Characterisation of Short-Segment Copy Number Inference Algorithms

Author:

Mohammad Zardbani

Master of Science (Bioinformatics) Thesis

University of Melbourne

Project Location:

Walter and Eliza Hall Institute of Medical Research (WEHI)

Project Supervisors:

Prof. Tony Papenfuss
Dr. Daniel Cameron

Acknowledgments

I would like to express my gratitude to my supervisors Prof Tony Papenfuss and Dr Daniel Cameron from the bioinformatics division at the Walter and Eliza Hall Institute of medical research (WEHI). I am grateful for all the valuable support and feedback generously provided by Prof Tony Papenfuss, who have been an immense source of knowledge and expertise throughout my master's thesis research project. I was also fortunate to learn crucial coding and analytical skills from the vast technical expertise of Dr Daniel Cameron and grateful for the opportunity to explore his research topic proposal in this project.

Contents

Frequently Used Abbreviations

SNP Single Nucleotide Polymorphism

SV Structural Variation

CNV Copy Number Variation

DP Discordantly Aligned Read-pairs

CI Confidence Interval

VAF Variant Allele Frequency

RD Read-depth

CN Copy-number

List of Figures

1 Introduction and Background

Genomic variations are alteration of genetic material which can range in size from small variations, collectively referred to as single nucleotide polymorphisms (SNPs), to large variations such as, structural variations (SVs) and copy number variations (CNVs) [1]. SNPs are the simplest and smallest form of genetic variation, generally less than 50 bases (b) in size, and are often only characterised by base substitution [2, 3]. SVs are defined as larger variation events constituting a wide range of genetic rearrangements such as deletions, insertions, duplications, inversions, translocations, and complex rearrangements (figure (1)), which would result in either producing balanced or unbalanced rearrangements [2, 4]. Variants disrupt continuity of genetic sequence in one or more breakpoint(s) and produce flanking sequences which surround SVs and CNVs. Identifying such breakpoints is supportive evidence for existing variants in a region. SV breakpoints also provide valuable quantitative information about allele frequencies at breakpoints. In balanced rearrangement events such as inversions and translocations, the amount of genetic material remains unchanged. However, unlike balanced rearrangements which are copy-number neutral, CNVs are defined as unbalanced rearrangements that result in genetic gain or loss, and consequently, alter the amount of DNA present in a cell. CNVs often change the number of genes and are associated with a wide range of diseases including autism [5], schizophrenia [6, 7], and cancer [2–4]. Hence, accurate SV and CNV genotyping is essential for discovering their underlying association with human diseases, evolution, and complex traits [2].

Variant detection methods can be broadly divided into experimental and computational approaches [8]. Traditionally, SVs and CNVs were detected and characterised using experimental methods such as fluorescence in situ hybridization (FISH), karyotyping, and array-based techniques (e.g., comparative genome hybridization (CGH)) [9–11]. However, there are considerable methodological constraints and drawbacks associated with experimental methods including, low resolution, inability to detect novel or rare variants, and low genomic coverage, which makes experimental variant detection methods unsuitable for precise and accurate genotyping of SVs and CNVs [12]. SVs were originally defined as variation events that are larger than 1 kilo base (kb) in size using experimental approaches. However, progress in sequencing technologies and development of better methodology in sequencing data analysis redefined SVs and CNVs as genomic variations larger than 50(b) [8]. Computational methods use sequencing data generated by massively parallel sequencing technologies for genomic variation inference (SV and CNV calling). Often computational based SV and CNV calling tools are referred to as SV and CNV callers respectively. Even though, it is possible to identify precise breakpoint locations of an SV with high resolution (i.e., single-base resolution) and in theory identify, resolve and characterise any SV or CNV using only sequencing data. Nonetheless, variant analysis methods are still limited by constraints of assembling error free sequences as input, high complexity of genomic rearrangements, and computational limitations [1].

SV and CNV callers rely on the information provided by single or paired-end sequencing read datasets for variant inference. Sample reads are typically aligned to a reference genome (control) and variations are inferred based on the position, orientation, and signal intensity of mapped reads. Read alignments can be characterised and labelled by variant callers to describe a genomic rearrangement event more effectively. Read-depth, split-read, and discordantly aligned paired-end reads (DPs) are commonly used read types for variant detection (figure (2)). Read-depth is simply the number of reads aligned to a given locus. Change in read-depth signal intensity is often an indicator of copy-number change and therefore, read-depth signal is widely used by CNV inference algorithms as an input [13]. The majority of sequencing reads are expected to have a full alignment, hence, reads that are partially aligned or mapped unexpectedly can be

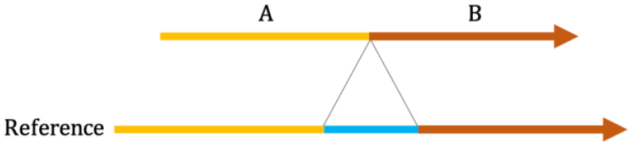
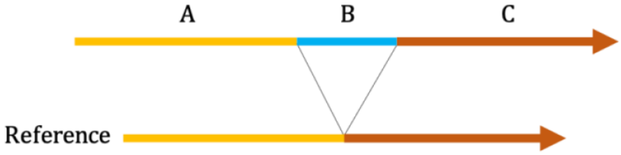
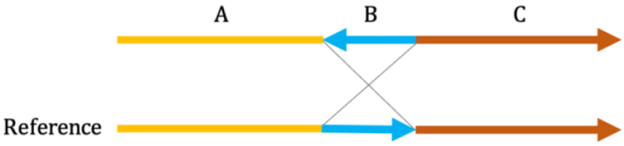
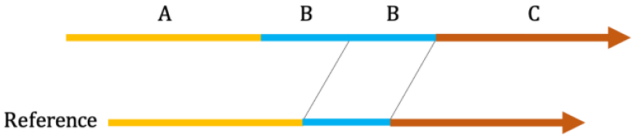
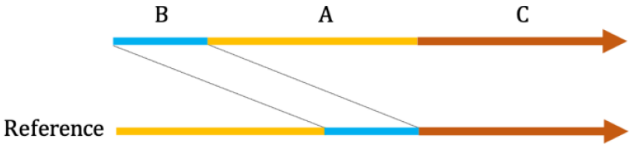
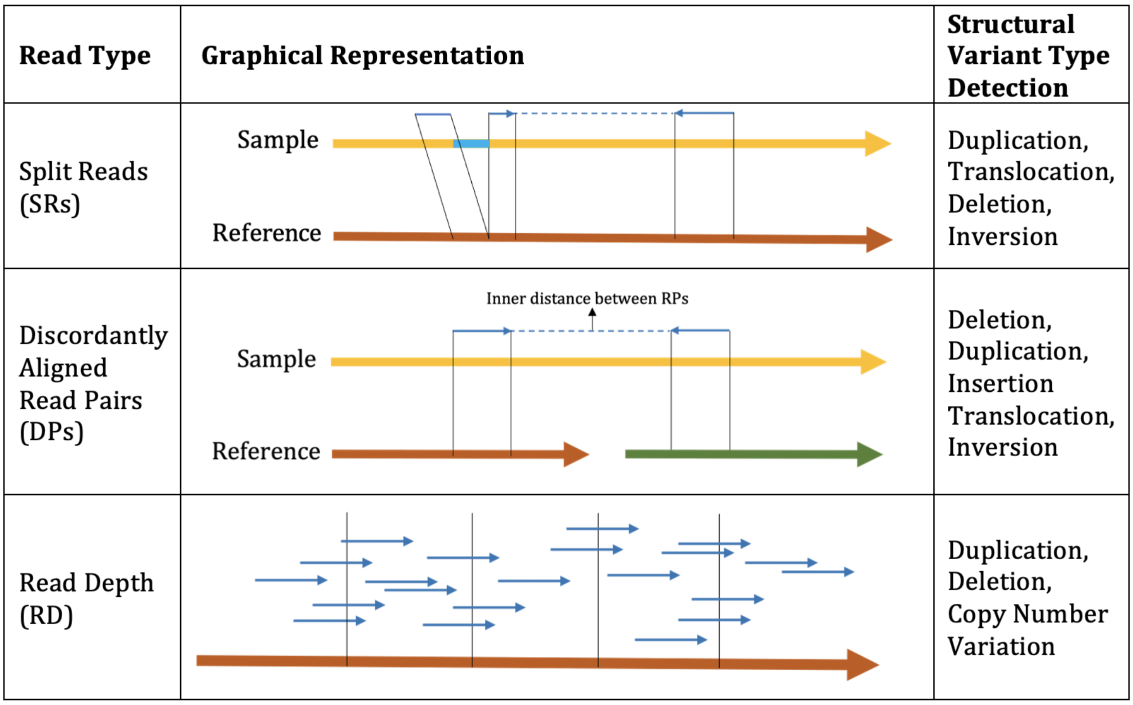
Structural Variation Type	Graphical Representation	Number of Breakpoints
Deletion		1
Insertion		2
Inversion		2
Duplication		1 or more
Translocation		2

Figure 1: *This figure presents basic structural variation types, their graphical representations, and generated number of breakpoints by each variation event on sequencing data. In deletion events parts of a query sequence is missing when compared to a reference genome. Query sequences with insertions contain unique extra sequencing data. An inversion event is characterised by an inverted region in the query sequence when compared to a reference genome. Duplication or in general, amplification events are characterised by presence of amplified sequencing data. Translocation events are shuffling of sequence fragments within the same sequencing data.*

used as evidence to infer SVs. Split-reads are partially mapped reads from single or paired-end read datasets that indicate the presence of a potential SV breakpoint. Fragment size (also referred to as inner-distance or insert size) and read length of paired-end reads are known values which can also be used to check whether the reads are mapped to the genome as expected (i.e., concordant reads) or not (i.e., discordantly aligned reads). DPs are paired-end reads with unexpected alignment features. For instance, genomic rearrangement events such as duplications and insertions may alter the inner distance between read pairs; or translocation events may result in alignment of each read to distantly situated loci (figure (2)) [14]. Split-read and DP read types are commonly used for SV analysis, and read-depth signal is used for copy-number analysis by most CNV callers. Large CNV events, up to chromosome levels, produce strong read-depth signals and therefore, most copy number inference algorithms use read-depth or coverage data to detect large CNVs with high confidence [15].



Read Type	Graphical Representation	Structural Variant Type Detection
Split Reads (SRs)	Sample Reference	Duplication, Translocation, Deletion, Inversion
Discordantly Aligned Read Pairs (DPs)	Inner distance between RPs Sample Reference	Deletion, Duplication, Insertion Translocation, Inversion
Read Depth (RD)		Duplication, Deletion, Copy Number Variation

Figure 2: *This figure represents only a subset of sequencing read types that could be used as evidence for genetic variant inference. Split-reads represent sequencing reads that contain structural variant supportive evidence along their read length. Discordantly aligned read-pairs represent a large subset of variant supportive reads that are abnormally mapped to the sequencing data because they were sequenced from parts of query sequence with variation. Read-depth is the number of reads (i.e., read counts)for a given locus along the query and reference sequence.*

Variant allele frequency (VAF) is the fraction of alleles that contain variation over total number of alleles for a given locus. Read-depth signal combined with VAF values derived from SNP calls are particularly important for cancer cell fraction (CCF) or purity calculations in cancer genomics [16]. However, VAF values can be used for CNV inference as changes in the ratio of major and minor allele frequencies provide evidence for changes in copy-number. Tumour ploidy directly influence sequence coverage which in turn deviates VAF from 0, 0.5, and 1 to other fractions (e.g., 0.3 and 0.6 for aneuploid regions) in case of aneuploidy [17]. VAF values can also be calculated at an SV breakpoint, referred to as variant allele frequency of SVs (SV-VAF), which provide information about copy-number change caused by SVs [18]. The ability to confidently define boundaries of SVs and CNVs depends on the resolution of calling tools that are used to detect the variants [1]. SV callers with single base resolution capability are needed to determine the variant allelic frequency at the SV breakpoint (i.e., allelic frequency of an SV (SV-VAF)). Most CNV calling algorithms primarily utilise read-depth for copy-number inference, however, read-depth signal becomes noisy as the length of a variant segment is reduced. Furthermore, in short variant segments read-depth signal becomes increasingly weak and effectively indistinguishable from the noise. As a result, it is challenging to accurately infer copy-number of short variant segments and often CNV calling algorithms are either incapable of handling such short segments by design or dismiss them all together to achieve better performance scores during benchmarking.

SNPs are extensively studied and a variety of curated databases and effective computational tools for detecting SNPs are currently available. However, despite a strong interest in SV and CNV related studies in the past decade, computational methods and tools that detect, annotate and analyse SVs and CNVs are still evolving to achieve better accuracy, speed and gain more reliability. Consequently, further progress in tool development for detecting and analysing larger structural variations, SVs and CNVs, is needed.

In this research project we explore and evaluate copy-number inference methods that utilise read-depth and SV-VAF to establish a baseline for short-segment CNV calling. Given that SV-VAF does not depend on variant segment length, CNV calling algorithms can potentially utilise SV-VAF instead of read-depth signal to infer copy-number in short segments. Nonetheless, information about effectiveness of SV-VAF for short-segment copy-number inference is limited. To the best of our knowledge, a comparison between SV-VAF and read-pair signals for short-segment copy-number inference have not been reported by previous studies. In this study, we investigate the effectiveness of using SV-VAF and read-depth signal for copy-number inference in short segments. We performed computer simulations to suggest the most suitable input signal for short-segment copy-number inference models. Furthermore, length threshold of short segments based on variant types and sizes were identified. We also investigate short-segment variant calling ability of select SV and CNV callers on real sequencing data to report a small-scale variant calling benchmarking analysis. Lastly, based on our results, we characterise variation event types and sizes, in order to make recommendations about utilising SV-VAF or read-depth signal that improve short-segment copy-number inference.

2 Research Project Objectives

2.1 Research Motivation

Currently established CNV inference methods and tools cover a wide range of statistical and computational approaches which are empirically or theoretically based. The literature surrounding copy-number calling is also populated with methodologies that underly such CNV inference algorithms in great detail. Regardless of the available repertoire however, variant calling tools are constantly challenged by existing and upcoming cancer genomics data due to presence of complex and novel rearrangements in tumour cells. CNV calling algorithms rely on sequencing signal input for copy-number inference and therefore, it is inevitable to have parts of input data incomplete, noisy, or statistically weak for high confidence genetic variation inference. Particularly, CNV calling in short-segments is challenging due to lack of sufficient statistical inference signal and presence of excessive noise. Our focus in this study is to develop a robust platform to detect and reliably infer copy-number of such CNVs. Knowing it is more likely to observe complex genomic rearrangement clusters and compound variation events in genomes of cancer samples, we integrate commonplace cancer genomics *in silico* experimental design in this study with the broad objective to advance variant calling in tumour sequencing data. Nonetheless, our methodology and findings are not limited to cancer genomics only, and can be applied to wide range of experimental and clinical variant calling settings.

2.2 Research Questions

We designed our methods to answer the following questions from generated results and reported findings. Our intended research questions are:

- What is the size range of short-segment copy-numbers?
- What is the most appropriate computational approach for copy-number inference in short-segments?
- Does SV-VAF result in better performance in short-segment copy-number calling compared to read-depth signal?
- What is the best strategy for short-segment copy-number inference based on variation events and sizes?
- How do current variant callers perform in short-segment variant calling?

2.3 Study Aims

Our aims for this study include:

- Replicate sequencing data equivalent SV-VAF and read-depth signal by computer simulations.
- Develop copy-number inference models based on read-depth and SV-VAF.
- Detect a size range for short-segments using developed models.
- Define a universal performance metric to evaluate model performance along identified short-segments.
- Investigate performance of read-depth and SV-VAF based variant inference models in short-segments.

- Investigate segment size thresholds where performance of read-depth and SV-VAF models cross over.
- Replicate CNVs on real sequencing data to characterise variation events and sizes by benchmarking variant callers.
- Report performance of a few select SV and CNV callers on short-segment variant calling.
- Establish a platform for further research in short-segment copy-number inference.

3 Literature Review

3.1 Introduction

The advent of high throughput sequencing about 15 years ago lead to discovery of a new range of SVs and CNVs which were previously undetectable due to methodological constraints. Since then, a broad range of variant types and sizes are characterised and numerous variant callers were developed. Variant callers rely on variant detection signals for SV and CNV inference and implement statistical or heuristic approaches in their algorithms. Nonetheless, the ever expanding variant analysis studies from sequencing data constantly demands for more efficient, accurate, and capable variant callers. This is specially the case in cancer genomics where tumour samples contain complex and detrimental genetic rearrangements involving small events that contain one to a few hundred nucleotides, to large scale events up to chromosome or even genome levels. Moreover, already established role of SVs and CNVs in evolution, population diversity, predisposition and onset of various diseases make them important targets in therapeutics and clinical studied. Consequently, development of variant callers with the capacity to accurately detect the entire range of variant types and sizes is crucial for further progress in genetic research.

The literature covering cancer genomics and variant calling tools also contain a diverse and rich repertoire of statistical and holistic methods (table 1). However, given the complexity of genetic variation in cancer samples, variant callers often need to be equiped with additional features. These features include the ability to estimate purity and ploidy of samples, capability to perform allele-specific inference, and high-resolution variant breakpoint detection. Hence, often new and novel algorithmic approaches and methods are essential to achieve high performance variant calling in cancer data. Furthermore, making improvements to variant callers in this space is challenging and slow. The switch from array-based variant detection methods to detecting genetic variants from sequencing data contributed significantly to our understanding of genetic variants and their role. Perhaps now we are at the verge of another breakthrough in genetic variant calling by seizing the ability to detect variants with single-nucleotide resolution and trace variation events along their track. Knowing such details about variation events would allow us to capture a broader perspective of genetic alterations and their mechanisms in genomic studies and paves the way for better understanding of their clinical implications. This will potentially mark a new era in precision medicine and has already provided life-saving clinical outcomes in therapeutics development [19].

In this literature review we first introduce a broad overview of important variant detection signals, variant inference algorithms and most commonly utilised methods. We then discuss the role of CNVs in disease, and review a wide range of CNV inference algorithms and their approach to copy-number inference. Our review of the current variant calling literature provides evidence that combining variant calling signals leads to both better detection rate and discovery of a wider range of variant types and sizes [20–23]. We aim to implement suggestions made by the literature in designing our methods and analysing our results.

3.2 Variant Calling

Variation inference methods from sequencing data can be broadly divided into five different approaches including, *de novo* assembly of genome, paired-end read mapping, split-read, read-depth (also referred to as depth of coverage), and a combination of mentioned approaches [8]. These approaches vary based on the input or algorithmic strategy employed to detect variations. SV callers generally employ paired-end read mapping, split-read, *de novo* assembly, or a combination approach. And CNV callers mostly rely only on read-depth signal as input for CNV inference. Insert size and orientation of reads are main sources of information in paired-end mapping based SV calling methods. Split-reads could also indicate presence of SVs and statistically significant clusters of split-reads are used for up to single-base resolution SV calling and identifying exact location of SV breakpoints. However, paired-end mapping and split-read methods lack suitable information for copy-number inference. Read-depth based CNV callers typically use the following four steps for copy-number inference: 1) mapping, 2) normalisation, 3) segmentation, and 4) copy-number estimation [8]. Read mapping and normalisation (i.e., GC bias and regions with repeated sequence adjustment) steps produce read-depth signals which are similar to the data generated by traditional CNV calling methods (e.g., array CGH). Therefore, often same algorithms such as circular binary segmentation (CBS) [24] which were used on array-based data, can be used with minor alterations to infer copy-numbers from sequencing data. Estimated copy-numbers are then used for segmentation of genome accordingly and CNVs are inferred for segments with alternate copy-numbers [8]. Defining segments efficiently is a challenging task in CNV calling and CNV callers often either consider model-free or model-based (statistical) approaches to tackle this difficulty. Model-free methods which were mostly employed by early CNV callers tend to have a high false discovery rate, hence, most modern tools rely on methods that incorporate statistical models such as CBS, mean shift-based, shifting level model, expectation maximisation, and hidden Markov models (HMM). Statistical models assume probability distributions and work with statistical measurements on read-depth which rely on parametric and non-parametric models for copy-number inference.

Statistical models that utilise read-depth for copy-number inference (Change-point methods), adapt to local read-depth signal change by dividing sequencing data into bins with equal widths (e.g., 1000(b)). Consequently, they are able to detect localised changes in read-depth signal and therefore, identify segments with alternative copy-number values (i.e., CNV segments). Various sequencing data parameters such as read-depth are then calculated per bin. This makes calculation of mean values localised and accounting for biases such as GC bias simpler. Moreover, over-segmentation of sequencing data can be prevented by merging adjacent segments with equal copy-number. Apart from binning the sequence, while performing variant calling, genome is also segmented according to variant change-points (i.e., breakpoints in case of SVs or copy-number change points in CNVs). Genome segmentation is crucial for many variant calling algorithms to perform inference with high accuracy and specificity. It is reported that the sensitivity of segmentation methods can be increased by matching segment boundaries to SV breakpoints along the sequence. Increased sensitivity results in capturing potential oscillating copy-number regions and small deletions better (supplementary note: [25]), [26].

Most modern callers such as GRIDSS [27] and PURPLE [28] offer better performance by combining multiple inputs such as read-pairs, split-reads for SV calls, and use read-depth signal along with SV breakpoint positions for CNV calling. State-of-the-art callers also utilise SV breakpoints which were passed down from SV calls as evidence to check for copy-number change (i.e., SV aware CNV calling). CNV calling from SV-VAF requires single nucleotide resolution SV calls to check for copy-number change when moving across SV breakpoints and calculation of SV-VAF values [18]. High performance of combination approach is inherited from selecting the most desirable features of other available methods to more reliably infer SVs.

GRIDSS [27] is a top-performing SV caller with the ability to accurately report single-base resolution variation breakpoints. It takes advantage of assembly methods and looks for evidence of SVs from various sequencing data inputs including split-reads and discordantly aligned read-pairs. Concordantly aligned read-pairs are often removed from paired-end read datasets in SV callers such as GRIDSS that do not use read-depth signal for SV or copy-number inference.

However, when read-depth or copy-number analysis is performed concordant reads are necessary for inferring CNVs and heterozygous deletions [14].

3.3 CNVs and Disease

Large deletion and duplication events lead to a change in the amount of DNA; amplification events result in increasing the copy-number in a segment, or a whole chromosome in extreme chromosome doubling events. Deletion events, consist of hemizygous deletions or homozygous deletion (i.e., loss of heterozygousity (LOH) cases) that cause a reduction in copy-number value of a segment. Events involving SVs with no effect on copy-number are called copy neutral events (e.g., inversions). Copy-number analysis of whole genome from sequencing data is described as 'digital karyotyping'. CNVs are detected in a range of serious health conditions including Alzheimer's disease, schizophrenia, autism, Parkinson's disease, and cancer [5–7, 29–31]. Furthermore, CNVs that are located in close proximity to one another (i.e., recurrent or compound CNVs) can be used to identify cancer genes [32]. One recently published study [33] attempted to model CNV states for predicting survival rates of cancer patients.

The evidence for importance of CNVs in diseases is already established. However, the challenge is now to detect and infer the copy-number in these regions, which requires high resolution CNV calling. Unlike SNPs which are commonly shared between populations and can be used as a reference point, shared CNVs between two populations are challenging to detect as CNV boundaries between individuals vary, and multiple genotypes for higher copy-number CNV regions exist [34, 35]. Mentioned complications and other contributing factors make studying CNVs from sequencing data more challenging than SVs. Identifying the entire CNV size range may assist with development in the areas of precision medicine and personalised pharmaceuticals.

3.4 CNV Inference Algorithms

The process of fitting a model on sequencing data to allocate copy-number values (e.g., genome-wide, allele specific, etc.) is called segmentation. Read-depth based CNV calling algorithms have the ability to detect large CNVs with high accuracy because large copy-number segments contain sufficient number of reads, read-depth signal is suitable for reliable copy-number inference. However, read-depth signal becomes noisy for smaller CNVs (typically $<$1000(b)), which negatively impacts accuracy of copy-number inference in short segments [8].

In cancer genomics, sequencing samples contain a mixture of normal and cancer cells. Estimating the ratio of cancer to normal cells (purity estimation) is crucial for tumour related data analysis including CNV calling from cancer data [36, 37]. Tumour samples are believed to follow an evolutionary clonal theory that accounts for prevalence level of acquired alterations in cells to construct phylogenetic trees. The clonal expansion of cells in a tumour helps with calculating purity of tumour samples. Often high-prevalence cell alteration events occur during early stages of tumour formation and in turn ancestral cells are located at the root of a phylogenetic tree. Similarly low-prevalence cell alteration events occur in cells that populate leaves of a phylogenetic tree. Using this information three types of cells can be considered: normal cells (i.e., cells without the aberration), tumour cells that contain the alteration, and tumour cells without the alteration. Hence, purity is also referred to as aberrant cell fraction (ACF). Tools and algorithms such as Sequenza [38], ABSOLUTE [37], and allele-specific copy number analysis of tumours (ASCAT) [36] infer allelic copy-number profiles (allele-specific copy-number calling) by accurate estimation of purity and ploidy of sequencing samples. Copy-number profiles can be inferred from data with sufficient coverage and depth by calculating read-depth ratios which is the relative number of mapped reads to a genome position. Furthermore, ploidy of tumour samples, the number of chromosomes in somatic cells, is another important parameter for allele-specific copy-number inference. Ideally studies that involve variant calling from tumour data would include paired sequencing data from normal and tumour samples in order to accurately calculate ploidy and purity of tumour samples. Ploidy values are used for correctly assigning reference copy-number values to normal sample.

Variant calls in tumour samples contain both germline and somatically acquired mutations. However, often only somatic mutations are clinically interesting and germline mutations are filtered out from SV call datasets. In paired tumour-normal studies, variants that are in the normal sample are considered germline mutations and therefore, excluded from tumour variant calls [14].

One of the earliest methodologies described for high-resolution CNV detection and its accompanying algorithm is described in SegSeq [39]. Statistical power analysis on sequencing data can be employed to detect CNVs and to determine boundaries of copy-number segments. Assuming that reads are randomly generated in a genome sample, the number of reads in a given segment follows a Poisson distribution. Knowing the distribution of read counts in each segment, then inferring copy-number of segments from read-depth signal becomes relatively simple.

ASCAT [36] is one the main traditional allele-specific copy number algorithms which was developed for SNP array data and later adopted for sequencing data. Array data provides two outputs which are primarily used for copy-number analysis: measure of total signal intensity (logR track), and allelic contrast (i.e., alternative or B allele frequency (BAF)). ASCAT employs Allele-specific Piecewise Constant Fitting (ASPCF) algorithm to reduce and seperate noise from data. ASPCF is a filtering and segmentation algorithm that fits piecewise constant regression functions to logR ratios and BAF values in 40 distinct regions of genome in order to force both change-points and segmentation aligned to the same position [36]. ASCAT then utilises ASPCF's output to estimate purity and ploidy of each segment and subsequently, infer copy-number of alleles.

CNVkit [40] is a package of computational methods for genome wide CNV calling and visualisation from targeted whole genome sequencing. In targetted DNA sequencing reads mapped to targets and other regions are called on- and off-reads respectively. Even though, off-reads are often not suitable for detecting SNPs, they can provide information about CNVs. CNVkit uses both on- and off-target reads to calculate logR values (i.e., construct read-depth signal) for copy-number inference. The default genome segmentation algorithm in CNVkit is CBS, however, HaarSeg [41] and Fused Lasso [42] can also be optionally specified as segmentation algorithms. Logarithm of tumour and normal read ratios (i.e., logR track) is replicated in sequencing data by measuring total copy-number of SNP loci. Moreover, BAF is adopted as a number between 0 and 1 indicating the allelic imbalance of an SNP, where values 0 and 1 indicate homozygous and 0.5 heterozygous regions respectively (e.g., ASPCF identifies regions with calculated BAF values of less than 0.3 or more than 0.7 as homozygous). As a result, ASCAT algorithm with minor adaptations is used to infer allele-specific copy-number from sequencing data for instance by `copynumber` [43], a popular R implementation of this algorithm utilised by other copy-number callers such as Sequenza [38] and PURPLE [28] to retrieve copy-number segmentations.

Sequenza [38] uses paired tumour-normal data to infer allelic copy-number profiles by utilising VAF and average depth ratios (normalised read-depth of tumour over read-depth of normal) in a probabilistic model. Sequenza's model is applied to segmented data with the following model parameters: tumour ploidy, purity, segment-specific copy-number, and minor allele copy-number. Sequenza segments data by determining heterozygous and homozygous regions from sequencing depth in normal sample and calculating variant alleles and allelic frequency of tumour samples. Given that a single gold standard for tumour data does not exist, performance evaluation of new tools are often conducted by direct comparison of state-of-the-art tools and calculating the correlation of produced results. Purity estimates produced by Sequenza and ASCAT in segments with similar ploidy estimates are strongly correlated. Consequently, inferred copy-number profiles of these segments were in agreement between the two.

Most allele-specific inference algorithms determine copy-number profiles relative to sample's purity and ploidy (i.e., relative allele-specific copy-number profiles output by Sequenza). In general relative copy-number profiles can be inferred from microarray, comparative genome hybridisation (CGH), and sequencing data. Given that relative copy-number profiles are sample specific, often inferred copy-number profiles of tumour samples cannot be used to compare results between studies or even between samples. Absolute copy-number of cancer cells (i.e., copy-number per cancer cell) on the other hand, are simple integer values which are easy to interpret and can be used to study cancer cell lineage more broadly. ABSOLUTE [37] is a copy-number caller capable of inferring absolute copy-number profiles from both total and allelic copy-ratio data derived from CGH, SNP microarray, and sequencing methods. ABSOLUTE takes copy-number segmented data along with precomputed models of recurrent cancer karyotypes and optional SNP allelic fraction values as input. It then computes the tumour ploidy and purity, and assigns integer values to clonal copy-number states and non-integer values to sub-clonal copy-number states, using a probabilistic model. ABSOLUTE's output consist of the Absolute copy-number of pre-defined segments and the number of mutated alleles at SNPs. ABSOLUTE is specifically designed to work with only tumour sequencing data, and its algorithm does not include a segmentation analysis. ABSOLUTE is one of the most widely used CNV callers in cancer genomics as it is compatible with both array and sequencing data, and works with whole genome and exome sequencing segmented copy-number data.

State-of-the-art copy-number calling algorithms often combine various effective methods to yield high quality copy-number calls and in most cases to minimise the trade off between sensitivity and specificity of their results. Incorporating various input signals (e.g., BAF and read-depth ratios by ASCAT and PURPLE) to gather evidence is one strategy for copy-number calling algorithms to take advantage of all available inputs from sequencing data. Another strategy is to adopt an ensemble approach and choose best performing copy-number callers based on input data and user or experiment requirements. Moreover, CNV calling algorithms in variant calling pipelines that contain connected clusters of genomic rearrangements including CNVs, also takes advantage of multiple sources for more accurate CNV calling. For instance, PURPLE CNV caller in GRIDDS, PURPLE and LINX [28] variant calling pipeline combines high quality SV calls from GRIDDS in the form of breakpoints with VAF values and read depth ratios to detect copy-number change along breakpoints and infer CNVs with single nucleotide resolution. LINX is a visualisation tools to track complex variations across chromosomes in a Circos [44] plot.

TITAN [45] infers CNVs including segments with LOH by using a probability model and taking into account all the three possible cell types (i.e., normal cells, aberrant and non aberrant containing cells) in a sample. Both read-depth and VAF values at SNP sites are utilised by TITAN, and assuming co-occuring of all cell types in each clone allows for using a clustering statistical approach which increases statistical power for weaker signals while inferring CNVs. Having a HMMs framework which takes advantage of adjacent loci along a sequence, TITAN is capable of maximising available statistical strength to take advantage of signals with varying strengths for inference. TITAN's performance is evaluated and benchmarked using multiple truth sets including *in-silico* engineered sample mixture analysis by simulation. As a result of powerful statistical modelling and use of multiple signals, TITAN's algorithm is more sensitive to sub-clonal events and performs better at de-convoluting available signals from sequencing data.

Many variant callers are capable of inferring sequencing segment copy-number profiles from impure samples. However, accurate estimation of sub-clonal purity values are needed for estimating copy-number of sub-clonal cell populations. This information is valuable for getting insight into tumour evolution and analysing phylogenetic tree of cancer cells. Sclust [25] is an allele-specific copy-number caller with the ability to infer clonal and sub-clonal specific copy-numbers. Sclust's copy-number analysis module performs copy-number inference and mutation clustering with high computational efficiency (mean runtime of <10 min for tested datasets). Read-depth ratios along with *in-silico* reconstructed VAF values at SNP points are model inputs for Sclust. A piecewise constant function is fitted to read-depth ratios and segmentation of genome is performed over the following three steps: 1) initial coarse segmentation to approximate purity and ploidy values (a process similar to CBS), 2) sensitive segmentation with over-segmentation, and 3) forming new segments from adjacent segments with equal copy-number value. The performance of Sclust was compared to Theta [46] and Battenberg (unpublished, available from `https://github.com/cancerit/cgpBattenberg`) using breast cancer genome PD4120a [47] dataset in [25] study; Battenberg, Theta [46] and Sclust all produced similar results for tumour purity estimates.

Most CNV calling tools are not designed to work with low quality tumour sequencing data often derived from low quality samples. Hence, a large amount of low quality clinical data (i.e., data with low purity and low coverage values) is discarded due to lack of specialised computational methods capable of accurate variant calling. Weak variant detection signals, and insufficient statistical power are main challenges for inference from such data. Accucopy [48] is an allele-specific CNV caller that works with low purity and coverage (as low as 2X) data. Accucopy combines information from both read-depth log ratios and allelic coverage ratios that are utilised in multiple statistical techniques and implemented into a complex probabilistic inference model. Accucopy's developers [48] claim to achieve better call accuracy on simulated and real data in comparison to ABSOLUTE, Sequenza, and Sclust. Superior performance of Accucopy results from reducing false positive heterozygous SNP calls with Strelka2 [49], use of simulated and available data to fine tune the inference model, and reducing coverage noise by smoothing.

Strelka2 [49] is a mixture-model-based indel caller, specialised in small-segment variant calling with high accuracy and computational efficiency. Mutation and error rate of indels are adaptively estimated by a mixture-model which makes Strelka2's indel calling context specific and increase accuracy of calls. Runtime of probability model in Strelka2 is improved by pre-computing read likelihoods from candidate alignments and the maximum alignment likelihood and therefore, avoiding computationally more costly models such as pair HMMs. Additionally, Strelka2 uses a random forest implementation to increase precision of calls by accounting for error not considered in the probability model, and auto tuning model parameters. Strelka2 demonstrated good performance for very short (larger than 1 base) indel calling, however, the datasets used for benchmarking do not reflect the extensive complexity of some tumour data with short-segment clustering events.

PURPLE [28] is a whole genome allele-specific copy-number caller that takes read ratios (derived from read-depth signal) and SNP BAF values as input to estimate purity and ploidy of samples. AMBER and COBALT algorithms supply BAF and read depth ratios to PURPLE respectively. AMBER generates BAF values for likely heterozygous SNP sites from variant call format (VCF) files. Knowing the likely SNP sites along the genome can then serve as evidence for reliable breakpoint detection. Furthermore, boundaries of copy-number segments can be adjusted using optional input of SV breakpoints. COBALT calculates read depth ratios of tumour and reference genome by segmenting the genome to 1000(b) windows and counting the reads in each window after performing quality control and GC normalisation. PURPLE uses Bioconductor's 'copynumber' package which performs piecewise constant fit on read depth ratio and BAF segments separately. Further smoothing is then performed on segmented data to reduce segmentation caused by noise which are unlikely to contain CNVs. Taking precise breakpoint variant positions from a single-base resolution SV caller as input is a unique optional feature in PURPLE. Moreover, PURPLE is one of the few copy-number callers that mentions read-depth signal loss in short-segment CNVs in its documentation and attempts to infer small CNVs by utilising precise breakpoint positions and VAF values at SV breakpoints. PURPLE is designed to be integrated in a single-base resolution variant calling pipeline such as GRIDSS, PURPLE, and LINX pipeline or input SV calls from Strelka, hence, it is not often run in isolation.

CNVnator [50] is a highly cited read-depth based copy-number caller which is used in both cancer and population genomic variation studies. CNVnator is one of the earliest read-depth based callers capable of both whole genome copy-number analysis and population genotyping. In addition to a mean-shift model, CNVnator is also calibrated using 1000 Genomes Project to detect abnormal CNVs for population genotyping. Furthermore, having a multiple bandwidth partitioning incorporated into its algorithm, a diverse range of copy-number segment sizes can be detected, with the shortest reported as a few hundred bases with relatively high breakpoint resolution. Although, this caller is primarily designed to detect germline copy-numbers, it can detect somatic copy-number changes where aberrant cell fraction is above 50%. However, it is not capable of inferring copy-number neutral events. CNVnator is frequently included in benchmarking studies to compare read-depth based methods performance with other approaches.

CNVpytor [51] is CNVnator's recent successor, which is reported to be faster, more efficient, and with the added ability to incorporate BAF values from known SNP sites to increase novel CNV detection rate. CNVpytor takes advantage of a mixed signal approach (i.e., read-depth and BAF values) to achieve better accuracy with typically much lower run times. There are also added visualisation features such as variant chromosome plots (similar to Circos [44] plots) showing copy-number variations across all chromosomes as well as quality control and relevant summary statistics graphs. CNVpytor provides a simple platform for copy-number analysis and it expands on flexibility of CNVnator which is used for copy-number analysis in various species, paired normal-tumour cancer studies, and analysis of population genomics. Additionally, it is capable of inferring copy-number neutral events.

Recent benchmarking studies of CNV callers suggest combining callers improves the detection rate of variants [20, 21]. Having a complementary set of detection signals such as read-depth, read-pairs, and split-reads contributes to gaining accuracy and sensitivity in calling. This is due to complementary nature of CNV detection signals. For instance in duplication events read-depth signal is naturally stronger as the number of reads in amplification segments are higher. In copy-number reduction events (i.e., deletion cases) however, read-depth signal is negatively impacted; and in turn, read-pairs may be used to detect deletion sides more accurately, specially in small deletion segments where read-depth signal is effectively lost. It is also reported that read-pair-based callers detect significantly higher number of CNVs 1 - 50(kb) in length compared to other methods [20]. Nonetheless, methods that are not based on read-depth detect higher number of CNVs in total. Consistent with the assumption that small deletion segments are captured more effectively by non-read-depth based methods, Delly [52], a read-pair based caller was reported to detect the most number of deletions (50(b) - 1(kb) in size) in [20] benchmarking study. Second highest small deletion detection rate was reported for Manta [53], which also utilise a combination of read-pairs and split-reads. Interestingly Delly and Manta also detected the highest number of amplifications 50(b) to 1(kb) in size. In turn, the detection rate of read-depth based CNV callers was highest in both deletions and amplifications with varying length between 1 - 50(kb) [20]. Small deletions are inferred with higher confidence using algorithms that utilise alternative to read-depth detection signals, compared to small amplifications. Hence, choosing the most suitable signal based on variant sizes may lead to better detection rate, accuracy, and more reliable calls. In [20] study however, call specificity of non-read-depth-based tools are reported considerably high with a significant number of false positive and doubtful calls recorded for 1 - 5(kb) amplification segments. Overall results of benchmarking studies [20–23] show that callers based on similar signals, perform similarly and this can be leveraged to take a targeted or ensemble approach for variant calling based on variant types and sizes. Moreover, taking advantage of multiple callers that utilise various detection signals is likely to improve call qualities and contribute to capturing larger sets of variants. Regardless of the approach, the trade-off between specificity and sensitivity need to be considered based on the study aims and needs.

3.5 Conclusion

Recently developed copy-number callers recognised the advantages of utilising variant specific signal, combining available signals to increase statistical power, and effective visualisation of variants. In cancer genomics, studies that focus on emerging ensemble-based approaches [54, 55], report superior performance compared to running copy-number callers in isolation and suggest ways to minimise shortcomings of one tool by overlapping results derived from multiple copy-number callers. For instance, if the results from a read-pair based caller contain a large number of false positive cases in small amplification cases, calls from a read-depth based caller can be used instead. Similarly, strength of read-pair based callers can be used to compensate for lack of accuracy in other variant calling scenarios.

As reported by PURPLE [28] (section 5 of PURPLE's documentation), in compound short-segment variations read-depth signal is often lost and alternative evidence such as VAF values at SV breakpoints can be utilised for copy-number inference. Furthermore, CNV inference algorithm literature is populated with a range of simple to complex inference methods, which include a diverse algorithmic approach for utilising available variant detection signals and breakpoint evidence. However, majority of copy-number inference algorithms rely on essentially only read-depth based methods. Therefore, introduction of a robust alternative copy-number inference signal such as VAF at SV breakpoints (SV-VAF) can positively impact the design and methodology of upcoming CNV callers. This is particularly applicable for copy-number inference in short variant segments.

Table 1: Description of inference algorithms and detection signals of select variant callers.

Variant Caller	**Signal**	**Inference Algorithm/Method**
ASCAT [36]	Read-depth	LOH analysis. GC and read count normalisation along with segmentation at SV breakpoints. Then read-depth is utilised to analyse copy number change across segments.
CNVkit [40]	Read-depth	Implemented CNV detection pipeline capable of exome-level on-target and high-resolution off-target region copy number inference utilising normalised coverage values.
Sequenza [38]	Read-depth and VAF	A probabilistic model based on a maximum a posteriori approach which takes two copies as preferred prior probabilities and infers copy number based on average depth ratio and VAF parameters at segment's level with a grid-based search.
ABSOLUTE [37]	Read-depth	Purity and ploidy of samples are calculated from input copy-number segments. Assigns clonal and sub-clonal copy-number states using a probabilistic model.
PURPLE [28]	Read-depth and SV-VAF	SV-VAF input values from AMBER and read-depth ratios input by COBALT. CNVs are inferred primarily using read-depth ratios. Small-segment CNVs are detected and inferred using SV-VAF values.
TITAN [45]	Read-depth and VAF	Using a probability model for inferring CNV and LOH events in normal cells aberrant and non aberrant containing cells in a sample.
Sclust [25]	Read-depth	Piecewise constant fitting function on read-depth ratios and performs sensitive and multi-step copy-number segmentation.
Theta [46]	Read-depth	Considers samples with tumour heterogenious tumour-normal admixture and selects an optimal solution for maximum likelihood mixture decomposition problem.
Accucopy [48]	Read-depth and allelic coverage ratios	Statistical learning from both read-depth and allelic coverage ratios used as input for an allele-specific probabilistic copy-number inference model.
Strelka2 [49]	Read-depth and read alignment	A mixture-model with pre-computed read likelihood of alignments.
CNVnator [50]	Read-depth	Mean-shift model for copy-number segmentation and pre-calibrated using 1000 genome project.
GRIDSS [27]	Split-reads - read-pair mapping and assembly	Single-nucleotide breakpoint resolution is achieved by a combination of breakend assembly and discordant paired-end reads and split-reads analysis.
CNVPytor [51]	Read-depth and VAF	Mean-shift model for copy-number segmentation with addition of VAF values at SNPs to increase detection rate.
Delly [52]	Split-reads and read-pair mapping	Mapping analysis of read-pairs and breakpoint position retrieval from split-reads.
Manta [53]	Split-reads and read-pair mapping	Breakend graph assembly and analysis

4 Methodology

4.1 Model Development

We designed two theoretical implementations of copy-number inference models based on read-depth signal and SV-VAF with naive initial assumptions, and gradually increased the complexity and accuracy of our models by adding features and accounting for real input sequencing data complexities. The read-depth based model was implemented using change-point method and the initial model was inspired by SegSeq [39] CNV caller due to its simple copy-number inference model design. To compare the suitability and performance of SV-VAF and read-depth signals in short CNV segments, we designed our models to work with only one primary inference method, either read-depth or SV-VAF. We then evaluated their performance over various simulated datasets.

The number of reads (i.e., read count) for each segment is the primary input for a read-depth based copy-number inference model. Read-depth signal is normally generated by calculating the number of bases covering a given locus. However, since we know the true copy-number value of a given segment in our simulations, we estimate the read counts assuming read count values of segments follow a probability distribution. In both models (i.e., read-depth and SV-VAF based copy-number inference models), the value of read counts in each segment is considered to follow a Poisson distribution, and the expected read count was calculated from the following formula:

$$\lambda = \frac{CN \times C \times EL}{L_{read\ length}} \tag{1}$$

$$\begin{aligned}
\lambda &: expected\ number\ of\ reads \\
CN &: copy\ number\ of\ the\ current\ segment \\
C &: sequencing\ coverage \\
EL &: effective\ length\ of\ current\ segment \\
L_{read\ length} &: read\ length
\end{aligned}$$

Above formula (1) provides flexibility to calculate expected read counts for both the whole genome and a specified region. This formula is used to calculate the expected number of reads in copy-number calculations by read-depth model. Moreover, read count values are also needed to calculate SV-VAF value (formula (6)). Therefore, in SV-VAF model, formula (1) is used to calculate the number of both SV and reference supportive reads. We considered two approaches for counting reads in a segment for our read-depth models. These models were labelled as read-depth ‘start-in’ and read-depth ‘overlap’ models.

4.2 Read-depth ‘start-in’ Model

Simulated sequencing reads that have their starting base position aligned within a segment are considered start-in reads. Based on that, read-depth ‘start-in’ model only counts start-in reads. This is the simplest implementation of read count values for utilising in read-depth based models to infer a segment’s copy-number. Read counts have a direct positive correlation with segment length. Hence, shorter segments contain lower read counts. We define the effective length as the length that is used to evaluate the number of bases present in a segment. The effective length is either the actual segment length value or summation of segment length and read length together. Consequently, the effective length is considered only the segment length when inferring copy-number by read-depth 'start-in' model. In other words, ($L_{effective}$) in formula (3), and EL in formula (1) take the value of segment length. This will result in the shortest current segment length consideration among our copy-number inference models. Most read-depth based CNV calling algorithms use this approach for copy-number inference.

4.3 Read-depth 'overlap' Model

We considered reads that share any aligned bases with boundaries of a segment as overlap reads. Read-depth 'overlap' model counts both reads that start-in and overlap a segment to calculate a segment's copy-number. This effectively increases the read count value to boost read-depth signal. Furthermore, counting overlapping reads is equivalent to extending segment's length by read-length. Based on that, read-depth 'overlap' model considers effective length as segment length plus read-length ($L_{read\ length}$); $L_{effective}$ in formula (3), and EL in formula (1). Read-depth 'start-in' and 'overlap' models both use formula (3) to calculate copy-number of a segment, CN_{RD}. Read counts in a segment, sequencing coverage, and read length were used to compute the copy-number of a segment using read-depth by the following formulas:

$$n_{bases} = n_{reads} \times L_{read\ length} \tag{2}$$

$$CN_{RD} = \frac{n_{bases}}{C \times L_{effective}} \tag{3}$$

$$\begin{aligned} n_{bases} &: the\ number\ of\ bases\ covered\ by\ reads \\ CN_{RD} &: copy\ number\ calculated\ from\ read - depth \\ n_{reads} &: the\ number\ of\ reads\ in\ a\ segment \\ L_{read\ length} &: read\ length \\ C &: coverage\ value \\ L_{effective} &: effective\ read\ length \end{aligned}$$

4.4 SV-VAF Models

Copy-number calculation method from SV-VAF (i.e., VAF of an SV breakpoint), depends on the number of reads (represented by varying widths of coloured rectangles in figure (3)) and SV orientation of the current segment with respect to the flanking segment figure (3). Given the following equations:

$$Decreasing\ orientation : CN_{flanking\ segment} = CN_{current\ segment} + CN_{SV} \tag{4}$$

$$Increasing\ orientation : CN_{current\ segment} = CN_{flanking\ segment} + CN_{SV} \tag{5}$$

Since we only use SV-VAF values and not VAF at SNPs, we refer to SV-VAF in our models and calculations simply as VAF from here onwards in this project. Total copy-number across a breakpoint remains constant, and VAF value is used to calculate the current segment's copy-number. We assumed there are two types of reads at a breakpoint: reads that contain supportive evidence for presence of an SV, $n_{SV supportive}$; and reads that support the reference allele, $n_{Ref supportive}$. VAF value at an SV breakpoint is calculated by the following formula:

$$VAF = \frac{n_{SV supportive}}{n_{Ref supportive}} \tag{6}$$

$$\begin{aligned} n_{SV supportive} &: number\ of\ SV\ supportive\ reads \\ n_{Ref supportive} &: number\ of\ reference\ supportive\ reads \end{aligned}$$

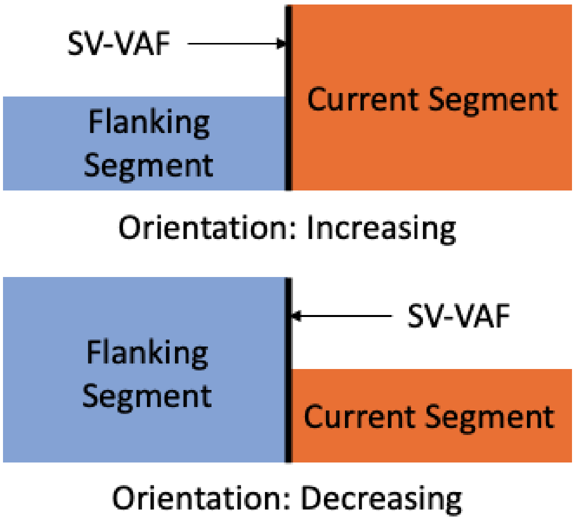


Figure 3: *SV Orientation to calculate copy-number of current segment. Widths of coloured rectangles represent the number of reads present in the corresponding segment.*

VAF model calculates the current segment's copy-number based on the orientation of SVs, using below formulas:

When SV orientation is increasing:

$$CN_{current\ segment} = \frac{CN_{flanking\ segment}}{1 - VAF} \tag{7}$$

When SV orientation is decreasing:

$$CN_{current\ segment} = CN_{flanking\ segment} \times (1 - VAF) \tag{8}$$

$CN_{FlankingSegment}$ is calculated by CN_{RD} formula (3) as our implementation of VAF model takes the flanking segment's copy-number as an input, formulae (7 and 8). The evidence for presence of an SV is captured within reads. Consequently, we assumed the size of an SV is determined by SV supportive reads. Hence, the interval length used for read alignment at a breakpoint can either be just read length ('split-read') or fragment length ('read-pair'). In 'split-read' case, SVs are detected just by split-reads, and maximum length of an SV is limited to just the length of that read, $L_{read\ length}$. Additionally, SV's length can also be considered as read-pair which is the library fragment size ($2 \times L_{read\ length} + read\ inner\ distance$). As a result, two approaches for copy-number inference using VAF models were labelled as VAF 'split-read' and VAF 'read-pair'. Effective length (EL) for expected reads in formula (1) is then subsequently considered as read length ($L_{read\ length}$) for VAF 'split-read' model, and fragment size for VAF 'read-pair' model. Our theoretical models were implemented for simulation in R and all simulations were carried out using RStudio. Our code and data is available on the following GitHub repository: `https://github.com/MoMoTPark/masters_thesis`

4.5 Simulation

We developed our simulation code in R programming language (version 4.1.2) using RStudio (version 1.4.1717). We used a combination of provided tools from both 'tidyverse' library and R's base package for code development. A test driven development (TDD) approach was used for developing simulation code in R; where small test case scenarios were first written and the concept code was tested throughout the development of the simulation procedures.

We designed the simulations to generate 20,000 (10,000 inferred copy-number values per model) copy-number observations for each simulation run with the following input parameters:

Read length: The length of sequencing reads

Coverage: Sequencing coverage of a given locus

Flanking segment length: The length of flanking sequence (flanking segment is located adjacent to the current segment, and considered the segment with a known copy-number value, figure (3))

Current segment length: The length of current segment (copy-number of current segment is unknown and we are interested to infer copy-number value of this segment)

Actual/known flanking segment copy-number value: Copy-number value of the flanking segment

Actual/known current segment copy-number value: Copy-number value of the current segment given as a reference to calculate confidence interval from inferred and actual copy-number values.

Fragment size: Fragment length of paired-end sequencing reads that contains both sequencing reads plus inner distance length of reads

Read-depth based model: Read-depth model to be used to infer copy-number of current segment from read-depth signal

VAF-based model: VAF model to be used to infer copy-number of current segment from VAF values

We generated four read datasets per simulation: SV supportive reads, reference supportive reads, reads located in the flanking segment, and reads located in the current segment. Read datasets were generated from a Poisson random variable generator '`rpois(10000, λ)`' in R; λ calculated using formula (1). Copy-number of each segment was then inferred using segment specific read dataset values. In read-depth based models, these read datasets were used to reconstruct read-depth signal and calculate copy-number values from formulas (2) and (3). Similarly, in VAF model, after calculating VAF from formula (6), copy-number of current segment is inferred from formulas (4, 5, 7, 8).

We first simulated a generalised dataset to analyse the confidence interval (CI) trend of our models and evaluate their performance given various copy-number states. Our simulated dataset contained 969,600 observations and for each datapoint contained 9 input parameters. We also calculated VAF, and model specific CI width values, representing a 95% CI, from inferring 10,000 copy-number values per model in each simulation run. Copy-numbers were generated by considering all possible permutations of the following parameters:

Read length: 150(b)

Coverage: 60x

Flanking segment length: array of length values [from:10, to:10000, by:100]

Current segment length: array of length values [from:0, to:1000, by:10]

Actual flanking segment copy-number: array of copy-number values [1, 2, 3, 4]

Actual current segment copy-number: array of copy-number values [1, 2, 3, 4]

Fragment size: 600(b)

Read-depth based model: both 'start-in' and 'overlap'

VAF-based model: both 'split-read' and 'read-pair'

Then multi-facet style (correlation) plots (section 5.2) were generated to investigate the impact of current and flanking segment length values on performance of our models. We also replicated above simulation with the exception of changing flanking-segment-length parameter to be a constant value of 10,000(b) to confirm expected behaviour of our models (section 8.1).

Plots for CI and VAF values along with a wide range of copy-number states for a single current segment length were generated to investigate the effect of amplification and deletion events on predictive power of VAF-based model. The following 8 input parameters were used to generate a dataset for these plots (figures (7 and 22)):

Read length: 150(b)

Coverage: 60x

Flanking segment length: array of length values [100, 10000, 10000]

Current segment length: 50(b)

Actual flanking segment copy number: array of copy-number values [from:0, to:4, by:0.2]

Actual current segment copy number: array of copy-number values [from:0, to:4, by:0.2]

Fragment size: 600(b)

VAF-based model: 'read-pair'

Performance similarity and dissimilarity of our models were investigated by looking at cross-over points (i.e., range of datapoints where both models perform similar). Therefore, we dedicated the largest simulated dataset consisting of 1,438,200 observations to distinguish the characteristics of parameter values and the type of events that result in different performance of our models. We generated this dataset by permutation of the following parameters:

Read length: 150(b)

Coverage: 60x

Flanking segment length: array of length values [from:50, to:10000, by:100]

Current segment length: array of length values [from:0, to:1000, by:20]

Actual flanking segment copy number: array of copy-number values [1, 2, 3, 4, 5, 10, 20, 30, 50, 70]

Actual current segment copy number: array of copy-number values [from:1, to:10, by:1, and added values 20, 30, 40, 55, 70]

Fragment size: 600(b)

Read-depth based-model: 'start-in'

VAF-based model: 'read-pair'

Observations with model CI difference of less than 0.05 (overall significance level) were considered similar in performance. Basic descriptive statistics of parameters for deletion and amplification cases were visualised. We also investigated the specific deletion cases where VAF-based model performed worse than read-depth based model, and amplification cases where read-depth based model performed worse than VAF-based model figures (8, 9, 10, and 11). Furthermore, we generated a heat map (figure (12)) in order to simultaneously visualise the impact of flanking and current segment length parameters on CI difference of inferred copy-numbers on a wide range of copy-number values. Only the subset of observations with significant performance gap between the two models, defined as absolute CI difference of more than 0.05, were used in the heat map. Our metric to highlight regions with significant performance gap consisted of the log of absolute CI difference between the models ($log\ |CI\ width\ _{read-depth\ model} - CI\ width\ _{VAF\ model}|$). We then investigated regions with the most performance difference between models. These regions were highlighted in figures (12 and 23) with log CI difference values of above 0 (i.e., more than 1 absolute CI value difference). We further filtered our heat map dataset into two subsets containing observations with one model always performing better than the other and plotted descriptive statistics of flanking and current segment lengths, and copy-numbers for each dataset as box and whiskers plots, figures (13, 14, 15, and 16).

4.6 Model Performance Measurement

The number of reads (n) generated from a Poisson distribution (where parameter λ is considered expected number of reads at any given locus) is approximately equal to total number of inferred copy-numbers for each simulation run. For instance, if we choose to simulate reads for current segment of only 150(b) in length, then approximately n copy-numbers are generated for this specified current segment length along other parameters which are constant values. Since a large array of inferred copy-numbers for each simulation run is available, 95% CI of this array was calculated from subtracting two inferred copy-number values stored in approximately 97.25^{th} and 2.25^{th} percentile location of the array. This value provides a general and agnostic performance metric to compare model performance. CI value of copy-numbers in each run then acts as one datapoint for producing graphs with a Y axis which indicates CI width values and an X axis that represents any other chosen model parameter (e.g., current segment length).

Each simulation run for a given range of constant model parameter values typically result in two CI values, one for the chosen read-depth model, and another for SV-VAF model. These values are then directly compared to represent the performance of each model; smaller CI values in a model indicate better model performance as inferred copy-numbers vary less and therefore, have a lower standard deviation.

4.7 Variant Calls on Real Sequencing Data

We intended to replicate a paired tumour-normal sample variant calling analysis study from real sequencing data. Given we are looking at novel and rare short-segment CNVs, we replicated tumour samples by adding pre-defined variants to a normal sample data. This way tumour sample is derived directly from normal sample which contains all germline variants in the normal sample plus added CNVs. Consequently, we were able to input our paired tumour-normal samples to variant callers that are capable of somatic variant calling and filter for implemented CNVs. Detection rate and copy-number inference accuracy of calls were considered as primary metrics for evaluating the performance of utilised variant callers.

Sample data: Our normal sample consisted of whole genome sequencing data from chromosome 21 of Ashkenazim trio son sample (HG002, Illumina HiSeq 60x)(section 8.3); described at the

'Genome in a Bottle' (GIAB) consortium [56]. We only used chromosome 21 (chr21) slice of HG002 data to reduce run-times and computation cost. We replicated tumour sample sequencing data by adding pre-defined somatic variants to known relatively variant-free segments on our normal sample using `BAMSurgeon` [57].

We surveyed suitable relatively variant-free regions of our sample data with IGV and designed 3 distinct variation clusters and 1 large-scale short-segment CNV containing sample in different regions. Each specified truth set below was implemented as a separate aligned sequencing data tumour sample (in bam format):

Truth set 1:

Deletion cluster position: chr21: 42,541,540 - 42,544,300

Truth set 2:

Amplification cluster position: chr21: 13,306,000 - 13,316,500

Truth set 3:

Mix of deletion and amplification cluster position: chr21: 27,677,600 - 27,688,100

Truth set 4:

Copy-number variation integration per 10(kb) bins position: chr21: 14,106,800 - 34,372,600

`BAMSurgeon` [57] is a somatic variant replicating tool for simulating tumour sequencing data to benchmark variant calling tools. We created compound variant equivalent deletion and amplification clusters, with variants ranging between 50 - 1000(b) in size in truth sets 1 and 2 (consecutively positioned). Our mixed variant clusters (truth set 3) consisted of various deletions and amplifications ranging between 55 - 9000(b) in size (non-consecutive positioned). Truth set 4 tumour dataset contained the most variants, and included 50 - 1000(b) variant size range. `BAMSurgeon` (version 1.3) was used to generate 4 tumour replicating sample sequencing datasets (output as bam files)(cluster positions listed above) that contained variant truth sets added to existing chromosome 21 of HG002 Ashkenazim trio son (i.e., normal sample) aligned sequencing read data (bam file) for benchmarking our variant calling pipelines. `BAMSurgeon` provides a variant call format (vcf) output which we used as reference for true position and size of newly incorporated variants. Apart from added new variants, our tumour sequencing samples were identical to chromosome 21 of the son (HG002) in Ashkenazim trio from GIAB study [56]. Furthermore, all generated tumour bam files were sorted and indexed with `samtools` (version 1.15.1), and variant call results (outputs formatted as vcf) were analysed with `bcftools` (version 1.15).

GRIDSS (version 2.13.2) was used as our primary SV caller to detect precise breakpoint locations of pre-existing and implemented variations in our sample data. We used GRIDSS with default parameters and provided exclusion regions (bed format) for our specific sample rather than using a generic reference specific exclusions file. Total SV call outputs were then filtered to only include somatic SVs by provided somatic-filtering tool in GRIDSS package. Filtered SVs were also further divided into high-confidence and high-and-low-confidence subsets which were utilised as input for PURPLE. PURPLE calls were generated using default (recommended) parameters with additional SV calls input from GRIDSS and outputs from AMBER (version 3.9) and COBALT (version 1.13). Calls from both GRIDSS and PURPLE were compared to the truth set using a custom Python script. Accuracy, detection rate, and detection rate for CNVs under 1000(b) in length were our metrics for comparing callers. Given CNVpytor's simple copy-number analysis workflow, the ability to infer various copy-number segment sizes with high accuracy and speed, and flexibility to work with a different range of genomic data, we chose to also perform a copy-number analysis with CNVpytor. Knowing our simulated results, we aimed to test our hypotheses using adjusted real sequencing data and run multiple variant calling analysis workflows. In addition, this provides a benchmark for length threshold of some frequently used variant callers.

We inspected truth set 4 positions (length of 20,265,800(b)) with IGV to check for existing coverage. Then segmented this section into 10(kb) non-overlapping bins, and added a CNV (i.e., amplification or deletion) approximately positioned in the middle of each bin. Generated tumour sample in truth set 4 contained 1658 variation events, each within a 10(kb) long bin. We used our described pipeline to run GRIDSS and PURPLE callers on this dataset. However, we did not use CNVpytor to call CNVs in this dataset. Added variation event counts in truth set 4 are listed below:

Total added CNV events: 1658

Total number of added duplication events: 803

Total number of added deletion events: 855

Total added CNV events under 1000(b) in length: 1354

The number of added duplication events under 1000(b) in length: 647

The number of added deletion events under 1000(b) in length: 707

We developed various Python scripts to automate analysis of the large number of added CNVs implemented in truth set 4. Our scripts compared the number and type of calls made by GRIDSS and PURPLE to our true variant records file (i.e., vcf file output by `BAMSurgeon`). We then performed a number of validation test case scenarios to check for correct behaviour of our developed Python code. The scripts successfully extracted and filtered variants by their size and type; moreover, we counted the number of variation events called by each caller and calculated corresponding caller metrics.

We were also looking to detect and characterise any short segment copy-number clusters that are present on the sample data; to further investigate whether these clusters can be detected with high performing variant callers. Variant calls and records of added variants for each truth set are available from:
`https://github.com/MoMoTPark/masters_thesis`

CNVpytor (version 1.0) has a combined call option (currenlty in prototype stage), which takes VAF values of SNP positions to increase the accuracy and detection rate of CNVs. Combined call feature utilises both read-depth signal and VAF values, which ideally would result in improving the resolution of copy-number events. Hence, in addition to running CNVpytor's default model (i.e., read-depth only model), we also performed a combined call on all clusters. We prepared SNP calls for our clusters and normal sample using `bcftools` (version 1.15) `mpileup` and `call` with multi-allelic model. SNP calls were filtered and used as input for combined calls in CNVpytor.

5 Results and Discussion

5.1 Performance Metrics

Inferred copy-number values of four samples with current segment length sizes of 50(b), 100(b), 150(b), and 200(b) were plotted in figure (4) to demonstrate and confirm that CI metrics are behaving as expected and correctly indicate the performance of each model without bias. Each box and whiskers plot (i.e., one box and whiskers plot per model for a given current segment length), contains 10,000 estimated copy-number values for the given current segment size. Interquartile range along with outliers are shown for each model. The interquartile range represents the CI width, which is the difference between 97.25th and 2.25th percentiles (as described in section 4.6). Furthermore, it confirms that reported CI values that are used as model performance metrics, are half CI length of inferred copy-numbers. For example if the current segment copy-number is inferred as 1 with a reported CI width of 1, then 95% CI equivalent range of this datapoint is between 0.5 and 1.5 (i.e., the actual copy-number value can vary between 0.5 and 1.5). Figure (4) also demonstrates that noise and the number of potential outliers produced are higher for lower current segment length sizes in read-depth model. Further indicating that as the current segment length is reduced, read-depth signal becomes noisier, resulting in larger CI values for the inferred segment copy-numbers.

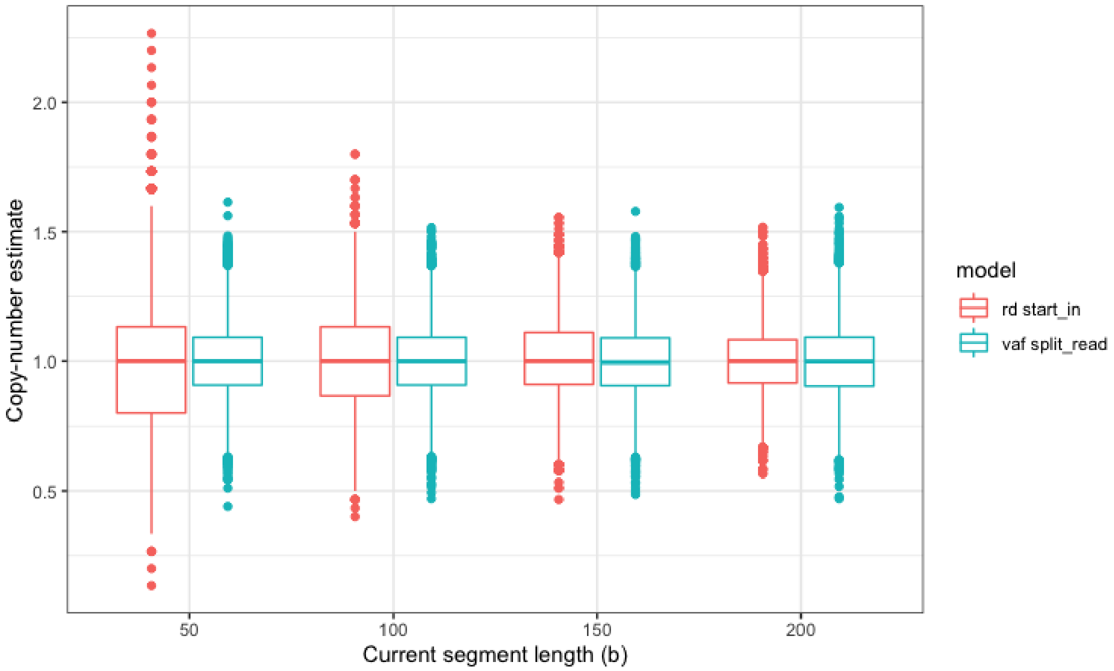


Figure 4: *Box and whiskers plots of estimated copy-number values from current segment length of different sizes.*

5.2 General Trend of Performance Across Multiple Copy-number States

We consistently observed that higher copy-number values result in worse performance of our models. Regardless of chosen parameters and model inference approach, higher copy-number values negatively impact CI of inferred copy-numbers because they amplify random error produced during random variable generation in our simulations. This was expected as the number of observed reads for each segment is a product of read counts and read-length (formula (3)) which are directly affected by read-depth, namely, the copy-number of a segment. Hence, the error is amplified further with higher copy-number values and CI width is increased as a result.

The effect of copy-number increase is less apparent in VAF-based model, as only the copy-number of flanking segment is calculated with read-depth signal. This does not apply to read-depth based model as copy-number of current segment are inferred using read-depth signal and therefore, significantly influenced by copy-number values.

As it is demonstrated in figure (5), current segment length impacts the performance of read-depth 'start-in' model drastically (this also extends to read-depth 'overlap' model, described in section 8.1, figure (20)). And as expected, VAF model is unaffected by current segment length parameter with CI values remaining constant across all length values. Read-depth based model's performance (figure (5), represented by red lines) degrades while fluctuating as current segment length is reduced. Slight fluctuations seen are the result of having multiple flanking segment values for a single current segment length, given parameters are permutated over the entire specified flanking segment length during simulation. This then contributes to having a non-linear decrease in read-depth model's performance and observing small peaks. Therefore, we do not interpret such fluctuations as true representation of read-depth model performance as we could demonstrate, peaks can be minimised or eliminated if a constant flanking segment value is chosen for all simulations in figure (19) (section 8.1). Given our specified parameters for presented results in figure (5), read-depth model's performance drastically declines when current segment length goes below 175(b). Similar behaviour can be seen with different parameter sets and the threshold for rapid decline of read-depth model's performance is current segment length values of below 200(b).

Indicated performance of VAF model also varies across different flanking segment length values. As mentioned, in our read-depth based models, current segment's copy-number is inferred using observed bases in the segment, which is directly influenced by the copy-number of current segment. Nonetheless, read-depth models remain unaffected by flanking segment length parameter. This can be confirmed looking at the performance of two models in figure (6). CI values of read-depth model remain unchanged across all given flanking segment values. On the other hand, copy-number of flanking segment in VAF-based model is calculated using read-depth based copy-number inference method and therefore, flanking segment's length affects the performance of VAF-based models. Similar to fluctuations observed in read-depth model (red line in figure (5)), VAF-based model fluctuations (cyan line in figure (6)) are not considered to reflect true performance. This is because at a given flanking segment length, multiple inferred current segment copy-number values exist which cause these fluctuations shape across the graphs.

Although, various model performance cross-over boundaries were observed between read-depth and VAF based models in our correlation plots, we were mainly interested in investigating the general trend of model performance (using CI values as proxies for model performance) given current and flanking segment length values. Having multiple parameters influencing the CI of the inferred copy-numbers, we relied on generating additional plots and summary statistics of our simulated datasets to further analyse trends and possible performance cross-over thresholds.

The performance of VAF model has a direct positive correlation with the flanking segment length parameter values indicated in figures (6 and 7). As expected, model's performance declines with lower flanking segment values, however, the behaviour of the model over different VAF values given copy-number of flanking and current segments is our main interest for analysis of plotted values in figure (7). We observed that predictive power of VAF model is split based on amplification and deletion cases. Namely, regardless of a given dataset, even if flanking segment's length is less than current segment's length (demonstrated in section 8.1, figure (22)), VAF-based

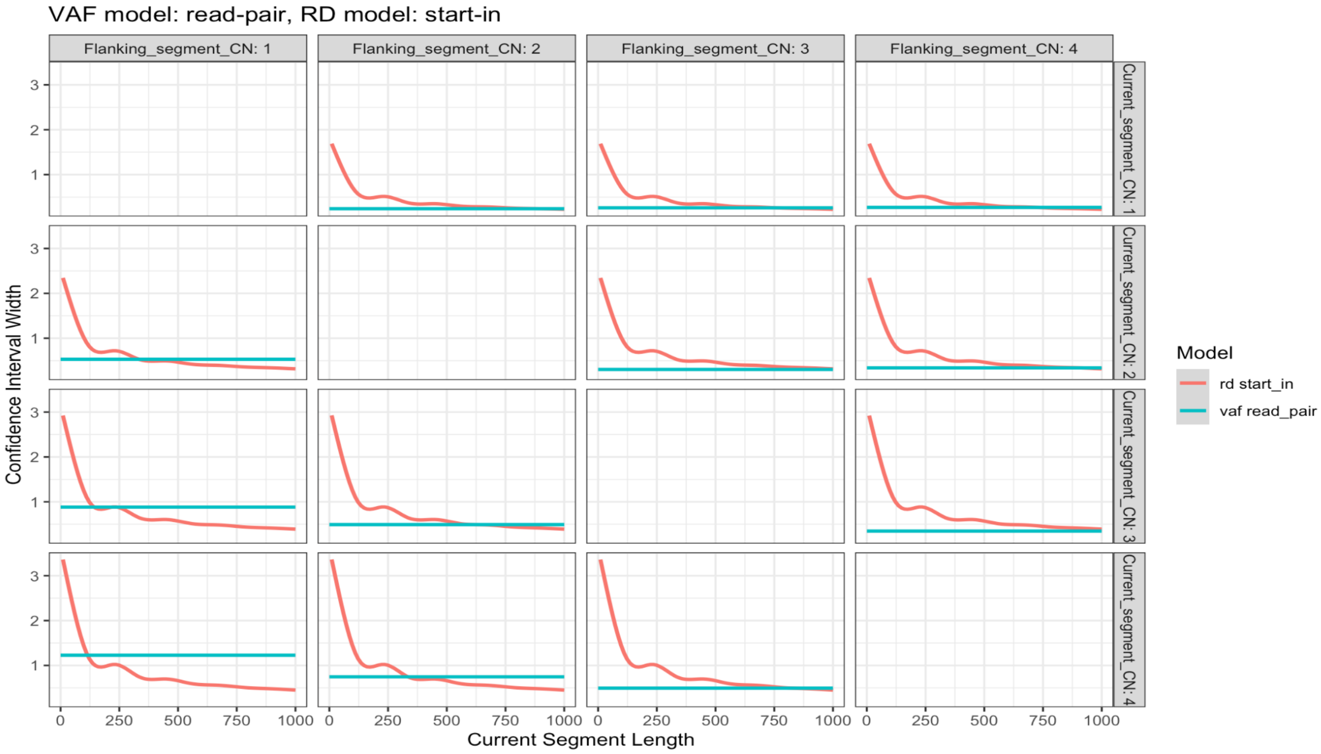


Figure 5: *Confidence interval trend for short current segment length values of 0 - 1000(b) in 12 different copy-number states in presented in a correlation style plot. VAF model's performance remains unaffected by change in current segment line. This is confirmed as the line representing CI of VAF model remains constant throughout each sub-plot. As expected, performance of read-depth model changes based current segment's length. A general tendency for worse read-depth model performance is captured for current segment length values shorter than 200(b).*

model performs better in deletion cases compared to amplification cases. This behaviour is not interpreted as VAF-based model performs better than read-depth model in deletion cases. Based on our simulations, we confirm that VAF-based model holds more copy-number predictive power for a deletion event, when compared to a case with the same parameters (apart from the direction of copy-number change) for an amplification event.

Figure (8) only includes observations from read-depth model inferred copy-numbers with CI width values lower than VAF model (i.e., simulated observations with better read-depth model performance). We demonstrate that majority of reads are skewed towards higher length values, which intuitively provide more statistical power for read-depth signal; hence, inferred copy-numbers have lower CI.

Read-depth based model relies on current segment's read-depth signal, and in amplification cases this signal is amplified, in turn positively contributing to copy-number inference with higher confidence. Figure (9) clearly confirms this assumption with a strong trend for having lower copy-number values in the flanking segment compared to current segment. Particularly, given that only datapoints with lower read-depth based model CI values were considered to generate figure (9).

We also looked at deletion cases with lower read-depth model CI values, to investigate the distribution of copy-numbers in flanking and current segments, given we have already observed

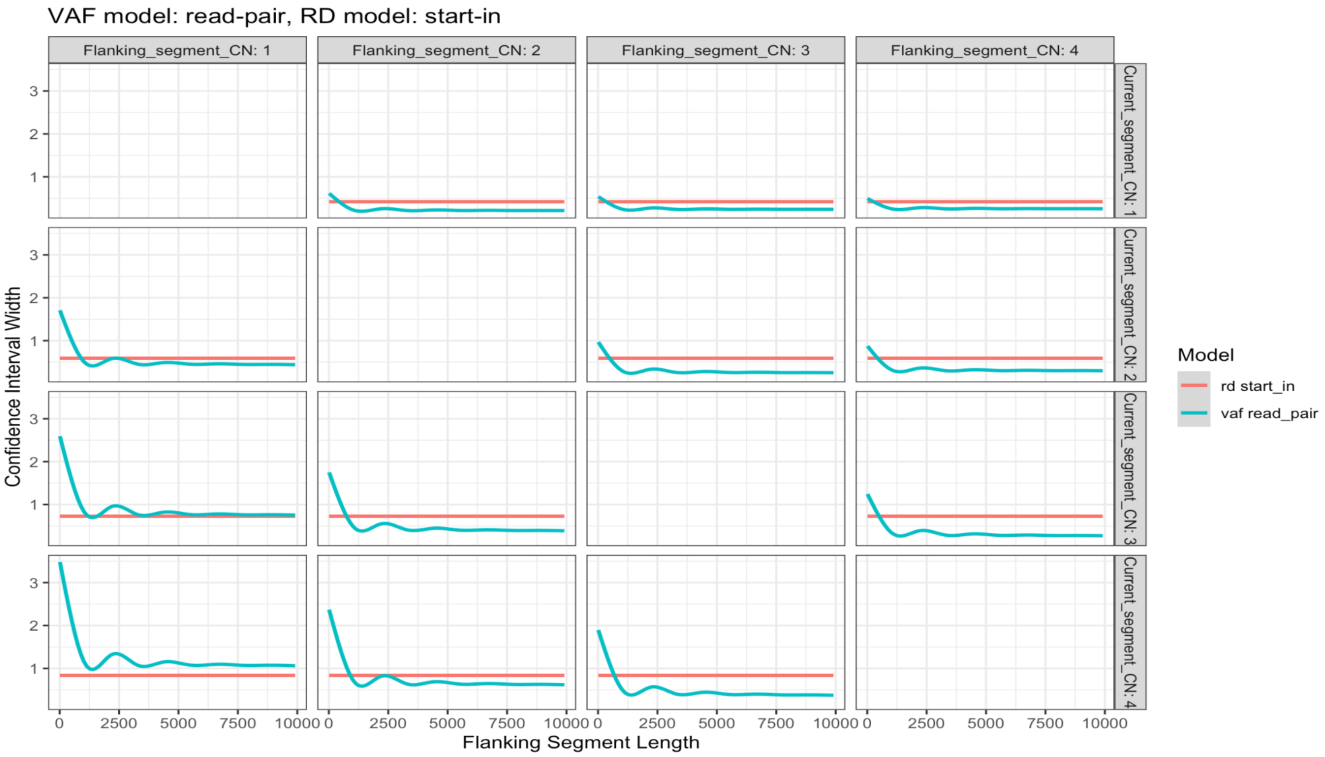


Figure 6: *General trend of confidence interval for various flanking segment length values of 0 - 10,000(b) in 12 different copy-number states are presented in a correlation style plot. Read-depth model does not rely on flanking segment length parameter. Therefore, performance of read-depth model remains unaffected by this parameter as indicated by flat red lines in all sub-plots. Given that copy-number of flanking segment in VAF-based model is inferred from read-depth signal, performance of VAF model is affected by this parameter and degrades as the flanking segment is reduced in length.*

that CI of inferred copy-numbers by VAF model is better in most deletion cases. Figure (10) shows a significant gap between flanking and current segment copy-numbers in such cases. We suspect the poor performance of VAF model is then resulted from having a large number of reference supportive reads and insufficient SV supportive reads (calculated by formula (6)), which reduces the statistical power of VAF-based copy-number inference.

Knowing that our read-depth model performs better in amplification cases, we also investigated amplification scenarios where VAF model resulted in a lower CI for the inferred copy-numbers; namely, read-depth model did not perform as expected. Hence, we looked at the distribution of flanking segment lengths (figure (11)). Even though, flanking segment length values were slightly skewed towards higher values in our dataset in figure (11), no significant difference could be reported. We then visualised the current segment length in these cases and observed that majority of length values are below 500(b) (figure (11)); it is highly likely that the short current segment length distribution in this dataset is responsible for worse performance of read-depth model.

Performance of our models fluctuate depending on values of four main parameters: flanking and current segment lengths, flanking and current segment copy-numbers. We investigated the impact of these parameters simultaneously by plotting flanking and current segment length (inclusive of all copy-number permutations), and generated a heat map based on logarithm of

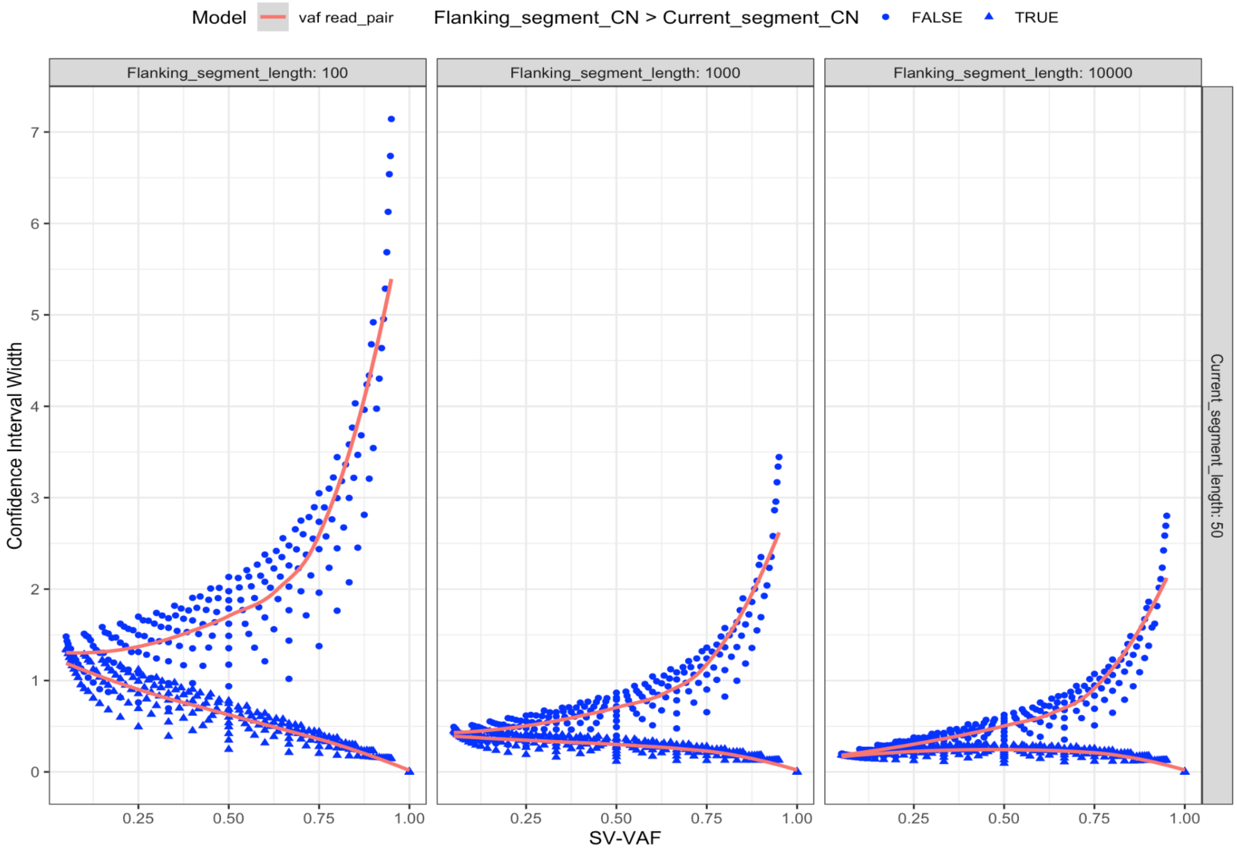


Figure 7: *Confidence interval of VAF model and corresponding VAF values for various copy-number states are represented for deletion (Flanking-segment-copy-number > Current-segment-copy-number)(blue triangles) and amplification (blue circles) events. Lower CI values were recorded for higher flanking segment lengths, and model's performance is better in deletion cases.*

absolute difference between inferred copy-number CI values of the two models, figure (12). We also only included the subset of observations with a significant difference (>0.05) between model CIs to narrow our focus for better visualisation. In figure (12) similarity of performance (smaller log of CI difference values) is represented by dark shades of red, and difference in performance (larger log of CI difference values) is highlighted by light shades of yellow. We observed a clear relationship and positive correlation between flanking and current segment length parameters. Also, a diagonal directed trend was detected at flanking and current segment length values that approximately overlap; along with a region that absolute performance difference of two models were consistently about 1 (i.e., $log(1) = 0$), highlighted by shade corresponding to log CI difference of 0. It is interesting to observe the gradual increase in model performance difference when going away from the diagonal shapes, suggesting in the regions with one segment length as extreme (either short length of flanking or current segment), one of the models perform significantly better than the other. Since performance of read-depth model is impacted by current segment's length, and performance of VAF-based model affected by the flanking segment's length, identified trends in figure (12) are consistent with the design of our models. Therefore, the regions of interest for further analysis would be lightly shaded areas (with log of CI difference of above 0) where performance of two models are significantly different. We also generated a second heat map plot that includes a larger dataset of flanking and current segment length values (demonstrated by figure (23) in section (8.1)). Figure (23) is considered the extended version

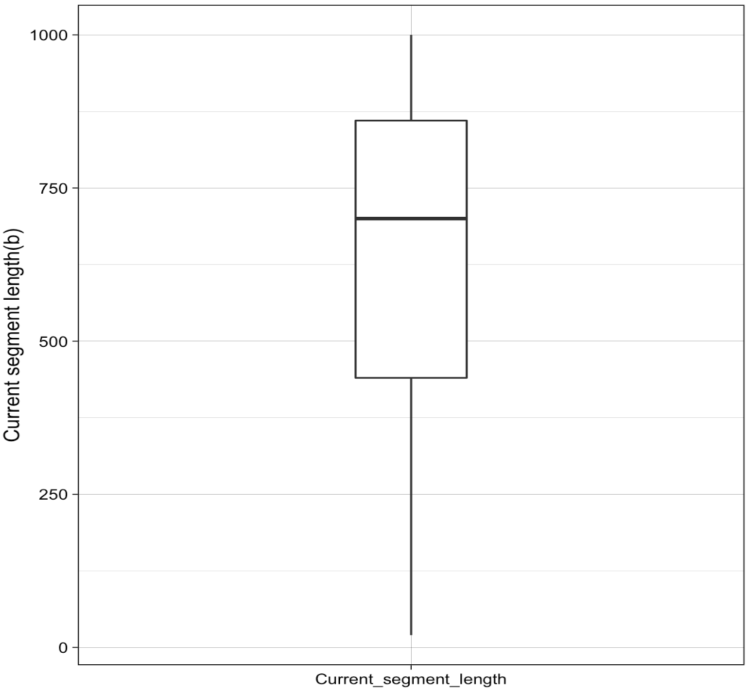


Figure 8: *Range of current segment lengths contained in a filtered dataset with lower read-depth based model inferred copy-number CI values. There is a shift to higher length values as more information is contained in longer segments.*

of figure (12), and indicates a more general trend in performance difference and the correlation between current and flanking segment length values across various copy-numbers.

We investigated regions with significant performance difference (i.e., with log of absolute CI difference values ≥ 0) shown in generated heat map (figure (12)) by looking at two subsets of our dataset. First subset consisted of events with significantly lower CIs inferred by VAF model; and second subset containing events with significantly better read-depth model performance. We observed that VAF model performs significantly better when flanking segment length range is distributed towards longer sizes, and current segment length is very short (majority of segments are less than 150(b)), figure (13). However, no considerable trend was detected from flanking and current segment copy-number distributions, figure (14). Therefore, we recommend using a VAF-based model for events with a long flanking segment (above 750(b) in our specific dataset) and a very short (below 150(b)) current segment, yields a more reliable copy-number inference.

Figure (15) visualises segment length distributions for events with significantly better read-depth model performance. As expected, given a read-depth based model relies on longer segment sizes for reliable inference, the distribution of current segment length is skewed towards higher length values. Additionally, flanking segment length distribution does not affect read-depth based copy-number inference directly; however, it represents flanking segment length range where VAF model did not perform as well as read-depth model due to sufficient size range of current segment length for read-depth based copy-number inference.

The distribution of copy-numbers in cases with superior read-depth model performance indicate a very strong bias towards large amplification cases, figure (16). Flanking segment copy-numbers are distinctly confined to only values under 10. While current segment copy-numbers predominantly consist of values above 20. Therefore, we suggest use of a read-depth based method

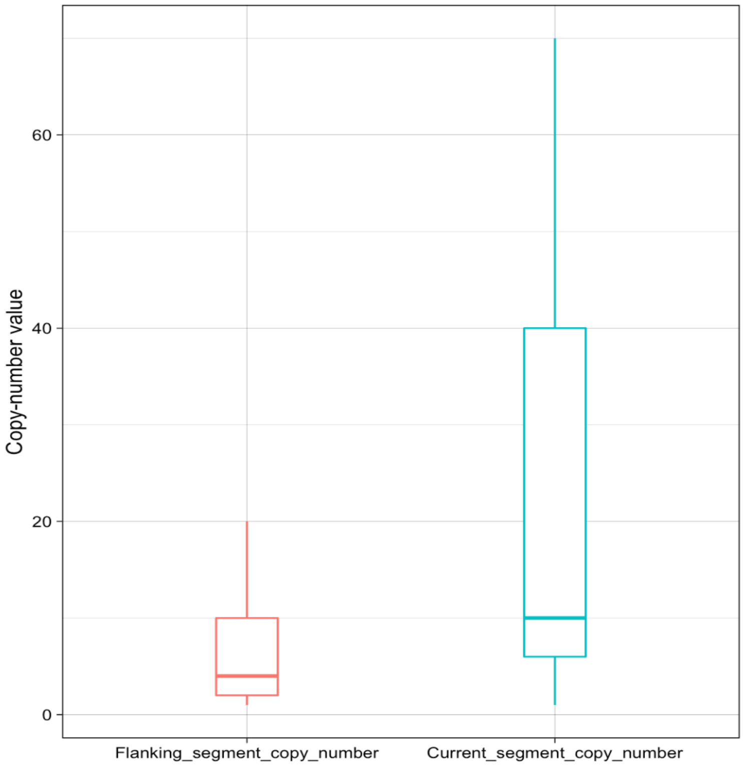


Figure 9: *Copy-number distribution of flanking and current segment for datapoints with lower read-depth based CI values (i.e., observations with better read-depth model performance) indicate a clear trend for lower values in flanking segment copy-number and higher values in current segment copy-number.*

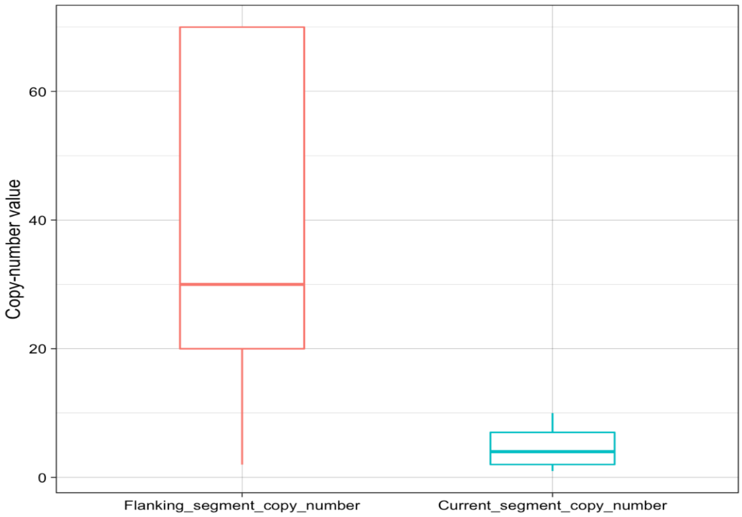


Figure 10: *Copy-number distribution for read-depth inferred deletion cases indicate a significant gap between flanking and current segment copy-numbers. Having large flanking segment copy-numbers coupled with small current segment copy-numbers negatively impacts VAF-based model's performance. Hence, read-depth model generated lower inferred copy-number CI values.*

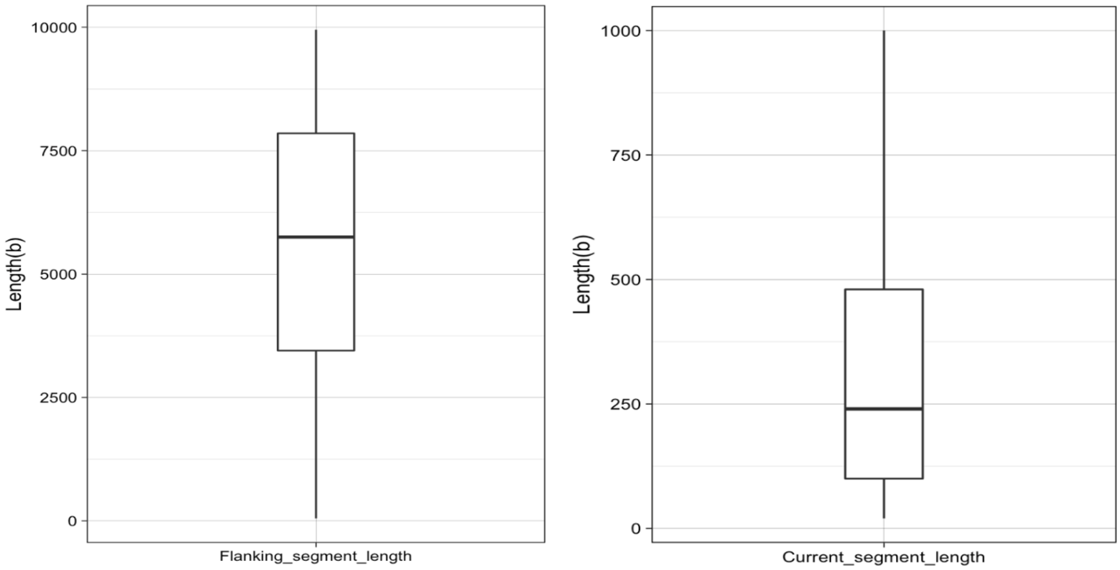


Figure 11: *Flanking and current segment length distribution for VAF inferred amplification cases were visualised to investigate potential challenging segment length ranges for read-depth model. Short current segment values, majority located below whole dataset current segment mean length, and slightly above average length flanking segment values were observed.*

for copy-number inference in current segment length of less than 1000(b) when a similar strong amplification copy-number trend along with sufficient current segment length values (typically above 250(b)) is observed.

5.3 Copy-number analysis on sample sequencing data

Knowing our simulated results, we aimed to test our hypotheses using adjusted real sequencing data and run multiple variant calling analysis pipelines to investigate the detection rate and accuracy of current callers for capturing variants less than 1000(b) in size. In addition, this provides a benchmark for length threshold of some frequently used variant callers. GIRDSS was used as our primary SV caller to detect precise breakpoint locations of pre-existing and implemented CNVs in our sample data. GRIDSS accurately detected 35 out of 42 events (missed 7 amplification events) from our BAMSurgeon generated variation clusters in truth sets 1, 2, and 3 (i.e., deletion, amplification, and mixed clusters). GRIDSS does not utilise read-depth signal, it instead gathers evidence from a combination of breakpoint detection approaches including discordant read-pair, split-read and break-end assembly. Based on that, consistent with our simulation results that a read-depth based approach is most effective for capturing amplification events, variation events that were not detected by GRIDSS were all amplification events (table (2) in section (8.2)). Furthermore, GRIDSS was able to infer exact breakpoint of variants in truth sets 1, 2, and 3 with great accuracy and resolution. In most cases reported variant breakpoint positions were either exactly as the true variant's position or with only a few bases difference. Nonetheless, GRIDSS only reports precise breakpoints and not copy-number of variant segments.

PURPLE is a read-depth based copy-number caller that takes precise SV breakpoint positions from an upstream SV caller, typically GRIDSS, for allele-specific copy-number inference. Even though, PURPLE's documentation strongly recommends using PURPLE with SV calls input from a single-base resolution SV caller such as GRIDSS, we first used PURPLE in isolation with

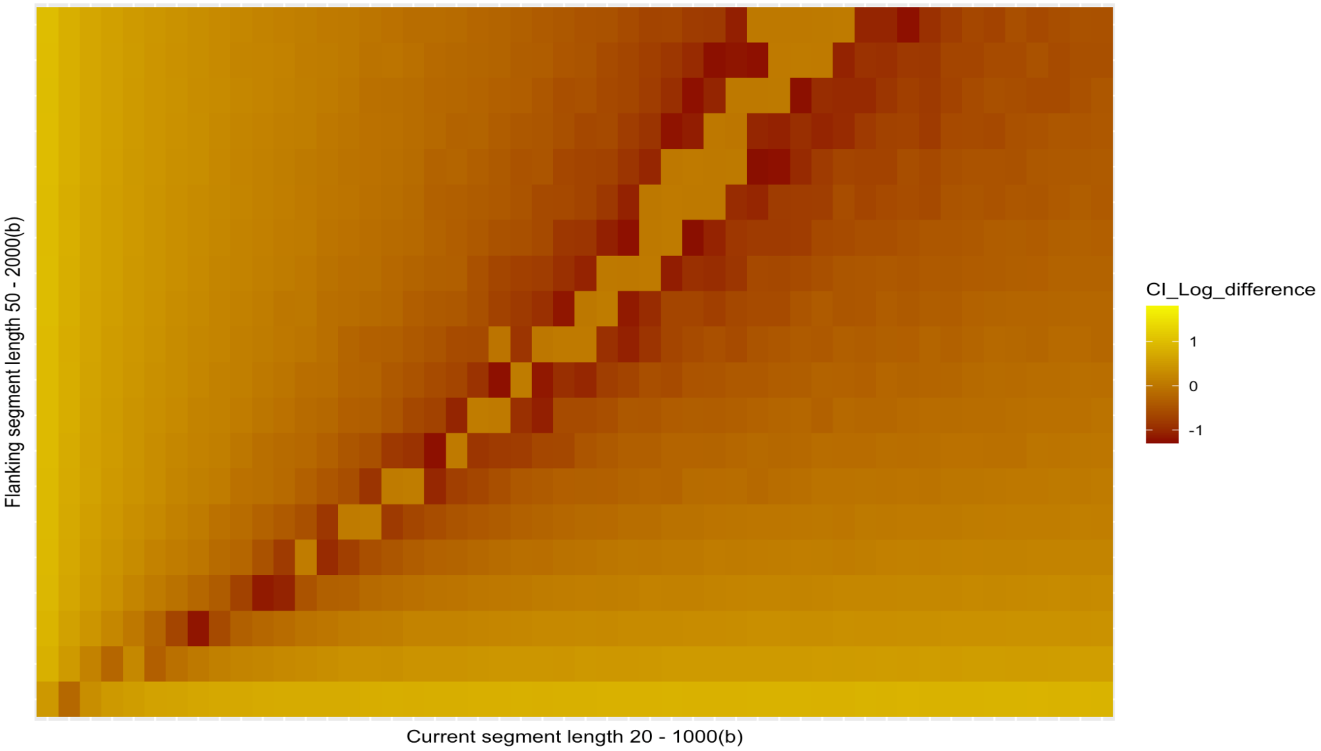


Figure 12: *Heat map based on log of CI difference between models and their corresponding flanking and current segment length values for a wide range of inferred copy-numbers is visualised by above heat map. Lower CI-Log-difference values, represented by dark shades of orange, show simulation observations with similar performance between VAF and read-depth models. Areas represented by lighter shades of orange, higher CI-log-difference values, show observations with significantly different recorded performance between VAF and read-depth inferred copy-numbers.*

default parameters to observe if it is capable of detecting any of our implemented variants. As a result, PURPLE called a very limited number of somatic CNVs and was unable to capture any variants from our clusters. However, PURPLE performed exceptionally well in capturing most variants from our clusters when high quality SV calls and precise breakpoint locations were passed on from GRIDSS as additional evidence for copy-number inference. PURPLE accurately inferred copy-numbers of 8 out of 15 CNVs in deletion cluster (truth set 1), 8 out of 17 CNVs in amplification cluster (truth set 2), and 7 out of 10 in mix cluster (truth set 3). 13 out of 23 total calls made by PURPLE were amplifications and the remaining 10 calls were deletions. Some partial calls were also made which are likely due to reassembly of reads by BAMSurgeon in our clusters that caused ambiguity by having duplicate read names and potentially confusing PURPLE's segmentation algorithm. We noticed a short-segment duplication call with length of 54(b) and inferred copy-number of 10.5 which was accurately detected and inferred by both GRIDSS and PURPLE. Furthermore, a short-segment deletion event with length of 69(b) was correctly detected by GRIDSS with accurate single-base resolution; however, PURPLE did not infer the copy-number of this deletion event. Although, our sequencing variant calling results only show a limited set of observations, our suggestion about suitability of read-depth signal in short-segment amplification copy-number inference have been supported by PURPLE's output. Furthermore, GRIDSS seems to detect precise breakpoint of deletion events more frequently compared to amplification events.

1658 CNVs were added to HG002 normal sample sequencing data to generate truth set 4 in

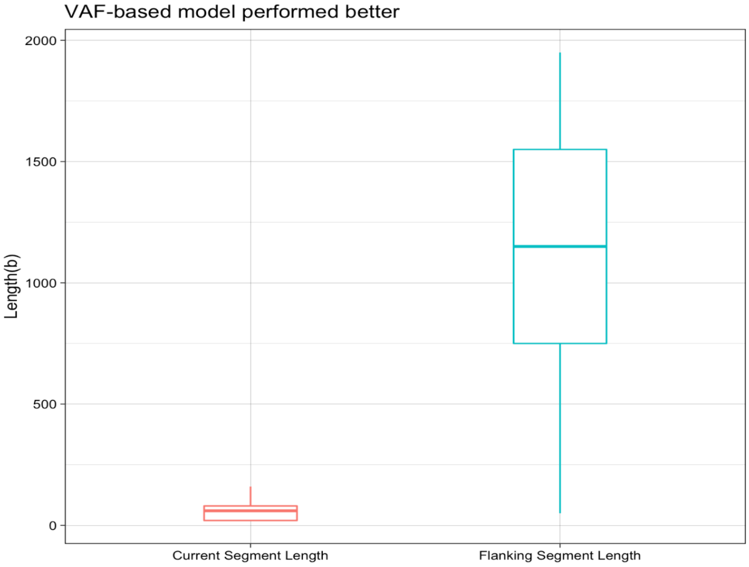


Figure 13: *Distribution of flanking and current segment length values for subset of observations with VAF-based model inferred log of absolute CI difference values* ≥ 0. *VAF-based model performed significantly better for short current segment length of less than 150(b) and longer flanking segment length range. Additionally, the distribution of current segment copy number values appeared to be highly right skewed. This is likely due to presence of many outliers towards the right end tail of this distribution.*

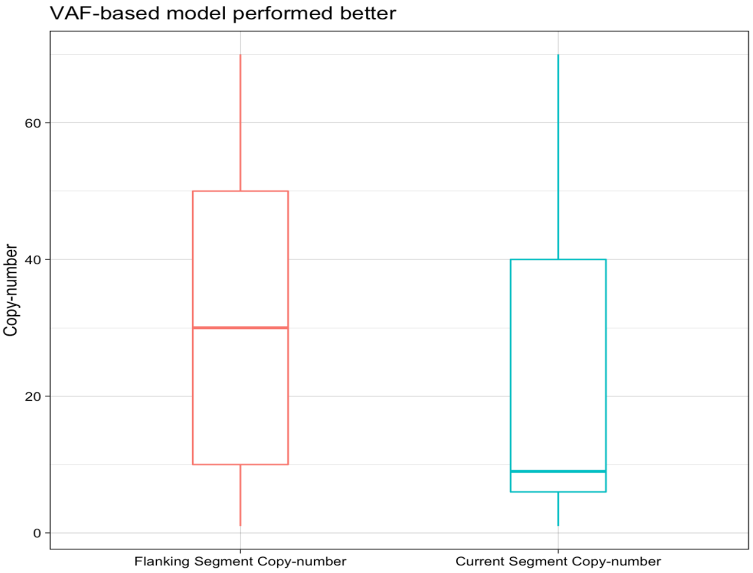


Figure 14: *Distribution of flanking and current segment copy-numbers for subset of observations with VAF-based model inferred log of absolute CI difference values* ≥ 0. *No general trend was detected for copy-number distributions of flanking and current segments when VAF-based model performed significantly better.*

order to investigate the ability and detection rate of GRIDSS and PURPLE in small-segment CNVs. The number of implemented deletion and amplification events is described in methods section 4.7 in detail. GRIDSS and PURPLE detected 0.78 (1301 out of 1658) and 0.68 (1140 out

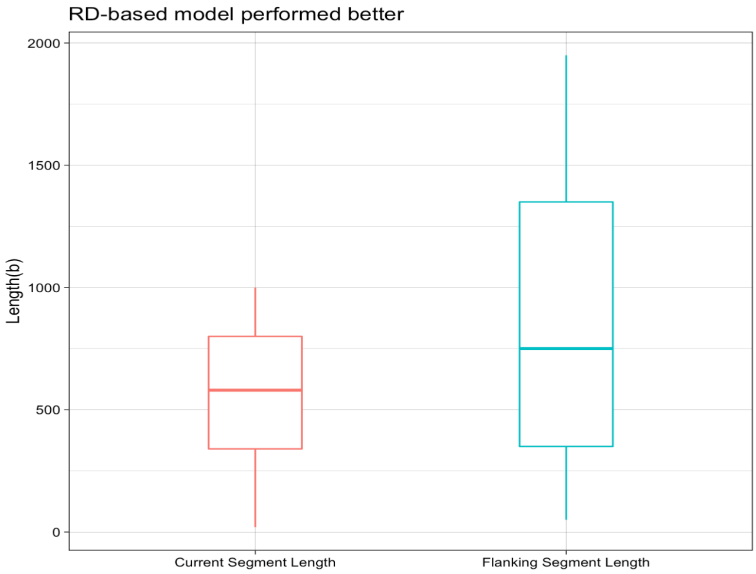


Figure 15: *Distribution of flanking and current segment length values for subset of observations with read-depth model inferred log of absolute CI difference values* ≥ 0*. Current segment values are skewed towards larger length values (above 250(b)) with median falling above 500(b). Meanwhile, flanking segment lengths cover a wide range of values but do not indicate any deviation or skewness toward any extremes. Since read-depth model performed significantly better for these length distributions, we can confirm that read-depth model's performance has a strong positive correlation with current segment length and it particularly performs better for current segment length values above 250(b).*

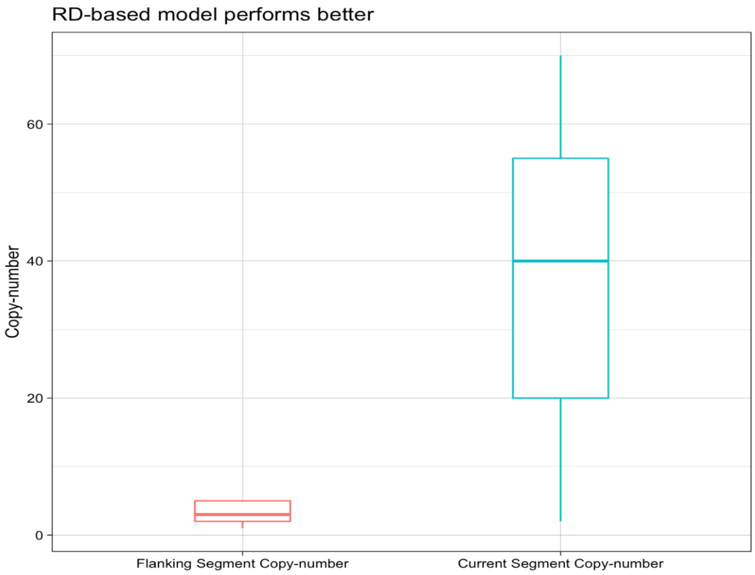


Figure 16: *Distribution of flanking and current segment copy numbers for subset of observations with read-depth model inferred log of absolute CI difference values* ≥ 0*. A distinct strong amplification trend is observed with flanking copy-number ranging in values lower than 10, and current segment copy-numbers of mostly larger than 20. Furthermore, the difference between flanking and current segment copy-numbers are large and on average more than 10.*

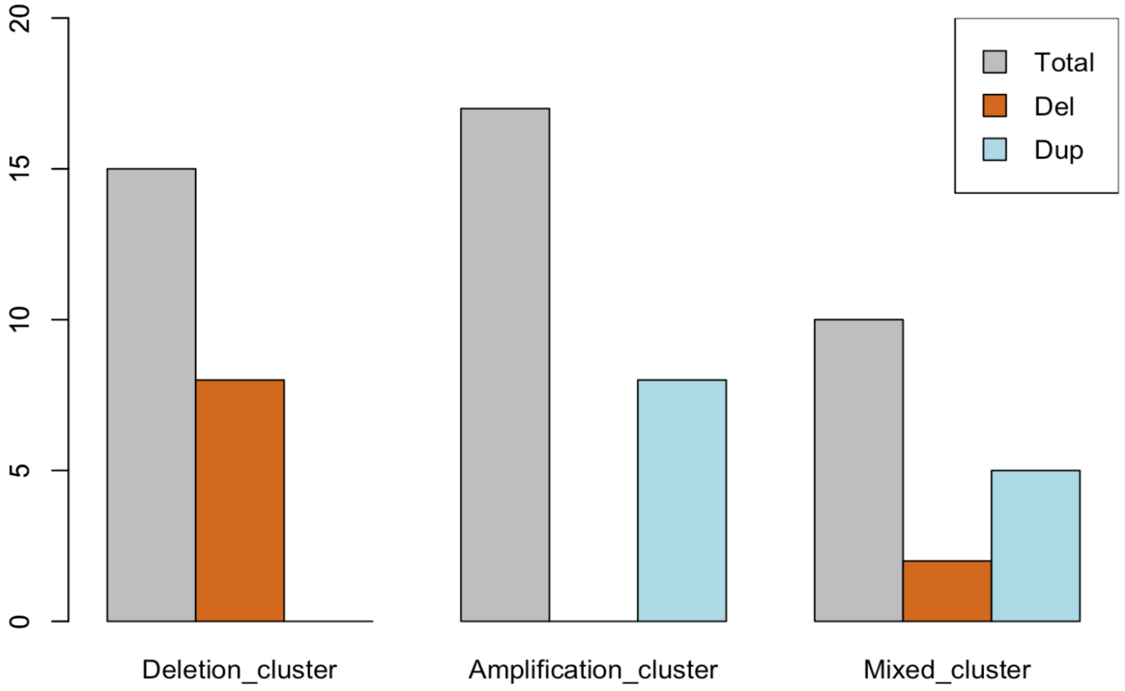


Figure 17: *Total number of variants for each cluster is reported by the first bar of each cluster. The number of 'Del' and 'Dup' calls made by PURPLE is reported by second and third bars respectively. PURPLE accurately detected the same number of variants (8 calls for each cluster) in deletion and amplification clusters. More amplifications were called (5 calls) compared to deletions (2 calls) in the mix cluster. A short-segment duplication event with detected length of 54(b) and inferred copy-number value of 10.5 was also successfully called by PURPLE. This segment exactly corresponds to an amplification event from the generated variant set we added with BAMSurgeon.*

of 1658) of total CNVs respectively. 1354 CNVs contained in truth set 4 were less than 1000(b) in length. Consequently, detection rate of GRIDSS and PURPLE for Under 1000(b) in length CNVs were 0.76 (1030 out of 1354) and 0.67 (914 out of 1354) respectively figure (18). GRIDSS detection rate was higher than PURPLE in both total calls and calls for less than 1000(b) in length; this is consistent with our assumption that alternative signals to read-depth are more effective in capturing shorter than 1000(b) CNVs. GRIDSS uses a combination approach and attempts to assemble breakpoints for a more accurate variation event reconstruction. We suspect assembly methods such as the one GRIDSS utilises (break-end assembly), are highly effective for capturing small variation events with higher resolution. It is interesting to observe that GRIDSS detected more duplication events than deletion events under 1000(b) in length. We were expecting GRIDSS would detect more deletion events as it does not use read-depth signal for variation inference. Perhaps this is due to the structure and characteristics of implemented variants in truth set 4 dataset.

We performed variant calling only on truth sets 1, 2, and 3 with CNVpytor due to its poor performance on our variant datasets. CNVpytor uses read-depth as its primary variant detection signal and it is capable of copy-number inference in short bins down to 100(b). CNVpytor missed precise breakpoints of our deletion cluster, and instead of calling individual deletion events, it misclassified most of these regions as a single duplication event with the reported length of 4600(b) in truth set 1. Only 2 amplification events were detected by CNVpytor in

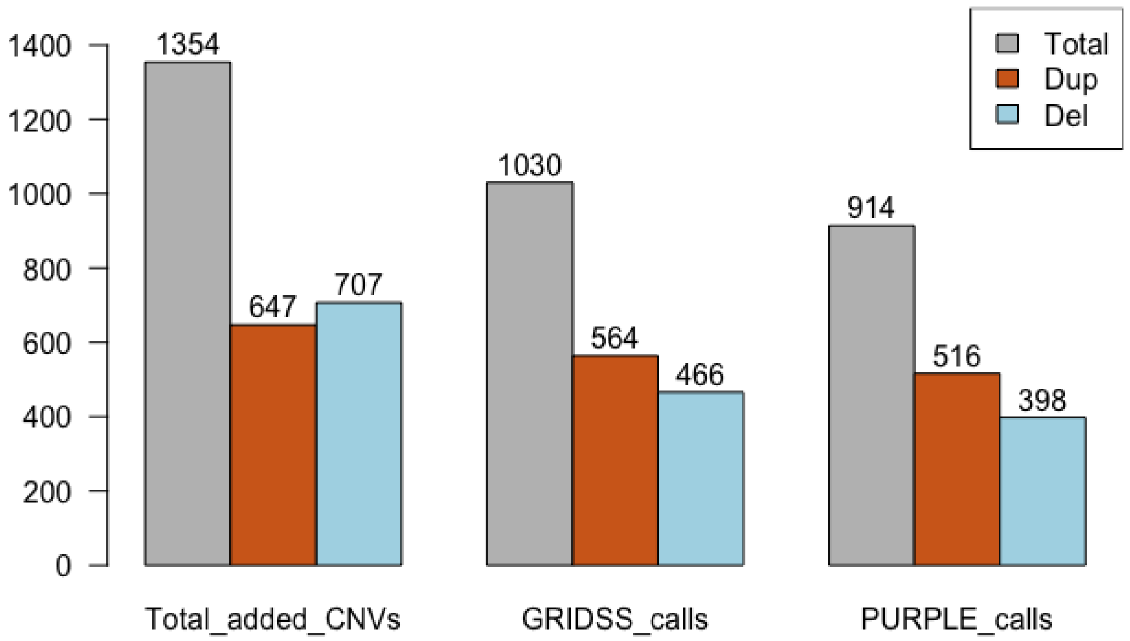


Figure 18: *1354 CNVs under 1000(b) were added to our normal sample (chromosome 21 of HG002 data), containing 647 and 707 duplication 'Dup' and deletion 'Del' events respectively. GRIDSS detected 116 more CNVs compared to PURPLE. GRIDSS and PURPLE both detected more duplications than deletions even though, total number of added deletions were higher than duplications.*

the amplification cluster (truth set 2), which were misclassified as one event; inferred as an amplified segment with a length of 6500(b), 200(b) longer than actual true length of the two events combined. Similarly, our mixed cluster (truth set 3) which contained a combination of duplications and deletions with varying sizes was incorrectly called as one single duplication event 1300(b) in length.

Optional VAF values at SNP locations were also provided to CNVpytor as input in order to perform a combined copy-number call on truth sets 1, 2, and 3. We aimed to improve detection rate and segmentation resolution of CNVpytor by adding SNP VAF values, given its poor performance using read-depth only model. Nonetheless, combined call feature is still in prototype phase as specified in CNVpytor's documentation. Combined calling did not improve the accuracy or detection rate of copy-number calls on any of our clusters despite much longer run-times and extra input SNP locations. Interestingly, CNVpytor's combined calls reported larger CNV segment length values for the same calls made by read-depth only model. For instance, in amplification and mix clusters (truth set 1 and 2), both combined calls reported on average 1500(b) larger segment sizes. Although, this issue may be addressed in the final release added VAF values option. Also we suspect CNVpytor's difficulty to detect precise breakpoints and differentiate events is caused by having complex compound rearrangements in our datasets. Overall performance of CNVpytor on our three variation clusters (truth sets 1, 2, and 3) was poor and quite different compared to calls made by GRIDSS and PURPLE.

6 Conclusion

This study presents foundational methodology and generated simulation data for comparing methods of copy-number inference from VAF values and read-depth signal. We believe this is the first study to propose alternative methods and investigate performance of models based on VAF at SV breakpoints for inferring copy-number in challenging short-segments and compounding variant clusters. Therefore, this study serves as a foundational study for further methods and tools development in this topic. Even though, suggestions for suitability of read-depth or VAF based approach in copy-number inference can be made for specific variation event types and sizes; our results demonstrated it is still challenging to rely on a single inference signal for accurate and reliable copy-number evaluation in short CNVs and compounding variant clusters. Nonetheless, the growing trend for ensemble-based and combination approach variant calling in recent years provides better insight into more effective variation inference methods in challenging genomic rearrangements.

We recognised the following described shortcomings in this project for model design and simulations which could be addressed by future studies. Our assumption of a Poisson distribution for segment read counts may not reflect true biological sequencing read count distributions. This is because mean and variance in a Poisson distribution are equal values; however, real sequencing data tend to have higher variation that result in, a larger variation value than mean. It is demonstrated that variation in GC content of local genomic positions affect the sequencing coverage and therefore, normalisation of data against GC content (in most cases per bin or segment) is required during sequence read-depth signal analysis. We assumed GC bias-free data in our simulations and did not adjust coverage based on GC bias. However, adjusting for GC and other normalisation steps before reconstructing read-depth signal may influence the statistical inference power of read-depth based models, which is not taken into account in this study. Furthermore, often library sizes between tumour and normal samples vary, which needs to be considered when working with averaged values such as average read-depth or read-depth ratios. We assumed our tumour and normal samples are completely pure and therefore, did not adjust for purity. Adjusting with purity may reduce available inference signals in tumour samples and affect copy-number inference models. Lastly, our models assumed input data parameters are absolute true values, which is not the case with input parameters from real sequencing data that often contain some level of error.

Due to missing copy-number information of short arm on chromosomes 13, 14, 15, 21, and 22 in human genome, the copy-number of the long arm is extended to the short arm. We extracted only chromosome 21 from our sample sequencing data as it is a smaller dataset to work with and to make our data processing less time consuming; however, not having any information about short arm copy-number of chromosome 21 may influence the performance of our selected variant callers negatively. Consequently, in future studies perhaps a more comprehensive whole genome data which includes total chromosome count would reflect a more accurate representation of performance of variant callers and make variant analysis more thorough.

Based on our simulations and real sequencing data variant calling analysis we make the following suggestions. We suggest read-depth based copy-number inference models generally perform well in amplification events due to increase in segment read-counts. Additionally, read-depth signal is likely to produce more accurate inferred copy-numbers in amplification events with segment length size of higher than 250(b); this is particularly the case with highly amplified segments where copy-number change between flanking and current segment is more than 10 (i.e., $|CN_{flanking\ segment} - CN_{current\ segment}| \geq 10$). VAF based copy-number inference models are unaffected by current segment length but sensitive to drastic copy-number change and short flanking segment length values. Therefore, VAF based models generally perform better in deletion events where read-depth signal is often weak due to low number of read counts. Hence, VAF is often more suitable for copy-number inference in segments less than 150(b) in size. Compounding variant clusters are particularly challenging for both VAF and read-depth models. This is because VAF value of each segment in a cluster depends on its corresponding flanking segment VAF value; namely, one VAF value is inferred based on the VAF value of its flanking segment which is also a variant segment in a cluster of compound variations. Often inaccurate inferred VAF value of previous segments would result in poor performance of VAF based model in the entire cluster. Moreover, read-depth signal is not capable of resolving segments in such clusters with high resolution and therefore, segment specific copy-number inference is not possible. Additionally, segment length sizes that are less than read-length size cause complications in calculating read-depth; this lead to loss of read-depth signal over such compounding variant clusters. Consequently, copy-number inference in segments located in variant clusters require further investigation to develop more accurate and effective inference methods. We also demonstrated that variant callers capable of single-base breakpoint resolution such as GRIDSS can in fact detect short-segment CNVs; however, GRIDSS detection rate for variant calling in SVs under 1000(b) in length requires improvement. Copy-number inference in short-segments by PURPLE proven to be a more challenging task compared to just detecting precise breakpoint positions. Furthermore, PURPLE's output was not consistent enough to be reliably used for segments under 1000(b) in length. Given the complexity of inferring copy-number in under 1000(b) length CNV segments, perhaps development of a specialised short-segment CNV caller is the most feasible approach to tackle this task. Particularly, considering the inability of current variant callers for copy-number inference in short-segments.

7 References

1. Carson, A. R., Feuk, L., Mohammed, M. & Scherer, S. W. Strategies for the detection of copy number and other structural variants in the human genome. *Human Genomics* **2,** 403–414. ISSN: 1473-9542. doi:10.1186/1479-7364-2-6-403 (June 2006).

2. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. en. *Nature Reviews Genetics* **12,** 363–376. ISSN: 1471-0064. doi:10.1038/nrg2958 (May 2011).

3. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. en. *Nature Reviews Genetics* **17,** 224–238. ISSN: 1471-0064. doi:10.1038/nrg.2015.25 (Apr. 2016).

4. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biology* **20,** 246. ISSN: 1474-760X. doi:10.1186/s13059-019-1828-7 (Nov. 2019).

5. Sebat, J. *et al.* Strong Association of De Novo Copy Number Mutations with Autism. EN. *Science.* doi:10.1126/science.1138659 (Apr. 2007).

6. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. en. *Nature Genetics* **41,** 1223–1227. ISSN: 1546-1718. doi:10.1038/ng.474 (Nov. 2009).

7. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. en. *Nature* **471,** 499–503. ISSN: 1476-4687. doi:10.1038/nature09884 (Mar. 2011).

8. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14,** S1. ISSN: 1471-2105. doi:10.1186/1471-2105-14-S11-S1 (Sept. 2013).

9. Buysse, K. *et al.* Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. en. *European Journal of Medical Genetics* **52,** 398–403. ISSN: 1769-7212. doi:10.1016/j.ejmg.2009.09.002 (Nov. 2009).

10. Carter, N. P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature genetics* **39,** S16–S21. ISSN: 1061-4036. doi:10.1038/ng2028 (July 2007).

11. Kallioniemi, O. P. *et al.* Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. eng. *Seminars in Cancer Biology* **4,** 41–46. ISSN: 1044-579X (Feb. 1993).

12. Snijders, A. M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. en. *Nature Genetics* **29,** 263–264. ISSN: 1546-1718. doi:10.1038/ng754 (Nov. 2001).

13. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. en. *Bioinformatics*

28, 2711–2718. ISSN: 1460-2059, 1367-4803. doi:10.1093/bioinformatics/bts535 (Nov. 2012).

14. Hayes, M. en. in *Cancer Bioinformatics* (ed Krasnitz, A.) 65–83 (Springer, New York, NY, 2019). ISBN: 9781493988686. doi:10.1007/978-1-4939-8868-6_3.

15. Liu, B. *et al.* Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. en. *Oncotarget* **6,** 5477–5489. ISSN: 1949-2553. doi:10.18632/oncotarget.3491 (Mar. 2015).

16. Goh, G., McGranahan, N. & Wilson, G. A. en. in *Cancer Bioinformatics* (ed Krasnitz, A.) 217–226 (Springer, New York, NY, 2019). ISBN: 9781493988686. doi:10.1007/978-1-4939-8868-6_13.

17. Fortier, N., Rudy, G. & Scherer, A. in *Copy Number Variants* (ed Bickhart, D. M.) 115–127 (Springer New York, New York, NY, 2018). ISBN: 9781493986651 9781493986668. doi:10.1007/978-1-4939-8666-8_9.

18. Cmero, M. *et al.* Inferring structural variant cancer cell fraction. en. *Nature Communications* **11,** 730. ISSN: 2041-1723. doi:10.1038/s41467-020-14351-8 (Feb. 2020).

19. Kim, J. *et al.* Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. en. *New England Journal of Medicine.* doi:10.1056/NEJMoa1813279 (Oct. 2019).

20. Coutelier, M. *et al.* Combining callers improves the detection of copy number variants from whole-genome sequencing. en. *European Journal of Human Genetics* **30,** 178–186. ISSN: 1476-5438. doi:10.1038/s41431-021-00983-x (Feb. 2022).

21. Gabrielaite, M. *et al.* A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. en. *Cancers* **13,** 6283. ISSN: 2072-6694. doi:10.3390/cancers13246283 (Dec. 2021).

22. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. en. *Genome Biology* **20,** 117. ISSN: 1474-760X. doi:10.1186/s13059-019-1720-5 (Dec. 2019).

23. Zhang, L., Bai, W., Yuan, N. & Du, Z. Comprehensively benchmarking applications for detecting copy number variation. en. *PLOS Computational Biology* **15** (ed Ioshikhes, I.) e1007069. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1007069 (May 2019).

24. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. en. *Biostatistics* **5,** 557–572. ISSN: 1465-4644, 1468-4357. doi:10.1093/biostatistics/kxh008 (Oct. 2004).

25. Cun, Y., Yang, T.-P., Achter, V., Lang, U. & Peifer, M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. en. *Nature Protocols* **13,** 1488–1501. ISSN: 1750-2799. doi:10.1038/nprot.2018.033 (June 2018).

26. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144,** 27–40. ISSN: 0092-8674. doi:10.1016/j.cell.2010.11.055 (Jan. 2011).

27. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research* **27,** 2050–2060. ISSN: 1088-9051. doi:10.1101/gr.222109.117 (Dec. 2017).

28. Cameron, D. L. *et al.* *GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number* en. Tech. rep. Type: article (bioRxiv, Sept. 2019), 781013.

29. Stankiewicz, P. & Lupski, J. R. Structural Variation in the Human Genome and its Role in Disease. en. *Annual Review of Medicine* **61,** 437–455. ISSN: 0066-4219, 1545-326X. doi:10.1146/annurev-med-100708-204735 (Feb. 2010).

30. De Pagter, M. S. & Kloosterman, W. P. in *Chromosomal Instability in Cancer Cells* (eds Ghadimi, B. M. & Ried, T.) 165–193 (Springer International Publishing, Cham, 2015). ISBN: 9783319202907 9783319202914. doi:10.1007/978-3-319-20291-4_8.

31. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. en. *Nature Genetics* **45,** 1134–1140. ISSN: 1546-1718. doi:10.1038/ng.2760 (Oct. 2013).

32. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. en. *Proceedings of the National Academy of Sciences* **104,** 20007–20012. ISSN: 0027-8424, 1091-6490. doi:10.1073/pnas.0710052104 (Dec. 2007).

33. Alqahtani, K., Taylor, C. C., Wood, H. M. & Gusnanto, A. Sparse modelling of cancer patients' survival based on genomic copy number alterations. en. *Journal of Biomedical Informatics* **128,** 104025. ISSN: 1532-0464. doi:10.1016/j.jbi.2022.104025 (Apr. 2022).

34. Zhou, Y., Bickhart, D. M. & Liu, G. E. en. in *Copy Number Variants: Methods and Protocols* (ed Bickhart, D. M.) 49–59 (Springer, New York, NY, 2018). ISBN: 9781493986668. doi:10.1007/978-1-4939-8666-8_3.

35. Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. en. *Nature Genetics* **47,** 296–303. ISSN: 1546-1718. doi:10.1038/ng.3200 (Mar. 2015).

36. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. en. *Proceedings of the National Academy of Sciences* **107,** 16910–16915. ISSN: 0027-8424, 1091-6490. doi:10.1073/pnas.1009843107 (Sept. 2010).

37. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. en. *Nature Biotechnology* **30,** 413–421. ISSN: 1546-1696. doi:10.1038/nbt.2203 (May 2012).

38. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. eng. *Annals of Oncology: Official Journal of the European Society for Medical Oncology* **26,** 64–70. ISSN: 1569-8041. doi:10.1093/annonc/mdu479 (Jan. 2015).

39. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. en. *Nature Methods* **6,** 99–103. ISSN: 1548-7105. doi:10.1038/nmeth.1276 (Jan. 2009).

40. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. en. *PLOS Computational Biology* **12,** e1004873. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1004873 (Apr. 2016).

41. Ben-Yaacov, E. & Eldar, Y. C. A fast and flexible method for the segmentation of aCGH data. en. *Bioinformatics* **24,** i139–i145. ISSN: 1367-4803, 1460-2059. doi:10.1093/bioinformatics/btn272 (Aug. 2008).

42. Tibshirani, R. & Wang, P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. en. *Biostatistics* **9,** 18–29. ISSN: 1465-4644, 1468-4357. doi:10.1093/biostatistics/kxm013 (Jan. 2008).

43. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13,** 591. ISSN: 1471-2164. doi:10.1186/1471-2164-13-591 (Nov. 2012).

44. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. en. *Genome Research* **19,** 1639–1645. ISSN: 1088-9051. doi:10.1101/gr.092759.109 (Sept. 2009).

45. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. en. *Genome Research* **24,** 1881–1893. ISSN: 1088-9051, 1549-5469. doi:10.1101/gr.180281.114 (Nov. 2014).

46. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. en. *Genome Biology* **14,** R80. ISSN: 1465-6906. doi:10.1186/gb-2013-14-7-r80 (2013).

47. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. English. *Cell* **149,** 994–1007. ISSN: 0092-8674, 1097-4172. doi:10.1016/j.cell.2012.04.023 (May 2012).

48. Fan, X., Luo, G. & Huang, Y. S. Accucopy: accurate and fast inference of allele-specific copy number alterations from low-coverage low-purity tumor sequencing data. *BMC Bioinformatics* **22,** 23. ISSN: 1471-2105. doi:10.1186/s12859-020-03924-5 (Jan. 2021).

49. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. en. *Nature Methods* **15,** 591–594. ISSN: 1548-7105. doi:10.1038/s41592-018-0051-x (Aug. 2018).

50. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. en. *Genome Research* **21,** 974–984. ISSN: 1088-9051, 1549-5469. doi:10.1101/gr.114876.110 (June 2011).

51. Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. en. *GigaScience* **10,** giab074. ISSN: 2047-217X. doi:10.1093/gigascience/giab074 (Nov. 2021).

52. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339. ISSN: 1367-4803. doi:10.1093/bioinformatics/bts378 (Sept. 2012).

53. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. eng. *Bioinformatics (Oxford, England)* **32,** 1220–1222. ISSN: 1367-4811. doi:10.1093/bioinformatics/btv710 (Apr. 2016).

54. Anzar, I., Sverchkova, A., Stratford, R. & Clancy, T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. en. *BMC Medical Genomics* **12,** 63. ISSN: 1755-8794. doi:10.1186/s12920-019-0508-5 (Dec. 2019).

55. Wang, M. *et al.* SomaticCombiner: improving the performance of somatic variant calling based on evaluation tests and a consensus approach. en. *Scientific Reports* **10,** 12898. ISSN: 2045-2322. doi:10.1038/s41598-020-69772-8 (Dec. 2020).

56. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. en. *Scientific Data* **3,** 160025. ISSN: 2052-4463. doi:10.1038/sdata.2016.25 (Dec. 2016).

57. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. en. *Nature Methods* **12,** 623–630. ISSN: 1548-7105. doi:10.1038/nmeth.3407 (July 2015).

8 Appendices

8.1 Appendix 1: Accompanying Simulation Figures

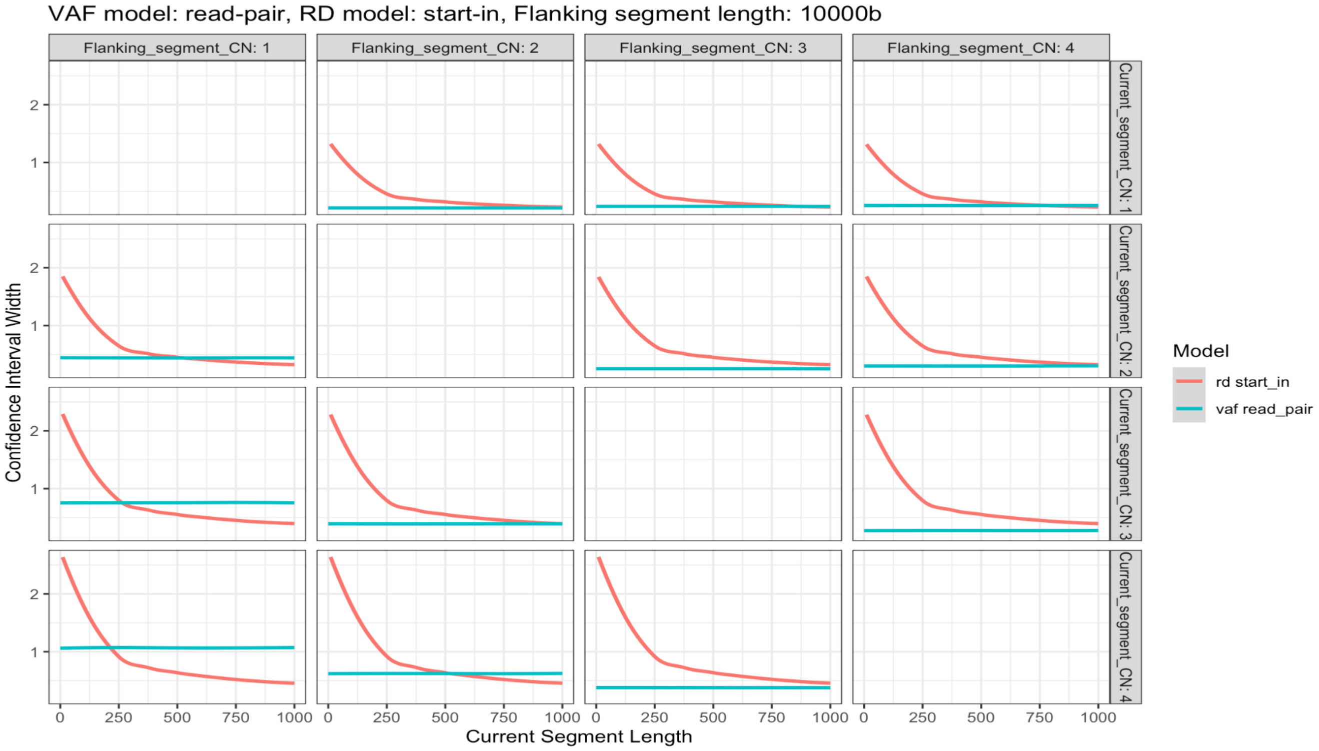


Figure 19: *CI trend for short current segment length(b) values with a single flanking value of 10,000(b) in 12 different copy-number states. The stable trend of CI line confirms that when a single flanking segment length is chosen, localised fluctuations of CI line is resolved and therefore, our interpretation of local fluctuations in figure (5) is correct.*

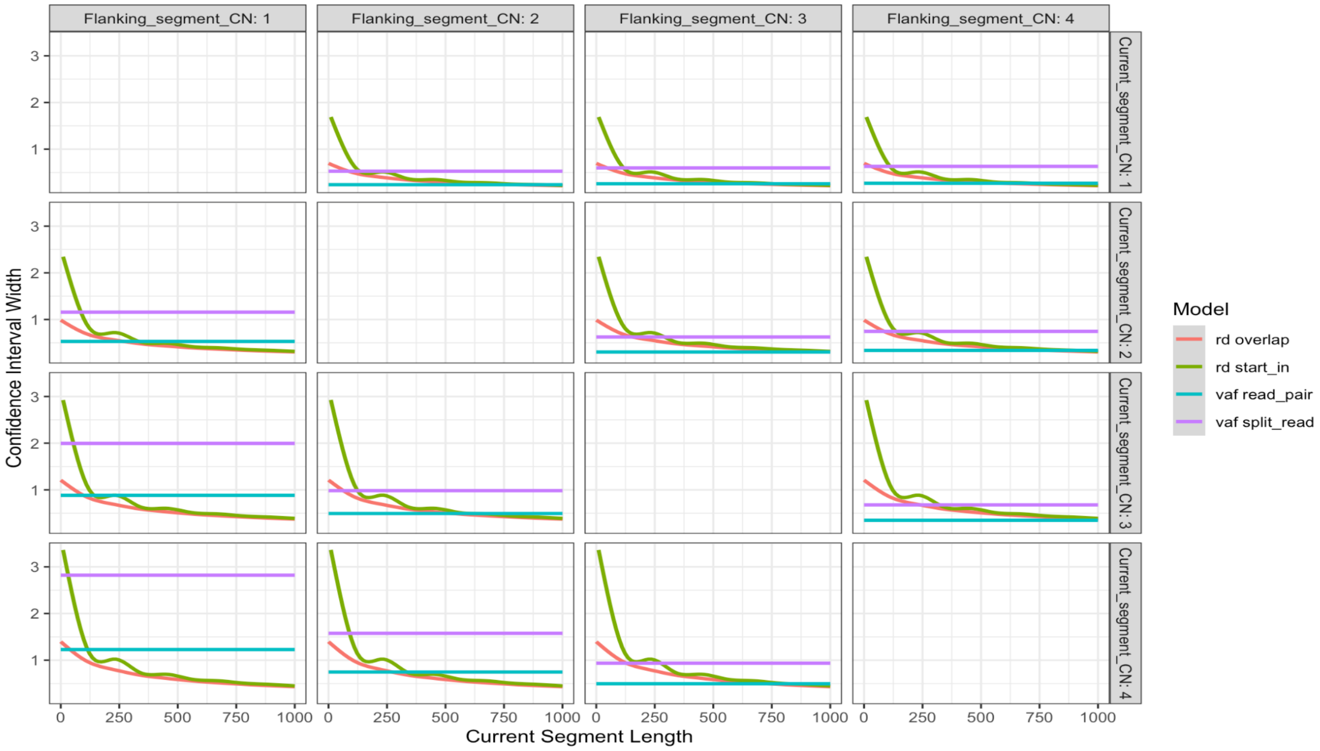


Figure 20: *CI trend for short current segment length(b) values for all copy-number inference models in 12 different copy-number states*

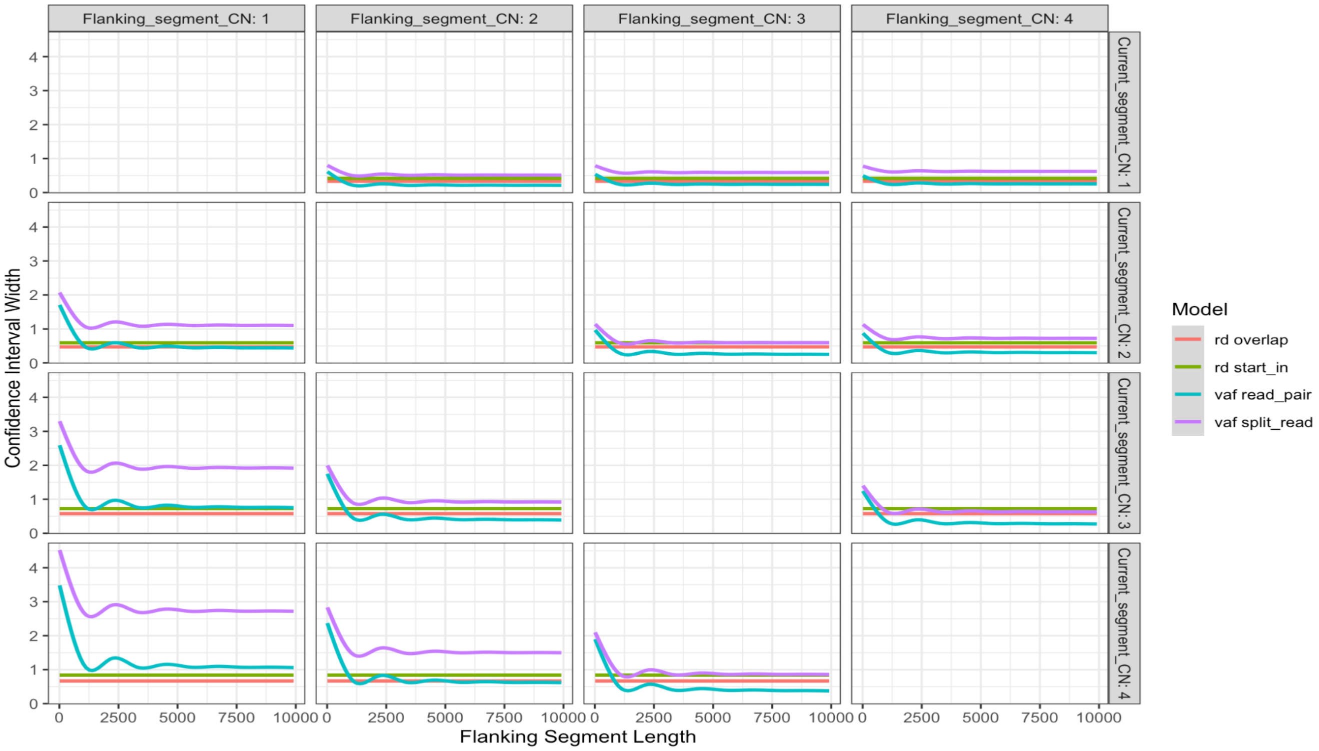


Figure 21: *CI trend for various flanking segment length(b) values for all copy-number inference models in 12 different copy-number states*

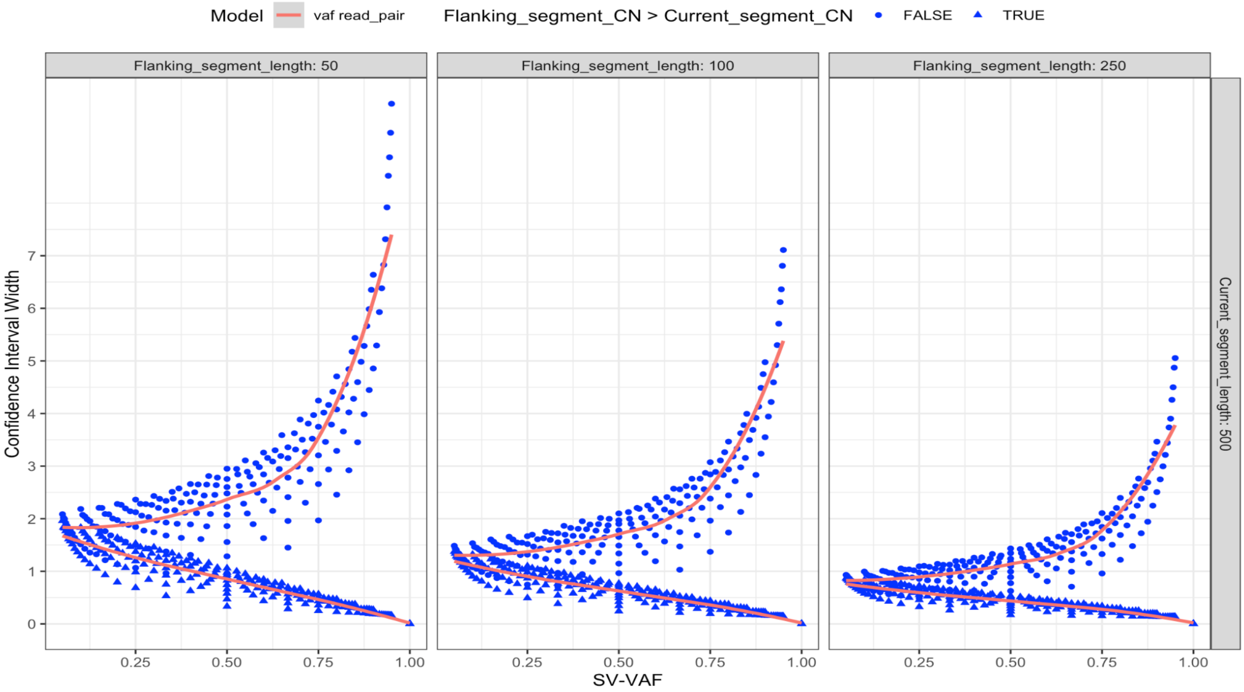


Figure 22: *CI of VAF model and corresponding VAF values for various copy-number states are represented for deletion (Flanking-segment-copy-number > Current-segment-copy-number)(blue triangle) and amplification (blue circles) events. Lower CI width values were recorded for higher flanking segment lengths, even though, current segment length remains higher than flanking segment length, the model's performance is still consistently better in deletion cases.*

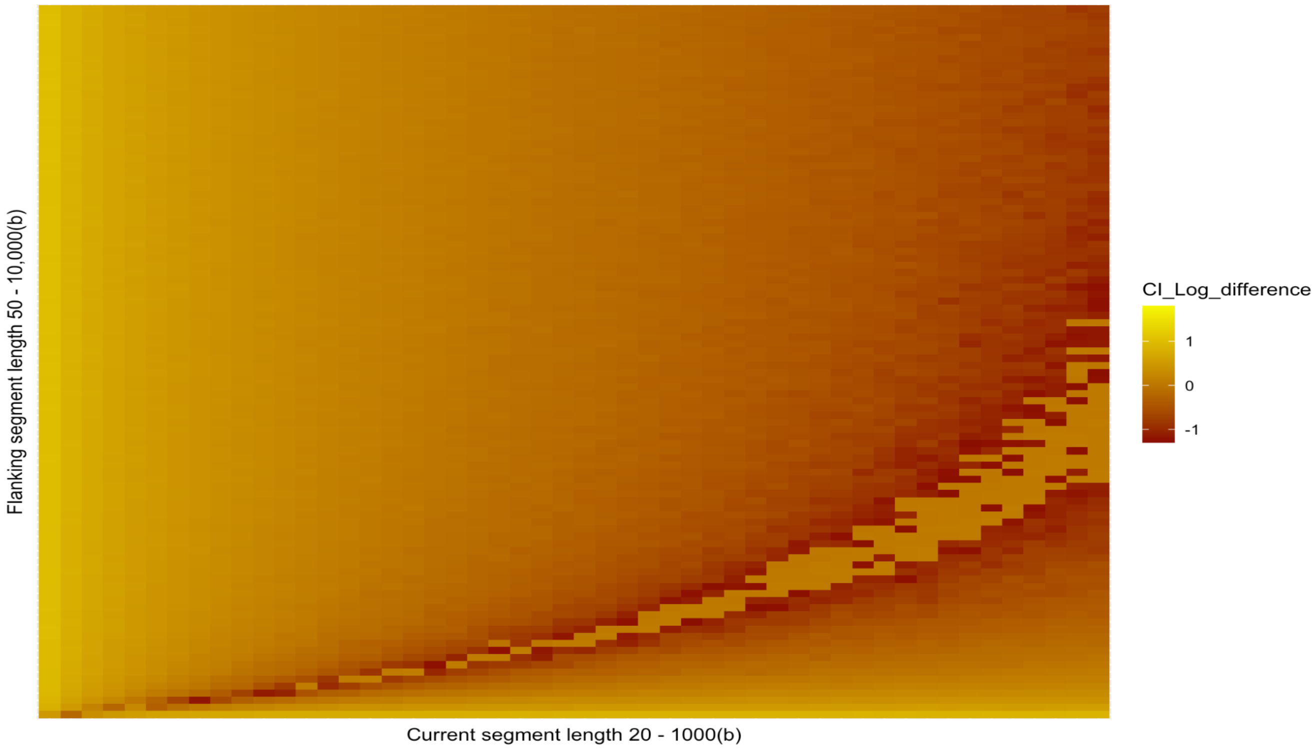


Figure 23: *Heat map based on log of CI difference between models and their corresponding flanking and current segment length values for a wide range of inferred copy-numbers. This heat map contains the same subset as the heat map figure in our results section; however, it shows the entire range of flanking segment lengths (up to 10,000) to show the general trend of model performance difference.*

8.2 Appendix 2: Sequencing Data Variant Calls

GRIDSS successfully detected most variants in truth sets 1, 2, and 3 with single-base resolution and high breakpoint position accuracy. In, table (2) 'DEL' and 'DUP' keywords refer to deletion and amplification events respectively. 'Length' refers to the detected length by GRIDSS and may be slightly different to the actual defined variant length which can be obtained by subtracting end from start positions. 'Event ID' corresponds to specified ID in GRIDSS variant call format file to match both breakpoints for a given event.

Table 2: GRIDSS variant calls on truth set 1, 2, and 3 variant clusters

Chromosome	**Start Position**	**End Position**	**Variant Type**	**Event ID**	**Length**
chr21	42,542,590	42,542,940	DEL	gridss282fb_2378	351
chr21	42,541,979	42,542,965	DEL	gridss282fb_2372	987
chr21	42,541,979	42,543,040	DEL	gridss282fb_2373	1062
chr21	42,544,290	42,544,840	DEL	gridss282fb_2384	551
chr21	42,545,440	42,546,090	DEL	gridss282fb_2386	651
chr21	42,543,790	42,544,290	DEL	gridss282fb_2382	501
chr21	42,546,090	42,546,790	DEL	gridss282fb_2387	701
chr21	42,541,979	42,543,340	DEL	gridss282fb_2375	1362
chr21	42,541,979	42,543,165	DEL	gridss282fb_2374	1187
chr21	42,544,840	42,545,440	DEL	gridss282fb_2385	601
chr21	42,546,790	42,547,540	DEL	gridss282fb_2388	751
chr21	42,548,340	42,549,190	DEL	gridss282fb_2392	851
chr21	42,543,340	42,543,790	DEL	gridss282fb_2381	451
chr21	42,542,290	42,542,590	DEL	gridss282fb_2377	301
chr21	42,542,040	42,542,290	DEL	gridss282fb_2376	251
chr21	13,309,900	13,310,550	DUP	gridss279bf_152	649
chr21	13,304,700	13,305,416	DUP	gridss279bf_137	715
chr21	13,310,225	13,311,073	DUP	MISSED	
chr21	13,309,300	13,309,900	DUP	gridss279bf_148	599
chr21	13,312,272	13,313,800	DUP	gridss279bf_154	1527
chr21	13,313,625	13,314,579	DUP	gridss279bf_157	953
chr21	13,309,500	13,311,073	DUP	gridss279bf_151	1572
chr21	13,304,625	13,305,416	DUP	gridss279bf_135	790
chr21	13,313,650	13,314,550	DUP	gridss279bf_158	899
chr21	13,306,300	13,306,500	DUP	gridss279bf_139	199
chr21	13,306,500	13,306,750	DUP	MISSED	
chr21	13,307,400	13,307,800	DUP	gridss279bf_144	399
chr21	13,306,750	13,307,050	DUP	gridss279bf_140	299
chr21	13,308,250	13,308,750	DUP	MISSED	
chr21	13,307,050	13,307,400	DUP	gridss279bf_143	349
chr21	13,312,800	13,313,650	DUP	gridss279bf_156	849
chr21	13,308,750	13,309,300	DUP	gridss279bf_146	549
chr21	27,677,600	27,677,655	DUP	gridss280bf_135	54
chr21	27,677,655	27,678,913	DEL	gridss280fb_5436	1259
chr21	27,678,913	27,679,214	DEL	gridss280fb_5439	302
chr21	27,679,214	27,679,766	DUP	MISSED	
chr21	27,679,766	27,680,456	DEL	gridss280fb_5440	691
chr21	27,680,456	27,685,456	DUP	MISSED	
chr21	27,685,456	27,687,768	DUP	MISSED	
chr21	27,687,768	27,687,836	DEL	gridss280fb_5443	69
chr21	27,687,836	27,696,836	DUP	MISSED	
chr21	27,696,836	27,696,941	DUP	gridss280bf_142	104

8.3 Appendix 3: Sequencing Data Resources

GIAB Ashkenazim trio son, HG002 sample (`HG002.GRCh38.60x.1.bam`)**sequencing data retrieved from:**
`https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novoalign_bams/HG002.GRCh38.60x.1.bam`

Corresponding reference genome retrieved from:
`https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set.fna.gz`

Sample specific exclusion regions for our reference genome were retrieved from:
`https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_GRC_exclusions.bed`