Selection Scan: Bayenv2

Environmental Correlation Analysis

PBIO381

Contents

- BAYENV2 Model
  - Null Model - Neutral Population Structure
  - Standardized Allele Frequencies
    - X^TX, a F_{ST} analog
- Data
  - Genome-Wide SNPs (balsam poplar)
  - Environmental: Lat, Long, Elev, Complex bioclim variables
- Analysis
  - Estimate Covariance matrix of neutral population structure
  - Test SNP x Environment correlation as a test of selection
- Draw inferences from results

Bayenv2

- Does not assume that populations are evolutionarily independent
- Estimates a null model of the covariance in allele frequencies between subpopulations (i.e. neutral population structure)
- Accounts for this covariance when inferring significant correlations between gene frequencies and the environment

Bayenv2:

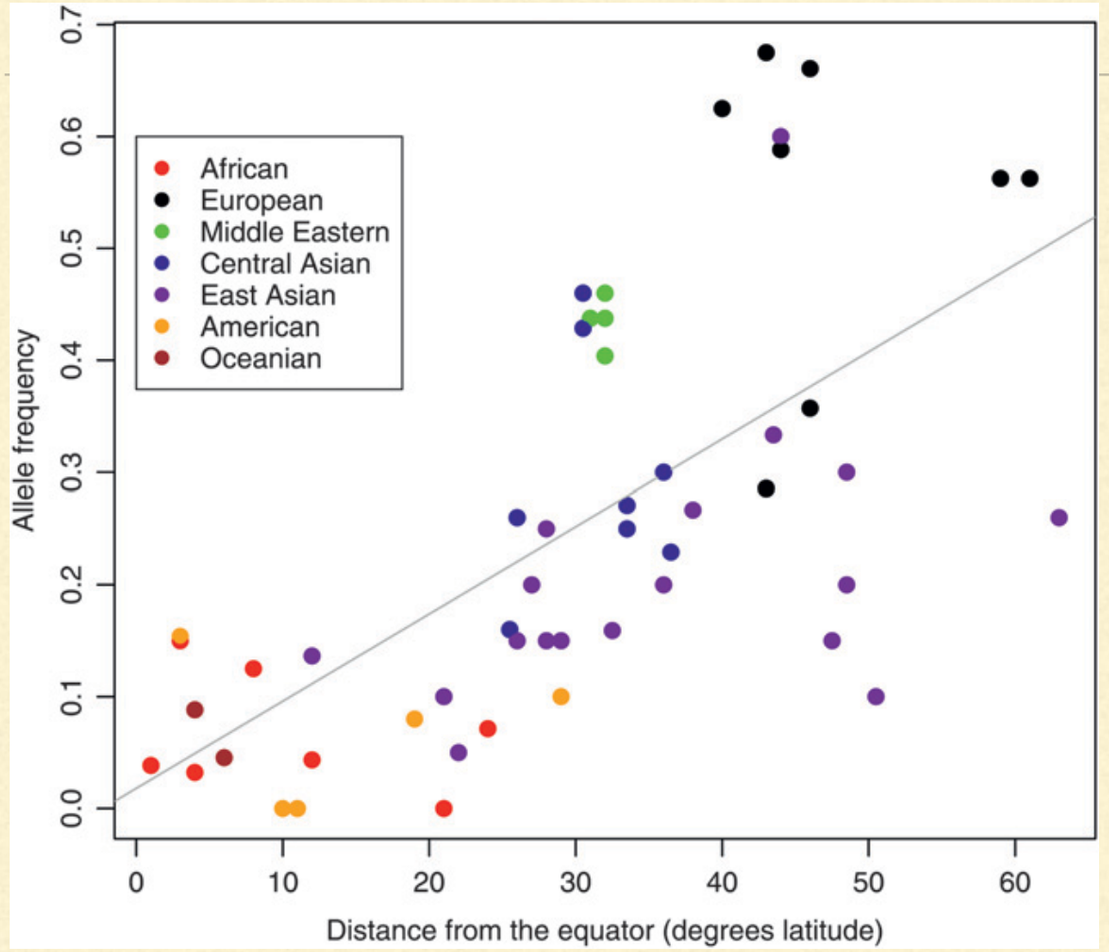


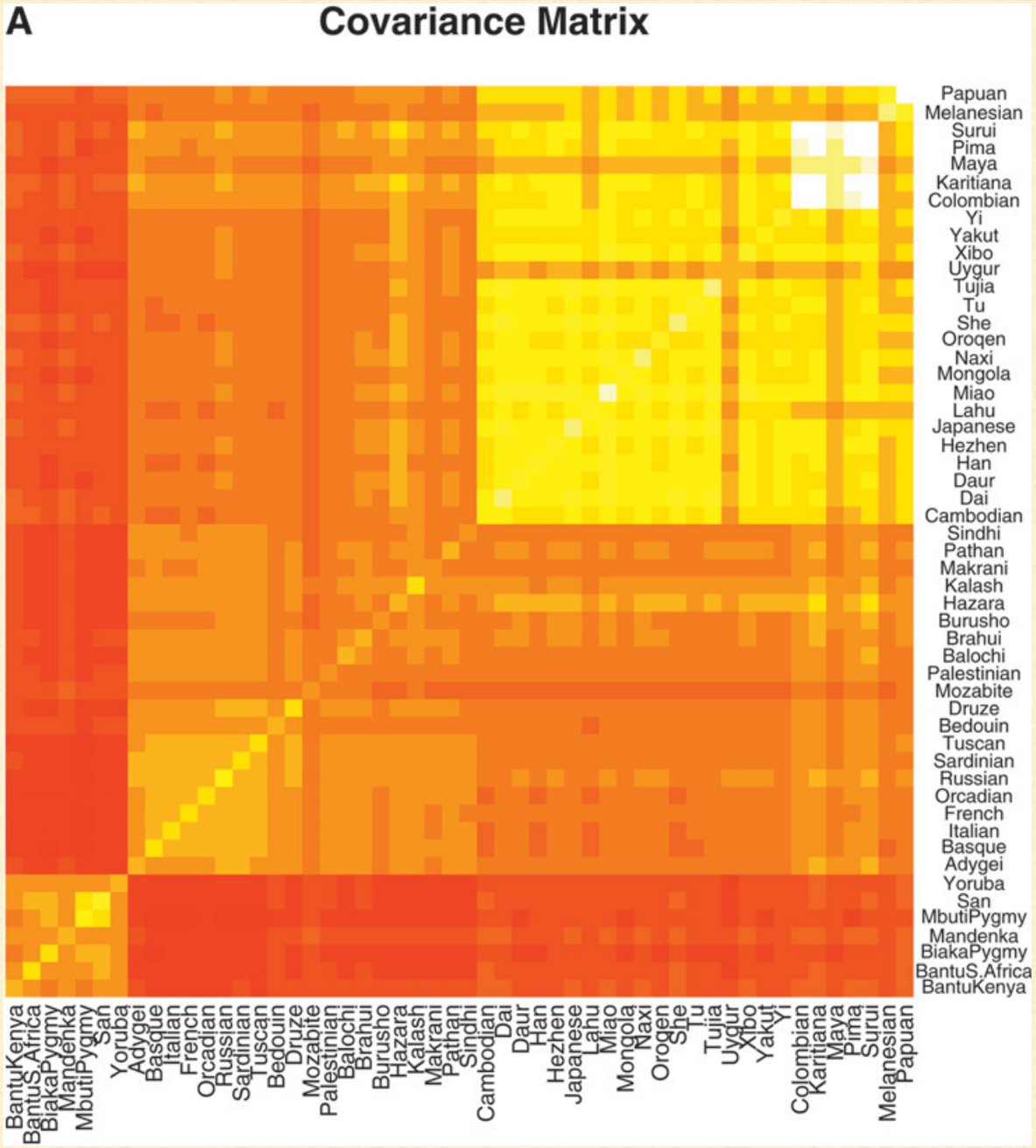
FIGURE 1.—The distance from the equator for each of 52 human populations, plotted against sample allele frequencies for the SNP AGT M235T in each population. The points are colored according to the geographic region each population belongs to, following region definitions of ROSENBERG *et al.* (2002). The data were generated using HGDP samples by THOMPSON *et al.* (2004) and are replotted on the basis of a figure in that article.

Coop et al, 2010

Selection or Drift?

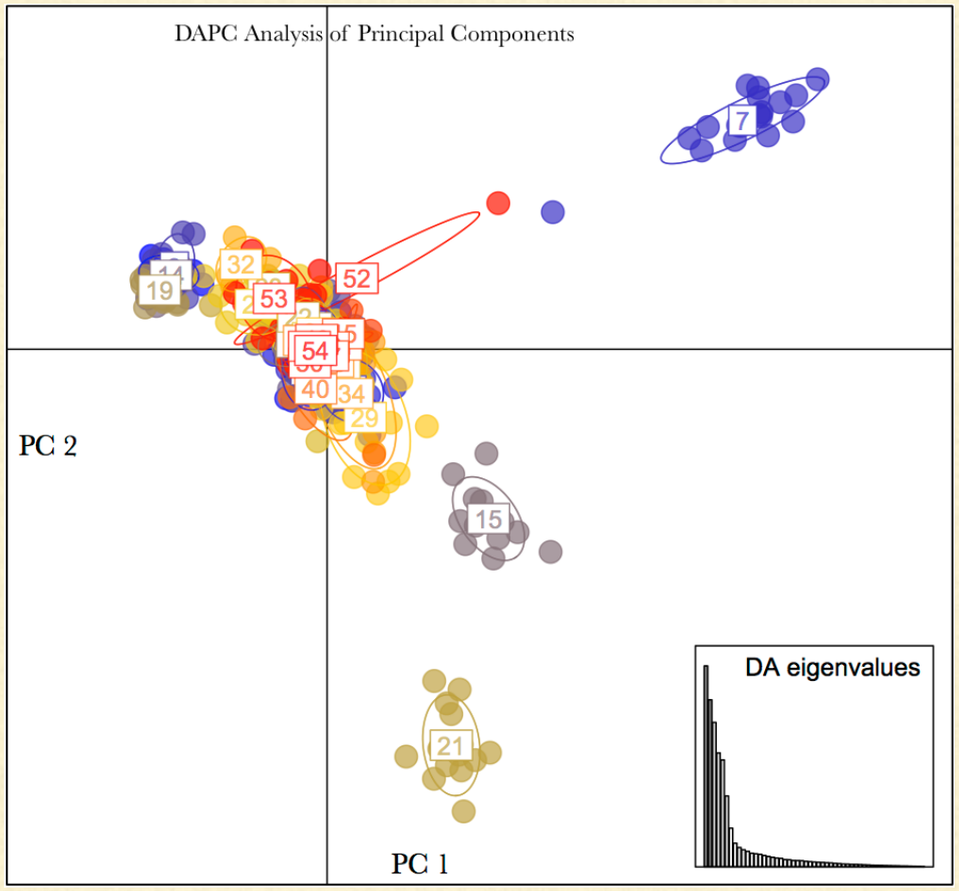
Sodium Retention in Humans

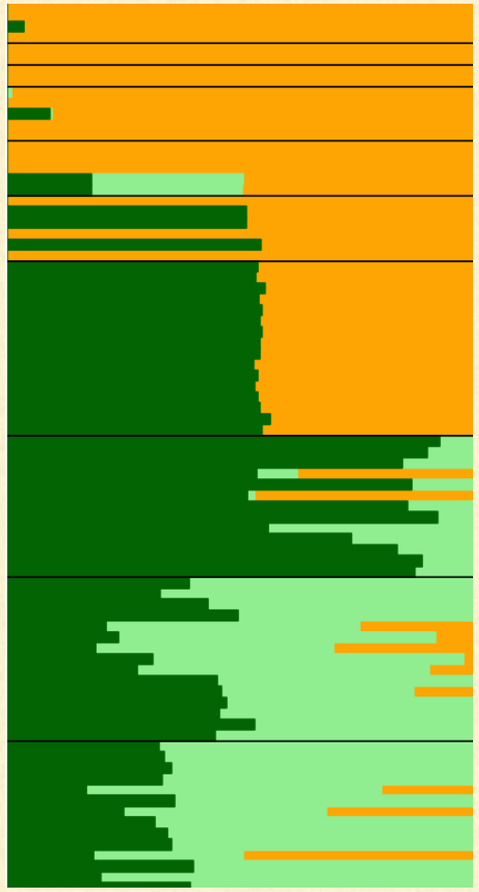
Bayenv2: Covariance of Gene Frequencies



52 Human Populations

Lighter colors represent higher values





Are Rear-Edge Populations a Concern for Climate Mitigation? Harnessing Genome Scans for Understanding Climate Adaptation in Range-Wide Populations of a Widespread Boreal Tree *Populus balsamifera*

Vikram E. Chhatre, Karl C. Fetter, Matthew C. Fitzpatrick, Stephen R. Keller

The University of Vermont

NSF

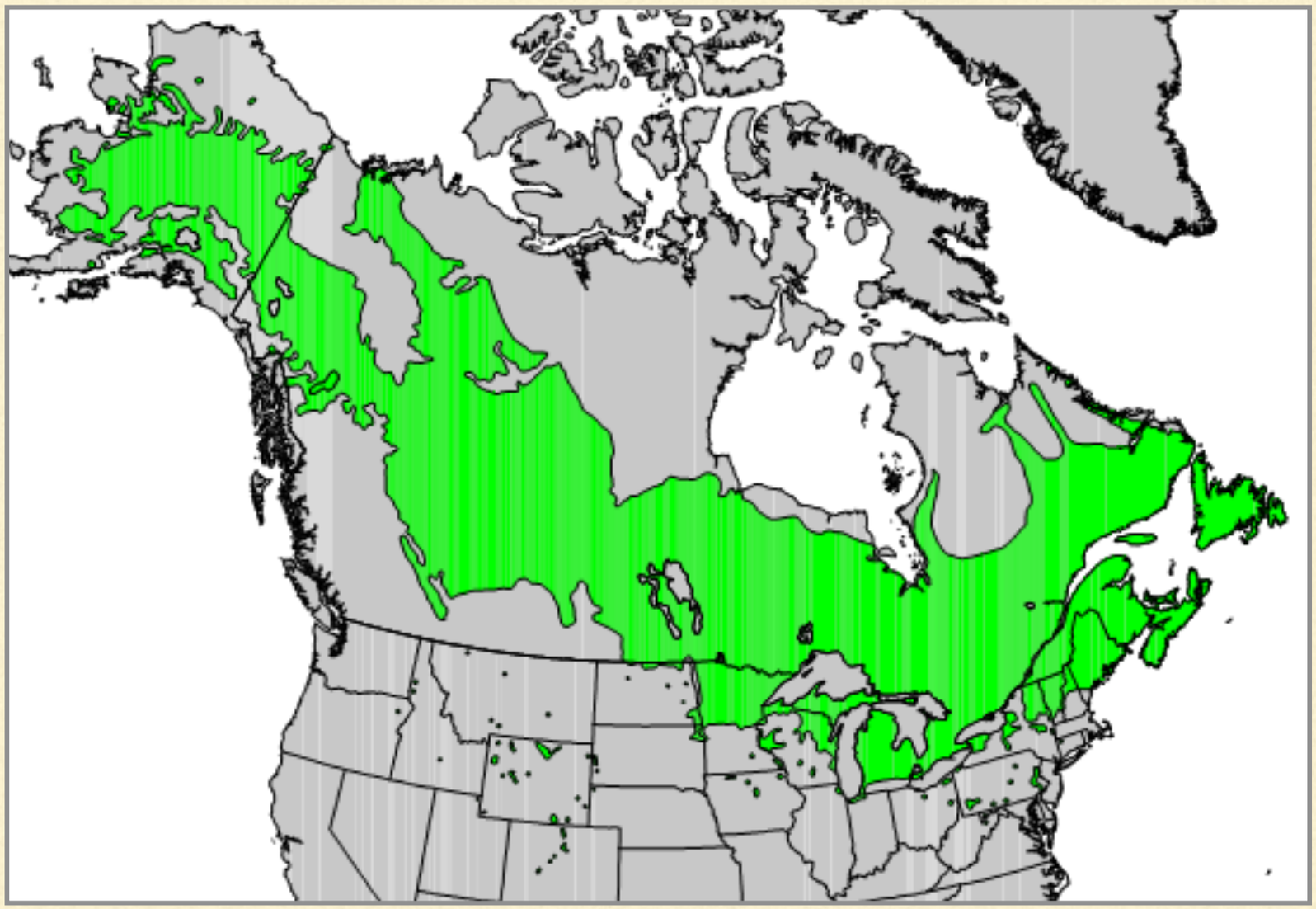
University of Maryland
Center for Environmental Science

Ecological genomics of climate adaptation in trees


Balsam Poplar

- How did climates of the past shape standing genetic variation?

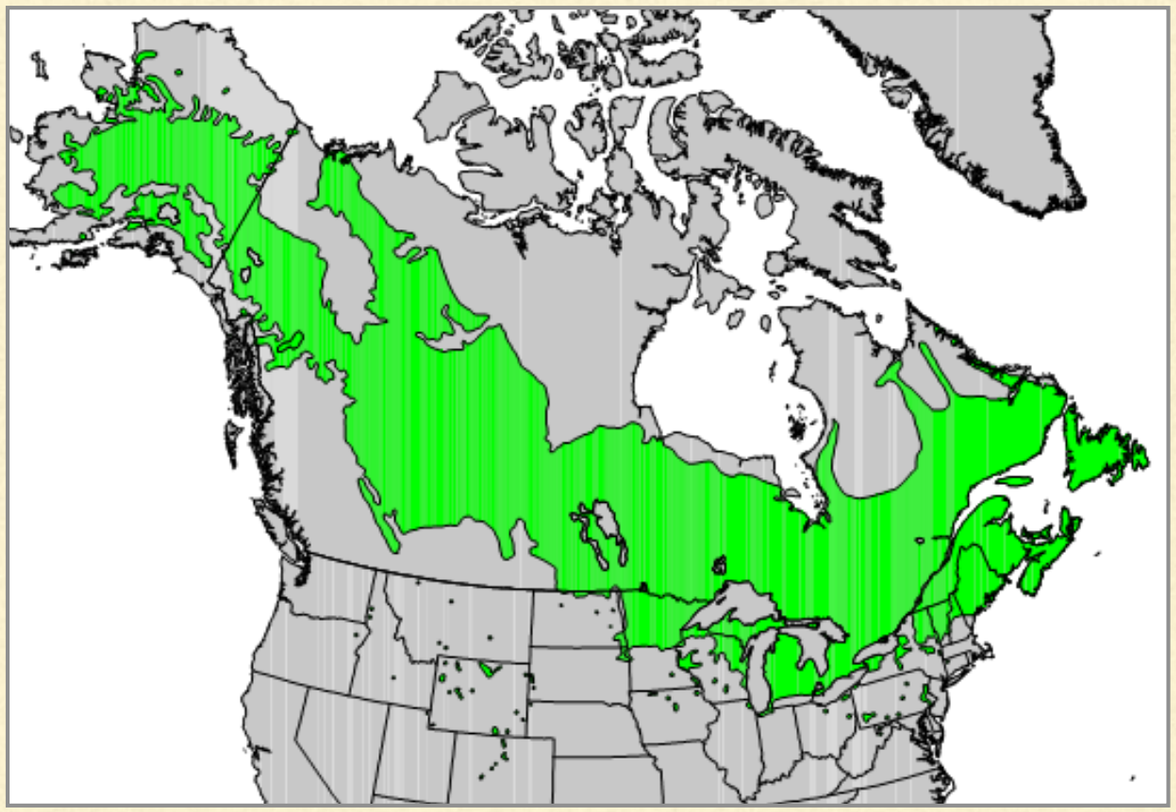
Populus balsamifera



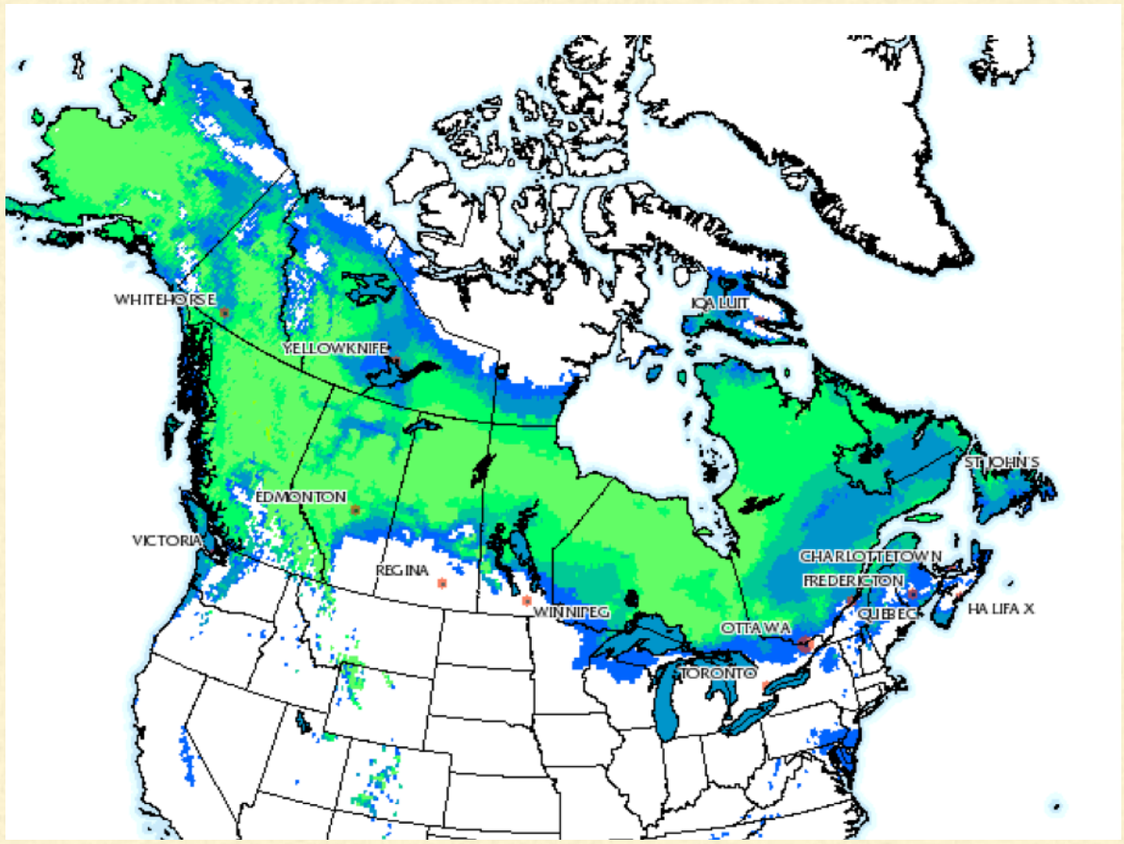
Range map (Little)

- One of the most widely distributed tree species in North America
- Occupies areas well above continental tree line
- Northernmost populations may be very sensitive to climate change

Conservation of adaptive genetic variation at the rear edge



Current distribution of *P. balsamifera*



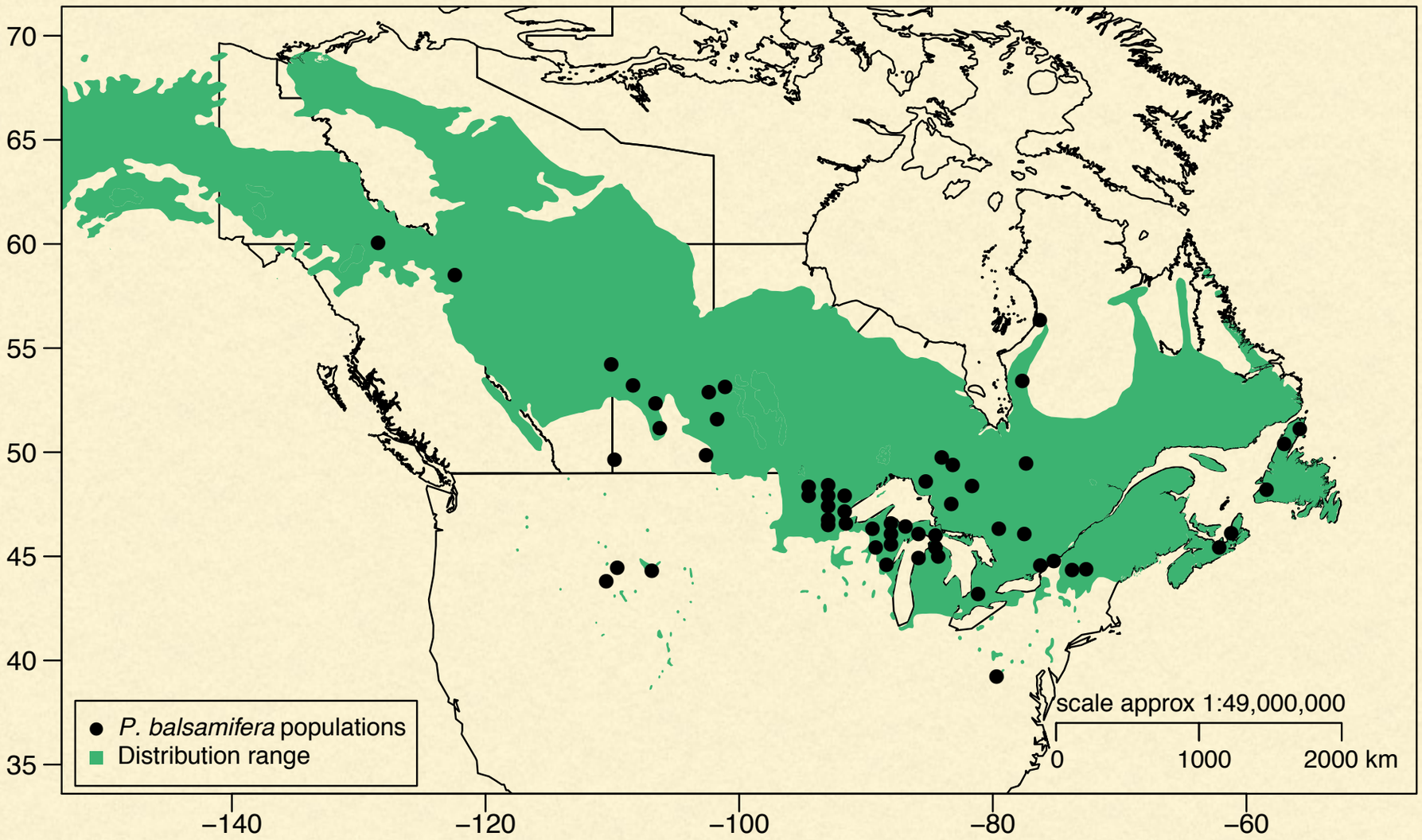
Future species distribution prediction

We need population samples from the rear edge to understand adaptation to warmer climate in the standing variation

Ecological genomics of climate adaptation

Bayenv2 Data Set

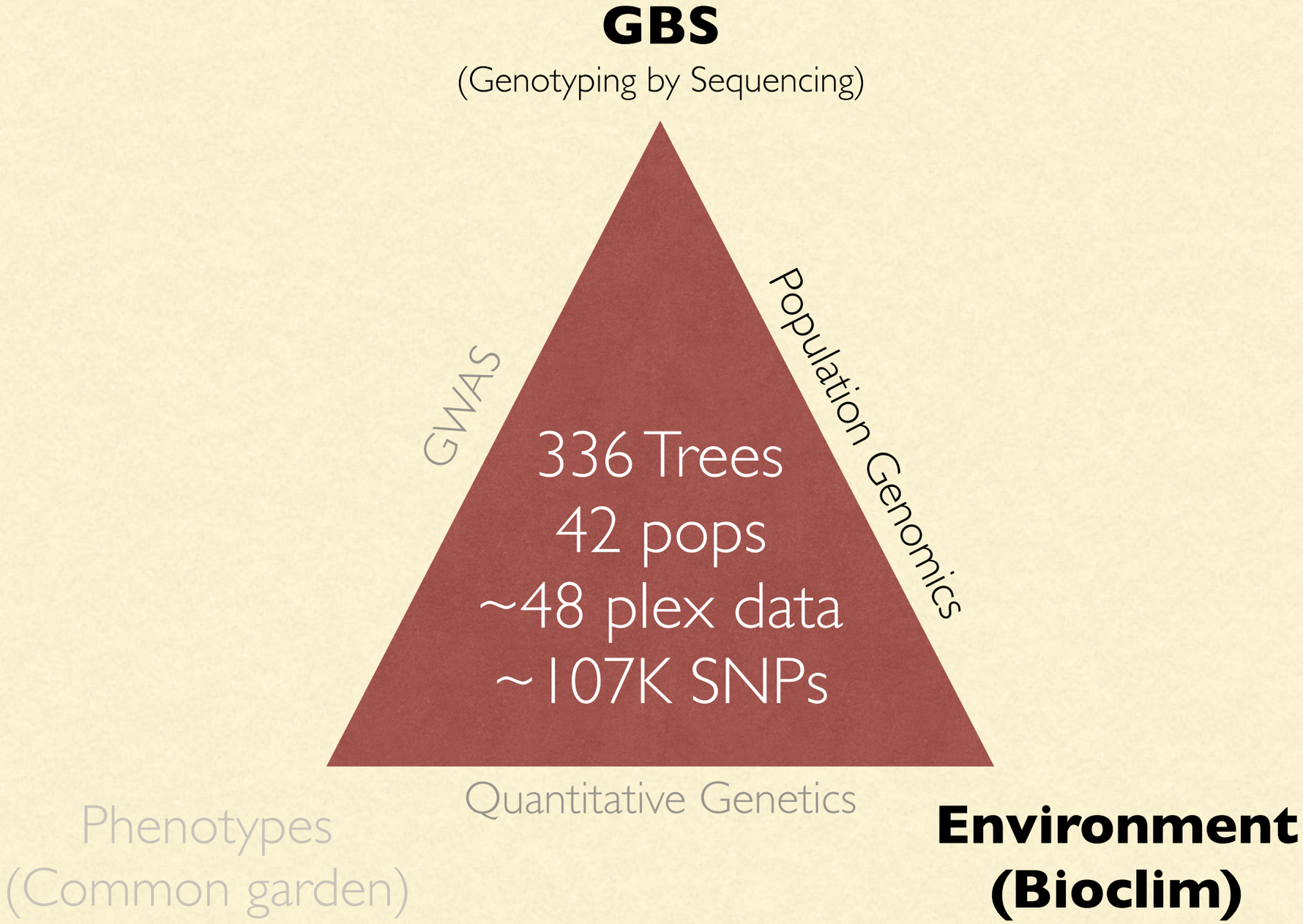
42 Populations, 336 Trees, 107K SNPs



Objectives

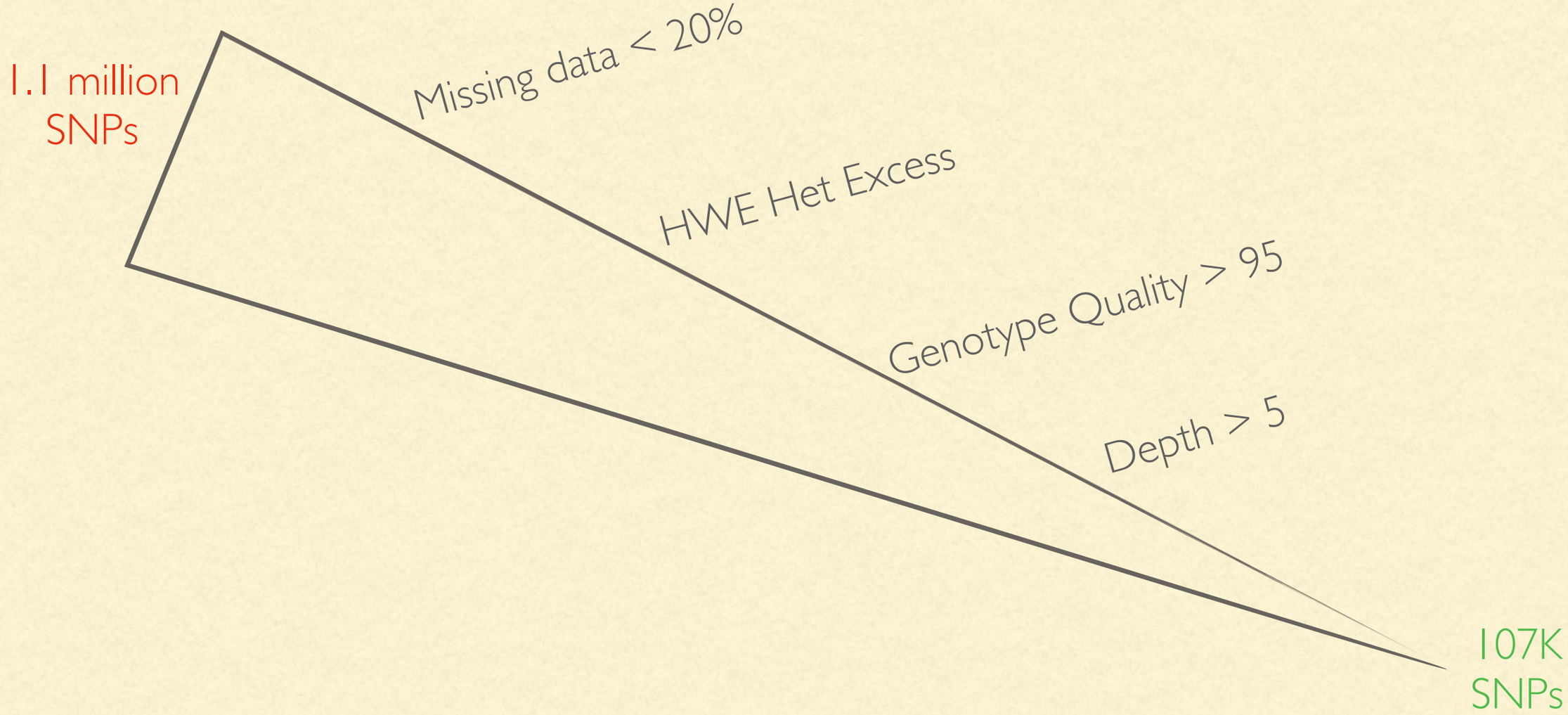
- Estimate a null model of neutral population structure
- Can we detect local adaptation manifest as gene+environment correlations along climatic and/or geographical gradients?

Methods: population genomics



GBS Data Filtering

48 Plex Sequencing for 336 Trees
Illumina Platform - Tassel GBS Pipeline



ENVIRONMENTAL DATA

WorldClim - Global Climate Data

Free climate data for ecological modeling and GIS

Bioclim

BIOCLIM

Bioclimatic variables are derived from the monthly temperature and rainfall values in order to generate more biologically meaningful variables. These are often used in ecological niche modeling (e.g., BIOCLIM, GARP). The bioclimatic variables represent annual trends (e.g., mean annual temperature, annual precipitation) seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., temperature of the coldest and warmest month, and precipitation of the wet and dry quarters). A quarter is a period of three months (1/4 of the year).

They are coded as follows:

BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter

Latitude
Longitude
Elevation

ENVIRONMENTAL DATA

Environmental Variables may be highly **correlated!**

SOLUTION?

Principal Component Analysis

Bayenv2 File Format

Genetic Data

Population Allele Counts

Why use allele counts when
we are interested in frequencies?

Environmental Data

Standardized

i.e. Subtract the Mean & Divide by STDEV

Our Timeline

Monday, November 2

Start estimating COVARIANCE matrix

Wednesday, November 4

Visualize & Understand COVARIANCE matrix
Begin Environmental Correlation Analysis

Monday, November 9

Understand Program Output
Determine significance using Bayes Factors
Spearman's rho from std. alle. freq.
How could you use X^TX, the population differentiation estimator