Bioinformatics, 31(13), 2015, 2123–2130
doi: 10.1093/bioinformatics/btv118
Advance Access Publication Date: 24 February 2015
Original Paper

OXFORD

Gene expression

Guidance for RNA-seq co-expression network construction and analysis: safety in numbers

S. Ballouz*, W. Verleyen and J. Gillis*

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard Woodbury, NY 11797, USA

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on June 19, 2014; revised on January 28, 2015; accepted on February 19, 2015

Abstract

Motivation: RNA-seq co-expression analysis is in its infancy and reasonable practices remain poorly defined. We assessed a variety of RNA-seq expression data to determine factors affecting functional connectivity and topology in co-expression networks.

Results: We examine RNA-seq co-expression data generated from 1970 RNA-seq samples using a Guilt-By-Association framework, in which genes are assessed for the tendency of co-expression to reflect shared function. Minimal experimental criteria to obtain performance on par with microarrays were >20 samples with read depth >10 M per sample. While the aggregate network constructed shows good performance (area under the receiver operator characteristic curve ~0.71), the dependency on number of experiments used is nearly identical to that present in microarrays, suggesting thousands of samples are required to obtain 'gold-standard' co-expression. We find a major topological difference between RNA-seq and microarray co-expression in the form of low overlaps between hub-like genes from each network due to changes in the correlation of expression noise within each technology.

Contact: jgillis@cshl.edu or sballouz@cshl.edu

Supplementary information: Networks are available at: http://gillislab.labsites.cshl.edu/supplements/rna-seq-networks/ and supplementary data are available at *Bioinformatics* online.

1 Introduction

While RNA-seq offers major advantages over microarrays, including higher dynamic range and whole transcriptome assessment, it is not free of technical and biological biases. These include dependencies on library preparation, sequencing depth, gene lengths, GC content and transcript lengths, as well as other laboratory-specific practices (Hitzemann *et al.*, 2013). In the absence of careful experimental design, distinguishing gene expression variability due to technical artifacts and biological signal is challenging (Auer and Doerge, 2010). In addition, the downstream impact of biases in RNA-seq analysis will depend on the purpose to which it is being put. Our focus is on the development of appropriate standards and controls for RNA-seq co-expression analysis.

Co-expression networks use the correlation (or related measures) of gene expression profiles across multiple samples to ascertain common regulation and thus common functions (Eisen *et al.*, 1998). Co-expression network analysis in microarrays exists in broadly two discrete categories. In the first case, descended from the first co-expression analyses, there are studies which are highly targeted toward conditions of interest with networks derived from relatively small numbers of samples (dozens to the low hundreds) and often focusing on broad network changes in some condition of interest (Voineagu *et al.*, 2011). With increasing data, meta-analysis across many datasets became more common, with samples numbering in the many hundreds or thousands (Wren, 2009). Typically, the quality of co-expression networks is measured using some variant of the Guilt-By-Association (GBA) principle (Oliver, 2000). GBA states that genes with similar functional properties will tend to interact or exhibit similar profiles in network data, such as co-expression. Even 'gold-standard' expression datasets perform relatively poorly under

GBA when assessed in benchmarking exercises (e.g. the Mousefunc competition for genome-wide gene function prediction—Pena-Castillo *et al.*, 2008). However, meta-analysis across many datasets often improves performance dramatically (Gillis and Pavlidis, 2011a; Lee *et al.*, 2004; Wren, 2009). Very few analyses to date have been performed on co-expression networks from RNA-seq data (although see, e.g. Iancu *et al.*, 2012; Sekhon *et al.*, 2013), with fewer conclusions drawn about their utility, nor has their presumed novelty been fully assessed.

In the following, we attempt to provide a better understanding of appropriate standards for RNA-seq co-expression analysis. Methodologically, we begin with estimating expression from raw data, move to constructing individual networks, examining aggregate properties and then finally consider a variety of machine learning methods applied to the network. At each stage, we are concerned with what factors drive variation in constructed co-expression networks. First, we are concerned with the functional connectivity in the networks, which we measure as a GBA function prediction task, using the cross-validation results to assess performance. Second, we wish to assess network features which may be important but not map to previously characterized function, and for this we focus on node degree. Node degree is central to network topology and a major factor in numerous other properties (Gillis and Pavlidis, 2011b; Newman, 2003).

At each stage of our analysis, we define methods that yield reasonable results for the downstream method. Our intent is not to re-characterize, e.g. normalization methods in RNA-seq data, nor to claim that we have identified an ideal means for future RNA-seq co-expression analysis. Rather, we attempt to define a habitable zone of methods which may be reasonably used without introducing major artifacts and which will allow the field to progress as individual researchers interpret their own data or conduct meta-analyses. We observe a particularly strong role for number of samples and aggregation (Sections 3.2 and 3.3) in GBA assessments while mapping and machine learning methods (Sections 3.1 and 3.4) affect primarily topology and not GBA. We explain the discrepancy by finding topologically important genes vary readily (Section 3.5), but primarily due to noise in the data (Section 3.6).

2 Methods

2.1 RNA-seq datasets

We obtained the raw sequence reads in SRA format for 10 RNA-seq datasets from the SRA database (Leinonen *et al.*, 2011). These were converted into fastq files using 'fastq-dump' from the SRA Toolkit (https://github.com/ncbi/sratoolkit), filtered using the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and then aligned with Bowtie2 using the default settings (Langmead and Salzberg, 2012). Sequences that mapped to multiple locations in the genome were discarded. The expression levels in terms of FPKM were determined using Cufflinks (Trapnell *et al.*, 2010) and the GENCODE annotations file (version 18) (Harrow *et al.*, 2012). For comparison, we used RSEQtools (Habegger *et al.*, 2011) as a secondary expression level estimation pipeline. We also obtained 43 RNA-seq datasets from the Gemma resource (Zoubarev *et al.*, 2012) that contained 10 or more samples and had detected at least 17 000 protein coding transcripts and had been analyzed using RSEM (Li and Dewey, 2011). We downloaded expression levels from the BrainSpan atlas (2011) of their brain tissue analysis set, across 16 tissues and 42 patients, totaling 578 samples. We obtained gene count files from the ReCount database (Frazee *et al.*, 2011) for two large human experiments. We calculated the FPKM (Mortazavi *et al.*, 2008) using local R scripts. We used 50 unique RNA-seq experiments for our aggregate analyses with an overlapping set of 30 705 nodes, of which 18 292 were protein-coding nodes. We ignored alternate isoforms in our aggregates and performance measures, and used the median value in case of multiple measures.

2.2 Microarray datasets

We downloaded 43 separate experiments across 5134 individual microarrays with at least 20 samples, all performed on the 'Affymetrix Human Genome U133 Plus 2.0 Array' (GPL570) from GEO using the 'GEOquery' R package (Davis and Meltzer, 2007). For each expression dataset, the data were quantile normalized with the 'limma' R package (Smyth, 2004). Probes on the platform that targeted multiple genes were discarded while the median expression level was taken for genes that had multiple probes, leaving 20 283 genes using the NCBI 'Homo_sapiens.gene_info.gz' data file (March 2013).

2.3 GO, KEGG and Reactome annotations

In order to assess how well-known biological information was captured in our co-expression networks, we obtained gene annotations from the Gene ontology (GO) (Ashburner *et al.*, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and the Reactome (Croft *et al.*, 2011) databases. GO annotations for human genes were obtained from the GO Consortium 'gene_association.goa_ref_human.gz' (March 2013). Gene annotations labeled with IEA evidence codes were removed, and the remaining terms were propagated in the ontology graph using the transitive property (genes of a child node acquire the annotations of the parent terms). GO terms were limited to groups containing 20–1000 genes, leaving 2953 GO groups and 11 859 associated genes. Pathways were downloaded from the KEGG website (http://www.genome.jp/kegg/). Reactome pathways were extracted from the 'gene_association.reactome' file (March 19, 2013), which we downloaded from the Reactome website (http://www.reactome.org/download/). No further parsing of the KEGG or Reactome data were performed, other than the pathway needed to contain at least two annotated genes, leaving 1530 Reactome pathways with 6252 associated genes and 259 KEGG pathways with 6398 associated genes. For our machine learning analysis, we used GO slim, a version of GO that filters analyses to 109 GO terms, broadly representative of the ontology overall.

2.4 Co-expression networks and network aggregation

A co-expression network was generated for each RNA-seq experiment by calculating the correlation of each gene pair across the samples of an experiment, and then treating the rank of the coefficient measure within the network as an edge weight (Gillis and Pavlidis, 2011a). Changes in the distribution of actual correlations (induced by sample size dependencies) have no impact on our results; only the relative strength of gene–gene correlations within an experiment alters network topology. We tested the effects of methodology by using the Spearman or Pearson correlation coefficients for calculating the weight edges. We also used the WGCNA method to test the effect of thresholding the networks (Zhang and Horvath, 2005). Further to using correlation coefficients as weights, we also tested using mutual information (MI) with the parmigene R package (Sales and Romualdi, 2011), and used four algorithms to infer the networks: ARACNE additive and multiplicative (Margolin *et al.*, 2006), CLR (Faith *et al.*, 2007) and MRNET (Peng *et al.*, 2005). Our aggregate network was generated from the individual networks

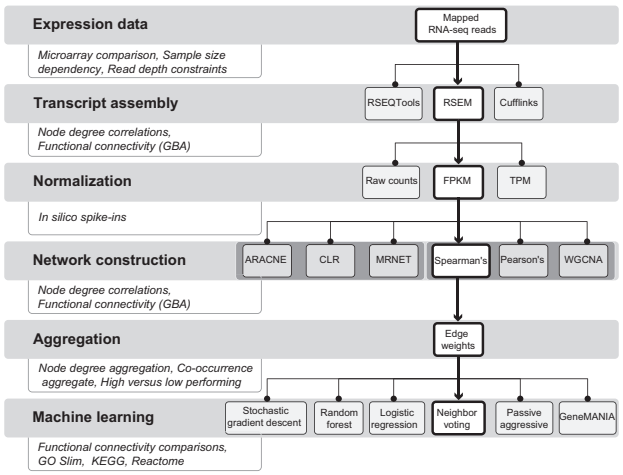

Fig 1. RNA-seq co-expression pipeline. The branched tree shows the differing methods and tools applied. The dark paths are the default procedures and methods used when assessing a secondary feature. The italicized text in the white boxes describes the different tests and validations used in our assessments

through summation of the ranked edge weights. We also resampled across datasets to ensure the aggregated network exhibited robust properties. Overall, our RNA-seq aggregate included 50 networks and 1970 samples. In a similar fashion, we generated an aggregate of the microarrays on a random sample of 30 datasets using 3320 individual microarrays. The experiments selected for both networks are available in the Supplementary Tables S1 and S2.

2.5 Network topology and functional connectivity assessments

We calculated node degree as the summation of all the weights connected to a given node. We used the GBA principle and a neighbor voting algorithm to measure functional connectivity of the network as described by 'known' biological properties. In brief, GBA states that genes that are connected carry similar functions. By using known annotations as labels for the genes (here we used GO, KEGG and Reactome labeled gene sets), one can determine how well the network recapitulates this information (i.e. how many genes with the same function are connected) by using a machine learning algorithm to classify genes as belonging to a given function based on their connectivity within the network. A very simple but effective method is neighbor voting where genes are given the labels of other genes in their neighborhood (i.e. their connections); in our case, their score is the fraction of genes to which they are connected which have a given functional property (Gillis and Pavlidis, 2011a). Through performing a 3-fold cross validation and calculating the average area under the receiver operator characteristic curve (AUROC), we used this as a measure of the functional connectivity of the network, termed 'performance'. We then also used five more sophisticated machine learning methods covering a range of the most popular approaches which include logistic regression (Fan *et al.*, 2008), random forest (Breiman, 2001), support vector machine (SVM) through stochastic gradient descent (Zhang, 2004), the passive aggressive method (Crammer *et al.*, 2006) and GeneMANIA (Mostafavi *et al.*, 2008).

3 Results

We analyze co-expression from the raw expression data, through transcript assembly and normalization to network construction, and then on to meta-analysis through aggregation and machine learning of gene function. One difficult point to address is the interplay between method choices at each stage of our analysis. To use every expression estimate pipeline in combination with every network construction technique in combination with every machine learning algorithm and so forth is not practicable. In general, we use what we consider to be a mainstream method in other stages of analysis and then examine methodological variation within a given stage across a range of possibilities to determine robustness and dependencies as summarized in Figure 1.

3.1 RNA-seq expression estimates and co-expression measures affect network topology but not functional connectivity

To begin, we assessed the use of three standard pipelines to obtain RNA-seq expression levels as a possible source of variance in our co-expression analyses. We used a sample of 10 representative experiments, comparing the estimation expressions of Cufflinks, RSEQtools and RSEM and constructed co-expression networks using our default procedures (highlighted in Fig. 1 and described in Section 2). Network topology was not fully robust to the methods used (Supplementary Fig. S1A), with the correlations of node degree between the networks created ranging between $\rho = -0.39$ and $\rho = 0.88$.

We then measured a network's functional significance in its ability to capture functional information as encoded by GO using a cross-validation approach; essentially asking if a genes' function is the same as its neighbors. Using this metric, we evaluated the three expression estimation methods to determine which produced the best networks on the 10 experiments. We see that these methods generated networks with similar performances (Supplementary Fig. S1B), even though the expression profiles and networks were modestly correlated (Supplementary Fig. S1C).

We next evaluated the use of the Spearman and Pearson correlation coefficients as means of determining co-expression between gene pairs, along with MI in the same subset of 10 experiments (Supplementary Fig. S1D). The MI network performs below the correlation coefficients, and closer to the networks generated including the negative correlation (average AUROC = 0.56). When we also tried the alternative MI inference methods such as ARACNE, CLR and MRNET, the networks also performed considerably worse than the correlation coefficient inference methods (e.g. average AUROCs for ARACNE-additive = 0.52, ARACNE-multiplicative = 0.53, CLR = 0.55 and MRNET = 0.55). Also, the WGCNA soft thresholding acts in a similar fashion, lowering performance (average AUROC = 0.55). In summary, the computationally simplest and most transparent method for calculating co-expression—simple correlation—was also the highest performing.

3.2 Many samples and greater read depth increases the functional connectivity of the networks

Across our corpus of individual experiments, co-expression networks constructed exhibit only modest performance, being fairly tightly clustered around an average AUROC of 0.585. While some GO groups exhibited good performance across all networks, they were a relative minority, with only 0.7% of GO terms having average performances above 0.8 (AUROC, Fig. 2A). Moreover, there were few high performing GO terms outside of this minority. While in theory it is possible that individual networks work well for different functions, this did not occur to any substantial extent. If we assign each GO terms the maximum value it took across all of

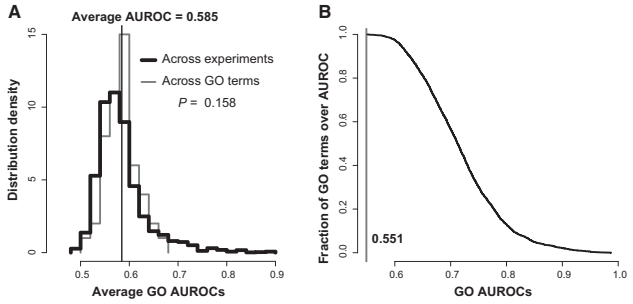

Fig 2. RNA-seq co-expression experiment performance at predicting gene function. (**A**) Distribution of all the AUROCs of the individual GO terms (2953 GO terms) averaged across experiments (thick line, SD ~0.081). Averaging the AUROCs across the GO terms in each experiment (50 RNA-seq datasets) shows a similar but shorter range (mean SD ~0.067). (**B**) Fraction of GO terms ever achieving the given AUROC. Few GO terms are predictable at a high level (AUROC ~0.9) even once

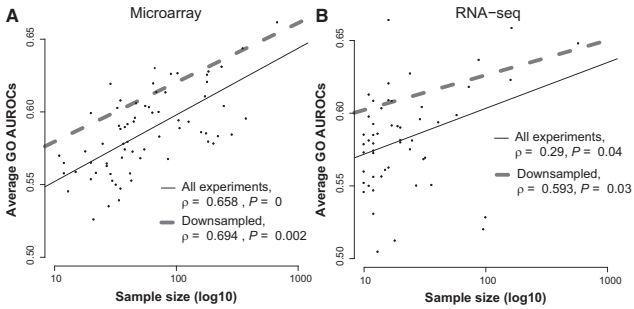

Fig 3. Dependency of GO functional connections on sample sizes. Network derived from (**A**) microarray experiments and (**B**) RNA-seq experiments are measured for GBA performance and plotted against sample size. The microarray-derived networks have a much higher correlation with the sample size ($\rho = 0.66$, $P < 2.2\,e{-}16$) than RNA-seq ($\rho = 0.29$, $P \sim 0.04$)

the co-expression data, we can see that only 2% of GO terms ever perform better than AUROC 0.9 in any of the assessed networks (Fig. 2B).

Networks constructed using more samples were generally higher performing, and the trend is surprisingly consistent in strength between microarrays and RNA-seq (Fig. 3A and B). Samples assayed by microarrays show a linear relationship between sample size and network performance (Fig. 3A, $\rho = 0.66$) however RNA-seq co-expression networks exhibit more variability in performance as a function of sample count than in microarrays (Fig. 3B, $\rho = 0.29$, $P \sim 0.04$), possibly due to variation in read depth. Controlling for this by sampling from within large experiments to calculate the dependency on sample size shows a very similar performance trend, albeit shifted upward (AUROC gain of ~0.02) for both microarray and RNA-seq. A similar, albeit weaker trend is visible with increasing read depth (Supplementary Fig. S2C) and suggests using RNA-seq co-expression to obtain even modest performance (AUROC >0.55) requires sample sizes >20 and read depth per sample close to 100 M reads. The apparent performance dependencies suggest that to move from an AUROC = 0.55–0.6 requires either thousands of samples, or millions of more sequenced reads, which is likely not feasible in a single experiment. We repeated the same functional connectivity experiments using KEGG and Reactome as references, and had essentially identical results (Supplementary Figs. S3 and S4).

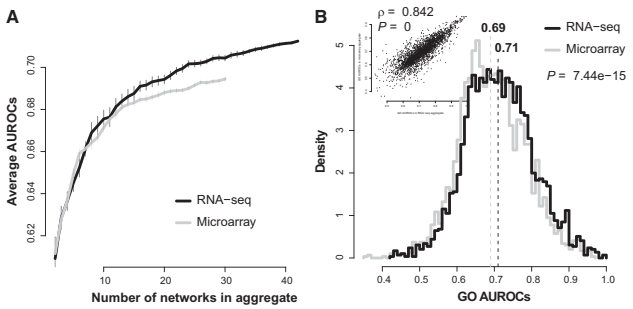

Fig 4. GO functional connectivity improves with aggregation. (**A**) We measure the GBA-based performance of the aggregated network in microarrays (grey) and RNA-seq (black) and see a similar dependence on number of networks involved. As more experiments are aggregated, the performance as recorded by the average AUROC increases. All the aggregation curves are averaged across 10 different orderings of experiments (using a sample of 43 networks for the RNA-seq, and 43 networks for the microarray). (**B**) The average GO term AUROCs across both the two networks is significant ($P \sim 7.44e{-}15$). Inset shows the correlation between the two GO AUROCs distributions plotted ($\rho = 0.84$, $P < 1e{-}16$), showing that GO groups that perform well in one, perform equally well in the other network

3.3 Aggregating across many networks improves performance and removes noise

As mentioned, a solution to the relatively poor performance of individual co-expression networks has traditionally been meta-analysis across multiple datasets. We performed this across both RNA-seq and microarray networks, by rank standardizing within each weighted network and then summating. We perform this in a stepwise manner (Fig. 4A) and evaluate performance after the addition of each network. A notable point is that both RNA-seq and microarray networks appear to converge to similar performance measures (average AUROC ~0.7). While ~20 separate experiments (with moderate depth and sample count >10) is sufficient to obtain a high-quality network (AUROC >0.6) as captured by GO, to comprehensively capture unknown functional information may require >50 separate experiments. This is further suggested by a slower convergence to a fixed node degree under aggregation (Supplementary Fig. S5A). Aggregation across experiments improves performance well beyond that attained by even the largest individual experiments with far fewer samples.

One concern relating to the use of aggregate network data is that it is harder to accumulate data to specifically tune the network to any particular experimental context (e.g. tissue), which is one major advantage of co-expression over, e.g. protein–protein interaction data. If individual experiments are tuned for some function, we should expect their performance to rise above the aggregate over some fraction of the assessed functions. However, a majority (56%, Fig. 5A) of GO terms never outperform the aggregate in any experiment. This effect is particularly strong when we consider that the individual networks could simply outperform the aggregate due to noise combined with their having multiple tests for each function. Indeed, only 46 (out of 2953) GO groups perform better on average in the individual datasets than the aggregate, and these functions are ones with very poor performance in both the aggregate and underlying data (average AUROC <0.52 in the individual datasets, Supplementary Table S3). Similarly, for any given experiment, the fraction of its GO terms which outperform the aggregate is very small, reaching a maximum of 17% only in very large experiments and averaging only 5% of GO terms (again, those with low performance, Fig. 5B).

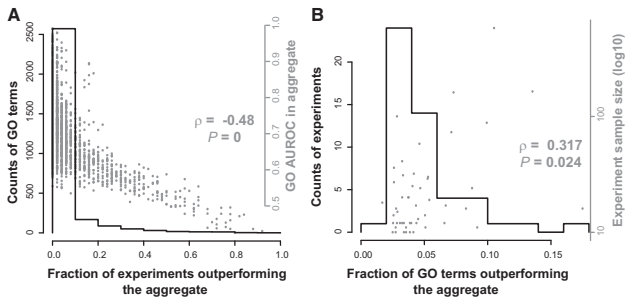

Fig. 5. Comparing individual RNA-seq GO term performances to the aggregate. (**A**) For each GO term, we compared the number of times that GO term performed better in the 50 experiments than within the aggregate. There were a few GO terms where nearly all the experiments outperformed the aggregate, but these occur only at very low AUROCs overall. (**B**) Examining the number of experiments where a GO term outperforms the aggregate shows that no individual experiment has a large fraction of GO terms doing so. There is a weak correlation with sample size ($\rho = 0.317$, $p = 0.024$), likely explained by the higher performance of such networks overall

As another check to our use of the particular expression estimators (i.e. Cufflinks, RSEQtools and RSEM), we generated aggregates of the initial 10 subset experiments, and compared their node degrees and performance amongst each other. Even though the individual experiments had weaker correlations, the correlations between the node degrees of the aggregates of each of the methods were generally high ($\rho = 0.54$–0.84). The aggregates of only those 10 experiments also all had similar functional connectivity as determined by GO, with average AUROCs at 0.67. The correlations between the GO AUROCs were also high, ranging between $\rho = 0.86$ and 0.88 (Supplementary Table S5).

In order to check whether or not high-quality networks are necessary to generate an aggregate that has high performance, we split 46 RNA-seq networks into two groups based on their individual GO performances (i.e. their average AUROC). The top 23 performing networks were then aggregated into a 'top aggregate network' and this aggregate's performance was compared with the aggregate of the remaining 23 performing networks. We compared these results with the randomly sampled sets of 23 networks that were generated for the previous analyses. The average AUROC performance of the random experimental sets ranged between 0.69 and 0.71 (mean average AUROC = 0.70, SD ~0.01). The top group had the highest performance (average AUROC = 0.71), only just above the range of the random sampled, while the bottom 23 group had a much lower performance (average AUROC = 0.66). Even though this is lower than the random sets of experiments, it still outperforms almost all of the individual networks (equal to the best performing), indicating that even the weakest networks are useful in aggregate (Supplementary Fig. S5B).

To provide a better mechanistic explanation of aggregation's utility, we examined GO groups that benefited strongly from aggregation and attempted to trace the topological dependencies underlying their performance; we discuss an exemplary case in the following, but we believe the trends are likely general. The 'collagen' complex (GO: 0005581) gives moderate performance in the underlying datasets (mean AUROCs ~0.7), but benefits strongly from aggregation (AUROC >0.9). It is characteristic of groups benefiting from aggregation in exhibiting moderate performance even before aggregation and, interestingly, in being a protein complex. Taking all the protein complex subgroups in GO in our analyses, their performance on average in the aggregate was AUROC ~0.77, significantly greater than the remaining terms (Wilcoxon test, $p \sim 5.423e{-7}$) and complexes were enriched when ranking terms by the degree to which they benefited from aggregation (Wilcoxon test, $p \sim 0.01$).

Performance in the collagen complex is very broadly distributed both between gene pairs and across different datasets. In the aggregate, almost half (47%) of the gene pairs exhibit high ranking (top 10%) co-expression, while a proportionately smaller fraction (~2%) exhibit very highly ranked (top 1%) co-expression. This suggests that performance primarily arises through weakly distributed interactions, which we verified by removing the connections ranked in the top 1% and seeing no significant drop in performance whereas removal of the connections in the top 10% had a large effect (AUROC ~0.6). This diffuse improvement between gene-pairs depends on a large fraction of the data for each connection. On average, a given connection in the aggregate is only more tightly connected than in 35 out of the 50 underlying datasets (where connection strength is standardized by rank within all networks). This suggests, once again, that the very high performance in aggregate is distributed throughout many moderately significant pairwise relationships. Altogether, our examination of the collagen complex supports the view that aggregation will work best in functions where pairwise relationships are expected to be distributed (e.g. protein complexes rather than dynamic pathways) and weak aggregate effects can summate to provide a strong total signal.

3.4 Machine learning methods do not confound network performance results

We re-assessed our aggregate network using a series of other machine learning algorithms to establish whether the performance of the co-expression network would be affected by the methodology used to extract the biological signal. It is possible that more biological data might be present if we look to subtler data features, using more sophisticated algorithms. We used logistic regression, random forest, SVM using stochastic gradient descent, passive aggressive and GeneMANIA, as previously analyzed in yeast data (Verleyen *et al.*, 2014). We ran each of these algorithms on the aggregate RNA-seq co-expression network with GO slim (109 terms), and assessed the performance using the average AUROC as previously described. The average AUROC varied slightly between methods (Supplementary Table S4), with the neighbor voting, logistic regression and stochastic gradient descent giving the highest performance (average AUROC = 0.69), and the random forest performing the worst of the six methods (average AUROC = 0.63). However, all the methods perform well on similar GO groups as neighbor voting (performances correlated at $\rho = 0.87$–0.96, $P < 2.2e{-16}$), implying that the results are not strongly method independent. While additional data will doubtless render new properties of genes learnable, our complete list of reported performances (Supplementary Table S3) provides a useful prior as to how easily different functions may be expected to be read out from network data.

3.5 Microarray and RNA-seq-derived networks have different hub genes

Although the aggregate networks between the microarray and RNA-seq technologies appear to converge as apparent through their performance metrics, we wished to ascertain how similar the connectivity was between the two. To our surprise, the correlation between the node degree between the RNA-seq aggregate and the microarray aggregate was negative ($\rho < 0$) across all methodologies, implying that many highly connected genes in one network were now no longer well connected in the second. Indeed, hub-like genes

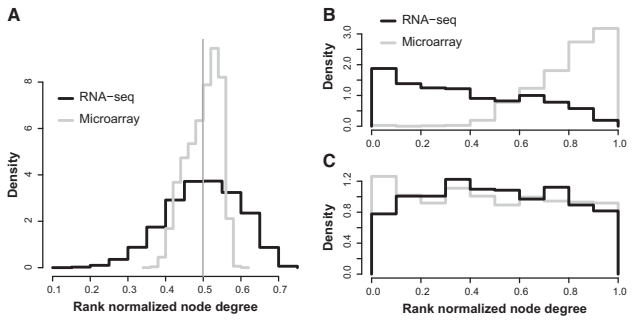

Fig. 6. Network topology differences are due to low expressing genes in RNA-seq. (**A**) The range of node degrees between the two aggregate networks is much higher in the RNA-seq co-expression aggregate (black lines) than microarray (grey lines), with an SD ∼0.266 compared with SD ∼0.144, respectively. (**B**) We can see the difference quite pronounced in the bottom third expressed genes (730 genes) in the RNA-seq experiments that overlapped across the 50 individual experiments. (**C**) Taking a comparable subset of low expressed genes from the microarray, we do not see the same effect of low expressing genes on node degree

(e.g. top 3%, ∼500 genes) were significantly dissimilar in the two networks ($P < 1.4e{-}9$). Even so, both networks capture individual GO functions in a similar way ($P = 0.84$, Fig. 4B and inset). This dissimilarity suggests again that while inference methods average away topological noise, the actual networks exhibit it to a substantial degree and this could be easily altered by varying methodology in other ways.

In order to determine if this aggregate trend toward topological reordering affects individual experiments, we assessed the correlation between the node degrees of the aggregate network and each of the individual networks (in both platforms). We used the 30 networks from the microarray aggregate, and the 50 networks of the RNA-seq aggregates, and calculated the node degree correlations to the final aggregate and cross platform aggregate (excluding the dataset itself from the self-platform aggregate as a form of cross validation). Overall, cross platform correlations are much lower (RNA-seq networks to microarray aggregate average $\rho = -0.02$, microarray networks to RNA-seq aggregate average $\rho = 0.06$) than the correlations between the aggregate and the individual networks of the same platform (RNA-seq networks to RNA-seq aggregate average $\rho = 0.38$, microarray networks to microarray aggregate average $\rho = 0.26$). This indicates that the negative correlation between microarrays and RNA-seq data is a recurring feature of the individual networks, and not just the aggregates, and suggests a feature due to technological biases (in conjunction with network construction methods, Supplementary Fig. S5C).

Unlike in microarrays where we definitively know some genes are not measured or present on the platform, the situation is more ambiguous in RNA-seq. As a null network to our aggregate, we generated a co-occurrence network by weighting edges based on the co-occurrence of pairs of genes detected in the network, disregarding their correlations; in essence, a network reflecting only the fact that two genes were both measurable in a given experiment (and then aggregated across experiments). The correlation between node degrees of this network and the true co-expression network ($\rho = 0.45$, Supplementary Fig. S5D) implies that there is a strong influence of the presence of a gene pair contributing to a higher node degree correlation. Although the node degrees of the co-expression and co-occurrence networks correlate, the functional connectivity of the co-occurrence network is very poor (average AUROC = 0.56, Supplementary Fig. S2H).

3.6 Noisy expression variation dominates topology outside of functional connectivity

Our results up to this point exhibit a dichotomy: RNA-seq co-expression and microarray co-expression are very similar in their functional properties but very dissimilar in their overall topology (Fig. 6A). Moreover, RNA-seq is most robust in its topology where artifacts likely dominate, such as in within given pipelines or co-occurrence in meta-analysis. One possibility to explain the difference between the microarray and RNA-seq aggregate networks would simply be that different data went into each of their construction. We tested this by looking specifically to a large dataset replicated in both microarray and RNA-seq, the BrainSpan atlas (2011). Here, again, performances are similar in microarray and RNA-seq (mean GO AUROCs of 0.64 and 0.65, respectively), but topology is starkly different with a correlation of $\rho = -0.66$ between the node degrees of the two networks constructed. Altogether, this strongly suggests that technical variation drives the topological differences between microarray and RNA-seq.

To test this hypothesis, we used the recent MAQC-III (SEQC) results providing consensus measures of replicability in expression analysis (Li *et al.*, 2014; SEQC/MAQC-III Consortium, 2014). SEQC provides criteria under which transcript expression estimates are not expected to be reliable in RNA-seq pipelines (with moderately low expression—the bottom one-third of transcripts—being a major threshold). For each expression dataset, we determined which transcripts would fail their threshold, and we then examined the topological properties of the subset of transcripts which would fail in at least 80% of the experiments underlying the aggregate. This set of transcripts exhibits starkly different topological properties in the RNA-seq co-expression and microarray data, being very high node degree within microarray and very low in RNA-seq (Fig. 6B). Because these transcripts contribute many hub-like genes to the microarray data, this effect is fully sufficient to account for the lack of overlap we see. If we use the same criterion to define a set of transcripts (failing SEQC guidelines but in microarray), we see no such difference in node degree between microarray and RNA-seq co-expression (Fig. 6C). Altogether, this strongly suggests that the topological reordering we see between microarray and RNA-seq reflects noisy expression, and further, that RNA-seq is both more sensitive at detecting when this occurs (when expression is too low to be reliable) and in not generating spurious correlation among these noisy values (low node degree).

4 Discussion

The appropriate way to assemble and assess RNA-seq data is still a topic of considerable controversy. While it might seem that these methodological discussions would need to be resolved to put RNA-seq data to use in co-expression analysis, we suggest that is not so. Co-expression has frequently relied on the aggregation of diverse data and so many of the problems that affect RNA-seq as a pure measure of transcriptomic activity may not have downstream importance in co-expression use. Our results give a perhaps surprisingly consistent message: simple, basic approaches were among the highest performing, from measuring network connectivity (Spearman correlation) to functional inference (neighbor voting). Similarly, a simple rank-and-summate aggregation across multiple datasets generates a more robust network, capturing known biology with high efficacy.

In addition, we have demonstrated that individual RNA-seq networks exhibit a sample size dependency very similar to microarray data, suggesting that technical issues are not at the heart of the 'noisiness' of co-expression data. Also important is the dependence on read

depth in RNA-seq co-expression. Our study suggests that 1 M reads per sample (with many samples) is sufficient to build a network that is slightly divergent from random and adding such networks has no negative impact on the performance of the aggregate network. Thus, when aggregating data, it appears sensible to be extremely lenient in constructing inclusionary criteria. Recent suggestions for best RNA-seq practices in differential expression (Liu *et al.*, 2014) highlight a power improvement in adding biological samples in a tradeoff with read depth. Users of co-expression data cannot naively apply those recommendations because the minimum number of samples is already relatively high, and robustly measuring correlations may require an even greater increase. In co-expression analysis, increasing depth past 100 M aligned reads per sample and even thousands of samples continue to add value if our focus is on genome-wide functional inference for a given dataset. So profound is the performance difference between aggregate and individual networks that it is difficult to recommend any minimum depth or sample count which is adequate; certainly none in our case come close to producing the same apparent performance. However, more 'focused' uses (e.g. studying changes in expression across brain development) may be permitted by more narrowly constructed datasets. While our analysis does not find evidence to support the view that such uses will be higher performing in any way, this may reflect our reliance on GO as a reference, even for narrowly defined functions.

Since we have principally been concerned with measuring our network data in ways which generalize broadly, our use of the GO as a way of defining 'function' is perhaps the most easily critiqued, albeit hard to avoid and common. While it is very encouraging that many of the methodological variations in RNA-seq analysis do not negatively affect functional inference or render it less robust, it is natural to ask whether our measure of function is adequate. Since our results were virtually identical using KEGG and Reactome as references, we suggest our results are likely robust to most choice of reference data. Further to this, the machine learning algorithm selected has little influence on the results, which also implies that the features being selected to make predictions, determine connectivity and assess the networks are adequate for the task at hand. In addition, we think it is important to keep in mind that our evaluation of co-expression GBA is intended to be a proxy for comparative estimates of functional properties within the network in general. It would have been relatively easy for us to inflate our performance by incorporating a stronger prior based on semantic data or GO itself, for example. Instead we used methods that we think are strongly likely to generalize to novel tasks at some cost to absolute performance.

Disease studies that use a co-expression network to derive interaction partners and likely disease gene candidates typically do so through measures such as node degree and modularity (for instance, see Parikshak *et al.*, 2013). Our interpretation of the topological shuffling we see between RNA-seq and microarray is not that they are robustly different (even though the significance is high), but rather that the measurements are easily influenced by weak biases (e.g. gene size). The question is whether the bias is specific to one technology, and if so, which one. Our analysis suggests that the reordering reflects the tendency for low-expressing genes to be correlated in microarray data, but that RNA-seq allows us to identify which genes are 'too' low to be trusted with much greater specificity. Further, the low impact of variation in node degree on semi-supervised learning (e.g. through GBA) suggests that even without the appropriate filtering, many methods adequately identify the absence of learnable features in noisy expression.

Beyond simple read depth and sample size cut-offs, we suggest a few 'bright line' rules would likely advance network analysis. First, any single dataset analysis intended to present targeted biological results should provide comparative assessment on an aggregate network as a control. This would help disentangle data and methodological dependencies when presenting results. Second, the guidelines provided by SEQC should be applied as a default for individual datasets to diminish noise and ensure stability. In meta-analyses, we suggest using an approximation of the SEQC guidelines as a uniform filter; e.g. all transcripts with expression in the bottom third for 80% or more of the samples be removed from the experiment. Third, for smaller sample sized experimental data, more cautious statistical control can help better guard against artifacts; we particularly recommend the use of non-parametric statistics and resampling as baseline methods when an argument in favor of robustness cannot rest on meta-analysis.

In summary, RNA-seq offers unique virtues but is not a technical panacea. Co-expression analyses will still require many samples and a reliable quality reference network as a control and comparative measure. Because of our reliance on pre-existing knowledge (GO) or data (microarray) to provide reference knowledge, we cannot readily assess RNA-seq where its application is likely to prove most novel and exciting, e.g. novel or more diverse transcript assessment. Unsupervised interpretation of RNA-seq data connectivity is unlikely to be robust to underlying methodological choices without careful filtering, while methods that extract signals using training data (supervised or semi-supervised methods) appear to safely recover known information. We make our networks available as a resource for other researchers at http://gillislab.labsites.cshl.edu/supplements/rna-seq-networks/.

Acknowledgements

We thank Paul Pavlidis for helpful comments on the manuscript and Sanja Rogic for assistance using the Gemma RNA-seq data.

Funding

J.G., S.B. and W.V. were supported by a grant from T. and V. Stanley. No funding source played any role in the design, in the collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

BrainSpan atlas (2011) BrainSpan: Atlas of the Developing Human Brain [Internet]. Available from: http://developinghumanbrain.org (9 March 2015, date last accessed)

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Crammer,K. *et al.* (2006) Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, **7**, 551–585.

Croft,D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

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

Faith,J.J. *et al.* (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.

Fan,R.E. *et al.* (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Frazee,A. *et al.* (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

Gillis,J. and Pavlidis,P. (2011a) The impact of multifunctional genes on 'guilt by association' analysis. *PLoS One*, **6**, e17258.

Gillis,J. and Pavlidis,P. (2011b) The role of indirect connections in gene networks in predicting function. *Bioinformatics*, **27**, 1860–1866.

Habegger,L. *et al.* (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, **27**, 281–283.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.

Hitzemann,R. *et al.* (2013) Genes, behavior and next-generation RNA sequencing. *Genes Brain Behav.*, **12**, 1–12.

Iancu,O.D. *et al.* (2012) Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics*, **28**, 1592–1597.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Leinonen,R. *et al.* (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.

Li,B. and Dewey,C. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,S. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotech.*, **32**, 888–895.

Liu,Y. *et al.* (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304.

Margolin,A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Mostafavi,S. *et al.* (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9**(Suppl. 1), S4.

Newman,M.E. (2003) The structure and function of complex networks. *SIAM Rev.*, **45**, 167–256.

Oliver,S. (2000) Proteomics: guilt-by-association goes global. *Nature*, **403**, 601–603.

Parikshak,N.N. *et al.* (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, **155**, 1008–1021.

Pena-Castillo,L. *et al.* (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.*, **9**, S2.

Peng,H. *et al.* (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.

Sales,G. and Romualdi,C. (2011) parmigene—a parallel R package for mutual information estimation and gene network reconstruction. *Bioinformatics*, **27**, 1876–1877.

Sekhon,R.S. *et al.* (2013) Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PLoS One*, **8**, e61005.

SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotech.*, **32**, 903–914.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Verleyen,W. *et al.* (2015) Measuring the wisdom of the crowds in network-based gene function inference. *Bioinformatics*, **31**, 745–752.

Voineagu,I. *et al.* (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380–384.

Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–1701.

Zhang,A. (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *International conference on Machine Learning*. Proceeding. ICML '04 Proceedings of the twenty-first international conference on Machine learning, p. 116. http://dl.acm.org/citation.cfm?doid=1015330.1015332.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, 1128.

Zoubarev,A. *et al.* (2012) Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, **28**, 2272–2273.