Advances in statistical methods to handle large data sets for GWAS in crop breeding

Boby Mathew[1], Mikko J. Sillanpää[2] and Jens Léon[1]

[1]Institute of Crop Science and Resource Conservation,
University of Bonn, 53115, Bonn, Germany

[2]Department of Mathematical Sciences and Biocenter Oulu,
University of Oulu, FIN-90014, Finland

June 4, 2018

Introduction

Mathew et al. (2018b) Quantitative trait loci (QTL) analysis is a well known statistical exercise in biological research to identify genetic loci associated with a quantitative trait/phenotype of interest. QTL mapping studies utilize molecular markers to locate the genomic regions that affect the phenotype. Two of the most commonly used QTL mapping approaches are linkage analysis (LA)(also known as family-based linkage mapping or QTL mapping) and association mapping (linkage disequilibrium (LD) mapping). Linkage mapping considers the linkage disequilibrium that exits with in families in order to map the region, whereas the association mapping needs markers that are in LD with a potential QTL across the entire population. Association mapping is based on the assumption that the alleles which influence on the trait are inherited from a single common ancestor in the past. Table 1 summarizes a comparison between association and linkage mapping.

Even though, association and linkage mapping are viewed as fundamentally different approaches, both methods tries to make use of the recombination events. Over the decades

Table 1 A comparison between association and linkage mapping

Property of the mapping method	Linkage mapping	Association Mapping
Mapping populations	Close relatives	Unrelated or related individuals
Marker density	Moderate marker density	High marker density
Mapping resolution	Long (< 5 MB)	Short(< 1 Mb)
Susceptible to population stratification	No	Yes
Biological basis	Recent recombination	Historical recombination
Suitable phenotypes	Rare phenotypes	Common phenotypes
Parameter of interest	Recombination fraction	Statistical association
Controlled experiment	Yes	No

many LA studies (i.e., QTL mapping in offspring population resulting from a simple line crossing experiments) have reported hundreds of QTLs in various plant species and only a few identified QTLs were targeted at gene level (Patterson et al. (2006)). Recent advances in low cost high throughput DNA sequencing technologies have helped genome-wide association mapping (GWAS) to emerge as an alternative to linkage mapping and which offers high mapping resolution and is more time-efficient. However, before starting, one should make sure that in order to fully utilize all the potential available in GWAS, the association mapping should be optimally performed in multiparental populations with enough number of generations (to accumulate enough recombination events). In this chapter we shortly discuss some of the main challenges for GWAS studies with large datasets.

Single-locus association model

Despite the availability of large number of single-nucleotide polymorphisms (SNPs), standard GWAS analysis methods consider one SNP at a time and identify the marker-trait association using a single-locus model. Single-locus model is the simplest and most commonly used model to identify associations between SNPs and continuous trait of interest. However, it is already known that hidden population structure due to LD and of sample structure (cryptic relatedness) leads to inflated test statistics and that may lead to false positive and negative associations between marker and trait. Plenty of correction methods have been proposed especially for single marker association testing where a phenotypic relevance of a single putative gene position is tested at a time in isolation of other putative loci (*e.g*, Principal component analysis, (Price et al. 2006); Mixed-model approach, (Kang et al. 2008a, Müller et al. 2011, Yu et al. 2006); Structured association, (Pritchard et al. 2000)), for a review see (Sillanpää 2011). Mixed model method including a random polygenic term, which describes relationships between individuals, is performing well and is most widely used method in plant, animal and human datasets. It is now generally accepted that it can correct confounded (spurious) associations due to both: close relatives and population structure in the dataset. However, its general drawback is that it may loose statistical power (by over-correcting the structure) or it may lead to wrong findings if the candidate SNP is included in the calculation of genomic relationship matrix (Würschum and Kraft (2015)). A single-locus mixed model with the polygenic random effect can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{v} + \mathbf{Z}\mathbf{g} + \boldsymbol{\epsilon}. \tag{1}$$

Here $\mathbf{Y}$=$\{Y_i\}_{i=1}^n$ is a vector of phenotype values for n lines and $\boldsymbol{\beta}$ is the vector of fixed effects with known incidence matrix $\mathbf{X}$, whereas, $\mathbf{W}$ is the incidence matrix for the marker being tested for the association. Moreover $\mathbf{g}$ is an $n \times 1$ vector of polygenic effects with the incidence matrix $\mathbf{Z}$ and $\mathbf{g} \sim N(\mathbf{0}, \mathbf{K}\sigma_a^2)$. Here, $\mathbf{K}$ defines the covariance structure that describes the relatedness among individuals and can be calculated either from the marker information or with the pedigree. Additionally, $\boldsymbol{\epsilon}$ corresponds to the vector of residual, following a normal distribution as $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_e^2)$. In single-locus model, the marker association is tested for one marker at a time with the null hypothesis that is $v = 0$ against the alternative hypothesis is that $v \neq 0$. Different mixed-model methods are mainly differing in how they implement the required speed up of the computation (refitting the model and estimating the polygenic variance separately for each candidate SNP) which makes it applicable for large genome-wide datasets.

Some of the interesting packages based on single-locus association model are: PLINK (Purcell et al. 2007); TASSEL (Bradbury et al. 2007); CMLM (Zhang et al. 2010); ECMLM (Li et al. 2014); FaST-LMM (Lippert et al. 2011); EMMA (Kang et al. 2008b); GEMMA (Zhou and Stephens (2012)); GRAMMAR-Gamma (Svishcheva et al. 2012); rrBLUP

(Endelman (2011)).

Multilocus association model

Single-locus model test a single SNP at a time and known to have some drawbacks. Firstly, it is hard to locate the region contain the QTL in to a small region because a number of SNPs can be in LD with the QTL, in this case the significant SNP can span a wide range on the chromosome (Pryce et al. (2010)). Secondly, effect of a single SNP may be quite small, but may have strong joint effects and by considering all SNPs simultaneously will improve the power to detect their joint activity. One solution to these problems is to jointly fit all SNPs using a multilocus association model. Another interesting benefit of multilocus model is their capability of automatically correcting/controlling the confounding due to population structure/relatedness without having polygenic term in the model (for example, Pikkuhookana and Sillanpää (2009), Würschum and Kraft (2015), Kärkkäinen and Sillanpää (2012)). Moreover, they are also relatively robust for model misspecification. The basic multilocus association model can be defined as:

$$\mathbf{Y} = \mu + \sum_{j=1}^{m} \mathbf{M_{.j}} \mathbf{b_j} + \boldsymbol{\epsilon}. \tag{2}$$

Here $\mathbf{Y}$=$\{Y_i\}_{i=1}^{n}$ is a vector of phenotype values of n lines, m is the total number of markers, $\mathbf{M}_{ij}$ (note that $\mathbf{X}$ is commonly used notation for genotypes, however to avoid the confusion with the fixed effect in linear mixed model we use the notation $\mathbf{M}$ here) is the genotypic value of individual i at marker j coded as 0, 1, 2 for the genotype AA, Aa, aa respectively, $\mathbf{b_j}$ (note that $\boldsymbol{\beta}$ is the commonly used notation for marker effect and to avoid the confusion with the fixed effect term in Eq. 1 we use the notation $\mathbf{b}$) is the random marker effect associated with marker j, and $\boldsymbol{\epsilon}$ corresponds to the residual, following a normal distribution as $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_e^2)$. With multilocus model regression methods are generally used to estimate marker effects ($\mathbf{b}$).

Some of the interesting packages based on multilocus association model are: MLMM (Segura et al. 2012); FASTmrEMMA (Wen et al. 2017); MRMLM (Wang et al. 2016;

E-Bayes (Xu 2007).

High dimensional data space in GWAS

Nowadays GWAS studies involves thousands of SNPs owes much to the recent advances in genotyping technologies. This availability of high-throughput genomic data leads to the ‘large p, small n’ (here p corresponds to the number of marker effects and n is for the number of samples) problem in GWAS. The so called ‘large p, small n’ occur when the number of parameters to be estimated (marker effects $\mathbf{b}$ in model 2) is much larger than the samples (phenotype values) and biologists have to deal with large data space. Additionally, strong LD among SNPs poses additional challenges in GWAS studies. Multilocus models are more prone to ‘large p, small n’ problem because joint analysis of all SNPs together is computationally challenging, whereas the one-dimensional genome scan by testing a single SNP at a time can handle a large number of markers without much problems. Two of the most commonly used methods to deal with large data space are regularization and variable selection methods.

Variable selection and shrinkage/regularization

Identifying the relevant predictive variables is a fundamental problem in statistical learning. Most standard regression methods may fail in such cases where the number of markers is much larger than the sample size. Variable selection and shrinkage (regularization) methods are commonly applied to such problems to select the best subset of predictors. Stepwise procedures (forward selection and backward elimination) are commonly used for variable selection in ‘large p, small n’ problems. Backward-forward variable selection methods are only applicable with a couple of tens of markers and become quickly impractical as the number of predictors increases. The single-locus model can also be extended to the multilocus framework by applying stepwise (forward/backward) regression approach proposed by Segura et al. (2012). Stepwise regression is an iterative procedure, where in each step, a SNP is added to the model as a cofactor based on predefined criteria. Then the p-values for all added cofactors are re-estimated together with the variance components.

The process of adding significant SNPs to the model is repeated until the benefit of adding new terms to the model comes sufficiently close to zero. Shrinkage/regularization methods are another class of estimation approach used to solve the 'large p, small n' problems. Regularization methods attempt to estimate all the genetic effects, while the effects of irrelevant covariates (spurious effects) are automatically shrunken toward zero.

Frequentist regularization approaches

Ridge regression and LASSO (least absolute shrinkage and selection operator)(Tibshirani (1996)) are prominent shrinkage methods in the classical framework, and fall under the umbrella of penalized likelihood regression models, with the penalty being imposed on the L2-norm for the former and on the L1 -norm for the latter (see Schmidt (2005) for more details). In frequentist framework, LASSO and its extensions adaptive LASSO (Zou (2006)), elastic net (Zou and Hastie (2005)) and adaptive elastic net (Zou and Zhang (2009)) have been widely used for association mapping or genomic selection studies (Chen and Chen (2008); Wang et al. (2011); Waldmann et al. (2013); Sokolov et al. (2016)). Fan and Li (2001) showed that the LASSO shrinkage produces biased estimates for the large coefficients, and Zou (2006) proposed an extension of LASSO called adaptive LASSO in order to alleviate this bias. Another limitation of LASSO approach is that when there exist high correlations among predictors LASSO will arbitrarily choose one and drop the other predictors. To remedy this problem, elastic net (ENET) was proposed as an extension of LASSO and ENET is robust to high correlations among predictors. Later, Zou and Zhang (2009) proposed adaptive elastic net as a combination of the adaptive LASSO and the elastic net to deal with the collinearity problem and improved performance with high-dimensional data. See Ogutu et al. (2012) and Li and Sillanpää (2012) for a comprehensive review about the frequentist regularization procedures in association mapping and genomic selection. It has been long argued that the classical shrinkage methods (LASSO and its extensions) cannot identify a number of non-zero effects exceeding the sample size. This is a major shortcoming when dealing with genome-wide dense sets of markers and Bayesian formulations of the regularization methods can overcome this with the help of prior distributions.

Bayesian regularization approaches

For the Bayesian inference with model 2 (Eq. 2), prior distributions must be specified for the unknown parameter such as $\mathbf{b_j}$ and σ_e^2. In the Bayesian framework, regularization is achieved by imposing specific prior distribution on the random marker effects and the priors shrink unimportant marker effects toward zero. Normal distribution with a common variance is the simplest prior one can be assume for SNP effects and this is equivalent to the ridge regression BLUP. One disadvantage of using normal distribution for SNP effects is that finally large number of SNP effects will have non-zero values. As a solution to this problem some heavy-tailed distribution, like t-distribution, can be used as a prior distribution for the SNP effects (BayesA; Meuwissen et al. (2001)). Another commonly used heavy-tailed shrinkage distribution for SNP effects is Laplace (double exponential) distribution, which is sharply peaked around zero. This is known as Bayesian LASSO: Park and Casella (2008), Li et al. (2010) and its adaptive counterpart: Extended Bayesian LASSO; Mutshinda and Sillanpää (2010)). Many other variants exist including Meuwissen et al. 2001 (BayesB) and Habier et al. 2011 (BayesC and BayesCπ). These multilocus models can be used both for association mapping and for genomic prediction.

The parameter estimation in most of the Bayesian hierarchical shrinkage methods is based on Markov chain Monte Carlo (MCMC) sampling which may not be optimal for high dimensional data due to their computational complexity. Deterministic methods such as *maximum a posteriori* (MAP) estimation can be a used as an fast alternative to sampling based algorithms. However, MAP estimation methods can produce good point estimates but their accuracy estimates are usually badly underestimated (*i.e.,* the estimated posterior uncertainty around the point estimate is much too narrow). MAP estimation is mainly based on numerical optimization (Gelman et al. (2014)) or different variants of expectation maximization (EM)(Dempster et al. (1977)) algorithm. Variational Bayes (VB) estimation (Jaakkola and Jordan (2000)) offers another class of MAP estimation technique in multilocus models (Li and Sillanpää (2012)) but also their uncertainty measures are too narrow. Variational Bayes can be considered as the extension of traditional expectation-maximization (EM) algorithm and is computationally less intensive than MCMC counterparts for the shrinkage estimation. Many MAP implementations exist for large data sets (for example, Sun et al. (2010), Zhang and Xu (2005), Huang et al. (2015), Mutshinda and Sillanpää (2012), Li and Sillanpää (2012)).

Significance threshold for association

Even though GWAS studies have great potential to pinpoint the single nucleotide polymorphisms underlying quantitative traits, false discoveries are a major concern in associations studies. In a single-locus model-based GWAS study, one is typically screening through thousands of markers by testing association one at a time which may lead to many false positive findings. One important question is which significance level (α) should be chosen in order to reduce the number of false alarms due to multiple testing (i.e., high number of tests performed). Bonferroni correction is one of the most commonly used method to correct for multiple testing in GWAS studies. Bonferroni adjustment treats all markers as independent, even though the markers are likely to be in LD with each other. Thus the Bonferroni adjustment may be too conservative for extremely large number of markers. An alternative to Bonferroni correction, FDR (false discovery rate), which is designed to capture the portion of false positives to the number of total positive test results is also widely applied in GWAS studies (Devlin et al. (2003)). The quantile-quantile (Q-Q) plot (which is a graphical representation of the proportion of significant SNPs compared to the expected number of significant SNPs based on p-values) is also used in GWAS studies based on single-locus model to monitor the number of false positives.

Bonferroni correction and Q-Q plot are suitable when the loci are independent. Thus, it is hard to apply these methods for the multilocus association models because such models search loci jointly and their combinations can be correlated. Permutation test proposed by Churchill and Doerge (1994) is commonly applied for choosing the significance level for association in both single and multilocus association analysis (Xu (2003)). As an interesting alternatives credible interval approach (Li et al. (2010)) and Wald-statistic

(Yang and Xu (2007)) can also be used in multilocus association testing to decide which signals are true. But all these approaches are generally sensitive to the collinearity in the marker data and among these methods permutation test seems to suffers less from the correlated predictors. Another interesting approach is to use the estimates from many MCMC chains with different starting values, and consider only the SNPs (stable signals) that appear in all MCMC iterations as the significant ones (Mathew et al. (2018a); Wei et al. (2014)). For more alternatives, see Chen et al. (2017).

Dimensionality reduction methods

The cost of high-throughput genotypying/phenotyping is no longer a major hurdle for GWAS studies and the biologists have entered the era of Big Data. Variable selection and shrinkage methods are mainly designed to moderate the number of predictors to hundreds or thousands. However with Big Data (ultra-high p small n), these methods may be computationally infeasible and statistically inaccurate. Another problem associated with high dimensional data is that many unimportant predictors can be highly correlated with the response variable and variable selection alone might be difficult in such cases. While making statistical inference with Big data (high dimensional data space) one can use dimensionality reduction approach to reduce the number of predictors (p) close to the sample size before applying variable selection/regularization methods. Collinearity, which is a condition where some of the predictors are highly correlated due to LD with each other is a major problem with the high dimensional data space. In such cases LD pruning can be applied to remove the highly correlated SNPs and preselect a subset of SNPs which are uncorrelated with each other. Then the selected subset of SNPs can be further analyzed using a multilocus association model. The PLINK software offers features for SNP pruning based on LD. SNP tagging (Lin and Altman (2004); Meng et al. (2003)) and SNP binning (Xu (2013)) based on haplotypes are another useful approaches to reduce the dimensionality (by selecting an informative sets of SNPs) in GWAS studies involving millions of SNPs. Sure Independence Screening (SIS) (Fan and Lv (2008)) is an efficient method to reduce the dimensionality of high dimensional data space from ultra-high to a relatively large scale. SIS can preselect the most important predictors based on their marginal correlation with the response variable. Recent studies (Kärkkäinen et al. (2015); Mathew et al. (2018a)) showed that SIS can be applied to preselect the important predictors for multilocus association in very high dimensional cases. SIS is based on univariate screening step and one of the drawback of SIS is that the important predictors that are marginally nearly uncorrelated with the response variable could be missed out because of this univariate screening approach. In order to over come this drawback Fan and Lv (2008) also proposed an iterative procedure called iterative sure independence screening (ISIS). The ISIS procedure iterates the SIS procedure conditional on the previously selected predictors, thus the procedure can capture the important predictors that are marginally nearly uncorrelated with the response variable.

Conclusion

High throughput genotyping technologies are capable of generating enormous set of high density SNP markers with low cost and that enables whole-genome association mapping in natural/breeding populations. Multilocus association mapping approaches, known to have some advantages over conventional QTL-mapping, may importantly have more power to detect QTLs and to control the number of false positives. However, multilocus association analysis involving high-density markers need to apply variable selection or shrinkage approaches in order to identify the best subset of relevant predictor variables. Even though most of the variable selection/shrinkage approaches presented in plant or animal genetics context are primary designed for genomic prediction purposes, they can also be applied for gene mapping. When such methods are applied for association analysis based on multilocus association models one need to perform additional statistical tests for association between the SNPs and the trait of interest to fully control false positives.

References

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635.

Chen, C., Steibel, J. P., and Tempelman, R. J. (2017). Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods. *Genetics*, 206(4):1791–1806.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38.

Devlin, B., Roeder, K., and Wasserman, L. (2003). False discovery or missed discovery? *Heredity*, 91(6):537–538.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, 4:250–255.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1):186.

Huang, A., Xu, S., and Cai, X. (2015). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity*, 114(1):107.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008a). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008b). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Kärkkäinen, H. P., Li, Z., and Sillanpää, M. J. (2015). An efficient genome-wide multilocus epistasis search. *Genetics*, 201(3):865–870.

Kärkkäinen, H. P. and Sillanpää, M. J. (2012). Robustness of Bayesian multilocus association models to cryptic relatedness. *Annals of Human Genetics*, 76(6):510–523.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2010). The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., Buckler, E. S., and Zhang, Z. (2014). Enrichment of statistical power for genome-wide association studies. *BMC biology*, 12(1):73.

Li, Z. and Sillanpää, M. J. (2012). Estimation of quantitative trait locus effects with epistasis by variational bayes algorithms. *Genetics*, 190(1):231–249.

Li, Z. and Sillanpää, M. J. (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, 125(3):419–435.

Lin, Z. and Altman, R. B. (2004). Finding haplotype tagging SNPs by use of principal components analysis. *The American Journal of Human Genetics*, 75(5):850–861.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835.

Mathew, B., Léon, J., Sannemann, W., and Sillanpää, M. J. (2018a). Detection of Epistasis for Flowering Time Using Bayesian Multilocus Estimation in a Barley MAGIC Population. *Genetics*, 208(2):525–536.

Mathew, B., Léon, J., and Sillanpää, M. J. (2018b). A novel linkage-disequilibrium corrected genomic relationship matrix for snp-heritability estimation and genomic prediction. *Heredity*, 120(4):356–368.

Meng, Z., Zaykin, D. V., Xu, C.-F., Wagner, M., and Ehm, M. G. (2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *The American Journal of Human Genetics*, 73(1):115–130.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Müller, B., Stich, B., and Piepho, H. (2011). A general method for controlling the genome-wide type i error rate in linkage and association mapping experiments in plants. *Heredity*, 106(5):825–831.

Mutshinda, C. M. and Sillanpää, M. J. (2010). Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*, 186(3):1067–1075.

Mutshinda, C. M. and Sillanpää, M. J. (2012). Swift block-updating EM and pseudo-EM procedures for Bayesian shrinkage analysis of quantitative trait loci. *Theoretical and Applied Genetics*, 125(7):1575–1587.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC Proceedings*, volume 6, page S10. BioMed Central.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190.

Pikkuhookana, P. and Sillanpää, M. (2009). Correcting for relatedness in bayesian models for genomic data association analysis. *Heredity*, 103(3):223–237.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181.

Pryce, J., Bolormaa, S., Chamberlain, A., Bowman, P., Savin, K., Goddard, M., and Hayes, B. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science*, 93(7):3331–3345.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Schmidt, M. (2005). Least squares optimization with l1-norm regularization. *CS542B Project Report*, pages 14–18.

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7):825–830.

Sillanpää, M. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, 106(4):511–519.

Sokolov, A., Carlin, D. E., Paull, E. O., Baertsch, R., and Stuart, J. M. (2016). Pathway-based genomics prediction using generalized elastic net. *PLoS Computational Biology*, 12(3):e1004790.

Sun, W., Ibrahim, J. G., and Zou, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics*, 185(1):349–359.

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., and Aulchenko, Y. S. (2012). Rapid variance components–based method for whole-genome association analysis. *Nature Genetics*, 44(10):1166.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., and Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, 4:270.

Wang, D., Eskridge, K. M., and Crossa, J. (2011). Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(2):170–184.

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., and Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, 6:19444.

Wei, W.-H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722.

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., Wang, S.-B., Dunwell, J. M., Zhang, Y.-M., and Wu, R. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings in Bioinformatics*, pages 1–13.

Würschum, T. and Kraft, T. (2015). Evaluation of multi-locus models for genome-wide association studies: a case study in sugar beet. *Heredity*, 114(3):281.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801.

Xu, S. (2007). An empirical bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 63(2):513–521.

Xu, S. (2013). Genetic mapping and genomic selection using recombination breakpoint data. *Genetics*, 195(3):1103–1115.

Yang, R. and Xu, S. (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics*, 176(2):1169–1185.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208.

Zhang, Y.-M. and Xu, S. (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity*, 95(1):96.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42(4):355–360.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733.