BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 13 2014, pages 1940–1941
doi:10.1093/bioinformatics/btu149

Genetics and population analysis

Advance Access publication March 14, 2014

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,†] and Anne-Louise Leutenegger[1,4,†]

[1] Inserm, U946, Genetic variability and human diseases, Paris, 75010, [2]Université Paris Sud, Kremlin-Bicêtre, 94270, [3]Fondation Jean Dausset CEPH, Paris, 75010, [4]Université Paris-Diderot, UMR 946, Institut Universitaire d'Hématologie, Paris, 75475, [5]Inserm, U1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest, 29218 and [6]Centre Hospitalier Régional Universitaire de Brest, Brest, 29200, France

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: FSuite is a user-friendly pipeline developed for exploiting inbreeding information derived from human genomic data. It can make use of single nucleotide polymorphism chip or exome data. Compared with other software, the advantage of FSuite is that it provides 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 and implementation: 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 or steven.gazal@inserm.fr

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

Received on November 18, 2013; revised on February 26, 2014; accepted on March 10, 2014

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 Genome-wide Complex Trait Analysis (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). However, none of these applications provides 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, which 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.*, 2012), i.e. homozygosity mapping on inbred cases from genome-wide association study (GWAS), 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 two 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

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

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 (i) inferring an individual as inbred by comparing the maximized likelihood with the one to be outbred with a likelihood ratio test, (ii) estimating the proportion of mating types in a population and (iii) 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 is calculated. 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. When applied on GWAS data, this allows to perform HBD-GWAS strategy directly (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 Fig. 1 for different examples). Graphs similar to the ones of Figures 1B and 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 single nucleotide polymorphism (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 (Supplementary Information).

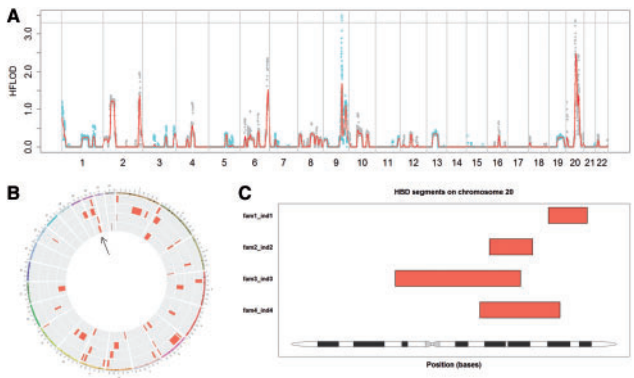

Fig. 1. FSuite graphical outputs. These graphs were generated with FSuite example dataset. Graph (A) shows the genome-wide HFLOD plot. Dots are HFLOD at each marker. The solid line is a moving average, to remove the impact of a submap with a false positive signal. In graph (B), the HBD segments are represented 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

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 the user's computer. FSuite pipeline includes a detailed documentation and a simulated example dataset. It also includes an example dataset of five cases affected with a rare disease genotyped on 180 160 SNPs, among whom four 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 not only 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-point estimates, ROH-based estimates and HMMs modeling LD (Gazal *et al.*, 2014). 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.

Funding: The plateforme de génomique constitutionnelle (Faculté de médecine, Université Paris-Diderot, Paris, France) to S.G.

Conflict of Interest: none declared.

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.* (2014) Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. *Hum. Hered*, **77**, doi:10.1159/000358224.

Genin,E. *et al.* (2012) Could inbred cases identified in GWAS data succeed in detecting rare Rercessive 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.