Missing ROH ?

Recommendations for tuning PLINK in runs of homozygosity analyses on medium density genotypes

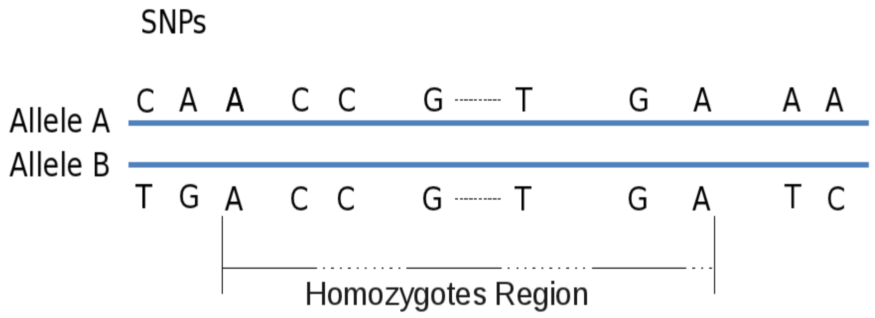
Roel Meyermans, W. Gorssen, N. Buys and S. Janssens

Session 54, Thursday 29th of August 2019

EAAP 2019, Ghent

Runs Of Homozygosity

- Homozygous segments assumed to arise from a common ancestor



- State-of-the-art method for inbreeding analyses

- Detection of ROH islands as signatures of selection

→ PLINK (v1.9, Chang et al. 2015) is commonly used for ROH detection

PLINK ROH detection algorithm

Scanning window approach:

1. Scanning window definition

   (-window-snp, -window-missing and -window-het)

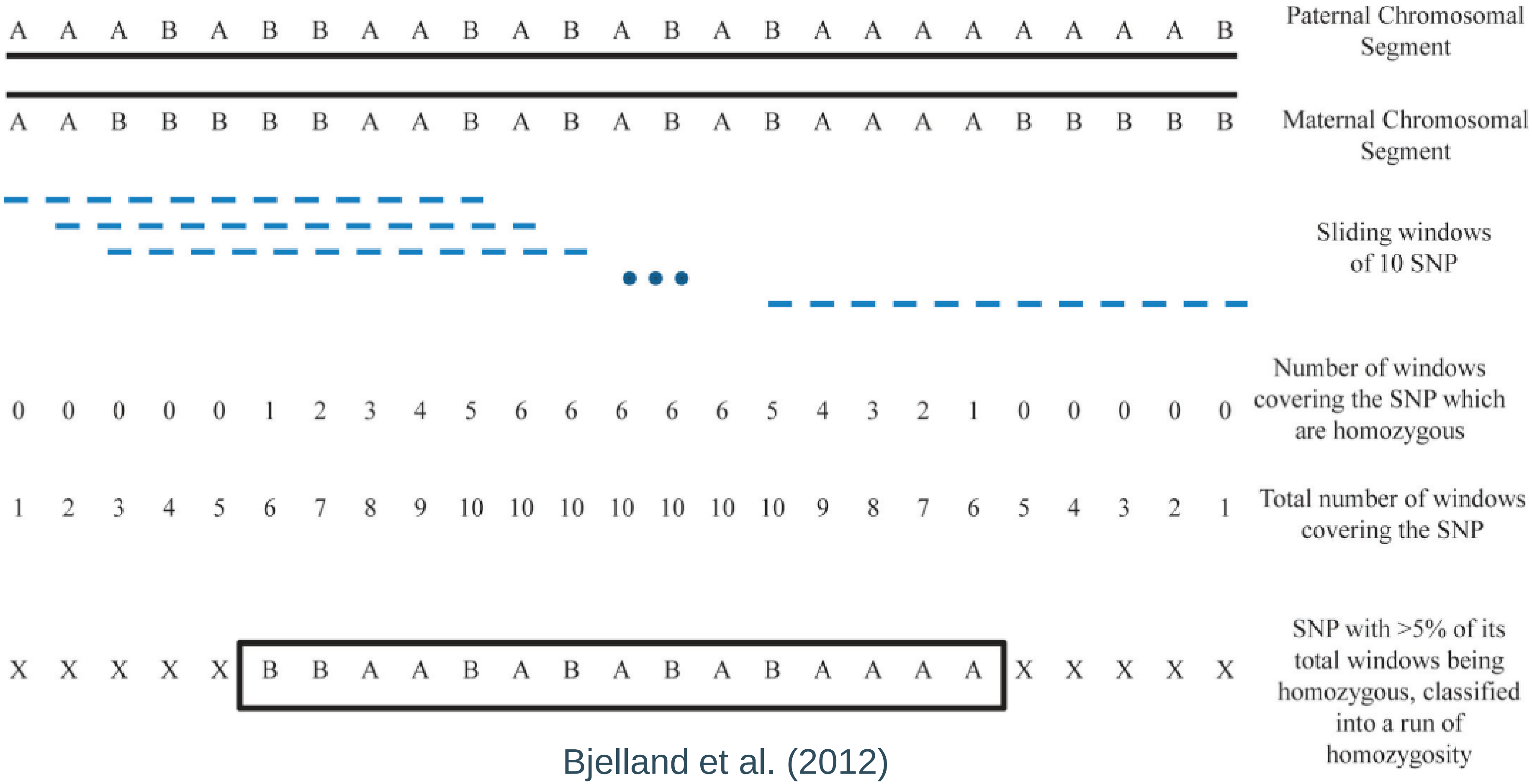
2. Every individual SNP's proportion of appearance in homozygous windows is calculated

3. SNPs passing threshold → potential ROH

4. Extra constraints to identify ROH

   (-gap, -density, -snp, -kb and -het)

PLINK ROH detection algorithm



A	A	A	B	A	B	B	A	A	B	A	B	A	B	A	B	A	A	A	A	A	A	A	A	B	Paternal Chromosomal Segment
A	A	B	B	B	B	B	A	A	B	A	B	A	B	A	B	A	A	A	A	B	B	B	B	B	Maternal Chromosomal Segment
																									Sliding windows of 10 SNP
0	0	0	0	0	1	2	3	4	5	6	6	6	6	6	5	4	3	2	1	0	0	0	0	0	Number of windows covering the SNP which are homozygous
1	2	3	4	5	6	7	8	9	10	10	10	10	10	10	10	9	8	7	6	5	4	3	2	1	Total number of windows covering the SNP
X	X	X	X	X	B	B	A	A	B	A	B	A	B	A	B	A	A	A	A	X	X	X	X	X	SNP with >5% of its total windows being homozygous, classified into a run of homozygosity

Bjelland et al. (2012)

Impact of (PLINK) settings on ROH detection

Data quality control

➡ Pruning for low MAF

➡ Pruning for LD

PLINK settings

Scanning window

- --homozyg-window-snp
- --homozyg-window-het
- --homozyg-window-missing
- --homozyg-window-threshold

ROH definition

- --homozyg-snp
- --homozyg-kb
- ➡ --homozyg-density
- --homozyg-gap
- --homozyg-het

Material and Methods

- Review of recent papers (pigs, cattle, sheep, horses,…)
- Test all (previously undiscussed) settings independently
- Use own and publicly availabel data of different species and populations

Material and Methods

Basic testing conditions

- No MAF pruning
- No LD pruning
- Density 200 $\frac{kb}{SNP}$
- Gap 2 Mb
- Threshold 0.05
- Scanning window size 20 SNPs
- Minimal ROH length 1 Mb
- Minimal number of SNPs calculated by Purfield et al. 2012
- Maximum 1 missing SNP and no heterozygous SNPs

Genome Coverage

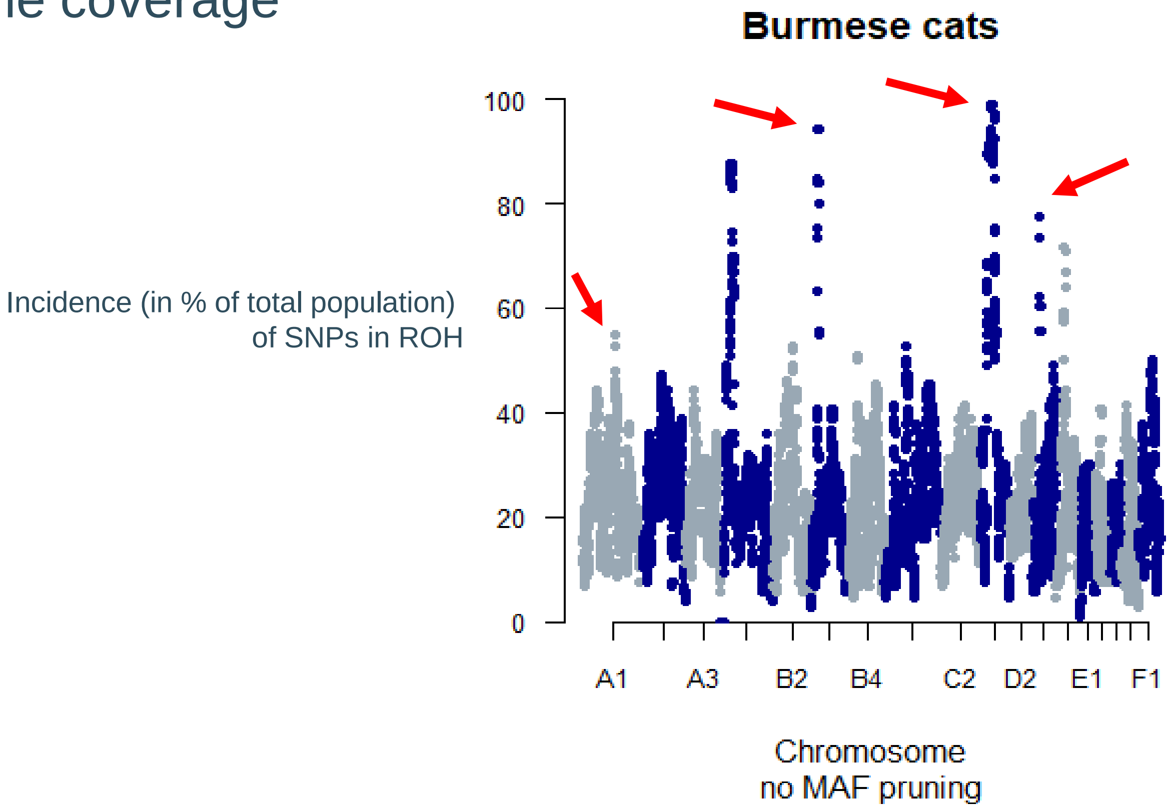
$$Genome\ coverage = \frac{ROH_{max}}{Total\ genome\ length}$$

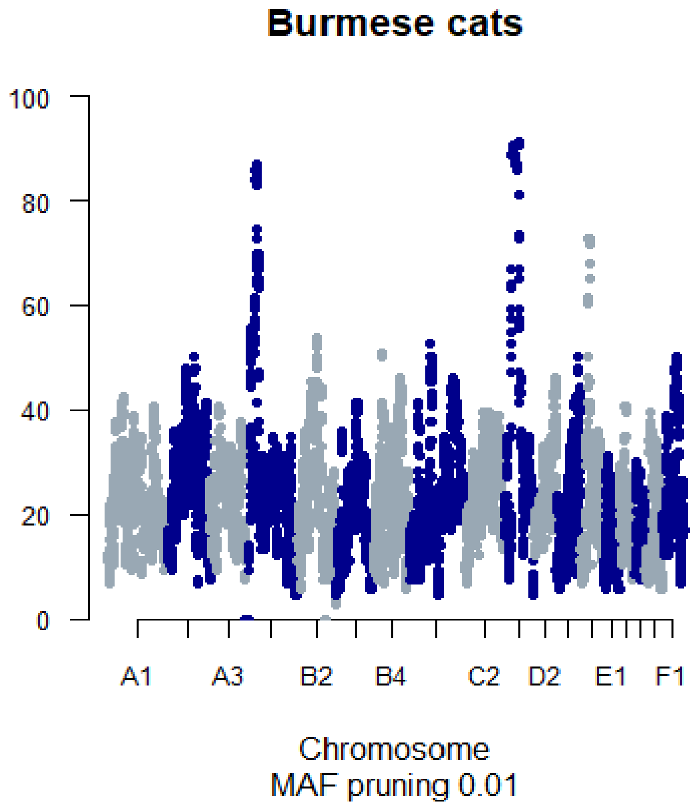
Where ROH_{max} is calculated as the total ROH length of a completely homozygous individual using the current analysis settings.

Correct settings would give ± 100% genome coverage (F_{ROH}=100%)

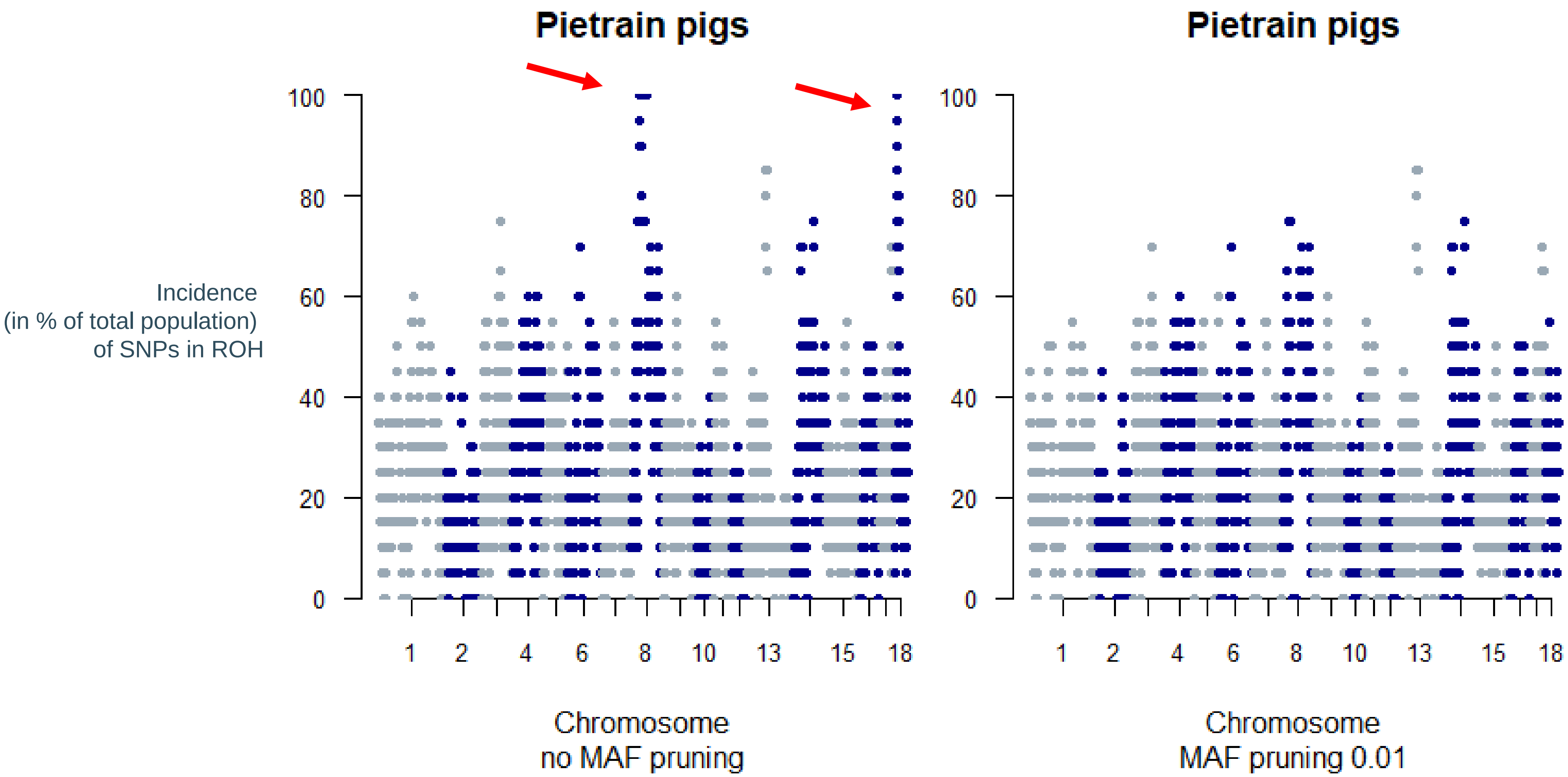
Pruning for low MAF

Discarding SNPs with low minor alleles leads to undetected ROH (islands) and a drop in genome coverage



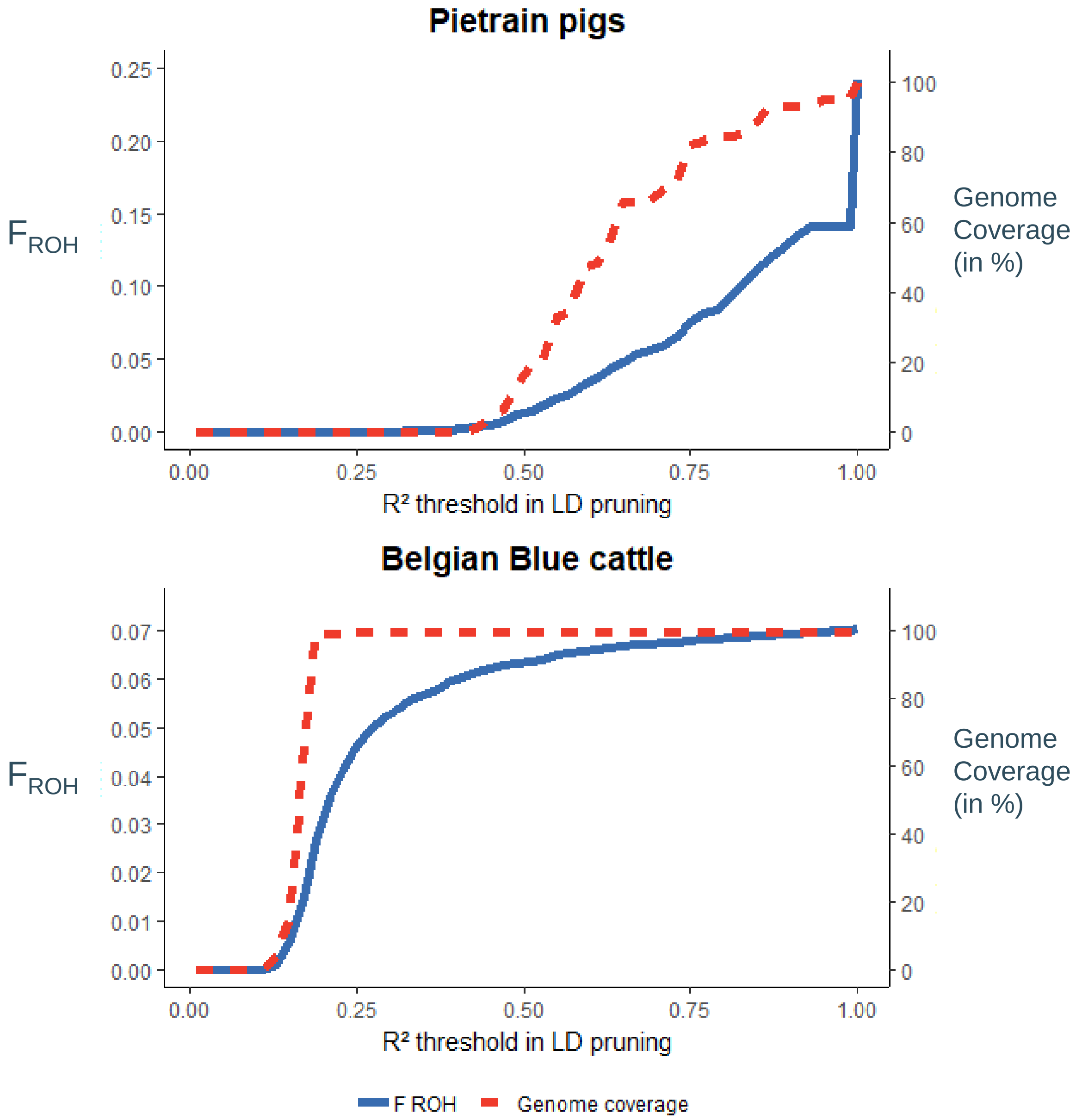


Pruning for low MAF



Pruning for LD

- Implications on ROH detection
- Genome coverage drop
- Effect is population dependent



Pruning for LD and MAF

For GWAS:

- Rare alleles (MAF < 0.05) are of little interest
- Highly correlated SNPs (LD) only slow down the analysis

For ROH detection:

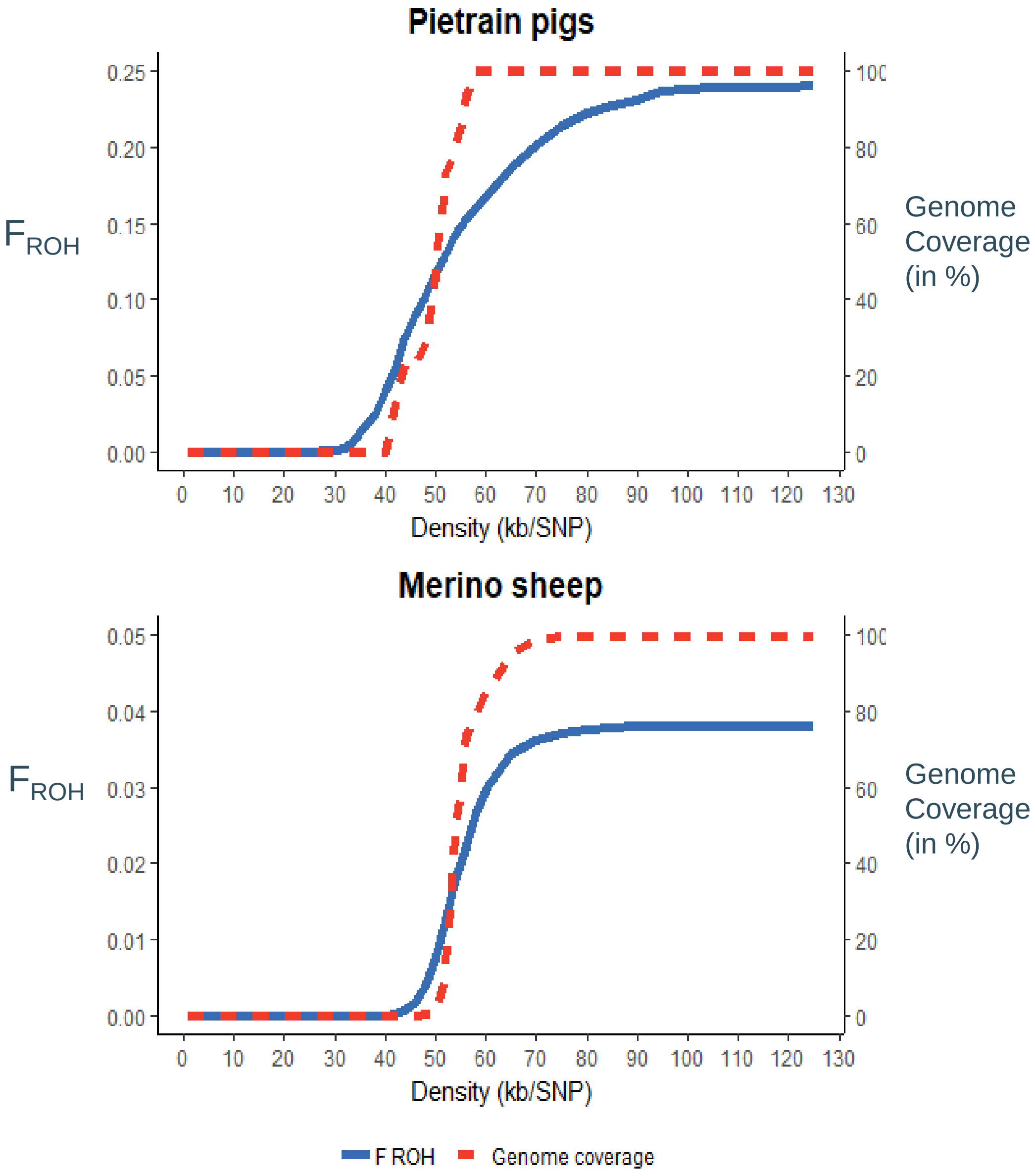
- No harm in including rare alleles
- Computation time is not that critical using medium density genotypes

Minimal SNP Density

- SNP density expressed in minimal $\frac{\mathrm{kb}}{\mathrm{SNP}}$

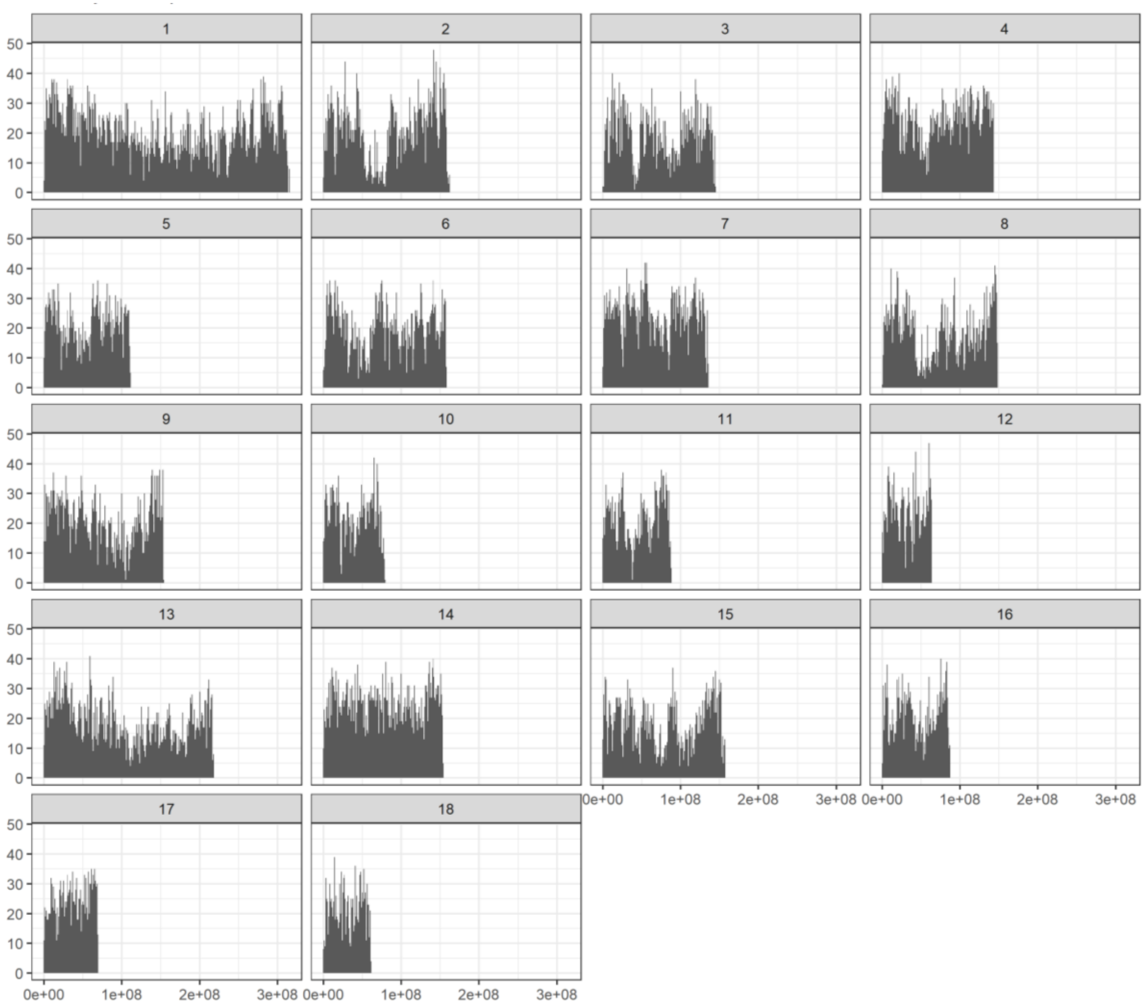
- Default PLINK setting is minimal 50 $\frac{kb}{SNP}$

- Population and array dependent



Local SNP density differences

SNP density (in # SNPs/Mb)



Location in the genome

Conclusions

- MAF and LD pruning affects ROH detection and is perhaps unnescessary in ROH detection
- Low minimal density (in kb/SNP) can lead to low genome coverage
- Calculating genome coverage helps to detect problems
- Report all PLINK settings in your publications

Acknowledgements

Complete results and discussion are reported in
Meyermans & Gorssen et al., under review with BMC Genomics

Dr. C. Chang, Dr. I. Curik and Dr. J. Sölkner for their input

SB PhD grant of the Research Foundation Flanders (FWO) (1S37119N)

Thank you for your attention

roel.meyermans@kuleuven.be

Minimal ROH length (SNP)

The minimal number of SNPs in a ROH was determined by the formula proposed by Lencz et al. and adapted by Purfield et al.:

$$L = \frac{log_e \frac{\alpha}{n_s n_i}}{log_e(1 - het)},$$

with n_s the number of genotyped SNPs per individual, n_i the number of genotyped individuals, α the percentage of false positive ROH (0.05) and het the mean heterozygosity across all SNPs.