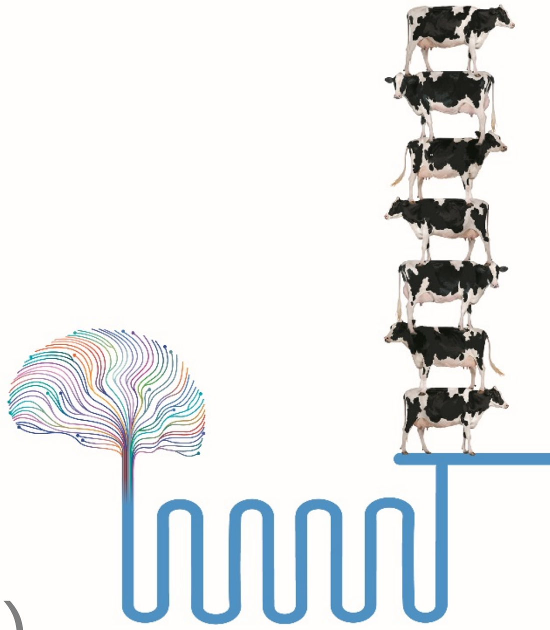
Selecting SNP for Genomic prediction

Ric Sherlock,
M.Keehan, K.Tiplady,
T.Johnson, C.Couldrey,
M.Littlejohn, S.Davis,
B.Harris

Genomic prediction at LIC

- Currently
  - 35K SNP from Illumina Bovine 50k
  - Hybrid Single Step method
- Developing
  - Sequence-derived, trait-specific SNP sets
  - Single Step Marker Model (SSMM)



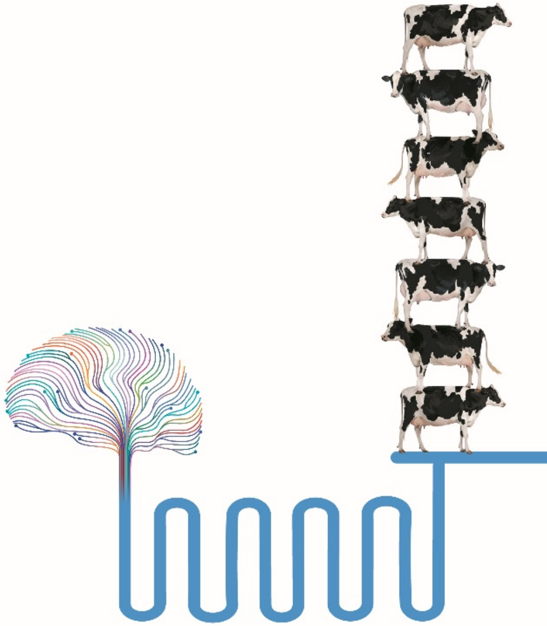
Phenotypes

- Breeding Worth (BW) Traits
  - Fat, Protein, Milk volume, Live weight, Fertility

- Type Traits
  - Udder support

Phenotypes

- Produced phenotypic records for genotyped males and females
  - – Males
    (non-genotyped progeny)
  - – Females
    (original record + weight)



Training data set

1990 2009 2011

5,863 ♂

59,386 ♀

*Sires born before 2009

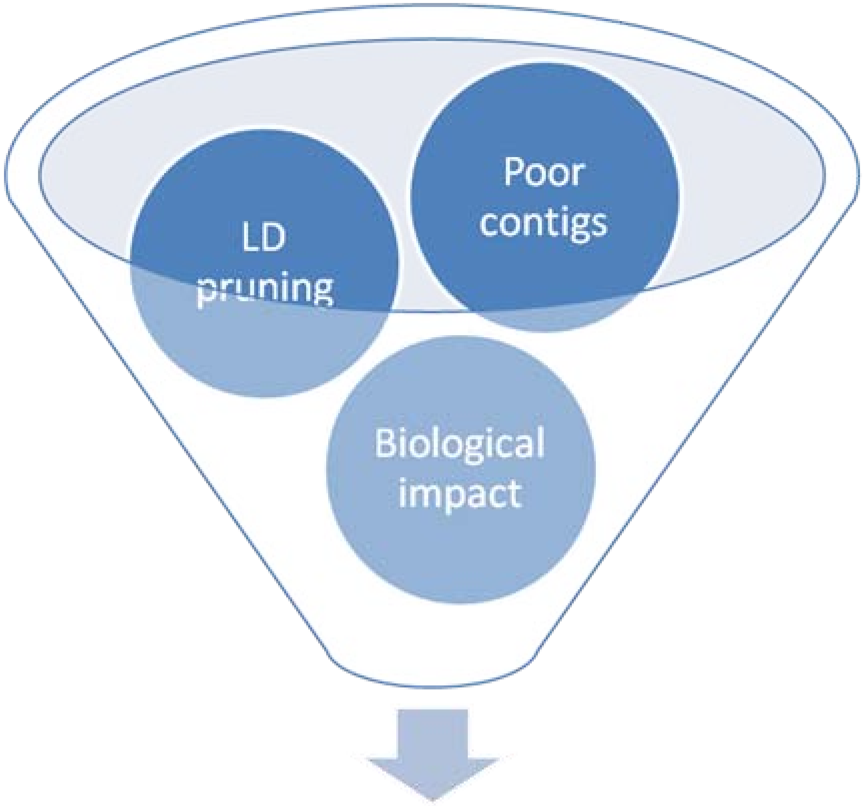
Choosing trait-specific markers

- 19.5 million whole-genome sequence variants

- Identifying marker sets for each trait
  - – Which markers?
  - – How many markers?

Marker pre-selection

19.5 million sequence variants

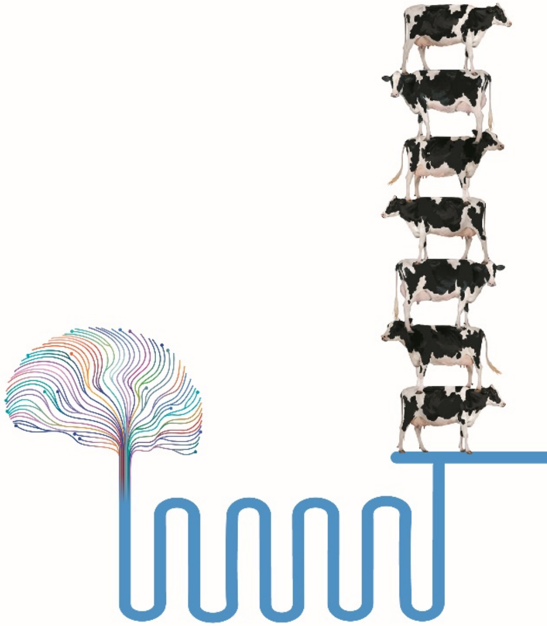


LD pruning

Poor contigs

Biological impact

1.6 million subset



BayesRC

Include prior belief of biological significance of markers

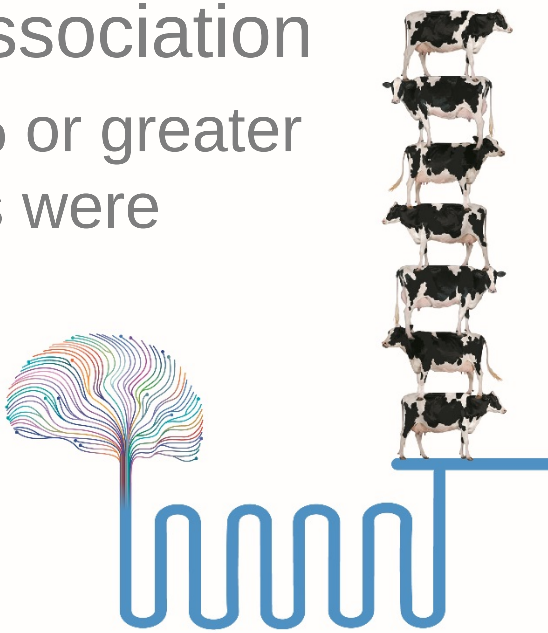
Variant class	# Variants	Definition
1	26820	Considerable published evidence of causation
2	39804	Correlated with differences in RNA expression; Predicted by SNPEff to have functional effects and were located in genes we have seen expressed
3	337603	Not captured in Class 2 and were located in regulatory regions or in genes we hadn't seen expressed
4	1144151	Other variants not yet in a class

GWAS

- Method of Analysis
  - BayesRC (MacLeod et. al, 2016)
  - SNPs had unit variances and mean zero
  - Animal Model — SNP Markers (mixture distribution) and polygenic random effects
  - Fixed effects — overall mean and breed proportions

GWAS

- Two step approach:
  - – Chromosome by chromosome association
    - SNP with a prior probability of 1.5% or greater not being in the zero variance class were selected by trait
  - – GWAS of all selected SNP



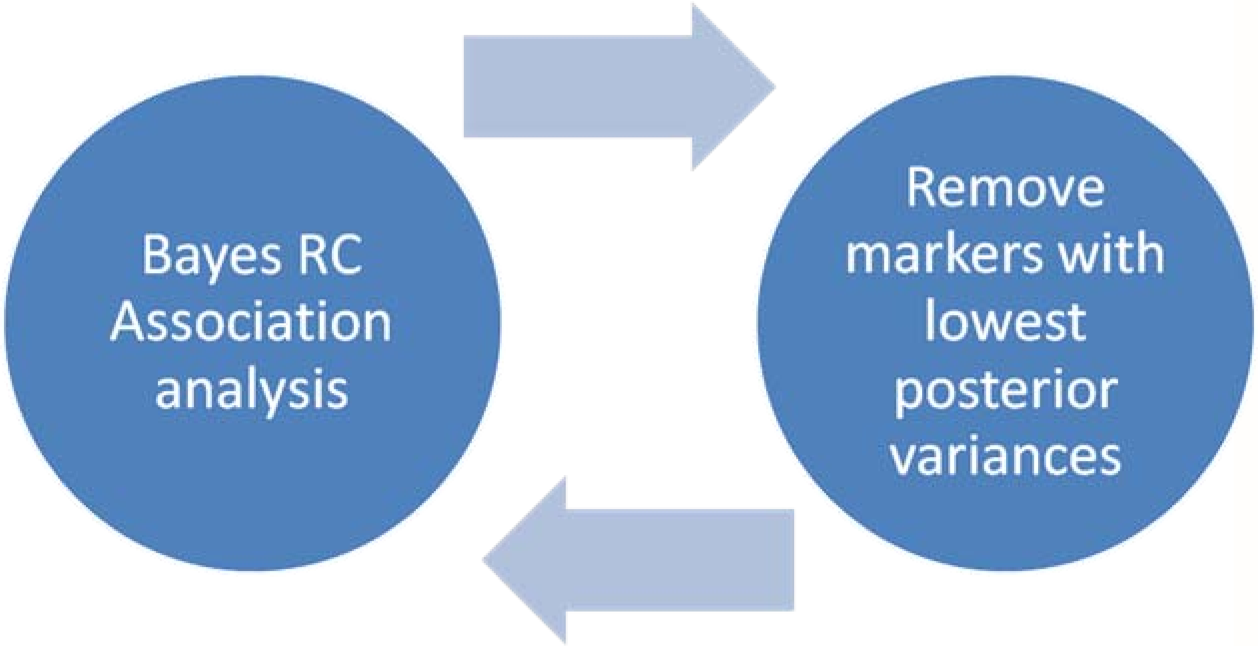
Number of SNP

Iteration	Milk	Fat	Prot	Lwt	Fert	Udder support
0	185051	147100	158619	170166	79363	95429
1	141398	103568	110111	118895	25554	41459

How many markers to use?

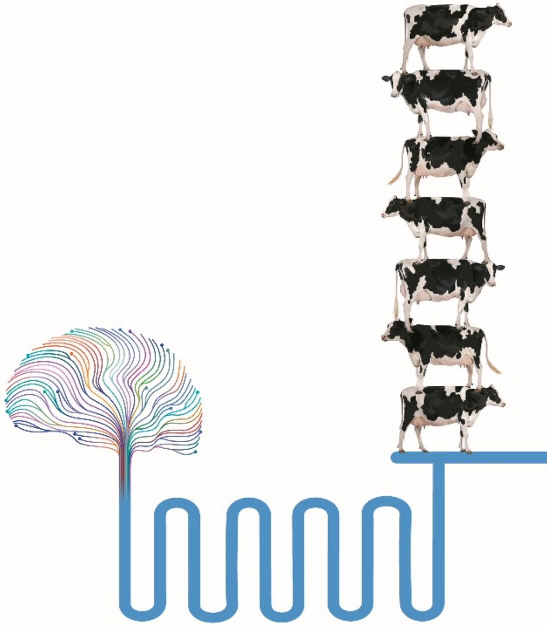
- Can we simply truncate the top N markers?
- What happens to signal captured by truncated markers?

GWAS iterations



Bayes RC Association analysis

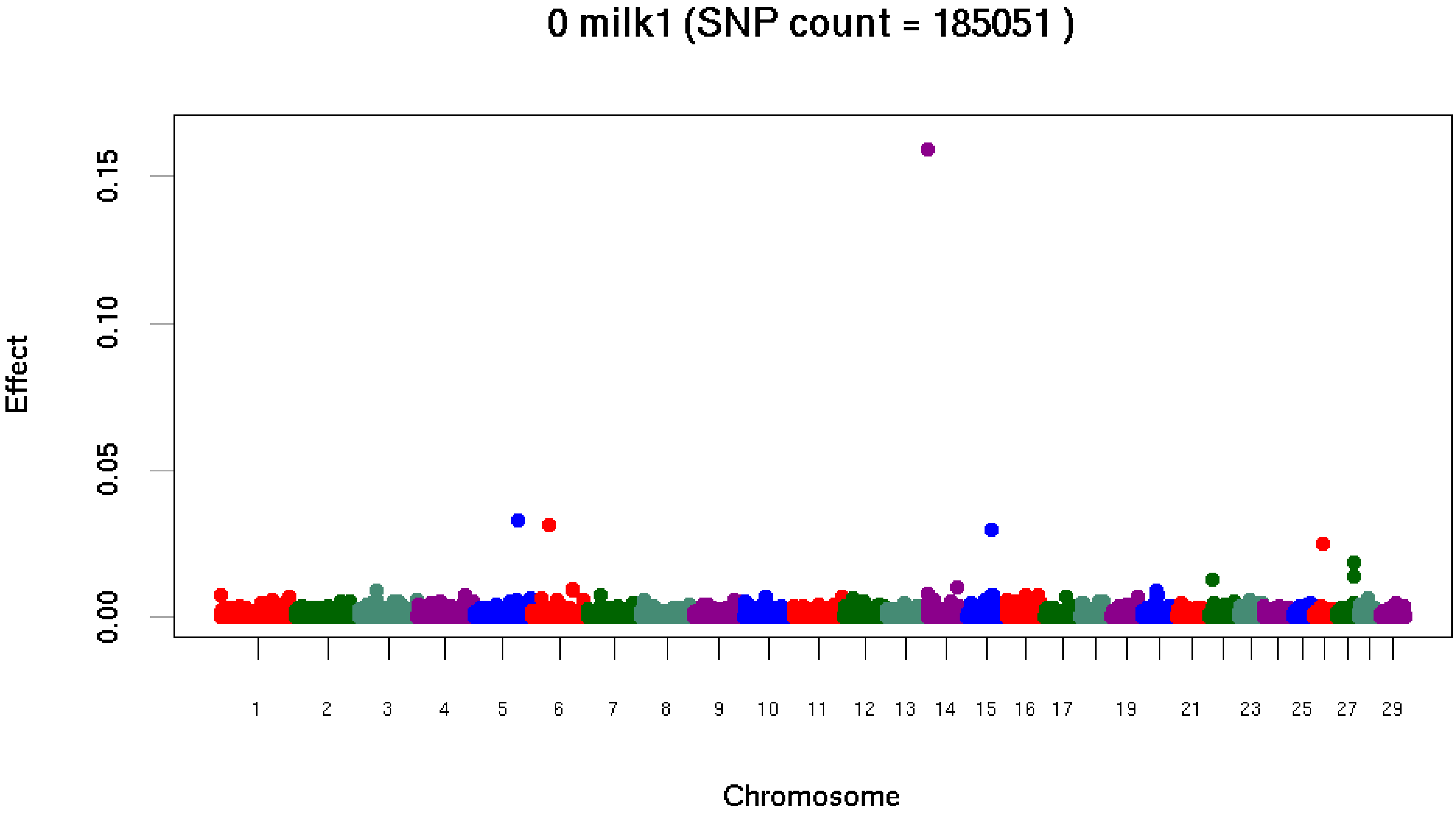
Remove markers with lowest posterior variances



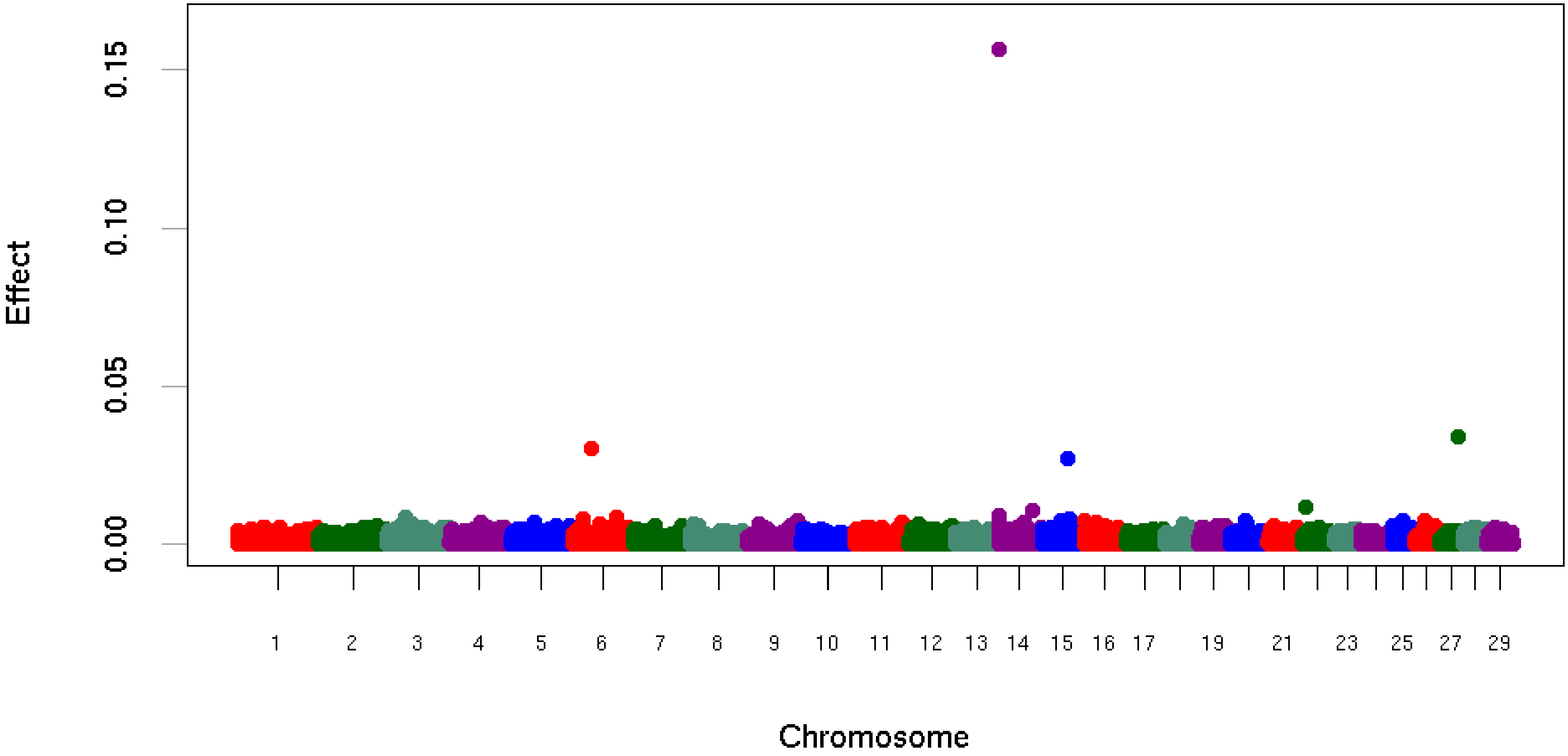
Stop once number of markers is less than 1000

Number of SNP

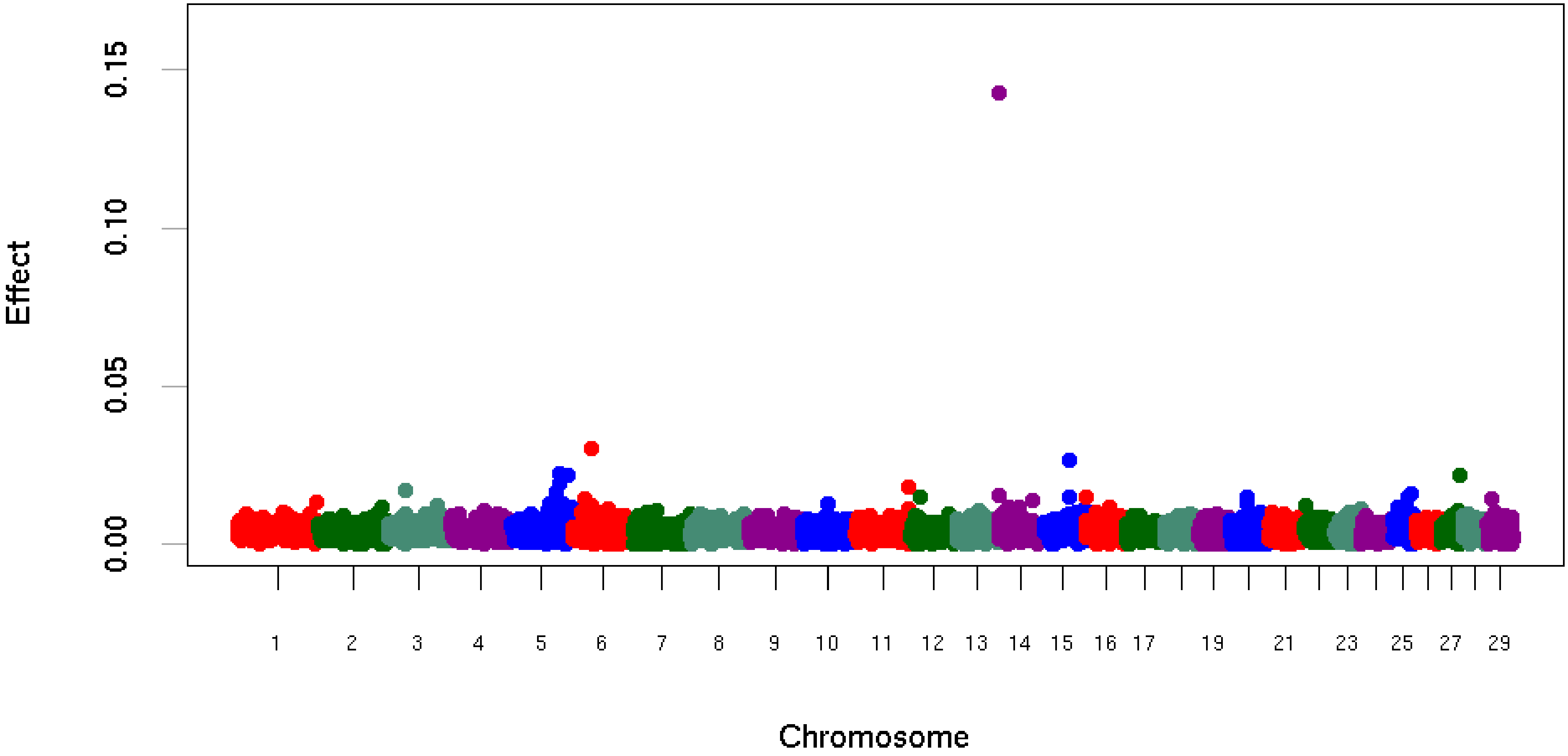
Iteration	Milk	Fat	Prot	Lwt	Fert	Udder support
0	185051	147100	158619	170166	79363	95429
1	141398	103568	110111	118895	25554	41459
2	113119					
3	84839					
4	56560					
5	28280					
6	14140					
7	7070					
8	3535					
9	1767					
10	883					


0 milk1 (SNP count = 185051)
Effect
0.15
0.10
0.05
0.00
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
23
25
27
29
Chromosome

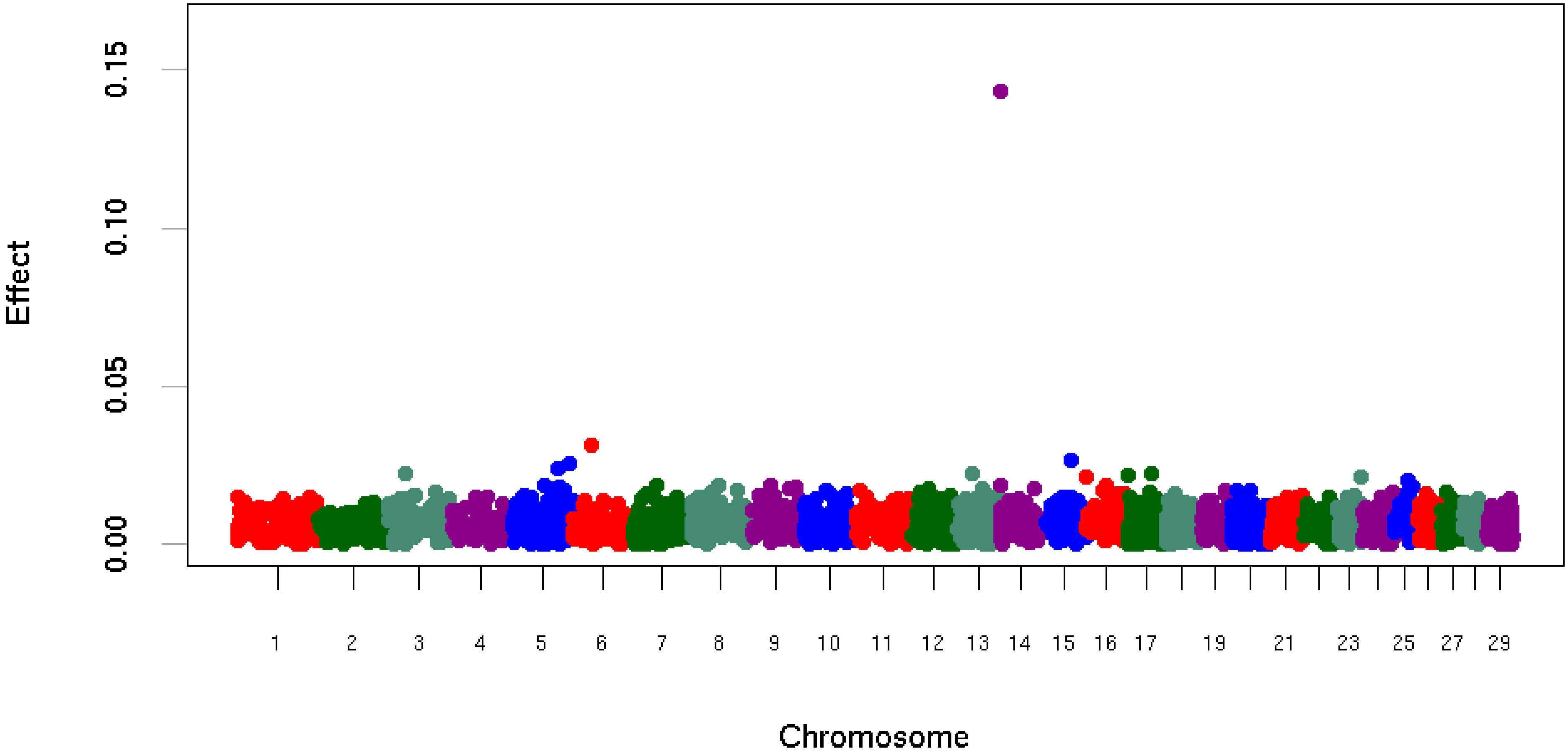
4 milk1 (SNP count = 56560)

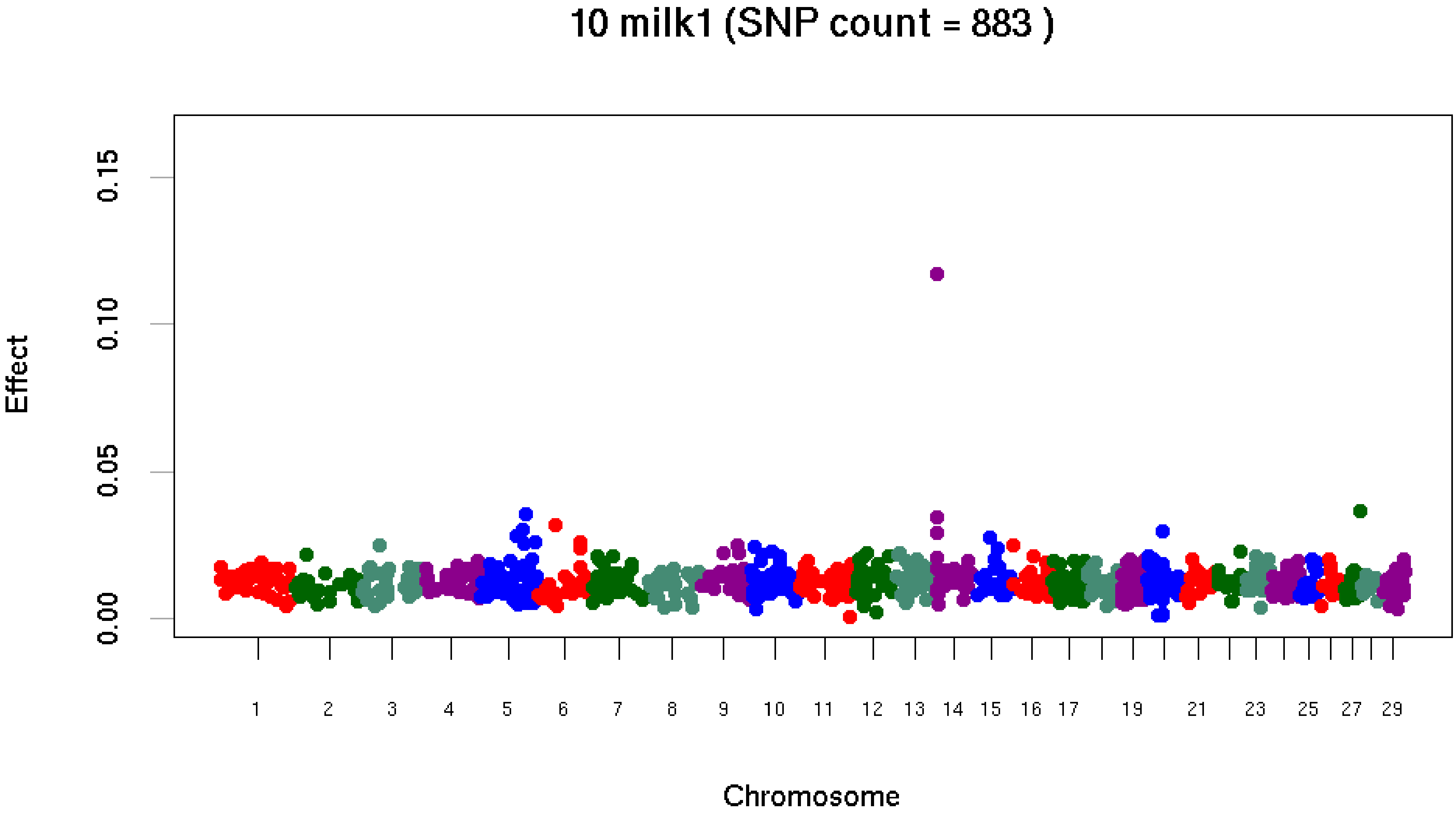


7 milk1 (SNP count = 7070)



8 milk1 (SNP count = 3535)



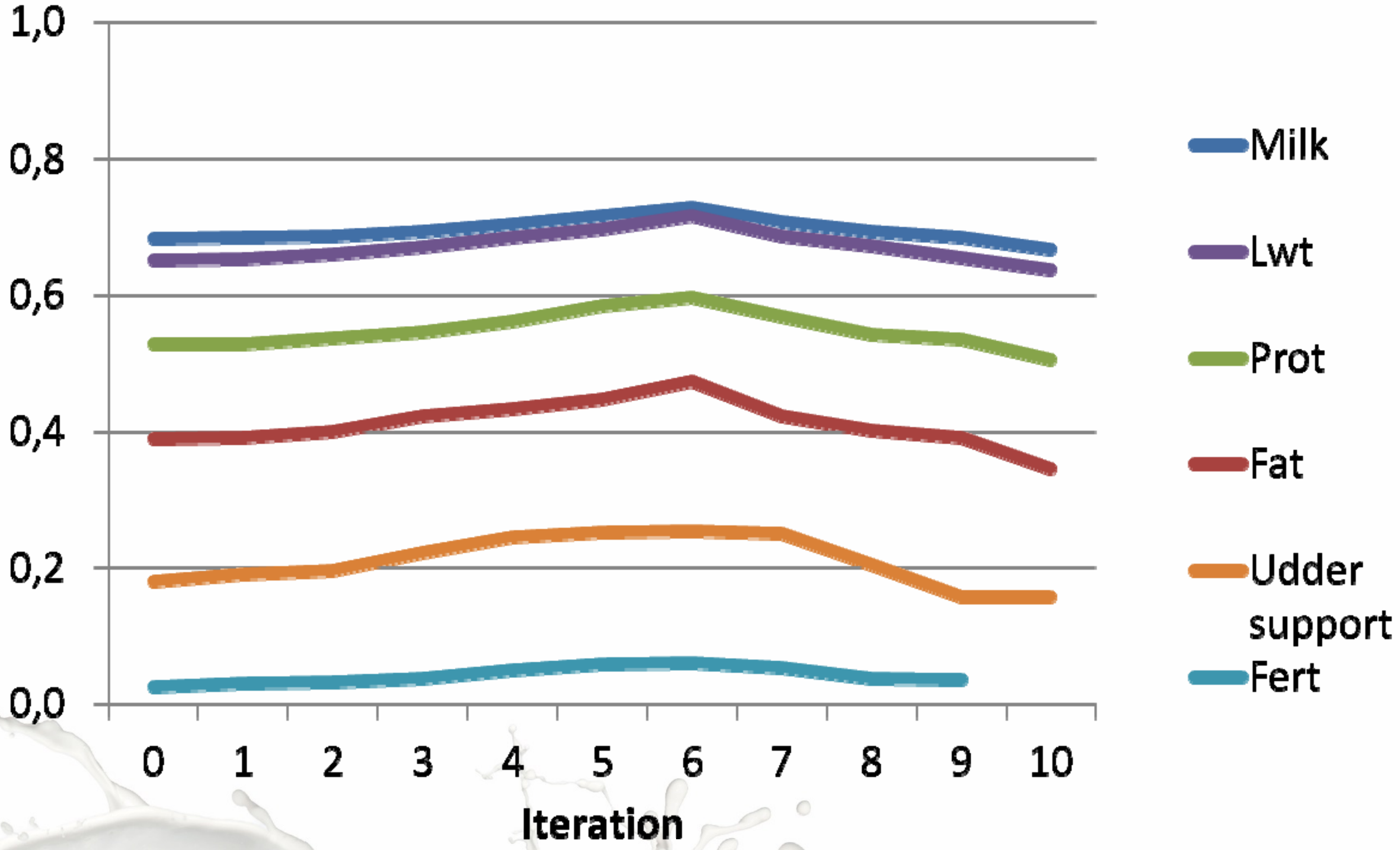

10 milk1 (SNP count = 883)
Effect
0.00
0.05
0.10
0.15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
23
25
27
29
Chromosome

Number of SNP

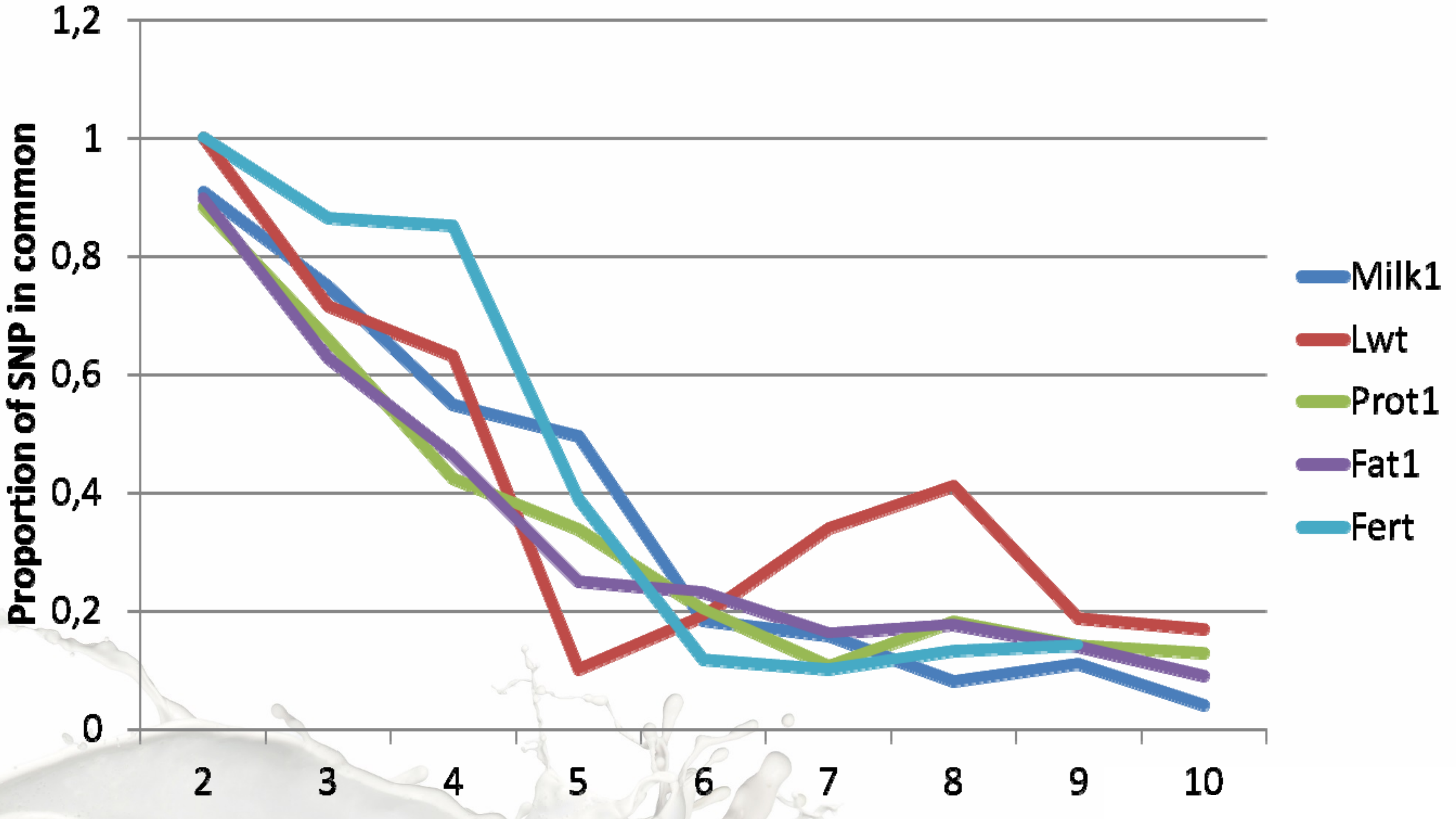
Iteration	Milk	Fat	Prot	Lwt	Fert	Udder support
0	185051	147100	158619	170166	79363	95429
1	141398	103568	110111	118895	25554	41459
2	113119	82855	88089	95116	20444	33168
3	84839	62141	66067	71337	15333	24876
4	56560	41428	44045	47558	10222	16584
5	28280	20714	22023	23779	5111	8292
6	14140	10357	11011	11889	2555	4146
7	7070	5178	5505	5944	1277	2073
8	3535	2589	2752	2972	638	1036
9	1767	1294	1376	1486	500	518
10	883	647	688	743		500

Which marker set is best?

Estimated Heritability

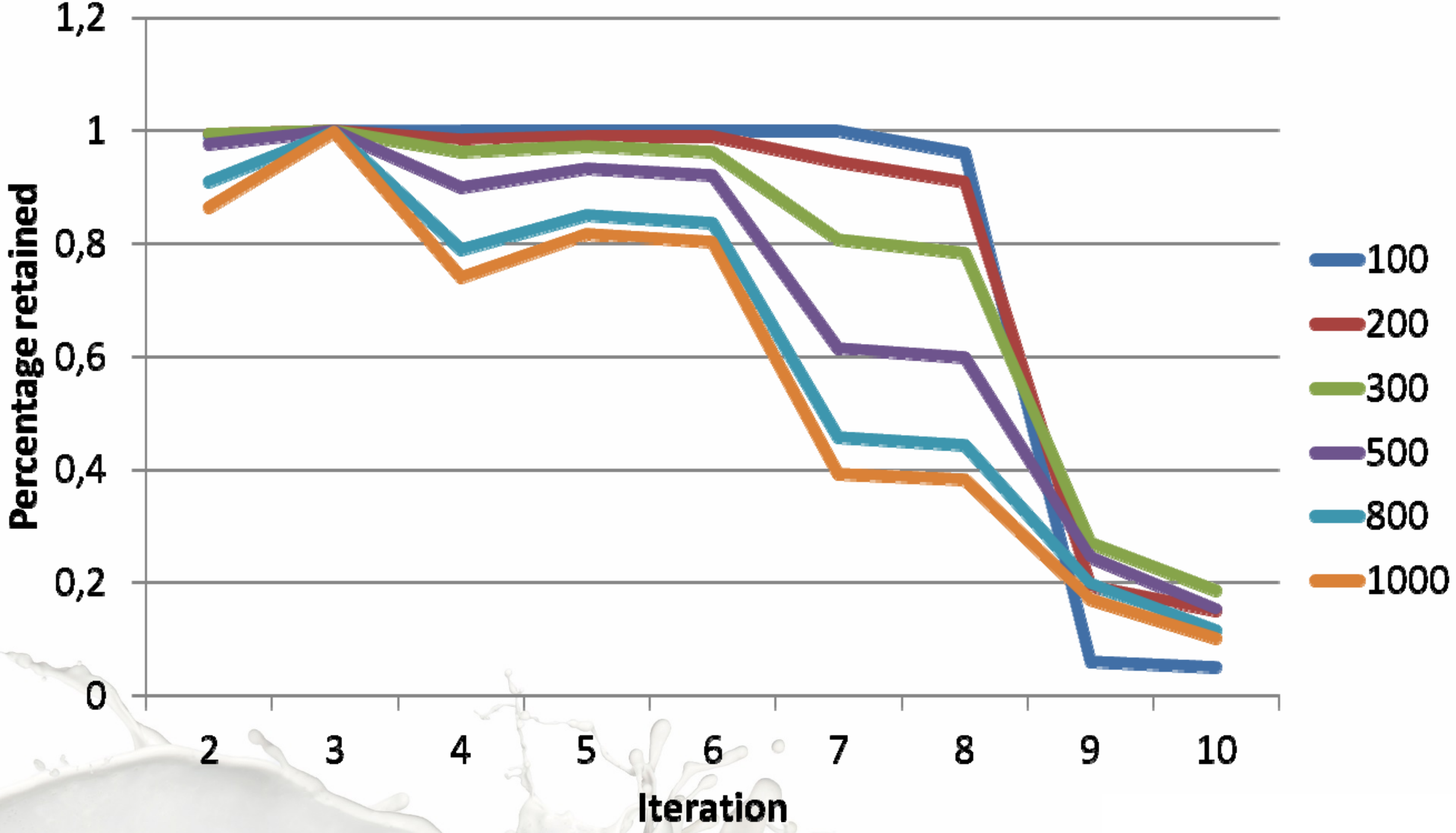


Truncation vs Iterated Pruning



Proportion of Top SNP retained

(prot1)



Conclusions

- Iterative pruning generated significantly different sets of variants compared to truncation selection.
- Do the SNP sets results in different genomic predictions?