Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics

Florian Privé[1]

[1]National Centre for Register-based Research, Aarhus University, Aarhus, 8210, Denmark.

Contact:

- florian.prive.21@gmail.com

Abstract

The UK Biobank project is a prospective cohort study with deep genetic and phenotypic data collected on almost 500,000 individuals from across the United Kingdom. Within this dataset, we carefully define 17 distinct ancestry groups from all four corners of the world. Using allele frequencies derived from these global reference groups, we are now able to effectively measure diversity from summary statistics of any genetic dataset. Measuring genetic diversity is an important problem because increasing genetic diversity is key to making new genetic discoveries, while also being a major source of confounding to be aware of in genetics studies.

Several projects have focused on providing genetic data from diverse populations, such as the HapMap project, the 1000 genomes project, the Simons genome diversity project, and the human genome diversity project (International HapMap 3 Consortium *et al.* 2010; 1000 Genomes Project Consortium *et al.* 2015; Mallick *et al.* 2016; Bergström *et al.* 2020). However, these datasets do not contain many individuals per population and therefore are not large enough to provide accurate allele frequency estimates for diverse worldwide populations. The UK Biobank project is a prospective cohort study with deep genetic and phenotypic data collected on almost 500,000 individuals from across the United Kingdom. Despite being a cohort from the UK, this dataset is so large that it includes individuals that were born in all four corners of the world. Here we carefully use information on self-reported ancestry, country of birth, and genetic similarity to define 17 distinct ancestry groups from the UK Biobank to be used as global reference populations. This includes five groups with genetic ancestries from Europe, four from Africa, three from South Asia, three from East Asia, one from the Middle East, and one from South America (Table 1). The detailed procedure used to define these groups is presented in the Supplementary Materials. We then provide reference allele frequencies for 5,816,590 genetic variants across these 17 diverse ancestry groups.

Then, as a use-case for the new reference set of allele frequencies we provide, we propose to estimate global ancestry proportions from a cohort based on its allele frequencies only (i.e. summary statistics). We propose to find the convex combination of ancestry proportions α_k (positive and sum to 1) which minimize the following problem: $\sum_{j=1}^{M}\left(w_j f_j^{(0)} - \sum_{k=1}^{K}\alpha_k w_j f_j^{(k)}\right)^2$, where M is the number of variants, K the number of reference populations, $f_j^{(k)}$ is the frequency of variant j in population k, and $f_j^{(0)}$ is the frequency of variant j in the cohort of interest. This is similar to Summix (Arriaga-MacKenzie *et al.* 2021), except for two differences. First, we introduce weights $w_j = 1/\max\left(\sqrt{f_j^{(0)}(1-f_j^{(0)})},\ 0.1\right)$ to account for more variability in more common variants. Second, and most importantly, we use the new reference set of allele frequencies we present here (instead of the five continental 1000 Genomes populations) for measuring genetic diversity with more precision. This optimization problem is a standard quadratic programming problem that can be solved using R package quadprog (Turlach *et al.* 2019), and that we implement in function `snp_ancestry_summary` in our R package bigsnpr (Privé *et al.* 2018).

We download several GWAS summary statistics which reports allele frequencies and apply this method to both validate our new reference set we provide and show one of its potential use. We first apply function `snp_ancestry_summary` to more homogeneous samples as an empirically validation. When applying this function (using the new reference set we provide) to the Biobank Japan (Japanese cohort), FinnGen (Finnish), a Peruvian cohort, a Qatari cohort and Sub-Saharan African cohort, the ancestry proportions estimates obtained match expectations (Table 1). When comparing our estimates with reported ancestries, for example PAGE is composed of 44.6% Hispanic-Latinos, 34.7% African-Americans, 9.4% Asians, 7.9% Native Hawaiians and 3.4% of some other ancestries (self-reported), whereas our estimates are of 24.9% South American, 21.9% European, 35.4% African, 2.0% South Asian, 11.3% East Asian, 3.9% Filipino, and 0.5% Middle East. GWAS summary statistics from either European ancestries or trans-ethnic ancestries all have a substantial proportion estimated from European ancestry groups, while ancestries from other continents are still largely underrepresented (Table 1).

Here we provide an unprecedentedly large and diverse reference set of allele frequencies using the UK Biobank. We then show how to effectively measure diversity from GWAS summary statistics using this new

Population	N	BBJ	FinnGen	Perú	Qatar	Africa	UKBB	GERA	PAGE	BrCa	PrCa	CAD	body fat	covid	eczema	epilepsy	urate
Africa (West)	735					36.7	1.4	1.9	26.4	0.2	0.2	2.4	1	4	0.3	0.6	3.6
Africa (South)	449					63.3	0.3	0.4	6.8		0.1	0.9	1.1	1.2	0.1		4
Africa (East)	276				11.4				0.2				0.2	0.1			1.4
Africa (North)	268				22.1			0.4	2							0.3	
Middle East	523				63.5		1.3	3.4	0.5	2	2.2	5.2	2.9	1.5	1.6	3.2	2.8
Ashkenazi	1975				1.9		1.4	6	0.9	1.3	3.3	2.2	2.3	1.3	2	3.6	1.5
Europe (North East)	320		8.9				3.8	7.7	0.7	9.6	10	10	11	11.7	10.1	8.9	7.5
Finland	143 (+ 99)		85.2				2.7	3.5	0.7	11.9	14.8	8.2	10.5	15.4	13.1	7.1	4.3
British-Irish Isles	4000		5.1				79.1	53	8.9	59.8	56.1	46.6	49.2	54.8	61.7	62	36.6
Europe (South West)	603						7.7	12.2	10.7	11.2	11.2	7.5	9.5	5.3	10.1	11.2	7.3
South America	473 (+ 84)		0.4	99.1			0.1	4.5	24.9	0.9	0.6	1.4	0.7	1.8	0.4	0.1	0.2
Japan	240 (+ 104)	96.1	0.4	0.9				1.8	8.8	0.2	0.2	2.1	3	0.4	0.4		25.9
Asia (East)	961	3.9					0.2	3.7	2.5	0.6	0.3	2.5	1.8	0.2		3	2.3
Philippines	295						0.1	1.4	3.9					0.2	0.1	0.1	0.5
Bangladesh	223 (+ 86)						0.6		1.2	1.1	0.8	2.8	2.8	0.9			1.4
Sri Lanka	372				1.1		0.4		0.8	0.8	0.2	2.2	1.6	0.4			0.7
Pakistan	400						0.8			0.4		6	2.5	0.7			

Table 1: Reference population with their size (N) and corresponding ancestry proportions in several GWAS datasets. Citations for the allele frequencies used: the Biobank Japan (BBJ, Sakaue *et al.* (2021)), FinnGen (`https://r5.finngen.fi/`), GWAS in Peruvians (Asgari *et al.* 2020), GWAS in Qataris (Thareja *et al.* 2021), GWAS in Sub-Saharan Africans (Africa, Chen *et al.* (2019)), the UK Biobank (UKBB, Bycroft *et al.* (2018)), GERA (Hoffmann *et al.* 2018), PAGE (Wojcik *et al.* 2019), breast cancer (BrCa, Michailidou *et al.* (2017)), prostate cancer (PrCa, Schumacher *et al.* (2018)), coronary artery disease (CAD, Nikpay *et al.* (2015)), body fat percentage (Lu *et al.* 2016), COVID-19 (The COVID-19 Host Genetics Initiative and Ganna 2021), eczema (Paternoster *et al.* 2015), epilepsy (The International League Against Epilepsy Consortium on Complex Epilepsies 2018), and serum urate (Tin *et al.* 2019). Several of them have been downloaded through the NHGRI-EBI GWAS Catalog (MacArthur *et al.* 2017).

set of reference allele frequencies. Measuring genetic diversity is an important problem because increasing genetic diversity is key to making new genetic discoveries, while also being a major source of confounding to be aware of in genetics studies. Our work has limitations though. First, although we managed to define ancestry groups with more than 240 individuals, it seems that these groups are still not large enough for the ancestry deconvolution method to recover 100% of one ancestry (e.g. the case for the FinnGen and Biobank Japan cohorts which are likely very homogenous). In this case, we also capture a small proportion from nearby ancestry groups. Second, it is not fully clear whether we can capture any ancestry, e.g. Native Hawaiians in the PAGE study; it seems that they are partly captured by the "Philippines" ancestry group we define. Third, with the 17 ancestry groups we defined, we think we are capturing a large proportion of the genetic diversity in both Europe and Africa, but we may still be missing more fine-grained diversity in both South Asia and South America. Fourth, when we use the frequencies reported in the GWAS summary statistics, it is not clear whether they were recomputed from all individuals (e.g. before performing any quality control and filtering), and for meta-analyses of binary traits, whether they were computed as weighted average using total sample sizes or effective sample sizes. Anyway, despite these limitations, we envision that the new set of reference allele frequencies we provide will have many useful applications.

Software and code availability

The newest version of R package bigsnpr can be installed from GitHub (see `https://github.com/privefl/bigsnpr`). The set of reference allele frequencies for 5,816,590 genetic variants across 17 diverse ancestry groups we create here can be downloaded from `https://figshare.com/ndownloader/files/31218445`. All code used for this paper is available at `https://github.com/privefl/freq-ancestry/tree/main/code`. We have extensively used R packages bigstatsr and bigsnpr (Privé *et al.* 2018) for analyzing large genetic data, packages from the future framework (Bengtsson 2021) for easy scheduling and parallelization of analyses on the HPC cluster, and packages from the tidyverse suite (Wickham *et al.* 2019) for shaping and visualizing results.

Acknowledgements

F.P. is supported by the Danish National Research Foundation (Niels Bohr Professorship to John McGrath) and by a Lundbeck Foundation Fellowship (R335-2019-2339 to Bjarni J. Vilhjálmsson). The author thanks GenomeDK and Aarhus University for providing computational resources and support that contributed to these research results. This research has been conducted using the UK Biobank Resource under Application Number 58024.

Declaration of Interests

The author declares no competing interests.

References

1000 Genomes Project Consortium *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68.

Arriaga-MacKenzie, I. S., Matesi, G., Chen, S., Ronco, A., Marker, K. M., Hall, J. R., Scherenberg, R., Khajeh-Sharafabadi, M., Wu, Y., Gignoux, C. R., *et al.* (2021). Summix: A method for detecting and adjusting for population structure in genetic summary data. *The American Journal of Human Genetics*.

Asgari, S., Luo, Y., Akbari, A., Belbin, G. M., Li, X., Harris, D. N., Selig, M., Bartell, E., Calderon, R., Slowikowski, K., *et al.* (2020). A positively selected FBN1 missense variant reduces height in peruvian individuals. *Nature*, **582**(7811), 234–239.

Bengtsson, H. (2021). A Unifying Framework for Parallel and Distributed Processing in R using Futures. *The R Journal*.

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., *et al.* (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, **367**(6484).

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203–209.

Chen, J., Sun, M., Adeyemo, A., Pirie, F., Carstensen, T., Pomilla, C., Doumatey, A. P., Chen, G., Young, E. H., Sandhu, M., *et al.* (2019). Genome-wide association study of type 2 diabetes in Africa. *Diabetologia*, **62**(7), 1204–1211.

Hoffmann, T. J., Theusch, E., Haldar, T., Ranatunga, D. K., Jorgenson, E., Medina, M. W., Kvale, M. N., Kwok, P.-Y., Schaefer, C., Krauss, R. M., *et al.* (2018). A large electronic-health-record-based genome-wide study of serum lipids. *Nature genetics*, **50**(3), 401–413.

International HapMap 3 Consortium *et al.* (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**(7311), 52.

Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., Cousminer, D. L., Dastani, Z., Drong, A. W., Esko, T., *et al.* (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications*, **7**(1), 1–15.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., *et al.* (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, **45**(D1), D896–D901.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., *et al.* (2016). The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, **538**(7624), 201–206.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., *et al.* (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**(7678), 92.

Nikpay, M., Goel, A., Won, H.-H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., *et al.* (2015). A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, **47**(10), 1121.

Paternoster, L., Standl, M., Waage, J., Baurecht, H., Hotze, M., Strachan, D. P., Curtin, J. A., Bønnelykke, K., Tian, C., Takahashi, A., *et al.* (2015). Multi-ethnic genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*, **47**(12), 1449.

Privé, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, **34**(16), 2781–2787.

Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., *et al.* (2021). A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics*, pages 1–10.

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., *et al.* (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature Genetics*, **50**(7), 928.

Thareja, G., Al-Sarraj, Y., Belkadi, A., Almotawa, M., Suhre, K., and Albagha, O. M. (2021). Whole genome sequencing in the middle eastern Qatari population identifies genetic associations with 45 clinically relevant traits. *Nature Communications*, **12**(1), 1–10.

The COVID-19 Host Genetics Initiative and Ganna, A. (2021). Mapping the human genetic architecture of covid-19 by worldwide meta-analysis. *MedRxiv*.

The International League Against Epilepsy Consortium on Complex Epilepsies (2018). Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nature communications*, **9**.

Tin, A., Marten, J., Kuhns, V. L. H., Li, Y., Wuttke, M., Kirsten, H., Sieber, K. B., Qiu, C., Gorski, M., Yu, Z., *et al.* (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature Genetics*, **51**(10), 1459–1474.

Turlach, B. A., Weingessel, A., and Moler, C. (2019). *quadprog: Functions to Solve Quadratic Programming Problems*. R package version 1.5-8.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., *et al.* (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**(43), 1686.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., *et al.* (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, **570**(7762), 514–518.