Genome-wide co-expression distributions as a metric to prioritize genes of functional importance

Pâmela A. Alexandre1*, Nicholas J. Hudson2, Sigrid A. Lehnert1, Marina R.S. Fortes3, Marina Naval-Sánchez4, Loan T. Nguyen5, Laercio R. Porto-Neto1 and Antonio Reverter1

1CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia;
2School of Agriculture and Food Sciences, The University of Queensland, QLD 4072, Australia.
3School of Chemistry and Molecular Biosciences, The University of Queensland, QLD 4072, Australia.
4Institute for Molecular Bioscience, The University of Queensland, St Lucia, Queensland 4067, Australia;
5Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, Queensland 4067, Australia;

*Corresponding author

E-mail: pamela.alexandre@csiro.au (PAA)
Abstract

Genome-wide gene expression is routinely used as a tool to gain a systems-level understanding of complex, biological processes. Numerical approaches that have been used to highlight influential genes include abundance, differential expression, differential variation, network connectivity and differential connectivity. Network connectivity tends to be built on a small subset of extremely high co-expression signals that are deemed significant, but this overlooks the vast majority of pairwise signals. Here, we aimed to assess a complementary strategy, namely whether the entire shape of the distribution of genome-wide co-expression values contains a meaningful biological signal that has hitherto remained hidden from view.

We have developed a computational pipeline to assign one of 8 distributions (including normal, skewed, bimodal, kurtotic, inverted) to every gene. We then used a hypergeometric enrichment process to determine if particular genes (regulators versus non-regulators) and properties (differentially expressed or not) tend to be associated with particular distributions greater than would be expected by chance. Examination of several distinct data sets spanning 4 species indicates that there is indeed an additional biological signal present in the genome-wide distribution of co-expression values which would be overlooked by currently adopted approaches.
High-throughput technologies, such as RNA-Seq, enables access to a vast amount of data. Here, we describe a new approach to interrogate these data and extract further information to help researchers to understand complex phenotypes. Our method is based on gene-level co-expression distributions which were compared to eight possible template shapes to group genes with similar behaviours. The method was tested using five different datasets and the consistency of the results indicate it can be used as a complementary strategy to analyse transcriptomic data.
Introduction

The uncovering of genetic architecture behind complex phenotypes involves analysing a large variety of genes that interact with each other, and with other molecules, to respond to environmental stimuli [1]. Therefore, gene co-expression studies are becoming increasingly popular in the quest of going beyond differential expression (DE) and recovering more of the functional information from relevant tissues [2]. A gene co-expression study requires the computation of the co-expression correlation coefficient between a given gene and all the other genes under scrutiny, potentially numbered into the thousands for a genome-wide investigation. The result is often a large and tangled network of nodes (genes) connected by edges (significant correlations) from which one expects to extract biological insights and identify regulatory genes. However, defining which gene-level features are relevant to the biological question is not a simple task and main strategies to prioritise genes include focusing on hub (highly connected) genes, using the literature/biological criteria, overlaying other information onto the network such as patterns of differential expression and even exploiting variation in expression around the mean [3]. There is merit in developing additional data driven metrics to dissect co-expression networks and prioritize genes.

One system-wide topological feature of those co-expression networks is a structure comprising many nodes with few connections and few central nodes (hubs) with many connections, in a non-random fashion characterized by a scale-free power-law distribution [4]. The implication is that, within the network, different genes present different “behaviours” (i.e. different abilities to influence other molecules in the network), represented by the strength of their correlation coefficients. We have previously concluded that mRNA present in highly inter-connected modules encode proteins collectively involved in tightly regulated processes such as ribosomal biogenesis (Hudson et al., 2009). However, the typical approach of reverse engineering a co-expression network built from significant pairwise correlations actually
reflects only a tiny proportion of the total gene-gene interactions potentially at play across the whole system.

With this in mind, we propose that individual genes may possess diagnostic genome-wide distributions of co-expression values that may be similar or different to what one observes when examining all pairs. Furthermore, these individual distributions may respond to environmental condition and/or physiological state, producing different distributions in different biological circumstances. Hudson et al. (2009) demonstrated that these differences certainly exist at a whole system level by plotting frequency distributions of all pairwise correlation coefficients for six transcriptional landscapes of bovine skeletal muscle, considering different breeds, nutrition and physiological states, each one yielding a distinct distribution. Similarly, Remondini et al. (2005) explored the dynamic properties of all pairwise gene correlation distributions, such as skewness, using a time series expression data of rat fibroblast cell line expressing a conditional Myc-estrogen receptor oncoprotein. Their analysis was able to identify the cascade of c-myc-activated genes within the network. Therefore, current evidence supports the concept that correlation dynamics in co-expression networks have biological meaning.

Given previous studies have explored co-expression distributions and differential co-expression at the level of all gene-pairs, the distribution of correlation coefficients from a single gene across all its competing partners has been neglected. In the approach described here, we aim to determine if those genome-wide patterns are the source of untapped biological information. Based on a template of eight possible distribution shapes, from those likely to arise by chance alone (null distributions) to those apparently subject to some poorly understood external forcing (skewed and bimodal), we evaluated datasets of gene co-expression from human, cattle, duck and fly to determine if correlation distributions can be used as a metric to
prioritize genes of functional importance. We anticipate this new metric improving our ability to extract further meaningful information from co-expression networks.

Methods

The central idea of the methodology developed here was to group individual genes based on them sharing the distribution of genome-wide correlation coefficients. For that purpose, eight shapes were used as templates (Fig 1), varying between unimodal, bimodal, skewed or symmetrical, representing different proportions of positive and negative correlations. Shapes 1 and 2 will be often referred to as null distributions; Shapes 3 and 7 as positively skewed distributions; and Shapes 4 and 8 as negatively skewed distributions.

**Fig 1. Distribution shapes template.** Individual genes were assigned to one of these eight distributions shapes based on their correlation coefficients to all other expressed genes. Distributions were based on the proportion of correlations falling in each of the eight 0.25-bins of the plots.
Starting with a log2 normalized expression matrix, the Pearson correlation coefficient was computed for each gene pair across the samples. For each gene, the number of correlations fell within the eight 0.25-bins of the distributions was recorded (Fig 1) and summary statistics were calculated (i.e.; mean, SD, skewness and kurtosis). Then, based on the proportion of correlations falling in each bin and the summary statistics, and using a naïve chi-square classifier, the Euclidian distance between each gene distribution to each of the eight template distribution shapes was computed. We did this by comparing the observed values for a given gene to the expected values for each template shape listed in Table 1. In algebraical terms, the distance of the \(i\)-th gene to the \(j\)-th template shape \(D_{i,j}\) was computed as follows:

\[
D_{i,j} = \sqrt{\sum_{k=1}^{8} \left( \frac{O_{BP_{i,k}} - E_{BP_{j,k}}}{O_{BP_{i,k}} + E_{BP_{j,k}}} \right)^2 + \left( \frac{O_{MN_{i}} - E_{MN_{j}} - E_{MN_{j}}}{} \right)^2 + \left( \frac{O_{SD_{i}} - E_{SD_{j}}}{O_{SD_{i}} + E_{SD_{j}}} \right)^2 + \left( \frac{O_{SK_{i}} - E_{SK_{j}}}{O_{SK_{i}} + E_{SK_{j}}} \right)^2 + \left( \frac{O_{KU_{i}} - E_{KU_{j}}}{O_{KU_{i}} + E_{KU_{j}}} \right)^2 }^{1/2}
\]

Where subscripts \(i, j\) and \(k\) indicate gene, template distribution shape and 0.25-bin within shape, respectively; \(O_{BP_{i,k}}\) is the observed bin proportion of the \(i\)-th gene in the \(k\)-th bin, that is the proportion of all the co-expression correlation coefficients from the \(i\)-th gene that fall within the \(k\)-th 0.25 bin; \(E_{BP_{j,k}}\) is the expected bin proportion of the \(k\)-th bin in the \(j\)-th template distribution shape, for instance the expected proportion for the 1\(^{st}\) bin in the 1\(^{st}\), 2\(^{nd}\) and 3\(^{rd}\) template shapes are 5%, 2% and 15%, respectively; \(O_{MN_{i}}\) is the observed mean of all the co-expression correlation coefficients from the \(i\)-th gene; \(E_{MN_{j}}\) is the expected mean of all the co-expression correlation coefficients in the \(j\)-th template distribution shape, for instance the expected mean is zero for symmetric distributions such as Shapes 1, 2, 5 and 6; \(O_{SD_{i}}\) is the observed standard deviation of all the co-expression correlation coefficients from the \(i\)-th gene; \(E_{SD_{j}}\) is the expected standard deviation of all the co-expression correlation coefficients in the \(j\)-th template distribution shape; \(O_{SK_{i}}\) is the observed skewness of all the co-expression
correlation coefficients from the $i$-th gene; $ESK_j$ is the expected skewness of all the co-expression correlation coefficients in $j$-th template distribution shape; $OKU_i$ is the observed kurtosis of all the co-expression correlation coefficients from the $i$-th gene; $EKU_j$ is the expected kurtosis of all the co-expression correlation coefficients in $j$-th template distribution shape.

Using linear mapping, distances were transformed into similarities ($S_{i,j}$) of a given gene to belong to each of the template distribution shape as follows:

$$S_{i,j} = 1.0 - \frac{D_{i,j} - \text{Min}\{D_{i,j}\}}{\text{Max}\{D_{i,j}\} - \text{Min}\{D_{i,j}\}}$$  \hspace{1cm} (2)

Where $\text{Min}\{D_{i,j}\}$ and $\text{Max}\{D_{i,j}\}$ are the minimum and maximum $D_{i,j}$, respectively.

Finally, similarities were transformed to probabilities ($P_{i,j}$) of a given gene to belong to each template distribution shape so that the sum of all probabilities for a given gene added to one:

$$P_{i,j} = \frac{S_{i,j}}{\sum_{j=1}^{8} S_{i,j}}$$  \hspace{1cm} (3)

A gene was assigned to the $j$-th distribution template shape if its $P_{i,j}$ was the largest across all $j$’s. The analyses were implemented using shell scripts to manipulate the data and our own programs in FORTRAN for the fast computation and assignment of within-gene co-expression distributions to template shapes. Scripts and source code are available from the corresponding author upon request.
Table 1. Expected bin proportions, mean, standard deviation (SD), skewness and kurtosis for the 8 shapes.

| Expected 0.25-Bin Proportions, % | Mean | SD | Skewness | Kurtosis |
|----------------------------------|------|----|----------|----------|
|                                  | Bin 1 | Bin 2 | Bin 3 | Bin 4 | Bin 5 | Bin 6 | Bin 7 | Bin 8 |          |          |          |
| Shape 1                          | 5     | 10   | 15    | 20    | 20    | 15    | 10    | 5     | 0        | 0.456    | 0        |
| Shape 2                          | 2     | 5    | 15    | 28    | 28    | 15    | 5     | 2     | 0        | 0.354    | 0.065 |
| Shape 3                          | 15    | 30   | 25    | 19    | 9     | 3     | 2     | 1     | -0.387   | 0.382    | 0.857   |
| Shape 4                          | 1     | 2    | 3     | 9     | 15    | 25    | 30    | 15    | 0.387    | 0.382    | -0.857  |
| Shape 5                          | 28    | 15   | 5     | 2     | 2     | 5     | 15    | 28    | 0        | 0.752    | -1.771 |
| Shape 6                          | 5     | 25   | 15    | 5     | 5     | 15    | 25    | 5     | 0        | 0.566    | -1.535 |
| Shape 7                          | 5     | 30   | 20    | 10    | 5     | 10    | 15    | 5     | -0.137   | 0.532    | 0.5     |
| Shape 8                          | 5     | 15   | 10    | 5     | 10    | 20    | 30    | 5     | 0.137    | 0.532    | -0.5    |

To gain insight into the biological drivers of gene co-expression distribution, genes were categorized according to their reported biological relevance (e.g.; DE, transcription factors/regulators). A hypergeometric test was applied to identify enriched or depleted categories in each shape, using the function “phyper” in the R environment [7]. Therefore, we compared the within-shape proportion of genes in a given category to the proportion of overall genes in that category. To test the association between the categories and the types of distribution, a chi-square test of independence was applied. Results were considered significant if P-value ≤ 0.05.

To investigate the relationship between number of connections per gene (degree) and distribution shapes, we used the same datasets as input to a PCIT analysis [8]. The PCIT algorithm combines partial correlation coefficients with information theory to determine significant correlations between two genes after accounting for all the other genes under scrutiny. We then computed the number of significant correlations of each gene to evaluate: 1) if there was a relationship between distribution shape and average number of connections per gene using a one-way ANOVA; 2) if the top and the bottom 5% genes based on degree would be enriched in specific shapes.
Finally, although it is not in the scope of this work to discuss specific biological aspects of each dataset, we were interested in evaluating if the different distribution shapes would capture some general biological process. For that purpose, we used the online platform GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion tool - http://cbl-gorilla.cs.technion.ac.il/) to tested a list of genes falling in each shape for each dataset against all genes considered for analysis in that dataset. GOrilla uses hypergeometric test and multiple test correction (FDR – false discovery rate) to determine significant enriched gene ontology terms (Padj<0.05). For this analysis, we focused on cell components.

Datasets used for experimental evaluation

We applied our methodology to publicly available datasets comprised of RNA sequencing (RNA-seq) data from four organisms (cattle, duck, human and fly) in different physiological conditions, from different tissues. These four datasets focused on different biological questions and were chosen so that we could better explore the utility of new metric. Tables 2 summarizes the characteristics of each dataset.

| Dataset                      | Number of tissues | Number of libraries | Number of genes | Gene filtering criteria                                                                 |
|------------------------------|-------------------|---------------------|-----------------|----------------------------------------------------------------------------------------|
| Cattle feed efficiency       | 5                 | 86                  | 11,662          | Average log2FPKM > 1 across all tissues                                                  |
| Cattle puberty               | 5                 | 60                  | 16,978          | Average FPKM ≥ 0.2 in at least 1 tissue                                                  |
| Drosophila embryogenesis     | 1 (14 time points)| 68 averaged by time point | 7,771         | More than 0 counts in at least 20 samples and more than 100 counts on average           |
| Duck preadipocyte            | 1 (6 time points) | 36 averaged by time point | 13,322         | FPKM > 0 in all samples                                                                  |
| Human                        | 1                 | 15                  | 19,206          | TPM > 0 in all samples                                                                  |

FPKM - fragments per kilobase of gene per million mapped reads
TPM – transcripts per million reads
The cattle feed efficiency dataset is publicly available in the European Nucleotide Archive under the study ID PRJEB27337 (https://www.ebi.ac.uk/ena/data/view/PRJEB27337). Further information about animal management and phenotypic measurements can be found in Alexandre et al. (2015), and information about RNA-Seq libraries and initial processing can be found in Alexandre et al. (2019). Briefly, RNA was extracted from samples of the adrenal gland, hypothalamus, liver, skeletal muscle and pituitary from nine Nellore bulls from each extreme of feed efficiency (evaluated by residual feed intake, Koch et al., 1963). Eighty-six RNA libraries were sequenced on Illumina HiSeq2500 equipment (2x100 pb), reads were aligned to the bovine reference genome (UMD3.1) and expression values were estimated as log2 FPKM (fragments per kilobase of gene per million mapped reads). For this study, we used expression data of 11,662 genes with mean expression > 1 log2(FPKM) across all tissues. Moreover, we classified genes as DE (382 genes) according to Alexandre et al. (2019) and as regulators (REG, 1,072 genes) based on the transcription factors and co-factors indicated in the Animal Transcription Factor Database 2.0 [14].

The cattle puberty dataset is available at EMBL-EBI BioSamples repository (www.ebi.ac.uk/biosamples) under the submission identifiers GSB-113 and GSB-8708. It is comprised of five tissues (hypothalamus, pituitary, ovary, uterus and liver) from 6 pre- and 6 post-pubertal Brahman (Bos indicus) heifers, totalizing 60 samples. More details about animal management and data generation can be found in Fortes et al. (2016), Nguyen et al. (2017), Fortes et al. (2018) and Nguyen et al. (2018). In short, libraries were paired-end sequenced using an Illumina HiSeq200 equipment and reads were aligned to the bovine reference genome (UMD3.1). A total of 16,978 genes that presented average FPKM ≥ 0.2 in at least 1 tissue were used for the analysis. Genes were classified as DE (2,335 genes) based on the four
aforementioned works and as REG (1,584 genes) also based on the Animal Transcription Factor Database 2.0 [14].

The duck subcutaneous preadipocyte differentiation dataset is available in the Additional File 3 of the source article, where more information about cell culture and RNA-Seq libraries can be found [9]. Briefly, mRNA of six biological replicates of preadipocytes cultured in differentiation medium were collected at -48h, 0h, 12h, 24h, 48h, and 72h and sequenced on an Illumina HiSeq X-10 with PE150. The paired-end reads were aligned to the duck reference genome (Anas_platyrhynchos.BGI_duck_1.0) and gene expression was estimated as FPKM. We kept for analysis only genes presenting counts in all samples (13,322 genes) which were then log2-transformed. Genes were classified as DE (3,321 genes) based on the list provided by the authors in additional file 4 (Wang et al., 2019) and as REG (675 genes) based on the Animal Transcription Factor Database 3.0 [19].

The drosophila embryogenesis dataset is available in NCBI’s Gene Expression Omnibus and is accessible through GEO Series accession numbers GSE121160 and GSE121161. Further information about sample collection and RNA libraries generation can be found in Becker et al. (2018). The mentioned study generated a paired transcriptome/proteome time course dataset with 14 time points during Drosophila melanogaster embryogenesis, i.e. 0, 1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18, 20 hours. The 68 embryo RNA libraries were pooled in equimolar ratio and sequenced on 8 lanes of a HiSeq2500 (1× 51 cycles plus 7 cycles for the index read). Reads were mapped to the BDGP6 fly reference genome and gene expression was estimated as read counts. Data was averaged within each time point and log2 transformed prior to implementation. Genes were classified in the categories defined by Becker et al. (2018), based on pairwise comparison of genes up or down regulated (relative to the first time point - 0h) in mRNA and protein data, resulting in groups named to represent the gene status on mRNA/protein data: up/up (511 genes), down/up (1770 genes), down/down (1048 genes) and
up/down (326 genes). Genes were also classified as regulators (791 genes) based on the list provided by Rhee et al. (2014) consisting of essential genes involved in replication and transcription, splicing, DNA repair and cell division.

The human dataset was downloaded from The Genotype-Tissue Expression (GTEx) Project V8 (https://www.gtexportal.org/) which contain data of non-diseased tissue sites across nearly 1000 individuals [10]. We used liver RNA-Seq data from 15 individuals, provided as TPM counts. We kept for analysis only genes presenting non-zero counts in all samples, which were then log2-transformed. Genes were classified as REG (1,153 genes) and tissue enriched (TE, 231 genes) according to information provided by The Human Protein Atlas [22]. Moreover, genes were defined as DE (793 genes) if they were identified by Crow et al. (2019) as having high probability (>0.95) of being DE in any experiment, based on a meta-analysis of over 600 differential expression studies.

Results and discussion

Overall co-expression distribution

Although our methodology aims to evaluate co-expression distributions at a gene-level, we did calculate all the correlations across all genes for each of the five RNA-Seq datasets we evaluated: Cattle Feed Efficiency, Cattle Puberty, Drosophila Embryogenesis, Duck Preadipocyte and Human. The overall frequency distributions in each dataset that give us an overview of co-expression patterns can be found in Fig 2. The higher number of positive correlations, even though discrete in some datasets, was already expected and documented on previous research [5,24]. The number of positive correlations is specially elevated in the Cattle Feed Efficiency dataset. This can be due to the high inflammatory response found in the liver of those animals, which remain strong even when analysing all five tissues together and results in a set of highly positively co-expressed genes [12]. On the other side, among the 5 tissues
analysed in the Cattle Puberty Dataset, only ovary and uterus showed great differences between pre- and post-puberty and the effect of the coordinated mechanisms regulating those differences are not so strong in the overall frequency distribution. Both Drosophila Embryogenesis and Duck Preadipocyte datasets representing developmental processes through time-series data present similar shapes. They show strong positive and negative correlation as a consequence of the tightly coordinated processes the datasets represent. The Human dataset is the one with the frequency distribution closer to a bell shape, but still, the higher presence of positive correlations can be observed.

**Fig 2. Frequency distributions of all co-expression correlation coefficients in each of the five RNA-Sequence datasets.** Datasets include: 1) Cattle Feed Efficiency (11,662 genes and 67,995,291 correlations); 2) Cattle Puberty (16,978 genes and 144,117,753 correlations); 3) Drosophila Embryogenesis (7,771 genes and 30,190,335 correlations); 4) Duck Preadipocyte
(13,322 genes and 88,731,181 correlations); and 5) Human (19,311 genes and 186,447,705 correlations).

**Co-expression distribution in datasets with contrasting phenotypes**

Considering co-expression distribution at gene-level, the proportion of genes falling in each distribution shape considering the two cattle datasets with contrasting phenotypes can be found in Fig 3. When comparing the proportion of each category of genes with that of all genes within individual distributions, for the cattle feed efficiency dataset we identified an over-representation of REG in negatively skewed distributions (i.e., with an overabundance of positive co-expressions - Shapes 4 and 8) and an under-representation of those genes in null distributions (Shapes 1 and 2) and in a positively skewed distribution (Shape 3). On the other hand, DE genes were under-represented in shape 4 and over-represented in null distributions (Shapes 1 and 2) and in bimodal skewed distributions (Shapes 7 and 8). The exact number of genes falling in each distribution shape and the significance of enrichment analysis can be found in S1 Table.

Similarly, the results using the cattle puberty dataset showed an over-representation of REG in a bimodal negative skewed distribution (Shape 8) and an under-representation of those genes in null distributions (Shapes 1 and 2) as well as in a positively skewed distribution (Shape 7). Differentially expressed genes also behave similarly to the previous analysis, being over-represented in null distributions (Shapes 1 and 2) and in a bimodal positively skewed distribution (Shape 7).
Fig 3. Proportion of genes falling in each template shape in datasets with contrasting phenotypes. The proportion of genes classified as differentially expressed (DE), regulator (REG) or both (DE-REG) are compared to the overall (All) proportion of genes within each shape.

Considering that several genes are expected to present different behaviour according to the contrasting condition tested, we applied our pipeline again using the two cattle datasets split by phenotype. We then identified genes that were assigned to different shapes by comparing high to low feed efficiency and pre to post puberty. From the 11,662 genes tested using the cattle feed efficiency dataset, 2,032 genes were assigned to different shapes depending on the condition, being 133 DE and 158 REG. Likewise, 2,740 out of 16,978 genes from the cattle puberty dataset were assigned to different shapes depending on the phenotype, being 620 DE and 216 REG. The shift in the proportion of each class of gene depending on the phenotypic condition can be seen in Fig 4.
Fig 4. Proportion of genes falling in each template shape in contrasting phenotypic conditions within cattle datasets. The proportion of genes classified as differentially expressed (DE), regulator (REG) or both (DE-REG) are compared to the overall (All) proportion of genes within each shape.

The enrichment analysis for DE and REG among the genes changing behaviour according to the phenotype were again concordant, with REG being under-represented and DE being over-represented (P<0.01) in both datasets. Most REG genes, such as transcription factors and co-factors, are expected to be central genes in the network, presenting co-expression with many genes as a consequence of their regulatory role in highly coordinated biological processes. This can be one reason why they are not particularly enriched among the genes changing co-expression distribution between conditions. Nevertheless, the few REG that do change co-expression distribution between conditions are defiantly worth to be further explored.
as potential key regulators. On the other hand, a gene identified as DE, being either central to a specific function or the final product of an altered pathway, are more prone to be condition-specific/enriched, not only regarding their expression level but also regarding their relationship with other genes. By identifying DE genes of central or peripheral function in the network based on their enrichment in null or non-null distributions, and exploring changes in their behaviour, one can gain insight into the molecular dynamics behind the phenotype regulation.

In general, genes changing behaviour between conditions might be DE genes that fail the significance threshold in the differential expression analysis. They can also be genes that, although not differentially expressed between conditions, play different roles depending on the overall gene expression pattern, a feature already widely explored using differential connectivity measures [25]. The advantage of comparing distribution shapes is that it considers the direction (positive or negative) of the correlations and the proportion of the correlation falling in each bin of the distribution, corresponding to the correlation’s strength. One can explore all genes changing behaviour or focus on specific distribution changes, for instance, from unimodal to bimodal or from symmetric to skewed. In our datasets, 38% of genes changing shapes in Feed Efficiency and 77% in Puberty represents a change from null to non-null distributions or vice-versa (which, in this case, also represents changes from symmetrical to skewed). Respectively, 37% and 76% genes changing shapes are moving from unimodal to bimodal distributions or vice-versa. Although percentages are similar, different sets of genes are selected depending on the criteria.

The enrichment analysis for DE and REG in each shape show similar results between phenotypic conditions (S2 and S3 Tables). Nevertheless, it is possible to observe a shift in the proportion of genes falling in each shape, particularly from bimodal to unimodal shapes, implicating in a loss/gain of genes presenting both positive and negative correlations at the same time.
Co-expression distribution in time series datasets

Both time series datasets, duck fat differentiation and drosophila embryogenesis, follow the same pattern identified in the cattle data sets (Fig 5), with REG being under-represented in a null distribution (Shape 1) and over-represented in a bimodal skewed distribution (Shape 7 or 8, S4 Table). Although in the Drosophila dataset the DE genes were subdivided into different classes according to their concordance between mRNA and protein expression data, it is possible to observe in both datasets the enrichment of DE genes being split between null and non-null distributions. Quite remarkably, considering non-null distribution in the Drosophila dataset, genes consistently down-regulated (down-down) are enriched in the negatively-skewed bimodal distribution (Shape 8) while genes consistently up-regulated (up/up) are enriched in the opposite shape - the positively-skewed bimodal distribution (Shape 7). Moreover, those two classes are under-represented in the symmetrical bimodal distribution (Shape 6) where up/down and down/up are both enriched.

Fig 5. Proportion of genes falling in each template shape in time series datasets. The proportion of genes classified as differentially expressed (DE), regulator (REG) or both (DE-REG) are compared to the overall (All) proportion of genes within each shape. In Drosophila dataset, DE genes were clustered into four groups (down/down, down/up, up/down and up/up) – refer to methods for more information.
Another curious observation is that, while both bovine datasets (representing contrasting phenotypes) present no genes in Shape 6, both time-series datasets not only present gene falling in this shape but also there is an enrichment of DE genes. It is important to notice at this point the impact of species, tissues and even filtering criteria in the proportion of genes assigned to each shape in each dataset. Although some patterns can be identified, each dataset has its idiosyncrasies and different aspects may be worth investigating.

Co-expression distribution in a physiological baseline dataset

In contrast to the other datasets, the Human dataset consists of several datapoints all representing a single “non-disease” baseline state and this fact is reflected in the results (Fig 6). The REG, found in other datasets to be over-represented in skewed distributions, are over-represented in a null distribution (Shape 2) and under-represented in skewed bimodal distributions (Shapes 7 and 8). In the previous datasets, DE genes already showed enrichment in null distributions, but they also showed enrichment in other non-null shapes. Without the effect of contrasting conditions, genes with high probability of being differentially expressed were enriched in null distributions (Shapes 1 and 2) and in Shape 3 but with very few genes falling in this last shape (43 genes – 0.2% of the total). Tissue enriched genes was the only category enriched in non-null distribution (Shape 7). That is probably because genes particular to tissue’s specific activities tend to be tightly correlated, as they need to be expressed in a coordinated pattern to keep tissues' functions despite physiological conditions. That behaviour can be clearly observed in studies that constructed co-expression networks using multiple tissue transcriptomic data, where tissue-specific genes push genes to cluster by tissue [12,26].
Fig 6. Proportion of genes falling in each template shape in human dataset. The proportion of genes classified as differentially expressed (DE), regulator (REG) or both (DE-REG) are compared to the overall (All) proportion of genes within each shape.

Relationship between gene categories and distribution shapes

Considering the similarities observed between the first three datasets and the fact that without contrasting phenotypes genes with regulatory potential appear enriched in null distributions (i.e., Shapes 1 and 2), we evaluated the dependence between DE or REG genes and null vs non-null (Shapes 3 to 8), unimodal (Shapes 1 to 4) vs bimodal (Shapes 5 to 8) and symmetric (Shapes 1, 2, 5, and 6) vs skewed (Shapes 3, 4, 7 and 8) distributions. Again, we found a consistent pattern among most of the datasets, with DE genes being found more frequently than expected (P<0.05) in null, unimodal and symmetrical distributions and REG being found more frequently than expected (P<0.05) in non-null, bimodal and skewed distributions (Fig 7). On the other hand, for the human dataset, both DE and REG were found more frequently than expected (P<0.05) in null, unimodal and symmetrical distributions.
### Figure 7. Number of genes assigned to different distribution shapes

More especially, the tables show the number of differentially expressed genes (DE) and regulator (REG) genes assigned to null (Shapes 1 and 2), unimodal (Shapes 1 to 4) and symmetric (Shapes 1, 2, 5 and 6) distributions. Blue cells represent the type of distribution in which DE or REG were found in higher number than expected by chance where variables were found to be dependent (P<0.05).

**Relationship between gene degree and distribution shapes**

Because genes that are expected to be highly connected to others (such and REG and TE) appear enriched in non-null distributions, we explored the relationship between number of significant correlations and distributions shapes. Table 3 shows that, for all datasets, there

| Cattle Feed Efficiency Dataset | Null | Unimodal | Symmetric |
|-----|------|---------|-----------|
| DE  | Yes  | 157     | 371       | 157       |
|     | No   | 2,671   | 11,063    | 2,671     |
| REG | Yes  | 180     | 1,046     | 180       |
|     | No   | 2,648   | 10,388    | 2,648     |

| Cattle Puberty Dataset      | Null | Unimodal | Symmetric |
|-----------------------------|------|---------|-----------|
| DE                          | Yes  | 1,463   | 1,463     | 1,463     |
|                             | No   | 5,855   | 5,855     | 5,855     |
| REG                         | Yes  | 567     | 568       | 567       |
|                             | No   | 6,751   | 6,754     | 6,751     |

| Drosophyla Embryogenesis Dataset | Null | Unimodal | Symmetric |
|----------------------------------|------|---------|-----------|
| DE                               | Yes  | 842     | 842       | 881       |
|                                  | No   | 689     | 699       | 3,040     |
| REG                              | Yes  | 111     | 117       | 578       |
|                                  | No   | 1,420   | 1,424     | 5,343     |

| Duck Preadipocyte Dataset       | Null | Unimodal | Symmetric |
|---------------------------------|------|---------|-----------|
| DE                              | Yes  | 636     | 636       | 984       |
|                                 | No   | 1,609   | 1,613     | 2,337     |
| REG                             | Yes  | 105     | 105       | 183       |
|                                 | No   | 2,140   | 2,144     | 3,487     |

| Human Non-diseased Dataset      | Null | Unimodal | Symmetric |
|---------------------------------|------|---------|-----------|
| DE                              | Yes  | 503     | 510       | 503       |
|                                 | No   | 8,443   | 8,487     | 8,443     |
| REG                             | Yes  | 618     | 621       | 618       |
|                                 | No   | 8,328   | 8,376     | 8,328     |
is a significant difference in the average number of connections per gene in each distribution shape.

Table 3. Average number of connections per gene (Log10) falling in each shape according to the dataset and P-value for the one-way ANOVA.

| Dataset                 | Average number of connections per gene (Log10) | P-value |
|-------------------------|-----------------------------------------------|--------|
|                         | Shape 1 | Shape 2 | Shape 3 | Shape 4 | Shape 5 | Shape 6 | Shape 7 | Shape 8 |
| Cattle feed efficiency  | 3.04    | 2.81    | 3.30    | 3.49    | -       | -       | 3.15    | 3.20    | <2E-16 |
| Cattle puberty          | 3.14    | 2.52    | 1.40    | 1.87    | -       | -       | 3.57    | 3.58    | <2E-16 |
| Drosophila embryogenesis| 2.58    | 2.26    | 2.39    | -       | -       | 3.03    | 2.92    | 2.87    | <2E-16 |
| Duck preadipocyte       | 2.33    | 2.06    | 1.82    | -       | -       | 2.60    | 2.60    | 2.66    | <2E-16 |
| Human                   | 2.60    | 2.34    | 2.46    | 2.40    | -       | -       | 3.17    | 3.44    | <2E-16 |

As anticipated, the bottom 5% genes regarding degree (less connected genes) were enriched in both null distributions (Shapes 1 and 2) for all datasets (S6 Table). The few enrichments out of those 2 shapes were due to the low number of genes assigned to that shape. On the other hand, bottom genes were depleted in non-null distributions, particularly in bimodal shapes. The top 5% genes (corresponding to genes with the higher number of significant correlations) were enriched in different shapes depending of the dataset, but exclusively in non-null distribution. Except for the Human dataset, those enrichments were found in shapes that were also enriched for functionally important categories of genes (DE or REG). This overlap, particularly between transcription factors and highly connected genes have been reported before [5]. For all datasets, top genes were depleted in null distributions. This results reinforce that genes related to tightly coordinated processes are more often found in non-null distribution.
The functional enrichment of genes falling in each shape, although not consistently among the different datasets, demonstrate strong biological signals resulting from our methodology (Table 4), with some of the enrichments as significant as q-value of 1.51E-153 (nucleoplasm). To the best of our knowledge, this is the first time it is demonstrated genes grouped according to the distribution of their co-expression correlations can capture specific gene functions. Here, we reported enrichment results at Cellular Component level, as it is not the focus of this work to discuss particular biological mechanisms behind each dataset. Nevertheless, we also have found shape-specific enrichment for Biological Process and Molecular Function in our test datasets (data not shown).

Table 4. Hypergeometric gene set enrichments for each distribution within each data set (FDR q-values) based on Cellular Component.

|                      | Feed efficiency | Puberty | Drosophila | Duck                              | Human                                      |
|----------------------|-----------------|---------|------------|-----------------------------------|--------------------------------------------|
| **Shape 1**          | Ribosomal unit  | Membrane part | Cytoplasmic part | Condensed chromatin outer | Plasma membrane part |
|                      | (7.84E-15)      | (9.17E-32) | (1.67E-8)   | chromatin outer kinetochore      | (1.95E-5)                                |
|                      | Mitochondrial part |         | Mitochondrion | (3.2E-3)                       |                                           |
|                      | (1.61E-14)      |          | (1.89E-4)   |                                   |                                           |
| **Shape 2**          | Mitochondrial part | Condensed | No functional | No functional | Immunoglobulin complex |
|                      | (3.82E-13)      | chromatin outer | enrichment | enrichment | (2.98E-21)                 |
|                      | ribosomal unit  | kinetochore  |                       |                       |                                           |
|                      | (1.09E-3)       | (3.91E-2)   |             |                       |                                           |
| **Shape 3**          | Contractile fibre | No functional | No functional | No functional | No functional |
|                      | (7.55E-24)      | enrichment | enrichment | enrichment |                             |
| **Shape 4**          | Nucleoplasm     | No functional | No functional | No functional | No functional |
|                      | (2.31E-27)      | enrichment | enrichment | enrichment |                             |
| **Shape 5**          | -               | -        | -          | -                                 |                                           |
| **Shape 6**          | -               | -        | Nuclear part | Proteasome complex | - |
|                      |                 |          | (7.64E-26) | (6.71E-3)                         |                                           |
| **Shape 7**          | No functional | Cytosolic part | Ribosomal subunit | Peroxisomal | - |
|                      | enrichment     | (9.75E-27)| (5.04E-16) | matrix (3.83E-8)                 |                                           |
| **Shape 8**          | Neuron projection | Nucleoplasm | Cytoplasm | Golgi apparatus | Intracellular organelle part |
|                      | (2.51E-2)       | (1.51E-153)| (1.45E-6) | (2.4E-2)                       | (1.29E-108)                         |

24
Conclusion

We started with the premise that in a co-expression network, different genes present different co-expression distributions and can be grouped according to those. Indeed, considering the five vastly different datasets we analyzed, genes were assigned consistently to pre-defined distribution shapes, regarding the enrichment of differentially expressed and regulatory genes, in situations involving contrasting phenotypes, time series or physiological baseline data.

Admittedly, there is some arbitrary subjectivity in the creation of the proposed 8 template shapes. However, we believe 8 is the minimum number of shapes required to capture in a balanced way the symmetry vs. skewness contrast in one dimension, and the uni- vs. multi-modality contrast in another dimension. Similarly, the use of more or less bins within distributions (eg., having 10 0.10-bins instead of 8 0.25-bins) is worthy of further research. Indeed, across the five datasets analyzed, no gene was allocated to Shape 5.

In conclusion, the results highlight that the distribution shape of correlation coefficients can be used as a novel metric to prioritise genes of functional importance and to further explore topological characteristics of gene networks. By considering that highly connected genes will be assigned to particular distribution shapes according to the experimental design underlying the gene co-expression networks, regulatory genes can even be identified in datasets that do not represent physiological contrasts or time series.

Acknowledgments

The authors thank Dr. Yutao Li and Dr. Sonja Dominik for reviewing the manuscript.
References

1. Swami M. Networking complex traits. Nat Rev Genet. 2009;10: 2566.
doi:10.1038/ng.332

2. Hudson NJ, Dalrymple BP, Reverter A. Beyond differential expression: the quest for
causal mutations and effector molecules. BMC Genomics. 2012. doi:10.1186/1471-
2164-13-356

3. Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, Silburn PA, et al.
Variance of gene expression identifies altered network constraints in neurological
disease. Gibson G, editor. PLoS Genet. 2011;7: e1002207.
doi:10.1371/journal.pgen.1002207

4. Barabási A-L, Oltvai ZN. Network biology: understanding the cell’s functional
organization. Nat Rev Genet. 2004;5: 101–13. doi:10.1038/nrg1272

5. Hudson NJ, Reverter A, Wang YH, Greenwood PL, Dalrymple BP. Inferring the
transcriptional landscape of bovine skeletal muscle by integrating co-expression
networks. PLoS One. 2009;4. doi:10.1371/journal.pone.0007249

6. Remondini D, O’Connell B, Intrator N, Sedivy JM, Neretti N, Castellani GC, et al.
Targeting c-Myc-activated genes with a correlation method: Detection of global
changes in large gene expression network dynamics. Proc Natl Acad Sci. 2005;102:
6902–6906. doi:10.1073/pnas.0502081102

7. R Core Team. R: A language and environment for statistical computing. R Foundation
for Statistical Computing. Vienna; 2019. Available: https://www.r-project.org/

8. Reverter A, Chan EKF. Combining partial correlation and an information theory
approach to the reversed engineering of gene co-expression networks. Bioinformatics.
2008;24: 2491–2497. doi:10.1093/bioinformatics/btn482

9. Wang Z, Yin Z-T, Zhang F, Li X-Q, Chen S-R, Yang N, et al. Dynamics of
transcriptome changes during subcutaneous preadipocyte differentiation in ducks.
BMC Genomics. 2019. doi:10.1186/s12864-019-6055-9

10. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-
Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653

11. Alexandre PA, Kogelman LJA, Santana MHA, Passarelli D, Pulz LH, Fantinato-Neto
P, et al. Liver transcriptomic networks reveal main biological processes associated
with feed efficiency in beef cattle. BMC Genomics. 2015;16. doi:10.1186/s12864-015-
2292-8

12. Alexandre PA, Naval-Sanchez M, Porto-Neto LR, Ferraz JBS, Reverter A, Fukumasu
H. Systems biology reveals NR2F6 and TGFBI as key regulators of feed efficiency in
beef cattle. Front Genet. 2019;10: 1–16. doi:10.3389/fgene.2019.00230

13. Koch RM, Swiger LA, Chambers D, Gregory KE. Efficiency of Feed Use in Beef
Cattle. J Anim Sci. 1963;22: 486–494.

14. Zhang HM, Liu T, Liu CJ, Song S, Zhang X, Liu W, et al. AnimalTFDB 2.0: A
resource for expression, prediction and functional study of animal transcription factors.
Nucleic Acids Res. 2015;43: D76–D81. doi:10.1093/nar/gku887
15. Nguyen LT, Reverter A, Cánovas A, Venus B, Anderson ST, Islas-Trejo A, et al. STAT6, PBX2, and PBRM1 Emerge as Predicted Regulators of 452 Differentially Expressed Genes Associated With Puberty in Brahman Heifers. Front Genet. 2018;9: 87. doi:10.3389/fgene.2018.00087

16. Fortes MRS, Zacchi LF, Nguyen LT, Raidan F, Weller MMDCA, Choo JJY, et al. Pre- and post-puberty expression of genes and proteins in the uterus of Bos indicus heifers: the luteal phase effect post-puberty. Anim Genet. 2018;49: 539–549. doi:10.1111/age.12721

17. Nguyen LT, Reverter A, Cánovas A, Venus B, Islas-Trejo A, Porto-Neto LR, et al. Global differential gene expression in the pituitary gland and the ovaries of pre- and postpubertal Brahman heifers. J Anim Sci. 2017;95: 599–615. doi:10.2527/jas.2016.0921

18. Fortes MRS, Nguyen LT, Weller MMDCA, Cánovas A, Islas-Trejo A, Porto-Neto LR, et al. Transcriptome analyses identify five transcription factors differentially expressed in the hypothalamus of post- versus prepubertal Brahman heifers. J Anim Sci. 2016;94: 3693–3702. doi:10.2527/jas.2016-0471

19. Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019. doi:10.1093/nar/gky822

20. Becker K, Bluhm A, Casas-Vila N, Dinges N, Dejung M, Sayols S, et al. Quantifying post-transcriptional regulation in the development of Drosophila melanogaster. Nat Commun. 2018;9: 4970. doi:10.1038/s41467-018-07455-9

21. Rhee DY, Cho DY, Zhai B, Slattery M, Ma L, Mintseris J, et al. Transcription factor networks in Drosophila melanogaster. Cell Rep. 2014;8: 2031–2043. doi:10.1016/j.celrep.2014.08.038

22. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80- ). 2015;347: 1260419–1260419. doi:10.1126/science.1260419

23. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci. 2019;116: 6491–6500. doi:10.1073/pnas.1802973116

24. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Res. 2004;14: 1085–1094. doi:10.1101/gr.1910904

25. Reverter A, Hudson NJ, Nagaraj SH, Pérez-Enciso M, Dalrymple BP. Regulatory impact factors: Unraveling the transcriptional regulation of complex traits from expression data. Bioinformatics. 2010;26: 896–904. doi:10.1093/bioinformatics/btq051

26. Cánovas A, Reverter A, DeAtley KL, Ashley RL, Colgrave ML, Fortes MRS, et al. Multi-Tissue Omics Analyses Reveal Molecular Regulatory Networks for Puberty in Composite Beef Cattle. PLoS One. 2014;9: e102551. doi:10.1371/journal.pone.0102551
S1 Table. Number of genes assigned to each distribution shape in datasets with contrasting phenotypes. Numbers in parentheses represent the P-values for enrichment (above) and depletion (below). The categories include differentially expressed (DE) and regulator (REG) genes.

| Shape   | Cattle Feed Efficiency Dataset | Cattle Puberty Dataset |
|---------|--------------------------------|------------------------|
|         | REG (P-value) | DE (P-value) | DE-REG (P-value) | Others (P-value) | Total (P-value) | REG (P-value) | DE (P-value) | DE-REG (P-value) | Others (P-value) | Total (P-value) |
| Shape 1 | 122 (1.62E-07) | 91 (1.0) | 6 (0.83) | 1,784 (0.98) | 2,003 (1) | 449 (4.87E-13) | 1,223 (1) | 82 (1) | 5,003 (1) | 6,757 (4.19E-21) |
| Shape 2 | 48 (3.61E-04) | 56 (1) | 4 (0.96) | 717 (0.23) | 825 (1) | 31 (4.39E-03) | 153 (1) | 5 (1) | 372 (1) | 561 (1.82E-10) |
| Shape 3 | 31 (2.10E-08) | 40 (0.95) | 1 (0.45) | 708 (1) | 780 (1) | 0 (0.84) | 0 (0.76) | 0 (0.98) | 2 (1) | 2 (1) |
| Shape 4 | 818 (5.94E-17) | 157 (1) | 16 (0.86) | 6,835 (0.98) | 7,826 (1) | 1 (0.16) | 0 (1) | 0 (1) | 1 (1) | 2 (1) |
| Shape 5 | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) |
| Shape 6 | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) | 0 (1) |
| Shape 7 | 7 (0.74) | 6 (7.63E-111) | 0 (1) | 80 (0.76) | 93 (1) | 182 (4.22E-11) | 636 (1) | 50 (1) | 2,315 (1) | 3,183 (7.78E-14) |
| Shape 8 | 19 (0.03) | 5 (2.47E-08) | 0 (1) | 111 (0.98) | 135 (1) | 781 (5.86E-38) | 183 (1) | 3 (1) | 5,506 (1) | 6,473 (1.43E-75) |
| Total   | 1,045 (0.74) | 355 (7.63E-111) | 27 (1) | 10,235 (0.76) | 11,662 (1) | 346 (1) | 236 (1) | 9 (1) | 1,952 (1) | 16,978 (7.78E-14) |
### S2 Table. Number of genes assigned to each distribution shape in the cattle feed efficiency dataset split by phenotype.

Numbers in parentheses represent the P-values for enrichment (above) and depletion (below). The categories include differentially expressed (DE) and regulator (REG) genes.

| Shape  | Low Feed Efficiency | High Feed Efficiency |
|--------|---------------------|----------------------|
|        | REG | DE   | DE-REG | Others | Total   | REG | DE   | DE-REG | Others | Total   |
| Shape 1|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (7.67E-07) | (0.05) | (0.12) | 2,137   | (1) | (3.18E-05) | (0.58) | (0.01) | 1,729   |
|        | (3.49E-07) | (1) | (0.98) | (0.89) |         | (2.97E-08) | (1) | (0.63) | (0.99) |         |         |
| Shape 2|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (1.39E-10) | (0.56) | (0.96) | 800     | (0.97) | (7.37E-08) | (0.04) | (0.97) | 830     |
|        | (0.01) | (1) | (0.72) | (0.05) |         | (0.04) | (1) | (0.99) | (0.03) |         |         |
| Shape 3|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (0.04) | (0.83) | (1.96E-04) | 732   | 1 | 0.01 | 0.85 | 5.89E-04 |         |
|        | (2.14E-08) | (0.98) | (0.49) | (1) |         | (3.66E-08) | (0.99) | (0.45) | (1) |         |         |
| Shape 4|     |      |        |        |         |     |      |        |        |         |
|        | (3.75E-18) | (1) | (0.80) | (0.97) |         | (3.66E-16) | (1) | (0.87) | (0.99) |         |         |
|        | (1) | (1.16E-23) | (0.32) | (0.03) |         | (1) | (1.42E-16) | (0.24) | (0.02) |         |         |
| Shape 5|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (1) | (1) | (1) |         | (1) | (1) | (1) | (1) |         | (1) | (1) | (1) | (1) |         |
| Shape 6|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (1) | (1) | (1) |         | (1) | (1) | (1) | (1) |         | (1) | (1) | (1) | (1) |         |
| Shape 7|     |      |        |        |         |     |      |        |        |         |
|        | (1) | (2.08E-04) | (1) | (0.52) |         | (1) | (4.21E-04) | (1) | (0.90) |         | (1) | (1) | (1) | (1) |         |
|        | (4.71E-03) | (1) | (0.56) | (0.56) |         | (0.04) | (1) | (0.93) | (0.21) |         |         |
| Shape 8|     |      |        |        |         |     |      |        |        |         |
|        | (0.56) | (0.52) | (1) | (0.44) |         | (0.01) | (1) | (1) | (0.99) |         | (1) | (0.59) | (0.96) | (0.05) |         |
|        | (0.50) | (0.58) | (0.27) | (0.61) |         | (1) | (0.79) | (0.96) | (0.05) |         | (1) | (0.79) | (0.96) | (0.05) |         |
| Total  | 1,045 | 355 | 27 | 10,235 | 11,662 | 1,045 | 355 | 27 | 10,235 | 11,662 |
S3 Table. Number of genes assigned to each distribution shape in the cattle puberty dataset split by phenotype. Numbers in parentheses represent the P-values for enrichment (above) and depletion (below). The categories include differentially expressed (DE) and regulator (REG) genes.

| Shape   | Pre-puberty       | Post-puberty      |
|---------|-------------------|-------------------|
|         | REG (1) | DE 6.66E-66 | DE-REG 2.05E-05 | Others 1 | Total 6,272 | REG (1) | DE 4.66E-64 | DE-REG 6.50E-08 | Others 1 | Total 6,866 |
| Shape 1 | 397     | 1,179     | 76          | 4,620    | (1.61E-15) | 478     | 1,254     | 88         | 5,046    | (1.32E-09) |
|         | (1)     | (1)       | (1)        | (1)      |            | (1)     | (1)       | (1)       | (1)      |            |
| Shape 2 | 31      | 169       | 4          | 412      | (9.98E-23) | 23      | 118       | 2         | 314      | (4.12E-14) |
|         | (1)     | (1)       | (1)        | (1)      | (5.45E-04) | (1)     | (1)       | (1)       | (1)      | (3.77E-06) |
| Shape 3 | 0       | 2         | 0          | 3        | (0.64)     | 0       | 4         | 0         | 3        | (0.54)     |
|         | (0.84)  | (0.98)    | (0.96)     | (0.31)   |            | (0.92)  | (1)       | (0.94)    | (0.99)   | (0.22)     |
| Shape 4 | 0       | 2         | 0          | 2        | (0.02)     | 0       | 1         | 0         | 0        | (1)        |
|         | (1)     | (1)       | (1)        | (1)      | (1)        | (1)     | (1)       | (1)       | (1)      | (1)        |
| Shape 5 | 0       | 0         | 0          | 0        | (1)        | 0       | 0         | 0         | 0        | (1)        |
|         | (1)     | (1)       | (1)        | (1)      | (1)        | (1)     | (1)       | (1)       | (1)      | (1)        |
| Shape 6 | 0       | 0         | 0          | 0        | (1)        | 0       | 0         | 0         | 0        | (1)        |
|         | (1)     | (1)       | (1)        | (1)      | (1)        | (1)     | (1)       | (1)       | (1)      | (1)        |
| Shape 7 | 196     | 646       | 57         | 2,373    | (3.96E-35) | 167     | 622       | 48        | 2,323    | (3.76E-33) |
|         | (1)     | (1)       | (1)        | (1)      | (1.70E-09) | (1)     | (1)       | (1)       | (1)      | (1.27E-15) |
|         | (1)     | (1)       | (1)        | (1)      | (7.56E-81) | (1)     | (1)       | (1)       | (1)      | (7.00E-36) |
| Shape 8 | 819     | 198       | 3          | 5,791    | (5.71E-265)| 775     | 197       | 2         | 5,513    | (7.66E-240)|
|         | (1)     | (1)       | (1)        | (1)      | (6.54E-27) | (1)     | (1)       | (1)       | (1)      | (1.47E-26) |
| Total   | 1,443   | 2,196     | 140        | 13,199   | 16,978     | 1,443   | 2,196     | 140       | 13,199   | 16,978     |
### S4 Table. Number of genes assigned to each distribution shape in time series datasets.

Numbers in parentheses represent the P-values for enrichment (above) and depletion (below). The categories include differentially expressed (DE) and regulator (REG) genes – refer to methods for specific categories in Drosophila dataset.

| Shape  | Down/down | Down/up | Up/down | Up/up | REG | Others | Total |
|--------|------------|---------|---------|-------|-----|--------|-------|
|         | (4.29E-16) | (0.94)  | (0.01)  | (8.74E-04) | (1) | (1)    | 1,433 |
| Shape 1 | 292        | 305     | 76      | 122   | 106 | 638    |       |
|         | (1)        | (0.07)  | (0.99)  | (1)   | (4.17E-05) | (8.45E-13) |       |
|         | (0.16)     | (0.96)  | (0.02)  | (0.63) | (0.98) | (0.69) |       |
| Shape 2 | 17         | 16      | 9       | 6     | 5   | 50     | 98    |
|         | (0.90)     | (0.08)  | (0.99)  | (0.53) | (0.06) | (0.39) |       |
|         | (0.14)     | (0.41)  | (0.65)  | (0.51) | (0.34) | (0.31) | (0.87) |
| Shape 3 | 3          | 3       | 0       | 0     | 0   | 4      | 10    |
|         | (0.96)     | (0.83)  | (0.65)  | (0.51) | (0.34) | (0.31) | (1)   |
| Shape 4 | 0          | 0       | 0       | 0     | 0   | 0      |       |
|         | (1)        | (1)     | (1)     | (1)   | (1) | (1)    |       |
| Shape 5 | 0          | 0       | 0       | 0     | 0   | 0      |       |
|         | (1)        | (1)     | (1)     | (1)   | (1) | (1)    |       |
| Shape 6 | 468        | 1,159   | 202     | 210   | 467 | 2,351  | 4,390 |
|         | (8.76E-17) | (1)     | (0.98)  | (3.46E-13) | (0.94) | (0.89) |       |
|         | (1.38E-18) | (0.02)  | (1)     | (0.07) | (0.12) |       |       |
| Shape 7 | 1          | 16      | 38      | 171   | 115 | 564    | 790   |
|         | (3.27E-51) | (1.6E-69)| (0.84) | (1)   | (1) | (1)    |       |
|         | (1.43E-29) | (0.01)  | (1)     | (0.85) | (1) | (1)    | (1)   |
| Shape 8 | 267        | 271     | 1       | 2     | 98  | 509    | 1,050 |
|         | (1)        | (0.99)  | (5.10E-20) | (1.48E-30) | (0.18) | (9.72E-04) |       |
|         | (1.43E-29) | (0.01)  | (1)     | (0.85) | (1) | (1)    | (1)   |
| Total   | 1,048      | 1,770   | 326     | 511   | 791 | 4,116  | 7,771 |
|         | (1.43E-29) | (0.01)  | (1)     | (0.85) | (1) | (1)    | (1)   |
S5 Table. Number of genes assigned to each distribution shape in human dataset. Numbers in parentheses represent the P-values for enrichment (above) and depletion (below). The categories include differentially expressed (DE), regulator (REG) and tissue enriched (TE) genes.

| Shape | REG | DE  | DE-REG | TE  | Others | Total |
|-------|-----|-----|--------|-----|--------|-------|
| 1     | (0.55) | (2.48E-04) | (0.02) | (0.09) | (1) | 3,881 |
|       | 215 | 182 | 24 | 55 | 3,411 | 3,881 |
| 2     | (9.31E-06) | (3.28E-10) | (2.92E-04) | (0.99) | (1.53E-03) | 5,065 |
|       | 344 | 262 | 35 | 46 | 4,382 | 5,065 |
| 3     | (0.92) | (4.82E-03) | (0.16) | (1) | (0.96) | 43 |
|       | 1 | 6 | 1 | 0 | 35 | 43 |
| 4     | (0.37) | (1) | (1) | (1) | (0.79) | 8 |
|       | 1 | 0 | 0 | 0 | 7 | 8 |
| 5     | (0.93) | (0.74) | (0.97) | (0.91) | (0.60) | 0 |
|       | 0 | 0 | 0 | 0 | 0 | 0 |
| 6     | (1) | (1) | (1) | (1) | (1) | 1 |
|       | 0 | 0 | 0 | 0 | 0 | 1 |
| 7     | (1) | (1) | (0.88) | (7.21E-34) | (9.40E-01) | 1,860 |
|       | 79 | 48 | 5 | 91 | 1,640 | 1,860 |
| 8     | (4.00E-03) | (2.94E-03) | (0.22) | (1) | (0.07) | 8,427 |
|       | 435 | 217 | 13 | 39 | 7,733 | 8,427 |
| Total | (0.99) | (1) | (1) | (1) | (1) | (3.58E-24) | 17,208 | 19,284 |
S6 Table. Number of genes assigned to each distribution shape in all datasets based on genes being on the top or bottom 5% when ranked by degree (number of significant correlations to other genes). Numbers in parentheses represent the P-values for enrichment (above) and depletion (below).

| Feed Efficiency | Puberty | Drosophila | Duck | Human |
|-----------------|---------|------------|------|-------|
|                 | Top     | Bottom     | Top  | Bottom |
| Shape 1         | (1)     | (4.72E-24) | (1)  | (1.56E-37) | (1)  | (7.1E-126) | (1)  | (9.3E-110) |
|                 | 0       | 198        | 0    | 518    | 0    | 280        | 0    | 335        |
|                 | (8.44E-50) | (1)   | (3.27E-194) | (1)  | (3.74E-36) | (1)  | (1.18E-48) |
|                 | (1)     | (4.27E-252) | (1)  | (8.64E-275) | (1)  | (8.84E-44) | (1)  | (2.3E-280) |
|                 | 0       | 335        | 0    | 314    | 0    | 53         | 0    | 231        |
|                 | (8.41E-20) | (1)   | (1.93E-13) | (1)  | (6.31E-03) | (1)  | (1.27E-06) |
| Shape 2         | (1)     | (1.38E-94) | (1)  | (2.50E-03) | (1)  | (2.68E-06) | (1)  | (6.19E-06) |
|                 | 5       | 18         | 0    | 2      | 0    | 6          | 0    | 4          |
|                 | (1.54E-12) | (6.76E-05) | (0.90) | (1)   | (5.98E-01) | (1)  | (0.81)     | (1)  | (0.11)     |
|                 | (1)     | (6.97E-233) | (1)  | (2.50E-03) | (1)  | (1)        | (1)  | (1)        |
| Shape 3         | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | 0       | 0          | 0    | 0      | 0    | 0          | 0    | 0          |
|                 | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
| Shape 4         | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | 578     | 31         | 0    | 2      | 0    | 0          | 0    | 0          |
|                 | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
| Shape 5         | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | 0       | 0          | 0    | 0      | 0    | 0          | 0    | 0          |
|                 | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
| Shape 6         | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | 0       | 0          | 0    | 0      | 389  | 27         | 0    | 1          |
|                 | (1)     | (1)        | (1)  | (1)    | (1)  | (1)        | (1)  | (1)        |
|                 | 1       | (0.99)     | (1)  | (1.54E-31) | (1)  | (1)        | (0.99) | (1)        |
|                 | (8.32E-03) | (0.05) | (1)  | (5.83E-70) | (1)  | (2.48E-19) | (0.01) | (2.35E-36) |
| Shape 7         | (1)     | (1)        | (1.77E-58) | (1)  | (1)        | (1.31E-64) | (1)  | (<1E-999) |
|                 | 0       | 0          | 0    | 0      | 550  | 8          | 0    | 10         |
|                 | (9.45E-04) | (9.45E-04) | (1)  | (3.03E-166) | (1)  | (6.20E-26) | (1)  | (8.6E-118) |
| Shape 8         | (1)     | (1)        | (1)  | (1)    | (1.31E-64) | (1)  | (<1E-999) |
|                 | 0       | 0          | 0    | 0      | 524  | 50         | 965  | 0          |
|                 | (9.45E-04) | (9.45E-04) | (1)  | (3.03E-166) | (1)  | (6.20E-26) | (1)  | (8.6E-118) |
| Total           | 583     | 583        | 849  | 849    | 389  | 389        | 666  | 666        |

Total 583 583 849 849 389 389 666 666