JASs Journal of Anthropological Sciences
Open Access
forum Vol. 99 (2021), pp. 159 - 161
doi. 10.4436/jass.99012

Biobank-scale datasets and their relevance to anthropology/popgen

Francesco Montinaro

Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Estonia;
Department of Biology-Genetics, University of Bari, Bari, Italy
e-mail:francesco.montinaro@gmail.com

Since the completion of the Human Genome Project, the analysis of genome-wide data has become crucial under many different subjects, enormously increasing our knowledge in medical, epidemiological and clinical perspectives. In the attempt to increase the understanding of the biological basis of diseases, or, more generally, phenotypic variation, many national and international Institutions put a substantial effort into the creation of "Biobanks" - a collection of a large number of biological samples associated to a series of experimental analysis and other types of data.

The use of biobank data gave a further boost to the "race for personalised medicine" that, despite many important advancements, is still far to be universally achieved.

The availability of thousands of genome-wide individual data, often from the same geographic area, and their association to multiple and diverse kinds of data has also raised the possibility of studying evolutionary and anthropological aspects of human populations that were impossible to tackle using limited datasets.

In this context, knowing the fine-scale genetic composition of a population, addressing both its temporal and geographical variation, is not only important from an epidemiological perspective but has profound implications in the understanding of the connection between social, cultural and biological changes in populations, which is essential in an XXI century evolving society. As an example, American individuals are among the most thoroughly studied populations, and although many important features of their recent evolution have been unlocked well before the so-called Genomic Era, many interesting and crucial details were added analysing large datasets. The growing availability of samples from both sides of the Atlantic sea revealed the extremely diverse picture of the American genome mosaic, adding insights to some of the most tremendous events of Human history such as the Atlantic Slave Trade and the Colonial Era. It has recently emerged that continental contributions to American populations are different quantitatively and qualitatively, with subtle but significant differences observed also at a microgeographic scale. These differences are also evident when admixture recombination decay pattern is exploited to provide a temporal dimension to the complex admixture history, revealing that the genomic contribution of different African or European populations is chronologically diverse, providing, for the first time, a "genomic stratigraphy" of the continents' influences (Ongaro et al. 2019; Micheletti et al. 2020). Furthermore, the analysis of demographic trajectory through time has revealed the differential ancestry specific demographic impact of the colonial Era in different areas (Ongaro et al. 2019). Interestingly, it has been recently suggested that the ancestry proportion differences and pattern of relatedness currently present in the US nowadays might be the results of recent long and short-range movements, including the forced relocation of "native" American in the country (Bryc et al. 2015). It is important to note that for most of the researches cited above, only a limited amount of genealogical, social, biological and cultural data linked to participants were available, and although the large sample size of the analysed dataset provided

useful insights, the huge potential of a multilayered investigation has not been fully unlocked. In this perspective, including as much as possible data related to participants' family (including health history) and sociocultural factors may fill the gap towards a complete understanding of recent movements and admixture, fully recovering the evolutionary and biological impact of this complex phenomenon. As an example, the combined analysis of genetic and genealogical data for ~770,000 US individuals of European ancestry allowed the reconstruction of subtle, complex demographic forces in shaping the patterns of genetic variation among contemporary North Americans (Han et al. 2017). Similarly, the analysis of thousands of individuals from many different countries, coupled with haplotype-based methods allowed the recognition of important differences in closely related human groups, shedding light on a rich history of recent migrations and diverse admixture histories. For example, the analysis of the Identity by Descent (IBD) blocks sharing allowed the reconstruction of the population size changes through time, revealing the impact of global and local historical events, such as plague pandemics or geo-political phenomena (Abdellaoui et al. 2013; Pankratov et al. 2020).

The availability of massive genetic datasets provided important insights also in the characterisation of very distant events. Virtually all non-African individuals harbour a variable amount (1~4%) of genomes derived by a complex history of interbreeding with archaic humans, such as Neanderthals and Denisovans.

Given the limited number of archaic derived fragments in any single genome, the analysis of biobank-scale dataset is essential to achieve a detailed knowledge of the interbreeding dynamics, in the attempt to resolve many long-standing questions, such as the number of archaic encounters and/or their relative impact in different populations, together with the characterisation of our extinct cousins. In fact, resurrecting archaic human fragments interspersed in the genomes of more than 25,000 Icelandic individuals led to the observation of the indirect Denisova influence on Europe and the refinement of the biological impact of archaic introgression (Skov et al. 2020) Moreover, the comparison of mutation patterns in archaic and sapiens genomic fragments revealed the existing of differences in the relative occurrence of mutation types, suggesting differences in sex-specific generation intervals between the species. Similarly, the comparison of thousands of genome-wide data and phenotypes or Electronic Health Records in individuals of European ancestry contributed to uncover the association of archaic-derived polymorphism with many different traits such as skin and hair color, immune response and psychological traits (Simonti et al. 2016; Dannemann et al. 2017; Dannemann and Kelso 2017). In this context, the analysis of many biobank-scale datasets from different populations will help to clarify the extent of these effects, helping further characterise the interaction between archaic genetic legacy and environment in the 50,000 years following the interbreeding.

One of the most challenging tasks in human population genetics is represented by the identification of genetic markers that have been the target of recent selective pressure and adaptation, given the confounding effect of demography and admixture in selection analysis. Furthermore, putative selected variants identified from genomic scan for natural selection are seldom replicated, limiting our knowledge of populations adaptive history, and its application into evolutionary-guided epidemiological, translational and medical studies which constitutes a promising approach towards a full knowledge of the biology of our organism. So far, only a limited number of surveys harnessed genome-wide dataset composed by thousand individuals to identify putative selected polymorphism. The availability of high coverage sequence data led to the development of a novel singleton distribution-based (SDS, Singleton Density Score) approach that infers changes in allele frequencies that occurred at a recent time scale (Field et al. 2016). The application of this approach on British populations confirmed previous adapted candidates and suggested very recent adaptation for alleles involved in blue eyes and blond hair.

Furthermore, the analysis of polygenic selection patterns suggested that selection has driven the changes of frequency for alleles related to height along the genomes, together with other complex traits, informing on the importance of polygenic adaptation in shaping human genetic variation. Interestingly, the combination of population genetics and SDS suggested the existence of different selective patterns among different regions in Estonia, emphasizing the importance of considering hidden genetic structure both for evolutionary studies and medical applications (Pankratov et al. 2020)

Sadly, as previously observed for translational studies (Sirugo et al. 2019), virtually all of the established Biobanks including genome data are almost exclusively focusing on Eurasians and some American populations, jeopardizing the medical characterisation of some populations and possibility to achieve a global picture of the demographic and adaptive history of our species. Collecting and analysing biological, phenotypic and medical data from understudied areas, embracing a truly ethical and collaborative approach with national and Local institutions should be a globally shared priority for the whole scientific community.

References

Abdellaoui A, Hottenga J-J, de Knijff P, et al (2013) Population structure, migration, and diversifying selection in the Netherlands. Eur J Hum Genet 21:1277–1285. https://doi.org/10.1038/ejhg.2013.48

Bryc K, Durand EY, Macpherson JM, et al (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet 96:37–53. https://doi.org/10.1016/j.ajhg.2014.11.010

Dannemann M, Kelso J (2017) The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. Am J Hum Genet 101:578–589. https://doi.org/10.1016/j.ajhg.2017.09.010

Dannemann M, Prüfer K, Kelso J (2017) Functional implications of Neandertal introgression in modern humans. Genome Biol 18:61. https://doi.org/10.1186/s13059-017-1181-7

Field Y, Boyle EA, Telis N, et al (2016) Detection of human adaptation during the past 2000 years. Science 354:760–764. https://doi.org/10.1126/science.aag0776

Han E, Carbonetto P, Curtis RE, et al (2017) Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat Commun 8:14238. https://doi.org/10.1038/ncomms14238

Micheletti SJ, Bryc K, Ancona Esselmann SG, et al (2020) Genetic Consequences of the Transatlantic Slave Trade in the Americas. Am J Hum Genet 107:265–277. https://doi.org/10.1016/j.ajhg.2020.06.012

Ongaro L, Scliar MO, Flores R, et al (2019) The Genomic Impact of European Colonization of the Americas. Curr Biol 29:3974–3986.e4. https://doi.org/10.1016/j.cub.2019.09.076

Pankratov V, Montinaro F, Kushniarevich A, et al (2020) Differences in local population history at the finest level: the case of the Estonian population. Eur J Hum Genet 28:1580–1591. https://doi.org/10.1038/s41431-020-0699-4

Simonti CN, Vernot B, Bastarache L, et al (2016) The phenotypic legacy of admixture between modern humans and Neandertals. Science 351:737–741. https://doi.org/10.1126/science.aad2149

Sirugo G, Williams SM, Tishkoff SA (2019) The Missing Diversity in Human Genetic Studies. Cell 177:26–31. https://doi.org/10.1016/j.cell.2019.02.048

Skov L, Coll Macià M, Sveinbjörnsson G, et al (2020) The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes. Nature 582:78–83. https://doi.org/10.1038/s41586-020-2225-9