FSuite: exploiting inbreeding in dense SNP chip and exome data

Steven Gazal, Mourad Sahbatou, Marie-Claude Babron, Emmanuelle Génin, Anne-Louise Leutenegger

FSuite: exploiting inbreeding in dense SNP chip and exome data

Steven Gazal[1,2*], Mourad Sahbatou[3], Marie-Claude Babron[1,4], Emmanuelle Génin[5,6§], Anne-Louise Leutenegger[1,4§]

[1]Inserm, U946, Genetic variability and human diseases, Paris, France, [2]Univ Paris Sud, Kremlin-Bicêtre, France, [3]Fondation Jean Dausset CEPH, Paris, France, [4]Univ Paris-Diderot, Institut Universitaire d'Hématologie, UMR 946, Paris, France, [5]Inserm, U1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest, France and [6]Centre Hospitalier Régional Universitaire de Brest, France. [§]These authors contributed equally to this work.

ABSTRACT

Summary: FSuite is a user friendly pipeline developed for exploiting inbreeding information derived from human genomic data. It can make use of SNP chip or exome data. Compared to other software, the advantage of FSuite is to provide a complete suite of scripts to describe and use the inbreeding information. It includes a module to detect inbred individuals and estimate their inbreeding coefficient, a module to describe the proportion of different mating types in the population and the individual probability to be offspring of different mating types that can be useful for population genetic studies. It also allows the identification of shared regions of homozygosity between affected individuals (homozygosity mapping) that can be used to identify rare recessive mutations involved in monogenic or multifactorial diseases.

Availability: FSuite is developed in Perl and uses R functions to generate graphical outputs. This pipeline is freely available under GNU GPL license at:
http://genestat.cephb.fr/software/index.php/FSuite.

Contact: fsuite.software@gmail.com

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 INTRODUCTION

Inbreeding is a central concept in genetics. In population studies, the inbreeding coefficient *f* of individuals is evaluated to characterize mating habits. In rare disease studies, homozygosity mapping (Lander and Botstein, 1987) has been widely and successfully used to localize variants with strong recessive effect by searching for regions homozygous by descent (HBD) on pedigrees with inbred cases.

With the availability of dense genome-wide genetic data, it is now possible to study inbreeding for individuals without genealogical information. Several software packages are available to estimate inbreeding coefficients using genetic data. Some of these provide single-point estimates of the inbreeding coefficient such as PLINK (Purcell, et al., 2007) and GCTA (Yang, et al., 2011). Other programs allow the detection of HBD segments in individuals such as PLINK runs of homozygosity (ROHs), BEAGLE (Browning and Browning, 2010) and IBDLD (Han and Abney, 2013). None of these applications however provide an integrative solution to exploit inbreeding information.

FSuite is a user friendly pipeline composed of several functions integrating FEstim software (Leutenegger, et al., 2003). It estimates *f* and proposes population genetic statistics that are not available in other software, such as detecting inbred individuals, and inferring parental mating types. A homozygosity mapping statistic with heterogeneity, HFLOD, is proposed, that gives more importance to HBD segments on individuals with small *f* (Leutenegger, et al., 2006). With these features, it is easy to perform the HBD-GWAS strategy (Genin, et al.,

*To whom correspondence should be addressed.

2012), i.e. homozygosity mapping on inbred cases from genome-wide association study (GWAS), in order to detect Mendelian sub-entities of complex diseases. FSuite also provides graphical outputs to facilitate interpretations of homozygosity mapping results.

2 METHODS

FSuite needs files in PLINK format: a "map" and a "ped" file, or PLINK binary files. They should contain genome-wide data for the 22 autosomes. If the sample is large enough, FSuite can estimate allele frequencies. Otherwise, frequencies estimated on a reference sample should be furnished in a PLINK "frq" format.

2.1 Creation of random submaps

FEstim uses a hidden Markov model (HMM) to model the HBD process of an individual. It requires the markers to be in minimal linkage disequilibrium (LD), which is not the case of actual dense genetic data. Running FEstim on multiple random sparse genome maps (submaps) has been proposed to remove LD (Leutenegger, et al., 2011). FSuite allows the creation of such submaps, based on genetic or physical positions. The default option is to create 100 random submaps from the map file, with one marker every 0.5 cM. An alternative option is to randomly select a marker between recombination hotspots (McVean, et al., 2004; Winckler, et al., 2005).

2.2 Population genetic studies

The FEstim model depends on 2 parameters f and a, where f is the individual inbreeding coefficient and $1/(a(1-f))$ is the expected length of HBD segments (here cM), that are estimated by maximum likelihood. When several submaps are considered, FSuite estimates the inbreeding coefficient as the median value of the f estimates obtained on the different maps. FSuite also fixes FEstim HMM parameters to compute the likelihood of different mating types. These likelihoods can be used for: 1) inferring an individual as inbred by comparing the maximized likelihood with the one to be outbred with a likelihood ratio test, 2) estimating the proportion of mating types in a population, and 3) estimating the probability for an individual to be offspring of different mating types (Leutenegger, et al., 2011). The ones considered in FSuite are first cousin, second cousin, double first cousin, avuncular and unrelated. FSuite outputs the median p-values/probabilities obtained on the multiple submaps.

2.3 Homozygosity mapping and HBD-GWAS strategy

HBD posterior probabilities and homozygosity mapping score (FLOD) are calculated per individual at each marker. If a marker is present in several submaps, then an average of the HBD posterior probabilities and FLOD are calculated. Note that FSuite accepts single individuals and nuclear families. Finally, to allow for heterogeneity at a locus, a heterogeneity FLOD (HFLOD) is maximized over a grid of α values (the proportion of cases linked to this locus).

By default, FSuite performs these steps on cases previously inferred as inbred, to remove non-informative outbred individuals for homozygosity mapping. Applied on GWAS data, this allows performing directly HBD-GWAS strategy (Genin, et al., 2012).

2.4 Graphical outputs

For a quicker and easier understanding of the results, FSuite produces some graphical outputs generated with R or Circos (Krzywinski, et al., 2009), in addition to text output files (see Figure 1 for different examples). Note that graphs similar to the ones of Figures 1-B and 1-C can also be plotted for runs of homozygosity detected by PLINK.

2.5 Application to whole exome sequencing data

Using submaps has shown good results on SNP data. Whole exome sequencing data are now more and more available, but do not uniformly cover the genome. However, we observed through simulation studies, that FSuite also provides accurate f estimates when applied to SNP genotypes extracted from exome data (see supplementary information).

2.6 Requirement and documentation

FSuite needs Perl, R and some of its packages, Circos (optional), PLINK and Merlin (Abecasis, et al., 2002) to be installed on your computer. FSuite pipeline includes a detailed documentation and a simulated example dataset. It also includes an example dataset of 5 cases affected with a rare disease genotyped on 180,160 SNPs, among whom 4 are inbred.

3 DISCUSSION

FSuite is a user friendly pipeline developed for population genetic studies, rare disease studies and multifactorial disease studies. It implements a set of statistics that are not available in other existing genetic software. It can be used on SNP data but also on whole exome sequencing data. FSuite does not model LD, but takes it into account by using several sparse submaps. It provides robust results, and can be applied on small datasets (where modeling LD is not accurate) if reference allele frequencies are available. In addition, on larger datasets, it has been shown that this strategy gives less bias than available single-points estimates, ROH based estimates, and HMMs modeling LD (Gazal, et al., in press). FSuite offers the possibility to exploit inbreeding information derived from genomic data and will help investigators to better explore their SNP-chip and exome data.

ACKNOWLEDGEMENTS

SG is funded by the plateforme de génomique constitutionnelle (Faculté de médecine, Univ Paris-Diderot, Paris, France).

REFERENCES

Abecasis, G.R., *et al*. (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees, *Nat Genet*, **30**, 97-101.

Browning, S.R. and Browning, B.L. (2010) High-resolution detection of identity by descent in unrelated individuals, Am J Hum Genet, **86**, 526-539.

Gazal, S., *et al.* (in press) Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III, *Hum Hered*.

Genin, E., *et al.* (2012) Could Inbred Cases Identified in GWAS Data Succeed in Detecting Rare Recessive Variants Where Affected Sib-Pairs Have Failed?, *Hum Hered*, **74**, 142-152.
Han, L. and Abney, M. (2013) Using identity by descent estimation with dense genotype data to detect positive selection, *Eur J Hum Genet*, **21**, 205-211.
Krzywinski, M., et al. (2009) Circos: an information aesthetic for comparative genomics, *Genome Res*, **19**, 1639-1645.
Lander, E.S. and Botstein, D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children, *Science*, **236**, 1567-1570.
Leutenegger, A.L., *et al.* (2006) Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome, *Am J Hum Genet*, **79**, 62-66.
Leutenegger, A.L., *et al.* (2003) Estimation of the inbreeding coefficient through use of genomic data, *Am J Hum Genet*, **73**, 516-523.
Leutenegger, A.L., *et al.* (2011) Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us?, *Eur J Hum Genet*, **19**, 583-587.
McVean, G.A., *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome, *Science*, **304**, 581-584.
Purcell, S., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am J Hum Genet*, **81**, 559-575.
Winckler, W., *et al.* (2005) Comparison of fine-scale recombination rates in humans and chimpanzees, *Science*, **308**, 107-111.
Yang, J., *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis, *Am J Hum Genet*, **88**, 76-82.

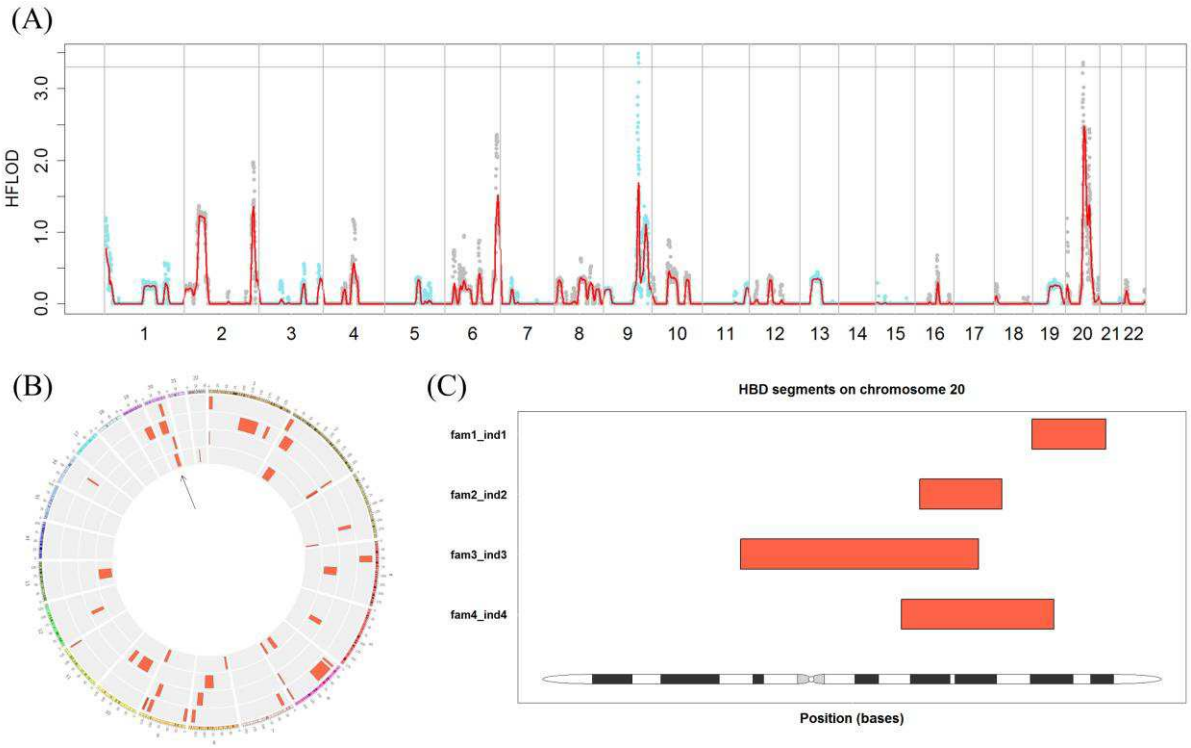

Figure 1. FSuite graphical outputs. These graphs were generated with FSuite example dataset. Graph (A) shows the genome-wide HFLOD plot. Dots are HFLOD at each marker. The solid line is a moving average, to remove the impact of a submap with a false positive signal. In graph (B), the HBD segments are represented genome-wide as a Circos plot. The arrow pinpoints chromosome 20, where the maximum HFLOD is reached. Graph (C) shows the HBD segments for each case on this chromosome.

Supplementary information: FSuite accuracy on whole exome sequencing data

Accuracy of FSuite on whole exome sequencing (WES) data was checked by simulating 100 offspring of first-cousin using European (EUR) 1000 Genomes (1000G) haplotypes.

1. Simulations

Individual genomes of 1,000 first-cousin offspring were simulated by gene-dropping with Genedrop program of MORGAN2.9 (www.stat.washington.edu/thompson/Genepi/MORGAN).

To have realistic genome patterns, we downloaded 758 available EUR 1000G autosomal haplotypes phased with SHAPEIT version 2 (Delaneau, et al., 2013) as reference haplotypes. They were randomly drawn without replacement for each chromosome and were assigned to founders to construct the genotype data of each individual of the family.

We kept 1000G variants coming from two map designs. First, we kept the variants present in the Affymetrix 250K chip. Second, to simulate a WES dataset, we kept variants that are referenced in the exome variant server database (EVS), and that have a minor allele frequency (MAF) $\geq$ 5% in EUR of 1000G to keep frequent polymorphisms. This cutoff is motivated by the fact that variants with small MAF are numerous but only informative for very few individuals. So, if these variants are selected in a submap, they will bring many frequent uninformative homozygous genotypes. In addition, the frequency of rare alleles is difficult to estimate. A total of 316,963 variants from 1000G were thus kept for this simulation study: 248,290 present in Affymetrix 250K chip, 71,206 in EVS and 2,533 common to both.

To define true HBD in the simulated inbred individuals, founder haplotype labels of the 316,963 variants were used. The true inbreeding coefficient (f_{true}) of each of the 1,000 first-cousin offspring was calculated by dividing the genome length that is HBD (in cM) by the total genome length obtained by adding the genetic distance between the first and the last variants on each autosome (in cM).

2. Methods

Different submap selections were performed on data using Affymetrix 250K markers, or our selection of frequent WES polymorphisms. Three methods were compared:

1) **FSuite_dft:** FSuite with default option (creating 100 random submaps with one marker every 0.5 cM)
2) **FSuite_1hot:** FSuite with 1 random submap created from recombination hotspots.
3) **FSuite_100hot:** FSuite with 100 random submaps created from recombination hotspots.

All these methods used the allele frequencies of EUR of 1000G.

3. Simulation results

For each method, we plotted the *f* estimates obtained using either Affymetrix 250K data or WES data as a function of the f_{true} values for the 1,000 first-cousin offspring (Figure S1).

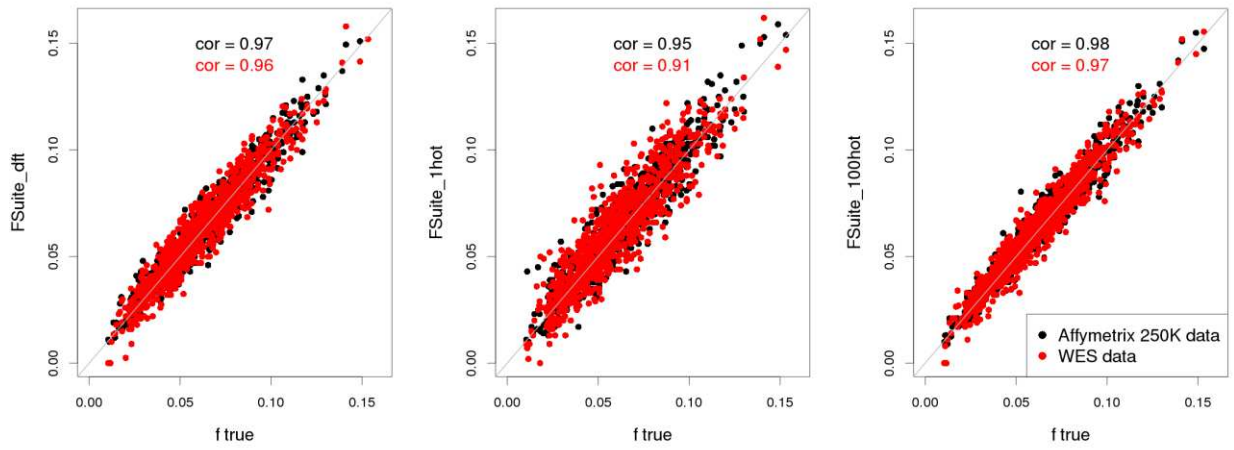

Figure S1: FSuite estimations with Affymetrix 250K SNP array and whole exome sequencing (WES) data. Each graph plots the *f* estimates on both Affymetrix 250K data (black) and WES data (red) as a function of their f_{true} values. The correlations have been calculated between the f_{true} values and the *f* estimates on Affymetrix 250K data (text in black), and between the f_{true} values and the *f* estimates on WES data (text in red).

The results obtained with Affymetrix 250K data and WES data were very similar, whatever the submap selection method, suggesting that FSuite can be used with WES data that do not uniformly cover the genome.

Note that the best correlations were obtained for the multiple submaps based on recombination hotspots, followed by the default option and finally the single submap based on recombination hotspots.

4. Resources

1000 Genomes haplotypes and frequencies were downloaded at

http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html.

EVS frequencies were downloaded at http://evs.gs.washington.edu/EVS/.

References

Delaneau, O., Zagury, J.F. and Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies, *Nat Methods*, **10**, 5-6.