
Inferring population structure

See Pritchard, Stephens & Donnelly, Genetics, 2000

Significance

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

From 23andme

Chromosome View

-  Solid segments indicate that both chromosomes come from the same geographic region. [See a Cambodian Woman's painting.](#)
-  Dual-colored segments indicate chromosomes from different geographic regions. [See an African American Man's painting.](#)

Select a person: African American Woman ▾



African American Woman ?

Most African Americans today trace a large part of their ancestry to sub-Saharan Africa as a result of the slave trade. Over the generations since, both Europeans and Native Americans have intermarried with African Americans and contributed ancestry, as seen in the ancestry painting of this woman, who identified herself as African American.

	Africa 65%
	Europe 29%
	Asia 7%
	Not Genotyped

Worldwide Examples

Click on the icons in the map below to see example paintings of individuals from across the globe.



Model I

(Without admixture)

- K populations
- Distinct populations characterized by different allele frequencies

X : The genotypes of the samples individuals. $(x_i^{(i,1)}, x_i^{(i,2)})$ is the genotype of the i^{th} individual at the j^{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_{klj} , the frequency of the j^{th} allele of the j^{th} locus in population k

Objective

Simultaneously

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

Joint distribution of Z and F , given the data

$$P(Z, F|X) \propto P(X|Z, F)P(Z), P(F)$$

Assuming *a priori* independence of priors for Z and F

Probability of a genotype, given the population of origin and allele frequencies

Assume H-W equilibrium within populations

$$P(x_i^{(i,a)} = j | Z, P) = f_{z(i)j} \text{ for } a = 1, 2$$

Priors on Z and F

Before observing any data all populations of origin equally likely

$$P(z^{(i)} = k) = \frac{1}{K}$$

Use a Beta distribution for the prior over allele frequencies. More generally, for multiallelic loci, a Dirichlet distribution

$$f_{kl} \sim \mathcal{D}(\lambda_1, \dots, \lambda_J)$$

In application Pritchard *et al.* set $\lambda_j = 1$ for all j (i.e. a uniform distribution over allele frequencies).

Gibbs sampler

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

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

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 $\mathbf{Q}$: $q_k^{(i)}$ = proportion of ancestry of individual i from population k .

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

We now have

$$P(x_l^{(i,a)} = j | Z, F, Q) = f_{z_l^{(i,a)} j}$$

and

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

Use a Dirichlet distribution for the admixture proportions of each individual so that

$$q^{(i)} = \mathcal{D}(\alpha, \alpha, \dots, \alpha)$$

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

- Sample $F^{(m)}, Q^{(m)}$ from $P(F, Q|X, Z^{(m-1)})$
- Sample $Z^{(m)}$ from $P(Z|X, P^{(m)}, Q^{(m)})$
- Metropolis-Hastings step (to integrate α out)

Modes of the parameter space

The populations are not numbered, therefore equivalent solutions can be obtained with different population labelling (e.g. in two-population case with 20 individuals, equivalence of the solution involving individuals 1-10 in population 1 and 11-20 in population 2 with 1-10 in population 2 and 11-20 in population 1).

In practice the sampler does not cross between modes, so this does not create a problem.

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.

Simulations

- 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.

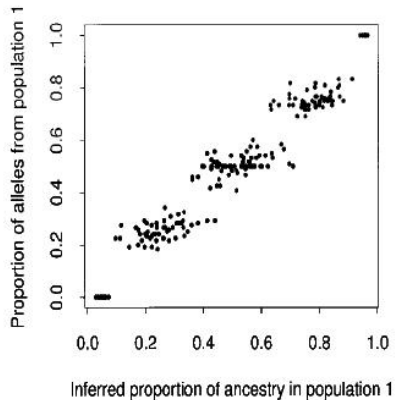
Results of simulations

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

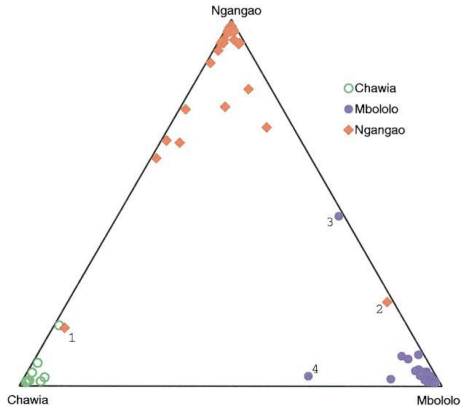
Data from the endangered Taita thrush (*Turdus helleri*) from four locations in Kenya.

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



Extension by Falush *et al.*, *Genetics* 2003

- The above analysis assumed independence of loci
- Useful with sparse data but with dense genotype data successive loci are non-independent
- Takes account of admixture linkage disequilibrium (non-independence of ancestry at nearby loci)

Model with admixture linkage disequilibrium

Ancestry of entire blocks of loci from the same source population.
Consider breaks between these blocks as a Poisson process, with rate r . Model ancestry proportions along a chromosome as a Markov chain:

$$P(z_1^{(i)} = k | r, Q) = q_k^{(i)}$$

and

$$P(z_{l+1}^{(i)} = k' | z_l^{(i)} = k, r, Q) = \begin{cases} e^{(-d_l r)} + (1 - e^{(-d_l r)})q_{k'}^{(i)} & \text{if } k' = k \\ (1 - e^{(-d_l r)})q_{k'}^{(i)} & \text{otherwise} \end{cases}$$