Inference of population structure and local adaptation using population genomic data

A tutorial using the R package LEA

Olivier.Francois@imag.fr

Computational and Mathematical Biology group
Université Grenoble-Alpes, France

March, 2015

Outline: Part 1

- Theory (40min): Inference of population structure and individual ancestry coefficients
- Estimates based on sparse matrix factorization methods
- Detection of outlier loci
- Practice (40min): Data analysis using `R`

Outline: Part 2

- Theory (40min): Ecological association tests.
- Estimates based on latent factor models
- Detection of genetic variation linked to ecological gradients
- Practice (40min): Data analysis using `R`

Data: Genotypic matrices

- Population genetic data for a diploid species are recorded in a genotypic matrix

$$X = \begin{pmatrix} 0 & 0 & 1 & \dots \\ 1 & 2 & 0 & \dots \\ 0 & 2 & 0 & \dots \\ \dots & & & \dots \end{pmatrix}$$

  with n rows (individual genotypes) and L columns (genomic position or locus).

- Each locus is called an SNP (*Single Nucleotide Polymorphism*), and each value (0,1,2) encodes the number of mutant (or derived) nucleotides at locus ℓ.

Inference of population structure

- Sampled genotypes in X **share ancestry from K ancestral populations, where K is unknow**.
- The ancestry coefficients are stored in the **Q**-matrix, and displayed using barplot representations.
- $\mathbf{q}_{ik}$ is the *fraction of individual i's genome that originates from ancestral population k*.
- A **F**-matrix records allele frequencies at each locus and in all K ancestral populations.

Nonnegative Matrix Factorization: Rationale

- The observed frequencies are a combination of ancestral frequencies, with weights given by $\mathbf{Q}$

$$\mathbf{p}(x_{i\ell} = j) = \sum_{k=1}^{K} \mathbf{q}_{ik} \mathbf{f}_{k\ell}(j)\,, \quad j = 0, 1, 2.$$

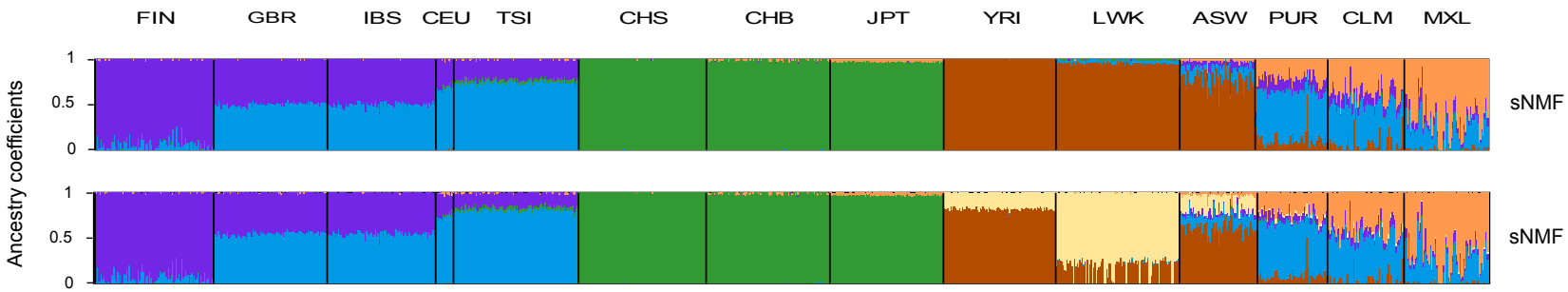
- which writes as

$$\mathbf{P} = \mathbf{QF}\,.$$

- The matrices Q and F are factors of the P matrix that can be estimated using matrix factorization methods.

Visualization of ancestry coefficients.

- The individual ancestry coefficients, stored in the matrix $\mathbf{Q}$, can be displayed using a barplot representation as follows



Q-matrix computed from the 1000 Genomes project data ($n = 1,092$ individuals and $L = 2.2M$ SNPs)

NMF: Principles

- **NMF** (Nonnegative Matrix Factorization) computes least-square estimates of admixture proportions
- For the normalized data matrix $\mathbf{X}$, it finds matrices $\mathbf{Q}$ and $\mathbf{F}$ that minimize

$$\mathrm{LS}(\mathbf{Q},\mathbf{F}) = \|\mathbf{X} - \mathbf{Q}\mathbf{F}\|_{\mathrm{F}}^2\,, \quad \mathbf{Q},\mathbf{F} \geq 0\,.$$

- We used additional constraints

$$\sum_{k=1}^{K} \mathbf{q}_{ik} = 1\,, \quad \sum_{j=0}^{2} \mathbf{f}_{k\ell}(j) = 1\,, \quad j = 0,1,2.$$

Sparse NMF (sNMF)

- sNMF computes **regularized** least-square estimates of admixture proportions

$$\mathrm{LS}(\mathbf{Q},\mathbf{F}) = \|\mathbf{X} - \mathbf{Q}\mathbf{F}\|_{\mathrm{F}}^2 + \sqrt{\alpha}\sum_{i=1}^{n}\|\mathbf{q}_{i\cdot}\|_1^2\,, \quad \mathbf{Q},\mathbf{F} \geq 0\,,$$

  where α is a non-negative **regularization** parameter that penalizes intermediate values of ancestry coefficients.

- We solved the LS problem by using the **Alternating Non-negativity-constrained Least Squares algorithm with the Active Set** method (ANLS-AS, Kim and Park 2011).

Cross-entropy criterion

- We employ a **cross-validation** technique to evaluate the prediction error of ancestry estimation algorithms
- Our criterion provides an estimate of **cross-entropy**, measuring the capability of an algorithm to correctly predict missing (or masked) genotypes.

Choosing K and α: Cross-entropy criterion

- The cross-entropy criterion provides an estimate of the following quantity

$$H(p^{\text{sample}}, p^{\text{pred}}) = -\sum_{j=0}^{2} p^{\text{sample}}(j) \log p_{i\ell}^{\text{pred}}(j), \quad j = 0, 1, 2.$$

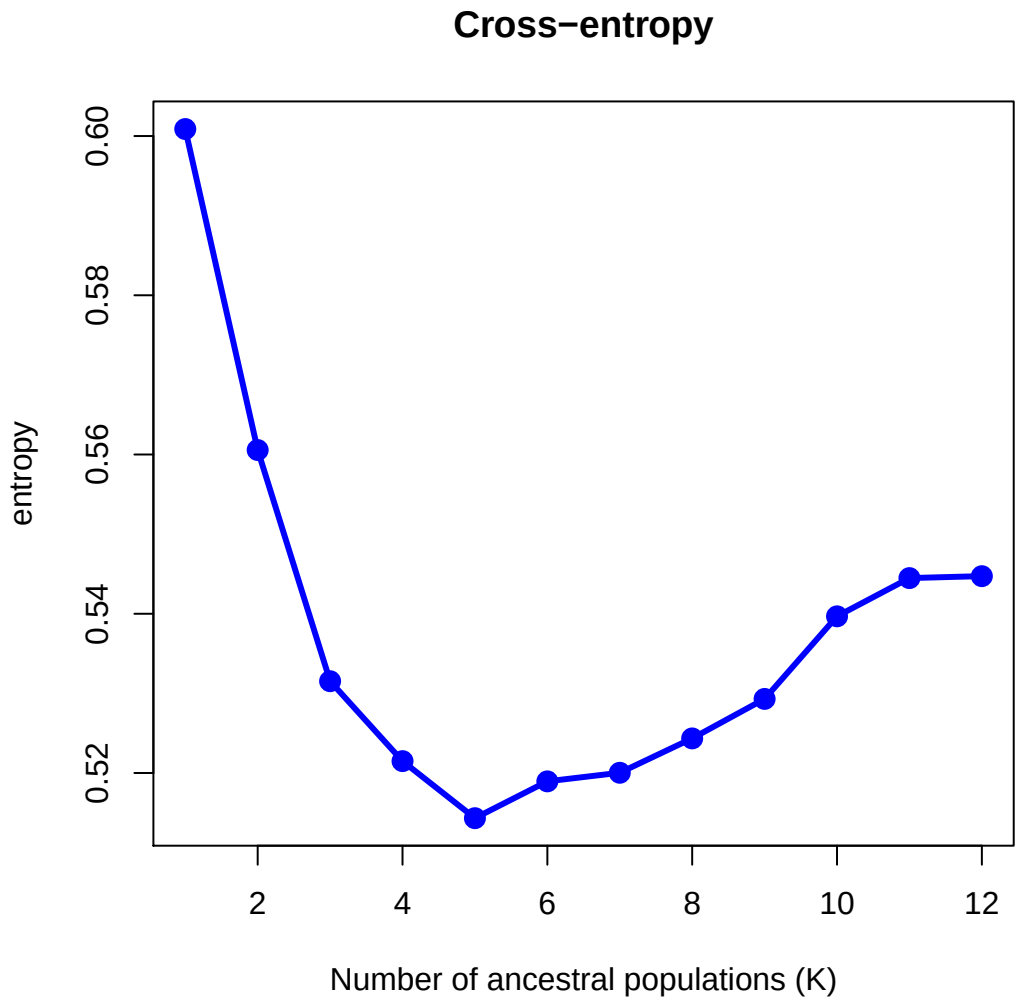
Cross-entropy for sNMF

- Running sNMF example for $K = 1 - 12$ ($\alpha = 100$)

```
project=snmf(genotype,K=1:12,alpha=100,entropy=T)

entropy=sapply(1:12,FUN=function(i)
cross.entropy(project,K = i))
```

Cross-entropy for sNMF



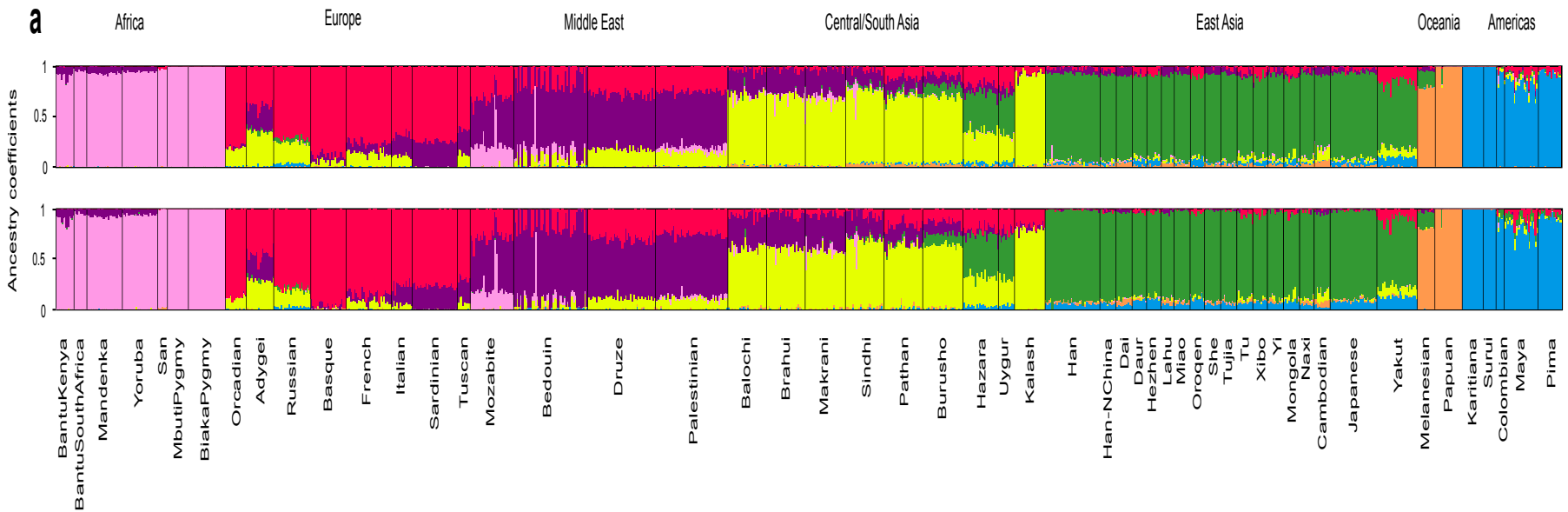
Cross-entropy as a function of K. Lower values indicate better models

Test assay used by Frichot et al. (2014)

Data set	Sample size	Number of SNPs
HGDP00778	934	78K
HGDP00542	934	48.5K
HGDP00927	934	124K
HGDP00998	934	2.6K
HGDP01224	934	10.6K
HGDP-CEPH	1,043	660K
1000 Genomes	1,092	2.2M
A. thaliana	168	216K

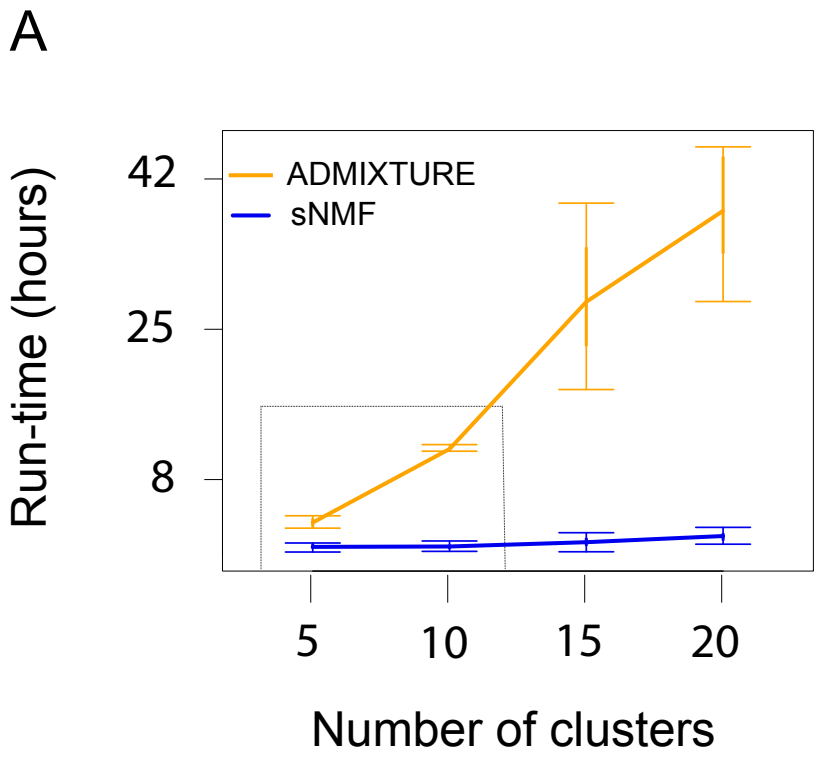
Ref: Patterson et al. 2012; Li et al. 2008;The 1000 Genomes Project Consortium 2012; Atwell et al. 2010

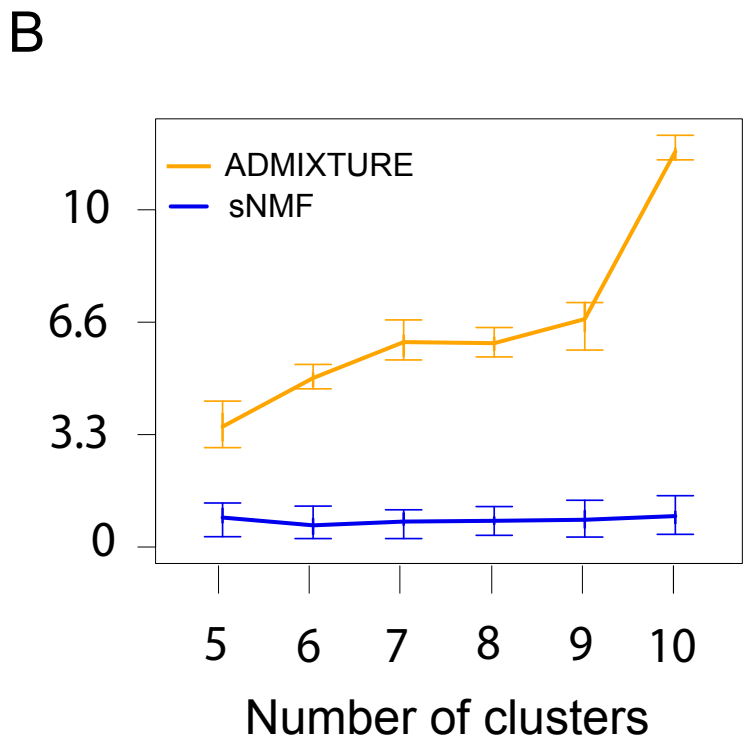
HGDP-CEPH (660K SNPs): sNMF estimates are very close to those of likelihood-based approaches



Ancestry estimates using ADMIXTURE *(top) and* sNMF *(down)*

sNMF runs faster than likelihood-based approaches (HGDP-CEPH data, 660K SNPs)





Detecting outlier loci: genome scan for selection

- Rationale: Locally adaptive loci exhibit larger differences in allele frequency than selectively neutral loci.
- Common approaches to separate selective from neutral processes focus on the evaluation of population differentiation statistics across loci (Lewontin and Krakauer 1975). The most classical measure of population differentiation is

$$F_{\mathrm{ST}} = 1 - \frac{\sigma_S^2}{\sigma_T^2},$$

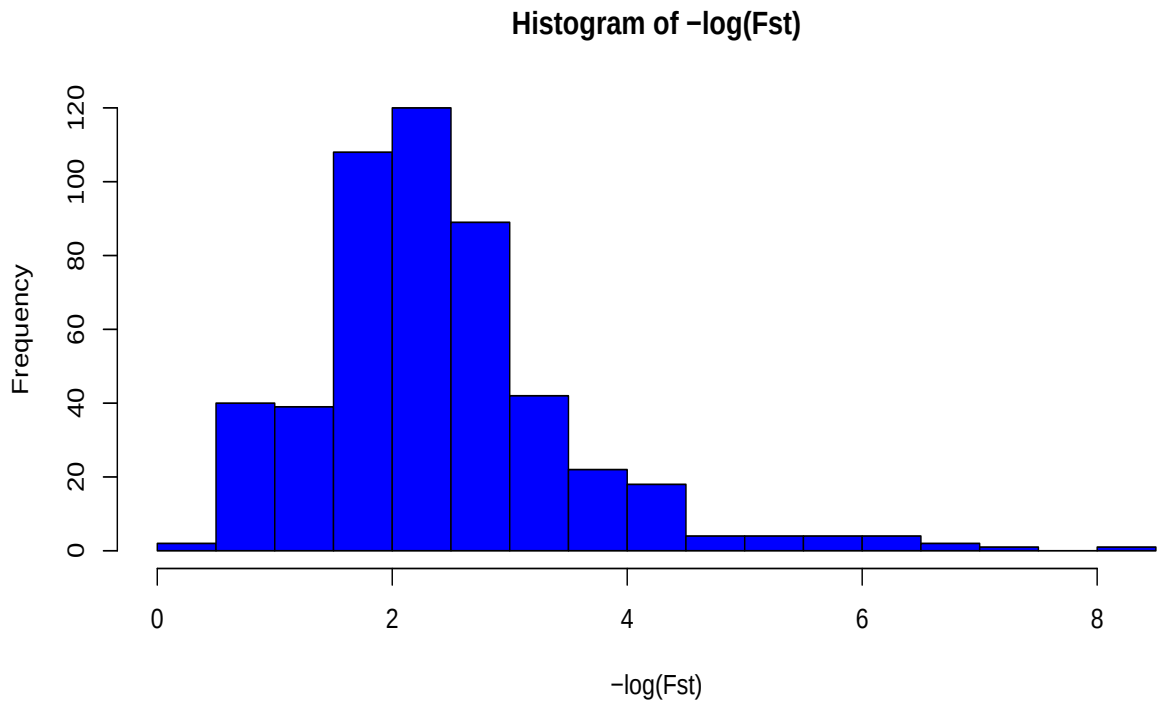
where σ_S^2 is the variance of allele frequencies within subpopulations (or ancestral populations) and σ_T^2 is the variance in the total population.

Statistical testing problem

- Null Hypothesis: For each locus, we want to test the null-hypothesis

$$F_{\mathrm{ST}} = F_{\mathrm{ST}}^{\mathrm{neutral}}$$

where $F_{\mathrm{ST}}^{\mathrm{neutral}}$ is the (unknown) population differentiation value for non-adaptive alleles.



Multiple testing and test calibration issues

- False Discovery Rate (FDR) is defined as

$$\text{FDR} = \text{prob}(\text{False Discovery} \mid \text{Positive test}) = q.$$

- The Benjamini-Hochberg algorithm provides a list of candidate loci with expected FDR $= q$ (see `R` scripts).
- The algorithm requires that the test are correctly calibrated, ie, the distribution of p-values is uniform when we assume that the null hypothesis, H_0, is correct.

Converting F_{ST}'s into significance values

- Theory: F_{ST} corresponds to the squared correlation coefficient in a regression model where allele frequencies are regressed on population labels. We defined

$$t - \text{scores} = \sqrt{\frac{(n-2)F_{\mathrm{ST}}}{1-F_{\mathrm{ST}}}} \sim t(n-2)$$

  where $t(n-2)$ is the Student distribution with $n-2$ degrees of freedom.

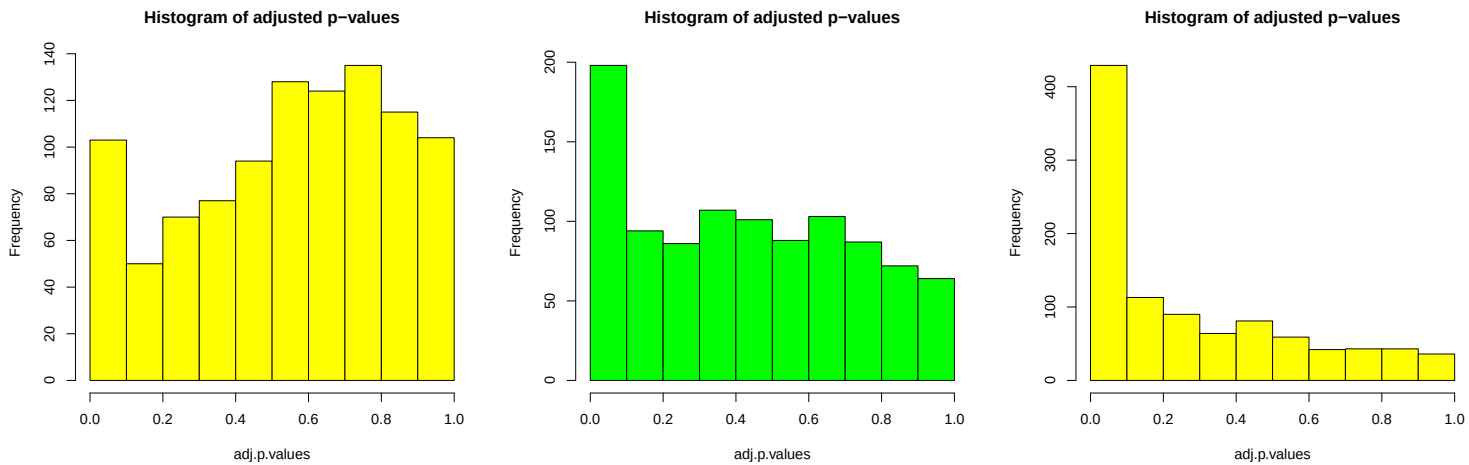
- Using `R`, significance values can be computed from t-scores as follows

```
p.values=2-2*pt(t.scores, df=n-2, lower=T)
```

Test calibration

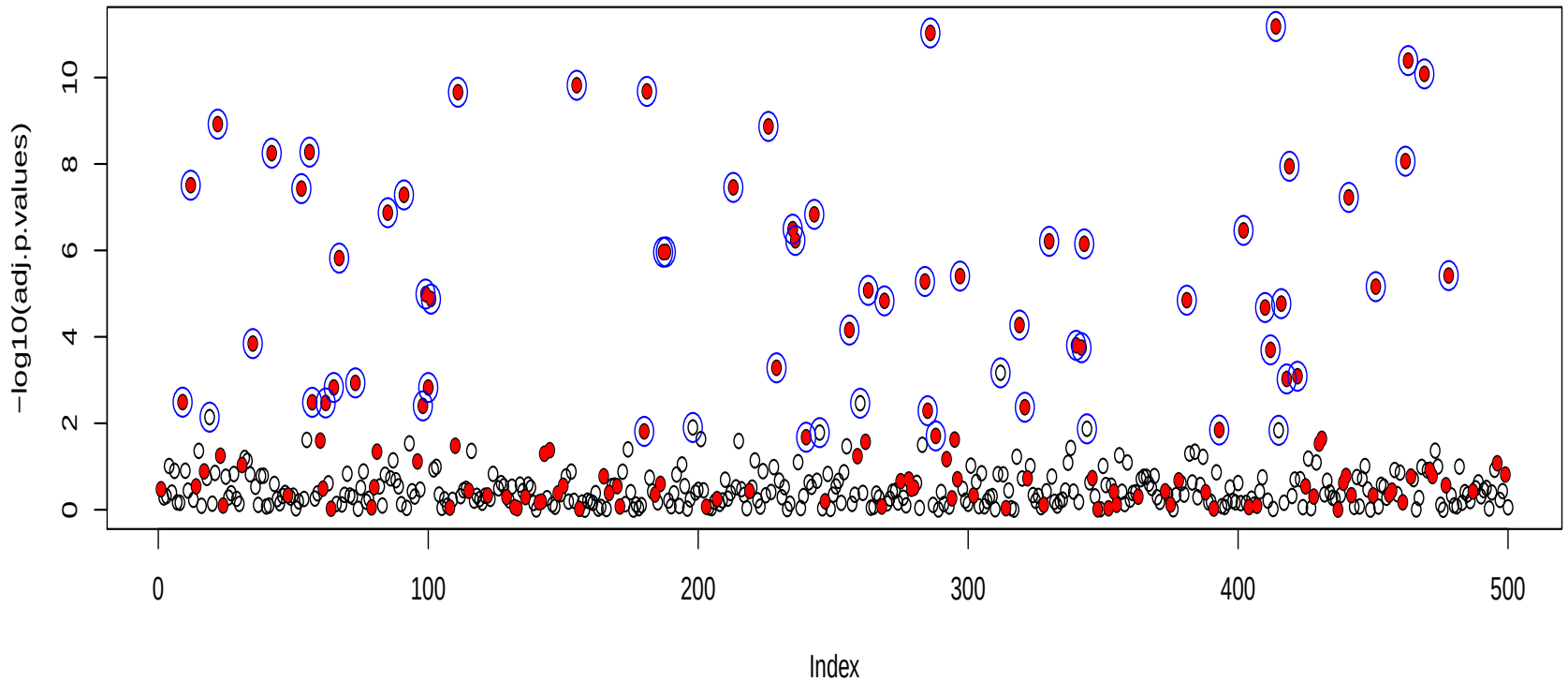
- Inflation/Deflation factor: Tests can be calibrated by rescaling t-scores as follows

```
adj.p.values=2-2*pt(t.scores/lambda,df=n-2,lower=T)
```



Conservative, well-calibrated, and liberal test histograms

Genome scan result (simulated data with "known outliers")



Significance values for outlier tests (log scale, 500 loci). Red points correspond to loci targeted by natural selection. Larger circles correspond to loci in the candidate list.
Expected FDR = .15 – Observed FDR = .10

Summary

- Sparse non-negative matrix factorization/ANLS-AS is a fast and accurate method for estimating individual ancestry coefficients.
- Without loss of accuracy, `sNMF` computed estimates of ancestry coefficients within run-times approximately 10 to 100 times faster than those of the likelihood-based approach `ADMIXTURE`.
- Estimation of ancestral allele frequencies enables genome scan for adaptive allele based on population differentiation statistics.
- Using `R` standard statistical tools can be applied to the control of the FDR when detecting outlier loci.

Practice (40 minutes)

- Instructions, scripts and data sets are available from Olivier Francois' webpage. Add
  `~/LEA/Tutorial-LEA.zip`
  to the webpage `URL`.
- Read install instructions from the `snmf.r` script
- Install the LEA from `R` Bioconductor (Example running Linux)

```
install.packages("LEA0.2.tar.gz",repos=NULL,
type="source")
```

Part 2

- Theory (40min): Ecological association tests.
- Estimates based on latent factor models
- Detection of genetic variation linked to ecological gradients
- Practice (40min): Analysis of data using `R`

Introduction

- Tests to identify associations between loci and environmental or ecological gradients
- Context: *ecological genomics* and health applications
- Correction for population structure, demography and other confounding factors
- Applications to human genomics data sets

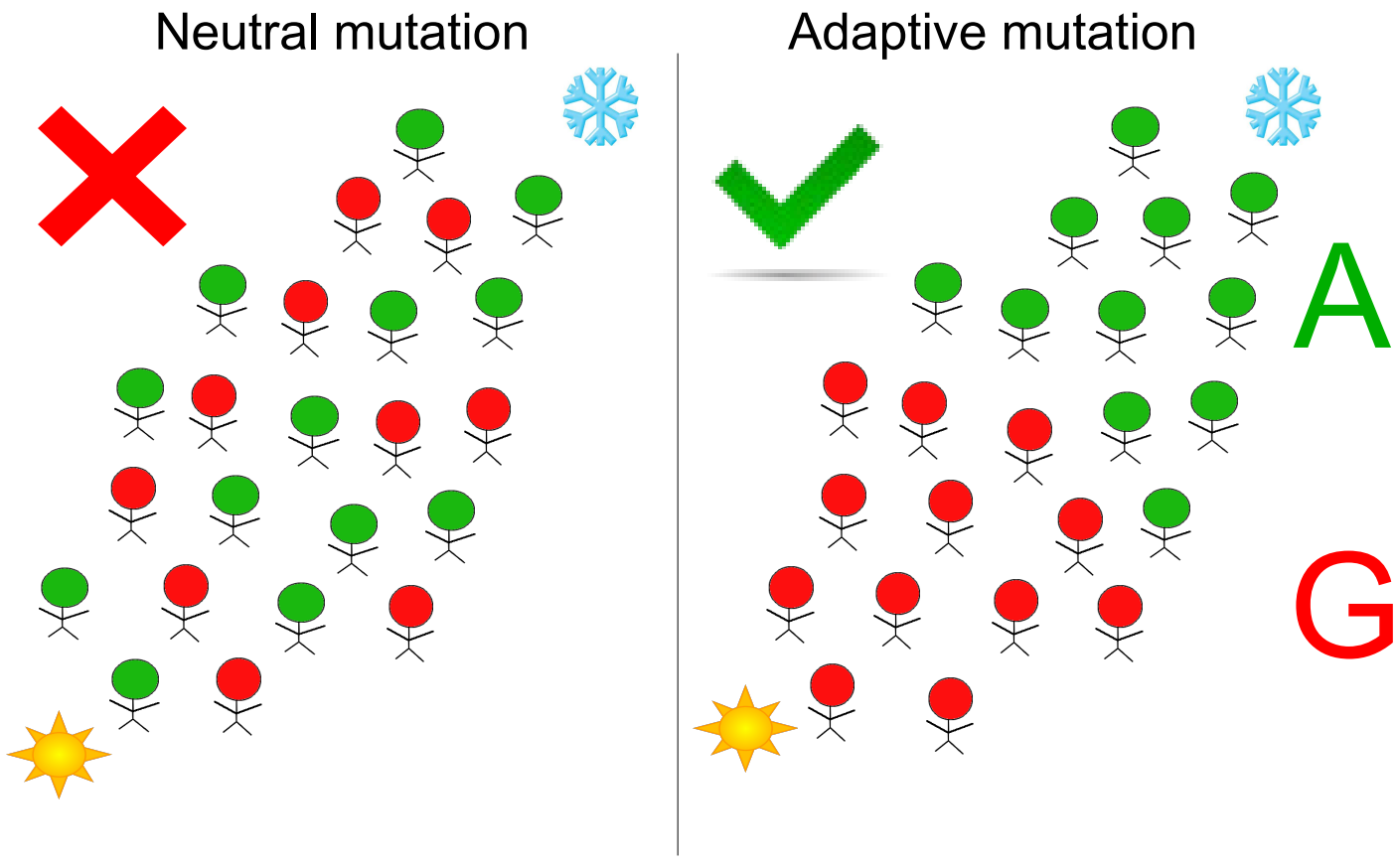
Human ecological genomics

- Objective: Evaluating the effects of interactions between humans and their environments (climate, diet, pathogens) on health and well-being.
- Understanding the genetic origins of chronic diseases (diabetes, asthma, etc).
- Examples:
  - Excess of genes associated with autoimmune diseases such as celiac disease, type 1 diabetes, and multiples sclerosis in response to pathogenic densities.
  - EGLN1 and PPARA confer tolerance to hypoxia and adaptation to high altitude in tibetans (Simonsen et al. 2011).

Local adaptation

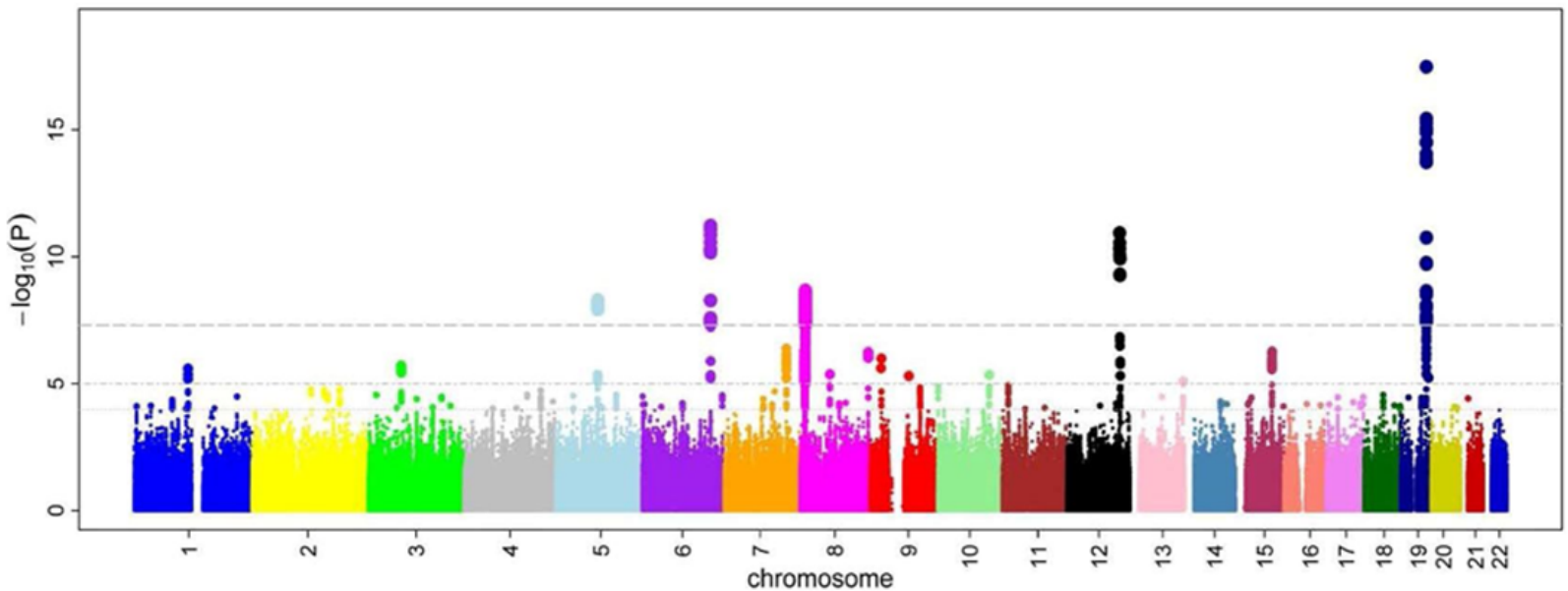
- Local adaptation through natural selection plays a central role in shaping human genetic variation.
- A way to investigate signatures of local adaptation in genomes is to identify allele frequencies that exhibit high correlation with environmental variables (Novembre and Di Rienzo 2009).

Signatures of local adaptation



Genome scans

- Loci with high correlations or (z-scores) are potentially under selection



Caveat

- Inflated number of false positives caused by population structure and isolation by distance patterns.

Model for testing associations between loci and ecological gradients

- We proposed to combine linear regression and factor models (Frichot et al. 2013).
- Latent Factor Mixed model (LFMM):

$$Y_{i\ell} = \mu_\ell + B_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell} \quad (1)$$

- B_ℓ is a d-dimensional vector of regression coefficients.

Rationale

- The matrix $U^T V$ estimates the part of genetic variation that cannot be explained by adaptation to the environment.

$$Y_{i\ell} = \mu_\ell + B_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell}$$

- $U = (U_i)_{i=1,\ldots,K}$ contains K unknown factors (U is of dimension $n \times K$.
- ϵ is the residual error from low-rank approximation ($K \leq n$).
- The number of factors K can be chosen by evaluating the number of *ancestral populations* in ancestry estimation programs.

Background literature on LFMMs

- They can be viewed as *Structural Equation Models* (Sanchez et al. JASA 2005)
- EM algorithms (Sammel and Ryan, Biometrics 1996; An et al. Stat. Med. 2013)
- Bayesian factor regression models (West, Bayesian Stat. 2003; Woodward et al. Biometrics 2014)
- Use of control gene lists (Listgarden et al. PNAS 2010)
- Population genetics (Frichot et al MBE 2013)

Model proposed

- Bayesian Hierarchical model with (weakly) informative prior distributions
- Prior distribution on regression coefficients $B \sim \mathrm{N}(0, \Lambda)$
- Prior distribution on factor $U_i \sim \mathrm{N}(0, \sigma_U^2 I_K)$ and $V_\ell \sim \mathrm{N}(0, I_K)$
- Hyperprior distributions on Λ and σ_U^2 are Inv-Gamma distributions (sparsity).
- Fixed number of latent factors K.

Estimation algorithm

- Stochastic algorithm: Gibbs sampler based on alternating regression models.
- Simple Monte-Carlo estimate for the standard deviation of regression coefficients
- Computation of locus-specific z-scores and p-values.

Human Genome Diversity Project (SNP arrays)

- Worldwide sample of DNA from 1,043 individuals in 52 populations
- The genotypes were generated on Illumina 650K arrays
- Climatic data for each of the 52 population samples from the WorldClim database at 30 arcsecond ($1km^2$) resolution
- These data included 11 bioclimatic variables interpolated from global weather station data collected during a 50 year period (1950-2000), and were summarized with their PC1

Results

- A total of 2,624 (0.4%) SNPs obtained z-scores > 5.

Results

- A total of 2,624 (0.4%) SNPs obtained z-scores > 5.
- A total of 508 (0.08%) SNPs obtained z-scores > 6.

Results

- A total of 2,624 (0.4%) SNPs obtained z-scores > 5.
- A total of 508 (0.08%) SNPs obtained z-scores > 6.
- A total of 65 (0.007%) SNPs obtained z-scores > 7.

Results

- Among loci with z-scores greater than 5, 28 were GWAS-SNPs with known disease or trait association.
- Among the 65 SNPs with z-scores greater than 7, 31 were intra-genic SNPs.

GWAS-SNPs associated with environmental predictors.

Gene	Trait association	$-\log_{10} P$-value
OCA2/HERC2	Eye and hair color, pigmentation	9.15
DHCR7	Vitamin D levels	7.78
SLC45A2	Hair color	6.90
Intergenic MUC7	Alcoholism	8.91
ZMIZ1	Crohn's disease	8.77
KLK3	Prostate Cancer	8.61
ICOSLG	Celiac disease	7.02
HLA-DRA	Systemic sclerosis	6.97
NCAPG-LCORL	Height	9.43
BOK	Brain structure and development	9.43

Genic SNPs associated with environmental predictors.

Gene	Annotation (dbSNPs)	$-\log_{10} p$-value
EPHB4	Heart morphogenesis and angiogenesis	16.54
NRG1	Nervous system development, cell proliferation	16.21
RBM19	Regulation of embryonic development	15.98
EYA2	Eye development and DNA repair	15.9
POLA1	Mitotic cell cycle and cell proliferation	15.87

Summary

- Fast algorithms based on low rank approximations (ML and Gibbs Sampler algorithms)
- Separate neutral from adaptive variation
- Many new adaptive SNPs with functions associated to multicellular organ development

Using LFMM in practice

- Narrow the range of the number of factors, K, by investigating plausible values using `sNMF`
- Multiple runs of LFMM increase the power to detect significant associations.

```
project=lfmm("example.lfmm","gradient.env",K=8,
it=10000,b=5000,rep=5)
```

- Get the z-scores and combine them using the median value

```
zs=z.scores(project,K=8)
zs.combined=apply(zs,MARGIN=1,median)
```

Test calibration using inflation factor correction

- For each locus ℓ, p-values are calibrated using the following formula

$$p_\ell = \text{prob}(\chi_1^2 > z_\ell^2/\lambda) \quad \ell = 1, \ldots, L.$$

- Suggested value for λ close to the genomic inflation factor

$$\lambda = \text{median}(z^2)/0.456$$

($\lambda \approx 1$ indicates that the histogram of p-values is flat).

- Visual inspection of histograms or QQ-plots may help as well.

Multiple testing issue and candidate list

- Candidate list of genes associated with ecological gradients are provided by the Benjamini-Hochberg FDR control algorithm.

Practice (40 minutes)

- Instructions, scripts and data sets are available from Olivier Francois' webpage. Add
  `~/LEA/Tutorial-LEA.zip`
  to the webpage `URL`.
- Read instructions from the `lfmm.r` script

Acknowledgments

- Eric Frichot
- Sean Schoville, François Mathieu, Théo Trouillon, Guillaume Bouchard
- This work received support from the “Région Rhône-Alpes”, Xerox Research, National Science Foundation USA, and Ensimag-Grenoble INP.