fraction seeds on the stability of the final solution. For WisecondorX and Ion Reporter Optimized Workflow, performance was compared

across 0.5, 1.0, and 2.0 Mb. Additionally, WisecondorX was evaluated using two different mappers: the native TMAP and Bowtie2 (v. 2.5.5). This was done to address developer findings that Bowtie2-mapped data yields superior results within the WisecondorX framework [28].

The first tier of analysis measured global accuracy by mapping all log-ratio results to a standardized 1.0 Mb master grid based on the hg38 reference. To compare different workflow bin sizes against the A549 truth set, a base-pair-weighted average was used to assign log2 ratios to the grid. This approach weighted segments by their genomic length when spanning bin boundaries to provide a precise comparison.

Log2 ratios were clipped to a range of ±3 to mitigate the influence of technical outliers. For workflows providing alteration-only files, non-reported regions were assigned a neutral value (log2 = 0) to allow for a full-genome comparison. Signal agreement between the test runs and the truth set was then quantified using the Pearson Correlation Coefficient (*r*) and Root Mean Square Error (RMSE).

The second tier evaluated the accuracy of discrete CNA calls (Gain, Loss, or Neutral) through a bin-wise confusion matrix analysis. This utilized the same 1.0 Mb master grid as the signal correlation. While workflow outputs were provided as alteration segment lists, they were mapped onto the grid by assigning “Gain” or “Loss” labels to bins where the segment covered more than 50% of the bin area. All remaining bins were classified as “Neutral” to provide a complete genomic background, matching the full-profile of the A549 truth set.

Each 1.0 Mb bin was cross-reference between the truth set and the experimental results to populate the confusion matrix. Sensitivity and precision were calculated independently for both the “Gain” and “Loss” categories. Finally, a Macro F1-Score was calculated by averaging the F1 scores of both gains and losses. This provided a balanced summary of diagnostic performance, ensuring that detection accuracy for both types of alterations were represented equally.

3.3.6 Ion Reporter Analysis

Following workflow optimization, copy number analysis was performed on the total patient cohort (n = 24). Sample inclusion for initial processing was primarily determined by the manufacturer-defined input quality control metric of > 100,000 reads for a sample. However, one sample from the Initial sub-cohort with 99,575 reads was retained, as the < 0.5% deviation from the 100,000-read threshold was considered negligible for 1.0 Mb resolution analysis. Individual read counts for the entire sample cohort are detailed in Table B. 1 (Appendix B).

Because this study focuses on liquid biopsy pipeline performance, quality filtering was evaluated on a per-sample basis to maximize plasma data retention. Plasma cfDNA samples clearing QC were retained for standalone bioinformatic analysis regardless of the QC status or their matching tissue sample, whereas patients lacking viable cfDNA data were excluded. Consequently, a total of 19 patients retained fully matched tissue-plasma pairs, while one additional patient was retained exclusively for standalone cfDNA evaluation, yielding a final plasma analytical cohort of n = 20.

Following the application of these criteria, the remaining viable plasma samples were processed as follows:

- **Initial Clinical Cohort** (n = 13 plasma remaining): Analyzed using the Initial Workflow at 1.0 and 2.0 Mb resolutions.
- **New Patient Cohort** (n = 7 plasma remaining): Analyzed using the Optimized Workflow at 1.0 and 0.5 Mb resolutions.

3.3.7 ichorCNA Analysis

Copy number analysis was then performed on the QC-passed patient cohort (n = 20) using ichorCNA (v. 0.3.2) at 1.0 Mb and 0.5 Mb resolutions. Input BAM files were converted to WIG files using the HMMcopy *readCounter* command. A Panel of Normals (PoN), was constructed using the healthy gDNA control cohort (n = 17) to provide a custom reference baseline. The pipeline was configured and executed within a Linux environment via the command-line interface.

Analysis was conducted using the parameter values specified in the official documentation:

- **Normal Fraction Candidates**: [0.5, 0.6, 0.7, 0.8, 0.9]
- **Subclonal Detection:** True
- **Transition Probability**: 0.9999

3.3.8 WisecondorX Analysis

The QC-passed patient cohort (n = 20) were further evaluated using WisecondorX (v. 1.3.1) at 1.0 Mb, and 0.5 Mb resolutions. FASTQ files were mapped to the hg38 reference genome using the mapper Bowtie2. The resulting BAM files were then converted to NPZ files using the WisecondorX *convert* command, and a reference baseline was generated from the healthy gDNA control cohort (n = 17). Similar to ichorCNA, the workflow was deployed within a Linux environment and executed using the same command-line environment. Analysis was conducted using default parameters:

- **Significance Threshold**: 1×10^{-4}
- **Z-score Cutoff**: 5

An overview of the sample filtering criteria, cohort stratification, and parallel pipeline processing configurations for the entire bioinformatic workflow is visualized in Figure 3.7.

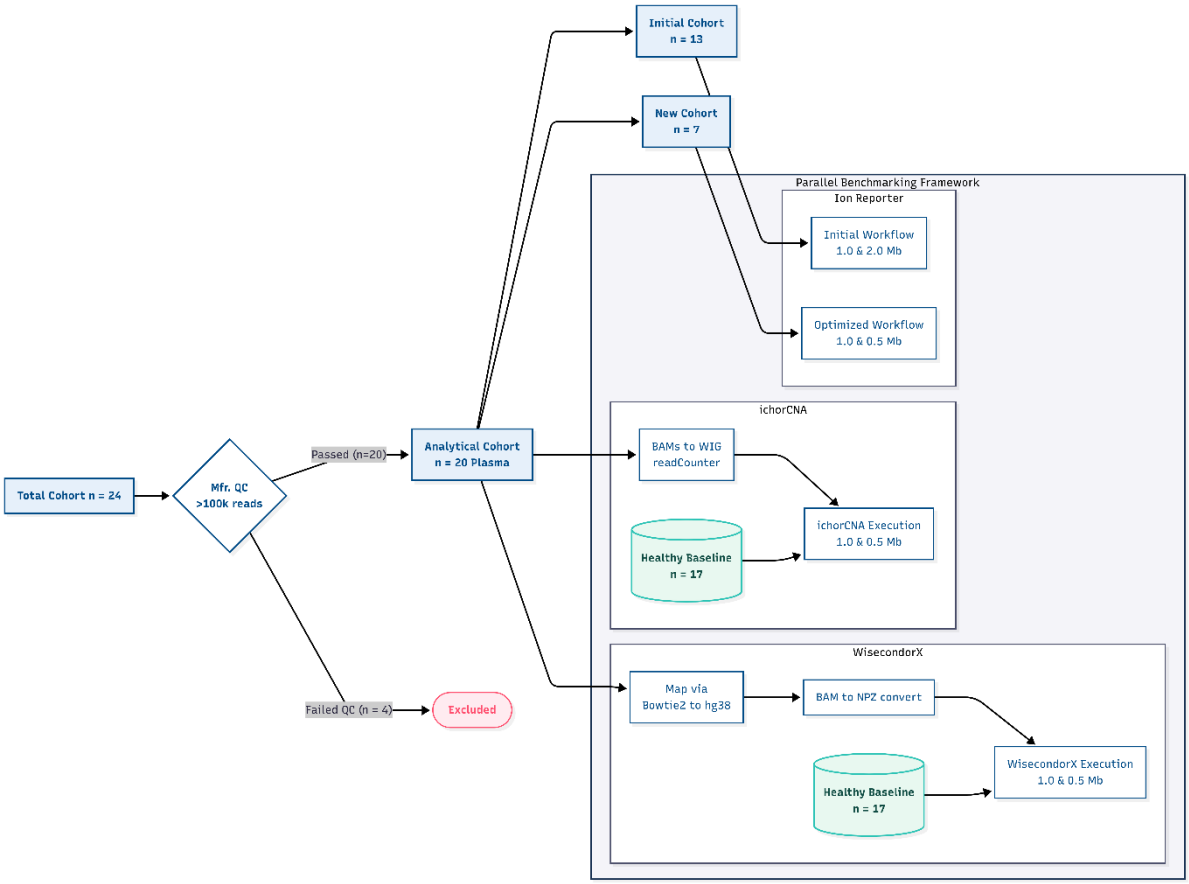

Figure 3.7: Methodological Pipeline and Cohort Processing Flowchart.
The diagram illustrates the sample quality control funnel from the initial cohort (n = 24) to the final plasma analytical cohort (n = 20). The parallel benchmarking framework highlights the independent algorithmic tracks, resolution settings, and healthy control baseline inputs (n = 17) utilized across Ion Reporter, ichorCNA, and WisecondorX. *Flowchart designed by the author.*

4 Results and Discussion

This chapter presents the experimental and computational findings of the study. It covers the pre-analytical optimization of the sequencing framework, the performance benchmarking of the three bioinformatic pipelines, and the final application of these workflows to CNAs in the clinical NSCLC patient cohort.

4.1 Sample Inclusion and Quality Filtering

While a total pool of 24 real-world NSCLC patient cases was available for evaluation, manufacturer-defined quality control filtering (> 100,000 reads) resulted in a final clinical cohort of 20 viable plasma cfDNA samples, which included 19 fully matched tissue-plasma pairs. Consequently, these quality-filtered samples were used in the downstream bioinformatic pipelines for copy number profiling and comparative analysis.

4.2 Pre-Analytical Optimization

4.2.1 Sequencing Run Performance across Development Phases

To establish a robust sequencing framework for low-input cfDNA profiling, the pre-analytical workflow was optimized across four iterative phases. The experimental parameters tested, including shifting pooling strategies, loading concentrations, and the number of samples per chip, are summarized in Table 4.1. These are presented alongside the primary instrument metrics generated automatically by the Ion Torrent software (Total Reads, Final Library, and Usable Sequence), which reflect raw sequencing output independent of downstream alignment. Additionally, to account for technical variance, Table 4.1 includes the average genome coverage calculated directly from the individual sample reads assigned by the instrument for both the raw dataset and a filtered subset applying the manufacturer-defined QC threshold (> 100,000 reads per sample).

Table 4.1: Sequencing Performance Metrics Across Iterative Pre-Analytical Optimization Phases.

Phase	Samples per Chip	Loading Conc. (pM)	Total Reads	Avg. Genome Coverage (Unfiltered) (x)	Avg. Genome Coverage (>100k Filtered) (x)	Final Library (%)	Usable Sequence (%)
1	16[a]	80	1,993,832 ± 136,382	0.0044 ± 0.0010	0.0046 ± 0.0008	63.8 ± 4.5	36.8 ± 3.3
2	8[a]	80	1,958,695	0.0079 ± 0.0011	0.0079 ± 0.0011	67	35
3	8[b]	30	873,881	0.0037 ± 0.0025	0.0054 ± 0.0024	69	51
4	8[a]	60/30[c]	2,348,412 ± 111,877*	0.0098 ± 0.0037*	0.0101 ± 0.0033*	77.3 ± 1.2*	44.3 ± 3.5*

[a] Equivolume pooling. [b] Equimolar pooling. [c] Target loading concentration set to 60 pM for cfDNA and tumor DNA samples, and 30 pM for healthy controls. * Statistical significance compared to Phase 1 ($p < 0.05$, Welch's t-test).

Comparison between the final configuration in Phase 4 against the configuration in Phase 1 shows a statistically significant increase across all key instrument parameters, including total reads, final library percentage, and usable sequence percentage ($p < 0.05$, Welch's t-test). Most notably, the protocol modifications yielded a significant increase in both unfiltered average genome coverage ($p = 3.93\times10^{-7}$, Welch's t-test) and filtered average coverage ($p = 1.04\times10^{-7}$, Welch's t-test).

While these data demonstrate a statistically robust relative improvement over Phase 1, the absolute average coverage stabilized at approximately 0.01x (0.0101 ± 0.0033x). This depth sits below the 0.1x to 1x range typical of standard sWGS copy number protocols [17], [28], [29]. This ultra-low coverage baseline likely reflects the technical constraints inherent to processing highly sparse, low-input cfDNA material where template preparation efficiency is physically limited. Rather than indicating a systematic failure, this low-coverage setting serves to evaluate whether downstream bioinformatics pipelines can successfully extract meaningful biological signals from minimal input data.

4.2.2 Sample-Type Read Allocation

While the clinical PGT workflow relies on equivolume pooling due to uniform sample inputs, an equimolar pooling strategy based on Qubit fluorometric quantification was initially trialed to achieve a more precise read distribution. However, initial results demonstrated highly skewed read allocations across individual samples. Following an initial sequence run that yielded low total outputs due to incorrect, manufacturer-supplied sequencing reagents incompatible with the library preparation kit, the remaining volume of the same physical library pool was re-sequenced in a separate run using the correct reagents. Comparing these two runs allowed for an evaluation of whether this uneven read distribution was a stochastic artifact of the sequencing process itself or a result of the physical pool composition. The resulting comparison of relative read allocation between the two runs is illustrated in Figure 4.1 (individual sample read counts are available in Table B. 2, Appendix B).

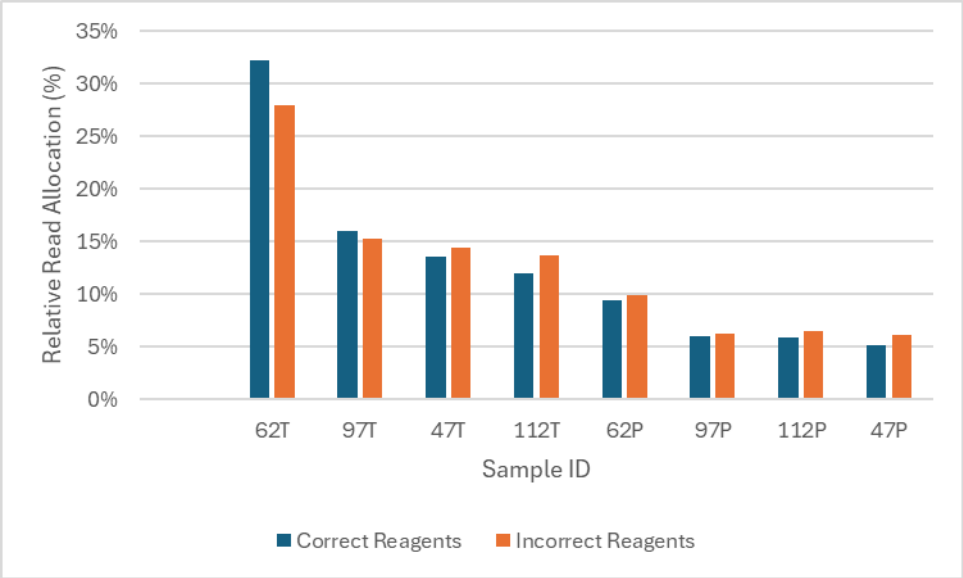

Figure 4.1: Relative Read Allocation Profiles between Independent Sequencing Executions of an Identical Library Pool.
Comparison demonstrates a highly conserved, non-random distribution of sequencing reads across individual multiplexed samples (n = 8), confirming that sample-to-sample coverage variance is a function of physical pool composition rather than automated sequencing stochasticity.

Despite the initial run suffering from severe under-sequencing, both runs exhibit a nearly identical relative read distribution profile across the individual multiplexed samples. This direct alignment indicates that the observed sample-to-sample read variation is determined by the physical pool composition rather than random distribution during the sequencing process. Because this pattern indicated that the skewed read distribution was introduced during pre-analytical pool preparation rather than during sequencing, the protocol was subsequently reverted to an equivolume pooling strategy to minimize quantification-induced variability.

A closer inspection of this pre-analytical skewing reveals a distinct trend based on sample origin: the FFPE tumor tissue samples consistently claimed a higher proportion of sequencing reads than the cfDNA samples. This pattern could be due to a mathematical limitation introduced by the standard molar conversation formula. Because the fluorometric Qubit assay only measures total mass concentration (ng/µL), converting these values to molarity (nM) using a fixed constant assumes that every sample shares an identical average fragment size.

While a fixed conversion constant is entirely reasonable for standard applications like PGT, where uniform cell inputs and a controlled lysis step during library preparation yield highly consistent fragment lengths, it can introduce inaccuracies when dealing with highly variable samples. FFPE tumor tissue DNA exhibits highly heterogeneous fragment distributions due to harsh chemical fixation and tissue degradation, which can result in shorter average fragment lengths than the formula assumes. In this scenario, the sample would contain a higher number of individual, sequenceable library fragments per nanogram of mass. Consequently, relying on a static conversion factor could have caused these FFPE samples to be under-diluted, leaving them disproportionately over-represented on the sequencing chip. This observation suggests

that relying on Qubit mass values alone may represent a “blind” normalization when applied to mixed cohorts. To optimize an equimolar approach for such datasets, incorporating a secondary fragment size analysis, such as a Bioanalyzer assay, could allow for calculating true molarity based on each sample’s unique fragment profile.

4.3 Ion Reporter Optimization

Following the two-phase bisection optimization workflow, TP values were established for the Initial and Optimized workflows. Within each phase, the selected TP was held constant across the evaluated resolutions to allow for a direct, controlled comparison of how bin size influences CNA calling.

4.3.1 Initial Workflow (10-Sample Baseline)

Using the baseline of 10 healthy controls from the Initial sub-cohort, the clinical baseline TP of -3.0 was incrementally adjusted to increase sensitivity. The bisection strategy settled on a final optimized TP value of -2.82, which achieved zero false positives across the validation controls at both the 2.0 Mb and 1.0 Mb resolutions. However, narrowing the bin size from 2.0 to 1.0 Mb expectedly increased the average MAPD value from 0.179 ± 0.033 to 0.236 ± 0.034 ($p = 1.88\times10^{-8}$, Welch’s t-test).

4.3.2 Optimized Workflow (Expanded 17-Sample Baseline)

Expanding the custom baseline to 17 samples by integrating higher-depth healthy controls successfully reduced background noise. This added stability allowed the workflow to tolerate a more sensitive parameter configuration. Using the same bisection protocol, the TP was lowered to a final optimized value of -2.56.

The expanded baseline significantly stabilized the sequencing data at the 1.0 Mb resolution, decreasing the average MAPD from 0.236 ± 0.034 to 0.182 ± 0.035 ($p = 5.22\times10^{-5}$, Welch’s t-test, Table 4.2). This statistically significant improvement highlights the impact of higher-depth healthy controls in suppressing baseline variance while maintaining zero false positives. Because this expanded baseline provided sufficient data stability and low background noise at the 1.0 Mb resolution, further evaluation at the larger 2.0 Mb resolution was deemed less critical for the primary objectives and was omitted from subsequent patient cohort analysis.

4.3.3 Exploratory 0.5 Mb Resolution

To explore the detection limits of the optimized workflow, the more sensitive TP from Phase 2 (-2.56) was applied to a high-resolution 0.5 Mb bin size using the expanded 17 sample baseline. Decreasing the bin size expectedly increased background variance, raising the average MAPD to 0.234 ± 0.021 compared to the 1.0 Mb resolution ($p = 9.44\times10^{-5}$, Welch’s t-test, Table 4.2).

While the cohort’s average MAPD of 0.234 remains below the standard 0.3 QC threshold for acceptable run stability, the increased background variance combined with the sensitive TP, introduced a single false-positive CNA call within the validation controls (Table 4.2). This

event occurred within the lower-depth validation sample (159,540 reads), whereas the higher-depth validation control (362,837 reads) maintained zero false positives. This lower-depth control exhibited an individual MAPD of 0.325, surpassing the 0.3 QC threshold and suggesting a localized reduction in data stability due to low read count. Based on these boundary observations, a minimum threshold of 200,000 reads was established for 0.5 Mb analysis to provide a security margin of 25% above the observed failure point.

Table 4.2: Summary of Ion Reporter Parameter Mapping, Background Noise, and Validation Specificity.

Workflow	**Bin Size (Mb)**	**Baseline Size (Samples)**	**TP**	**Avg. MAPD**	**False Positives in Validation Samples**
Initial	2.0	10	-2.82	0.179 ± 0.033	0
	1.0			0.236 ± 0.034	0
Optimized	1.0	17	-2.56	0.182 ± 0.035	0
	0.5			0.234 ± 0.021	1

4.3.4 Cross-Pipeline False Positive Assessment in Validation Controls

To contextualize the specificity of the optimized Ion Reporter workflow, the same two healthy validation controls were cross-analyzed using WisecondorX and ichorCNA. Across both alternative platforms, a persistent susceptibility to baseline artifacts was observed, particularly within the lower-depth control sample.

When evaluating WisecondorX, the lower-depth control uniformly generated exactly one false-positive call across all tested resolutions (2.0, 1.0, and 0.5 Mb) in both the TMAP- and Bowtie2-aligned workflows. Conversely, the higher-depth validation control remained entirely clean (zero false positives) across all target resolutions for both alignment workflows, except for a single artifact appearing in the Bowtie2-aligned pipeline at the 0.5 Mb resolution.

For ichorCNA, the baseline noise trends were more pronounced. At the 1.0 Mb resolution, ichorCNA yielded two false-positive calls in both the lower-depth and higher-depth validation controls. When evaluated at the 0.5 Mb resolution, these artifacts increased, resulting in eight false positives in the lower-depth control and one false positive in the higher-depth control.

Ultimately, these cross-pipeline findings suggest that baseline noise in healthy controls may represent a shared, depth-dependent challenge across low-pass analytical platforms rather than an issue isolated to specific software architecture. This highlights a practical constraint often encountered in low-pass sequencing workflows, indicating that minimizing false-positive calls at higher resolutions may be heavily dependent on expanding the normal reference baseline or ensuring higher sequencing coverage across the processed samples. The final cross-pipeline specificity performance across both validation controls is visually summarized in Figure 4.2.

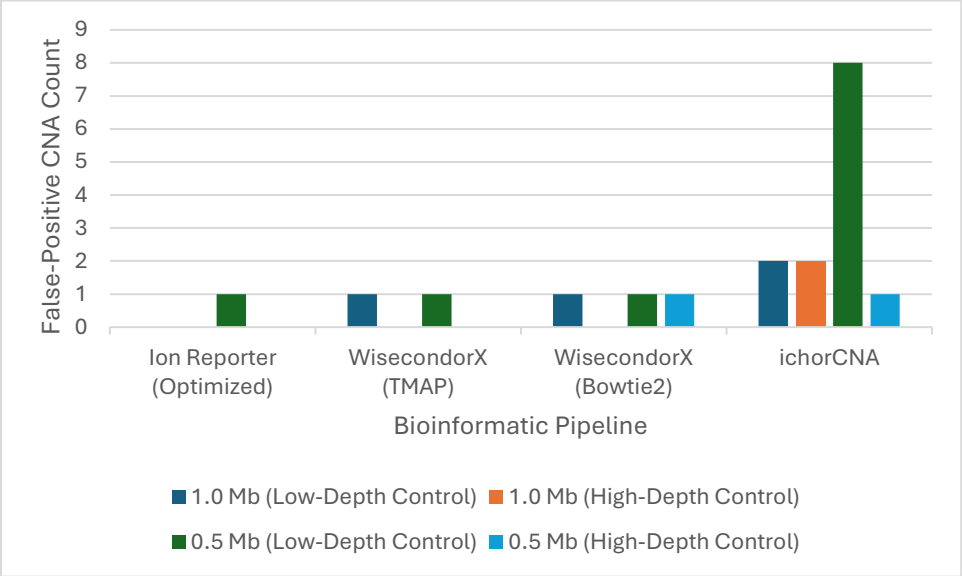

Figure 4.2: False-Positive CNA Call Rates Across Copy Number Pipelines and Validation Control Depths.
Categorical count of false-positive CNA calls mapped across Ion Reporter (Optimized), WisecondorX (compiled across TMAP and Bowtie2 alignments), and ichorCNA at 1.0 Mb and 0.5 Mb resolutions. Profiling is split between the lower-depth (159k reads) and higher-depth (362k reads) independent healthy validation controls. *Figure designed by the author.*

4.4 Benchmarking

To evaluate the analytical utility of the optimized Ion Reporter workflow, its performance was benchmarked against two open-source CNA detection pipelines: ichorCNA [7] and WisecondorX [28]. The evaluation spanned all tested resolutions (2.0 Mb, 1.0 Mb, and 0.5 Mb) across all three platforms. Additionally, for the WisecondorX pipeline, benchmarking included a comparative analysis of upstream alignment software. Performance was evaluated using BAM files generated via the native Ion Torrent aligner (TMAP) alongside BAM files independently re-mapped from raw FASTQ data using Bowtie2, following developer recommendations for optimal signal processing. Each workflow combination was assessed based on three primary statistical metrics: call accuracy (Macro F1-score), log-ratio signal correlation against the reference (Pearson *r*), and overall signal deviation (RMSE).

4.4.1 Comparative Performance Summary

To provide a clear, high-level overview of how the platforms compare at their absolute best, the single top-performing configuration from each pipeline's benchmarking matrix was selected for this summary (Table 4.3). The results for all the configurations are provided in Appendix C (Table C. 1and Table C. 2).

The results reveal that no single pipeline outperformed the others in every category. Instead, there is a divide between final call accuracy and raw signal quality. WisecondorX achieved the

highest definitive calling accuracy, yielding the top Macro F1-score (0.548) and the lowest overall error rate (RMSE = 0.417).

While the optimized Ion Reporter workflow secured the strongest log-ratio signal correlation to the reference standard (Pearson $r = 0.580$), this apparent advantage in signal fidelity is likely driven by a technical artifact. Unlike open-source pipelines that output continuous, unsegmented copy number signals across every individual genomic bin, Ion Reporter only reports a filtered list of discrete, called variant regions. To prevent severe mathematical skewing, non-reported genomic regions were assigned a log value of 0 for this benchmark. This strategy was employed because the vast majority of the genome is copy number neutral, and omitting these regions would restrict the correlation analysis to a few isolated variant coordinates, rendering the Pearson *r* metric incapable of evaluating whole-genome signal fidelity. While this zero-filling was methodologically necessary to run a fair genome-wide analysis, it artificially eliminated natural baseline sequencing noise, likely mathematically inflating the Ion Reporter's Pearson *r* score and lowering its RMSE.

Table 4.3: Benchmarking Summary of Top-Performing Pipeline Configurations.

Pipeline	**Best Configuration**	**Macro F1 (Call Accuracy)**	**Pearson *r* (Signal Correlation)**	**RMSE (Signal Deviation)**
WisecondorX	1.0 Mb, Bowtie2-Mapped, Default Settings	0.548	0.553	0.417
ichorCNA	0.5 Mb, Default Settings	0.498	0.398	0.464
Ion Reporter (Optimized Workflow)	0.5 Mb, Tuned TP & Buffer	0.441	0.580	0.451

The calling accuracy in Table 4.3 is evaluated using the macro F1-score. The F1-score itself is the harmonic mean of sensitivity and precision, serving as a robust metric that balances both the pipeline's ability to catch true variants (sensitivity) and its ability avoid false positives (precision). In a standard global F1 calculation (micro F1), all genomic variant calls are pooled into a single pool, which allows a high performance in a dominant variant category to mathematically mask poor performance in another. In contrast, a macro F1-score calculates an independent F1-score for genomic gains and losses separately, and then takes their unweighted arithmetic mean. This applies equal mathematical weight to both variant types, ensuring that a performance deficit in either category is explicitly exposed in the final summary metric rather than diluted by a global average.

4.4.2 Call Accuracy for Genomic Gains and Losses

Because the A549 positive control cell line is natively hypotriploid, its overall genomic baseline is elevated and the neutral background signal is shifted upward relative to standard reference

expectations. During benchmarking, this baseline shift led the copy number callers to misinterpret normal, non-variant regions of the A549 genome as an abundance of gains. This systematic bias resulted in a significant over-calling of copy number gains, yielding a high sensitivity (> 0.93) but a low precision (< 0.15) due to the accumulation of false positives (Table 4.4).

Table 4.4: Stratified Sensitivity and Precision for Copy Number Gains and Losses.

Pipeline	**Variant Type**	**Sensitivity**	**Precision**
Ion Reporter	Gains	0.959	0.133
	Losses	0.487	0.965
ichorCNA	Gains	0.934	0.140
	Losses	0.761	0.745
WisecondorX	Gains	0.934	0.147
	Losses	0.926	0.772

In contrast, metrics for copy number losses were significantly better balanced. For ichorCNA and WisecondorX, copy number losses were resolved with high sensitivity (0.761 and 0.926, respectively) and stable precision (~ 0.75 - 0.77). Conversely, the optimized Ion Reporter workflow exhibited a pronounced trade-off, demonstrating a lower sensitivity of 0.487 alongside a very high precision of 0.965. The sensitivity and precision scores for both genomic gains and losses are visually compared in Figure 4.3.

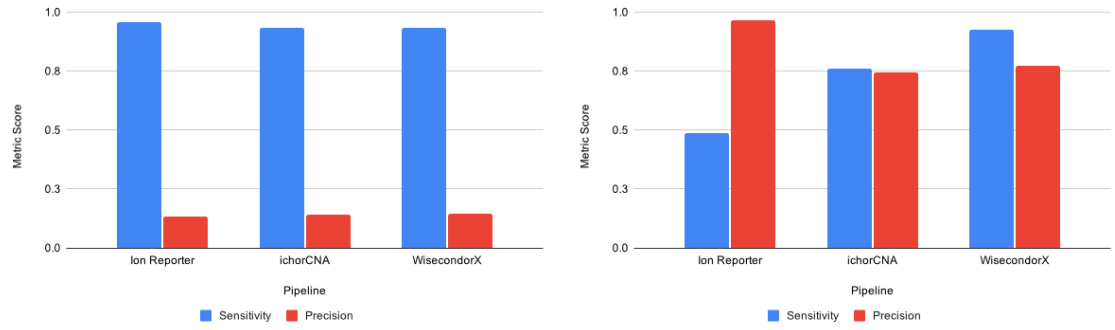

Figure 4.3: Stratified Sensitivity and Precision Performance for Copy Number Variants
Comparison of diagnostic metrics across the top-performing pipeline configurations evaluated against the A549 reference standard, split into the left panel (genomic gains) and right panel (genomic losses). *Figure designed by the author.*

As a result of these disparities, a standard global average would mask the large amount of false-positive gain calls due to the high sensitivity observed across both categories. By utilizing a

macro F1-score, equal mathematical weight is applied to both variant types. This ensures that the severe precision drop for gains caused by the triploid baseline directly and symmetrically penalizes the final summary performance metrics.

4.4.3 ichorCNA Tumor Fraction and Ploidy Estimation

Unlike Ion Reporter and WisecondorX, ichorCNA utilizes a probabilistic HMM to simultaneously estimate global sample attributes, including tumor fraction, baseline ploidy, and subclone fraction. This architecture generates multiple competing mathematical solutions ranked by log likelihood. While the tool provides global metrics for all generated solutions, it exclusively outputs read-depth ratios and copy number calls for the single top-ranked model.

Table 4.5 displays the top two parameter estimation solutions generated by the pipeline for the undiluted, hypotriploid A549 positive control.

Table 4.5: Competing ichorCNA Parameter Estimation Solutions for the A549 Positive Control.

Solution	**Tumor Fraction**	**Ploidy**	**Subclone Fraction**	**Log-Likelihood**
Top-Ranked	0.4127	2.28	0.901	704
Second-Ranked	0.8773	2.63	0.13	701

The marginal 3-point log-likelihood difference between the top-ranked solution (704) and second-ranked solution (701) highlights a critical limitation in ichorCNA's automated model selection framework, creating a mathematical tie between two mutually exclusive global biological profiles. Given that the undiluted positive control possesses a near 100% tumor fraction and a hypotriploid genome, the second-ranked solution (TFx: 0.8773, ploidy: 2.63) much more closely mirrors the true biological baseline. However, the pipeline automatically selected the inaccurate top-ranked solution as its primary output due to its slight log likelihood advantage.

Because both models explain the underlying data almost equally well, this ambiguity fundamentally compromises the entire output. In uncharacterized clinical patient samples, such algorithmic ambiguity introduces a severe diagnostic risk. Since the algorithm is unable to tell the models apart, users cannot reliably distinguish true genomic landscapes from mathematically equivalent artifacts, potentially introducing critical genomic misinterpretations.

This risk is compounded by an architectural constraint: ichorCNA only generates downstream data files, such as bin-level read-depth ratios, call lists, and plots, for the single top-ranked model. Accessing alternative solutions requires manual data extraction from backend *.RData* files paired with modifications to the primary execution script, or an iterative process of adjusting initialization parameters to manually force the desired model to the top [35].

Consequently, this requirement for manual intervention indicates a notable operational constraint for standardized screening applications. These findings suggest that for the pipeline to serve as a reliable tool in biomarker discovery, a more transparent or automated method for interacting with competing models would be highly beneficial. Without direct visibility into these alternative structures, the current architecture presents a risk of user oversight, suggesting it requires additional optimization before serving as a user-friendly foundation for identifying downstream genetic variations.

4.5 Copy Number Analysis of Clinical Patient Samples

Following workflow optimization, copy number profiling was executed across the quality-filtered clinical NSCLC dataset. As detailed in the cohort stratification flowchart (Figure 4.4), Out of the initial 24 patient cases, the 1.0 Mb workflow yielded 20 viable plasma cfDNA samples, including 19 fully matched tissue-plasma pairs. At the 0.5 Mb resolution, the cohort comprised 6 viable plasma samples, including 5 matching tissue-plasma pairs.

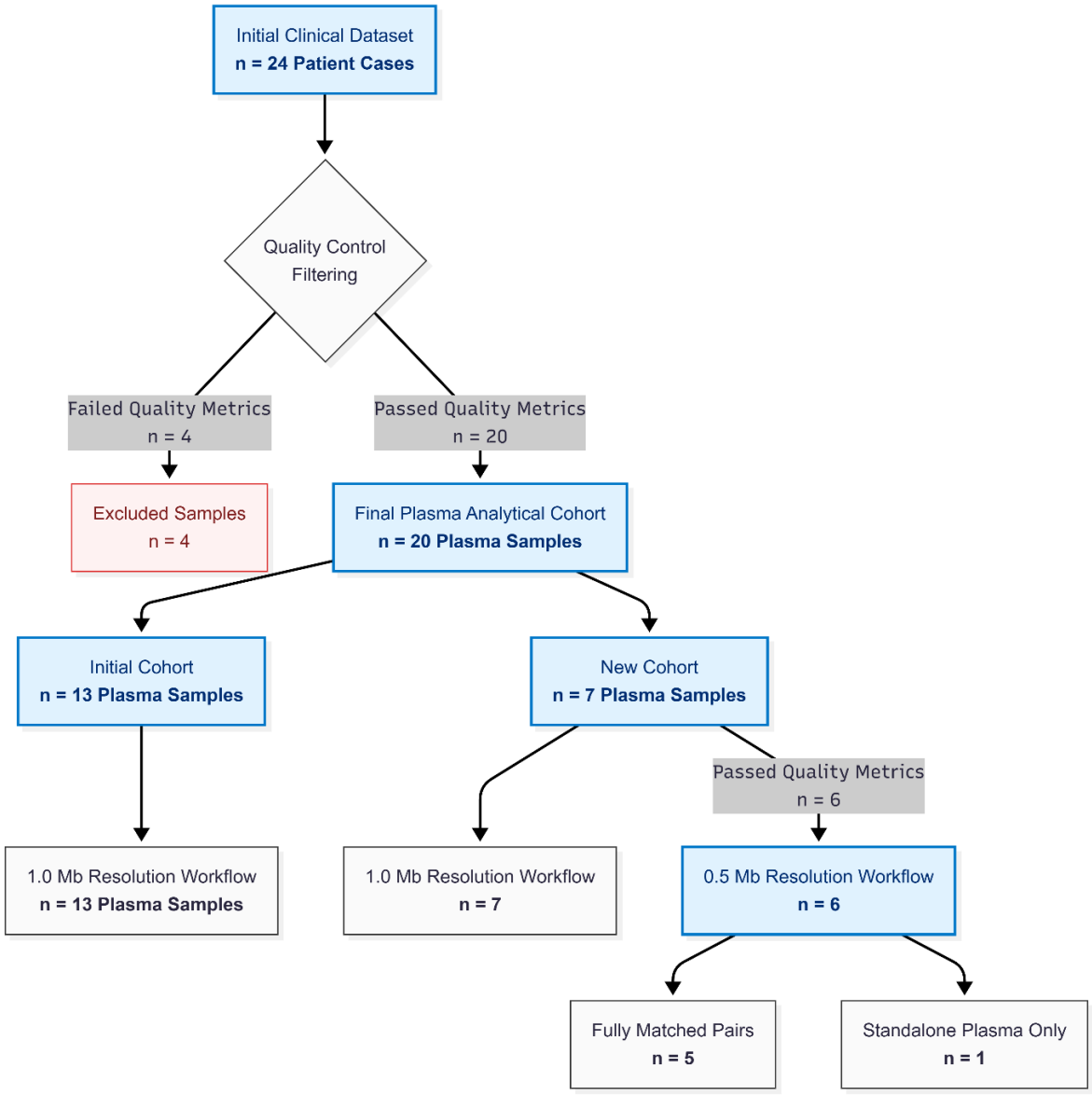

Figure 4.4: Flowchart of Clinical NSCLC Cohort Stratification and Sample Tracking across Processing Resolutions.
The diagram illustrates the initial filtering of the 24 patient cases down to the 20 analytical plasma samples, tracking their subsequent selection and allocation into the 1.0 Mb and 0.5 Mb copy number profiling workflows. *Flowchart designed by the author.*

The complete catalog of genomic plots for the new cohort is compiled in Appendix D (Figure D. 1 to Figure D. 7), featuring both tumor tissue and plasma cfDNA profiles for Ion Reporter alongside plasma-only profiles for the ichorCNA and WisecondorX. For the representative case studies highlighted in this section, copy number profile evaluations focused on two primary levels of comparison. First, the biological concordance between matched solid tissue biopsies and plasma cfDNA profiles generated within the Ion Reporter workflow. Second, the cross-pipeline technical concordance among the three parallel plasma pipelines evaluated at 0.5 Mb resolution.

4.5.1 Case Study 1: High-Concordance Profile (LCG29)

To evaluate whether ctDNA could serve as a reliable proxy for solid tumor tissue in identifying CNAs, copy number profiling was conducted across the 19 fully matched clinical patient pairs that successfully cleared global quality control filtering. Within this cohort, a single patient case (LCG29) possessed a sufficiently robust ctDNA signal to allow for a direct tissue-plasma concordance comparison.

As demonstrated in the Ion Reporter Optimized workflow (Figure 4.5), the plasma cfDNA copy number profile mirrors the genomic landscape identified in the solid tumor tissue, with several matching copy number calls. Notably, the y-axis scales vary considerably between the profiles. The plasma sample exhibits a lower signal amplitude compared to the solid tumor tissue, reflecting the dilution of tumor-derived DNA by healthy background cfDNA. Due to this compressed signal amplitude, certain CNAs fell below the strict detection thresholds enforced to eliminate false positives within the validation controls during optimization.

This high concordance between these independent sequencing runs serves as a preliminary proof of concept. It suggests that when a sample possesses an adequate tumor fraction, the optimized workflow is capable of capturing tumor-derived copy number signals from a standard blood draw, even at ultra-low sequencing depths.

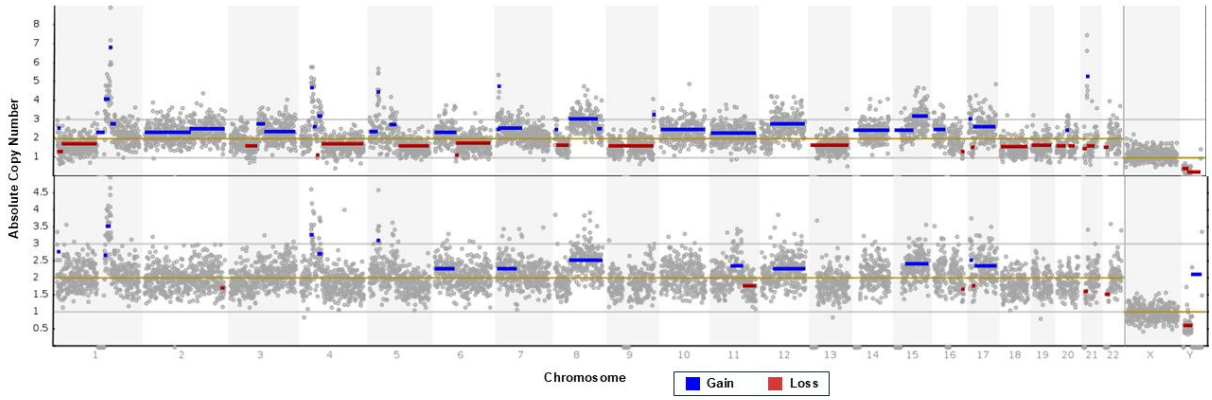

Figure 4.5: Copy Number Profiling Concordance between Matching Tumor Tissue (top panel) and cfDNA (bottom panel) for Patient LCG29.
The profiles show high concordance across the genome at a 0.5 Mb resolution. *Figure captured from the Ion Reporter built-in genome viewer interface using an ENPB of 1.8-2.2.*

To evaluate how these biological signals manifest across different bioinformatics architectures, the high-resolution 0.5 Mb plasma cfDNA profiles generated for patient LCG29 were evaluated across all three parallel pipeline workflows. Because all samples in the clinical cohort were processed through the same three workflows, the data from patient LCG29 can be directly compared across these independent pipelines. This comparison allows for an assessment of technical concordance within the liquid biopsy medium itself, testing whether variations in baseline filtering, normalization algorithms, and segmentation strategies impact the final genomic interpretation.

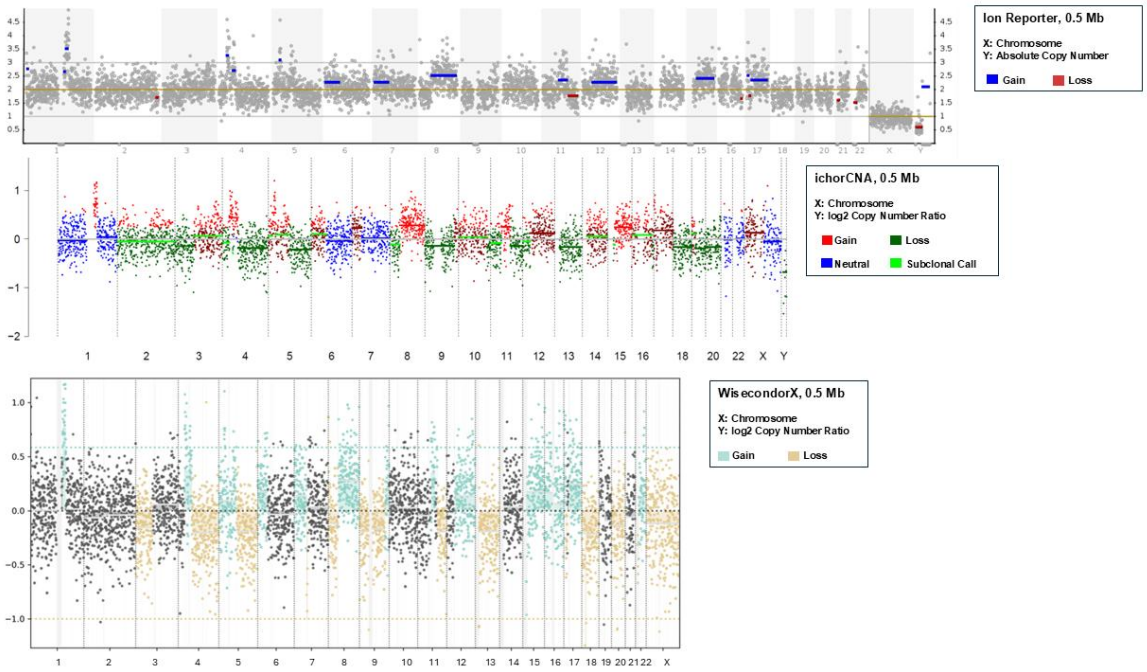

Figure 4.6: High-Resolution 0.5 Mb Copy Number Profiling across Independent ctDNA pipelines for Patient LCG29.
The panels display plasma cfDNA profiles generated by Ion Reporter (top), ichorCNA (middle), and WisecondorX (bottom). *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

As illustrated in Figure 4.6, the resulting plasma profiles show relatively consistent results across the workflows. The main structural patterns detected by Ion Reporter are generally tracked by both ichorCNA and WisecondorX. This technical agreement suggests that the core copy number calls are reasonably reproducible across different software environments when the biological signal is strong. However, because LCG29 represents an ideal, high-signal case for this cohort, these results may not reflect pipeline performance on the remaining samples where the ctDNA fraction is much lower.

4.5.2 Case Study 2: Low-Concordance Profile (LCG65)

Because the clinical cohort consists of real-world NSCLC patients, the fraction of tumor-derived DNA within the total cfDNA pool is highly variable. For a substantial portion of the analyzed plasma samples, copy number profiling revealed effectively neutral profiles with no detectable alterations. Crucially, this lack of copy number calls was not merely an artifact of conservative software calling thresholds failing to report variations. Instead, the raw data points on the graphs consistently formed a flat cloud of data points uniformly centered around the copy neutral baseline (CN 2), indicating a low or absent ctDNA fraction rather than a complete biological absence of genomic alterations.

While these low-signal profiles generally demonstrated complete genomic neutrality in the plasma, the call algorithm occasionally generated minor, low-amplitude copy number calls driven by subtle baseline fluctuations. Such a case is illustrated in patient case LCG65 (Figure 4.7), which shows a tissue-plasma pair at 0.5 Mb resolution. The solid tumor tissue profile shows a complex genomic landscape with distinct CNAs across multiple chromosomes. In contrast, the matching plasma sample displays an effectively flat distribution, failing to capture the genomic alterations seen in the tissue.

This stark divergence highlights a critical limitation when interpreting liquid biopsy data from a single sampling timepoint. When analyzing an isolated plasma sample snapshot, an effectively flat profile is fundamentally ambiguous: it cannot be reliably determined whether the patient's tumor lacks CNAs entirely, or if the ctDNA signal is simply too faint to pass the detection thresholds. While a multi-sample longitudinal design would allow one to track a treatment-induced decline of a previously visible tumor signal, a single snapshot provides no such baseline.

However, because this study utilizes matching solid tumor tissue, the underlying biological reality is known. The tissue profile confirms that complex CNAs are present within the patient's tumor DNA. Even though longitudinal biological evolution may have occurred between the solid tissue biopsy and the single plasma snapshot, tumor progression typically drives genomes toward greater complexity rather than reverting them to a balanced, copy-neutral state. Consequently, the flat plasma profile is highly unlikely to be attributed to a biological absence of CNAs in the tumor cells themselves. Instead, it likely represents an extreme biological dilution of ctDNA within the healthy background cfDNA pool, driven either by a highly successful therapeutic intervention that cleared the circulating tumor burden, or by an inherent low-shedding phenotype where the active tumor signal rests below the pipeline's technical limit of detection.

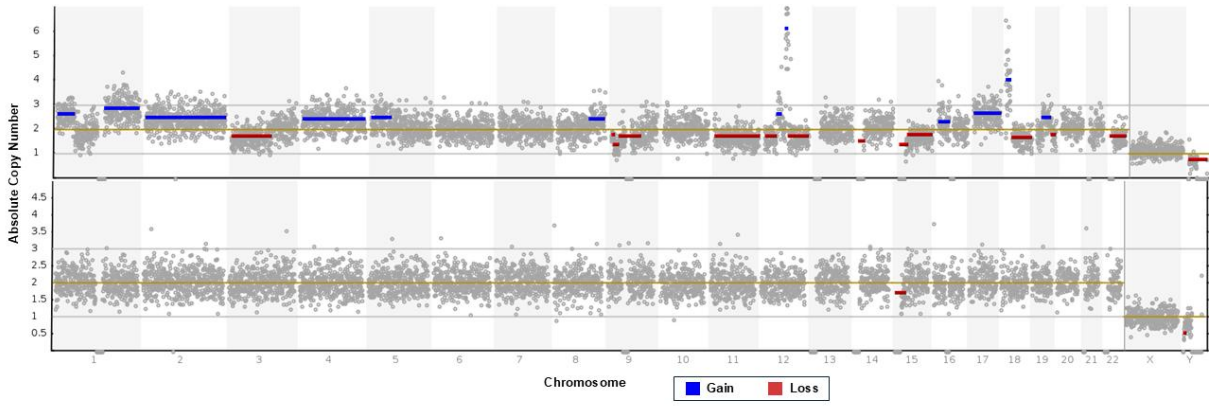

Figure 4.7: Copy Number Profiling Concordance between Matching Tumor Tissue (top panel) and cfDNA (bottom panel) for Patient LCG65.
The cfDNA profile shows an effectively flat distribution with borderline, low-amplitude software calls at a 0.5 Mb resolution. *Figure captured from the Ion Reporter built-in genome viewer interface using an ENPB of 1.8-2.2.*

Following the initial tissue-plasma concordance analysis, the cross-pipeline performance for patient LCG65 was evaluated using the high-resolution 0.5 Mb profiles across all three parallel workflows (Figure 4.8).

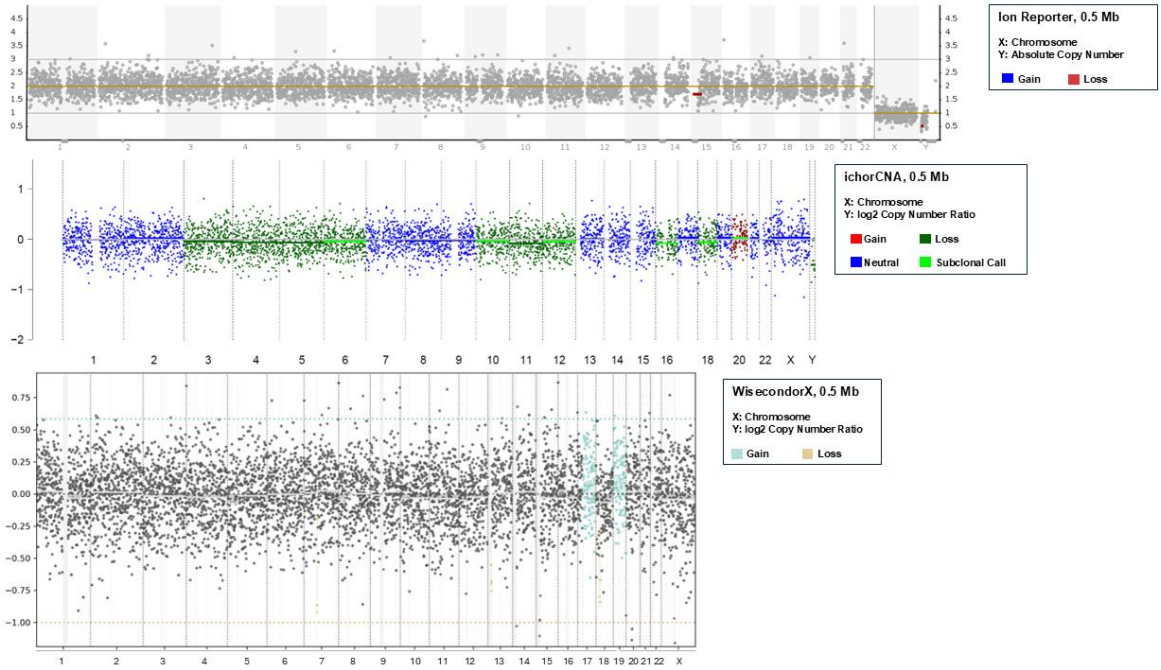

Figure 4.8: High-Resolution 0.5 Mb Copy Number Profiling across Independent ctDNA Pipelines for Patient LCG65.
The panels display plasma cfDNA profiles generated by Ion Reporter (top), ichorCNA (middle), and WisecondorX (bottom). *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

As illustrated in Figure 4.8, while the raw data points across all three pipelines consistently form a flat cloud indicative of a copy-neutral genome, the downstream segmentation and calling algorithms handle this baseline noise with varying degrees of stringency. Both Ion Reporter and WisecondorX demonstrate relatively conservative behavior: Ion Reporter generates only two borderline copy number losses, and WisecondorX identifies two borderline copy number gains.

In contrast, ichorCNA demonstrates a much higher sensitivity to baseline fluctuations in this low-shedder sample. While its raw data profile is similarly flat, the calling algorithm generates a high volume of CNAs, including ten unreliable calls, several of which are flagged as subclonal. This marked divergence underscores that in low-signal or zero-signal scenarios, cross-pipeline concordance drops significantly at the algorithmic level. While the raw inputs are uniformly flat, ichorCNA's mathematical modeling is prone to over-interpreting technical noise as subclonal events, whereas Ion Reporter and WisecondorX maintain a more stable, conservative baseline.

4.5.3 Case Study 3: Sex Chromosome Discrepancy (LCG43)

When evaluating the sex chromosome profiles of male patients within the New clinical cohort, a discrepancy between the solid tumor tissue and matching plasma cfDNA was observed in two cases, aligning with somatic Loss of Y (LOY) which is a recognized genetic event in NSCLC. This phenomenon is demonstrated in patient case LCG43:

- **Solid Tumor Tissue**: Exhibited a complete omission of the Y chromosome track on the genomic plot (Figure 4.9, left panel) and was officially assigned an automated "Female" gender label by the software.
- **Matching Plasma cfDNA**: Correctly received an automated "Male" gender label by the software, exhibiting a normal, copy-neutral Y chromosome baseline (Figure 4.9, right panel).

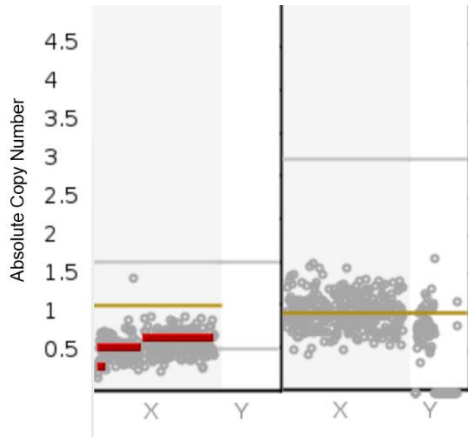

Figure 4.9: Discrepant Sex Chromosome Profiles Reflecting Somatic Loss of Y in Tumor Tissue for Patient LCG43.
The tumor tissue (left panel) shows complete omission of the Y chromosome track, resulting in an automated female classification. The plasma cfDNA (right panel) retains a normal male sex chromosome profile due to a biological absence of major tumor shedding. *Figure captured from the Ion Reporter built-in genome viewer interface using an ENPB of 1.8-2.2 at a 0.5 Mb resolution.*

This discrepancy is likely driven by differences in sample-specific tumor content. In solid tumor tissue with a high fraction of malignant cells, the somatic LOY dominates the signal. Conversely, in a liquid biopsy with low tumor fraction, the ctDNA fragments carrying this deletion are likely obscured by the background of healthy cfDNA that retain the Y chromosome. Ultimately, such discrepancies do not necessarily indicate a technical failure of the pipeline, but rather underscore that liquid and solid biopsies capture different biological sources and may not yield identical profiles in patients with low tumor-shedding characteristics.

To evaluate whether this normal plasma male baseline remains technically consistent across different bioinformatics architecture, the 0.5 Mb plasma cfDNA data for patient LCG43 was evaluated across all three parallel pipeline workflows.

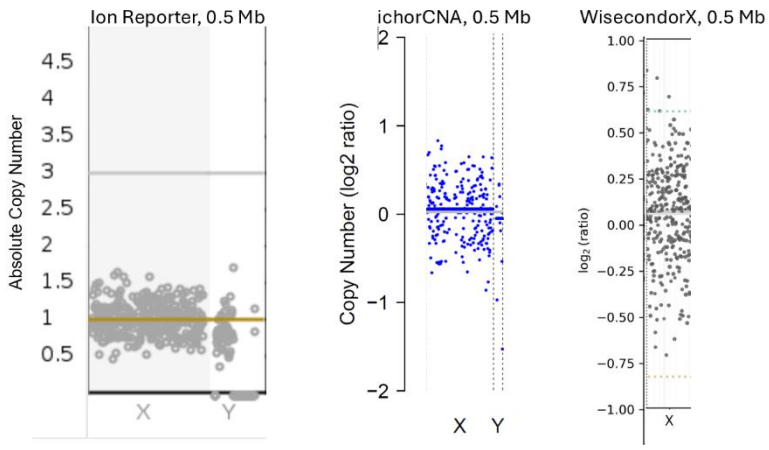

Figure 4.10: High-Resolution 0.5 Mb Copy Number Profiling across Independent ctDNA Pipelines for the Sex Chromosome Discrepancy Case (LCG43).
The panels display plasma cfDNA profiles generated by Ion Reporter (left), ichorCNA (middle), and WisecondorX (right). Note: In the WisecondorX panel, the chromosome Y track is completely omitted from the visual output due to reference baseline constraints, meaning only chromosome X is displayed. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

As illustrated in Figure 4.10, the cross-pipeline evaluation of the plasma sex chromosomes reveals an architectural divergence driven by different pipeline designs and baseline dependencies. While Ion Reporter utilizes a sex-matching calibration that correctly maps the male haploid state (CN 1) for both the X and Y tracks, and ichorCNA successfully plots both chromosomes on its log-ratio scale, WisecondorX omits the Y chromosome track entirely and displays only chromosome X.

This total omission of chromosome Y is a direct consequence of the reference baseline configuration. To comply with the strict operational constraints of Ion Reporter, an exclusively male healthy control cohort (n = 17) was used to construct the baseline panel. While ichorCNA successfully handles these sex chromosome tracks on its log-ratio baseline, WisecondorX's internal gender-prediction algorithm relies on a mixed reference panel containing both male and female samples to mathematically scale and calibrate its sex chromosome boundaries. Without female control genomes to compute a comparative baseline threshold, WisecondorX's sex-chromosome processing fails to clear its internal quality checks for the Y chromosome, causing the software to drop the track completely from its visual output. This outcome demonstrates that a reference panel optimized specifically for one tool's requirements can inadvertently limit how well another pipeline displays the data.

4.5.4 Classification of ichorCNA Tumor Fraction Estimates

To conclude the clinical cohort evaluation, a manual classification framwork was established to categorize the tumor fraction estimates generated by ichorCNA. This categorization was modeled after benchmarks defined by the software developers regarding detection limits and calling reliability. Samples were stratified into three distinct tiers: high confidence (≥ 10% TFx), representing the threshold required for reliable downstream CNA calling; low confidence (3-10% TFx), spanning the boundary zone below 10% where manual inspection is

typically recommended; and undetected (< 3% TFx or failed QC), aligning with the software's absolute theoretical limit of detection. Importantly, while these developer-defined parameters are established for sequencing coverages down to 0.1x, this analysis applied them to the current dataset's ultra-low coverage to test the operational boundary limits of the software.

Because the average sequencing depths differed substantially between the Initial Clinical Cohort (averaging 0.0053x) and the new cohort (averaging 0.01x), the cohorts were analyzed separately to minimize depth-associated detection biases. Furthermore, due to its low coverage, the Initial sub-cohort was only analyzed at 1.0 Mb resolution (n = 14 plasma, n = 12 tissue). In contrast, the increased depth of the New cohort permitted analysis at both 1.0 Mb (n = 7 plasma, n = 7 tissue) and 0.5 Mb (n = 6 plasma, n = 5 tissue) resolutions. The resulting distribution across both cohorts is presented in Table 4.6.

Table 4.6: Categorization of Initial and New Cohorts by Developer-Defined ichorCNA Thresholds.

Cohort & Avg. Coverage	**Resolution (Mb)**	**Sample Type**	**High Confidence (≥ 10% TFx)**	**Low Confidence (3-10% TFx)**	**Undetected (< 3% TFx or failed QC)**
Initial (~0.0053x)	1.0	cfDNA (n = 13)	4 (30.8%)	8 (61.5%)	1 (7.7%)
		Matching tumor DNA (n = 12)	11 (91.7%)	1 (8.3%)	
New (~0.01x)	1.0	cfDNA (n = 7)	1 (14.3%)	6 (85.7%)	
		Matching tumor DNA (n = 7)	7 (100%)		
	0.5	cfDNA (n = 6)	1 (16.7%)	5 (83.3%)	
		Matching tumor DNA (n = 5)	4 (80.0%)		1 (20.0%)

Note: Percentages calculated relative to the total number of samples within each specific sample type row. Row totals may deviate slightly from 100.0% due to rounding.

These stratified data indicate a sharp distinction in classification confidence based on sample origin. Solid tumor DNA samples consistently clustered within the high-confidence tier across both cohorts, reflecting the robust somatic signal present in direct tissue extractions. Conversely, the plasma cfDNA samples predominantly shifted toward the low-confidence and

undetected tiers, mirroring the highly variable tumor-shedding heterogeneity typical of real-world clinical cohorts.

Crucially, because these tiers classify the software's raw statistical estimations rather than a verified biological ground truth, they evaluate model stability rather than absolute diagnostic accuracy. Without independent orthogonal quantification, these outputs are unverified. Any individual estimation across the high-confidence, low-confidence, or undetected spans could be biologically correct or incorrect. For example, a "high confidence" assignment may still represent an algorithmic overestimation driven by technical artifacts, while a "low confidence" or "undetected" classification may fail to resolve a genuinely high biological tumor fraction due to the ultra-low coverage. Therefore, these manual tiers do not necessarily reflect the actual biological tumor fraction in the patient's blood. Instead, they simply demonstrate how stable or unstable the software's mathematical models become when processing ultra-low coverage data.

When pushing the New cohort to a higher 0.5 Mb resolution, a slight increase in undetected samples was observed for both tissue and plasma, which likely reflects the increased background noise introduced by narrower bin sizes (fewer reads per bin) under ultra-low coverage conditions.

This stratification, where direct tissue extractions cluster within the high-confidence tier while plasma cfDNA trends toward the low-confidence and undetected categories, presents a distribution pattern that broadly aligns with recent benchmarks described by Allsopp et al. (2023) [42]. In their study, an Ion ReproSeq workflow was paired with ichorCNA to perform sWGS on low-input plasma cfDNA. Even within an advanced cohort of 105 metastatic breast cancer patients, the authors reported a sCNA detection rate of 39% (41/105). While the article does not explicitly define the status of the remaining 61% of cases, it can be inferred that these samples yielded copy-neutral or sub-threshold profiles where any ctDNA signal remained below the software's analytical detection limits (3% TFx). For the 39% of patients with detectable signals in the Allsopp et al. cohort, ichorCNA resolved a median tumor fraction of 34%, spanning a range from 17% to 58%. This indicates that their positive detections occurred in distinct, high-shedding environments above the software's minimum limits. By contrast, the workflow in this study detected tumor content in 95% (19/20) of plasma samples. However, a notable number of these detected cases (14/19) belong to the low-confidence boundary zone (3-10% TFx). For the total clinical cohort, evaluation of the 1.0 Mb plasma cfDNA data revealed that ichorCNA resolved a median tumor fraction of 5% among detected samples, with values ranging from 3% to 27%.

This very low median highlights that within the current dataset, the detected tumor fraction estimates remain heavily skewed toward the absolute analytical floor. Because these values lack orthogonal verification, it remains unclear whether this distribution reflects true low-shedding clinical profiles or merely forcing a minimal non-zero mathematical fit due to a lack of data points. This analytical limitation is likely driven by the ultra-low sequencing depths evaluated in this study, especially given that ichorCNA was originally optimized for coverages down to 0.1x. While Allsopp et al. successfully operated at a 0.01x depth [42] that matches this study's New cohort, their pipeline benefited from a robust baseline reference consisting of 50 healthy

controls. By contrast, the average coverage of the plasma samples included in this comparison was 0.0065x, and the reference panel was only consisting of 17 healthy controls.

Ultimately, while these classification patterns demonstrate that ichorCNA can extract parameter trends beneath its recommended coverage specifications, these boundary findings represent preliminary indicators. Further study using scaled cohorts and independent validation methods would be required to evaluate the absolute accuracy of these algorithmic estimations.

4.6 Methodological Considerations in Pipeline Selection

The structural and algorithmic differences between ichorCNA and WisecondorX heavily influenced the analytical boundaries of this study, using the proprietary Ion Reporter workflow as the baseline clinical standard. While all three pipelines are highly dependent on user-defined parameter optimization, meaning their stringency and sensitivity can be shifted via tuning, they operate under fundamentally different architectural baselines. Ion Reporter served as a stable empirical control representing a closed, commercial software platform with pre-integrated analytical configurations.

In contrast, the open-source frameworks are bound by distinct developer-defined constraints. While ichorCNA provides a highly comprehensive set of global genomic metrics, its strict operational reliance on higher tumor fractions (> 10%) for automated calling led to model overfitting when evaluating low tumor-fraction cfDNA samples. Conversely, by leveraging an algorithmic architecture optimized for the lower biological baselines native to NIPT, WisecondorX demonstrated greater robustness at lower fractions. These differences suggest that while ichorCNA remains a well-established framework for advanced profiling of high tumor content samples, workflows targeting early-stage or low-fraction liquid biopsies benefit significantly from NIPT-derived calling architectures.

The conservative copy number calling behavior observed within the Ion Reporter workflow is heavily dictated by its proprietary architecture, which forces specificity to be evaluated at a sample level rather than a genomic bin level. Because the optimization process relied on only two validation controls, specificity can only be calculated as 50% or 100%. Importantly, even if the underlying read-depth ratios for each bin were available, accepting a baseline like 5% false-positive bins in healthy controls would not be comparable or applicable to false-positive calls within clinical patient samples. While background noise in controls presents as isolated, random bin fluctuations, the calling algorithm requires a consecutive sequence of several altered bins to call an alteration in a patient sample.

4.7 Study Limitations and Future Perspectives

While the benchmarking framework successfully highlighted the operational boundaries of all three evaluated pipelines, several technical limitations inherent to the study design should be considered to properly contextualize the findings.

The primary technical constraint encountered in this study was the ultra-low sequencing depth, which introduced substantial data sparsity into the workflow. At this low coverage, the signal-

to-noise ratio is severely reduced, directly limiting copy number calling accuracy. Consequently, the pipelines frequently lack the necessary genomic resolution to confidently distinguish true somatic CNAs from background sequencing noise, a challenge that is compounded in clinical samples presenting with low tumor fractions.

In addition to sequencing depth constraints, the constrained size of the reference baseline cohort restricted the precision of background noise correction algorithms across all evaluated tools. Because CNA calling software relies on reference panels to identify and mask systematic biases, utilizing a limited baseline cohort diminishes the software's efficacy in filtering out systematic, run-specific sequencing artifacts. As established in the methodology, the exact same cohort of healthy controls was utilized to build the baselines for all three pipelines to ensure an identical comparison. Consequently, the background noise that could not be adequately smoothed out by this restricted reference pool uniformly affected the analytical sensitivity across the entire benchmarking matrix, impacting the proprietary workflow alongside the open-source tools. This shared vulnerability highlights the need to expand the cohort of healthy samples to enhance noise-correction precision. Furthermore, expanding the number of independent healthy validation controls is also important to confidently fine-tune parameters, allowing for a controlled reduction in specificity to ultimately increase clinical sensitivity and avoid overfitting.

Finally, the total absence of technical replicates within the dataset limited the ability to mathematically average out or mitigate stochastic, run-to-run sequencing noise. Because stochastic variance is an inherent characteristic of liquid biopsy sequencing workflows across both proprietary and open-source platforms, the reliance on single-run evaluations introduces an element of analytical volatility to the benchmarking results.

Ultimately, addressing the technical bottlenecks identified in this study is crucial, as constraints such as a limited reference baseline and the lack of technical replicates directly impact the resolution and reliability of the analytical outputs. While resolving these constraints is ideally required to minimize background noise, doing so in practice often depends on institutional resources and budget availability. Overcoming these practical boundaries provides two distinct pathways for future work. From a technical validation standpoint, future studies could utilize a controlled dilution series by titrating a positive control, such as the A549 cell line, with healthy donor DNA. This setup would allow researchers to empirically map the precise detection limits of each pipeline across known tumor fractions. Alternatively, if the current pipeline resolution is deemed operationally sufficient, future perspectives could shift toward evaluating the biological significance of the called CNAs. Correlating these analytical outputs across all evaluated pipelines with clinical outcomes would be essential to assess their real-world diagnostic and predictive utility, given that a workflow optimized for a specific clinical application may not inherently translate to reliable oncology liquid biopsy profiling.

5 Conclusion

This study advanced the laboratory sequencing protocol and evaluated downstream bioinformatic frameworks for CNA analysis under ultra-low coverage conditions. By anchoring the analysis at an extreme sequencing depth of approximately 0.01x, this work outlines the baseline requirements for sample preparation and algorithmic performance when operating at the absolute lower boundaries of sWGS.

The experimental phase identified an optimal chip loading concentration for the sequencing platform, which stabilized data usability and sequencing coverage within the strict technical constraints of the wet lab. By mitigating the risks of chip underloading and overloading, this protocol adjustment prevented poor data quality and stabilized the overall data yield, providing a consistent raw data foundation for low-coverage genomic profiling. Within this established laboratory framework, the study demonstrated that a clinical Preimplantation Genetic Testing (PGT) workflow could be adapted for liquid biopsy profiling, achieving a detection resolution of 0.5 Mb in the highest-coverage samples evaluated.

Furthermore, the bioinformatic benchmarking framework highlighted the distinct architectural trade-offs that persist across pipelines at this extreme depth. Reflecting its intentional optimization to eliminate false-positive calls in the validation controls, the Ion Reporter workflow operated on the conservative side, proving highly reliable at minimizing false positives. In contrast, WisecondorX achieved the highest overall call accuracy within the positive control benchmarking matrix, while ichorCNA offered the greatest sensitivity in real-world clinical samples at the cost of increased vulnerability to background sequencing noise. Ultimately, these integrated findings demonstrate that while 0.01x coverage sits well below standard sWGS thresholds, it remains a viable framework for capturing true tumor-derived copy number signals, providing a foundational benchmark for future technical optimization and low-cost non-invasive genomic monitoring.

6 References

[1] N. Giustini and L. Bazhenova, "Recognizing Prognostic and Predictive Biomarkers in the Treatment of Non-Small Cell Lung Cancer (NSCLC) with Immune Checkpoint Inhibitors (ICIs)," *LCTT*, vol. Volume 12, pp. 21–34, Mar. 2021, doi: 10.2147/LCTT.S235102.

[2] A. Hallqvist, A. Rohlin, and S. Raghavan, "Immune checkpoint blockade and biomarkers of clinical response in non–small cell lung cancer," *Scand J Immunol*, vol. 92, no. 6, p. e12980, Dec. 2020, doi: 10.1111/sji.12980.

[3] L. Wang, Y. Hu, S. Wang, J. Shen, and X. Wang, "Biomarkers of immunotherapy in non-small cell lung cancer (Review)," *Oncol Lett*, vol. 20, no. 5, pp. 1–1, Aug. 2020, doi: 10.3892/ol.2020.11999.

[4] E. A. Eklund *et al.*, "Comprehensive genetic variant analysis reveals combination of KRAS and LRP1B as a predictive biomarker of response to immunotherapy in patients with non-small cell lung cancer," *J Exp Clin Cancer Res*, vol. 44, no. 1, p. 75, Feb. 2025, doi: 10.1186/s13046-025-03342-6.

[5] G. Rossi *et al.*, "Precision Medicine for NSCLC in the Era of Immunotherapy: New Biomarkers to Select the Most Suitable Treatment or the Most Suitable Patient," *Cancers*, vol. 12, no. 5, p. 1125, Apr. 2020, doi: 10.3390/cancers12051125.

[6] D. Chakravarty and D. B. Solit, "Clinical cancer genomic profiling," *Nat Rev Genet*, vol. 22, no. 8, pp. 483–501, Aug. 2021, doi: 10.1038/s41576-021-00338-8.

[7] V. A. Adalsteinsson *et al.*, "Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors," *Nat Commun*, vol. 8, no. 1, p. 1324, Nov. 2017, doi: 10.1038/s41467-017-00965-y.

[8] E. Lakatos, H. Hockings, M. Mossner, W. Huang, M. Lockley, and T. A. Graham, "LiquidCNA: Tracking subclonal evolution from longitudinal liquid biopsies using somatic copy number alterations," *iScience*, vol. 24, no. 8, p. 102889, Aug. 2021, doi: 10.1016/j.isci.2021.102889.

[9] B. Miles and P. Tadi, "Genetics, Somatic Mutation," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2026. Accessed: May 26, 2026. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557896/

[10] "Definition of driver mutation - NCI Dictionary of Cancer Terms - NCI." Accessed: May 26, 2026. [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/driver-mutation

[11] V. Vashisht, A. Vashisht, A. K. Mondal, J. Woodall, and R. Kolhe, "From Genomic Exploration to Personalized Treatment: Next-Generation Sequencing in Oncology," *CIMB*, vol. 46, no. 11, pp. 12527–12549, Nov. 2024, doi: 10.3390/cimb46110744.

[12] A. redaktör: R. Webbredaktion, "Nationell kvalitetsrapport Lungcancer för 2024." Accessed: May 26, 2026. [Online]. Available: https://cancercentrum.se/utvecklingsarbeteutbildning/statistikrapporter/rapporter/rapporter/nationellkvalitetsrapportlungcancerfor2024.10316.html

[13] "Immune Checkpoint Inhibitors - NCI." Accessed: May 26, 2026. [Online]. Available: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors

[14] C. for D. E. and Research, "FDA approves pembrolizumab for adults and children with TMB-H solid tumors," *FDA*, Sep. 2024, Accessed: May 26, 2026. [Online]. Available: https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors

[15] R. Esposito Abate *et al.*, "Harmonization of tumor mutation burden testing with comprehensive genomic profiling assays: an IQN Path initiative," *J Immunother Cancer*, vol. 12, no. 2, p. e007800, Feb. 2024, doi: 10.1136/jitc-2023-007800.

[16] M. J. Friedrich, "Going With the Flow: The Promise and Challenge of Liquid Biopsies," *JAMA*, vol. 318, no. 12, pp. 1095–1097, Sep. 2017, doi: 10.1001/jama.2017.10203.

[17] S. O. Hasenleithner and M. R. Speicher, "A clinician's handbook for using ctDNA throughout the patient journey," *Mol Cancer*, vol. 21, no. 1, p. 81, Mar. 2022, doi: 10.1186/s12943-022-01551-7.

[18] C. Rolfo *et al.*, "Liquid Biopsy for Advanced NSCLC: A Consensus Statement From the International Association for the Study of Lung Cancer," *Journal of Thoracic Oncology*, vol. 16, no. 10, pp. 1647–1662, Oct. 2021, doi: 10.1016/j.jtho.2021.06.017.

[19] F. Mouliere, "A hitchhiker's guide to cell-free DNA biology," *Neuro-Oncology Advances*, vol. 4, no. Supplement_2, pp. ii6–ii14, Nov. 2022, doi: 10.1093/noajnl/vdac066.

[20] R. Fu, H. A. Su, Y. Zhao, Y. Tian, H. Chen, and D. Lu, "Dissecting cell-free DNA fragmentation variation in tumors using cell line-derived xenograft mouse," *PLoS One*, vol. 20, no. 7, p. e0327483, Jul. 2025, doi: 10.1371/journal.pone.0327483.

[21] E. Heitzer, L. Auinger, and M. R. Speicher, "Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living," *Trends in Molecular Medicine*, vol. 26, no. 5, pp. 519–528, May 2020, doi: 10.1016/j.molmed.2020.01.012.

[22] Pcauchy, *Français : Illustration de la digestion par MNase. La MNase (bleu) est activée par addition de Ca2+, et digère l'ADN linker. Elle est stoppée par l'addition d'EDTA ou d'EGTA. Ensuite, la digestion des nucléosomes (suite à digestion de l'ARN) est effectuée par la protéinase K, résultant en des brins d'ADN correspondant majoritairement à des mononucléosomes.* 2016. Accessed: May 26, 2026. [Online]. Available: https://commons.wikimedia.org/wiki/File:Digestion_MNase.jpg

[23] L. Liu *et al.*, "Comparison of Next-Generation Sequencing Systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, pp. 1–11, 2012, doi: 10.1155/2012/251364.

[24] J. S. Mandlik, A. S. Patil, and S. Singh, "Next-Generation Sequencing (NGS): Platforms and Applications," *Journal of Pharmacy and Bioallied Sciences*, vol. 16, no. Suppl 1, pp. S41–S45, Feb. 2024, doi: 10.4103/jpbs.jpbs_838_23.

[25] S. Morganti *et al.*, "Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life," *Critical Reviews in Oncology/Hematology*, vol. 133, pp. 171–182, Jan. 2019, doi: 10.1016/j.critrevonc.2018.11.008.

[26] "Semiconductor Sequencing Technology - SE." Accessed: May 26, 2026. [Online]. Available: https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

[27] P. Hupé, *English: From second to fourth-generation sequencing, illustration on TAGGCT template*. 2012. Accessed: May 26, 2026. [Online]. Available: https://commons.wikimedia.org/wiki/File:From_second_to_fourth-generation_sequencing,_illustration_on_TAGGCT_template.svg

[28] L. Raman, A. Dheedene, M. De Smet, J. Van Dorpe, and B. Menten, "WisecondorX: improved copy number detection for routine shallow whole-genome sequencing," *Nucleic Acids Research*, vol. 47, no. 4, pp. 1605–1614, Feb. 2019, doi: 10.1093/nar/gky1263.

[29] J. E. Han and H. J. Cho, "Exploring the prognostic value of ultra-low-pass whole-genome sequencing of circulating tumor DNA in hepatocellular carcinoma," *Clin Mol Hepatol*, vol. 30, no. 2, pp. 160–163, Apr. 2024, doi: 10.3350/cmh.2024.0141.

[30] W. Kuśmirek, "Different Strategies for Counting the Depth of Coverage in Copy Number Variation Calling Tools," *Bioinform Biol Insights*, vol. 16, p. 11779322221115534, Jan. 2022, doi: 10.1177/11779322221115534.

[31] FREX Consortium *et al.*, "Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation," *Eur J Hum Genet*, vol. 29, no. 1, pp. 99–109, Jan. 2021, doi: 10.1038/s41431-020-0672-2.

[32] B.-J. Yoon, "Hidden Markov Models and their Applications in Biological Sequence Analysis," *CG*, vol. 10, no. 6, pp. 402–415, Sep. 2009, doi: 10.2174/138920209789177575.

[33] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, Oct. 2004, doi: 10.1093/biostatistics/kxh008.

[34] "Introduction to Ion Reporter™ Software." Accessed: May 26, 2026. [Online]. Available: https://ionreporter.thermofisher.com/ir/help/GUID-4A207E04-2113-4633-968D-4B7A65A1D64A.html

[35] *broadinstitute/ichorCNA*. (May 21, 2026). R. Broad Institute. Accessed: May 26, 2026. [Online]. Available: https://github.com/broadinstitute/ichorCNA

[36] *CenterForMedicalGeneticsGhent/WisecondorX*. (Apr. 14, 2026). Python. Center For Medical Genetics Ghent. Accessed: May 26, 2026. [Online]. Available: https://github.com/CenterForMedicalGeneticsGhent/WisecondorX

[37] "DepMap Cell Line Summary." Accessed: May 26, 2026. [Online]. Available: https://depmap.org/portal/cell_line/ACH-000681?tab=overview

[38] R. W. K. Chiu, L. L. M. Poon, T. K. Lau, T. N. Leung, E. M. C. Wong, and Y. M. D. Lo, "Effects of Blood-Processing Protocols on Fetal and Total DNA Quantification in Maternal Plasma," *Clinical Chemistry*, vol. 47, no. 9, pp. 1607–1613, Sep. 2001, doi: 10.1093/clinchem/47.9.1607.

[39] QIAGEN, *EZ1&2™ ccfDNA Kit Handbook*. Hilden, Germany: QIAGEN, 2026. Accessed: Feb. 12, 2026. [Online]. Available: https://www.qiagen.com/us/resources/kithandbook/hb-2915-003-hb-ez1and2-ccfdna-0823-ww

[40] Thermo Fisher Scientific, *Qubit™ 1X dsDNA HS Assay Kits USER GUIDE*. in Pub. No. MAN0017455, Rev. C.0. Waltham, MA, USA: hermo Fisher Scientific, 2020. Accessed: Feb. 12, 2026. [Online]. Available: https://documents.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017455_Qubit_1X_dsDNA_HS_Assay_Kit_UG.pdf

[41] Thermo Fisher Scientific, *Ion ReproSeq™ PGS Kits – Ion S5™/Ion GeneStudio™ S5 Systems USER GUIDE*. in Pub. No. MAN0016712, Rev. K. Waltham, MA, USA: Thermo Fisher Scientific, 2025. Accessed: Feb. 12, 2026. [Online]. Available: https://documents.thermofisher.com/TFS-Assets/LSG/manuals/MAN0016712_IonReproSeqPGS_S5_UG.pdf

[42] R. C. Allsopp *et al.*, "A Rapid, Shallow Whole Genome Sequencing Workflow Applicable to Limiting Amounts of Cell-Free DNA," *Clinical Chemistry*, vol. 69, no. 5, pp. 510–518, Apr. 2023, doi: 10.1093/clinchem/hvac220.

Appendix A Benchmarking R Scripts

This appendix presents the custom R scripts used to process, normalize, and evaluate the copy number datasets generated in this study.

```
# ============================================================================
# Thesis Appendix A.1: Signal Correlation and Structural Deviation Analysis
# Purpose: This script normalizes continuous log2 copy number ratio data from
#          disparate platforms into uniform 1.0 Mb genomic bins, executing
#          Pearson correlation (r) and Root Mean Square Error (RMSE)
#          benchmarking against a ground-truth reference (DepMap).
# Dependencies: readxl, tidyverse, GenomicRanges
# ============================================================================

library(readxl)
library(tidyverse)
library(GenomicRanges)

# --- 1. SETUP Environment & Target Path Directories ---
# Configure directory paths for the dataset input files and reference controls
setwd("path/to/project_folder")
truth_file  <- "DepMap A549.xlsx"
chrom_file  <- "GRCh38.p2.xlsx"
input_folder <- "pearson"

# --- 2. Build Reference Coordinates: Chromosome Sizes & Genomic Grid Alignment ---
# Import standard GRCh38 chromosome boundaries and strip string formatting
chrom_sizes <- read_excel(chrom_file, col_names = FALSE) %>%
  filter(...2 != "length") %>%
  select(chr = 1, size = 2) %>%
  mutate(size = as.numeric(size),
         chr = gsub("(?i)chr", "", as.character(chr))) %>%
  filter(!is.na(size))

# Segment the reference genome into contiguous 1.0 Mb (1e6 bp) resolution bins
master_grid <- chrom_sizes %>%
  group_by(chr, size) %>%
  reframe(start = seq(1, size, by = 1e6)) %>%
  mutate(end = pmin(start + 1e6 - 1, size)) %>%
  ungroup()

# Cast the data frame to a GenomicRanges object to facilitate spatial intersection matching
grid_gr <- GRanges(seqnames = master_grid$chr, ranges = IRanges(master_grid$start,
master_grid$end))

# --- 3. BP-Weighted Spatial Normalization Function ---
# Maps fragmented platform segments to the standardized 1.0 Mb grid, weighting data by
base-pair overlap
map_to_grid_weighted <- function(df_clean, master_grid, grid_gr) {
  out_vals <- rep(0, nrow(master_grid))
  if(nrow(df_clean) == 0) return(out_vals)

  # Enforce dynamic range capping ([-3, 3]) to eliminate extreme amplification or deletion
artifacts
  d_sub <- df_clean %>%
    transmute(c = chr, s = start, e = end, v = pmax(pmin(value, 3), -3))

  # Compute overlap coordinates between standard grid bins and platform-specific segments
  seg_gr <- GRanges(seqnames = d_sub$c, ranges = IRanges(d_sub$s, d_sub$e))
  hits <- findOverlaps(grid_gr, seg_gr)

  if(length(hits) > 0) {
    q_idx <- queryHits(hits)
```

```
    s_idx <- subjectHits(hits)

    # Calculate exact base-pair overlap width for accurate proportional averaging
    ov_start <- pmax(master_grid$start[q_idx], d_sub$s[s_idx])
    ov_end   <- pmin(master_grid$end[q_idx], d_sub$e[s_idx])
    ov_bp    <- ov_end - ov_start + 1

    # Aggregate values via weighted means for bins intersecting multiple segmented blocks
    df_hit <- data.frame(bin = q_idx, val_bp = ov_bp * d_sub$v[s_idx], bp = ov_bp) %>%
      group_by(bin) %>%
      summarise(v_sum = sum(val_bp, na.rm = TRUE), w_sum = sum(bp, na.rm = TRUE), .groups =
"drop")

    out_vals[df_hit$bin] <- df_hit$v_sum / df_hit$w_sum
  }
  return(out_vals)
}

# --- 4. Uniform Excel Data Extraction Utility ---
# Standardizes arbitrary spreadsheet layout formats into explicit data structures
get_clean_data <- function(f_path) {
  df <- read_excel(f_path, col_names = FALSE, .name_repair = "minimal")

  # Standard column indexes mapped to source outputs: Col 1=Chr, 2=Start, 3=End, 7=Log2
ratio
  c_idx <- 1; s_idx <- 2; e_idx <- 3; v_idx <- 7

  # Programmatically skip file headers if descriptive text strings are encountered
  if (!is.numeric(df[[1, s_idx]])) { df <- df[-1, ] }

  # Consolidate spatial variables and clean data formatting rules
  res <- data.frame(
    chr   = gsub("(?i)chr", "", as.character(df[[c_idx]])),
    start = as.numeric(df[[s_idx]]),
    end   = as.numeric(df[[e_idx]]),
    value = as.numeric(df[[v_idx]]),
    stringsAsFactors = FALSE
  ) %>% filter(!is.na(value), !is.na(start), !is.na(end))

  return(res)
}

# --- 5. Iterative Pipeline Execution & Statistical Benchmarking ---
# Process baseline ground-truth matrix
cat("Processing Truth...\n")
truth_data <- get_clean_data(truth_file)
truth_binned <- map_to_grid_weighted(truth_data, master_grid, grid_gr)

# Discover and iterate across all configuration targets inside the designated folder
file_list <- list.files(path = input_folder, pattern = "\\.xlsx$", full.names = TRUE)

results_pearson <- map_df(file_list, function(f_path) {
  fname <- basename(f_path)
  cat("Correlating:", fname, "\n")

  tryCatch({
    # Load and execute grid-normalization for the target test configuration
    test_data <- get_clean_data(f_path)
    test_binned <- map_to_grid_weighted(test_data, master_grid, grid_gr)

    comp_df <- data.frame(ref = truth_binned, test = test_binned) %>%
      filter(is.finite(ref), is.finite(test))

    # Pearson Correlation Check: Handle invariant profiles to avoid divide-by-zero errors
    if(sd(comp_df$test, na.rm = TRUE) == 0) {
      cor_val <- 0
```

```
    } else {
      cor_val <- cor(comp_df$ref, comp_df$test, method = "pearson")
    }

    # Calculate Root Mean Square Error to evaluate quantitative copy-number amplitude
deviation
    rmse_val <- sqrt(mean((comp_df$ref - comp_df$test)^2))

    tibble(File = fname, Pearson_r = round(cor_val, 4), RMSE = round(rmse_val, 4))
  }, error = function(e) {
    message("Error in ", fname, ": ", e$message)
    return(NULL)
  })
})

# --- 6. DISPLAY COMPARATIVE BENCHMARK MATRIX ---
# Output results ordered descending by linear signal correlation to evaluate performance
tiers
print(results_pearson %>% arrange(desc(Pearson_r)))
```

```
# ==============================================================================
# Thesis Appendix A.2: Binary Classification and Segment Concordance Framework
# Purpose: This script evaluates the performance of discrete genomic copy number
#          alteration (CNA) calls. It maps platform-specific segmented outputs
#          to a standard 1.0 Mb reference grid using a majority overlap rule,
#          subsequently calculating Sensitivity, Precision, and Macro-F1 scores
#          against ground-truth classifications.
# Dependencies: readxl, tidyverse, GenomicRanges
# ==============================================================================

library(readxl)
library(tidyverse)
library(GenomicRanges)

# --- 1. SETUP Environment & Target Path Directories ---
# Configure directory paths for the dataset input files and reference controls
setwd("path/to/project_folder")
truth_file   <- "DepMap A549.xlsx"
chrom_file   <- "GRCh38.p2.xlsx"
input_folder <- "confusion"

# --- 2. Build Reference Coordinates: Chromosome Sizes & Genomic Grid Alignment ---
cat("Building 1Mb Master Grid...\n")
chrom_sizes <- read_excel(chrom_file, col_names = FALSE) %>%
  filter(...2 != "length") %>%
  select(chr = 1, size = 2) %>%
  mutate(size = as.numeric(size),
         chr = gsub("(?i)chr", "", as.character(chr))) %>%
  filter(!is.na(size))

master_grid <- chrom_sizes %>%
  group_by(chr, size) %>%
  reframe(start = seq(1, size, by = 1e6)) %>%
  mutate(end = pmin(start + 1e6 - 1, size)) %>%
  ungroup()

# Cast the data frame to a GenomicRanges object to facilitate spatial intersection matching
grid_gr <- GRanges(seqnames = master_grid$chr, ranges = IRanges(master_grid$start,
master_grid$end))

# --- 3. Discrete Categorization Mapping Function ---
# Assigns exactly one discrete label ("Gain", "Loss", "Neutral") to each standard 1.0 Mb
bin
map_calls_to_grid <- function(df_clean, master_grid, grid_gr) {
  out_types <- rep("Neutral", nrow(master_grid))
  if(nrow(df_clean) == 0) return(out_types)

  # Map cleaned, discrete segment ranges to genomic intervals
  seg_gr <- GRanges(seqnames = df_clean$chr, ranges = IRanges(df_clean$start,
df_clean$end))
  hits <- findOverlaps(grid_gr, seg_gr)

  if(length(hits) > 0) {
    q_idx <- queryHits(hits)
    s_idx <- subjectHits(hits)

    # Calculate overlap width (base pairs) to determine the dominant label inside the bin
    ov_start <- pmax(master_grid$start[q_idx], df_clean$start[s_idx])
    ov_end   <- pmin(master_grid$end[q_idx], df_clean$end[s_idx])
    ov_bp    <- ov_end - ov_start + 1

    # Enforce a dynamic majority threshold rule (> 500,000 bp overlap) to declare a
structural call
    best_per_bin <- data.frame(bin = q_idx, type = df_clean$type[s_idx], bp = ov_bp) %>%
      group_by(bin, type) %>%
      summarise(total_bp = sum(bp), .groups = "drop") %>%
```

```
      group_by(bin) %>%
      slice_max(order_by = total_bp, n = 1, with_ties = FALSE) %>%
      ungroup() %>%
      filter(total_bp > 500000)

    out_types[best_per_bin$bin] <- best_per_bin$type
  }
  return(out_types)
}

# --- 4. Standardization and Platform-Specific Parsing Utility ---
# Translates varying column layouts and categorical string calls from different pipelines
into uniform states
get_clean_calls <- function(f_path, is_truth = FALSE) {
  fname <- basename(f_path)
  df <- read_excel(f_path, col_names = FALSE, .name_repair = "minimal")

  # Condition 4a: Determine internal data structures based on source software names
  if (grepl("ichorCNA", fname, ignore.case = TRUE)) {
    c_idx <- 2; s_idx <- 3; e_idx <- 4; v_idx <- 12
  } else {
    c_idx <- 1; s_idx <- 2; e_idx <- 3
    if (grepl("IonReporter", fname, ignore.case = TRUE)) v_idx <- 4
    if (grepl("WisecondorX|bowtie", fname, ignore.case = TRUE)) v_idx <- 6
  }

  # Programmatically skip file headers if descriptive text strings are encountered
  if (!is.numeric(df[[1, s_idx]])) { df <- df[-1, ] }

  temp_res <- data.frame(
    chr   = gsub("(?i)chr", "", as.character(df[[c_idx]])),
    start = as.numeric(df[[s_idx]]),
    end   = as.numeric(df[[e_idx]]),
    stringsAsFactors = FALSE
  )

  # Condition 4b: Map multi-class state attributes into standard Gain/Loss/Neutral criteria
  if (is_truth) {
    # Reference database classification maps (DepMap integer designations)
    val <- as.numeric(df[[4]])
    temp_res$type <- ifelse(val %in% c(4, 5), "Gain",
                            ifelse(val %in% c(1, 2), "Loss", "Neutral"))
  } else if (grepl("IonReporter", fname, ignore.case = TRUE)) {
    # Ion Reporter absolute copy number thresholds
    val <- as.numeric(df[[v_idx]])
    temp_res$type <- ifelse(val > 2, "Gain", "Loss")
  } else if (grepl("WisecondorX|bowtie", fname, ignore.case = TRUE)) {
    # WisecondorX structural aberration string outputs
    val <- tolower(as.character(df[[v_idx]]))
    temp_res$type <- ifelse(val == "gain", "Gain",
                            ifelse(val == "loss", "Loss", "Neutral"))
  } else if (grepl("ichorCNA", fname, ignore.case = TRUE)) {
    # ichorCNA sub-clonal and high-level integer segment calls
    val <- toupper(as.character(df[[v_idx]]))
    temp_res$type <- ifelse(val %in% c("GAIN", "HLAMP", "AMP"), "Gain",
                            ifelse(val %in% c("HETD", "HETLOSS", "HOMD"), "Loss",
"Neutral"))
  }

  # FINAL STEP: Flatten contiguous overlapping segments of identical types using reduce()
  # to prevent double-counting artifact regions across chromosomes
  clean_res <- temp_res %>%
    filter(!is.na(type), type != "Neutral", !is.na(start), !is.na(end)) %>%
    group_by(type, chr) %>%
    reframe(
      gr = as.data.frame(reduce(GRanges(seqnames = chr, ranges = IRanges(start, end))))
```

```
    ) %>%
    unnest(gr) %>%
    select(chr = seqnames, start, end, type) %>%
    ungroup()

  return(clean_res)
}

# --- 5. Iterative Pipeline Execution & Contingency Matrix Assessment ---
# Process baseline ground-truth matrix
cat("Processing Truth...\n")
truth_raw <- get_clean_calls(truth_file, is_truth = TRUE)
truth_binned <- map_calls_to_grid(truth_raw, master_grid, grid_gr)

# Discover and iterate across all configuration targets inside the designated folder
file_list <- list.files(path = input_folder, pattern = "\\.xlsx$", full.names = TRUE)

results_final <- map_df(file_list, function(f_path) {
  fname <- basename(f_path)
  cat("Analyzing:", fname, "\n")

  tryCatch({
    # Load and execute grid-normalization for the target test configuration
    test_raw <- get_clean_calls(f_path)
    test_binned <- map_calls_to_grid(test_raw, master_grid, grid_gr)

    comp_df <- data.frame(truth = truth_binned, test = test_binned)

    # Calculate Gain Performance Metrics (True Positive, False Positive, False Negative)
    tp_g <- sum(comp_df$truth == "Gain" & comp_df$test == "Gain")
    fp_g <- sum(comp_df$truth != "Gain" & comp_df$test == "Gain")
    fn_g <- sum(comp_df$truth == "Gain" & comp_df$test != "Gain")

    sens_g <- ifelse((tp_g + fn_g) > 0, tp_g / (tp_g + fn_g), 0)
    prec_g <- ifelse((tp_g + fp_g) > 0, tp_g / (tp_g + fp_g), 0)
    f1_g   <- ifelse((sens_g + prec_g) > 0, 2*(sens_g*prec_g)/(sens_g+prec_g), 0)

    # Calculate Loss Performance Metrics (True Positive, False Positive, False Negative)
    tp_l <- sum(comp_df$truth == "Loss" & comp_df$test == "Loss")
    fp_l <- sum(comp_df$truth != "Loss" & comp_df$test == "Loss")
    fn_l <- sum(comp_df$truth == "Loss" & comp_df$test != "Loss")

    sens_l <- ifelse((tp_l + fn_l) > 0, tp_l / (tp_l + fn_l), 0)
    prec_l <- ifelse((tp_l + fp_l) > 0, tp_l / (tp_l + fp_l), 0)
    f1_l   <- ifelse((sens_l + prec_l) > 0, 2*(sens_l*prec_l)/(sens_l+prec_l), 0)

    # Compute unweighted Macro-F1 score to balance performance measurement across both
aberration types
    macro_f1 <- (f1_g + f1_l) / 2

    tibble(
      File = fname,
      Sens_Gain = round(sens_g, 4), Prec_Gain = round(prec_g, 4), F1_Gain = round(f1_g, 4),
      Sens_Loss = round(sens_l, 4), Prec_Loss = round(prec_l, 4), F1_Loss = round(f1_l, 4),
      Macro_F1  = round(macro_f1, 4)
    )

  }, error = function(e) {
    message("Error in ", fname, ": ", e$message); return(NULL)
  })
})

# --- 6. DISPLAY COMPARATIVE PERFORMANCE MATRIX ---
# Output final configuration ranking ordered descending by the macro-averaging threshold
print(results_final %>% arrange(desc(Macro_F1)))
```

Appendix B Sequencing Read Allocations

This appendix contains the sequencing read totals for the evaluated patient cohort, organized by sample ID and sample type.

Table B. 1: Total Raw Sequence Read Counts for the Evaluated Cohort.

Sample ID	Sample Type	Reads
A549	Positive Control	487262
NM14	Healthy Control	112651
NM15	Healthy Control	147968
NM16	Healthy Control	117341
NM17	Healthy Control	125385
NM18	Healthy Control	120547
NM21	Healthy Control	119971
NM22	Healthy Control	144333
NM30	Healthy Control	110428
NM64	Healthy Control	129082
NM66	Healthy Control	159540
NM67	Healthy Control	114557

NM68	Healthy Control	115249
NM70	Healthy Control	126998
NM72	Healthy Control	294593
NM77	Healthy Control	243082
NM85	Healthy Control	180583
NM87	Healthy Control	245933
NM91	Healthy Control	362837
NM93	Healthy Control	175800
LCG2	Tumor Tissue	130177
LCG16	Tumor Tissue	118301
LCG21	Tumor Tissue	291483
LCG22	Tumor Tissue	141929
LCG26	Tumor Tissue	127285
LCG29	Tumor Tissue	214087
LCG38	Tumor Tissue	183592

LCG42	Tumor Tissue	131311
LCG43	Tumor Tissue	252048
LCG44	Tumor Tissue	155536
LCG45	Tumor Tissue	99575
LCG47	Tumor Tissue	435493
LCG53	Tumor Tissue	149781
LCG62	Tumor Tissue	389584
LCG63	Tumor Tissue	122405
LCG65	Tumor Tissue	325985
LCG75	Tumor Tissue	182760
LCG84	Tumor Tissue	276432
LCG97	Tumor Tissue	277923
LCG105	Tumor Tissue	170701
LCG112	Tumor Tissue	139169
LCG128	Tumor Tissue	84345

LCG139	Tumor Tissue	240193
LCG141	Tumor Tissue	255979
LCG2	cfDNA	93758
LCG16	cfDNA	149655
LCG21	cfDNA	366299
LCG22	cfDNA	163613
LCG26	cfDNA	118974
LCG29	cfDNA	442285
LCG38	cfDNA	60138
LCG42	cfDNA	167458
LCG43	cfDNA	324457
LCG44	cfDNA	119795
LCG45	cfDNA	111049
LCG47	cfDNA	114321
LCG53	cfDNA	204296
LCG62	cfDNA	54697
LCG63	cfDNA	126524
LCG65	cfDNA	370976
LCG75	cfDNA	250131
LCG84	cfDNA	192246
LCG97	cfDNA	361423
LCG105	cfDNA	51704
LCG112	cfDNA	296382

LCG128	cfDNA	204920
LCG139	cfDNA	254816
LCG141	cfDNA	211842

Note: Samples with a read count falling below the quality control threshold of 100,000 reads were excluded from downstream copy-number profiling and pipeline benchmarking. An exception was made for plasma sample LCG45 (99,575 reads), since its minimal deviation (< 0.5%) from the threshold was considered negligible for analysis at 1.0 Mb resolution.

Table B. 2: Read Count Comparison Between Initial and Re-sequenced Runs

Sample ID	Sample Type	Incorrect Reagents	Correct Reagents
LCG47	Tissue	16820	115195
LCG47	cfDNA	7104	43842
LCG62	Tissue	32710	273219
LCG62	cfDNA	11561	80092
LCG97	Tissue	17830	135441
LCG97	cfDNA	7276	50808
LCG112	Tissue	15953	101439
LCG112	cfDNA	7619	49816

Appendix C Full Benchmarking Matrix

This appendix contains the complete evaluation datasets tracking the performance of all eleven pipeline configurations against the established ground-truth baseline.

Table C. 1: Signal Correlation and Error Metrics Across All Eleven Pipeline Configurations.

Pipeline	**Resolution**	**Pearson_r**	**RMSE**
WisecondorX (Bowtie2)	2.0 Mb	0.597	0.402
Ion Reporter	0.5 Mb	0.580	0.451
Ion Reporter	1.0 Mb	0.571	0.511
WisecondorX (Bowtie2)	1.0 Mb	0.553	0.417
Ion Reporter	2.0 Mb	0.511	0.523
ichorCNA	0.5 Mb	0.398	0.464
ichorCNA	1.0 Mb	0.373	0.467
WisecondorX (TMAP)	0.5 Mb	0.296	0.510
WisecondorX (TMAP)	2.0 Mb	0.291	0.502
WisecondorX (Bowtie2)	0.5 Mb	0.289	0.512
WisecondorX (TMAP)	1.0 Mb	0.286	0.510

Table C. 2: CNA Detection Performance Across All Eleven Pipeline Configurations.

Pipeline	Resolution	Sens_Gain	Prec_Gain	F1_Gain	Sens_Loss	Prec_Loss	F1_Loss	Macro_F1
WisecondorX (Bowtie2)	1.0 Mb	0.9339	0.1467	0.2535	0.9258	0.7715	0.8417	0.5476
WisecondorX (Bowtie2)	2.0 Mb	0.7314	0.1192	0.205	0.9033	0.769	0.8307	0.5179
ichorCNA	0.5 Mb	0.9339	0.1403	0.2439	0.7606	0.7452	0.7528	0.4984
ichorCNA	1.0 Mb	0.9339	0.1387	0.2416	0.7512	0.7442	0.7477	0.4946
WisecondorX (TMAP)	1.0 Mb	0.876	0.1324	0.2301	0.7446	0.7411	0.7429	0.4865
WisecondorX (Bowtie2)	0.5 Mb	0.8802	0.1251	0.219	0.7455	0.7372	0.7414	0.4802
WisecondorX (TMAP)	0.5 Mb	0.843	0.119	0.2085	0.7418	0.7411	0.7414	0.475
Ion Reporter	0.5 Mb	0.9587	0.1333	0.2341	0.4873	0.9647	0.6475	0.4408
WisecondorX (TMAP)	2.0 Mb	0.3182	0.0605	0.1017	0.7465	0.73	0.7382	0.4199
Ion Reporter	1.0 Mb	0.9587	0.1339	0.2349	0	0	0	0.1175
Ion Reporter	2.0 Mb	0.8182	0.1153	0.2021	0	0	0	0.1011

Appendix D Supplemental CNA Profiles

This appendix contains the copy number profiles for the patients within the new cohort whose samples cleared the QC threshold of >100,000 reads (n = 7).

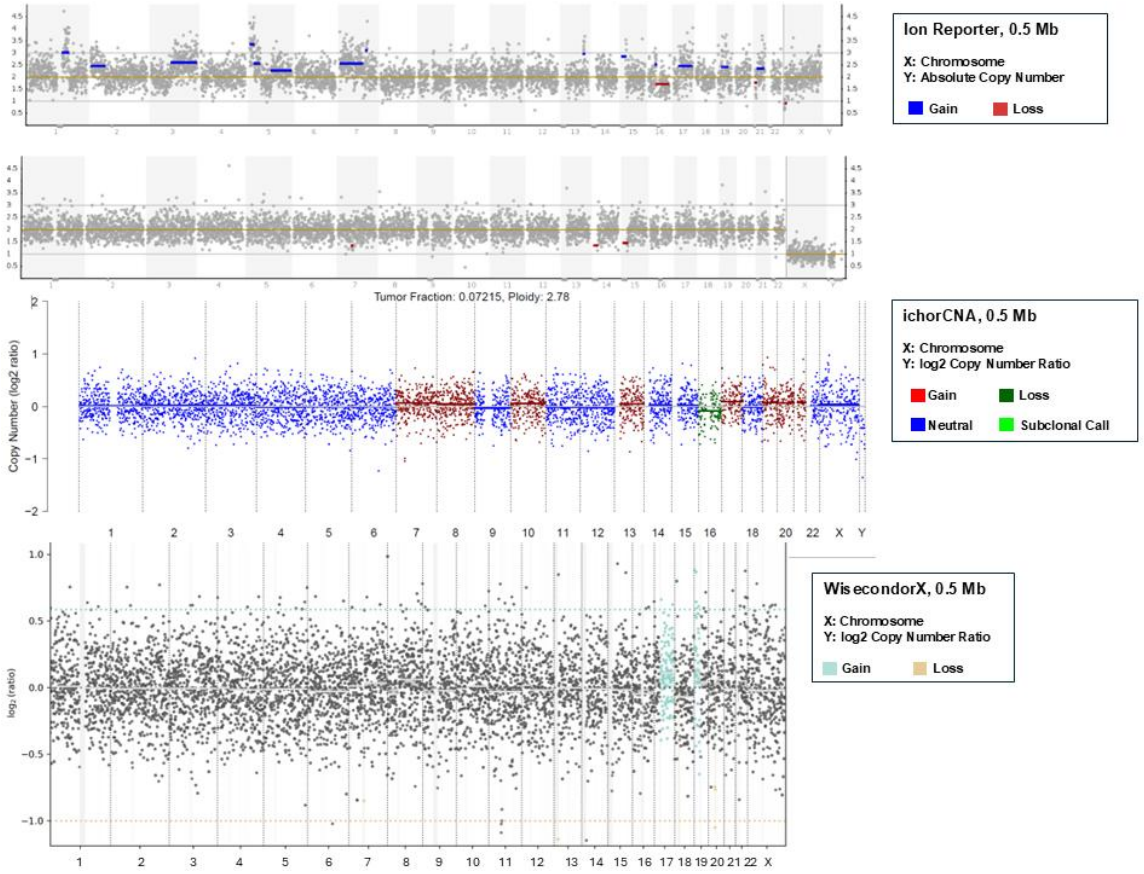

Figure D. 1: Copy Number Profiles of Patient LCG21.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

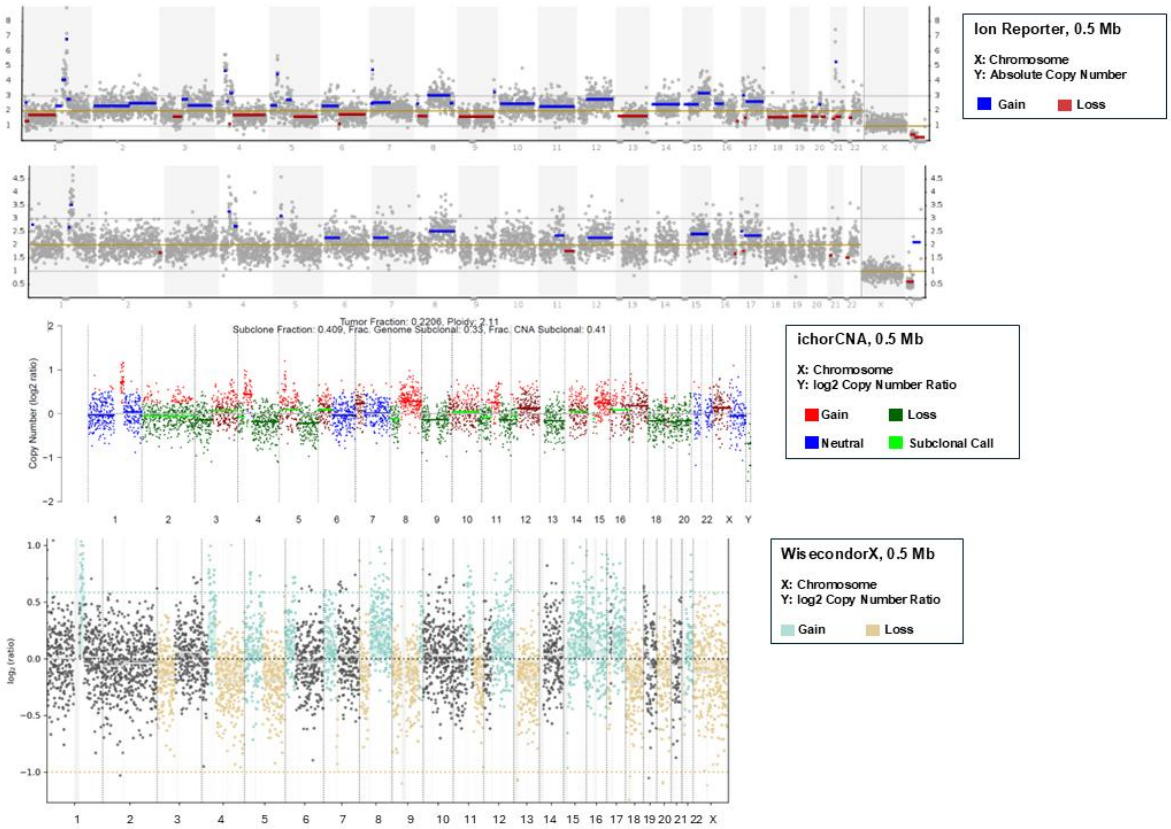

Figure D. 2: Copy Number Profiles of Patient LCG29.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

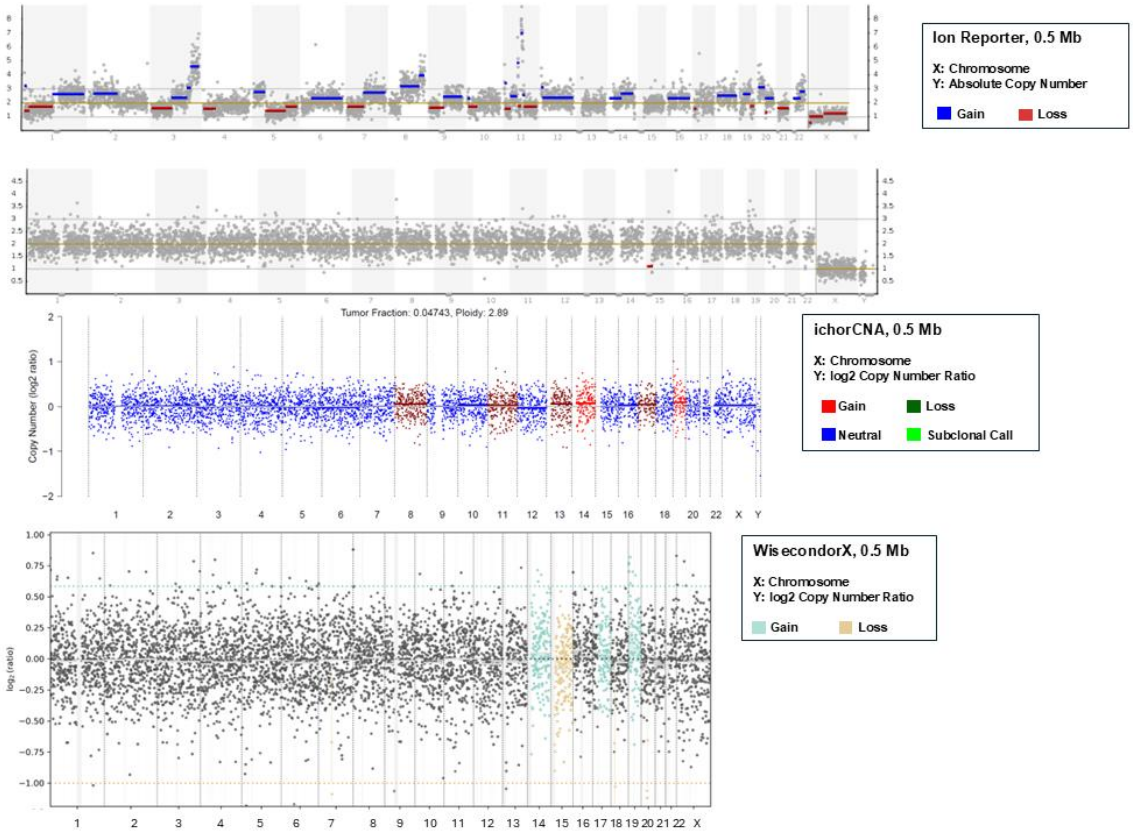

Figure D. 3: Copy Number Profiles of Patient LCG43.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

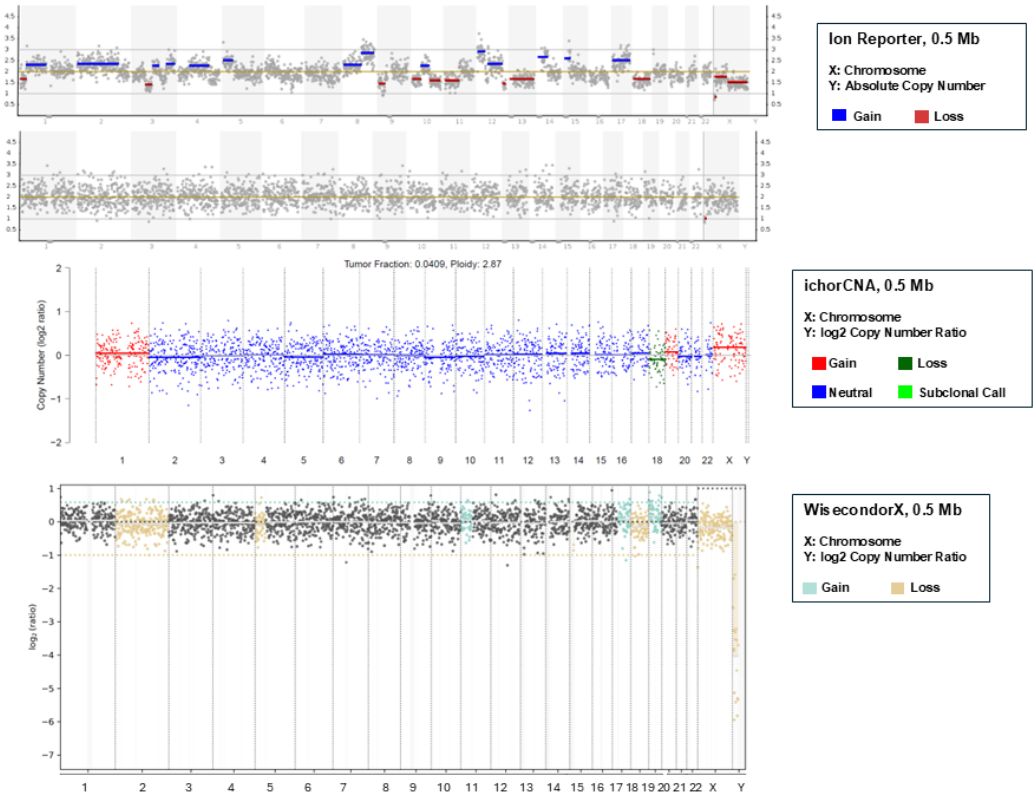

Figure D. 4: Copy Number Profiles of Patient LCG47.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

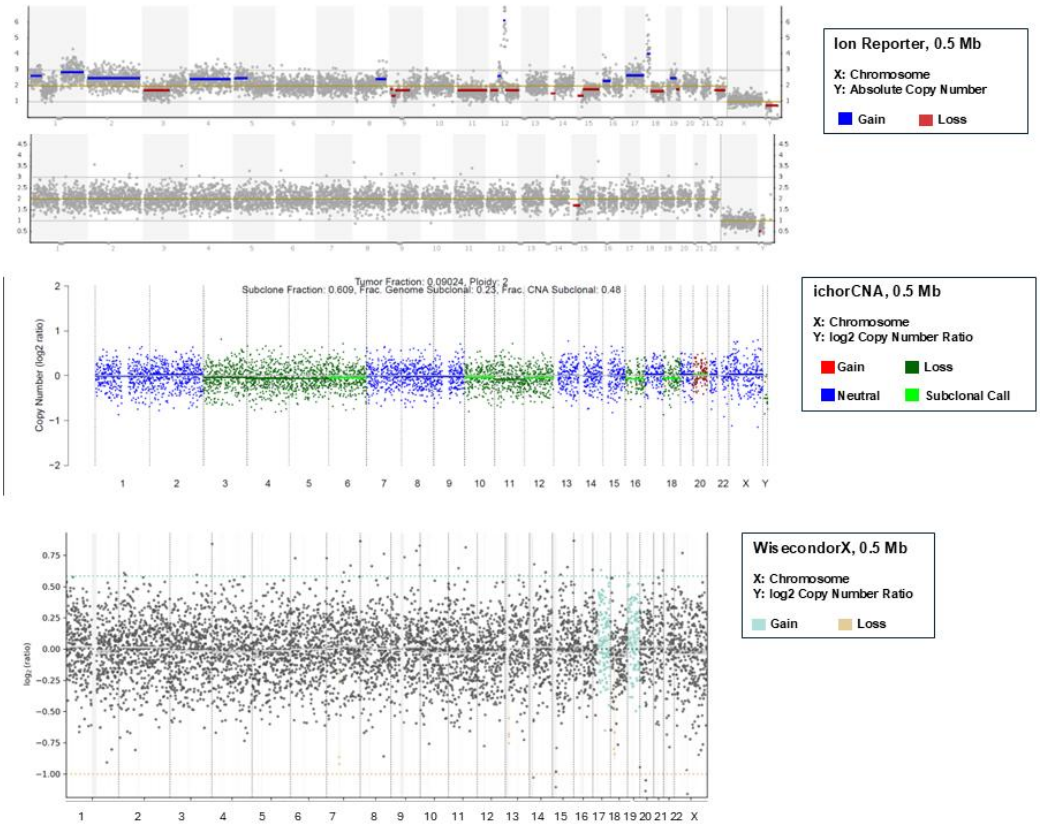

Figure D. 5: Copy Number Profiles of Patient LCG65.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

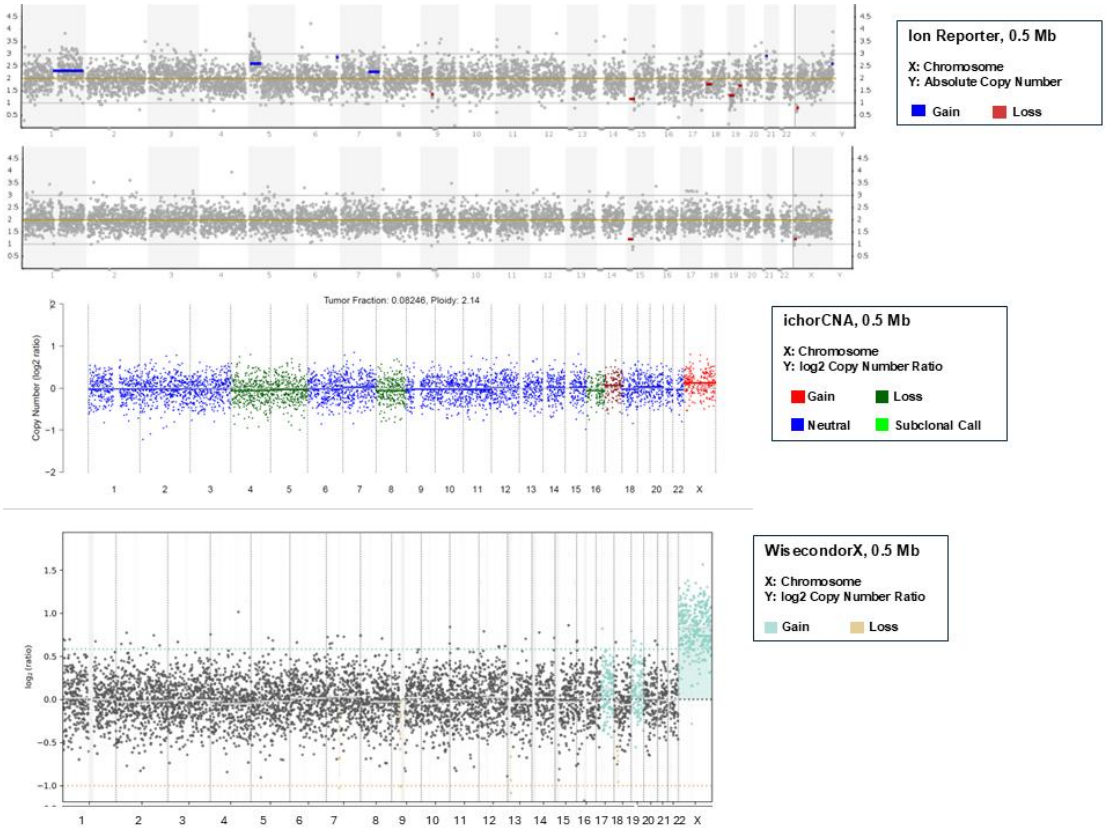

Figure D. 6: Copy Number Profiles of Patient LCG97.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

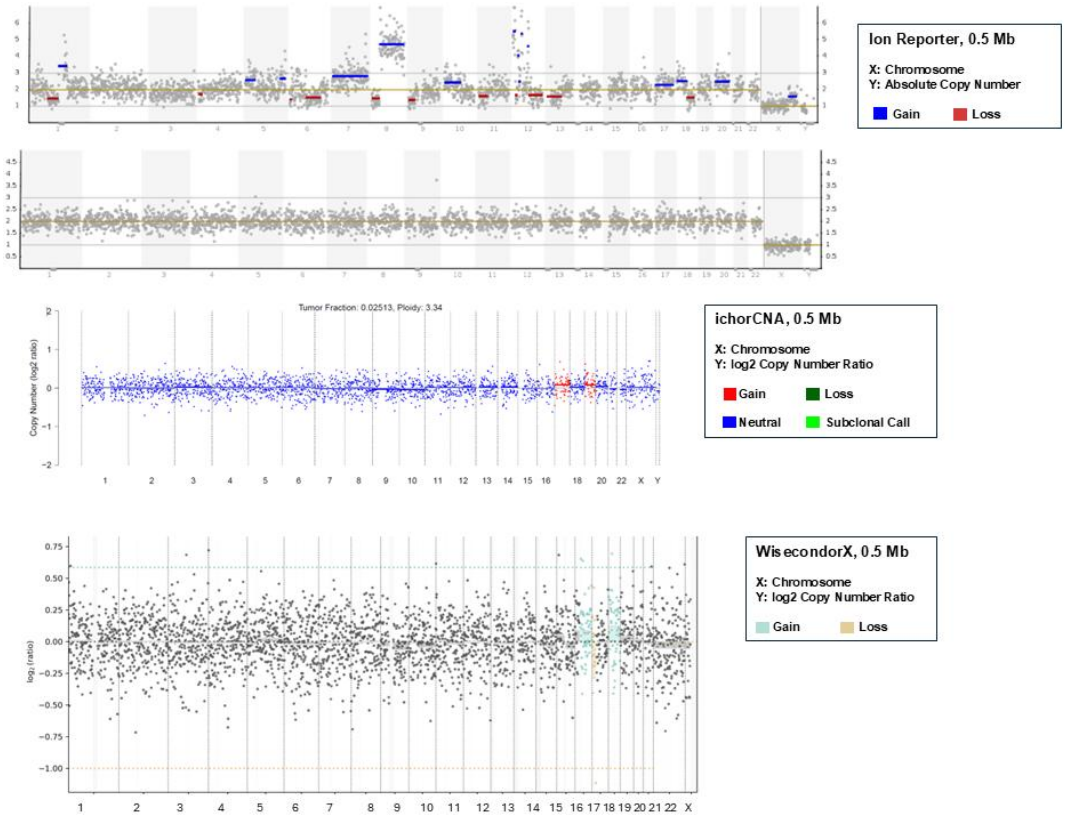

Figure D. 7: Copy Number Profiles of Patient LCG112.
Genomic copy number profiles across different sample types and bioinformatic pipelines. Panels from top to bottom display: (1) tumor tissue analyzed via Ion Reporter, (2) plasma cfDNA analyzed via Ion Reporter, (3) plasma cfDNA analyzed via ichorCNA, and (4) plasma cfDNA analyzed via WisecondorX. *Figures captured from the pipeline visual outputs. Ion Reporter uses an ENPB of 1.8-2.2.*

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY