doi:10.20944/preprints202004.0394.v1

Population Genetic Considerations for Using Biobanks as International Resources in the Pandemic Era and Beyond

Hannah Carress[1], Daniel John Lawson[2], and Eran Elhaik[3]

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom
[2] Department of Mathematics, University of Bristol, Bristol, United Kingdom
[3] Department of Biology, Lund University, Lund, Sweden
* Please address all correspondence to Eran Elhaik at eran.elhaik@bio.lu.se

ABSTRACT

The past years saw the rise of genomic biobanks and mega-scale meta-analysis of genomic data that promise to reveal the genetic underpinnings of health and disease. However, the over-representation of Europeans in genomic studies not only limit the global understanding of disease risk and intervention efficacy, but also inhibit viable research into the genomic differences between carriers and patients. Whilst the community has agreed that more diverse samples are required, it is not enough to blindly increase diversity; the diversity must be quantified, compared, and annotated to lead to insight. Genetic annotations from separate biobanks need to be comparable, computable, operate without access to raw data due to privacy concerns. Comparability is key both for regular research and to allow international comparison in response to pandemics. Here, we evaluate the appropriateness of commonly used genomic tools used to depict population structure in a standardized and comparable manner. The end goal is to reduce the effects of confounding and learn from genuine variation in genetic effects on phenotypes across populations, which will improve the value of biobanks, locally and internationally, increase the accuracy of association analyses, and inform developmental efforts.

Keywords: bioinformatics; population structure; population stratification bias; genomic medicine; biobanks

doi:10.20944/preprints202004.0394.v1

INTRODUCTION

Box 1: Glossary

Admixed Population
A population of individuals with ancestors from two or more relatively distinct populations relatively recent in human history.

Admixture mapping
Gene mapping of susceptibility alleles for genetic disease that show differential risk by ancestry, correlating the degree of ancestry near to genomic regions with greater disease risk.

Bayesian clustering
Assignment of individuals to clusters based on genetic similarity without assuming predefined populations, using statistical methods that allow inferences to be drawn from the data and prior information.

Expectation-Maximisation (EM) algorithm
An iterative method to find maximum likelihood estimates (MLE) of parameters in statistical models, altering between an expectation step (creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters) and a maximisation step (computes parameters maximizing the output of the expectation step).

Genetic Relatedness Matrix
The GRM represents the genomic similarities among all individuals. Each cell in the matrix measures the genotypic correlation between a pair of individuals (the rows and columns). The GRM can be used with a phenotypic distance matrix to estimate heritability without estimating the phenotypic effect of individual SNPs.

Hidden Markov Model (HMM)
A statistical Markov model, that is a randomly changing system assumed to consist of future states that only depend on current states, whereby states are unobservable (hidden).

Markov Chain Monte Carlo (MCMC)
A simulation method used in Bayesian calculations, incorporating a class of algorithms that can obtain a sample of the desired distribution by observing several steps of the Markov chain, which is a sequence of a probability of events that depend only on the state of the previous event and has the desired distribution at its equilibrium.

Maximum Likelihood Estimate (MLE)
Method for selecting the best hypothesis from a set of alternatives on the basis of which contains the fewest evolutionary changes.

Maximum optimisation
Maximising a function by choosing input values from a specified set and using this to calculate the value of the function.

Mixed Linear Models (MLM) or Linear Mixed Models (LMMs)
Mixed linear models that incorporate both fixed and random effects where there is non-independence in the data. These models are used as a form of "global correction" as they

doi:10.20944/preprints202004.0394.v1

account for both ancestry and relatedness.

Multidimensional Scaling (MDS)
A type of multivariate analysis which allows multidimensional information to be displayed graphically (usually two dimensions) with minimum loss of information whilst preserving distance between data points. MDS uses pairwise distances between points as input.

Principal Components Analysis (PCA)
A type of multivariate analysis that allows multidimensional information to be displayed graphically with minimum loss of information while preserving covariance of data. PCA is typically applied to genotype data as a form of "global correction" and the derived PCs are used for further analyses.

Association studies and biobanks

Association studies aim to detect whether genetic variants found in different individuals are associated with a trait or disease of interest, by comparing the DNA of individuals that vary in relation to the phenotypes (Byun *et al.*, 2017). This article addresses how ancestry influences on genetic associations could be robustly compared across international biobank projects. Whilst associations are scientifically invaluable for a range of questions we will outline below, as we completed this review the world plunged into the COVID-19 pandemic. If our recommendations had been in place, biobanks would be better positioned to answer why some people develop severe disease, and others do not, whilst accounting *separately* for ancestry differences and other co-morbidity differences between populations. For example, since the major-histocompatibility-complex antigen loci (HLA) are the prototypical candidates that modulate the genetic susceptibility to infectious diseases, association studies aim at identifying their types may provide valuable information for strategizing prevention, treatment, vaccination, and clinical approaches (Shi *et al.*, 2020). Such cardinal questions striking the core differences between individuals, families, communities, and populations necessitated genomic biobanks.

The completion of the human genome allowed genomic biobanks to be envisioned. After the International HapMap Project, practicality the first international biobank (Belmont *et al.*, 2005) facilitated routine collection of data for genome wide association studies (GWAS) (Visscher *et al.*, 2012), these became the leading genetic tool for phenotype-genotype investigations. Over time, GWAS have been used to identify associations between thousands of variants for a wide variety of traits and diseases, with mixed results. GWAS drew much criticism concerning their validity, error rate, interpretation, application, biological causation (Ikegawa, 2012) and replication (Palmer and Pe'er, 2017). Since much of this criticism concerned small sample sizes with reduced the power of association analyses and yield spurious associations, major efforts were taken to recruit thousands and tens of thousands of participants into studies where their biological data and prognosis were collected. These collections served as the basis for what is considered today as a (genomic) biobank (Somiari and Somiari, 2015).

Today, biobanks are known as massive scale datasets containing many hundreds of thousands of participants from specified populations. Biobanks have brought enormous power to association studies. Their potential to enable personalised treatment became clear even before the technology and matured and both private and government-sponsored banks began amassing tissues and data (Kaiser, 2002). For example, Generation Scotland (Smith *et al.*, 2006) includes DNA, tissues, and phenotypic information from nearly 30,000 Scots (Generation Scotland, 2016), the 100,000 Genomes Project, sequenced the genomes of 100,000 NHS patients with rare diseases aiming to understand the aetiology of their conditions from their genomic data (Caulfield *et al.*, 2017) and the UK Biobank project, sequenced the complete genomes of over half a million individuals (Sudlow *et al.*, 2015) with the aim of improving the prevention, diagnosis, and treatment of a wide range of diseases (Bycroft *et al.*, 2018). Pending projects include the Genome Russia Project, which aims to fill the gap in the mapping of human populations by providing the whole-genome sequences of some 3,000 people from a variety of regions of Russia (Oleksyk *et al.*, 2015). Biobanks are not without controversy. In Iceland, deCODE genetics has created the world's most extensive and comprehensive population data collection on genealogy, genotypes and phenotypes of a single population. However, the economic value of the genomic data remained largely inaccessible and the company filed for bankruptcy (Kaiser, 2009). The experience of deCODE highlighted the risks in entrusting private companies to manage genomic databases, promoting similar efforts to have at least a partial government control in the dozens of newly founded biobanks (reviewed in (Dubow and Marjanovic, 2016) (Figure 1). Moreover, as the use of biobanks is expanding beyond their locality, for example, in the case of rare conditions where samples need to be pooled from multiple biobanks, the view of biobanks should be changed from localized managed resources to more global resources that should adhere to international standards to increase the accuracy of association studies and the use of biobanks (Scudellari, 2013).

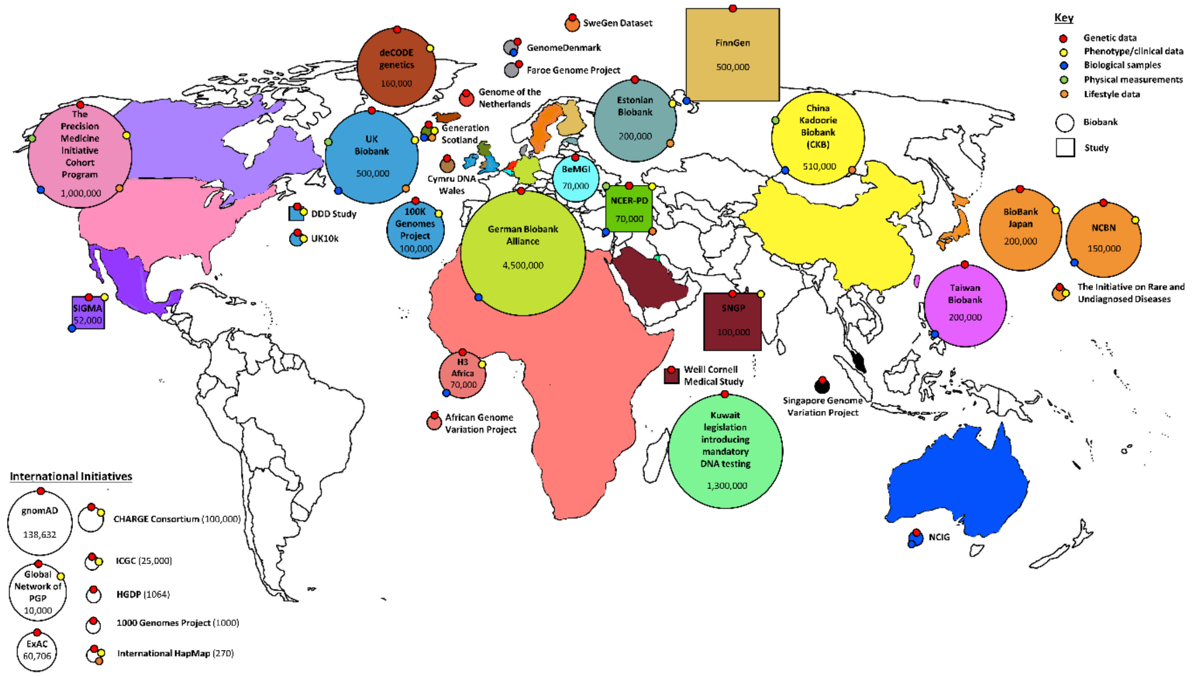

SweGen Dataset
GenomeDenmark
Faroe Genome Project
deCODE genetics 160,000
FinnGen 500,000
Key
Genetic data
Phenotype/clinical data
Biological samples
Physical measurements
Lifestyle data
Biobank
Study
The Precision Medicine Initiative Cohort Program 1,000,000
UK Biobank 500,000
Genome of the Netherlands
Generation Scotland
Estonian Biobank 200,000
China Kadoorie Biobank (CKB) 510,000
Cymru DNA Wales
BeMGI 70,000
NCER-PD 70,000
DDD Study
UK10k
100K Genomes Project 100,000
German Biobank Alliance 4,500,000
BioBank Japan 200,000
NCBN 150,000
Taiwan Biobank 200,000
The Initiative on Rare and Undiagnosed Diseases
SIGMA 52,000
SNGP 100,000
H3 Africa 70,000
Weill Cornell Medical Study
Singapore Genome Variation Project
African Genome Variation Project
Kuwait legislation introducing mandatory DNA testing 1,300,000
NCIG
International Initiatives
gnomAD 138,632
CHARGE Consortium (100,000)
ICGC (25,000)
Global Network of PGP 10,000
HGDP (1064)
1000 Genomes Project (1000)
ExAC 60,706
International HapMap (270)

Figure 1: *The distribution of national biobanks throughout the world, colour-coded by country or continent. International initiatives have been listed separately. The size of the shapes represents the intended size of the biobank/study (intended number of samples/individuals). Coloured dots around the edge of the shapes represent the type of information that the biobank contains (ref key).*

Even past the formation of biobanks, many associations results failed to replicate (e.g., Border *et al.*, 2019) or differ in effect worldwide - e.g. BMI (Lawson *et al.*, 2020), Schizophrenia (Li *et al.*, 2015) Hypertension (Wain, 2014) and Parkinsons' disease (Nalls *et al.*, 2014). Although strong associations between genetic variants and a phenotype typically replicated within the population that was studied, they may not have been replicated elsewhere.

This leads naturally to further questioning the value and cost-effectiveness of association studies and biobanks (Chalmers *et al.*, 2016) – what do the associations mean, and what are they useful for? How can we decide whether the association is relevant for different individuals, particularly those of mixed origins or those who may not know their origins? What are the considerations when designing a new biobank or merging data from multiple biobanks? We argue that understanding population structure is a key component to answering these questions and contribute to the sustainability of biobanks. In the following, we review the current state of knowledge on the importance of population structure to association studies and biobanks and the implications to downstream analyses. We then review biobank relevant models that describe population structure. We end with the challenges and benefits of the tools that implements these models.

Population diversity

Human genetic variation is a significant contributor to phenotypic variation among individuals and populations, with single-nucleotide polymorphisms (SNPs) being the most common form of genetic variation. Of the entire human genomic variation, only a paucity (12%) is between continental populations and even less genetic variation (1%) is between intra-continental populations (Elhaik, 2012). In other words, a relatively small group of SNPs are geographically differentiated, whilst a much larger group of SNPs vary among individuals irrespective of geography. However, most of these variants are rare and non-functional (Kamm *et al.*, 2019). Both common and functional variants are strong predictors of geography, phenotypes and cultural practices that may be linked with the risk for a disease. Thereby, geographical and ancestral origins can not only inform of us of what risk of disease an individual has, but also modify the effect of treatment (Yusuf and Wittes, 2016). In general, and with the clear exception for high admixture or migration followed by relative isolation (Ramachandran *et al.*, 2005; Das *et al.*, 2016; Marshall *et al.*, 2016), most associations between geographic location and genetic similarity are expected to hold worldwide (e.g., Elhaik *et al.*, 2014), due to the exchange of genes and migrants between geographically proximate populations (e.g., Rosenberg *et al.*, 2002; Mountain and Risch, 2004; Jakobsson *et al.*, 2008; Li and Yu, 2008; Xing *et al.*, 2009). These relationships are also expected to hold for common and rare variants (Altshuler *et al.*, 2012). The geographic differentiation between populations underlies their genetic variation or population structure, and studies in the field aim to analyse, describe, or account for the genetic variation in time and space within and among populations.

Unfortunately, worldwide diversity misrepresented in GWAS studies (Sirugo *et al.*, 2019). By 2009, 96% of individuals represented in GWAS were of European descent (Need and Goldstein, 2009). This over-representation was rationalized by the interest to focus on ancestrally “homogenous” populations to avoid *population stratification bias*, i.e., systematic ancestry differences due to different allele frequencies in the studied cohorts that produced false positives (Hindorff *et al.*, 2018). Consequent efforts to carry out studies on non-Europeans met with some success; by 2016, the proportion of Europeans included in GWAS declined to 81% (Popejoy and Fullerton, 2016) and further to 78% in 2019 (Sirugo *et al.*, 2019). However, even then, 71.8% of GWAS individuals are recruited from only three countries: the US, UK, and Iceland (Mills and Rahal, 2019).

Not all major genetic datasets are equally diverse, and most are skewed towards individuals of European ancestry (Figure 2). For example, 61% of the samples in the Exome Aggregation Consortium (ExAC) dataset (60,252 individuals) (Lek *et al.*, 2016), 59% of the Genome Aggregation Database (gnomAD) (141,456 individuals) (Karczewski *et al.*, 2019), 94%

of the UK Biobank database (500,000 individuals) (Sudlow *et al.*, 2015) and an estimated 97.6% of the deCODE database are Europeans (Jurczak, 2017). The UK Biobank was designed to be representative of the general population of the United Kingdom, however, that makeup is only 85% "White" (Tutton, 2009). Such misrepresentation of the global population structure has a detrimental impact on genomic medicine studies in England and international studies that rely on their results for several reasons: firstly, they promote a simplified view of "Europeans" as "homogeneous" (Elhaik *et al.*, 2014); secondly, ignorance of the global population structure also prevents properly correcting the studies for *stratification bias*; and thirdly, in many trials the number of participants enrolled in each country remains small, which increases the risk for false positives due to chance or undetected population structure, even if a correction is attempted (Lawson *et al.*, 2020). This, in turn, discourages further studies of under-represented populations. These limitations were underscored during the COVID-19 pandemic as the UK biobank data were shared internationally (Dyer, 2020) to improve the response to the virus and protect the public represented in the biobank (Figure 2).

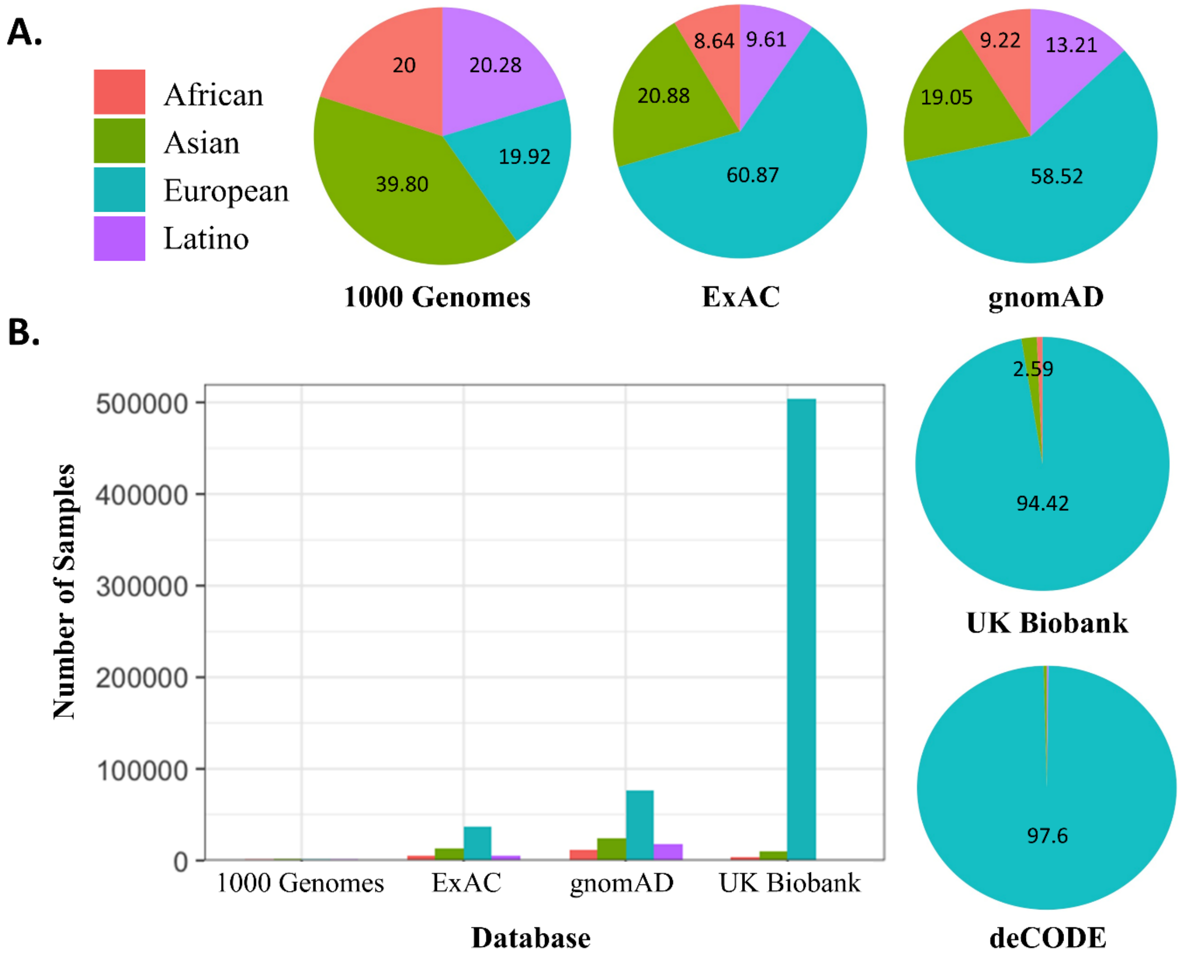


Figure 2: *The (a) percentage and (b) number of samples in the 1000 Genomes Project, the ExAC browser, the UK Biobank and the gnomAD browser categorised into five ancestry groups: European, South Asian, African, East Asian and Latin (https://www.nature.com/articles/nature15393; http://exac.broadinstitute.org/faq; https://gnomad.broadinstitute.org/faq). The deCODE database has been circled in (a) and excluded in (b) because when contacted deCODE genetics were unable to disclose any information regarding the ethnicity or number of samples; however, it can assumeed that the database is roughly 97.6% European based on the finding of the recent consensus where 97.6% of the Icelandic population was defined as European (93% Icelandic and 3.1% Polish)* (Jurczak, 2017).

Population stratification may bias GWAS through two routes: the choice of cohort, and association analysis. Currently, individuals are matched and grouped mainly using self-reported "race," rather than genomic ancestry. This criterion is believed to account for the participants genetic background and supposedly allow controlling for population genetic structure (e.g. Baughn *et al.*, 2018, 2020). A numerical example of how a false positive association can be created due to population stratification is shown in Table 1 (Hellwege *et al.*, 2017). However, grouping based on demographics alone does not account for differences in genetic ancestry between individuals, which leads to biased interpretation of the results, false negative or positive results (Campbell *et al.*, 2005; Wang, Localio and Rebbeck, 2006; Chikhi *et al.*, 2010; Yusuf and Wittes, 2016; Elhaik and Ryan, 2019).

Genomic medicine and diversity

Personalised medicine is thought of as the utilisation of epidemiological knowledge to a granular classification of patients into cohorts that differ in their disease susceptibility, disease prognosis, or response to treatment and considered the epitome of the 21st century medicine (Lesko and Woodcock, 2004). To facilitate the accurate identification and classification of individuals into cohorts it is necessary to consider their genomes, which lends credence to the development or *genomic medicine* and its aspired derivation *personalised genomic medicine.*

Genomic medicine seeks to deploy the insights that the genetic revolution has brought about in medical practice (Feero, Guttmacher and Collins, 2010). The ability to predict individual risk of disease development, guide intervention, and direct the treatment is the core of genomic medicine (Guttmacher and Collins, 2002). Most applications outside of simple Mendelian diseases start by considering known associations and testing for them in the sequence of the patient. Harnessing the knowledge gained from a small fraction of patients into the routine care of new patients has the potential to expand diagnoses outside of rare diseases, determine optimal drug therapy and effectiveness through targeted treatment, and allow for a more accurate prediction of an individual's susceptibility to disease – the pillars of the genomic medicine vision (Johnson *et al.*, 2019).

Personalised genomic medicine aims to tailor a treatment to an individuals' genetic needs and is expected to revolutionise disease treatment by using targeted therapy and treatment tailored to the individual to achieve the most effective outcome (Figure 3) (NHS, 2016). This form of genomic medicine was made feasible due to advances in computational biotechnology and its implementation into the health care system (Figure 4) (Pasic *et al.*, 2013) alongside biological advancements that include the mapping of human genetic variation across the world, through parallel global efforts (Brieger *et al.*, 2019). However, it remains a futuristic vision rather than an everyday reality, due to the multiple obstacles that genetic studies face in deciphering the complex genotype-phenotype relationships (Manolio *et al.*, 2013; Weitzel *et al.*, 2016). One of the notorious difficulties in the field is the variation among population subgroups, which is often due to their genomic background (Yusuf and Wittes, 2016). Personalisation to the ancestral group-level is a more realistic short-term goal, yet being well-represented in genomic datasets is the exception rather than the rule. For example, an individual of Aramean ancestry living in the UK would be matched to only a handful of individuals in the UK Biobank. Similarly, individuals from Transcaucasia may be considered either "Europeans" or "Asians", and poorly represented by either, as their populations resemble an older admixture between these continental groups (Elhaik *et al.*, 2014; De Barros Damgaard *et al.*, 2018). Personalised medicine and drug discovery are two areas that are particularly affected by a lack of diversity in biobanks.

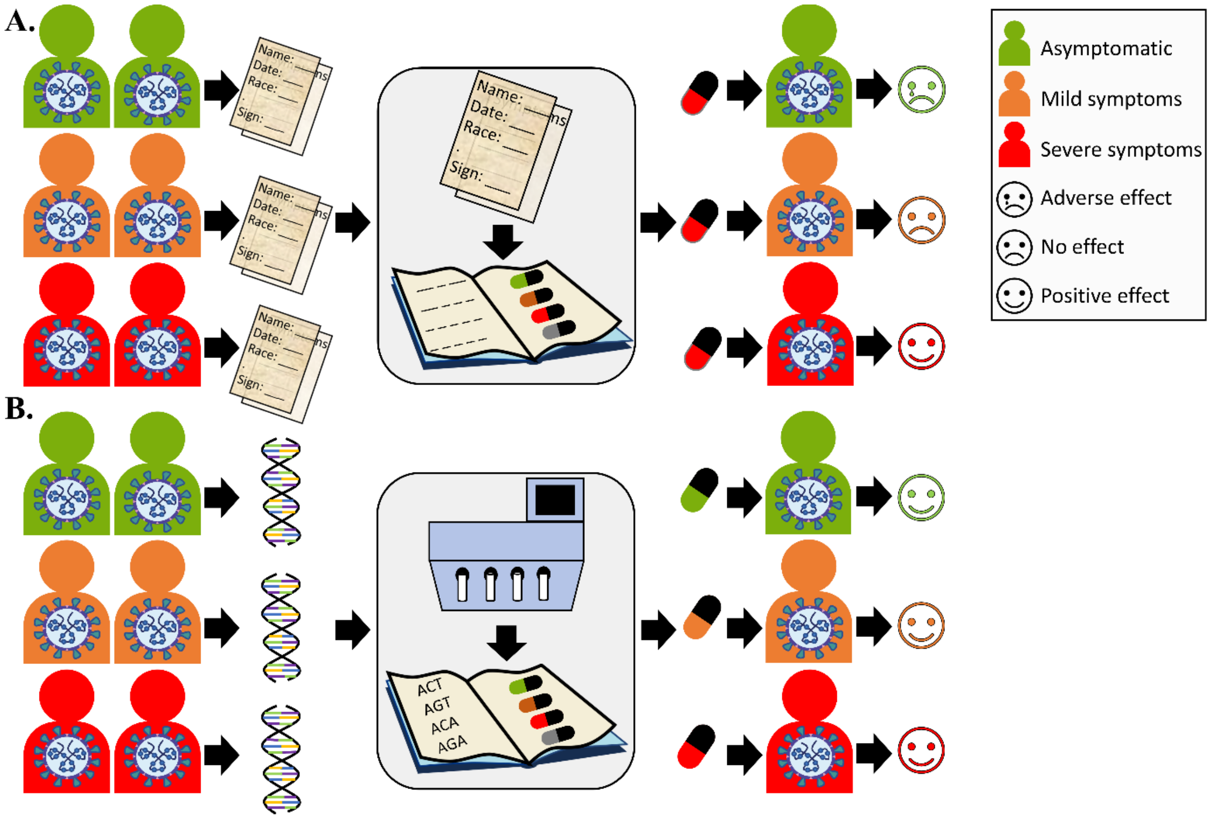
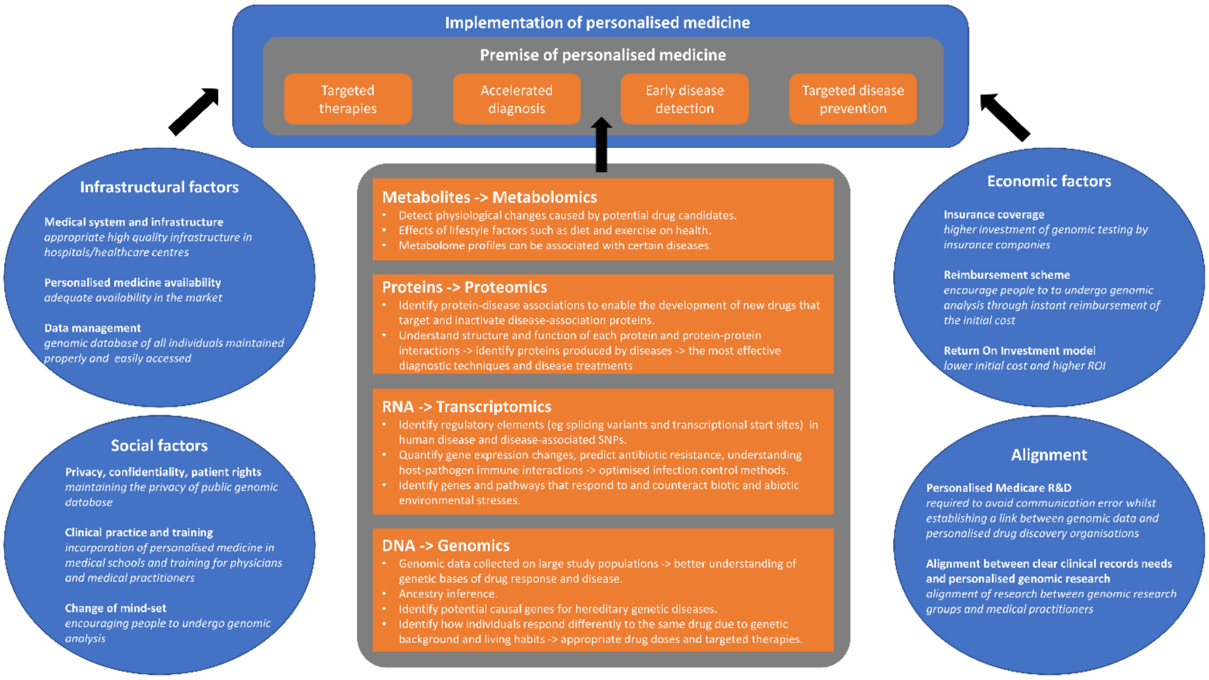


Figure 3: Using the example of COVID-19: (a) The current method of treatment whereby all patients with the same disease receive the same treatment. (b) Personalised medicine, whereby treatment is tailored to an individual to increase effectiveness.

Similar problems emerged in clinical trials, where individuals are matched using demographic criteria, paired and then randomly assigned to treatment and control groups to determine a drug's efficacy (Elhaik and Ryan, 2019). It is well recognized that pharmaceutical research and development is in crisis due to the soaring costs of drugs, which is primarily due to the regulatory process governing Phase-III clinical trials where the successful results of the smaller previous trials may not be reproduced (Roy, 2012). A chief cause for this irreproducibility is population stratification bias caused by the uneven distribution of ancestries and other covariates (such as variation in age distribution, BMI, etc between countries influencing COVID-19 mortality) in the treatment and control groups.

This should be unsurprising as these subjects are linked. Drugs are developed for Europeans, and consequently results in unfavourable healthcare outcomes for non-Europeans. Due to the poor understanding of associations in many populations and their under representation in clinical databases, poor drug choices are made and genetic screening of non-Europeans is more likely to produce ambiguous or false-positive diagnoses (Petrovski and Goldstein, 2016; Cruz-Correa *et al.*, 2017).



Implementation of personalised medicine
Premise of personalised medicine
Targeted therapies
Accelerated diagnosis
Early disease detection
Targeted disease prevention

Infrastructural factors

Medical system and infrastructure
appropriate high quality infrastructure in hospitals/healthcare centres

Personalised medicine availability
adequate availability in the market

Data management
genomic database of all individuals maintained properly and easily accessed

Social factors

Privacy, confidentiality, patient rights
maintaining the privacy of public genomic database

Clinical practice and training
incorporation of personalised medicine in medical schools and training for physicians and medical practitioners

Change of mind-set
encouraging people to undergo genomic analysis

Metabolites -> Metabolomics

- Detect physiological changes caused by potential drug candidates.
- Effects of lifestyle factors such as diet and exercise on health.
- Metabolome profiles can be associated with certain diseases.

Proteins -> Proteomics

- Identify protein-disease associations to enable the development of new drugs that target and inactivate disease-association proteins.
- Understand structure and function of each protein and protein-protein interactions -> identify proteins produced by diseases -> the most effective diagnostic techniques and disease treatments

RNA -> Transcriptomics

- Identify regulatory elements (eg splicing variants and transcriptional start sites) in human disease and disease-associated SNPs.
- Quantify gene expression changes, predict antibiotic resistance, understanding host-pathogen immune interactions -> optimised infection control methods.
- Identify genes and pathways that respond to and counteract biotic and abiotic environmental stresses.

DNA -> Genomics

- Genomic data collected on large study populations -> better understanding of genetic bases of drug response and disease.
- Ancestry inference.
- Identify potential causal genes for hereditary genetic diseases.
- Identify how individuals respond differently to the same drug due to genetic background and living habits -> appropriate drug doses and targeted therapies.

Economic factors

Insurance coverage
higher investment of genomic testing by insurance companies

Reimbursement scheme
encourage people to to undergo genomic analysis through instant reimbursement of the initial cost

Return On Investment model
lower initial cost and higher ROI

Alignment

Personalised Medicare R&D
required to avoid communication error whilst establishing a link between genomic data and personalised drug discovery organisations

Alignment between clear clinical records needs and personalised genomic research
alignment of research between genomic research groups and medical practitioners

Figure 4: *the road to personalised medicine. How the use of omics can be used to create the premise of personalised medicine (orange), which can be implemented into the healthcare system through the adoption of a variety of different factors (blue).*

CURRENT BIOBANK STANDARDS REPRESENTING GENETIC VARIATION

Accounting for population differences requires a reliable and global population structure model. Regrettably, despite the vast amount of genetic data currently available, no unified population structure model has been developed. Instead, population genetic studies typically describe variation in the data they study, sometimes with respect to related populations, defined in a rudimentary way, for example using the 14 (or even just the original four) HapMap populations (Altshuler *et al.*, 2010) or 26 of the 1000 Genomes populations (Altshuler *et al.*, 2012). Unsurprisingly, without a model, correcting for population stratification remains strenuous.

Many association studies ignore population stratification or implicitly assume its redundancy if the data were collected from continental groups (e.g., Ooi *et al.*, 2019). Groups are assigned either by self-identified ethnicity or inferred by comparison to the Hapmap or 1000 Genomes populations, and analyse each cluster independently (e.g., Ooi *et al.*, 2019). This approach is implausible, not accounting for the existence of fine-scale structure everywhere (Lawson *et al.*, 2020), and cannot be applied to more admixed populations, which is important where recent massive migrations have occurred, such as in the Americas.

PCs and GRMs

Currently, “global correction” of such populations, using either Principal Components Analysis (PCA see Box 1, e.g. Price *et al.*, 2006) and/or mixed linear models (MLM, Box 1, e.g. (Zhang and Pan, 2015)) starts with the Genetic Relatedness Matrix (GRM, Box 1) (Jiang and Wang, 2018) as the *de-facto* standard used to describe ancestry of large-scale genetic datasets. PCA corrects for the largest variation components of the GRM whilst MLM corrects for the whole matrix, accounting for recently related individuals.

These tools view the genome as a set of independent loci whose effect can be simply added up. Unfortunately, depending on sampling and genetic drift, this can yield spurious results (Novembre and Stephens, 2008; McVean, 2009; Arenas *et al.*, 2013; Elhaik and Ryan, 2019), including representing individuals with two ancestrally different parents as similar to populations that resemble this mixture. For example, an individual with one European and one Asian parent may be incorrectly corrected as a Middle Eastern individual (Elhaik and Ryan, 2019).

Both PCA and MLMs are used for meta-analysis of a large number of independent studies (e.g., BMI (Locke *et al.*, 2015)). Meta-analysis importantly demonstrates replication of effects genetic risk loci and hence minimizes individual cohort bias. However, the effect size estimate of meta-analysis is the averaged effect of the SNP on outcome across several populations. The assumption that the effects of a SNP are equal across populations with different allele frequencies is unlikely to hold for three main reasons. Firstly, many SNPs identified in GWAS are not the causal variant but rather in linkage disequilibrium with one or more causal variants and linkage disequilibrium patterns differ between populations (Clark *et al.*, 2005). Secondly, gene-environment interactions (Purcell, 2002) may contribute to the overall effect of a SNP and these may differ by population (for example, in BMI and exercise, Rask-Andersen *et al.*, 2017). Thirdly, statistical artefacts can arise from differential correction power for stratification across studies (Lawson *et al.*, 2020). The resulting bias is problematic because many downstream applications use summary statistics from GWAS and do not access the original dataset.

IMPLICATIONS OF POPULATION STRUCTURE

Detecting associations between genetics and phenotypes is only the beginning of the process. Different applications are, to various degrees, affected by a bias in the estimates of an effect, which is typically subjected to very large variance for all but the strongest associations.

The assessment of lifestyle factors (e.g., smoking and alcohol intake) on health is the most elementary aspect of health research. Currently, the National Institute for Clinical Excellence (NICE) and other bodies responsible for changing health care practice require Randomised Controlled Trials (RCT) evidence to demonstrate causality between risk factors and disease before putting out findings into practice. RCTs are thereby the "gold standard" for testing a new treatment or identifying causal risk factors for disease. In such trials, participants are randomly exposed either to an exposure (e.g., treatment) or a placebo, to assess the effect of the exposure on disease. However, RCTs may be too costly, time consuming, unfeasible, or unethical to carry-out, which requires alternative solutions.

Causal analysis using Mendelian Randomisation

First outlined by Katan (2004) and further developed by Davey-Smith and Ebrahim, (2003), Mendelian Randomisation (MR) is a statistical approach in which genetic variants associated with an exposure of interest are used to examine the causal effect of said exposure on disease. Because genotype is assigned at conception and common genetic variants are typically not associated with other lifestyle factors, these variants can be used as "instruments" for causal inference, limiting the problems of confounding and reverse causality that otherwise plague observational epidemiology. MR may, therefore, offer an affordable and faster alternative to traditional RCTs (Lippman *et al.*, 2009; Mokry *et al.*, 2015). However, MR assumes that there is no confounding between the genetic polymorphism (which is a proxy for the exposure) and the disease outcome. If population stratification occurs due to mismatched ancestries, then this assumption will be violated, and any estimates will be biased. For instance, common genetic polymorphism in the CHRNA5-A3-B4 gene cluster that is related to nicotine dependence, are often used as an instrument for tobacco smoke exposure (Rask-Andersen *et al.*, 2017). Assume that two alleles, *A* and *C*, exist at this polymorphic site, with those carrying the *A* allele exhibiting a tendency to smoke more cigarettes. Europeans without cryptic African/East Asian ancestry are unlikely to have the *A* allele regardless of their smoking practices, which may bias the MR study if ancestry is not properly accounted for in the study design. Within single studies where researchers have access to individual level data, ancestry may be accounted for, to some extent, by adjusting for principle components. However, MR requires very large sample sizes which necessitates collaboration across studies and meta-analysis, which may introduce genetic heterogeneity. MR's susceptibility to population stratification is a well-recognised bias (Smith, 2010; Hayeck *et al.*, 2015) in case-control pharmacogenetics studies where differences in ancestry affect the results (e.g., weekly warfarin dose requirement to maintain a therapeutic effect varies by ancestry, likely due to genetic variation). Other MR limitations include a reliance on large GWAS, horizontal pleiotropy, and canalisation (Timpson *et al.*, 2018).

Two-sample Mendelian Randomisation (MR), in which the SNP-exposure association is estimated in one study and the SNP-outcome association is estimated in another, is important because it allows sharable summary statistics to be used for causal inference. Often one or both associations are determined using summary statistics and the researcher does not access the primary data (Burgess and Thompson, 2015). Importantly, summary statistics are usually meta-analysed to determine an "average" SNP-exposure estimate across studies, and similarly further studies are meta-analysed to determine the SNP-outcome estimate. Whilst in one step MR, there is an assumption that the effect of the SNP on the outcome and the effect of the SNP on the exposure is uniform across the populations included in any meta-analyses – Two-sample MR makes a further assumption that the population in which the SNP-exposure estimate is determined is representative of the population in which the SNP-outcome association is determined (or that any differences are negligible). This assumption is questionable when combining an exposure GWAS from Han Chinese and an outcome GWAS from a Caucasian population, from which MR may produce biased results (Scheinfeldt *et al.*, 2016; Koellinger and De Vlaming, 2019). Even the induced bias of using two different Caucasian populations (e.g., an exposure GWAS in a Scandinavian population and an outcome measured in a southern England population), is largely unknown. That bias would be most severe for rare conditions and small cohorts that include diverse individuals.

Recently MR studies using a two-sample approach (Bergholdt *et al.*, 2015) have been automated using online platforms (Hemani *et al.*, 2016). In an analysis which is limited to summary data (e.g. Ooi *et al.*, 2019), population stratification bias is difficult to identify, and the analysis is often run without adjustment for possible population differences. Sometimes, the homogeneity of the dataset is assumed due to the continental affiliation of the cohort (e.g., in the case of (Ooi *et al.*, 2019), (Pickrell *et al.*, 2016) analysed third party summary statistics calculated for "Europeans"). Linkage Disequilibrium (LD) score regression (Loh *et al.*, 2015) can estimate the sample overlap between summary statistics, but this is reliant on relatively large samples and often not used in MR pipelines. MR assumptions, and their consequent estimates, would undoubtedly be more trustable if the underlying GWAS estimates were more universal and less population specific.

Polygenic scores

Similar concerns were raised by multiple groups concerning polygenic scores. Sohail et al. (2019) reported that polygenic adaptation signals based on large numbers of SNPs below genome-wide significance were found to be extremely sensitive to biases due to uncorrected population stratification. Berg et al. (2019) analysed the UK Biobank and showed that previously reported signals of selection were strongly attenuated or absent and were due to population stratification. Both papers found that methods for correcting for population stratification in GWAS were not always sufficient for polygenic trait analyses and doubted the strength of the conclusions based on polygenic and advised caution in their interpretation. Further concerns about polygenetic scores were raised by other groups (Bulik-Sullivan *et al.*, 2015; Wellenreuther and Hansson, 2016; Khan, Cooper and Greenland, 2020).

Drug discovery

Association tests are also used to identify drug targets (Finan *et al.*, 2017). Whilst it is not required that the effect sizes are large, they must be associated with an underlying biological pathway (Tam *et al.*, 2019). Other reasons for an association with a trait of interest limit the utility of biobanks for discovery; these include differences in lifestyle between populations, genetic interactions and linkage. The power to detect these unintended associations is set to grow with biobank size, and therefore correcting for population stratification will aid in the reduction of false-positive drug target leads. Since genetic variation is partly geographically differentiated, the frequencies of variants that affect risk and progression or drug response may differ based on geographic location (Ioannidis, Ntzani and Trikalinos, 2004; Daar and Singer, 2005; Adeyemo and Rotimi, 2009; Schärfe *et al.*, 2017). For pharmacogenetics to propel the practice of individualized drug therapy to become the standard of care (Lewis, 2005) accurate genetic profiles should be constructed (Ortega and Meyers, 2014; Yusuf and Wittes, 2016) and genetic tools must be developed and verified that account for confounding effects using DNA sequencing analysis.

MODELS FOR POPULATION STRUCTURE

There are two cases to consider for modelling population structure: when the individual data for all populations are available and when they are not. With access to the individual data, a wide range of options exist, which can be broadly split again to within-dataset and cross-dataset analysis. Within-dataset analysis for biobanks must scale to hundreds of thousands of samples, though need not naturally be comparable. Cross-dataset analyses would typically reference standard datasets, creating a comparable statistic for each individual. Depending on the usage, these references may not themselves be biobank scale. Meta-analysis using summary statistics resembles a cross-dataset analysis with the further requirement of the creation of sharable summary statistics that remains meaningful without individual-level data.

This section summarises the current state of these methods, whilst the USAGE section describes the challenges and benefits of the various tools that are available for each function.

doi:10.20944/preprints202004.0394.v1

DESCRIBING GENETIC VARIATION WITHIN A SINGLE DATASET

Markers for Ancestry

Genomic ancestry inference may employ specialized markers, such as ancestry informative markers (AIMs), which have significant differences in allele frequencies between populations. For instance, the T allele of the SNP rs316598 is very rare in Africans (3.3%) but common elsewhere and can be used to differentiate Africans from non-Africans. AIMs, combined with other methods, can thereby be used to identify the origins of samples provided that the genomes of worldwide reference populations are available (Elhaik *et al.*, 2013, 2014; Elhaik and Ryan, 2019).

One key advantage in using such markers to intensify the ancestry information, which can lead to its identification using downstream tools, is that frequencies of a particular dataset are sometimes already available as summary statistics. If frequency information has been released, then useful ancestry summaries can be extracted. However, to perform such an analysis in practice requires a global model to combine data together and form a meaningful and comparable report on ancestry for each dataset. This will typically require examination of the methods to follow.

Low dimensional representations

PCA aims to reduce the dimensionality of the SNP dataset by reducing the genetic markers into principal components (PCs) (Price *et al.*, 2006; Chang *et al.*, 2015). For population genetic inference, the results are hard to interpret as the PCs do not mean anything intrinsically and often require more than two dimensions to correctly visualise. Important population structure is not always in the top PCs, especially under uneven sampling or genetic drift (McVean, 2009; Elhaik and Ryan, 2019). However, it is fast to compute, amenable to averaging across individuals, and performed empirically well in stratification correction. These characteristics contributed to the popularity of PCA as a method of first choice. However, the results of PCA strongly depend on the choice of markers and samples, and interpretation is subjective without an actual measure of "close to" or "cluster with." Since it has been suggested that PCs portray some geographic similarity within Europe (Novembre *et al.*, 2008), modified methods have been proposed, albeit with limited success. The Spatial admixture analysis (SPA) (Yang *et al.*, 2012), for example, had a biogeographical prediction accuracy of 2% at the country level (Elhaik *et al.*, 2014). These methods cannot be readily applied to biobank-style data as they don't scale.

PCA, despite its failings, is fast and robust. Given enough samples that were carefully chosen, it may be tweaked to form patterns that exhibit similarity to geography (e.g., over 100K carefully chosen samples identified several broad regions in the UK (Figure 1 in (Galinsky *et al.*, 2016)). However, there are alternative and complementary approaches that bear consideration.

Genetic Relatedness Matrices (GRM)

Identity-by-state (IBS) is often used to represent population structure and further represents relatedness (see below). IBS is the proportion of SNPs that are shared between each pair of individuals and therefore forms an *N* by *N* matrix of genetic similarity. Similarly, the association literature uses the "Genetic Relatedness Matrix" (GRM) (Jiang and Wang, 2018) in which SNPs centred by their frequency and weighted by their variance. The GRM is an important tool in mixed linear models that jointly address population structure and relatedness, perhaps the most common tools being GCTA (Yang *et al.*, 2011) and GTAK (Van der Auwera *et al.*, 2013). The GRM can be shown to contain the same information as used by both admixture models and PCA (e.g., (Lawson *et al.*, 2012) supplementary material). The advantage of correcting for the complete matrix (rather than the low-rank approximation used in PCA) is that it retains the relatedness information. Otherwise these procedures are asymptotically equivalent. Fast implementations exist, such as Bolt-LMM (Loh *et al.*, 2018), but these may implicitly exploit low-rank structure and hence lower correction power. Implementations like LMM-OPS (Conomos *et al.*, 2018) attempt to correct increased type-I error rates and a loss of power due to heterogenous ancestry.

Ancestry as a mixture

Admixture or admixture-like analyses originated in the popular program *STRUCTURE* (Pritchard, Stephens and Donnelly, 2000; Falush, Stephens and Pritchard, 2003). Here, the ancestry of each individual is modelled as a proportion of K admixture components, which are learned automatically, and represent "historical populations" in the model. Whilst computation was historically a concern, fast enough implementations now exist (e.g., *faststructure* (Raj, Stephens and Pritchard, 2014) and *terrastructure* (Gopalan *et al.*, 2016)). However, the ancestral interpretation is often misleading (Lawson, van Dorp and Falush, 2018), since sampling and genetic drift can also create the same representation for different true histories. Further, the choice of "proper" K is unclear (Weiss and Lambert, 2014; Lawson, van Dorp and Falush, 2018) and can have significant effects on the inference. Conceptually, admixture uses the same information as PCA (Lawson *et al.*, 2012) and hence suffer from the same limitations.

Sibship, Kinship, and Clanship via Identity by Descent

Since related individuals do not represent statistically independent samples and may lead to false positive associations, association analyses containing related individuals require special care (Moltke and Albrechtsen, 2014; Kuhn, Jakobsson and Günther, 2018). The degree of relatedness is different between individuals and ranges between the largely known sibship, the often-known kinship, and the typically unknown clanship. Relatedness can be identified from DNA, as it is inherited in segments from one's ancestors, whereby long segments are shared between more recently related individuals. This inheritance pattern is used to define genetic regions for two pairs of individuals that are Identical By Descent (IBD).

However, the difficulty in estimating relatedness increases as the relation becomes more distant because the IBD DNA segment between individuals are shorter and more difficult to distinguish from DNA segments that are identical by state (IBS). Typically, all the IBD segments that are more recent that a chosen (average) age of the pairwise relationship are sought by thresholding the lengths of segments.

Alone, IBD is not a measure of ancestry, though its results can be summarised into a Kinship matrix analogously to the GRM. Kinship is the probability that two homologous alleles drawn from each of two individuals are IBD (Thompson, 1975; Speed and Balding, 2015). The value of IBD for detecting associations in biobanks has not been explored, likely due to the complexity of the calculation (N by N analyses), which is time consuming. One possibility is to create an unbiased random sample of genes and traits by sampling only one version of each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 April 2020 doi:10.20944/preprints202004.0394.v1

IBD tract, since the two copies are clearly dependent. Other possibilities are to treat long IBD as a sparse property, reducing the need to generate a full pairwise matrix.

Haplotypes

IBD matches may overlap and may ignore some parts of the genome entirely. An alternative approach is to identify the closest relative for every individual at each position on the genome. This is the approach taken in Chromosome painting (Lawson *et al.*, 2012), which allows the identification of fine-scale population structure, beyond the detection limit of PCA or related approaches (Leslie *et al.*, 2015). Chromosome painting is applicable for samples up to thousands but cannot be used at biobank scale (Pan *et al.*, 2017), because of the same problem of producing an *N* by *N* matrix. Considering large matrices of pairwise haplotype information (throughout the genome) is not trivial and remains a challenge for biobanks.

Local Ancestry

The purpose of Local Ancestry Inference (often abbreviated LAI) is to analyse individual segments of DNA to establish changes in ancestral origin. Being able to assign a SNP as having originated in a particular ancestry, association testing can, in principle, be carried out in each ancestry as if it were a single sample population.

Conceptually, such methods examine a stretch of DNA and use a model related to the mixture approaches to identify the source population. The approaches vary in how appropriate stretches of DNA are defined and how they are matched to the sources. Many approaches use a Hidden Markov Model (HMM) strongly related to Chromosome Painting to assign genomic segments to specified reference populations by exploiting Linkage Disequilibrium.

Current implementations may scale to the thousands (see USAGE) but are limited in scale for learning population structure and are likely to only form a part of a biobank population model when describing external populations. Additionally, the biological parameters needed (e.g., genetic maps, recombination and mutation rate, average ancestry coefficients, and the average number of generations since admixture) may be unknown and are difficult to learn (Dias-Alves, Mairal and Blum, 2018). Considerable effort for biobanks would be required to store, report and use the per-SNP ancestry information returned.

DESCRIBING GENETIC VARIATION WITH AN EXTERNAL REFERENCE

Markers for Ancestry and Projecting PCs

The use of AIMs to represent genetic diversity within a biobank is not well developed. Because AIMs themselves are indicative but not diagnostic of particular population and are a biased sample of the genome (towards ancestry), it is hard to arrive at an ancestry mixture or other measure of structure. However, with efforts in calibration for external datasets, the information required to assess large-scale structures is clearly present in AIMs, which are standard in all commercial microarrays (Illumina, 2013).

It is straight forward to project an individual into the genetic variation of a reference dataset when the reference is described by Principal Components. Associated with each SNP and PC is a weighting, and these must simply be summed. This approach is common in the study of ancient populations, which due to the high missingness of their data are often described in terms of modern variation (Lazaridis *et al.*, 2014).

This has not been performed for biobanks because they contain large variation. However, as discussed above, meta-analysis of many small populations leads to incomplete correction for stratification. Since there is no standard reference, the results of the projection would also be dependent on the choice of the reference populations. Thereby, they can be easily manipulated and are incomparable across studies.

Mixtures of known populations

The ancestry models described above can all be structured to allow comparison of a sample dataset with respect to a reference dataset. *ADMIXTURE* (Alexander and Lange, 2011) is the most popular tool to make "supervised" inference in this way.

When an individual receives ancestry from different sources, they inherit SNPs and haplotypes in proportion to their ancestry from each source. Therefore, significant power can be obtained by considering not only SNPs, but also Haplotypes, quantified either by IBD, Chromosome Painting, or some other technique. These methods describe kinship or haplotype sharing with the reference. This, in turn, can be used to learn an individual's ancestry mixture, which is routinely done for example via Non-Negative Least Squares (NNLS) (Hellenthal *et al.*, 2014; Pagani *et al.*, 2016) or SOURCEFIND (Chacón-Duque *et al.*, 2018). Because the computational cost of these approaches is linear in the size of the target dataset, they can be used at the biobank scale. However, the value of the resulting mixture has yet to be established.

Gene pool models

Frequently, we do not have samples from the underlying ancestral components that led to modern populations. "Gene pool" models allow inferred putative ancestral populations to be used, in place of fixed reference populations. Ancestral populations are first generated from the allele frequencies of a worldwide panel of individuals that correspond to chosen *K* splits, produced by *ADMIXTURE* or alike program. These "populations" correspond to the ancestral populations of all individuals in the dataset. The advantage of creating these populations from a diverse panel of global individuals is that they can be used as reference to infer the admixture components (e.g., through a *supervised ADMIXTURE*) of further individual without changing the model. The admixture components can be used to correct for population stratification (Elhaik and Ryan, 2019), in the same manner as principal components are used, excepting that they model admixture directly, whereas PCA does not. This approach, first employed for biogeography (Elhaik *et al.*, 2014), has been routinely for population genetic investigations and was shown to be applicable to both modern and ancient populations (Flegontov *et al.*, 2016; Marshall *et al.*, 2016; Das *et al.*, 2017). Despite its premise it has yet been implemented in biobanks; the barriers resemble those of Mixture Models in that a "correct" set of gene pools is hard to establish.

Local ancestry models

A local ancestry model can be defined by constructing a reference dataset and applying the local ancestry models to identify ancestry structure within the reference. These approaches have not been widely applied to biobanks in the past, due to issues of scaling. However, as with the genome-wide haplotype approaches, local ancestry can be learned at scale – efficient approaches scale linearly in the Biobank size.

Local genomic ancestry tools are typically used to investigate ancestry on a granular scale, which is necessary when analysing highly admixed individuals, such as African Americans, Latinos or Ashkenazic Jews (Altshuler *et al.*, 2012; Das *et al.*, 2016). The genomes of these individuals constitute a mosaic of geographically and genetically distinct ancestral populations, and local ancestry tools aim to identify the chromosomal boundaries associated with each ancestral population.

However, the promise of comparing to a standard reference simultaneously allows the methods to scale sufficiently and allows comparison across datasets. The key unsolved questions, above those for unlinked methods, are around value. This approach generates extremely large datasets of ancestry information potentially at each SNP. Storing and exploiting such information is a considerable ongoing challenge. Would a fine-scale representation of ancestry help understand the distribution of traits? Does it replace, or complement, the simpler approach of representing ancestry as a proportion of the genome?

USAGE

Markers for Ancestry

AIMs are identified by finding SNPs that are particularly associated with particular populations or geographic regions. Although many sets of AIMs have been published (Elhaik *et al.*, 2013), they were obtained from a handful of populations and their specificity was not validated on other populations. To identify AIMs, it is critical to first assemble a worldwide panel of populations. The search for AIMs is typically performed genome-wide. The putative AIMs should then be evaluated for their specificity and sensitivity in identifying a fine population structure, ideally using a different panel (Esposito *et al.*, 2018). Finally, global ancestry tools can average the ancestry of each contributing population across the individual's AIMs and report the average proportion contributed by each ancestral or parental population.

Ancestry as a mixture

Global genomic ancestry tools can be categorised as shown in (Figure S1) (Table S1). Whilst *STRUCTURE* was initially the most popular approach, it suffered from several disadvantages. First, its accuracy and reliability have been a source of concern (Kalinowski, 2011; Lombaert, Guillemaud and Deleury, 2018). When the diversity of the native population is low, *STRUCTURE* was shown to produce particularly misleading results (Lombaert, Guillemaud and Deleury, 2018). Finally, *STRUCTURE* is a notoriously slow tool, which was soon replaced by dramatically faster implementations.

FRAPPE and *ADMIXTURE* are based on a similar approach to *STRUCTURE,* but both use a maximum likelihood estimate approach to optimise the likelihood for allele frequencies and group memberships, using slightly different algorithms. By default, *ADMIXTURE* uses a block relaxation algorithm that allows for fast convergence and highly accurate parameter estimates (Alexander, Novembre and Lange, 2009) and has an optional Expectation-Maximisation (EM) algorithm. *FRAPPE* uses solely the EM algorithm (Tang *et al.*, 2005), which optimises the likelihood for both allele frequencies and fractional group memberships (Tang *et al.*, 2005). *FRAPPE* has been demonstrated to not only be much more computationally efficient than *STRUCTURE*, but also to produce significantly fewer biased estimates (Tang *et al.*, 2005). However, due to its strict convergence criteria, its EM algorithm is computationally intensive and slower than *ADMIXTURE* (Alexander, Novembre and Lange, 2009), which was reported to have a higher accuracy than *STRUCTURE* and *FRAPPE*.

Spatial approaches, exemplified by *GENELAND* (Guillot, Mortier and Estoup, 2005), *TESS* (Durand *et al.*, 2009) and *BAPS* (Corander, Waldmann and Sillanpää, 2003) are conceptually similar to STRUCTURE, but consider geographical coordinates in their prior distributions, allowing identification of the spatial location of genetic variants between populations. Therefore, these software do not only group individuals genetically into clusters, but are also able to estimate the spatial distribution of these clusters (Corander, Waldmann and Sillanpää, 2003; Guillot, Mortier and Estoup, 2005; Durand *et al.*, 2009). Mitigating privacy concerns, this has the advantage of replacing a real location with a genetically induced one. Yet the approaches are currently rather inaccurate (perhaps due to population structure being more complex than a simple mixture). There are also no scalable implementations.

Bayesian clustering models have been known to have different strengths and weaknesses that depend on the spatial genetic patterns present and on factors such as gene flow, dispersal distance and demography. *GENELAND* (Guillot *et al.*, 2005) has been demonstrated to be highly efficient when gene flow is low and genetic discontinuities correspond to simple shaped boundaries (Chen *et al.*, 2007; Safner *et al.*, 2011; Blair *et al.*, 2012), however it is sensitive to the level of genetic differentiation (Coulon *et al.*, 2006; Ball *et al.*, 2010) and its accuracy (Latch *et al.*, 2006; Chen *et al.*, 2007) and speed in analysing large datasets (Guillot, Mortier and Estoup, 2005) were criticised. Alternative tools like TESS and BAPS were shown to outperform *GENELAND* and each other under some scenarios, but not in others (Corander, Waldmann and Sillanpää, 2003; Guillot *et al.*, 2005; Latch *et al.*, 2006). Interestingly, Bayesian clustering models are known to overestimate genetic structure in the presence of IBD (Frantz *et al.*, 2009; Safner *et al.*, 2011), which highlights the importance of accounting for other types of structure in the data such as cryptic relatedness. Attention should be given to the priors used in Bayesian analyses and their effect on the final results (García-Pérez, 2019).

Local ancestry and haplotypes

Local ancestry and haplotype tools can be divided into four categories. Here, we will discuss the four most popular tools (Figure S2): *HAPMIX*, *ChromoPainter*, *LAMP*, *and LAMP-LD* (Table S2).

HAPMIX (Price *et al.*, 2009) is a popular approach that was limited to only two source populations and is unsuitable to biobanks. The biological parameters that *HAPMIX* requests (e.g., genetic maps, recombination and mutation rate, average ancestry coefficients, and the average number of generations since admixture) are typically unknown (Dias-Alves, Mairal and Blum, 2018). The more recent *MOSAIC* (Salter-Townshend and Myers, 2019), places an HMM over the haplotype estimation performed by *Chromopainter*, to learn how frequently haplotypes from different ancestries appear in unadmixed ancestries. It is therefore plausible to run at biobank scales in principle, though considerable effort would be required to report and use the per-SNP ancestry information returned.

LAMP and *LAMP-LD* work effectively with three-way admixture and gain a computational advantage by ascribing ancestry to pre-defined windows, though neither scales beyond hundreds of samples or tens of thousands of SNPs (Baran *et al.*, 2012) and are hence both inapplicable for biobanks (Dias-Alves, Mairal and Blum, 2018).

ChromoPainter is part of the *fineSTRUCTURE* pipeline (Lawson *et al.*, 2012) which allows the identification of fine-scale population structure, that cannot be identified by PCA or related approaches (Leslie *et al.*, 2015) Chromosome painting is applicable for samples up to thousands but cannot be used at biobank scale (Pan *et al.*, 2017) to examine variation within a sample. It can however be used to compare large biobank datasets to standardize references. There is an unpublished fast approximation in the PBWT package (Durbin, 2014) that can handle hundreds of thousands of samples for analysing within-sample variation.

These methods allow characterisation of LAI and gain power and resolution through analysis of haplotypes. One typical assumption is that admixture tract lengths are independent and exponentially differentiated; therefore, they are less effective when the admixture is strong because the admixture tracts are longer than expected under an exponential distribution (Schraiber and Akey, 2015). Further, many require phased data and is therefore susceptible to phasing errors.

Overall, the popular local ancestry tools are positioned along the extreme ends of limited models. On the one end, are mostly HMM-based that either do not consider LD or are limited to two or three reference populations. On the other hand, are more robust tools that aim to identify haplotypes, but their high memory consumption limits their usage. An additional limitation of local ancestry approach is the challenging evaluation of the results in follow up analyses. Local ancestry approach should be preferred when the loci or region of interest are known, however in an exploratory GWA or MR analyses it is unclear how to analyse a large number of segments associated with various ancestral populations. In this case, grouping the ancestral populations into geographical regions may be an appropriate compromise between accuracy and power considerations.

Sibship, Kinship, and Clanship

Relatedness inference tools exploit different statistical approaches in analysing IBD segments and identifying the correct level of relatedness. We will discuss the six most popular tools: *PLINK, KING, fastIBD, GERMLINE, PC-Relate and REAP* (Figure S3) (Table S3).

Kinship can be inferred by kinship coefficient estimation or identity by descent (IBD) detection. Kinship coefficient is a classic measurement of relatedness and can be defined as the probability that two homologous alleles drawn from each of two individuals are identical by descent (Thompson, 1975; Speed and Balding, 2015). Software that estimate the kinship coefficient often use relatedness estimators to calculate the kinship coefficient, which fall into two categories based on the method that they use: moment estimators used by *KING, REAP* and *PC-Relate*, and maximum likelihood (ML) estimators (Wang, 2002).

ML estimators mostly use an EM algorithm to estimate the *K* coefficients, whilst moment estimators, used by *KING*, *REAP* and *PC-Relate*, use the statistical method of moments to estimate the realised k coefficients; the proportion of genome at which two individuals share 0, 1, or 2 IBD genes (Wang, Sverdlov and Thompson, 2016). *KING* can produce reliable inference for large sample sizes (millions of unrelated and thousands of relative pairs) (Manichaikul *et al.*, 2010). However, *KING* is prone to biased estimates in admixed populations and in the presence of population structure, due to the violation of simplifying assumptions that do not hold in the presence of population structure and/or ancestry admixture (Conomos *et al.*, 2016). Conversely, *REAP* (Thornton *et al.*, 2012) and *PC-Relate* (Conomos *et al.*, 2016) are able to account for different ancestry background of admixed individuals by using individual-specific allele frequencies derived from model-based population structure analysis methods (e.g., *ADMIXTURE)*. Bias in these allele frequencies can lead to significantly biased relatedness estimates (Conomos *et al.*, 2016). Despite this, *PC-Relate* has an advantage over *KING* and *REAP*, because they, unlike *PC-Relate*, have difficulty separating unrelated individuals from more distantly related ones (Moltke and Albrechtsen, 2014). Both tools have relatively high accuracy for first through third degree classification; however, their accuracy decreased substantially to below 50% for fourth through seventh and unrelated classification (Table 4) (Ramstetter *et al.*, 2017). Overall, PC-Relate appears to be the most robust kinship coefficient estimation tool when compared with *KING* and *REAP*, due to its ability to work effectively with admixed populations, whilst also being able to distinguish between unrelated individuals and more distantly related ones.

Methods for IBD detection identify similarity between haplotypes that are statistically unlikely to occur in the absence of IBD sharing (Durand, Eriksson and Mclean, 2014). *PLINK* (Purcell *et al.*, 2007) incorporates a method of moments approach using an HMM to infer underlying IBD in chromosomal segments based on observed IBS states. *PLINK* was criticised for producing a high level of false positives (individuals who are unrelated based on IBS sharing, but are called as related) for second-degree relationships (Stevens *et al.*, 2011). *fastIBD* (Browning and Browning, 2011) and *GERMLINE* (Gusev *et al.*, 2009) detect “seeds” of identical short haplotype matches and extend them to nearby sites. *fastIBD* can be applied to large sample sizes across genome-wide SNP data; however, it is obliged to carry out haplotype phasing and is therefore susceptible to phasing errors, particularly if the SNP set is small. Computer memory capacities may also limit the number of individuals that can be phased at one time; therefore, in practicality, it is computationally unfeasible to analyse over than 100,000 individuals. Whereas *fastIBD* is based on shared haplotype frequency, GERMLINE is based on shared haplotype length.

Ramstetter *et al.* (2017) tested the accuracy of relationship inference software on SNP data of large Mexican American pedigrees spanning up to six generations. They showed that there are no “one size fits all” IBD tool and that tools vary in their sensitivity to the IBD segment length, which correspond to the degree of relatedness. The main reason for this is that haplotype based IBD segment detection methods struggle to detect long IBD segments if the shared haplotype has discordant alleles due to genotype or phasing error. One solution is to use tools like Refined IBD (Browning and Browning, 2013) and recover the long IBD segment by mending smaller ones using an external tool (Browning and Browning, 2013). Concurrent methods generally rely on diploid genotype data, which make them ineffective when dealing with ancient data which has a low concentration of endogenous DNA and fragmentation (Kuhn, Jakobsson and Günther, 2018). Since all tools underperformed in inferring remote relatedness (over 3rd degree) in diverse samples Ramstetter *et al.* (2017), further efforts should be made towards the development and testing of more robust tools.

DISCUSSION

The rise of genomic biobanks and biological and computational biotechnology advancements have allowed for significant developments in the field of personalised medicine, making the vision of targeted therapies, accelerated diagnosis and early disease detection become more of a reality. However, the geographic differentiation of human genetic variation (population genetic structure) suggests that the frequencies of certain disease-causing genetic variants and variants in drug-metabolising genes may differ depending on geographic location, leading to geographic disparities in the susceptibility of an individual to a disease and/or specific drug treatment. Therefore, the significant over-representation of Europeans in genomic studies currently limits the global understanding of disease risk and intervention efficacy.

It is widely accepted that increased samples from a much more diverse range of populations is required. However, diversity needs to be quantified, compared and annotated within and between biobanks in order to lead to insight. Biobanks must therefore contain genetic annotations that are comparable, computable and compatible across datasets. Whereas previous studies explored the applicability of bioinformatics tools for association studies (e.g., Hellwege *et al.*, 2017), this review focussed on whether tools are conceptually comparable, and whether they scale. It therefore assesses the confounding effects of stratification bias through the identification of population genetic structure in a standardised and comparable manner, with a view to improving biobanks, increasing the accuracy of association analyses, and informing developmental efforts. These tools vary in their strengths and limitations; therefore, it is vital to review these characteristics in order to apply them appropriately.

Genomic ancestry inference encompasses tools that are able to identify the ancestry of an individual, by utilising specialised markers to compare the genetic similarity of an individual's DNA to other individuals sampled from a variety of populations or geographic regions. Global genetic ancestry tools assess the average proportion contributed by each ancestral population across the whole of the individual's genome, whilst local ancestry inference tools identify the ancestry of distinct segments within chromosomes.

Simple descriptions such as Ancestry Informative Markers (AIMs) and Low Dimensional Representation with PCA are useful but insufficient. Current best-practice includes correcting for kinship using the Genetic Relatedness Matrix, which may be valuable but does not provide a framework for interpreting external datasets.

For global genetic ancestry inference, some tools do scale well enough to be considered for biobanks. The limitations include unrealistic assumptions, a tendency to mistake cryptic relatedness for genetic structure, conceptual issues in the interpretation of admixture, and a lack of prior research into how global ancestry can be usefully applied for association studies.

GRM approaches can jointly represent population genetic structure and cryptic relatedness, which can avoid consequent false positive associations in GWAS, within a single dataset. More fine-scaled representations exist in the form of kinship (measuring IBD) and haplotype similarity (Chromosome Painting) matrices, which are scalable. In all cases these capture an inherently noisy and hence statistical property. Consequently, further efforts should be made towards the development of more robust tools for remote degrees of relatedness (over 3rd degree) in diverse samples, especially in the case of cross-dataset comparison. Studies are needed explore the value of these fine-scale approaches for biobanks.

Local ancestry inference tools are still slower, though can be deployed similarly to phasing and imputation should a compelling use case be found. Efforts should be made to develop a new approach that addresses the common limitations, including a requirement of phased data and consequent susceptibility to phasing errors, ability to model LD, and restrictions in terms of number of populations. The local ancestry approach is clearly deployable when the region of interest is known. It may be useful to group the ancestral

doi:10.20944/preprints202004.0394.v1

populations by geographic region as a way of compromising between accuracy and power considerations. Correctly deployed, local ancestry could correct for local genetic correlations in a way that is much more powerful than simple correlations as captured by LD, though the value for association studies is yet to be determined.

Furthermore, the rise of paleogenomic medicine and rapid accumulation of ancient genomes have already shed light on several conditions (e.g., (Cassidy *et al.*, 2016) also requires the development of specialised kinship inference software that are capable of handling ancient DNA. At the moment, however most current methods rely on diploid genotype data making them ineffective when handling ancient DNA.

With rapid advances in technology and the dense amount of genetic variation data available, we can continue to expect development of new inference software and enhancements of existing ones. For example, there is much scope for improved modelling of LD to reduce error rates and improve ability to detect subtle population structure. However, a challenge for the future will be to develop inference methods that are computationally efficient and applicable to large sample sizes, whilst being able to fully exploit the rich information available in the form of haplotypes. There is also currently a lack of representation of non-European populations in genetic studies, and unless populations of diverse ancestries are included therefore incorporating a more equal knowledge of genetic variation across ancestry groups, it could contribute further to health disparities and negatively impact genomic interpretation. Efforts should therefore be made to include data from more diverse populations in GWAS and develop robust population structure models that can reduce or eliminate the *stratification bias* from the cohort. Not only this, but biobanks must begin to incorporate individual level genetic annotations that are comparable, computable, and compatible across datasets. Clinicians must also be properly trained to understand their output so that they can make an informed decision as to whether or not a genetic variant may be causative or whether the association is likely the result of population stratification.

Overall, with increased availability of large genomic datasets, a more equal representation of genetic variation across ancestry groups, continued improvement and development of genetic inference software, population structure inference will occur with finite detail and more effectively distinguish between closely related populations, in turn allowing for individual-level genetic annotations to be incorporated in biobanks and increasing the accuracy of association analyses.

In summary, we have identified a gap in the literature concerning the design and standardization of biobanks. Started as localized initiatives, the progress in sequencing technologies sparred the rapid growth of biobanks in size, diversity, and geography, although conceptually they are still thought of as local datasets. This perception limits the usefulness of biobanks and prevents banking on their resources in joint analyses. To overcome this limitation, it is critical to develop a holistic solution to the problem of population structure. Current strategies implemented in the various tools aim to expose different aspects of the data by ubiquitously mapping the ancestry of individuals, though none could be used as a complete solution to ancestry. One unsolved challenge is to create representations that are useful for meta-analysis without sharing individual-level data. Natural summaries of PCA or admixtures can be created from means and variances, but it is an open question to establish whether these are sufficiently accurate and whether alternative representations can protect privacy whilst maximising research benefits. PCA's accuracy, in specific, has been challenged by several groups. Yet other tools also suffer from limitations related either to their design, which affect their speed and accuracy or their basic assumptions concerning human populations, which, in turn, affect the usefulness of their output to the population genetics. These shortcomings, often unacknowledged, limit our ability to interpret the results and increase the burden of evidence when using these tools. Further efforts should be made to explore the limitations of these tools and optimal usage on global and massive datasets as well as to divide new approaches that overcome the most common limitations of running time, identification of admixture, and high specificity among human populations.

We end this review with five take-away messages: Firstly, more diverse data are needed, worldwide, both from populous populations who will benefit from inclusion in datasets *en masse* as well as pockets of genetic diversity that may shed light on biological processes that would otherwise remain undiscovered. Secondly, the methodology to interpret and harmonise results from diverse datasets is not ready. Thirdly, the main barrier is in the creation of sharable and comparable summary statistics from diverse data. Fourthly, these summary datasets should be carefully designed to allow effective association correction, as well as meta-analysis, which we argue requires placing the genetics into some type of model. Finally, clinicians, geneticists and epidemiological researchers will all have to learn how to exploit the information that comes from the genomic diversity revolution, when it comes.

As this review is written at the heat of the COVID-19 pandemic and biobank data are internationally shared to improve diagnosis and treatment outcome, practically transforming our vision of international biobanks into reality, we hope that the our study would serve to improve the accuracy, reliability, and replicability of association studies and biobanks.

COMPETING INTERESTS

EE is consultant to DNA Diagnostic Centre.

AUTHORS' CONTRIBUTIONS

EE initiated the study. HC did the research. HC, DL and EE wrote the paper.

ACKNOWLEDGEMENTS

This work was partially supported by the MRC (MR/R025126/1) to EE.

REFERENCES

Adeyemo, A. and Rotimi, C. (2009) 'Genetic variants associated with complex human diseases show wide variation across multiple populations', *Public Health Genomics*, 13(2), pp. 72–79. doi: 10.1159/000218711.

Alexander, D. H. and Lange, K. (2011) 'Enhancements to the ADMIXTURE algorithm for individual ancestry estimation', *BMC Bioinformatics*, 12. doi: 10.1186/1471-2105-12-246.
Alexander, D. H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals.', *Genome research*, 19(9), pp. 1655–1664. doi: 10.1101/gr.094052.109.

Altshuler, D. M. *et al.* (2010) 'Integrating common and rare genetic variation in diverse human populations', *Nature*, 467(7311), pp. 52–58. doi: 10.1038/nature09298.

Altshuler, D. M. *et al.* (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491, pp. 56–65. doi: 10.1038/nature11632.

Arenas, M. *et al.* (2013) 'Influence of admixture and paleolithic range contractions on current European diversity gradients', *Molecular Biology and Evolution*, 30(1), pp. 57–61. doi: 10.1093/molbev/mss203.

Van der Auwera, G. A. *et al.* (2013) 'From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline', *Current Protocols in Bioinformatics*, (SUPL.43). doi: 10.1002/0471250953.bi1110s43.

Ball, M. C. *et al.* (2010) 'Integrating multiple analytical approaches to spatially delineate and characterize genetic population structure: An application to boreal caribou (Rangifer tarandus caribou) in central Canada', *Conservation Genetics*, 11(6), pp. 2131–2143. doi: 10.1007/s10592-010-0099-3.

Baran, Y. *et al.* (2012) 'Fast and accurate inference of local ancestry in Latino populations', *Bioinformatics*, 28(10), pp. 1359–1367. doi: 10.1093/bioinformatics/bts144.

De Barros Damgaard, P. *et al.* (2018) '137 ancient human genomes from across the Eurasian steppes', *Nature*, 557(7705), pp. 369–374. doi: 10.1038/s41586-018-0094-2.

Baughn, L. B. *et al.* (2018) 'Differences in genomic abnormalities among African individuals with monoclonal gammopathies using calculated ancestry', *Blood Cancer Journal*, 8(10). doi: 10.1038/s41408-018-0132-1.

Baughn, L. B. *et al.* (2020) 'The CCND1 c.870G risk allele is enriched in individuals of African ancestry with plasma cell dyscrasias', *Blood Cancer Journal*, 10(3). doi: 10.1038/s41408-020-0294-5.

Belmont, J. W. *et al.* (2005) 'A haplotype map of the human genome', *Nature*, 437(7063), pp. 1299–1320. doi: 10.1038/nature04226.

Berg, J. J. *et al.* (2019) 'Reduced signal for polygenic adaptation of height in UK biobank', *eLife*, 8. doi: 10.7554/eLife.39725.

doi:10.20944/preprints202004.0394.v1

Bergholdt, H. K. M., Nordestgaard, B. G. and Ellervik, C. (2015) 'Milk intake is not associated with low risk of diabetes or overweight-obesity: A Mendelian randomization study in 97,811 Danish individuals', *American Journal of Clinical Nutrition*, 102(2), pp. 487–496. doi: 10.3945/ajcn.114.105049.

Blair, C. *et al.* (2012) 'A simulation-based evaluation of methods for inferring linear barriers to gene flow', *Molecular Ecology Resources*, 12(5), pp. 822–833. doi: 10.1111/j.1755-0998.2012.03151.x.

Border, R. *et al.* (2019) 'No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples', *American Journal of Psychiatry*, 176(5), pp. 376–387. doi: 10.1176/appi.ajp.2018.18070881.

Brieger, K. *et al.* (2019) 'Genes for Good: Engaging the Public in Genetics Research via Social Media', *American Journal of Human Genetics*, 105(1), pp. 65–77. doi: 10.1016/j.ajhg.2019.05.006.

Browning, B. L. and Browning, S. R. (2011) 'A fast, powerful method for detecting identity by descent', *American Journal of Human Genetics*, 88(2), pp. 173–182. doi: 10.1016/j.ajhg.2011.01.010.

Browning, B. L. and Browning, S. R. (2013) 'Improving the accuracy and efficiency of identity-by-descent detection in population data', *Genetics*, 194(2), pp. 459–471. doi: 10.1534/genetics.113.150029.

Bulik-Sullivan, B. *et al.* (2015) 'LD score regression distinguishes confounding from polygenicity in genome-wide association studies', *Nature Genetics*, 47(3), pp. 291–295. doi: 10.1038/ng.3211.

Burgess, S. and Thompson, S. G. (2015) *Mendelian randomization: Methods for using genetic variants in causal estimation*, *CRC Press*. doi: 10.1201/b18084.

Bycroft, C. *et al.* (2018) 'The UK Biobank resource with deep phenotyping and genomic data', *Nature*, 562(7726), pp. 203–209. doi: 10.1038/s41586-018-0579-z.

Byun, J. *et al.* (2017) 'Ancestry inference using principal component analysis and spatial analysis: A distance-based analysis to account for population substructure', *BMC Genomics*, 18(789), pp. 1–12. doi: 10.1186/s12864-017-4166-8.

Campbell, C. D. *et al.* (2005) 'Demonstrating stratification in a European American population', *Nature Genetics*, 37(8), pp. 868–872. doi: 10.1038/ng1607.

Cassidy, L. M. *et al.* (2016) 'Neolithic and Bronze Age migration to Ireland and establishment of the insular atlantic genome', *Proceedings of the National Academy of Sciences of the United States of America*, 113(2), pp. 368–373. doi: 10.1073/pnas.1518445113.

Caulfield, M. *et al.* (2017) *The 100,000 Genomes Project Protocol*, *Genomics England*. doi: 10.6084/M9.FIGSHARE.4530893.V2.

Chacón-Duque, J. C. *et al.* (2018) 'Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance', *Nature Communications*, 9(1). doi: 10.1038/s41467-018-07748-z.

Chalmers, D. *et al.* (2016) 'Has the biobank bubble burst? Withstanding the challenges for sustainable biobanking in the digital era Donna Dickenson, Sandra Soo-Jin Lee, and Michael Morrison', *BMC Medical Ethics*, 17(1). doi: 10.1186/s12910-016-0124-2.

Chang, C. C. *et al.* (2015) 'Second-generation PLINK: Rising to the challenge of larger and richer datasets', *GigaScience*, 4(1). doi: 10.1186/s13742-015-0047-8.

Chen, C. *et al.* (2007) 'Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study', *Molecular Ecology Notes*, 7(5). doi: 10.1111/j.1471-8286.2007.01769.x.

Chikhi, L. *et al.* (2010) 'The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes', *Genetics*, 186(3), pp. 983–995. doi: 10.1534/genetics.110.118661.

Clark, A. G. *et al.* (2005) 'Ascertainment bias in studies of human genome-wide polymorphism', *Genome Research*, 15(11), pp. 1496–1502. doi: 10.1101/gr.4107905.

Conomos, M. P. *et al.* (2016) 'Model-free Estimation of Recent Genetic Relatedness', *American Journal of Human Genetics*, 98(1), pp. 127–148. doi: 10.1016/j.ajhg.2015.11.022.

Conomos, M. P. *et al.* (2018) 'Genome-Wide Control of Population Structure and Relatedness in Genetic Association Studies via Linear Mixed Models with Orthogonally Partitioned Structure', *bioRxiv*, p. 409953. doi: 10.1101/409953.

Corander, J., Waldmann, P. and Sillanpää, M. J. (2003) 'Bayesian analysis of genetic differentiation between populations.', *Genetics*, 163(1), pp. 367–374.

Coulon, A. *et al.* (2006) 'Genetic structure is influenced by landscape features: Empirical evidence from a roe deer population', *Molecular Ecology*, 15(6), pp. 1669–1679. doi: 10.1111/j.1365-294X.2006.02861.x.

Cruz-Correa, M. *et al.* (2017) 'Clinical Cancer Genetics Disparities among Latinos', *Journal of Genetic Counseling*, pp. 379–386. doi: 10.1007/s10897-016-0051-x.

Daar, A. S. and Singer, P. A. (2005) 'Pharmacogenetics and geographical ancestry: Implications for drug development and global health', *Nature Reviews Genetics*, pp. 241–246. doi: 10.1038/nrg1559.

Das, R. *et al.* (2016) 'Localizing Ashkenazic Jews to primeval villages in the ancient Iranian lands of Ashkenaz', *Genome Biology and Evolution*, 8(4), pp. 1132–1149. doi: 10.1093/gbe/evw046.

Das, R. *et al.* (2017) 'The origins of Ashkenaz, Ashkenazic Jews, and Yiddish', *Frontiers in Genetics*, 8(JUN). doi: 10.3389/fgene.2017.00087.

doi:10.20944/preprints202004.0394.v1

Davey-Smith, G. D. and Ebrahim, S. (2003) '"Mendelian randomization": Can genetic epidemiology contribute to understanding environmental determinants of disease?', *International Journal of Epidemiology*, 32(1), pp. 1–22. doi: 10.1093/ije/dyg070.

Dias-Alves, T., Mairal, J. and Blum, M. G. B. (2018) 'Loter: A software package to infer local ancestry for a wide range of species', *Molecular Biology and Evolution*, 35(9), pp. 2318–2326. doi: 10.1093/molbev/msy126.

Dubow, T. and Marjanovic, S. (2016) *Population-scale sequencing and the future of genomic medicine: Learning from past and present efforts*, *Population-scale sequencing and the future of genomic medicine: Learning from past and present efforts*. doi: 10.7249/rr1520.

Durand, E. *et al.* (2009) 'Spatial inference of admixture proportions and secondary contact zones', *Molecular Biology and Evolution*, 26(9), pp. 1963–1973. doi: 10.1093/molbev/msp106.

Durand, E. Y., Eriksson, N. and Mclean, C. Y. (2014) 'Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis', *Molecular Biology and Evolution*, 31(8), pp. 2212–2222. doi: 10.1093/molbev/msu151.

Durbin, R. (2014) 'Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)', *Bioinformatics*, 30(9), pp. 1266–1272. doi: 10.1093/bioinformatics/btu014.

Dyer, C. (2020) 'Covid-19: Rules on sharing confidential patient information are relaxed in England', *Bmj*, p. m1378. doi: 10.1136/bmj.m1378.

Elhaik, E. (2012) 'Empirical Distributions of FST from Large-Scale Human Polymorphism Data', *PLoS ONE*, 7(11). doi: 10.1371/journal.pone.0049837.

Elhaik, E. *et al.* (2013) 'The GenoChip: A new tool for genetic anthropology', *Genome Biology and Evolution*, 5(5), pp. 1021–1031. doi: 10.1093/gbe/evt066.

Elhaik, E. *et al.* (2014) 'Geographic population structure analysis of worldwide human populations infers their biogeographical origins', *Nature Communications*, 5. doi: 10.1038/ncomms4513.

Elhaik, E. and Ryan, D. M. (2019) 'Pair Matcher (PaM): fast model-based optimization of treatment/case-control matches', *Bioinformatics*, 35(13), pp. 2243–2250. doi: 10.1093/bioinformatics/bty946.

Esposito, U. *et al.* (2018) 'Ancient ancestry informative markers for identifying fine-scale ancient population structure in eurasians', *Genes*, 9(12). doi: 10.3390/genes9120625.

Falush, D., Stephens, M. and Pritchard, J. K. (2003) 'Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.', *Genetics*, 164(4), pp. 1567–1587.

Feero, W. G., Guttmacher, A. E. and Collins, F. S. (2010) 'Genomic medicine - An updated primer', *New England Journal of Medicine*. doi: 10.1056/NEJMra0907175.

Finan, C. *et al.* (2017) 'The druggable genome and support for target identification and validation in drug development', *Science Translational Medicine*, 9(383). doi: 10.1126/scitranslmed.aag1166.

Flegontov, P. *et al.* (2016) 'Genomic study of the Ket: A Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry', *Scientific Reports*. doi: 10.1038/srep20768.

Frantz, A. C. *et al.* (2009) 'Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance?', *Journal of Applied Ecology*, 46(2), pp. 493–505. doi: 10.1111/j.1365-2664.2008.01606.x.

Galinsky, K. J. *et al.* (2016) 'Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure', *American Journal of Human Genetics*, 99(5), pp. 1130–1139. doi: 10.1016/j.ajhg.2016.09.014.

García-Pérez, M. Á. (2019) 'Bayesian Estimation with Informative Priors is Indistinguishable from Data Falsification', *The Spanish journal of psychology*, 22. doi: 10.1017/sjp.2019.41.

Generation Scotland (2016) 'Generation Scotland : Facts and Figures', (July).

Gopalan, P. *et al.* (2016) 'Scaling probabilistic models of genetic variation to millions of humans', *Nature Genetics*, 48(12), pp. 1587–1590. doi: 10.1038/ng.3710.

Guillot, G. *et al.* (2005) 'A spatial statistical model for landscape genetics', *Genetics*, 170(3), pp. 1261–1280. doi: 10.1534/genetics.104.033803.

Guillot, G., Mortier, F. and Estoup, A. (2005) 'GENELAND: A computer package for landscape genetics', *Molecular Ecology Notes*, 5(2), pp. 712–715. doi: 10.1111/j.1471-8286.2005.01031.x.

Gusev, A. *et al.* (2009) 'Whole population, genome-wide mapping of hidden relatedness', *Genome Research*, 19, pp. 318–326. doi: 10.1101/gr.081398.108.

Guttmacher, A. E. and Collins, F. S. (2002) 'Genomic medicine - A primer', *New England Journal of Medicine*, 347(19), pp. 1512–1520. doi: 10.1056/NEJMra012240.

Hayeck, T. J. *et al.* (2015) 'Mixed model with correction for case-control ascertainment increases association power', *American Journal of Human Genetics*, 96(5), pp. 720–730. doi: 10.1016/j.ajhg.2015.03.004.

Hellenthal, G. *et al.* (2014) 'A genetic atlas of human admixture history', *Science*, 343(6172), pp. 747–751. doi: 10.1126/science.1243518.

Hellwege, J. N. *et al.* (2017) 'Population Stratification in Genetic Association Studies', *Current protocols in human genetics*, 95, pp. 1.22.1-1.22.23. doi: 10.1002/cphg.48.

Hemani, G. *et al.* (2016) 'MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations', *bioRxiv*, p. 078972. doi: 10.1101/078972.

Hindorff, L. A. *et al.* (2018) 'Prioritizing diversity in human genomics research', *Nature*

Reviews Genetics, 19(3), pp. 175–185. doi: 10.1038/nrg.2017.89.

Ikegawa, S. (2012) 'A Short History of the Genome-Wide Association Study: Where We Were and Where We Are Going', *Genomics & Informatics*, 10(4), p. 220. doi: 10.5808/gi.2012.10.4.220.

Illumina (2013) *Illumina Microarray Solutions, 370-2013-003*. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/applications/genotyping/Microarray_Solutions.pdf.

Ioannidis, J. P. A., Ntzani, E. E. and Trikalinos, T. A. (2004) '"Racial" differences in genetic effects for complex diseases', *Nature Genetics*, 36(12), pp. 1312–1318. doi: 10.1038/ng1474.

Jakobsson, M. *et al.* (2008) 'Genotype, haplotype and copy-number variation in worldwide human populations.', *Nature*, 451(7181), pp. 998–1003. doi: 10.1038/nature06742.

Jiang, D. and Wang, M. (2018) 'Recent developments in statistical methods for gwas and high-throughput sequencing association studies of complex traits', *Biostatistics and Epidemiology*, 2(1), pp. 132–159. doi: 10.1080/24709360.2018.1529346.

Johnson, S. B. *et al.* (2019) 'Rethinking the ethical principles of genomic medicine services', *European Journal of Human Genetics*. doi: 10.1038/s41431-019-0507-1.

Jurczak, K. (2017) *Ethnic groups and nationalities in Iceland*, *WorldAtlas*.

Kaiser, J. (2002) 'Population databases boom, From Iceland to the U.S.', *Science*, 298(5596), pp. 1158–1161. doi: 10.1126/science.298.5596.1158.

Kalinowski, S. T. (2011) 'The computer program STRUCTURE does not reliably identify the main genetic clusters within species: Simulations and implications for human population structure', *Heredity*, 106(4), pp. 625–632. doi: 10.1038/hdy.2010.95.

Kamm, J. *et al.* (2019) 'Efficiently Inferring the Demographic History of Many Populations With Allele Count Data', *Journal of the American Statistical Association*, pp. 1–16. doi: 10.1080/01621459.2019.1635482.

Karczewski, K. J. *et al.* (2019) 'Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes', *bioRxiv*, p. 531210. doi: 10.1101/531210.

Katan, M. B. (2004) 'Apolipoprotein E isoforms, serum cholesterol, and cancer', *International Journal of Epidemiology*, 33(1), p. 9. doi: 10.1093/ije/dyh312.

Khan, S. S., Cooper, R. and Greenland, P. (2020) 'Do Polygenic Risk Scores Improve Patient Selection for Prevention of Coronary Artery Disease?', *JAMA - Journal of the American Medical Association*, 323(7), pp. 614–615. doi: 10.1001/jama.2019.21667.

Koellinger, P. D. and De Vlaming, R. (2019) 'Mendelian randomization: The challenge of unobserved environmental confounds', *International Journal of Epidemiology*, 48(3), pp. 665–671. doi: 10.1093/ije/dyz138.

Kuhn, J. M. M., Jakobsson, M. and Günther, T. (2018) 'Estimating genetic kin relationships in prehistoric populations', *PLoS ONE*, 13(4). doi: 10.1371/journal.pone.0195491.

Latch, E. K. *et al.* (2006) 'Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation', *Conservation Genetics*, 7, pp. 295–302. doi: 10.1007/s10592-005-9098-1.

Lawson, D. J. *et al.* (2012) 'Inference of population structure using dense haplotype data', *PLoS Genetics*, 8(1). doi: 10.1371/journal.pgen.1002453.

Lawson, D. J. *et al.* (2020) 'Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity?', *Human Genetics*, 139(1), pp. 23–41. doi: 10.1007/s00439-019-02014-8.

Lawson, D. J., van Dorp, L. and Falush, D. (2018) 'A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots', *Nature Communications*, 9(1). doi: 10.1038/s41467-018-05257-7.

Lazaridis, I. *et al.* (2014) 'Ancient human genomes suggest three ancestral populations for present-day Europeans', *Nature*, 513(7518), pp. 409–413. doi: 10.1038/nature13673.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536(7616), pp. 285–291. doi: 10.1038/nature19057.

Lesko, L. J. and Woodcock, J. (2004) 'Translation of pharmacogenomics and pharmacogenetics: A regulatory perspective', *Nature Reviews Drug Discovery*, 3(9), pp. 763–769. doi: 10.1038/nrd1499.

Leslie, S. *et al.* (2015) 'The fine-scale genetic structure of the British population', *Nature*, 519(7543), pp. 309–314. doi: 10.1038/nature14230.

Lewis, L. D. (2005) 'Personalized drug therapy; the genome, the chip and the physician', *British Journal of Clinical Pharmacology*, 60(1), pp. 1–4. doi: 10.1111/j.1365-2125.2005.02457.x.

Li, Q. and Yu, K. (2008) 'Improved correction for population stratification in genome-wide association studies by identifying hidden population structures', *Genetic Epidemiology*, 32(3), pp. 215–226. doi: 10.1002/gepi.20296.

Li, Z. *et al.* (2015) 'Loci with genome-wide associations with schizophrenia in the Han Chinese population', *British Journal of Psychiatry*, 207(6), pp. 490–494. doi: 10.1192/bjp.bp.114.150490.

Lippman, S. M. *et al.* (2009) 'Effect of selenium and vitamin E on risk of prostate cancer and other cancers: The selenium and vitamin E cancer prevention trial (SELECT)', *JAMA - Journal of the American Medical Association*, 301(1), pp. 39–51. doi: 10.1001/jama.2008.864.

Locke, A. E. *et al.* (2015) 'Genetic studies of body mass index yield new insights for obesity biology', *Nature*, 518(7538), pp. 197–206. doi: 10.1038/nature14177.

Loh, P. R. *et al.* (2015) 'Efficient Bayesian mixed-model analysis increases association power in large cohorts', *Nature Genetics*, 47(3), pp. 284–290. doi: 10.1038/ng.3190.

Loh, P. R. *et al.* (2018) 'Mixed-model association for biobank-scale datasets', *Nature Genetics*, 50(7), pp. 906–908. doi: 10.1038/s41588-018-0144-6.

Lombaert, E., Guillemaud, T. and Deleury, E. (2018) 'Biases of STRUCTURE software when exploring introduction routes of invasive species', *Heredity*, 120(6), pp. 485–499. doi: 10.1038/s41437-017-0042-1.

Manichaikul, A. *et al.* (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26(22), pp. 2867–2873. doi: 10.1093/bioinformatics/btq559.

Manolio, T. A. *et al.* (2013) 'Implementing genomic medicine in the clinic: The future is here', *Genetics in Medicine*, 15, pp. 258–267. doi: 10.1038/gim.2012.157.

Marshall, S. *et al.* (2016) 'Reconstructing Druze population history', *Scientific Reports*, 6. doi: 10.1038/srep35837.

McVean, G. (2009) 'A genealogical interpretation of principal components analysis', *PLoS Genetics*, 5(10). doi: 10.1371/journal.pgen.1000686.

Mills, M. C. and Rahal, C. (2019) 'A scientometric review of genome-wide association studies', *Communications Biology*, 2(1). doi: 10.1038/s42003-018-0261-x.

Mokry, L. E. *et al.* (2015) 'Mendelian randomisation applied to drug development in cardiovascular disease: A review', *Journal of Medical Genetics*, 52(2), pp. 71–79. doi: 10.1136/jmedgenet-2014-102438.

Moltke, I. and Albrechtsen, A. (2014) 'RelateAdmix: A software tool for estimating relatedness between admixed individuals', *Bioinformatics*, 30(7), pp. 1027–1028. doi: 10.1093/bioinformatics/btt652.

Mountain, J. L. and Risch, N. (2004) 'Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups', *Nature Genetics*, 36. doi: 10.1038/ng1456.

Nalls, M. A. *et al.* (2014) 'Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease', *Nature Genetics*, 46(9), pp. 989–993. doi: 10.1038/ng.3043.

Need, A. C. and Goldstein, D. B. (2009) 'Next generation disparities in human genomics: concerns and remedies', *Trends in Genetics*, 25(11), pp. 489–494. doi: 10.1016/j.tig.2009.09.012.

NHS (2016) *Improving Outcomes Through Personalised Medicine.*, *NHS England*.

Novembre, J. *et al.* (2008) 'Genes mirror geography within Europe', *Nature*, 456(7218), pp. 98–101. doi: 10.1038/nature07331.

Novembre, J. and Stephens, M. (2008) 'Interpreting principal component analyses of spatial population genetic variation', *Nature Genetics*, 40(5), pp. 646–649. doi: 10.1038/ng.139.

Oleksyk, T. K., Brukhin, V. and O'Brien, S. J. (2015) 'The Genome Russia project: Closing the largest remaining omission on the world Genome map', *GigaScience*, 4(1), p. 53. doi: 10.1186/s13742-015-0095-0.

Ooi, B. N. S. *et al.* (2019) 'The genetic interplay between body mass index, breast size and breast cancer risk: a Mendelian randomization analysis', *International Journal of Epidemiology*, 48(3), pp. 781–794. doi: 10.1093/ije/dyz124.

Ortega, V. E. and Meyers, D. A. (2014) 'Pharmacogenetics: Implications of race and ethnicity on defining genetic profiles for personalized medicine', *Journal of Allergy and Clinical Immunology*, 133(1), pp. 16–26. doi: 10.1016/j.jaci.2013.10.040.

Pagani, L. *et al.* (2016) 'Genomic analyses inform on migration events during the peopling of Eurasia', *Nature*, 538(7624), pp. 238–242. doi: 10.1038/nature19792.

Palmer, C. and Pe'er, I. (2017) 'Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies', *PLoS Genetics*, 13(7). doi: 10.1371/journal.pgen.1006916.

Pasic, M. D., Samaan, S. and Yousef, G. M. (2013) 'Genomic medicine: New frontiers and new challenges', *Clinical Chemistry*, 59(1), pp. 158–167. doi: 10.1373/clinchem.2012.184622.

Petrovski, S. and Goldstein, D. B. (2016) 'Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine', *Genome Biology*, 17(1), p. 157. doi: 10.1186/s13059-016-1016-y.

Pickrell, J. K. *et al.* (2016) 'Detection and interpretation of shared genetic influences on 42 human traits', *Nature Genetics*, 48(7), pp. 709–717. doi: 10.1038/ng.3570.

Popejoy, A. B. and Fullerton, S. M. (2016) 'Genomics is failing on diversity', *Nature*, 538(7624), pp. 161–164. doi: 10.1038/538161a.

Price, A. L. *et al.* (2006) 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics*, 38, pp. 904–909. doi: 10.1038/ng1847.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) 'Inference of population structure using multilocus genotype data', *Genetics*, 55(2), pp. 945–959.

Purcell, S. (2002) 'Variance components models for gene-environment interaction in twin analysis', *Twin Research*, 5(6), pp. 554–571. doi: 10.1375/136905202762342026.

Purcell, S. *et al.* (2007) 'PLINK: a tool set for whole-genome association and population-based linkage analyses.', *American journal of human genetics*, 81(3), pp. 559–75. doi: 10.1086/519795.

Raj, A., Stephens, M. and Pritchard, J. K. (2014) 'FastSTRUCTURE: Variational inference of population structure in large SNP data sets', *Genetics*, 197(2), pp. 573–589. doi: 10.1534/genetics.114.164350.

Ramachandran, S. *et al.* (2005) 'Support from the relationship of genetic and geographic in human populations for a serial founder effect originating in Africa', *Proceedings of the National Academy of Sciences of the United States of America*, 102(44), pp. 15942–15947. doi: 10.1073/pnas.0507611102.

Ramstetter, M. D. *et al.* (2017) 'Benchmarking relatedness inference methods with genome-wide data from thousands of relatives', *Genetics*, 207(1), pp. 75–82. doi: 10.1534/genetics.117.1122.

Rask-Andersen, M. *et al.* (2017) 'Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status', *PLoS Genetics*, 13(9). doi: 10.1371/journal.pgen.1006977.

Rosenberg, N. A. *et al.* (2002) 'Genetic structure of human populations', *Science*, 298(5602), pp. 2381–2385. doi: 10.1126/science.1078311.

Roy, A. S. a (2012) 'STIFLING NEW CURES: The True Cost of Lengthy Clinical Drug Trials', *Project FDA Report*, pp. 1–11.

Safner, T. *et al.* (2011) 'Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics', *International Journal of Molecular Sciences*, 12(2), pp. 865–889. doi: 10.3390/ijms12020865.

Salter-Townshend, M. and Myers, S. (2019) 'Fine-scale inference of ancestry segments without prior knowledge of admixing groups', *Genetics*, 212(3), pp. 869–889. doi: 10.1534/genetics.119.302139.

Schärfe, C. P. I. *et al.* (2017) 'Genetic variation in human drug-related genes', *Genome Medicine*, 9(1). doi: 10.1186/s13073-017-0502-5.

Scheinfeldt, L. B. *et al.* (2016) 'Challenges in translating GWAS results to clinical care', *International Journal of Molecular Sciences*, 17(8). doi: 10.3390/ijms17081267.

Schraiber, J. G. and Akey, J. M. (2015) 'Methods and models for unravelling human evolutionary history', *Nature Reviews Genetics*, 16, pp. 727–740. doi: 10.1038/nrg4005.

Scudellari, M. (2013) 'Biobank managers bemoan underuse of collected samples', *Nature Medicine*. doi: 10.1038/nm0313-253a.

Shi, Y. *et al.* (2020) 'COVID-19 infection: the perspectives on immune responses', *Cell Death & Differentiation*. doi: 10.1038/s41418-020-0530-3.

Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) 'The Missing Diversity in Human Genetic Studies', *Cell*, 177(1), pp. 26–31. doi: 10.1016/j.cell.2019.02.048.

Smith, B. H. *et al.* (2006) 'Generation Scotland: The Scottish Family Health Study; a new resource for researching genes and heritability', *BMC Medical Genetics*, 7. doi: 10.1186/1471-2350-7-74.

Smith, G. D. (2010) 'Mendelian randomization for strengthening causal inference in observational studies: Application to gene × environment interactions', *Perspectives on Psychological Science*, 5(5), pp. 527–545. doi: 10.1177/1745691610383505.

Sohail, M. *et al.* (2019) 'Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies', *eLife*, 8. doi: 10.7554/eLife.39702.

Somiari, S. B. and Somiari, R. I. (2015) 'The future of biobanking: A conceptual look at how biobanks can respond to the growing human biospecimen needs of researchers', in *Advances in Experimental Medicine and Biology*, pp. 11–27. doi: 10.1007/978-3-319-20579-3_2.

Speed, D. and Balding, D. J. (2015) 'Relatedness in the post-genomic era: Is it still useful?', *Nature Reviews Genetics*, 16(1), pp. 33–44. doi: 10.1038/nrg3821.

Stevens, E. L. *et al.* (2011) 'Inference of relationships in population data using identity-by-descent and identity-by-state', *PLoS Genetics*, 7(9). doi: 10.1371/journal.pgen.1002287.

Sudlow, C. *et al.* (2015) 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLoS Medicine*, 12(3). doi: 10.1371/journal.pmed.1001779.

Tam, V. *et al.* (2019) 'Benefits and limitations of genome-wide association studies', *Nature Reviews Genetics*, 20(8), pp. 467–484. doi: 10.1038/s41576-019-0127-1.

Tang, H. *et al.* (2005) 'Estimation of individual admixture: Analytical and study design considerations', *Genetic Epidemiology*, 28(4), pp. 289–301. doi: 10.1002/gepi.20064.

Thompson, E. A. (1975) 'The estimation of pairwise relationships', *Annals of Human Genetics*, 39(2), pp. 173–188. doi: 10.1111/j.1469-1809.1975.tb00120.x.

Thornton, T. *et al.* (2012) 'Estimating kinship in admixed populations', *American Journal of Human Genetics*, 91(1), pp. 122–138. doi: 10.1016/j.ajhg.2012.05.024.

Timpson, N. J. *et al.* (2018) 'Genetic architecture: The shape of the genetic contribution to human traits and disease', *Nature Reviews Genetics*, 19(2), pp. 110–124. doi: 10.1038/nrg.2017.101.

Tutton, R. (2009) *Race/Ethnicity: Multidisciplinary Global Contexts*.
Visscher, P. M. *et al.* (2012) 'Five years of GWAS discovery', *American Journal of Human Genetics*, 90(1), pp. 7–24. doi: 10.1016/j.ajhg.2011.11.029.

Wain, L. V. (2014) 'Blood Pressure Genetics and Hypertension: Genome-Wide Analysis and Role of Ancestry', *Current Genetic Medicine Reports*, 2(1), pp. 13–22. doi: 10.1007/s40142-014-0032-z.

Wang, B., Sverdlov, S. and Thompson, E. (2016) 'Efficient estimation of realized kinship from SNP genotypes', *Genetics*, 205(3), pp. 1–23. doi: 10.1534/genetics.116.197004.

Wang, J. (2002) 'An estimator for pairwise relatedness using molecular markers', *Genetics*, 160(3), pp. 1203–1215.

doi:10.20944/preprints202004.0394.v1

Wang, Y., Localio, R. and Rebbeck, T. R. (2006) ‘Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions’, *Cancer Epidemiology Biomarkers and Prevention*, 15(1), pp. 124–132. doi: 10.1158/1055-9965.EPI-05-0304.

Weiss, K. M. and Lambert, B. W. (2014) ‘What type of person are you? Old-fashioned thinking even in modern science’, *Cold Spring Harbor Perspectives in Biology*, 6(1). doi: 10.1101/cshperspect.a021238.

Weitzel, K. W. *et al.* (2016) ‘The IGNITE network: A model for genomic medicine implementation and research’, *BMC Medical Genomics*, 9(1). doi: 10.1186/s12920-015-0162-5.

Wellenreuther, M. and Hansson, B. (2016) ‘Detecting Polygenic Evolution: Problems, Pitfalls, and Promises’, *Trends in Genetics*, 32(3), pp. 155–164. doi: 10.1016/j.tig.2015.12.004.

Xing, J. *et al.* (2009) ‘Fine-scaled human genetic structure revealed by SNP microarrays’, *Genome Research*, 19(5), pp. 815–825. doi: 10.1101/gr.085589.108.

Yang, J. *et al.* (2011) ‘GCTA: A tool for genome-wide complex trait analysis’, *American Journal of Human Genetics*, 88(1), pp. 76–82. doi: 10.1016/j.ajhg.2010.11.011.

Yang, W. Y. *et al.* (2012) ‘A model-based approach for analysis of spatial structure in genetic data’, *Nature Genetics*, 44(6), pp. 725–731. doi: 10.1038/ng.2285.

Yusuf, S. and Wittes, J. (2016) ‘Interpreting Geographic Variations in Results of Randomized, Controlled Trials’, *New England Journal of Medicine*, 375(23), pp. 2263–2271. doi: 10.1056/nejmra1510065.

Zhang, Y. and Pan, W. (2015) ‘Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements?’, *Genetic Epidemiology*, 39(3), pp. 149–155. doi: 10.1002/gepi.21879.