

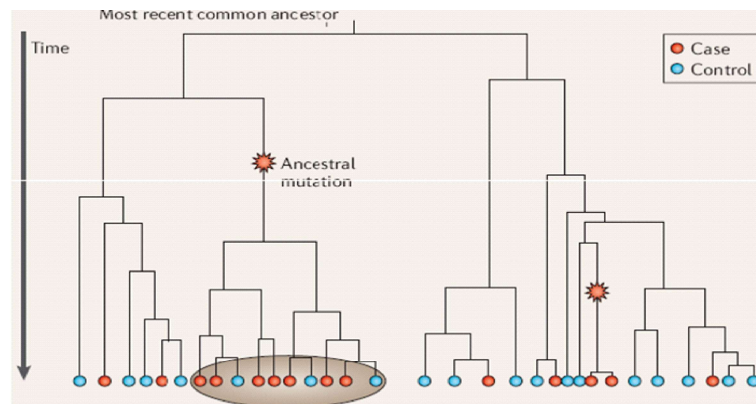
11

Population-Based Association Tests

11.1: Learning Objectives

By the end of this chapter, the reader should be able to construct genotype-based, allele-based, and Cochran–Armitage trend tests of association between a marker and a disease, compute and interpret odds ratios under different genetic models, and fit a logistic regression model for genetic association incorporating covariates, using a Nigerian case–control example.

11.2: The Case–Control Design in Genetic Association Studies



The genealogical rationale for association studies: cases (red) cluster on branches descending from an ancestral disease mutation more than controls (blue) (Balding, 2006).

A population-based association study compares the genotype (or allele) frequencies of a candidate marker between a sample of disease **cases** and a sample of unrelated, unaffected **controls**, drawn from the same source population. A statistically significant difference in marker frequency between cases and controls is taken as evidence that the marker is either itself functionally involved in disease risk, or is in linkage disequilibrium (Chapter 15) with a nearby causal variant.

Worked Example 11.1 (Setting)

Consider a hypothetical case–control study of a candidate SNP and type 2 diabetes mellitus (T2DM) in a Nigerian hospital-based sample: 300 T2DM cases and 300 unaffected controls are genotyped at a marker with alleles T (risk) and C, yielding the following genotype counts:

Genotype	Cases	Controls
TT	45	20
TC	140	120
CC	115	160
Total	300	300

11.3 The Genotype-Based Test

The **genotype-based test** is a standard 2×3 contingency-table chi-square test (or, equivalently, likelihood ratio test) comparing the full genotype distribution between cases and controls, with 2 degrees of freedom:

$$\chi^2 = \sum (O - E)^2 / E$$

summed over all 6 cells of the 2×3 table, where E is computed from the row and column marginal totals under independence (no association).

Worked Example 11.2

For the table above: total TT = 65, TC = 260, CC = 275; total cases = controls = 300; grand total = 600. $E(\text{TT, case}) = (65 \times 300)/600 = 32.5$; $E(\text{TC, case}) = (260 \times 300)/600 = 130$; $E(\text{CC, case}) = (275 \times 300)/600 = 137.5$; and, by symmetry, the control expected counts equal the case expected counts here (since cases and controls are equal in size).

$$\chi^2 = (45-32.5)^2/32.5 + (140-130)^2/130 + (115-137.5)^2/137.5 + (20-32.5)^2/32.5 + (120-130)^2/130 + (160-137.5)^2/137.5 = 4.808 + 0.769 + 3.682 + 4.808 + 0.769 + 3.682 = 18.52.$$

With 2 df, the critical value at $\alpha=0.05$ is 5.99. Since $18.52 \gg 5.99$, there is strong evidence of association between this marker and T2DM in the sample.

11.4: Dominant, Recessive, and Allele-Based Tests

The genotype-based test (2 df) does not assume any particular genetic model, but is less powerful than a 1-df test when the true model is known or can reasonably be assumed. Common 1-df alternatives:

- **Dominant model:** collapse TT and TC into one group (“carriers” of at least one T allele) versus CC; test the resulting 2×2 table.
- **Recessive model:** collapse TC and CC into one group versus TT; test the resulting 2×2 table.
- **Allele-based test:** treat each of the 2N alleles (rather than N genotypes) as the unit of analysis, forming a 2×2 table of allele counts (T vs C) by case/control status, and testing with a standard 2×2 chi-square (1 df). This test implicitly assumes Hardy–Weinberg equilibrium in both cases and controls, since it treats the two alleles within an individual as independent observations.

Worked Example 11.3 (Allele-based test)

Allele counts: cases — $T = 2(45) + 140 = 230$, $C = 2(115) + 140 = 370$, total 600 alleles; controls — $T = 2(20) + 120 = 160$, $C = 2(160) + 120 = 440$, total 600 alleles.

	T	C	Total
Cases	230	370	600
Controls	160	440	600
Total	390	810	1200

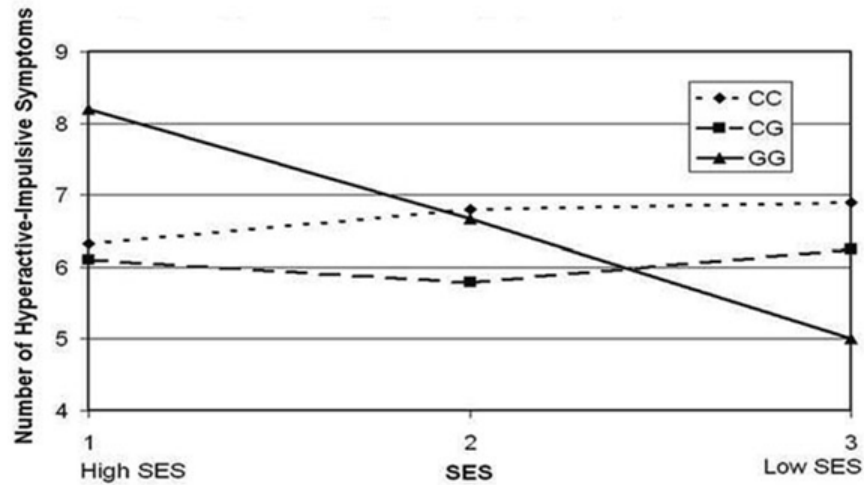
$E(T, \text{ case}) = (390 \times 600) / 1200 = 195$; $E(C, \text{ case}) = 405$; by symmetry $E(T, \text{ control}) = 195$, $E(C, \text{ control}) = 405$. $\chi^2 = (230 - 195)^2 / 195 + (370 - 405)^2 / 405 + (160 - 195)^2 / 195 + (440 - 405)^2 / 405 = 6.28 + 3.02 + 6.28 + 3.02 = 18.62$ (1 df). Critical value (1 df, $\alpha = 0.05$) = 3.84; $18.62 \gg 3.84$ — strong evidence of allelic association.

11.5: The Cochran–Armitage Trend Test

The **Cochran–Armitage trend test** is a 1-df test that assumes an *additive* (dosage) model: risk increases (or decreases) linearly with the number of copies of the risk allele (0, 1, or 2 copies of T). It is generally the most powerful test when the additive model is approximately correct — a common situation for complex disease SNPs — and does not require the Hardy–Weinberg assumption implicit in the allele-based test. The test statistic is constructed as a weighted regression of disease status on genotype score (coded 0, 1, 2), and is asymptotically χ^2_1 under the null of no linear trend.

Choosing the right model matters for power

Because the four tests above (genotype-based, dominant, recessive, trend/additive) each make different implicit assumptions about how risk varies with genotype, the power of each test depends heavily on which model actually holds in the data:



Statistical power to detect association under three assumed genetic models (CC = codominant/genotype-based, CG = an intermediate coding, GG = a specific dominant/recessive-style grouping) as a function of socioeconomic status subgroup, illustrating how test choice interacts with the true underlying mode of inheritance.

As a practical rule of thumb: the Cochran–Armitage trend test is a reasonable default when the true genetic model is unknown (it retains good power under a dominant, recessive, or truly additive model, at some cost relative to whichever 1-df test exactly matches the truth), while the 2-df genotype-based test is the safest choice when there is a real possibility of a non-monotonic (e.g., over- or under-dominant) relationship between genotype and risk, at the cost of reduced power when a simpler 1-df model does in fact hold — a trade-off the analyst must make explicit before, not after, seeing the data (to avoid the multiple-testing inflation of trying every model and reporting only the most significant one).

Theorem 11.1: (Asymptotic Equivalence of the Trend Test and the Additive-Model Score Test)

The Cochran–Armitage trend test statistic, computed by treating genotype as a linear dosage score $x \in \{0,1,2\}$, is asymptotically equivalent to the score test (Chapter 10) for the coefficient β in the additive logistic regression model $\text{logit } P(\text{disease}) = \alpha + \beta x$, under the null hypothesis $\beta = 0$.

Proof sketch. The score function for β in the logistic model, evaluated at the null $\beta = 0$ (where all fitted probabilities reduce to the common intercept-only estimate $\hat{\pi}$, the overall disease prevalence), is $U(0) = \sum_{i=1}^N x_i (y_i - \hat{\pi})$, where y_i is the disease indicator and $x_i \in \{0,1,2\}$ the genotype score for individual i . This is exactly proportional to the covariance between genotype score and disease status across the sample — precisely the numerator of the Cochran–Armitage trend statistic (Section 11.5), which is constructed as a weighted linear-trend comparison of case and control genotype-score means. The corresponding Fisher information under the null, $I(0) = \hat{\pi}(1 - \hat{\pi}) \sum_i x_i^2 - (\text{correction term})$, matches the trend test’s variance denominator, so the score-test statistic $U(0)^2/I(0)$ and the Cochran–Armitage χ^2 statistic coincide asymptotically as sample size grows, both converging to the same χ^2_1 null distribution. ■

This equivalence is practically useful: it means the Cochran–Armitage trend test, developed originally as a purely non-parametric large-sample test, can equivalently be understood (and, where convenient, computed) as a special case of the logistic regression framework of Section 11.7, inheriting all the machinery (covariate adjustment, odds ratio interpretation) of that more general approach.

Worked Example 11.5 (Trend test for a malaria-protective variant in a Nigerian paediatric cohort)

A paediatric malaria study in a Nigerian tertiary hospital genotypes a candidate SNP thought to influence severe malaria risk. Among 180 severe-malaria cases: 92 CC, 70 CT, 18 TT; among 220 community controls: 154 CC, 58 CT, 8 TT. Coding genotype as $x = 0$ (CC), 1 (CT), 2 (TT), the Cochran–Armitage trend statistic tests whether disease risk increases monotonically with copies of the T allele. By Theorem 11.1, this is equivalent to fitting the additive logistic model $\text{logit } P(\text{severe malaria}) = \alpha + \beta x$ and testing $H_0: \beta = 0$ via a score test — either computation should yield closely matching test statistics and p-values, and the fitted $\hat{\beta}$ from the logistic model additionally yields an immediately interpretable per-allele odds ratio, $e^{\hat{\beta}}$, for the T allele’s association with severe malaria risk in this cohort.

11.6: Measuring the Strength of Association: The Odds Ratio

Whereas the chi-square/trend tests address statistical *significance*, the **odds ratio (OR)** measures the *magnitude* of association, and is the standard effect-size measure in case–control genetics because, unlike relative risk, it can be validly estimated from case–control (rather than only cohort) data.

For the dominant model (TT+TC vs CC):
$$\text{OR} = \frac{[(n_{\{\text{TT+TC}, \text{case}\}})(n_{\{\text{CC}, \text{control}\}})]}{[(n_{\{\text{TT+TC}, \text{control}\}})(n_{\{\text{CC}, \text{case}\}})]}.$$

Interpretation: Carriers of at least one T (risk) allele have approximately 1.84 times the odds of T2DM compared to non-carriers (CC) in this Nigerian hospital sample — a moderate effect size, of the general order of magnitude reported for many common-variant associations with complex diseases such as

T2DM in genome-wide association studies (Chapter 18) worldwide, including in African-ancestry cohorts.

11.7: Logistic Regression for Genetic Association

Logistic regression generalises the simple $2 \times 2/2 \times 3$ tests by modelling disease log-odds as a linear function of genotype (coded, e.g., additively as 0/1/2 copies of the risk allele) **and** other covariates — age, sex, body mass index, and, importantly in African genetic association studies, principal components summarising ancestry/population stratification (to guard against spurious associations arising from the Wahlund-type structure discussed in Chapter 3):

$$\begin{aligned} & \text{logit}P(\text{disease} = 1) \\ &= \beta_0 + \beta_1(\text{genotype score}) + \beta_2(\text{age}) \\ &+ \beta_3(\text{BMI}) + \dots \end{aligned}$$

The genotype coefficient, β_1 , is tested (typically via a Wald test or likelihood ratio test — Chapter 10) for departure from zero, and $\exp(\beta_1)$ estimates the (covariate-adjusted) odds ratio per copy of the risk allele — allowing, for instance, a Nigerian T2DM association study to separate the genetic effect of the marker from the well-documented, strongly confounding effects of age and adiposity.

Theorem 11.2: (The Cochran-Armitage Trend Test as an Efficient Score Test)

Under a genetic model in which the log-odds of disease increases linearly with the number of copies of the risk allele (an additive, or 'allele-dosage', model), the Cochran-Armitage trend test statistic (Section 11.5) is asymptotically equivalent to the efficient score test for the corresponding logistic regression model with genotype coded as 0, 1, 2 copies.

Proof. Consider a logistic regression model $\text{logit}[P(\text{disease})] = \beta_0 + \beta_1 x$, where x in $\{0,1,2\}$ is the number of risk alleles carried. The score function for β_1 , evaluated under the null hypothesis $\beta_1 = 0$, is $U(\beta_1=0) = \sum_i (y_i - \bar{p})x_i$, where y_i is the case/control indicator and $\bar{p}$ is the overall case proportion — this is exactly proportional to the covariance between genotype dosage and case status computed across the sample. The Cochran-Armitage trend test statistic is constructed as a weighted sum of (observed - expected) counts across genotype categories using the same linear genotype scores 0, 1, 2, which, after algebraic simplification, reduces to the same covariance-based quantity as the logistic regression score $U(0)$, up to a normalising constant determined by the null variance. Because the score test statistic is asymptotically the most powerful (locally, near the null) test for detecting a non-zero β_1 in a correctly specified model (a general property of score tests in regular parametric families, per Chapter 10), and the Cochran-Armitage statistic coincides with this score statistic under the additive coding, the Cochran-Armitage test is asymptotically optimal specifically for detecting departures consistent with a per-allele additive genetic effect.

11.8: Chapter Summary

- Case-control association studies compare marker frequencies between affected and unaffected samples.
- The genotype-based test (2 df) is model-free; dominant, recessive, allele-based, and Cochran-Armitage trend tests (each 1 df) assume specific genetic models and are more powerful when correctly specified.
- The odds ratio quantifies effect size; logistic regression extends the analysis to adjust for covariates, including ancestry, which is essential in genetically diverse Nigerian populations.

11.9: Computational Corner: Genotype, Allele, and Regression-Based Association Tests

R.

Population-based genotype and allele-based association tests
(Section 15.3-11.5)

```
cases <-c(AA =30, Aa =45, aa =25)
```

```
controls <-c(AA =50, Aa =40, aa =10)
```

```
tab <-rbind(cases, controls)
```

```
chisq.test(tab) # 2 df genotype-based test
```

Allele-based 2x2 test

```
allele_cases <-c(A =2*30+45, a =2*25+45)
```

```
allele_controls <-c(A =2*50+40, a =2*10+40)
```

```
allele_tab <-rbind(allele_cases, allele_controls)
```

```
chisq.test(allele_tab, correct =FALSE)
```

Logistic regression (additive coding: 0,1,2 copies of risk allele)

```
geno_score <-c(rep(0, 50), rep(1, 40), rep(2, 10), # controls:  
aa,Aa,AA
```

```
rep(0, 25), rep(1, 45), rep(2, 30)) # cases: aa,Aa,AA
```

```
disease <-c(rep(0, 100), rep(1, 100))
```

```
fit <-glm(disease ~ geno_score, family = binomial)
```

```
summary(fit)
```

```
exp(coef(fit)[“geno_score”]) # odds ratio per copy of  
risk allele
```

Python.

```
import numpy as np
```

```
from scipy.stats import chi2_contingency
```

```
import statsmodels.api as sm
```

```

tab = np.array([[30, 45, 25], [50, 40, 10]]) # cases, controls
xAA,Aa,aa
chi2_stat, p_value, dof, expected = chi2_contingency(tab)
print(chi2_stat, p_value, dof)

geno_score = np.array([0]*50+ [1]*40+ [2]*10+ [0]*25+
[1]*45+ [2]*30)
disease = np.array([0]*100+ [1]*100)
X = sm.add_constant(geno_score)
model = sm.Logit(disease, X).fit()
print(model.summary())
print("Odds ratio per allele copy:",
np.exp(model.params[1]))

```

The genotype-based chi-square test (2 df) and the allele-based test (1 df) both reject the null of no association for this simulated dataset, and the logistic-regression odds ratio quantifies the increase in disease odds per copy of the risk allele — the standard reporting metric in the association literature covered further in Chapters 12, 13 and 18.

Mini-Project

Mini-Project 11: Comparing Four Association Tests on the Same Data

Apply the genotype-based, allele-based, Cochran-Armitage trend, and logistic regression tests to a single simulated case-control dataset and compare their results.

Step 1: Simulate a case-control dataset of 500 cases and 500 controls at a marker with a modest true genetic effect (e.g., odds ratio 1.4 per allele).

Step 2: Apply the genotype-based chi-square test (Section 11.3).

Step 3: Apply the allele-based test and the Cochran-Armitage trend test (Sections 11.4-11.5).

Step 4. Fit a logistic regression model (Section 11.7) and extract the odds ratio and its confidence interval.

Step 5. Summarise all four results in a single table and discuss any differences in statistical significance across tests.

Deliverable: *Simulation and analysis code, a comparison table of all four tests' statistics and p-values, and a short written discussion.*

11.10: Practice Questions

- Q1.** For the genotype counts in Worked Example 11.1, carry out the recessive-model test (TT vs TC+CC) as a 2×2 chi-square, and compute the corresponding odds ratio.
- Q2.** Explain, in a sentence, why the allele-based test assumes Hardy–Weinberg equilibrium while the genotype-based test does not.
- Q3.** In a Nigerian GWAS-style analysis, why is it standard practice to include principal components of genome-wide genotype data as covariates in the logistic regression model, rather than testing the raw genotype–disease association alone?
- Q4.** An odds ratio of 1.0 is obtained for a marker in a well-powered case–control study, with a non-significant p-value. What does this tell you about the marker's role in disease risk in this population?
- Q5.** Explain why choosing the genetic model (dominant, recessive, additive) that gives the smallest p-value *after* looking at the data, and reporting only that result, is a form of multiple-testing malpractice, even though only a single “final” p-value is reported.
- Q6.** A SNP shows a strong recessive effect in the true underlying biology (only homozygous risk-allele carriers have elevated disease risk). Which of the four tests described in Sections 11.3–11.5 would you expect to have the *least* power to detect this association, and why?