

APPENDIX D

Master Answer Key to End-of-Chapter Exercises

Full worked solutions to every end-of-chapter practice question are collected here, organised by chapter, so that each chapter's main text can flow directly from its Computational Corner into its exercises without interruption. Where a question is computational, a short verification script in both R and Python is given immediately after the worked solution; purely conceptual or explanatory questions are answered in prose alone.

Chapter 1: Basic Biological Concepts in Genetics

Answer to Q1. Genotype is the pair of alleles carried at the locus (e.g., AA, AS, or SS at HBB); phenotype is the observable outcome (e.g., normal red cells, sickle-cell trait, or sickle-cell disease) that results from the genotype.

Answer to Q2. (a) The baby's phenotype is sickle-cell disease (SCD). (b) The baby is homozygous, since both alleles at the locus are S.

Answer to Q3. (a) A male has only one X chromosome and therefore only one allele at an X-linked locus; he is hemizygous, so he is either affected or unaffected — there is no heterozygous “carrier” state possible for him at that locus. (b) Because G6PD is X-linked

recessive, a male needs only one copy of the deficient allele to be affected, whereas a female needs two copies (one from each parent); since the deficient allele is not extremely rare in many Nigerian populations, hemizygous males are affected far more often than homozygous-deficient females.

Answer to Q4. The mother contributes I^A or i (each with probability $1/2$); the father contributes I^B or i (each with probability $1/2$). Possible offspring genotypes: $I^A I^B$ (phenotype AB), $I^A i$ (phenotype A), $I^B i$ (phenotype B), and ii (phenotype O), each with probability $1/4$.

Answer to Q5. A genetic marker is a polymorphic DNA segment (or, classically, a polymorphic protein/antigen) of known chromosomal location, used as a reference point in mapping. Example: the ABO or Rhesus blood group systems, both widely typed in Nigerian blood banks and used historically as genetic markers before DNA-based markers became available.

Answer to Q6. A microsatellite with only three alleles distinguished purely by repeat length carries no information about which specific DNA sequence change is functionally responsible for a disease — it is simply a polymorphic “signpost,” not a candidate functional variant itself, so genotyping it cannot directly reveal the causal mutation. However, because it is highly polymorphic (many distinguishable alleles, giving high heterozygosity and hence high statistical power to detect co-segregation with a trait in a pedigree), and because its chromosomal location is precisely known, it is an excellent tool for **linkage mapping** (Chapter 17): tracking whether a particular microsatellite allele is co-inherited with a disease

within families localises the unknown causal gene to the chromosomal region near the marker, even though the marker itself is never the cause.

Answer to Q7. For a cross of AB $\times$ OO: the AB parent contributes allele I^A or I^B (each with probability 1/2); the OO parent contributes only allele i (probability 1). Offspring genotypes: I^A i (phenotype A) with probability 1/2, and I^B i (phenotype B) with probability 1/2 — no AB and no O offspring are possible. Running the simulation code of Section 1.9 with mother = ["A","B"] and father = ["O","O"] should confirm approximately a 50:50 split between phenotypes A and B, with zero simulated AB or O offspring.

```

mother <-c("A", "B"); father <-c("O", "O")
offspring <-replicate(2000, c(sample(mother, 1),
sample(father, 1)))
table(apply(t(offspring), 1, function(g) if (all(g
%in%c("A","O")) "A"elseif (all(g %in%c("B","O"))
"B"elseif (setequal(g, c("A","B"))) "AB"else"O"))

```

Chapter 2: Mendel's Laws of Inheritance

Answer to Q1. Let X = number of SS children among 4, $X \sim \text{Binomial}(n = 4, p = 1/4)$. $P(X = 1) = C(4,1) (1/4)^1 (3/4)^3 = 4 \times 0.25 \times 0.421875 = 0.4219$, i.e., about 42.2%.

Answer to Q2. At HBB: father AS, mother AA $\Rightarrow P(\text{child AS}) = 1/2$. At ABO: father $I^A B$, mother $I^A i$ $\Rightarrow$ offspring genotypes $I^A I^A$, $I^A i$, $I^B i$, ii each with probability $1/4$; only ii gives blood group O, so $P(\text{group O}) = 1/4$. By independent assortment, $P(\text{AS and group O}) = 1/2 \times 1/4 = 1/8 = 0.125$.

Answer to Q3. Expected counts under 3:1: $E_{\text{dominant}} = 150$, $E_{\text{recessive}} = 50$. $\chi^2 = (160-150)^2/150 + (40-50)^2/50 = 100/150 + 100/50 = 0.667 + 2.000 = 2.667$. Critical value ($df = 1$, $\alpha = 0.05$) = 3.84. Since $2.667 < 3.84$, we fail to reject H_0 — the data are consistent with a 3:1 ratio.

Answer to Q4. At the molecular level, both haemoglobin A and haemoglobin S are produced and can be detected (e.g., by electrophoresis) in an AS individual — this is co-dominance. At the clinical level, however, an AS individual does not manifest the disease phenotype (sickling crises, chronic haemolysis) that only appears when both alleles are S — so with respect to the *disease* phenotype, S behaves as recessive to A.

R and Python verification (Q1, Q3).

```
dbinom(1, size =4, prob =1/4)           # Q1:  $P(X = 1)$  SS
      child of 4)
```

```
chisq.test(c(160, 40), p =c(0.75, 0.25))   # Q3:
      3:1 goodness-of-fit
```

```
fromscipy.statsimport binom, chisquare
print(binom.pmf(1, 4, 1/4))                # Q1
print(chisquare([160, 40], f_exp=[150, 50])) # Q3
```

Chapter 3: Hardy–Weinberg Equilibrium

Answer to Q1. $N = 500$. $\hat{p}(B) = [2(350) + 130] / [2(500)] = 830/1000 = 0.83$. $\hat{q}(b) = 1 - 0.83 = 0.17$.

Answer to Q2. $E_{BB} = 500(0.83)^2 = 344.5$; $E_{Bb} = 2(500)(0.83)(0.17) = 141.1$; $E_{bb} = 500(0.17)^2 = 14.45$. $\chi^2 = (350 - 344.5)^2/344.5 + (130 - 141.1)^2/141.1 + (20 - 14.45)^2/14.45 = 0.088 + 0.874 + 2.132 = 3.09$. Critical value (1 df, $\alpha=0.05$) = 3.84. Since $3.09 < 3.84$, we do not reject H_0 — consistent with HWE (though the fit is closer to the boundary, illustrating that G6PD data can show mild departures, e.g. due to selection).

Answer to Q3. $P(O) = q^2$ (where $q = \text{freq}(i) = 0.36 \Rightarrow q = 0.6$. Then $\text{freq}(I^A) = p = 1 - 0.6 = 0.4$.

Answer to Q4. This is the Wahlund effect: pooling two subpopulations with different allele frequencies (say p_1, p_2) into one sample yields an average homozygote frequency of $E[p_i^2] = \bar{p}^2 + \text{Var}(p)$, which always exceeds $\bar{p}^2$ (the homozygote frequency expected if the pooled population were a single randomly mating unit) whenever $\text{Var}(p) > 0$. In other words, hidden population structure inflates apparent homozygosity even without inbreeding, selection or assortative mating at the individual level.

R and Python verification (Q1, Q2).

```
n_BB<-350; n_Bb<-130; n_bb<-20; N<-500
p_hat <- (2*n_BB + n_Bb) / (2*N); q_hat <- 1 - p_hat
E <- N * c(p_hat^2, 2*p_hat*q_hat, q_hat^2)
chisq_stat <- sum((c(n_BB, n_Bb, n_bb) - E)^2 / E)
c(p_hat = p_hat, q_hat = q_hat, chisq = chisq_stat, p_value
  =pchisq(chisq_stat, 1, lower.tail=FALSE))
```

```

import numpy as np
from scipy.stats import chi2
n_BB, n_Bb, n_bb = 350, 130, 20; N = 500
p_hat = (2*n_BB + n_Bb) / (2*N); q_hat = 1 - p_hat
E = N * np.array([p_hat**2, 2*p_hat*q_hat, q_hat**2])
chisq_stat = np.sum((np.array([n_BB, n_Bb, n_bb]) - E)**2 /
E)
print(p_hat, q_hat, chisq_stat, chi2.sf(chisq_stat, 1))

```

Answer to Q5. The expected count of the rare homozygote (TT) is only $500 \times 0.005^2 \approx 0.0125$, far below the conventional rule-of-thumb minimum of 5 expected counts per cell required for the chi-square approximation to be reliable; with so small an expected count, the chi-square test's p-value would not be trustworthy. An **exact test** (Section 3.9), which computes the exact probability of the observed genotype configuration under HWE given the fixed allele counts, without relying on any large-sample approximation, should be used instead.

Answer to Q6.

```

library(HardyWeinberg)
HWExact(c(AA =990, AB =10, BB =0), verbose
=TRUE)

```

This scenario has an even smaller minor allele frequency ($\hat{q} = 10/2000 = 0.005$, same as Worked Example 3.3a, but a larger sample), so the exact test remains the appropriate choice; the resulting p-value should be qualitatively similar in spirit to Worked Example 3.3a (no evidence against HWE, since a single heterozygote-only rare-allele pattern with zero rare homozygotes is entirely consistent with a rare allele under random mating), while a naively-applied chi-square test would again be unreliable here for exactly the same small-expected-count reason as in Q5.

Chapter 4: Estimation of Allele Frequency: Approach to Equilibrium for an X-Linked Gene

Answer to Q1. If $p_m(0) = p_f(0)$, then $d(0) = 0$, so $d(t) = 0$ for all t : the two sexes start at, and remain at, the same allele frequency in every subsequent generation — the population is already at its X-linked equilibrium.

Answer to Q2. $p_\infty = (1/3)(0.05) + (2/3)(0.25) = 0.01667 + 0.16667 = 0.1833$. $p_m(1) = p_f(0) = 0.25$. $p_f(1) = \frac{1}{2}(0.25) + \frac{1}{2}(0.05) = 0.15$. $p_m(2) = p_f(1) = 0.15$. $p_f(2) = \frac{1}{2}(0.15) + \frac{1}{2}(0.25) = 0.20$.

Answer to Q3. $d(t+1) = p_f(t+1) - p_m(t+1) = [\frac{1}{2}p_f(t) + \frac{1}{2}p_m(t)] - p_f(t) = -\frac{1}{2}p_f(t) + \frac{1}{2}p_m(t) = -\frac{1}{2}[p_f(t) - p_m(t)] = -\frac{1}{2}d(t)$. This is a first-order geometric recurrence with ratio $-\frac{1}{2}$, so by induction $d(t) = (-\frac{1}{2})^t d(0)$ for all $t \geq 0$.

Answer to Q4. We need $(1/2)^t < 0.01 \Rightarrow t > \log(0.01)/\log(0.5) = (-2)/(-0.301) \approx 6.64$, so $t = 7$ generations (the smallest integer satisfying the inequality).

R and Python verification (Q2, Q4).

```
iterate_xlinked <- function(q_f0, q_m0, generations = 8) {
  q_f <- q_f0; q_m <- q_m0
  for (t in 1:generations) { new_f <- 0.5*(q_f+q_m); q_m <- q_f;
    q_f <- new_f }
  c(q_f = q_f, q_m = q_m)
}
iterate_xlinked(0.05, 0.25, 1) # Q2: generation 1
iterate_xlinked(0.05, 0.25, 2) # Q2: generation 2
log(0.01)/log(0.5) # Q4: generations to |d(t)| < 0.01
import numpy as np
def iterate_xlinked(qf, qm, gens):
```

```

for _ in range(gens):
    qf, qm = 0.5*(qf+qm), qf
return qf, qm
print(iterate_xlinked(0.05, 0.25, 1))
print(iterate_xlinked(0.05, 0.25, 2))
print(np.log(0.01) / np.log(0.5))

```

Answer to Q5. The neutral recurrence of Section 4.3 only redistributes an existing allele frequency difference between the sexes toward a common weighted-average equilibrium — it cannot, on its own, *increase* the overall allele frequency in the population, since neither equation adds new copies of the allele; the long-run equilibrium $p_\infty = (1/3)p_m(0) + (2/3)p_f(0)$ is always a weighted average of (and therefore bounded between) the two starting frequencies. To explain the allele rising from a low starting frequency to 20%, an additional force — natural **selection** favouring carriers of the allele (here, most plausibly, a survival or fertility advantage for G6PD-deficient heterozygous females or hemizygous males in malaria-endemic environments, analogous to the sickle-cell heterozygote advantage of Chapter 3) — must be incorporated into the model, since only selection (or migration introducing additional copies, or mutation, both quantitatively minor over the relevant timescale) can change the population-wide allele frequency itself.

Answer to Q6.

```

iterate_xlinked(0.10, 0.60, 1); iterate_xlinked(0.10,
0.60, 5); iterate_xlinked(0.10, 0.60, 10)

for g in [1, 5, 10]:
    print(g, iterate_xlinked(0.10, 0.60, g))

```

Starting from a large sex difference ($d(0) = 0.10 - 0.60$)

= -0.50), the geometric decay $d(t) = (-1/2)^t d(0)$ means the difference roughly halves each generation; by around generation 8–10 the two curves should be visually indistinguishable on a typical plot (the residual difference, $(1/2)^{10} \times 0.50 \approx 0.0005$, being far smaller than the plot's visual resolution), consistent with the general rate result of Section 4.6.

Chapter 5: Non-Random Mating: Phenotypic Assortative Mating

Answer to Q1. Allele frequency is a property of the whole gene pool and is only changed by processes that add, remove, or differentially transmit alleles (mutation, migration, selection, drift) — assortative mating merely changes *who mates with whom*, not the total count of A and a alleles in the population, so p and q are unchanged. Heterozygosity falls because “like with like” mating preferentially creates AA×AA and aa×aa pairs (which can only produce homozygous offspring), reducing the proportion of Aa×Aa and other mixed pairings that generate heterozygotes, relative to random mating.

Answer to Q2. $V_A^*/V_A = 1/(1 - rh^2) = 1/(1 - (0.15)(0.40)) = 1/(1 - 0.06) = 1/0.94 = 1.064$, i.e., about a 6.4% inflation in additive genetic variance.

Answer to Q3. This pattern is more consistent with consanguinity (inbreeding). Because inbreeding reflects shared ancestry, it elevates homozygosity across the entire genome, not merely at loci tied to a specific assorted trait, and it specifically raises the incidence of rare autosomal recessive disorders (since related parents are more likely to both carry the same

rare recessive allele identical by descent) — exactly the pattern described.

Answer to Q4. With $r = 0$, $V_A^*/V_A = 1/(1-0) = 1$, i.e., no inflation: additive genetic variance remains at its random-mating value. This makes sense because assortative mating only distorts the transmission of genetic variance when mate choice is correlated with the trait's genetic value; with zero spousal correlation, mating is effectively random with respect to the trait, so there is no assortative-mating effect to inflate the variance.

R and Python verification (Q2).

```
r <-0.15; h2 <-0.40
1/ (1- r*h2)
r, h2 =0.15, 0.40
print(1/ (1- r*h2))
```

Answer to Q5. This pattern suggests **phenotypic assortative mating on pigmentation** rather than inbreeding: the elevated homozygosity is confined to the trait-associated locus (and, presumably, loci in LD with it), while background genome-wide homozygosity — which inbreeding would elevate uniformly across the whole genome — remains at its Hardy–Weinberg-expected level, exactly the diagnostic distinction drawn in Section 5.7.

Answer to Q6.

```
result_strong <-simulate_assortative(p0=0.5, alpha=0.8)
tail(result_strong, 1)
result_weak <-simulate_assortative(p0=0.5, alpha=0.3)
tail(result_weak, 1)
result_strong = simulate_assortative(0.5, 0.8)
print(result_strong.tail(1))
result_weak = simulate_assortative(0.5, 0.3)
```

```
print(result_weak.tail(1))
```

Increasing alpha from 0.3 to 0.8 should produce a markedly lower equilibrium heterozygote (Aa) proportion after 10 generations, since stronger positive assortment more aggressively sorts individuals into like-genotype pairs, accelerating the loss of heterozygosity relative to the weaker-assortment case — while the allele frequency p remains approximately 0.5 in both simulations, exactly as Q1's theoretical result predicts.

Chapter 6: Pedigree Analysis: Structure, Notation, and Relatedness

Answer to Q1. For ordinary first cousins, $r = 1/8$ (Section 6.3). Double first cousins share two independent lines of cousin relationship (one through each pair of grandparents), so the contributions add: $r = 1/8 + 1/8 = 1/4$.

```
r_single_cousin <- 1/8
r_double_cousin <- r_single_cousin + r_single_cousin
r_double_cousin
r_single_cousin = 1/8
r_double_cousin = r_single_cousin + r_single_cousin
print(r_double_cousin)
```

Answer to Q2. For an avuncular (uncle/aunt–niece/nephew) pair, $r = 1/4$ (Section 6.3), so $F(\text{child}) = r/2 = 1/8 = \mathbf{0.125}$. This is exactly double the risk from a first-cousin union ($F = 1/16$, Worked Example 6.4), reflecting the closer relationship between uncle/aunt and niece/nephew ($r = 1/4$) than between first cousins ($r = 1/8$).

```
F_uncle_niece <- (1/4) / 2
F_first_cousins <- (1/8) / 2
c(F_uncle_niece = F_uncle_niece, F_first_cousins =
F_first_cousins, ratio = F_uncle_niece / F_first_cousins)
F_uncle_niece = (1/4) / 2
F_first_cousins = (1/8) / 2
```

```
print(F_uncle_niece, F_first_cousins, F_uncle_niece /
      F_first_cousins)
```

Answer to Q3. Generation I: I-1 (square, filled — affected male, G6PD-deficient) mated to I-2 (circle, unfilled — unaffected, non-carrier or untested female). Generation II: two daughters, II-1 and II-2 (circles, unfilled but obligate carriers, since each inherits her father's only X chromosome), each married to an unrelated, unaffected male. Generation III: II-1's son, III-1, is drawn as a filled square (affected), having inherited the deficient X allele from his carrier mother; II-1's other children (if any) and II-2's children would be unaffected or carriers with probability $1/2$ each, depending on sex. Note the classic X-linked recessive signature: the trait passes from an affected grandfather, through obligate-carrier daughters (never affected themselves), to affected grandsons — with no father-to-son transmission anywhere in the pedigree.

Answer to Q4. Running the R or Python code of Section 6.7 for individuals 6 and 7 (both children of the same parents, individuals 3 and 5) gives $\phi(6,7) = 0.25$, exactly the full-sibling kinship coefficient, confirming $r = 2\phi = 0.5$ — matching the path-counting result of Section 6.3 (two paths, through the shared mother and shared father, each contributing $(1/2)^2 = 1/4$, summing to $1/2$).

Answer to Q5. The kinship coefficient between an individual and an unrelated founder is zero because, by definition, an unrelated founder has no ancestors in common with the individual anywhere in the pedigree, so there is no path along which an allele could be identical by descent between them. A non-inbred founder's kinship coefficient with themselves is $\phi = 1/2$ (not 1) because ϕ_{ii} is defined as the probability that

two alleles drawn *at random*, one from each of two independent copies of the individual's genotype (conceptually, from two of the individual's own gametes) are identical by descent; a non-inbred individual's own two alleles at a locus are, by definition, not IBD (they came from two different, unrelated ancestral lineages via the two parents), so this probability works out to $1/2$, and the corresponding coefficient of relationship with oneself is $r = 2\phi = 1$ as expected (a person shares 100% of their own genotype with themselves), while $F = 0$ confirms no inbreeding.

Answer to Q6. An offspring allele matching neither parental allele at a locus indicates a **Mendelian inheritance inconsistency** — biologically, this cannot arise from ordinary transmission (barring a new mutation at that exact site, which is very rare), and instead usually signals a genotyping error, a sample-labelling mix-up, or non-paternity/non-maternity. The modern quality-control procedure built on this logic is **Mendelian error checking**, a standard automated step in genotype quality control (Chapter 19) that scans every genotyped parent–offspring trio for such inconsistencies and flags or removes the implicated markers or individuals before any association or linkage analysis proceeds.

Answer to Q7. Half-first-cousins share exactly one grandparent (rather than two, as for ordinary first cousins), so there is only **one** path connecting them, of length 4 meioses (child → shared grandparent → other child → half-cousin, i.e., 4 links), contributing $(1/2)^4 = 1/16$. With only one such path (compared to the two paths for ordinary first cousins), $r = 1/16$ — half the

ordinary first-cousin value of $1/8$, reflecting the halved degree of shared ancestry.

Chapter 7: Segregation Analysis and Genetic Model Fitting

Answer to Q1. $\Sigma(a_i - 1) = (3 - 1) + (2 - 1) + (1 - 1) = 2 + 1 + 0 = 3$.
 $\Sigma(s_i - 1) = (6 - 1) + (4 - 1) + (3 - 1) = 5 + 3 + 2 = 10$. $\hat{p} = 3/10 = 0.30$.

```
fam <- data.frame(s = c(6,4,3), a = c(3,2,1))
p_hat <- sum(fam$a - 1) / sum(fam$s - 1)
p_hat

s = [6, 4, 3]; a = [3, 2, 1]
p_hat = sum(ai - 1 for ai in a) / sum(si - 1 for si in s)
print(p_hat)
```

Answer to Q2. Every ascertained family is, by construction, guaranteed to contain at least one affected member (the proband); families with zero affected offspring are never sampled at all, no matter how common they truly are. Including every proband, without removing at least one affected individual per family, therefore systematically inflates the apparent proportion of affected offspring relative to the population's true underlying segregation ratio, since the "luckiest" families (zero affected) are entirely excluded from the naive calculation while every included family's count is biased upward by at least its proband.

Answer to Q3. Using the R/Python code of Section 7.6 with
`fam <- data.frame(s = c(6,4,3), a = c(3,2,1)):`

```
negloglik <- function(p, fam) {
  -sum(lchoose(fam$s - 1, fam$a - 1) + (fam$a - 1)
    * log(p) + (fam$s - fam$a) * log(1 - p))
}
```

```

fit <-optim(0.3, negloglik, fam = fam, method
="Brent", lower =1e-4, upper =1-1e-4)
ll_alt <--negloglik(fit$par, fam); ll_null <--
negloglik(0.25, fam)
LRT <-2* (ll_alt - ll_null)
c(p_mle = fit$par, LRT = LRT, p_value =pchisq(LRT,
df =1, lower.tail =FALSE))

```

```

import numpy as np
from scipy.optimize import minimize_scalar
from scipy.stats import chi2
from math import lgamma, log
s = [6, 4, 3]; a = [3, 2, 1]
def lchoose(n, k): return lgamma(n+1) - lgamma(k+1)
- lgamma(n-k+1)
def negloglik(p):
return-sum(lchoose(si-1, ai-1) + (ai-1)*log(p) + (si-
ai)*log(1-p) for si, ai inzip(s, a))
res = minimize_scalar(negloglik, bounds=(1e-4, 1-1e-
4), method="bounded")
ll_alt =-negloglik(res.x); ll_null =-negloglik(0.25)
LRT =2*(ll_alt - ll_null)
print(res.x, LRT, chi2.sf(LRT, 1))

```

With only 3 small families, the LRT statistic will typically be small and the p-value well above 0.05, so the data — while giving a point estimate (0.30) somewhat above the Mendelian 0.25 — do not provide statistically significant evidence against the recessive model at this sample size.

Answer to Q4. Under complete (truncate) ascertainment, essentially every affected individual in the population is certain to be identified and to bring their family into the sample, so it is no longer true that families are ascertained through a single, low-probability proband

event; instead, a family with *any* affected children is certain to be ascertained, and a family could in principle be “found” via more than one of its affected members simultaneously. Simply removing one proband per family (the single-ascertainment correction) under-corrects in this setting, because more than one affected child effectively acted as an independent “entry point” into the sample; the correct (truncate) likelihood must instead condition on the fact that the family was ascertained *at all* — formally, conditioning the family’s contribution to the likelihood on the event “at least one affected offspring,” rather than on “exactly one designated proband.”

Answer to Q5. $LRT = 2 \times (-16.9 - (-18.4)) = 2 \times 1.5 = 3.0$.

Comparing to χ^2_1 , the critical value at $\alpha = 0.05$ is 3.84; since $3.0 < 3.84$, the result is not statistically significant at the conventional 5% level ($p \approx 0.083$), so we do not reject the simple autosomal recessive model ($p = 0.25$) in favour of the unrestricted model, despite the unrestricted model achieving a somewhat higher likelihood.

```
LRT <- 2 * (-16.9 - (-18.4)); pchisq(LRT, df = 1,
lower.tail = FALSE)
```

```
from scipy.stats import chi2
```

```
LRT = 2 * (-16.9 - (-18.4)); print(LRT, chi2.sf(LRT, 1))
```

Answer to Q6. Each individual small family contributes only a modest amount of statistical information (few informative meioses), so its likelihood curve as a function of the segregation ratio p tends to be flat/wide, giving weak, imprecise evidence about the true value of p and a wide confidence interval. Because families are ascertained independently, their

individual log-likelihoods can be **summed** to form a combined log-likelihood (exactly as independent likelihood contributions combine in Section 7.4), which sharpens (narrows) the combined likelihood curve around the true parameter value — the same statistical principle of information accumulation that underlies combining LOD scores across families in linkage analysis (Chapter 17) and combining studies in GWAS meta-analysis (Chapter 18).

Chapter 8: Estimation of the Recombination Fraction

Answer to Q1. $\theta = 8/50 = 0.16$. $SE(\theta) = \sqrt{[0.16(0.84)/50]} = \sqrt{[0.1344/50]} = \sqrt{0.002688} = 0.0518$.

Answer to Q2. $Z(\theta) = \log_{10}[\theta^8 (1-\theta)^{42} / 0.5^{50}]$ at $\theta=0.16$. $(0.16)^8$: $\ln(0.16) = -1.8326$, $\times 8 = -14.661$, $e^{(-14.661)} \approx 4.29 \times 10^{-7}$. $(0.84)^{42}$: $\ln(0.84) = -0.1744$, $\times 42 = -7.324$, $e^{(-7.324)} \approx 6.59 \times 10^{-4}$. Numerator $\approx 4.29 \times 10^{-7} \times 6.59 \times 10^{-4} \approx 2.83 \times 10^{-10}$. $(0.5)^{50} \approx 8.88 \times 10^{-16}$. Ratio $\approx 2.83 \times 10^{-10} / 8.88 \times 10^{-16} \approx 3.19 \times 10^5$. $Z = \log_{10}(3.19 \times 10^5) \approx 5.50$. Since $Z = 5.50 \geq 3$, this is conventionally significant evidence for linkage.

Answer to Q3. $\theta = 0.5$ means recombinant and non-recombinant gametes are equally likely — exactly what independent assortment (Mendel's Second Law) predicts for loci on different chromosomes (or very far apart), i.e. knowing which allele was transmitted at one locus gives no information about the other. $\theta = 0$ means every gamete is non-recombinant, i.e., the two loci are always transmitted together — consistent with the loci being so close together (or identical) that crossing-over between them essentially never occurs.

Answer to Q4. A LOD score of -3 at $\theta = 0.5$ is not the usual way LOD is reported ($Z(0.5)$ is by definition 0, since $L(0.5)/L(0.5) = 1$); what is meant here is a LOD score of -3 at some trial θ substantially less than 0.5, or more precisely a summary statistic strongly favouring $\theta = 0.5$. A LOD of -3 (odds of 1000:1 against linkage at the tested θ) provides significant evidence *against* linkage at that trial distance — i.e., strong evidence that the two loci assort independently (are unlinked, or are very far apart).

R and Python verification (Q1, Q2).

```
r_recomb <- 8; n <- 50
theta_hat <- r_recomb / n
se_theta <- sqrt(theta_hat*(1-theta_hat)/n)
lod <- function(theta, r, n) (r*log10(theta) + (n-r)*log10(1-
theta)) - (n*log10(0.5))
c(theta_hat = theta_hat, se = se_theta, LOD = lod(theta_hat,
r_recomb, n))

import numpy as np
r_recomb, n = 8, 50
theta_hat = r_recomb / n
se_theta = np.sqrt(theta_hat*(1-theta_hat)/n)
lod = (r_recomb*np.log10(theta_hat) + (n-
r_recomb)*np.log10(1-theta_hat)) - n*np.log10(0.5)
print(theta_hat, se_theta, lod)
```

Answer to Q5. Because LOD scores are additive across independent families at a fixed θ , the combined LOD score at $\theta = 0.15$ is simply $1.6 + 1.6 = 3.2$. This crosses the conventional significance threshold of $Z \geq 3$, so while neither family alone provided significant evidence, their combined analysis does provide significant evidence of linkage at $\theta = 0.15$.

Answer to Q6.

```

lod <-function(theta, r, n) (r*log10(theta) + (n-
r)*log10(1-theta)) - (n*log10(0.5))
# Family 1: 8 recombinants out of 50; Family 2: 5
recombinants out of 30
lod_fam1 <-lod(0.2, 8, 50); lod_fam2 <-lod(0.2, 5, 30)
lod_pooled <-lod(0.2, 8+5, 50+30)
c(sum_individual = lod_fam1 + lod_fam2, pooled =
lod_pooled) # should match

deflod(theta, r, n):
return (r*np.log10(theta) + (n-r)*np.log10(1-theta)) -
n*np.log10(0.5)
lod_fam1 = lod(0.2, 8, 50); lod_fam2 = lod(0.2, 5, 30)
lod_pooled = lod(0.2, 13, 80)
print(lod_fam1 + lod_fam2, lod_pooled)

```

The two quantities match exactly, confirming the additivity property algebraically as well as numerically: summing log-likelihood-ratio contributions family-by-family is mathematically identical to computing the LOD score directly from the pooled recombinant/non-recombinant counts.

Chapter 9: Inheritance of Quantitative Traits

Answer to Q1. If $t = 0.99$ (the regression slope directly estimates h^2 under the mid-parent regression model). This is plausible because birth weight, unlike adult height, is strongly influenced by maternal environmental factors during pregnancy (maternal nutrition, placental function, gestational age, maternal illness such as malaria or hypertension), which inflate V_E relative to V_A and hence lower $h^2 = V_A/(V_A+V_E+\dots)$.

Answer to Q2. $t = \sigma^2_B/(\sigma^2_B + \sigma^2_W) = 45/(45+15) = 45/60 = 0.75$. $\hat{h}^2 = 2t = 1.50$. (Note: an estimate above 1.0 is

not biologically meaningful as a heritability — it signals that the simplifying assumptions behind $t \approx \frac{1}{2}h^2$ [no dominance variance, no shared sibling environment, no assortative mating] are likely violated in this sample, most plausibly because siblings also share a common household/environment that inflates the between-family variance component beyond what shared genes alone would produce; in practice such estimates would be treated as evidence of shared-environment effects rather than literal heritability, and the model would need extension, e.g., using twin or adoption designs.)

Answer to Q3. Parent–offspring regression works because a parent transmits, on average, exactly half of their *additive* genetic value to each offspring in an unbiased way — the additive effects of individual alleles pass through segregation predictably. Dominance effects (interaction between the two specific alleles at a locus) and epistatic effects (interaction between alleles at different loci) depend on particular allele *combinations* within an individual; these combinations are broken apart and reassembled anew at each generation by segregation and independent assortment/recombination (Chapters 2 and 7), so a parent’s dominance/epistatic deviation is not reliably passed on — hence these components inflate H^2 (which counts all genetic variance) but do not inflate the parent–offspring regression slope, which isolates $h^2 = V_A/V_P$.

Answer to Q4. Under the threshold model, liability to the defect is treated as a continuous, roughly normally distributed trait influenced by many genes and environmental factors, with the defect appearing only above a high threshold. Full siblings of an affected

child share, on average, half of the many genes contributing to liability, so their own liability distribution is shifted somewhat closer to the threshold than the general population's — increasing the proportion of siblings who exceed the threshold (3%) relative to the general population (0.5%), without requiring that any single gene, on its own, determines the defect in a simple Mendelian fashion.

R and Python verification (Q1, Q2).

```
sigma2_B <-45; sigma2_W <-15
t <- sigma2_B / (sigma2_B + sigma2_W); h2_hat <-2*t
c(t = t, h2_hat = h2_hat)
sigma2_B, sigma2_W =45, 15
t = sigma2_B / (sigma2_B + sigma2_W)
print(t, 2*t)
```

Answer to Q5. Each individual gene of small effect contributes a small, discrete increment (or decrement) to an individual's genotypic value, and the Central Limit Theorem states that the *sum* of a large number of independent (or weakly correlated) random contributions tends toward a normal distribution as the number of contributions grows, regardless of the distribution of each individual contribution — so even though a single gene's effect might follow an arbitrary (even a simple two- or three-valued) distribution, the sum across hundreds of such genes, plus environmental variation, converges toward the familiar smooth, symmetric, bell-shaped normal curve.

Answer to Q6. $H^2 = V_G/V_P = (0.35+0.05+0.02)/1.00 = 0.42/1.00 = \mathbf{0.42}$. $h^2 = V_A/V_P = 0.35/1.00 = \mathbf{0.35}$. Since V_A is only one (non-negative) component of the total genetic variance $V_G = V_A + V_D + V_I$, and $V_D, V_I \geq 0$, it always holds that $V_A \leq V_G$,

and hence $h^2 \leq H^2$ — narrow-sense heritability can never exceed broad-sense heritability, with equality only when dominance and epistatic variance are both exactly zero.

```
V_A<-0.35; V_D<-0.05; V_I<-0.02; V_E<-0.58; V_P
<- V_A+V_D+V_I+V_E
c(H2=(V_A+V_D+V_I)/V_P, h2=V_A/V_P)
```

Chapter 10: Likelihood Methods in Statistical Genetics

Answer to Q1. $l(\theta) = n_R \log \theta + n_{NR} \log(1-\theta)$. $U(\theta) = n_R/\theta - n_{NR}/(1-\theta)$. Setting $U(\theta)=0$: $n_R(1-\theta) = n_{NR} \theta \Rightarrow n_R - n_R\theta = n_{NR}\theta \Rightarrow n_R = \theta(n_R+n_{NR}) \Rightarrow \theta = n_R/(n_R+n_{NR})$, matching the estimator quoted in Chapter 8.

Answer to Q2. $\hat{p} = (2 \times 100 + 200)/(2 \times 400) = 400/800 = 0.5$. At $\hat{p}=0.5$, HWE-expected proportions are $p^2=0.25$, $2pq=0.50$, $q^2=0.25$, i.e., expected counts 100, 200, 100 — exactly equal to the observed counts. So the LRT statistic Λ would be (very close to) 0, and we would strongly fail to reject H_0 : this sample fits HWE essentially perfectly.

Answer to Q3. Consistency: $\hat{\psi} \rightarrow \text{true } \psi$ as sample size grows. Asymptotic normality : $\hat{\psi}$ is approximately normally distributed for large n , permitting standard confidence intervals and tests. Asymptotic efficiency: the MLE has the smallest possible variance among consistent, asymptotically normal estimators. Efficiency matters most precisely when samples are *limited* (e.g., a rare Nigerian pedigree disorder with only a handful of informative families), because an inefficient estimator wastes the little information available, whereas the MLE extracts the maximum possible precision from a small dataset.

Answer to Q4. Because $\theta = 0.5$ is the *boundary* of the biologically meaningful parameter space for θ (which by convention ranges over $[0, 0.5]$) when testing for linkage ($\theta < 0.5$) rather than an interior null value; the standard chi-square asymptotic derivation for the LRT assumes the true parameter lies in the interior of the parameter space under H_0 . When the null value is on the boundary, the correct asymptotic null distribution of the (one-sided) LRT statistic is instead a 50:50 mixture of a point mass at zero and a χ^2_1 distribution.

R and Python verification (Q2).

```
n_AA<-100; n_Aa<-200; n_aa<-100; N<-400
p_hat<- (2*n_AA+n_Aa)/(2*N)
E<- N*c(p_hat^2, 2*p_hat*(1-p_hat), (1-p_hat)^2)
sum((c(n_AA,n_Aa,n_aa)-E)^2/E) # LRT/chi-square ~ 0:
perfect HWE fit
```

import numpy as np

```
n_AA, n_Aa, n_aa=100, 200, 100; N=400
p_hat=(2*n_AA + n_Aa)/(2*N)
E=N*np.array([p_hat**2, 2*p_hat*(1-p_hat), (1-p_hat)**2])
print(np.sum((np.array([n_AA,n_Aa,n_aa]) - E)**2/E))
```

Answer to Q5. A flat likelihood surface near its peak means that a wide range of parameter values are all nearly as consistent with the observed data as the MLE itself — the data do not sharply discriminate between them — corresponding to a small curvature (second derivative) at the peak and hence, via the observed/expected information matrix (Section 10.4), a *large* standard error for the estimate. A sharply peaked surface, by contrast, means only a narrow range of parameter values is compatible with the data, corresponding to large curvature and a *small* standard error — the data pin down the parameter precisely.

Answer to Q6.

```
newton_raphson(0.99)
newton_raphson(0.5)
print(newton_raphson(0.99))
print(newton_raphson(0.5))
```

Because the score and information functions here are well-behaved (smooth and monotonic) over the interior of (0,1), Newton–Raphson should still converge to the same correct MLE even from the poor starting value of 0.99, though it will typically require a few more iterations to travel the greater distance back toward the true optimum than when starting from a central value like 0.5 — an illustration of the general lesson that Newton–Raphson’s fast (quadratic) local convergence is not guaranteed to be fast, or even guaranteed to converge at all, far from the optimum or near a function’s boundary.

Chapter 11: Population-Based Association Tests

Answer to Q1. Recessive-model table: TT — cases 45, controls 20; TC+CC — cases 255, controls 280. Row/column totals: TT total=65, TC+CC total=535, cases=300, controls=300, grand total=600. $E(\text{TT}, \text{case}) = (65 \times 300) / 600 = 32.5$; $E(\text{TC}+\text{CC}, \text{case}) = 267.5$; by symmetry

$E(\text{TT}, \text{control}) = 32.5$ $E(\text{TC}+\text{CC}, \text{control}) = 267.5$. $\chi^2 =$
 $(45-32.5)^2/32.5 + (255-267.5)^2/267.5 +$
 $(20-32.5)^2/32.5 + (280-267.5)^2/267.5 =$
 $4.808+0.584+4.808+0.584 = 10.78$ (1 df) —
 significant (critical value 3.84). OR =
 $(45 \times 280) / (20 \times 255) = 12600 / 5100 = 2.47$.

Answer to Q2. The allele-based test treats each individual’s two alleles as two independent observations in the 2×2 table; this independence between the two alleles within a

genotype is precisely the assumption of Hardy–Weinberg equilibrium (random union of gametes), so if HWE does not hold (e.g., due to population stratification), the allele-based test’s nominal type-I error rate is no longer valid, whereas the genotype-based test treats each individual (not each allele) as the independent unit and makes no such assumption.

Answer to Q3. Nigeria’s population is genetically diverse (multiple ethnic groups with somewhat different allele frequencies at many loci, an instance of the Wahlund-type structure of Chapter 3); if cases and controls are not perfectly matched for ancestry, a marker that simply differs in frequency between ethnic groups can appear spuriously associated with disease if disease prevalence also differs by group, purely due to confounding by ancestry rather than any true genetic effect on the disease — a phenomenon called population stratification. Including ancestry principal components as covariates statistically adjusts for this confounding, protecting against false-positive associations.

Answer to Q4. An OR of 1.0 with a non-significant p-value in a well-powered study provides reasonably strong evidence that this marker has **little or no** effect on disease risk in this population — i.e., the data are consistent with the null hypothesis of no association, and (because the study is well-powered) a true effect of at least moderate size can likely be ruled out, though very small effects can never be entirely excluded by any finite study.

R and Python verification (Q1).

```
tab <-matrix(c(45,255,20,280), nrow=2, byrow=TRUE,
dimnames=list(c("case","control"),c("TT","TC_CC")))
chisq.test(tab, correct=FALSE)
(45*280)/(20*255)    # odds ratio
```

```

import numpy as np
from scipy.stats import chi2_contingency
tab = np.array([[45, 255], [20, 280]])
chi2_stat, p, dof, exp = chi2_contingency(tab,
correction=False)
print(chi2_stat, p)
print((45*280)/(20*255))

```

Answer to Q5. Testing multiple genetic models and reporting only the smallest resulting p-value is equivalent, in its statistical effect, to performing several correlated hypothesis tests and selectively reporting the most extreme outcome — exactly the multiple-testing problem addressed formally in Chapters 12 and 18, even though the analyst presents only a single “final” number. The true probability of obtaining at least one nominally significant result by chance, when several models are tried, is higher than the nominal per-test α level (e.g., trying 3 roughly independent models at $\alpha = 0.05$ gives a combined false-positive rate closer to $1 - 0.95^3 \approx 14\%$, not 5%), so the reported p-value understates the true false-positive risk unless the model choice is pre-specified (before seeing the data) or a multiple-testing correction is applied to account for the number of models compared.

Answer to Q6. The **dominant model test** would be expected to have the least power. Under a true recessive effect, heterozygotes have the same risk as homozygous non-carriers; the dominant-model test incorrectly groups these unaffected-risk heterozygotes together with the truly high-risk homozygotes into a single “carrier” category, diluting the genuine signal (since the “carrier” group is now a mix of truly elevated-risk and normal-risk individuals) and thereby reducing the

test's ability to detect the real association, compared with the recessive-model test (which correctly isolates the homozygous risk group) or the 2-df genotype-based test (which does not force any particular grouping).

Chapter 12: Candidate Gene Association Studies

Answer to Q1. Bonferroni threshold = $\alpha/m = 0.05/8 = \mathbf{0.00625}$.

Only SNPs with a per-test p-value below 0.00625 would be declared significant.

0.05/8

print(0.05/8)

Answer to Q2. When tested SNPs are in strong LD, they are not independent tests, so the simple Bonferroni correction (which assumes m independent tests) is overly conservative — it treats correlated tests as if they each carried a full, independent unit of multiple-testing “penalty,” inflating the effective threshold far more than necessary and losing statistical power. A permutation test instead empirically derives the null distribution of the minimum p-value across all tested SNPs *while preserving their actual correlation structure* (since the same LD pattern is present in every permuted dataset), giving an accurate, appropriately less conservative family-wise significance threshold.

Answer to Q3. Two distinct (non-multiple-testing) reasons for the historically poor replication of candidate-gene findings: (i) small sample sizes typical of early candidate-gene studies gave low statistical power and highly variable effect-size estimates, so many “significant” findings were simply the winning tail of sampling variability (the “winner’s curse”); (ii) inadequate correction for population stratification (Chapter 14) allowed confounding by ancestry to produce spurious associations that appeared

statistically robust within a single study but did not reflect a true, generalisable biological effect and therefore failed to replicate elsewhere.

Answer to Q4. Because allele frequencies at a genetic locus can differ substantially between populations with different evolutionary and demographic histories, a variant’s relevance as a disease “risk allele” is population-specific: a variant that is common and clinically important in one population (here, *APOL1* risk alleles, common in West Africa due to historical selection for trypanosomiasis resistance) may be rare or absent, and therefore clinically irrelevant, in another population (here, European ancestry) — so risk-allele findings, and the clinical or research conclusions built on them, cannot be assumed to transfer across populations without independent validation in the population of intended application.

Answer to Q5. Setting `beta_true <- c(0, 0.30, 0, 0.30, 0)` (SNP2 and SNP4 non-zero) and re-running the simulation:

```
set.seed(55)
n <- 400
snp <- matrix(rbinom(n * 5, 2, prob = 0.3), nrow = n, ncol = 5)
beta_true <- c(0, 0.30, 0, 0.30, 0)
trait <- 120 + snp %*% beta_true + rnorm(n, 0, 5)
pvals <- sapply(1:5, function(j) summary(lm(trait ~ snp[,
j]))$coefficients[2, 4])
print(pvals)
print(pvals < 0.05/5)

import numpy as np, 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, 0.30, 0, 0.30, 0])
```

```

trait = 120 + snp @ beta_true + rng.normal(0, 5, n)
pvals = np.array([sm.OLS(trait, sm.add_constant(snp[:,
j])).fit().pvalues[1] for j in range(5)])
print(pvals, pvals < 0.05/5)

```

Both SNP2 and SNP4 should generally be correctly flagged by the Bonferroni threshold, though with only $n = 400$ and a modest effect size the exact outcome depends on the random seed and effect size chosen — reflecting the broader lesson that statistical power, not just correct methodology, determines whether a true signal is actually detected in any single dataset.

Answer to Q6. Because each tag SNP was chosen specifically so that every other common SNP in the region is correlated with at least one tag at $r^2 > 0.9$ (Section 12.3), a common causal variant — even if not itself directly genotyped — will almost always be very strongly correlated with (and therefore produce a strong, detectable association signal through) its nearest tag SNP; the tag SNP effectively “captures” most of the causal variant’s statistical signal by proxy. Testing only 6 well-chosen tags therefore retains nearly all the power to detect an association *somewhere* in the gene that testing all 40 SNPs directly would have provided, while requiring far less genotyping — the central economic rationale of LD-based tagging (Chapter 15).

Chapter 13: Family-Based Association Tests: The TDT and FBAT

Answer to Q1. $TDT = (118 - 82)^2 / (118 + 82) = 36^2 / 200 = 1296 / 200 = 6.48$. Comparing to the χ^2_1 critical value at $\alpha = 0.01$ (6.63), $6.48 < 6.63$, so this result is **not quite significant** at the 1% level (though it is significant at the conventional 5% level, critical value 3.84).

```

b <- 118; c <- 82
tdt <- (b - c)^2 / (b + c)

```

```
c(tdt = tdt, p_value =pchisq(tdt, df=1, lower.tail
=FALSE))
```

```
fromscipy.statsimport chi2
b, c =118, 82
tdt = (b - c)**2/ (b + c)
print(tdt, chi2.sf(tdt, 1))
```

Answer to Q2. The TDT statistic is built entirely from *within-family* transmissions — comparing, for each heterozygous parent, the allele actually transmitted to the affected child against the allele that could equally well have been transmitted (probability 1/2 each under Mendel’s first law, Chapter 2) — and never uses the population allele frequency, nor compares to any externally sampled control group. Because Mendelian transmission from a heterozygous parent is 1/2:1/2 regardless of which ethnic subgroup that parent belongs to, mixing trios from two ethnic groups with very different allele frequencies does not bias the null expectation of the TDT in any way; the test remains valid exactly because it never relies on the population’s overall allele frequencies matching between any two groups.

Answer to Q3. $TDT = (10-5)^2/(10+5) = 25/15 = 1.67$, well below the χ^2_1 critical value of 3.84 (not significant). With only 15 informative trios, the chi-square approximation to the true binomial-based null distribution is questionable (a common rule of thumb wants $b+c$ to be reasonably large, often $\geq 20-30$, for the chi-square approximation to be reliable); an **exact test** — a two-sided exact binomial test of $b = 10$ successes out of $b+c = 15$ trials against a null probability of 0.5 — would be preferable here and

would not rely on the large-sample chi-square approximation.

```
binom.test(10, 15, p=0.5)
```

```
from scipy.stats import binomtest
print(binomtest(10, 15, p=0.5))
```

Answer to Q4. For a single affected child at a bi-allelic marker, T_i (the phenotype-derived weight in the FBAT statistic) equals 1 for the (necessarily affected) offspring in each informative trio, and X_i is the offspring's observed count of the tested allele (0, 1, or 2 copies); conditioning on the sufficient statistic H (the parental genotypes) then reduces the general FBAT statistic exactly to a comparison of transmitted vs. non-transmitted alleles from heterozygous parents — precisely McNemar's test, i.e. the classical TDT statistic of Section 13.3.

Answer to Q5.

```
set.seed(1)
n_reps <- 1000
tdt_stats <- replicate(n_reps, {
  transmissions <- rbinom(90, 1, 0.5)
  b <- sum(transmissions == 1); c <- sum(transmissions == 0)
  (b - c)^2 / (b + c)
})
mean(tdt_stats > 3.84)
import numpy as np
rng = np.random.default_rng(1)
n_reps = 1000
tdt_stats = []
for _ in range(n_reps):
  transmissions = rng.binomial(1, 0.5, 90)
  b = np.sum(transmissions == 1); c = np.sum(transmissions
    == 0)
```

```
tdt_stats.append((b - c)**2 / (b + c))
print(np.mean(np.array(tdt_stats) > 3.84))
```

The simulated proportion of replications exceeding 3.84 should lie close to the nominal 0.05 type-I error rate, confirming that the TDT's asymptotic chi-square(1) null distribution is well calibrated for this sample size (90 heterozygous parents) — a direct simulation-based validation of the theory of Section 13.3, in the spirit of Chapter 20.

Answer to Q6. The classical TDT strictly requires both parents' genotypes for every trio, since its statistic is built entirely from the transmitted/non-transmitted alleles of heterozygous parents — with no parental genotype at all, there is nothing to compute. The general FBAT framework (Section 13.4), however, is built around a sufficient statistic that can, in some cases, be reconstructed from the genotypes of *other* family members (e.g., unaffected siblings) even when a parent's genotype is missing, using Mendelian transmission logic to partially infer the missing parental information; families with missing parents but informative siblings can therefore still contribute (partial) information to the FBAT statistic, whereas they would have to be discarded entirely under the simple trio-only TDT.

Chapter 14: Population Substructure, Stratification, and Admixture

Answer to Q1. $\bar{p} = 0.6(0.70) + 0.4(0.40) = 0.42 + 0.16 = 0.58$. Naive-HWE expected heterozygosity $= 2(0.58)(0.42) = 0.4872$. Wahlund deficit term $= 2(0.6)(0.4)(0.70 - 0.40)^2 = 0.48 \times 0.09 = 0.0432$. Actual pooled heterozygosity $= 0.4872 - 0.0432 = \mathbf{0.4440}$.

```
w1<-0.6; w2<-0.4; p1<-0.70; p2<-0.40
p_bar <- w1*p1 + w2*p2
naive_het <- 2*p_bar*(1-p_bar)
```

```

deficit <- 2*w1*w2*(p1-p2)^2
actual_het <- naive_het - deficit
c(p_bar=p_bar, naive_het=naive_het, deficit=deficit,
  actual_het=actual_het)

w1, w2, p1, p2 = 0.6, 0.4, 0.70, 0.40
p_bar = w1*p1 + w2*p2
naive_het = 2*p_bar*(1-p_bar)
deficit = 2*w1*w2*(p1-p2)**2
actual_het = naive_het - deficit
print(p_bar, naive_het, deficit, actual_het)

```

Answer to Q2. Population stratification refers to a sample being drawn from multiple genetically distinct subgroups that remain distinguishable at the group level — for example, a Nigerian national genetic study that separately recruits Yoruba participants in Ibadan and Hausa/Fulani participants in Kano, each group internally fairly homogeneous but differing from each other. Cryptic relatedness instead refers to some sampled individuals being more closely related to one another than expected under random sampling, without this being known — for example, a “population-based” case sample recruited from a single rural community in which, unknown to the investigators, several participants are in fact second cousins of one another through unrecorded family ties.

Answer to Q3. $\lambda_{GC} = 1.35$ suggests systematic inflation of the genome-wide test statistics beyond what would be expected under the null, consistent with — though not conclusive proof of — inadequately corrected population stratification or cryptic relatedness. Before concluding this is purely stratification, a careful analyst should examine a **principal component analysis** of the genome-wide genotype data (Section

14.5): if the leading PCs show clear sub-population clustering that aligns with the observed inflation (and including those PCs as covariates substantially reduces λ_{GC}), stratification is the likely explanation; if instead PCs show no clear structure and the inflation persists after PC adjustment, an alternative explanation — genuine, highly **polygenic** architecture (many true small-effect associations spread across the genome, which also mildly inflates λ_{GC}) — should be considered instead.

Answer to Q4. The leading principal components of genome-wide genotype data capture the major axes of ancestry variation in the sample (Section 14.5); because an individual’s estimated position along these axes is a good proxy for their ancestral background, including the PCs as covariates in the regression model statistically “holds ancestry constant” when estimating each marker’s effect on the trait, removing the part of the marker–trait association that is attributable to ancestry differences between cases and controls (Section 14.4) rather than to any true effect of the marker itself.

Answer to Q5. Increasing the allele-frequency difference (e.g. $p_1 = 0.9$, $p_2 = 0.3$, so $|p_1 - p_2| = 0.6$ instead of 0.3) increases the Wahlund deficit term $2w_1w_2(p_1 - p_2)^2$, since this term grows with the *square* of the frequency difference between the subpopulations; quadrupling the frequency difference ($0.3 \rightarrow 0.6$, a doubling) quadruples the deficit term. This confirms that the more differentiated the mixed subpopulations are at a locus, the larger the mechanical heterozygote deficit (and hence the larger the apparent, spurious departure from HWE) that results purely from pooling them, with no true inbreeding, selection, or non-random mating required.

```
w1<-0.5; w2<-0.5; p1<-0.9; p2<-0.3
deficit <-2*w1*w2*(p1-p2)^2
deficit
```

```
w1, w2, p1, p2 =0.5, 0.5, 0.9, 0.3
print(2*w1*w2*(p1-p2)**2)
```

Answer to Q6. An individual whose position falls between two population clusters most plausibly represents an **admixed individual** — someone whose recent ancestry draws from both underlying populations (for example, a person with one parent from each source population, or with more distant mixed ancestry), so that their genome-wide genotype pattern is, in effect, a weighted blend of both clusters' typical genotype patterns, placing them at an intermediate position along the principal component axis that separates the two groups.

Chapter 15: Linkage Disequilibrium and Haplotype Inference

Answer to Q1. $p_A = 0.45 + 0.05 = 0.50$; $p_B = 0.45 + 0.05 = 0.50$. $D = 0.45 - (0.50)(0.50) = 0.45 - 0.25 = 0.20$. $D_{\max} (D > 0) = \min(p_A p_b, p_a p_B) = \min[(0.50)(0.50), (0.50)(0.50)] = 0.25$. $D' = 0.20 / 0.25 = 0.80$. $r^2 = (0.20)^2 / [(0.5)(0.5)(0.5)(0.5)] = 0.04 / 0.0625 = 0.64$. Both D' (0.80) and r^2 (0.64) indicate strong linkage disequilibrium between these two loci.

Answer to Q2. Nigerian (African) populations generally have larger and more ancient effective population sizes than populations founded more recently from smaller groups (e.g., European populations, which experienced population bottlenecks during out-of-Africa migration). Larger, older populations have

accumulated more generations of recombination events, breaking down ancestral haplotype associations more thoroughly, so LD extends over shorter physical distances (decays faster with distance) in Nigerian/African populations than in populations with more recent, smaller founding effective population sizes.

Answer to Q3. Required sample size scales approximately as $1/r^2$ relative to direct typing: $N_{\text{marker}} \approx N_{\text{direct}} / r^2 = 1000/0.5 = 2000$ samples.

Answer to Q4. Clark's algorithm is a deterministic, parsimony-based approach: it starts from individuals whose haplotypes can be read off unambiguously, then resolves remaining ambiguous individuals one at a time by preferentially assigning them haplotypes already seen in the sample, aiming to minimise the total number of distinct haplotypes — it does not use a formal probability model. The EM algorithm is a likelihood-based, probabilistic approach: it treats the unresolved phase for each individual as missing data, and iteratively computes the expected probability of each possible haplotype-pair resolution given current estimated population haplotype frequencies (E-step), then updates those frequency estimates by maximum likelihood (M-step), converging to formal maximum likelihood estimates of haplotype frequencies across the whole sample simultaneously.

R and Python verification (Q1, Q3).

```
pA<-0.50; pB<-0.50; pAB<-0.45
D<- pAB - pA*pB
Dmax <-min(pA*(1-pB), (1-pA)*pB)
Dprime <- D/Dmax
```

```

r2 <- D^2/(pA*(1-pA)*pB*(1-pB))
c(D=D, Dprime=Dprime, r2=r2)
1000/0.5# Q3: sample size scaling by 1/r^2

pA, pB, pAB = 0.50, 0.50, 0.45
D = pAB - pA*pB
Dmax = min(pA*(1-pB), (1-pA)*pB)
Dprime = D/Dmax
r2 = D**2/(pA*(1-pA)*pB*(1-pB))
print(D, Dprime, r2)
print(1000/0.5)

```

Answer to Q5. Association power depends on r^2 (not merely physical proximity or D') because r^2 directly determines the effective sample size penalty for testing the marker instead of the true causal variant (Section 15.4's sample-size-inflation relationship, $n_{\text{marker}} \approx n_{\text{causal}} / r^2$); a marker with small r^2 to the causal locus, even if physically adjacent, requires a much larger sample to achieve the same power as directly testing the causal variant, and in the extreme (r^2 near 0) essentially no amount of additional sample size at the marker will efficiently recover the causal signal, however close the marker is physically.

Answer to Q6.

```

pA<-160/320; pB<-160/320; pAB<-150/320
D<- pAB - pA*pB
Dmax <-min(pA*(1-pB), (1-pA)*pB)
Dprime<- D/Dmax
r2 <- D^2/(pA*(1-pA)*pB*(1-pB))
c(D=D, Dprime=Dprime, r2=r2)
pA, pB, pAB =160/320, 160/320, 150/320
D = pAB - pA*pB
Dmax =min(pA*(1-pB), (1-pA)*pB)
Dprime = D/Dmax

```

```

r2 = D**2/(pA*(1-pA)*pB*(1-pB))
print(D, Dprime, r2)

```

This haplotype configuration (dominated by AB and ab, with only small amounts of the recombinant Ab and aB haplotypes) gives D^2 very close to 1 (near-complete disequilibrium relative to the maximum possible given the allele frequencies) and a correspondingly high r^2 as well, since both allele frequencies are equal (0.5) — illustrating the special case where, when $p_A = p_B$, D^2 and r^2 tend to agree closely; in general, however, D^2 can be high while r^2 is much lower whenever the two allele frequencies differ substantially (Section 15.4), which is why r^2 (not D^2) is the measure directly relevant to statistical power for association testing (Q5).

Chapter 16: Coalescent Models

Answer to Q1. $E(T) = 2N = 2(5,000) = 10,000$ generations = $10,000 \times 25 = 250,000$ years.

Answer to Q2. $E(T_{\text{MRCA}}) = 4N(1 - 1/n) = 4(8,000)(1 - 1/10) = 32,000 \times 0.9 = 28,800$ generations.

Answer to Q3. As n grows, most of the *additional* lineages coalesce very quickly near the “tips” (recent past) of the genealogy, because the coalescence rate $C(k,2)/(2N)$ is large when k (the number of remaining distinct lineages) is large — so adding more samples mainly adds shallow, recent branching, not additional depth. The final, slowest step — the coalescence of the last two remaining lineages, at rate $1/(2N)$, expected waiting time $2N$ generations — always dominates the total time to the MRCA, regardless of how large n is, which is why $E(T_{\text{MRCA}})$ converges to (rather than grows without bound past) $4N$ generations.

Answer to Q4. $E(S) = \theta \times \sum_{k=1}^{n-1} 1/k = 0.008 \times 3.548 = 0.0284$. (Note: this is the expectation per site typically applied across a genomic region with many independent sites/base pairs in practice; for a single locus with θ this small, an expected value below 1 simply reflects a low probability of observing any segregating site in a very short sequence, consistent with a low value of θ at that specific locus.)

R and Python verification (Q1, Q2, Q4).

```
2*5000*(1-1/10)      # Q1: E[T] generations (2N
convention check)
4*8000*(1-1/10)      # Q2: E[T_MRCA] = 4N(1-1/n)
theta <-0.008
theta *sum(1/(1:9))   # Q4: expected segregating sites,
harmonic sum to n-1=9
import numpy as np
print(2*5000*(1-1/10))
print(4*8000*(1-1/10))
theta =0.008
print(theta *sum(1/k for k in range(1, 10)))
```

Answer to Q6. The bottleneck sharply reduced the effective population size, N , of the founding out-of-Africa lineage for an extended period; since expected coalescence time scales directly with N ($E(T) = 2N$ generations for a pair, Section 16.4), a smaller N during the bottleneck mechanically forces lineages to coalesce much faster during that period than they would have in the larger ancestral African population. Faster coalescence means fewer independent branches survive to accumulate distinct mutations over time, so less genetic variation is preserved through the bottleneck; because mutation since the bottleneck has not yet had enough generations to fully restore the lost diversity, non-African populations — all

descended from this reduced founding group — retain systematically lower genetic diversity today than African populations, whose ancestors never passed through this size reduction.

Chapter 17: Parametric and Non-Parametric Linkage Analysis

Answer to Q1. $\theta = R/(R+N_{-R}) = 2/20 = 0.10$. $\theta^{R(1-\theta)}\{N_{-R}\} = (0.10)^{2(0.90)}\{18\} = 0.01 \times 0.1501 = 0.001501$. $0.5^{20} = 9.537 \times 10^{-7}$. $Z(\theta) = \log_{10}(0.001501/9.537 \times 10^{-7}) = \log_{10}(1573.9) = 3.197$. Since $Z \approx 3.20 \geq 3$, this constitutes conventionally significant evidence of linkage.

Answer to Q2. Combined LOD = $Z_1(0.1) + Z_2(0.1) = 1.2 + 0.9 = 2.1$. This simple addition is valid because LOD scores at a fixed θ are, by construction, log-likelihood-ratio contributions from independent families (log of a product of independent likelihoods = sum of logs), so summing the LOD scores across independent pedigrees, evaluated at the same θ , gives exactly the LOD score for the combined dataset at that θ .

Answer to Q3. Mean IBD = $[0(10)+1(20)+2(30)]/60 = (0+20+60)/60 = 80/60 = 1.333$. This exceeds the null expectation of 1.0, suggesting possible linkage between the marker and an asthma-susceptibility locus, though (as in Worked Example 17.2) a formal significance test would be needed before concluding definitively.

Answer to Q4. Type 2 diabetes is a complex, polygenic trait with an unknown, almost certainly non-Mendelian mode of inheritance (multiple contributing genes, each of modest effect, interacting with environmental/lifestyle factors) — specifying a single dominant/recessive genetic model with

a fixed penetrance, as parametric analysis requires, would misrepresent the true biology and could severely reduce power or even mask true linkage signals; non-parametric (allele-sharing) analysis makes no such assumption and is therefore more appropriate. Sickle-cell disease, by contrast, has a well-established, simple autosomal recessive mode of inheritance with essentially complete penetrance (Chapters 1–2), so a correctly specified parametric model can be used, and — because the correct model is known — parametric analysis will generally deliver higher statistical power than a non-parametric approach for this disorder.

R and Python verification (Q1, Q2, Q3).

```
R<-2; NR<-18; n<-20
theta_hat <- R/(R+NR)
lod <- -log10((theta_hat^R)*(1-theta_hat)^NR) -log10(0.5^n)
c(theta_hat=theta_hat, LOD=lod)          # Q1
1.2+0.9# Q2: combined LOD across families
(0*10+1*20+2*30)/60# Q3: mean IBD sharing
R, NR, n=2, 18, 20
theta_hat = R/(R+NR)

import numpy as np
lod = np.log10((theta_hat**R)*(1-theta_hat)**NR) -
np.log10(0.5**n)
print(theta_hat, lod)
print(1.2+0.9)
print((0*10+1*20+2*30)/60)
```

Answer to Q5. An observed mean IBD sharing of 1.35 among affected sibling pairs, compared with the null expectation of exactly 1.0 under no linkage, is consistent with, and provides evidence *for*, linkage between the marker and a disease locus: affected relatives sharing more genetic material IBD than chance alone would predict is precisely

the signature that non-parametric linkage analysis (Section 17.7) is designed to detect. Whether the excess of 0.35 above the null value of 1.0 is *statistically significant* would need to be assessed formally via the MLS or an equivalent NPL test statistic, accounting for the sample size of 20 pairs.

Answer to Q6. The null hypothesis of no linkage corresponds to $(z_0, z_1, z_2) = (1/4, 1/2, 1/4)$, and under the alternative of true linkage, sharing can only be pushed in one direction (z_2 increased, z_0 decreased) — biologically, affected relatives cannot share *less* than the null-expected amount of IBD due to a disease locus, since sharing more of the disease-linked region is what excess co-inheritance produces. Because the alternative hypothesis parameter space is therefore restricted to only one side of the null value (a “boundary” null hypothesis, exactly the same situation as the one-sided variance-component tests discussed in Chapter 10), the standard regularity conditions for the usual χ^2 asymptotic result do not hold, and the correct null distribution is instead the non-standard 50:50 mixture of a point mass at zero (representing the cases where the constrained MLE falls exactly at the null value) and a χ^2_1 distribution (representing the cases where it falls strictly inside the alternative region).

Chapter 18: Genome-Wide Association Studies (GWAS)

Answer to Q1. The threshold is unchanged by the number of SNPs actually tested in this particular study; it is fixed at $\alpha = 0.05/1,000,000 = 5 \times 10^{-8}$ under the standard convention (based on the estimated total number of effectively independent common-variant tests across the genome). A SNP with $p = 2 \times 10^{-7}$ does **not** meet

this threshold ($2 \times 10^{-7} > 5 \times 10^{-8}$), so it would be classed as, at best, “suggestive,” not genome-wide significant.

Answer to Q2. A systematic, genome-wide inflation of significance (rather than isolated peaks) together with λ_{GC} substantially above 1.0 (here, 1.35) most likely indicates inadequately corrected population stratification (or other systematic confounding, e.g., uncorrected relatedness/cryptic family structure among samples) rather than a genuine excess of true genetic associations. The population-structure assessment and correction step (Section 18.3, step 3 — typically PCA-based ancestry adjustment, possibly needing additional principal components or a mixed-model approach) should be revisited.

Answer to Q3. Because LD decays faster, and haplotype patterns differ, in African (including Nigerian) populations relative to the smaller-founding-population history typical of many European populations (Section 15.6), a genotyped SNP panel designed and validated around European LD patterns will, on average, “tag” (capture, via LD) a smaller fraction of the total genetic variation carried by nearby, untyped variants in a Nigerian sample — and imputation from a reference panel lacking sufficient African haplotype diversity will similarly perform less accurately — both effects reducing statistical power and completeness of genomic coverage relative to a comparably sized European-ancestry GWAS using the same array/panel.

Answer to Q4. A single common-variant SNP with a modest odds ratio (1.15) explains only a small fraction of the overall heritable risk of a complex, polygenic trait like T2DM (Chapter 9’s heritability discussion; Section 18.7’s polygenic picture), so it has very limited value on its own

for individual risk prediction — most patients carrying the risk allele will never develop the disease, and most patients who do develop it will not carry this particular allele. A more informative alternative is a polygenic risk score aggregating the effects of very many (often thousands of) SNPs genome-wide; however, the clinician should further be cautioned that most existing polygenic risk scores were derived predominantly from European-ancestry discovery GWAS and are known to transfer poorly (systematically less accurately) to Nigerian and other African-ancestry patients (Section 18.6), so a polygenic score ideally derived from, or at minimum properly recalibrated for, African-ancestry data should be sought wherever possible.

R and Python verification (Q1).

```
5e-8# genome-wide significance threshold (fixed convention,
independent of study SNP count)
2e-7<5e-8# does  $p = 2 \times 10^{-7}$  meet the threshold? (FALSE)
threshold = 5e-8
p_snp = 2e-7
print(threshold, p_snp < threshold)
```

Answer to Q5. Under severe uncorrected stratification with no true associations, the Q–Q plot would show the *entire* bulk of points lifting away from the diagonal fairly uniformly across the whole range of p-values (not just at the extreme tail), reflecting a general inflation of test statistics genome-wide (high λ_{GC}) with no distinct outliers at the top — a diffuse, whole-plot departure from the diagonal. By contrast, a well-controlled study with a handful of true associations shows the bulk of SNPs (the vast majority, which are truly null) tracking the diagonal line very closely, with only a small number of points at the extreme upper-right departing sharply upward — precisely the pattern shown in the Q–Q plot of Section 18.5.

Answer to Q5. Under severe uncorrected stratification with no true associations, the Q–Q plot would show the *entire* bulk of points lifting away from the diagonal fairly uniformly across the whole range of p-values (not just at the extreme tail), reflecting a general inflation of test statistics genome-wide (high λ_{GC}) with no distinct outliers at the top — a diffuse, whole-plot departure from the diagonal. By contrast, a well-controlled study with a handful of true associations shows the bulk of SNPs (the vast majority, which are truly null) tracking the diagonal line very closely, with only a small number of points at the extreme upper-right departing sharply upward — precisely the pattern shown in the Q–Q plot of Section 18.5.

Answer to Q6. The 5×10^{-8} threshold was calibrated to control the family-wise error rate for the roughly 1 million effectively independent common-variant tests typical of a standard SNP genotyping array (Section 18.4). A whole-genome sequencing study tests many more variants — tens of millions, including a large proportion of rare variants that are less correlated with one another (weaker LD, Chapter 15, Section 15.6) — so the effective number of independent tests is substantially larger, requiring a correspondingly more stringent (smaller) significance threshold to maintain the same overall false-positive control; in practice, and especially once rare-variant burden/gene-based tests (Chapter 20) are also considered, thresholds around 10^{-9} or gene-based thresholds like the 2.5×10^{-6} figure of Appendix A are used instead.

Chapter 19: Genotype Imputation

Answer to Q1. A “hard-called” imputed genotype reports a single, most-probable genotype (e.g., 0, 1, or 2 copies of the alternative allele) at the untyped position, discarding information about how confident that call actually was. A “dosage” instead reports the full posterior expectation of the allele count (a continuous value between 0 and 2, e.g. 1.85) reflecting the underlying genotype-probability distribution produced by the imputation algorithm. Association testing on dosages is preferred because it correctly propagates the imputation uncertainty into the test’s standard errors — a poorly imputed variant (low INFO/r^2 , Section 19.5) will have dosages clustered less tightly around integer values, appropriately down-weighting its contribution and protecting against overconfident false-positive or false-negative conclusions that a hard-call approach could produce by treating every imputed genotype as if it were directly, perfectly observed.

Answer to Q2. (i) An early, largely European reference panel contains comparatively little of the additional haplotype diversity present in West African populations (which, having larger and more ancient effective population sizes, carry more distinct haplotypes overall — Chapter 15, Section 15.6), so many haplotypes actually present in a Nigerian study sample would simply have no closely matching counterpart anywhere in the panel. (ii) Faster LD decay in African populations (Section 15.6) means that, even where a broadly similar haplotype background exists, the physical distance over which a genotyped tag SNP remains strongly correlated with a nearby untyped variant is shorter, so a reference panel and genotyping array designed and validated primarily on European LD patterns

will, on average, tag a smaller fraction of the variation actually present in a Nigerian genome.

Answer to Q3. No — a variant imputed with $\text{INFO}/r^2 = 0.25$ falls below the commonly used inclusion thresholds (typically 0.3 or, more stringently, 0.8) described in Section 19.5, indicating substantial uncertainty in its imputed genotypes; such a poorly imputed variant would ordinarily be **excluded** from the primary association analysis (or, at minimum, flagged for cautious secondary interpretation only), since including it risks contributing mostly noise, and — at very low quality — can introduce spurious associations rather than genuine signal.

Answer to Q4. Strengthening the LD between SNP1 (the tag) and SNP2 (the untyped marker being imputed) — for instance by increasing the coefficient linking $p2$ to snp1_ref in the simulation — makes each SNP1 genotype class a more reliable predictor of the corresponding SNP2 genotype, because the two loci’s alleles co-occur on the same haplotypes more consistently (higher r^2 , Chapter 15, Section 15.4). This directly increases the accuracy of the simple “modal genotype” imputation rule, and hence increases the masked-genotype concordance rate observed in the study sample — exactly mirroring why real imputation methods perform best for variants in strong LD with genotyped markers and perform poorly for weakly-tagged variants.

```
set.seed(19)
n_ref <- 500
snp1_ref <- rbinom(n_ref, 2, 0.4)
# stronger LD: larger coefficient linking snp2 to snp1
snp2_ref <- pmin(pmax(round(rbinom(n_ref, 2,
pmin(pmax(0.4*0.95+ (snp1_ref-0.8)*0.30/2, 0), 1))), 0), 2)
```

```

lookup <- tapply(snp2_ref, snp1_ref, function(x)
as.integer(names(sort(table(x), decreasing=TRUE))[1]))
n_study <- 300
snp1_study <- rbinom(n_study, 2, 0.4)
snp2_study_true <- pmin(pmax(round(rbinom(n_study, 2,
pmin(pmax(0.4*0.95+ (snp1_study-0.8)*0.30/2, 0), 1))), 0), 2)
snp2_imputed <- lookup[as.character(snp1_study)]
mean(snp2_imputed == snp2_study_true)

```

```

import numpy as np
from collections import Counter
rng = np.random.default_rng(19)
n_ref = 500
snp1_ref = rng.binomial(2, 0.4, n_ref)
p2 = np.clip(0.4*0.95+ (snp1_ref -0.8)*0.30/2, 0, 1)
snp2_ref = np.clip(np.round(rng.binomial(2, p2)), 0, 2)
lookup = {g: Counter(snp2_ref[snp1_ref==
g]).most_common(1)[0][0] for g in [0,1,2] if
np.any(snp1_ref==g)}
n_study = 300
snp1_study = rng.binomial(2, 0.4, n_study)
p2s = np.clip(0.4*0.95+ (snp1_study -0.8)*0.30/2, 0, 1)
snp2_study_true = np.clip(np.round(rng.binomial(2, p2s)), 0, 2)
snp2_imputed = np.array([lookup[g] for g in snp1_study])
print(np.mean(snp2_imputed == snp2_study_true))

```

Answer to Q5. Imputation solves a “cost” problem because it allows a cheaply genotyped array (a few hundred thousand to a million SNPs) to yield inferred genotypes at tens of millions of variants, in silico, at a small fraction of the cost of directly sequencing every participant (Section 19.2). It solves a “meta-analysis harmonisation” problem because studies genotyped on different arrays (with different, only partially overlapping SNP sets) can each be imputed up to the same dense reference panel, producing

a common set of variants that can be jointly analysed or meta-analysed across studies, rather than being limited to the small intersection of SNPs common to every array.

Answer to Q6. Including both the “high confidence” (5.2 million) and “usable with caution” (1.9 million) tiers gives 7.1 million of the 8 million imputed variants, i.e. $7.1/8 = 88.75\%$. The main risk of this more permissive threshold is that a meaningful fraction of the additional 1.9 million included variants have less reliable imputed genotypes (more residual uncertainty, INFO between 0.3 and 0.8), increasing the chance that noise from poorly imputed genotypes contributes false-positive or attenuated-effect associations to the downstream analysis — the fundamental completeness-versus-reliability trade-off discussed in Section 19.5.

Chapter 20: Genome Sequencing, Rare-Variant Methods, and Simulation-Based Approaches

Answer to Q1. A genotyping array can only measure genotypes at a fixed, pre-specified panel of previously known SNPs chosen when the array was designed (Section 19.2/20.2); by construction it cannot report anything about a variant that was unknown, and therefore not included on the array, at design time. Whole-genome sequencing instead reads the DNA sequence directly, base by base, and so will detect and report *any* variant present in the sequenced sample relative to the reference genome — including a truly novel variant never previously observed in any human — because detection does not depend on the variant having been anticipated in advance.

Answer to Q2. A SKAT-type variance-component test would be expected to have greater power here (Section 20.4).

Because the gene's rare variants act in **mixed directions** (10 increasing and 5 decreasing risk), a simple burden test — which collapses all variants into a single directional score, implicitly assuming they act consistently — would have its true signal partially cancelled out, whereas a variance-component test evaluates whether the *variance* of variant effects departs from zero without requiring a consistent direction, and is therefore considerably more robust to exactly this kind of mixed-effect scenario.

Answer to Q3. Increasing the sample size from 200 to 2,000 individuals (400 to 4,000 haplotypes) will substantially increase the **total number** of variants observed, because a larger sample captures more of the rare variation segregating in the population (many rare variants, by definition, are carried by very few haplotypes and are simply missed entirely in a small sample). However, the **proportion** of variants classified as rare ($\text{MAF} < 1\%$) should increase further still, not merely stay constant, because coalescent theory (Chapter 16) predicts that the site-frequency spectrum is heavily weighted toward rare variants — most mutations in a genealogy occur on short, low-frequency branches near the tips of the tree — so a larger sample disproportionately reveals additional rare, previously unobserved variants relative to the (already largely captured) common variants.

```
import msprime, numpy as np
for n in [200, 2000]:
    ts = msprime.sim_ancestry(samples=n,
    population_size=10_000, sequence_length=1_000_000,
    recombination_rate=1e-8,
    random_seed=20)
    mts = msprime.sim_mutations(ts, rate=1e-8,
    random_seed=21)
```

```

geno = mts.genotype_matrix()
freqs = geno.mean(axis=1)
n_rare = np.sum((freqs > 0) & (freqs < 0.01))
print(n, mts.num_sites, n_rare, n_rare / mts.num_sites if
mts.num_sites else 0)

```

Answer to Q4. Lowering *odds_ratio* from 2.0 toward 1.2 should **decrease** the estimated power monotonically, since a smaller true effect size produces a smaller difference in disease probability between carriers and non-carriers of the gene's rare-variant burden, making the association harder to distinguish from chance in any dataset of fixed size. This illustrates the general, universal relationship in statistical power analysis: for a fixed sample size and significance threshold, statistical power increases with the true effect size (and, separately, with sample size and allele/carrier frequency), and every rare-variant study design must contend with the fact that plausible true effect sizes for individual genes are often modest, requiring very large sequenced sample sizes to achieve adequate power.

for orv in [2.0, 1.6, 1.2]:

print(orv, simulate_burden_power(odds_ratio=orv))

Answer to Q5. The coalescent is far more efficient for neutral variation in a single large population because it only needs to simulate the genealogical history of the individuals actually *sampled* — typically a few hundred to a few thousand lineages — tracing their ancestry backward only until they have all coalesced to their most recent common ancestor(s), regardless of how large the true underlying population was; it never needs to represent the millions of individuals in the population who left no descendants in the sample. Forward-in-time simulation, by contrast, must simulate the entire population, every individual, every generation, moving forward from the ancestral population

to the present — a computational burden that scales with total population size and number of generations, not just sample size. Forward-in-time simulation becomes necessary, however, whenever the process being modelled depends on information the backward-only coalescent cannot represent — most importantly, **natural selection** acting on a locus (since an individual's probability of leaving descendants, and hence the very shape of the genealogy, depends on their fitness, which must be tracked forward through the generations as it changes), as well as complex, explicitly-modelled pedigree or non-neutral demographic scenarios (Chapter 6).

Answer to Q6. Prediction accuracy increases with heritability because heritability, $h^2 = V_A/V_P$ (Chapter 9, Section 9.4), directly measures the proportion of total phenotypic variance that is attributable to additive genetic effects — the only component that genomic prediction methods can, even in principle, hope to capture from genotype data. As h^2 increases, a larger share of the observed phenotypic variation is genuinely genetic (and therefore predictable from genotypes) rather than environmental noise (which no genotype-based method can predict), so every prediction strategy's achievable accuracy rises toward its ceiling as heritability approaches 1; conversely, as h^2 approaches 0, genetic prediction accuracy for any method must approach zero as well, since there is almost no genetic signal left in the phenotype to predict.