the hapFLK method

María Inés Fariello, **Simon Boitard**, Magali San Cristobal, Bertrand Servin

INRA, GenPhySE, Toulouse
simon.boitard@toulouse.inra.fr

Doctoral course "Environmental genomics", May 23-27, 2016.

Outline

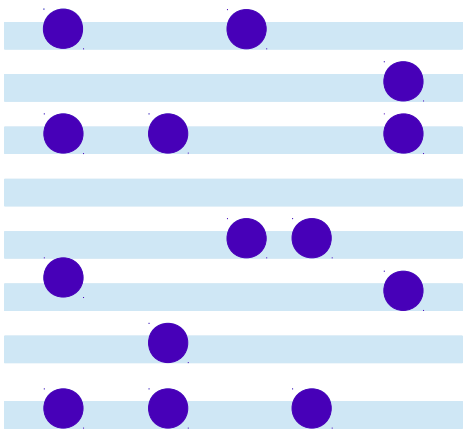
Outline

Genome scans for selection

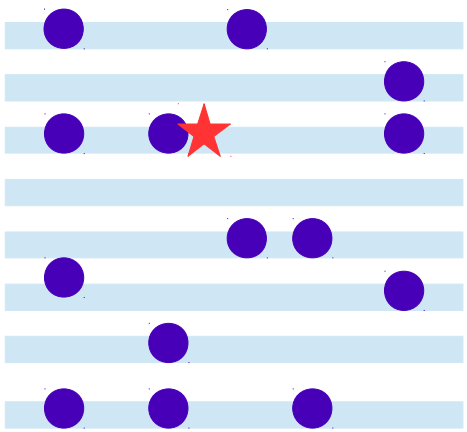
- Most genomic regions are neutral, but some of them are (or have been) under selection (natural or artificial).
- Detecting the regions under selection is important for theory (evolution) and applications (medicine, agronomy).
- Genome wide scans for selection now possible from dense genotyping (SNP chips) or sequencing (NGS) data.
- Focus on positive (adaptative) selection.

Genetic diversity around a positively selected mutation

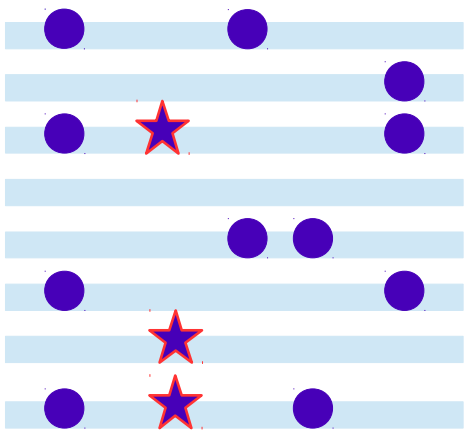
No selection



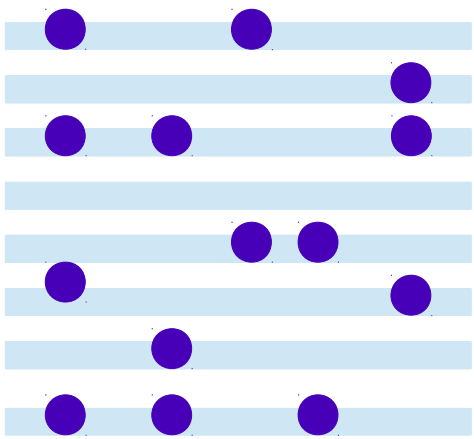
Selection on
a new variant



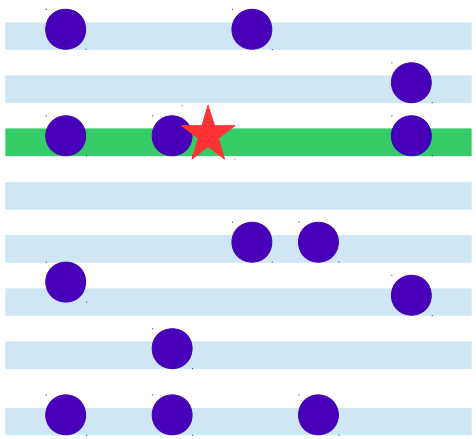
Selection on
standing variation



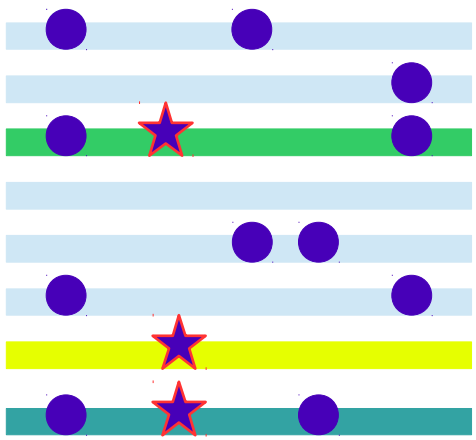
Genetic diversity around a positively selected mutation



One single haplotype



Several haplotypes

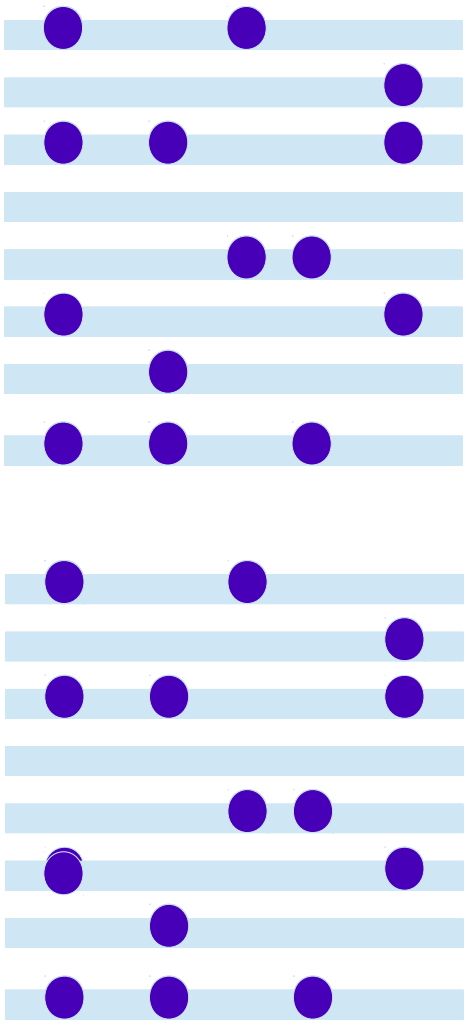


Genetic diversity around a positively selected mutation

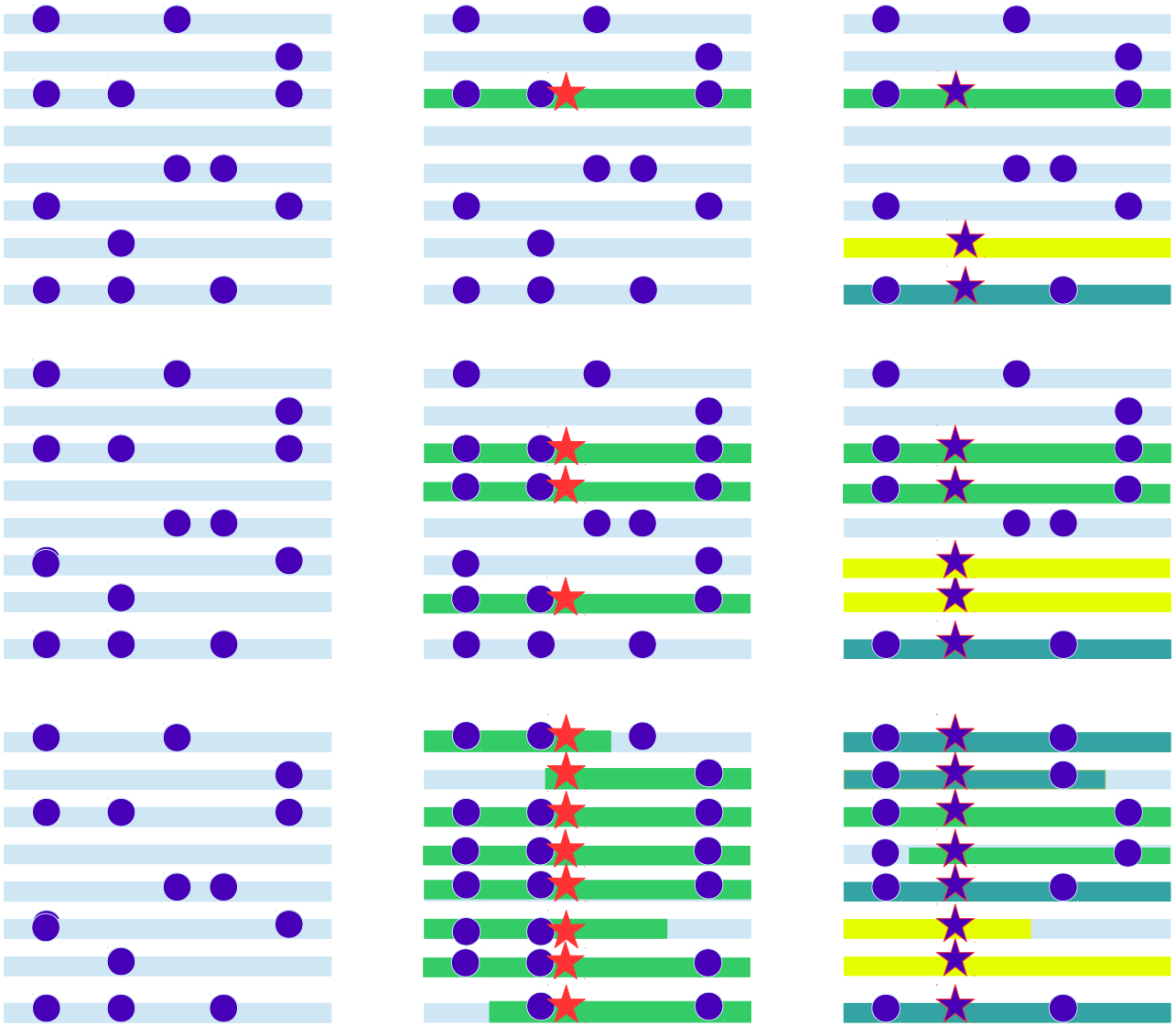
Random drift evolution

One haplotype increases in frequency

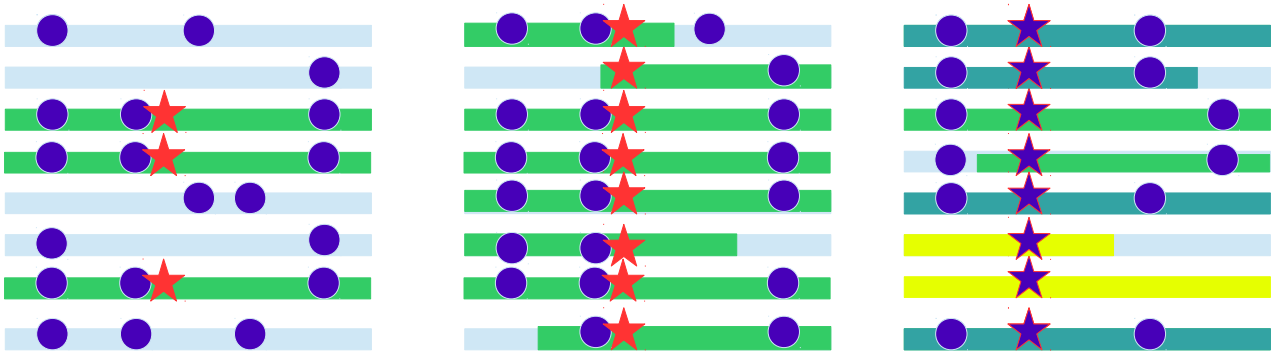
Several haplotypes increase in frequency



Genetic diversity around a positively selected mutation



Different sweep signatures



Sweep scenario	partial	hard	soft
Allele frequencies	elevated	extreme	intermediate / extreme
Haplotype frequencies	elevated for one haplotype	one fixed haplotype	elevated for several haplotypes

Detection of selective sweeps from genomic data

Detection power may be increased by:

- Comparing neutral vs selected populations.
- Account for population history, in particular their hierarchical structure (FLK).
- Using haplotype information (hapFLK).

Outline

Outline

F_{ST} based tests

$p = (p_1, \ldots, p_i, \ldots, p_n)$: allele frequencies at one SNP in several populations.

$\overline{p}$ and s_p^2: observed mean and variance of p.

$$F_{ST} = \frac{s_p^2}{\bar{p}(1-\bar{p})}$$

- H_0 : "neutral evolution" (genetic drift)
  vs H_1 : "positive selection in one (or more) population ".
- H_0 rejected if F_{ST} too large.

Lewontin et Krakauer (LK) test (1973)

$$T_{LK}^{\ell} = \frac{n-1}{\bar{F}_{ST}} F_{ST}^{\ell}$$

- T_{LK} distribution under H_0 is χ^2 if :

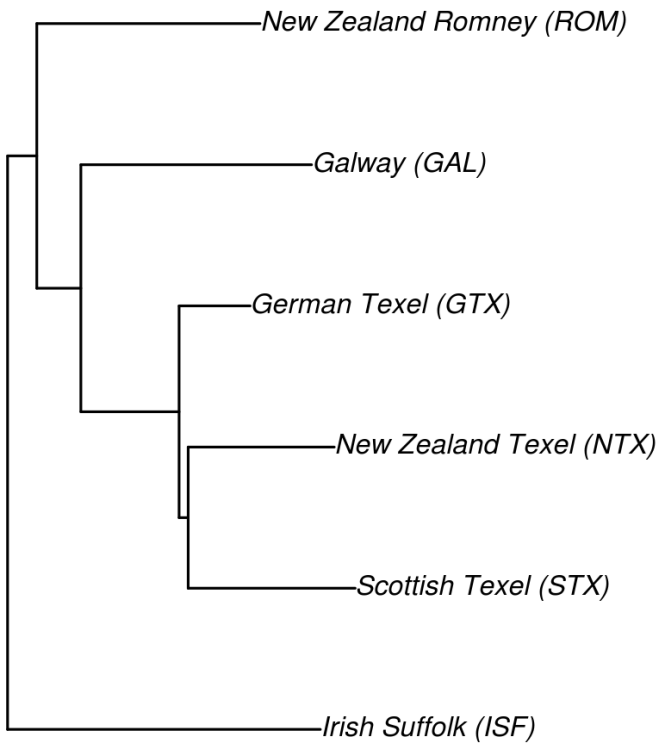
$$Var(p_i) = F_{ST} p_0 (1 - p_0), \quad Cov(p_i, p_j) = 0$$

- Only true if populations have a star like phylogeny with equal population sizes.

FLK test (Bonhomme *et al*, 2010)

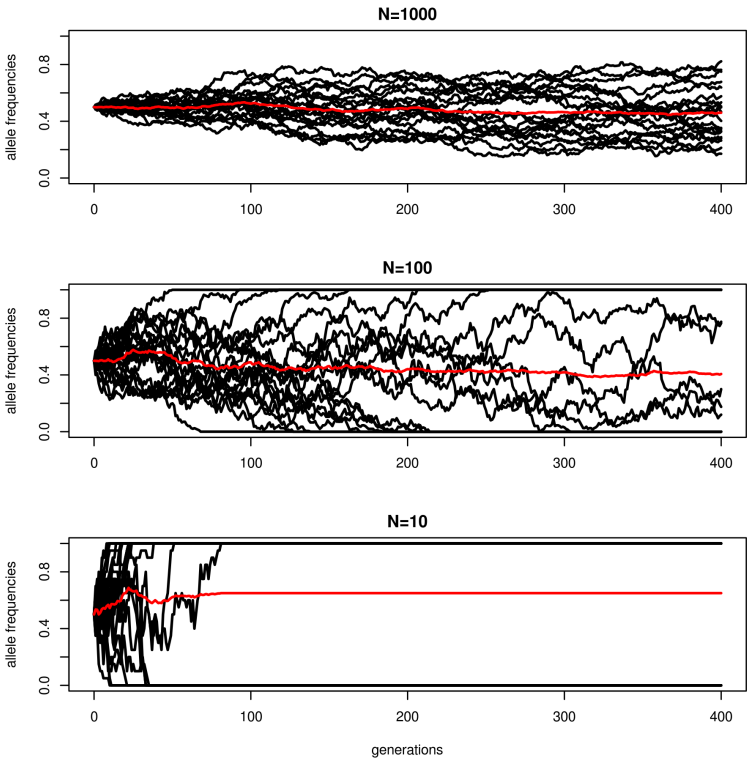
Extension of F_{ST} accounting for

- differences in effective size between populations.
- differences in correlations between population pairs.



(first estimated from genome wide data)

Genetic drift in one population



$$\mathbb{E}(p(t)) = p_0 \quad (1)$$

$$Var(p(t)) = F_t p_0(1-p_0) \quad (2)$$

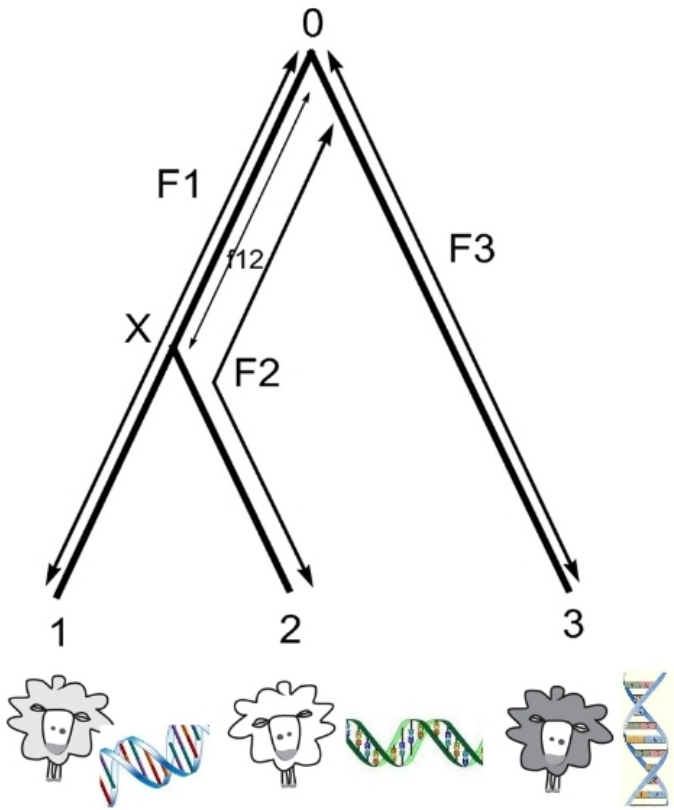
Wright-Fisher fixation index

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \approx \frac{t}{2N}$$

N: effective population size

Extension to several populations with arbitrary phylogeny

The distribution of p under H_0 can be modelled using the kinship matrix F.



$$Var(p_i) = F_i p_0(1-p_0)$$
$$Cov(p_i, p_j) = f_{ij} p_0(1-p_0)$$

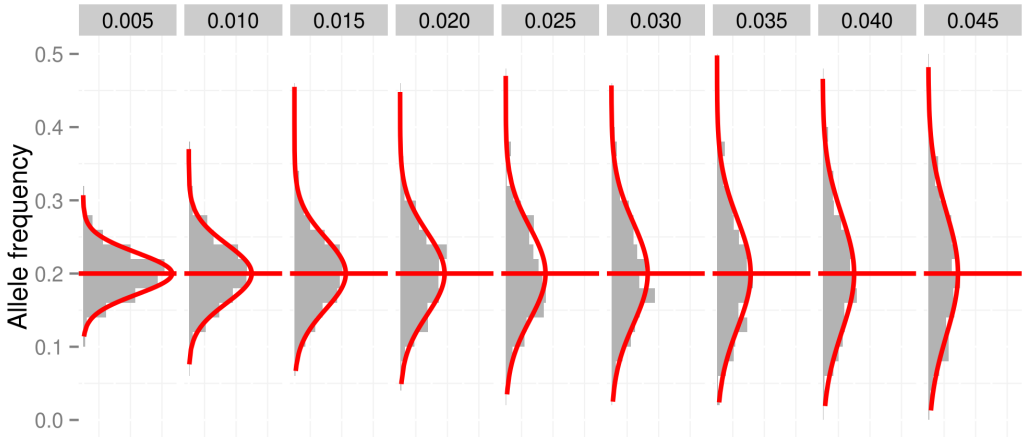
$$F_3 = 1 - \left(1 - \frac{1}{2N_3}\right)^t \approx \frac{t}{2N_3}$$
$$f_{12} = 1 - \left(1 - \frac{1}{2N_{12}}\right)^{t_{12}} \approx \frac{t_{12}}{2N_{12}}$$

Kinship matrix

$$F = \begin{pmatrix} F_1 & f_{12} & 0 \\ f_{12} & F_2 & 0 \\ 0 & 0 & F_3 \end{pmatrix}$$

$$\rightarrow Var(p) = F p_0(1-p_0)$$

Normal approximation



- Provided F_t is small, we can model:

$$p(t) \sim \mathcal{N}(p0, F_t p0(1 - p0))$$

See Nicholson *et al.* (2002).

FLK test (Bonhomme *et al.* 2010)

bi-allelic markers (SNP)

$$T_{F-LK} = (\mathbf{p} - \hat{p}_0 \mathbf{1}_n)' \widehat{Var(\mathbf{p})}^{-1} (\mathbf{p} - \hat{p}_0 \mathbf{1}_n)$$

$$\hat{p}_0 = \frac{\mathbf{1}_n' F^{-1} \mathbf{p}}{\mathbf{1}_n' F^{-1} \mathbf{1}_n}, \quad \widehat{Var(\mathbf{p})} = F \hat{p}_0 (1 - \hat{p_0})$$

multi-allelic markers

$A > 2$ allèles

- $\mathbf{p}_0$ vector of size A.
- $\mathbf{p}$ vector of size $n \times A$.
- $Var(\mathbf{p})$ can be written as a function of F and $\mathbf{p} - \mathbf{p}_0$

Distribution under H_0 is $\chi^2((A-1)(n-1))$

FLK test (Bonhomme *et al.* 2010)

bi-allelic markers (SNP)

$$T_{F-LK} = (\mathbf{p} - \hat{p}_0 \mathbf{1}_n)' \widehat{Var(\mathbf{p})}^{-1} (\mathbf{p} - \hat{p}_0 \mathbf{1}_n)$$

$$\hat{p}_0 = \frac{\mathbf{1}_n' F^{-1} \mathbf{p}}{\mathbf{1}_n' F^{-1} \mathbf{1}_n}, \quad \widehat{Var(\mathbf{p})} = F \hat{p}_0 (1 - \hat{p_0})$$

multi-allelic markers

$A > 2$ allèles

- $\mathbf{p}_0$ vector of size A.
- $\mathbf{p}$ vector of size $n \times A$.
- $Var(\mathbf{p})$ can be written as a function of F and $\mathbf{p} - \mathbf{p}_0$

Distribution under H_0 is $\chi^2((A-1)(n-1))$

Estimation of the population kinship matrix

- The **Reynolds genetic distance** $\mathcal{D}$ (Reynolds, Weir and Cockerham, 1983) between two populations i and j has expectation:

$$E(\mathcal{D}_{ij}) = \frac{F_i + F_j}{2}$$

  see Laval *et al.* (2002).
- The matrix of Reynolds distances is computed over many ($\sim 10^4$) SNPs. Assumes majority of them are neutral.
- The population tree is built using the neighbour joining algorithm on this matrix.

Outline

Accounting for correlation between SNPs

Cumulating single SNP tests

Windowing approach average / max of SNP statistics over genome windows (*Weir et al. 2005*)

Composite likelihood Product of SNP likelihoods within genome windows (XP-CLR : *Chen et al. 2010*)

Bayesian hierarchical models autoregressive component in the model (*Guo et al. 2009*)

Choice of window size? fixed?

Bayesian methods computer intensive (MCMC).

Accounting for correlation between SNPs

Cumulating single SNP tests

Windowing approach average / max of SNP statistics over genome windows (*Weir et al. 2005*)

Composite likelihood Product of SNP likelihoods within genome windows (XP-CLR : *Chen et al. 2010*)

Bayesian hierarchical models autoregressive component in the model (*Guo et al. 2009*)

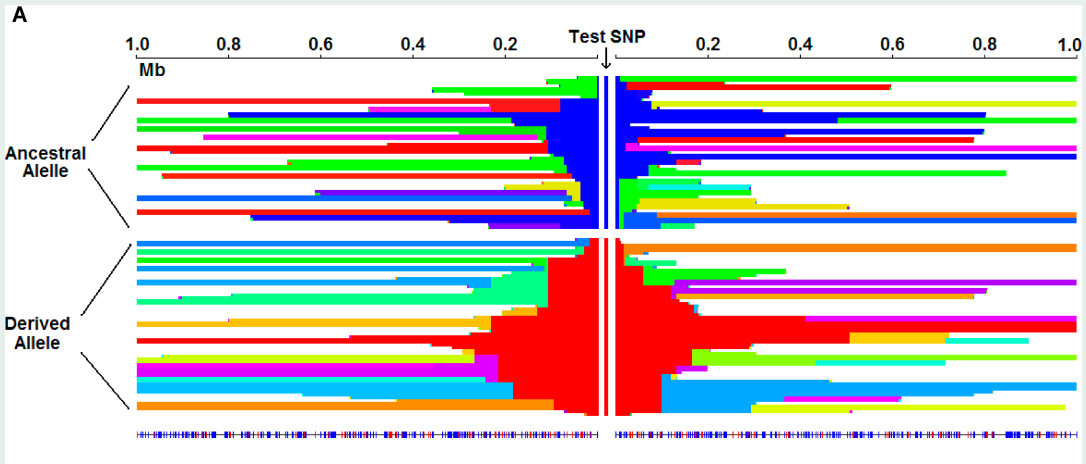
Choice of window size? fixed?

Bayesian methods computer intensive (MCMC).

Accounting for correlation between SNPs

Using haplotype length

Within population EHH (*Sabeti et al. 2002*), iHS (*Voight et al. 2006*)



Between populations XP-EHH *(Sabeti et al. 2007)*

Limited to two populations.

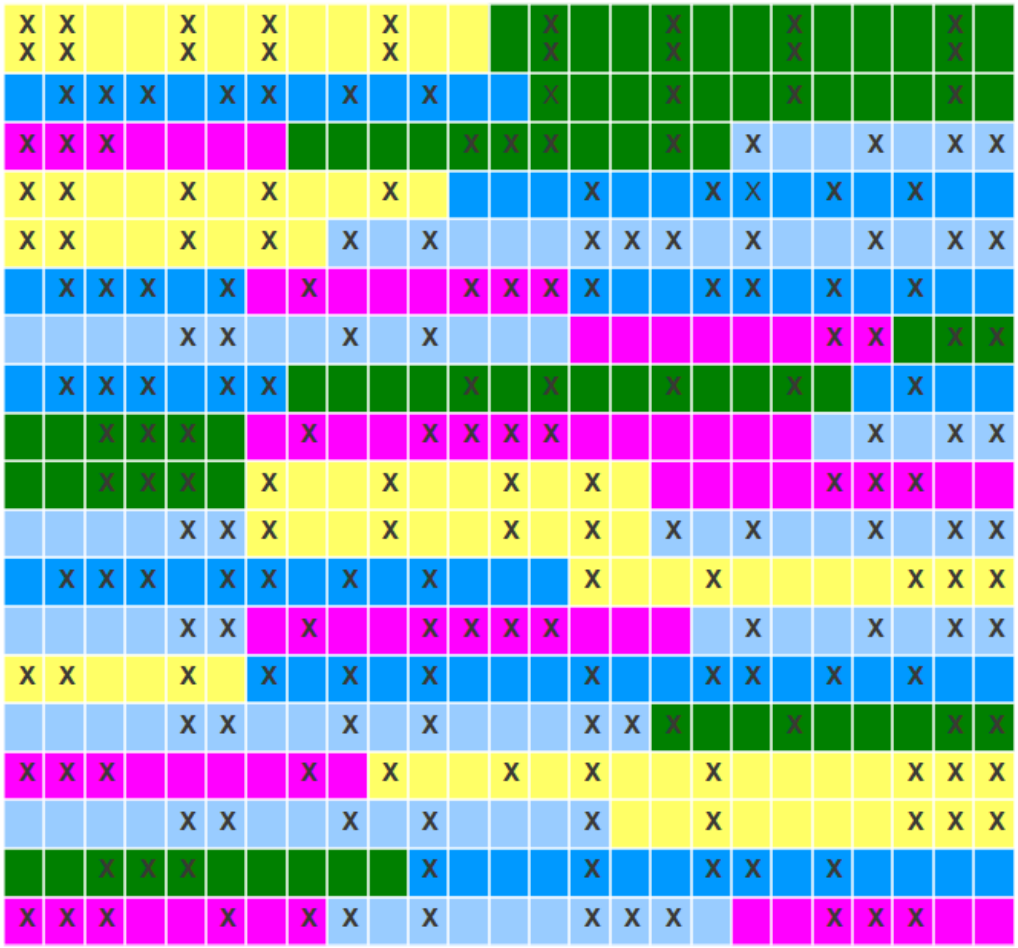
Our approach

FLK test at the haplotype level

1. Define haplotypes using a continuous model over the genome (no fixed window)
2. FLK test from these haplotypes (multi-allelic version)

Local haplotype clustering (*Scheet and Stephens* 2006)

Genetic similarity between individuals evolves continously over the genome due to ancestral recombinations.

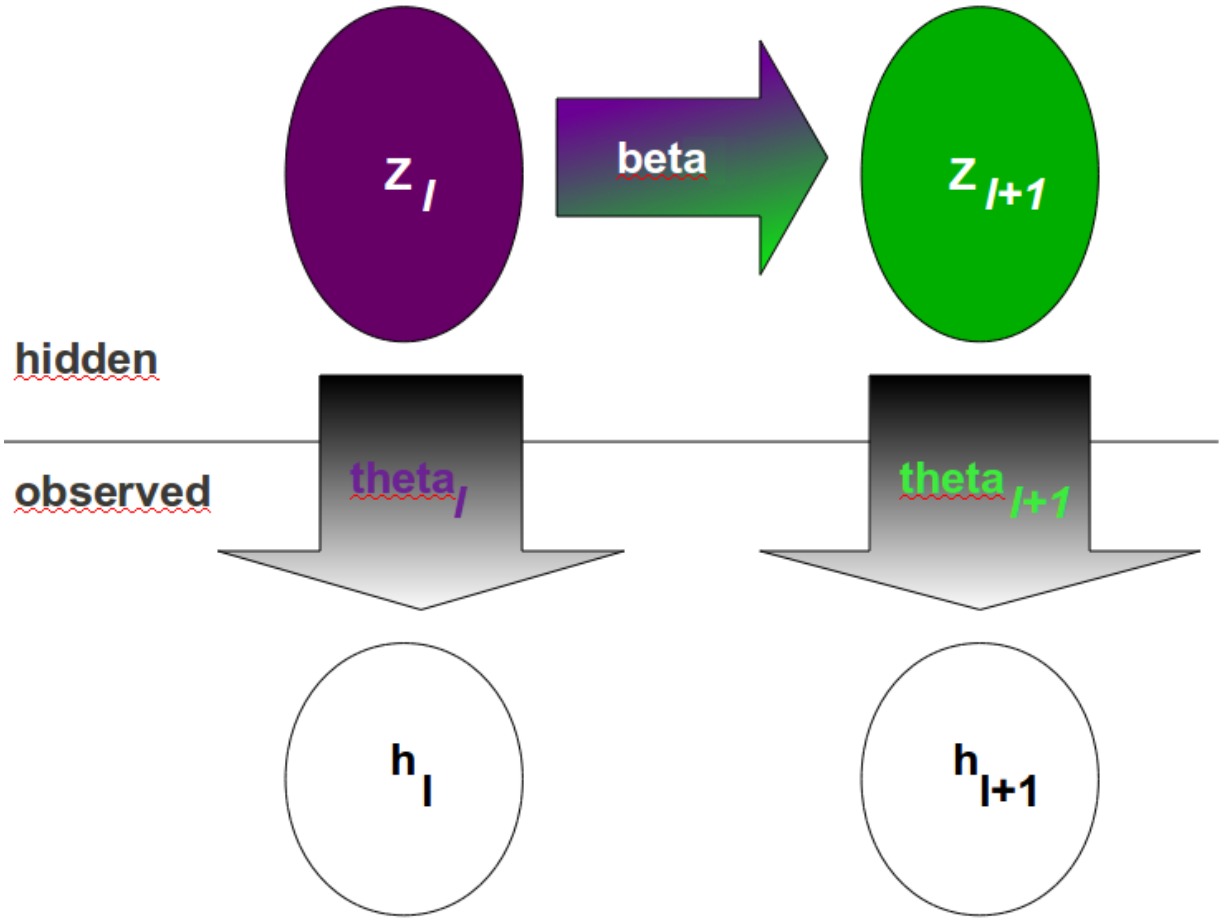


Example: 10 individuals

lines: haplotypes

columns: SNPs

Hidden Markov Model

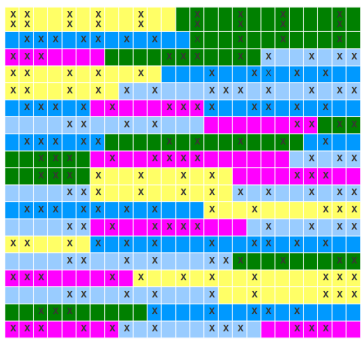


$z_{i\ell}$ cluster of haplotype i at SNP ℓ (hidden state)

$h_{i\ell}$ allele of haplotype i at SNP ℓ (observed state)

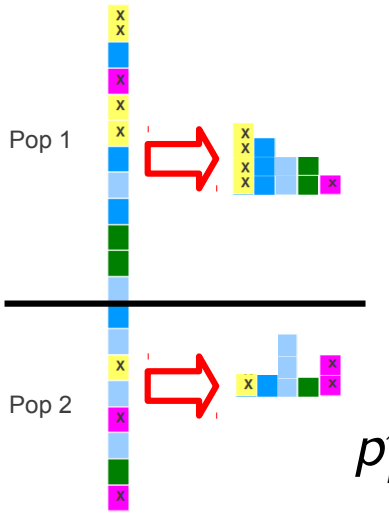
hapFLK test (Fariello *et al*, 2013)

Cluster estimation



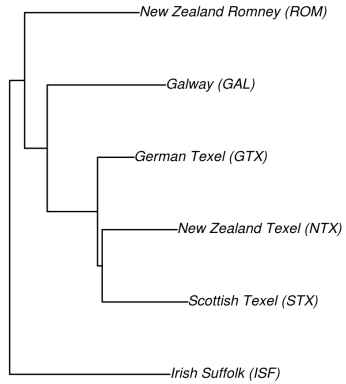
EM algorithm

Estimation of cluster frequencies for each SNP ℓ and population j:



$$p_{kj}^{\ell} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbb{P}(z_{ik}^{\ell} | \Theta)$$

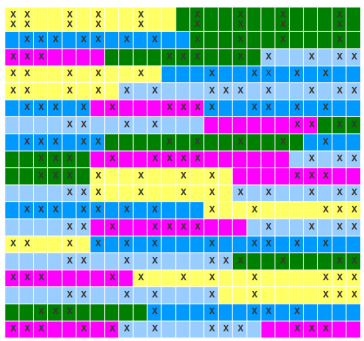
Computation of T_{F-LK}, using clusters as alleles.



Average of T_{F-LK} over EM iterations provides hapFLK

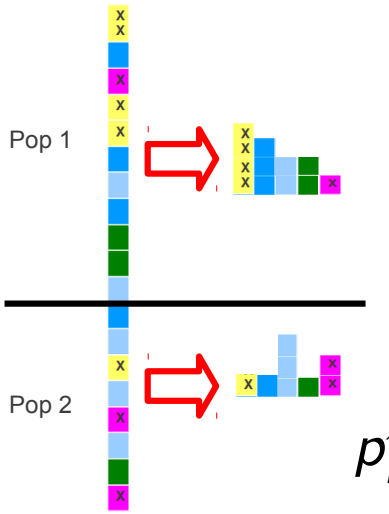
hapFLK test (Fariello *et al*, 2013)

Cluster estimation



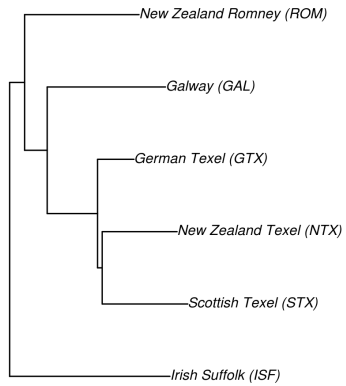
EM algorithm

Estimation of cluster frequencies for each SNP ℓ and population j:



$$p_{kj}^{\ell} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbb{P}(z_{ik}^{\ell} | \Theta)$$

Computation of T_{F-LK}, using clusters as alleles.



Average of T_{F-LK} over EM iterations provides hapFLK

Our approach

FLK test at the haplotype level

1. Define haplotypes using a continuous model over the genome (no fixed window)
2. FLK test from these haplotypes (multi-allelic version)

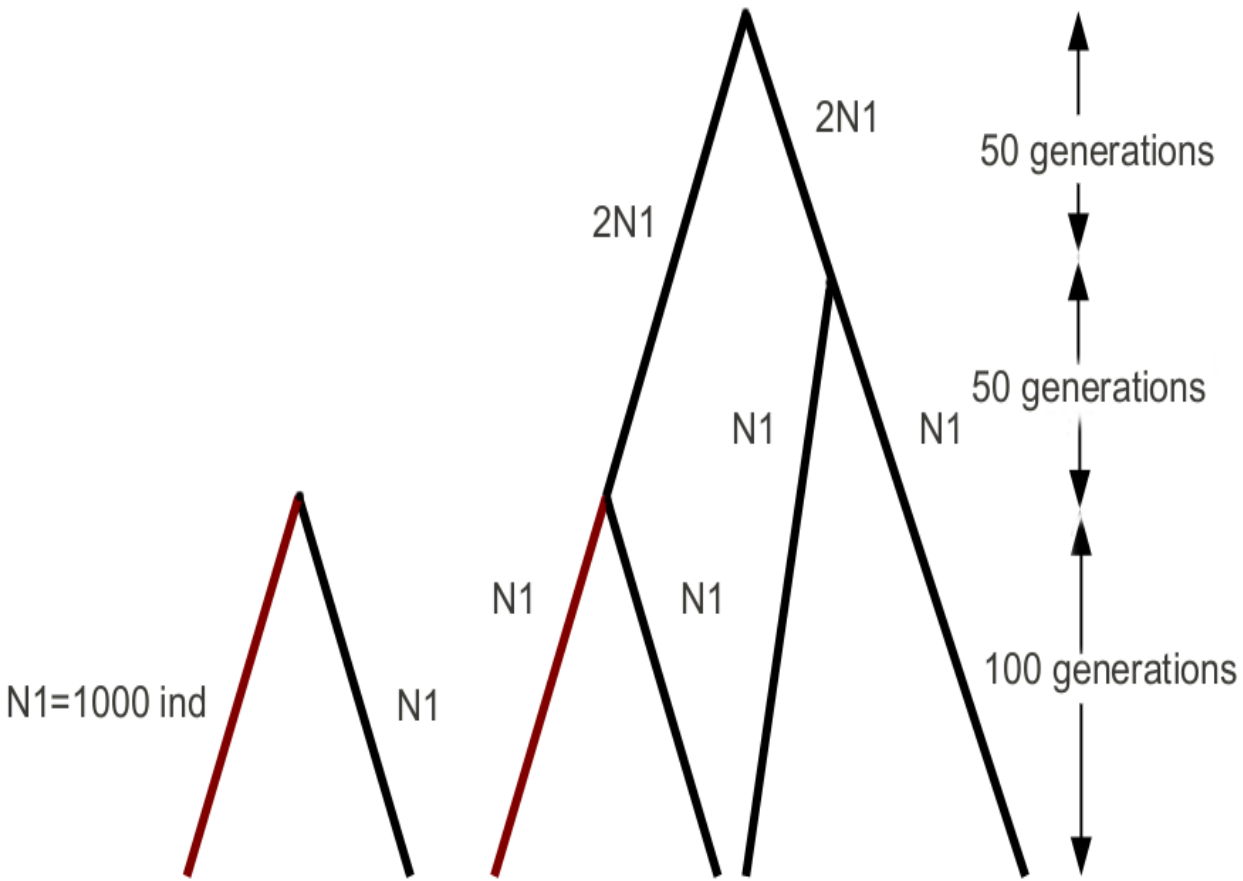
- Any number of populations.
- Accounts for populations hierarchical structure / unequal population sizes.
- Genotype data allowed.
- Missing data allowed.

Outline

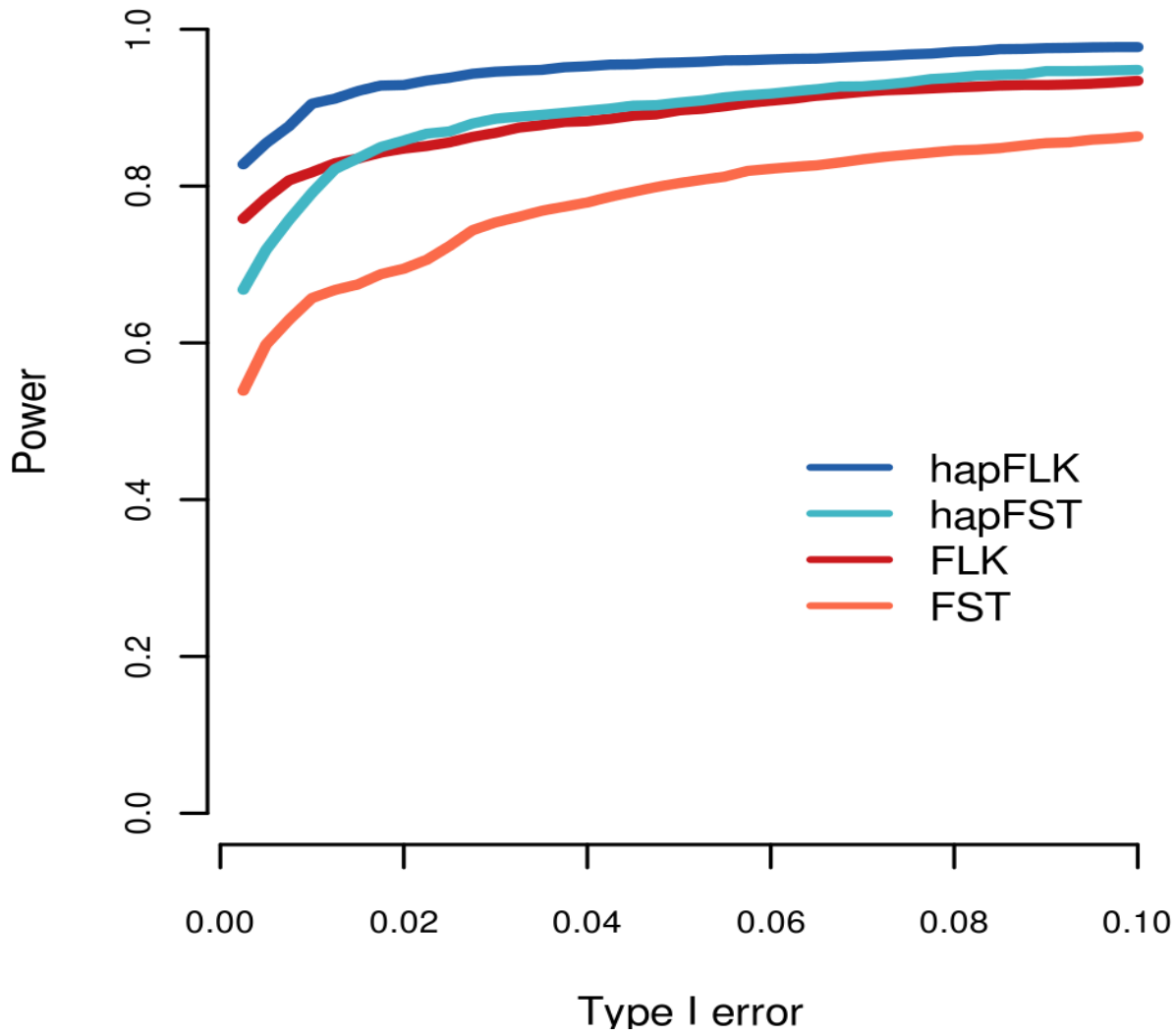
Outline

Simulation results

Simulation of 5Mb segments with 100 SNPs (dense genotyping) or 300 SNPs (full sequencing) with MAF $> 5\%$.

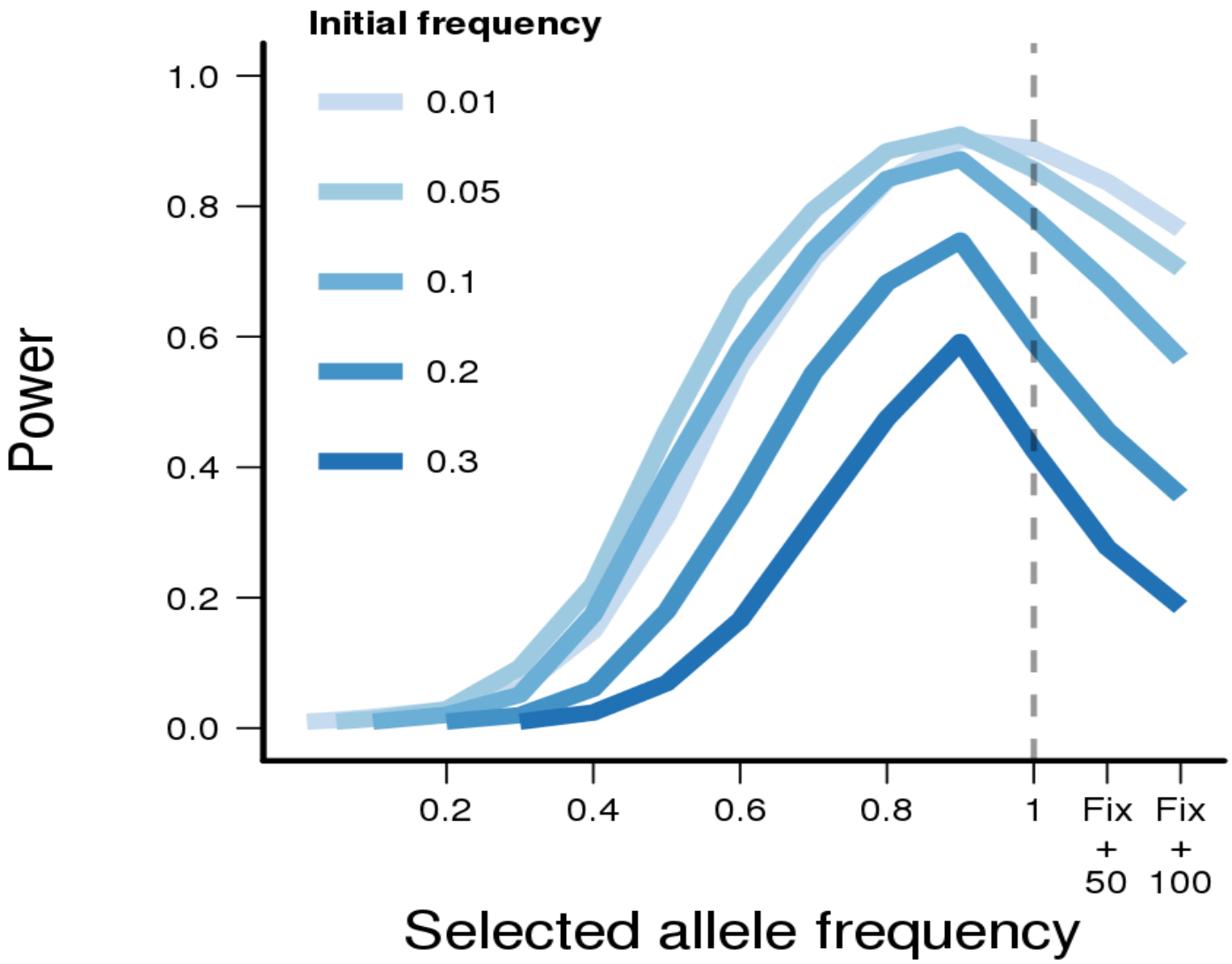


Increased power for haplotype based tests



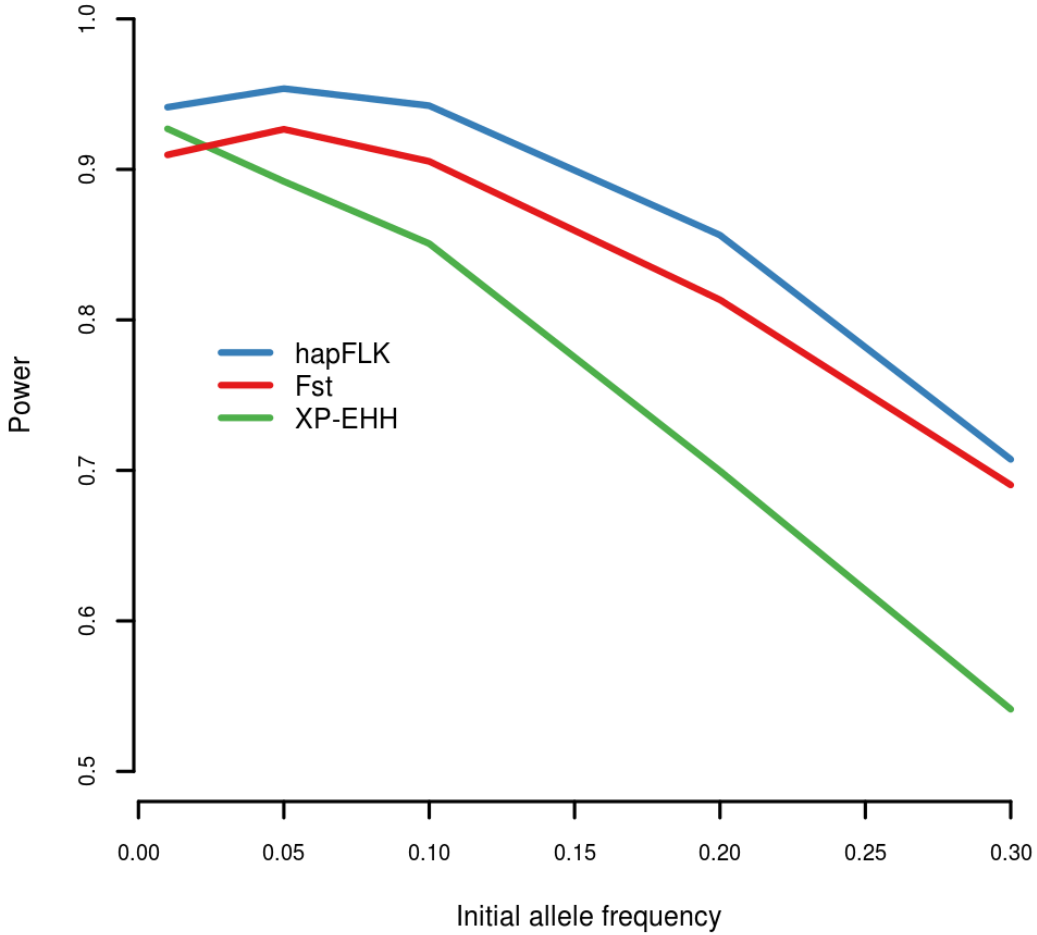
Hard sweep scenario, 4 populations, genotyping data.

Soft and incomplete sweeps can be detected



2 populations, genotyping data.

Increased power for soft sweeps compared to XP-EHH

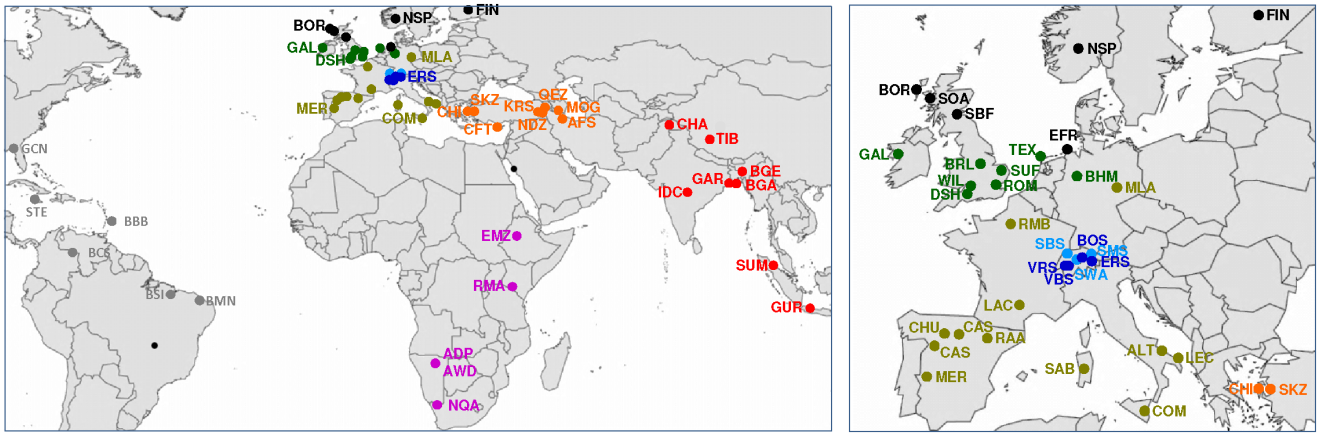


2 populations, genotyping data, type I error $= 5\%$.

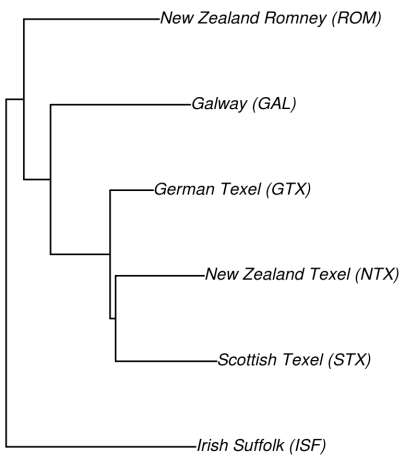
Outline

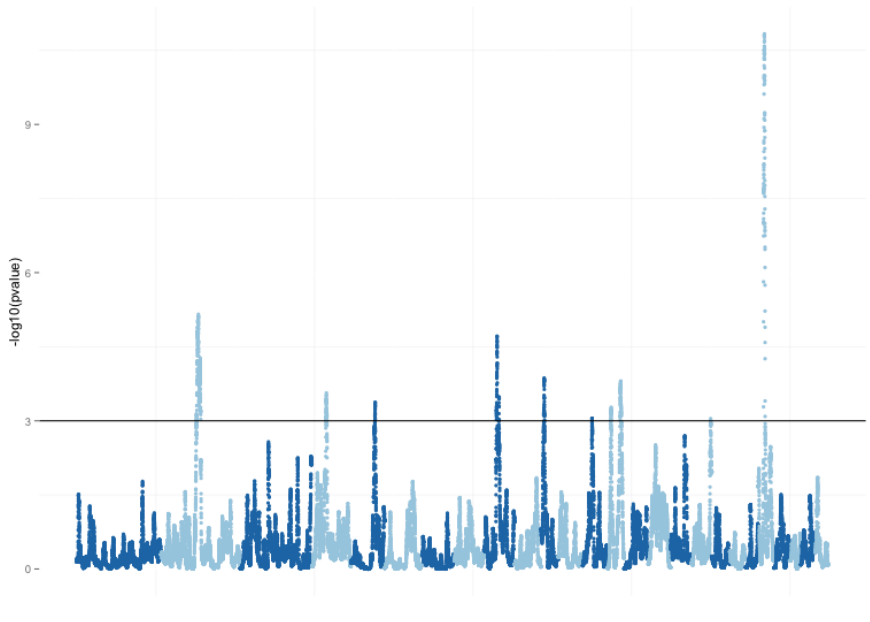
The Sheep HapMap data

74 populations, 50K SNPs (Kijas *et al*, 2012)

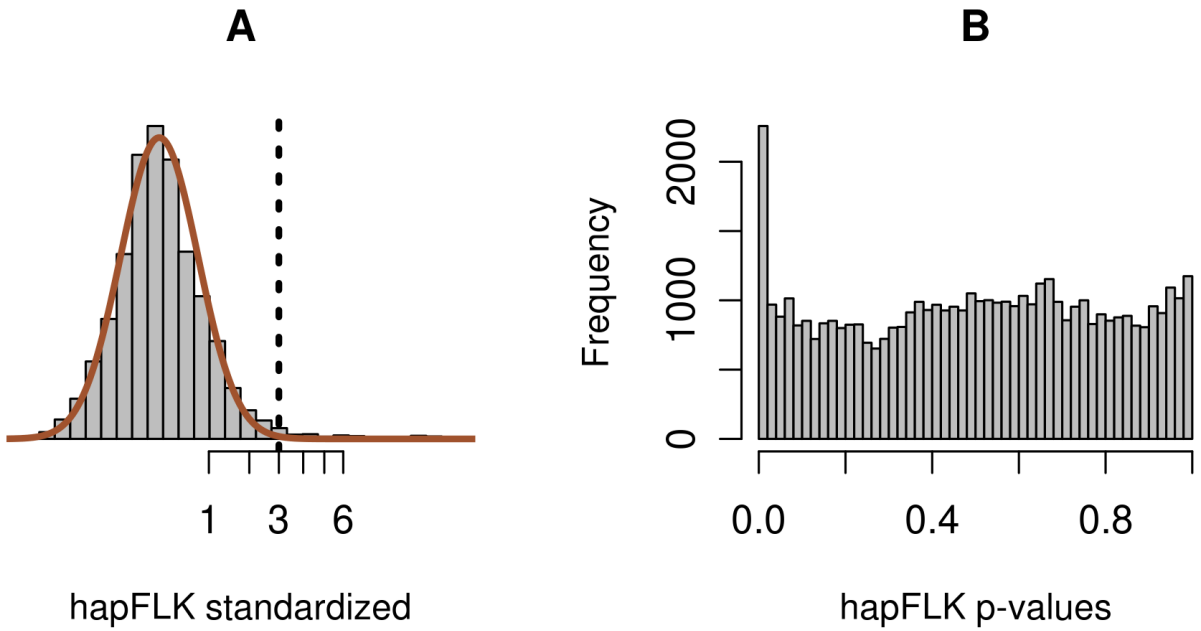


Focus on northern Europe





Computation of p-values

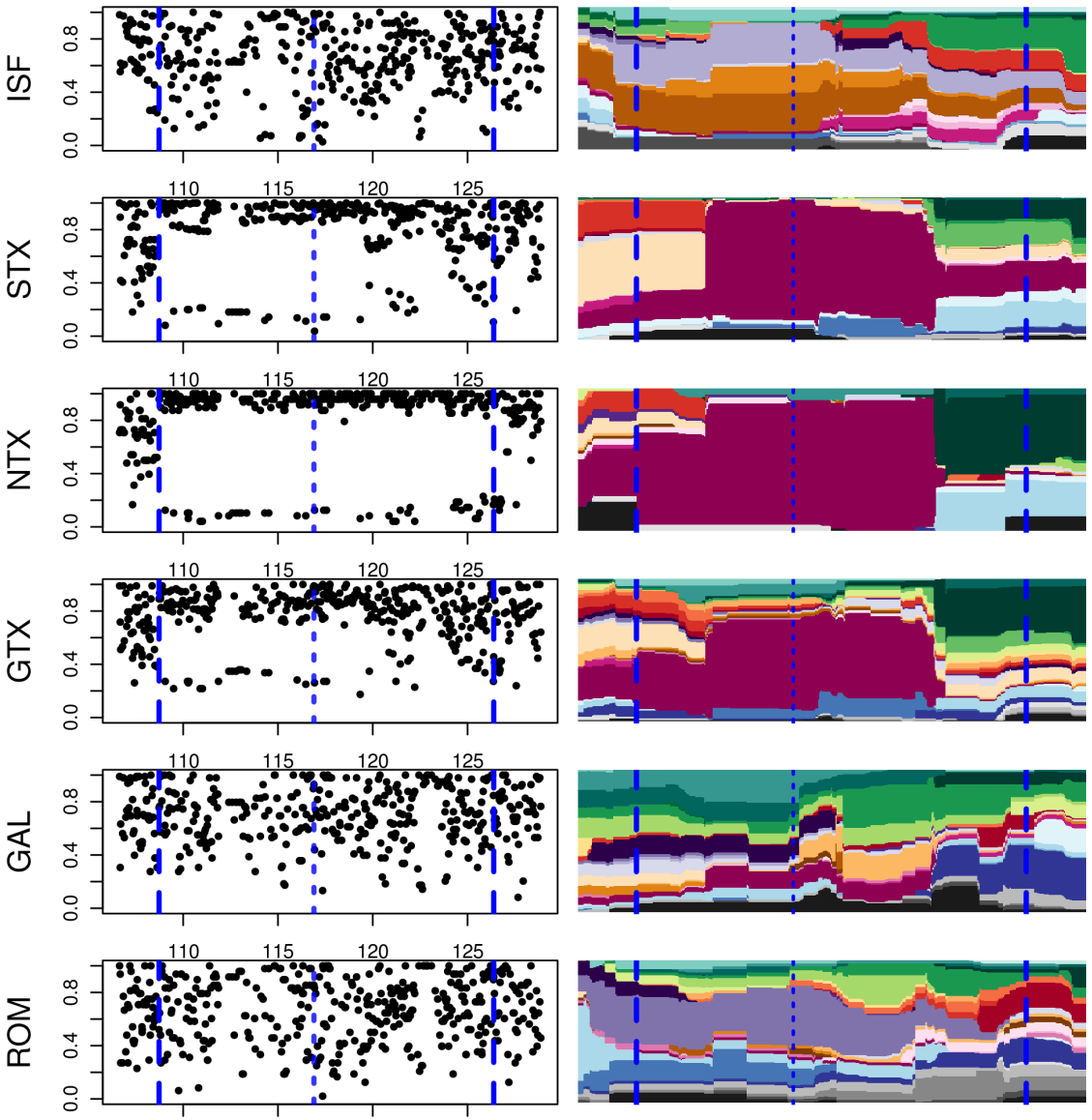


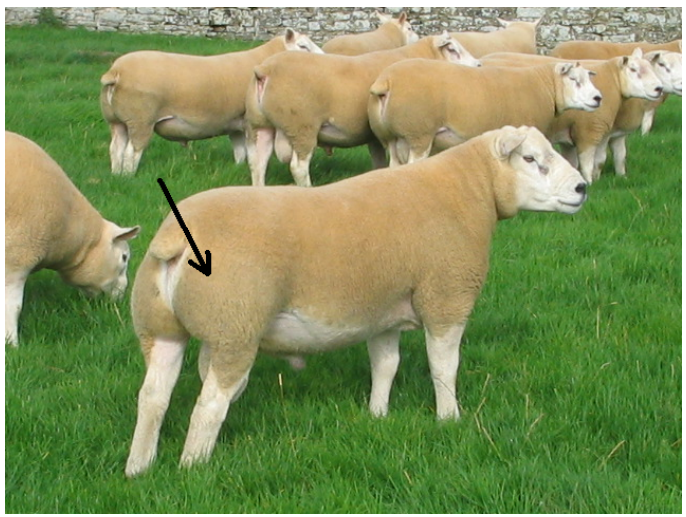
- *hapFLK* approximately gaussian, with some outliers.

$\Rightarrow$ Normalization using robust mean and variance estimates (function *rlm* in R)

Hard sweep signal in Texel Sheep

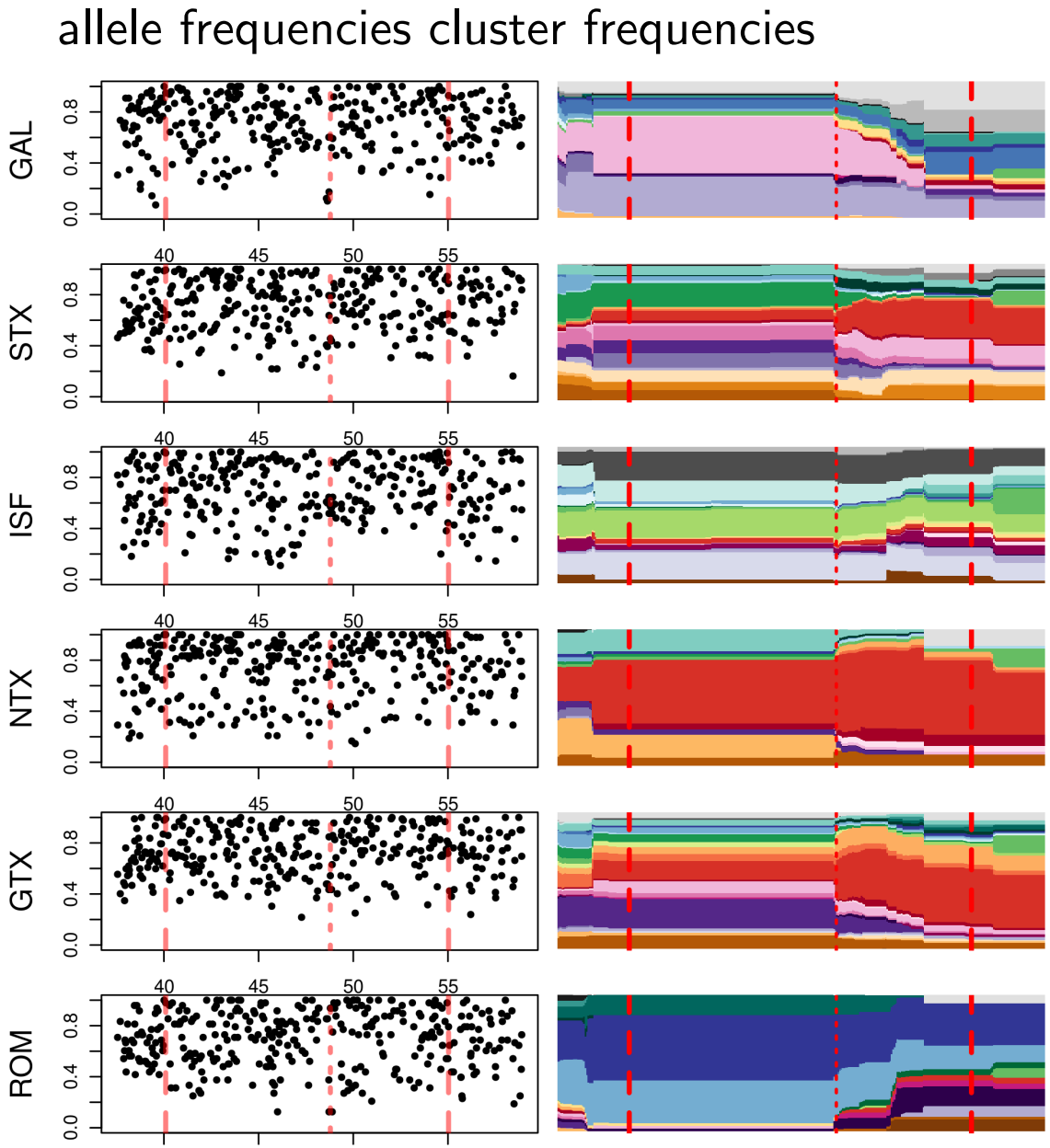
allele frequencies cluster frequencies





Candidate mutation in MSTN

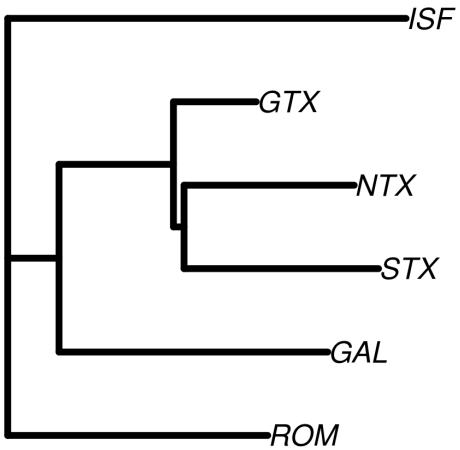
Soft / Incomplete sweep signal in New Zealand Sheep



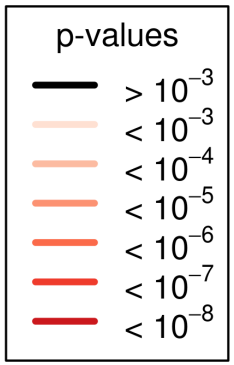
- One cluster at relatively high frequency in New Zealand Texel (NTX)
- Two clusters at fixation in New Zealand Rommey (ROM).

New Zealand breeds are the ones under selection

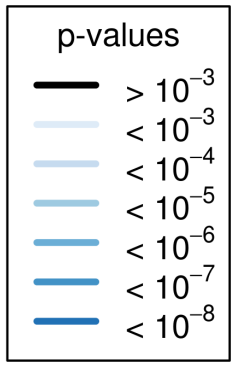
Whole genome tree



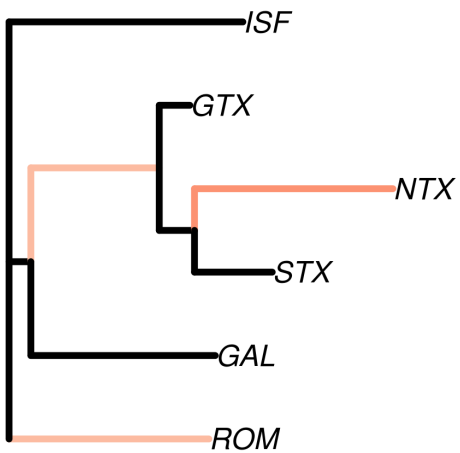
SNP trees

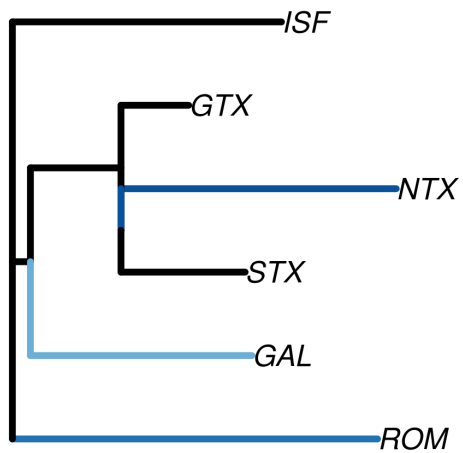


Haplotype trees



OAR 14

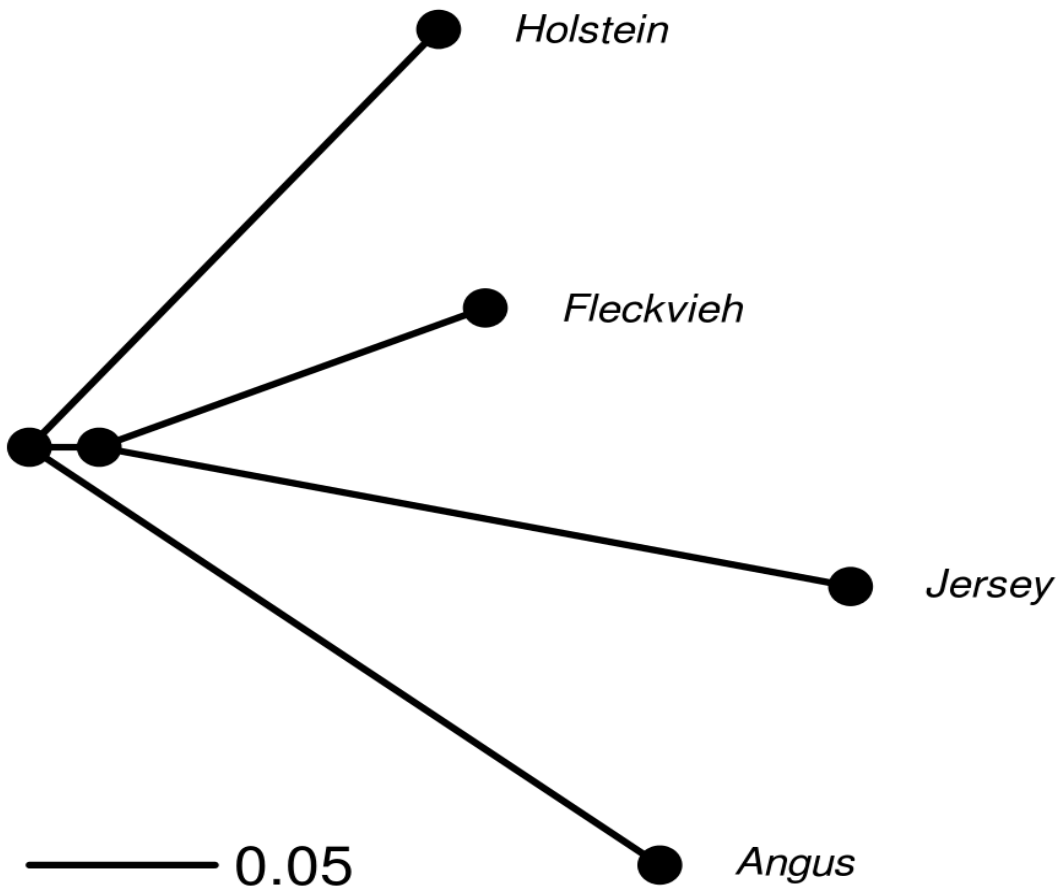




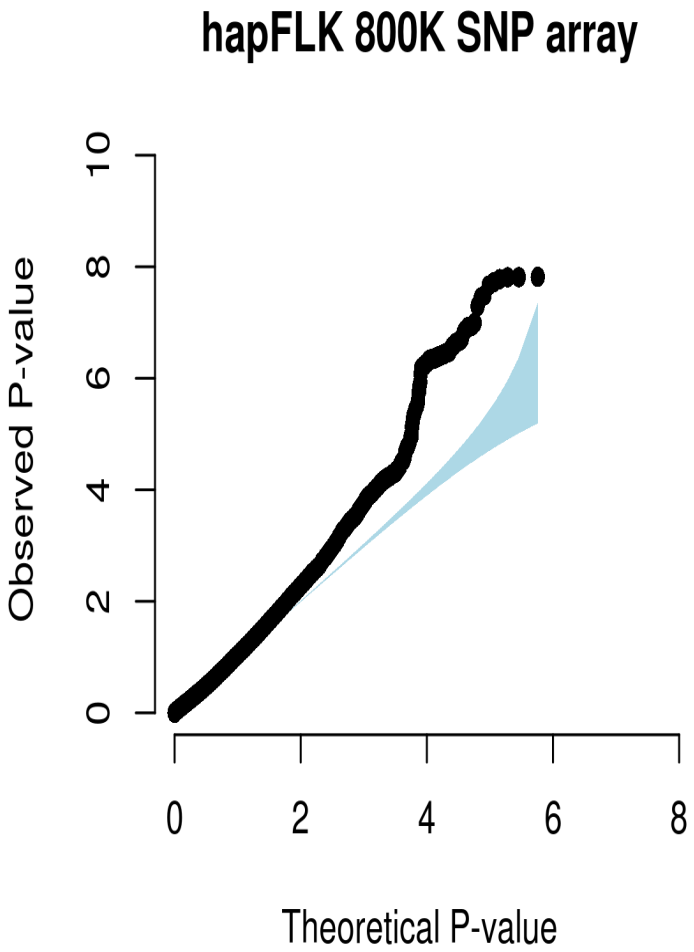
Outline

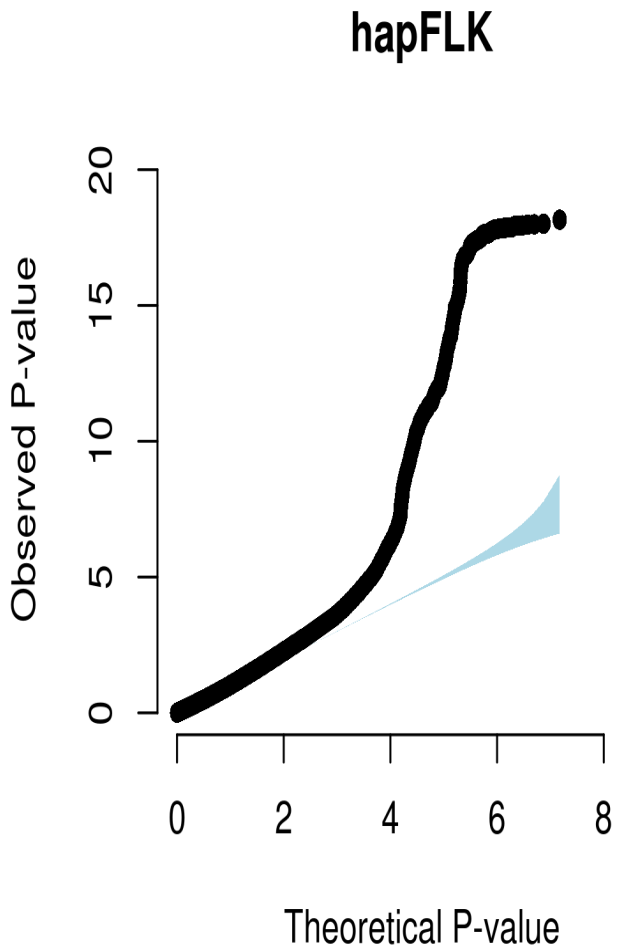
The 1000 bulls genome project, run2

- 234 sequences from 4 breeds (90 used).
- 29 millions bi-allelic variants (SNPs and indels)

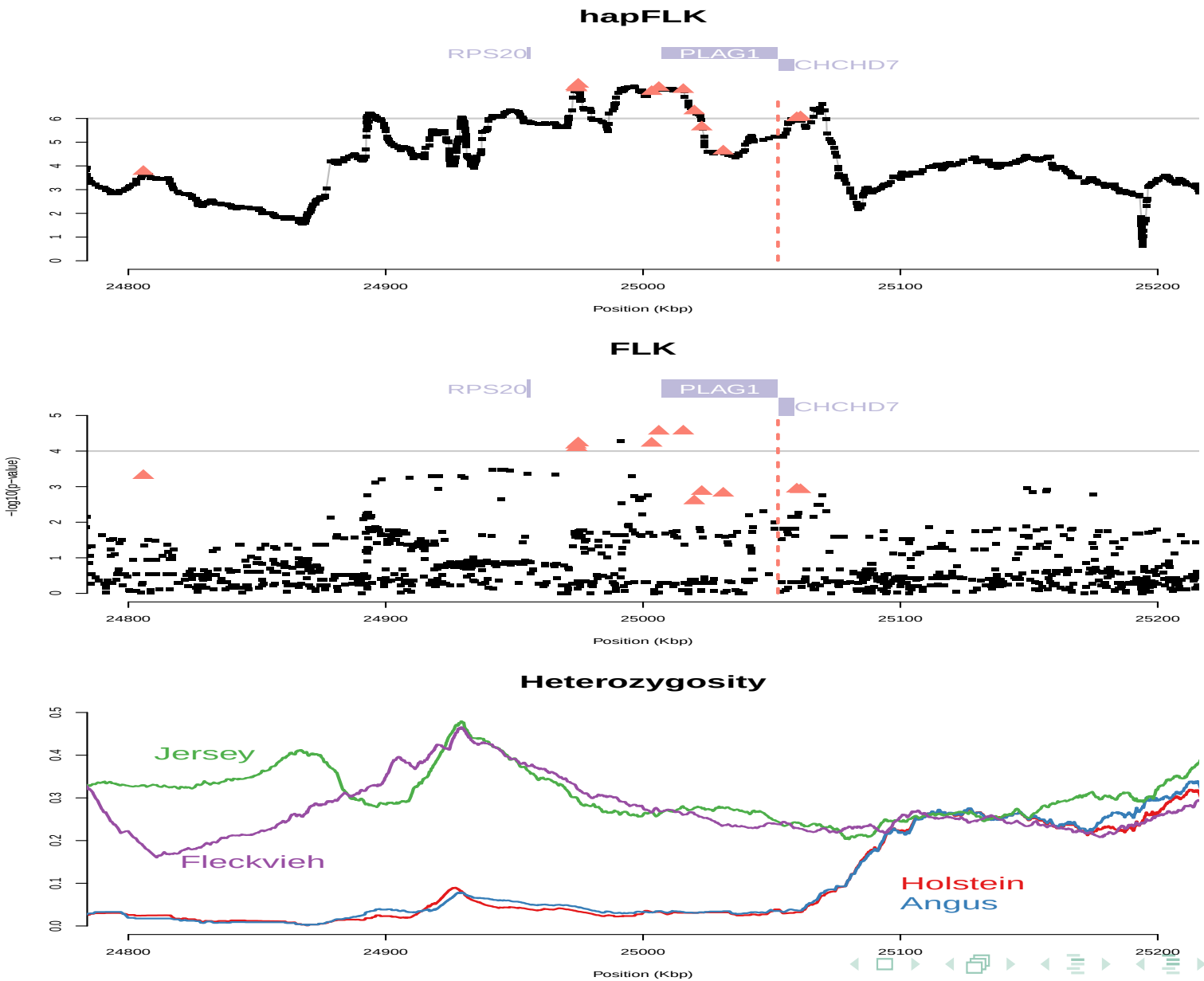


Increased power from NGS data





PLAG1 region - locate causal mutation



Outline

Conclusion

- The hapFLK approach:
  - Detection of positive selection from multi-population samples.
  - Accounts for population size heterogeneity and hierarchical structure of populations.
  - Uses haplotype information.

  $\rightarrow$ increased detection power
- Other advantages:
  - No need for sliding window.
  - Run from unphased genotype data, missing data allowed.
  - Soft and incomplete sweeps can be detected.
- Limitations:
  - Pure drift model.
  - Outlier approach.

References

- Methods:
  - M. Bonhomme, C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, M. San Cristobal (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186: 241-26.
  - M. I. Fariello, S. Boitard, H. Naya, M. SanCristobal, B. Servin (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193: 929-941.
- Applications:
  - M.I. Fariello, B. Servin, G. Tosser-Klopp, R. Rupp, C. Moreno, International Sheep Genome Consortium, M. San Cristobal and S. Boitard (2014). Selection Signatures in Worldwide Sheep Populations. *PLoS ONE* 9(8), e103813.
  - P.F. Roux, S. Boitard, Y. Blum et al (2015). Combined QTL and Selective Sweep Mappings with Coding SNP Annotation and Cis-eQTL Analysis Revealed PARK2 and JAG2 as New Candidate Genes for Adiposity Regulation. *G3* 5(4) 517-529.
  - S. Boitard, M. Boussaha, A. Capitan, D. Rocha and B. Servin (2016). Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. *Genetics* 203: 433-450.

Outline

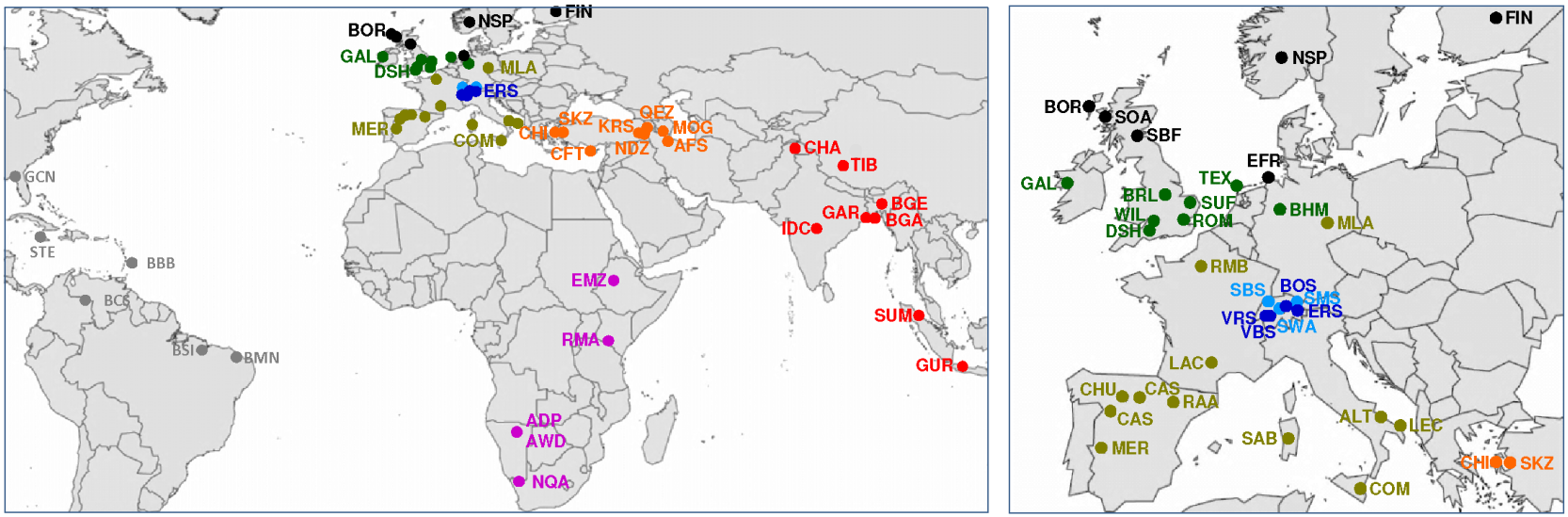
The `hapflk` software

- Builds and plots population trees (using `R` packages)
- Performs FLK and hapFLK tests on SNP data
- Computes associated p-values
- Plots cluster frequencies

Technical information

- Webpage: `https://forge-dga.jouy.inra.fr/projects/hapflk`
- Free software (GPL).
- Binaries for Linux 64bits and MacOSX 10.6+
- Source package available (requires `python`, `numpy`, `scipy`, `C` compiler).
- Other needs for additional routines: `R` with `ape`, `phangorn` and `ggplot2` packages, `python statsmodels` package.

Example data: sheep from Northern Europe



Kijas *et al.* (2012) PLoS Biology
6 Populations + Outgroup (Soay), 388 individuals, 50K SNPs
Available at `http://www.sheephapmap.org`

Before diving in ...

- Remember assumptions underlying the neutral model:
  - Population tree
  - Pure drift model (no mutations, no admixture)
  - Small F_i (say < 0.2)
- This means
  - Discard strongly bottleneck-ed or admixed populations
  - Discard low frequency variants (at the meta-population level), likely to have appeared after population spit.
- Perform a diversity analysis before:
  - Population structure (STRUCTURE, PCA, treemix ...) to remove outliers.
  - Within population kinship between individuals to identify a set of "unrelated" individuals

Main steps of the analysis

1. Compute kinship matrix and FLK.
2. Perform hapFLK test chromosome by chromosome.
3. Merge hapFLK results in a single file.
4. Compute hapFLK p-values.
5. Call significant regions and plot p-values.
6. Annotate interesting regions (e.g. myostatin region on chr 2).
   - Plot haplotype cluster frequencies.
   - Plot local population trees.