1 A probability foundation for population genomics

In the lessons that I will be teaching, my goal is to give you an introduction to some of the probability and 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. I will also provide the pdf of my notes after the presentations.

Please interrupt me with your questions. I will continue with this material this afternoon in my next session, and we'll add in more hands-on work there.

1.1 Why do we need probability theory for genomics?

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

1. (Example: genotype probabilities from 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 (P(genotype|reads)). e.g.: AA (2): 0.99 AT (1): 0.01 TT (0): 0.00 (discrete probability distribution).

   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. (Example: population allele frequencies): we do not observe allele frequency directly, but learn about it from a sample of individuals. For statistics.

3. (Example: theoretical models of allele frequency in finite populations). Additionally, in some cases we have probability theory for the expectations of observations that derive from theory.

1.2 Estimation of allele frequencies

Let's consider the statistical usage first.

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 parameter: $x = 63 \times 2 + 34 = 160$, and $n = 200$, 160/200 = 0.8 (this is the maximum likelihood estimate)

Could the true allele frequency be 0.79 or 0.85? Yes! We have a finite sample from the true population. We can build a probability model for alternative values of p.

1. probability estimate of p – However, what if we want to obtain a correct probability distribution for the parameter p that reflects our uncertainty? What do we need to do? We will need to calculate a Bayesian posterior probability density. Let's do this, so that we can study the effect of sampling more and to lay the groundwork for stuff that comes later.

We'll use Bayesian estimation (no MCMC required) and Bayes' theorem.

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

(a) $P(\text{data}|p)$ – 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*?
Ask students to draw what they believe the distribution of p likely is.
Yes, but for today, 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.

(c) In this example we won't need to calculate $P(\text{data})$, but mathematically it's there. It normalizes expression to insure left side sums to 1.

2. 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 with parameters α and β: $P(p) = Cp^{\alpha-1}(1-p)^{\beta-1}$, 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 α and $\beta = 1$). So the posterior distribution is: beta(x+1, n-x+1)

Rcode (line 16 in buerkle_code.R):

```
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
```

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

Bayesian point estimate (mean) is: $E(p) = \frac{\alpha}{\alpha+\beta} = \frac{161}{161+41} = 0.797$. Had we used a different prior, we would likely have given less prior weight to $p = 0.5$.

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

3. This is a rare case in pop genomics where there is an analytical solution to the Bayesian model. Otherwise we use MCMC to obtain samples from posterior distribution.

4. (Reflection and encouragement about allele frequency estimation modeling we just did) – What was the goal of the math modeling exercise? — to develop a probabilty model for our estimate of p, given our observations. The problem is that we are uncertain about p, because we sampled from the population and we do not know the truth.

   Suppose we had a population in a hatchery and one in a lake where we introduced fish. We could observe the same alleles in each. We are sure to get different $\frac{x}{n}$ observations in the pair. Does that mean the allele frequencies are different and the populations have evolved? How would we know the true p_1 and p_2, and therefore the true $p_1 - p_2$? — we need a model for the truth, given our observations

   Questions? — break?

1.3 Theoretical and statistical models for allele frequencies

Given the model that I have introduced so far, I would like to develop this further, to build hierarchical models in which information can be shared among loci, or frequencies of alleles can be correlated among populations.

1. single locus model – beta-binomial – as above

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

   Loci exist in a genome, so they share their evolutionary history to some extent. We should be able to use information from multiple loci to learn about population history (we will do that in the second multilocus model below). I am leaving denominator (P(data) out for convenience, which is why I am writing $\propto$ rather than $=$.

2. 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 for 160 A out of 200 total copies sampled.

   $\sim$ means "is distributed as".

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

   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.

   This has the same analytical solution as the single-locus model, because it just a collection of single-locus models.

3. Multilocus model for allele frequencies and diversity

   This is a model we'll use in our hands-on session this afternoon.

   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: $P(p_j|\theta) \sim \text{beta}(\theta, \theta)$

   Hyperprior for θ: $P(\theta) \sim \text{Uniform}(0.001, c)$

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

   Draw different between symmetrical beta distributions (with $\alpha = \beta$, and $\alpha < 0, = 0, > 0$).

   (b) In theory, $\theta \sim 4N\mu$ – 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 θ, 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.

   Draw corresponding beta distribution with large and small theta and HWE quadratics.

   (c) Transformation – in Bayesian analysis 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.

   (d) The beta distribution of allele frequencies – what shape (θ) do you think holds for real populations? This is a parametric version of the site frequency distribution.

   show Nelson et 2012 ("An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People" – 10.1126/science.1217876).

   In terms of the parameters of the beta distribution, real populations have something like $\alpha = \beta << 1$.

1.4 The F-model of population differentiation

Given those fundamentals, let's consider another hierarchical model. The model can be used as a theoretical, generative model for allele frequencies, or we can use it as a statistical model to learn about allele frequencies in populations.

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

- effective population sizes (N_e) and genetic drift
- mutation rate

- gene flow
- selection
- recombination and gene conversion

Estimates of genetic differentiation can potentially inform us a bit on these underlying processes. F_{ST} is commonly used to quantify differentiation and is a measure of the variance in allele frequencies among populations.

There are multiple definitions (implementations) and not all people mean precisely the same thing by F_{ST}.

One major distinction is whether F_{ST} is a simple summary statistic of allele frequencies (also, a fixed effects parameter) or it is an evolutionary (random effects) parameter.

1. deterministic, fixed effects parameter – F_{ST} can be a simple deterministic summary of allele frequencies (e.g., Nei's G_{ST}). For example, $G_{ST} = \frac{H_T - H_S}{H_T}$. Here, all uncertainty in F_{ST} is due to uncertainty in allele frequencies (finite sample). NB: this approach does not define a multi-locus estimate of F_{ST}.
2. Random effects, evolutionary parameter – the same evolutionary parameter can give rise to different allele frequencies (a random draw from the process). Uncertainty in F_{ST} is due to finite sampling of population and limited sampling of the evolutionary process (evolutionary sampling).

   Weir and Cockerham's F_{ST} estimator (sometimes written θ_{ST}) and various "F-models" implement this random effects estimate.

We will mostly consider F_{ST} as an evolutionary parameter and on F-models.

1.4.1 Theoretical, generative F-model

The F-model we will consider is an elaboration of our beta distribution of allele frequencies at loci and introduces a parameter for the variance (or correlation) of allele frequencies at a locus among populations.

The F-model posits that the distribution of allele frequencies at a locus among populations is beta(α, β), where the parameters $\alpha = \pi\theta$ and $\beta = (1-\pi)\theta$, $\theta = \frac{1}{F_{ST}} - 1$ and π is the expected (mean) allele frequency.

Recall that the mean of any beta distribution is given by $\hat{x} = \frac{\alpha}{\alpha+\beta}$. After substituting $\alpha = \pi\theta$ and $\beta = (1-\pi)\theta$, through rearrangement we obtain $\hat{x} = \pi$.

This reparameterization of the beta to use θ and π is used in other settings also and is useful because now one parameter corresponds to the mean (π) and the second parameter (θ) is a multiplier that corresponds to the precision (inverse of the variance).

The F-model arises (approximately) under two conditions (formal population genetic models):

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

2. Divergence from a common ancestor – when populations diverge simultaneously from a common ancestor, the distribution of allele frequencies in the descendant populations is approximately beta, where π is the ancestral allele frequency and θ is inversely proportional to the effect of drift following divergence, and is a function of time and N_e.

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

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

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

R code (line 26 in buerkle_code.R):

```
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="purple")
abline(v=0.5, col="blue") # ancestral allele frequency
```

Repeat this for an initial frequency of 0.1. Experiment with different parameters.

1. What effect does a larger F_{ST} have on resulting allele frequencies?
2. What effect does the starting frequency have on the distribution of allele frequencies and the variance introduced by F_{ST}?
3. Which results best illustrate the lack of one-to-one correspondence between process (here F_{ST}) and empirical pattern?

   - 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.

So we have a model for how drift can influence allele frequencies. Let's put this to use! Let's use our generative model to create reasonable population genetic data and investigate the relationship between global allele frequency and F_{ST}.

We will use Hudson's F_{ST} as a summary statistic on the variance in allele frequencies, but we could also use other definitions of F_{ST}.

Activity: using the code that starts on line 38, let's learn about the relationship between global allele frequency and F_{ST}

Have students do this, walk through the code as a group and answer student questions.

What are the implications of this? – for a single value of simulated process Fst that generates variance in allele frequencies, we get a wide range of observed Fst in the data. Fst is enormously constrained by global (ancestral) allele frequency.

How will this play out across the genome?

1.4.2 Statistical F-model

In addition to its use as a generative, theoretical model, we can use the F-model for statistics.

Different versions of the F-model have been used as a foundation for several important Bayesian population genomic models and software: e.g., 1) Foll and Gaggiotti's F-model and `Bayescan` software (uses a locus specific model) and 2) Pritchard et al. F-model (correlated allele frequencies) in `structure` software (simple F-model with one F_{ST} for all loci).

Specification of the F-model: Let us consider the specification of the F-model stepwise, rather than facing the full equation all at once.

1. Likelihood term: as before, the data x_{ij} are the count of the reference allele at each locus (i), but are now also indexed by the population that is being considered (j), with n allele copies sampled from each population ($n = 2\times$ the number of diploid 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 distributions (their joint probability):

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

   Remember: $\sim$ means "is distributed as".

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, $\theta = \frac{1}{F_{ST}} - 1$ (i.e., θ is a transformation of F_{ST}). Taking the product across loci:

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

3. Priors on π and F_{ST}:

   (a) π – For simplicity, we'll place independent priors on each π_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). For now, let us define F_{ST} as beta(1,1) and assume that all loci in the genome share the same F_{ST}.

With these components, we can write an F-model:

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

1.5 Activity: statistical population 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).

1.5.1 Algorithms and software for MCMC

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

1. Gibbs – following sufficient burn-in to a stationary distribution, each and every sample will be from the posterior distribution. Used when the parameter being updated involves a product of two distributions that is known analytically (i.e., the are conjugate).

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 a long time and there is no way to be completely certain of convergence to the posterior distribution. We use diagnostics to get a sense of mixing and convergence. We discard initial samples as a burn-in and run multiple chains, each long enough to obtain a good number of independent samples from the posterior and to gauge convergence among the replicate chains.

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.

1.5.2 Implementation of MCMC for our multilocus model for allele frequencies and diversity

Hands-on — Please execute and experiment with the code in buerkle_code.R (lines 79–201).

The questions and tasks are embedded in the code comments and they are also in the webpage for this activity.

1. Examine the plots for estimates of θ at each of the 500 retained steps from each of your three replicate chains.

   (a) Does it appear that you have proper mixing within chains?

   (b) What would an example of poor mixing look like?

   (c) Does it appear that you good convergence among the three chains?

   (d) What would an example of poor convergence look like?

2. (a) Does the 95% credible interval for θ include the true value that you simulated (`sim.theta` on line 83 of the code)?

   (b) What does this mean?

3. For a simulation that sets `sim.theta` to 5:

   (a) how well do the estimates agree if you simulated 20 individuals, compared to their agreement if you sampled 100 individuals?

   You'll need to simulate new data for each (`nind`←20 and `nind`←100) and run the MCMC on each simulated data set.

(b) What do you think is the cause of the difference?

4. Compare a simulation with `sim.theta`←5 to a simulation with `sim.theta`←0.1.

   (a) How does this affect the genome-wide mean H_e statistic on the data?

   (b) Does this fit with your expectations about how θ and H_e relate to diversity?

1.5.3 Implementation of MCMC for F-model

Exercise: continue on lines 203–271 in the code and answer the corresponding questions for this section on the webpage (statistical F-model).

1. How realistic is it to assume all populations and loci share a single F_{ST} parameter?

2. How could we relax this assumption?

3. Please explain the plot on line 22 for a simulation with `simFst` set to 0.01 and again for a simulation with `simFst` set to 0.3. What are the noteworthy aspects?

4. (a) (line 65) Does the 95% credible interval for F_{ST} include the true value that you simulated (`simFst`=0.01 on line 8 of the program)?

   (b) What does this mean?

5. For a simulation that sets `simFst` to 0.3:

   (a) Does the 95% credible interval for F_{ST} include the true value of $F_{\mathrm{ST}} = 0.3$?

   (b) What do you think is a likely cause of the discrepancy?

6. Compare a plot of a PCA of populations simulated with `simFst`←0.01 to populations simulated with `simFst`←0.3.

   (a) In which of these simulations are individuals more readily assigned to clusters in the PC2 vs. PC1 plot?

   (b) Does the percentage of variation explained by PC1 scale with `simFst`?

7. Compare a plot of a PCA of populations simulated with `simFst`←0.01 to a PCA of the very similar simulation to which a small number of loci were added that were more differentiated (line 107). What is the effect on clustering?