Detection of regulator genes and eQTLs in gene networks

Lingfei Wang[1] and Tom Michoel[1,*]

[1]Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, Scotland, United Kingdom

[*]Corresponding author, E-mail: tom.michoel@roslin.ed.ac.uk

Abstract

Genetic differences between individuals associated to quantitative phenotypic traits, including disease states, are usually found in non-coding genomic regions. These genetic variants are often also associated to differences in expression levels of nearby genes (they are "expression quantitative trait loci" or eQTLs, for short) and presumably play a gene regulatory role, affecting the status of molecular networks of interacting genes, proteins and metabolites. Computational systems biology approaches to reconstruct causal gene networks from large-scale omics data have therefore become essential to understand the structure of networks controlled by eQTLs together with other regulatory genes, as well as to generate detailed hypotheses about the molecular mechanisms that lead from genotype to phenotype. Here we review the main analytical methods and softwares to identify eQTLs and their associated genes, to reconstruct co-expression networks and modules, to reconstruct causal Bayesian gene and module networks, and to validate predicted networks *in silico*.

Contents

1 Introduction **2**

2 Genetics of gene expression **4**

3 Co-expression networks and modules **6**

3.1 Co-expression gene networks 6

3.2 Clustering and co-expression module detection 7

4 Causal gene networks **10**

4.1 Using genotype data to prioritize edge directions in co-expression networks . 10

4.2 Using Bayesian networks to identify causal regulatory mechanisms 11

4.3 Using module networks to identify causal regulatory mechanisms 13

4.4 Illustrative example 14

5 *In silico* validation of predicted gene regulation networks 14

6 Future perspective: Integration of multi-omics data 17

7 Conclusions 17

1 Introduction

Genetic differences between individuals are responsible for variation in the observable phenotypes. This principle underpins genome-wide association studies (GWAS), which map the genetic architecture of complex traits by measuring genetic variation at single-nucleotide polymorphisms (SNPs) on a genome-wide scale across many individuals [1]. GWAS have resulted in major improvements in plant and animal breeding [2] and in numerous insights into the genetic basis of complex diseases in human [3]. However, quantitative trait loci (QTLs) with large effects are uncommon and a molecular explanation for their trait association rarely exists [1]. The vast majority of QTLs indeed lie in non-coding genomic regions and presumably play a gene regulatory role [4, 5]. Consequently, numerous studies have identified *cis*- and *trans*-acting DNA variants that influence gene expression levels (i.e., "expression QTLs"; eQTLs) in model organisms, plants, farm animals and human (reviewed in [6–10]). Gene expression programmes are of course highly tissue- and cell-type specific, and the properties and complex relations of eQTL associations across multiple tissues are only beginning to be mapped [11–14]. At the molecular level, a mounting body of evidence shows that *cis*-eQTLs primarily cause variation in transcription factor (TF) binding to gene regulatory DNA elements, which then causes changes in histone modifications, DNA methylation and mRNA expression of nearby genes; *trans*-eQTLs in turn can usually be attributed to coding variants in regulatory genes or *cis*-eQTLs of such genes [15].

Taken together, these results motivate and justify a systems biological view of quantitative genetics ("systems genetics"), where it is hypothesized that genetic variation, together with environmental perturbations, affects the status of molecular networks of interacting genes, proteins and metabolites; these networks act within and across different tissues and collectively control physiological phenotypes [16–22]. Studying the impact of genetic variation on gene regulation networks is of crucial importance in understanding the fundamental biological mechanisms by which genetic variation causes variation in phenotypes [23], and is expected to lead to the discovery of novel disease biomarkers and drug targets in human and veterinary medicine [24]. Since direct experimental mapping of genetic, protein–protein or protein–DNA interactions is an immensely challenging task, further exacerbated by the cell-type specific and dynamic nature of these interactions [25], comprehensive, experimentally verified molecular networks will not become available for multi-cellular organisms in the foreseeable future. Statistical and computational methods are therefore essential to reconstruct trait-associated causal networks by integrating diverse omics data [18, 19, 26].

A typical systems genetics study collects genotype and gene, protein and/or metabolite expression data from a large number of individuals segregating for one or more traits of interest. After raw data processing and normalization, eQTLs are identified for each of the expression data types, and a co-expression matrix is constructed. Causal Bayesian gene networks, co-expression modules (i.e. clusters) and/or causal Bayesian module networks are

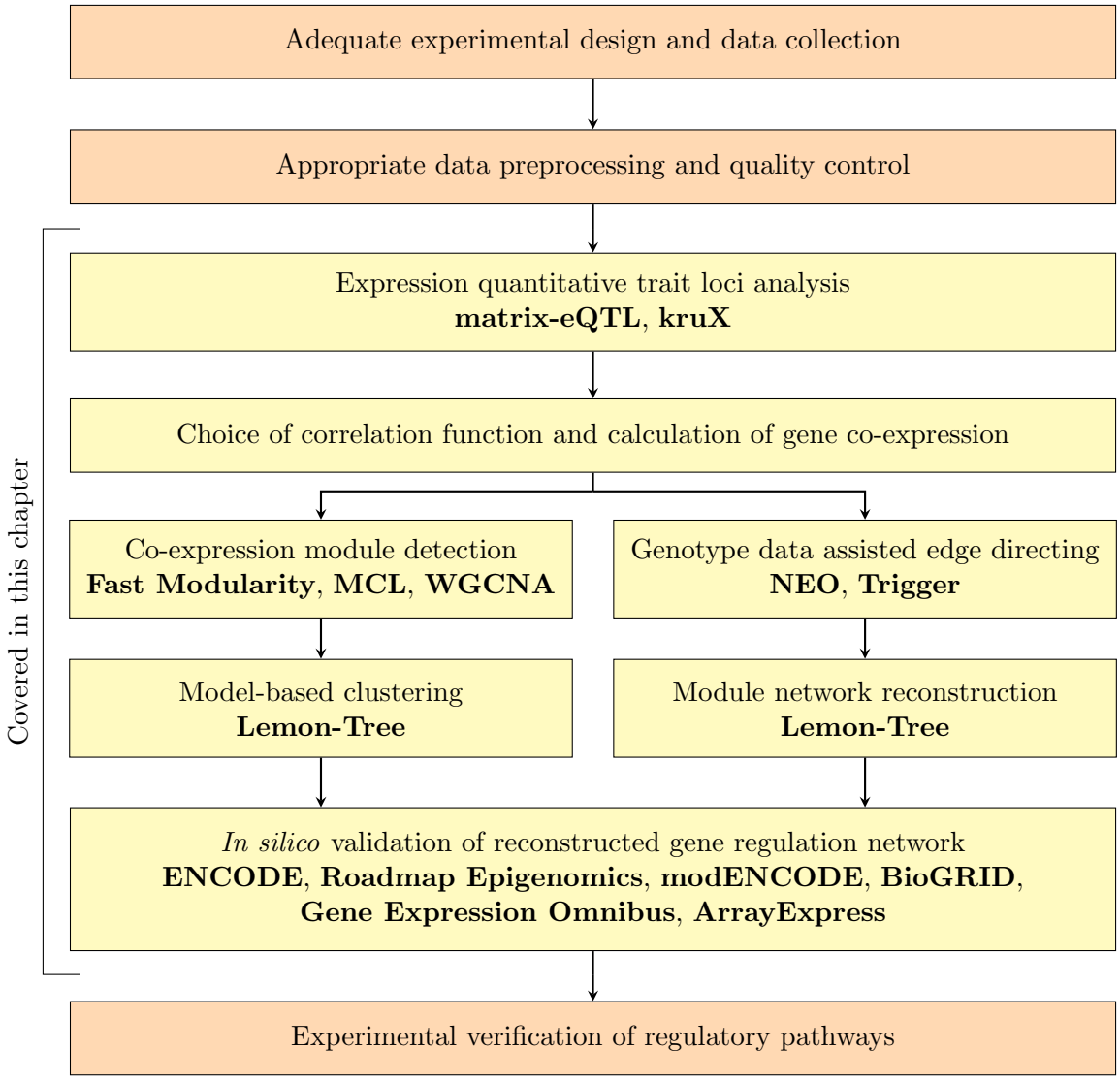

Figure 1: A flow chart for a typical systems genetics study and the corresponding softwares. Steps in light yellow are covered in this chapter.

then reconstructed. *In silico* validation of predicted networks and modules using independent data confirms their overall validity, ideally followed by experimental validation of the most promising findings in a relevant cell line or model organism (Figure 1). Here we review the main analytic principles behind each of the steps from eQTL identification to *in silico* network validation, and present a selection of most commonly used methods and softwares for each step. Throughout this chapter, we tacitly assume that all data has been quality controlled, pre-processed and normalized to suit the assumptions of the analytic methods presented here. For expression data, this usually means working with log-transformed data where each gene expression profile is centred around zero with standard deviation one. We also assume that the data has been corrected for any confounding factors, either by regressing out known covariates and/or by estimating hidden factors [27].

2 Genetics of gene expression

A first step towards identifying molecular networks affected by DNA variants is to identify variants that underpin variations in eQTLs of transcripts [8], proteins [28] or metabolites [29] across individuals. When studying a single trait, as in GWAS, it is possible to consider multiple statistical models to explicitly account for additive and/or dominant genetic effects [30]. However, when the possible effects of a million or more SNPs on tens of thousands of molecular abundance traits need to be tested, as is common in modern genetics of gene expression studies, the computational cost of testing SNP-trait associations one-by-one becomes prohibitive. To address this problem, new methods have been developed to calculate the test statistics for the parametric linear regression and analysis of variance (ANOVA) models [31] and the non-parametric ANOVA model (or Kruskal-Wallis test) [32] using fast matrix multiplication algorithms, implemented in the softwares **matrix-eQTL** (`http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/`) [31] and **kruX** (`https://github.com/tmichoel/krux`) [32].

In both softwares, genotype values of s genetic markers and expression levels of k transcripts, proteins or metabolites in n individuals are organized in an $s \times n$ genotype matrix $\mathbf{G}$ and $k \times n$ expression data matrix $\mathbf{X}$. Genetic markers take values $0, 1, \ldots, \ell$, where ℓ is the maximum number of alleles ($\ell = 2$ for biallelic markers), while molecular traits take continuous values. In the linear model, a linear relation is tested between the expression level of gene i and the genotype value (i.e. the number of reference alleles) of SNP j. The corresponding test statistic is the Pearson correlation between the ith row of $\mathbf{X}$ and the jth row of $\mathbf{G}$, for all values of i and j. Standardising the data matrices to zero mean and unit variance, such that for all i and j,

$$\sum_{l=1}^{n} X_{il} = \sum_{l=1}^{n} G_{jl} = 0 \quad \text{and} \quad \sum_{l=1}^{n} X_{il}^2 = \sum_{l=1}^{n} G_{jl}^2 = n,$$

it follows that the correlation values can be computed as

$$R_{ij} = \sum_{l=1}^{n} X_{il} G_{jl} = (\mathbf{X}\mathbf{G}^T)_{ij},$$

where $\mathbf{G}^T$ denotes the transpose of $\mathbf{G}$. Hence, a single matrix multiplication suffices to compute the test statistics for the linear model for all pairs of traits and SNPs.

The ANOVA models test if expression levels in different genotype groups originate from the same distribution. Therefore, ANOVA models can account for both additive and dominant effects of a genetic variant on expression levels. In the parametric ANOVA model, suppose the test samples are divided into $\ell + 1$ groups by the SNP j. The mean expression level for gene i in each group m can be written as

$$\overline{X_i^{(m,j)}} = \frac{1}{n^{(m,j)}} \sum_{\{l:\, G_{jl}=m\}} X_{il},$$

where $n^{(m,j)}$ is the number of samples in genotype group m for SNP j.

Again assuming that the expression data is standardised, the F-test statistic for testing gene i against SNP j can be written as

$$F_i^{(j)} = \frac{n-\ell-1}{\ell} \frac{SS_i^{(j)}}{n - SS_i^{(j)}},$$

where $SS_i^{(j)}$ is the sum of squares between groups,

$$SS_i^{(j)} = \sum_{m=0}^{\ell} n^{(m,j)} \overline{X_i^{(m,j)}}^2.$$

Let us define the $n \times s$ indicator matrix $\mathbf{I}^{(m)}$ for genotype group m, i.e. $\mathbf{I}_{lj}^{(m)} = 1$ if $G_{jl} = m$ and 0 otherwise. Then

$$\sum_{\{l:\, G_{jl}=m\}} X_{il} = \left(\mathbf{X}\mathbf{I}^{(m)}\right)_{ij}.$$

Hence, for each pair of expression level X_i and SNP G_j, the sum of squares matrix $SS_i^{(j)}$ can be computed via $\ell - 1$ matrix multiplications [1].

In the non-parametric ANOVA model, the expression data matrix is converted to a matrix $\mathbf{T}$ of data ranks, independently over each row. In the absence of ties, the Kruskal-Wallis test statistic is given by

$$S_{ij} = \frac{12}{n(n+1)} \sum_{m=0}^{\ell} n^{(m,j)} \overline{T_i^{(m,j)}}^2 - 3(n+1),$$

where $\overline{T_i^{(m,j)}}$ is the average expression rank of gene i in genotype group m of SNP j, defined as

$$\overline{T_i^{(m,j)}} = \frac{1}{n^{(m,j)}} \sum_{\{l:\, G_{jl}=m\}} T_{il},$$

which can be similarly obtained from the $\ell - 1$ matrix multiplications.

[1] There are only $\ell - 1$ matrix multiplications, because the data standardization implies that $\mathbf{X}\mathbf{I}^{(0)} = 1 - \sum_{m=1}^{\ell-1} \mathbf{X}\mathbf{I}^{(m)}$.

There is as yet no consensus about which statistical model is most appropriate for eQTL detection. Non-parametric methods were introduced in the earliest eQTL studies [33, 34] and have remained popular, as they are robust against variations in the underlying genetic model and trait distribution. More recently, the linear model implemented in matrix-eQTL has been used in a number of large-scale studies [14, 35]. A comparison on a dataset of 102 human whole blood samples showed that the parametric ANOVA method was highly sensitive to the presence of outlying gene expression values and SNPs with singleton genotype group. Linear models reported the highest number of eQTL associations after empirical False Discovery Rate (FDR) correction, with an expected bias towards additive linear associations. The Kruskal-Wallis test was most robust against data outliers and heterogeneous genotype group sizes and detected a higher proportion of non-linear associations, but was more conservative for calling additive linear associations than linear models [32].

In summary, when large numbers of traits and markers have to be tested for association, efficient matrix multiplication methods can be employed to calculate all test statistics at once, leading to a dramatic reduction in computation time compared to calculating these statistics one-by-one for every pair using traditional methods. Matrix multiplication is a basic mathematical operation which has been purposely studied and optimized for tens of years [36]. Highly efficient packages, such as **BLAS** (http://www.netlib.org/blas/) and **LAPACK** (http://www.netlib.org/lapack/), are available for use on generic CPUs, and are indeed employed in most mainstream scientific computing softwares and programming languages, such as Matlab and R. In recent years, Graphics Processor Unit (GPU)-accelerated computing, such as CUDA, has revolutionised scientific calculations that involve repetitive operations in parallel on bulky data, offering even more speedup than the existing CPU-based packages. The first applications of GPU computing in eQTL analysis have already appeared (e.g. [37]), and more can be expected in the future.

Lastly, for pairs exceeding a pre-defined threshold on the test statistic, a p-value can be computed from the corresponding test distribution, and these p-values can then be further corrected for multiple testing by common procedures [31, 32].

3 Co-expression networks and modules

3.1 Co-expression gene networks

The Pearson correlation is the simplest and computationally most efficient similarity measure for gene expression profiles. For genes i and j, their Pearson correlation can be written as

$$C_{ij} = \sum_{l=1}^{n} X_{il} X_{jl} \,. \tag{1}$$

In matrix notation, this can be combined as the matrix multiplication

$$\mathbf{C} = \mathbf{X}\mathbf{X}^T.$$

Gene pairs with large positive or negative correlation values tend to be up- or down-regulated together, due to either a direct regulatory link between them, or being jointly co-regulated by a third, often hidden, factor. By filtering for correlation values exceeding a significance

threshold determined by comparison with randomly permuted data, a discrete co-expression network is obtained. Assuming that a high degree of co-expression signifies that genes are involved in the same biological processes, graph theoretical methods can be employed, for instance, to predict gene function [38].

One drawback of the Pearson correlation is that by definition it is biased towards *linear* associations. To overcome this limitation, other measures are available. The Spearman correlation uses expression data ranks (cf. Section 2) in Equation (1), and will give high score to *monotonic* relations. Mutual information is the most general measure and detects both linear and non-linear associations. For a pair of discrete random variables A and B (representing the expression levels of two genes) taking values a_l and b_m, respectively, the mutual information is defined as

$$MI(A,B) = H(A) + H(B) - H(A,B),$$

where

$$\begin{aligned} H(A) &= -\sum_l P(a_l) \log P(a_l), \\ H(B) &= -\sum_m P(b_m) \log P(b_m), \\ H(A,B) &= \sum_{lm} P(a_l, b_m) \log P(a_l, b_m), \end{aligned}$$

are the individual and joint Shannon entropies of A and B, and $P(a_l) = P(A = a_l)$, and likewise for the other terms. Since gene expression data are continuous, mutual information estimation is non-trivial and usually involves some form of discretisation [39]. Mutual information has been successfully used as a co-expression measure in a variety of contexts [40–42].

3.2 Clustering and co-expression module detection

It is generally understood that cellular functions are carried out by "modules", groups of molecules that operate together and whose function is separable from that of other modules [43]. Clustering gene expression data (i.e. dividing genes into discrete groups on the basis of similarities in their expression profiles) is a standard approach to detect such functionally coherent gene modules. The literature on gene expression clustering is vast and cannot possibly be reviewed comprehensively here. It includes "standard" methods such as hierarchical clustering [44], k-means [45], graph-based methods that operate directly on co-expression networks [46], and model-based clustering algorithms which assume that the data is generated by a mixture of probability distributions, one for each cluster [47]. Here we briefly describe a few recently developed methods with readily available softwares.

Modularity maximization Modularity maximization is a network clustering method that is particularly popular in the physical and social sciences, based on the assumption that intra-module connectivity should be much denser than inter-module connectivity [48, 49]. In the context of co-expression networks, this method can be used to identify gene modules directly from the correlation matrix $\mathbf{C}$ [50]. Suppose the genes are grouped into N modules M_l, $l = 1, \ldots, N$. Each module M_l is a non-empty set that can contain any combination of

the genes $i = 1, \ldots, k$, but each gene is contained by exactly one module. Also define M_0 as the set containing all genes. The modularity score function is defined as

$$S(M) = \sum_{l=1}^{N} \left(\frac{W(M_l, M_l)}{W(M_0, M_0)} - \left(\frac{W(M_l, M_0)}{W(M_0, M_0)} \right)^2 \right),$$

where $W(A, B) = \sum_{i \in A, j \in B, i \neq j} w(C_{ij})$ is a weight function, summing over all the edges that connect one vertex in A with another vertex in B, and $w(x)$ is a monotonic function to map correlation values to edge strengths. Common functions are $w(x) = |x|$, $|x|^\beta$ (power law) [51], $e^{\beta|x|}$ (exponential) [50], or $1/(1 + e^{\beta x})$ (sigmoid) [52].

A modularity maximization software particularly suited for large networks is **Fast Modularity** (http://www.cs.unm.edu/~aaron/research/fastmodularity.htm) [53].

Markov Cluster algorithm The Markov Cluster (MCL) algorithm is a graph-based clustering algorithm, which emulates random walks among gene vertices to detect clusters in a graph obtained directly from the co-expression matrix $\mathbf{C}$. It is implemented in the **MCL** software (http://micans.org/mcl/) [54,55]. The MCL algorithm starts with the correlation matrix $\mathbf{C}$ as the probability flow matrix of a random walk, and then iteratively suppresses weak structures of the network and performs a multi-step random walk. In the end, only backbones of the network structure remain, essentially capturing the modules of co-expression network. To be precise, the MCL algorithm performs the following two operations on $\mathbf{C}$ alternatingly:

- **Inflation:** The algorithm first contrasts stronger direct connections against weaker ones, using an element-wise power law transformation, and normalizes each column separately to sum to one, such that the element C_{ij} corresponds to the dissipation rate from vertex X_i to X_j in a single step. The inflation operation hence updates $\mathbf{C}$ as $\mathbf{C} \to \mathbf{\Gamma}_\alpha \mathbf{C}$, where the contrast rate $\alpha > 1$ is a predefined parameter of the algorithm. After operation $\mathbf{\Gamma}_\alpha$, each element of $\mathbf{C}$ becomes

$$C_{ij} \to \mathbf{\Gamma}_\alpha C_{ij} = |C_{ij}|^\alpha / \sum_{p=1}^{k} |C_{pj}|^\alpha.$$

- **Expansion:** The probability flow matrix $\mathbf{C}$ controls the random walks performed in the expansion phase. After some integer $\beta \geq 2$ steps of random walk, gene pairs with strong direct connections and/or strong indirect connections through other genes tend to see more probability flow exchanges, suggesting higher probabilities of belonging to the same gene modules. The expansion operation for the β-step random walk corresponds to the matrix power operation

$$\mathbf{C} \to \mathbf{C}^\beta.$$

The MCL algorithm performs the above two operations iteratively until convergence. Non-zero entries in the convergent matrix $\mathbf{C}$ connect gene pairs belonging to the same cluster, whereas all inter-cluster edges attain the value zero, so that cluster structure can be obtained directly from this matrix [54,55].

Weighted Gene Co-expression Network Analysis With higher than average correlation or edge densities within clusters, genes from the same cluster typically share more neighbouring (i.e. correlated) genes. The weighted number of shared neighbouring genes hence can be another measure of gene function similarity. This information is captured in the so-called topological overlap matrix $\mathbf{\Omega}$, first defined in [56] for binary networks as

$$\omega_{ij} = \frac{A_{ij} + \sum_u A_{iu} A_{uj}}{\min(k_i, k_j) + 1 - A_{ij}},$$

where A is the (binary) adjacency matrix of the network and $k_i = \sum_u A_{iu}$ is the connectivity of vertex X_i. The $\sum_u A_{iu} A_{uj}$ term represents vertex similarity through neighbouring genes, and the rest of terms normalise the output as $0 \leq \omega_{ij} \leq 1$. This concept was later extended onto networks with weighted edges by applying a "soft threshold" pre-process on the correlation matrix, for example as

$$A_{ij} = \left| \frac{1 + C_{ij}}{2} \right|^{\alpha},$$

or

$$A_{ij} = \left| C_{ij} \right|^{\alpha},$$

such that $0 \leq A_{ij} \leq 1$ [57]. Note that in the first case only positive correlations have high edge weight, whereas in the second case positive and negative correlations are treated equally. The parameter $\alpha > 1$ is determined such that the weighted network with adjacency matrix A has approximately a scale-free degree distribution [57].

In principle, any clustering algorithm (including the aforementioned ones) can be applied to the topological overlap matrix $\mathbf{\Omega}$. In the popular **WGCNA** software (`http://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/Rpackages/WGCNA/`) [51], which is a multi-purpose toolbox for network analysis, hierachical clustering with a dynamic tree-cut algorithm [58] is employed.

Model-based clustering Model-based clustering approaches assume that the observed data is generated by a mixture of probability distributions, one for each cluster, and takes explicitly into account the noise of gene expression data. To infer model parameters and cluster assignments, techniques such as Expectation Maximization (EM) or Gibbs sampling are used [59]. A recently developed method assumes that the expression levels of genes in a cluster are random samples drawn from a mixture of normal distributions, where each mixture component corresponds to a clustering of samples for that module, i.e. it performs a two-way co-clustering operation [60]. The method is available as part of the **Lemon-Tree** package (`https://github.com/eb00/lemon-tree`) and has been successfully used in a variety of applications [61].

The co-clustering is carried out by a Gibbs sampler which iteratively updates the assignment of each gene and, within each gene cluster, the assignment of each experimental condition. The co-clustering operation results the full posterior distribution, which can be written as

$$p(\mathcal{C} \mid \mathbf{X}) \propto \prod_{l=1}^{N} \prod_{u=1}^{L_l} \iint p(\mu, \tau) \prod_{i \in \mathcal{M}_l} \prod_{m \in \mathcal{E}_{l,u}} p(X_{im} \mid \mu, \tau) \, d\mu d\tau,$$

where $\mathcal{C} = \{M_l, \mathcal{E}_{l,u} : l = 1, \ldots, N; u = 1, \ldots, L_l\}$ is a co-clustering consisting of N gene modules M_l, each of which has a set of L_m sample clusters as $\mathcal{E}_{l,u}$; $p(X_{im} \mid \mu, \tau)$ is a normal distribution function with mean μ and precision τ; and $p(\mu, \tau)$ is a non-informative normal-gamma prior. Detailed investigations of the convergence properties of the Gibbs sampler showed that the best results are obtained by deriving consensus clusters from multiple independent runs of the sampler. In the **Lemon-Tree** package, consensus clustering is performed by a novel spectral graph clustering algorithm [62] applied to the weighted graph of pairwise frequencies with which two genes are assigned to the same gene module [61].

4 Causal gene networks

4.1 Using genotype data to prioritize edge directions in co-expression networks

Pairwise correlations between gene expression traits define undirected co-expression networks. Several studies have shown that pairs of gene expression traits can be causally ordered using genotype data [63–69]. Although varying in their statistical details, these methods conclude that gene A is causal for gene B, if expression of B associates significantly with A's eQTLs and this association is abolished by conditioning on expression of A and on any other known confounding factors. In essence, this is the principle of "Mendelian randomization", first introduced in epidemiology as an experimental design to detect causal effects of environmental exposures on human health [70], applied to gene expression traits.

To illustrate how these methods work, let A and B be two random variables representing two gene expression traits, and let E be a random variable representing a SNP which is an eQTL for gene A and B. Since genotype cannot be altered by gene expression (i.e. E cannot have any incoming edges), there are three possible regulatory models to explain the joint association of E to A and B:

1. $E \rightarrow A \rightarrow B$: the association of E to B is indirect and due to a causal interaction from A to B.
2. $E \rightarrow B \rightarrow A$: idem with the roles of A and B reversed.
3. $A \leftarrow E \rightarrow B$: A and B are independently associated to E.

To determine if gene A mediates the effect of SNP E on gene B (model 1), one can test whether conditioning on A abolishes the correlation between E and B, using the partial correlation coefficient

$$\mathrm{cor}(E, B \mid A) = \frac{\mathrm{cor}(E, B) - \mathrm{cor}(E, A)\mathrm{cor}(B, A)}{\sqrt{(1 - \mathrm{cor}(E, A)^2)(1 - \mathrm{cor}(B, A)^2)}.}$$

If model 1 is correct, then $\mathrm{cor}(E, B \mid A)$ is expected to be zero, and this can be tested for example using Fisher's Z transform to assess the significance of a sample correlation coefficient. The same approach can be used to test model 2, and if neither is significant, it is concluded that no inference on the causal direction between A and B can be made (using SNP E), i.e. that model 3 is correct. For more details, see [65], who have implemented this approach in the **NEO** software (http://labs.genetics.ucla.edu/horvath/htdocs/aten/NEO/).

Other approaches are based on the same principle, but use statistical model selection to identify the most likely causal model, with the probability density functions (PDF) for the models below:

- $p(E, A, B) = p(E)p(A \mid E)p(B \mid A)$,
- $p(E, A, B) = p(E)p(B \mid E)p(A \mid B)$,
- $p(E, A, B) = p(E)p(A \mid E)p(B \mid E, A)$,

where the dependence on A in the last term of the last model indicates that there may be a residual correlation between B and A not explained by E. The minimal additive model assumes the distributions are [66]

$$\begin{aligned} E &\sim \text{Bernoulli}(q), \\ A \mid E &\sim \text{N}(\mu_{A|E}, \sigma_A^2), \\ B \mid A &\sim \text{N}\left(\mu_B + \rho\frac{\sigma_B}{\sigma_A}(A - \mu_A), (1-\rho^2)\sigma_B^2\right), \\ B \mid E, A &\sim \text{N}\left(\mu_{B|E} + \rho\frac{\sigma_B}{\sigma_A}(A - \mu_{A|E}), (1-\rho^2)\sigma_B^2\right), \end{aligned}$$

so that E fulfills a Bernoulli distribution, $A \mid E$ undergoes a normal distribution whose mean depends on E, and that $B \mid A$ has a conditional normal distribution whose mean and variance are contributed in part by A. For $(B \mid E, A)$, the mean of B also depends on E. The parameters of all distributions can be estimated by maximum likelihood, and the model with the highest likelihood is selected as the most likely causal model. The number of free parameters can be accounted using penalties like the Akaike information criterion (AIC) [66].

The approach has been extended in various ways. In [64], likelihood ratio tests, comparison to randomly permuted data, and false discovery rate estimation techniques are used to convert the three model scores in a single probability value $P(A \rightarrow B)$ for a causal interaction from gene A to B. This method is available in the **Trigger** software (https://www.bioconductor.org/packages/release/bioc/html/trigger.html). In [69] and [68], the model selection task is recast into a single hypothesis test, using F-tests and Vuong's model selection test respectively, resulting in a significance p-value for each gene-gene causal interaction.

It should be noted that all of the above approaches suffer from limitations due to their inherent model assumptions. In particular, the presence of unequal levels of measurement noise among genes, or of hidden regulatory factors causing additional correlation among genes, can confuse causal inference. For example, excessive error level in the expression data of gene A, may mistake the true structure $E \rightarrow A \rightarrow B$ as $E \rightarrow B \rightarrow A$. These limitations are discussed in [18, 71].

4.2 Using Bayesian networks to identify causal regulatory mechanisms

Bayesian networks are probabilistic graphical models which encode conditional dependencies between random variables in a directed acyclic graph (DAG). Although Bayesian network cannot fully reflect certain pathways in gene regulation, such as self-regulation or feedback loops, they still serve as a popular method for modelling gene regulation networks, as

they provide a clear methodology for learning statistical dependency structures from possibly noisy data [72–74].

We adopt our previous convention in Section 2, where we have the gene expression data $\mathbf{X}$ and genetic markers $\mathbf{G}$. The model contains a total of k vertices (i.e. random variables), X_i with $i = 1, \ldots, k$, corresponding to the expression level of gene i. Given a DAG $\mathcal{G}$, and denoting the parental vertex set of X_i by $\mathbf{Pa}^{(\mathcal{G})}(X_i)$, the acyclic property of $\mathcal{G}$ allows to define the joint probability distribution function as

$$p(X_1, \ldots, X_k \mid \mathcal{G}) = \prod_{i=1}^{k} p(X_i \mid \mathbf{Pa}^{(\mathcal{G})}(X_i)). \tag{2}$$

In its simplest form, we model the conditional distributions as

$$p\big(X_i \mid \mathbf{Pa}^{(\mathcal{G})}(X_i)\big) = N\left(\alpha_i + \sum_{X_j \in \mathbf{Pa}^{(\mathcal{G})}(X_i)} \beta_{ji}(X_j - \alpha_j), \sigma_i^2\right),$$

where (α_i, σ_i) and β_{ji} are parameters for vertex X_i and edge $X_j \to X_i$ respectively, as part of the DAG structure $\mathcal{G}$. Under such modelling, the Bayesian network is called a linear Gaussian network.

The likelihood of data $\mathbf{X}$ given the graph $\mathcal{G}$ is

$$p(\mathbf{X} \mid \mathcal{G}) = \prod_{i=1}^{k} \prod_{l=1}^{n} p(X_{il} \mid \{X_{jl}, X_j \in \mathbf{Pa}^{(\mathcal{G})}(X_i)\}).$$

Using Bayes' rule, the log-likelihood of the DAG $\mathcal{G}$ based on the gene expression data $\mathbf{X}$ becomes

$$\log p(\mathcal{G} \mid \mathbf{X}) = \log p(\mathbf{X} \mid \mathcal{G}) + \log p(\mathcal{G}) - \log p(\mathbf{X}),$$

where $p(\mathcal{G})$ is the prior probability for $\mathcal{G}$, and $p(\mathbf{X})$ is a constant when the expression data is provided, so the follow-up calculations do not rely on it.

Typically, a locally optimal DAG is found by starting from a random graph and randomly ascending the likelihood by adding, modifying, or removing one directed edge at a time [72–74]. Alternatively, the posterior distribution $p(\mathcal{G} \mid \mathbf{X})$ can be estimated with Bayesian inference using Markov Chain Monte Carlo (MCMC) simulation, allowing us to estimate the significance levels at an extra computational cost. The parameter values of α, β, and σ, as part of $\mathcal{G}$, can be estimated with maximum likelihood.

When Bayesian network is modified by a single edge, only the vertices that receive a change would require a recalculation, whilst all others remain intact. This significantly reduces the amount of computation needed for each random step. A further speedup is achievable if we constrain the maximum number of parents each vertex can have, either by using the same fixed number for all nodes, or by pre-selecting a variable number of potential parents for each node using, for instance, a preliminary L_1-regularisation step [75].

Two DAGs are called Markov equivalent if they result in the same PDF [74]. Clearly, using gene expression data alone, Bayesian networks can only be resolved up to Markov equivalence. To break this equivalence and uncover a more specific causal gene regulation network, genotype data is incorporated in the model inference process. The most straightforward

approach is to use any of the methods in the previous section to calculate the probability $P(X_i \to X_j)$ of a causal interaction from X_i to X_j [63,76–78], for example by defining the prior as $p(\mathcal{G}) = \prod_{X_i} \left(\prod_{X_j \in \mathbf{Pa}^{(\mathcal{G})}(X_i)} P(X_j \to X_i) \prod_{X_j \notin \mathbf{Pa}^{(\mathcal{G})}(X_i)} (1 - P(X_j \to X_i)) \right)$. A more ambitious approach is to jointly learn the eQTL associations and causal trait (i.e. gene or phenotype) networks. In [79], EM is used to alternatingly map eQTLs given the current DAG structure, and update the DAG structure and model parameters given the current eQTL mapping. In [80], Bayesian networks are learned where SNPs and traits both enter as variables in the model, with the constraint that traits can depend on SNPs, but not vice versa. However, the additional complexity of both methods means that they are computationally expensive and have only been applied to problems with a handful of traits [79, 80].

A few additional "tips and tricks" are worth mentioning:

- First, when the number of vertices is much larger than the sample count, we may break the problem into independent sub-problems by learning a separate Bayesian network for each co-expression module (Section 3.1 and [78]). Dependencies between modules could then be learned as a Bayesian network among the module eigengenes [81], although this does not seem to have been explored.
- Second, Bayesian network learning algorithms inevitably result in locally optimal models which may contain a high number of false positives. To address this problem, we can run the algorithm multiple times and report an averaged network, only consisting of edges which appear sufficiently frequent.
- Finally, another technique that helps in distinguishing genuine dependencies from false positives is *bootstrapping*, where resampling with replacement is executed on the existing sample pool. A fixed number of samples are randomly selected and then processed to predict a Bayesian network. This process is repeated many times, essentially regarding the distribution of sample pool as the true PDF, and allowing to estimate the robustness of each predicted edge, so that only those with high significance are retained [82]. In theory, even the whole pipeline of Figure 1 up to the *in silico* validation could be simulated in this way. Although bootstrapping is computationally expensive and mostly suited for small datasets, it could be used in conjunction with the separation into modules on larger datasets.

4.3 Using module networks to identify causal regulatory mechanisms

Module network inference is a statistically well-grounded method which uses probabilistic graphical models to reconstruct modules of co-regulated genes and their upstream regulatory programs, and which has been proven useful in many biological case studies [61, 83, 84, 105]. The module network model was originally introduced as a method to infer regulatory networks from large-scale gene expression compendia, as implemented in the **Genomica** software (`http://genomica.weizmann.ac.il`) [83]. Subsequently the method has been extended to integrate eQTL and gene expression data [52, 85, 86]. The module network model starts from the same formula as Equation (2). It is then assumed that genes belonging to the same module share the same parents and conditional distributions; these conditional distributions are parameterized as decision trees, with the parental genes on the internal

(decision) nodes and normal distributions on the leaf nodes [83]. Recent algorithmic innovations decouple the module assignment and tree structure learning from the parental gene assignment and use Gibbs sampling and ensemble methods for improved module network inference [60, 87]. These algorithms are implemented in the **Lemon-Tree** software (`https://github.com/eb00/lemon-tree`), a command line software suite for module network inference [61].

4.4 Illustrative example

We have recently identified genome-wide significant eQTLs for 6,500 genes in seven tissues from the Stockholm Atherosclerosis Gene Expression (STAGE) study [12], and performed co-expression clustering and causal networks reconstruction [88]. To illustrate the above concepts, we show some results for a co-expression cluster in visceral fat (88 samples, 324 genes) which was highly enriched for tissue development genes ($P = 5 \times 10^{-10}$) and contained 10 genome-wide significant eQTL genes and 25 transcription factors, including eight members of the homeobox family (Figure 2a).

A representative example of an inferred causal interaction is given by the co-expression interaction between HAP1 (huntingtin-associated protein 1, chr17 q21.2-21.3) and FOXG1 (forkhead box G1, chr14 q11-q13). The expression of both genes is highly correlated ($\rho = 0.85$, $P = 4.4 \times 10^{-24}$, Figure 2b). HAP1 expression shows a significant, non-linear association with its eQTL rs1558285 ($P = 1.2 \times 10^{-4}$); this SNP also associates significantly with FOXG1 expression in the cross-association test ($P = 0.0024$), but not anymore after conditioning FOXG1 on HAP1 and its own eQTL rs7160881 ($P = 0.67$) (Figure 2c). In contrast, although FOXG1 expression is significantly associated with its eQTL rs7160881 ($P = 0.0028$), there is no association between this SNP and HAP1 expression ($P = 0.037$), and conditioning on FOXG1 and HAP1's eQTL has only a limited effect ($P = 0.19$) (Figure 2d). Using conditional independence tests (Section 4.1), this results in a high-confidence prediction that HAP1 $\rightarrow$ FOXG1 is causal.

A standard greedy Bayesian network search algorithm [75] was run on the aforementioned cluster of 324 genes. Figure 2e shows the predicted consensus sub-network of causal interactions between the 10 eQTLs and 25 TFs. This illustrates how a sparse Bayesian network can accurately represent the fully connected co-expression network (all 35 genes have high-mutual co-expression, cf. Figure 2a).

Figure 2f shows a typical regulatory module inferred by the **Lemon-Tree** software, also from the STAGE data. Here a heatmap is shown of the genotypes of an eQTL (top), the expression levels of a regulatory gene (middle), predicted to regulate a co-expression module of 11 genes (bottom). The red lines indicate sample clusters representing separate normal distributions inferred by the model-based co-clustering algorithm (Section 3.2).

5 *In silico* validation of predicted gene regulation networks

Gene regulation networks reconstructed from omics data represent hypotheses about the downstream molecular implications of genetic variations in a particular cell or tissue type.

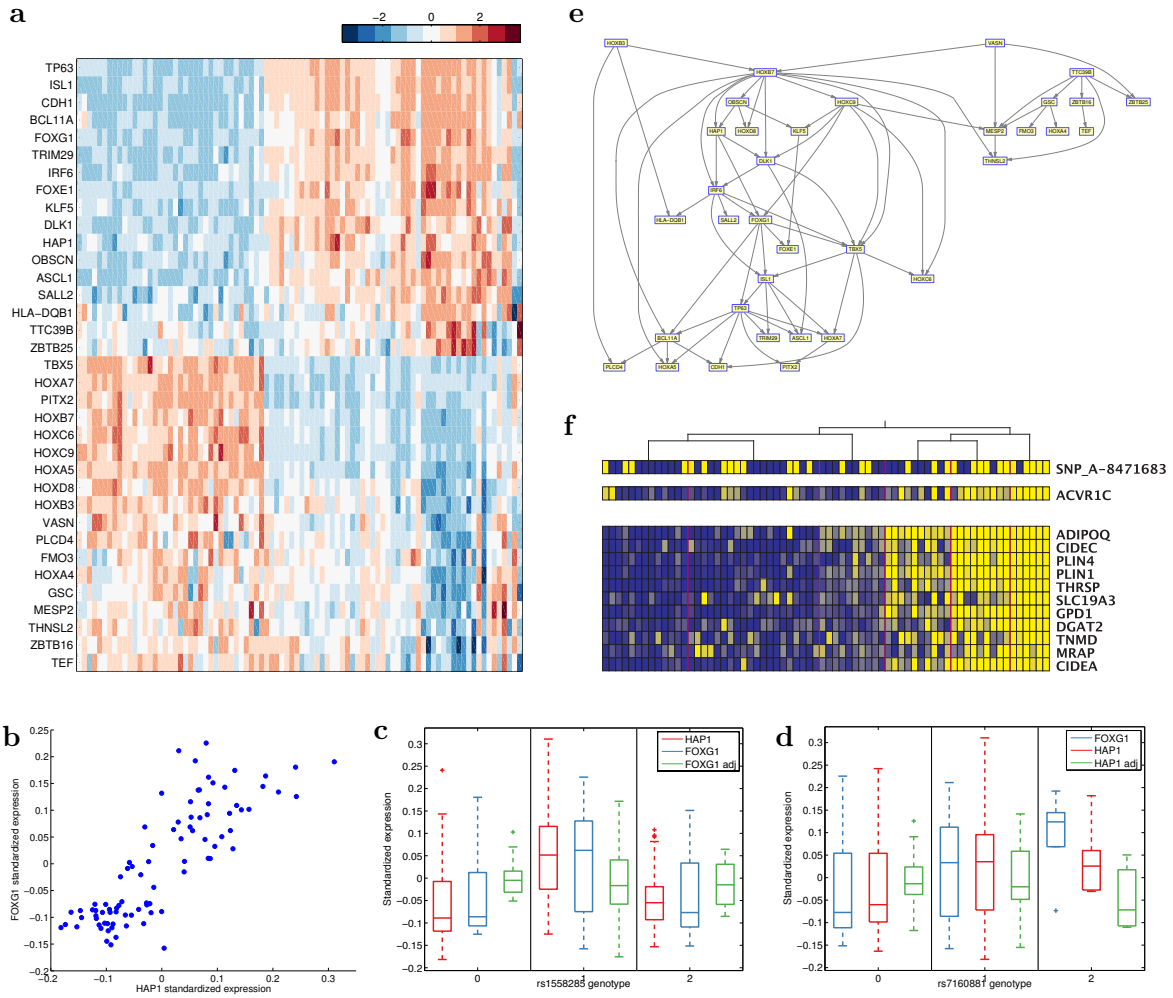

Figure 2: **(a)** Heatmap of standardized expression profiles across 88 visceral fat samples for 10 eQTL genes and 25 TFs belonging to a co-expression cluster inferred from the STAGE data. **(b)** Co-expression of HAP1 and FOXG1 across 88 visceral fat samples. **(c)** Association between HAP1's eQTL (rs1558285) and expression of HAP1 (red), FOXG1 (blue) and FOXG1 adjusted for HAP1 and FOXG1's eQTL (green). **(d)** Association between FOXG1's eQTL (rs7160881) and expression of FOXG1 (blue), HAP1 (red), and HAP1 adjusted for FOXG1 and HAP1's eQTL (green). **(e)** Causal interactions inferred between the same genes as in (a) using Bayesian network inference. **(f)** Example of a regulatory module inferred by **Lemon-Tree** from the STAGE data. See Section 4.4 for further details.

An essential first step towards using these networks in concrete applications (e.g. discovering novel candidate drug target genes and pathways) consists of validating them using independent data. The following is a non-exhaustive list of typical *in silico* validation experiments.

Model likelihood comparison and cross-validation. When different algorithms are used to infer gene network models, their log-likelihoods can be compared to select the best one. (With the caveat that the same data that was used to learn the models is used to compare them, this comparison is meaningful only when the algorithms optimize *exactly* the same (penalized) log-likelihood functions.) In a K-fold cross-validation experiment, the available samples are divided into K subsets of approximately equal size. For each subset, models are learned from a dataset consisting of the $K-1$ other subsets, and the model likelihood is calculated using only the unseen data subset. Thus, cross-validation is used to test the generalisability of the inferred network models to unseen data. For an example where model likelihood comparison and cross-validation were used to compare two module network inference strategies, see [87].

Functional enrichment. Organism-specific gene ontology databases contain structured functional gene annotations [89]. These databases can be used to construct gene signature sets composed of genes annotated to the same biological process, molecular function or cellular component. Reconstructed gene networks can then be validated by testing for enriched connectivity of gene signature sets using a method proposed by [76]. For a given gene set, this method considers all network nodes belonging to the set and their nearest neighbours, and from this set of nodes and edges, the largest connected sub-network is identified. Then the enrichment of the gene set in this sub-network is tested using the Fisher exact test and compared to the enrichment of randomly selected gene sets of the same size.

Comparison with physical interaction networks. Networks of transcription factor - target interactions based on ChIP-sequencing data [90] from diverse cell and tissue types are available from the **ENCODE** [91], **Roadmap Epigenomics** [92] and **modENCODE** [93–95] projects, while physical protein-protein interaction networks are available for many organisms through databases such as the **BioGRID** [96]. Due to indirect effects, networks predicted from gene expression data rarely show a significant overlap with networks of direct physical interactions. A more appropriate validation is therefore to test for enrichment for short connection paths in the physical networks between pairs predicted to interact in the reconstructed networks [61].

Gene perturbation experiments. Gene knock-out experiments provide the ultimate gold standard of a causal network intervention, and genes differentially expressed between knock-out and control experiments can be considered as true positive direct or indirect targets of the knocked-out gene. Predicted gene networks can be validated by compiling relevant (i.e. performed in a relevant cell or tissue type) gene knock-out experiments from the **Gene Expression Omnibus** (`http://www.ncbi.nlm.nih.gov/geo/`) or **ArrayExpress** (`https://www.ebi.ac.uk/arrayexpress/`) and comparing the overlap between gene sets responding to a

gene knock-out and network genes predicted to be downstream of the knocked-out gene. Overlap significance can be estimated by using randomized networks with the same degree distribution as the predicted network.

6 Future perspective: Integration of multi-omics data

Although combining genotype and transcriptome data to reconstruct causal gene networks has led to important discoveries in a variety of applications [21], important details are not incorporated in the resulting network models, particularly regarding the causal molecular mechanisms linking eQTLs to their target genes, and the relation between variation in transcript levels and protein levels, with the latter ultimately determining phenotypic responses. Several recent studies have shown that at the molecular level, *cis*-eQTLs primarily cause variation in transcription factor binding to gene regulatory DNA elements, which then causes changes in histone modifications, DNA methylation and mRNA expression of nearby genes (reviewed in [15]). Although mRNA expression can be used as a surrogate for protein expression, due to diverse post-transcriptional regulation mechanisms, the correlation between mRNA and protein levels is known to be modest [97,98], and genetic loci that affect mRNA and protein expression levels do not always overlap [28,99]. Thus, an ideal systems genetics study would integrate genotype data and molecular measurements at all levels of gene regulation from a large number of individuals.

Human lymphblastoid cell lines (LCL) are emerging as the primary model system to test such a approach. Whole-genome mRNA and micro-RNA sequencing data are available for 462 LCL samples from five populations genotyped by the 1000 Genomes Project [35]; protein levels from quantitative mass spectrometry for 95 samples [99]; ribosome occupancy levels from sequencing of ribosome-protected mRNA for 50 samples [100]; DNA-occupancy levels of the regulatory TF PU.1, the RNA polymerase II subunit RBP2, and three histone modifications from ChIP-sequencing of 47 samples [101]; and the same three histone modifications from ChIP-sequencing of 75 samples [102]. These population-level datasets can be combined further with three-dimensional chromatin contact data from Hi-C [103] and ChIA-PET [102], knock-down experiments followed by microarray measurements for 59 transcription-associated factors and chromatin modifiers [104], as well as more than 260 ENCODE assays (including ChIP-sequencing of 130 TFs) [91] in a reference LCL cell line (GM12878). Although the number of samples where all measures are simultaneously available is currently small, this number is sure to rise in the coming years, along with the availability of similar measurements in other cell types. Despite the challenging heterogeneity of data and analyses in the integration of multi-omics data, web-based toolboxes, such as **GenomeSpace** (`http://www.genomespace.org`) [105] can prove helpful to non-programmer researchers.

7 Conclusions

In this chapter we have reviewed the main methods and softwares to carry out a systems genetics analysis, which combines genotype and various omics data to identify eQTLs and their associated genes, reconstruct co-expression networks and modules, reconstruct causal

Bayesian gene and module networks, and validate predicted networks *in silico*. Several method and software options are available for each of these steps, and by necessity a subjective choice about which ones to include had to be made, based largely on their ability to handle large datasets, their popularity in the field, and our personal experience of using them. Where methods have been compared in the literature, they have usually been performed on a small number of datasets for a specific subset of tasks, and results have rarely been conclusive. That is, although each of the presented methods will give somewhat different results, no objective measurements will consistently select one of them as the "best" one. Given this lack of objective criterion, the reader may well prefer to use a single software that allows to perform all of the presented analyses, but such an integrated software does not currently exist.

Nearly all of the examples discussed referred to the integration of genotype and transcriptome data, reflecting the current dominant availability of these two data types. However, omics technologies are evolving at a fast pace, and it is clear that data on the variation of TF binding, histone modifications, and post-transcriptional and protein expression levels will soon become more widely available. Developing appropriate statistical models and computational methods to infer causal gene regulation networks from these multi-omics datasets is surely the most important challenge for the field.

Acknowledgements

The authors' work is supported by the BBSRC [BB/M020053/1] and Roslin Institute Strategic Grant funding from the BBSRC [BB/J004235/1].

References

[1] Mackay TF, Stone EA and Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**:565–577 (2009).

[2] Goddard ME and Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**:381–391 (2009).

[3] Manolio TA. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics* **14**:549–558 (2013).

[4] Hindorff LA *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**:9362–9367 (2009).

[5] Schaub MA *et al.* Linking disease associations with regulatory information in the human genome. *Genome Research* **22**:1748–1759 (2012).

[6] Rockman MV and Kruglyak L. Genetics of global gene expression. *Nature Reviews Genetics* **7**:862–872 (2006).

[7] Georges M. Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annu Rev Genomics Hum Genet* **8**:131–162 (2007).

[8] Cookson W *et al.* Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**:184–194 (2009).

[9] Cheung VG and Spielman RS. Genetics of human gene expression: mapping dna variants that influence gene expression. *Nature Reviews Genetics* **10**:595–604 (2009).

[10] Cubillos FA, Coustham V and Loudet O. Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. *Current Opinion in Plant Biology* **15**:192–198 (2012).

[11] Dimas AS *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**:1246–1250 (2009).

[12] Foroughi Asl H *et al.* Expression quantitative trait loci acting across multiple tissues are enriched in inherited risk of coronary artery disease. *Circulation: Cardiovascular Genetics* **8**:305–315 (2015).

[13] Greenawalt DM *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Research* **21**:1008–1016 (2011).

[14] Ardlie KG *et al.* The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**:648–660 (2015).

[15] Albert FW and Kruglyak L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16**:197–212 (2015).

[16] Williams RW. Expression genetics and the phenotype revolution. *Mammalian Genome* **17**:496–502 (2006).

[17] Kadarmideen HN, von Rohr P and Janss LL. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome* **17**:548–564 (2006).

[18] Rockman MV. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature* **456**:738–744 (2008).

[19] Schadt EE. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**:218–223 (2009).

[20] Schadt EE and Björkegren JL. New: network-enabled wisdom in biology, medicine, and health care. *Science Translational Medicine* **4**:115rv1–115rv1 (2012).

[21] Civelek M and Lusis AJ. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* **15**:34–48 (2014).

[22] Björkegren JL *et al.* Genome-wide significant loci: How important are they?: Systems genetics to understand heritability of coronary artery disease and other common complex disorders. *Journal of the American College of Cardiology* **65**:830–845 (2015).

[23] Chen Y *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**:429–435 (2008).

[24] Schadt EE, Friend SH and Shaywitz DA. A network view of disease and compound screening. *Nat Rev Drug Disc* **8**:286–295 (2009).

[25] Walhout AJ. Unraveling transcription regulatory networks by protein–DNA and protein–protein interaction mapping. *Genome Research* **16**:1445–1454 (2006).

[26] Ritchie MD *et al.* Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics* **16**:85–97 (2015).

[27] Stegle O *et al.* Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* **7**:500–507 (2012).

[28] Foss EJ *et al.* Genetic basis of proteome variation in yeast. *Nature Genetics* **39**:1369–1375 (2007).

[29] Nicholson G *et al.* A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genetics* **7**:e1002270 (2011).

[30] Laird N and Lange C. *The Fundamentals of Modern Statistical Genetics* (Springer2011).

[31] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**:1353–1358 (2012).

[32] Qi J *et al.* kruX: Matrix-based non-parametric eQTL discovery. *BMC Bioinformatics* **15**:11 (2014).

[33] Brem RB *et al.* Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**:752–755 (2002).

[34] Schadt EE *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**:e107 (2008).

[35] Lappalainen T *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**:506–511 (2013).

[36] Golub GH and Van Loan CF. *Matrix computations* (The Johns Hopkins University Press1996), third edn.

[37] Hemani G *et al.* Detection and replication of epistasis influencing transcription in humans. *Nature* **508**:249–253 (2014).

[38] Sharan R, Ulitsky I and Shamir R. Network-based prediction of protein function. *Molecular Systems Biology* **3**:88 (2007).

[39] Daub CO *et al.* Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics* **5**:118 (2004).

[40] Butte A and Kohane I. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomputing* **5**:415–426 (2000).

[41] Basso K *et al.* Reverse engineering of regulatory networks in human b cells. *Nat Genet* **37**:382–390 (2005).

[42] Faith JJ *et al.* Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**:e8 (2007).

[43] Hartwell LH *et al.* From molecular to modular cell biology. *Nature* **402**:C47–C52 (1999).

[44] Eisen MB *et al.* Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**:14863–14868 (1998).

[45] Tavazoie S *et al.* Systematic determination of genetic network architecture. *Nature Genetics* **22**:281–285 (1999).

[46] Sharan R and Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. In *Proc Int Conf Intell Syst Mol Biol*, vol. 8, 16 (2000).

[47] Medvedovic M and Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**:1194–1206 (2002).

[48] Newman MEJ and Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* **69**:026113 (2004).

[49] Newman MEJ. Modularity and community structure in networks. *PNAS* **103**:8577–8582 (2006).

[50] Ayroles JF *et al.* Systems genetics of complex traits in drosophila melanogaster. *Nat Genet* **41**:299–307 (2009).

[51] Langfelder P and Horvath S. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559 (2008).

[52] Lee SI *et al.* Learning a prior on regulatory potential from eqtl data. *PLoS Genetics* **5**:e1000358 (2009).

[53] Clauset A, Newman MEJ and Moore C. Finding community structure in very large networks. *Phys Rev E* **70**:066111 (2004).

[54] Van Dongen SM. Graph clustering by flow simulation (2001).

[55] Enright AJ, Van Dongen S and Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**:1575–1584 (2002).

[56] Ravasz E *et al.* Hierarchical organization of modularity in metabolic networks. *Science* **297**:1551–1555 (2002).

[57] Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**:17 (2005).

[58] Langfelder P, Zhang B and Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics* **24**:719–720 (2008).

[59] Liu JS. *Monte Carlo strategies in scientific computing* (Springer2002).

[60] Joshi A, Van de Peer Y and Michoel T. Analysis of a Gibbs sampler for model based clustering of gene expression data. *Bioinformatics* **24**:176–183 (2008).

[61] Bonnet E, Calzone L and Michoel T. Integrative multi-omics module network inference with Lemon-Tree. *PLoS Computational Biology* **11** (2015).

[62] Michoel T and Nachtergaele B. Alignment and integration of complex networks by hypergraph-based spectral clustering. *Physical Review E* **86**:056111 (2012).

[63] Zhu J *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res* **105**:363–374 (2004).

[64] Chen LS, Emmert-Streib F and Storey JD. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology* **8**:R219 (2007).

[65] Aten JE *et al.* Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Systems Biology* **2**:34 (2008).

[66] Schadt EE *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37**:710–717 (2005).

[67] Neto EC *et al.* Inferring causal phenotype networks from segregating populations. *Genetics* **179**:1089–1100 (2008).

[68] Neto EC *et al.* Modeling causality for pairs of phenotypes in system genetics. *Genetics* **193**:1003–1013 (2013).

[69] Millstein J *et al.* Disentangling molecular relationships with a causal inference test. *BMC Genetics* **10**:23 (2009).

[70] Smith GD and Ebrahim S. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**:1–22 (2003).

[71] Li Y *et al.* Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in Genetics* **26**:493–498 (2010).

[72] Friedman N, Nachman I and Peér D. Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, 206–215 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA1999).

[73] Friedman N *et al.* Using bayesian networks to analyze expression data. *Journal of Computational Biology* **7**:601–620 (2000).

[74] Koller D and Friedman N. *Probabilistic Graphical Models: Principles and Techniques* (The MIT Press2009).

[75] Schmidt M, Niculescu-Mizil A and Murphy K. Learning graphical model structure using L1-regularization paths. In *AAAI*, vol. 7, 1278–1283 (2007).

[76] Zhu J *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics* **40**:854–861 (2008).

[77] Zhu J *et al.* Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biology* **10**:e1001301 (2012).

[78] Zhang B *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**:707–720 (2013).

[79] Neto EC *et al.* Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics* **4**:320 (2010).

[80] Scutari M *et al.* Multiple quantitative trait analysis using Bayesian networks. *Genetics* **198**:129–137 (2014).

[81] Langfelder P and Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**:54 (2007).

[82] Friedman N, Goldszmidt M and Wyner A. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 196–205 (Morgan Kaufmann Publishers Inc.1999).

[83] Segal E *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**:166–167 (2003).

[84] Friedman N. Inferring cellular networks using probabilistic graphical models. *Science* **308**:799–805 (2004).

[85] Lee S *et al.* Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci USA* **103**:14062–14067 (2006).

[86] Zhang W *et al.* A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computational Biology* **6**:e1000642 (2010).

[87] Joshi A *et al.* Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics* **25**:490–496 (2009).

[88] Talukdar H *et al.* Cross-tissue regulatory gene networks in coronary artery disease. *Cell Systems* **2**:196–208 (2016).

[89] Ashburner M *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**:25–29 (2000).

[90] Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* **13**:840–852 (2012).

[91] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74 (2012).

[92] Kundaje A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**:317–330 (2015).

[93] Gerstein M *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**:1775–1787 (2010).

[94] Roy S *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**:1787–1797 (2010).

[95] Yue F *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**:355–364 (2014).

[96] Chatr-aryamontri A *et al.* The BioGRID interaction database: 2015 update. *Nucleic acids research* gku1204 (2014).

[97] Lu P *et al.* Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature Biotech* **25**:117–124 (2007).

[98] Schwanhausser B *et al.* Global quantification of mammalian gene expression control. *Nature* **473**:337–342 (2011).

[99] Wu L *et al.* Variation and genetic control of protein abundance in humans. *Nature* **499**:79–82 (2013).

[100] Cenik C *et al.* Integrative analysis of rna, translation and protein levels reveals distinct regulatory variation across humans. *Genome Research* doi:10.1101/gr.193342.115 (2015).

[101] Waszak SM *et al.* Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**:1039–1050 (2015).

[102] Grubert F *et al.* Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**:1051–1065 (2015).

[103] Rao SS *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:1665–1680 (2014).

[104] Cusanovich DA *et al.* The functional consequences of variation in transcription factor binding. *PLoS Genetics* **10**:e1004226 (2014).

[105] Qu K *et al.* Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nature Methods* **13**:245–247 (2016).