VEGAS: *Ve*rsatile *G*ene-based *A*ssociation *S*tudy

v0.7.30

July 30, 2010

Introduction

VEGAS is a free program for performing gene-based tests for association using the results from genetic association studies. It annotates SNPs to corresponding genes, produces a gene-based test statistic, and then uses simulation to calculate an empirical gene-based p-value.

By default, the program uses the HapMap2 CEU (Central Europeans, Utah) population to estimate patterns of linkage disequilibrium for each gene. Other ready-to-use HapMap populations are also available to download. A custom set of individuals can also be used if individual genotype information is available.

VEGAS was developed by Jimmy Liu at the Queensland Institute of Medical Research. For the latest version, see http://gump.qimr.edu.au/VEGAS. Please send any bug-reports, suggestions or feedback to jimmy.liu@qimr.edu.au.

For a detailed description of how VEGAS works, see

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, AMFS Investigators, Hayward NK, Montgomery GW, Visscher PM, Martin NG, MacGregor S. (2010). A Versatile Gene-Based Test for Genome-wide Association Studies. American Journal of Human Genetics, 87. [doi]

Please also cite this paper if you have used VEGAS in your research.

Changes from v0.6.28

1. To enable quick comparison between individual SNP-association and gene-based association results, the VEGAS output file now includes two additional columns indicating the most significant SNP and its corresponding p-value for each gene.
2. There was a bug in the previous version where the top % test did not produce a p-value for genes that required 10^6 simulations. This section of the program has been modified with the hope of correcting this, although some issues still remain (see “Known issues” section).

Installation

VEGAS is written for Linux/Unix and depends on the following to be installed and accessible through `$PATH`:

- Perl (should be included in most Linux/Unix distributions)
- R 2.9.2 or later (with corpcor and mvtnorm packages)
- PLINK

To install VEGAS, extract `vegas-0.7.30-hapmapCEU.tar.gz` by typing:

```
tar -xvzf vegas-0.7.30-hapmapCEU.tar.gz
```

Quickstart:

To run VEGAS using the example files, type:

```
./vegas example.txt –pop hapmapCEU –out example
```

Usage

VEGAS is a command-line program written in Perl and calls R functions and PLINK as required. All commands follow the basic format:

```
./vegas [input file] [parameters]
```

`[input file]` is a two-column, white-space delimited file with of genetic association results. The first column lists each SNP (rs name) and the second column the corresponding p-values. This file should not have a header. See `example.txt` for the correct format.

There is one required parameter and several optional ones:

Required reference-set parameter:

Either:

```
–pop [dir]
```

`[dir]` is the name of the folder containing the reference-population set. For example:

```
–pop hapmapCEU
```

Or:

```
–custom [genotypes]
```

`[genotypes]` is the name of the file with a custom set of individual genotypes. The current version only supports genotypes in binary PLINK PED format. For example:

```
–custom genotypes
```

will look for individual genotype files `genotypes.bed`, `genotypes.bim` and `genotypes.fam`. The PLINK website describes how to convert between different genotype formats. See the 'Using custom individual genotypes' section below for more information.

Optional parameters:

Run VEGAS for genes on a single chromosome:

```
–chr [n]
```

Run VEGAS only on genes specified in a text file:

```
–genelist [file]
```

Specify maximum number of simulations. The default is 1,000,000:

```
–max [n]
```

Specify name of output file. VEGAS will automatically add the `.out` extension:

```
-out [file]
```

Run VEGAS considering a given percentage of the most significant SNPs from each gene:

```
-top [n]
```

Run VEGAS considering only the single most significant SNP from each gene:

```
-topsnp
```

Some examples:

Run VEGAS for all genes using HapMap CEU population as reference:

```
./vegas example.txt -pop hapmapCEU -out default-test
```

Run VEGAS considering the 10% most significant SNPs from each gene:

```
./vegas example.txt -pop hapmapCEU -top 10 -out top10-test
```

Run VEGAS for genes on chromosome 5:

```
./vegas example.txt -pop hapmapCEU -chr 5 -out chr5-test
```

Run VEGAS for genes listed in example_genelist.txt:

```
./vegas example.txt -pop hapmapCEU -genelist
example_genelist.txt -out genelist-test
```

Run VEGAS considering only the single most significant SNP in each gene:

```
./vegas example.txt –pop hapmapCEU –topsnp –out topsnp-test
```

Run VEGAS with the maximum number of simulations set at 10,000,000:

```
./vegas example.txt –pop hapmapCEU –max 10000000 –out max-test
```

All the optional parameters can be used in conjunction with each other (for example, `–topsnp` and `–chr 5`), with the following exceptions:

```
-chr -genelist
```

```
-top -topsnp
```

Once running, VEGAS will create a unique time-stamp folder to store temporary working files. By default, this folder will be deleted once the analysis is finished. If you want to keep this folder (useful for debugging), use the `–keeptimestamp` option.

In terms of time, running VEGAS using genome-wide association results imputed up to 2.3million HapMap SNPs may take ~12-16 hours. Users who have access to a computing cluster environment can split their analysis into separate chromosomes using the `–chr` option to run VEGAS in parallel.

Output file

Once the analysis is finished, VEGAS will store the results in the file specified by the `-out` option. If `-out` is not in effect, the results will be stored in `genebased-output.out`. This is a space delimited text file and can be viewed in any spreadsheet program. See `example.out` for an example output file. The column labels are:

`Chr`: Chromosome

`Gene`: Gene name

`nSNPs`: Number of SNPs in the input file that maps to the gene

`nSims`: Number of simulations performed for this gene

`Start`: Start position of this gene (not including 50kb gene-boundary)

`Stop`: Stop position of this gene (not including 50kb gene-boundary)

`Test statistic`: The sum of the individual chi-squared 1 degree of freedom SNP-association test statistics

`Pvalue`: The gene-based p-value considering the full set of SNPs

`Top-[n]-pvalue`: If the `-top` option is in effect, this column gives the gene-based p-value considering the top n% of SNPs in gene

`Top-SNP-pvalue`: If the `-topsnp` option is in effect, this gives the gene-based p-value considering the single most significant SNP in each gene.

`Best-SNP`: The name of the most significant SNP within the gene.

`SNP-pvalue`: The original association p-value for the best SNP within the gene.

Using custom individual genotypes

If available, individual genotype data can be used instead of a HapMap population to estimate patterns of linkage disequilibrium. However, we do not recommend that users include their entire set of individuals in their analysis as this will slow down VEGAS considerably. Instead, we suggest that users randomly select ~200 unrelated individuals from their study sample to be used as the reference set. A sample size of ~200 individuals is sufficient for estimating linkage disequilibrium for the majority of common SNPs.

Using a custom set of genotypes also allows users to specify gene boundaries. By default, VEGAS use ±50,000bp from the first and last exon of a gene. Boundaries can be specified using the `-lower [n]` and `-upper [n]` options.

For example, if the genotypes of ~200 individuals are in `genotypes.bed`, `genotypes.bim` and `genotypes.fam` and the user wishes to use 30kb boundaries, type:

```
./vegas input_file -custom genotypes -lower 30000 -upper 30000
```

Known issues

PLINK Debian package

If PLINK was installed from a Debian package, the executable may be named `p-link` or `snplink`. VEGAS assumes that it is named `plink`. To correct this (in the case of `p-link`), type:

```
mv vegas vegas-backup

sed 's/plink/p-link/g' vegas-backup > vegas
```

Top % test not producing p-values

Some genes that require 10^6 simulations may not produce a top-% gene-based p-value due to memory allocation problems in R. This mainly occurs for genes with more than ~100 SNPs.