Comparison of primary glioblastoma Single-cell and Tissue RNA-seq co-expression networks - Submission to PLOS Journals

Brian Arand[2], Raghu Machiraju[1, 2], Kun Huang[1, 2]

1 Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, The United States of America
2 Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, The United States of America

Abstract

This work offers an initial glimpse at the potential of single-cell co-expression network module enrichment analysis. We run two primary glioblastoma transcriptome datasets, one single-celled and one bulk tissue sample, through analogous co-expression module analysis pipelines. We find that stringent filtering at multiple stages of the single-celled pipleine is vital to control for the noisy nature of single-celled data. Comparison of output modules at the module level and at the gene-level reveals that the two pipelines produce comparable results for many expected modules, as well as a number of notable differences that may illustrate key differences in the underlying technologies. More interestingly, we find that single-celled modules lack a strong signal for extracellular biological process ontologies and tumor microenvironment like extracelluar matrix organization, collagen fiber organization, immune response, cellular adhesion. Small signals for anti-apoptosis were only detected in the single-cell data, and a stronger cell cycle/nuclear division signal was detected in the single-cell results than in the bulk cell results. These findings reinforce some assumptions about the domains specific to both RNA sequencing and single-cell RNA sequencing technologies.

Introduction

The isolation of individual cells' transcriptome profiles has been largely a theoretical concept to bioinformaticians. And accordingly, transcriptomic inquiry has been limited to those questions regarding tissues, potentially composed of a heterogeneous hodgepodge of cellular types, subtypes, and states. But with the advent of Single-Cell RNA sequencing (RNASeq) technology, comes the potential for refined resolution in transcriptomic datasets. And expectedly, recent publications suggest a peaking interest in this new landscape of informatics. It has been shown that many bioinformatics techniques that were developed for bulk-cell tissue samples can be effectively applied to single-cellular datasets. However, co-expression network analysis has largely been an unexplored area of analysis in regards to single-cell RNASeq data. To fill this gap, we leverage this new technology to construct and analyze gene co-expression networks for primary glioblastoma single-cell samples. Glioblastoma is widely known to be a heterogeneous cancer, making it a prime candidate for single-cellular inquiries. For instance, we hypothesized that the averaging of single-cells' profiles within a tissue sample may mask or otherwise confound downstream gene correlations based analysis. Correlation between two genes may exist across tissue samples purely due to changing proportion of cellular subtypes within those samples. However, a single-cellular perspective, of the same tumors may theoretically filter out those artificial tissue-level correlations. And so correlation based analyses, like co-expression network analysis, require study. In our work, we begin this journey by looking at network mining, module detection, and gene enrichment analysis at both the single-cell and bulk cell (tissue sample) levels. The final goal of this work is to shed light on the convoluted intricacies of inter-cellular genomic landscape of glioblastoma tissue from a single-cellular perspective.

Materials and Methods

We analyzed two glioma datasets: GSE57872, the data presented in Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma [1], and GSE48865, the data presented in RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas [2]. GSE57872 consists of single-cell samples whereas GSE48865 consists of more-traditional, bulk samples. From here on, GSE57872 and GSE48865 samples will be referred to as 'single-cell' and 'bulk' samples respectively. After preprocessing each dataset independently (details explained in following sections), we step each dataset through a coexpression-network analysis workflow. An overview of this workflow is shown in Fig. 1. Details of and rational for data set dependent variations of this workflow are explained in the sections to follow.

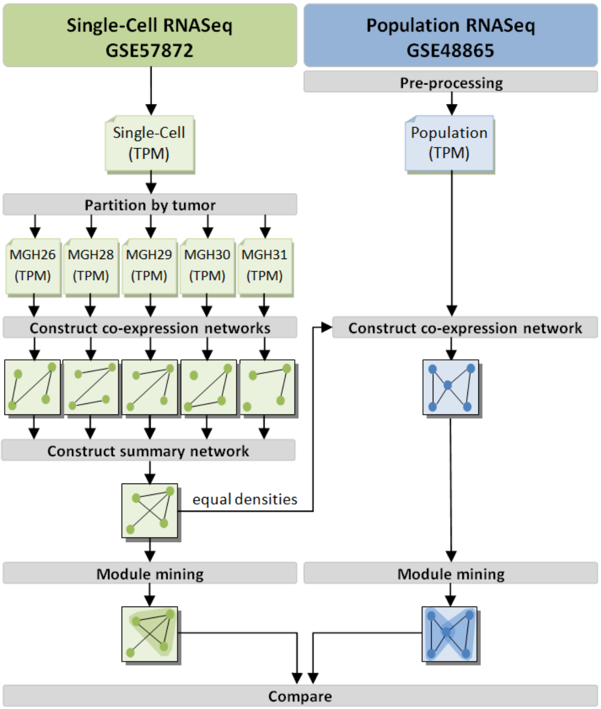


Figure 1. Network comparison workflow. Note that there is no partition nor aggregation step in the bulk RNASeq pipeline. Moreover, note that edges of the constructed bulk co-expression network are filtered to achieve the closest network density possible to that realized in the Single-Cell pipeline.

Single-Cell Data

Single-Cell RNASeq Samples

In this section, we discuss the handling of the GSE57872, single-cell data. This data set is comprised of normalized gene expression values for 5,948 genes for 430 single-cell samples selected by flow cytometry cell sorting and micromanipulation from 5 different primary glioblastomas labeled: MGH26, MGH28, MGH29, MGH30, and MGH31. Patel et al sequenced the cells using SMARTseq protocol [3]. Alignment to hg19 was performed with Bowtie (version 1.1.1) [4] and the authors calculated TPM (transcripts per million) values using RSEM (version 1.2.3) [5].The final reported values were log-transformed and mean-shifted per gene. More formally, let x_{si} be the TPM enrichment value for the i^{th} gene of sample s. And let N_t be the sample size of tumor t, then the analogous, final reported value, y_{si}, would be:

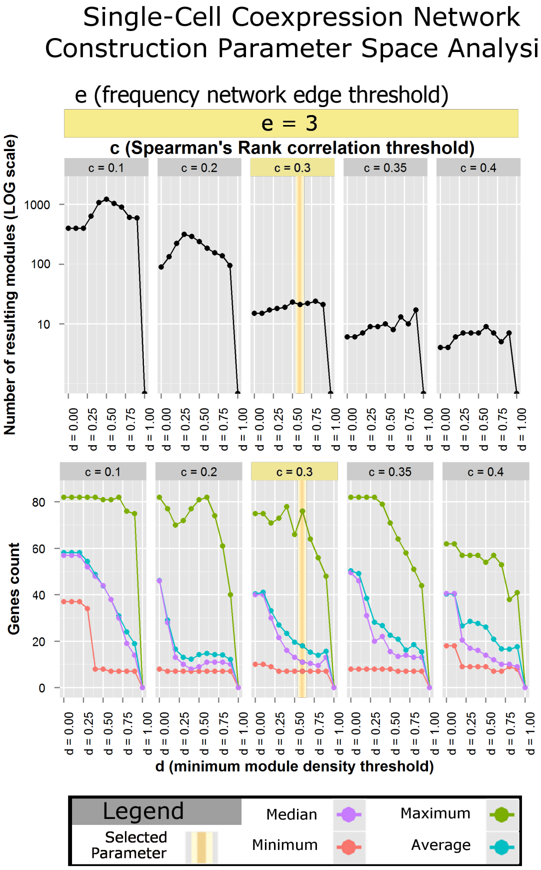


Figure 2. Parameterization space analysis. A comparison of module statistics (a) number of modules output (b) distribution statistics of module size in gene count for modules output in one run of CODENSE. Values of the selected parameterization are highlighted in yellow.

$$y_i = \frac{\sum_s \log_2(x_{si} + 1)}{N_t} \tag{1}$$

$$y_{si} = \sum_s \log_2(x_{si} + 1) - y_i \tag{2}$$

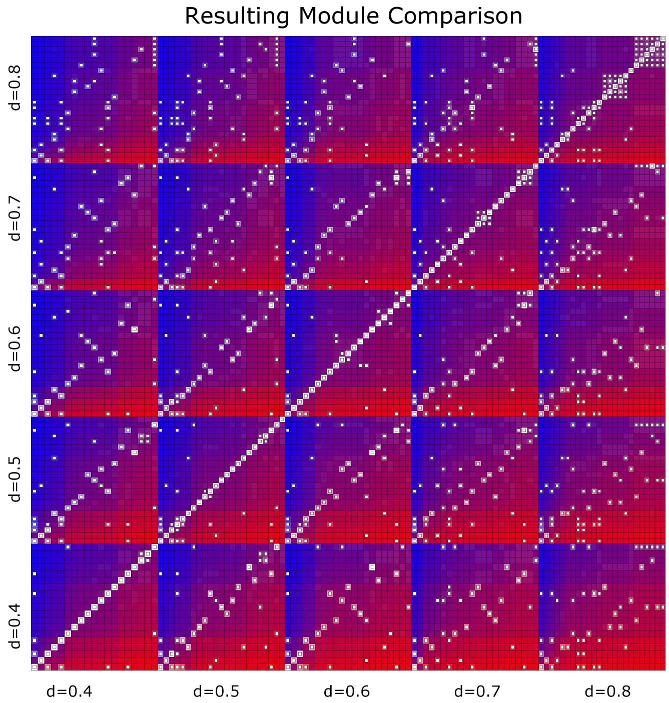


Figure 3. Parameterization space starry night map. Here we display the exhaustive gene-set by gene-set enrichment profiles for those modules output by 5 different runs of CODENSE for different values of the minimum density threshold d. The color of each square encodes the size of some gene set A, some gene set B, and the size of $A \cap B$ relative to $A \cup B$. These three values are encoded into the red, blue, and green channels of the square respectively. Thus, the whiter a square is the more similar A and B are. Furthermore, a Fisher's exact test was performed for significant enrichment per gene set pair. Placement of a small, white token indicates significant overlap between A and B. Note that the central diagonal is all white. Which makes sense because this is the comparison of all sets with themselves. This diagonal pattern is most stable when comparing the selected characterization d=0.6 with each of it's neighboring parameterizations than between other neighboring steps shown above

0.0.1 Network Construction

In our analysis, single-cell samples were grouped by tumor of origin, t. For each group, a gene-by-gene Spearman's rank correlation matrix was calculated:

$$P_t = \begin{bmatrix} \rho_{11t} & \cdots & \rho_{1kt} \\ \vdots & \ddots & \vdots \\ \rho_{k1t} & \cdots & \rho_{kkt}i \end{bmatrix} \qquad (3)$$

Where ρ_{ijt} is the Spearman's correlation between the i^{th} and j^{th} gene of tumor t. Co-expression matrices were then created by filtering at the same Spearman's rank threshold, c, to produce binary matrices, B_t, per tumor.

$$B_t = \begin{bmatrix} b_{11t} & \dots & b_{1kt} \\ \vdots & \ddots & \vdots \\ b_{k1t} & \dots & b_{kkt}i \end{bmatrix}, b_{ijt} = \{ \begin{array}{cc} 1, & |\rho_{ijt}| \geq c \\ 0, & o.w. \end{array} \tag{4}$$

This is analogous to cutting a co-expression network at a given edge threshold. After analyzing the distribution of resulting networks' densities 2 we settled on a threshold of 0.3 with an associated frequency network density of 0.04. We point out that this threshold is low, but we argue that it is a necessary trade-off given the distribution of correlations. It is expected to see such non-concordant data, because current single-cell RNASeq technology is known to produce noisy data (need a reference for this regarding pcr amplification of single cell).

The binary matrices were then aggregated into a single frequency matrix, F. This aggregation is performed in an attempt to attenuate patient-specific signals and focus, rather, on glioblastoma-specific patterns of gene co-expression.

$$F = \begin{bmatrix} \sqsubset \sum_t b_{11t} & \dots & \sum_t b_{1kt} \\ \vdots & \ddots & \vdots \\ \sum_t b_{k1t} & \dots & \sum_t b_{kkt} \end{bmatrix} \tag{5}$$

0.0.2 Module Detection

It is this matrix F that is fed into the CODENSE software to be converted into a summary network, S, and then mined for high-density modules [6].

$$S = (V, E), \text{ where } v_i, v_j \in V \text{ and } (v_i, v_j) \in E \leftrightarrow f_{ij} \geq e \tag{6}$$

Frequency threshold threshold e was set to be 3 meaning that in order for a given edge representing a gene-gene correlation to appear in the final summary network S, that same pair of genes had to appear in a majority of the individual B_t networks.

CODENSE then reported the modules that meet some minimum density threshold, d. At the suggestion of the CODENSE authors and after analyzing the effect perturbing d has on the number and size of the resulting modules, we chose a minimum density threshold of 0.6. We wanted to maximize d to ensure dense modules to ensure their biological significance. Note that in Fig. 2, heightened values of d can drastically affect the output number and size of modules. Our selection of $d = 0.6$ marks the upper limit of where this non-robust behavior starts in our selected parameter space sampling resolution. At the selected parameterization, CODENSE reported 21 dense, first-order modules. These relatively high thresholds guarantee that our resulting modules will be dense and therefore will capture only the strongest signals found in the data.

CODENSE Parameterization Validation

We make a considerable effort to ensure that the resulting modules output by CODENCE at the selected parameterization are robust to change withing the algorithm's parameterization space. Fig. 2 shows that the size and number of output modules are relatively stable at the selected values for e, c, and d (highlighted in yellow). Furthermore, 3 is a custom visualization we developed to compare set-by-set co-enrichment. The central diagonal is visualized all in white, indicating that the sets compared with themselves ais all did not change the gene membership of the sets much.

0.1 Bulk Cell Samples

Our bulk cell sample control comes from GEO accession GSE48865. 274 glioma tissue samples make up this dataset, of which, 59 were identified as primary, stage-4 glioblastoma tissue samples. We filtered out any reads from the 59 samples with more than 10% of its nucleotides designated as unknown bases, and any read with more than 50% of its bases having Sanger phred+33 quality scores less than 5. After filtering, we retained 1,388,064,773 reads in total.

Each filtered sample was then aligned to hg19 transcriptome (GRCh37) using bow-tie (version 1.1.1 with default parameters provided by RSEM) and TPM values per gene were estimated using RSEM (version 1.2.3). 73% of the reads were aligned successfully either uniquely or multi-mapped. TPM values were adjusted again as in equation 2. Genes were then filtered to retain only those that were highly expressed across the 59 samples. The i^{th} gene was filtered out if $y_i <= 4.5$. We retained 6,932 genes after filtering, 3,927 of which had also been retained in the analogous steps of the single-cell workflow as illustrated in Fig. **??**.

0.1.1 Network Construction and Module Detection

Again, as in equation 3, Spearman's Rank correlation was calculated for all genes pair-wise. The Spearman's Rank threshold, c, from equation 4, was set to 0.02 so as to achieve the same density as the single-cell network. No further network aggregation step was required for this data set.

Modules were also found using the CODENSE algorithm using the same parameterization used in the single-cell workflow described above.

0.2 Enrichment

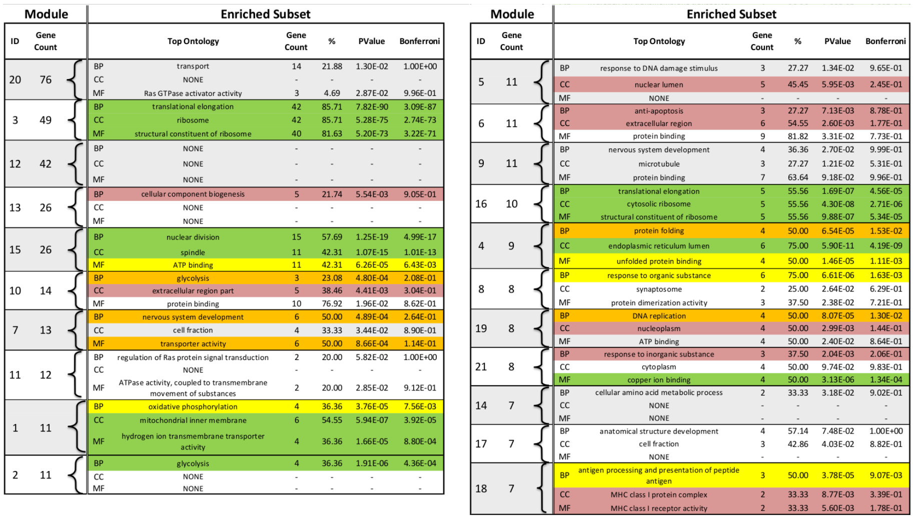
Module sets derived from both the single-cell network and bulk tissue sample network were then ontology enriched. Modules were enriched for all terms in Gene Ontology Consortium's biological process (BP), cellular component (CC), and molecular function (MF) human ontologies using the David Bioinformatics Resources (version 6.7) Functional Annotation Tool [7]. The most significant (in terms of bonferroni adjusted p-value) ontology per category was reported for each module.

Results

Using the parameters described above, our workflow produced 21 modules for the single-cell data and 48 modules for the bulk data set. A breakdown of these modules labeled with their most significant biological process ontology, cellular component ontology, and molecular function ontology is shown for the single-cell data in fig. 4 and for the bulk data in fig. 5.

We compared these output module sets at the module-level and the gene-level. At the module-level, we find that the single-celled modules are enriched for many expected biological process ontologies such as translation elongation, nuclear division, glycolysis, oxidative phosphorylation, protien folding, and DNA replication. Likewise, we find expected enriched ontologies reported by the bulk pipeline–namely: translational elongation, cell cycle, extracellular matrix organization, antigen processing, immune response, RNA splicing, and others.

Amid these expected ontologies, we found a number of enrichments that were exclusive to either one workflow or the other. Bulk modules enriched for several ontologies that single-cell modules did not–extracelluar matrix organization, collagen fiber organization, immune response, and cellular adhesion. These ontologies represent



Module		Enriched Subset					
ID	Gene Count		Top Ontology	Gene Count	%	PValue	Bonferroni
20	76	BP	transport	14	21.88	1.30E-02	1.00E+00
		CC	NONE	-	-	-	-
		MF	Ras GTPase activator activity	3	4.69	2.87E-02	9.96E-01
3	49	BP	translational elongation	42	85.71	7.82E-90	3.09E-87
		CC	ribosome	42	85.71	5.28E-75	2.74E-73
		MF	structural constituent of ribosome	40	81.63	5.20E-73	3.22E-71
12	42	BP	NONE	-	-	-	-
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
13	26	BP	cellular component biogenesis	5	21.74	5.54E-03	9.05E-01
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
15	26	BP	nuclear division	15	57.69	1.25E-19	4.99E-17
		CC	spindle	11	42.31	1.07E-15	1.01E-13
		MF	ATP binding	11	42.31	6.26E-05	6.43E-03
10	14	BP	glycolysis	3	23.08	4.80E-04	2.08E-01
		CC	extracellular region part	5	38.46	4.41E-03	3.04E-01
		MF	protein binding	10	76.92	1.96E-02	8.62E-01
7	13	BP	nervous system development	6	50.00	4.89E-04	2.64E-01
		CC	cell fraction	4	33.33	3.44E-02	8.90E-01
		MF	transporter activity	6	50.00	8.66E-04	1.14E-01
11	12	BP	regulation of Ras protein signal transduction	2	20.00	5.82E-02	1.00E+00
		CC	NONE	-	-	-	-
		MF	ATPase activity, coupled to transmembrane movement of substances	2	20.00	2.85E-02	9.12E-01
1	11	BP	oxidative phosphorylation	4	36.36	3.76E-05	7.56E-03
		CC	mitochondrial inner membrane	6	54.55	5.94E-07	3.92E-05
		MF	hydrogen ion transmembrane transporter activity	4	36.36	1.66E-05	8.80E-04
2	11	BP	glycolysis	4	36.36	1.91E-06	4.36E-04
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-

Module		Enriched Subset					
ID	Gene Count		Top Ontology	Gene Count	%	PValue	Bonferroni
5	11	BP	response to DNA damage stimulus	3	27.27	1.34E-02	9.65E-01
		CC	nuclear lumen	5	45.45	5.95E-03	2.45E-01
		MF	NONE	-	-	-	-
6	11	BP	anti-apoptosis	3	27.27	7.13E-03	8.78E-01
		CC	extracellular region	6	54.55	2.60E-03	1.77E-01
		MF	protein binding	9	81.82	3.31E-02	7.73E-01
9	11	BP	nervous system development	4	36.36	2.70E-02	9.99E-01
		CC	microtubule	3	27.27	1.21E-02	5.31E-01
		MF	protein binding	7	63.64	9.18E-02	9.96E-01
16	10	BP	translational elongation	5	55.56	1.69E-07	4.56E-05
		CC	cytosolic ribosome	5	55.56	4.30E-08	2.71E-06
		MF	structural constituent of ribosome	5	55.56	9.88E-07	5.34E-05
4	9	BP	protein folding	4	50.00	6.54E-05	1.53E-02
		CC	endoplasmic reticulum lumen	6	75.00	5.90E-11	4.19E-09
		MF	unfolded protein binding	4	50.00	1.46E-05	1.11E-03
8	8	BP	response to organic substance	6	75.00	6.61E-06	1.63E-03
		CC	synaptosome	2	25.00	2.64E-02	6.29E-01
		MF	protein dimerization activity	3	37.50	2.38E-02	7.21E-01
19	8	BP	DNA replication	4	50.00	8.07E-05	1.30E-02
		CC	nucleoplasm	4	50.00	2.99E-03	1.44E-01
		MF	ATP binding	4	50.00	2.40E-02	8.64E-01
21	8	BP	response to inorganic substance	3	37.50	2.04E-03	2.06E-01
		CC	cytoplasm	4	50.00	9.74E-02	9.83E-01
		MF	copper ion binding	4	50.00	3.13E-06	1.34E-04
14	7	BP	cellular amino acid metabolic process	2	33.33	3.18E-02	9.02E-01
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
17	7	BP	anatomical structure development	4	57.14	7.48E-02	1.00E+00
		CC	cell fraction	3	42.86	4.03E-02	8.82E-01
		MF	NONE	-	-	-	-
18	7	BP	antigen processing and presentation of peptide antigen	3	50.00	3.78E-05	9.07E-03
		CC	MHC class I protein complex	2	33.33	8.77E-03	3.39E-01
		MF	MHC class I receptor activity	2	33.33	5.60E-03	1.78E-01

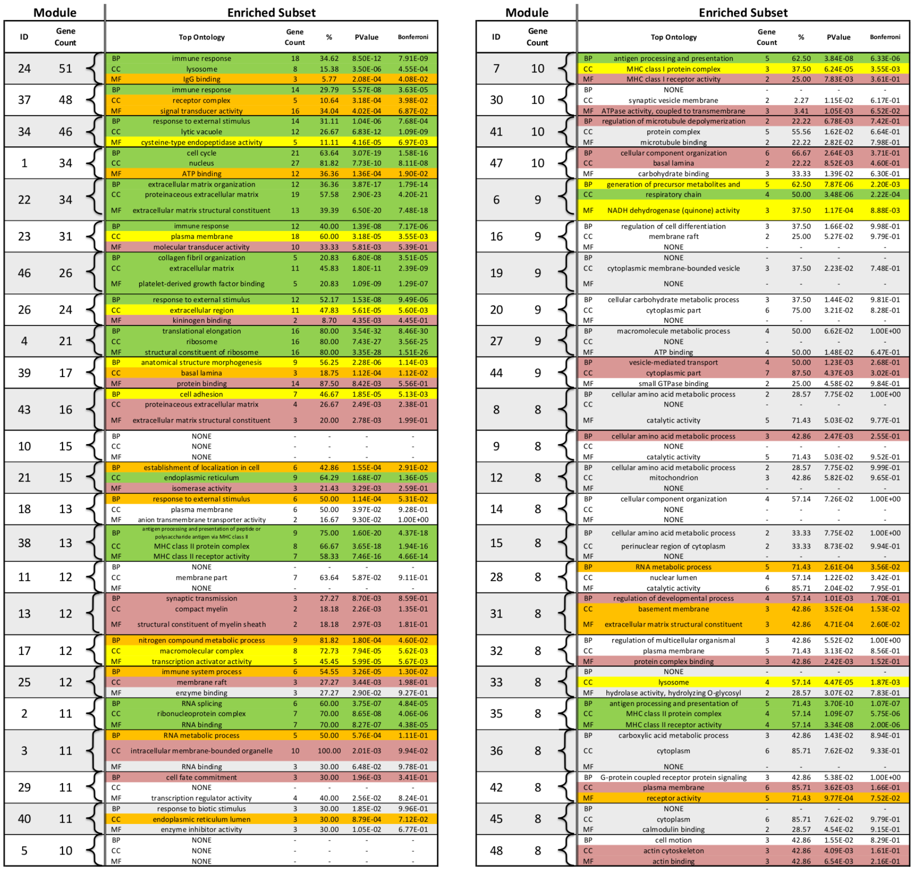
Figure 4. Single-cell network module ontology enrichment. A list of the modules output from the CODENSE algorithm ordered by module size, each enriched for BP, CC, and MC ontologies. Row color denotes a certain statistical significance as demarcated in the figure legend.

extracellular and/or tumor micro environments. Analogously there were also a couple of weaker ontology signals that only single-celled data reported that are worth mentioning– nervous system development, and anti-apoptosis. These provide potential candidate genes that may be averaged out in the bulk cell data that only the resolution of single-celled data can obtain.

We then profiled the two lists of output ontology-enriched modules to get another view of both the consensus and differing ontologies between the two data sets. When using exact equality to define the intersection between the two sets of ontologies, we find that only two ontologies were identified by both single-cell and bulk pipelines: translation elongation and cellular amino acid metabolic process. To get a better understanding of the similarities of the pipelines' output, we relax our definition of equality and report what we will be calling 'contextual ontologies' in the intersection. We aim to summaries ontologies reported by both pipelines by seeking close common ancestors within the GO hierarchical database of ontologies. More concretely, a contextual ontology is the root of the shortest possible subtree with height no greater than 2 that connects at least one ontology that was reported by the single-cell pipeline with at least one ontology that was reported by the bulk pipeline. These contextual ontologies, along with both the differing and consensus ontologies are reported in fig. 6.

A gene-level comparison of the resulting modules of the two data sets is also provided in fig. 7. We find 3 or 4 groups of modules within the bulk output that are co-enriched for each other. Meaning that a given subgroup of modules share a significant number of genes with each other. We hypothesis that this co-enrichment occurred as a result of the specific parameterization used to construct and mine the network for modules. By either raising the correlation threshold, c, or raising the minimum module density threshold, d, these co-enriched modules may fuse into a larger module. We further hypothesis, by taking cues from the ontology enrichment of the co-enriched modules, that the resulting larger modules would likely enrich for immune response, metabolic processes, and extracellular matrix ontologies.

At the gene level, we find significant co-enrichment between the module outputs of the two pipelines; 5 pairs of which are of interest. The breakdown of these 5 co-enrichments can be found in fig. 8. Note that there was a drop in p-value between



Module		Enriched Subset					
ID	Gene Count		Top Ontology	Gene Count	%	PValue	Bonferroni
24	51	BP	immune response	18	34.62	8.50E-12	7.91E-09
		CC	lysosome	8	15.38	3.50E-06	4.55E-04
		MF	IgG binding	3	5.77	2.08E-04	4.08E-02
37	48	BP	immune response	14	29.79	5.57E-08	3.63E-05
		CC	receptor complex	5	10.64	3.18E-04	3.98E-02
		MF	signal transducer activity	16	34.04	4.02E-04	6.87E-02
34	46	BP	response to external stimulus	14	31.11	1.04E-06	7.68E-04
		CC	lytic vacuole	12	26.67	6.83E-12	1.09E-09
		MF	cysteine-type endopeptidase activity	5	11.11	4.16E-05	6.97E-03
1	34	BP	cell cycle	21	63.64	3.07E-19	1.58E-16
		CC	nucleus	27	81.82	7.73E-10	8.11E-08
		MF	ATP binding	12	36.36	1.36E-04	1.90E-02
22	34	BP	extracellular matrix organization	12	36.36	3.87E-17	1.79E-14
		CC	proteinaceous extracellular matrix	19	57.58	2.90E-23	4.20E-21
		MF	extracellular matrix structural constituent	13	39.39	6.50E-20	7.48E-18
23	31	BP	immune response	12	40.00	1.39E-08	7.17E-06
		CC	plasma membrane	18	60.00	3.18E-05	3.55E-03
		MF	molecular transducer activity	10	33.33	5.81E-03	5.39E-01
46	26	BP	collagen fibril organization	5	20.83	6.80E-08	3.51E-05
		CC	extracellular matrix	11	45.83	1.80E-11	2.39E-09
		MF	platelet-derived growth factor binding	5	20.83	1.09E-09	1.29E-07
26	24	BP	response to external stimulus	12	52.17	1.53E-08	9.49E-06
		CC	extracellular region	11	47.83	5.61E-05	5.60E-03
		MF	kininogen binding	2	8.70	4.35E-03	4.45E-01
4	21	BP	translational elongation	16	80.00	3.54E-32	8.46E-30
		CC	ribosome	16	80.00	7.43E-27	3.56E-25
		MF	structural constituent of ribosome	16	80.00	3.35E-28	1.51E-26
39	17	BP	anatomical structure morphogenesis	9	56.25	2.28E-06	1.14E-03
		CC	basal lamina	3	18.75	1.12E-04	1.12E-02
		MF	protein binding	14	87.50	8.42E-03	5.56E-01
43	16	BP	cell adhesion	7	46.67	1.85E-05	5.13E-03
		CC	proteinaceous extracellular matrix	4	26.67	2.49E-03	2.38E-01
		MF	extracellular matrix structural constituent	3	20.00	2.78E-03	1.99E-01
10	15	BP	NONE	-	-	-	-
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
21	15	BP	establishment of localization in cell	6	42.86	1.55E-04	2.91E-02
		CC	endoplasmic reticulum	9	64.29	1.68E-07	1.36E-05
		MF	isomerase activity	3	21.43	3.29E-03	2.59E-01
18	13	BP	response to external stimulus	6	50.00	1.14E-04	5.31E-02
		CC	plasma membrane	6	50.00	3.97E-02	9.28E-01
		MF	anion transmembrane transporter activity	2	16.67	9.30E-02	1.00E+00
38	13	BP	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	9	75.00	1.60E-20	4.37E-18
		CC	MHC class II protein complex	8	66.67	3.65E-18	1.94E-16
		MF	MHC class II receptor activity	7	58.33	7.46E-16	4.66E-14
11	12	BP	NONE	-	-	-	-
		CC	membrane part	7	63.64	5.87E-02	9.11E-01
		MF	NONE	-	-	-	-
13	12	BP	synaptic transmission	3	27.27	8.70E-03	8.59E-01
		CC	compact myelin	2	18.18	2.26E-03	1.35E-01
		MF	structural constituent of myelin sheath	2	18.18	2.97E-03	1.81E-01
17	12	BP	nitrogen compound metabolic process	9	81.82	1.80E-04	4.60E-02
		CC	macromolecular complex	8	72.73	7.94E-05	5.62E-03
		MF	transcription activator activity	5	45.45	5.99E-05	5.67E-03
25	12	BP	immune system process	6	54.55	3.26E-05	1.30E-02
		CC	membrane raft	3	27.27	3.44E-03	1.98E-01
		MF	enzyme binding	3	27.27	2.90E-02	9.27E-01
2	11	BP	RNA splicing	6	60.00	3.75E-07	4.84E-05
		CC	ribonucleoprotein complex	7	70.00	8.65E-08	4.06E-06
		MF	RNA binding	7	70.00	8.27E-07	4.38E-05
3	11	BP	RNA metabolic process	5	50.00	5.76E-04	1.11E-01
		CC	intracellular membrane-bounded organelle	10	100.00	2.01E-03	9.94E-02
		MF	RNA binding	3	30.00	6.48E-02	9.78E-01
29	11	BP	cell fate commitment	3	30.00	1.96E-03	3.41E-01
		CC	NONE	-	-	-	-
		MF	transcription regulator activity	4	40.00	2.56E-02	8.24E-01
40	11	BP	response to biotic stimulus	3	30.00	1.85E-02	9.96E-01
		CC	endoplasmic reticulum lumen	3	30.00	8.79E-04	7.12E-02
		MF	enzyme inhibitor activity	3	30.00	1.05E-02	6.77E-01
5	10	BP	NONE	-	-	-	-
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
7	10	BP	antigen processing and presentation	5	62.50	3.84E-08	6.33E-06
		CC	MHC class I protein complex	3	37.50	6.24E-05	3.55E-03
		MF	MHC class I receptor activity	2	25.00	7.83E-03	3.61E-01
30	10	BP	NONE	-	-	-	-
		CC	synaptic vesicle membrane	2	2.27	1.15E-02	6.17E-01
		MF	ATPase activity, coupled to transmembrane	3	3.41	1.05E-03	6.52E-02
41	10	BP	regulation of microtubule depolymerization	2	22.22	6.78E-03	7.42E-01
		CC	protein complex	5	55.56	1.62E-02	6.64E-01
		MF	microtubule binding	2	22.22	2.82E-02	7.98E-01
47	10	BP	cellular component organization	6	66.67	2.64E-03	3.71E-01
		CC	basal lamina	2	22.22	8.52E-03	4.60E-01
		MF	carbohydrate binding	3	33.33	1.39E-02	6.30E-01
6	9	BP	generation of precursor metabolites and	5	62.50	7.87E-06	2.20E-03
		CC	respiratory chain	4	50.00	3.48E-06	2.22E-04
		MF	NADH dehydrogenase (quinone) activity	3	37.50	1.17E-04	8.88E-03
16	9	BP	regulation of cell differentiation	3	37.50	1.66E-02	9.98E-01
		CC	membrane raft	2	25.00	5.27E-02	9.79E-01
		MF	NONE	-	-	-	-
19	9	BP	NONE	-	-	-	-
		CC	cytoplasmic membrane-bounded vesicle	3	37.50	2.23E-02	7.48E-01
		MF	NONE	-	-	-	-
20	9	BP	cellular carbohydrate metabolic process	3	37.50	1.44E-02	9.81E-01
		CC	cytoplasmic part	6	75.00	3.21E-02	8.28E-01
		MF	NONE	-	-	-	-
27	9	BP	macromolecule metabolic process	4	50.00	6.62E-02	1.00E+00
		CC	NONE	-	-	-	-
		MF	ATP binding	4	50.00	1.48E-02	6.47E-01
44	9	BP	vesicle-mediated transport	4	50.00	1.23E-03	2.68E-01
		CC	cytoplasmic part	7	87.50	4.37E-03	3.02E-01
		MF	small GTPase binding	2	25.00	4.58E-02	9.84E-01
8	8	BP	cellular amino acid metabolic process	2	28.57	7.75E-02	1.00E+00
		CC	NONE	-	-	-	-
		MF	catalytic activity	5	71.43	5.03E-02	9.77E-01
9	8	BP	cellular amino acid metabolic process	3	42.86	2.47E-03	2.55E-01
		CC	NONE	-	-	-	-
		MF	catalytic activity	5	71.43	5.03E-02	9.52E-01
12	8	BP	cellular amino acid metabolic process	2	28.57	7.75E-02	9.99E-01
		CC	mitochondrion	3	42.86	5.82E-02	9.65E-01
		MF	NONE	-	-	-	-
14	8	BP	cellular component organization	4	57.14	7.26E-02	1.00E+00
		CC	NONE	-	-	-	-
		MF	NONE	-	-	-	-
15	8	BP	cellular amino acid metabolic process	2	33.33	7.75E-02	1.00E+00
		CC	perinuclear region of cytoplasm	2	33.33	8.73E-02	9.94E-01
		MF	NONE	-	-	-	-
28	8	BP	RNA metabolic process	5	71.43	2.61E-04	3.56E-02
		CC	nuclear lumen	4	57.14	1.22E-02	3.42E-01
		MF	catalytic activity	6	85.71	2.04E-02	7.95E-01
31	8	BP	regulation of developmental process	4	57.14	1.01E-03	1.70E-01
		CC	basement membrane	3	42.86	3.52E-04	1.53E-02
		MF	extracellular matrix structural constituent	3	42.86	4.71E-04	2.60E-02
32	8	BP	regulation of multicellular organismal	3	42.86	5.52E-02	1.00E+00
		CC	plasma membrane	5	71.43	3.13E-02	8.56E-01
		MF	protein complex binding	3	42.86	2.42E-03	1.52E-01
33	8	BP	NONE	-	-	-	-
		CC	lysosome	4	57.14	4.47E-05	1.87E-03
		MF	hydrolase activity, hydrolyzing O-glycosyl	2	28.57	3.07E-02	7.83E-01
35	8	BP	antigen processing and presentation of	5	71.43	3.70E-10	1.07E-07
		CC	MHC class II protein complex	4	57.14	1.09E-07	5.75E-06
		MF	MHC class II receptor activity	4	57.14	3.34E-08	2.00E-06
36	8	BP	carboxylic acid metabolic process	3	42.86	1.43E-02	8.94E-01
		CC	cytoplasm	6	85.71	7.62E-02	9.33E-01
		MF	NONE	-	-	-	-
42	8	BP	G-protein coupled receptor protein signaling	3	42.86	5.38E-02	1.00E+00
		CC	plasma membrane	6	85.71	3.62E-03	1.66E-01
		MF	receptor activity	5	71.43	9.77E-04	7.52E-02
45	8	BP	NONE	-	-	-	-
		CC	cytoplasm	6	85.71	7.62E-02	9.79E-01
		MF	calmodulin binding	2	28.57	4.54E-02	9.15E-01
48	8	BP	cell motion	3	42.86	1.55E-02	8.29E-01
		CC	actin cytoskeleton	3	42.86	4.09E-03	1.61E-01
		MF	actin binding	3	42.86	6.54E-03	2.16E-01

Figure 5. Bulk network module ontology enrichment.A list of the modules output from the CODENSE algorithm ordered by module size, each enriched for BP, CC, and MC ontologies. Row color denotes a certain statistical significance as demarcated in the figure legend.

the the cell cycle enriched ontology from the bulk cell dataset compared to the nuclear division enriched ontology from the single-cell dataset (co-enrichment 1). It is possible that this washing out may be due to the averaging out of gene signals when measuring at the tissue-level. This same phenomenon may explain the even more extreme drop seen in co-enrichment 2, transnational elongation. Classically, these two signals are dominant among transcriptomic signals. These two co-enrichment comparisons give us additional candidate genes to investigate for signs of this averaging phenomenon aforementioned.

Discussion

Our analysis illustrates the potential of single-cell co-expression network module enrichment analysis as a potential tool for biological inquiry. We find that, when taking the contextual approach to comparing the sets of reported ontologies, most of the more significantly enriched ontologies reported by the bulk data were related to those reported by single-celled data. Furthermore, when taking a more refined, gene-level approach to this comparison a handful of modules are still in agreement.

Between the module and gene-level comparisons of the two workflows' outputs, we are confident that the single-cell data is picking up on the same biological signals as the traditional bulk cell pipeline. But, there were also noticable differences. Single-celled data innately lacks extracellular signals and this is evident in our analysis by the absence of extracellular matrix organization and cellular adhesion ontologies and genes from the single-celled output. This may be of particular interest to studies that are specifically interested in intracellular environments rather than entire microenvironments.

Further comparisons and contrast may indeed be possible. Conservative filtering in the single-celled pipline may be limiting the results in this work. Accounting for noise in single-celled data could strengthen the claims possible in co-expression module enrichment comparisons in the future.

References

1. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–1401. Available from: http://www.sciencemag.org/content/344/6190/1396.abstract.

2. Bao ZS, Chen HM, Yang MY, Zhang CB, Yu K, Ye WL, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Research. 2014 Aug;24(11):1765–1773. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216918/.

3. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotech. 2012 Aug;30(8):777–782. Available from: http://dx.doi.org/10.1038/nbt.2282.

4. Langmead B, Trapnell C, Pop M, Salzberg S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10(3):R25. Available from: http://genomebiology.com/2009/10/3/R25.

5. Li B, Dewey C. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323. Available from: http://www.biomedcentral.com/1471-2105/12/323.

6. Hu H, Yan X, Huang Y, Han J, Zhou XJ. Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics. 2005;21(suppl 1):i213–i221. Available from: http://bioinformatics.oxfordjournals.org/content/21/suppl_1/i213.abstract.

7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols. 2008 Dec;4(1):44–57. Available from: http://dx.doi.org/10.1038/nprot.2008.211.

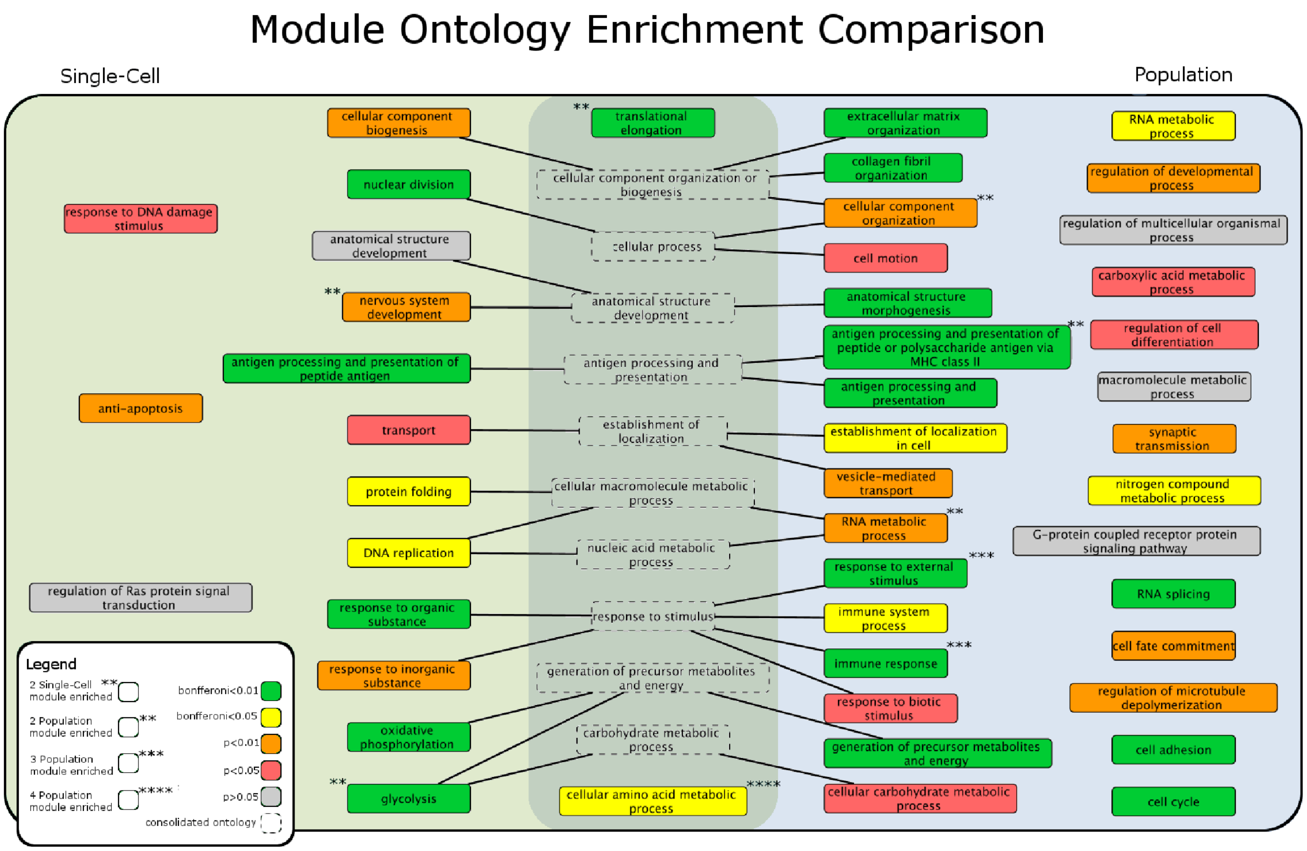

Module Ontology Enrichment Comparison
Single-Cell
Population
cellular component biogenesis
nuclear division
response to DNA damage stimulus
anatomical structure development
** nervous system development
antigen processing and presentation of peptide antigen
anti-apoptosis
transport
protein folding
DNA replication
regulation of Ras protein signal transduction
response to organic substance
response to inorganic substance
oxidative phosphorylation
** glycolysis
** translational elongation
cellular component organization or biogenesis
cellular process
anatomical structure development
antigen processing and presentation
establishment of localization
cellular macromolecule metabolic process
nucleic acid metabolic process
response to stimulus
generation of precursor metabolites and energy
carbohydrate metabolic process
cellular amino acid metabolic process ****
extracellular matrix organization
collagen fibril organization
cellular component organization **
cell motion
anatomical structure morphogenesis
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II **
antigen processing and presentation
establishment of localization in cell
vesicle-mediated transport
RNA metabolic process **
response to external stimulus ***
immune system process
immune response ***
response to biotic stimulus
generation of precursor metabolites and energy
cellular carbohydrate metabolic process
RNA metabolic process
regulation of developmental process
regulation of multicellular organismal process
carboxylic acid metabolic process
regulation of cell differentiation
macromolecule metabolic process
synaptic transmission
nitrogen compound metabolic process
G-protein coupled receptor protein signaling pathway
RNA splicing
cell fate commitment
regulation of microtubule depolymerization
cell adhesion
cell cycle
Legend
2 Single-Cell module enriched ** □
2 Population module enriched □**
3 Population module enriched □***
4 Population module enriched □****
bonfferoni<0.01 🟩
bonfferoni<0.05 🟨
p<0.01 🟧
p<0.05 🟥
p>0.05 ⬜
consolidated ontology ⬚

Figure 6. Enriched module comparison

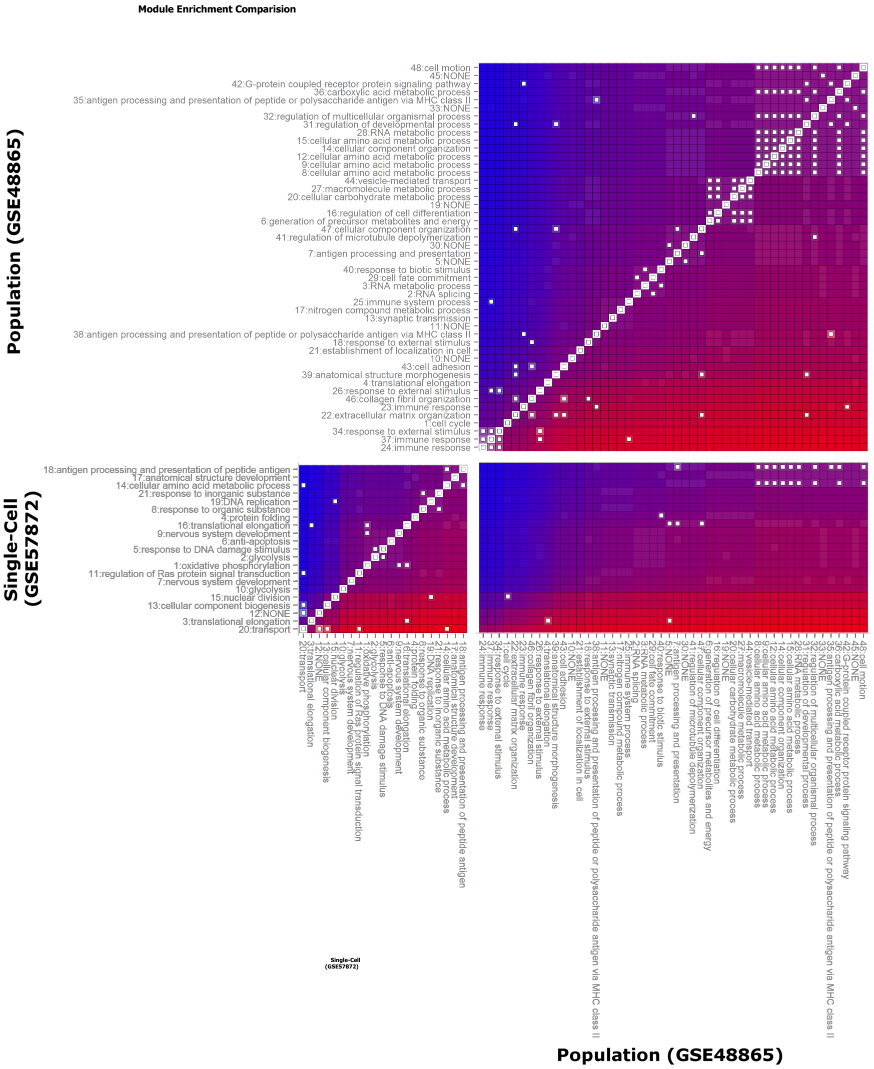
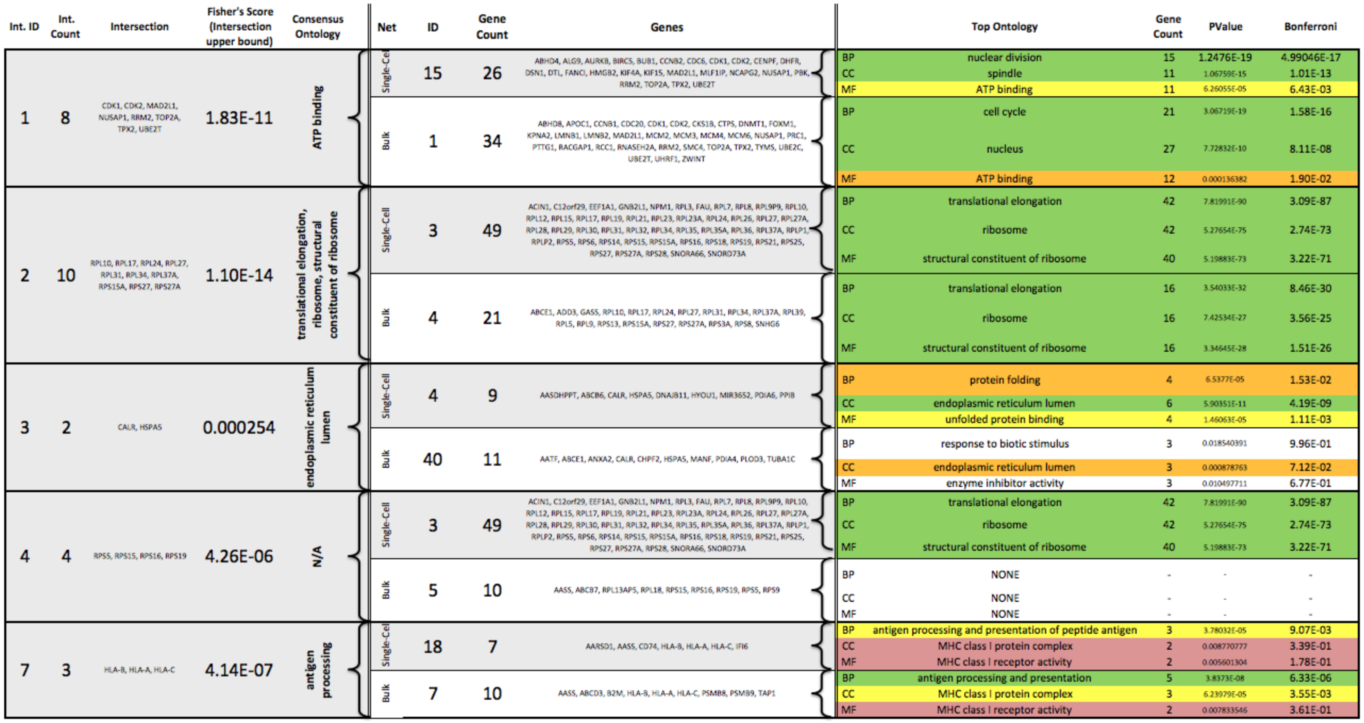


Figure 7. Gene-level starry map comparison of output modules. A small white box is placed in all module by module comparison that achieved a Fisher's exact significance of less than or equal to 0.05.



Int. ID	Int. Count	Intersection	Fisher's Score (Intersection upper bound)	Consensus Ontology	Net	ID	Gene Count	Genes		Top Ontology	Gene Count	PValue	Bonferroni
1	8	CDK1, CDK2, MAD2L1, NUSAP1, RRM2, TOP2A, TPX2, UBE2T	1.83E-11	ATP binding	Single-Cell	15	26	ABHD4, ALG9, AURKB, BIRC5, BUB1, CCNB2, CDC6, CDK1, CDK2, CENPF, DHFR, DSN1, DTL, FANCI, HMGB2, KIF4A, KIF15, MAD2L1, MLF1IP, NCAPG2, NUSAP1, PBK, RRM2, TOP2A, TPX2, UBE2T	BP	nuclear division	15	1.2476E-19	4.99046E-17
									CC	spindle	11	1.06759E-15	1.01E-13
									MF	ATP binding	11	6.26055E-05	6.43E-03
					Bulk	1	34	ABHD8, APOC1, CCNB1, CDC20, CDK1, CDK2, CKS1B, CTPS, DNMT1, FOXM1, KPNA2, LMNB1, LMNB2, MAD2L1, MCM2, MCM3, MCM4, MCM6, NUSAP1, PRC1, PTTG1, RACGAP1, RCC1, RNASEH2A, RRM2, SMC4, TOP2A, TPX2, TYMS, UBE2C, UBE2T, UHRF1, ZWINT	BP	cell cycle	21	3.06719E-19	1.58E-16
									CC	nucleus	27	7.72832E-10	8.11E-08
									MF	ATP binding	12	0.000136382	1.90E-02
2	10	RPL10, RPL17, RPL24, RPL27, RPL31, RPL34, RPL37A, RPS15A, RPS27, RPS27A	1.10E-14	translational elongation, ribosome, structural constituent of ribosome	Single-Cell	3	49	ACIN1, C12orf29, EEF1A1, GNB2L1, NPM1, RPL3, FAU, RPL7, RPL8, RPL9P9, RPL10, RPL12, RPL15, RPL17, RPL19, RPL21, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL29, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL37A, RPLP1, RPLP2, RPS5, RPS6, RPS14, RPS15, RPS15A, RPS16, RPS18, RPS19, RPS21, RPS25, RPS27, RPS27A, RPS28, SNORA66, SNORD73A	BP	translational elongation	42	7.81991E-90	3.09E-87
									CC	ribosome	42	5.27654E-75	2.74E-73
									MF	structural constituent of ribosome	40	5.19883E-73	3.22E-71
					Bulk	4	21	ABCE1, ADD3, GAS5, RPL10, RPL17, RPL24, RPL27, RPL31, RPL34, RPL37A, RPL39, RPL5, RPL9, RPS13, RPS15A, RPS27, RPS27A, RPS3A, RPS8, SNHG6	BP	translational elongation	16	3.54033E-32	8.46E-30
									CC	ribosome	16	7.42534E-27	3.56E-25
									MF	structural constituent of ribosome	16	3.34645E-28	1.51E-26
3	2	CALR, HSPA5	0.000254	endoplasmic reticulum lumen	Single-Cell	4	9	AASDHPPT, ABCB6, CALR, HSPA5, DNAJB11, HYOU1, MIR3652, PDIA6, PPIB	BP	protein folding	4	6.5377E-05	1.53E-02
									CC	endoplasmic reticulum lumen	6	5.90351E-11	4.19E-09
									MF	unfolded protein binding	4	1.46063E-05	1.11E-03
					Bulk	40	11	AATF, ABCE1, ANXA2, CALR, CHPF2, HSPA5, MANF, PDIA4, PLOD3, TUBA1C	BP	response to biotic stimulus	3	0.018540391	9.96E-01
									CC	endoplasmic reticulum lumen	3	0.000878763	7.12E-02
									MF	enzyme inhibitor activity	3	0.010497711	6.77E-01
4	4	RPS5, RPS15, RPS16, RPS19	4.26E-06	N/A	Single-Cell	3	49	ACIN1, C12orf29, EEF1A1, GNB2L1, NPM1, RPL3, FAU, RPL7, RPL8, RPL9P9, RPL10, RPL12, RPL15, RPL17, RPL19, RPL21, RPL23, RPL23A, RPL24, RPL26, RPL27, RPL27A, RPL28, RPL29, RPL30, RPL31, RPL32, RPL34, RPL35, RPL35A, RPL36, RPL37A, RPLP1, RPLP2, RPS5, RPS6, RPS14, RPS15, RPS15A, RPS16, RPS18, RPS19, RPS21, RPS25, RPS27, RPS27A, RPS28, SNORA66, SNORD73A	BP	translational elongation	42	7.81991E-90	3.09E-87
									CC	ribosome	42	5.27654E-75	2.74E-73
									MF	structural constituent of ribosome	40	5.19883E-73	3.22E-71
					Bulk	5	10	AASS, ABCB7, RPL13AP5, RPL18, RPS15, RPS16, RPS19, RPS5, RPS9	BP	NONE	-	-	-
									CC	NONE	-	-	-
									MF	NONE	-	-	-
7	3	HLA-B, HLA-A, HLA-C	4.14E-07	antigen processing	Single-Cell	18	7	AARSD1, AASS, CD74, HLA-B, HLA-A, HLA-C, IFI6	BP	antigen processing and presentation of peptide antigen	3	3.78032E-05	9.07E-03
									CC	MHC class I protein complex	2	0.008770777	3.39E-01
									MF	MHC class I receptor activity	2	0.005601304	1.78E-01
					Bulk	7	10	AASS, ABCD3, B2M, HLA-B, HLA-A, HLA-C, PSMB8, PSMB9, TAP1	BP	antigen processing and presentation	5	3.8373E-08	6.33E-06
									CC	MHC class I protein complex	3	6.23979E-05	3.55E-03
									MF	MHC class I receptor activity	2	0.007833546	3.61E-01

Figure 8. Interesting significant Gene-Level module comparisons. This is a detailed look at 5 significant module co-enrichments between the bulk and single-cell output. Co-enrichemnt was calculated with the fisher's exact score using only the genes present in the intersection of both bulk and single-cell co-expression networks.