

12

Candidate Gene Association Studies

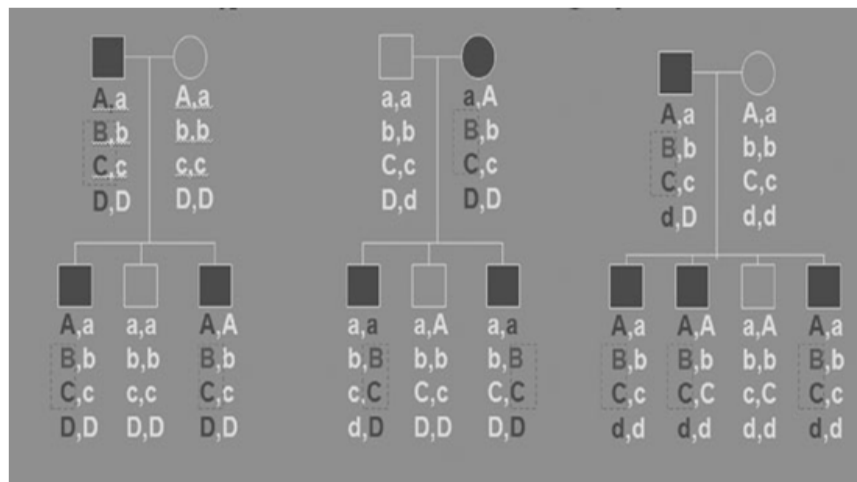
12.1: Learning Objectives

By the end of this chapter, the reader should be able to explain the rationale and design of a candidate-gene association study, distinguish it from genome-wide approaches (Chapter 18), describe how candidate genes and SNPs are selected, discuss the advantages and well-documented pitfalls of the candidate-gene approach (including the replication crisis of the 2000s), and carry out a candidate-gene regression analysis with an appropriate multiple-testing correction in R and Python.

12.2: What Is a Candidate-Gene Study?

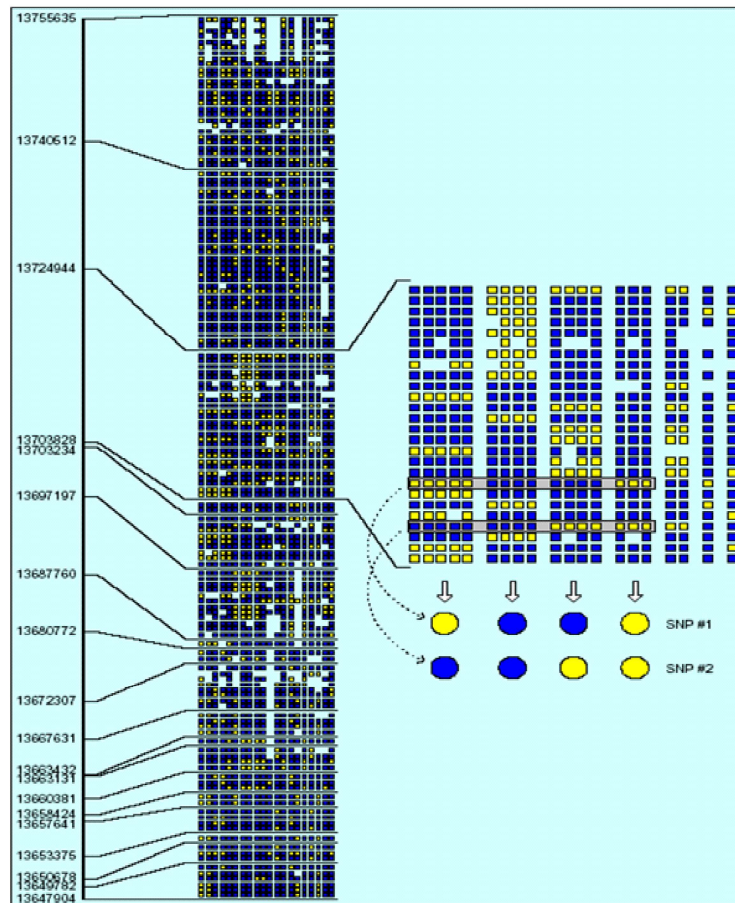
A **candidate-gene association study** tests one or a small number of genes — selected in advance because of known or suspected biological relevance to the trait or disease under study — for statistical association with that trait, using the same basic statistical machinery (chi-square tests, logistic regression) developed in Chapter 11. Candidates are typically chosen from one or more of the following sources: (i) genes encoding proteins in a biologically plausible pathway (e.g., genes of the renin–angiotensin system for hypertension); (ii) genes previously implicated by linkage analysis (Chapter 17) as lying within a chromosomal region linked to the trait; (iii) genes highlighted by animal-model or functional studies; or (iv) genes already reported in the published literature, often from a different population, motivating a **replication study**.

This targeted design contrasts with the **genome-wide association study (GWAS)**, which tests hundreds of thousands to millions of markers spread across the entire genome without any prior assumption about which genes are relevant (Chapter 18). Historically, candidate-gene studies dominated human genetics from the early 1990s until the mid-2000s, when the falling cost of genome-wide genotyping arrays made GWAS the default design for common, complex traits.



A chromosome ideogram marking the physical locations of several candidate genes and their surrounding marker loci — the kind of genome-wide “parts list” from which candidate genes are selected for targeted study.

12.3: Selecting Candidate SNPs Within a Gene



A candidate gene's linkage-disequilibrium structure and physical map, used to select a minimal set of tag SNPs for genotyping.

Having settled on a candidate gene, the investigator must decide how much of it — and which specific variants — to genotype. Genotyping every possible SNP within and around a gene is neither necessary nor economical, because SNPs in strong linkage

disequilibrium (Chapter 15) carry largely redundant information. Practical strategies include:

- **Tagging SNPs:** choosing a minimal subset of SNPs such that every other common SNP in the region has r^2 above some threshold (often a higher threshold, e.g. $r^2 > 0.9$, than is typical in GWAS chip design, since fewer SNPs are being genotyped overall and denser coverage of a single small region is affordable) with at least one tag SNP — see Section 12.3 of the LD chapter’s tagging discussion (Chapter 15).
- **Functional SNPs:** variants with a plausible biological mechanism — non-synonymous coding changes, splice-site variants, or variants in known promoter or enhancer elements.
- **Previously reported SNPs:** variants already associated with the trait (or a related trait) in the literature, included specifically to attempt replication.
- **Minor allele frequency (MAF) thresholds:** very rare variants are usually excluded from a standard candidate-gene study (which typically has limited power for anything below MAF $H^* 0.05$) and are instead the target of the rare-variant and sequencing-based methods of Chapter 20.

The region genotyped is usually extended beyond the gene’s coding sequence to include the promoter, untranslated regions, and a margin (often 5–20 kb) upstream and downstream, since regulatory variants affecting expression may lie outside the transcribed sequence itself.

12.4: Advantages and Disadvantages of the Candidate-Gene Design

Advantages: A candidate-gene study genotypes far fewer markers than a GWAS, so it is considerably cheaper, and the multiple-testing burden (Section 12.5) is much smaller, giving greater statistical power per genotyped SNP for a fixed sample size — an

important practical consideration for studies of Nigerian and other African cohorts, where sample sizes are frequently smaller than those available in well-funded biobanks elsewhere.

Disadvantages: The approach is limited in scope by construction: a true causal gene that was not on the candidate list can never be discovered. Historically, this design also proved highly susceptible to publication of **false-positive findings** that failed to replicate — a well-documented crisis in the candidate-gene literature of the 1990s and 2000s, in fields ranging from psychiatric genetics to cardiovascular risk, traced to a combination of small sample sizes, inadequate correction for population stratification (Chapter 14), publication bias favouring positive results, and biologically plausible but ultimately spurious a priori hypotheses. For this reason, contemporary reviewers and journals require independent replication, ideally in more than one population, before a candidate-gene finding is accepted, and increasingly regard genome-wide-significant findings from unbiased GWAS or sequencing studies (Chapters 18 and 20) as more credible evidence than an isolated candidate-gene report.

Worked Example 12.1: (APOL1 and kidney disease risk in African populations)

A well-established and now widely replicated example of a genuine candidate-gene finding relevant to African populations is the association between coding variants in the *APOL1* gene (the G1 and G2 risk alleles, which arose under selective pressure from resistance to trypanosomiasis) and non-diabetic chronic kidney disease, an association discovered through a combination of linkage mapping in African-American families followed by targeted candidate-gene fine-mapping of the linked region. Because *APOL1* risk alleles are common in West African populations (including Nigeria) but virtually absent in European populations, this is a clear illustration of why candidate-gene (and GWAS) findings from one population cannot automatically be assumed

to generalise to another — the practical motivation for Chapter 14’s discussion of population substructure, and a caution against uncritically importing “risk alleles” identified elsewhere into Nigerian clinical practice without local validation.

12.5: The Multiple-Testing Problem in a Candidate-Gene Study

Even though a candidate-gene study tests far fewer markers than a GWAS, testing multiple SNPs within the same gene (and, often, multiple genes) still inflates the **family-wise type I error rate** if uncorrected. The simplest correction is the **Bonferroni correction**: for m independent tests at nominal significance level α , declare significance only if the per-test p-value is below α/m .

Theorem 12.1: (Bonferroni Inequality)

For any m hypothesis tests, each conducted at per-test significance level α/m , the probability of at least one false positive among the m tests (the family-wise error rate, FWER), under the global null hypothesis that none of the m associations is real, satisfies $\text{FWER} \leq \alpha$ — regardless of any dependence structure among the tests.

Proof. Let E_i denote the event that test i produces a false positive (rejects its null when that null is true), for $i = 1, \dots, m$, so $P(E_i) = \alpha/m$ for each i under the Bonferroni-corrected threshold. By Boole’s inequality (the union bound), for any collection of events,

$$P\left(\bigcup_{i=1}^m E_i\right) \leq \sum_{i=1}^m P(E_i).$$

The left-hand side is precisely the family-wise error rate (the probability that *at least one* of the m tests is a false positive). Substituting $P(E_i) = \alpha/m$:

$$\text{FWER} = P\left(\bigcup_{i=1}^m E_i\right) \leq \sum_{i=1}^m \frac{\alpha}{m} = m \times \frac{\alpha}{m} = \alpha.$$

Crucially, Boole's inequality requires no assumption about independence or any other dependence structure among the events E_i , so this bound holds universally — whether the m SNPs are in strong LD (Chapter 15), completely independent, or anything in between. ■

Because the Bonferroni inequality holds regardless of dependence, it is always *valid* — but, as the proof makes clear, it can be very *conservative* (far below the true FWER of α) whenever the tests are strongly positively correlated (as with SNPs in high LD), since the union bound then substantially overstates the true probability of at least one false positive; this is precisely why the permutation-based alternative of this section, which directly accounts for the actual dependence structure in the data, achieves the same α -level FWER control with less loss of power in an LD-rich candidate-gene region.

Worked Example 12.2 (Bonferroni correction for a Nigerian hypertension candidate-gene panel).

A Nigerian cardiovascular genetics unit tests 12 SNPs across three candidate genes (AGT, ACE, and NOS3, all previously implicated in blood pressure regulation) for association with hypertension in a Lagos-based case-control sample. To control the family-wise error rate at the conventional $\alpha = 0.05$ using Theorem 12.1, each individual SNP must meet the Bonferroni-corrected threshold of $\alpha/m = 0.05/12 \approx 0.00417$. Suppose the smallest observed p-value, for a NOS3 variant, is 0.0021 — this passes the Bonferroni threshold ($0.0021 < 0.00417$) and would be declared significant, with the FWER guaranteed (by Theorem 12.1) not to exceed 5% across all 12 tests, regardless of how correlated the SNPs

Theorem 12.2 (The Bonferroni Bound on Family-Wise Error Rate)

If m hypothesis tests are each conducted at significance level α/m , then the probability of at least one false positive among all m tests (the family-wise error rate) is at most α , regardless of the dependence structure among the tests.

Proof. Let $A_1, \dots, A_m$ denote the events that test i produces a false positive (rejects a true null hypothesis), with $P(A_i) \leq \alpha/m$ for each i by construction. By Boole's inequality (the union bound), for any collection of events, $P(A_1 \cup A_2 \cup \dots \cup A_m) \leq \sum_{i=1}^m P(A_i)$. Substituting the per-test bound gives $P(\text{at least one false positive}) \leq \sum_{i=1}^m (\alpha/m) = m \text{ times } (\alpha/m) = \alpha$. Crucially, Boole's inequality requires no assumption about independence or any other dependence structure among the events A_i , so this bound holds even when the m genetic markers being tested are in strong linkage disequilibrium with each other (Section 12.5) — a property that also explains why the Bonferroni correction, while always valid, can be conservative (overly strict) precisely in the correlated-marker setting typical of a candidate-gene study, motivating the permutation-based alternative of Section 12.7.

12.6: Chapter Summary

- A candidate-gene study tests a small, pre-specified set of genes chosen for biological plausibility, prior linkage evidence, or previous literature reports.
- Tag-SNP selection, functional annotation, and MAF thresholds guide which variants within a candidate gene are genotyped.
- The design is cheaper and better powered per marker than GWAS, but is limited to the genes considered and has a

well-documented history of non-replicating false positives; independent replication is essential.

- The *APOL1*–kidney disease association illustrates both a genuine, population-specific candidate-gene finding and the need for local validation of any risk allele before it is applied in a Nigerian or West African clinical setting.
- Multiple testing within a gene is usually corrected with a Bonferroni bound or, more accurately given LD among SNPs, a permutation test.

12.7: Computational Corner: Candidate-Gene Regression with Permutation-Based Multiple Testing

R.

```
# Simulated candidate-gene study: 5 tag SNPs genotyped (0/1/2
# coded), one trait
set.seed(55)
n <- 400
snp <- matrix(rbinom(n * 5, 2, prob = 0.3), nrow = n, ncol = 5)
colnames(snp) <- paste0("SNP", 1:5)
beta_true <- c(0, 0, 0.35, 0, 0) # only SNP3 truly
affects the trait
trait <- 120 + snp %*% beta_true + rnorm(n, 0, 5) # e.g.,
systolic BP (mmHg)

pvals <- sapply(1:5, function(j) summary(lm(trait ~ snp[,
j]))$coefficients[2, 4])
names(pvals) <- colnames(snp)
print(pvals)

# Bonferroni correction
bonferroni_sig <- pvals < 0.05/5
print(bonferroni_sig)
```

```

# Permutation-based family-wise p-value for the minimum
observed p-value
n_perm <-2000
min_p_obs <-min(pvals)
perm_min_p <-replicate(n_perm, {
  trait_perm <-sample(trait)
  min(sapply(1:5, function(j) summary(lm(trait_perm ~ snp[,
j]))$coefficients[2, 4]))
})
empirical_p <-mean(perm_min_p <= min_p_obs)
cat("Empirical (permutation) family-wise p-value:",
empirical_p, "\n")

```

Python.

```

import numpy as np
import statsmodels.api as sm

rng = np.random.default_rng(55)
n =400
snp = rng.binomial(2, 0.3, size=(n, 5))
beta_true = np.array([0, 0, 0.35, 0, 0])
trait =120+ snp @ beta_true + rng.normal(0, 5, n)

def snp_pvalues(trait_vec, snp_mat):
    pvals = []
    for j in range(snp_mat.shape[1]):
        X = sm.add_constant(snp_mat[:, j])
        model = sm.OLS(trait_vec, X).fit()
        pvals.append(model.pvalues[1])
    return np.array(pvals)

pvals = snp_pvalues(trait, snp)
print(pvals)
print("Bonferroni significant:", pvals <0.05/5)

```

```

n_perm = 2000
min_p_obs = pvals.min()
perm_min_ps = []
for _ in range(n_perm):
    trait_perm = rng.permutation(trait)
    perm_min_ps.append(snp_pvalues(trait_perm, snp).min())
perm_min_ps = np.array(perm_min_ps)
empirical_p = np.mean(perm_min_ps <= min_p_obs)
print("Empirical (permutation) family-wise p-value:",
      empirical_p)

```

In this simulated study, only SNP3 was given a genuine effect on the trait; both the Bonferroni-corrected threshold and the permutation-based family-wise p-value should correctly flag SNP3 while controlling the overall false-positive rate across all five tested SNPs, illustrating in miniature the multiple-testing discipline required of every candidate-gene report before it is accepted as more than a chance finding (Section 12.4).

Mini-Project

Mini-Project 12: A Full Candidate-Gene Study with Multiple Testing Correction

Simulate a candidate-gene study covering multiple SNPs within one gene and correctly apply a multiple-testing correction appropriate to their correlation structure.

Step 1: Simulate genotypes for 10 correlated SNPs (in modest linkage disequilibrium with each other) across 600 case-control individuals, with one true causal SNP.

Step 2: Test each SNP for association individually.

Step 3: Apply both a Bonferroni correction and a permutation-based correction (Section 12.7) accounting for the SNPs' correlation.

Step 4: Compare how many SNPs remain significant under each correction method.

Step 5. Write a short paragraph explaining why the permutation-based correction differs from Bonferroni when SNPs are correlated.

Deliverable: *Simulation code, per-SNP test results, and a comparison of Bonferroni vs. permutation-based significant SNP counts.*

12.8: Practice Questions

- Q1.** A candidate-gene study genotypes 8 SNPs in the *TCF7L2* gene region for association with type-2 diabetes in a Lagos cohort. At what per-SNP p-value threshold would a Bonferroni correction declare significance at an overall $\alpha = 0.05$?
- Q2.** Explain, in two or three sentences, why a permutation-based multiple-testing correction is generally preferred over a simple Bonferroni correction when the tested SNPs are in strong linkage disequilibrium with one another.
- Q3.** Give two reasons, distinct from multiple testing, why candidate-gene studies of the 1990s and 2000s produced an unusually high proportion of findings that failed to replicate.
- Q4.** Why does the high frequency of *APOL1* risk alleles in West African populations, contrasted with their near-absence in European populations, illustrate a general limitation of transferring “risk allele” findings across populations without local validation?
- Q5.** Using the R or Python code of Section 12.7, change *beta_true* so that two SNPs (say SNP2 and SNP4) have a true non-zero effect, re-run the simulation, and report whether the Bonferroni correction and the permutation test both correctly identify both SNPs.

- Q6.** A candidate-gene study genotypes 6 tag SNPs, chosen to represent 40 total common SNPs in the gene region at $r^2 > 0.9$. Explain why testing only the 6 tag SNPs, rather than all 40 SNPs, does not meaningfully reduce the study's ability to detect a true association somewhere in the gene, provided the causal variant is itself common.

