A hierarchical bayesian model for measuring the extent of local adaptation from haplotype data.

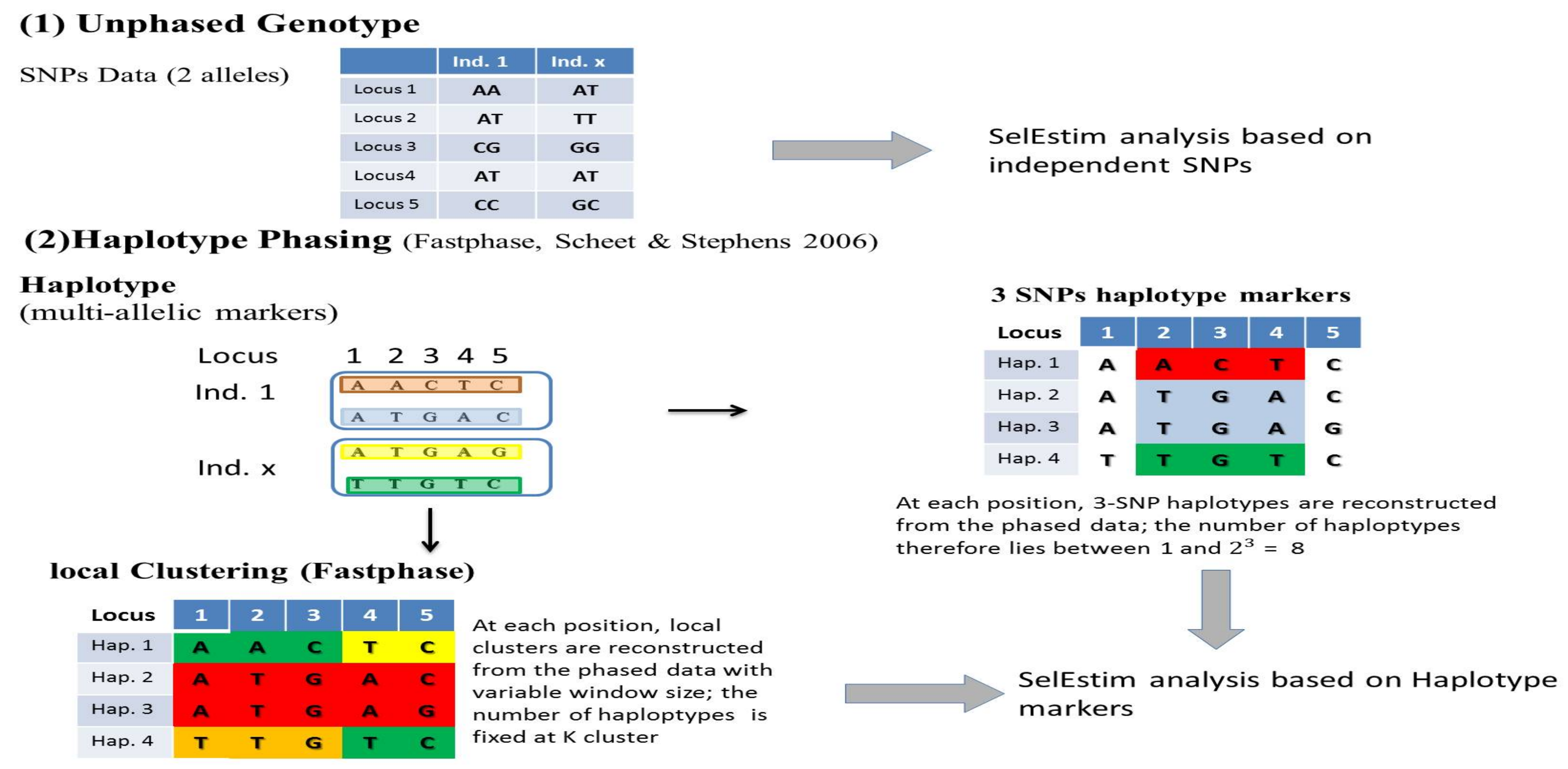
Valentin Hivert[1,2], Mathieu Gautier[1,2], and Renaud Vitalis[1,2]

[1]INRA, UMR CBGP, F-34988 Montferrier-sur-Lez, France
[2]Institut de Biologie Computationnelle, 34095 Montpellier, France

Introduction & Objectives

- The recent advent of high throughput sequencing and genotyping technologies (Next Generation Sequencing, NGS) enables the comparison of patterns of polymorphisms at a very large number of markers, which makes it possible to characterize genomic regions involved in the adaptation of organisms to their environment. Here, we present some recent developments to SelEstim (Vitalis et al. 2014), a hierarchical bayesian model that identifies and measures genomic signatures of selection from gene frequency data.
- we extend the model to analyse multi-allelic markers. Considering haplotype blocks as multi-allelic markers, this allows to account for the information brought by linkage desequilibrium.

Genetic data & Analysis



(1) Unphased Genotype

SNPs Data (2 alleles)

	Ind. 1	Ind. x
Locus 1	AA	AT
Locus 2	AT	TT
Locus 3	CG	GG
Locus4	AT	AT
Locus 5	CC	GC

SelEstim analysis based on independent SNPs

(2)Haplotype Phasing (Fastphase, Scheet & Stephens 2006)

Haplotype
(multi-allelic markers)

3 SNPs haplotype markers

Locus	1	2	3	4	5
Hap. 1	A	A	C	T	C
Hap. 2	A	T	G	A	C
Hap. 3	A	T	G	A	G
Hap. 4	T	T	G	T	C

At each position, 3-SNP haplotypes are reconstructed from the phased data; the number of haploptypes therefore lies between 1 and $2^3 = 8$

local Clustering (Fastphase)

Locus	1	2	3	4	5
Hap. 1	A	A	C	T	C
Hap. 2	A	T	G	A	C
Hap. 3	A	T	G	A	G
Hap. 4	T	T	G	T	C

At each position, local clusters are reconstructed from the phased data with variable window size; the number of haploptypes is fixed at K cluster

SelEstim analysis based on Haplotype markers

Fig. 1: Analysis pipeline. Genetic data were simulated from an island model with 8 demes of size $N = 1000$ and $F_{\mathrm{ST}} = 0.1$. Selection (s) is targeting a single position, with one allele selected for in 2 demes and the alternative allele selected for in 2 other demes. Three chromosomes of 5Mb were simulated with a 1cM/Mb recombination rate.

SelEstim

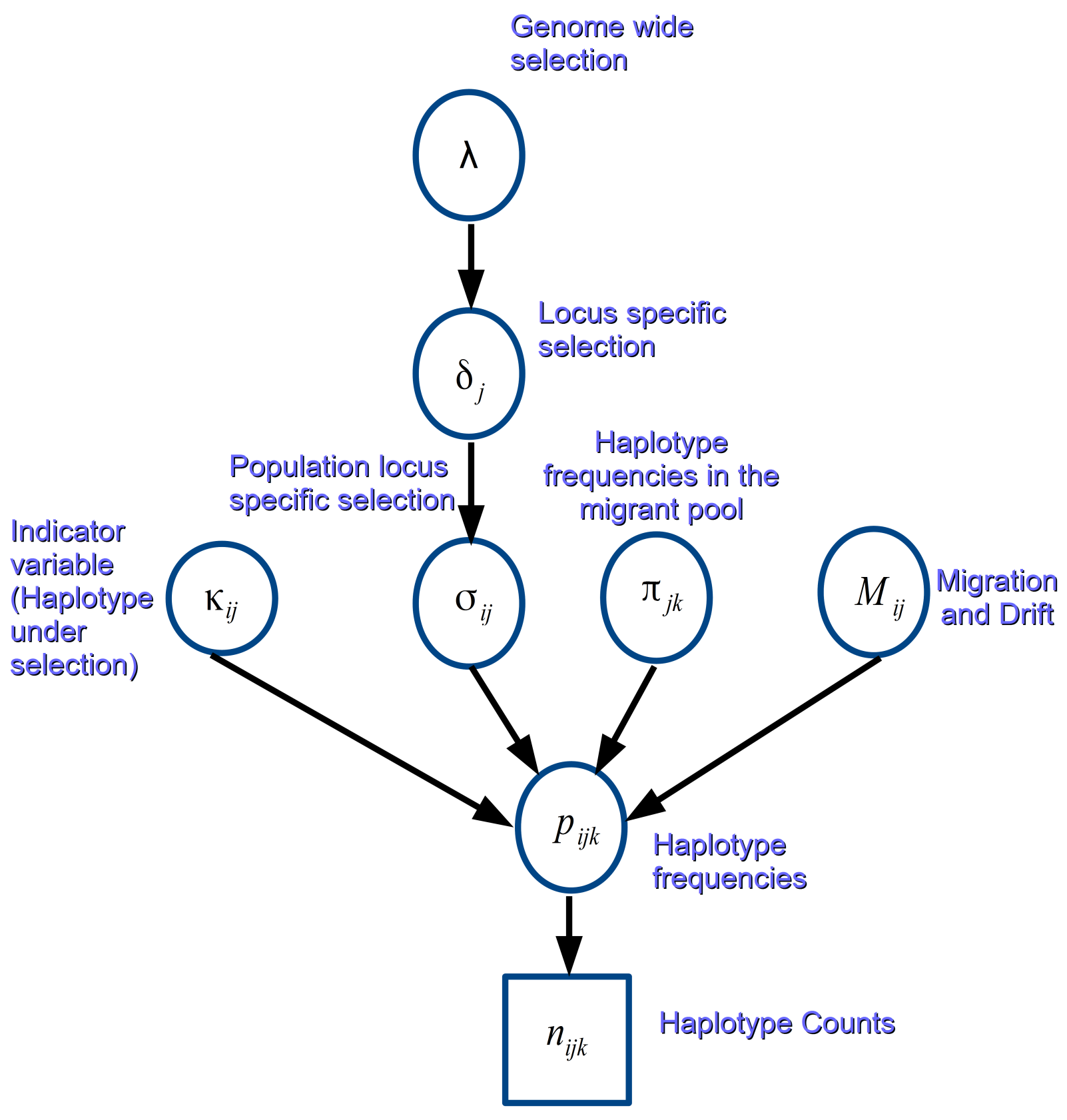


Fig. 2: Directed Acyclic Graph of SelEstim

Results

Locus-specific selection coefficient along the three simulated chromosomes with strong selection ($2Ns = 100$, see Fig. 3) and weaker selection ($2Ns = 50$, see Fig. 4). Results are given for analyses with bi-alellic data (top), 3-SNP haplotypes (middle) and local clustering (bottom). The position targeted by selection is indicated with a red arrow.

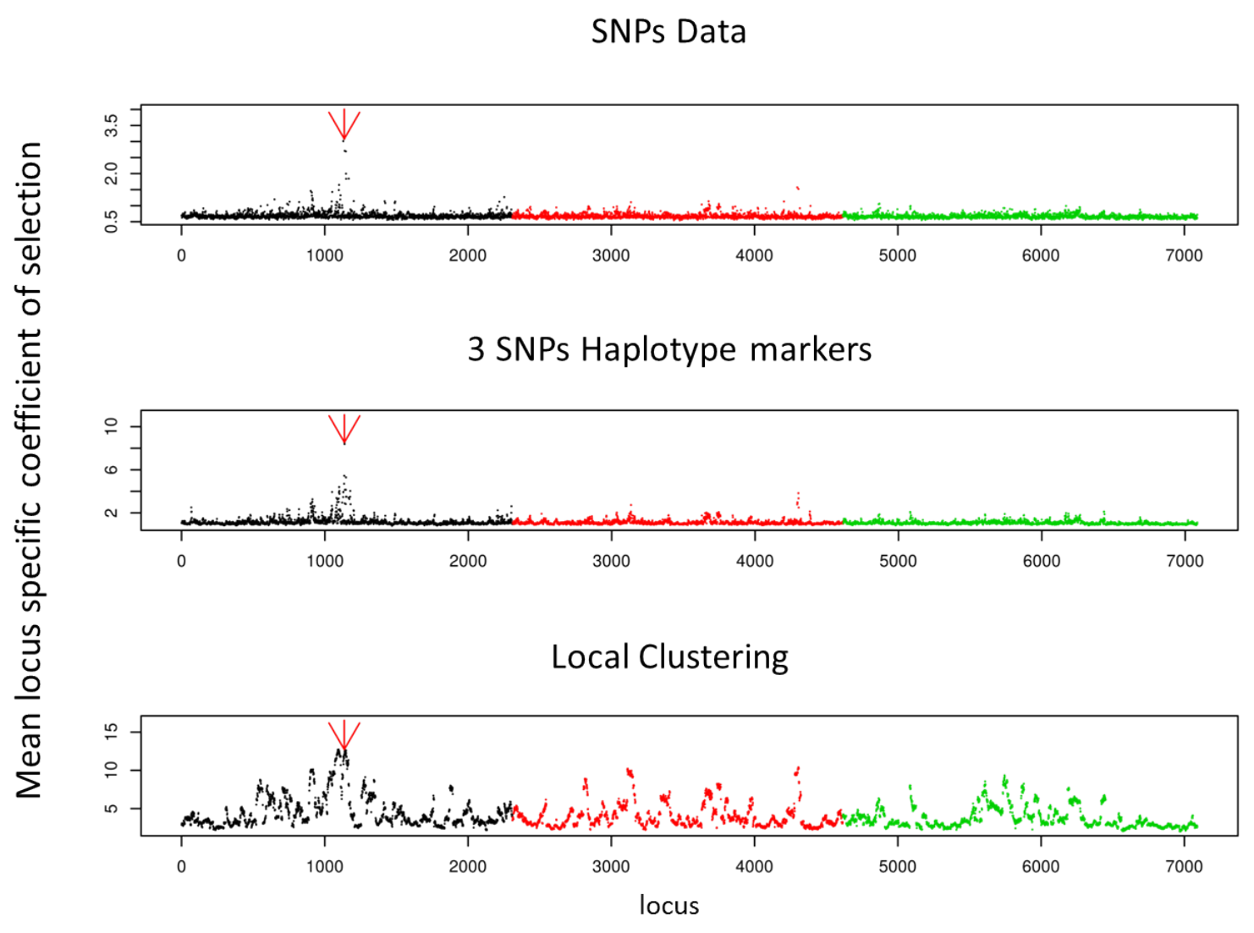


Fig. 3: Mean locus specific selection coefficient for simulation with strong selection ($2Ns = 100$)

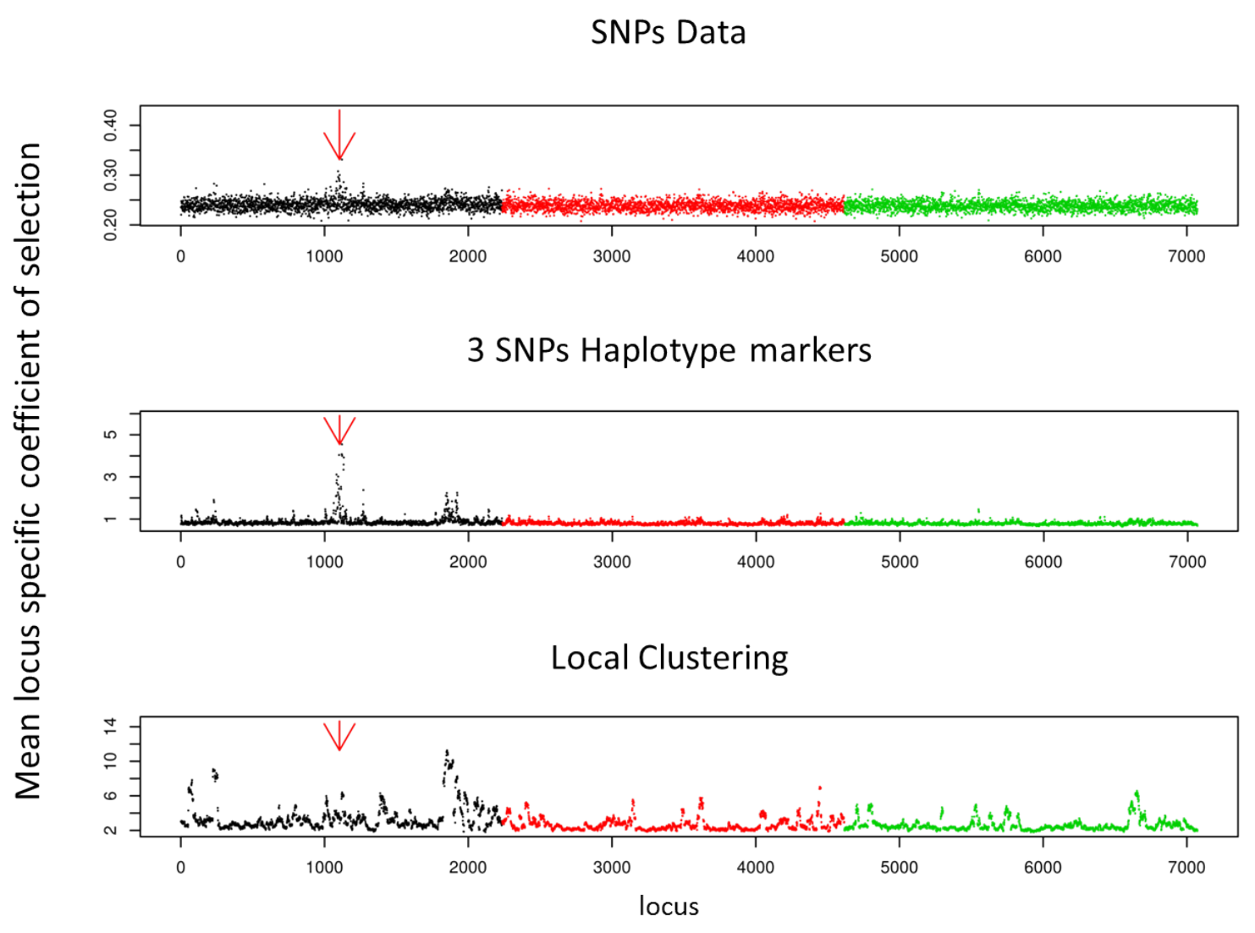


Fig. 4: Mean locus specific selection coefficient for simulation with weak selection ($2Ns = 50$)

Application example on human data

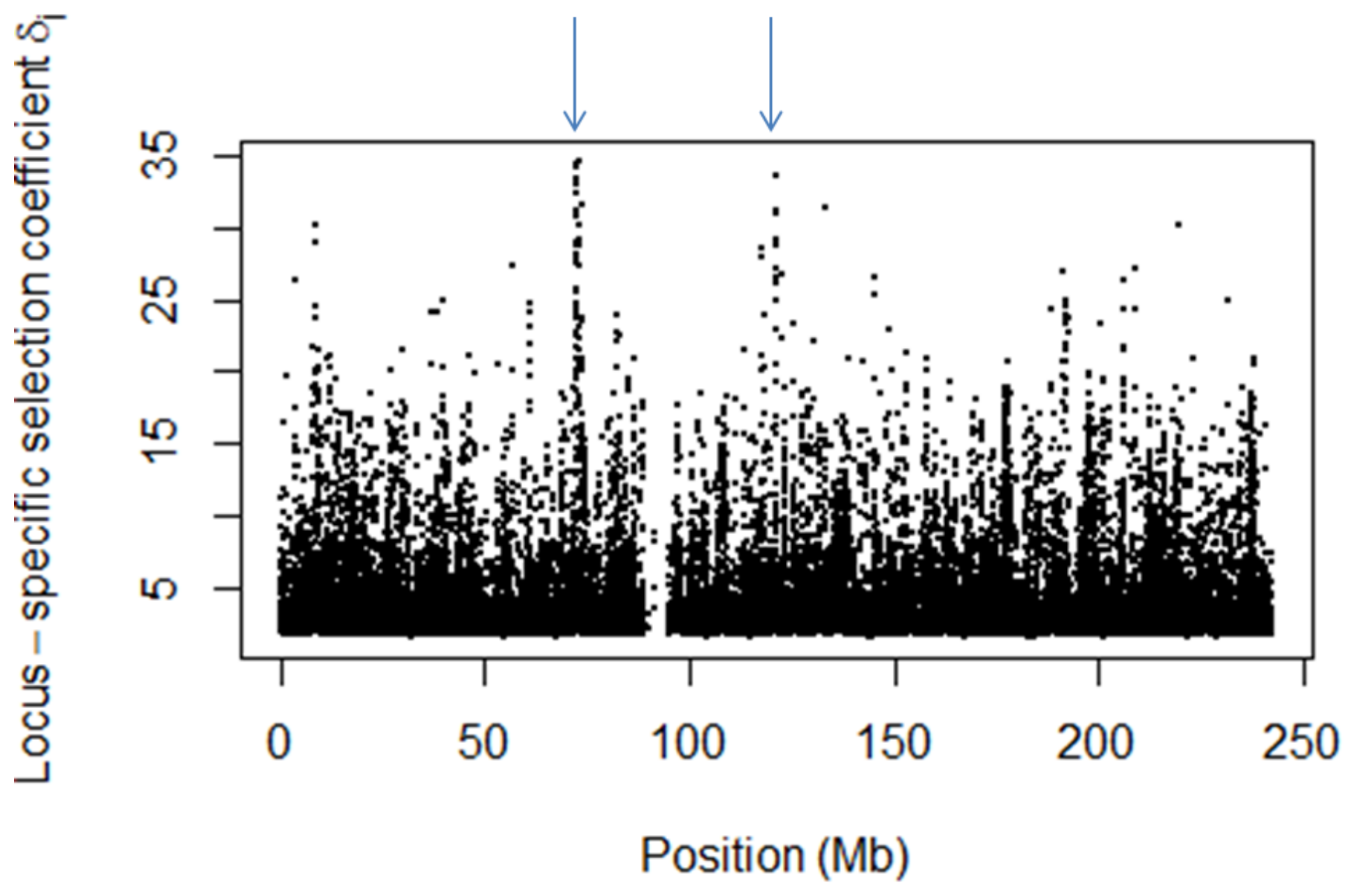


Fig. 5: SelEstim analysis of HapMap phased data for nine worldwide populations. The analysis was conducted for chromosome 2, with 49,906 SNPs recoded as 3-SNP haplotypes.

Conclusion

- Linkage disequilibrium (LD) information brought by haplotype data increases the power to detect genomic regions targeted by selection
- 3-SNP haplotypes seem more efficient to capture LD information than local clustering of haplotypes

Acknowledgements

Valentin Hivert's PhD is funded by the ERA-Net BiodivERsA project EXOTIC and the French National Institute for Agricultural Research (”Plant Health and Environment” division)

References

Scheet, P. and M. Stephens (2006). “A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.” In: *American journal of human genetics* 78.4, pp. 629–644.
Vitalis, R. et al. (2014). “Detecting and measuring selection from gene frequency data”. In: *Genetics* 196.March, pp. 799–817.

LaTeX TikZposter