

18

Genome-Wide Association Studies (GWAS)

18.1: Learning Objectives

By the end of this chapter, the reader should be able to describe the design and analytical pipeline of a genome-wide association study, define genome-wide significance and explain the multiple-testing correction it embodies, interpret Manhattan and quantile–quantile (Q–Q) plots, describe the problem of population stratification and its correction, and discuss the specific opportunities and challenges of GWAS in Nigerian and African populations.

18.2: From Candidate-Gene to Genome-Wide Studies

The association tests of Chapter 11 were framed around a single candidate marker, chosen in advance on biological grounds. **Genome-wide association studies (GWAS)** instead test hundreds of thousands to millions of markers (SNPs), spread across the entire genome, for association with a trait or disease — an approach made possible by high-throughput SNP genotyping arrays (and, increasingly, whole-genome sequencing) and motivated by the recognition that most common-disease susceptibility loci could not have been confidently predicted in advance from biological knowledge alone.

The essential logic of GWAS rests on **linkage disequilibrium** (Chapter 11): because genotyped SNPs are in LD with many nearby, untyped variants, a relatively modest, well-

chosen panel of markers can indirectly “tag” and test the effect of most common genetic variation across the genome, without needing to genotype every possible variant directly.

18.3: The GWAS Analytical Pipeline

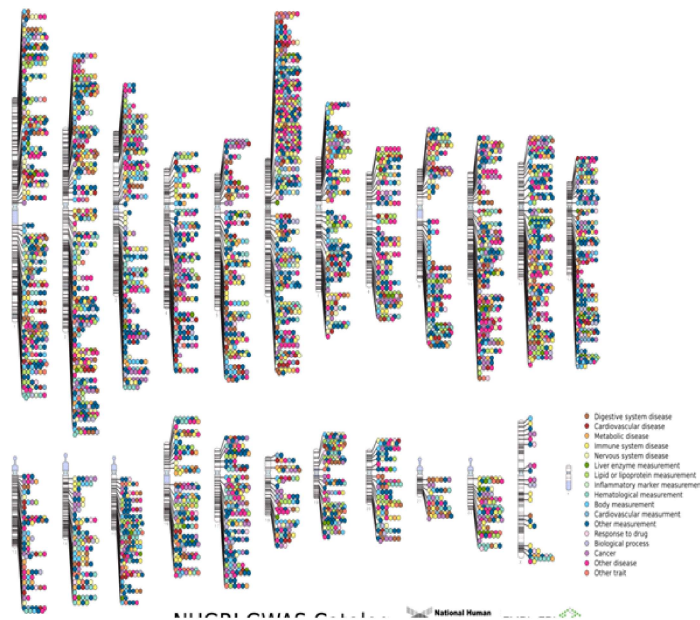
A typical GWAS proceeds through the following stages:

1. **Sample and phenotype collection:** cases and controls (or a quantitative-trait cohort), with rigorous phenotype definition — an area demanding particular care in Nigerian studies, e.g., standardising the definition and measurement of hypertension or T2DM across multiple recruiting hospital sites.
2. **Genotyping and quality control (QC):** removing SNPs and samples with high missingness, SNPs deviating from Hardy–Weinberg equilibrium in controls (Chapter 3, often used as a QC filter, since strong HWE departure frequently signals genotyping artefact rather than true biology), samples with discordant sex or unexpectedly high relatedness/duplication, and SNPs with very low minor allele frequency (which provide little statistical power).
3. **Population structure assessment:** principal component analysis (PCA) of genome-wide genotypes to identify and adjust for ancestry differences between cases and controls (Chapter 11, Section 11.7) — an especially critical step in Nigeria, given substantial genetic differentiation among ethnolinguistic groups (e.g., Yoruba, Igbo, Hausa-Fulani, and numerous minority groups), documented by projects such as the International HapMap Project’s Yoruba (YRI) panel and the H3Africa consortium.
4. **Imputation:** statistically inferring genotypes at untyped variants, using a reference haplotype panel and LD patterns (Chapter 11), to boost the effective density of tested variants — historically a particular challenge for Nigerian/African studies because widely used reference panels (e.g.,

early phases of 1000 Genomes, HapMap) under-represented African haplotype diversity, motivating panels built specifically from African genomes.

5. **Single-marker association testing:** fitting, at each SNP in turn, a regression model (logistic regression for binary disease traits, as in Chapter 11, Section 11.7; linear regression for quantitative traits) of phenotype on genotype (usually coded additively, 0/1/2 copies of one allele), adjusted for covariates including ancestry principal components.
6. **Genome-wide significance thresholding and replication** (Section 14.4–14.5).

18.4: Genome-Wide Significance: The Multiple-Testing Problem



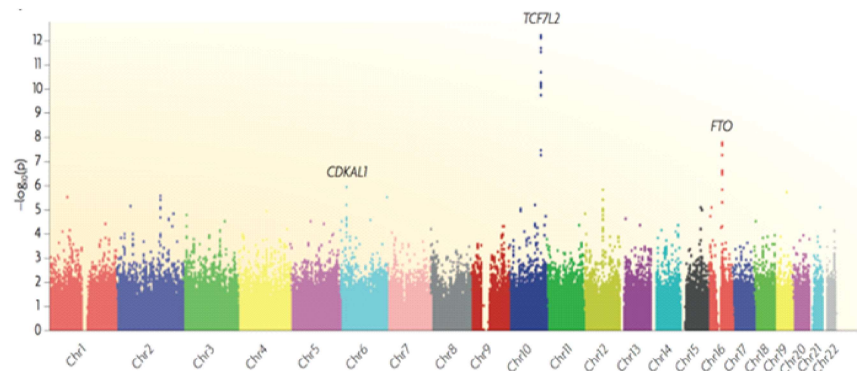
Chromosome ideogram showing the genomic locations of GWAS-discovered variants across the genome, coloured by associated trait category.

Testing hundreds of thousands to millions of SNPs simultaneously creates a severe multiple-testing problem: if each individual test used the conventional $\alpha = 0.05$ threshold, an enormous number of false-positive associations would be expected purely by chance. A simple, widely used (if somewhat conservative) **Bonferroni correction**, assuming approximately one million independent tests (a common approximation for the effective number of independent common-variant tests across the genome in populations of predominantly non-African ancestry — see Section 18.6 for the African-specific complication), gives the now-standard **genome-wide significance threshold**:

$$\alpha(\text{genome-wide}) = 0.05/1,000,000 = 5 \times 10^{-8}$$

A SNP association with $p < 5 \times 10^{-8}$ is conventionally reported as **genome-wide significant**; p between roughly 10^{-7} and 10^{-5} is often termed “suggestive,” analogous to the suggestive/significant LOD-score distinction of Chapter 17.

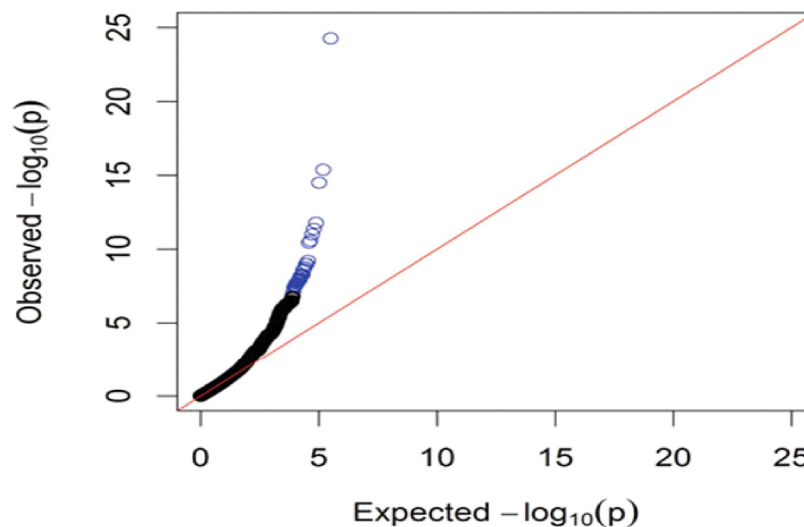
18.5: Visualising GWAS Results: Manhattan and Q-Q Plots



A Manhattan plot: “ $-\log_{10}(p)$ plotted against chromosomal position, with SNPs coloured by chromosome, revealing genome-wide association signals as sharp towers.

A **Manhattan plot** displays $-\log_{10}(\text{p-value})$ for every tested SNP on the y-axis, against chromosomal position on the x-axis (SNPs ordered by chromosome and position), producing a characteristic skyline in which genuine association signals appear as sharp “towers” (clusters of highly significant SNPs, reflecting LD among neighbouring markers, Chapter 15) rising above the genome-wide significance line.

A **quantile–quantile (Q–Q) plot** compares the observed distribution of $-\log_{10}(\text{p-values})$ across all tested SNPs to the distribution expected under the null hypothesis of no association anywhere in the genome (which should follow a uniform distribution of raw p-values, hence a straight diagonal line on the Q–Q plot).



A genome-wide Q–Q plot: observed “ $-\log_{10}(p)$ ” values (black dots) plotted against their expected values under the null (diagonal red line). The bulk of SNPs track the diagonal closely, indicating good control of confounding, while a small number of SNPs at the upper right depart sharply above the line — the true association signals.

A genome-wide Q–Q plot: observed $-\log_{10}(p)$ values (black dots) plotted against their expected values under the null (diagonal red line). The bulk of SNPs track the diagonal closely, indicating good control of confounding, while a small number of SNPs at the upper right depart sharply above the line — the true association signals.

Systematic deviation of the bulk of the Q–Q plot away from the diagonal (particularly, an inflation of small p-values across nearly all SNPs, not just a few clear outliers) is a standard diagnostic for uncorrected **population stratification** or other systematic confounding, and is commonly summarised by the **genomic inflation factor**, λ_{GC} (ideally close to 1.0; substantially greater than 1.0 signals inadequate stratification correction).

Theorem 18.1 (Genomic Control Correction)

If, under a global null hypothesis of no true association at m tested SNPs, uncorrected population stratification inflates every SNP's chi-square test statistic by a common multiplicative factor λ , i.e. $\chi^2_{i,\text{observed}} = \lambda \cdot \chi^2_{i,\text{true} - \text{null}}$ for each SNP i , then dividing every observed statistic by $\hat{\lambda} = \text{median}(\chi^2_{\text{observed}}) / \text{median}(\chi^2_1)$ restores the corrected statistics to their nominal χ^2_1 null distribution.

Proof sketch. Under the stated inflation model, each $\chi^2_{i,\text{true} - \text{null}}$ genuinely follows a χ^2_1 distribution (since the vast majority of SNPs, under the global null, have no true association), whose theoretical median is $\text{median}(\chi^2_1) \approx 0.455$. Since $\chi^2_{i,\text{observed}} = \lambda \chi^2_{i,\text{true} - \text{null}}$ for every SNP, the *sample* median of the observed statistics across all m SNPs satisfies $\text{median}(\chi^2_{\text{observed}}) = \lambda \cdot \text{median}(\chi^2_{\text{true} - \text{null}}) \approx \lambda \times 0.455$ (medians scale linearly under a constant multiplicative transformation). Solving for λ gives the genomic control estimator

$\hat{\lambda} = \frac{\text{median}(\chi^2_{\text{observed}})}{\text{median}(\chi^2_1)}$

$\hat{\lambda} = \frac{\text{median}(\chi^2_{\text{observed}})}{\text{median}(\chi^2_1)}$

$\hat{\lambda} = \text{median}(\chi_{\text{observed}}^2)/0.455 = \text{median}(\chi_{\text{observed}}^2)/\text{median}(\chi_1^2)$. Dividing each observed statistic by this estimated $\hat{\lambda}$: $\chi_{i,\text{observed}}^2/\hat{\lambda} = \lambda\chi_{i,\text{true-null}}^2/\hat{\lambda} \approx \chi_{i,\text{true-null}}^2$ (since $\hat{\lambda} \rightarrow \lambda$ as $m \rightarrow \infty$ by the consistency of the sample median), recovering statistics that are, asymptotically, correctly distributed as χ_1^2 under the null. ■

The median (rather than the mean) is deliberately used in $\hat{\lambda}$ because the bulk of SNPs are null under any realistic genome-wide scan, so the median is robust to the small number of truly associated SNPs at the extreme tail (which would otherwise inflate a mean-based estimator and lead to over-correction of genuine signals) — a practical design choice that follows directly from the robustness properties of the median as a summary statistic.

Worked Example 18.2 (Genomic control correction for a multi-ethnic Nigerian GWAS)

A GWAS of type 2 diabetes recruits cases and controls from across Nigeria without fully matching for ethnic background, testing 500,000 SNPs. The median chi-square statistic across all SNPs is found to be 0.57 (compared with the theoretical null median of 0.455), suggesting stratification-driven inflation. By Theorem 18.1, $\hat{\lambda} = 0.57/0.455 \approx 1.25$. A SNP with an uncorrected chi-square statistic of 32 (nominal $p \approx 1.5 \times 10^{-8}$, just past genome-wide significance) is corrected to $32/1.25 = 25.6$ (corrected $p \approx 4.2 \times 10^{-7}$) — no longer meeting the genome-wide significance threshold of 5×10^{-8} once the estimated stratification inflation is accounted for, a sobering illustration of why Nigerian and other multi-ethnic African GWAS must budget carefully for ancestry principal components (Section 14.5) and genomic control correction before reporting any “discovery,” since

failing to do so can convert a spurious, stratification-driven signal into an apparently genome-wide-significant (but false) finding.

18.6: Special Considerations for GWAS in Nigerian and African Populations

Several features of Nigerian and African population genetics have direct, practical consequences for GWAS design and interpretation, extending themes introduced in Chapters 15–16:

- **Greater genetic diversity and faster LD decay** (Section 15.6) mean that African populations require denser SNP panels (or high-quality imputation from African-inclusive reference panels) to achieve comparable genomic coverage to studies in European or East Asian populations — a naive application of a genotyping array or reference panel designed primarily around non-African LD patterns will under-capture true genetic variation and reduce statistical power in Nigerian samples.
- **Greater within-Africa population substructure:** Nigeria alone hosts substantial ethnolinguistic diversity, and even modest allele-frequency differentiation among groups (e.g., Yoruba, Igbo, Hausa-Fulani, and others) can generate confounding via population stratification (Chapter 3's Wahlund effect, in its GWAS-scale form) if not properly modelled with ancestry principal components or mixed models.
- **The effective number of independent tests** underlying the standard 5×10^{-8} threshold was originally calibrated largely on non-African (mainly European-ancestry) LD patterns; because African genomes carry more genetic variation and less redundant LD, some methodological work suggests that a modestly more stringent genome-wide threshold may, in principle, be appropriate for African-

ancestry GWAS — though 5×10^{-8} remains the most widely applied convention for comparability across studies.

- **The historical under-representation of African genomes** in reference panels, discovery cohorts, and downstream clinical-genetic risk tools (e.g., polygenic risk scores) means that genetic findings and prediction tools developed almost entirely in European-ancestry cohorts frequently transfer poorly to Nigerian populations — underscoring the scientific and public-health importance of dedicated African GWAS efforts, such as those coordinated through the H3Africa consortium, in which Nigerian institutions (including the University of Ibadan and University College Hospital, Ibadan) have played a substantial role, for conditions of major regional relevance including sickle-cell disease modifiers, hypertension, and infectious-disease susceptibility (e.g., malaria and Lassa fever).

18.7: Worked Example 18.1 (Interpreting a GWAS Result)

A GWAS of hypertension in a Nigerian cohort of 5,000 cases and 5,000 controls, testing 800,000 genome-wide SNPs after quality control, reports a lead SNP with $p = 3.2 \times 10^{-9}$ and odds ratio 1.22 (per copy of the risk allele), located within an intron of a gene previously implicated in renal sodium handling.

(a) Is this result genome-wide significant? Since $p = 3.2 \times 10^{-9} < 5 \times 10^{-8}$, yes, this SNP reaches conventional genome-wide significance.

(b) What additional steps would be expected before this finding is considered robust? Replication of the association in an independent Nigerian (or other African-ancestry) cohort; confirmation that the genomic inflation factor λ_{GC}

for the overall study is close to 1.0 (ruling out stratification as an alternative explanation, Section 18.5); and, ideally, fine-mapping (using the dense local LD structure, Chapter 11) and functional follow-up to distinguish the tested SNP from other, more strongly causal variants in LD with it.

(c) Comment on effect size. An odds ratio of 1.22 is a modest effect size, typical of common-variant GWAS findings for complex traits: it explains only a small fraction of the total heritable variation in hypertension risk (Chapter 9's heritability framework), consistent with the now widely accepted view that complex traits are influenced by very many variants of individually small effect (the **polygenic** picture, echoing Fisher's infinitesimal model introduced in Chapter 9), so that clinically useful genetic prediction for such traits in Nigerian populations will likely require **polygenic risk scores** aggregating very many loci — built, ideally, from Nigerian/African discovery data rather than transferred uncorrected from non-African studies (Section 14.6).

Theorem 18.2 (Expected Genomic Control Inflation Factor Under No Stratification)

If a genome-wide association study genuinely has no population stratification and no systematic genotyping artefacts, and the vast majority of tested markers have no true association with the phenotype, then the expected value of the genomic control inflation factor λ (the ratio of the observed median chi-square test statistic to the expected median of a chi-square(1) distribution) is 1.

Proof. Under the null hypothesis of no association at a given marker, and in the absence of any confounding population structure, the association test statistic (e.g., from Section 11.3 or 11.7) is asymptotically distributed as chi-square with 1 degree of freedom, by the general asymptotic theory of likelihood-based

(or score-based) tests established in Chapter 10. If the overwhelming majority of the m tested markers are truly null (a reasonable assumption for most complex traits, where only a small number of markers have genuine effects), then the empirical distribution of the m observed test statistics is, to close approximation, a sample from the chi-square(1) distribution. The median of a chi-square(1) distribution is a fixed, known constant (approximately 0.4549). By the consistency of the sample median as an estimator of the population median (a standard result from order statistics theory), the observed median test statistic converges to this same constant as m grows large, so their ratio, λ , converges to 1 in expectation. A value of λ substantially greater than 1 in real data is therefore attributable to a violation of one of the theorem's assumptions — most commonly, uncorrected population stratification (Chapter 14) — which is precisely why λ is used diagnostically in Section 18.9.

18.8: Chapter Summary

- GWAS test genome-wide panels of SNPs for association with disease/trait, exploiting linkage disequilibrium (Chapter 15) to indirectly capture most common genetic variation.
- The pipeline includes QC, population-structure adjustment (PCA), imputation, single-marker regression testing, and stringent multiple-testing correction (genome-wide significance, conventionally $p < 5 \times 10^{-8}$).
- Manhattan and Q–Q plots are the standard visual summaries; the genomic inflation factor diagnoses residual stratification.
- African, including Nigerian, populations present both distinctive challenges (denser marker/reference-panel requirements, greater substructure, historical under-

representation) and distinctive scientific opportunities (greater genetic diversity, finer LD-based fine-mapping resolution) for GWAS — motivating dedicated African genomics efforts such as H3Africa.

18.9: Computational Corner: Manhattan Plots, Q-Q Plots, and Genomic Control

R.

```
# Manhattan and Q-Q plots for simulated GWAS summary
# statistics (Section 18.5)
set.seed(42)
n_snp <- 5000
chr <- sort(sample(1:22, n_snp, replace = TRUE))
pos <- ave(chr, chr, FUN = function(x) seq_along(x))
pvals <- runif(n_snp)
pvals[sample(n_snp, 5)] <- runif(5, 1e-9, 1e-7) # plant a few
# "true hits"

# genomic control lambda
chisq_stat <- qchisq(1 - pvals, df = 1)
lambda_gc <- median(chisq_stat) / qchisq(0.5, df = 1)
cat("Genomic control lambda:", lambda_gc, "\n")

plot(-log10(pvals), col = chr %% 2 + 2, pch = 20,
     xlab = "SNP index (chromosome order)", ylab = expression(-
log[10](p)))
abline(h = -log10(5e-8), col = "red", lty = 2) # genome-wide
# significance

qqplot(-log10(ppoints(n_snp)), -log10(sort(pvals)),
     xlab = "Expected -log10(p)", ylab = "Observed -log10(p)")
abline(0, 1, col = "red")
```

Python.

```

import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

rng = np.random.default_rng(42)
n_snp = 5000
chrom = np.sort(rng.integers(1, 23, n_snp))
pvals = rng.uniform(size=n_snp)
pvals[rng.choice(n_snp, 5, replace=False)] =
rng.uniform(1e-9, 1e-7, 5)

chisq_stat = chi2.isf(pvals, df=1)
lambda_gc = np.median(chisq_stat) / chi2.ppf(0.5, df=1)
print("Genomic control lambda: ", lambda_gc)

plt.scatter(range(n_snp), -np.log10(pvals), c=chrom %2,
cmap="coolwarm", s=5)
plt.axhline(-np.log10(5e-8), color="red", linestyle="—")
plt.xlabel("SNP index"); plt.ylabel("-log10(p)")

expected = -np.log10(np.linspace(1/n_snp, 1, n_snp))
observed = -np.log10(np.sort(pvals))
plt.figure()
plt.scatter(expected, observed, s=5)
plt.plot([0, max(expected)], [0, max(expected)], color="red")

```

Both scripts reproduce, on simulated data with a handful of planted “true” associations, the two standard GWAS diagnostic plots of Section 18.5 — the Manhattan plot, which visually locates genome-wide significant peaks against the 5×10^{-8} threshold, and the Q–Q plot together with the genomic control inflation factor λ_{GC} , which checks whether the bulk of the genome-wide test statistics follow their null distribution, a critical check for undetected population stratification (Chapter 18) before any peak is taken seriously.

Mini-Project

Mini-Project 18: A Complete Mini-GWAS from Simulation to Manhattan Plot

Run a complete, simplified GWAS pipeline end to end on simulated genome-wide data.

- Step 1:** Simulate genotype data for 5,000 SNPs across several chromosomes in 1,000 cases and 1,000 controls, with one true causal SNP.
- Step 2:** Test every SNP for association using logistic regression (Section 11.7).
- Step 3:** Produce a Manhattan plot (Section 18.5) of $-\log_{10}(p)$ against genomic position.
- Step 4:** Produce a Q-Q plot and compute the genomic control inflation factor λ (Section 18.9).
- Step 5:** Determine whether your true causal SNP surpasses genome-wide significance (Section 18.4) and discuss what sample size or effect size would be needed if it does not.

Deliverable: *Full pipeline code, a Manhattan plot, a Q-Q plot with λ , and a short written interpretation.*

18.10: Practice Questions

- Q1.** A GWAS tests 600,000 SNPs. Using the standard Bonferroni approach with an assumed 1,000,000 effectively independent tests genome-wide (Section 18.4), what p-value threshold should be used for genome-wide significance, and would a SNP with $p = 2 \times 10^{-7}$ from this study meet that threshold?
- Q2.** A Q-Q plot from a Nigerian GWAS shows the bulk of observed p-values systematically more significant than expected under the null, across almost all SNPs genome-wide (not just a handful of top hits), with $\lambda_{GC} = 1.35$.

What does this most likely indicate, and what step in the analytical pipeline (Section 18.3) should be revisited?

- Q3.** Explain why a genotyping array and imputation reference panel developed and validated mainly in European-ancestry samples may perform less well in a Nigerian GWAS than in a European GWAS, referencing Chapter 15's discussion of LD.
- Q4.** A lead GWAS SNP for T2DM has odds ratio 1.15. A clinician asks whether this single SNP can be used, on its own, to usefully predict an individual Nigerian patient's T2DM risk. How would you respond, and what alternative approach would you suggest?
- Q5.** Sketch (in words) what a Q–Q plot would look like for a genome-wide scan that suffers from severe, uncorrected population stratification but contains no true associations at all, and contrast this with the Q–Q plot of a well-controlled study that does contain a handful of true, strong associations.

