Computational problems in multi-tissue models of health and disease

Manikandan Narayanan[1]

July 2017

[1] Systems Genomics and Bioinformatics Unit, Laboratory of Systems Biology, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Bethesda, MD, USA 20892. Email: manikandan.narayanan@nih.gov

Abstract

A modern development at the interface of computer science and systems biology is being fostered by high-dimensional molecular data emerging on multiple tissues of the same individual collected across large groups of healthy/diseased individuals. We review computational and statistical problems that arise in analyzing such multi-tissue genomic datasets, specifically problems posing new challenges compared to their single-tissue counterparts, such as ones related to missing data imputation, statistical learning of high-dimensional network models capturing gene-gene correlations within/across tissues, and graph algorithms to identify genes clustering across many tissue networks. A recurring research theme is the potential to integrate or pool information from across tissues to enhance power of detecting signals shared across tissues while also accounting for tissue-specific differences. We show how methods harnessing this integrative potential to address multi-tissue problems ranging from correlation/causal network inference to graph algorithms are ushering in an era of integrated, whole-system modeling of life processes.

Keywords: Bioinformatics, Computational systems biology, Genomic data science, Multi-tissue data, Biomolecular networks, Gene networks, Intra/inter-tissue networks, Graph algorithms, Whole-body/system models.

1 Introduction

We are at a juncture where modern computer science encourages taking a "computational or data science lens" to probe high-dimensional data in other sciences. Computational and statistical methods for instance play a central role in modern biological sciences by enabling systematic analysis of large biomolecular datasets, and thereby help build quantitative/predictive models of complex life processes and uncover their molecular underpinnings [8]. Genome-wide (also known as genomic) data on genes, proteins or other biomolecules have mostly been measured in a single tissue or cell type across several individuals, however biomolecular data measured in multiple tissues of large groups of healthy/diseased individuals are very recently and rapidly emerging. Such multi-tissue datasets open up rich problems relating to high-dimensional statistical inference/learning (two terms used in-

terchangeably in this text) and computational algorithms like graph algorithms, especially when we utilize the data to infer and analyze whole-system network models of complex behaviors (Figure 1).

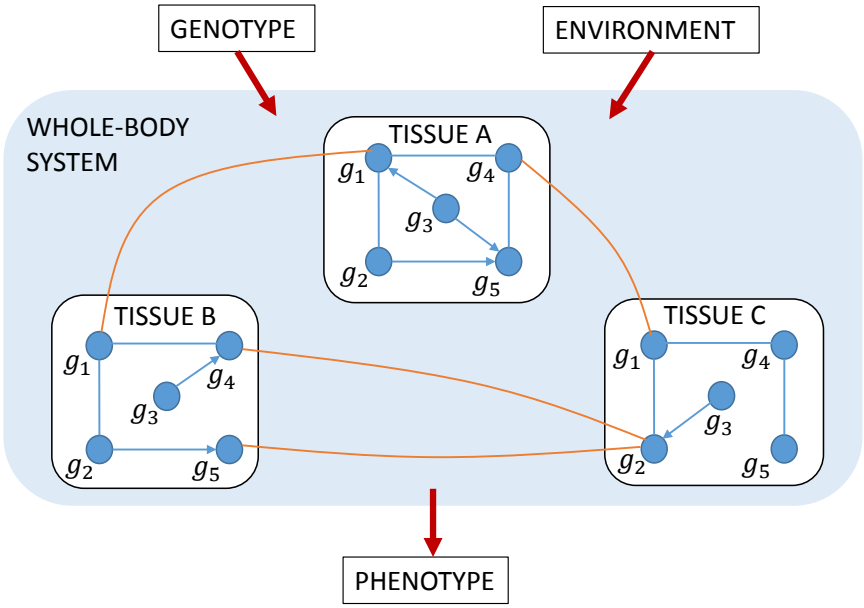


Figure 1: A multi-tissue biological system viewed as a "genotype × environment → phenotype" map. Genetic (genotype) and environmental factors influence how genes and other biomolecules ($g_1, g_2, \ldots$) in different tissues interact with and regulate each other as part of a whole-system network (shown as links between genes) to produce a healthy trait or disease endpoint (phenotype). Note that gene activity or expression data is measured in each tissue as the same gene could be regulated differently in different tissues, whereas the genetic data is assumed identical across tissues as all cells in the body having descended from the same embryo are genetically identical (except for acquired mutations as in cancer).

Multi-tissue analysis faces unique computational and statistical challenges stemming from the *high-dimensional*, *noisy*, *transposable* and *heterogeneous* nature of multi-tissue data. Some of these properties but not all are shared with single-tissue data.

- *High-dimensionality* of genomic data refers to the "small n, large p" issue, where the number n of samples is typically much fewer than the number p of features (e.g., tens of thousands of biomolecules or millions of genetic variants). So, we need additional *a priori* assumptions [3] to infer the large number of system parameters (e.g., relationship strength between all pairs of biomolecules) from small n, and computationally efficient methods that can scale well to inferring and analyzing large systems.
- *Noise* arising from measurement noise, sampling/coverage bias, false positive/negative errors and especially batch effects (i.e., bias from having to generate a large dataset over an extended period of time in distinct batches) affect data from large-scale genomic experiments more so than data from small-scale reference experiments, and hence experimental design to data analysis strategies in genomics need careful thought.
- *Transposability* of a multi-tissue data matrix refers to the property that both its rows (biomolecules) and columns (tissues) are fea-

tures with dependencies [35], due to coordination of genes acting in the same biological process, developmental history of tissues, etc. Two-way dependent or transposable data may require extension of classic statistical methods, which typically deal with one-way dependent data matrices (e.g., single-tissue data with rows as dependent features/biomolecules and columns as independent samples measuring the features).
- Multi-tissue *data integration* is challenging due to *heterogeneity* (variation) in the molecular processes governing different tissues, and incomplete knowledge on the extent (or lack thereof) of this heterogeneity. So new methods are needed to quantify the heterogeneity of biomolecular activities across tissues, strengthen inferences of activities within a tissue by borrowing information from other tissues found to have similar signals, and approach system-wide understanding by also inferring tissue-tissue communication from data.

This review highlights the diverse types of computational problems that aim for non-trivial integration of data from different tissues and individuals by addressing the challenges above. Outside the scope of this review are works that repeat a single-tissue analysis on each tissue separately and then summarize/compare results in the end - while these works are useful, we would like to focus on problems unique to multi-tissue data integration. We also do not cover single-individual, multi-tissue data (also called "expression atlas") in detail, as we are more interested in inferring network models from several individuals' data. For each of the problem types unique to multi-tissue analysis that we discuss, we highlight one or two recent methods that illustrate the problem best in our opinion to maintain brevity, and encourage interested readers to follow the illustrative methods' publications and references therein to explore other related methods in this burgeoning field.

2 Biomedical motivation and background

We first provide biomedical motivation behind multi-tissue studies and the nature and scale of multi-tissue data, so as to provide a concrete context for understanding the computational problems.

2.1 Rationale for multi-tissue network models

Good health emerges from proper integrated functioning of all organs and tissues within our body. At the basis of healthy tissue function and proper inter-tissue (tissue-tissue) communication are the myriad genes, proteins and other biomolecules that interact with and regu-

late each other in a concerted fashion. A collection or network of all such interactions among biomolecules, which reside either inside cells composing various tissues or in surrounding body fluids connecting the tissues, could move us towards an integrated, whole-system view of complex life processes [13] (Figure 1).

Many complex diseases affect multiple tissues, and a whole-system network model could also offer an integrated view of such a disease and thereby promote new strategies for its prevention, diagnosis and treatment. For instance, type 2 diabetes is a chronic human disease that is not localized to any one tissue - the glucose-insulin imbalance is a systemic property of the disease involving interactions among various tissues including pancreas, liver, muscle and adipose among others. A whole-system network model of diabetes would explain the complex cascade of dysregulation events in the disease (by clarifying for instance which genes in which tissues cause early stages of the disease, which gene products act at tissue-tissue interfaces to maintain/worsen/propagate the disease, and which genes get affected in later disease stages), and enable simulations of the effect of perturbations like potentially new medicines or unstudied genetic variants.

2.2 Emerging data in multi-tissue genomics

Modern biology enables a data-driven approach to build increasingly system-wide (if not yet whole-system) network models of health and disease. With the help of modern high-throughput technologies, genome-wide biomolecular data on multiple tissues of an individual, collected across large groups of healthy or diseased individuals, are now being rapidly generated and released in the public domain. These genomic data collected from each individual could range from DNA sequences (genetic variants or genotype) of the individual to his/her tissue-specific data on expression (activity level) of all genes encoded in the genome, proteins translated from the expressed genes, or small-molecule metabolites. Technologies that enable high-throughput measurement of DNA sequences or gene expression levels include microarrays and next generation sequencing (DNAseq, RNAseq), and that of protein or metabolite expression levels include mass spectrometry.

In humans, the most recent and popular example of such multi-tissue genomics data is from the Genotype-Tissue Expression (GTEx) consortium - the pilot phase of this project released expression measurements of the roughly 20,000 genes in the human genome across 175 individuals in 43 tissues [18]. Not all tissues could be profiled in each individual due to varying RNA quality in the postmortem tissue samples, however 9 high-priority tissues (adipose, artery, heart, lung, muscle, nerve, skin, thyroid and blood) have been profiled more fre-

quently than other tissues. Another example of a multi-tissue dataset comes from a study of Alzheimer's disease - genome-wide gene expression data in multiple brain regions (cerebellum, visual cortex and prefrontal cortex tissue samples collected postmortem) are available for 100s of healthy and diseased individuals [40]. In another study, 856 twins were expression profiled using samples from three tissues (adipose, lymphoblastoid cell lines and skin tissues) [17]. Moving beyond gene expression, metabolite expression data is also available for 2251 metabolites across 374 participants in multiple body fluids (plasma, urine, and saliva) [11]. Many of these studies with expression data also have matched genetic data (genetic variants assayed or imputed at 6+ million variant sites across the genome) and clinical data (endpoints such as disease status, blood sugar or lipid levels, etc.), so that heritability i.e., genetic control of expression and clinical traits can be studied. Similar multi-tissue genomic datasets are also available for other model organisms such as mouse [39, 20, 37].

Such multi-tissue data can be accessed from project-specific websites (e.g., `https://www.gtexportal.org` for the GTEx project) or public database repositories where data from different projects get posted. Gene expression data, the most openly available data type, can be accessed from the NCBI GEO or EBI ArrayExpress repositories. Genetic and phenotypic data can be accessed from the NCBI dbGAP or EBI EGA repositories in a controlled fashion, which helps protect privacy of study subjects, by researchers who formally apply for data access. Meta-data capturing experimental and clinical context of these datasets may also be in these repositories, but could be incomplete or harder to retrieve due to fewer standards around reporting them. Still, it is important to strive to obtain meta-data such as processing date of each sample in order to mitigate batch effects that critically impact large-scale studies [24], and patient medication history to properly interpret analysis results. The scientific community is realizing the power of open data and meta-data in accelerating biomedical research, and many funders/journals are also mandating public data release when projects are funded/published by them. This trend bodes well for current and future research in the highly data-driven field of multi-tissue analysis.

3 Challenging multi-tissue problems

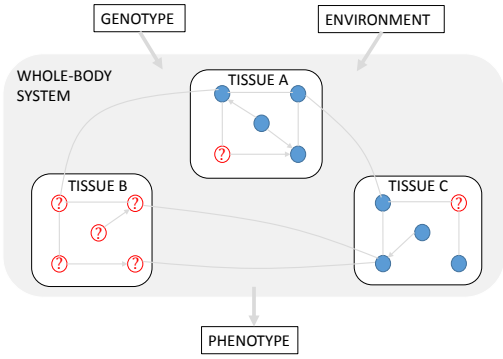
3.1 Multi-tissue data representation, imputation and testing

A multi-tissue dataset can be represented as a 3D matrix $\boldsymbol{D}$, with $\boldsymbol{D}_{ijk}$ indicating an individual i's measurements of the expression level or activity of biomolecule j in tissue k. Biomolecules are often genes in

the problems we discuss, but could also be proteins or metabolites (so sometimes when the context is clear, we simply use the term gene to refer to a gene or any other biomolecule). If we consider each individual as a random sample from the population of all individuals (or subpopulations such as a healthy or diseased group of individuals), then 2D slices of this matrix corresponding to different individuals are a more natural representation for statistical analysis. In this scenario, individual i's data is represented as the 2D slice $\boldsymbol{X}^{(i)}=\boldsymbol{D}_{i**}$ and is considered as one realization of the matrix-valued random variable $\boldsymbol{X}$.

Data imputation is a key problem in many multi-tissue studies since all tissues are often not measured in all individuals (due to variability in accessibility of tissues, postmortem tissue quality, etc.), and imputing such unmeasured tissues (different columns of different $\boldsymbol{X}^{(i)}$ matrices) to obtain complete data matrices can simplify and strengthen several downstream analyses. Single-tissue imputation methods like k-nearest neighbor genes, applied as is to multi-tissue data, would only borrow information from other genes (based on gene-gene or row-row correlations across individuals in $\boldsymbol{X}$). Multi-tissue setting offers the flexibility to borrow information from other tissues as well (based on tissue-tissue or column-column correlations in $\boldsymbol{X}$) and further utilize genetic data on the same individuals if available. MixRF is a recently proposed imputation method that operates in this integrative fashion [36] - it couples a robust random forest approach with a mixed-effects model for each gene, whose outcome variable is the expression vector of the gene across all tissues in an individual and predictor variables are genetic factors controlling gene expression, known covariates like gender of the individual, and top principal components of every tissue (which borrow information across genes to capture environmental/developmental factors encoded in multiple tissues). MixRF was able to impute several unmeasured tissues in the GTEx data and could be applied to other studies beyond GTEx using GTEx data as the reference. The running time of MixRF scales linearly with the number of genes and predictor variables, and as $O(n_t \log n_t)$ with the number n_t of observed tissues summed across all individuals.

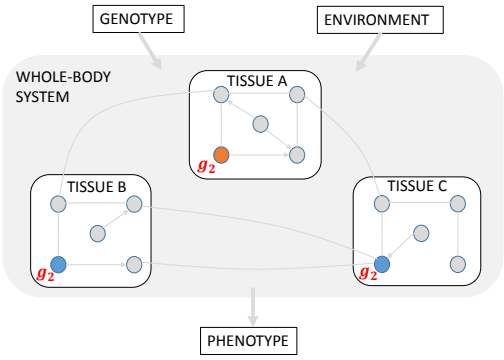
An earlier imputation study took a different approach to exploit the transposable or two-way dependence structure (row-row and column-column correlations) of $\boldsymbol{X}$ [1]. It assumed the expression data $\boldsymbol{X}$ to follow a matrix-variate normal distribution, with the two-way dependence parameterized using separate row and column covariance matrices. Specifically, the covariance matrix of a strung-out vector of $\boldsymbol{X}$ is the Kronecker product of the positive-definite row and column covariance matrices $\boldsymbol{\Sigma^{(r)}}$ and $\boldsymbol{\Sigma^{(c)}}$ (i.e., $Cov(\boldsymbol{X}_{jk}, \boldsymbol{X}_{j'k'}) = \boldsymbol{\Sigma}^{(\boldsymbol{r})}_{jj'} \times \boldsymbol{\Sigma}^{(\boldsymbol{c})}_{kk'}$). There is also a matrix $\boldsymbol{M}$ of constants to specify the mean parameter of gene j in tissue k (i.e., $\mathbb{E}(\boldsymbol{X}_{jk}) = \boldsymbol{M}_{jk}$). Under these multivariate



Problem: Data imputation

normality assumptions and a penalized maximum likelihood approach whose penalties encourage sparsity or shrinkage of the inverse covariance matrices to deal with the high-dimensionality i.e., "small n, large p" issue, regularized estimators for the row/column parameters are derived and an EM-type algorithm developed to impute missing multi-tissue data [1]. The imputation algorithm has a prohibitive running time as it scales cubically with the number of rows or columns of $\boldsymbol{X}$. The key contribution of this study however is the transposable regularized covariance models (TRCM), which paved the way for fruitful works on other multi-tissue problems.

Differential expression analysis or testing is one such multi-tissue problem that could be addressed with TRCM-like models. It concerns statistical hypothesis tests about the mean parameter $\boldsymbol{M}$ in the model for $\boldsymbol{X}$ - we may want to know if the mean expression of a gene is the same or different between two groups of tissues (e.g., identify genes that are differentially expressed in all brain-related tissues compared to the rest say, and thereby understand which genes and biological processes are brain-specific). Classic statistical tests like t-test of a gene's mean in two tissue groups assume independence of tissue samples, and classic multiple-testing procedures for controlling error rates across all tested genes also assume independence or weak dependence between the genes. Multi-tissue data could exhibit strong two-way dependences. One possible solution to this problem is to sphere or decorrelate the data before applying classic tests - a work transformed $\boldsymbol{X}$ using TRCM model estimates to make its rows and columns approximately independent, and found that applying classic statistical methods on the transformed or decorrelated $\boldsymbol{X}$ yielded test statistics that better followed null distributions and multiple-testing procedures that better controlled error rates across all tested genes [2]. Another work on hypothesis tests about the means of different row or column subgroups of $\boldsymbol{X}$ generalized the TRCM model along a non-parametric direction (see references in [35]). One of their contributions is the development of new regularized covariance estimators and hypothesis tests about the mean and also covariance matrices, assuming a transposable covariance structure as in TRCM models but without assuming normality [35]. Some related open problems mentioned in these studies include hypothesis tests about the mean of not just row or column subgroups but simultaneously predefined row and column subgroups, and efficient tests of whether the covariance of a multi-tissue data indeed satisfies a Kronecker product structure to justify the application of TRCM-based methods.



Problem: Differential expression analysis or testing

3.2 Statistical inference of multi-tissue models

Intra-tissue problems like inferring network models of biomolecular interactions within a specific tissue, which can be addressed using single-tissue data, permit new approaches that borrow information from other tissues in a multi-tissue setting. Inter-tissue problems like inferring interactions between genes in different tissues, which are simply inconceivable using only single-tissue data, are natural to pose in a multi-tissue study. We highlight both classes of multi-tissue problems here, and focus on methods that infer networks of correlated biomolecules (coexpression networks) from multi-tissue data and open problems that go beyond correlation to causation. As these problems benefit from the rich information and causal structure provided by simultaneously obtained genetic data, we start with problems on genetic control of tissue activities, an interesting topic in its own right.

Inferring genetic control of tissue gene expression

Quantitative Trait Loci (QTL) is a locus or site in the genome whose variation at the DNA level across a population of individuals correlates with a specific phenotype like height. Finding QTLs for a disease phenotype (disease case-control status or related clinical endpoint like blood cholesterol level), as done in several Genome-Wide Association Studies (GWAS), could be a powerful first step to discover genes proximal to the QTL that cause disease. Such causal discovery from observed correlations is justified if confounding factors are accounted for in the GWAS analysis or do not affect the individuals under independent random sampling assumptions. As such, GWAS offers a systematic approach to causality mapping in a species like humans where perturbation experiments for causal discovery are limited for ethical reasons. GWAS analyses need to be computationally efficient as millions of genetic factors in the genome are scanned for QTL association to each studied phenotype. A genetic factor here refers to any site in the genome harboring a genetic variant - the most popular one being a Single Nucleotide Polymorphism or SNP. The human genome contains millions of SNPs, with each measured or imputed SNP in an individual typically coded as 0, 1 or 2 (or respectively as aa, aA or AA) to indicate which parental variant was inherited by the individual at this SNP.

A genetic factor whose DNA variation across individuals correlates with the expression level of a gene is called an expression QTL or eQTL (Figure 2). Many multi-tissue studies collect genetic data on the same individuals from whom gene expression data are collected. This multi-tissue setting enables mapping of tissue eQTLs for each tissue with gene expression data, and raises natural questions on

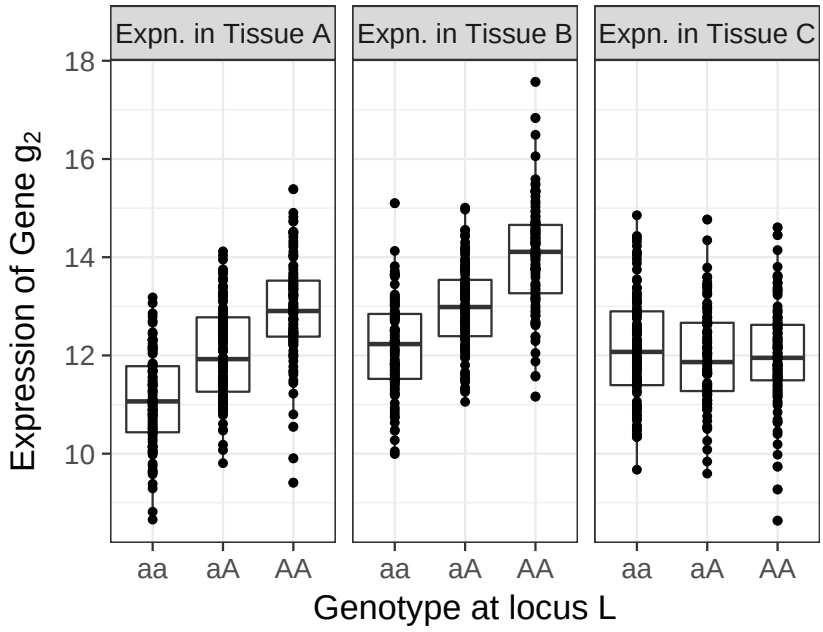
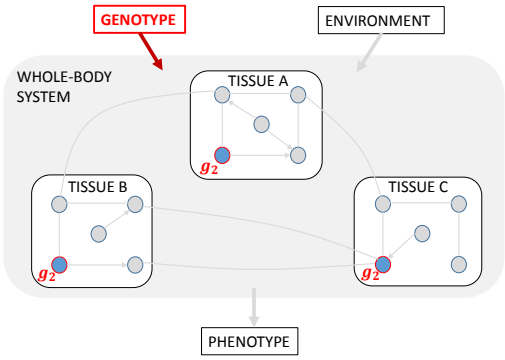


Figure 2: Multi-tissue eQTL example with hypothetical data. The genetic factor L is an eQTL for gene g_2 (i.e., correlated with the gene's expression across individuals in a population) in Tissue A and B but not C. In other words, this eQTL-gene association is shared between tissues A and B but not C. Note that L is a genomic locus/site with the variant inherited by each individual (each dot in the plot) at this site indicated as aa, aA or AA as explained in the text. Expn. stands for expression.

whether eQTLs for the same gene in different tissues are shared vs. distinct, and whether statistical power to detect shared eQTLs of a gene could be improved by joint analysis of tissues. This leads to the *Multi-tissue eQTL mapping* problem, where the goal is to find eQTLs for a gene within each tissue, with a special emphasis on borrowing information from other tissues that share eQTLs for this gene (Figure 2). The problem is challenging because the identity of tissues that share eQTLs with the query tissue is not known *a priori* and has to be inferred from the data simultaneously when detecting the eQTLs.

A Bayesian Model Averaging (BMA) approach was taken to address this problem in a study that models a tested eQTL as active or inactive in each of K input tissues for a total of 2^K possible configurations [14]. Conditioned on a configuration vector that identifies the subset of tissues in which the eQTL is active, the authors developed a hierarchical linear regression model where the outcome variable is multi-tissue expression of a gene and predictor variable is the tested eQTL (SNP). The heterogeneity of the genetic effect of the eQTL across all active tissues is modeled using a prior distribution on the regression coefficient of the SNP. Bayesian inference in this model resulted in increased power for detecting shared eQTLs, and also yielded direct estimates of the proportion of eQTLs shared by any number of different tissues (which is tricky to estimate from single-tissue eQTL analysis repeated on each tissue due to incomplete power considera-



Problem: Multi-tissue eQTL mapping

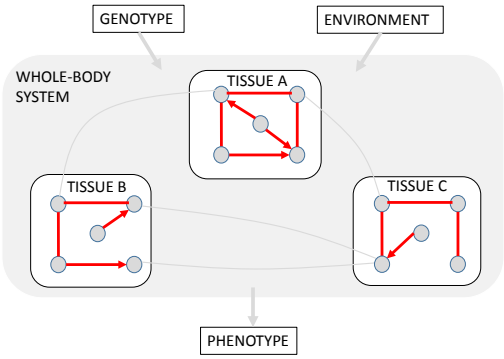
tions). When the number of tissues K is large, the authors present a model that scales better than 2^K under some restricting assumptions that an eQTL is active either in all tissues or only in one tissue. Developing multi-tissue eQTL models that use weaker assumptions to achieve tractable computational and statistical performance with large K is an open problem. Another important and actively researched open problem is to use multi-tissue eQTLs and their patterns of sharing and specificity across different tissues to predict the relevant causal tissue(s) underlying a QTL for disease risk [32].

Gene network inference within and across tissues

Network representation of gene-gene and other biomolecular interactions have proven useful in various contexts, including discovery of new disease genes and processes [40]. Interactions in a network could be physical/chemical in nature (e.g., experimentally-observed direct protein-protein or protein-DNA interactions) or statistical/functional (e.g., data-driven gene-gene correlation or causal relations, which reveal coordinated functional regulation of the interacting genes). We focus on the latter here. For functional network inference from single-tissue data, there are mature methods that address the high-dimensionality challenge (of learning millions of network parameters, one per pair of genes, from a small number of samples). For instance, Gaussian Graphical Model (GGM) methods can use a penalized likelihood approach to learn a sparse network with few parameters [9], and Bayesian network methods [3] can use prior information on physical/chemical or genetic interactions to compensate small sample sizes.

Extending network inference to a multi-tissue setting brings new opportunities and challenges, similar in spirit to the eQTL mapping problem seen above. *Intra-tissue network inference* refers to the problem of utilizing multi-tissue data to infer a functional network that comprises correlation/coexpression relations among biomolecules within a particular query tissue, and encourages borrowing of information from other tissues with shared signals (Figure 3). Power to detect coexpression relations could be improved by integrating data from multiple tissues, while accounting for differences in these relations across tissues.

Inspired by earlier work on multi-class network inference [9], a recent method addresses this problem by extending the GGM approach from single-tissue to multi-tissue data [33]. This method uses a tree-based hierarchy of relations between multiple tissues (derived from data or prior knowledge) to "transfer" network relations learnt in one tissue to other nearby tissues in the hierarchy. To scale to thousands of genes in 35 GTEx tissues, they augmented their multi-tissue GGM



Problem: Intra-tissue network inference

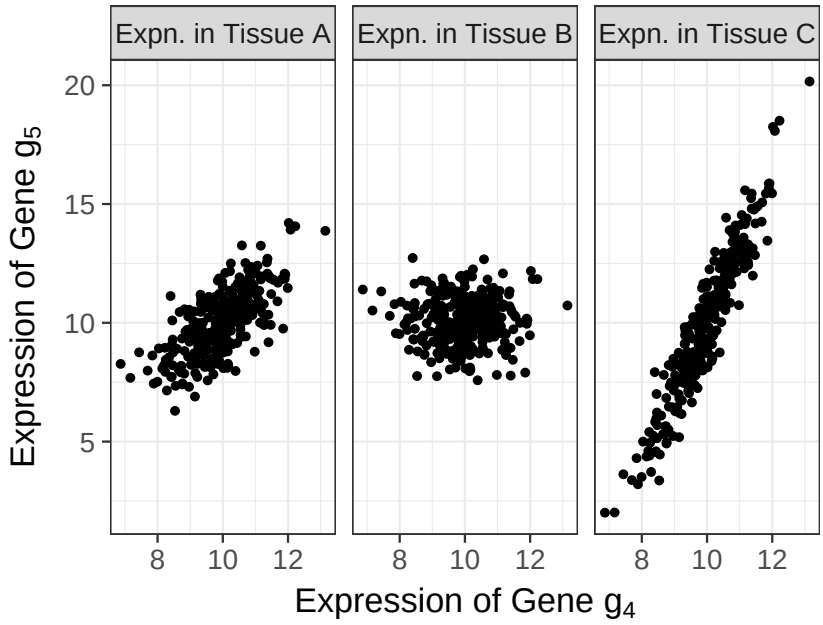


Figure 3: Intra-tissue network coexpression example with hypothetical data. The expression levels of genes g_4 and g_5 are correlated across individuals (dots) in Tissue A and C but not B. Among the tissues A and C sharing this association signal (i.e., significant gene-gene correlation), the exact strength of correlation/coexpression can vary. Expn. stands for expression.

algorithm with principled heuristics and identified steps that could be run in parallel; as a result, meaningful networks were inferred even for tissues with very few samples in GTEx data by pooling information from related tissues. In detail, their multi-tissue GGM method jointly learns all GGMs, one per tissue, by optimizing a single objective function with these additive components: single-tissue GGM objective functions, one per tissue; and L2 "transfer" penalties, one per pair of nearby nodes in the tissue hierarchy, to encourage similarity of inferred networks between related tissues. Note that the single-tissue GGM objective function is of the form $\frac{n}{2}(\log\{\det(\boldsymbol{S})\} - tr(\boldsymbol{\Sigma}\boldsymbol{S})) - \lambda\|\boldsymbol{S}\|_1$, with the first term capturing the log likelihood of a multivariate Gaussian distribution that the genes are assumed to follow (n denotes the sample size, $\boldsymbol{S}$ the inverse covariance matrix parameter and $\boldsymbol{\Sigma}$ the empirical covariance matrix), and the second term being a λ-weighted L1 penalty that enforces $\boldsymbol{S}$ and therefore the intra-tissue networks defined by the non-zero entries of $\boldsymbol{S}$ to be sparse (note that $\boldsymbol{S}_{jj'}$ entry captures the partial correlation between gene j and gene j' conditioned on all other genes) [19]. This objective function can be optimized by the so-called graphical lasso algorithm, which takes $O(p^4)$ running time to learn a dense network defined over p genes and $O(p^3)$ time for a reasonably sparse network (under the high-dimensional setting $p \gg n$ with density determined by the value of the penalty parameter λ) [15, 27].

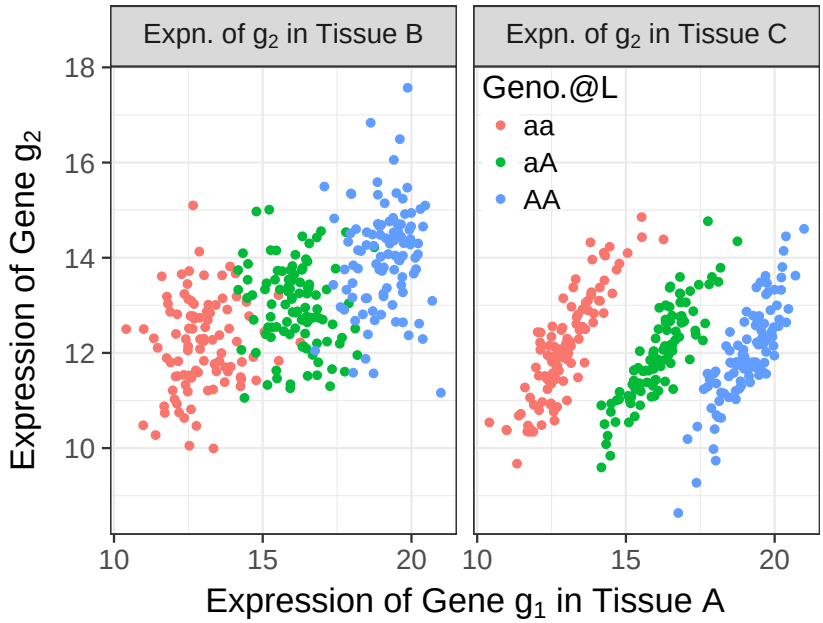
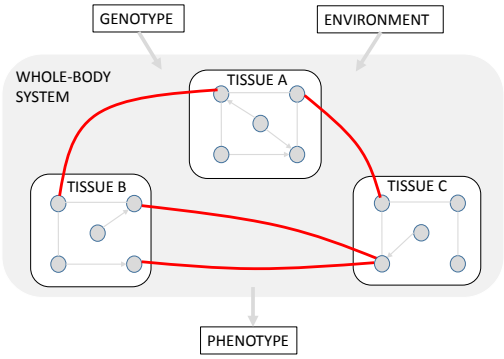


Figure 4: Inter-tissue network coexpression example with hypothetical data. Gene g_1 in Tissue A is correlated with gene g_2 in Tissue **B** solely due to a confounding genetic factor (shared eQTL L), as there is no evidence for correlation within each genotype group aa, aA or AA. But, the same g_1 in Tissue A is correlated with g_2 in Tissue **C** within each genotype group, suggesting that factors other than L (e.g., inter-tissue communication between tissues A and C specifically) may drive this correlation structure. Expn. stands for expression and Geno.@L for genotype at locus L, with both also shown in Figure 2.

A natural goal of any multi-tissue study is to discover biomolecular interactions that mediate tissue-tissue or inter-tissue communication, but confounding factors pose a challenge. Two genes in two different tissues could be correlated to each other due to confounding from common genetic factors (shared eQTLs driving the genes independently in their corresponding tissues; Figure 4, left panel) or common environmental factors (food intake of the individual, circadian rhythm cues from the brain, etc., that affect both tissues) [25]. Control of such confounding in order to infer genes in two different tissues that are co-expressed (Figure 4, right panel) due to direct communication between the two tissues via exchange of hormones or other biomolecular signals constitute the *inter-tissue network inference* problem.



Problem: Inter-tissue network inference

One approach to this problem is to adjust gene expression data to remove contributions from confounding genetic factors, and use the adjusted data to derive inter-tissue coexpression relations employing standard procedures [25]. One could for instance include all shared eQTLs driving gene j in tissue k (whose expression is denoted $\boldsymbol{X}_{jk}$) and j' in another tissue k' (whose expression is $\boldsymbol{X}_{j'k'}$) as covariates in a separate linear regression of $\boldsymbol{X}_{jk}$ or $\boldsymbol{X}_{j'k'}$, and take the fit residuals as the adjusted expression data. But mapping shared eQTLs for every gene pair is a non-trivial problem on its own. An alternate approach adjusts each gene's expression for all genetic factors in the genome by modeling them in a linear mixed model (LMM) as a random effect

term, estimating the term's covariance matrix using the observed genetic similarity between every pair of individuals, and performing these steps using computationally efficient LMM methods from GWAS literature (which typically take $O(n^3 + mn^2 + kpn^2)$ time with n individuals, m genetic factors and p genes measured in k tissues [38]) in order to eventually build inter-tissue networks [25]. Open problems on addressing important LMM pitfalls [38] in an efficient fashion for all gene pairs remain. One pitfall concerns the LMM assumption that all genetic factors have small additive effects on expression - so we may decide to model a large-effect eQTL as a fixed instead of random effect term in the LMM, and this decision may be gene or gene-pair specific.

The network inference problems discussed thus far refer to gene-gene correlation/coexpression networks, and it could prove quite useful to extend them to infer causal relations from large observational datasets (rather than data from small-scale perturbation experiments traditionally used for causal discovery). Researchers have developed tests of causality between two correlated genes within a single tissue based on utilizing natural genetic variation across individuals as "causal anchors" or "instrumental variables". A genetic variant driving gene expression (eQTL) could serve as an anchor or instrument, since the variant is determined at conception after random assortment of the parental variants (the so-called Mendelian randomization concept that mimics randomized controlled trial setup) and therefore not affected by any environmental or phenotypic confounding factors (with some exceptions like in cancer) [10]. Multi-tissue causality inference is an open field, as there are no current methods to the best of our knowledge that can strengthen inference of gene-gene causal relations within a tissue by borrowing shared signals from other tissues, or build inter-tissue causal networks by viewing eQTLs as instruments for causal discovery and not just as confounders of inter-tissue co-expression relations as seen above. Building such intra/inter-tissue causal networks could help resolve questions like whether a disease affects genes in a tissue-independent fashion vs. inter-tissue inter-linked fashion, and thereby reveal key causal tissues for disease intervention.

3.3 Computational/algorithmic analysis of multi-tissue models

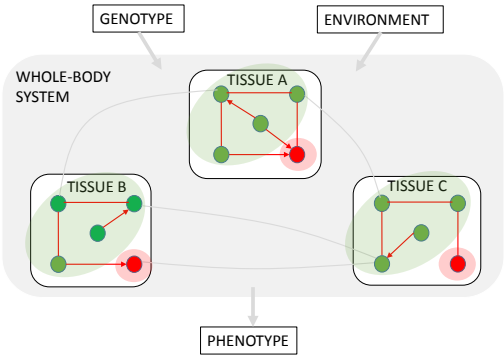
Inferred models in genomics like the coexpression networks seen above often contain hundreds to thousands of variables (genes). Extracting biological insights from a large genomic model, unlike a statistical model over a few variables, requires further computational analyses that organize, visualize and dissect the model structure. Such analyses have been developed to study biological networks mostly in a

single-tissue/celltype setting. Popular network analysis examples in this setting, which are based on graph-theoretic concepts and reviewed in detail elsewhere [28], include: partitioning all genes in a network into smaller modules of well-connected genes corresponding to tightly-regulated biological processes via graph clustering, finding key regulator genes in a network using graph-theoretic measures like betweenness centrality, searching for "active" subnetworks that optimally connect predefined disease genes via Steiner-like graph problems, and mining for overrepresented network substructures like feedback motifs to uncover organizing principles of networks using triangle-counting or similar graph algorithms. There are introductory books on algorithms for such single-network analysis problems arising not only in biology but also in other sciences [30].

Network analysis problems take a new role in the multi-tissue context, as they call for integration not only of multiple networks but also of intra-tissue with inter-tissue networks to find for instance key genes mediating tissue-tissue crosstalk. Multi-tissue network analysis can benefit from emerging nascent research in the broader field of multi-layer networks, where many layers each containing an intra-layer network are coupled to each other via inter-layer connections [23, 4]. Some early works on multi-tissue network analysis exist [12, 11], however this research area is not as mature as its single-tissue counterpart and so ripe for new developments.

To give a flavor of the challenges in integrating even just the intra-tissue networks, we consider extending graph clustering from single to multiple graphs. *Simultaneous clustering* is the problem of jointly clustering multiple graphs - each encoding a distinct set of edges (gene-gene relations within a tissue in our case) over the same set of nodes (genes) - where the aim is to identify subsets of nodes that form well-connected clusters across the collection of graphs. This problem applied on the intra-tissue networks would reveal gene clusters that are robustly coexpresssed within multiple tissues. One may wonder if this problem could be trivially solved by single-graph clustering of a shared network that is derived by a simple edge-by-edge overlap of all intra-tissue networks. But incomplete power of detecting coexpression relations would preclude this approach by making the intra-tissue networks non-overlapping at the edge level but robustly overlapping at higher levels of organization like clusters [33].

The biological criteria used to assess the well-connectedness or quality of a cluster could determine the computational complexity of the simultaneous clustering problem [29] and a related problem from literature called network alignment [34]; hence approaches ranging from provably-efficient algorithms to principled heuristics have been developed for these problems in a different context of multi-environment

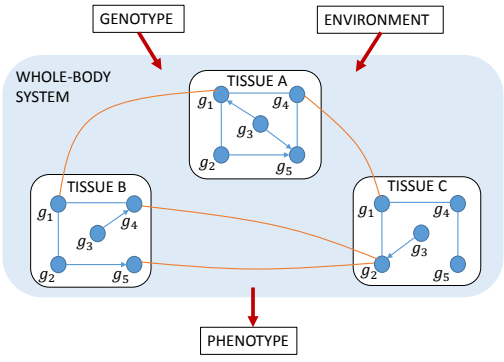


Problem: Simultaneous clustering

or multi-species networks [34]. One approach called JointCluster is a provably-efficient approximation algorithm when cluster quality is measured by conductance [29]. A cluster is well-connected with conductance at least α if every two-way partition or cut of the cluster has conductance at least α, with conductance defined as the ratio of edges crossing the cut to the edges incident at the smaller side of the cut and α being an user-specified cutoff. JointCluster offers an approximation solution to the optimal partitioning of the nodes in the graph into clusters that satisfies two criteria: conductance of each cluster is at least α and the fraction ϵ of edges lost between the clusters is minimized. The algorithm extends spectral techniques from an earlier single-graph clustering study [21] to the case of multiple graphs to obtain theoretical guarantees on the detected clustering's quality and running time. These guarantees, specifically poly-logarithmic approximation guarantees on the two-criteria quality measure involving α and ϵ and worst-case running time bound of $O(kp^3)$ (suppressing factors of $\log p$) for k graphs each defined over p genes, coupled with an effective scaling heuristic and the flexibility to handle multiple heterogeneous networks, would make JointCluster applicable to integration of intra-tissue networks. Still, many open problems exist in simultaneous clustering of multi-tissue networks, such as integrative clustering of both intra- and inter-tissue networks and explicitly modeling the differences besides similarities between tissue networks via a development-based hierarchy of tissues. Combined information in intra- and inter-tissue networks may also prompt fresh rethinking of network analysis problems beyond clustering such as key regulator and active subnetwork analyses.

3.4 Towards whole-body models

An ultimate goal that all multi-tissue problems seen so far converge upon is to build a whole-body or whole-system model that is both predictive of a complex phenotype like disease and mechanistic or non-black-box-like in terms of revealing its underlying biomolecular network structure. To manage the complexity of the *whole-body or whole-system model building* problem, researchers have taken a component/object-oriented modeling approach, wherein they first build model components like within-tissue models for all biomolecules and tissues relevant to the studied phenotype, and then integrate these components into a single model using additional information like inter-tissue communication. To clarify, we provide a concrete example of a system-wide multi-tissue model of human metabolism based on ODEs (Ordinary Differential Equations, with one ODE per relevant metabolite/biomolecule to model its rate of production/loss).



Problem: Whole-body model building

Many whole-system modeling efforts have focused on metabolism, since genome-wide ODE models of single-celltype or single-tissue metabolism are quite mature and popular. For instance, one can: i) reconstruct the network of all metabolic reactions in an organism from well-studied metabolic networks in microbial organisms by exploiting the high evolutionary conservation of metabolic reactions and their enzymes encoded by genes, ii) write down the ODEs readily from the reconstructed metabolic network by applying law of mass action, and iii) circumvent the infeasible estimation of thousands of kinetic ODE parameters using Flux Balance Analysis (FBA), which derives the rate or flux of all reactions under the assumption that cellular metabolism is in steady-state and optimizes a biological objective like growth rate or a tissue endpoint [5]. FBA can be implemented using linear programming methods, since the steady-state and other constraints and the biological objective function can be written in terms of linear combinations of the reaction fluxes in most cases. FBA can also be viewed as a generalization of the maximum flow problem in graphs, since we are essentially optimizing the flow of metabolites through the metabolic network subject to conservation of flow (steady-state) and other constraints.

Researchers have built a large-scale model of steady-state human metabolism involving liver, adipose, and muscle tissues by tailoring a human metabolic network to each of these tissues using tissue-specificity of proteins, integrating them in a non-trivial fashion that goes beyond a trivial union of the three tissue-specific models by adding a blood compartment with buffers, and validating the integrated model via FBA of known human metabolic cycles that utilize these tissues [5]. Analyzing the resulting system-wide model using multi-tissue expression data from obese vs. diabetic obese individuals revealed differentially active metabolic reactions that could not be found from expression data alone. This whole-system undertaking [5] and similar work on plant systems (described in an easily comprehensible paper [16]) suggest several open problems such as ones related to integrated models of metabolism and gene/protein expression [26], personalized models of metabolism [6] tailored to an individual's multi-tissue expression data, and analyzing dynamical instead of steady state behaviors of whole-body systems [31].

The field of whole-body modeling is much broader than the system-wide models of metabolism and gene expression discussed above, as life processes can happen at multiple space/time scales from molecular to cellular, tissue-level and organismal. A multi-scale model of heart function for instance may involve modeling not only metabolism/expression within heart cells, but also biophysics of heart rhythm, dynamics of blood flow, and interactions with lung or other tissues [41] - building

it would be a huge collaborative undertaking involving clinicians, experimentalists and analysts with many open challenges, however the incentive would also be huge in delivering a model that can predict the effect of any drug or molecular perturbation on heart function and failure [41]. A more comprehensive discussion of multi-scale modeling approaches can be found in special journal issues on this topic [22, 7].

4 Conclusion

We reviewed computational and statistical problems pertaining to analysis of multi-tissue genomic data, which comprises genome-wide data on biomolecular activities collected from multiple tissues of several individuals. These problems constitute an active area of research as large multi-tissue studies with this new type of data are recently emerging, and progress in methods solving these problems are ushering in an era of whole-system predictive models - the holy grail of systems biology.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, NIAID.

References

[1] Genevera I. Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790, June 2010.

[2] Genevera I. Allen and Robert Tibshirani. Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):721–743, September 2012.

[3] Peter J. Bickel, James B. Brown, Haiyan Huang, and Qunhua Li. An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 367(1906):4313–4337, November 2009.

[4] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, November 2014.

[5] Aarash Bordbar, Adam M. Feist, Renata Usaite-Black, Joseph Woodcock, Bernhard O. Palsson, and Iman Famili. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Systems Biology*, 5:180, October 2011.

[6] Aarash Bordbar, Douglas McCloskey, Daniel C. Zielinski, Nikolaus Sonnenschein, Neema Jamshidi, and Bernhard O. Palsson. Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics. *Cell Systems*, 1(4):283–292, October 2015.

[7] Jean-Louis Coatrieux, Alejandro F. Frangi, Grace C. Y. Peng, David Z. D'Argenio, Vasilis Z. Marmarelis, and Anushka Michailova. Editorial: TBME Letters special issue on multiscale modeling and analysis in computational biology and medicine–part-2. *IEEE Transactions on Bio-Medical Engineering*, 58(12):3434–3439, December 2011.

[8] National Research Council, Division on Engineering and Physical Sciences, Board on Mathematical Sciences and Their Applications, and Committee on Mathematical Sciences Research for DOE's Computational Biology. *Mathematics and 21st Century Biology*. National Academies Press, Washington, DC, July 2005.

[9] Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 76(2):373–397, March 2014.

[10] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, August 2007.

[11] Kieu Trinh Do, Gabi Kastenmüller, Dennis O. Mook-Kanamori, Noha A. Yousri, Fabian J. Theis, Karsten Suhre, and Jan Krumsiek. Network-based approach for analyzing intra- and interfluid metabolite associations in human blood, urine, and saliva. *Journal of Proteome Research*, 14(2):1183–1194, February 2015.

[12] Radu Dobrin, Jun Zhu, Cliona Molony, Carmen Argman, Mark L. Parrish, Sonia Carlson, Mark F. Allan, Daniel Pomp, and Eric E. Schadt. Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biology*, 10(5):R55, 2009.

[13] Ilia A. Droujinine and Norbert Perrimon. Defining the interorgan communication network: systemic coordination of organismal cellular processes under homeostasis and localized stress. *Frontiers in Cellular and Infection Microbiology*, 3:82, 2013.

[14] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS genetics*, 9(5):e1003486, May 2013.

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.

[16] Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Pedro A. Saa, and Lars K. Nielsen. A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Frontiers in Plant Science*, 6:4, 2015.

[17] Elin Grundberg, Kerrin S. Small, Ã. . . sa K. Hedman, Alexandra C. Nica, Alfonso Buil, Sarah Keildson, Jordana T. Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, James Nisbett, Magdalena Sekowska, Alicja Wilk, So-Youn Shin, Daniel Glass, Mary Travers, Josine L. Min, Sue Ring, Karen Ho, Gudmar Thorleifsson, Augustine Kong, Unnur Thorsteindottir, Chrysanthi Ainali, Antigone S. Dimas, Neelam Hassanali, Catherine Ingle, David Knowles, Maria Krestyaninova, Christopher E. Lowe, Paola Di Meglio, Stephen B. Montgomery, Leopold Parts, Simon Potter, Gabriela Surdulescu, Loukia Tsaprouni, Sophia Tsoka, Veronique Bataille, Richard Durbin, Frank O. Nestle, Stephen O'Rahilly, Nicole Soranzo, Cecilia M. Lindgren, Krina T. Zondervan, Kourosh R. Ahmadi, Eric E. Schadt, Kari Stefansson, George Davey Smith, Mark I. McCarthy, Panos Deloukas, Emmanouil T. Dermitzakis, Tim D. Spector, and Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, October 2012.

[18] GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[19] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, volume (Chapter 17 on Graphical Models). Springer, New York, NY, 2009.

[20] Guo-Jen Huang, Sagiv Shifman, William Valdar, Martina Johannesson, Binnaz Yalcin, Martin S. Taylor, Jennifer M. Taylor, Richard Mott, and Jonathan Flint. High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Research*, 19(6):1133–1140, June 2009.

[21] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, May 2004.

[22] Michael R. King and Scott L. Diamond. Multiscale systems biology: a special issue devoted to understanding biology and medicine across multiple scales. *Annals of Biomedical Engineering*, 40(11):2293–2294, November 2012.

[23] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, September 2014.

[24] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[25] Quan Long, Carmen Argmann, Sander M. Houten, Tao Huang, Siwu Peng, Yong Zhao, Zhidong Tu, GTEx Consortium, and Jun Zhu. Inter-tissue coexpression network analysis reveals DPP4 as an important gene in heart to blood communication. *Genome Medicine*, 8(1):15, February 2016.

[26] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*, 10(4):e1003580, April 2014.

[27] Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13:781–794, March 2012.

[28] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews. Genetics*, 14(10):719–732, October 2013.

[29] Manikandan Narayanan, Adrian Vetta, Eric E. Schadt, and Jun Zhu. Simultaneous Clustering of Multiple Gene Expression and Physical Interaction Datasets. *PLoS Computational Biology*, 6(4), April 2010.

[30] Mark Newman. *Networks : an introduction*. Oxford University Press, Oxford New York, 2010.

[31] Elin Nyman, Cecilia Brännmark, Robert Palmér, Jan Brugård, Fredrik H. Nyström, Peter Strålfors, and Gunnar Cedersund. A hierarchical whole-body modeling approach elucidates the link between in Vitro insulin signaling and in Vivo glucose homeostasis. *The Journal of Biological Chemistry*, 286(29):26028–26041, July 2011.

[32] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos Panousis, Alexandra C Nica, GTEx Consortium, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *bioRxiv*, 074682, 2016.

[33] Emma Pierson, GTEx Consortium, Daphne Koller, Alexis Battle, Sara Mostafavi, Kristin G. Ardlie, Gad Getz, Fred A. Wright, Manolis Kellis, Simona Volpi, and Emmanouil T. Dermitzakis. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS computational biology*, 11(5):e1004220, May 2015.

[34] Roded Sharan and Trey Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24(4):427–433, April 2006.

[35] Anestis Touloumis, John C. Marioni, and Simon Tavaré. HDTD: analyzing multi-tissue gene expression data. *Bioinformatics (Oxford, England)*, 32(14):2193–2195, July 2016.

[36] Jiebiao Wang, Eric R. Gamazon, Brandon L. Pierce, Barbara E. Stranger, Hae Kyung Im, Robert D. Gibbons, Nancy J. Cox, Dan L. Nicolae, and Lin S. Chen. Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx. *American Journal of Human Genetics*, 98(4):697–708, April 2016.

[37] Yibo Wu, Evan G. Williams, Sébastien Dubuis, Adrienne Mottis, Virginija Jovaisaite, Sander M. Houten, Carmen A. Argmann, Pouya Faridi, Witold Wolski, Zoltán Kutalik, Nicola Zamboni, Johan Auwerx, and Ruedi Aebersold. Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell*, 158(6):1415–1430, September 2014.

[38] Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed model association methods. *Nature Genetics*, 46(2):100–106, February 2014.

[39] Xia Yang, Eric E. Schadt, Susanna Wang, Hui Wang, Arthur P. Arnold, Leslie Ingram-Drake, Thomas A. Drake, and Aldons J. Lusis. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Research*, 16(8):995–1004, August 2006.

[40] Bin Zhang, Linh Tran, Valur Emilsson, and Jun Zhu. Characterization of Genetic Networks Associated with Alzheimer's Disease. *Methods in Molecular Biology (Clifton, N.J.)*, 1303:459–477, 2016.

[41] Yanhang Zhang, Victor H. Barocas, Scott A. Berceli, Colleen E. Clancy, David M. Eckmann, Marc Garbey, Ghassan S. Kassab, Donna R. Lochner, Andrew D. McCulloch, Roger Tran-Son-Tay, and Natalia A. Trayanova. Multi-scale Modeling of the Cardiovascular System: Disease Development, Progression, and Clinical Intervention. *Annals of Biomedical Engineering*, 44(9):2642–2660, September 2016.