

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

Brian Arand<sup>2</sup>, Raghu Machiraju<sup>1, 2</sup>, Kun Huang<sup>1, 2</sup>

**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

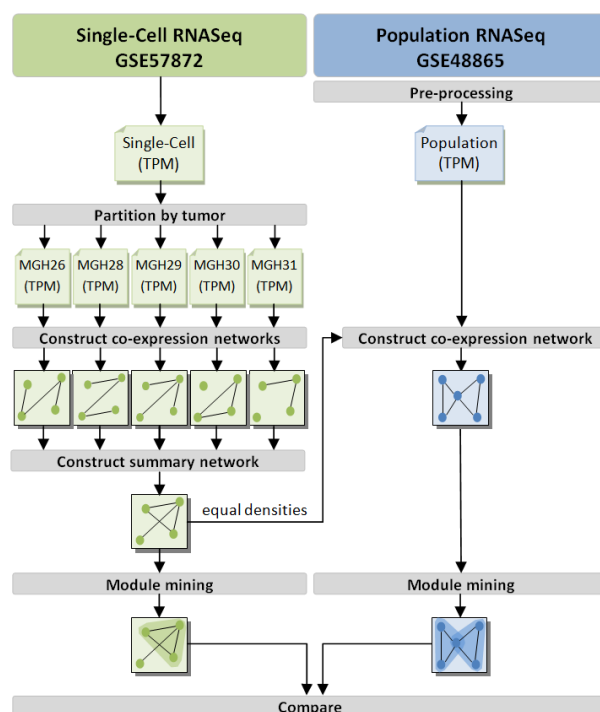
We compare our resulting networks both at the ontology enriched module level and at the gene level...

## 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 population-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 population of 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, population-of-cell (or tissue) samples. From here on, GSE57872 and GSE48865 samples will be referred to as 'single-cell' and 'population' 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.



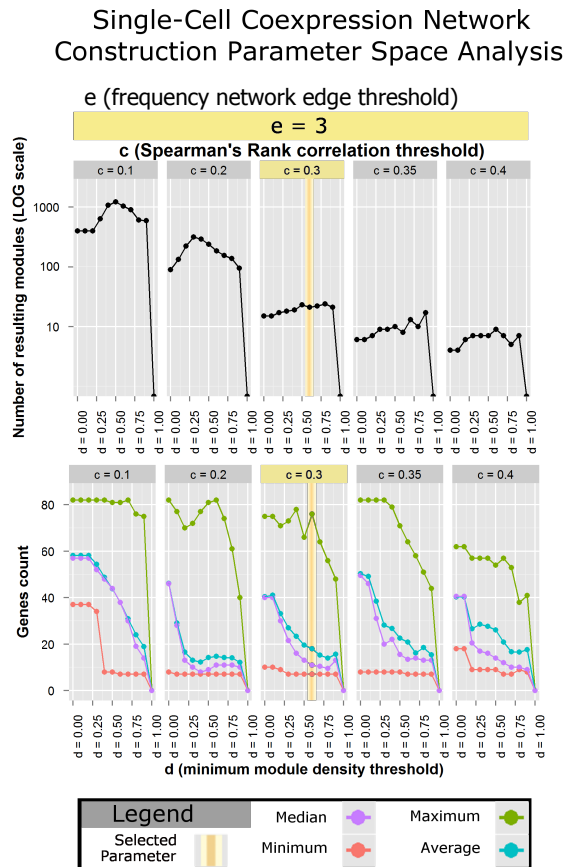
**Figure 1. Network comparison workflow.** Note that there is no partition nor aggregation step in the population RNASeq pipeline. Moreover, note that edges of the constructed population co-expression network are filtered to achieve the closest network density possible to that realized in the Single-Cell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

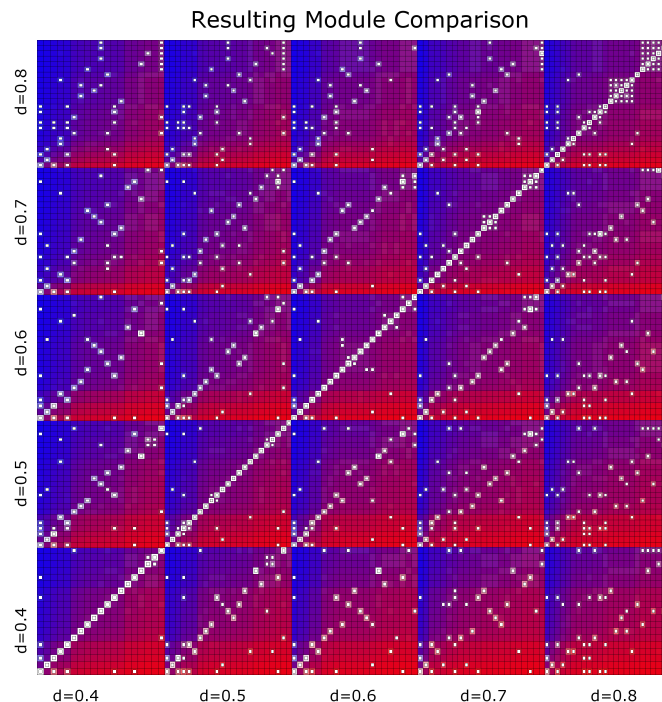
$$y_i = \frac{\sum_s \log_2(x_{si} + 1)}{N_t} \quad (1)$$

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

### 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:

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



**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

Where  $\rho_{ijt}$  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.

$$B_t = \begin{bmatrix} b_{11t} & \dots & b_{1kt} \\ \vdots & \ddots & \vdots \\ b_{k1t} & \dots & b_{kkt} \end{bmatrix}, b_{ijt} = \begin{cases} 1, & |\rho_{ijt}| \geq c \\ 0, & o.w. \end{cases} \quad (4)$$

This is analogous to cutting a co-expression network at a given edge threshold. After analyzing the distribution of resulting networks' densities 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.

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

## 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].

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

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

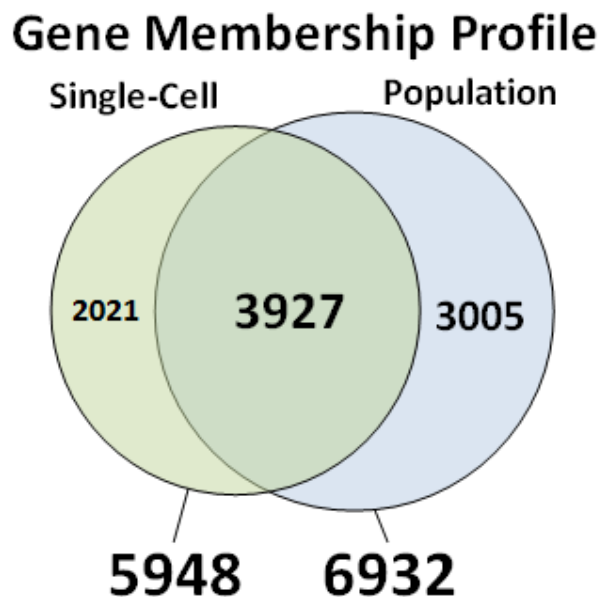
We make a considerable effort to ensure that the resulting modules output by CODENSE at the selected parameterization are robust to change withing the algorithm's parameterization space. Fig. 2 shows that the size and number of output modules are relatively stable at the selected values for  $e$ ,  $c$ , 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 all did not change the gene membership of the sets much.

## 0.1 Population Tissue Samples

Our population-tissue 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  $y_i \leq 4.5$ . 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. 4.



**Figure 4. Sample gene profile.**

### 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 population 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 population data set.

We compared these output module sets at the module level and the gene level. 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. 5 and for the population data in fig. 6.

We then profiled the two lists of output ontology enriched modules to find both consensus and marginal ontologies between the two data sets. Fig. 8 illustrates this module level comparison. When using exact equality to define the intersection between the two sets of ontologies only find two ontologies were identified by both single-cell and population pipelines. However, many of the output modules share common ancestral ontologies. To get a better understanding of the similarities of the pipelines' output, we report what we will be calling 'consolidated ontologies' in the intersection. Consolidated ontologies are ancestral ontologies from the Gene Ontology Consortium's ontology hierarchy database that link at least one ontology reported by the single-cell workflow to at least one ontology reported by the population workflow that is no more than two generations ancestor to the output ontologies from which the consolidated ontology was inferred. If two consolidated ontologies were inferred from the same set of output ontologies, only the most specific consolidated ontology is reported.

Module		Enriched Subset						Module		Enriched Subset					
ID	Gene Count	Top Ontology	Gene Count	%	PValue	Bonferroni		ID	Gene Count	Top Ontology	Gene Count	%	PValue	Bonferroni	
20	76	BP	transport	14	21.88	1.30E-02	1.00E+00	5	11	BP	response to DNA damage stimulus	3	27.27	1.34E-02	9.65E-01
		CC	NONE	-	-	-	-			CC	nuclear lumen	5	45.45	5.95E-03	2.45E-01
		MF	Ras GTPase activator activity	3	4.69	2.87E-02	9.96E-01			MF	NONE	-	-	-	-
3	49	BP	translational elongation	42	85.71	7.82E-06	3.09E-01	6	11	BP	anti-apoptosis	3	27.27	7.13E-03	8.78E-01
		CC	ribosome	42	85.71	5.28E-05	2.74E-01			CC	extracellular region	6	54.55	2.60E-03	3.77E-01
		MF	structural constituent of ribosome	40	81.63	5.20E-03	3.22E-01			MF	protein binding	9	81.82	3.31E-02	7.73E-01
12	42	BP	NONE	-	-	-	-	9	11	BP	nervous system development	4	36.36	2.70E-02	9.99E-01
		CC	NONE	-	-	-	-			CC	microtubule	3	27.27	1.21E-02	5.31E-01
		MF	NONE	-	-	-	-			MF	protein binding	7	63.64	9.18E-02	9.96E-01
13	26	BP	cellular component biogenesis	5	21.74	5.54E-03	9.05E-01	16	10	BP	translational elongation	5	55.56	1.69E-02	4.56E-01
		CC	NONE	-	-	-	-			CC	cytosolic ribosome	5	55.56	4.30E-03	2.71E-01
		MF	NONE	-	-	-	-			MF	structural constituent of ribosome	5	55.56	9.88E-02	5.34E-01
15	26	BP	nuclear division	15	57.69	1.25E-19	4.99E-17	4	9	BP	protein folding	4	50.00	6.54E-05	1.53E-02
		CC	signaling	11	42.31	1.07E-15	1.01E-13			CC	endoplasmic reticulum lumen	6	75.00	5.90E-11	4.19E-09
		MF	ATP binding	11	42.31	6.26E-05	6.43E-03			MF	unfolded protein binding	4	50.00	1.46E-05	1.11E-03
10	14	BP	glycolysis	3	23.08	4.80E-04	2.08E-01	8	8	BP	response to organic substance	6	75.00	6.61E-06	1.63E-03
		CC	extracellular region part	5	38.46	4.41E-03	3.04E-01			CC	synaptosome	2	25.00	2.64E-02	6.29E-01
		MF	protein binding	10	76.92	1.96E-02	8.62E-01			MF	protein dimerization activity	3	37.50	2.38E-02	7.21E-01
7	13	BP	nervous system development	6	50.00	4.89E-04	2.64E-01	19	8	BP	DNA replication	4	50.00	8.07E-03	1.30E-02
		CC	cell fraction	4	33.33	3.44E-02	8.90E-01			CC	nucleoplasm	4	50.00	2.99E-03	3.44E-01
		MF	transporter activity	6	50.00	8.66E-04	1.14E-01			MF	ATP binding	4	50.00	2.40E-02	8.64E-01
11	12	BP	regulation of Ras protein signal transduction	2	20.00	5.82E-02	1.00E+00	21	8	BP	response to inorganic substance	3	37.50	2.04E-03	2.06E-01
		CC	NONE	-	-	-	-			CC	cytoplasm	4	50.00	9.74E-02	9.83E-01
		MF	ATPase activity, coupled to transmembrane movement of substances	2	20.00	2.85E-02	9.12E-01			MF	coppper ion binding	4	50.00	3.31E-04	3.94E-01
1	11	BP	oxidative phosphorylation	4	36.36	3.76E-05	7.56E-03	14	7	BP	cellular amino acid metabolic process	2	33.33	3.18E-02	9.02E-01
		CC	mitochondrial inner membrane	6	54.55	5.94E-07	3.92E-05			CC	NONE	-	-	-	-
		MF	hydrogen ion transmembrane transporter activity	4	36.36	1.66E-05	8.80E-04			MF	NONE	-	-	-	-
2	11	BP	glycolysis	4	36.36	1.91E-06	4.36E-04	17	7	BP	anatomical structure development	4	57.14	7.48E-02	1.00E+00
		CC	NONE	-	-	-	-			CC	cell fraction	3	42.86	4.03E-02	8.82E-01
		MF	NONE	-	-	-	-			MF	NONE	-	-	-	-
		BP	antigen processing and presentation of peptide antigen	3	50.00	3.78E-05	9.07E-03	18	7	CC	MHC class I protein complex	2	33.33	8.77E-03	3.39E-01
		CC	NONE	-	-	-	-			MF	MHC class I receptor activity	2	33.33	5.60E-03	1.78E-01
		MF	NONE	-	-	-	-								

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

## Discussion

## References

- 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>.
- 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/>.
- 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.



Module		Enriched Subset				
ID	Gene Count	Top Ontology	Gene Count	%	PValue	Benfoni
24	51	BP	18	34.82	8.50E-12	7.91E-09
		CC	8	15.38	3.50E-08	4.55E-04
		MF	3	5.77	2.08E-04	4.08E-02
37	48	BP	14	28.76	5.57E-08	8.69E-05
		CC	5	10.44	3.18E-04	3.88E-02
		MF	16	34.04	4.02E-04	6.87E-02
34	46	BP	14	11.11	1.04E-06	7.68E-04
		CC	12	26.67	6.83E-12	1.09E-09
		MF	5	11.11	4.16E-05	6.97E-03
1	34	BP	21	63.64	3.07E-19	1.58E-16
		CC	27	81.82	7.73E-10	8.11E-08
		MF	12	36.36	1.86E-04	1.90E-02
22	34	BP	12	36.36	3.97E-17	1.79E-14
		CC	19	57.58	2.90E-23	4.20E-21
		MF	13	39.39	6.50E-20	7.48E-18
23	31	BP	12	40.00	1.39E-08	7.17E-06
		CC	18	60.00	3.18E-05	3.55E-03
		MF	10	33.33	3.81E-03	5.19E-01
46	26	BP	5	20.83	6.80E-08	3.51E-05
		CC	11	45.83	1.80E-11	2.39E-09
		MF	5	20.83	1.09E-09	1.29E-07
26	24	BP	12	52.17	1.53E-08	8.49E-06
		CC	11	47.91	5.61E-05	5.60E-03
		MF	2	8.70	4.35E-03	4.45E-01
4	21	BP	16	80.00	3.54E-32	8.46E-30
		CC	16	80.00	7.43E-27	3.56E-25
		MF	10	50.00	1.35E-28	1.51E-26
39	17	BP	9	54.25	2.29E-06	1.14E-03
		CC	3	18.75	1.12E-04	1.12E-02
		MF	14	87.50	8.42E-03	5.54E-01
43	16	BP	4	46.67	1.85E-05	5.13E-03
		CC	4	26.67	2.49E-03	7.38E-01
		MF	3	20.00	2.78E-03	1.99E-01
10	15	BP	-	-	-	-
		CC	-	-	-	-
		MF	-	-	-	-
21	15	BP	6	42.86	1.55E-04	2.91E-02
		CC	9	64.29	1.68E-07	1.36E-05
		MF	3	21.43	1.29E-03	4.52E-01
18	13	BP	6	50.00	1.14E-04	5.31E-02
		CC	6	50.00	3.97E-02	9.28E-01
		MF	2	16.67	9.30E-02	1.05E-00
38	13	BP	9	75.00	1.00E-20	4.37E-18
		CC	8	66.67	3.65E-18	1.94E-16
		MF	7	58.33	7.46E-16	4.66E-14
11	12	BP	-	-	-	-
		CC	7	63.64	5.87E-02	9.11E-01
		MF	-	-	-	-
13	12	BP	3	27.27	8.70E-03	8.59E-01
		CC	3	18.18	2.24E-03	1.35E-01
		MF	2	18.18	2.97E-03	1.81E-01
17	12	BP	9	81.82	1.80E-04	4.60E-02
		CC	8	72.73	7.94E-05	5.62E-03
		MF	5	45.45	5.99E-05	5.67E-03
25	12	BP	6	54.55	3.26E-05	1.30E-02
		CC	3	27.27	3.44E-03	1.98E-01
		MF	3	27.27	2.90E-02	9.37E-01
2	11	BP	6	60.00	3.75E-07	4.84E-05
		CC	7	70.00	8.85E-08	4.06E-06
		MF	7	70.00	8.27E-07	4.38E-05
3	11	BP	5	50.00	5.76E-04	1.11E-01
		CC	10	100.00	2.01E-03	9.94E-02
		MF	3	30.00	6.48E-02	9.78E-01
29	11	BP	3	30.00	1.96E-03	3.41E-01
		CC	-	-	-	-
		MF	4	40.00	2.54E-02	8.24E-01
40	11	BP	3	30.00	8.95E-02	9.56E-01
		CC	3	30.00	6.79E-04	7.12E-02
		MF	3	30.00	1.05E-02	6.77E-01
5	10	BP	-	-	-	-
		CC	-	-	-	-
		MF	-	-	-	-

Module		Enriched Subset				
ID	Gene Count	Top Ontology	Gene Count	%	PValue	Benfoni
7	10	BP	5	42.86	1.84E-08	6.13E-06
		CC	3	37.50	6.24E-05	3.55E-03
		MF	2	25.00	7.81E-01	3.61E-01
30	10	BP	-	-	-	-
		CC	2	2.27	1.15E-02	6.17E-01
		MF	3	3.41	1.05E-03	6.52E-02
41	10	BP	2	22.22	6.29E-03	7.45E-01
		CC	5	55.56	1.62E-02	6.64E-01
		MF	6	66.67	2.64E-03	3.74E-01
47	10	BP	2	22.22	8.52E-03	4.65E-01
		CC	3	33.33	1.39E-02	6.30E-01
		MF	3	42.50	1.81E-06	2.23E-03
6	9	BP	4	50.00	3.48E-06	2.22E-04
		CC	3	37.50	1.17E-04	8.88E-01
		MF	-	-	-	-
16	9	BP	3	37.50	1.64E-02	9.98E-01
		CC	2	25.00	5.27E-02	9.79E-01
		MF	-	-	-	-
19	9	BP	-	-	-	-
		CC	3	37.50	2.23E-02	7.48E-01
		MF	-	-	-	-
20	9	BP	3	37.50	1.44E-02	9.81E-01
		CC	6	75.00	3.21E-02	8.26E-01
		MF	-	-	-	-
27	9	BP	4	50.00	6.62E-02	1.00E+00
		CC	-	-	-	-
		MF	4	50.00	1.48E-02	6.47E-01
44	9	BP	4	50.00	1.23E-03	2.68E-01
		CC	7	87.50	4.37E-03	3.02E-01
		MF	2	25.00	4.58E-02	9.84E-01
8	8	BP	2	28.57	7.75E-02	1.00E+00
		CC	-	-	-	-
		MF	5	71.43	5.03E-02	9.77E-01
9	8	BP	3	42.86	2.47E-03	2.55E-01
		CC	-	-	-	-
		MF	5	71.43	5.01E-02	9.52E-01
12	8	BP	2	28.57	7.75E-02	9.98E-01
		CC	3	42.86	5.82E-02	9.65E-01
		MF	-	-	-	-
14	8	BP	4	57.14	7.26E-02	1.00E+00
		CC	-	-	-	-
		MF	2	33.33	8.73E-02	9.94E-01
15	8	BP	2	33.33	7.75E-02	1.00E+00
		CC	3	37.50	1.13E-02	8.56E-01
		MF	1	12.50	4.47E-03	1.37E-01
28	8	BP	5	71.43	2.41E-04	3.56E-02
		CC	4	57.14	1.22E-02	3.42E-01
		MF	6	85.71	2.04E-02	7.95E-01
31	8	BP	4	57.14	1.01E-03	1.78E-01
		CC	3	42.86	3.52E-03	1.53E-02
		MF	3	42.86	4.71E-04	2.60E-02
32	8	BP	3	42.86	5.52E-02	1.00E+00
		CC	5	71.43	1.13E-02	8.56E-01
		MF	3	42.86	2.42E-02	1.52E-01
33	8	BP	-	-	-	-
		CC	4	57.14	4.47E-05	3.87E-03
		MF	2	28.57	3.07E-02	7.85E-01
35	8	BP	5	71.43	3.70E-10	1.07E-07
		CC	4	57.14	1.08E-07	5.79E-06
		MF	4	57.14	3.34E-08	2.08E-06
36	8	BP	3	42.86	1.43E-02	8.94E-01
		CC	6	85.71	7.62E-02	9.33E-01
		MF	-	-	-	-
42	8	BP	3	42.86	5.38E-02	1.00E+00
		CC	6	85.71	3.42E-03	1.64E-01
		MF	5	71.43	9.77E-04	7.52E-02
45	8	BP	-	-	-	-
		CC	6	85.71	7.62E-02	9.79E-01
		MF	2	28.57	4.54E-02	9.15E-01
48	8	BP	3	42.86	1.55E-02	8.25E-01
		CC	3	42.86	4.09E-03	1.61E-01
		MF	3	42.86	6.54E-03	2.10E-01

Figure 6. Population network module ontology enrichment.

Nat Biotech. 2012 Aug;30(8):777–782. Available from: <http://dx.doi.org/10.1038/nbt.2282>.

- 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>.
- 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>.
- 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](http://bioinformatics.oxfordjournals.org/content/21/suppl_1/i213.abstract).
- 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>.



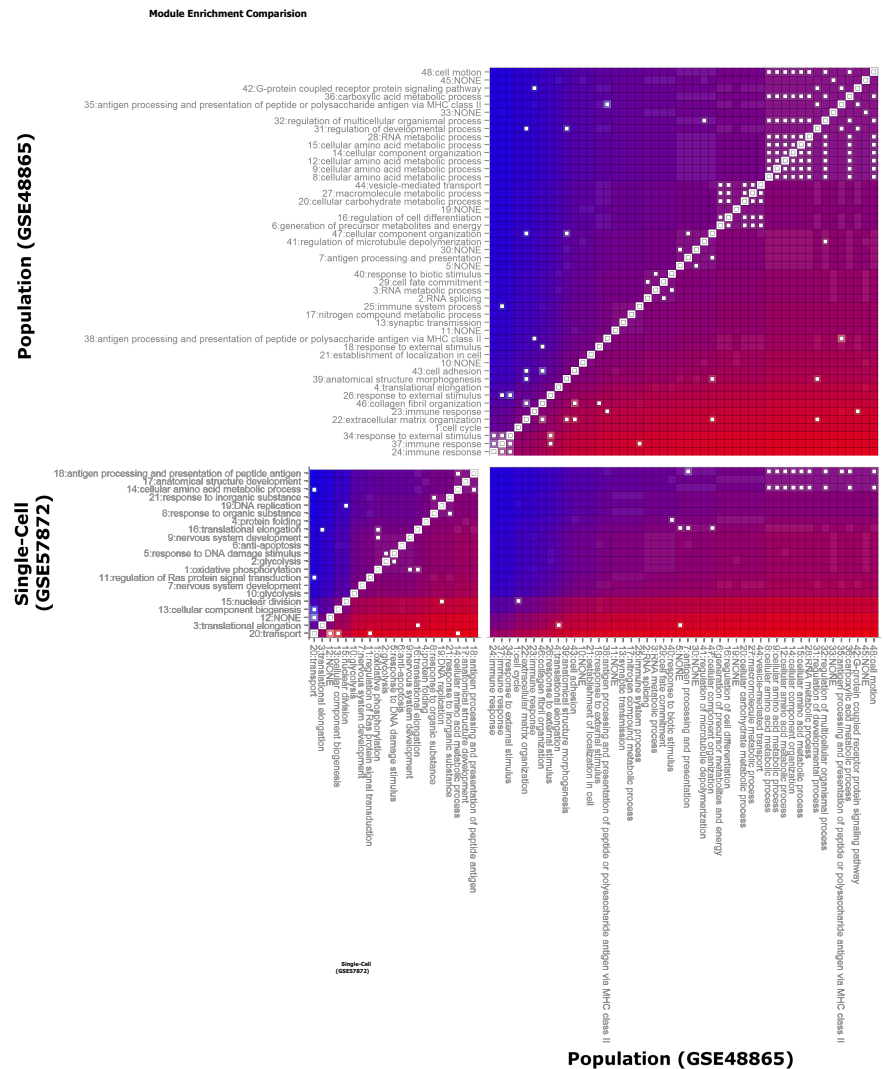
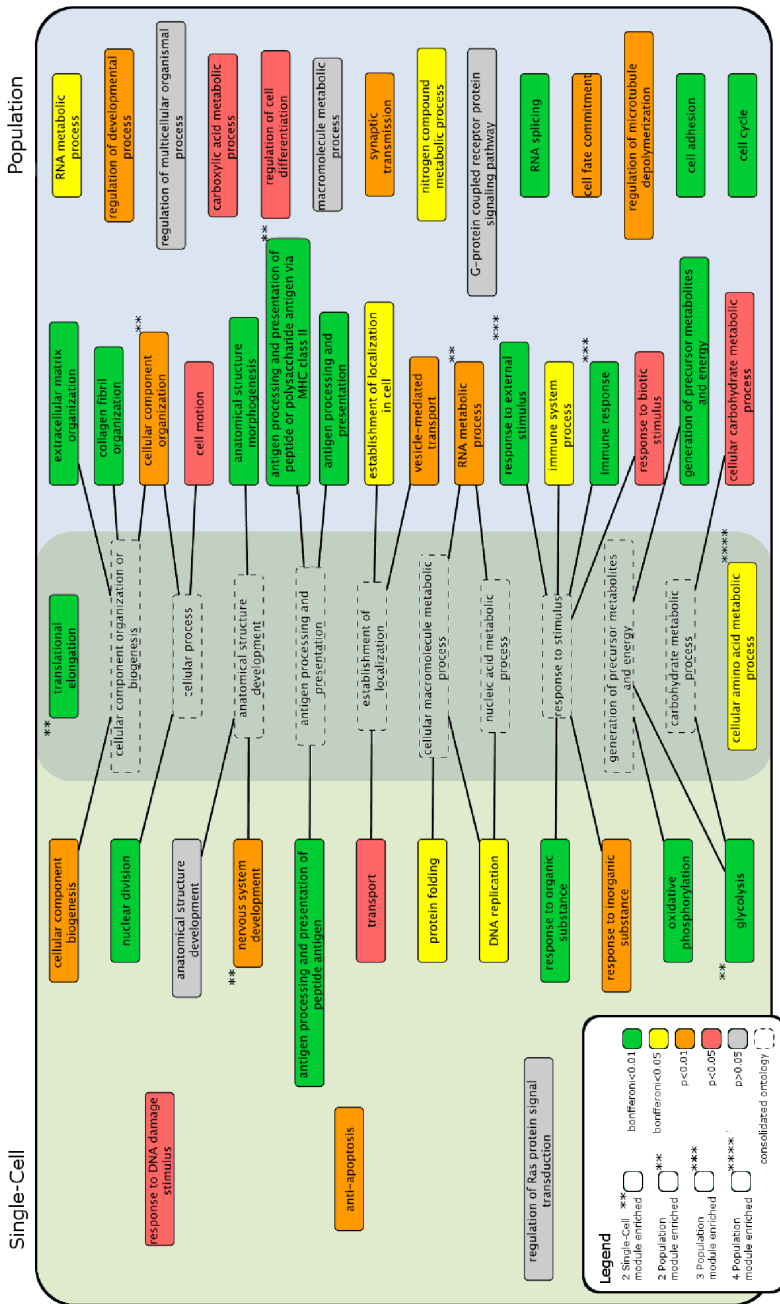


Figure 7. Gene-level starry map comparison of output modules.



**Figure 8.** Enriched module comparison