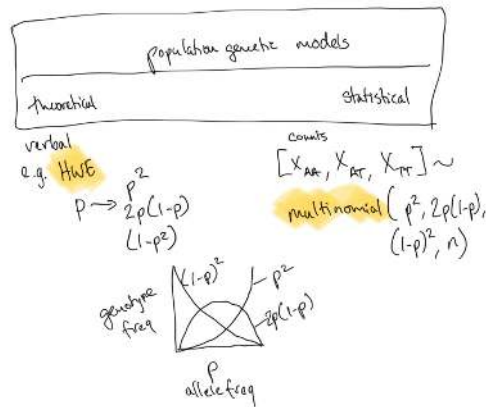


A probability foundation for population genomics

1.1 Why do we need probability theory for genomics?

- ⇒ estimate parameters and model parameters
 - a. example: genotype probabilities/likelihoods to represent uncertainty after seeing seq. reads
 - b. population allele frequencies
 - c. theoretical model of allele freq.



1.2 Estimation of allele frequencies

Suppose 100 individuals: 63 AA, 34 AT, 3 TT

↳ What is ' p ', the frequency of 'A' allele?

$$63 \times 2 + 34 = 160 \quad \frac{160}{200} = 0.8 \leftarrow \text{maximum likelihood estimate}$$

Build probability model for allele freq

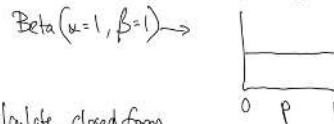
↳ Bayesian probability estimate of $p \rightarrow$ posterior probability density (no MCMC)

$$P(p | \text{data}) = \frac{P(\text{data} | p) \cdot P(p)}{P(\text{data})}$$

Annotations: $P(p | \text{data})$ is posterior probability; $P(\text{data} | p)$ is likelihood; $P(p)$ is prior probability; $P(\text{data})$ is normalizing value.

(a) $P(\text{data} | p) \rightarrow$ Binomial - set of ' n ' Bernoulli trials

(b) $P(p) \rightarrow [0, 1]$
uniform - equally likely



2. Calculate closed form
Solution for the posterior distribution proportional

$$(a) P(p | x, n) \propto P(x | p, n) \cdot P(p)$$

Annotations: $x=160$, $n=200$

$$C = \binom{n}{x} = \frac{n!}{x!(n-x)!} \quad C p^x (1-p)^{n-x} \cdot K p^{\alpha-1} (1-p)^{\beta-1}$$

$$P(p | x, n) = C \cdot K p^{x+\alpha-1} (1-p)^{n-x+\beta-1}$$

Beta($x+\alpha, n-x+\beta$)

Beta($160+1, 40+1$)

$$E(p) = \frac{\alpha}{\alpha + \beta} = \frac{161}{161 + 41}$$

1.3 Theoretical and statistical models for allele frequencies

1. single locus model — beta-binomial

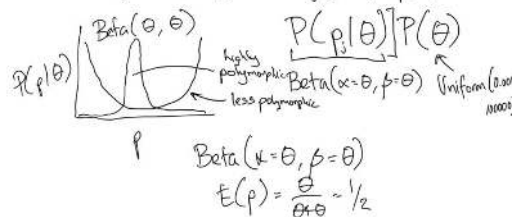
$$P(p|x, n) \propto P(x|p, n) \cdot P(p)$$

2. multilocus model j -loci

$$P(\vec{p} | \vec{x}, \vec{n}) \propto \prod_j \underbrace{P(x_j | p_j, n_j)}_{\text{binomial}} \underbrace{P(p_j)}_{\text{Beta}(1,1)}$$

3. multilocus model for allele frequency and diversity

$$P(\vec{p}, \theta | \vec{x}, \vec{n}) \propto \prod_j [P(x_j | p_j, n_j)]$$



in theory $\rightarrow \theta = 4Nm$ if drift and mutation are the only processes on allele freq. $\rightarrow$ the pop will equilibrate to a beta(θ, θ)

Dirichlet

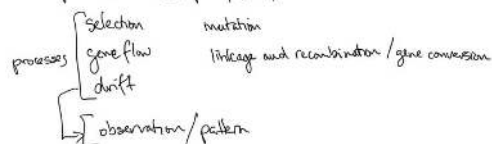
$$P(p|\theta)$$

What if you were interested in heterozygosity?

When calculate $P(p|\theta)$, we can estimate of transformation of 'p': $H = 2p(1-p)$

1.4 The F-model of population differentiation

processes that shape pop diff:



pop diff $\rightarrow F_{ST}$ is a measure of variance in allele freq. among pop's

many definitions of F_{ST}

1. deterministic, fixed effects parameter that is a transformation of allele freq. $\rightarrow G_{ST} = (H_T - H_S) / H_T$

2. random effect parameter, evolutionary parameter

Weir and Cockerham's $F_{ST} = \theta_{ST}$

F-models

1.4.1 Theoretical, generative F-model

$$P(p | \alpha: \pi \theta, \beta: (1-\pi) \theta) \quad \text{F-model}$$

$$E(p) = \frac{\pi \theta}{\pi \theta + (1-\pi) \theta} = \frac{\pi \theta}{\theta} = \pi \quad \leftarrow \text{mean}$$

$$\theta = \frac{1}{F_{ST}} - 1$$

$$\theta = \text{precision} = \frac{1}{\text{variance}}$$

1. Infinite-island model $\rightarrow$ where π is the allele freq in the migrant pool and $\theta = 4Nm$

2. Divergence from a common ancestor

