SG_LB_BMEG591E_FinalProject

2023-04-05

Introduction:

Our analysis is based on the paper published by Ha et al. (2023) entitled: "Reduced expression of alanyl aminopeptidase is a robust biomarker of non-familial adenomatous polyposis and non-hereditary nonpolyposis colorectal cancer syndrome early-onset colorectal cancer" (https://doi.org/10.1002/cam4.5675). The authors have made their sequencing data (fastqs) and read counts file publicly available on the GEO database: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213092. We chose this analysis because it involved many concepts that we covered in class, but also went beyond these and required learning some new techniques. Cancer biomarker discovery is a research interest of Liam's lab, so this paper is relevant to potential projects he may carry out down the road.

This study aimed to identify differential gene expression markers to distinguish early-onset colorectal cancer from late-onset colorectal cancer of the subtypes non-familial adenomatous polyposis and non-hereditary non-polyposis.The group performed RNA-seq of 49 early-onset and 50 late-onset colorectal cancer samples. They analyzed differentially-expressed genes between these groups and filtered them based on log2FC, logCPM, and P-values. They validated the identified differentially-expressed genes using data from TCGA (The Cancer Genome Atlas) database, based on similar expression profiles. Based on this validation, they identified the gene analyl aminopeptidase (ANPEP) as significantly downregulated in early-onset colorectal cancer patients in both their cohort and the TCGA cohort. This association was further supported by methylation data and information from the GTEx and GSE196006 datasets. The study concluded that the gene ANPEP was significantly down-regulated in early-onset colorectal cancer, and could serve as a novel biomarker.

The scope of our analysis is to carry out the methods described in section 2.2: Bulk RNA sequencing and data analysis. In these steps, the authors used the fastq files output from RNAseq to obtain normalized read counts, and used these to produce the heatmap shown in Figure 1A and the volcano plot shown in Figure 2B. These steps were used to identify genes that had significant differential expression between the early-onset and late-onset groups in the 99 patient samples. These steps precede the comparison with TCGA data that ultimately led the authors to single out ANPEP as a novel biomarker.

Methods and Results:

Part I: Pre-processing fastqs to obtain read counts

To demonstrate the pre-processing steps for generating a read counts file from fastq files as performed by the authors, we use only 4 fastq files, two from the early-onset group and two from the late-onset group. As there are 99 samples with 2 fastq files each in the whole dataset, the analysis would be too time-consuming if we used all the available files.

The files are named as follows:

LateOnset_SRR21523090_pass_1.fastq.gz LateOnset_SRR21523090_pass_2.fastq.gz

LateOnset_SRR21523091_pass_1.fastq.gz LateOnset_SRR21523091_pass_2.fastq.gz

EarlyOnset_SRR21523117_pass_1.fastq.gz EarlyOnset_SRR21523117_pass_2.fastq.gz

EarlyOnset_SRR21523118_pass_1.fastq.gz EarlyOnset_SRR21523118_pass_2.fastq.gz

In the following script, we tried to carry out all pre-processing steps only for the sample EarlyOnset_SRR21523117. We then incorporated all of these steps into a pipeline that could be used to process any number of samples.

1. Check quality of raw fastqs using fastqc

As shown below, the fastq file failed the per base sequence content and adapter content quality metrics. This showed that we needed to trim adapters and indexes.

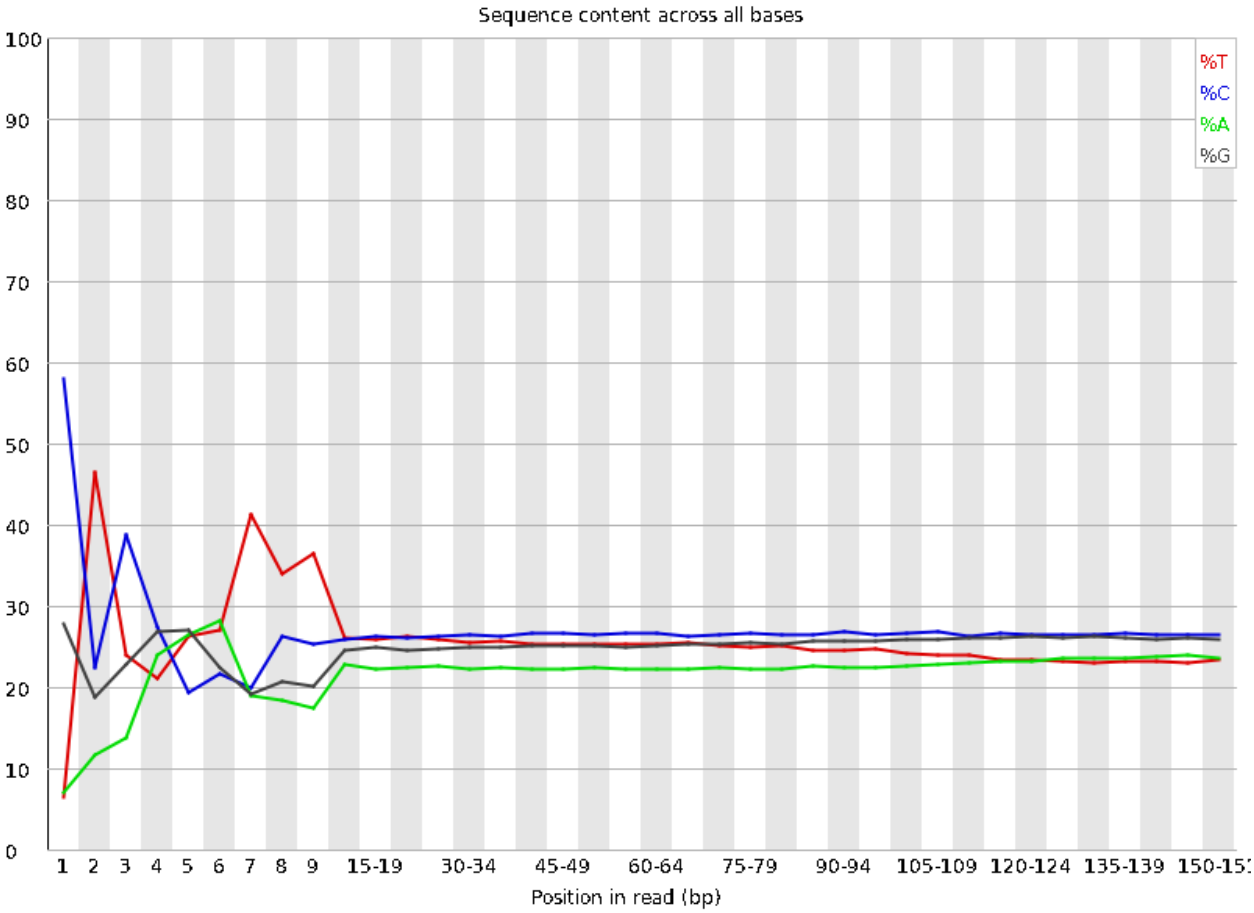
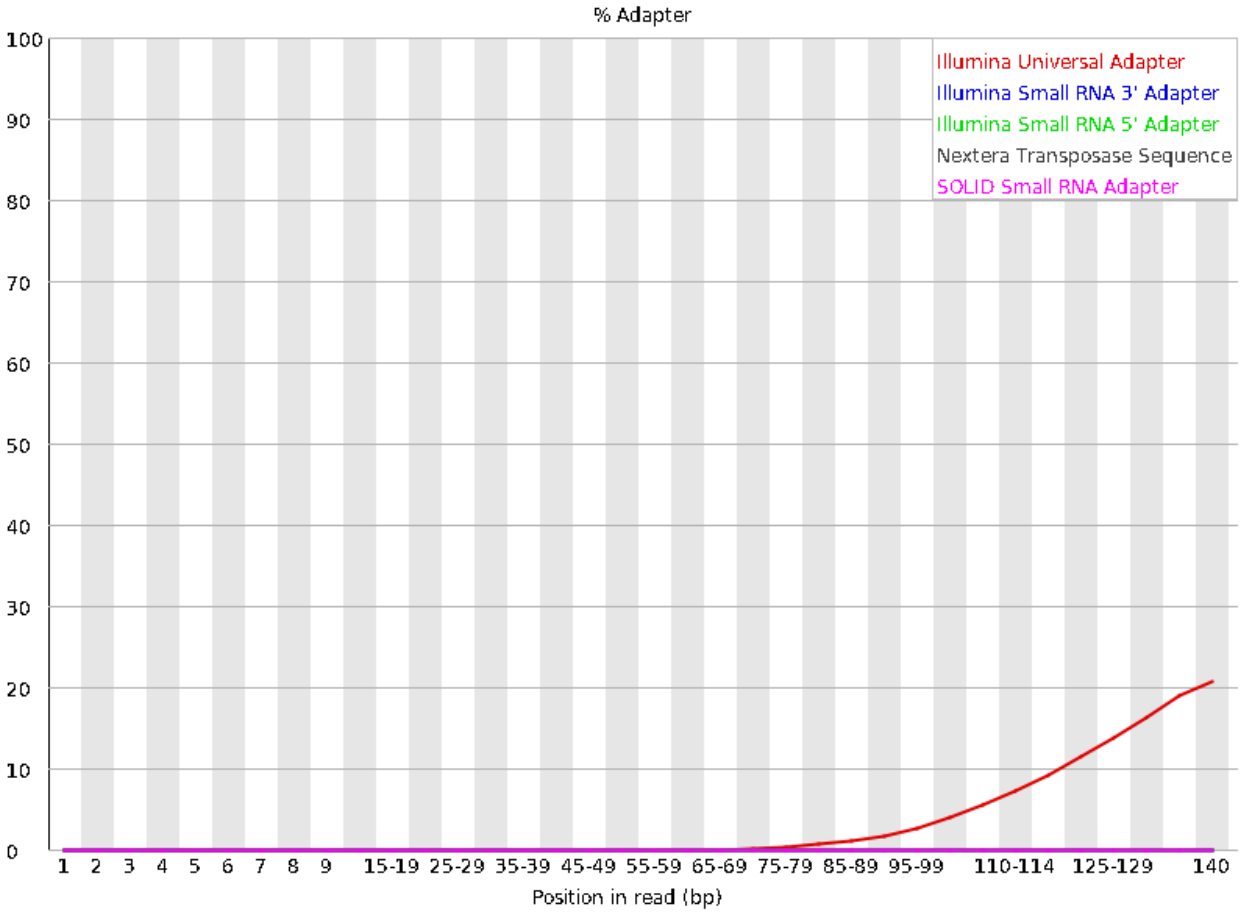


Figure 1: sequence content



2. Trim adapters using cutadapt

This study used the TruSeq Stranded mRNA LT Sample Prep Kit to prepare samples.

According to the Illumina Adapter Sequences manual (Document # 1000000002694 v17 pg. 48): https://support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html

The sequence that can be used for TruSeq universal adapter trimming is:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

The authors were not clear about the adapter sequences they used, but fastqc detected universal adapter sequences in this sample, so we will assume they used this sequence and use cutadapt to trim it.

Trimming improved the per base sequence content quality metric by removing the first 10 bases from the reads in the fastqs (as shown above), but the adapter content did not change, which means that a different adapter sequence was used. Since we do not know what this sequence was, we will move on and use the resulting trimmed fastqs for subsequent steps.

3. Align fastqs to hg38 reference genome using bowtie2 and convert to BAM file

4. Produce a counts file using featureCounts

After this step, we should have a text file that displays the number of reads assigned to each given gene ID. This will tell us the expression levels of each gene.

The output of this script is a read counts file with 28397 rows:

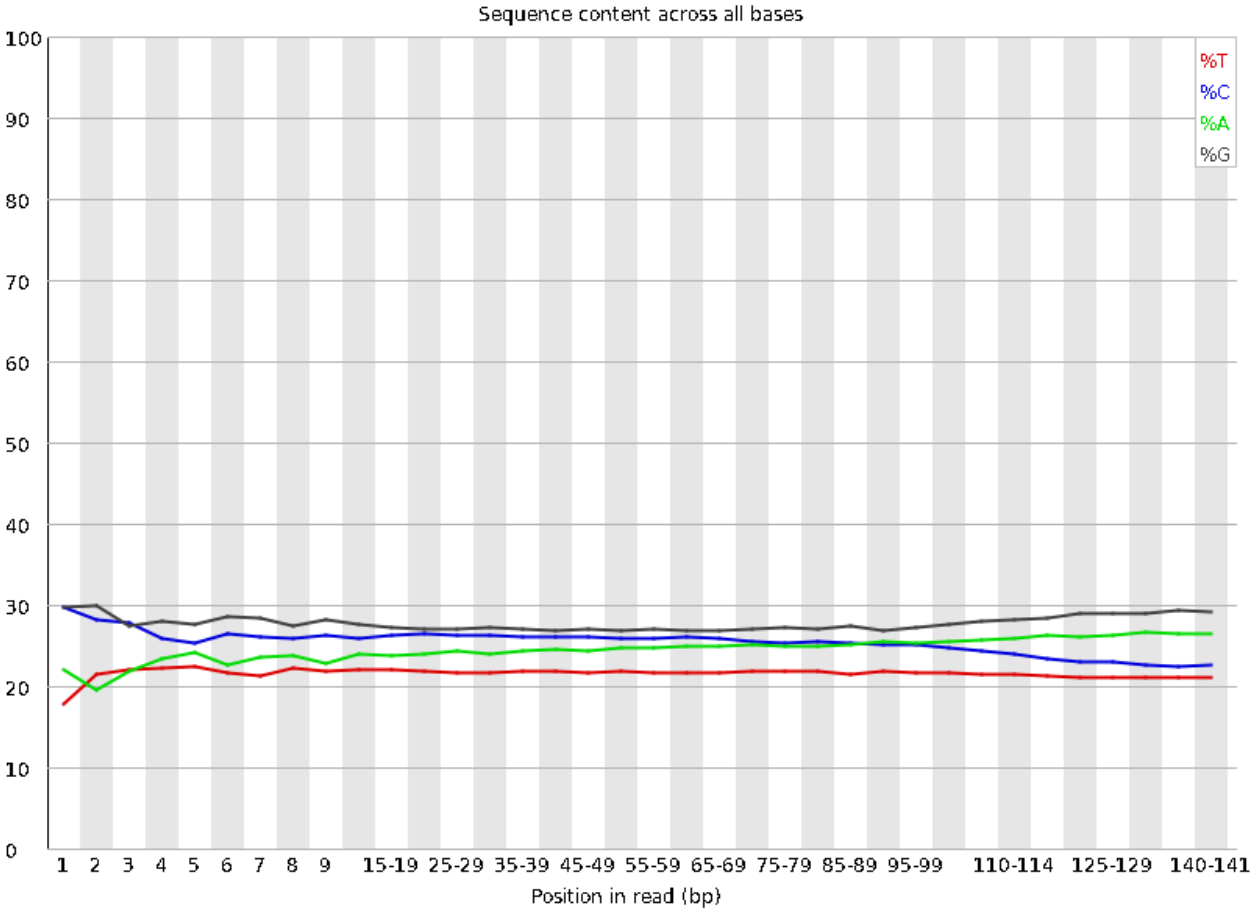


Figure 2: adapter content

The featureCounts program states that a total of 56732052 alignments occurred, and 21202936 (37.4%) of alignments were successfully assigned. It is likely that the performance of the alignment would have been better if we had removed the correct adapter sequences using cutadapt.

Once we had verified these steps for one sample, we incorporated them into a pipeline.

We made a task file called sample_names.txt, with the following contents:

LateOnset_SRR21523090 LateOnset_SRR21523091 EarlyOnset_SRR21523117 EarlyOnset_SRR21523118

We used the script runTheseJobsSerially from assignment 2, made a pipeline file called fastqToReadCounts.sh, and tested the pipeline as follows:

We then ran featureCounts on all the BAM files generated.

Part II: Analysis of read counts data to identify differentially expressed genes (DEG) between early-onset and late-onset groups

To replicate the steps which produced the heatmap in Figure 1A and volcano plot in Figure 2B, we use the read counts file available from the GEO database that had been generated from all the fastq files. We also compared plots produced using our filtered DEG list versus the filtered DEG list provided in supplementary appendix 1 by the authors.

Part II A: Generating a normalized DEG list from the read counts file

It appears that many genes match between our list and the authors' list, and the values in the corresponding columns are usually close. A notable difference is that the top 4 logFC candidates in our table (SFTA3, SFTPB, SFRP1, SLC5A8) and the bottom 4 (LYPD2, TSIX, ITLN2, and LOC102723453) are not in the authors' table. Some of these have very large |logFC| values (e.g. SFTA3: 6.796528, LYPD3: -8.039594). These 8 genes met the filtering criteria, so we think they should have been included in the analysis. Furthermore, these genes have very low FDR values, which further favors their inclusion. If the authors had other reasons to exclude these genes, we think that they should have mentioned them in the paper. For comparison, we provide heat maps based on both our DEG list and the authors' DEG list.

Part II B: Generating a heatmap of differentially expressed genes.

Figure 1 in the paper displays a heatmap of DEGs, comparing the early-onset to the late-onset sample groups.

In this heatmap, we have plotted the z-scores (as discussed) for the cpm gene expression values of our top logFC ranked genes for all patient samples. We have sorted the patient samples into early-onset (left) and late-onset (right).We have left the sample names and gene names in the heatmap, so that it can be verified that the samples are sorted into early and late-onset groups (left half: early-onset, right half: late-onset), and the gene list matches our gene list.

It appears that in both our heatmap and the heatmap shown in the paper, there are differences in expression between the early- and late-onset cohorts. Both heatmaps have higher expression levels visible in the top left and bottom right quadrants, and lower expression in the bottom left and top right quadrants. This gives some evidence that we correctly selected genes that were differentially expressed between early-onset and late-onset groups. However, there are some genes in our list with high expression across all samples that overwhelm the signal, and there appear to be no genes with z-scores below -2 in our heatmap. To determine whether the issue is with the list of genes we used, we made another heatmap using the DEG list from the authors' appendix table 1.

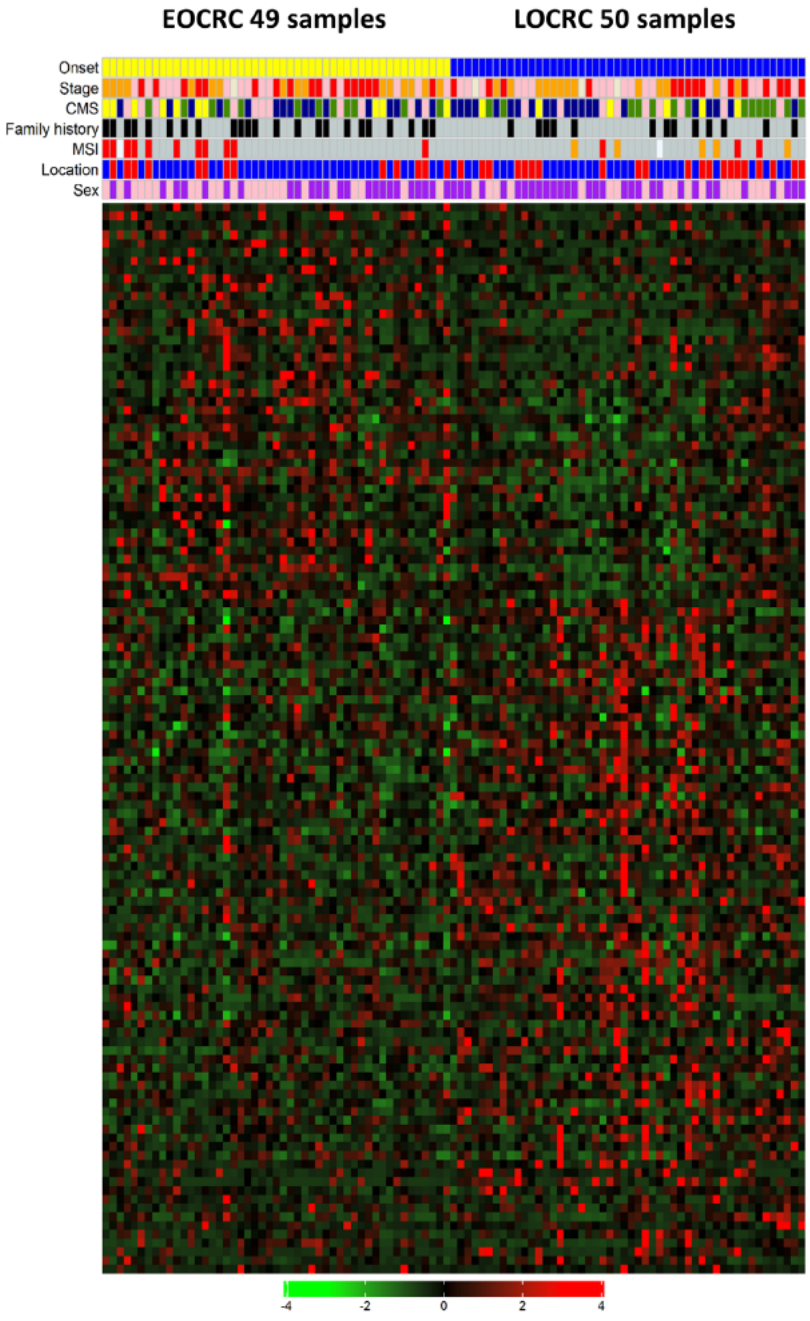


Figure 3: sequence content

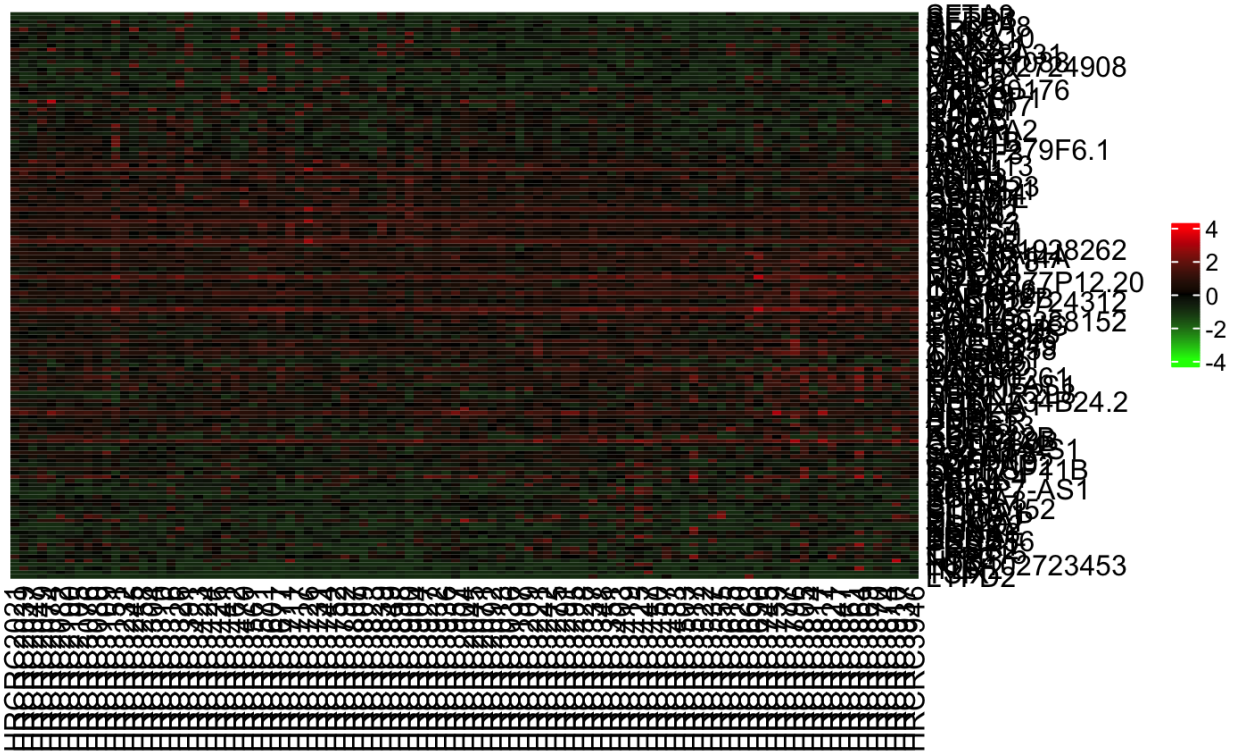
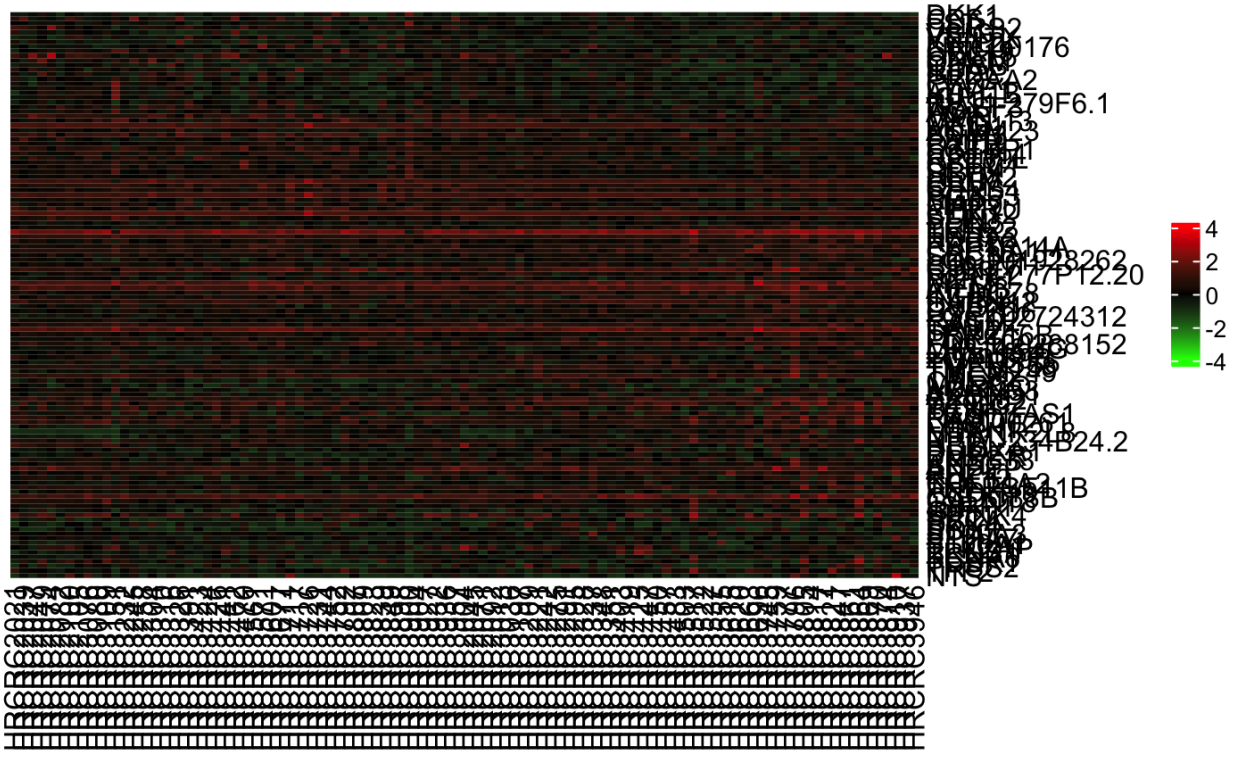


Figure 4: sequence content



In general, this heatmap looks quite similar to the heatmap generated from genes in out DEG list. This heat map also has some visible differences between the early onset and late onset groups, but it has the

same issue with globally overexpressed genes as the first heatmap.

This suggests that the difference between our heatmap and the paper's likely results mainly from our normalization steps, rather than our gene list, since the gene list for the second heatmap was the same as the one used in the paper. It is worth noting that we have sorted the samples by cohort (early-onset and late-onset), but we do not know whether they are sorted within each cohort in the same way as in the paper, so this could contribute to differences in appearance between the paper's heatmap and ours. The paper's description of the normalization method is brief (for example, z-score normalization was not mentioned), so the order in which steps were carried out is unclear.

Possible further approaches to replicate the paper's results might include: - Calculating the z scores before calculating cpm, or only calculating z scores without cpm. - Removing high-signal genes and re-scaling (say, removing the top 6 genes with the highest sum of values across rows). - Performing one or more of the normalization steps (TMM, cpm, and/or z-scores) after filtering the genes instead of doing these all before filtering. - Filtering based on false discovery rate (FDR) instead of P-value and logFC (as discussed). - Contacting the authors to inquire about their normalization steps

In the interest of time, we will not experiment with altering the order of these steps, but the above could be a good approach.

Part IIC: Generating a volcano plot from the DEG list

The authors made a volcano plot of -log10 P-value versus Log2FC.These volcano plots display the genes that meet the authors' defined threshold values for |log2FC| and -log10 P-value in the top right and left squares.

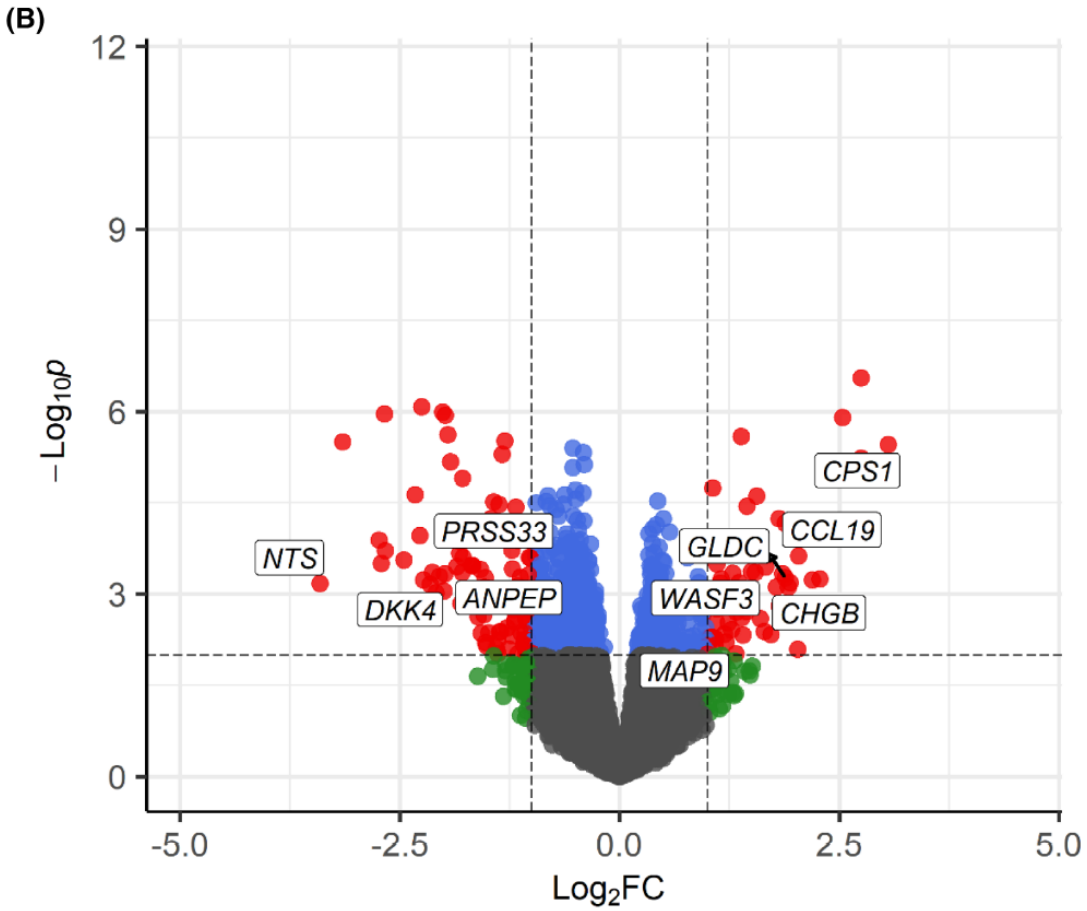


Figure 5: sequence content

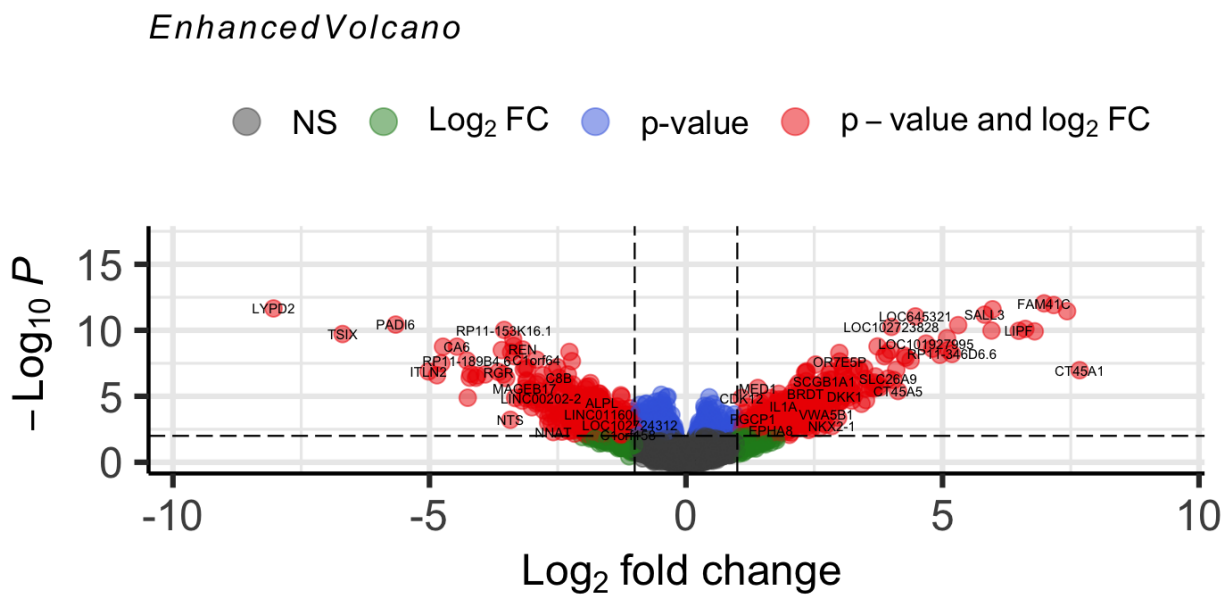


Figure 6: sequence content

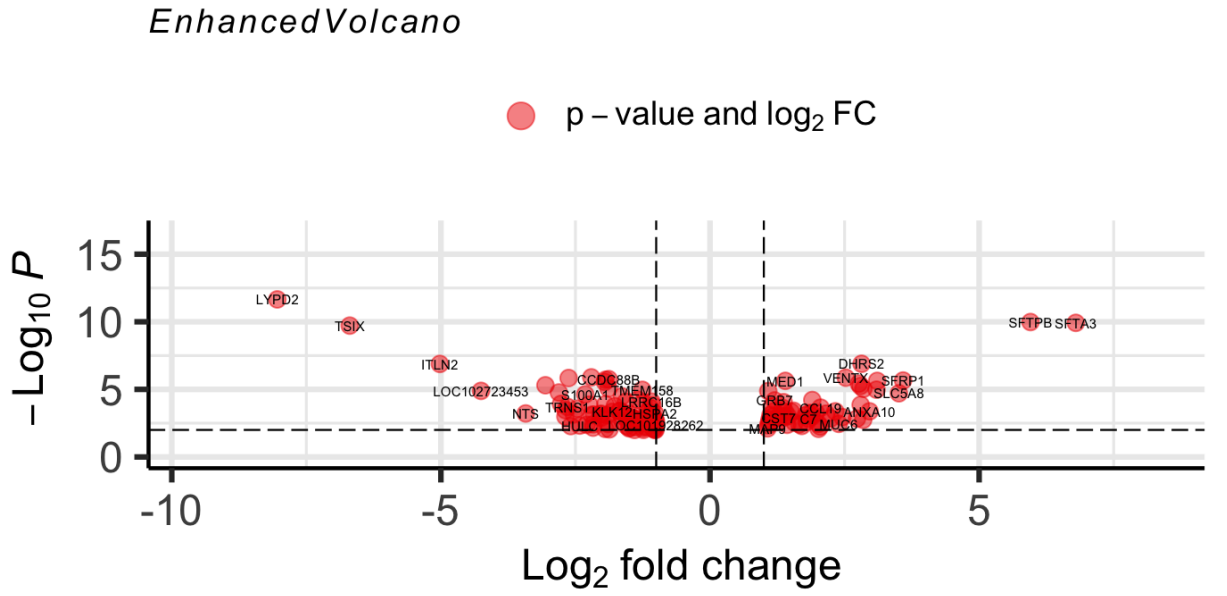


Figure 7: sequence content

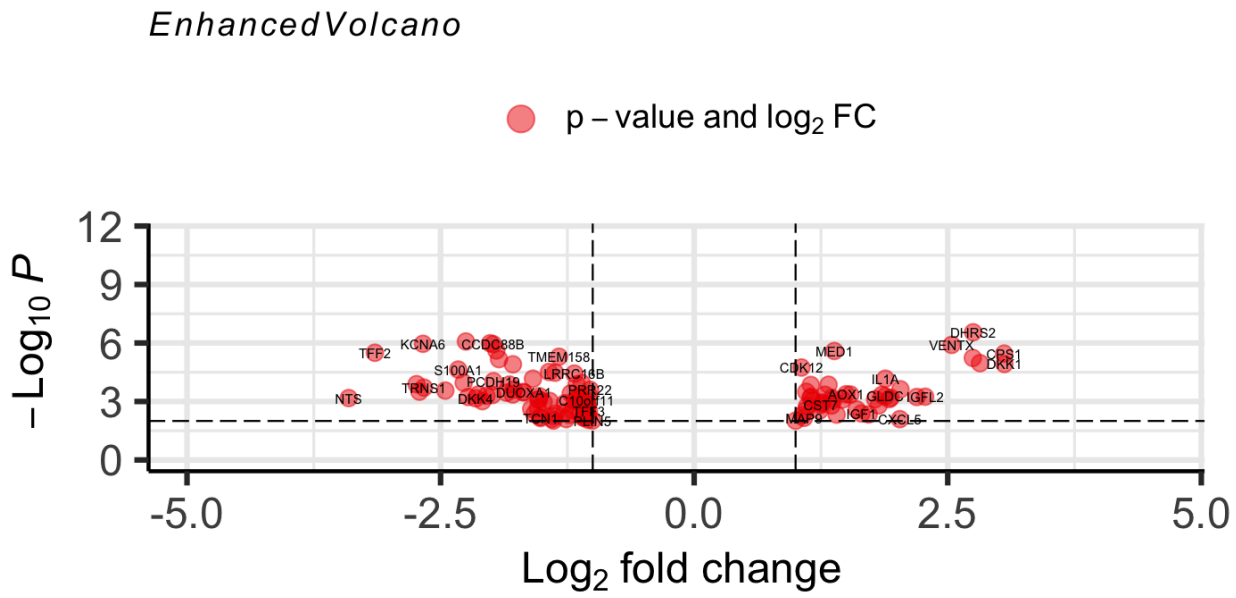


Figure 8: sequence content

Since the filtered volcano plots only include the genes that already met the threshold values, they only show genes in the top right and left squares. The authors only included their filtered gene list in the appendix, so we cannot compare their unfiltered list to ours.

Several genes appear to match in position between these four plots, including NTS, DKK4, CCL19, and PRSS33, which gives evidence that our replication of Figure 2 is successful. Other genes are not labelled in the paper but appear in the same positions between the 3 plots (e.g. DHRS1, CCDC88B, CDK12). Our filtered volcano displays some genes with much higher |log2FC| and -log10P values than those identified by the authors (as discussed earlier, e.g. the top 4 and bottom 4 log2FC genes). It remains unclear why the authors excluded these genes from the analysis given their low FDR values, and this bears further investigation.

Conclusion:

Our pre-processing steps generated read counts from fastq files provided on the GEO database. The authors did not write out a detailed description of their pre-processing steps, so we followed their general description as well as we could. When we performed differential gene expression analysis using the counts file provided on the GEO database, a large number of genes matched between our filtered DEG list and the authors' list, and the values obtained for logP, log2FC, and FDR were generally very close. Notably, the 8 most differentially expressed genes we found were not present in the authors' list. Since these genes also have very low FDR values, we conclude that they bear further investigation. Our heatmap displayed visible differences between early and late onset colorectal cancer groups, but these differences were less obvious than in the original paper, and we had issues with globally highly expressed genes. Comparison with a heatmap that used the authors' DEG list led us to conclude that the differences in our heatmap were more related to the normalization steps than the DEG list. We proposed further steps that could be taken to rectify our normalization. Comparison of the three volcano plots that we generated gave good evidence that we properly replicated the authors' methods for producing a DEG list, and further demonstrates that the highly differentially expressed genes on our list that were not in the authors' list warrant further investigation.

Overall, we were able to successfully repeat the steps of the authors' analysis using the data they had made publicly available on GEO, and our results were comparable to theirs, though there are some interesting discrepancies that could benefit from further analysis.

Authors: Liam Brockley (26182865) and Sakshi Goyal (72365547)

References:

1. Ha, Ye Jin, Yun Jae Shin, Ka Hee Tak, Jong Lyul Park, Jeong Hwan Kim, Jong Lyul Lee, Yong Sik Yoon, Chan Wook Kim, Seon Young Kim, and Jin Cheon Kim. "Reduced Expression of Alanyl Aminopeptidase Is a Robust Biomarker of Non-Familial Adenomatous Polyposis and Non-Hereditary Nonpolyposis Colorectal Cancer Syndrome Early-Onset Colorectal Cancer." Cancer Medicine n/a, no. n/a. Accessed April 5, 2023. https://doi.org/10.1002/cam4.5675.

2. Gu, Z. Complex Heatmap Visualization. iMeta 2022.

3. Blighe K, Rana S, Lewis M (2022). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.16.0, https://github.com/kevinblighe/EnhancedVolcano.