Modularity in gene expression across animal development

Alfredo Rago[1]*, Jack Werren[2], John K. Colbourne[1]

November 12, 2014

[1]Environmental Genomics group, School of Biosciences, University of Birmingham, UK
[2]Department of Biology, University of Rochester, Rochester, New York, USA
*AXR280@bham.ac.uk

1 Abstract

What are the units of evolution? From its inception, evolutionary theory focused on individual organisms as the units under natural selection. Yet there is much evidence that selection also acts at lower levels of biological organization, optimizing the functions of individual gene products independently of the whole organism. By contrast, high-throughput gene expression and knock-out studies now demonstrate that virtually every gene requires interactions with products of other genes and external environment to carry out its functions (Mackay and Anholt, 2006). If true, the ability of each gene to be selected independently of other genes will be constrained by its need to form stable functional interactions (Papakostas et al., 2014). It is time to break new ground by unifying the dynamics of how genes are co-regulated with those of selection in order to identify the genetically encoded functional modules that are targets for evolution. Current tests of significant changes in gene expression operate under the pan-reductionist framework by detecting changes in mean expression between different conditions for every gene (Soneson and Delorenzi, 2013; Smyth, 2005). Alternatively, the most common multivariate methods employed in gene expression studies de-convolute gene modules based on their correlation and generate network graphs that display their relative association (Allen et al., 2012). Both approaches present a theoretical hindrance for a biologically meaningful interpretation of their results when applied to gene expression data: Gene-specific linear models do not include explicit tests for gene-gene interactions; machine learning and correlation clustering are unable to discriminate the relative contributions of our conditions of interest to the observed gene-gene correlations. Here, we implement a method for condition-specific detection of gene co-regulation based on cluster modularity. This method relies on a correlation based clustering approach to identify gene-gene interactions and on permutation-based linear models to test for variation of relevant network concepts across experimental conditions. Further to that, I employ this condition-specific network-based framework to provide an interpretation of gene expression patterns in sexual

development of the Jewel wasps (Nasonia *vitripennis*) from both a gene regulation and gene evolution perspective.

2 Data source and quality controls

The data used in this study consists of a developmental time series of transcriptional activity of whole animals in males and females of the jewel wasp Nasonia *vitripennis*. The experimental design comprises five distinct developmental stages from early embryo to sexually mature adult, for both males and females. Each of these conditions was sampled in triplicate. All animals used for data collection come from an inbred strain with minimal amounts of genetic diversity. Further to that, Nasonia species have an haplodiploid sex-determination mechanism and lack sex-specific chromosomes. Observed differences in the expression profile of males and females are thus likely to be exclusively caused by differential gene regulation rather than by genetic differences between individuals.

All of the data used was generated trough the application of single-channel tiling-path microarrays. Prior to our analyses these data were subject to rigorous quality control and normalization procedures. The signal from each probe was normalized to the 99th quantile of expression from the random Markov probes present on each array, which reflect the intrinsic noise levels on each chip. Further to that, probes were assigned to specific exons according to the latest release of the Official Nasonia Gene Set (OGS2.0, Gilbert and Rago et al., in preparation). After these steps the transcriptional status of our samples is expressed as a count table whose entries represent expression of each exon across every sample. Our data is therefore analogous to those produced *via* RNA-seq, although expressed as a continuous log transformed signal to noise ratio rather than as a discrete number of read counts or proportion.

3 Transcript deconvolution

Whereas most studies assess the level of expression at the gene level, this approach has a high chance of producing misleading data as it ignores the additional diversity that is generated at the transcriptomic level. Eukaryotic genes are comprised by a continuous series of intronic and exonic sequences. The former are excised during or shortly after transcription. The latter are retained in the mature mRNA either as protein-coding or as untranslated regions (UTRs). Differential inclusion or exclusion of exons (alternative splicing) is a widespread feature of eukaryotic gene regulation. Transcripts of the same gene comprising different combinations of exons (alternative transcripts or isoforms) can trigger different effects despite constant gene expression at the gene level.

A more biologically meaningful alternative is to quantify expression levels of individual transcripts. While more accurate, this latter approach faces several technical challenges. Transcript-based analyses increase the dimensionality of datasets by an average of 1-2 orders of magnitude as they analyze independently each isoform rather than reducing them to a single gene expression value. They also require reliable assignment of expression values to individual transcripts. Transcript-specific expression is typically obtained by averaging

the signals of transcript-aspecific exons to match those provided by transcript-specific exons (Trapnell et al., 2010). This implies the ability to trace an *a priori* distinction between transcript specific and transcript aspecific exons, a task delegated to publicly available gene sets. There is however abundant evidence that even the most complete gene set annotations available are just now approaching completeness (Brown et al., 2014). It is therefore preferable adopt data-specific estimation of constitutive and transcript-specific exons in order to enable the detection of novel alternative splicing events relevant to the design of interest.

In order to address these issues we implemented a modified version of the method proposed by Patrick et al. (2013). Briefly, we perform hierarchical correlation based clustering of exons within each gene. Strong correlations among exons arise from their coexpression as part of a single transcript. Therefore, every cluster will represent either a transcript or the subset of exons that are present across all isoforms (constitutive exons). Constitutive exon clusters can be distinguished from transcript specific exon clusters from their total expression value. Since constitutive exons are present across all transcripts their average expression levels will be always greater than each other individual cluster.

Correlation based clustering is an intuitive and easily implemented way of generating experiment-specific transcript assignments but its results depend on the threshold chosen as a cutoff for independent expression. Whereas Patrick et al. (2013) justify their choice of an arbitrary threshold by matching with previously published data, we find this solution counterproductive to the goal of detecting experiment-specific splicing events, especially in species which currently lack a comprehensive annotation of their transcriptional diversity.

As such, we developed a gene-by-gene bottom up iterative algorithm choosing as a termination clause the presence of a single cluster whose expression is always ranked first or tied. Our choice of termination criterion is based on the biological observation that the vast majority of genes comprise a core set of exons included in all transcripts (Chen, 2013). As constitutive exons are present across all transcripts, their signal will approximate the sum of all alternative transcripts and thus be greater than each one individually. In the special case where only one alternative transcript is present in our sample of interest, constitutive exons will be tied first with the only group of alternative transcripts present in the sample. To respect these assumptions a cluster assignment will have to identify a single exon cluster whose expression level is consistently higher than those of all exon groups or tied with the best ranking one. We can thus interpret our termination criterion as the indication that our clustering method has been able to find a valid group of constitutive exons with the lowest correlation cutoff possible. Other groups of exons are classified as belonging to the same transcripts only if their reciprocal correlations are higher or equal to those observed amongst constitutive exons of their own gene. In the scenario where no alternative isoforms are observed for our gene of interest within our experiment the algorithm will converge to a single cluster comprising all exons (all exons are constitutive).

To further improve the flexibility of our algorithm we also include two biologically interpretable parameters that enable fine-tuning of the sensitivity to robustness tradeoff on the termination clause. The sensitivity used generating rankings can be set to exclude an excessive number of significant digits, which likely represent stochastic variation in gene expression measurements. It is also possible to specify the number of samples in which constitutive exons can violate the termination clause, or expected false negatives.

The final output of our clustering method is a reduced count table, with averaged expression values among highly correlated exons. More importantly, each cluster of exons is representative of either total gene expression (in the case of constitutive exon clusters) or isoform specific expression (in all other cases).

We removed negative values (which fall below the noise threshold) before calculating correlations as they might cause spurious signal during clustering, excluded genes which did not display expression for any exon in any of our samples, allowed one exception to the termination clause and reduced the sensitivity to the first three significant digit of gene expression values. We were able to assign 76.162 exons (42%) to constitutive clusters and 90.181 (49%) to alternative clusters, which were then reduced to 55.486 clusters of which 18.231 constitutive (33%) and 37.255 alternative (66%). We observed an even split between genes with at least one assigned isoform (11.501) and genes that did not show sign of alternative splicing in our experiment (11.648). Genes with alternative splicing were assigned a median of 3 isoforms, consistently with the highly skewed power-distribution of isoforms per gene reported in other animals.

4 Network construction and cluster assignment

Measuring the absolute amount of each transcript provides an accurate representation of the cell's transcriptional status but it also restricts our ability to investigate the molecular bases of transcript regulation. An increase in a transcript's level can be achieved either by raising the overall RNA production of its gene or biasing splicing towards a specific isoform. Since these two processes are partially interdependent assessing the degree by which they can be selectively targeted to modulate transcript expression is an interesting biological question, especially since different molecular mechanisms might mediate and regulate different steps of the complex RNA maturation process.

To the authors' knowledge, comparisons between transcript regulation *via* gene regulation and alteration of alternative splicing has so far not been tackled with an integrative approach. The few studies that address splicing as a focus for network analyses do so trough exon specific analyses (Dai et al., 2012; Chen and Goldhamer, 1999) and without reference to the regulation of their parent gene. Exon-based approaches are able to identify clusters of co-spliced exons across the experimental treatments but lack the ability to address the relationships between gene expression and alternative splicing as they focus exclusively on the latter, therefore impeding the detection of cross-regulation between the two processes. Further to that the number of exons within a transcriptome is usually larger than those of observed alternative transcripts and imposes an additional computational load.

A popular approach to disentangle signal from gene expression from alternative splicing is to rely on splicing ratios: the relative abundance of a single transcript (or exon) over the total gene expression level (Monlong et al., 2014). Splicing ratios are a continuous 0-1 bound variables that indicate the degree of alternative splicing irrespective of gene expression levels. Calculating splicing ratios requires the ability to measure both the gene-wide and transcript-specific expression levels. We solve this issue trough the application of correlation-based clustering (see previous section), which generates clusters of constitutive and alternative ex-

ons. Constitutive exons are included in all isoforms observed in our gene and comprise an unbiased estimator of total RNA production, or gene expression. Clusters of alternative exons represent the subset of RNAs that are allocated specifically to each isoform. We can thus reconstruct splicing ratios by dividing the expression value of each non constitutive cluster by its genes' constitutive cluster.

Due to noise in the experimental measurements, two types of exceptions to the expected distribution are possible. Expression ratios higher than 1 can be caused by the relative increase in the error term that occurs when splicing ratios approach unity, an hypothesis confirmed by manual spot-checking and summary statistics. Consequently, we replace splicing ratios above the unity with 1. Divide by zero errors arise when a non-expressed gene contains a group of isoform-specific exons with sufficient noise to breach the cutoff threshold. As these errors are caused by non-expressed genes, we replace divide by zero values with zeroes. The special case of isoforms comprised exclusively by baseline exons is addressed implicitly as it corresponds to $1-\sum E_i$, where $E_i$is the isoform-specific expression ratio for each of the genes' other transcripts.

The final dataset used for our network generation consists of two distinct types of data entries. Gene expression values, which reflect total gene expression, and splicing ratio values, that indicate the allocation of total RNA production to individual transcripts. The overall dimensionality of the dataset is given by the sum of the number of genes measured, plus the number of putative transcripts represented by each splicing ratio, multiplied by the number of samples. To further simplify analyses, we replaced values below noise threshold with zeroes and removed 13.370 entries (24%) which showed zero variance and are most likely due to genes and exons whose expression is not observed in our dataset. This resulted in a dataset with a total of 42.116 entries, 23% the size of the total number of exons annotated for the Nasonia gene set, and 20% the size of the total gene+exon dataset.

We generated a network comprising all data using the WGCNA R package (Langfelder and Horvath, 2008), which performs signed undirected weighted network construction *via* power-scaled correlation. We employed biweight midclustering correlations to ensure robustness to outliers and selected a power-scaling threshold of 20 after verifying that the generated network was consistent with the expected scale-free topology. Following network construction, we applied hierarchical clustering based on topological dissimilarity to detect groups of coexpressed genes together with co-spliced isoforms. Since our aim is to find condition-specific modules, we maximized the sensitivity of the clustering split parameter and set our minimal cluster size at 20 transcripts. We subsequently collapsed all clusters with a distance of 10% of lower.

Our final output generated a total of 341 clusters, with a median size of 63 genes or isoforms.

5 Resampling and GLM testing of network parameters

The most common method for characterizing gene clusters is testing the changes of their mean expression levels under different treatments via correlation or linear models . These methods enable characterization of clusters trough their expression profile but impede in-

vestigation of changes in network structure as they reduce transcriptional complexity to a single value per cluster (mean expression). By contrast network construction methods collapse all available data to produce a single network in which connections between genes are an average of all treatments. Using all available data to construct the network is often the wisest choice, as the increased diversity produces a more accurate separation between functionally distinct clusters. However, it also implies that no further data is available to test the relative effect of different treatments to the observed interactions.

Due to these challenges, the field of transcriptional network comparison is still in its early stages. Most published methods on network comparisons consider gene-gene interactions as the actualization of a constant underlying architecture. As a result the role of the diverse conditions under which different transcriptional interactions and network topologies emerge is often overlooked. An example of this mindset is the tensor-based recurrently dense subnetwork approach (Dai et al., 2012; Li et al., 2011; Yan et al., 2007), which uses diversity only as a means to achieve higher resolution between interacting transcripts. This approach is in contrast with biological data, which highlights that molecular interactions are highly contingent on the environmental parameters. In order for us to understand how this contingency is enacted, we need a method to discriminate treatment-specific changes in network topologies.

To overcome the limitations of network-based approaches we employ boostrapping to generate multiple networks based on random subsets of the dataset. We record the topological parameters of interest across the permuted networks as well as the type of sample treatments that have been excluded in each permutation. Finally, we evaluate how removal of specific treatments affects the network parameters of interest trough the use of linear models.

The core of our method lies in the resampling strategy applied during bootstrapping, which needs to generate both random and non-random removals of samples in regard to our treatments of interest. In our case, we apply two different resampling strategies to address the presence of stage-specific modules and the presence of sex-specific modules within each stage. In the case of stage-specific modules, we want to exclude possible spurious effects from sex-bias while preserving randomness in regard to stage. As such, we constrain bootstrapping to preserve a 1:1 ratio of male:female samples removed from within each stage and remove a total of 6 samples for each permutation. This creates a series of datasets with a constant number of missing samples and sex ratio, but with a number of removed samples from each stage that varies from 0 to 6 (all of the samples in a single stage). In the case of sex-specific modules within each stage we allowe variation of sex ratios removing three samples from each stage in each permutation. This generates a series of datasets in which any single stage lacks 3 samples of varying sex ratio.

Since the main focus of our study is the presence of functional (and therefore evolutionarily) cohesive modules, we record two network parameters for each of the modules identified during global network construction (see previous section). Both parameters are modular variants of network density, or the proportion of observed connections over the total possible ones. The first parameter is the network density of all transcripts within a cluster, or integration coefficient. The second parameter is the the network density of transcripts within a cluster with those outside of it, or pleiotropy coefficient. In the biological interpretation, in-

tegration reflects the degree by which genes are coregulated under a specific treatment, while pleiotropy the residual interactions with transcripts outside of its own functional module.

As densities, both integration and pleiotropy coefficients are 0-1 bound proportions and can be converted to a normally distributed variable trough the application of a logit transformation. We therefore employed generalized linear models (GLMs) with a logit link function and a Gaussian error distribution to analyze the changes in integration and pleiotropy coefficients in relation to our treatments of interest. We further confirmed the appropriateness of distributional assumptions via diagnostic plots.

6 Future work

In this paper we demonstrate proof of concept for a method of detecting dynamic changes in topological properties of transcriptional networks. Our method is designed to take advantage of the increasing diversity of transcriptomic studies while reducing their complexity to biologically interpretable units. We took special care to enable flexibility in all steps of data treatment. Sensitivity during exon clustering can be finely tuned depending on data quality. The clustering methods employed can be replaced with user-defined ones or even manually specified cluster assignments. Our method is also able to accommodate for arbitrarily complex questions trough appropriate design of the resampling strategy. While its initial application is promising, we still need to perform adequate quality control steps with benchmarking and synthetic datasets to assess the relative impact of noise, platform-specific and sample-specific challenges.

References

Mackay, T. F. C.; Anholt, R. R. H. *Annual review of genomics and human genetics* **2006**, *7*, 339–67.

Papakostas, S.; Vø llestad, L. A. r.; Bruneaux, M.; Aykanat, T.; Vanoverbeke, J.; Ning, M.; Primmer, C. R.; Leder, E. H. *Nature communications* **2014**, *5*, 4071.

Soneson, C.; Delorenzi, M. *BMC bioinformatics* **2013**, *14*, 91.

Smyth, G. K. In *Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor*; Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W., Eds.; Springer: New York, 2005; pp 397–420.

Allen, J. D.; Xie, Y.; Chen, M.; Girard, L.; Xiao, G. *PloS one* **2012**, *7*, e29348.

Trapnell, C.; Williams, B. a.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. *Nature biotechnology* **2010**, *28*, 511–5.

Brown, J. B. et al. *Nature* **2014**, 1–7.

Patrick, E.; Buckley, M.; Yang, Y. H. *BMC bioinformatics* **2013**, *14*, 31.

Chen, F.-C. *Briefings in bioinformatics* **2013**, *15*, 542–551.

Dai, C.; Li, W.; Liu, J.; Zhou, X. J. *BMC systems biology* **2012**, *6 Suppl 1*, S17.

Chen, J. C.; Goldhamer, D. J. *Cell and tissue research* **1999**, *296*, 213–9.

Monlong, J.; Calvo, M.; Ferreira, P. G.; Guigó, R. *Nature Communications* **2014**, *5*, 4698.

Langfelder, P.; Horvath, S. *BMC bioinformatics* **2008**, *9*, 559.

Li, W.; Liu, C.-C.; Zhang, T.; Li, H.; Waterman, M. S.; Zhou, X. J. *PLoS computational biology* **2011**, *7*, e1001106.

Yan, X.; Mehan, M. R.; Huang, Y.; Waterman, M. S.; Yu, P. S.; Zhou, X. J. *Bioinformatics (Oxford, England)* **2007**, *23*, i577–86.