# 1 Bayesian and Likelihood-based inference

In the lessons that I will be teaching, my goal is to give you an introduction to some of the modeling frameworks that come into play in population and speciation genomics, rather than teaching you specific software. This will be an introduction and a point of departure for more learning.

I will be writing my notes on screen, so that you can take notes as we go. Please interrupt me with your questions.

## 1.1 Why do we need probability theory for genomics?

We want to estimate parameters in genomics and probability gives us a basis for estimation and inference.

### 1.1.1 Example: genotype likelihoods/probabilities

1. Fundamental issue with sequence reads: the true genotype at a locus sequenced from a sample individual is an unknown parameter and we want to estimate a probability associated with the possible genotypes at a locus. e.g.: AA (2): 0.99 AT (1): 0.01 TT (0): 0.00

   genotype probabilities have become a common question with the random reads generated from DNA sequencers and the necessity of modeling sequencing error. Common software for variant calling results in genotype likelihoods.

2. Two ways of using genotype probabilities:

   (a) we can retain and utilize information for genotypes only (code them as 0,1,2) when probabilities exceed a threshold

   (b) we can build models that work directly with genotype probabilities. Some software does this, as will be discussed some in this school. This will make use of all of the data, rather than throwing some of it away.

### 1.1.2 Population genomic parameters more generally

We are typically interested in genotypes because they tell us something about individuals, populations and species. They support biological inferences because we can use them to estimate population genomic parameters.

Ideally we'll use an inferential and modeling framework that recognizes that genotypes come from individuals, within populations, within space and species. Likewise, diploid genotypes also come from within a genome, with chromosomes (with linkage, assortment, recombination, gene conversion, etc.).

We should strive for a modeling framework that utilizes mutual information between these levels of organization and sampling: this properly accounts for our uncertainty across the framework, and also captures parameters of interest in the hierarchy.

Examples of population genetic parameters:

1. mutation rate ($4N\mu$) – population – individuals (and loci)
2. estimates of allele frequencies
3. $F_{ST}$
4. admixture coefficients (structure's $q$)
5. $F_{IS}$

Ask students: first parameter here is for a population, what about the other four? Where does the mutual information arise to estimate these parameters (individuals, loci, or both)?

### 1.1.3 Parameter estimation

parameter estimation is central but often disregarded in biology. Not so in population genetics; there we often have theory for a parameter but not much for statistics.

Oddly, much of what many of us know about probability and statistics is instead concerned with the question of how improbable our data are, given a null hypothesis that were not interested in. Consequently, many biologists give little thought to parameter estimation itself, but instead on ways to get p-values.

So we need some different statistical tools that address the questions we're interested in and allow us to make decisions about evidence.

1. framework for model comparison and choice – confronting models with data

   We are interested in:

   (a) comparing alternative parameter values and finding values that have the most support from observed data.

   (b) contrasting models with different numbers and types of parameters to support inferences about biology.

   For example, we might want to compare $F_{ST}$ estimates or models of genetic architectures for a trait (numbers of causal loci and effect sizes).

These points, and other considerations, suggest that we will prefer Bayesian or likelihood methods for parameter estimation, rather than typical frequentist methods. So I will outline some of the essentials of Bayesian estimation, with asides about likelihood.

## 1.2 Essentials of Bayesian estimation

1. Key elements of Bayesian inference

   (a) Bayesian methods result in a probability density for the model parameters, given the data (the posterior probability).

(b) The probability density fully describes all information about the parameters and our uncertainty about them.

(c) Measure of uncertainty for each individual parameter incorporates uncertainty for all parameters.

(d) Computers make it possible to obtain a very good description of the probability density of the parameters, given the data

(e) Densities can readily be obtained for estimates of transformed parameters

(f) Probabilities provide a robust framework for scientific inference – model choice, strength of evidence, etc.

2. Bayesian and likelihood: Bayesian inference is one of two principal approaches to inference and parameter estimation. The other being Likelihood analysis. We'll prefer Bayes for the attributes in the previous list. Most of these things are not true for Likelihood.

   Given this very basic motivation, we need to make sure we have a foundation in probability.

3. Bayes theorem: equation for the posterior probability

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

   Label the different components – prior, likelihood and P(data), and discuss what each represents.

   (a) analytical solutions – available in simple cases. "Closed form" solution for the posterior probability density in the form of an integral. Quantiles (0.025, median, 0.975) and expectation (mean) can be calculated directly, as for any probability distribution.

   (b) simulation (MCMC) solutions – stochastic sampling from the posterior – when analytical solutions are not available, there are algorithms (MCMC) to draw samples from posterior and fully characterize it. This is typical in population genomics.

   (c) $P(\text{data})$: it is a normalizing constant for the sum $P(\text{data})$ across all possible values of $\theta$. Makes it so that the posterior density sums to one and is a proper probability distribution. We typically will not be able to calculate $P(\text{data})$. MCMC methods allow us to draw samples from the posterior to converge on a proper pdf that sums to one.

   $P(\theta|\text{data}) \propto P(\text{data}|\theta) \cdot P(\theta)$

   (d) Obtain quantiles (2.5% to 97.5% is the 95% ETPI or credible interval for true parameter) and expectation (mean) from pdf.

   (e) ABC – Approximate Bayesian Computation – solution for when equation of $P(Data|\theta)$ is not known, but it can be simulated, for example with the coalescent (but not just with the coalescent).

4. An example and application:

   Suppose we have genotypic data for 100 individuals from a population and we want to estimate the frequency of the two SNP alleles in the population ($p$ for 'A' allele). Our sample data contain: 63 AA, 34 AT, 3 TT.

Non-Bayesian point estimate of $p$: $x = 63 \times 2 + 34 = 160$, and $n = 200$, $160/200 = 0.8$ (this is the ML estimate)

However, if we want to obtain a posterior probability distribution for $p$, what do we need to do?

$$P(p|\text{alleledata}) \propto P(\text{alleledata}|p)P(p)$$

(a) $P(\text{alleledata}|p)$ – binomial – Process suggests we should model the allele data as binomially distributed (i.e., binomial probability function; a set of Bernoulli trials; discrete samples from a discrete process) – this is the likelihood.

(b) $P(p)$ – We need to place a prior probability distribution on $p$.

i. $p = [0, 1]$ – p can only take on values between 0 and 1.

ii. Are certain values of $p$ more likely *a priori*? Perhaps, but lets assume all equally likely.

iii. beta – The desirable prior for a binomial is a beta distribution, because of their mathematical relationship (beta is conjugate prior to binomial).
So, lets chose a beta prior $beta(\alpha = 1, \beta = 1)$, although we could chose a uniform prior and use simulations/MCMC instead.
Not all probability distributions have a conjugate prior, but many common simple parameter probability distributions do.

5. Closed form solution for posterior distribution

(a) Full model: $P(p|x, n) \propto P(x|p, n)P(p)$.

(b) Binomial likelihood: $P(x|p, n) = Cp^x(1 - p)^{n-x}$, where C is a constant that does not depend on $p$ (binomial coefficient).

(c) beta prior: $P(p) = Cp^{\alpha-1}(1 - p)^{\beta-1}$, where C is again a constant that does not depend on $p$.

(d) $P(x|p, n)P(p) = Cp^{x+\alpha-1}(1 - p)^{n-x+\beta-1}$

(e) This function is the probability density function for a beta distribution with parameters $x + \alpha$ and $n - x + \beta$ (our $\alpha$ and $\beta = 1$). So the posterior distribution is: beta(x+1, n-x+1)
Rcode:

```
p<-seq(0,1,0.001)
plot(p, dbeta(p, shape1=160+1, shape2=200-160+1),
     type ="l", xlab="p", ylab="density")
abline(v=qbeta(p=c(0.025, 0.975), shape1=160+1, shape2=200-160+1))
qbeta(p=c(0.025, 0.975), shape1=160+1, shape2=200-160+1)
## [1] 0.7390267 0.8494647
```

Bayesian point estimate is: $E(p) = \frac{\alpha}{\alpha+\beta} = \frac{161}{161+41} = 0.797$
Futher demo: Common question is how sample size affects confidence in an allele frequency and how many individuals should I sample from a population. Adjust model in Rcode to examine how confidence for 100 individuals differs from sample of 10 individuals.

```
qbeta(p=c(0.025, 0.975), shape1=16+1, shape2=20-16+1)
## [1] 0.5809340 0.9178241
```

Note the larger 95% confidence interval for the allele frequency parameter. Interpretation is that the true parameter lies in this interval with 95% confidence.

Instead we could also use MCMC simulation methods to draw samples from the posterior distribution and obtain equivalent estimates. This is a rare case in pop genomics where there is an analytical solution.

## 1.3 Probability distributions

We have seen a bit of application of probability to population genomics. There are a few probability distributions that you need to know as a foundation for modeling and I'll provide a brief overview of these here.

Draw these and discuss.

Make distinction between discrete and continuous distributions. Mention generating processes if possible.

1. binomial (Bernoulli trials) – count of events in reference category, discrete samples from discrete trials
2. multinomial (with one draw, for genotypes) – count of events in categories, discrete samples from discrete trials
3. beta – continuous distribution on [0,1]
   Illustrate betas with different parameters (we want to be able to refer to this below)

   ```
   p<-seq(0,1,0.01)
   par(mfrow=c(1,3))
   plot(p, dbeta(p, shape1=1, shape2=1),
   type="l", main="1,1", xlab="p", ylab="Density")
   plot(p, dbeta(p, shape1=100, shape2=100),
   type="l", main="100,100", xlab="p", ylab="Density")
   plot(p, dbeta(p, shape1=0.1, shape2=0.1),
   type="l", main="0.1,0.1", xlab="p", ylab="Density")
   ```

4. Dirichlet (multivariate generalization of the Beta)
5. uniform e.g., (0,10000]

### 1.3.1 AFS: Student discussion regarding beta

. Given that beta is a reasonable prior for allele frequency, it can be used to learn about the genome-wide distribution of for allele frequency at loci. What type of shape do you think characterizes the allele frequency spectrum and the beta for allele frequency prior?

Relate this to Nelson et al. 2012 paper (14002 individuals, 202 genes). Open it on screen, describe briefly, show top of Fig. 1.; DOI: `10.1126/science.1217876`

Also newer paper (lek16.pdf) that I am carrying with me.

# 2 Hierarchical models for allele frequencies

We want to work on some fundamental models associated with genotypic and next-generation sequence data that are useful and allow us to understand some of the modeling choices and possibilities.

We will develop three models and use these as a basis for understanding more complicated models.

1. Multilocus model for allele frequencies

   Suppose we have genotypic data for several loci and individuals. Let's generalize our single locus model for allele frequencies at each locus ($j$).

   $P(\vec{p}|x) \propto \prod_j P(x|p_j)P(p)$

   Likelihood: $P(x|p_j) \sim$ binomial$(p_j)$ – This is one binomial for all allele copies sampled from the population, as we did before 160 A out of 200 total copies sampled.

   Prior: $P(p) \sim$ beta$(1, 1)$ – constant and used for all loci

   What are we assuming in this model? We are assuming all loci are independent. But in reality loci share a genome and some history (for example history of drift). We will incorporate this in the next model.

2. Multilocus model for allele frequencies and diversity

   We could allow the prior on allele frequencies to be a parameter that we estimate from the data. We could make a hierarchical model that has a beta prior for allele frequencies and a hyperprior for the parameter of the beta.

   $P(\vec{p}, \theta|x) \propto \prod_j P(x|p_j)P(p_j|\theta)P(\theta)$

   Likelihood: $P(x|p_j) \sim$ binomial$(p_j)$ – as before, a binomial for all allele copies sampled from the population

   Conditional prior for $p_j$: $P(p_j|\theta) \sim$ beta$(\theta, \theta)$

   Hyperprior for $\theta$: $P(\theta) \sim$ Uniform$(0.001, c)$

   (a) The $\theta$ parameter describes diversity at loci. $P(p_i|\theta)$ is a version of the allele frequency spectrum. Recall beta with small value for $\theta$ would indicate that most loci have allele frequencies that are near one or zero. So $\theta$ is an interesting parameter itself.

   (b) $\theta \sim 4N\mu$ – if drift and mutation were the only the processes that affect diversity and they are constant, allele frequencies will equilibrate to a beta distribution with parameter $\theta$, which under these circumstances is an estimate of $4N\mu$. That's interesting. A parameter in a conditional prior for allele frequencies can be the population size-scaled mutation rate.

   (c) Transformation – Recall that we said that we can transform parameter estimates and the distribution of the transformed estimates will be a posterior distribution for the transformed value. In this case we are estimating allele frequencies with $P(p_j|\theta)$ and getting a posterior distribution for $p_j$. So we can calculate expected heterozygosity $H_e = 2p(1-p)$ (a transformation of $p$), as we estimate $p$, and get a posterior distribution for $H_e$.

3. Multilocus model for allele frequencies and diversity, with data that involve genotype uncertainty (individual data). The unobserved genotype is now an unknown. Could also incorporate sequence error, but we will not here.

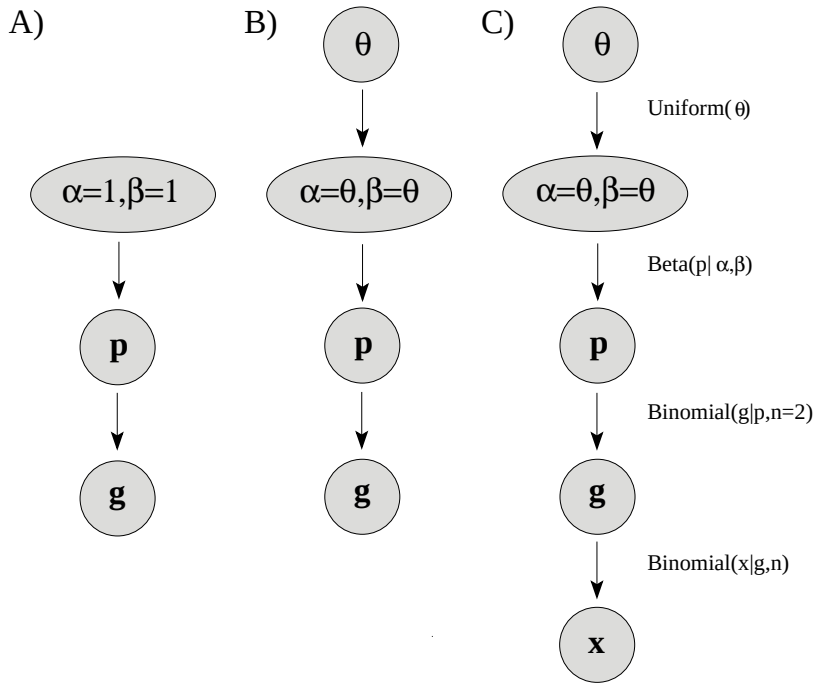
   $P(\vec{p}, \vec{g}, \theta|\vec{x}, \vec{n}) = \prod_i \prod_j P(x_{ij}|g_{ij}, n_{ij}) P(g_{ij}|p_j, n = 2) P(p_j|\theta) P(\theta)$

   Likelihood: $P(x|g_{ij}, n_{ij}) \sim \text{binomial}(g_{ij}/2, n_{ij})$ – where $g_{ij} = (0, 1, 2)$

   Conditional prior for genotype: $P(g_{ij}|p_j, n = 2) \sim \text{binomial}(p_j, n = 2)$

   The rest is the same as previous model.

   Alternatively, rather than working we read data directly ($x$), one could multiply by each of genotype likelihood obtained from other models (e.g., bcftools).



# 3 F-models for population differentiation

The amount of genetic variation within populations and differentiation among populations are determined by evolutionary processes:

- effective population sizes ($N_e$) and genetic drift
- mutation rate
- gene flow
- selection
- recombination and gene conversion

Estimates of genetic differentiation might shed light on these underlying processes

## 3.1 Quantifying differentiation

$F_{ST}$ commonly used to quantify differentiation.

1. $F_{ST}$ is generally a measure of the variance in allele frequencies among populations.
2. Ambiguity in usage of $F_{ST}$ – not all people mean the same thing by $F_{ST}$. One major distinction is whether $F_{ST}$ is a simple deterministic summary (fixed effects parameter) of allele frequencies or it is an evolutionary (random effects) parameter.
   (a) deterministic, fixed effects parameter – $F_{ST}$ can be a simple deterministic summary of allele frequencies (e.g., Nei's $G_{ST}$), where $F_{ST} = \frac{H_T - H_S}{H_T}$.
   All uncertainty in $F_{ST}$ is due to uncertainty in allele frequencies (finite sample). NB: no definition of multi-locus estimate of $F_{ST}$.
   (b) Random effects, evolutionary parameter – the same evolutionary parameter can give rise to different allele frequencies. Uncertainty in $F_{ST}$ is due finite sampling of population and limited sampling of the evolutionary process (evolutionary sampling).
   Weir and Cockerham's $F_{ST}$ and various "F-models". We will focus on $F_{ST}$ as an evolutionary parameter and on F-models.

## 3.2 F-models overview

1. beta distribution of allele frequencies – The F-model posits that the distribution of allele frequencies among populations is beta with parameters $\alpha = \pi\theta$ and $\beta = (1-\pi)\theta$, where $\theta = \frac{1}{F_{ST}} - 1$ and $\pi$ is the expected allele frequency.
2. The F-model arises (approximately) under two conditions:

(a) Infinite-island model – when many populations exchange migrants, the equilibrium allele frequency (equilibrium between drift and gene flow) is beta where $\pi$ is the migrant gene pool allele frequency and $\theta = 4Nm$, which at equilbrium has a direct relationship to $F_{ST}$.

(b) Divergence from a common ancestor – when populations diverge simultaneously from a common ancestor, the distribution of allele frequencies is approximately beta, where $\pi$ is the ancestral allele frequency and $\theta$ is inversely proportional to the effect of drift following divergence, and is a function of time and $N_e$.

Draw ancestor ($\pi$) and three descendant populations connected by $\theta = \frac{1}{F_{ST}} - 1$ amount of evolution.

(c) when these conditions are not met, model is still useful – distribution of allele frequencies can still be modelled as beta, where $\theta = \frac{1}{F_{ST}} - 1$ is a measure of genetic differentation (the variance in allele freq. among populations ) and $\pi$ is the expected allele frequency.

3. F-model has been used as a foundation for several Bayesian population genomic models and software. E.g., 1) Foll and Gaggiotti's F-model and `Bayescan` and 2) Pritchard et al. F-model (correlated allele frequencies) in `structure`

4. Use R to plot the distribution of allele frequencies across populations with: $\pi = 0.1, 0.5$ and $F_{ST} = 0.01, 0.4$

   R code:

```
p<-seq(0,1,0.01)
plot(p, dbeta(p, shape1=0.5 *  (-1 + 1/0.4), # Fst 0.4
     shape2=(1-0.5)* (-1 + 1/0.4)), col="red",  type="l", ylab="density",
     xlab="p", ylim=c(0,10))
lines(p, dbeta(p, shape1=0.5 *  (-1 + 1/0.01), # Fst 0.01
     shape2=(1-0.5)* (-1 + 1/0.01)), col="blue")
abline(v=0.5, col="red")
```

   Have students repeat this for an initial frequency of 0.1. Experiment with this.

Discuss the F-model results for changes in allele frequencies.

- Allele frequency variation from a intermediate frequency variant: low and high drift.
- Allele frequency variation from a very common or rare variant (two sides of same coin, typical case). Large magnitude allele frequency variation (drift) often will not be that evident, because it will happen to common/rare alleles and result in fixation/loss after small shift in allele frequency.

  Illustration of lack of one-to-one between pattern and process.

## 3.3 Locus-specific F-model

$F_{ST}$ can vary among loci and often we're interested in this variation and finding extraordinarily differentiated loci. In the above F-model, we assumed all loci had the same $F_{ST}$.

Alternatively, $F_{ST}$ could also vary among populations. Could also write nested models, but there is limited information to estimate parameters in these.

We will assume genotypes are known, but as we've seen, this assumption could be relaxed and we could easily model genotype uncertainty.

Assume pair (or larger group) of populations that share locus-specific $F_{ST}$.

1. Likelihood term: as before, the data are the count of allele at each locus ($i$) and population ($j$), $x_{ij}$, with $n$ allele copies sampled from each population ($n = 2\times$ the number of individuals).

   The probability of the data is a function of allele frequencies in each population and locus. Thus, the likelihood is a product of binomial distribution (their joint probability):

   $P(x_{ij}|p,n) \sim \prod_i \prod_j \text{binom}(x_{ij}|p_{ij}, n_{ij})$

2. Conditional prior for allele frequencies: the allele frequencies for each locus follow a beta distribution with parameters $\alpha = \pi_i\theta_i$ and $\beta = (1-\pi_i)\theta_i$. Recall, $\theta = \frac{1}{F_{ST}} - 1$.

   Taking the product across loci:

   $P(p|\pi,\theta) = \prod_i \text{beta}(\pi_i\theta_i, (1-\pi_i)\theta_i)$

3. Priors on $\pi$ and $F_{ST}$:

   (a) $\pi$ – For simplicity, we'll place independent priors on each $\pi_i$ as beta(1,1). Clearly we could incorporate another layer in the hierarchy and share information among loci as we did in earlier allele frequency models.

   (b) $F_{ST}$ – Various possibilities. A natural choice would be beta, because it is constrained to the scale of $F_{ST}$ and can assume many shapes (does not need to be symmetrical, and can be uni- or bimodal).

   We will assume $F_{ST} \sim \text{beta}(\psi S, (1-\psi)S)$, where $\psi$ is the expected value of $F_{ST}$ and $S$ is the precision (1/var) in $F_{ST}$.

4. Uniform, uninformative priors on $\psi$ and $S$.

5. Full model for locus-specific $F_{ST}$:

$$P(p,\pi,F_{ST},\psi,S|x,n) \propto \prod_i \prod_j [P(x_{ij}|p_{ij},n)P(p_{ij}|\pi_i,F_{ST_i})] \prod_i [P(\pi_i)P(F_{ST_i}|\psi,S)]P(\psi)P(S)$$

   binomial beta beta(1,1) beta , beta(1,1) , uniform(0.001,1000)

6. Discuss definition of outliers in this context: $P(F_{ST_i}|\psi,S)$ is a beta distribution and we can estimate the quantiles for $F_{ST_i}$ within the genome-wide distribution. Draw picture of beta and individual deviates.

7. Questions:

   (a) How realistic is it to assume all populations share a locus-specific $F_{\mathrm{ST}_i}$?
   When will this assumption be most reasonable? – replicate experimental populations, pairs of populations

   (b) The alternative is that all (or sets of) loci share an $F_{ST}$, but that populations vary. How realistic and useful is that?

8. draw graph of the part of the model that would use sets of loci to learn about population-specific $F_{ST}$

# 4 Exercise: simple F-models in JAGS

Software for Bayesian parameter estimation in population genomics uses Markov Chain Monte Carlo methods. These are methods to obtain samples from the posterior distribution, particularly when there is no analytical solution for it (the typical situation).

## 4.1 Algorithms for MCMC

Generally there are two types of algorithms for new values in a chain (ultimately from the posterior distribution):

1. Gibbs – every sample will be from the posterior distribution. Used when a step is based on conjugate distributions.

2. Metropolis (one variant is Metropolis-Hastings)

   In Metropolis, there are independence chains and random-walk chains. We need to monitor mixing in updates that use Metropolis (the rate of acceptance of new values).

   The Metropolis-Hastings algorithm meets criteria that ensure we will eventually converge to and sample the posterior distribution. Unfortunately, this could take millions of years and there is no way to be completely certain of convergence to the posterior distribution. We use diagnostics to get a sense. We discard initial samples as a burnin and run the chain long enough to obtain a good number of independent samples from the posterior.

Discussion of covergence, mixing, and proposal distributions.

## 4.2 Illustration with software for MCMC

JAGS is software that implements methods to generate stochastic samples from Markov chains. It is easy to specify the models, and the JAGS software determines what algorithms to use for updating chain.

Example: Implement locus-specific F-model for an analysis with known genotypes, but it is trivial to add a new likelihood to the model hierarchy to incorporate genotype uncertainty.

```
## simulate allele frequencies at 100 loci in ancestral population
nloci <- 100
nind <- 25

## generate loci that are likely to be variable with beta(15,15).
## Invariant loci give JAGS trouble.  Alternatively, drop
## invariant loci in sim.g below.  Try resimulating data (and avoid
## invariant loci) if you get JAGS error: "Slicer stuck at value with infinite density"
sim.pi <- rbeta(nloci, 15, 15)

## simulate allele frequencies in three derived (Fst=0.01) populations
sim.p <- matrix(nrow=nloci, ncol=3)
for(k in 1:3){
  sim.p[,k] <- rbeta(nloci, sim.pi * (-1 + 1/0.01), (1-sim.pi) * (-1 + 1/0.01)) }

#plot(sim.pi, sim.p[,1]) ## compare ancestral and derived allele frequencies

sim.g <- array(0, dim=c(nloci,nind,3))
for(k in 1:3){
  sim.g[,,k] <- matrix(rbinom(nloci*nind, 2, prob=sim.p[,k]),
  nrow=nloci, ncol=nind)
}

## Use R to JAGS interface to estimate Fst from these data
library(rjags)

mod.jags <- jags.model("locusFmodel.jags", data=list(nind=nind,
    nloci=nloci, npop=3, g=sim.g), n.chains=2)
mod.sam <- jags.samples(model=mod.jags, variable.names=c("Fst", "psi"), n.iter=2000, thin=2)

plot(mod.sam$Fst[2,,1]) ## for locus 2
mean(mod.sam$Fst[2,,1])

plot(c(mod.sam$psi[1,,1], mod.sam$psi[1,,2])) ## plot both chains for genome-wide Fst (psi)
```

The JAGS model should be in a separate text object (or file).

```
model{
  for(i in 1:nloci){
    for(j in 1:nind){
      ## binomial likelihood for genotype = number allele copies of reference allele
      for(k in 1:npop){
        g[i,j,k] ~ dbinom(p[i,k], 2)
      }
    }
  }
  for(i in 1:nloci){
    for(k in 1:npop){
      ## population allele frequency in sample populations
      p[i,k] ~ dbeta(0.001+pi[i]*theta[i], 0.001+(1-pi[i])*theta[i])
    }
    theta[i] <- -1 + 1/Fst[i]
    Fst[i] ~  dbeta(0.001+psi*S, 0.001+(1-psi)*S)
    pi[i] ~  dbeta(1,1)
  }
  psi ~ dbeta(1, 1)
```

```
  S ~ dunif(0.01, 1000)
}
```

# 5 Models for ancestry and introgression in hybrid zones

The Pritchard et al. (2000) and Falush et al. (2003) papers have played a central role in analysis of population genetics, particularly when we're interested in mixed ancestry of individuals that results from distant crosses or admixture. These and related models are referred to as structure-like models.

They assume that there are demes from which individuals are sampled whole (the no admixture model) or with hybridization (the admixture model). They are closely related to the allele frequency models that we discussed and only require minor modifications relative to these.

1. The Pritchard et al 2000 paper has two models: the no-admixture model and the admixture model.

   (a) no-admixture model:interested in assigning individuals to populations of ancestry, without allowing for possibility of admixture. We would also estimate population allele frequencies along the way.
   $P(Z, P|X) \propto P(X|Z, P)P(Z)P(P)$ (Eq. 1 in Pritchard et al 2000).
   Example: assignment of fish caught at sea to their freshwater populations of origin.

   (b) admixture model: allows individuals to have mixed ancestry and now is a model for $P(Z, P, Q, \alpha|X)$. In this case, we're interested in the estimate of admixture proportion for each individual ($Q$), along with $P$ and $Z$.

2. Round 1 of questions for the students to answer:

   (a) In the case of the no-admixture model, how could each component on the right-hand side of the equation distributed?
   Answers:
      i. $P(X|Z, P) \sim$ multinomial$(p_{zl}, 1)$ or Bernoulli$(p_{zlj})$
      ii. $P(Z) =$ multinomial$(1/K, n = 1)$
      iii. $P(P) \sim$ Dirichlet$(1, ..., 1)$, which is equivalent to beta$(1, 1)$ for SNPs.

   (b) Below is the full equation for the posterior in the case of the admixture model. How could each component distributed?
   Answer: $P(Q, Z, P, \alpha|X) \propto P(X|Z, P)P(Z|Q)P(Q|\alpha)P(\alpha)P(P)$
   multinom multinom(q,n=1) Dirichlet($\alpha$) Unif[0,10] Dirichlet(1,,1)

   (c) How could F-models be brought into this model? (they have been in `structure`.

   Discuss these as a class before going on to the next set of questions (if time allows).

3. Round 2 of questions for the students to answer:

   (a) The Pritchard et al 2000 model assumes that all loci introgress equally: $P(Z|Q) \sim q$. Is this reasonable? How could we relax this assumption?

   Answers:

   i. For understanding overall make-up of individuals, this might be an ok simplification. Overall admixture should be estimated fine. But if you want to understand locus-specific behavior and introgression, it is probably not reasonable.

   ii. One could take $q$ as the expectation, but write a function that allows for locus-specific deviations from the genome-wide expectation. Gompert and Buerkle have implemented such a function in `bgc`, `finestructure` by Falush, and there are many different methods that allow for locus-specific ancestry.