<https://helda.helsinki.fi>

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

Li, Zitong

2018-07

Li , Z , Kemppainen , P , Rastas , P & Merilä , J 2018 , ' Linkage disequilibrium clustering-based approach for association mapping with tightly linked genomewide data ' , Molecular Ecology Resources , vol. 18 , no. 4 , pp. 809-824 . <https://doi.org/10.1111/1755-0998.12893>

<http://hdl.handle.net/10138/321979>

<https://doi.org/10.1111/1755-0998.12893>

acceptedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

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

3
4 **Keywords:** GWAS, quantitative trait loci, principal component regression, multi-locus method,
5 four-way cross

6
7 Zitong Li^{1*}, Petri Kemppainen^{1*}, Pasi Rastas¹, Juha Merilä¹

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

11 *Equal contribution

12
13 **Correspondence to** Zitong Li, Tel. + 358-40 8169003; E-mail: lizitong1985@gmail.com

Formatted: Finnish

14 **Abstract**

15 Genome-wide association studies (GWAS) aim to identify genetic markers strongly associated with
16 quantitative traits by utilizing linkage disequilibrium (LD) between candidate genes and markers.
17 However, because of LD between nearby genetic markers, the standard GWAS approaches
18 typically detect a number of correlated SNPs covering long genomic regions, making corrections
19 for multiple testing overly conservative. Additionally, the high dimensionality of modern GWAS
20 data poses considerable challenges for GWAS procedures such as permutation tests, which are
21 computationally intensive. We propose a cluster-based GWAS approach that first divides the
22 genome into many large non-overlapping windows, and uses linkage disequilibrium network
23 analysis in combination with principal component (PC) analysis as dimensional reduction tools to
24 summarize the SNP data to independent PCs within clusters of loci connected by high LD. We then
25 introduce single- and multi-locus models that can efficiently conduct the association tests on such
26 high dimensional data. The methods can be adapted to different model structures, and used to
27 analyse samples collected from the wild or from bi-parental F₂ populations, which are commonly
28 used in ecological genetics mapping studies. We demonstrate the performance of our approaches
29 with two publicly available data sets from a plant (*Arabidopsis thaliana*) and a fish (*Pungitius*
30 *pungitius*), as well as with simulated data.

31 Introduction

32 A central problem in quantitative genetics is to understand the relationship between genotypes and
33 quantitative traits. A Genome-wide association study (GWAS; Balding 2006; Korte and Farlow
34 2013) is a population-based approach to identify a set of candidate loci associated with complex
35 traits from a genome-wide set of genetic variants. Another closely related approach is quantitative
36 trait locus (QTL) mapping (Mackay et al. 2009), which utilizes experimental crosses or pedigree
37 data. The major difference between the GWAS and QTL approaches is that the former utilizes
38 historical recombination events, whereas the latter relies on recent recombination events to detect
39 association / linkage between genetic markers and phenotypes. Nevertheless, both approaches tend
40 to use similar types of statistical methods, such as linear regression, to identify phenotype-genotype
41 associations (Ernst and Steibel 2013). Therefore, although the main focus of this methodological
42 paper is on statistical analysis of GWAS data, we will also demonstrate how the developed
43 approach can be utilized with QTL mapping data.

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

50 Next generation sequencing techniques have provided a cost-effective access to large
51 genomic data sets, such as high-resolution SNP panels. The accessibility of such panels in GWAS
52 and QTL studies provides an opportunity to fine-map the casual loci underlying phenotypes but
53 such high dimensional data sets also pose great challenges. First, in many ecological GWAS and
54 QTL-mapping studies, sample sizes are often limited to few hundreds of individuals due to logistic
55 or budgetary limitations. However, the number of SNPs in these studies may reach hundreds of
56 thousands or even several million, creating what statisticians know as a ' p much larger than n '

57 problem (i.e. number of SNPs is much larger than the number of individuals; Hastie et al. 2009).
58 Second, another feature of large genomic data sets is that SNPs which are physically close to each
59 other are often in linkage disequilibrium (i.e. correlated). This high dimensionality and correlation
60 structure of population genomic data sets pose difficulties for both single- and multi-locus mapping
61 approaches to identify QTL (Xu 2013a). First, single-locus mapping approaches rely on multiple-
62 testing corrections to reduce the rate of false positives. The most conventional and widely used
63 approach is the Bonferroni correction (Dudbridge and Koeleman 2004), which works best when the
64 multiple hypothesis tests are independent from each other. Thus, the Bonferroni correction typically
65 becomes overly conservative when the tests are positively correlated, which is likely to be the case
66 when LD is prevalent in the data.

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

82 be gained if groups of SNPs in high LD were analyzed jointly by either single- or multi-locus
83 mapping approaches.

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

103 Recently, Kemppainen *et al.* (2015) proposed to use network analytical tools (LD network
104 analysis: LDna) to study genome wide LD-patterns in population genomic data sets. This
105 unsupervised method effectively partitions genomic data into sets of loci that have similar
106 phylogenetic signals irrespective of their physical position in the genome. As such, the LDna

107 approach could provide a useful tool for flexible dimensionality reduction in gene mapping
108 approaches utilizing large genomic datasets.

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

120 An additional novel methodological contribution of this work is that the single locus-based
121 linear regression approach of Hibar et al. (2011) was generalized to a single- and multi-locus linear
122 mixed model (LMM) context with the possibility to include a random effect to control for spurious
123 effects of population structure. Consequently, the method is suitable for analyzing data sets with
124 hidden family and population structure, including data collected from the wild. We illustrate the
125 utility of the novel approach using two publicly available data sets: 278 nine-spined sticklebacks
126 (*Pungitius pungitius*) genotyped for 74 078 SNPs (Yang et al. 2016; Li et al. 2017; Rastas et al.
127 2017), and 337 thale cresses (*Arabidopsis thaliana*) genotyped for 200 121 SNPs (Atwell et al.
128 2010; Baxter et al. 2010) as well as simulated data.

129

130 **Materials and Methods**

131 *Single-locus models for association mapping*

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

$$137 \quad y_i = \beta_0 + x_{ij}\beta_j + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad (1)$$

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

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

$$144 \quad y_i = \beta_0 + x_{ij}\beta_j + z_{ij}\gamma_j + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad (2)$$

145 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
 146 is the dominance effect, and all other notations are the same as in (1).

147 To test if a SNP is significantly associated with a trait, one can test the null hypothesis:

148 $\beta_j = 0$ against the alternative hypothesis: $\beta_j \neq 0$. Standard procedures including t - and F -tests can
 149 be used (Kutner et al. 2004). Since many hypothesis tests are simultaneously conducted, it is
 150 important to adjust the p -values (i.e. adjust the significance threshold α) to control for false
 151 positives. Bonferroni correction (Shaffer 1995) – simply adjusting the significance threshold (α) by
 152 dividing it by the number of SNPs (p ; i.e. α/p) – is a conventional and popular way to control the
 153 family wise error (FWER): the probability of having one incorrectly rejected null hypothesis among
 154 all the hypotheses (Efron 2010). The drawback of the Bonferroni correction is that the multiplicity
 155 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:

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

where the random effect u_i is specified as $\mathbf{u} = [u_1, \dots, u_n] \sim \text{MVN}(0, \sigma_g^2 \mathbf{A})$ with known $n \times n$ sized relationship matrix $\mathbf{A}$ and unknown variance σ_g^2 . The random effect $\mathbf{u}$ 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 $\mathbf{A}$ can be estimated from molecular marker information as (van Raden 2008):

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

180 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
181 individuals i and k ($i, k=1, \dots, n$) at the SNP j . Alternatively, one may also estimate the relationship
182 matrix from the known pedigree of the individuals.

183 Restricted maximum likelihood (REML) based programs such as EMMA (Kang et al.
184 2008) and EMMAX (Kang et al. 2010) have been widely used to evaluate the regression parameters
185 and variance components as described by Equation (3). The EMMA approach refers to a
186 computational procedure which uses REML to estimate the variance components repeatedly for
187 each SNP. In contrast, EMMAX estimates the variance components once based on an intercept
188 model, and then fixes them to evaluate the effect and statistical significance of the SNPs.
189 Consequently, the EMMAX approach is much faster and simpler to use on large data sets, and both
190 simulation and empirical studies have shown that the EMMAX approach can have the same
191 statistical power and ability to control for false positives than the more precise EMMA method
192 (Kang et al. 2010). Therefore, we will consider EMMAX as the default method for mixed model
193 analysis in this work.

194 In a linear mixed model, the hypothesis testing can be conducted using t - or F -tests in a
195 similar way as in the case of standard linear regression. Bonferroni correction can also be
196 straightforwardly used for multiple testing. However, the permutation test procedure used for
197 standard linear model (1) is not applicable for the mixed model. The reason is that the standard
198 permutation test randomly reshuffles phenotypes, which is equivalent to sampling phenotype data
199 from a uniform distribution, and this implementation will remove any among-individual correlation
200 from the data. Clearly, this violates the assumption of dependency structure among individuals in
201 the mixed model, and might yield spurious statistical results (Joo et al. 2016). A correct way to
202 conduct permutation tests on the basis of the mixed model would be to draw a sufficient number of
203 independent samples from a multivariate normal distribution $MVN(\mathbf{0}, \hat{\sigma}_g^2 \mathbf{A} + \hat{\sigma}_e^2 \mathbf{I})$, and then use
204 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), F₁ offspring of a hybrid cross generated from two heterozygous parents (Van Ooijen 2009), and in the case of an outbred F₂ design (Xu 2013b), there might be up to four possible alleles, A₁, A₂, B₁ and B₂ originating from two different breeds: dam and sire (A₁ and A₂ from the dam, and B₁, B₂ from the sire). In such a case, the QTL model can be specified as

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

where β_{dj} is the substitution effect of alleles A₁ and A₂ of the dam at the locus j ($j=1, \dots, p$), β_{sj} is the substitution effect of B₁ and B₂, 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):

$$\begin{array}{l} \left| \begin{array}{ccc} +1 & +1 & +1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ -1 & -1 & +1 \end{array} \right| \begin{array}{l} \text{for genotype } A_1B_1, \\ \text{for } A_1B_2, \\ \text{for } A_2B_1, \\ \text{for } A_2B_2. \end{array} \end{array}$$

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, $\beta_j = \alpha_j$. 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:

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

Here the effect size β_j of the j th 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:

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

where $\mathbf{y}$ is a vector of the phenotype data y_i , $\mathbf{X}$ is the design matrix of genotypes x_{ij} , and β 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 $\lambda \sum_{j=1}^p |\beta_j|$ ($\lambda > 0$) 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 $MVN(\mathbf{0}, \sigma_g^2 \mathbf{A})$, and the Equation (7) becomes equivalent to

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

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 (LDI) 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 $\geq LDI$. 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 $\geq LD2$. 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:

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

and

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

as single and multi-locus models, respectively. The notation W_{il} ($l=1, \dots, m_k$) represents the PCs in the k th window ($k=1, \dots, M$; M is the total number of the windows), θ_0 is the intercept, and θ_{lk} is the

302 regression parameter of the given PC, u_i is the random effect defined in the same way as in (3) and
303 the kinship matrix is calculated as in (4) using the original SNP data.

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

313
314 *Arabidopsis thaliana* data set

315 The *A. thaliana* GWAS data set originates from Baxter et al. (2010), who used it to identify genetic
316 variants associated with leaf sodium accumulation. A total of 337 individuals were genotyped using
317 an Affymetrix SNP array to generate around 250 000 SNPs as described in Atwell et al. (2010).
318 After removing SNPs with a minor allele frequency < 0.05 done by using our own R script, 200 121
319 SNPs distributed over five chromosomes of 18-30 Mb in length remained to be used here. We used
320 two sets of threshold values for LD-clustering: *low*, with $LD1=0.1$ and $LD2=0.3$ and *high*, with
321 $LD1=0.3$ and $LD2=0.5$. The threshold value for the subsequent PC regression step was kept at 80%
322 for both sets of analyses. To reduce the computational burden of LDn-clustering, based on putative
323 recombination hot spots (see above) window break points were chosen such that window size was
324 approximately 1000 SNPs, and pairwise r^2 values were only calculated within a window size of 100
325 SNPs (as LDna requires a pairwise a matrix of all r^2 -values for each window the remaining values
326 were set to 0).

327

328 *Pungitus pungitus* data set

329 The *P. pungitus* F₂ inter-population cross data of 283 individuals was originally generated by
330 crossing a female from the Baltic Sea (Helsinki; 60°13'N, 25°11'E) and a male from a northeastern
331 Finnish pond (Rytilampi; 66°23'N, 29°19'E). Detailed information about the origin, maintenance,
332 genotyping and phenotyping of the crosses can be found from earlier publications (e.g. Laine et al.
333 2013; Yang et al. 2016; Li et al. 2017).

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

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

348 Finally, the markers were ordered within each linkage group with OrderMarkers2 module
349 with default parameters. OrderMarkers2 was run twice on each chromosome using
350 informativeMask=13 and informativeMask=23, removing either markers only informative in the
351 mother or father, respectively. This created two maps for each chromosome, one having more
352 maternal markers and the other having more paternal markers, both having on average 2/3 markers
353 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).

380 However, for these simulations any random set of 300 consecutive polymorphic SNPs would have
 381 sufficed. Two of the first combinations were the same as those used for the original data set and in
 382 the two latter, *LD2* was further reduced to 0.2 but with two different *PC*-thresholds: 0.8 and 0.9.
 383 This was done to investigate how over-merging LD clusters (when LD thresholds are low)
 384 potentially can be compensated by extracting more PCs during the PC regression step. For each data
 385 set a phenotype was generated on the basis of the multiple-locus model in Equation (6). The effect
 386 size of five QTL were independently simulated from a normal distribution $N(0,1)$. The random
 387 effect $\mathbf{u}$ is simulated from a multivariate normal distribution $MVN(\mathbf{0}, \sigma_g^2 \mathbf{A})$, with $\sigma_g^2 = 10$, and the
 388 residual error is simulated from a normal distribution $N(0,1)$ with narrow sense heritability, h^2 ,
 389 between 0.2 and 0.3. The five QTL were either randomly chosen among the 300 SNPs (*random*) or
 390 within a window of 50 bps (*clustered*). As the main aim of these simulations was to compare the
 391 power to detect significant QTL with the SNP- and LD-cluster based approaches in a small data set,
 392 the data set size and generation of phenotypic values were not aimed to necessarily be biologically
 393 realistic. Analyses were performed on 1000 sets of simulated phenotypic values for the four
 394 threshold settings as well as for a data set where each SNP was analysed independently (*'no*
 395 *clustering'*). EMMAX was used for the GWAS analyses as described above. Statistical power was
 396 estimated as the proportion of significant QTL among all 5×1000 causal SNPs in the simulated data
 397 sets, after Bonferroni correction for multiple testing (performed separately for each simulated data
 398 set). Confidence intervals for the proportion of significant QTL was estimated as the 95% quantiles
 399 from 1000 bootstrap replicates. False negative rates for LDn-clustered data sets were well below
 400 0.05% for all threshold settings and were thus not considered further here. False negative rates for
 401 the *no clustering* data sets were not considered either as we would have needed to take into account
 402 that non-causal loci can be significant also due to LD, and thus defining false negatives would have
 403 been somewhat arbitrary.

404 For the *P. pungitus* genome we focused on a single chromosome (chromosome I) using
 405 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 heritabilities 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 F₂-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

483 data (38% of QTL detected). The multi-locus method also illustrated better ability to control the
484 number of false positives than the single-locus approach (Table 2).

485
486 *Analysis of leaf sodium accumulation in A. thaliana*

487 The standard SNP-based single-locus association mapping with Bonferroni correction identified 23
488 significant SNPs, with 22 located in Chr4 (ranging from 6381929 bp to 7581539 bp in the *A.*
489 *thaliana* genome), and a single SNP located in Chr3 (18095036 bp; Fig. 3a). The permutation test
490 identified 28 SNPs located in the same genomic regions as the Bonferroni test (Fig. 3c). The multi-
491 locus approach identified only three significant SNPs in Chr4 (located at 6392280, 6418442 and
492 6742032 bp, respectively; Fig. 3e), which are a subset of the SNPs detected by the single locus
493 mapping.

494 The cluster-based single-locus mapping (data generated with the parameter $LD1=0.3$ and
495 $LD2=0.5$) with Bonferroni and permutation tests detected four, six and 21 significant genomic
496 regions in Chr4, respectively (Fig. 3b, d). The window-based multi-locus approach identified one
497 region (6415034-6418442 bp) and two singleton QTL at 6392280 and 6455695 in the same
498 chromosome (Fig. 3f). For all the methods, the signal with the highest statistical significance was
499 detected at the SNP located at 6392280 bp of Chr 4.

500
501 *Analysis of P. pungitus data*

502 In the QTL analysis of the *P. pungitus* data, the SNP-based single-locus approach with Bonferroni
503 correction did not identify any significant loci (Fig. 4a; Fig. S8a, Supporting Information). This was
504 also the case in the multi-locus analysis (Fig. 4e; Fig. S8e, Supporting Information). In contrast, the
505 permutation test based on the single-locus mapping identified multiple significant loci in three
506 chromosomes (Chr 9, 20 and 21) when the allele substitution and dominance effects were tested in a
507 group (Fig. 4c). In separate testing of the allele substitution effects, a number of loci in Chr 9, 20,
508 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 F₂-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 accidentally 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

561 The performance of conventional SNP-based single- and multi-locus approaches is influenced by
562 the LD pattern of the data. In the case of the *A. thaliana* GWAS data set with a fast LD decay over
563 the genome, the single-locus mapping with either Bonferroni or permutation tests identified a
564 similar set of more than 20 SNPs in the same genomic regions in Chr 3 and Chr 4. In contrast, the
565 LASSO based multi-locus approach only identified three SNPs in Chr4. One of them (Chr4:
566 6392280) is located within the region of the gene AtHKT1_1: (Chr4:6391984–6395877), which has
567 been shown to be functionally associated with sodium leaf accumulation in *A. thaliana* (Baxter et al.
568 2010). This difference between single and multi-locus mapping results can be explained by the fact
569 that the multi-locus method relies on conditional hypothesis testing. When the strength of the
570 association for a single SNP is tested, all other correlated SNPs' associations have already been
571 accounted for. Therefore, the multi-locus test is stricter than the single-locus test.

572 In the bi-parental *P. pungitus* data with high levels of LD extending considerable distances
573 over the linkage map, the Bonferroni correction became too conservative to identify any significant
574 SNPs. This was expected: Bonferroni becomes overly conservative when the multiple tests are
575 positively correlated with each other (Goeman and Solari 2014). In contrast, the permutation test,
576 which can effectively account for the correlation structure in the data, was still able to identify a
577 number of significant loci with the significant allele assignable to one of the grandparents. That the
578 detected QTL had allele substitution effects from the grandfather (originating from the pond
579 population), but not from the grandmother, indicates that the grandparental genotypes were AB and
580 AA, respectively, and the allele 'B' originating from the pond environment caused the phenotypic
581 variation observed in the F₂ generation. The four-way cross model applied here was able to detect
582 more significant QTL for the focal trait than the MapQTL approach applied to the same data by
583 Yang et al. (2016). In addition, the four-way cross model helps elucidate from which population the
584 allele effects on the phenotypes originate from.

585 The multi-locus mapping with the original SNP data also failed to identify any significant
586 QTL in the *P. pungitus* SNP data. One possible explanation is that the widely used coordinate

587 descent algorithm used to solve the LASSO penalized regression may work poorly and converge
588 extremely slowly for highly correlated data sets (Kim et al. 2016). Another possible reason is that
589 the stability selection as a multiple testing approach involves a data sub-sampling step, which may
590 result in reduced statistical power when the sample size is small. Regarding the hypothesis tests, a
591 de-biased LASSO approach (Javanmard and Montanari 2014; Li et al. 2017) can be performed on
592 the whole data set without any re-sampling of the data, and therefore might have better power to
593 detect QTL. Unfortunately, we discovered that the de-biased LASSO could not be applied to this
594 high dimensional data set with over 200 000 regression parameters due to its high computational
595 cost. Nevertheless, as discussed below, the de-biased LASSO can easily be applied to the LDn-
596 clustered data set.

597

598 *Cluster-based gene mapping*

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

608 In the simulated data (focusing on 300 polymorphic SNPs spanning 122 kb around the most
609 significant QTL for sodium leaf accumulation) we saw no differences in the proportion of
610 significant QTL between SNP-based gene mapping and cluster-based gene mapping. This was
611 expected; due to the fast LD decay across *A. thaliana* chromosomes, the number of independent test
612 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., Nederbragt 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).

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

647 Finally, from the computational point of view, the cluster-based approach appears to have
648 a distinct advantage over mapping with individual SNPs. For instance, in the case of the *A. thaliana*
649 data, the original SNP data of over 200 000 SNPs (or alleles) can be summarized with only 90 000
650 PCs in a high LD data set. This leads to a substantial reduction of the computational complexity.
651 For example, conducting a permutation test on the *A. thaliana* data set takes about seven days by
652 using 5 000 replications on a 64-bit Windows 7 desktop computer with a 3.4-GHz Intel (i7) CPU
653 and 32.0 GB of RAM (note that computational time estimates for all the methods were implemented
654 on a single core). Using the same set up, the cluster-based permutation test takes only 6-7 hours.
655 The stability selection took about three hours on the *A. thaliana* data set and only 30 minutes on the
656 *P. pungitus* data set. The de-biased LASSO approach consumed about 30 days for the clustered data
657 set, and might take several months for the full SNP data. The cluster-based approach can also be
658 used for other computationally intensive GWAS models such as the Bayesian LASSO (Li et al.
659 2011; Pasanen et al. 2015) and Elastic net (Huang et al. 2015) to improve their computational
660 efficiency. The LDn-clustering algorithm took <20 min for the *A. thaliana* data set and <10 min for
661 the *P. pungitus* data set, and can be parallelised over many computer clusters (each
662 window/chromosome can be processed independently) for use in whole genome data sets where this
663 kind of dimensionality reduction is likely to be most beneficial.

664

665 *Concluding remarks and future directions*

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

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

690

691 **Acknowledgements**

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

698 **References**

699 Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., ..., Nordborg, M.
700 (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines.
701 *Nature*, 465(7298), 627-631. doi: 10.1038/nature08800

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

704 Baxter, I., Brazelton J. N., Yu D., Huang, Y. S., Lahner, B., Yakubova, E., ..., Salt, D. E. (2010) A
705 Coastal Cline in Sodium Accumulation in *Arabidopsis thaliana* Is Driven by Natural Variation
706 of the Sodium Transporter AtHKT1;1. *PLoS Genetics*, 6(11), e1001193. doi:
707 10.1371/journal.pgen.1001193

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

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

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

715 Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A.,...,McVean, G.,
 716 1000 Genomes Project Analysis Group (2011) The Variant Call Format and VCFtools.
 717 Bioinformatics, 27(15), 2156-2158.
 718
 719 Dudbridge, F., & Koeleman, B. P. C. (2004) Efficient computation of significance levels for
 720 multiple associations in large studies of correlated data, including genomewide association
 721 studies. American Journal of Human Genetics, 75 (3), 424–435, 2004. doi: 10.1086/423738
 722 Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and*
 723 *Prediction*. Cambridge University Press: Cambridge.
 724 Ernst, C. W., & Steibel, J. P. (2013) Molecular advances in QTL discovery and application in pig
 725 breeding. Trends in Genetics, 29(4), 215-224. doi: 10.1016/j.tig.2013.02.002
 726 Fan, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., & Xiong, M. (2013) Functional
 727 linear models for association analysis of quantitative traits. Genetic Epidemiology 37(7), 726-
 728 742. doi: 10.1002/gepi.21757
 729 Friedman, J., Hastie, T., Tibshirani, R. (2010) Regularization paths for generalized linear models
 730 via coordinate descent. Journal of Statistical Software, 33(1), 1.
 731 Ge, T., Smoller, J. W., Sabuncu, M. R. (2016) Kernel machine regression in neuroimaging genetics.
 732 Machine Learning and Medical Imaging, 31-68. [https://doi.org/10.1016/B978-0-12-804076-](https://doi.org/10.1016/B978-0-12-804076-8.00002-5)
 733 [8.00002-5](https://doi.org/10.1016/B978-0-12-804076-8.00002-5)
 734 Goeman, J. J., & Solari, A. (2014) Multiple hypothesis testing in genomics. Statistics in Medicine
 735 33(11), 1946-1978. doi: 10.1002/sim.6082
 736 Hastie, T., Tibshirani, R., & Friedman, J. (2009) *Elements of Statistical Learning (Second Edition)*.
 737 Springer: New York.
 738 Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L.,..., & Thompson, P.
 739 M. (2011) Voxelwise gene-wide association study (vGeneWAS) multivariate gene-based

740 association testing in 731 elderly subjects. *Neuroimaging*, 56(4), 1875-1891.
741 doi:10.1016/j.neuroimage.2011.03.077.

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

744 Husby, A., Kawakami, T., Rönnegård, L., Smeds, L., Ellegren, H., & Qvarnström, A. (2015)
745 Genome-wide association mapping in a wild avian population identifies a link between genetic
746 and phenotypic variation in a life-history trait. *Proceedings of the Royal Society B*. doi:
747 10.1098/rspb.2015.0156

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

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

753 Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., Eskin, E. (2008)
754 Efficient control of population structure in model organism association mapping. *Genetics*
755 178(3), 1709–1723. doi: 10.1534/genetics.107.080101

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

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

762 Kim, B., Yu, D., Won, J-H. (2016) Comparative study of computational algorithms for the Lasso
763 with high-dimensional, highly correlated data. *Applied Intelligence*, in press. doi:
764 10.1007/s10489-016-0850-7.

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

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

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

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

775 Laine, V. N., Shikano, T., Herczeg, G., Vilkki, J., & Merilä, J. (2013) Quantitative trait loci for
776 growth and body size in the nine-spined stickleback *Pungitius pungitius* L. *Molecular Ecology*,
777 22 (23), 5861-5876. doi: 10.1111/mec.12526

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

780 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., March, G., Abecasis, G.,
781 Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009) The Sequence
782 Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. doi:
783 10.1093/bioinformatics/btp352

784 Li, Z., & Sillanpää, M. J. (2012) Overview of LASSO-related penalized regression methods for
785 quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, 125(3),
786 419–435. doi: 10.1007/s00122-012-1892-9

787 Li, Z., Guo, B., Yang, J., Herczeg, G., Gonda, A., Balázs, G., Shikano, T., Calboli, F. C. F., &
788 Merilä, J. (2017) Deciphering the genomic architecture of the stickleback brain with a novel
789 multi-locus gene-mapping approach. *Molecular Ecology*, 26(6), 1557-1575. doi:
790 10.1111/mec.14005

791 Liang, Y., & Kelemen, A. (2008) Statistical advances and challenges for analyzing correlated high
792 dimensional SNP data in genomic study for complex diseases. *Statistics Surveys*, 2: 43-60. doi:
793 10.1214/07-SS026

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

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

798 Morgenthaler, S., Thilly, W. G., 2007. A strategy to discover genes that carry multi-allelic or mono-
799 allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation*
800 *Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615: 28-56. doi:
801 10.1016/j.mrfmmm.2006.09.003

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

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

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

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

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

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

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

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

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

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

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

826 Xu, S. (1996) Mapping quantitative trait loci using four-way crosses. *Genetics Research*, 68(2),
827 175-181. doi: 10.1017/S0016672300034066

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

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

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

834 Yi, H., Breheny, P., Imam, N., Liu, Y., & Hoeschele, I. (2015) Penalized multimarker vs. single-
835 marker regression methods for genome-wide association studies of quantitative traits. *Genetics*,
836 199: 205-222. doi: 10.1534/genetics.114.167817

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

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

842 Zhang, K., Calabrese, P., Nordborg, M., Sun, F.Z. (2002) Haplotype block structure and its

843 applications to association studies: Power and study designs. *American journal of human*
844 *genetics*, **71**, 1386–1394.

845 Zheng X, Levine D, Shen J *et al.* (2012) A high-performance computing toolset for relatedness
846 and principal component analysis of SNP data. *Bioinformatics (Oxford, England)*, **28**, 3326–
847 3328.

848

849

850 **Data accessibility**

- 851 - The *A. thaliana* SNP dataset is available at: <https://github.com/Gregor-Mendel-Institute/atpolydb>.
- 852 The phenotype data is available from the original publication (Baxter et al. 2010).
- 853 - The *P. pungitus* phenotype data is available from Yang et al. (2016).
- 854 -The *P. pungitus* SNP data set as well as the R source codes for implementing all the statistical
855 methods introduced in the paper will be available in Dryad upon acceptance.
- 856 -LDn-clustering is available as an additional function in an updated version of the existing LDna R-
857 package (<https://github.com/petrikemppainen/LDna/tree/v.63>).

858

859 **Conflicts of interest**

860 Authors declare no conflict of interests

861

862 **Figure legends**

863 Figure 1. LDn-clustering. Shown is an example of how LDn-clustering can account for the mosaic-
864 like pattern of LD in population genomic data by grouping loci (within windows) based on LD
865 regardless of their physical position in the genome. Each LD-cluster has a unique colour
866 combination [colours between (a) and (b) do not necessarily match] and line height along the y-axis
867 (upper panel). In each LD-cluster the minimum LD between all loci in the cluster is above (a) 0.1 or
868 (b) 0.3 and the median LD among all pairwise LD values in each LD-cluster is above (a) 0.3 or (b)
869 0.5. Loci not connected to any other SNPs by these thresholds are considered as independent
870 (singleton-clusters). There are 15 and 25 unique LD-clusters in (a) and (b), respectively. Positions
871 of the vertical lines (along the x-axis) match the positions of loci in the lower LD heatmap figure.
872 The figure is based on 63 consecutive SNPs from *A. thaliana* data set Chr 4 (starting from SNP-
873 position 6237655).

874
875 Figure 2. Results from simulated study with subsets of data. Panel (a) shows mean number of
876 significant QTL (five in each simulated data set), for four different threshold settings for LDn-
877 clustering (values in the legend separated by ‘;’ represent threshold values *LD1*, *LD2* and *PC*,
878 respectively) when QTL are randomly sampled among all SNPs (*Random*), or from 50 consecutive
879 SNPs (*Clustered*) along the chromosome ($h^2 = 0.2 - 0.3$). Panel (b) shows the time taken to conduct
880 GWAS on clustered (yellow) and non-clustered data (grey) for the *A. thaliana* simulated data. Panel
881 (c) shows the proportion of significant QTL from *P. pungitus* linkage group 1 (one QTL in each
882 data set) for different heritabilities when GWAS was performed on all 4344 SNPs (Clustering=No)
883 or when GWAS was performed on 12 clusters produced by LDn-clustering (Clustered=Yes). Panel
884 (d) show the time taken to conduct GWAS on clustered (yellow) and non-clustered data (grey) for
885 the *P. pungitus* simulated data. Data are based on 1000 simulated sets of phenotypic values, and
886 error bars in (a) and (c) represent 95% bootstrap confidence intervals (1000 bootstrap replicates).

887

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

897
898 Figure 4. Quantitative trait locus mapping of the *P. pungitus* data. Results of SNP- and cluster-
899 based QTL analysis are shown on the left (a, c, e) and right (b, d, f) panels, respectively. The allele
900 substitution effects of two founders and the dominance effects are tested jointly in the same model.
901 In (a) and (b), dots (blue or green coloured) represent the p -values from the association test
902 calculated by single -locus mapping, and the red curve the significance threshold after Bonferroni
903 correction. In (c) and (d), dots represent the adjusted p -values (0.05) calculated by the permutation
904 test in single-locus mapping, and red lines the significance threshold (0.05). In (e) and (f), dots
905 present the selection probability calculated by the multi-locus stability selection method, and the red
906 line the corresponding significance threshold (guaranteeing the expected number of false positives
907 to be < 1).

909 Figure 1.

910

911
912 Figure 2.

913

914

915

916

917

918

919

Figure 3

924 **Tables**

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

927	Setting	<i>LD1</i>	<i>LD2</i>	<i>PC</i>	PCs	Clusters
928	1	0.1	0.3	0.8	143	111
929	2	0.3	0.5	0.8	172	168
930	3	0.1	0.2	0.8	130	85
931	4	0.1	0.2	0.9	140	85

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

934 **Table 2.** The average performance of single- and multi-locus QTL-mapping methods with SNP or
 935 cluster based analyses in a simulation study of genome-wide *A. thaliana* data. Number of false
 936 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

937

MOLECULAR ECOLOGY RESOURCES

Supporting Information for:

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

Zitong Li, Petri Kempainen, Pasi Rastas, Juha Merilä

Table of Contents:

Figure S1	Page 2
Figure S2	Page 4
Figure S3	Page 5
Figure S4	Page 6
Figure S5	Page 7
Figure S6	Page 8
Figure S7	Page 9
Figure S8	Page 10
Appendix S1	Page 11
References	Page 13

MOLECULAR ECOLOGY RESOURCES

MOLECULAR ECOLOGY RESOURCES

Figure S1. Linkage disequilibrium network clustering. (a) A toy example of the two complete linkage clustering steps (*step 1* and *step 2*) in LDn-clustering is given. In LD networks (top), vertices represent loci and edges represent r^2 -values above a given threshold value (in this case threshold is ≥ 0.05 , but all loci are connected to each other with some [low] level of LD in the data). When viewing the LD-network with only edges with r^2 -values ≥ 0.05 , the initial cluster is ‘long and elongated’ within which we wish to identify more ‘compact’ and ‘spherical’ sub-clusters. Spherical clusters represent a set of highly interconnected nodes, which in the framework of LD-networks signify sets of loci all connected by high LD. In more elongated clusters, the last locus in the cluster has typically very low LD with the first, and only adjacent loci are connected by high LD. Thus, a complete linkage clustering algorithm in which a cluster is defined by its weakest edge is employed. If we remove (*step 1*) the two weakest edges (indicated in blue) that are below the LD threshold $LD1$ (set to 0.1), three ‘spherical’ clusters remain (black, green and red); all loci within these clusters are interconnected by at least one edge $>LD1$. Next, loci in clusters where median LD between all loci in the cluster is below the LD threshold $LD2$ (set to 0.3) enter to *step 2*. The complete linkage clustering procedure is repeated, but now removing all edges $<LD2$ (indicated in red). In *step 2*, one cluster (green) and three singleton-clusters (blue, red and yellow) are identified (all loci within these clusters are interconnected by at least one edge $>LD2$). Clusters from *step 1* and *step 2* are combined and PC regression is used to summarize the genetic information among loci in the clusters. However, note that if the median LD for the green cluster in *step 1* would have been $>LD2$ (i.e. would have contained more strongly connected loci) this cluster would not have entered into the second step and the additional blue cluster would not have been identified. In other words, this procedure finds larger clusters than just setting $LD1=LD2=0.3$, but still ensures that weakly connected SNPs from high recombination regions are treated independently in subsequent GWAS. For instance, in the *A. thaliana* example data set (300 SNPs) clustering with $LD1=0.1$ and $LD2=0.3$ identified 111 clusters whereas $LD1=LD2=0.3$ identified 143 clusters. This two-step procedure can also be demonstrated with clustering trees (data is the same as for the *A. thaliana* example data set with 300 SNPs). First, all clusters (b) whose ancestral branches cross $LD1$ (red dashed line) are identified (*step 1*); in these clusters minimum LD between any pair of loci in the cluster is $\geq LD1$. A second clustering step (*step 2*) is performed with loci from clusters from *step 1* with median LD $< LD2$, but now clusters with minimum LD between any pair of loci in the cluster $\geq LD2$ are identified. Loci from clusters that do not meet this criterion are treated as individual SNPs in subsequent analyses (singleton-clusters). Clusters in (a) and (b) that meet the LD criteria are indicated in red, else they are grey (i.e. tree in [b] is based on loci represented by grey branches in tree [a]).

MOLECULAR ECOLOGY RESOURCES

Figure S2. Linkage disequilibrium networks. Shown are clustering solutions for (a) *low* and (b) *high* (b) threshold settings in the *A. thaliana* data (Atwell et al. 2010; Baxter et al. 2010). Data is based on the first 1000 SNPs on Chr 1. Edges represent r^2 -values between loci (nodes) above given threshold values which are 0.5 for (a) and (b), and 0.9 for (c); loci not connected to any other loci by LD above the thresholds are not shown. Each colour combination represents a unique LD-cluster (195 and 363 clusters for *low* and *high*, respectively, but not all are visible in these networks) and black indicates singleton-clusters. In (c) linkage group 15 from the *P. pungitus* F₂-intrecross data (Yang et al. 2016; Li et al. 2017; Rastas et al. 2017) is shown with nine LD-clusters. While some ‘sub-networks’ are ‘elongated’ and ‘chain-like’, within each of them the LDn-binning finds sub-clusters that are more ‘spherical’ and ‘compact’ (where all loci are strongly interconnected by high levels of LD).

MOLECULAR ECOLOGY RESOURCES

Figure S3. Linkage disequilibrium decay in (a) *A. thaliana* and (b) *P. pungitius* data for chromosome 1. Black dots represent pairwise r^2 -estimates (function ‘snpGdsLDMat’; R-package: ‘SNPRelate’; Zheng et al. 2012) between the 500 randomly selected SNPs from the first chromosomes of each of the species, and the red solid curves depict a LD decay curve estimated using a non-parametric regression (“penalized B-splines”; Eilers and Marx 1996). Briefly, the

penalized B-spline is defined as $\min_a [LD(d) - \sum_{k=1}^M \varphi_k(d) \alpha_k]^2 + \lambda \sum_{i=3}^M [\alpha_{i-2} - 2\alpha_{i-1} + \alpha_i]^2$, where

$LD(d)$ represents the pairwise r^2 measure among the SNPs, and d is the absolute map distance between the two SNPs, φ_k represent B-spline bases, which are a series of truncated power
bases defined in the domain of d , α_k represents the model parameters, and

$$\lambda \sum_{i=3}^M [\alpha_{i-2} - 2\alpha_{i-1} + \alpha_i]^2$$

is a roughness penalty used here to induce smoothness in the curve fitting to reduce the noise.
The optimal value of the tuning parameter λ was chosen by cross validation.

Field Code Changed

Formatted: Font: (Default) Times New Roman

Formatted: Font: Italic

Field Code Changed

Formatted: Font: Italic

Field Code Changed

Field Code Changed

Field Code Changed

Formatted: Font: (Default) Calibri, (Asian) Chinese (PRC)

Figure S4. Distribution of number of SNPs in linkage disequilibrium network clusters from *A. thaliana* full data set. *Low* and *high* refer to LD threshold settings in LDn-clustering, see main text for details. Distributions are Gamma shaped with *high* threshold settings producing more clusters with fewer loci than *low* settings.

MOLECULAR ECOLOGY RESOURCES

Figure S5. Proportion variance explained by PC versus median LD in LD-clusters. Data is from two LDn-clustering threshold settings (*low* and *high*) from the full *A. thaliana* data set. Color represents number of SNPs in an LD-cluster on log2 scale. Horizontal line depicts threshold value for *PC* (the minimum cumulative amount of genetic variation PCs explain in a cluster), and vertical lines threshold values for *LD2* for *high* (0.5, yellow) and *low* (0.3, blue) LD threshold settings for LDn-clustering.

MOLECULAR ECOLOGY RESOURCES

Figure S6. LD network of *P. pungitus* chromosome 12. Each color represents a unique LD-cluster. One outlier LD-cluster (blue) containing only six SNPs is not connected to the rest of the chromosome at LD threshold: $r^2=0.9$ in this network. The exact shape of the network is to some extent arbitrary.

MOLECULAR ECOLOGY RESOURCES

Figure S7. Quantitative trait locus mapping of the *P. pungitus* data: the results of de-biased LASSO on the cluster-based data set for (a) the allele effects of two founders and the dominance effects tested separately (the top, middle and bottom panels, respectively) and (b) tested jointly as a group. Dots (in blue or green color) represent p-values from the association test calculated by single-locus mapping, and red lines depict the Bonferroni adjusted significance thresholds (0.05).

MOLECULAR ECOLOGY RESOURCES

Figure S8. 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 effects of two founders and the dominance effects were tested separately, and were illustrated in the top, middle and bottom panels of each sub-figure. In (a) and (b), dots (blue or green colour) represent the p -values from the association tests calculated by single-locus mapping, and the red line the Bonferroni adjusted significance threshold (0.05). In (c) and (d), dots represent the adjusted p -values 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 from the multi-locus stability selection method, and the red line is the corresponding significance threshold (guaranteeing the expected number of false positives to be < 1).

MOLECULAR ECOLOGY RESOURCES

Appendix S1 Group hypothesis testing procedure for the cluster-based association mapping

Single-locus approach

In the single locus mixed effect model (as defined in Equation 9 in the main text)

$$y_i = \theta_0 + \sum_{l=1}^{m_k} W_{il} \theta_{lk} + u_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad (\text{S1})$$

where W_{il} ($i=1, \dots, n$; n is the total number of individuals; $l=1, \dots, m_k$; where m_k is the number of principal components calculated on the basis of the original SNPs in the cluster) are the principal components (PCs) in the k th cluster ($k=1, \dots, q$; q is the total number of the clusters defined in the genome), θ_0 is the intercept, and θ_l is the regression parameter of the given PC, u_i is the random effect having $\mathbf{u} = [u_1, \dots, u_n] \sim \text{MVN}(0, \sigma_g^2 \mathbf{A})$, the kinship matrix $\mathbf{A}$ is calculated using Equation (4) in the main text, and ε_i is the model residual. We want to evaluate the strength of association between a quantitative trait and a genomic region $\mathbf{W}_k = (W_{i1}, \dots, W_{im_k})$ consisting of m_k number of principal components from the cluster. This can be done by comparing the model (S1) and a null model with only an intercept term:

$$y_i = \theta_0 + u_i + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad (\text{S2})$$

which leads to an F -statistic:

$$F = \frac{RSS_{\text{intercept}} - RSS_{\text{full}}}{df_{\text{intercept}} - df_{\text{full}}} \bigg/ \frac{RSS_{\text{full}}}{df_{\text{full}}}. \quad (\text{S3})$$

In (S3), RSS represents the residual sum of squares of the intercept model (S1):

$$RSS_{\text{intercept}} = \sum_{i=1}^n (y_i - \hat{\theta}_0 - \hat{u}_i)^2,$$

and that of the full model (S2):

$$RSS_{\text{full}} = \sum_{i=1}^n (y_i - \hat{\theta}_0 - \sum_{l=1}^{m_k} W_{il} \hat{\theta}_l - \hat{u}_i)^2,$$

where $\hat{\theta}_0$, $\hat{\theta}_l$ and $\hat{u}_i$ are the estimates of the regression parameters obtained by EMMAX (Kang et al. 2010). The degrees of freedom df are defined as the number of fixed regression parameters in (S1) or (S2), so we have $df_{\text{intercept}}=1$, and $df_{\text{full}}=m_k+1$. The P -value is calculated as $1-P_F(F, df_{\text{intercept}}, df_{\text{full}})$, where P_F is the cumulative probability function of the F distribution (Kutner et al. 2004). The multiple adjustment is conducted using the Bonferroni or Permutation test for mixed models as described in the main text (Methods and Materials section).

Multi-locus approach

The cluster-based LASSO regression is defined as:

$$\min_{\theta} \frac{1}{2n} (\mathbf{y} - \mathbf{W}\theta)^T \Sigma^{-1} (\mathbf{y} - \mathbf{W}\theta) + \lambda \sum_{k=1}^M \sum_{l=1}^{m_k} |\theta_{lk}|, \quad (\text{S4})$$

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Subscript

Formatted: Font: Italic

MOLECULAR ECOLOGY RESOURCES

where $\mathbf{y}$ is a vector of the phenotype data y_i , $\mathbf{W}$ is the design matrix of the PCs in all the cluster, $\boldsymbol{\theta} = (\theta_{11}, \dots, \theta_{m_1 1}, \dots, \theta_{1q}, \dots, \theta_{m_q q})$ is the vector of the regression coefficients of the PCs, and $\boldsymbol{\Sigma} = \sigma_g^2 \mathbf{A} + \sigma_e^2 \mathbf{I}$. Stability selection is used to test if a genomic region is associated with the phenotype. Half of the individuals are first randomly selected and analyzed by the LASSO regression defined by (S4). The LASSO shrinks most of the unimportant coefficients to zero. A cluster $\mathbf{W}_k = (W_{1l}, \dots, W_{im_k})$ ($k=1, \dots, q$) is claimed to be selected into the model if at least one of its component has a non-zero effect. The tuning parameter was chosen so that a total number of q ($q < M^{1/2}$) clusters were selected into the model. This procedure was repeated thousands of times, and a selection probability P_l , summarizing the frequency of the cluster being over the replicates, was obtained. The following criterion was used for QTL identification: if $P_l > \frac{1}{2} + \frac{q^2}{2 \times M}$, the cluster $\mathbf{W}_l$ is judged to be significantly associated with the quantitative trait. This guarantees that the expected number of falsely selected loci is less than one (Meinshausen and Bühlmann 2010).

MOLECULAR ECOLOGY RESOURCES

References

- 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
- 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
- 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
- Eilers, P. H. C., & Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–121. http://www.stat.washington.edu/courses/stat527/s13/readings/EilersMarx_StatSci_1996.pdf
- 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
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004) *Applied Linear Regression Models*. New York: McGraw-Hill.
- 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
- 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
- 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
- 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.