Genome analysis

**De novo** construction of polyploid linkage maps using discrete graphical models

Pariya Behrouzi 1,* and Ernst C. Wit 2

1Department of Plant Sciences, Wageningen University and Research, Wageningen 6708PB, The Netherlands and 2Institute of Computational Science, Università della Svizzera italiana, Lugano 6900, Switzerland

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on March 29, 2018; revised on August 16, 2018; editorial decision on August 30, 2018; accepted on September 3, 2018

Abstract

**Motivation:** Linkage maps are used to identify the location of genes responsible for traits and diseases. New sequencing techniques have created opportunities to substantially increase the density of genetic markers. Such revolutionary advances in technology have given rise to new challenges, such as creating high-density linkage maps. Current multiple testing approaches based on pairwise recombination fractions are underpowered in the high-dimensional setting and do not extend easily to polyploid species. To remedy these issues, we propose to construct linkage maps using graphical models either via a sparse Gaussian copula or a non-paranormal skeptic approach.

**Results:** We determine linkage groups, typically chromosomes, and the order of markers in each linkage group by inferring the conditional independence relationships among large numbers of markers in the genome. Through simulations, we illustrate the utility of our map construction method and compare its performance with other available methods, both when the data are clean and contain no missing observations and when data contain genotyping errors. Our comprehensive map construction method makes full use of the dosage SNP data to reconstruct linkage map for any bi-parental diploid and polyploid species. We apply the proposed method to three genotype datasets: barley, peanut and potato from diploid and polyploid populations.

**Availability and implementation:** The method is implemented in the R package netgwas which is freely available at https://cran.r-project.org/web/packages/netgwas.

**Contact:** pariya.behrouzi@wur.nl

**Supplementary information:** Supplementary data are available at Bioinformatics online.

1 Introduction

A linkage map provides a fundamental resource to understand the order of markers for the vast majority of species whose genomes are yet to be sequenced. Furthermore, it is an essential ingredient in the often used quantitative trait loci (QTL) mapping (Pang et al., 2017; Wu et al., 2015) of genetic diseases and particularly in identifying genes responsible for heritable or other types of diseases in humans or traits such as disease resistance in plants or animals.

Recent advances in sequencing technology make it possible to comprehensively sequence huge numbers of markers, construct dense maps and ultimately create a foundation for studying genome structure and genome evolution, identifying QTLs and understanding the inheritance of multi-factorial traits. Next-generation sequencing (NGS) techniques offer massive and cost-effective sequencing throughput. However, they also bring new challenges for constructing high-quality linkage maps. NGS data can suffer from high rates of genotyping errors, as the observed genotype for an individual is not necessarily identical to its true genotype. Under such circumstances, constructing high-quality linkage maps can be difficult (Buetow, 1991; Ronin et al., 2015).

Each species is categorized as diploid or polyploid by comparing its chromosome number. Diploids have two copies of each chromosome. For diploid species many algorithms for constructing linkage maps have been proposed. Some of them have been implemented...
into user-friendly software, such as JOINMAP (Jansen et al., 2001), R/qtl (Broman et al., 2003), OneMap (Margarido et al., 2007) and MSTmap (Wu et al., 2008). Among the algorithms for constructing genetic maps, R/qtl estimates genetic maps and identifies genotyping errors in relatively small sets of markers. JOINMAP is a commercial software widely used in the scientific genetics community. It uses two methods to construct genetic maps: one is based on regression (Stam, 1993) and the other uses a Monte Carlo multi-point maximum likelihood (Jansen et al., 2001). OneMap has been reported to construct linkage maps in non-inbred populations. However, it is computationally expensive. MSTMap is a fast genetic map algorithm that determines the order of markers by computing the minimum spanning tree of an associated graph.

Polyploid organisms have more than two chromosome sets. Polyploidy is very common in flowering plants and in different crops such as watermelon, potato and bread wheat, which contain three (triploid), four (tetraploid) and six (hexaploid) sets of chromosomes, respectively. Despite the importance of polyploid species, statistical tools for construction of their linkage map are underdeveloped (Grandke et al., 2017). However, Grandke et al. (2017) recently developed a method for this purpose. Their method is based on calculating recombination frequencies between marker pairs, then using hierarchical clustering and an optimal leaf algorithm to detect chromosomes and order markers. Nevertheless, this method can be computationally expensive even for a small numbers of markers. PolymapR (Bourke et al., 2017) is another software that constructs a genetic map from bi-parental populations of outcrossing auto-polyploids. It clusters markers over a range of LOD thresholds; it requires users to select the LOD threshold that best clusters the data. Then it uses weighted linear regression or multi-dimensional scaling for ordering markers. Most literature has focused on constructing genetic linkage maps for tetraploids, but these are limited only to auto-tetraploid species. TetraploidsSNPMap, is a software for this situation (Hackett et al., 2017), but because it needs manual interaction and visual inspection its application is limited. For example, user needs to specify how many linkage groups (chromosomes) the algorithm should detect. Furthermore, current approaches to polyploid map construction are based mainly on estimation of recombination frequency and LOD scores (Wang et al., 2016), which does not use the full multi-variante information in the data.

Different diploid and polyploid map construction methods have made substantial steps toward building better-quality linkage maps. However, the existing methods still suffer from low quality genetic mapping performance, in particular when ratios of genotyping errors and missing observations are high. The main contribution of this paper is to introduce, for both diploid and polyploid species, a novel linkage map algorithm to overcome the difficulties arising routinely in NGS data. With the proposed method we aim to build high-density and high-quality linkage maps using the statistical property called conditional dependence relationships, which reveals direct relations among genetic markers. For diploid scenarios, we evaluated the performance of the proposed method and the other methods in several comprehensive simulation studies, both when the input data were clean and had no missing observations and when the input data were noisy. We measured the performance of the methods in accuracy scores of grouping and ordering. In addition, we studied the performance of our method and an alternative method in constructing linkage maps for tetraploid peanut. Furthermore, we applied the map construction method in netgwas (Behrouzi and Wit, 2017b) to construct maps for two genotype datasets: barley and potato from diploid and tetraploid populations, respectively.

2 Genetic background on linkage map

A linkage map is the linear order of genetic markers on a chromosome. Geneticists use it to study the association between genes and traits. In this section, we describe the relationship between a linkage map and single nucleotide polymorphism (SNP) markers. For the moment, we assume that each allele can take only one of two values, $A$ or $a$. This assumption can be relaxed without requiring any methodological adjustments; more will follow in the discussion. Here, we are dealing with markers from high-throughput data such as NGS and SNP arrays.

2.1 Linkage map for diploids and polyploids

Diploid organisms contain two sets of chromosomes, one from each parent, whereas polyploids contain more than two sets of chromosomes. In polyploids the number of chromosome sets reflects their level of ploidy: triploids have three sets, tetraploids have four, pentaploids have five and so forth. Here, we refer to diploids and polyploids as q-ploid $q \geq 2$, where in diploids $q = 2$, triploids $q = 3$, tetraploids $q = 4$ and so on.

The genotype of any q-ploid organism can be homozygous or heterozygous at each single locus on the genome. Different genotype forms of the same gene are called alleles. Alleles can lead to different traits. Alleles are commonly represented by letters; for example, for the gene related to the trait, the allele could be called $A$ and $a$. In q-ploid individuals there are $q$ copies of allele. If all $q$ allele copies of an organism are identical, the organism is in the homozygous state at that locus; otherwise it is in the heterozygous state. For instance, a tetraploid individual is homozygous for two size alleles, $A$ and $a$, if all 4 allele copies are either $A$, or $a$, which correspond with the genotypes $AAAA$ and $aaaa$, respectively. If a tetraploid individual is heterozygous the following three genotypes could appear: one copy of the $A$ allele and three copies of $a$ (e.g. $Aaaa$), two copies of $A$ and two copies of $a$ (e.g. $AAaa$), or three copies of $A$ and one copy of $a$ (e.g. $AAAd$). Unlike existing methods, our method works not only for diploid organisms but also for all polyploids. Obviously, our method can also be used to analyze simple haploid organisms such as haploid yeast cells.

2.2 Mapping population

Mating between two parental lines with recent common biological ancestors is called inbreeding. Mating between parental lines with no common ancestors up to e.g. 4–6 generations is called outcrossing. In both cases, the genomes of the derived progenies are random mosaics of the genomes of the parents. As a consequence of inbreeding parental alleles are attributable to each parental line in the genome of the progeny, whereas in outcrossing this is not the case.

Inbreeding progenies derive from two homozygous parents. Some inbreeding designs, such as double haploid (DH), lead to a homozygous population where the derived genotype data include only homozygous genotypes of the parents, namely $AA$ and $aa$ (conveniently coded as 0 and 1). However, some other inbreeding designs such as F2 lead to a heterozygous population, where the derived genotype data contain both heterozygous and homozygous genotypes, namely $AA$, $Aa$ and $aa$ (conveniently coded as 0, 1 and 2). Although many other experimental designs are being used in genetic studies, not all existing methods for linkage mapping support all inbreeding experimental designs. However, our proposed algorithm constructs a linkage map for any type of biparental inbreeding experimental designs. In fact, unlike other existing methods, our approach does not require specifying the population type because it
is broad and handles any population type that contains at least two distinct genotype states.

Outcrossing or outbred experimental designs, such as full-sib families, derive from two non-homozygous parents. Thus the genome of the progenies includes a mixed set of many different marker types, including fully informative markers and partially informative markers (e.g. missing markers). Markers are fully informative when all of the resulting gamete types can be phenotypically distinguished on the basis of their genotypes; they are called partially informative when the gamete types have identical phenotypes.

### 2.3 Meiosis and Markov dependence

During meiosis, chromosomes pair and exchange genetic material (crossover). In diploids, pairing at meiosis occurs between two chromosomes. In polyploids the q chromosome copies may form different types of multi-valent pairing. For example, in tetraploids all four chromosome copies may pair at meiosis. Assume a sequence of ordered SNP markers X_1,X_2,...,X_q along chromosome c in a q-ploid species. We describe the Markov dependence structure between markers for different population schemes (Fig. 1):

- Scheme (i): During meiosis in inbred populations, one complication arises when in the genotype data we cannot identify each homologue due to heterozygous genotypes. Q-ploid (q ≥ 2) heterozygous inbred populations, like F2, are examples of such cases, where we define X_{jk} as

\[
X_{jk} = \begin{cases} 
1 & \text{if marker } j \text{ on homologue } k \text{ is of type } A, \\
0 & \text{otherwise}
\end{cases}
\]

where A is one of the two possible alleles at that specific location. Here, X_{jk} represents the allele at homologue k of a chromosome, where the genotype in that location can be written as X_j = \{X_1, ..., X_{jq}\}. For example, at marker location j, X_j = Aaaa is one possible genotype for a tetraploid species (q = 4); it includes one copy of the desirable allele A where X_{1j} = 1, X_{2j} = 0, X_{3j} = 0 and X_{4j} = 0 represent the alleles in the first, second, third and fourth homologues, respectively. The other possible genotypes which include one copy of the desired allele A are aAaa, aAaa, aAaa, and AAAA. Because it is typically impossible to distinguish between genotypes with the same number of copies of a desired allele (e.g. Aaaa, aAaa, aAaa, aAaa), we therefore take a random variable Y_j as the observed number of A alleles at location j:

\[
Y_j = \sum_{k=1}^{q} X_{jk}. 
\]

Table 1 shows an example of correspondence between Y_j and X_j for a q-ploid species when q = 4. We note that a q-ploid species

| Y_j | X_j |
|-----|-----|
| 0   | aaaa|
| 1   | Aaaa, aAaa, aAaa, aAaa |
| 2   | Aaaa, Aaaa, aAaa, Aaaa, aAaa, aAaa |
| 3   | Aaaa, aAaa, Aaaa, aAaa, aAaa |
| 4   | Aaaa |

Note: Relation between different genotypes, X_j, and allele dosage, Y_j, for a tetraploid individual, where A is the reference allele.

Fig. 1. Cartoon example of conditional dependence pattern between neighboring markers in different population schemes: (i) homozygous, (ii) inbred, (iii) outcrossing (outbred) populations, where ordered markers Y_1, ..., Y_q reside on chromosome 1 and Y_2, ..., Y_q on chromosome 2.
contains \(q + 1\) genotype states at location \(j\), as shown in Table 1 for a tetraploid species.

Due to **genetic linkage**, the sequence of ordered SNP markers \(Y_1, Y_2, \ldots, Y_g\) forms a Markov chain as Equation (1) with state space \(S\) which contains \(q + 1\) states. Therefore, the conditional (in)dependence relationship (2) between neighboring markers is held. Figure 1(iii) presents a cartoon image of the conditional independence graph for this scheme.

Scheme (iii): In outcrossing (outbred) populations, unlike inbred populations, the meaning of ‘parental’ is either unknown or not well defined. In other words, markers in the genome of the progenies can not easily be assigned to their parental homologues. For example, if both non-homzygous parents contain \(A_i A_j A_k\) genotype at marker location \(j\), then offspring will also have \(A_i A_j A_k\) genotype at marker location \(j\). But we do not know whether that genotype belongs to the paternal or maternal homologue, since both parents have \(A_i A_j A_k\) genotype at marker location \(j\). So, in this case we define \(X_{jk}\) as follows

\[
X_{jk} = \begin{cases} 
1 & \text{if marker } j \text{ on homologue } k \text{ is of type } A_i \\
0 & \text{otherwise}
\end{cases}
\]

where \(A_i\) is one of the possible parental alleles at location \(j\). So, random variable \(Y_j\), which represents the dosage of alleles, can be defined as Equation (3).

Furthermore, in polyploids the **linkage** depends on how a single chromosome pairs during meiosis to produce gametes. In this regard, if both polyploid parents have an \(A_i\) allele in all \(q\) haploids, then offspring will also have it, and this will not co-vary with neighboring markers. The possibility of different pairing models during meiosis makes the situation more complex. In diploids, the four homologue chromosomes pair up and form a bivalent, then crossover before recombination occurs. But polyploid meiosis can occur in various ways; in tetraploids four homologue chromosomes can during meiosis form either two separate bivalents, each of which contributes one haploid, like diploids; or, alternatively, in a more complex situation, the four homologue chromosomes can form quadrivalents, so that cross-over occurs between eight haploids. In both pairing models, bivalent or quadrivalent, crossover events result in recombined haploids that are mosaics of parental chromosomes. Outbred progenies are genetically diverse and highly heterozygous, whereas inbred individuals have little or no genetic variation.

The term (1) partially holds for the scheme (iii), where a discrete graphical model can be defined for a multi-nominal variable \(Y_j = \{0, 1, \ldots, q\}\). We use conditional independence to construct linkage maps in outbred populations. However, in this type of population, due to a mixed set of different marker types, the conditional independence relationship between neighboring markers may be more complicated. Many genetic assumptions made in traditional linkage analyses (e.g., known parental linkage phases throughout the genome) do not hold here. For example, when both parents have \(A_i\) allele, then their offspring will also have it; however this will not covary with neighboring markers. Figure 1(iii) shows a cartoon example of such conditional independence graphs.

To summarize, term (1) holds for schemes (i) and (ii) and partially (iii) because transition probability from a genotype at location \(j\) to a genotype at location \(j + 1\) depends on the recombination frequency between the two locations \(j\) and \(j + 1\), which is independent of recombination in the other locations. This can be modeled by a discrete Markov process \(\{Y_j\}_{j=1}^{g}\) with state space \(S\) which contains \(q + 1\) genotype states and a transition matrix, which, in case of polyploids \((q \geq 3)\), can be calculated with respect to the mode of chromosomal pairing (e.g., bivalent or quadrivalent). The Markov structure of the SNP markers in all three schemes yields a graphical model with as many nodes as markers in a genome. The random variable \(Y_j\) follows a discrete graphical model whereby the joint distribution \(P(Y_j)\) can be factorized as,

\[
P(Y) = \prod_{i=1}^{p} \prod_{j=1}^{q} \pi_{i,j}^{(c)}(Y_i^{(c)}, Y_j^{(c)}).
\]

where \(C\) defines the number of chromosomes in a genome, and \(p_c\) stands for the number of markers in chromosome \(c\). The outer multiplication of (4) shows the **independent assortment law**, and the inner multiplication represents the **genetic linkage** between markers within a chromosome, where the factor \(\pi_{i,j}^{(c)}\) indicates the conditional dependence between adjacent markers, given the rest of the markers. Through this probabilistic insight, the inferred conditional (in)dependence relationship between markers provides a high-dimensional space for the construction of a linkage map.

### 3 Algorithm to detect linkage map

We propose to build a linkage map in two steps; first, we reconstruct an undirected graph for all SNP markers on a genome, and second, we determine the correct order of markers in the obtained linkage groups from the first step. We also show how our method handles genotyping errors and missing observations in reconstructing a linkage map.

#### 3.1 Estimating marker-marker network

To reconstruct an undirected graph between SNP markers in a q-ploid species we propose two methods: the sparse ordinal glasso approach (Behrouzi and Wit, 2017a) and the non-paranormal skeptic approach (Liu et al., 2012) (the latter discussed under Supplementary Material). The former method can deal with missing values, whereas the latter is computationally faster.

An undirected graphical model for the joint distribution (4) of a random vector \(Y = (Y_1, \ldots, Y_g)\) is associated with a graph \(G = (V, E)\), where each vertex \(j\) corresponds to a variable \(Y_j\). The pair \((j, l)\) is an element of the edge set \(E\) if and only if \(Y_j\) is dependent of \(Y_l\), given the rest of the variables. In the graph estimation problem, we have \(n\) samples of the random vector \(Y\), and it is our aim to estimate the edge set \(E\). Depending on how various mapping populations are produced, \(Y\) represents either binary variables \(Y = \{0, 1\}\), as in homozygous populations, or multi-nominal variables \(Y = \{0, 1, \ldots, q\}\) where \(q\) is the ploidy level. For example in diploids \(q\) is 2 and in tetraploids 4.

#### 3.1.1 Sparse ordinal glasso

A relatively straightforward approach to discover the conditional (in)dependence relationship among markers is to assume underlying continuous variables \(Z_1, \ldots, Z_p\) for markers \(Y_1, \ldots, Y_p\), which can not be observed directly. In our modeling framework, \(Y_i\) and \(Z_i\) define observed rank and true latent value, respectively, where each latent variable corresponds to one observed variable. The relationship between \(Y_i\) and \(Z_i\) is expressed by a set of cut-points \((-\infty, C_i^{(d)}], (C_i^{(d)}, \infty)\) for \(d = 0, 1, \ldots, q\), which is obtained by partitioning the range of \(Z_i\) into \(q + 1\) disjoint intervals. Thus, \(Y_i^{(d)}\), which represents the genotype of the \(i\)-th sample for the \(j\)-th marker, can be written as follows
where we define $\mathcal{D} = \{ x_j^{(0)} \in \mathbb{R} | C_{q,1}^{(0)} < x_j^{(0)} \leq C_j^{(0)} \}$. We use a high dimensional Gaussian copula with discrete marginals. We assume

$$
Z \sim N_p(0, \Sigma)
$$

where the $p \times p$ precision matrix $\Theta = \Sigma^{-1}$ contains all the conditional independence relationships between the latent variables. Given our parameter of interest $\Theta$, we non-parametrically estimate the cutpoints for each $i=1, \ldots, n$ as follows

$$
C_{q}^{(i)} = \begin{cases} 
-\infty & \text{if } q = 0; \\
\Phi^{-1}\left(\frac{1}{n} \sum_{j=1}^{n} I(y_j^{(i)} \leq q)\right) & \text{if } q = 1, \ldots, q_i - 1; \\
+\infty & \text{if } q = q_i.
\end{cases}
$$

### 3.1.2 Penalized EM algorithm

In genotype datasets we commonly encounter situations where the number of genetic markers $p$ exceeds the number of samples $n$. To solve this dimensionality problem we propose to impose an $l_1$ norm penalty on the likelihood consisting of the absolute value of the elements of the precision matrix $\Theta$. Furthermore, to be able to deal with commonly occurring missing values in genotype data we implement an EM algorithm (McLachlan and Krishnan, 2007), which iteratively finds the penalized maximum likelihood estimate $\hat{\Theta}_q$. This algorithm proceeds by iteratively computing the conditional expectation of complete log-likelihood and optimizing it. In the E-step we compute the conditional expectation in the penalized log-likelihood

$$
Q_q(\Theta; (\hat{\Theta}^{(m)}) = 
\sum_{j=1}^{n} \log(\hat{\Theta}) - tr\left(\frac{1}{n} \sum_{j=1}^{n} E_{Z^{(j)}}(Z^{(j)}(\hat{\Theta})^{(m)}, D)\hat{\Theta}\right) - p\log(2n)
$$

where $\lambda$ is a non-negative tuning parameter. To calculate the conditional expectation $R = \frac{1}{n} \sum_{j=1}^{n} E_{Z^{(j)}}(Z^{(j)}(\hat{\Theta}^{(m)}), D)$ we propose two different approaches, namely Gibbs sampling and an approximation method (Behrouzi and Wit, 2017a; Guo et al., 2015). Further details on the calculation of the conditional expectation are provided in the Supplementary Material. The M-step is a maximization problem which can be solved efficiently using either graphical lasso (Friedman et al., 2008)

$$
\hat{\Theta}^{(m+1)}_{\text{glasso}} = \arg\max_{\Theta} \left\{ \log(\Theta) - tr(R\Theta) - \lambda\|\Theta\|_1 \right\}
$$

or the CLIME estimator (Cai et al., 2011)

$$
\hat{\Theta}^{(m+1)}_{\text{CLIME}} = \arg\min_{\Theta} \|\Theta\|_1 \quad \text{subject to} \quad \|R \Theta - I_p\|_{\infty} \leq \lambda,
$$

where $I_p$ is a $p$-dimensional identity matrix.

In large-scale genotyping studies, it is common to have missing genotype data. Before determining the number of linkage groups and ordering markers, we handle the missing data within the E-step of the EM algorithm, where we calculate the conditional expectation of true latent variables given the observed ranks. If an observed value, $y_j^{(0)}$, is missing, we take the unconditional expectation of the corresponding latent variable. In the EM framework we can easily handle high ratios of missingness in the data.

### 3.2 Determining linkage groups

A group of loci that are correlated defines a linkage group (LG). Depending on the density and proximity of the underlying markers each LG corresponds to a chromosome or part of a chromosome.

The number of discovered linkage groups is controlled by the tuning parameter $\lambda$ (Section 3.1). We use the extended Bayesian criterion (eBIC), which has successfully been applied by Yin and Li (2011) in selecting sparse Gaussian graphical models for genomic data to determine the number of linkage groups. The eBIC is defined as

$$
eBIC(\lambda) = -2\ell(\hat{\Theta}_q) + (\log n + 4; \log p)df(\lambda),
$$

where $\ell(\hat{\Theta}_q)$ is the non-penalized likelihood and $\gamma \in [0, 1]$ is an additional parameter. And $df(\lambda) = \sum_{i \leq i < j \leq p} I(\hat{\theta}_{q,i} \neq 0)$ where $\hat{\theta}_{q,i}$ is the $(i,j)$th entry of the estimated precision matrix $\hat{\Theta}_q$, and $I$ is the indicator function. In case of $\gamma = 0$ the classical BIC is obtained. Typical values for $\gamma$ are 0 and 1. We select the value of $\gamma$ that minimizes (7) for $\gamma = \frac{1}{2}$. We note that in practice there is an opportunity that linkage groups have been selected manually given a prior knowledge.

It is notable that in existing map construction methods the construction of linkage groups is usually done by manually specifying a threshold for pairwise recombination frequencies; this, however, influences the output map, whereas our method detects linkage groups automatically in a data-driven way.

Some genotype studies suffer from low numbers of samples or they contain signatures of epistatic selection (Behrouzi and Wit, 2017a), which may cause bias in determining the linkage groups. To address this problem, besides the model selection step, we use the fast-greedy algorithm to detect the linkage groups in the inferred graph. This community detection algorithm reflects the two biological concepts of genetic linkage and independent assortment in a sense that it defines communities which are highly connected within, and have few links between communities.

### 3.3 Ordering markers

Assume that a set of $d$ markers has been assigned to the same linkage group. Let $G(V^{(d)}, E^{(d)})$ be a sub-graph on the set of unordered $d$ markers, where $V^{(d)} = \{1, \ldots, d\}$, $d \leq p$ and the edge set $E^{(d)}$ represents the estimated edges among $d$ markers where $E^{(d)} \subseteq E$. We remark that the precision matrix $\hat{\Theta}_q^{(d)}$, a sub-matrix of $\hat{\Theta}_q$, contains all conditional dependence relations between the set of $d$ markers. Depending on the type of mating between the parental lines we introduce two methods to order markers, one based on dimensionality reduction and another based on bandwidth reduction. Both methods result in a one-dimensional map.

#### 3.3.1 Inbred

In inbred populations, markers in the genome of the progenies can be assigned to their parental homologues, resulting in a simpler conditional independence pattern between neighboring markers. In the case of inbreeding, we use multi-dimensional scaling (MDS) to represent the original high-dimensional space in a one-dimensional map while attempting to maintain pairwise distances.

We define the distance matrix $D$ which is a $d \times d$ symmetric matrix where $D_{ii} = 0$ and $D_{ij} = -\log(p_{ij})$ for $i \neq j$. Here, the matrix $p$ represents the conditional correlation among $d$ objects which can be obtained as $p_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_i \theta_j}}$, where $\theta_{ii}$ is the $i$th element of the precision matrix $\Theta$.

We aim to construct a configuration of $d$ data points in a one-dimensional Euclidean space by using information about the distances...
between the \(d\) nodes. Given the distance matrix \(D\), we define a linear ordering \(L\) of \(d\) elements such that the distance \(D\) between them is similar to \(D\). We consider a metric MDS, which minimizes
\[
L = \arg\min_{L} \sum_{i=1}^{d} \sum_{j=1}^{d} (D_{ij} - \bar{D})^2 \tag{1}
\]
across all linear orderings.

### 3.3.2 Outbred

An outbred population derived from mating two non-homozygous parents results in markers in the genome of progenies that can not easily be assigned to their parental homologues. Neighboring markers that vary only on different haploids will appear as independent, therefore requiring a different ordering algorithm (Fig. 1c).

In that case, to order markers we use the reverse Cuthill-McKee (RCM) algorithm (Cuthill and McKee, 1969). This algorithm is based on graph models. It reduces the bandwidth of the associated adjacency matrix, \(A_{d \times d}\), for the sparse matrix \(\Theta_{d}^{[\Delta]}\). The bandwidth of the matrix \(A\) is defined by \(b = \max_{i \neq j}|i - j|\). The RCM algorithm produces a permutation matrix \(P\) such that \(PA^TP^T\) has a smaller bandwidth than does \(A\). The bandwidth is decreased by moving the non-zero elements of the matrix \(A\) closer to the main diagonal. The way to move the non-zero elements is determined by relabeling the nodes in graph \(G(V, E)\) in consecutive order. Moreover, all of the non-zero elements are clustered near the main diagonal.

### 3.4 Uncertainty in map construction

Both empirical estimation of marginals and selection of the tuning parameter produce uncertainty in the map construction procedure. We compute the uncertainty associated with the estimated linkage map through a non-parametric bootstrap.

We replicate \(B\) datasets that are created by sampling with replacement \(n\) samples from the dataset \(Y_{n \times p}\). We run the entire map construction procedure to each bootstrap dataset. Each estimated map is associated with an adjacency matrix. The average of the B bootstrap adjacency matrices for the bootstrap samples reflects the underlying uncertainty in the estimation procedure of the linkage map construction.

We have applied this procedure to evaluate the uncertainty associated with the estimation of the linkage map for the example dataset \(Cvi \times Col\) in \textit{A. thaliana}. This well-studied experiment is derived from a RIL cross between Columbia-0 (Col-0) and the Cape Verde Island (Cvi-0), where 367 individual plants were genotyped across 90 genetic markers (Simón et al., 2008). The \(Cvi - 0 \times Col - 0\) RIL is a diploid population with three possible genotypes, where the genotypes are coded as \(0, 1, 2\), where 0 and 2 represent two homozygous genotypes (AA resp. BB) from Col-0 and Cvi-0, and 1 defines the heterozygous genotype (AB). We generate 100 independent bootstrap samples from the Col-0 and Cvi-0 cross. For each 100 bootstrap samples, we apply the map construction algorithm. Figure 2 presents the certainty associated with the estimated linkage map for a sub-sample of \(n = 50\) plants. The line type shows the estimated certainty associated with each link. For example, the gray dotted between marker ‘c2.003593’ from chromosome 2 and marker ‘c1.2989998’ from chromosome 1 has the certainty value of 0.01, likewise for markers ‘c3.221457’ and ‘c4.00012’ from chromosomes 3 and 4. To sum up, for the links in the original dataset we obtain a 56% certainty that the links are really there, whereas for the non-links in the original dataset we are 98% certain that they are not there. We remark that when we use all \(n = 367\) individuals, all the 100 bootstrap samples estimate an identical linkage map, which is the reason why for illustrative purpose, we used a sub-sample \(n = 50\).

### 4 Simulation study

In this section, we study the performance of the proposed method for different diploids and polyploids. In Section 4.1 we perform a comprehensive simulation study to compare the performance of the proposed algorithm with other available tools in diploid map constructions, namely JoinMap (Jansen et al., 2001) and MSTMap (Wu et al., 2008). The former is based on Monte Carlo maximum likelihood and the latter uses a minimum spanning tree of a graph.

In Section 4.2, we compare our method with an alternative method to examine their performance on polyploids. At this moment the proposed method is the only one that constructs linkage maps for polyploid species in data-driven way without any manual adjustment.

#### 4.1 Diploid species

We simulate genotype data from an inbred \(F2\) population. This population type generates discrete random variables with values \(Y = \{0, 1, 2\}\) associated with the three distinct genotype states, AA, Aa and aa at each marker. The procedure in generating genotype data is as follows: first, two homozygous parental lines are simulated with genotypes AA and aa at each locus. A given number of markers, \(p\), are spaced along the pre-defined chromosomes. Then, two parental lines are crossed to give an \(F1\) population with all heterozygous genotypes Aa at each marker location. Finally, a desired number of individuals, \(n\), are simulated from the gametes produced by the \(F1\) population.

A genotyping error means that the observed genotype for an individual is not identical to its true genotype, for example, observing genotype AA when Aa is the true genotype. Genotyping errors can distort the final genetic map, especially by incorrectly ordering markers and inflating map length. Therefore, to order markers that contain genotyping errors is an essential task in constructing high-quality linkage maps. To investigate this, we create genotyping errors in the simulated datasets by randomly flipping the
heterozygous loci along the chromosomes to either one of the homozygous allele.

For each simulated dataset, we compare the performance of the map construction in netgwas with two other models: JOINMAP, and MSTMap. We compute two criteria: grouping accuracy (GA) and ordering accuracy (OA), to assess the performance of the above mentioned tools in estimating the correct map. The former measures the closeness of the estimated number of linkage groups to the correct number, and the latter calculates the ratio of markers that are correctly ordered. We define the grouping accuracy as follows:

\[ GA = \frac{1}{1 + \left(\frac{LG - LG^*}{LG}\right)^2} \]

where \( LG \) stands for actual number of linkage groups and \( LG^* \) is the estimated number of linkage groups. The GA criterion is a positive value with a maximum of 1. A high value of GA indicates good performance in determining the correct number of linkage groups. To compute ordering accuracy, we calculate the Jaccard distance, \( d_J \), which measures mismatches between the estimated order and the true order. We define the ordering accuracy of the estimated map as \( OA = \frac{1}{1 + d_J} \). This measurement lies between 0 and 1, where 1 and 0 stand for a perfect and a poor ordering, respectively.

In terms of computational burden, it is worth noting that netgwas runs in parallel. In the performed simulations, we ran the map construction functions, both in netgwas and MSTMap on a Linux machine with 24 2.5 GHz Intel Xeon processors and 128 GB memory. JOINMAP runs only on Windows. We ran it on a Windows machine with 3.20 GHz Intel Xeon processors and 8 GB RAM memory.

When the data are clean and complete, netgwas and MSTMap are both efficient in terms of running time. For 1000 markers over 5 chromosomes and 200 F2 individuals, they run in 1.4 and 0.20 min, respectively, on Intel i7 laptop with 16 Gb RAM. However, as shown in Figure 4 in Wu et al. (2008), when the input data is noisy, the running time for MSTMap increases, whereas the noise ratio does not affect running time in netgwas.

**Evaluation of estimated maps in presence of genotyping errors**

We studied the accuracy of the estimated linkage maps using two methods: netgwas and MSTMap where genotyping errors are randomly distributed across the genetic markers. The simulated data contained 300 markers for both \( n = 100 \) and \( n = 200 \) individuals where the genotyping error rates ranged from 0 up to 0.45. In these sets of simulations, we activated the error-detection feature in MSTMap.

Figure 3 evaluates the accuracy of estimated maps in terms of grouping (Fig. 3, top) and ordering accuracies (Fig. 3, bottom). In general, this figure shows that netgwas constructed significantly better maps than MSTMap across the full range of genotyping error rates. More specifically, for a moderate number of individuals, \((n = 200)\), Figure 3 (top) shows that netgwas correctly estimated the actual number of linkage groups for the full range of genotyping error rates. When \( n = 100 \) netgwas perfectly estimated the actual number of linkage groups up to 10% genotyping errors, and very
accurately ($\geq 0.95$) estimated the number of linkage groups for error rates between 10% and 30%. With more than 30% genotyping errors the accuracy diminished. MSTMAP always made significantly poorer estimates of the actual number of linkage groups than did netgwas; its performance immediately began to drop as soon as there was some level of genotyping errors. Surprisingly, it estimated the number of linkage groups better when $n = 100$ than $n = 200$. As Wu et al. (2008) mentioned in their paper, choosing $\epsilon$ remains a critical issue in MSTMAP to detect correct number of linkage groups. In their clustering approach, $\epsilon$ does not only depends on $n$, but it also depends on the Hamming distance between linkage groups. Therefore, it might be that the estimation of the linkage group or clustering is better for small $n$ rather than for big $n$.

Figure 3 (bottom) shows the ordering accuracy within each correctly estimated linkage group. Ordering quality in netgwas was significantly better than MSTMAP for both $n = 100$ and $n = 200$. This is because conditional independence is an effective way to recover relationships among genetic markers. More specifically, when $n = 200$ and the error rate equaled zero, netgwas ordered markers perfectly (100% accuracy) and MSTMAP orders markers with a high accuracy (95%). In addition, with increased genotyping error rates, the map construction in netgwas outperformed that of the MSTMAP in ordering markers within each LG. Based on our simulations, we remark that with both netgwas and MSTMAP erroneous markers remain in the estimated linkage map. However, netgwas orders them in the correct LG (Fig. 3), whereas MSTMAP performs poorly in detecting LGs as well as in correctly ordering markers. We note that, if one is interested in identifying erroneous markers, netgwas uses the Lincoln and Lander (1992) approach to detect markers that have genotyping errors. In their approach, an error LOD score is calculated for each individual at each marker; large scores ($> 4$) indicate likely genotyping errors. The netgwas uses Rqtl package (Broman et al., 2003) to detect genotyping errors.

4.2 Polyploid species

The peanut (Arachis hypogaea L.) is an important oilseed crop and food legume grown in tropical or sub-tropical regions (Bertioli et al., 2013). Linkage map for tetraploid peanut have already been generated in Bertioli et al. (2013) using MapMaker (Lander et al., 1987).

In total, 771 markers were mapped into 20 linkage groups. This population consisted of 89 F6 individuals derived from a cross between A. hypogaea cv. Runner IAC 886 and a colchicine-induced tetraploid. We applied netgwas to construct linkage map for the same population based on the same sample.

Figure 4 compares the estimated maps from netgwas and MapMaker. The netgwas builds the tetraploid peanut map by using marker-marker partial correlations. It detects all the peanut chromosomes correctly, except for chromosome TA9, which it detected as two smaller linkage groups containing 10 and 42 markers. Bertioli et al. (2013) used MapMaker to map 771 markers into 20 linkage groups with minimum LOD score of 3.0 and a maximum recombination fraction of 0.35. In Bertioli et al. (2013), they manually tune the number of LGs to the known information about the peanut chromosome, whereas netgwas detects number of LGs in a data-driven way. In terms of ordering the markers were similar, except in chromosomes TA9 and TA10. In addition, netgwas runs in 22.44 s on an Intel i7 laptop with 16 GB RAM. The run-time of MapMaker is unknown because the authors of the peanut dataset did not provide computational times.

5 Construction of linkage map for diploid Barley

In the literature (Wu et al., 2008), a barley genotyping dataset is used to compare different map construction methods for real-world diploid data. This genotyping dataset is generated from a doubled haploid population, which results in homozygous individual plants, $Y_i \in \{0, 1\}$. Barley genotype data are the result of crossing Oregon Wolfe Barley Dominant with Oregon Wolfe Barley Recessive (see http://wheat.pw.usda.gov/ggpages/maps/OWB). The Oregon Wolfe Barley (OWB) data include $p = 1328$ markers that were genotyped on $n = 175$ individuals of which 0.02% genotypes are missing. The barley dataset is expected to yield 7 linkage groups, one for each of the 7 barley chromosomes.

As shown in Figure 5, through estimating $\Theta$, which contains conditional (in)dependence relationships between barley markers, we were able to correctly detect the seven barley chromosomes as sub-graphs in the estimated undirected graph. Furthermore, using

---

**Table 1: Estimated number of linkage groups (LGs) for OWB dataset**

| Linkage group (LG) | Estimated # LG | Size of the LGs |
|--------------------|----------------|-----------------|
| netgwas            | 7              | 140, 199, 211,187, 236, 182, 173 |
| MSTMap             | 1              | 1328            |

**Comparison of ordering accuracy between netgwas and MSTMap.**

In this Table is assumed MSTMap correctly estimated the number of LGs in the OWB dataset.

| Linkage group (LG) | Sensitivity score |
|--------------------|------------------|
|                    | netgwas | MSTMap  |
| 1                  | 0.86     | 0.92    |
| 2                  | 0.78     | 0.52    |
| 3                  | 0.78     | 0.92    |
| 4                  | 0.74     | 0.49    |
| 5                  | 0.71     | 0.38    |
| 6                  | 0.61     | 0.50    |
| 7                  | 0.70     | 0.61    |
| average            | 0.74     | 0.63    |

**Fig. 5. Summary of comparison between netgwas and MSTMAP in barley data.**

Table on top summarizes estimated number of LGs (chromosomes) and size of markers within each LG. Below, ordering accuracy scores for the two methods. Below figure shows the estimated undirected graph in netgwas for the barley data. This graph consists of seven sub-graphs, each showing a chromosome.
the conditional correlation matrix as distance in the multidimensional scaling approach helped us to order markers with high accuracy. In addition, Figure 5 reports the result of applying the two methods: netgwas and MSTMAP, to construct a linkage map for the barley data. The top part of Figure 5 shows that our method correctly estimated the true number of chromosomes. Also, the size of markers within each chromosome is consistent with the number of markers that reported in Gistué et al. (2011). MSTMAP was not able to estimate the true number of chromosomes and grouped all 1328 markers as one linkage group. The bottom of Figure 5 shows the accuracy of estimated marker order in seven barley chromosomes. To be able to compare marker order in both methods we used the actual map to cluster markers in the map resulting from MSTMAP. Thus, at the bottom of Figure 5 it is assumed that MSTMAP has estimated correct number of chromosomes. Average ordering of accuracy scores across the linkage groups in netgwas is higher than those in MSTMAP except with chromosomes 1 and 3.

To investigate the influence of segregation distortion on map construction in netgwas, we constructed a linkage map for a Double Haploid (DH) wheat population of size 599 markers that genotyped on 218 individuals (the dataset is available at ASMap package). This genotype data contains a set of markers with segregation distortion. The netgwas ordered distorted markers in the correct linkage group (see Supplementary Table S1). Based on this result, we remark that segregation distortion seems to have little influence on netgwas map construction.

6 Construction of linkage map for tetraploid potato

World-wide, the potato is the third most important food crop (Bradshaw and Bonierbale, 2010). However, the complex genetic structure of tetraploid potatoe’s (Solanum tuberosum L.) makes it difficult to improve important traits such as disease resistance in this crop. Thus there is a great interest in constructing linkage maps in the potato to identify markers related to disease resistance genes.

The full-sib mapping population MSL603 consists of 156 F1 plants resulting from a cross between female parent Jacqueline Lee and male parent MSG227-2. The obtained genotype data contain 1972 SNP markers (Massa et al., 2015) with five allele dosages which are associated with the random variables \( Y_j \in \{ 0, 1, \ldots, 4 \} \) for \( j = 1, \ldots, 1971 \).

Figure 6 represents the result of applying the proposed map construction method to the unordered potato genotype data. Figure 6a shows the estimated sparse precision matrix for the unordered genotype data. Figure 6b represents the estimated precision matrix after ordering markers; it reveals the number of potato chromosomes as blocks across the diagonal. The potato genome contains 12 chromosomes. The proposed method correctly identifies all 12 chromosomes. The estimated linkage map contains 1957 markers. Figure 6c compares the estimated order in netgwas versus the estimated order in Massa et al. (2015) using the TetraploidSNPMap software (Hackett et al., 2017). Each dashed line shows the estimated linkage group (LG), where netgwas estimates LGs using the eBIC criteria in (7) and in TetraploidSNPMap the number of LGs should be specified manually. Given that the ordering of markers has always been a challenging task in linkage map constructions, and in particular for polyploid species, both methods ordered markers with similar precision except in chromosome 9 where TetraploidSNPMap suggests a different ordering.

Using simulated tetraploid data with 3000 markers over five chromosomes and 207 F1 individuals netgwas produced maps within 5.5 min, whereas TetraploidSNPMap took 15 min on an Intel i7 laptop with 16 GB RAM.

7 Conclusion

Construction of linkage maps is a fundamental and necessary step for detailed genetic study of diseases and traits. A high-quality linkage map provides opportunities for greater throughput gene manipulation and phenotype improvement.

Here, we have introduced a novel method for constructing linkage maps from high-throughput genotype data where the number of genetic markers exceeds the number of individuals. The proposed method makes full use of SNP dosage data to construct a linkage map for any bi-parental diploid or polyploid population. We propose to build linkage maps in two steps: (i) inferring conditional independence relationships between markers on the genome; (ii) ordering markers in each linkage group, typically a chromosome. In the first step of the proposed method we used the Markov properties of adjacent markers: the genotype of an individual haploid at marker \( Y_j \) given its genotype at \( Y_{j-1} \) or \( Y_{j+1} \) is conditionally independent from the genotype at any other marker location. This property defines a graphical model for discrete random variables.
We employed a Gaussian copula graphical model combined with a penalized EM algorithm to estimate a sparse precision matrix $\Theta$. This method iteratively computes the conditional expectation of the complete penalized log-likelihood, and optimizes it to estimate $\Theta$. The method can also deal with missing values, which are very common in genotype datasets. The non-paranormal skeptic is an alternative approach that is computationally faster but can not deal with missing genotypes. The number of linkage groups is determined via the information criteria, eBIC. Detection of linkage groups in the existing map construction software is usually done by manual tuning; this, however, influences the output map, whereas our method detects linkage groups automatically in a data-driven way.

Depending on the type of mapping population, inbred or outbred, we use either a multi-dimensional scaling approach or the Cuthill-McKee algorithm, respectively, in step 2 of the proposed linkage map construction. Both ordering algorithms result in a one-dimensional map. We noted that in outcrossing populations it is difficult to order markers because a clear definition of the parental genotype is lacking.

We performed several simulation studies to compare the performance of the proposed method with other commonly used diploid map construction tools. To address the challenges in the construction of a linkage map from genotype data, we studied the performance of the proposed method on simulated data with high ratios of genotyping errors. As shown in our simulation studies, our method, called netgwas, outperformed the commonly available linkage map tools, when the input data were noisy.

As outlined in Cervantes-Flores et al. (2008), constructing linkage maps in polyploids, with outcrossing behavior, is a challenging task. So far, based on our experience, no method has been developed to construct polyploid linkage maps for a large number of different marker types without any manual adjustment and/or visual inspection. Based on the simulated polyploids with outcrossing behavior, the proposed method detected the true number of linkage groups with high accuracy, and ordered markers with reasonable precision.

We applied the proposed method to two genotype studies involving barley and potato. In the barley map construction, we correctly detected its seven chromosomes, whereas other method grouped all markers in one linkage group. The netgwas method ordered markers with higher accuracy in most of the chromosomes. The method detected all the potato chromosomes, although it identified chromosome 10 as two linkage groups. Its ordering of markers within each chromosome was a substantial improvement of what has been possible up until now. We remark that the proposed map construction method uses all possible marker types, unlike the other map construction methods, which use a subset of markers (Grandke et al., 2017).

Although modern sequencing methods might be able to create accurate physical maps, there is an important role for methods such as ours that creates a linkage map. Despite its apparently appeal, a physical map merely orders the nucleotides (ATCG) of a chromosome, which defines the physical distance, but it does not give information on the genetic distance between markers. For breeders, knowing about genetic distance is more relevant than physical distance, as it will be more obvious which parts of the genotype will tend to be inherited together. Also, a linkage map is the first requirement for estimating the genetic background of phenotypic traits in quantitative trait loci (QTL) studies. In practice, many software packages for performing QTL analysis require linkage maps, not physical maps. Another reasons why linkage map construction methods continue to be developed is that they enable us to determine linkage disequilibrium structure between polymorphisms (Behrouzi and Wit, 2017a; Matise et al., 2007; Rodriguez-Fontenla et al., 2014) and they are very useful in assembling the genome of a particular species.

Acknowledgement

The authors acknowledge the contribution of the COST Action CA15109 (COSTNET).

Conflict of Interest: none declared.

References

Behrouzi, P. and Wit, E.C. (2017a) Detecting epistatic selection with partially observed genotype data by using copula graphical models. J. R. Stat. Soc. Ser. C (Appl. Stat.).

Behrouzi, P. and Wit, E.C. (2017b) netgwas: an r package for network-based genome-wide association studies. arXiv Preprint arXiv:1710.01236.

Bertola, D.J. et al. (2013) The use of snp markers for linkage mapping in diploid and tetraploid peanut. G3: Genes, Genomes, Genetics, 4, 89–96.

Bourke, P. et al. (2017) polymap: linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. In: BioRxiv, 228817.

Bradshaw, J.E. and Bonierbale, M. (2010) Potatoes. In: Root and Tuber Crops. Springer, pp. 1–52.

Broman, K.W. et al. (2003) R/qtl: qtl mapping in experimental crosses. Bioinformatics, 19, 889–890.

Buotew, K.H. (1991) Influence of aberrant observations on high-resolution linkage analysis outcomes. Am. J. Human Genet., 49, 985.

Cai, T. et al. (2011) A constrained l1 minimization approach to sparse precision matrix estimation. J. Am. Stat. Assoc., 106, 594–607.

Cervantes-Flores, J.C. et al. (2008) Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose asp markers. Mol. Breeding, 21, 511–532.

Castel, L. et al. (2011) Comparative mapping of the Oregon Wolfe barley using doubled haploid lines derived from female and male gametes. Theor. Appl. Genet., 122, 1399–1410.

Cuthill, E. and McKee, J. (1969) Reducing the bandwidth of sparse symmetric matrices. In: Proceedings of the 1969 24th National Conference, ACM, New York, NY, USA, pp. 157–172.

Friedman, J. et al. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9, 432–441.

Grandke, F. et al. (2017) Pergola: fast and deterministic linkage mapping of polyploids. BMC Bioinformatics, 18, 12.

Guo, J. et al. (2015) Graphical models for ordinal data. J. Comput. Graph. Stat., 24, 183–204.

Hackett, C.A. et al. (2017) Tetraploidsnpmap: software for linkage analysis and qtl mapping in autotetraploid populations using snp dosage data. J. Heredity, 108, 438–442.

Jansen, R. et al. (2001) Constructing dense genetic linkage maps. Theor. Appl. Genet., 102, 1113–1122.

Lander, E.S. et al. (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics, 1, 174–181.

Lincoln, S.E. and Lander, E.S. (1992) Systematic detection of errors in genetic linkage data. Genomics, 14, 604–610.

Liu, H. et al. (2012) High-dimensional semiparametric gaussian copula graphical models. Ann. Stat., 40, 2293–2326.

Margarido, G. et al. (2007) Onemap: software for genetic mapping in outcrossing species. Hereditas, 144, 78–79.

Massa, A.N. et al. (2015) Genetic linkage mapping of economically important traits in cultivated tetraploid potato (solanum tuberosus l). G3: Genes Genomes Genet., 5, 2357–2364.

Matise, T.C. et al. (2007) A second-generation combined linkage–physical map of the human genome. Genome Res., 17, 1783–1000.

Mclachlan, G. and Krishnan, T. (2007) The EM Algorithm and Extensions, Vol 382. John Wiley & Sons.
Linkage mapping in polyploids using discrete graphical models

Pang, M. et al. (2017) Quantitative trait loci mapping for feed conversion efficiency in crucian carp (carassius auratus). *Sci. Rep.*, 7, 16971.

Rodriguez-Fontenla, C. et al. (2014) Genetic distance as an alternative to physical distance for definition of gene units in association studies. *BMC Genomics*, 15, 408.

Ronin, Y. et al. (2015) Building ultra-dense genetic maps in the presence of genotyping errors and missing data. In: *Advances in Wheat Genetics: From Genome to Field*, Springer, Tokyo, pp. 127–133.

Simon, M. et al. (2008) QTL mapping in five new large RIL populations of arabidopsis thaliana genotyped with consensus SNP markers. *Genetics*, 178, 2253–2264.

Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: JOIN MAP. *Plant J.*, 3, 739–744.

Wang, H. et al. (2017) The potential of probabilistic graphical models in linkage map construction. *Theor. Appl. Genet.*, 130, 433–444.

Wu, Q.-H. et al. (2015) High-density genetic linkage map construction and QTL mapping of grain shape and size in the wheat population yanda1817 × bei-nong6. *PLoS One*, 10, e0118144.

Wu, Y. et al. (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.*, 4, e1000212.

Yin, J. and Li, H. (2011) A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.*, 5, 2630.