A unified framework for inferring population genetic structure and gene-environment associations

Olivier.Francois@imag.fr

Computational and Mathematical Biology group
University Joseph Fourier Grenoble, France

MCEB 2012

Outline

- New tests to identify associations between loci and environmental or ecological gradients
- Correction for population structure and demography
- New: Latent factor mixed models
- Application to (large) SNP data sets

No trees!

- Gene trees are unobserved data
- No reconstruction is attempted
- Trees translate into covariance structure for the data.

Signatures of local adaptation

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

Signatures of local adaptation

- This method is useful when many beneficial alleles have weak phenotypic effects or in case of soft sweeps (Pritchard *et al.* 2010).
- Population structure can confound interpretation of these associations.

Basic principles

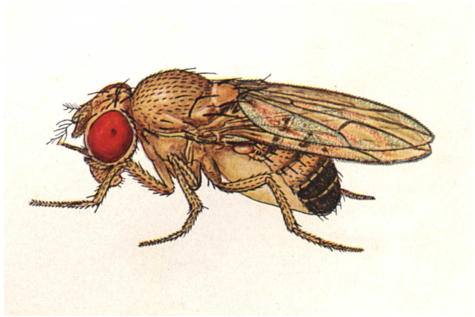
- For allele frequencies ($Y_{i\ell}$) and a set of environmental variables (X_i), standard tests are based on regression models

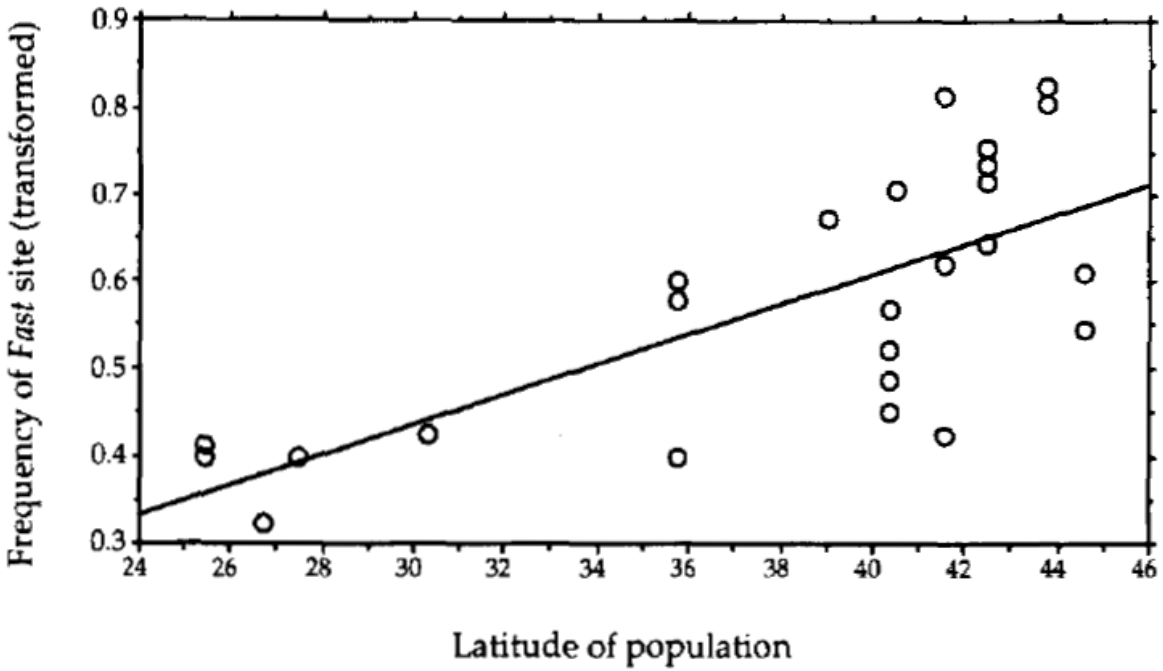
$$Y_{i\ell} = \mu_\ell + B_\ell^T X_i + \epsilon_{i\ell}\,, \qquad i = 1, \ldots, n,$$

  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

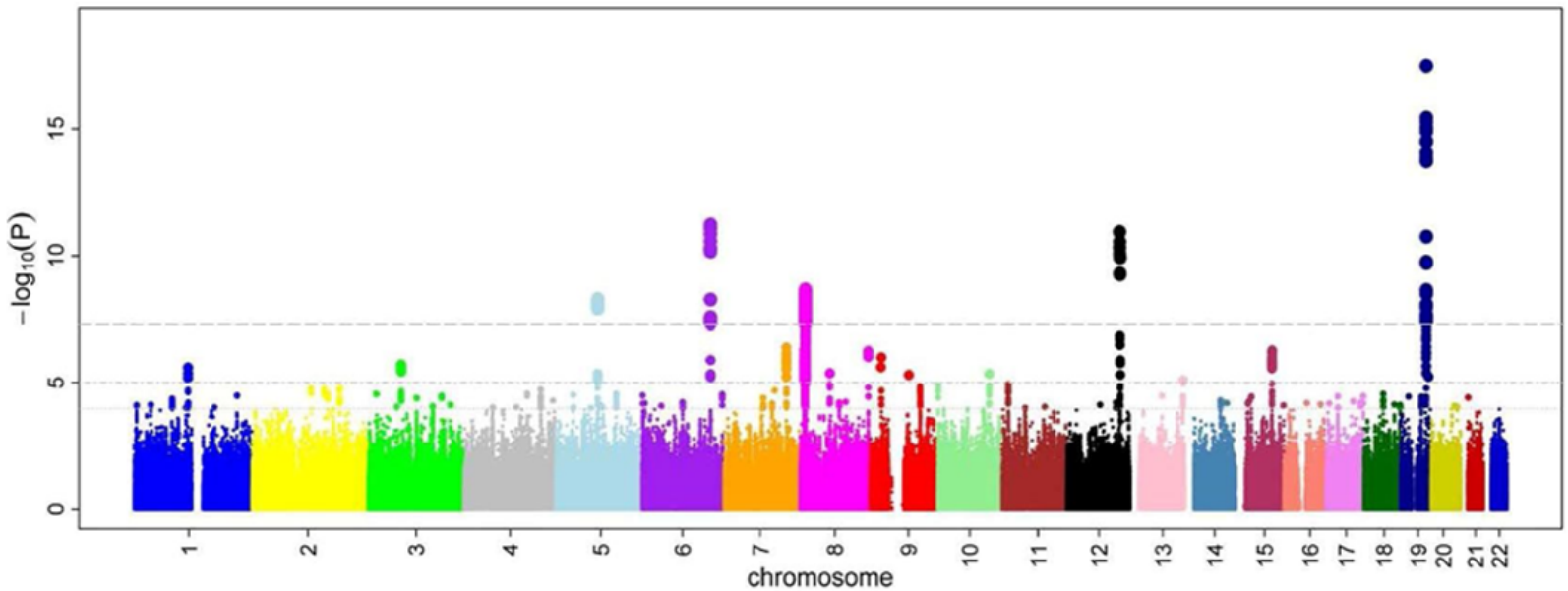




frequency of *Adh-F* (square-root, arcsine transformed) on the latitude of each sample

Genome scans

- Loci with high Z-scores are potentially under selection



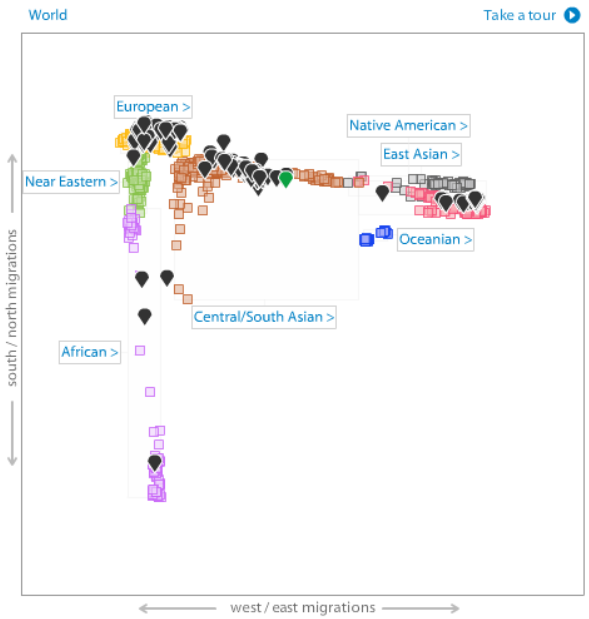
.

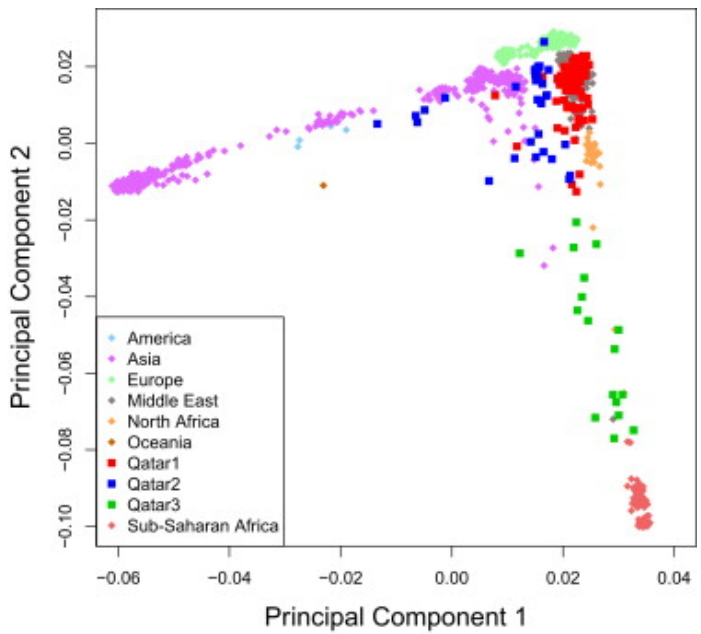
Caveat

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

Inference of population structure

- Principal Component Analysis (PCA) is commonly used to describe population structure





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.
- 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 neglected dimensions ($K \leq n$).

Model for testing associations between loci and environmental gradients

- Latent Factor Mixed model (LFMM):

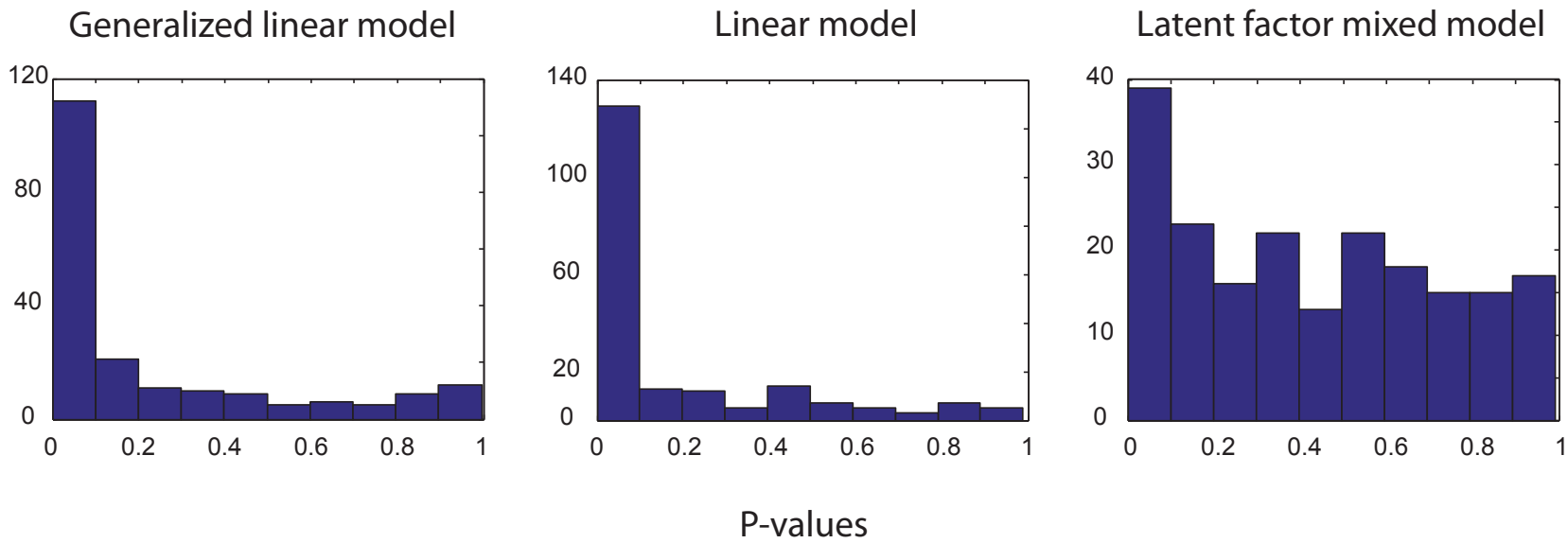
$$Y_{i\ell} = \mu_\ell + B_\ell^T X_i + U_i^T V_\ell + \epsilon_{i\ell} \quad (1)$$

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

Model for testing associations between loci and environmental gradients

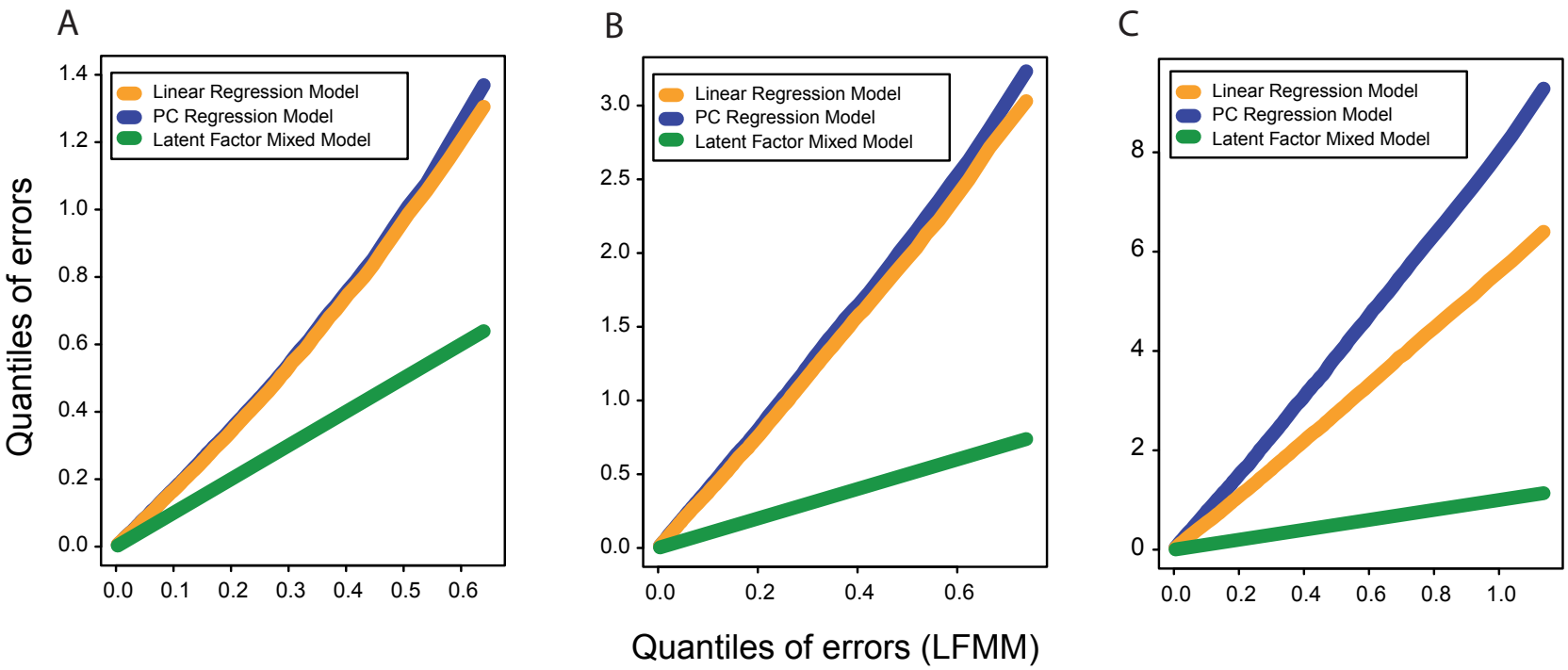
- Rationale: The matrix $U^T V$ estimates the part of genetic variation that cannot be explained by adaptation to the environment.
- ϵ is the residual error from low-rank approximation ($K \leq n$).
- Fast algorithms for ML or Bayesian inference.

Comparisons under neutral "isolation by distance" models



Comparisons with the standard linear and PC regression models

- Relative statistical errors decrease with the number (K) of hidden factors



A) $K = 2$, B) $K = 20$, C) $K = 100$

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

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

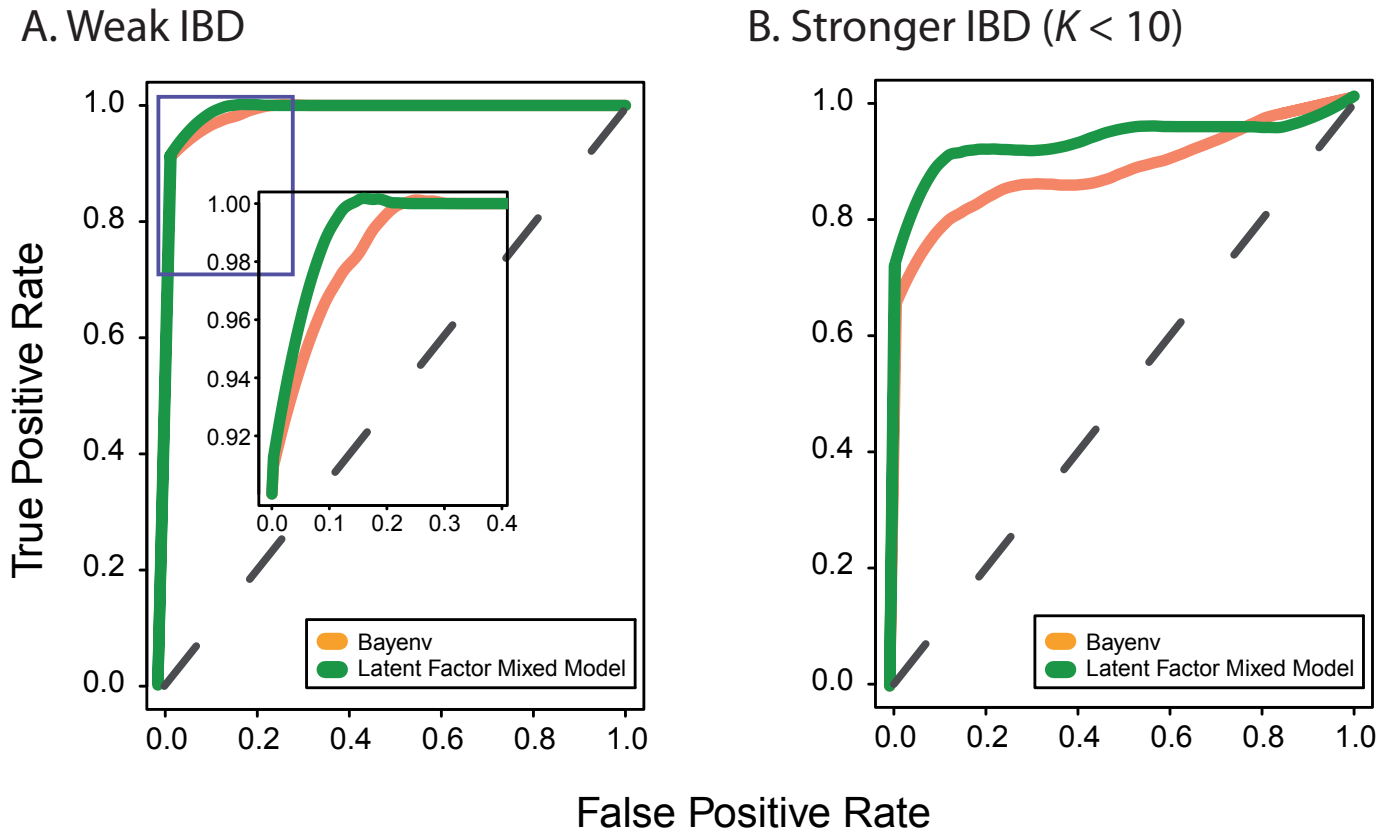
- Environmental variables are fixed effects and population structure is introduced as random effects (Hancock *et al.* 2008; Coop *et al.* 2010).

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
- We used $K = 50$ latent factors

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?

Acknowledgments

- Eric Frichot
- Sean Schoville
- Guillaume Bouchard
- This work received support from “région Rhône-Alpes”, NSF, Xerox Research and Grenoble INP.
- We seek highly motivated applicants to EU PhD fellowship in the lab!