Genomic runs and resampled predictive models as alternative/complement to GWAS

Filippo Biscarini

(Biostatistician, bioinformatician, quantitative geneticist)

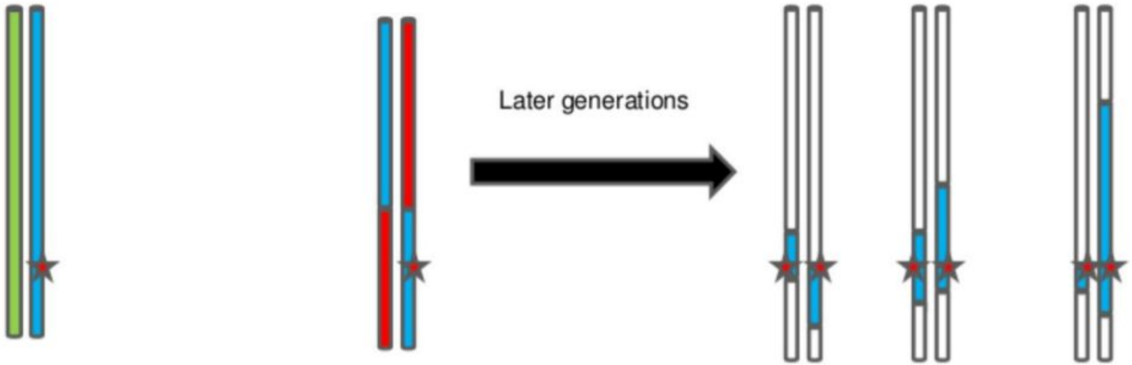
CNR-IBBA, Milan (Italy)

Genomic runs

genomic runs - ROH

Runs of homozygosity (ROH)

- **stretches** of **homozygous SNP** genotypes
- autozygosity (IBD vs IBS)



genomic runs - ROH

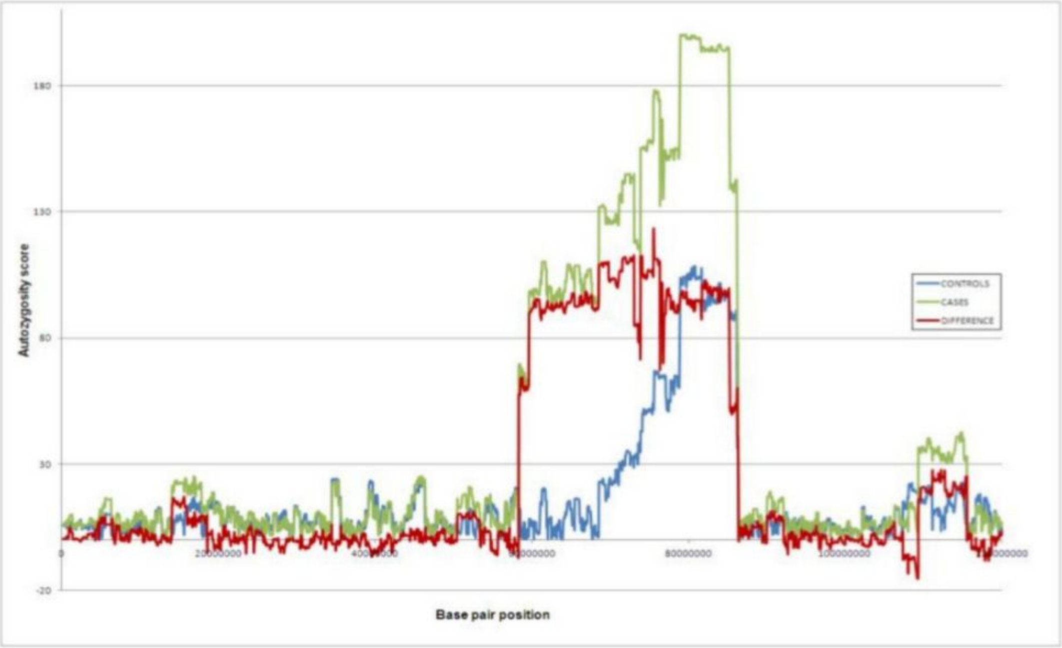
ROH-based approaches for association studies

FECUND

- **ROH** may be used to **localize mutations**
- Mainly, **recessive mutations** (not lethal)
- Homozygosity (**ROH**) **patterns** are:
  - different in **cases vs controls** (> homozygosity in cases)
  - different **around the mutation** compared to the rest of the genome (> homozygosity around the mutation)

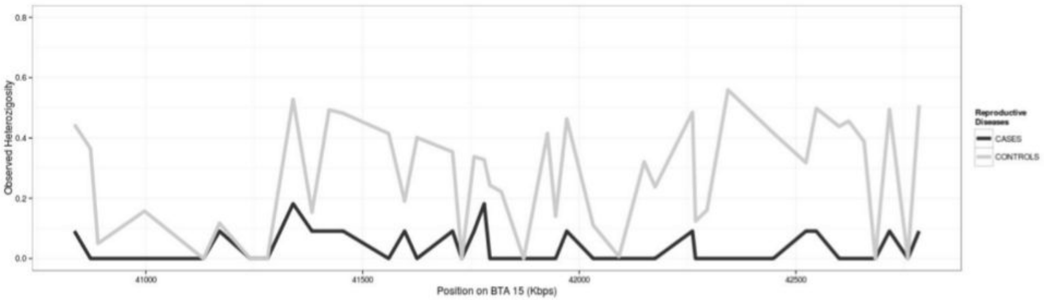
genomic runs - ROH

Perinatal mortality calves – BTA4 (Pollot 2012)

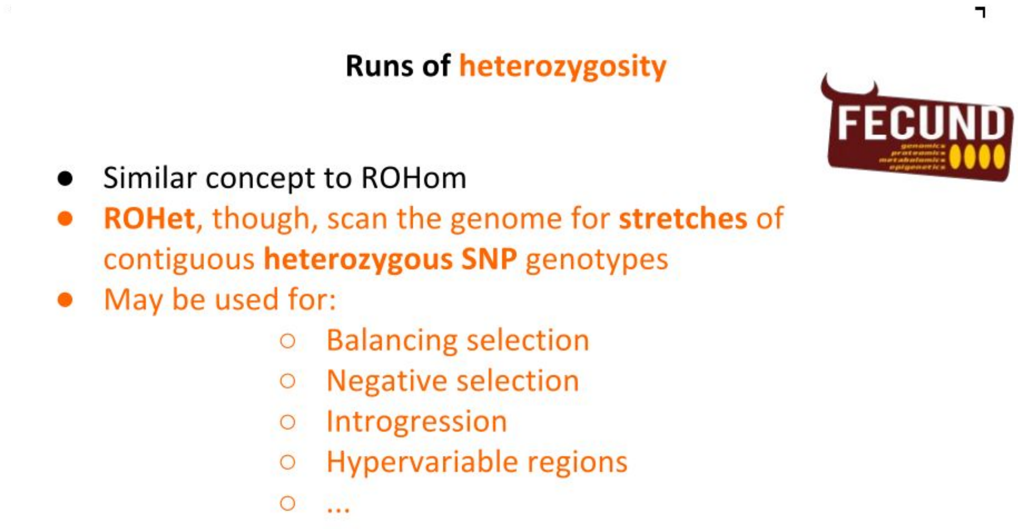


genomic runs - ROH

Reproductive diseases Holsteins – BTA15 (Biscarini et al. 2015)



genomic runs - “ROHet”



Runs of heterozygosity

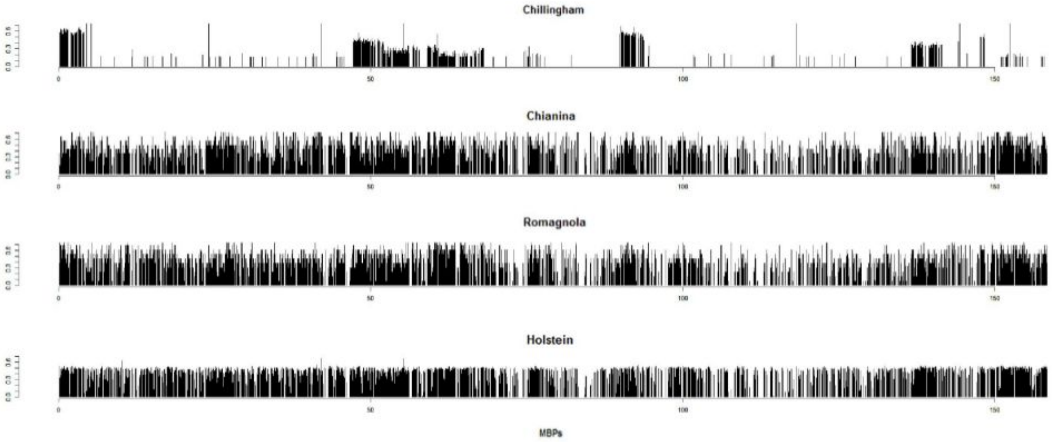
FECUND

- Similar concept to ROHom
- **ROHet**, though, scan the genome for **stretches** of contiguous **heterozygous SNP** genotypes
- May be used for:
    - ○ Balancing selection
    - ○ Negative selection
    - ○ Introgression
    - ○ Hypervariable regions
    - ○ ...

genomic runs - “ROHet”

ROHet in Chillingham cattle (Williams et al 2015)

FECUND
genomics
proteomics
metabolomics
epigenetics



genomic runs - “ROHet”

ROHet in Chillingham and deleterious haplotypes

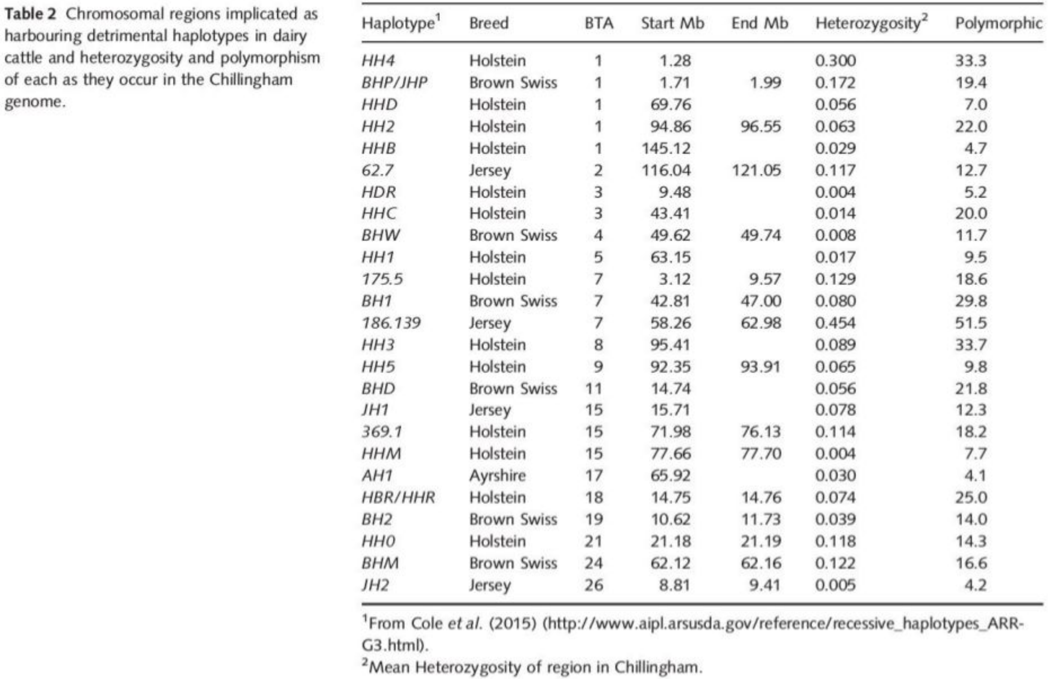


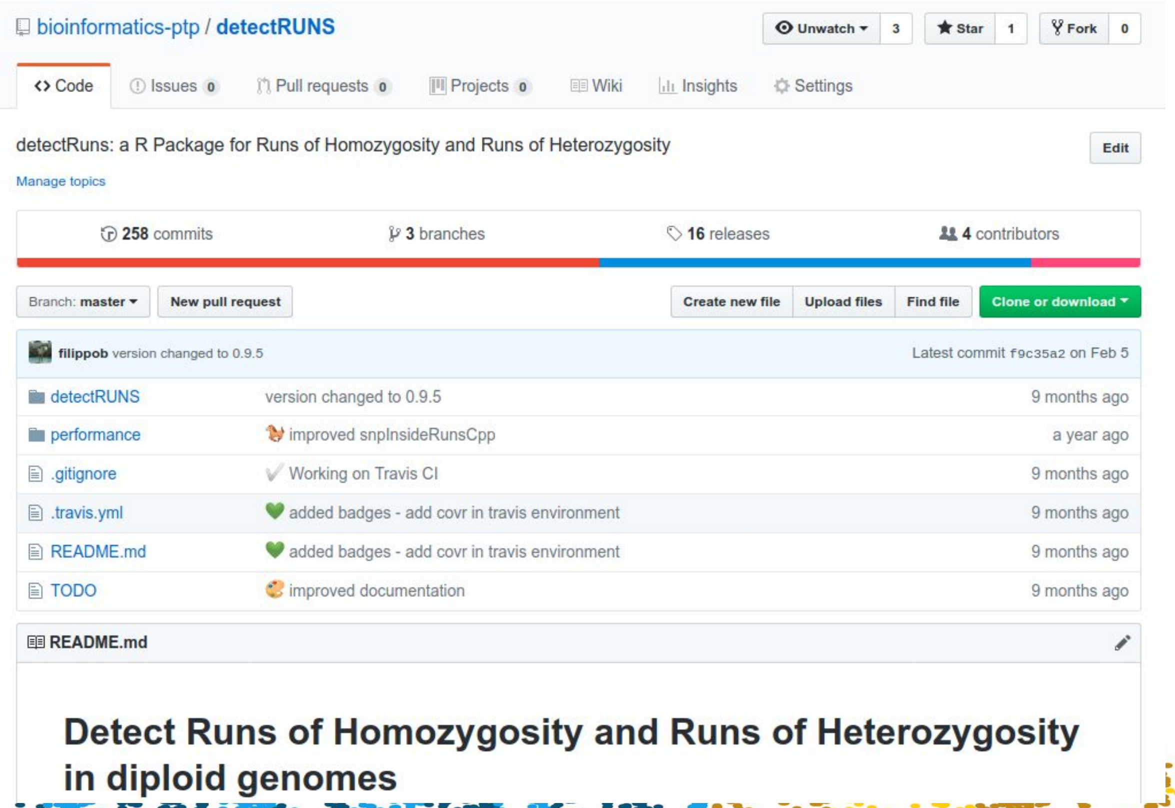
Table 2 Chromosomal regions implicated as harbouring detrimental haplotypes in dairy cattle and heterozygosity and polymorphism of each as they occur in the Chillingham genome.

Haplotype[1]	Breed	BTA	Start Mb	End Mb	Heterozygosity[2]	Polymorphic
HH4	Holstein	1	1.28		0.300	33.3
BHP/JHP	Brown Swiss	1	1.71	1.99	0.172	19.4
HHD	Holstein	1	69.76		0.056	7.0
HH2	Holstein	1	94.86	96.55	0.063	22.0
HHB	Holstein	1	145.12		0.029	4.7
62.7	Jersey	2	116.04	121.05	0.117	12.7
HDR	Holstein	3	9.48		0.004	5.2
HHC	Holstein	3	43.41		0.014	20.0
BHW	Brown Swiss	4	49.62	49.74	0.008	11.7
HH1	Holstein	5	63.15		0.017	9.5
175.5	Holstein	7	3.12	9.57	0.129	18.6
BH1	Brown Swiss	7	42.81	47.00	0.080	29.8
186.139	Jersey	7	58.26	62.98	0.454	51.5
HH3	Holstein	8	95.41		0.089	33.7
HH5	Holstein	9	92.35	93.91	0.065	9.8
BHD	Brown Swiss	11	14.74		0.056	21.8
JH1	Jersey	15	15.71		0.078	12.3
369.1	Holstein	15	71.98	76.13	0.114	18.2
HHM	Holstein	15	77.66	77.70	0.004	7.7
AH1	Ayrshire	17	65.92		0.030	4.1
HBR/HHR	Holstein	18	14.75	14.76	0.074	25.0
BH2	Brown Swiss	19	10.62	11.73	0.039	14.0
HH0	Holstein	21	21.18	21.19	0.118	14.3
BHM	Brown Swiss	24	62.12	62.16	0.122	16.6
JH2	Jersey	26	8.81	9.41	0.005	4.2

[1] From Cole *et al.* (2015) (http://www.aipl.arsusda.gov/reference/recessive_haplotypes_ARR-G3.html).

[2] Mean Heterozygosity of region in Chillingham.

genomic runs - detectRUNS



detectRuns: a R Package for Runs of Homozygosity and Runs of Heterozygosity

258 commits · 3 branches · 16 releases · 4 contributors

filippob version changed to 0.9.5 — Latest commit f9c35a2 on Feb 5

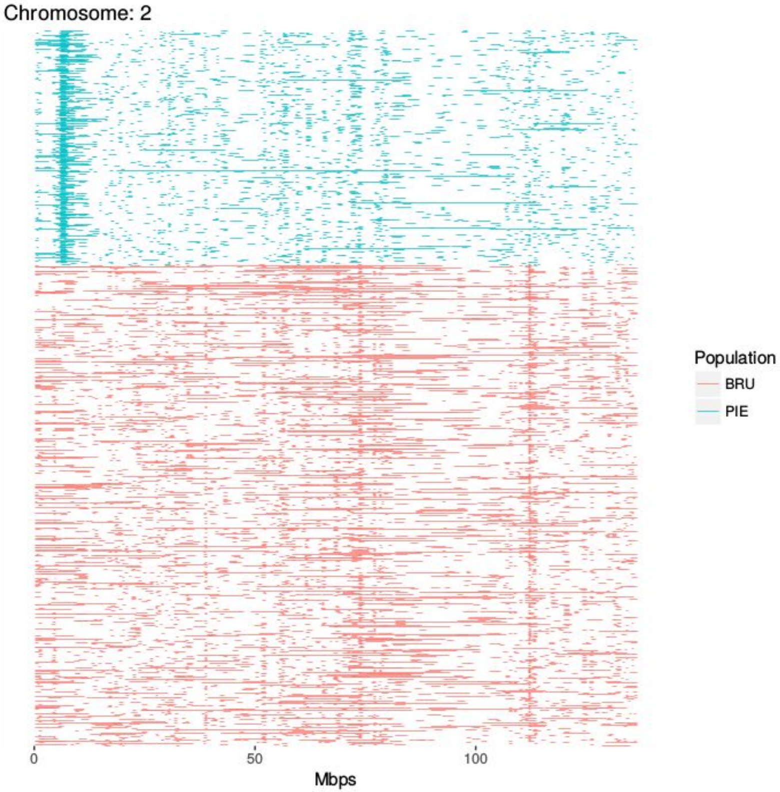
detectRUNS	version changed to 0.9.5	9 months ago
performance	improved snpInsideRunsCpp	a year ago
.gitignore	Working on Travis CI	9 months ago
.travis.yml	added badges - add covr in travis environment	9 months ago
README.md	added badges - add covr in travis environment	9 months ago
TODO	improved documentation	9 months ago

README.md

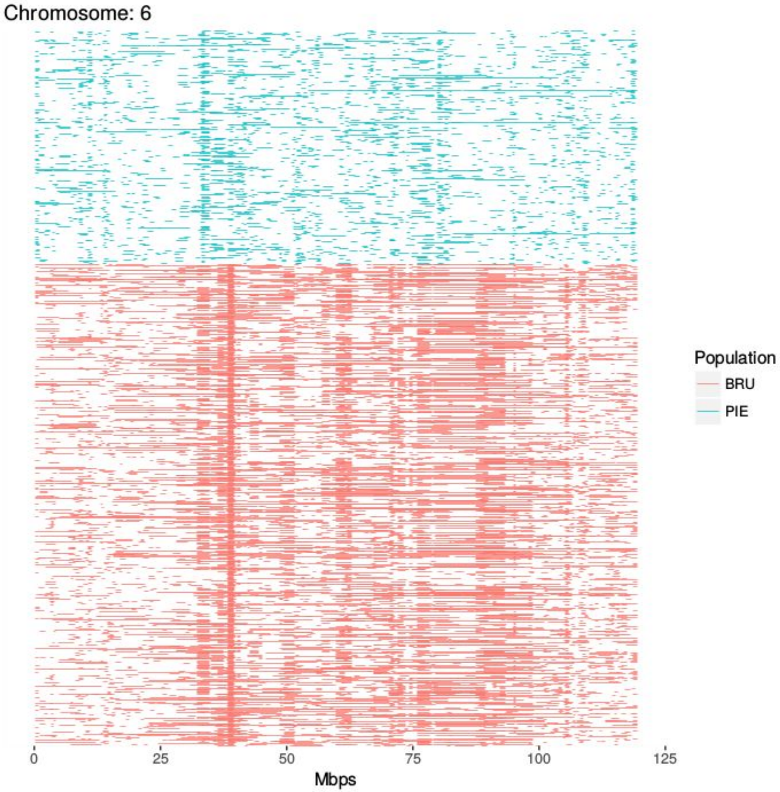
Detect Runs of Homozygosity and Runs of Heterozygosity in diploid genomes

- on CRAN
- two methods
    - sliding windows
    - window-less
- output files (runs per individual, per chromosome)
- pots
    - plot runs
    - plot stacked runs
    - plot n. of times SNP are in runs

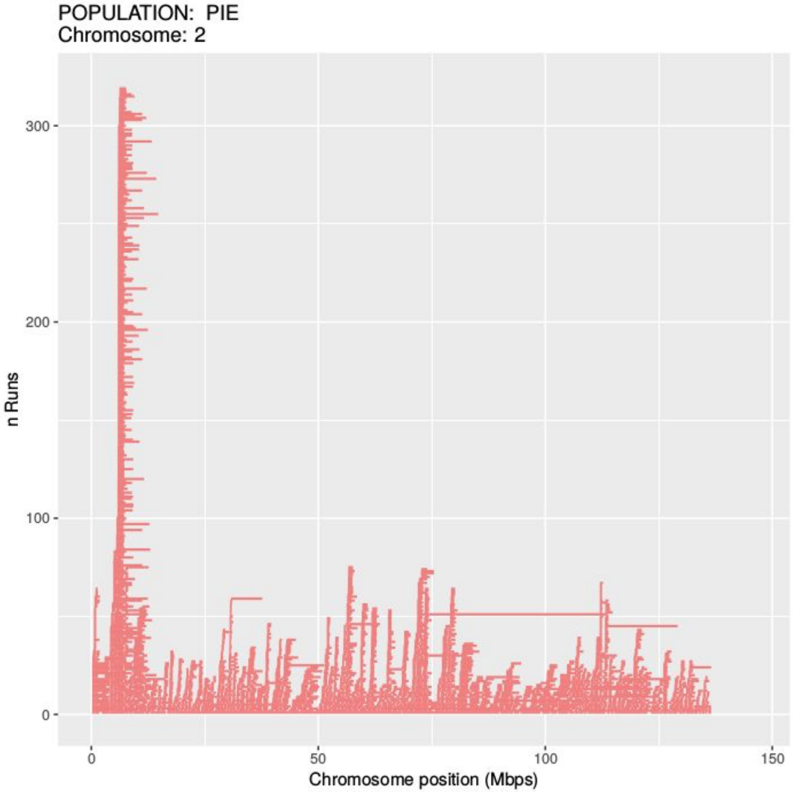
genomic runs - detectRUNS



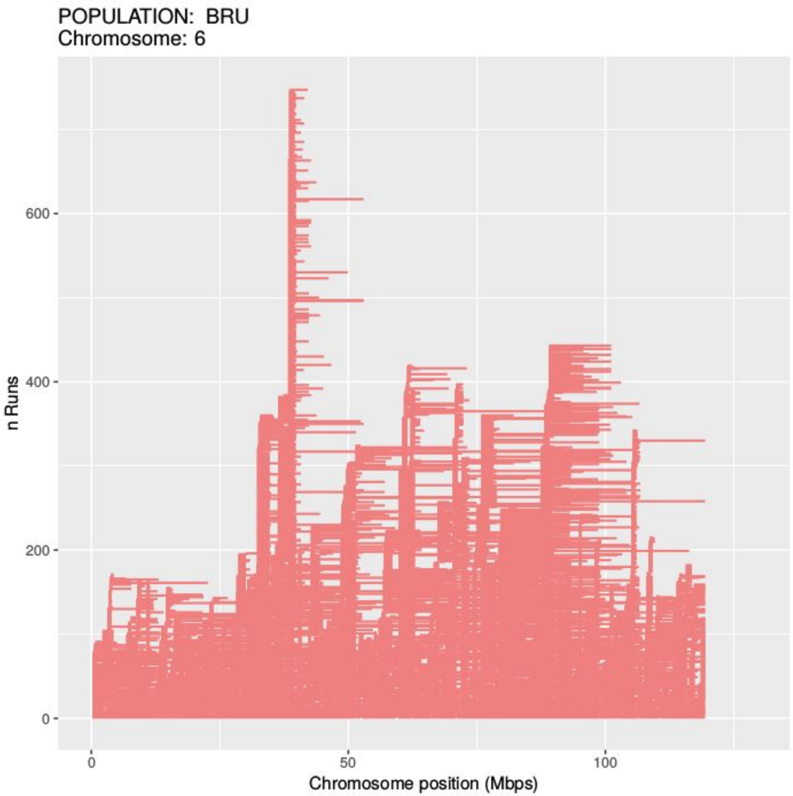
genomic runs - detectRUNS



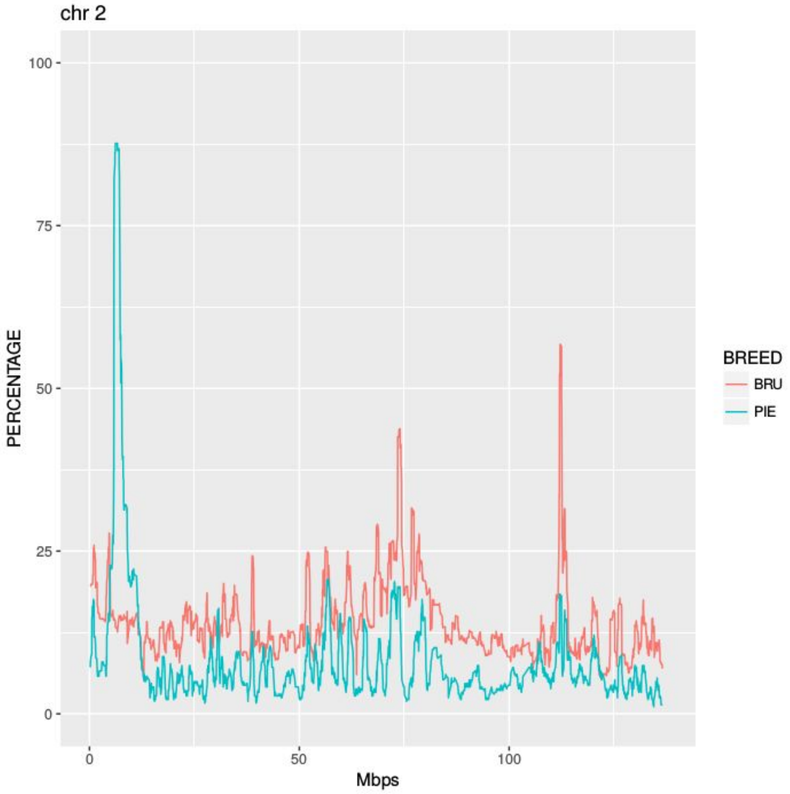
genomic runs - detectRUNS



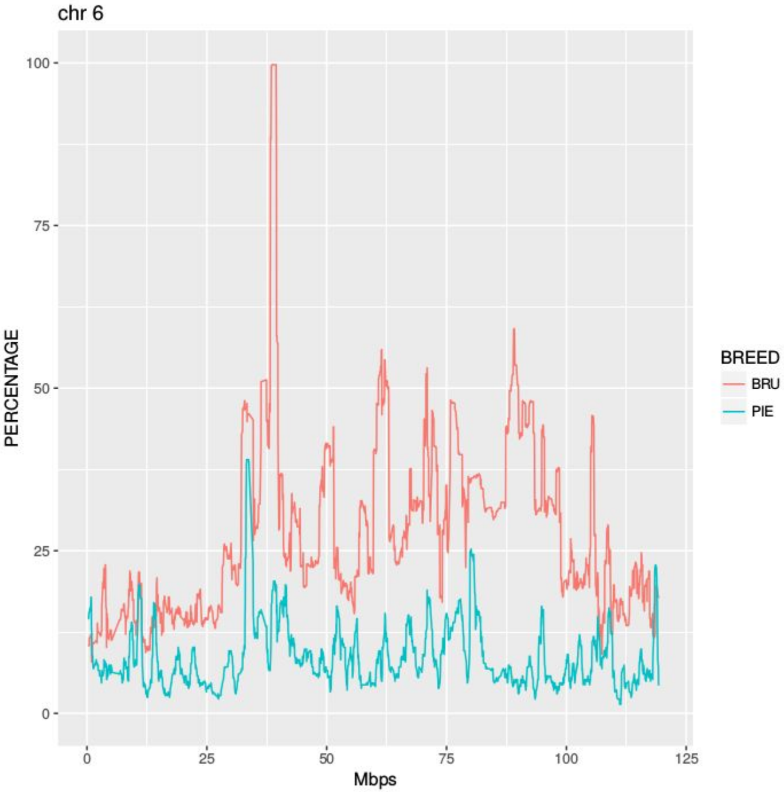
genomic runs - detectRUNS



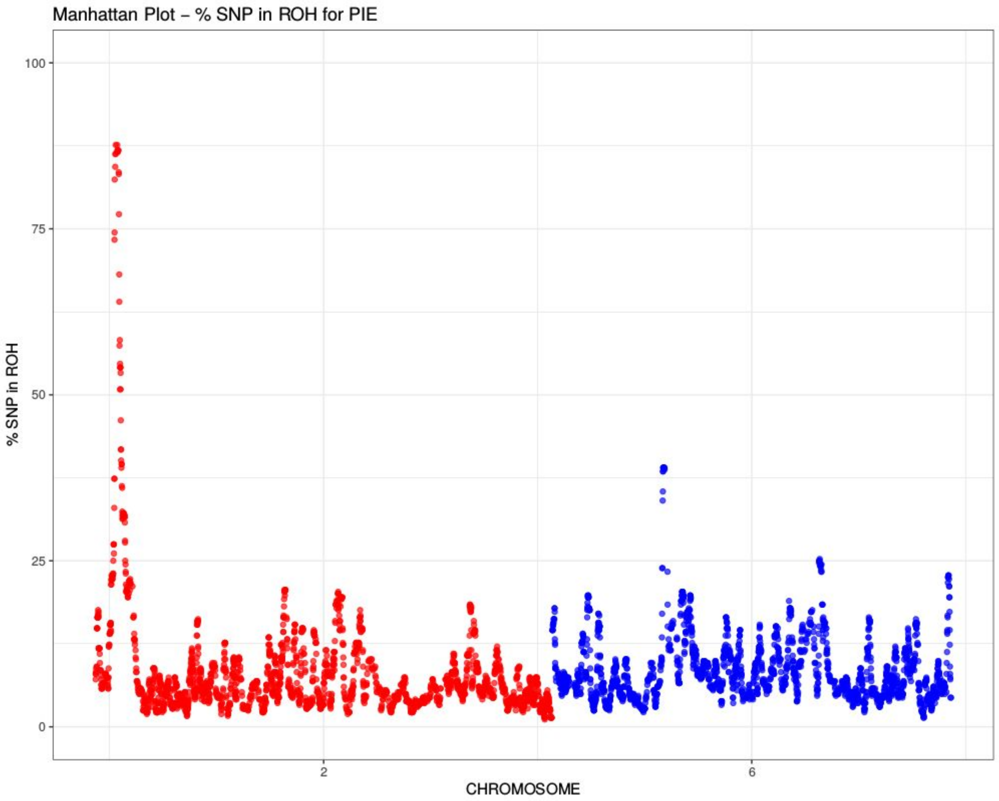
genomic runs - detectRUNS



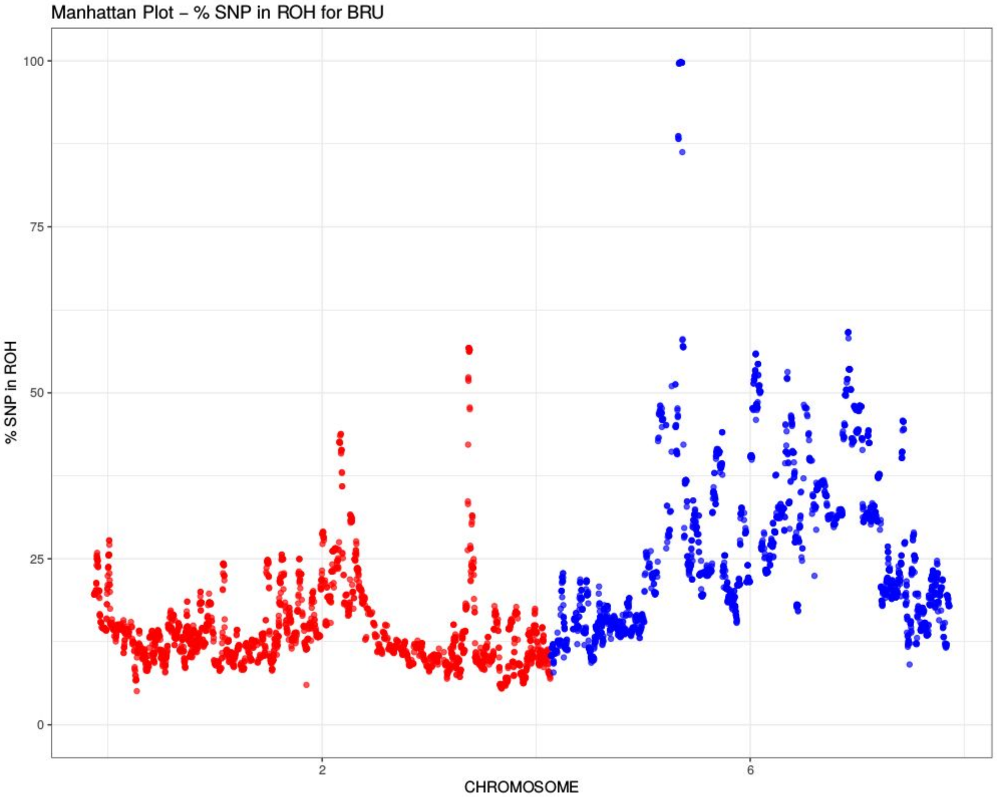
genomic runs - detectRUNS



genomic runs - detectRUNS



genomic runs - detectRUNS



Resampled predictive models

resampled predictive models

- models for genomic predictions typically run hundreds of times
    - □a. **cross-validation** schemes
    - □b. resampling the training set
- from **resampled replicates** of the predictive model, most relevant predictors can be extracted
- some methods operate directly such variable selection (e.g. **Lasso-penalized models**)

resampled predictive models

- **each replicate** of the predictive model (e.g. 1000) → different subset of relevant SNPs
- 1000 subsets of relevant SNPs
- **SNP frequency** (in the 1000 subsets of predictors) against **SNP position** along the genome → detecting associated variants!
- the **accuracy** of prediction is relevant!

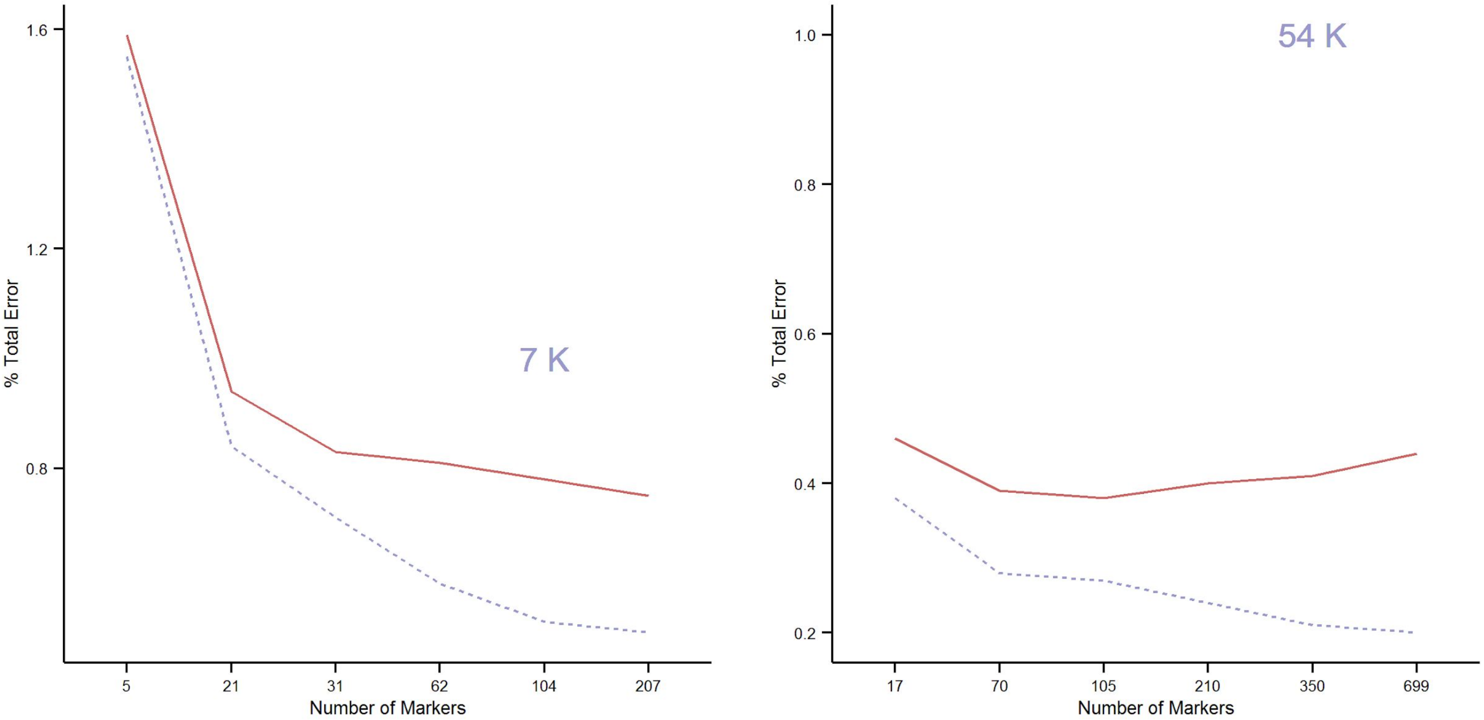
prediction of haplotype carriers

BH2 haplotype on **BTA19** associated with stillbirth and reduced cow fertility in **Brown Swiss** cattle

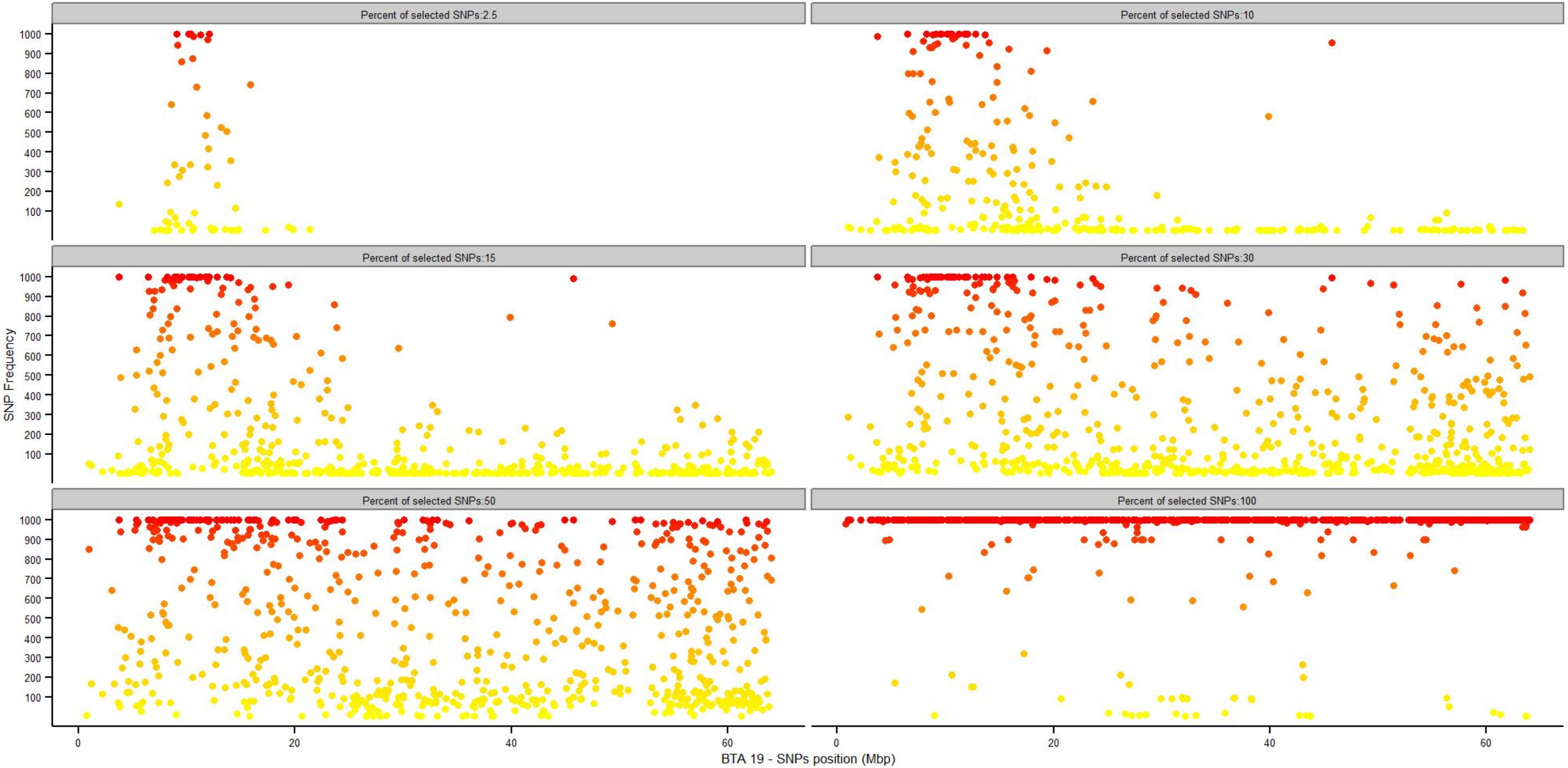
- **3645 Brown Swiss** bulls & cows (513 carriers), **1512 SNPs** on BTA19 (from 50K SNP chip), **211 SNP** (from 7K SNP chip)
- LDA: **linear discriminant analysis**
- **Decreasing subsets** of SNPs
- 10-fold **cross-validation** x 100 times

[Biffani et al. 2015 *Genet Sel Evol*]

prediction of haplotype carriers



resampling-based mutation mapping



TUBD1 mutation in Brown Swiss cattle

- **BH2**: Braunvieh haplotype 2, on **BTA19** associated with high peri- and postnatal calf losses in Braunvieh cattle
- missense mutation in the gene ***TUBD1*** (recessive)
- homozygous calves are underweight at birth and suffer from recurrent respiratory diseases
- ultrastructural abnormalities of cilia in the respiratory tract

prediction of mutation carriers

***TUBD1* mutation** behind the BH2 haplotype on BTA19

Missense mutation → **microtubular defects** in **airway** cilia

→ chronic respiratory disease

→ stillbirth, perinatal mortality in calves

[Schwarzenbacher et al. 2016 *BMC Genomics*]

Predict directly mutation carriers rather than haplotype carriers!

Haplotype-mutation concordance is 99% though

Not so in other cases (e.g. CVM)

prediction of mutation carriers

392 Brown Swiss and **3116 Fleckvieh** cows and bulls, **1512 SNPs** on BTA19 (from 50K SNP chip)

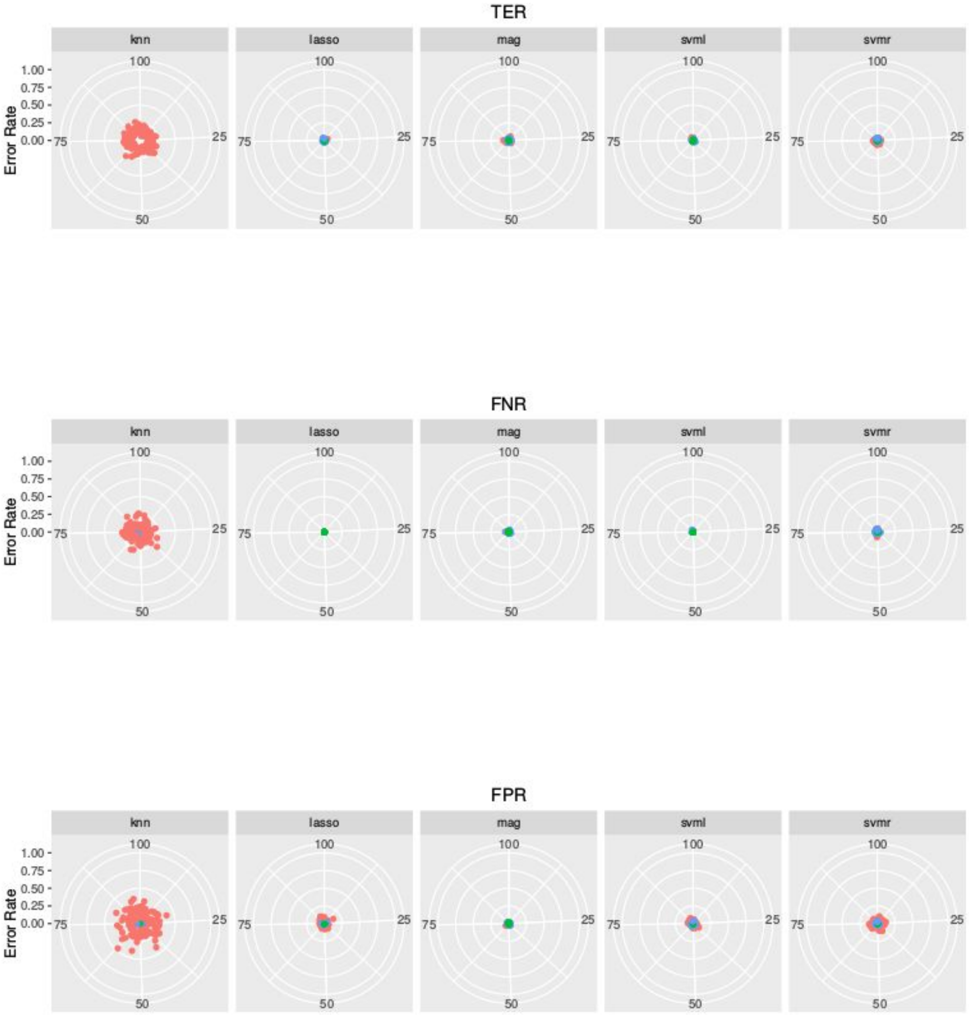
Five **classification** methods:

- **KNN**
- **Lasso-penalized LR**
- **SVM-Linear** kernel
- **SVM-Radial** kernel
- **MAG** (Multi Allelic Gene prediction) [haplotype-based method]

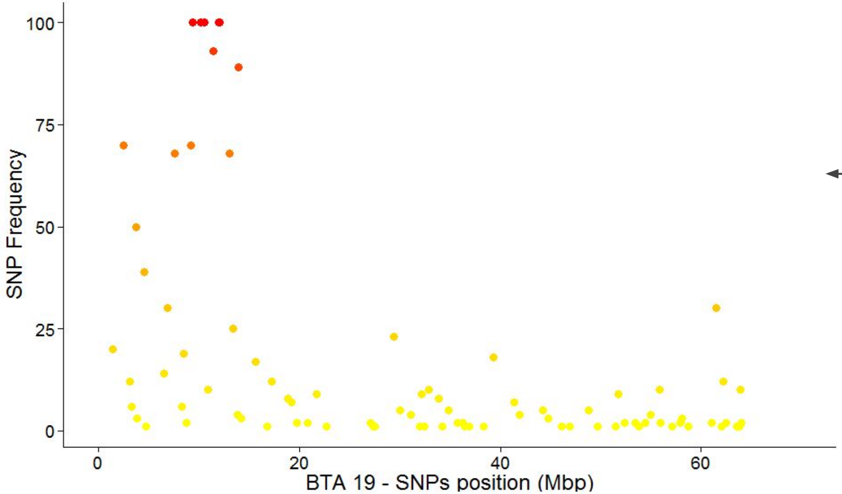
10-fold **cross-validation** x 100 times

[Biscarini et al. 2016 *BMC Genomics*]

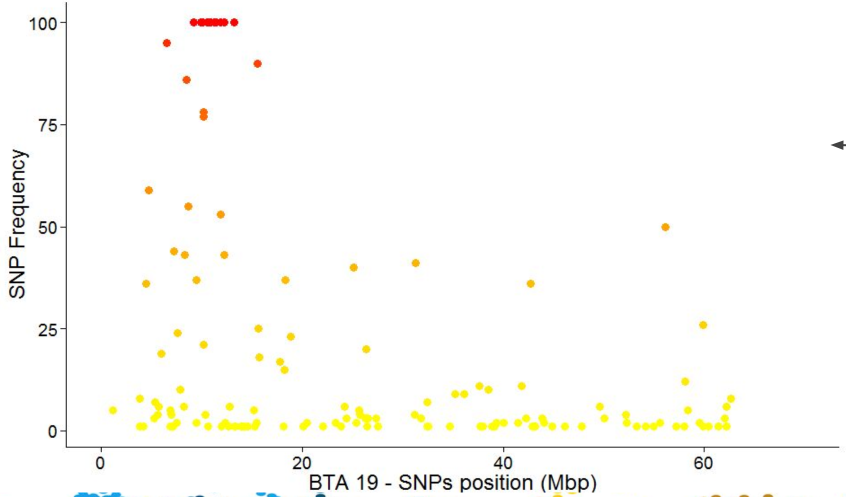
prediction of mutation carriers: accuracy



prediction of mutation carriers: resampled models

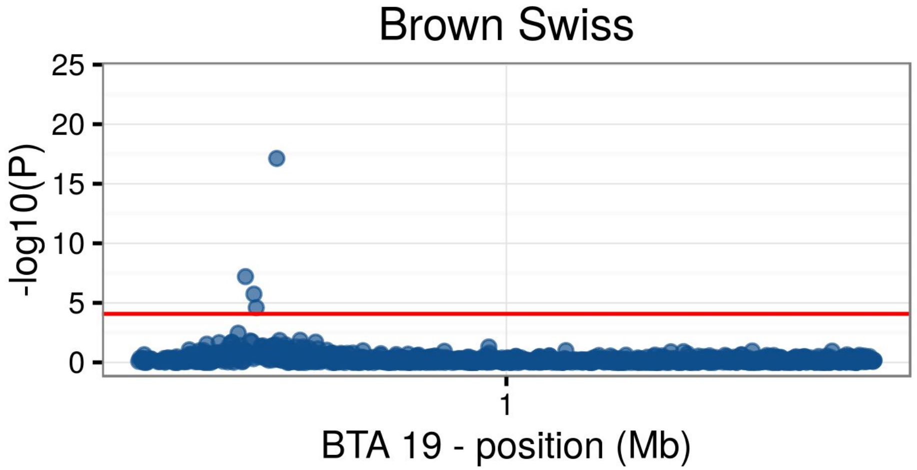


Brown Swiss

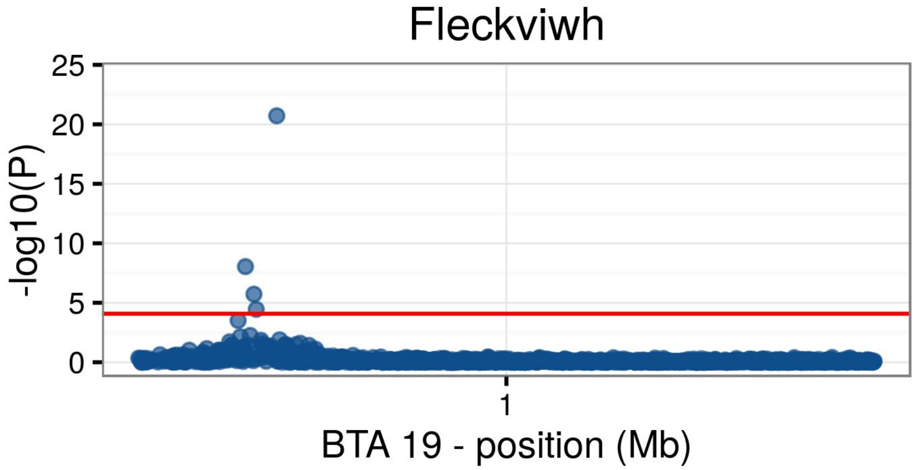


Fleckvieh

TUBD1 mutation: GWAS



Brown Swiss



Fleckviwh

Physalia Courses

NEXT COURSE

Deep learning for biologists