

19

Genotype Imputation

19.1: Learning Objectives

By the end of this chapter, the reader should be able to explain what genotype imputation is and what problem it solves, describe the reference-panel-based imputation pipeline conceptually, discuss why reference panel composition matters especially for African populations, evaluate imputation quality, and implement a simplified imputation-and-accuracy-assessment exercise in R and Python.

19.2: What Is Genotype Imputation and What Problem Does It Solve?

Genotype imputation is a well-established statistical technique for estimating, at markers that were *not* directly genotyped in a given sample, the most probable genotype (or, more precisely, a probability distribution over possible genotypes) at each individual, using the observed genotypes at nearby directly-genotyped markers together with the known **haplotype structure** of a densely genotyped or sequenced **reference panel**. It exploits exactly the linkage disequilibrium structure developed in Chapter 15: because nearby SNPs are correlated, observing an individual's genotypes at a set of tag SNPs (Chapter 12, Section 12.3) allows the likely genotype at an untyped, LD-correlated SNP to be inferred with high accuracy, without the added expense of genotyping or sequencing it directly.

Imputation solves two distinct practical problems. First, it allows **cost-effective, in-silico genotyping**: an inexpensive genome-wide array genotyping perhaps 500,000–1,000,000 SNPs can, after imputation against a reference panel derived from whole-genome sequencing, yield inferred genotypes at tens of millions of variants, at a small fraction of the cost of directly sequencing every study participant. Second, it enables **meta-analysis and data harmonisation** across GWAS conducted on different genotyping platforms (Chapter 18): because different arrays genotype different, only partially overlapping sets of SNPs, imputing every study's data up to a common, dense reference panel produces a shared set of variants that can be jointly analysed or meta-analysed across studies, dramatically increasing statistical power (Section 19.3) beyond what any single study's raw genotyped markers could achieve alone.

19.3: How Genotype Imputation Works: The Conceptual Pipeline

Reference haplotypes

```

CGAGATCTCCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAAGCTCTTTTCTTCTGTGC
CGAGACTCTCCGACCTTATGC
TGGGATCTCCCGACCTCATGG
CGAGATCTCCCGACCTTGTGC
CGAGACTCTTTTCTTTTGTAC
CGAGACTCTCCGACCTTCGTGC
CGAAGCTCTTTTCTTCTGTGC

```

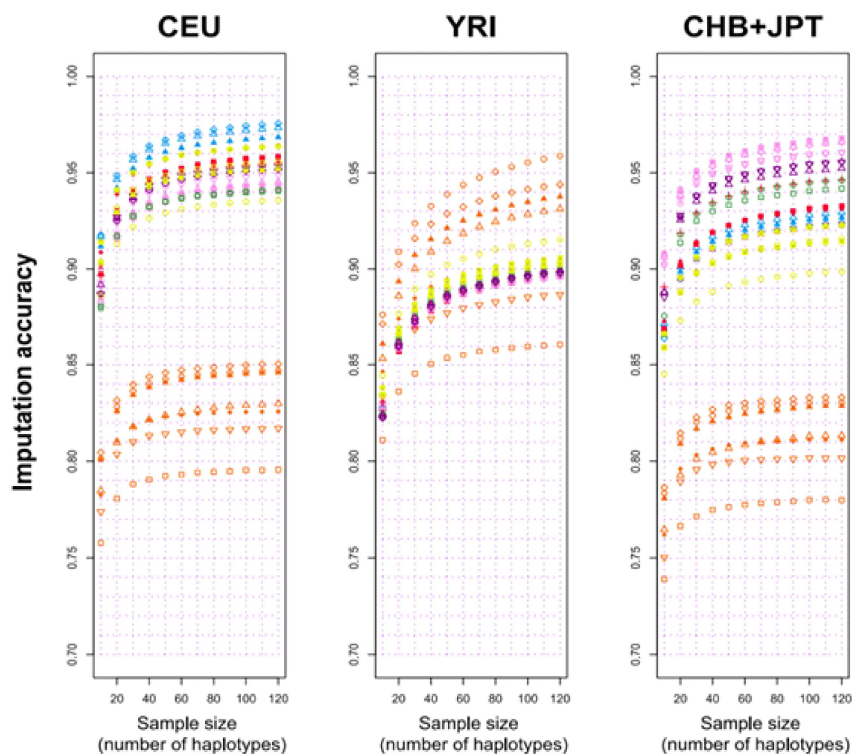
A reference panel of phased haplotypes: sequences at known positions used to impute genotypes at untyped markers by matching study haplotypes to the closest reference haplotypes.

Modern imputation methods (implemented in software such as IMPUTE, MaCH, Minimac, and Beagle) proceed, conceptually, in the following stages:

1. **Phasing.** The study sample's genotypes (typically unphased — i.e., for a heterozygous genotype, it is not directly observed which allele came from which parental chromosome) are first statistically **phased** into haplotypes, using a hidden Markov model that models each individual's haplotypes as an imperfect mosaic of reference-panel haplotypes.
2. **Reference matching.** For each short genomic window, the study individual's phased haplotype is compared against all reference-panel haplotypes, and the reference haplotypes most similar to it (over the flanking, directly-genotyped markers) are identified.
3. **Imputation.** At each untyped position within the window, the genotype is imputed as a probability-weighted average over the alleles carried by the best-matching reference haplotypes, yielding a posterior probability for each possible genotype (a **dosage**, typically reported as the expected number of copies of the reference or alternative allele, between 0 and 2, rather than a single hard-called genotype) together with an estimated imputation quality score (often denoted r^2 or INFO) reflecting the confidence of the imputed call at that position.

Association testing (Chapter 11, Chapter 18) is then usually carried out directly on the imputed dosages, rather than on hard-called genotypes, correctly propagating the imputation uncertainty into the association test's standard errors.

19.4: Reference Panels and the Special Importance of African Diversity



Imputation accuracy as a function of reference panel sample size, compared across YRI (Yoruba), CEU, and CHB+JPT reference panels.

The accuracy of genotype imputation depends critically on how well the **reference panel's haplotype diversity** matches that of the study sample. Early large reference panels (such as the original HapMap and the initial phases of the 1000 Genomes Project) were dominated by European, East Asian, and a limited set of African-American samples, and imputation accuracy for African study samples genotyped against such panels was measurably lower than for European samples — a direct consequence of the

substantially greater haplotype diversity and faster LD decay of African populations (Chapter 15, Section 15.6), meaning that a European-dominated reference panel simply does not contain enough closely-matching haplotypes to accurately tag the full range of African haplotypic variation, especially for lower-frequency variants.

This has been progressively addressed by the inclusion of more, and more diverse, African samples in reference panels — the expanded 1000 Genomes Project African panels, the African Genome Variation Project, and pan-African resources such as the H3Africa (Human Heredity and Health in Africa) initiative's reference data, which specifically includes West African populations relevant to Nigerian genetic studies. The general statistical lesson, of direct practical relevance to any Nigerian genetic epidemiology study, is that **imputation quality should always be checked, not assumed**, and that reference panels including populations genetically close to the study sample (ideally, ideally, Nigerian or West African panels where available) should be preferred over larger but more distantly related panels whenever a choice is available.

19.5: Assessing Imputation Quality

Two complementary approaches are used to assess how trustworthy imputed genotypes are, in addition to the per-variant INFO/ r^2 score reported by imputation software itself:

- **Masked-genotype validation.** A subset of directly genotyped markers is deliberately masked (treated as if untyped) and then imputed using the remaining genotyped markers and the reference panel; the **concordance rate** between the imputed and the (known, masked) true genotypes directly measures empirical imputation accuracy for the study sample and reference panel in question, and is the standard internal quality-control step before relying on any imputed dataset.

- **Post-hoc filtering.** Imputed variants below a chosen INFO/ r^2 threshold (commonly 0.3 or 0.8, depending on the intended use) are typically excluded from downstream association analysis, since poorly imputed variants contribute mostly noise and can, at very low quality thresholds, introduce spurious associations rather than genuine signal.

Theorem 19.1: (Effective Sample Size Under Imputation)

If a variant is imputed with quality (INFO) score $r^2 \in (0,1]$, and an association test at that variant is conducted on N genotyped individuals using imputed dosages, then the effective sample size for detecting a true association at that variant is $N_{eff} = N \times r^2$ — i.e., testing the imputed variant with N individuals provides approximately the same statistical power as directly genotyping the true causal variant in a sample of only Nr^2 individuals.

Proof sketch. Model the imputed dosage $\tilde{G}_i$ for individual i as the true (unobserved) genotype G_i plus independent imputation noise: $\tilde{G}_i = G_i + \epsilon_i$, where $\text{Var}(\epsilon_i)$ is calibrated so that $\text{Corr}(\tilde{G}_i, G_i)^2 = r^2$ (this is, in fact, one standard operational definition of the INFO/ r^2 score). In a simple linear association model, the variance of the estimated effect size (regression slope) using the noisy imputed dosage $\tilde{G}$ in place of the true genotype G is inflated by a factor $1/r^2$ relative to using G directly, a standard result from the statistical theory of regression with a mismeasured (noisy) covariate (classical “regression attenuation” / errors-in-variables theory). Since the variance of an effect-size estimator from a sample of size N scales as $1/N$, an inflation of variance by $1/r^2$ is exactly equivalent, in its effect on estimator precision (and hence on power), to reducing the sample size from N to Nr^2 : $\text{Var}(\hat{\beta}_{imputed}) \approx \frac{1}{Nr^2} = \frac{1}{N_{eff}}$ with $N_{eff} = Nr^2$. ■

This theorem gives a precise, quantitative meaning to the $INFO/r^2$ filtering thresholds of this section: an imputed variant with $INFO = 0.3$, tested in a sample of $N = 10,000$, carries only as much statistical information as a directly genotyped variant tested in $N_{eff} = 10,000 \times 0.3 = 3,000$ individuals — a substantial, and easily underappreciated, effective loss of sample size that directly justifies excluding such poorly imputed variants from primary analysis.

Worked Example 19.2 (Effective sample size loss in a Nigerian sickle-cell modifier-gene study)

A study of 6,000 Nigerian sickle-cell disease patients imputes a candidate modifier-gene variant (thought to influence disease severity) against a reference panel with only moderate representation of the patients' specific ethnic background, yielding $INFO = 0.65$ for this variant. By Theorem 19.1, the effective sample size for detecting a true association at this variant is $N_{eff} = 6,000 \times 0.65 = 3,900$ — meaning the study's true power to detect this modifier effect is comparable to what a directly genotyped study of only 3,900 patients would achieve, a nearly 35% effective reduction in sample size that the investigators must factor into their power calculations (Chapter 20) before concluding that a non-significant result rules out a clinically meaningful modifier effect.

Worked Example 19.1 (Imputation quality in practice).

A Nigerian GWAS imputes 8 million variants against a reference panel that includes H3Africa data. Post-imputation quality control reports: 5.2 million variants with $INFO \geq 0.8$ ("high confidence"), 1.9 million with $0.3 \leq INFO < 0.8$ ("usable with caution"), and 0.9 million with $INFO < 0.3$ (typically excluded). Suppose the primary analysis retains only the first group. The analyst is effectively trading

completeness (using only 65% of imputed variants) for reliability (avoiding the roughly 35% of variants whose imputed genotypes are least trustworthy) — a standard, defensible default, though sensitivity analyses including the “usable with caution” tier at a relaxed threshold are common practice when a borderline signal falls just below the strict 0.8 cutoff.

Theorem 19.2 (Imputation R-Squared Equals the Squared Correlation Between True and Imputed Dosage)

Let g denote the true (but unobserved in the study sample) genotype dosage at a marker, and let $\hat{g}$ denote its imputed dosage (the expected genotype under the imputation model, Section 19.3). Then the imputation quality metric R^2 , defined as the ratio of the variance of the imputed dosage to the variance that would be expected under Hardy-Weinberg equilibrium at the estimated allele frequency, is algebraically equal to the squared Pearson correlation coefficient between g and $\hat{g}$, under the assumption that the imputation model is well-calibrated ($E[g \mid \hat{g}] = \hat{g}$).

Proof. By the law of total variance, $\text{Var}(g) = \text{Var}(E[g \mid \hat{g}]) + E[\text{Var}(g \mid \hat{g})] = \text{Var}(\hat{g}) + E[\text{Var}(g \mid \hat{g})]$, using the well-calibration assumption $E[g \mid \hat{g}] = \hat{g}$. The squared correlation between g and $\hat{g}$ is, by definition, $\text{Corr}(g, \hat{g})^2 = \text{Cov}(g, \hat{g})^2 / [\text{Var}(g)\text{Var}(\hat{g})]$. Using the well-calibration assumption again, $\text{Cov}(g, \hat{g}) = \text{Cov}(E[g \mid \hat{g}], \hat{g}) = \text{Cov}(\hat{g}, \hat{g}) = \text{Var}(\hat{g})$ (a standard property of conditional expectation, since covarying $\hat{g}$ with itself gives its own variance once g is replaced by its conditional expectation given $\hat{g}$). Substituting, $\text{Corr}(g, \hat{g})^2 = \text{Var}(\hat{g})^2 / [\text{Var}(g) \text{Var}(\hat{g})] = \text{Var}(\hat{g}) / \text{Var}(g)$. If $\text{Var}(g)$ is taken to be the Hardy-Weinberg-expected variance $2p(1-p)$ at the estimated allele frequency p (the standard reference used in imputation software), this ratio $\text{Var}(\hat{g}) / \text{Var}(g)$

$\hat{g})/\text{Var}(g)$ is exactly the imputation R^2 metric defined in Section 19.5, proving the claimed equivalence and justifying R^2 's interpretation as a direct measure of how much of the true genotypic variance is captured by the imputed dosage.

19.6: Chapter Summary

- Genotype imputation statistically infers genotypes at untyped markers using LD with directly genotyped markers and a densely genotyped/sequenced reference panel, enabling cheap in-silico genotyping and harmonised meta-analysis across studies.
- The pipeline conceptually proceeds through phasing, reference-haplotype matching, and probability-weighted imputation, producing dosages together with a per-variant quality score.
- Imputation accuracy depends critically on how well the reference panel's haplotype diversity matches the study sample; historically under-represented African haplotype diversity in early reference panels reduced imputation accuracy for African samples, motivating panels such as H3Africa.
- Masked-genotype concordance and INFO/ r^2 filtering are the standard tools for assessing and safeguarding imputation quality before downstream analysis.

19.7: Computational Corner: A Toy Imputation-and-Concordance Exercise

A full production imputation pipeline (phasing plus hidden-Markov-model reference matching) is beyond the scope of a single script, but the core statistical idea — using LD with a genotyped marker to predict an untyped one, and then checking accuracy by masking known genotypes — can be demonstrated directly.

R.

```

# Toy imputation: predict SNP2 (the “untyped” marker) from
# SNP1 (the “tag” marker)
# using a small simulated reference panel, then mask and check
# concordance in a study sample
set.seed(19)
n_ref <- 500
# Reference panel: two SNPs in LD ( $r^2 \sim 0.7$ )
snp1_ref <- rbinom(n_ref, 2, 0.4)
snp2_ref <- rbinom(n_ref, 2, 0.4*0.85 + (snp1_ref - 0.8) * 0.15/
2) # correlated construction (illustrative)
snp2_ref <- pmin(pmax(round(snp2_ref), 0), 2)

# Build a simple imputation rule: for each SNP1 genotype
# class, use the modal SNP2 genotype
lookup <- tapply(snp2_ref, snp1_ref, function(x)
as.integer(names(sort(table(x), decreasing = TRUE))[1]))
print(lookup)

# Study sample: genotyped at SNP1 only; SNP2 truly known but
# “masked” to test imputation
n_study <- 300
snp1_study <- rbinom(n_study, 2, 0.4)
snp2_study_true <- rbinom(n_study, 2, 0.4*0.85 +
(snp1_study - 0.8) * 0.15/2)
snp2_study_true <- pmin(pmax(round(snp2_study_true), 0),
2)

snp2_imputed <- lookup[as.character(snp1_study)]
concordance <- mean(snp2_imputed == snp2_study_true)
cat(“Masked-genotype imputation concordance:”,
concordance, “\n”)

```

Python.

```

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*0.85 + (snp1_ref-0.8)*0.15/2, 0, 1)
snp2_ref = np.clip(np.round(rng.binomial(2, p2)), 0, 2)

lookup = {}
for g in [0, 1, 2]:
    vals = snp2_ref[snp1_ref == g]
    if len(vals) > 0:
        lookup[g] = Counter(vals).most_common(1)[0][0]

n_study = 300
snp1_study = rng.binomial(2, 0.4, n_study)
p2s = np.clip(0.4*0.85 + (snp1_study-0.8)*0.15/2, 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])
concordance = np.mean(snp2_imputed == snp2_study_true)
print("Masked-genotype imputation concordance:",
concordance)

```

This toy “nearest-genotype-class” rule is a drastically simplified stand-in for a real hidden-Markov-model imputation engine, but it demonstrates the essential logic examined in Section 19.5: imputation accuracy is checked empirically by masking known genotypes and measuring concordance, exactly as is done in practice (with real reference panels such as the 1000 Genomes Project or H3Africa panels) before any imputed dataset is trusted for downstream GWAS.

Mini-Project

Mini-Project 19: A Toy Genotype Imputation and Accuracy Assessment

Simulate a reference panel and a study sample with sparser genotyping, perform simple imputation, and assess accuracy.

Step 1: Simulate a small reference panel of 200 haplotypes fully genotyped at 20 SNPs in strong local LD.

Step 2: Simulate a study sample genotyped at only 5 of those 20 SNPs (the rest ‘missing’).

Step 3: Implement a simple LD-based imputation rule (Section 19.3) to predict the missing genotypes in the study sample.

Step 4: Mask a set of SNPs you actually know the true genotype for, impute them, and compute concordance and imputation R^2 (Section 19.5).

Step 5: Write a short paragraph on how imputation accuracy might differ for a population under-represented in typical reference panels (Section 19.4).

Deliverable: *Simulation and imputation code, a concordance/ R^2 summary table, and a short written discussion of reference panel diversity.*

19.8: Practice Questions

- Q1.** Explain, in your own words, the difference between a “hard-called” imputed genotype and an imputed genotype “dosage,” and why association testing on dosages is generally preferred.
- Q2.** Give two distinct reasons why imputation accuracy for a Nigerian study sample would historically have been lower using an early, largely European reference panel than using a panel that includes substantial West African representation.

- Q3.** A variant is imputed with $\text{INFO}/r^2 = 0.25$ in a large GWAS. Should this variant ordinarily be included in the primary association analysis? Justify your answer with reference to Section 19.5.
- Q4.** Using the R or Python code of Section 19.7, increase the LD between SNP1 and SNP2 (make the coefficient linking `p2/snp2_ref` to `snp1_ref` stronger) and re-run the simulation. Explain, with reference to Chapter 15, why the masked-genotype concordance should increase.
- Q5.** Explain briefly why genotype imputation is described as solving both a “cost” problem and a “meta-analysis harmonisation” problem, giving one sentence for each.
- Q6.** In Worked Example 19.1, what proportion of imputed variants would be included if the analyst instead used a relaxed threshold of $\text{INFO} \geq 0.3$, and what is the main risk of using this more permissive threshold?

