
Inferring population structure

See Pritchard, Stephens & Donnelly, Genetics, 2000

Significance

- Of general scientific/anthropological interest
- Major potential confounder in GWAS studies
- Commercial interest

Modules used in the STRUCTURE software package

This lecture is based primarily on the following paper:

Inference of Population Structure Using Multilocus Genotype Data
Pritchard, Stephens and Donnelly
Genetics 2000

Model I

(Without admixture)

- K populations
- Distinct populations characterized by different allele frequencies

X : The genotypes of the samples individuals. $(x_m^{(i,1)}, x_m^{(i,2)})$ is the genotype of the i^{th} individual at the m^{th} locus

Z : The populations of origin of the sampled individuals. $z^{(i)}$, the population of origin of individual i

F : The population-specific allele frequencies. f_{kmj} , the frequency of the j^{th} allele of the m^{th} locus in population k

Objective

Simultaneously

- estimate allele frequencies in the K populations and
- assign individuals to populations

We will do this by sampling from $P(Z, F|X)$ using a Gibbs sampler.

Gibbs sampler

Used Gibbs sampling to sample from $P(Z, F|X)$ as follows:

- sample F^{m+1} from $P(F|X, Z^{(m)})$
- sample Z^{m+1} from $P(Z|X, F^{(m+1)})$

Both of these sampling steps require Bayes' Rule.

To sample F^{m+1} from $P(F|X, Z^{(m)})$ we use $P(F|X, Z) = \frac{P(X|Z, F)P(F)}{P(X)}$.

Note that $P(x_m^{(i,a)} = j|Z, F) = f_{Z^{(i)}j}$ for $a = 1, 2$

The assumption of H-W equilibrium and linkage equilibrium then allows us to calculate $P(X|Z, F)$.

A Beta prior is used for $P(F)$ (for theoretical and convenience reasons)

Gibbs sampler

Used Gibbs sampling to sample from $P(Z, F|X)$ as follows:

- sample F^{m+1} from $P(F|X, Z^{(m)})$
- sample Z^{m+1} from $P(Z|X, F^{(m+1)})$

To sample Z^{m+1} from $P(Z|X, F^{(m+1)})$ we use

$$P(z_i = k|x_i, F) = \frac{P(x_i|z_i = k, F)P(z_i = k)}{P(x_i|F)}$$

The prior probability $P(z_i = k)$ is set to $1/k$

(i.e. before observing any data all populations of origin equally likely - could cause problems for very imbalanced sample)

Gibbs sampler

Used Gibbs sampling to sample from $P(Z, F|X)$ as follows:

- sample F^{m+1} from $P(F|X, Z^{(m)})$
- sample Z^{m+1} from $P(Z|X, F^{(m+1)})$

For sufficiently large m and c

$$(Z^{(m)}, F^{(m)}), (Z^{(m+c)}, F^{(m+c)}), (Z^{(m+2c)}, F^{(m+2c)}) \dots$$

will be approximately independent samples from $P(Z, F|X)$

Model II

(With admixture)

Admixed individuals have non-zero proportions of their ancestry from multiple populations. Define $Q : q_k^{(i)} =$ proportion of ancestry of individual i from population k .

Redefine Z so that e.g. $z_m^{(i,a)}$ is the population of origin of allele copy a at locus m in individual i

We now have

$$P(x_m^{(i,a)} = j | Z, F, Q) = f_{z_m^{(i,a)} mj}$$

and

$$P(z_l^{(i,a)} = k | F, Q) = q_k^{(i)}$$

Sampling from $P(F, Q, Z|X)$

- Sample $F^{(m+1)}, Q^{(m+1)}$ from $P(F, Q|X, Z^{(m)})$
- Sample $Z^{(m+1)}$ from $P(Z|X, F^{(m)}, Q^{(m+1)})$

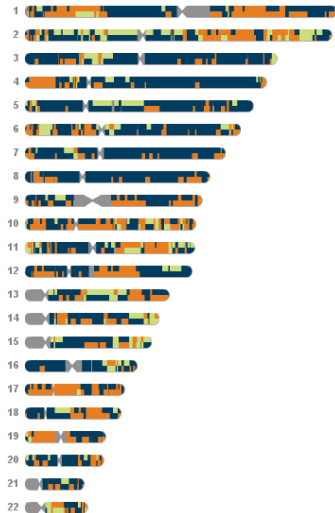
Estimating K

Can be estimated from

$$P(K|X) \propto P(X|K)P(K)$$

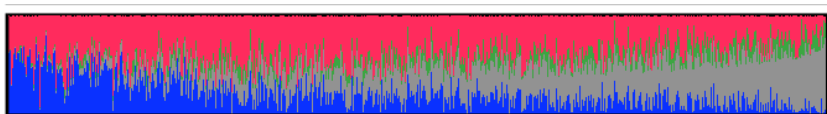
Given prior assumptions for K. In practice $P(X|K)$ is difficult to calculate, computationally and is approximated.

Examples



From 23andme

Results of Structure, applied to an admixed population ('Cape Coloured' population)



Simulations (from Pritchard *et al.*)

- Model 1: Single randomly mating population
- Model 2: Two randomly mating populations of size $2N$, separated N generations ago, with no admixture (2 separate datasets - with 5 and 15 loci (denoted 2A and 2B, respectively))
- Model 3: Admixture - the two populations from simulation 2 are fused to produce a single randomly mating population. Samples collected 2 generations after admixture.

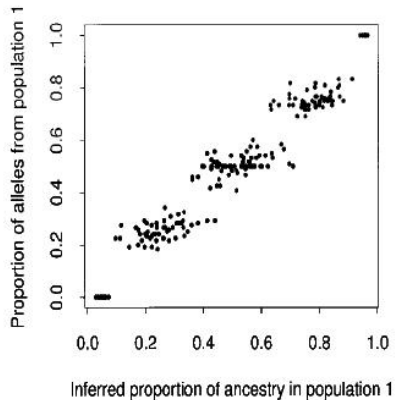
Results of simulations

TABLE 1

Estimated posterior probabilities of K , for simulated data sets 1, 2A, 2B, and 3 (denoted X_1 , X_{2A} , X_{2B} , and X_3 , respectively)

K	$\log P(K X_1)$	$P(K X_{2A})$	$P(K X_{2B})$	$P(K X_3)$
1	~ 1.0	~ 0.0	~ 0.0	~ 0.0
2	~ 0.0	0.21	0.999	~ 1.0
3	~ 0.0	0.58	0.0009	~ 0.0
4	~ 0.0	0.21	~ 0.0	~ 0.0
5	~ 0.0	~ 0.0	~ 0.0	~ 0.0

Results of simulations (Dataset 3)



Application to data

Data from the endangered Taita thrush (*Turdus helleri*) from four locations in Kenya.

TABLE 3
Inferring the value of K , the number of populations,
for the *T. helleri* data

K	$\log P(X K)$	$P(K X)$
1	-3144	~ 0
2	-2769	~ 0
3	-2678	0.993
4	-2683	0.007
5	-2688	0.00005

The values in the last column assume a uniform prior for K ($K \in \{1, \dots, 5\}$).

Application to data

