ngsAssociation
flexible association mapping using pooled or unpooled next-generation sequencing data

Tyler Linderoth

Contents

1 Introduction

ngsAssociation is a program for association mapping using next-generation sequencing (NGS) data. It accurately estimates population allele frequencies by modeling sources of NGS genotyping error such as sampling chromosomes with replacement, variable coverage, and sequencing error. The implemented methods are based largely on those explained in Kim *et al.* (2010). *ngsAssociation* comprises two subprograms; *association* calculates a likelihood ratio (LR) for a test of allelic association, and *summarize*, which estimates the MAF calculates a LR for a test of whether a site is variable.

2 Installation

ngsAssociation can be downloaded from
`https://github.com/tplinderoth/ngsAssociation`
or directly fetched from github using the terminal command:
`git clone https://github.com/tplinderoth/ngsAssociation`

To install *ngsAssociation* issue the following commands:

```
cd ngsAssociation
make
```

This will create the ngsAssociation executable. *ngsAssociation* has been successfully compiled and tested on Linux and MacOsX operating systems.

3 Input File Format

3.1 Main Pileup Input

ngsAssociation takes a single Pileup format file containing sequencing information for all pools to be analyzed as input. Each line of the Pileup input must be tab-delimited and should specify, in the following order:

(1) sequence identifier
(2) position in sequence
(3) reference nucleotide
(4) coverage at the position
(5) read bases at the position
(6) base quality scores

Fields 4-6 should be repeated for each pool.

Example Pileup input for 4 pools at 3 sites:

```
chr1 1 C 2 ,.  JH 4 ,...  A5<E 4 .TtT F@=D 4 ..,, BB:C
chr1 23 G 1 .  E 4 cC..  BBE> 3 .,, HGH 5 C.c, D>>G
chr1 55 T 6 ,.Aaa, FHH=;H 2 aA 6> 4 .A,A CFFF 2 ,, AB
```

3.2 Treatment Identifier File

To run the *association* routine of *ngsAssociation*, you must supply a text file of treatment identifiers to the `-treatments` option so that the program knows the phenotypes of the different pools in the input Pileup. Each line of the treatments file specifies the treatment identifier of a pool. The order of identifiers in the treatments file corresponds to the order of pools in the input Pileup.

Example `treatments.text` file:

```
control
control
case
case
control
case
case
control
```

The above example corresponds to Pileup input with 8 pools, representing two different phenotypes: pools 1-2 in the Pileup have the 'control' phenotype, 3-4 have the 'case' phenotype, 5 is a control, 6-7 are cases, and 8 is a control. Note that SAMtools outputs individuals/pools in Pileup format in the corresponding order in which input BAM files are supplied or listed in a BAM list.

4 Running the Program

Running *ngsAssociation* without arguments or with `-help` outputs information about its two subprograms *association* and *summarize*:

```
./ngsAssociation -help
```

4.1 association

The *association* routine estimates the minor allele frequency (MAF) for the total population, such that the frequency of this allele is the same in each treatment-specific population. This is the null hypothesis, versus the alternative hypothesis in which the frequency of the total population's minor allele is different in each treatment, and so the frequency of this allele is estimated respective of treatment. The likelihood under the null is compared to the alternative likelihood by means of a LR for a test of allelic association. The LRT statistics are approximately distributed according to a χ^2 with one degree of freedom. The test becomes conservative under this distribution when the MAF is greater than 5%, as p-values become slightly skewed away from small values. When the MAF is smaller than 0.5% and less than the sequencing error rate, the LRs depart from the χ^2_1.

To run the *association* routine use:

```
./ngsAssociation association [options]
```

Issue the above command without options to print help information for *association*.

4.2 summarize

The *summarize* routine calculates a LR for whether a site is variable by comparing the likelihood under the null hypothesis of the population MAF being zero to the alternative hypothesis likelihood for which the MAF is estimated using maximum likelihood. The LRT statistics are approximately distributed as a 50:50 mixture of a χ^2_0 and χ^2_1. Therefore, the *summarize* routine can be used for obtaining ML estimates of the MAF and statically calling SNPs. In addition, *summarize* produces a summary of the sequencing coverage and base quality information for each site.

To run the *summarize* routine use:

```
./ngsAssociation summarize [options]
```

Issue the above command without options to print help information for *summarize*.

4.3 Input Options

The table below explains the command-line options for running *ngsAssociation*. Options marked with $\mathcal{A}$ or $\mathcal{S}$ are specific to the *association* and *summarize* routines, respectively. Default values are enclosed in [ ].

option	*input type*	*description*
`-infile`	file|−	Pileup format file of reads and base quality scores. '`-infile -`' will read from STDIN. See section 3.1.
`-outfile`	file	Name of output file. If not supplied, results are printed to STDOUT.
`-treatments`	file	File of treatment identifiers for pools. See section 3.2. $\mathcal{A}$
`-poolsz`	integer	Haploid sample size of each pool. [2]
`-Qoffset`	float	Minimum possible ASCII decimal value used to encode base quality scores. This is the amount that will be subtracted from the quality score decimal values before they are translated into sequencing error probabilities according to $P(error) = 10^{\frac{-Q}{10}}$. [33]
`-minQ`	float	Minimum base quality score to retain read. [20]
`-minpooln`	integer	Minimum number of covered pools to retain site. [1]
`-mincov`	integer	Minimum number of reads from a pool at the current site for the pool to be considered 'covered '. [1]
`-printIndiv`		If specified, output coverage and base quality information for each pool. $\mathcal{S}$

4.4 Output

4.4.1 association output

The output from running the *association* routine by field is:

(1) sequence identifier
(2) position in sequence (1-based index)
(3) -log likelihood of null hypothesis (MAF is the same among treatments)
(4) -log likelihood of alternative hypothesis (MAF is different between treatments)
(5) likelihood ratio of an allelic association
(6) total population MAF (from the null hypothesis)
(7+) MAF in each treatment (from the alternative hypothesis)

The treatment-specific MAFs are listed in the order in which the treatments uniquely appear in the treatment identifier file.

4.4.2 summarize output

The output from running the *summarize* routine by field is:

(1) sequence identifier
(2) position in sequence (1-base indexed)
(3) reference allele
(4) total site depth;reference allele count;alternate allele count
(5) count of each allele at site: A;C;G;T;indel
(6) ML MAF estimate
(7) LR that site is variable
(8) pool coverage
(9) read bases for pool
(10) base quality scores for pool reads delimited by ;

Fields 8-10 are printed for each pool by specifying the `-printIndiv` option.

5 Examples

All of the files for running examples can be found in the `ngsAssociation/examples` directory. The examples that follow assume that you are working from that directory.

5.1 Association Test

The two input files used for calculating LRs for association testing are `assoc_example.pileup` and `poolid.txt`. `assoc_example.pileup` is the Pileup format sequencing data for 200 case and 200 control pools at a pooling size of 5 diploid individuals. The average coverage per pool and sequencing error rate is 5X and 1%, respectively. The file `assoc_example.param` has information about the true total population and treatment-specific MAFs used to generate `assoc_example.pileup`. The contents of `assoc_example.param` according to field are:

(1) sequence identifier
(2) position in sequence
(3) total population MAF
(4) control MAF
(5) case MAF

`poolid.txt` is the treatment identifier file. It denotes the first 200 pools in the Pileup as controls and the following 200 pools as cases. Note that the labels used are arbitrary, so that 'control' and 'case' could just as easily be 'red' and 'green', '0' and '1', etc.

To calculate LRs for the example association data and output them to a file called `assoc_lr.txt`, issue the following commands:

```
../ngsAssociation association -infile assoc_example.pileup
-outfile assoc_lr.txt -treatments poolid.txt -poolsz 10 -Qoffset 33 -minQ 13
-minpooln 200 -mincov 2
```

Note that `-poolsz` requires the *haploid* size of each pool, which is 10, when 5 diploid individuals are used to construct each pool. For non-pooled designs (i.e. each 'pool' is comprised of 1 individual), `-poolsz` would be set to the organism's ploidy (2 for diploids). `-Qoffset 33` indicates that 33 should be subtracted from the ASCII decimal values of the base quality scores (see section 4.3) before they are interpreted as error probabilities. `-minQ 13` tells the program to throw out read bases if their quality is less than 13, and the combination of `-minpooln 200` and `-mincov 2` causes entire sites to be discarded if fewer than 200 of the 400 total pools are covered by less than 2 sequencing reads.

5.2 SNP Calling and Data Summary

The only input file necessary for running the *summarize* routine is `snp_example.pileup`, which contains the Pileup format sequencing data for 40 pools at a pooling size of 5 diploid individuals and an average sequencing depth of 5X per pool. The average sequencing error rate is 1%. The file `snp_example.param` contains the true MAF for each one of the sites in `snp_example.pileup`.

To output an estimate of the MAF for each site and the LRs for whether the sites are variable to the file `snp_lr.txt`, run the following command:

```
../ngsAssociation summarize -infile snp_example.pileup -outfile snp_lr.txt
-poolsz 10 -Qoffset 33 -minQ 13 -minpooln 20 -mincov 2
```

An explanation of the input is similar to that of the association test example (section 5.1). In addition to outputting the MAF and LR of each site being variable, this routine also produces detailed, site-wide coverage information (section 4.4.2). To additionally output the coverage and base quality information for each pool add `-printIndiv` to the command:

```
../ngsAssociation summarize -infile snp_example.pileup
-outfile snp_lr_indivinfo.txt -poolsz 10 -Qoffset 33 -minQ 13 -minpooln 20
-mincov 2 -printIndiv
```

The outputted file `snp_lr_indivinfo.txt` will now contain the extra information about each pool.

6 Author Details

ngsAssociation was written by Tyler Linderoth.
Contact: tylerp.linderoth@gmail.com

References

[1] Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Perderson, O., Wang, J., and Nielsen, R. (2010). Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet. Epidemiol.*, **34**(5), 479-491.