A model-based approach to characterize individual autozygosity at both global and local genomic scale

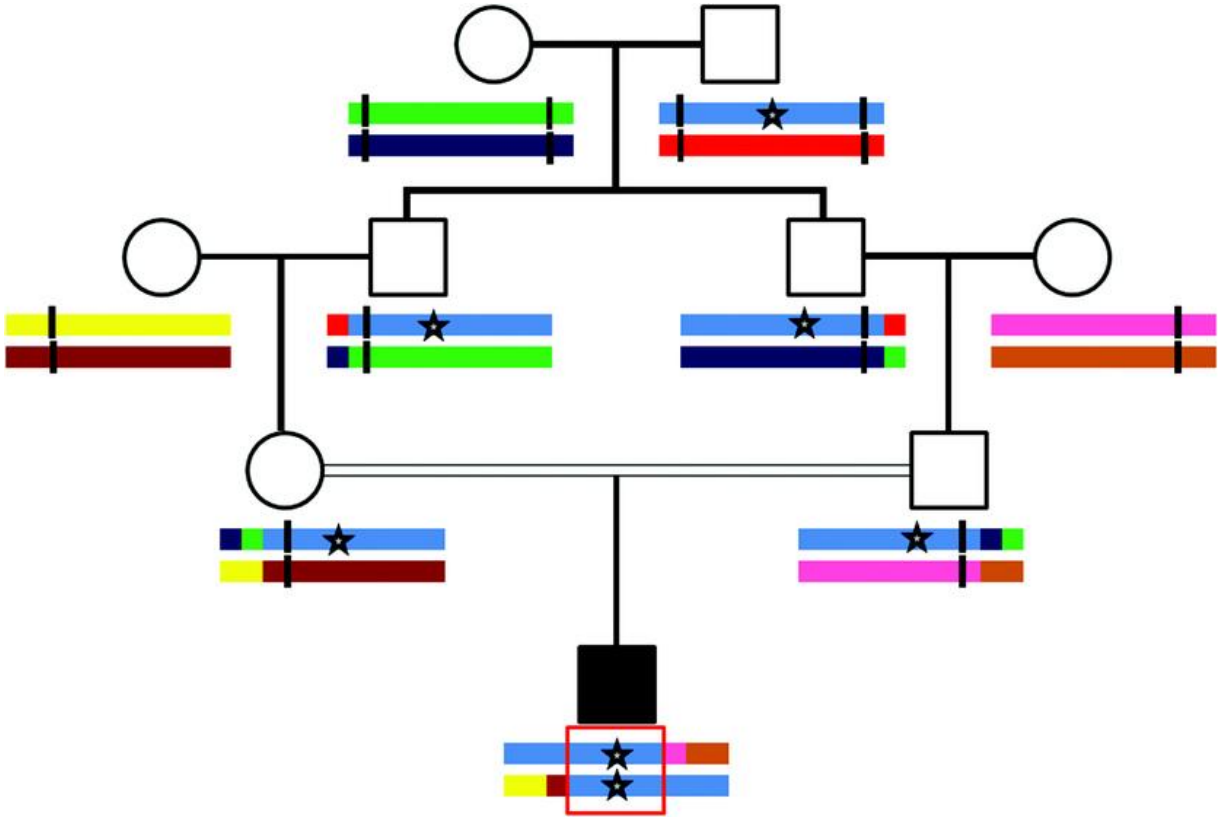
Tom Druet & Mathieu Gautier
Unit of Animal Genomics, GIGA-R, University of Liège, Belgium
Centre de Biologie pour la Gestion des Populations, INRA, France

Outline

- Autozygosity, homozygosity-by-descent (HBD)
- Origin of HBD segments in the genome
- Model-based approach to identify HBD segments
- Two applications of HBD identification with reduced information (low-fold sequencing, low density)

Homozygosity-by-descent

- Autozygous segment, IBD in one individual: homozygous-by-descent (HBD)



Applications

- Identification of HBD segments (or ROH)
  - Estimate inbreeding coefficient
  - Study inbreeding depression
  - Homozygosity mapping (recessive effects)
  - Measure genetic diversity
  - Reveal population demographic history
  - Identify signatures of selection

Origin of HBD segments in the genome

Origin of HBD segments

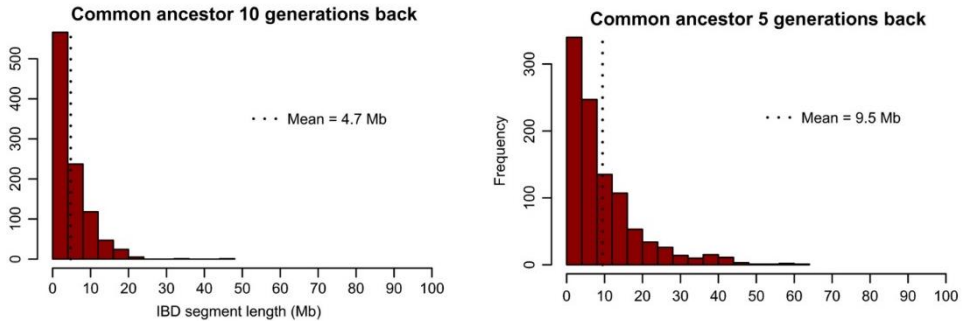
- Positions in the genome can be HBD (autozygous) or non-HBD (allozygous)

Origin of HBD segments

- HBD segments are generated through a complex process
  - Several ancestors $\{A_i, A_j, A_k, \ldots\}$ contribute to autozygosity
  - Each ancestor A_i has it own contribution C_i to autozygosity

Origin of HBD segments

- HBD segments are generated through a complex process
  - Several ancestors {A_i, A_j, A_k,…} contribute to autozygosity
  - Each ancestor A_i has it own contribution C_i to autozygosity
  - The length (L_i) of HBD segments are ancestor specific
    - Function of the size of the inbreeding loop associated with A_i
    - The 'age' of A_i measured in generations



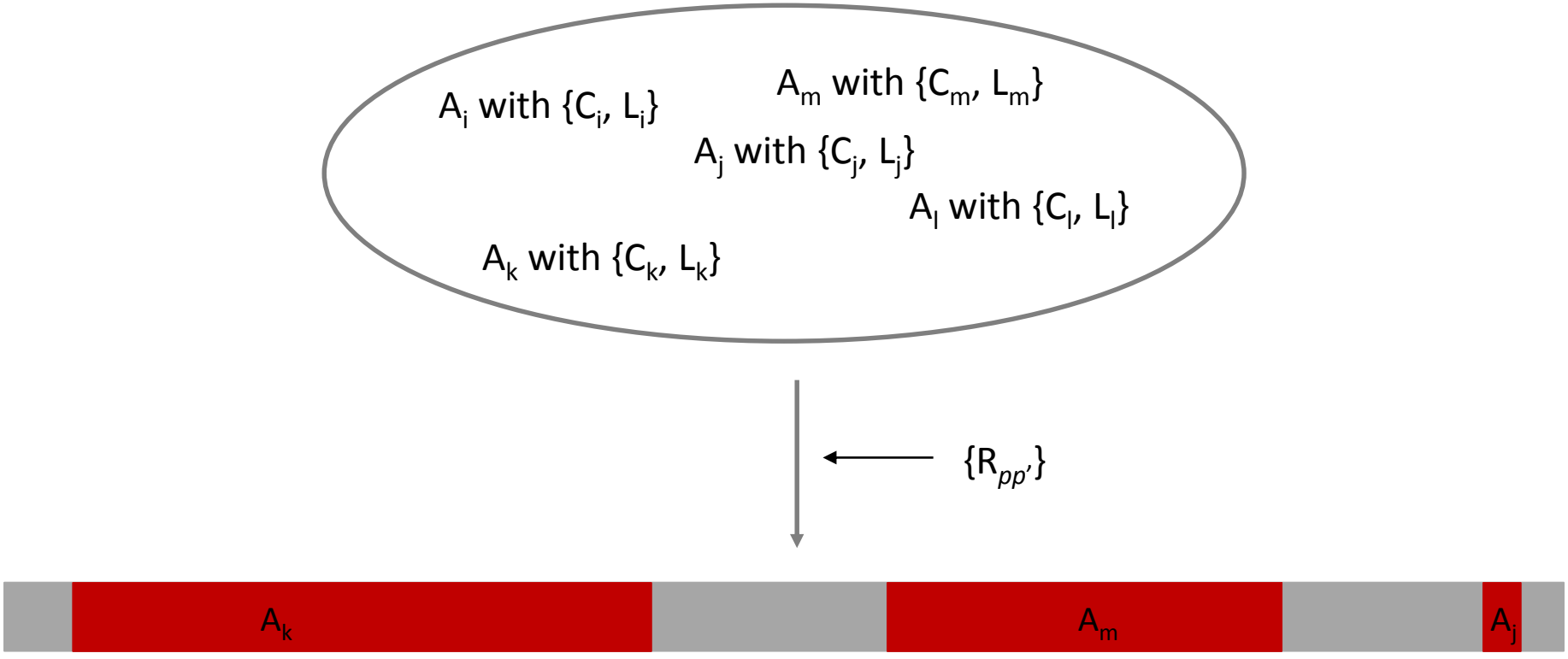
Kardos et al., Evol. Appl. 2016

Origin of HBD segments

- HBD segments are generated through a complex process
  - Several ancestors {A_i, A_j, A_k,…} contribute to autozygosity
  - Each ancestor A_i has it own contribution C_i to autozygosity
  - The length (L_i) of HBD segments are ancestor specific
  - Recombination rates are variable along the genome
  - Stochastic processes

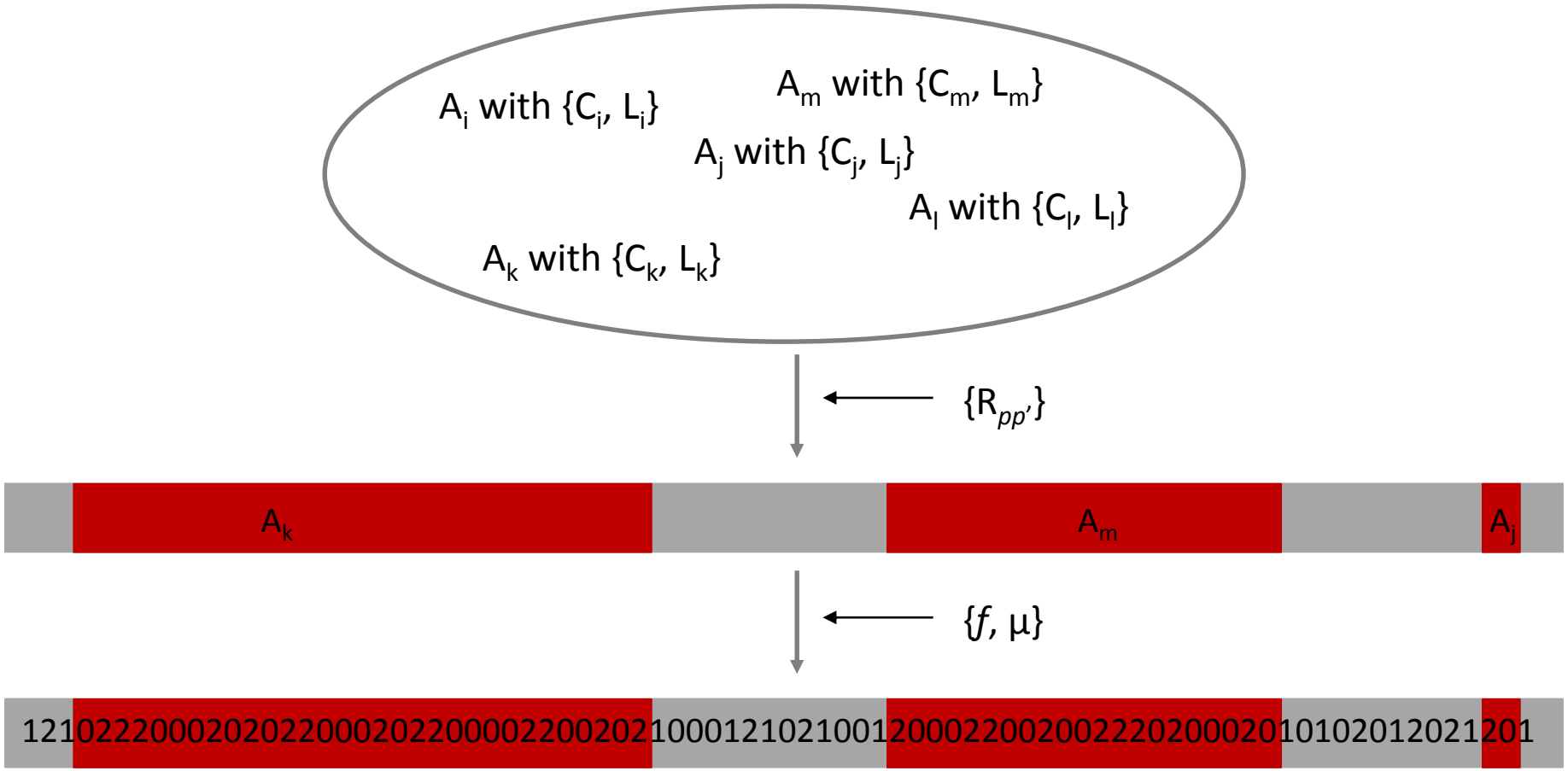
Origin of HBD segments

- HBD segments are generated through a complex process

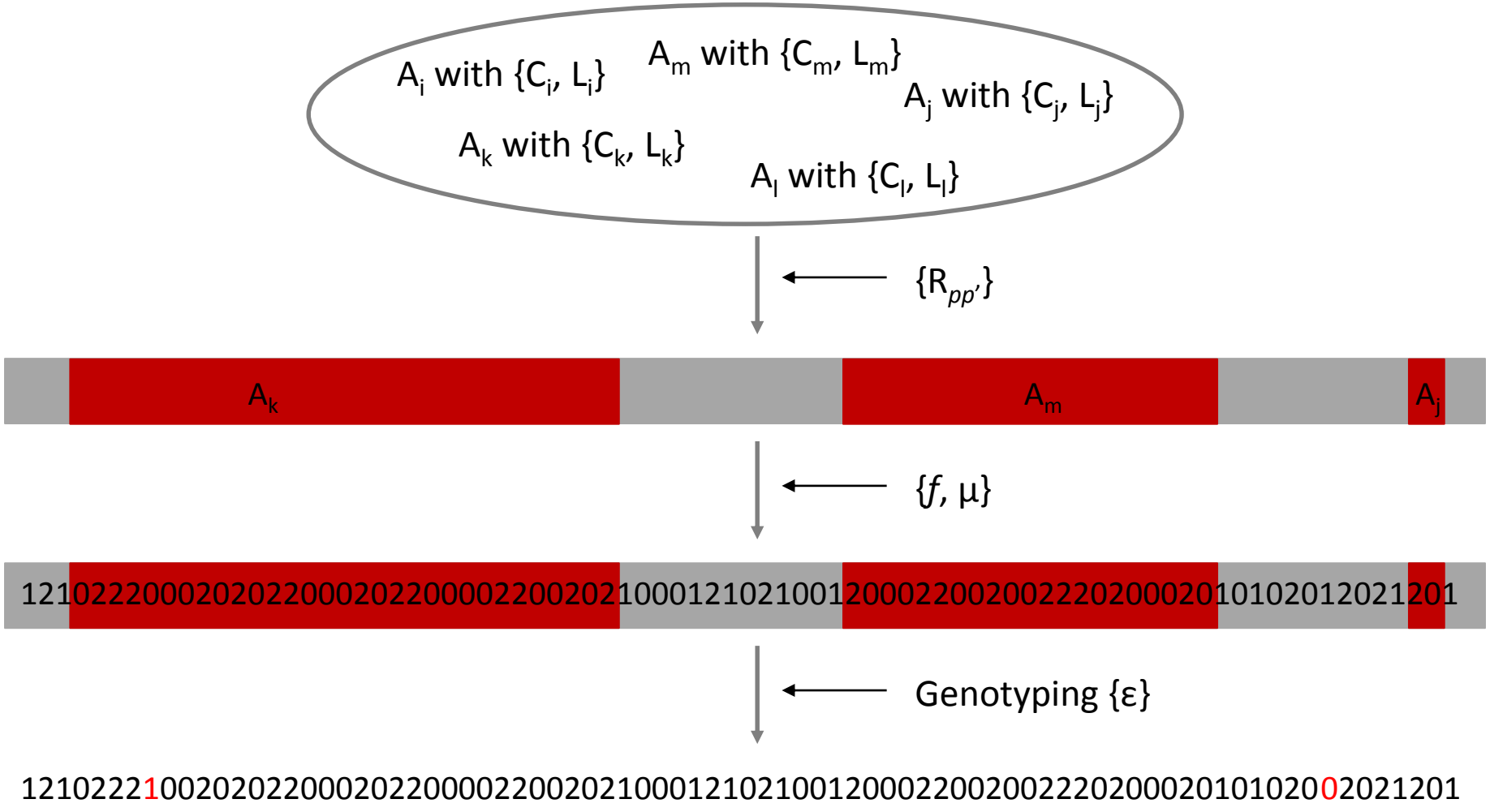


Genotypes in different segments

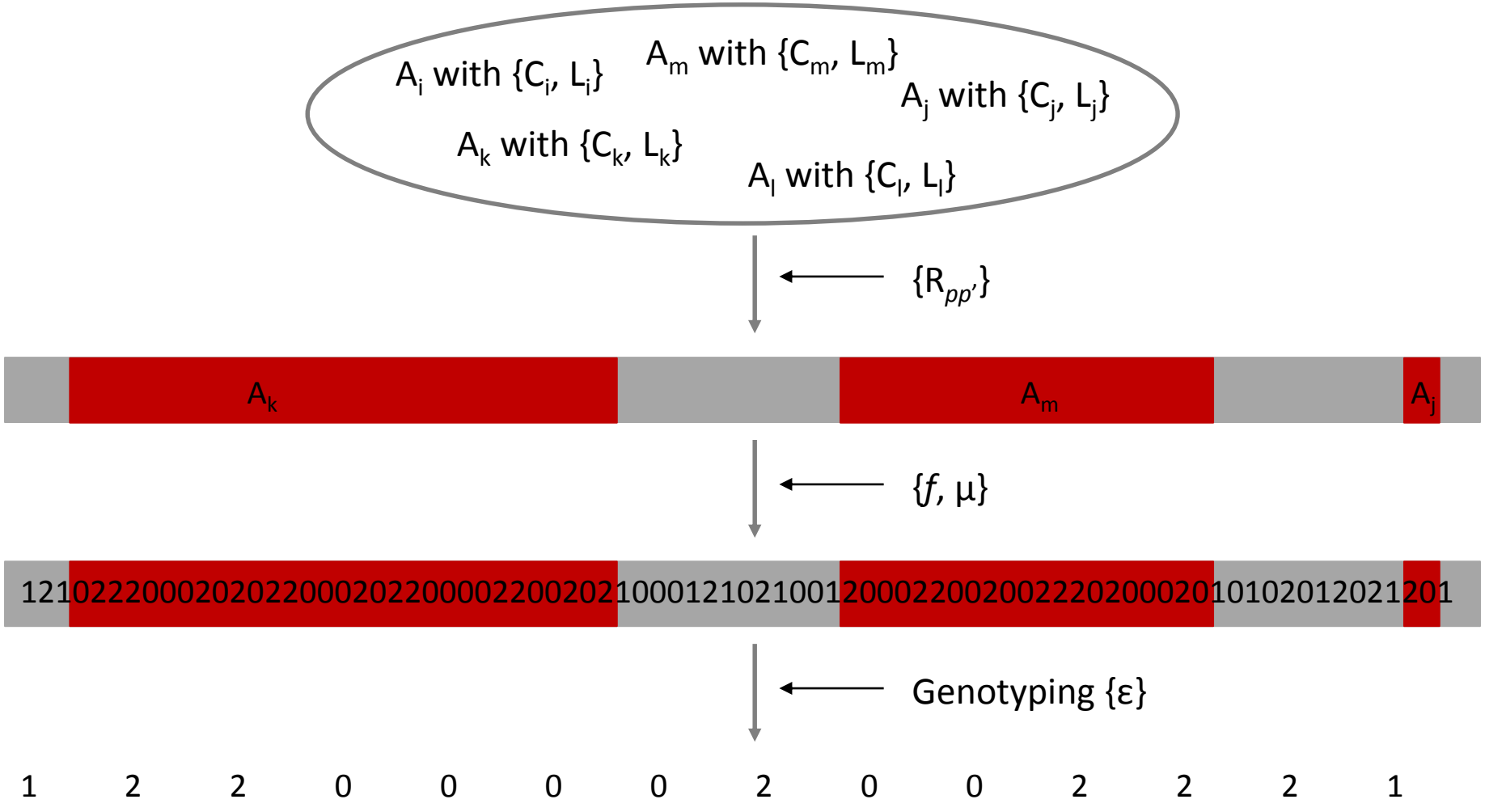
- HBD segments are not directly observed: data required



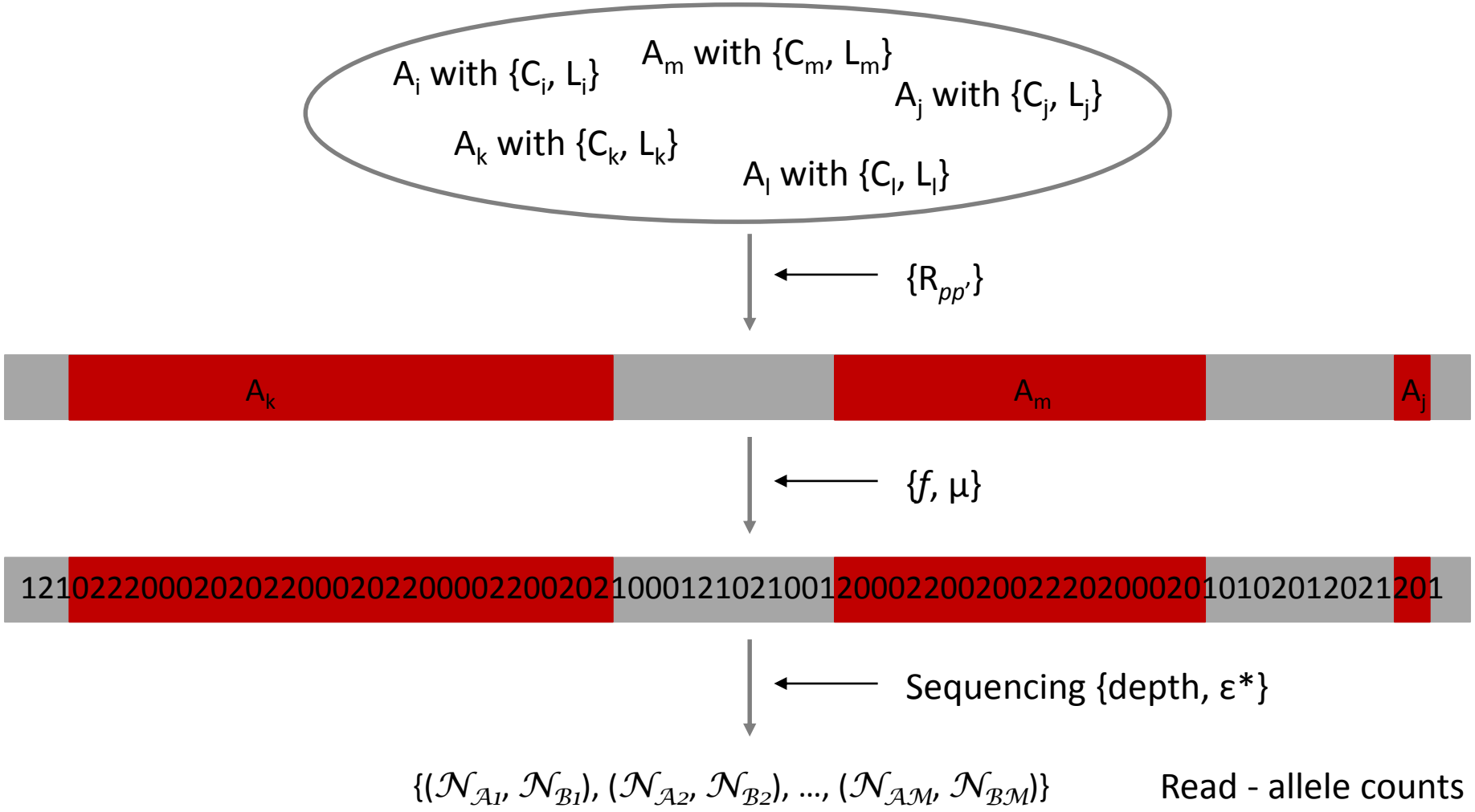
Genotyping arrays



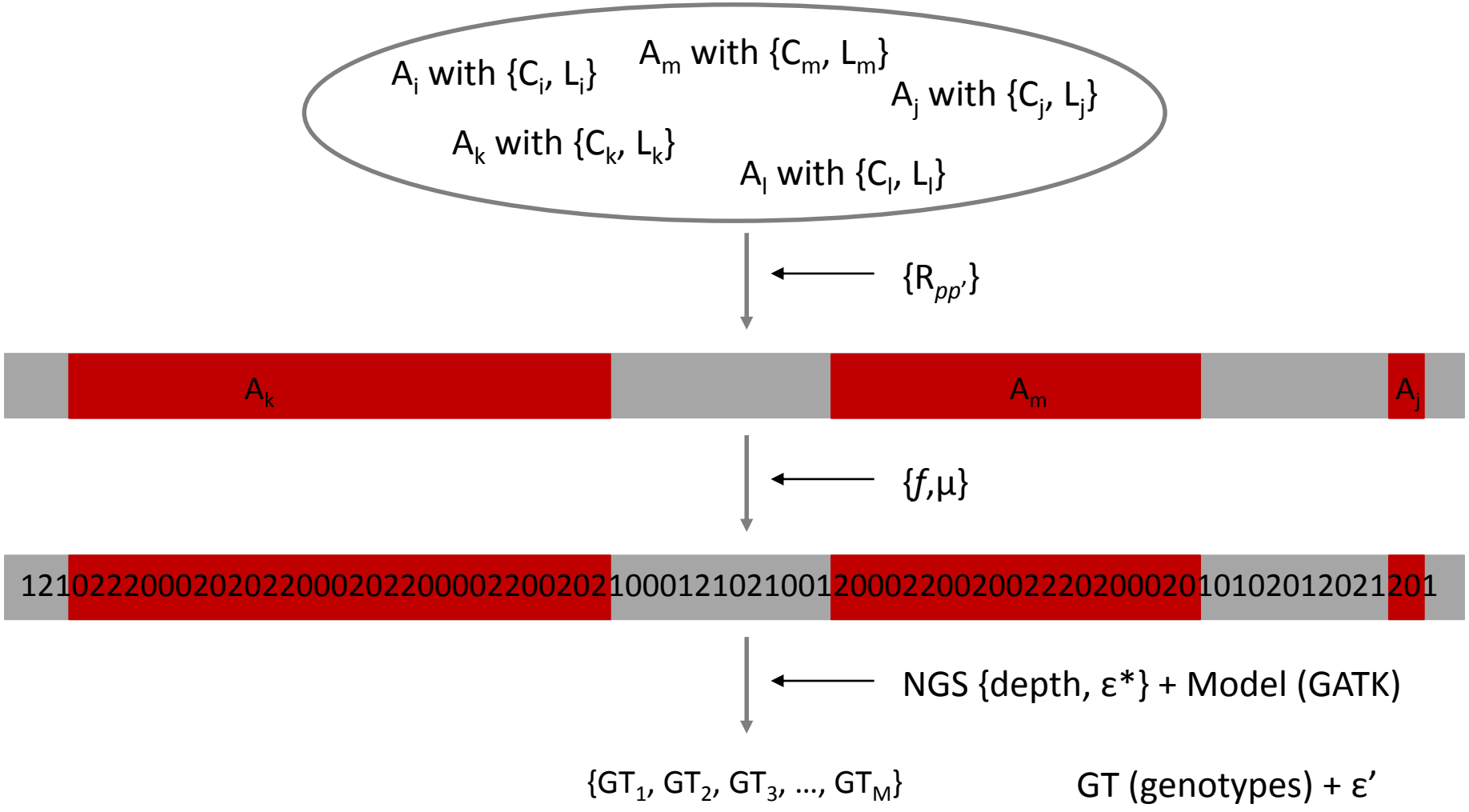
Genotyping arrays (low-density)



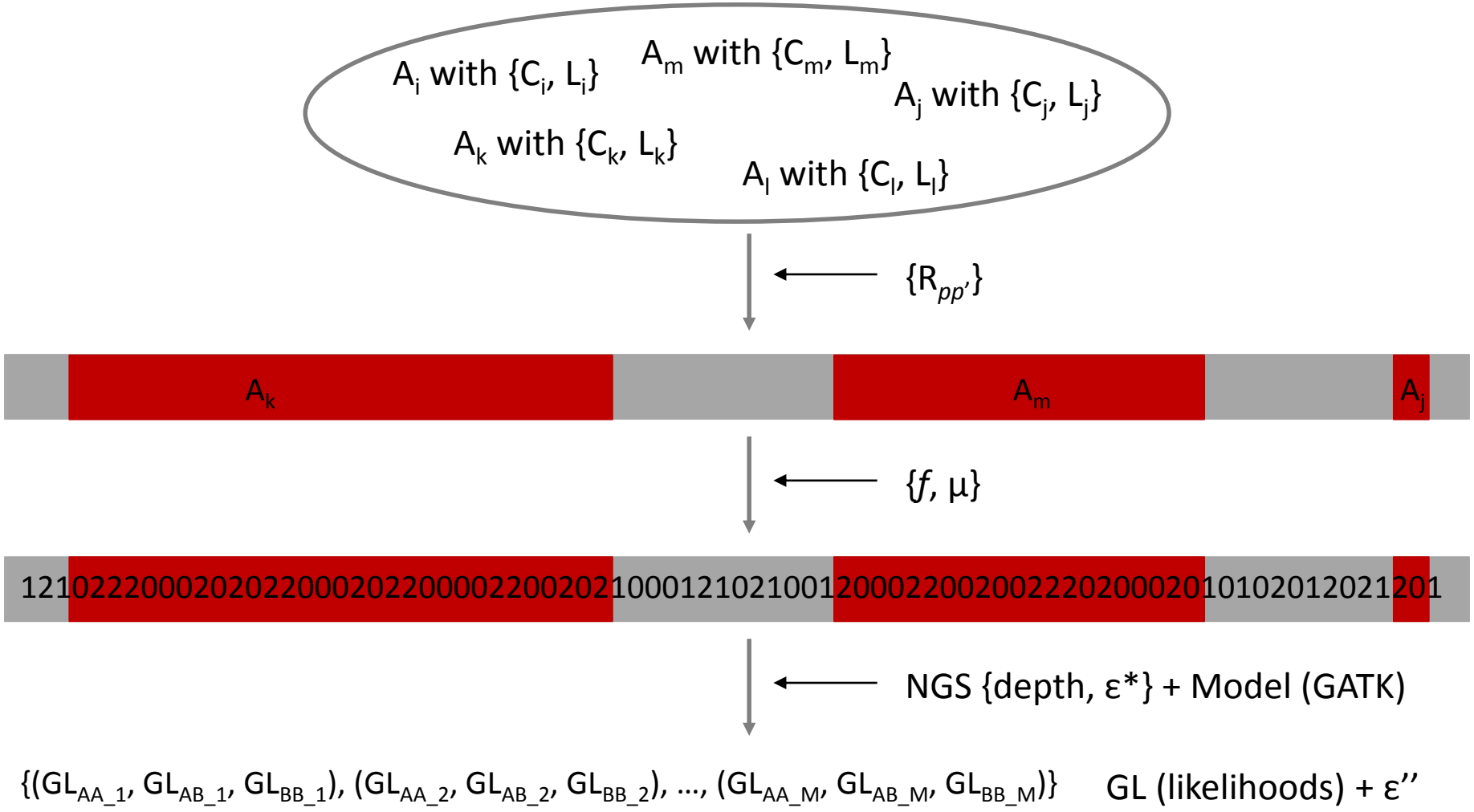
Sequencing data



Sequencing data



Sequencing data



Identification of HBD segments

HBD identification

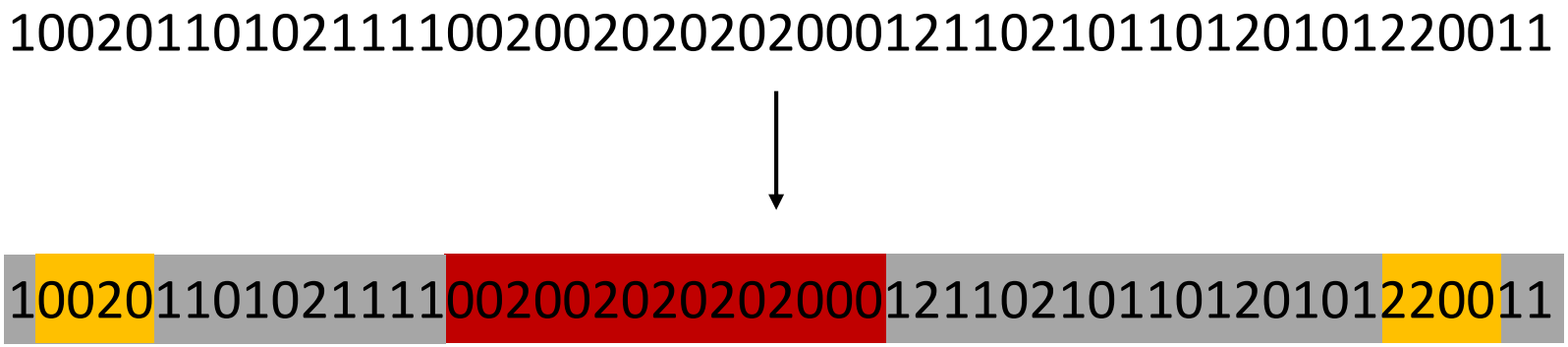
- Complex to identify HBD segments and infer parameters
  - Few markers per segments (density, ancient ancestors)
  - Border of segments
  - Low-fold sequencing
  - Uncertain genotypes (high ‘genotyping’ error rate)
- Additional noise
  - HBD segments can overlap
  - Recent HBD masks more ancient segments

Hidden Markov model

- Each position in the genome is HBD or non-HBD
- Positions are assigned to *K* HBD and non-HBD classes
- Length distributions and frequencies vary
  - Length of segments are exponentially distributed with rate R_k
  - Expected length is $1/R_k$ in Morgans
  - Frequency of classes function of the mixing coefficients ρ_k

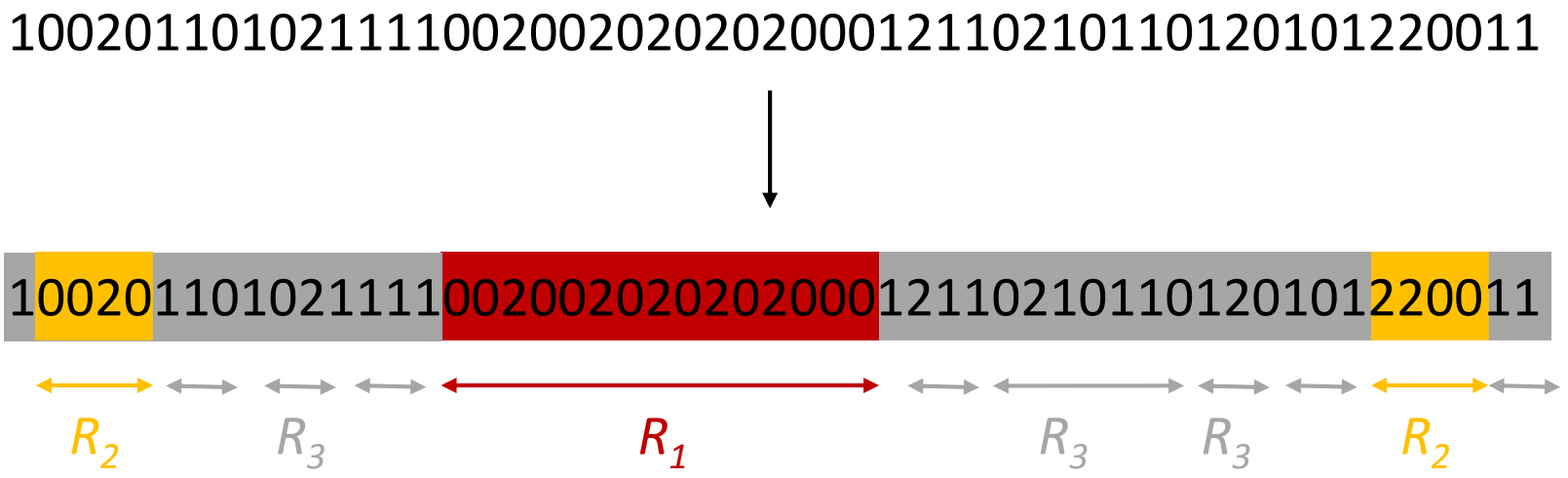
Hidden Markov model

- Positions are assigned to *K* HBD and non-HBD classes



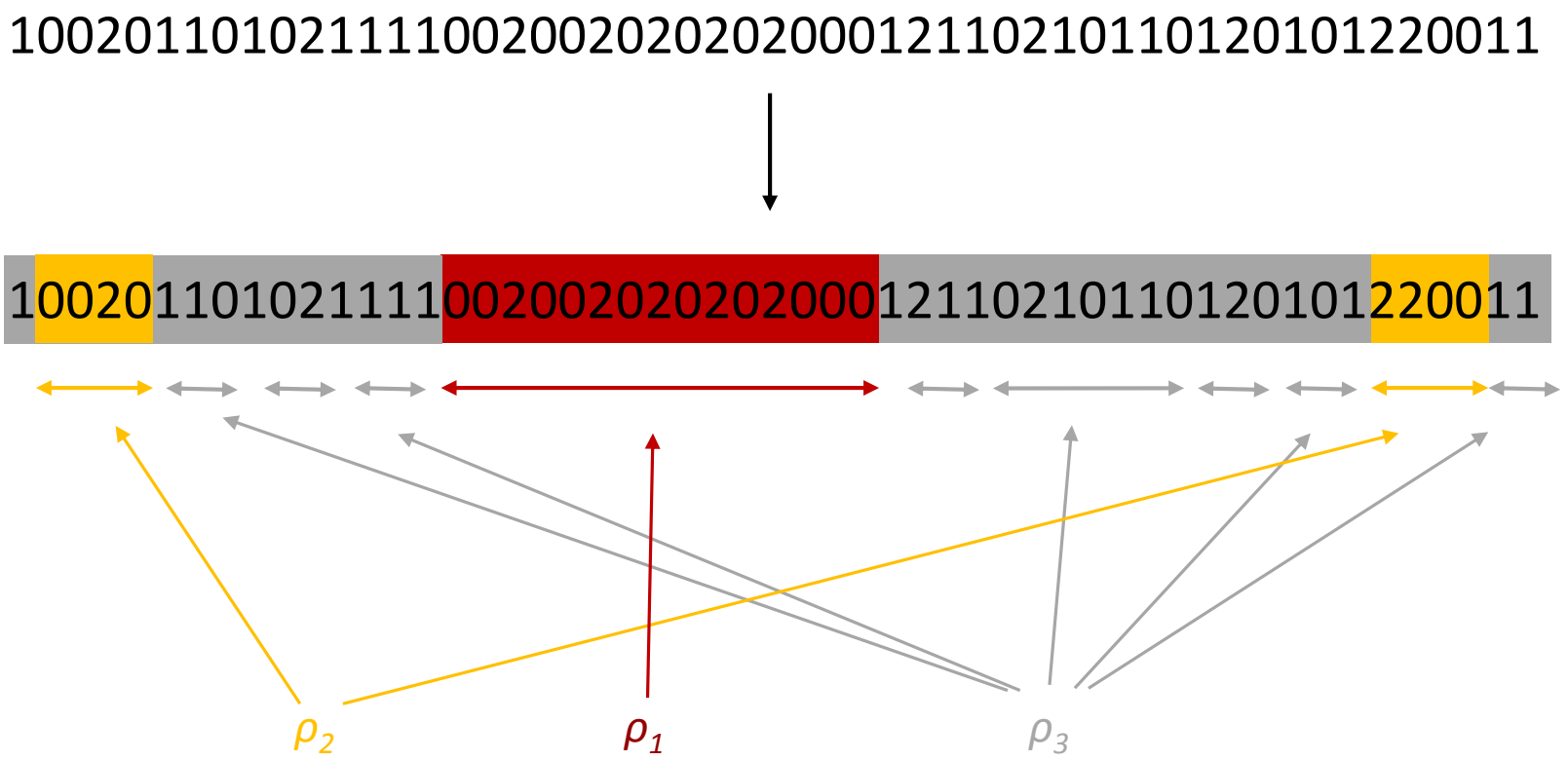
Hidden Markov model

- Positions are assigned to *K* HBD and nn-HBD classes



Hidden Markov model

- Positions are assigned to *K* HBD and nn-HBD classes



Transition probabilities

- No coancestry change between markers *d* Morgans apart
  - The HBD / non-HBD segment extends

	HBD_1	HBD_2	Non-HBD
HBD_1	$e^{-R_1 d}$		
HBD_2		$e^{-R_2 d}$	
Non-HBD			$e^{-R_K d}$

Transition probabilities

- Coancestry change between markers *d* Morgans apart
  - The HBD / non-HBD segment stops

	HBD_1	HBD_2	Non-HBD
HBD_1	$1 - e^{-R_1 d}$		
HBD_2		$1 - e^{-R_2 d}$	
Non-HBD			$1 - e^{-R_K d}$

Transition probabilities

- After coancestry change
  - New segment starts in state *k* with probability ρ_k

	HBD_1	HBD_2	Non-HBD
HBD_1	$(1-e^{-R_1 d})\rho_1$	$(1-e^{-R_1 d})\rho_2$	$(1-e^{-R_1 d})\rho_K$
HBD_2	$(1-e^{-R_2 d})\rho_1$	$(1-e^{-R_2 d})\rho_2$	$(1-e^{-R_2 d})\rho_K$
Non-HBD	$(1-e^{-R_K d})\rho_1$	$(1-e^{-R_K d})\rho_2$	$(1-e^{-R_K d})\rho_K$

Transition probabilities

- Resulting transitions probabilities
  - With and without coancestry change

	HBD_1	HBD_2	Non-HBD
HBD_1	$e^{-R_1 d} + (1 - e^{-R_1 d})\rho_1$	$(1 - e^{-R_1 d})\rho_2$	$(1 - e^{-R_1 d})\rho_K$
HBD_2	$(1 - e^{-R_2 d})\rho_1$	$e^{-R_2 d} + (1 - e^{-R_2 d})\rho_2$	$(1 - e^{-R_2 d})\rho_K$
Non-HBD	$(1 - e^{-R_K d})\rho_1$	$(1 - e^{-R_K d})\rho_2$	$e^{-R_K d} + (1 - e^{-R_K d})\rho_K$

Emission probabilities

- Probability of genotype given HBD status
  - Identical for all HBD classes (does not depend on R_k)
- Non-HBD classes: Hardy-Weinberg proportions
- HBD classes: homozygotes (error, mutation)

	HBD	Non-HBD
A_iA_i	$(1-\varepsilon)f_i$	f_i^2
A_iA_j	ε	$2f_if_j$

Extension to WGS data

- Emission probabilities with genotype likelihoods
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls
- Integration over the three possible genotypes:

$$P(Data \mid HBD) = P(Data \mid A_iA_i) \times P(A_iA_i \mid HBD) + P(Data \mid A_jA_j) \times P(A_jA_j \mid HBD) + P(Data \mid A_iA_j) \times P(A_iA_j \mid HBD)$$

Extension to WGS data

- Emission probabilities with genotype likelihoods
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls
- Integration over the three possible genotypes:

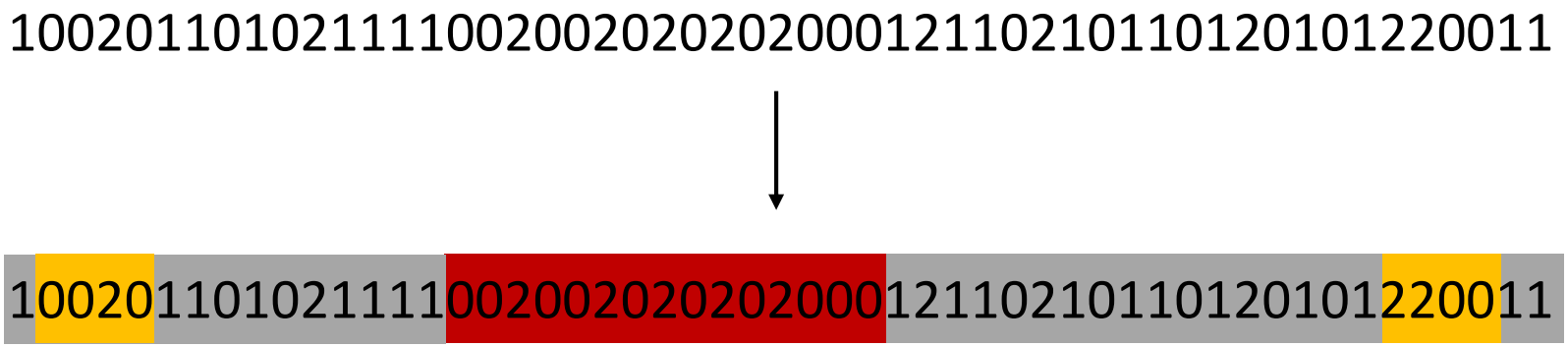
$$P(\text{Data} \mid \text{HBD}) = P(\text{Data} \mid A_iA_i) \times P(A_iA_i \mid \text{HBD}) + P(\text{Data} \mid A_jA_j) \times P(A_jA_j \mid \text{HBD}) + P(\text{Data} \mid A_iA_j) \times P(A_iA_j \mid \text{HBD})$$

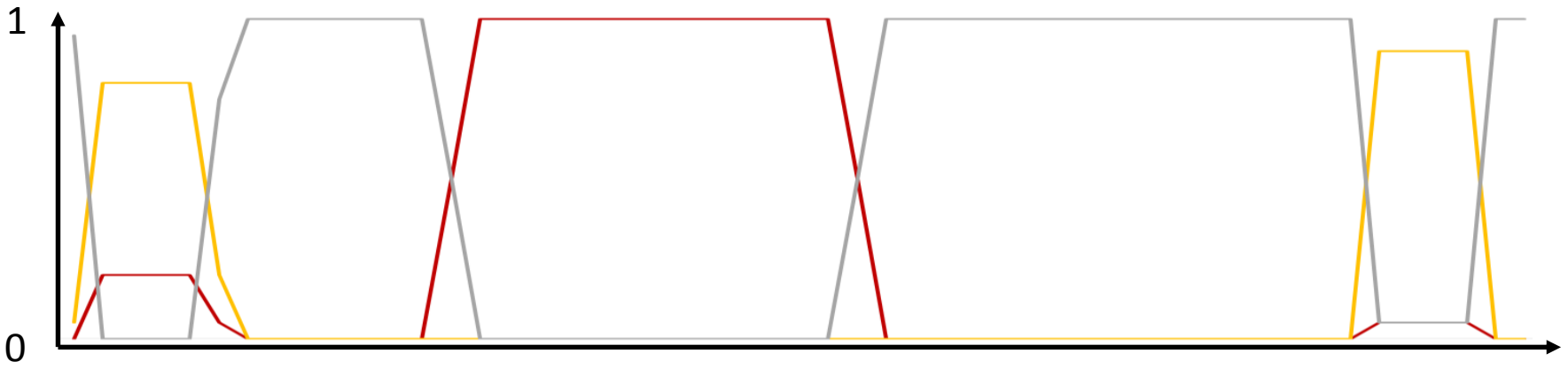
Get information in VCF file

$$\begin{cases} P(AD_l = \{N_{l1}, N_{l2}\} | A_{l1}, A_{l1}) = (1-\varepsilon^*)^{N_{l1}} (\varepsilon^*)^{N_{l2}} \\ P(AD_l = \{N_{l1}, N_{l2}\} | A_{l1}, A_{l2}) = (0.5)^{N_{l1}+N_{l2}} \\ P(AD_l = \{N_{l1}, N_{l2}\} | A_{l2}, A_{l2}) = (\varepsilon^*)^{N_{l1}} (1-\varepsilon^*)^{N_{l2}} \end{cases}$$

Hidden Markov model

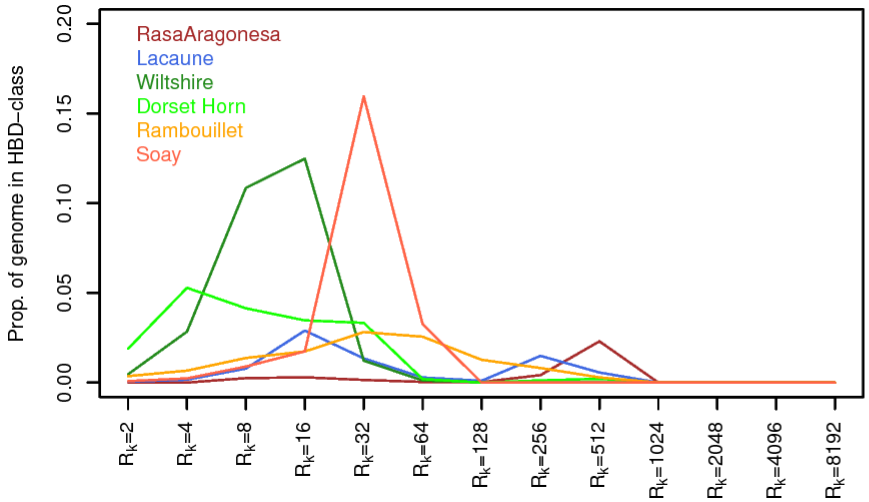
- Positions are assigned to *K* HBD and non-HBD classes

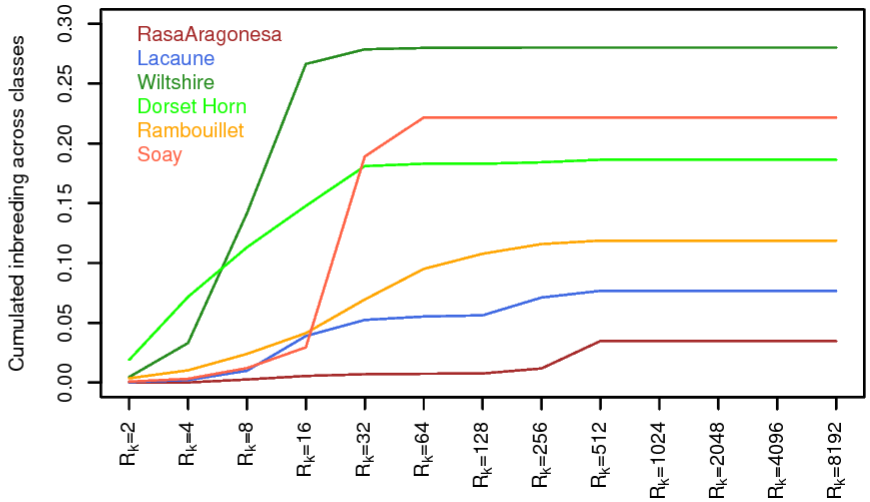




Summary output

- Average HBD probability per class
  - Average over the genome
- Cumulative values
  - All class with rate ≤ T
  - F with different base pop.





A model for HBD identification

- Main features
  - Using genotypes, genotype probabilities, read counts
  - Allele frequencies, error rates (mutation), genetic map
  - Multiple HBD classes (with different length and frequencies)
  - Integration over all windows sizes (HMM framework)
  - Global and local contribution of each class to the genome
  - Probabilistic output

Method evaluation

- Simulations studies (Druet & Gautier, Mol. Ecol. 2017)
  - HBD prob., rates and mixing coefficients
  - Simple and more complex scenarios
  - Different marker densities, allele frequency spectrums, error rates, low-fold sequencing (1x), variable recombination rates
  - Efficiency decreases with informativity
  - Compared to other methods (including likelihood-based ROH)

- Most useful when limited information
  - HBD probabilities (not binary classification)
  - Illustration with two such applications

Low-fold sequencing in Belgian Blue cattle

Low-fold sequencing in cattle

- 47 Belgian Blue sires sequenced
  - Paired-End sequencing 2 x 100, cover > 10x
  - Nextera Mate-Pair 2 x 75, cover 0.45x (first run)
  - Nextera Mate-Pair 2 x 75, cover 0.90x (two runs)
- Genotyped on BovineHD
  - 7K – Markers from BovineLD array
  - 32K – Markers from 50K array
  - 585K – Markers from BovineHD array

Low-fold sequencing in cattle

- Some statistics
  - 5,667,384 SNPS selected in 10x data

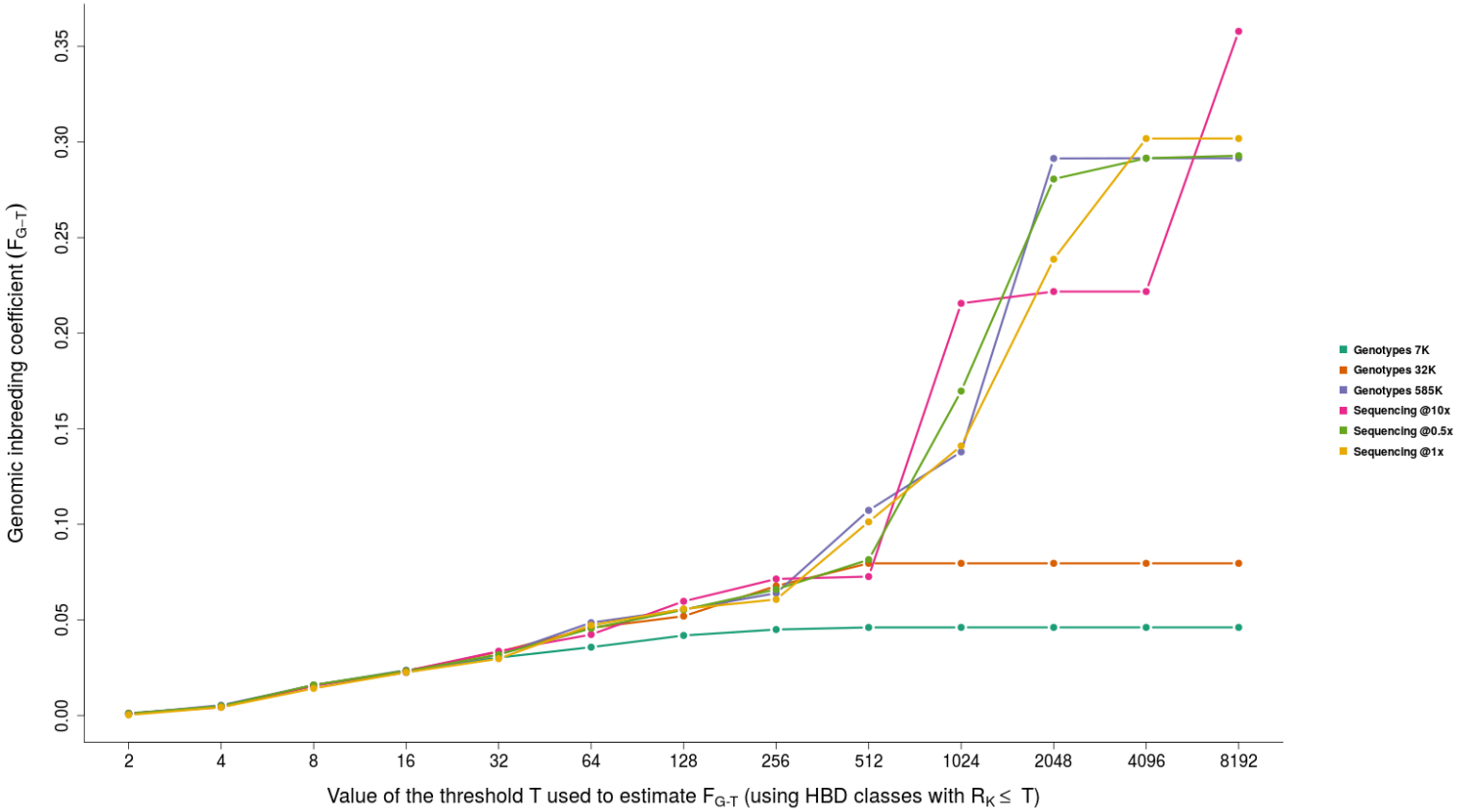
	NMPI	NMPII
Number of SNPs	2,345,312	3,996,864
Cover	0.42x	0.87x
Positions with > 1 read	173,100	880,200
Positions with > 3 reads	3372	60,950

*1 read non-informative, emission prob. equal to f_i

**4 reads less informative than SNPs, prob. to observe one allele in heterozygotes is still 0.125

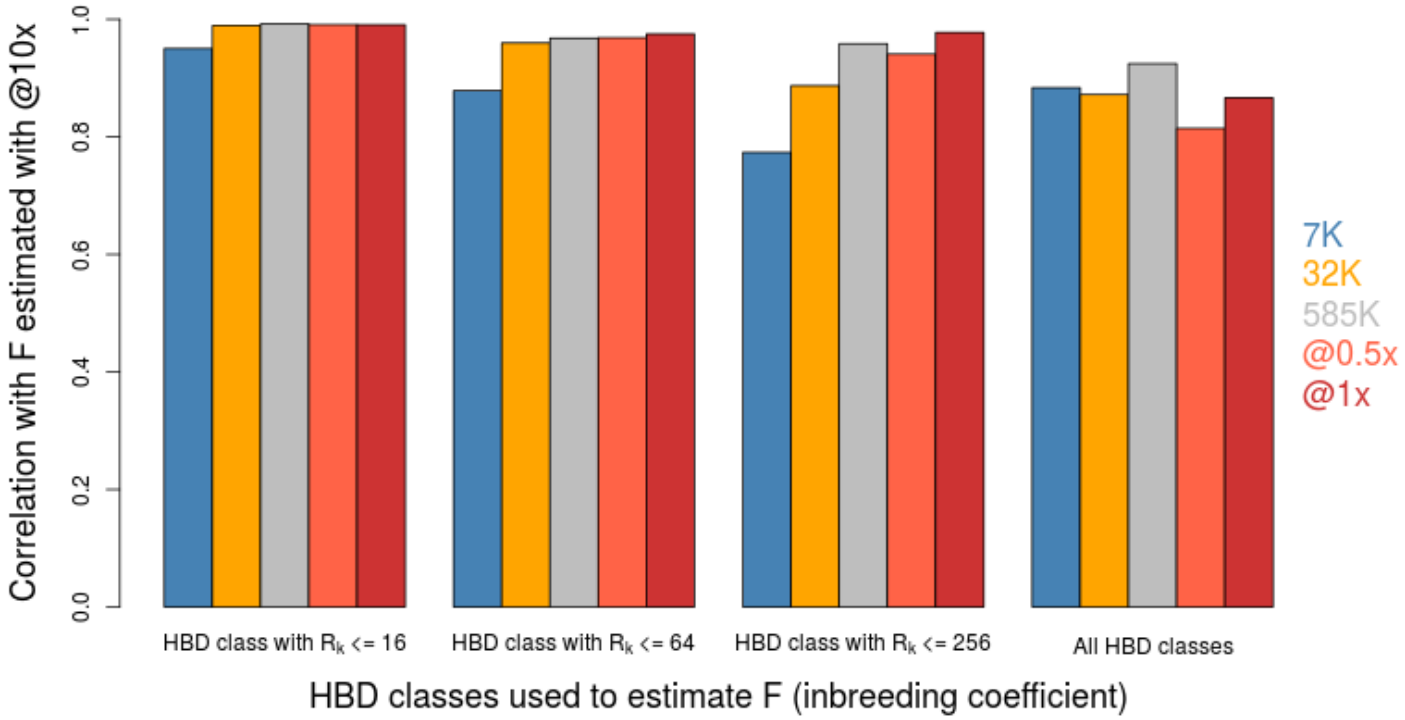
Autozygosity levels

- Characterization with a model with 13 HBD classes:



Individual autozygosity levels

- Correlations with whole-genome sequencing data (> 10x)



Identification of HBD segments

Low-fold sequencing in cattle

- Works with real low-fold sequencing data (0,5x)
- Formula based on AD gives similar results
- Using allele frequency estimates from 10x or 1x gives similar results
- Differences more pronounced for ancient autozygosity, smaller segments

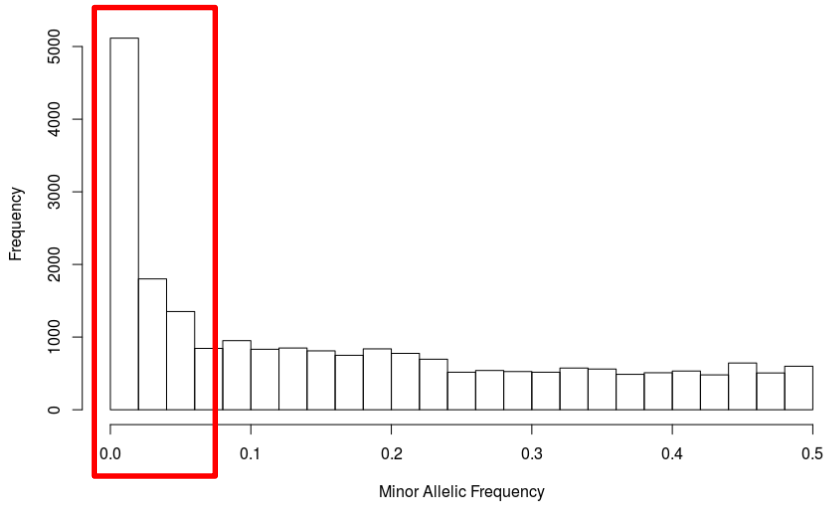
Inbreeding in European Bison (Wisent)

Inbreeding in European Bison

- Extinct in the wild (beginning 20th century)
- Restoration from 12 founders
- Two distinct genetic lines:
  - Lowland line (Ll), 7 founders without Caucasian blood
  - Lowland-Caucasian line (LC), 12 founders (one Caucasian subsp.)
- ~2,000 lowland in the Bialowieza forest (Poland)
- Drastic bottleneck (also reduction after WWII)

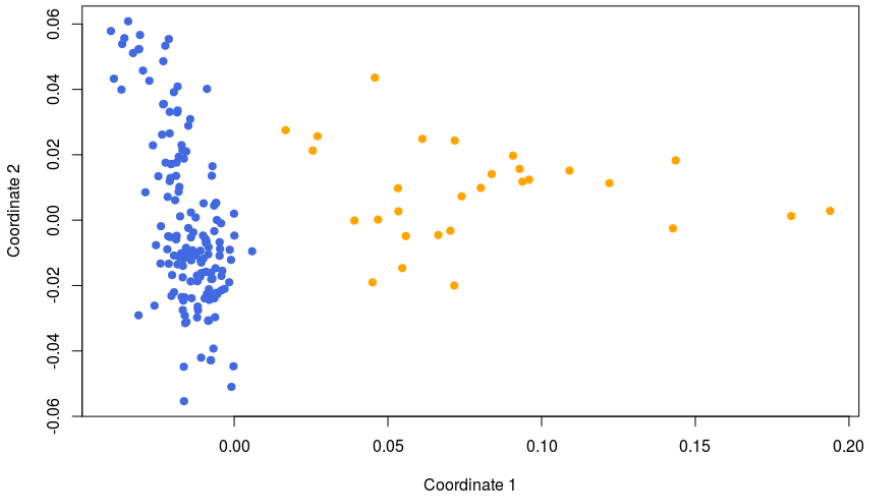
Genotyping data

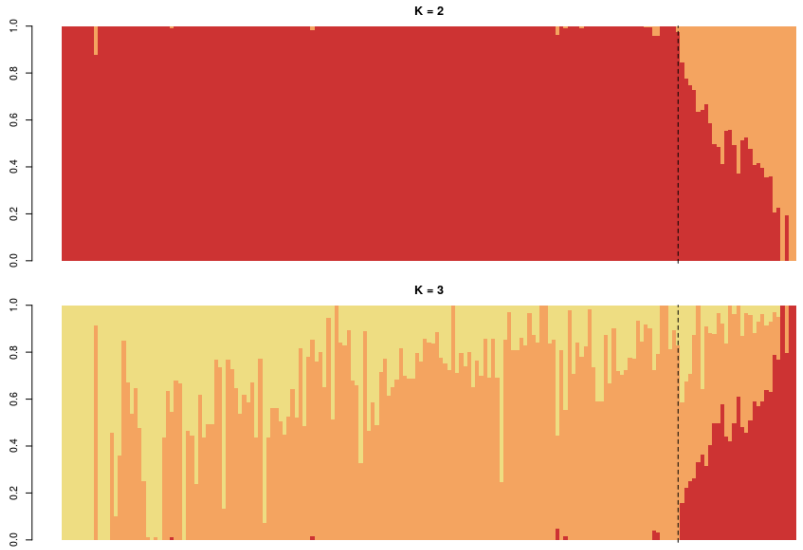
- 154 Ll and 29 LC individuals
  - Sampled at Mammal Research Institute in Bialowieza (+INRA)
- Genotyped with BovineHD array (Illumina, CA)
  - 710,964 mapping on autosomes (Bovine build)
  - After filtering (monomorphic, call rate): 22,602 SNPs
  - Low informativity: MAF, LD



Population structure

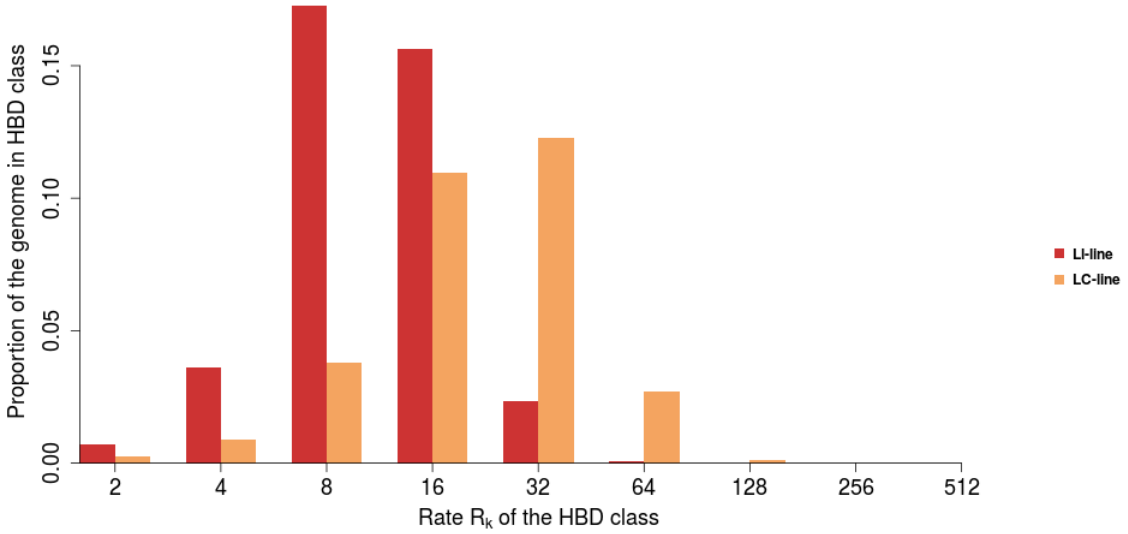
- Structure identified with SNPs correspond to the two genetic lines
- MDS analysis (PLINK)
- ADMIXTURE





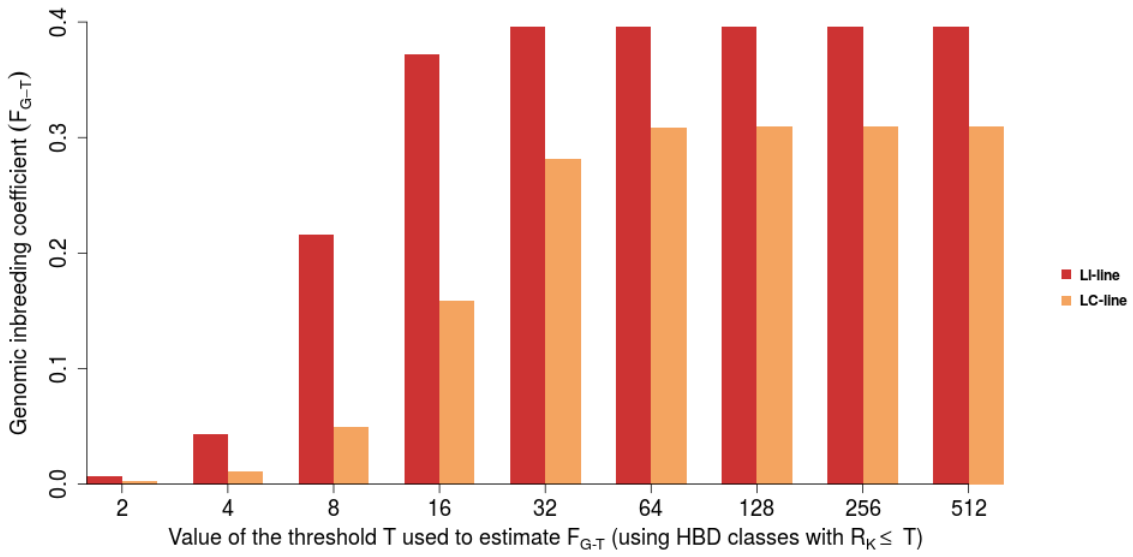
Individual autozygosity

- Characterization with a model with 10 HBD classes:
  - Major contribution of HBD classes with R_k = 8 and 16 in LI
  - Major contribution of HBD classes with R_k = 16 and 32 in LC
  - Rate ~ size of inbreeding loop (generations)



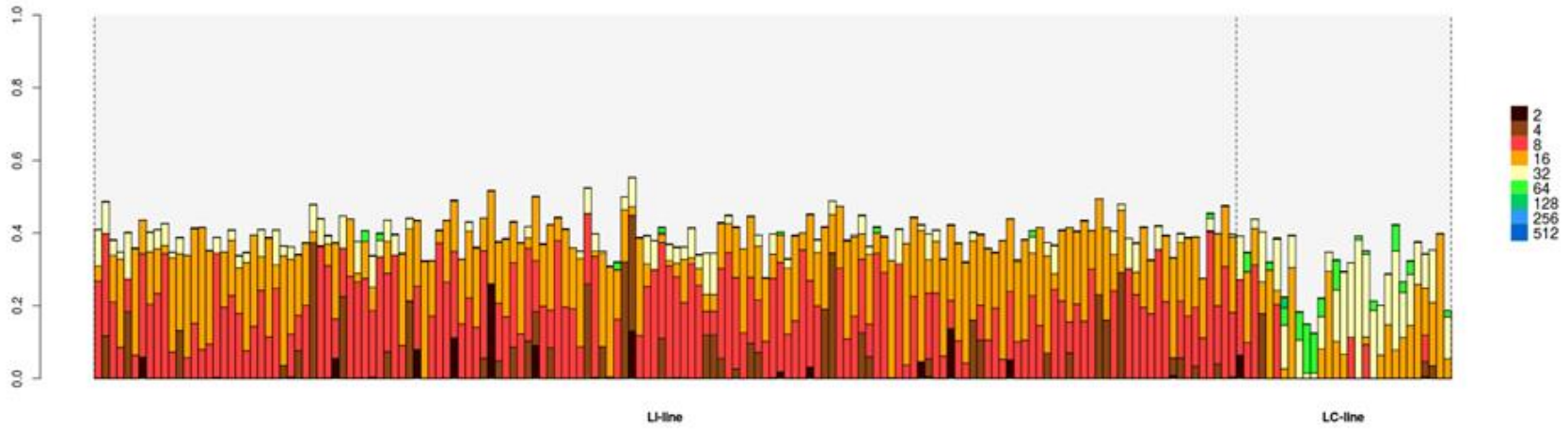
Individual inbreeding

- Characterization with a model with 10 HBD classes:
  - Recent autozygosity is 40% in Ll and 30% in LC



Partitioning per individual

- Percentage of the genome in each HBD class (y-axis)
  - LI dominated by HBD classes with R_k from 2 to 16
  - LC dominated by HBD classes with R_k from 16 to 64



Distribution of HBD segments

- Longer segments in the Lowland-line

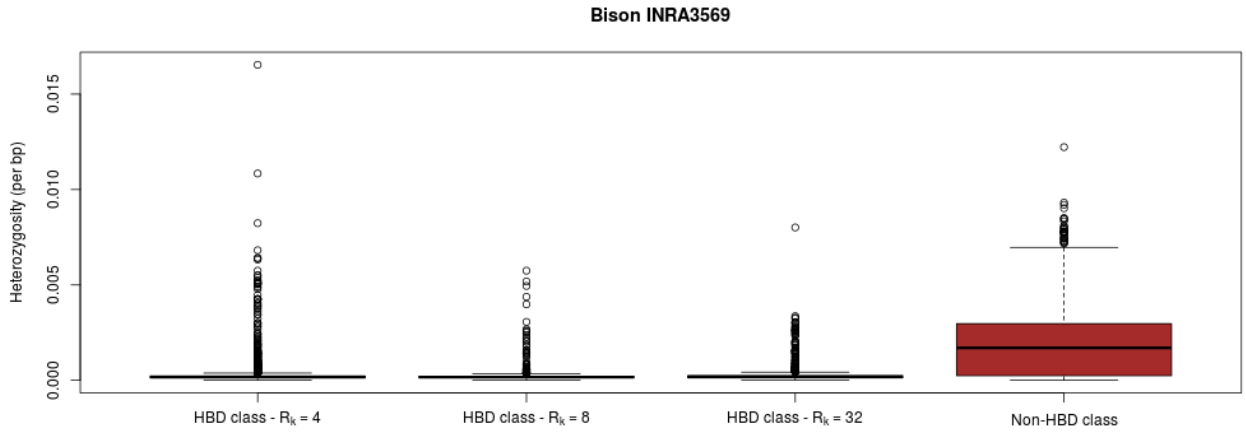
Number of segments per individual	Lowland line	Lowland-Caucasian line
Total	72.5	83.0
. < 5 Mb	20.2	38.8
5 Mb ≤ . < 10 Mb	19.1	21.7
10 Mb ≤ . < 20 Mb	19.5	15.9
20 Mb ≤ . < 50 Mb	12.5	5.9
50 Mb ≤ .	1.1	0.6
Average length	12.7 Mb	8.2 Mb
Max. length	123.7 Mb	90.7 Mb

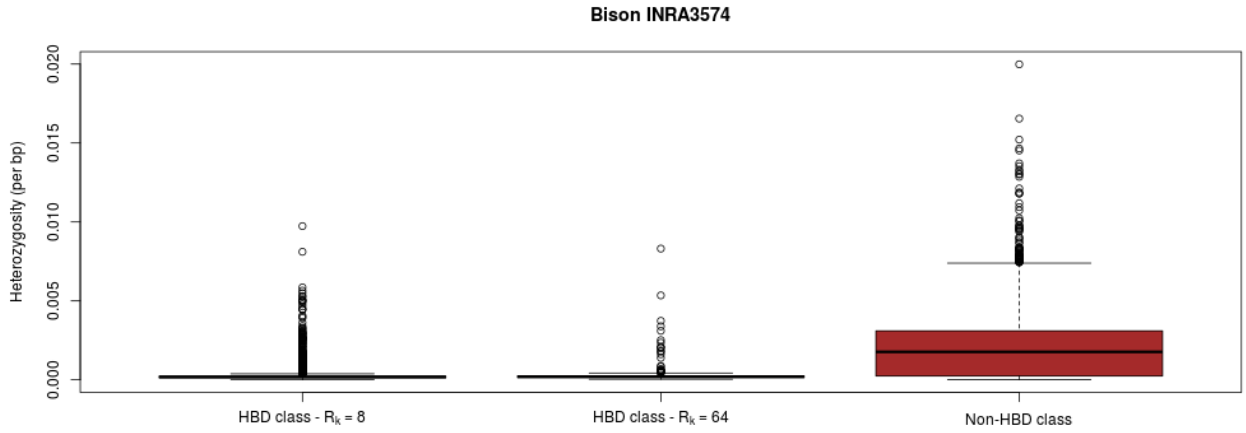
Validation with NGS data

- Inbreeding was characterized with few markers
- Use of NGS data for two Ll individuals
  - Sequencing cover ~ 8x
  - Measure average heterozygosity in 100 kb windows around marker positions

Validation with NGS data

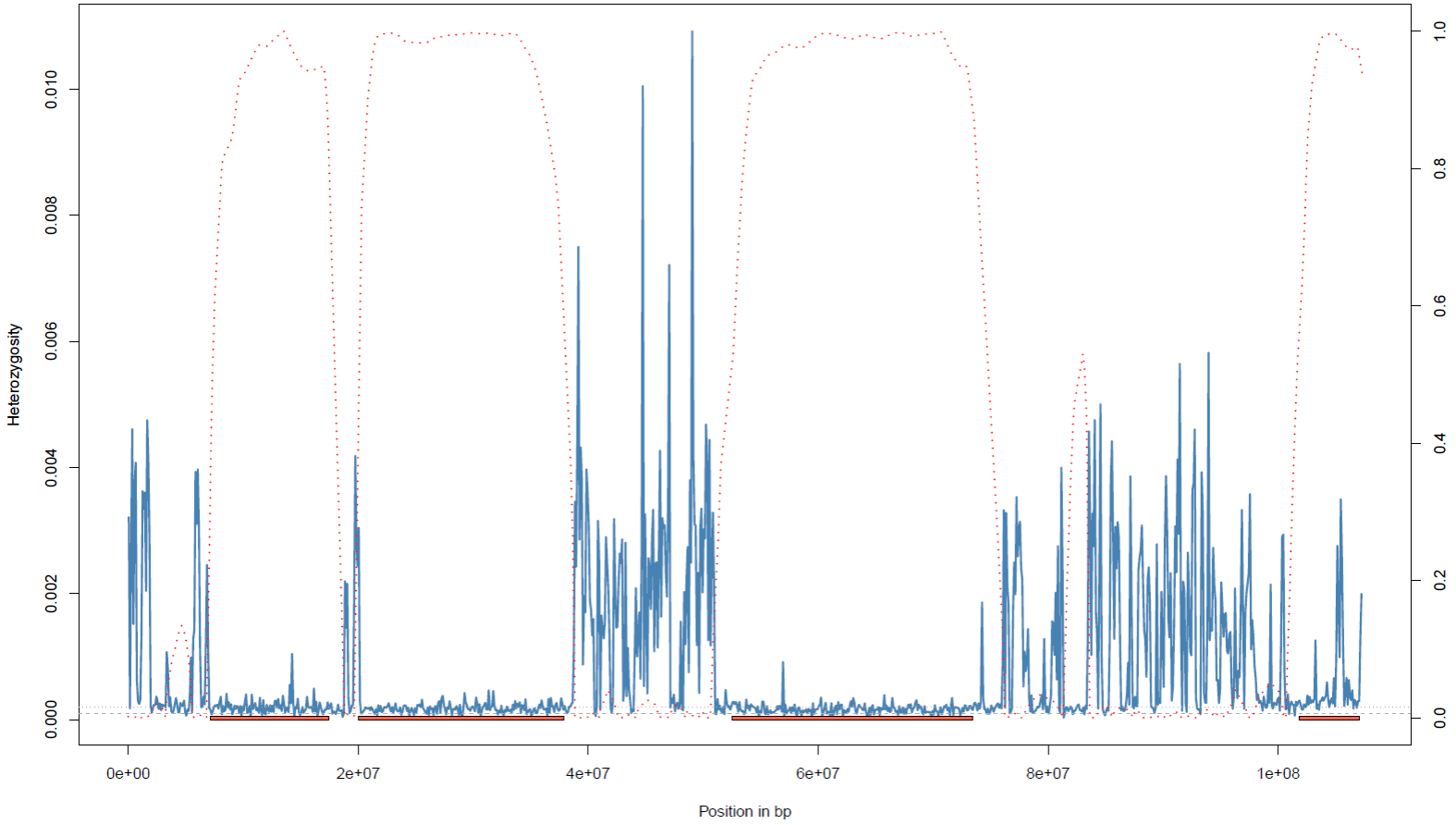
- HBD classes present a ten-fold heterozygosity reduction





Validation with NGS data

- Regions of reduced heterozygosity have high HBD prob.



Using LD genotypes in Bison

- The model-based approach allows to characterize recent autozygosity with a limited number of markers that are not extremely polymorphic
- Identified HBD segments present strong heterozygosity reduction in NGS data

Summary

- Model based approach:
  - Using genotypes, genotype probabilities, read counts
  - Allele frequencies, error rates (mutation), genetic map
  - Global and local contribution of each class to the genome
  - HBD probability in as output
- Important when information is weaker:
  - Low marker density, less informative genotypes (low-fold sequencing, errors), short HBD segments, border, etc.

Implementation

- Fortran program (Github) and R package (cran)

RZooRoH: Partitioning of Individual Autozygosity into Multiple Homozygous-by-Descent Classes

Functions to identify Homozygous-by-Descent (HBD) segments associated with runs of homozygosity (ROH) and to estimate individual autozygosity (or inbreeding coefficient). HBD segments and autozygosity are assigned to multiple HBD classes with a model-based approach relying on a mixture of exponential distributions. The rate of the exponential distribution is distinct for each HBD class and defines the expected length of the HBD segments. These HBD classes are therefore related to the age of the segments (longer segments and smaller rates for recent autozygosity / recent common ancestor). The functions allow to estimate the parameters of the model (rates of the exponential distributions, mixing proportions), to estimate global and local autozygosity probabilities and to identify HBD segments with the Viterbi decoding. The method is fully described in Druet and Gautier (2017) <doi:10.1111/mec.14324>.

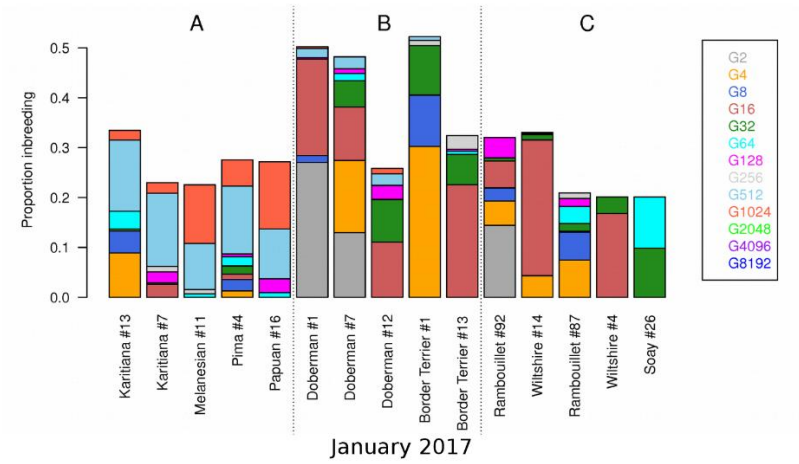
Version:	0.1.1
Depends:	R (≥ 3.2.0), methods
Imports:	foreach, doParallel, parallel, data.table, RColorBrewer, iterators
Suggests:	knitr, rmarkdown
Published:	2018-06-23
Author:	Tom Druet, Naveen Kumar Kadri, Amandine Bertrand and Mathieu Gautier
Maintainer:	Tom Druet <tom.druet at uliege.be>
License:	GPL-3
NeedsCompilation:	yes
CRAN checks:	RZooRoH results

Downloads:

Reference manual:	RZooRoH.pdf
Package source:	RZooRoH_0.1.1.tar.gz
Windows binaries:	r-devel: RZooRoH_0.1.1.zip, r-release: RZooRoH_0.1.1.zip, r-oldrel: RZooRoH_0.1.1.zip
OS X binaries:	r-release: RZooRoH_0.1.1.tgz, r-oldrel: RZooRoH_0.1.1.tgz
Old sources:	RZooRoH archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=RZooRoH to link to this page.



Collaborators

- Amandine Bertrand, Naveen Kumar Kadri
  - Uliege (Belgium)
- Stanislaw Kaminski, Kamil Olénski
  - University of Warmia and Mazury in Olsztyn (Poland)
- Malgorzata Tokarska
  - Mammal Research Institute, Polish Academy of Sciences
- Laurence Flori
  - SELMET, INRA, CIRAD, Montpellier Supagro, Univ. Montpellier (France)