BMC Genomics

REVIEW Open Access

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[1,3*]

Abstract

The past years have seen the rise of genomic biobanks and mega-scale meta-analysis of genomic data, which promises to reveal the genetic underpinnings of health and disease. However, the over-representation of Europeans in genomic studies not only limits the global understanding of disease risk but also inhibits 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 and computable and to 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 the most common 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

Background

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 [1]. For example, the major-histocompatibility-complex antigen loci are the prototypical candidates that modulate the genetic susceptibility to infectious diseases. As a result, association studies aim to identify which loci may provide valuable information for strategising prevention, treatment, vaccination and clinical approaches [2]. 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. The International HapMap Project, practically the first international biobank [3], facilitated the routine collection of data for genome-wide association studies (GWAS) [4]. GWAS to improve clarity soon after 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 [5] and replication [6]. Since much of this criticism was due to spurious associations yielded from small sample sizes with reduced power of association analyses, major efforts were taken to recruit

* Correspondence: eran.elhaik@biol.lu.se
[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK
[3] Department of Biology, Lund University, Lund, Sweden
Full list of author information is available at the end of the article

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 [7].

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. Although it was unclear whether these new databases would deliver their most ambitious promises, the potential of biobanks in enabling personalised treatment was noted before the technology matured. It was initially expected that these databases would lead to the rapid discovery of a better genetic understanding of complex disorders, allowing for personalised treatments [8]. However, it is now clear that this expectation was exaggerated [8]. For example, a comprehensive review of the genomics of hypertension on its way to personalised medicine concluded that despite the wealth of identified genomic signals, actionable results are lacking [9]. No new drugs for the treatment of hypertension were approved for more than two decades. Moreover, the tailoring of therapy to each patient has not progressed beyond considering self-reported African ancestry and serum renin levels [9]. Another example is autism, the most extensively studied (40 years) and heavily funded ($2.4B in NIH funding over the past ten years [10]) mental disorder with nearly three dozen biobanks [11]. Despite these major efforts at understanding the disorder, there is still no single genetic test for autism, not to mention genetic treatment [12]. These gloomy reports of the state of knowledge in two of the most studied complex disorders, which typically harness massive biobanks, were not what the biobank enthusiasts envisioned at the beginning of the century [8].

Back then, both private and government-sponsored banks began amassing tissues and data. For example, Generation Scotland [13] includes DNA, tissues and phenotypic information from nearly 30,000 Scots [14]; the 100,000 Genomes Project sequenced the genomes of over 100,000 NHS patients with rare diseases, aiming to understand the aetiology of their conditions from their genomic data [15]; and the UK Biobank project sequenced the complete genomes of over half a million individuals [16] with the aim of improving the prevention, diagnosis and treatment of a wide range of diseases [17]. 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 3000 people, from a variety of regions of Russia [18]. 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 [19]. The experience of deCODE highlighted the risks in entrusting private companies to manage genomic databases, promoting similar efforts to have at least partial government control in the dozens of newly founded biobanks (reviewed in [20]), as illustrated in Fig. 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 locally-managed resources to more global resources. These should adhere to international standards to increase the accuracy of association studies and the use of biobanks [21].

Even past the formation of biobanks, many associations results failed to replicate (e.g., [22]) or show a difference in the effect across worldwide populations, in traits and disorders like body-mass index (BMI) [23], schizophrenia [24], hypertension [25] and Parkinsons' disease [26]. 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 [27] – 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 contributing to the usefulness of biobanks and their ability to serve the general population [28–30]. 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 implement these models.

Main text

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 [31]. In other words, a relatively small group of SNPs are geographically differentiated, whilst a much larger group of SNPs vary among individuals, irrespective of

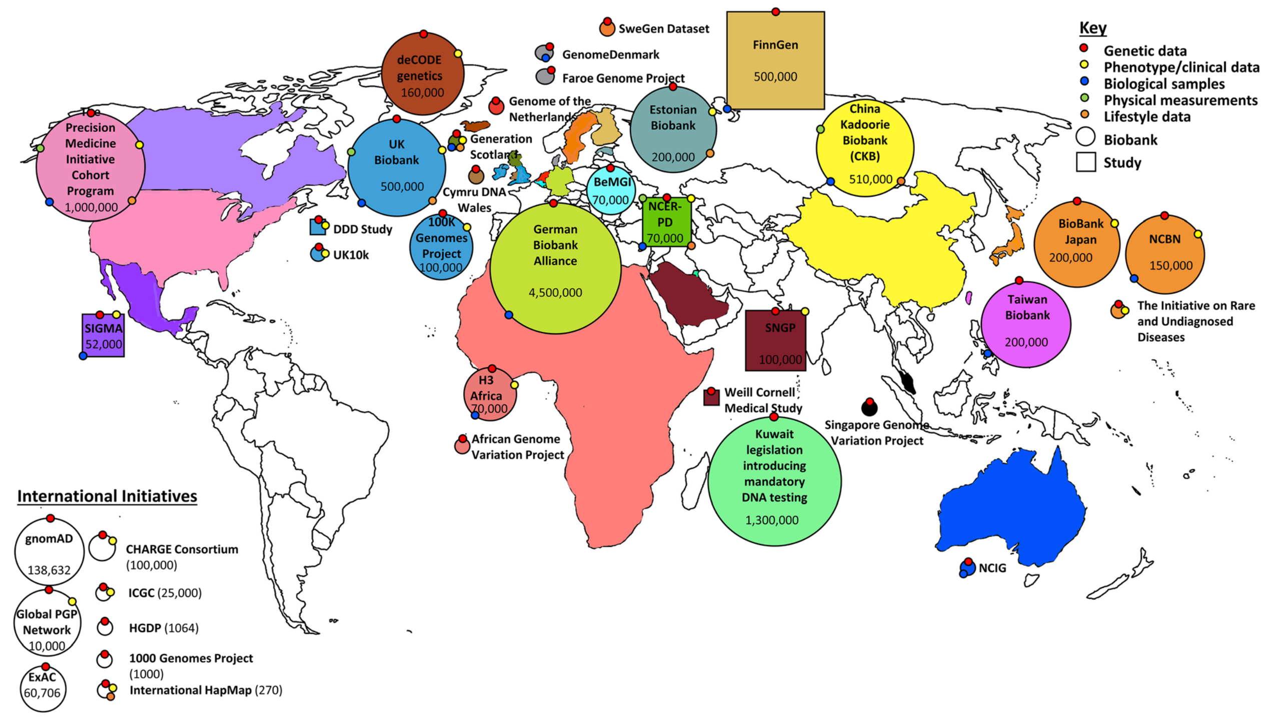

Key
Genetic data
Phenotype/clinical data
Biological samples
Physical measurements
Lifestyle data
Biobank
Study
The Precision Medicine Initiative Cohort Program 1,000,000
deCODE genetics 160,000
UK Biobank 500,000
SweGen Dataset
GenomeDenmark
Faroe Genome Project
Genome of the Netherlands
Generation Scotland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

Fig. 1 Global genomic biobanks (circles) and studies (squares). Databases vary by the type of data (see key) and their size. The map was created using R (v3.6) package 'rworldmap' (v1.3–6)

geography. However, most of these variants are rare and non-functional [32]. 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 us of what risk of disease an individual has, but also modify the effect of treatment [30]. In general, and with the clear exception for high admixture or migration followed by relative isolation [33–35], most associations between geographic location and genetic similarity are expected to hold worldwide (e.g., [36]). This is due to the exchange of genes and migrants between geographically proximate populations (e.g., [37–41]). These relationships are also expected to hold for common and rare variants [42]. 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 is widely misrepresented in GWAS studies [43]. By 2009, 96% of individuals represented in GWAS were of European descent [44]. This over-representation was rationalised 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 [45]. Consequent efforts to carry out studies on non-Europeans were met with some success; by 2016, the proportion of Europeans included in GWAS declined to 81% [46] and further to 78% in 2019 [43]. However, even then, 71.8% of GWAS individuals are recruited from only three countries: the US, UK and Iceland [47].

Not all major genetic datasets are equally diverse, and most are skewed towards individuals of European ancestry (Fig. 2). For example, 61% of the samples in the Exome Aggregation Consortium (ExAC) dataset (60,252 individuals) [48], 59% of the Genome Aggregation Database (gnomAD) (141,456 individuals) [49], 94% of the UK Biobank database (500,000 individuals) [16] and an estimated 97.6% of the deCODE database are Europeans [50]. The UK Biobank was designed to be representative of the general population of the United Kingdom; however, that makeup is only 85% "White" [51]. 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" [36]; secondly, ignorance of the global population structure prevents properly correcting the studies for *stratification bias*; and thirdly, the unequal representation of diversity within major genetic datasets increases the risk for false positives, due to chance or undetected population structure, and current methods to attempt to correct this

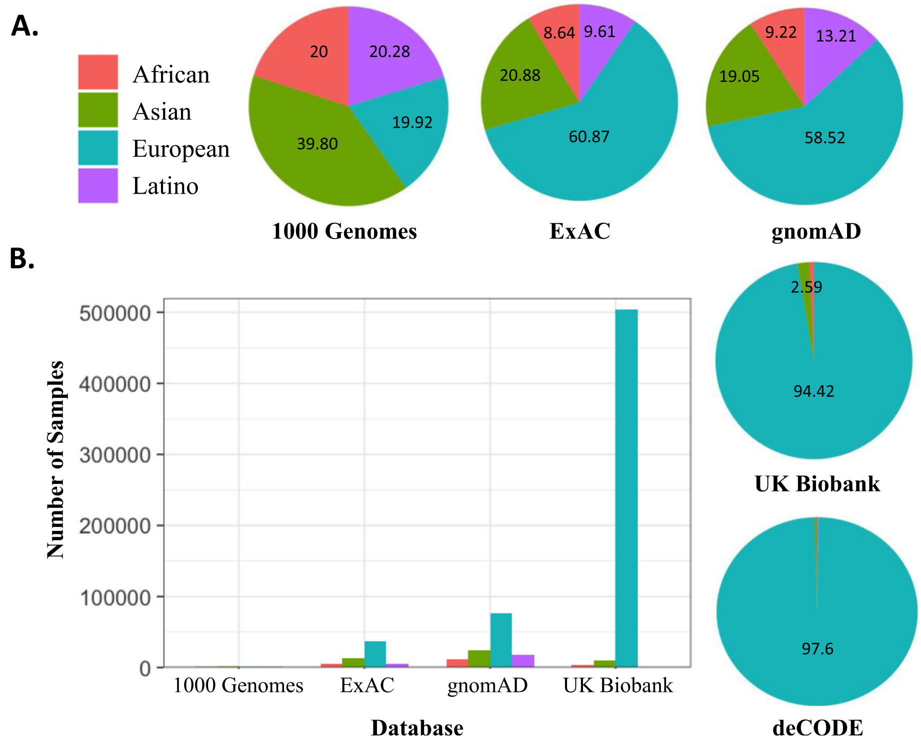


Fig. 2 The **a** percentage and **b** number of samples in the 1000 Genomes Project, the ExAC browser, the UK Biobank and the gnomAD browser categorised into five ancestry groups: European, South Asian, African, East Asian and Latin (https://www.nature.com/articles/nature15393; http://exac.broadinstitute.org/faq; 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 ancestry or number of samples; however, it can assumed 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) [50]

underlying population structure are inadequate [23]. These limitations were highlighted during the COVID-19 pandemic, as the UK biobank data were shared internationally [52] to improve the response to the virus and protect the public represented in the biobank.

Population stratification may bias GWAS through two routes: the choice of the 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., [53, 54]). A numerical example of how a false positive association can be created due to population stratification is demonstrated by Hellwege et al. [55].

However, grouping based on demographics alone does not account for differences in genetic ancestry between individuals, which leads to biased interpretation of the results or false negative or positive results [30, 56–59].

Genomic medicine and diversity

Personalised medicine is thought of as the utilisation of epidemiological knowledge to produce a granular classification of patients into cohorts. These cohorts differ in their disease susceptibility, disease prognosis or response to treatment. It is considered the epitome of twenty-first century medicine [60]. To facilitate the accurate identification and classification of individuals into cohorts, it is necessary to consider their genomes, which lends credence to the development of *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 [61]. The ability to predict individual risk of disease development, guide intervention and direct the treatment are the core principles of genomic medicine [62]. 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 [63].

Personalised genomic medicine aims to tailor a treatment to an individuals' genetic needs. This is expected to revolutionise disease treatment by using targeted therapy and treatment tailored to the individual to achieve the most effective outcome [64], as illustrated in Fig. 3. This form of genomic medicine was made feasible due to advances in computational biotechnology and its implementation into the health care system [65], illustrated in Fig. 4, alongside biological advancements that include the mapping of human genetic variation across the world through parallel global efforts [66]. However, it remains a futuristic vision rather than an everyday reality, due to the multiple obstacles that genetic studies face in deciphering complex genotype-phenotype relationships [67, 68]. One of the notorious difficulties in the field is the variation among population subgroups, which is often due to their genomic background [30]. 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 [36, 69]. The development of personalised medicine is, therefore, an area particularly affected by a lack of diversity in biobanks.

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 [70] or 26 of the 1000 Genomes populations [42]. 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., [71]). Groups are assigned either by self-identified ancestry or inferred by

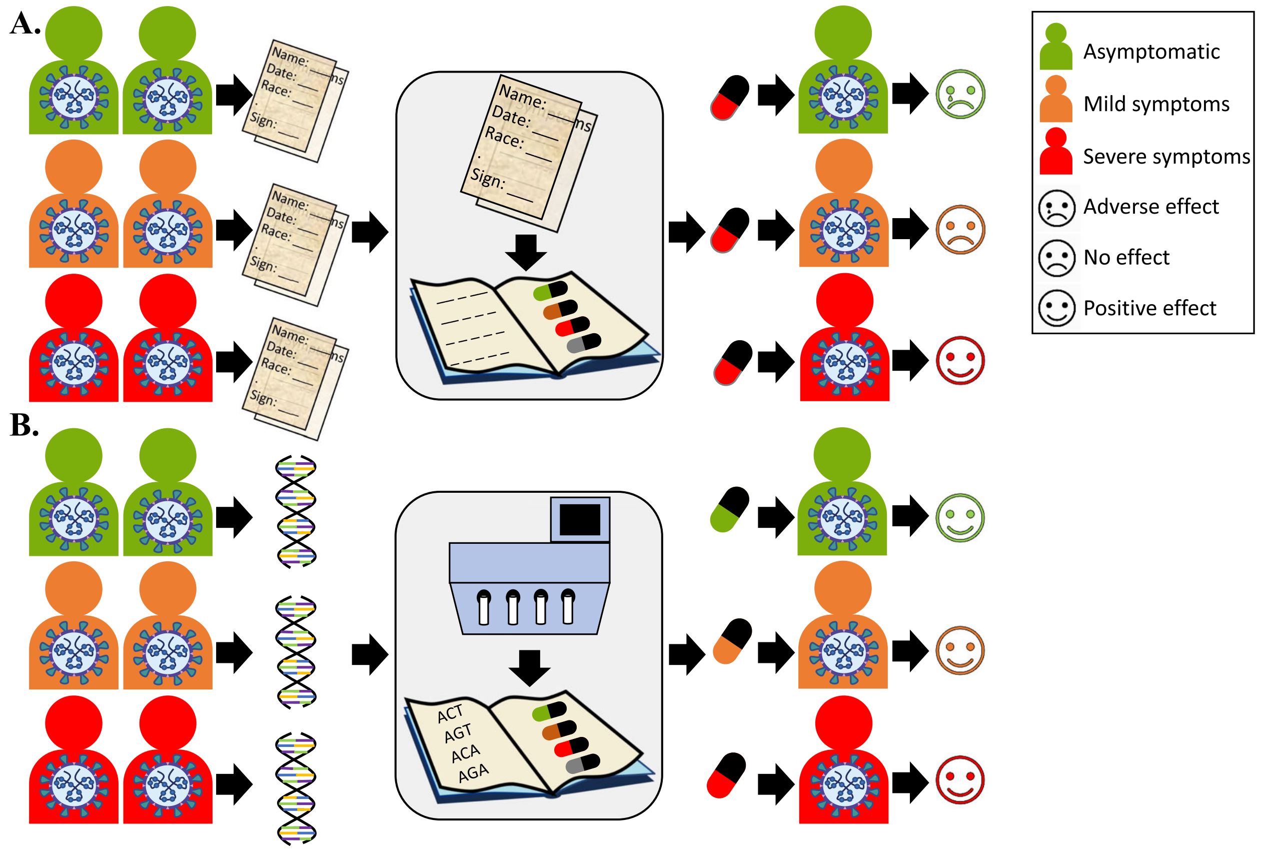
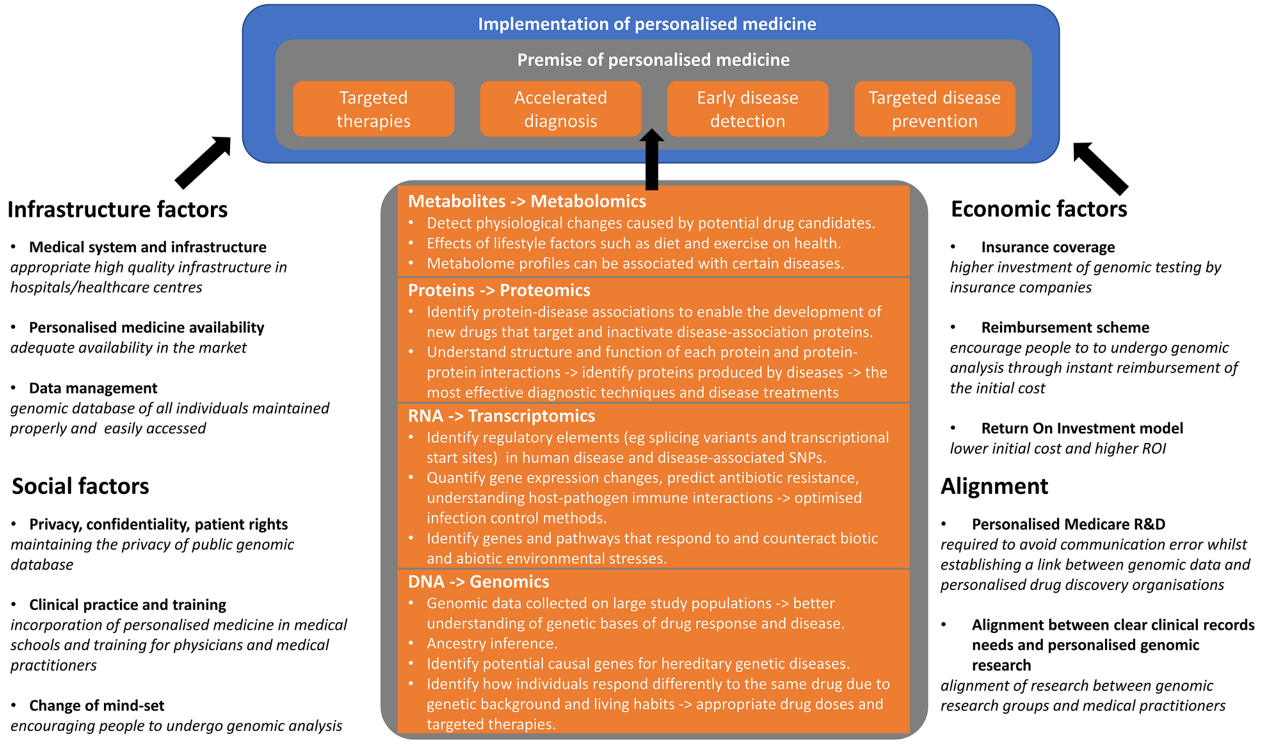


Fig. 3 Using the example of COVID-19: **a** The current method of treatment whereby all patients with the same disease receive the same treatment. **b** Personalised medicine, whereby treatment is tailored to an individual to increase effectiveness



Implementation of personalised medicine

Premise of personalised medicine

Targeted therapies

Accelerated diagnosis

Early disease detection

Targeted disease prevention

Infrastructure 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

Fig. 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)

comparison to the HapMap or 1000 Genomes populations, and each cluster is analysed independently (e.g., [71]). This approach does not account for the existence of fine-scale structure [23] 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 Supplementary Text S1, e.g., [72]) and/or mixed linear models (MLM, Supplementary Text S1, e.g. [73]) start with the Genetic Relatedness Matrix (GRM, Supplementary Text S1) [74] as the de-facto standard used to describe ancestry of large-scale genetic datasets. PCA aims to correct for the largest variation components of the GRM, whilst MLM aims to correct 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 [58, 75–77] 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 labelled as a Middle Eastern individual [58].

Both PCA and MLMs are used for meta-analyses of a large number of independent studies (e.g., BMI [78]). Meta-analysis demonstrates replication of effects of genetic risk loci and hence minimises individual cohort bias. However, the effect size estimate of meta-analysis is the averaged effect of the SNP on outcomes across several populations. The assumption that the effects of an 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 causal variants, but rather are in linkage disequilibrium (LD) with one or more causal variants, and LD patterns differ between populations [79]. Secondly, gene-environment interactions [80] may contribute to the overall effect of an SNP and these may differ by population (for example, in BMI and exercise, [81]). Thirdly, statistical artifacts can arise from differential correction power for stratification across studies [23]. 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 genotypes 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 the very large variance for all but the strongest associations.

Causal analysis using Mendelian randomisation

First outlined by Katan [82] and further developed by Davey-Smith and Ebrahim, [83], 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 the 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 reversing causality that otherwise plagues observational epidemiology. MR may, therefore, offer an affordable and faster alternative to traditional RCTs [84, 85]. 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 is often used as an instrument for tobacco smoke exposure. 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 principal 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 [86, 87] in case-control pharmacogenetics studies where differences in ancestry affect the results (e.g., weekly warfarin dose required 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 [88].

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 [89]. 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 [90, 91]. 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 [92] have been automated using online platforms [93]. In an analysis that is limited to summary data (e.g., [71]), 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., [71, 94] analysed third-party summary statistics calculated for "Europeans"). LD score regression [95] 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 trustworthy 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. [96] reported that polygenic adaptation signals based on large numbers of SNPs below genome-wide significance were found to be extremely sensitive to bias due to uncorrected population stratification. Berg et al. [97] 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. Both papers, therefore, advised caution in their interpretation. Further concerns about polygenetic scores were raised by other groups [98–100].

Drug discovery

GWAS are also used to identify druggable target genes [101]. Whilst it is not essential that the effect sizes are large, they must be associated with an underlying biological pathway [102]. There may be several reasons that limit the utilities of biobanks to identify drug targets, i.e., an association between a trait and genomic variant, like differences in lifestyle between populations, genetic interactions and genetic linkage. Since genetic variation is partly geographically differentiated, the frequencies of certain disease-causing genetic variants and variants in drug-metabolising genes may differ based on geographic location, leading to geographic disparities in the susceptibility of an individual to a disease and/or specific drug treatment [103–106]. As a result, the power to detect these unintended associations with a trait of interest is expected to grow with biobank size; therefore, correcting for population stratification will aid in the reduction of false-positive drug target leads. For pharmacogenetics to propel the practice of individualised drug therapy to become the standard of care [107], accurate genetic profiles should be constructed [30, 108] and genetic tools must be developed and verified toaccount for confounding effects using DNA sequencing analyses.

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 scales. 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.

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, thereby, 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 [36, 58, 109].

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 an 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) [72, 110]. 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 [75]. Importantly, population structure is not always in the top PCs, especially under uneven sampling or genetic drift [58, 76]. However, it is fast to compute, which contributed to the popularity of PCA as a method of first choice. 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 [111], modified methods have been proposed, albeit with limited success. The *Spatial admixture analysis (SPA)* [112], for example, had a biogeographical prediction accuracy of 2% at the country level [36]. These methods cannot be readily applied to biobank-style data as they don’t scale [75].

Unfortunately, PCA results may not be reliable, robust, or replicable as the field assumes [75]. Given enough samples having been carefully chosen, it may be tweaked to form patterns that exhibit similarity to geography (e.g., over 100 K carefully chosen samples identified several broad regions in the UK (Fig. 1 in [113])). There are alternative and complementary approaches that require consideration.

Identity-by-state (IBS) and the Genetic relatedness matrix (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) [74], in which SNPs are 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* [114] and *GTAK* [115]. The GRM can be shown to contain the same information as used by both admixture models and PCA (e.g., [116] 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* [117], but these may implicitly demonstrate the low-rank structure and hence lower correction power. Implementations like *LMM-OPS* [118] attempt to correct increased type-I error rates and a loss of power due to heterogeneous ancestry.

Ancestry as a mixture

Admixture or admixture-like analyses originated in the popular program *STRUCTURE* [119, 120]. 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* [121] and *terrastructure* [122]). However, the ancestral interpretation is often misleading [123], since sampling and genetic drift can also create the same representation for different true histories. Further, the choice of "proper" K is unclear [123, 124] and can have significant effects on the inference. Conceptually, ADMIXTURE uses the same information as PCA [116] and hence suffers from the same limitations [75].

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 [125, 126]. 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).

The difficulty in estimating relatedness increases as the relationship becomes more distant because the IBD DNA segments between individuals are shorter and more difficult to distinguish from DNA segments that are IBS. Typically, all the IBD segments that are more recent than 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 [127, 128]. 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 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 [116], which allows the identification of fine-scale population structure beyond the detection limit of related approaches [129]. Chromosome Painting is applicable for samples up to thousands but cannot be used at biobank scale [130] 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 (LAI) is to analyse individual segments of DNA to establish changes in ancestral origin. Being able to assign an 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), which is strongly related to Chromosome Painting to assign genomic segments to specified reference populations by exploiting LD

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 average number of generations since admixture) may be unknown and are difficult to learn [131]. A 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 a 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 [132].

It is straightforward 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 [133].

This has not been performed for biobanks because they contain large variations. However, as discussed above, a 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 not comparable across studies [75].

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* [134] is the most popular tool to make "supervised" inferences 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) [135, 136] or *SOURCEFIND* [137]. 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 putative 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 a reference to infer the admixture components (e.g., through a *supervised ADMIXTURE*) of other individuals without changing the model. The admixture components can be used to correct for population stratification [58] in the same manner as principal components are used, accepting that they model admixture directly, whereas PCA does not. This approach, first employed for biogeography [36], has been routinely used in population genetic investigations and was shown to be applicable to both modern and ancient populations [34, 138, 139]. Despite its promise, it is yet to be 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 structures within the reference. These approaches have not been widely applied to biobanks in the past due to issues of scale. 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 [33, 42]. 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 associated with particular populations or geographic regions. Although many sets of AIMs have been published [109], 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 [140]. 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 [141, 142]. When the diversity of the native population is low, *STRUCTURE* was shown to produce particularly misleading results [142]. 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 estimation 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 [143] and has an optional Expectation-Maximisation (EM) algorithm. *FRAPPE* uses solely the EM algorithm [144], which optimises the likelihood for both allele frequencies and fractional group memberships [144]. *FRAPPE* has been demonstrated to not only be much more computationally efficient than *STRUCTURE* but also to produce significantly fewer biased estimates [144]. However, due to its strict convergence criteria, its EM algorithm is computationally intensive and slower than *ADMIXTURE* [143], which was reported to have higher accuracy than *STRUCTURE* and *FRAPPE*.

Spatial approaches, exemplified by *GENELAND* [145, 146], *TESS* [147] and *BAPS* [148], 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 not only group individuals genetically into clusters but are also able to estimate the spatial distribution of these clusters [145–148]. Mitigating privacy concerns 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. *GENELA ND* [145, 146] has been demonstrated to be highly efficient when gene flow is low and genetic discontinuities correspond to simple shaped boundaries [149–151]; however, it is sensitive to the level of genetic differentiation [152, 153], and its accuracy [150, 154] and speed in analysing large datasets [145, 146] were criticised. Alternative tools like *TESS* and *BAPS* were shown to outperform *GENELAND* and each other under some scenarios but not in others [145, 146, 148, 154]. Interestingly, Bayesian clustering models are known to overestimate genetic structure in the presence of IBD [151, 155], 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 [156].

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 [157] 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 [131]. *MOSAIC* [158] 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 [159] and are hence both inapplicable for biobanks [131].

ChromoPainter is part of the *fineSTRUCTURE* pipeline [116], which allows the identification of fine-scale population structure that cannot be identified by PCA or related approaches [129]. Chromosome Painting is applicable for samples up to thousands but cannot be used at a biobank scale [130] 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 [160] 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 [161]. Further, many require phased data and are therefore susceptible to phasing errors.

Overall, the popular local ancestry tools are positioned along the extreme ends of limited models. At one end are mostly HMM-based tools, that either do not consider LD or are limited to two or three reference populations. At the other end are more robust tools that aim to identify haplotypes, but their high memory consumption limits their usage. An additional limitation of the local ancestry approach is the challenging evaluation of the results in follow-up analyses. The 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 IBD detection. Kinship coefficient is a classic measurement of relatedness and can be defined as the probability that two homologous alleles are drawn from each of two individuals are identical by descent [127, 128]. Software that estimate the kinship coefficient often use relatedness estimators to calculate the kinship coefficient, which falls into two categories based on the method that they use: likelihood approach to determine the likelihood of a pair of individuals having a relationships (e.g., half-sibs, full-sibs, etc.) and a relatedness-based approach to evaluate the probability of IBD [162].

ML estimators mostly use an EM algorithm to estimate the K coefficients (the proportion of genome at which two individuals share 0, 1, or 2 IBD genes); *KING, REAP* and *PC-Relate*, use the statistical method of moments to estimate the realised k coefficients [163]. *KING* can produce reliable inferences for large sample sizes (millions of unrelated and thousands of relative pairs) [164]. 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 [165]. Conversely, *REAP* [166] and *PC-Relate* [165] are able to account for different ancestry backgrounds 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 [165]. 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 [126]. 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 [167]. 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 the similarity between haplotypes that are statistically unlikely to occur in the absence of IBD sharing [168]. *PLINK* [169] 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 [170]. *fastIBD* [171] and *GERMLINE* [172] 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 100,000 individuals. Whereas *fastIBD* is based on shared haplotype frequency, *GERMLINE* is based on shared haplotype length.

Ramstetter et al. [167] 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 tools and that tools vary in their sensitivity to the IBD segment length, which corresponds 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* [173] and recover the long IBD segment by mending smaller ones using an external tool [173]. Concurrent methods generally rely on diploid genotype data, which makes them ineffective when dealing with ancient data which have a low concentration of endogenous DNA and fragmentation [125]. Since all tools underperformed in inferring remote relatedness (over 3rd degree) in diverse samples [167], further efforts should be made towards the development and testing of more robust tools.

Conclusions

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 lack of representation of diversity in genomic studies poses a limitation in the current 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., [55]), this review focussed on whether tools are conceptually comparable and whether they scale. Therefore, it assesses the confounding effects of stratification bias through the identification of population genetic structure in a standardised and comparable manner with a goal of 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 (GRM) 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 in order to explore the value of these fine-scale approaches for biobanks.

Local ancestry inference tools are still slower, though they 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 the 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 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., [174] 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 the 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 the ability to detect subtle population structures. 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. Unless populations of diverse ancestries are included, therefore incorporating an equal knowledge of genetic variation across ancestry groups, it could contribute further to health disparities and negatively impact genomic interpretation. Efforts should 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, an equal representation of genetic variation across ancestry groups and continuous improvement and development of genetic inference software, population structure inference will occur with finite detail. This will allow for more effective differentiation 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 localised 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 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 shareable 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 height 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 our study would serve to improve the accuracy, reliability and replicability of association studies and biobanks.

Abbreviations

AIM: Ancestry Informative Marker; EM algorithm: Expectation-Maximisation algorithm; GAI: Global Ancestry Inference; HMM: Hidden Markov Model; MCMC: Markov Chain Monte Carlo; MLE: Maximum Likelihood Estimate; MLM: Mixed Linear Model; LMM: Linear Mixed Model; MDS: Multidimensional Scaling; PCA: Principal Components Analysis; GWAS: Genome-Wide Association Study; BMI: Body-Mass Index; SNP: Single Nucleotide Polymorphism; RCT: Randomised Controlled Trial; MR: Mendelian

Randomisation; LD: Linkage Disequilibrium; IBD: Identical By Descent; IBS: Identical by State; LAI: Local Ancestry Inference

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07618-x.

Additional file 1: Text S1. concepts used in the study. **Table S1**. A summary of the advantages and limitations of seven of the most popular global genomic ancestry tools, as discussed in this review. **Table S2**. A summary of the advantages and limitations of the five of the most popular selected local ancestry inference tools. **Table S3**. A summary of the advantages and limitations of the six most popular of the selected relatedness inference tools, as discussed in this review. **Figure S1**. The popularity (normalised number of citations) of the different GAI software, separated into model-based (red) and non-parametric (blue) tools. The tools were found using existing review papers and free search, using the search engines 'Google,' 'Google Scholar' and the journal 'Bioinformatics' to search for keywords including: 'software,' 'tools,' 'inference,' 'biogeographic,' 'ancestry,' 'kinship' and 'haplotype.' The number of citations for the paper proposing the tool, taken from 'Google Scholar,' were compared for each tool within each domain. To account for the differences in the number of years since publication, the number of citations was normalised by dividing the number of citations by the number of years since publication. **Figure S2**. The popularity (normalised number of citations) of the different LAI software, separated into their technologies: Hidden Markov Model (HMM) (green), Chromosome Painting (red) and Statistical Learning Algorithm (SLA) (blue). Finding the tools and calculating the normalised citation number was done as in Figure S1. **Figure S3**. The popularity (normalised number of citations) of the different kinship inference software, separated into their software strategies: Identity-By-Descent (IBD) detection (red) and Kinship Coefficient Estimation (blue). Finding the tools and calculating the normalised citation number was done as in Figure S1

Acknowledgements
Not Applicable.

Authors' contributions
EE initiated the study. HC did the research. HC, DL and EE wrote the paper. All authors have read and approved the manuscript.

Funding
EE was partially supported by the MRC (MR/R025126/1), Crafoord Foundation, the Swedish Research Council (2020–03485), and Erik Philip-Sörensen Foundation (G2020–011). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lund, partially funded by the Swedish Research Council through grant agreement no. 2018–05973. The funders had no role in the study. Open Access funding provided by Lund University.

Availability of data and materials
All our data and materials are publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not Applicable.

Competing interests
EE is a consultant to DNA Diagnostic Centre.

Author details
[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK. [2]School of Mathematics and Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [3]Department of Biology, Lund University, Lund, Sweden.

Received: 16 October 2020 Accepted: 14 April 2021
Published online: 17 May 2021

References

1. Byun J, Han Y, Gorlov IP, Busam JA, Seldin MF, Amos CI. Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. BMC Genomics. 2017; 18(789):1–12. https://doi.org/10.1186/s12864-017-4166-8.
2. Shi Y, Wang Y, Shao C, Huang J, Gan J, Huang X, et al. COVID-19 infection: the perspectives on immune responses. Cell Death Differ. 2020;27(5):1451–4. https://doi.org/10.1038/s41418-020-0530-3.
3. Belmont JW, et al. A haplotype map of the human genome. Nature. 2005; 437(7063):1299–320. https://doi.org/10.1038/nature04226.
4. Visscher PM, et al. Five years of GWAS discovery. Am J Hum Genet. 2012; 90(1):7–24. https://doi.org/10.1016/j.ajhg.2011.11.029.
5. Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. Genomics Informatics. 2012;10(4):220. https://doi.org/10.5808/gi.2012.10.4.220.
6. Palmer C, Pe'er I. Statistical correction of the Winner's curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet. 2017;13(7):e1006916. https://doi.org/10.1371/journal.pgen.1006916.
7. Somiari SB, Somiari RI. The future of biobanking: a conceptual look at how biobanks can respond to the growing human biospecimen needs of researchers. Adv Exp Med Biol. 2015:11–27. https://doi.org/10.1007/978-3-319-20579-3_2.
8. Kaiser J. Population databases boom, from Iceland to the U.S. Science. 2002; 298(5596):1158–61. https://doi.org/10.1126/science.298.5596.1158.
9. Padmanabhan S, Dominiczak AF. Genomics of hypertension: the road to precision medicine. Nat Rev Cardiol. 2020;18(4):235–50. https://doi.org/10.1038/s41569-020-00466-4.
10. NIH RePORT. Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC); 2021. https://report.nih.gov/funding/categorical-spending#/. (Last Accessed 14 Feb 2021).
11. Al-jawahiri R, Milne E. Resources available for autism research in the big data era: a systematic review. PeerJ. 2017;10(7717):e2880. https://doi.org/10.7717/peerj.2880.
12. Thapar A, Rutter M. Genetic advances in autism. J Autism Dev Disord. 2020. https://doi.org/10.1007/s10803-020-04685-z.
13. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, et al. Generation Scotland: the Scottish family health study; a new resource for researching genes and heritability. BMC Med Genet. 2006;7(1). https://doi.org/10.1186/1471-2350-7-74.
14. Generation Scotland. Generation Scotland : Facts and Figures; 2016.
15. Caulfield M, et al. The 100,000 genomes project protocol. Genomics England. 2017. https://doi.org/10.6084/M9.FIGSHARE.4530893.V2.
16. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779.
17. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK biobank resource with deep phenotyping and genomic data. Nature. 2018; 562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z.
18. Oleksyk TK, Brukhin V, O'Brien SJ. The genome Russia project: closing the largest remaining omission on the world genome map. GigaScience. 2015; 4(1):53. https://doi.org/10.1186/s13742-015-0095-0.
19. J. Kaiser, Cash-Starved deCODE Is Looking For a Rescuer for Its Biobank. Science. 2009;325(5944):1054.
20. Dubow T, Marjanovic S. Population-scale sequencing and the future of genomic medicine: learning from past and present efforts; 2016. https://doi.org/10.7249/rr1520.
21. Scudellari M. Biobank managers bemoan underuse of collected samples. Nat Med. 2013;19(3):253. https://doi.org/10.1038/nm0313-253a.
22. Border R, Johnson EC, Evans LM, Smolen A, Berley N, Sullivan PF, et al. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. Am J Psychiatr. 2019;176(5):376–87. https://doi.org/10.1176/appi.ajp.2018.18070881.
23. Lawson DJ, Davies NM, Haworth S, Ashraf B, Howe L, Crawford A, et al. Is population structure in the genetic biobank era irrelevant, a challenge, or

an opportunity? Hum Genet. 2020;139(1):23–41. https://doi.org/10.1007/s00439-019-02014-8.

24. Li Z, Xiang Y, Chen J, Li Q, Shen J, Liu Y, et al. Loci with genome-wide associations with schizophrenia in the Han Chinese population. Br J Psychiatry. 2015;207(6):490–4. https://doi.org/10.1192/bjp.bp.114.150490.
25. Wain LV. Blood pressure genetics and hypertension: genome-wide analysis and role of ancestry. Curr Genet Med Rep. 2014;2(1):13–22. https://doi.org/10.1007/s40142-014-0032-z.
26. Nalls MA, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet. 2014;46(9):989–93. https://doi.org/10.1038/ng.3043.
27. Chalmers D, Nicol D, Kaye J, Bell J, Campbell AV, Ho CWL, et al. 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 Medl Ethics. 2016;17(1):39. https://doi.org/10.1186/s12910-016-0124-2.
28. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210–7. https://doi.org/10.1016/j.cell.2010.03.031.
29. Nunes K, Aguiar VRC, Silva M, Sena AC, de Oliveira DCM, Dinardo CL, et al. How ancestry influences the chances of finding unrelated donors: an investigation in admixed Brazilians. Front Immunol. 2020;11:584950. https://doi.org/10.3389/fimmu.2020.584950.
30. Yusuf S, Wittes J. Interpreting geographic variations in results of randomized, controlled trials. N Engl J Med. 2016;375(23):2263–71. https://doi.org/10.1056/nejmra1510065.
31. Elhaik E. Empirical distributions of FST from large-scale human polymorphism data. PLoS One. 2012;7(11):e49837. https://doi.org/10.1371/journal.pone.0049837.
32. Kamm J, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. J Am Stat Assoc. 2019;115(531):1–16. https://doi.org/10.1080/01621459.2019.1635482.
33. Das R, Wexler P, Pirooznia M, Elhaik E. Localizing Ashkenazic Jews to primeval villages in the ancient Iranian lands of Ashkenaz. Genome Biol Evol. 2016;8(4):1132–49. https://doi.org/10.1093/gbe/evw046.
34. Marshall S, Das R, Pirooznia M, Elhaik E. Reconstructing Druze population history. Sci Rep. 2016;6(1). https://doi.org/10.1038/srep35837.
35. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005;102(44):15942–7. https://doi.org/10.1073/pnas.0507611102.
36. Elhaik E, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat Commun. 2014;5(1):3513. https://doi.org/10.1038/ncomms4513.
37. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451(7181):998–1003. https://doi.org/10.1038/nature06742.
38. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol. 2008;32(3):215–26. https://doi.org/10.1002/gepi.20296.
39. Mountain JL, Risch N. Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups. Nat Genet. 2004;36(S11):S48–53. https://doi.org/10.1038/ng1456.
40. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298(5602):2381–5. https://doi.org/10.1126/science.1078311.
41. Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, Thara R, et al. Fine-scaled human genetic structure revealed by SNP microarrays. Genome Res. 2009;19(5):815–25. https://doi.org/10.1101/gr.085589.108.
42. Altshuler DM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. https://doi.org/10.1038/nature11632.
43. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177(1):26–31. https://doi.org/10.1016/j.cell.2019.02.048.
44. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25(11):489–94. https://doi.org/10.1016/j.tig.2009.09.012.
45. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet. 2018;19(3):175–85. https://doi.org/10.1038/nrg.2017.89.
46. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161–4. https://doi.org/10.1038/538161a.
47. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019;2(1):9. https://doi.org/10.1038/s42003-018-0261-x.
48. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91. https://doi.org/10.1038/nature19057.
49. Karczewski KJ, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. 2019:531210. https://doi.org/10.1101/531210.
50. Jurczak K. Ethnic groups and nationalities in Iceland. In: WorldAtlas; 2017.
51. Tutton R. Race/ethnicity: multidisciplinary global contexts; 2009.
52. Dyer C. Covid-19: rules on sharing confidential patient information are relaxed in England. BMJ. 2020:m1378. https://doi.org/10.1136/bmj.m1378.
53. Baughn LB, Pearce K, Larson D, Polley MY, Elhaik E, Baird M, et al. Differences in genomic abnormalities among African individuals with monoclonal gammopathies using calculated ancestry. Blood Cancer J. 2018;8(10):96. https://doi.org/10.1038/s41408-018-0132-1.
54. Baughn LB, et al. The CCND1 c.870G risk allele is enriched in individuals of African ancestry with plasma cell dyscrasias. Blood Cancer J. 2020;10(3). https://doi.org/10.1038/s41408-020-0294-5.
55. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. Curr Protoc Hum Genet. 2017;95(1):1.22.1–1.22.23. https://doi.org/10.1002/cphg.48.
56. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, et al. Demonstrating stratification in a European American population. Nat Genet. 2005;37(8):868–72. https://doi.org/10.1038/ng1607.
57. Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics. 2010;186(3):983–95. https://doi.org/10.1534/genetics.110.118661.
58. Elhaik E, Ryan DM. Pair matcher (PaM): fast model-based optimization of treatment/case-control matches. Bioinformatics. 2019;35(13):2243–50. https://doi.org/10.1093/bioinformatics/bty946.
59. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiol Biomark Prev. 2006;15(1):124–32. https://doi.org/10.1158/1055-9965.EPI-05-0304.
60. Lesko LJ, Woodcock J. Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective. Nat Rev Drug Discov. 2004;3(9):763–9. https://doi.org/10.1038/nrd1499.
61. Feero WG, Guttmacher AE, Collins FS. Genomic medicine - an updated primer. N Engl J Med. 2010;362(21):2001–11. https://doi.org/10.1056/NEJMra0907175.
62. Guttmacher AE, Collins FS. Genomic medicine - A primer. N Engl J Med. 2002;347(19):1512–20. https://doi.org/10.1056/NEJMra012240.
63. Johnson SB, Slade I, Giubilini A, Graham M. Rethinking the ethical principles of genomic medicine services. Eur J Hum Genet. 2019;28(2):147–54. https://doi.org/10.1038/s41431-019-0507-1.
64. NHS. Improving Outcomes Through Personalised Medicine. England: NHS; 2016.
65. Pasic MD, Samaan S, Yousef GM. Genomic medicine: new frontiers and new challenges. Clin Chem. 2013;59(1):158–67. https://doi.org/10.1373/clinchem.2012.184622.
66. Brieger K, Zajac GJM, Pandit A, Foerster JR, Li KW, Annis AC, et al. Genes for good: engaging the public in genetics research via social media. Am J Hum Genet. 2019;105(1):65–77. https://doi.org/10.1016/j.ajhg.2019.05.006.
67. Manolio TA, Chisholm RL, Ozenberger B, Roden DM, Williams MS, Wilson R, et al. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15(4):258–67. https://doi.org/10.1038/gim.2012.157.
68. Weitzel KW, et al. The IGNITE network: a model for genomic medicine implementation and research. BMC Med Genet. 2016;9(1). https://doi.org/10.1186/s12920-015-0162-5.
69. De Barros Damgaard P, et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018;557(7705):369–74. https://doi.org/10.1038/s41586-018-0094-2.
70. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52–8. https://doi.org/10.1038/nature09298.
71. Ooi BNS, Loh H, Ho PJ, Milne RL, Giles G, Gao C, et al. The genetic interplay between body mass index, breast size and breast cancer risk: a Mendelian randomization analysis. Int J Epidemiol. 2019;48(3):781–94. https://doi.org/10.1093/ije/dyz124.

72. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9. https://doi.org/10.1038/ng1847.
73. Zhang Y, Pan W. Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? Genet Epidemiol. 2015;39(3):149–55. https://doi.org/10.1002/gepi.21879.
74. Jiang D, Wang M. Recent developments in statistical methods for gwas and high-throughput sequencing association studies of complex traits. Biostat Epidemiol. 2018;2(1):132–59. https://doi.org/10.1080/24709360.2018.1529346.
75. Elhaik, E. Why most Principal Component Analyses (PCA) in population genetic studies are wrong. bioRxiv. 2021;2021.2004.2011.439381. https://doi.org/10.1101/2021.04.11.439381.
76. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009;5(10):e1000686. https://doi.org/10.1371/journal.pgen.1000686.
77. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 2008;40(5):646–9. https://doi.org/10.1038/ng.139.
78. Locke AE, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197–206. https://doi.org/10.1038/nature14177.
79. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15(11):1496–502. https://doi.org/10.1101/gr.4107905.
80. Purcell S. Variance components models for gene-environment interaction in twin analysis. Twin Res. 2002;5(6):554–71. https://doi.org/10.1375/136905202762342026.
81. Rask-Andersen M, Karlsson T, Ek WE, Johansson Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. PLoS Genet. 2017;13(9):e1006977. https://doi.org/10.1371/journal.pgen.1006977.
82. Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Int J Epidemiol. 2004;33(1):9. https://doi.org/10.1093/ije/dyh312.
83. Davey-Smith GD, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol. 2003;32(1):1–22. https://doi.org/10.1093/ije/dyg070.
84. Lippman SM, et al. Effect of selenium and vitamin E on risk of prostate cancer and other cancers: the selenium and vitamin E cancer prevention trial (SELECT). JAMA. 2009;301(1):39–51. https://doi.org/10.1001/jama.2008.864.
85. Mokry LE, Ahmad O, Forgetta V, Thanassoulis G, Richards JB. Mendelian randomisation applied to drug development in cardiovascular disease: a review. J Med Genet. 2015;52(2):71–9. https://doi.org/10.1136/jmedgenet-2014-102438.
86. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, et al. Mixed model with correction for case-control ascertainment increases association power. Am J Hum Genet. 2015;96(5):720–30. https://doi.org/10.1016/j.ajhg.2015.03.004.
87. Smith GD. Mendelian randomization for strengthening causal inference in observational studies: application to gene × environment interactions. Perspect Psychol Sci. 2010;5(5):527–45. https://doi.org/10.1177/1745691610383505.
88. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2018;19(2):110–24. https://doi.org/10.1038/nrg.2017.101.
89. Burgess S, Thompson SG. Mendelian randomization: methods for using genetic variants in causal estimation. London, UK: Chapman & Hall/CRC Press; 2015. https://doi.org/10.1201/b18084.
90. Koellinger PD, De Vlaming R. Mendelian randomization: the challenge of unobserved environmental confounds. Int J Epidemiol. 2019;48(3):665–71. https://doi.org/10.1093/ije/dyz138.
91. Scheinfeldt LB, et al. Challenges in translating GWAS results to clinical care. Int J Mol Sci. 2016;17(8). https://doi.org/10.3390/ijms17081267.
92. Bergholdt HKM, Nordestgaard BG, Ellervik C. Milk intake is not associated with low risk of diabetes or overweight-obesity: a Mendelian randomization study in 97,811 Danish individuals. Am J Clin Nutr. 2015;102(2):487–96. https://doi.org/10.3945/ajcn.114.105049.
93. Hemani G, et al. MR-base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv. 2016:078972. https://doi.org/10.1101/078972.
94. Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709–17. https://doi.org/10.1038/ng.3570.
95. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284–90. https://doi.org/10.1038/ng.3190.
96. Sohail M, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife. 2019;8. https://doi.org/10.7554/eLife.39702.
97. Berg JJ, et al. Reduced signal for polygenic adaptation of height in UK biobank. eLife. 2019;8. https://doi.org/10.7554/eLife.39725.
98. Bulik-Sullivan B, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3): 291–5. https://doi.org/10.1038/ng.3211.
99. Khan SS, Cooper R, Greenland P. Do polygenic risk scores improve patient selection for prevention of coronary artery disease? JAMA. 2020;323(7):614–5. https://doi.org/10.1001/jama.2019.21667.
100. Wellenreuther M, Hansson B. Detecting polygenic evolution: problems, pitfalls, and promises. Trends Genet. 2016;32(3):155–64. https://doi.org/10.1016/j.tig.2015.12.004.
101. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. Sci Transl Med. 2017;9(383):eaag1166. https://doi.org/10.1126/scitranslmed.aag1166.
102. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8): 467–84. https://doi.org/10.1038/s41576-019-0127-1.
103. Adeyemo A, Rotimi C. Genetic variants associated with complex human diseases show wide variation across multiple populations. Public Health Genomics. 2009;13(2):72–9. https://doi.org/10.1159/000218711.
104. Daar AS, Singer PA. Pharmacogenetics and geographical ancestry: implications for drug development and global health. Nat Rev Genet. 2005; 6(3):241–6. https://doi.org/10.1038/nrg1559.
105. Ioannidis JPA, Ntzani EE, Trikalinos TA. "Racial" differences in genetic effects for complex diseases. Nat Genet. 2004;36(12):1312–8. https://doi.org/10.1038/ng1474.
106. Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation in human drug-related genes. Genome Med. 2017;9(1):117. https://doi.org/10.1186/s13073-017-0502-5.
107. Lewis LD. Personalized drug therapy; the genome, the chip and the physician. Br J Clin Pharmacol. 2005;60(1):1–4. https://doi.org/10.1111/j.1365-2125.2005.02457.x.
108. Ortega VE, Meyers DA. Pharmacogenetics: implications of race and ethnicity on defining genetic profiles for personalized medicine. J Allergy Clin Immunol. 2014;133(1):16–26. https://doi.org/10.1016/j.jaci.2013.10.040.
109. Elhaik E, Greenspan E, Staats S, Krahn T, Tyler-Smith C, Xue Y, et al. The GenoChip: a new tool for genetic anthropology. Genome Biol Evol. 2013; 5(5):1021–31. https://doi.org/10.1093/gbe/evt066.
110. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1):7. https://doi.org/10.1186/s13742-015-0047-8.
111. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98–101. https://doi.org/10.1038/nature07331.
112. Yang WY, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. Nat Genet. 2012;44(6):725–31. https://doi.org/10.1038/ng.2285.
113. Galinsky KJ, Loh PR, Mallick S, Patterson NJ, Price AL. Population structure of UK biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. Am J Hum Genet. 2016;99(5):1130–9. https://doi.org/10.1016/j.ajhg.2016.09.014.
114. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.
115. Van der Auwera GA, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;(SUPL.43). https://doi.org/10.1002/0471250953.bi1110s43.
116. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453. https://doi.org/10.1371/journal.pgen.1002453.
117. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018;50(7):906–8. https://doi.org/10.1038/s41588-018-0144-6.

118. Conomos MP, et al. Genome-wide control of population structure and relatedness in genetic association studies via linear mixed models with orthogonally partitioned structure. bioRxiv. 2018:409953. https://doi.org/10.1101/409953.
119. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567–87.
120. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;55(2):945–59.
121. Raj A, Stephens M, Pritchard JK. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89. https://doi.org/10.1534/genetics.114.164350.
122. Gopalan P, Hao W, Blei DM, Storey JD. Scaling probabilistic models of genetic variation to millions of humans. Nat Genet. 2016;48(12):1587–90. https://doi.org/10.1038/ng.3710.
123. Lawson DJ, van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat Commun. 2018;9(1):3258. https://doi.org/10.1038/s41467-018-05257-7.
124. Weiss KM, Lambert BW. What type of person are you? Old-fashioned thinking even in modern science. Cold Spring Harb Perspect Biol. 2014;6(1). https://doi.org/10.1101/cshperspect.a021238.
125. Kuhn JMM, Jakobsson M, Günther T. Estimating genetic kin relationships in prehistoric populations. PLoS One. 2018;13(4):e0195491. https://doi.org/10.1371/journal.pone.0195491.
126. Moltke I, Albrechtsen A. RelateAdmix: a software tool for estimating relatedness between admixed individuals. Bioinformatics. 2014;30(7):1027–8. https://doi.org/10.1093/bioinformatics/btt652.
127. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16(1):33–44. https://doi.org/10.1038/nrg3821.
128. Thompson EA. The estimation of pairwise relationships. Ann Hum Genet. 1975;39(2):173–88. https://doi.org/10.1111/j.1469-1809.1975.tb00120.x.
129. Leslie S, et al. The fine-scale genetic structure of the British population. Nature. 2015;519(7543):309–14. https://doi.org/10.1038/nature14230.
130. Pan X, Wang Y, Wong EHM, Telenti A, Venter JC, Jin L. Fine population structure analysis method for genomes of many. Scientific Reports. 2017; 7(1).
131. Dias-Alves T, Mairal J, Blum MGB. Loter: a software package to infer local ancestry for a wide range of species. Mol Biol Evol. 2018;35(9):2318–26. https://doi.org/10.1093/molbev/msy126.
132. Illumina Microarray Solutions, 370–2013-003; 2013. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/applications/genotyping/Microarray_Solutions.pdf.
133. Lazaridis I, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513(7518):409–13. https://doi.org/10.1038/nature13673.
134. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics. 2011;12(1). https://doi.org/10.1186/1471-2105-12-246.
135. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science. 2014;343(6172):747–51. https://doi.org/10.1126/science.1243518.
136. Pagani L, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538(7624):238–42. https://doi.org/10.1038/nature19792.
137. Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, et al. Latin Americans show wide-spread Converso ancestry and imprint of local native ancestry on physical appearance. Nat Commun. 2018;9(1):5388. https://doi.org/10.1038/s41467-018-07748-z.
138. Flegontov P, Changmai P, Zidkova A, Logacheva MD, Altınışık NE, Flegontova O, et al. Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient north Eurasian ancestry. Sci Rep. 2016; 6(1). https://doi.org/10.1038/srep20768.
139. Das R, et al. The origins of Ashkenaz, Ashkenazic Jews, and Yiddish. Front Genet. 2017;8(JUN). https://doi.org/10.3389/fgene.2017.00087.
140. Esposito U, Das R, Syed S, Pirooznia M, Elhaik E. Ancient ancestry informative markers for identifying fine-scale ancient population structure in eurasians. Genes. 2018;9(12). https://doi.org/10.3390/genes9120625.
141. Kalinowski ST. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity. 2011;106(4):625–32. https://doi.org/10.1038/hdy.2010.95.
142. Lombaert E, Guillemaud T, Deleury E. Biases of STRUCTURE software when exploring introduction routes of invasive species. Heredity. 2018;120(6):485–99. https://doi.org/10.1038/s41437-017-0042-1.
143. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10.1101/gr.094052.109.
144. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol. 2005;28(4): 289–301. https://doi.org/10.1002/gepi.20064.
145. Guillot G, Estoup A, Mortier F, Cosson JF. A spatial statistical model for landscape genetics. Genetics. 2005a;170(3):1261–80. https://doi.org/10.1534/genetics.104.033803.
146. Guillot G, Mortier F, Estoup A. GENELAND: a computer package for landscape genetics. Mol Ecol Notes. 2005b;5(2):712–5. https://doi.org/10.1111/j.1471-8286.2005.01031.x.
147. Durand E, Jay F, Gaggiotti OE, Francois O. Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol. 2009;26(9):1963–73. https://doi.org/10.1093/molbev/msp106.
148. Corander J, Waldmann P, Sillanpää MJ. Bayesian analysis of genetic differentiation between populations. Genetics. 2003;163(1):367–74.
149. Blair C, et al. A simulation-based evaluation of methods for inferring linear barriers to gene flow. Mol Ecol Resour. 2012;12(5):822–33. https://doi.org/10.1111/j.1755-0998.2012.03151.x.
150. Chen C, et al. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes. 2007;7(5):747–56. https://doi.org/10.1111/j.1471-8286.2007.01769.x.
151. Safner T, Miller MP, McRae BH, Fortin MJ, Manel S. Comparison of Bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. Int J Mol Sci. 2011;12(2):865–89. https://doi.org/10.3390/ijms12020865.
152. Ball MC, Finnegan L, Manseau M, Wilson P. Integrating multiple analytical approaches to spatially delineate and characterize genetic population structure: an application to boreal caribou (Rangifer tarandus caribou) in Central Canada. Conserv Genet. 2010;11(6):2131–43. https://doi.org/10.1007/s10592-010-0099-3.
153. Coulon A, et al. Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. Mol Ecol. 2006;15(6):1669–79. https://doi.org/10.1111/j.1365-294X.2006.02861.x.
154. Latch EK, Dharmarajan G, Glaubitz JC, Rhodes OE Jr. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conserv Genet. 2006;7(2):295–302. https://doi.org/10.1007/s10592-005-9098-1.
155. Frantz AC, Cellina S, Krier A, Schley L, Burke T. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? J Appl Ecol. 2009;46(2):493–505. https://doi.org/10.1111/j.1365-2664.2008.01606.x.
156. García-Pérez MÁ. Bayesian estimation with informative priors is indistinguishable from data falsification. Span J Psychol. 2019;22:E45. https://doi.org/10.1017/sjp.2019.41.
157. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S, Pritchard JK. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics. 2009;5(6):e1000519.
158. Salter-Townshend M, Myers S. Fine-scale inference of ancestry segments without prior knowledge of admixing groups. Genetics. 2019;212(3):869–89. https://doi.org/10.1534/genetics.119.302139.
159. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics. 2012;28(10):1359–67. https://doi.org/10.1093/bioinformatics/bts144.
160. Durbin R. Efficient haplotype matching and storage using the positional burrows-wheeler transform (PBWT). Bioinformatics. 2014;30(9):1266–72. https://doi.org/10.1093/bioinformatics/btu014.
161. Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. Nat Rev Genet. 2015;16(12):727–40. https://doi.org/10.1038/nrg4005.
162. Wang J. An estimator for pairwise relatedness using molecular markers. Genetics. 2002;160(3):1203–15. https://doi.org/10.1093/genetics/160.3.1203.
163. Wang B, Sverdlov S, Thompson E. Efficient estimation of realized kinship from SNP genotypes. Genetics. 2016;205(3):1–23. https://doi.org/10.1534/genetics.116.197004.

164. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–73. https://doi.org/10.1093/bioinformatics/btq559.
165. Conomos MP, et al. Model-free estimation of recent genetic relatedness. Am J Hum Genet. 2016;98(1):127–48. https://doi.org/10.1016/j.ajhg.2015.11.022.
166. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012;91(1):122–38. https://doi.org/10.1016/j.ajhg.2012.05.024.
167. Ramstetter MD, Dyer TD, Lehman DM, Curran JE, Duggirala R, Blangero J, et al. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. Genetics. 2017;207(1):75–82. https://doi.org/10.1534/genetics.117.1122.
168. Durand EY, Eriksson N, Mclean CY. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. Mol Biol Evol. 2014;31(8):2212–22. https://doi.org/10.1093/molbev/msu151.
169. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
170. Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, Pevsner J. Inference of relationships in population data using identity-by-descent and identity-by-state. PLoS Genet. 2011;7(9):e1002287. https://doi.org/10.1371/journal.pgen.1002287.
171. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. Am J Hum Genet. 2011;88(2):173–82. https://doi.org/10.1016/j.ajhg.2011.01.010.
172. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, et al. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 2009;19(2):318–26. https://doi.org/10.1101/gr.081398.108.
173. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194(2):459–71. https://doi.org/10.1534/genetics.113.150029.
174. Cassidy LM, Martiniano R, Murphy EM, Teasdale MD, Mallory J, Hartwell B, et al. Neolithic and bronze age migration to Ireland and establishment of the insular Atlantic genome. Proc Natl Acad Sci U S A. 2016;113(2):368–73. https://doi.org/10.1073/pnas.1518445113.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.