Estimation of relationships and inbreeding from sequence data in presence of admixture

Steven Gazal[1,2], Mourad Sahbatou[3], Marie-Claude Babron[4,5], Emmanuelle Génin[6,7], Anne-Louise Leutenegger[4,5]

Summary

The 1000 Genomes Project (TGP) provides a unique source of whole genome sequencing data for studies of human population genetics and human diseases.

Because of the presence of admixed populations, we performed simulations to study the robustness of the inbreeding coefficient estimation in the presence of admixture. We found that our multi-point approach (FSuite) was quite robust to admixture unlike single-point methods (PLINK, Purcell et al. 2007).

We then estimated the genomic inbreeding coefficient of each individual and found an unexpected high level of inbreeding in TGP. Inbred individuals were found in each of the 26 populations, with some populations showing proportions above 50%. We also detected 227 previously unreported pairs of close relatives (up to and including 1st-cousins).

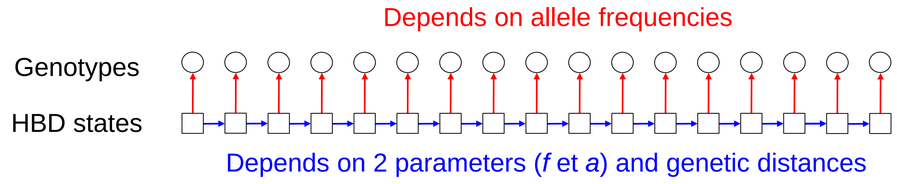
Objectives

1. **Study the impact of admixture on FSuite estimates**
2. **Apply FSuite on final release of 1000 genomes data**

FSuite

1) A multi-point model

- Model the genome of an individual as a hidden Markov chain



- Assumptions :
  1. Marker alleles are independent conditionally on homozygosity-by-descent (HBD) state
  2. *f* is the probability to be HBD at a locus
  3. 1/*a* is the expected HBD segment length
  4. Homogeneous population, i.e. no admixture
- Estimation of *f* and *a* by maximum likelihood
- Implemented in FSuite/FEstim software (Leutenegger et al. 2003; Gazal et al 2014)

2) In presence of admixture

- Admixture causes an excess of heterozygote genotypes
- Admixture biases single-point methods estimating kinship and inbreeding coefficients (Thornton et al. 2012, Moltke and Albrechtsen 2013)
- Is it still true for multi-point methods? (Thompson and Kuhner 2014)

Simulation study

Simulation of individuals with mixed ancestry from Europe and Africa using:

- A sample of 300 individuals: 6 x 1C, 6 x 2C, 18 x 3C, 30 x 4C, 240 OUT
  1st (1C); 2nd (2C); 3rd (3C), 4th cousin (4C) or outbred (OUT) offspring
- HapMap3 haplotypes for founder individuals: 232 CEU, 226 YRI haplotypes of 987k SNPs

➢ f_{true} : true inbreeding coefficient

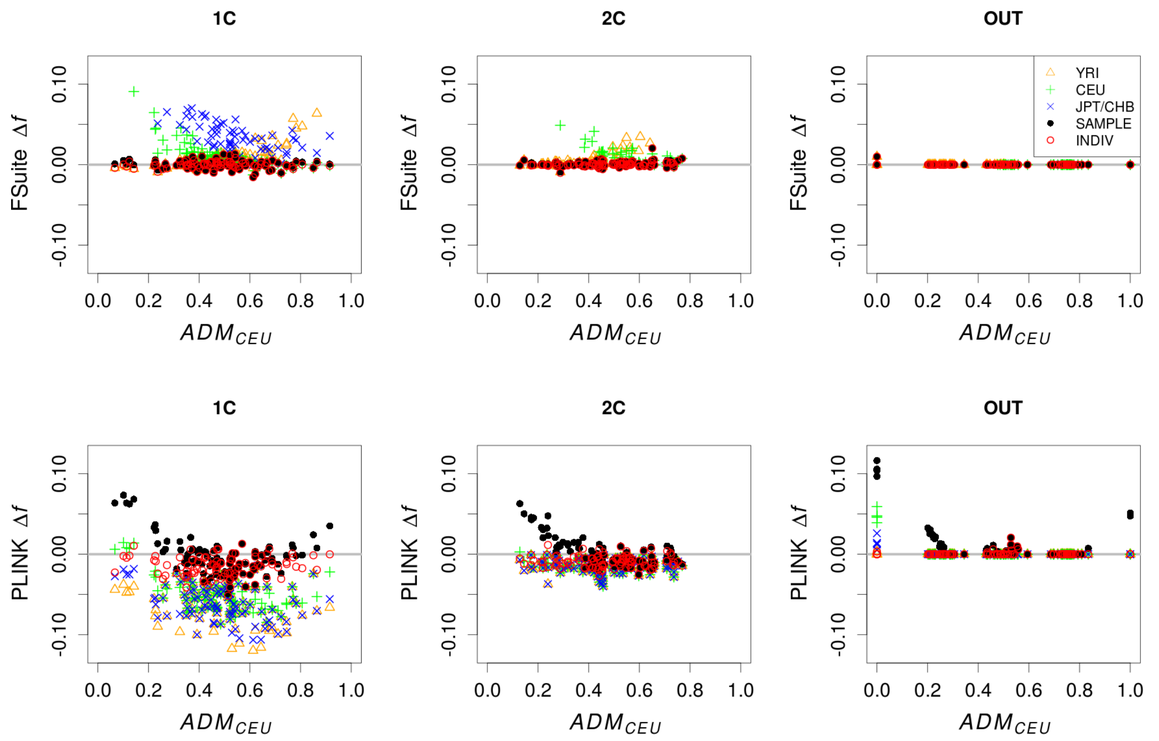
➢ ADM_{CEU} : true proportion of CEU ancestry

➢ Estimation of *f* using different sets of allele frequencies: by default from the data (SAMPLE), Europe (CEU), Africa (YRI), Asia (JPT/CHB), weighting CEU and YRI according to the individual ADM_{CEU} (INDIV)

$$\Delta f = f_{estimated} - f_{true}$$

➢ 100 replicates

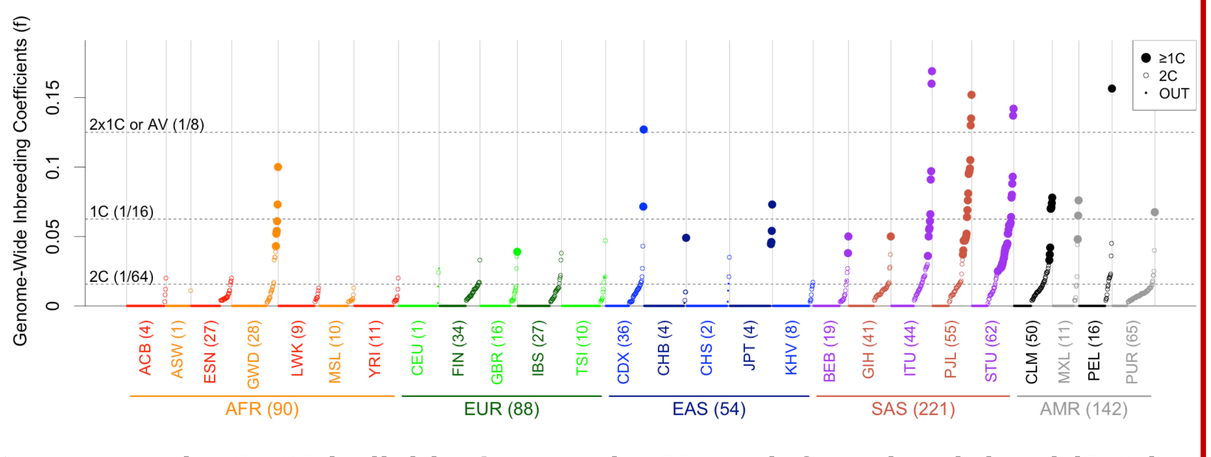
Simulation results



➢ **FSuite provides reliable *f* estimates even when there are some admixed individuals in the studied population**

1000 Genomes Project (TGP)

	Final Phase
African (AFR)	**661**
African Caribbean in Barbados (ACB)	96
African Ancestry in Southwest United States (ASW)	61
Esan in Nigeria (ESN)	99
Gambian in Western Division, The Gambia (GWD)	113
Luhya in Webuye, Kenya (LWK)	99
Mende in Sierra Leone (MSL)	85
Yoruba in Ibadan, Nigeria (YRI)	108
European (EUR)	**503**
Utah residents with European ancestry (CEU)	99
Finnish in Finland (FIN)	99
British in England and Scotland (GBR)	91
Iberian populations in Spain (IBS)	107
Toscani in Italy (TSI)	107
East Asian (EAS)	**504**
Chinese Dai in Xishuangbanna, China (CDX)	93
Han Chinese in Bejing, China (CHB)	103
Southern Han Chinese, China (CHS)	105
Japanese in Tokyo, Japan (JPT)	104
Kinh in Ho Chi Minh City, Vietnam (KHV)	99
South Asian (SAS)	**489**
Bengali in Bangladesh (BEB)	86
Gujarati Indian in Houston,Texas (GIH)	103
Indian Telugu in the United Kingdom (ITU)	102
Punjabi in Lahore, Pakistan (PJL)	96
Sri Lankan Tamil in the United Kingdom (STU)	102
Admixed American (ADM)	**347**
Colombian in Medellin, Colombia (CLM)	94
Mexican Ancestry in Los Angeles, California (MXL)	64
Peruvian in Lima, Peru (PEL)	85
Puerto Rican in Puerto Rico (PUR)	104
TOTAL	**2,504**



➢ **Among the 2,497 individuals tested, 595 are inferred as inbred (24%)**
94 individuals are likely to be 1C (68), AV (1) or 2x1C (25)

Conclusion

- Multi-point methods provide reliable *f* estimates even when there are some admixed individuals in the studied population.
  Fsuite **http://genestat.cephb.fr/software/index.php/FSuite**
- An application on final release of 1000 Genomes Project reveals a high proportion of inbred individuals, especially in South Asian populations.

1 INSERM, IAME, UMR 1137, F-75018 Paris; 2 Plateforme de génomique constitutionnelle du GHU Nord, APHP, Hôpital Bichat, F-75018 Paris; 3 Inserm, U946, Genetic variability and human diseases, Paris; 4 Univ Paris-Diderot, IUH, Paris; 5 Fondation Jean Dausset CEPH, Paris; 6 Inserm, U1078, Génétique, Génomique fonctionnelle et Biotechnologies, Brest; 7 Centre Hospitalier Régional Universitaire de Brest;

Anne-Louise Leutenegger, INSERM U946, Genetic Variability and Human Diseases, 27 rue Juliette Dodu, 75010 Paris – anne-louise.leutenegger@inserm.fr