Practicalities

Using genetic associations with the environment to infer positive selection across genomes

Angela Hancock
January 30, 2018
Practical Matters

• Environmental Data Sets
• Methods
  • SAM
  • Dealing with confounding due to population structure
  • BayEnv
  • LFMM
  • Other Mixed model methods
• Simulation-based comparisons of methods
ISRIC – world soil information database

http://www.isric.org/
Datasets are available in the following categories:

- Temperature (5°×5° gridded versions)
- Precipitation (5°×5° and 2.5°×3.75° gridded versions)
- Pressure and Circulation Indices
- UK Climate Indices
- Mediterranean climate
- Alpine climate data
- High-resolution gridded datasets
- NCEP/NCAR Reanalysis data - May 2011: updated for 2010
- Paleoclimate
- Drought indices
FAO GeoNetwork

• Agriculture and Livestock
• Applied Ecology
• Base Maps, Remote Sensing
• Biological and Ecological Resources
• Climate
• Fisheries and Aquaculture
• Forestry
• Human Health
• Hydrology and Water Resources
• Infrastructures
• Land Cover and Land Use
• Population and Socio-Economic Indicators
• Soils and Soil Resources
• Topography

WORLDCLIM Project provides variables at several resolutions

<table>
<thead>
<tr>
<th>variable</th>
<th>10 minutes</th>
<th>5 minutes</th>
<th>2.5 minutes</th>
<th>30 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum temperature (°C)</td>
<td>tmin 10m</td>
<td>tmin 5m</td>
<td>tmin 2.5m</td>
<td>tmin 30s</td>
</tr>
<tr>
<td>maximum temperature (°C)</td>
<td>tmax 10m</td>
<td>tmax 5m</td>
<td>tmax 2.5m</td>
<td>tmax 30s</td>
</tr>
<tr>
<td>average temperature (°C)</td>
<td>tavg 10m</td>
<td>tavg 5m</td>
<td>tavg 2.5m</td>
<td>tavg 30s</td>
</tr>
<tr>
<td>precipitation (mm)</td>
<td>prec 10m</td>
<td>prec 5m</td>
<td>prec 2.5m</td>
<td>prec 30s</td>
</tr>
<tr>
<td>solar radiation (kJ m⁻² day⁻¹)</td>
<td>srad 10m</td>
<td>srad 5m</td>
<td>srad 2.5m</td>
<td>srad 30s</td>
</tr>
<tr>
<td>wind speed (m s⁻¹)</td>
<td>wind 10m</td>
<td>wind 5m</td>
<td>wind 2.5m</td>
<td>wind 30s</td>
</tr>
<tr>
<td>water vapor pressure (kPa)</td>
<td>vapr 10m</td>
<td>vapr 5m</td>
<td>vapr 2.5m</td>
<td>vapr 30s</td>
</tr>
</tbody>
</table>
Bioclim variables are derived from monthly WORLDCLIM data to create meaningful variables

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
BIO19 = Precipitation of Coldest Quarter
An early method: SAM (spatial analysis method)

• Simple linear model method
• Use geo-referenced environmental data and marker data with a focus on microsatellite data (for each possible state, set to 0 or 1)
• Test association between each allele and environmental variable using logistic regression
• Assess significance using two methods:
  • Likelihood ratio test
  • Wald test

\[ G = -2 \ln \frac{L}{L'} \]

\[ W = \frac{\beta_i}{\sigma(\beta_i)} \]

Joost et al., 2007, 2008, 2009
But confounding due to population structure may arise if structure correlates with the environmental variable... 

...even when the SNP has no functional effect
Population Structure causes correlations across the genome

Controlling for population structure can provide power to separate the signal from noise
Some methods to deal with Population Structure

• Genomic control: Scale down the test-statistic so that its median becomes the expected median.

• Use the first $n$ principle components of the genotype matrix (Price et al., 2006).

• Model the genotype effect as a random term in a mixed model, by explicitly describing the covariance structure between the individuals.
BayEnv: a linear mixed model method to assess evidence for correlations with environment

- Models the joint distribution of allele frequencies across populations for a variant as a function of
  - Population ‘history’ (null model)
  - Population ‘history’ + environment (alternative model)
- Then asks whether there is evidence a variant is an adaptation to a particular climate variable by comparing these two models in a Bayesian framework
Population history

• Demographic history is included in the model via a covariance matrix of populations

  • This is different from the assumption of quantitative trait mapping approaches, which include the kinship matrix to control for other loci that contribute to the trait (infinitesimal model)!

• The covariance structure is modeled under the assumption that transformed population allele frequencies have a multivariate normal distribution
Bayenv method

\[ H_0: \ y = \beta_0 + \mu + \varepsilon \]
\[ H_1: \ y = \beta_0 + \beta_1 x + \mu + \varepsilon \]

where \( y \) is the vector of allele frequencies,
\( \beta_0 \) is the intercept,
\( \mu \) is the random effect term due to population history, and
\( \varepsilon \) is the random error
\( x \) is the environmental variable,
\( \beta_1 \) is the effect size of environmental variable on allele frequencies,

\[ BF = \frac{Pr(D|M_1)}{Pr(D|M_0)} \]
Bayenv uses the (predicted) variance/covariance matrix to control for population structure.
Generating the kinship matrix

Since the population allele frequency is drawn from a normal distribution, it could be <0 or >1, which doesn’t make sense, therefore, a simple transformation is used:

\[ x_{kl} = g(\theta_{kl}) = \begin{cases} 
0 & \text{if } \theta_{kl} < 0 \\
\theta_{kl} & 0 \leq \theta_{kl} \leq 1 \\
1 & \theta_{kl} > 1.
\end{cases} \]

Population allele frequency variable, not constrained to be between 0 and 1.
Generating the kinship matrix

Joint posterior over all loci

\[ P(\Omega, \theta_1, \ldots, \theta_L, \varepsilon_1, \ldots, \varepsilon_L|n_1, m_1, \ldots, n_L, m_L) \propto \]

- Prior on the allele counts
- Prior on the vector of a.f.s
- Prior on the covariance matrix
- Prior on the ancestral frequency at a locus

- MCMC to explore the sample space and sequentially update parameters
- Decide whether to accept \( \theta' \) based on the ratio of the alternative to the null posterior
The Bayes factor

\[ BF = \frac{Pr(D|M_1)}{Pr(D|M_2)} \]

Interpreting the Bayes factor

<table>
<thead>
<tr>
<th>K</th>
<th>dB</th>
<th>bits</th>
<th>Strength of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1:1</td>
<td>&lt; 0</td>
<td></td>
<td>Negative (supports ( M_2 ))</td>
</tr>
<tr>
<td>1:1 to 3:1</td>
<td>0 to 5</td>
<td>0 to 1.6</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>3:1 to 10:1</td>
<td>5 to 10</td>
<td>1.6 to 3.3</td>
<td>Substantial</td>
</tr>
<tr>
<td>10:1 to 30:1</td>
<td>10 to 15</td>
<td>3.3 to 5.0</td>
<td>Strong</td>
</tr>
<tr>
<td>30:1 to 100:1</td>
<td>15 to 20</td>
<td>5.0 to 6.6</td>
<td>Very strong</td>
</tr>
<tr>
<td>&gt; 100:1</td>
<td>&gt; 20</td>
<td>&gt; 6.6</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

*Jeffreys 1961*

In practice, BayEnv authors recommend using a ranking approach rather than trusting the BF$s are well-calibrated.
Comparison of Bayenv to other methods

Power to detect a correlation between allele frequency and climate

Latitude

Summer precipitation

Practicalities: Methods
Bayenv2

• Allows calculation of a standardized set of allele frequencies by removing the covariance among populations and making the residuals available for further analyses.

• Use these to:
  • Conduct non-model based tests of population differentiation
  • Non-parametric tests of correlation (e.g., Spearman's rho)
Latent factor mixed model approach (LFMM)

• Similar to BayEnv, but uses *factors derived from* the covariance matrix to model population history

• Individual-based rather than population-based

• *Simultaneously models correlation with population structure and environment*, so could gain some power when structure is correlated with the environment

*Frichot et al., 2013*
LFMM: The Model

\[ G_{il} = \mu_l + \beta_l^TX_i + U_i^TV_l + \epsilon_{il} \]

where

- \( G \) is a response variable in a Bayesian regression model
- Gaussian prior distributions on \( \mu \) and \( \beta_l \)
- \( U_i \) and \( V_l \) are scalar vectors with Gaussian priors
- \( B_l \) is a vector of regression coefficients

- Use Gibbs sampler to move through sample space
- Use a stochastic algorithm to compute standard deviations and \( z \)-scores for the environmental effects.
- Compare each locus to the genomic background and retained loci with \( z \)-scores exhibiting the highest absolute values
Comparison among methods

Simulated genetic data under different models:

Highly Structured Isolation with Migration  Isolation Migration Model  Stepping Stone Model

Used 4 approaches:  Population Differentiation (Bayescan)  Naive regression  LFMM  Bayenv

De Villemereuil et al., 2014
FDR vs. Significance

correlated
Highly Structured
Isolation with
Migration

uncorrelated
Highly Structured
Isolation with
Migration

Isolation
Migration Model

Stepping
Stone Model
Statistical Power vs. Significance

**correlated**
Highly Structured Isolation with Migration

**uncorrelated**
Highly Structured Isolation with Migration

Isolation Migration Model

Stepping Stone Model
FDR vs. Significance (polygenic case)

- correlated
  - Highly Structured Isolation with Migration
- uncorrelated
  - Highly Structured Isolation with Migration

Practicalities: Methods

(A) Isolation Migration Model

(B) Stepping Stone Model
Power vs. Significance (polygenic case)

correlated
Highly Structured Isolation with Migration

Isolation Migration Model

uncorrelated
Highly Structured Isolation with Migration

Stepping Stone Model
Simulation-based comparison of methods under different migration models and selfing vs outcrossing

De Mita et al., 2013
Included several sampling schemes across a grid

A  Indexes and gradient

B  Sampling scheme S1

C  Sampling scheme S2

D  Sampling scheme S3

E  Sampling scheme S4

F  Sampling scheme S5
Diverse methods included in analysis, but useful to see how BayEnv (CWDRP) compares to others.

<table>
<thead>
<tr>
<th>Method and reference</th>
<th>Technique</th>
<th>Underlying model</th>
<th>Env. variable</th>
<th>Control loci</th>
<th>Sampling</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>GLM</td>
<td>Independence of observations</td>
<td>Yes</td>
<td>No</td>
<td>Individuals</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Joost et al. (2007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEE</td>
<td>GEE</td>
<td>Independence of clusters</td>
<td>Yes</td>
<td>No</td>
<td>Several individuals per population</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Porcoet et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CWDRP</td>
<td>MCMC</td>
<td>Island model</td>
<td>Yes</td>
<td>Yes</td>
<td>Frequencies</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Coop et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLK</td>
<td>Forward simulations</td>
<td>Multiple divergence model</td>
<td>No</td>
<td>Yes</td>
<td>Frequencies</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Bonhomme et al.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BN</td>
<td>Coalescent simulations</td>
<td>Island model</td>
<td>No</td>
<td>Yes</td>
<td>Frequencies</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>EHF</td>
<td>Coalescent simulations</td>
<td>Hierarchical island model</td>
<td>No</td>
<td>Yes</td>
<td>Frequencies</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Excoffier et al. (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VDB</td>
<td>Coalescent simulations</td>
<td>Pairwise divergence model</td>
<td>No</td>
<td>Yes</td>
<td>Frequencies</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Vitalis et al. (2001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FG</td>
<td>RJ-MCMC</td>
<td>Island model</td>
<td>No</td>
<td>No</td>
<td>Frequencies</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Foll &amp; Gaggiotti (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

De Mita et al., 2013
There is some variation in the performance of different methods across demographic models.

De Mita et al., 2013
Several methods perform very poorly in models with selfing.

**Practicalities: Method comparisons**

![Graph showing method comparisons](image)
Depending on the migration model and sampling scheme, different methods perform best.

*IM*: isolation migration

*IBD*: isolation by distance

1R: expansion from 1 refugium

2R: expansion from 2 refugia

Lotterhos and Whitlock 2015
PCA Adapt also performs well
Sampling and Scale

Linear model-based methods assume the residuals are normally distributed and have a constant variance.

Cases where a single sample or population is divergent from the others genetically and resides in a divergent environment are especially problematic and can strongly affect the results.

Possible solutions:

• try transforming the data
• leave out outliers
• use a non-parametric method (e.g., BayEnv, Partial Mantel)
How does power compare across different sampling schemes?

Random vs. paired vs. transects

Lotterhos and Whitlock 2015
Paired > transects > random

IM: isolation migration
IBD: isolation by distance
1R: expansion from 1 refugium
2R: expansion from 2 refugia
For some migration models, BayEnv has power at a low selection coefficient.
Genotype-phenotype association studies ("GWAS") are similar to genotype-environment association studies

- **Genotype-phenotype association:**
  Calculate a correlation between a SNP and a phenotype while controlling for other SNPs in the genome

- **Genotype-environment association:**
  Calculate a correlation between a SNP and an environmental variable controlling for population structure

\[ Y = X\beta + u + \epsilon, \quad u \sim N(0, \sigma_g K), \quad \epsilon \sim N(0, \sigma_e I) \]
Genotype-phenotype association studies ("GWAS") are similar to genotype-environment association studies

- Genotype-phenotype association:
  
  Calculate a correlation between a SNP and a phenotype while controlling for other SNPs in the genome

Mixed model approach for genotype-phenotype mapping

\[ Y = X\beta + u + \epsilon, \; u \sim N(0, \sigma_g K), \; \epsilon \sim N(0, \sigma_e I) \]

| Phenotype effect | SNP effect | ‘Error’ terms | Kinship matrix | Other random error |
Q: But why is a covariance matrix used in G-P association mapping to represent other SNPs contributing to the phenotype??

A: Fisher’s infinitesimal model states that traits are shaped by many many small effect loci scattered across the genome.

This means that the error term in a G-P mixed model is similar to the error terms used in G-E associations.
Q: Why is this cool?

A: Because a lot more work has been done to speed up G-P association methods compared to G-E association methods.

Using G-E methods facilitates large-scale genome-wide analyses.
GEMMA

• We will use GEMMA for conducting climate correlation analyses in the tutorial

• GEMMA uses a linear mixed model approach to remove the effects of kinship before estimating the correlation between a SNP and a phenotype (here climate variable)

• GEMMA is based on the earlier EMMA software and gives equivalent results, but is much faster (linear in the number of individuals versus quadratic).

• This speed is accomplished by replacing the eigen decomposition of the K(inship) matrix with a set of recursion equations
GEMMA

• GEMMA provides an estimate of $\beta$ (PVE) and can conduct several tests to assess significance for the explanatory power of the SNP:
  • LRT requires calculation of ML estimate, but is generally considered more reliable than Wald or score
  • Wald (A Wald test is conducted by comparing the coefficient's estimated value with the estimated standard error for the coefficient – assumes normality)
  • Score test (Cochran-Armitage test for trend assuming additive effect)