OA Genetics
Open Access

Linkage Analysis & Genetic Mapping

Critical review

# From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics

A Huang*, D Liu

## Abstract

**Introduction**

Mendelian theory of inheritance explains qualitative traits that are controlled by single genetic locus, while modern quantitative genetic theory states that complex traits are influenced by many factors including main effects of many quantitative trait loci, epistatic effects involving more than one quantitative trait loci, environmental effects and the effects of gene–environment interactions. Quantitative trait locus mapping identifies important quantitative trait loci that reveal genetic basis of complex traits, serve as a map for functional gene cloning, and assist breed line selection. The aim of this study was to discuss the quantitative trait locus method and the resulting genomic selection from quantitative trait locus mapping.

**Discussion**

We introduced different breeding populations, genetic models and statistical methods for quantitative trait locus mapping. Genetic background, model principles and computational algorithms and techniques were reviewed. We specifically discussed the whole-genome quantitative trait locus mapping that simultaneously estimates genetic effects associated with markers of entire genome. The fact that the whole-genome approach avoids evaluation of multiple models and model selection while enables genomic selection that predicts genetic merits for a quantitative trait has drawn great attention in research community, and will be an effective tool in breeding line selection.

**Conclusion**

While traditional quantitative trait locus mapping was designed with the availability of marker density and model capability in recent two decades, whole-genome quantitative trait locus mapping is the result of advancement in generating high-density molecular markers and development in high-dimensional sparse modelling algorithms. Whole-genome quantitative trait locus mapping and genomic selection together lead to a systematic genetics that will increase genetic gain and revolutionise crop and livestock breeding.

*Corresponding author
Email: a.huang1@umiami.edu

Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, USA

## Introduction

The genotype of an individual organism is the unordered allele pairs at one or more loci. As a result of genetically programmed biological developments, individual organisms exhibits observable characteristics called phenotypes or traits, which may be quantitative such as height and weight, or qualitative such as gender and disease status. The understanding of genotype/phenotype relationship is of paramount importance for both scientific research and social economics. For example, use of detected quantitative trait loci (QTLs) has been proved as an effective tool to increase food production, resistance to diseases and pests, tolerance to heat, cold and draught, and to improve nutrient contents in animal and plant breeding during the last two decades[1].

Mendelian theory of inheritance explains qualitative traits that are controlled by single genetic locus. If complex traits in animal/plant breeding or diseases of human beings are controlled by many genetic loci, individual effect of each locus can hardly be distinguished. Alternatively, the quantitative inheritance theory assumes that complex traits are resulted from multiple gene factors, gene–gene interactions as well as environmental effects[2]. Each of the main and interaction effects exhibits only a modest effect on the phenotype and it is difficult to dissect their individual effect. The quantitative inheritance theory is applicable for all complex traits in different organisms, even for prokaryotes. For example, it has been shown that genes belonging to specific/non-specific membrane channels, oxidative stress response and osmotic stress response are involved in conferring bacterial resistance to high arsenic level, and chasing for a single gene or single regulon may end up with no meaningful result[3,4]. While systematic sampling and genotyping is still lacking for prokaryotes, genotype/phenotype association in animal/plant breeding can be modelled by QTL mapping in breeding lines. In human populations, this relationship is studied by examining the association of phenotypes with the natural occurring genetic variations such as single-nucleotide polymorphisms (SNPs).

QTL is a region in genome that is responsible for variation in the phenotype of interest. In animal/plant breeding, molecular markers are selected in even space throughout entire genome and QTL mapping is to infer which genetic loci are strongly associated with the complex trait



FOR CITATION PURPOSES: *Huang A, Liu D*. From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics. OA Genetics 2013 May 01;1(1):4.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Genetics
Open Access



and to estimate the genetic effects of these loci. Two inbred lines with different traits of interest are chosen to cross and the first generation ($F_1$) will have identical genetic markers that show complete linkage disequilibrium (the non-random association of alleles at different loci) for genes differing between the breeding lines. Starting from the $F_1$, a number of designs have been proposed for QTL mapping. For example, a backcross design is to cross the $F_1$ individuals to one of the two parental lines; an intercross design is to cross between siblings among $F_1$ individuals; a doubled haploid design is to develop individuals from pollens of an $F_1$ plant through another culture and chromosome doubling; and a recombinant inbred lines design is to cross between sibling individuals for many generations start from $F_2$ till almost all of the segregating loci come to be homozygous. The different experiment designs produce different breeding populations for QTL mapping, and the $F_2$ population provides the most of genetic information among different types of mapping populations[5].

With the advent of new DNA sequencing technologies, high density markers can be easily generated along the genome. However, it is still very likely that true causal markers are not captured due to the large amount of genomic variants in living organisms[6]. On the other hand, with the large amount of available genetic markers, researchers usually have no idea about the number, location and effect of the markers involved in the inheritance of target phenotypes. Therefore, the correlation among genetic markers and oversaturated models are two common properties in QTL mapping and SNP-based association studies. In this review, we will discuss traditional QTL mapping methods and novel statistical methods that enable whole-genome QTL mapping, meanwhile introduce the idea of genomic selection(GS) resulted from whole-genome QTL mapping.

## Discussion

The authors have referenced some of their own studies in this review. The protocols of these studies have been approved by the relevant ethics committees related to the institution in which they were performed.

### Traditional QTL mapping methods

Techniques for QTL mapping include single marker mapping, interval mapping, multiple loci mapping as well as composite interval mapping. Principles in these mapping techniques are generally the same and methods used in one population can be extended to other experiment populations. Single marker test examines the segregation of quantitative or qualitative traits with respect to the examining genotype at a single locus, while multiple loci mapping considers multiple makers, possible high order maker interactions as well as environment factors simultaneously. The interval mapping[7] and composite interval mapping[8] are extensions of single marker test and multiple loci test, respectively. In the interval mapping, consecutive two testing markers (or several markers in a testing window in composite interval mapping) are ordered according to their physical location, and those peak testing values in single tests are declared as QTLs[7]. The various techniques test the genotype and phenotype association with different genetic effects such as additive and dominance effect. In regression models, they can be examined simultaneously by adding dummy variables to encode different effects. A widely used genetic model is the Cockerham model[9] that defines the values of the additive effect as –1, 0 and 1 for the three genotypes and the values of the dominance effect as –0.5 and 0.5 for homozygotes and heterozygotes, respectively.

Similar to QTL mapping, the goal of association study in natural population is to identify SNPs in individuals that are systematically associated with different disease states. Using the natural occurring DNA variations as markers to trace inheritance in families are similar to QTL mapping, while extra cares are required to handle population structures in genome-wide association study of common human diseases. In the ensuing sections, we will focus on the general computational methods that have been applied to both problems.

### Multiple-variant methods

The single variant approach is simple and straightforward for most models such as *t*-test, analysis of variance and simple linear regression for quantitative traits, the Cochran–Armitage test, Pearson $\chi^2$ test and generalised regression for qualitative traits. However, single marker methods require multiple test correction such as Bonferroni correction and false discovery rate[10], to control the overall type I error rate. Given the large number of possible markers, multiple test correction leads to such stringent criterion that most modest effects could not pass the threshold. Moreover, since complex phenotype is controlled by multiple genetic factors and their interactions with environmental effects, marker identified with a single marker method may only explain a small proportion of the phenotype variation.

Providing the promising of high power and reasonable type I error rate, multiple-variant approach needs to take care of several challenges. Firstly, traditional ordinary least square method fails for the case of $p >> n$. When taking into account of interactions between loci, variable selection is even more demanding. Secondly, it is likely that the increased degree of freedom in



FOR CITATION PURPOSES: *Huang A, Liu D*. From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics. OA Genetics 2013 May 01;1(1):4.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Genetics
Open Access

Critical review

a multiple-variant model jeopardises the power gain over single-locus models. Thirdly, in binary-trait or case–control association analysis, complete or semi-complete separation can be a serious problem due to the discrete nature of both marker data and binary outcomes[11]. Accurate variable selection and sparse model inference methods become critical for the multiple-variant approach, especially when epistatic effects are considered. Variable selection and shrinkage are two techniques in handling the selection problems, and they typically produce sparse models.

Variable selection methods include forward selection, backward elimination and forward stage-wise selection. However, these greedy selection methods may result in suboptimal subset, and are computational expensive even for small number of variables[12]. On the other hand, variable shrinkage method includes all variables in the model and applies a penalty function or appropriate prior distributions on the variables to automatically shrink most non-effects towards zero. Moreover, including all possible effects into a single model results in whole-genome QTL mapping, which overcomes limitations of the genetic model considered by traditional mapping methods and prevents all problems of model selection[13].

## Whole-genome QTL mapping and genomic selection

Strictly speaking, whole-genome QTL mapping includes additive and dominance main effects and all pairwise interactions. For a QTL model includes $k$ markers, the total number of effects is $p = 2k + 4k(k - 1)$. However, in practice, most multiple QTL models cannot handle such large number of effects and epistatic effects are usually not considered (which is named as genome-wide QTL mapping)[14]. State-of-the-art whole-genome QTL model can handle up to $10^8$ effects including all main and pairwise interaction effects on a personal computer[15]. Nevertheless, next generation sequencing technologies can easily genotype up to $10^7$ SNPs for most model organisms and whole-genome QTL mapping at this scale is computationally too demanding and causes memory overflow. In such cases, markers carrying duplicated information content need to be deleted[16].

Another natural outcome of whole-genome QTL mapping is GS, which is originally proposed in genome-wide QTL mapping that mapping results enable predictions of estimated breeding values (EBV)[14]. GS is rooted from marker-assisted breeding, and can be viewed as a new generation of breeding method different from traditional breeding that relies on phenotypic selection and relative information[17]. In GS, individual genetic merits are estimated by simultaneously accounting all markers and all types of marker effects. From biological point of view, a marker map must cover all genomic positions such that all QTLs can be assessed for their contributions to EBV, although in reality some adjustment for linkage disequilibrium might be required to avoid colinearity. From computational point of view, the QTL model must simultaneously estimate all genetic effects including main and epistatic effects, genetic and environmental interactions and effects of rare alleles, such that the breeding values are predicted based on all significant effects[18]. Whole-genome QTL mapping provides more accurate marker effects and phenotypic variance estimation, which results in better performance of predicting genomic EBV.

## Variable shrinkage methods

Two well-known methods with ability in handling large number of variables are regularised regression and Bayesian shrinkage method[19]. Consider the general multiple linear regression problem with $n$ samples and $p$ variables:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector collecting $n$ individual phenotype, $\mathbf{X}$ is an $n \times p$ designed matrix for $p$ marker effects, $\boldsymbol{\varepsilon}$ is the residual error that follows a normal distribution with zero-mean and covariance $\sigma_0^2\mathbf{I}$: $\varepsilon \sim N(0, \sigma_0^2\mathbf{I})$. The generic penalised regression method infers the model parameters as:

$$\hat{\beta} = \arg_\beta \min \left\| \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \eta(\boldsymbol{\beta}) \right\|_q^q, \qquad (2)$$

where $\eta(\beta) \geq 0$, and $\left\| \eta(\boldsymbol{\beta}) \right\|_q^q = \sum_{j=1}^{p} \eta(\beta_j)^q$ is the $l_q$ penalty term. Here, the intercept is absorbed into matrix $\mathbf{X}$, and it is not penalised. Without loss of generality, variables are assumed to be re-scaled to have variance 1 to obtain invariant penalties. Notice that when $\eta(\beta_j) = \begin{cases} 1, \text{ if } \beta_j \neq 0 \\ 0, \text{ otherwise} \end{cases}, \forall j,$ the model penalises the number of effects, and the solution of (2) is equivalent to the AIC (for $\lambda = 2$) or BIC (for $\lambda = \ln(n)$) in forward selection and backward elimination, and the estimates are the same as the least square solutions. From this point of view, variable shrinkage is a generalisation of variable selection. Let $\eta(\beta_j) = |\beta_j|, j = 1, 2, \cdots, p$ to penalise the effect size, (2) becomes the bridge regression, which is a generalisation of two most popular penalised regression methods, namely the $l_2$ penalty of the ridge regression[20], and the $l_1$ penalty of the least absolute shrinkage and selection operator (Lasso) that implicitly enforces a sparse solution[19].

In the Bayesian shrinkage approach, a prior distribution $p(\boldsymbol{\beta}|\lambda)$ is assigned to $\boldsymbol{\beta}$. The posterior distribution given the observed data takes the form of

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \lambda) \propto p(\boldsymbol{y}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\lambda). \qquad (3)$$



FOR CITATION PURPOSES: *Huang A, Liu D*. From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics. OA Genetics 2013 May 01;1(1):4.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Genetics
Open Access



If we are looking for $\boldsymbol{\beta}$ that maximises the posterior distribution, the problem is equivalent to finding

$$\hat{\boldsymbol{\beta}} = \arg_{\beta}\max\{\log p(\boldsymbol{y}|\boldsymbol{\beta}) + \sum_{j=1}^{p}\log p(\beta_j|\lambda)\}, \quad (4)$$

which is the penalised ML method. Then $\hat{\boldsymbol{\beta}}$ is the mode of the posterior distribution and this method is referred as *maximum a posteriori* (MAP) approach.

The Bayesian interpretation of penalised regression methods has sparked interests in developing Bayesian hierarchical regression models, which can be called as a Bayesian Lasso approach. A direct gain of Bayesian Lasso is to incorporate variance information to facilitate significance test, and different techniques are employed for different designs to achieve sparse estimation. Theoretically, prior distributions with spike finite limit at zero and flat tails at two ends can be penalty of the log posterior distribution (Figure 1). In practice, conjugate priors are often chosen. Among them the Normal + *inv*-$\chi^2$ (*t*-family) distribution[21] and NEG hierarchical priors[22-24] are most oftenly used with both variable shrinkage and computational feasibility considerations. Table 1 listed several popular Bayesian Lasso methods designed.

For Bayesian estimation, Markov Chain Monte Carlo (MCMC) can be employed to draw samples from posterior distribution for each parameter. However, for high dimensional data the MCMC method is known to be computationally intensive. The MAP is a modal representation of the posterior distribution that achieves faster computation and easier interpretation. Efficient methods for MAP estimation have been developed, which further integrate out the variance components and employ numerical methods to obtain similar sparse results as Lasso (named as HyperLasso)[22]. Another method stands in between MCMC and

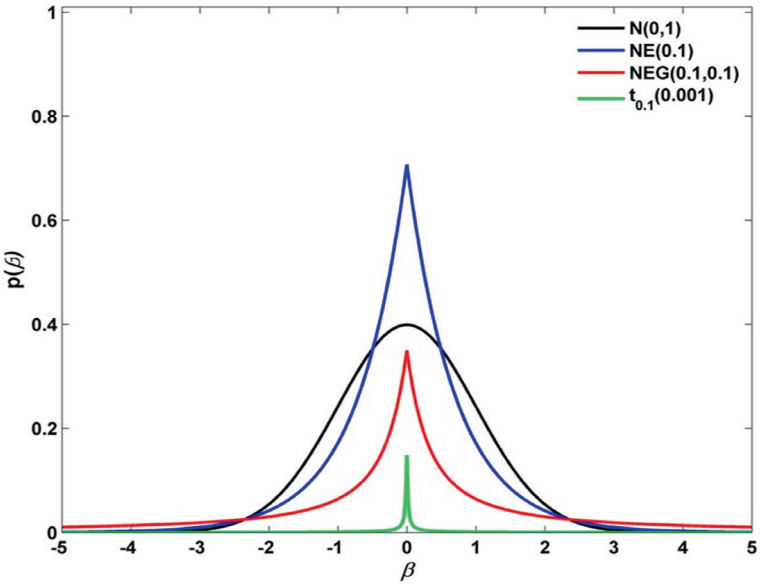


***Figure 1:*** Prior distributions that penalise posterior distributions.

**Table 1 Popular Bayesian Lasso methods**

| Algorithm | Prior distribution | Inference method | Reference |
|---|---|---|---|
| Bayesian Lasso | NE[a] | MCMC[c] | Park T, Casella G[27] |
| Bayesian Lasso | Normal + inv_$\chi^2$ | MCMC | Yi N, Xu S[28] |
| Bayesian Lasso | NEG[b] | MCMC | Li J et al.[29] |
| BhGLM | Normal + inv_$\chi^2$ | EM[d] | Yi N, Banerjee S[25] |
| Bayesian HyperLasso | NEG | EM | Griffin JE, Brown PJ[30] |
| Bayesian Lasso | Normal + Cauchy | EM | Gelman A et al.[11] |
| RVM[e] | Normal + Uniform | Empirical Bayes | Tipping ME, Faul AC[31] |
| EBlasso | NE/NEG | Empirical Bayes | Cai X et al.[23]<br>Huang A et al.[24] |

[a]Normal and exponential hierarchical prior distribution.
[b]Normal, exponential and gamma hierarchical prior distribution.
[c]Morkov chain Monte Carlo.
[d]Expectation-maximization algorithm.
[e]Relevant vector machine.

MAP is the expectation-maximization (EM) algorithm that estimates posterior mean and variance simultaneously through iterative expectation (E-step) and maximization (M-step). However, convergence become a serious problem when model dimension increases[25]. Recently, more efficient algorithm that does not rely on MCMC yet infers the posterior distribution is achieved by empirical Bayesian Lasso (EBlasso) method[23,24].

Different from the iterative EM algorithm, EBlasso first finds the marginal posterior distribution of the variance components. Due to the shrinkage applied from the prior distribution on the variance components, most variables will have zero variance that maximises the marginal posterior distribution, and only those variables with none-zero variance will stay in the model, result in a sparse presentation. Next, the posterior means of the



FOR CITATION PURPOSES: *Huang A, Liu D*. From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics. OA Genetics 2013 May 01;1(1):4.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*

OA Genetics
Open Access



non-zero variables are estimated with the give variance. Along with other algorithmic techniques, the EBlasso approach is very efficient and is able to handle a model with several million of variables[26]. Our previous studies demonstrated that EBlasso outperformed several other multiple QTL mapping methods including the empirical Bayes method in[21], the Bayesian hierarchical generalised linear models (BhGLM)[25], HyperLasso[22] and Lasso[19]. EBlasso has also been applied to whole-genome QTL mapping[15] and can be easily extended to GS.

## Conclusion

A correct QTL model is the one that includes all true QTLs and estimates their effects simultaneously. In real data analyses, we would like to perform a significance test since true QTLs are unknown. Although Lasso and HyperLasso both can handle large models, they only yield a point estimate without variance information. Bootstrapping, refitting to an ordinary least square model, as well as covariance test statistics developed during the Lasso selection path have been associated with the two methods for a significance test. The EBlasso method, on the other hand, can handle a large model with a speed comparable with that of Lasso, and estimate the posterior distribution for a sparse model. The availability of high-dimensional sparse models with capability in handling few million effects enables whole-genome QTL mapping and GS. In fact, with high-density markers available in many animal and plant organisms collecting from both inbred lines and natural populations, GS can play a significant role in improving breading technologies.

## References

1. Balding DJ, Bishop M, Cannings C. Handbook of statistical genetics. UK: Wiley; 2008.
2. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. Boston: Addison-Wesley; 1996.
3. Huang A, Teplitski M, Rathinasabapathi B, Ma L. Characterization of arsenic-resistant bacteria from the rhizosphere of arsenic hyperaccumulator Pteris vittata. Can J Microbiol. 2010 Mar;56(3):236–46.
4. Huang A. Characterization of arsenic resistant bacterial communities in the rhizosphere of an arsenic hyperaccumulator Pteris Vittata L: University of Florida; 2009.
5. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, et al. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987 Oct;1(2):174–81.
6. Consortium TIH. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct;449(7164):851–61.
7. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics. 1989 Jan;121(1):185–99.
8. Zeng ZB. Precision mapping of quantitative trait loci. Genetics. 1994 Apr; 136(4):1457–68.
9. Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics. 1954 Nov;39(6):859–82.
10. Dale R N. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004 Apr;74(4):765–9.
11. Gelman A, Jakulin A, Pittau MG, Su YS. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat. 2008 Dec;2(4):1360–83.
12. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
13. Xu S. Estimating polygenic effects using markers of the entire genome. Genetics. 2003 Feb;163(2):789–801.
14. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001 Apr;157(4):1819–29.
15. Huang A, Xu S, Cai X. Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice. PLoS ONE. 2014.
16. Xu S. Genetic mapping and genomic selection using recombination breakpoint data. Genetics. 2013 Nov;195(3):1103–15.
17. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 2009 Feb; 92(2):433–43.
18. Jonas E, de Koning D-J. Does genomic selection have a future in plant breeding? Trends Biotechnol. 2013 sep;31(9): 497–504.
19. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1): 267–88.
20. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67.
21. Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics. 2007 Jun; 63(2):513–21.
22. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet. 2008 Jul;4(7):e1000130.
23. Cai X, Huang A, Xu S. Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics. 2011 May;12(1):211.
24. Huang A, Xu S, Cai X. Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC Genet. 2013;14(5):5.
25. Yi N, Banerjee S. Hierachical generalized linear models for multiple quantitative trait locus mapping. Genetics. 2009 Mar;181(3):1101–33.
26. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R. A significance test for the lasso. 2012;arXiv:13017161.
27. Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc. 2008 Jun;103(482):681–6.
28. Yi N, Xu S. Bayesian LASSO for quantitative trait loci mapping. Genetics. 2008 Jun;179(2):1045–55.
29. Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. Bioinformatics. 2011 Feb;27(4):561–23.
30. Griffin JE, Brown PJ. Bayesian hyperlassos with non-convex penalization. Aust N Z J Stat. 2011 Dec;53(4):423–42.
31. Tipping ME, Faul AC. Fast marginal likelihood maximisation for sparse Bayesian models. Proc 9th International Workshop on Artificial Intelligence and Statistics. 2003.



FOR CITATION PURPOSES: *Huang A, Liu D*. From quantitative trait locus mapping to genomic selection: the roadmap towards a systematic genetics. OA Genetics 2013 May 01;1(1):4.

*Competing interests: none declared. Conflict of interests: none declared.*
*All authors contributed to the conception, design, and preparation of the manuscript, as well as read and approved the final manuscript.*
*All authors abide by the Association for Medical Ethics (AME) ethical rules of disclosure.*