Genome wide associations study for fertility and longevity in cattle

Methodological focus

Patrik Waldmann, Gábor Mészáros, Johann Sölkner

Introduction

- Genome wide association study (GWAS) – a well-established technique for identifying genetic variants of interest
- The challenge is to find methods that:
  - Identify true associations
  - Provide a satisfactory results in terms of false positives and false negatives in large-scale GWAS
- The work is a part of our current project "Genome wide association study for functional longevity and related traits in dairy cows"

Methods

- Lasso, elastic net, ridge regression
  - Account for the correlated nature of the predictor variables
  - “Related” to each other

- Single SNP regression
  - The “classical” approach

Methods

$$\hat{\beta}_0, \widehat{\boldsymbol{\beta}} = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \beta_0 - \sum_{j=1}^{p} \boldsymbol{\beta}_j \boldsymbol{X}_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\boldsymbol{\beta}_j^2 + \alpha \left| \boldsymbol{\beta}_j \right| \right] \right\}$$

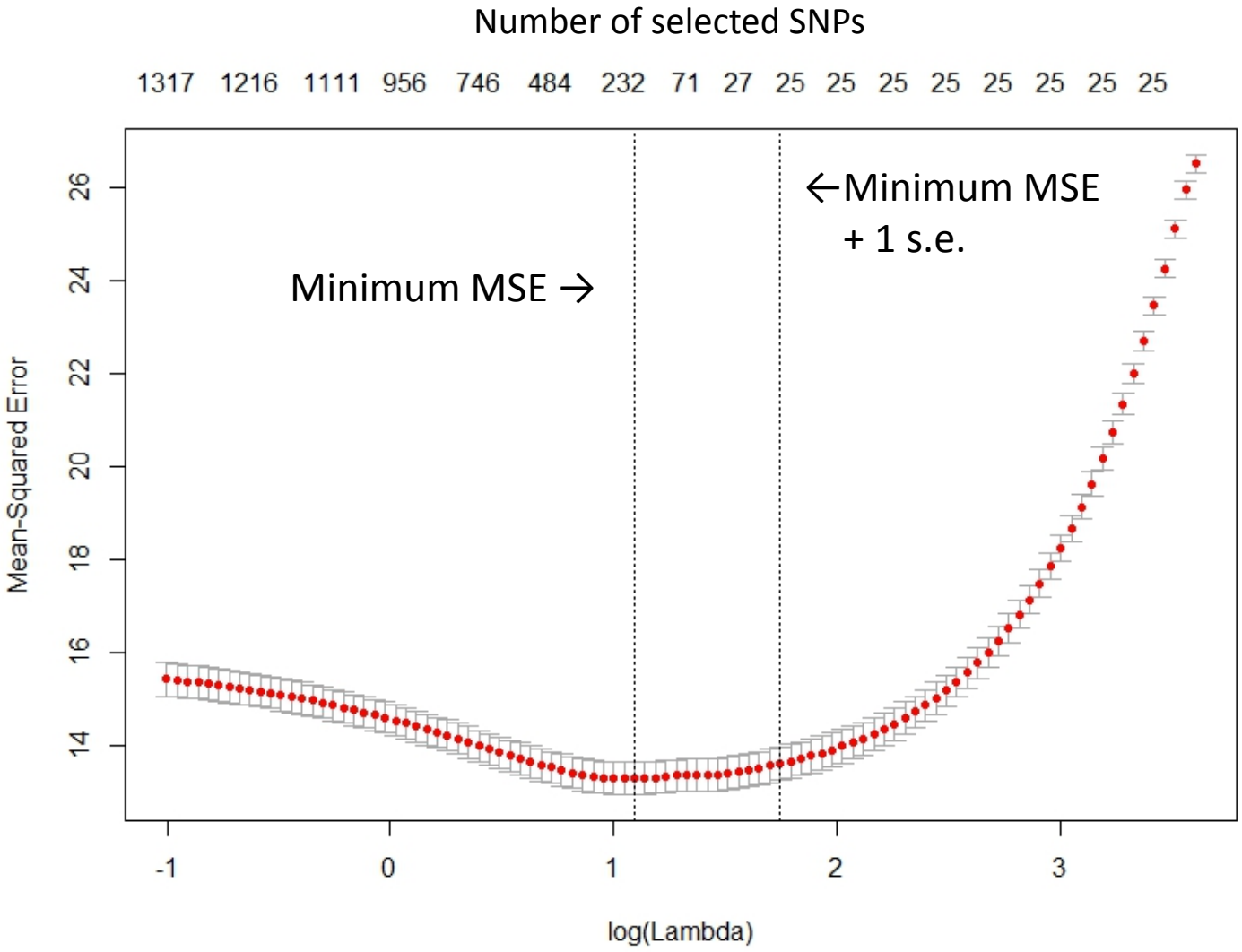
Tuning parameter alpha:

- Lasso if α = 1
- Ridge regression if α = 0
- Elastic net α between 0 and 1, i.e. mixture of lasso and ridge regression

Methods

- Lasso and elastic net using *glmnet* R package (Friedman et al., 2010)
  - $\alpha = \langle 1, 0.75, 0.5, 0.3, 0.1, 0.05, 0.01 \rangle$
- Estimation of population structure using spectral graphs
  - Number of significant eigenvectors from GemTools R package (Klei et al., 2011)

MSE plot – glmnet



Data sets

- Simulation study
  - 1,000 individuals – 50,000 markers – 25 QTLs
- QTLMAS 2010 data set
  - 3,226 individuals – 10,031 markers – 37 QTLs
- Cattle data
  - 4,900 individuals – 33,556 markers – ??? QTLs

Small simulation study

- Simulated 1,000 animals and 50,000 markers
  - 25 QTLs – centered around positions 1,000; 10,000; 20,000; 30,000 and 40,000
- Scenarios:
  - High correlation (LD) between all markers
  - 10 medium and 15 highly correlated markers
  - Medium correlation between all markers
- 100 replicates for each setting

At minimum MSE

		Lasso	EN075	EN05	EN03	EN01	EN005	EN001
High LD (r = 0.94)	Correct out of 25	3	6	10	16	25	25	25
	False positive	3	3	5	6	18	38	734
	MSE	12.76	12.80	12.87	13.01	13.26	13.35	14.06
Mixed LD (r = 0.75)	Correct out of 25	3	5	9	14	20	24	25
	False positive	4	4	5	6	14	30	892
	MSE	12.86	12.89	12.98	13.07	13.33	13.49	14.28
Medium LD (r = 0.55)	Correct out of 25	20	20	21	22	25	25	25
	False positive	6	7	9	12	38	84	1186
	MSE	13.91	13.92	13.92	13.87	13.96	14.06	15.04

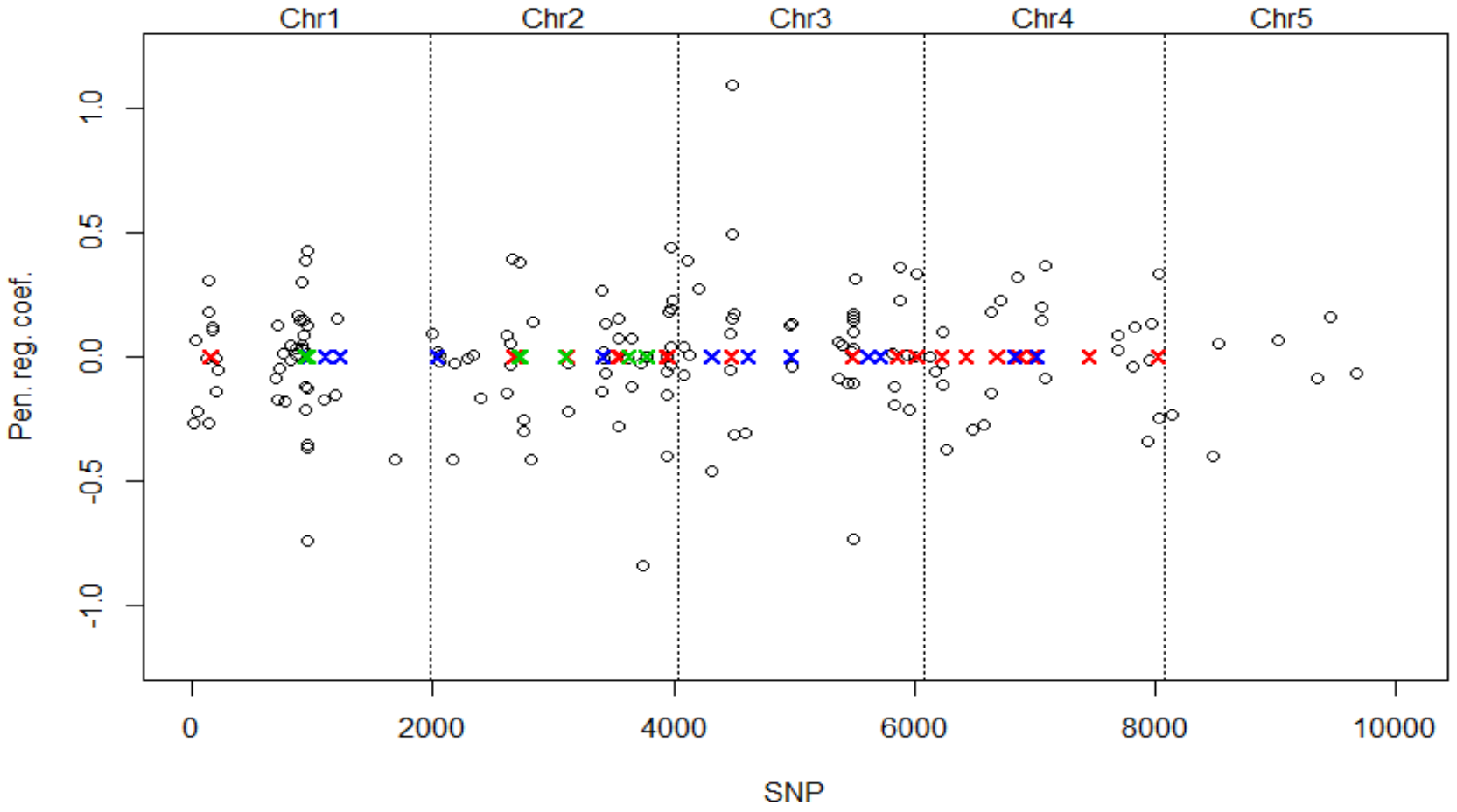
At minimum MSE + 1 s.e.

		Lasso	EN075	EN05	EN03	EN01	EN005	EN001
High LD (r = 0.94)	Correct out of 25	3	7	12	19	25	25	25
	False positive	0	0	0	0	0	0	50
	MSE	13.11	13.21	13.27	13.37	13.63	13.78	14.47
Mixed LD (r = 0.75)	Correct out of 25	2	6	10	15	20	24	25
	False positive	0	0	0	0	0	0	111
	MSE	13.23	13.29	13.38	13.45	13.76	13.88	14.65
Medium LD (r = 0.55)	Correct out of 25	19	20	21	23	25	25	25
	False positive	0	0	0	0	0	0	220
	MSE	14.32	14.35	14.28	14.23	14.38	14.49	15.44

QTLMAS 2010 data set

- More complex structure
- 3,226 individuals (5 generation pedigree)
- Five autosomal chromosomes – 100 Mb each
- 10,031 markers – 37 QTLs

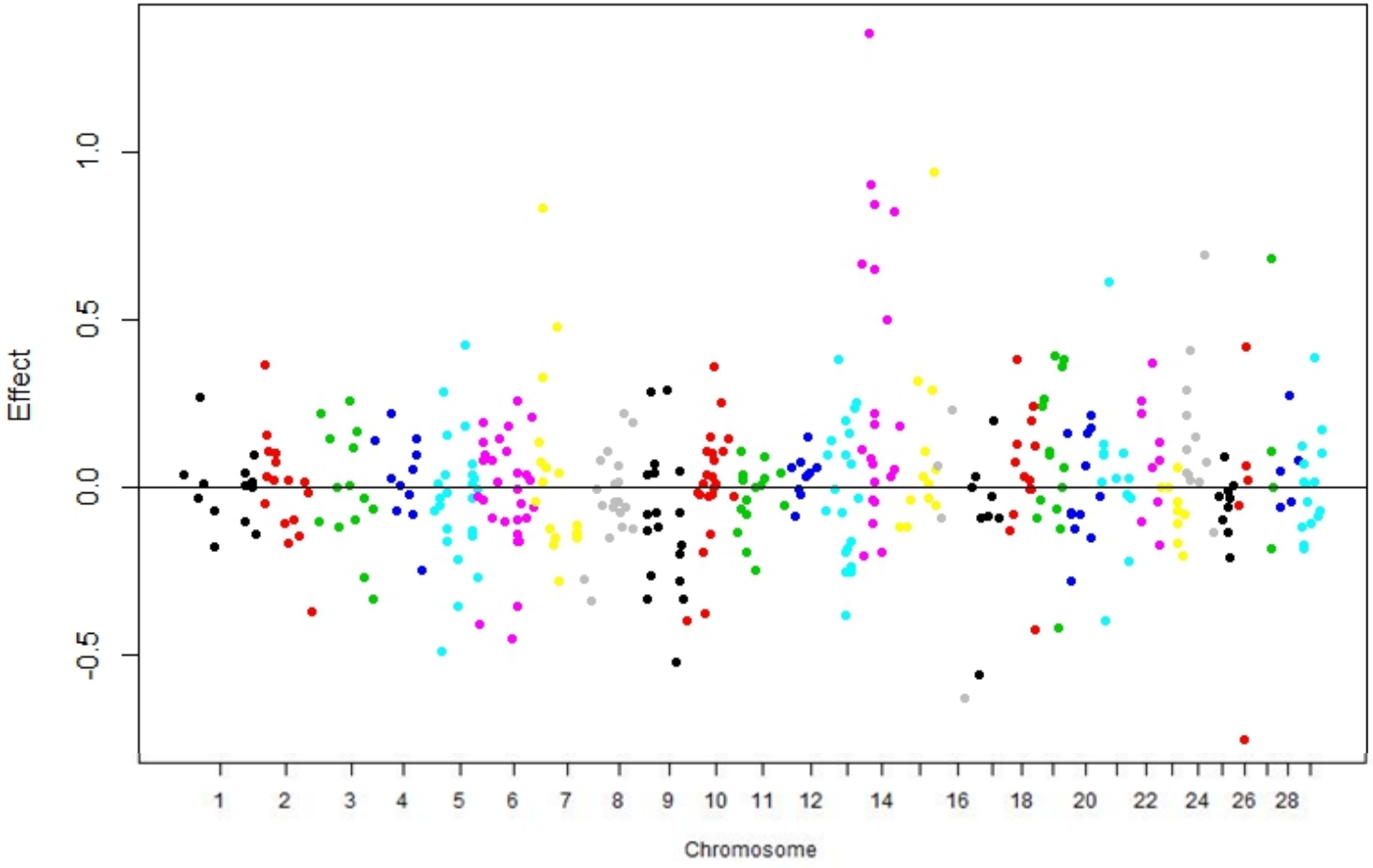
Elastic net (α=0.1) with pop. str.



Cattle data

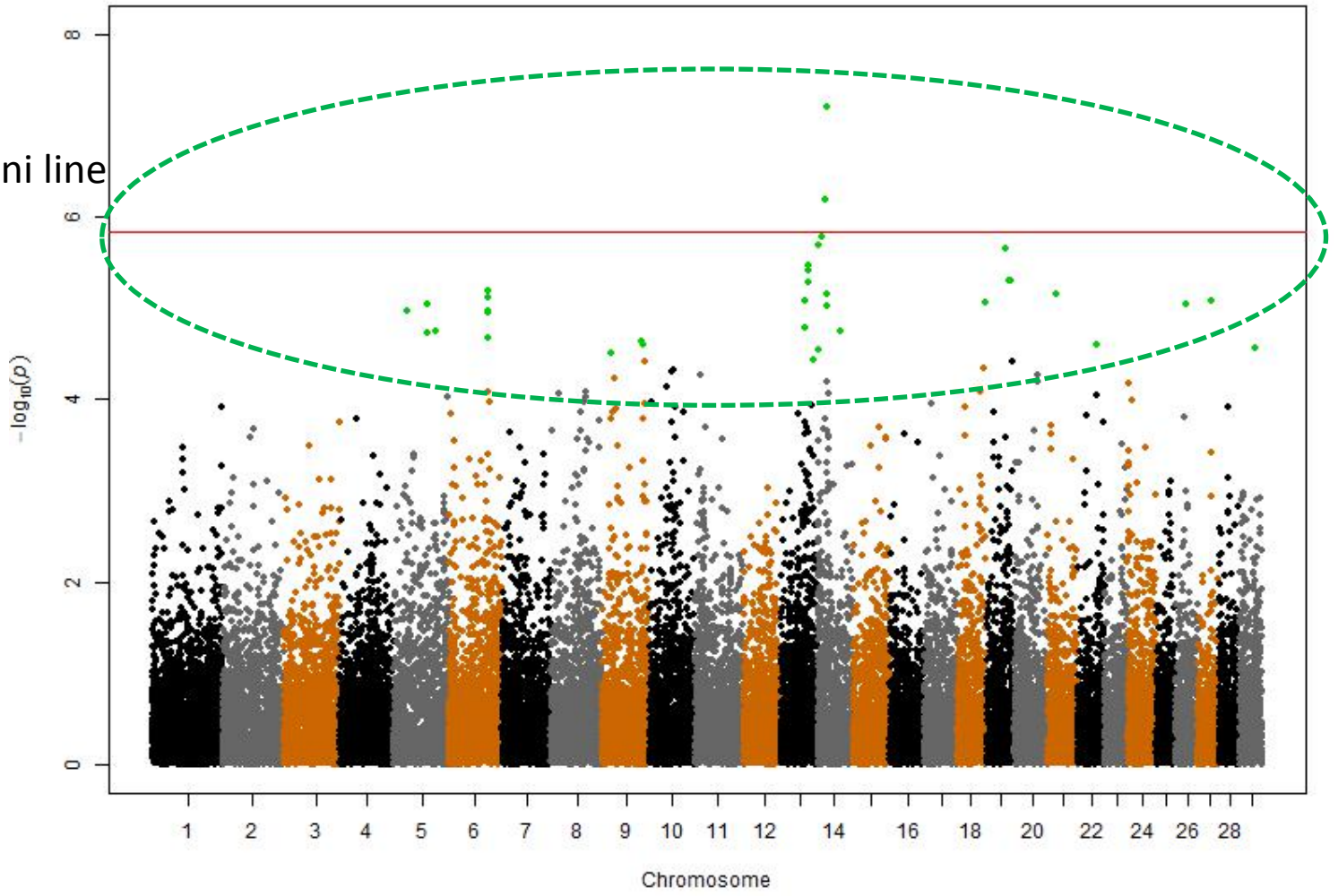
- German – Austrian Fleckvieh pool data (50k)
- Only unambiguously mapped SNPs
- After quality control: 33,556 SNPs
- Phenotype: deregressed breeding values
- Longevity: 4,887 bulls; Fertility: 4,905 bulls
- Population structure - spectral clustering techniques

Longevity – Elasic net (α= 0.1)

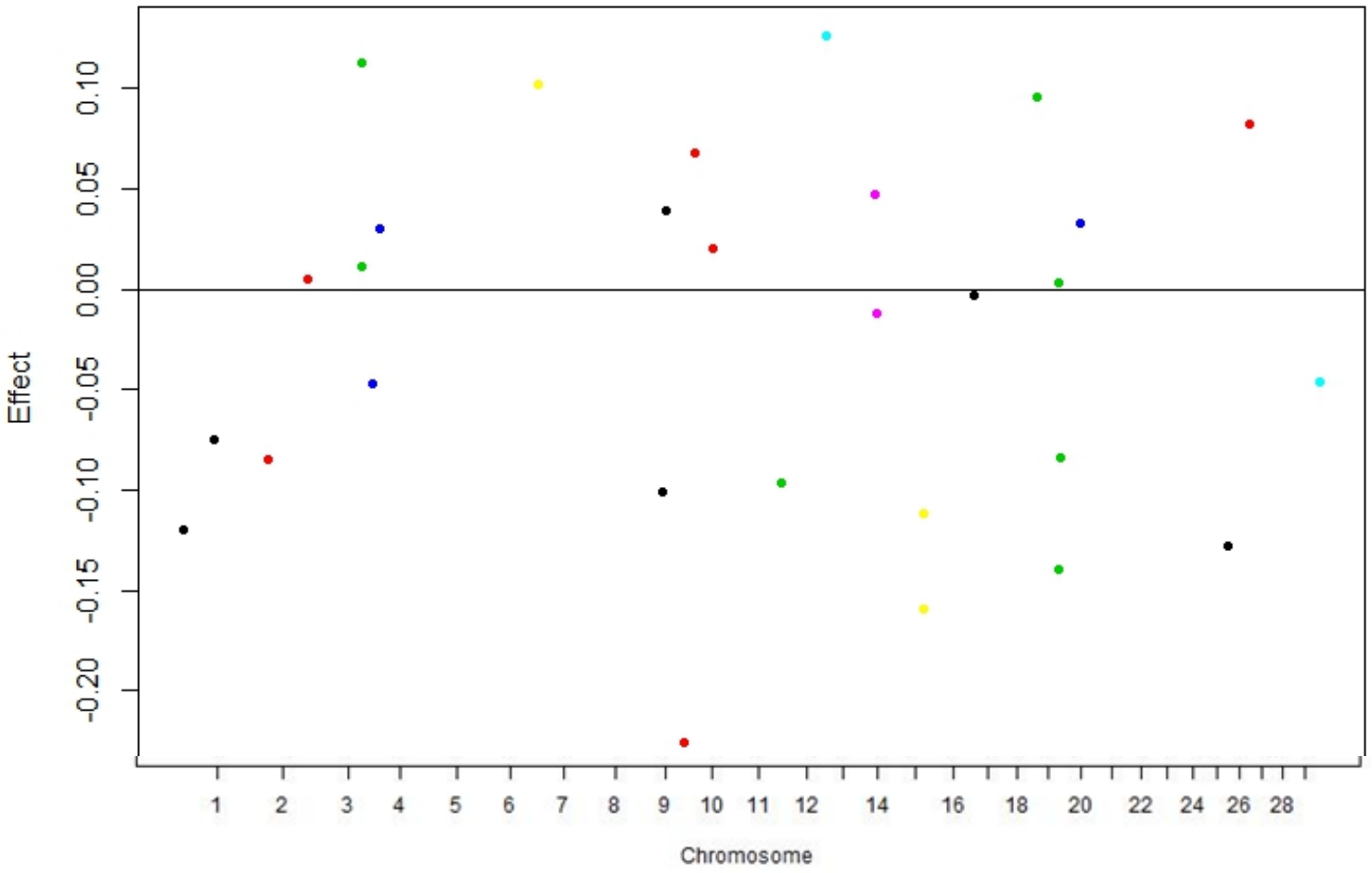


Single SNP regression - Longevity

Bonferroni line

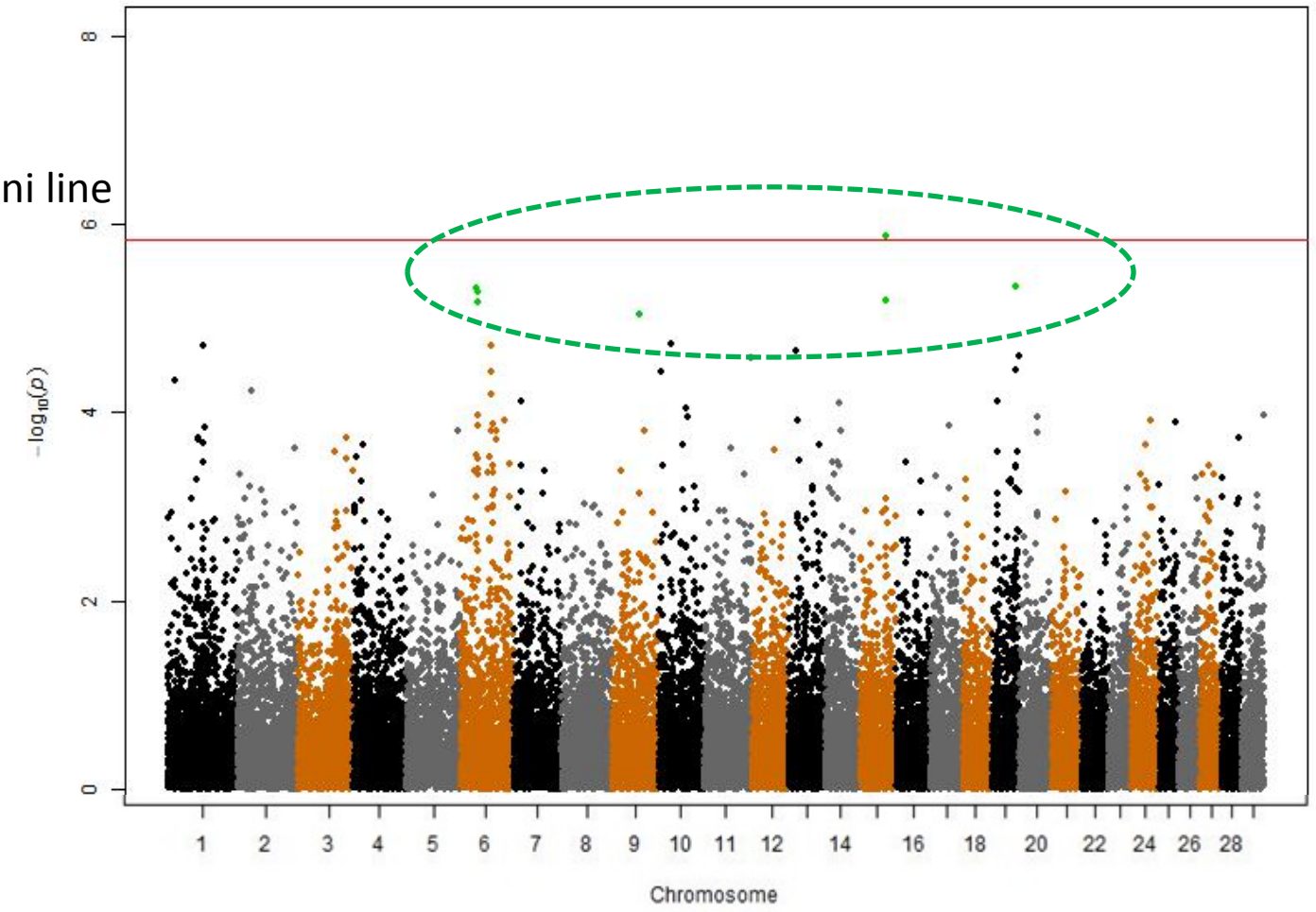


Fertility – Elasic net (α= 0.1)



Single SNP regression - Fertility

Bonferroni line



Results outcome

- Elastic net works well in simulated data
- Real data – large differences in number of "significant" SNPs
- Possible reasons:
  - Elastic net might be sensitive to the lack of error variance in phenotype values
  - More input information – there were no SNPs picked with 2k genotypes (ref. Abstract book)
- Complex problem, not yet solved

Conclusions

- Lasso and elastic net – possibility to perform variable selection with correlated predictor variables
- Elastic net with $\alpha \sim 0.1$ gave the best result in the simulations
- Further study of the variable selection criteria in case of highly multi-collinear data is needed

Acknowledgements

Data from breeding organizations

- Förderverein Biotechnologieforschung
- Rinderbesammungsgenossenschaft Memmingen
- Gesellschaft zur Förderung der Fleckviezucht in Niederbayern
- Nutzvieh GmbH Miesbach
- Rinderunion Baden-Württemberg eG
- Zentrale Arbeitsgemeinschaft Österreichischer Rinderzüchter
- Arbeitsgemeinschaft Süddeutscher Rinderzucht- und Besamungsorganisationen

Breeding values

- ZuchtData EDV-Dienstleistungen GmbH

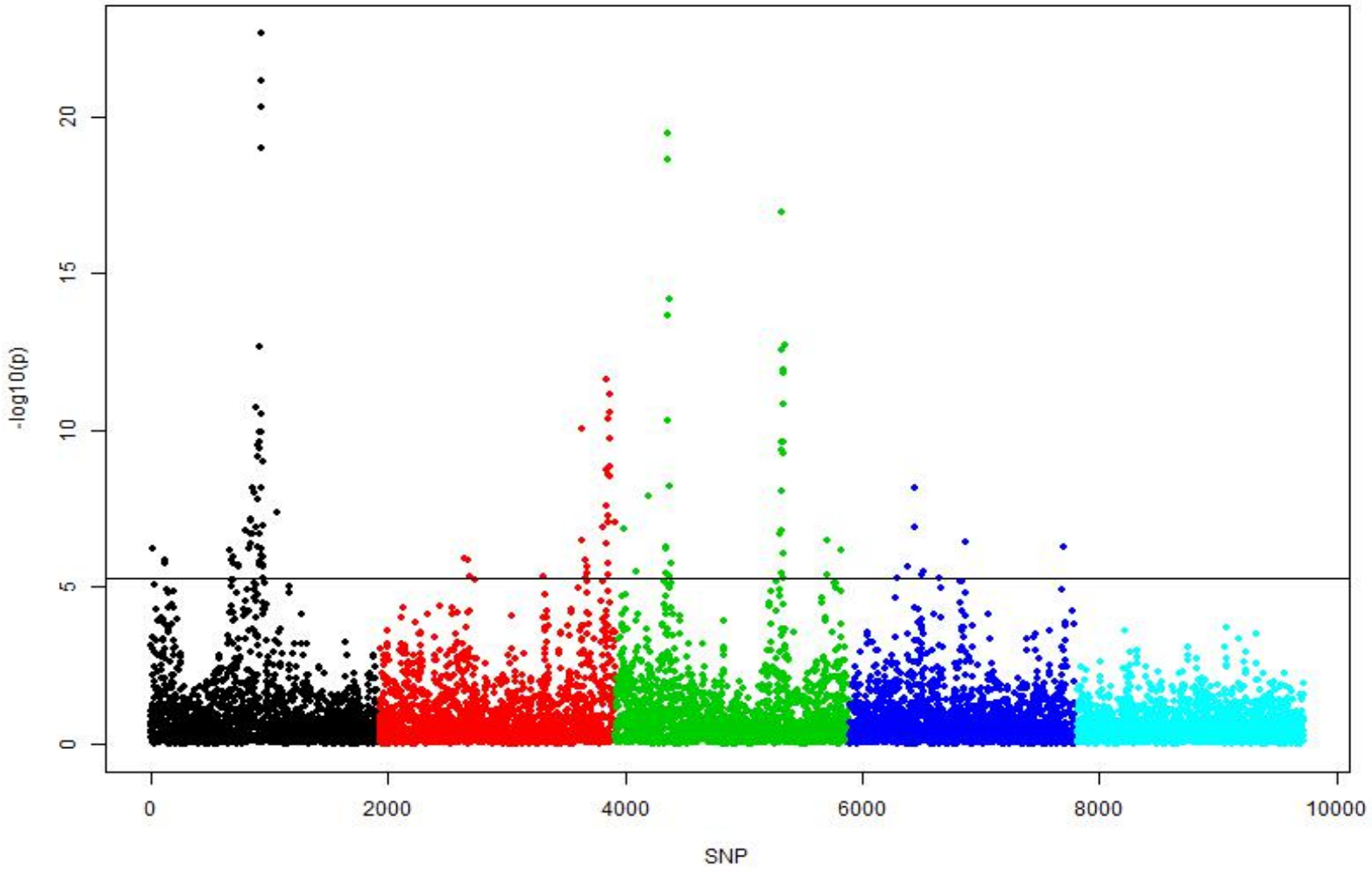
Funding

- Austrian Science Fund (FWF) “Genome wide association study for functional longevity and related traits in dairy cows” TRP 46-B19

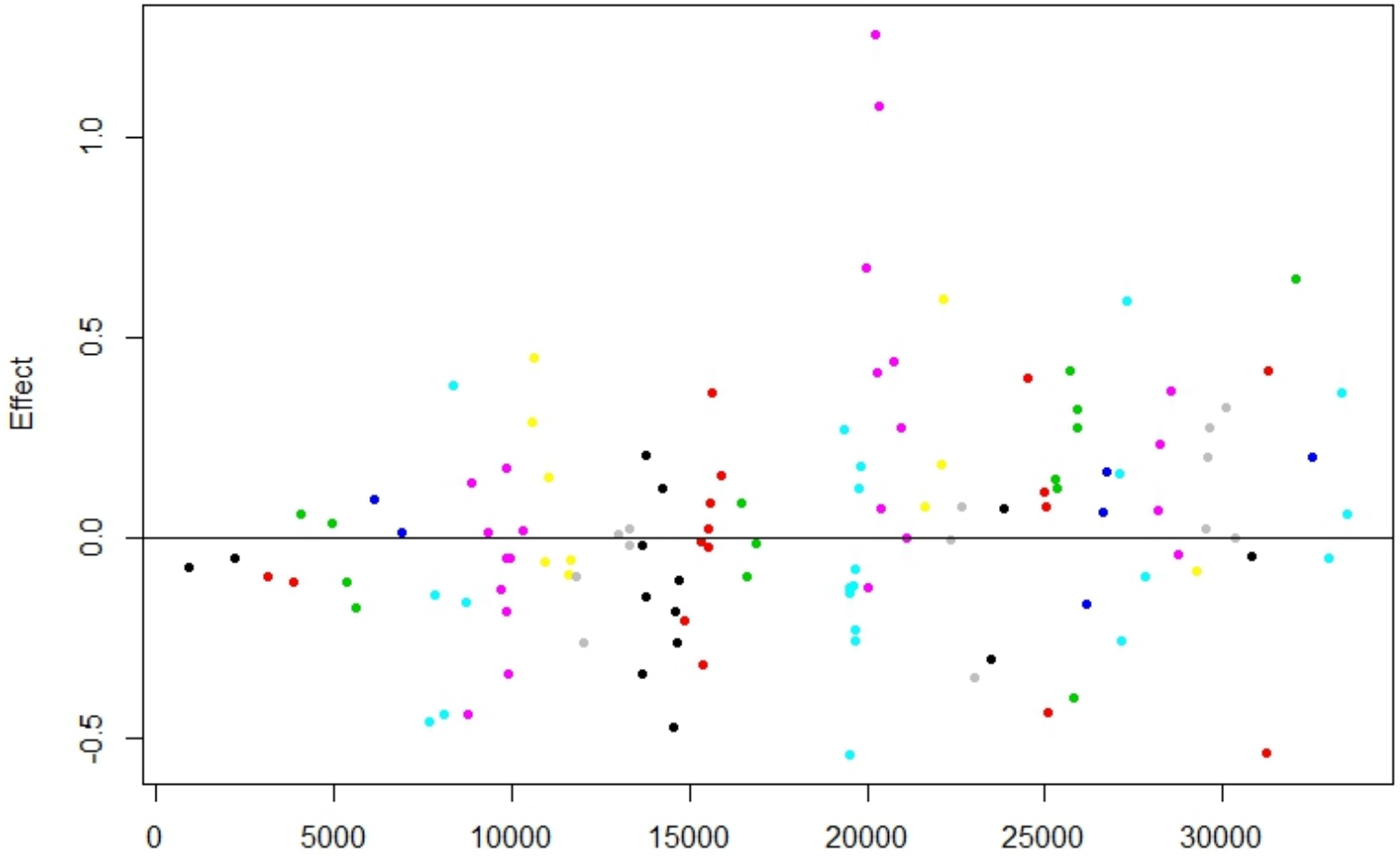
Thanks for your attention!

gabor.meszaros@boku.ac.at

Single SNP regression – QTLMAS2010



Longevity – Elasic net (α= 0.99)



QTLMAS2010 – selected SNPs

		Lasso	EN09	EN075	EN05	EN03	EN01	EN005
No pop. struct. corr.	Selected SNPs	163	164	181	213	229	305	436
	minMSE + 1 SE	60.80	60.82	60.77	60.79	60.75	60.94	60.95
Pop. struct. corr.	Selected SNPs	87	88	90	91	102	149	240
	minMSE + 1 SE	60.36	60.42	60.45	60.84	60.53	61.19	61.40

At minimum MSE

		Lasso	EN075	EN05	EN03	EN01	EN005	EN001
High LD	Correct	3 (0.78)	6 (1.53)	10 (2.09)	16 (2.29)	25 (0.55)	25 (0)	25 (0)
	False positive	3 (15.1)	3 (13.1)	5 (23.8)	6 (20.0)	18 (43.2)	38 (76.8)	734 (414)
	MSE	12.76 (0.281)	12.80 (0.273)	12.87 (0.274)	13.01 (0.259)	13.26 (0.265)	13.35 (0.270)	14.06 (0.310)
Mixed LD	Correct	3 (0.82)	5 (1.47)	9 (1.82)	14 (1.97)	20 (1.50)	24 (0.93)	25 (0)
	False positive	4 (10.1)	4 (11.1)	5 (16.6)	6 (21.5)	14 (33.1)	30 (56.1)	892 (377)
	MSE	12.86 (0.335)	12.89 (0.341)	12.98 (0.322)	13.07 (0.310)	13.33 (0.292)	13.49 (0.304)	14.28 (0.340)
Low LD	Correct	20 (1.31)	20 (1.34)	21 (1.48)	22 (1.37)	25 (0.51)	25 (0)	25 (0)
	False positive	6 (29.3)	7 (25.2)	9 (23.5)	12 (29.4)	38 (49.1)	84 (91.5)	1186 (395)
	MSE	13.91 (0.300)	13.92 (0.301)	13.92 (0.285)	13.87 (0.300)	13.96 (0.273)	14.06 (0.280)	15.04 (0.357)

At minimum MSE + 1 s.e.

		Lasso	EN075	EN05	EN03	EN01	EN005	EN001
High LD	Correct	3 (0.72)	7 (1.46)	12 (2.07)	19 (1.92)	25 (0)	25 (0)	25 (0)
	False positive	0 (0.51)	0 (0.46)	0 (0.76)	0 (2.42)	0 (0.85)	0 (1.38)	50 (110)
	MSE	13.11 (0.297)	13.21 (0.292)	13.27 (0.293)	13.37 (0.276)	13.63 (0.280)	13.78 (0.303)	14.47 (0.307)
Mixed LD	Correct	2 (0.85)	6 (1.40)	10 (1.66)	15 (1.38)	20 (1.51)	24 (0.89)	25 (0)
	False positive	0 (0)	0 (0.14)	0 (0.71)	0 (0.14)	0 (0.83)	0 (2.33)	111 (139)
	MSE	13.23 (0.348)	13.29 (0.368)	13.38 (0.361)	13.45 (0.323)	13.76 (0.316)	13.88 (0.342)	14.65 (0.361)
Low LD	Correct	19 (1.78)	20 (1.67)	21 (1.56)	23 (1.16)	25 (0.36)	25 (0)	25 (0)
	False positive	0 (1.12)	0 (2.54)	0 (2.62)	0 (0.83)	0 (3.25)	0 (6.80)	220 (197)
	MSE	14.32 (0.346)	14.35 (0.309)	14.28 (0.295)	14.23 (0.342)	14.38 (0.307)	14.49 (0.283)	15.44 (0.357)