Maddy Barry

Lauren Driscoll Lehner

Nvolve Biomarker Project II

Part two of the biomarker project involved a lot of discovery and working through challenges. This project was a continuation of the first part of the biomarker project, where we analyzed a breast cancer dataset and discovered an interesting overlap in the HER2 positive and HER2 negative groups that we wanted to explore further. A general outline was supplied, including a skeleton code that would produce relevant material to put into a Gene Set Enrichment Analysis tool. Our goal was to use the GSEA tool to learn more about what the overlap in groups consisted of, such as if any interesting genes were upregulated or downregulated. Although as a whole this could be perceived as an intimidating project, we were able to work through it by taking it piece by piece. We broke it down into manageable parts that we would figure out, and once we felt comfortable with that particular aspect of the project, we moved on to the next task. This approach proved to be an effective way to navigate through the project and successfully reach milestones.

Technically, one of us was familiar with coding, and the other was not experienced in code, so we could work together to interpret and utilize the code written to analyze our biomarker dataset further. We practiced critical thinking and worked through roadblocks in using GSEA, as we had never used it before. We spent a lot of time attempting to manually enter our dataset and adjust the parameters accordingly, and eventually were able to run our dataset successfully. This was probably the most time-consuming aspect, and in retrospect, we could have asked for further assistance. However, it was beneficial to be able to work through it on our own and tackle the challenge. We were able to get a better understanding of how GSEA functions, compared to if we were able to just plug in the dataset and not have to research the purpose behind utilizing each of the parameters.

The first breast cancer biomarker project was done using the gene expression data from human primary tumor samples obtained from the Gene Expression Omnibus (GEO). The database consists of curated datasets that have been submitted and validated by GEO. It opens the door for data mining and finding a large quantity of data in an area of interest. With a focus on the HER2 biomarker, we were able to determine a set of three probes that were able to detect the EBBR gene and resulted in either HER2 positive or HER2 negative classification. During our initial analysis, an interesting overlap in the two classification groups occurred leading us to question what was driving this potential third classification. Our goal for this project was to determine if there were a set or sets of genes causing samples originally placed in the HER2 negative category to display overexpression of HER2 in the same threshold as the HER2 positive group.

To do this, we took the original gene expression data from GEO and used an R notebook given to us by our coach. The R script analyzed the data to determine the interquartile ranges of the different probes. We filtered based on variance because it is more unlikely to be influenced by outliers and is the best measurement for skewed data or distribution. We refined our range by only selecting probes that showed an interquartile range greater than 2. This was done because a smaller range shows more variability which could indicate the central data is more

closely clustered together. It also indicates that the results are more reliable and consistent. We performed an unsupervised hierarchical clustering technique to determine if there was a particular probe that was driving the correlation. Based on the heat map produced, there was not a distinct probe driving the overlap in the original data analysis. Figure 1 shows the 208 probes with an interquartile range of less than 2 and the 156 tumor samples.

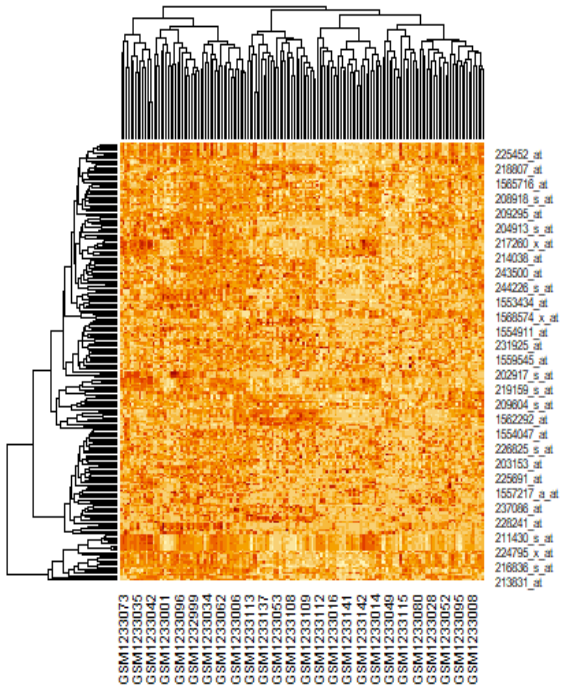

Figure 1: Heat Map obtained from R script depicting the correlation of probes and samples. The heat map does not show any strong bands of samples or probes grouped together. A strong block of either yellow or red would indicate that there was a higher output across all the samples for that particular gene and it could be of interest. However, this was not shown here.

To further analyze the data we utilized the Gene Set Enrichment Analysis software (GSEA). GSEA is a computational method that allows for multiple gene set comparisons against a particular dataset. This analysis could give insight into overlaps of gene datasets and potentially reveal patterns not originally seen. These connections could lead to novel discoveries and enhance insight into biological similarities between genes. To do this, GSEA compiles a ranked list of gene sets and calculates enrichment scores (ES) that determine the degree at which a gene set is overrepresented in a ranked list of genes. A positive ES indicates ranked genes at the top of the compiled list, while a negative ES are gene sets ranked at the bottom. This data is then visualized through graphs and distinct peaks, at either the beginning or end of the graph, which will potentially highlight gene sets of interest. GSEA also determines a normalized enrichment score (NES) and is used as the primary statistic for examining gene set enrichment results. We found we got the best results by running our dataset against gene sets in the C4 category, which consists of 858 computational gene sets formed by mining vast

amounts of cancer-oriented microarray data. Of these 858 gene sets, there were similarities with 3.

After running the GSEA on the two phenotype (HER2 positive, HER2 negative) groups, the results indicated that there were 10 core enrichment genes for module 117, 12 for module 6 and 9 for module 84. A few of the top genes of interest were CEACAM6 (cell adhesion molecule 6), S100A8 (S100 calcium binding protein A8) and apolipoprotein D. CEACAM6 and S100A8 are known to be overexpressed in breast cancer, whereas upregulation of ApoD has been suggested to possibly be related to breast cancer progression in recent research.

Fig 1: Enrichment plot: MODULE_117
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [plain text format]

	SYMBOL	TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	CEACAM6	CEA cell adhesion molecule 6 [Source:HGNC Symbol;Acc:HGNC:1818]	8	0.454	0.0886	Yes
2	CLCA2	chloride channel accessory 2 [Source:HGNC Symbol;Acc:HGNC:2016]	11	0.427	0.2129	Yes
3	SCGB2A2	secretoglobin family 2A member 2 [Source:HGNC Symbol;Acc:HGNC:7050]	14	0.392	0.3257	Yes
4	APOD	apolipoprotein D [Source:HGNC Symbol;Acc:HGNC:612]	24	0.281	0.3508	Yes
5	AGR2	anterior gradient 2, protein disulphide isomerase family member [Source:HGNC Symbol;Acc:HGNC:328]	34	0.204	0.3504	Yes
6	JCHAIN	joining chain of multimeric IgA and IgM [Source:HGNC Symbol;Acc:HGNC:5713]	35	0.202	0.4163	Yes
7	MFAP5	microfibril associated protein 5 [Source:HGNC Symbol;Acc:HGNC:29673]	38	0.188	0.4628	Yes
8	PIP	prolactin induced protein [Source:HGNC Symbol;Acc:HGNC:8993]	46	0.173	0.4672	Yes
9	LTF	lactotransferrin [Source:HGNC Symbol;Acc:HGNC:6720]	56	0.113	0.4372	Yes
10	CXCL9	C-X-C motif chemokine ligand 9 [Source:HGNC Symbol;Acc:HGNC:7098]	57	0.111	0.4734	Yes
11	CCL8	C-C motif chemokine ligand 8 [Source:HGNC Symbol;Acc:HGNC:10635]	87	0.010	0.2617	No
12	CILP	cartilage intermediate layer protein [Source:HGNC Symbol;Acc:HGNC:1980]	97	-0.021	0.2020	No

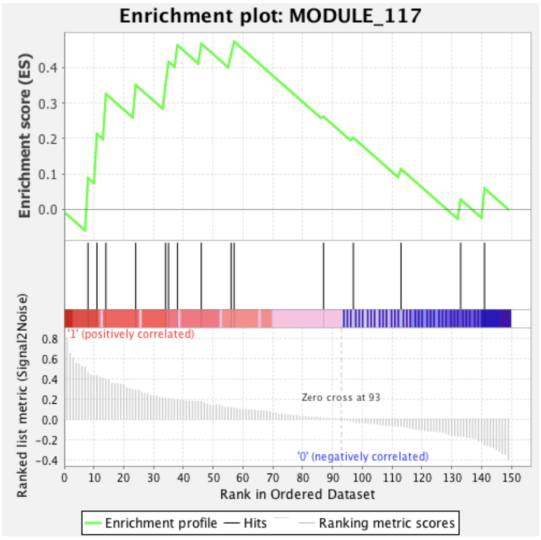

Fig 1: Enrichment plot: MODULE_117

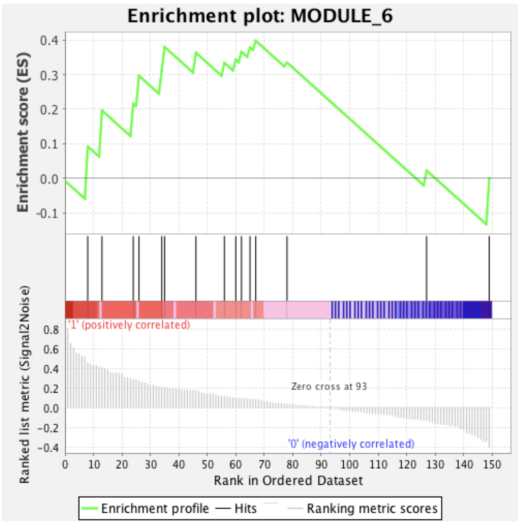

Fig 1: Enrichment plot: MODULE_6

Fig 1: Enrichment plot: MODULE_6
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [plain text format]

	SYMBOL	TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	CEACAM6	CEA cell adhesion molecule 6 [Source:HGNC Symbol;Acc:HGNC:1818]	8	0.454	0.0920	Yes
2	S100A8	S100 calcium binding protein A8 [Source:HGNC Symbol;Acc:HGNC:10498]	13	0.401	0.1961	Yes
3	APOD	apolipoprotein D [Source:HGNC Symbol;Acc:HGNC:612]	24	0.281	0.2158	Yes
4	TFAP2B	transcription factor AP-2 beta [Source:HGNC Symbol;Acc:HGNC:11743]	26	0.264	0.2963	Yes
5	AGR2	anterior gradient 2, protein disulphide isomerase family member [Source:HGNC Symbol;Acc:HGNC:328]	34	0.204	0.3123	Yes
6	JCHAIN	joining chain of multimeric IgA and IgM [Source:HGNC Symbol;Acc:HGNC:5713]	35	0.202	0.3797	Yes
7	PIP	prolactin induced protein [Source:HGNC Symbol;Acc:HGNC:8993]	46	0.173	0.3631	Yes
8	LTF	lactotransferrin [Source:HGNC Symbol;Acc:HGNC:6720]	56	0.113	0.3340	Yes
9	AQP3	aquaporin 3 (Gill blood group) [Source:HGNC Symbol;Acc:HGNC:636]	60	0.094	0.3430	Yes
10	GPRC5A	G protein-coupled receptor class C group 5 member A [Source:HGNC Symbol;Acc:HGNC:9836]	62	0.091	0.3660	Yes
11	PROM1	prominin 1 [Source:HGNC Symbol;Acc:HGNC:9454]	65	0.082	0.3784	Yes
12	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1 [Source:HGNC Symbol;Acc:HGNC:4944]	67	0.080	0.3977	Yes
13	TNFRSF10B	TNF receptor superfamily member 10b [Source:HGNC Symbol;Acc:HGNC:11905]	78	0.034	0.3348	No

Fig 1: Enrichment plot: MODULE_84
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: GSEA details [plain text format]

	SYMBOL	TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
1	S100A8	S100 calcium binding protein A8 [Source:HGNC Symbol;Acc:HGNC:10498]	13	0.401	0.0799	Yes
2	APOD	apolipoprotein D [Source:HGNC Symbol;Acc:HGNC:612]	24	0.281	0.1294	Yes
3	LYZ	lysozyme [Source:HGNC Symbol;Acc:HGNC:6740]	25	0.281	0.2529	Yes
4	PDE4B	phosphodiesterase 4B [Source:HGNC Symbol;Acc:HGNC:8781]	32	0.214	0.3023	Yes
5	JCHAIN	joining chain of multimeric IgA and IgM [Source:HGNC Symbol;Acc:HGNC:5713]	35	0.202	0.3763	Yes
6	IGKC	immunoglobulin kappa constant [Source:HGNC Symbol;Acc:HGNC:5716]	47	0.160	0.3651	Yes
7	SPP1	secreted phosphoprotein 1 [Source:HGNC Symbol;Acc:HGNC:11255]	54	0.116	0.3714	Yes
8	LTF	lactotransferrin [Source:HGNC Symbol;Acc:HGNC:6720]	56	0.113	0.4134	Yes
9	CXCL9	C-X-C motif chemokine ligand 9 [Source:HGNC Symbol;Acc:HGNC:7098]	57	0.111	0.4621	Yes
10	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1 [Source:HGNC Symbol;Acc:HGNC:4944]	67	0.080	0.4307	No

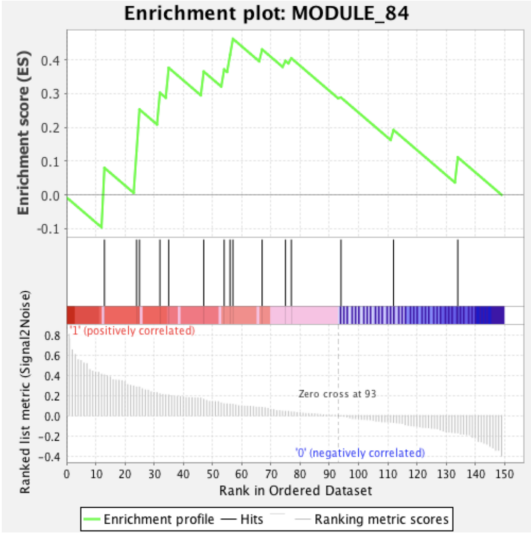

Fig 1: Enrichment plot: MODULE_84

There were not strong signals on the graphs which is consistent with the lack of strong clusters on the heat map.

A few limitations and considerations must be considered when analyzing the results obtained from the Gene Set Enrichment Analysis. For instance, the GSEA software was difficult to use and ensure the files needed were properly formatted. Potential labeling errors could lead to insignificant results. Furthermore, the original data analysis done in the first project could have been done incorrectly, leading to the overlapping interquartile ranges in HER2+ and HER- samples.

For this particular project, we would suggest implementing certain milestones and potentially provide an example dataset for people to practice utilizing GSEA with (that is known to run well and give ‘ideal’ results). For the milestones, the first would be having a dataset to work with. Another milestone could be researching GSEA and how it works, like we did in this experiment. The milestone that we did not have but may be beneficial in the future is having the group practice using GSEA with a sample dataset to get used to how GSEA works. This would give them an idea of what results can or should look like. Once they have practiced using that,

they can run their own dataset and have a better understanding of what they did right or wrong and hopefully have a better understanding of why their dataset did or did not have interesting results. To continue this project, further analysis can be done on gene sets not previously connected to cancer. We could also do a deeper dive into the different categories of gene sets provided by GSEA, such as the C3 category that provides different regulatory target gene sets. However, a quick analysis showed several gene sets with indicators of false positive results. Overall, the analysis done using GSEA showed no strong indication of novel genes driving the overlap in expression values in our previous breast cancer biomarker project.