A hidden Markov model to estimate inbreeding from whole genome sequence data

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

Introduction

- Controlling inbreeding in livestock species or in small populations
  - Recessive defects, inbreeding depression, etc.
- Genomic data
  - Observation of realized inbreeding
  - Pedigree sometimes unavailable

Genomic inbreeding *F*

- Estimation with genomic relationship matrix (GRM)
  - Reference population
  - Independent SNPs
  - Global estimate
- Runs of homozygosity (ROH)
  - Parameter definitions
  - Allele frequencies not used
  - Inappropriate for low-fold sequencing

Hidden Markov models

- Models the genome as a mosaic of IBD (inbred) and non-IBD segments (e.g., Leutenegger, 2003 - AJHG)

1002011010211110020020202121101211021011012010121001 1

Hidden Markov models

- Models the genome as a mosaic of IBD (inbred) and non-IBD segments (e.g., Leutenegger, 2003 - AJHG)

1002011010211110020020202121101211021011012010121001 1

Emission probabilities

- Probability of genotype given IBD status (emission prob.):

	IBD	Non-IBD
A_iA_i	p_i	p_i^2
A_iA_j	ε	$2p_ip_j$

Transition probabilities

- Absence of coancestry change is $e^{-\alpha}$ (α is the transition rate: recombination rate & time to common ancestor)
- Prob. new coancestry is IBD is F
- Prob. New coancestry is non-IBD equals $(1-F)$

Transition probabilities

- Transition matrix:

	IBD	Non-IBD
IBD		$(1-e^{-\alpha})(1-F)$
Non-IBD	$(1-e^{-\alpha})F$	

Transition probabilities

- Transition matrix:

	IBD	Non-IBD
IBD	$e^{-\alpha}$	$(1-e^{-\alpha})(1-F)$
Non-IBD	$(1-e^{-\alpha})F$	$e^{-\alpha}$

Transition probabilities

- Transition matrix:

	IBD	Non-IBD
IBD	$e^{-\alpha} + (1-e^{-\alpha})F$	$(1-e^{-\alpha})(1-F)$
Non-IBD	$(1-e^{-\alpha})F$	$e^{-\alpha} +(1-e^{-\alpha})(1-F)$

Extension to WGS data

- Replace genotypes in emission probabilities:
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF):

$$P(\text{Data} \mid \text{IBD}) = p_i\, P(A_iA_i \mid \text{Data}) + p_j\, P(A_jA_j \mid \text{Data}) + \varepsilon\, P(A_iA_j \mid \text{Data})$$

Extension to WGS data

- Replace genotypes in emission probabilities :
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  - Use allele counts (allele depth – AD)

$$P(AD \mid IBD) = p_i\, P(AD \mid A_iA_i) + p_j\, P(AD \mid A_jA_j)$$

ε included

Extension to WGS data

- Replace genotypes in emission probabilities :
  - Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  - Use allele counts (allele depth – AD)

- Recent implementations:
  - BCFtools / RoH (Narasimhan *et al.* – Bionformatics, 2016)
  - ngsF-HMM (Viera *et al.* – Bionformatics, 2016)

Limitation

- Assumes a single inbreeding event (one ancestor)
  - Still a single reference population
- In livestock species, complex inbreeding
  - Many common ancestors over many generations
  - Variable Ne over time (including bottlenecks)

Mixture of inbreeding classes

- Mixture of several IBD and nonIBD with different age (G)
- Emission probabilities unchanged
- Transition probabilities same principle
  - Each distribution with its own mixing proportions

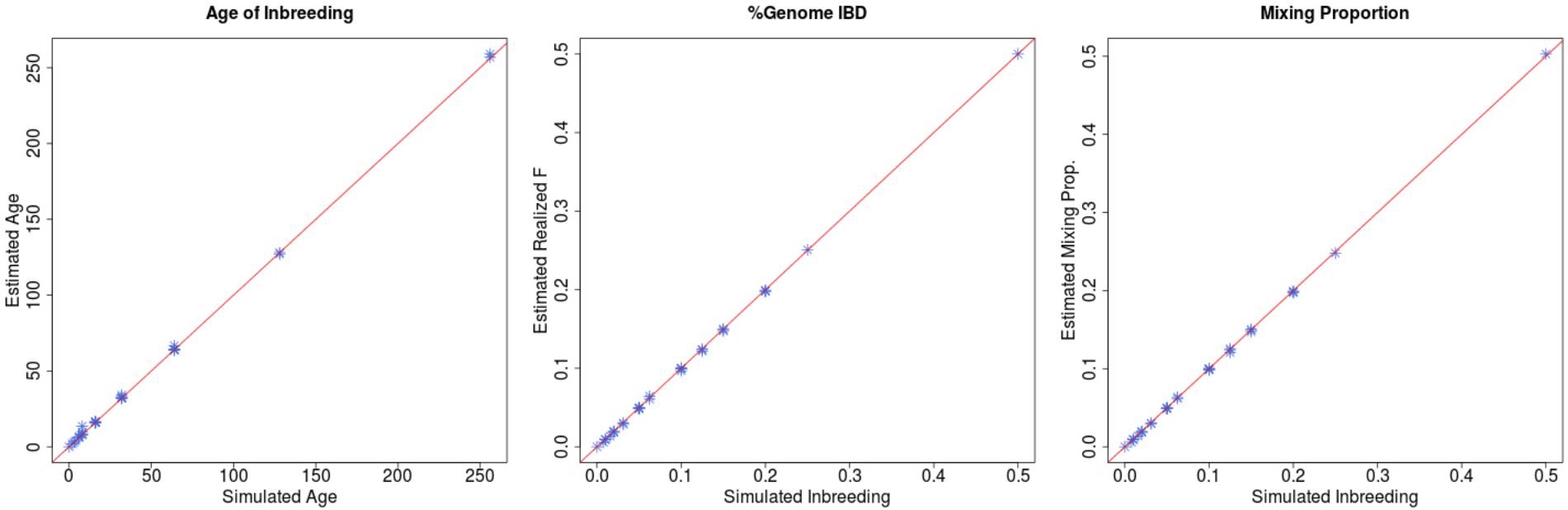
Mixture of inbreeding classes

- Mixture of several IBD and nonIBD with different age (G)
- Emission probabilities unchanged
- Transition probabilities same principle
  - Each distribution with its own mixing proportions

1002011010211110020020202020001211021011012010122001 1

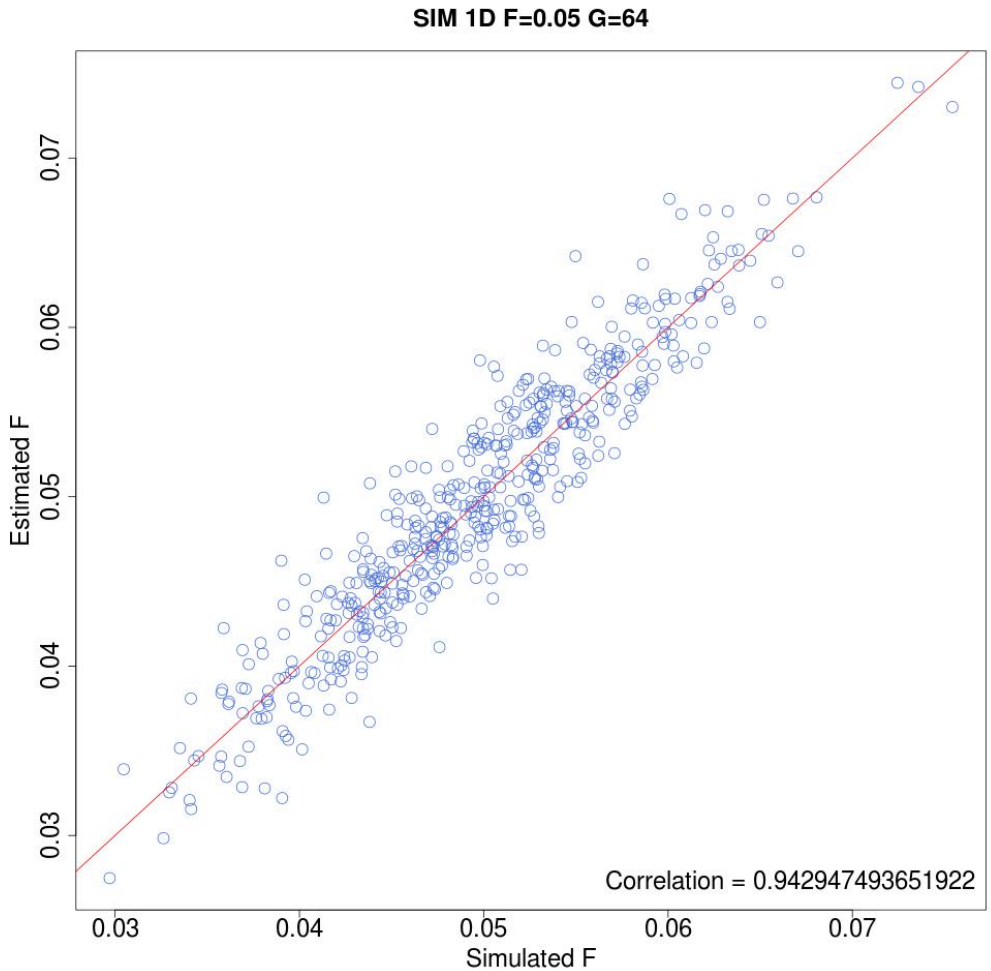
Testing with simulations

- One distribution (1 age), 500 individuals, medians



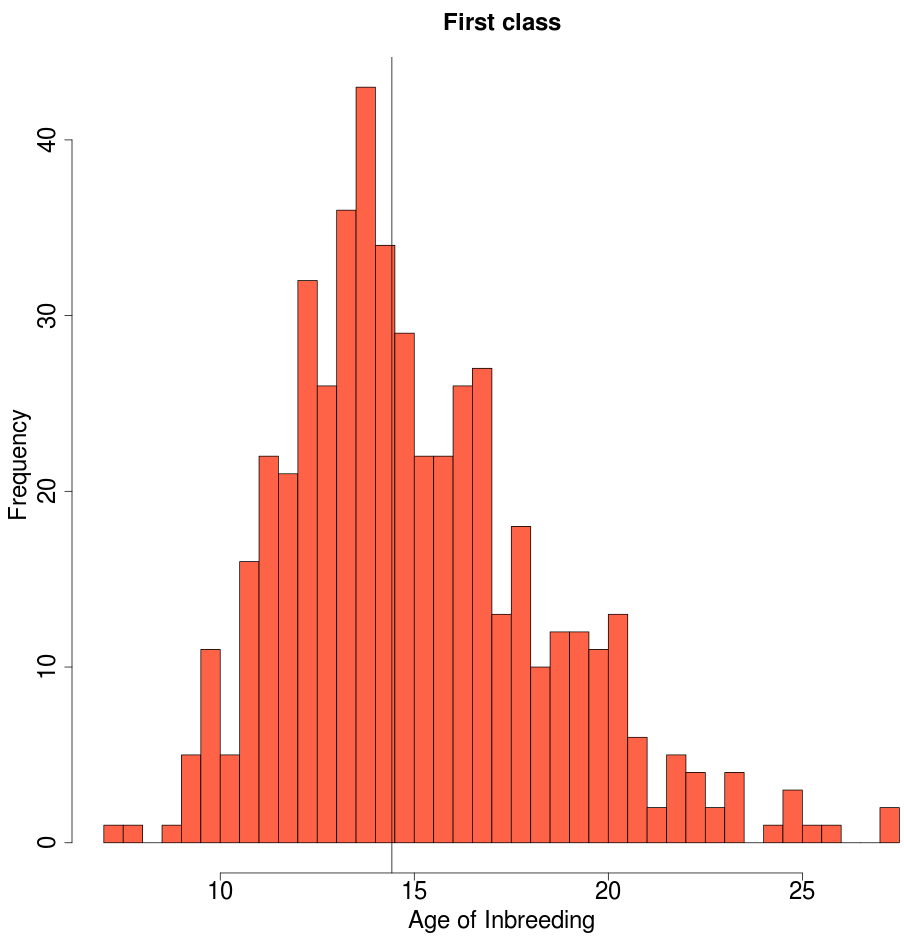
Estimated F ~ Simulated F

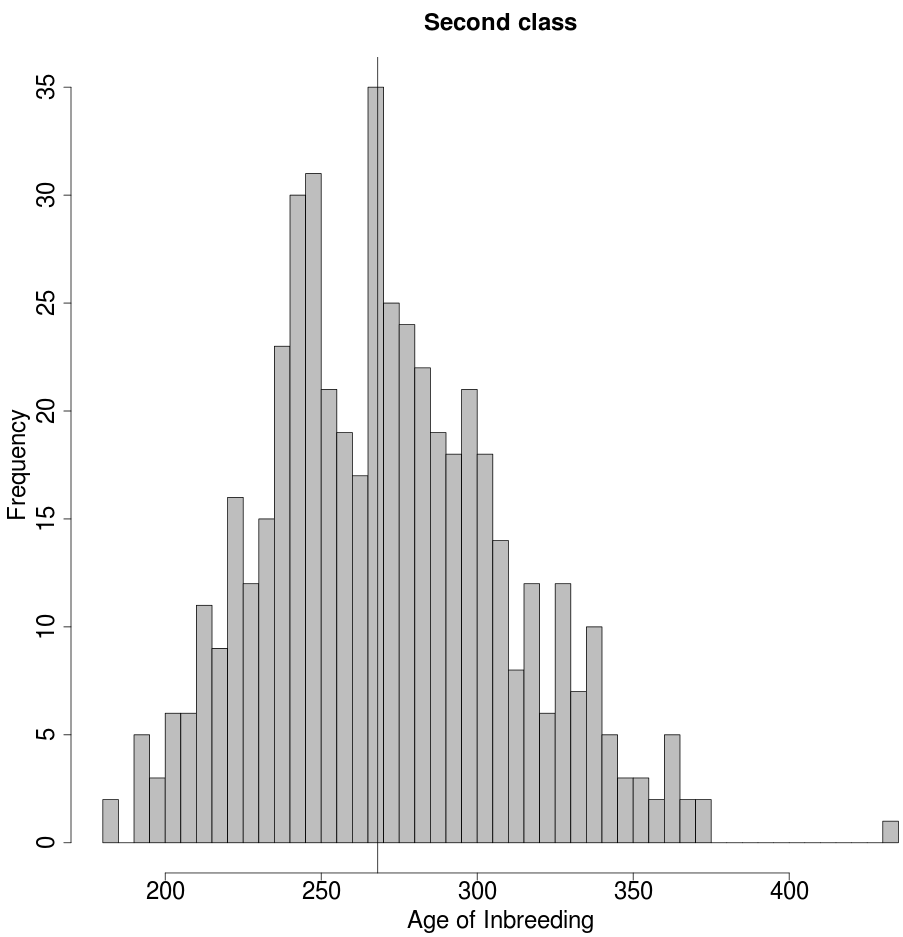
- Simulated F = 0.05 and G = 64



Two simulated distributions

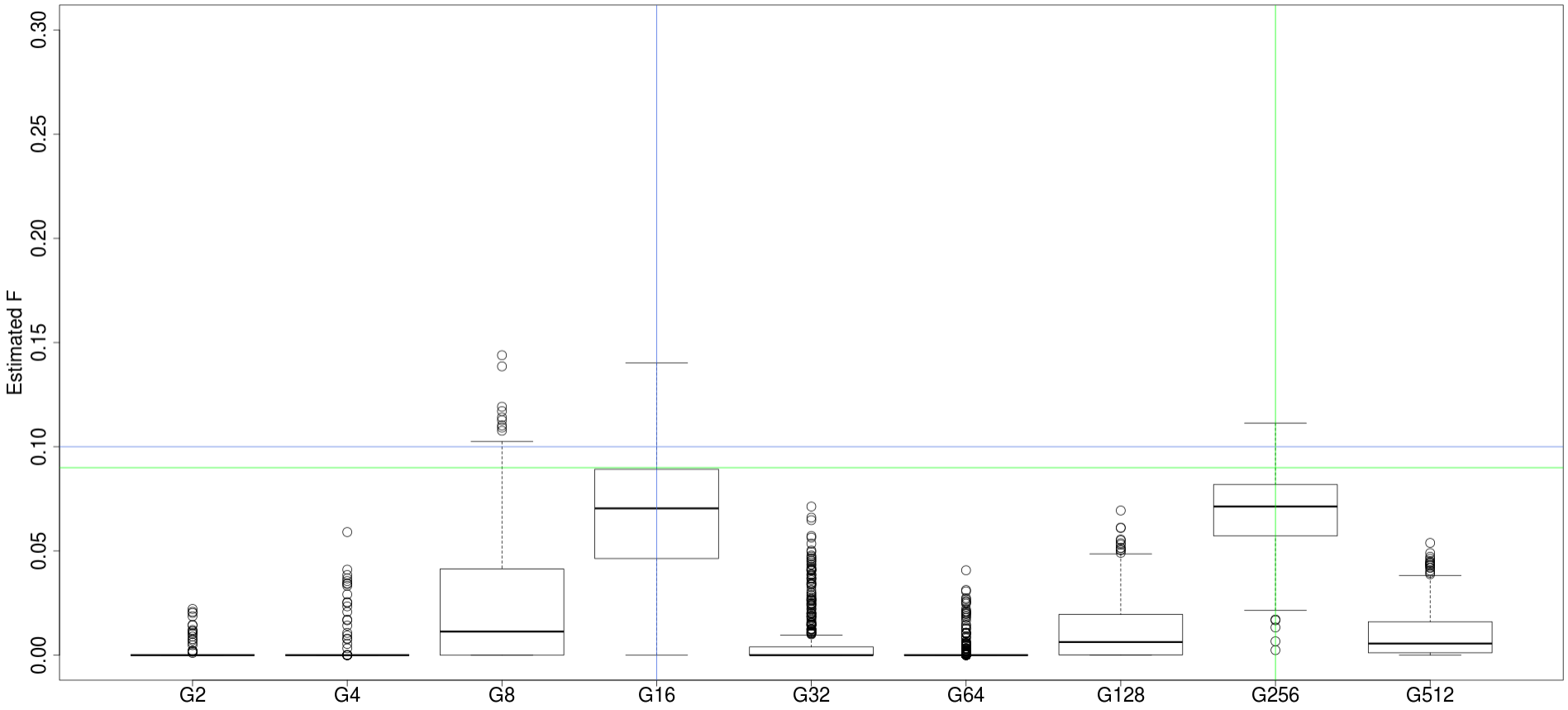
- Simulated Age, G1 = 16 & G2 = 256





Two simulated distributions

- Mixture of 10 predefined classes (9 IBD, 1 nonIBD)

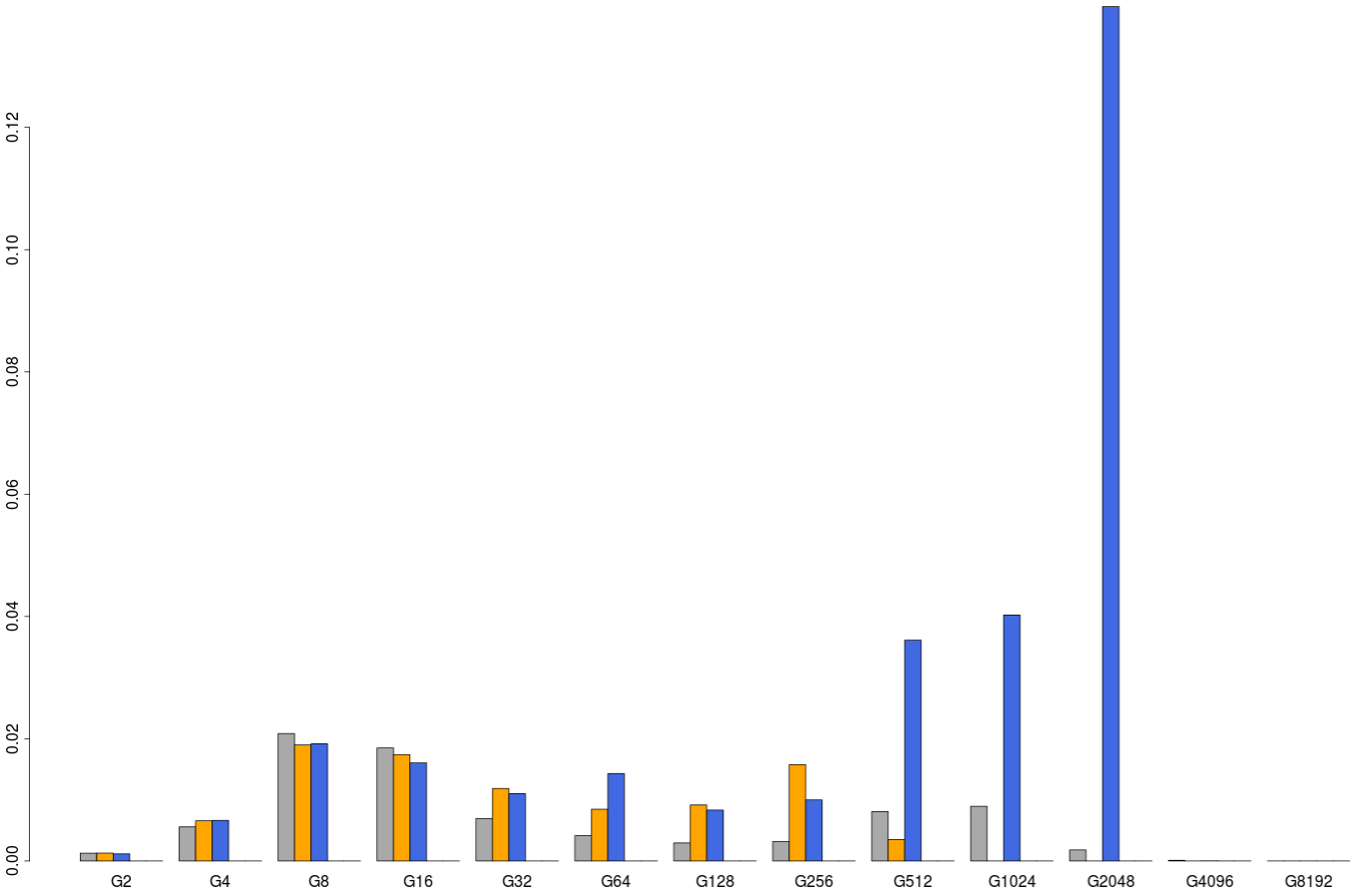


Summary of simulations

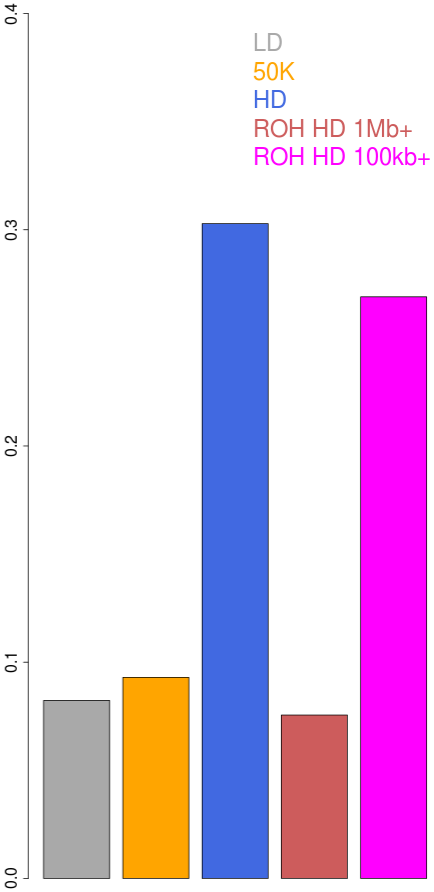
- Simulations with varying age, number of distributions, type of markers, low-fold sequencing data, errors
- Assessing with estimated age, mixing (1 dist.), global F, , local F, population and individual estimates, estimating K
- Better when younger F, larger F, more markers, higher MAF, higher cover, large age differences

Belgian Blue cattle (634 bulls)

Proportion inbreeding per age class

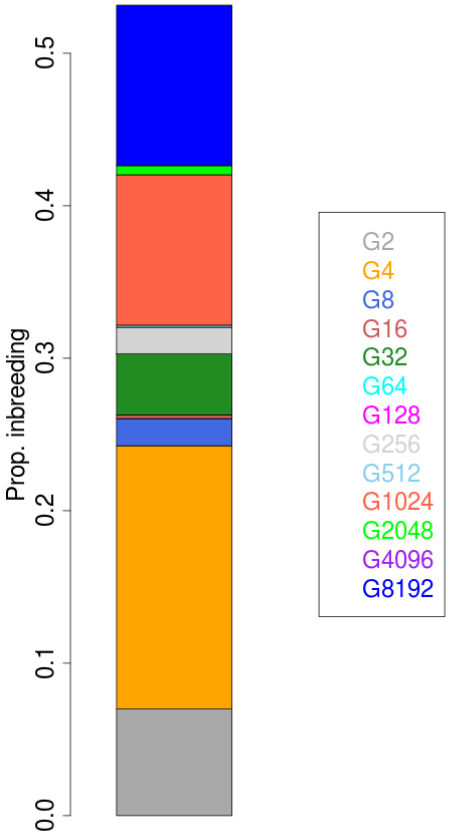


Total F



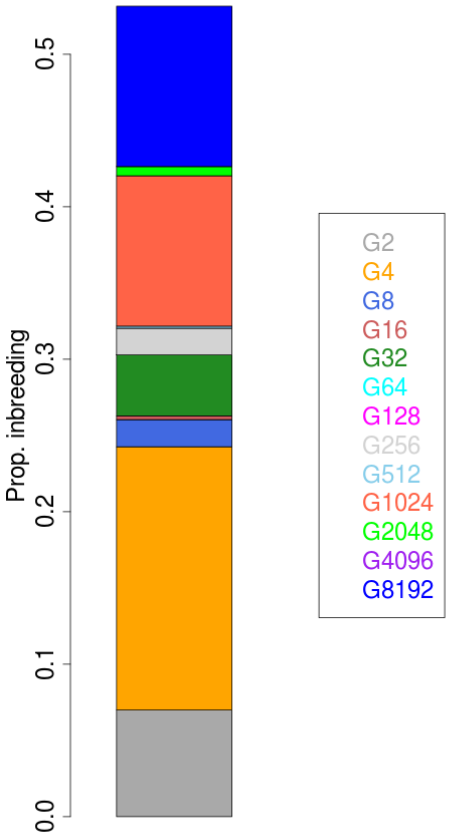
WGS data (high cover @114x)

- Sire x MGS mating: expected 25% at G3



WGS data (high cover @114x)

- Sire x MGS mating: expected 25% at G3



Chr	Length (Mb)	#het snps	#snps	Prop. het
2	92.385886	23	192567	1.2e-4
1	51.469735	0	117044	0
21	46.047682	1	107278	9.3e-6
16	44.281690	0	81934	0
2	34.592319	13	80042	1.6e-4
4	33.943960	4	84630	4.7e-5
4	32.406205	0	64784	0
20	30.317150	6	70982	8.4e-5
10	27.445232	2	62643	3.2e-5
23	26.648470	1	74953	1.3e-5

BBB WGS (@10-15x)

- Longest IBD segments for one sire

Chr	Lenght (HD)	#Het	#SNPs	Length (WGS geno)	#Het	#SNPs	Prop. Het	Lenght (Gen. Lik)
9	94.6	2	23298	84.6	375	182480	0.0025	94.6
22	46.4	1	11834	34.1	82	69465	0.0012	45.2
13	34.0	0	7031	31.3	141	59879	0.0023	34.1
20	20.6	0	5418	20.5	127	48748	0.0026	20.7
8	16.2	0	3331	9.3	41	19566	0.0021	16.2

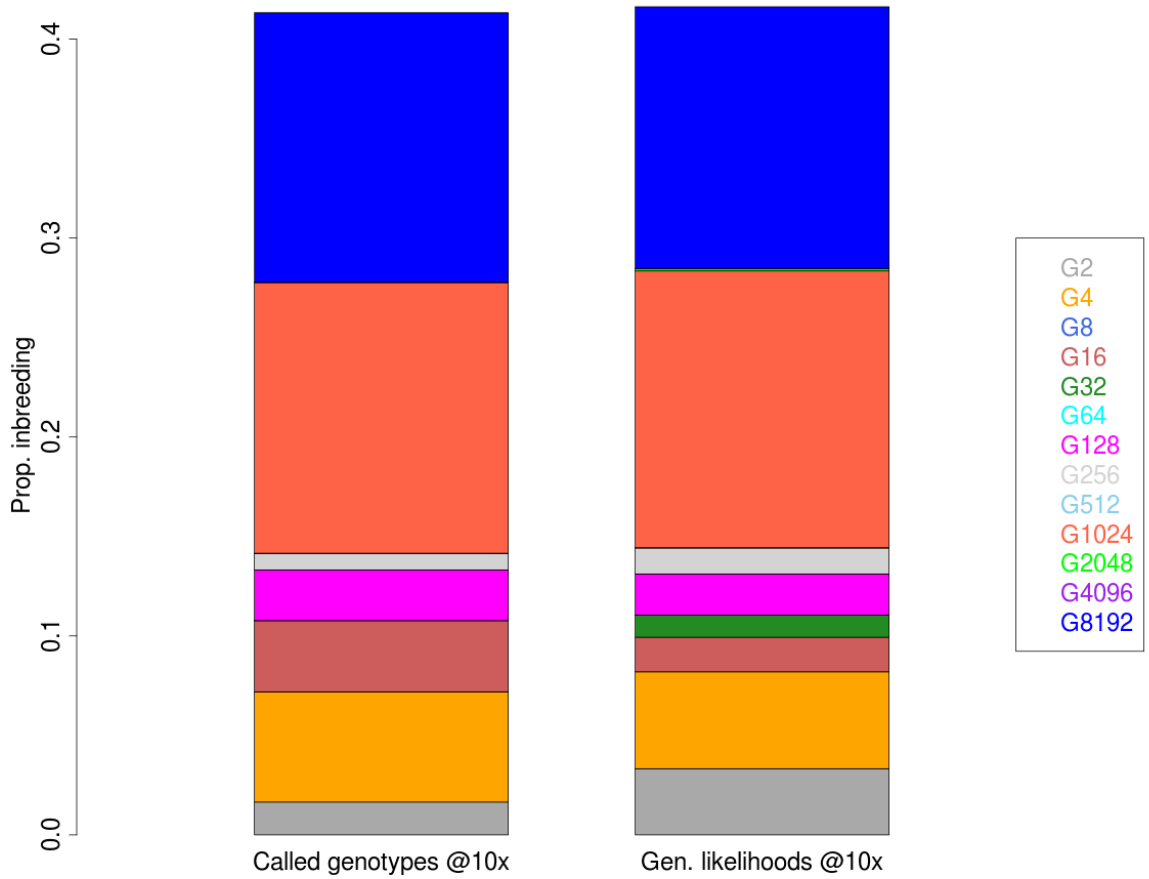
BovineHD

WGS called genotypes

WGS likelihoods

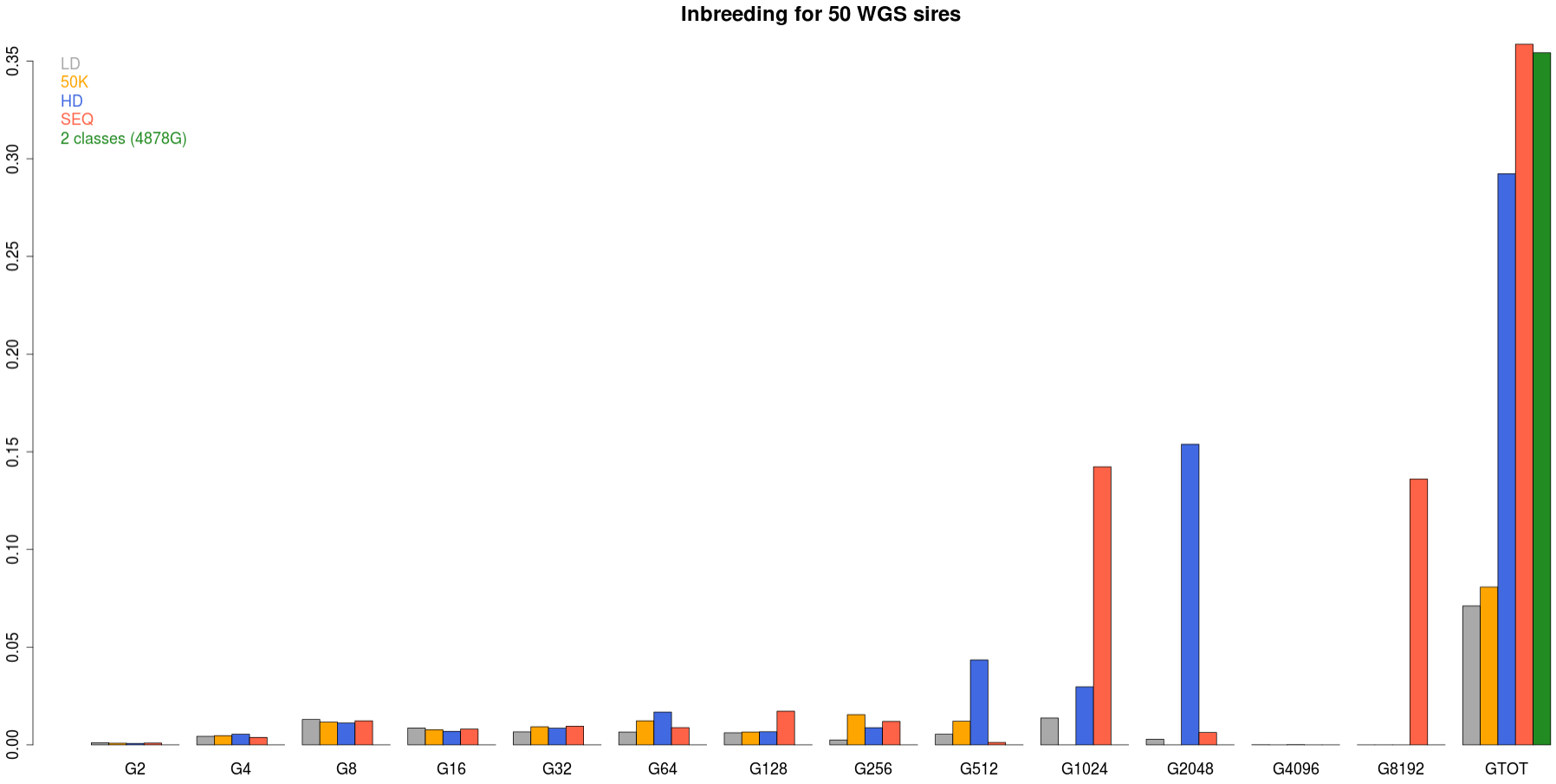
BBB WGS (@10-15x)

- Repartition in IBD classes (geno vs gen. likelihoods)



Whole Genome Sequence

- 50 sequenced Belgian Blue sires



Conclusions

- The model uses all the information
  - Sequence of genotypes, allele frequencies, error rates
- The model classifies inbreeding in different age classes
  - Better than just one (open perspectives)
- The model estimates local and global inbreeding
- The model can work with genotyping arrays and sequence data
  - With different allelic spectra