c2019487_MET583 - Population Genetics

Technical Report: Ancestry Determination and Sex Checking Analysis Pipelines

c2019487

Word Count: 2303

Background

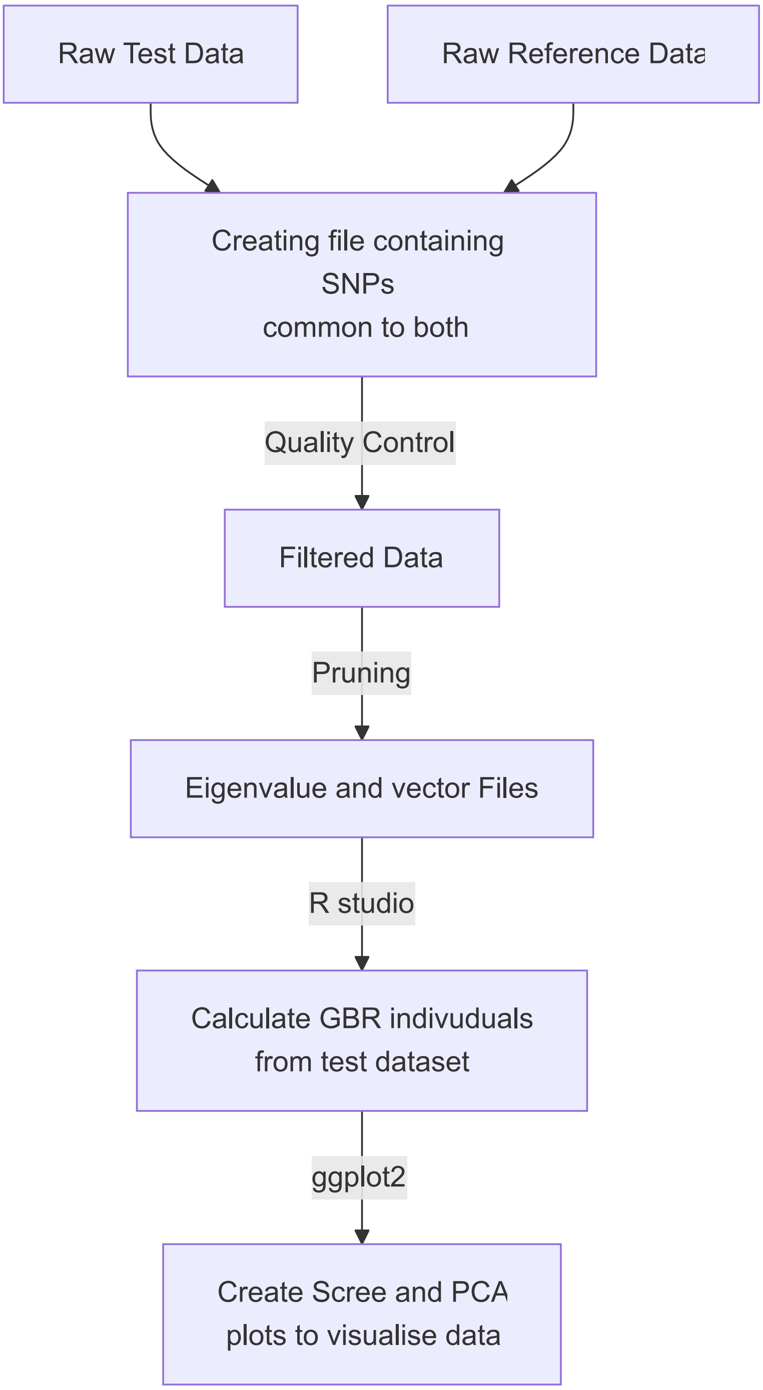
Population genetics characterises genomic individualism and is a product of generations of recombination, evolutionary pressures, and genetic mutations. The 1000 Genomes Project provides a comprehensive reference of human genomic variation within and between global populations, highlighting allele frequencies and structural differences found in all humans (The 1000 Genomes Project Consortium 2015). Principal component analysis (PCA) is a common analysis method for visualising and clustering individuals' ancestry related variation based on shared genetic variants. PCA transforms correlated genetic variants into a smaller set of principal components which capture the most relevant sources of variation and helps to increase computational efficacy, reduce noise and prevent false positives (Cavalli-Sforza and Feldman 2003; Patterson et al. 2006).

Eigenvalues represent the scale of variance shown by each principal component in the PCA. They are derived from the eigen decomposition of the correlation matrix of the data, where each PC corresponds to an eigenvector, and the eigenvalue associated with it quantifies the proportion of total variance captured by that component. The larger the eigenvalue, the more variance the corresponding principal component explains, making eigenvalues crucial for identifying the most informative PCs and designing PCA plots. Only PCs with high eigenvalues are useful as those with low eigenvalues represent noise or redundant variation: PCs with low eigenvalues are significantly less effective in PCA clustering (Jolliffe and Cadima 2016). A scree plot is used to visualise eigenvalues by plotting them against their respective PC number. This helps identify the elbow point, where eigenvalues start tailing off, which shows increasingly less significant variation in PCs. PCs beyond this point have a low effect on variance and can be ignored to reduce dimensionality while still maintaining the dataset's structure. Scree plots provide an effective way to decide the PCs applicable for downstream analysis and are a crucial preliminary step for clustering populations in PCA.

PLINK is a standard command line tool for genome wide association studies (GWAS) as it is very effective at analysing SNP-based ancestry and sex determination with its inbuilt sex-check commands and quality control methods including linkage disequilibrium (LD), minor allele frequency (MAF) thresholds, and Hardy-Wienberg equilibrium (HWE) filtering for SNP selection (Purcell et al. 2007; Anderson et al. 2010). Sex determination in PLINK relies on X-chromosome heterozygosity, which is given a metric of F = 0-1 . Females have an F ≈ 0 and males F ≈ 1, this is due to the high X-chromosome heterozygosity loss observed in male individuals owing to the fact that they possess a single X-chromosome (Purcell et al. 2007).

Methods: Workflow Description

Ancestry Determination


Raw Test Data
Raw Reference Data
Creating file containing SNPs common to both
Quality Control
Filtered Data
Pruning
Eigenvalue and vector Files
R studio
Calculate GBR indivuduals from test dataset
ggplot2
Create Scree and PCA plots to visualise data

Flowchart 1 Overview of PLINK and R-studio ancestry-matching pipeline.

A crucial step for ensuring that the dataset is one of high integrity is to identify and exclude ambiguous SNPs. An SNP is considered ambiguous when its alleles are A/T or C/G because these pairs are complementary strands of one another, making it difficult to determine whether the alleles are consistent across datasets without strand information. This ambiguity can lead to errors in data merging and in allele frequency comparisons, potentially causing GWAS misclassification. It is therefore important to remove A/T and C/G alleles from the dataset. This command (Code Block 1) operates on the `.bim` file, which contains information about SNPs in the PLINK dataset. The `awk` script selects SNPs where the reference and alternate alleles form an A/T or C/G pair. The `cut -f` function then extracts the SNP IDs (column 2), creating an exclusion list that can be applied in downstream analyses. By removing ambiguous SNPs before conducting PCA, the risk of allele misalignment is reduced, which ensures more accurate population clustering and sex determination analyses.

```
# CODE BLOCK 1
awk \
'($5=="T"&&$6=="A")|| \
($5=="A"&&$6=="T")|| \
($5=="C"&&$6=="G")|| \
($5=="G"&&$6=="C")' \
"${OUTDIR}/reference_common_maf5.bim" | \
cut -f 2 > "${OUTDIR}/ambiguous.exclude"
```

MAF refers to the occurrence of less common alleles at specific SNP loci, which are very useful variants for determining rare diseases or traits. However in ancestry determination and sex checking, rare variants are less applicable and can be filtered out. The MAF filter threshold of 5% (`--maf 0.05`) removes any rare SNPs which are seen in less than 5% of the population, in other words only the top 95% common SNPs are retained in the dataset. This 5% threshold ensures that the analysis retains statistical robustness, while also removing a small number of SNPs which contribute little to the PCA and may introduce noise to the GWAS (Code Block 2).

```
# CODE BLOCK 2
$PLINK \
--bfile "${OUTDIR}/reference_common" \
--maf 0.05 \ # Define MAF threshold - 5%
--make-bed \
--out "${OUTDIR}/reference_common_maf5"
```

The use of the R package `dplyr` allows the population data to be merged with the principal component dataset, adding a regional code corresponding to each individual from the reference dataset. Firstly, this population data can be used to colour and label the individuals on the PCA, allowing clusters to be visualised (figure 2). Secondly, the mean and standard deviation (SD) for each population, and each PC can be calculated with the purpose of using these results to predict a GBR subset in the test dataset. The test GBR subgroup can be identified as all of the individuals which fall within ±2 SD of the reference GBR subgroup; Using ±2 SD as a threshold is effective for predicting a subgroup within a population because it encompasses a statistically significant proportion of data (Almirantis 1999). This method of calculating the GBR subgroup from ±2 SD produces a dataset with 1492 GBR and 2504 unknown individuals.

In stark contrast to this, the calculations were also run with a threshold of ±1 SD which produced a dataset with far fewer predicted GBR individuals, 408 in total. To investigate these large differences in prediction power between 1 and 2 standard deviations, a random forest model was run on the reference dataset with the aim of more accurately predicting the GBR samples in the test dataset. The area under the receiver operating characteristic curve for this random forest model, which uses 1000 decision trees, was 0.8 (suggesting the model makes a correct prediction 80% of the time) (Figure 4). This model can be considered a relatively good fit. The random forest model predicted that from the test dataset, 498 individuals belong to the GBR ancestry subgroup. It can therefore be concluded that while using a ±2 SD threshold is more widely regarded, in the context of the work outlined in this report a ±1 SD threshold predicts the ancestry subgroup to a much higher degree of accuracy (Code Block 3).

```
# CODE BLOCK 3
# Calculate the mean and SD for GBR
gbr_stats <- pca_population %>%
  filter(POP == "GBR") %>%
  summarise(
    mean_PC1 = mean(PC1, na.rm = TRUE),
    sd_PC1 = sd(PC1, na.rm = TRUE),
    mean_PC2 = mean(PC2, na.rm = TRUE),
    sd_PC2 = sd(PC2, na.rm = TRUE),
    mean_PC3 = mean(PC3, na.rm = TRUE),
    sd_PC3 = sd(PC3, na.rm = TRUE)
  )

# Identify NA rows and modify their population value based on SD range of GBR
# X represents either 1 or 2 standard deviations: 1 is coded in the full pipeline
pca_eGBR <- pca_population %>%
  mutate(
    POP = if_else(
      is.na(POP) &  # Only change where population isnt defined
      (PC1 >= gbr_stats$mean_PC1 - X * gbr_stats$sd_PC1 & PC1 <= gbr_stats$mean_PC1 + X * gbr_stats$sd_PC1) &
      (PC2 >= gbr_stats$mean_PC2 - X * gbr_stats$sd_PC2 & PC2 <= gbr_stats$mean_PC2 + X * gbr_stats$sd_PC2) &
      (PC3 >= gbr_stats$mean_PC3 - X * gbr_stats$sd_PC3 & PC3 <= gbr_stats$mean_PC3 + X * gbr_stats$sd_PC3),
      "eGBR",  # Mark those within the range as "eGBR"
      as.character(POP)  # Keep the original value of POP if not NA or not within the range
    )
  )
```

Linkage Disequilibrium

Linkage Disequilibrium (LD) defines the non-random association of alleles at different points on the chromosome, these associations occur more frequently than would be expected by random chance. This is due to a number of factors including genetic drift, recombination patterns and natural selection, allowing linkage disequilibrium to be the focus of inheritance markers and quantitative characters present in the genome (Slatkin 2008). An R^2 threshold of 0.1 was used to prune SNPs in linkage disequilibrium (Code Block 5), this threshold was chosen as it was essential to reduce redundancy whilst retaining SNPs applicable for downstream analysis (Wray 2005). A window size and step size of 50 and 10 respectively avoid over-pruning whilst ensuring the pipeline is computationally efficient. These LD threshold values correspond to an accurate and informed predicted GBR population cluster (Figure 2). The differences in pruning thresholds are crucial to produce accurate downstream ancestry calls. As well as drawing on previous literature, the differences in LD R^2 have also been visualised and 0.1 is concluded as the more stringent and comprehensive correlation coefficient for performing pruning, whilst also retaining sufficient variants across all chromosomes (Figure 1).

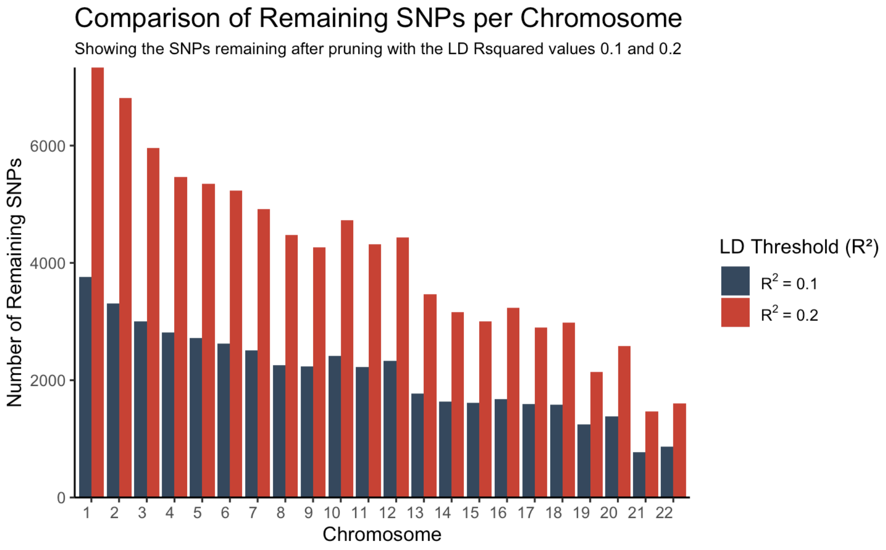
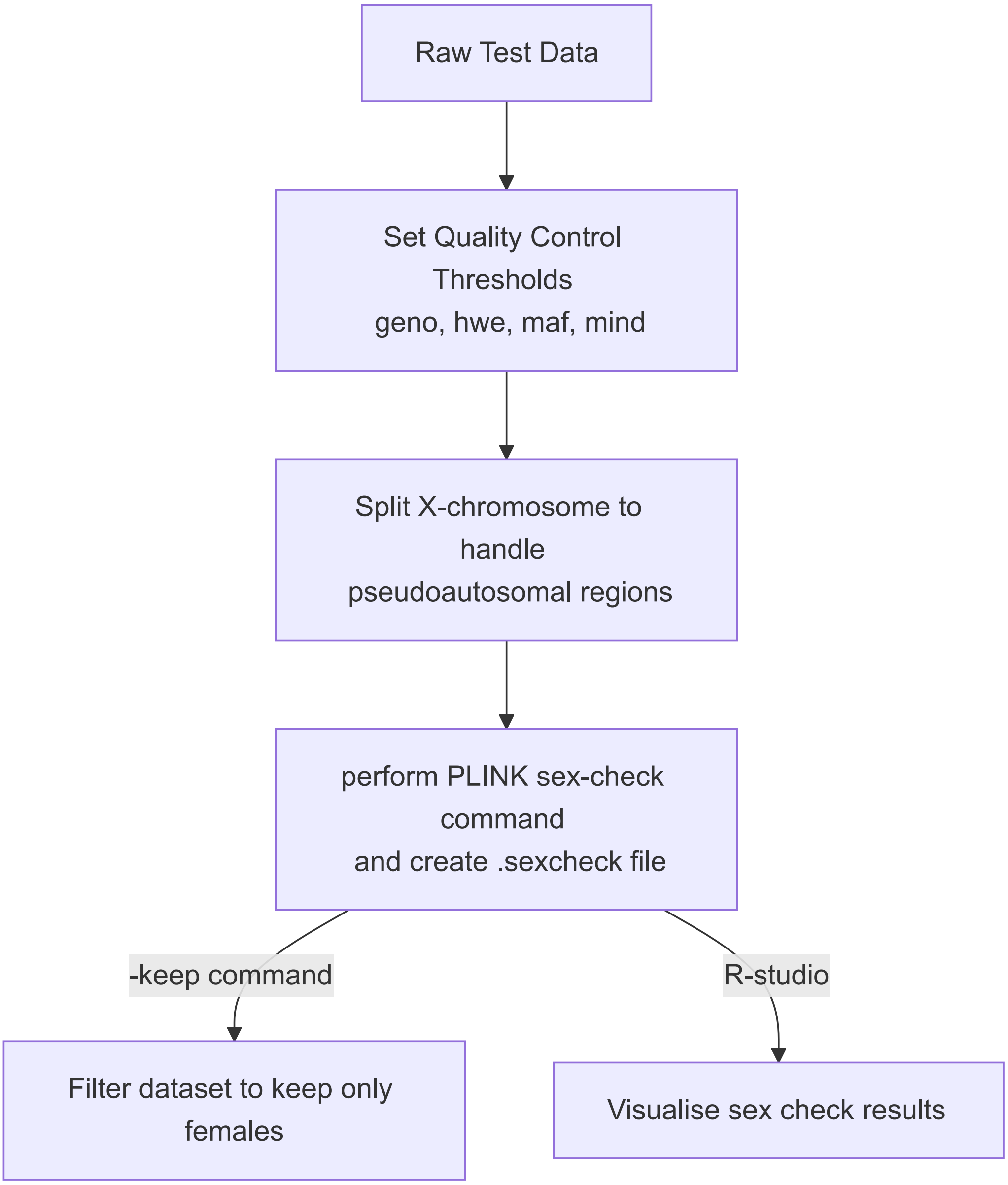


Figure 1: Retained SNPs after pruning with different linkage disequilibrium correlation thresholds.

Sex Checking


Raw Test Data
Set Quality Control Thresholds geno, hwe, maf, mind
Split X-chromosome to handle pseudoautosomal regions
perform PLINK sex-check command and create .sexcheck file
-keep command
R-studio
Filter dataset to keep only females
Visualise sex check results

Flowchart 2 Overview of PLINK and R-studio sex checking pipeline including .keep file creation and X-chromosome splitting

```
# CODE BLOCK 4
$PLINK
--bfile "$TEST" \
--geno 0.05 \
--mind 0.1 \
--maf 0.01 \
--hwe 1e-6 \
--make-bed \
--out "$OUTDIR/test_qc"
```

This command (Code Block 4) performs quality control on the test data ensuring that only high-quality variants and individuals are retained for further analyses. The command applies several QC filters. `geno 0.05` removes SNPs with a missing genotype rate greater than 5% across all individuals, which helps eliminate variants with bad genotyping quality. `mind 0.1` filter excludes individuals with over 10% missing genotypes: higher than 10% missing genotypes may indicate sample contamination, low DNA quality, or genotyping failures. The `maf 0.01` filter removes SNPs with a minor allele frequency below 1%, as low-frequency variants contribute little statistical power and may introduce biases due to sequencing artefacts. The `hwe 1e-6` filter excludes variants deviating from HWE at a significance level of 1×10^{-6} as large deviations in a population may indicate genotyping errors or population substructure. These QC steps are essential to ensure the reliability of sex-check verification and pave the way for further QC steps later on in the pipeline.

```
# CODE BLOCK 5
$PLINK \
--bfile "${OUTDIR}/reference_common_maf5_noWS_aims_noLD" \
--indep-pairwise 50 10 0.1 \
--out "${OUTDIR}/ld_independent"
```

Splitting the X-chromosome using `--split-x` ensures proper separation of pseudoautosomal regions (PAR) from non-PAR regions, as they follow different inheritance patterns. This distinction is critical for accurate sex checks, as non-PAR regions reflect sex-specific genotypes, while PAR regions behave autosomally, potentially skewing results if not handled correctly. `--split-x` must also be performed with the correct reference genome build so variants can be correctly assigned to the PARs, for this 1000 genomes reference data the genome assembly GRCh37 is used (`--split-x hg19`).

The sex determination pipeline concludes with PLINK's `--sex-check` command which outputs a 6 column table, and column 5 contains the inbreeding coefficient for each individuals X-chromosome. Due to previous QC (including `--split-x`) the results show a strong split between male and female individuals in the dataset, with no samples falling between F = 0.2 and 0.8 which shows that there is very low aneuploidy in the dataset. The male subgroup are exclusively F=1, whereas the female subgroup range from F = -0.1<0.2 (Figure 3). Negative F values in the female subgroup can be explained in a number of ways, including naturally high heterozygosity, random genotyping errors, or sample contamination (Chang 2025). A focus for future research would be to further investigate the different causes of negative F values. After checking for discrepancies between the reported sex (.sexcheck file PEDSEX column) and the F value predicted sex (.sexcheck F column) the file containing any discrepancies contained 0 lines, showing that there was 100% agreement between the two columns. The commands `awk '$6 == "F" && $3 != "2"' "${OUTDIR}/test.sexcheck" > "${OUTDIR}/test_sex_check_discrepancies.txt"` and `awk '$6 == "F" && $3 != "2"'`

`"${OUTDIR}/reference.sexcheck" > "${OUTDIR}/ref_sex_check_discrepancies.txt"` were run to check any discrepancies in the `.sexcheck` file.

Results and Discussion

PCA was conducted to observe and characterise the genetic structure of the test group, specifically the GBR subgroup, using the 1000 genomes project as a reference. PLINK was used to conduct thorough QC, which culminated in the creation of combined eigenvalue and eigenvector files. From this, a scree plot (Figure 5) and subsequent PCA plots (Figure 2) were constructed of PC1 and 2, PC 1 and 3, and PC 2 and 3.

The PCA plots are very clear in their clustering, which have be coloured according to the population data from the reference `.super-population` file (continental population), which is a result of the PLINK QC measures and the utility of PCs 1-3. The predicted GBR subgroup from the test data which contains 408 individuals has also been plotted on the PCA and can be seen in bright orange. As well as validating these results with a random forest model (0.8 AUC, Figure 4), the predicted GBR population also falls in the centre of the "EUR" or European cluster, forming a tight cluster. It has been previously stated that there are shared variants between closely related European populations, including GBR, Spanish (IBS) and Finnish (FIN) which appear in the 1000 genomes reference data, and that these populations form tight clusters within a wider European cluster in PCA (Fedorova et al. 2016; Elhaik 2022). The remaining unidentified samples from the test data are also mapped on the PCA as the light grey points (Figure 2), further analysis could determine likely populations super populations of these individuals using similar techniques, they appear to be spread across a number of populations including EUR, SAS and AMR. This is an exciting area of future research and would help to further validate the GBR predictions of this report.

While stringent QC filtering ensured high-quality genotype data, potential limitations could include the removal of individuals with ancestry overlap due to strict ±1 SD thresholds.
It has not be quantified in this report, however it is likely that should all the unknown samples' populations be predicted, a portion of individuals would overlap and fall into multiple regions. To overcome this, a more stringent machine learning model could be used to help determine the perfect SD threshold, however due to various factors including genetic mutations, and historic immigration patterns it is unlikely that a 100% accurate prediction model is possible. An example of this would be a Gaussian mixture model which works to effectively predict the distribution of genetic variance, the model captures dynamic genetic subpopulations and has been shown to enhance the accuracy of genotype and phenotype prediction (Au et al. 2011).

The use of a Histogram to visualise the F values derived from the sex determination shows clearly the split between the two sexes. Males show near-zero heterozygosity (F > 0.99) due to hemizygosity on the X-chromosome (Mohandas et al. 1992) after PLINKs removal of the pseudoautosomal region (`--split-x`). When the PAR is split from the X-chromosome, the F-values are generally 0.99 or above due to the incredibly high loss of heterozygosity seen in the male population. There is more variance in the female group due to the limited heterozygosity loss, which is a result of females having two X-chromosomes. The histogram shows a good sex determination for a number of reasons: the X-chromosome has been effectively split, removing the PARs; no individuals fall inside the 0.2>0.8 region, and the female subgroup shows higher variation than the males. There were no discrepancies between the F value sex determination and the sex reported in the metadata. Additionally, sex classification relied on predetermined F value thresholds, which might not fully account for X-chromosome variability in certain populations, and genotyping errors including poor-quality SNP calls or incorrect reference mapping, could wrongly increase or decrease F values.

Data

- Reference genotype data was obtained from the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015).
- High LD ranges (high-ld-b37) was obtained from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015).
- Test data is unknown
-

Software

- R version 4.4.1 (2024-06-14)
- Plink 1.9

References

Almirantis, Y. 1999. A Standard Deviation Based Quantification Differentiates Coding from Non-coding DNA Sequences and gives Insight to their Evolutionary History. *Journal of Theoretical Biology* 196(3), pp. 297–308. doi: https://doi.org/10.1006/jtbi.1998.0840.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. and Zondervan, K.T. 2010. Data quality control in genetic case-control association studies. *Nature Protocols* 5(9), pp. 1564–1573. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025522/ [Accessed: 14 November 2019].

Au, K., Lin, R. and Foulkes, A.S. 2011. Mixture Modelling as an Exploratory Framework for Genotype–Trait Associations. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 60(3), pp. 355–375. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC3285383/ [Accessed: 5 February 2025].

Cavalli-Sforza, L.L. and Feldman, M.W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics* 33(S3), pp. 266–275. doi: https://doi.org/10.1038/ng1113.

Chang, C. 2025. *Basic Statistics - PLINK 1.9*. Available at: https://www.cog-genomics.org/plink/1.9/basic_stats [Accessed: 29 January 2025].

Elhaik, E. 2022. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports* 12(1), p. 14683. Available at: https://www.nature.com/articles/s41598-022-14395-4.

Fedorova, L., Qiu, S., Dutta, R. and Fedorov, A. 2016. Atlas of Cryptic Genetic Relatedness among 1000 Human Genomes. *Genome Biology and Evolution* 8(3), pp. 777–790. Available at: https://academic.oup.com/gbe/article/8/3/777/2574139/ [Accessed: 5 February 2025].

Jolliffe, I.T. and Cadima, J. 2016. Principal component analysis: a review and recent developments. _Philosophical Transactions of the Royal Society A: Mathematical, Physical and

Engineering Sciences_374(2065), p. 20150202. Available at: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

Mohandas, T.K., Speed, R.M., Passage, M.B., Yen, P.H., Chandley, A.C. and Shapiro, L.J. 1992. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *American Journal of Human Genetics* 51(3), p. 526. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC1682713/ [Accessed: 5 February 2025].

Patterson, N., Price, A.L. and Reich, D. 2006. Population Structure and Eigenanalysis. *PLoS Genetics* 2(12), p. e190. doi: https://doi.org/10.1371/journal.pgen.0020190.

Purcell, S. et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81(3), pp. 559–575. doi: https://doi.org/10.1086/519795.

Slatkin, M. 2008. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9(6), pp. 477–485. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124487/.

The 1000 Genomes Project Consortium. 2015. A Global Reference for Human Genetic Variation. _Nature_526(7571), pp. 68–74. doi: https://doi.org/10.1038/nature15393.

Wray, N.R. 2005. Allele Frequencies and the *r_2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies. _Twin Research and Human Genetics* 8(2), pp. 87–94. doi: https://doi.org/10.1375/1832427053738827.