Colin Sheehan
3-19-20

Final Project: GSEA analysis comparing high versus low expressing tumors for a gene of interest, using publically available TCGA dataset

Gene set enrichment analysis is a powerful bioinformatics technique used frequently among biomedical researchers to infer the possible involvement of a biological process in large datasets. Specifically, this technique looks at the over-representation or under-representation of a gene set (usually some experimentally verified list of genes involved in a particular biological process) in some sample being tested, and it allows for hypothesis testing. The data input into the analysis normally takes the form of the change in gene expression between two samples, and ranked according change. For example, one might want to know if some biological process like glycolysis is predicted to be over/under-active in tumors from male versus female patients, or in tumors possessing a particular mutation versus wild-type, etc. Furthermore, some packages that perform the GSEA don't require a pre-decided gene set to test against the queried dataset; instead, it will run a comprehensive analysis using several gene sets and will return the most important biological processes from the data. Despite this powerful approach to make meaning from large amounts of data, or to validate the clinical significance of experimental data, GSEA is seldom used among cellular biologists without some level of computational background. Here, I perform a GSEA analysis in R using publically available TCGA data, comparing tumors with high versus low expression for a gene of interest to my lab, a transcription factor called BACH1. BACH1 is a heme-dependent transcription factor implicated in breast cancer, and high expression of BACH1 is associated with a very poor prognosis in patients.

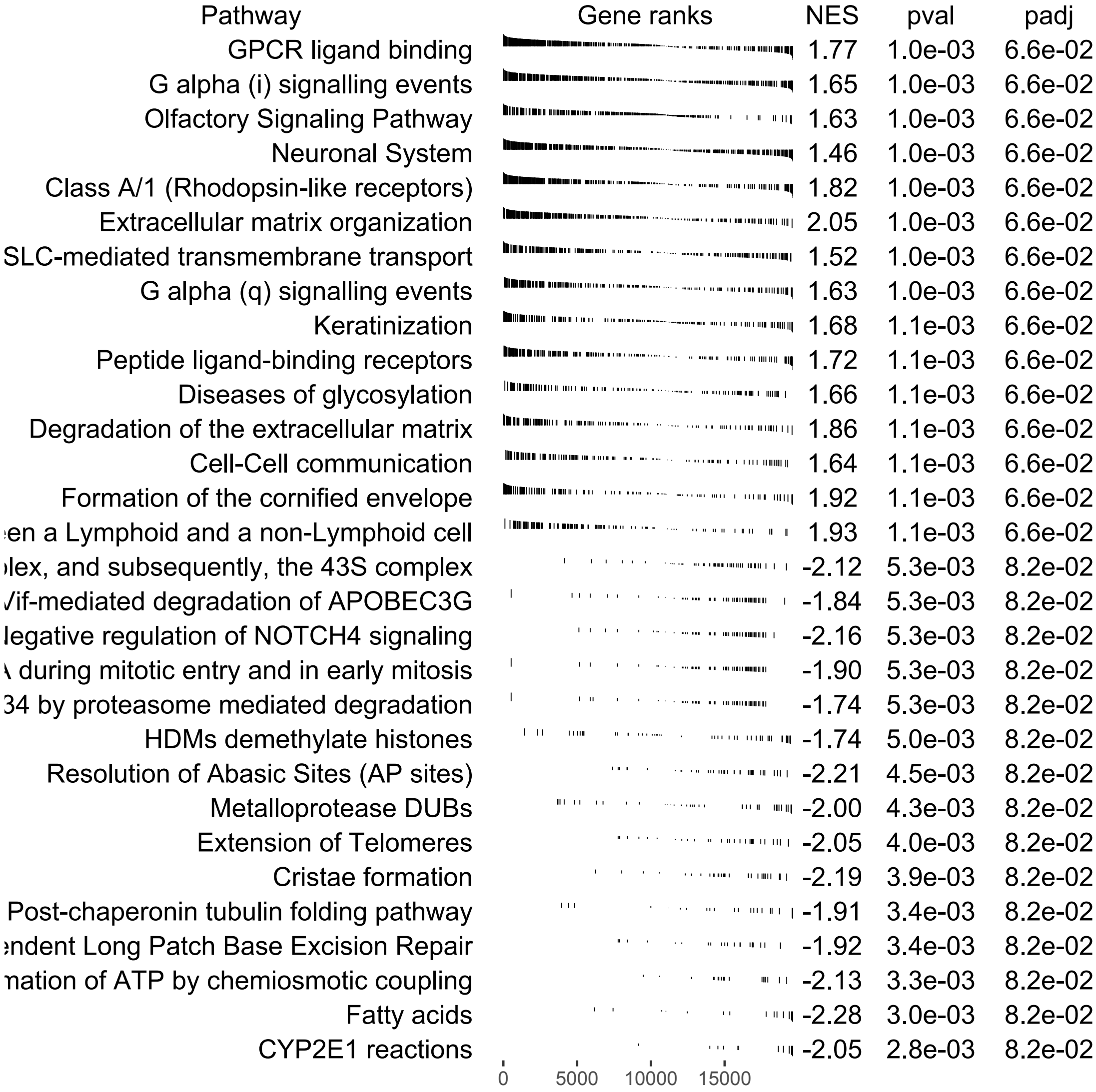
To conduct the GSEA analysis, I downloaded the breast cancer dataset (BRCA) from the publically available, cancer genome atlas program (TCGA); this contains FPKM-normalized mRNA expression values for over 60,000 genes and non-coding regulatory elements and over 1200 patient tumor samples. Samples were then organized into either high-BACH1 expression or low-BACH1 expression, defined as merely being greater or less than the BACH1 mean for all 1200 samples. The difference in the mean gene expression between these two groups was collected into a dataframe, and genes were ranked according to the difference in expression. Finally, the gene set enrichment analysis was performed using the fgsea package in R, which takes as input a named vector of the ranked genes along with their gene ID, as well as a vector containing the gene sets to query against. For the initial analysis, the reactome.db package was used to provide a comprehensive list of potential pathways to query based on the genes that were going to be input for my GSEA analysis.

One somewhat challenging aspect of this project was related to gene identification. The fgsea package uses genes in the form of an Entrez_ID name; however, the TCGA data provides genes in the form of an Ensemble_ID name. Because of this issue, I had to convert all the Ensemble_IDs from the TCGA data into their Entrez_ID form to perform the analysis. Luckily, this is not an uncommon issue among researchers; biomaRt is an R package that can be used to convert the gene names from a query into any other specified ID form. Unfortunately, the yield of successfully converted names was relatively low when the raw TCGA Ensemble IDs were input to biomRT. This may be partially due to the TCGA database using very uncommon Ensembl gene names. For example, if you were to look up the Ensemble ID name for BACH1, the ID that most often appears is ENSG00000156273.16. However, the ID for BACH1 that is used by the TCGA database is ENSG00000156273.14. As a result, it is not surprising that BioMart was unable to produce a conversion for this query. To resolve this issue—as suggested from an online resource where I discovered BioMart—the additional digits were removed to produce a more generalized Ensemble ID to be converted by BioMart (for example, using ENSG00000156273 as the ID for BACH1)("TCGA biomart…"). This greatly increased the yield of successfully converted genes; although for future work, I may just want to write a script to do this using a local dictionary. Despite the increased number of genes converted, there remained many genes without an Entrez ID as well as a few matching multiple Entrez IDs. To address this, I wrote a for-loop to iterate over the Ensemble genes from the TCGA dataframe and deal with each issue on an individual basis. Genes with successful conversion were matched with their Entrez ID. Genes missing a conversion were matched with an NA to signify missing data. Finally, the cases where genes had multiple IDs were simply matched to the first Entrez ID of the group. Ultimately, this produces a list vector containing all the appropriate pairings, which was then added onto the TCGA dataframe using the cbind command.

After performing the GSEA analysis, the fgsea package has many different ways of visualizing the results. Of the most immediate interest, I had it return the 15 most enriched gene sets along with the 15 most under-enriched, in the form of a gene rank plot (figure 1). Of the four most enriched processes were GPCR ligand binding, G protein signaling, Olfactory Signaling Pathway and Neuronal System. It is very interesting that genes related to neuronal processes were enriched with the BACH1 high samples; this is a theme that has previously appeared in our lab, although the BACH1 involvement in nervous system physiology is an area entirely unexplored.

A common visualization method for looking at the enrichment of individual gene sets is the enrichment plot. This plot provides a visualization of the enrichment score as a random walk down the ranked gene list. The enrichment score is represented by where the random walk produces the greatest deviation from 0 in the data (Subramanian et al., 2005). An example of an enrichment plot for one gene set with intercellular communication is shown in figure 2. I was interested in stem cell regulation, since the literature suggests BACH1 is implicated with a stem cell phenotype. An enrichment plot was produced from the "stem cell regulators" gene set provided from the reactome.db package. However, this result was not statistically significant (p-value = 0.515). This gene set also had a very limited number of potential genes to query relative to other pathways. In this instance, the user may want to create a custom gene set to query for enrichment analysis. For this, I produced a new gene set based on the genes included in a gene ontology term (GO Term), stem cell population maintenance, which has a list of 171 genes. To do this, I again used the biomaRt to acquire the entrez ID names for all the genes contained in this GO term, "GO:0019827." These were then formatted into the proper vector to run the fgsea analysis again. The results of this are displayed as the enrichment plot in figure 4. Further, these were statistically significant, with a p-value of 0.025.

In summary, I was able to perform a GSEA analysis to explore predicted biological processes implicated from differential gene expression in breast cancer data. The provided R script can be very easily amended to perform the GSEA on any gene of interest, or any particular TCGA dataset—lung, colon, breast, etc. In addition, I became versatile with the biomaRt package to convert list of genes to other types of identifiers, as well as quickly interface with other informatics methods like gene ontology (GO) terms.



Pathway	Gene ranks	NES	pval	padj
GPCR ligand binding		1.77	1.0e-03	6.6e-02
G alpha (i) signalling events		1.65	1.0e-03	6.6e-02
Olfactory Signaling Pathway		1.63	1.0e-03	6.6e-02
Neuronal System		1.46	1.0e-03	6.6e-02
Class A/1 (Rhodopsin-like receptors)		1.82	1.0e-03	6.6e-02
Extracellular matrix organization		2.05	1.0e-03	6.6e-02
SLC-mediated transmembrane transport		1.52	1.0e-03	6.6e-02
G alpha (q) signalling events		1.63	1.0e-03	6.6e-02
Keratinization		1.68	1.1e-03	6.6e-02
Peptide ligand-binding receptors		1.72	1.1e-03	6.6e-02
Diseases of glycosylation		1.66	1.1e-03	6.6e-02
Degradation of the extracellular matrix		1.86	1.1e-03	6.6e-02
Cell-Cell communication		1.64	1.1e-03	6.6e-02
Formation of the cornified envelope		1.92	1.1e-03	6.6e-02
en a Lymphoid and a non-Lymphoid cell		1.93	1.1e-03	6.6e-02
lex, and subsequently, the 43S complex		-2.12	5.3e-03	8.2e-02
/if-mediated degradation of APOBEC3G		-1.84	5.3e-03	8.2e-02
legative regulation of NOTCH4 signaling		-2.16	5.3e-03	8.2e-02
\ during mitotic entry and in early mitosis		-1.90	5.3e-03	8.2e-02
34 by proteasome mediated degradation		-1.74	5.3e-03	8.2e-02
HDMs demethylate histones		-1.74	5.0e-03	8.2e-02
Resolution of Abasic Sites (AP sites)		-2.21	4.5e-03	8.2e-02
Metalloprotease DUBs		-2.00	4.3e-03	8.2e-02
Extension of Telomeres		-2.05	4.0e-03	8.2e-02
Cristae formation		-2.19	3.9e-03	8.2e-02
Post-chaperonin tubulin folding pathway		-1.91	3.4e-03	8.2e-02
endent Long Patch Base Excision Repair		-1.92	3.4e-03	8.2e-02
nation of ATP by chemiosmotic coupling		-2.13	3.3e-03	8.2e-02
Fatty acids		-2.28	3.0e-03	8.2e-02
CYP2E1 reactions		-2.05	2.8e-03	8.2e-02

Figure 1: Gene rank plot of top over/under-enriched gene sets for the BACH1 high vs. BACH1 low breast cancers

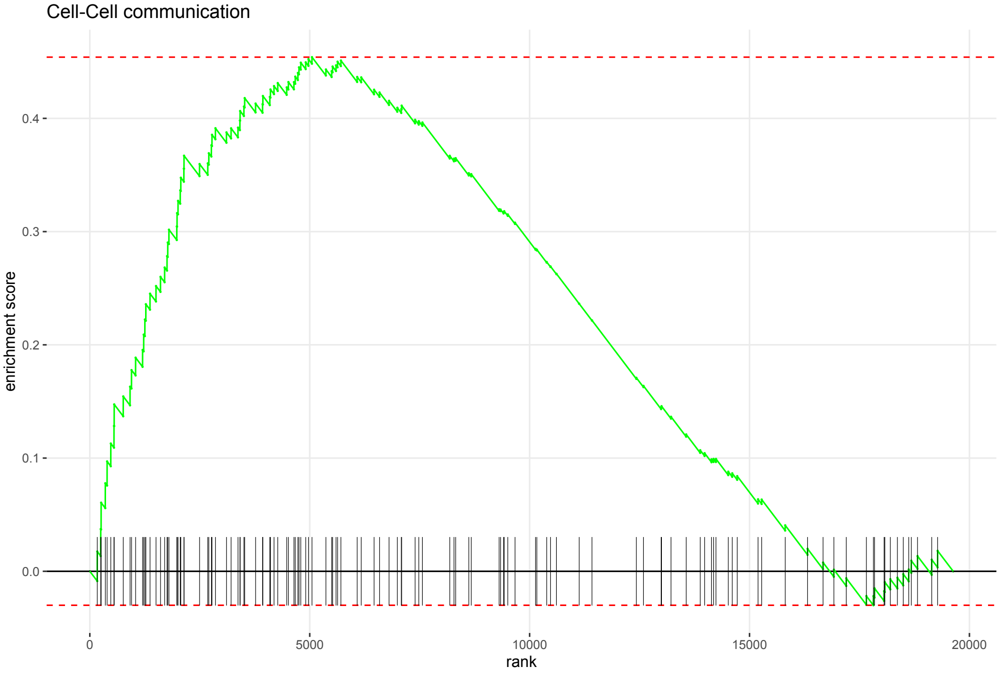


Figure 2: Gene Enrichment plot for Intercellular Communication gene set

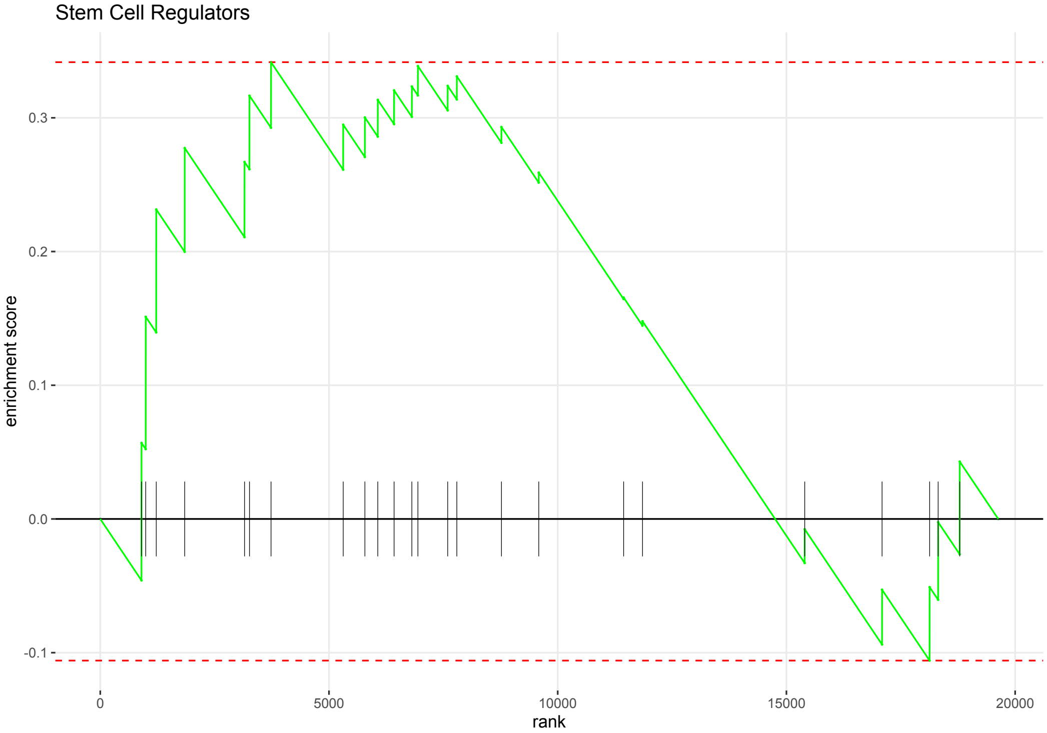


Figure 3: Gene Set Enrichment Plot for Stem Cell Regulators. Note the low number of genes included in this gene set, visualized as the vertical black lines along the x axis of the enrichment plot.

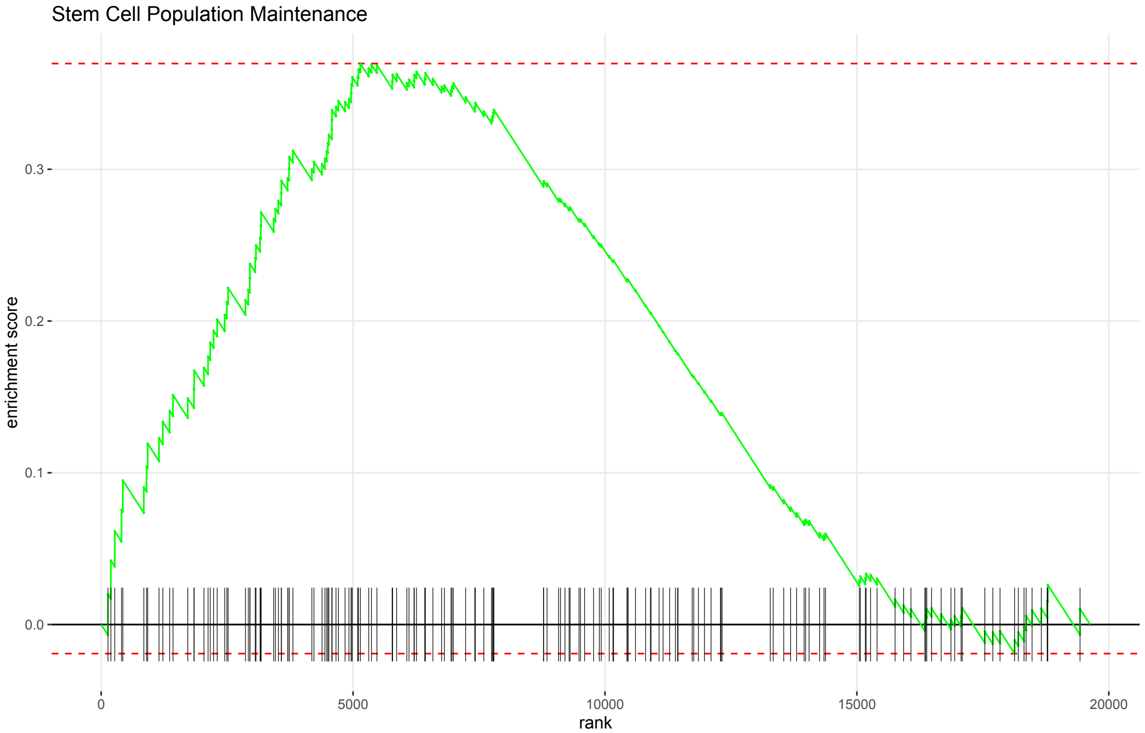


Figure 4: Gene Set Enrichment Plot for Stem Cell Population Maintenance GO Term. Note the improved sample size for genes to query with this GO term.

Sources:

Source for using the BiomaRt package:

Durinck, S., Huber, W., Davis, S., Pepin, F., Buffalo, V.S., Smith, M., 2020. biomaRt: Interface to BioMart databases (i.e. Ensembl). Bioconductor version: Release (3.10). https://doi.org/10.18129/B9.bioc.biomaRt

Original paper describing the GSEA method:

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Online forum with suggestion to remove additional digits from Ensemble ID name for query:
TCGA biomart conversion [WWW Document], n.d. URL https://support.bioconductor.org/p/110041/ (accessed 3.20.20).

TCGA data was downloaded from the Xena Data hub portal:
UCSC Xena [WWW Document], n.d. URL https://xenabrowser.net/datapages/ (accessed 3.20.20).

Tutorial on how to use the fsgea package for GSEA analysis and visualization methods:
Using fgsea package [WWW Document], n.d. URL https://bioconductor.org/packages/release/bioc/vignettes/fgsea/inst/doc/fgsea-tutorial.html (accessed 3.20.20).