Feature selection for genomic prediction

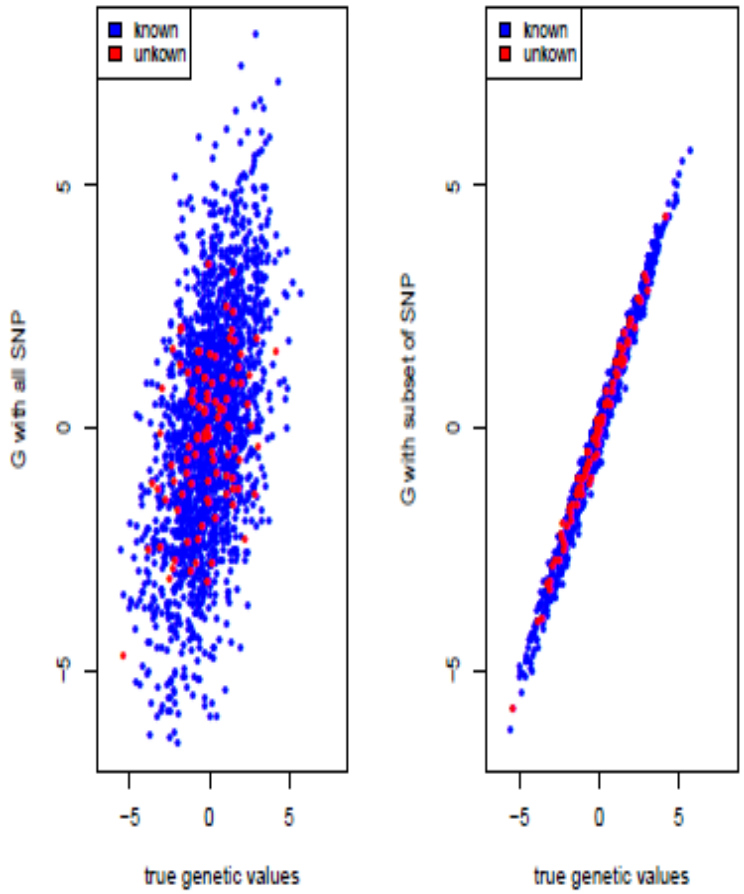
Cedric Gondro

The big picture

- how well can we predict
  - effects of methods
  - effects of population
  - effects of architecture
- do we need to know the truth
- heritability X predictability
- prediction as feature selection

Ideally…

- In a perfect world we would know the true SNP associated to a trait or even better, the functional causal variants
- We would know the variants of large effect but also all the ones with small effects
- And we would use only them for making predictions…

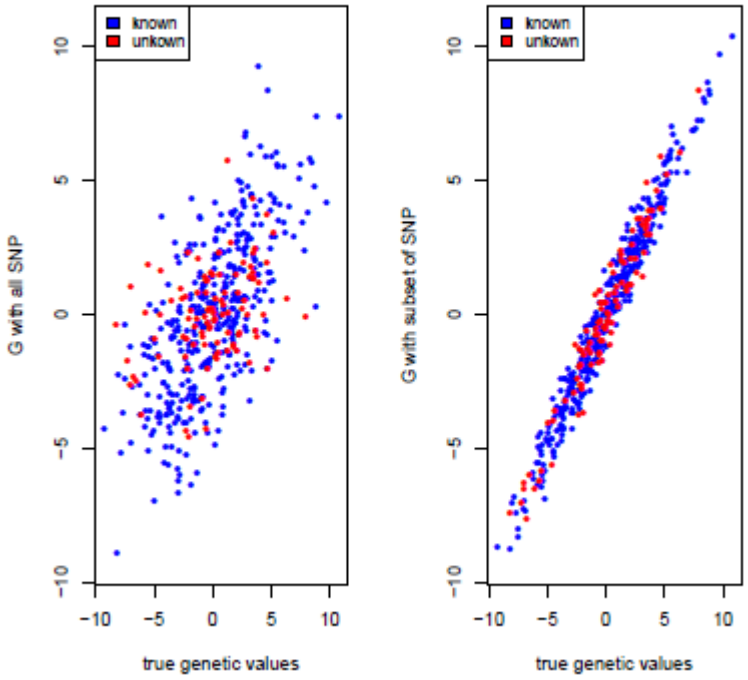


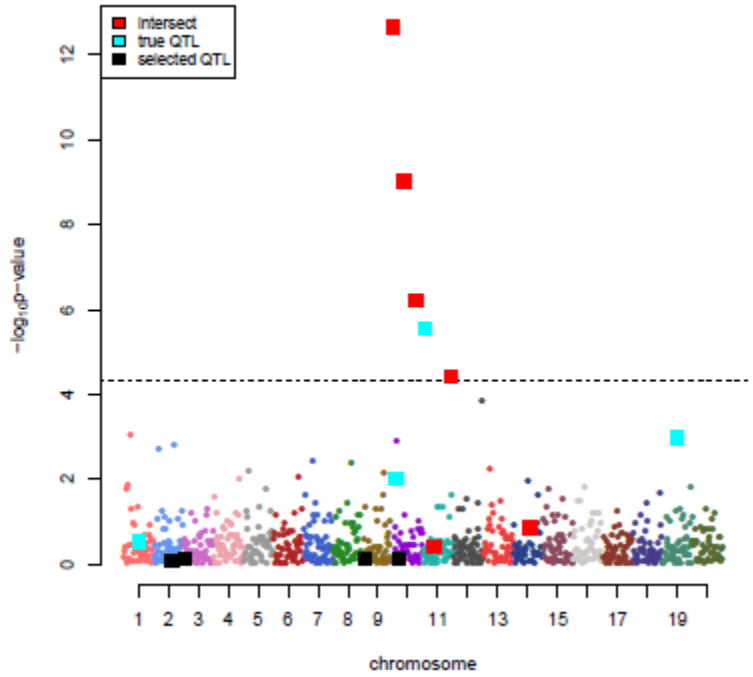
Can we develop a smaller, cheaper and better panel?

tBLUP

Use *trait G* instead of G
trait relationship matrix – TRM

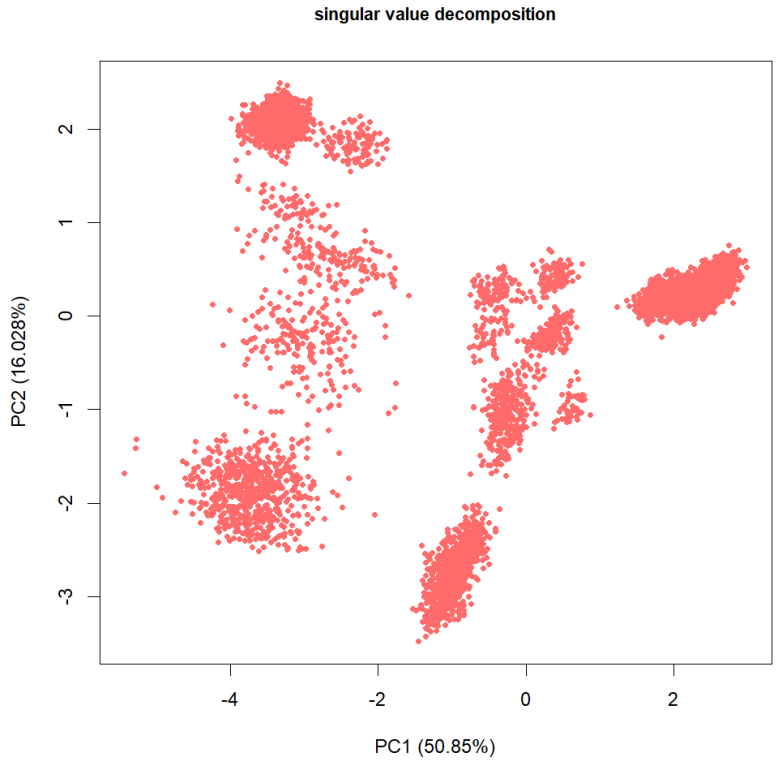
gBLUP using only functional markers



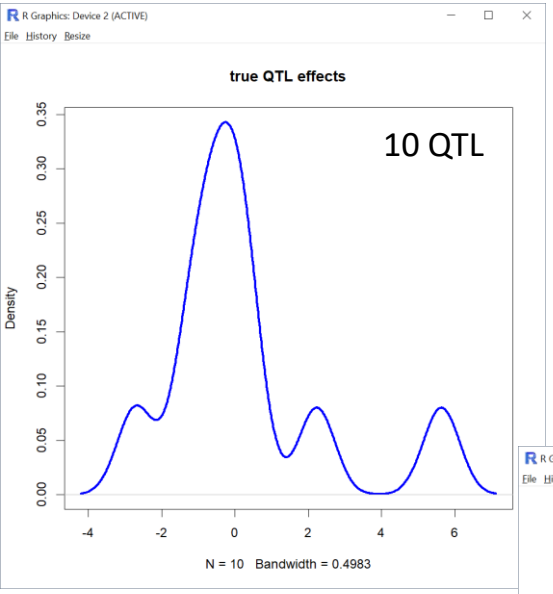


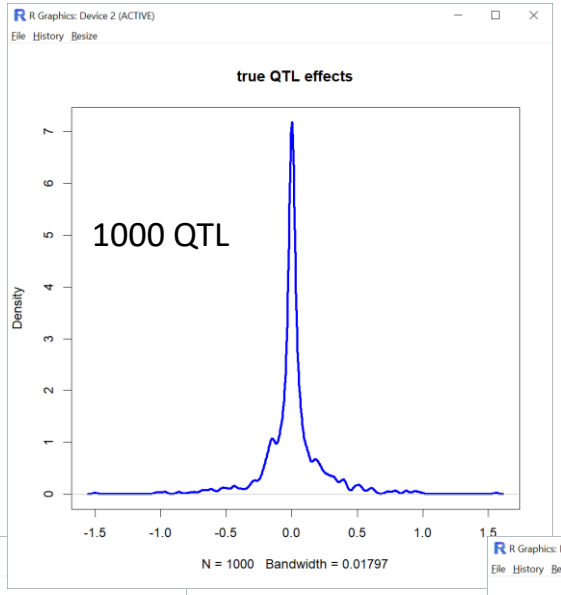
Backdrop…
Some simulations

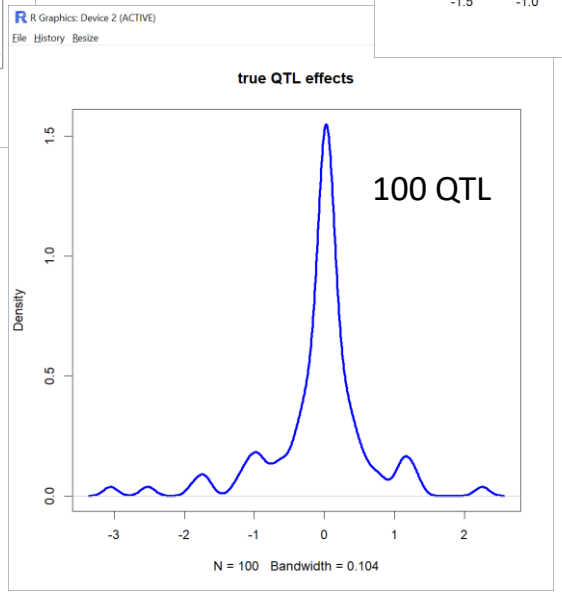
- 50k data
- 7539 animals
- multibreed
- simple additive model
- QTL included in data
- h^2=0.4, V_a=50

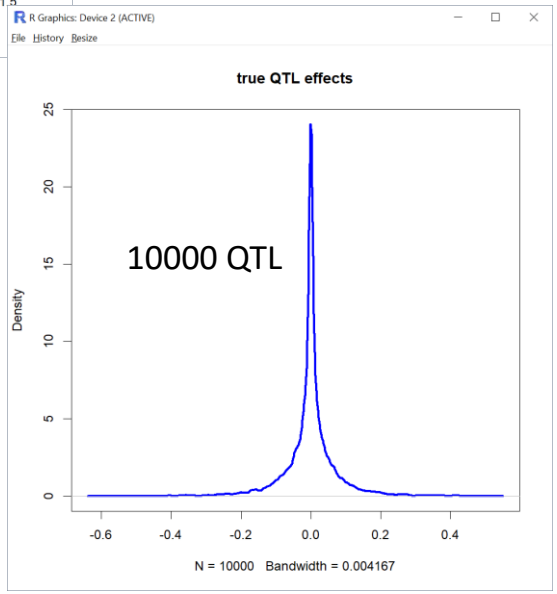


Additive marker effects

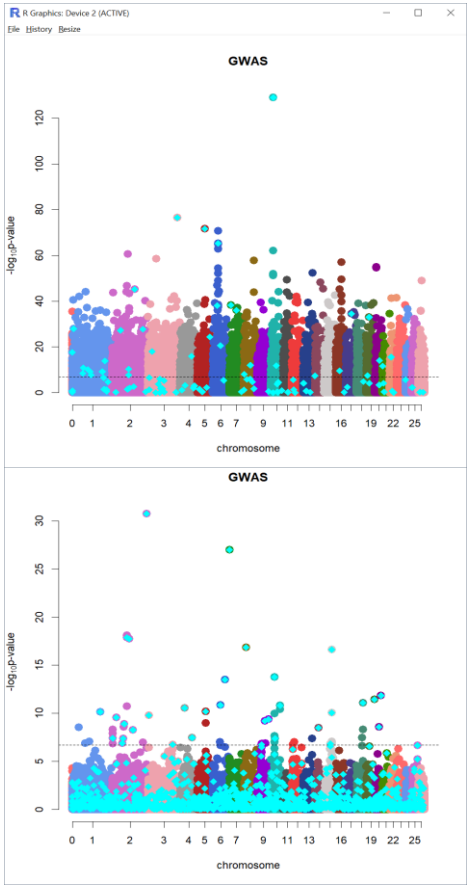








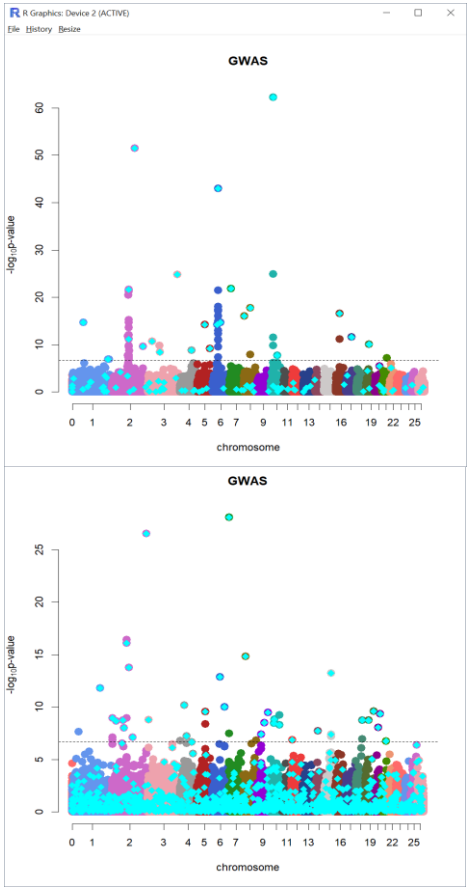
Simple GWAS



100 QTL

before adjusting for population structure

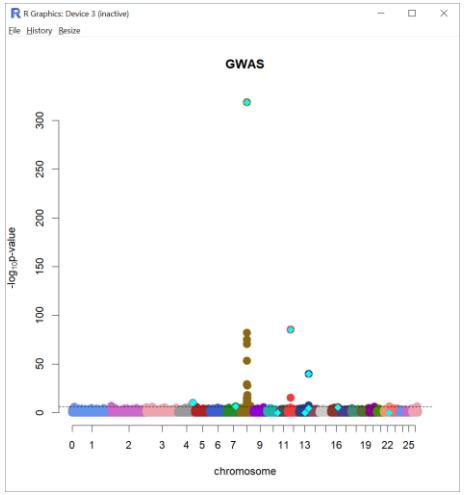
1000 QTL

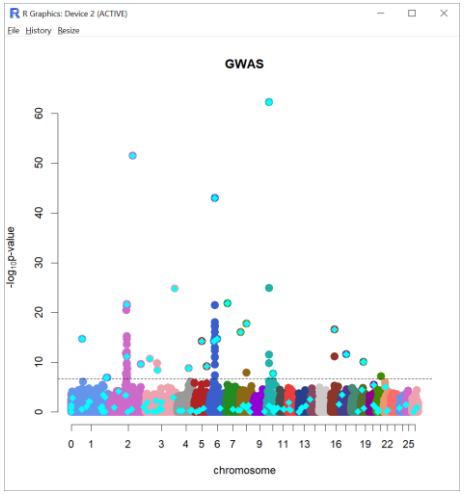


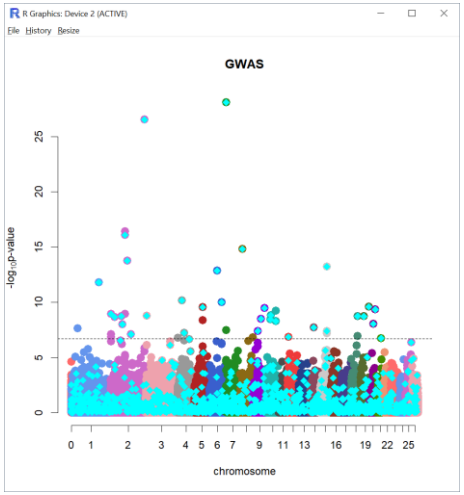
Correction effect is larger when there are fewer QTL

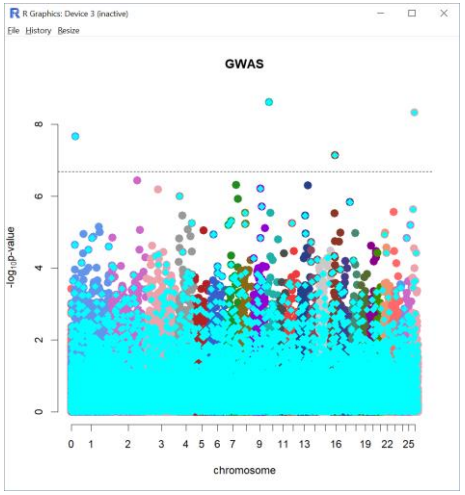
after adjusting for population structure

GWAS





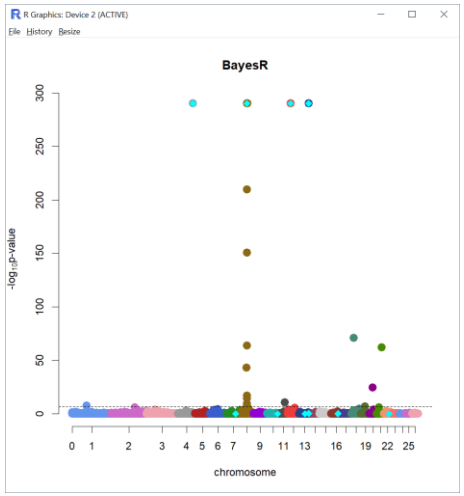


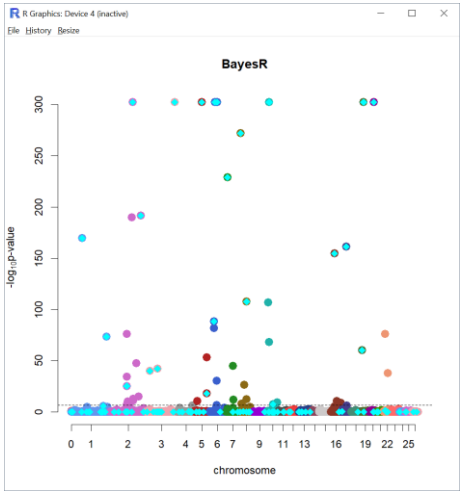


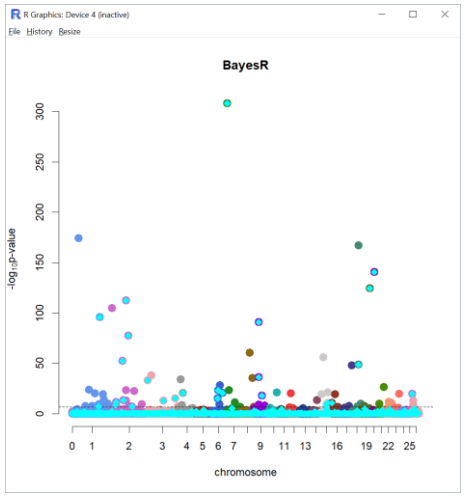
LD plays a large role overestimates regions

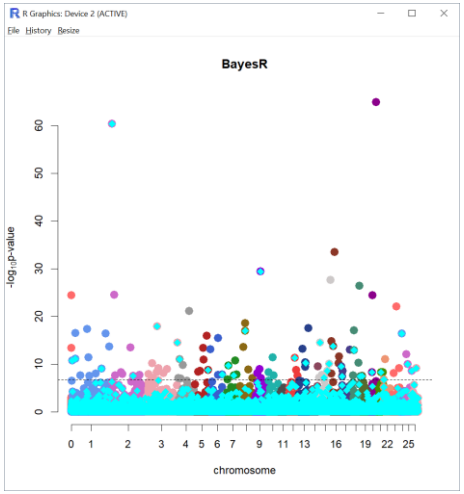
10 -> 100 -> 1000 -> 10000 QTL

BayesR





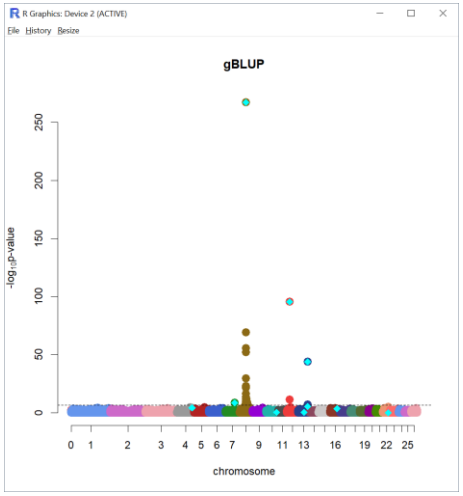


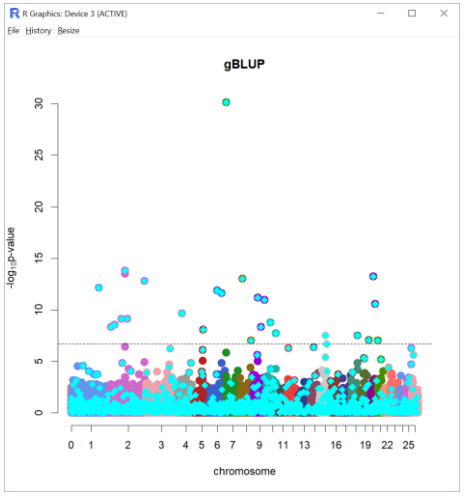


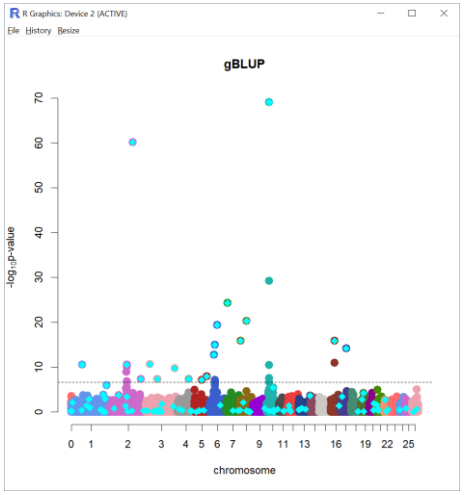
accounts better for LD
still many false positives

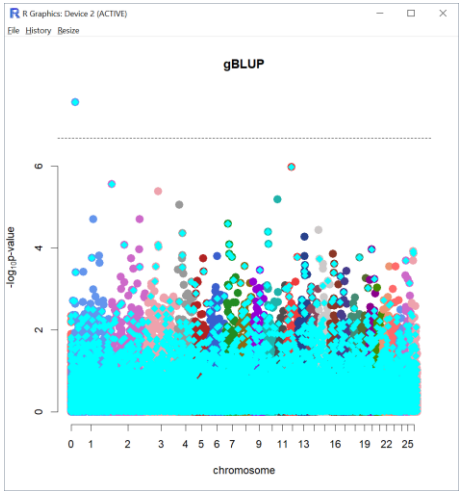
10 -> 100 -> 1000 -> 10000 QTL

GBLUP









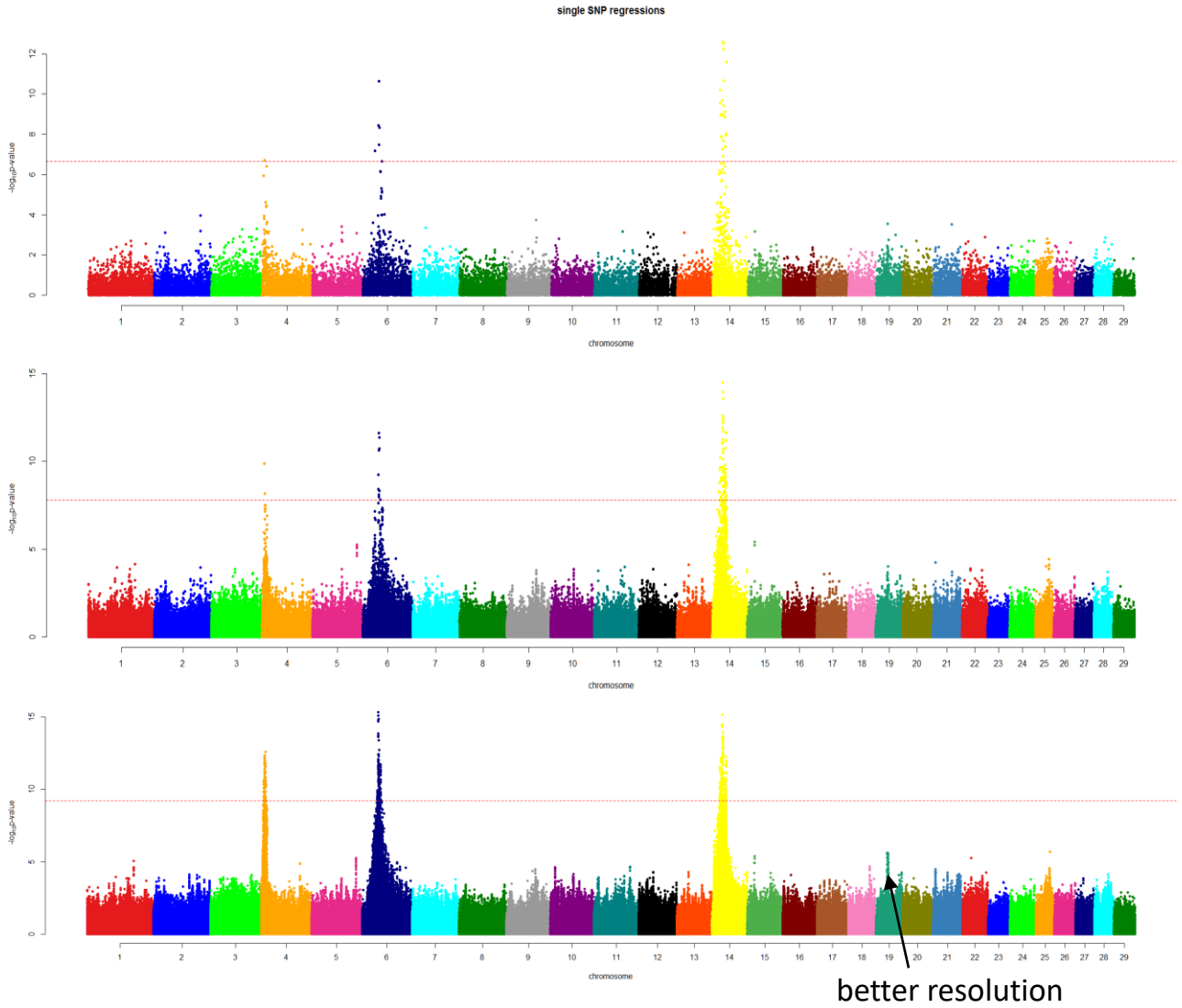
accounts well(ish) for LD
less false positives
more conservative

10 -> 100 -> 1000 -> 10000 QTL

50k

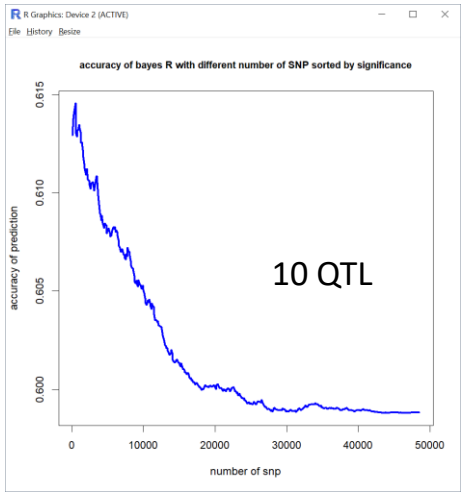
700k

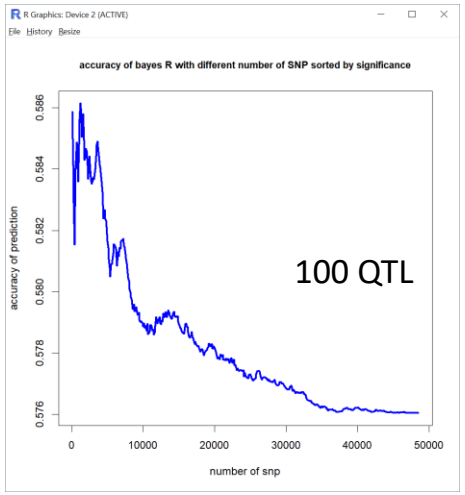
seq

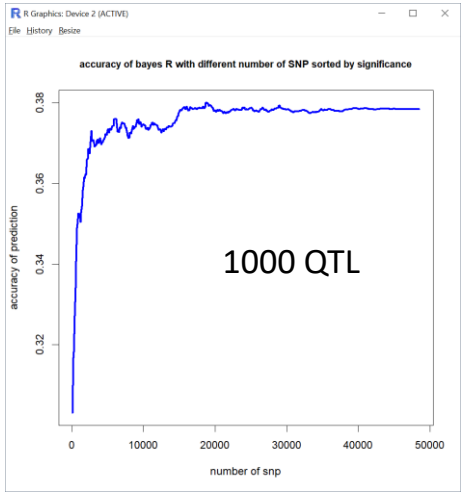


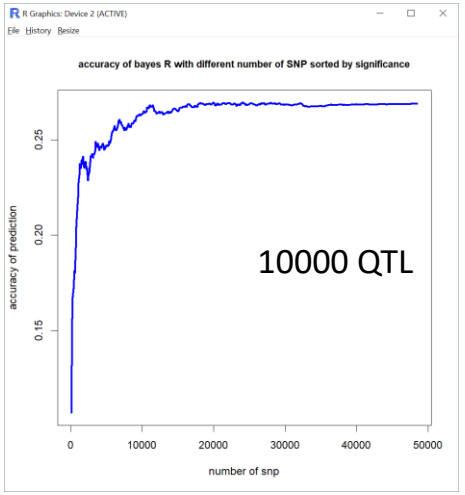
GWAS and marker numbers – cattle data, h^2=0.6

A first look at feature selection









With few QTL the estimates of true effects are more correct – the top subset is more predictive

Adding more SNP only adds noise and reduces prediction accuracy

But change in accuracy is still marginal

With many QTL the true effects are all wrong, basically approximates GBLUP

Needs to use all SNP to improve prediction accuracy

accuracy of prediction using only top SNP (1%, 5% and 10%)
original – original effect estimates used directly for prediction
new – effects re-estimated with GBLUP and then used for prediction
still need to re-test subsets with BayesR

method and number of SNP	*10*		*100*		*1000*		*10000*	
	new	*original*	*new*	*original*	*new*	*original*	*new*	*original*
gwas 485	0.597	0.478	0.535	0.387	0.460	0.358	0.252	0.241
gwas 2427	0.494	0.368	0.452	0.313	0.443	0.326	0.281	0.286
gwas 4854	0.468	0.333	0.446	0.275	0.412	0.304	0.316	0.288
gblup 485	0.508	0.502	0.499	0.455	0.477	0.428	0.260	0.237
gblup 2427	0.426	0.459	0.419	0.436	0.419	0.401	0.289	0.300
gblup 4854	0.398	0.431	0.416	0.442	0.378	0.383	0.288	0.301
bayesr 485	0.564	0.615	0.543	0.584	0.401	0.334	0.193	0.182
bayesr 2427	0.469	0.611	0.459	0.584	0.406	0.368	0.237	0.229
bayesr 4854	0.434	0.608	0.448	0.582	0.395	0.373	0.250	0.247

best

somehow we know the SNP and use them to estimate effects and predict

the magic full truth – QTL and real effects are known

or we have the global effects, somehow found the QTL and used the original effects with just the QTL

theoretical maximum accuracy: 0.632

accuracy with true QTL using GBLUP

10	0.659
100	0.633
1000	0.568
10000	0.356

accuracy with true QTL effects

10	0.659
100	0.643
1000	0.634
10000	0.629

SNP	GWAS	GBLUP	BAYESR
10	0.655	0.655	0.629
100	0.621	0.585	0.605
1000	0.501	0.540	0.366
10000	0.315	0.346	0.280

How does sub-setting compare to using all SNP?

method and number of SNP	*10*		*100*		*1000*		*10000*	
	new	*original*	*new*	*original*	*new*	*original*	*new*	*original*
gwas 485	0.597	0.478	0.535	0.387	0.460	0.358	0.252	0.241
gwas 2427	0.494	0.368	0.452	0.313	0.443	0.326	0.281	0.286
gwas 4854	0.468	0.333	0.446	0.275	0.412	0.304	**0.316**	0.288
gblup 485	0.508	0.502	0.499	0.455	**0.477**	0.428	0.260	0.237
gblup 2427	0.426	0.459	0.419	0.436	0.419	0.401	0.289	0.300
gblup 4854	0.398	0.431	0.416	0.442	0.378	0.383	0.288	0.301
bayesr 485	0.564	**0.615**	0.543	**0.584**	0.401	0.334	0.193	0.182
bayesr 2427	0.469	0.611	0.459	0.584	0.406	0.368	0.237	0.229
bayesr 4854	0.434	0.608	0.448	0.582	0.395	0.373	0.250	0.247

	all SNP		top SNP	
SNP	GBLUP	BAYESR	GBLUP	BAYESR
10	0.389	0.599	0.508	0.615
100	0.441	0.576	0.499	0.584
1000	0.347	0.378	0.477	0.406
10000	0.311	0.269	0.301	0.250

… but maximum of 4854 SNP used and there are 10000 QTL
if top 20% (~10k) used then GBLUP accuracy = 0.319 original (or 0.306 with re-estimated effects)
BayesR still lower at 0.264 original and 0.274 re-estimate

GWAS was already higher with 4854 SNP (0.316); 0.321 with top 20%

A case for feature selection

using only significant SNP (1% Bonferroni)

	10			*100*			*1000*			*10000*		
	SNP	new	original	SNP	new	original	SNP	new	original	SNP	new	original
GWAS	39	0.648	0.517	65	0.363	0.411	45	0.211	0.379	4	0.127	0.125
GBLUP	26	0.648	0.541	29	0.589	0.522	26	0.414	0.391	1	NA	0.05
BAYESR	21	0.648	0.613	48	0.603	0.582	84	0.371	0.295	113	0.147	0.118
		✓			✓			✓			✗	

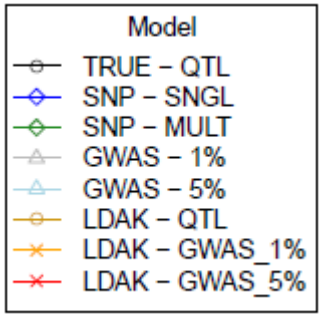
	all SNP		top SNP
SNP	GBLUP	BAYESR	best in class
10	0.389	0.599	0.615
100	0.441	0.576	0.584
1000	0.347	0.378	0.477
10000	0.311	0.269	0.316

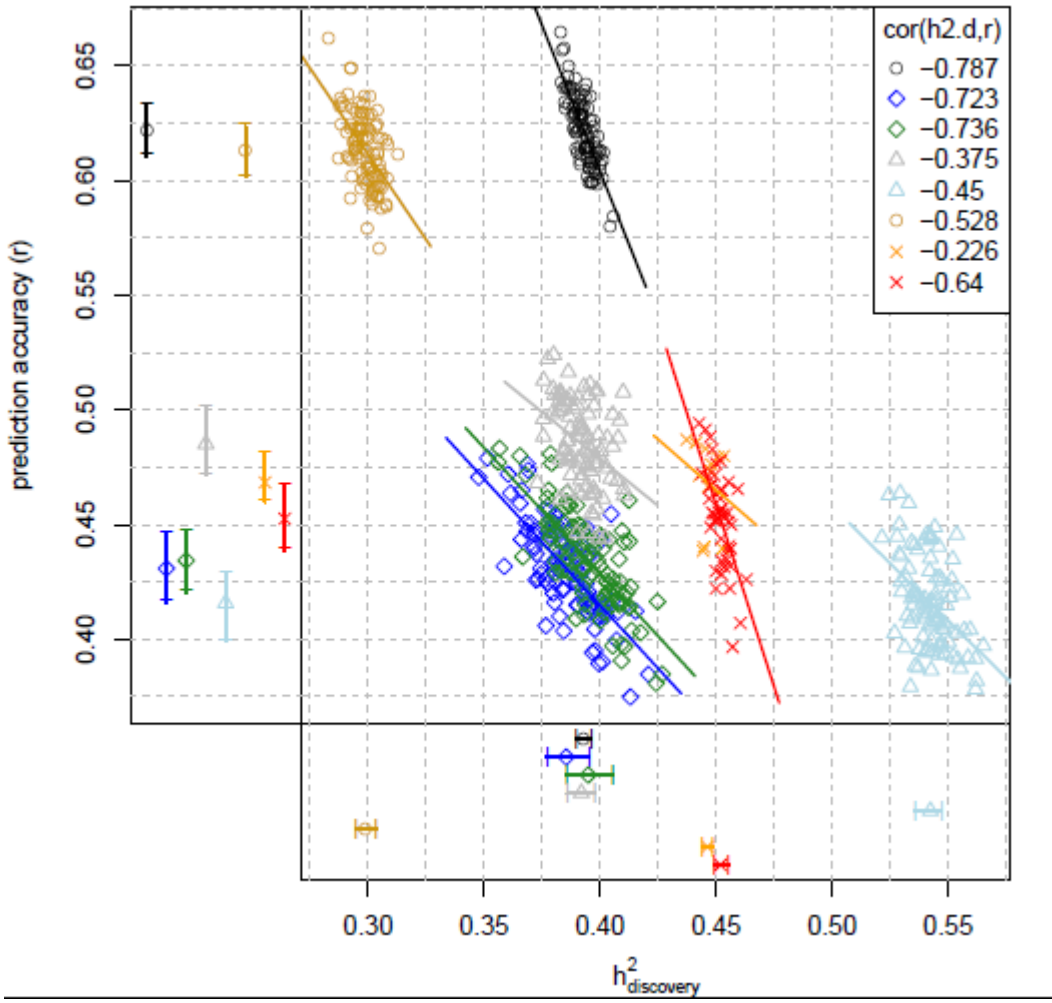
theoretical maximum accuracy: 0.632

accuracy with true QTL using GBLUP

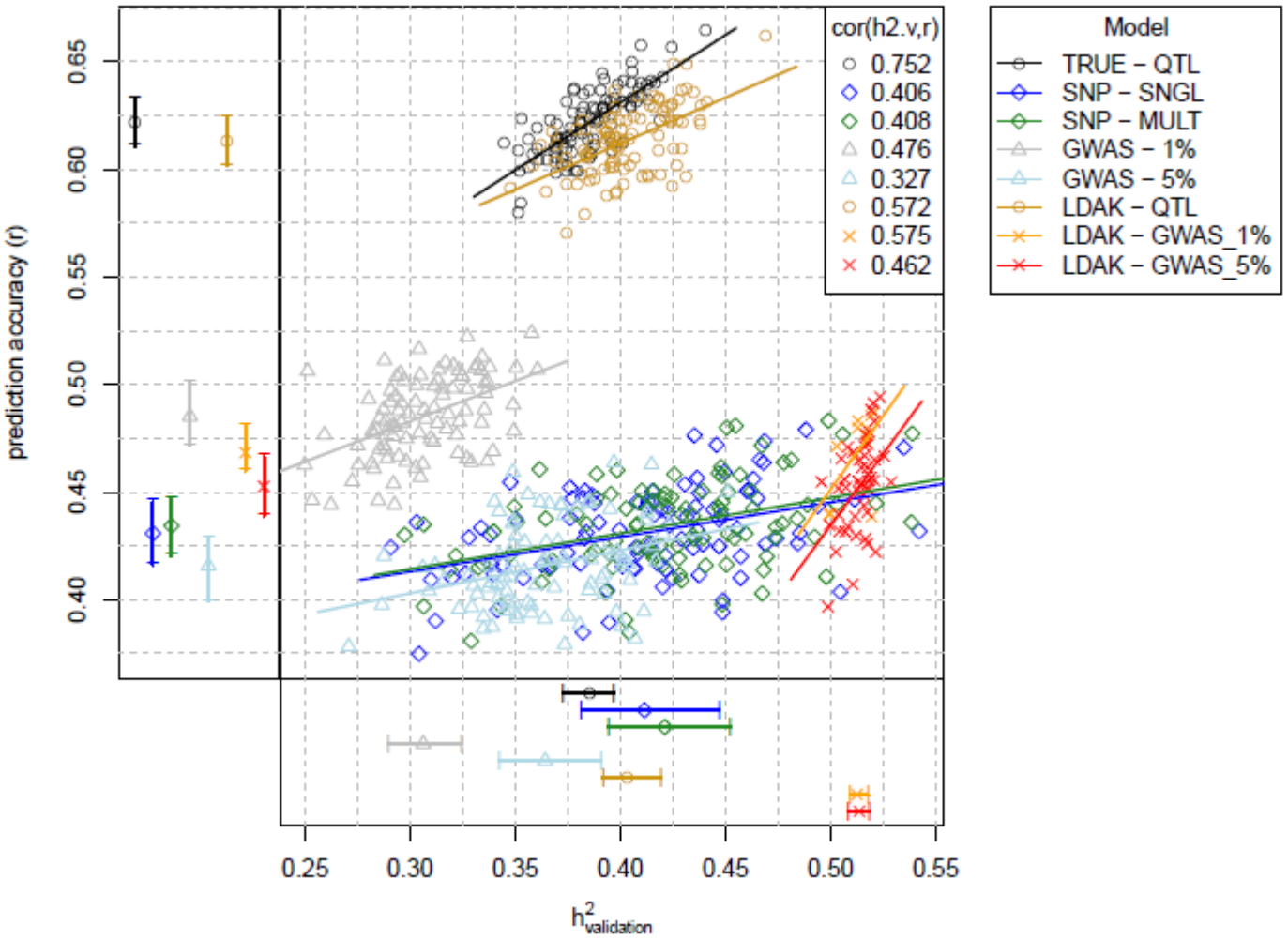
10	0.659
100	0.633
1000	0.568
10000	0.356

Heritability x predictability

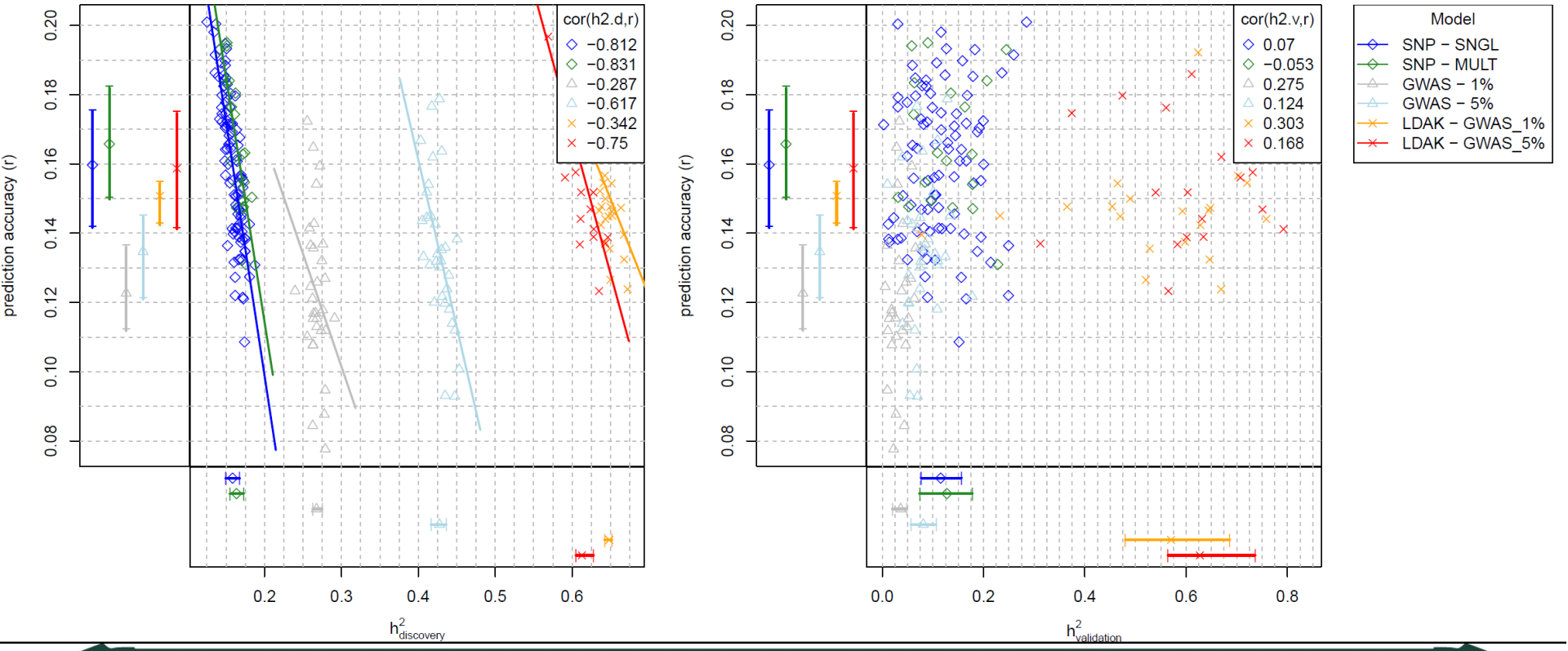




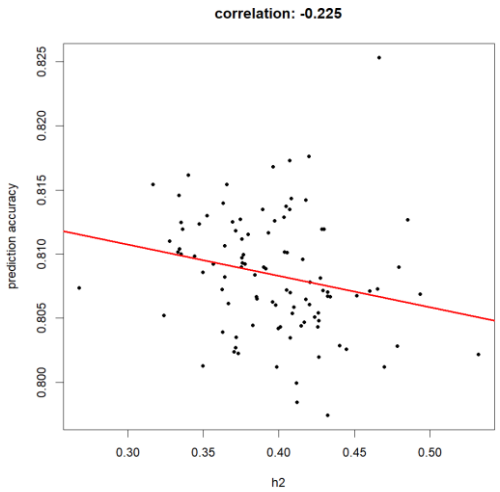
Heritability x predictability



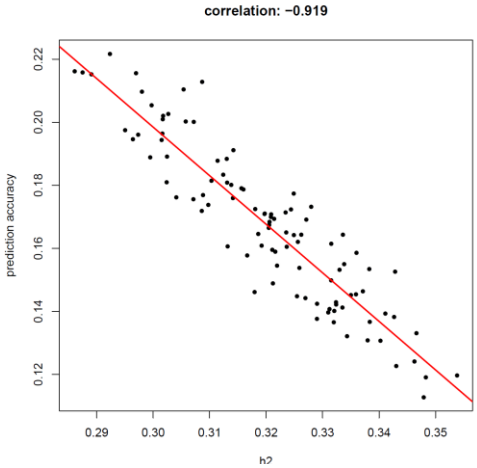
Real data


prediction accuracy (r)
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.2
0.3
0.4
0.5
0.6
$h^2_{discovery}$
cor(h2.d,r)
−0.812
−0.831
−0.287
−0.617
−0.342
−0.75
0.0
0.2
0.4
0.6
0.8
$h^2_{validation}$
cor(h2.v,r)
0.07
−0.053
0.275
0.124
0.303
0.168
Model
SNP − SNGL
SNP − MULT
GWAS − 1%
GWAS − 5%
LDAK − GWAS_1%
LDAK − GWAS_5%

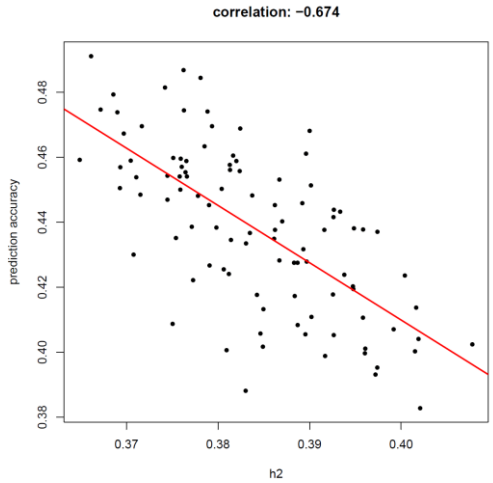
Heritability x predictability – population structures



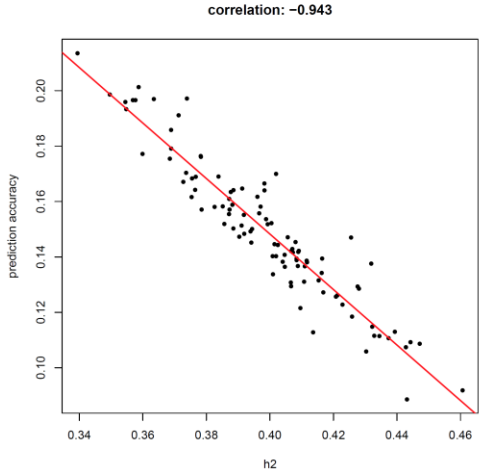
cattle



human

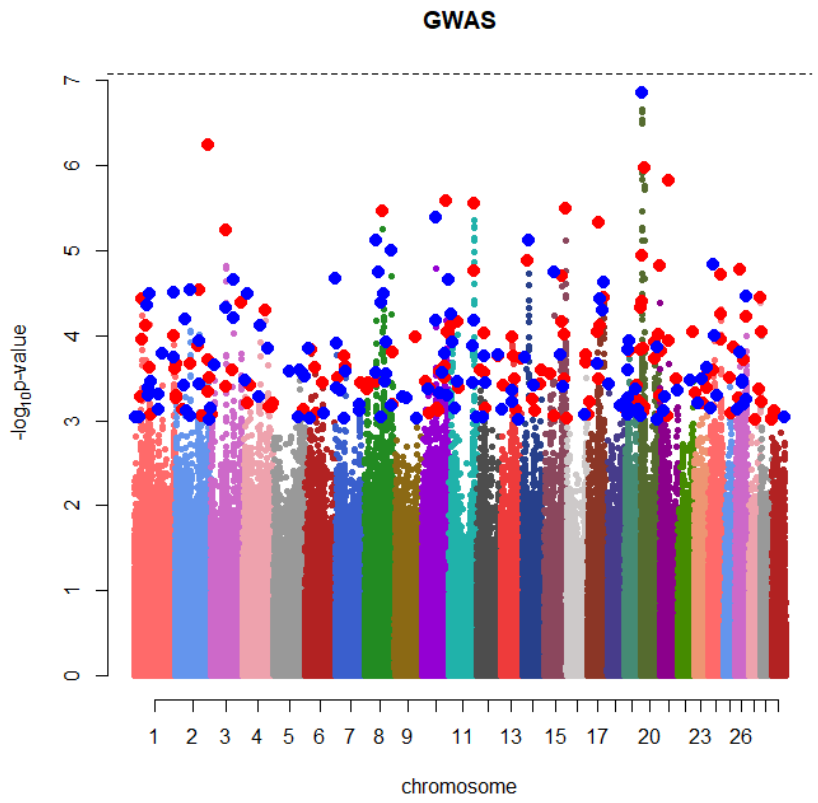


sheep



random

AI optimization – GWAS + LD + local search

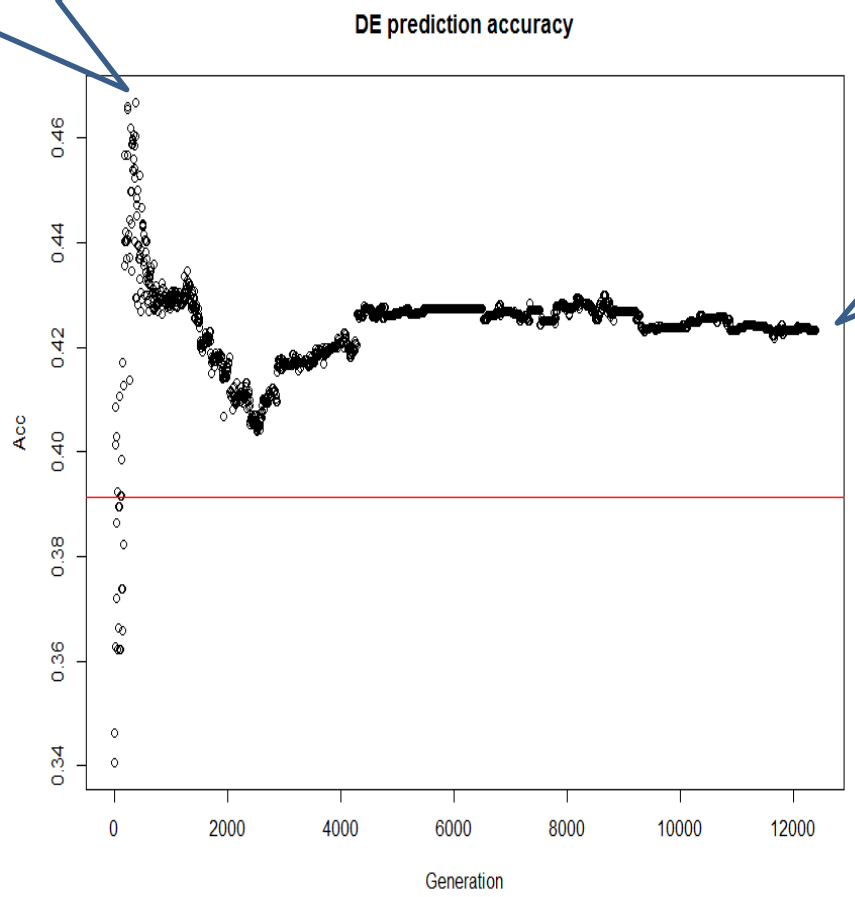


- Accuracy of prediction with 588k SNP = 0.17
- Accuracy of prediction with 218 informative SNP = 0.24

Filter and wrapper methods - sequential backward selection

Overfitting problem

Maximum Accuracy 46.7% using 515 SNP

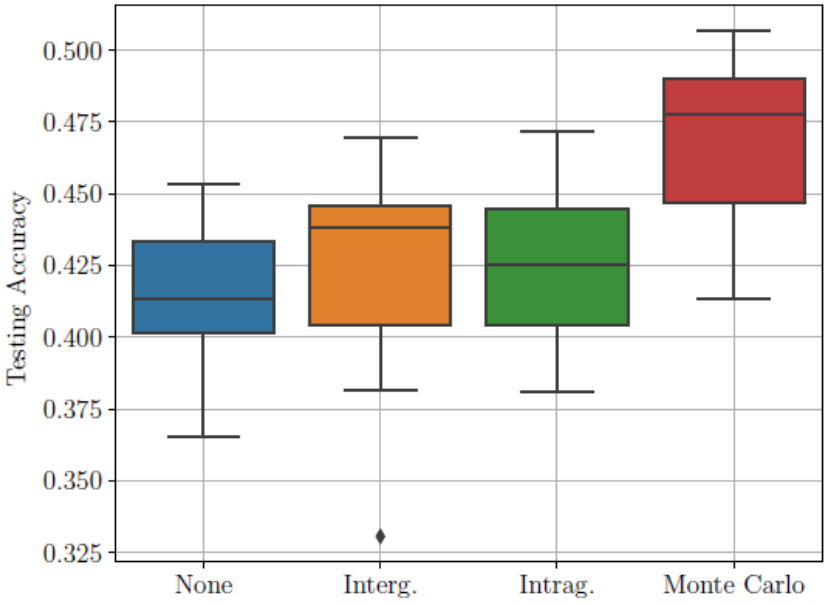


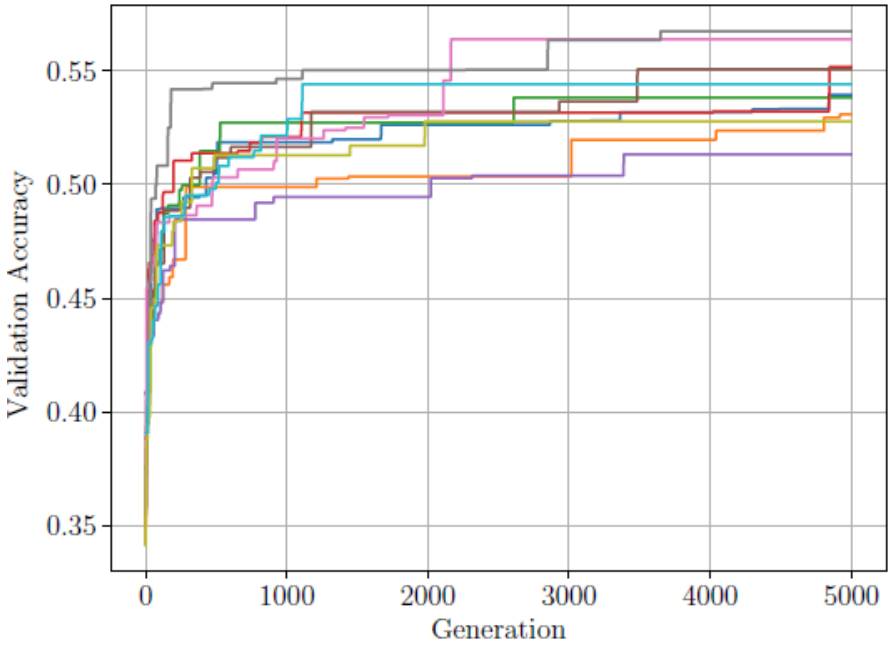
Accuracy after 12,500 generations: 42.3% using 690 SNP

2065 pigs
250 validation
1815 Training
#SNP ~ 42k
h2 ~ 0.41
max acc ~ 0.64

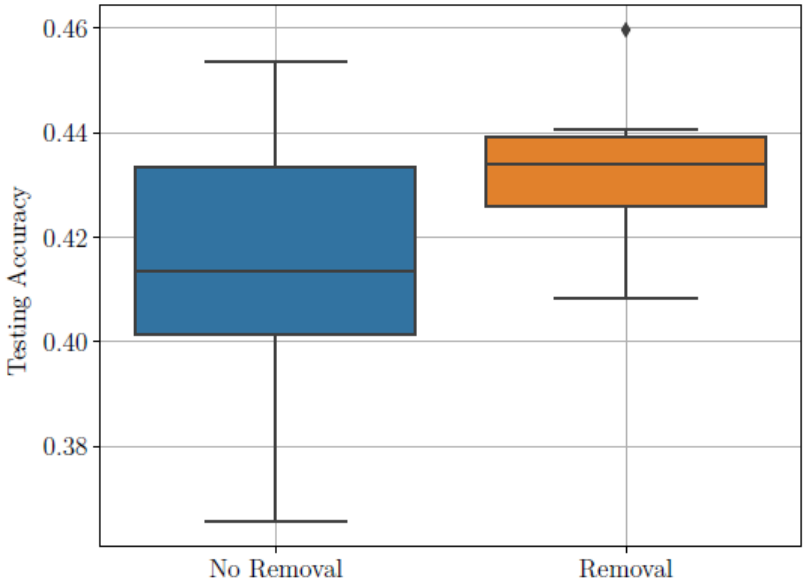
GBLUP accuracy 39.1%

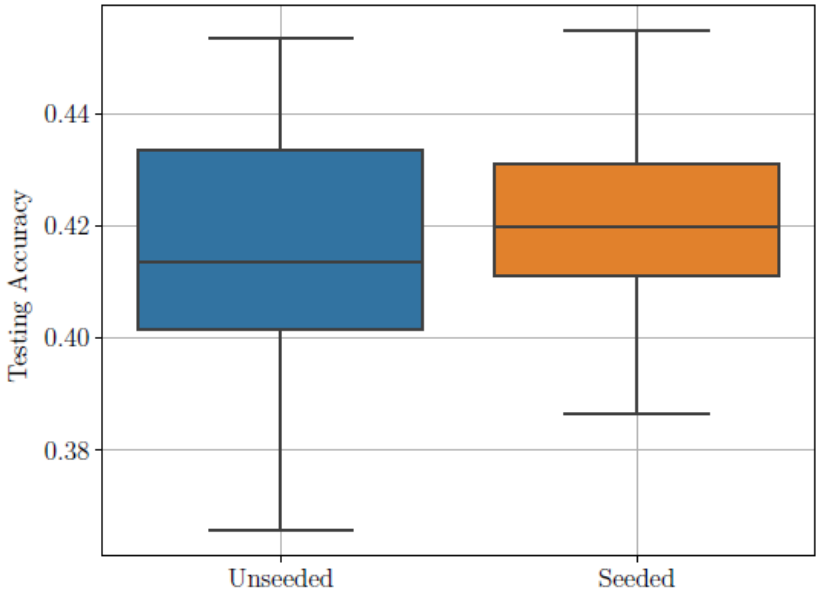
Avoid overfitting



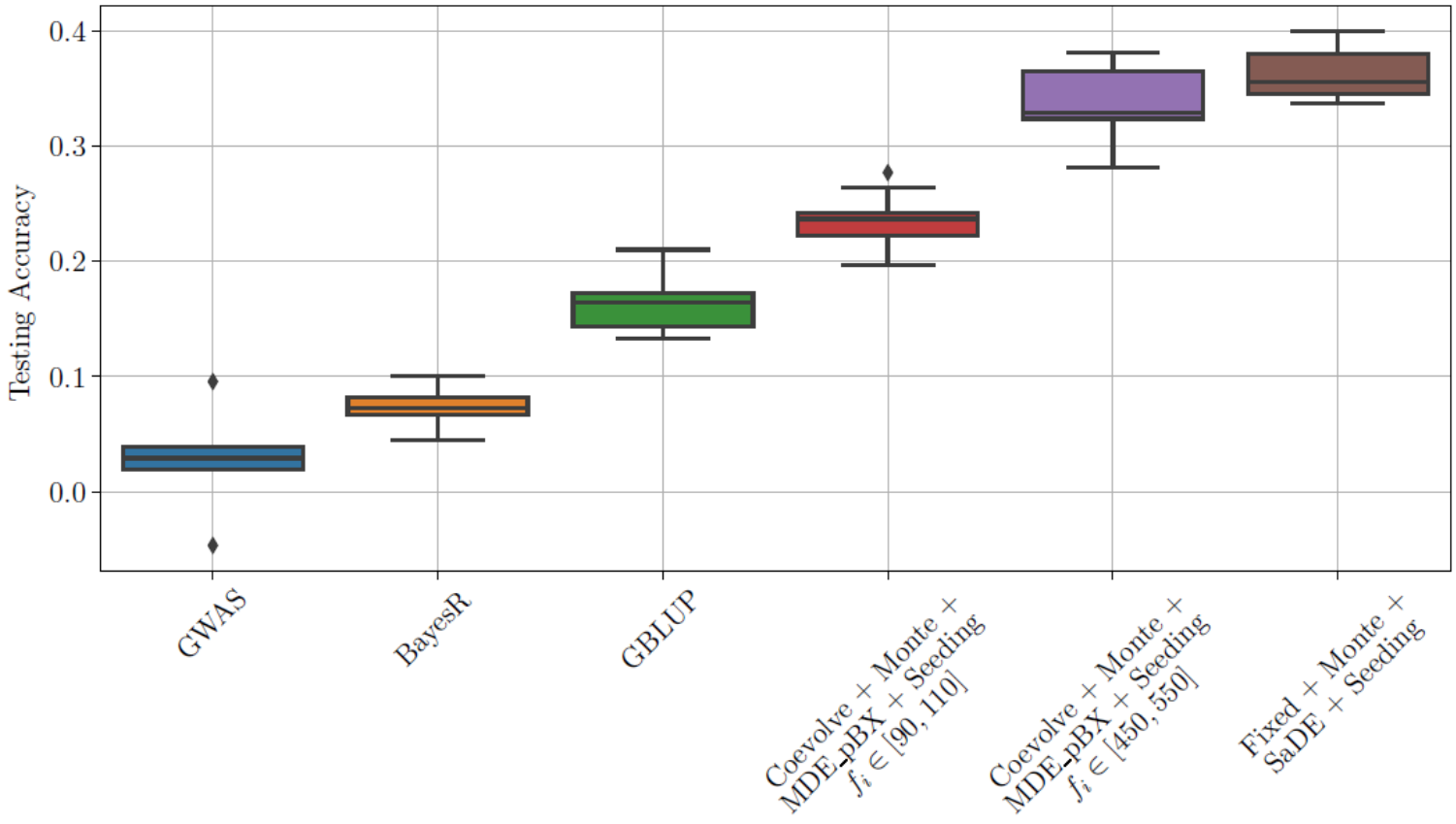


Improving prediction





Comparative performance – real data



theoretical maximum 0.41 (h2=0.17)

Summary

- Sub-setting always better but no clear winning method
- Removing noise increases accuracy
- Heritability is not a good measure of predictability within a trait
- Not even QTL knowledge is enough if effects are small and many
- bayesR better with more discrete genetic architectures
- GBLUP better with polygenic architectures

How well can we separate signal from noise?

num QTL	*10*			*100*			*1000*			*10000*		
method	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*
true positive	5	4	4	23	19	23	37	31	31	6	1	41
false positive	48	25	17	48	17	27	26	2	60	4	0	100
true negative	48483	48506	48514	48393	48424	48414	47515	47539	47481	38537	38541	38441
false negative	5	6	6	77	81	77	963	969	969	9994	9999	9959

num QTL	*10*			*100*			*1000*			*10000*		
method	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*	*GWAS*	*GBLUP*	*BAYESR*
true positive	50.0	40.0	40.0	23.0	19.0	23.0	3.7	3.1	3.1	0.1	0.0	0.4
false positive	90.6	86.2	81.0	67.6	47.2	54.0	41.3	6.1	65.9	40.0	0.0	70.9
true negative	100.0	100.0	100.0	99.8	99.8	99.8	98.0	98.0	98.0	79.4	79.4	79.4
false negative	50.0	60.0	60.0	77.0	81.0	77.0	96.3	96.9	96.9	99.9	100.0	99.6

10 QTL, h2=0.4
perfect acc = 0.659

	numSNP	h2_disc	h2_valid	accuracy	prop_h2	prop_acc	cor_G
gblup	48541	0.393	0.388	0.389	0.981	1.000	1.000
bayesr	48541	0.480	NA	**0.599**	1.212	1.541	1.000
qtl	10	0.376	0.405	0.659	1.022	1.696	0.193
top_gwas	485	0.321	0.400	0.597	1.009	1.535	0.554
top_gwas1	2427	0.338	0.479	0.494	1.210	1.271	0.848
top_gwas2	4854	0.396	0.506	0.468	1.277	1.203	0.921
top_gblup	485	0.379	0.373	0.508	0.942	1.306	0.581
top_gblup1	2427	0.523	0.475	0.426	1.198	1.096	0.864
top_gblup2	4854	0.637	0.515	0.398	1.300	1.023	0.930
top_bayes	485	0.380	0.428	0.564	1.080	1.451	0.635
top_bayes1	2427	0.459	0.504	0.469	1.273	1.207	0.871
top_bayes2	4854	0.528	0.513	0.434	1.294	1.117	0.934
random	10	0.010	0.069	0.212	0.175	0.545	0.137
random1	10	0.004	0.006	0.080	0.016	0.205	0.086
random2	10	0.004	0.008	0.108	0.019	0.278	0.121
random3	10	0.007	0.023	0.150	0.058	0.385	0.130
random4	10	0.007	0.033	0.120	0.082	0.310	0.140

100 QTL, h2=0.4
perfect acc = 0.643

	numSNP	h2_disc	h2_valid	accuracy	prop_h2	prop_acc	cor_G
gblup	48541	0.376	0.500	0.441	1.250	1.000	1.000
bayesr	48541	0.540	NA	**0.576**	1.350	1.305	1.000
qtl	100	0.391	0.392	0.633	0.980	1.434	0.357
top_gwas	485	0.303	0.324	0.535	0.812	1.213	0.536
top_gwas1	2427	0.338	0.425	0.452	1.064	1.024	0.835
top_gwas2	4854	0.399	0.466	0.446	1.166	1.011	0.917
top_gblup	485	0.378	0.340	0.499	0.852	1.130	0.576
top_gblup1	2427	0.533	0.423	0.419	1.057	0.950	0.864
top_gblup2	4854	0.641	0.488	0.416	1.222	0.944	0.930
top_bayes	485	0.381	0.352	0.543	0.881	1.231	0.627
top_bayes1	2427	0.462	0.422	0.459	1.055	1.040	0.877
top_bayes2	4854	0.512	0.455	0.448	1.137	1.015	0.935
random	100	0.046	0.100	0.245	0.251	0.555	0.370
random1	100	0.045	0.121	0.277	0.304	0.627	0.337
random2	100	0.045	0.113	0.241	0.282	0.546	0.288
random3	100	0.043	0.088	0.274	0.219	0.621	0.318
random4	100	0.051	0.089	0.272	0.223	0.617	0.358

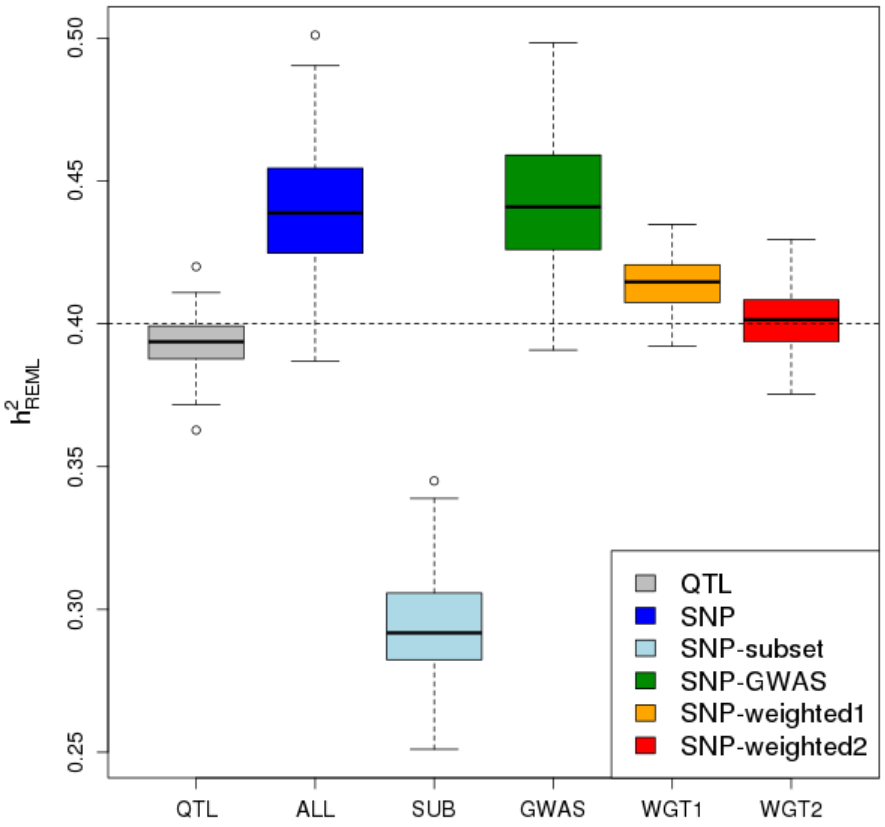
1000 QTL, h2=0.4
perfect acc = 0.634

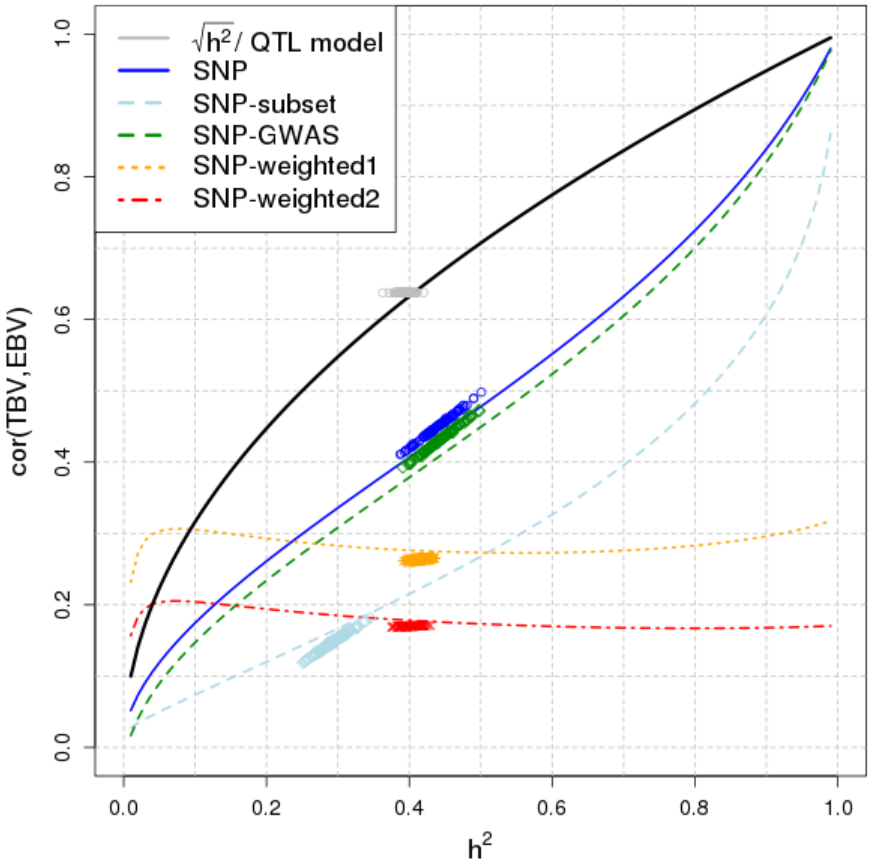
	numSNP	h2_disc		h2_valid	accuracy	prop_h2	prop_acc	cor_G
gblup	48541	0.423		0.462	0.347	1.164	1.000	1.000
bayesr	48541	0.623	NA		0.378	1.572	1.090	1.000
qtl	1000	0.427		0.459	0.568	1.158	1.637	0.774
top_gwas	485	0.269		0.309	0.460	0.778	1.324	0.575
top_gwas1	2427	0.351		0.332	0.443	0.838	1.275	0.851
top_gwas2	4854	0.405		0.408	0.412	1.030	1.188	0.920
top_gblup	485	0.384		0.323	**0.477**	0.815	1.375	0.591
top_gblup1	2427	0.541		0.425	0.419	1.072	1.207	0.866
top_gblup2	4854	0.640		0.442	0.378	1.115	1.090	0.929
top_bayes	485	0.326		0.261	0.401	0.658	1.154	0.620
top_bayes1	2427	0.481		0.383	0.406	0.967	1.169	0.871
top_bayes2	4854	0.560		0.404	0.395	1.018	1.139	0.933
random	1000	0.079		0.088	0.190	0.222	0.546	0.757
random1	1000	0.098		0.171	0.127	0.431	0.366	0.752
random2	1000	0.079		0.140	0.195	0.352	0.563	0.763
random3	1000	0.080		0.101	0.158	0.255	0.454	0.767
random4	1000	0.094		0.098	0.113	0.246	0.327	0.752

10000 QTL, h2=0.4
perfect acc = 0.629

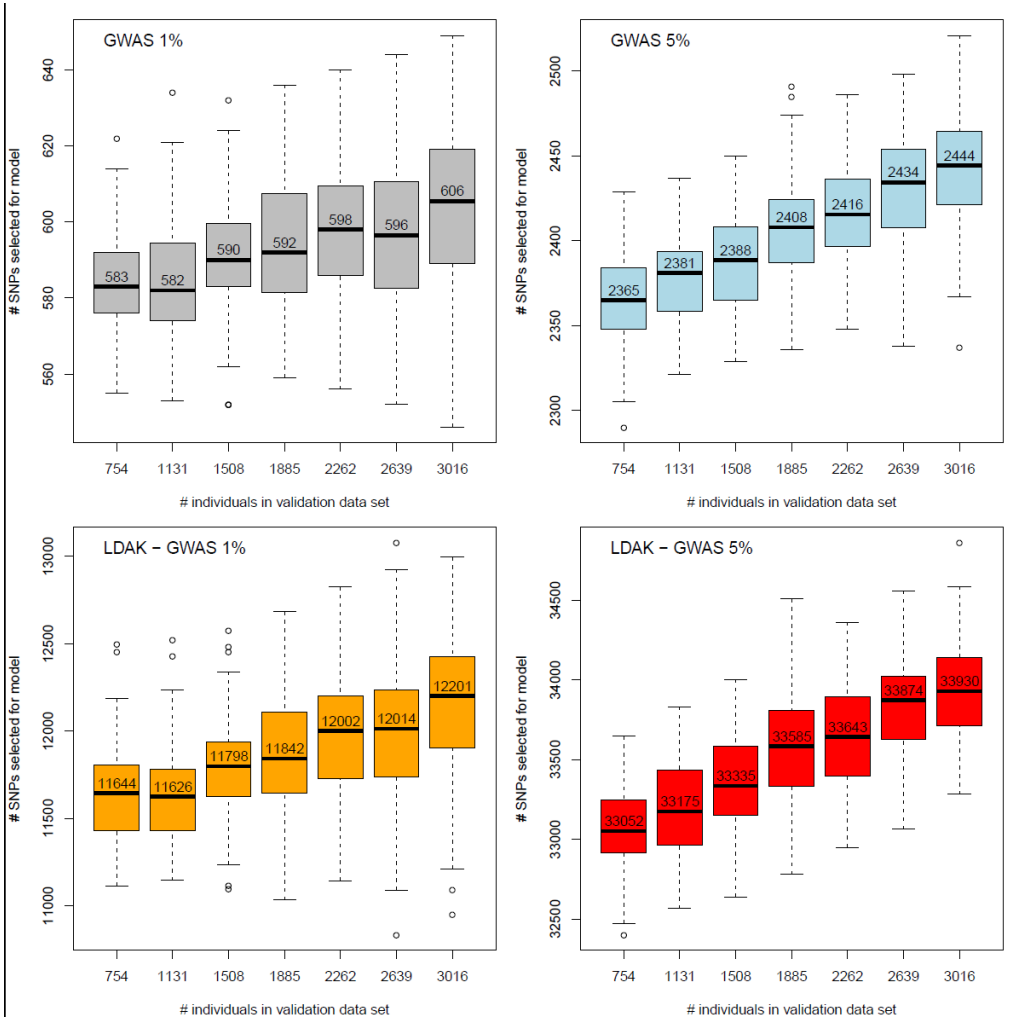
	numSNP	h2_disc		h2_valid	accuracy	prop_h2	prop_acc	cor_G
gblup	48541	0.403		0.265	0.311	0.647	1.000	1.000
bayesr	48541	0.626	NA		0.269	1.530	0.864	1.000
qtl	10000	0.388		0.301	0.356	0.736	1.144	0.973
top_gwas	485	0.182		0.167	0.252	0.408	0.811	0.571
top_gwas1	2427	0.302		0.257	0.281	0.629	0.903	0.838
top_gwas2	4854	0.364		0.279	**0.316**	0.681	1.015	0.918
top_gblup	485	0.318		0.186	0.260	0.453	0.835	0.585
top_gblup1	2427	0.529		0.261	0.289	0.637	0.928	0.862
top_gblup2	4854	0.629		0.251	0.288	0.613	0.927	0.927
top_bayes	485	0.246		0.150	0.193	0.368	0.622	0.609
top_bayes1	2427	0.448		0.249	0.237	0.608	0.762	0.873
top_bayes2	4854	0.540		0.293	0.250	0.716	0.805	0.936
random	10000	0.288		0.201	0.294	0.491	0.944	0.971
random1	10000	0.289		0.252	0.280	0.617	0.901	0.972
random2	10000	0.262		0.220	0.311	0.537	0.999	0.972
random3	10000	0.277		0.246	0.279	0.601	0.896	0.971
random4	10000	0.284		0.236	0.237	0.576	0.761	0.972

Bea's plots

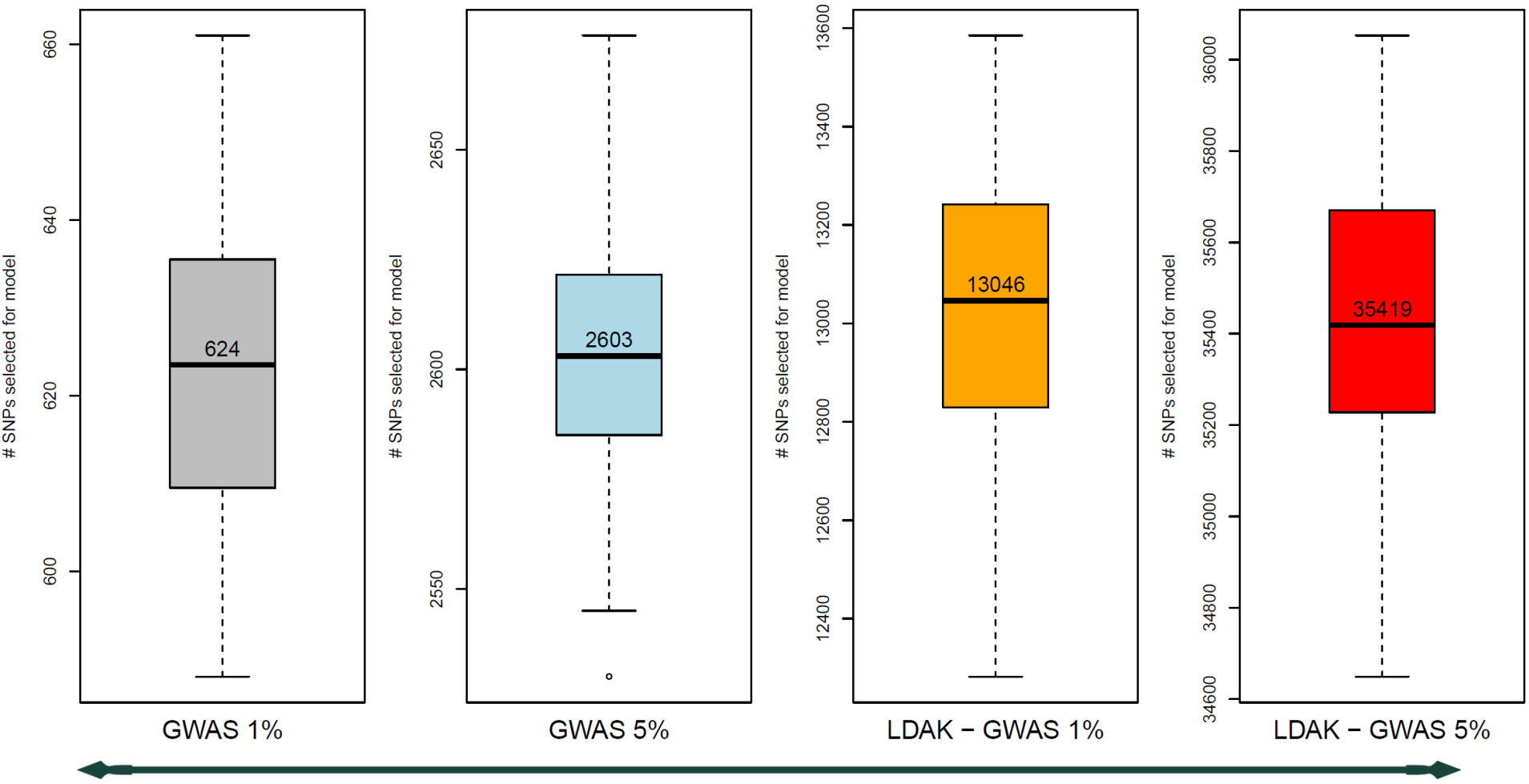

h^2_{REML}
0.25
0.30
0.35
0.40
0.45
0.50
QTL
ALL
SUB
GWAS
WGT1
WGT2
QTL
SNP
SNP-subset
SNP-GWAS
SNP-weighted1
SNP-weighted2


cor(TBV,EBV)
0.0
0.2
0.4
0.6
0.8
1.0
h^2
$\sqrt{h^2}$ / QTL model
SNP
SNP-subset
SNP-GWAS
SNP-weighted1
SNP-weighted2

Number of SNP selected – sheep 100 QTL



Number of SNP selected – sheep FEC



Boxplot prediction accuracy – sheep 100 QTL