Bioinformatics, 36(6), 2020, 1807–1813
doi: 10.1093/bioinformatics/btz805
Advance Access Publication Date: 5 November 2019
Original Paper

OXFORD

Systems biology

Model-based clustering of multi-tissue gene expression data

Pau Erola [1,2,*], Johan L. M. Björkegren[3,4] and Tom Michoel [1,5,*]

[1]Division of Genetics and Genomics, The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, UK, [2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK, [3]Department of Genetics and Genomic Sciences, Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA, [4]Integrated Cardio Metabolic Centre (ICMC), Karolinska Institutet, Huddinge 141 57, Sweden and [5]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen N-5020, Norway

*To whom correspondence may be addressed.

Associate Editor: Inanc Birol

Received on November 17, 2018; revised on September 5, 2019; editorial decision on October 24, 2019; accepted on October 31, 2019

Abstract

Motivation: Recently, it has become feasible to generate large-scale, multi-tissue gene expression data, where expression profiles are obtained from multiple tissues or organs sampled from dozens to hundreds of individuals. When traditional clustering methods are applied to this type of data, important information is lost, because they either require all tissues to be analyzed independently, ignoring dependencies and similarities between tissues, or to merge tissues in a single, monolithic dataset, ignoring individual characteristics of tissues.

Results: We developed a Bayesian model-based multi-tissue clustering algorithm, revamp, which can incorporate prior information on physiological tissue similarity, and which results in a set of clusters, each consisting of a core set of genes conserved across tissues as well as differential sets of genes specific to one or more subsets of tissues. Using data from seven vascular and metabolic tissues from over 100 individuals in the STockholm Atherosclerosis Gene Expression (STAGE) study, we demonstrate that multi-tissue clusters inferred by revamp are more enriched for tissue-dependent protein-protein interactions compared to alternative approaches. We further demonstrate that revamp results in easily interpretable multi-tissue gene expression associations to key coronary artery disease processes and clinical phenotypes in the STAGE individuals.

Availability and implementation: Revamp is implemented in the Lemon-Tree software, available at https://github.com/eb00/lemon-tree

Contact: pau.erola@bristol.ac.uk or tom.michoel@uib.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Clustering gene expression data into groups of genes sharing the same expression profile across multiple conditions remains one of the most important methods for reducing the dimensionality and complexity of large-scale microarray and RNA-sequencing datasets (Andreopoulos *et al.*, 2008; D'haeseleer, 2005; van Dam *et al.*, 2017). Coexpression clusters group functionally related genes together, and reveal how diverse biological processes and pathways respond to the underlying perturbation of the biological system of interest. Traditionally, clustering is performed by collecting data from multiple experimental treatments (Eisen *et al.*, 1998), time points (Spellman *et al.*, 1998), cell or tissue types (Freeman *et al.*, 2007), or genetically diverse individuals (Ghazalpour *et al.*, 2006) in a single data matrix from which meaningful patterns are extracted using any of a whole range of statistical and algorithmic approaches. More recently, it has become feasible to probe systems along two or more of these dimensions simultaneously. In particular, we are interested in multi-tissue data, where gene expression profiles are obtained from multiple tissues or organs sampled from dozens to hundreds of individuals (Foroughi Asl *et al.*, 2015; Franzén *et al.*, 2016; Fu *et al.*, 2012; Greenawalt *et al.*, 2011; Grundberg *et al.*, 2012; GTEx Consortium, 2017; Hägg *et al.*, 2009; Keller *et al.*, 2008). These data can potentially reveal the similarity and differences in (co)expression between tissues as well as the tissue-specific variation in (co)expression across individuals.

However, when traditional clustering methods are applied to this type of data, important information is lost. For instance, if each tissue-specific sub-dataset is clustered independently, the resulting sets of clusters will rarely align, and to compare clusters across tissues, one will be faced with the general problem of determining cluster preservation statistics (Langfelder *et al.*, 2011). If instead the

data are concatenated 'horizontally' in a single gene-by-sample matrix, a common set of clusters will be found, but these will be biased heavily towards house-keeping processes that are coexpressed in all tissues. A potentially more promising approach is to concatenate data 'vertically' in a tissue-gene-by-individual matrix, where the entities being clustered are 'tissue-genes', the tissue-specific expression profiles of genes (Dobrin *et al.*, 2009; Talukdar *et al.*, 2016). However, in studies with a large number of tissues, the number of individuals with available data in *all* tissues is typically very small, i.e. a large number of samples will have to be discarded to obtain a tissue-gene-by-individual matrix without missing data.

Dedicated clustering algorithms for multi-tissue expression data are scarce and mostly based on using the higher-order generalized singular value decomposition or related matrix decomposition techniques to identify common and differential clusters across multiple conditions (Li *et al.*, 2011; Ponnapalli *et al.*, 2011; Xiao *et al.*, 2014). However, these methods either require that all tissues have the same number of one-to-one matching samples (Ponnapalli *et al.*, 2011), or that tissue-specific coexpression networks are reconstructed for each tissue separately as a preliminary step (Li *et al.*, 2011; Xiao *et al.*, 2014). Bayesian model-based clustering methods, which model the data as a whole using mixtures of probability distributions (Fraley and Raftery, 2002; Ickstadt *et al.*, 2017; Si *et al.*, 2014), are an attractive alternative approach for clustering multi-tissue data, because they would allow, at least in principle, to account for different noise levels and sample sizes in different tissues and to incorporate prior information on the relative similarity between certain tissues based on their known physiological function.

Here we present a novel statistical framework and inference algorithm for model-based clustering of multi-tissue gene expression data, which can incorporate prior information on tissue similarity, and which results in a set of clusters, each consisting of a core set of genes conserved across tissues as well as differential sets of genes specific to one or more subsets of tissues.

2 Materials and methods

2.1 Approach

In model-based clustering, a partitioning of genes into non-overlapping clusters parametrizes a probabilistic model from which the expression data is assumed to have been generated, typically in the form of a mixture distribution where each cluster corresponds to one mixture component. Using Bayes' theorem, this can be recast as a probability distribution on the set of all possible clusterings parameterized by the expression data, from which maximum-likelihood solutions can be obtained using expectation-maximization or Gibbs sampling.

Our approach to clustering multi-tissue data combines ideas from existing ordinary ('single-tissue') and multi-species model-based clustering methods. We use the generative model of Qin (2006) and Joshi *et al.* (2008) to obtain the posterior probability for a (single-tissue) clustering given a (single-tissue) dataset. From Roy *et al.* (2013) we use the idea that a multi-tissue clustering consists of a set of *linked* clusters, where cluster k in one tissue corresponds to cluster k in any other tissue, and each cluster k contains a *core* set of genes, belonging to cluster k in *all* tissues, and a *differential* set of tissue-specific genes, belonging to cluster k in one or more, but not all, tissues. Like Roy *et al.* (2013), we assume that the data from one tissue can influence the clustering in another tissue, albeit via a simpler mechanism as we do not aim to reconstruct any phylogenetic histories among tissues. In brief, we assume that the posterior probability distribution of clusterings in tissue t is given by its ordinary single-tissue distribution given the expression data for tissue t, multiplied by a tempered distribution for observing that same clustering given the expression data for all other tissues $t' \neq t$. The degree of tempering determines the degree of influence of one tissue on another, and can be used to model known prior relationships between tissues. For instance, we expect *a priori* that coexpression clusters will be more similar between vascular tissues, than between vascular and metabolic tissues.

2.2 Statistical model for single-tissue clustering

Our method is based on previous single-tissue, model-based clustering algorithms (Joshi *et al.*, 2008; Qin, 2006). In brief, for an expression data matrix $\mathbf{X} \in \mathbb{R}^{G\times N}$ for G genes and N samples, a clustering $\mathcal{C}$ is defined as a partition of the genes into K non-overlapping sets C_k. We assume that the data points for the genes in each cluster and each sample are normally distributed around an unknown mean and unknown variance/precision. Given a clustering $\mathcal{C}$ and a set of means and precisions (μ_{kn}, τ_{kn}) for each cluster k and sample n, we obtain a distribution on expression data matrices as

$$p(\mathbf{X}|\mathcal{C}, \{\mu_{kn}, \tau_{kn}\}) = \prod_{k=1}^{K}\prod_{n=1}^{N}\prod_{g\in C_k} p(x_{gn}|\mu_{kn}, \tau_{kn}).$$

Assuming a uniform prior on the clusterings $\mathcal{C}$ and independent normal-gamma priors on the normal distribution parameters, we can use Bayes' rule to find the marginal posterior probability of observing a clustering $\mathcal{C}$ given data $\mathbf{X}$, upto a normalization constant:

$$P(\mathcal{C}|\mathbf{X}) \propto \prod_{k=1}^{K}\prod_{n=1}^{N}\int\int p(\mu,\tau)\prod_{g\in C_k} p(x_{gn}|\mu,\tau)\,d\mu d\tau. \tag{1}$$

Note that we use a capital 'P' to indicate that this is a *discrete* distribution. $p(\mu,\tau) = p(\mu|\tau)p(\tau)$ is the normal-gamma prior, with

$$p(\mu|\tau) = \left(\frac{\lambda_0\tau}{2\pi}\right)^{1/2} e^{-\frac{\lambda_0\tau}{2}(\mu-\mu_0)^2}, \quad p(\tau) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\tau^{\alpha_0-1}e^{-\beta_0\tau},$$

$\alpha_0, \beta_0, \lambda_0 > 0$ and $-\infty < \mu_0 < \infty$ being the parameters of the normal-gamma prior distribution. We use the values $\alpha_0 = \beta_0 = \lambda_0 = 0.1$ and $\mu_0 = 0.0$, resulting in a non-informative prior. The double integral in (1) can be solved exactly in terms of the sufficient statistics $T_{kl}^{(\alpha)} = \sum_{i\in\mathcal{C}_k}\sum_{n=1}^{N} x_{in}^{\alpha}$ $(\alpha = 0, 1, 2)$ for each cluster, see Joshi *et al.* (2008) for details.

For computational purposes, the decomposition of Eq. (1) into a product of independent factors, one for each cluster and sample, is important. We write the log-likelihood or Bayesian score accordingly as:

$$\mathcal{S}(\mathcal{C}) = \log P(\mathcal{C}|\mathbf{X}) = \sum_{k=1}^{K}\sum_{n=1}^{N}\mathcal{S}_{kn}. \tag{2}$$

2.3 Statistical model for multi-tissue clustering

Next, we assume that expression data $\mathbf{X} = [\mathbf{X}_1 \in \mathbb{R}^{G\times N_1}, \ldots, \mathbf{X}_T \in \mathbb{R}^{G\times N_T}]$ is available for G genes in T tissues, with N_t samples in each tissue $t \in \{1, \ldots, T\}$. We define a multi-tissue clustering as a collection of single-tissue clusterings $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_T\}$, and assume that the probability of observing $\mathcal{C}$ given data $\mathbf{X}$ is given by

$$\begin{aligned} P(\mathcal{C}|\mathbf{X}) &= P(\mathcal{C}_1, \ldots, \mathcal{C}_T|\mathbf{X}_1, \ldots, \mathbf{X}_T) \\ &= \frac{1}{Z}\prod_{t=1}^{T}\left\{ aP(\mathcal{C}_t|\mathbf{X}_t)\prod_{t'\neq t} P(\mathcal{C}_t|\mathbf{X}_{t'})^{\lambda_{t,t'}}\right\}, \end{aligned} \tag{3}$$

where Z is a normalization constant which we henceforth will ignore, each factor $P(\mathcal{C}_t|\mathbf{X}'_t)$ is a single-tissue posterior probability distribution defined in Eq. (1), and $\lambda_{t,t'} \in [0, 1]$ is a set of hyperparameters that define the prior tissue similarities; for notational convenience we define $\lambda_{t,t} = 1$.

Note that $P(\mathcal{C}_t|\mathbf{X}_{t'})$ is a discrete distribution measuring how well clustering $\mathcal{C}_t$ is supported by data $\mathbf{X}_{t'}$. Raising a discrete distribution to a power less than 1 has the effect of making the distribution more uniform. Hence in Eq. (3), we are asking that clustering $\mathcal{C}_t$ is supported predominantly by data $\mathbf{X}_t$ from its own tissue, but also, albeit

to a lesser extent depending on the values of $\lambda_{t,t'}$, by data from the other tissues.

Optimizing Eq. (3) across all multi-tissue clusterings is challenging. A considerable simplification is obtained if we constrain the problem to multi-tissue clusterings with the *same* number of clusters K in each tissue. Denoting by $\mathcal{I}_t$ the set of samples/individuals in tissue t and by $N = \sum_t N_t$ the total number of samples, the decomposition in Eq. (2) allows to write:

$$\begin{aligned}\log P(\mathcal{C}|\mathbf{X}) &= \sum_{t=1}^{T}\sum_{t'=1}^{T}\lambda_{t,t'}\log P(\mathcal{C}_t|\mathbf{X}_{t'}) \\ &= \sum_{t=1}^{T}\sum_{t'=1}^{T}\lambda_{t,t'}\sum_{k=1}^{K}\sum_{n\in\mathcal{I}_{t'}}\mathcal{S}_{kn}^{(t)} \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_n^{(t)}\mathcal{S}_{kn}^{(t)},\end{aligned} \tag{4}$$

where we used $\lambda_{t,t} = 1$, defined $\gamma_n^{(t)} \equiv \lambda_{t,t(n)}$, with $t(n)$ the tissue to which sample n belongs, and wrote $\mathcal{S}_{kn}^{(t)}$ to denote the Bayesian score of clustering $\mathcal{C}_t$ with respect to sample n.

Two extremal choices for the hyper-parameters are of interest. If $\lambda_{t,t'} = 1$ for all t, t', then the Bayesian score

$$\mathcal{S}^{(t)} = \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_n^{(t)}\mathcal{S}_{kn}^{(t)} \tag{5}$$

is the same for each tissue t and identical to Eq. (2) for the concatenated data matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_T]$. Hence this is equivalent to clustering the entire dataset as if it came from a single-tissue ('horizontal' data concatenation). If $\lambda_{t,t'} = 0$ for $t' \neq t$, then Eq. (3) decomposes as a product of independent single-tissue factors. This is equivalent to clustering each tissue sub-dataset independently.

2.4 Optimization algorithm

To find a local maximum of the Bayesian score in Eq. (4), the following heuristic, greedy optimization algorithm was used:

1. **Data standardization:** Using appropriately normalized gene expression data, each gene is standardized to have mean zero and standard deviation one on the concatenated data $\mathbf{X}$.
2. **Determine the number of clusters:** K-means clustering is run on the concatenated data with the number of clusters ranging from 2 to 100. The optimal number K is selected by visual inspection of an elbow plot.
3. **Initialize multi-tissue clustering:** Starting from the k-means clustering output at the selected number of clusters, genes are reassigned until a local optimum is reached for the single-tissue score Eq. (2) on the concatenated data $\mathbf{X}$. All $\mathcal{C}_t$ are initialized by this clustering.
4. **Optimize multi-tissue clustering:** For each tissue t, optimize $\mathcal{C}_t$ by finding a local maximum for the Bayesian score Eq. (5) using single-gene reassignments; only gene reassignments improving the score by a minimum threshold ϵ are considered.

Note that even in the case $\lambda_{t,t'} = 0$ for $t' \neq t$, which removes all tissue dependencies in the Bayesian score (4), this algorithm still results in a multi-tissue clustering with linked clusters, due to each tissue being initialized by the same clustering and converging to a local optimum.

2.5 Implementation

The statistical model and optimization algorithm have been implemented in Java, as an extension of the 'task' revamp in the Lemon-Tree software (Bonnet *et al.*, 2015; Erola *et al.*, 2019), available at https://github.com/eb00/lemon-tree.

2.6 The Stockholm Atherosclerosis Gene Expression dataset

In the STockholm Atherosclerosis Gene Expression (STAGE) study, 612 tissue samples from 121 individuals were obtained during coronary artery bypass grafting surgery from the atherosclerotic arterial wall (AAW, $n = 73$), internal mammary artery (IMA, $n = 88$), liver ($n = 87$), skeletal muscle (SM, $n = 89$), subcutaneous fat (SF, $n = 72$) and visceral fat (VF, $n = 98$) of well-characterized CAD patients; fasting whole blood (WB) was obtained for isolation of DNA ($n = 109$) and RNA ($n = 105$) and biochemical analyses. Gene expression profiles from RNA samples of different tissues were jointly normalized to enable comparison across tissues (Foroughi Asl *et al.*, 2015; Hägg *et al.*, 2009; Talukdar *et al.*, 2016). 4956 genes with variance greater than 1 across all 612 samples were selected for further analysis, and subsequently standardized to have mean zero and standard deviation one, again across all 612 samples.

2.7 Multi-tissue clustering methods for comparison

We ran four multi-tissue clustering methods (see Supplementary Fig. S1):

- Revamp with reassignment threshold $\epsilon = 0.005$ and prior tissue similarities $\lambda_{t,t'} = \rho_{t,t'}^{\alpha}$, where $\rho_{t,t'}$ is the average correlation coefficient between samples from tissue t and t' measured in the same individual and $\alpha = 0.25$ is a dissipation parameter to scale the correlation values. Here we suggest to derive the similarity coefficients using Pearson's correlation, but other distance measures could be used.
- Revamp with reassignment threshold $\epsilon = 0.005$ and prior tissue similarities $\lambda_{t,t'} = 0$.
- An alternative method, which treats the expression profile of each gene g in each tissue t as a separate (gene, tissue) variable and clusters the resulting (gene, tissue)-by-individual expression matrix using the single-tissue clustering algorithm (Section 2.2). This results in a single set of clusters, which are disentangled into a set of linked clusters, by assigning gene g to cluster m in tissue t whenever (g, t) belongs to original cluster m. This method was called 'vertical data concatenation' before, and relies on having expression data from multiple tissues in the *same* individual. In STAGE, 21 individuals had data in all 7 tissues.
- Single-tissue clustering on the entire dataset of 612 samples (called 'horizontal data concatenation' before). This results in an identical clustering across all tissues. It is not a true multi-tissue clustering method, but is used as an overall benchmark to determine the relevance of a multi-tissue approach.

2.8 Validation data

To evaluate the biological relevance of each multi-tissue clustering method, we used the following approach:

- We performed GSEA using first the GOSlim ontology, that gives a broad overview of the ontology content without the detail of the specific fine-grained terms (http://www.geneontology.org/page/go-slim-and-subset-guide), and after on GO terms (http://www.geneontology.org/page/download-ontology).
- We assigned sets of 'regulators' to each of the modules considering as candidate regulators the tissue-specific sets of genes with significant eQTLs identified in Foroughi Asl *et al.* (2015) (2464 AAW, 3209 IMA, 4491 liver, 2534 SM, 2373 SF, 2994 VF and 5691 WB genes).
- We obtained human tissue protein–protein interaction (PPI) networks from Barshir *et al.* (2013). Specifically, we used TissueNet v2 networks consisting of curated experimentally detected PPIs between proteins expressed in Genotype-Tissue Expression dataset tissues 'Artery Aorta', 'Liver', 'Muscle Skeletal', 'Adipose Subcutaneous',

'Adipose Visceral' and 'Whole Blood', available for download at http://netbio.bgu.ac.il/labwebsite/? q=tissuenet2-download.

2.9 Validation methods

We tested for GO functional enrichment using the task go_annotation in the Lemon-Tree software, and task regulators were used to identify gene 'regulators' using a probabilistic scoring (Joshi *et al.*, 2009).

To test for enrichment of known PPIs in a given clustering, we calculated the fold-change enrichment as

$$\mathrm{FC} = \frac{\frac{\text{Number of co-clustered gene pairs with PPI}}{\text{Total number of PPI}}}{\frac{\text{Total number of co-clustered gene pairs}}{\text{Total number of gene pairs}}}.$$

All clustering methods were run on the seven available STAGE tissues, and the results for six tissues were used for validation (IMA did not have a matching tissue in the TissueNet database). To evaluate the clustering of a particular tissue, we used all PPIs for that tissue. To evaluate the core gene set of a cluster (for cluster m, the set of genes belonging to m in all tissues), we used the set of PPIs shared across all tissues.

Because the fold-change value is influenced by the number of clusters (more clusters results in fewer co-clustered pairs), we used the same number ($k = 12$) of clusters for all compared methods (Section 2.7).

3 Results

3.1 Multi-tissue clustering with revamp produces mappable clusters with tunable overlap levels

To identify co-expression clusters that reflect biological similarities and differences across tissues, we analyzed samples from seven tissues from the STAGE study. First we initialized revamp with the partition obtained from clustering all tissue samples using k-means with $k = 12$ clusters for all our analyses, as this value was near the inflection point of the elbow plots in all tissues (Supplementary Fig. S2). Then we updated the cluster assignments for each tissue independently using our Bayesian model-based score that depends on a set of hyper-parameters $\lambda_{t,t'}$, expressing prior beliefs on pairwise tissue similarities (Section 2.3), using a greedy optimization algorithm that has one free parameter ϵ, the minimum gain in Bayesian score for reassigning a gene from one cluster to another (Section 2.4). The resulting multi-tissue clustering consists of a set of linked clusters, where cluster k in one tissue corresponds to cluster k in any other tissue. Genes that belong to a particular cluster k in *all* tissues form a core set of genes with conserved coexpression across tissues, whereas genes that belong to cluster k in one or more, but not all, tissues form tissue-specific sets of genes that are *differentially* coexpressed with the core of cluster k.

To test the influence of the method parameters, we systematically tested a large space of parameter combinations (Supplementary Fig. S3). Both the reassignment threshold ϵ and tissue similarities $\lambda_{t,t'}$ ultimately govern the degree of overlap across tissues of the linked clusters, with small thresholds and near-zero similarities leading to nearly tissue-independent clusterings, and large thresholds and/or near-one similarities leading to nearly identical clusterings. Although ϵ and $\lambda_{t,t'}$ are to some extent interchangeable (i.e. a smaller threshold value can be compensated by a uniform increase in similarity values), setting ϵ to a small, non-zero value is recommended to avoid spurious reassignments due to numerical round-off errors in the Bayesian score calculation.

When comparing this partitioning with clustering tissues independently, the cluster quality is improved (Supplementary Table S1) and the similarities between tissues are stronger. The functional enrichment analysis revealed that a larger proportion of functional enriched categories were shared across two or more tissues (Supplementary Fig. S4). Moreover, similarity heatmaps showed that the degree of shared enrichment between tissues in our clustering was able to reflect the degree of overall expression similarity (Supplementary Fig. S5). Yet it is noteworthy to mention that multi-tissue clustering methods, and in particular revamp when using prior tissue similarities that is optimized based on Eq. (5), may show fuzzy borders when assessed with traditional validation methods like silhouette scores (see Supplementary Fig. S6).

3.2 Revamp multi-tissue clustering is more enriched for tissue protein–protein interactions than other approaches

To evaluate the performance of revamp, we ran four different multi-tissue clustering methods (see Methods), testing for each one for the enrichment of human tissue protein-protein interactions (PPIs) from the TissueNet database (Barshir *et al.*, 2013) among co-clustered genes, using six tissues that matched between STAGE and TissueNet.

On a tissue-by-tissue basis, running revamp with or without prior tissue similarity values resulted in similar fold-change enrichment values for tissue PPIs (average fold-change over 6 tissues of 1.49 and 1.48, respectively) as running single-tissue clustering on all samples together (average fold-change 1.50), and considerably higher enrichment than using vertically concatenated data (average fold-change 1.22) (Fig. 1). For a baseline reference, we also calculated enrichment for each tissue clustered individually using the single-tissue clustering method. Consistent with the assumption that analyzing data integratively using multi-tissue clustering should improve biological relevance, single-tissue clustering resulted in lower fold-change values (average fold-change 1.31) (Fig. 1).

We further reasoned that genes assigned consistently to the same cluster across all tissues ('core' cluster genes) should reflect tissue-independent interactions between these genes. To test this hypothesis, we calculated enrichment of tissue-independent PPIs (i.e. PPIs present in all six tissue PPI networks) among core cluster genes. For revamp with prior tissue similarity values, a significant increase in enrichment for tissue-independent PPIs was observed (fold-change 1.72), whereas for revamp without prior tissue similarities and horizontal data concatenation no difference was observed compared to all tissue PPIs (fold-changes 1.47 and 1.57, respectively) (Fig. 1). Vertical data concatenation resulted in very small core gene sets, containing no known tissue-independent PPIs (see also Supplementary Table S2).

3.3 Functional predictions by Revamp clusters and gene regulators associated with CAD

To test whether the clustering algorithm accurately captures the higher-level biological process represented by each module we first performed gene ontology enrichment analysis (see top enrichments in Supplementary Table S3). Network analysis revealed three

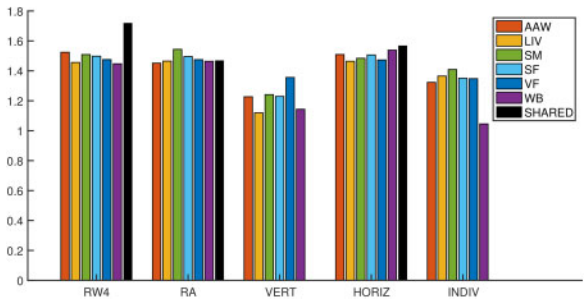

Fig. 1. Fold-change enrichment of tissue PPIs in tissue clusters for four multi-tissue clustering methods and individual single-tissue clustering. RW4—revamp with prior tissue similarities set according to their overall expression correlation, RA—revamp with prior tissue similarities set to zero, VERT—vertical data concatenation, HORIZ—horizontal data concatenation, INDIV—each tissue clustered individually. Each colored bar shows the fold-change overlap of tissue PPIs in clusters for the matching tissue; the black bar shows the fold-change overlap of tissue-shared PPIs in tissue-shared genes of linked clusters. See Section 2 for details. (Color version of this figure is available at *Bioinformatics* online.)

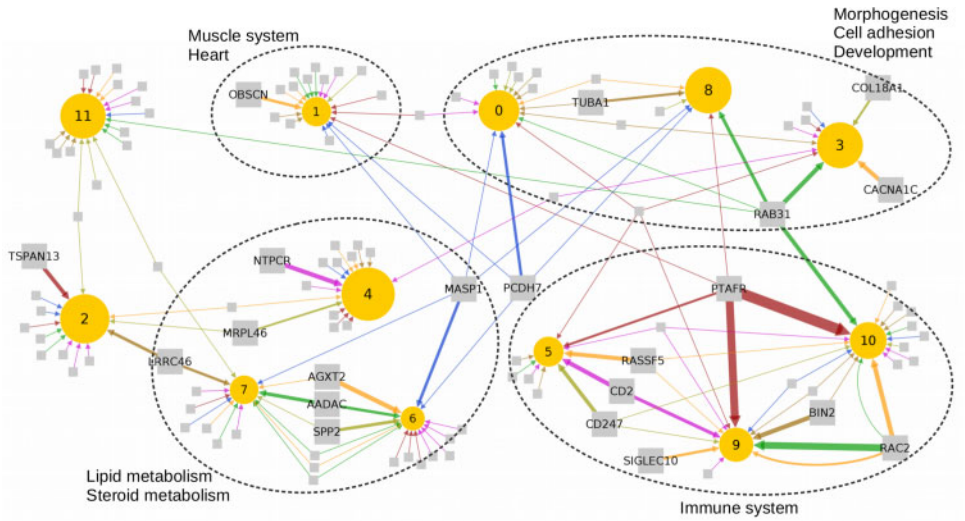

Fig. 2. Module regulatory network for all seven tissues. Regulators are presented as squares and clusters as circles with size proportional to the number of genes in the cluster. Only the regulators with a score greater than 20 in the regulators task are represented, and we named those with a score above 60. Edges are colored per tissue as per Figure 3, and their width is proportional to the regulator score. (Color version of this figure is available at *Bioinformatics* online.)

connected components: clusters 5, 9 and 10 were related with immune system response; the lipid metabolic process was enriched in clusters 4, 6 and 7; and clusters 0 and 8 were associated with cell adhesion and extracellular matrix organization.

Then we ran independently on each tissue the regulator probabilistic scoring task (see Section 2.9) to predict upstream regulatory genes, considering as candidate regulators the tissue-specific genes with genetic variants in their regulatory regions affecting gene expression ('*cis*-eQTL effects'). The regulatory network of the most significant regulators for the inferred modules is depicted in Figure 2.

The development of atherosclerosis is in large part mediated by the inflammatory cascade (Crowther, 2005). Our results indicated that the inflammatory response in AAW may be regulated by PTAFR, a mediator in platelet aggregation and the inflammatory response (Perisic *et al.*, 2016; Rastogi *et al.*, 2008). SF and VF were shown to be regulated by SIGLEC10 and CD247, respectively, genes that have been previously associated with CAD (Ammirati *et al.*, 2008; Shen *et al.*, 2013). Other tissues were linked to the previously identified inflammatory regulators BIN2 (Liao *et al.*, 2011), CD2 (Hansson and Libby, 2006), RAC2, that also directs plaque osteogenesis (Ceneri *et al.*, 2017), and the pro-apoptotic regulator of RAS protein, RASSF5 (Dejeans *et al.*, 2010).

Lipid metabolism also plays a key role in the development of atheroma plaques. Metabolism-related clusters 6 and 7 were found to be regulated by AGXT2 and SPP2, in SF and VF respectively. AGXT2 polymorphisms were identified as risk for CAD in Asian populations (Yoshino *et al.*, 2014; Zhou *et al.*, 2014), and SPP2 may contribute to the atheroprotective effects of HDL (Abdel-Latif *et al.*, 2015). AADAC, that controls the export of sterols (Tiwari *et al.*, 2007), may also be a regulator in SM. In WB, we found MASP1, a gene associated with a decreased lectin pathway activity in acute myocardial infarction patients (Yan *et al.*, 2016).

The atherogenic pathway involves the inflammation of the arterial wall, injury of the intima, lipid infiltration and activation of the angiogenic signaling, processes that involve a dysfunction in the cell adhesion (Sun, 2014). Our analysis showed that RAB31, which induces lipid accumulation in atheroma plaques (Fu *et al.*, 2002), regulates the morphogenesis-related clusters 3 and 8 in SM. Cluster 3 was also shown to be regulated by CACNA1C in SF, a gene involved in calcium channels and associated with inherited cardiac arrhythmia (Kawashiri *et al.*, 2014), and COL18A1 in VF, that may control angiogenesis and vascular permeability (Moulton *et al.*, 1999). The expression levels of PCDH7, gene involved in cell adhesion, and TUBA1 were also previously correlated with CAD (Chittur *et al.*, 2008; Eyster *et al.*, 2011; Sinnaeve *et al.*, 2009).

3.4 Revamp discovers multi-tissue clusters underlying CAD phenotypes

The systems genetics paradigm says that genetic variants in regulatory regions affect nearby gene expression ('*cis*-eQTL effects'), which then causes variation in downstream gene networks ('*trans*-eQTL effects') and clinical phenotypes. Ultimately, gene-gene interactions across metabolic and vascular tissues will enable information flow to the end stage phenotypic changes in CAD. We therefore used regression analysis to identify associations between module gene expression and CAD phenotypes (see Talukdar *et al.*, 2016), as presented in Figure 3.

The aggregated results revealed that AAW and SF are the main tissues associated with very-low-density lipoprotein (VLDL) and low-density lipoprotein (LDL) cholesterol levels, while the liver was the main tissue associated with high-density lipoprotein (HDL) cholesterol. Fat has been previously identified as the main contributor of CAD heritability, and the top regulatory networks in CAD have shown to be strongly enriched in associations with plasma levels of HDL, LDL and pro-insulin (Zeng *et al.*, 2019), as it is depicted in the left part of Figure 3.

Besides that, IMA was found to be associated in cluster 3 with the thyroid-stimulating hormone, that causes many hemodynamic effects and influences the structure of the heart and circulatory system (Grais and Sowers, 2014), and alcohol consumption in clusters 5 and 9, whose associations with cardiovascular diseases are heterogeneous (Bell *et al.*, 2017).

On the other hand, the results showed that the phenotypes related to anthropometric measurements are mostly associated with SM, liver and IMA, and with less significance with WB and AAW, but not with SF and VF. If we focus on clusters related to body weight, as a typical example of a trait regulated by, and affecting multiple tissues, we can find gene regulators such as PTAFR (in AAW) and CD2 (in IMA) which have been described to affect food intake and body weight, apart from the inflammatory response (Are Hanssen *et al.*, 2004; Li and McIntyre, 2015). In SM, RAC31 may influence on the body weight by mediating the insulin-stimulated glucose uptake (Lyons *et al.*, 1999). Last, also the candidate regulators BIN2 and RAC2 have been associated with obesity and metabolic syndrome (Aguilera *et al.*, 2013; Zhang *et al.*, 2005).

4 Conclusion

Herein we proposed a Bayesian model-based multi-tissue clustering algorithm, revamp, which incorporates prior information on

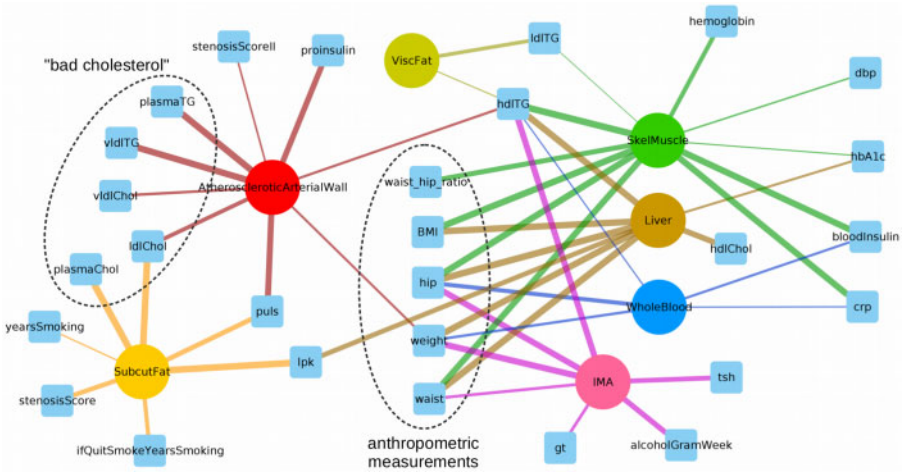

Fig. 3. Network representation of the correlation between the eigengenes, the first principal component of a given module, and relevant CAD phenotypes (squares), aggregated per tissue (circles). Edge width is inversely proportional to the correlation *P*-value

physiological tissue similarity, and which results in a set of clusters consisting of a core set of genes conserved across tissues as well as differential sets of genes specific to one or more subsets of tissues. Using data from seven vascular and metabolic tissues from the STAGE study, we demonstrated that our method resulted in multi-tissue clusters with higher enrichment of tissue-specific protein-protein interactions than comparable clustering algorithms. Moreover, the multi-tissue clusters highlighted the ability of revamp to link together regulatory genes, biological processes and clinical patient characteristics in a meaningful way across multiple tissues, and we believe this makes it an attractive and statistically sound method for analyzing multi-tissue gene expression datatsets in general. Revamp is implemented and freely available in the Lemon-Tree software at https://github.com/eb00/lemon-tree.

Funding

This work was supported by BBSRC [Roslin Institute Strategic Programme, BB/P013732/1] and the NIH [NHLBI R01HL125863]. P.E. has been partially supported by CRUK [C18281/A19169].

Conflict of Interest: none declared.

References

Abdel-Latif,A. *et al.* (2015) Lysophospholipids in coronary artery and chronic ischemic heart disease. *Curr. Opin. Lipidol.*, **26**, 432–437.

Aguilera,C.M. *et al.* (2013) Genetic susceptibility to obesity and metabolic syndrome in childhood. *Nutr. Hospital.*, **28**, 44–55.

Ammirati,E. *et al.* (2008) Expansion of T-cell receptor zeta dim effector T cells in acute coronary syndromes. *Arterioscl. Thrombosis Vasc. Biol.*, **28**, 2305–2311.

Andreopoulos,B. *et al.* (2008) A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief. Bioinf.*, **10**, 297–314.

Are Hanssen,S. *et al.* (2004) Costs of immunity: immune responsiveness reduces survival in a vertebrate. *Proc. Biol. Sci.*, **271**, 925–930.

Barshir,R. *et al.* (2013) The TissueNet database of human tissue protein–protein interactions. *Nucleic Acids Res.*, **41**, D841–D844.

Bell,S. *et al.* (2017) Association between clinically recorded alcohol consumption and initial presentation of 12 cardiovascular diseases: population based cohort study using linked health records. *BMJ (Clinical Research ed.)*, **356**, j909.

Bonnet,E. *et al.* (2015) Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput. Biol.*, **11**, e1003983.

Ceneri,N. *et al.* (2017) Rac2 modulates atherosclerotic calcification by regulating macrophage interleukin-1β production. *Arterioscl. Thrombosis Vasc. Biol.*, **37**, 328–340.

Chittur,S.V. *et al.* (2008) Histone deacetylase inhibitors: a new mode for inhibition of cholesterol metabolism. *BMC Genomics*, **9**, 507.

Crowther,M.A. (2005) Pathogenesis of atherosclerosis. *Hematol. Am. Soc. Hematol. Educ. Program*, **2005**, 436–441.

Dejeans,N. *et al.* (2010) Modulation of gene expression in endothelial cells by hyperlipaemic postprandial serum from healthy volunteers. *Genes Nutr.*, **5**, 263–274.

D'haeseleer,P. (2005) How does gene expression clustering work? *Nat. Biotechnol.*, **23**, 1499.

Dobrin,R. *et al.* (2009) Multi-tissue coexpression networks reveal unexpected subnetworks associated with disease. *Genome Biol.*, **10**, R55.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Erola,P. *et al.* (2019) Learning differential module networks across multiple experimental conditions. In: Sanguinetti,G., Huynh-Thu,V. (eds.) *Gene Regulatory Networks. Methods in Molecular Biology*, Vol. 1883. Humana Press, New York, NY, 303–321.

Eyster,K.M. *et al.* (2011) Gene expression signatures differ with extent of atherosclerosis in monkey iliac artery. *Menopause (New York, N.Y.)*, **18**, 1087–1095.

Foroughi Asl,H. *et al.* (2015) Expression quantitative trait loci acting across multiple tissues are enriched in inherited risk of coronary artery disease. *Circ. Cardiovasc. Genet.*, **8**, 305–315.

Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Franzén,O. *et al.* (2016) Cardiometabolic risk loci share downstream *cis* and *trans* genes across tissues and diseases. *Science*, **353**, 827–830.

Freeman,T.C. *et al.* (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput. Biol.*, **3**, e206.

Fu,J. *et al.* (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.*, **8**, e1002431.

Fu,Y. *et al.* (2002) The adipocyte lipid binding protein (ALBP/aP2) gene facilitates foam cell formation in human THP-1 macrophages. *Atherosclerosis*, **165**, 259–269.

Ghazalpour,A. *et al.* (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.*, **2**, e130.

Grais,I.M. and Sowers,J.R. (2014) Thyroid and the heart. *Am. J. Med.*, **127**, 691–698.

Greenawalt,D.M. *et al.* (2011) A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.*, **21**, 1008–1016.

Grundberg,E. *et al.* (2012) Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084.

GTEx Consortium. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204.

Hägg,S. *et al.* (2009) Multi-organ expression profiling uncovers a gene module in coronary artery disease involving transendothelial migration of leukocytes and LIM domain binding 2: the Stockholm Atherosclerosis Gene Expression (STAGE) study. *PLoS Genet.*, **5**, e1000754.

Hansson,G.K. and Libby,P. (2006) The immune response in atherosclerosis: a double-edged sword. *Nat. Rev. Immunol.*, **6**, 508–519.

Ickstadt,K. *et al.* (2017) Toward integrative Bayesian analysis in molecular biology. *Annu. Rev. Stat. Its Appl.*, **5**, 141–167.

Joshi,A. *et al*. (2008) Analysis of a Gibbs sampler for model based clustering of gene expression data. *Bioinformatics*, **24**, 176–183.

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

Kawashiri,Ma. *et al*. (2014) Current perspectives in genetic cardiovascular disorders: from basic to clinical aspects. *Heart Vessels*, **29**, 129–141.

Keller,M.P. *et al*. (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res*., **18**, 706–716.

Langfelder,P. *et al*. (2011) Is my network module preserved and reproducible? *PLoS Comput. Biol*., **7**, e1001057.

Li,W. and McIntyre,T.M. (2015) Platelet-activating factor receptor affects food intake and body weight. *Genes Dis*., **2**, 255–260.

Li,W. *et al*. (2011) Integrative analysis of many weighted co-expression networks using tensor computations. *PLoS Comp. Biol*., **7**, e1001106.

Liao,Y.C. *et al*. (2011) BRAP activates inflammatory cascades and increases the risk for carotid atherosclerosis. *Mol. Med. (Cambridge, Mass.)*, **17**, 1065–1074.

Lyons,W.E. *et al*. (1999) Brain-derived neurotrophic factor-deficient mice develop aggressiveness and hyperphagia in conjunction with brain serotonergic abnormalities. *Proc. Natl. Acad. Sci. USA*, **96**, 15239–15244.

Moulton,K.S. *et al*. (1999) Angiogenesis inhibitors endostatin or TNP-470 reduce intimal neovascularization and plaque growth in apolipoprotein E-deficient mice. *Circulation*, **99**, 1726–1732.

Perisic,L. *et al*. (2016) Gene expression signatures, pathways and networks in carotid atherosclerosis. *J. Internal Med*., **279**, 293–308.

Ponnapalli,S.P. *et al*. (2011) A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PLoS One*, **6**, e28072.

Qin,Z. (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, **22**, 1988–1997.

Rastogi,P. *et al*. (2008) Potential mechanism for recruitment and migration of CD133 positive cells to areas of vascular inflammation. *Thrombosis Res*., **123**, 258–266.

Roy,S. *et al*. (2013) Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res*., **23**, 1039–1050.

Shen,H. *et al*. (2013) Processes of sterile inflammation. *J. Immunol. (Baltimore, Md.: 1950)*, **191**, 2857–2863.

Si,Y. *et al*. (2014) Model-based clustering for RNA-seq data. *Bioinformatics*, **30**, 197–205.

Sinnaeve,P.R. *et al*. (2009) Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One*, **4**, e7037.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*., **9**, 3273–3297.

Sun,Z. (2014) Atherosclerosis and atheroma plaque rupture: normal anatomy of vasa vasorum and their role associated with atherosclerosis. *Sci. World J*., **2014**, 285058.

Talukdar,H. *et al*. (2016) Cross-tissue regulatory gene networks in coronary artery disease. *Cell Syst*., **2**, 196–208.

Tiwari,R. *et al*. (2007) An acetylation/deacetylation cycle controls the export of sterols and steroids from *S.cerevisiae*. *EMBO J*., **26**, 5109–5119.

van Dam,S. *et al*. (2017) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinf*., **19**, 575–592.

Xiao,X. *et al*. (2014) Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genet*., **10**, e1004006.

Yan,W. *et al*. (2016) Depletion of complement system immunity in patients with myocardial infarction. *Mol. Med. Rep*., **14**, 5350–5356.

Yoshino,Y. *et al*. (2014) Missense variants of the alanine: glyoxylate aminotransferase 2 gene correlated with carotid atherosclerosis in the Japanese population. *J. Biol. Regul. Homeostat. Agents*, **28**, 605–614.

Zeng,L. *et al*. (2019) Contribution of gene regulatory networks to heritability of coronary artery disease. *J. Am. College Cardiol*., **73**, 2946–2957.

Zhang,X. *et al*. (2005) High dietary fat induces NADPH oxidase-associated oxidative stress and inflammation in rat cerebral cortex. *Exp. Neurol*., **191**, 318–325.

Zhou,J.P. *et al*. (2014) Association of the AGXT2 V140I polymorphism with risk for coronary heart disease in a Chinese population. *J. Atheroscl. Thrombosis*, **21**, 1022–1030.