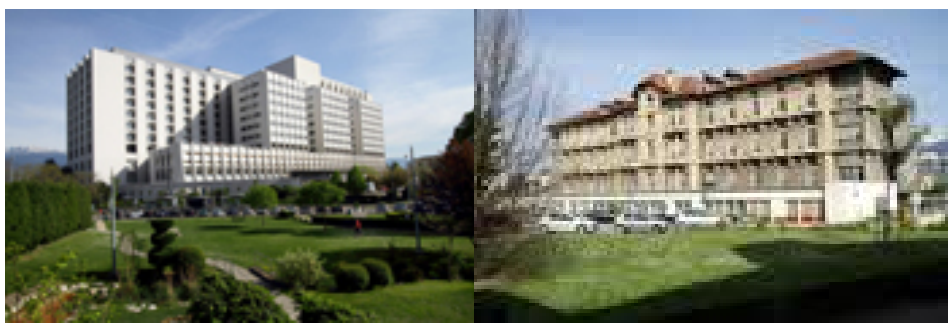
Latent variable models for genome scans for selection

Olivier Francois
Computational and Mathematical Biology Lab
TIMC-IMAG – Université Grenoble-Alpes, France

Paris, June 18, 2015

Research at TIMC-IMAG, Université Grenoble-Alpes

- Department of Biomedical Engineering (promotes interactions between MDs and scientists).
- BCM lab: ten members (3 mathematicians, 4 bioinformaticians, 3 clinicians)
- Focus on *Big Data* analysis for health and well-being applications



Outline

- Context: *ecological genomics* and health applications
- Latent factor models for population genetic data
- New tests to identify associations between loci and environmental or ecological gradients
- Correction for population structure, demography and other confounding factors
- Applications to (large) human genomic data sets

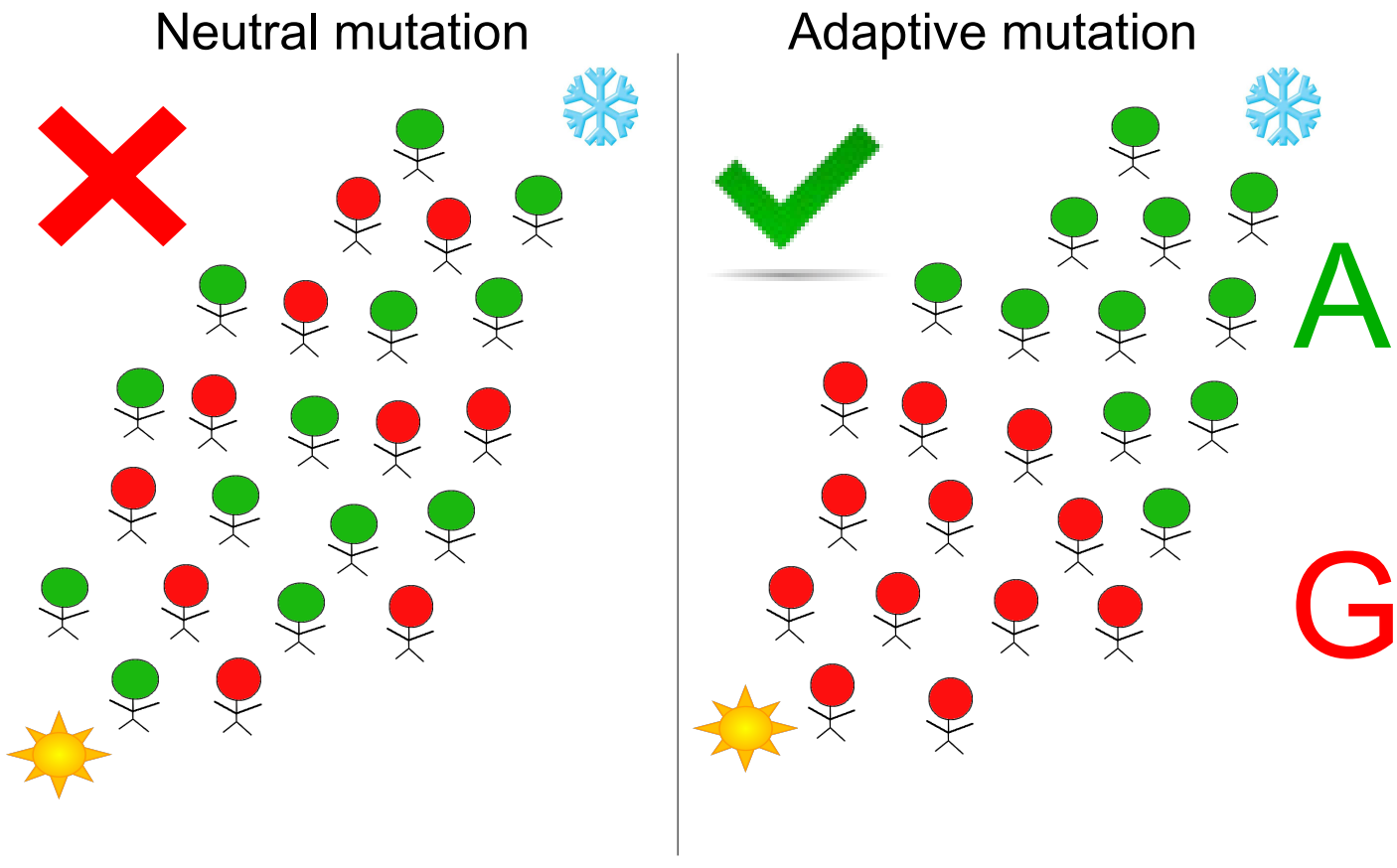
Human ecological genomics

- 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 (Fumagalli et al. 2012).
  - *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



Signatures of local adaptation

- Detection alleles correlated with ecological gradients can be useful when many beneficial alleles have weak phenotypic effects or in case of selection on standing variation (soft sweeps, Pritchard *et al.* 2010).
- Unobserved factors such as geographic population structure, genetic background, sequencing platforms, etc, can confound interpretation of these associations.

Basic principles

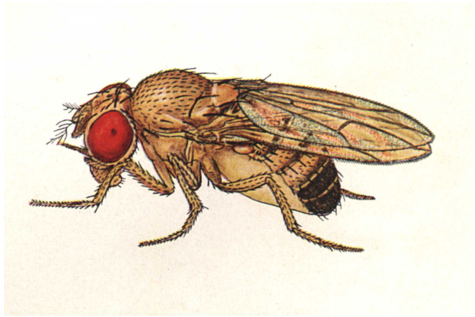
- For allele frequencies ($Y_{i\ell}$) and a set of environmental variables (X_i), standard tests are based on regression models

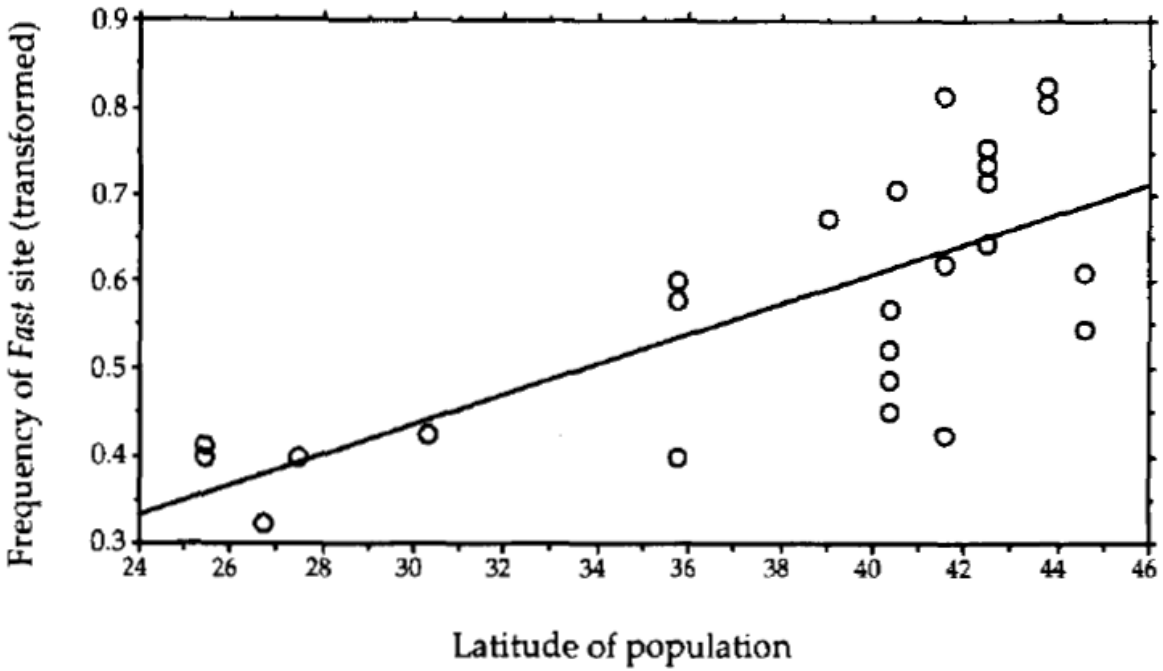
$$Y_{i\ell} = \mu_\ell + B_\ell^T X_i + \epsilon_{i\ell}\,, \qquad i = 1, \ldots, n,$$

  where $Y_{i\ell}$ is the allele frequency at locus ℓ in population or individual i.

- B_ℓ represents environmental effects, ϵ_ℓ are uncorrelated residuals.

Example of evidence for selection at the Adh locus





frequency of *Adh-F* (square-root, arcsine transformed) on the latitude of each sample

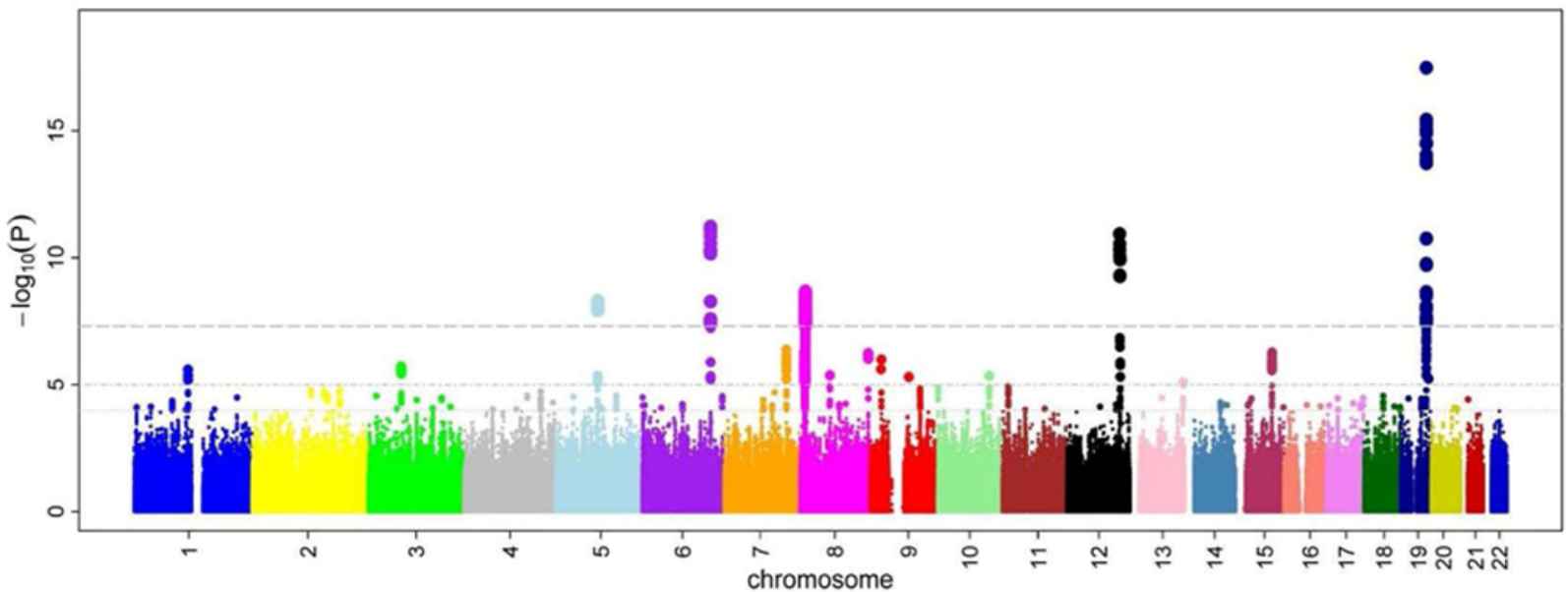
Databases

- Human genomics projects: HGPD (Li et al. Science 2008), 1000 Genomes (Nature 2012)
- Large SNP[1] genotypic matrices: $(Y_{i\ell}) \sim$ 1–5 Giga entries ($Y_{i\ell}$ = 0,1,2).
- Environmental databases: Worldclim, WHO
- Bioinformatic data bases: dbSNP, etc

[1] SNP = Single Nucleotide Polymorphism (DNA locus that exhibits variation among human populations)

Genome scans

- Loci with high Z-scores are potentially under selection

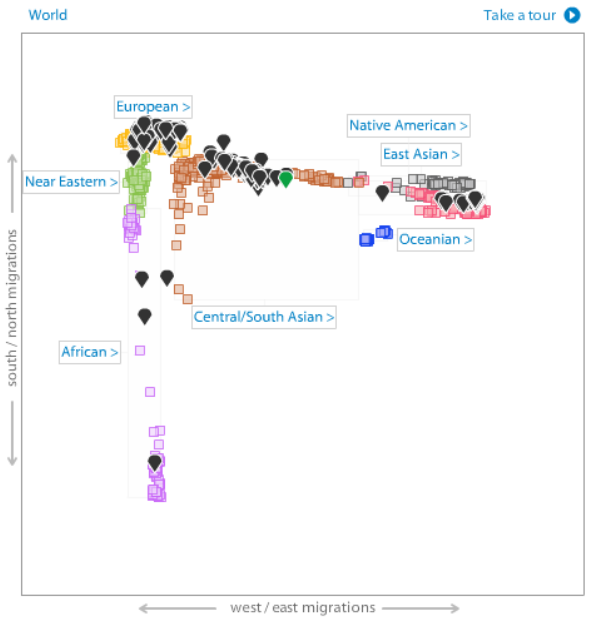


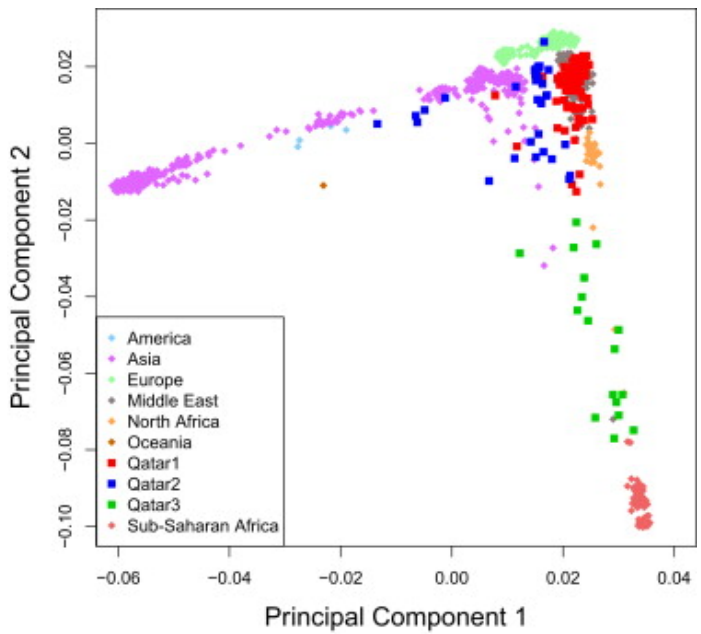
Caveat

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

Inference of population structure using PCA

- Principal Component Analysis (PCA) is often used as an ancestry estimation method.





PCA and factor analysis

- PCA is related to factor analysis via maximum likelihood estimates (Tipping and Bishop 1999; Engelhardt and Stephens 2010)

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

  where $Y_{i\ell}$ is the allele frequency at locus ℓ in population or individual i (0, 1/2, 1).
- U_i and V_ℓ are independent Gaussian vectors with K dimensions corresponding to PC scores and loadings ($\sigma_V^2 = 1$).
- $\epsilon_{i\ell}$ are uncorrelated residuals corresponding to dimensions greater than K.

Model for testing associations between loci and ecological gradients

- A combination of 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.

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

- 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 clusters or *ancestral populations* in ancestry estimation programs.

Background literature on LFMMs

- *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)

ML estimation in LFMMs

- The log-likelihood for LFMM parameters is defined as follows :

$$-\ell(B, U, V, \sigma^2) = \frac{1}{2\sigma^2}\|Y - XB^T - UV^T\|^2 + \frac{nL}{2}\log(2\pi\sigma^2)$$

- Non-identifiablity: ML estimates satisfy

$$B^T = (X^T X)^{-1} X^T Y - CV^T$$

where C is any $d \times K$ matrix and V can be obtained by singular value decomposition

$$U\Sigma V^T = Y - X(X^T X)^{-1} X^T Y$$

.

Using regularized estimates (Ridge regresssion)

- Regularized least-squares estimates of LFMM parameters (B, U, V) minimize

$$\frac{1}{\sigma^2}\|Y - XB^T - UV^T\|^2 + \mathrm{tr}(B^T\Lambda^{-1}B)$$

where $\Lambda > 0$ is a $d \times d$ diagonal matrix of regularization coefficients.

Using regularized estimates (Ridge regresssion)

- Identifiability: The estimator for the regression coefficients is given by

$$B^T = (X^T X)^{-1} X^T (Y - UV^T)$$

where

$$UV^T = C^{-1}\mathrm{svd}(CY)$$

and C is the Cholesky factor in the decomposition of the *oblique* projection on $\mathrm{vect}(X)^{\perp}$

$$CC^T = I - X(X^T X + \Lambda^{-1})^{-1} X^T.$$

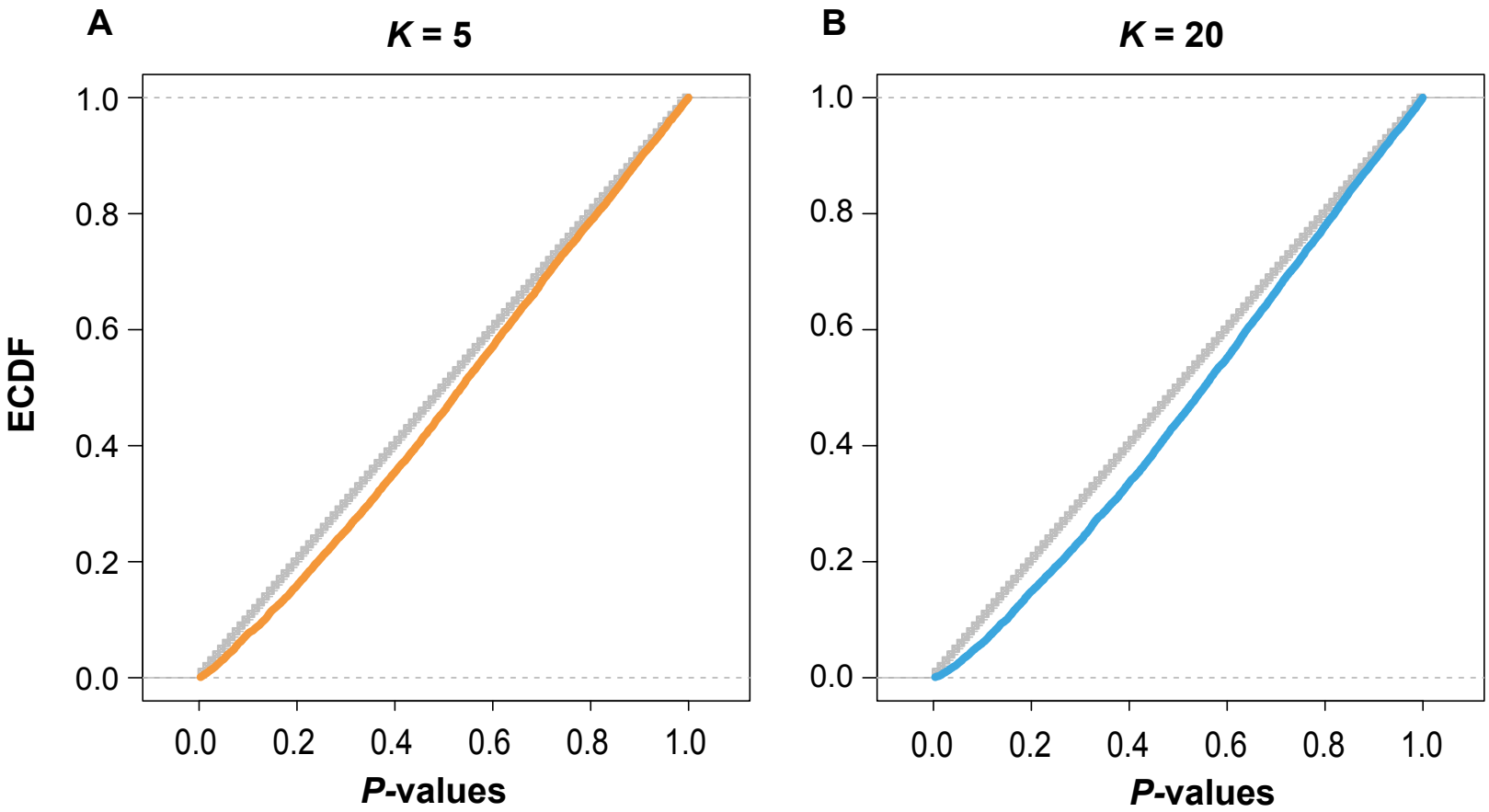
Model proposed

- 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 regressions).
- Multi-threated version $\Longrightarrow$ Acceptable run-times.
- Simple Monte-Carlo estimate for the standard deviation of regression coefficients
- Computation of locus-specific z-scores and p-values.

Distribution of *p*-values under generative simulation models



Comparisons under neutral "isolation-by-distance" models

- Isolation-by-distance models are classical population genetic models in which individuals close to each other are more related than individuals far apart (Kimura and Weiss, 1964).
- We considered equilibrium and non-equilibrium models. In non-equilibrium models, populations expanded in the northward direction from a single source population.

Comparisons under neutral "isolation-by-distance" models

- The choice of K was based on the computation of the genomic inflation factor

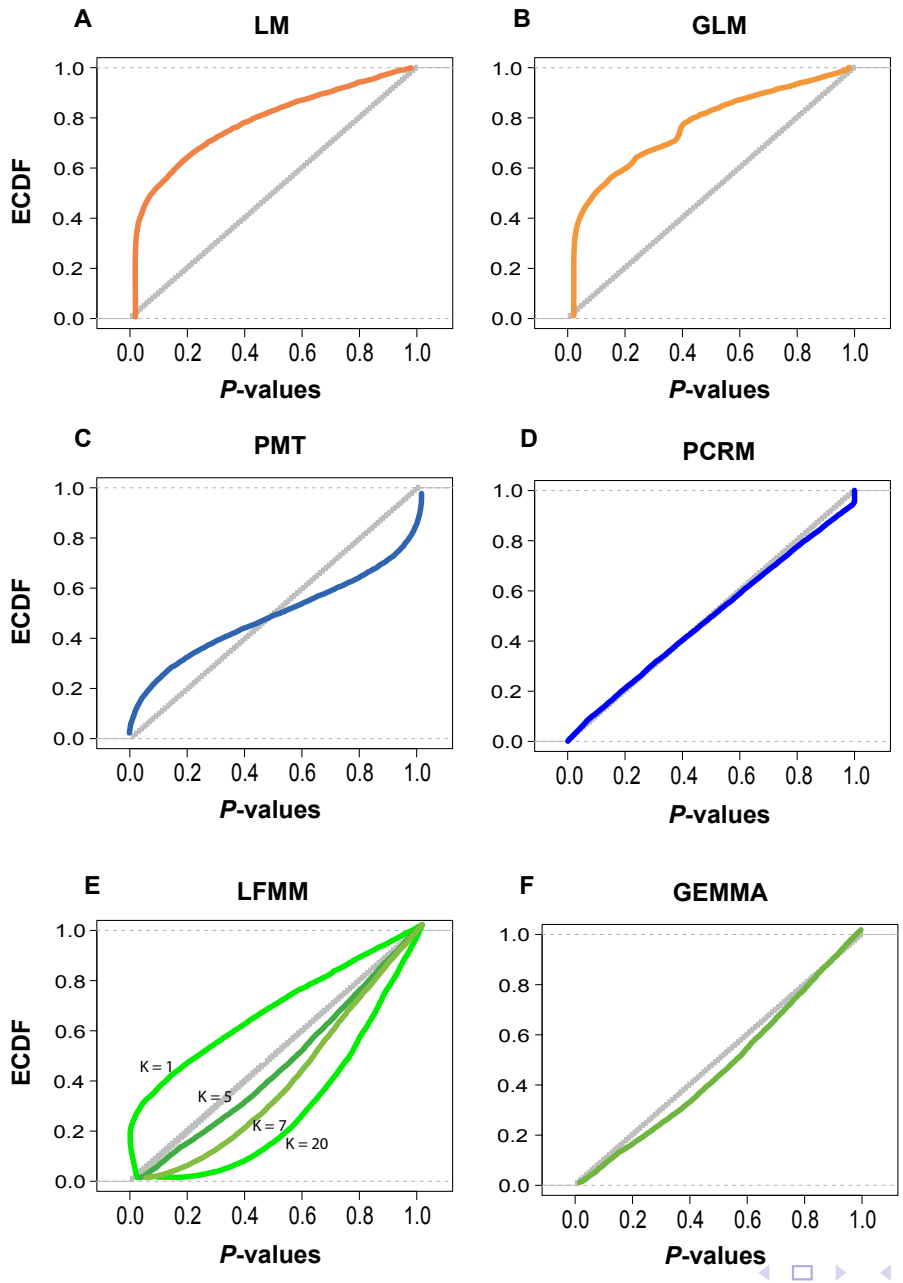
$$\lambda = \text{median}(z^2)/0.456$$

($\lambda \approx 1$ indicates that the p-values are correctly calibrated).

- Then p-values are computed as

$$p_\ell = p(\chi_1^2 > z_\ell^2/\lambda)$$

Distribution of *P*-values under neutral "isolation-by-distance" models



Power study – strong selection gradient, parallel to the main axis of variation.

- Rates of false negative (FN) and false positive (FP) association for tests based on linear models (LM), principal component regression (PCRM), standard linear mixed models (GEMMA), Partial Mantel correlations (PMT) and LFM models (LFMM).

FN (FP)	LM	GLM	PCRM	GEMMA	PMT	LFMM
type I error:						
$-\log_{10}\alpha = 3$	0% (33%)	0% (24%)	100 % (3%)	100 % (2%)	99% (6.8%)	4% (5%)
$-\log_{10}\alpha = 4$	0% (27%)	0% (19%)	100 % (0%)	100 % (0%)	100% (3.4%)	14% (3%)

Alternative ways of correcting for population structure

- For allele frequencies ($Y_{i\ell}$) and a set of environmental variables (X_i), the test is based on a regression model

$$Y = \mu + B^T X + \eta\,,$$

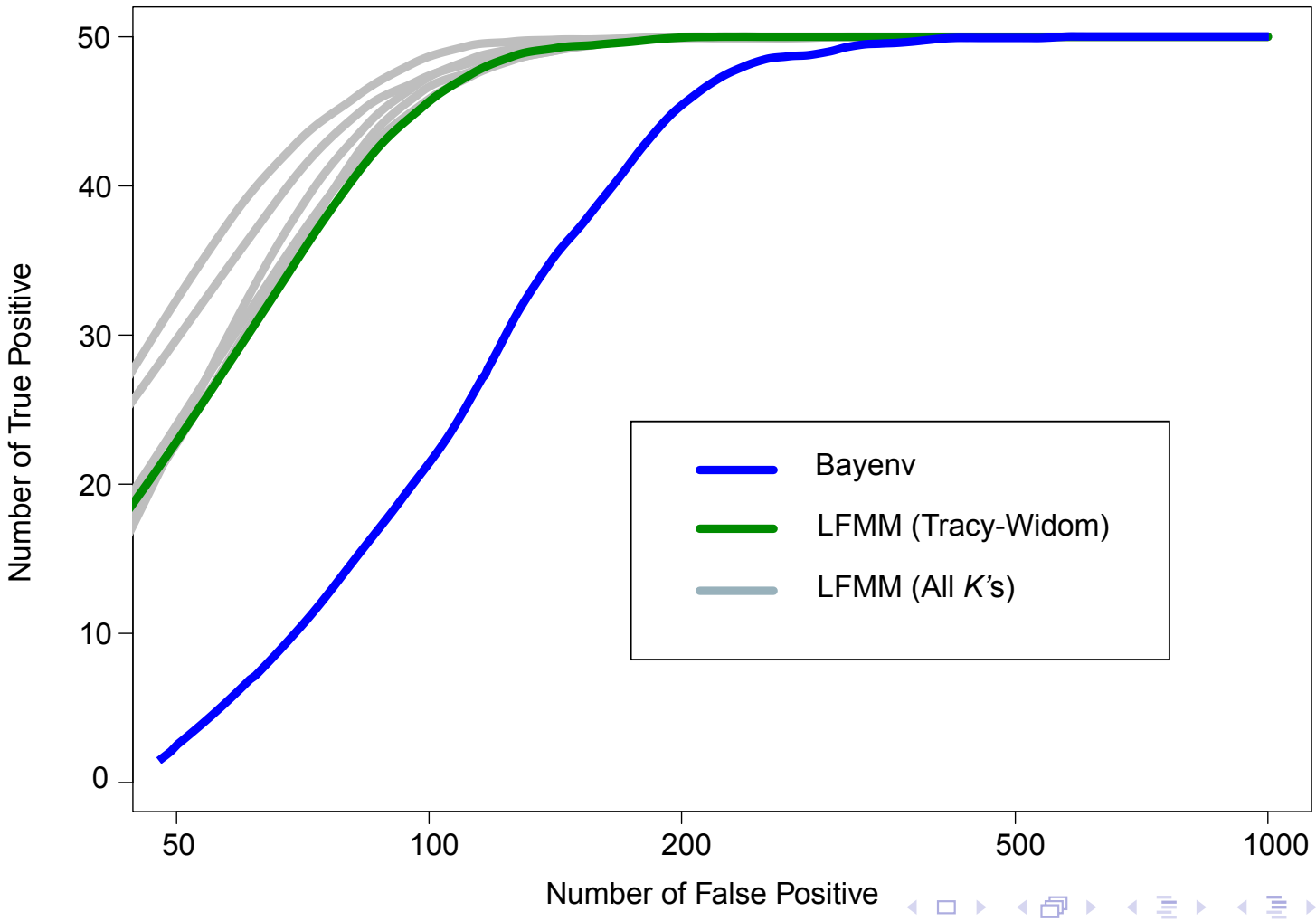
- Environmental variables are fixed effects and population structure is introduced as random effects (Hancock *et al.* 2008; Coop *et al.* 2010).

Alternative ways of correcting for population structure

- In the `Bayenv` model (Coop *et al.* 2010), the covariance matrix of the random effects, η, is set to the empirical covariance matrix.
- This makes the implicit assumption that the covariance structure is not influenced by local adaptation.

Comparison with Bayenv

- Simulation context: Neutral population structure is generated by an isolation by distance mechanism. Association with an environmental gradient is generated at a few loci (5%).



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
- Soft sweeps were frequent during human evolution?

LEA

LEA: an R package for Landscape and Ecological Association Studies

Bioconductor version: Release (3.1)

LEA is an R package dedicated to landscape genomics and ecological association tests. LEA can run analyses of population structure and genome scans for local adaptation. It includes statistical methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (snmf, pca); and identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures (lfmm), and controlling the false discovery rate. LEA is mainly based on optimized C programs that can scale with the dimension of very large data sets.

Author: Eric Frichot <eric.frichot at gmail.com>, Olivier Francois <olivier.francois at imag.fr>

Acknowledgments

- Eric Frichot
- Sean D. Schoville
- Guillaume Bouchard
- This work received support from “La région Rhône-Alpes” and from National Science Foundation USA.