VEGAS2 version 2

By
Dr. A. Mishra[1,2] and **Asso. Prof. S. MacGregor**[1]
[1]QIMR Berghofer Medical Research Intitute, Brisbane, Australia
[2]INSERM Centre U1219, Bordeaux, France
Aniket.mishra@qimrberghofer.edu.au
aniket.mishra@u-bordeaux.fr
Stuart.macgregor@qimrberghofer.edu.au

Content

- Installation
- Usage: Gene-based analysis
- Usage: Pathway-based analysis

Installation- System Requirements

- The VEGAS2 version 2 is developed for Unix, Linux and Mac operating systems
- Make sure plink 1.90 and R (**compiled files not links**) are accessible in your systems $PATH. To check you can type

```
$ which plink; which R
    /usr/local/bin/plink
    /usr/local/bin/R
```

- Then check whether the original or linked files are present

```
$ ls –l /usr/local/bin/plink
$ ls –l /usr/local/bin/R
    lrwxr-xr-x  1 root  wheel  47  4 Jul 21:25 /usr/local/bin/
    R -> /Library/Frameworks/R.framework/Resources/bin/R
```

- If linked file is present as the case for R, include the path of original R executable file in your $PATH variable as follows

```
$ PATH = $PATH:/Library/Frameworks/R.framework/
Resources/bin/R
```

Installation- System Requirements

- Make sure R packages mvtnorm and corpcor are installed in you system
- You can install them from CRAN repository as follows

```
$ R
> install.packages("mvtnorm")
> install.packages("corpcor")
```

Usage: Gene-based Analysis

- For gene-based analysis user needs three input files
  - snpandp file, which is a two column text file with rsIDs and association p-values. Remember this file should not contain header or NAs, -9 etc.
  - plink binary formatted genotype file (.bed, .bim and .fam). Please refer to plink webpage https://www.cog-genomics.org/plink2 for details.
  - gene location file. It should a four column file with Chromosome, Transcription Start, Transcription Stop and GeneID (or Symbol)

Usage: Gene-based Analysis

- To perform gene based test the first parameter should be **–G** followed by **–snpandp** and input text file, then user has to provide a genotype file and a gene location file using **–custom** and **–glist** parameters respectively. Basic command is as follows:

```
$ cd VEGAS2v2example
$ vegas2v2 –G –snpandp example.txt –custom /Users/aniketmishra/Desktop/VEGAS2making/VEGAS2v2example/example –glist example.glist
```

Note: Make sure you provide detailed path of plink binary file

Usage: Gene-based Analysis

- By default, vegas2v2 computes gene-based p-value considering association statistics of all variants within a gene. This version also provides test considering only top association statistics using parameters **–top** and **–topsnp**.
- User can also very flexible gene-boundary to using **–upper** and **–lower** parameters.
- By default vegas2v2 performs 1E6 simulations to compute gene-based p-values, which is sufficient for multiple testing correction of around 25000 tests. But user can use **–max** parameter to compute more accurate p-value by increasing the limit of maximum number of simulations to more than 1E6.

Usage: Gene-based Analysis

- By default, vegas2v2 will perfom test on all genes provided in gene location file using **–glist** parameter. User can also choose to perform test on small subset of genes using parameter **–genelist** as follows

```
$ vegas2v2 –G –snpandp example.txt –custom /Users/
aniketmishra/Desktop/VEGAS2making/VEGAS2v2example/
example –glist example.glist –genelist
example.genelist –out TESTsubset
```

- User can also provide output file name using **–out** parameter.

Top-percent and Best-snp Tests

- Furthermore users can perform top-percentage and topsnp tests using following respective commands

```
$ vegas2v2 –G –snpandp example.txt –
custom /Users/aniketmishra/Desktop/
VEGAS2making/VEGAS2v2example/example –
glist example.glist –top 10 –out Top10TEST
```

```
$ vegas2v2 –G –snpandp example.txt –
custom /Users/aniketmishra/Desktop/
VEGAS2test/VEGAS2v2example/example –glist
example.glist –topsnp –out TopSNPTEST
```

Usage: Pathway-based Analysis

- For pathway-based analysis user needs two input files
  - geneandp file, which is a two column text file with rsIDs and association p-values. Remember this file should not contain header or NAs, -9 etc. After getting gene-basedoutput.out file user can use awk to make geneandp file as follows

```
$ awk '{print $2,$8}' gene–basedoutput.out |
grep –v Gene|sed 's/"//g'> Example.geneandp
```

  - gene pathway annotation file which a text file with first column of gene ids (Symbol) and second column with the names of genesets.

Usage: Pathway-based Analysis

- To perform pathway-based test the first parameter should be **–P** followed by **–geneandp** and geneandp file then user has to provide gene-pathway annotation file using parameter **–geneandpath** as follows:

```
$ vegas2v2 –P –geneandp Example.geneandp –geneandpath Example.vegas2pathSYM –glist example.glist
```

- By default VEGAS2v2 performs maximum 1E6 resamples to compute pathway's association p-value, which is enough to identify associated pathways after correcting for multiple tests performed for around 10000 pathways. Here we also provide **–maxsample** parameter which can be used by users to compute more accurate pathway p-value.

If you do use VEGAS2 software do cite

- VEGAS2Pathway publication:

Mishra, A., Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Cancer Family Registry (CCFR), & MacGregor, S. (2017) A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility, Twin Research and Human Genetics, 20(1), 1-9. Pubmed ID: 28105966

- VEGAS2 publication:

Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet*. 2015 Feb;18(1):86-91. doi: 10.1017/thg.2014.79. Epub 2014 Dec 18. Pubmed ID: 25518859

- VEGAS publication:

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Investigators A, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* 2010, 87:139-145. Pubmed ID: 20598278