SNAPE-pooled

emanuele.raineri@gmail.com

1. introduction

SNAPE-pooled computes the probability distribution for the frequency of the minor allele in a certain population, at a certain position in the genome of the population. If you decide to use SNAPE-pooled, you should first read the accompanying paper which describes the formulae used in it.

2. input format

The input data must be formatted according to the pileup specifications [see http://samtools.sourceforge.net/pileup.shtml], *i.e.* the following fields must be present:

1. a chromosome field, part of the genomic coordinate
2. an integer, specifying the position along the chromosome
3. the reference nucleotide, *i.e.* the content of the reference genome for the population at that position of the given chromosome
4. the coverage, that is the number of bases in the pileup
5. the pileup, a list of all the nucleotides aligned with the position specified in (1) and (2). Each nucleotide comes from a different read, each read might (or not) come from a different individual.
6. the quality pileup, that is a quality symbol for each of the nucleotides in (4).

3. command line options

It is also necessary to specify some of the parameters used in the calculations, which can be done through a set of command line options. These are:

`nchr`	Number of different individuals in the pool
`theta`	θ the nucleotide diversity
`D`	Prior genetic difference between reference genome and population
`priortype`	Can be informative or flat
`fold`	folded or unfolded
`noextremes`	excludes $f = 0, 1$ from the computation of $E(f)$
`spectrum`	If present, print the full pdf for the minor allele frequency.

if `-spectrum` is not specified, ony summary values will be printed, see following section.

4. output format

the output contains a minimum of 10 fields, `TAB`-separated, as in the following list:

1. chr (1) and (2) are the genomic coordinates
2. position along the chromosome
3. reference
4. # reference nucleotides
5. # number of minor (alternative) nucleotides
6. average quality of the reference nucleotides
7. average quality of the alternative nucleotides
8. first and second most frequent nucleotides in the pileup
9. $1 - p(0)$ where $p(f)$ is the probability distribution function for the minor allele freqeuncy
10. $p(1)$
11. $E(f)$ mean value of f

In addition, if `-spectrum` is specified on the command line, the full pdf for f is printed after the fields listed above.

5. example

A typical command line:

```
./snape-pooled -nchr 9 -theta 0.1 -D 0.1 -priortype flat -fold folded < input_file.pool
```