Recommendations for the analysis of gene expression data to identify intrinsic differences between similar tissues

Abbassi Daloii, T.; Kan, H.E.; Raz, V.; Hoen, P.A.C. 't

Genomics 112 (2020) 3157–3165

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Recommendations for the analysis of gene expression data to identify intrinsic differences between similar tissues

Tooba Abbassi-Daloii[a], Hermien E. Kan[b,c], Vered Raz[a], P.A.C. 't Hoen[a,d,*]

[a] Department of Human Genetics, Leiden University Medical Center, the Netherlands
[b] C.J. Gorter Center for High Field MRI, Department of Radiology, Leiden University Medical Center, the Netherlands
[c] Duchenne Center Netherlands, the Netherlands
[d] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center

ARTICLE INFO

Keywords:
Differential expression
Module detection
Co-expressed genes
Prior knowledge
Skeletal muscle

ABSTRACT

Identifying genes involved in functional differences between similar tissues from expression profiles is challenging, because the expected differences in expression levels are small. To exemplify this challenge, we studied the expression profiles of two skeletal muscles, deltoid and biceps, in healthy individuals. We provide a series of guides and recommendations for the analysis of this type of studies. These include how to account for batch effects and inter-individual differences to optimize the detection of gene signatures associated with tissue function. We provide guidance on the selection of optimal settings for constructing gene co-expression networks through parameter sweeps of settings and calculation of the overlap with an established knowledge network. Our main recommendation is to use a combination of the data-driven approaches, such as differential gene expression analysis and gene co-expression network analysis, and hypothesis-driven approaches, such as gene set connectivity analysis. Accordingly, we detected differences in metabolic gene expression between deltoid and biceps that were supported by both data- and hypothesis-driven approaches. Finally, we provide a bioinformatic framework that support the biological interpretation of expression profiles from related tissues from this combination of approaches, which is available at github.com/tabbassidaloii/AnalysisFrameworkSimilarTissues.

1. Introduction

Gene expression profiling technologies such as expression microarrays or RNAseq have successfully been applied to improve the understanding of the molecular basis of distinct tissue functions [1–4]. However, it is still not sufficiently clear whether differences between similar tissues, such as different types of adult skeletal muscles, are entrenched in the gene expression pattern [1,2,5]. Differential expression analysis (DEA) is the initial step in identifying genes that could discriminate between tissue functions. Application of DEA can be a challenge when studying similar tissues as expression level differences could be small. If molecular differences are small, those might be obscured by technical variation, and large sample sizes may be required for DEA [6]. Another limitation of DEA is that genes are studied in isolation, whilst genes and their gene products operate in networks [7]. Genes participating in the same molecular and biological processes tend to show correlated expression patterns (co-expression), because they can be under control of the same transcriptional regulations [8]. Considering this, numerous methods for gene co-expression network analysis have been developed to complement DEA. These methods cluster groups of co-expressed genes into modules [7,9–11]. These modules could be unique to a specific condition or tissue [12].

Weighted gene co-expression network analysis (WGCNA), the most widely used approach for co-expression analysis, constructs a network that is based on the pairwise correlations between genes expression levels [12,13]. WGCNA assumes a scale-free network following a power law distribution. This distribution, which seems to reflect many biological phenomena, is characterized by a small fraction of highly connected and a large fraction of lowly connected genes [12,14]. To achieve this power law distribution, sparsity is introduced in the network by a power transformation on the correlations followed by a soft-threshold. The power transformation emphasizes strong pairwise correlations and downtones weak correlations [12]. A hierarchical clustering algorithm groups co-expressed genes together in increasingly large modules. Different thresholds used by the Dynamic Tree Cut algorithm determine the size of the modules [15]. There are other parameters, discussed in [15], that can affect gene co-expression network construction, network connectivity and the number of modules.

* Corresponding author at: CMBI 260, Radboud University Medical Center, PO Box 9101, 6500 HB Nijmegen, the Netherlands.
E-mail address: Peter-Bram.tHoen@radboudumc.nl (P.A.C. 't Hoen).

https://doi.org/10.1016/j.ygeno.2020.05.026
Received 12 February 2020; Received in revised form 5 May 2020; Accepted 26 May 2020
Available online 30 May 2020

However, how to optimally select the combination of WGCNA parameters that create modules with the highest biological coherence, is still unsolved. It has been reported that using prior knowledge to construct the gene co-expression network improves the performance of gene co-expression analysis for the extraction of biologically meaningful modules [16–18]. Therefore, here we used an analysis framework where the prior knowledge of gene interactions from a pathway database was considered as a criterion to select an optimal set of WGCNA parameters. We assessed this analysis framework using a human expression dataset containing paired deltoid and biceps muscles from healthy individuals [19]. These two muscles are connected to the humerus bone but have different biomechanical function and muscle architecture [20]. We demonstrate the complementarity of the data-driven approaches, DEA and WGCNA, and a hypothesis-driven approach, gene set connectivity analysis, to gain insight into the functional differences between deltoid and biceps. Therefore, we suggest applying our analysis framework could improve functional interpretation of molecular expression profiles of related tissues.

2. Results

2.1. Differential gene expression analysis

We utilized a published dataset of healthy individuals from deltoid and biceps [19] to illustrate the complementary use of data-driven and hypothesis-driven approaches (Suppl Fig. 1) for the discovery of molecular signatures that are distinct between these two muscles. Data from two outlier individuals were removed. In total, the expression data of ten individuals (aged 48.5 ± 8.75 yr) was analyzed to investigate gene contributing to the intrinsic differences between muscles. Consistent with the previous study [19] principal component analysis (PCA) indicated a strong batch effect (Suppl. Fig. 2). Muscle samples from the same individuals were proceeded in the same batch. Accordingly, we corrected for the batch effect by paired differential expression analysis (DEA) of deltoid versus biceps from the same individual. Our DEA showed a small number of differentially expressed genes between deltoid and biceps which is in agreement with the previous study [19]. A total of 130 genes (out of 18,217 expressed genes) were differentially expressed (FDR of 0.05; Suppl. Table 1), and only six genes showed differences higher than 2-fold (Fig. 1). There was no enrichment of biological or molecular processes for the differentially expressed genes (FDR < 0.05). This confirms that the contrast between deltoid and biceps is not sufficiently high for the DEA. Therefore, we subsequently performed the gene co-expression network analysis.

2.2. Gene co-expression network analysis

2.2.1. The batch effect correction

We applied unsupervised weighted gene co-expression network analysis (WGCNA) [12] to identify gene modules whose expression pattern differ between the two muscles. Using dynamic tree-cutting algorithm, the genes were clustered into modules based on their expression values across all samples in both muscles. Subsequently, the association of each module with muscle type, age or batch was evaluated. Nine out of 10 modules (containing 79% of genes) were significantly associated with the batch effect, while none was associated with muscle type (Suppl. Fig. 3). We regressed out the batch effect using linear models and the residuals of the linear models were used for WGCNA. The correction for batch effect as a technical variation can preserve the meaningful biological signals [21,22].

2.2.2. Use of a knowledge database to select the most optimal WGCNA parameters

Different WGCNA settings determine the nature, size, number and connectivity of the modules [13]. It is a priori unclear which settings would provide modules with the highest biological coherence.

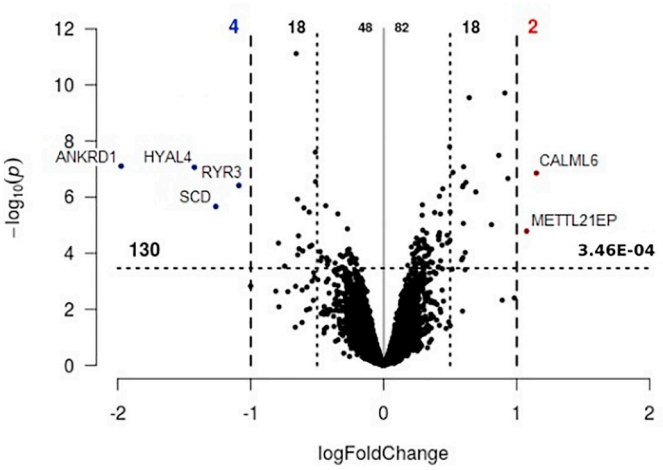

Fig. 1. Volcano plot displaying expression differences between biceps and deltoid muscles. On the y-axis is the inverted 10log p-value for differential expression between biceps and deltoid, and the x-axis shows the log2 fold change (logFC). The genes with higher expression in biceps are on the right and those with higher expression in deltoid are on the left. The dashed horizontal line marks the p-value cut-off. Points above the line are differentially expressed genes with adjusted p-values (FDR) lower than 0.05. Genes with logFC > 1 are labelled.

Therefore, we performed a full parameter sweep, testing various combinations of settings for power, minClusterSize, deepSplit and CutHeight. In order to assess the performance of these different parameters, the co-expressed pairs (CPs) and knowledge pairs (KPs) were determined using the edges (Fig. 7c). The knowledge network was obtained from the Reactome database. In the knowledge network, there is an edge between genes when they have at least one pathway in common. Only genes with annotation in the Reactome database were considered to define the CPs (8504 genes in total). Then, the enrichment factors, the ratio between overlapping and nonoverlapping pairs, were calculated to score different sets of parameters (Fig. 7c; Suppl. Table 2). The settings for which more than 30% of the genes were not assigned to any module were excluded, as this would limit the possibilities to assign biological processes to genes. The enrichment factor values showed highest changes across different settings for power and CutHeight, whereas for minClusterSize and deepSplit changes were lower (Fig. 2a–d). The power and cutHeight parameters were the most important parameters determining the consistency of modules in the co-expression networks with the KPs of the knowledge database. Generally, the settings with higher numbers of modules and smaller module sizes had a higher enrichment factor (Fig. 2e). These results suggest that the consistency of the co-expression network with a knowledge network can be considered to determine the optimal set of WGCNA parameters.

2.2.3. Selection of a knowledge database

The optimal WGCNA setting may depend on the knowledge database selected. To validate this dependency, we calculated the enrichment factors with a knowledge network that was obtained from Human Phenotype Ontology (HPO). The HPO captures different information than the Reactome. We then compared the enrichment factors for different settings between HPO and Reactome (Fig. 2f). The optimal WGCNA settings obtained from HPO were different from those obtained from Reactome (Fig. 2f; Suppl. Table 2). Accordingly, the correlation coefficient (0.58) was modest. This underscores that the optimal WGCNA settings depends on the selected knowledge base.

To assess if our framework could be used to identify modules delineating deltoid or biceps muscle groups and to explore the molecular differences between these two muscles, we considered the co-expression network with the highest overlap with the Reactome knowledge database. The selected settings were: power = 8, minClusterSize = 15,

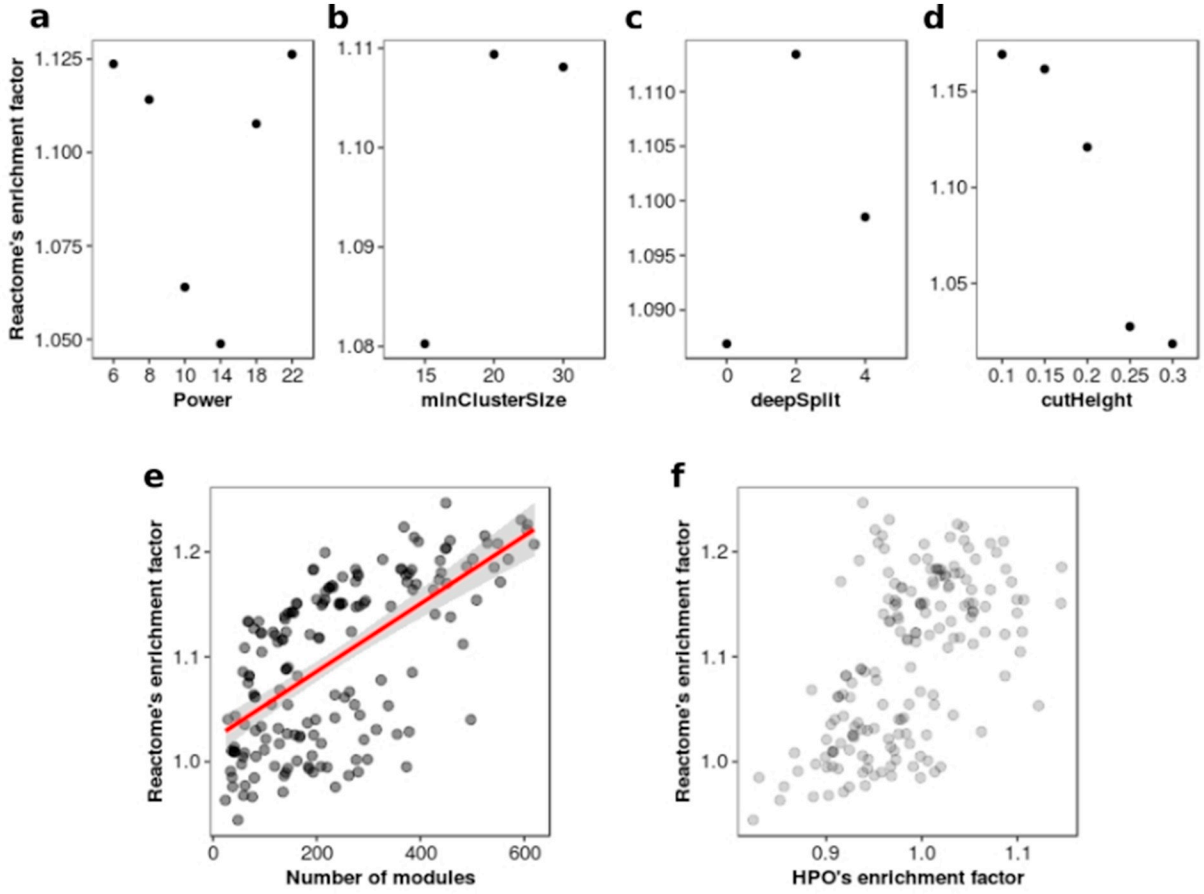

Fig. 2. Usage a knowledge database to determine the most optimal WGCNA settings. a–d) The x-axis shows different settings for a) power, b) deepSplit, c) minClusterSize, and d) cutHeight. The y-axis represents the enrichment factor, which is defined as the ratio between number of overlapping and nonoverlapping pairs in the co-expression and knowledge networks. Each dot shows the median of enrichment factors from the combination of the given parameter on the x-axis and all other parameters. e) The enrichment factor increases as a function of the number of modules. Each dot represents a different WGCNA setting. X-axis and y-axis represent the number of modules and the enrichment factor, respectively. f) The optimal WGCNA settings differ when evaluating with the Reactome or HPO knowledge network. Each dot represents a different WGCNA setting. X-axis and y-axis represent the Reactome and HPO's enrichment factors, respectively. For panel (e) and panel (f), only settings which assigned > 70% of genes to a module (< 30% genes in the 'grey' module) were included in the calculation.

deepSplit = 2 and CutHeight = 0.2.

2.2.4. Muscle-related modules identified with linear mixed-effect models

The module eigengene (ME) is the first principal component of the expression levels of the genes in a module. The ME was calculated for each module to represent the gene expression profiles of the genes in the module. Then, linear mixed-effect models were fitted to evaluate how the eigengene depended on muscle type and age (modeled as fixed effects) and the individual (modeled as random effect). No module was found to be associated with age. On the contrary, we identified 18 (out of 449) modules delineating the two muscle types (Fig. 3, size range: 15–148; containing 1017 genes in total). These 18 muscle-related modules contained 98 out of 130 significantly differentially expressed genes. Seven out of the 18 muscle-related modules demonstrated higher expression levels in deltoid, and the remaining 11 modules demonstrated higher expression levels in biceps (Fig. 3a). The 18 modules contained a larger gene list than obtained with DEA (Fig. 3b). Subsequently, we explored molecular differences between the two muscle using genes in the 18 modules, as described in the following paragraphs.

2.3. Data-driven identification of mitochondrial metabolism-related differences between deltoid and biceps

To identify functional differences between deltoid and biceps muscles, each muscle-related module was annotated using g:Profiler which compiles pathways and gene annotations from several knowledge databases. The significantly enriched biological processes are listed in Suppl. Table 3. In three modules with higher expression in deltoid (M.125, M.367 and M.54), the mitochondria term was enriched (Suppl. Table 3). Fig. 4 shows the expression level of mitochondrial genes in each module with overall higher expression levels in deltoid than biceps. From these 51 genes, only BSG and FAM110B were found to be differentially expressed using DEA (FDR < 0.05). This explains why mitochondrial pathways were not discovered in the enrichment analysis of differentially expressed genes.

To further investigate a role for these 51 genes in the mitochondrial function, they were mapped to the mitochondrial pathways (Fig. 5). The oxidative phosphorylation (OXPHOS), tricarboxylic acid (TCA) cycle and beta-oxidation were the most prominent.

2.4. Hypothesis-driven identification of differences in aerobic metabolism between muscles

The enrichment analysis suggested differences in aerobic metabolism between deltoid and biceps. To further investigate this metabolism differences, we performed a hypothesis-driven gene set connectivity analysis for the genes involved in respiratory electron transport and TCA cycle. The genes annotated for these pathways were retrieved from Reactome (R-HSA-611105 and R-HSA-71403). We determined the

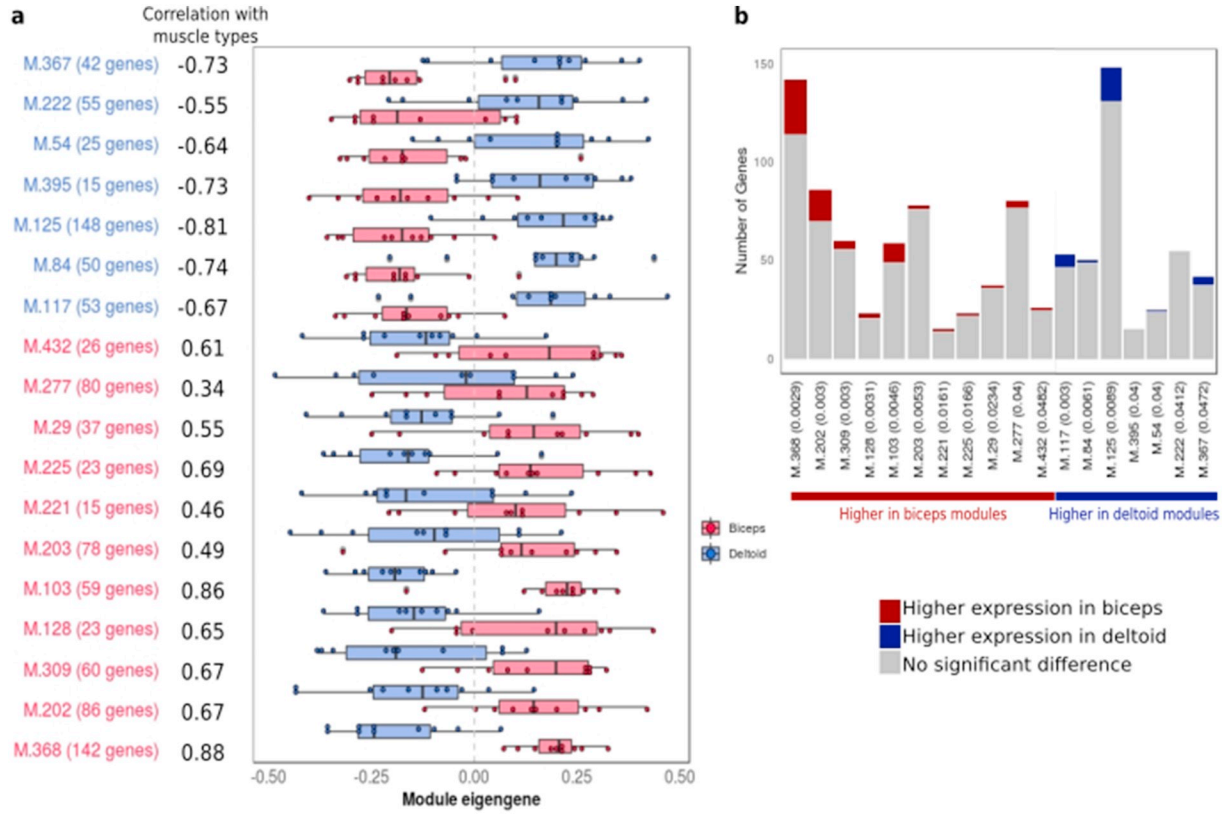

Fig. 3. Muscle-related gene co-expression modules. a) Boxplot shows gene co-expression modules with overall higher expression in biceps (red) or deltoid (blue). The loading of each individual on the eigengene of each module is represented by a dot, and the boxes represent the median and interquartile range. The module size is given between parentheses. The Pearson correlation with the muscle type without adjusting for age and individual is shown; b) The representation of the differentially expressed genes from DEA (FDR $<$ 0.05) in the significant modules. Numbers between parentheses represent FDRs related to the association with muscle types. The differentially expressed genes that have higher expression levels in biceps or deltoid are highlighted in red and blue, respectively, the non-differentially expressed genes in gray. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

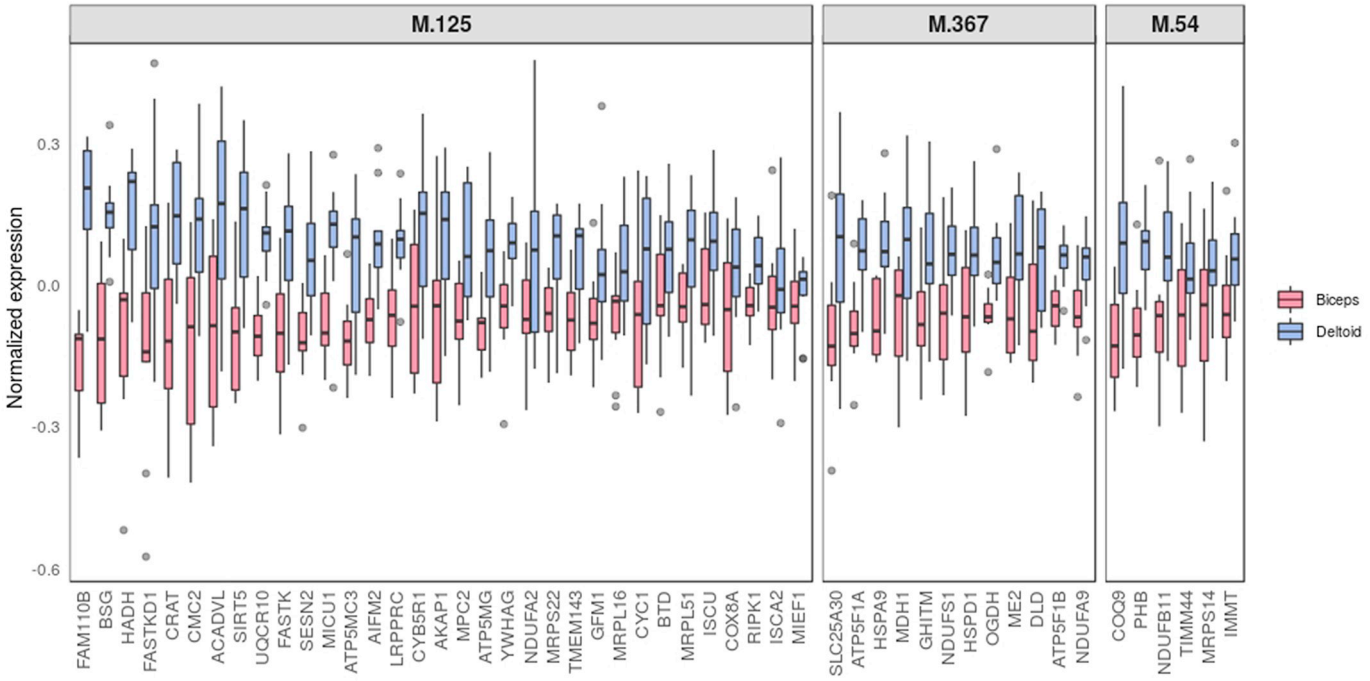

Fig. 4. Expression levels of mitochondrial genes in the three modules that are higher in deltoid. Expression levels of genes that are assigned to mitochondria (GO:0005739) in M.125 and M.367, and mitochondrial inner membrane (GO:0005743) in M.54. The y-axis shows normalized expression levels and the boxes reflect the median and interquartile range with overall higher expression in deltoid (blue) than biceps (red). The genes are ordered by their absolute LogFC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

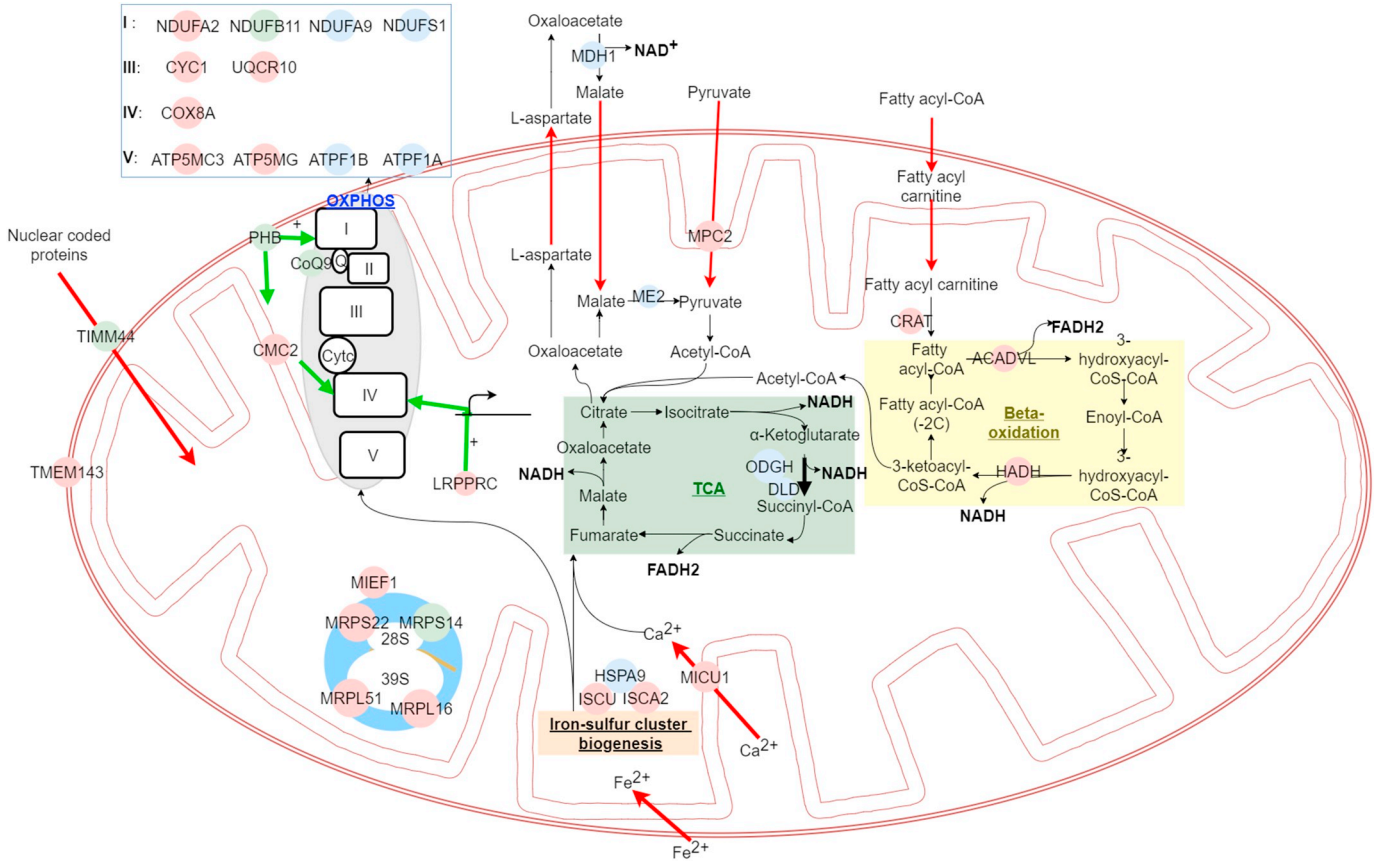

Fig. 5. A schematic representation of the function of genes with higher expression in deltoid in the mitochondrial structures and pathways. 34 (out of the 51) mitochondrial genes are shown and colored according to their modules: M.125 in pink, M.367 in blue and M.56 in green. The OXPHOS (gray), ribosome (blue), iron-sulfur cluster biogenesis (orange), TCA cycle (green) and beta-oxidation (yellow) are highlighted. The genes related to different main protein complexes of OXPHOS are shown in the rectangle out of the mitochondria. In the TCA cycle rectangle, the bold arrow shows the rate-limiting step. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

topological overlaps of aerobic metabolism genes in the co-expression network (Fig. 6a). Topological overlap defines as a similarity measure between each pair of genes in relation to all other genes in the network. High topological overlaps indicate that genes share the same neighbors in the co-expression network [26]. Many genes in the aerobic metabolism subnetwork were highly connected but they were assigned to 58 different modules (Fig. 6a). We checked whether genes in this subnetwork were more strongly connected than would be expected by chance. A bootstrapping approach was used to sample 1000 subnetworks with the same number of nodes (98 nodes), and the mean of topological overlap of all the nodes for each random subnetwork was calculated (Fig. 6b). The topological overlap of the nodes in the subnetwork of metabolic genes was significantly higher than for the random networks showing that the genes related to aerobic metabolism are co-expressed (Fig. 6b).

As fast and slow myofibers differ in mitochondria content, we hypothesized that myofiber types might also be distinct between the two muscles. The slow-twitch myofibers show higher aerobic activity and possess more mitochondria than the fast-twitch myofibers [23]. To investigate this hypothesis, the expression profiles of fast-twitch and slow-twitch genes were assessed (Suppl. Fig. 4). The list of fast-twitch and slow-twitch genes was compiled from [24], and contained mostly genes coding for sarcomeric proteins. Six out of the nine fast-twitch genes, including myosin heavy chain *MYH1*, were higher in biceps. Whereas, six out of ten slow-twitch markers, including myosin heavy chain *MYH7*, were higher in deltoid (Suppl. Fig. 4). A relatively higher expression of *MYH7* and *MYH2* in deltoid than biceps suggests a higher aerobic activity in this muscle [25]. We next assessed the topological overlaps of myofiber type genes in the co-expression network. In general, the topological overlaps among fast-twitch genes was distinct from the slow-twitch genes but they were assigned to 13 different modules (Fig. 6c). The module enrichment analysis did not reveal differences in sarcomeric gene expression (Suppl. Table 3). Therefore, we checked whether genes in myofiber type subnetwork were significantly connected. 1000 subnetworks with 19 nodes were sampled and the mean of topological overlap of all the nodes for each random subnetwork was calculated (Fig. 6d). The genes in the subnetwork of myofiber type genes had significantly higher topological overlap than genes in random networks.

Further, to investigate the relation between mitochondrial and myofiber type gene expression, we created a heatmap based on the Pearson correlation coefficient between these genes. In general, the slow-twitch genes, including *MYH7* specific slow-twitch marker, were correlated with mitochondrial genes (Suppl. Fig. 5). This suggests a higher aerobic metabolism in deltoid. In contrast, the majority of the fast-twitch genes showed no correlation with mitochondrial genes (Suppl. Fig. 5). Taken together, the hypothesis-driven gene set connectivity analysis supported the differences in aerobic metabolism suggested by data-driven approach.

3. Discussion

We present a gene expression analysis framework to gain insight into the molecular signatures that drive functional differences between similar tissues, such as different skeletal muscles. We show that the modest differences between muscle types could not be determined

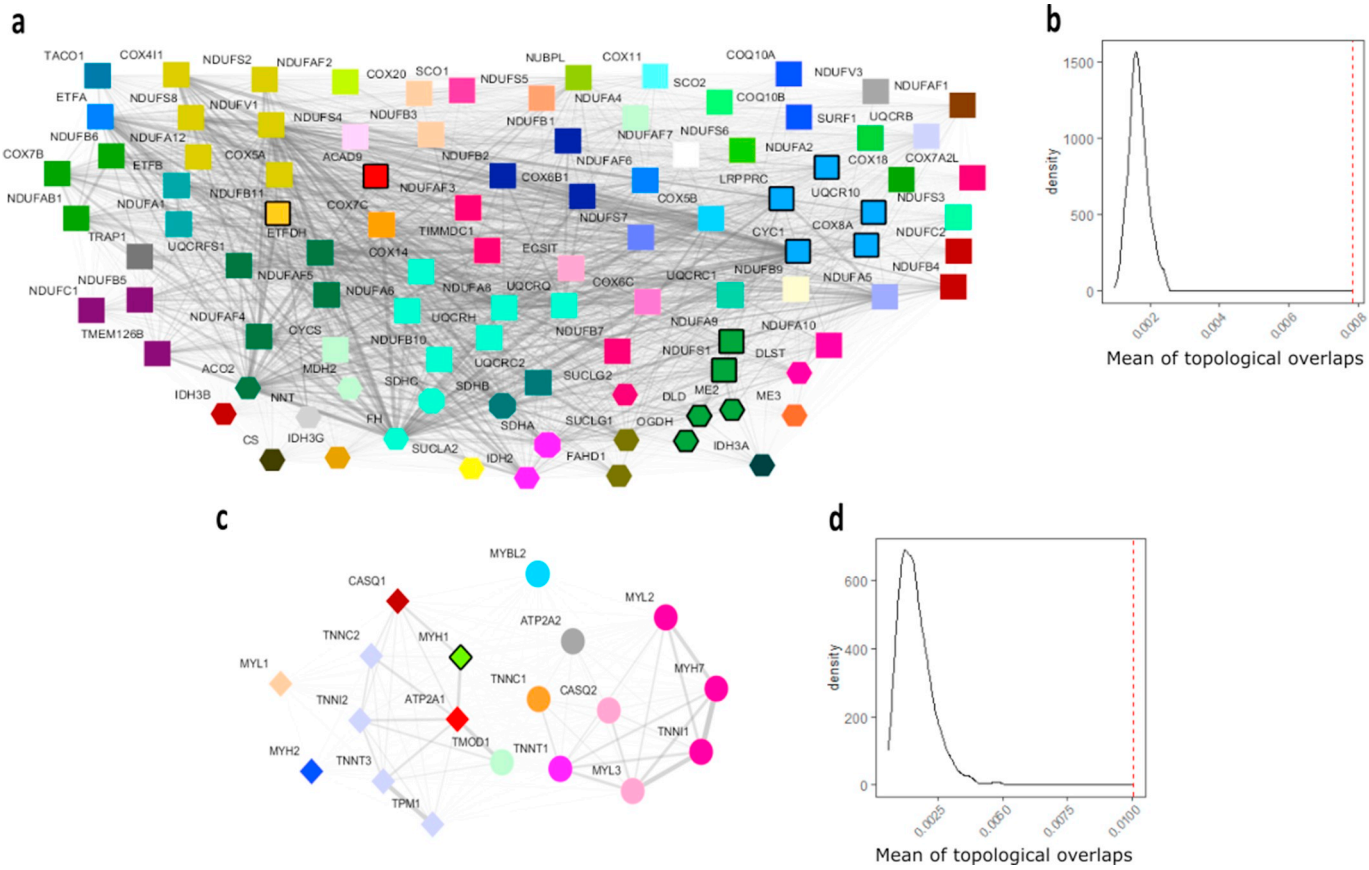

Fig. 6. A gene set connectivity analysis supports differences in aerobic metabolism between deltoid and biceps. a) The subnetwork of genes involved in respiratory electron transport and TCA cycle pathways. Square, hexagon and octagon indicate genes related to respiratory electron transport, TCA cycle and both pathways, respectively. The edge thickness reflects the degree of topological overlap. The colors represent the different modules to which the genes belong. The genes in the muscle-related modules have a black border. The muscle-related modules include M.125 (blue), M.54 (orange), M.367 (green) and M.203 (red) b) The distribution of mean topological overlap among gene pairs in 1000 random subnetworks containing 98 genes. The vertical dashed red line shows the mean topological overlap of the aerobic metabolism subnetwork. c) The subnetwork of the fast-twitch and slow-twitch genes. Diamonds and circles indicate genes related to fast-twitch and slow-twitch myofibers, respectively. The edge thickness reflects the degree of topological overlap. The colors represent the different modules to which the genes belong. The gene in the muscle-related modules (M.277) is green and has a black border. d) The distribution of mean topological overlap among gene pairs in 1000 random subnetworks containing 19 genes. The vertical dashed red line shows the mean topological overlap in the myofiber type gene subnetwork. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

through differential expression analysis alone but became more apparent by unsupervised gene co-expression network analysis and hypothesis-driven approaches. Based on our analysis framework, we make the following practical recommendations to study gene expression profiles from similar tissues with subtle differences in expression level: (1) DEA should be complemented by unsupervised gene co-expression network analysis to identify biological processes with small but coherent differences. (2) Correcting for the known sources of unwanted (technical) variations is a crucial data preparation step for network analysis in order to better preserve the meaningful biological signals. This is particularly true in the case biological differences between conditions are small and may be hidden by technical sources of variation. (3) WGCNA parameters can be adjusted through evaluation of consistency with known biological pathways or networks. This improves the biological interpretation of a WGCNA-derived network compared to WGCNA networks obtained with default or an arbitrary selection of parameters. (4) The knowledge database should be consciously selected based on the research purposes as it has effect on scoring different settings. (5) Data-driven approaches can be complemented by the hypothesis-driven approach to gain a better understanding of functional differences.

A unique aspect of our work is the parameter sweep that was employed to construct various co-expression networks using different WGCNA set of parameters. While it is recommended to define the soft-threshold power based on a scale-free topology criterion [12], this is a rule-of-thumb and based on visual inspection. CEMiTool [27] is another network analysis tool that provides an automated module detection workflow follows the standard WGCNA steps. However, the soft-threshold power selection in CEMiTool is based on a linear regression fit that quantifies whether the degree distribution of the genes in the network follows a power law distribution. This is a more objective approach for the selection of the soft power threshold than visual inspection of the curve reflecting the scale-free topology. However, the effect of other parameters (e.g. minimum module size) on the network construction cannot be tuned by CEMiTool [27]. We performed the parameter sweep followed by an evaluation of the consistency with established knowledge networks, as a more robust evaluation of the most optimal WGCNA settings.

Identifying modules with different expression between groups of samples is different from differential co-expression analysis (DCA) [28]. The DCA is an extension of co-expression analysis and identifies genes with different co-expression patterns [8,29]. Differences in co-expression patterns may arise from differential regulation of gene expression, as for example shown for cancer vs normal tissues [30,31]. Modules that are subject to similar regulation but show difference in abundance between the groups, will remain undetected by the DCA [29]. Those modules, however, will be identified with our approach involving linear (mixed) models of the association of the eigengene with the group

membership. DCA is therefore complementary to our approach. However, we could not apply DCA because it requires a bigger number of samples per group than available in our study [32].

In our framework, WGCNA followed by pathway enrichment analysis of muscle-related modules suggested the involvement of mitochondrial gene expression. Mitochondrial pathways were not enriched in differentially expressed genes. Mitochondrial genes showed only small differences in expression between muscles but a consistent higher expression in deltoid. From those, *DLD* and *OGDH* encode components of the α-ketoglutarate dehydrogenase complex (KGDHC), a rate-limiting enzyme of the citric acid (TCA) cycle [33–35]. The collective results point at a slightly higher aerobic activity in deltoid compared to biceps muscles. This illustrates that our analysis framework can identify biological relevant differences between similar tissues such as deltoid and biceps.

In summary, we successfully applied both data- and hypothesis-driven approaches to determine genes discriminating between related tissues, and to identify biological processes which could not be detected using differential expression analysis.

4. Methods

4.1. Preprocessing of expression data

A published normalized microarray dataset was obtained from Gene Expression Omnibus (ncbi.nlm.nih.gov/geo), GSE36398, containing paired deltoid and biceps samples collected from healthy adult individuals [19]. Probe identifiers of Affymetrix Human Gene 1.0 ST Array were annotated with human Ensembl gene using biomaRt, an R package interface with the BioMart database [36]. Multiple probes that assigned to the same genes were considered separately in all analyses. Generally, we referred to probes as genes. Genes whose expression values were below the first quantile were filtered out. Samples with a standardized sample connectivity below a threshold of −2.5 were considered outliers, where the standardized sample connectivity was defined as the overall Euclidean distances between a given sample and all other samples [37]. Both samples from the outlier individual were excluded. We removed samples from one more individual (aged 24 yr) to create a dataset with a smaller age range. The final datasets contained ten individuals with paired biceps and deltoid samples.

4.2. Differential expression analysis (DEA)

Differentially expressed genes were identified using linear models with Limma R package version 3.26.9 [38]. Gender and muscle type (deltoid or biceps) were considered in the linear models. The samples were processed in different arrays leading to a batch effect noted already by the authors of the original paper describing the dataset [19]. However, both samples from each individual were included in the same batch. In our DEA, we corrected for this batch effect by preforming paired comparison of biceps and deltoid from the same individuals. The paired analysis was specified using the duplicate correlation argument implemented in limma's linear models. The Benjamini-Hochberg false-discovery rate (FDR) was applied to adjust for multiple testing. Enrichment analysis for functional groups was performed using the g:ProfileR R package (version 0.6.7). This package collects pathways and gene annotations from several knowledge databases including Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), Reactome and Human Phenotype Ontology (HPO) [39].

4.3. Weighted co-expression networks, module detection and co-expressed pairs

We used linear regression to correct for the batch effect and used the residuals from these linear regression models to perform the weighted gene co-expression network analysis using WGCNA R package: version 1.61 [13]. To calculate co-expression, we used the biweight mid-correlation (median-based) function in WGCNA of the signed hybrid type. In this type of network, the similarity between a pair of genes equals to their correlation if it is positive and is zero otherwise. The co-expression values were used to build the adjacency matrix. We employed a parameter sweep to construct various co-expression networks using different set of parameters. For this, a parallel computational framework (multithreading over different CPUs) was developed. The adjacency matrix was raised to different powers (6, 8, 10, 14, 18 and 22). The output matrices were then converted into topological overlap matrices (TOM) and TOM dissimilarity matrices (dissTOM = 1-TOM). The dissTOMs were used as the inputs for agglomerative hierarchical clustering using the average linkage method [40]. The dynamic tree-cutting algorithm was applied to define modules from the resulting clustering trees [15]. As part of the parameter sweep, different values were assigned for both the minimum module size parameter (minClusterSize = 15, 20 and 30) and the deepSplit parameter that controls how finely the branches should be split (deepSplit = 0, 2 and 4). For each module, a summary expression measure called the module eigengene (ME) was calculated as the first principal component of the scaled module expression profiles [12]. The modules with similar expression profiles were merged at different cut heights (CutHeight = 0.1, 0.15, 0.2, 0.25 and 0.3). Genes that did not assign into any specific module (the “grey” module) were not considered in the analysis. The co-expressed pairs (CPs) in each network were defined as any pair of genes assigned to the same module (Fig. 7a).

4.4. Knowledge databases and knowledge pairs

We annotated all the genes in the expression dataset using g:profiler R package. Then, the annotations from Reactome were extracted to define knowledge pairs (KPs): any pair of genes participating in at least one common pathway (Fig. 7b). The KPs extracted from this database were used to evaluate co-expressed pairs in different WGCNA networks constructed by various parameter settings.

4.5. Best WGCNA parameter setting, co-expressed and knowledge pairs overlaps

We scored the different set of parameters (Section 2.3) by comparing the observed co-expressed pairs with knowledge pairs. First, all possible pairs were counted (N(N-1)/2 where N denotes number of genes in co-expression networks having at least one annotation in Reactome) in each co-expression network. Then, those pairs were assigned to 4 different groups: **SM-SP** (the pairs of genes present in both CPs and KPs), **nSM-SP** (the pairs of genes present only in KPs), **SM-nSP** (the pairs of genes present only in CPs) and **nSM-nSP** (the pairs of genes present neither in CPs nor in KPs) (Suppl. Fig. 6). The enrichment factor, defined as the ratio between overlapping and nonoverlapping pairs (SM-SP × nSM-nSP / nSM-SP × SM-nSP), was used to evaluate and rank different sets of parameters (Fig. 7c).

4.6. Module-trait association

We fitted linear-mixed models to the module's eigengenes (eigenvectors) using the lme function (nlme R package: version 3.1–131 [41]) to determine modules associated with either deltoid or biceps. These models included age and muscle type as fixed effects and individual as a random effect. We performed the enrichment analysis for each module using the g:ProfileR R package. If multiple probes of a single gene were assigned to the same module, we considered that gene only once for the enrichment analysis.

4.7. The gene set connectivity analysis

The topological overlap of genes in an a priori defined gene set were

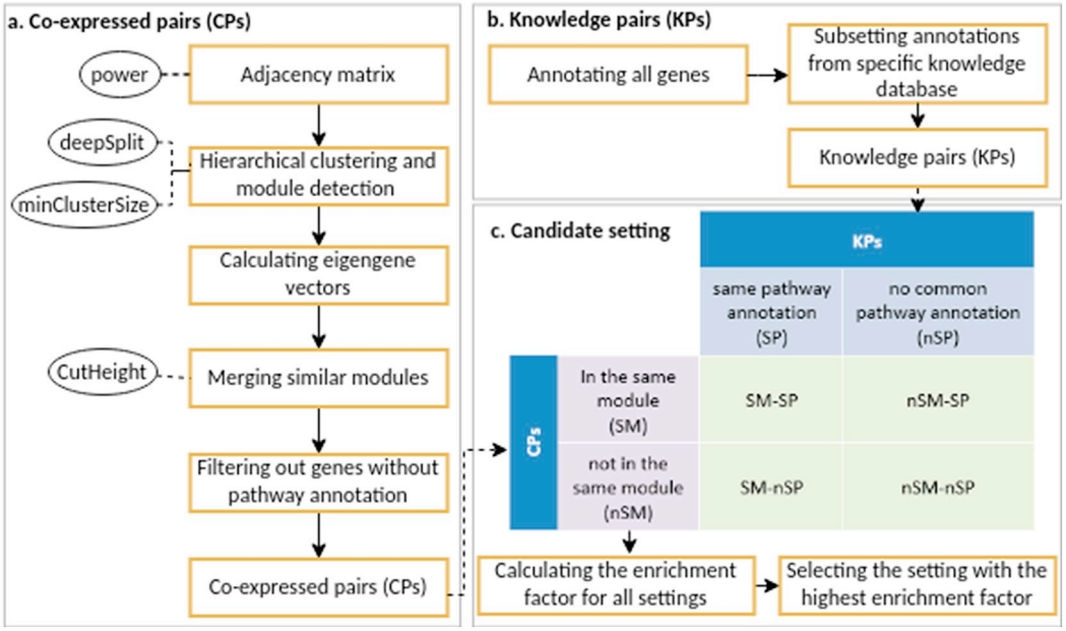

Fig. 7. Evaluating the gene co-expression networks using a knowledge database. a) Constructing various co-expression networks with different sets of parameters from WGCNA and defining co-expressed pairs; b) Defining knowledge pairs; c) Evaluating and ranking the best set of parameters based on consistency with knowledge database.

extracted from topological overlap matrices in WGCNA to perform gene set connectivity analysis. The subnetworks were exported and visualized in Cytoscape: version 3.7.1 [42]. The significance of node connectivity in the subnetworks were evaluated by a bootstrapping approach. In summary, 1000 subnetworks with the same number of nodes as in the original gene set were sampled. For each random subnetwork, the mean of topological overlap of all the nodes was calculated. The mean topological overlap of genes in the subnetwork was compared with the mean in random subnetworks. In summary, one-sided empirical *p*-value was computed as the proportion of the subnetworks with higher mean topological overlap compared to mean topological overlap of genes in our gene set. The assumption was that the empirical distribution created by sampling is under null meaning no strong topological overlaps between nodes were expected. Finally, the empirical p-value was calculated by **x/n** where x denotes the number of times that mean topological overlap of a random subnetwork is bigger than gene set topological overlap and n is number of random subnetworks. We, also, corrected for finite sampling bias by using **(x + 1)/(n + 1)**.

Suppl. Fig. 1 summarizes the data- and hypothesis-driven approaches used in our bioinformatics framework. *All scripts are publicly available on GitHub:* github.com/tabbassidaloii/AnalysisFrameworkSimilarTissues.

Authors contribution

TAD: Conceptualization, Formal analysis, Writing - original draft. PACH: Supervision, Conceptualization. PACH, VR and HEK: Writing - review & editing. All authors have read and approved the final version of this manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

This study was funded by the Netherlands Organization for Scientific Research (NWO), under research program VIDI, project 'similar but not the same', number 917.164.90.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2020.05.026.

References

[1] E. Pierson, et al., Sharing and specificity of co-expression networks across 35 human tissues, PLoS Comput. Biol. 11 (5) (2015) e1004220.
[2] A.R. Sonawane, et al., Understanding tissue-specific gene regulation, Cell Rep. 21 (4) (2017) 1077–1088.
[3] F. Aguet, et al., The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues, (2019), p. 787903.
[4] M. Uhlén, et al., Tissue-based Map of the Human Proteome, 347(6220) (2015), p. 1260419.
[5] G.T. Consortium, The genotype-tissue expression (GTEx) project, Nat. Genet. 45 (6) (2013) 580–585.
[6] C. Wei, J. Li, R.E. Bumgarner, Sample size for detecting differentially expressed genes in microarray experiments, BMC Genomics 5 (2004) 87.
[7] M.B. Eisen, et al., Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U. S. A. 95 (25) (1998) 14863–14868.
[8] A. de la Fuente, From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases, Trends Genet. 26 (7) (2010) 326–333.
[9] K. Eren, et al., A comparative analysis of biclustering algorithms for gene expression data, Brief. Bioinform. 14 (3) (2013) 279–292.
[10] A. Sturn, J. Quackenbush, Z. Trajanoski, Genesis: cluster analysis of microarray data, Bioinformatics 18 (1) (2002) 207–208.
[11] L.J. Heyer, S. Kruglyak, S. Yooseph, Exploring expression data: identification and analysis of coexpressed genes, Genome Res. 9 (11) (1999) 1106–1115.
[12] B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol. 4 (2005) 17.
[13] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinformatics 9 (2008) 559.
[14] A.L. Barabasi, E. Bonabeau, Scale-free networks, Sci. Am. 288 (5) (2003) 60–69.
[15] P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R, Bioinformatics 24 (5) (2008) 719–720.
[16] Y. Blum, M. Houee-Bigot, D. Causeur, Sparse factor model for co-expression networks with an application using prior biological knowledge, Stat. Appl. Genet. Mol. Biol. 15 (3) (2016) 253–272.
[17] P. Reshetova, et al., Use of prior knowledge for the analysis of high-throughput transcriptomics and metabolomics data, BMC Syst. Biol. 8 (Suppl. 2) (2014) S2.
[18] Z. Wang, et al., Incorporating prior knowledge into Gene Network Study, Bioinformatics 29 (20) (2013) 2633–2640.
[19] F. Rahimov, et al., Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers, Proc. Natl. Acad. Sci. U. S. A. 109 (40) (2012) 16234–16239.
[20] S.L. Peterson, G.M. Rayan, Shoulder and upper arm muscle architecture, J. Hand.

Surg. [Am.] 36 (5) (2011) 881–889.
[21] P. Parsana, et al., Addressing confounding artifacts in reconstruction of gene co-expression networks, Genome Biol. 20 (1) (2019) 94.
[22] J. Somekh, S.S. Shen-Orr, I.S. Kohane, Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset, BMC Bioinformatics 20 (1) (2019) 268.
[23] P. Mishra, et al., Mitochondrial dynamics is a distinguishing feature of skeletal muscle fiber types and regulates organellar compartmentalization, Cell Metab. 22 (6) (2015) 1033–1044.
[24] L.R. Smith, G. Meyer, R.L. Lieber, Systems analysis of biological networks in skeletal muscle function, Wiley Interdiscip. Rev. Syst. Biol. Med. 5 (1) (2013) 55–71.
[25] M. Wang, et al., Myostatin facilitates slow and inhibits fast myosin heavy chain expression during myogenic differentiation, Biochem. Biophys. Res. Commun. 426 (1) (2012) 83–88.
[26] A.M. Yip, S. Horvath, Gene network interconnectedness and the generalized topological overlap measure, BMC Bioinformatics 8 (2007) 22.
[27] P.S.T. Russo, et al., CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses, BMC Bioinformatics 19 (1) (2018) 56.
[28] D. Kostka, R. Spang, Finding disease specific alterations in the co-expression of genes, Bioinformatics 20 (2004) 194–199.
[29] S. van Dam, et al., Gene co-expression analysis for functional classification and gene-disease predictions, Brief. Bioinform. 19 (4) (2018) 575–592.
[30] Y. Lai, et al., A statistical method for identifying differential gene-gene co-expression patterns, Bioinformatics 20 (17) (2004) 3146–3155.
[31] J.K. Choi, et al., Differential coexpression analysis using microarray data and its application to human cancer, Bioinformatics 21 (24) (2005) 4348–4355.
[32] S. Ballouz, W. Verleyen, J. Gillis, Guidance for RNA-seq co-expression network construction and analysis: safety in numbers, Bioinformatics 31 (13) (2015) 2123–2130.
[33] G.E. Gibson, K.F. Sheu, J.P. Blass, Abnormalities of mitochondrial enzymes in Alzheimer disease, J. Neural Transm. (Vienna) 105 (8–9) (1998) 855–870.
[34] G.E. Gibson, et al., The alpha-ketoglutarate-dehydrogenase complex: a mediator between mitochondria and oxidative stress in neurodegeneration, Mol. Neurobiol. 31 (1–3) (2005) 43–63.
[35] N.M. Anderson, et al., The emerging role and targetability of the TCA cycle in cancer metabolism, Protein Cell 9 (2) (2018) 216–237.
[36] S. Durinck, et al., Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt, Nat. Protoc. 4 (8) (2009) 1184–1191.
[37] M.C. Oldham, P. Langfelder, S. Horvath, Network methods for describing sample relationships in genomic datasets: application to Huntington's disease, BMC Syst. Biol. 6 (2012) 63.
[38] M.E. Ritchie, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.
[39] J. Reimand, et al., g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments, Nucleic Acids Res. 35 (2007) W193–W200 (Web Server issue).
[40] P. Langfelder, S. Horvath, Fast R functions for robust correlations and hierarchical clustering, J. Stat. Softw. 46 (11) (2012).
[41] J. Pinheiro, et al., nlme: Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-131, (2017).
[42] M. Kohl, S. Wiese, B. Warscheid, Cytoscape: software for visualization and analysis of biological networks, Methods Mol. Biol. 696 (2011) 291–303.