User Manual

Genetic and population analysis

VariantExplorer: rapid annotation of sequence variants for allele frequencies and ENCODE features

Rudi Alberts[1,2*], Marijn C. Visschedijk[1,2], Cleo C. van Diemen[2] and Rinse K. Weersma[1]

[1]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Centre Groningen, Groningen, the Netherlands

[2]Department of Genetics, University of Groningen and University Medical Centre Groningen, Groningen, the Netherlands

Contents

1 QUICK INSTALL GUIDE FOR LINUX/UNIX AND MAC

1.1 Quick install guide for Linux/Unix

Open a terminal. Type `perl -v` to see whether Perl has been installed. If not, type
`sudo yum install perl`

In these instructions, replace yum by your package manager, depending on your distribution. Type:
```
sudo yum install mysql-workbench
sudo perl -MCPAN -e 'install DBD::mysql'
```

Download the Linux install script by typing:
`wget https://raw.github.com/rudi2013/variantexplorer/master/installVariantExplorerLinux.txt`

or download it using your browser. Run this script by typing in the terminal:
`source installVariantExplorerLinux.txt`

This will download, unzip, install and run VariantExplorer in the folder ~/variantexplorer.

1.2 Quick install guide for MAC

Open a terminal and type perl -v to see whether Perl has been installed. If not, go to `http://www.perl.org/get.html`. Click the Download ActivePerl under the Mac OS X section and downloading the Disk Image File, e.g. `ActivePerl-5.16.3.1603-darwin-10.8.0-296746.dmg`. Once downloaded, click on the .dmg file (in `Downloads`) and double click on the ActivePerl package icon to install the program.

Download and install the Mysql Community Server DMG Archive for Mac OS from here:
`http://dev.mysql.com/downloads/mysql/`
There is no need to sign up or login. At the download page, click "No thanks, just start my download". Copy and paste this to the terminal:
```
export PATH=/usr/local/mysql/bin/:$PATH
export DYLD_LIBRARY_PATH=/usr/local/mysql/lib
```

If Perl had already been installed, type:
```
sudo perl -MCPAN -e 'install DBI'
sudo perl -MCPAN -e 'install DBD::mysql'
```

If you installed ActivePerl, type:
```
sudo ppm install DBI
sudo ppm install DBD::mysql
```

Download the Mac install script by typing:
`curl -O https://raw.github.com/rudi2013/variantexplorer/master/installVariantExplorerMac.txt`

or download it using your browser. Run this script by typing in the terminal:
`source installVariantExplorerMac.txt`

This will download, unzip, install VariantExplorer into the folder ~/variantexplorer and run it.

2 DOWNLOAD AND INSTALLATION FOR UNIX/LINUX

2.1 Perl

On most UNIX-like systems Perl has already been installed. To verify this, open a terminal (under Applications – Accessories – Terminal) and copy and paste the following command into it (then press the 'return' key):

```
$ perl -v
```

If Perl has been installed, this will display a message similar to this:

```
This is perl 5, version 12, subversion 4 (v5.12.4) built for darwin-thread-multi-
2level
```

If you get a similar message, Perl has been installed and you can proceed to the next section. If you have an old version, you can update Perl to the newest version. In Linux, you can use a package manager like yum or apt to do this, depending on your distribution. For example, if you have a CentOS system, you can update Perl by typing

```
$ sudo yum update perl
```

If you have an Ubuntu system, you type

```
$ sudo apt-get update perl
```

If Perl cannot be found, you need to install it. You can install Perl using your package manager. For example, if you have a CentOS system, you can install Perl by typing:

```
$ sudo yum install perl
```

If you have an Ubuntu system you can use:

```
$ sudo apt-get install perl
```

Try whether Perl is properly installed by typing:

```
$ perl -v
```

As an alternative, Perl can also be downloaded and installed by hand. For this, go to this website:

```
http://www.perl.org/get.html
```

For Linux, click the Download ActivePerl button under Unix. On the next page, click the appropriate package, e.g. AS package for Linux. This will download a file such as

```
ActivePerl-5.16.3.1603-i686-linux-glibc-2.3.6-296746.tar.gz
```

In the terminal, change the directory to the location where the file has been saved. If this is, for example, /home/john/programs, type

```
$ cd /home/john/programs
```

To see the contents of this folder, type

```
$ ls
```

Extract the downloaded file by typing

```
$ tar xzvf ActivePerl-5.16.3.1603-i686-linux-glibc-2.3.6-296746.tar.gz
```

substituting the filename with the name of the downloaded file. Now install Perl by typing:

```
$ cd ActivePerl-5.16.3.1603-i686-linux-glibc-2.3.6-296746

$ sh install.sh
```

Verify whether the installation has succeeded by typing:

```
$ perl –v
```

2.2 MySQL

To install the Ensembl Perl API, the MySQL workbench need to be installed first. They can be removed after the Ensembl Perl API has been installed. Install it using:

```
$ sudo yum install mysql-workbench
```

Replace yum by your package manager (e.g. apt-get) if needed. Next, install the Perl DBD:mysql module by typing:

```
$ sudo perl -MCPAN -e 'install DBD::mysql'
```

Press `yes` or Enter a couple of times.

2.3 The Ensembl Perl API

Create a folder to install VariantExplorer and the Ensembl Perl API into. Open a terminal and type:

```
$ cd
$ mkdir variantexplorer
$ cd variantexplorer
$ mkdir src
$ cd src
```

Download the following 5 files to this folder:

```
$ wget http://www.ensembl.org/cvsdownloads/ensembl-72.tar.gz
$ wget http://www.ensembl.org/cvsdownloads/ensembl-compara-72.tar.gz
$ wget http://www.ensembl.org/cvsdownloads/ensembl-variation-72.tar.gz
$ wget http://www.ensembl.org/cvsdownloads/ensembl-functgenomics-72.tar.gz
$ wget http://bioperl.org/DIST/old_releases/bioperl-1.2.3.tar.gz
```

Unpack the downloaded files by typing:

```
$ tar xzvf ensemble-72.tar.gz
```

Substitute the name of each file to unpack them all.

You have to tell Perl where to find the modules you just installed. You can do this by using the use lib clause in your script but if you want to make these modules available for all your scripts, the best way is to add them into the PERL5LIB environment variable.

Under `bash`, `ksh`, or any `sh`-derived shell, type:

```
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/bioperl-1.2.3
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-compara/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-variation/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-functgenomics/modules
export PERL5LIB
```

Under `csh` or `tcsh`:

```
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/bioperl-1.2.3
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-compara/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-variation/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-functgenomics/modules
```

Hint: if you copy these 5 or 6 lines to the bottom of your ~/.bashrc file, the environment variable will be automatically loaded each time you login. This will work from the next time you login.

For more information, see the Ensembl Perl API Installation website:

```
http://www.ensembl.org/info/docs/api/api_installation.html
```

2.4 VariantExplorer

In the terminal, type:

```
$ cd
$ cd variantexplorer
```

Download the program into this folder by copying these lines to the terminal:

```
wget https://github.com/rudi2013/variantexplorer/raw/master/config.txt
wget https://github.com/rudi2013/variantexplorer/raw/master/input.vcf
wget https://github.com/rudi2013/variantexplorer/raw/master/manual.pdf
wget https://github.com/rudi2013/variantexplorer/raw/master/output.vcf
wget https://github.com/rudi2013/variantexplorer/raw/master/test.pl
wget https://github.com/rudi2013/variantexplorer/raw/master/test2.pl
wget https://github.com/rudi2013/variantexplorer/raw/master/variantexplorer.pl
```

Go to Chapter 4 for information on how to run the program.

3 DOWNLOAD AND INSTALLATION FOR MAC

3.1 Perl

On most Mac systems Perl has already been installed. To verify this, open a terminal (under Applications – Utilities - Terminal) and copy and paste the following command into it (then press the 'return' key):

```
$ perl -v
```

If Perl has been installed, this will display a message similar to this:

```
This is perl 5, version 12, subversion 4 (v5.12.4) built for darwin-thread-multi-
2level
```

If you get a similar message, Perl has been installed and you can proceed to the next section. If you have an old version or if Perl has not been installed, go to this website:

```
http://www.perl.org/get.html
```

Click the Download ActivePerl under the Mac OS X section and downloading the Disk Image File, e.g. `ActivePerl-5.16.3.1603-darwin-10.8.0-296746.dmg`. There is no need to sign up or login. At the download page, click "No thanks, just start my download". Once downloaded, click on the .dmg file (in `Downloads`) and double click on the ActivePerl package icon to install the program.

To check whether Perl works, open a new terminal and type:

```
$ perl -v
```

3.2 MySQL

Download and install the Mysql Community Server DMG Archive for Mac OS from here:
`http://dev.mysql.com/downloads/mysql/`

Copy and paste these lines into the terminal:

```
export PATH=/usr/local/mysql/bin/:$PATH
export DYLD_LIBRARY_PATH=/usr/local/mysql/lib

sudo perl -MCPAN -e 'install DBI'
sudo perl -MCPAN -e 'install DBD::mysql'
```

3.3 The Ensembl Perl API

Create a folder to install both VariantExplorer and the Ensembl Perl API into. Open a terminal and type:

```
$ cd
$ mkdir variantexplorer
$ cd variantexplorer
$ mkdir src
$ cd src
```

Download the following 5 files to this folder using:

```
curl -O http://www.ensembl.org/cvsdownloads/ensembl-72.tar.gz
curl -O http://www.ensembl.org/cvsdownloads/ensembl-compara-72.tar.gz
curl -O http://www.ensembl.org/cvsdownloads/ensembl-variation-72.tar.gz
curl -O http://www.ensembl.org/cvsdownloads/ensembl-functgenomics-72.tar.gz
curl -O http://bioperl.org/DIST/old_releases/bioperl-1.2.3.tar.gz
```

Unpack the downloaded files by typing:

```
tar xzvf ensemble-72.tar.gz
```

Substitute the name of each file to unpack them all.

You have to tell Perl where to find the modules you just installed. You can do this by using the use lib clause in your script but if you want to make these modules available for all your scripts, the best way is to add them into the PERL5LIB environment variable.

Under `bash`, `ksh`, or any `sh`-derived shell, type:

```
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/bioperl-1.2.3
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-compara/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-variation/modules
PERL5LIB=${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-functgenomics/modules
export PERL5LIB
```

Under `csh` or `tcsh`:

```
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/bioperl-1.2.3
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-compara/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-variation/modules
setenv PERL5LIB ${PERL5LIB}:${HOME}/variantexplorer/src/ensembl-functgenomics/modules
```

For more information, see the Ensembl Perl API Installation website:

```
http://www.ensembl.org/info/docs/api/api_installation.html
```

3.4 VariantExplorer

In the terminal, type:

```
$ cd
$ cd variantexplorer
```

Download the program into this folder by copying these lines to the terminal:

```
curl -O https://github.com/rudi2013/variantexplorer/raw/master/config.txt
curl -O https://github.com/rudi2013/variantexplorer/raw/master/input.vcf
curl -O https://github.com/rudi2013/variantexplorer/raw/master/manual.pdf
curl -O https://github.com/rudi2013/variantexplorer/raw/master/output.vcf
curl -O https://github.com/rudi2013/variantexplorer/raw/master/test.pl
curl -O https://github.com/rudi2013/variantexplorer/raw/master/test2.pl
curl -O https://github.com/rudi2013/variantexplorer/raw/master/variantexplorer.pl
```

Go to Chapter 4 for information on how to run the program.

4 DOWNLOAD AND INSTALLATION FOR WINDOWS

4.1 Perl

On most Windows systems Perl has not been installed. To verify whether Perl has been installed, open an MS-DOS command prompt (under Start – Programs - Accessories) and copy and paste the following command into it (then press the 'return' key):

```
$ perl –v
```

Tip: the command prompt can also be started by clicking Start – Run, entering `cmd` in the box and clicking OK.

If Perl has been installed, this will display a message similar to this:

```
This is perl 5, version 12, subversion 4 (v5.12.4) built for MSWin32-x86-multi-
thread
```

If you get a similar message, Perl has been installed and you can proceed to the next section. If you have an old version or if Perl has not been installed, go to this website:

```
http://www.perl.org/get.html
```

Click the Download ActivePerl button under the Windows section and download the Windows Installer Package such as `ActivePerl-5.16.3.1603-MSWin32-x86-296746.msi`. Once downloaded, right-click on the .msi file and click Install and follow the procedure with default settings.

To check whether Perl works, open a new terminal and type:

```
$ perl -v
```

4.2 MySQL

Install the Perl module DBD::mysql by typing on the command prompt:

```
ppm install DBD::mysql
```

By doing this, the module will be downloaded and installed by Perl.

4.3 The Ensembl Perl API

Create a folder to install the Ensembl Perl API into. Open a command prompt and type:

```
cd \
mkdir src
cd src
```

Download the following 5 files to this folder:
```
http://www.ensembl.org/cvsdownloads/ensembl-72.tar.gz
http://www.ensembl.org/cvsdownloads/ensembl-compara-72.tar.gz
http://www.ensembl.org/cvsdownloads/ensembl-variation-72.tar.gz
```

```
http://www.ensembl.org/cvsdownloads/ensembl-functgenomics-72.tar.gz
http://bioperl.org/DIST/old_releases/bioperl-1.2.3.tar.gz
```

Extract the files by right-clicking them and choosing 'extract here'. You can also use WinZip or WinRar. After extracting, you should see the following five folders in the src folder, after typing `dir` on the command line:

```
<DIR> bioperl-1.2.3
<DIR> ensembl
<DIR> ensembl-compara
<DIR> ensembl-functgenomics
<DIR> ensemble-variation
```

Make sure after extracting that each of the folders contain subfolders like `docs`, `modules` and `scripts`. It might happen that, after extracting, you obtain a folder ensemble-72 that contains one folder named ensembl. This ensembl folder then contains the `docs`, `modules` and `scripts` folders. In this case, you need to move this ensembl folder one level up (from \src\ensemble-72 move it to \src using windows explorer).

You have to tell Perl where to find the modules you just installed. You can do this by using the use lib clause in your script- but if you want to make these modules available for all your scripts, the best way is to add them into the PERL5LIB environment variable. Copy this to the command prompt:

```
set PERL5LIB=C:\src\bioperl-1.2.3;C:\src\ensembl\modules;C:\src\ensembl-
compara\modules;C:\src\ensembl-variation\modules;C:\src\ensembl-
functgenomics\modules
```

For more information, see the Ensembl Perl API Installation website:

```
http://www.ensembl.org/info/docs/api/api_installation.html
```

4.4 VariantExplorer

Create a folder to install VariantExplorer into. Open an MS-DOS prompt and type:

```
cd \
mkdir variantexplorer
cd variantexplorer
```

Download the program files into this folder from here:

```
https://github.com/rudi2013/variantexplorer/raw/master/config.txt
https://github.com/rudi2013/variantexplorer/raw/master/input.vcf
https://github.com/rudi2013/variantexplorer/raw/master/manual.pdf
https://github.com/rudi2013/variantexplorer/raw/master/output.vcf
https://github.com/rudi2013/variantexplorer/raw/master/test.pl
https://github.com/rudi2013/variantexplorer/raw/master/test2.pl
https://github.com/rudi2013/variantexplorer/raw/master/variantexplorer.pl
```

5 RUNNING THE PROGRAM

5.1 Test scripts

To check whether everything has been properly installed, one can run the test script test.pl. This will connect to the Ensembl database and retrieve information about three transcripts. To run it, go to the variantexplorer folder in a terminal and type:

```
$ perl test.pl
```

The output looks like this:

```
ENSG00000167207   NOD2  protein_coding    nucleotide-binding oligomerization domain
containing 2
ENSG00000270120   RP11-327F22.6     sense_intronic
ENSG00000225285   RP4-758J18.10     lincRNA
```

A second test script retrieves genes, transcripts and exons for a specific genomic location. Run it as:

```
$ perl test2.pl
```

The output looks like this:

```
ENSG00000117620: 1:-30933-26257 (+1)
      ENST00000370155: 1:-30933-22731 (+1) biotype: protein_coding
            ENSE00001451960: 1:-30933--30560 (+1)
            ENSE00003569706: 1:-7185--6981 (+1)
            ENSE00003491408: 1:-1461--1307 (+1)
            ENSE00003488106: 1:6312-6434 (+1)
            ENSE00003675499: 1:10643-10811 (+1)
            ENSE00003622900: 1:14580-14698 (+1)
            ENSE00003604622: 1:16960-17093 (+1)
            ENSE00001810895: 1:21664-22731 (+1)
      ENST00000465289: 1:-30933-26170 (+1) biotype: protein_coding
            ENSE00001451960: 1:-30933--30560 (+1)
            ENSE00003569706: 1:-7185--6981 (+1)
            ..........
```

5.2 Input file

The input file for VariantExplorer is a VCF file. This file should at least contain 8 columns, separated by tabs. This is the default for VCF files. The file may contain comment lines, starting with ##. Also it contains one line with column headers, starting with #CHROM. Chromosome names are in the first column. Basepair positions are in the second column. Empty fields contain a dot. Example input file `input.vcf`:

```
##fileformat=VCFv4.0
##comment line
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
1       1138913 .       .       .       .       .       data in col9      data in col10
1       1139202 .       .       .       .       .       data in col9      data in col10
1       1139498 .       .       .       .       .
1       1138913 .       .       .       .       .
```

```
1       1142150 .       .       .       .       .
1       223316685       .       .       .       .       .   tri-allelic SNP
7       98655221        .       .       .       .       .       .
1       1365570 .       .       .       .       .
1       1365925 .       .       .       .       .
1       1366179 .       .       .       .       .
```

5.3 Configuration file

The VariantExplorer folder contains a configuration file named `config.txt`. By editing this file, the user can switch different options on or off and specify which data to include. Each line contains one option, and options are switched on (off) by removing (adding) a # symbol at the beginning of the line. The `config.txt` file looks like this:

```
### General options ###

addgeneinfo
addencodedata

### 1000 Genomes options ###

1000GENOMES:phase_1_ALL
#1000GENOMES:phase_1_EUR
#1000GENOMES:phase_1_AFR
#1000GENOMES:phase_1_ASN
#1000GENOMES:phase_1_AMR
#1000GENOMES:phase_1_CEU
#1000GENOMES:phase_1_ASW
#1000GENOMES:phase_1_MXL
#1000GENOMES:phase_1_CLM
#1000GENOMES:phase_1_GBR
#1000GENOMES:phase_1_FIN
#1000GENOMES:phase_1_IBS
#1000GENOMES:phase_1_YRI
#1000GENOMES:phase_1_CHB
#1000GENOMES:phase_1_JPT
#1000GENOMES:phase_1_LWK
#1000GENOMES:phase_1_TSI
#1000GENOMES:phase_1_PUR

### ESP options ###

#ESP6500:African_American
ESP6500:European_American

### HapMap options ###

CSHL-HAPMAP:HapMap-CEU
#CSHL-HAPMAP:HapMap-HCB
#CSHL-HAPMAP:HapMap-JPT
#CSHL-HAPMAP:HapMap-YRI
#CSHL-HAPMAP:HAPMAP-ASW
#CSHL-HAPMAP:HAPMAP-CHB
#CSHL-HAPMAP:HAPMAP-CHD
```

```
#CSHL-HAPMAP:HAPMAP-GIH
#CSHL-HAPMAP:HAPMAP-LWK
```

With this configuration file, allele frequencies for `1000GENOMES:phase_1_ALL`, `ESP6500:European_American` and `CSHL-HAPMAP:HapMap-CEU` will be collected by the program. If you wish to add for example the `1000GENOMES:phase_1_EUR` data, you just need to remove the # symbol in front of it, save the configuration file and run the VariantExplorer program again. Don't change the name of the `config.txt` file. Similarly, the ENCODE options can be switched on and off under the General options section, by adding or removing the # symbol before 'addencodedata'.

5.4 Running VariantExplorer

Run VariantExplorer by typing:

```
$ perl variantexplorer.pl input.vcf output.vcf
```

This will read the genome positions in the input file, `input.vcf`, and write the results to `output.vcf`. The input and output files can have any name. All integrated data will be added to the 8th column (the INFO column) of the input file. This ensures that the result file is also a VCF file. Hence, it can be used with other tools.

When adding the `-t` switch to the command, the integrated data will be added to the file as separate columns. All data will appear between the 8th and the 9th column of the input file. Column headers are added to the header line starting with #CHROM. Run the program like this:

```
$ perl variantexplorer.pl –t input.vcf output.txt
```

In this example, the output is written to `output.txt`. This file can be easily opened in a spreadsheet program (e.g. Excel). The output looks like this (all in one big table):

Overlap of SNV positions with ENCODE features DNAseI sites, histone modification sites, Polymerase II sites and TFBSs:

#CHROM	POS	DNASE1	HISTONE	POLYMERASE	TFBS
1	1138913	DNase1/DNase1/	H3K27me3/H3K27me3/I		
1	1139202	DNase1/DNase1/	H3K27me3/H3K27me3/I		
1	1139498	DNase1/DNase1	H3K27me3/H3K27me3/I		
1	1138913	DNase1/DNase1/	H3K27me3/H3K27me3/I		
1	1142150	DNase1/DNase1/	H3K27me3/H3K27me3/I		ZEB1/EBF/ZEB1/EBF
1	223316685	DNase1/DNase1/	H3K27me3/H3K27me3/I	PolII/PolII/PolII/PolII	CTCF/CTCF/CTCF/E2F1/E2F
7	98655221				
1	1365570	DNase1/DNase1/	H3K27ac	PolII	CTCF/CTCF
1	1365925	DNase1/DNase1/	H3K27ac	PolII	CTCF/CTCF/CTCF/CTCF/CTC
1	1366179	DNase1/DNase1/	H3K27ac	PolII/PolII	

Cell types for each of the ENCODE features:

DNASE1_CELLTYPES	HISTONE_CELLTYPES	POLYMERASE_CELLTYPES	TFBS_CELLTYPES
DNase1 - HUVEC Enriched Site	H3K27me3 - NHEK Enriched Site/H3K27		
DNase1 - K562 Enriched Site/I	H3K27me3 - NHEK Enriched Site/H3K27		
DNase1 - H1ESC Enriched Site	H3K27me3 - NHEK Enriched Site/H3K27		
DNase1 - HUVEC Enriched Site	H3K27me3 - NHEK Enriched Site/H3K27		
DNase1 - H1ESC Enriched Site	H3K27me3 - H1ESC Enriched Site/H3K2		ZEB1 - GM12878 Enriched Site/EB
DNase1 - NHEK Enriched Site/	H3K4me3 - NHEK Enriched Site/H3K4me	PolII - K562 Enriched Site/PolII - K562	CTCF - NHEK Enriched Site/CTCF -
DNase1 - HUVEC Enriched Site	H3K4me2 - HepG2 Enriched Site/H3K4n	PolII - HepG2 Enriched Site	CTCF - HUVEC Enriched Site/CTCF
DNase1 - HUVEC Enriched Site	H3K4me2 - HepG2 Enriched Site/H3K4n	PolII - HepG2 Enriched Site	CTCF - HUVEC Enriched Site/CTCF
DNase1 - HUVEC Enriched Site	H3K4me2 - HepG2 Enriched Site/H3K4n	PolII - HepG2 Enriched Site/PolII - Hep	

Gene names, IDs and biotype:

GENE_ID	GENE_NAME	BIOTYPE
ENSG00000186891	TNFRSF18	protein_coding
ENSG00000186891	TNFRSF18	protein_coding
ENSG00000186891	TNFRSF18	protein_coding
ENSG00000198742	SMURF1	protein_coding
HUVEC Enriched Site		
ENSG00000225285	RP4-758J18.10	lincRNA
ENSG00000225285	RP4-758J18.10	lincRNA

Reference and alternative alleles and allele frequencies in 1000 genomes data:

1000GENOMES:phase_1_ALL_REF	1000GENOMES:ph	1000GEN	1000GENOMES:	1000GENOMES:phase_1_EUR_REF	1000GENOMES:p	1000GENOM	1000GENOM E
T	0.811355311	C	0.188644689	T	0.94591029	C	0.05408971
T	0.812271062	C	0.187728938	T	0.94591029	C	0.05408971 T
C	0.992216117	T	0.007783883	C	0.992084433	T	0.00791557 C
T	0.811355311	C	0.188644689	T	0.94591029	C	0.05408971
G	0.886904762	A	0.113095238	G	0.947229551	A	0.05277045
C	0.918498168	A	0.081501832	C	0.964379947	A	0.03562005
TTG	0.525641026	-	0.474358974	TTG	0.568601583	-	0.43139842
A	0.399267399	C	0.600732601	A	0.783641161	C	0.21635884
G	0.998168498	A	0.001831502	G	0.994722955	A	0.00527704

Reference and alternative alleles and allele frequencies in ESP and HAPMAP data:

ESP6500:European_American_REF	ESP6500:Eur	ESP6500:	ESP6500:Eur	CSHL-HAPMAP:HapMap-CEU_REF	CSHL-HAPMA	CSHL-HAPM	CSHL-HAPMAP:
				T	0.964602	C	0.0353982
T	0.948394	C	0.051606				
C	0.994289	T	0.00571096				
				T	0.964602	C	0.0353982
				A	0.818584	C	0.181416

6 EXTENDING THE PROGRAM

VariantExplorer is a Perl script that makes use of the Ensembl Perl API. The Perl script runs on a computer that is connected to the internet. Using the API, data is retrieved from Ensembl.

Ensembl has an excellent website with documentation about their Perl API (http://www.ensembl.org/info/docs/api/index.html). The databases are divided into 4 parts:

1. Core
2. Comparative genomics
3. Variation
4. Regulation

By browsing the Database schemes and Tutorials for these 4 sections, one can find out from which part one wants to add data. In the tutorial sections, many examples of working code are given that can be used.

6.1 Example

This example shows how gene IDs, names and biotypes overlapping with the SNVs in the input file have been added to VariantExplorer. Other Ensembl data can be added in a similar way.

Gene IDs, names and biotypes are stored in the Core section. First we need to get the Core adapter:

```
# adaptor for Core
my $sa = $reg->get_adaptor("human", "core", "slice");
```

Now we need to create a slice indication which part of the genome should be queried.

```
# get the slice for this position
my $slice = $sa->fetch_by_region('chromosome', $chromosome, $position, $position);
```

We can use the get_all_Genes() function to retrieve the genes and information about them in this slice:

```
# retrieve gene ID, gene name, biotype from Ensembl
my $genes = $slice->get_all_Genes();
```

The slice can contain multiple genes. We create 3 empty arrays to store the IDs, names and biotypes:

```
my @geneids;
my @geneextnames;
my @genebiotype;
```

Next, we loop through $genes and store each geneID, name and biotype in the corresponding array:

```
while ( my $gene = shift @{$genes} ) {
      push(@geneids, $gene->stable_id());
      push(@geneextnames, $gene->external_name);
      push(@genebiotype, $gene->biotype());
}
```

Finally, we need to add these results to the $result variable. If the -t option is not set, all results are added to $result separated by ";". If the -t option is set, all results are added to $result separated by a tab. This is achieved using this code:

```
# add gene annotation to result
```

```
if ($opt_t) {
        $result.="\t".join("/",@geneids)."\t".join("/",@geneextnames)."\t".join("/",@genebiotype);
} else {
        $result.=";".join("/",@geneids).";".join("/",@geneextnames).";".join("/",@genebiotype);
}
```