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.

[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.

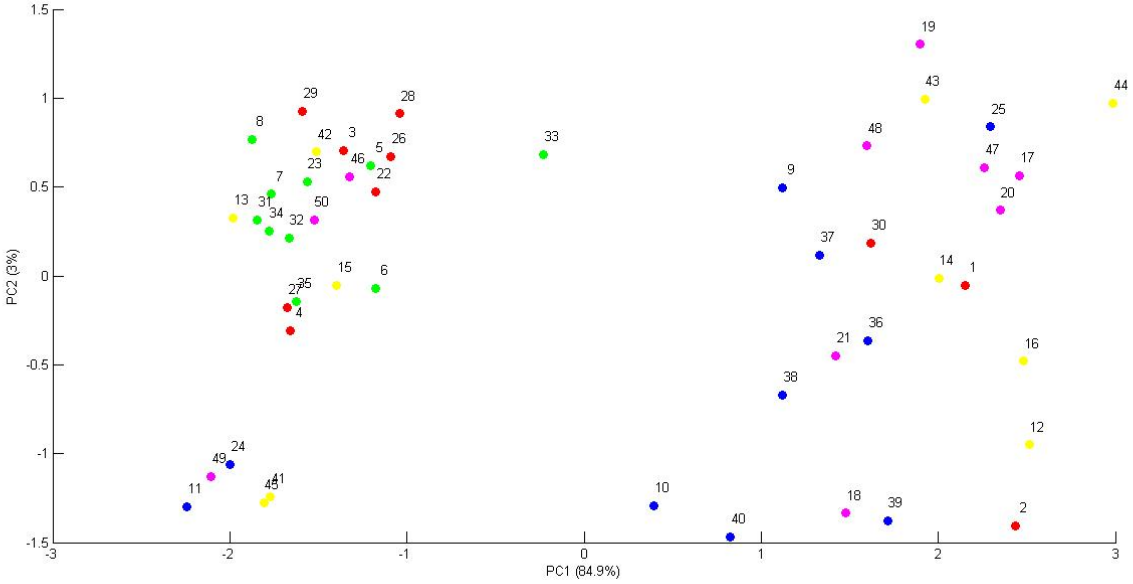

Fig. 1. This is a test image