1.1 Why do we need probability theory for genomics?

Estimate parameters and model parameters

a. Example: genotype probabilities/inclination to express uncertain after seeing seq. reads
b. Population allele frequencies

c. Theoretical model of allele freq

<table>
<thead>
<tr>
<th>Population genetic models</th>
<th>Statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotypic</td>
<td>Genotypic</td>
</tr>
<tr>
<td>Morphological</td>
<td>Phenotypic</td>
</tr>
<tr>
<td>(p, 2q(p), (1-p), r)</td>
<td>(x_0, x_1, x_2, x_3)</td>
</tr>
<tr>
<td></td>
<td>\text{maximized} (p, 2q(p), (1-p), r)</td>
</tr>
<tr>
<td></td>
<td>P(\text{allele})</td>
</tr>
</tbody>
</table>

1.2 Estimation of allele frequencies

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

L: What is the frequency of 'A' allele?

63 + 2 x 34 = 160

\[ \hat{p} = \frac{160}{200} = 0.8 \]

Maximum likelihood estimate

The probability model for allele freq

L: Bayesian probability estimate of \( p \rightarrow \) posterior probability density (no MCMC)

\[ P(p | \text{data}) \propto P(\text{data} | p) \cdot P(p) \]

(a) \( P(\text{data} | p) \rightarrow \text{Binomial} \sim \text{set of } n' \)

Bernoulli trials

(b) \( P(p) \rightarrow \text{Beta}(u+1, \beta+1) \)

uniform - equally likely

\[ \text{Beta}(u+1, \beta+1) \rightarrow \frac{1}{BP(u, \beta)} \]

2. Calculate closed form

Solution for the posterior distribution

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

\[ C_n(x) = \frac{n!}{x!(n-x)!} \]

\[ P(x | p, n) \propto p^x (1-p)^n-x \cdot \text{Beta}(x+k, n-x+\beta) \]

\[ P(p | x, n) \propto \frac{C_n(x)}{K_p(p)} \]

\[ E(p) = \frac{x + \alpha}{\alpha + \beta} = \frac{161}{160 + 41} \]
1.3 Theoretical and statistical models for allele frequencies

1. Single locus model — beta-binomial
   \[ P(x|n) \propto P(x|p,n) \cdot P(\theta) \]

2. multilocus model
   \[ P(x|n) \propto \prod_{j=1}^{J} P(x_j|n_j) \cdot P(\theta) \]

3. multilocus model for allele frequency and diversity
   \[ P(x, \theta | n) \propto \prod_{j=1}^{J} P(x_j|n_j) \cdot P(\theta) \]

\[ \theta \sim \text{Beta}(\alpha, \beta) \]
\[ E(\theta) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(\theta) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \]

in theory \( \Theta \approx 4N_{A/m} \)

Dirichlet \( P(p|\theta) \)

What if we were interested in heterozygosity?

When calculating \( P(p|\theta) \), we can estimate \( \Theta \) from the observed diversity.

1.4 The F-model of population differentiation

processes that shape pop diff:

- Selection
- Drift
- Gene flow
- Drift and recombination / gene conversion

\[ p_{TOT} \rightarrow F_{ST} \text{ is a measure of variance in allele freq among pops} \]

Many definitions of \( F_{ST} \)

1. deterministic, fluid effect parameter that is a transformation of allele freq:
   \[ Q = (1 - 2\theta) / \theta \]

2. random effect parameter, variance in allele freq
   \[ \text{Var}(p) = \frac{Q}{1 - Q} \]

1.4.1 Theoretical, gene-rescue F-model

\[ P(p | \pi, \theta, \Delta) \]

\[ \theta = \frac{1}{4N_{m}} \]

1. Infinite island model \( \Rightarrow \) where \( \Delta \) is the allele freq on the migrant pool and \( \Theta = 4N_{m} \)

2. Disagreement from a common ancestor

\[ \pi \rightarrow \text{constant allele freq} \]

\[ \Theta = \frac{1}{4N_{m}} - 1 \]