Structure

Ryan Hernandez

Inference of Population Structure Using Multilocus Genotype Data

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

McGill

ryan.hernandez@me.com

Goals

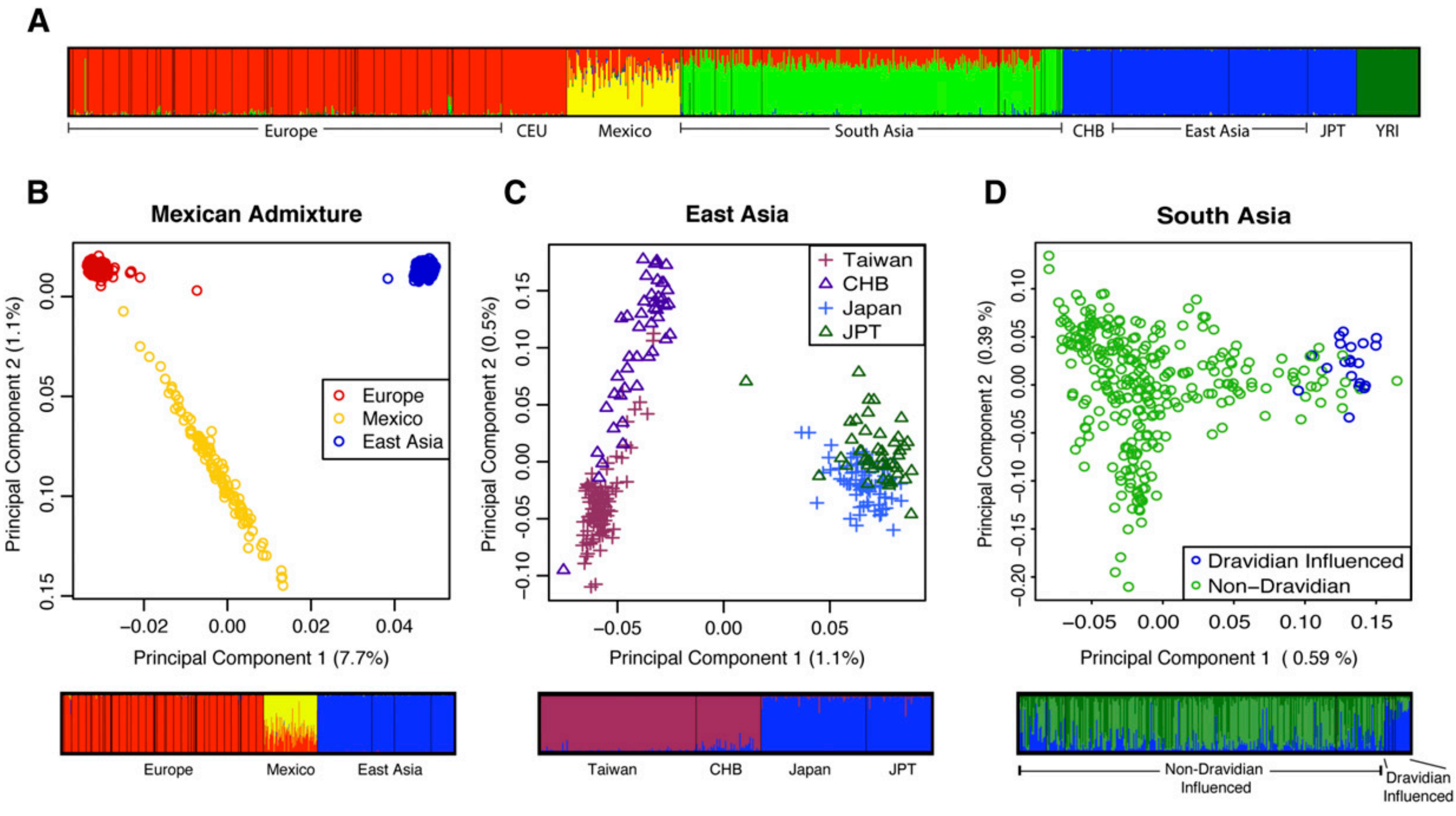
- How does the algorithm work?

Comparing populations

There are in general two ways to compare populations:

- Distance-based methods
  - Fst
  - Neighbor-joining
  - Principal Component Analysis (PCA)
- Model-based methods
  - STRUCTURE

Genomic Structure of Admixture



Auton et al. Gen. Res. (2009)

Structure

- In this paper, multiple algorithms are proposed for inferring admixture parameters.
- The ultimate goal is to learn how population structure has impacted genetic variation.
- This is done using MCMC, a common approach to solving Bayesian Inference problems.
- This was one of the first applications of MCMC in genetics.

Parameters

- X: **Our data**, the genotypes.
- Z: **What we want**, the populations of origin
- P: **What we need**, the allele freq's in all populations.

- Ultimately, we want to calculate: $Pr(Z,\ P \mid X)$
    - This is the posterior probability of the population of origin (and their frequencies) for all samples.

Assumptions

- Hardy-Weinberg Equilibrium (HWE).
- SNPs are independent (linkage equilibrium)
- We know how many populations contribute to our sample: K (though some statistics can be helpful).
- We know nothing about the population of origin
    - P(Z) = 1/K

Overview of the model

- Suppose we knew P and Z
    - i.e., we know the allele frequencies in each population, and which population each individual came from.

- It would then be easy to calculate Pr(X | Z, P)
    - The probability of an observed allele is just its frequency in the population of origin, **and** we multiply across sites.

Overview of the model

- Suppose we knew Z, but not P
    - i.e., which population each individual came from, but not the allele frequencies in those populations.
- It would then be easy to estimate P from X, Z
    - The maximum likelihood estimate for the allele frequency for a population is just given by the frequency of the allele in the sampled individuals from that population.
    - We can add a probability distribution on this using the so-called Dirichlet distribution (a continuous distribution between 0 and 1 in n dimensions) with mean given by the MLE.

Overview of the model

- Suppose we knew P, but not Z
  - i.e., we know the allele frequencies in each population, but not which population the individuals come from.

- It would then be easy to calculate Pr(Z | X, P)
  - This is just the relative probability of each population.
  - i.e., $Pr(Z{=}1 \mid X,P) = P(X \mid Z{=}1,P) / \sum_i P(X \mid Z{=}i, P)$

Key to the algorithm

- Assume you know everything by guessing, then update your guess!
- **Step 0:** Make random guess for population of origin
- **Step 1:** Given population of origin, calculate allele frequencies.
    - Let N[k,] be the number of chromosomes in population k with a particular allele (at each SNP).
    - The probability distribution for SNP i is $\mathrm{Beta}(1+N[k,i],\ 1+n-N[k,i])$.
        - n is the total number of chromosomes in population k.
    - `P[k,] = rbeta(nsnps, 1+P[k,], 1+2*n-P[k,])`
- **Step 2:** Given population allele frequencies, update population of origin
    - For each individual, calculate log-likelihood of data for each population.
    - Choose a population randomly according to relative probabilities (R)
    - `Z[i] = sample(1:K, size=1, prob=R)`
- **Step 3:** Repeat steps 1 & 2, keeping the results every c iterations, until m samples are drawn.