

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

© Copyright 2013, Ali Al-Shahib and Anthony Underwood

The software package is provided "as is" without warranty of any kind. In no event shall the author or his employer be held responsible for any damage resulting from the use of this software, including but not limited to the frustration that you may experience in using the package. The program package, including source codes, example data sets, executables, and this documentation, is distributed free of charge for academic use only. Permission is granted to copy and use programs in the package provided no fee is charged for it and provided that this copyright notice is not removed.

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

Not much, you just need:

- **Unix.** 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).
- **ruby** version 1.8.7 and above.
- 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. Your database 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

Database 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 queried data.

First, you need to tell snp-search what you want out. You have several options:

- 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, --cutoff_snp_qual

SNP quality cutoff, (default = 90)

-g, --cutoff_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
```

- Querying the database to output all SNPs without SNPs in specified features in the database (e.g. phages). This is a way of ignoring SNPs in genes (likely to be mobile element genes) that are not needed for SNP analysis. The user has the option of generating a core SNP tree Newick file for SNP phylogeny (if -F option was used to output 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 -n -F/T -o)

-F, --fasta

output fasta file format (default)

-T, --tabular

output tabular file format

-c, --cutoff_snp_qual

SNP quality cutoff, (default = 90)

-g, --cutoff_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 (separate by comma e.g. S1,S4,S8). Only used with -f option

-l, --ignore_snps_on_annotation

The name of the feature(s) to ignore. Features should be separated by comma (e.g. phages,insertion,transposons)

-o, --out

Name of output file. Required

You will get three output files here: 1) Concatenated FASTA file, 2) A SNPs positions file, and 3) a tsv file with tab-delimited SNPs and their positions.

-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). If you decide to use this parameter, you must provide the full path of Fasttree (e.g. /usr/local/Fasttree).

- Output all SNPs with information. 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, --cutoff_snp_qual

SNP quality cutoff, (default = 90)

-g, --cutoff_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 name 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@phe.gov.uk or anthony.underwood@phe.gov.uk

Have fun snp-searching!