Supplementary information and figures
LMM-Lasso: A Lasso Multi-Marker Mixed Model for Association Mapping with Population Structure Correction

Barbara Rakitsch[1,2,*], Christoph Lippert[1,2,*], Oliver Stegle[1,2,*], Karsten Borgwardt[1,2,3]

1 Max Planck Institute for Developmental Biology, Tübingen, Germany

2 Max Planck Institute for Intelligent Systems, Tübingen, Germany

3 Eberhard Karls Universität Tübingen, Germany

$\ast$ E-mail: {barbara.rakitsch, christoph.lippert, oliver.stegle}@tuebingen.mpg.de

1 Semi-empirical dataset

1.1 Simulation procedure

As basis for our simulation we used real genomic data from *Arabidopsis thaliana*. Genotype data for 1,196 plants is available from [3]. For simulating population-driven effects, we used the real phenotype leaf number at flowering time (LN, 16°C, 16 hrs daylight) which is available for 176 plants. Univariate analyses as done in [1] have shown that the phenotype has an excess of associations when population structure is not accounted for. On the other hand, after correction the p-values are approximately uniformly distributed. First, to determine the fraction of genetic and residual variance, we fit a random effects model to LN, which we subsequently used to predict the population structure for the remaining 1,120 plants. We then simulated the phenotypes as follows:

$$\mathbf{y} = \sigma_{\text{sig}}\mathbf{y}_{\text{sig}} + (1-\sigma_{\text{sig}})[\sigma_{\text{pop}}\mathbf{y}_{\text{pop}} + (1-\sigma_{\text{pop}})\boldsymbol{\psi}],$$

where $\mathbf{y}_{\text{sig}} = \mathbf{X}^{(k)}\mathbf{w}$, $\mathbf{X}^{(k)}$ is the SNP data for the k causal SNPs, $\mathbf{w} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. The first two causal SNPs are drawn such that they are in close linkage (distance between 1kb and 10kb), the remaining causal SNPs are randomly drawn from the complete genome.

The initial settings used for the simulation experiments were $\sigma_{\text{sig}} = 0.7$, $\sigma_{\text{pop}} = 0.5$ and $k = 100$. To determine the influence of the population strength, we considered $\sigma_{\text{sig}} = 0.5$, $k = 20$ and varied $\sigma_{\text{pop}} \in \{0.0, 0.3, 0.5, 0.7, 0.9, 1.0\}$. To experimentally assess the impact of the overall noise, we fixed $k = 100, \sigma_{\text{pop}} = 0.5$, and let σ_{sig} vary in $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. Finally, we considered different numbers of causal SNPs $k \in \{10, 20, 500, 100, 300, 1000\}$ and fixed $\sigma_{\text{sig}} = 0.7, \sigma_{\text{pop}} = 0.5$. For the linkage experiments, we used the $\sigma_{\text{sig}} = 0.7$, $\sigma_{\text{pop}} = 0.5$ and $k = 10$. We simulated 30 phenotypes for all settings.

1.2 Evaluation

Since in many cases the causal loci might not be genotyped and stochastic effects might cause larger correlations between strongly correlated SNPs and the phenotype, we consider a SNP called positive by a model as a True Positive if it is in close proximity to the known causal SNP (+/- 10kb). On the other hand, if a SNP called by a model is not close to a causal SNP, it is considered a False Positive.

A fair comparison between the univariate and multivariate methods is difficult as the univariate methods select blocks of linked markers, whereas the multivariate methods select only one representative marker per block. To account for this principle difference, we employed a post-processing procedure to sparsity the solutions of all methods in a comparable manner. For this purpose, we iteratively selected the most associated marker genome-wide. To ensure that the next-based marker is not in LD to this SNP, we ignored neighboring markers (+/- 10kb) and proceeded with selection the next SNP. This process was repeated until the no marker above the threshold was left.

2 Relationship to Stepwise Regression

The difference between the Lasso and Forward Selection can be easiest seen by going over Forward Stagewise Linear Regression. In Forward Selection, we start with the SNP having the largest effect size. We then iteratively add SNPs that can explain most of the phenotype conditioned on the SNPs that have already been selected. In Forward Stagewise Linear Regression instead, one moves only a small step in the direction of the most correlated SNP and then re-estimates the most correlated SNP on the remaining phenotype which is far less greedy. In [2], it is shown that there is a close relationship between Forward Stagewise Linear Regression and Lasso resulting in nearly identical solutions.

3 Multivariate models better differentiate multiple causal loci from correlation due to linkage

Previously, step-wise regression models that include genetic variants in the order of effect sizes have been considered to differentiate between true genetic heterozygosity and local correlation due to linkage [4]. Here, we show that LMM-Lasso can be successfully applied for the same task, however with the additional benefit that a step-wise order of including genetic markers as co-factors is not needed (Figure 3). The comparison includes true genetic heterogeneity where two loci within linkage disequilibrium (LD) jointly regulate the phenotype (left) as well as a single genetic effect that is broadened by LD (right). The LMM-Lasso model and Lasso are able to differentiate between the two types of genetic architectures reliably, whereas univariate models suffer from correlation due to linkage.

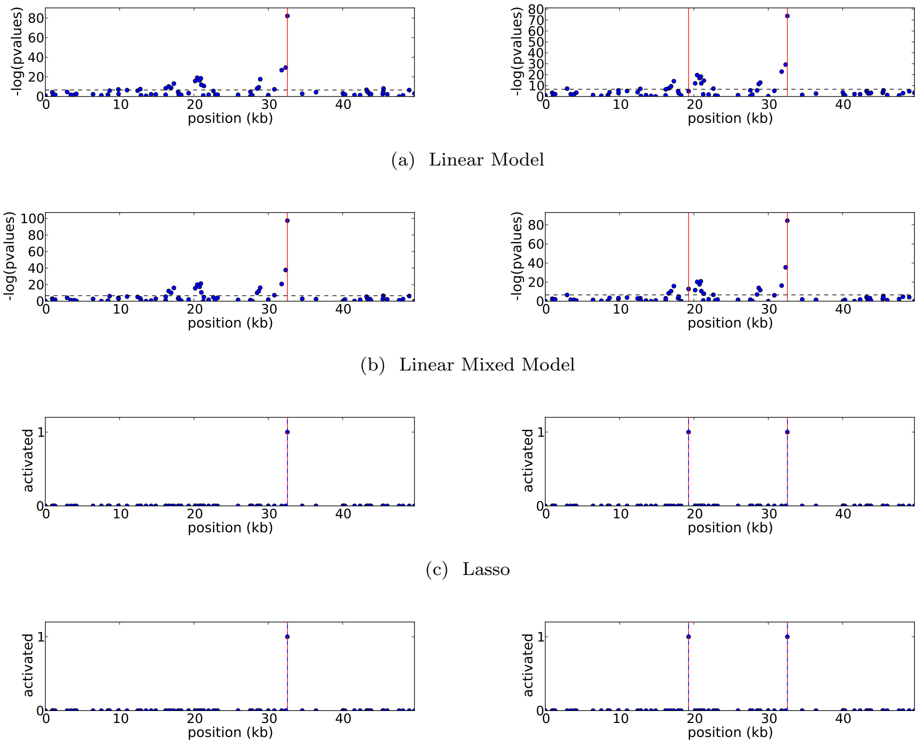
References

1. Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, Rong Jiang, N Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R Ecker, Nathalie Faure, Joel M Kniskern, Jonathan D G Jones, Todd Michael, Adnane Nemri, Fabrice Roux, David E Salt, Chunlao Tang, Marco Todesco, M Brian Traw, Detlef Weigel, Paul Marjoram, Justin O Borevitz, Joy Bergelson, and Magnus Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, pages 1–5, 2010.

2. Bradley Efron, Trevor Hastie, Lain Johnstone, and Robert Tibshirani. Least angle regression. *Ann Stat*, 32:407–499, 2004.

3. M. W. Horton, A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjalmsson, M. Nordborg, J. O. Borevitz, and J. Bergelson. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.*, 44(2):212–216, Feb 2012.

4. J. Yang, T. Ferreira, A.P. Morris, S.E. Medland, P.A.F. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.N. Weedon, R.J. Loos, et al. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–375, 2012.



(a) Linear Model

(b) Linear Mixed Model

(c) Lasso

(d) LMM-Lasso

Figure 1. Differentiation between multiple causal loci from spurious correlation due to linkage on simulated data. Left: A single SNP with a strong effect in an LD block. **Right:** Same as before, however with an additional SNP with weaker effect size in the opposite direction. While all methods detect the SNP with large effect size, the second one is only recovered by the multivariate methods. The red lines indicate the causal SNPs, the blue dots the score assigned from the methods to the SNPs.

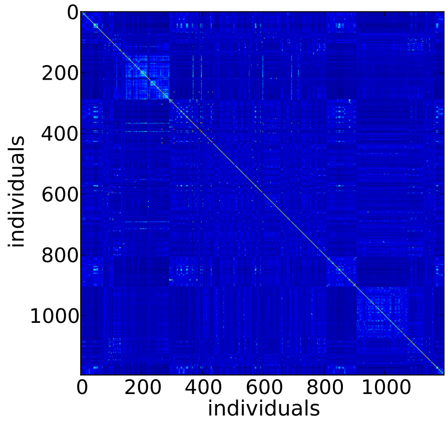
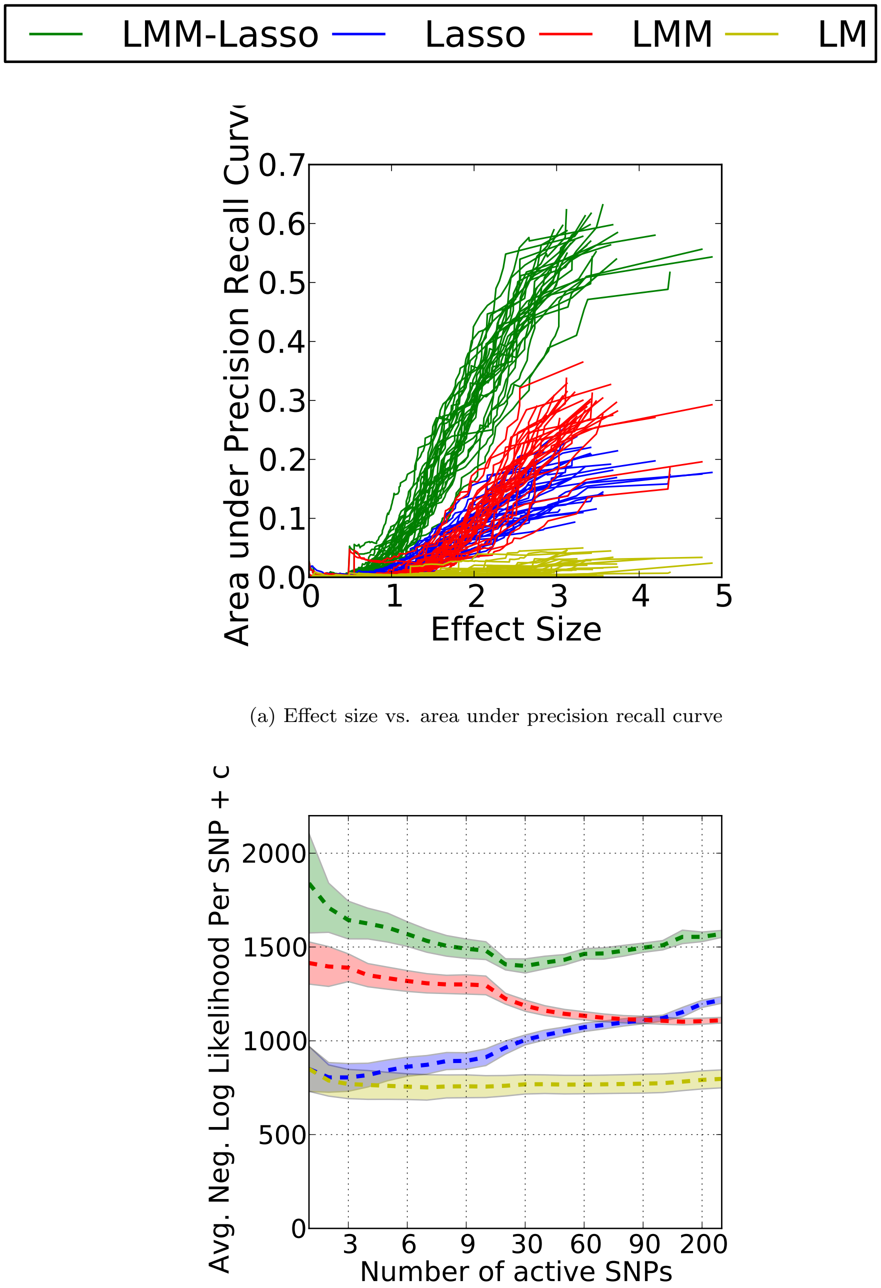


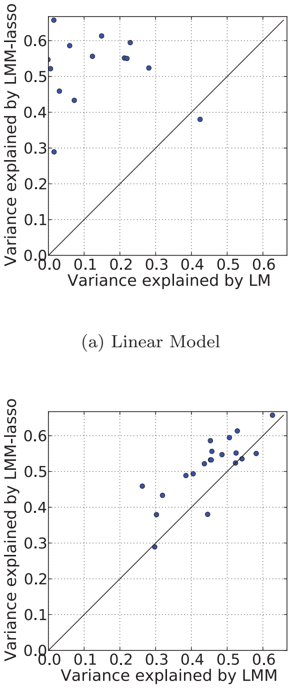
Figure 2. Realized Relationship Matrix from the 1196 plants of *Arabidopsis thaliana*, available from [3]. The relatedness between the individuals is complex and very strong as the matrix is deeply structured.



(a) Effect size vs. area under precision recall curve

(b) Averaged neg. log-likelihood vs. number of active SNPs

Figure 3. Evaluation of alternative methods on semi-empirical GWAS dataset. a) Area under the precision recall curve as a function of the total effect size of all causal SNPs. b) Average negative log-likelihood of each selected SNPs under the multivariate normal distribution $\mathcal{N}(0, \mathbf{K})$ as a function of the number of SNPs that are active in the model. The smaller the log likelihood is, the more the SNPs are correlated with the population structure. For the LMM-Lasso and the Lasso active SNPs have been selected by following the regularization path. For linear mixed model (LMM) and linear model (LM), the set of active SNPs have been obtained in ascending order of the p-value obtained. In the beginning, Lasso and the linear model choose SNPs that heavily reflect the population structure, while the mixed model approaches don't. In both figures the number of causal SNPs was 100.



(a) Linear Model

(b) Linear Mixed Model

Figure 4. Predictive power and sparsity of the proposed models for quantitate traits in *Arabidopsis thaliana*. Maximal explained variance on an independent test set a) LMM-Lasso vs. linear model including the top associated SNP. b) LMM-Lasso vs. linear mixed model including the top associated SNP.

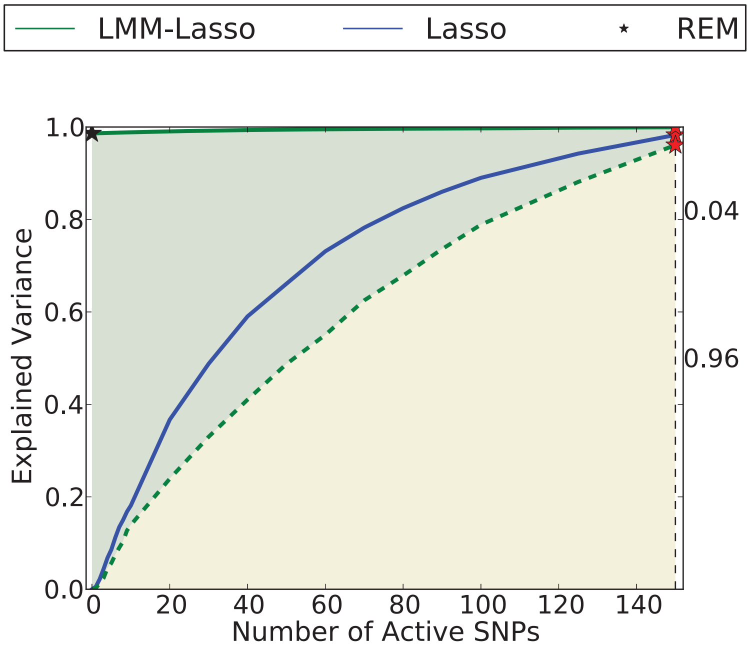


Figure 5. Variance dissection into individual SNP effects and global genetic background driven by population structure. Shown is the explained variance on the training set as a function of the number of active SNPs for the flowering phenotype (10°) in *Arabidopsis thaliana*. In blue, the predict the explained variance of the Lasso as a function of the number of SNPs in the model. In green, the total predictive variance of LMM-Lasso for different sparsity levels. The shaded area indicates the fraction of variance LMM-Lasso explains by means of individual SNPs (yellow) and population structure (green). LMM-Lasso without additional SNPs in the model corresponds to a genetic random effect model as in common usage (black star).

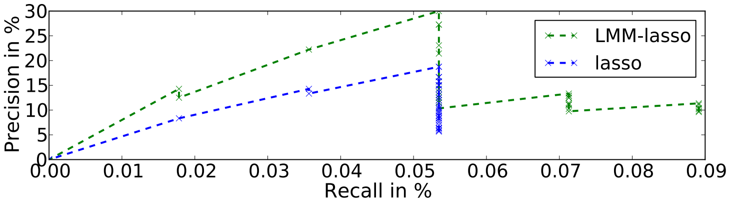


Figure 6. Evaluation of the Lasso Methods for FLC gene expression in *Arabidopsis thaliana.* Precision-Recall Curve for recovering SNPs in proximity to known candidate genes using alternative methods. Shown is precision (TP/(TP+FP)) as a a function of the recall (TP/(TP+FN)). Each point in the plot corresponds to a specific selection threshold.

Table 1. SNPs close to candidate genes detected by the LMM-Lasso

Phenotype	Chrom.	Pos.	GeneID	Dist(Gene)	P_{sel}(LMM-Lasso)	P_{sel}(Lasso)	$-\log p$(LMM)	$-\log p$(LM)
LD	4	454542	AT4G01060	5930	**0.90**	**0.94**	6.41	**10.74**
LD	4	466307	AT4G01060	5222	**0.80**	0.00	3.37	2.55
LD	4	466800	AT4G01060	5715	**0.59**	0.20	4.20	6.40
LD	5	3169764	AT5G10140	3734	**0.60**	0.00	4.74	3.83
LD	5	3188328	AT5G10140	8879	**0.71**	**0.96**	4.77	**9.03**
LDV	1	24341937	AT1G65480	4340	**0.72**	0.01	3.03	4.24
LDV	2	9965680	AT2G23380	0	**0.89**	0.00	3.92	1.16
LDV	2	10579146	AT2G24790	3922	**0.54**	0.08	3.34	6.49
LDV	3	21241923	AT3G57390	2755	**0.57**	0.13	3.28	5.38
LDV	5	5186340	AT5G15850	8440	**0.66**	**1.00**	4.34	**11.02**
SD	1	24341937	AT1G65480	4340	**0.71**	0.00	2.65	3.55
SD	3	17951698	AT3G48430	0	**0.54**	0.00	3.21	4.07
SD	4	458226	AT4G01060	2246	**0.90**	0.02	5.11	3.35
SDV	1	24341937	AT1G65480	4340	**0.84**	0.06	2.79	4.68
SDV	3	1409077	AT3G05040	874	**0.62**	0.00	3.48	2.74
SDV	4	10994321	AT4G20370	6461	**0.79**	0.00	3.92	2.42
SDV	5	1164843	AT5G04240	4702	**0.57**	0.16	3.74	3.57
SDV	5	26794176	AT5G67100	44	**0.77**	0.11	5.35	5.40
FT10	1	24341937	AT1G65480	4340	**1.00**	0.00	5.02	3.69
FT16	1	24341937	AT1G65480	4340	**0.54**	0.00	2.40	1.94
FT16	4	5735680	AT4G08920	8430	**0.53**	0.00	2.85	3.92
FT16	5	3188328	AT5G10140	8879	**0.95**	**1.00**	4.52	**12.34**
FT22	1	24341937	AT1G65480	4340	**0.55**	0.00	2.33	1.76
FT22	5	3188328	AT5G10140	8879	**0.55**	**1.00**	4.09	**10.96**
2W	4	454542	AT4G01060	5930	**0.95**	**0.90**	**6.03**	**8.29**
2W	4	460246	AT4G01060	226	**0.51**	0.00	4.26	1.89
2W	5	3188328	AT5G10140	8879	**0.67**	**0.89**	3.74	**9.11**
8W	4	9479396	AT4G16845	0	**0.71**	0.01	3.05	3.00
8W	5	5186340	AT5G15850	8440	**0.81**	**0.94**	4.48	**7.97**
FLC	4	205170	AT4G00450	0	**0.88**	0.18	5.01	**6.88**
FLC	4	210657	AT4G00450	0	**0.56**	0.00	4.78	5.40
FLC	4	454542	AT4G01060	5930	**0.85**	**0.92**	5.03	**9.49**
FLC	4	1115702	AT4G02560	7788	**0.70**	0.00	4.10	2.32
FLC	5	15897391	AT5G39660	0	**0.86**	0.00	4.32	3.51
FRI	4	268809	AT4G00650	217	**1.00**	**1.00**	**17.45**	**20.91**
FRI	4	268990	AT4G00650	36	**1.00**	**1.00**	**13.65**	**15.13**
FRI	4	276143	AT4G00650	4640	**0.85**	**0.96**	**14.37**	**17.36**
8WGHFT	1	24341937	AT1G65480	4340	**0.53**	0.00	2.65	3.16
8WGHFT	1	25508306	AT1G68050	4033	**0.82**	0.02	4.12	4.68
8WGHFT	2	10579146	AT2G24790	3922	**0.54**	0.35	3.72	**6.76**
8WGHFT	4	5735680	AT4G08920	8430	**0.93**	**0.56**	4.13	5.14
8WGHLN	4	269962	AT4G00650	0	**0.59**	0.00	4.47	1.53
0WGHFT	2	19026329	AT2G46340	2917	**0.80**	0.07	3.87	5.34
0WGHFT	4	269962	AT4G00650	0	**0.89**	0.00	6.31	4.78
0WGHFT	4	454542	AT4G01060	5930	**0.86**	**0.74**	**6.34**	**10.08**
0WGHFT	4	1250359	AT4G02780	5546	**0.52**	0.00	3.62	5.85
FTField	1	23372416	AT1G63030	341	**0.62**	0.00	3.54	3.43
FTField	4	196902	AT4G00450	6569	**0.81**	0.00	3.23	2.85
FTField	4	9211041	AT4G16280	0	**0.54**	0.00	3.17	3.01
FTField	5	8361420	AT5G24470	2544	**0.57**	**0.97**	4.03	**8.25**
FTDiameterField	5	26809133	AT5G67100	6803	**0.53**	**1.00**	4.33	**11.10**
FTGH	1	24341923	AT1G65480	4326	**0.59**	0.18	3.26	5.72
LN10	1	21416665	AT1G57820	1170	**0.87**	0.04	4.03	3.57
LN10	1	24341937	AT1G65480	4340	**0.96**	0.07	4.25	4.23
LN10	4	9479350	AT4G16845	0	**0.62**	0.00	2.71	2.04
LN16	4	280774	AT4G00650	9271	**0.51**	0.00	4.27	5.97
LN16	4	454542	AT4G01060	5930	**0.73**	**0.95**	5.73	**11.25**
LN22	2	19027744	AT2G46340	1502	**0.63**	0.00	3.03	2.54
LN22	4	454542	AT4G01060	5930	**0.55**	**0.84**	5.91	**10.52**
LN22	4	1232950	AT4G02780	4817	**0.56**	0.00	3.97	3.22
LN22	5	3178615	AT5G10140	0	**0.66**	0.33	4.68	**7.52**

List of all SNPs that are in close proximity to candidate genes for all phenotypes related to flowering time of *Arabidopsis thaliana*. We report the $-\log_{10}$ transformed p-values for the univariate methods and the frequency with which the SNP is subsampled by the multivariate methods.

Table 2. Candidate genes containing multiple associations

Phenotype	Chromosome	Position	GeneID	LM	LMM
LD	4	(466307,466800)	AT4G01060	(2.55,6.40)	(3.37,4.20)
2W	4	(454542,460246)	AT4G01060	(8.29,1.89)	(6.03,4.26)
FLC	4	(205170,210657)	AT4G00450	(6.88,5.40)	(5.01,4.78)
FRI	4	(268809,268990)	AT4G00650	(20.91,15.13)	(17.45,13.65)
FRI	4	(268990,276143)	AT4G00650	(15.13,17.36)	(13.65,14.37)

List of all candidate genes that have two activated SNPs in close proximity for all phenotype related to flowering time of *Arabidopsis thaliana.* The last two columns show the $-\log_{10}$ transformed p-values for the linear and the linear mixed model.