HAPIH Hrvatska agencija za poljoprivredu i hranu | Croatian Agency for Agriculture and Food | L'Agence Croate pour l'Agriculture et l'Alimentation

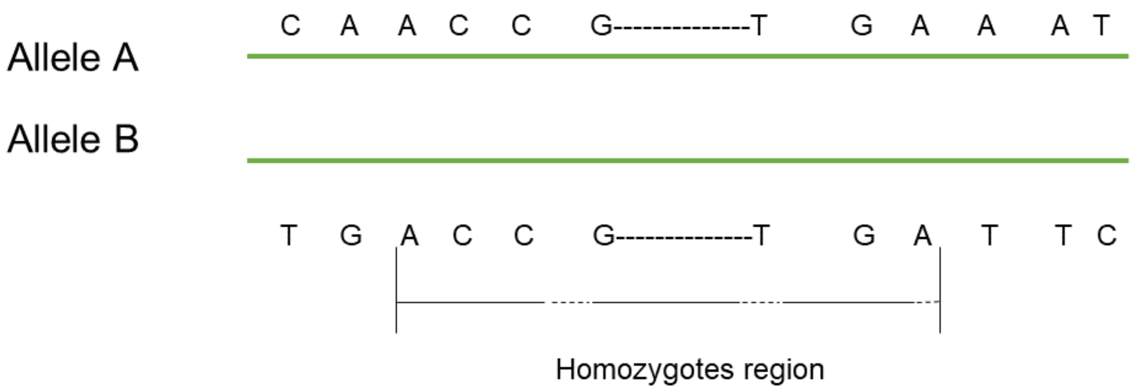
Estimation of genomic inbreeding (F_{ROH}) in Croatian Holstein population

Marija Špehar[1], Dora Ceranac[1], Zdenko Ivkić[1], Drago Solić[1], Ante Kasap[2]

[1]Croatian Agency for Agriculture and Food, Svetošimunska 25, 10000 Zagreb, Croatia
[2]University of Zagreb Faculty of Agriculture, Svetošimunska 25, 10000 Zagreb, Croatia

Introduction

- Runs of homozygosity (ROH)
- Segments from both parents are the same in the offspring



C A A C C G--------------T G A A A T

Allele A

Allele B

T G A C C G--------------T G A T T C

Homozygotes region

Runs of homozygosity (ROH)

- Number and length of ROH reflect individual demographic history
  - Long ROHs – recent inbreeding
  - Short ROHs – ancient inbreeding
- Investigation of genetic architecture of complex traits or diseases - opportunity to dissect genomic regions affecting complex traits
- Complementary to GWAS, inbreeding depression studies and genomic prediction

Objective

- To estimate the inbreeding using ROH (F_{ROH}) in Croatian Holstein population

  - – In order to provide a valuable clue for future research

Material

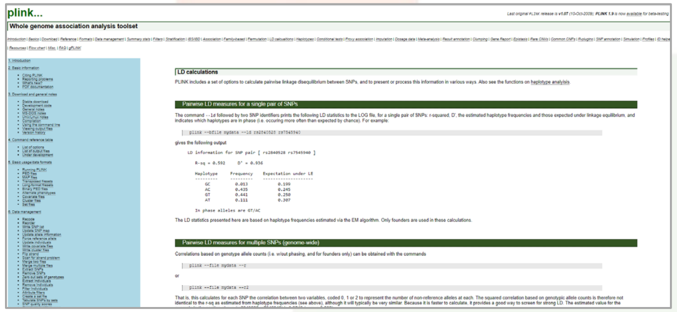
- 89 Holstein females
  - from 29 herds
  - progenies of 31 sires (CAN, DNK, DEU, FRA, NLD, USA)
- Illumina BovineSNP50K BeadChip – 52,445 SNPs
- Quality control
  - autosomal SNPs only
  - call rate: SNP 0.05 & animal 0.10
  - GC score: >0.7
  - **45,661 SNPs and 85 animals**

Method

- Identifying ROH (Howrigan et al., 2011 ,Ferenčaković et al., 2013; Meyermans et al., 2020, Gorssen et al., 2021)
- Sliding window approach
  - Homozyg Length (1 Mb)
  - SNP ROH density (1 SNP/150 Kb)
  - Homozyg minSNP (44) (Lencz et al., 2007, Purfield et al., 2012)
  - SNP gap (1 Mb)
  - Homozyg het (0)
  - Homozyg miss SNP (1)

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

Method cont…



plink...

Last original PLINK release is v1.07 (10-Oct-2009); PLINK 1.9 is now available for beta-testing

Whole genome association analysis toolset

Introduction | Basics | Download | Reference | Formats | Data management | Summary stats | Filters | Stratification | IBS/IBD | Association | Family-based | Permutation | LD calculations | Haplotypes | Conditional tests | Proxy association | Imputation | Dosage data | Meta-analysis | Result annotation | Clumping | Gene Report | Epistasis | Rare CNVs | Common CNPs | R-plugins | SNP annotation | Simulation | Profiles | ID helper | Resources | Flow chart | Misc. | FAQ | gPLINK

1. Introduction
2. Basic information
   - Citing PLINK
   - Reporting problems
   - What's new?
   - PDF documentation
3. Download and general notes
   - Stable download
   - Development code
   - General notes
   - MS-DOS notes
   - Unix/Linux notes
   - Compilation
   - Using the command line
   - Viewing output files
   - Version history
4. Command reference table
   - List of options
   - List of output files
   - Under development
5. Basic usage/data formats
   - Running PLINK
   - PED files
   - MAP files
   - Transposed filesets
   - Long-format filesets
   - Binary PED files
   - Alternate phenotypes
   - Covariate files
   - Cluster files
   - Set files
6. Data management
   - Recode
   - Reorder
   - Write SNP list
   - Update SNP map
   - Update allele information
   - Force reference allele
   - Update individuals
   - Write covariate files
   - Write cluster files
   - Flip strand
   - Scan for strand problem
   - Merge two files
   - Merge multiple files
   - Extract SNPs
   - Remove SNPs
   - Zero out sets of genotypes
   - Extract individuals
   - Remove individuals
   - Filter individuals
   - Attribute filters
   - Create a set file
   - Tabulate SNPs by sets
   - SNP quality scores

LD calculations

PLINK includes a set of options to calculate pairwise linkage disequilibrium between SNPs, and to present or process this information in various ways. Also see the functions on haplotype analyisis.

Pairwise LD measures for a single pair of SNPs

The command `--ld` followed by two SNP identifiers prints the following LD statistics to the LOG file, for a single pair of SNPs: r-squared, D', the estimated haplotype frequencies and those expected under linkage equilibrium, and indicates which haplotypes are in phase (i.e. occuring more often than expected by chance). For example:

```
plink --bfile mydata --ld rs2840528 rs7545940
```

gives the following output

```
  LD information for SNP pair [ rs2840528 rs7545940 ]

    R-sq = 0.592     D' = 0.936

    Haplotype     Frequency    Expectation under LE
    ---------     ---------    --------------------
       GC          0.013           0.199
       AC          0.435           0.245
       GT          0.441           0.250
       AT          0.111           0.307

    In phase alleles are GT/AC
```

The LD statistics presented here are based on haplotype frequencies estimated via the EM algorithm. Only founders are used in these calculations.

Pairwise LD measures for multiple SNPs (genome-wide)

Correlations based on genotype allele counts (i.e. w/out phasing, and for founders only) can be obtained with the commands

```
plink --file mydata --r
```

or

```
plink --file mydata --r2
```

That is, this calculates for each SNP the correlation between two variables, coded 0, 1 or 2 to represent the number of non-reference alleles at each. The squared correlation based on genotypic allele counts is therefore not identical to the r-sq as estimated from haplotype frequencies (see above), although it will typically be very similar. Because it is faster to calculate, it provides a good way to screen for strong LD. The estimated value for the

- PLINK
  - – Quality control analysis
  - – Estimation of lengths of the ROH segments of different size
  - – ROH length category (1-2 Mb, 2-4 Mb, 4-8 Mb, 8-16, Mb >16 Mb)

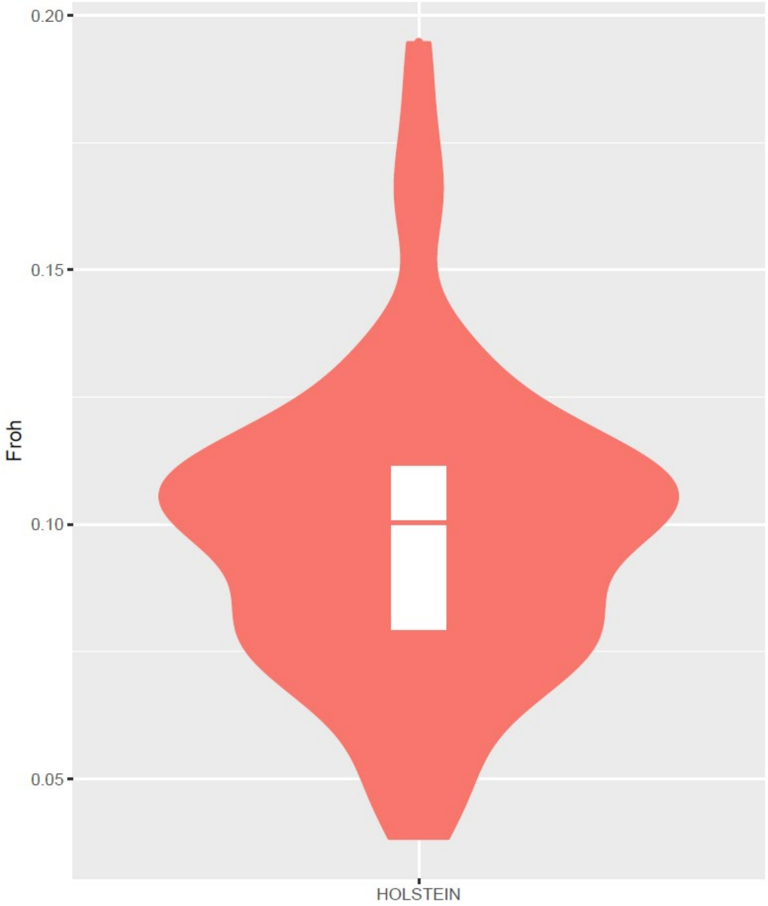
- R programme
  - – Post hoc analysis i.e. estimation of genomic $F_{ROH(>4Mb)}$

$$F_{ROH(>4Mb)} = \frac{\sum_i L(ROH_i)}{L\ (total\ length\ of\ genome\ covered\ by\ SNP)}$$

Results

Results (1)

- Individual F_{ROH} – from 4.33% to 19.74%
- **Average** – **10.25%**

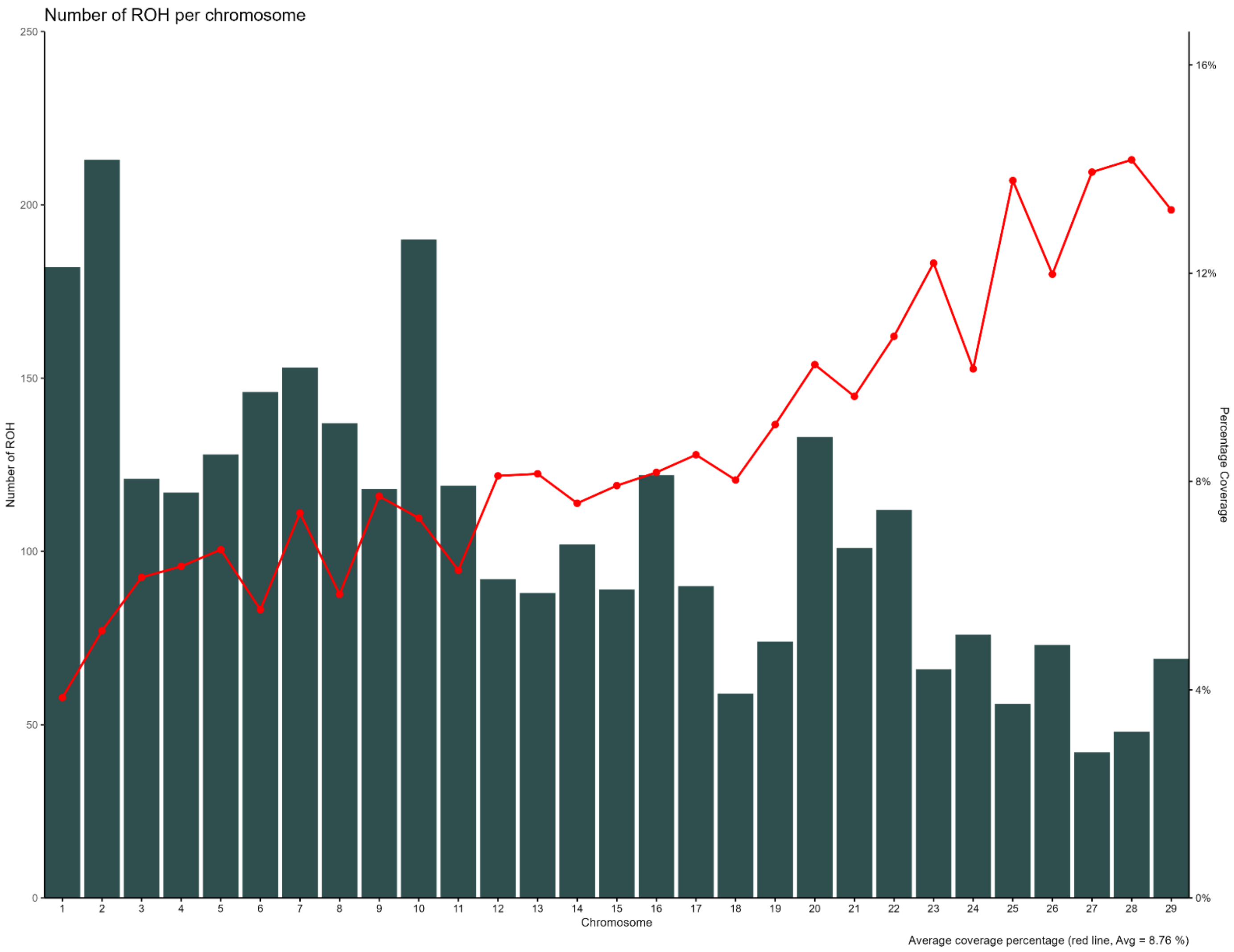


Results (2)

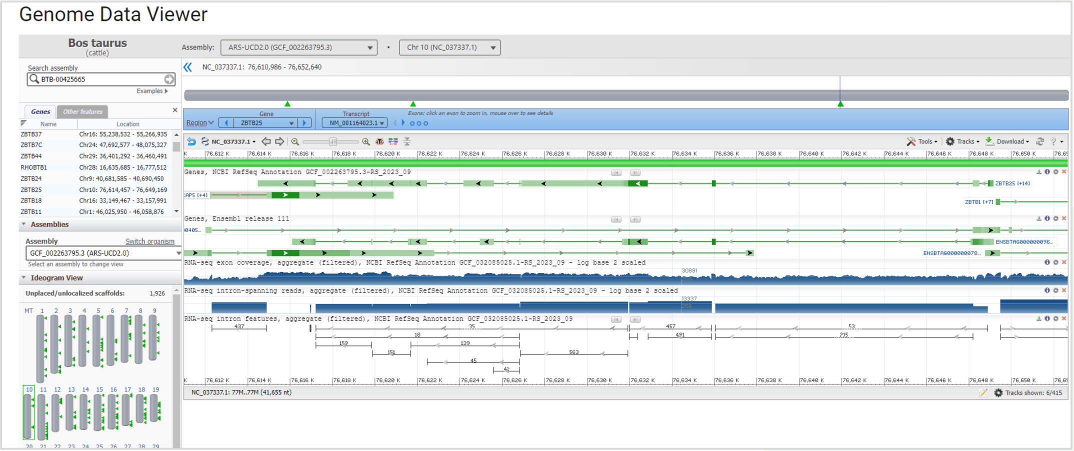
- Total of 3,116 autosomal ROHs - from 1.0 to 59.5 Mb
- ~37 ROHs per animal with an average length of ~6.94 Mb

ROH Category (Mb)	Frequency (%)	Length proportion of category/total length of the genome covered by ROHs (%)
1-2	2.7	0.2
2-4	33.9	12.2
4-8	37.0	30.9
8-16	19.3	31.5
> 16	7.1	25.2

Number of ROH per chromosome and average percentage of each chromosome covered by ROH



ROH incidence plot



Genome Data Viewer

Bos taurus (cattle)

Assembly: ARS-UCD2.0 (GCF_002263795.3) · Chr 10 (NC_037337.1)

Search assembly: BTB-00425665

NC_037337.1: 76,610,986 - 76,652,640

Name	Location
ZBTB37	Chr16: 55,238,532 - 55,266,935
ZBTB7C	Chr24: 47,692,577 - 48,075,327
ZBTB44	Chr29: 36,401,292 - 36,460,491
RHOBTB1	Chr28: 16,635,685 - 16,777,512
ZBTB24	Chr9: 40,681,585 - 40,690,450
ZBTB25	Chr10: 76,614,457 - 76,649,169
ZBTB18	Chr16: 33,149,467 - 33,157,991
ZBTB11	Chr1: 46,025,950 - 46,058,876

Assembly: GCF_002263795.3 (ARS-UCD2.0)

Select an assembly to change view

Unplaced/unlocalized scaffolds: 1,926

Gene: ZBTB25 · Transcript: NM_001164023.1

Exons: click an exon to zoom in, mouse over to see details

Genes, NCBI RefSeq Annotation GCF_002263795.3-RS_2023_09

Genes, Ensembl release 111

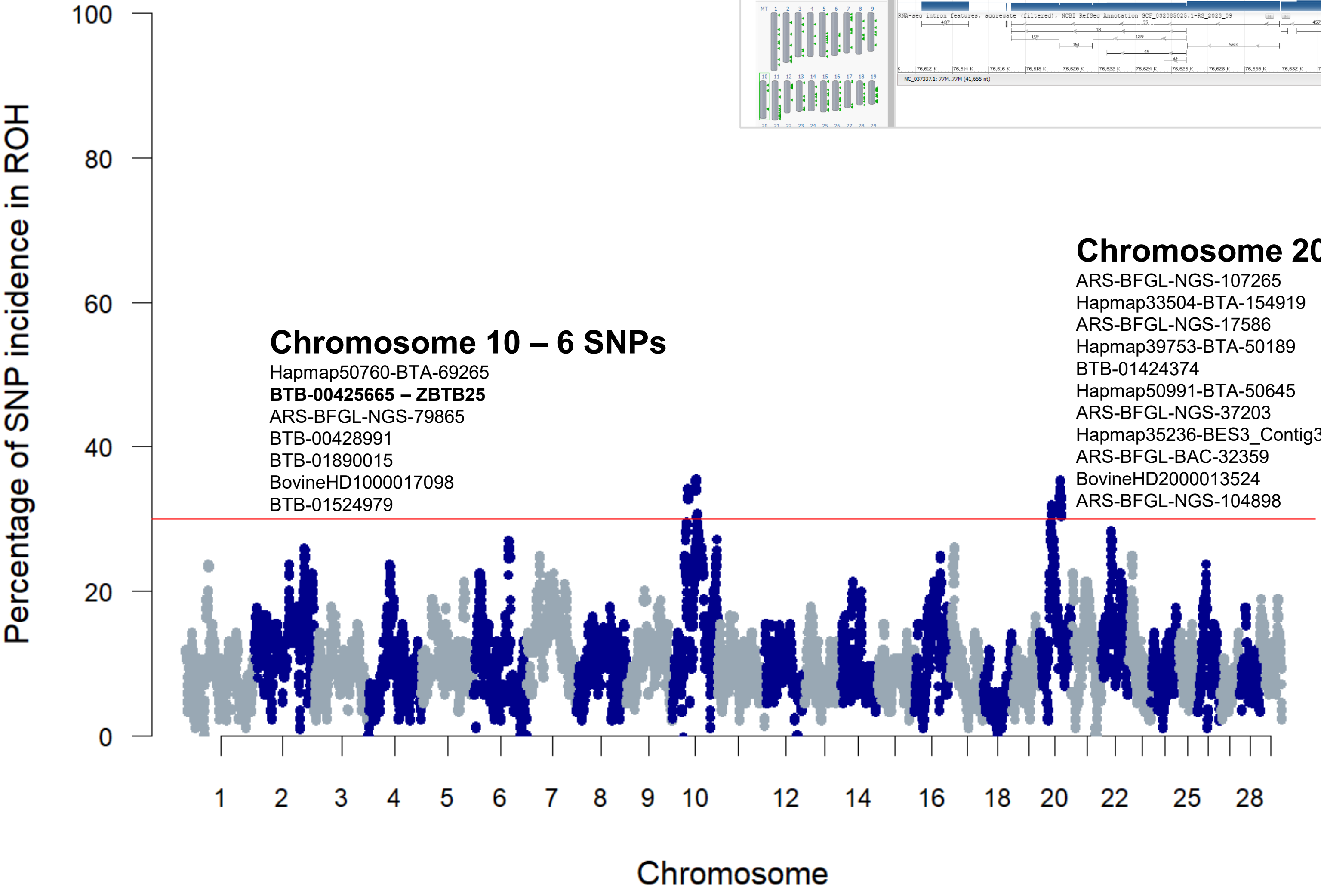
RNA-seq exon coverage, aggregate (filtered), NCBI RefSeq Annotation GCF_032085025.1-RS_2023_09 - log base 2 scaled

RNA-seq intron-spanning reads, aggregate (filtered), NCBI RefSeq Annotation GCF_032085025.1-RS_2023_09 - log base 2 scaled

RNA-seq intron features, aggregate (filtered), NCBI RefSeq Annotation GCF_032085025.1-RS_2023_09

NC_037337.1: 77M..77M (41,655 nt)

Tracks shown: 6/415



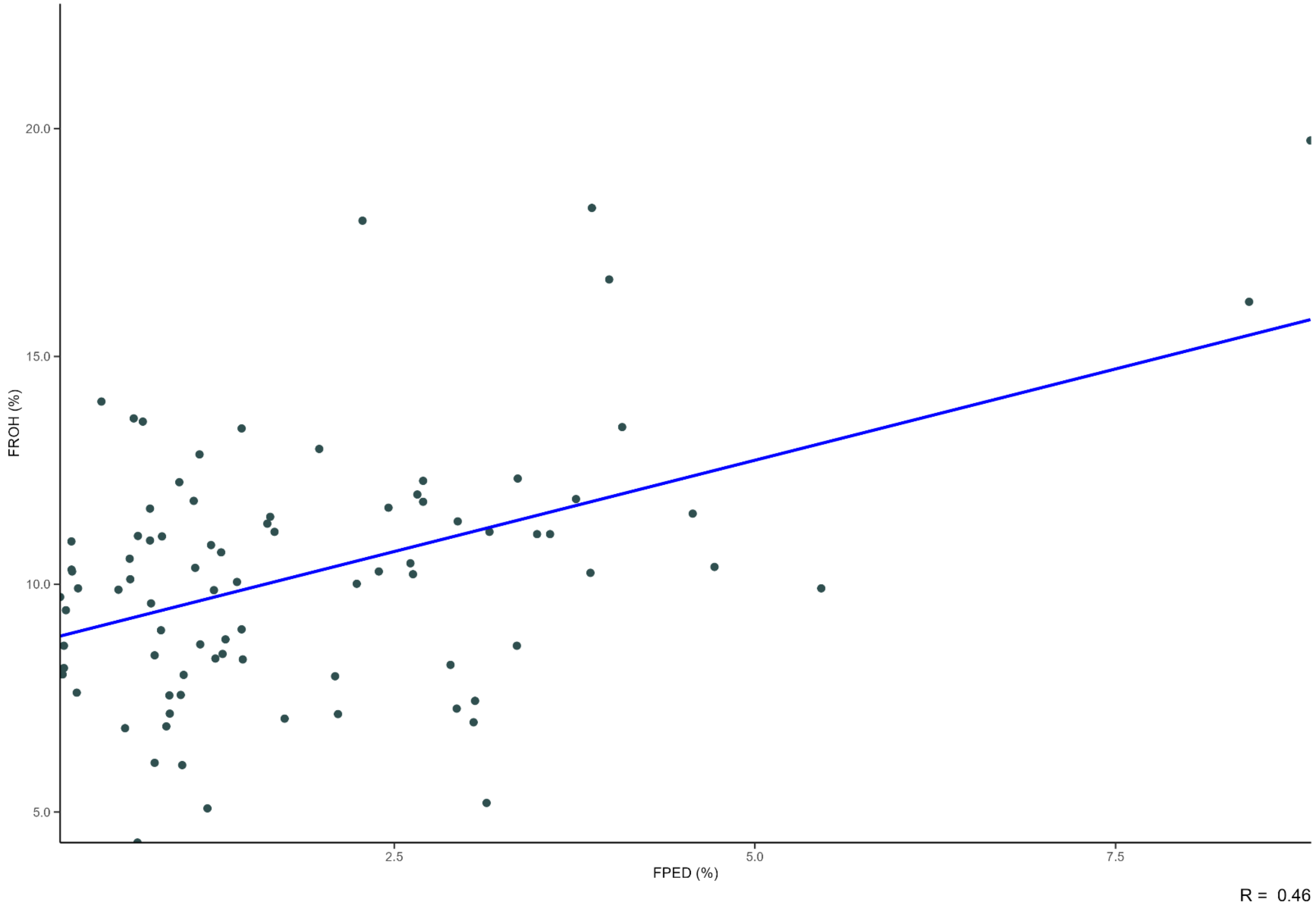
Chromosome 10 – 6 SNPs

Hapmap50760-BTA-69265
BTB-00425665 – ZBTB25
ARS-BFGL-NGS-79865
BTB-00428991
BTB-01890015
BovineHD1000017098
BTB-01524979

Chromosome 20 – 11 SNPs

ARS-BFGL-NGS-107265
Hapmap33504-BTA-154919
ARS-BFGL-NGS-17586
Hapmap39753-BTA-50189
BTB-01424374
Hapmap50991-BTA-50645
ARS-BFGL-NGS-37203
Hapmap35236-BES3_Contig374_812
ARS-BFGL-BAC-32359
BovineHD2000013524
ARS-BFGL-NGS-104898

Correlation between F_{PED} and F_{ROH}



Conclusions

- The estimated level of genomic inbreeding in this population calls for more attention in designing of mating schemes to minimize inbreeding rates
- More samples should be used to get more information on this issue (meanwhile 386 available)
- Other breeds for comparison
  - Simmental

Future perspective

- Larger number of novel factors will come into play during sire or mate selection
  - Inclusion of dominance effects
  - Exploitation of variation in recombination across the genome
  - Increase in use of advanced reproductive technologies
  - Future introduction of gene-editing procedures

Thank You For Listening

https://www.icbf.com/?p=5831