User Manual

snp-search:

simple processing, manipulation and searching of SNPs from high-throughput technologies

Version 2.8.1 (September 2013)

Ali Al-Shahib

Anthony Underwood

What is snp-search?

Snp-search is an easy to use tool for management of SNPs generated from haploid next generation sequencing data. Given a vcf file, snp-search stores the SNPs generated by the variant calling algorithm into a sqlite database. snp-search can then be used to extract useful information from the database. For example, by running snp-search using the command line syntax, the user can extract unique SNPs for a specified set of strains; generate a SNP phylogeny and provide detailed information about each individual SNP

Obtaining and installing the code

SNPsearch is written in Ruby and operates in a Unix environment. It is made available as a gem.

To install snp-search, do

```
gem install snp-search
```

Requirements

- **Unix**.
- **ruby** version 1.8.7 and above. Once snp-search is installed, all the necessary gems to run snp-search will also be installed from Rubygems (note that Rubygems requires admin privileges. If you do not have admin privileges then we suggest you install RVM: (beginrescueend.com/rvm/install/) and then gem install snp-search).
- **sqlite** (http://www.sqlite.org/) Many UNIX installations will come with this by default
- Optional: FastTree 2. If you require a tree output in Newick format, you must install FastTree from www.microbesonline.org/fasttree/#Install.

Thats it!

Running snp-search

1- The first thing you need to do is to create the database (snp-search -create)

Two files are needed to create the SQLite3 database:

a. Variant Call Format (.vcf) file (which contains the SNP information)
b. The reference genome that you used to generate your .vcf file (in genbank or embl format, the script will automatically detect the format).

You need the following parameters:

-d
Name of your database (note that this is a required field in all commands).

-v
Variant Call Format (VCF) file. See
http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40

-r
Reference genome (The same file that was used in generating the .vcf file). This should be in genbank or embl format.

Optional: -A AD ratio cutoff (default 0.9)

Usage:

```
snp-search -create -d my_snp_db.sqlite3 -r my_ref.gbk -v
my_vcf_file.vcf
```

Note: The strain names in your database will be taken from your vcf file so make sure they are named appropriately in your vcf file.

2- Now that you have created the database (my_snp_db.sqlite3) you can use snp-search to output several data in several formats.

First, you need to tell snp-search the type of data you want out. You have several options:

1. Querying the Database to select the number of unique SNPs within the list of the strains/samples provided (list_of_my_strains.txt). The output is a text file with a list of the unique SNPs and information about each SNP (e.g. if its synonymous or non-synonymous SNP).

   -u, --unique_snps
   Query for unique snps in the database
   -c, --cuttoff_snp_qual
   SNP quality cutoff, (default = 90)
   -g, --cuttoff_genotype
   Genotype quality cutoff (default = 30)
   -s, --strain
   The strains/samples you like to query (only used with -unique_snps flag)
   -o, --out
   Name of output file, Required

   Usage:

```
snp-search -O -u -d my_snp_db.sqlite3 -s list_of_my_strains.txt -
o unique_snps.out
```

2. Querying the database to output all SNPs except those in specified features in the reference file annotation (e.g. phages). This is a way of ignoring SNPs in genes (likely to be evolving through horizontal gene transfer) and may confound evolutionary analysis. The user has the option of generating a core SNP tree Newick file for the SNP phylogeny (if -F option was used to ouput fasta file).

-f, --all_or_filtered_snps
SNPs from specified features in the database (if you do not want to ignore any SNPs, just use this option with -F/T -o)
-F, --fasta
output fasta file format (default)
-T, --tabular
output tabular file format
-c, --cuttoff_snp_qual
SNP quality cutoff, (default = 90)
-g, --cuttoff_genotype
Genotype quality cutoff (default = 30)
-R, --remove_non_informative_snps
Only output informative SNPs. Only used with -e option
-e, --ignore_snps_in_range
A list of position ranges to ignore e.g 10..500,2000..2500. Only used with -e option
-a, --ignore_strains
A list of strains to ignore (seperate by comma e.g. S1,S4,S8). Only used with -f option
-I, --ignore_snps_on_annotation
The product annotation(s) to ignore. Annotations should be separated by comma (e.g. phages,inserstion,transposons)
-o, --out
Name of output file, Required
-t, --tree
Generate SNP phylogeny (only used with -fasta option)
-p, --fasttree_path
Full path to the FastTree tool (e.g. /usr/local/bin/FastTree. only used with -tree option)

Usage:

```
snp-search -O -F -f -n my_snp_db.sqlite3 -a
phage,insertion,transposon -R -o snps_without_phages.fasta
```

Note: The algorithm FastTree is used to generate the nwk file. FastTree can be downloaded from http://www.microbesonline.org/fasttree/#Install (see above)

3. Output all SNPs with information. (This equivalent of specifying –f –T without any filtering options) Information for each SNP includes whether the SNP is synonymous or non-synonymous, gene function, whether it is a pseudogene and other useful information. These information will be tab-separated.

   -i, --info
   Output various information about SNPs
   -c, --cuttoff_snp_qual
   SNP quality cutoff, (default = 90)
   -g, --cuttoff_genotype
   Genotype quality cutoff (default = 30)
   -o, --out
   Name of output file, Required

   Usage:
   ```
   snp-search -O -info -d my_snp_db.sqlite3 –o
   snps_all_with_info.txt
   ```

View database in Unix or in a GUI

Your database will be in sqlite3 format. If you like to view your table(s) and perform direct queries you can type the following in your command prompt:

```
sqlite3 snp_db.sqlite3
```

Alternatively, you may download a SQL tool to view your database (e.g. SQLite sorcerer).

Example

The following are some examples of commands that can be used while using snp-search. We have a vcf file called ecoli.vcf and genbank file called ecoli.gbk:

Create the database:

```
snp-search -C -r ecoli.gbk -v ecoli.vcf -d ecoli.sqlite3
```

Now we have the database, we can query the database using snp-search. So, to ignore specific features in the database and produce a filtered concatenated fasta file of the SNPs in the DB we run the following command:

```
snp-search -O -f -F -d ecoli.sqlite3 -R -I phage,insertion,transposon
-o ecoli_concatenated_snps_filtered.fasta
```

If we have a certain number of strains that we are interested in and we like to know the number of SNPs shared only between these strains we first prepare a text file (called ecoli_strains.txt) with the strains names (matching those in the vcf file) separated by a new line, e.g.

Ecoli1

Ecoli2

Ecoli3

We then run the following command:

```
snp-search -O -u -d ecoli.sqlite3 -s ecoli_strains.txt -o
ecoli_unique_snps_strains.txt
```

If we require all the SNPs in a tabular format with further information provided (such as whether the SNP is synonymous or non-synonymous and gene information) we run the following command:

```
snp-search -output -info -d ecoli.sqlite3 -o ecoli_snp_info.txt
```

Contact

If you have any comments, questions or suggestions, please email
ali.al-shahib@hpa.org.uk or anthony.underwood@hpa.org.uk

Have fun snp-searching!