Riding the whole-genome data tsunami: a landscape genomic study of local adaptation in Moroccan sheep and goats

Sylvie Stucki, Kevin Leempoel & Stéphane Joost
& the NEXTGEN Consortium

Laboratory of Geographic Information Systems (LASIG)

June 18, 2014

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

nextgen

Outline

Introduction

Small ruminants in Morocco

- Key importance of local breeds for population livelihood
- NEXTGEN project: Whole-genome sequencing techniques $\Rightarrow$ improve sustainable breeding practices
- Local adaptation in Moroccan sheep and goats

Landscape genomics

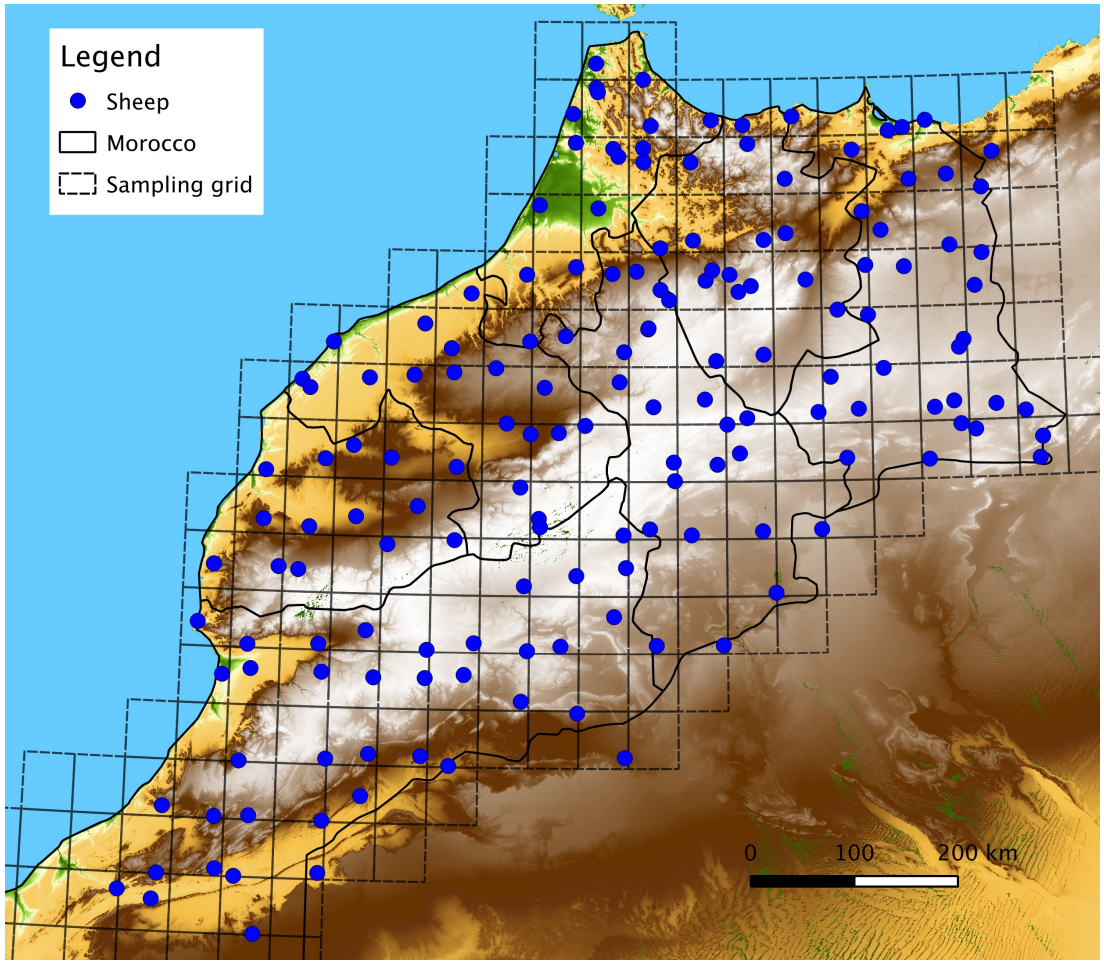
- Individuals are adapted to their habitat
- Detect selection signatures using genome-environment associations

Whole genome sequencing

- Computational workload

Outline

Sheep samples to be sequenced



Molecular and environmental data

Whole-genome sequencing

Sheep

160 samples
40.7M SNPs
2.8M indels

Goats

161 samples
29.6M SNPs
2.1M indels

Molecular and environmental data

Whole-genome sequencing

Sheep

160 samples
40.7M SNPs
2.8M indels

Goats

161 samples
29.6M SNPs
2.1M indels

Climate and topography-related data

WorldClim

monthly values for
precipitation and temperature

Shuttle Radar Topography Mission (SRTM)

Slope, curvature, day duration,
total insolation (solstices), . . .

Molecular and environmental data

Whole-genome sequencing

Sheep

160 samples
40.7M SNPs
2.8M indels

Goats

161 samples
29.6M SNPs
2.1M indels

Climate and topography-related data

WorldClim

monthly values for precipitation and temperature

Looking at correlations

Shuttle Radar Topography Mission (SRTM)

Slope, curvature, day duration, total insolation (solstices), ...

15 environmental variables

Outline

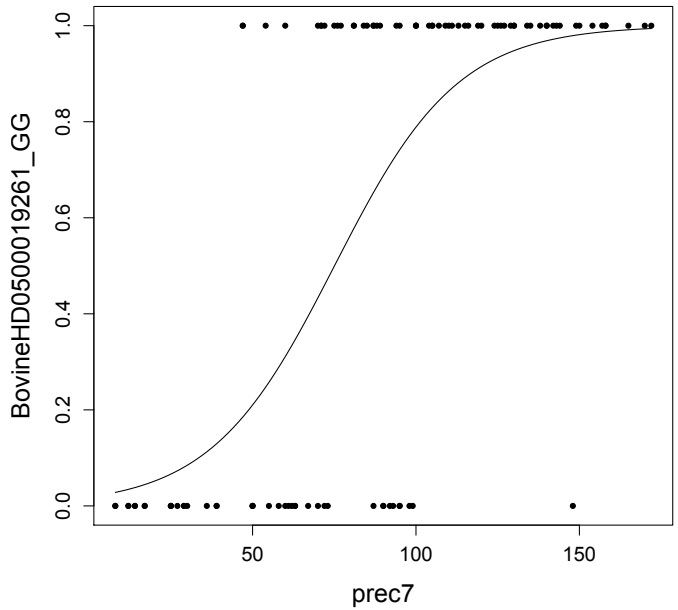
Detection of selection signatures

Logistic regressions

Maximum likelihood

G and Wald tests

Bonferroni correction



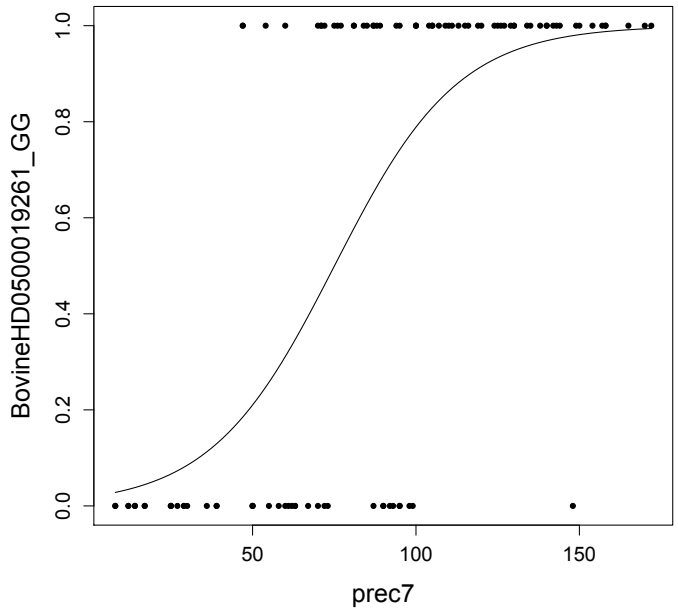
Samβada

Logistic regressions

Maximum likelihood

G and Wald tests

Bonferroni correction



`lasig.epfl.ch/sambada`

Samβada

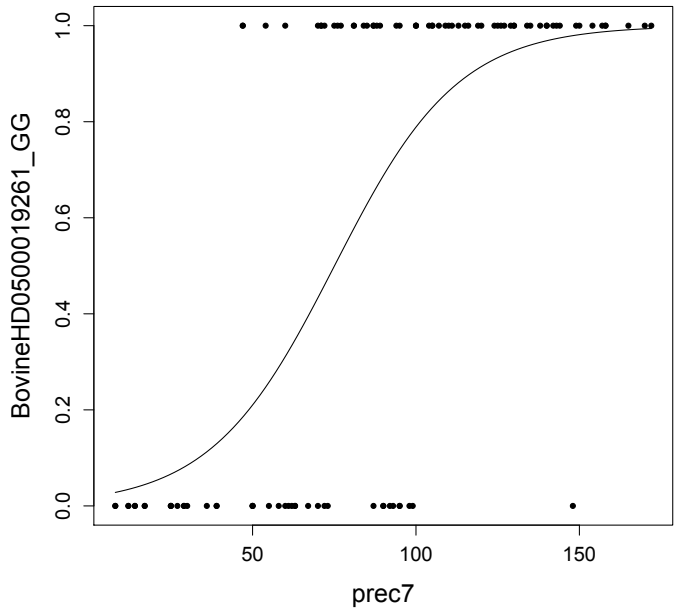
Logistic regressions

Maximum likelihood

G and Wald tests

False discovery rate

(Storey, 2003)



`lasig.epfl.ch/sambada`

Samβada

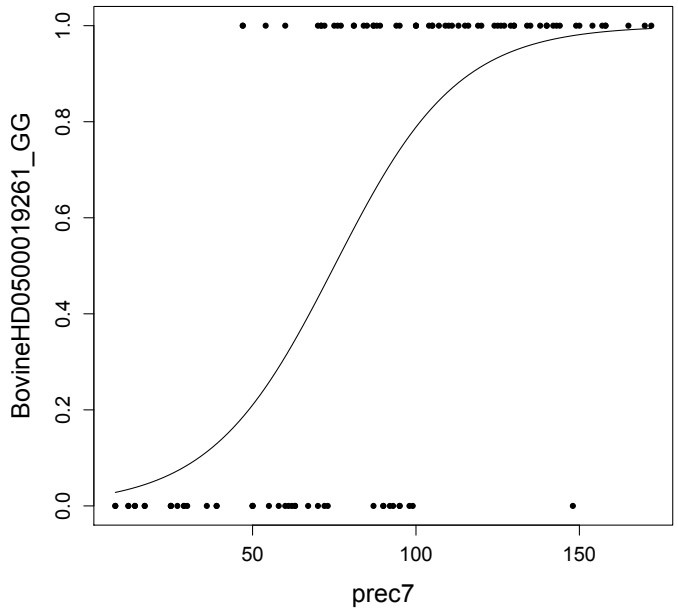
Logistic regressions

Maximum likelihood

G and Wald tests

False discovery rate

(Storey, 2003)



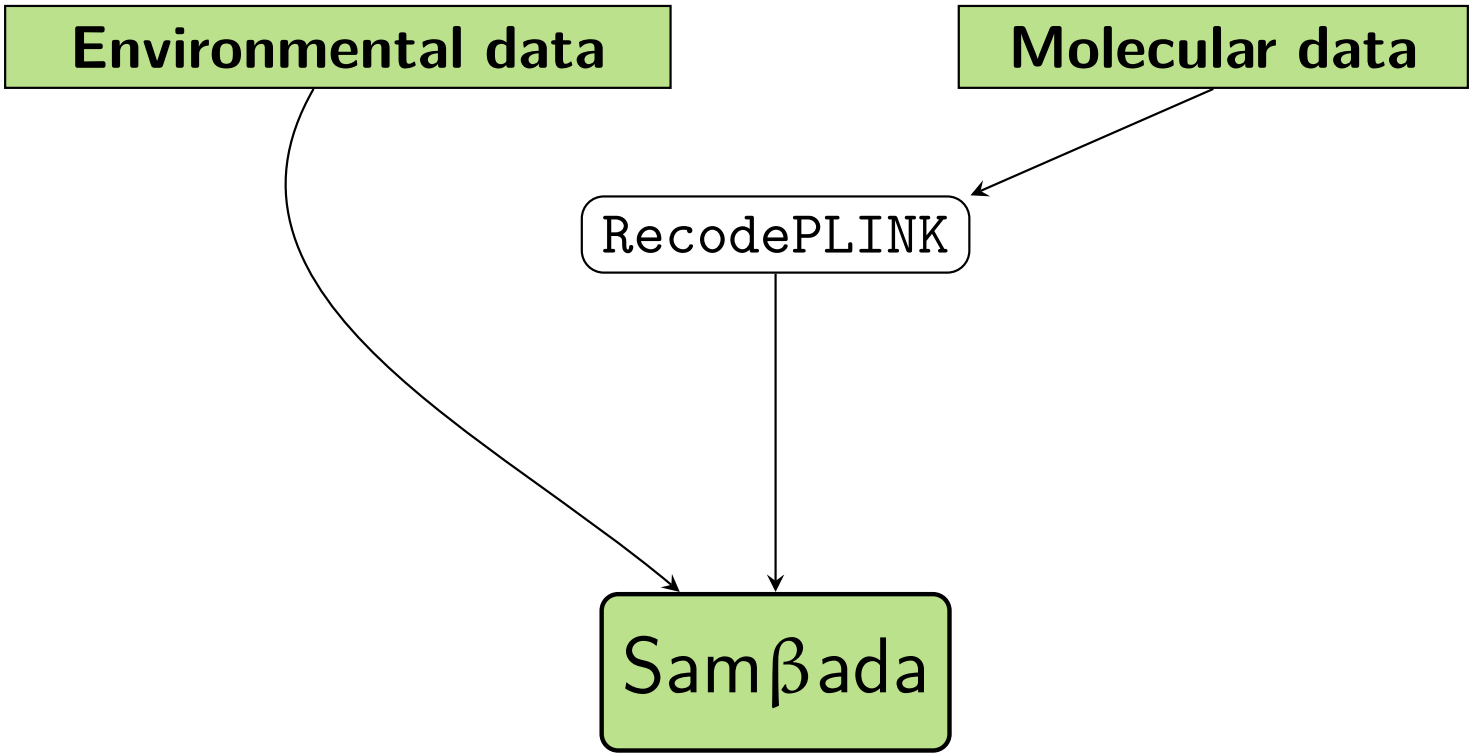
Multivariate analysis

Spatial autocorrelation

`lasig.epfl.ch/sambada`

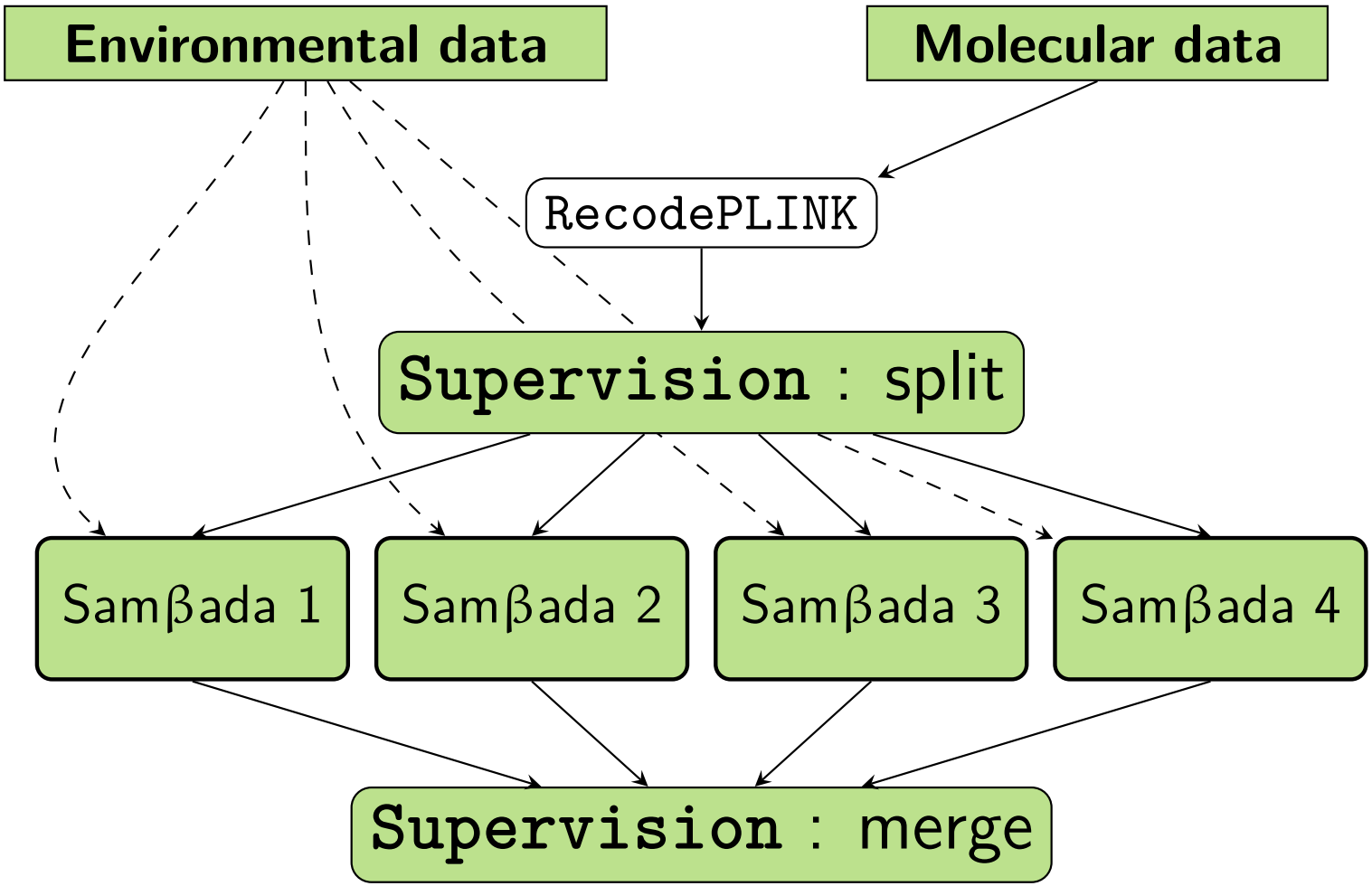
Samβada's workflow

One process



Samβada's workflow

Distributed computing



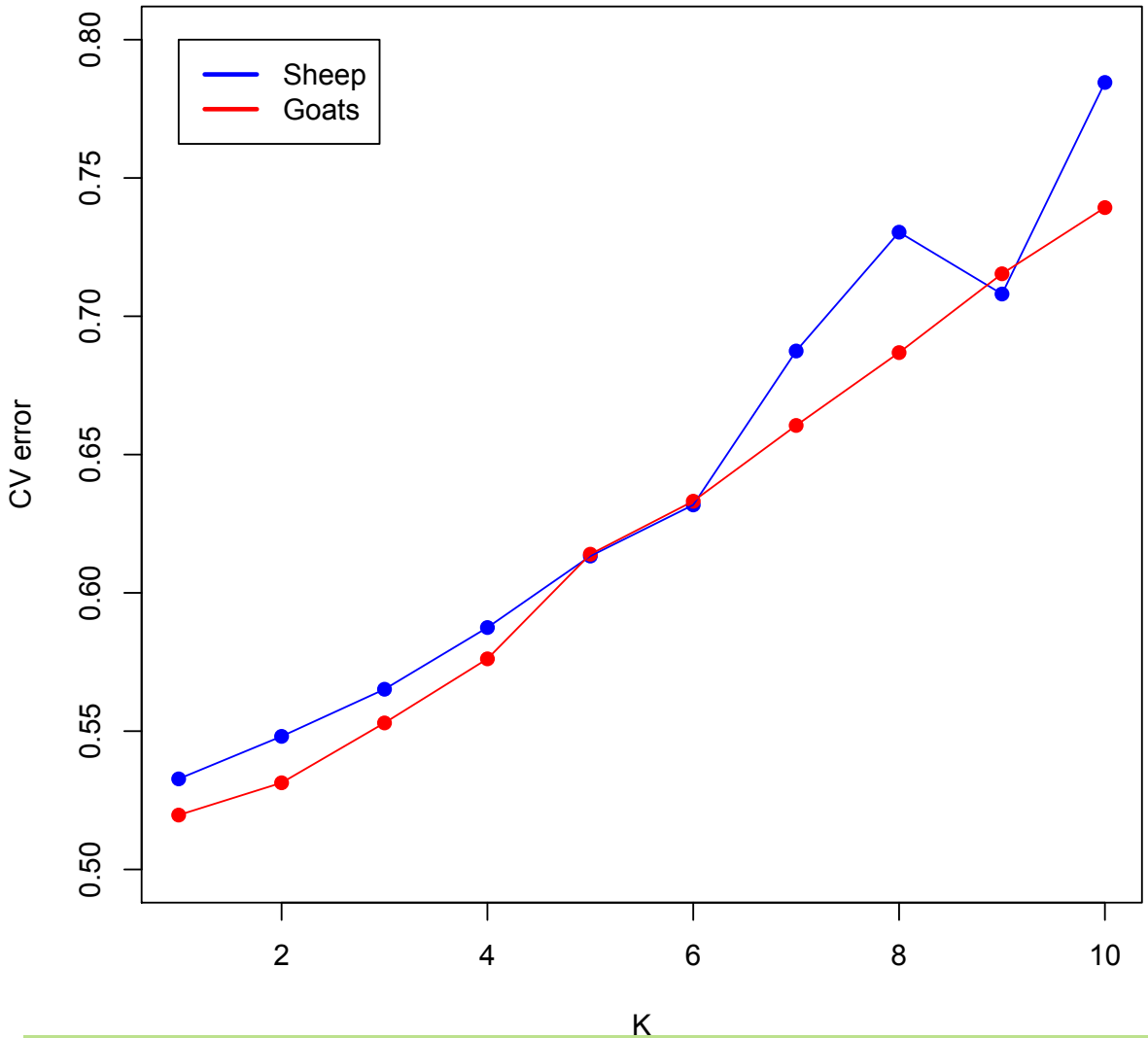
Overview of analysis

1. Prune markers so LD$\leq$0.2
   `PLINK 1.9`: –indep-pairwise 50 5 0.2
2. Prune markers for loci and individual call rates and MAF
   `PLINK 1.9`: –geno=0.05 –mind=0.05 –maf 0.05
3. Analyse population structure with `Admixture`
4. Remove chromosomes X, Y

   sheep 1'799'364 markers (SNPs and indels)
   goats 1'757'210 markers (SNPs and indels)
5. Recode markers for Samβada
   biallelic markers, recoded as {A, G} SNPs.

Outline

Population structure (`Admixture`; Alexander, 2009)



FDR selection of models

Sheep

Marker	Env.	G score	Wald score	q-value
23:43794976_GG	prec_3	44.30	24.29	0.0011
23:43812782_GG	prec_3	44.30	24.29	0.0011
23:43874160_GG	prec_3	42.56	23.39	0.0017
1:38304177_GG	bio_15	37.37	18.29	0.0185
23:43847594_GG	prec_3	33.73	23.09	0.0957
7:48256781_GG	bio_15	32.00	18.87	0.1669
7:48262822_GG	bio_15	32.00	18.87	0.1669
23:43861704_GG	prec_3	31.53	21.43	0.1855
1:190582_AA	tmean_7	30.68	18.69	0.2501
1:137076394_GG	tmean_7	30.52	17.00	0.2501

FDR selection of models

Goats

Marker	Env.	G score	Wald score	q-value
24:19436980_GG	bio_15	38.00	26.21	0.0197
6:12259667_AA	bio_15	37.78	25.96	0.0197
6:12254244_AA	bio_15	37.78	25.96	0.0197
20:4481114_GG	bio_7	34.48	19.65	0.0803
6:12242353_GG	bio_15	33.07	24.82	0.1322
9:14309947_GG	bio_7	32.55	19.19	0.1322
2:133961081_GG	tmean_7	32.33	22.90	0.1322
6:47914533_AA	prec_3	32.16	16.84	0.1322
11:15823825_GG	bio_7	31.93	23.69	0.1322
4:95035251_GG	bio_15	31.52	20.56	0.1471

Outline

What about the WGS analysis?

You promised!

Everything was prepared. . .

- Data was split by chromosome
- Automation scripts were designed
- Analysis was run successfully

What about the WGS analysis?

You promised!

Everything was prepared...

- Data was split by chromosome
- Automation scripts were designed
- Analysis was run successfully

Monday, 6pm (Lausanne time)

MISTAKE spotted!

What about the WGS analysis?

You promised!

What about the WGS analysis?

You promised!

Monday, 6pm - Thuesday, 2pm (Lausanne time)

- Correct mistake
- Run analysis on pruned dataset
- Launch analysis on whole dataset

What about the WGS analysis?

You promised!

Monday, 6pm - Thuesday, 2pm (Lausanne time)

- Correct mistake
- Run analysis on pruned dataset
- Launch analysis on whole dataset

Other tasks

- Online C++ assessment
- Job interview

What about the WGS analysis?

You promised!

Wednesday, 5pm (Caerdydd time)

WGS Analysis just finished :-)

What about the WGS analysis?

You promised!

Wednesday, 5pm (Caerdydd time)

WGS Analysis just finished :-)

Computation time

Sheep

2M SNPs $\times$ 15 env. var.	28M SNPs $\times$ 15 env. var.
1 desktop computer, 8 cores	2 desktop computers, 8 cores
$\sim$ 2 hours	$\sim$ 18 hours

The story so far...

No population structure in Moroccan sheep and goats

- Use simple landscape genomic models

Samβada can analyse WGS data

- Signal of selection is weak
- Maybe adapt our method?

Next steps

- Map detections on the genome
- Analyse local spatial autocorrelation
- Compare sheep and goats

Thank you!

sylvie.stucki@a3.epfl.ch

Appendix

False discovery rate

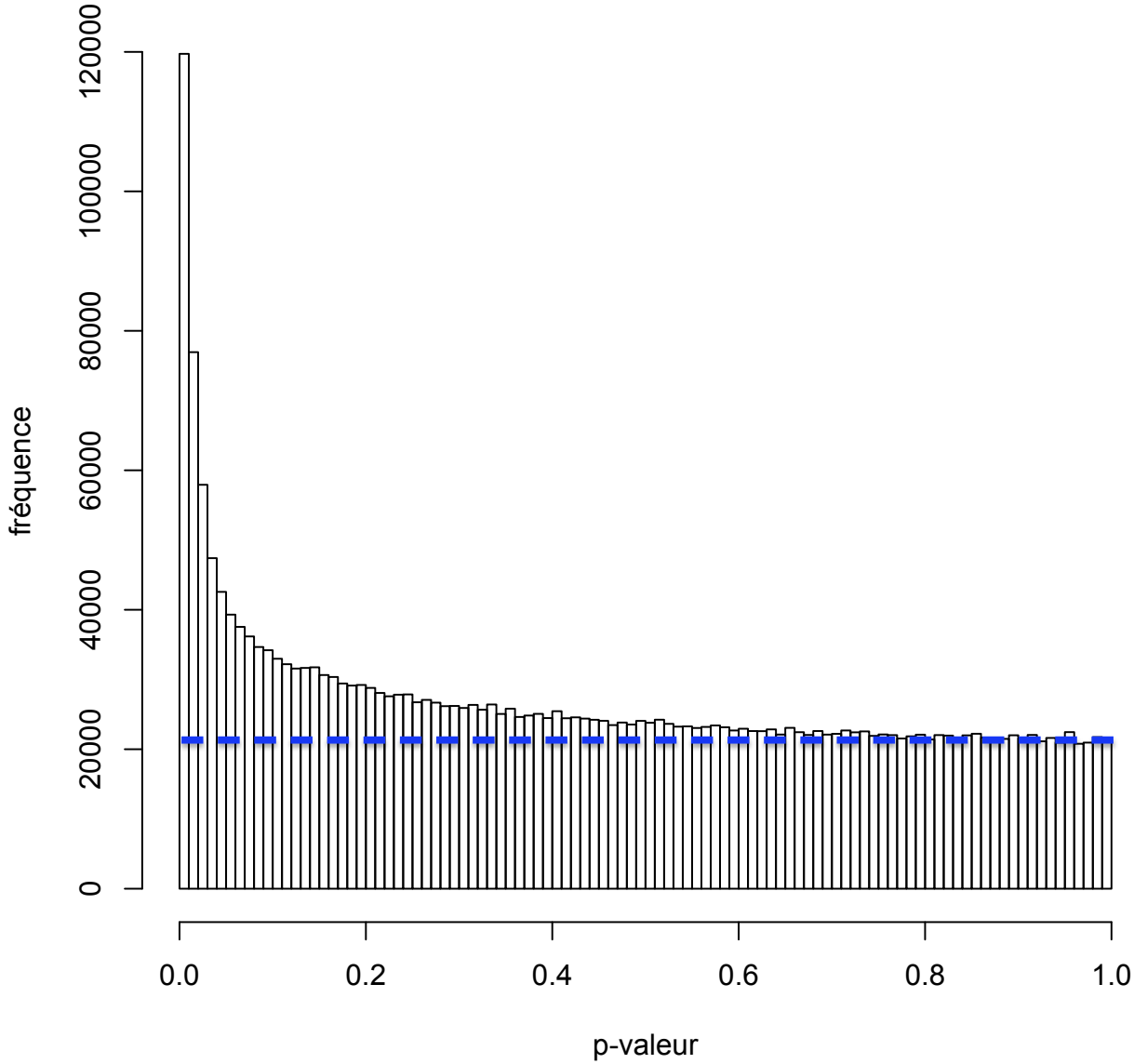
Outline

False discovery rate according to Storey (2003)

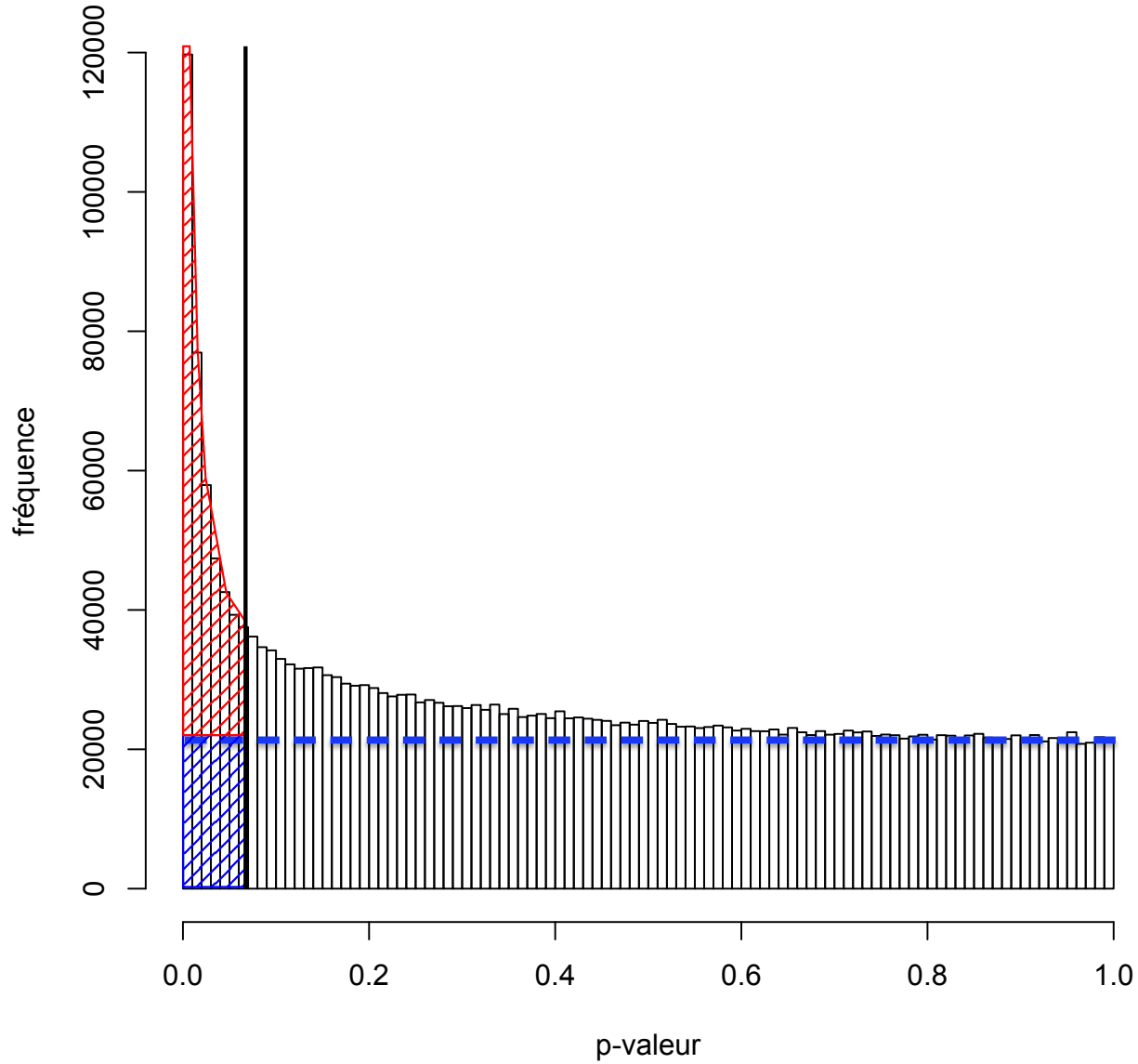
Neutral markers Uniform distribution of p-values

Markers under selection Small p-values

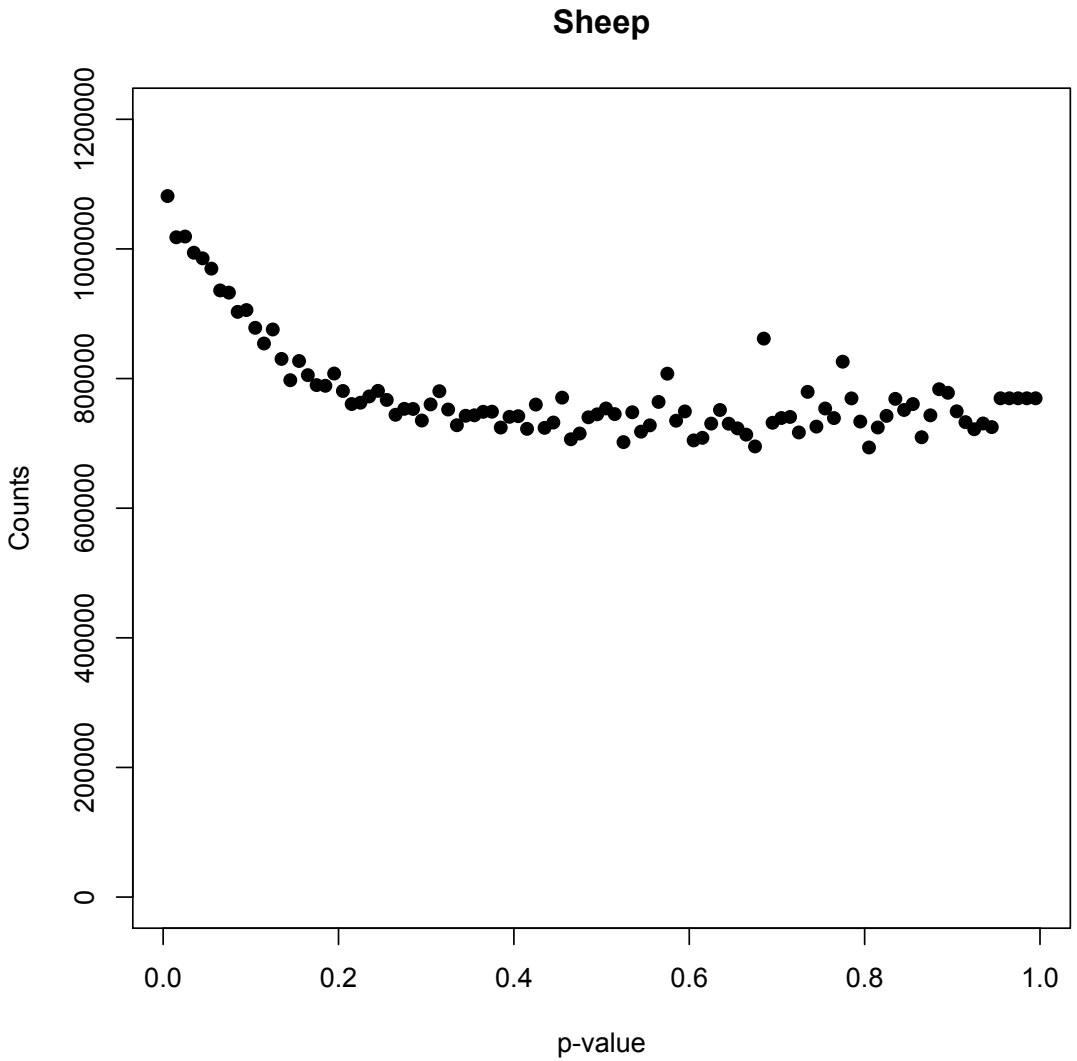
False discovery rate according to Storey (2003)



False discovery rate according to Storey (2003)



Samβada's results : p-values



Samβada's results : p-values