Overview of the Structure Approach

Data:

X: genotypes of sampled individuals (known)
Z: population origins of individuals (unknown)
P: population allele frequencies (unknown)

Assumptions:

Hardy-Weinberg Equilibrium (HWE) *within* populations
Linkage Equilibrium (LE) between loci *within* populations

Model:

Each allele [at each locus, in each genotype] is a random draw from a probability distribution, Pr(*X*|*Z,P*)

Population structure is imposed on the data to account for any Disequilibrium (HWD and/or LD)

Goal:

Perform statistical inference on the parameters *Z* & *P*

Method:

A Bayesian approach using MCMC to approximate samples from Pr(*Z,P* | *X*)

X, Z, & P

$(x_l^{(i,1)}, x_l^{(i,2)})$ = Genotype of i^{th} individual At l^{th} locus

$z^{(i)}$ = Pop. from which individual *i* originated

p_{klj} = freq. of allele *j* at locus *l* in population *k*

$i = 1, 2, ..., N$ $\quad l = 1, 2, ..., L$
$j = 1, 2, ..., J_l$ $\quad k = 1, 2, ..., K$

We want Pr(*Z,P* | *X*) and can calculate Pr(*X* | *Z,P*) directly

Bayes rule gives

$$\Pr(Z, P \mid X) \propto \Pr(Z)\Pr(P)\Pr(X \mid Z, P)$$

where Pr(*Z*) and Pr(P) are called *priors*

Pr(*Z,P* | *X*) can't be computed exactly,
but MCMC can give an approximate sample from the distribution

$$(Z^{(1)}, P^{(1)}), (Z^{(2)}, P^{(2)}), ..., (Z^{(M)}, P^{(M)})$$

Summary statistics based on this sample (e.g., *posterior means*) give estimates of the parameters.

e.g., $$E(p_{klj} \mid X) \approx \frac{1}{M}\sum_{m=1}^{M} p_{klj}^{(m)}$$

Step 0. Specify initial starting values $Z^{(0)}$ for population origins (e.g., random assignment according to a uniform probability distribution)

Step 1a. Sample $P^{(m)}$ from Pr(*P* | *X*, $Z^{(\text{m-1})}$)

Step 2a. Sample $Z^{(m)}$ from Pr(*Z* | *X*, $P^{(\text{m})}$)

Step1b: estimate allele freqs. for each population assuming population origins are known

Step 2b: estimate population origins for each individual assuming population allele freqs. are known

$$(Z^{(m)}, P^{(m)}), (Z^{(m+c)}, P^{(m+c)}), (Z^{(m+2c)}, P^{(m+2c)}), ...$$

are approximately independent draws from Pr(*Z,P* | *X*) for sufficiently large *m* and *c*

Mutation Models for Microsatellites

X = size of an Msat allele in repeat units

Single-step Stepwise Mutation Model (SSMM)

X → (X + 1) with probability *0.5*
(X – 1) with probability *0.5*

Multi-step Stepwise Mutation Model (MSMM)

X → X +/- 1 with probability *p*
X +/- U with probability *1-p*

Where U ~ Geometric(λ) [1/ λ = 1.5 is commonly used]

ANOVA $y_{gki} = p + a_g + b_{gk} + w_{gki}$

Source	**SS**	**DF**	**MS**
Among Groups	SS(AG)	G-1	SS(AG)/(G-1)
Among Pops Within Group	SS(AP)	K-G	SS(AP)/(K-G)
Among individuals Within Population	SS(WP)	N-K	SS(WP)/(N-K)
Total	SS(Tot)	N-1	

$K = \sum_{g=1}^{G} k_g \qquad N = \sum_{g=1}^{G} \sum_{k=1}^{k_g} n_{gk}$

$$SS(Tot) = \sum_{g=1}^{G} \sum_{k=1}^{k_g} \sum_{i=1}^{n_{gk}} (y_{gki} - \overline{y}_{...})^2$$

$$SS(WP) = \sum_{g=1}^{G} \sum_{k=1}^{k_g} \sum_{i=1}^{n_{gk}} (y_{gki} - \overline{y}_{gk.})^2$$

AMOVA uses the relationship below to compute the SS

$$SS(z) = \sum_{i=1}^{N} (z_i - \overline{z})^2 = \frac{1}{2N} \sum_{i=1}^{N} \sum_{j=1}^{N} (z_i - z_j)^2$$

Which allows you to specify a distance measure between different alleles.

➔

$$SS(WP) = \sum_{g=1}^{G} \sum_{k=1}^{k_g} \frac{1}{2n_{gk}} \sum_{i=1}^{n_{gk}} \sum_{j=1}^{n_{gk}} \delta^2_{(gk)ij}$$

$\delta^2_{(gk)ij} = (y_{gki} - y_{gkj})^2$ is meaningful for Msats

Percentages of variance explained are ratios of variance components, e.g.,

$$\sigma_a^2 / \sigma_{Tot}^2 , \qquad \sigma_b^2 / \sigma_{Tot}^2 , \qquad \sigma_w^2 / \sigma_{Tot}^2$$

ANOVA can be used separately for each allele (variant 1 vs. not variant 1), and then combined over alleles.

Let x_{kji} be the j^{th} allele in the i^{th} individual in the k^{th} population

and y_{kji} be an 0/1 indicator variable for the variant "A"

Then, E(y_{kji}) = *p* and Var(y_{kji}) = *p(1-p)*

AMOVA extends to multiple loci (e.g., *m* different loci)

$$SS(WP) = \sum_{g=1}^{G} \sum_{k=1}^{k_g} \frac{1}{2n_{gk}} \sum_{i=1}^{n_{gk}} \sum_{j=1}^{n_{gk}} (y_{gki} - y_{gkj})' W (y_{gki} - y_{gkj})$$

where ***y*** is now an m-vector and ***W*** is a square matrix of weights that defines the relationships between loci and/or their weights.