NPStat v1: Manual
Population genetics from pooled NGS data

Luca Ferretti

May 31, 2016

The statistical tests implemented in this code can be found in the paper "*Population genomics from pool sequencing*" by Luca Ferretti, Sebastian E. Ramos-Onsins and Miguel Perez-Enciso, published in *Molecular Ecology* (2013), DOI: 10.1111/mec.12522.

I would be glad to help with any problem or receive suggestions for additional functionalities. Please write me at **luca.ferretti@gmail.com**.

1 Requirements

The only requirements are

- a C compiler;
- an installed version of the **Gnu Scientific Libraries (GSL)**. They can be found at `http://www.gnu.org/software/gsl/` and can be installed in all operating system. Under Ubuntu or Debian, you can simply launch the command (from root): `apt-get install gsl-bin libgsl0-dev`.

2 Install

Under Unix, simply type

```
make
```

Under other operative systems you can compile the C code file `NPStat-v1.c` with your compiler (replacing 1 by the version number). Be sure to include

the math library (*math.h*) and the GSL libraries. Rename the executable file as `npstat`.
For example, to compile the code with gcc, call it as

```
gcc -o npstat NPStat-v1.c -lgsl -lgslcblas -lm
```

3 Input format

The main input of the program is in **pileup format**. In the version 0.1.19 of SAMtools, this can be generated from the aligned .bam file with the command

```
samtools mpileup -r SCAFFOLD_NAME INPUT_FILENAME.bam > FILE.pileup
```

where `INPUT_FILENAME.bam` is the input file and `SCAFFOLD_NAME` is the scaffold or chromosome to analyze.

Important! The file should contain data from a **single population** and from a **single chromosome** or scaffold.

Other three types of files could be useful:

- Outgroup sequence in FASTA format. The sequence should be aligned with the reference used to align the .bam file. This FASTA file should contain just the sequence for the chromosome or scaffold analyzed.
- File with a list of positions of filtered SNPs for the chromosome or scaffold analyzed. This could be useful to analyze only the SNPs called by some SNP calling software for pools.
- Annotation file in GFF3 format. This allows to perform the McDonald-Kreitman test. Also this file should contain the annotation just for the chromosome or scaffold analyzed.

4 How to use

From the working directory, use the command

```
./npstat -n sample_size -l window_length [options] FILE.pileup
```

The possible options are:

`-n sample_size` : haploid sample size

`-l window_length` : window length in bases

`-mincov minimum_coverage` : filter on minimum coverage (default 4)

`-maxcov maximum_coverage` : filter on maximum coverage (default 100)

`-minqual minimum_base_quality` : filter on base quality (default 10)

`-nolowfreq m` : filter on minimum allele count > m (default 1)

`-outgroup file.fa` : outgroup file in FASTA

`-snpfile file.snp` : consider SNPs only if present in file.snp

`-annot file.gff3` : annotation file in GFF3 format

An important option is `-nolowfreq` m. This specifies how many alleles of low frequency are discarded. The default option is m=1, which means that alleles appearing in only 1 read will be discarded. Data at high coverage or high error rate would need higher values, e.g. m=2 above read depth 50, m=3 above read depth 100, etc. Use m=0 only if the SNPs have already been called by an external SNP caller and passed to the program through the option `-snpfile`.

5 Output files

The output file has the same name as the input file, with `.stats` in the end. The file is tab-separated. Each row corresponds to the statistics of a single window. The file contains the following columns:

1. window number,
2. number of bases covered by sequences,
3. number of bases covered and with known outgroup allele,
4. average read depth,
5. number of segregating sites S,
6. Watterson estimator of θ,
7. Tajima's Π estimator of heterozygosity,
8. Tajima's D,
9. unnormalized Fay and Wu's H,
10. normalized Fay and Wu's H,

11. variance of the number of segregating sites,
12. variance of the Watterson estimator,
13. divergence per base (from outgroup),
14. nonsynonimous polymorphisms,
15. synonimous polymorphisms,
16. nonsynonimous divergence,
17. synonimous divergence,
18. α (fraction of substitutions fixed by positive selection).

All these statistics are computed after filtering for minimum read depth, qualities and allele count.

The HKA test can be obtained by composing data from S (columns 5), $\mathrm{Var}(S)$ (column 11) and divergence (column 13). The McDonald-Kreitman test can be obtained by composing synonimous/nonsynonimous polymorphism/divergence data from columns 14-17 in a 2×2 contingency table.

Note that we approximate all aminoacids to be 4-fold degenerate (i.e. nonsynonimous and synonimous sites actually correspond to the 1st/2nd base and 3rd base in the codon, respectively).