Leveraging consanguinity in the UK Biobank cohort to identify rare recessive variants involved in complex traits

Sidonie Foulon*,[1], Margot Derouin[1], Marie-Sophie Ogloblinsky[2], Steven Gazal[3], Hervé Perdry[4] and Anne-Louise Leutenegger[1]

Introduction

- GWAS Genome-wide association studies identify genetics variants involved in complex traits[(1,2)]. GWAS are most powerful at detecting common variants with an additive effect but these variants do not explain all the phenotypic variability. To study the role of rare recessive variants, other strategies need to be developed.
- Consanguinity Offspring of relatives are consanguineous. In absence of family data, a consanguineous individual can be identified through his genome which carries homozygous-by-descent (HBD) segments (Fig.1). Rare recessive variants have higher chances to be found in HBD segments.

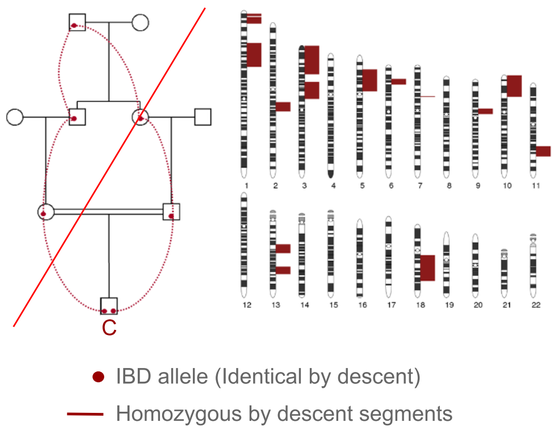

Fig.1: Consanguinity of an individual descended from first cousins.

- HBD-GWAS The HBD-GWAS method[(3)] proposes to analyze **consanguineous cases only** to identify regions where cases share more HBD segments than expected based on their level of consanguinity.
- When controls are available, we observe that **consanguineous controls** can also share HBD segments (Fig.2).

➔ HBD-LOGISTIC We propose here an approach that **relies both on consanguineous cases and controls** . It searches the genome for regions with an excess of HBD segments shared among cases compared to controls.

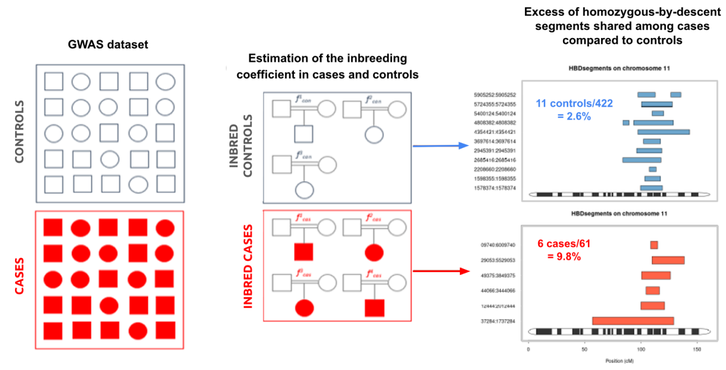

Fig.2: Comparison of shared HBD segments between cases and controls.

Methods

- The first steps of the methods are common for HBD-GWAS and HBD-LOGISTIC strategies
  - For each individual : using maximum likelihood in a hidden Markov model, estimation of the inbreeding coefficient f , and test of H0: f = 0 vs H1: f > 0
  - For consanguineous individuals : computation of HBD probabilities (pHBD) and FLOD score (depending on pHBD and f) along the genome of each individual
- The last step is specific to HBD-LOGISTIC
  - Logistic regression analysis with either FLOD or pHBD as main explanatory covariate. Results can be plotted through quantile-quantile plot and Manhattan plot.
- Fantasio HBD-GWAS and HBD-LOGISTIC are implemented in Fantasio, an R package (github.com/genostats/Fantasio) coded in C++ and R. Fantasio relies on R package Gaston (on CRAN).
- UK Biobank cohort We evaluate the model on UK Biobank prospective cohort (~500,000 individuals living in the United-Kingdom). We construct the phenotypic status (diabetes phenotype) for all individuals using fields available in UK Biobank (diabetes diagnosed by a doctor, self reported illness, age at diagnosis). We infer the origin of the individuals in 6 groups of genetic ancestry using a random forest classifier based on PCA from 1000 Genomes and HGDP-CEPH reference panels.
- Sensitivity We evaluate our approach under the hypothesis of non association (H0) in order to assess the sensitivity of the model. With this simulation, we also aim to choose which explanatory variable is the most suitable. The simulation is based on permutations of the observed phenotypes. We also construct a phenotype depending on the value of f (individuals with higher f are more likely to be affected) to assess the properties of the method when the disease is associated with inbreeding, without any specific region of the genome involved (H0 of non association).
- Application We illustrate HBD-LOGISTIC and compare it with HBD-GWAS with an application to diabetes in Middle-East individuals from UK Biobank.

Results

- Consanguinity and diabetes cases in UK Biobank We find an excess of consanguineous individuals among diabetes cases for some populations: Middle-East, Africa, Central South Asia (Tab.1).

Population	Africa (n = 5 557)	America (n = 968)	Central South Asia (n = 8 191)	East Asia (n = 2 407)	Middle East (n = 1 212)	Europe (n = 430 578)
# Cases (proportion in population)	591 (11 %)	39 (4 %)	1 332 (16 %)	129 (5 %)	133 (11 %)	20 804 (5 %)
# Consanguineous (proportion in population)	620 (11 %)	103 (11 %)	3 378 (42 %)	275 (11 %)	487 (40 %)	10 317 (2 %)
# Consanguineous in cases (proportion in cases)	87 (15 %)	4 (10 %)	582 (44 %)	16 (12 %)	62 (48 %)	519 (2.6 %)
# Consanguineous in controls (proportion in controls)	533 (11%)	99 (11%)	2 796 (41%)	259 (11%)	426 (40%)	9 798 (2%)

Tab.1: Distribution of UKBB individuals, inbred and diabetics for each population

- Sensitivity We test the analysis of Middle-East data with simulated phenotypes, using FLOD and pHBD+f explanatory variables: (A) & (B) for phenotypes obtained by permuting the observed phenotypes, and (C) & (D) for the phenotypes depending on f (Fig.3).

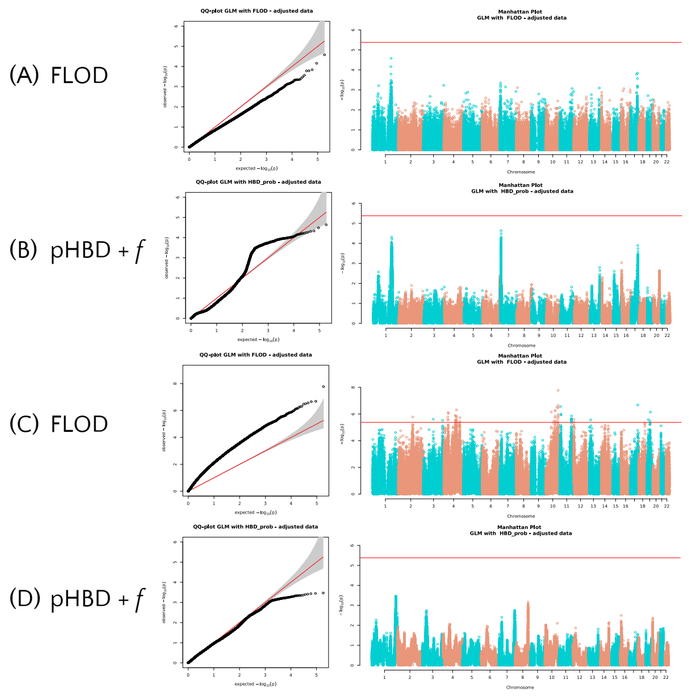

Fig.3: HBD-LOGISTIC with simulated phenotypes

- Application to UK Biobank data We compare the results given by HBD-LOGISTIC on case-control data with what we obtained with the initial HBD-GWAS on case only data for the Middle-East population origin (Fig.4).

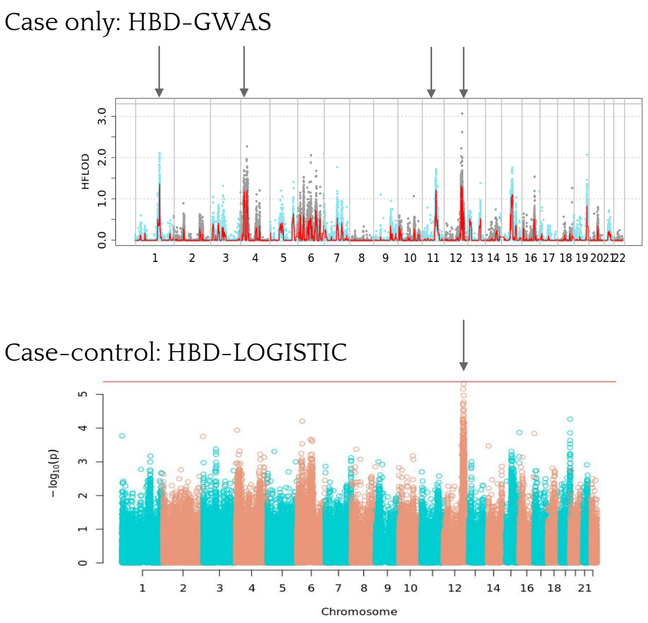

Fig.4: Comparison between HBD-GWAS and HBD-LOGISTIC

Conclusion

Evaluation of the method allows us to conclude that

- sufficient amount of consanguineous individuals are available in UK Biobank, some origins offer better results
- out of the several associations found by HBD-GWAS, we only keep the ones where the difference of proportion of shared HBD segments in cases vs. controls is high.
- however, the current implementation of HBD-LOGISTIC does not produce the expected distribution of p-values for data simulated under the hypothesis H0 of non-association, even more if the phenotype depends on f. Changes and extensions of the implementation are planned.

1 NeuroDiderot, Inserm, Université Paris Cité, UMR1141, 48 bd Sérurier, 75019, Paris, France
2 GGB, Inserm, UBO, UMR1078, 22 avenue Camille Desmoulins, 29200, Brest, France
3 University of Southern California, 1450 Biggy Street, 90033, Los Angeles, USA
4 CESP Inserm U1018, Université Paris Saclay, 16 Avenue Paul-Vaillant-Couturier, 94807, Villejuif, France

Submitting author email: sidonie.foulon@inserm.fr

(1) Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42**, D1001–D1006 (2014).
(2) Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat Commun* **11**, 5900 (2020).
(3) Génin, E. *et al.* Could inbred cases identified in GWAS data succeed in detecting rare recessive variants where affected sib-pairs have failed? *Hum Hered* **74**, 142–152 (2012).

Acknowledgments : This research was conducted using the UK Biobank mediacl database under application 59366 – Method Developments for the genetic analysis of complex traits and was funded by the inserm cross-cutting program GOLD (GenOmics varaibility in Health and Disease)

GitHub