VEGAWES: Variational Segmentation on Whole Exome Sequencing

Samreen Anjum Michele Ceccarelli

Contents

1 Overview

This document describes classes and functions of the VEGAWES (Variational Estimator of Genomic Aberrations on Whole Exome Sequencing data) package. The package consists of a pipeline to process Whole Exome Sequencing data BAM files and compute copy number segmentation based on the variational model, inspired from VEGA [2]. The algorithm implemented in this package is described in "VEGAWES: Variational segmentation on Whole Exome Sequencing " [1].

In this package, we read in paired tumor BAM files as input, and compute average read counts for both normal and tumor BAM files using GATK (Genome Analysis ToolKit). Then, the average read counts are normalized and preprocessed to remove GC bias. The log ratio (LRR) values of the read counts is calculated by taking the ratio of the tumor average read counts and the normal average read counts and passed on as input to perform VEGAWES segmentation.

The package includes pre-computed GC Content values on the exon list and the reference genome, both mentioned in the Installation section.

2 Installation

Install VEGAWES on your computer by using the following command:

```
library(devtools)
install_github("samreenanjum/VEGAWES")
```

In addition, the following tools and data must be downloaded:

- The reference gonome used in the development of this package can be found at

  `ftp://ftp.ncbi.nlm.nih.gov/sra/reports/Assembly/GRCh37-HG19_Broad_variant/Homo_sapiens_assembly19.fasta`

- Genome Analysis Toolkit (GATK) jar file
- Normal and Tumor BAM files

3 Run VEGAWES

In order to use this package, all the items mentioned in the installation must be downloaded and Java is assumed to be in the current path. The exon list used in this setup has been downloaded from UCSC Genome Bioinformatics table for human genome reference hg19 and is included in the package.

In order to run the pipeline, the user needs to execute the `runVEGAWES`. This function requires two arguments: the input file ("parameters.txt ") and the list of chromosomes that need to be analyzed (see Appendix A for more details). The package consists of a "parameters.txt" file that should be modified by the user before running the package.

The parameters.txt file requires the following information:

- Path to the working directory
- Reference genome file path
- Path to the GATK jar
- File containing list of exons to run GATK (format: chr:probe_start-probe_end; The file 'targets.interval_list' is included in the package under /inst/extdata)
- Name of the sample (to create an output folder to save GATK results and the segmentation results, for example: '06-0125')
- Normal input BAM file path
- Tumor input BAM file path
- Path to GC Content Data folder (The data is included in the package under /data as 'GCContent.<chromosome>.RData')

Example of 'parameters.txt':
/home/samreen/VEGAWES
/home/samreen/VEGAWES/Homo_sapiens_assembly19.fasta
/home/samreen/tools/GATK/GenomeAnalysisTK.jar
/home/samreen/VEGAWES/inst/extdata/targets.interval_list
06-0125
/home/samreen/data/TCGA/GBM/TCGA-06-0125/6f138046-a805-489a-a335-6686173d6505/C484.TCGA-06-0125-10A-01D-1490-08.6.bam
/home/samreen/data/TCGA/GBM/TCGA-06-0125/ced14238-6593-4648-b3bb-ca7711f71346/C484.TCGA-06-0125-01A-01D-1490-08.6.bam
/home/samreen/VEGAWES/data/GCC

The first step in the pipleline is to run GATK which takes about 30-40 minutes per file and saves the average read coverage results in the sample output folder. This needs to be run once per BAM file, therefore, `runVEGAWES` also checks for the presence of the GATK results in the sample output folder, and runs GATK only if the files are absent. The second step in the pipeline is to remove GC content bias from the original average read coverage values. The GC content for the exon list using the reference genome has been pre-computed and included in the 'data' folder. In order to compute segmentation on all chromosomes of the input files:

```
> runVEGAWES("parameters.txt", chr.list = c(1:22))
```

The `chr.list` argument can be modified to analyze specific chromosomes. For example, if the user would like to run VEGAWES algorithm on the chromosomes 1 and 7, then the following command can be used:

```
> runVEGAWES("parameters.txt", chr.list = c(1,8))
```

The computed segmentation is saved in a tab delimited file in the output folder and called 'Segmentation.<chromosome>.txt' (example, 'Segmentation.1.txt '). For more detail about this file, please refer to the next subsection (3.1). Other two parameters can be chosen by the user: `alpha` and `beta` for more details on these parameters see Appendix A.

3.1 Output File Format

The segmentation results are stored in a file in the output folder. This file has a row for each segmented region and for each of them it has seven features (columns of the file):

Chromosome: the chromosome containing the region

bp Start: the genomic start position (in bp) of the region

bp End: the genomic end position (in bp) of the region

Num of Markers: the number of markers contained in the region

Mean: the mean value of the LRR of all probes contained in the region

Label: indicates the computed copy number label of the region: loss (1), normal (2) and gain (3).

A VEGAWES: Function Description

A.1 runVEGAWES

The function `runVEGAWES` processes and computes the segmentation on tumor BAM files. The function first initializes the parameters and paths using the input file and then runs GATK to compute the average read coverage. It then processes each chromosome to remove the GC bias content and then performs the VEGAWES segmentations. The header of `runVEGAWES` is as follows:

```
runVEGAWES(inputfile, chr.list, alpha=0.001, beta=0.5)
```

`inputfile:` This argument is the 'parameters.txt'file the contains the required filepaths and samplename. For more details see Section (3) .

`chr.list:` This argument is used to list the chromosomes that have to be processed. By using `c(1:22)` all chromosomes will be segmented.

`alpha:` (default value 0.001) This argument is used to define the weight given to the distance between the exons factor. Setting this to 0 will allow the usage of the original VEGA segmenation (see [1] for more details).

`beta:` (default value 0.5) This argument is used to define the stop condition of VEGAWES algorithm (see [1] for more details).

A.2 initializeParams

This function reads in the parameters file, creates an output folder named after the sample name and returns an object `params` that contains all the required paths and parameters as described in Section (3)

```
initializeParams(inputfile)
```

`inputfile:` This argument is the "parameters.txt" file the contains the required filepaths and samplename. For more details see Section (3).

A.3 runGATK

This function runs the GATK command with the DepthOfCoverage utility to compute the average read coverage of the normal and tumor BAM files and saves it in the output folder named after the sample name. For more details see Section (3).

```
runGATK(params)
```

`params:` This argument is the output object from `initializeParams` containing the path to the GATK jar, the exon list, the reference genome, the sample name as well the input BAM files.

A.4 read.from.gatk

This function reads in the GATK file and creates an object containing the required information.

```
run.from.gatk(gatk.file)
```

`gatk.file:` This argument is the GATK output file saved in the output folder

A.5 `calculateCN`

This function computes the copy number value for the given `logR` values.

```
calculateCN(logR)
```

`logR`: This argument is vector containing all the logR values computed by the segmentation algorithm

A.6 `correct.GCContent`

This function adjusts the average read coverage values based on the GC Content. The approach is based on the median normalization approach described in [3].

```
correct.GCContent(average.coverage,GCContent,step)
```

`average.coverage`: This argument is the original average read coverage values

`GCContent`: This argument is the percentage of GCContent in an exon

`step`: The argument is the interval of GC percentage values chosen to compute the median to adjust each exon

References

[1] Anjum S. *et al.* (2015).VEGAWES: Variational segmentation on Whole Exome Sequencing, *Submitted*.

[2] Morganella S. *et al.* (2010).VEGA: Variational segmentation for copy number detection, *Bioinformatics*.

[3] Yoon S. *et al.* (2009).Sensitive and accurate detection of copy number variants using read depth of coverage, *Genome research 19.9 1586-1592*.