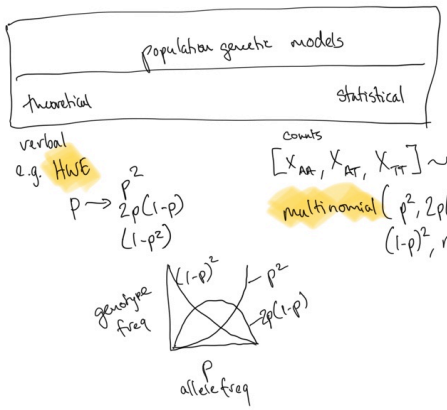
A probability foundation for population genomics

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?

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

Build probability model for allele freq

↳ Bayesian probability estimate of p → posterior probability density (no MC/MC)

likelihood

$$P(p \mid data) = \frac{P(data \mid p)\,P(p)}{P(data)}$$

posterior probability

prior probability

normalizing value

(a) $P(data \mid p) \rightarrow$ Binomial – set of 'n' Bernoulli trials

(b) $P(p) \rightarrow [0, 1]$

uniform – equally likely

Beta($\alpha=1, \beta=1$) →

2. Calculate closed form solution for the posterior distribution

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

160 200

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

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

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

$$\text{Beta}(160+1,\ 40+1)$$

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

1.3 Theoretical and statistical models for allele frequencies

1. single locus model — beta-binomial

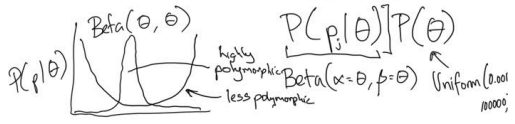
$$P(p|x,n) \propto P(x|p,n) \cdot P(p)$$

2. multiloc us model — j-loci

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

3. multiloc us model for allele frequency and diversity

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



$$\text{Beta}(\alpha=\theta,\beta=\theta)$$

$$E(p) = \frac{\theta}{\theta+\theta} = 1/2$$

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

Dirichlet

$P(p|\theta)$

What if you were interested in heterozygosity?

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

1.4 The F-model of population differentiation

processes that shape pop diff:

- processes
  - selection
  - gene flow
  - drift
  - mutation
  - linkage and recombination / gene conversion
- → observation / pattern

pop diff → 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. → $G_{ST} = (H_T - H_S)/H_T$
2. random effect parameter, evolutionary parameter
   - Weir and Cockerham's $F_{ST} = \theta_{ST}$
   - F-models

1.4.1 Theoretical, generative F-model

$$\boxed{P(p|\alpha=\pi\theta,\ \beta=(1-\pi)\theta)}$$ F-model

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

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

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

1. Infinite-island model → where π is the allele freq in the migrant pool and $\theta = 4Nm$
2. Divergence from a common ancestor