**BIOM 200C Final Project: Re-analysis of GSE129647**

Sahan, Ayse; Shing, Brian

**Background and Significance**

In breast cancer, metastasis of primary tumor cells to other organs is one of the leading causes of death (1). Despite aggressive chemotherapy, 40% of breast cancer patients relapse and succumb to metastasis (2). The interplay between the epigenetic landscape and gene expression and its effect on the metastatic potential of primary tumor cells has not been well characterized (3). As such, this experiment combines analyses of RNA, ChIP-, and ATAC-seq of MDA-MB-231 breast cancer cells to elucidate the connection between epigenetic marks and metastatic gene expression profiles to identify potential metastasis biomarkers.

**Analysis Steps and Methods**

RNA-seq and acetylation (H3K27ac) ChIP-seq fastq files for parental and brain tissue tumor cells from GSE 129647 were re-analyzed. Parental and BrM2 brain metastatic MDA-MB-231 breast cancer cells were collected for analysis. RNA-seq samples were sequenced on an Illumina HiSeq4000 while ChIP-seq samples were sequenced on an Illumina HiSeq2000. The RNA-seq reads were aligned to the hg38 genome index and gene counts were quantified using STAR aligner and visualized on the Integrative Genome Viewer to confirm read alignment. Differentially expressed genes (padj<0.01) were considered for analysis. Upregulated (> 2) and downregulated (< -1.5) genes were identified by comparing the log2 fold change and run through metascape and Enrichr pathway enrichment. MultiQC was run to generate an aggregate quality control report. For the ChIP-seq data, BWA was used to align reads to the genome. Reads with MAPQ > 10 were kept for analysis. Deeptools was used to create bedGraph and bigWig files to visualize the ChIP-seq data. MACS2 was used to find peaks, and the GREAT web portal was used to identify genes near identified peaks and elucidate the gene ontology and pathway enrichment. Peak location relative to transcription start site (TSS) and enriched pathways were identified using the Cistrome-GO web portal. featureCounts and DESeq2 were used to identify significantly differentially enriched peaks. Data integration was done through comparing ChIP-seq peaks and RNA-seq transcripts in IGV and pathway enrichment of each dataset.

**Notable Results**

Through RNA-seq analysis, we identified transcripts involved in neuron projection, regulation of cell adhesion, and chemical synaptic transmission as upregulated in the brain metastatic cells as compared to parental cells. Down-regulated pathways included extracellular matrix organization, EGFR signaling, and microglial cell proliferation. The top upregulated genes included matrix metalloproteinase 1 (MMP1), melanoma-associated antigen 6 (MAGEA6), ADAM metallopeptidase with thrombospondin type 1 motif 1 (ADAMTS1). All of these genes have been implicated in several types of cancer progression (6,7,8). The top down-regulated genes were retinoic acid receptor responder 2 (RARRES2) and adhesion G protein-coupled receptor F1 (ADGRF1). Regions with significantly down-regulated acetylation corresponded to MUC1, SEMA3B, and, in alignment with the RNA-seq analysis, ADGRF1. Regions correlated to Kirre Like Nephrin Family Adhesion Molecule 3 (KIRREL3) and semaphorin 3B (SEMA3B) had increased acetylation in the brain metastatic cells as compared to parental cells. In both RNA-seq and ChIP-seq gene ontology pathway enrichment analyses, we identified commonly upregulated pathways in the brain metastatic cell samples, including cell junction, cell motility, angiogenesis, ECM binding, growth factor response, and cell-adhesion molecule binding pathways.

**Discussion**

We found agreement in our findings in terms of up/down-regulated genes and pathway enrichment analyses. Interestingly, brain metastasized cells were enriched in neuronal projection and chemical synaptic transmission. This may indicate that metastasized cells either “infect” surrounding cells to spread or exhibit stem-like characteristics and express proteins suiting their microenvironment, which would explain the metastasized cells being enriched in brain tissue gene expression pathways. A future direction would include checking if a similar behavior is observed in cells metastasized to other organs, such as the lung.

Many of the top differentially expressed genes identified in the RNA-seq analysis are linked to cancer metastasis or invasiveness. MMP1 is overexpressed in triple-negative breast cancer, which is one of the most invasive forms of the disease (6). As one of the top upregulated genes in the brain metastatic cells, MMP1 may be involved in metastasis and motility of breast cancer cells. The MAGE family of proteins has been implicated in tumor metastasis and overall worse prognosis (7). ADAMTS1 is another metalloproteinase that has been implicated in tumor metastasis and progression (8,9). Several of the other significantly up- or down-regulated genes may be indicative of pathways involved in promoting metastasis. Through data integration, we observed ADGRF1 was hypo-acetylated and down-regulated in transcription. According to the Human Protein Atlas, ADGRF1 is a favorable prognostic marker for renal cancer, and the majority of tested cancers do not express it. Therefore, an interesting future direction may be to study ADGRF1 as a candidate tumor suppressor. A compelling future direction would be to collect parental and metastasized cells at different stages of cancer to extract transcriptomic and epigenetic data at various time-points to identify if there is epigenetic priming occurring at an early stage in the parental cells that may be correlated to: potential for cell metastasis, metastasis organ specificity, and clinical prognosis. This could yield new A close up of a logo

Description automatically generateddiagnostic and prognostic practices and lead to the identification of novel therapy targets.

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated.

A screenshot of a cell phone

Description automatically generatedFigure 1. Differential gene expression of brain metastatic vs. parental breast cancer cells in a heatmap (right), volcano plot (center), and gene ontology upregulated (left top) and downregulated (left bottom) pathways.

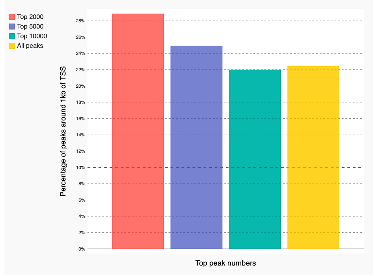
A screenshot of a cell phone

Description automatically generatedA close up of text on a white background

Description automatically generatedA picture containing map, text

Description automatically generated

A close up of a map

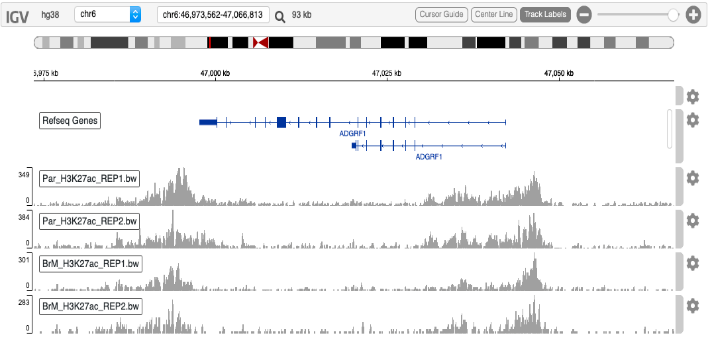
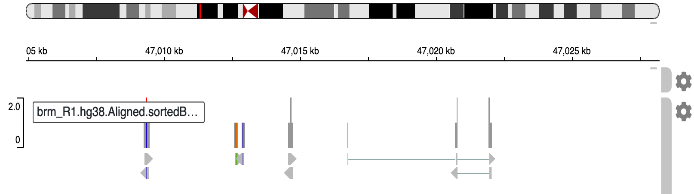
Description automatically generatedFigure 2. More than 25% of the top peaks were located around 1kb of a TSS in parental (top right) and brain metastatic (bottom right) cells (from Cistrome-GO). Overlap of peak signals to TSS for replicates (left).

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3. Motif enrichment of peaks from parental (left) and brain metastatic (right) cells from ChIP-seq.



A close up of text on a white background

Description automatically generated

Figure 4. IGV tracks of ChIP-seq peaks (top right) and RNA-seq transcripts for brain metastatic (bottom right) and parental breast cancer (left) cells.

**Supplemental Analysis**

**Data source and alignment**

All analyses were conducted on UC San Diego’s Triton Shared Computing Center. RNA- and acetylation ChIP-seq fastq files for parental and brain tissue tumor cells were downloaded from Gene Omnibus (GSE129647) using fastq-dump. STAR and BWA were used to generate the hg38 genome index and align the RNA- and ChIP-seq reads, respectively, to the genome.

**RNA-seq analysis**

Gene counts were quantified using featureCounts and the gencode v33 release GTF file. BAM and BAI files were visualized on the Integrative Genome Viewer to confirm sequence read alignment using genes, such as GAPDH, for reference. R package DESeq2 was utilized to identify statistically significant differentially expressed genes (padj<0.01), and heatmap.2 was used to generate heatmaps. As a new analysis tool, R package EnhancedVolcano was used to generate a volcano plot and identify the top significant gene hits. Upregulated and downregulated genes were identified by comparing the log2 fold change. Genes that were upregulated >2 fold or down-regulated <-1.5 fold were run through metascape and EnrichR to determine pathway enrichment. MultiQC was run to generate an aggregate quality control report.

**ChIP-seq analysis**

SAM files following BWA alignment were processed by Samtools. Reads that did not uniquely map were discarded, and the SAM files were converted to sorted BAM and BAI files. Deeptools was utilized to create bedGraph and bigwig files for visualization on the Integrative Genome Viewer and UCSC genome browser.

MACS2 was utilized to find ChIP-seq peaks with the respective input experiments serving as controls. Peaks from different samples were merged using bedtools. featureCounts was used to quantify reads and DESeq2 was used to identify statistically significant (p-adj < 0.05) differentially enriched peaks. Peak positions were linked to differential enrichment information. Peaks with log2 fold change >1 were identified. The resulting peaks were annotated on the GREAT web server to assign peaks to nearby genes and perform pathway enrichment analysis. As another new tool, the peaks were inputted into the Cistrome-GO web portal, which graphed peak location relative to TSS and provided information on associated enriched pathways. Enriched de novo DNA motifs were found using findMotifsGenome.pl tool from Homer.

**ChIP-seq and RNA-seq data integration**

Enriched pathways in both datasets were identified and compared. The annotatePeaks.pl tool was used to assign genes to each peak based on location, and this information was joined to differentially expressed genes identified by RNA-seq. The top 5 differentially expressed genes identified from DESeq2 analysis of RNA-seq data were visualized in the Integrative Genome Viewer (IGV) and compared to bigwig files from the ChIP-seq data to identify acetylation patterns corresponding to gene counts.

**References**

1. Klastersky J, Daneau D, & Verhest A. 1972. Causes of death in patients with cancer. *European Journal of Cancer* (1965) 8:149–154.

2. Weigelt B, Peterse JL, & van’t Veer LJ. 2005. Breast cancer metastasis: markers and models. *Nat Rev Cancer* 5:591–602.

3. Cai, W.L., Greer, C.B., Chen, J.F. *et al.* 2020. Specific chromatin landscapes and transcription factors couple breast cancer subtype with metastatic relapse to lung or brain. *BMC Med Genomics* 13, 33.

4. Soto MS, et al. Functional role of endothelial adhesion molecules in the early stages of brain metastasis. *Neuro Oncol* 2014. 16(4):540-551.

5. Siesser PF, Manness PF. L1 cell adhesion molecules as regulators of tumor cell invasiveness. *Cell Adh Migr.* 2009. 3(3):275-277.

6. Wang QM, et al. MMP1 is overexpressed in triple-negative breast cancer tissues and the knockdown of MMP-1 expression inhibits tumor cell malignant behaviors *in vitro*. *Oncology Letters* 2019. 17(2):1732-1740.

7. Weon JL, Potts PR. The MAGE protein family and cancer. *Current Opinion in Cell Biology* 2015. 37:1-8.

8. Fernandez-Rodriguez R, et al. Stroma-derived but not tumor ADAMTS1 is a main driver of tumor growth and metastasis. *Oncotarget* 2016. 7(23):34507-34519.

9. Tan Ide A, et al. The metalloproteinase ADAMTS1: a comprehensive review of its role in tumorigenic and metastatic pathways. *International Journal of Cancer* 2013. 133(10):2263-2276.