PEAK Predicts Gene Regulatory Network Linkages during Sea Urchin Development with High Sensitivity from Gene Expression Data Alone

Jingyi Zhang
Virginia Tech: Virginia Polytechnic Institute and State University
Farhan Ibrahim
Virginia Tech: Virginia Polytechnic Institute and State University
Doaa Altarawy
Virginia Tech: Virginia Polytechnic Institute and State University
Lenwood S Heath
Virginia Tech: Virginia Polytechnic Institute and State University
Sarah Tulin (tulins@canisius.edu)
Canisius College https://orcid.org/0000-0001-8365-0559

Research

Keywords: Strongylocentrotus purpuratus, gene expression, gene regulatory networks, prediction, machine learning

Posted Date: January 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-142579/v1

Abstract

Background

Gene regulatory network (GRN) inference can now take advantage of powerful machine learning algorithms to predict the entire landscape of gene-to-gene interactions with the potential to complement traditional experimental methods in building gene networks. However, the dynamical nature of embryonic development – representing the time-dependent interactions between thousands of transcription factors, signaling molecules, and effector genes – is one of the most challenging arenas for GRN prediction.

Results

In this work, we show that successful GRN predictions for developmental systems *from gene expression data alone* can be obtained with the Priors Enriched Absent Knowledge (PEAK) network inference algorithm. PEAK is a noise-robust method that models gene expression dynamics via ordinary differential equations and selects the best network based on information-theoretic criteria coupled with the machine learning algorithm Elastic net. We test our GRN prediction methodology using two gene expression data sets for the purple sea urchin (*S. purpuratus*) and cross-check our results against existing GRN models that have been constructed and validated by over 30 years of experimental results. Our results found a remarkably high degree of sensitivity in identifying known gene interactions in the network (maximum 76.32%). We also generated 838 novel predictions for interactions that have not yet been described, which provide a resource for researchers to use to further complete the sea urchin GRN.

Conclusions

GRN predictions that match known gene interactions can be produced using gene expression data alone from developmental time series experiments.

Background

In the 1950s and 1960s, it began to be clear to researchers that gene products representing the sum of the entire genome were not present in every cell and, therefore, gene expression was being regulated (Mirsky 1951; Britten and Davidson 1969). The exciting discovery of the role of transcription factors in the process of gene regulation set off a new era of exploration of the roles of genes in phenotypes, disease, evolution, and embryonic development (Jacob and Monod 1961). Gene regulation can be organized and modeled as a hierarchical network, a Gene Regulatory Network (GRN), as first proposed by Eric Davidson (Arnone and Davidson 1997; Davidson et al. 2002a). GRN models are now routinely used to follow the causal links between regulatory genes that lead to cell fate decisions or cell activities. GRN models are also used to create hypotheses about the function of players in a regulatory program.

Beginning with the construction of the sea urchin endomesoderm GRN (Davidson et al. 2002a, 2002b), many GRNs in different model systems have been compiled through extensive experimental

perturbations often involving a combination of knockdown techniques combined with visualization of changes in gene expression. Accuracy of the GRN is improved when further experiments confirm *cis*-regulatory interactions at the level of transcription factor binding sites. However, there is now a pressing need to facilitate GRN modeling using computational prediction tools to help fill in the gaps in existing GRNs and to help create new GRN models with testable interaction predictions, especially in systems with limited resources.

Various GRN prediction algorithms from gene expression data have been proposed (Delgado and Gómez-Vela 2019), and several have been compared and tested through the DREAM consortium (Marbach et al. 2012a). The overall accuracy of these methods using gene expression data alone is usually below 50% even with a consensus of several computational methods. The use of additional prior information, such as transcription factor binding data, Gene Ontology (GO) annotation, functional association, protein-protein interactions (PPIs), and experimentally verified relations, was shown to improve prediction accuracy (Altarawy et al. 2017). However, this information is not always available for every new model system that could benefit from GRN inference.

For our analysis, we use the Priors Enriched Absent Knowledge (PEAK) network inference algorithm to reconstruct GRN interactions from gene expression data (Altarawy et al. 2017). PEAK uses differential equations, CLR (Context Likelihood of Relatedness), and the machine learning method Elastic net to predict the likely interactions between target genes and transcription factors. PEAK consists of two phases, a coarse-grained phase and a fine-grained phase, to predict a GRN. In the coarse-grained phase, potential regulators for each gene are extracted using mixed context likelihood of relatedness (mixed CLR). In the fine-grained phase, two modified versions of Elastic net are used to refine the predictions and to integrate curated or noisy prior knowledge, when available. Prior knowledge about the network can also be added if available; however, none has been used in this study.

We chose two sea urchin embryonic GRNs governing endomesoderm and ectoderm specification as test networks, because they are widely regarded as some of the most well-supported developmental GRN models with many *cis*-regulatory interactions verified at the base-pair level. We also obtained two sea urchin gene expression data sets to use as input. The California purple sea urchin, *Strongylocentrotus purpuratus*, is a marine invertebrate. The sea urchin is a member of the phylum Echinodermata, which, along with the Hemichordata, are the closest known sister groups to the Chordates, the phylum to which humans belong. The genome of *S. purpuratus* was sequenced in 2006, which produced an estimated gene count of 23,000 (Sea Urchin Genome Sequencing Consortium et al. 2006). The sea urchin develops rapidly from a single-cell (the fertilized egg) to a late gastrula embryo in 48 hours at 14^{o}C and then into a prism-shaped larvae by 72 hours. As an indirect developing animal, the sea urchin larvae must undergo metamorphosis to achieve its final adult structure. Over the last 20 years, GRN models describing the regulation of embryonic development of *S. purpuratus* have been built by experimentation and collaboration of sea urchin researchers around the world. The GRN models describing development are divided into the network describing the ectodermal tissue layer, which will give rise to the nervous system and outer tissues of the larvae, and the network controlling the endomesodermal tissue layer, which will

give rise to the gut and the larval skeleton. The most recent versions of the two GRN models are hosted at BioTapestry, accessible at http://grns.biotapestry.org/SpEndomes/ for the endomesoderm network and at http://grns.biotapestry.org/SpEcto/ for the ectoderm network (Longabaugh et al. 2005).

The goal of our approach is to employ PEAK to predict gene regulatory interactions from gene expression data alone. For many emerging model systems crucial for understanding evolution, gene expression data from developmental time series is readily available but, due to limited resources, these systems lack extensive prior information regarding transcription factor binding sites, ChIPseq data, and proteomics. Therefore, these emerging model systems have a demonstrated need for GRN inference to guide future experiments on regulatory interactions during embryonic development (Tulin et al. 2013; Fernandez-Valverde et al. 2018).

Previous large scale assessment of network inference methods aimed at predicting gene networks using gene expression data alone have a maximum of 50% precision using 800 microarray experiments (Marbach et al. 2012a) and much less accuracy using 300 experiments. Our method of using the PEAK algorithm on gene expression data from developmental time series data alone was able to achieve a maximum of 76.32% sensitivity using 98 experiments.

Results

Differential gene expression analysis

A complete GRN model would need to include the entire set of expressed transcripts whose expression is controlled by the regulatory apparatus. To begin to define the regulatory gene set to input into the PEAK machine learning algorithm, we started with the sea urchin RNAseq transcript data set covering 0-72hpf (Tu et al. 2014). The RNAseq transcript data set was then filtered to identify the set of transcripts that are differentially expressed during embryonic development and are regulative in nature.

We used three programs (NOISeq, EdgeR, and GFold) to determine the set of differentially expressed genes (DEGs) with the parameters described in the methods section. There was variation in the number of DEGs as determined by NOISeq, EdgeR, and GFold on the transcriptome RNAseq data. We compared the overlap of genes above the threshold identified by each method to obtain a core set of DEGs (Fig. 1). There were 20,422 total unique differentially expressed transcripts determined by 5 methods (2 different thresholds for both GFold and NOISeq, and one threshold for EdgeR). Only 10,627 genes were consistently specified as differentially expressed no matter which method was used. We found that the result from NOISeq ($\lambda_1 = 0.9$) contained the most overlap and the least difference with results from the other methods while maintaining a more selective number of genes determined to be differentially expressed (Table 1). Only .01% of the genes determined by NOISeq ($\lambda_1 = 0.9$) are unique to that method; in contrast, more than 5% of the genes from the result of EdgeR are unique to EdgeR. All the genes determined by GFold ($\lambda_4 = \mp 1.5$) overlapped with the results from other methods, but the GFold DEG set was missing 2,755 genes that were identified as differentially expressed by the other 4 methods.

Table 1
Differential gene expression analysis results summary from three programs (NOISeq, GFold, and EdgeR) using two different thresholds for NOISeq and GFold.

Program	Total DEGs	# Unique Genes	% Unique Genes	# Overlap Genes	% Overlap Genes
NOISeq ($\lambda_1 = 0.9$)	15496	2	0.01%	15494	99.99%
NOISeq ($\lambda_2 = 0.85$)	16996	411	2.42%	16585	97.58%
GFold 1 ($\lambda_3 = \mp 1$)	17046	16	0.09%	17030	99.91%
GFold 2 ($\lambda_4 = \mp 1.5$)	12950	0	0.00%	12950	100.00%
EdgeR	19166	988	5.15%	18178	94.85%

Gene Ontology Filter

The NOISeq filter q_value > 0.9 reduced the 21,092 total transcripts to a set of 15,496 differentially expressed transcripts. However, the set of DEGs identified by NOISeq ($\lambda_1 = 0.9$) is still too large to be effectively used as input into the PEAK prediction algorithm. Therefore, we applied a second filter to the gene set to achieve a more appropriate number of regulatory genes. The second filter was a Gene Ontology (GO) filter for genes related to transcription and signaling. We used the custom GO annotation generated by the authors of the transcriptome (Tu et al. 2012). The GO filter identified 1,038 transcripts that are regulatory in nature, of which 544 were also identified as differentially expressed (Additional file 1). The 544 transcripts represent 504 unique gene models annotated by a single SPU_ID.

PEAK analysis

We next applied the PEAK GRN prediction algorithm on the filtered gene set of 504 DEGs as determined by NOISeq that are also identified by the GO filter. We specified to PEAK the set of 258 Transcription Factors (TFs) used during embryonic development according to a compilation of genes specified as TFs by Materna et al. (Materna et al. 2010) and genes specified as TFs according to the custom GO annotation by Tu et al. (Tu et al. 2012) (Additional file 2). We used the filtered gene set and TF specification file as inputs (Additional file 3) into PEAK, and set all the other parameters as default except the “Repeat” as 1. The output from PEAK contained 14,802 predicted interactions in total (Additional file 4). Among the set of predictions from PEAK representing the top 5 hits for each gene there are both known interactions matching the ground truth network and new predictions (Fig. 2).

To evaluate the performance of PEAK, we repeated the PEAK analysis with data relating only to genes present in the ground truth sea urchin GRNs using the same procedure and parameters as with the set of all DGE genes that fit the GO filter. In this analysis, we used three measures to perform the evaluation. *Sensitivity* represents the proportion of our predicted gene interactions that hit the corresponding ground

truth GRNs, indicating the percentage of gene interactions that are correctly identified by the PEAK algorithm among the total known ones. Because there are still numerous undiscovered gene interactions and other complexities of GRNs such as transient binding, multiple upstream regulators, and indirect interactions (Van den Broeck et al. 2020), *new predictions* designates the proportion of new gene interactions predicted by the algorithm which is not yet known in the ground truth GRN. And the *miss rate* represents the proportion of known gene interactions not discovered by the algorithm.

We evaluated the gene set from the ectoderm GRN and the endomesoderm GRN separately and compared the results as summarized in Table 2. For the known ectoderm GRN, there are 36 genes out of 39 that are identified as DEGs in the transcriptome RNAseq data. When the gene expression data from the transcriptome RNAseq data set is used as input for PEAK, the algorithm successfully predicted 45 of 78 gene-to-gene interactions present in the ground truth GRN, yielding 57.69% sensitivity (Table 2). PEAK failed to predict 33 known connections but provided 555 new possible ones. The sea urchin endomesoderm GRN is currently a larger network model, with 58 genes and 146 edges. We found 57 of those 59 genes were identified as being differentially expressed by NOISEq. When TFs were specified, 83 of the 142 connections were correctly predicted for a sensitivity of 58.45% (Table 2).

Table 2
Statistic result for transcriptome data and the Nanostring data compared with ground truth GRNs.

Data set	**Transcriptome RNAseq data**		**Nanostring data**	
Ground Truth GRN (GT)	Ectoderm GRN	Endomesoderm GRN	Ectoderm GRN	Endomesoderm GRN
True predictions (TP)	45	83	29	58
Sensitivity	57.69%	58.45%	50.88%	76.32%
Miss rate	42.31%	41.55%	49.21%	23.68%
New predicted edges	555	1110	412	838
Total predictions	600	1193	441	896

To achieve a higher sensitivity, we considered what limitations might be present in the approach. A limitation of the transcriptome data when applied to the ground truth GRN is that only 10 timepoints were sampled in total and only 5 of those timepoints covered the period of time during development which the ground truth GRN models describe. Therefore, we sought out a second data set with more timepoints at closer sampling intervals. For the second data set, we chose a high-density embryonic data set where 161 genes critical to early development were sampled once an hour for 48 hours in duplicate and gene expression was quantified with Nanostring technology (Materna et al. 2010). Because there are genes in the ground truth GRNs that do not have a match in this second data set, we only retained the genes that appear in both the Nanostring data set and the ground truth GRNs. For the ectoderm gene set, we had a match for 28 of 39 genes, so the number of relevant gene interactions in the ground truth ectoderm GRN was reduced from 85 to 57. With the small number of input genes, we found that PEAK gave similar

prediction results. The result was 29 known gene-to-gene interactions were successfully predicted out of 57, yielding a sensitivity of 50.88% (Table 2).

For the endomesoderm data, 41 genes were present in both the ground truth GRN and the Nanostring data set, which reduced the 146 total known interactions to 76 known interactions. When TFs were specified, a total of 58 gene-to-gene interactions were successfully predicted out of 76, for a sensitivity of 76.32% (Table 2). Because more known connections in the endomesoderm ground truth GRN are supported by solid evidence at base-pair resolution and the Nanostring data set had far more experimental time points, it was expected that PEAK predictions using the Nanostring data set would yield the highest sensitivity for the endomesoderm GRN.

We further tested the impact of adjusting the parameter describing transcript turnover due to maternal and zygotic degradation mechanisms in terms of transcript half-life on the prediction results. Recent estimates of transcript turnover in sea urchin are in the range of 6 to 9 hours (Gildor et al. 2016). We explored the performance of PEAK for a range of median half-life times from 3 hours to 15 hours (Fig. 3). In general, the sensitivity of the predicted results did not have any obvious increase or decrease trend with the increase of half-life. We obtained the highest sensitivity on the transcriptome data when the half-life was set to 7 hours for ectoderm genes. But when the half-life was set to 3 hours, we obtained the highest sensitivity on the Nanostring data set. Particularly when we used the set of 45 endomesoderm genes present in the Nanostring data, the sensitivity of the prediction was 76.32%.

Discussion

The computational prediction of a GRN has been explored in many organisms. In bacteria, for example, researchers used an integrative method to predict a GRN in *B. subtilis* with a large amount of input data (Arrieta-Ortiz et al. 2015). The *B. subtilis* study used more than 600 gene expression experiments and incorporated prior knowledge from the ground truth network to improve accuracy. The sensitivity of their GRN prediction is 74%, and they predicted 2,258 new regulatory interactions. The scale of experiments used in the Arrieta-Ortiz study is difficult to achieve in multicellular organisms. As is true with most other machine learning applications, the more prior knowledge available to train the algorithm, the better predictions one can expect the algorithm to produce. However, our approach using PEAK was able to yield 76% sensitivity using only gene expression data and no other prior knowledge.

Some organisms benefit from extremely large research communities producing extensive prior knowledge data sets in the form of ChIP assays, protein-protein interaction databases, functional gene annotation, extensive tissue expression information, TF binding sites, and known regulatory interactions. In these cases, the richness of prior knowledge has been harnessed to produce high quality novel network interaction predictions (Marbach et al. 2012b). However, the future of many disciplines will require knowledge from a wider variety of model species to uncover universal truths and mechanisms. The first “-omics" step in many research programs establishing new model species is often the creation of a transcriptome, which inherently creates a quality source of gene expression data (Henry et al. 2010, 2017;

Du et al. 2012; Helm et al. 2013; Tulin et al. 2013; Chen et al. 2014). New model species are especially important in evodevo as these emerging species are all poised to increase our understanding of how metazoan body plans evolved. The investigation of how changes in the gene regulatory program controlling embryonic development have shaped evolution is a particularly active research question and one that can benefit most directly from using GRN inference approaches to create model networks. Therefore, our approach of only using gene expression data as input to generate regulatory interaction predictions is vital to the success of future research.

One goal of GRN prediction is to narrow the search space for edges to a subset of promising interactions to be further studied and validated experimentally. With 23,300 gene models in the sea urchin, there are ~ 540 million possible gene-to-gene interactions. By using time series gene expression data as input into the PEAK prediction algorithm for the 547 transcripts most likely to be a part of the regulatory program, we generated 14,802 predicted interactions that can be ordered by confidence or searched for specific regulators or target genes. One caveat to our approach is the limitation that PEAK cannot predict self-interactions where genes turn on or off their own expression. These self-interactions are known to be important to the precision and robustness of regulatory programs. Experimental design should take this limitation into account and make sure to check for self-regulation when investigating the targets of regulatory genes.

We also tested the ability of PEAK to make predictions using our data sets without specifying the set of TFs. Surprisingly, the sensitivity was not dramatically different than when the list of TFs was specified. Since the performance of PEAK was not overtly affected when the list of TF was not specified to the algorithm, those model systems where functional gene annotations are incomplete can still make use of our method. It is also possible to specify the transcript half-life to PEAK; however, we found that using the half-life value that most closely corresponds to previously published values for the organism did not always result in the best prediction result. Therefore, our recommendation is to use a range of half-life values to find the best fit to the data experimentally.

Conclusion

We found that using gene expression data as the sole input for machine learning GRN predictions was sufficient to generate predictions that match known regulatory interactions when using the PEAK program. Our approach also generated new possible gene-to-gene interactions that are not currently described. The new predictions from our pioneer dataset will serve as a resource for the sea urchin community; a relatively small, but influential group, in the areas of *cis*-regulatory biology and developmental regulatory networks. Our future research will include applying a similar approach to an emerging model species that transcriptomic gene expression data but does not yet have a developmental GRN model in order to generate predictions for new regulatory interactions. Our method is broadly applicable to any organism and is more accessible due to the ability to start with gene expression data that is often readily available. Although no prior knowledge is required, PEAK can accept many forms of prior knowledge to improve the quality of predictions (Altarawy et al. 2017). Especially in emerging

evodevo model systems, where gene expression datasets are often the first to be produced, PEAK shows promise to be able to assist network building efforts.

Methods

Data sets

Two gene expression data sets were obtained to use as input into the PEAK algorithm. The first data set comes from the sea urchin transcriptome project (Tu et al. 2012) where 10 embryonic timepoints (labeled with time in units of hours-post-fertilization (hpf)) were assayed for transcript expression by RNAsEq. The sequenced transcripts in the transcriptome data set derived from cDNA collected from: (1) the unfertilized egg, 0 hpf; (2) cleavage stage, 10 hpf; (3) hatched blastula stage, 18 hpf; (4) mesenchyme blastula, 24 hpf; (5) the early gastrula, 30 hpf; (6) mid-gastrula stage, 40 hpf; (7) late-gastrula stage, 48 hpf; (8) prism stage, 56 hpf; (9) late prism stage, 64 hpf (10) the pluteus stage, 72 hpf. All embryonic samples were obtained from a single male and female mating pair, except the 24hr sample which was done separately as a pilot experiment. Only a single replicate was sequenced for each time point, presumably due to limitations of the amount of material that can be obtained from a single spawning event and the decision not to introduce biological variation due to individual differences if multiple urchins were used. Each sample generated 36.5M reads, of which 79% mapped to the S.purpuratus genome v3. Gene models were built by Cufflinks, and, after quality filtering, 21,092 transcript models were defined and assigned an 8-digit WHL ID number beginning with “22.” These models have been incorporated into the annotated sea urchin gene database and assigned to previously established “SPU ID” numbers. Initially, values of expression for each gene model were expressed in FPKMs (Fragments Per Kilobase of transcript per Million mapped reads) as determined by Cufflinks. The gene expression values were then converted into transcripts per embryo by the authors (Tu et al. 2012), and made available online in searchable form at http://bouzouki.bio.cs.cmu.edu:3838/quantdev/. We obtained the full data set with gene models identified by WHL ID and SPU ID and expression values for each timepoint in transcripts per embryo. We converted the expression values into RPKMs (Reads Per Kilobase of transcript per Million mapped reads) for our analysis.

The existing GRN models for sea urchin development cover the time period from 0 to 30 hpf. Of the 10 time points sampled in the transcriptome data set from Tu, et al. only 5 are represented in the range of 0-30hpf. Therefore, we decided to also use data from a high density time series containing 49 time points, one time point per hour over the first 48 hours of development and the unfertilized egg (Materna et al. 2010). The data set from Materna et al. uses Nanostring technology, which is often used as a gold standard for absolute quantitation due to the fact that it measures RNA directly as opposed to using an enzymatic reaction (Geiss et al. 2008). Even the Tu, et al. transcriptome data set was validated using independent Nanostring quantitation (Tu et al. 2014). A key difference between RNAseq and Nanostring is that Nanostring will only produce data for specific known gene models for which probes were designed, whereas RNAseq surveys the entire transcriptome. The sea urchin Nanostring data set queried 161 regulatory gene products, including 130 transcription factors and 31 molecules involved in signaling. This

gene set overlaps nicely with the gene set present in the ground truth sea urchin ectoderm and endomesoderm GRN models. Specifically, 68 genes in total overlapped between the Nanostring probe set and the ground truth GRN genes. The overlap included 31 of 39 genes overlapping with the ectoderm GRN and 44 of 58 genes overlapping with the endomesoderm GRN. There were 7 genes in the Nanostring probe set that appear in both the ectoderm and endomesoderm GRNs (namely, *not, otxa, foxA, eve, myc, bra*, and *hnf6*). The normalized RNA counts produced by the Nanostring's Ncounter were used directly in our analysis.

Sea urchin GRN models

The current versions of the *S. purpuratus* GRNs for endomesoderm and ectoderm development are hosted by the Institute for Systems Biology and can be accessed online using the web application BioTapestry Interactive Network Viewer. The endomesoderm GRN can be found at http://grns.biotapestry.org/SpEndomes/, and the ectoderm GRN can be found at http://grns.biotapestry.org/SpEcto/. These two GRN models were built by a collaboration of sea urchin labs over the last thirty-some years. Each regulatory interaction is depicted as a directional line connecting two gene nodes, and each interaction is supported by experimental evidence, which can be accessed in the BioTapestry viewer directly. We obtained lists of the genes present in each network and a list of every gene-to-gene interaction present in the current version of the models from the BioTapestry director William Longabaugh. There are 39 genes represented in the ectoderm GRN and 58 genes represented in the endomesoderm GRN. The interaction list we obtained includes direct interactions and indirect interactions that are driven by signaling molecule intermediates. Interactions derived from signaling intermediates were not used in our comparison list. We also removed interactions where a gene regulates its own expression because the PEAK algorithm is not designed to be able to predict this type of interaction. The final list used in our analysis contained 85 unique gene-to-gene interactions present in the ectoderm GRN (Additional file 5) and 146 unique gene-to-gene interactions present in the endoderm GRN (Additional file 6). These unique interactions made up our ground truth GRN models, which were used for comparison to the interactions predicted by the PEAK algorithm. The number of genes and connections included in our analysis when requiring a match between gene expression data set and corresponding ground truth network are described in Table 3.

Table 3
Summary of data sets used in the evaluation of PEAK's predictions. For each data set, we only used the genes that both appear in the gene expression data and the ground truth GRN data.

Data set	Transcriptome RNAseq data	Transcriptome RNAseq data	Nanostring data set	Nanostring data set
Ground Truth GRN (GT)	Ectoderm GRN	Endomesoderm GRN	Ectoderm GRN	Endomesoderm GRN
Method	RNAseq	RNAseq	Nanostring	Nanostring
Timepoints (T)	10	10	48	48
Genes (N)	37	58	28	41
Edges in GT	82	146	57	76

Preprocessing and differential gene expression determination

The RNAseq data set was constructed with only one biological replica. Multiple methods have been developed to perform differential gene expression analysis on RNAseq data when only a single biological replica is available. These methods include: NOISeq (Tarazona et al. 2011), based on the multinomial distribution; GFold (Feng et al. 2012), based on the posterior distribution of log fold change; and EdgeR (Robinson et al. 2010), based on the negative binomial (NB) distribution. To discover quantitative changes in expression levels between experimental time points, we first applied NOISeq, GFold, and EdgeR to determine the set of differentially expressed genes for further analysis.

For NOISeq, we normalized the data by RPKM (Reads Per Kilobase of transcript per Million mapped reads), which takes into account that more sequencing reads are generated from longer transcripts. The length of each transcript was obtained from the sea urchin database, Echinobase (https://www.echinobase.org). We omitted genes that had no record in the gene database. We set the simulation parameters as recommended in the NOISeq handbook, where the percentage (*pnr*) of the sequencing depth is $pnr = 0.2$, the number of samples to be simulated (*nss*) for each condition is $nss = 5$ and a small variability (v) is $v = 0.02$. We selected the differentially expressed genes with the higher NOISeq probabilities based on our chosen thresholds $\lambda_1 = 0.9$, $\lambda_2 = 0.85$. For GFold, we set the thresholds for the GFold value to $\lambda_3 = \mp 1$, $\lambda_4 = \mp 1.5$, since the GFold value is similar to the log2 fold change that is reliably used to select differentially expressed genes. For EdgeR, we set the log2 fold change (*log2fc*) cutoff as 2 and the edgeR dispersion as 0.01.

We used the gene database at Echinobase to map all WHL IDs to SPU_IDs to ensure that the IDs we analyzed before and after are consistent and unique.

Computational GRN prediction

PEAK was used for our computational GRN predictions. PEAK can be accessed as a front-end web application that can be used to submit and retrieve predictions, available here: http://detangle.cs.vt.edu/. In our analysis, default parameters were used for PEAK, except the “Repeat” was set as 1. We also experimentally tested different half-life values. PEAK returns predicted interactions for each gene that score above the confidence threshold set by the ‘PEAK value’. Each gene has up to 30 ranked predictions. Predictions are marked positive or negative when representing an enhancing interaction or a repressive interaction, respectively. Each interaction is assigned a confidence score, allowing users to sort the interactions with the highest confidence or the top-5 or top-10 predicted interactions for each gene.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated and analyzed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Funding

ST was supported by institutional funds from Canisius College Department of Biology. LH and JZ were supported by institutional funds from Virginia Tech Department of Computer Science.

Authors' contributions

ST conceived of the study. JZ and FI completed the DGE analysis. JZ performed the experiments. ST manually curated gene lists. DA consulted and fine-tuned PEAK's handling of our files. JZ, DA, LH, and ST analysed the results. JZ, DA, LH, and ST wrote the manuscript. JZ made the figures and final file formats.

Acknowledgements

The authors would like to thank Andy Cameron and Greg Wray for help obtaining the transcriptome database files. We would also like to thank Bill Longabaugh and Isabelle Peters for lists of interactions from the sea urchin GRNs.

References

Altarawy D, Eid F-E, Heath LS. PEAK: Integrating Curated and Noisy Prior Knowledge in Gene Regulatory Network Inference. J Comput Biol. 2017 Sep;24(9):863–73.

Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. Development. 1997 May;124(10):1851–64.

Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol. 2015 Nov 17;11(11):839.

Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969 Jul 25;165(3891):349–57.

Chen S-H, Li K-L, Lu I-H, Wang Y-B, Tung C-H, Ting H-C, et al. Sequencing and analysis of the transcriptome of the acorn worm Ptychodera flava, an indirect developing hemichordate. Mar Genomics. 2014 Jun;15:35–43.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, et al. A genomic regulatory network for development. Science. 2002a Mar 1;295(5560):1669–78.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, et al. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Dev Biol. 2002b Jun 1;246(1):162–90.

Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. Artif Intell Med. 2019 Apr;95:133–45.

Du H, Bao Z, Hou R, Wang S, Su H, Yan J, et al. Transcriptome sequencing and characterization for the sea cucumber Apostichopus japonicus (Selenka, 1867). PLoS One. 2012 Mar 12;7(3):e33311.

Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics. 2012 Nov 1;28(21):2782–8.

Fernandez-Valverde SL, Aguilera F, Alexander Ramos-Diaz R. Inference of Developmental Gene Regulatory Networks Beyond Classical Model Systems: New Approaches in the Post-genomic Era. In: Symposium on From Small and Squishy to Big and Armored - Genomic, Ecological and Paleontological Insights into the

Early Evolution of Animals at the Annual Meeting of the Society-for-Integrative-and-Comparative-Biology. San Francisco, CA: OXFORD UNIV PRESS INC; 2018. p. 640–53.

Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008 Mar;26(3):317–25.

Gildor T, Malik A, Sher N, Ben-Tabou de-Leon S. Mature maternal mRNAs are longer than zygotic ones and have complex degradation kinetics in sea urchin. Dev Biol. 2016 Jun 1;414(1):121–31.

Helm RR, Siebert S, Tulin S, Smith J, Dunn CW. Characterization of differential transcript abundance through time during Nematostella vectensis development. BMC Genomics. 2013 Apr 19;14:266.

Henry JJ, Perry KJ, Fukui L, Alvi N. Differential localization of mRNAs during early development in the mollusc, Crepidula fornicata. Integr Comp Biol. 2010 Nov;50(5):720–33.

Henry JQ, Lesoway MP, Perry KJ, Osborne CC, Shankland M, Lyons DC. Beyond the sea: Crepidula atrasolea as a spiralian model system. Int J Dev Biol. 2017;61(8-9):479–93.

Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961 Jun;3:318–56.

Longabaugh WJR, Davidson EH, Bolouri H. Computational representation of developmental genetic regulatory networks. Dev Biol. 2005 Jul 1;283(1):1–16.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012a Jul 15;9(8):796–804.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, et al. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Res. 2012b Jul;22(7):1334–49.

Materna SC, Nam J, Davidson EH. High accuracy, high-resolution prevalence measurement for the majority of locally expressed regulatory genes in early sea urchin development. Gene Expr Patterns. 2010;10(4-5):177–84.

Mirsky AE. Genetics in the Twentieth Century. LC Dunn, The McMillan Co, New York. 1951;128–33.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

Sea Urchin Genome Sequencing Consortium, Sodergren E, Weinstock GM, Davidson EH, Cameron RA, Gibbs RA, et al. The genome of the sea urchin Strongylocentrotus purpuratus. Science. 2006 Nov 10;314(5801):941–52.

Tarazona S, García F, Ferrer A, Dopazo J, Conesa A. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. EMBnet.journal. 2011;17(B):18–9.

Tulin S, Aguiar D, Istrail S, Smith J. A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems [Internet]. Vol. 4, EvoDevo. 2013. p. 16. Available from: http://dx.doi.org/10.1186/2041-9139-4-16

Tu Q, Cameron RA, Davidson EH. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. Dev Biol. 2014 Jan 15;385(2):160–7.

Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson EH. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res. 2012 Oct;22(10):2079–87.

Van den Broeck L, Gordon M, Inzé D, Williams C, Sozzani R. Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. Front Genet. 2020 May 25;11:457.

Figures

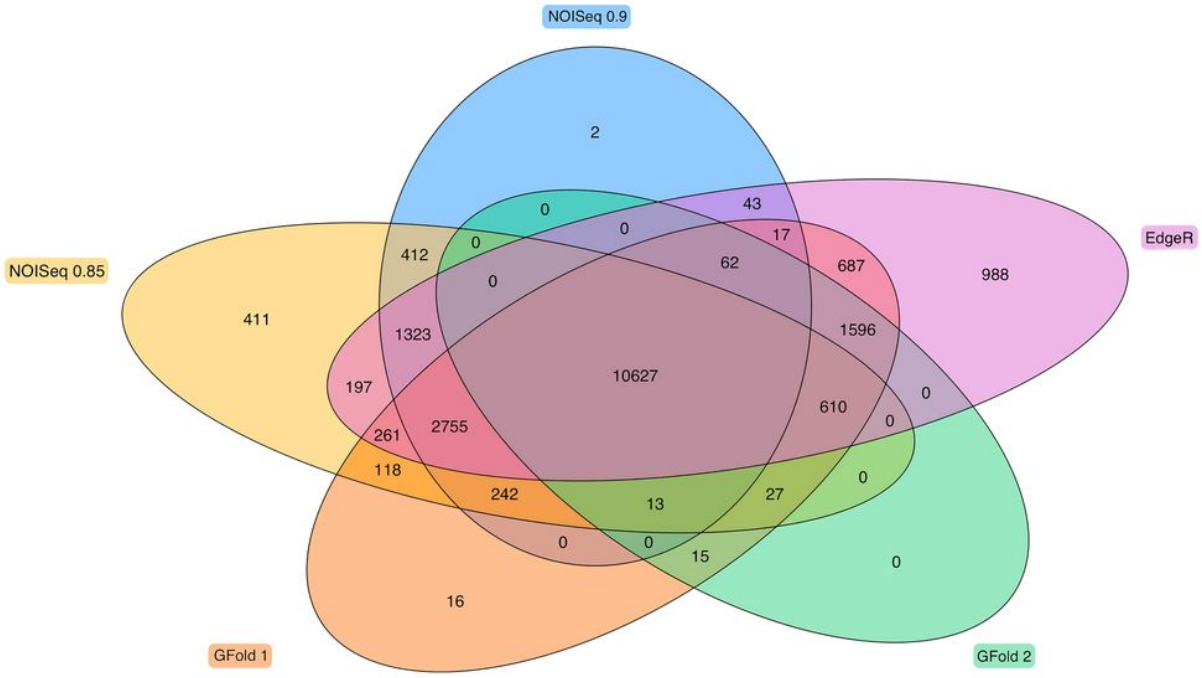

Figure 1

Intersection of unique differentially expressed genes determined by NOISeq, EdgeR and GFold. The intersection of all 5 methods contains 10,627 unique genes specified as differentially expressed. NOISeq (λ1= 0.9) and GFold2 (λ4= ∓ 1.5) identify the fewest unique genes at 2 and 0, respectively. The intersection of the other 4 programs was increased by 2,755 when GFold2 ((λ4= ∓ 1.5)) was removed

from the overlap. NOISeq (λ1= 0.9) had the best balance of containing the most agreed upon gene set while adding the fewest unique genes.

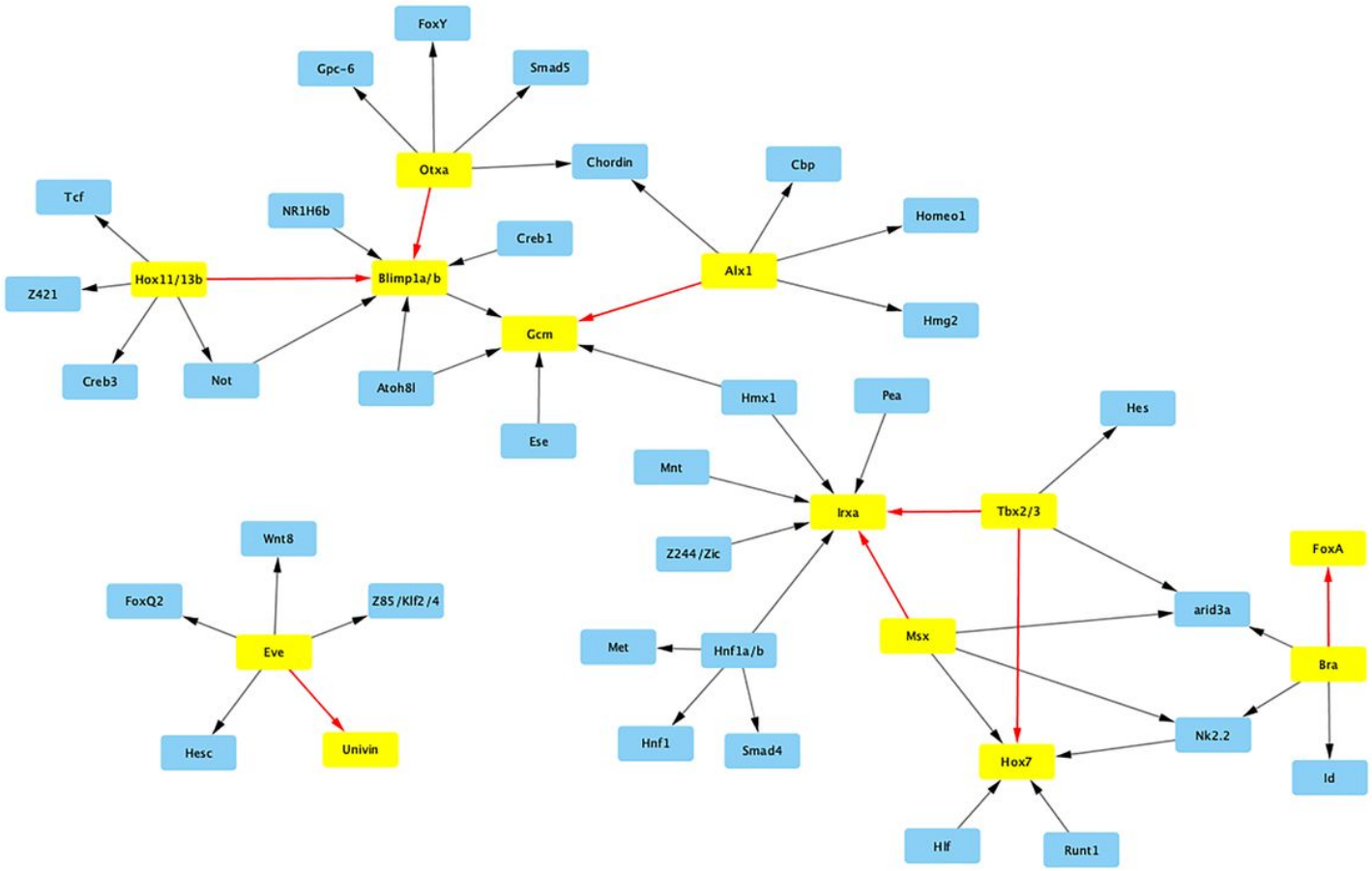

Figure 2

A subset of PEAK predicted interactions for genes in the ground truth network with the top-5 predictions with the highest confidence scores for each gene. Both known interactions (red arrows) and new predicted interactions (black arrows) are included among these predictions.

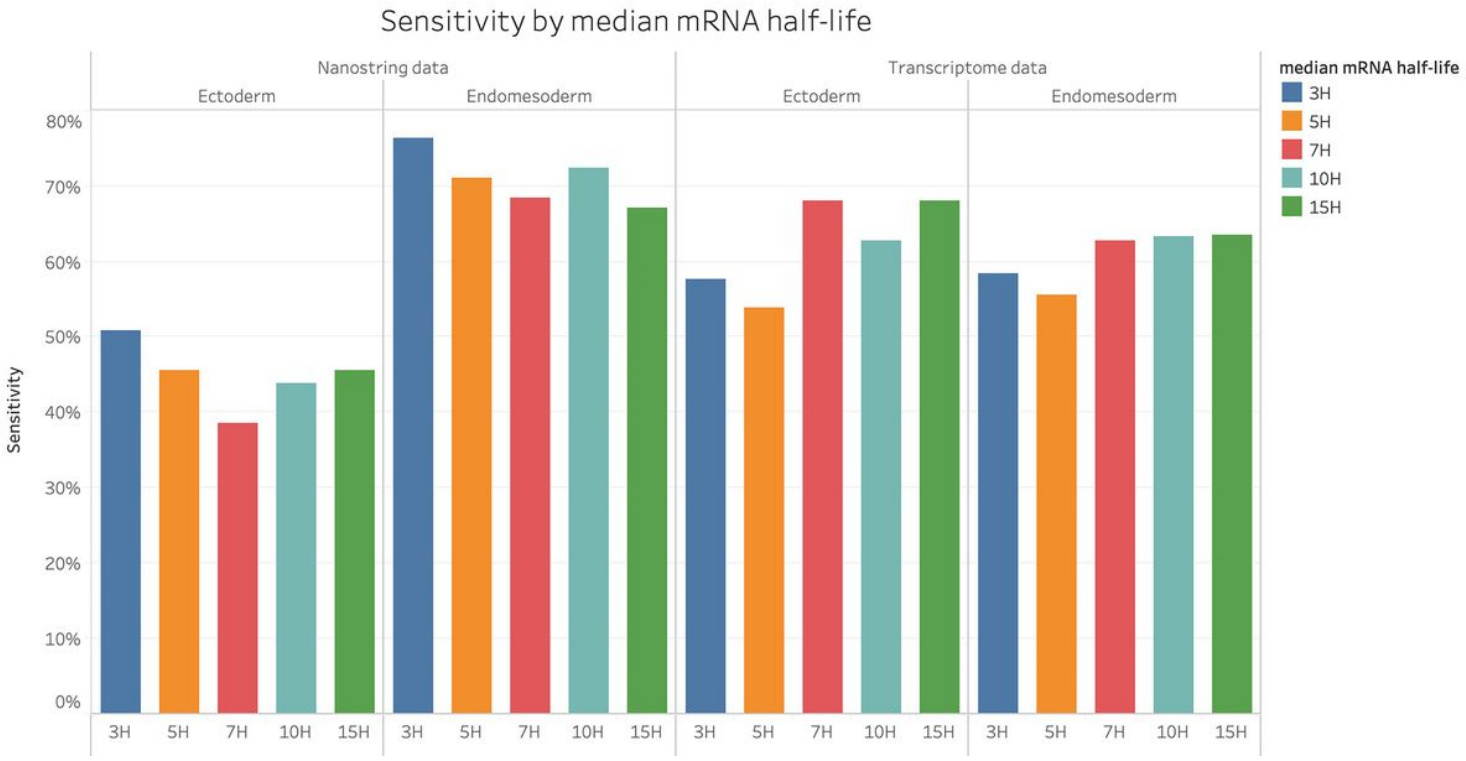

Figure 3

Sensitivity as a measure of accuracy for the prediction of ectoderm and endomesoderm gene regulatory relations calculated with 5 different median mRNA half-life settings (3hrs, 5hrs, 7hrs, 10hrs, 15hrs) on the transcriptome data and the nanostring data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile2.csv
- AdditionalFile4.csv
- AdditionalFile5.csv
- AdditionalFile6.csv
- Additionalfile1.xlsx
- ExperimentsMetaData.tsv
- GeneExpression.tsv
- TranscriptionFactors.tsv