Genome-Wide Association Studies(GWAS)

B.Sc. Biotechnology -II year

Subject-Computational Biology & Bioinformatics

Mamta Sagar
Assistant Professor, Department of Bioinformatics
UIET & IBSBT, CSJMU University, Kanpur

Introduction

GWAS, or Genome-Wide Association Studies, are responsible for the deluge of discoveries in terms of the genetic risk factors for common disease that have been pouring out of research labs recently. What you do for a genome-wide association study is find a lot of people who have the disease, a lot of people who don't, and who are otherwise well matched.

And then, searching across the entire genome using SNPs, you try to find a place where there is a consistent difference. And if you're successful--and [you've] got to be really careful about the statistics here, so that you don't jump on a lot of false positives--it allows you to zero in on a place in the genome that must be involved in disease risk without having to guess ahead of time what kind of gene you're going to find.

The beauty of GWAS is it got us past the candidate gene approach, which had been pretty frustrating because most of the candidates turned out not to be right, to be able to say the whole set of genes are your candidates, let's consider all of them, and here's a strategy that's comprehensive enough to allow you to do that.

- Genome-wide association studies (GWAS) test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease.
- This methodology has generated a myriad of robust associations for a range of traits and diseases, and the number of associated variants is expected to grow steadily as GWAS sample sizes increase.

GWAS results have a range of applications, such as gaining insight into a phenotype’s underlying biology, estimating its heritability, calculating genetic correlations, making clinical risk predictions, informing drug development programmes and inferring potential causal relationships between risk factors and health outcomes.

Genome-wide association studies (GWAS) aim to identify associations of genotypes with phenotypes by testing for differences in the allele frequency of genetic variants between individuals who are ancestrally similar but differ phenotypically. GWAS can consider copy-number variants or sequence variations in the human genome, although the most commonly studied genetic variants in GWAS are single-nucleotide polymorphisms (SNPs).

GWAS typically report blocks of correlated SNPs that all show a statistically significant association with the trait of interest, known as genomic risk loci.

Results from GWAS can be used for a range of applications. For example, trait-associated genetic variants can be used as control variables in epidemiology studies to account for confounding genetic group differences Benjamin (D. J. et al., 2012)

Selecting study populations.

GWAS often require very large sample sizes to identify reproducible genome-wide significant associations and the desired sample size can be determined using power calculations in software tools such as CaTS14 or GPC15. In addition, one can choose between population-based and family-based designs.

The choice of data resource and study design for a GWAS depends on the required sample size, the experimental question and the availability of pre-existing data or the ease with which new data can be collected.

GWAS can be conducted using data from resources such as biobanks or cohorts with disease-focused or population-based recruitment, or through direct to consumer studies.

Assembling data sets of a sufficient size to run a well-powered GWAS for a complex trait requires major investments of time and money that go beyond the capacity of most individual laboratories.

However, there are several excellent public resources available that provide access to large cohorts with both genotypic and phenotypic information, and the majority of GWAS are conducted using these pre-existing resources.

Even when new data have been collected inhouse, these will typically be co-analysed with data from pre-existing resources; collecting new data is usually required when more refined phenotyping is desired.

For all study designs, recruitment strategies must be carefully considered as these can induce collider bias and other forms of bias in the resultant data.

Genotyping. Genotyping of individuals is typically done using microarrays for common variants or nextgeneration sequencing methods such as WES or WGS that also include rare variants.

Genotyping

Microarray-based genotyping is the most commonly used method for obtaining genotypes for GWAS owing to the current cost of nextgeneration sequencing.

However, the choice of genotyping platform depends on many factors and tends to be guided by the purpose of the GWAS; for example, in a consortium-led GWAS, it is usually wise to have all individual cohorts genotyped on the same genotyping platform.

Ideally, WGS — which determines nearly every genotype of a full genome — is preferred over WES and microarrays, and is expected to become the method of choice over the next couple of years with the increasing availability of low-cost WGS technology.

Data processing.

Input files for a GWAS include anonymized individual ID numbers, coded family relations between individuals, sex, phenotype information, covariates, genotype calls for all called variants and information on the genotyping batch.

Following input of the data, generating reliable results from GWAS requires careful quality control. Some example steps include removing rare or monomorphic variants, removing variants that are not in Hardy–Weinberg equilibrium, filtering SNPs that are missing from a fraction of individuals in the cohort, identifying and removing genotyping errors, and ensuring that phenotypes are well matched with genetic data, often by comparing self-reported sex versus sex based on the X and Y chromosomes.

Software tools such as PLINK have been specifically designed to analyse genetic data and can be used to conduct many of these quality control steps20 (further software for quality control analysis and other stages of GWAS are summarized in Table 1).

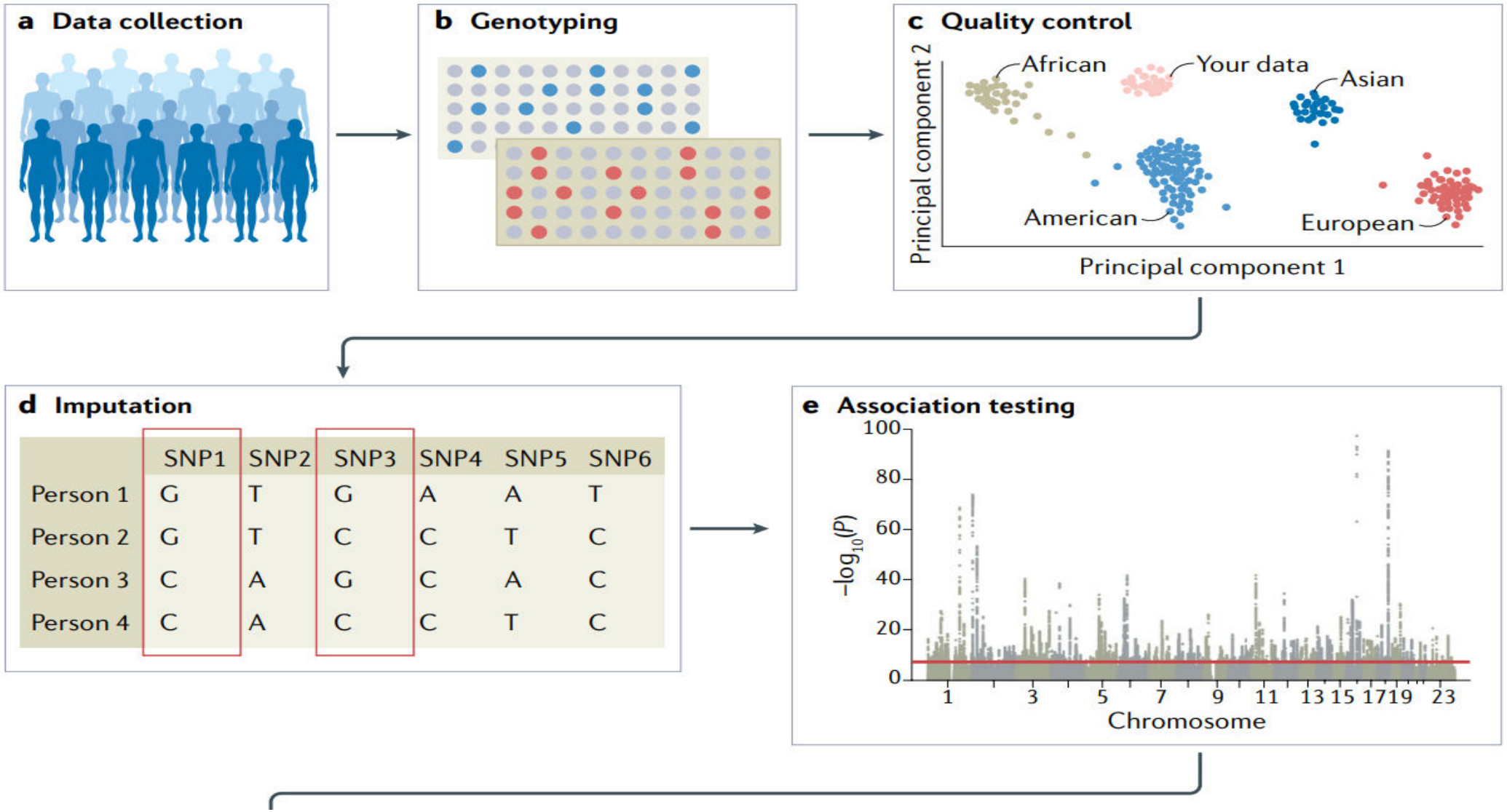
Once sample and variant quality control have been performed on GWAS array data, variants usually undergo phasing and are imputed using through an iterative process using principal component analysis; the genotypes of all individuals are used to define clusters of individuals with similar genotypes. This is done first to identify and exclude outliers, and then to compute and include principal components as covariates in subsequent GWAS regression models33.

Testing for associations.

The theory of genetic association is based on the biometrical model (see Supplementary Note for more details). Typically in GWAS, linear or logistic regression models are used to test for associations, depending on whether the phenotype is continuous (such as height, blood pressure or body mass index) or binary (such as the presence or absence of disease), respectively.

Covariates such as age, sex and ancestry are included to account for stratification and avoid confounding effects from demographic factors, with the caveat that this may reduce statistical power for binary traits in ascertained samples34. Including an additional random effect term — which is individual-specific in linear or logistic mixed models to account for genetic relatedness among individuals — can improve statistical power for genomic discovery and increase control for stratification at the cost of requiring greater computational resources35,36 (although this limitation can be addressed by using tools such as fastGWA37).

Overview of steps conducting GWAS



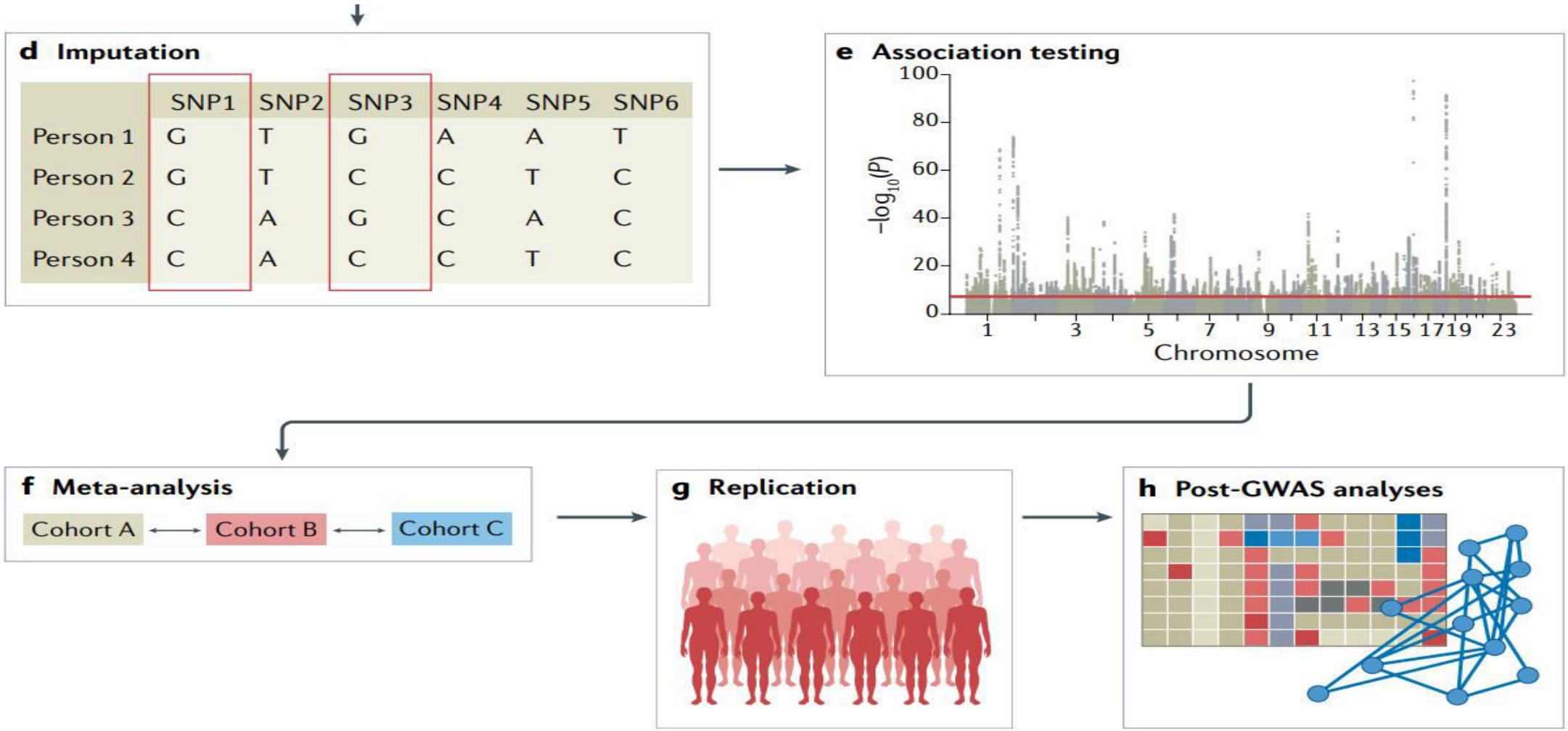


Fig. 1 | **Overview of steps for conducting GWAS. a** | Data can be collected from study cohorts or available genetic and phenotypic information can be used from biobanks or repositories. Confounders need to be carefully considered and recruitment strategies must not introduce biases such as collider bias. **b** | Genotypic data can be collected using

References

- **Genome-wide association studies,** Emil Uffelmann, , Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada,
- Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen & Danielle Posthuma
- https://www.genome.gov/genetics-glossary/Genome-Wide-Association Studies#:~:text=%3D,the%20presence%20of%20a%20disease.
- Benjamin, D. J. et al. The promises and pitfalls of genoeconomics. Annu. Rev. Econ. 4, 627–662 (2012).