

FEstim

Authors: Anne-Louise Leutenegger, INSERM, anne-louise.leutenegger@inserm.fr,
Steven Gazal, INSERM, steven.gazal@inserm.fr

Version: 1.3 – July 2010

Copyright (C) 2003, 2006, 2010, Anne-Louise Leutenegger

If you encounter any problem, please send a bug report to AL Leutenegger at the following e-mail:
anne-louise.leutenegger@inserm.fr

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <<http://www.gnu.org/licenses/>>.

Contents

FEstim.....	1
Contents	2
1 Introduction.....	3
2 Release history	3
3 Running FEstim	3
4 Description of inputs.....	4
4.1 Markerfile	4
4.1.1 FEstim format - original format.....	4
4.1.2 PLINK format - format used by software PLINK	4
4.2 Datafile	4
4.3 Sample type	5
4.4 LODnoInbreeding file (option --samplewise or --familywise)	5
4.5 Output format.....	6
4.6 Supplementary options	6
4.6.1 Fixed (F , A)	6
4.6.2 Standard error	6
4.6.3 Other options	7
5 Output files	7
5.1 FEstim.out.....	7
5.2 inbreedings.out.....	7
5.3 conditional_proba.out, bestIBD.out & locusIBD.out	7
5.3.1 If option --out1 is used, 2 files are produced	7
5.3.2 If default option --out2 is used, only 1 file (locusIBD.out) is produced	8
5.4 lodscore.out.....	8
5.5 family_lodscore.out	9
6 Examples.....	9
6.1 SNPs and families.....	9
6.1.1 Samplewise lodscores	9
6.1.2 Familywise lodscores.....	9
6.2 SNPs and independent individuals	9
6.2.1 Estimating (F, A) for each individual	10
6.2.2 Using fixed (F, A) provided by the user for each individual.....	10
6.3 Microsatellites.....	10

1 Introduction

FEstim is a C program that implements the individual inbreeding coefficient estimation from genome-wide genotype data (Leutenegger et al, AJHG, 2003) and the genomically controlled homozygosity mapping lodscores (Leutenegger et al, AJHG, 2006).

If you use this program please cite the following references:

1. Leutenegger AL, Prum B, Genin E, Verny C, Lemaître A, Clerget-Darpoux F, Thompson EA. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet.* 73:516-23
2. Leutenegger AL, Labalme A, Genin E, Toutain A, Steichen E, Clerget-Darpoux F, Edery P (2006) Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits; Application to the Taybi-Linder syndrome. *Am J Hum Genet.* 79:62-6

2 Release history

Version 1.3:

- A new command line to specify the inputs and the options
- An option to read the PLINK map file format (using both “.map” and “.frq” files)
- New output format, with inference of the hidden IBD statuses (“bestIBD” by Viterbi algorithm)
- Bug fix for FLOD computation in presence of sibship data

Version 1.2:

- The number of individuals, the number of markers and the number of nuclear families present in the input files do not need to be specified on the command line.
- In the datafile, parents do not need to precede their offspring anymore, and isolated affected individuals do not need to precede the nuclear families.

3 Running FEstim

At the prompt, depending on the sample option chosen (see below) for FEstim type:

```
./FEstim --mfest markerfile --ped datafile > FEstim.out
```

or

```
./FEstim --mfest markerfile --ped datafile --samplewise  
LODnoInbreeding_file > FEstim.out
```

Three executables are provided: FEstim (Linux 32-bits), FEstim.cygwin (Cygwin for Windows), FEstim.mac (Mac OS X). Source codes are in src/ directory.

The maximum numbers of individuals and markers that can be analysed in a single run depend on the user's computer.

For instance, on a Pentium IV 3GHz with 1024MB of RAM: if sample option is "nuclear", it is ~500 individuals genotyped for a 5cM microsatellite map (~660 markers).

4 Description of inputs

4.1 Markerfile

- The marker alleles have to be recoded (from 1 to nb of alleles, 0 for missing) as in Linkage, Genhunter, Allegro, etc ... Do not use allele length as allele name (for microsatellites).
- The sum of the allele frequencies at a marker can be slightly different from 1 to allow for rounding errors, but within a default tolerance range of 0.001. The frequencies are then re-scaled to sum up to 1 exactly.
- If an allele is present in one of the individuals in datafile, its frequency has to be different from 0.
- The program handles missing marker genotypes.
- Make sure that there are no co-localised markers (i.e. with the same position in cM)

4.1.1 FEestim format - original format

- **--mfest** Name of the file (along with path if necessary) containing genetic marker map information

File format: each row of the file corresponds to a marker

- 1st column: marker name
- 2nd column: chromosome (1 to 22, no sex chromosome)
- 3rd column: position of the marker on the chromosome (in centiMorgan)
Note that the first marker does not have to be at position 0 cM. If your positions are in Morgan, add the option **--Morgan**.
- 4th column: number of alleles
- 5th, 6th columns; etc... : 1, frequency of allele "1"; etc...

4.1.2 PLINK format - format used by software PLINK

- **--mlink** Name of the marker and map file name (along with path if necessary) without the extension.

The program expects two files with the extensions provided by PLINK (Purcell et al, AJHG: 81, 2007): "markerfile.map" and "markerfile.frq" (without the header line!). In PLINK, make sure to use the **--recode12** options. If your positions are in Morgan, add the option **--Morgan**.

4.2 Datafile

- **--ped** Name of the file containing the individual data in Linkage format

File format: each row corresponds to an individual

- 1st column: family id (alpha-numeric)
- 2nd column: individual id (alpha-numeric)
- 3rd column: father id (alpha-numeric)
- 4th column: mother id (alpha-numeric)
- 5th column: sex (1=male, 2=female)
- 6th column: disease status (0=unknown, 1=unaffected, 2=affected)
- 7th & 8th column; etc...: allele1 & allele2 at 1st marker locus; etc...

Notes:

- Only give the information for individuals and/or parents and/or sibship (=nuclear family structure), no inbreeding loop information should be specified.
- FEStim does not check for Mendelian errors in nuclear family data. This should be done beforehand.
- The inbreeding coefficient will be estimated for all the genotyped individuals present in the file.
- If sample type (see below) is "independent", the father and mother ids of individuals need to be present but are ignored.
- If sample type (see below) is "nuclear":
 - o Isolated individuals should have "0" for both their father and mother ids (i.e. be founder individuals).
 - o Parents of a sibship should have "0" for their own father and mother ids (i.e. be founder individuals).

4.3 Sample type

Individual sample information:

- **--independent** (default option)
All the individuals in the datafile will be treated as independent of each other regardless of the specified family structures
- Otherwise, individuals will be analysed taking into account the nuclear family structures. You will need to provide a lodfile (see next paragraph for details):
 - o **--samplewise** LODnoInbreeding_file
If the user is only interested in computing a sample FLOD and not having a FLOD for each nuclear family
 - o **--familywise** LODnoInbreeding_file
If the user needs an FLOD for each nuclear family (usually in the case where genetic heterogeneity is suspected) in addition to a sample FLOD

4.4 LODnoInbreeding file (option --samplewise or --familywise)

Lodscores obtained by ignoring the consanguinity in the nuclear families:

- If option **--samplewise** : name of the file containing the combined family lodscores without inbreeding, i.e. the sum of the lodscore of each nuclear family computed ignoring the inbreeding.
- If option **--familywise**: name of the file containing the list of the filenames (with the path if they are not in the current directory) for each nuclear family lodscores computed ignoring the inbreeding.

The list of the filenames should have a line per family for all the families (including isolated individuals) following the order in which they appear in the datafile:

- o 1st column: family id (alpha-numeric, limited to 10 characters)
- o 2nd column: the name of the file containing the family lodscore,
 - or, "ISO" (no quotes) for isolated individuals
 - or, "SINGLE_OFFSPRING" (no quotes) for single offspring with parents because the family lodscores are zero in these cases

For all the lodscore files, the format is the same: one marker on each row

- o 1st column: marker name
- o 2nd column: chromosome

- o 3rd column: position (in centiMorgan) of the marker on the chromosome
Note the first 3 columns should be the same as in markerfile
- o 4th column: (family or combined) lodscore ignoring inbreeding at this position

Notes:

- These lodscores can be obtained with standard computer package -Genhunter, Allegro, Merlin, ...- using the same fully penetrant recessive genetic model and disease allele frequency fD (0.00001 by default) as in FEstim.
- Only the lodscore at the marker positions should be provided. A check is done to ensure that this file contains lodscores computed at the same markers as the ones provided in markerfile. This check is done using the marker name, not the marker position. A comment is issued if the positions differ but the program stops only if the marker names differ.
- In case of -infinity lodscore: FEstim translates it to -9999 if it reads -inf, -INF, -Inf (and also the complete word version), the user can otherwise change the values to -9999 beforehand.

4.5 Output format

The format in which the output files are provided (see section 5 for output):

- **--out1**
original FEstim output files: FEstim.out, inbreedings.out, conditional_proba.out, lodscore.out, family_lodscore.out; plus an additional one (starting from version 1.3) bestIBD.out
- **--out2** (default option)
output files are the same, except that conditional_proba.out & bestIBD.out are merged into one file only, named locusIBD.out with a different layout

4.6 Supplementary options

4.6.1 Fixed (F , A)

- **--FnA** Name of the file containing the list of (F,A) values for each individuals

It allows the user to specify its own (F,A) values for each individual, hence not doing any estimation of the parameters.

File format: each row corresponds to an individual *in the same order as in datafile (--ped)*

- 1st column: family id (alpha-numeric)
- 2nd column: individual id (alpha-numeric)
- 3rd column: F value
- 4th column: A value

4.6.2 Standard error

Standard error computation option:

- **--SE**
If the user wants to estimate standard errors for the estimated inbreeding and IBD rate of change
- **--noSE** (default option)
If the user does not want any standard error estimation

4.6.3 Other options

- **--help** gives the list of the options
- **--eps** genotype error rate (default = 0.001)
- **--tolfq** tolerance for marker allele frequency (default = 0.001)
- **--msmark** maximum size of the marker name (default = 20)
- **--msid** maximum size of the individual name (default = 20)
- **--msfam** maximum size of the family name (default = 20)
- **--mnfsib** maximum number of family with a sibship (default = 10)
- **--fd** susceptibility allele frequency (default = 0.00001)
- **--tolg** tolerance of Gemini algorithm (default = 0.00001)
- **--iterMC** number of MC iterations for Information Matrix estimation (default = 8000)
- **-f** starting value for estimation of inbreeding parameter (default = 0.05). Do not use boundary values 0 or 1 as a starting value.
- **-a** starting value for estimation of the IBD rate of change parameter (default = 0.05). Do not use the boundary value 0 or large values (e.g. >1) as a starting value.

5 Output files

5.1 *FEstim.out*

Raw output of the parameter estimation with the value of the parameters and the associated loglikelihood for each Gemini iteration.

Some details about the information matrix estimation.

Note: Error messages are always output to the screen

5.2 *inbreedings.out*

List of the individuals (famid_id) with their inbreeding (F), IBD rate of change (A) and associated standard errors (StdE).

If the sample option is **--samplewise** or **--familywise**, the median F and A for the nuclear family will also be printed.

Note that even if the parents are typed and present in the datafile, the median is only taken over the sibship.

For individuals not genotyped, no estimation is done, "NA" is printed out.

5.3 *conditional_proba.out, bestIBD.out & locusIBD.out*

5.3.1 If option **--out1** is used, 2 files are produced

5.3.1.1 *conditional_proba.out*

List of the IBD probabilities conditional on the observed genotypes (Y) **for each individual and each marker** of the genome scan.

- 1st column: chromosome
- 2nd column: marker position on chromosome
- 3rd column: $P(\text{IBD}=0|Y)$

- 4th column: $P(\text{IBD}=1|Y)$
- 5th column: $P(\text{IBD}=0|Y)+P(\text{IBD}=1|Y)$, should always be 1
- 6th column: genotype at the marker k (Y_k)
- 7th column: multipoint lodscore (FLOD) for this individual using his (F, A) assuming a rare recessive trait model

5.3.1.2 bestIBD.out

List of the inferred IBD statuses (using Bellman-Viterbi algorithm) based on the observed genotypes (Y) **for each individual and each marker** of the genome scan.

- 1st column: chromosome
- 2nd column: marker position on chromosome
- 3rd column: genotype at the marker k (Y_k)
- 4th column: inferred IBD status (0 = non-IBD, 1 = IBD) at position k (BestIBD)

5.3.2 If default option --out2 is used, only 1 file (locusIBD.out) is produced

List of the IBD probabilities conditional on the observed genotypes (Y), FLOD and inferred IBD status (using Bellman-Viterbi algorithm) **for each individual and each marker** of the genome scan.

- 1st column: famid_id
- 2nd column: marker name
- 3rd column: chromosome
- 4th column: marker position on chromosome
- 5th column: genotype at the marker k (Y_k)
- 6th column: $P(\text{IBD}=1|Y)$
- 7th column: multipoint lodscore (FLOD) for this individual using his (F, A) assuming a rare recessive trait model
- 8th column: inferred IBD status (0 = non-IBD, 1 = IBD) at position k (BestIBD)

5.4 lodscore.out

Multipoint lodscore (FLOD) along the genome for the sample of individuals provided in the datafile taking into account the FEstim estimates.

- If the option is --independent (by default): this lodscore is computed as the sum of each individual FLOD. Each individual FLOD is itself computed using the individual's F and A estimates.
- If the sample option is --samplewise or --familywise: for each family, this lodscore is computed using the median sibship F and A on the first affected individual of the sibship. The overall sample FLOD is computed as the sum of all these lodscores and the value provided by the user in the LODscoreNuclear file.

Note: The lodscores are computed assuming a fully penetrant recessive model with disease allele frequency (f_D) of 0.00001.

5.5 *family_lodscore.out*

Multipoint lodscore (FLOD) along the genome for each isolated individual and each nuclear family separately.

For a nuclear family, the family FLOD is calculated by adding the lodscore ignoring inbreeding provided by the user in the List_familyLODscoreNuclear file with the lodscore computed using the median sibship F and A on the first affected individual of the sibship.

For an isolated individual, the individual FLOD is

$$\left[\frac{P(\text{IBD}=1|Y) + fD \cdot P(\text{IBD}=0|Y)}{F + fD \cdot (1-F)} \right]$$

Notes: This output is disabled if lodfile option --samplewise is used.

And if default option --independent is used, it is the same FLOD as in conditional_proba.out

6 Examples

The example files are in the Example/ directory.

6.1 *SNPs and families*

Example with a map of SNPs and 10 individuals (2 isolated patients, 1 patient with parents, 1 sibship with parents).

6.1.1 Samplewise lodscores

- command line if the sample lodscore is provided ("samplewise"):

```
./FEstim --mfest map_snp.in --ped data_snp_mixnuclear.in --samplewise  
sample_mixnuclear_lodnoinbreed.in --out1> FEstim_mixnuclearS.out
```

- outputs were renamed:
inbreedings_mixnuclearS.out, conditional_proba_mixnuclearS.out,
lodscore_mixnuclearS.out, bestIBD_mixnuclearS.out

6.1.2 Familywise lodscores

- command line if the family lodscores are provided ("familywise"):

```
./FEstim --mfest map_snp.in --ped data_snp_mixnuclear.in --familywise  
list_fam_mixnuclear_lodnoinbreed.in --out1 > FEstim_mixnuclearF.out
```

- outputs were renamed:
inbreedings_mixnuclearF.out, conditional_proba_mixnuclearF.out,
family_lodscore_mixnuclearF.out, lodscore_mixnuclearF.out, bestIBD_mixnuclearF.out

6.2 *SNPs and independent individuals*

Example with a map of SNPs and 5 individuals:

6.2.1 Estimating (F, A) for each individual

- command line:

```
./FEstim --mfest map_snp.in --ped data_snp.in --SE > FEstim_snp.out  
OR  
./FEstim --mplink plink_snp --ped data_snp.in --SE > FEstim_snp.out
```

- outputs were renamed:
inbreedings_snp.out, locusIBD_snp.out, family_lodscore_snp.out, lodscore_snp.out

6.2.2 Using fixed (F, A) provided by the user for each individual

- command line:

```
./FEstim --mfest map_snp.in --ped data_snp.in --FnA list_fixedFnA.in --  
out1 > FEstimNoEstim_snp.out
```

- outputs were renamed:
inbreedings_NoEstim_snp.out, conditional_proba_NoEstim_snp.out,
family_lodscore_NoEstim_snp.out, lodscore_NoEstim_snp.out, bestIBD_NoEstim_snp.out

6.3 *Microsatellites*

Example with a map of microsatellites and 1 individual:

- command line:

```
./FEstim --mfest map_msat.in --ped data_msat.in --SE --out1 >  
FEstim_msat.out
```

- outputs were renamed:
inbreedings_msat.out, conditional_proba_msat.out, family_lodscore_msat.out,
lodscore_msat.out, bestIBD_msat.out