User's guide to FESTA

Shyam Gopalakrishnan
Zhaohui Qin
Gonçalo Abecasis

Department of Biostatistics
University of Michigan

March 25, 2005

Contents

1 Introduction to FESTA

This documentation explains the use of FESTA. FESTA stands for **F**ragmented **E**xhaustive **S**earch for **TA**gSNPs. It can be used to obtain tagSNPs for any given SNP set provided pairwise LD measurements are available.
Description: The algorithm partitions the set of SNPs into disjoint precincts based on the threshold value of the LD measure. Subsequently, it 'solves' each of these precincts to identify the tagSNPs. The tagSNPs are obtained using a mixture of greedy and comprehensive search techniques. First, the greedy algorithm is used to identify an initial solution. Next, an exhaustive search is carried out to improve this solution. A user specified constant limits the number of combinations evaluated for the exhaustive search calculations.

FESTA provides many options, such as variable computation limit, greedy versus greedy+exhaustive search, double cover etc., which can be used to tailor the algorithm. These options and their purposes will be discussed later in the document.

2 FESTA Package

The FESTA package comes with the executable for the UNIX/Linux environment. It also contains this freely distributed manual and sample input and output files. To start using the package, type

```
gunzip FESTA.tar.gz
```

at the command prompt followed by

```
tar xvf FESTA.tar.
```

These commands will create a directory called FESTA in the current directory. This directory will contain the aforementioned files.

3 FESTA Options

FESTA provides the user with multiple options so that the user has a large amount of control over the execution of the program. The list of available options is given below along with a brief explanation for each.

Usage:

FESTA -d data-file -t threshold [-o output-file] [-p mandatory-file] [-e exclude-file] [–limit comp-limit] [-T text-file] [-a algorithm] [-m map-file] [-f freq-file] [-z] [–double-cover] [–DClimit dc-limit] [–ignore-tags] [-c] [–c5]

This table details the basic options for FESTA:

Option	**Explanation**
-d data-file [1]	This options must be used to tell FESTA which data file to use. The data file must contain SNP names and their pairwise LD values. If a pair of SNPs is not present in the file, it will be assumed that the LD measure between them is 0. A sample input file is included.
-t threshold [1]	This is the threshold for the LD measure. Any pair of SNPs that have a LD measure less than the threshold are treated as not connected. 0.5 and 0.8 are two common choices for the threshold.
–limit comp-limit	This option sets the computation threshold which is used to decide whether the cluster can be searched comprehensively or the solution space needs to be pruned before searching it comprehensively. If the number of combinations, i.e. $\binom{n}{k}$ *or* C_n^k, that have to be checked are greater than comp-limit, then greedy algorithm is used to reduce the size of thee solution space until it is within the comp-limit threshold. If this condition is not specified, a default limit of 1M is assumed.
-o output-file	This option can be used to designate an ouput file. If not specified, output is sent to a file named '*tagSNP.out*' in the current directory. **NOTE: If such a file (or the specified output file) already exists, it will be overwritten.**
-p mandatory-file	The mandatory file is a file which contains the SNPs (marker names) that must be included in the final tagSNP set. A sample mandatory file is present in the FESTA package.

[1] These arguments are mandatory.

-e exclude-file	The exclude file lists the SNPs (marker names) that cannot be included in the final tagSNP set. The format of the exclude file is identical to the mandatory file. Mandatory file supercedes the exclusion file. If a marker is declared mandatory, then its presence in the exclude-file is ignored (and a WARNING message is output).
-c	Chooses 4 solutions for single cover using the 4 criteria described below: 1. Maximize average LD value between tagSNPs and non-tagSNPs. 2. Maximize lowest LD value between tagSNPs and non-tagSNPs. 3. Minimize average LD value between tagSNPs. 4. Maximize average LD value between tagSNPs. Each of the solutions is stored in a file called criteria_{criteria_no}.out. So the single cover corresponding to the first criteria is output to a file called criteria_1.out and so on. Any previously existing file of the same name will be overwritten.

This table enumerates the advanced options for FESTA:

Option	**Explanation**
–cols p1,p2,p3	This option can be used to make the input file format flexible. This option is used to specify the column numbers in the data file that contain the required data. Here, p1 denotes the column number which contains the name of the first marker, p2 denotes the column number that contains the name of the second marker and p3 denotes the column number that contains the LD measure between the two markers. All three must be integers. The default values are p1 = 1, p2 = 2, p3 = 3.
-T text-file	When this option is specified, an additional output file is generated containing all markers and their neighbors. This output file is ordered by precinct. The output format is dicussed in the next section.

-a algorithm	If algorithm type is 1 then the simple greedy algorithm is used. For *any other* argument, greedy+comprehensive search algorithm is used. It is important to note that, irrespective of the algorithm type, the map is always clustered before the search for tagSNPs is conducted. If the algorithm type is 1, implying that if only greedy search is conducted for tagSNPs, extra options requiring the solutions from comprehensive search, such as, '–double-cover', '-D', '-c', '-N', '–ignore-tags' etc. may NOT be given.[2]
-m map-file	This option is used to specify the map-file. The map file contains the physical position of the SNPs given in the data file. A SNP present in the data file must be included in the map-file, otherwise erroneous results may be generated.
-f freq-file	This option is used to give the frequency file, which contains the minor allele frequency (MAF) of the SNPs. Again, if the MAF of the SNP is not present in the frequency file, it will default to a value of 0.
-z	This option calculates the physical size of each precinct and outputs as additional information in the output file..
–double-cover	All the SNPs will be double covered, i.e. there will be at least 2 chosen SNPs that tag each SNP. Even the tagSNPs are double covered. In case, there exist some excluded markers, due to which all the markers cannot be double covered, a warning message will be ouput indicating the same.
–c5	This option is an extension of the '-c' switch. Whereas the '-c' switch gives 4 output files in accordance with the 4 aforementioned criteria, '–c5' provides 5 solutions, 4 that are given by the '-c' switch and an additional solution that *maximizes average minor allele frequency of tagSNPs*. Please note that this option must be used only in conjunction with the '-f' switch.

[2]If any extra parameters are specified, they are simply ignored

4 File Formats

In this section, we will take a look at the file formats for the input and output files. In addition to this section, there are example files in the FESTA package which can be referred to, as a case in point.

4.1 Input Files

There are three kinds of input files that are required. They are the disequilibrium file, the map file and the frequency file. Let us examine their formats one by one.

Disequilibrium File: The disequilibrium file lists all the SNPs and their pairwise LD measure. As mentioned earlier, any pairwise LD measure that is missing will be assumed to be zero. This file can be generated using LDMAX. The file is expected to be in the format given below. There must be at least three (single) tab seperated columns. The first line must be a comment line (it must not contain data to be used). The column numbers specified using the '–cols' switch must contain the names of the SNPs and the pairwise LD measure. If the '–cols' switch is not used, then the first 2 columns are assumed to contain the name of the SNPs and the third column must contain the LD measure. The first three lines of the sample disequilibrium file are included (Here p1 = 1, p2 = 2, p3 = 3, to indicate column numbers if one were to use the '–cols' switch).

```
LABEL1  LABEL2  DELTASQ
rs10199046:119  rs10221549:119  0.77778
rs10199046:119  rs10221616:119  0.77778
```

On the second line *rs10199046:119* and *rs10221549:119* are SNP names and the LD measure is 0.77778. Also observe that the first line is a header and does not contain any information.

Map File: The map file contains the physical postions of all the SNPs that have been enumerated in the data file. To stress the point, if any SNP does not appear in the map file but does appear in the data file, its position is defaulted to 0. This may lead to erroneous results. The map file must contain exactly three (single) tab seperated columns. The second column must contain the SNP name and the third column must contain the physical position of the SNP. As an example, the first two lines of the sample map file are given below.

```
2       rs10199046:119  51.644128
2       rs10202962:119  51.641796
```

It is important to note that, unlike the data file, there is no header line in the map file. The first line need not be a header line. In this case the first column contains the choromosome number, but since it is not used, it can contain anything.

Frequency File: This file contains the Allele frequencies of the SNPs. The format of this file is best explained with an example.

```
M rs10199046:119
A   4 0.87500
A   2 0.12500
```

As with the map file, this file also does not have any header line. The three lines give the allele frequencies of the marker *rs10199046:119*. The line containing the SNP name must begin with an 'M' followed by whitespace. The allele frequencies must follow this line and must begin with an 'A' followed by whitespace. The allele frequency must be the third token (word) on the line. In the above example, the second line is the major allele frequency of the marker. It is 'A' followed by spaces followed by the number followed by spaces followed by the allele frequency. The number of whitespaces is irrelevant, but this must appear on new line. The sample frequency file should make the format sufficiently clear.

Mandatory and Exclude file: The mandatory/exclude file contains the SNPs (markers) that must be include/excluded from the final tagSNP (solution) set. These files contain the names of the markers, one on each line. A sample mandatory file is contained in the FESTA package. A portion of the file is reproduced below.

```
rs6706917:116.1
rs6707563:116.1
rs6734029:116.1
```

4.2 Output Files

There are three types of output files. These are the result output file containing information about the results of the greedy/comprehensive search for tagSNPs in the map, the criteria output files containing the information on the five solutions chosen on the basis of the previously mentioned criteria and the cluster output file which is a textual representation of the clustered SNP map.

Result Output File: This file contains the results of the greedy and/or comprehensive search for the tagSNPs. The output file format is given and explained below.

```
Cl no Cl size Gr set size GrEx set size No.Sols Double Cover size
1     17      1           1             17      1
2     9       1           1             9       1
3     13      1           1             13      1
```

This is sample output file. The first line is the header line. The first column is the precinct number, the second column is the precinct size, the third column is the result given by greedy algorithm, the fourth column is the result given by greedy+comprehensize hybrid search, the fifth column is the number of such

solutions found by the hybrid search. The last column is the size of the double cover (additional SNPs required over the hybrid result). If the user requests that the physical size of the precinct be displayed, then another column will be added to display the same. If double cover is not requested, the appropriate column will not appear. The result file will contain only data that has been requested. There are other sample output files as part of the FESTA package. They also contain the command that was run to obtain them.

Some precincts cannot be 'solved' comprehensively. An example of one such precinct is given below (Not from sample output).

```
Cl no.  Cl size Gr set size     GrEx set size   No.Sols
...
...
18      88      5               5               2
...
...
```

Here, the number of combinations is $\binom{88}{5} = 39175752$, which is greater than the default threshold of 1M. Therefore, greedy algorithm will be used to prune the solution space before comprehensively searching for tagSNPs.

At the bottom of the result file is a summary of results, as given below:

```
Time taken = 0.120000 seconds

Number of markers: 50
No of cluster: 11
Number of clusters where greedy marker removal had to be
performed before exhaustive search: 0
Total tagSNP size of Greedy solution: 11
Total tagSNP size of Greedy Exhaust solution: 11
```

The summary is self-explanatory.

Criteria Output Files: The criteria output files is just a list of SNPs that are a part of the single cover which was chosen using the desired criteria. The first few lines of a sample criteria output file are given.

```
Single cover SNPs selected by criteria 1

rs1206397:100.2
rs1528800:116.2
rs2216132:111.3
```

All five sample criteria output files are included in the package.

Connection Information File: This output file contains the textual representation of the SNP map after it has been clustered. It lists all the SNPs in the precinct and the neighbors of each SNP are listed after the SNP name as well. So a line in this output file looks like this.

```
SNP name :: SNP neighbor 1, SNP neighbor 2 ...
```

A portion of a sample graph output File is given.

```
Precinct number 9
rs2540989:116.2 ::

Precinct number 10
BI112308:0 ::   rs1206413:116.2,
rs1206413:116.2 ::      BI112308:0,

Precinct number 11
rs7349275:116.1 ::      rs7349348:116.2,
rs7349348:116.2 ::      rs7349275:116.1,
```

5 Updates

The updates to this program will be available at www.sph.umich.edu/qin/FESTA. This site also has some online documentation and the referenced paper, along with sample files. We also plan to release the source code once the program is stable.

6 References

Please cite the following manuscript when you publish results obtained from FESTA.
An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. Qin ZS, Gopalakrishnan S, Abecasis G (in preparation).

7 Contact Information

In case of any queries about the program, please feel free to contact Shyam at gopalakr@umich.edu.