De novo construction of polyploid linkage maps using discrete graphical models

Behrouzi, P., & Wit, E. C.

Behrouzi, P., & Wit, E. C. (2019). De novo construction of polyploid linkage maps using discrete graphical models. Bioinformatics, 35(7), 1083-1093. https://doi.org/10.1093/bioinformatics/bty777

De novo construction of polyploid linkage maps using discrete graphical models

P. Behrouzi
Wageningen University and Research, The Netherlands
pariya.behrouzi@wur.nl

E. C. Wit
Universit della Svizzera italiana, Switzerland

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 nonparanormal 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: 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

Keywords: Linkage mapping; Diploid; Polyploid; Graphical models; Gaussian copula; High-density genotype data.

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 (Wu et al., 2015, Pang et al., 2017) 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 multipoint 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 construct a genetic map from bi-parental populations of outcrossing autopoly-

ploids. 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 autotetraploid species. TetraploidSNPMap, 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., 2017), which does not use the full multivariate 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 very 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.

1 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.

1.1 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. AAAa). 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.

1.2 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 called 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.

1.3 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 multivalent pairing. For example, in tetraploids all four chromosome copies may pair at meiosis. Assume a sequence of ordered SNP markers $X_1^c, X_2^c, \ldots, X_d^c$ along chromosome c in a q-ploid species. We describe the Markov dependence structure between markers for different population schemes (see Figure 1):

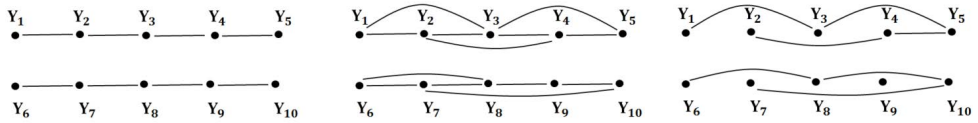
Scheme (i): During meiosis in inbred populations, genetic material from one of the two parents is copied into the offspring in a sequential fashion, i.e. reading along the genome, until the copying switches in a random fashion to the other parent. Thus, the genome of the offspring is a random but piecewise continuous mosaic of the genomes of its parents. The genotype state at each chromosomal region, or locus, of the offspring is either homozygous maternal, heterozygous, or homozygous paternal. For instance, as a result of genetic linkage and crossover a homozygous maternal genotype will typically be followed by a heterozygous genotype before being able to be followed by a homozygous paternal genotype.

Genetic linkage means that markers located close to one another on a chromosome are linked and tend to be inherited together during meiosis. Another key biological fact is that during meiosis markers on different chromosomes segregate independently; this is called the *independent assortment law*.

For example, in scheme (i) consisting of only a homozygous population, the random variable Y_j which represents the genotype of an individual at location j can be defined as

$$Y_j = \begin{cases} 1 & \text{paternal marker at locus } j \text{ on homologue k,} \\ 0 & \text{otherwise.} \end{cases}$$

This scheme occurs in inbred homozygous populations that include only two genotype states, namely homozygous maternal and homozygous paternal. Mapping populations, such as backcrossing, are included in this scheme. Then, under the assumption



(i) (ii) (iii)

Figure 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 $Y_1, \ldots, Y_5$ reside on chromosome 1, and $Y_6, \ldots, Y_{10}$ on chromosome 2.

of no crossover interference – meaning when a crossover has formed, other crossovers are not prevented from forming – the recombination frequency between the two locations j and $j+1$ is independent of recombination at the other locations on the genome. So, the following holds

$$\begin{aligned} Pr(Y_{j+1} = y_{j+1} \mid Y_j = y_j, Y_{j-1} = y_{j-1}, \ldots, Y_1 = y_1) \\ = Pr(Y_{j+1} = y_{j+1} \mid Y_j = y_j) \end{aligned} \tag{1}$$

This equation indicates that the genotype of a marker at location $j+1$ is conditionally independent of genotypes at locations $j-1, j-2, \ldots, 1$ given a genotype at location j. This can be written as

$$Y_{j+1} \perp\!\!\!\perp (Y_1, \ldots, Y_{j-1}) \mid Y_j \tag{2}$$

This defines a discrete graphical model $G = (V, E)$ which consists of vertices $V = \{1, \ldots, p\}$ and edge set $E \subseteq V \times V$ with a binary random variable $Y_j \in \{0, 1\}^p$. Given the above property between neighboring markers, we construct linkage maps using conditional (in)dependence models. Figure 1(i) shows a cartoon image of conditional (in)dependencies for this scheme.

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

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

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 $X_{j\cdot} = \{X_{j1} \ldots X_{jq}\}$. 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 $X_{j1} = 1$, $X_{j2} = 0$, $X_{j3} = 0$, and $X_{j4} = 0$ 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 $aAaa$, $aaAa$, $aaaA$. 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:

$$Y_j = \sum_{k=1}^{q} X_{jk}. \tag{3}$$

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 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 $Y_1, Y_2, \ldots, Y_d$ 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(ii) 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-homozygous parents contain $A_jA_jA_jA_j$ genotype at marker location j, then offspring will also have $A_jA_jA_jA_j$ genotype at marker location j. But we do not know whether that genotype belongs to the paternal or maternal homologue, since both parents have $A_jA_jA_jA_j$ genotype at marker location j. So, in this case we define X_{jk} as follows

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

Table 1
Number of copies (dosage) of a reference allele. Relation between different genotypes, $X_{j.}$, and allele dosage, Y_j, for a tetraploid individual, where A is the reference allele.

Y_j	$X_{j.}$
0	$aaaa$
1	$Aaaa$, $aAaa$, $aaAa$, $aaaA$
2	$AAaa$, $AaAa$, $aAAa$, $AaaA$, $aaAA$, $aAaA$
3	$AAAa$, $AaAA$, $AAaA$, $aAAA$
4	$AAAA$

where A_j 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 generate gametes. In this regard, if both polyploid parents have an A_j allele in all q haploids, then the 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 two homologue chromosomes pair up and form a bivalent, then cross-over before recombinations occur. 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 multinomial 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_j 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 $\{Y_j\}_{j=1,\ldots,d}$ 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 $P(Y)$ can be factorized as,

$$P(Y) = \prod_{c=1}^{C} \prod_{j=1}^{p_c - 1} f_{j,j+1}^{(c)}(Y_j^{(c)}, Y_{j+1}^{(c)}), \tag{4}$$

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 $f_{j,j+1}^{(c)}$ 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.

2 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.

2.1 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 nonparanormal skeptic approach (Liu et al., 2012) (the latter discussed under Supplementary Materials). 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 $Y = (Y_1, \ldots, Y_p)$ 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 multinomial variables $Y = \{0, 1, \ldots, q\}$ where q is the ploidy level. For example in diploids q is 2 and in tetraploids 4.

Sparse ordinal glasso. A relatively straightforward approach to discover the conditional (in)dependence relation 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_j and Z_j define observed rank and true latent value, respectively, where each latent variable corresponds to one observed variable. The relationship between Y_j and Z_j is expressed by a set of cut-points $(-\infty, C_1^{(j)}], (C_1^{(j)}, C_2^{(j)}], \ldots,$ $(C_q^{(j)}, \infty)$, which is obtained by partitioning the range of Z_j into $q_j - 1$ disjoint intervals. Thus, $y_j^{(i)}$, which represents the genotype of the i-th sample for the j-th marker,

can be written as follows

$$y_j^{(i)} = \sum_{k=1}^{q} k \times 1_{\{C_{q-1}^{(j)} < z_j^{(i)} \leq C_q^{(j)}\}} \qquad i = 1, 2, \ldots, n, \tag{5}$$

where we define $\mathcal{D} = \{z_j^{(i)} \in \mathbb{R} \mid C_{q-1}^{(j)} < z_j^{(i)} \leq C_q^{(j)}\}$. 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 Θ, we non-parametrically estimate the cut-points for each $j = 1, \ldots, p$ as follows

$$\widehat{C}_q^{(j)} = \begin{cases} -\infty & \text{if } q = 0 \ ; \\ \Phi^{-1}(\sum_{i=1}^{n} I(y_j^{(i)} \leq q)/n) & \text{if } q = 1, \ \ldots, \ q_j - 1; \\ +\infty & \text{if } q = q_j. \end{cases}$$

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 Θ. 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 $\widehat{\Theta}_\lambda$. 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

$$\begin{aligned} &Q_\lambda(\boldsymbol{\Theta} \mid \widehat{\boldsymbol{\Theta}}^{(m)}) = \\ &\frac{n}{2}\Big[\log|\boldsymbol{\Theta}| - tr(\frac{1}{n}\sum_{i=1}^{n} E_{Z^{(i)}}(Z^{(i)}Z^{(i)t}|y^{(i)}, \widehat{\boldsymbol{\Theta}}^{(m)}, \widehat{\mathcal{D}})\boldsymbol{\Theta}) - p\log(2\pi)\Big] \\ &- \lambda||\Theta||_1 \end{aligned} \tag{6}$$

where λ is a nonnegative tuning parameter. To calculate the conditional expectation $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} E_{Z^{(i)}}(Z^{(i)}Z^{(i)t}|y^{(i)}, \widehat{\Theta}^{(m)}, \widehat{\mathcal{D}})$ 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 Materials. The M-step is a maximization problem which can be solved efficiently using either graphical lasso (Friedman et al., 2008)

$$\widehat{\boldsymbol{\Theta}}_{glasso}^{(m+1)} = \arg\max_{\boldsymbol{\Theta}} \Big\{ \log|\boldsymbol{\Theta}| - tr(\bar{R}\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1 \Big\}$$

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

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

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, $y_j^{(i)}$ 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.

2.2 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 λ (section 2.1). 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

$$eBIC(\lambda) = -2\ell(\widehat{\Theta}_\lambda) + (\log n + 4\gamma \log p)\text{df}(\lambda), \tag{7}$$

where $\ell(\widehat{\Theta}_\lambda)$ is the non-penalized likelihood and $\gamma \in [0, 1]$ is an additional parameter. And $df(\lambda) = \sum_{1\leq i<j\leq p} I(\widehat{\theta}_{ij,\lambda} \neq 0)$ where $\widehat{\theta}_{ij,\lambda}$ is (i, j)th entry of the estimated precision matrix $\widehat{\Theta}_\lambda$ and I is the indicator function. In case of $\gamma = 0$ the classical BIC is obtained. Typical values for γ are $1/2$ and 1. We select the value of λ 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.

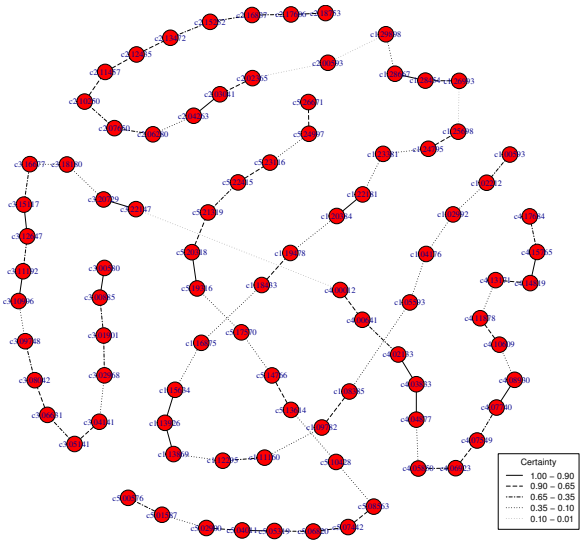


Figure 2: The certainty associated with the linkage map estimation in A.thaliana using the non-parametric bootstrap.

2.3 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 $\widehat{\Theta}_{\lambda}^{(d)}$, a submatrix of $\widehat{\Theta}_{\lambda}$, 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.

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 multidimensional 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 $D_{ij} = -\log(\rho_{ij})$ for $i \neq j$. Here, the matrix ρ represents the conditional correlation among d objects which can be obtained as $\rho_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}}\sqrt{\theta_{jj}}}$, where θ_{ij} is the ij-th element of the precision matrix Θ.

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 dis-

tance $\widehat{D}$ between them is similar to D. We consider a metric MDS, which minimizes $\widehat{L} = \arg\min_L \sum_{i=1}^{d}\sum_{j=1}^{d}(D_{ij} - \widehat{D}_{ij})^2$ across all linear orderings.

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 [see Figure 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 $\widehat{\Theta}_{\lambda}^{(d)}$. The bandwidth of the matrix A is defined by $\beta = \max_{\theta_{ij}\neq 0}|i-j|$. The RCM algorithm produces a permutation matrix P such that PAP^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 $G(V_d, E_d)$ in consecutive order. Moreover, all of the nonzero elements are clustered near the main diagonal.

2.4 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 data set $Cvi \times Col$ in *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 (Simon 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 subsample of $n = 50$ plants. The line type shows the estimated certainty associated with each link. For example, the gray dotted between marker "c2.00593" from chromosome 2 and marker "c1.298998" from chromosome 1

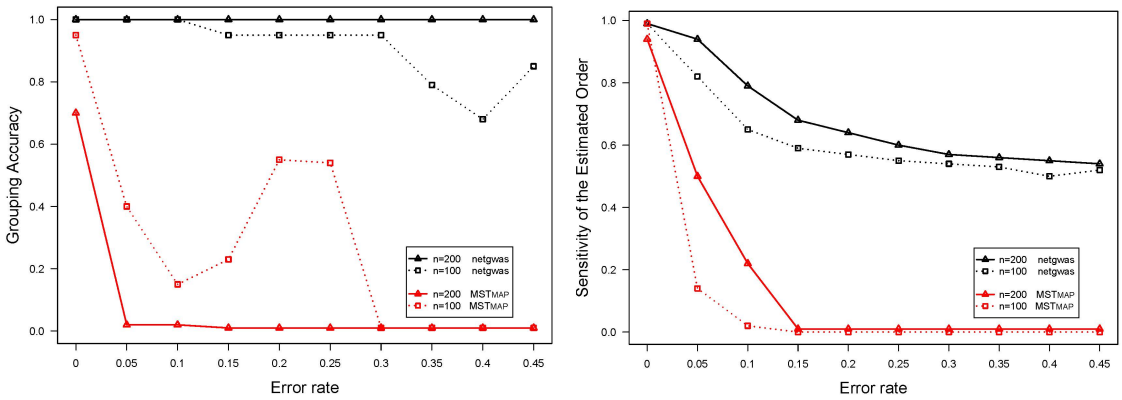


Figure 3: Comparison of performance between map construction in `netgwas` and MSTMAP for different genotyping error rates. Simulated data contain 300 markers for different numbers of individual n. Top figure reports average grouping, and bottom figure shows average ordering accuracy scores over 50 independent runs.

has the certainty value of 0.01, likewise for markers "c3.22147" 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 subsample $n = 50$.

3 Simulation study

In this section, we study the performance of the proposed method for different diploids and polyploids. In section 3.1 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 3.2 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.

3.1 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 predefined 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: $GA = \frac{1}{1+(LG-\widehat{LG})^2}$, where LG stands for actual number of linkage groups and $\widehat{LG}$ 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 2.5GHz Intel Xeon processors and 128GB 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 five chromosomes and 200 F2 individuals, they run in 1.4 and 0.20 min, respectively, on Intel i7 laptop with 16Gb 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`.

Table 2
Summary of performance measures of linkage map construction in simulated F2 populations for netgwas, MSTMAP and JOINMAP at different rates of missingness and genotyping errors. The Tables presents average grouping and ordering accuracy scores for 50 independent runs and standard deviation in parentheses. Best scores are boldfaced.

		Grouping Accuracy			Ordering Accuracy		
Missing rate	Error rate	netgwas	MSTMap	JOINMAP	netgwas	MSTMap	JOINMAP
p=1000 & n=200							
0	0	**1.00** (0.00)	0.61 (0.36)	0.00 (0.00)	**1.00** (0.00)	0.91 (0.06)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.04 (0.03)	0.00 (0.00)	**0.56** (0.00)	0.51 (0.09)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.44 (0.16)	0.00 (0.00)	0.52 (0.00)	**0.78** (0.02)	0.00 (0.00)
0.15	0.15	**1.00** (0.01)	0.05 (0.00)	0.00 (0.00)	**0.52** (0.00)	0.60 (0.13)	0.00 (0.00)
p=1000 & n=100							
0	0	**1.00** (0.00)	0.74 (0.35)	0.00 (0.00)	**1.00** (0.00)	0.82 (0.08)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.13 (0.07)	0.00 (0.00)	**0.53** (0.01)	0.50 (0.04)	0.00 (0.00)
0.10	0.10	**0.95** (0.16)	0.01 (0.00)	0.00 (0.00)	**0.52** (0.01)	0.13 (0.16)	0.00 (0.00)
0.15	0.15	**0.95** (0.15)	0.00 (0.00)	0.00 (0.00)	**0.49** (0.04)	0.00 (0.00)	0.00 (0.00)

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 (Figure 3, top) and ordering accuracies (Figure 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 (≥ 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 ϵ remains a critical issue in MSTMAP to detect correct number of linkage groups. In their clustering approach, ϵ does not only depends on n, but it also depends on the Hamming distance between linkage groups. Therefore, it it

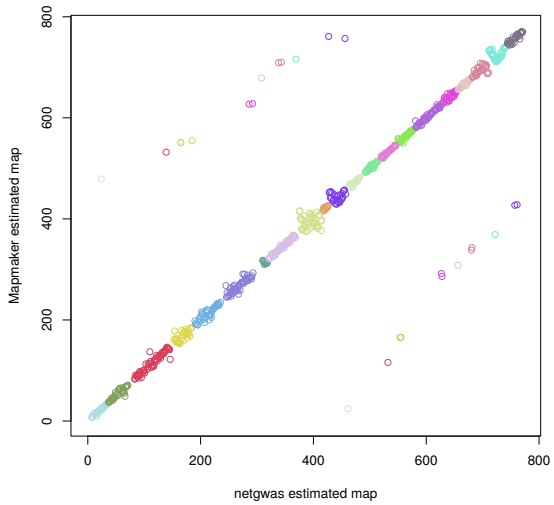


Figure 4: Tetraploid Peanut linkage map comparison. Performance of `netgwas` and `MapMaker` on map construction for tetraploid peanut. There is a high level of consistency between MapMaker and netgwas.

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 (see Figure 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 `R\qtl` package (Broman et al., 2003) to detect genotyping errors.

Evaluation of estimated maps for incomplete and noisy data

In Table 4 we simulate inbred F2 populations with 10 linkage groups that contain different rates of missing and genotyping error. Here, we report average grouping

and ordering accuracy scores over 50 independent simulated datasets. In all scenarios, netgwas detects linkage groups with higher accuracy. Furthermore, the netgwas performs well in correctly ordering markers as its ordering accuracy scores are higher compared to the other two methods, except in one case where MSTmap performs better.

We remark that the ordering accuracy scores in Table 4 should be interpreted carefully, as inversion in the order of flanking markers reduces the number of correct ordering and ultimately decreases orderings accuracy scores. Furthermore, determining linkage groups (LGs) in JoinMap requires an input parameter to be specified by the user, whereas the other two methods determine LGs in data-driven way. Thus, to treat all the three methods equally we used a conservative LOD score threshold as suggested by Stam (2012) to detect LGs. The fact that JoinMap scores so badly is the result of these automated selection of the tuning parameters. The zero grouping accuracy is because markers were incorrectly assigned to many more linkage groups than the true number of linkage groups, where the zero ordering accuracy comes from the large Jacard distance between the estimated order of JoinMap and the true order. Practitioners have reported better results when manually tuning the various parameters of JoinMap.

3.2 Polyploid species

The peanut (*Arachis hypogaea L.*) is an important oilseed crop and food legume grown in tropical or subtropical 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 markers the maps were similar, except in chromosomes TA9 and TA10. In addition, netgwas runs in 22.44 seconds on an Intel i7 laptop with 16GB RAM. The run-time of MapMaker is unknown because the authors of the peanut dataset did not provide computational times.

Estimated number of linkage groups (LGs) for OWB data set

	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 assumed MSTMAP has estimated correctly the number of LGs in the OWB data set.

Linkage Group (LG)	Sensitivity Score	
	netgwas	MSTMap
1	0.86	**0.96**
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

Figure 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 estimated undirected graph in `netgwas` for the barley data. This consists of 7 sub–graphs, each showing a chromosome.

4 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, $Y_{ij} \in \{0, 1\}$. 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 $\widehat{\Theta}_\lambda$, which contains conditional (in)dependence relationships between barley markers, we were able to correctly detect the 7 barley chromosomes as sub–graphs in the estimated undirected graph. Furthermore, using the conditional correlation matrix as distance in the multi-dimensional 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 Cistué 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 7 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 Table 1 in supplementary materials). Based on this result, we remark that segregation distortion seems to have little influence on `netgwas` map construction.

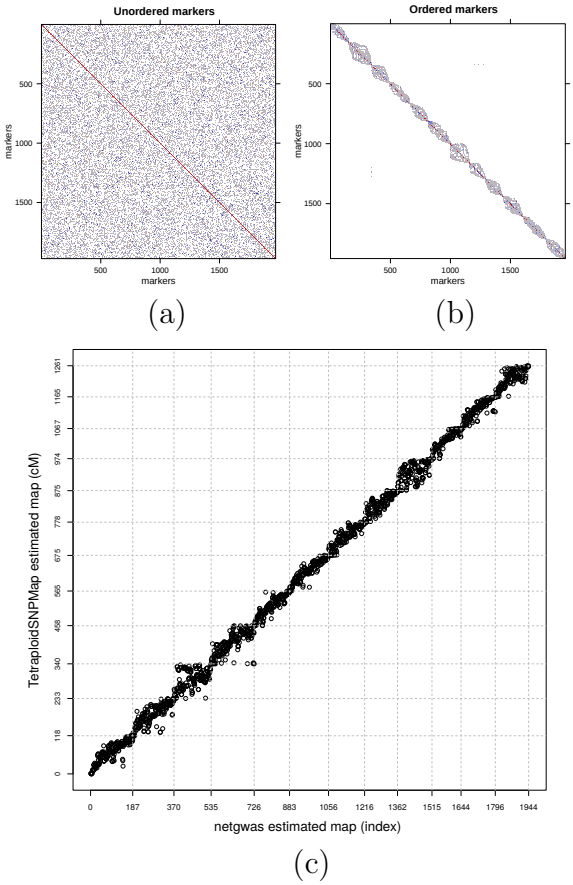
5 Construction of linkage map for tetraploid potato

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



(a) (b)

(c)

Figure 6: Construction linkage map in potato. (a) Estimated precision matrix for unordered genotype data of tetraploid potato. (b) Estimated precision matrix after ordering markers. (c) Estimated order of markers across potato genome, versus estimated order in tetraploidSNPmap software. Each dashed line represents a chromosome. All potato chromosomes were detected correctly in netgwas.

207 F1 individuals netgwas produced maps within 5.5 minutes, whereas TetraploidSNPMap took 15 minutes on an Intel i7 laptop with 16GB RAM.

6 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 $\widehat{\Theta}_\lambda$. This method iteratively computes the conditional expectation of the complete penalized log-likelihood, and optimizes it to estimate $\widehat{\Theta}_\lambda$. The method can also deal with missing values, which are very common in genotype datasets. The nonparanormal 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 7 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 (Matise et al., 2007, Rodriguez-Fontenla et al., 2014, Behrouzi and Wit, 2017a) and they are very useful in assembling the genome of a particular species.

References

Behrouzi, P. and E. C. Wit (2017a). Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

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

Bertioli, D. J., P. Ozias-Akins, Y. Chu, K. M. Dantas, S. P. Santos, E. G. Gouvea, P. M. Guimarães, S. C. M. Leal-Bertioli, S. J. Knapp, and M. C. Moretzsohn (2013). The use of snp markers for linkage mapping in diploid and tetraploid peanut. *G3: Genes, Genomes, Genetics*, g3–113.

Bourke, P., G. van Geest, R. E. Voorrips, J. Jansen, T. Kranenburg, A. Shahin, R. G. Visser, P. Arens, M. J. Smulders, and C. Maliepaard (2017). polymapr: linkage analysis and genetic map construction from f1 populations of outcrossing polyploids. *bioRxiv*, 228817.

Bradshaw, J. E. and M. Bonierbale (2010). Potatoes. In *Root and tuber crops*, pp. 1–52. Springer.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics 19*(7), 889–890.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *American journal of human genetics 49*(5), 985.

Cai, T., W. Liu, and X. Luo (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cervantes-Flores, J. C., G. C. Yencho, A. Kriegner, K. V. Pecota, M. A. Faulk, R. O. Mwanga, and B. R. Sosinski (2008). Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose aflp markers. *Molecular Breeding 21*(4), 511–532.

Cistué, L., A. Cuesta-Marcos, S. Chao, B. Echávarri, Y. Chutimanitsakun, A. Corey, T. Filichkina, N. Garcia-Mariño, I. Romagosa, and P. M. Hayes (2011). Comparative mapping of the oregon wolfe barley using doubled haploid lines derived from female and male gametes. *Theoretical and applied genetics 122*(7), 1399–1410.

Cuthill, E. and J. McKee (1969). Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pp. 157–172. ACM.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Grandke, F., S. Ranganathan, N. van Bers, J. R. de Haan, and D. Metzler (2017). Pergola: fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics 18*(1), 12.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics 24*(1), 183–204.

Hackett, C. A., B. Boskamp, A. Vogogias, K. F. Preedy, and I. Milne (2017). Tetraploidsnpmap: software for linkage analysis and qtl mapping in autotetraploid populations using snp dosage data. *Journal of Heredity 108*(4), 438–442.

Jansen, J., A. De Jong, and J. Van Ooijen (2001). Constructing dense genetic linkage maps. *Theoretical and Applied Genetics 102*(6-7), 1113–1122.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg (1987). Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics 1*(2), 174–181.

Lincoln, S. E. and E. S. Lander (1992). Systematic detection of errors in genetic linkage data. *Genomics 14*(3), 604–610.

Liu, H., F. Han, M. Yuan, J. Lafferty, L. Wasserman, et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics 40*(4), 2293–2326.

Margarido, G., A. Souza, and A. Garcia (2007). Onemap: software for genetic mapping in outcrossing species. *Hereditas 144*(3), 78–79.

Massa, A. N., N. C. Manrique-Carpintero, J. J. Coombs, D. G. Zarka, A. E. Boone, W. W. Kirk, C. A. Hackett, G. J. Bryan, and D. S. Douches (2015). Genetic linkage mapping of economically important traits in cultivated tetraploid potato (solanum tuberosum l.). *G3: Genes— Genomes— Genetics 5*(11), 2357–2364.

Matise, T. C., F. Chen, W. Chen, M. Francisco, M. Hansen, C. He, F. C. Hyland, G. C. Kennedy, X. Kong, S. S. Murray, et al. (2007). A second-generation combined linkage–physical map of the human genome. *Genome research 17*(12), 000–000.

McLachlan, G. and T. Krishnan (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons.

Pang, M., B. Fu, X. Yu, H. Liu, X. Wang, Z. Yin, S. Xie, and J. Tong (2017). Quantitative trait loci mapping for feed conversion efficiency in crucian carp (carassius auratus). *Scientific reports 7*(1), 16971.

Rodriguez-Fontenla, C., M. Calaza, and A. Gonzalez (2014). Genetic distance as an alternative to physical distance for definition of gene units in association studies. *BMC genomics 15*(1), 408.

Ronin, Y., D. Minkov, D. Mester, E. Akhunov, and A. Korol (2015). Building ultradense genetic maps in the presence of genotyping errors and missing data. In *Advances in Wheat Genetics: From Genome to Field*, pp. 127–133. Springer.

Simon, M., O. Loudet, S. Durand, A. Bérard, D. Brunel, F. Sennesal, M. Durand-Tardif, G. Pelletier, and C. Camilleri (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. *The Plant Journal 3*(5), 739–744.

Stam, P. (2012). Lecture notes on construction of genetic linkage maps.

Wang, H., F. A. van Eeuwijk, and J. Jansen (2017). The potential of probabilistic graphical models in linkage map construction. *Theoretical and Applied Genetics 130*(2), 433–444.

Wu, Q.-H., Y.-X. Chen, S.-H. Zhou, L. Fu, J.-J. Chen, Y. Xiao, D. Zhang, S.-H. Ouyang, X.-J. Zhao, Y. Cui, et al. (2015). High-density genetic linkage map construction and qtl mapping of grain shape and size in the wheat population yanda1817× beinong6. *PloS one 10*(2), e0118144.

Wu, Y., P. R. Bhat, T. J. Close, and S. Lonardi (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS genetics 4*(10), e1000212.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics 5*(4), 2630.

Supplementary Materials

Computing conditional expectation

We calculate $\bar{R}$ in equation (6) of the paper as

$$E\left[Z^{(i)}Z^{(i)t}|Y^{(i)},\widehat{\Theta}^{(m)}\right]=E\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right]E\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right]^{t}+cov\left[Z^{(i)}|Y^{(i)},\widehat{\Theta}^{(m)}\right] \tag{8}$$

The conditional random variable $Z|Y$ follows a truncated p-variate normal distribution. Wilhelm et al,.(2010) provided the analytical solution to compute moments of truncated multivariate normal distribution. However, their approach is feasible for only very few variables. Here, we propose instead to simulate a large number of samples from the truncated p-variate normal distribution and compute the sample conditional covariance matrix and sample conditional mean to estimate $E\big[Z^{(i)}Z^{(i)t}|Y^{(i)},\widehat{\Theta}^{(m)}\big]$ using the equation (8).

Alternatively, we use an efficient approximate estimation algorithm, which is implemented in Behrouzi and Wit (2017) and Guo et al,. (2015). The variance elements in the conditional expectation matrix can be calculated through the second moment of the conditional $Z_j^{(i)} \mid Y^{(i)}$, and the rest of the elements in this matrix can be approximated through $E(Z_j^{(i)}Z_{j'}^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}}) \approx E(Z_j^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}})\ E(Z_{j'}^{(i)} \mid y^{(i)};\widehat{\Theta},\widehat{\mathcal{D}})$ using mean field theory. The first and second moment of $z_j^{(i)}|y^{(i)}$ can be written as

$$E(Z_j^{(i)} \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) = E[E(Z_j^{(i)} \mid z_{-j}^{(i)},y_j^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}], \tag{9}$$

$$E((Z_j^{(i)})^2 \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) = E[E((Z_j^{(i)})^2 \mid z_{-j}^{(i)},y_j^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}) \mid y^{(i)},\widehat{\Theta},\widehat{\mathcal{D}}], \tag{10}$$

where $z_{-j}^{(i)} = (z_1^{(i)},\ldots,z_{j-1}^{(i)},z_{j+1}^{(i)},\ldots,z_p^{(i)})$. The inner expectations in (9) and (10) are relatively straightforward to calculate. $z_j^{(i)} \mid z_{-j}^{(i)},y_j^{(i)}$ follows a truncated Gaussian distribution on the interval $[c^{(j)}_{y_j^{(i)}},c^{(j)}_{y_j^{(i)}+1}]$ with parameters $\mu_{i,j}$ and $\sigma_{i,j}^2$ given by

$$\mu_{ij} = \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1}z_{-j}^{(i)t},$$

$$\sigma_{i,j}^2 = 1-\widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1}\widehat{\mathbf{\Sigma}}_{-j,-j}.$$

Let $r_{k,l}=\frac{1}{n}\sum_{i=1}^{n}E(Z_k^{(i)}Z_l^{(i)} \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}})$ be the (k,l)-th element of empirical correlation matrix $\bar{R}$, then to obtain the $\bar{R}$ two simplifications are required.

$$\begin{aligned}
&E(Z_k^{(i)}Z_l^{(i)t} \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}}) \approx E(Z_k^{(i)} \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}})E(Z_l^{(i)} \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}}) && \text{if } 1\le k\neq l\le p,\\
&E(Z_k^{(i)}Z_l^{(i)t} \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}}) = E((Z_k^{(i)})^2 \mid y^{(i)},\widehat{\mathbf{\Theta}},\widehat{\mathcal{D}}) && \text{if } k=l.
\end{aligned}$$

Applying the results in the appendix to the conditional $z_j^{(i)} \mid z_{-j}^{(i)}, y_j^{(i)}$ we obtain

$$E(Z_j^{(i)} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) = \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^t} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) + \frac{\phi(\widehat{\delta}^{(i)}_{j,y_j^{(i)}}) - \phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}})}\tilde{\sigma}_j^{(i)}, \tag{11}$$

$$\begin{aligned} E((Z_j^{(i)})^2 \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}}) &= \widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^t} Z_{-j}^{(i)} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1}\widehat{\mathbf{\Sigma}}_{j,-j}^{t} + (\tilde{\sigma}_j^{(i)})^2 \\ &+ 2\frac{\phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}}) - \phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}})}[\widehat{\mathbf{\Sigma}}_{j,-j}\widehat{\mathbf{\Sigma}}_{-j,-j}^{-1} E(Z_{-j}^{(i)^t} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})]\tilde{\sigma}_j^{(i)} \\ &+ \frac{\delta^{(i)}_{j,y_j^{(i)}}\phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}}) - \tilde{\delta}^{(i)}_{j,y_j^{(i)}+1}\phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1})}{\Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}+1}) - \Phi(\tilde{\delta}^{(i)}_{j,y_j^{(i)}})}(\tilde{\sigma}_j^{(i)})^2, \end{aligned} \tag{12}$$

where $Z_{-j}^{(i)} = (Z_1^{(i)}, \ldots, Z_{j-1}^{(i)}, Z_{j+1}^{(i)}, \ldots, Z_p^{(i)})$ and $\tilde{\delta}^{(i)}_{j,y_j^{(i)}} = [c_j^{(i)} - E(\tilde{\mu}_{ij} \mid y^{(i)}; \widehat{\mathbf{\Theta}}, \widehat{\mathcal{D}})]/\tilde{\sigma}_{ij}$. In this way, an approximation for $\bar{R}$ is obtained as follows:

$$\tilde{r}_{kl} = \begin{cases} \frac{1}{n}\sum_{i=1}^{i=n} E(Z_k^{(i)} \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}}) E(Z_l^{(i)} \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}}) & \text{if } 1 \leq k \neq l \leq p \\ \frac{1}{n}\sum_{i=1}^{i=n} E((Z_k^{(i)})^2 \mid y^{(i)}, \widehat{\mathbf{\Theta}}^{(m)}, \widehat{\mathcal{D}}) & \text{if } k = l. \end{cases}$$

The latent graphical model discussed in the paper, though it is a natural approach, is computationally expensive for a large number of variables ($p > 2000$). We therefore describe here an alternative method to construct high–dimensional undirected graphical models.

Nonparanormal SKEPTIC

As alternative, we use the nonparanormal skeptic approach (Liu et al., 2012) to estimate the penalized concentration matrix Θ. In this approach, instead of using the transformed data to estimate precision matrix Θ, a sample correlation matrix Γ can be computed from pairwise rank correlations, such as Kendall's tau and Spearman's rho. For the random vector $y_j^{(1)}, \ldots, y_j^{(n)}$ the Kendall's tau and Spearman's rho are given, respectively, by

$$\widehat{\tau}_{jl} = \frac{2}{n(n-1)} \sum_{i,\acute{i}=1}^{n} \text{sign}(y_j^{(i)} - y_j^{(\acute{i})})(y_l^{(i)} - y_l^{(\acute{i})})$$

and

$$\widehat{\rho}_{jl} = \frac{\sum_{i=1}^{n}(r_j^i - \bar{r}_j)(r_l^i - \bar{r}_l)}{\sqrt{\sum_{i=1}^{n}(r_j^i - \bar{r}_j)^2 . \sum_{i=1}^{n}(r_l^i - \bar{r}_l)^2}}$$

$$\widehat{\Gamma}_{jl} = \begin{cases} \sin(\frac{\pi}{2}\widehat{\tau}_{jl}) & j \neq l \\ 1 & j = l \end{cases} \quad ; \quad \widehat{\Gamma}_{jl} = \begin{cases} 2\sin(\frac{\pi}{6}\widehat{\rho}_{jl}) & j \neq l \\ 1 & j = l. \end{cases}$$

To estimate the sparse precision matrix and the graph, one can use either the graphical lasso

$$\widehat{\Theta}_{\text{glasso}} = \arg\max_{\boldsymbol{\Theta}} \left\{ \log|\boldsymbol{\Theta}| - tr(\Gamma\boldsymbol{\Theta}) - \lambda||\boldsymbol{\Theta}||_1 \right\} \tag{13}$$

or CLIME estimator, with $\widehat{\Gamma}$ as input

$$\widehat{\Theta}_{\text{CLIME}} = \arg\min_{\Theta} ||\Theta||_1 \qquad \text{subject to} \qquad ||\widehat{\Gamma}\Theta - I_p||_\infty \leq \lambda, \tag{14}$$

Although both methods involve convex optimization problems, these can be efficiently solved.

Evaluation of estimated maps for incomplete and noisy data

In Table 4 we simulate inbred F2 populations with 10 linkage groups that contain different rates of missing and genotyping error. Here, we report average grouping and ordering accuracy scores over 50 independent simulated datasets. In all scenarios, netgwas detects linkage groups with higher accuracy. Furthermore, the netgwas performs well in correctly ordering markers as its ordering accuracy scores are higher compared to the other two methods, except in one case where MSTMAP performs better.

We remark that the ordering accuracy scores in Table 4 should be interpreted carefully, as inversion in the order of flanking markers reduces the number of correct ordering and ultimately decreases orderings accuracy scores. Furthermore, determining linkage groups (LGs) in JOINMAP requires an input parameter to be specified by the user, whereas the other two methods determine LGs in data-driven way. Thus, to treat all the three methods equally we used a conservative LOD score threshold as suggested by Stam (2012) to detect LGs. The fact that JOINMAP scores so badly is the result of these automated selection of the tuning parameters. The zero grouping accuracy is because markers were incorrectly assigned to many more linkage groups than the true number of linkage groups, where the zero ordering accuracy comes from the large Jacard distance between the estimated order of JOINMAP and the true order. Practitioners have reported better results when manually tuning the various parameters of JOINMAP.

Table 3
Influence of distorted markers on map construction for wheat population in `netgwas`. We used a Double Haploid wheat population, which contains a set of markers with segregation distortion. The `netgwas` ordered distorted markers in the correct linkage group (LG).

distorted marker(s)	known LG	netgwas LG	marker order in netgwas
1A.m.34 1A.m.37	1A	1	1A.m.1, 1A.m.2, 1A.m.3, 1A.m.13, 1A.m.4, 1A.m.5, 1A.m.6, 1A.m.7 1A.m.8, 1A.m.9, 1A.m.10, 1A.m.11, 1A.m.12, 1A.m.15, 1A.m.14 1A.m.18, 1A.m.16, 1A.m.17, 1A.m.19 1A.m.21 1A.m.20 1A.m.22, 1A.m.23, 1A.m.24, 1A.m.25, 1A.m.26, 1A.m.27, 1A.m.28, 1A.m.29, 1A.m.30, 1A.m.32, 1A.m.33, **1A.m.34**, 1A.m.31, 1A.m.35, 1A.m.36, **1A.m.37**, 1A.m.38, 1A.m.39, 1A.m.40, 1A.m.41
3B.m.15 3B.m.16 3B.m.17 3B.m.18 3B.m.19 3B.m.20	3B	9	3B.m.1, 3B.m.2, 3B.m.4, 3B.m.3, 3B.m.5, 3B.m.6, 3B.m.7, 3B.m.8, 3B.m.9, 3B.m.10, 3B.m.13, 3B.m.12, 3B.m.11, 3B.m.14, **3B.m.15**, **3B.m.18**, **3B.m.16**, **3B.m.19**, **3B.m.17**, **3B.m.20**, 3B.m.21, 3B.m.22, 3B.m.23, 3B.m.24, 3B.m.26, 3B.m.25, 3B.m.27, 3B.m.28, 3B.m.29, 3B.m.30
6D.m.12	6D	20	6D.m.5, 6D.m.1, 6D.m.2, 6D.m.3, 6D.m.4, 6D.m.6, 6D.m.7, 6D.m.8, 6D.m.9, 6D.m.10, 6D.m.11, **6D.m.12**, 6D.m.14, 6D.m.13, 6D.m.15
7B.m.6	7B	22	7B.m.1, 7B.m.2, **7B.m.6**, 7B.m.3, 7B.m.4, 7B.m.7, 7B.m.5, 7B.m.8, 7B.m.9, 7B.m.12, 7B.m.10, 7B.m.11, 7B.m.13, 7B.m.17, 7B.m.14, 7B.m.15, 7B.m.16, 7B.m.18, 7B.m.19, 7B.m.20, 7B.m.21, 7B.m.22, 7B.m.23, 7B.m.24, 7B.m.25, 7B.m.26 7B.m.27, 7B.m.28, 7B.m.29, 7B.m.30, 7B.m.32, 7B.m.31, 7B.m.33, 7B.m.34, 7B.m.36, 7B.m.35, 7B.m.37

Table 4
Summary of performance measures of linkage map construction in simulated F2 populations for netgwas, MSTMAP and JOINMAP at different rates of missingness and genotyping errors. This Table presents average grouping and ordering accuracy scores for 50 independent runs and standard deviation in parentheses. Best scores are boldfaced.

		Grouping Accuracy			Ordering Accuracy		
Missing rate	Error rate	netgwas	MSTMap	JOINMAP	netgwas	MSTMap	JOINMAP
p=1000 & n=200							
0	0	**1.00** (0.00)	0.61 (0.36)	0.00 (0.00)	**1.00** (0.00)	0.91 (0.06)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.04 (0.03)	0.00 (0.00)	**0.56** (0.00)	0.51 (0.09)	0.00 (0.00)
0.10	0.10	**1.00** (0.00)	0.44 (0.16)	0.00 (0.00)	0.52 (0.00)	**0.78** (0.02)	0.00 (0.00)
0.15	0.15	**1.00** (0.01)	0.05 (0.00)	0.00 (0.00)	**0.52** (0.00)	0.60 (0.13)	0.00 (0.00)
p=1000 & n=100							
0	0	**1.00** (0.00)	0.74 (0.35)	0.00 (0.00)	**1.00** (0.00)	0.82 (0.08)	0.00 (0.00)
0.05	0.05	**1.00** (0.00)	0.13 (0.07)	0.00 (0.00)	**0.53** (0.01)	0.50 (0.04)	0.00 (0.00)
0.10	0.10	**0.95** (0.16)	0.01 (0.00)	0.00 (0.00)	**0.52** (0.01)	0.13 (0.16)	0.00 (0.00)
0.15	0.15	**0.95** (0.15)	0.00 (0.00)	0.00 (0.00)	**0.49** (0.04)	0.00 (0.00)	0.00 (0.00)