Using Probabilistic Graphical Models to Reconstruct Biological Networks and Linkage Maps

Huange Wang

Thesis committee

Promotor
Prof. Dr F.A. van Eeuwijk
Professor of Applied Statistics
Wageningen University & Research

Co-promotor
Dr J. Jansen
Senior Research Scientist, Biometris
Wageningen University & Research

Other members
Prof. Dr J.J. Houwing-Duistermaat, University of Leeds, UK
Prof. Dr A.K. Smilde, University of Amsterdam
Prof. Dr E.C. Wit, University of Groningen
Prof. Dr R.F. Veerkamp, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Production Ecology & Resource Conservation (PE & RC)

Using Probabilistic Graphical Models to Reconstruct Biological Networks and Linkage Maps

Huange Wang

Thesis
submitted in fulfilment of the requirement for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 18 May 2017
at 11 a.m. in the Aula.

Huange Wang

Using probabilistic graphical models to reconstruct biological networks and linkage maps,
150 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with summary in English

ISBN 978-94-6343-153-8
DOI http://dx.doi.org/10.18174/410875

Contents

Chapter 1 General introduction 1

Chapter 2 A new method to infer causal phenotype networks using QTL and phenotypic information 9

Chapter 3 Genotype-phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and QTLs in ripe tomatoes 35

Chapter 4 The potential of probabilistic graphical models in linkage map construction 65

Chapter 5 A comparative simulation study of the PC algorithm and the Metropolis-Hastings algorithm in constructing random and scale-free Bayesian networks 103

Chapter 6 General discussion 131

Summary 143

Acknowledgement 145

List of publications 147

Curriculum vitae 149

Chapter 1

General introduction

1.1 Mining for genotype-phenotype relationships

An integrative view of diversity and singularity in the living world requires a better understanding of the intricate link between genotypes and phenotypes (Orgogozo et al. 2015). Genotype, i.e. the “internally coded, inheritable information” carried by a living organism, holds the critical instructions that are used and interpreted by cellular machinery to produce a phenotype, i.e. the “outward, physical manifestation” of the organism. However, the relationship of genotype to phenotype is not one-to-one; phenotypes typically result from interactions between the products of multiple genes (Lehner 2007).

Existing methods for unraveling the genetic architecture of complex traits mainly identify genomic regions associated with phenotypic variation through standard quantitative trait locus (QTL) analysis. But between genes and the final phenotypes, there exist a few intermediary substances such as proteins and metabolites, which have a quantitative nature and vary among individuals within populations. Successfully linking variations at intermediate levels to allelic variations on the one hand and to phenotypic variations on the other hand is expected to improve the prediction and manipulation of complex traits, which are crucial in plant and animal breeding.

1.2 Inferring causal relationships among phenotypic traits

Estimates of phenotypic correlations are widespread throughout the literature of plant and animal breeding (Searle 1961). According to quantitative genetic theory, genetic and environmental causes of correlation combine together to produce the phenotypic correlation (Waitt and Levin 1998); hence the phenotypic variance is conventionally partitioned into genetic and environmental components.

However, beyond phenotypic correlations, causal relationships among phenotypes have attracted increasing research interest in recent years. Investigation on the causal structure of genetic data used to be ignored simply because it is well-known that genotype is the cause and phenotype is the effect, and the reverse scenario has been proven to be unlikely (Li et al. 2006). Of late, it has been realized that phenotypic traits may exert causal effects between them (Rosa et al. 2011), especially when intermediate phenotypes are involved. For example, two molecular traits can serve, respectively, as the upstream and downstream elements in a biochemical pathway. Then the upstream one is considered as a cause, and the downstream one is considered as an effect.

Causal inference in phenotypic traits, or equivalently, the construction of causal phenotype networks was so far mainly based upon logic that involves underlying QTLs (Jansen et al. 2009). In practice, it has become fashionable to map QTLs for phenotypes of interest via genome-wide scans, since genotyping has become cheaper and easier thanks to the advancement of genome sequencing technologies. Contrariwise, phenotyping, especially that of late-emerging traits in long-lived species, remains time-consuming and thus expensive (Monneveux et al. 2012). This imposes limits on the sample size (Ríos 2015), i.e., leads to limited sample sizes that provide insufficient power for detecting small to medium sized QTLs. Thus, it is often the case that no QTL is identified for some of the traits of interest. In such cases, the state-of-the-art methods for inferring causal relationships among phenotypes become invalid because they require at least one unique QTL for each trait studied (Logsdon and Mezey 2010; Neto et al. 2008).

1.3 Genetic mapping

Genetic map construction remains an important prerequisite for QTL analysis in organisms for which genomic sequence is not available (Broman 2010). In view of the fact that the closer the two markers are on a chromosome, the more likely they are to be passed together on to the next generation, the "co-segregation" patterns of markers are believed useful for genetic mapping. For this reason, current approaches to map construction are mainly based on the estimates of recombination frequencies between genetic markers.

These approaches, though theoretically true, can produce reasonably informative genetic maps; their practical performances have proven largely dependent on the quality of the marker data. It is known that genotyping errors inflate the number of apparent recombinations, and thus expand map distances and reduce the proportion of correctly ordered maps, especially when marker density increases (Göring and Terwilliger 2000; Hackett and Broadfoot 2003; Shields et al. 1991). Markers exhibiting high nearest-neighbour stress (N.N.Stress) are generally considered to have genotyping errors (Van Ooijen and Jansen 2013) and are therefore often removed from constructed genetic maps (Farré et al. 2011; Ting et al. 2013). Nonetheless, it should be noted that this post-hoc filtering is inherently biased because it is applied to marker orders that are obtained under the assumption of no error.

Moreover, few methods have been proposed for genetic map construction in the case of chromosomal rearrangements such as reciprocal translocations. A reciprocal translocation refers to an even exchange of DNA fragments between two non-homologous chromosomes. Recombination between loci around the translocation breakpoints is severely suppressed. As a consequence, markers in these regions become ‘pseudo-linked’, that is, markers that lie on different chromosomes involved in the translocation will be mapped into a single linkage group (Farré et al. 2011).

1.4 Probabilistic graphical models

Reverse-engineering of biological networks is a central research problem in computational and systems biology. Earlier approaches to solving this problem mainly resort to clustering and correlation analysis, which are rather straightforward but with limited effectiveness. More specifically, clustering is able to uncover the modular topology but cannot explore in depth the fine architecture of each module (Hanisch et al. 2002; Jiang and Singh 2010; Muff et al. 2005; Ravasz et al. 2002); correlation analysis is known to not only confound direct and indirect associations but also provide no means to distinguish between cause and effect (Opgen-Rhein and Strimmer 2007).

Recent studies show that probabilistic graphical models (PGMs), which combine the graph theory and probability theory to give a multivariate statistical modeling, have been successfully used to reconstruct a wide range of biological networks (Airoldi 2007; Friedman 2004). PGMs are categorized into two general types based on the nature of edges in the resulting network: undirected graphical models and directed graphical models. A representative subgroup of undirected graphical models is Gaussian graphical models (GGMs), which allow the identification of conditional independence relationships between variables under the assumption of multivariate Gaussian distribution. The most commonly used directed graphical model are Bayesian Network (BNs), which decompose a joint probability distribution over multiple variables into a set of conditional and marginal distributions on low-dimensional subspaces. Thanks to their factorization of multivariate probability distributions, BNs are an efficient tool for reasoning under uncertainty, i.e. exploring the dependence structure of variables to facilitate reasoning in multidimensional domains under probabilistic settings (Campos 2006; Silander et al. 2008).

Werhli et al. performed a comparative study of correlation analysis, GGMs and BNs in the reconstruction of gene regulatory networks, using both laboratory data from cytometry experiments and synthetic data from gold-standard networks (Werhli et al. 2006). Their results showed that: first, GGMs and BNs outperformed correlation analysis on Gaussian observational data; second, there was no significant difference between GGMs and BNs on observational data in general; third, for interventional data, BNs outperformed GGMs and correlation analysis, especially when taking the edge directions rather than just the skeletons of the graphs into account. The last point in particular suggested that active interventions in the form of gene knockouts and over-expressions would be helpful to exploit the full potential of BNs.

1.5 Study objectives and outline of the thesis

The work in this thesis aims at exploring the use of PGMs in quantitative genetics and systems biology for plants, and further developing computational strategies for reverse-engineering of biological networks and genetic mapping. These objectives are achieved in the following four chapters, each of which is deliberately self-contained and can be read individually with no loss of understanding.

Chapter 2 proposes a novel method called the QTL+phenotype supervised orientation (QPSO) algorithm. As its name indicates, this algorithm is designed to infer directionality of edges in undirected phenotype networks utilizing phenotypic interactions in addition to QTL information. The QPSO algorithm outperforms other existing methods as it is applicable to cases where some phenotypes of interest come without QTLs. This makes it suitable for a much broader range of real-life studies, especially those involving expensive phenotypes.

In chapter 3, we present an integrative method for simultaneous modeling of multilevel phenotypic responses to DNA variations. More specifically, we combine three GGM approaches with the QPSO algorithm to model genotype-phenotype relationships with consideration for variations on intermediate molecular phenotypes, e.g. metabolites. The inferred dependency network which, though not essentially representing biological pathways, suggests how the effects of allele substitutions propagate through multilevel phenotypes. Such simultaneous study of the underlying genetic architecture and multifactorial interactions is expected to enhance the prediction and manipulation of complex traits.

Chapter 4 shows that PGMs have great potential in reliable reconstruction of genetic maps. We prove both theoretically and practically that PGMs can be used to construct genetic maps in the face of data perturbations caused by genotyping errors. Moreover, we demonstrate empirically that PGMs offer a promising solution to genetic map construction in the case of a reciprocal translocation.

Chapter 5 is dedicated to a comparative investigation of the two most common approaches to constructing PGMs, i.e. the PC algorithm and the Metropolis-Hastings algorithm. In view of the fact that BNs become an effective tool for causal network inference and most biological systems exist in the form of random network or scale-free network, here we compare the performance of the two algorithms in constructing both random and scale-free BNs. With this study, we aim to provide an informative guide to choosing the appropriate method depending on the application background and further selecting the proper related parameters.

Finally, chapter 6 presents a general discussion and a few concluding remarks.

References

Airoldi EM (2007) Getting started in probabilistic graphical models. PLoS Comput Biol 3:e252

Broman KW (2010) Genetic map construction with R/qtl. Technical Report# 214. University of Wisconsin-Madison, Department of Biostatistics & Medical Informatics

Campos LMd (2006) A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. Journal of Machine Learning Research 7:2149-2187

Farré A, Benito IL, Cistué L, De Jong J, Romagosa I, Jansen J (2011) Linkage map construction involving a reciprocal translocation. Theoretical and applied genetics 122:1029-1037

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799-805

Göring HH, Terwilliger JD (2000) Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. The American Journal of Human Genetics 66:1107-1118

Hackett C, Broadfoot L (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90:33-38

Hanisch D, Zien A, Zimmer R, Lengauer T (2002) Co-clustering of biological networks and gene expression data. Bioinformatics 18:S145-S154

Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM (2009) Defining gene and QTL networks. Current opinion in plant biology 12:241-246

Jiang P, Singh M (2010) SPICi: a fast clustering algorithm for large biological networks. Bioinformatics 26:1105-1111

Lehner B (2007) Modelling genotype–phenotype relationships and human disease with genetic interaction networks. Journal of Experimental Biology 210:1559-1566

Li W, Wang M, Irigoyen P, Gregersen PK (2006) Inferring causal relationships among intermediate phenotypes and biomarkers: a case study of rheumatoid arthritis. Bioinformatics 22:1503-1507

Logsdon BA, Mezey J (2010) Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. PLoS Comput Biol 6:e1001014

Monneveux P, Jing R, Misra SC (2012) Phenotyping for drought adaptation in wheat using physiological traits.

Muff S, Rao F, Caflisch A (2005) Local modularity measure for network clusterizations. Physical Review E 72:056107

Neto EC, Ferrara CT, Attie AD, Yandell BS (2008) Inferring causal phenotype networks from segregating populations. Genetics 179:1089-1100

Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC systems biology 1:37

Orgogozo V, Morizot B, Martin A (2015) The differential view of genotype–phenotype relationships. Frontiers in genetics 6:179

Ríos RO (2015) Plant breeding in the Omics Era. Springer

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551-1555

Rosa GJ, Valente BD, de los Campos G, Wu X-L, Gianola D, Silva MA (2011) Inferring causal phenotype networks using structural equation models. Genetics Selection Evolution 43:6

Searle S (1961) Phenotypic, genetic and environmental correlations. Biometrics 17:474-480

Shields D, Collins A, Buetow K, Morton N (1991) Error filtration, interference, and the human linkage map. Proceedings of the National Academy of Sciences 88:6501-6505

Silander T, Roos T, Kontkanen P, Myllymäki P (2008) Factorized normalized maximum likelihood criterion for learning Bayesian network structures. Proceedings of the 4th European workshop on probabilistic graphical models (PGM-08). Citeseer, pp 257-272

Ting N-C, Jansen J, Nagappan J, Ishak Z, Chin C-W, Tan S-G, Cheah S-C, Singh R (2013) Identification of QTLs associated with callogenesis and embryogenesis in oil palm using genetic linkage maps improved with SSR markers. PLoS One 8:e53076

Van Ooijen JW, Jansen J (2013) Genetic mapping in experimental populations. Cambridge University Press

Waitt DE, Levin DA (1998) Genetic and phenotypic correlations in plants: a botanical test of Cheverud's conjecture. Heredity 80:310-319

Werhli AV, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22:2523-2531

This chapter was published as: *Wang H., van Eeuwijk F. (2014) A new method to infer causal phenotype networks using QTL and phenotypic information. PLoS One 9(8), e103997*

Chapter 2

A new method to infer causal phenotype networks using QTL and phenotypic information

Abstract

In the context of genetics and breeding research on multiple phenotypic traits, reconstructing the directional or causal structure between phenotypic traits is a prerequisite for quantifying the effects of genetic interventions on the traits. Current approaches mainly exploit the genetic effects at quantitative trait loci (QTLs) to learn about causal relationships among phenotypic traits. A requirement for using these approaches is that at least one unique QTL has been identified for each trait studied. However, in practice, especially for molecular phenotypes such as metabolites, this prerequisite is often not met due to limited sample sizes, high noise levels and small QTL effects. Here, we present a novel heuristic search algorithm called the QTL+phenotype supervised orientation (QPSO) algorithm to infer causal directions for edges in undirected phenotype networks. The two main advantages of this algorithm are: first, it does not require QTLs for each and every trait; second, it takes into account associated phenotypic interactions in addition to detected QTLs when orienting undirected edges between traits. We evaluate and compare the performance of QPSO with another state-of-the-art approach, the QTL-directed dependency graph (QDG) algorithm. Simulation results show that our method has broader applicability and leads to more accurate overall orientations. We also illustrate our method with a real-life example involving 24 metabolites and a few major QTLs measured on an association panel of 93 tomato cultivars. Matlab source code implementing the proposed algorithm is freely available upon request.

2.1 Introduction

In animal and plant breeding, selection of superior genotypes for further crossing is an important objective. To achieve this objective, identification of quantitative trait loci (QTLs) can be a first step in the development of a breeding strategy; alternatively nowadays, estimation of genomic breeding values can be considered to form another initial step. Whether a breeding strategy is based on QTLs or genomic breeding values, multi-trait approaches offer clear advantages over single-trait approaches (Calus and Veerkamp 2011; Jiang and Zeng 1995). In multi-trait models, correlations, or associations, between traits have a symmetrical nature and are not supposed to convey information about causal relationships. Nonetheless, causal inference in correlated traits has been attracting growing research interest since it allows predicting effects of external interventions, where the effects of QTLs on phenotypic traits can be interpreted to represent a specific class of interventions (Rosa et al. 2011; Valente et al. 2013).

Causal inference in correlated traits, or equivalently, the construction of directed phenotype networks was so far mainly based upon logic that involves underlying QTLs (Jansen et al. 2009). For the simplest system with two traits (T_1, T_2) and one QTL (Q), Schadt et al. (2005) and Li et al. (2010) presented different implementations of triad analysis to determine whether the three entities are interconnected in, what they called, causal ($Q \rightarrow T_1 \rightarrow T_2$), reactive ($Q \rightarrow T_2 \rightarrow T_1$) or independent ($T_1 \leftarrow Q \rightarrow T_2$) manner. Further research efforts concerned the investigation of multi-locus and multi-trait systems. Aten et al. (2008) developed a network edge orienting (NEO) method and software to 1) perform genetic marker selection for each trait and 2) infer pairwise relationships between traits, using local-structure edge orienting (LEO) scores. Specifically, the LEO scores were calculated according to the likelihoods of local structural equation models (SEMs), which integrated two traits and the markers selected for each of them. Li et al. (2006) introduced another systematic method to first infer genetic architecture of multiple traits and then iteratively assess and refine the path model by means of covariance-based SEM. Neto et al. (2008) proposed a QTL-directed dependency graph (QDG) approach that requires a priori estimation of QTLs for the traits and executes the following two steps: 1) learn an undirected network from phenotypic data; 2) infer causal direction for every edge in the undirected phenotype network by conditioning on detected QTLs. In the QDG algorithm, QTL mapping is treated independently from the construction of phenotype network. In contrast, a QTL-driven phenotype network (QTLnet) method was introduced to jointly infer a directed phenotype network and the associated genetic architecture for a set of correlated traits (Neto et al. 2010). An adaptive lasso (AL) based method was presented to infer a gene regulatory network from gene expression and expression quantitative trait loci (eQTLs) data (Logsdon and Mezey 2010). In their simulation studies, Logsdon and Mezey (2010) compared the performance of five algorithms, i.e. the PC algorithm (Spirtes et al. 2000), the NEO algorithm, the QDG algorithm, the QTLnet algorithm and the AL algorithm. The results indicated that in the setting of tens of traits and QTLs, the QDG and the AL algorithms exhibited comparable performance but consistently outperformed the other three methods. Logsdon and Mezey (2010) also considered a couple of other algorithms including the one proposed by Li et al. (2006), but they were deemed computationally expensive. Therefore, the QDG and the AL algorithms will be regarded as two state-of-the-art methods in this field.

In practice, it has become fashionable to map QTLs for phenotypes of interest via genome-wide scans, since genotyping has become cheaper and easier thanks to the advancement of genome sequencing technologies. Contrariwise, phenotyping, and especially metabolic profiling and sensory assessment, is still expensive and time-consuming (Gagneur et al. 2011). Thus, for phenotypic traits such as metabolites and sensory attributes, it is hard to obtain large sample sizes that provide sufficient power for detecting small to medium sized QTLs. And it is often the case that, given high-dimensional phenotypic and genetic data (i.e. large numbers of traits and QTLs vs. small numbers of samples), significant QTLs cannot be identified for each and every trait (Hill et al. 2013; Joosen et al. 2013). In such cases, both the QDG and AL algorithms become inapplicable as they require at least one unique QTL for each trait studied (Logsdon and Mezey 2010; Neto et al. 2008).

To construct directed phenotype networks, especially when some traits come without QTLs, we present in this paper a QTL+phenotype supervised orientation (QPSO) algorithm. Compared with the benchmark QDG algorithm, our proposed method is likewise based on a priori determination of an undirected phenotype network and QTLs for the traits, where we recommend estimation of initial QTLs using multi-trait QTL mapping methods (Alimi et al. 2013; Jiang and Zeng 1995; Malosetti et al. 2008). Our QPSO algorithm implements a heuristic search different from that of the QDG algorithm and investigates a more comprehensive local structure at each step. More specifically, the QPSO algorithm takes into account the related phenotypic interactions in addition to QTLs when orienting an undirected edge between two traits. As a result, it can orient multiple undirected edges simultaneously. The performance of the QPSO and the QDG algorithms is compared through a series of simulations. The results show that our method has broader applicability and produces more accurate overall orientations. To demonstrate the QPSO algorithm empirically, we use it in combination with the PC-skeleton (Spirtes et al. 2000) to build a partially directed network that sheds light on causal relationships between 24 metabolites in ripe fruits of a tomato association panel.

2.2 Method

2.2.1 Causal inference in two correlated traits

Assume Y_1 and Y_2 are two correlated traits connected by an undirected edge in a phenotype network. The causal direction of $Y_1 - Y_2$ should follow one of two scenarios: $Y_1 \rightarrow Y_2$ or $Y_1 \leftarrow Y_2$. The two causal models are considered likelihood equivalent because $p(Y_1)p(Y_2|Y_1) = p(Y_1,Y_2) = p(Y_2)p(Y_1|Y_2)$. Thus, it is impossible to distinguish between $Y_1 \rightarrow Y_2$ and $Y_1 \leftarrow Y_2$, i.e. to orient $Y_1 - Y_2$, using a maximum-likelihood criterion.

Neto et al. (2008) presented a smart way to solve the problem of causal inference in two correlated traits. They introduced QTLs to Y_1 and Y_2 so as to get two expanded directed graphs as shown in **Figure 2.1**. The two expanded directed graphs are not likelihood equivalent since $p(\boldsymbol{Q}_1)p(Y_1|\boldsymbol{Q}_1)p(\boldsymbol{Q}_2)p(Y_2|Y_1,\boldsymbol{Q}_2) \neq p(\boldsymbol{Q}_2)p(Y_2|\boldsymbol{Q}_2)p(\boldsymbol{Q}_1)p(Y_1|Y_2,\boldsymbol{Q}_1)$, which can be further simplified as $p(Y_1|\boldsymbol{Q}_1)p(Y_2|Y_1,\boldsymbol{Q}_2) \neq p(Y_2|\boldsymbol{Q}_2)p(Y_1|Y_2,\boldsymbol{Q}_1)$. In this context, it is feasible to infer the causal direction of $Y_1 - Y_2$ according to the maximum-likelihood criterion. More specifically, $Y_1 - Y_2$ should be oriented in favor of the direction present in the model with higher likelihood, i.e. $Y_1 \rightarrow Y_2$ if $p(Y_1|\boldsymbol{Q}_1)p(Y_2|Y_1,\boldsymbol{Q}_2) > p(Y_2|\boldsymbol{Q}_2)p(Y_1|Y_2,\boldsymbol{Q}_1)$ while $Y_1 \leftarrow Y_2$ if $p(Y_1|\boldsymbol{Q}_1)p(Y_2|Y_1,\boldsymbol{Q}_2) < p(Y_2|\boldsymbol{Q}_2)p(Y_1|Y_2,\boldsymbol{Q}_1)$.

2.2.2 Causal inference in local generalized phenotype networks

In the context of **Figure 2.1**, $Y_1 - Y_2$ is oriented by introducing parent nodes to Y_1 and Y_2, where the parent nodes are restricted to earlier identified QTLs. However, it is known that many molecular traits, such as metabolites and proteins, do interact with one another. This means that in addition to QTLs, some other traits may also have causal effects on Y_1 and Y_2. Therefore, these traits should also be included in the parent nodes of Y_1 and Y_2; or, at least, their potential effects on Y_1 and Y_2 should be taken into account when one is attempting to orient $Y_1 - Y_2$. To make a comprehensive consideration of the local structure regarding Y_1 and Y_2, we present here the concept of local generalized phenotype network (LGPN) (**Figure 2.2A**), in which we include 1) QTLs identified for Y_1 and Y_2, 2) traits that have been determined as parent nodes of Y_1 and Y_2, 3) traits that are directly connected to Y_1 and Y_2 by undirected edges (these traits are hereinafter referred to as neighbouring traits of Y_1 and Y_2).

It has been demonstrated that the maximum-likelihood criterion can be employed to infer the direction of $Y_1 - Y_2$ in the context of **Figure 2.1**. Inspired by this, we find a feasible solution to the problem of causal inference in LGPNs that meet the following two conditions: 1) both Y_1 and Y_2 have parents nodes and at least one of Y_1 and Y_2 has unique parent nodes; 2) each neighboring trait of Y_1 is nonadjacent to at least one of the parent nodes of Y_1, and the same is true of Y_2. Assume in such a LGPN there are n undirected edges including $Y_1 - Y_2$. As every undirected edge has two optional directions (i.e. either forward or backward), the total number of candidate directed graphs derived from that LGPN is then 2^n. Verma and Pearl (Verma and Pearl 1990) have proved a theorem for the characterization of equivalent graphical models:

Theorem: *Two directed acyclic graphs (DAGs) are likelihood equivalent if and only if they have the same skeletons and the same v-structures (A v-structure in a DAG G is an ordered triple of nodes (X, Y, Z) such that G contains the directed edges X→Y and Z→Y, and X and Z are not adjacent in G).*

According to the theorem, we find that under the two aforementioned conditions, each of the 2^n candidate directed graphs possesses a distinct set of v-structures (for detailed explanation please refer to **Supplementary material**) and thus returns a distinct log-likelihood score $\sum_{i=1}^{N}\log_{10}\left(f(y_{1i}|\boldsymbol{pa}(y_{1i}))f(y_{2i}|\boldsymbol{pa}(y_{2i}))\right)$, where N is the sample size, $\boldsymbol{pa}(X)$ represents the parent nodes of trait X, and $f()$ is a conditional probability density function with parameters replaced by the corresponding maximum-likelihood estimates. Accordingly, the locally optimal directed graph (LODG) among the 2^n candidates should be the one with the highest log-likelihood score.

All undirected edges involving in a LGPN can be oriented simultaneously in the light of the corresponding LODG. These newly determined directed edges will then be employed to infer directions of some remaining undirected edges in the entire phenotype network. This leads to a heuristic search process, which will be described in detail in the following section. In the process of heuristic search, it might happen that some of the traits have never been assigned parent nodes in all of the previous steps. In cases where only Y_1 or Y_2, say Y_1, has been determined with parent nodes, the maximum likelihood criterion is able to identify the LODG for a reduced LGPN (**Figure 2.2B**), and the log-likelihood score should be reformulated as $\sum_{i=1}^{N}\log_{10}f(y_{1i}|\boldsymbol{pa}(y_{1i}))$. In particular cases where neither Y_1 nor Y_2 has unique parent nodes, the maximum likelihood criterion fails to infer direction of $Y_1—Y_2$. This means that the consideration of LGPN regarding Y_1 and Y_2 becomes a bit pointless and should be skipped.

In this study, we restrict ourselves to quantitative phenotypic traits and categorical QTL data, i.e., QTLs are represented by closest markers that can take one of two or three genotypes at that locus, depending on the type of population. Missing values in phenotypic and marker data are assumed to be estimated or imputed before that causal inference is applied. We also assume that a LGPN is a conditional linear Gaussian (CLG) model, in which discrete variables are not allowed to have continuous parents, and the joint distribution of continuous variables for every instantiation of discrete variables is multivariate Gaussian (Shenoy 2006).

2.2.3 Causal inference in an entire undirected phenotype network

A LODG may introduce new parent nodes to some of the traits. As illustrated in **Figure 2.3**, Y_1 is the newly determined parent node of C_1 and C_4. This updated causal information might subsequently enable or improve the orientation of the remaining undirected edges connecting to C_1 and C_4. Therefore, iterative implementation of causal inference in sequential LGPNs can finally orient as many edges as possible in an undirected phenotype network. This is, however, a typical heuristic search technique that has to be rerun from different starting points a number of times to avoid getting stuck in local optima. To this end, we exploit the Bayesian information criterion (BIC) score as a global evaluation metric to find the most likely fully or partially directed phenotype network obtained in multiple runs. The BIC score is a well-known penalized likelihood criterion that is often used to prevent overfitting the training data. It is formally defined as $LL(\boldsymbol{D}|G) - 0.5 * \log(N)\,|G|$, where G is the learnt network, $\boldsymbol{D}$ is the training data, $LL(\boldsymbol{D}|G)$ is the maximum log-likelihood, N is the sample size, and $|G|$ denotes the dimension of G (Schwarz 1978).

In summary, our QPSO algorithm executes the following steps to perform causal inference in an entire undirected phenotype network, where we assume that the QTLs have been identified earlier by a multi-trait QTL mapping method like the ones described by Malosetti et al. (2008) and Alimi et al. (2013).

(1) Randomly choose a pair of traits that simultaneously satisfy two conditions: first, they are connected by an undirected edge; second, both of them have parent nodes and at least one of them has unique parent nodes.
(2) Extract the LGPN (as illustrated in **Figure 2.2A**) with respect to these two traits.
(3) Identify the LODG from all candidate directed graphs derived from that LGPN; update the phenotype network (i.e. orient all the corresponding undirected edges) according to the LODG.
(4) Repeat steps (1), (2) and (3) until no more traits satisfying the two conditions mentioned in step (1) remain.
(5) If the resulting phenotype network is partially directed, randomly choose a pair of traits that simultaneously satisfies two conditions: first, the traits are connected by an undirected edge; second, only one of them has parent nodes.
(6) Extract the LGPN (as illustrated in **Figure 2.2B**) with respect to these two traits.
(7) Identify the LODG from all candidate directed graphs derived from that LGPN; update the phenotype network according to the LODG.
(8) Repeat steps (5) (6) and (7) until no more undirected edges can be oriented; store the overall orientation of the entire phenotype network.
(9) Repeat steps (1) through (8) a number of times (this number is hereinafter referred to as the number of iterations); use the BIC score to evaluate each overall orientation and return the one with the highest score.

An implementation of the QPSO algorithm has been realized in Matlab. Thereinto, the probability density function of the CLG distribution and the BIC score are computed by calling functions in Bayes Net Toolbox (https://code.google.com/p/bnt/). Matlab source code is available from the authors upon request.

2.3 Results

2.3.1 Synthetic phenotypic and QTL data

We followed the same protocol used by Neto et al. (2008) to generate synthetic data for a simulation study creating phenotypic and marker data for an F2 population. A directed network composed of 65 nodes and 74 edges (**Figure 2.4**) was created by the *randomDAG* function in the R package ‘*pcalg*’ (http://cran.r-project.org/web/packages/pcalg/index.html). In this network, 34 nodes denoted phenotypic traits while the other 31 nodes represented QTLs. QTLs were randomly selected among 50 markers, with 5 markers unevenly distributed on each of 10 chromosomes. Observations of a trait were generated on the basis of linear regression model $y = \boldsymbol{\alpha}^{\boldsymbol{T}}\boldsymbol{q} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{x} + \varepsilon$, where $\boldsymbol{q}$ is a vector of marker scores (QTLs), $\boldsymbol{x}$ is a vector of traits, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the regression coefficients corresponding to $\boldsymbol{q}$ and $\boldsymbol{x}$, and ε is the residual. To simplify exposition, we assumed quantitative traits and categorical QTL data, and allowed only additive genetic effects with an increment of 0.1 per allele. Specifically, QTL genotypes *aa*, *Aa* and *AA* were respectively encoded as 1, 2 and 3; the regression coefficient for genotype *aa* was uniformly drawn from [0.2, 0.4]; the coefficients for genotypes *Aa* and *AA* were then given by adding 0.1 and 0.2, respectively. Besides, the regression coefficient of a phenotype on one another was chosen uniformly from [0.5, 1], and standard deviation of ε was randomly drawn from [0.1, 0.5].

2.3.2 Simulation results

Our QPSO method is applicable to pre-learnt undirected or partially directed phenotype networks. There are a number of ways to learn undirected graphical models from data, including marginal and partial correlation analyses, as well as conditional independence tests. We consider the QDG algorithm still to represent a benchmark algorithm with which to compare our QPSO approach. The QDG algorithm uses an undirected phenotype network as reconstructed by the PC-skeleton algorithm as the starting configuration for edge orientation. For the comparative simulations, we also took the PC-skeleton as the method to arrive at an undirected phenotype network.

In a first set of 20 simulation runs, we evaluated the performance of the PC-skeleton algorithm using two indicators, recall and precision. Each simulation run was based on a distinct phenotypic dataset. Recall, also called true positive rate or sensitivity, measures the proportion of true edges that are retrieved in relation to the full set of true edges. Precision, or positive predictive value, measures the proportion of true (positive) edges in the set of identified edges (true and false positives). The higher recall and precision, the better the reconstruction of the network is. The results of our first set of simulations are shown in **Table 2.1**, where means and standard deviations for recall and precision are given. With increasing sample size, both recall and precision improved with respect to their means across simulation runs, while their standard deviations remained at a low level. In particular, given that in practice 100 individuals is a representative sample size for biological data like metabolites, a recall of 0.86 and a precision of 0.97 on average, is very encouraging. High mean value and low standard deviation indicate that the PC-skeleton algorithm can accurately and consistently recover an undirected network, using a reasonable sample size.

Given an undirected phenotype network pre-learnt by the PC-skeleton algorithm, our next step was to infer causal directions for edges in the network by exploiting associated QTLs. Both the QDG and QPSO algorithms are applicable to this problem when at least one QTL has been identified for each and every trait. In a second set of simulations we then made a comparative evaluation of the two edge orientation algorithms using the full set of QTL data and the earlier reconstructed undirected phenotype network. Results are presented in **Table 2.2**, where we give mean and standard deviation of the proportion of true positive edges that were correctly oriented for QDG using all QTLs and QPSO using all QTLs over 20 independent simulation runs. To achieve consistent results (i.e. small standard deviations) from multiple runs, the QDG algorithm claimed 1000 iterations (Neto et al. 2008) while our QPSO method required only about 10 iterations for each individual run. Two conclusions regarding the effectiveness of the two algorithms can be drawn from the comparative study. First, along with the increase of samples, the overall orientations obtained by both methods became increasingly accurate and consistent. Second, given the same sample size, the QPSO algorithm produced more accurate overall orientation than the QDG method, since the former always possessed a higher mean proportion of correctly oriented true edges combined with a comparable or slightly lower standard deviation.

The major advantage of the QPSO algorithm lies in the ability of inferring causal relationships between correlated traits when some or more of the traits do not have QTLs. To demonstrate this, in a third set of simulations, we blanked out a number of detected QTLs and then investigated the performance of the QPSO algorithm. We assumed that QTLs corresponding to the clear rectangular nodes in **Figure 2.4** were not available for the reconstruction of the directed phenotype network, i.e., these QTLs were removed from the input of the QPSO algorithm. Results of this particular simulation study are summarized in the columns of **Table 2.2** that are labelled QPSO using partial QTLs. It is obvious that QPSO still has reasonably good edge orientation even when a substantial proportion of the traits come without QTLs. **Table 2.2** learns that given the same sample size, the overall orientation obtained by the QPSO algorithm with partial QTLs is getting refined when sample size increases, and is slightly inferior to the one obtained by the same algorithm with full QTLs, but nevertheless superior to the one obtained by the QDG algorithm with full QTLs.

To demonstrate the robustness of the QPSO algorithm, we elaborated on the results of the third set of simulations reported above. We selected five edges from the simulated phenotype network that differed with respect to the configuration of parent nodes for two correlated traits: (1) between traits 2 and 18, with the two traits having one common QTL (C8m1) and trait 2 having a unique QTL (C3m5); (2) between traits 1 and 16, with each trait having a unique QTL (C2m1 for trait 1 and C4m4 for trait 16); (3) between traits 16 and 26, with trait 16 having a unique QTL (C4m4) and trait 26 having no QTL; (4) between traits 13 and 26, with trait 13 having a unique QTL (C6m1) and trait 26 having no QTL; (5) between traits 26 and 31, with neither trait having QTL. We investigated the accuracy of orientations obtained by the QPSO algorithm for the five edges (**Table 2.3**). When sample size increased from 100 to 500, the two edges 2─18 and 13─26 were almost 100% correctly oriented, the average percentages of correct orientations improved from 65 to 95% for the edge 1─16, from 25 to 70% for the edge 16─26 and from 35 to 100% for the edge 26─31. The declining performance of our method on edge 16─26 than 1─16 was mainly due to error propagation in orientations. If an incorrect direction has been assigned to edge

1─16 in a previous step, it will affect the accuracy of orientation regarding edge 16─26. Likewise, an incorrect direction inferred for edge 16─26 will subsequently harm the orientation of edge 26─31. However, results in **Table 2.3** indicate that our QPSO algorithm possesses higher accuracy in orientation of edge 26─31 than of 16─26. This is because the algorithm makes a full consideration on the neighborhood of trait 26 (i.e. the interactions between traits 13, 15, 16 and 26 were all taken into account) when orienting the edge 26─31, so that the negative impact of incorrect orientation of edge 16─26 can be counterbalanced, to some extent, by the positive effect of correct orientation of edge 13─26.

Each run of the QPSO algorithm selects the best model according to the maximum-likelihood criterion. Nonetheless, in many cases, several models may have very close likelihoods, meaning that they are all compatible with the data. Therefore, it is critical to check the consistency of those competing models. Also based on the third set of simulations, we compared the best two models obtained by a single run of the QPSO algorithm for different sample sizes. The results (**Table 2.4**) show that for a given sample size, the best two models indeed possess very close BIC scores; but, more importantly, they are substantially the same, except for a handful of edges that are assigned with opposite directions in the two models. In view of the high consistency that exists between the best two models, we believe it will suffice to return only the best model as final output. All simulations were implemented in a 32 bit Intel(R) Core(TM) i5-2410M 2.30GHz 4GB RAM machine. The computing time of a single run of the QPSO algorithm for each sample size studied is also included in **Table 2.4**.

As explained in the Method section, the QPSO method returns fully or partially directed phenotype networks depending on the number of available QTLs. The PC algorithm, which is a further extension of the PC-skeleton algorithm, also returns partially directed phenotype networks but without using QTLs. To demonstrate the advantage of our QPSO method over the PC algorithm, in a final set of simulations we assessed the performance of the PC algorithm in the reconstruction of the simulated phenotype network. Results are shown in the last two columns of **Table 2.2**. The last four columns of **Table 2.2** support the conclusion that for a given sample size and undirected phenotype network, the QPSO algorithm with partial QTLs orients correctly far more edges than the PC algorithm, which orients edges without using QTL information.

2.3.3 Metabolic and QTL data collected in ripe tomato fruits

Metabolic data were collected from ripe fruits of 93 tomato cultivars, an association panel provided by five breeding companies involved in the Centre for BioSystems Genomics tomato quality program (http://www.cbsg.nl/tomato.aspx). According to morphological characteristics of ripe tomato fruits, the 93 cultivars were categorized into three groups, labelled as beef, cherry and round. The three groups made up approximately 25%, 25% and 50% of the total collection. Metabolic profiling of cultivars was based on pooled fruit samples, where the sample for each beef or round cultivar mixed 12 fruits while the sample for each cherry cultivar contained 18 fruits. Sugars and acids were measured using the technique described by Roessner-Tunali et al. (2003). Volatiles were quantified using the method presented by Tikunov et al. (2005). In this study, we investigated a subset of 24 metabolites of special interest. The same set of metabolic data was studied by Ursem et al. (2008), where a detailed description of the measurements and the data can be found. Most of the metabolites strongly discriminated between cherry and non-cherry (i.e. beef and round) tomatoes, as was found by both principal component analysis and discriminant analysis (Ursem et al. 2008). Application of the PC-skeleton algorithm to reconstruct a phenotypic network between the 24 metabolites led to a network with 17 edges (**Figure 2.5**). The reconstruction was done choosing a rather strict test level of 0.01 for the conditional independence tests to arrive at a sparse but high confidence phenotypic network.

To find a list of QTLs driving the variation in the 24 metabolites, association analysis was performed using 600 SNPs in a multi-trait mixed model association mapping procedure that allowed for trait specific effects of pleiotropic QTLs. In addition, this mixed model contained intercept terms for the cherry and non-cherry groups to correct for this obvious type of population structure. To investigate the susceptibility of the QPSO algorithm to the amount of QTL information for orienting edges between metabolites, we selected QTLs at three levels significance. The more liberal the threshold, the greater the number of selected QTLs is. We adopted three closely together thresholds for the significance of the test for a QTL with an effect on any of the 24 metabolites at a given marker locus, corresponding to $-\log_{10}$(P-value) = 4.5, 5.0, 5.5. At the strictest level of $-\log_{10}$(P-value) > 5.5, 11 QTLs were identified for seven metabolites (**Figure 2.5A**), with two QTLs that had pleiotropic effect on two metabolites. Of the 24 metabolites, 17 remained without QTL. Lowering the $-\log_{10}$(P-value) for QTL detection to 5.0 led to four additional QTLs and more QTLs with pleiotropic effects: eight metabolites came with one or more QTLs, 16 stayed without QTLs (**Figure 2.5B**). At a $-\log_{10}$(P-value) threshold of 4.5, a total of 19 QTLs were detected for 10 metabolites (**Figure 2.5C**).

2.3.4 Causal relationships among tomato metabolites

The QPSO algorithm was used to orient undirected edges between the metabolites. The results corresponding to QTLs selected at the three thresholds of $-\log_{10}$(P-value) = 4.5, 5.0 and 5.5, are shown in **Figure 2.5A**, **B** and **C**, respectively. Comparison of the three graphs indicates that when more QTLs with relatively small effects enter the model, more traits tend to be associated with at least one QTL, and accordingly more undirected edges between traits can be oriented. The 11 QTLs for the seven metabolites in **Figure 2.5A** allowed 11 of the 17 edges to be oriented. For the 15 QTLs and 8 metabolites in **Figure 2.5B** and the 19 QTLs and 10 metabolites in **Figure 2.5C**, 13 edges out of the 17 could be oriented.

Among the 17 undirected edges between metabolites, 11 were oriented throughout the three graphs. We examined the consistency of the inferred directions of the 11 edges and found that only the edge connecting 1-penten-3-one and trans-2-

hexenal came varied in direction across the test levels for QTLs. The directions of the other 10 edges were invariant to the changes in the amount of QTL information. This invariance of edge orientation provides a modest demonstration of the robustness of the QPSO algorithm.

After reconstruction of the directed network, an investigation of pleiotropic QTLs is possible in a post hoc analysis of the network. For example, in **Figure 2.5C**, initially the two QTLs rs4494 and rs4715 were pleiotropic for 3-methylbutanol and 2-methylbutanol. Simultaneously, 3-methylbutanol was identified to be a direct upstream metabolite of 2-methylbutanol. Did the two QTLs have pleiotropic effects on both traits, or, were their effects on 2-methylbutanol mediated via 3-methylbutanol? To answer this question, we used the BIC scoring metric to evaluate and compare the two models shown in **Figure 2.6A** and **B**, where Q denotes rs4494 or rs4715, Y_1 and Y_2 represent respectively 3-methylbutanol and 2-methylbutanol. It turned out that with respect to either of the two QTLs, the simplified model in **Figure 2.6B** possessed a higher BIC score, thereby providing a better fit to the observed data. Thus, we deleted from **Figure 2.5C** the two edges pointing from rs4494 and rs4715 to 2-methylbutanol. The same concern can be raised with respect to the QTL effect of NSG1 on methyl salicylate and 2-methoxyphenol. In this case, we failed to infer the causal relationship between the two metabolites due to lack of unique QTL. To this type of specific problems, Neto et al. (2008) suggested a possible solution by comparing the likelihoods of the three models shown in **Figure 2.6B**, **C** and **D**. Here, we exploited the BIC score again and let Q, Y_1 and Y_2 denote NSG1, methyl salicylate and 2-methylbutanol, respectively. Comparative results indicated that the data best supported the pleiotropic model in **Figure 2.6D**, therefore the local structure of NSG1, methyl salicylate and 2-methylbutanol in **Figure 2.5C** should remain the same. The investigation to the reality of observed pleiotropic relations as described for **Figure 2.5C** was equally applied to **Figure 2.5A** and **B**.

Given the structure of the network, we estimated effects of traits on one another and of QTLs on traits. To that end, we regressed metabolites on QTLs and adjacent upstream metabolites. We discriminated between positive and negative associations among the metabolites according to the signs of fitted regression coefficients. The signs of QTL effects were not considered as they are somewhat arbitrary in the context of association mapping and binary markers such as SNPs.

The above directed network can be compared with undirected networks constructed on the basis of marginal and partial correlations, like a correlation network and a graphical Gaussian model (GGM), see Figure 5 and 9 as presented by Ursem et al. (2008). Both these graphs look very dense despite the fact that only strongly significant correlations were displayed (q<0.05, as a false discovery rate procedure was chosen). From a dense graph with many variables incorporated, it is hard to arrive at meaningful interpretations. Compared with the results reported by Ursem et al. (2008), our findings obtained by the PC-skeleton algorithm in combination with the QPSO algorithm comprised a much sparser graph, with the additional advantages of showing (partial) directedness between traits and the influence of QTLs on traits. It should be remarked that between the three graphs, a central backbone coincided.

Although we reconstructed a directed network on a set of metabolites, the resulting network cannot be interpreted as an approximation to a metabolic network, a major reason being the absence of time course data. The metabolic data we analysed represented mean metabolite abundances obtained from grinding a number of fruits for a set of tomato genotypes. To get insight in biological pathways, we should measure series of chemical reactions occurring over relative short time frames within a cell, but the measurement and analysis of such time series still presents large challenges (Blair et al. 2012). The value of a directed network like that of **Figure 2.5** is that it allows to correctly quantifying the effects of QTL allele substitutions, say genetic interventions or perturbations, at a number of phenotypic traits simultaneously. For instance, changes at locus rs7213 will have an effect on the concentration of 1-penten-3-one, which will subsequently affect the concentration of cis-3-hexenal. In contrast, variations in the concentration of 1-penten-3-one will not influence the level of trans-2-hexenal, as trans-2-hexenal is an upstream metabolite of 1-penten-3-one. Another representative example is that if we attempt to control the concentration of 2-methylbutanol, we should be cautious about the allelic composition at loci rs4715, rs8396, rs8340, rs7143 and rs8233, since any genetic perturbation leading to an alteration in the concentration of 3-methylbutanol will then change the concentration of 2-methylbutanol.

From a biological point of view, **Figure 2.5A**, **B** and **C** present several interesting clusters. It is noteworthy that the major carbohydrates glucose and fructose are linked to sucrose and citric acid via myo-inositol. Whilst myo-inositol is synthesized from glucose, the recovery of the indirect link is remarkable, also considering that myo-inositol is linked to sucrose which can be broken down into glucose and fructose or alternatively into UDP- glucose and fructose. Another remarkable link is the one between beta-damascenone and beta-ionone both of which are break-down products of carotenoids (Baldermann et al. 2010). Interestingly 6-methyl-5-hepten-2-one was not linked to these, despite being a carotenoid class volatile. This indicates that the latter open chained form likely stems from lycopene (Gao et al. 2008), potentially explaining why it is not linked to any of the former two metabolites. Furthermore, the negative correlation between aspartic acid and glutamic acid might be explained by the action of aspartate aminotransferase converting glutamate oxaloacetate to 2-oxoglutarate and aspartate. It is clear that the C5 and C6 volatiles were grouped together. Whilst intriguing that these are likely produced from the same precursors via lipoxygenases (Rambla et al. 2014), one would speculate that the C5 and C6 volatiles should probably be disconnected, making the 1-penten-3-one (C5) mini hub linked to many C6 volatiles worth further investigation. Incidentally, both C5 and C6 volatiles were also found in different clusters previously (Mathieu et al. 2009). Regarding the metabolites 3- and 2-methylbutanol, they both are likely leucine/isoleucine derived compounds and they were found linked to 2-isobutylthiazole before (Mathieu et al. 2009).

In summary, our partially directed network for the 24 tomato metabolites is clearly more concise and informative than those of conventional marginal and partial correlation analyses and allowed discriminating between direct and indirect metabolic responses to particular genetic perturbations in tomatoes. Following Valente et al. (2013), it is exactly the type of information that is needed for predicting the effects of genetic interventions on sets of correlated phenotypic traits.

2.4 Discussion

The QPSO algorithm is applied to pre-learnt undirected or partially directed phenotype networks. Correlation networks and GGMs are the most common models used to learn undirected graphs from biological data (Krumsiek et al. 2011; Ma et al. 2007; Ursem et al. 2008). Bayesian networks (BNs) are considered a promising tool to recover partially directed biological networks (Gavai et al. 2009; Hodges et al. 2010; Iyer et al. 2013). Formally, a BN is a DAG that represents probabilistic conditional independence structures for a set of interacting variables. Two mainstream approaches regarding BN structure learning are the constraint-based and the score-based methods. However, due to their inherent limitations, in many cases the two approaches can only return partially directed graphs rather than DAGs. Please refer to Mahdi and Mezey (2013) and Chickering (1996) for details. A comparative evaluation of correlation networks, GGMs and BNs has been made in the reconstruction of gene regulation networks (Werhli et al. 2006). The results indicated that GGMs performed comparably to BNs on general observations, and both GGMs and BNs outperformed correlation networks on Gaussian observations.

Besides the construction of undirected or partially directed phenotype networks, QTL mapping for the traits is also a prerequisite for using the QPSO algorithm. Standard QTL mapping methods, including association mapping and linkage mapping, process phenotypic traits in a parallel fashion without paying attention to the underlying dependence structure of traits. Neto et al. (2010) claimed that QTL mapping conditional on the phenotype network should lead to a better estimated genetic architecture, and a better genetic architecture should in turn result in a better inferred phenotype network. Accordingly, they developed a statistical framework, named QTLnet, to jointly infer a causal phenotype network and the associated genetic architecture for a set of correlated phenotypes. The QTLnet method is actually a Metropolis–Hastings algorithm that integrates QTL mapping and the sampling of directed phenotype networks at each step. However, like many other Markov Chain Monte Carlo approaches, this method shows slow mixing of the resulting Markov chains and requires considerable computation time. Its implementation in R can handle no more than 20 traits at this point (Neto et al. 2010).

The QPSO algorithm treats QTL mapping independent from phenotype network reconstruction and cannot correct misspecified edges in undirected phenotype networks pre-learnt by the PC algorithm. In this sense, it would be considered less robust than the QTLnet method. We observed, however, that the QPSO algorithm performed well in the reconstruction of directed phenotype networks: 1) the results of our first set of simulations and also the ones shown by Neto et al. (2008) implied that given relatively sufficient samples (say, ≥100 for a network composed of 34 phenotypes and 27 edges, or, ≥300 for a network composed of 100 phenotypes and 107 edges), the undirected phenotype networks recovered by the PC-skeleton algorithm were fairly reliable (with recall>0.85 and precision>0.90); 2) the simulation results presented by Logsdon and Mezey (2010) indicated that in small-scale phenotype networks (to which the QTLnet method is only applicable), the QTLnet method was outperformed by the QDG algorithm that was used as benchmark in this study; 3) the results of our second and third sets of simulations showed that compared with the benchmark QDG algorithm, our proposed method was applicable to more general cases and led to more accurate overall orientations. In summary, we have confidence that the QPSO algorithm is of great potential in practical applications.

In simulation experiments, the QPSO algorithm was applied to a random network consisting of dozens of nodes and edges. Theoretically, this method has no limit to the scale of either random networks or scale-free networks, since it always decomposes a whole network into a finite number of LGPNs and makes causal inferences in the LGPNs using a heuristic search strategy. Scale-free networks show power-law degree distributions that are very different from the Poisson degree distributions of random networks. More specifically, in scale-free networks, most nodes have relatively few links while only a few nodes (called hubs) have a large number of links; contrariwise, in random networks nodes are more evenly connected. Here we would like to point out that node degree distribution is believed to have some effect on the efficiency of the QPSO algorithm, but the extent of this impact is hard to evaluate. Recall from the Method section that the LODG is selected from 2^n candidates, where n would be a big number if either or both of Y_1 and Y_2 are hubs. A large n means that, on the one hand, an enormous computational effort has to be made when scanning for the LODG; on the other hand, the number of LGPNs decomposed from the whole network is significantly reduced as a great number of undirected edges are assigned to the same LGPN. These two effects will counterbalance each other to some extent; but on the whole, the overall efficiency of the algorithm will vary a lot depending on the specific circumstances, including sample size, the number of nodes, and the node degree distribution. In addition, computer memory and processor speed are practical factors that can also affect the scalability and efficiency of the algorithm.

As explained previously, the QPSO algorithm returns fully or partially directed phenotype networks depending on the number of available QTLs. Its exhaustive search for LODGs is based on the distinction between non-equivalent DAGs, each of which has a unique set of v-structures. Thus, there is no directed cycle in a LODG. However, the QPSO algorithm is overall a heuristic method. It takes a random walk from one LODG to another. The integration of all LODGs does not necessarily lead to a complete DAG. That is, in some cases, it is possible that certain edges in two or more LODGs form a directed cycle. Please note that the benchmark QDG algorithm has substantially the same property.

We developed our methodology in the first place for data from plant breeding experiments, in which advanced experimental designs are common that include local control of error variation at multiple levels and in multiple directions. As genotypes for population types like doubled haploids and recombinant inbred lines are replicated in such experiments, reconstruction of networks take place at genotypic means obtained from mixed model analyses of one or more experiments. These genotypic means will have small standard errors and that will contribute to the stability of reconstructed directed networks. For metabolic assessments, usually pooled samples of fruits stemming from multiple replicates in an experiment are used. Pooling is another way of reducing measurement error. Therefore, it will beneficial to bring phenotypic traits to the aggregation level of genotypic means before trying to reconstruct a phenotype network. The QPSO algorithm is applicable to a complete data matrix of genotypes (samples) by traits. Pre-processing of phenotypic data by converting them to genotypic means by mixed model analyses provides a straightforward and accurate way of imputing missing phenotypic values.

In conclusion, we have presented a novel heuristic search algorithm, named QPSO, to infer causal relationships between correlated traits. This algorithm allows some traits to come without QTLs, and it takes into account associated phenotypic interactions in addition to QTLs when orienting undirected edges between traits. Thanks to these two properties, the QPSO algorithm has much broader applicability and produces more accurate overall orientations, compared to the benchmark QDG algorithm.

References

Alimi NA, Bink MCAM, Dieleman JA, Magan JJ, Wubs AM, Palloix A, van Eeuwijk FA (2013) Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theoretical and Applied Genetics 126:2597-2625

Aten JE, Fuller TF, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: The NEO software. Bmc Systems Biology 2

Baldermann S, Kato M, Kurosawa M, Kurobayashi Y, Fujita A, Fleischmann P, Watanabe N (2010) Functional characterization of a carotenoid cleavage dioxygenase 1 and its relation to the carotenoid accumulation and volatile emission during the floral development of Osmanthus fragrans Lour. J Exp Bot 61:2967-2977

Blair RH, Kliebenstein DJ, Churchill GA (2012) What Can Causal Networks Tell Us about Metabolic Pathways? Plos Computational Biology 8

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution 43

Chickering DM (1996) Learning equivalence classes of Bayesian network structures. Uncertainty in Artificial Intelligence:150-157

Gagneur J, Elze MC, Tresch A (2011) Selective Phenotyping, Entropy Reduction, and the Mastermind game. Bmc Bioinformatics 12

Gao HY, Zhu HL, Shao Y, Chen AJ, Lu CW, Zhu BZ, Luo YB (2008) Lycopene accumulation affects the biosynthesis of some carotenoid-related volatiles independent of ethylene in tomato. J Integr Plant Biol 50:991-996

Gavai AK, Tikunov Y, Ursem R, Bovy A, van Eeuwijk F, Nijveen H, Lucas PJF, Leunissen JAM (2009) Constraint-based probabilistic learning of metabolic pathways from tomato volatiles. Metabolomics 5:419-428

Hill CB, Taylor JD, Edwards J, Mather D, Bacic A, Langridge P, Roessner U (2013) Whole-Genome Mapping of Agronomic and Metabolic Traits to Identify Novel Quantitative Trait Loci in Bread Wheat Grown in a Water-Limited Environment. Plant Physiology 162:1266-1281

Hodges AP, Dai DJ, Xiang ZS, Woolf P, Xi CW, He YQ (2010) Bayesian Network Expansion Identifies New ROS and Biofilm Regulators. Plos One 5

Iyer SP, Shafran I, Grayson D, Gates K, Nigg GT, Fair DA (2013) Inferring functional connectivity in MRI using Bayesian network structure learning with a modified PC algorithm. NeuroImage 75:165-175

Jansen RC, Tesson BM, Fu JY, Yang YJ, McIntyre LM (2009) Defining gene and QTL networks. Current Opinion in Plant Biology 12:241-246

Jiang CJ, Zeng ZB (1995) Multiple-Trait Analysis of Genetic-Mapping for Quantitative Trait Loci. Genetics 140:1111-1127

Joosen RVL, Arends D, Li Y, Willems LAJ, Keurentjes JJB, Ligterink W, Jansen RC, Hilhorst HWM (2013) Identifying Genotype-by-Environment Interactions in the Metabolism of Germinating Arabidopsis Seeds Using Generalized Genetical Genomics. Plant Physiology 162:553-566

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. Bmc Systems Biology 5

Li RH, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA (2006) Structural model analysis of multiple quantitative traits. Plos Genetics 2:1046-1057

Li Y, Tesson BM, Churchill GA, Jansen RC (2010) Critical reasoning on causal inference in genome-wide linkage and association studies. Trends in Genetics 26:493-498

Logsdon BA, Mezey J (2010) Gene Expression Network Reconstruction by Convex Feature Selection when Incorporating Genetic Perturbations. Plos Computational Biology 6

Ma SS, Gong QQ, Bohnert HJ (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Research 17:1614-1625

Mahdi R, Mezey J (2013) Sub-Local Constraint-Based Learning of Bayesian Networks Using A Joint Dependence Criterion. Journal of Machine Learning Research 14:1563-1603

Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk FA (2008) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). Euphytica 161:241-257

Mathieu S, Cin VD, Fei ZJ, Li H, Bliss P, Taylor MG, Klee HJ, Tieman DM (2009) Flavour compounds in tomato fruits: identification of loci and potential pathways affecting volatile composition. J Exp Bot 60:325-337

Neto EC, Ferrara CT, Attie AD, Yandell BS (2008) Inferring causal phenotype networks from segregating populations. Genetics 179:1089-1100

Neto EC, Keller MP, Attie AD, Yandell BS (2010) Causal Graphical Models in Systems Genetics: A Unified Framework for Joint Inference of Causal Network and Genetic Architecture for Correlated Phenotypes. Annals of Applied Statistics 4:320-339

Rambla JL, Tikunov YM, Monforte1 AJ, Bovy AG, Granell A (2014) The expanded tomato fruit volatile landscape. J Exp Bot

Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiology 133:84-99

Rosa GJM, Valente BD, de los Campos G, Wu XL, Gianola D, Silva MA (2011) Inferring causal phenotype networks using structural equation models. Genetics Selection Evolution 43

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang CS, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang LM, Castle J, Zhu HY, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics 37:710-717

Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6:461-464

Shenoy PP (2006) Inference in hybrid Bayesian networks using mixtures of Gaussians. the 22nd conference on uncertainty in artificial intelligence, pp 428-436

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge, Mass.

Tikunov Y, Lommen A, de Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiology 139:1125-1137

Ursem R, Tikunov Y, Bovy A, van Berloo R, van Eeuwijk F (2008) A correlation network approach to metabolic data analysis for tomato fruits. Euphytica 161:181-193

Valente BD, Rosa GJM, Gianola D, Wu XL, Weigel K (2013) Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? Genetics 194:561-572

Verma T, Pearl J (1990) Equivalence and synthesis of causal models. the Sixth Conference on Uncertainty in Artificial Intelligence, pp 220-227

Werhli AV, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22:2523-2531

Figure 2.1 Candidate solutions to causal inference in two correlated traits. Y_1 and Y_2 are two traits correlated with each other; $\boldsymbol{Q}_1=\{Q_{11},\ldots,Q_{1k}\}$ and $\boldsymbol{Q}_2=\{Q_{21},\ldots,Q_{2l}\}$ denote QTLs for Y_1 and Y_2, respectively.

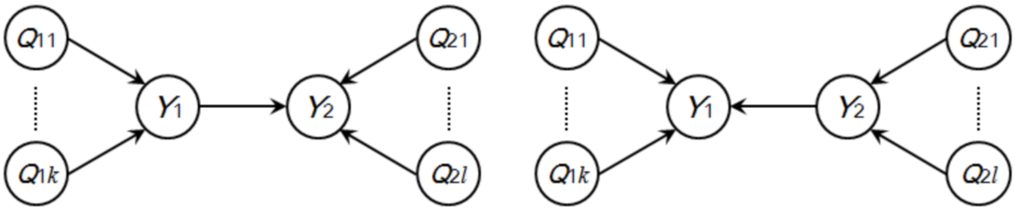


Figure 2.2 The general representations of resolvable LGPNs. Y_1 and Y_2 are two correlated traits; $\boldsymbol{P}_1=\{P_{11},\ldots,P_{1k}\}$ and $\boldsymbol{P}_2=\{P_{21},\ldots,P_{2l}\}$ are, respectively, the unique parent nodes of Y_1 and Y_2; $\boldsymbol{P}_{12}=\{P_1,\ldots,P_s\}$ are the common parent nodes of Y_1 and Y_2; $\boldsymbol{C}_1=\{C_{11},\ldots,C_{1u}\}$ and $\boldsymbol{C}_2=\{C_{21},\ldots,C_{2v}\}$ are the unique neighboring traits of Y_1 and Y_2; $\boldsymbol{C}_{12}=\{C_1,\ldots,C_t\}$ are the common neighboring traits of Y_1 and Y_2. Note that each of the neighboring traits of Y_1 is nonadjacent to at least one of the parent nodes of Y_1, and the same is true of Y_2. Also note that $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and $\boldsymbol{P}_{12}$ are allowed to have three different compositions: (1) a pure set of QTLs, if only genetic factors have been identified for Y_1 and/or Y_2; (2) a mixed set of QTLs and traits, if some traits in addition to QTLs have been determined to have causal effects on Y_1 and/or Y_2; (3) a pure set of traits, if only some traits have been found as causal factors of Y_1 and/or Y_2; in contrast, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$ and $\boldsymbol{C}_{12}$ only refer to those traits that are directly connected to Y_1 and/or Y_2 by an undirected edge. (A) The general representation of LGPNs where both Y_1 and Y_2 have parent nodes, and at least one of them has unique parent nodes; (B) the general representation of LGPNs where only Y_1 has parent nodes.

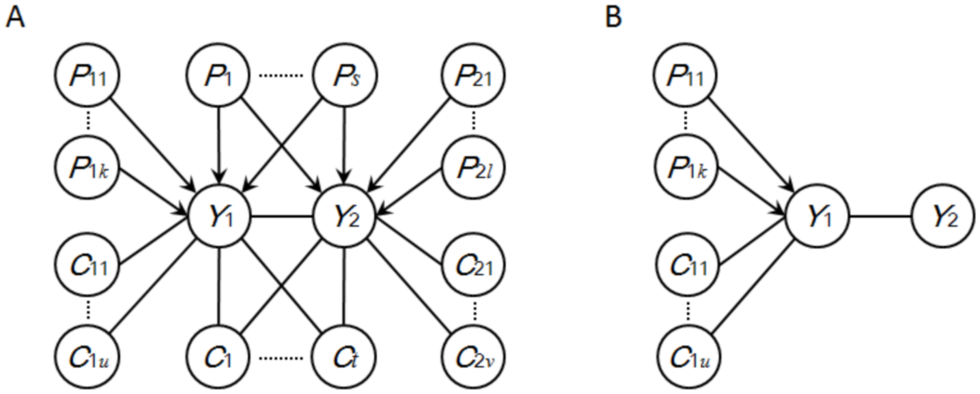


Figure 2.3 An example of LODG. Y_1 and Y_2 are two correlated traits; C_2 and C_3 are two traits that have been newly determined as parent nodes of Y_1; Y_1, C_3 and C_5 are three traits newly determined as parent nodes of Y_2; Y_1 is a newly determined parent node of traits C_1 and C_4; Y_2 is a newly determined parent node of traits C_4 and C_6.

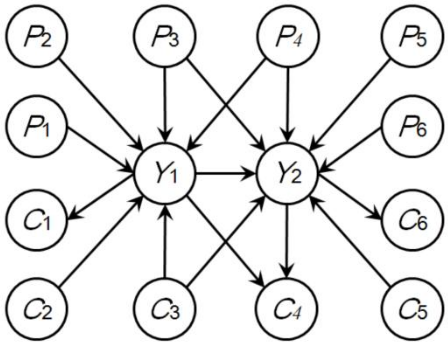


Figure 2.4 A synthetic QTL-phenotype network. This network consists of 31 QTLs, 34 traits and 74 directed edges. Traits are ordered by numerical numbers and QTLs are labelled in the form of ‘C*i*m*j*’ indicating the *j*-th marker on the *i*-th chromosome. Because only a part of QTLs were used in a third set of simulations, the nodes are further classified as follows: shaded rectangular nodes – QTLs present in the third set of simulations; clear rectangular nodes – QTLs absent in the third set of simulations; shaded circular nodes – traits provided with QTLs in the third set of simulations; clear circular nodes – traits provided without QTLs in the third set of simulations.

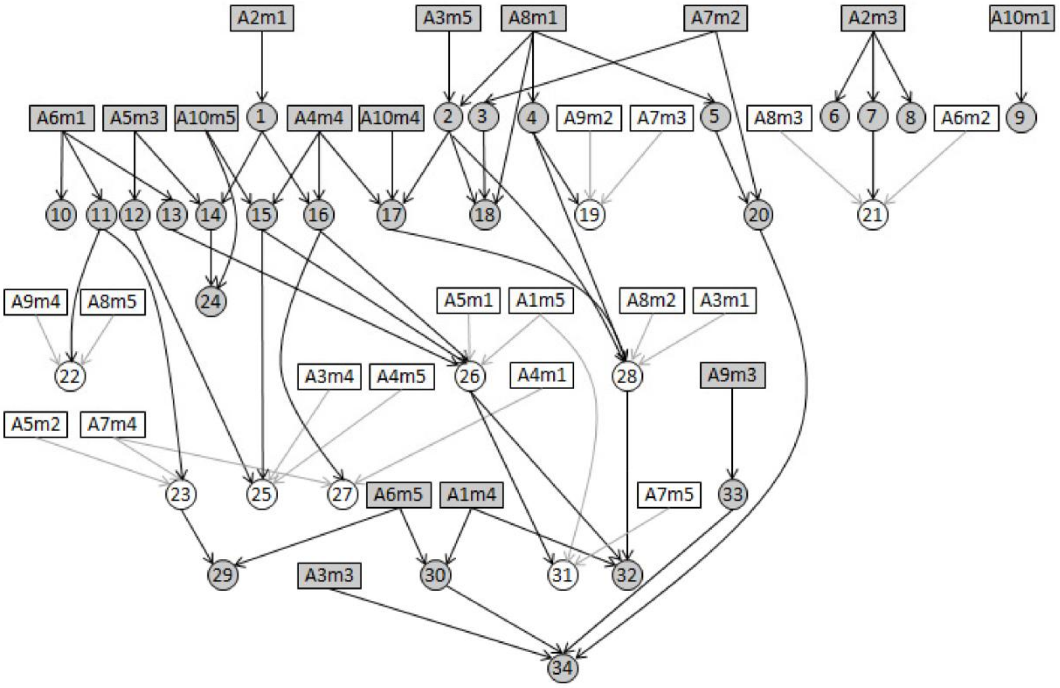
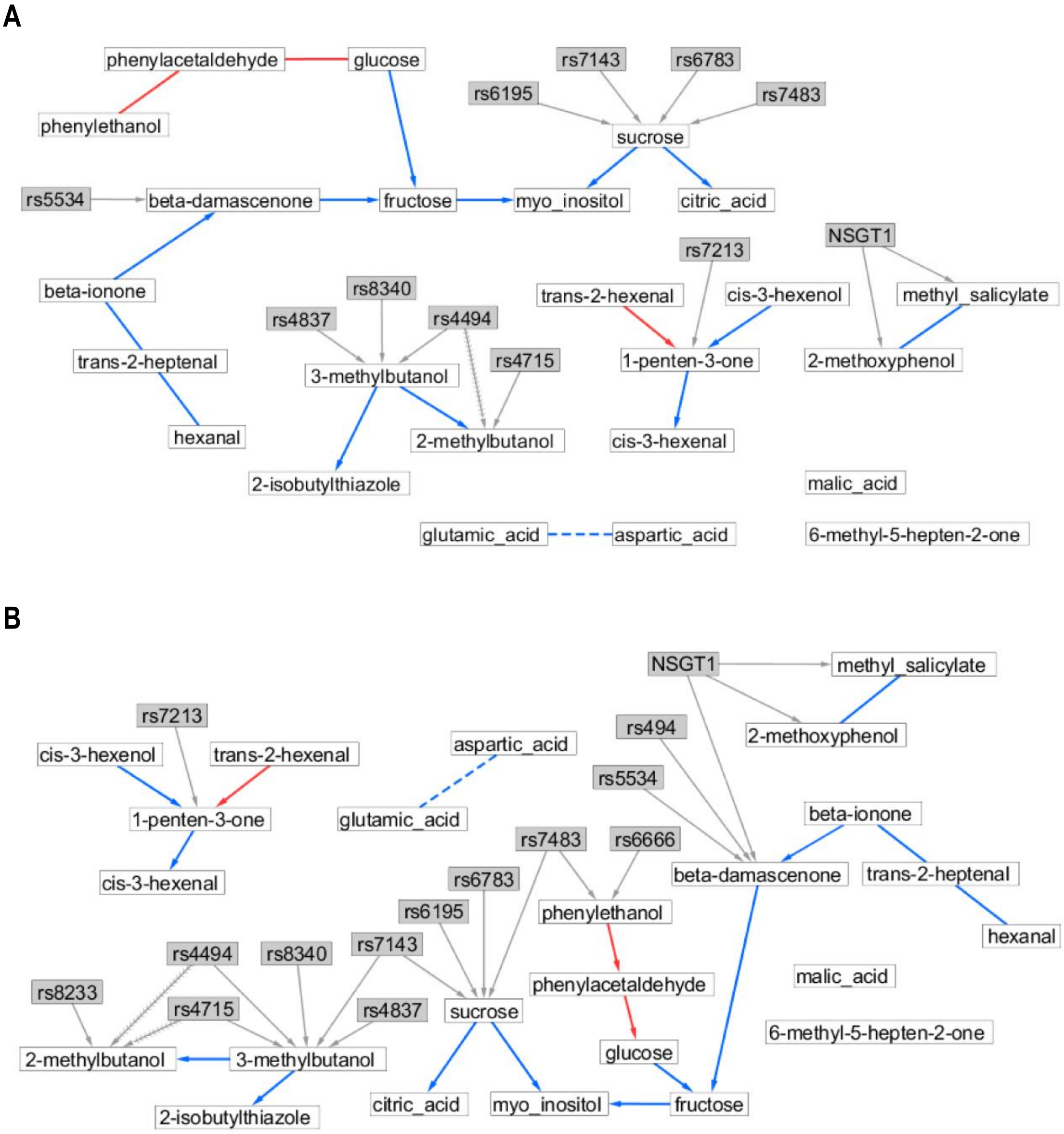


Figure 2.5 Three partially directed graphs describing the relationships among 24 metabolites in ripe tomato fruits. Clear nodes represent metabolites; shaded nodes denote QTLs identified for the metabolites. QTLs in (A), (B) and (C) were selected on the basis of $-\log_{10}$(p-value) thresholds 5.5, 5.0 and 4.5, respectively. Grey edges link QTLs to the corresponding metabolites. Blue and red edges, without regard to their directions, were learnt by the PC-skeleton algorithm; their directions, if any, were inferred by the QPSO algorithm. Blue edges occur consistently throughout the three graphs representing different test levels for QTLs, while red edges do not. Solid and dashed edges indicate positive and negative correlations, respectively; fishbone edges are removed by post hoc causal reasoning.



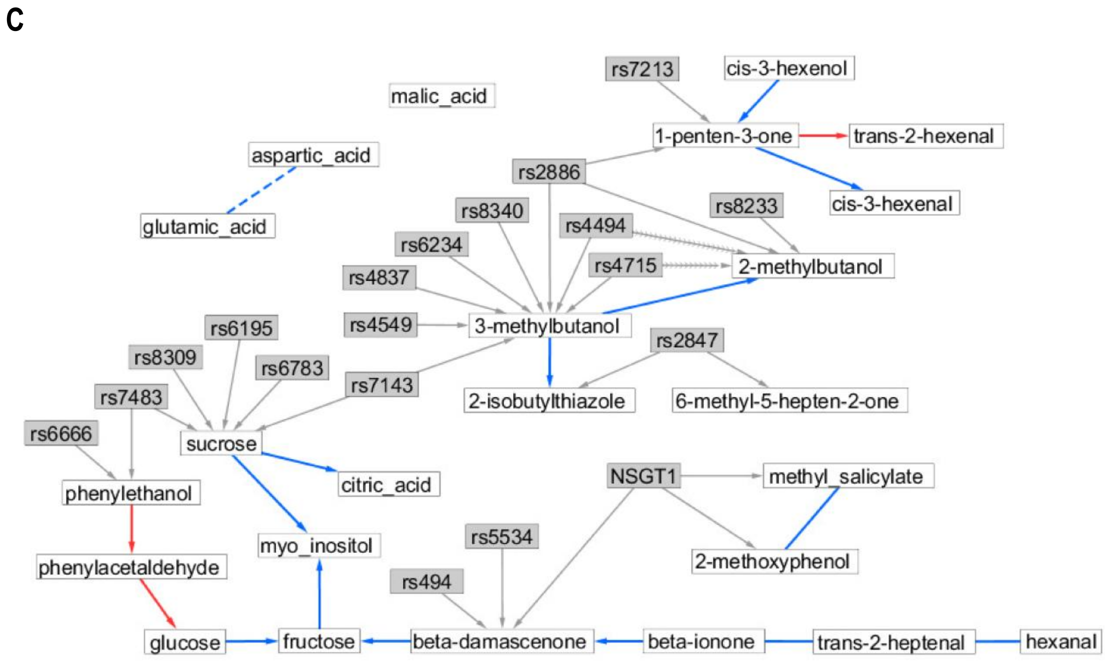


Figure 2.6 Test models in triad analysis. (A) a QTL Q has pleiotropic effects on two traits Y_1 and Y_2, Y_1 is also a causal factor of Y_2; (B) Q is identified for Y_1, Y_1 has a causal effect on Y_2; (C) Q is identified for Y_2, Y_2 has a causal effect on Y_1; (D) Q is identified for both Y_1 and Y_2, but the causal relationship between Y_1 and Y_2 is unclear.

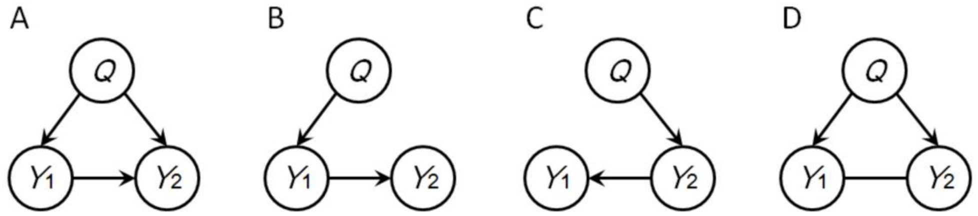


Table 2.1 Performance of the PC-skeleton algorithm in reconstructing the synthetic phenotype network across a series of 20 simulations.

Sample size	Recall		Precision	
	mean	sd	mean	sd
100	0.86	0.06	0.97	0.03
200	0.94	0.03	0.97	0.03
300	0.96	0.03	0.98	0.03
400	0.98	0.03	0.98	0.03
500	0.99	0.03	0.98	0.02

The significance level of conditional independent tests used in the PC-skeleton algorithm was set at 0.01.

Table 2.2 Comparative evaluation of three algorithms in overall orientation of the synthetic phenotype network. Sample size, means and standard deviations of the proportion of true positive edges that were correctly oriented across a series of 20 simulations.

Sample size	QDG (using full QTLs)		QPSO (using full QTLs)		QPSO (using partial QTLs)		PC (no use of QTLs)	
	mean	sd	mean	sd	mean	sd	mean	sd
100	0.72	0.11	0.78	0.07	0.77	0.09	0.49	0.11
200	0.74	0.08	0.81	0.07	0.80	0.08	0.57	0.06
300	0.75	0.06	0.81	0.06	0.81	0.07	0.60	0.05
400	0.77	0.05	0.82	0.06	0.82	0.06	0.60	0.05
500	0.78	0.05	0.84	0.05	0.83	0.06	0.60	0.05

Table 2.3 Demonstration of the robustness of the QPSO algorithm. Proportion of correct edge orientation across 20 simulations for edges with varying parent configurations. Node numbers refer to Figure 2.4.

Sample	Proportions of correct orientations of five edges				
size	2→18	1→16	16→26	13→26	26→31
100	1.00	0.75	0.45	1.00	0.45
200	1.00	0.75	0.50	1.00	0.80
300	1.00	0.80	0.60	1.00	0.95
400	1.00	0.85	0.65	1.00	1.00
500	1.00	0.95	0.70	1.00	1.00

Decimal numbers were the average values deduced from 20 independent runs in the third set of simulations.

Table 2.4 Comparison between the best two models obtained by a single run of the QPSO algorithm

Sample	The best model			The second-best model			Computing
size	BIC score	different edges		BIC score	different edges		time (h)
100	-6.8990e+03	14→24 (√)		-6.9057e+03	14←24 (×)		0.89
200	-1.3308e+04	1→14 (√)	1←16 (×)	-1.3325e+04	1←14 (×)	1→16 (√)	1.27
		16→27 (√)	14←24 (×)		16←27 (×)	14→24 (√)	
300	-1.9663e+04	11→23 (√)	1←16 (×)	-1.9693e+04	11←23 (×)	1→16 (√)	2.18
		16→26 (√)	11←22 (×)		16←26 (×)	11→22 (√)	
			14←24 (×)			14→24 (√)	
400	-2.5944e+04	11→23 (√)	11←22 (×)	-2.5988e+04	11←23 (×)	11→22 (√)	2.76
		15→25 (√)	15←26 (×)		15←25 (×)	15→26 (√)	
500	-3.2333e+04	1→14 (√)	4←28 (×)	-3.2459e+04	1←14 (×)	4→28 (√)	3.30
		4→19 (√)	11←22 (×)		4←19 (×)	11→22 (√)	
		11→23 (√)			11←23 (×)		
		17→28 (√)			28←17 (×)		

Computing time was measured on a 32 bit Intel(R) Core(TM) i5-2410M CPU 2.30GHz machine with 4GB RAM. “different edges” were assigned with opposite directions in the best two models. (√) means the direction of that edge was inferred correctly, whereas (×) applies to the opposite case.

Supplementary material

Figure 2A shows a local generalized phenotype network (LGPN), in which 1) both traits Y_1 and Y_2 have parent nodes and at least one of Y_1 and Y_2 has unique parent nodes; 2) each neighboring trait of Y_1 is nonadjacent to at least one of the parent nodes of Y_1, and the same is true of Y_1.

Theorem [Verma and Pearl, 1990]*: Two directed acyclic graphs (DAGs) are likelihood equivalent if and only if they have the same skeletons and the same v-structures (A v-structure in a DAG G is an ordered triple of nodes (X, Y, Z) such that G contains the directed edges X→Y and Z→Y, and X and Z are not adjacent in G).*

According to this theorem, we can deduce that given the directed edge pointing from P_{11} to Y_1, the two candidate directions of the undirected edge between C_{11} and Y_1 (i.e., $C_{11}\rightarrow Y_1$ and $Y_1\rightarrow C_{11}$) will form two nonequivalent structures: $P_{11}\rightarrow Y_1\leftarrow C_{11}$ and $P_{11}\rightarrow Y_1\rightarrow C_{11}$. The reason is quite straightforward: since C_{11} is nonadjacent to P_{11}, $P_{11}\rightarrow Y_1\leftarrow C_{11}$ is a v-structure whereas $P_{11}\rightarrow Y_1\rightarrow C_{11}$ is not. The same is true if P_{11} is replaced by any other node $\in\{P_{11},\ldots,P_{1\mathrm{k}}\}\cup\{P_1,\ldots,P_s\}$, and, C_{11} is replaced by any other node $\in\{C_{11},\ldots,C_{1u}\}\cup\{C_1,\ldots,C_t\}$.

Similarly, we can deduce that given the directed edge pointing from P_{21} to Y_2, $P_{21}\rightarrow Y_2\leftarrow C_{21}$ and $P_{21}\rightarrow Y_2\rightarrow C_{21}$ are nonequivalent. And the same is true if P_{21} is replaced by any other node $\in\{P_{21},\ldots,P_{2l}\}\cup\{P_1,\ldots,P_s\}$, and, C_{21} is replaced by any other node $\in\{C_{21},\ldots,C_{2v}\}\cup\{C_1,\ldots,C_t\}$.

Lastly, let's consider the orientation of the undirected edge between Y_1 and Y_2. We restrict ourselves to the cases where at least one of Y_1 and Y_2 has unique parent nodes (please note the unique parent nodes are not limited to QTLs, that is, traits that have been previously determined as parent nodes of Y_1 and Y_2 are also taken into account). In **Figure 2**A, P_{11} is a unique parent node of Y_1, which indicates that P_{11} is nonadjacent to Y_2. Therefore, we know that $P_{11}\rightarrow Y_1\leftarrow Y_2$ and $P_{11}\rightarrow Y_1\rightarrow Y_2$ are nonequivalent as the former forms a v-structure while the latter does not. The same is true if P_{11} is replaced by any other node $\in\{P_{11},\ldots,P_{1\mathrm{k}}\}$.

Similarly, we have $P_{21}\rightarrow Y_2\leftarrow Y_1$ and $P_{21}\rightarrow Y_2\rightarrow Y_1$ are nonequivalent and the same is true if P_{21} is replaced by any other node $\in\{P_{21},\ldots,P_{2l}\}$.

In conclusion, if a LGPN satisfies the aforementioned two conditions and there are a total of n undirected edges involved in it, we know that each of the 2^n candidate directed graphs possesses a distinct set of v-structures and is therefore not equivalent to the others.

This chapter was published as: *Wang H., Paulo J., Kruijer W., Boer M., Jansen H., Tikunov Y., Usadel B., van Heusden S., Bovy A., van Eeuwijk F. (2015) Genotype–phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and QTLs in ripe tomatoes. Mol Biosyst 11:3101–3110*

Chapter 3

Genotype-phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and QTLs in ripe tomatoes

Abstract

Modeling genotype-phenotype relationships is a central objective in plant genetics and breeding. In standard QTL mapping, as applied to plant breeding populations, variations in phenotypic traits are modeled in relation to variations at the genomic level, regardless of intermediate levels of biological variations. Here we present an integrative method for simultaneous modeling of multilevel phenotypic responses to DNA variations. Specifically, for ripe tomato fruits, the dependencies of 24 sensory traits on 29 metabolites and the dependencies of all the sensory and metabolic traits further on 21 QTLs were investigated by graphical modeling and causal inference techniques. The inferred dependency network which, though not essentially representing biological pathways, suggests how the effects of allele substitutions propagate through multilevel phenotypes. Such simultaneous study of the underlying genetic architecture and multifactorial interactions is expected to enhance the prediction and manipulation of complex traits.

3.1 Introduction

Elucidating the genetic architecture of complex traits is a key objective in plant genetics. Existing methods mainly identify genomic regions associated with phenotypic variation through single- or multi-trait quantitative trait locus (QTL) analysis. However, between DNA and final phenotype, there exist multilevel intermediate substances such as proteins and metabolites, which possess a quantitative nature and vary among individuals within populations. Successfully linking variations at intermediate levels to DNA variations on the one hand and to phenotypic variations on the other hand should enhance the prediction and manipulation of complex traits.

Associations between and within multilevel omics data can be jointly learnt by probabilistic graphical models (PGMs), which typically unravel probabilistic conditional independence structures of multiple variables. A particular type of PGMs, namely Gaussian graphical models (GGMs, also known as “covariance selection” or “concentration graph” models) (Drton and Perlman 2004), has become popular in computational systems biology. GGMs are superior to the well-known correlation networks (also called “relevance networks”), as they are based on partial correlations and thereby distinguish between direct and indirect associations (Krumsiek et al. 2011).

The metabolome is of great importance in crop plants, as metabolite concentrations reflect the developmental stage of plants and determine to a great extent many quality traits such as nutritional value and sensory attributes. Recent advances in plant metabolite profiling, including gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) and nuclear magnetic resonance (NMR), have enabled large-scale analyses that reveal quantitative variation in the metabolic content of various species (Carreno-Quintero et al. 2013). Accordingly, it has become feasible to investigate associations between metabolites.

Beyond associations, dependencies among metabolites are of interest to plant biologists for understanding adaptation and survival in relation to primary and secondary metabolism. The metabolome is recognized as a highly interactive system, where a metabolite variation may lead to a chain reaction: changes in the concentration of a metabolite alter the concentrations of some other metabolites through specific regulatory pathways. A few methods have been presented to uncover dependencies among associated traits, using previously determined QTLs (Aten et al. 2008; Cai et al. 2013; Li et al. 2006; Logsdon and Mezey 2010; Neto et al. 2008; Neto et al. 2010). All these approaches require at least one unique QTL for each trait studied. In practice, however, this prerequisite is often not satisfied. To cope with more general scenarios where some of the traits come without QTL or unique QTL, a QTL + phenotype supervised orientation (QPSO) algorithm has recently been proposed (Wang and van Eeuwijk 2014). This algorithm looks promising in learning dependencies between metabolites, whose profiling is still expensive and time-consuming, with small sample sizes limiting the power of QTL mapping.

In this paper, we combine three GGM approaches with the QPSO algorithm to model genotype-phenotype relationships with consideration for the intermediate metabolite variations. Our integrative method is demonstrated through a practical case study, in which we obtain a dependency network involving 24 sensory traits, 29 metabolites and 21 QTLs identified for those sensory traits and metabolites in ripe tomato fruits. In the first place, a high-confidence true positive undirected network, which represents direct associations within and between metabolites and sensory traits, is learnt by the three GGM approaches including: (i) lasso-based neighborhood selection (Meinshausen and Buhlmann 2006) (LBNS) in combination with a stability approach to regularization selection (Liu et al. 2010) (StARS), (ii) the PC-skeleton algorithm (Spirtes et al. 2000) and (iii) the Lasso (Tibshirani 1996) in combination with stability selection (Meinshausen and Buhlmann 2010) (SS). In the second place, given the undirected network and QTLs previously identified for the sensory traits and metabolites, edge directions (i.e., the directions of associations) are inferred by the QPSO algorithm. In the third place, each sensory trait and metabolite is regressed on its QTLs and inferred parent nodes (i.e., nodes with outgoing edges pointing to this sensory trait or metabolite). The fitted regression coefficients provide more details regarding the estimated dependencies: “+” – positive, “-” – negative, and their absolutes values – the strength of dependencies.

It is known that tomato sensory traits are co-determined by metabolites (Abegaz et al. 2004; Carli et al. 2009; Tandon et al. 2003). A major concern of plant breeders and physiologists is, thus, how to identify metabolites and the underlying genomic regions responsible for certain sensory traits of interest, and thereby come up with targeted strategies for simultaneous improvement of those traits. Our proposed method provides a way to investigate the dependencies within and between metabolites and sensory traits. The estimated dependencies which, though not equal to biological pathways, suggest how the effects of allele substitutions propagate through metabolites to sensory traits. This information should help breeders and physiologists to predict and manipulate the target traits.

3.2 Materials

3.2.1 Tomato populations and phenotypic data

The data were collected on ripe fruits of four F_2 segregating populations developed in the tomato program of a consortium that was called the Centre for BioSystems Genomics (CBSG; http://www.cbsg.nl/tomato.aspx). Four contrasting tomato cultivars were selected as parental lines, namely C074 (cherry fruit type), C085 (cherry fruit type), R075 (round fruit type) and R104 (round fruit type). Crosses between the parental lines were made following a half-diallel mating design. The F_1 plants were selfed and the subsequent F_2 generation included four cherry×round populations: C074×R075, C074×R104, C085×R075 and C085×R104. For each cherry×round population, plants of 48 offspring genotypes were grown.

On all plants, 29 metabolites and 24 sensory traits were scored on ripe fruits, which were harvested and prepared as described in Tikunov et al. (2005). Metabolic profiling was carried out in two ways: volatiles were measured using a head space Solid Phase Microextraction – Gas Chromatography – Mass Spectrometry (SPME-GC-MS) (Tikunov et al. 2005); sugars and acids were quantified using the method of GC-MS of trimethylsylil ester derivatives (Roessner-Tunali et al. 2003). All metabolites were identified at level 1 annotation (Sumner et al. 2007) using authentic chemical standards analyzed at identical experimental conditions, except beta-damascenone, which has a level 2 identity: NIST mass spectral library 2010 (Mainlib) match 911 (0-1000) and the library retention index deviation of 4 (http://www.nist.gov/srd/nist1a.cfm). All metabolites have corresponding CAS ID numbers. Sensory profiles were obtained by a trained panel of judges for just 16 out of the 48 genotypes for each cherry×round population. The judges evaluated each genotype for a set of sensory traits including smell, taste, aftertaste, and mouthfeel experience. All sensory attributes were scored on a scale of 0 to 100. In addition to the metabolites and sensory traits, brix was measured for each genotype using a refractometer (GMK-701R; Nie-Co Products, Aalsmeer, NL). Metabolite abundances were transformed to log10 scale for statistical analysis. Prior to network reconstruction, genotypic means for the sensory traits, brix and metabolites were calculated using mixed models, which contained corrections for measurement time (for brix and metabolites), judge (for the sensory traits), population (for all traits) and the presence/absence of the Rin mutation (for all traits). Rin is the recessive ripening-inhibitor mutation that inhibits ripening (Vrebalov et al. 2002), and was present in all crosses involving parent R075. Fruits from plants that are homozygous for Rin do not ripen and have lower concentrations of metabolites. The corrected genotypic means were used for further analysis.

3.2.2 Genotypic data and QTL analysis

A set of 6000 SNP markers was available from the Infinium BeadArray. A selection of the markers was used to produce a high quality genetic linkage map. The obtained linkage map contained 600 SNP markers, 50 markers per chromosome, evenly spread at about 2cM.

A multi-trait QTL mapping strategy was implemented following the idea described in Malosetti et al. (2008) and Alimi et al. (2013). This strategy assumes that a single biparental offspring population was present. We turned the four cherry × round F2 populations into a single biparental F2 population by interpreting the two cherry parents to represent a first single parent and the two round tomato parents to represent a second single parent. Phenotypes were then regressed on genetic predictors, i.e. independent variables expressing molecular marker information. Genetic predictors were based on the expected number of alleles coming from the round parents, i.e. conditional QTL probabilities given flanking marker information using a Hidden Markov model (Jiang and Zeng 1997). Parametrization was such that positive regression coefficients, QTL allele substitution effects, would point to the round allele as increasing the level of the trait, whereas negative QTL effects would imply that the cherry allele increased the trait. In comparison to Malosetti et al. (2008) and Alimi et al. (2013), for the current multi-trait QTL model we took care to allow for population specific intercepts for each trait. Another deviation from Malosetti et al. (2008) and Alimi et al. (2013) was that we included a trait specific correction for the presence/absence of the Rin mutation. Our multi-trait QTL model for a vector of trait responses was therefore as follows: traits = population specific trait intercepts + trait specific RIN corrections + trait specific QTLs + trait specific residuals. The trait specific residuals were modeled with trait specific variances and correlations. Multi-trait QTL models were fitted on each of three groups of traits: 1) volatiles; 2) sugars and acids; 3) sensory attributes. The multi-trait QTL modeling was done in GenStat 16 (http://www.vsni.co.uk/software/genstat/). Positions of QTLs identified for the traits studied are summarized in **Table 3.1**.

3.3 Methods

3.3.1 Outline approach to dependency network reconstruction

Figure 3.1 illustrates our integrative method for learning dependency network from the sensory, metabolic and QTL data. First, two GGM approaches, (i) LBNS + StARS and (ii) the PC-skeleton algorithm, were used to obtain the consensus of direct associations among metabolites (**Figure 3.S1B** vs. **3.S3B**) and that among sensory traits (**Figure 3.S2B** vs. **3.S3D**). Second, the Lasso + SS was implemented in addition to the above two approaches to get the consensus of dependencies of sensory traits on metabolites (**Figure 3.S4A-C**). Please note that here brix was taken into account, since it is a major intermediate between metabolites and sensory traits. Specifically, brix was treated as a response of metabolites and a predictor for sensory traits in the Lasso + SS. The reason for taking multiple ways to network reconstruction is because the common findings of various methods are considered to be true positive with high-confidence. Third, given (i) the dependencies obtained in the second step and (ii) QTLs previously identified for the metabolites and sensory traits, the directions of associations were inferred by the QPSO algorithm. Last, each metabolite and sensory trait was regressed on its QTLs and estimated parent nodes, respectively. It is worth noting that parent nodes of a metabolite should only be metabolites, while parent nodes of a sensory trait could consist of metabolites and sensory traits. Signs of the fitted coefficients discriminated between positive and negative dependencies. This is particularly helpful to decipher whether cherry or round allele contributed to the alteration of a trait. As in **Figure 3.2-4**, positive QTL effects (solid red lines) mean that the round allele increased the level of a trait whereas the cherry allele led to a decrease; conversely, negative QTL effects (dashed red lines) mean that the cherry allele produced an increase while the round allele a decrease. The absolute values of the fitted coefficient implied the strength of dependencies, which were depicted by the edge thickness in **Figure 3.2-4**.

3.3.2 Gaussian Graphical Models (GGMs)

GGMs are a class of undirected graphs that present only direct associations among multivariate Gaussian random variables. Under the assumption that all involved variables have a multivariate Gaussian distribution, two variables are said to be conditionally independent, i.e. not directly associated, if and only if their partial correlation is zero. Partial correlation measures the degree of correlation between two variables after removing the effects of other variables. It is known that zero entries in the inverse covariance matrix, also known as concentration matrix or precision matrix, correspond to zero partial correlations. In sum, under multivariate normality, non-zero entries of the concentration matrix imply direct associations between pairs of variables, and thereby define the presence of edges in GGM.

3.3.3 Lasso-Based Neighborhood Selection (LBNS) + Stability Approach to Regularization Selection (StARS)

For high-dimensional data with more variables than samples, the concentration matrix cannot be directly estimated from the sample covariance matrix as the latter is non-invertible (singular). In such a case, estimating a sparse concentration matrix is a prerequisite to constructing GGM. To this end, Meinshausen and Buhlmann (2006) proposed the LBNS scheme. This scheme first fits a lasso model (Tibshirani 1996) to each variable separately, using all other variables as predictors. It then sets an entry in the concentration matrix, say p_{ij}, to be non-zero if the estimated coefficient of variable i on j and/or the estimated coefficient of variable j on i is non-zero.

A major challenge when applying lasso-based approaches to graphical modeling is to specify the regularization parameter that controls the sparsity of the resulting graph: larger amounts yield sparser graphs whereas smaller amounts lead to denser graphs. To come up with a general solution that is especially suited to high-dimensional problems, Liu et al. (2010) proposed StARS: a stability approach to regularization selection. StARS implements subsampling (Politis et al. 1999) to draw a finite number of subsamples (overlapping subsamples are allowed) and constructs a GGM for each subsample. StARS starts with a large regularization and gradually reduces it until the resulting graphs are simultaneously sparse and replicable across all subsamples. An implementation of LBNS in combination with StARS is available in the R package ‘huge’, which involves a variability threshold with two alternatives 0.1 and 0.05 (Zhao et al. 2012). Application of the two thresholds to both metabolic and sensory data suggested that 0.1 would be a better choice in this study (see section 5.2 for details).

3.3.4 The PC-skeleton algorithm

The PC algorithm, named after its inventors Peter Spirtes and Clark Glymour, consists of two steps: first, learn an undirected graph from observational data through a series of conditional independence tests; second, orient as many edges as possible according to the estimated conditional independencies and the acyclic constraint. Here we only used the first step, which is referred to as the PC-skeleton algorithm. It starts with a complete graph and removes redundant edges one by one if pairs of corresponding variables are found conditionally independent. For proper implementation of conditional independence tests on different types of data, the PC-skeleton algorithm uses Fisher's z-transformation of the partial correlation for quantitative data and the G^2 statistic for categorical data (Colombo et al. 2014). In this study, the significance level of conditional independence tests was set at 0.05. The reason for this will be given in detail in section 5.2.

3.3.5 The Lasso + SS

Though GGMs can effectively reveal direct associations among substances of the same nature, they perform poorly in the identification of associations between substances of different nature. This is mainly because substances of different nature are usually obtained by different measuring techniques and thus have medium to low absolute correlations. This phenomenon was also observed in the present study for associations between metabolic and sensory traits. For this reason, we performed Lasso regression (Tibshirani 1996) of each sensory trait on metabolites as a supplement to the implementation of LBNS + StARS and the PC-skeleton algorithm. The proper amount of regularization in the Lasso was chosen by SS. More specifically, the Lasso was applied to each of a hundred half-size subsamples. The first four predictor metabolites that entered the regularization path for each sensory trait were selected. The final selection retained those predictors that were selected for at least $\pi * 100$ percent of the subsamples. π was chosen such that the expected number of false positives, i.e. $4^2/(p * (2\pi - 1))$, was bounded at 1, where p is the number of metabolites (Meinshausen and Buhlmann 2010).

3.3.6 The QPSO algorithm

Inferring causal phenotype networks contributes to predicting the effects of external interventions on traits (Valente et al. 2013), and thereby attracts a surge of research interest (Rosa et al. 2011). Current approaches mainly exploit previously determined QTLs to learn about causal relationships between traits. These methods require at least one unique QTL for each and every trait. This prerequisite, however, is often not met in practice due to various reasons such as limited samples sizes, small QTL effects and high noise levels. To get rid of this unrealistic prerequisite, the QPSO algorithm has been presented very recently (Wang and van Eeuwijk 2014). This algorithm is applied to a pre-learnt undirected phenotype network, based on which it searches for the optimal causal phenotype network through a heuristic strategy. A major advantage of the QPSO algorithm is that it takes into account the relevant phenotypic interactions in addition to the detected QTLs when orienting an undirected edge between two traits. As a result, it is applicable to general cases where some traits lack unique QTLs, or, come without QTL.

3.4 Results

3.4.1 A dependency network involving 29 metabolites and 14 QTLs

Figure 3.2 presents a dependency network involving 29 metabolites in ripe tomatoes and the most significant 14 QTLs (p-value<0.01) identified by multi-trait mixed model analysis for the metabolites. Except two QTLs, rs6495 and rs8314, which were responsible for beta-damascenone and cis-3-hexenol respectively, all other QTLs were found associated with multiple metabolites. In particular, rs2050 had pleiotropic effects on many metabolites, including eleven volatiles, two sugars and three acids. For two metabolites, eugenol and trans-2-hexenal, no QTL was identified. Another ten metabolites were, respectively, associated with one QTL. Each of the remaining metabolites was associated with two or more QTLs.

Figure 3.2 indicates a separation between primary and secondary metabolism, i.e., sugars and acids on the left whereas volatiles on the right. Further, (1) sugars and a sugar alcohol, myo-inositol, were grouped together; (2) acids were gathered and linked to sugars; (3) most volatiles interacted, and a few of them were connected with sugars/acids.

Metabolic profiling of ripe tomatoes was carried out at single time points after harvest, that is, it did not produce time series data. The dependency network (**Figure 3.2**) learnt from non-sequential metabolic data cannot be interpreted as metabolic pathways; instead, it represented directed associations at the level of mean metabolite abundances. These dependencies, though essentially different from pathways, still provide hints on how the effects of allele substitutions propagate through metabolites. For example, genotypic changes at locus rs6691 shall alter the concentration of 1-penten-3-one. This will probably subsequently affect the concentrations of beta-ionone, cis-3-hexenal and aspartic acid. Conversely, variations in the concentration of 1-penten-3-one are unlikely to affect the concentration of trans-2-hexenal, since trans-2-hexenal was found a parent node of 1-penten-3-one in the dependence network.

A better understanding of the dependence structure underlying multiple traits contributes to a better manipulation of those traits. Assume we want to regulate the concentration of beta-ionone, we should control genotypes at loci rs6691 and rs3540 in addition to those at rs2050 and rs6254. The reason is that any allele substitution leading to a change in the concentration of 1-penten-3-one might then alter the concentration of beta-ionone.

3.4.2 A dependency network involving 24 sensory traits and 7 QTLs

Figure 3.3 shows a dependency network involving 24 sensory traits in ripe tomatoes and the most significant 7 QTLs (p-value<0.01) identified by multi-trait mixed model analysis for the sensory traits. Among the 7 QTLs, rs8591 and rs8016 were respectively responsible for one sensory trait; each of the remaining QTLs was associated with multiple sensory traits. From another perspective, 7 sensory traits came without QTLs, while every other trait was identified with at least one QTL.

Figure 3.3 is helpful to predict the simultaneous influence of various allele substitutions on multiple sensory traits. Assume that a genotypic change at locus rs7448 raises the level of scent_tomato. Accordingly there might be a decrease in scent_smoky, and further, an increase in scent_sweet. However, to finely predict one or more phenotypes, a comprehensive consideration of multiple allele substitutions is usually required. For instance, an increase in scent_tomato is not necessarily coupled with a decrease in scent_smoky. This is because apart from QTL rs7448, which had direct negative effect on scent_tomato and, subsequently, indirect positive influence on scent_smoky, scent_smoky was found also being regulated by another two QTLs rs7775 and rs8016. Analogously, scent_sweet was directly or indirectly determined by 5 QTLs, including rs7089, rs8434, rs7448, rs7775 and rs8016.

3.4.3 A dependency network involving brix, 29 metabolites, 24 sensory traits and 21 QTL

Figure 3.4 shows a dependency network involving brix as well as all metabolites, sensory traits and QTLs mentioned above. Brix was found to be dependent on two sugars sucrose and fructose and the sugar alcohol myo-inositol; meanwhile, it was found a main factor influencing taste_sweet. This does not come as a surprise, as brix whilst being a measure of total soluble solids content is most often used to measure sugar content. Indeed, silencing an invertase had a strong influence on brix (Fridman et al. 2000; Zanor et al. 2009). Citric acid was involved in the determination of taste_sour, aftertaste_fresh and taste_tomato. Sucrose, in addition to citric acid, was also a predictor of taste_tomato. Scent_smoky was driven by methyl_salicylate, which was positively affected by guaiacol. This coincides with the previous findings that both methyl_salicylate and guaiacol contribute to the smokey smell of tomatoes (Buttery et al. 1987; Buttery et al. 1990), though a recent study indicates that guaiacol is probably a more important contributor (Tikunov et al. 2013).

In addition to the aforementioned positive directed associations between metabolites and sensry traits, three negative dependencies were respectively found between eugenol and aftertaste_fresh, 2-methyl-1-butanol and aftertaste_chemical, as well as 2-methyl-1-butanol and aftertaste_sweet. The latter two are in agreement with the fact that 2-methyl-1-butanol is often found in fruits (NCBI PubChem) and that it seems to improve or partially impart an Italico-cheese flavor (US 3978242 A), which would not be perceived as a chemical taste but rather associated with natural products.

By taking into account the directed associations from metabolites to sensory traits, we were able to get a more realistic estimation of the dependence structure underlying those sensory traits. An example is that in **Figure 3.3** aftertaste_sour is present as a parent node of taste_sour, while in **Figure 3.4** a reversed dependency, which seems more logical, is achieved simply because an additional determinant citric_acid has been introduced to taste_sour.

3.5 Discussion

3.5.1 Comparison with known metabolic reactions

As noted above, though the metabolic part of network was learnt from non-sequential data and thus intrinsically not a representation of metabolic pathways, it is still to some extent informative about the regulatory mechanisms underlying those metabolites.

There was a separation between primary and secondary metabolites. This of course makes sense considering the structural function of primary metabolites and the auxiliary function of secondary metabolites. Interestingly, within the primary metabolites, sucrose was the parent of fructose which in turn was the parent of glucose. This may be due to the enzymatic action of invertase which splits sucrose into glucose and fructose. And the direct link between sucrose and glucose was recovered as an indirect one is potentially due to the additional action of Sucrose Synthase utilizing fructose and UDP-glucose.

It is noteworthy that in **Figure 3.4** fructose and glutamic acid were present as parent nodes of myo-inositol which in turn was the parent of sucrose. Metabolically myo-inositol is synthesized from glucose-6-phosphate via D-myo-inositol 3-phosphate (Hegeman et al. 2001). But since neither glucose-6-phosphate nor D-myo-inositol 3-phosphate were quantified in this study, the network reconstruction and orientation algorithms might have compacted the network. Whilst this leaves the link from glutamic acid unexplained, it seems like a good testable hypothesis for the sugars and the sugar alcohol myo-inositol, which could be explored.

For glutamic acid a direct and strong influence was observed from aspartic acid. Metabolically this might be explained by the enzymatic action of aspartate aminotransferase that converts glutamic acid and oxaloacetate into 2-oxoglutarate and aspartate. Indeed, aspartate aminotransferase has already been implicated in glutamate content in red tomato fruits (Boggio et al. 2000). Comparatively, the impact of malic acid on aspartic acid seems less obvious. That said, an RNAi approach against PEPCK revealed strongly increased aspartic acid levels coinciding with reduced malate levels. However, silencing of NADP-malic enzyme in the same study showed less aspartic acid and somewhat lower malic acid levels in one transgenic line (Osorio et al. 2013).

Turning to volatiles as flavor carrying compounds it is obvious that the most-likely carotenoid derived volatiles beta-damascenone (Mathieu et al. 2009) and beta-ionone (Baldermann et al. 2010) were linked because of the common precursor beta-carotene. However, the deduced influence of one on the other might only be explained by hidden variables such as the actual enzyme activities and actual carotenoid concentrations not measured here. Also it is intriguing that 6-methyl-5-hepten-2-one and geranylacetone, both being interconnected, were not linked to the former pair of volatiles despite them also being carotenoid volatiles. The different differential behaviors of these two pairs of volatiles were also observed in earlier studies (Mathieu et al. 2009), and it has been reported that 6-methyl-5-hepten-2-one likely stems from lycopene (Gao et al. 2008). We therefore suspect the difference is attributed to distinct precursors. Apart from these carotenoid derived metabolites, the synthesis of phenylethyl alcohol from benzeneacetaldehyde (Sakai et al. 2007) was recovered in our analysis.

Regarding the linked metabolites 3-methyl-1-butanol and 3-methylbutanal, they are most likely leucine derived, whilst the associated 2-methyl-1-butanol likely stems from isoleucine. Also the association between 2-isobutylthiazole and 3-methyl-1-butanol was observed before (Mathieu et al. 2009; Tikunov et al. 2005). Thus this whole sub-cluster of metabolites is derived from or associated to branched chain amino acids. The current model for the biosynthesis of leucine-derived flavor imparting compounds assumes a decarboxylation to an aldehyde followed by a reduction. The truth, however, is that the alcohols should derive from the aldehydes.

3.5.2 Choice of methods and parameters

The most straightforward way to construct biological networks is the correlation network (also known as relevance network), which is based on unconditional pairwise correlations. However, though strong correlations are good indicators of dependencies, they cannot distinguish between direct and indirect associations. Thus, correlation networks are typically dense graphs, from which definitive conclusions can hardly be drawn (see examples in **Figure 3.S5A** and **B**). To learn less dense but more informative graphs, especially from high-dimensional data with limited sample size, here we used three approaches to graphical modeling: LBNS + StARS, the PC-skeleton algorithm, and the Lasso + SS.

StARS has been shown to outperform the conventional regularization parameter selection methods, including AIC, BIC and cross-validation, in the reconstruction of high-dimensional graphs (Liu et al. 2010). In view of this, we exploited StARS to set regularization in LBNS. The R package “huge” implements StARS with two optional variability thresholds: 0.1 and 0.05. We tested both thresholds on the metabolic and sensory data separately, and found that 0.05 led to a bit sparser graph than 0.01 (**Figure 3.S1A** vs. **B**, **Figure 3.S2A** vs. **B**). As we aimed to extract a consensus network, the variability threshold of 0.1 was then used in StARS to ensure that given the same dataset, edges obtained by LBNS can overlap, to a large extent, with those learnt by the PC-skeleton algorithm.

The PC-skeleton algorithm also requires a pre-specified parameter, i.e. the significance level of conditional independence tests. We tested the two most common significance levels, 0.01 and 0.05, on the metabolic and sensory data separately. Results on the same datasets indicated that the significance level of 0.05 recovered a few more edges than the level of 0.01 (**Figure 3.S3A** vs. **B**, **Figure 3.S3C** vs. **D**). Again, to reach as many as possible consensus edges, we took the significance level of 0.05 in this study.

Our strategy, which first overfits an undirected graph by LBNS + StARS and then screens out the unlikely edges by comparison with the outcome of the PC-skeleton algorithm, was also tried on the mixture of metabolic and sensory data. Surprisingly, only a few links between metabolites and sensory traits were discovered by either method (see black edges in **Figure 3.S4A** and **B**). After discarding edges unique to one graph, we were left with merely eight common links (see the boldfaced black edges in **Figure 3.S4A** or **B**). This implies that the above strategy, when being used to decipher the relationships between substances of different nature, is very likely to produce an underfitted graph. We then tried a third method, i.e. regressing every sensory trait on the metabolites by the Lasso + SS, to get the directed graph in **Figure 3.S4C**. To draw safe conclusions but without losing too much useful information, we extracted those edges that appeared between metabolites and sensory traits at least twice over **Figure 3.S4A**, **B** and **C**. Finally, 12 edges satisfying this criterion were reported (see black edges in **Figure 3.4**).

3.5.3 Other aspects

Multi-trait analysis is in general preferred over single-trait analysis for QTL mapping. This is because: (1) multi-trait analysis takes into account the genetic correlations among traits and thus increases the power of detecting QTLs (Jiang and Zeng 1995); (2) it allows a more straightforward assessment of pleiotropic effects of QTLs (Alimi et al. 2013; Malosetti et al. 2008). Nonetheless, the outputs of multi-trait QTL analyses not necessarily fully encompass the results of single-trait analyses. That is, a QTL identified by single-trait analysis can be missed in multi-trait analysis, though this rarely happens. In this study we missed a QTL for scent smoky on chromosome 9, whereas this QTL was clearly identified in another study with the same material (Tikunov et al. 2013). We were able to detect the QTL when rerunning a single-trait analysis for scent smoky. A limited multi-trait analysis on scent smoky and some volatiles that are known to be related to scent smoky produced the QTL as well.

We have identified a total of 21 QTLs for the 29 metabolites and 24 sensory traits. Most of the QTLs were found to have pleiotropic effects; in particular, a few of them, such as rs2050, rs6687, rs7089 and rs7448, served as hubs in the resulting dependency network (**Figure 3.4**). A particularly noteworthy phenomenon was that a number of directed triangles appeared in **Figure 3.4**, especially around the hubs. One may doubt whether the QTL really affects so many traits? Does its impact on a downstream trait actually pass through the upstream traits? Moreover, will two directly associated traits become independent of each other given their common QTL? A possible solution to these detailed questions is the triad analysis, which aims at identifying causal relationships in configurations consisting of two traits and one QTL (Li et al. 2010; Schadt et al. 2005).

Though both SS and StARS can choose a proper regularization for high-dimensional sparse linear regression, they are essentially different. Given the same training dataset, StARS tolerates false positives (false edges in the reconstructed graph) but not false negatives (true edges absent in the reconstructed graph) and thus leads to a dense graph with high recall but relatively low precision (in the context of graphical modeling, recall refers to the fraction of true edges that are recovered in the resulting graph; precision refers to the fraction of recovered edges that are actually true); SS, contrariwise, allows false negatives but not false positives and therefore results in a sparse graph with high precision but comparatively low recall.

3.6 Conclusion

We have investigated the utility of existing methods for GGM reconstruction in combination with the QPSO algorithm for dependency inference between 29 metabolites and 24 sensory traits scored on ripe tomatoes. The resulting network provides hints on how the sensory traits depend upon the metabolites and further upon the detected QTLs. This integrative approach does not require the identification of QTLs for each and every trait studied, and thus has broad applicability across a number of practical settings. Furthermore, it is applicable to a range of population structures, including offspring populations from crosses between inbred parents and outbred parents, association panels and natural population. The novel dependencies emerged in this study form the hypotheses that can be individually tested in the future.

References

Abegaz EG, Tandon KS, Scott JW, Baldwin EA, Shewfelt RL (2004) Partitioning taste from aromatic flavor notes of fresh tomato (Lycopersicon esculentum, Mill) to develop predictive models as a function of volatile and nonvolatile components. Postharvest Biology and Technology 34:227-235

Alimi NA, Bink MCAM, Dieleman JA, Magan JJ, Wubs AM, Palloix A, van Eeuwijk FA (2013) Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. Theor Appl Genet 126:2597-2625

Aten JE, Fuller TF, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: The NEO software. Bmc Systems Biology 2

Baldermann S, Kato M, Kurosawa M, Kurobayashi Y, Fujita A, Fleischmann P, Watanabe N (2010) Functional characterization of a carotenoid cleavage dioxygenase 1 and its relation to the carotenoid accumulation and volatile emission during the floral development of Osmanthus fragrans Lour. J Exp Bot 61:2967-2977

Boggio SB, Palatnik JF, Heldt HW, Valle EM (2000) Changes in amino acid composition and nitrogen metabolizing enzymes in ripening fruits of Lycopersicon esculentum Mill. Plant Sci 159:125-133

Buttery RG, Ling LC, Light DM (1987) Tomato Leaf Volatile Aroma Components. J Agr Food Chem 35:1039-1042

Buttery RG, Takeoka G, Teranishi R, Ling LC (1990) Tomato Aroma Components - Identification of Glycoside Hydrolysis Volatiles. J Agr Food Chem 38:2050-2053

Cai XD, Bazerque JA, Giannakis GB (2013) Inference of Gene Regulatory Networks with Sparse Structural Equation Models Exploiting Genetic Perturbations. Plos Computational Biology 9

Carli P, Arima S, Fogliano V, Tardella L, Frusciante L, Ercolano MR (2009) Use of network analysis to capture key traits affecting tomato organoleptic quality. J Exp Bot 60:3379-3386

Carreno-Quintero N, Bouwmeester HJ, Keurentjes JJB (2013) Genetic analysis of metabolome-phenotype interactions: from model to crop species. Trends in Genetics 29:41-50

Colombo D, Hauser A, Kalisch M, Maechler M (2014) Package 'pcalg'.

http://cran.r-project.org/web/packages/pcalg/pcalg.pdf

Drton M, Perlman MD (2004) Model selection for Gaussian concentration graphs. Biometrika 91:591-602

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. P Natl Acad Sci USA 97:4718-4723

Gao HY, Zhu HL, Shao Y, Chen AJ, Lu CW, Zhu BZ, Luo YB (2008) Lycopene accumulation affects the biosynthesis of some carotenoid-related volatiles independent of ethylene in tomato. J Integr Plant Biol 50:991-996

Hegeman CE, Good LL, Grabau EA (2001) Expression of D-myo-inositol-3-phosphate synthase in soybean. Implications for phytic acid biosynthesis. Plant Physiol 125:1941-1948

Jiang CJ, Zeng ZB (1995) Multiple-Trait Analysis of Genetic-Mapping for Quantitative Trait Loci. Genetics 140:1111-1127

Jiang CJ, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101:47-58

Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. Bmc Systems Biology 5

Li RH, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA (2006) Structural model analysis of multiple quantitative traits. Plos Genetics 2:1046-1057

Li Y, Tesson BM, Churchill GA, Jansen RC (2010) Critical reasoning on causal inference in genome-wide linkage and association studies. Trends in Genetics 26:493-498

Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (StARS) for high dimensional graphical models. Advances in Neural Information Processing Systems

Logsdon BA, Mezey J (2010) Gene Expression Network Reconstruction by Convex Feature Selection when Incorporating Genetic Perturbations. Plos Computational Biology 6

Malosetti M, Ribaut JM, Vargas M, Crossa J, van Eeuwijk FA (2008) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (Zea mays L.). Euphytica 161:241-257

Mathieu S, Cin VD, Fei ZJ, Li H, Bliss P, Taylor MG, Klee HJ, Tieman DM (2009) Flavour compounds in tomato fruits: identification of loci and potential pathways affecting volatile composition. J Exp Bot 60:325-337

Meinshausen N, Buhlmann P (2006) High-dimensional graphs and variable selection with the Lasso. Annals of Statistics 34:1436-1462

Meinshausen N, Buhlmann P (2010) Stability selection. Journal of the Royal Statistical Society Series B-Statistical Methodology 72:417-473

Neto EC, Ferrara CT, Attie AD, Yandell BS (2008) Inferring causal phenotype networks from segregating populations. Genetics 179:1089-1100

Neto EC, Keller MP, Attie AD, Yandell BS (2010) Causal Graphical Models in Systems Genetics: A Unified Framework for Joint Inference of Causal Network and Genetic Architecture for Correlated Phenotypes. Annals of Applied Statistics 4:320-339

Osorio S, Vallarino JG, Szecowka M, Ufaz S, Tzin V, Angelovici R, Galili G, Fernie AR (2013) Alteration of the Interconversion of Pyruvate and Malate in the Plastid or Cytosol of Ripening Tomato Fruit Invokes Diverse Consequences on Sugar But Similar Effects on Cellular Organic Acid, Metabolism, and Transitory Starch Accumulation. Plant Physiol 161:628-643

Politis DN, Romano JP, Wolf M (1999) Subsampling, 1st edn. Springer

Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol 133:84-99

Rosa GJM, Valente BD, de los Campos G, Wu XL, Gianola D, Silva MA (2011) Inferring causal phenotype networks using structural equation models. Genetics Selection Evolution 43

Sakai M, Hirata H, Sayama H, Sekiguchi K, Itano H, Asai T, Dohra H, Hara M, Watanabe N (2007) Production of 2-phenylethanol in roses as the dominant floral scent compound from L-phenylalanine by two key enzymes, a PLP-Dependent decarboxylase and a phenylacetaldehyde reductase. Biosci Biotech Bioch 71:2408-2419

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang CS, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang

LM, Castle J, Zhu HY, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nature Genetics 37:710-717

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge, Mass.

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3:211-221

Tandon KS, Baldwin EA, Scott JW, Shewfelt RL (2003) Linking sensory descriptors to volatile and nonvolatile components of fresh tomato flavor. J Food Sci 68:2366-2371

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 58:267-288

Tikunov Y, Lommen A, de Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol 139:1125-1137

Tikunov YM, Molthoff J, de Vos RCH, Beekwilder J, van Houwelingen A, van der Hooft JJJ, Nijenhuis-de Vries M, Labrie CW, Verkerke W, van de Geest H, Zamora MV, Presa S, Rambla JL, Granell A, Hall RD, Bovy AG (2013) NON-SMOKY GLYCOSYLTRANSFERASE1 Prevents the Release of Smoky Aroma from Tomato Fruit. Plant Cell 25:3067-3078

Valente BD, Rosa GJM, Gianola D, Wu XL, Weigel K (2013) Is Structural Equation Modeling Advantageous for the Genetic Improvement of Multiple Traits? Genetics 194:561-572

Vrebalov J, Ruezinsky D, Padmanabhan V, White R, Medrano D, Drake R, Schuch W, Giovannoni J (2002) A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (Rin) locus. Science 296:343-346

Wang HG, van Eeuwijk FA (2014) A New Method to Infer Causal Phenotype Networks Using QTL and Phenotypic Information. Plos One 9

Zanor MI, Osorio S, Nunes-Nesi A, Carrari F, Lohse M, Usadel B, Kuhn C, Bleiss W, Giavalisco P, Willmitzer L, Sulpice R, Zhou YH, Fernie AR (2009) RNA Interference of LIN5 in Tomato Confirms Its Role in Controlling Brix Content, Uncovers the

Influence of Sugars on the Levels of Fruit Hormones, and Demonstrates the Importance of Sucrose Cleavage for Normal Fruit Development and Fertility. Plant Physiol 150:1204-1218

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2012) The huge Package for High-dimensional Undirected Graph Estimation in R. J Mach Learn Res 13:1059-1062

Figure 3.1 A schematic diagram of the proposed integrative method for learning dependency networks from the sensory, metabolic and QTL data.

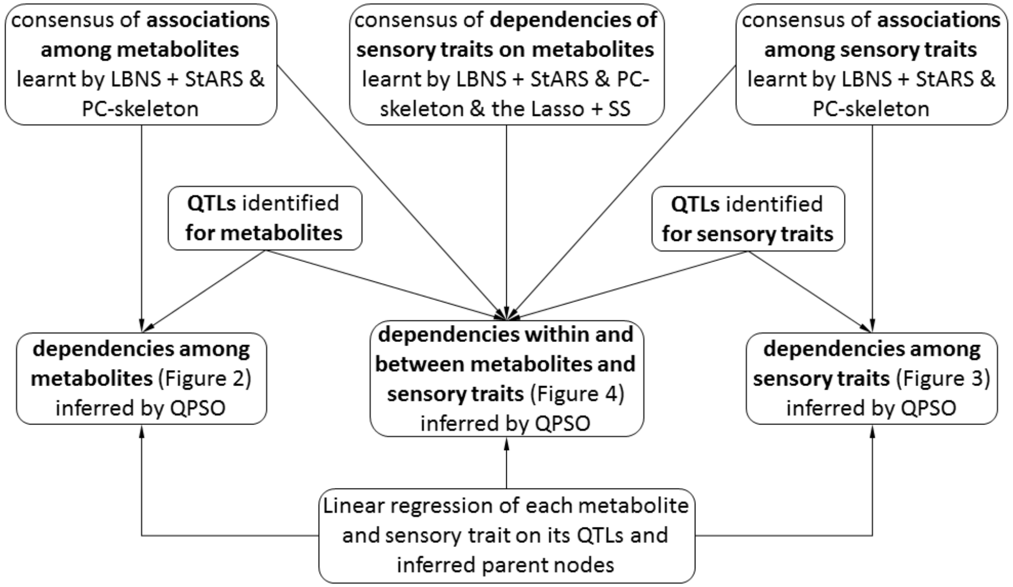

consensus of **associations among metabolites** learnt by LBNS + StARS & PC-skeleton
consensus of **dependencies of sensory traits on metabolites** learnt by LBNS + StARS & PC-skeleton & the Lasso + SS
consensus of **associations among sensory traits** learnt by LBNS + StARS & PC-skeleton
QTLs identified **for metabolites**
QTLs identified **for sensory traits**
dependencies among metabolites (Figure 2) inferred by QPSO
dependencies within and between metabolites and sensory traits (Figure 4) inferred by QPSO
dependencies among sensory traits (Figure 3) inferred by QPSO
Linear regression of each metabolite and sensory trait on its QTLs and inferred parent nodes

Figure 3.2 A dependency network of 29 metabolites and 14 QTLs detected in ripe tomatoes. Red edges connect QTLs to their target traits; blue edges represent the dependencies between metabolites. Line style and thickness are determined by the fitted coefficients of each metabolite being regressed on its QTLs and inferred parent nodes. Specifically, thicker lines indicate stronger dependencies; positive and negative dependencies are distinguished by solid and dashed lines. In particular, a solid red edge indicates the round allele at the QTL increases the trait, while a dashed red edge indicates the cherry allele at the QTL increases the trait.

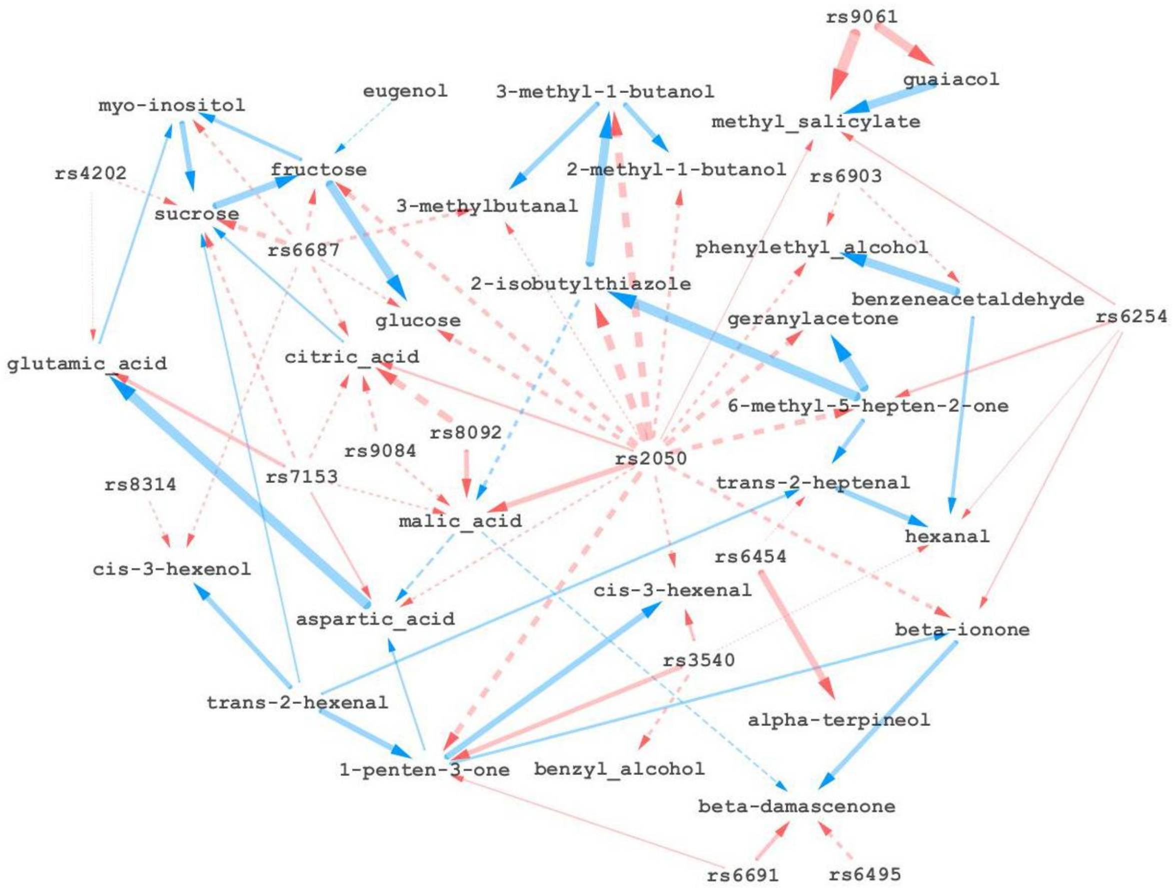


Figure 3.3 A dependency network of 24 sensory traits and 7 QTLs detected in ripe tomatoes. Red edges connect QTLs to their target traits; green edges represent the dependencies between sensory traits. Line style and thickness are determined by the fitted coefficients of each sensory trait being regressed on its QTLs and inferred parent nodes. Solid and dashed linestyle schemes are identical to those in Figure 3.2.

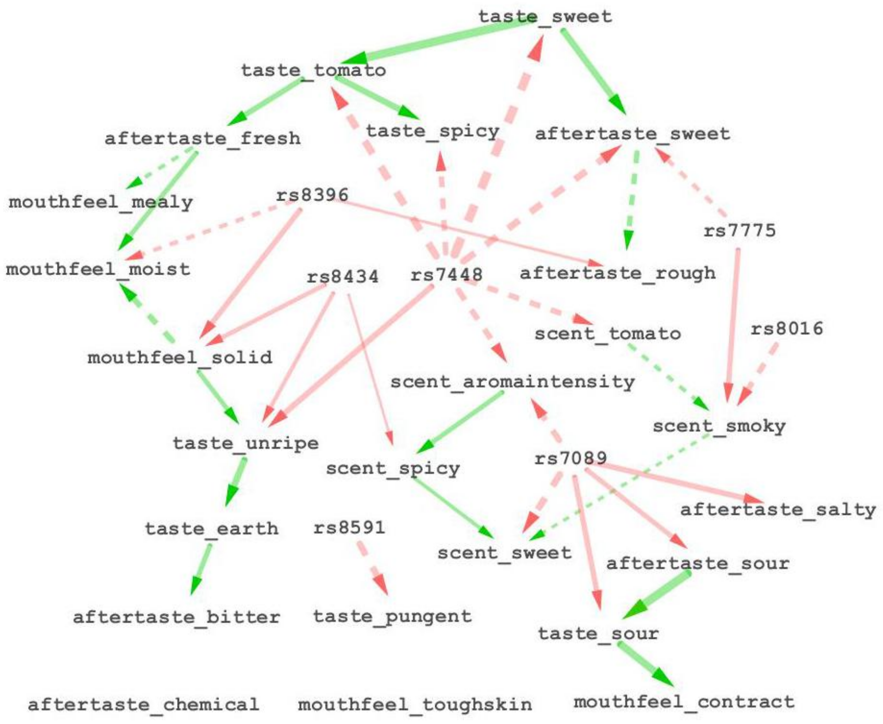


Figure 3.4 A dependency network of brix, 29 metabolites, 24 sensory traits and 21 QTLs detected in ripe tomatoes. Black edges represent dependencies of sensory trait on metabolites (via brix). Red, blue and green edges, together with their linestyle and thickness are identical to those in Figure 3.2 and 3.

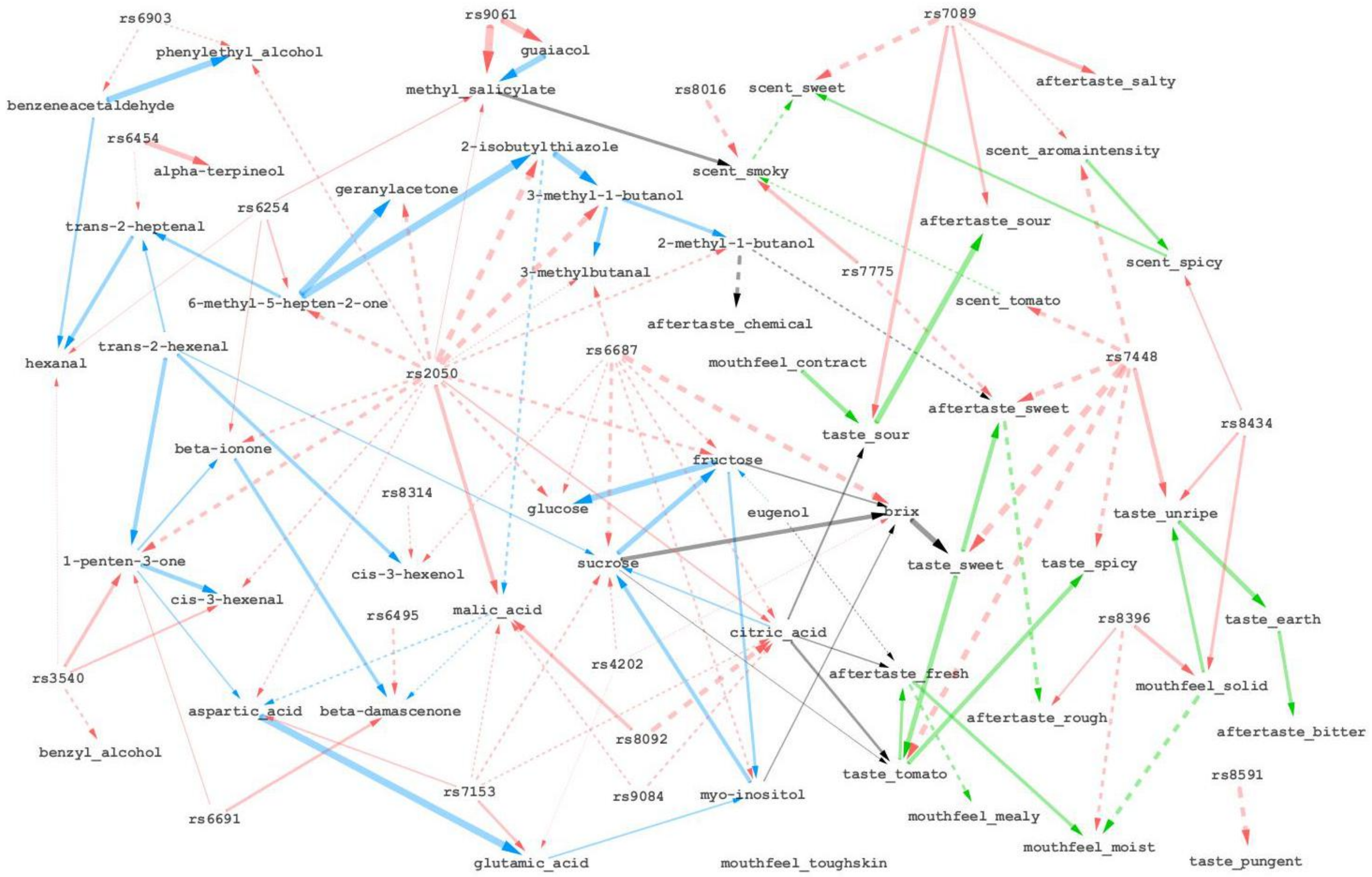


Figure 3.S1 Two correlation networks involving 29 metabolites or 24 sensory traits in ripe tomatoes. Nodes denote metabolites in (A) and sensory traits in (B). Edges represent the significant pairwise Pearson correlations (the p-value associated with the t-test < 0.05) between the nodes. Solid and dashed edges indicate positive and negative correlations, respectively.

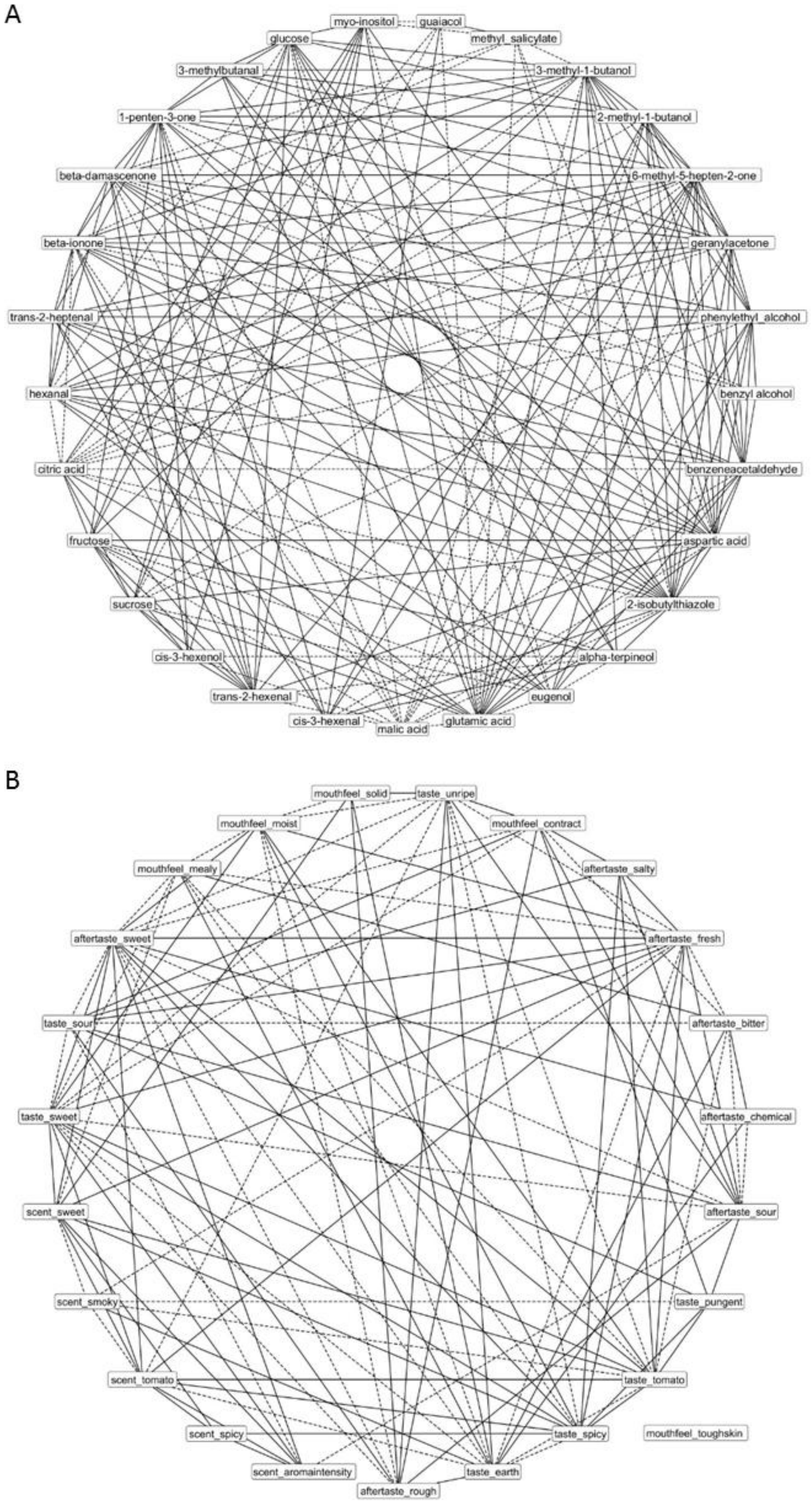


Figure 3.S2 Two undirected networks involving 29 metabolites in ripe tomatoes. Nodes denote metabolites. Edges represent the strong symmetric associations between nodes. Both (A) and (B) are learnt by LBNS + StARS. The variability threshold used in StARS to construct (A) is 0.05 while the threshold used to construct (B) is 0.1.

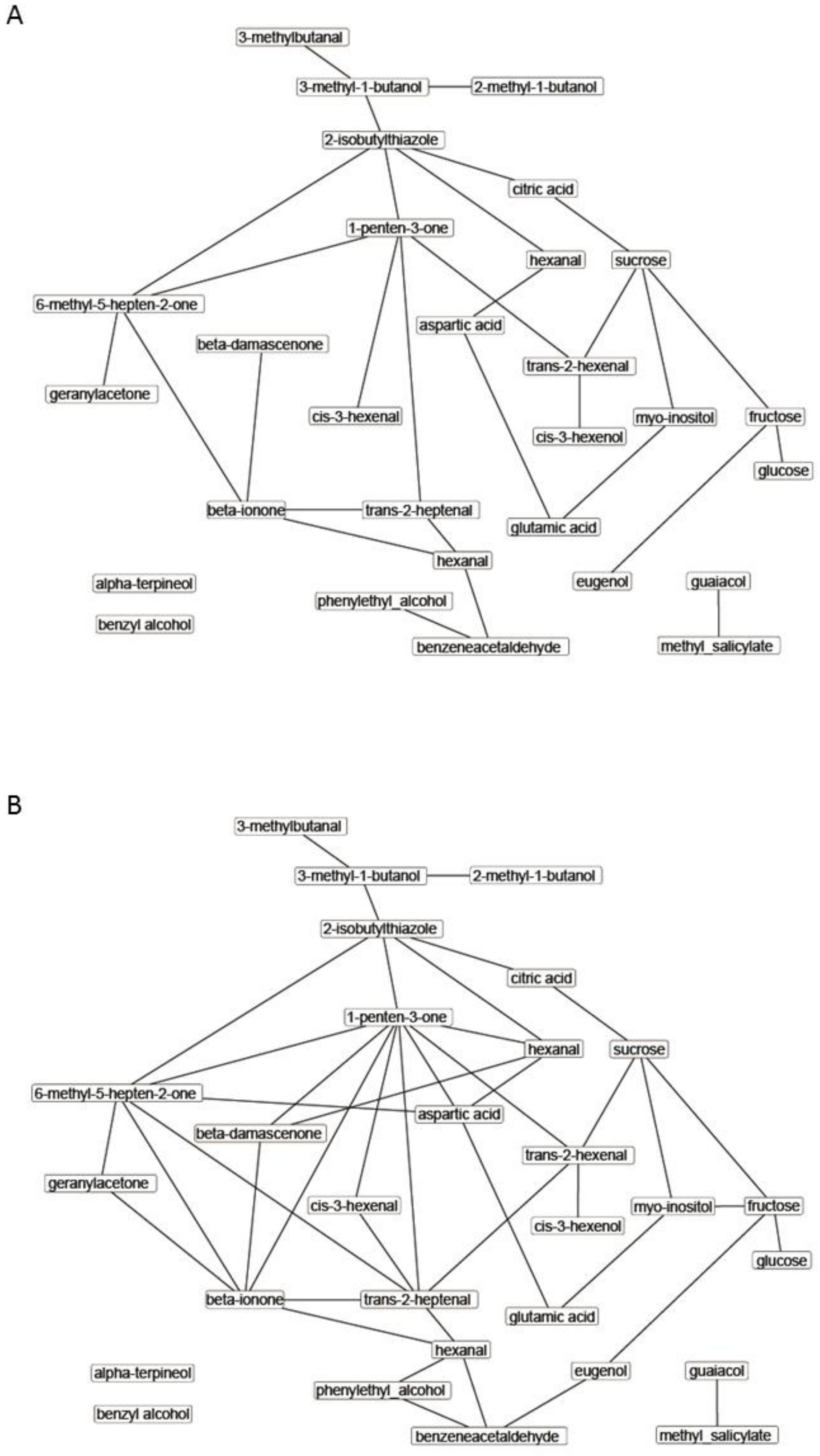


Figure 3.S3 Two undirected networks involving 24 sensory traits in ripe tomatoes. Nodes denote sensory traits. Edges represent the strong symmetric associations between nodes. Both (A) and (B) are learnt by LBNS + StARS. The variability threshold used in StARS to construct (A) is 0.05 while the threshold used to construct (B) is 0.1.

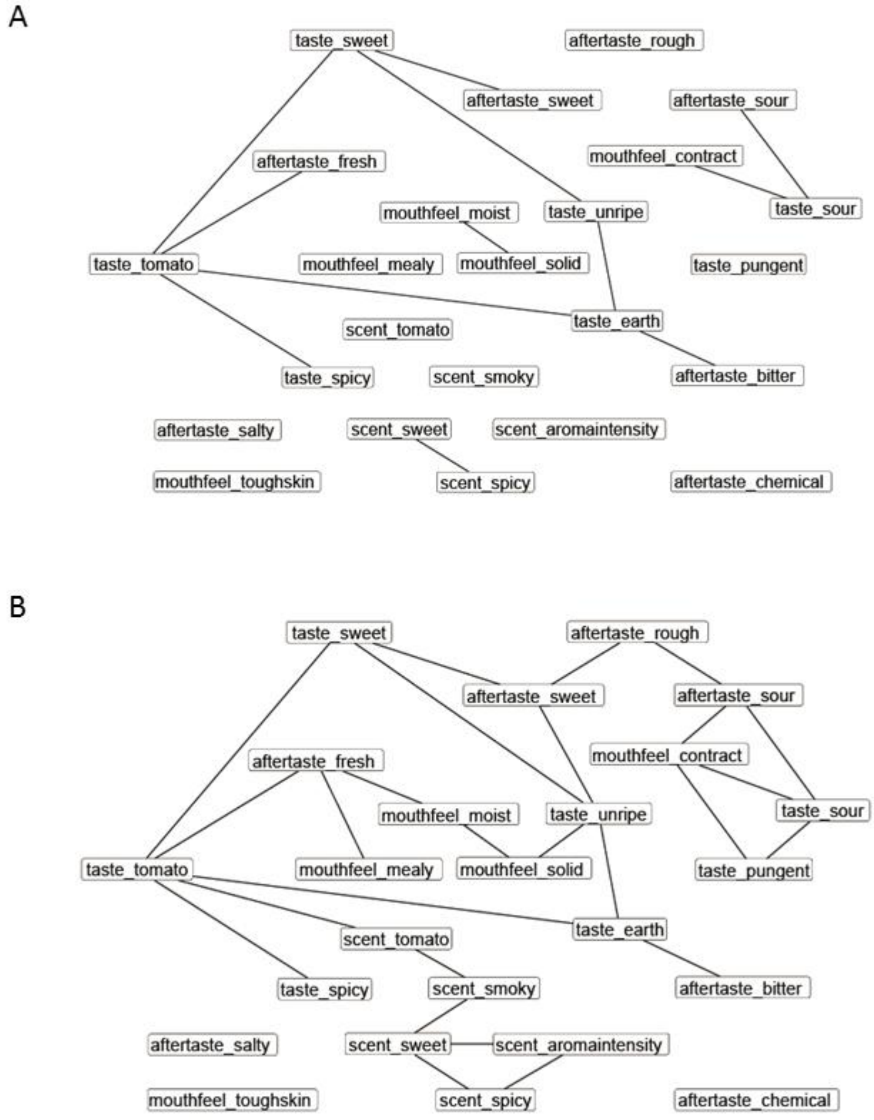


Figure 3.S4 Four undirected networks involving 29 metabolites or 24 sensory traits in ripe tomatoes. Nodes denote metabolites in (A) and (B), and sensory traits in (C) and (D). Edges represent the strong symmetric associations between nodes. All four graphs are learnt by the PC-skeleton algorithm, where the significance level of conditional independence tests used to construct (A) and (C) is 0.01 and the one used to construct (B) and (D) is 0.05.

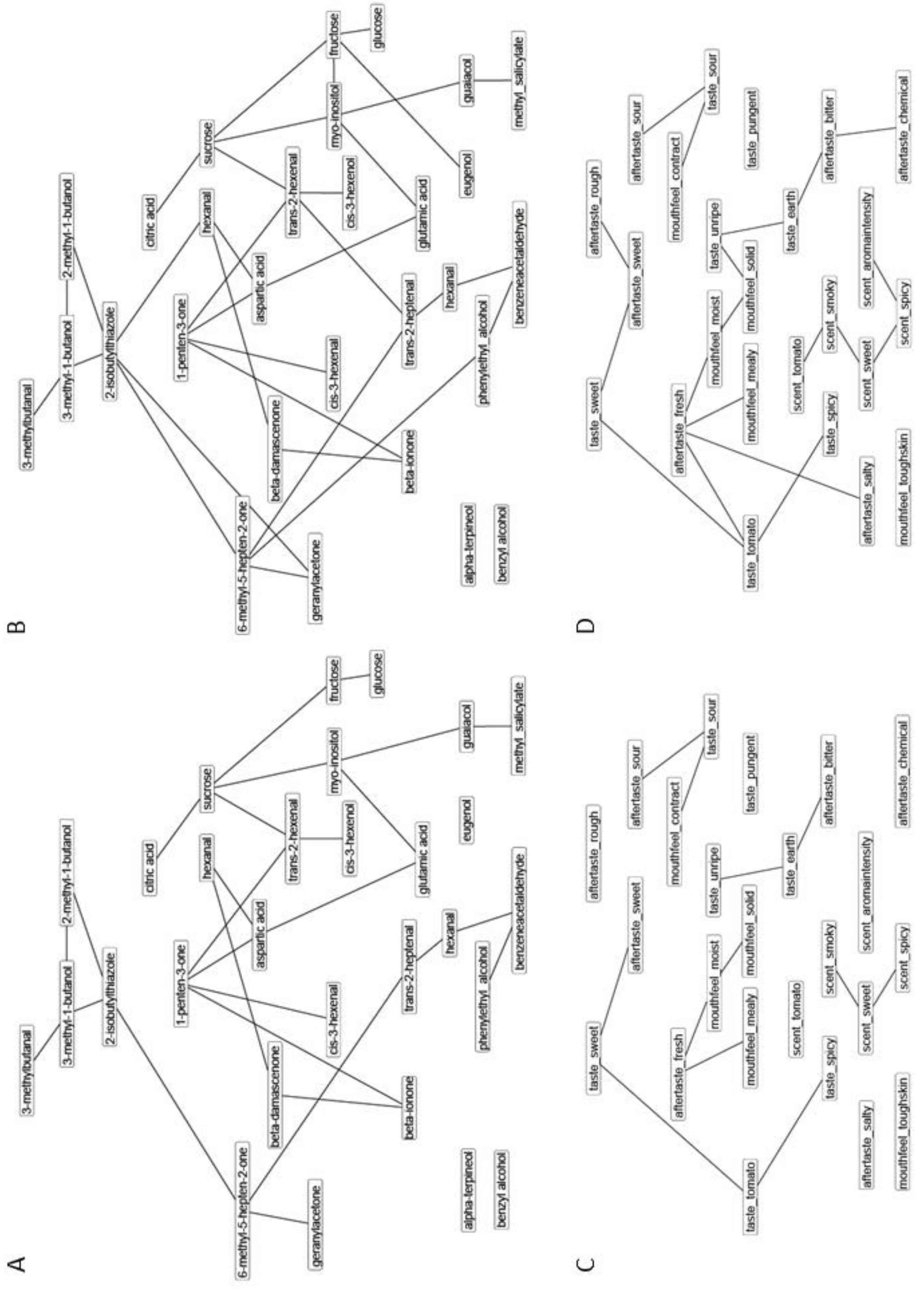


Figure 3.S5 Three networks regarding the relationships between 29 metabolites, brix and 24 sensory traits in ripe tomatoes. Blue, green and lavender nodes are used to distinguish between metabolites, sensory traits and brix. Edges in (A) and (B) are, respectively, learnt by LBNS + StARS (variability threshold set at 0.1) and the PC-skeleton algorithm (significance level of 0.05 for conditional independence tests). Edges in (C) are learnt by the Lasso + SS (where explanatory variables selected over 77.6% of 100 half-sized subsamples are returned for each response variable). Particularly, black edges highlight strong dependencies between metabolites and sensory traits (some via brix), where the bold black edges are consistent in (A) and (B).

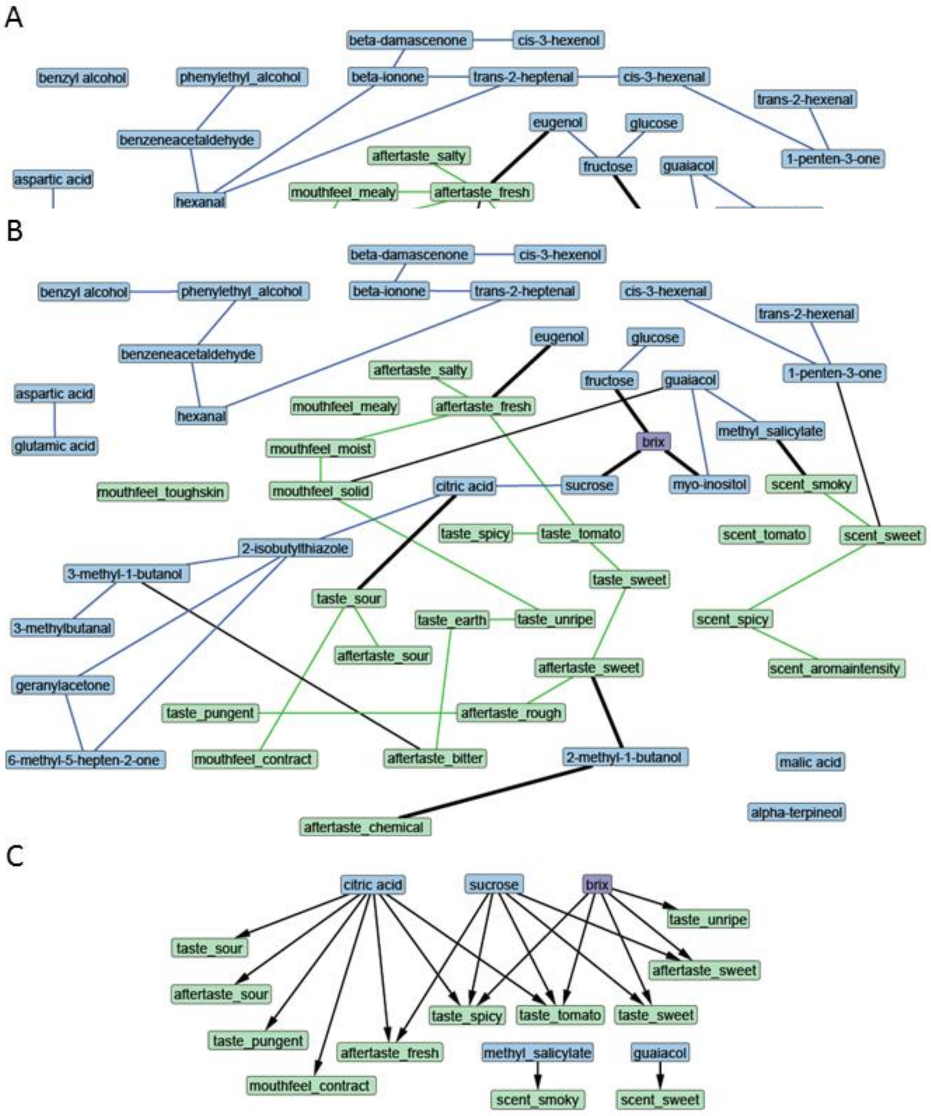


Table 3.1. Position information of QTLs (coincide with SNP identifiers) identified for the 29 metabolites and 24 sensory traits. Map positions obtained from an integrated map of four tomato populations following from crosses between cherry and round tomato parent lines (see Material section).

SNP/QTL	Chromosome	Position	Metabolite	SNP/QTL	Chromosome	Position	Metabolites	SNP/QTL	Chromosome	Position	Sensory Traits
rs8314	1	10.35	cis-3-hexenol	rs2050	5	39.26	6-methyl-5-hepten-2-one	rs7448	2	63.5	scent_aromaintensity
rs6454	1	128.97	trans-2-heptenal	rs2050	5	39.26	geranylacetone	rs7448	2	63.5	scent_tomato
rs6454	1	128.97	alpha-terpineol	rs2050	5	39.26	aspartic acid	rs7448	2	63.5	taste_sweet
rs6687	2	76.3	cis-3-hexenol	rs2050	5	39.26	citric acid	rs7448	2	63.5	taste_tomato
rs6687	2	76.3	myo-inositol	rs2050	5	39.26	malic acid	rs7448	2	63.5	taste_unripe
rs6687	2	76.3	3-methylbutanal	rs2050	5	39.26	fructose	rs7448	2	63.5	taste_spicy
rs6687	2	76.3	citric acid	rs2050	5	39.26	glucose	rs7448	2	63.5	aftertaste_sweet
rs6687	2	76.3	glucose	rs4202	6	4.03	glutamic acid	rs8396	3	103.79	mouthfeel_moist
rs6687	2	76.3	fructose	rs4202	6	4.03	sucrose	rs8396	3	103.79	mouthfeel_solid
rs6687	2	76.3	sucrose	rs4202	6	4.03	brix	rs8396	3	103.79	aftertaste_rough
rs6687	2	76.3	brix	rs3540	6	14.14	cis-3-hexenal	rs7775	7	56.23	scent_smoky
rs6691	4	55.68	1-penten-3-one	rs3540	6	14.14	1-penten-3-one	rs7775	7	56.23	aftertaste_sweet
rs6691	4	55.68	beta-damascenone	rs3540	6	14.14	hexanal	rs8016	8	11.04	scent_smoky
rs7153	4	86.64	sucrose	rs3540	6	14.14	benzyl alcohol	rs8591	8	58.98	taste_pungent
rs7153	4	86.64	malic acid	rs6254	6	47.72	beta-ionone	rs7089	10	7.73	scent_aromaintensity
rs7153	4	86.64	aspartic acid	rs6254	6	47.72	6-methyl-5-hepten-2-one	rs7089	10	7.73	scent_sweet
rs7153	4	86.64	citric acid	rs6254	6	47.72	methyl_salicylate	rs7089	10	7.73	taste_sour
rs7153	4	86.64	glutamic acid	rs6254	6	47.72	hexanal	rs7089	10	7.73	aftertaste_sour
rs2050	5	39.26	methyl_salicylate	rs8092	6	56.64	citric acid	rs7089	10	7.73	aftertaste_salty
rs2050	5	39.26	phenylethyl_alcohol	rs8092	6	56.64	malic acid	rs8434	11	25.94	scent_spicy
rs2050	5	39.26	cis-3-hexenal	rs9061	9	87.48	methyl_salicylate	rs8434	11	25.94	taste_unripe
rs2050	5	39.26	3-methylbutanal	rs9061	9	87.48	guaiacol	rs8434	11	25.94	mouthfeel_solid
rs2050	5	39.26	1-penten-3-one	rs9084	9	94.11	citric acid				
rs2050	5	39.26	beta-ionone	rs9084	9	94.11	malic acid				
rs2050	5	39.26	3-methyl-1-butanol	rs6903	11	86.04	phenylethyl_alcohol				
rs2050	5	39.26	2-methyl-1-butanol	rs6903	11	86.04	benzeneacetaldehyde				
rs2050	5	39.26	2-isobutylthiazole	rs6495	12	26.59	beta-damascenone				

This chapter was published as: *Wang, H., van Eeuwijk, F. A., & Jansen, J. (2016). The potential of probabilistic graphical models in linkage map construction. Theoretical and Applied Genetics, 1-12.*

Chapter 4

The potential of probabilistic graphical models in linkage map construction

Abstract

It has been shown that linkage map construction can be hampered by the presence of genotyping errors and chromosomal rearrangements such as inversions and translocations. Here, we report a novel method for linkage map construction using probabilistic graphical models. The method is proven, both theoretically and practically, to be effective in filtering out markers that contain genotyping errors. In particular, it carries out marker filtering and ordering simultaneously, and is therefore superior to the standard post-hoc filtering using nearest-neighbour stress. Furthermore, we demonstrate empirically that the proposed method offers a promising solution to linkage map construction in the case of a reciprocal translocation.

4.1 Introduction

Genetic maps greatly facilitate a variety of genetic and genomic studies, including the genetic dissection of complex traits, comparative genomic analyses, and genome assembly (Bowers et al. 2012; Liu et al. 2014). Current approaches to map construction are mainly based on estimation of recombination frequency, and they aim to achieve three core objectives: (1) grouping, *i.e.* assigning markers to linkage groups; (2) ordering, *i.e.* finding the correct order of markers within each linkage group; (3) spacing, *i.e.* estimating the map distances between pairs of adjacent markers (Cheema and Dicks 2009; Wu et al. 2008b).

Grouping is usually done by setting a threshold either directly on the pairwise recombination frequencies or on a statistic based on the pairwise recombination frequencies, *e.g.* the LOD scores (Van Os et al. 2005). Ordering can be viewed as an optimization problem. It typically involves two essential elements: (1) a scoring function that quantifies the quality of a given marker order, *e.g.* the likelihood (Cartwright et al. 2007; Jansen et al. 2001), the sum of adjacent recombination frequencies (SARF) (Falk 1989), the sum of adjacent LOD scores (SALOD) (Weeks and Lange 1987), the product of adjacent recombination fractions (PARF) (Wilson 1988) and weighted least squares (WLS) (Stam 1993); (2) a search strategy that reduces the space of candidate marker orders, *e.g.* simulated annealing (Cartwright et al. 2007; Jansen et al. 2001), ant colony optimization (ACO) (Iwata and Ninomiya 2006), genetic algorithms (Gaspin and Schier 1998), evolutionary algorithms (Mester et al. 2003) or greedy and Lin-Kernighan heuristics (Van Os et al. 2005). The optimal marker order is the one that optimizes the scoring function. The map distance is measured in centiMorgan (cM), which is a unit that describes a recombination frequency of 1%. For complete data, spacing is straightforward once ordering is done (Wu et al. 2008b). For incomplete data, multi-point maximum likelihood estimates of recombination frequencies between adjacent markers can be obtained by the EM algorithm using the theory of hidden Markov models (Lander and Green 1987).

It has been recognized that genotyping errors tend to inflate map lengths and reduce the proportion of correctly ordered maps, particularly as marker density increases (Hackett and Broadfoot 2003; Shields et al. 1991). Markers exhibiting high nearest-neighbour stress (N.N.Stress, a quantity measuring the difference between estimated and observed recombination frequencies for the directly neighbouring loci with respect to a particular locus on the map) generally contain genotyping errors (Van Ooijen and Jansen 2013) and are therefore often removed from constructed genetic maps (Farré et al. 2011; Ting et al. 2013). Nonetheless, this post-hoc filtering is inherently biased in terms of predictive validity, as it is applied to marker orders that are obtained under the assumption of no error.

Marker orders have been shown to be relatively robust against both missing data and genotyping errors for widely spaced markers (10cM intervals) (Hackett and Broadfoot 2003). This essentially coincides with the proposition made in Vision et al. (2000). In their study, Vision et al. demonstrated that it is neither necessary nor desirable to genotype all markers in every individual of a large mapping population to get a high-density genetic map. Instead, genotyping a limited number of markers, which are evenly and sparsely distributed throughout the genome, is sufficient for constructing a high-confidence framework map. Afterwards, additional markers can be added to the framework map by certain fine-mapping strategies, so as to avoid the loss in map resolution.

Few methods have been proposed for linkage map construction in the case of reciprocal translocations. A reciprocal translocation refers to an even exchange of DNA fragments between two non-homologous chromosomes. Recombination between loci around the translocation breakpoints is severely suppressed. As a consequence, markers in these regions become ‘pseudo-linked’, *i.e.* markers that lie on different chromosomes involved in the translocation will be mapped onto a single linkage group (Farré et al. 2011).

Probabilistic graphical models (PGMs) combine graph theory and probability theory to give a multivariate statistical modelling framework. A PGM depicts a set of random variable as nodes or vertices in a graph, and encodes the conditional independence between variables through edges in the graph where a lack of an edge between two nodes indicates that the two variables are conditionally independent. Beyond existing successful applications of PGMs in the reconstruction of various biological networks (Airoldi 2007; Friedman 2004), we show here that they can also serve as a map construction method that does not suffer from wrong marker orders as a consequence of genotyping errors and reciprocal translocations. More specifically, we demonstrate both theoretically and empirically that linkage map construction using PGMs can achieve marker filtering and ordering at the same time effectively. Moreover, PGMs allow accurate positioning of the translocation breakpoint and correct ordering of markers on the distal parts of the two chromosomes.

4.2 Materials and Methods

4.2.1 Partial correlation coefficient vs. N.N.Stress in identifying markers having genotyping errors

The partial correlation coefficient provides a measure of conditional independence between variables, which forms the basis for construction of PGMs. Here we demonstrate, theoretically, that the partial correlation coefficient can serve as an alternative to N.N.Stress to identify markers with genotyping errors. To begin with, a few basic concepts are briefly reviewed. The *recombination frequency* θ refers to the probability of observing a gamete with a recombinant haplotype in a single meiosis of a heterozygous parent. In this study, we mainly consider the recombination frequency between marker loci. For each marker, the two parental alleles are denoted by a and b, respectively. The *genotyping error rate* ε is the probability of observing allele a when b is the true allele, or vice versa. An observation on a set of markers is referred to as a phenotype. We investigate the probabilities associated with all possible phenotypes for an ordered triplet of markers M_1-M_2-M_3 (**Table 4.1**). The genotypic frequencies are obtained under two assumptions:

(1) recombination events occurring in adjacent intervals are statistically independent;

(2) the alleles *a* and *b* occur with equal probability (0.5).

For mathematical simplicity, we replace alleles *a* and *b* by the values -1 and 1, respectively. By doing so, the mean and variance of each marker, hereafter considered as a random variable and denoted by $M_{k\ (k=1,2,3)}$, become 0 and 1, respectively. This will greatly facilitate the derivations on (partial) correlation coefficients presented below.

4.2.1.1 Partial correlation coefficient

Under the settings mentioned above, the correlation coefficient between markers M_i and M_j, r_{ij}, is equal to the expectation value $E[M_i \times M_j]$. Let θ_{ij} ($0<\theta_{ij}<0.5$) denote the recombination frequency between markers M_i and M_j; ε_{M1}, ε_{M2} and ε_{M3} denote locus-specific genotyping error rates. When $\varepsilon_{M1} = \varepsilon_{M3} = 0$ and $\varepsilon_{M2} = \varepsilon$ ($0<\varepsilon<0.5$), we obtain

$r_{12} = (1–2\theta_{12})(1–2\varepsilon)$

$r_{23} = (1–2\theta_{23})(1–2\varepsilon)$

$r_{13} = (1–2\theta_{12})(1–2\theta_{23})$

It is obvious that $(1–2\theta_{12})(1–2\varepsilon) < 1–2\theta_{12}$ and $(1–2\theta_{23})(1–2\varepsilon) < 1–2\theta_{23}$. This shows that if M_2 contains errors, r_{12} and r_{23} decrease when ε increases, whereas r_{13} remains unchanged.

The partial correlation coefficient $\rho_{MiMj|Mk}$ measures the correlation between markers M_i and M_j after removing the effect of marker M_k. It can be computed as

$$\rho_{MiMj|Mk} = \frac{r_{ij} - r_{ik} \times r_{jk}}{\sqrt{1-r_{ik}^2}\sqrt{1-r_{jk}^2}}$$

It follows that

$$\rho_{M1M2|M3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}, \quad \rho_{M1M3|M2} = \frac{r_{13} - r_{12} \times r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

We have derived that when $\varepsilon_{M1} = \varepsilon_{M3} = 0$ and $\varepsilon_{M2} = \varepsilon$ ($0<\varepsilon<0.5$), $\rho_{M1M2|M3}$ (and analogously, $\rho_{M2M3|M1}$) is a monotonically decreasing function of ε, whereas $\rho_{M1M3|M2}$ is a monotonically increasing function of ε (please refer to **Supplementary material** for detailed derivation). This indicates that the association between a marker containing genotyping error and each of its flanking markers decreases with increasing error rate, whereas the association between the two flanking markers increases with increasing error rate.

In **Table 4.2** we have summarized the values of r_{12}, r_{23} and r_{13} with respect to eight different settings of ε_{M1}, ε_{M2} and ε_{M3}. Accordingly, we have derived the following relationships:

$$\rho_{M1M3|M2} = \rho_{\tilde{M}1M3|M2} = \rho_{M1\tilde{M}3|M2} = \rho_{\tilde{M}1\tilde{M}3|M2} = 0$$

$$0 < \rho_{\tilde{M}1\tilde{M}3|\tilde{M}2} < \rho_{\tilde{M}1M3|\tilde{M}2} = \rho_{M1\tilde{M}3|\tilde{M}2} < \rho_{M1M3|\tilde{M}2} < 1, \text{ when } \theta_{12} = \theta_{23}$$

$$0 < \rho_{\tilde{M}1\tilde{M}3|\tilde{M}2} < \rho_{\tilde{M}1M3|\tilde{M}2} < \rho_{M1\tilde{M}3|\tilde{M}2} < \rho_{M1M3|\tilde{M}2} < 1, \text{ when } \theta_{12} > \theta_{23}$$

$$0 < \rho_{\tilde{M}1\tilde{M}3|\tilde{M}2} < \rho_{M1\tilde{M}3|\tilde{M}2} < \rho_{\tilde{M}1M3|\tilde{M}2} < \rho_{M1M3|\tilde{M}2} < 1, \text{ when } \theta_{23} > \theta_{12}$$

where M_i and $\tilde{M}_i$ denote a locus genotyped without and with error, respectively.

4.2.1.2 N.N.Stress

Genotyping errors that occur at a marker will increase the observed recombination frequencies between that marker and its flanking markers (Goring and Terwilliger 2000). When $\varepsilon_{M1} = \varepsilon_{M3} = 0$ and $\varepsilon_{M2} = \varepsilon$ ($0<\varepsilon<0.5$),

$$\rho_{12} = \theta_{12}+\varepsilon(1-2\theta_{12})$$

$$\rho_{23} = \theta_{23}+\varepsilon(1-2\theta_{23})$$

$$\rho_{13} = \theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}$$

where ρ_{ij} denote the observed recombination frequency between two markers M_i and M_j. Let d_{ij} denote the distance (in Morgans) between two markers M_i and M_j. Applying Haldane's mapping function, $d_{ij} = -0.5ln(1-2\rho_{ij})$, gives

$$d_{12} = -0.5ln[(1-2\theta_{12})(1-2\varepsilon)]$$

$$d_{23} = -0.5ln[(1-2\theta_{23})(1-2\varepsilon)]$$

$$d_{13} = -0.5ln[(1-2\theta_{12})(1-2\theta_{23})]$$

The N.N.Stress of marker M_2 given M_1 and M_3 is computed as

$$d_{12} + d_{23} - d_{13} = -ln(1-2\varepsilon)$$

Given that $0 < \varepsilon < 0.5$, $-ln(1-2\varepsilon)$ is a monotonically increasing function of ε. This indicates that markers genotyped with high error rate exhibit large N.N.Stress.

Analogously, we have investigated and listed the values of ρ_{12}, ρ_{23} and ρ_{13} with respect to eight different settings of ε_{M1}, ε_{M2} and ε_{M3} in **Table 4.3**. Further, we have derived the relationships below:

$$N.N.Stress_{M2|M1,M3} = N.N.Stress_{M2|\tilde{M}1,M3} = N.N.Stress_{M2|M1,\tilde{M}3} = N.N.Stress_{M2|\tilde{M}1,\tilde{M}3} = 0$$

$$0 < N.N.Stress_{\tilde{M}2|\tilde{M}1,\tilde{M}3} = N.N.Stress_{\tilde{M}2|\tilde{M}1,M3} = N.N.Stress_{\tilde{M}2|M1,\tilde{M}3} = N.N.Stress_{\tilde{M}2|M1,M3} = -ln(1-2\varepsilon)$$

where $N.N.Stress_{Mj|Mi,Mk}$ denote the N.N.Stress of M_j given its flanking markers M_i and M_k.

In view of the similarity between relationships revealed by partial correlation and N.N.Stress, we are able to draw the following conclusions:

1. When the marker data contain no genotyping errors, the partial correlations between physically non-adjacent markers are all equal to 0, whereas the absolute partial correlations between physically adjacent markers are close to 1. It implies that, ideally, marker ordering can be carried out through diagonalization of the partial correlations matrix.
2. In addition to its application to marker ordering, partial correlation coefficient can also serve as an alternative to N.N.Stress, *i.e.* it can be used to identify markers

involving genotyping errors. More specifically, if $\rho_{M1M3|M2}$ is larger than a certain threshold, for the conditioning marker M_2 one of the two situations holds:

(i) M_2 is not genetically located between M_1 and M_3;

(ii) M_2 is indeed between M_1 and M_3, but contains genotyping error (alternatively, the error rate of M_2 is much greater than the error rates of M_1 and M_3).

Notably, a large $\rho_{M1M3|M2}$ always comes with small r_{12} and r_{23}, which indicates, in the context of PGMs, that M_1 and M_3 are, highly likely, directly connected to each other; whereas M_2 is, quite possibly, disconnected from M_1 and M_3. This naturally provides a simultaneous graphical representation of two situations:

(i) the non-intermediate marker M_2 is excluded from the connection between M_1 and M_3;

(ii) the intermediate marker M_2 that involves big genotyping error is excluded from the connection between M_1 and M_3.

3. If not only M_2 but also M_1 or/and M_3 have genotyping errors (alternatively, the error rates of M_2, M_1 or/and M_3 are comparable), the increment of $\rho_{M1M3|M2}$ decreases while $N.N.Stress_{M2|M1,M3}$ does not change. This suggests in the application of partial correlation for identifying markers with genotyping errors, smaller cut-off values are preferable so that minor increases caused by genotyping errors of at least two markers in a triplet can still be captured.
4. When M_2 has no genotyping error, there is no increase of $\rho_{M1M3|M2}$, despite of genotyping errors occurring on either or both of M_1 and M_3. This shows partial correlation is limited to filtering out markers that simultaneously satisfy three requirements:

   (i) they are taken as conditioning variables;
   (ii) they are intermediates in triplets of markers;
   (iii) they have high error rates.

   However, this limitation can be overcome by iterative implementation of partial correlation estimation on sequential triplets of markers. Specifically, assume that M_1-M_2-M_3-M_4 is the true order of four markers, of which M_3 has a high error rate. Then, the problematic marker M_3 can be filtered out by investigating $\rho_{M2M4|M3}$ instead of $\rho_{M1M3|M2}$.

4.2.2 The PC-stable algorithm

In the construction of PGMs, the conditional independence relationships among a set of variables are typically represented in the form of an undirected graph. The PC algorithm (Spirtes et al. 2000) was originally designed to learn a Markov equivalence class of directed acyclic graphs that can be uniquely described as a completed partially directed acyclic graph (CPDAG) (Hauser and Buhlmann 2012). Its learning process consists of two phases: first, construct an undirected graph by means of a series of well-structured conditional independence tests; second, assign directions to certain edges according to the determined v-structures and the acyclic constraint, so that the undirected graph is transformed into a CPDAG.

It should be noted that only the first phase of the PC algorithm is applicable to linkage map construction, since in such a context the directionality of edges between markers is meaningless. However, it has been pointed out that the first phase of the PC algorithm returns order-dependent skeletons (Colombo and Maathuis 2014). That is, the resulting undirected graph is subject to the order of variables present in the input data. For this reason, a modified version of the PC algorithm, which is referred to as the PC-stable algorithm, has been presented to overcome the order-dependent issue (Colombo and Maathuis 2014). The PC-stable algorithm is implemented in the R package *pcalg*.

4.2.3 Frequentist diagonal ordering

In the application of the PC-stable algorithm to linkage mapping, the resulting undirected graphs usually capture the connectivity of markers to a large extent. Nonetheless, the linearity of markers could be a bit ambiguous at certain detailed parts. To eliminate such minor ambiguities, here we've proposed a frequentist diagonal ordering algorithm, which serves as a complement to the PC-stable algorithm for fine-ordering of markers. The logic behind this algorithm is rather straightforward: first, represent the undirected graph achieved by the PC-stable algorithm in the form of an adjacency matrix, which is typically a (0,1)-matrix with entries "1" indicating the corresponding two (row & column) variables are directly connected in the graph; second, restructure the adjacency matrix so that as many "1" entries as possible are located on the first super diagonal of the new adjacency matrix; third, convert the new adjacency matrix into input of a network visualization tool (*e.g.* Cytoscape), and let the relationships between markers be presented graphically. Essentially, this algorithm is to extract a marker string, as long as possible, from the constructed PGM. The related Matlab source code is available at: https://github.com/Huange/Frequentist-diagonal-ordering.

4.2.4 Simulated data

A doubled-haploid population was simulated using the R package *hypred*. Two homozygous parental lines with genotypes *aa* and *bb* at each of 200 loci, which were evenly distributed along a single chromosome of 300cM, were simulated initially. The two parental lines were then crossed to give an F1 population with heterozygous genotype *ab* at each locus. Subsequently, 300 doubled-haploid individuals were simulated from the gametes produced by the F1 generation. No interference was simulated, and so Haldane's mapping function was applicable to the marker data. The markers were numerically labelled from 1 to 200 according to their relative positions along the chromosome. Among them, six markers, 34, 51, 63, 128, 155 and 184, were set to have genotyping errors at rates of 1%, 3%, 5%, 1%, 3% and 5%, respectively.

4.2.5 Cucumber data

This set of marker data was obtained from a RIL population derived from an inter-subspecific cross between the North American processing market type cucumber cultivar Gy14 (C. sativus var. sativus) and the wild accession PI 183967 (C. sativus var. hardwickii) originating from India. The RIL population consisted of 77 F_6-F_8 individuals, each of which was genotyped with 995 SSR markers. For more details see Ren et al. (2009). To deal with missing values in the marker scores, we used a hidden Markov model approach (Jiang and Zeng 1997) implemented in Genstat to estimate the marker genotypes. It appeared interesting to investigate these data because our pre-processing results showed that genotyping errors were widely present across the whole dataset; besides, redundant markers existed in the sense that some markers were located on more or less the same locus.

4.2.6 Barley data

This set of marker data is obtained from DH_1 population developed from a cross between the barley varieties ‘Albacete’ and ‘Barberousse’. ‘Albacete’ is known for containing a reciprocal translocation between chromosomes 1H and 3H. The dataset consisted of 231 lines and 30 markers, of which 13 markers were located on chromosome 1H and 17 markers on chromosome 3H. For more details see Farré et al. (2011).

4.3 Results

4.3.1 Simulated data

By applying the PC-stable algorithm to the simulated marker data, we obtained a linkage map as shown in **Figure 4.1a**. In the map, all the six markers having genotyping errors (*i.e.*, nodes coloured in red) were successfully identified, as they were pulled aside from the linear string formed by the vast majority of all other markers. Meanwhile, a couple of markers without genotyping errors (*i.e.*, nodes coloured in cyan) were also pulled aside from the linear string. This should be attributed to the inherently weak connectivity between those markers and their flanking markers.

For comparison, we also reconstructed a linkage map from the simulated data with JoinMap 4.1. **Table 4.S1** gives the map position and the N.N.Stress of each marker in cM. It shows that the 200 markers were perfectly sequentially ordered, though a few markers possessed high N.N.Stress and thus should be removed from the reconstructed map. **Table 4.4** lists the top six markers with the highest N.N.Stress, which are, as expected, exactly those markers designed with genotyping errors. Furthermore, a minimum spanning tree (MST) was constructed with Genstat from the same dataset (**Figure 4.1b**), since MST has been claimed as another promising tool for efficient and accurate reconstruction of linkage maps (Wu et al. 2008b). Similarly, the 200 markers were substantially linearly arranged in the MST, except that only three markers 51, 63 and 184 were clearly shown in branches, indicating that they should be excluded from the reconstructed linkage map.

4.3.2 Cucumber data

4.3.2.1 Data pre-processing

In a single seed descent (SSD) procedure, the percentage of heterozygotes is halved each generation. In the cucumber data, the proportion of heterozygotes was according to expectation for most individuals, but high for about 10% of the individuals **(Figure 4.S1)**. Considering that the intention of SSD is to make all heterozygotes disappear eventually, we made all heterozygous scores missing and treated the entire population as a RIL_∞ population, *i.e.* a RIL population obtained after infinitely many generations of SSD. This might lead to some individuals coming with a high proportion of missing data. Afterwards, we first excluded markers with more than four (>5.2%) missing data. This concerned 132 markers, leaving 863 markers for further analysis. We then excluded individuals with >10% missing data. These concerned only two individuals, leaving 75 individuals for further analysis. It should be noted that the number of individuals is small for accurate map construction.

4.3.2.2 Forming linkage groups

With a threshold of 0.2 for the recombination frequency, two linkage groups were formed, consisting of 719 and 144 markers, respectively. With a threshold of 0.15, the linkage group consisting of 144 markers remained intact, while the linkage group consisting of 719 markers was split into five subgroups, consisting of 340, 108, 107, 95 and 69 markers, respectively. With a threshold of 0.10, the linkage group consisting of 340 markers was further split into three groups of 177, 162 and 1 markers, respectively. With a threshold of 0.10, also the linkage group consisting of 69 markers was split into three groups of 38, 30 and 1 markers, respectively. Given the estimated six linkage groups obtained with a threshold of 0.15, we used the ML algorithm of JoinMap (Stam 1993) five times to check the stability of the resulting genetic maps. The results indicated that only the linkage group consisting of 340 markers should be split into two groups **(Figure 4.S2)**: the first 177 markers at the upper part of the map (0 ~ 350cM), and the remaining 163 markers at the lower part of the map (370 ~ 750cM). The reason is that although there was a small gap between the two groups, there was no exchange of markers between the two groups in repeated runs of the ML algorithm. In summary, the 863 markers could be divided into seven linkage groups consisting of 177, 163, 144, 108, 107, 95 and 69 markers, respectively. Notably, this grouping was consistent with the one shown by the data providers, who assigned indicators Chr.6, Chr.3, Chr.5, Chr.2, Chr.1, Chr.4 and Chr.7 to the seven linkage groups, respectively. **Table 4.5** offers, for each linkage group, a summary of the total number of markers, the number of unique markers, average map length across five mapping runs, and the highest value of N.N.Stress. The lengths of the preliminary maps constructed for each linkage group were fairly consistent over five mapping runs. Nonetheless, they were always large and especially so for Chr.3 and Chr.6. Also, the highest N.N.Stress is generally quite high. Both phenomena are indicators of genotyping errors in the marker data. Genotyping errors inflate pairwise recombination frequencies between markers (Goring and Terwilliger 2000), and subsequently inflate map lengths and harm the accuracy of marker ordering (Hackett and Broadfoot 2003; Shields et al. 1991).

In this study, we will focus on the map construction for Chr.5, which is an example involving issues of genotyping errors in combination with locally high marker density. The original 144 markers of Chr.5 contained 104 markers, which were unique when accounting for the pattern of missing data alongside with the observed marker phenotypes. After missing data imputation 64 unique markers remained. Hereafter, we will use the imputed data of the 64 markers (but the marker numbers refer to the set of 104 markers) to illustrate our method.

4.3.2.3 Identifying representative markers having genotyping errors for Chr.5

Initially, we focused on a subset of 20 markers that were representative for Chr.5. The 20 markers were obtained as cluster centres of a K-medoids clustering as implemented in the QMKSELECT procedure of Genstat. According to expectation, the cluster centres should either be:

(i) high-quality markers (*i.e.* markers virtually without errors), in which case markers assigned to be a cluster are similar to the cluster centre, with a few more errors;

(ii) low-quality markers (*i.e.* markers with many errors), in which case the cluster is equivalent to its centre. Indeed, we observed that some of the 20 markers only represented themselves, that is, clusters of size 1.

For the 20 markers, we constructed a MST with Genstat and a PGM with the PC-stable algorithm, respectively (**Figure 4.2**; a linearized version of the MST is shown in **Figure 4.S3**). Most links present in the two graphs were consistent, except that markers 44 and 59, 59 and 65 were connected while markers 77 and 80 were disconnected in the PGM. We also constructed a series of linkage maps with JoinMap 4.1 by sequentially deleting the markers with the highest, positive N.N.Stress (**Figure 4.S4**). The deleted markers shown at the top of **Figure 4.S4** were almost identical to the problematic markers revealed in **Figure 4.2**, *i.e.* markers deviating from the linear tree. Notably, N.N.Stress analysis indicated that marker 77 had large genotyping error and thus should be excluded from an accurate linkage map. In this regard, the obtained PGM is considered a bit more precise than the MST, since in the former a string of markers was disconnected from marker 77, whereas the latter did not uncover the error issue underlying marker 77.

4.3.2.4 Constructing framework map for Chr.5

Instead of being restricted to the 20 representative markers, we then investigated the set of 64 unique markers on Chr.5 after missing data imputation. A graphical display of all pairwise recombination frequencies implied that some of the 64 markers were genetically closely or completely coinciding (**Figure 4.S5a**). Results of five independent mapping runs in JoinMap 4.1 further showed that the majority of genetically similar markers were located on the first half of Chr.5 and they led to chaos in the ordering of markers (**Figure 4.S5b**). An MST and a PGM were constructed respectively from the same set of marker data (**Figure 4.3;** a linearized version of the MST is shown in **Figure 4.S6**). The connectivity patterns revealed in the two graphs were generally similar to each other. Specifically, the lower parts of both graphs had roughly vertical linear structures, whereas the upper parts expanded horizontally instead of vertically and there was no obvious clue to the linearity of markers in this region. By further applying frequentist diagonal ordering to the adjacency matrix of the PGM, we obtained the graph shown in **Figure 4.4**. The long string on the left of the graph clearly indicated the linearity of markers at the second half of Chr.5. The short strings at the upper right of the graph were mainly extracted from the nested part of the PGM, *i.e.* the first half of Chr.5. Though each of the short strings revealed, to some extent, the linearity between a couple of markers, as a whole they failed to form a coherent string and thus were not very informative to an accurate map construction. Isolated markers at the lower right of the graph should be excluded from map construction anyway, because of the fact that they occurred either with big genotyping errors or they were genetically very similar to other markers. Again, a series of linkage maps were constructed by sequentially deleting the markers with the highest, positive N.N.Stress (**Figure 4.S7**). Not surprisingly, the deleted markers shown at the top of **Figure 4.S7** overlapped, to a large extent, with those markers excluded from the long string in **Figure 4.4**. It is worth noting that in addition to the first half of Chr.5, a few other problematic markers on the second part of this chromosome, *i.e.* markers 75, 77, 86 and 97, were also unanimously diagnosed by all three approaches.

4.3.3 Barley data - Linkage map construction involving a reciprocal translocation

Figure 4.5 presents the PGM constructed from the barley data by the PC-stable algorithm in combination with frequentist diagonal ordering. Impressively, instead of finding a 'pseudo-linkage' between markers of chromosomes 1H and 3H, as obtained with standard methods, we obtained a cross-like configuration between the given markers. The translocation breakpoint was located around markers 12, 19, 20 that belong to chromosome 1H and markers 1, 22 that belong to chromosome 3H. Moreover, markers on the distal parts of the two chromosomes were perfectly linearly ordered. Our findings were in full agreement with the reference map (**Table 4.S2**) supplied by the data providers.

4.4 Discussion

Our proposed method in principle can be applied to linkage mapping involving large numbers of markers. More generally, whatever the number of markers is, a three-step framework for achieving an accurate genetic map is as follows. First, cut up the set of markers into a number of linkage groups corresponding to the number of a single set of chromosomes. Second, for markers within a single linkage group, whatever the size, use K-medoids clustering to produce a limited set of clusters corresponding to the number of markers required for a framework map for that linkage group. Probably best to define the number of clusters slightly larger than the number of markers required for the framework map, so that it is possible to throw out clusters that are small or consist of isolated markers. Third, take the cluster centres, *i.e.* representative markers, of the larger groups, and start with the construction of PGMs at that point.

We have shown through the barley example that it is possible to simultaneously realize marker grouping and ordering with PGMs, which are constructed through a series of well-structured conditional independence tests, *e.g.* the PC-stable algorithm. Of course, the estimated number of linkage groups is subject to the significance level α adopted in the conditional independence tests. Empirically, smaller values of α tend to lead to sparser graphs (Colombo and Maathuis 2014) that are equivalent to conservative grouping of markers, *i.e.*, more linkage groups of smaller size.

By definition the map distance is measured in cM; 1 cM approximately corresponds to 1% recombination frequency. Constructing PGMs from the observed genotype data involves the calculation of partial correlation coefficients, which essentially measure the combined effect of recombination frequencies between all markers, error rates of markers and marker order. Consequently, once marker grouping and ordering have been achieved using PGMs, one still has to calculate recombination frequencies and genetic distances to obtain a complete genetic map.

Like most existing approaches to linkage map construction, our method is based on the assumption of independent recombination events. In reality, however, chiasma interference (hereafter simply referred to as interference) occurs when the occurrence of one crossover (or chiasma) influences the probability of another crossover occurring nearby, especially in regions of high marker density (Weeks et al. 1994). Assuming no interference simplifies the construction of linkage maps but it leads to considerable overestimation of map distances (Speed and Waterman 1996). In contrast to Haldane's mapping function that is applicable in the absence of interference, Kosambi's mapping function has been invented and empirically verified to well describe the mathematical relation between recombination frequency and map distance in the case of interference. And yet, the performance of PGMs in constructing linkage maps in the face of interference together with data perturbations caused by genotyping errors and reciprocal translocations is currently unclear and deserve further investigation.

A few other studies have also applied graph-theoretic approaches to genetic map construction for plant species (Ronin et al. 2012; Wu et al. 2008a; Yap et al. 2003).

However, they all concentrated on map integration, aiming at producing a consensus genetic map using maps from different populations. We have shown that PGMs present great potential for constructing a reliable genetic map for a single population, by constructing a genetic map in combination with tackling problems that are caused by genotyping errors and reciprocal translocations in the data.

References

Airoldi EM (2007) Getting started in probabilistic graphical models. Plos Comput Biol 3:2421-2425

Bowers JE, Bachlava E, Brunick RL, Rieseberg LH, Knapp SJ, Burke JM (2012) Development of a 10,000 Locus Genetic Map of the Sunflower Genome Based on Multiple Crosses. G3-Genes Genom Genet 2:721-729

Cartwright DA, Troggio M, Velasco R, Gutin A (2007) Genetic mapping in the presence of genotyping errors. Genetics 176:2521-2527

Cheema J, Dicks J (2009) Computational approaches and software tools for genetic linkage map estimation in plants. Brief Bioinform 10:595-608

Colombo D, Maathuis MH (2014) Order-Independent Constraint-Based Causal Structure Learning. J Mach Learn Res 15:3741-3782

Falk CT (1989) A Simple Scheme for Preliminary Ordering of Multiple Loci - Application to 45 Cf Families. Prog Clin Biol Res 329:17-22

Farré A, Benito IL, Cistue L, de Jong JH, Romagosa I, Jansen J (2011) Linkage map construction involving a reciprocal translocation. Theor Appl Genet 122:1029-1037

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799-805

Gaspin C, Schier T (1998) Genetic algorithms for genetic mapping. Lect Notes Comput Sc 1363:145-155

Goring HHH, Terwilliger JD (2000) Linkage analysis in the presence of errors II: Marker-locus genotyping errors modeled with hypercomplex recombination fractions. Am J Hum Genet 66:1107-1118

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90:33-38

Hauser A, Buhlmann P (2012) Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. J Mach Learn Res 13:2409-2464

Iwata H, Ninomiya S (2006) AntMap: Constructing genetic linkage maps using an ant colony optimization algorithm. Breeding Sci 56:371-377

Jansen J, de Jong AG, van Ooijen JW (2001) Constructing dense genetic linkage maps. Theor Appl Genet 102:1113-1122

Jiang C, Zeng ZB (1997) Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101:47-58

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84:2363-2367

Liu DY, Ma CX, Hong WG, Huang L, Liu M, Liu H, Zeng HP, Deng DJ, Xin HG, Song J, Xu CH, Sun XW, Hou XL, Wang XW, Zheng HK (2014) Construction and Analysis of High-Density Linkage Map Using High-Throughput Sequencing Data. PloS one 9

Mester D, Ronin Y, Minkov D, Nevo E, Korol A (2003) Constructing large-scale genetic maps using an evolutionary strategy algorithm. Genetics 165:2269-2282

Ren Y, Zhang Z, Liu J, Staub JE, Han Y, Cheng Z, Li X, Lu J, Miao H, Kang H, Xie B, Gu X, Wang X, Du Y, Jin W, Huang S (2009) An integrated genetic and cytogenetic map of the cucumber genome. PloS one 4:e5795

Ronin Y, Mester D, Minkov D, Belotserkovski R, Jackson BN, Schnable PS, Aluru S, Korol A (2012) Two-phase analysis in consensus genetic mapping. G3 2:537-549

Shields DC, Collins A, Buetow KH, Morton NE (1991) Error Filtration, Interference, and the Human Linkage Map. P Natl Acad Sci USA 88:6501-6505

Speed TP, Waterman MS (1996) Genetic mapping and DNA sequencing. Springer, New York

Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge, Mass.

Stam P (1993) Construction of Integrated Genetic-Linkage Maps by Means of a New Computer Package - Joinmap. Plant J 3:739-744

Ting NC, Jansen J, Nagappan J, Ishak Z, Chin CW, Tan SG, Cheah SC, Singh R (2013) Identification of QTLs Associated with Callogenesis and Embryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers. PloS one 8

Van Ooijen JW, Jansen J (2013) Genetic Mapping in Experimental Populations. Cambridge University Press

Van Os H, Stam P, Visser RG, Van Eck HJ (2005) RECORD: a novel method for ordering loci on a genetic linkage map. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 112:30-40

Vision TJ, Brown DG, Shmoys DB, Durrett RT, Tanksley SD (2000) Selective mapping: A strategy for optimizing the construction of high-density linkage maps. Genetics 155:407-420

Weeks DE, Lange K (1987) Preliminary ranking procedures for multilocus ordering. Genomics 1:236-242

Weeks DE, Ott J, Lathrop GM (1994) Detection of genetic interference: simulation studies and mouse data. Genetics 136:1217-1226

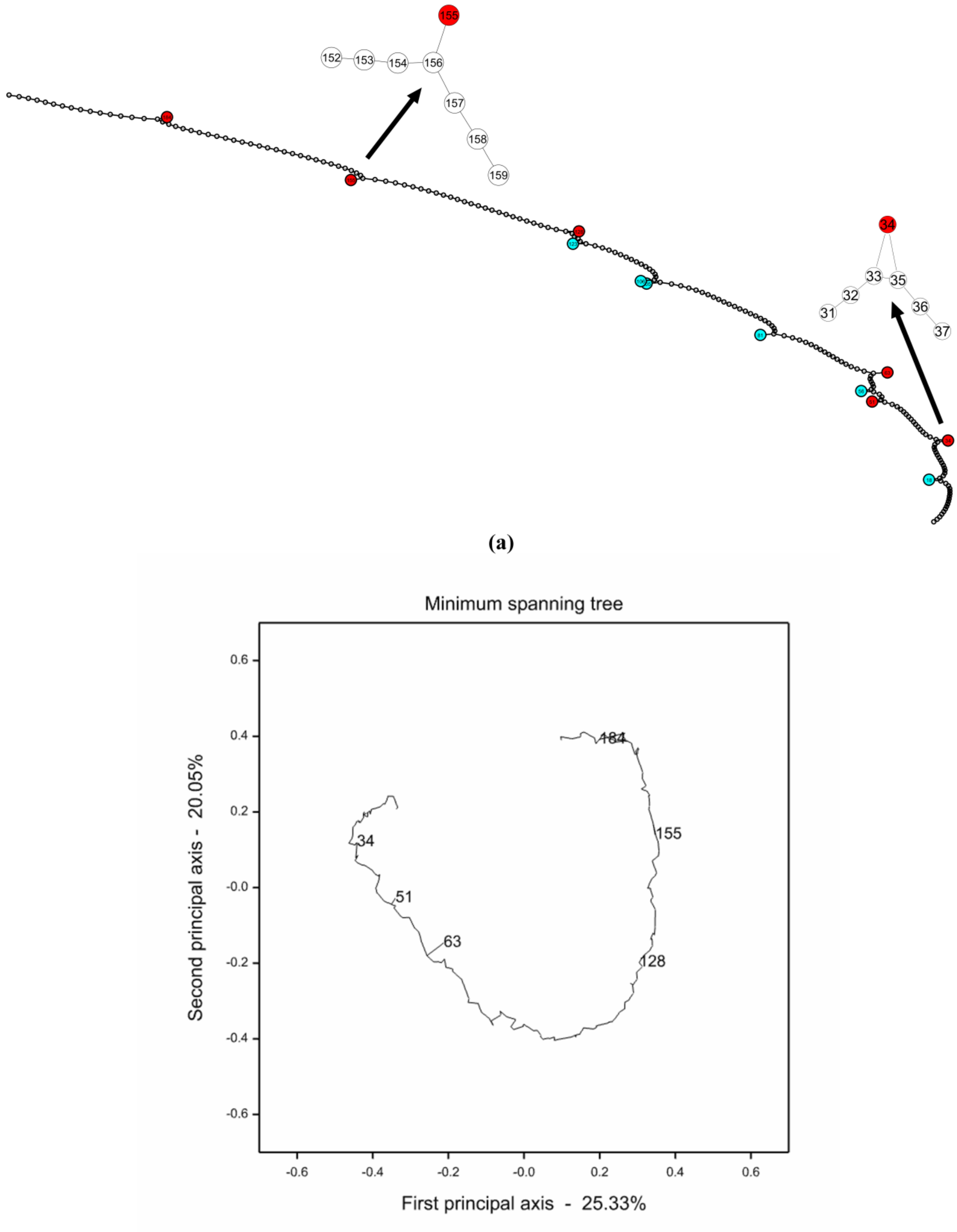
Wilson SR (1988) A major simplification in the preliminary ordering of linked loci. Genetic epidemiology 5:75-80

Wu Y, Close TJ, Lonardi S (2008a) On the accurate construction of consensus genetic maps. Computational systems bioinformatics / Life Sciences Society Computational Systems Bioinformatics Conference 7:285-296

Wu YH, Bhat PR, Close TJ, Lonardi S (2008b) Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. Plos Genet 4

Yap IV, Schneider D, Kleinberg J, Matthews D, Cartinhour S, McCouch SR (2003) A graph-theoretic approach to comparing and integrating genetic, physical and sequence-based maps. Genetics 165:2235-2247

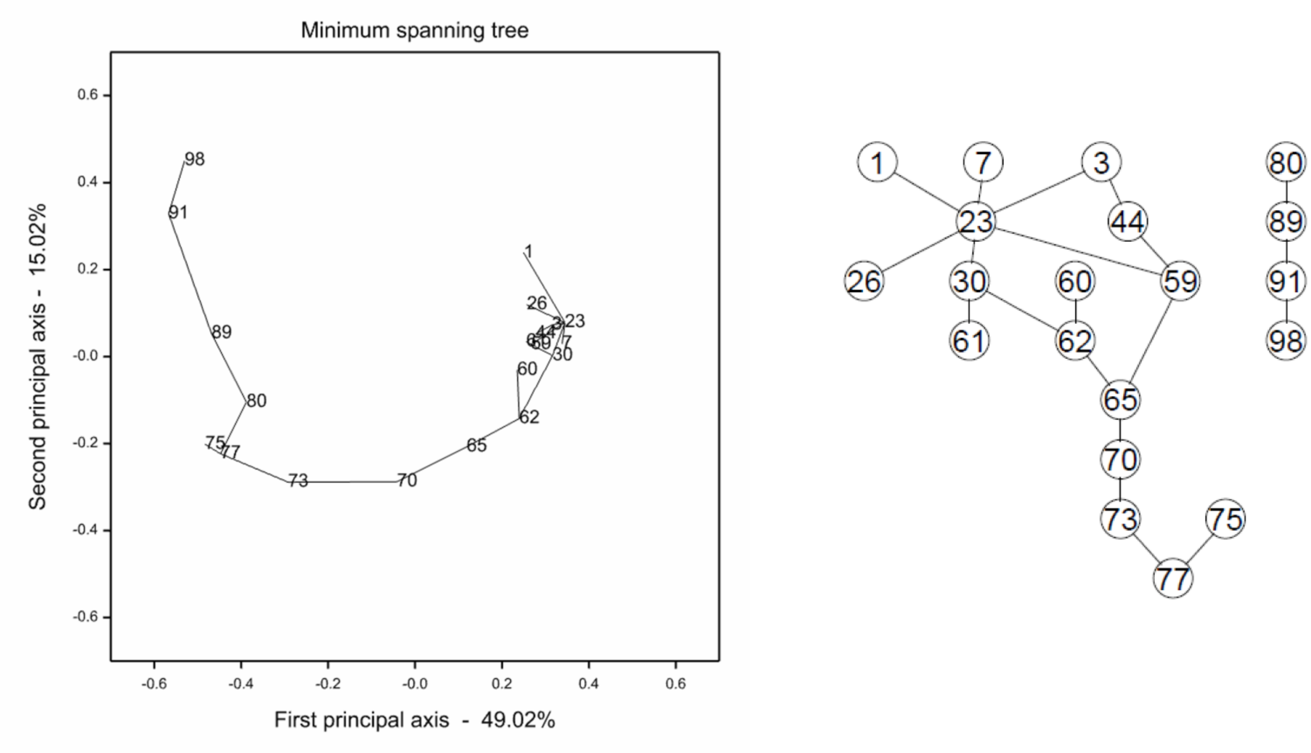
Figure 4.1 (a) A PGM constructed with the PC-stable algorithm for the simulated data. The six markers designed with genotyping errors are pulled aside from the linear string and coloured in red, another six markers pulled aside from the linear string are coloured in cyan. Enlargements of two detailed parts of the PGM are given above the linear string, though the whole graph itself can be enlarged dramatically to show all details clearly. **(b)** A MST constructed with Genstat for the simulated data. The diagram was projected on the first two principal axes obtained by a principal coordinate analysis. Only the six markers designed with genotyping errors are marked out and coloured in red.



(a)

(b)

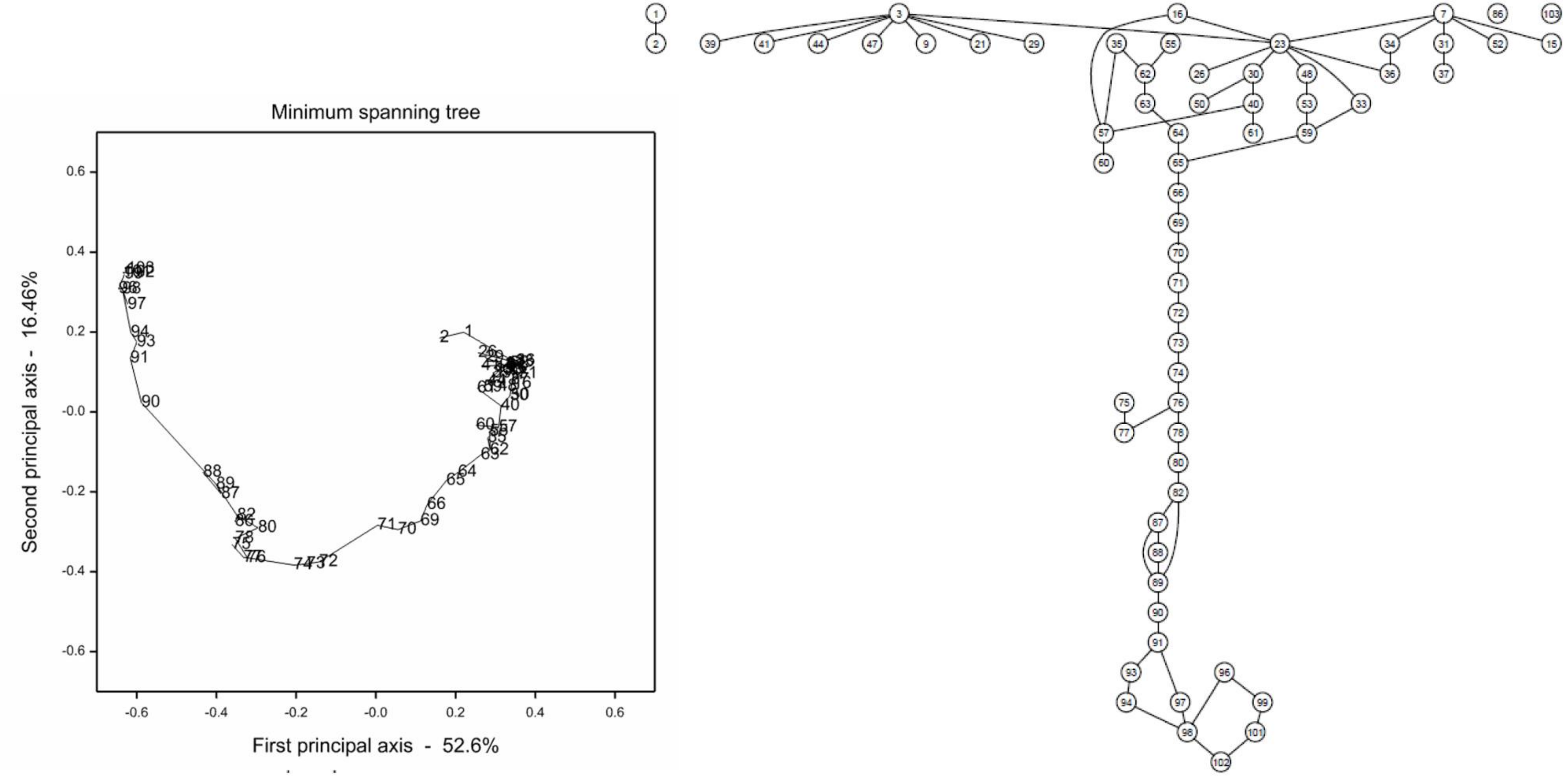
Figure 4.2 (a) A MST constructed with Genstat for 20 representative markers of Chr.5. The diagram was projected on the first two principal axes obtained by a principal coordinate analysis. **(b)** A PGM constructed with the PC-stable algorithm for the same set of 20 markers. The significance level for conditional independence tests was set at 0.05.



(a)

(b)

Figure 4.3 (a) A MST constructed with Genstat for 64 unique markers of Chr.5. The diagram was projected on the first two principal axes obtained by a principal coordinate analysis. **(b)** A PGM constructed with the PC-stable algorithm for the same set of 64 markers. The significance level for conditional independence tests was set at 0.05.



(a) (b)

Figure 4.4 An adjusted PGM obtained by further applying frequentist diagonal ordering to the adjacency matrix of the PGM shown in Figure 4.3b.

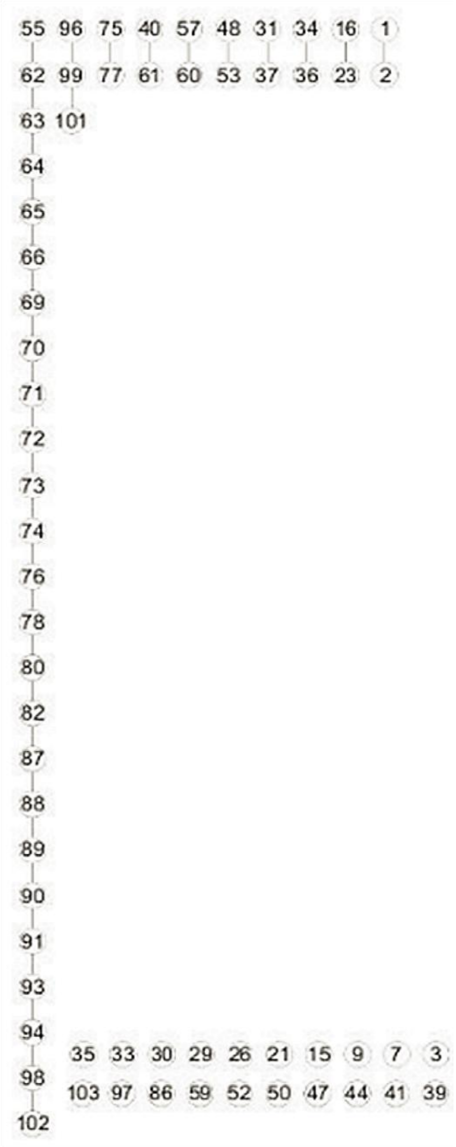


Figure 4.5 A PGM constructed from the barley data by the PC-stable algorithm in combination with frequentist diagonal ordering. Yellow nodes stand for markers on chromosome 1H and green nodes stand for markers on chromosome 3H. The significance level for conditional independence tests was set at 0.05.

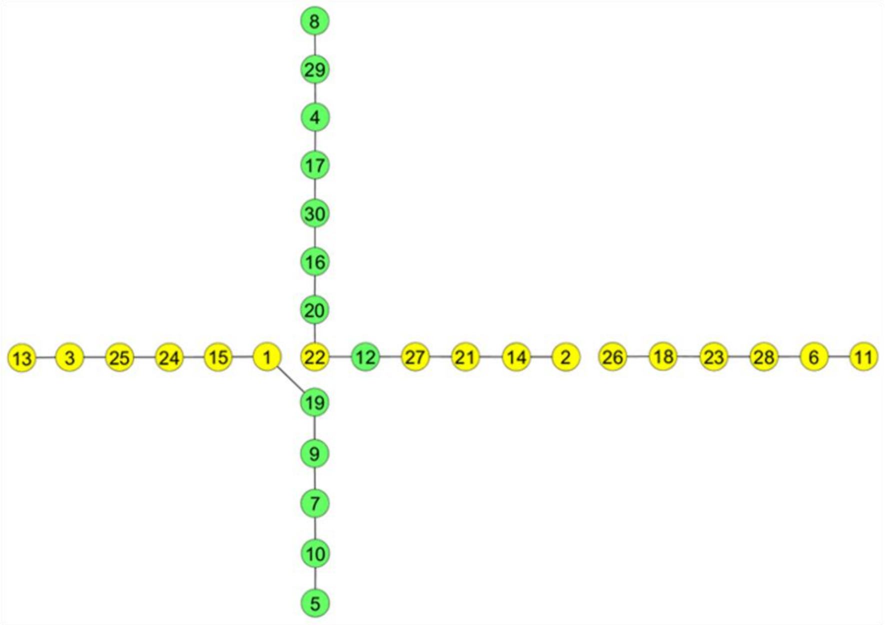


Figure 4.S1 The proportion of heterozygous scores for individuals in the cucumber data set.

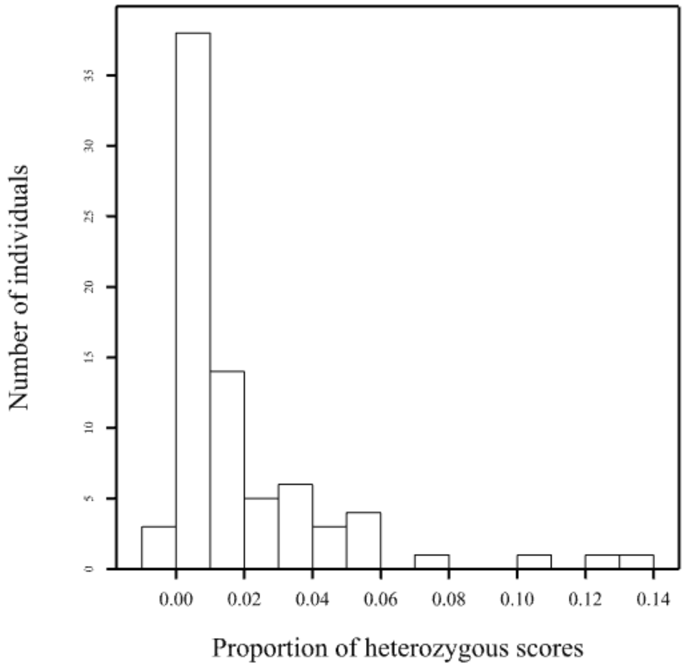


Figure 4.S2 Five genetic maps generated repeatedly by the ML algorithm of JoinMap for a linkage group consisting of 340 markers that was obtained with a threshold of 0.15 for the recombination frequency. According to the consistent (small) gap at approx. 360 cM, this linkage group could be split into two linkage groups of 177 and 163 markers, respectively.

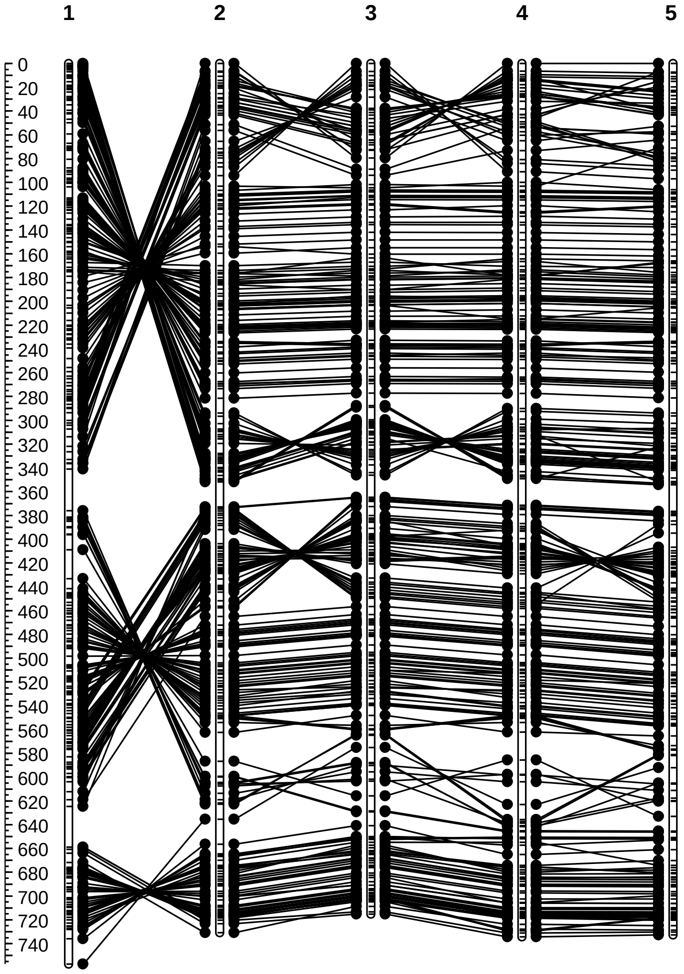


Figure 4.S3 Linearized MST for 20 representative markers of Chr.5. The numbers between connected markers represent the number of recombinations and simple matching coefficient of similarity, respectively.

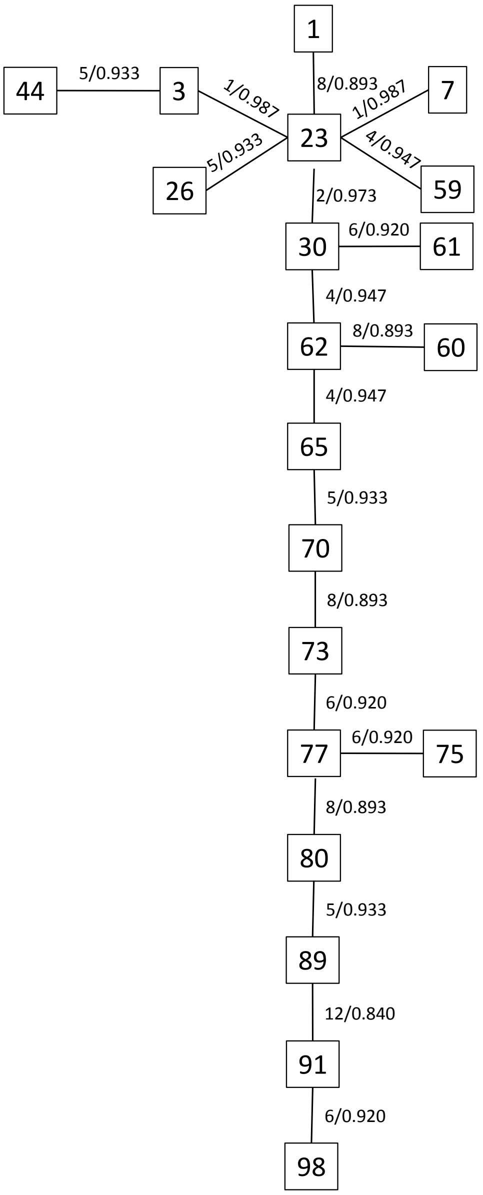


Figure 4.S4 Linkage maps of Chr.5 obtained by sequentially deleting markers with the highest, positive N.N.Stress from the set of 20 representative markers. The deleted markers are shown above the linkage maps; the associated N.N.Stress is given between brackets. Comparison of the last two maps indicates regions (shown in red) where markers have been deleted and the associated reductions in map length.

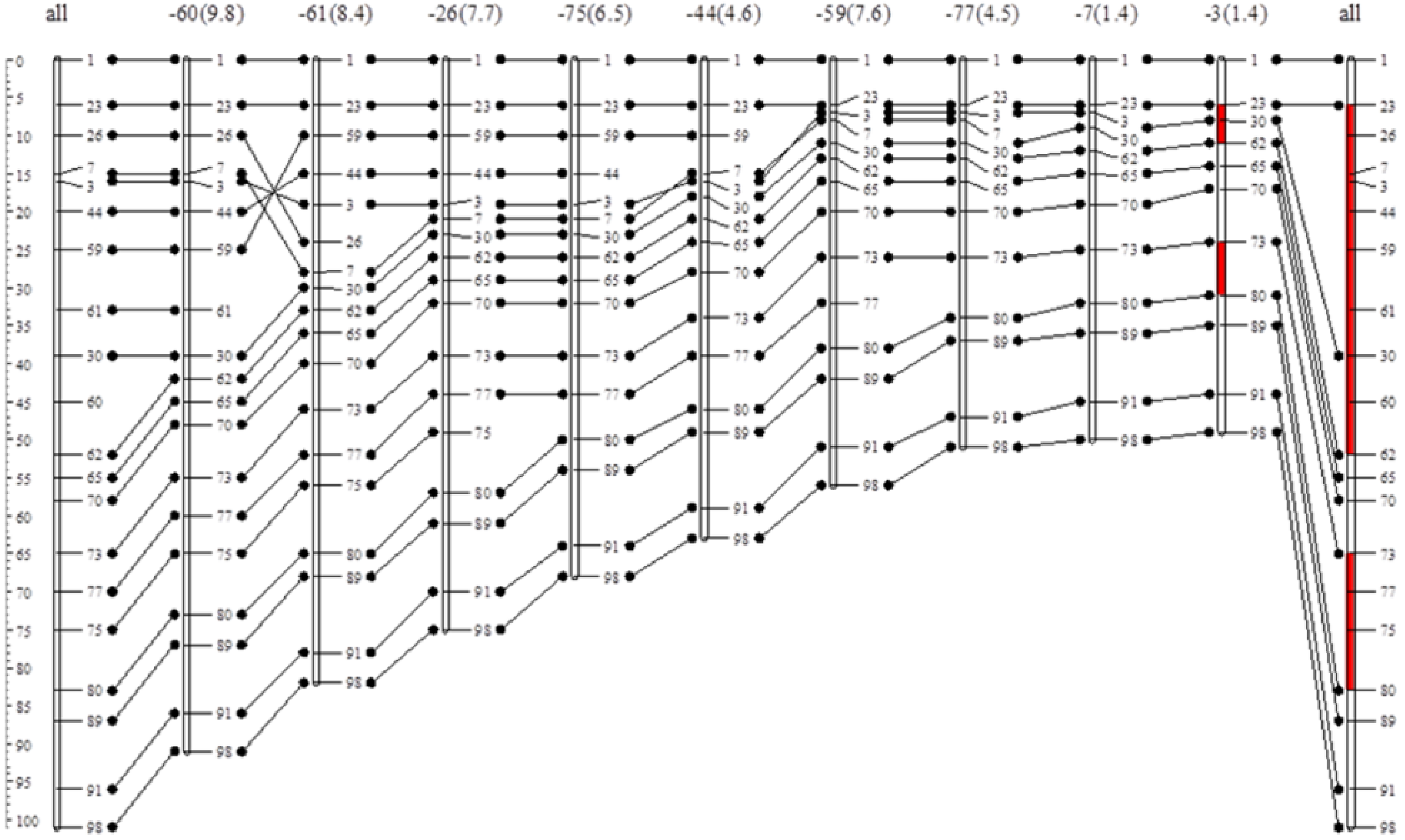
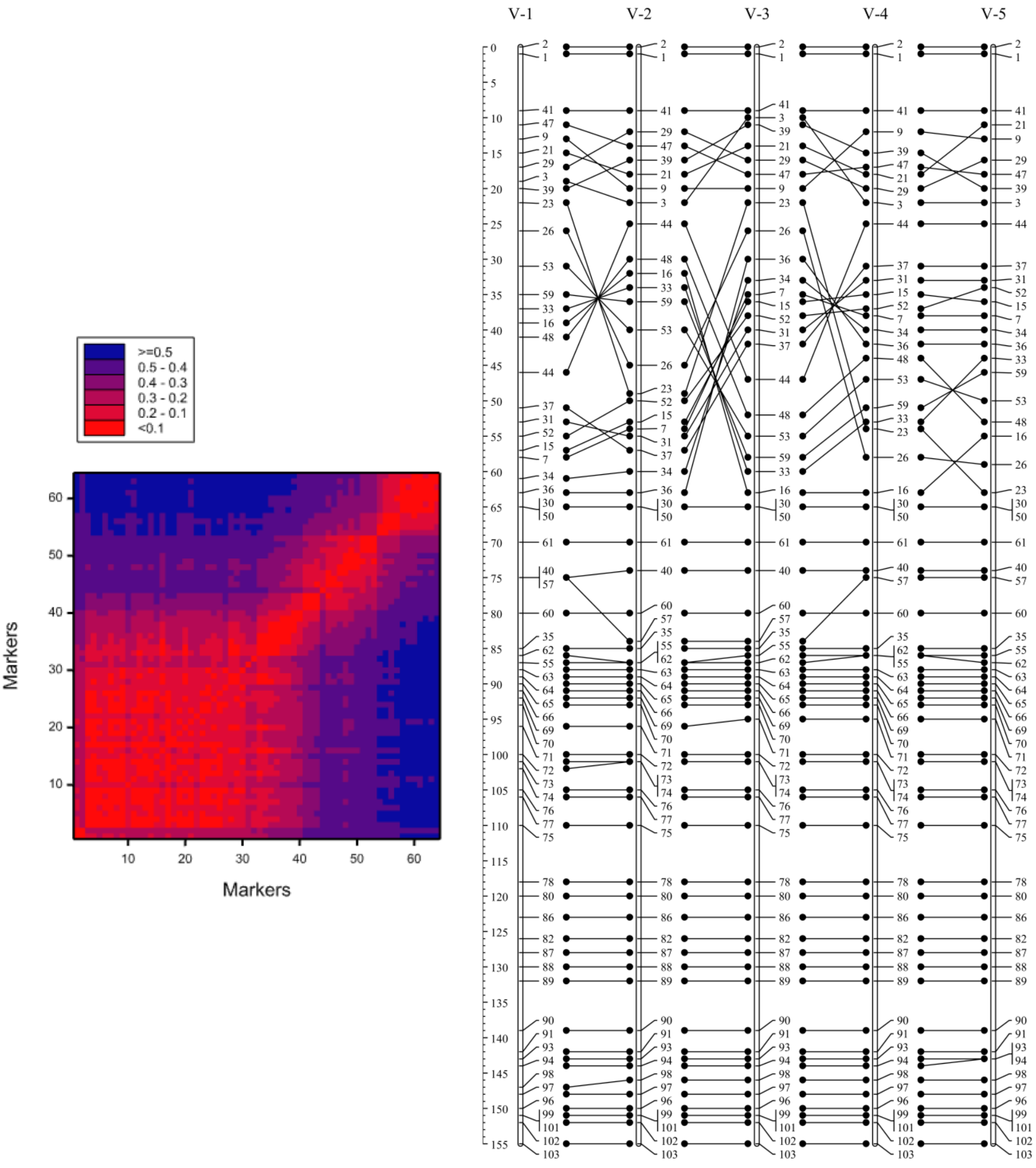


Figure 4.S5 (a) Pairwise recombination frequencies estimated for 64 unique markers of Chr.5. Markers are sorted according to their numerical labels. **(b)** Five linkage maps obtained by independent mapping runs in JoinMap 4.1 for the 64 markers.



(a) (b)

Figure 4.S6 Linearized MST for 64 unique markers of Chr.5.

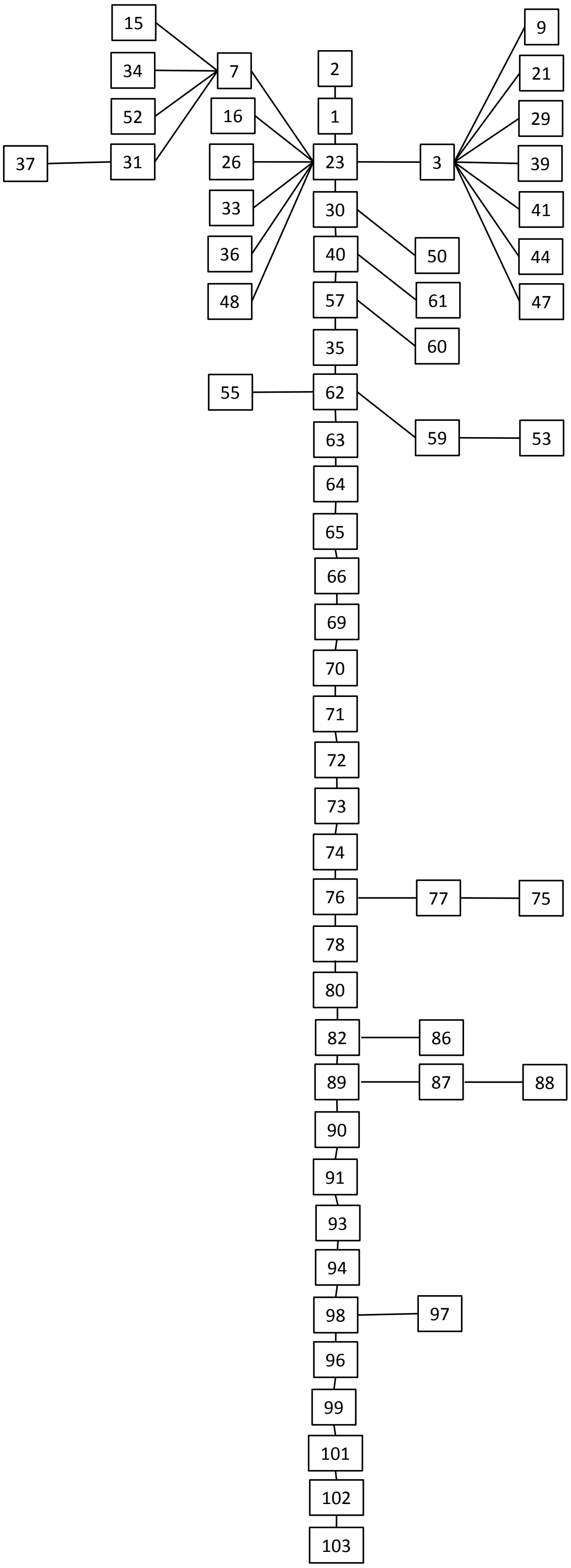


Figure 4.S7 Linkage maps obtained by sequentially deleting the unique markers with the highest, positive N.N.Stress on Chr.5. The deleted markers are shown above the linkage maps.

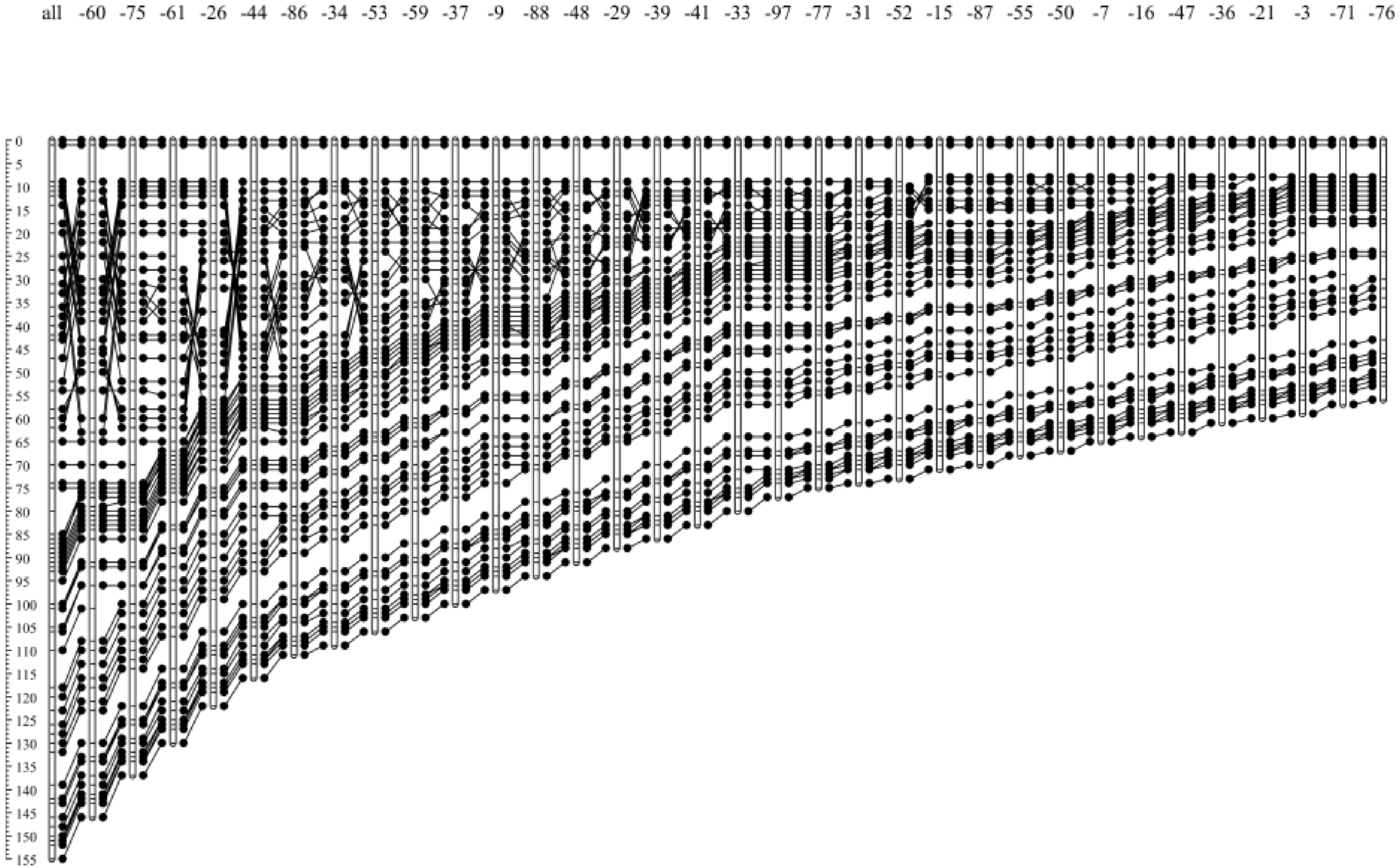


Table 4.1 Genotypic frequencies for ordered triplet of markers M_1-M_2-M_3. θ_{ij} ($0<\theta_{ij}<0.5$) denote the recombination frequency between markers M_i and M_j; ε_{M1}, ε_{M2} and ε_{M3} denote locus-specific genotyping error rates, $0<\varepsilon<0.5$. Numeric values -1 and 1 in the first three columns represent marker types *a* and *b*, respectively.

Marker type			Genotypic frequency
M1	*M2*	*M3*	($\varepsilon_{M1} = \varepsilon_{M3} = 0$, $\varepsilon_{M2} = \varepsilon$)
-1	*-1*	*-1*	$0.5\times[(1-\varepsilon)(1-\theta_{12})(1-\theta_{23})+\varepsilon\theta_{12}\theta_{23}]$
-1	*-1*	*1*	$0.5\times(1-\varepsilon)(1-\theta_{12})\theta_{23}+\varepsilon\theta_{12}(1-\theta_{23})$
-1	*1*	*1*	$0.5\times\varepsilon(1-\theta_{12})\theta_{23}+(1-\varepsilon)\theta_{12}(1-\theta_{23})$
-1	*1*	*-1*	$0.5\times[\varepsilon(1-\theta_{12})(1-\theta_{23})+(1-\varepsilon)\theta_{12}\theta_{23}]$
1	*1*	*1*	$0.5\times[(1-\varepsilon)(1-\theta_{12})(1-\theta_{23})+\varepsilon\theta_{12}\theta_{23}]$
1	*1*	*-1*	$0.5\times(1-\varepsilon)(1-\theta_{12})\theta_{23}+\varepsilon\theta_{12}(1-\theta_{23})$
1	*-1*	*-1*	$0.5\times\varepsilon(1-\theta_{12})\theta_{23}+(1-\varepsilon)\theta_{12}(1-\theta_{23})$
1	*-1*	*1*	$0.5\times[\varepsilon(1-\theta_{12})(1-\theta_{23})+(1-\varepsilon)\theta_{12}\theta_{23}]$

Table 4.2 Pairwise correlation coefficients for ordered triplet of markers M_1-M_2-M_3. Denotations of θ_{12}, θ_{23}, ε_{M1}, ε_{M2}, ε_{M3} and ε are identical to those in Table 4.1.

	r_{12}	r_{23}	r_{13}
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = 0$	$1-2\theta_{12}$	$1-2\theta_{23}$	$(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = 0$	$(1-2\theta_{12})(1-2\varepsilon)$	$1-2\theta_{23}$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} =\varepsilon$	$1-2\theta_{12}$	$(1-2\theta_{23})(1-2\varepsilon)$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = \varepsilon$	$(1-2\theta_{12})(1-2\varepsilon)$	$(1-2\theta_{23})(1-2\varepsilon)$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)^2$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = 0$	$(1-2\theta_{12})(1-2\varepsilon)$	$(1-2\theta_{23})(1-2\varepsilon)$	$(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = 0$	$(1-2\theta_{12})(1-2\varepsilon)^2$	$(1-2\theta_{23})(1-2\varepsilon)$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} =\varepsilon$	$(1-2\theta_{12})(1-2\varepsilon)$	$(1-2\theta_{23})(1-2\varepsilon)^2$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = \varepsilon$	$(1-2\theta_{12})(1-2\varepsilon)^2$	$(1-2\theta_{23})(1-2\varepsilon)^2$	$(1-2\theta_{12})(1-2\theta_{23})(1-2\varepsilon)^2$

Table 4.3 The observed pairwise recombination frequencies for ordered triplet of markers M_1-M_2-M_3. Denotations of θ_{12}, θ_{23}, ε_{M1}, ε_{M2}, ε_{M3} and ε are identical to those in Table 4.1.

	ρ_{12}	ρ_{23}	ρ_{13}
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = 0$	θ_{12}	θ_{23}	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = 0$	$\theta_{12}+\varepsilon(1-2\theta_{12})$	θ_{23}	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+\varepsilon(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = \varepsilon$	θ_{12}	$\theta_{23}+\varepsilon(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+\varepsilon(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = 0$, $\varepsilon_{M3} = \varepsilon$	$\theta_{12}+\varepsilon(1-2\theta_{12})$	$\theta_{23}+\varepsilon(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+2\varepsilon(1-\varepsilon)(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = 0$	$\theta_{12}+\varepsilon(1-2\theta_{12})$	$\theta_{23}+\varepsilon(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = 0$	$\theta_{12}+2\varepsilon(1-\varepsilon)(1-2\theta_{12})$	$\theta_{23}+\varepsilon(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+\varepsilon(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = 0$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = \varepsilon$	$\theta_{12}+\varepsilon(1-2\theta_{12})$	$\theta_{23}+2\varepsilon(1-\varepsilon)(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+\varepsilon(1-2\theta_{12})(1-2\theta_{23})$
$\varepsilon_{M1} = \varepsilon$, $\varepsilon_{M2} = \varepsilon$, $\varepsilon_{M3} = \varepsilon$	$\theta_{12}+2\varepsilon(1-\varepsilon)(1-2\theta_{12})$	$\theta_{23}+2\varepsilon(1-\varepsilon)(1-2\theta_{23})$	$\theta_{12}+\theta_{23}-2\theta_{12}\theta_{23}+2\varepsilon(1-\varepsilon)(1-2\theta_{12})(1-2\theta_{23})$

Table 4.4 The top six markers with the highest N.N.Stress obtained by JoinMap 4.1 from the simulated marker data.

	Locus	Position	N.N. Stress (cM)
1	marker63	96.145	11.01
2	marker184	296.883	10.939
3	marker155	243.352	6.318
4	marker51	70.273	5.629
5	marker128	195.223	2.046
6	marker34	43.34	2.041

Table 4.5 A summary of the total number of markers, the number of unique markers, the average map length across five mapping runs, and the highest values of N.N.Stress for each of the seven linkage groups constructed from the cucumber data (before missing data imputation).

Linkage group	Number of markers	Number of unique markers	Average map length in 5 runs (cM)	Highest N.N.Stress (cM)
Chr.1	107	103	195.1	8.0
Chr.2	108	103	269.8	11.7
Chr.3	163	151	343.9	22.3
Chr.4	95	67	115.4	11.8
Chr.5	144	104	155.0	9.1
Chr.6	177	157	333.1	12.0
Chr.7	69	66	176.7	9.5
Total	863	751		

Table 4.S1 The map position and the N.N.Stress of each marker obtained with JoinMap 4.1 from the simulated marker data.

Nr	Locus	Position (cM)	N.N. Stress (cM)
1	marker1	0	
2	marker2	1.01	-0.028
3	marker3	2.362	-0.047
4	marker4	4.057	-0.023
5	marker5	4.728	-0.014
6	marker6	5.738	-0.028
7	marker7	7.09	-0.057
8	marker8	9.131	-0.057
9	marker9	10.482	-0.009
10	marker10	10.817	-0.012
11	marker11	12.512	-0.059
12	marker12	14.207	-0.035
13	marker13	15.217	-0.014
14	marker14	15.888	-0.014
15	marker15	16.898	-0.028
16	marker16	18.25	-0.047
17	marker17	19.945	-0.023
18	marker18	20.616	0.671
19	marker19	21.287	-0.028
20	marker20	23.328	-0.072
21	marker21	25.023	-0.047
22	marker22	26.375	-0.047
23	marker23	28.07	-0.023
24	marker24	28.741	-0.023
25	marker25	30.436	-0.023
26	marker26	31.107	-0.023
27	marker27	32.802	-0.023
28	marker28	33.473	-0.019
29	marker29	34.825	-0.077
30	marker30	37.565	-0.019
31	marker31	37.9	-0.007
32	marker32	38.91	-0.05
33	marker33	41.299	-0.102
34	marker34	43.34	2.041
35	marker35	45.73	-0.102
36	marker36	47.771	-0.014
37	marker37	48.105	-0.012
38	marker38	49.8	-0.047
39	marker39	51.152	-0.047
40	marker40	52.847	-0.023
41	marker41	53.518	-0.023
42	marker42	55.213	-0.035

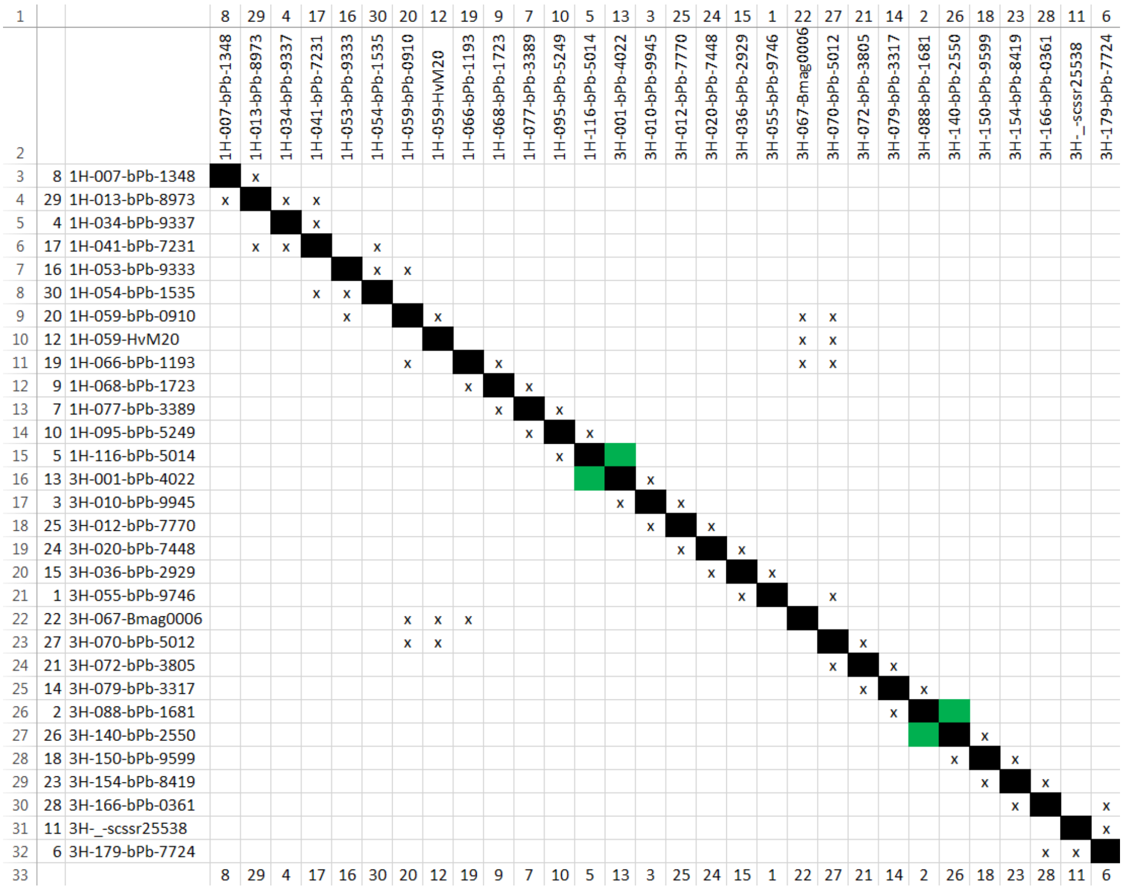
43	marker43	56.223	-0.035
44	marker44	57.918	-0.023
45	marker45	58.59	-0.019
46	marker46	59.941	-0.028
47	marker47	60.951	-0.028
48	marker48	62.303	-0.038
49	marker49	63.654	-0.038
50	marker50	65.005	-0.152
51	marker51	70.273	5.629
52	marker52	74.443	-0.059
53	marker53	75.114	-0.009
54	marker54	75.785	-0.033
55	marker55	78.174	-0.033
56	marker56	78.846	0.671
57	marker57	80.887	-0.057
58	marker58	82.238	-0.047
59	marker59	83.933	-0.047
60	marker60	85.285	-0.019
61	marker61	85.956	-0.009
62	marker62	86.627	-0.142
63	marker63	96.145	11.01
64	marker64	102.536	-0.139
65	marker65	103.546	0.666
66	marker66	104.898	-0.009
67	marker67	105.232	-0.005
68	marker68	105.903	-0.014
69	marker69	106.914	-0.028
70	marker70	108.265	-0.047
71	marker71	109.96	-0.047
72	marker72	111.312	-0.077
73	marker73	114.052	-0.118
74	marker74	116.093	-0.043
75	marker75	117.103	-0.035
76	marker76	118.798	-0.072
77	marker77	120.839	-0.087
78	marker78	122.88	0.642
79	marker79	124.576	-0.023
80	marker80	125.247	-0.019
82	marker82	126.598	-0.009
81	marker81	126.933	0.688
83	marker83	130.026	-0.043
84	marker84	130.697	-0.019
85	marker85	132.049	-0.047
86	marker86	133.744	-0.023
87	marker87	134.415	-0.023
88	marker88	136.11	-0.035

89	marker89	137.12	-0.043
90	marker90	139.161	-0.043
91	marker91	140.172	-0.043
92	marker92	142.213	-0.043
93	marker93	143.223	-0.021
94	marker94	144.233	-0.043
95	marker95	146.274	-0.028
96	marker96	146.945	-0.019
97	marker97	148.297	-0.057
98	marker98	150.338	-0.072
99	marker99	152.033	-0.035
100	marker100	153.043	-0.05
101	marker101	155.432	0.597
102	marker102	158.173	-0.159
103	marker103	160.913	-0.159
104	marker104	163.654	-0.058
105	marker105	164.664	-0.021
106	marker106	165.674	-0.014
107	marker107	166.345	-0.043
108	marker108	169.439	-0.228
109	marker109	172.889	-0.048
110	marker110	173.56	-0.009
111	marker111	174.231	-0.023
112	marker112	175.926	-0.012
113	marker113	176.26	-0.007
114	marker114	177.271	-0.028
115	marker115	178.622	-0.019
116	marker116	179.293	-0.014
117	marker117	180.303	-0.014
118	marker118	180.974	-0.023
119	marker119	182.669	-0.023
120	marker120	183.341	-0.009
121	marker121	184.012	-0.033
122	marker122	186.401	-0.05
123	marker123	187.411	0.671
124	marker124	188.083	-0.038
125	marker125	190.823	-0.058
126	marker126	191.833	-0.035
127	marker127	193.528	-0.059
128	marker128	195.223	2.046
129	marker129	197.964	-0.118
130	marker130	200.005	-0.043
131	marker131	201.015	-0.014
132	marker132	201.686	-0.028
133	marker133	203.727	-0.043
134	marker134	204.737	-0.021

135	marker135	205.747	0.662
136	marker136	207.789	-0.043
137	marker137	208.799	-0.05
138	marker138	211.188	-0.067
139	marker139	212.54	-0.028
140	marker140	213.55	-0.014
141	marker141	214.221	-0.038
142	marker142	216.961	-0.019
143	marker143	217.296	-0.005
144	marker144	217.967	-0.023
145	marker145	219.662	-0.137
146	marker146	223.47	-0.109
147	marker147	224.822	-0.038
148	marker148	226.173	-0.038
149	marker149	227.524	0.647
150	marker150	229.914	0.584
151	marker151	233.008	-0.11
152	marker152	234.703	-0.097
153	marker153	237.443	-0.058
154	marker154	238.453	-0.105
155	marker155	243.352	6.318
156	marker156	247.16	0.51
157	marker157	250.254	-0.065
158	marker158	251.264	-0.021
159	marker159	252.274	-0.021
160	marker160	253.285	-0.089
161	marker161	257.454	-0.119
162	marker162	258.805	-0.028
163	marker163	259.815	-0.028
164	marker164	261.167	-0.057
165	marker165	263.208	-0.087
166	marker166	265.249	-0.072
167	marker167	266.944	-0.047
168	marker168	268.295	-0.047
169	marker169	269.99	-0.084
170	marker170	272.38	-0.067
171	marker171	273.731	-0.038
172	marker172	275.083	-0.009
173	marker173	275.417	-0.009
174	marker174	276.769	0.666
175	marker175	277.779	-0.043
176	marker176	279.82	-0.057
177	marker177	281.171	-0.019
178	marker178	281.843	-0.019
179	marker179	283.194	-0.038
180	marker180	284.545	-0.057

181	marker181	286.587	0.642
182	marker182	288.282	-0.023
183	marker183	288.953	-0.116
184	marker184	296.883	10.939
185	marker185	304.424	-0.281
186	marker186	306.119	-0.084
187	marker187	308.509	-0.102
188	marker188	310.55	-0.028
189	marker189	311.221	-0.048
190	marker190	314.671	0.612
191	marker191	316.366	-0.097
192	marker192	319.106	-0.159
193	marker193	321.847	-0.138
194	marker194	324.236	-0.213
195	marker195	328.405	0.621
196	marker196	329.757	-0.038
197	marker197	331.108	-0.009
198	marker198	331.442	-0.019
199	marker199	334.183	0.612
200	marker200	336.224	

Table 4.S2 The neighbourhood obtained by the lasso approach for the 30 markers on chromosomes 1H and 3H of the barley data. An ‘x’ symbol indicates that the corresponding row marker is strongly linked to the column marker. Markers are ordered so that both chromosomes are linearly structured, and meanwhile the translocation breakpoint is clearly displayed off-diagonal.



1			8	29	4	17	16	30	20	12	19	9	7	10	5	13	3	25	24	15	1	22	27	21	14	2	26	18	23	28	11	6
2			1H-007-bPb-1348	1H-013-bPb-8973	1H-034-bPb-9337	1H-041-bPb-7231	1H-053-bPb-9333	1H-054-bPb-1535	1H-059-bPb-0910	1H-059-HvM20	1H-066-bPb-1193	1H-068-bPb-1723	1H-077-bPb-3389	1H-095-bPb-5249	1H-116-bPb-5014	3H-001-bPb-4022	3H-010-bPb-9945	3H-012-bPb-7770	3H-020-bPb-7448	3H-036-bPb-2929	3H-055-bPb-9746	3H-067-Bmag0006	3H-070-bPb-5012	3H-072-bPb-3805	3H-079-bPb-3317	3H-088-bPb-1681	3H-140-bPb-2550	3H-150-bPb-9599	3H-154-bPb-8419	3H-166-bPb-0361	3H-_-scssr25538	3H-179-bPb-7724
3	8	1H-007-bPb-1348	■	x																												
4	29	1H-013-bPb-8973	x	■	x	x																										
5	4	1H-034-bPb-9337			■	x																										
6	17	1H-041-bPb-7231		x	x	■		x																								
7	16	1H-053-bPb-9333					■	x	x																							
8	30	1H-054-bPb-1535				x	x	■																								
9	20	1H-059-bPb-0910					x		■	x												x	x									
10	12	1H-059-HvM20								■												x	x									
11	19	1H-066-bPb-1193							x		■	x										x	x									
12	9	1H-068-bPb-1723									x	■	x																			
13	7	1H-077-bPb-3389										x	■	x																		
14	10	1H-095-bPb-5249											x	■	x																	
15	5	1H-116-bPb-5014												x	■	🟩																
16	13	3H-001-bPb-4022													🟩	■	x															
17	3	3H-010-bPb-9945														x	■	x														
18	25	3H-012-bPb-7770															x	■	x													
19	24	3H-020-bPb-7448																x	■	x												
20	15	3H-036-bPb-2929																	x	■	x											
21	1	3H-055-bPb-9746																		x	■		x									
22	22	3H-067-Bmag0006							x	x	x											■										
23	27	3H-070-bPb-5012							x	x													■	x								
24	21	3H-072-bPb-3805																					x	■	x							
25	14	3H-079-bPb-3317																						x	■	x						
26	2	3H-088-bPb-1681																							x	■	🟩					
27	26	3H-140-bPb-2550																								🟩	■	x				
28	18	3H-150-bPb-9599																									x	■	x			
29	23	3H-154-bPb-8419																										x	■	x		
30	28	3H-166-bPb-0361																											x	■		x
31	11	3H-_-scssr25538																													■	x
32	6	3H-179-bPb-7724																												x	x	■
33			8	29	4	17	16	30	20	12	19	9	7	10	5	13	3	25	24	15	1	22	27	21	14	2	26	18	23	28	11	6

Supplementary material

When $\varepsilon_{M1} = \varepsilon_{M3} = 0$ and $\varepsilon_{M2} = \varepsilon$, $r_{M1M2} = (1–2\theta_{12})(1–2\varepsilon)$, $r_{M2M3} = (1–2\theta_{23})(1–2\varepsilon)$ and $r_{M1M3} = (1–2\theta_{12})(1–2\theta_{23})$.

Let $a = 1–2\theta_{12}$, $b = 1–2\theta_{23}$, $x = 1–2\varepsilon$.
$\because\ 0 < \theta_{12} < 0.5,\ \therefore\ 0 < a < 1$
$\because\ 0 < \theta_{23} < 0.5,\ \therefore\ 0 < b < 1$
$\because\ 0 < \varepsilon < 0.5,\ \therefore\ 0 < x < 1$

$$\rho_{M1M2|M3} = \frac{r_{M1M2} - r_{M1M3} \times r_{M2M3}}{\sqrt{1-r_{M1M3}^2}\sqrt{1-r_{M2M3}^2}}$$

$$= \frac{(1-2\theta_{12})(1-2\varepsilon)-(1-2\theta_{12})(1-2\theta_{23})(1-2\theta_{23})(1-2\varepsilon)}{\sqrt{1-(1-2\theta_{12})^2(1-2\theta_{23})^2}\sqrt{1-(1-2\theta_{23})^2(1-2\varepsilon)^2}}$$

$$= \frac{ax-ab^2x}{\sqrt{1-a^2b^2}\sqrt{1-b^2x^2}} = \frac{a(1-b^2)}{\sqrt{1-a^2b^2}}\left(x^{-2}-b^2\right)^{-\frac{1}{2}}$$

$$\because\ \frac{d\rho_{M1M2|M3}}{dx} = \frac{a(1-b^2)}{\sqrt{1-a^2b^2}}\left(-\frac{1}{2}\right)\left(x^{-2}-b^2\right)^{-\frac{3}{2}}(-2)x^{-3} = \frac{a(1-b^2)}{\sqrt{1-a^2b^2}}\left(1-b^2x^2\right)^{-\frac{3}{2}} > 0$$

$\therefore\ \rho_{M1M2|M3}$ is a monotonically increasing function of x.

Considering that x is a monotonically decreasing function of ε, $\rho_{M1M2|M3}$ is therefore a monotonically decreasing function of ε.

When $\varepsilon_{M1} = \varepsilon_{M3} = 0$ and $\varepsilon_{M2} = \varepsilon$, $r_{M1M2} = (1–2\theta_{12})(1–2\varepsilon)$, $r_{M2M3} = (1–2\theta_{23})(1–2\varepsilon)$ and $r_{M1M3} = (1–2\theta_{12})(1–2\theta_{23})$.

Let $a = 1–2\theta_{12}$, $b = 1–2\theta_{23}$, $x = (1–2\varepsilon)^2$.
$\because$ $0 < \theta_{12} < 0.5$, $\therefore$ $0 < a < 1$
$\because$ $0 < \theta_{23} < 0.5$, $\therefore$ $0 < b < 1$
$\because$ $0 < \varepsilon < 0.5$, $\therefore$ $0 < x < 1$

$$\rho_{M1M3|M2} = \frac{r_{M1M3} - r_{M1M2} \times r_{M2M3}}{\sqrt{1-r_{M1M2}^2}\sqrt{1-r_{M2M3}^2}}$$

$$= \frac{(1-2\theta_{12})(1-2\theta_{23}) - (1-2\theta_{12})(1-2\varepsilon)(1-2\theta_{23})(1-2\varepsilon)}{\sqrt{1-(1-2\theta_{12})^2(1-2\varepsilon)^2}\sqrt{1-(1-2\theta_{23})^2(1-2\varepsilon)^2}}$$

$$= \frac{ab-abx}{\sqrt{1-a^2x}\sqrt{1-b^2x}} = ab(1-x)\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}$$

$$\frac{d\rho_{M1M3|M2}}{dx} = -ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}} + ab(1-x)\left(-\frac{1}{2}\right)\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{3}{2}}\left[\left(-a^2\right)\left(1-b^2x\right) - b^2\left(1-a^2x\right)\right]$$

$$= -ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}} + ab(1-x)\frac{1}{2}\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{3}{2}}\left[a^2\left(1-b^2x\right) + b^2\left(1-a^2x\right)\right]$$

$$= ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}(1-x)\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-1}\left[a^2\left(1-b^2x\right)+b^2\left(1-a^2x\right)\right]\right\}$$

$$= ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}(1-x)\frac{a^2\left(1-b^2x\right)+b^2\left(1-a^2x\right)}{\left(1-a^2x\right)\left(1-b^2x\right)}\right\}$$

$$= ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}(1-x)\left[\frac{a^2}{1-a^2x}+\frac{b^2}{1-b^2x}\right]\right\}$$

$$= ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}\left[\frac{a^2(1-x)}{1-a^2x}+\frac{b^2(1-x)}{1-b^2x}\right]\right\}$$

$$< ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}\left(a^2+b^2\right)\right\} < ab\left[\left(1-a^2x\right)\left(1-b^2x\right)\right]^{-\frac{1}{2}}\left\{-1+\frac{1}{2}\times 2\right\} = 0$$

Thus, $\rho_{M1M2|M3}$ is a monotonically decreasing function of x. Considering that x is a monotonically decreasing function of ε, $\rho_{M1M2|M3}$ is accordingly a monotonically increasing function of ε.

Chapter 5

A comparative simulation study of the PC algorithm and the Metropolis-Hastings algorithm in constructing random and scale-free Bayesian networks

Abstract

Deciphering causal relationships from genetic and molecular phenotyping data sets has remained one of the central challenges in computational biology. Among existing approaches to the reconstruction of directed causal networks, Bayesian networks have proven to be promising both theoretically and practically. In particular, the PC algorithm and the Metropolis-Hastings algorithm, which are representatives of mainstream methods to the structure learning of Bayesian networks, are reported to have been successfully applied to the domain of biology. Most biological systems are considered to exist in the form of random network or scale-free network. The two types of networks are essentially different from each other in terms of node degree distribution. In view of these facts, here we compare the performance of the PC algorithm and the Metropolis-Hastings algorithm in constructing both random and scale-free Bayesian networks. Our simulation study shows that for either type of Bayesian network, the PC algorithm is superior to the M-H algorithm in terms of timeliness; the M-H algorithm is preferable to the PC algorithm when the completeness of reconstruction is emphasized; but when the fidelity of reconstruction is taken into account, the better one of the two algorithms varies from case to case. Moreover, whichever algorithm is adopted, larger sample sizes generally permit more accurate reconstructions, especially in regard to the completeness of the resulting networks.

5.1 Introduction

Constructing causal networks from genetic and molecular phenotyping data sets is still a major challenge in computational biology. Earlier approaches mainly resort to clustering and correlation analysis, which are rather straightforward techniques but with limited effectiveness. More specifically, clustering is able to uncover the modular topology of metabolic and protein interaction networks but cannot explore in depth the fine architecture of each module (Hanisch et al. 2002; Jiang and Singh 2010; Muff et al. 2005; Ravasz et al. 2002). The measures of correlation are known to not only confound direct and indirect associations but also provide no means to distinguish between cause and effect (Opgen-Rhein and Strimmer 2007). In particular, the most commonly used Pearson correlation coefficient applies rigorously only to linear associations with Gaussian noise (Numata et al. 2008).

Over the past two decades, there have been considerable attempts to construct diverse biological networks using more advanced approaches, among which Bayesian networks (BNs) have proven to be promising for causal network inference (Ellis and Wong 2008; Friedman et al. 2000; Heckerman 1998; Margaritis 2003). Formally, a BN consists of two components: (1) a directed acyclic graph (DAG) that encodes a set of conditional independence assertions about the variables of interest; (2) a conditional probability distribution (CPD) assigned to each variable given its parents in the DAG. A BN essentially represents a factorization of a multivariate probability distribution, that is, it decomposes a joint probability distribution over multiple variables into a set of conditional and marginal distributions on low-dimensional subspaces. This forms the basis for efficient reasoning under uncertainty, whose core idea is to explore the dependence structure of variables to facilitate reasoning in multidimensional domains under probabilistic settings (Reusch and Temme 2013; Wang et al. 2002).

Mainstream approaches to structure learning of BNs can be broadly divided into two categories: constraint-based search and score-based search. Constraint-based search is built on a well- organized set of conditional independence tests. It starts with a complete graph, and then deletes certain edges if the corresponding conditional independencies are detected from the training data. Representative algorithms of this category are the IC algorithm (Pearl and Verma 1995) and the PC algorithm (Spirtes et al. 1993). Score-based search benefits from research achievements in optimisation theory. It typically searches through a large model space with a heuristic strategy, and returns one or several most likely networks according to a scoring metric. Existing search strategies include greedy hill climbing, stochastic search such as Markov Chain

Monte Carlo (MCMC) and simulated annealing, other optimization methods such as genetic algorithms and ant colony optimization algorithms (Yuan and Malone 2013). Well-known scoring metrics are those based on information theory, such as *Akaike information criterion* (AIC) (Akaike 1998), *Bayesian information criterion* (BIC) (also known as *minimum description length* (MDL)) (Lam and Bacchus 1994; Rissanen 1978), *normalized maximum likelihood* (NML) (Silander et al. 2008) and *mutual information test* (MIT) (Campos 2006), and Bayesian scoring functions such as *Bayesian Dirichlet* (BD) (Heckerman et al. 1995) and its variants (K2, BDe and BDeu) (Buntine 1991; Cooper and Herskovits 1992; Heckerman et al. 1995).

Among various methods for structure learning of BNs, the two most common ones that have been successfully applied to the reconstruction of biological networks are the PC algorithm (Gavai et al. 2009; Mansmann and Jurinovic 2011; Schmidberger et al. 2011) and the MCMC approach (Husmeier 2003; Ram and Chetty 2009; Wu and Liu 2008; Zhou et al. 2004). In view of the promising expansion of the use of the two methods in systems biology, it becomes important for scientific researchers to better understand their relative strengths and weaknesses. Kalisch and Bühlmann (2007) evaluated the effectiveness of the PC algorithm in skeleton reconstruction of high-dimensional DAGs, and provided general guidance on parameter selection. Wu and Liu (2008) demonstrated MCMC had better accuracy and efficiency than greedy hill climbing search in the reconstruction of dynamic BNs. However, comparative evaluation of the PC algorithm and the MCMC approach has not yet been reported.

To fill the gap, here we compare the performance of the PC algorithm and the Metropolis-Hastings (M-H) algorithm (Metropolis et al. 1953), a representative MCMC method, in reconstructing synthetic causal networks of different complexities. Our primary motivation is, by running comparative simulations we hope to answer the question: under what circumstances will one algorithm outperform the other? Moreover, we aim to provide an informative guide to parameter selection in the use of each method. Complex networks were initially modelled using random networks introduced by Edgar Gilbert, Paul Erdős and Alfréd Rényi (Erdos and Rényi 1960; Gilbert 1959). A few recent studies have shown that many real-world networks exhibit scale-free behaviour in terms of node degree distribution (Albert 2005; Barabasi and Albert 1999; Jeong et al. 2000). Nonetheless, it is still under debate whether a wide variety of biological systems should be presented in the form of random network or scale-free network (Khanin and Wit 2006). Considering this, our comparative simulations are implemented on both random and scale-free BNs.

The rest of this chapter is organized as follows. Section 2 reviews the PC algorithm and the M-H algorithm. Section 3 describes the simulation settings. Results of comprehensive comparison followed by detailed discussion are presented in section 4. Finally, concluding remarks are drawn in section 5.

5.2 Methods

A BN is a DAG where the nodes {*1, 2*, …, *n*} denote random variables {$X_1, X_2, \ldots, X_n$} and the edges represent dependencies between the variables. By definition each node is assigned with a CPD given its parents. The directed acyclic structure in combination with all associated CPDs encodes a decomposition of the joint probability distribution (JPD) over all variables. Let $pa(X_i)$ be the parent nodes of X_i. According to the chain rule of probability, we have

$$P(X_1,\ldots,X_n)=\prod\nolimits_{i=1}^{n}P(X_i \mid pa(X_i)) \qquad (1)$$

A major advantage of BN lies in the fact that reverse derivations of various CPDs from the JPD enable probabilistic inference, which is to answer probabilistic queries in the form of $P(\boldsymbol{A}|\boldsymbol{B})$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are disjoint subsets of $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$. Real-world applications of probabilistic inference are, for example, to help clinicians diagnose diseases from symptoms and predict responses to treatments.

5.2.1 The PC algorithm

The PC algorithm is named after its inventors Peter Spirtes and Clark Glymour. It is designed to firstly construct an undirected network with conditional independence tests and then orient as many edges as possible according to certain orientation rules. It uses G^2 test for conditional independence of discrete variables, and uses Fisher’s z-transformation of the sample partial correlation to test for zero partial correlation of sets of normally distributed random variables. Note that a major, but often neglected, problem of the PC algorithm is that it is generally order-dependent, in the sense that the output depends on the order in which the variables are given (Cano et al. 2008; Dash and Druzdzel 1999). For this reason, Colombo and Maathuis have made a modification to the first phase of the PC algorithm to address the issue of order-dependence involved in the determination of *equivalent DAGs** (Colombo and Maathuis 2014).

**Equivalent DAGs: Two DAGs are said equivalent or in the same equivalence class if they show alternative ways of describing the same set of conditional dependencies. Pearl and Verma (1995) proved that DAGs are equivalent if and only if they have the* *same skeleton and the same set of v-structures (v-structure: two parent nodes converge on a child node).*

5.2.2 The M-H algorithm

The M-H algorithm is a commonly used MCMC method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult. In its application to BN structure learning, the state space of the resulting Markov chain consists of all plausible DAGs, whose stationary distribution is the desired posterior distribution. Statistical analysis on the Markov chain after burn-in will reveal general properties of the desired distribution, based on which one or several most likely DAGs can be returned as the result of structure learning. A detailed description of the M-H algorithm is given below.

5.2.2.1 Searching strategy

A Markov chain is a discrete random process holding the Markov property, which refers to the fact that the next state depends purely on the current state, i.e. the future is independent from the past given the present. An ergodic Markov chain will eventually converge to a stationary distribution, no matter which state the chain begins with. Let $\{M_1, M_2, \ldots, M_m\}$ denote the finite state space of a Markov chain and $P(M_l|M_k)$ represents the transition probability of going from state M_k to state M_l. The mathematical expressions of the two aforementioned properties are given as Eq.2-4, where D is the observational data and t is the step counter.

$$P_{t+1}(M_l) = \sum\nolimits_k P(M_l \mid M_k) P_t(M_k) \tag{2}$$

$$P_{t\to\infty}(M_l) \triangleq P_\infty(M_l) \tag{3}$$

$$P_\infty(M_l \mid D) = \sum\nolimits_k P(M_l \mid M_k) P_\infty(M_k \mid D) \tag{4}$$

$$P(M_l \mid M_k) \triangleq Q(M_l \mid M_k) A(M_l \mid M_k) \tag{5}$$

$$A(M_l \mid M_k) = \min\left\{ \frac{P(M_l \mid D) Q(M_k \mid M_l)}{P(M_k \mid D) Q(M_l \mid M_k)}, 1 \right\} \tag{6}$$

The M-H algorithm aims to generate a proper transition matrix, so that the Markov chain drawn accordingly can satisfy convergence (Eq.4). It defines $P(M_l|M_k)$ as the product of two terms: a proposal probability $Q(M_l|M_k)$ and an acceptance probability $A(M_l|M_k)$ (Eq.5). It allows only three elementary proposal moves for DAGs: (1) deletion of an edge, (2) reversal of an edge, and (3) creation of a new edge. Note that either of the last two moves may lead to graphs that violate the acyclic constraint and therefore have to be discarded. The neighbourhood of a DAG is restricted to all valid DAGs that can be reached from the given DAG with one of the three elementary moves. $Q(M_l|M_k)$ is then given by the reciprocal of the neighbourhood size of M_k, as each member within the neighbourhood has an equal chance of occurring in a random walk. $A(M_l|M_k)$ is calculated by Eq.6, which has proven to be a sufficient condition for Eq.4.

5.2.2.2 Bayesian scoring metrics

Bayesian scoring metrics define the best candidate DAG M^* as Eq.7, in which the posterior probability $P(M_k|D)$ can be computed by Eq.8, where $P(M)$ is the prior probability and $P(D|M_k)$ is the likelihood. Since all candidate DAGs are typically assigned with equal prior probabilities and the term $P(D)$ is identical, $P(D|M_k)$ is always calculated instead of $P(M_k|D)$ in practice.

$$P\left(M^* \mid D\right) \triangleq \max_k P\left(M_k \mid D\right) \tag{7}$$

$$P\left(M_k \mid D\right) = \frac{P\left(D \mid M_k\right) P\left(M_k\right)}{P\left(D\right)} \tag{8}$$

On the basis of four assumptions: multinomial samples, Dirichlet parameters, parameter independence and parameter modularity, Heckerman et al. (1995) proposed the Bayesian Dirichlet (BD) score (Eq.9),

$$P\left(D \mid M\right) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left(\frac{\Gamma\left(\alpha_{ij}\right)}{\Gamma\left(N_{ij} + \alpha_{ij}\right)} \times \prod_{k=1}^{r_i} \frac{\Gamma\left(N_{ijk} + \alpha_{ijk}\right)}{\Gamma\left(\alpha_{ijk}\right)} \right) \tag{9}$$

where N_{ijk} is the number of times the event $\{X_i = k, pa(X_i) = j\}$ occurs in D, r_i and q_i are the numbers of possible values for X_i and $pa(X_i)$, α_{ijk} denotes the hyper-parameter of Dirichlet distribution, $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Considering that the specification of α_{ijk} for all i, j and k is formidable in practice, Cooper and Herskovits (1992) suggests a simple uninformative assignment, i.e. $\alpha_{ijk} = 1$. This is referred to as the K2 score. In addition, Heckerman et al. (1995) also suggested another uninformative assignment over α_{ijk} by specifying an equivalent sample size N' and a prior probability distribution for each variable given its parents (Eq.10).

$$\alpha_{ijk} = N' \times P\left(X_i = k, pa\left(X_i\right) = j\right) \tag{10}$$

This metrics is called the BDe score since it possesses the property of likelihood equivalence, i.e. $P(D|M_k) = P(D|M_l)$ if the two DAGs M_k and M_l are equivalent. Buntine (1991) proposed a particular BDe score, named the BDeu score, where $P\left(X_i = k, pa\left(X_i\right) = j\right) = 1/\left(q_i \times r_i\right)$ and $N' = 1$, that is, $\alpha_{ijk} = 1/\left(q_i \times r_i\right)$.

5.2.2.3 Convergence diagnosis and structural feature selection

In the application of the M-H algorithm, it is critical to assess the convergence of the resulting Markov chains. Only then can one draw a safe conclusion on the desired posterior distribution. A common way for effective diagnosis is to implement multiple MCMC runs from overdispersed starting points, and then track a same convergence indicator for all the chains. Convergence is achieved when the tracks are well mixed. Werhli et al. (2006) found that a burn-in period of 20,000 steps followed by a sampling period of 80,000 steps, keeping samples in intervals of 200 MCMC steps, was usually sufficient.

It is generally not straightforward to select the most likely one or several DAGs from the posterior distribution of candidate models. This is because the number of DAGs has super-exponential growth of the number of variables, which implies the posterior distribution over all possible DAGs is rather scattered and few DAGs can be considered significant. Moreover, as mentioned previously, the most practical scoring metrics, BDe score, will assign the same score to equivalent DAGs. Therefore, it is advisable to consider the posterior distribution over equivalence classes instead of DAGs. A further way to make the posterior distribution less diffuse and more informative is to map the high-dimensional space of equivalence classes into the low-dimensional space of structural features. In this study, we extracted the remarkable high-frequency structural features to construct high-confidence (but incomplete) networks. Note that we only focused on a particular type of structural features, i.e. the parent-child relations. The posterior probability of a feature f is given by Eq.11, where $f(M_k) = 1$ if the network M_k satisfies the feature f, otherwise $f(M_k) = 0$.

$$P\left(f \mid D\right) = \sum_{k=1}^{m} f\left(M_k\right) P\left(M_k \mid D\right) \qquad (11)$$

5.3 Simulation setup

5.3.1 Random networks vs. scale-free networks

Two random network models have been commonly used in multiple disciplines. In the model introduced by Gilbert (1959), every possible edge occurs independently with equal probability. In the model characterised by Erdos and Rényi (1960), all possible random networks on a given set of variables with a fixed number of edges are equally likely. In comparison, Gilbert's model is more widely used in practice, due in part to the ease of analysis allowed by the independence of the edges.

It has recently been shown that the node connectivity of many large-scale networks follows a power-law distribution (Barabasi and Albert 1999). Such networks, unlike random networks, lack typical node degrees and are thus referred to as scale-free networks (Albert 2005). Scale-free networks possess some intriguing properties. First, they belong to the class of small-world networks, in which most nodes are not neighbours of one another but can be reached from every other through a short path. Second, though low-degree nodes are the majority in scale-free networks, nodes whose degrees are much higher than average, so-called hubs, exist as well. Third, scale-free networks are robust to random breakdowns, i.e. random node disruptions do not usually result in a major loss of connectivity. These properties are also possessed by various biological networks. Accordingly, there is a view that many biological networks are scale-free (Albert 2005; Jeong et al. 2000).

However, scientific researchers still debate on whether certain biological networks are indeed scale-free, since the scale-free assumption does not hold for a number of published datasets of various biological interactions (Khanin and Wit 2006). In view of this, in the present study we implement comparative simulations of the PC algorithm and the M-H algorithm on both random and scale-free BNs.

5.3.2 Synthetic data

We created (1) six different random BNs by function *randomDAG* in R package *pcalg* and (2) six different scale-free BNs by the Barabási-Albert (B-A) algorithm in Matlab. Each BN can be characterized by a two-dimensional vector (a, b), where a is the number of nodes and b is the number of edges in the graph. As shown in **Figure 5.1** and **2**, the six random BNs and the six scale-free BNs can be, respectively, sorted with respect to their complexities: (30, 30), (30, 55), (40, 40), (40, 75), (50, 50) and (50, 95). To facilitate analysis, for every BN, a Beta distribution was randomly assigned to each variable as its predetermined prior probability, or, conditional probability given the values of all its parents. That is, all variables were assumed to be binary.

In addition to network complexity, sample size of the observational data is another vital factor that should be taken into account in the evaluation of BN structure learning algorithms. We therefore investigated for each BN six different sample sizes, including 200, 350, 500, 1000, 2000 and 5000. Further, for every sample size, three replication datasets were generated in order to improve simulation accuracy.

5.3.3 Implementation and evaluation of the structure learning algorithms

In the use of the PC algorithm, the significance level of conditional independence tests was set at 0.01 and 0.05, respectively, for each set of data. The implementation of the PC algorithm is available in R package *pcalg*. In the use of the M-H algorithm, four independent Markov chains, each with length of 100,000 steps, were generated for every dataset. And for each Markov chain, the number of edges in the BN learnt at each step was tracked as the indicator of convergence. According to the mixture of the four tracks (see examples in **Figure 5.3**), the very first 20,000 steps were considered as the burn-in period and thus were discarded. We calculated the marginal posterior probability of every single edge, and then eliminated those edges whose probabilities were lower than a certain threshold. However, to our knowledge, there is no existing study reporting a golden rule for selecting such a threshold. Thus we evaluated the outcomes on the basis of a wide range of thresholds, which step between 0.1 and 0.9 by 0.1, to find out the most favourable threshold value. Open-source code of the M-H algorithm is available in Matlab package *Bayes Net Toolbox*.

We exploited a well-established criterion to evaluate and compare the simulation results. Each edge obtained after the structural feature selection mentioned previously was classified into one of the following four categories: (1) true positive (TP), i.e. the edge appears in both the true network and the selected high-frequency structural features; (2) false positive (FP), i.e. the edge appears in the selected high-frequency structural features but not in the true network; (3) true negative (TN), i.e. the edge occurs in neither the true network nor the selected high-frequency structural features; (4) false negative (FN), i.e. the edge occurs in the true network but not in the selected high-frequency structural features. After the events of TP, FP, TN and FN were counted, true positive rate (TPR), false positive rate (FPR) and precision can be computed as follows:

$$TPR = \frac{\#TP}{\#TP + \#FN}, FPR = \frac{\#FP}{\#FP + \#TN}, Precision = \frac{\#TP}{\#TP + \#FP}.$$

TPR, also known as sensitivity or recall in machine learning, measures completeness of the reconstruction. FPR, also known as the fall-out, is mathematically equal to the type I error rate. Precision measures the fidelity of the reconstruction. Apparently, a good graphical modelling algorithm should lead to high TPR and precision but low FPR. Here we used the ROC curve (a graphical plot of TPR vs. FPR) and precision-recall curve to compare (1) performance of different algorithms under the same simulation setting and (2) performance of the same algorithm under different simulation settings.

5.4 Results

To give a comprehensive representation of the simulation results, we first display the ROC and precision-recall curves obtained for each algorithm on every synthetic BN. Based on these diagrams, we are able to find out the optimal parameter setting for each algorithm. Afterwards, we compare the two algorithms with the selected parameters to see if one algorithm outperforms the other under certain circumstances.

5.4.1 Simulation results of the PC algorithm

Figure 5.4 shows, for every two networks that are of the same type and are composed of the same number of nodes but different numbers of edges, the ROC curves obtained by the PC algorithm at two significance levels, i.e. 0.01 and 0.05. Each line connects six data points that correspond to the six sample sizes in increasing order. We found that for any particular network and a given significance level, TPR grew successively as the sample size increased. For any particular network and a given sample size, higher TPR and FPR were obtained at a loose significance level, i.e. 0.05 rather than 0.01. Moreover, for any two networks that are of the same type and with the same number of nodes, given sample size and the significance level, the more complex the network (i.e. the higher the number of edges), the lower the TPR.

Figure 5.5 shows the precision-recall curves in the same layout as **Figure 5.4**. It is worth noting that for any particular network and a given sample size, higher recall but lower precision were obtained at a loose significance level, i.e. 0.05 rather than 0.01. Nonetheless, compared with recall, precision was much more sensitive to changes in the significance level. Thus we consider that a stricter significance level, e.g. 0.01 instead of 0.05, is preferable in the practical application of the PC algorithm.

5.4.2 Simulation results of the M-H algorithm

The six coloured lines in each subgraph of **Figure 5.6** and **7** display the ROC curves obtained by the M-H algorithm for a given synthetic network and six different sample sizes. Specifically, the nine data points involved in each curve correspond to increasing threshold of structural feature selection, which steps between 0.1 and 0.9 by 0.1. From the diagrams we concluded that, for any particular network and a given sample size, both TPR and FPR increased as the threshold decreased. And generally speaking, when the threshold reduced from 0.9 to 0.5, TPR grew rapidly while FPR maintained at very low levels (< 0.02); when the threshold further reduced from 0.5 to 0.1, TPR increased at a slow rate while FPR had successive growth. This indicates that thresholds above 0.5 lead to loss of TPR and thus should be considered too stringent, whereas thresholds below 0.5 result in increase of FPR and thus should be considered too lax. Consequently, 0.5 would be suggested as the optimal threshold value for structural feature selection on networks obtained by the M-H algorithm. In each subgraph of **Figure 5.6** and **7**, markers of different shapes are used to highlight the six data points that were obtained with fixed threshold 0.5 but different sample sizes. It is clear that for any particular network, FPR remained very low while TPR dropped remarkably along with the gradual reduction of sample size.

Figure 5.8 and **9** show the precision-recall curves in the same layout as **Figure 5.6** and **7**. For any particular network and a given sample size, when the threshold of structural feature selection decreased from 0.9 to 0.5, recall grew rapidly and precision remained at high levels; when the threshold further decreased from 0.5 to 0.1, recall did not increase too much while precision dropped significantly. This also implies that 0.5 would be the optimal threshold for structural feature selection as it offered the best trade-off between precision and recall. Again, in each subgraph of **Figure 5.8** and **9**, markers of different shapes are used to highlight the six data points that were obtained with fixed threshold 0.5 but different sample sizes. From the locations of the markers we can tell that, for any particular network, precision kept up high levels and recall increased dramatically along with the gradual increase in sample size.

5.4.3 Comparison between performance of the two algorithms

We concluded above that no matter random BNs or scale-free BNs are targeted, the lower significance level for conditional independence tests, say 0.01 instead of 0.05, is preferred in the use of the PC algorithm; and the median threshold value of 0.5 is optimal for structural feature selection following the use of the M-H algorithm. Below, we will make further comparison between the two algorithms given the suggested parameters.

Figure 5.10 and **11** present the precision-recall curves drawn for every synthetic network with increasing sample size. For any particular network and a given sample size, recall of the M-H algorithm was higher than that of the PC algorithm, which indicated the former outperformed the latter with respect to the completeness of reconstruction. As for the fidelity of reconstruction, it is hard to draw a straightforward conclusion since there was no clear pattern of precision over all subgraphs.

5.5 Discussion

Through the simulation we have noticed that, the M-H algorithm was much more time-consuming than the PC-algorithm. Moreover, the number of network nodes, rather than the number of network edges or the sample size, was the most critical factor affecting the running time of the M-H algorithm. Nonetheless, it has been reported that the bottleneck of applying the M-H algorithm to BN structure learning in practice is the memory requirement rather than the time requirement (Tamada et al. 2011). This is because the M-H algorithm needs to calculate the inverse of an adjacency matrix, such calculation typically has high demand for memory. In addition, during its learning process the M-H algorithm needs to store the intermediate optimal structures whose number increases super-exponentially with the number of network nodes. Consequently, the M-H algorithm is currently only applicable to networks of small or moderate sizes (with up to a few tens of nodes) on a typical PC. For plant species, the crop populations typically contain hundreds to thousands of individuals. In view of this and also the performance limit of the M-H algorithm, we therefore restricted the scale of simulations to 30~50 nodes in combination with 200~5000 samples.

5.6 Conclusion

In summary, we conclude that in the reconstruction of both random and scale-free BNs, the significance level for conditional independence tests involved in the PC algorithm is preferred at 0.01 rather than 0.05; and the optimal threshold for structural feature selection following the use of the M-H algorithm is 0.05. When the completeness of reconstruction is emphasized, the M-H algorithm is definitely preferable to the PC algorithm; but when the fidelity of reconstruction is taken into account, the better one of the two algorithms varies from case to case. In terms of timeliness, the PC algorithm is considered superior to the M-H algorithm, since the former is much less time-consuming than the latter. Last but not least, whichever algorithm is adopted, larger sample sizes generally permit more accurate reconstructions, especially in regard to the completeness of the resulting networks.

References

Akaike H (1998) Information theory and an extension of the maximum likelihood principle. Selected Papers of Hirotugu Akaike. Springer, pp 199-213

Albert R (2005) Scale-free networks in cell biology. Journal of cell science 118:4947-4957

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509-512

Buntine W (1991) Theory refinement on Bayesian networks. Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp 52-60

Campos LMd (2006) A scoring function for learning bayesian networks based on mutual information and conditional independence tests. Journal of Machine Learning Research 7:2149-2187

Cano A, Gómez-Olmedo M, Moral S (2008) A score based ranking of the edges for the PC algorithm. Proceedings of the Fourth European Workshop on Probabilistic Graphical Models, pp 41-48

Colombo D, Maathuis MH (2014) Order-independent constraint-based causal structure learning. Journal of Machine Learning Research 15:3741-3782

Cooper GF, Herskovits E (1992) A Bayesian method for the induction of probabilistic networks from data. Machine learning 9:309-347

Dash D, Druzdzel MJ (1999) A hybrid anytime algorithm for the construction of causal models from sparse data. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 142-149

Ellis B, Wong WH (2008) Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association 103:778-789

Erdos P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hungar Acad Sci 5:17-61

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology 7:601-620

Gavai AK, Tikunov Y, Ursem R, Bovy A, van Eeuwijk F, Nijveen H, Lucas PJ, Leunissen JA (2009) Constraint-based probabilistic learning of metabolic pathways from tomato volatiles. Metabolomics : Official journal of the Metabolomic Society 5:419-428

Gilbert EN (1959) Random graphs. The Annals of Mathematical Statistics 30:1141-1144

Hanisch D, Zien A, Zimmer R, Lengauer T (2002) Co-clustering of biological networks and gene expression data. Bioinformatics 18 Suppl 1:S145-154

Heckerman D (1998) A tutorial on learning with Bayesian networks. Learning in graphical models. Springer, pp 301-354

Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning 20:197-243

Husmeier D (2003) Reverse engineering of genetic networks with Bayesian networks. Biochemical Society transactions 31:1516-1518

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407:651-654

Jiang P, Singh M (2010) SPICi: a fast clustering algorithm for large biological networks. Bioinformatics 26:1105-1111

Kalisch M, Bühlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research 8:613-636

Khanin R, Wit E (2006) How scale-free are biological networks. Journal of computational biology 13:810-818

Lam W, Bacchus F (1994) Learning Bayesian belief networks: An approach based on the MDL principle. Computational intelligence 10:269-293

Mansmann U, Jurinovic V (2011) Biological feature validation of estimated gene interaction networks from microarray data: a case study on MYC in lymphomas. Briefings in bioinformatics 12:230-244

Margaritis D (2003) Learning Bayesian network model structure from data. US Army

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. The journal of chemical physics 21:1087-1092

Muff S, Rao F, Caflisch A (2005) Local modularity measure for network clusterizations. Physical review E, Statistical, nonlinear, and soft matter physics 72:056107

Numata J, Ebenhoh O, Knapp EW (2008) Measuring correlations in metabolomic networks with mutual information. Genome informatics International Conference on Genome Informatics 20:112-122

Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC systems biology 1:37

Pearl J, Verma TS (1995) A theory of inferred causation. Studies in Logic and the Foundations of Mathematics 134:789-811

Ram R, Chetty M (2009) MCMC Based Bayesian Inference for Modeling Gene Networks. IAPR International Conference on Pattern Recognition in Bioinformatics. Springer, pp 293-306

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551-1555

Reusch B, Temme K-H (2013) Computational intelligence in theory and practice. Springer Science & Business Media

Rissanen J (1978) Modeling by shortest data description. Automatica 14:465-471

Schmidberger M, Lennert S, Mansmann U (2011) Conceptual aspects of large meta-analyses with publicly available microarray data: a case study in oncology. Bioinformatics and biology insights 5:13-39

Silander T, Roos T, Kontkanen P, Myllymäki P (2008) Factorized normalized maximum likelihood criterion for learning Bayesian network structures. Proceedings of the 4th European workshop on probabilistic graphical models (PGM-08). Citeseer, pp 257-272

Spirtes P, Glymour CN, Scheines R (1993) Causation, prediction, and search. Springer-Verlag, New York

Tamada Y, Imoto S, Miyano S (2011) Parallel algorithm for learning optimal Bayesian network structure. Journal of Machine Learning Research 12:2437-2459

Wang H, Rish I, Ma S (2002) Using sensitivity analysis for selective parameter update in Bayesian network learning. Association for the Advancement of Artificial Intelligence (AAAI)

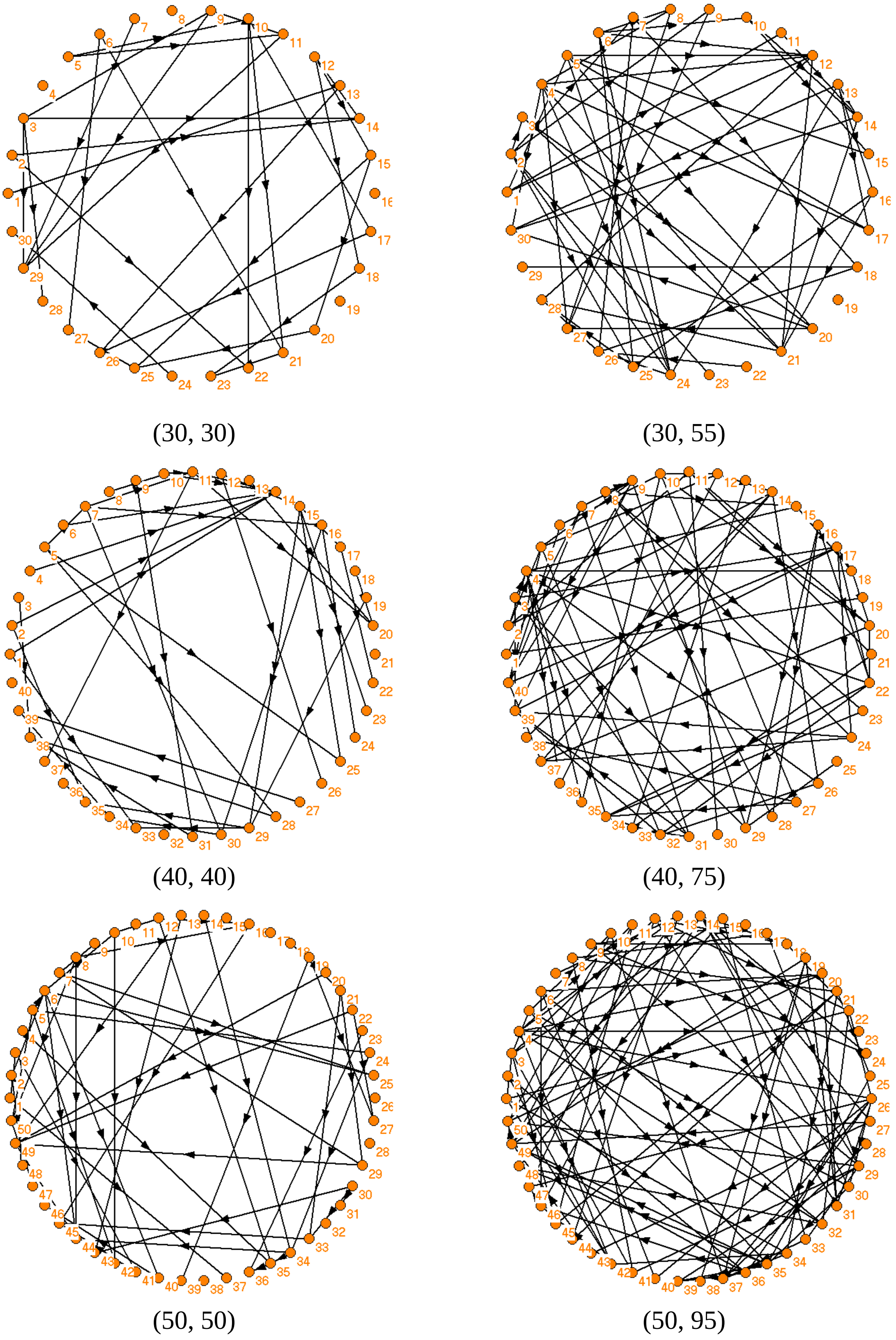
Werhli AV, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22:2523-2531

Wu H, Liu X (2008) Dynamic Bayesian networks modeling for inferring genetic regulatory networks by search strategy: comparison between greedy hill climbing and MCMC methods. Proc of World Academy of Science, Engineering and Technology, pp 224-234

Yuan C, Malone B (2013) Learning Optimal Bayesian Networks: A Shortest Path Perspective. J Artif Intell Res(JAIR) 48:23-65

Zhou X, Wang X, Pal R, Ivanov I, Bittner M, Dougherty ER (2004) A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. Bioinformatics 20:2918-2927

Figure 5.1 The six synthetic random BNs sorted according to their complexities. Network complexity is characterized by a two-dimensional vector (a, b), where a is the number of nodes and b is the number of edges in the graph.



(30, 30)

(30, 55)

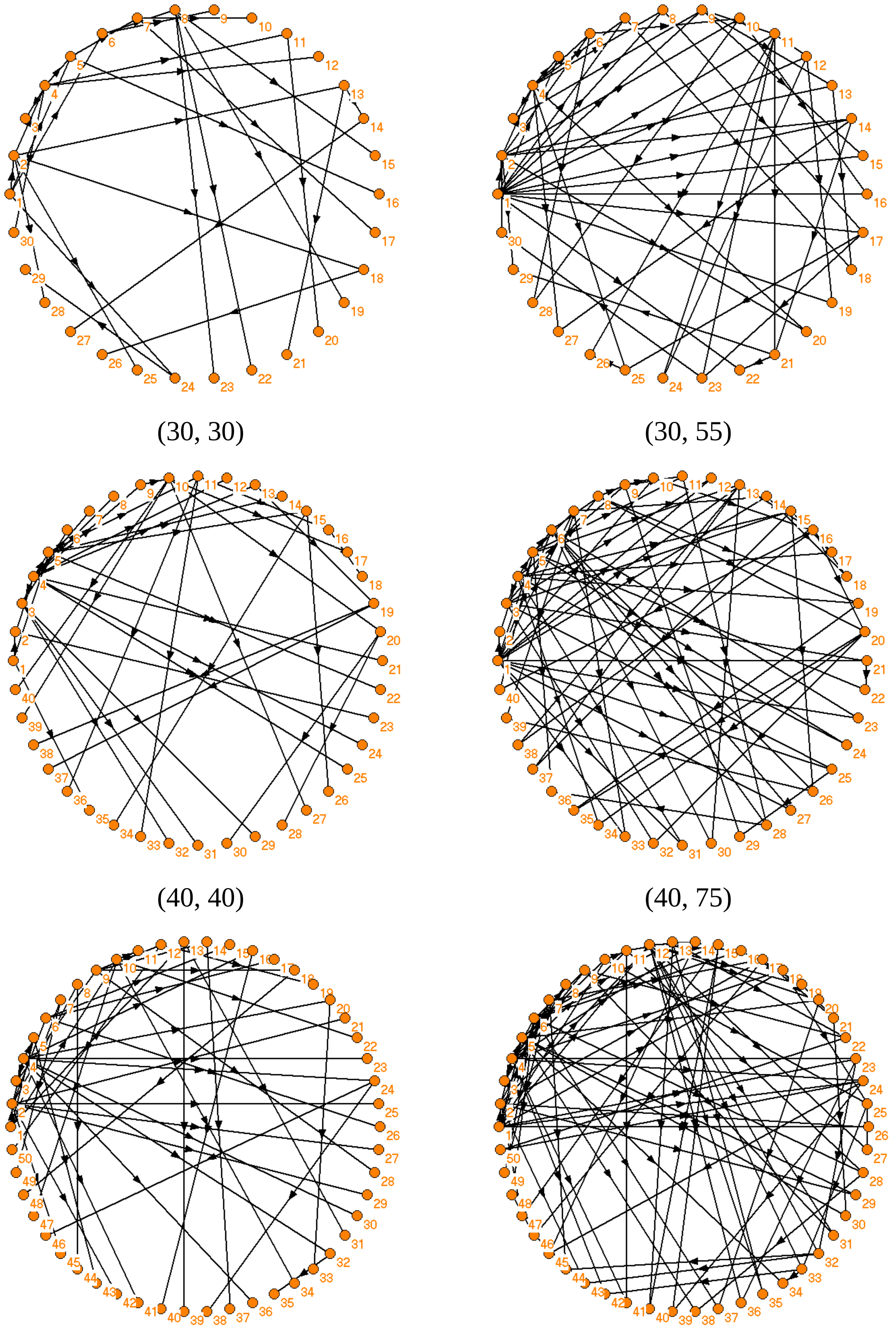
(40, 40)

(40, 75)

(50, 50)

(50, 95)

Figure 5.2 The six synthetic scale-free BNs sorted according to their complexities. Network complexity is characterized by a two-dimensional vector (a, b), where a is the number of nodes and b is the number of edges in the graph.



(30, 30) (30, 55)

(40, 40) (40, 75)

(50, 50) (50, 95)

Figure 5.3 Convergence diagnosis of four Markov chains by tracking the number of edges learned over 100,000 steps. The upper three subgraphs and the lower three subgraphs are shown, respectively, for the synthetic random and scale-free BNs (40, 40) with three different sample sizes 200, 500 and 2000.

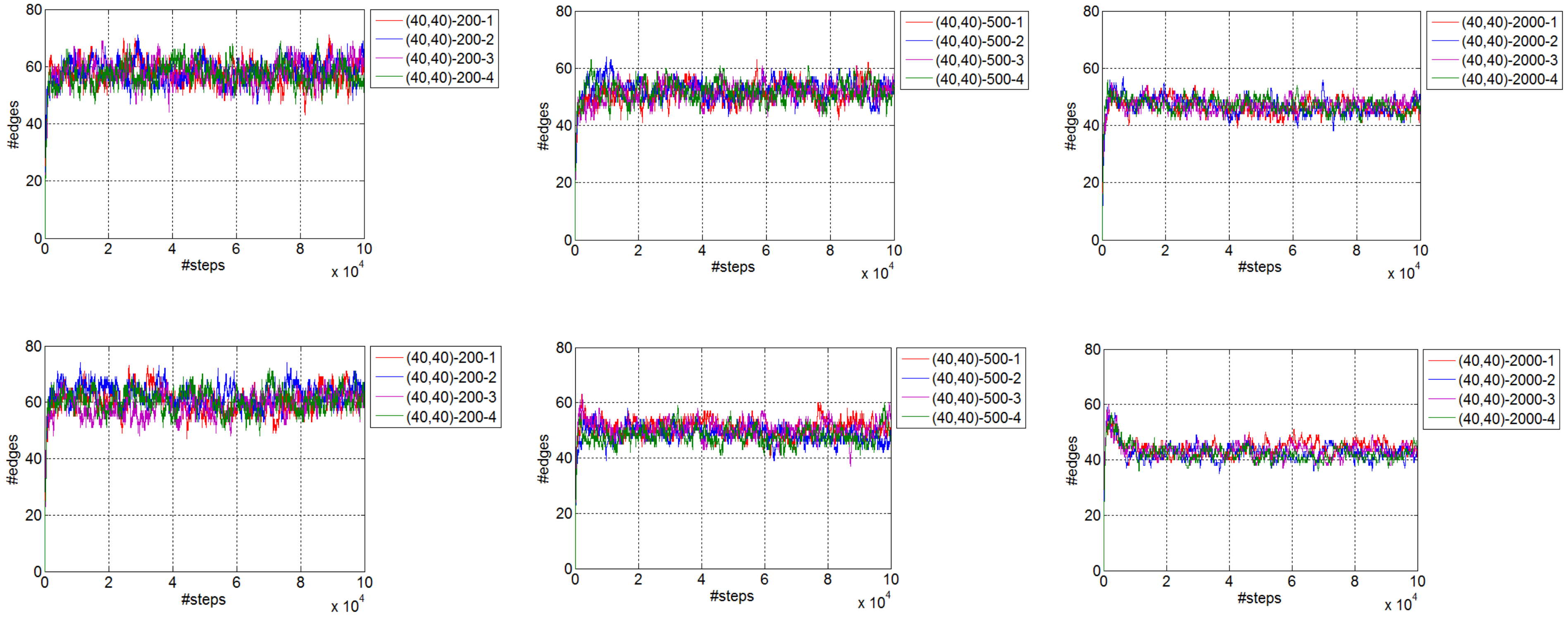


Figure 5.4 ROC curves of the PC algorithm. The upper three subgraphs and the lower three subgraphs are shown, respectively, for the synthetic random and scale-free BNs. In each subgraph, blue and red lines represent the curves obtained for networks with the same number of nodes but different numbers of edges: blue lines – fewer edges *vs.* red lines – more edges; solid and dashed lines represent the curves obtained at significance levels 0.01 and 0.05, respectively; the plus, square, dot, triangle, cross and diamond markers on each curve represent the TPR and FPR values obtained with sample sizes 200, 350, 500, 1000, 2000 and 5000, respectively.

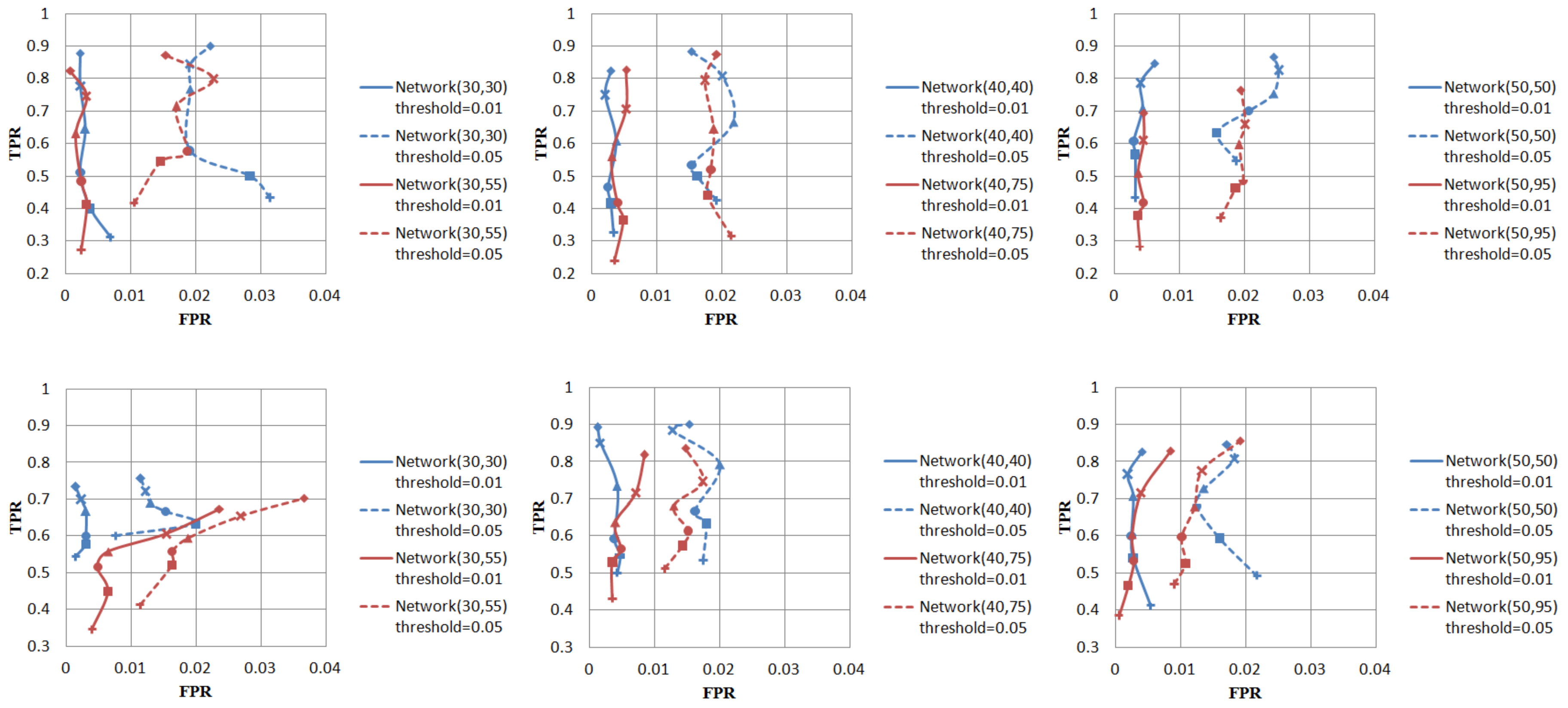


Figure 5.5 Precision-recall curves of the PC algorithm. The upper three subgraphs and the lower three subgraphs are shown, respectively, for the synthetic random and scale-free BNs. In each subgraph, blue and red lines represent the curves obtained for networks with the same number of nodes but different numbers of edges: blue lines – fewer edges *vs.* red lines – more edges; solid and dashed lines represent the curves obtained at significance levels 0.01 and 0.05, respectively; the plus, square, dot, triangle, cross and diamond markers on each curve represent the precision and TPR values obtained with sample sizes 200, 350, 500, 1000, 2000 and 5000, respectively.

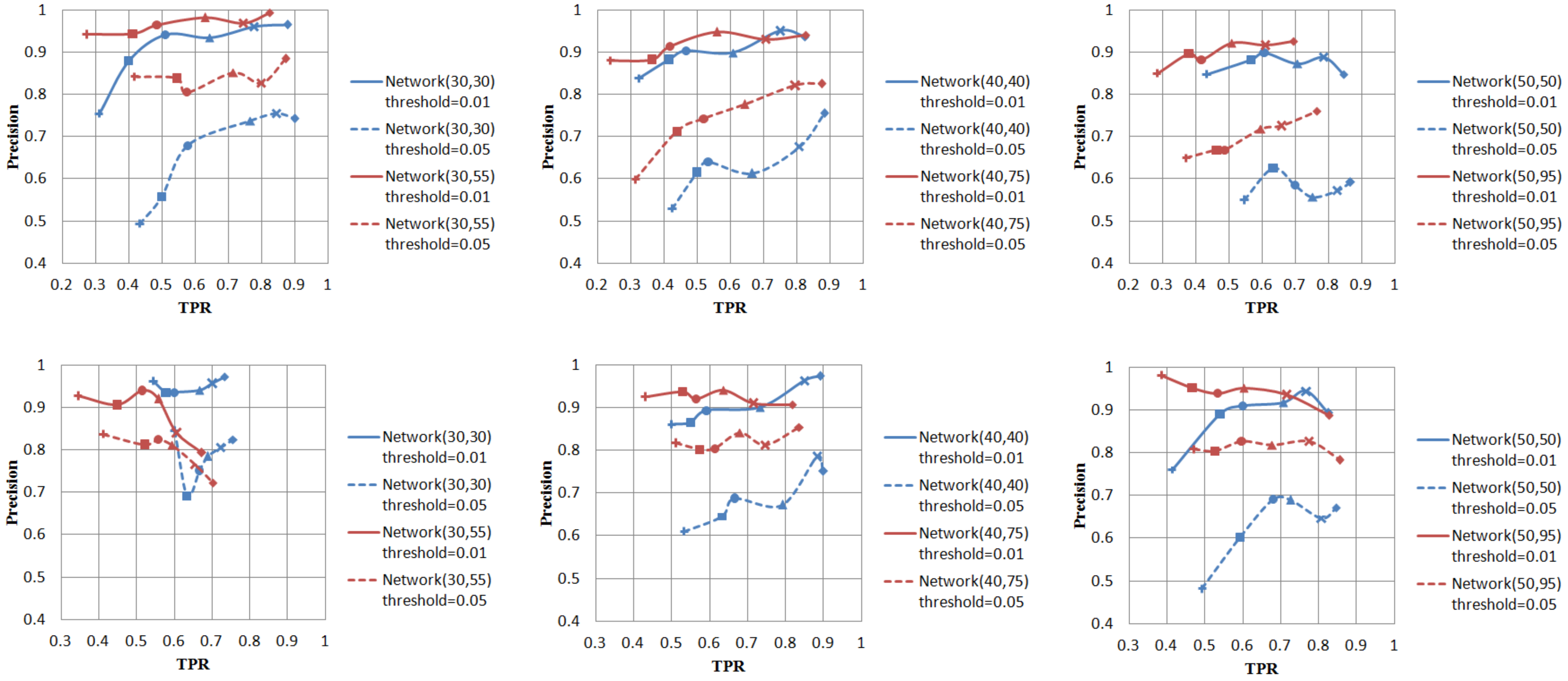


Figure 5.6 ROC curves obtained by the M-H algorithm for the six synthetic random BNs. A uniform colour scheme is applied to all subgraphs to discriminate the curves obtained for a given network with six different sample sizes. Each curve is a smooth connection of nine points obtained at different thresholds for structural feature selection (i.e. 0.1 by 0.1 to 0.9). In particular, the TPR and FPR values obtained at the threshold of 0.5 are marked on each curve (using a square, cross, dot, diamond, plus or triangle marker) for emphasis.

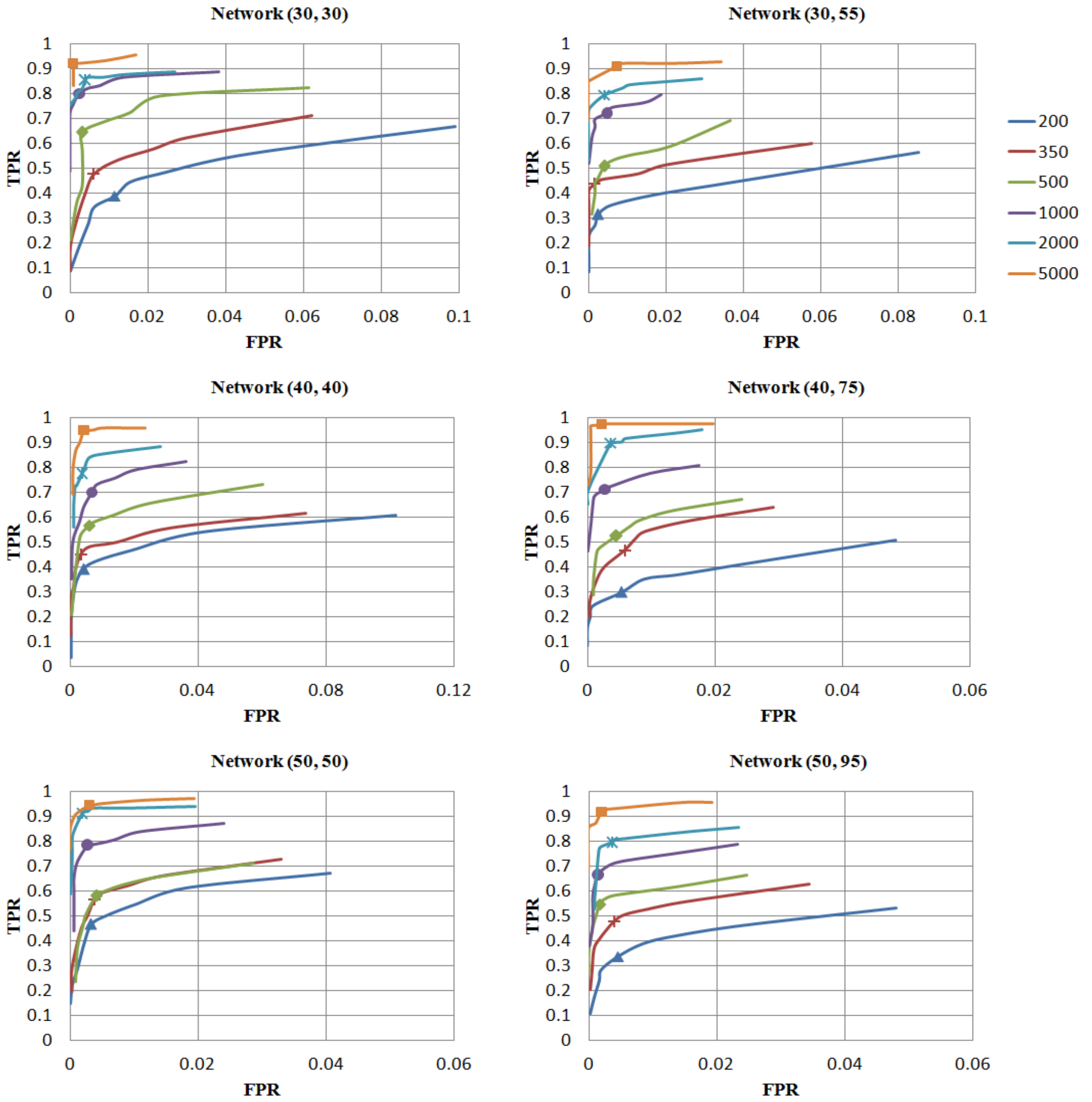


Figure 5.7 ROC curves obtained by the M-H algorithm for the six synthetic scale-free BNs. A uniform colour scheme is applied to all subgraphs to discriminate the curves obtained for a given network with six different sample sizes. Each curve is a smooth connection of nine points obtained at different thresholds for structural feature selection (i.e. 0.1 by 0.1 to 0.9). In particular, the TPR and FPR values obtained at the threshold of 0.5 are marked on each curve (using a square, cross, dot, diamond, plus or triangle marker) for emphasis.

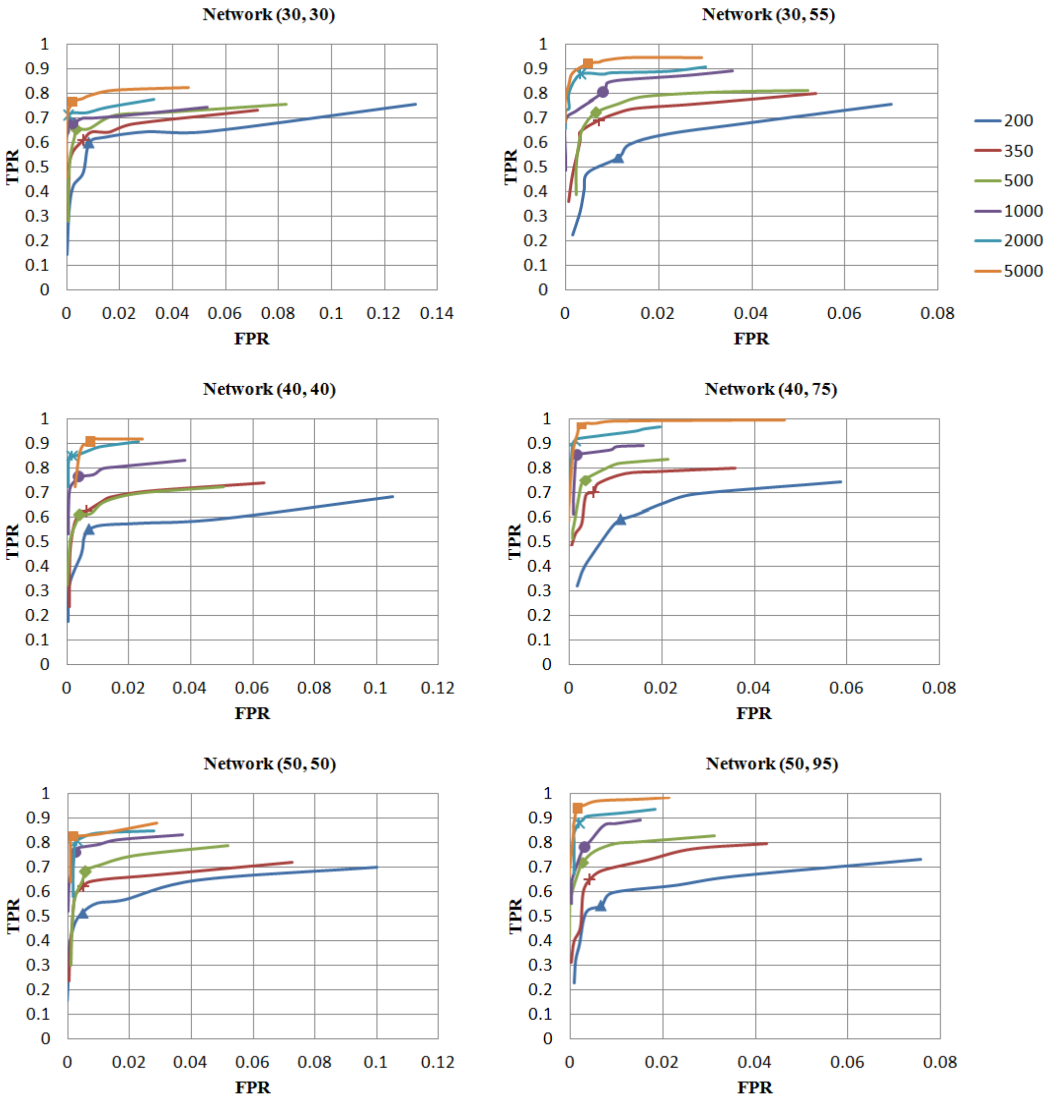


Figure 5.8 Precision-recall curves by the M-H algorithm for the six synthetic random BNs. A uniform colour scheme is applied to all subgraphs to discriminate the curves obtained for a given network with six different sample sizes. Every curve is a smooth connection of nine points obtained at different thresholds for structural feature selection (i.e. 0.1 by 0.1 to 0.9). In particular, the precision and TPR values obtained at the threshold of 0.5 are marked on each curve (using a square, cross, dot, diamond, plus or triangle marker) for emphasis.

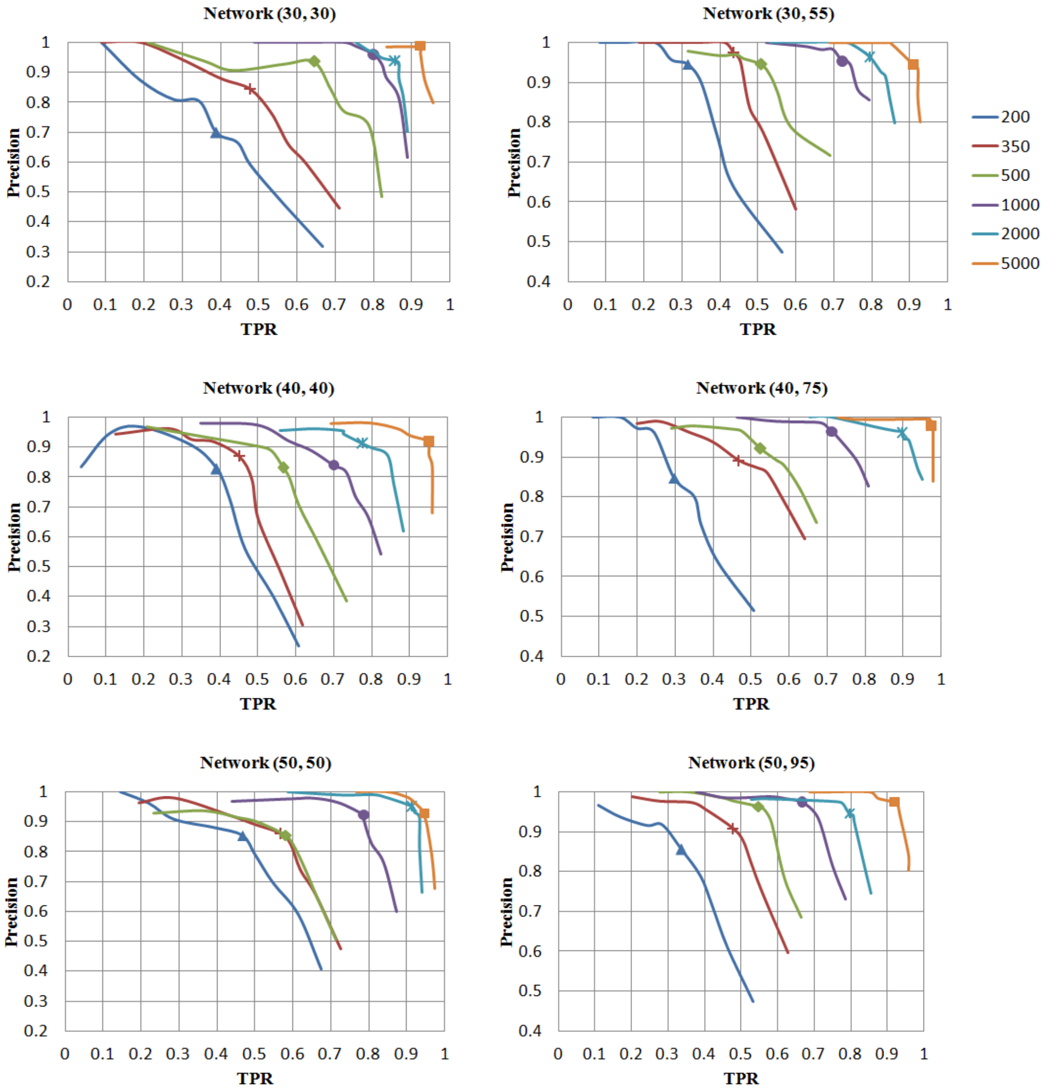


Figure 5.9 Precision-recall curves by the M-H algorithm for the six synthetic scale-free BNs. A uniform colour scheme is applied to all subgraphs to discriminate the curves obtained for a given network with six different sample sizes. Every curve is a smooth connection of nine points obtained at different thresholds for structural feature selection (i.e. 0.1 by 0.1 to 0.9). In particular, the precision and TPR values obtained at the threshold of 0.5 are marked on each curve (using a square, cross, dot, diamond, plus or triangle marker) for emphasis.

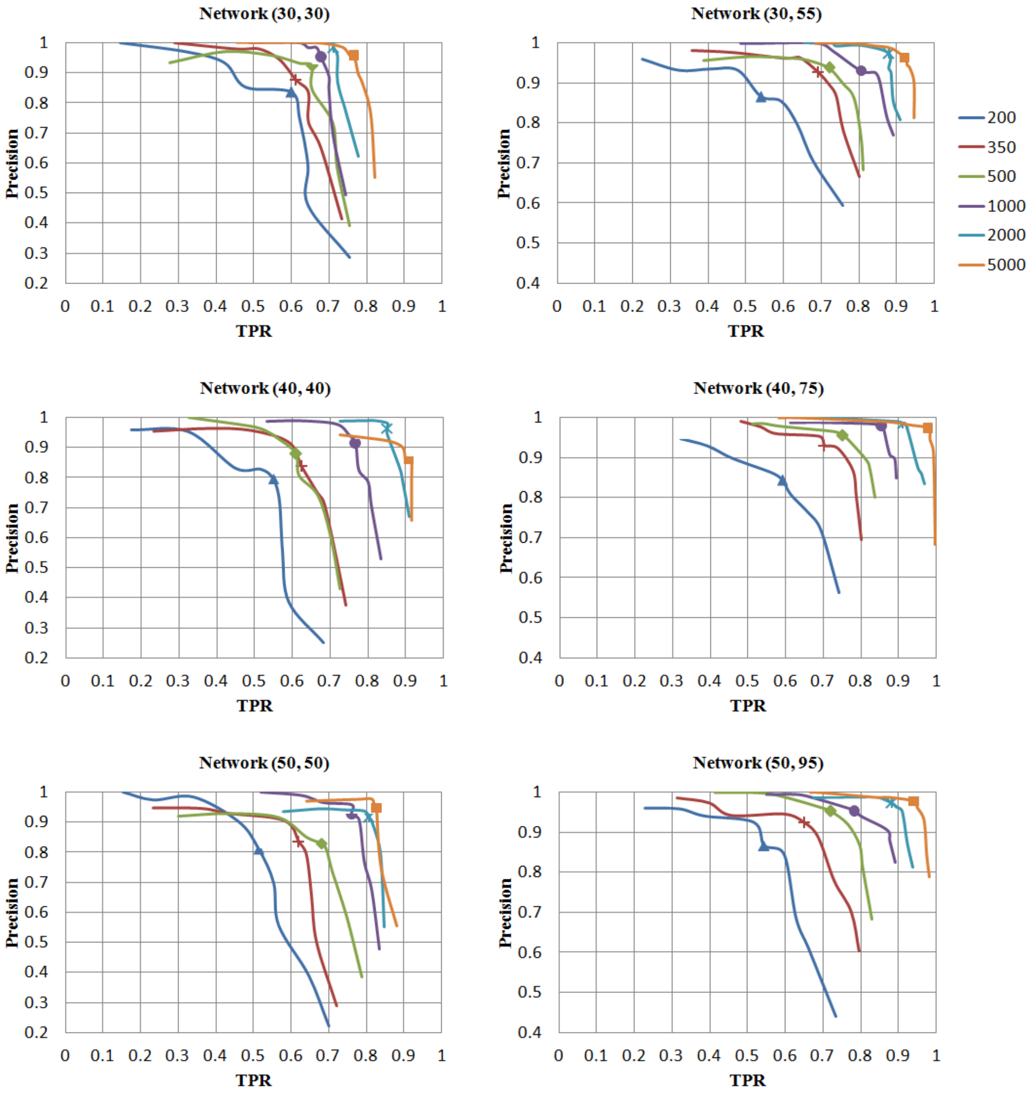


Figure 5.10 The comparison of precision-recall curves obtained by the M-H algorithm (the threshold of structural feature selection set at 0.5) and the PC algorithm (the significance level for conditional independence test set at 0.01) for the six synthetic random BNs. In each subgraph, the plus, square, dot, triangle, cross and diamond markers on each curve represent the precision and TPR values obtained with sample sizes 200, 350, 500, 1000, 2000 and 5000, respectively.

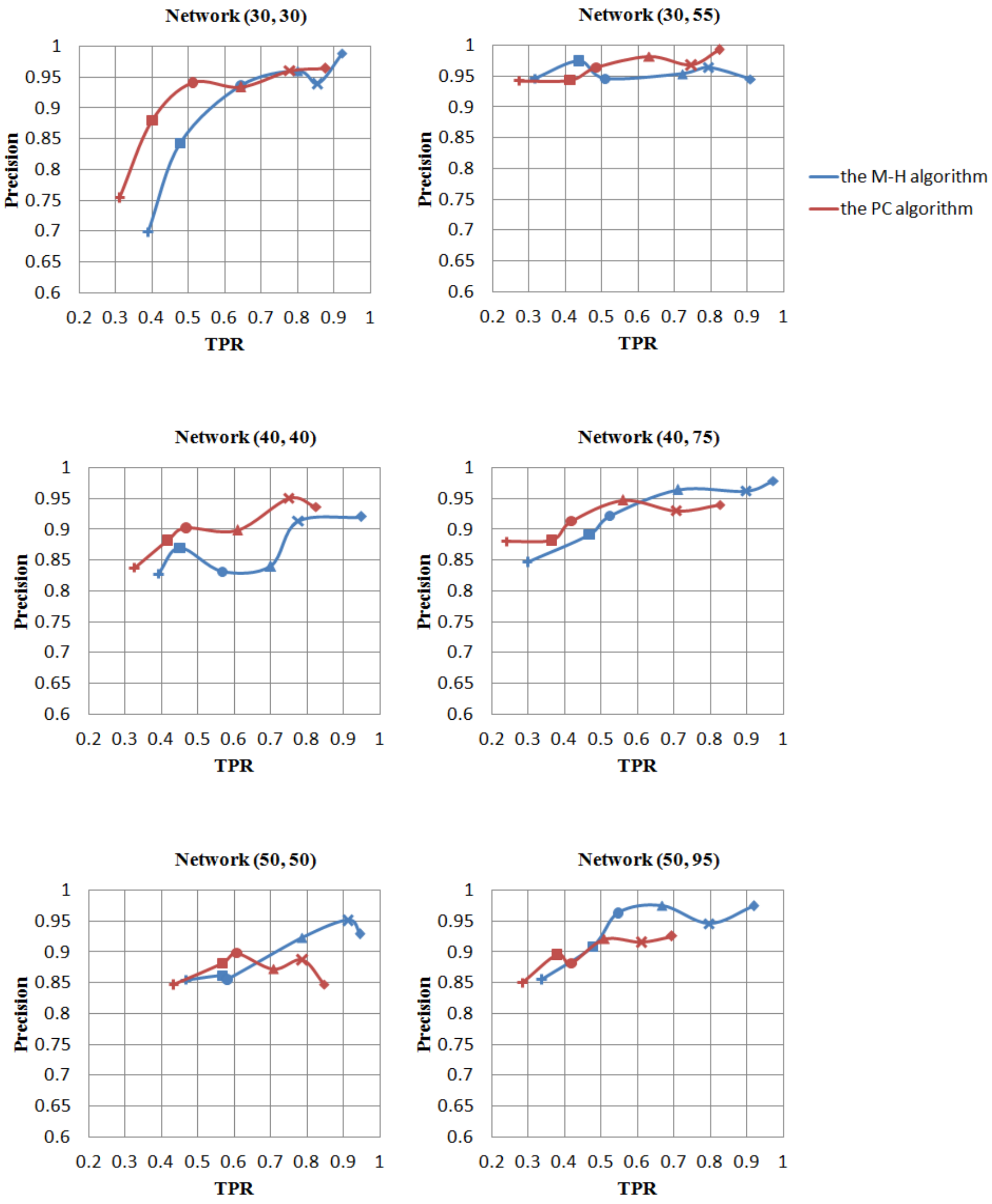
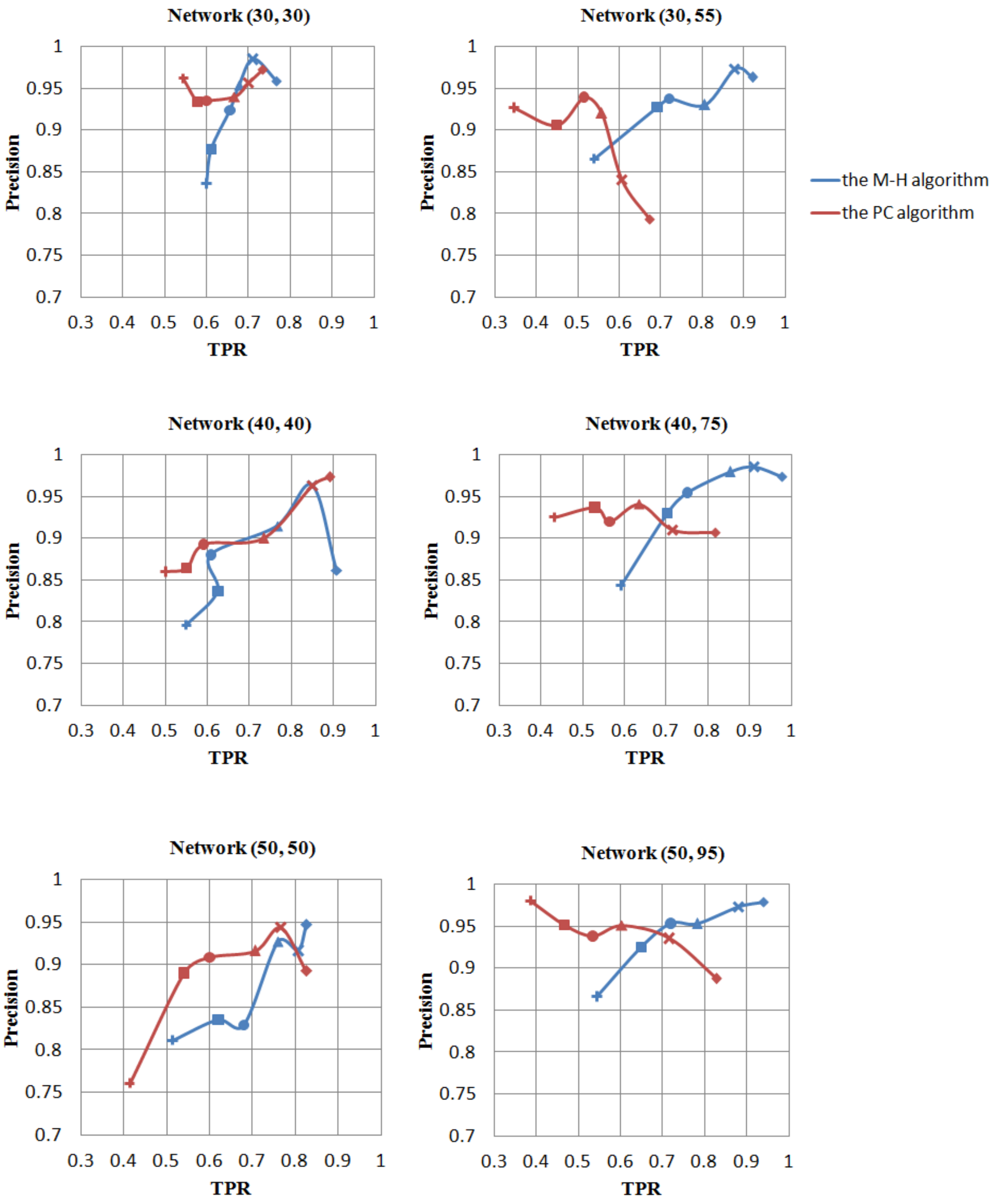


Figure 5.11 The comparison of precision-recall curves obtained by the M-H algorithm (the threshold of structural feature selection set at 0.5) and the PC algorithm (the significance level for conditional independence test set at 0.01) for the six synthetic scale-free BNs. In each subgraph, the plus, square, dot, triangle, cross and diamond markers on each curve represent the precision and TPR values obtained with sample sizes 200, 350, 500, 1000, 2000 and 5000, respectively.



Chapter 6

General discussion

The aim of this thesis is to explore the potential of probabilistic graphical models (PGMs) in systems biology and quantitative genetics. To achieve this goal, we conducted in-depth investigations on two categories of applications, i.e. gene-phenotype network reconstruction (**Chapter 2** and **3**) and linkage map reconstruction (**Chapter 4**). Furthermore, a comparative simulation study of two representative algorithms for structure learning of PGMs indicated that comprehensive evaluation of approaches to reconstructing PGMs would be rather complicated. This is due to the fact that a range of interactive factors, such as the size and complexity of the true networks, the completeness and fidelity of the reconstructed networks, as well as the running time of programs, have to be taken into account (**Chapter 5**). In this final chapter, we will further elaborate and discuss a few concepts, theories and findings related to the content of the thesis.

6.1 A theoretical summary of PGMs

The formalism of PGMs provides a unifying framework for capturing complex dependencies among random variables, and building large-scale multivariate statistical models (Wainwright and Jordan 2008)[1]. PGMs commonly seen in practical applications are categorized into two major types: Bayesian networks (BNs) and Markov random fields (MRFs).

6.1.1 Bayesian networks (BNs)

BNs, also known as Bayes nets or belief networks, are directed acyclic graphs (DAGs) that represent probabilistic causation. In contrast to deterministic causation (i.e. the occurrence of event *A* inevitably leads to event *B*), probabilistic causation characterizes cause-effect (causal) relations from the perspective of probability theory: the occurrence of event *A* increases the probability of event *B*. In statistics, it is generally accepted that observational studies, without further assumptions, can give hints but can never establish cause and effect, because associations do not logically imply probabilistic causation (Glasser 2008; Pearl 2009; Ward 2009). For instance, the observation that smokers are far more susceptible to lung cancer cannot establish smoking as a cause of increased cancer rate, since there may exist a certain genetic defect that causes both lung cancer and a craving for nicotine.

A BN allows a natural factorization of the joint probability distribution (JPD) over all variables in the graph into a set of conditional probability distributions (CPDs). It can be expressed mathematically as $P(X_1,...,X_n)=\prod_{i=1}^{n}P(X_i \mid pa(X_i))$, where $pa(X_i)$ denotes the set of parent variables of X_i. The factorization further implies a set of conditional independence relations in the form of $X_i \perp \sim de(X_i) \mid pa(X_i)$, where $\sim de(X_i)$ denotes the set of non-descendants of X_i. In words, each variable in a BN is conditionally independent of its non-descendants given its parent variables. Nonetheless, a BN is a non-unique representation of a given set of conditional independence relations. There are often many BNs that represent different causal structures but are said to be Markov equivalent (or, belong to the same Markov equivalence class) in the sense that they encode the same conditional independencies. For simplicity's sake, let’s take triplets of variables {*A*, *B*, *C*} as an example. Three different causal structures *A*→*C*→*B*, *A*←*C*←*B* and *A*←*C*→*B* are Markov equivalent since they entail the same set of relationships: (1) *A* and *C* are unconditionally dependent, as are *C* and *B*; (2) *A* and *B* are unconditionally dependent but conditionally independent given *C*. In contrast, the v-structure *A*→*C*←*B*, where *C* has converging arrows from *A* and *B* and there is no direct link between *A* and *B*, does not belong to the same Markov equivalence class because it reveals partially different relationships: (1) *A* and *C* are unconditionally dependent, as are *C* and *B*; (2) *A* and *B* are unconditionally independent but conditionally dependent given *C*. More generally, it has been proved that two BNs are Markov equivalent if and only if they have the same skeleton (the skeleton of a BN refers to the undirected graph resulted from removing the directions of all the edges in the network) and the same v-structures (Verma and Pearl 1991). Considering that only v-structures can be distinguished by the observed patterns of conditional independence and dependence, it is concluded that BNs between, but not within, Markov equivalence classes are distinguishable from one another. That is, causal inference based on conditional independence and dependence facts can actually identify a distinct Markov equivalence class instead of a distinct BN.

6.1.2 Markov random fields (MRFs)

A MRF, also known as a Markov network or undirected graphical model, represents conditional independence relations among a set of variables via graph separation as follows: for three subsets of variables $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, we say $\boldsymbol{C}$ separates $\boldsymbol{A}$ from $\boldsymbol{B}$ (or equivalently, $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$) if every path starting from a node in $\boldsymbol{A}$ and terminating in a node in $\boldsymbol{B}$ passes at least one node in $\boldsymbol{C}$. This is known as the global Markov property (for clarity, hereafter we restrict ourselves to Markov properties on MRFs, which are generally different from Markov properties on DAGs (Lauritzen 1996)). In a MRF, the smallest set of variables that makes a variable X conditionally independent of all other variables is called X's Markov blanket. One can show that, a node's Markov blanket is the set consisting of its immediate neighbors. This is known as the local Markov property, from which we can further deduce that any two non-adjacent variables are conditionally independent given all other variables. This is known as the pairwise Markov property.

Given the three Markov properties mentioned above, it is straightforward to unveil the conditional independencies encoded by a MRF. However, parameterization of MRFs is not as intuitive as that of BNs, since the parameters of MRFs are less interpretable and less modular due to their lack of probabilistic connotation (Koller et al. 2007; Murphy 2012). Specifically, one cannot use the chain rule of probability to factorize a JPD, because there is no topological ordering associated with a MRF (Murphy 2012). Instead of associating a CPD with each node, the practical parameterization method for a MRF is to associate a factor (also referred to as potential function) with each clique in the graph. The JPD is then defined as the normalized product of factors over all the maximal cliques. We will not go further into details on this aspect, as in this thesis we focus on the topological reconstruction of PGMs. For those readers interested in parametric inference in MRFs, please refer to (Koller et al. 2007).

It is obvious that for any MRF and any distribution over the variables, global Markov property implies local Markov property which in turn implies pairwise Markov property. What is less obvious, but nonetheless true, is that the three Markov properties are equivalent for a positive distribution (Koller and Friedman 2009; Murphy 2012). The importance of revealing the equivalence among the three Markov properties is that it is usually easier to empirically assess pairwise conditional independencies, which are commonly exploited to construct MRFs.

A particular type of MRFs, namely Gaussian graphical models (GGMs) or Gaussian Markov random fields (GMRFs), models a multivariate normal distribution in the form of an undirected graph, where nodes denote the variables and edges correspond to the non-zero entries in the precision matrix (i.e. the inverse covariance matrix) of the multivariate normal distribution. From a statistical point of view, the precision matrix indicates sophisticated and subtle relationships between variables: if an entry in the precision matrix is zero, then it means the two variables are conditionally independent given the other variables; furthermore, it means graphically there is no edge connecting the two nodes in the corresponding GGM. This transforms the problem of learning a GGM from observational data into the problem of estimating coefficients in the precision matrix, or alternatively, into the problem of estimating pairwise conditional independence relations among variables.

6.1.3 BNs vs. MRFs

Which type of PGMs has more “expressive power”, BNs or MRFs? The answer is that BNs are neither more nor less expressive than MRFs, as some conditional independence relations can be perfectly modeled by the former but not the latter, and vice versa. To be more specific, BNs are directed and acyclic, so that they are suitable for representing decomposable probabilistic causation; whereas MRFs are undirected, so that they are particular useful in modeling cases where the interactions between variables are symmetrical (e.g. associations) and one thus cannot naturally assign directions to edges in the graphs.

6.2 Proper evaluation of structure learning algorithms of PGMs

When developing and applying approaches to PGMs-based network reconstruction, it is of overriding importance to: (1) determine whether there exists a structure in the research problem for the algorithms to learn, since the answer can give an indication of how learnable the problem is; (2) test multiple structure leaning algorithms against the chosen performance measures, as this can tell which algorithms are worth tuning and which algorithms should not be considered further.

The theoretically ideal way to evaluate and compare the performance of different structure learning algorithms (as we have adopted in **Chapter 5**) is to generate synthetic data from an existing network and subsequently retrieve the network by applying the algorithms to the data. A major advantage of this approach is that each of the resulting networks obtained by the algorithms can be compared with the gold standard, i.e. the benchmark network used for synthetic data generation. But a critical element of this approach is the choice of performance measures that quantify the distance between an obtained network and the benchmark network: the smaller the distance, the better the algorithm performs in network reconstruction (de Jongh and Druzdzel 2009). The commonly used performance measures are: (1) recall, which represents the ratio of true positives to actual positives (i.e. the sum of true positives and false negatives); (2) precision, which represents the ratio of true positives to predicted positives (i.e. the sum of true positives and false positives). To put it in another way, recall measures the completeness whereas precision measures the fidelity of the reconstruction. In practice, it is easy to achieve either high recall or high precision, but rarely both simultaneously (see examples shown in **Figure 5.8** and **5.9**). For a given benchmark network, conservative algorithms will return structures with higher precision but lower recall; while liberal algorithms will return structures with higher recall but lower precision. For a given structure learning algorithm, higher recall but lower precision are obtained for denser networks; while higher precision but lower recall are obtained for sparser networks (Oyen et al. 2013). This shows that the optimal tradeoff between recall and precision really depends on the nature of the problem domain and what one aims to achieve.

As indicated above, the practical performance of a structure learning algorithm is subject to the application-specific circumstance, namely the topology of the benchmark/target network. The topology of a network mainly refers to two aspects: the size and the connectivity of a network. By size we simply mean the number of nodes in the network; with connectivity, we are talking about how easy or difficult it is for any two nodes to form a connection. These two factors together play a large part in how the network is analyzed and interpreted. Think about the modeling of a small rural community where residents are familiar with each other vs. the modeling of an urban metropolis where people usually know each other through complicated interpersonal relationships. Notably, a number of networks in nature, including some biological networks, exhibit a high degree of modularity. Networks with high modularity typically have dense connections between nodes within the same module but sparse connections between nodes belonging to different modules. Such heterogeneous connectivity makes structure learning algorithms that are based on global optimization suffer from a resolution limit, since it is hard to elaborate relatively small modules (Lancichinetti and Fortunato 2011; Nicolini and Bifone 2016). Given this, it is worthwhile to further expand our study in **Chapter 5** to investigate particularly the resolution limits of the two BN structure learning algorithms in the reconstruction of scale-free networks.

Moreover, a few nontrivial points need to be stressed. When causality come into play, as in the reconstruction of (multilayer) causal phenotype networks (**Chapter 2** and **3**), the orientation of edges becomes extremely important. Contrariwise, when establishing causal relations is not the essence of the task, one can be more relaxed about edge orientations (de Jongh and Druzdzel 2009). When real-time performance of an algorithm is of importance to the research problem (though fairly rare in the domain of biology), the runtime of the program has to be taken into account as well.

6.3 Practical problems in application of PGMs to systems biology

Systems biology is an inter-disciplinary field that integrates biology, computer science and engineering to decipher complex biological systems using holistic approaches (Calvert and Fujimura 2009). It is based on the understanding that the functions of a whole living organism are more than the sum of its parts (Hurlbut 2006). Accordingly, it requires the ability to obtain, integrate and analyze complex data sets from multiple sources.

The rapid evolution of high-throughput technologies, such as nucleotide sequencing, DNA-chips and protein mass spectrometry, have enabled extensive generation of multi-omics data. But, these data are typically heterogeneous and distributed in various databases, since they come from studies driven by different objectives and conducted on different platforms. This raises challenges in data access and integration. Although the acquisition of publicly available sources has been largely facilitated thanks to the current data explosion, discovering the appropriate data is often not straightforward due to the diversity of data types and formats. Besides, experimental biologists are still struggling to provide complete and non-redundant information collected from varied data sources. For instance, Pathguide has reported by 2013 a list of 547 biological pathways and molecular interaction related resources. These resources are not simply complementary, but often define similar signaling and metabolic pathways with different boundaries and components (Gomez-Cabrero et al. 2014).

Challenges beyond data access and integration lie in the integrative analysis of multi-omics data. A number of factors including data quality, the complexity of the target system and the characteristic of the technology employed come together to make integrative analysis not an easy job. It is notable that most of the systematic approaches developed so far are pipelines of analysis that apply several methods to carry out a sequence of tasks (Bersanelli et al. 2016). An example is the genotype-phenotype modeling scheme we have proposed in **Chapter 3**. Encouragingly, pipelines presented for addressing a particular problem can also be used, with minor modifications, to solve another problem, possibly with other types of omics (Bersanelli et al. 2016). For instance, although in **Chapter 3** we have only demonstrated that the proposed scheme is effective in inferring directed associations among metabolites and sensory traits given relevant QTLs, this scheme should also be applicable to the modeling of general hierarchical networks that represent multilevel phenotypic responses to DNA variations.

Identifying associations among entities within and across heterogeneous data sources is of great importance in most studies, as it is a straightforward and effective way to “glue” together pieces of information so as to provide a coherent view of the whole system. To establish multilevel associations, earlier studies often employed distance-based or correlation-based metrics, while recent studies tend to adopt more sophisticated modeling techniques such as PGMs. Whichever method is used, the conflicts between data measures must be handled beforehand. This includes missing-data imputation, data scaling, discretization, normalization, standardization, and etc. Nonetheless, even if data pre-processing has been done properly, it is noteworthy but often-overlooked that very few associations across heterogeneous data sets are usually revealed, compared to the great number of associations identified within the same data set (see, for example, Fig.4 in **Chapter 3**).

A rich body of literature supports the idea that associations in observational data can provide insights into causal relations among the measured variables (Blair et al. 2012; Pearl 2009; Shipley 2016). Nonetheless, it has been seen that causal relations are sensitive to subtle association patterns, which may be driven by other factors (e.g. environmental and experimental design factors) that do not reflect the underlying biological nature (Blair et al. 2012). In addition, graphical methods for causal inference from observational data, especially from observed associations, are admittedly subject to the existing theoretical constraints. As elaborated in **Chapter 5** and **Section 6.1**, constructing BNs on the basis of the BDe metric or conditional independence facts can end up with a distinct Markov equivalence class rather than a distinct BN. Also, **Chapter 2** has demonstrated that inferring causal relations from observed associations requires introducing extra known causal factors to the measured variables. For instance, causal inference in correlated traits (or equivalently, the construction of directed phenotype networks) is based upon logic that involves the underlying QTLs. The existing related algorithms request at least one unique QTL for each trait studied, though such prerequisite is hardly being met in reality. In comparison, the QPSO algorithm presented in **Chapter 2** is of more practical significance since it has a more realistic prerequisite – some traits can come without QTL. More encouragingly, as indicated in **Chapter 2** and **3**, the QPSO algorithm can be embedded into a bottom-up strategy to systematically model multilevel phenotypic responses to DNA variations.

Agreement between a mathematical or statistical model and the true underlying biology is vital to any practical study. It should be recognized that the extent to which a PGM derived from observational data can recapitulate the architecture of an underlying biological process is not yet well understood (Blair et al. 2012). On the one hand, observational data are often collected at single time points; on the other hand, biological processes typically display time varying dynamics. This conflict makes the interpretation of the reconstructed model challenging. Dynamic Bayesian networks (DBNs) are the time-generalization of BNs and relate variables to each other over discrete time points. Their major advantages lie in the ability to deal with multivariate time series data and permit presentation of cyclic causal relationships. Here, however, we will not discuss practical issues related to the use of DBNs, since analyzing time series data is beyond the scope of this thesis. For those who are interested in the computational power of DBNs, please refer to Ghahramani (1998), Murphy (2002) and Brulé (2016) for details.

6.4 Practical problems in application of PGMs to linkage mapping

Chapter 4 shows that PGMs are of great potential in the construction of linkage maps as they can achieve marker filtering and ordering simultaneously. By filtering, we mainly refer to the filter of markers with high genotyping error rates. Thanks to the natural structure of genetic markers, the topologies of PGMs reconstructed from sets of marker data are often nearly linear. But, for markers that are genetically very similar, the obtained PGMs expand horizontally instead of vertically, since there is no obvious clue to the linearity of those markers (see example in **Figure 4.3b**). For this reason, it is often the case that the ordering of markers by means of PGMs is susceptible to ambiguities in chromosomal regions with high marker density. To evade such ambiguities, we have proposed in **Chapter 4** a frequentist diagonal ordering algorithm, which serves as a complement to PGMs for fine-ordering of markers. Please note that after the use of PGMs followed by a frequentist diagonal ordering, markers that are pulled aside from the resulting linear linkage map belong to one of the following two categories: first, they are genetically close to one another; second, they have high genotyping error rates (see example in **Figure 4.4**).

6.5 Tips on the use of PGMs in the reconstruction of biological networks and linkage maps

1) As mentioned previously, PGMs of different natures apply in different circumstances: MRFs are ideal for identifying and representing conditional independence among measured variables; BNs are suitable for modeling of decomposable probabilistic causation.

2) Various algorithms have been presented for the modeling of PGMs. Each algorithm has its own pros and cons. The selection of the appropriate algorithm depends on the nature (e.g. size and complexity) of the target system and also the requirement (e.g. accuracy and completeness) of the reconstruction (see **Chapter 5**).

3) Most of the methods proposed for integrative analysis of multi-omics data employ a sequence of procedures. These methods, with minor changes to certain internal procedures, are often applicable to other analogous problems.

4) Networks learnt from multi-omics data typically show hierarchical structures. Notably, however, the interactions inferred from heterogeneous data are generally far less than those inferred from homogeneous data (see **Chapter 3**).

5) Detected associations may be further directed by means of introducing external causal factors to the measured variables. For instance, causal inference in correlated traits (or equivalently, the construction of directed phenotype networks) is based upon logic that involves underlying QTLs. Furthermore, the proposition of the QPSO algorithm indicates that the external causal factors are not necessarily required for each and every variable to fulfil the causal inference (see **Chapter 2**).

6) Causal inference is vulnerable to external influences such as experimental and analytical design factors. The inferred causal relations should be interpreted carefully, since some of them might be pseudo and some of them are simplified, but incomplete, representations of more complex causal flows (see **Chapter 3**).

7) In addition to biological network reconstruction, PGMs can also construct high-quality linkage maps in the face of data perturbations caused by genotyping errors and reciprocal translocations (see **Chapter 4**).

References

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L (2016) Methods for the integration of multi-omics data: mathematical aspects. BMC bioinformatics 17:167

Blair RH, Kliebenstein DJ, Churchill GA (2012) What can causal networks tell us about metabolic pathways? PLoS Comput Biol 8:e1002458

Brulé J (2016) The Computational Power of Dynamic Bayesian Networks. arXiv preprint arXiv:160306125

Calvert J, Fujimura JH (2009) Calculating life? EMBO reports 10:S46-S49

de Jongh M, Druzdzel MJ (2009) A comparison of structural distance measures for causal Bayesian network models. Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science series:443-456

Ghahramani Z (1998) Learning dynamic Bayesian networks. Adaptive processing of sequences and data structures. Springer, pp 168-197

Glasser SP (2008) Essentials of clinical research. Springer

Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J (2014) Data integration in the era of omics: current and future challenges. BMC systems biology 8:1

Hurlbut WB (2006) Framing the future: embryonic stem cells, ethics and the emerging era of developmental biology. Pediatric research 59:4R-12R

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press

Koller D, Friedman N, Getoor L, Taskar B (2007) 2 Graphical Models in a Nutshell. Statistical Relational Learning:13

Lancichinetti A, Fortunato S (2011) Limits of modularity maximization in community detection. Physical review E 84:066122

Lauritzen SL (1996) Graphical models. Clarendon Press

Murphy KP (2002) Dynamic bayesian networks: representation, inference and learning. University of California, Berkeley

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press

Nicolini C, Bifone A (2016) Modular structure of brain functional networks: breaking the resolution limit by Surprise. Scientific reports 6

Oyen D, Niculescu-Mizil A, Ostroff R, Stewart A, Clark VP (2013) Controlling the precision-recall tradeoff in differential dependency network analysis. arXiv preprint arXiv:13072611

Pearl J (2009) Causality. Cambridge university press

Shipley B (2016) Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R. Cambridge University Press

Verma TS, Pearl J (1991) Equivalence and synthesis of causal models. Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence, pp 220-227

Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning 1:1-305

Ward AC (2009) The role of causal criteria in causal inferences: Bradford Hill's. Epidemiologic Perspectives & Innovations 6:2

Summary

Probabilistic graphical models (PGMs) offer a conceptual architecture where biological and mathematical objects can be expressed with a common, intuitive formalism. This facilitates the joint development of statistical and computational tools for quantitative analysis of biological data. Over the last few decades, procedures based on well-understood principles for constructing PGMs from observational and experimental data have been studied extensively, and they thus form a model-based methodology for analysis and discovery. In this thesis, we further explore the potential of this methodology in systems biology and quantitative genetics, and illustrate the capabilities of our proposed approaches by several applications to both real and simulated omics data.

In quantitative genetics, we partition phenotypic variation into heritable, genetic, and non-heritable, environmental, parts. In molecular genetics, we identify chromosomal regions that drive genetic variation: quantitative trait loci (QTLs). In systems genetics, we would like to answer the question of whether relations between multiple phenotypic traits can be organized within wholly or partially directed network structures. Directed edges in those networks can be interpreted as causal relationships, causality meaning that the consequences of interventions are predictable: phenotypic interventions in upstream traits, *i.e.* traits occurring early in causal chains, will produce changes in downstream traits. The effect of a QTL allele can be considered to represent a genetic intervention on the phenotypic network. Various methods have been proposed for statistical reconstruction of causal phenotypic networks exploiting previously identified QTLs. In **chapter 2**, we present a novel heuristic search algorithm, namely the QTL+phenotype supervised orientation (QPSO) algorithm, to infer causal relationships between phenotypic traits. Our algorithm shows good performance in the common, but so far uncovered case, where some traits come without QTLs. Therefore, our algorithm is especially attractive for applications involving expensive phenotypes, like metabolites, where relatively few genotypes can be measured and population size is limited.

Standard QTL mapping typically models phenotypic variations observable in nature in relation to genetic variation in gene expression, regardless of multiple intermediate-level biological variations. In **chapter 3**, we present an approach integrating Gaussian graphical modeling (GGM) and causal inference for simultaneous modeling of multilevel biological responses to DNA variations. More specifically, for ripe tomato fruits, the dependencies of 24 sensory traits on 29

metabolites and the dependencies of all the sensory and metabolic traits further on 21 QTLs were investigated by three GGM approaches including: (i) lasso-based neighborhood selection in combination with a stability approach to regularization selection, (ii) the PC-skeleton algorithm and (iii) the Lasso in combination with stability selection, and then followed by the QPSO algorithm. The inferred dependency network which, though not essentially representing biological pathways, suggests how the effects of allele substitutions propagate through multilevel phenotypes. Such simultaneous study of the underlying genetic architecture and multifactorial interactions is expected to enhance the prediction and manipulation of complex traits. And it is applicable to a range of population structures, including offspring populations from crosses between inbred parents and outbred parents, association panels and natural populations.

In **chapter 4**, we report a novel method for linkage map construction using probabilistic graphical models. It has been shown that linkage map construction can be hampered by the presence of genotyping errors and chromosomal rearrangements such as inversions and translocations. Our proposed method is proven, both theoretically and practically, to be effective in filtering out markers that contain genotyping errors. In particular, it carries out marker filtering and ordering simultaneously, and is therefore superior to the standard post-hoc filtering using nearest-neighbour stress. Furthermore, we demonstrate empirically that the proposed method offers a promising solution to genetic map construction in the case of a reciprocal translocation.

In the domain of PGMs, Bayesian networks (BNs) have proven, both theoretically and practically, to be a promising tool for the reconstruction of causal networks. In particular, the PC algorithm and the Metropolis-Hastings algorithm, which are representatives of mainstream methods to BN structure learning, are reported to have been successfully applied to the field of biology. In view of the fact that most biological systems exist in the form of random network or scale-free network, in **chapter 5** we compare the performance of the two algorithms in constructing both random and scale-free BNs. Our simulation study shows that for either type of BN, the PC algorithm is superior to the M-H algorithm in terms of timeliness; the M-H algorithm is preferable to the PC algorithm when the completeness of reconstruction is emphasized; but when the fidelity of reconstruction is taken into account, the better one of the two algorithms varies from case to case. Moreover, whichever algorithm is adopted, larger sample sizes generally permit more accurate reconstructions, especially in regard to the completeness of the resulting networks.

Finally, **chapter 6** presents a further elaboration and discussion of the key concepts and results involved in this thesis.

Acknowledgements

The completion of this thesis has been made possible by the support and encouragement given by a number of colleagues, friends and family. I would like to take this opportunity to express my sincere gratitude to those who have contributed to this work.

First of all, I would like to thank my supervisor, Prof. Dr. Fred van Eeuwijk, for not only offering me the opportunity to study in his group but also being a great mentor. Fred has led me into the field of biostatistics and plant science. He encouraged me to pick up challenging research topics and motivated me to constantly pursue “more and better” - a profound understanding of the problem. His rigorous academic attitude and solid professional knowledge have deeply inspired me as a postgraduate student. His constructive guidance and suggestion has benefited me a lot in all these years of my PhD. I will keep seeking his valuable advice in my future professional career.

Secondly, I would like to say a big thank you to my co-supervisor, Dr. Hans (Johannes) Jansen, for his valuable time and efforts in this work. Despite his physical discomfort, Hans has always been kind and patient to discuss with me the various problems encountered in theory and practice and, more importantly, provided me with fruitful ideas. I also highly appreciate his very positive reference letter that helped me to win the job offer from CIMR, University of Cambridge.

I would also like to express my sincere appreciation to the people from Biometris. Those highly knowledgeable staffs, including Cajo, Joao, Willem, Chaozhi, Laura, Patricia, Marcos, Marco, Martin, Gerrit, Sabine and Maaike, have given me a lot of professional support. My fellow PhD students, including Thomas, Tahira, Alba, Daniela, Nadia, George, Nurudeen, Apri, Santosh and Meiyin, have shared so much experience with me. Without these people around, I would have never been able to complete the thesis.

I am very grateful to my friends in Wageningen. Yunlin, Siyu, Ting, Tao, Yijin, Xiaomei and many others, you have enriched my amateur life and your friendships mean a lot to me.

Lastly, my utmost thanks go to all the families for their unconditional love and support. To my beloved parents: thank you so much for always being there for me, giving me spiritual and material support, and in particular taking great efforts to help me raise the two kids. To my dear husband: it is your love, tolerance, unwavering support and encouragement to walk me through the very difficult phases of my PhD study. To my lovely son and daughter: Thank you for coming into my life. Your beautiful smiling faces and sweet kisses always free me from distress, even on my worst days.

List of publications

1. **Wang H.**, van Eeuwijk F. (2014) A new method to infer causal phenotype networks using QTL and phenotypic information. PLoS One 9(8), e103997
2. Xiao, D., **Wang, H.**, Basnet, R. K., Zhao, J., Lin, K., Hou, X., & Bonnema, G. (2014). Genetic dissection of leaf development in Brassica rapa using a genetical genomics approach. Plant physiology, 164(3), 1309-1325.
3. **Wang H.**, Paulo J., Kruijer W., Boer M., Jansen H., Tikunov Y., Usadel B., van Heusden S., Bovy A., van Eeuwijk F. (2015) Genotype–phenotype modeling considering intermediate level of biological variation: a case study involving sensory traits, metabolites and QTLs in ripe tomatoes. Mol Biosyst 11:3101–3110
4. **Wang, H.**, van Eeuwijk, F. A., & Jansen, J. (2016). The potential of probabilistic graphical models in linkage map construction. Theoretical and Applied Genetics, 1-12.
5. Schütte, J., **Wang, H.**, Antoniou, S., Jarratt, A., Wilson, N. K., Riepsaame, J., ... & Hannah, R. L. (2016). An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. Elife, 5, e11469.
6. Lim, C. Y., **Wang, H.**, Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., & Göttgens, B. (2016). BTR: training asynchronous Boolean models using single-cell expression data. BMC bioinformatics, 17(1), 355.

Curriculum vitae

Huange Wang was born on 30 January 1983 in Xi’an, China. She obtained her BSc degree in Electrical Engineering and Automation from Xi’an Jiaotong University, China, in 2005. Later that year, she commenced postgraduate study in Systems Engineering at Northwestern Polytechnical University, China. In 2007, she won a two-year scholarship offered by China Scholarship Council to pursue her theoretical study on Bayesian network learning and inference in Cranfield University, UK, where she was awarded an MSc degree in Applied and Computational Mathematics. From October 2009, she started her PhD research on the topic of reconstructing biological networks using probabilistic graphical models at Biometris, Wageningen University, Netherlands. She worked as a research associate in the Göttgens group, Cambridge Institute for Medical Research, University of Cambridge, from 2014 till 2016. Currently, she is enjoying the newborn stage of her second baby.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (6 ECTS)

- Review of reconstruction methods for biological networks

Writing of project proposal (3 ECTS)

- Reconstruction of biological networks using graphical modelling approaches

Post-graduate courses (6.3 ECTS)

- Uncertainty propagation in spatial and environmental modelling; PE&RC (2011)
- Identity by descent (IBD) approaches to genomic analyses of genetic traits; Biometris, WUR (2012)
- Pattern recognition; Netherlands Bioinformatics Centre (NBIC) (2013)
- Analysis of linear mixed models by ASREML-R with applications in plant breeding; Biometris, WUR & VSN International (2013)

Laboratory training and working visits (0.3 ECTS)

- Network reconstruction of QTLs and Phenotypic traits; Groningen Bioinformatics Centre (GBIC) (2012)

Deficiency, refresh, brush-up courses (3.6 ECTS)

- Biomolecular principles of the cell; Netherlands Consortium for Systems Biology (NCSB) (2011)
- EBI Roadshow: Databases, Sequence Alignment, ArrayExpress & Ensembl; NBIC Education (2013)

Competence strengthening / skills courses (3.6 ECTS)

- Scientific publishing; WGS (2012)
- Scientific writing; Language services, WUR (2012)
- Career perspectives; WGS (2013)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC Weekend (2011)
- PE&RC Day (2012)

Discussion groups / local seminars / other scientific meetings (6.5 ECTS)

- Biometris colloquium (2009-2014)

International symposia, workshops and conferences (5.4 ECTS)

- The XVIth Meeting of the EUCARPIA Section Biometrics in Plant Breeding; Stuttgart, Germany (2012)
- STATSEQ Meeting on Gene Network Inference with Systems Genetic Data and Beyond; Paris, France (2013)
- The 4th Channel Network Conference; St. Andrews, United Kingdom (2013)