Imperial College London

Read-depth based methods for copy number variation (CNV) detection using next-generation sequencing data

Hongye Wang

8/2020

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Research at Imperial College London

Formatted in the journal style of Frontiers in Genetics.
Submitted for the MRes in Computational Methods in Ecology and Evolution

Declaration

All the data were generated from the SimulateCNVs tool, and all the experiments were designed by myself. The source code of CNVnator and cn.MOPS were downloaded from website, and the HMMploidy was given by my supervisor Fumagalli Matteo. Also, during the writing process, I received comments and feedback from my supervisor.

Word count: 5305

Contents

Abstract

Copy number variation is a common form of genetic variation that causes abnormal numbers of copies in genomic regions on chromosomes. Studies have shown that CNVs can lead to differences in individual susceptibility to disease and have genetic effects. With the development of next-generation sequencing technology (NGS) in recent years, NGS-based CNV identification and analysis has been widely used in healthy and patient individual samples. Therefore, this clinical requirement promotes the development of algorithms and tools based on NGS CNV detection. In order to test the performance of these algorithms, I first simulated CNV samples supported by whole genome-sequencing (WGS), and then introduced two popular read-depth based CNV tools and one tool with CNV detection potential. These tools were tested with simulated CNV data to see their performance in CNV detection. In addition, I discussed the advantages and disadvantages of two CNV detectors and made suggestions for future development.

Introduction

1. Copy number variation

Genomic variation consists of small insertions or deletions (indels), single nucleotide variations (SNVs), copy number variations (CNVs), and large structural variations (SVs). The size of the alteration can vary from a single nucleotide position to large microscopically-visible chromosome anomalies (Alkan et al., 2011). CNV is a type of intermediate-scale structural variation and have been found genome-wide in humans. It could be recognized as a DNA fragment which is typically ranging from 1 kb to 5Mb in size. Compared with its reference genome, such altered DNA fragment presents a different copy number. The form of CNVs include deletions, duplications, and insertions (Feuk et al., 2006). The single nucleotide changes (SNPs) was once considered as the most important and prevalent form of genetic variation in human. Currently, some studies revealed that CNVs accounts for three times the total nucleotide content of SNPs. It has been estimated that 12% of the human genome pertains to copy number variable (Redon et al., 2006). Researchers found that genes are often compassed by CNVs, and these variations could affect genetic diversity. Thus, CNVs are likely to play important roles in human phenotypes and diseases susceptibility (McCarroll and Altshuler, 2007; Valsesia et al., 2013). It has been evidenced that CNVs can cause diseases or increase the risks of many complex diseases, such as cancer, autoimmune diseases, schizophrenia and obesity (Bochukova et al., 2009; Fanciulli et al., 2007). A striking observation from the analysis of cancer cell samples shows that the most important somatic aberration is shown as CNVs, because they found oncogene activation and suppressor gene inactivation were usually caused by abnormal chromosome copy number variation (amplification or deletion) (Varella-Garcia, 2010). Since the somatic CNVs play an important role in the prognosis and treatment improvement of cancer (Dancey et al., 2012), CNV research has become a hot spot in biomedicine in recent years.

2. CNV analysis by NGS data

The emergence and development of next-generation sequencing (NGS) technology has brought revolutionary progress to CNV detection research. NGS provides accurate sequence information to the base level (Zhang et al., 2011). Sequencing-based detection realizes the prediction of CNVs by comparing the short reads data with the reference sequence and counting signals based on the comparison result (Zhao et al., 2020).

Compared with traditional array-based method, where limited genomic regions are predefined by probes in CNV chip, the way of identifying copy numbers by NGS has the following potential advantages: 1) The breakpoints of CNV regions are able to be detected with higher accuracy, since there is no need of the predefined probes (Chiang et al., 2008); 2) large copy number estimation from NGS data is more accurate, because the coverage depth is linear proportional to the copy numbers (Alkan et al., 2009); 3) we can estimate the allele-specific copy number for the targeted alleles, whereas array-based approaches have the limitation that they can only detect predefined alleles. Researchers are becoming more and more interested in allele-specific copy numbers, because it allows determining whether alleles have functional integrity, which is very important, for example to identify mutations that lead to cancer development (Stratton et al., 2009).

Whole genome sequencing (WGS) and whole exome sequencing (WES) are the two major platforms for DNA sequencing (Zhao et al., 2013). WGS aims at sequencing the complete genome as its name suggests. In contrast, WES focuses on just the protein coding sequences, often with high coverage. In this project, I focused on the WGS simulation and copy number variants estimation. Compare to WES, WGS has the drawbacks of high cost and low speed, but there are many proponents of WGS, due to its less requirements on input data and more comprehensive CNV information (Meienberg et al., 2016).

3. Strategies for CNV detection

Due to the advantages of detecting copy number variations through NGS data, different kinds of CNV detectors have been developed rely on different attributes of NGS data. There are five main strategies for detecting CNVs: Paired-end Mapping (PEM) method, Split-Read (SR) method, Read-Depth (RD) method, *de novo* assembly (AS) method and comprehensive approaches (CB) (See Fig. 1) (Zhao et al., 2013). Although different strategies have different advantages and drawbacks, read-depth based approach is the major method to detect copy numbers in recent researches (Teo et al., 2012).

Based on the underlying hypothesis of read-depth method, researchers have developed some effective detection algorithms. Here, my project aims at evaluating three read-depth based detecting software: CNVnator, cn.MOPS and HMMploidy. The first two are popular and representative CNV detectors with different characteristics. The HMMploidy was created initially for inferring ploidy level but with potential of identifying CNVs (SamueleSoraggi/HMMploidy, 2020, unpublished), which was indicated in previous study. I hope to test whether the HMMploidy can identify CNVs and compare the detecting performance between three tools.

Methods

1. Simulation NGS data for WGS

To evaluate the performance of CNV detectors, here I used simulated datasets for further detections. There are two main advantages in using the simulated data: 1) I can alter the parameters easily to generate various data that meet my experimental requirements, like different depth of coverage, reads length and total CNV numbers. Notably, since the coverage of the short reads (or the total number of short reads) impacts the price of the NGS sequencing directly, it is important to know the minimum coverage of data needed for CNV detection; 2) the simulation will provide a known list of benchmark CNVs which shows the specific location and size of the simulated CNV regions along certain chromosome (Zhang et al., 2011). It is a gold standard for calculating the precision and recall of the CNV detectors. By simulating CNVs in WGS and WES datasets, the operational characteristics of existing and novel detection approaches can be comprehensively compared.

Many mature simulators have been developed to simulate copy number variants in NGS data. Here, I tried three different simulators, Xome-blender (Semeraro et al., 2018), CNV-sim (NabaviLab/CNV-Sim, 2020) and SimulateCNVs (Xing et al., 2018).

Compare with their simulation performances by simulating different depth samples in one chromosome, the SimulateCNVs tool performed best. SimulateCNVs is mainly based on python. It has many user-friendly features to simulate CNVs for both WGS data and WES data. Compared with SimulateCNVs, CNV-sim occasionally generates some unreasonable results, and most importantly, I found it was hard to use it for different depth simulation. As for Xome-blender, due to some bugs inside the source code, this tool cannot produce NGS data as I expected. SimulateCNVs obviously overcomes the known limitations of existing simulators and shows great power in simulation performance.

To generate shorts reads file for WGS sequence, one or more specific reference sequence file is needed. The reference chromosomes I used to simulate CNVs are chromosome 1 and chromosome 20, which are extracted from humanG1Kv37 references. On the basis of these two reference sequences, I determined the features of CNVs and the distribution of their copy numbers. For each dataset, 20 CNV fragments have been created on chromosome 20 and the location of CNV regions were randomly along the reference sequence. Totally 26 datasets were simulated from chromosome 20 with different coverages and CNV length as planned, for the purpose of testing the sensitivity and precision of each CNV detector. Chromosome 1 has fourfold length than chromosome 20. Totally, 150 CNV regions were generated on it randomly and 5 datasets were produced. They will be used to make a comprehensive comparison among all detectors.

2. Data processing

After obtained the paired-end short reads files, these raw data need to be performed a series of pre-processing to make them meet the conditions accepted by further detectors. The first step is mapping the short sequences against the reference genome. Although the short reads generated by the simulator originally came from an ordered genome, after DNA library construction and sequencing, the sequence relationship between different reads in the file has been lost. Therefore, we need to compare each short sequence with the reference genome of the species, find the position of each read on the reference genome, and then arrange them in order. Here I used BWA (Li and Durbin, 2009) to map the FASTQ files, it combines the BW(Burrows-Wheeler) compression algorithm with the suffix tree, which allows me to obtain accurate sequence comparison results at a quick time and small space cost. Next step is sorting the BAM file produced by BWA with the help of Samtools (Li et al., 2009). This step provide a sorted BAM file for downstream analysis. For some detection requirements, I filtered out all gaps, low quality reads and segmental duplications from the sorted data. The quality control report of the final modified data can be generated by Qualimap (García-Alcalde et al., 2012).

3. Read-depth based methods (RD)

The design idea of read-depth method is based on the correlation between the coverage depth of a genomic region and the copy number of this region (Teo et al., 2012). When the copy number changes in a certain region in chromosome, the number of short reads mapping this special region in sequencing data will increase or decrease significantly compared with other regions (See Fig.1) (Genomes. 2nd edition, 2002). If we compare RD -based approach to PEM and SR-based approach, it can be found that only RD-based method can predict the exact copy numbers, because the other two methods only use the position information. In addition, it is difficult for PEM/SR methods to identify large insertions as well as CNVs in complicated genomic area, however, using RD-based methods will make it easier and more accurate.

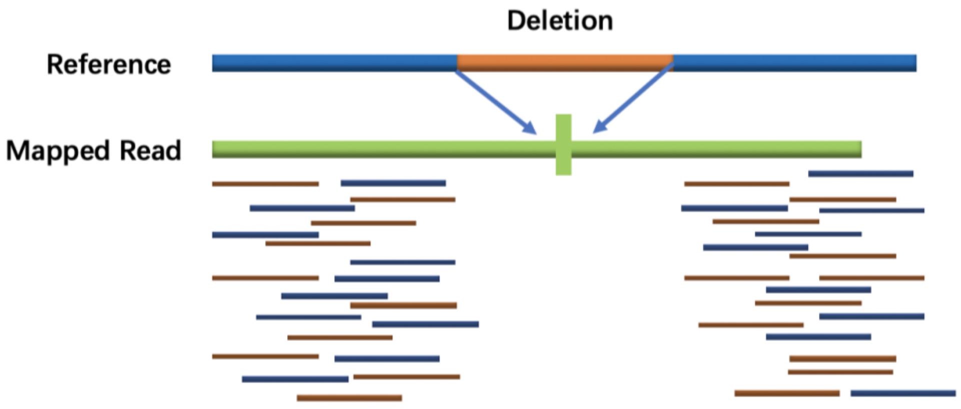

Fig. 1 Read depth-based method detects CNV. In this figure, there is a deletion region on reference sequence. Short reads are mapped to the reference genome except the deletion region, which gives a different depth distribution along the chromosome.

Based on the different research purposes, RD-based methods are often classified as three types: single sample, paired case, and a large population of samples (Pirooznia et al., 2015). In the single sample category, only the absolute copy number will be reported since there is no other subject available; in paired case, the report shows comparison between case and control; in terms of large population samples, the average RD from population will be calculated to estimate the copy number and predict the CNVs. There are some general steps when detecting CNVs by read-depth methods. Initially, mapping all the short reads to the reference sequence and calculate their read depth by a predefined window. The window size is an important factor that can affect the sensitivity and accuracy of CNV detectors (Boeva et al., 2010). Then the counts of read depth need a normalization to reduce GC bias as well as biases from repeating sequence. After that is a segmentation process for identifying a consecutive set of windows with similar amount of CNVs (Janevski et al., 2012). The last step is to estimate the significance, followed by a filtration. (Janevski et al., 2012; Zhao et al., 2013).

4. Estimating CNVs with CNVnator

CNVnator (Abyzov et al., 2011) is currently one of the most popular detection software. It uses the established mean-shift approach with addition corrections for multiplebandwidth partitioning and GC correction for more accurate CNV detection (Pirooznia et al., 2015). Mean shift is a hill climbing algorithm that involves iteratively moving the kernel to a higher density region until it converges (Yizong Cheng, 1995; Yang et al., 2003). Each shift is defined by an average shift vector. The average displacement vector always points in the direction of the greatest increase in density. We can think that the RD signal map of the chromosome sequence is an image that needs to be processed to identify different genomic CNV regions. Our goal is to find the maximum density in these mixtures. The mean shift process can move each data point along the mean shift vector to the place of these density maximums without directly estimating the density.

The workflow of calling CNVs by CNVnator include six steps (see Fig.2) and a reference genome file is needed with your NGS data. I designed three subsets of experiments for CNVnator to test 1) the effect of bin size on the sensitivity of detection under different coverage depths; 2) the difference of sensitivity between gains and loss at the same coverage level; 3) the sensitivity to detect CNVs of different size.

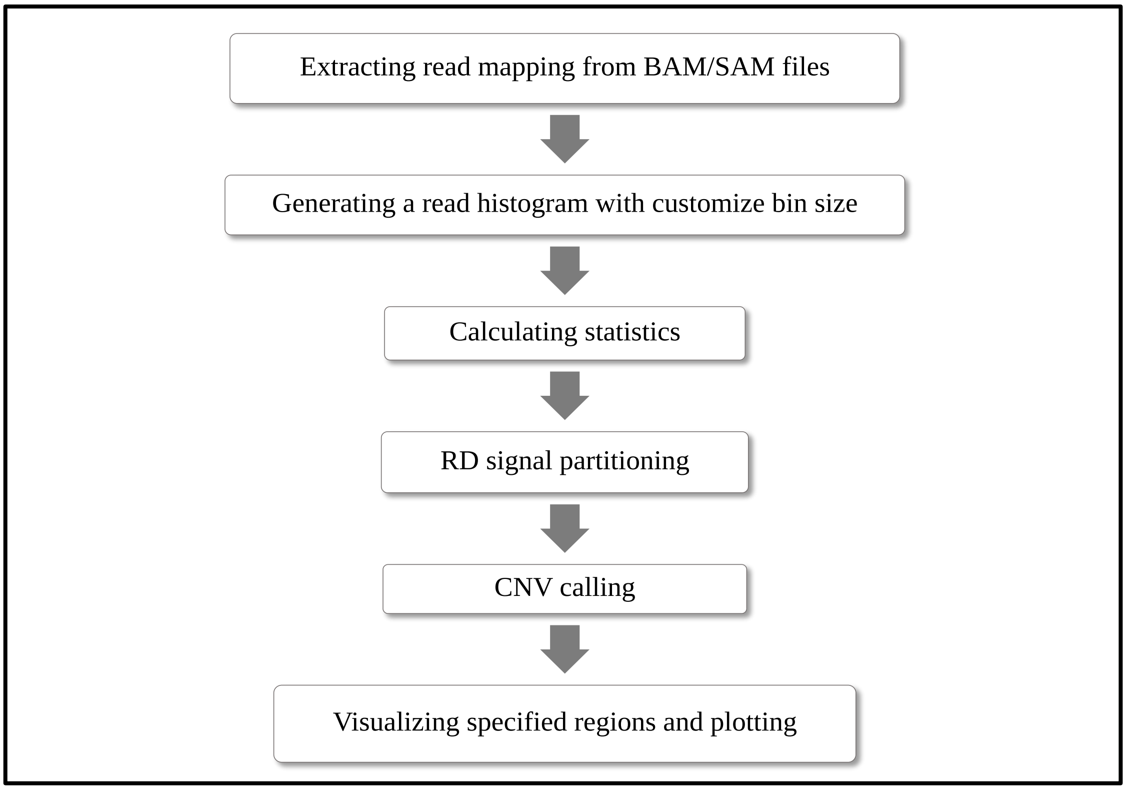
Extracting read mapping from BAM/SAM files
Generating a read histogram with customize bin size
Calculating statistics
RD signal partitioning
CNV calling
Visualizing specified regions and plotting

Fig. 2 Workflow of calling CNVs by CNVnator.

5. Estimating CNVs with cn.MOPS

cn.MOPS (Copy number estimation by a Mixture Of PoissonS) (Klambauer et al., 2012) is a CNV detector using multiple genome samples to identify CNVs by cross-comparison of their read counts along sequence. Using the Bayesian method, it utilizes mixed components and Poisson distribution to decompose the reads changes across different samples into integer copy number variants and noise (Zhang, Bai, Yuan and Du, 2019). In addition, the cross-comparison approach rises statistical power whilst reduces computational cost in the detection process (Pirooznia, Goes and Zandi, 2015). Different from the CNVnator, cn.MOPS does not need a referenced genome sequence as an input file, instead, a group of test files(more than 5) are required to run out a credible CNV calling result. The mechanism of cn.MOPS is shown in figure 3. For each genome sample, a detection for read depth will be applied across whole genome shown by curves. A segmentation algorithm will then be used for all samples to generate segments for cross-comparison between samples, which are green boxes. If cross-comparison finds variation across samples in certain segment, such segment will receive an informative/non-informative (I/NI) call, the abnormal curves within this segment will be recognized as CNVs shown as red box. The blue box marks those abnormal curves but with no significant variation across samples. (Klambauer et al., 2012).

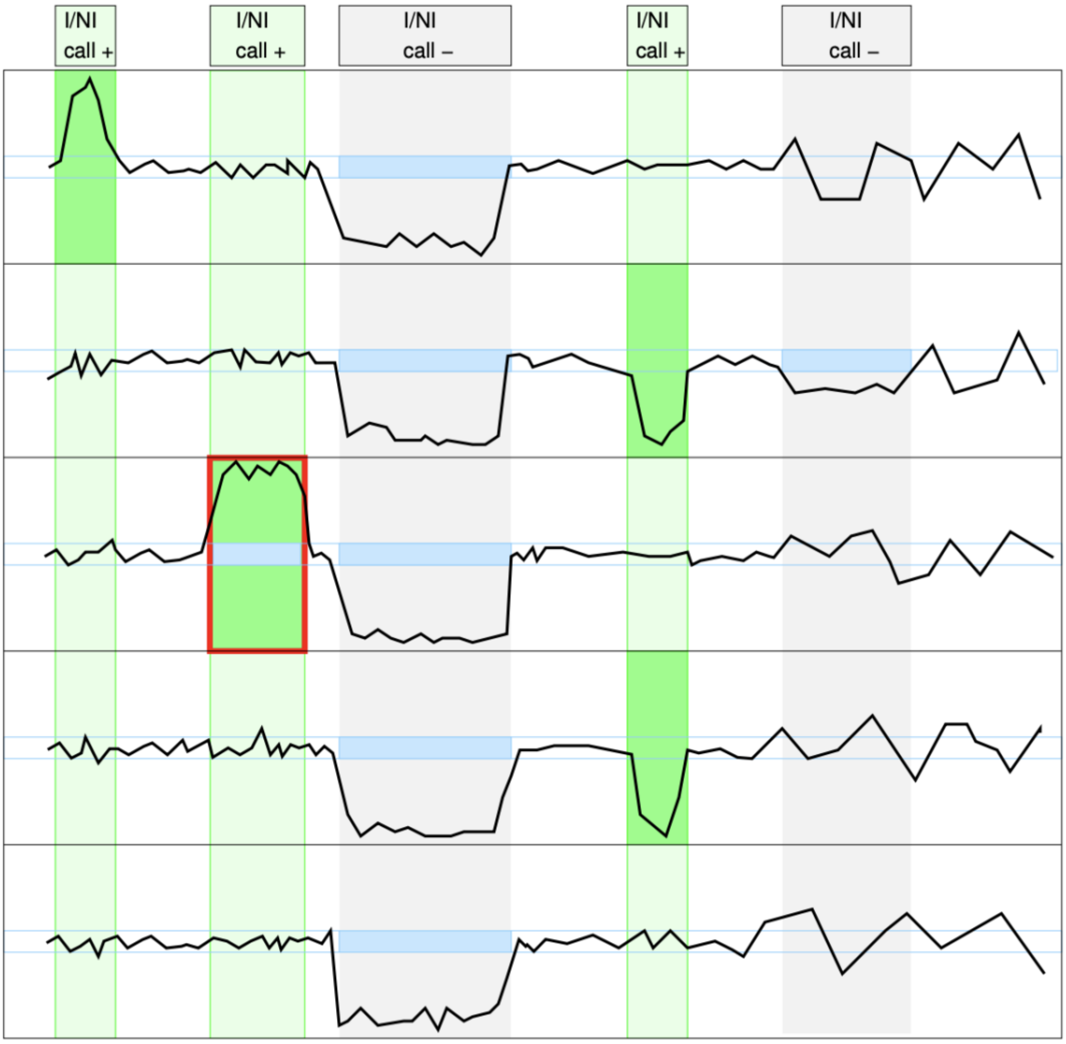

Fig. 3 Basic mechanism of cn.MOPS. Curves present read counts along each genome sample. Green boxes are the segmentation outcome for each sample. If cross-comparison reveals significant variation, the I/NI call will be positive and vice versa. The red box is the CNV call, that is both abnormal curves and variation across samples occur. The blue boxes mark the I/NI call negative segments with abnormal curves. (Klambauer et al., 2012)

The workflow of calling CNVs by cn.MOPS is showing below.(See Fig.4) Three subsets of experiments have been set on cn.MOPS to see: 1) the effect of window size on the sensitivity of detection under different coverage; 2) the difference of sensitivity between gains and loss at different coverage level; 3) the sensitivity to detect CNVs of different size.

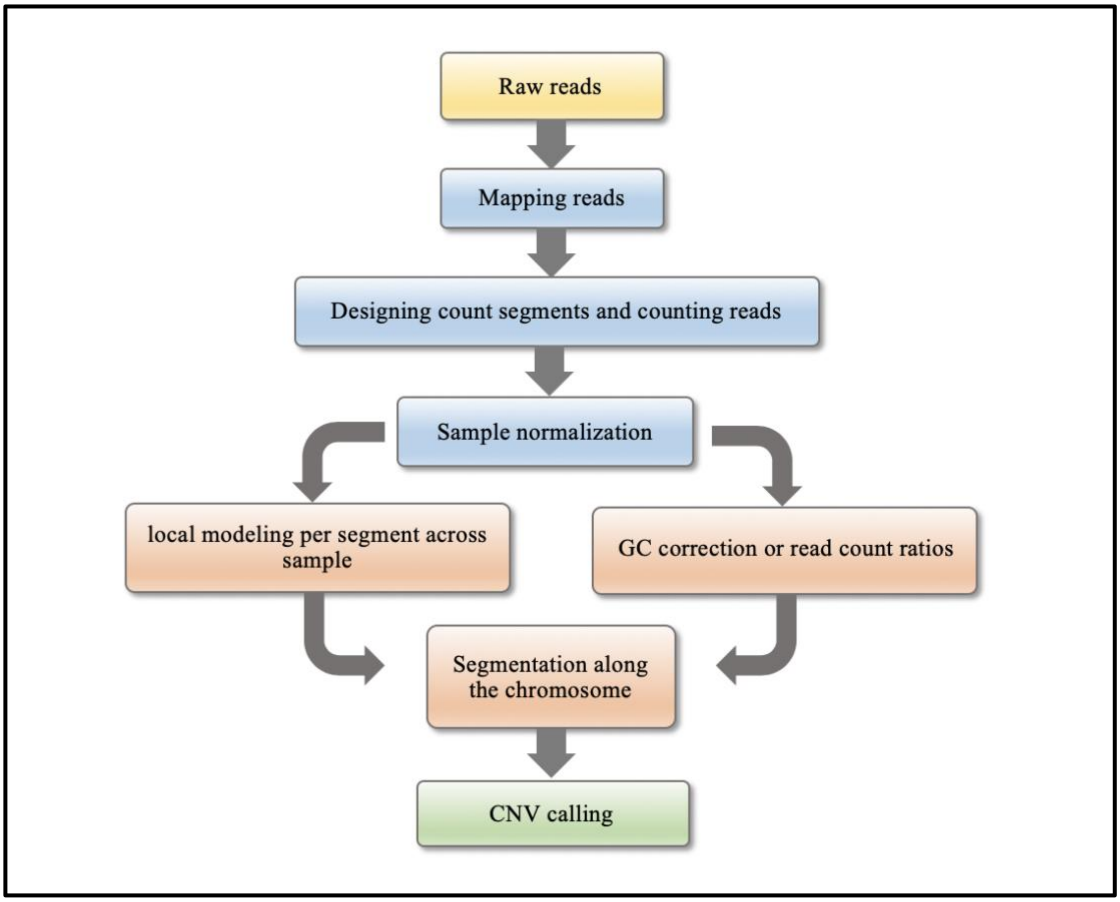
Raw reads
Mapping reads
Designing count segments and counting reads
Sample normalization
local modeling per segment across sample
GC correction or read count ratios
Segmentation along the chromosome
CNV calling

Fig. 4 workflow of calling CNVs by cn.MOPS

6. Estimating CNVs with HMMploidy

HMMploidy is a software which uses the combination of sequencing depth and genotype likelyhoods to infer ploidy variations from low sequencing data. HMMploidy comprises a Hidden Markov Model (HMM) (Rabiner, 1989) with observations being both sequencing depth levels and observed reads. From the further study result, researchers found an interesting phenomenon that some genome regions have an unusual depth alter of coverage in chromosome, so they conjectured that suggests the presence of copy number variants. Therefore, I want to test the probability of HMMploidy to estimate the copy numbers and the specific position and length of CNVs on chromosome (SamueleSoraggi/HMMploidy, 2020, unpublished).

7. Performance assessment

To determine the performance of each tool, the following measurements are used:

True positive rate (TPR, equal to sensitivity) and precision. These are defined as:

TPR = (true detected CNVs) / (all detected CNVs)

Precision = (true detected CNVs) / (number of true CNVs)

“True” CNVs were defined as those known by the simulation CNV list. Also, the true detected CNVs should have an overlap of 50% or more with a benchmark CNV region.

Results

1. Simulation results

From the quality control report produced by Qualimap, we can visualize same features of the simulation datasets. For example, see quality control result of simulation on chromosome 20 at mean depth of 10X below.

1.1 Depth of coverage

A genome’s coverage depth is from the equation of total amount of bases matched to the genome divided by the length of such genome. It shows how strong a genome is "covered" by sequenced fragments (short reads). However, there are some certain stretches of a genome contain sequence that is difficult to sequence, mainly due to repetitive regions, tracks of the same base, GC composition, closed DNA, etc. As a result, “gaps” with 0 or extremely low coverage are shown on the chromosome (See Fig.5) (Audano et al., 2019). Human chromosome 20 (chr20) currently has three unfinished gaps remaining on its q-arm (Minocherhomji et al., 2012). From the coverage across chromosome, we can also see that many peaks and valleys appear around the average coverage alternately. It means depth of these positions has changed more or less, implying the possibility of copy number variation at this site.

1.2 GC content

GC-content could be referred to a certain fragment of DNA/RNA or to an entire genome. GC bias in NGS data is known to aggravate genome assembly and GC bias may cause the coverage depth change, which can lower the accuracy of assembly (Piovesan et al., 2019).

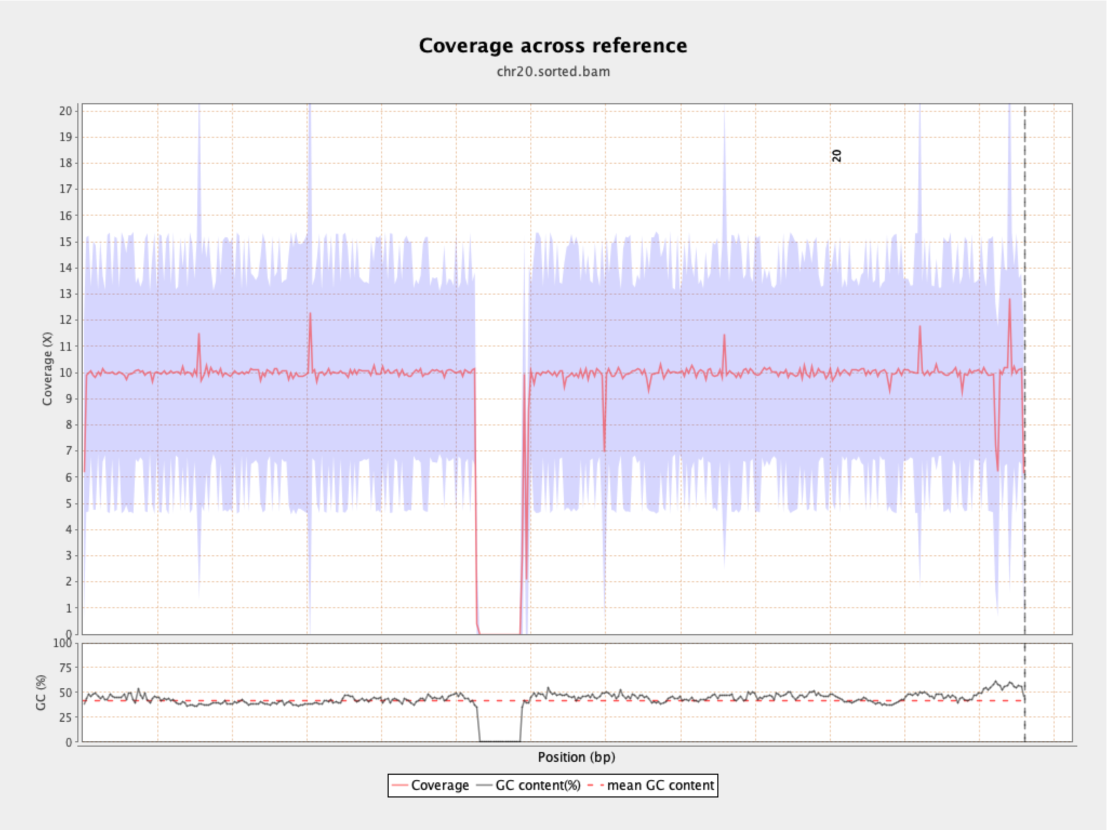

Fig. 5 Coverage across chromosome 20. There are two figures in the result plot. The upper picture gives the coverage distribution (red line) and coverage deviation (purple background) across the reference chromosome. The mean coverage is 10X in this sample. The lower picture gives GC content across the reference chromosome (black line) with its average value (red dotted line).

2. Performance of CNVnator on simulation data

2.1 Visualizing CNV regions

Figures 6 and 7 show two segments of deletion (10065742—10073363) and duplication (19330138—19336448) distribution on chromosome 20 detected by CNVnator. The changes of depth across the sequence can be seen very clearly, which presents CNV visually.

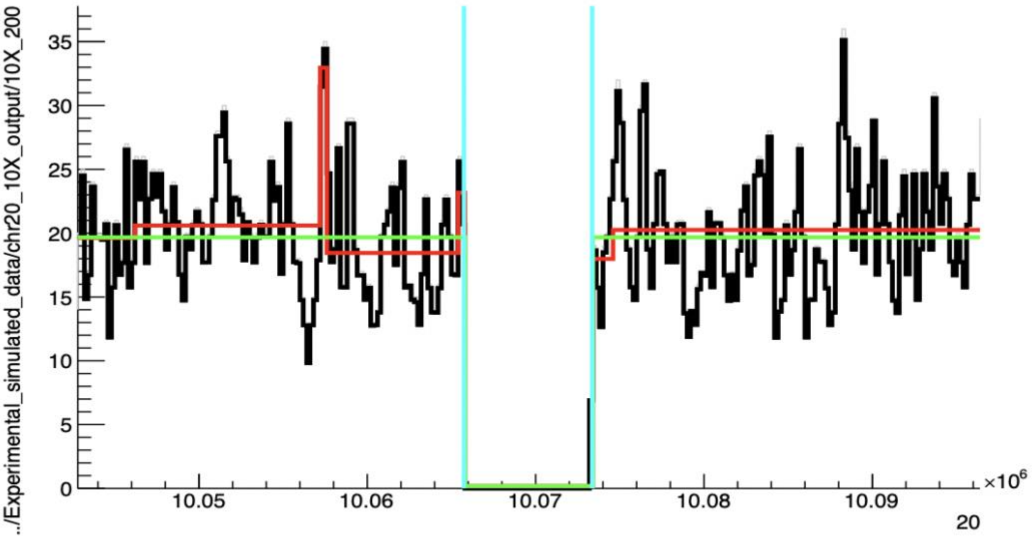

Fig. 6 distribution of one deletion region on chromosome 20 (10065742—10073363). The x-axis represents the positions on chromosome 20, the y-axis represents the coverage depth.

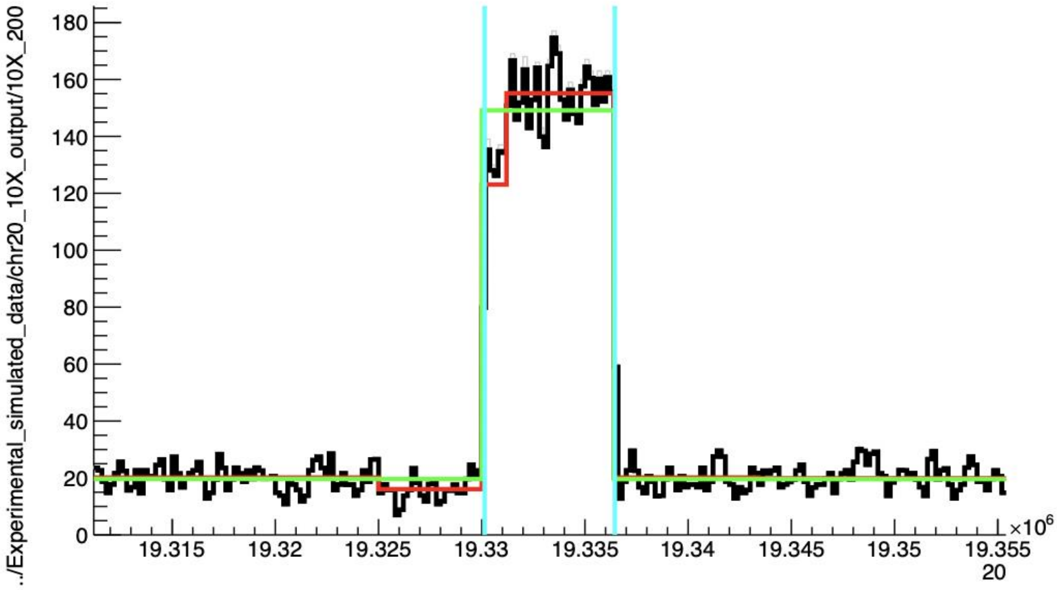

Fig. 7 distribution of one duplication region on chromosome 20 (19330138—19336448) The x-axis represents the positions on chromosome 20, the y-axis represents the coverage depth.

2.2 Bin size and coverage affect CNVnator estimation

Here, 4 groups datasets with the depth of 5X, 10X, 30X and 50X have been applied different bin sizes from 50bp to 2000bp respectively. Each dataset contains 20 CNV regions simulated from chromosome 20 randomly. The length of CNV region ranging from 1kb to 10kb. The results in Figure 9 show that as the bin size becomes larger, the TPR value decreases rapidly, and precision value basically does not change after the bin size becomes larger to 500 and remains at 100% level. This is because the smaller the bin size is, the more false positive signals or noise attend to be recorded, and the precision will be lower. When the window becomes larger, although the noise is ignored, some short-length CNVs will also be ignored, so the sensitivity will decrease. At this time, the detected CNVs are basically large CNV fragments. Figure 8 also shows that different coverage of datasets has different TPR. High coverage data has a better performance than lower coverage data, which means CNVnator is more suitable for detecting medium and high-depth reads. I tried the low coverage data in 1X and 3X as well, but the TRP always close to 0 since there are too many noises are counted. This reminds us to keep the coverage above 5X when using CNVnator to call CNVs.

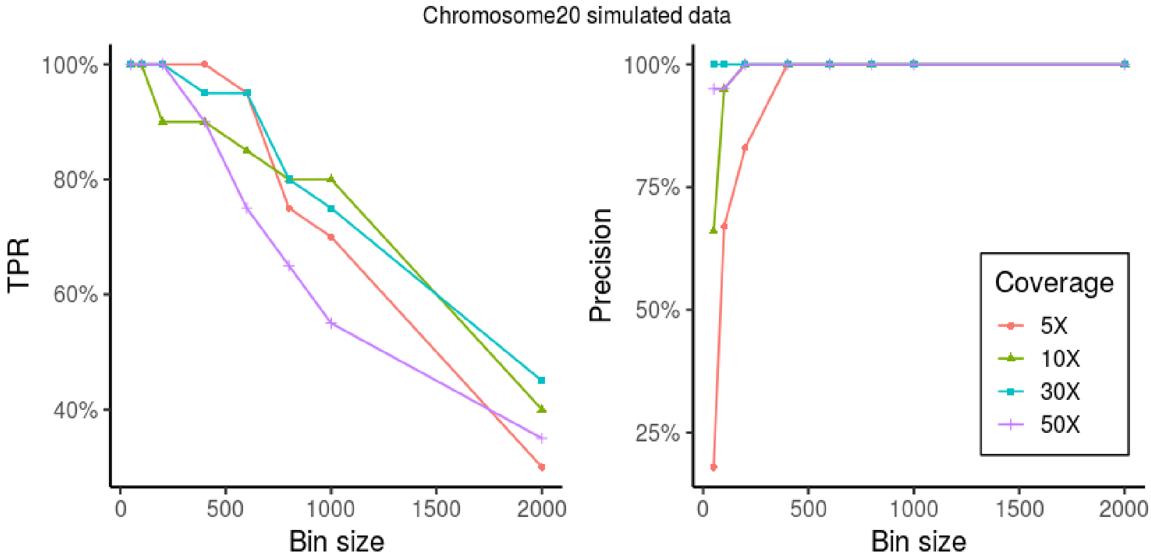

Fig. 8 Detection TPR (left) and precision (right) under different bin size in different coverage through CNVnator. Different coverages from 5X to 50X were shown in different colors. The x-axis of the coordinate system represents the level of bin size, the y-axis

represents the true positive rate (TPR) and precision respectively. The dataset was simulated from chromosome 20.

2.3 Bin size affects copy number gains and losses

When studying the effect of bin size on CNVnator sensitivity, I found an interesting phenomenon: with the bin size enlarged, more true CNVs were lost, but most of the lost CNVs are constituted by gains (duplication) segments, while the losses (deletion) will be remained even their length are shorter than gains. Therefore, 40 gains and losses segments were simulated on chromosome 20 respectively in the coverage of 10X, to see the sensitivity of CNV type detection with different bin size. Result shows in figure 9, it is obvious that the alter in bin size has a much greater impact on detecting losses than gains. As the bin size becomes larger, the number of losses remains at a high level, while the number of losses drops sharply. It indicates that CNVnator is more sensitive on detecting losses segments.

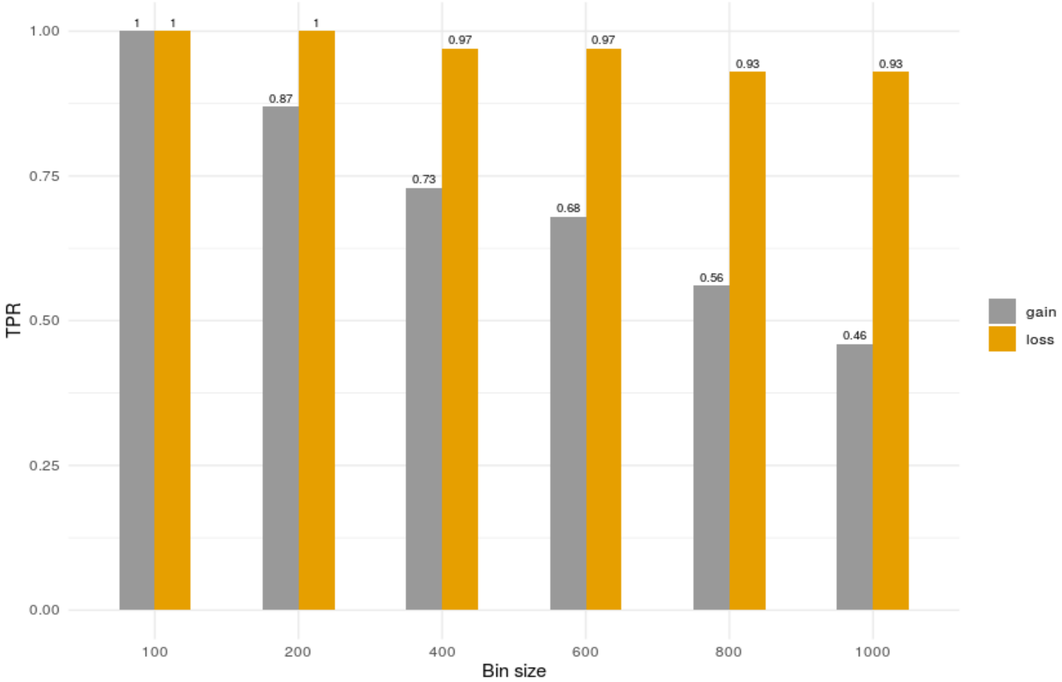

Fig. 9 Detection TPR of copy number gain and loss under different bin size through CNVnator. Two different colors show the true positive rate (TPR)of detected gains and losses.

2.4 Sensitivity of different CNV length

Random CNVs has been generated on reference sequence, the length of CNV segments ranged with 1000bp to 10,000bp. CNVnator shows good performance at detecting short CNV segments, even when the CNV length is around 1000bp, the TRP is still over 60%.

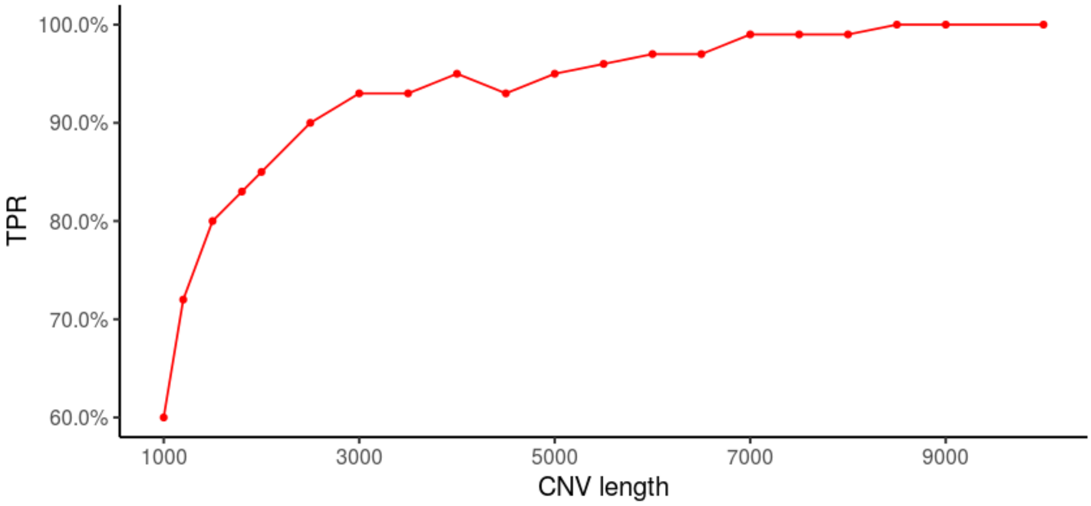

Fig. 10 Detection TRP of different CNV length at coverage of 10X through CNVnator. The x-axis of the coordinate system represents the level of CNV length, the y-axis represents the true positive rate (TPR).

3. Performance of cn.MOPS on simulation data

Similar experiments have been set on cn.MOPS. Four different coverage depth of simulation datasets have been used, including 1X,10X,30X and 50X. Each depth has 5 samples as a group to fill the input requirement. The other parameters follow the CNVnator.

cn.mops also allows for plotting the detected CNV regions (See Fig. 11). figure 12 shows that cn.MOPS has the ability of estimating CNVs at a very low coverage (1X), but performing better at the high coverage data. Also, with the bin size enlarged, the TPR value decreased dramatically while the precision increased at beginning and maintain a relatively stable level. From the figure 13, I find that cn.MOPS has almost the same sensitivity to gains and losses, but it is still slightly more sensitive to losses, which is related to the depth of NGS data. Finally, from the CNV length against detection sensitivity result, cn.MOPS is more inclined or better at detecting long CNV segments, preferably larger than 10kb.

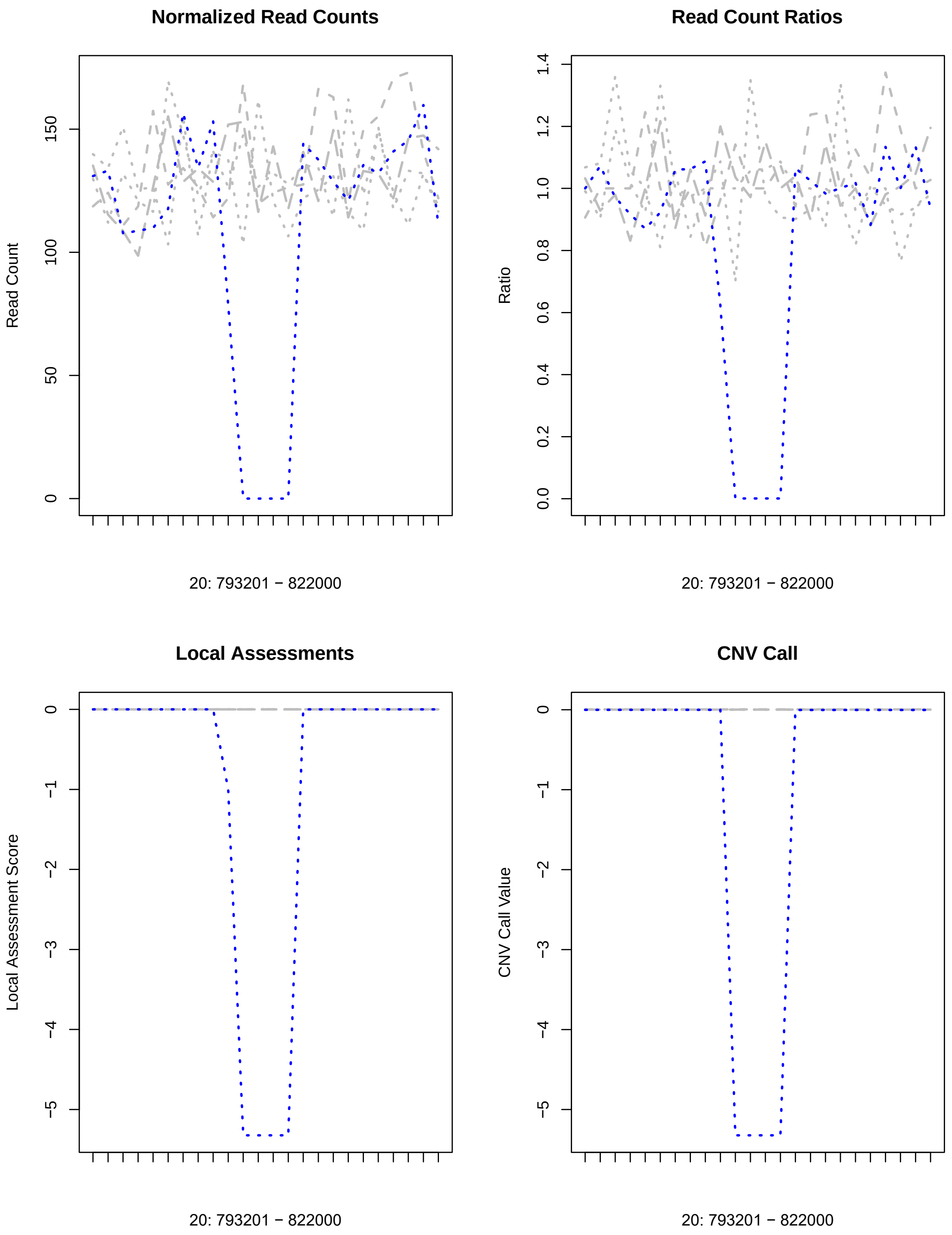

Fig. 11 one CNV region (deletion) detected by cn.MOPS: The x-axis represents the genomic position and on the y-axis we see the read counts (left), the call of the local model (middle) and the CNV call produced by the segmentation algorithm. Blue lines mark samples having a copy number loss.

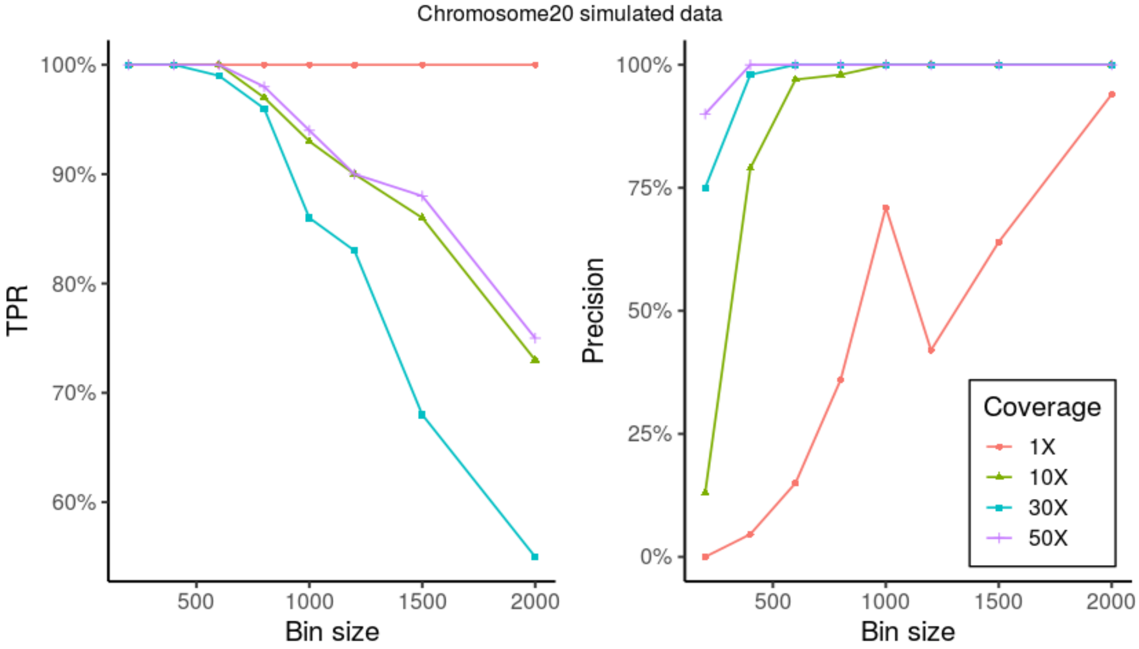

Fig. 12 Detection TPR (left) and precision (right) under different bin size in different coverage through cn.MPOS. Different coverages from 1X to 50X were shown in different colors. The x-axis of the coordinate system represents the level of bin size, the y-axis represents the true positive rate (TPR) and precision respectively. The dataset was simulated from chromosome 20.

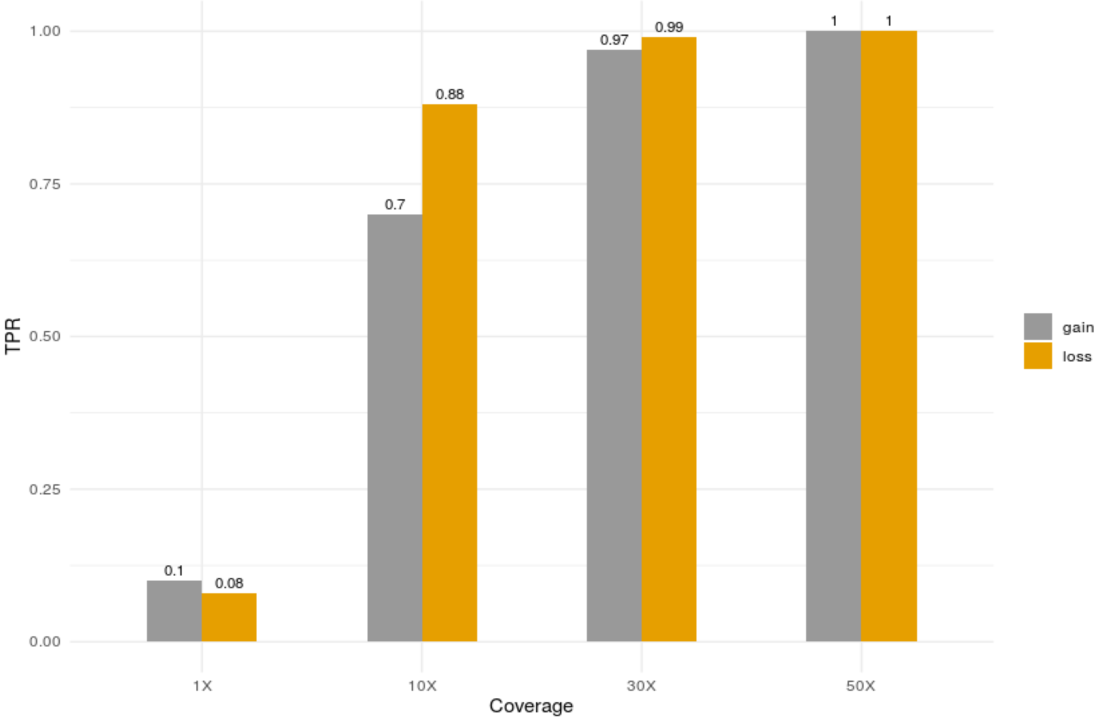

Fig. 13 Detection TPR of copy number gain and loss under different coverage through cn.MOPS. Two different colors show the true positive rate (TPR) of detected gains and losses. The x-axis of the coordinate system represents the level of coverage.

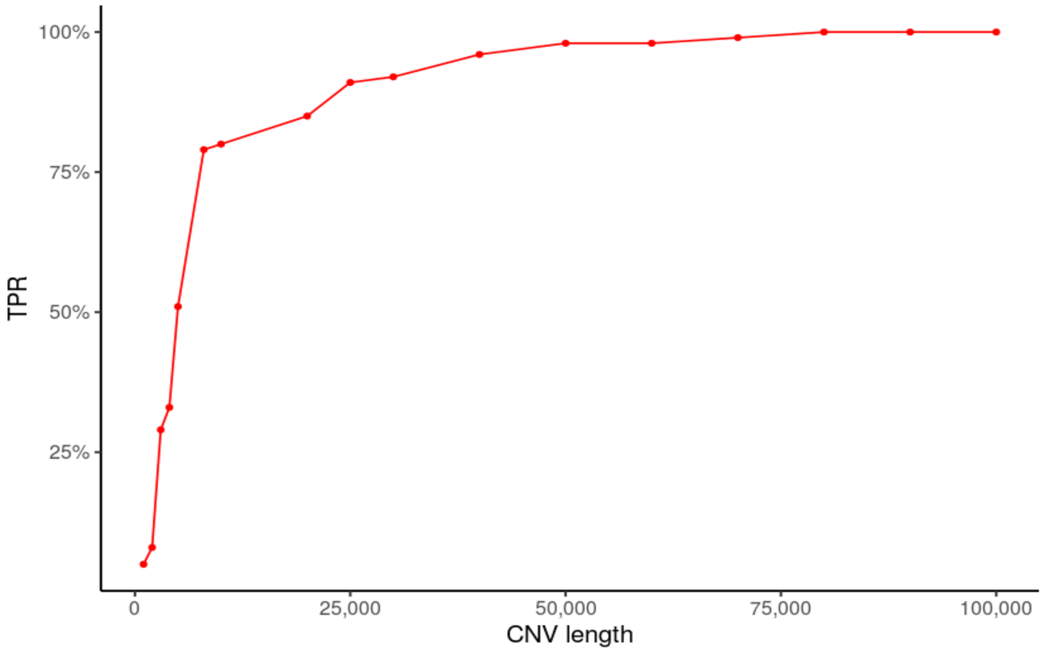

Fig. 14 Detection TRP of different CNV length at coverage of 10X through cn.MOPS. The x-axis of the coordinate system represents the level of CNV length, the y-axis represents the true positive rate (TPR).

4. Performance of HMMploidy on simulation data

Simulation datasets are set at a low and a mid-coverage (1X and 10X) to meet the requirement of HMMploidy, each dataset contains 20 CNVs ranging from 1kb to 10kb simulated by chromosome 20. The window size is fixed on 1000kb. Unfortunately, I did not get the positive result as expected. Although HMMploidy can show a certain degree of chromosome depth changes, it cannot accurately predict the length and position of CNV region. It is possibly due to a bug the main author is currently fixing.

5. Comparison between CNVnator and cn.MOPS

I randomly simulated 150 CNV segments on chromosome 1, 5 samples are produced. Then, use both CNVnator and cn.MOPS to run these samples. The distribution of CNV length and the CNV number of each length region in sample 1 are shown below (See, Table. 1). From the detected true positive (TP) CNVs and total number of CNVs, I calculate the TPR and precision for CNVnator and vn.MOPS. Although they are similar in TPR performance, CNVnator has a better precision than cn.MOPS obviously. This may be due to cn.MOPS's low sensitive in detecting short CNV fragments and the number of sample is not enough for cn.MOPS.

CNV length (kb)	**Number of CNVs**	**Identified by CNVnator**	**Identified by cn.MOPS**
< 3	8	4	0
3-7	23	22	15
7-10	15	15	12
10-50	8	8	7
50-100	10	10	9
100-500	36	36	21
500-1000	50	50	27
Total TP CNVs	150	145	91
Total detected CNVs	-	208	138
TPR	-	70%	66%
Precision	-	97%	61%

Table.1: comparison result between CNVnator and cn.MOPS performance on detecting CNVs for sample one. In the first column, I classified the CNV segments into seven ranges; the second column shows each group in sample one contains how many CNV segments, as well as the total CNV amount; the third and fourth column show how many true positive (TP) CNVs are detected by CNVnator and cn.MOPS within each range respectively, and the total amounts of detected true positive CNVs. In the last three rows, all the CNVs detected by CNVnator and cn.MOPS have been recorded, including true positive CNVs and false positive CNVs. Hence, I calculated the true positive rate (TPR) and the precision.

Discussion and conclusion

Evidence is accumulating that CNVs play important roles in human disease and CNV detection is a crucial task when clinically diagnosing the genetic diseases. Now the emergence of NGS provides a new method to detect CNV. Before implementing NGS-based CNV testing, researchers need to conduct extensive validation to evaluate the effectiveness of the new method (Feliubadaló et al., 2012). Many read-depth based CNV detection tools have been developed. I selected two representative well-known tools—CNVnator and cn.MOPS for testing and evaluation. CNVnator is a mature CNV detection software based on reading depth. Compared with other identification tools based on whole genome data, it can obtain better CNV boundary resolution. The main method of CNVnator is based on the mean shift theory. First, the detector segments the whole genome sequence into several non-overlapping bands with same size. Then consider the aligned reads of each band as the read-in depth signal. Finally, CNVnator predicts the copy number variation of each genome fragment by calculating the P value of the t-test of a sample to test whether the average RD signal of the fragment is simalar with the average at genome-scale. In an extensive comparative study, CNVnator also performed well in terms of breakpoint location and copy number estimation (Duan et al., 2013). Judging from the test results, there are still some short backs in CNVnator. CNVnator's detection results for low coverage data are not accurate, and it may only be suitable for use in medium or high coverage NGS data. Data detection of different depths is more obviously affected by the bin size, and the setting of the bin size requires investigating previous experiments or pre-experiments to find the best bin value that suits your data depth and read length. For the gap regions that naturally exist on the chromosome, if they are not removed in advance, CNVnator will judge them as deletions and output gap genome regions as the detected CNV results.

On the other hand, the main method of cn.MOPS is incorporating a probability model to decompose reads variants along genome sequences into integer form of copy numbers and noise by means of its Poisson distributions and mix components respectively. A Dirichlet prior is the key for cn.MOPS to control the FDR for CNV detection. The Dirichlet prior sets the default value of copy number as 2 for all samples at first, which is related to the null hypothesis. The probability of CNV appearing in the data is proportional to the moving distance of the posterior away from the Dirichlet prior. However, cn.MOPS also has its own weaknesses. First, single sample cannot be detected by cn.MOPS, because cn.MOPS need a cross-comparison between multiple samples to reduce biases and. noises. Second, cn.MOPS has insufficient sensitivity for detecting short CNV fragments.

After reading and comparing other CNV detectors evaluations, I found that the current detection methods mainly have the following common defects: 1) The selection of some key parameters in the detection process of the reading depth method directly affects the detection results, but the selection of these parameters is dependent on the experience of the researcher, such as: the size of the window, matching threshold, etc. Therefore, it is necessary to design a more intelligent method or a mechanism that can adaptively optimize parameter selection based on read information or matching information to improve the usability and operating efficiency of the tool; 2) Sequence-based detection methods first need to compare the short reads data. A small number of base mismatches are allowed during the comparison. When it is greater than the mismatch threshold, the short reads data will be discarded. In fact, these reads are not "junk data". Properly incorporating with Paired-End Mapping method and Split Read method could help RD based methods increase detecting power by offering boundary information for CNV detection, thereby the detection performance can be improved; 3) The correction of deviation is mainly GC correction, and current calculation of GC correction is relatively simple (Benjamini and Speed, 2012). It only considers the GC value of the target window and does not consider the situation of adjacent windows, which leads to some corrections that are too severe and insufficient smooth. These deficiencies require continuous improvement and improvement by software developers in future research to make this technology more mature.

The latest scientific research results show that PacBio RS single-molecule sequencing are able to generate long reads with a mean length of 1300bp (Chao et al., 2019), which is much greater than the reads of any current NGS platform. Longer reads could facilitate the aligning process of reads as well as the detecting results of CNVs in repeated regions of the reference chromosomes. These long reads can dramatically decrease mapping errors due to sequencing errors, moreover the more reads number means more statistical power of the RD method. The development of great efficient and accurate CNV detectors requires continuous updating techniques combined with novel computational algorithms. Although the completion of such tasks needs numerous endeavors from both industry and academia, a profound understanding of human CNV under health and disease will be super attractive.

Limitation

As a result of the COVID-19, I was unable to obtain the real data that should have been collected. Moreover, many software bugs cannot be repaired by developers currently, resulting in a significant reduction in the number of experimental tools that can be used.

Acknowledgements

Here, I would like to thank my supervisor, Fumagalli Matteo, for his help in the process of my project work. Also, I want to thank all the course lecturers as well as workers and classmates in Siwood Park Campus.

Data and code availability

I do not have empirical data. All my data could be simulated from SimulateCNVs tool. My code and results can be viewed or loaded in my github: https://github.com/Grace1016/CMEECourseWork

References

2002. *Genomes. 2Nd Edition*. Wiley Liss.

Abyzov, A., Urban, A., Snyder, M. and Gerstein, M., 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6), pp.974-984.

Alkan, C., Coe, B. and Eichler, E., 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), pp.363-376.

Alkan, C., Kidd, J., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J., Baker, C., Malig, M., Mutlu, O., Sahinalp, S., Gibbs, R. and Eichler, E., 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, 41(10), pp.1061-1067.

AllSeq. 2020. *WGS Vs. WES - Allseq*. [online] Available at: <https://allseq.com/kb/wgsvswes/> [Accessed 26 August 2020].

Audano, P., Sulovari, A., Graves-Lindsay, T., Cantsilieris, S., Sorensen, M., Welch, A., Dougherty, M., Nelson, B., Shah, A., Dutcher, S., Warren, W., Magrini, V., McGrath, S., Li, Y., Wilson, R. and Eichler, E., 2019. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176(3), pp.663-675.e19.

Benjamini, Y. and Speed, T., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10), pp.e72-e72.

Bochukova, E., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., Hurles, M. and Farooqi, I., 2009. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463(7281), pp.666-670.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J., Janoueix-Lerosey, I., Delattre, O. and Barillot, E., 2010. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, 27(2), pp.268-269.

Chao, Y., Yuan, J., Guo, T., Xu, L., Mu, Z. and Han, L., 2019. Analysis of transcripts and splice isoforms in Medicago sativa L. by single-molecule long-read sequencing. *Plant Molecular Biology*, 99(3), pp.219-235.

Chiang, D., Getz, G., Jaffe, D., O'Kelly, M., Zhao, X., Carter, S., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E., 2008. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods*, 6(1), pp.99-103.

Dancey, J., Bedard, P., Onetto, N. and Hudson, T., 2012. The Genetic Basis for Cancer Treatment Decisions. *Cell*, 148(3), pp.409-420.

Duan, J., Zhang, J., Deng, H. and Wang, Y., 2013. Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies. *PLoS ONE*, 8(3), p.e59128.

Fanciulli, M., Norsworthy, P., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J., Gough, S., de Smith, A., Blakemore, A., Froguel, P., Owen, C., Pearce, S., Teixeira, L., Guillevin, L., Graham, D., Pusey, C., Cook, H., Vyse, T. and Aitman, T., 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics*, 39(6), pp.721-723.

Feliubadaló, L., Lopez-Doriga, A., Castellsagué, E., del Valle, J., Menéndez, M., Tornero, E., Montes, E., Cuesta, R., Gómez, C., Campos, O., Pineda, M., González, S., Moreno, V., Brunet, J., Blanco, I., Serra, E., Capellá, G. and Lázaro, C., 2012. Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. *European Journal of Human Genetics*, 21(8), pp.864-870.

Feuk, L., Carson, A. and Scherer, S., 2006. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2), pp.85-97.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. and Conesa, A., 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20), pp.2678-2679.

GitHub. 2020. *Nabavilab/CNV-Sim*. [online] Available at: <https://github.com/NabaviLab/CNV-Sim> [Accessed 26 August 2020].

GitHub. 2020. *Samuelesoraggi/Hmmploidy*. [online] Available at: <https://github.com/SamueleSoraggi/HMMploidy> [Accessed 26 August 2020].

Janevski, A., Varadan, V., Kamalakaran, S., Banerjee, N. and Dimitrova, N., 2012. Effective normalization for copy number variation detection from whole genome sequencing. *BMC Genomics*, 13(Suppl 6), p.S16.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D., Mitterecker, A., Bodenhofer, U. and Hochreiter, S., 2012. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, 40(9), pp.e69-e69.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Li, H., Ruan, J. and Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), pp.1851-1858.

McCarroll, S. and Altshuler, D., 2007. Copy-number variation and association studies of human disease. *Nature Genetics*, 39(S7), pp.S37-S42.

Meienberg, J., Bruggmann, R., Oexle, K. and Matyas, G., 2016. Clinical sequencing: is WGS the better WES?. *Human Genetics*, 135(3), pp.359-362.

Piovesan, A., Pelleri, M., Antonaros, F., Strippoli, P., Caracausi, M. and Vitale, L., 2019. On the length, weight and GC content of the human genome. *BMC Research Notes*, 12(1).

Pirooznia, M., Goes, F. and Zandi, P., 2015. Whole-genome CNV analysis: advances in computational approaches. *Frontiers in Genetics*, 06.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), pp.257-286.

Redon, R., Ishikawa, S., Fitch, K., Feuk, L., Perry, G., Andrews, T., Fiegler, H., Shapero, M., Carson, A., Chen, W., Cho, E., Dallaire, S., Freeman, J., González, J., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J., Marshall, C., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D., Estivill, X., Tyler-Smith, C., Carter, N., Aburatani, H., Lee, C., Jones, K., Scherer, S. and Hurles, M., 2006. Global variation in copy number in the human genome. *Nature*, 444(7118), pp.444-454.

Semeraro, R., Orlandini, V. and Magi, A., 2018. Xome-Blender: A novel cancer genome simulator. *PLOS ONE*, 13(4), p.e0194472.

Stratton, M., Campbell, P. and Futreal, P., 2009. The cancer genome. *Nature*, 458(7239), pp.719-724.

Teo, S., Pawitan, Y., Ku, C., Chia, K. and Salim, A., 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21), pp.2711-2718.

Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J. and Kutalik, Z., 2013. The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. *Frontiers in Genetics*, 4.

Varella-Garcia, M., 2010. Chromosomal and genomic changes in lung cancer. *Cell Adhesion & Migration*, 4(1), pp.100-106.

Xing, Y., Dabney, A., Li, X. and Casola, C., 2018. SimulateCNVs: a novel software application for simulating CNVs in WES and WGS data.

Yang, C., Ramani Duraiswami, DeMenthon, D. and Davis, L., n.d. Mean-shift analysis using quasiNewton methods. Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429).

Yizong Cheng, 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8), pp.790-799.

Zhang, J., Chiodini, R., Badr, A. and Zhang, G., 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3), pp.95-109.

Zhang, L., Bai, W., Yuan, N. and Du, Z., 2019. Correction: Comprehensively benchmarking applications for detecting copy number variation. *PLOS Computational Biology*, 15(9), p.e1007367.

Zhao, L., Liu, H., Yuan, X., Gao, K. and Duan, J., 2020. Comparative study of whole exome sequencing-based copy number variation detection tools. *BMC Bioinformatics*, 21(1).