ICNV-TV: A read-depth based Copy Number Variation detection tool

Sriharsha Vogeti
CCNSB,IIIT-Hyderabad
vogetisri.harsha@research.iiit.ac.in

Prashanthi D
CCNSB,IIIT-Hyderabad
prashanthi.d@research.iiit.ac.in

Nita Parekh
CCNSB, IIIT-Hyderabad
nita@iiit.ac.in

***Abstract*—Copy Number Variation (CNV) is a form of structural variation contributing significantly to variations found in human genomes. A number of studies, so far have shown that CNVs are associated with complex diseases. Micro-array genomic comparative hybridization (arrayCGH)and fluorescence in situ hybridization (FISH) provided were limited by the low resolution offered. The advancement in sequencing technologies leading to the development of Next Generation Sequencing (NGS) techniques solved this problem by offering very high resolution. We develop a read-depth (RD) based pipeline called ICNVTV, having a total variation penalized least squares model for segmentation which was proposed by Duan et al 2013. A comprehensive testing, followed by comparison against other RD-based tools was done on simulated data and real data. We analysed results obtained from applying our method to five different populations belonging to South Asian region. Both the real data and population data were obtained from 1000Genome Project.**

***Keywords*—CNV Detection, CNVTV, Read-Depth Methods, Population Studies, NGS**

I. Introduction

Copy Number Variation (CNV) is a form of structural variation (SV) defined as a DNA segment that is 1kb or larger and is present at a variable copy number when compared to the reference genome [1]. Alkan et al [2] puts a much lower threshold on the size of a SV to be called as CNV at 50bp. It is estimated that about 12% of human genome across populations is subject to copy number variation [1]. CNVs have made important contributions to human evolution, diversity **7** and diseases.Several studies have shown the association of CNVs with complex diseases such as autism [3], schizophrenia [4], Alzheimer disease [5], cancer **6** etc. Initially for the detection of CNVs was carried out using FISH **9** comparative arrayCGH and SNP mircoarrays **10** The main disadvantages with these techniques were hybridization noise, limited coverage for genome and low resolution.The resolution offered by this techniques was around 10-25 kbp when 1 million probes where used **8** The development of Next generation Sequencing techniques has solved this problem. The main advantages of NGS techniques for sequencing are high throughput, low cost, high coverage and resolution **11**

There are five methodologies to detect CNVs from NGS data **11** They are a) Read Depth (RD) or Depth of Coverage (DOC) b) split read c) paired-end method (PEM) d) denovo of assembly of genome and e) combinatorial approach. Depth of coverage is the most widely used method for CNV detection.DOC assumes that the copy of a region in a genome is proportional to the read depth at that position.We developed a DOC based CNV detection method for whole genome data, which uses total variation penalized least square model by Duan et al. 2013 **12** for segmentation process. This method by Duan et al has been showed to detect CNVs of smaller sizes.

We validate our method on simulated data and real data from 1000 Genomes Project **16** and compare its performance with three other DOC method based tools viz., ReadDepth **13** CNVnator **14** and ControlFREEC **15** in terms of sensitivity and specificity. We demonstrate our method's high specificity with a relatively good sensitivity on simulated data. Our method is also successfully able to distinguish between homozygous deletion and gaps in the reference genome, which other methods under consideration fail to do. We then apply our method to five populations viz., *Bengali from Bangladesh* (BEB), *Gujarati Indian from Houston* (GIH), *Punjabi from Lahore* (PJL), *Sri Lankan Tamil from the UK* (STU) and *Indian Telugu from the UK* (ITU) belonging to South Asian super-population also from 1000 Genomes Project. We compare and contrast the results obtained in detail.

II. Methods

A. The Proposed Approach

In this section we describe about our method in detail. A general CNV detection method can be data preprocessing, segmentation and post-processing. Below we brief about each stage and our approach at each stage.

1) Data Preprocessing: Most of the tools use alignment file (BAM/SAM) as the input. The read depth (RD) values can be calculated from the alignment files. These RD values calculated are not necessarily proportional to the underlying copy number. This is because of biases such as GC-content bias and mappability bias and random variations. Such RD values obtained can give raise to false positive calls. Hence, it is necessary to remove such effects to obtain true RD values.

GC-bias Correction: GC-bias is the dependence between read depth and GC-content of region. GC-bias leads to under-representation of DNA fragments with high GC or high AT content. We use an algorithm given by Benjamani and Speed **17** They show that the GC-bias at a particular region is determined the GC composition of the full fragment to which it belongs. They use GC composition of the full fragment and then use it to correct the RD value at that position.

Mappability and Read Quality: Mappability for a region in the reference genome is defined as the probability that a read originating from it is unambiguously mapped back to it. Regions with repetitive elements have a low mappability. Reads generating from such regions are ambiguously mapped and have a low mappability score, leading to higher RD values. We ignore regions/bins with a mappability score (which lies between 0 and 1 including) of 0.5 for CNV detection. We also filter reads with a mapping quality score of zero (q0 reads). We used samtools **18** to filter out low quality reads.

Binning: Binning is the process in which RD values of a small intervals, called as bins are replaced by a value, generally average of RD values in that bin, as a representative of that bin. Binning is done to reduce the effects of random sequencing errors. Bin length (window length) is an optional parameter in our method, with a default value of 100bp. RD values were calculated using bedtools **19** We will describe in detail on how bin length affects the CNV detection in later sections.

2) Segmentation: Segmentation can be described as the process of divided the regions into segments such that bins in a region will have similar RD values. A majority of the CNV detection methods use statistical models such as Circular Binary segmentation, Mean Shift algorithm, Hidden Morkhov Model, etc. Duan J et al., 2011 proposes a total variation penalized least squares model for segmentation **20** They prove that their algorithm works better when compared to many tools in detecting CNVs with smaller window/bin size **12** We implement their algorithm in our method. The main reason for choosing it was its ability to detect smaller CNVs, given that the smaller SVs (between 50bp and 1000bp) were also being treated as CNVs.

3) Postprocessing: This stage involves identifying regions with abnormal copy number and determining their absolute copy numbers. For this the average RD value is taken as normal representing the copy number 2.To do so determination of upper and lower thresholds is done. For generally thresholds as simple as 1.5X and 0.5X of the average RD value are used as upper and lower threshold respectively. We use a slightly less strength viz., 1.45X and 0.55X of the average RD value as upper and lower thresholds respectively. The absolute copy number can be determining using the ratio of RD value of CNV region and average RD value.

III. Results

We have evaluated the proposed method on simulated data and population data from 1000 Genome Project. We use simulated data to compare proposed method with the other three CNV detection method. Our choice of method can be explained by the underlying segmentation process of each tool. ReadDepth uses Circular Binary Segmentation (CBS), while CNVnator uses Mean-shift algorithm and ControlFREEC uses ??. All of these methods do not require a control sample.

A. Simulated Data

A 10 Mbp segment of chromosome 4 of hg19 was taken as the starting sequencing. This very sequence is also used as the reference sequence. Now a total of 12 CNVs of different sizes (1000 bp, 2000 bp, 4000 bp) and copy number (1,3,4) were introduced to generate the sample sequence. Single-end reads were generated from the sample using ART simulator **21** at different depths (10X-60X), each depth having 30 samples. Bowtie2 **22** was used with default parameters for alignment to generate input alignment files in BAM format. Our method along with other three tools were tested on this data.We use a reciprocal overlap of 50% to call a certain prediction as a true positive. Any prediction that doest not satisfy this condition is taken as a false positive. Below are the results in tabulated form for our method.

TABLE I
My caption

Length (in bp)	Copy Number	10x	20x	30x	40x	50x	60x
987	1	2	11	19	26	22	21
	3	2	9	18	15	15	17
	4	30	30	30	30	30	30
2009	1	20	24	25	25	29	30
	3	18	18	20	28	25	26
	4	30	30	30	30	30	30
4024	1	27	29	29	30	30	30
	3	23	27	28	30	30	30
	4	30	30	30	30	30	30
4979	1	27	30	30	30	30	30
	3	21	26	30	30	30	29
	4	30	30	30	30	30	30
FP	–	0	0	0	0	0	0

We tested the others also on this simulated data and calculated the sensitivity and specificity values (Table II). Another important metric of a CNV detection is the prediction of breakpoints of the CNV regions. We calculated breakpoint error for samples at 30X and 60X for all the samples which can be found in table III.

TABLE II
A Simple Example Table

First	Next
1.0	2.0

B. Population Study

Populations of BEB, GIH, PJL, STU, ITU belonging to South Asian region were considered for the population analysis. For each population 10 samples of chromosome 11 were obtained from 1000 Genome Project, resulting in 50 samples for the study. We applied our method to detect CNVs. A bin size of 200 bp was used. Since, the data is of low coverage we considered all regions of the genome irrespective of its mappability. We then determined genes which fall in the reported CNVs. A total of 75 different genes where found to have an abnormal copy number across all the 50 samples.

IV. Discussion

V. Future work

Acknowledgments

References

[1] Redon et al., "Global variation in copy number in the human genome," *Nature*, 2006. DOI: 10.1038/nature05329.

[2] Alkan et al., "Genome structural variation discovery and genotyping," *Nat Rev Genet*, vol. 12, no. 5, pp. 363–376, May 2011. DOI: 10.1038/nrg2958.

[3] Sebat J et al., "Strong association of de novo copy number mutations with autism," *Science*, vol. 361, no. 5823, pp. 445–449, Mar. 2007. DOI: 10.1126/science.1138659.

[4] H. Stefansson et al., "Large recurrent microdeletions associated with schizophrenia," *Nature*, vol. 455, no. 7210, pp. 232–236, Sep. 2008. DOI: 10.1038/nature07229.

[5] F. later, "App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy," *Nat Genet.*, vol. 38, no. 1, pp. 24–26, Jan. 2006. DOI: 10.1038/ng1718.

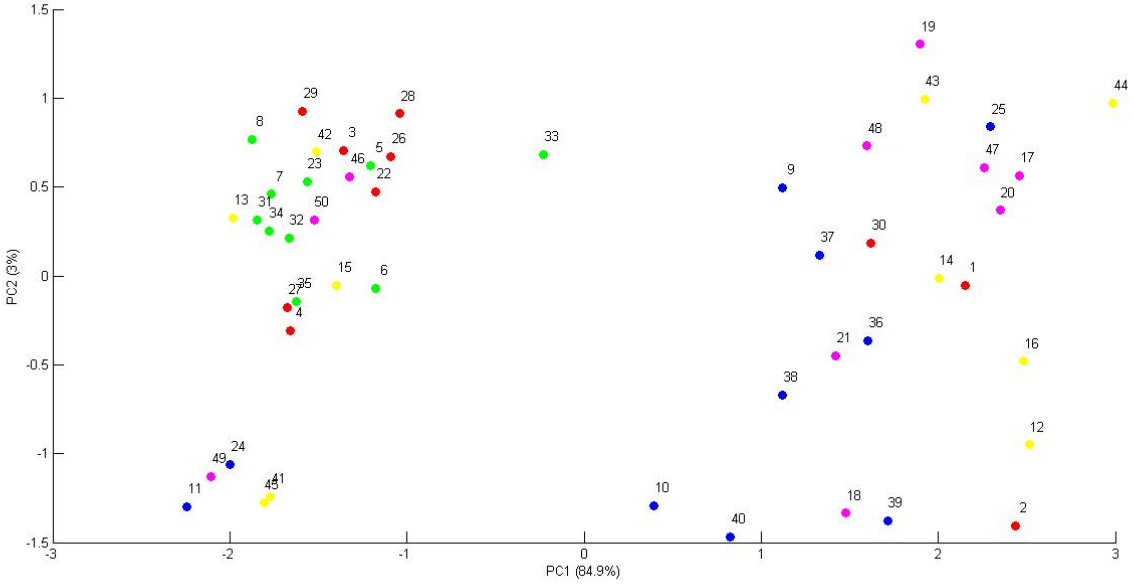

Fig. 1. This is a test image