Genome Browsers

I was exposed to the use of many different genome browsers throughout many courses in the JHU Individualized Genomics and Health program. Each has had its benefits and disadvantages, but each has been incredibly helpful in visualizing key information from datasets that would otherwise be missed. For my research paper as well as class work and projects, programs such as Integrative Genomics Viewer, the NCBI Genome Browser, and the viewer that Ensembl presents with every record have been of great use.

Integrative Genomics Viewer(IGV) is able to handle larger scale genomic datasets, and is able to be locally downloaded, with server access to multiple different reference genomes. IGV was used in the Bioinformatics: Tools for Genome Analysis course for the research paper and coursework. The following examples use BED files that were obtained via UCSC Table Browser or self-made in order to analyze dataset information in comparison to the appropriate reference genome.

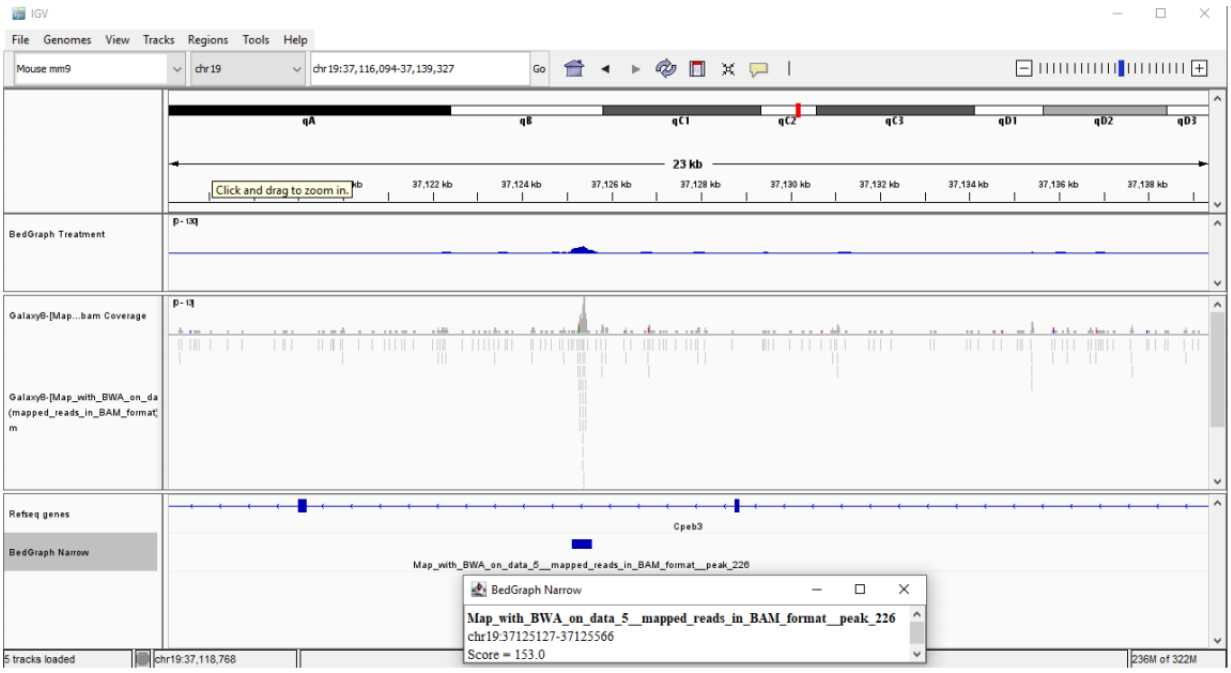
1. IGV was used to compare the differences in copy number variants (CNV) between cell lines with ovarian cancer (OV-TP) and breast cancer(BRCA-TP) for the 4 genes PKN2, GRXCR1, PRKN, and PPIAL4A. Difference would be indicated in the differences between color in those gene regions. Red indicated a loss of CNV, while blue indicated a gain of CNV. The results were obtained by examining the Cancer Genome Atlas Results on the reference genome of hg19. Below is a table demonstrating results found and an analysis of those findings.

Gene	Ovarian Cancer	Breast Cancer	Comparison of Results
PKN2	Majority Red	Slightly more Blue	Ovarian cancer shows a loss while breast cancer shows a gain, which could indicate that this gene is involved in the pathway for one of these diseases, but not the other.
GRXCR1	Slightly more Blue	Majority Blue	Since both show more blue here, this gene could be implicated in both cancers, with BC showing more copy number gain and therefore more effect from the gene in the development of that cancer type, BUT could also mean that this gene plays little role in the development of these diseases.
PRKN	Majority Blue	Majority Blue	As both are majority blue, it is similar to GRXCR1 where the gain of CN in PRKN plays a role in the development of both OC and BC, or since it shows similarly in both, doesn't play much of a role in the difference between the two diseases.
PPIAL4A	Completely Red	Completely Red	Both cancers show a loss of CNV in this region of the genome, indicating that PPIAL4A is implicated in both diseases, where the loss of this gene's expression could lead to cancerous conditions.

2. IGV was used to examine ENCODE tracks to examine areas of potential methylation related to specific genes in the HUVEC cell line. Results indicated that the TCF7 gene had elements from the DNase track(BED file DNase_HW_pt2.bed) with two located upstream of TSS, one within TSS, three within exon 1, one within an intron in the middle of the gene, and the last being in exon 9 indicating regions that were more sensitive to DNase activity. It also showed a region of H3K4me3 near the transcriptional start site, which might indicate that this gene was not

transcriptionally active in HUVEC cells.

3. IGV was used to analyze a ChIP-Seq dataset of reads from a mouse genome (mm9). Using Galaxy(see galaxy section file for more information about this workflow), the quality scores were analyzed with FASTQC, lower quality reads were removed using Trimmomatic, the reads were mapped to the mm9 genome using BWA, and finally MACS2 callpeaks was ran on the file to analyze which genes had ChIP peaks nearby.



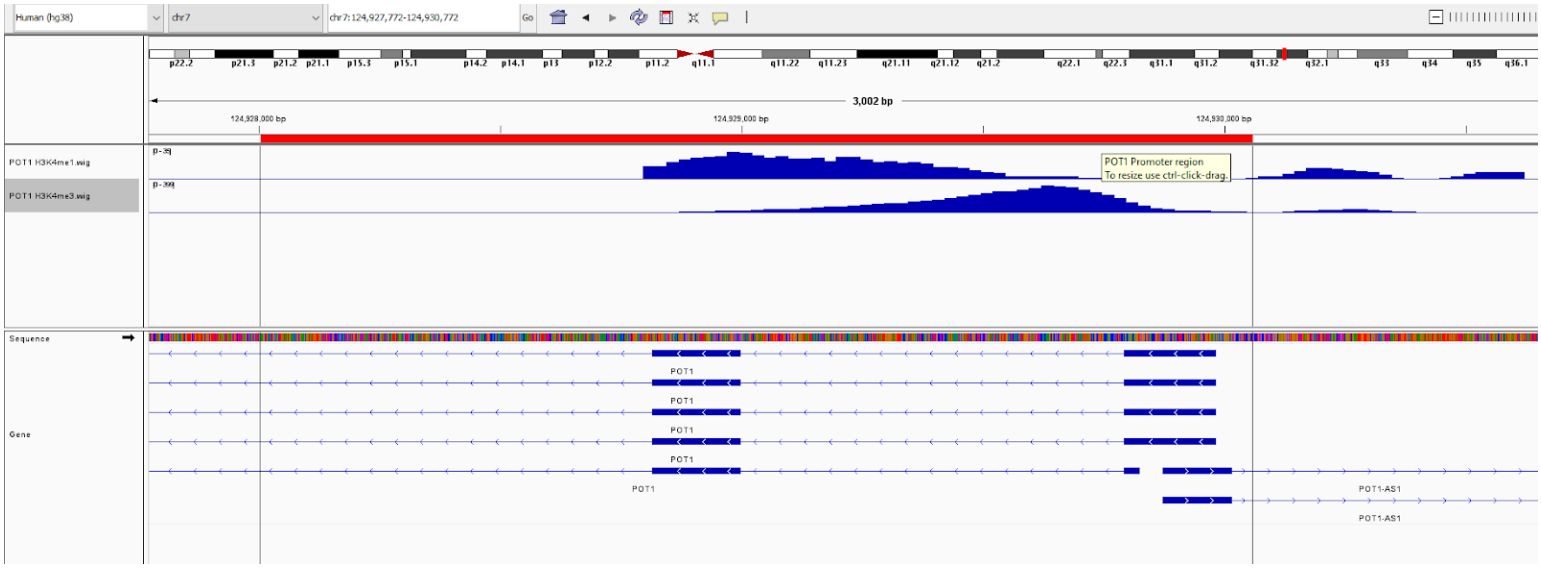
By using IGV, it was plain to see that the Cpeb3 gene was strongly associated with the ChIP-seq data set, with peaks from the MACS2 run being present within the Refseq gene loci.

4. Finally, IGV was also used to analyze methylation of the human POT1 gene, a gene that is the main focus of the research paper my peers and I are working on(specifically the SNP rs75932146's effect in relation to cancer). This was accomplished to identify the methylation state of this region to determine gene regulation patterns. The following excerpt is my interpretation of these findings that were submitted to our professor:

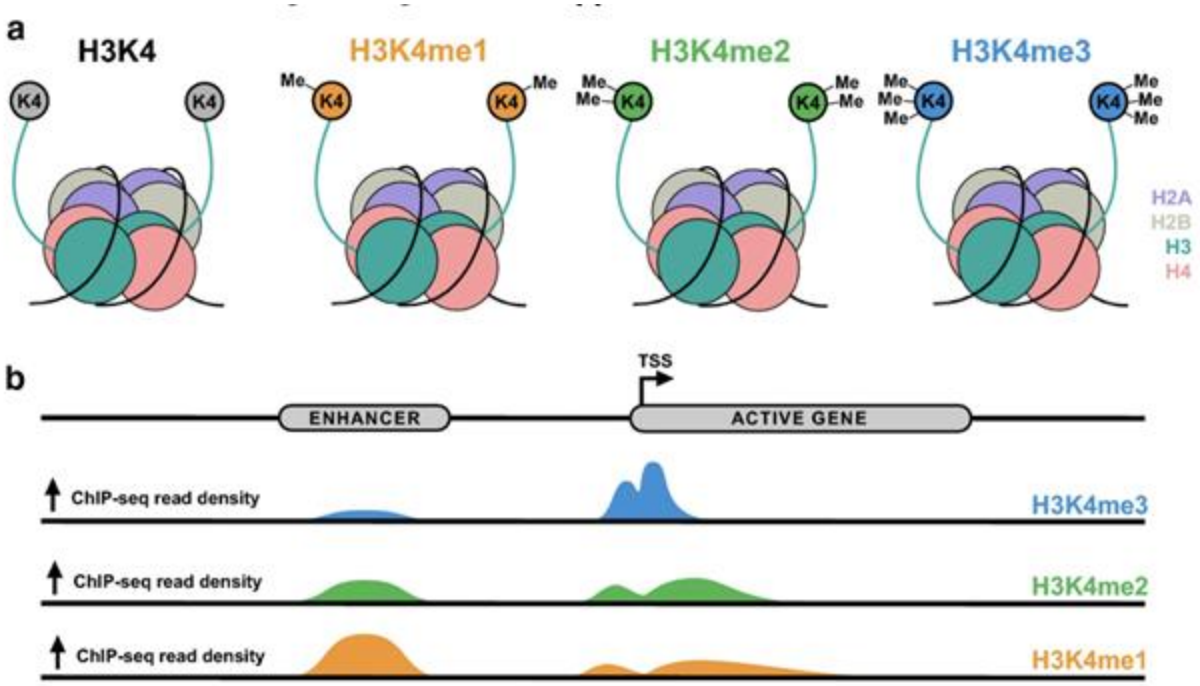
    Next, regulation of POT1's promoter region was examined via presence of H3K4me3 and H3K4me1 loci. These epigenetic markers hold a significant influence over both the expression and regulation of promoter regions for genes (Sharifi-Zarchi, A et al, 2017). H3K4me1 is a marker that indicates a region as a potential enhancer, and H3K4me3 indicates that region as a potential promoter (Sharifi-Zarchi, A et al, 2017). Using UCSC's table browser, the Regulation group and Layered H3K4me3 track was selected for HUVEC cells in the chr7: 124,928,800-124,930,601position. 69 regions were found.

    The same was done for Layered H3K4me1 in the same position. 65 regions were found.

    These regions were placed into IGV along with the promoter position in order to analyze the relationship between these methylation states and the gene's promoter.



The methylation markers at POT1's promoter show a distinct pattern, where only one of either methylation state peaks at a time.



From: Collins BE, Greer CB, Coleman BC, Sweatt JD. (2019). Histone H3 lysine K4 methylation and its role in learning and memory. Epigenetics Chromatin. doi: 10.1186/s13072-018-0251-8. PMID: 30616667; PMCID: PMC6322263

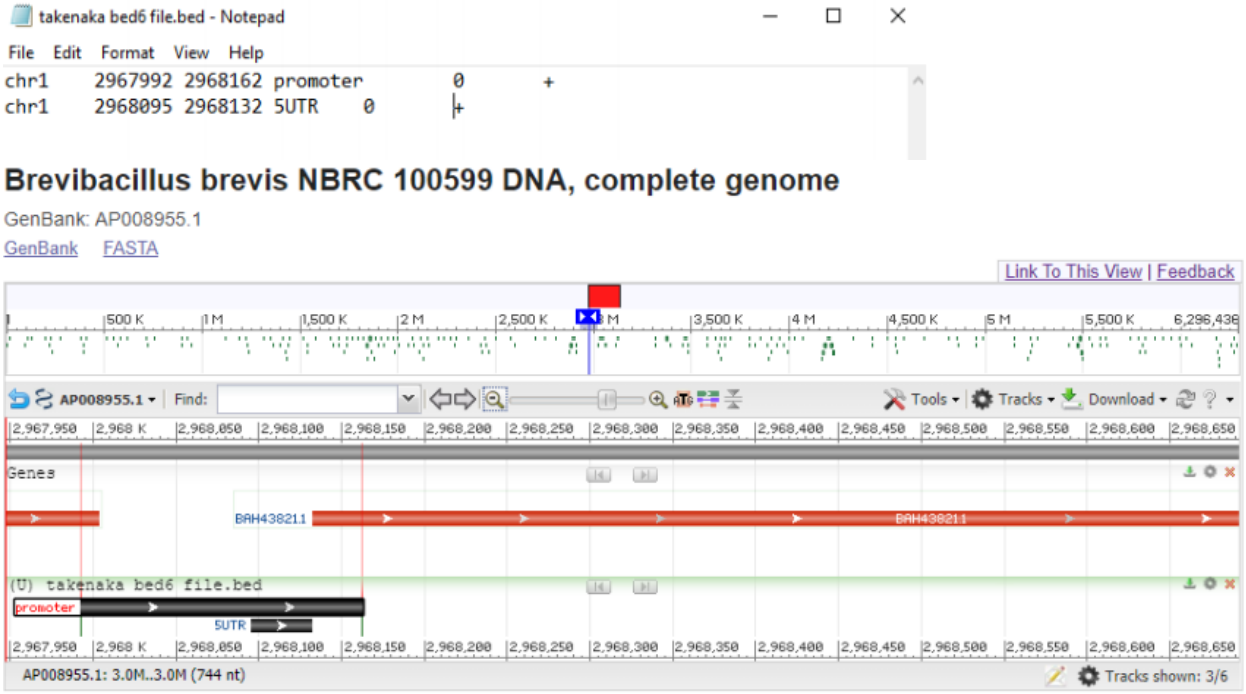
Comparing the IGV figure to this, the pattern seen in the highlighted region of the IGV results can be attributed to enhancers and promoters associated with the POT1 gene. The H3K4me1 track (top track from IGV figure in #4) shows resemblance to a potential enhancer region, while the H3K4me3 track (2nd track from IGV figure in #4) resembles the promoter.

Sharifi-Zarchi, A., Gerovska, D., Adachi, K., Totonchi, M., Pezeshk, H., Taft, R. J., Schöler, H. R., Chitsaz, H., Sadeghi, M., Baharvand, H., & Araúzo-Bravo, M. J. (2017). DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC genomics, 18(1), 964. https://doi.org/10.1186/s12864-017-4353-7

NCBI's Genome Viewer also allows for a user to browse reference genomes and input their own BED files for visualization on their online platform. The genome data viewer allows for the viewing/analysis o mostly NCBI-annotated assemblies.

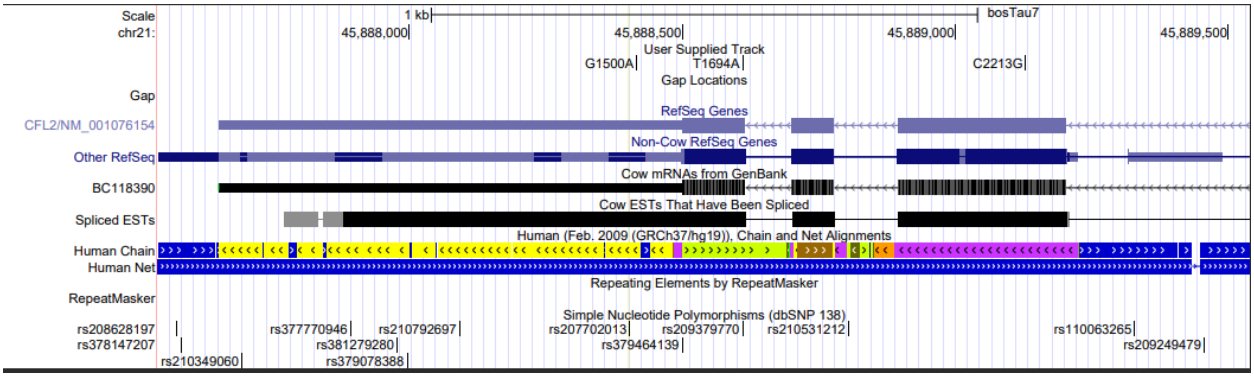
1. A BED6 file was created based on the findings from Takenaka et al 2015's research into *Brevibacillus brevis.* Based on the information provided, the BED file included both the promoter region and the 5' UTR as noted by the paper. NCBI's genome viewer was used to visualize the locations of these within the genome track of the bacteria.

    a. Takenaka, T., Ito, T., Miyahara, I., Hemmi, H., & Yoshimura, T. (2015). A new member of MocR/GabR-type PLP-binding regulator of D-alanyl-D-alanine ligase in Brevibacillus brevis. *The FEBS journal*, *282*(21), 4201–4217. https://doi.org/10.1111/febs.13415



UCSC Genome Browser is an online/downloadable genomic viewing tool maintained by the Genome Bioinformatics Group that allows for the importation of a user's own data as well as uses tracks from not only a reference genome but various other tracks that hold information about mapping and sequencing, genes and gene predictions, expression/regulation, and variation to name a few. With these additional tracks of information, examining a user's own dataset is complemented by this and compared to the other viewers, provides much more background information for the user if needed.

1. I created a BED file for the bosTau7 genome assembly of cows, in order to determine the location of three different SNPs in the CFL2 gene. The three positions were: C2213G, T1694A, and G1500A.



Once loaded into the track, UCSC allowed for a detailed examination of not only what component of the gene each SNP was found in, but also if they related to any known SNPs (from dbSNP 138). Using this information, it was shown that T1694A and C2213G intersected with with locations in coding exons, and could therefore impact the protein of this gene.