Outlier detection for high-dimensional data: application to population genomics

michael.blum@imag.fr
Université Grenoble Alpes

Nicolas Duforet-Frebourg,
former PhD student

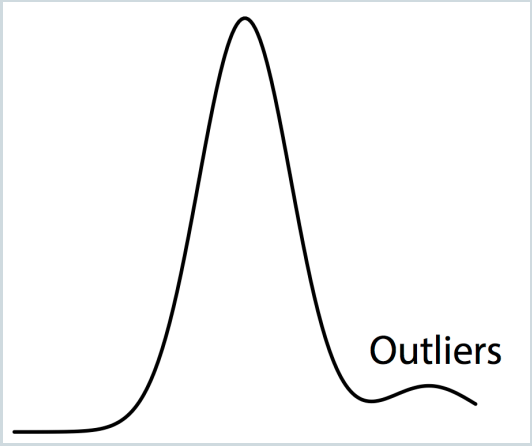
Keurcien Luu,
PhD student

How to map genes involved in natural selection using outlier detection?

- Genome-wide patterns are influenced by neutral processes.

  Migration, admixture, expansion

- Genes involved in natural selection are outliers.

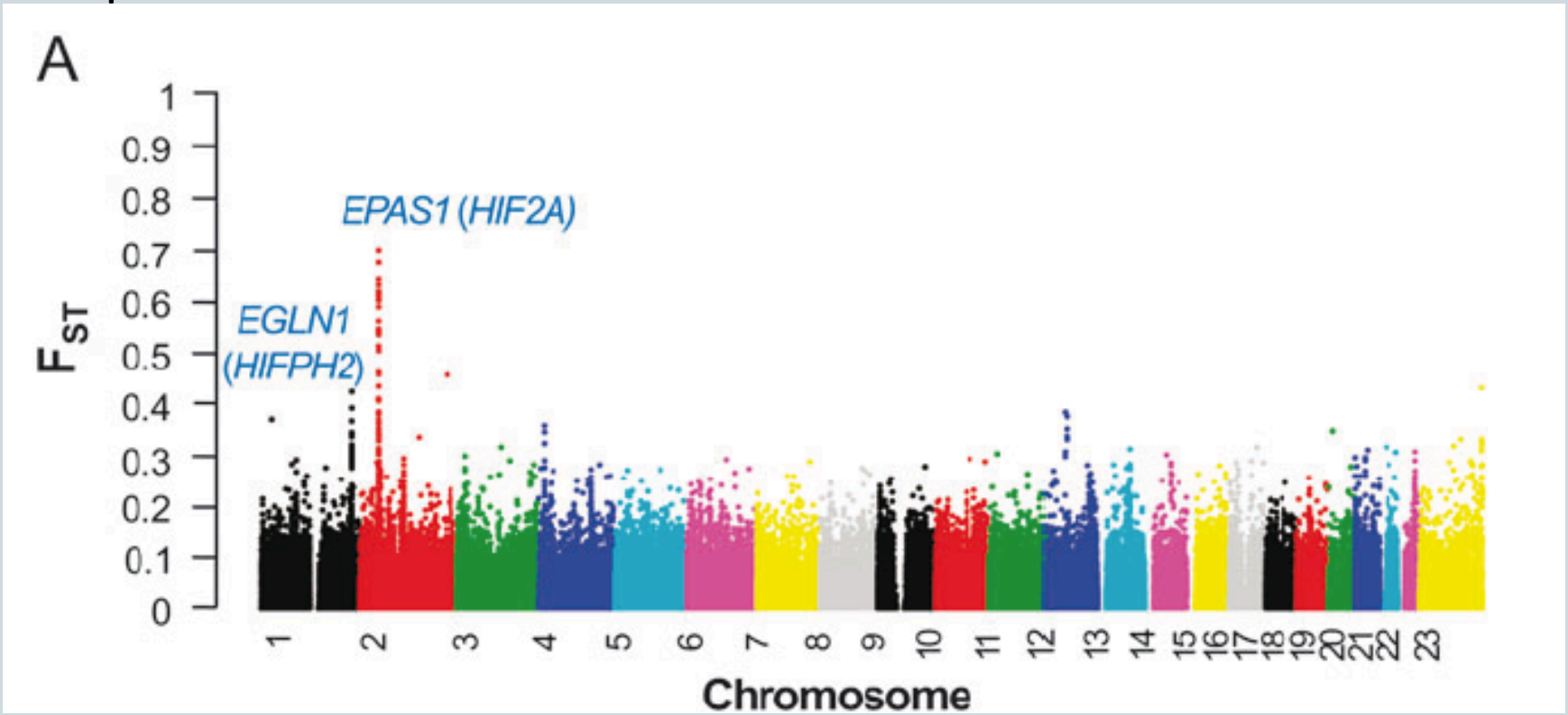


An example of natural selection in humans

- Tibetan populations got adapted to their high-altitude and low-oxygen environment thanks to increased respiratory rate and increased blood flow.
- These traits are transmitted from generation to generation.
- Tibetan plateau has been inhabited since ~ 3,000-20,000 years.

Mapping genes involved in natural selecion using outlier detection

Adaptation to altitude



Xu et al. MBE 2011

Single Nucleotide Polymorphism (SNP)

Data matrix Y (red and black pops)

	Locus 1 (Y_1)	Locus 2 (Y_2)	Locus 3 (Y_3)
Indiv 1	1	0	1
Indiv 2	0	2	1
Indiv 3	0	0	0
Indiv 4	0	1	0
Indiv 5	1	1	0

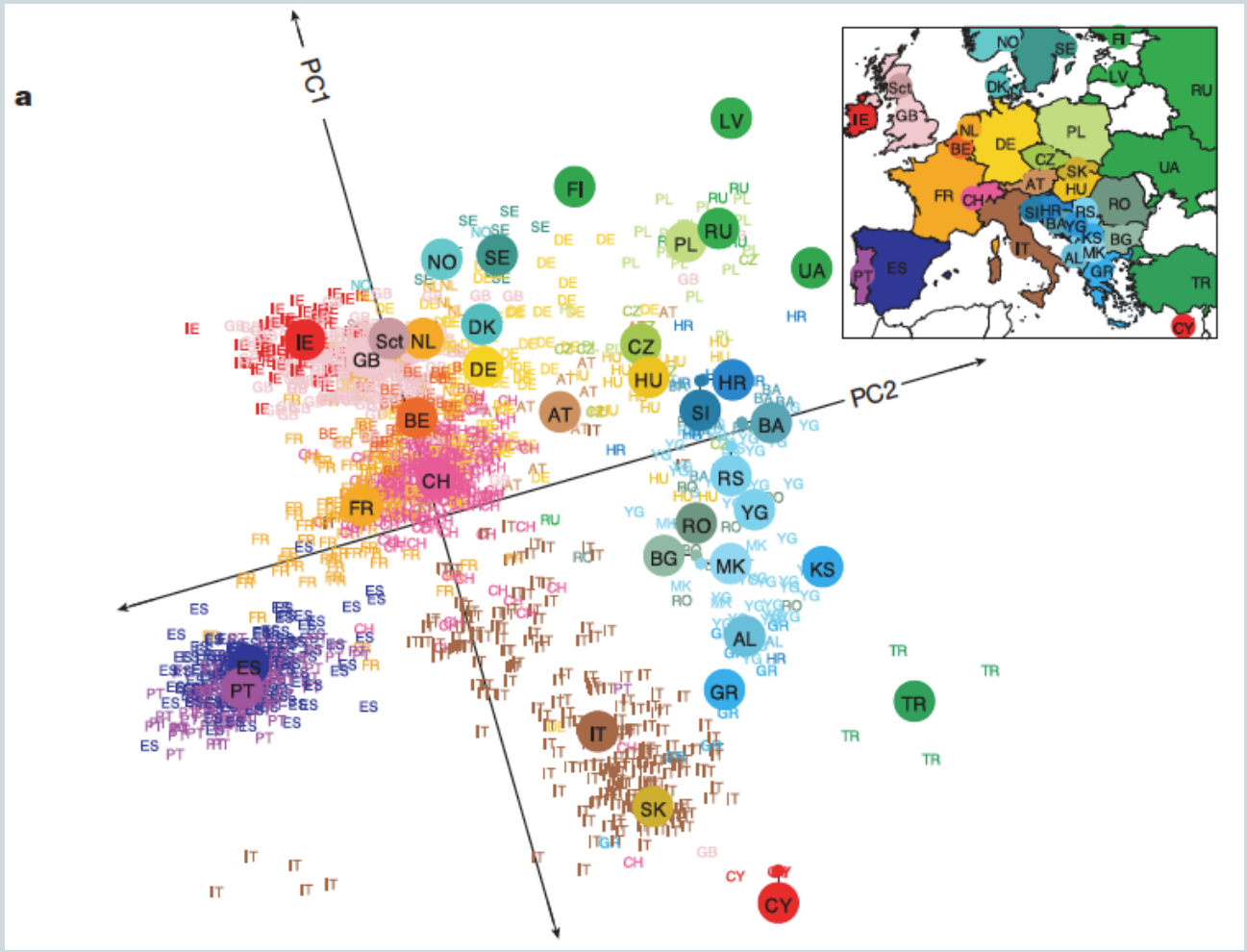
F_{ST}: A pervasive statistic to perform selection scan

$$G = \alpha + \beta P + \varepsilon,$$

where G denotes allele counts (0,1, or 2), and P denotes the population to which the individual is assigned (P=0 or 1 if there are 2 populations).

F_{ST} can be viewed as the proportion of variance (R^2) of G explained by P.

Sometimes, there are not 2 populations
Continuous population structure in Europe



Novembre et al. Nature 2008

Regression framework when population structure is continuous

$$G = \alpha + \beta_1 PC_1 + \cdots \beta_K PC_K + \varepsilon,$$

where G denotes allele counts (0,1, or 2) and PC_1, …,PC_k measure population structure and denotes the score of the k^{th} *PC.*

Please propose a test statistic to look for outliers (small survey among statisticians).

Regression framework when population structure is continuous

We consider the vector of z-scores (z_1,…,z_K) to detect outlier loci. The test statistic is a robust Mahalanobis distance.

$$D^2 = (z - \bar{z})^T \Sigma^{-1} (z - \bar{z}),$$

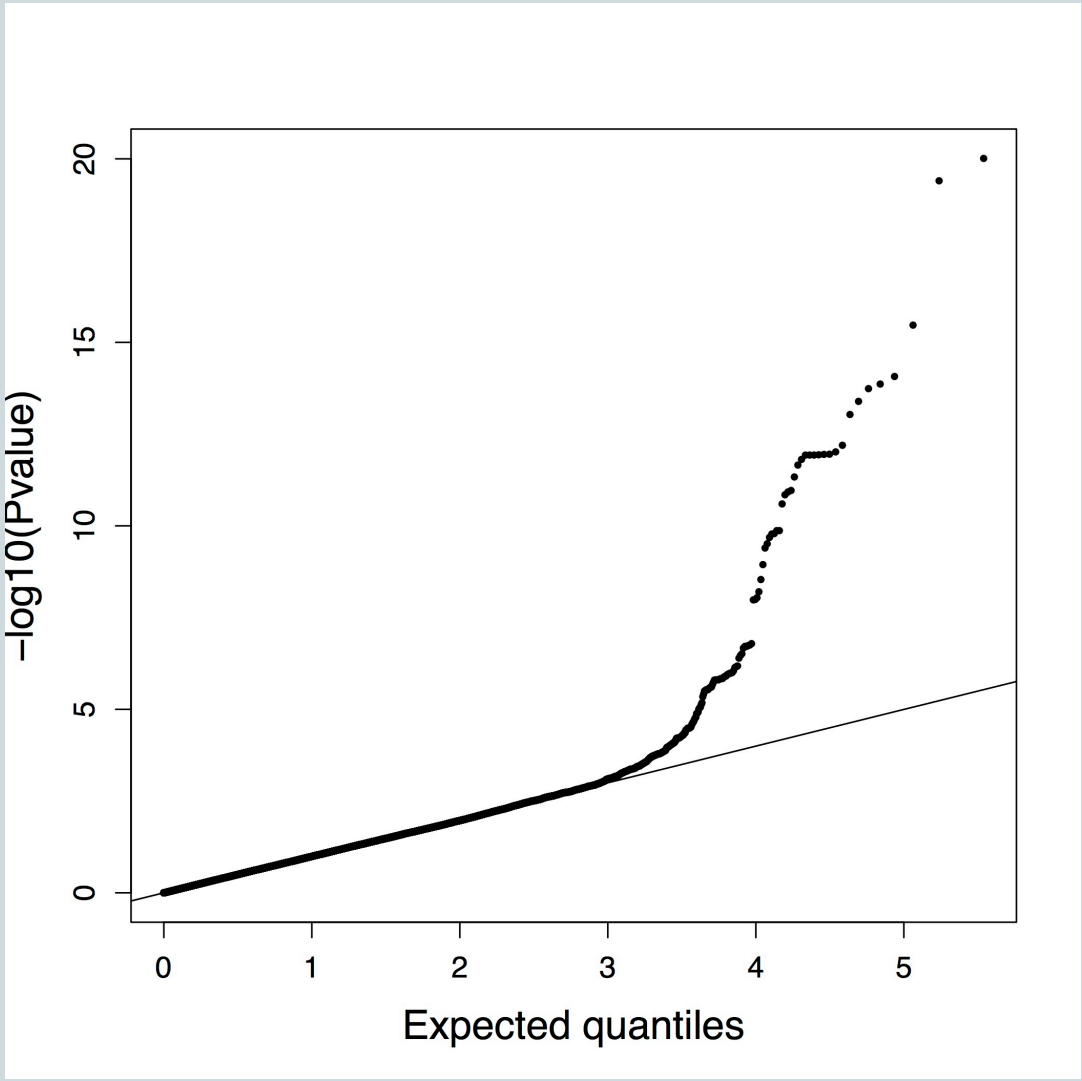
where $\bar{z}$ and Σ are robust estimates of the mean and covariance matrices of z-scores (Orthogonalized Gnanadesikan-Kettenring method, *Maronna and Zamar Technometrics 2012*).

Null distribution of the test statistic

If the vectors of z-scores were truly multivariate Gaussian, Mahalanobis distances should be chi-squared with K degrees of freedom.

In practice, Mahalanobis distances should be divided by a parameter to estimate (genomic inflation factor, empirical null distribution) to be approximated by a chi-square distribution with K degrees of freedom (*Maronna and Zamar Technometrics 2012*).

Qqplot
European data



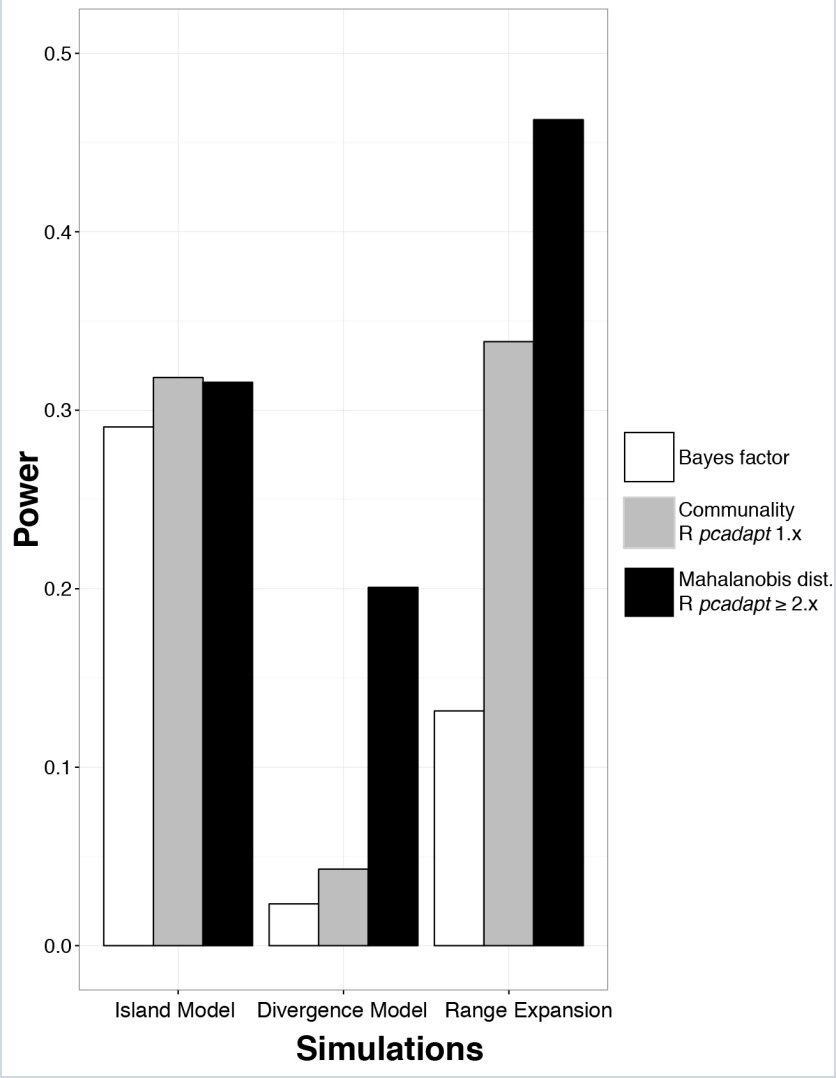
Simulation results

Power comparison of 3 statistics

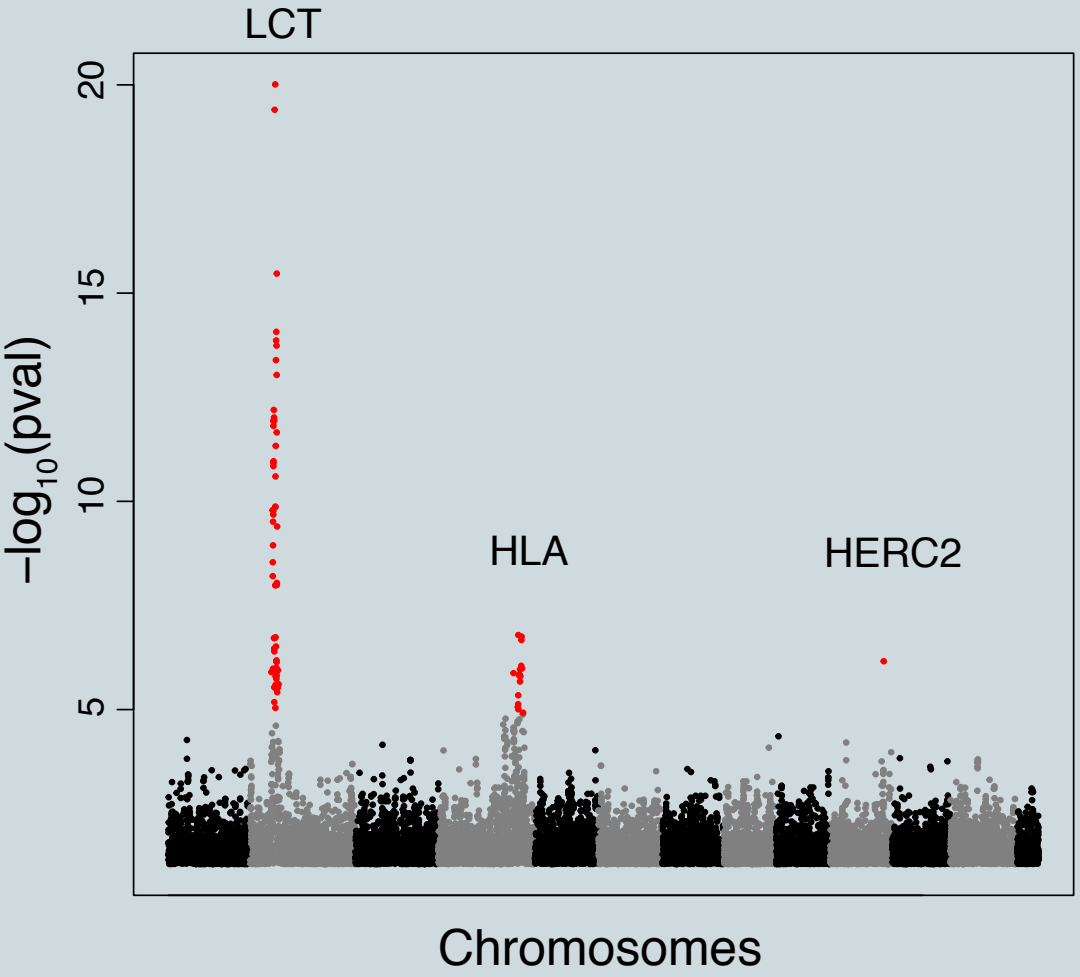
Stat 1: **Bayes factors** computed in a Bayesian factor model (*Duforet-Frebourg et al. MBE 2014*).

Stat 2: **Communality**, percentage of variance explained by the PCs (*Duforet-Frebourg et al. MBE 2016*).

Stat 3: **Mahalanobis distance** (*Luu et al. Mol Ecol Res 2016*).



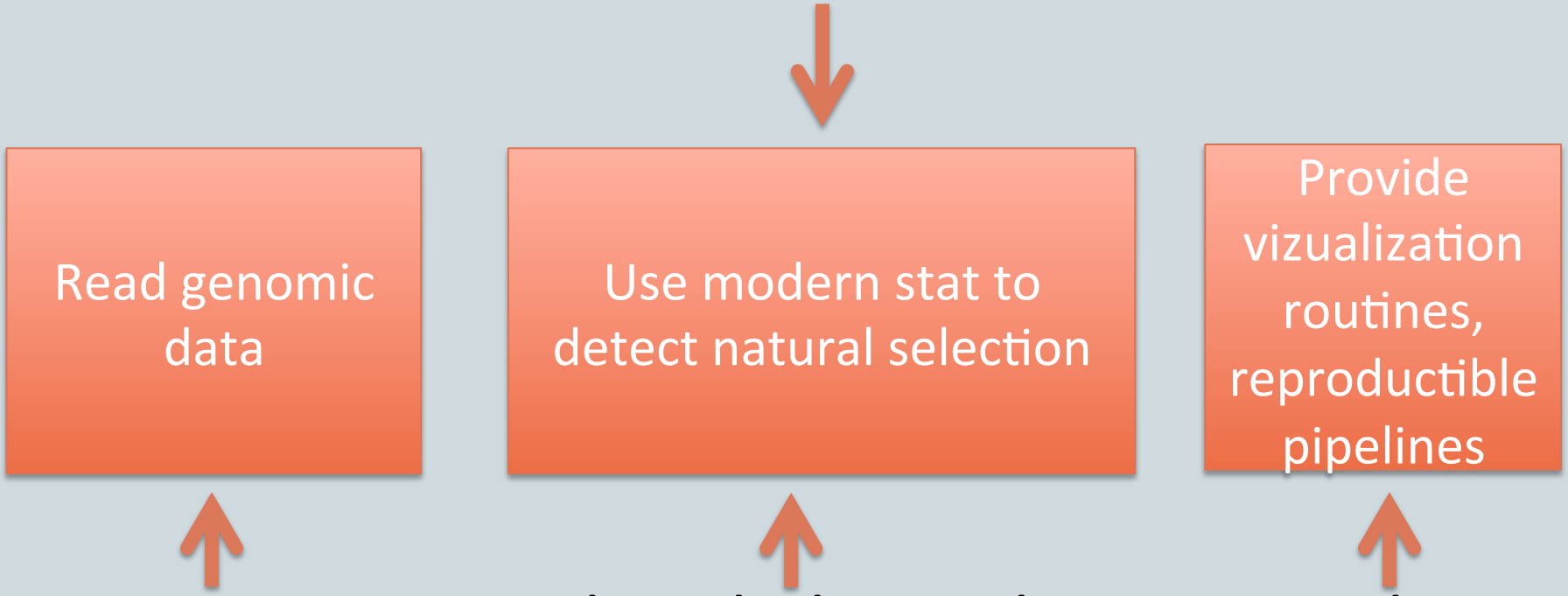
Natural selection in Europe detected with Mahalanobis distance.



Martins et al. Mol Ecol 2016

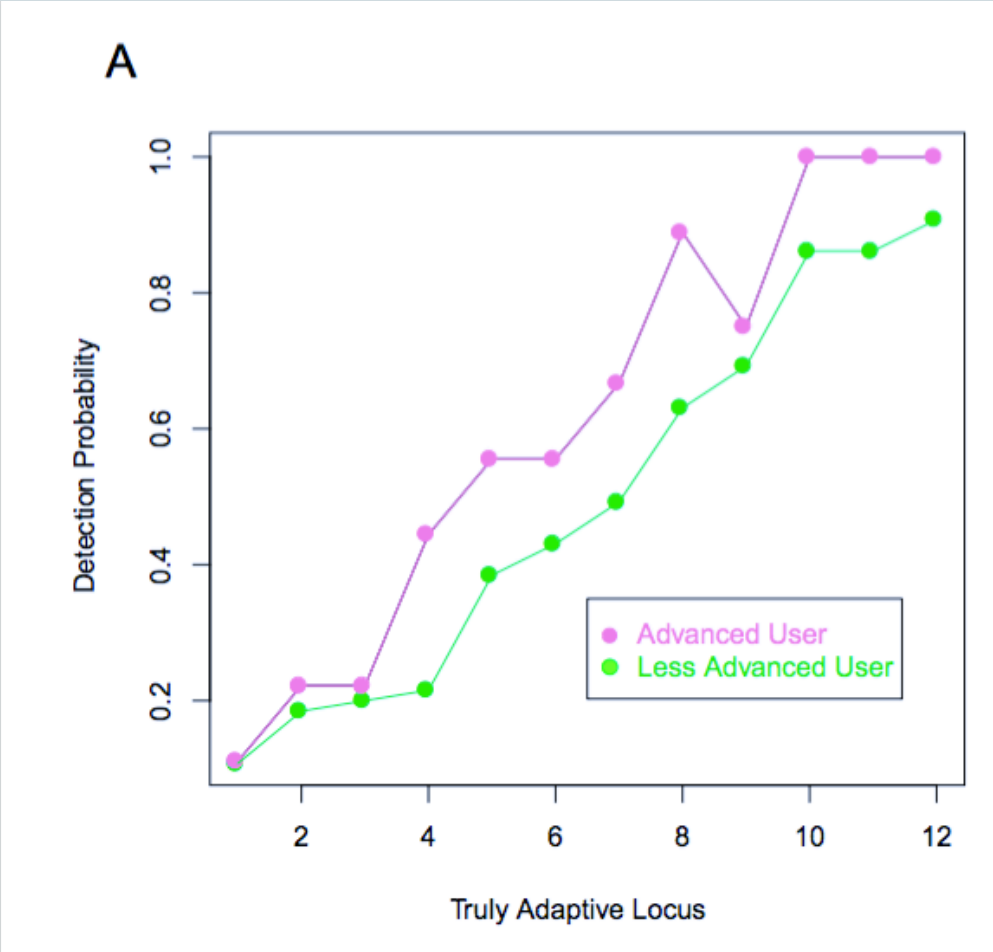
The difficulty of being a data scientist in biology.

We are are proud of this.



Geneticists want the whole pipeline in a single environment (R, Python) and they do not really care about the details of stat.

The SSMPG 2015 experiment



Lotterhos et al. BiorXiv 2016

Hadley Wickham's provocative statements about statistics

"There is a total disconnect between what people need to actually understand data and what was being taught."

"The fact that data science exists as a field is a colossal failure of statistics."

"Data munging and manipulation is hard and statistics has just said that's not our domain."