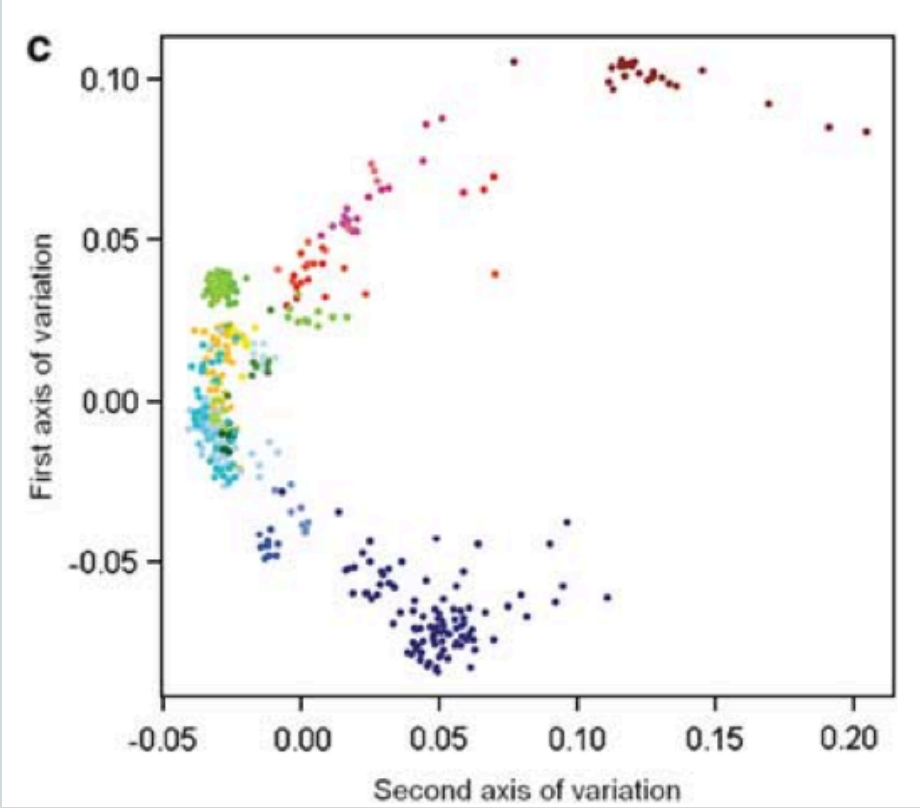
Bayesian robust principal component analysis to detect genomic regions involved in local adaptation

Michael GB Blum, Nicolas Duforet-Frebourg

CNRS, Université Joseph Fourier
Grenoble

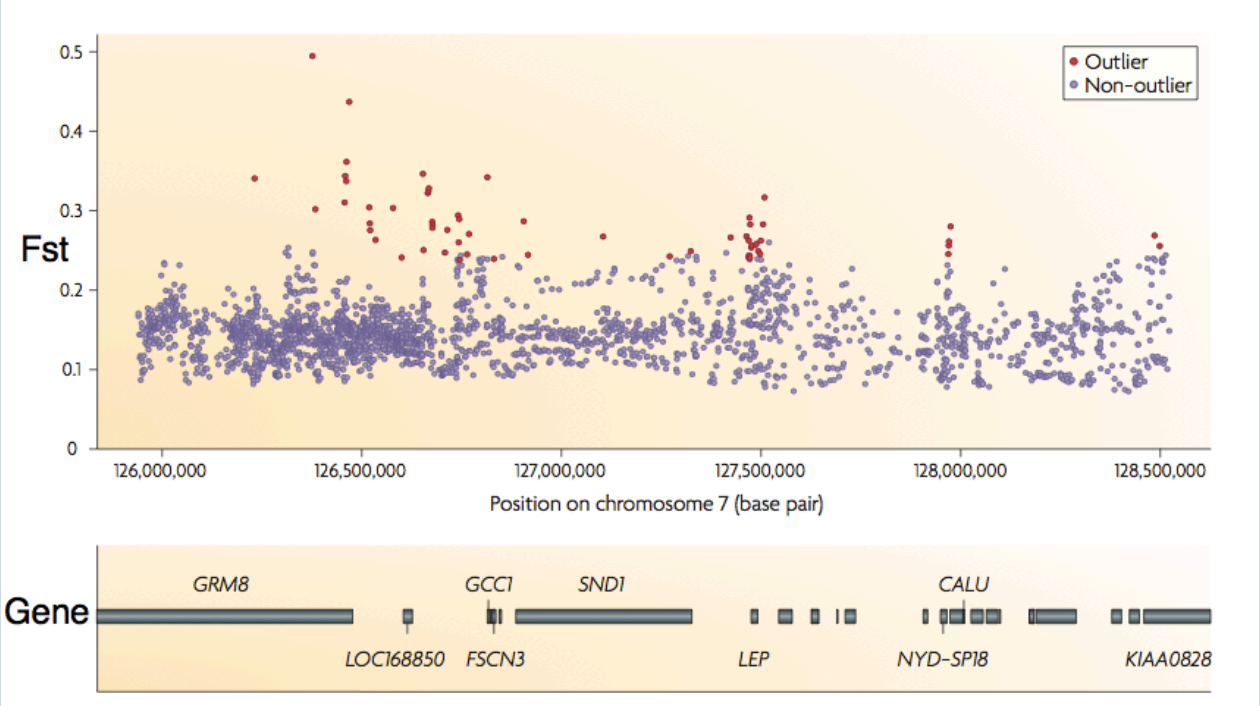
Principal component analysis in population genetics



Suo et al EJHG 2011

Local adaptation

- Definition: greater fitness of individuals in their local habitats due to natural selection
- Scanning genomes to look for the outlier regions, which have been involved in local adaptation



Holsinger and Weir, Nat Rev Genet 2009

Local adaptation
Does it matter for health?

- Local adaptation in humans

  The diversity of the local pathogenic environment is the predominant driver of local adaptation *Fumagalli et al. PLoS G 2011*

- Local adaptation in pathogens

  The loci involved in local adaptation are drug resistance loci in the Plasmodium falciparum malaria parasite *Park et al. PNAS 2012*

Principal component analysis (PCA)

PCA with K components is an optimal approximation of rank K for the matrix of genotypes X

$$X_i = \sum_{k=1}^{K} \lambda_i^k E_k$$

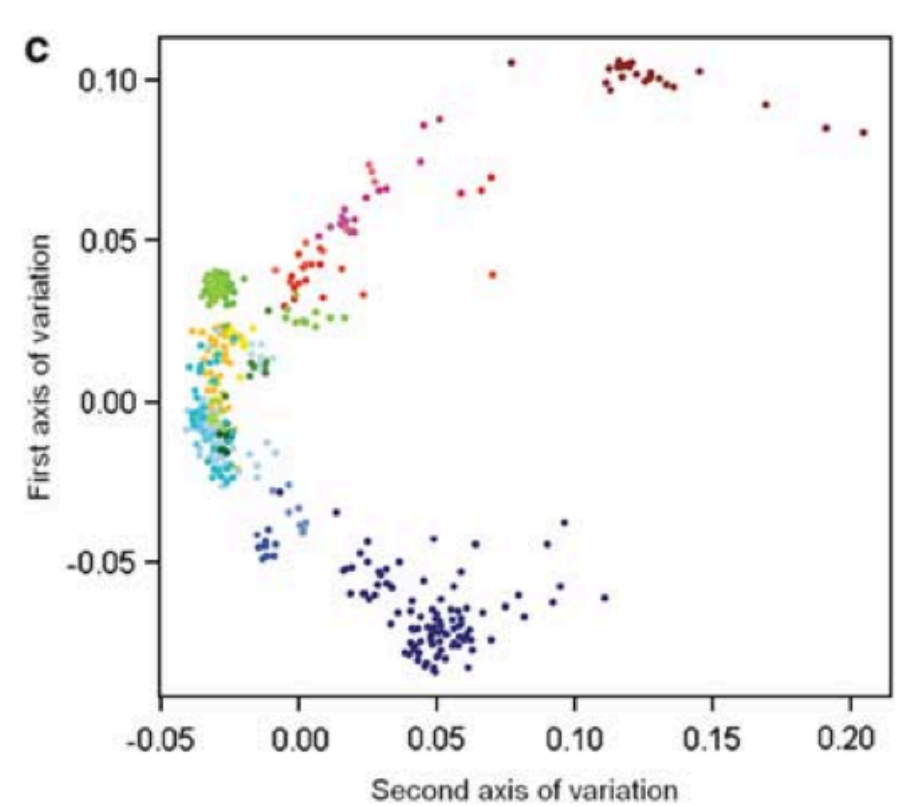
X_i: Genotype of the ith individual (0,1,1,2,0,0,.....)

E_k: vector of loadings $(E_k^1, E_k^2, E_k^3, \ldots)$ of the same length as X_i

Principal component analysis (PCA)

$$X_i = \sum_{k=1}^{K} \lambda_i^k E_k$$

In the example of East Asiatic data, the outlier loci in E_1 correspond to the regions involved in adaptation along the latitudinal gradient (*Laloe and Gautier, arXiv*)



Robust Bayesian principal component analysis

- A probabilistic version of PCA
  Tipping and Bishop JRSSB 1999

$$X_i = \sum_{k=1}^{K} \lambda_i^k E_k + \varepsilon_i$$

- The location-shift model for outlier detection
  Verdinelli and Wasserman 1991

$$p(E_{\cdot}^{j}) = (1-\pi)\ \mathrm{N}(0,\sigma^2) + \pi\ \mathrm{N}(\mathrm{A},\sigma^2)$$

where π is the genome-wide outlier probability.

Robust Bayesian principal component analysis

- Can account for different proportions of outliers for each PC

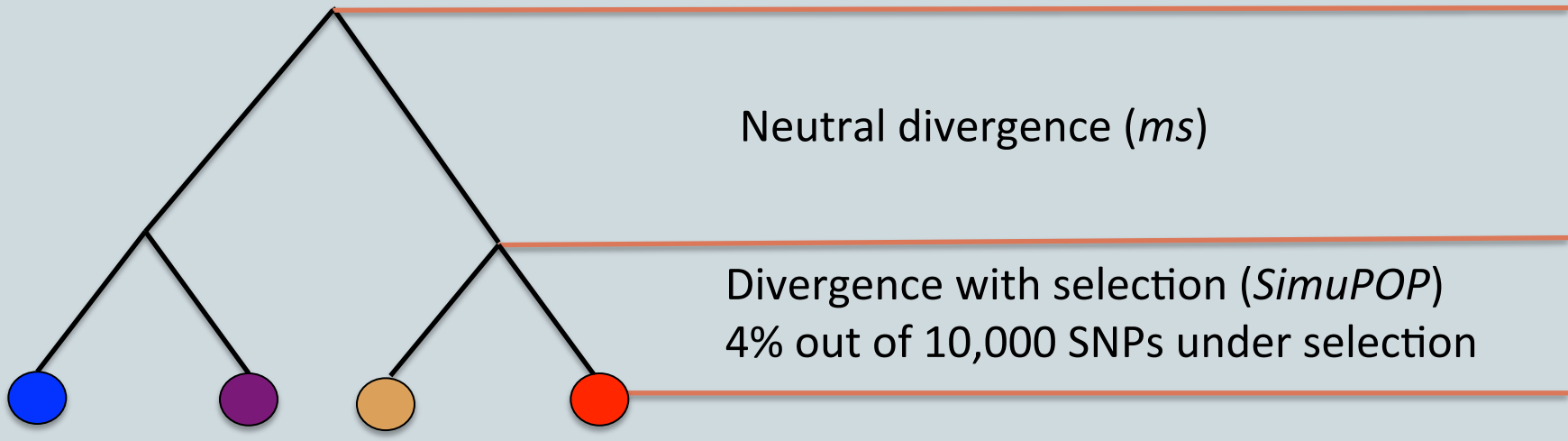
$$\pi = \pi_1 + \pi_2 + \ldots + \pi_K$$

where π_k is the probability to be outlier for the kth PC.

- A general advantage of Bayesian method for genome scan *Stephens and Balding Nat Rev Genet 2009*

  Provide an estimate of the false discovery rate because for each loci we have P(no selection|D) and P(selection|D)

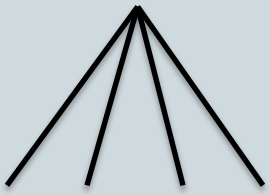
A simulation study in a divergence model



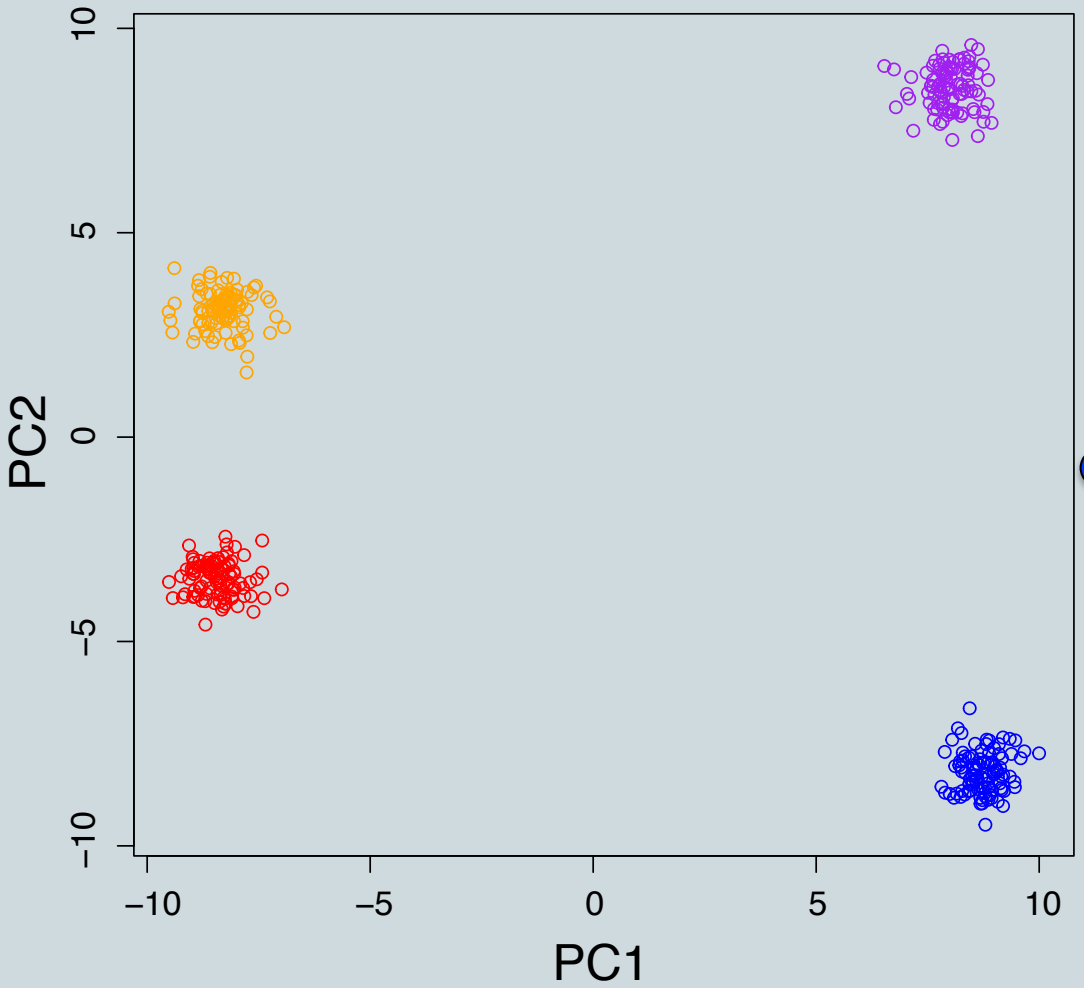
Methods for selection scan:

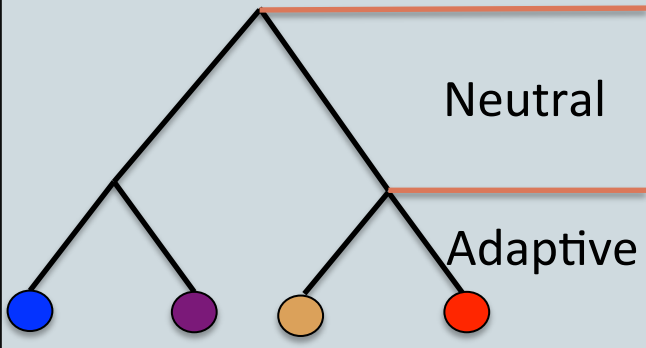
- F_{st} A measure of differentiation between populations
- Standard PCA and PCAdapt
- BayeScan (*Foll and Gaggiotti Genetics 2008*)

  Assumes a mechanistic model of instantaneous divergence

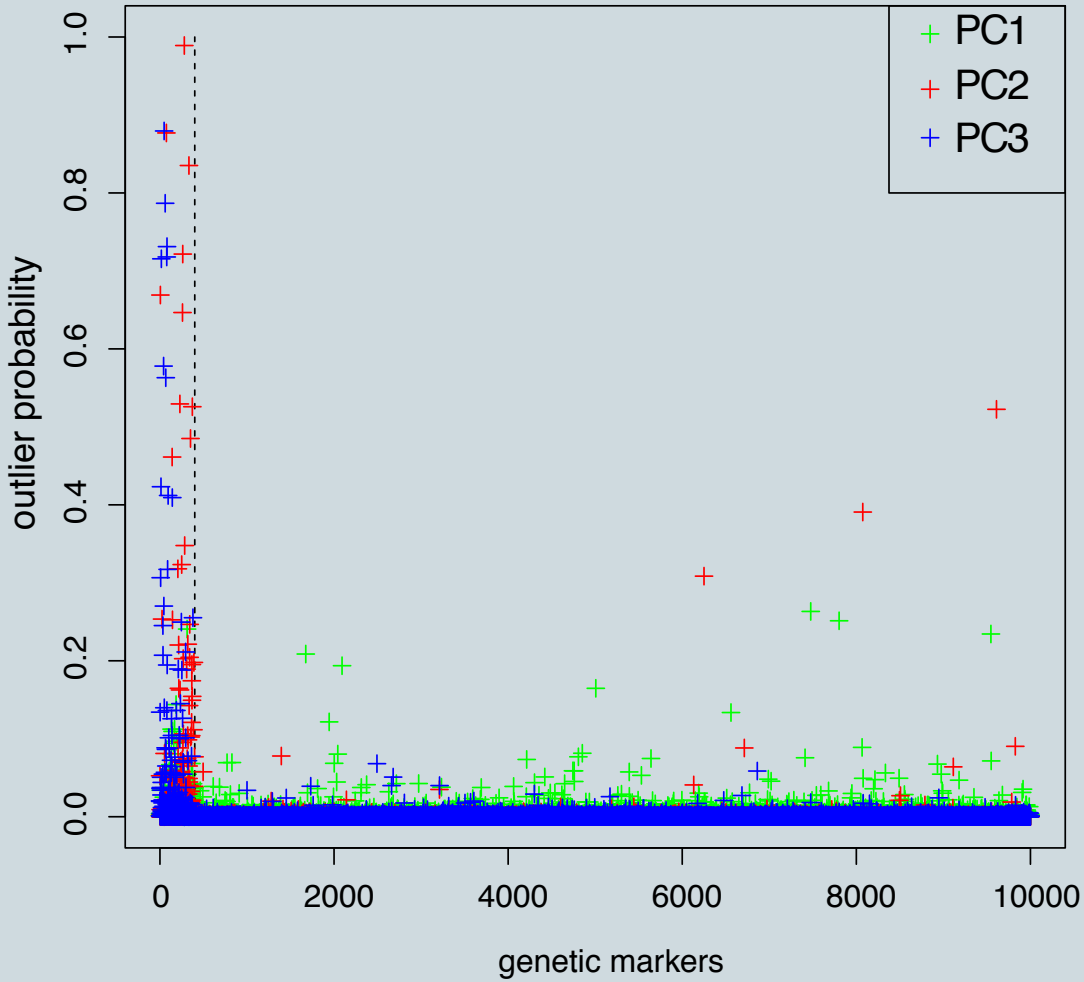


A simulation study in a divergence model

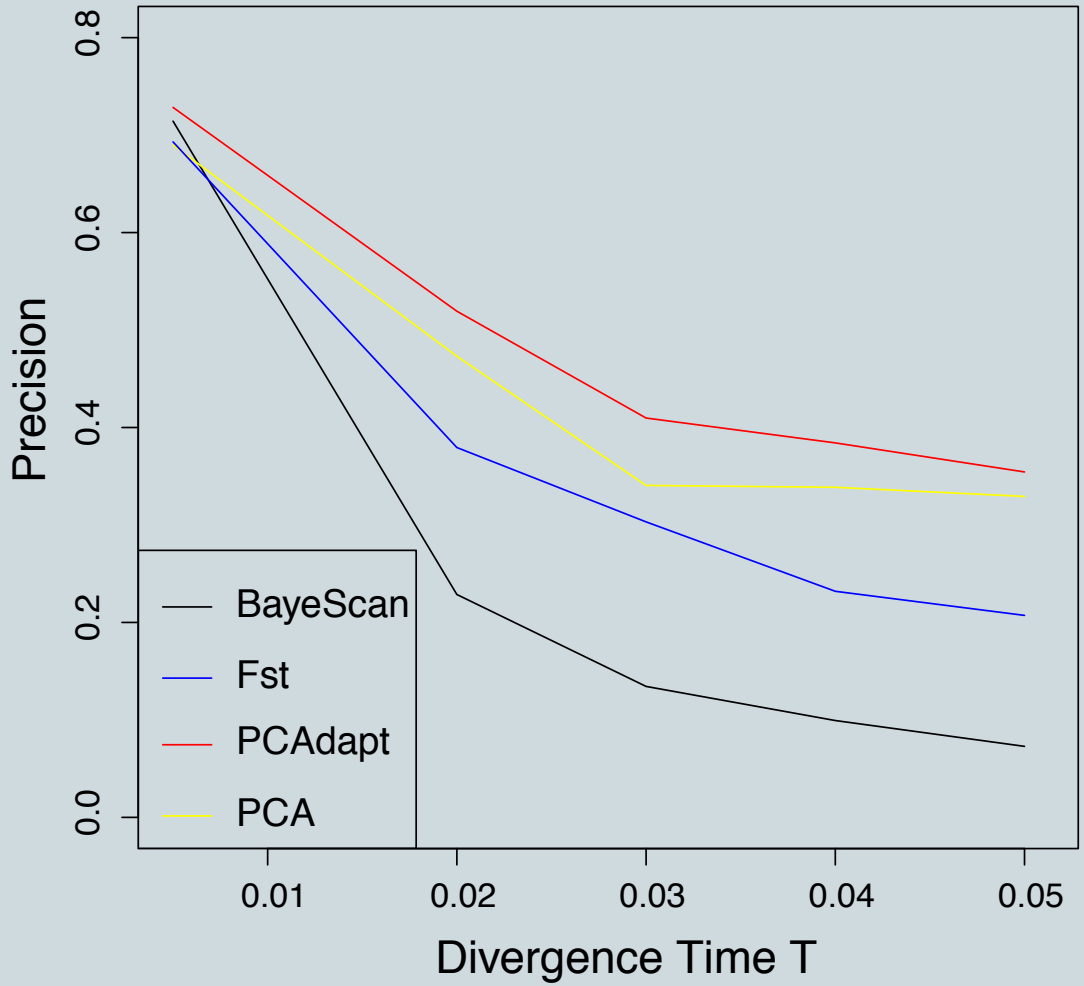


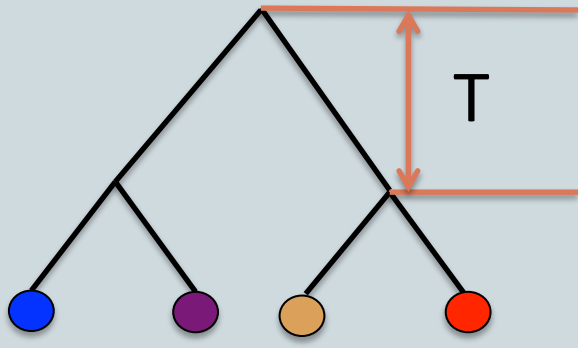


A simulation study in a divergence model



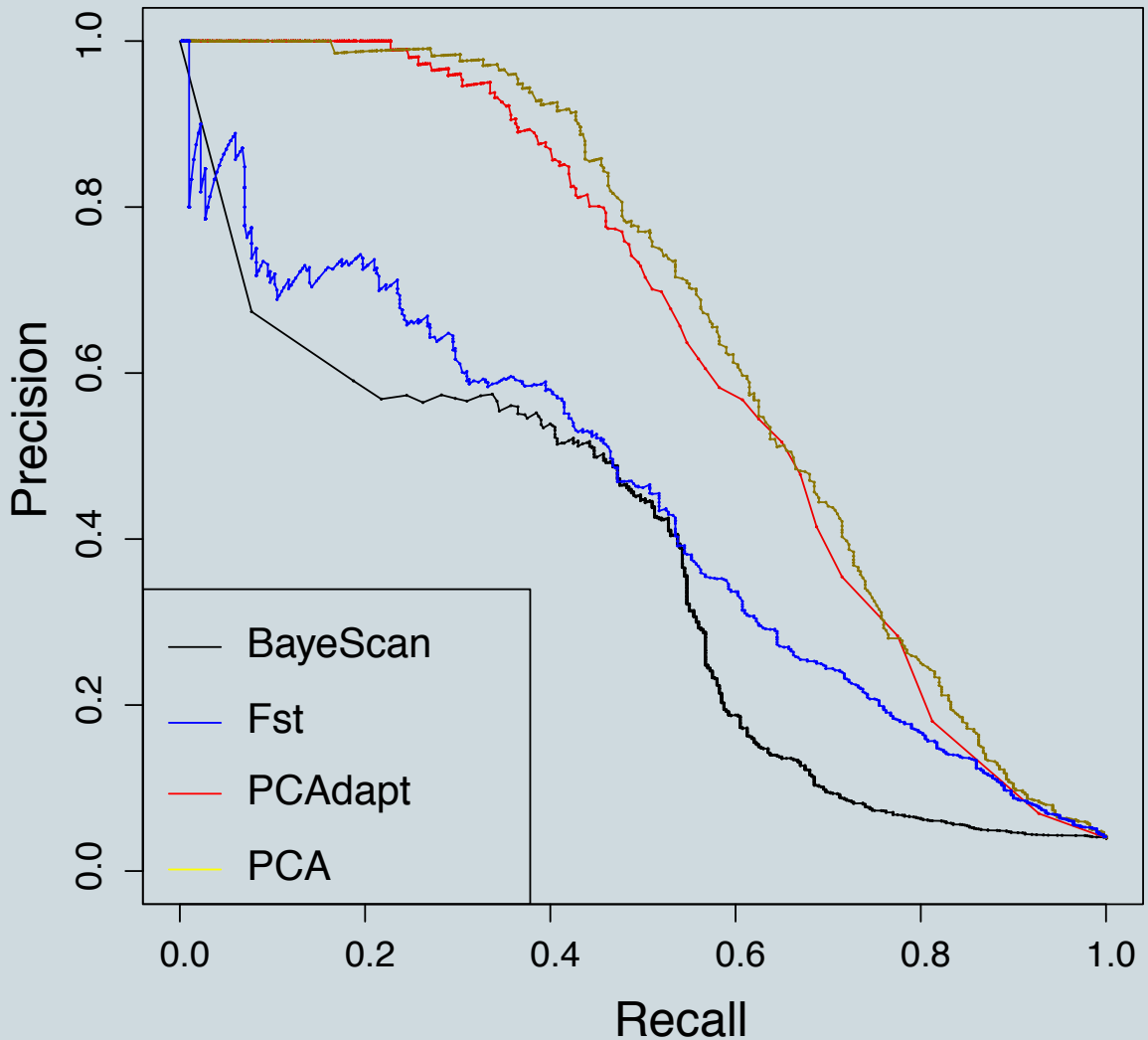
Comparing methods of selection scan





Precision =1-False Discov. Rate
Fraction of loci under selection among the significant loci

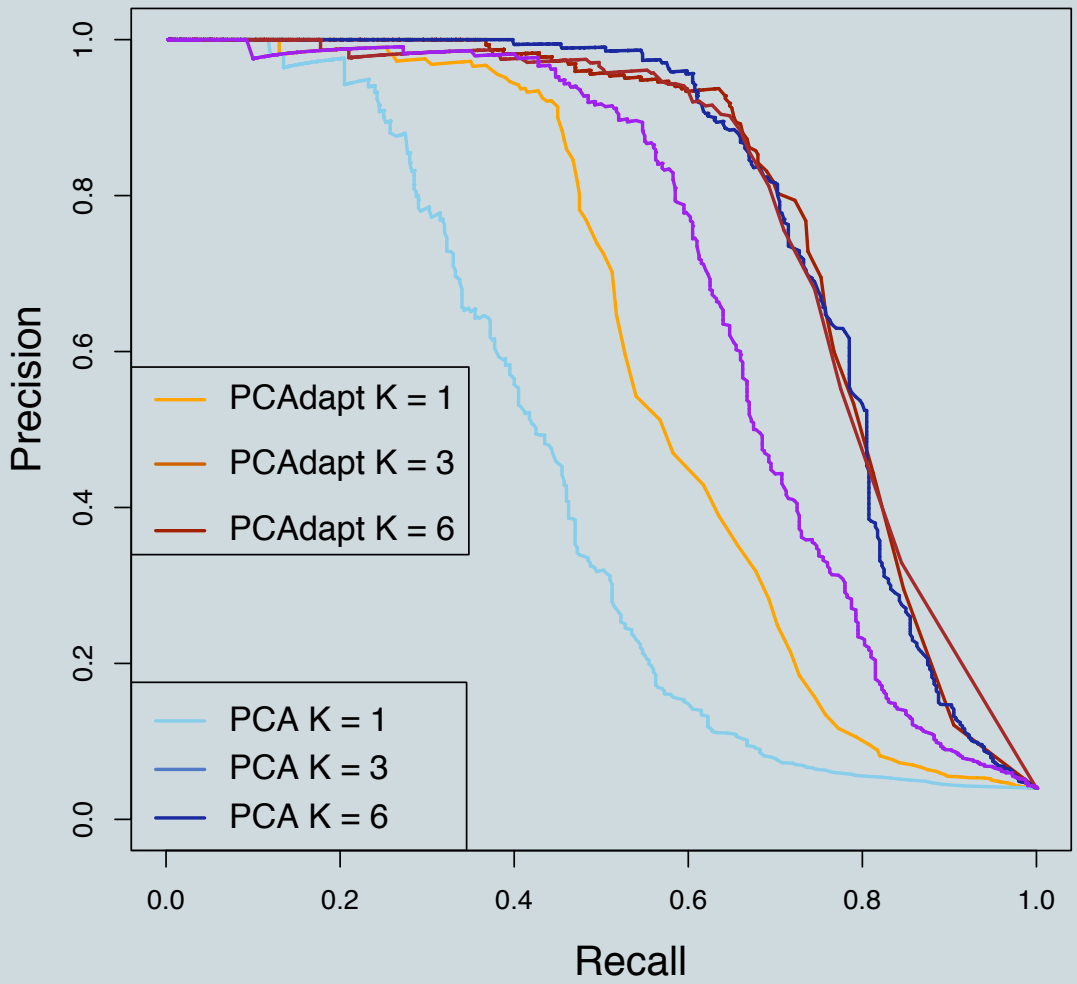
Comparing methods of selection scan



Precision
Fraction of loci under selection among the significant loci

Recall
Fraction of the loci under selection that are significant

Robustness w.r.t. the choice of K



Conclusions

- PCA and a Bayesian variant of PCA (PCAdapat) can be used for selection scan

- Robustness of (non-parametric) statistical models compared to a mechanistic model

$$X_i = \sum_{k=1}^{K} \lambda_i^k E_k$$ vs