#

The following explanation is the details for executing the method for identifying CNV’s after addressing the issue of mappability bias occurring in DNA regions affected with CNV.

GenSeg and MR-GenSeg on sample NA12878 on chromosome 22 of human DNA sequences. The same procedure is followed for the other samples and other chromosomes.

Software requirements:

1. Ubuntu

2. Samtools

3. Python

5. Java

4. A Hadoop real distributed cluster, with minimum 1 master and 1 slave. Atleast 16 GB RAM.

Step 1> The desired samples in .bam format is not included here, due to its large size. You can download the .bam files from the following url.

<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/>

Both low coverage and high coverage data for the desired samples are downloaded from the International Genome Sample Resource FTP site.

The standard reference file GRch37/hg19 has been obtained from 1000 genome project dataset.

Step 2>

Samtools is used to process the .bam file and the sorted .bam files are prepared and provided as input to the next step.

All the preprocessing step ,all alignments are present in the BWA-alignment.zip and preprocessing.zip

Step 3> Go to the Segmentation There execute the sortedcsv.py file to generate the base coverage and base quantification. The output of this is then provided to the next step.

Step 4>

Go to the file mrscore.py used for further preprocessing , filling missingvalues with 0 and generating the zscores. Output of this is provided to the next step.

Step 5>

This contains a file SegVar.py file which is the segmentation algorith GenSeg used to generate the segments of the genome.

The data samples have been obtained from phase3 and phase2 of 1000 genome project. Samples are high coverage files of H. Sapiens, the sample id and other details are as follows. (a) HG03007: A female with bengali as ethnicity, and Bangladesh as country of origin. (b) HG03006: A male having bengali ethnicity from Bangladesh. (c) NA12878: A female from UTAH ethnicity and USA as country of oigin. (d) NA12891: A male from UTAH ethnicity having USA as country of origin. The standard reference file GRch37/hg19 has been obtained from 1000 genome project dataset. Sample data sets are downloaded from the International Genome Sample Resource FTP site.