A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus Pungitius

Guo, Baocheng

2019-09-02

Guo , B , Fang , B , Shikano , T , Momigliano , P , Wang , C , Kravchenko , A & Merilä , J 2019 , ' A phylogenomic perspective on diversity, hybridization and evolutionary affinities in the stickleback genus Pungitius ' , Molecular Ecology , vol. 28 , no. 17 , pp. 4046-4064 . https://doi.org/10.1111/mec.

http://hdl.handle.net/10138/321981
https://doi.org/10.1111/mec.15204

acceptedVersion

Linkage disequilibrium clustering-based approach for association mapping with tightly linked genome-wide data

Keywords: GWAS, quantitative trait loci, principal component regression, multi-locus method, four-way cross

Zitong Li[1*], Petri Kemppainen[1*], Pasi Rastas[1], Juha Merilä[1]

[1]*Ecological Genetics Research Unit, Research Programme in Organismal and Evolutionary Biology, Faculty of Biological and Environmental Sciences, Department of Biosciences, University of Helsinki, P.O. Box 65, FI-00014 Helsinki, Finland*

*Equal contribution

Correspondence to Zitong Li, Tel. + 358-40 8169003; E-mail: lizitong1985@gmail.com

Formatted: Finnish

Abstract

Genome-wide association studies (GWAS) aim to identify genetic markers strongly associated with quantitative traits by utilizing linkage disequilibrium (LD) between candidate genes and markers. However, because of LD between nearby genetic markers, the standard GWAS approaches typically detect a number of correlated SNPs covering long genomic regions, making corrections for multiple testing overly conservative. Additionally, the high dimensionality of modern GWAS data poses considerable challenges for GWAS procedures such as permutation tests, which are computationally intensive. We propose a cluster-based GWAS approach that first divides the genome into many large non-overlapping windows, and uses linkage disequilibrium network analysis in combination with principal component (PC) analysis as dimensional reduction tools to summarize the SNP data to independent PCs within clusters of loci connected by high LD. We then introduce single- and multi-locus models that can efficiently conduct the association tests on such high dimensional data. The methods can be adapted to different model structures, and used to analyse samples collected from the wild or from bi-parental F_2 populations, which are commonly used in ecological genetics mapping studies. We demonstrate the performance of our approaches with two publicly available data sets from a plant (*Arabidopsis thaliana*) and a fish (*Pungitius pungitius*), as well as with simulated data.

Introduction

A central problem in quantitative genetics is to understand the relationship between genotypes and quantitative traits. A Genome-wide association study (GWAS; Balding 2006; Korte and Farlow 2013) is a population-based approach to identify a set of candidate loci associated with complex traits from a genome-wide set of genetic variants. Another closely related approach is quantitative trait locus (QTL) mapping (Mackay et al. 2009), which utilizes experimental crosses or pedigree data. The major difference between the GWAS and QTL approaches is that the former utilizes historical recombination events, whereas the latter relies on recent recombination events to detect association / linkage between genetic markers and phenotypes. Nevertheless, both approaches tend to use similar types of statistical methods, such as linear regression, to identify phenotype-genotype associations (Ernst and Steibel 2013). Therefore, although the main focus of this methodological paper is on statistical analysis of GWAS data, we will also demonstrate how the developed approach can be utilized with QTL mapping data.

The most widely used statistical approaches for GWAS belong to two classes: single-locus and multi-locus mapping methods (Yi et al. 2015). Single-locus methods utilize a marginal linear regression approach to map a quantitative trait to a single SNP at a time. In contrast, multi-locus approaches jointly estimate the effects of multiple SNPs on the trait. For both methods, hypothesis testing can be conducted to judge whether the SNPs are significantly associated with the trait, followed by correction for multiple testing to reduce the risk of calling false positive variants.

Next generation sequencing techniques have provided a cost-effective access to large genomic data sets, such as high-resolution SNP panels. The accessibility of such panels in GWAS and QTL studies provides an opportunity to fine-map the casual loci underlying phenotypes but such high dimensional data sets also pose great challenges. First, in many ecological GWAS and QTL-mapping studies, sample sizes are often limited to few hundreds of individuals due to logistic or budgetary limitations. However, the number of SNPs in these studies may reach hundreds of thousands or even several million, creating what statisticians know as a ‘*p* much larger than *n*’ problem (i.e. number of SNPs is much larger than the number of individuals; Hastie et al. 2009). Second, another feature of large genomic data sets is that SNPs which are physically close to each other are often in linkage disequilibrium (i.e. correlated). This high dimensionality and correlation structure of population genomic data sets pose difficulties for both single- and multi-locus mapping approaches to identify QTL (Xu 2013a). First, single-locus mapping approaches rely on multiple-testing corrections to reduce the rate of false positives. The most conventional and widely used approach is the Bonferroni correction (Dudbridge and Koeleman 2004), which works best when the multiple hypothesis tests are independent from each other. Thus, the Bonferroni correction typically becomes overly conservative when the tests are positively correlated, which is likely to be the case when LD is prevalent in the data.

Since a group of SNPs in high LD explain similar amounts of genetic variation in a given trait, it is reasonable to apply a dimensional reduction procedure before GWAS to exclude the redundant information from the data, and also to reduce the computational cost. Distance thinning (Danecek et al. 2011) is probably the most intuitive way for LD reduction, by simply extracting a subset of "unlinked" SNPs located within equal physical distance to each other. However, this approach does not account for the fact that the degree of LD among the loci can be unequal across the genome. A genome may consist of long LD blocks with hundreds of highly correlated SNPs, or it may contain singletons that are effectively unlinked even to nearby SNPs. In addition, unless recombination is entirely restricted between adjacent loci (e.g. due to an inversion) LD patterns across short physical distances are typically mosaic-like with potentially several distinct sets of loci connected by high LD overlapping in the genome (Daily et al. 2001; Zhang et al. 2002; Fig. 1). To account for this, some GWAS software, such as PLINK (Purcell et al. 2007), has implemented a LD pruning approach which first divides the genome into many (equal sized) windows, and then uses statistics to identify a few unlinked "tag" SNPs representative for the given window. These "tag" SNPs will then be used in the GWAS analyses. However, potentially much more information could be gained if groups of SNPs in high LD were analyzed jointly by either single- or multi-locus mapping approaches.

An alternative window-based approach aggregates information from multiple correlated SNPs and uses a few uncorrelated summary statistics to replace the original data (Ge et al. 2016). A benefit of this summary statistics-based approach is that it can reduce noise in the data due to sequencing errors (Beissinger et al. 2015). Xu (2013a) introduced this kind of window-based approach for QTL mapping. First, the chromosome was divided into many artificial (selected by the users) or natural windows (selected on the basis of breakpoints in the linkage map). Second, a numerical integration approach was used to aggregate the SNP data in every window, which revealed that this approach is equivalent to calculating the mean genotype value of multiple SNPs. Xu's (2013a) approach is related to the 'burden test' initially proposed in human genetics (Morgenthaler and Thilly 2007) to test a group of SNPs as a biological meaningful unit, such as a gene or a biochemical pathway. Within a functional unit, the SNPs were often summarized by dimensional reduction (Hibar et al. 2011) or smoothing techniques (Fan et al. 2013). For example, Hibar et al. (2011) proposed to use principal component analysis (PCA) for compressing SNP data prior to GWAS. The PCA is able to represent the original SNP data set with a set of independent principal components (i.e. orthogonal axes which explain the largest proportion of variation in the data). The chief benefit from the burden test-based approach is that it can maintain large amounts of the information in the data, while still effectively reducing the dimensionality. However, the burden test relies on prior knowledge of genome annotations, which may not be available for many species, especially for non-model organisms from the wild.

Recently, Kemppainen *et al*. (2015) proposed to use network analytical tools (LD network analysis: LDna) to study genome wide LD-patterns in population genomic data sets. This unsupervised method effectively partitions genomic data into sets of loci that have similar phylogenetic signals irrespective of their physical position in the genome. As such, the LDna approach could provide a useful tool for flexible dimensionality reduction in gene mapping approaches utilizing large genomic datasets.

The aim of this paper is introduce and test the performance of a novel cluster-based association mapping approach attempting to solve, or at least reduce, some of the problems faced by existing mapping approaches. This approach uses LD network clustering ('LDn-clustering') and PC regression as dimensionality reduction tools enhance computational efficiency of QTL detection. The first step of this approach involves an extension of the LDna approach (Kemppainen et al. 2015) and uses linkage disequilibrium network analysis for grouping loci connected by high LD in non-overlapping windows (i.e. small subsets of loci at time) along chromosomes. This LDn-clustering can define distinct sets loci connected by high LD even when the groups of loci are interspersed and/or physically overlapping along chromosomes (Fig. 1). The second step of the novel approach involves adoption of Hibar et al.'s (2011) strategy to use PCA as a method for dimensionality reduction in each cluster of loci connected by high LD ('LD-clusters).

An additional novel methodological contribution of this work is that the single locus-based linear regression approach of Hibar et al. (2011) was generalized to a single- and multi-locus linear mixed model (LMM) context with the possibility to include a random effect to control for spurious effects of population structure. Consequently, the method is suitable for analyzing data sets with hidden family and population structure, including data collected from the wild. We illustrate the utility of the novel approach using two publicly available data sets: 278 nine-spined sticklebacks (*Pungitius pungitius*) genotyped for 74 078 SNPs (Yang et al. 2016; Li et al. 2017; Rastas et al. 2017), and 337 thale cresses (*Arabidopsis thaliana*) genotyped for 200 121 SNPs (Atwell et al. 2010; Baxter et al. 2010) as well as simulated data.

Materials and Methods

Single-locus models for association mapping

Suppose we have a sample of individuals collected from a general population. A quantitative trait with phenotypic observations is denoted as y_i (i=1,…,n; n = total number of individuals), and bi-allele SNP genotypes are denoted as x_{ij} (j=1,…,p; p is the number of SNPs). A simple linear regression model for detecting an association between the phenotype and each single SNP is defined as

$$y_i = \beta_0 + x_{ij}\beta_j + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (1)$$

where β_0 is the population mean, and β_{j} is the marginal additive effect of the SNP j. The SNP data are typically coded as 1, 0 and -1 for three possible genotypes AA, AB and BB, respectively. When there are only two possible genotypes, as in the case of self-pollinating plants, the SNPs can be simply coded as 0 and 1. The residual error ε_i independently follows a normal distribution with zero mean and variance σ_e^2.

When the dominance effect is of interest, model (1) can be extended as

$$y_i = \beta_0 + x_{ij}\beta_j + z_{ij}\gamma_j + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (2)$$

where z_{ij} is an indicator of the dominance, coded as 0, 1 and 0 for AA, AB and BB for the SNP j; γ_j is the dominance effect, and all other notations are the same as in (1).

To test if a SNP is significantly associated with a trait, one can test the null hypothesis: $\beta_j = 0$ against the alternative hypothesis: $\beta_j \neq 0$. Standard procedures including t- and F-tests can be used (Kutner et al. 2004). Since many hypothesis tests are simultaneously conducted, it is important to adjust the p-values (i.e. adjust the significance threshold α) to control for false positives. Bonferroni correction (Shaffer 1995) – simply adjusting the significance threshold (α) by dividing it by the number of SNPs (p; i.e. α/p) – is a conventional and popular way to control the family wise error (FWER): the probability of having one incorrectly rejected null hypothesis among all the hypotheses (Efron 2010). The drawback of the Bonferroni correction is that the multiplicity adjustment procedure can be overly conservative, such that the test lacks the power to detect SNPs truly associated with traits. This happens, for instance, when the *p*-values are positively correlated (Goeman and Solari 2014), as in the case when the tested SNPs are in strong LD. A solution to circumvent this problem is to use permutation tests to control for the FWER. Here the phenotype data is randomly re-shuffled thousands of times, and the association analysis is conducted repeatedly on each re-shuffled data set. In this way, the empirical distribution of the test statistics can be obtained, and the adjusted *p*-values can be calculated based on these distributions to control the multiplicity (Westfall and Young 1993). The main benefit of a permutation test is that it can effectively account for the correlation structure among the multiple tests (Efron 2010), and yields less conservative thresholds and more power to detect true positive SNPs. However, the permutation approach is very time consuming for large GWAS data sets. Because of this, Bonferroni correction remains one of the most commonly used multiple testing approaches in GWAS studies (e.g. Goeman and Solari 2014; Segura et al. 2012; Husby et al. 2015).

Linear mixed models for controlling population structure

When there is hidden population and/or family structure in the data that may affect the association mapping, a linear mixed model can be applied to control for it:

$$y_i = \beta_0 + x_{ij}\beta_j + u_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (3)$$

where the random effect u_i is specified as $\mathbf{u} = [u_1,...,u_n] \sim \mathrm{MVN}(0,\sigma_g^2\mathbf{A})$ with known $n \times n$ sized relationship matrix **A** and unknown variance σ_g^2. The random effect **u** accounts for relatedness among the individuals, and it can help to reduce spurious effects caused by the population and/or family structure (Yu et al. 2006). The relationship (kinship) matrix **A** can be estimated from molecular marker information as (van Raden 2008):

$$A_{ik} = \frac{1}{p}\sum_{j=1}^{p}\frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1-p_j)}, \qquad (4)$$

where p_j is the minor allele frequency of the SNP j ($j=1,\dots,p$), x_{ij} and x_{kj} are the genotype values of individuals i and k (i, k=1,…,n) at the SNP j. Alternatively, one may also estimate the relationship matrix from the known pedigree of the individuals.

Restricted maximum likelihood (REML) based programs such as EMMA (Kang et al. 2008) and EMMAX (Kang et al. 2010) have been widely used to evaluate the regression parameters and variance components as described by Equation (3). The EMMA approach refers to a computational procedure which uses REML to estimate the variance components repeatedly for each SNP. In contrast, EMMAX estimates the variance components once based on an intercept model, and then fixes them to evaluate the effect and statistical significance of the SNPs. Consequently, the EMMAX approach is much faster and simpler to use on large data sets, and both simulation and empirical studies have shown that the EMMAX approach can have the same statistical power and ability to control for false positives than the more precise EMMA method (Kang et al. 2010). Therefore, we will consider EMMAX as the default method for mixed model analysis in this work.

In a linear mixed model, the hypothesis testing can be conducted using *t*- or *F*-tests in a similar way as in the case of standard linear regression. Bonferroni correction can also be straightforwardly used for multiple testing. However, the permutation test procedure used for standard linear model (1) is not applicable for the mixed model. The reason is that the standard permutation test randomly reshuffles phenotypes, which is equivalent to sampling phenotype data from a uniform distribution, and this implementation will remove any among-individual correlation from the data. Clearly, this violates the assumption of dependency structure among individuals in the mixed model, and might yield spurious statistical results (Joo et al. 2016). A correct way to conduct permutation tests on the basis of the mixed model would be to draw a sufficient number of independent samples from a multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \hat{\sigma}_g^2\mathbf{A} + \hat{\sigma}_e^2\mathbf{I})$, and then use EMMAX to calculate the test statistics on each sample (Joo et al. 2016). However, as in the case of standard linear regression, the permutation procedure will consume a considerable amount of computational time.

Single-locus models for four-way crosses

The linear models described by equations (1), (2) and (3) are standard choices for association analyses performed with bi-allele SNPs. In some circumstances, such as in the case of a four-way cross (Xu 1996), F_1 offspring of a hybrid cross generated from two heterozygous parents (Van Ooijen 2009), and in the case of an outbred F_2 design (Xu 2013b), there might be up to four possible alleles, A_1, A_2, B_1 and B_2 originating from two different breeds: dam and sire (A_1 and A_2 from the dam, and B_1, B_2 from the sire). In such a case, the QTL model can be specified as

$$y_i = \beta_0 + x_{dij}\beta_{dj} + x_{sij}\beta_{sj} + z_{ij}\gamma_j + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (5)$$

where β_{dj} is the substitution effect of alleles A_1 and A_2 of the dam at the locus j (j=1,…,p), β_{sj} is the substitution effect of B_1 and B_2, and γ_j is the dominance effect, and the coding system for [x_{1ij}, x_{2ij}, x_{3ij}] can be specified in the following matrix (Xu 2013b):

$$\begin{vmatrix} +1 & +1 & +1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \end{vmatrix} \begin{array}{l} \text{for genotype } A_1B_1, \\ \quad \text{for } A_1B_2, \\ \quad \text{for } A_2B_1, \\ \quad \text{for } A_2B_2. \end{array}$$

Note that the standard association mapping model in (2) is a special case of (5) where one cannot separate the allele A1 from A2 (or B1 from B2), and hence, β_j=α_j. In this sense, the model (5) has the benefit that it yields extra information about the sources of the observed QTL effects. However, the model (5) requires the knowledge of parental phasing, which is difficult to acquire in practice. Therefore, its application has been limited to certain experimental crosses (Xu 2013b).

Multi-locus model and LASSO

The single-locus mixed model (3) can easily be extended to a multiple regression problem by including all SNPs in the data in the same model:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + u_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (6)$$

Here the effect size β_j of the jth SNP is conditional on the effects of all other SNPs, which is different from the marginal effect size estimated by equation (3). Note that other kinds of single locus linear models as defined by Equations (2), (3) and (5), can be extended to a multi-locus context in a similar fashion by adding all the covariates (SNPs) into the same model.

When the number of SNPs p is larger than the number of individuals n, simultaneous estimation of the effects of multiple SNPs is intractable with the standard maximum likelihood. However, penalized regression, known as mixed LASSO (Wang et al. 2011), can handle this kind of high dimensional problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T \mathbf{K}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|, \qquad (7)$$

where $\mathbf{y}$ is a vector of the phenotype data y_i, $\mathbf{X}$ is the design matrix of genotypes x_{ij}, and $\boldsymbol{\beta}$ is the vector of the SNP effects β_j, and $\mathbf{K} = \sigma_g^2\mathbf{A} + \sigma_e^2\mathbf{I}$. The penalized term $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ (λ>0) shrinks the regression coefficient towards zero, keeping only a small number of SNPs with large effects in the model, excluding the likely irrelevant ones. As in the single locus model case, an EMMAX style algorithm (Kang et al. 2010) can be applied to first obtain REML estimates of the variance components as $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ based on an intercept model, and then fix the matrix to be $\hat{\mathbf{K}} = \hat{\sigma}_g^2\mathbf{A} + \hat{\sigma}_e^2\mathbf{I}$ in (7). Let $\tilde{\mathbf{y}} = \mathbf{K}^{-1/2}\mathbf{y}$ and $\mathrm{MVN}(\mathbf{0},\sigma_g^2\mathbf{A})$, and the Equation (7) becomes equivalent to

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|, \qquad (8)$$

which is the standard LASSO problem (Tibshirani 1996). An efficient coordinate descent algorithm (e.g. Friedman et al. 2010) can be applied to solve (8).

Several high dimension inference approaches have been proposed to conduct multiple testing on the basis of the LASSO estimates. Stability selection (Meinshausen and Bühlmann 2010) is a sampling-based approach similar to bootstrapping. In every run, it randomly sub-samples half of the individuals from the whole dataset, and performs LASSO regression on this partial data to select a set of SNPs. By repeating this procedure thousands of times, the selection probabilities of the SNPs are calculated, and a significance threshold can be derived to control for the multiplicity from the perspective of both false discovery rate and family-wise error. The benefits of stability selection over other approaches such as the de-biased LASSO method (Javanmard and Montanari 2014; Li et al. 2017) is that it can be efficiently used also on very large data sets. Therefore, in the following, we use the stability selection to compare the SNP- and Cluster-based approaches for multi-locus association testing.

Linkage disequilibrium network clustering

Association testing of groups of linked SNPs, rather than individual SNPs, starts with division of SNP data into units according to physical or linkage map information. We consider a simple window approach in which each chromosome is divided into many non-overlapping regions with roughly equal sized genomic segments. Window breakpoints are placed where LD (as estimated by r^2; function 'snpgdsLDMat'; R-package: 'SNPRelate'; Zheng et al. 2012) between adjacent SNPs is less than a threshold value (*LD1*) for ten consecutive SNPs in a row i.e. these regions mark putative recombination hot spots. When LD breaks down gradually along chromosomes, this result in 'long and elongated clusters', where LD between physically adjacent loci is high but the first locus in such clusters will not be in high LD (correlated) with the last locus (Fig. S1a, Supporting Information). Therefore, a complete linkage hierarchical clustering tree (using 1-r^2 as the distance measure; function '*hclust*' in R-package '*stats*'; R, core team) is constructed within each window, where clusters are extracted when the minimum LD between any pair of loci in the cluster is ≥ *LD1*. This breaks up 'long and elongated' clusters to 'spherical' clusters where all loci are interconnected by high LD (Fig. S1a, Supporting Information, see also documentation for R-function '*hclust*'). Such clusters can thus potentially be considered as independent units in a GWAS. For loci in clusters where median r^2 (between all pairwise loci within the cluster) is nevertheless > *LD2*, a second clustering step is performed. This time, the minimum r^2 between any pair of loci in the cluster is required to be ≥ *LD2.* All loci not part of clusters meeting this requirement are considered independently in a subsequent GWAS ('singleton-clusters'). This produces few but highly interconnected clusters (or individual SNPs), where all multi-locus clusters are compact and spherical (Fig. S2, Supporting Information) with median r^2 above *LD2,* (each containing a unique set of highly correlated SNPs), and all singleton-clusters are not in high LD with any adjacent SNPs within its window (Fig. S1a, Supporting Information).

For loci in each LD-cluster, we then apply a principal component analysis (PCA; Patterson et al. 2006), and extract the first few principal components (PCs) that captured the largest portion of variation (PCs explaining at least a threshold value, *PC,* of the total genetic variation in each LD-cluster) in the original data, and replace the original SNP data in the QTL model with these PCs (except for singleton-clusters which remain at their original state). With high threshold values for LD (producing many clusters with high LD), we expect most of the genetic variation to be explained by the first PC. However, when LD threshold values are low (producing fewer clusters with lower mean LD and with higher numbers of loci in each), the PCA step ensures that most of the genetic variation from each LD-cluster is still captured. The window-based regression model (also known as a "principal component regression", e.g. Hastie et al. 2009) can be formally defined as:

$$y_i = \theta_0 + \sum_{l=1}^{m_k} W_{il}\theta_l + u_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (9)$$

and

$$y_i = \theta_0 + \sum_{k=1}^{M}\sum_{l=1}^{m_k} W_{il}\theta_{lk} + u_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,\sigma_e^2), \qquad (10)$$

as single and multi-locus models, respectively. The notation W_{il} (l=1,…,m_k) represents the PCs in the *k*th window (k=1,…,*M*; *M* is the total number of the windows), θ_0 is the intercept, and θ_{lk} is the regression parameter of the given PC, u_i is the random effect defined in the same way as in (3) and the kinship matrix is calculated as in (4) using the original SNP data.

The same type of single-locus mixed model (or mixed LASSO) estimation procedure introduced above can be applied to solve Equations (9) and (10). Since in each window the multiple PCs represents a group of correlated SNPs likely to explain similar kinds of phenotypic variation, these PCs in the same window can be tested together instead of being tested individually. In this way, the total number of hypothesis tests is significantly reduced compared to the standard association mapping. In the single-locus mapping, the group testing is conducted with an *F*-test to compare a null intercept model with model (9) separately for every genomic region. In the multi-locus mapping, the stability selection can also be extended to calculate the selection probabilities of a group of variables. More technical details can be found in Appendix S1 (Supporting Information).

Arabidopsis thaliana data set

The *A. thaliana* GWAS data set originates from Baxter et al. (2010), who used it to identify genetic variants associated with leaf sodium accumulation. A total of 337 individuals were genotyped using an Affymetrix SNP array to generate around 250 000 SNPs as described in Atwell et al. (2010). After removing SNPs with a minor allele frequency < 0.05 done by using our own R script, 200 121 SNPs distributed over five chromosomes of 18-30 Mb in length remained to be used here. We used two sets of threshold values for LD-clustering: *low*, with *LD1*=0.1 and *LD2*=0.3 and *high*, with *LD1*=0.3 and *LD2*=0.5. The threshold value for the subsequent PC regression step was kept at 80% for both sets of analyses. To reduce the computational burden of LDn-clustering, based on putative recombination hot spots (see above) window break points were chosen such that window size was approximately 1000 SNPs, and pairwise r^2 values were only calculated within a window size of 100 SNPs (as LDna requires a pairwise a matrix of all r^2-values for each window the remaining values were set to 0).

Pungitus pungitus data set

The *P. pungitus* F_2 inter-population cross data of 283 individuals was originally generated by crossing a female from the Baltic Sea (Helsinki; 60°13’N, 25°11’E) and a male from a northeastern Finnish pond (Rytilampi; 66°23’N, 29°19’E). Detailed information about the origin, maintenance, genotyping and phenotyping of the crosses can be found from earlier publications (e.g. Laine et al. 2013; Yang et al. 2016; Li et al. 2017).

The RAD sequencing data used by Yang et al. (2016) and Li et al. (2017) were also used in this work, but the linkage mapping was re-conducted using the latest development of the LepMAP software: Lep-MAP3 (LM3; Rastas, 2017). A notable benefit of LM3 is its efficiency in inferring the parental/grandparental phase based on the dense SNP data, and this generates an opportunity to utilize the four-way cross QTL mapping (5). The input data was obtained by using the LM3 pipeline, first mapping individual fastq files to the genome using bwa mem (Li, 2013) followed by SAMtools mpileup (Li et al., 2009), and then running LM3 scripts pileupParser.awk and pileup2posterior.awk using the default parameters.

The mapping was done following the basic LM3 pipeline: First, ParentCall2 was used on the data of offspring, parents and grandparents. Then Filtering2 module was used with dataTolerance=0.001, filtering out markers segregating in a more distorted fashion than what would be expected by 1:1000 odds by chance. After this, SeparateChromosomes2 was run on the filtered data with lodLimit=75, followed by JoinSingles2All with lodLimit=60 and lodDifference=10 yielding 21 linkage groups with a total of over 89 000 markers assigned to these groups.

Finally, the markers were ordered within each linkage group with OrderMarkers2 module with default parameters. OrderMarkers2 was run twice on each chromosome using informativeMask=13 and informativeMask=23, removing either markers only informative in the mother or father, respectively. This created two maps for each chromosome, one having more maternal markers and the other having more paternal markers, both having on average 2/3 markers in common. The justification for constructing two maps is to remove the effect of markers informative only in one parent, as markers informative in different parents are not informative when compared against each other.

The phased data used for QTL analysis was the output from OrderMarkers2 with parameter outputPhasedData=1. The phases were converted into grandparental phase by first evaluating the final marker orders with option grandparentalPhase=1 and then matching the (parental) phased data with the grandparental one using phasematch.awk script of LM3. Thus, the parental phases were inverted, when needed, to obtain the grandparental phases for all markers. The only manual step involved removing clear errors from map-ends based on scatter plots of physical and map positions (Chakravarti 1991).

Ultimately, 278 individuals (5 individuals were found to be duplicated in the original data sets, and were therefore removed) genotyped for 74 078 SNPs distributed over 21 chromosomes with 66-111 cM (corresponding to 15-41 Mb in the physical map) length were used in the study. We used the combined map of the males and females to estimate r^2 for the LDn-clustering with a threshold value of 0.7 for both *LD1* and *LD2*. The threshold value for PC regression was set to 80%. Furthermore, we considered each chromosome as a window, and due to the much higher overall LD in this data set than in the *A. thaliana* data (Fig. S3, Supporting Information) we used all pairwise r^2-values within 2kb windows. For illustrative purposes, we focused on one particular quantitative trait: total lateral plate number analyzed earlier by Yang et al. (2016).

Simulation study - subsets of data

To investigate the effect of threshold values (*LD1*, *LD2* and *PC*) used for LDn-clustering on the power to detect significant QTL by GWAS, we simulated a region containing 300 polymorphic SNPs regions corresponding to 50 SNPs down- and up-stream of the most significant SNP in the *Arabidopsis* data set (corresponding to a 122 kb region spanning bps 6373268-6495751 on Chr4; see Results) with four combinations of threshold values (0.1;0.3;0.8, 0.3;0.5;0.8, 0.1;0.1;0.8 and 0.1;0.1;0.9, with values separated by ';' representing *LD1*, *LD2* and *PC* thresholds, respectively).

However, for these simulations any random set of 300 consecutive polymorphic SNPs would have sufficed. Two of the first combinations were the same as those used for the original data set and in the two latter, *LD2* was further reduced to 0.2 but with two different *PC*-thresholds: 0.8 and 0.9. This was done to investigate how over-merging LD clusters (when LD thresholds are low) potentially can be compensated by extracting more PCs during the PC regression step. For each data set a phenotype was generated on the basis of the multiple-locus model in Equation (6). The effect size of five QTL were independently simulated from a normal distribution N(0,1). The random effect **u** is simulated from a multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \sigma_g^2\mathbf{A})$, with $\sigma_g^2 = 10$, and the residual error is simulated from a normal distribution N(0,1) with narrow sense heritability, h^2, between 0.2 and 0.3. The five QTL were either randomly chosen among the 300 SNPs (*random*) or within a window of 50 bps (*clustered*). As the main aim of these simulations was to compare the power to detect significant QTL with the SNP- and LD-cluster based approaches in a small data set, the data set size and generation of phenotypic values were not aimed to necessarily be biologically realistic. Analyses were performed on 1000 sets of simulated phenotypic values for the four threshold settings as well as for a data set where each SNP was analysed independently ('*no clustering*'). EMMAX was used for the GWAS analyses as described above. Statistical power was estimated as the proportion of significant QTL among all 5×1000 causal SNPs in the simulated data sets, after Bonferroni correction for multiple testing (performed separately for each simulated data set). Confidence intervals for the proportion of significant QTL was estimated as the 95% quantiles from 1000 bootstrap replicates. False negative rates for LDn-clustered data sets were well below 0.05% for all threshold settings and were thus not considered further here. False negative rates for the *no clustering* data sets were not considered either as we would have needed to take into account that non-causal loci can be significant also due to LD, and thus defining false negatives would have been somewhat arbitrary.

For the *P. pungitus* genome we focused on a single chromosome (chromosome I) using the same clustering approach as for the full data set and compared it to a data set where each SNP was analysed independently (using all 4344 SNPs from chromosome I). We simulated phenotypes based on a single QTL with h^2 of 0.1, 0.025 or 0.05 and estimated statistical power as the proportion of data sets (n=1000) where the QTL was significant after Bonferroni correction for multiple testing. With the high power of QTL mapping in experimental crosses, such low heritabilites were necessary to discriminate between the SNP- and cluster-based methods with respect to the power to detect QTL. Bootstrap confidence intervals were estimated as above. For the above two simulations we also recorded the time to perform the EMMAX GWAS analyses on a 64-bit Windows 7 desktop computer with a 3.4-GHz Intel (i7) CPU and 32.0 GB of RAM.

Simulation study - genome wide data

The purpose of this simulation study was to evaluate and compare the performance of single- and multi-locus approaches combined with SNP or LD-cluster based genome-wide data. The simulation was based on the full genotype data set of *A. thaliana*. The LDn-clustering was conducted with the parameter *LD1*=0.3 and *LD2*=0.5 (*high*) to divide genome into 90496 LD-clusters, each considered as a locus. First, a single SNP not in high LD with any other loci (singleton-cluster) at the position 6932kb of Chr 4 was chosen as a QTL (QTL1), and its effect size was simulated from N(35,1), a normal distribution of mean 35 and variance 1, with 20%-30% heritability. Second, a single QTL (QTL2) was selected from a LD-cluster containing 20 correlated SNPs (16543kb-16517kb from Chr 4), and the effect size was simulated from N(20,1) explaining 20%-30% of the total phenotypic variation. Third, in an LD-cluster of 14 correlated SNPs (4663 kb-4658 kb) from Chr 2, five weak effect QTL (QTL3) were randomly chosen, and their effect sizes were simulated from a normal distribution N(5,1) with 5-10% heritability. This represents a scenario where adjacent QTL, in addition to being correlated, also individually explain some portion of the total phenotypic variation and is thus a more complex scenario compared to a single QTL correlated with nearby SNPs (QTL2) The random effect and residuals were simulated from MVN(0,100) and N(0,100), respectively for

50 replicate data sets with which the performance (proportion significant QTL and number of false positives) of SNP- and cluster-based single- and multi-locus methods were tested.

Results

LDn-clustering

For the *A. thaliana* data set the *low* and *high* threshold settings for LDn-clustering (0.1;0.3;0.8 and 0.3;0.5;0.8, respectively) reduced the number of independent tests in GWAS from 200,121 SNPs (original data set) to 57 148 and 90 496 clusters, respectively. Figure 1 shows examples of clustering solutions (upper panel) for *low* and *high* data sets; the heatmaps (lower panel) show that LDn-clustering can identify overlapping sets of loci in high LD when the LD pattern is highly mosaic-like. Figure S3a and b (Supporting Information), show examples of network representation of the clustering solutions for *low* and *high* data sets, respectively. The number of SNPs per cluster were Gamma distributed (Fig. S4, Supporting Information) with most clusters being singleton-clusters (51% and 67%, for *low* and *high* data sets, respectively) and few clusters containing many SNPs (up to 71 for both *low* and *high* data sets). Figure S5 (Supporting Information) shows the relationship between the proportion of genetic variance explained in each cluster by the first (upper panel) and the second (lower panel) PCs. This demonstrates that the higher the median LD in a cluster the more the first PC explains of the total genotypic variation in that cluster.

For the *Arabidopsis* data set, the first PC explained >80% of the variation in 73% and 97% of the clusters (for *low* and *high*, respectively), thus only one PC was extracted from these. In no cluster was it necessary to extract more than two PCs to explain at least 80% of the total genetic variation in each cluster (Fig. S5, Supporting Information).

In the *P. pungitus* data, LDn-clustering reduced the number of tests in GWAS from 75 484 to only 214. Because of the high LD in the experimental cross, the first PC from each cluster explained on average 97% of the genetic variation in each cluster (i.e. well above the *PC* threshold of 80%). LDn-clustering produced between eight and 14 clusters from the 21 chromosomes (mean=10.8), with each cluster containing on average 353 SNPs (range 40-1 858, with an outlier of only six SNPs for an LD-cluster on chromosome 12; Fig. S5, Supporting Information). Examples of network representation of LDn-clustering for *P. pungitus* chromosomes are shown in Figure S3c and in Figure S6 (Supporting Information).

Simulation study

In the simulated data based on 300 SNPs from the *A. thaliana* data set, the number of clusters and PCs extracted by the four different threshold settings for LDn-clustering are summarized in Table 1. There was no effect of these threshold settings on the power to detect significant QTL (Fig. 2a) using a single-locus approach. However, there was a moderate improvement in computational time between clustered and non-clustered data. For example, GWAS for LDn-clustered data with threshold settings 0.1;0.3;0.8 was on average 1.9 times faster than for non-clustered data (Fig. 2b). In contrast, in the *P. pungitus* data, power to detect significant QTL with clustered data was considerably higher than in non-clustered data when heritabilities were very low (h^2=0.01-0.025; Fig. 2c). In addition, for this F_2-generation experimental cross, GWAS analyses were on average 28 times faster in clustered data compared to non-clustered data (Fig. 2d). Note also that increasing the *PC* threshold from 0.8 to 0.9, increased the total number of PCs extracted from the data set (from 130 to 140), but not the total number of LD-clusters (Table 1).

Three different QTL effects were simulated in the genome-wide *A. thaliana* SNP data set. All methods (single- and multi-locus approaches using SNP- and LD-cluster-based analyses) detected significant QTL in >98% of the simulated data sets when large-effect QTL were simulated either in a singleton-cluster (loci not in high LD with any adjacent loci; QTL1) or a multi-locus cluster (a set of correlated SNPs from an LD cluster; QTL2; Table 2). However, when five linked QTL with smaller effects were simulated within a multi-locus cluster (QTL3), the performance of GWAS was lower. Among the methods, the multi-locus approach combined with LDn-clustered data shows the highest power (46% of QTL detected), followed by GWAS on single-locus SNP data (38% of QTL detected). The multi-locus method also illustrated better ability to control the number of false positives than the single-locus approach (Table 2).

Analysis of leaf sodium accumulation in A. thaliana

The standard SNP-based single-locus association mapping with Bonferroni correction identified 23 significant SNPs, with 22 located in Chr4 (ranging from 6381929 bp to 7581539 bp in the *A. thaliana* genome), and a single SNP located in Chr3 (18095036 bp; Fig. 3a). The permutation test identified 28 SNPs located in the same genomic regions as the Bonferroni test (Fig. 3c). The multi-locus approach identified only three significant SNPs in Chr4 (located at 6392280, 6418442 and 6742032 bp, respectively; Fig. 3e), which are a subset of the SNPs detected by the single locus mapping.

The cluster-based single-locus mapping (data generated with the parameter *LD1*=0.3 and *LD2*=0.5) with Bonferroni and permutation tests detected four, six and 21 significant genomic regions in Chr4, respectively (Fig. 3b, d). The window-based multi-locus approach identified one region (6415034-6418442 bp) and two singleton QTL at 6392280 and 6455695 in the same chromosome (Fig. 3f). For all the methods, the signal with the highest statistical significance was detected at the SNP located at 6392280 bp of Chr 4.

Analysis of P. pungitus data

In the QTL analysis of the *P. pungitus* data, the SNP-based single-locus approach with Bonferroni correction did not identify any significant loci (Fig. 4a; Fig. S8a, Supporting Information). This was also the case in the multi-locus analysis (Fig. 4e; Fig. S8e, Supporting Information). In contrast, the permutation test based on the single-locus mapping identified multiple significant loci in three chromosomes (Chr 9, 20 and 21) when the allele substitution and dominance effects were tested in a group (Fig. 4c). In separate testing of the allele substitution effects, a number of loci in Chr 9, 20, and 21 were identified as having significant allele substitution effects from the grandfather, and Chr

3, 6 and 8 having significant allele substitution effects from the grandmother (Fig. 4c). In the previous study, Yang et al. (2016) detected only two QTL (in Chr. 20 and 21) using the MapQTL software (Van Ooijen 2009).

When the QTL analysis was used to test the allele substitution and dominance effects jointly in the same model using the LD-cluster-based approach, single-locus mapping with Bonferroni correction identified two significant regions in Chr 20 (28-40cM) and 21 (32-53cM), respectively (Fig. 4b). When the effects were tested separately, Chr 20 and 21 were detected for the grandfather alleles, and Chr 8 for the grandmother alleles (Fig. S8b, Supporting Information). Permutation tests identified significant regions in the same chromosomes as the Bonferroni tests, but the former detected more genomic regions in each chromosome (Fig. 4d & S8d, Supporting Information). Finally, the stability selection approach identified only a single significant region in Chr 8 (Fig. 4f & S8f, Supporting Information).

Discussion

We have proposed a cluster-based gene mapping approach for analyzing quantitative traits that can be used with both single-locus and penalized regression-based multi-locus methods to conduct association tests. This approach uses network analyses to group (potentially physically overlapping) loci in high LD into clusters within non-overlapping windows. This approach is very general: it can be applied to various gene mapping problems, including data collected from the wild with unknown population structure, as well as data from F_2-generation experimental crosses (both inbred and outbred) by using slightly different model structures, but the same kind of parameter estimation and hypothesis testing methods. Even when only a draft genome is available, LDn-clustering could be performed separately for the available scaffolds.

Previous window-based approaches using equal sized windows (Xu 2013a) have been criticized, because they may accidently divide a meaningful region into separate adjacent windows, potentially resulting in the loss of power in QTL detection (e.g. Beissinger et al. 2015). This is solved in LDn-clustering by placing window-breakpoints in regions of low LD (lower than used for LD-clustering), which produces non-equal-sized windows. However, the main advantage of LDn-clustering is in its ability to distinguish many overlapping sets of SNPs in high LD interspersed along a chromosomal region. Thus, it can handle LD patterns that are highly mosaic-like where it would not otherwise be possible to define non-overlapping haplotype blocks without also grouping SNPs that are not connected by high LD. LDn-clustering is robust against threshold settings for clustering because in the event of over-merging of LD-clusters (due to too low LD-thresholds), the subsequent PC regression step will still ensure that most of the genetic variation from each cluster is captured. The two steps in LDn-clustering (LDn-clustering and PC regression) perform in some respect similar tasks; median LD in a cluster is positively correlated with the amount of genetic variation explained by the first PC (Fig. S5, Supporting Information). Thus, where LD-clusters produce more than one PC (the first explaining less than the threshold value *PC*), increasing LD threshold-values for those clusters would produce sub-clusters where the first PC is likely to explain a higher proportion of the total genotypic variance. The *low* and *high* threshold settings for the *A. thaliana* data set exemplifies this: *low* settings produced fewer clusters with more PCs compared to the *high* setting (Fig S5, Supporting Information). Since in our GWAS approach each cluster constitutes an independent test (rather than each PC), using lower LD-threshold settings are in theory expected to produce a stronger association test. However, the conducted simulations (Fig. 2a) show that the power to detect significant QTL did not differ between any of the four LDn-clustering threshold settings (two with even lower *LD2*-threshold values compared to *low*), and hence, this effect is likely to be marginal. Nevertheless, it may be easier to interpret data using high LD-threshold values, since in most cases, one PC is enough to explain most of the genetic variation in the resulting LD-clusters, yielding a reduced number of (more strongly correlated) SNPs for downstream analyses.

The impact of LD on SNP-based gene mapping

The performance of conventional SNP-based single- and multi-locus approaches is influenced by the LD pattern of the data. In the case of the *A. thaliana* GWAS data set with a fast LD decay over the genome, the single-locus mapping with either Bonferroni or permutation tests identified a similar set of more than 20 SNPs in the same genomic regions in Chr 3 and Chr 4. In contrast, the LASSO based multi-locus approach only identified three SNPs in Chr4. One of them (Chr4: 6392280) is located within the region of the gene AtHKT1_1: (Chr4:6391984–6395877), which has been shown to be functionally associated with sodium leaf accumulation in *A. thaliana* (Baxter et al. 2010). This difference between single and multi-locus mapping results can be explained by the fact that the multi-locus method relies on conditional hypothesis testing. When the strength of the association for a single SNP is tested, all other correlated SNPs' associations have already been accounted for. Therefore, the multi-locus test is stricter than the single-locus test.

In the bi-parental *P. pungitus* data with high levels of LD extending considerable distances over the linkage map, the Bonferroni correction became too conservative to identify any significant SNPs. This was expected: Bonferroni becomes overly conservative when the multiple tests are positively correlated with each other (Goeman and Solari 2014). In contrast, the permutation test, which can effectively account for the correlation structure in the data, was still able to identify a number of significant loci with the significant allele assignable to one of the grandparents. That the detected QTL had allele substitution effects from the grandfather (originating from the pond population), but not from the grandmother, indicates that the grandparental genotypes were AB and AA, respectively, and the allele 'B' originating from the pond environment caused the phenotypic variation observed in the F_2 generation. The four-way cross model applied here was able to detect more significant QTL for the focal trait than the MapQTL approach applied to the same data by Yang et al. (2016). In addition, the four-way cross model helps elucidate from which population the allele effects on the phenotypes originate from.

The multi-locus mapping with the original SNP data also failed to identify any significant QTL in the *P. pungitus* SNP data. One possible explanation is that the widely used coordinate descent algorithm used to solve the LASSO penalized regression may work poorly and converge extremely slowly for highly correlated data sets (Kim et al. 2016). Another possible reason is that the stability selection as a multiple testing approach involves a data sub-sampling step, which may result in reduced statistical power when the sample size is small. Regarding the hypothesis tests, a de-biased LASSO approach (Javanmard and Montanari 2014; Li et al. 2017) can be performed on the whole data set without any re-sampling of the data, and therefore might have better power to detect QTL. Unfortunately, we discovered that the de-biased LASSO could not be applied to this high dimensional data set with over 200 000 regression parameters due to its high computational cost. Nevertheless, as discussed below, the de-biased LASSO can easily be applied to the LDn-clustered data set.

Cluster-based gene mapping

In general, the LD-cluster-based approach shows higher or equivalent ability to identify significant QTL than the more conventional methods in the *A. thaliana* and *P. pungitus* data sets, as well as in the simulated data. In the case of the *A. thaliana* data, the single locus approach (with both Bonferroni and permutation tests) identified 6-20 significant genomic regions (or singletons) in Chr 4. Those regions overlapped with the region in which the 22 significant SNPs were detected by the individual SNP-based single-locus approach. The multi-locus cluster-based approach identified one significant region, and these findings were also similar to those obtained by using the SNP based approach. This suggest that the computationally efficient cluster-based approach has similar power as the SNP-based approaches to discover QTL in a data set with fast LD decay.

In the simulated data (focusing on 300 polymorphic SNPs spanning 122 kb around the most significant QTL for sodium leaf accumulation) we saw no differences in the proportion of significant QTL between SNP-based gene mapping and cluster-based gene mapping. This was expected; due to the fast LD decay across *A. thaliana* chromosomes, the number of independent test in the GWAS was only reduced by a factor of 3.5 and 2.2, using the *low* and *high* threshold settings for LDn-clustering, respectively. However, when we simulated multiple weak QTL in an LD-cluster comprising 14 highly correlated SNPs (QTL3), using both single- and multi-locus methods, we saw the highest power in the LD-cluster-based multi-locus approach (46% of QTL detected) followed by the SNP-based single-locus (38% of QTL detected) and conventional multi-locus approach (22% of QTL detected). Hence, the cluster-based (multi-locus) approach seems to have an advantage over SNP-based approaches when multiple weak (independent) QTL are correlated within a small physical region in the genome. However, more extensive simulations are required to fully test this.

In the *P. pungitus* QTL data set, the cluster-based single-locus approach also identified the same significant genomic regions as the individual SNP-based approach. However, in contrast to the SNP-based single locus analysis, even the conservative Bonferroni test appeared to have sufficient power to identify significant QTL in this data. The multi-locus approach with stability selection identified QTL only in a single chromosome, probably due to the use of sub-sampling in the hypothesis testing procedure. In fact, by switching the stability selection to de-biased LASSO (Fig. S7, Supporting Information), the multi-locus approach generally identified the same QTL as the single-locus approach. It is also worth noting that in the *P. pungitus* QTL data, each cluster consists of on average 336 SNPs (range 6-1858), which may include hundreds of genes according to the latest version of the nine-spined genome annotation (Varadharajan S., Nederbgragt L., Jacobssen K., Guo B., Löytynoja A., Rastas P. & Merilä J., unpublished data). Therefore, it might be difficult to locate the candidate genes in this data with any QTL method due to the very high LD in the data. More precise location of the QTL regions in this data would require fine-mapping with more individuals to increase resolution within identified candidate genomic regions. Alternatively, independent GWAS data or a multi-parental data set (e.g. Kover et al. 2009) with more recombination events and better resolution could be used.

In the simulated data for *P. pungitus*, we saw a clear advantage of the cluster-based approach in detecting single QTL effects, in particular when heritabilities were low (0.01-0.025).

With higher heritability (0.05), both the SNP-based and the cluster-based methods recovered close to 100% of simulated QTL. Possibly other, simpler, LD reduction methods (see introduction) would also work well for this data set. However, with LDn-clustering, one is guaranteed to not accidentally lose any vital genetic information by e.g. naively subsampling the data set at equal distances across the genome, while simultaneously having control over how strongly correlated SNPs are required to be in each cluster (LD-threshold: *LD2*). In addition, by plotting LD networks from experimental crosses, potentially interesting cases involving micro-chromosomes or mapping errors can be detected (Fig. S6, Supporting Information).

Finally, from the computational point of view, the cluster-based approach appears to have a distinct advantage over mapping with individual SNPs. For instance, in the case of the *A. thaliana* data, the original SNP data of over 200 000 SNPs (or alleles) can be summarized with only 90 000 PCs in a high LD data set. This leads to a substantial reduction of the computational complexity. For example, conducting a permutation test on the *A. thaliana* data set takes about seven days by using 5 000 replications on a 64-bit Windows 7 desktop computer with a 3.4-GHz Intel (i7) CPU and 32.0 GB of RAM (note that computational time estimates for all the methods were implemented on a single core). Using the same set up, the cluster-based permutation test takes only 6-7 hours. The stability selection took about three hours on the *A. thaliana* data set and only 30 minutes on the *P. pungitus* data set. The de-biased LASSO approach consumed about 30 days for the clustered data set, and might take several months for the full SNP data. The cluster-based approach can also be used for other computationally intensive GWAS models such as the Bayesian LASSO (Li et al. 2011; Pasanen et al. 2015) and Elastic net (Huang et al. 2015) to improve their computational efficiency. The LDn-clustering algorithm took <20 min for the *A. thaliana* data set and <10 min for the *P. pungitus* data set, and can be parallelised over many computer clusters (each window/chromosome can be processed independently) for use in whole genome data sets where this kind of dimensionality reduction is likely to be most beneficial.

Concluding remarks and future directions

In conclusion, we have introduced and tested the performance of a cluster-based association mapping approach that appears to be able to solve, or at least reduce, some of the problems faced by existing mapping approaches. Given the high dimensionality of modern GWAS data sets, the proposed cluster-based gene mapping approach that uses LDn-clustering and PC regression as a dimensionality reduction tool should prove useful for computationally efficient QTL detection in a variety of data and model structures. Our analyses of two empirical data sets and simulated data suggest that the cluster-based association approach has three major benefits over other types of association analyses. First, it provides a significant reduction of the dimensionality of the data, therefore also in the amount of computational time. Second, the new approach appears to be more efficient in detecting QTL due to less conservative correction for multiple statistical tests. Third, the usage of independent principal components (instead of highly correlated SNPs) likely increases the numerical stability of the computation, especially in the case of the multi-locus approach. The benefits of LDn-clustering are likely to be most useful for data sets from species with small effective population sizes (LD decays slowly with physical distance) and/or large numbers of genetic markers, including whole genome data. However, more detailed simulations are needed to fully understand the pros and cons of cluster-based association mapping approaches for the multitude of different single- and multi-locus approaches that are currently available.

An interesting direction for future research would be to extend the current cluster-based association approach for analysing gene-gene and gene-environment interactions (Yi 2015). In the standard association model, inclusion of these interaction terms significantly increases the dimensionality of the data (e.g. for 200 000 SNPs, there are about 2 000 billion pairwise G×G interaction terms). Since the computational requirement of such models is currently not possible to meet, a cluster-based approach able to reduce the data dimensionality could provide a solution and make analyses of such interactions possible.

Acknowledgements

We thank Anna Santure, Örjan Carlborg and an anonymous referee for useful suggestions that greatly improved the earlier version of this manuscript. We also thank Jacquelin DeFaveri proofreading the final version of the manuscript. Our research was supported by grants from the Academy of Finland (129662, 134728 and 218343 to JM), a grant form Helsinki Institute of Life Science (HiLIFE; to JM).

References

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., …, Nordborg, M. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature, 465(7298), 627-631. doi: 10.1038/nature08800

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. Nature Review Genetics, 7, 781-791. doi: 10.1038/nrg1916

Baxter, I., Brazelton J. N., Yu D., Huang, Y. S., Lahner, B., Yakubova, E,…, Salt, D. E. (2010) A Coastal Cline in Sodium Accumulation in Arabidopsis thaliana Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1. PLoS Genetics, 6(11), e1001193. doi: 10.1371/journal.pgen.1001193

Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R., Long, A. D. (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature, 467, 587-590. doi:10.1038/nature09352.

Chakravarti, A. (1991) A graphical representation of genetic and physical maps: the Marey map. Genomics, 11(1), 219-222.

Daly, M.J., Rioux, J.D., Schaffner, S.E., Hudson, T. J., Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nature genetics*, **29**, 229–232.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A.,…,McVean, G., 1000 Genomes Project Analysis Group (2011) The Variant Call Format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Dudbridge, F., & Koeleman, B. P. C. (2004) Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. American Journal of Human Genetics, 75 (3), 424–435, 2004. doi: 10.1086/423738

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press: Cambridge.

Ernst, C. W., & Steibel, J. P. (2013) Molecular advances in QTL discovery and application in pig breeding. Trends in Genetics, 29(4), 215-224. doi: 10.1016/j.tig.2013.02.002

Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., & Xiong, M. (2013) Functional linear models for association analysis of quantitative traits. Genetic Epidemiology 37(7), 726-742. doi: 10.1002/gepi.21757

Friedman, J., Hastie, T., Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1.

Ge, T., Smoller, J. W., Sabuncu, M. R. (2016) Kernel machine regression in neuroimaging genetics. Machine Learning and Medical Imaging, 31-68. https://doi.org/10.1016/B978-0-12-804076-8.00002-5

Goeman, J. J., & Solari, A. (2014) Multiple hypothesis testing in genomics. Statistics in Medicine 33(11), 1946-1978. doi: 10.1002/sim.6082

Hastie, T., Tibshirani, R., & Friedman, J. (2009) *Elements of Statistical Learning (Second Edition)*. Springer: New York.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L.,…, & Thompson, P. M. (2011) Voxelwise gene-wide association study (vGeneWAS) multivariate gene-based association testing in 731 elderly subjects. Neuroimaging, 56(4), 1875-1891. doi:10.1016/j.neuroimage.2011.03.077.

Huang, A., Xu, S., & Cai, X. (2015) Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity, 114(1), 107–115. doi: 10.1038/hdy.2014.79

Husby, A., Kawakami, T., Rönnegård, L., Smeds, L., Ellegren, H, & Qvarnström, A. (2015) Genome-wide association mapping in a wild avian population identifies a link between genetic and phenotypic variation in a life-history trait. Proceedings of the Royal Society B. doi: 10.1098/rspb.2015.0156

Javanmard, A., & Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research, 15, 2869–2909. http://jmlr.org/papers/v15/javanmard14a.html

Joo, J. W. J, Hormozdiari F., Han, B., & Eskin, E. (2016) Multiple testing correction in linear mixed models. Genome Biology, 17:62. doi: 10.1186/s13059-016-0903-6

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., Eskin, E. (2008) Efficient control of population structure in model organism association mapping. Genetics 178(3), 1709–1723. doi: 10.1534/genetics.107.080101

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. Nature Genetics, 42(4), 348–354. doi: 10.1038/ng.548

Kemppainen, P., Knight, C.G., Sarma, D.K. *et al.* (2015) Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular ecology resources*, **15**, 1031–1045.

Kim, B., Yu, D., Won, J-H. (2016) Comparative study of computational algorithms for the Lasso with high-dimensional, highly correlated data. Applied Intelligence, in press. doi: 10.1007/s10489-016-0850-7.

Korte. A., & Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods, 9, 29. doi: 10.1186/1746-4811-9-29

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., & Mott, R. (2009) A Multiparent advanced generation inter-Ccoss to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genetics 5(7), e1000551. doi:10.1371/journal.pgen.1000551

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004) *Applied Linear Regression Models*. New York: McGraw-Hill.

Li, J., Das, K., Fu, G., Li, R., & Wu, R. (2011) The Bayesian lasso for genome-wide association studies. 15, 27(4), 516-523. doi: 10.1093/bioinformatics/btq688

Laine, V. N., Shikano, T., Herczeg, G., Vilkki, J., & Merilä, J. (2013) Quantitative trait loci for growth and body size in the nine-spined stickleback *Pungitius pungitius* L. Molecular Ecology, 22 (23), 5861-5876. doi: 10.1111/mec.12526

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-Prints. https://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., March, G., Abecasis , G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, Z., & 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. doi: 10.1007/s00122-012-1892-9

Li, Z., Guo, B., Yang, J., Herczeg, G., Gonda, A., Balázs, G., Shikano, T., Calboli, F. C. F., & Merilä, J. (2017) Deciphering the genomic architecture of the stickleback brain with a novel multi-locus gene-mapping approach. Molecular Ecology, 26(6), 1557-1575. doi: 10.1111/mec.14005

Liang, Y., & Kelemen, A. (2008) Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. Statistics Surveys, 2: 43-60. doi: 10.1214/07-SS026

Mackay, T. F. C., Stone, E. A., Ayroles, J. F. (2009) The genetics of quantitative traits: challenges and prospects. Nature Review Genetics, 10(8), 565-577. doi: 10.1038/nrg2612

Meinshausen, N., & Bühlmann, P. (2010) Stability Selection. Journal of the Royal Statistical Society: Series B, 72(4), 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Morgenthaler, S., Thilly, W. G., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 615: 28-56. doi: 10.1016/j.mrfmmm.2006.09.003

Patterson, N., Price, A. L., & Reich, D. (2006) Population structure and eigenanalysis. PLoS Genetics 2: e190. doi: 10.1371/journal.pgen.0020190

Purcell, S., Neale, B., Todd-Brown, K, Thomas, L., Ferreira, M. A. R., Bender, D, …, & Sham, P. C. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81, 559-575.

van Raden, P. M. (2008) Efficient methods to compute genomic predictions. Journal of Dairy Science, 91, 4414-4423. doi: 10.3168/jds.2007-0980

R Core Team (2014) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rastas, P. (2017) Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. Bioinformatics, 33, 3726-3732. doi: 10.1093/bioinformatics/btx494

Pasanen, L., Holmström, L., & Sillanpää, M. J. (2015) Bayesian LASSO, scale space and decision making in association genetics. PLoS ONE 10: e0120017. doi: 10.1371/journal.pone.0120017

Shaffer, J. P. (1995) Multiple hypothesis testing. Annual Review of Psychology, 46, 561–584. doi:10.1146/annurev.ps.46.020195.003021

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature Genetics, 44, 825-830. doi:10.1038/ng.2314

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58, 267–288. https://statweb.stanford.edu/~tibs/lasso/lasso.pdf

Van Ooijen JW (2009) MapQTL v. 6.0: Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma BV, Wageningen, The Netherlands.

Westfall, P. H., and Young, S. S. (1993) *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Statistics. New York.

Xu, S. (1996) Mapping quantitative trait loci using four-way crosses. Genetics Research, 68(2), 175-181. doi: 10.1017/S0016672300034066

Xu, S. (2013a) Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195(3): 1103-1115. doi: 10.1534/genetics.113.155309

Xu, S. (2013b) *Principles of Statistical Genomics*. Springer: New York.

Yang, J., Guo, B., Shikano, T., Liu, X., Merilä, J. (2016) Quantitative trait locus analysis of body shape divergence in nine-spined sticklebacks based on high-density SNP-panel. *Scientific Reports*, **6**, 26632. doi: 10.1038/srep26632

Yi, H., Breheny, P., Imam, N., Liu, Y., & Hoeschele, I. (2015) Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics*, 199: 205-222. doi: 10.1534/genetics.114.167817

Yi, N. (2010) Statistical analysis of genetic interactions. *Genetic Research*, 92 (5-6), 443-459. doi:10.1017/S0016672310000595

Yu, J., Pressior. G., Briggs, W.H. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38 (2), 203-208. doi:10.1038/ng1702

Zhang, K., Calabrese, P., Nordborg, M., Sun, F.Z. (2002) Haplotype block structure and its applications to association studies: Power and study designs. *American journal of human genetics*, **71**, 1386–1394.

Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, **28**, 3326–3328.

Data accessibility

- The *A. thaliana* SNP dataset is available at: https://github.com/Gregor-Mendel-Institute/atpolydb. The phenotype data is available from the original publication (Baxter et al. 2010).

- The *P. pungitus* phenotype data is available from Yang et al. (2016).

-The *P. pungitus* SNP data set as well as the R source codes for implementing all the statistical methods introduced in the paper will be available in Dryad upon acceptance.

-LDn-clustering is available as an additional function in an updated version of the existing LDna R-package (https://github.com/petrikemppainen/LDna/tree/v.63).

Conflicts of interest

Authors declare no conflict of interests

Figure legends

Figure 1. LDn-clustering. Shown is an example of how LDn-clustering can account for the mosaic-like pattern of LD in population genomic data by grouping loci (within windows) based on LD regardless of their physical position in the genome. Each LD-cluster has a unique colour combination [colours between (a) and (b) do not necessarily match] and line height along the y-axis (upper panel). In each LD-cluster the minimum LD between all loci in the cluster is above (a) 0.1 or (b) 0.3 and the median LD among all pairwise LD values in each LD-cluster is above (a) 0.3 or (b) 0.5. Loci not connected to any other SNPs by these thresholds are considered as independent (singleton-clusters). There are 15 and 25 unique LD-clusters in (a) and (b), respectively. Positions of the vertical lines (along the x-axis) match the positions of loci in the lower LD heatmap figure. The figure is based on 63 consecutive SNPs from *A. thaliana* data set Chr 4 (starting from SNP-position 6237655).

Figure 2. Results from simulated study with subsets of data. Panel (a) shows mean number of significant QTL (five in each simulated data set), for four different threshold settings for LDn-clustering (values in the legend separated by ';' represent threshold values *LD1*, *LD2* and *PC*, respectively) when QTL are randomly sampled among all SNPs (*Random*), or from 50 consecutive SNPs (*Clustered*) along the chromosome (h^2 = 0.2 - 0.3). Panel (b) shows the time taken to conduct GWAS on clustered (yellow) and non-clustered data (grey) for the *A. thaliana* simulated data. Panel (c) shows the proportion of significant QTL from *P. pungitus* linkage group 1 (one QTL in each data set) for different heritabilities when GWAS was performed on all 4344 SNPs (Clustering=No) or when GWAS was performed on 12 clusters produced by LDn-clustering (Clustered=Yes). Panel (d) show the time taken to conduct GWAS on clustered (yellow) and non-clustered data (grey) for the *P. pungitus* simulated data. Data are based on 1000 simulated sets of phenotypic values, and error bars in (a) and (c) represent 95% bootstrap confidence intervals (1000 bootstrap replicates).

Figure 3. Genome-wide association mapping of the *A. thaliana* data. Results of SNP- and LD-cluster-based GWA analyses are shown on the left (a, c, e) and right (b, d, f) panels, respectively. In (a) and (b), dots (blue or green coloured) indicate *p*-values from the association test calculated by single-locus mapping, and the red line represents the significance threshold (0.05) adjusted by the Bonferroni correction. In (c) and (d), dots represent the adjusted *p*-values from the permutation test in single-locus mapping, and red lines the significance threshold (0.05). In (e) and (f), dots present the selection probability calculated by the multi-locus stability selection method, and the red line represents the corresponding significance threshold (guaranteeing the expected number of false positives to be < 1).

Figure 4. Quantitative trait locus mapping of the *P. pungitus* data. Results of SNP- and cluster-based QTL analysis are shown on the left (a, c, e) and right (b, d, f) panels, respectively. The allele substitution effects of two founders and the dominance effects are tested jointly in the same model. In (a) and (b), dots (blue or green coloured) represent the *p*-values from the association test calculated by single -locus mapping, and the red curve the significance threshold after Bonferroni correction. In (c) and (d), dots represent the adjusted *p*-values (0.05) calculated by the permutation test in single-locus mapping, and red lines the significance threshold (0.05). In (e) and (f), dots present the selection probability calculated by the multi-locus stability selection method, and the red line the corresponding significance threshold (guaranteeing the expected number of false positives to be < 1).

Figures

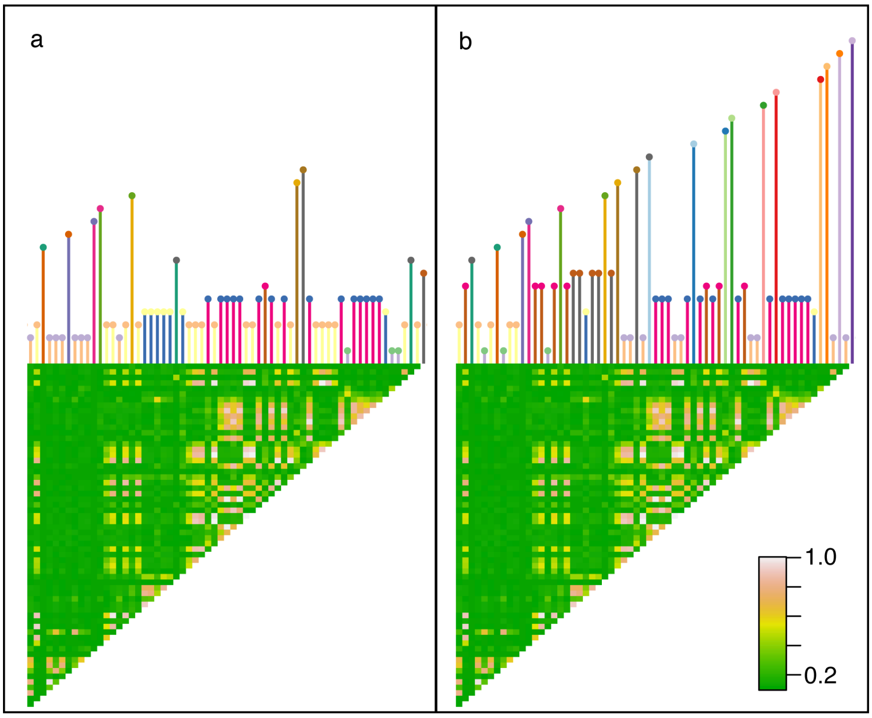
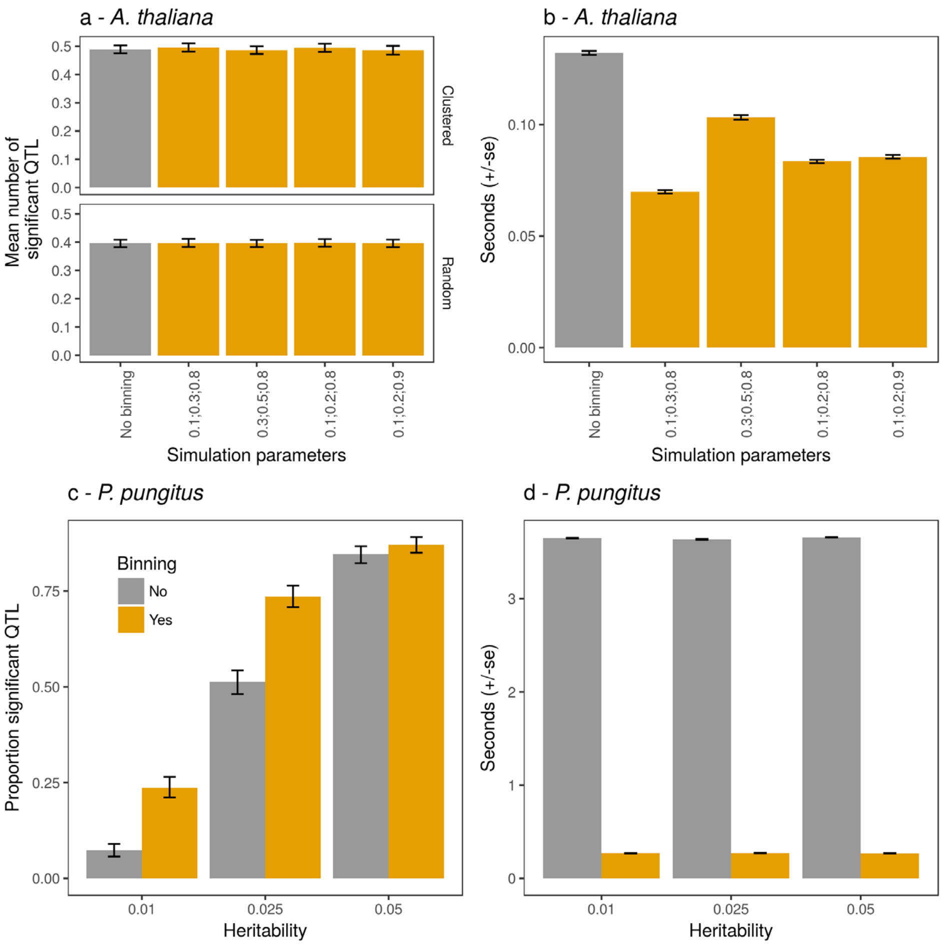


Figure 1.



a - *A. thaliana*
Mean number of significant QTL
0.0
0.1
0.2
0.3
0.4
0.5
Clustered
Random
No binning
0.1;0.3;0.8
0.3;0.5;0.8
0.1;0.2;0.8
0.1;0.2;0.9
Simulation parameters
b - *A. thaliana*
Seconds (+/-se)
0.00
0.05
0.10
No binning
0.1;0.3;0.8
0.3;0.5;0.8
0.1;0.2;0.8
0.1;0.2;0.9
Simulation parameters
c - *P. pungitus*
Proportion significant QTL
0.00
0.25
0.50
0.75
Binning
No
Yes
0.01
0.025
0.05
Heritability
d - *P. pungitus*
Seconds (+/-se)
0
1
2
3
0.01
0.025
0.05
Heritability

Figure 2.

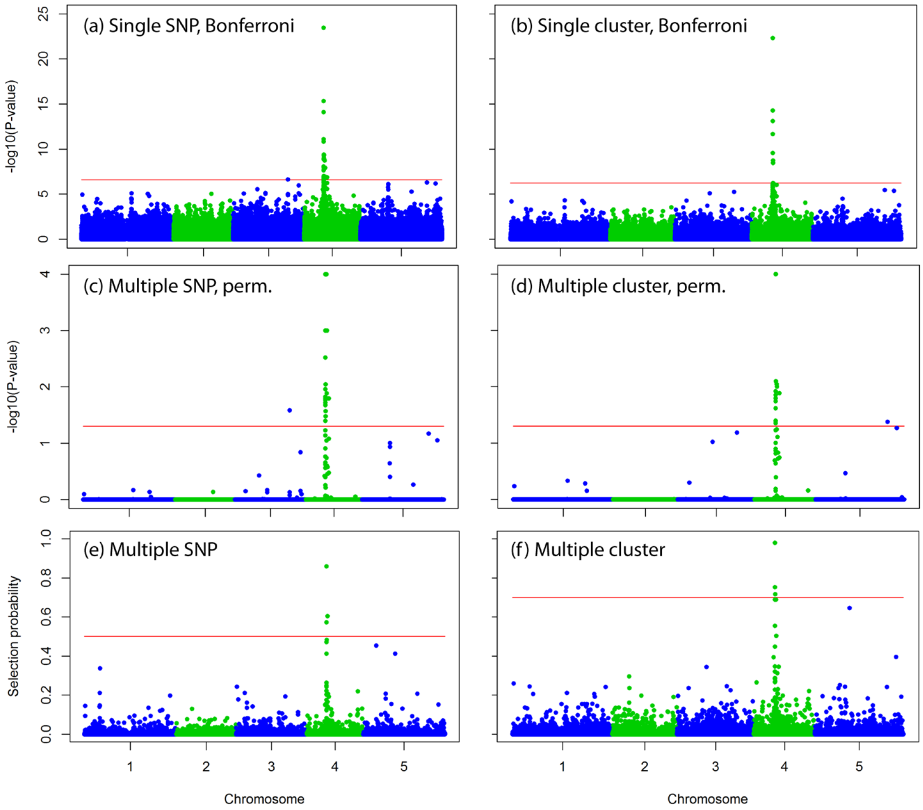

(a) Single SNP, Bonferroni
(b) Single cluster, Bonferroni
(c) Multiple SNP, perm.
(d) Multiple cluster, perm.
(e) Multiple SNP
(f) Multiple cluster
-log10(P-value)
-log10(P-value)
Selection probability
0
5
10
15
20
25
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
1
2
3
4
5
Chromosome
Chromosome

Figure 3

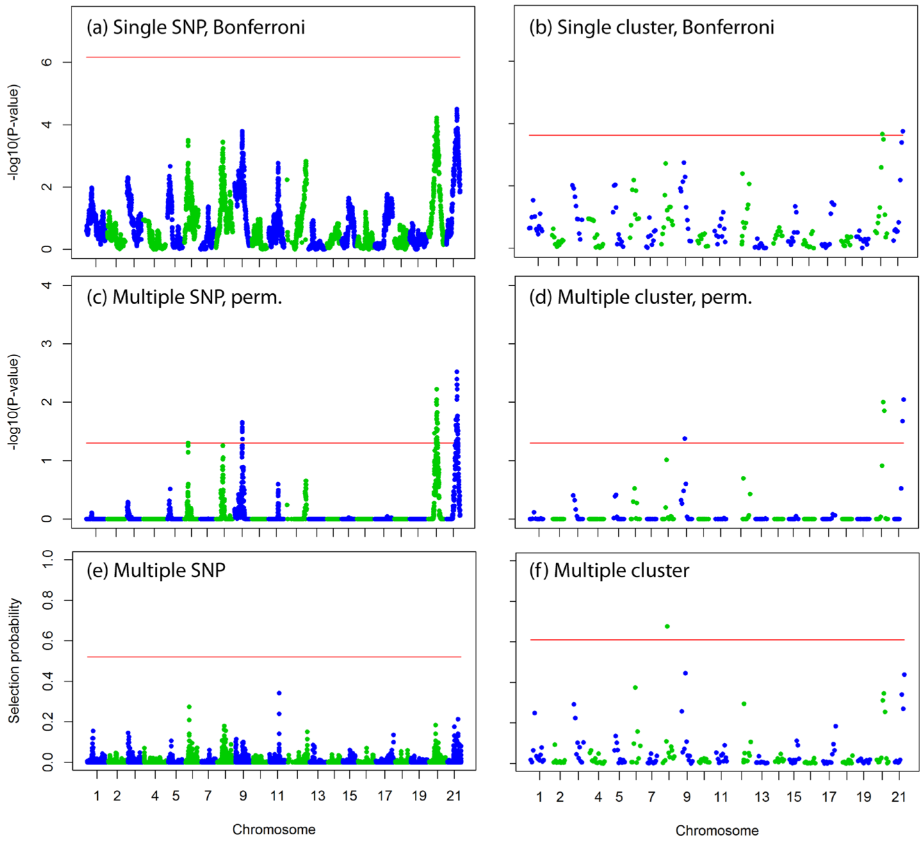

(a) Single SNP, Bonferroni
(b) Single cluster, Bonferroni
(c) Multiple SNP, perm.
(d) Multiple cluster, perm.
(e) Multiple SNP
(f) Multiple cluster
-log10(P-value)
-log10(P-value)
Selection probability
0
2
4
6
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
1 2 4 5 7 9 11 13 15 17 19 21
1 2 4 5 7 9 11 13 15 17 19 21
Chromosome
Chromosome

Figure 4

Tables

Table 1. Summary of LDn-clustering settings and results from *A. thaliana* 300 SNP simulation study.

Setting	*LD1*	*LD2*	*PC*	PCs	Clusters
1	0.1	0.3	0.8	143	111
2	0.3	0.5	0.8	172	168
3	0.1	0.2	0.8	130	85
4	0.1	0.2	0.9	140	85

LD1, *LD2* and *PC* refer to LDn-clustering threshold values used. PCs and Clusters refer to the total number of PCs and Clusters, respectively, extracted from the data.

Table 2. The average performance of single- and multi-locus QTL-mapping methods with SNP or cluster based analyses in a simulation study of genome-wide *A. thaliana* data. Number of false positives refers to average number of false positive QTL detected in simulations.

Simulated QTL	**Proportion of QTL detected by GWAS**			
	Single-locus		Multi-locus	
	SNP-based	Cluster-based	SNP-based	Cluster-based
QTL1	1	1	1	1
QTL2	1	0.98	1	0.98
QTL3	0.38	0.24	0.22	0.46
No. of false positives	2.7	1.6	0.1	0.6