BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 246–254
doi:10.1093/bioinformatics/bts679

Systems biology

Advance Access publication November 21, 2012

An effective framework for reconstructing gene regulatory networks from genetical genomics data

R. J. Flassig[1], S. Heise[1], K. Sundmacher[1,2] and S. Klamt[1,*]

[1]Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1 and [2]Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Systems Genetics approaches, in particular those relying on genetical genomics data, put forward a new paradigm of large-scale genome and network analysis. These methods use naturally occurring multi-factorial perturbations (e.g. polymorphisms) in properly controlled and screened genetic crosses to elucidate causal relationships in biological networks. However, although genetical genomics data contain rich information, a clear dissection of causes and effects as required for reconstructing gene regulatory networks is not easily possible.

Results: We present a framework for reconstructing gene regulatory networks from genetical genomics data where genotype and phenotype correlation measures are used to derive an initial graph which is subsequently reduced by pruning strategies to minimize false positive predictions. Applied to realistic simulated genetic data from a recent DREAM challenge, we demonstrate that our approach is simple yet effective and outperforms more complex methods (including the best performer) with respect to (i) reconstruction quality (especially for small sample sizes) and (ii) applicability to large data sets due to relatively low computational costs. We also present reconstruction results from real genetical genomics data of yeast.

Availability: A MATLAB implementation (script) of the reconstruction framework is available at www.mpi-magdeburg.mpg.de/projects/cna/etcdownloads.html

Contact: klamt@mpi-magdeburg.mpg.de

Received on June 26, 2012; revised on September 27, 2012; accepted on November 19, 2012

1 INTRODUCTION

Systems genetics seeks to reveal complex genetic interactions in biological systems by relating genetic variations to various phenotypic data from high-throughput measurements (Jansen and Nap, 2001; Jansen, 2003; Rockman and Kruglyak, 2006). In contrast to the classical ‘one gene perturbation at a time’ approach, systems genetics interprets naturally occurring multiple genetic variations [e.g. single nucleotide polymorphisms (SNP)] as multi-factorial perturbations from which causalities can be unraveled more efficiently (Jansen and Nap, 2001; Jansen, 2003). Systems genetics methods use properly controlled genetic crosses (segregating populations) such as recombinant congenic strains (RCS), recombinant inbred lines (RIL), advanced intercross lines (AIL) or chromosome substitution strains (CSS) to causally link genetic or chromosomal regions to observed phenotypic trait data (Jansen, 2003; Rockmann, 2008). Identifying a chromosomal region [the Quantitative Trait Locus (QTL)] that influences a certain phenotypic trait is known as QTL mapping.

In genetical genomics, a particular subclass of systems genetics, gene-expression levels are considered as phenotypic traits (called etraits), and identified QTLs (comprising single genes or gene regions) are referred to as expression-QTLs (eQTLs). One application of eQTL maps obtained from genetical genomics approaches is the reconstruction of gene regulatory networks (GRN). In the latter, nodes represent genes and edges represent interactions or dependencies between genes. GRN provide the basis for systems-level understanding of interacting genes and phenotype formation in living systems. They condense different types of molecular interactions on the signaling, metabolic and genetic level to a network representation of causalities. Therefore, GRN represent a causal projection of gene activities, neglecting detailed molecular mechanisms (Brazhnik *et al*., 2002; de la Fuente, 2010). Reconstructed gene regulatory networks can be used to narrow down genetic analysis by massively reducing the number of potential molecular interactions or locations of interaction sites. In the same way, GRN can be used to identify putative intervention points by relating genetic spots to pathologic phenotypes (Schadt *et al*., 2005).

According to Liu *et al*. (2010), the general GRN reconstruction pipeline for genetical genomics data consists of three major steps: (i) eQTL mapping, (ii) candidate regulator selection and (iii) network refinement. For step (i), there exist several eQTL mapping strategies: single-etrait-single-eQTL, multiple-etrait-single-eQTL, multiple-etrait-multiple-eQTL. A detailed review on the pipeline of eQTL mapping is, for instance, given by Michaelson *et al*. (2009). In step (ii), the eQTL map in combination with a genetic map is used to select single candidate (regulator) genes from the eQTLs as the latter often represent chromosomal regions due to genetic linkage. Methods used include conditional correlation (Bing and Hoeschele, 2005; Keurentjes *et al*., 2007), local regression (Liu *et al*., 2008) or analysis of between-strains SNPs (Li *et al*., 2005). In the third step (iii), network refinement methods are used to the topology obtained in step (ii), e.g., with the goal to identify and eliminate (false positive) edges arising from indirect effects. Here, Bayesian network approaches (Zhu *et al*., 2007) and structural equation modeling, SEM, (Liu *et al*., 2008) have been used. One disadvantage of the former is, that Bayesian networks cannot handle cycles if no time dimension is included (Friedman *et al*., 1998).

*To whom correspondence should be addressed.

Further, inferred dependencies in Bayesian networks do not necessarily represent causalities, as there may exist several alternative dependencies having the same joint probability (Pearl, 2000). As one approach for structural equation modeling, Liu *et al.* (2008) construct first an encompassing directed network from the eQTL analysis which is afterwards further refined by SEM. Similar as Baysian network inference, SEM is computationally expensive, which restricts the applicability to medium-sized networks.

In this work, we propose a novel strategy for reconstructing GRN from genetical genomics data. The workflow (Fig. 1; explained in detail in Section 2) encompasses implicitly the three major steps mentioned above but uses different techniques than used so far. The chosen methods were intentionally kept simple; realistic datasets show that they are nevertheless effective (Section 3). The initial GRN is constructed based on genotype–phenotype and phenotype–phenotype correlation analysis. Owing to genetic linkage, there are often groups of genetically adjacent regulator candidates that target the same gene resulting into eQTLs. To avoid a lot of false positive interactions, single candidate regulators are therefore identified from the eQTLs. Finally, as a method for network refinement, indirect path effects are removed by TRANSWESD, a recently introduced transitive reduction approach (Klamt *et al.*, 2010). Originally, TRANSWESD has been developed to prune perturbation graphs derived from single knock-out data. Here we show that TRANSWESD can also be used to prune perturbation graphs obtained from genetical genomics data, thus complementing Bayesian networks and SEM.

Overall, our framework constitutes simple modules with a total number of just two threshold parameters. We tested it on *in silico* recombinant inbred line data that have been provided by the DREAM initiative (DREAM5, Systems Genetics challenge 3 A; http://wiki.c2b2.columbia.edu/dream/index.php/D5c3; Prill *et al.*, 2010, 2011; Stolovitzky *et al.*, 2009; Vignes *et al.*, 2011). We find that the proposed modular reconstruction approach outperforms the best performer (Vignes *et al.,* 2011). We also present and discuss reconstruction results from real genetical genomics data of yeast (Brem and Kruglyak, 2005).

2 METHODS

2.1 Workflow overview

Figure 1 shows the general workflow of our reconstruction framework together with a simple illustrative example. Starting from a typical set of genetical genomics data that include genotyped markers, phenotyped genes and gene-to-marker association, marker linkage analysis and genotype assignment for each gene is performed in a preprocessing step. From these data an unweighted and unsigned perturbation graph G1 is derived using genotype–phenotype correlation in combination with an appropriated thresholding strategy. The nodes in the graph directly correspond to genes while linkage information is kept to allow later eQTL assignment for each gene. The perturbation graph G1 is refined to G2 by quantifying each identified edge with respect to edge sign and weight, which indicate activation/repression and interaction strength, respectively. Owing to genetic linkage, true regulators may be masked by other genes, e.g., on adjacent positions on the genetic map, resulting into eQTLs. The eQTLs of a given target gene t can be identified on the basis of all potential regulator genes of t and the marker linkage map. These relationships are captured in graph G3. Graph G4 is subsequently obtained by selecting one candidate regulator per eQTL based on the maximum of the edge weights. We call G4 the final perturbation graph, whose edges reflect direct and indirect effects between genes induced by genetic variations. To remove indirect edges that can be explained by the operation of sequences of edges (paths), we apply the transitive reduction method TRANSWESD resulting in the final graph G5. Optionally, if one is left to verify the interactions experimentally, it is desirable to have a list of edges sorted with respect to edge confidences. Such a list is also used by the DREAM5 evaluation procedures to assess the quality of a reconstructed network (Section 3.1). We generate such a sorted list based on the edge weights.

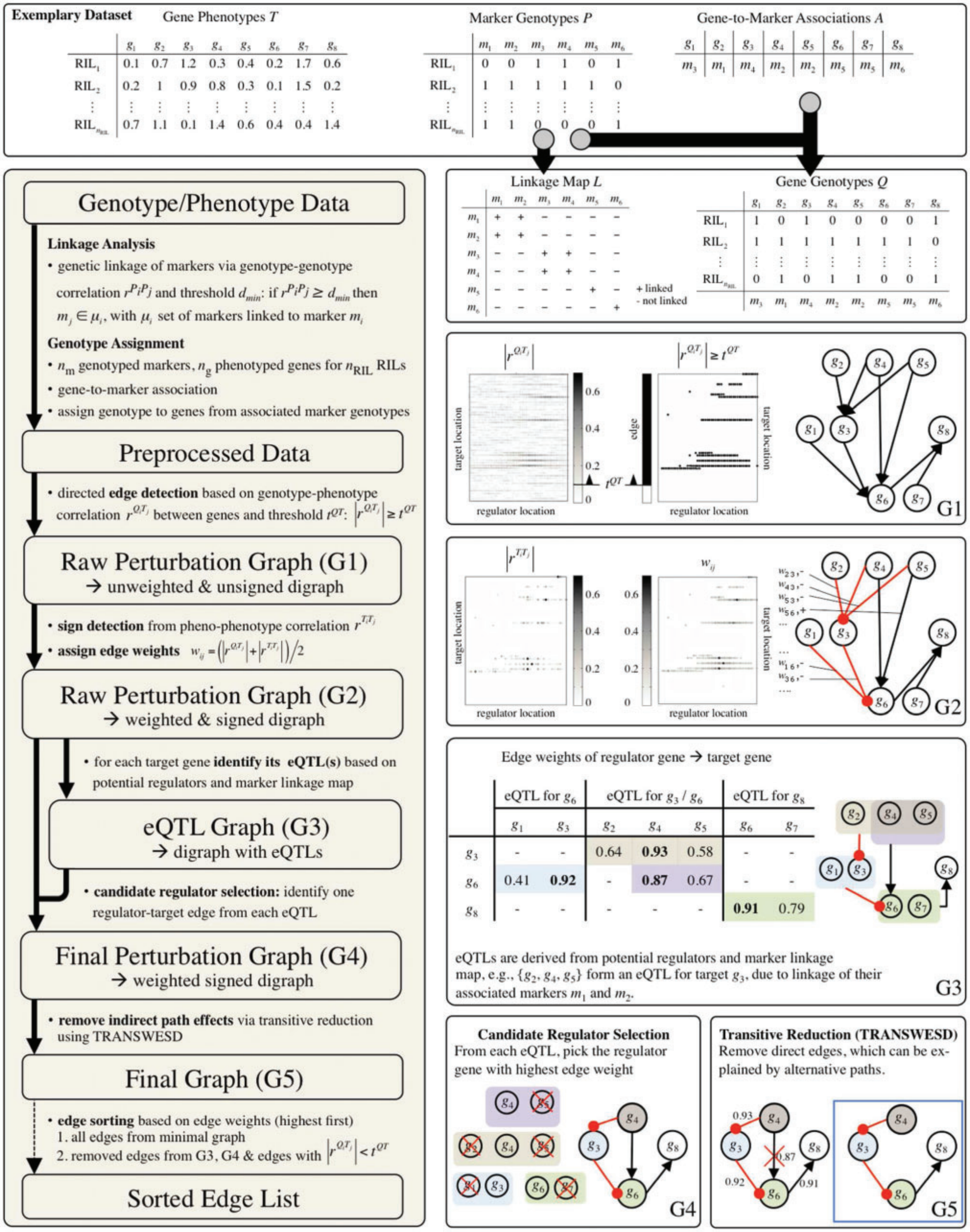
2.2 Preprocessing genetical genomics data

A genetical genomics dataset typically consists of the following measured/determined information (see Fig. 1): From a segregating population such as RILs, each segregant is genotyped for a set of polymorphic genetic markers that cover the genome or at least part of it. The genotype of each marker in each RIL is captured in a matrix P [e.g. two-valued (0/1) in the case of haploidic genomes]. Additionally, genes are expression profiled in each RIL (stored in a matrix T). In the typical case that several genes are associated to one marker, genes can be associated to a specific marker based on their position on the genome map (yielding the list A in Fig. 1). From this information, we extract by simple preprocessing steps, two additional matrices needed before the actual reconstruction process is started. First, the gene-to-marker association A is used to assign, for each RIL, an (approximated) genotype Q to each gene, which is taken from its associated marker genotype P. This genotype assignment is based on the assumption of genetic linkage between markers and genes. Further, genetic linkage of the markers needs to be taken into account to identify potential eQTLs in G3 at a later reconstruction step. If genetic linkage of the markers is unknown, a linkage analysis can be performed based on genotype–genotype Pearson correlation $r^{P_iP_j}$ of the markers m_i and m_j (P_i and P_j denoting their genotype). With a given threshold $d_{min} \in [0,1]$, if $r^{P_iP_j} \geq d_{min}$ then $m_j \in \mu_i$ with μ_i being the set of markers linked to marker m_i. By this procedure we obtain a linkage map L. The parameter d_{min} represents the minimal genotypic correlation at which two markers are considered to be linked. The threshold can be derived from (i) testing for significance of deviation from zero by a t-test (cf. Appendix of Bing and Hoeschele, 2005), whereas empirical significant levels can be derived from permutation tests (Churchill and Doerge, 1994; Carlborg *et al.*, 2005), (ii) the typical separation of candidate regulators in the eQTL map based on $r^{Q_iT_j}$ (see Fig. 1 right, panel of G1). Specifically, one may analyse the average number of eQTLs over the genome as a function of d_{min}. Regions of d_{min}, where the average number of eQTLs does not change much, indicate an optimal value. A similar thresholding strategy could be applied if the genetic distances (given in centiMorgan) between the markers are known *a priori.*

2.3 Generating the raw perturbation graph from genetical genomics data

The next step is the generation of the perturbation graphs G1 and G2 from the (preprocessed) genetical genomics data. The idea for detecting a potential regulator–target interaction is that a variation in a regulator gene's genotype causes a variation in the phenotype of the target gene. We use T_j to indicate the expression phenotype (etrait) of a gene j and Q_i for the genotype of a gene i (obtained from the marker genotype as described above). Based on the genotype–phenotype Pearson correlation coefficient $r^{Q_iT_j}$ (see Fig. 2), we assume an edge $i \rightarrow j$ to exist, if it exceeds a threshold value t^{QT}:

$$\left|r^{Q_iT_j}\right| \geq t^{QT}. \tag{1}$$



Exemplary Dataset

Gene Phenotypes T

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	0.1	0.7	1.2	0.3	0.4	0.2	1.7	0.6
RIL_2	0.2	1	0.9	0.8	0.3	0.1	1.5	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$RIL_{n_{RIL}}$	0.7	1.1	0.1	1.4	0.6	0.4	0.4	1.4

Marker Genotypes P

	m_1	m_2	m_3	m_4	m_5	m_6
RIL_1	0	0	1	1	0	1
RIL_2	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$RIL_{n_{RIL}}$	1	1	0	0	0	1

Gene-to-Marker Associations A

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

Linkage Map L

	m_1	m_2	m_3	m_4	m_5	m_6
m_1	+	+	–	–	–	–
m_2	+	+	–	–	–	–
m_3	–	–	+	+	–	–
m_4	–	–	+	+	–	–
m_5	–	–	–	–	+	–
m_6	–	–	–	–	–	+

+ linked
- not linked

Gene Genotypes Q

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
RIL_1	1	0	1	0	0	0	0	1
RIL_2	1	1	1	1	1	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$RIL_{n_{RIL}}$	0	1	0	1	1	0	0	1
	m_3	m_1	m_4	m_2	m_2	m_5	m_5	m_6

Genotype/Phenotype Data

Linkage Analysis

- genetic linkage of markers via genotype-genotype correlation $r^{P_iP_j}$ and threshold d_{min}: if $r^{P_iP_j} \geq d_{min}$ then $m_j \in \mu_i$, with μ_i set of markers linked to marker m_i

Genotype Assignment

- n_m genotyped markers, n_g phenotyped genes for n_{RIL} RILs
- gene-to-marker association
- assign genotype to genes from associated marker genotypes

Preprocessed Data

- directed **edge detection** based on genotype-phenotype correlation $r^{Q_iT_j}$ between genes and threshold t^{QT}: $|r^{Q_iT_j}| \geq t^{QT}$

Raw Perturbation Graph (G1)
→ unweighted & unsigned digraph

- **sign detection** from pheno-phenotype correlation $r^{T_iT_j}$
- **assign edge weights** $w_{ij} = \left(|r^{Q_iT_j}| + |r^{T_iT_j}|\right)/2$

Raw Perturbation Graph (G2)
→ weighted & signed digraph

- for each target gene **identify its eQTL(s)** based on potential regulators and marker linkage map

eQTL Graph (G3)
→ digraph with eQTLs

- **candidate regulator selection:** identify one regulator-target edge from each eQTL

Final Perturbation Graph (G4)
→ weighted signed digraph

- **remove indirect path effects** via transitive reduction using TRANSWESD

Final Graph (G5)

- **edge sorting** based on edge weights (highest first)
  1. all edges from minimal graph
  2. removed edges from G3, G4 & edges with $|r^{Q_iT_j}| < t^{QT}$

Sorted Edge List

Edge weights of regulator gene → target gene

	eQTL for g_6		eQTL for g_3 / g_6			eQTL for g_8	
	g_1	g_3	g_2	g_4	g_5	g_6	g_7
g_3	-	-	0.64	**0.93**	0.58	-	-
g_6	0.41	**0.92**	-	**0.87**	0.67	-	-
g_8	-	-	-	-	-	**0.91**	0.79

eQTLs are derived from potential regulators and marker linkage map, e.g., $\{g_2, g_4, g_5\}$ form an eQTL for target g_3, due to linkage of their associated markers m_1 and m_2.

Candidate Regulator Selection
From each eQTL, pick the regulator gene with highest edge weight

Transitive Reduction (TRANSWESD)
Remove direct edges, which can be explained by alternative paths.

Fig. 1. Workflow of the proposed framework for reconstructing gene regulatory networks from genetical genomics data (left) with an illustrative example (top panel and right). For detailed explanations see text

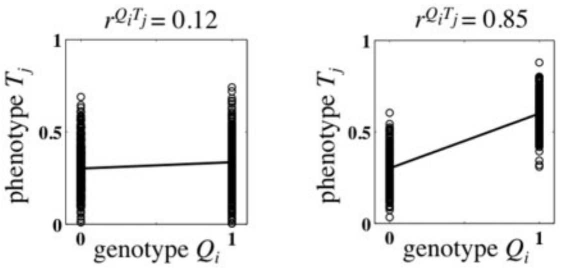


Fig. 2. Illustration for low (left) and high (right) correlation between the genotype Q_i of a gene i (with two genotypes; $Q_i \in \{0, 1\}$) and the expression phenotype T_j of a gene j. The example on the right-hand side indicates that gene i may regulate (directly or indirectly) gene j

The derived candidate edges reflect regulation of gene j by gene i by either *cis*, *cis–trans* or *trans* effects. In the case of $i = j$, it is most likely a *cis* effect, otherwise one has to condition on i: if gene i has a *cis* effect then gene j is *cis–trans* regulated else it is *trans* regulated. All three effects will result in increased correlations and Equation (1) can thus be used to derive the candidate edges for the GRN. The threshold $t^{QT} \in [0,1]$ can be selected based on a combination of several criteria, including (i) similar to the marker linkage analysis by P-values for rejecting $|r^{Q_iT_j}| > 0$ based on a t-test, (ii) minimal/maximal edge numbers one expects to find in the GRN and (iii) existing data. In the case of small sample size, Spearman correlation might be more appropriate. For diploidic genomes, where Q_i is three-valued, one may apply the same procedure for each pairwise combination of genotypes and merge the resulting networks to G1.

Importantly, the nodes in the obtained graph G1 directly correspond to genes (as required to eventually reconstruct gene regulatory networks); the eQTL (regions) will be assigned later in graph G3 based on the linkage map L. Beforehand, we assign an edge sign and edge weight to each detected candidate edge $i \rightarrow j$ in G1 resulting in G2 (Fig. 1). The edge sign s_{ij} is derived from $\mathrm{sgn}\left(r^{T_iT_j}\right)$, i.e., the correlation coefficient between expression levels of genes i and j. The strength w_{ij} of an edge is quantified by

$$w_{ij} = \left(|r^{Q_iT_j}| + |r^{T_iT_j}|\right)/2. \qquad (2)$$

The edge weight accounts for genotype–phenotype (QT-) and phenotype–phenotype (TT-) correlations by averaging both. This is especially important for (*cis*-)*trans*-regulated targets, as these are affected by both, geno- and phenotype of the potential regulator. We have also tested either QT- or TT-correlation alone, which both led to reconstructed GRN of significant lower quality (at least when applied to the DREAM5/3 A data). We also found that substituting the QT correlation coefficient in Equation (1) by the average TT- and QT-correlations as used in Equation (2) is not favourable, probably because then many high TT-correlation wrongly indicate a directed relationship (e.g., due to common upstream regulators).

2.4 Identification of eQTLs and candidate regulator selection

Owing to genetic linkage (correlated genotypes), a gene j that is found to be targeted by a gene i (i.e., an edge from i to j exists in G2) is typically also targeted by several other genes genetically adjacent to i resulting into an eQTL. An eQTL with respect to a given target gene j is identified by the set of all those genes that are potential regulators of j in G2 *and* that are genetically linked via their markers (see Fig. 1, G3: target gene g_3 has one eQTL formed by $\{g_2, g_4, g_5\}$). Importantly, two potential regulators g_x and g_y can be in the same eQTL, even in the case where their associated markers m_x and m_y are not linked in the linkage map L. This happens if there is another candidate regulator g_z whose marker m_z is linked to m_x and m_y in L. Note also that for each target gene, there may exist several eQTLs: in Fig. 1, gene g_6 has two eQTLs formed by genes $\{g_4, g_5\}$ and $\{g_1, g_3\}$. Once we have identified all eQTL(s) for each target gene, we arrive at the eQTL graph G3 (Fig. 1) in which the edges connect eQTLs with their target genes. G3 would represent the final result of classical eQTL mapping. If eQTL mapping was the envisioned goal, one could stop the procedure at this point.

However, if the reconstruction of a gene regulatory network is the ultimate goal then we need to select single candidate genes from each eQTL. Since there is a high probability that only a few or even only one of all the potential regulators of an eQTL are truely connected with the target gene, keeping all interactions in G2 that emanate from one and the same eQTL (the eQTLs being captured in G3) would result in many false positive predictions in the reconstructed network. We therefore select from each eQTL the candidate regulator with the maximal edge weight to be the 'true' regulator of the target gene, i.e., for each eQTL C we identify $i \rightarrow j$ with $w_{ij} = \max\left(w_{kj}\right),\ k \in C$, as the true edge and all other edges are removed from the eQTL C. We arrive then at the final perturbation graph G4 in Fig. 1 in which the nodes represent again genes.

2.5 Identifying and removing indirect effects (TRANSWESD)

Candidate regulator selection in the previous section leads to the reduced graph G4 where genetic linkage effects have been removed. We can now assume that the edges in G4 reflect true causalities. However, an edge may still represent an indirect effect induced by a chain of interactions [e.g. the effect of gene g_4 on gene g_6 in G5 of Fig. 1 is likely to be induced by the double-negative (thus positive) path $g_4 -| g_3 -| g_6$]. The goal of this final step is therefore to identify and eliminate edges that arise from indirect effects. To this end we use our recently introduced algorithm TRANSWESD, a TRANSitive reduction method for WEighted Signed Digraph, which is briefly described in the following (for further details see Klamt *et al.,* 2010). For TRANSWESD, we need association weights z_{ij} between nodes of the graph, which we directly derive from the edge weight via $z_{ij} = 1 - w_{ij}$, i.e., a low z_{ij} indicates a high association between i and j. An edge $i \rightarrow j$ with sign s_{ij} and weight w_{ij} is removed, if there is an alternative path P_{ij} $(i \Rightarrow j)$ connecting i and j with the following properties: (a) P_{ij} is simple, i.e., it does not contain a cycle; (b) P_{ij} does not involve edge $i \rightarrow j$; (c) the overall sign of P_{ij} (obtained by multiplying the signs of all its edges) is the same as s_{ij}; (d) the maximum weight of all edges on path P_{ij}, $\mathrm{z}_{\max}(\mathrm{P}_{ij})$, fulfills $\mathrm{z}_{\max}(\mathrm{P}_{ij}) < \alpha\ z_{ij}$. The confidence factor α is typically chosen close (but smaller) than unity, we use here 0.95 as in Klamt *et al.* (2010). If such a path P_{ij} exists, then it is considered as an explanation for the observed (indirect) effect of i upon j and the edge $i \rightarrow j$ is removed. If the graph is acyclic, the transitive reduction is unique and can easily be found. To deal with cyclic graphs, TRANSWESD uses some reasonable rules to resolve non-uniqueness [e.g., by removing edges with highest weight (lowest association) first] and also provides approximation variants for large networks (cf. Klamt *et al.*, 2010). For the computations made herein, we used the TRANSWESD implementation of *CellNetAnalyzer* (Klamt *et al.*, 2007). After applying TRANSWESD, we obtain the final reconstructed graph G5.

2.6 Sorted edge list

Optionally, a sorted list (ranking) of regulator–target interactions can be generated from the final graph G5, e.g., for prioritizing edges for experimental validation. We propose a sorted edge list which is made up of two parts. The first part of the sorted edge list contains all edges from the final graph, sorted according to edge weights with highest weights (= most significant) first. To also account for edges that have maybe wrongly been dropped during thresholding (not contained in G1), cluster removal, or by TRANSWESD, the second part contains all of these removed edges, also sorted according to their edge weights [Equation (2)] in descending order.

3 RESULTS

We applied the proposed method to (i) synthetic genetical genomics data that were provided for a recent Systems Genetics challenge of the DREAM initiative (DREAM 5, Challenge 3 A; http://wiki.c2b2.columbia.edu/dream/index.php/D5c3), and (ii) real genetical genomics data from yeast, which were originally published in Brem and Kruglyak (2005).

3.1 In silico genetical genomics data

The task of the Systems Genetics challenge of DREAM was to infer causal gene regulatory networks from phenotype expression data of a genotyped segregated population. Owing to lack of reliable experimental datasets for benchmarking different reconstruction algorithms, participants were given realistic *in silico* data which were generated by the SysGenSIM software (Pinna *et al*., 2011). The provided simulated data represent (noisy) data from homozygous RILs, whereas the genome of each individual consists of 1000 genes and is made up of 20 chromosomes with 50 genes each. Five different networks of modular scale-free topology had to be reconstructed for three different sample sizes (populations of 100, 300 and 999 RILs) eventually resulting in 15 reconstructed GRN. The haploidic genotype for all genes in all RILs was given as a binary vector (simulating the ideal situation of one marker per gene). The genotypes of adjacent genes were correlated mimicking genetic linkage. Each gene was assumed to have one functional genetic variant, either in the promotor (*cis* effect on the gene's expression rate) or coding region (*trans* effect on the target gene) of the gene. One motivation of the challenge was to analyse the reconstruction quality of participating methods when the population size becomes very small in comparison to the number of genes (e.g., 100 RILs/1000 genes). For each sample size, the reconstructed networks in form of a sorted edge list (last step in Fig. 1) are passed to the evaluation script of DREAM (for details see Stolovitzky *et al*., 2009 and www.the-dream-project.org). Since self-regulation was excluded by the challenge, we removed edges $i \rightarrow j$ where $i = j$. The evaluation script quantifies the reconstruction quality by comparing the reconstructed GRN to the gold standard resulting in AUROC and AUPR values. The resulting overall score is based on empirical *P*-values relatively computed to all submitted reconstructed GRNs (Stolovitzky *et al*., 2009). Applying our method to the described DREAM5/3 A data is straightforward and yields the results presented in Table 1. The preprocessing (Fig. 1) is reduced to generating the marker/gene linkage map *L*, since each gene had its own associated marker. When applying our framework, we need to specify the parameters t^{QT} and d_{min}. Two scenarios were considered for the threshold t^{QT}. First, based on the gold standards, we determined for each of the three RIL population sizes the optimal value (delivering the highest overall score for all networks of this size), which is then used for all five networks. It turns out that smaller population sizes require larger threshold values: 0.23 for 100 RILs; 0.15 for 300 and 0.09 for 999. These optimal values correspond to *P*-values smaller than 0.01, when assuming simple *t*-statistic. In addition, for an unbiased scenario, we sampled t^{QT} uniformly in the range of $0.05 \ldots 0.6$, which define a plausible range of maximal and minimal edge numbers contained in G1, and computed the average result over all networks (column G5* in Table 1). Regarding the parameter d_{min} required for identifying eQTLs, we chose $d_{min} = 0.5$ after inspecting geno–phenotype relationships in the data (see Fig. 1, panel G2). However, we found that the results are extremely robust w.r.t. changes of d_{min}. For example, the overall scores varied <1% when varying d_{min} in a large range of 0.3–0.8. In contrast, disregarding genetic linkage between markers by setting $d_{min} = 1$ leads to much lower reconstruction quality (see Table 1 overall scores of G2 versus G4). This also holds for the other extreme case $d_{min} = 0$, which would result in one large eQTL for all identified regulator candidates.

Several key observations can be made. Moving from the initial perturbation graph G2 to G4 by our candidate regulator selection approach, we see a clear improvement with respect to FP reduction at minimal loss of TPs by one order of magnitude. For example, for the 999 RIL individual scenario, when transforming G2 to G4, the number of false positive edges (FP) reduces on average from 51 644 down to 3368, whereas the number of true positives (TPs) reduces only from 2371 to 1844. Undesired removal of TPs may occur by selecting the wrong regulator gene of an eQTL or because several genes from an eQTL target the same gene concurrently. In the second pruning step from G4 to G5, TRANSWESD removes many indirect edges due to alternative paths found in G4, improving in almost all cases the AUPR value. As the precision [TP/(TP + FP)] of the reconstructed network increases significantly in all cases upon applying TRANSWESD, one could expect an even better relative improvement of the AUPR value. However, there is only a moderate increase of the AUPR because TRANSWESD removes mainly edges with lower edge weight and thus with lower confidence (and ranking position) in the edge list. Generally, TRANSWESD works better for networks with lower connectivity (in DREAM5/3 A, the edge density increases with increasing network index from approximately 2000 up to 5000 edges) and with larger sample size.

The made observations also hold for averaged scores from uniformly sampled (non-optimal) threshold parameters t^{QT} (see G5* in Table 1), i.e., the method is robust against threshold selection. Comparing the results of the proposed framework to the best performer of the DREAM5/3 A challenge for each RIL sample size and different network topologies (last column in Table 1), we see a clear improvement also for randomly chosen threshold parameters (G5* in Table 1). Even without applying any FP reduction, G2 is almost always better than the best performer, although it was constructed based on a simple eQTL mapping approach alone. G4/G5 obtained after pruning are always better than the best performer [this also holds for an improved version of the best performer method; cf. Vignes *et al*. (2011)]. Further, moving from large to small sample sizes, we see a clear relative improvement, i.e., increase of the overall score (e.g. of G5* averaged over the five different networks) with respect to all DREAM participant submissions. Consequently, our proposed method performs especially well for small sample sizes. At a given sample size, our method has averaged AUPR values which are up to three times larger than the best performer in the case of 100 samples (G5* versus best

Table 1. Reconstruction results for the DREAM5/3A: From G2 to G5 with $d_{min}=0.5$ and optimal threshold parameter t^{QT}

DREAM5	G2		G4		G5		G5*		Best performer DREAM5/3 A	
100/net1										
aupr auroc	0.175	0.839	0.241	0.842	**0.247**	**0.843**			0.085	0.754
TP/FP	815/32911		629/7785		**600/4210**					
100/net2										
aupr auroc	0.141	0.816	0.200	0.820	**0.203**	**0.821**			0.060	0.718
TP/FP	1034/35359		767/7689		**710/3859**					
100/net3										
aupr auroc	0.140	0.798	0.182	0.801	**0.186**	**0.802**			0.053	0.696
TP/FP	1174/35666		838/7691		**777/3791**					
100/net4										
aupr auroc	0.123	0.782	0.152	0.787	**0.158**	**0.788**			0.054	0.676
TP/FP	1259/35987		889/7770		**821/3861**					
100/net5										
aupr auroc	0.124	0.775	0.154	0.780	**0.158**	**0.781**			0.054	0.670
TP/FP	1409/38333		1016/8212		**919/3732**					
300/net1										
aupr auroc	0.279	0.924	0.462	0.928	**0.475**	**0.927**			0.211	0.855
TP/FP	1190/30910		1017/4216		**1000/3576**					
300/net2										
aupr auroc	0.209	0.889	0.355	0.892	**0.356**	**0.892**			0.144	0.793
TP/FP	1512/32197		1207/4256		**1167/3505**					
300/net3										
aupr auroc	0.205	0.885	0.309	0.890	**0.316**	**0.890**			0.141	0.786
TP/FP	1727/36869		1338/4757		**1274/3483**					
300/net4										
aupr auroc	0.188	0.868	0.291	0.854	**0.292**	**0.873**			0.132	0.759
TP/FP	1928/35180		1485/4538		**1409/3449**					
300/net5										
aupr auroc	0.191	0.849	0.291	0.854	**0.291**	**0.854**			0.113	0.737
TP/FP	2054/35607		1594/4481		**1504/3449**					
999/net1										
aupr auroc	0.288	0.965	0.601	0.969	**0.630**	**0.969**			0.358	0.933
TP/FP	1614/41139		1412/2830		**1385/2495**					
999/net2										
aupr auroc	0.247	0.936	0.441	0.942	**0.468**	**0.942**			0.258	0.885
TP/FP	2081/49406		1617/3359		**1561/2866**					
999/net3										
aupr auroc	0.232	0.920	0.435	0.927	**0.442**	**0.926**			0.195	0.844
TP/FP	2382/52813		1873/3328		**1779/2635**					
999/net4										
aupr auroc	0.223	0.908	0.377	0.916	**0.378**	**0.915**			0.183	0.821
TP/FP	2693/56059		2026/3626		**1872/2622**					
999/net5										
aupr auroc	0.225	0.891	0.381	0.899	**0.373**	**0.898**			0.178	0.813
TP/FP	3087/58801		2293/3697		**2075/2780**					
average over five different networks										
100/aupr auroc	0.140	0.802	0.186	0.806	**0.191**	**0.807**	0.166	0.807	0.061	0.703
300/aupr auroc	0.215	0.883	0.342	0.887	**0.346**	**0.887**	0.250	0.887	0.148	0.786
999/aupr auroc	0.243	0.924	0.447	0.930	**0.458**	**0.930**	0.294	0.928	0.234	0.859
100/TP/FP	1138/35651		828/7829		**765/3891**					
300/TP/FP	1682/34153		1328/4450		**1271/3504**					
999/TP/FP	2371/51644		1844/3368		**1734/2860**					
100/score	193.77		231.61		**236.22**		214.89		81.87	
300/score	170.54		237.72		**239.03**		189.42		89.40	
999/score	172.67		250.25		**251.81**		193.49		140.56	

Column G5* shows averaged results when sampling parameter t^{QT} equally distributed on the interval [0.05 . . . 0.6]. For more explanations see text.

Table 2. Reconstruction results for the yeast genetical genomics data set (Brem and Kruglyak, 2005) compared with the yeast gold standard of DREAM5/challenge 4.4

	G2	G4	G5
aupr/paupr/rank	0.0274/5.7e-11/4	0.0293/2.34e-14/3	0.0293/1.89e-14/3
auroc/pauroc/rank	0.5396/6.7e-28/1	0.5407/6.14e-30/1	0.5407/6.4e-30/1

The first row shows (separately for G2, G4 and G5) the AUPR, the ***P***-value of AUPR and the (virtual) rank within the DREAM5/challenge 4.4 AUPR performance ranking (total number of participants: 29). The second row gives the same values with respect to AUROC.

performer averaged over the five networks). This shows, that for small sample sizes, a rather simple method based on pure correlation measures in combination with FP reduction methods seems to be the best choice, keeping in mind that many different methods have been used by the 16 participants. However, even for the largest RIL populations provided in the DREAM5/3 A challenge, our method still achieves significantly higher scores.

3.2 Realistic genetical genomics data of yeast

To test our approach with real data, we use genotypic and expression data from 112 segregants obtained from a yeast cross between BY and RM strains of *Saccharomyces cerevisiae* (Brem and Kruglyak, 2005). The dataset was kindly provided in preprocessed form by Alberto de la Fuente (CRS4, Sardinia). Only 1573 of all 2956 markers were associated to at least one of the 5736 expression-profiled genes. Further, a gene-to-marker association list A was available. In contrast to the DREAM5/3 A data, there were thus much fewer markers than genes. We describe some of the steps in the workflow in more details: After preprocessing the data by computing the linkage map L from the marker genotypes in the RILs with $d_{min} = 0.5$ (we chose the same value as for the *in silico* data), we apply the reconstruction framework to the matrices T and Q to obtain G2. The correlation threshold has been set to $t^{QT} = 0.23$, corresponding to the determined optimal threshold for the *in silico* data with 100 RILs (as this is closest to the 112 RILs available in this study). The two parameters were thus unbiased and not specifically optimized for this dataset. Using the linkage map L, the eQTL graph G3 is obtained, which is used to derive the final perturbation graph G4 by selecting from each eQTL the gene–target interaction in G2 that has the highest edge weight.

In the literature and in some databases, one can find published (most likely sub-networks) of the yeast gene regulatory network, whose interactions have been identified from different sources, including ChIP–chip and motif finding studies (Lee *et al.*, 2002, Reimand *et al.*, 2010). However, a rigorous evaluation of our inferred network is not trivial, as the reliability of the gold standards from the sources mentioned above is unclear. We therefore pursue the following strategy to evaluate our reconstructed network. In another challenge of DREAM5 (challenge 4.4; not to be confused with the Systems Genetics challenge 3 A analysed in Section 3.1), the goal was to infer a subpart of the yeast GRN focusing on 333 candidate transcription factors (TF) and their interactions with 5950 (potential target) genes based on 536 microarrays each containing expression profiles for a given perturbation (e.g., specified gene knock-out or overexpression, including partial replicates). The evaluation in this challenge was based on a given gold standard containing interactions considered to exist between the genes of the given transcription factors and all other genes (it is not known to the authors how this gold standard was compiled), whereas self-regulation has been excluded. Most genes (5451) of the data from Brem and Kruglyak (2005) are present in the DREAM5/challenge 4.4 yeast gold standard. To compare our absolute performance with respect to this gold standard and relatively to the other 29 participants of this challenge, we evaluated a subgraph of our reconstructed yeast GRN, which was restricted to potential interactions between the 333 TFs as regulators and the 5451 target genes present in DREAM5/4.4. We used the specific DREAM5 evaluation script, which computes the AUPR and AUROC (and their *P*-values) of an inferred network with respect to the provided gold standard. The results are presented in Table 2. We first observe that even though our reconstruction is based on only 112 RILs (compared with the large number of 536 microarray experiments available to the participants of this challenge), the reconstruction belongs to the very best of the submissions (rank 1/3 for AUROC/AUPR). Therefore, as observed also for the *in silico* data (Section 3.1), the performance of our reconstruction framework proves again its suitability for small sample sizes. In Table 2 we also see that, when moving from G2 to G5 via G4, our applied FP-pruning strategies always improve the precision-recall and associated *P*-values. The same holds for the AUROC score, except that in G5 a minor reduction of the *P*-value can be observed due to the loss of some TPs during FP reduction.

Regarding the absolute quality of our reconstructed network, we notice that it does not meet the high scores of the *in silico* challenge DREAM5/3A. There are several possible reasons for this behaviour. First, the amount and resolution of the *in silico* data quality is much better, in particular, there was one marker per gene (whereas only 1573 markers for 5736 genes are available in the yeast data). Although noise has been added to the *in silico* data, it might be much higher under realistic conditions or/and other sources of uncertainty might also hamper the visibility of true interactions. Furthermore, the given gold standard for the yeast transcriptional network cannot automatically be considered to be the full truth. The similar low or even worse quality of reconstructed networks submitted by the other participants for DREAM5/4.4, may, at least partially, point to missing or false edges in the gold standard itself. To test the relevance of the inferred networks, it would therefore be interesting whether the top-ranked interactions of the reconstructions (not present in the gold standard) could be validated in experiments.

4 DISCUSSION AND CONCLUSION

In this article, we presented a simple, yet effective, modular framework for gene regulatory network reconstruction from genetical genomics data. We showed that in the case of the DREAM5/3A *in silico* data, our method (even in the non-tuned case) outperforms the best performers which applied a combination of Bayesian network analysis, LASSO and the Dantzig selector (Vignes *et al*., 2011). In the case of real data, we used the DREAM5/4.4 yeast GRN gold standard to assess the quality of the yeast GRN inferred by our method (and genetical genomics data) relatively to networks inferred by classical perturbation experiments and microarray data. The performance of our network reconstruction compares to the very best of the submissions, although we used just 112 RILs in contrast to DREAM5/4.4 submissions, which were based on 536 microarrays with well-defined perturbations. Our analysis indicates that simple correlation methods paired with subsequent FP-pruning strategies outperform complex methods, especially for small sample sizes (experimentally still most relevant). This is most likely due to a larger noise sensitivity of multi-locus method in contrast to univariate correlation analysis. Since correlation-based eQTL mapping yields many true positive, but also many false positive interactions, a local pruning based on linkage information and a global pruning based on path knowledge is important (see Table 1 best performer of DREAM5/3 A versus G2 and G4/5).

In view of our results, the proposed framework performs best on data with one marker per gene, which might be realistic for future ultra-high-throughput sequencing methods. As illustrated in Section 3.2, the framework can be readily applied to the general case where markers cover several genes and it can also be adapted easily to cases with more than two different genotypes. The presented local pruning approach for genetic linkage assumes that only one regulator gene is selected per eQTL, which therefore cannot account for the case that a target has several regulators within one eQTL. A relaxed selection strategy can be used based on partial correlation to potentially select more than one candidate regulator per eQTL, cf. Bing and Hoeschele (2005). In the case of complex traits, i.e., a trait is influenced by several eQTLs, the univariate approach to generate G1 based on correlating one Q_i with one T_j at a time possibly misses combinatorial effects and could therefore potentially result in a higher number of false-negative regulator–target interactions. Alternatively, a multi-locus method as the Lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005) or the random forests (Breitman, 2001) might be used. However, as has been pointed out by Michaelson *et al*. (2010), for very large data sets (millions of dense markers and phenotyped genes), multi-locus methods cannot be used due to computational overload. Here, the presented approach provides a computationally feasible and effective framework for filtering the most important interaction sites (on which multi-locus approaches can be applied) as it does not use any optimization algorithm.

In its presented form the framework has only two threshold parameters, which can be determined from the data in an optimal way by a combination of several approaches (as outlined in Section 2). The framework also allows an adjustment to other reconstruction methods, i.e., modules in the workflow (Fig. 1) may be replaced. For instance the candidate regulator selection based on the eQTL map G3 may be exchanged by other approaches, e.g., partial correlation or local regression.

Although our framework is based on a univariate analysis, it can provide reconstructed GRN at higher precision-recall level than advanced multi-locus methods, even with smaller sample sizes. This might not hold in the case when combinatorial or epigenetic effects (Brazhnik *et al*., 2002) are present, where multi-locus approaches may become advantageous (Michaelson *et al*., 2010). Therefore, in line with the key result of the DREAM initiative stating that community efforts based on many different reconstruction methods produce best results (Prill *et al*., 2011), we conclude that meta-methods, e.g., as proposed by Vignes *et al*. (2011), should combine both simple and complex methods.

ACKNOWLEDGEMENTS

We thank A. de la Fuente for providing us with the yeast data from Brem and Kruglyak (2005) in preprocessed form.

Funding: This work was partially supported by the Federal State of Saxony-Anhalt (Research Centre ‘Dynamic Systems: Biosystems Engineering’) and the German Federal Ministry of Education and Research (Grant 0315505A).

Conflict of Interest: none declared.

REFERENCES

Bing,N. and Hoeschele,I. (2005) Genetical genomic analysis of a yeast segregant population for transcription network inference. *Genetics*, **170**, 533–542.

Brazhnik,P. *et al*. (2002) Gene networks: how to put the function in genomics. *Trends Biotechnol*, **20**, 467–472.

Breitman,L. (2001) Random forests. *Machine Learning*, **45**, 5.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS*, **102**, 1572–1577.

Carlborg,O. *et al*. (2005) Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, **21**, 2383–2393.

Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Friedmann,N. *et al*. (1998) Learning the structure of dynamic probabilistic networks. In *Proceedings 14th Conference on Uncertainty in Artificial Intelligence*, 139–147.

de la Fuente,A. (2010) What are Gene Regulatory Networks?. In Das,S. *et al*. (eds.) *Computational Methodologies in Gene Regulatory Networks*. IGI Global, Hershey, PA, pp. 1–27.

Jansen,R. and Nap,N. (2001) Genetical genomics: the added value from segregation. *Trends Genet*., **17**, 388–391.

Jansen,R. (2003) Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet*., **4**, 145–151.

Keurentjes,J.J.B. *et al*. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl Acad. Sci. USA*, **104**, 1708–1713.

Klamt,S. *et al*. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biol*., **1**, 2.

Klamt,S. *et al*. (2010) TRANSWESD: inferring cellular networks with transitive reduction. *Bioinformatics*, **26**, 2160–2168.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Li,H. *et al.* (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.*, **14**, 1119–1125.

Liu,B. *et al.* (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**, 1763–1776.

Liu,B. *et al.* (2010) Inferring Gene Regulatory Networks from Genetical Genomics Data. In Das,S. *et al.* (eds.) *Computational Methodologies in Gene Regulatory Networks*. IGI Global, Hershey, PA, pp. 79–107.

Michaelson,J.J. *et al.* (2009) Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, **48**, 265–276.

Michaelson,J.J. *et al.* (2010) Data-driven assessment of eQTL mapping methods. *BMC Genomics*, **11**, 502.

Pearl,J. (2000) *Causality: Models, Reasoning, And Inference*. Cambridge University Press, Cambridge.

Pinna,A. *et al.* (2011) Simulating systems genetics data with SysGenSIM. *Bioinformatics*, **27**, 2459–2462.

Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 Challenges. *PLoS One*, **5**, 9202.

Prill,R.J. *et al.* (2011) Crowdsourcing network inference: the DREAM predictive signaling network challenge. *Sci. Signal.*, **4**, mr7.

Reimand,J. *et al.* (2010) Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res.*, **38**, 4768–4777.

Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.

Rockman,M.V. (2008) Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature*, **456**, 738–744.

Schadt,E.E. *et al.* (2005) An integraive genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.

Stolovitzky,G. *et al.* (2009) Lessons from the DREAM2 challenges. *Ann. NY Acad. Sci.*, **1158**, 159–195.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B (Methodol.)*, **58**, 267–288.

Vignes,M. *et al.* (2011) Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis. *PLoS ONE*, **6**, e29165.

Zhu,J. *et al.* (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput. Biol.*, **3**, e69.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B (Methodol.)*, **67**, 301–320.