MCB 516a Final Project Report

Alex Borowiec, Bryan Jacobs, Kate Johnson

Reference Paper: Habowski, A. N., Flesher, J. L., Bates, J. M., Tsai, C.-F., Martin, K., Zhao, R., Ganesan, A. K., Edwards, R. A., Shi, T., Wiley, H. S., Shi, Y., Hertel, K. J., & Waterman, M. L. (2020). Transcriptomic and proteomic signatures of stemness and differentiation in the colon crypt. *Communications Biology*, *3*(1). https://doi.org/10.1038/s42003-020-01181-z

Introduction:

The colonic epithelium's remarkable ability to regenerate itself every 3 – 5 days is maintained by balancing stem cell renewal with terminal differentiation.[6] Stem cells, and their associated supporting cells, are located in stem cell niches at the base of folds in the tissue, known as crypts. As stem cells differentiate into mature cell populations, they move up and out of the stem cell niche into the transient amplifying (TA) compartment of the colonic crypt where they adopt a progenitor cell fate. These progenitor cells within the TA compartment will then further differentiate into final cell fates within the colon, namely enterocytes, tuft cells, goblet cells, and enteroendocrine cells. Transition from a pluripotent state within the stem cell niche to terminally differentiated cells outside of the niche coincides with dramatic changes to mRNA expression in each cell population. However, it was unknown if changes in mRNA expression are an early event in stem cell differentiation (i.e. when stem cells first differentiate into progenitors within the TA compartment) or if other processes, such as mRNA processing, alternative polyadenylation, or protein expression have a more significant role.

This paper sought to understand these questions through a multiomics approach that looked at mRNA expression levels, mRNA splicing, polyadenylation, post-translational modifications, and proteomics across all cell populations within the colon. To achieve this multiplexing the authors optimized a Fluorescent Activated Cell Sorting (FACS) protocol using well established markers of each cell population on mouse colons to enrich each cell type found within the colon. In total, six populations of cells were sorted for downstream analysis; three of which were terminally differentiated mature cells (enterocytes, enteroendocrine cells, and tuft cells), two progenitor populations that give rise to mature differentiated cells (secretory progenitors and absorbative progenitors), as well as the colonic stem cells that give rise to the progenitor populations. With this optimized FACS protocol the authors then comprehensively analyzed each cell population using their multiomics approach. Interestingly, this paper did not use single cell RNA

sequencing, however their FACS approach allows for a pseudo-single cell analysis since each population that was enriched is one cell type.

Through these investigations the authors found that the initial transition from stem cell into progenitor cell is not driven by large changes in mRNA levels. Rather, alternative splicing, alternative polyadenylation, and protein changes dominate this early transition. For example, the authors showed that the transition from stem cell to absorbative progenitors coincides with over 900 skipped exon alternative splicing events. In contrast, their data showed that only 301 mRNAs showed significant changes in levels comparing stem and absorptive progenitor cells; 3-fold lower than alternative splicing events. Additionally, the authors were able to use their multiomics approach to tease out fine details about previously well studied signaling pathways in the intestines such as Notch signaling. It has been well established that the lateral inhibition from Notch signaling in stem cells influences whether they differentiate down a secretory or absorptive lineage.[5] Typically, secretory cells express Notch ligands whereas absorptive cells express Notch receptors.[7] The authors showed in this study that early on in commitment to these fates there is alternative splicing of the Notch regulator *Spen* between stem cells and their progenitors. Only 50% of stem cells maintain full length *Spen*, the other 50% are missing one RRM RNA binding domain. In contrast, 100% of progenitor cells show full length *Spen* – indicating a switch in splicing early on in the transition from stem to progenitor with a subsequent shift in mRNA levels as maturation of each cell type occurs.

The authors of this paper were also able to gain new insight into how the regenerative capacity of the colon is maintained during injury. One fascinating observation in this area is that stem cells are not strictly required to regenerate the intestines after injury.[2] Current models propose that some level of de-differentiation occurs in the colonic epithelium when the tissue is damaged, where differentiated cells revert into a plastic state and can replace stem cells that were lost during injury.[2] The authors found significant support for this model in their data stating that, *"…stem cells are not so much defined by what they express, but by what they do not express."* They argue, and their data supports, the view that classic markers of stem cells and pluripotency, such as *Lgr5*, *Smoc2*, and *Cd44*, are also expressed in differentiated cells. The difference between the populations is that mature differentiated cells express more mRNA compared to stem cells, and that this final difference in expression is driven by large changes in early alternative splicing.

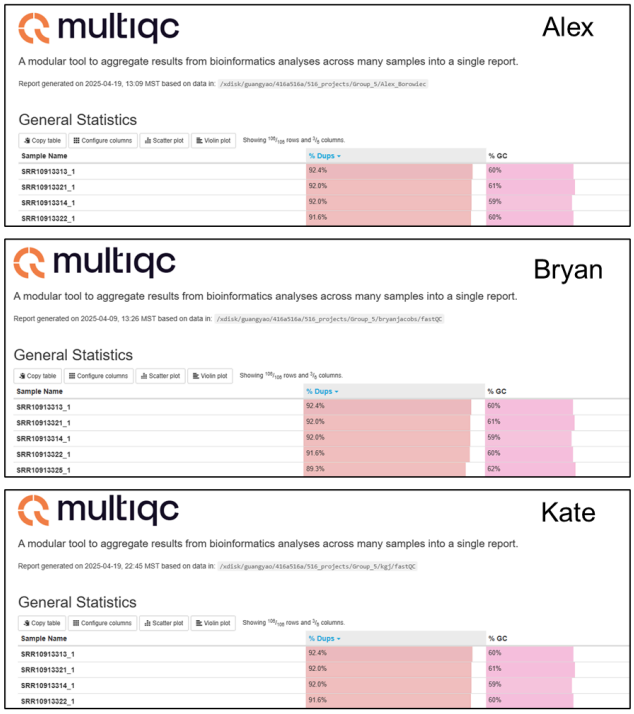
While the authors of this paper were focused on the initial transition of a stem cell into a progenitor cell, our group focused our analysis on comparing stem cells to fully differentiated enterocytes – a cell population that is classically thought of as the furthest from stemness. We chose this comparison since mRNA levels do not change dramatically when a stem cell becomes a progenitor cell. We reasoned that comparing stem cells and enterocytes would give us large changes in mRNA expression that we could run our analyses on. Our group was able to successfully reproduce figures and data from the reference paper with good agreement. We also found the data generated in this investigation to be robust to pre-processing and alignment – i.e., even though we changed parameters during RNAseq data processing our downstream analysis still somewhat recapitulated the results of the paper.

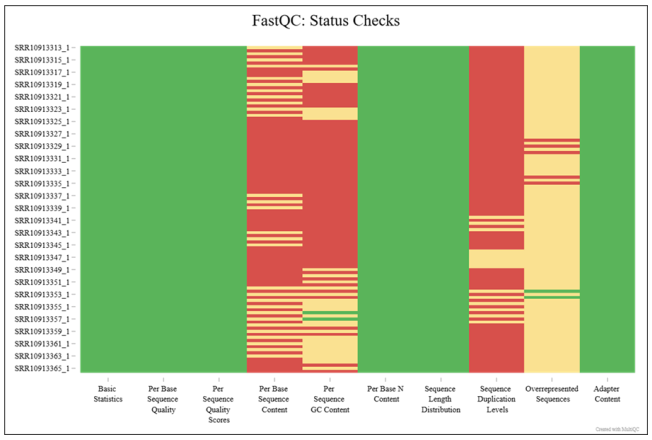
Results and Discussion:

HPC Analysis:

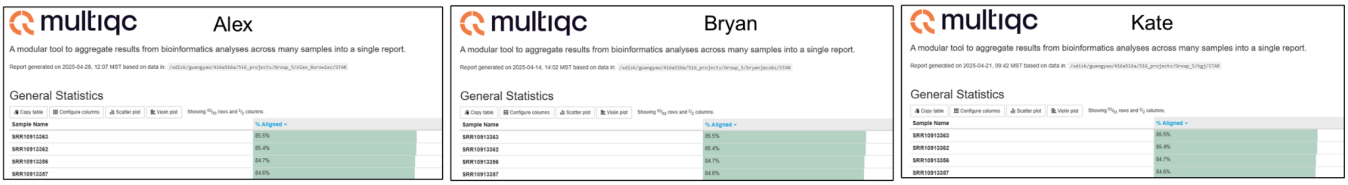
FastQC and STAR MultiQC

Each group member independently performed MultiQC on fastQC and STAR using default parameters. MultiQC reports for each group member are identical for both fastQC and STAR. For fastQC, percent duplicate reads ranged from 39.2% - 92.4% across all samples analyzed and total sequences ranged from 15.0 - 34.1 million. Overall, our fastQC results were normal for basic statistics, per base sequence quality, per sequence quality scores, per base N content, sequence length distribution, and adapter content. However, indication of abnormalities in the data was detected for per base sequence content, per sequence GC content, sequence duplication levels, and overrepresented sequences. For multiQC on STAR alignment using default parameters, the percentage of uniquely mapped reads ranged from 45.9 - 85.5% across all samples and uniquely mapped reads ranged from 7.5 - 21.9 million. The high percentage of duplicate reads from fastQC results could potentially be a result of how the RNA was prepared for sequencing in these experiments. We investigated the kits used to isolate the RNA as described in the paper's methods (Direct-zol RNA Micro-Prep kit Zymo #11-330 M & Clontech Low Input Pico Kit Takara #634940). As far as we can tell the authors did not specifically enrich mRNA or deplete other RNAs from the samples such as rRNA. Excessive rRNA in the sample preparation could lead to high percent duplicate reads.



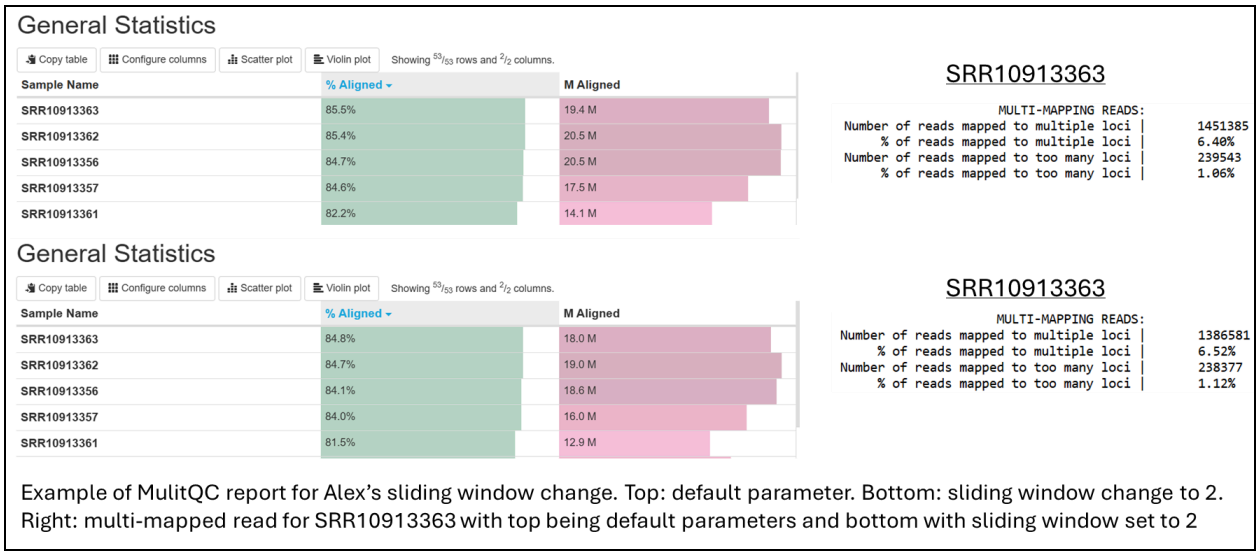


Left: Example MultiQC reports from all three group members for FastQC. Right: Summary of MultiQC for FastQC results



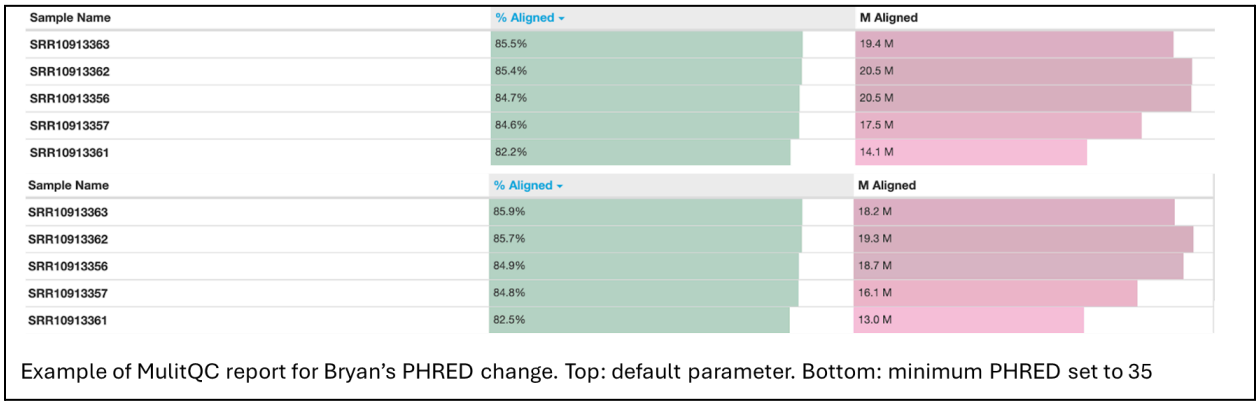
Example MultiQC reports for all group members for STAR default parameter results

Sliding Window Parameter Change. For the sliding window parameter in fastp, we chose to increase the window size from 1 to 2. The analysis should then average the quality of 2 consecutive bases instead of evaluating each base individually. This is a modest increase in size and from it we expected slightly less trimming of the reads. With this parameter change we did see modest changes to STAR alignment. For default parameters we saw 85.5% uniquely mapped reads for SRR10913363, with the increased window size we observed a decrease in that number to 84.4%. This decrease in uniquely mapped reads with the increased window coincides with an increase in the number of reads mapped to multiple loci: 6.40% in default versus 6.52% with parameter change. These results are consistent with how the sliding window affects trimming of reads. Increasing the window to 2 averages two bases which could lead to the retention of repetitive or ambiguous bases that would normally be trimmed. This could lead to the mapping of reads to multiple loci in downstream analysis.



Example of MulitQC report for Alex's sliding window change. Top: default parameter. Bottom: sliding window change to 2. Right: multi-mapped read for SRR10913363 with top being default parameters and bottom with sliding window set to 2

Min PHRED Score Parameter Change. We chose to increase the minimum PHRED score parameter in fastp to 35, from its default value of 20. While PHRED scores of 28 and above are generally considered high quality, we chose a much stricter threshold due to the majority of our reads falling above the threshold of 35. In general, with a threshold this high, we would expect to see very aggressive filtering and therefore a large decrease in the number of mapped reads in STAR alignment. However, due to the high quality of the data, we saw modest changes across samples. With default parameters, SRR10913363 had 19.4 million aligned reads, and 85.5% of reads aligned. With the PHRED score increased to 35, SRR10913363 had 18.2 million reads aligned and 85.9% of reads aligned. This trend was generally seen throughout the samples, and can be explained by the stricter PHRED threshold filtering out more reads, increasing the overall quality of the remaining reads. It is important to note that while the remaining data leads to a slight increase in alignment percentage, a PHRED threshold of this strictness has potential to harm the overall coverage of the data.



Example of MulitQC report for Bryan's PHRED change. Top: default parameter. Bottom: minimum PHRED set to 35

Max Read Length Parameter Change. The third parameter we changed for fastp was maximum read length using the command --max_len1 # --max_len2 #. Setting a limit for maximum read length means trimming all longer reads on the 3' end to the specified length, which removes any low quality bases for long reads. Reads shorter than the specified length remain unchanged. When we initially capped at 100 base pairs, there was no change in the fastp results or PCA or heatmap plots, suggesting all read lengths were less than 100 bp for the original PCA. We then tried capping at 10 bp, which made the reads too short for alignment using STAR. Next, we attempted capping max read length at 40 bp, which was also too short; we then realized reads need to be a minimum of 50 bp for STAR, and ideally between 75-100 bp. With this information, we finally settled on a maximum read length of 85 bp which subtly altered our fastp results and subsequent plots.

Sample Name	% Aligned	M Aligned
SRR10913363	85.5%	19.4 M
SRR10913362	85.4%	20.5 M
SRR10913356	84.7%	20.5 M
SRR10913357	84.6%	17.5 M
SRR10913361	82.2%	14.1 M

Sample Name	% Aligned	M Aligned
SRR10913363	85.1%	19.3 M
SRR10913362	84.8%	20.4 M
SRR10913356	84.2%	20.4 M
SRR10913357	84.1%	17.4 M
SRR10913361	81.7%	14.0 M

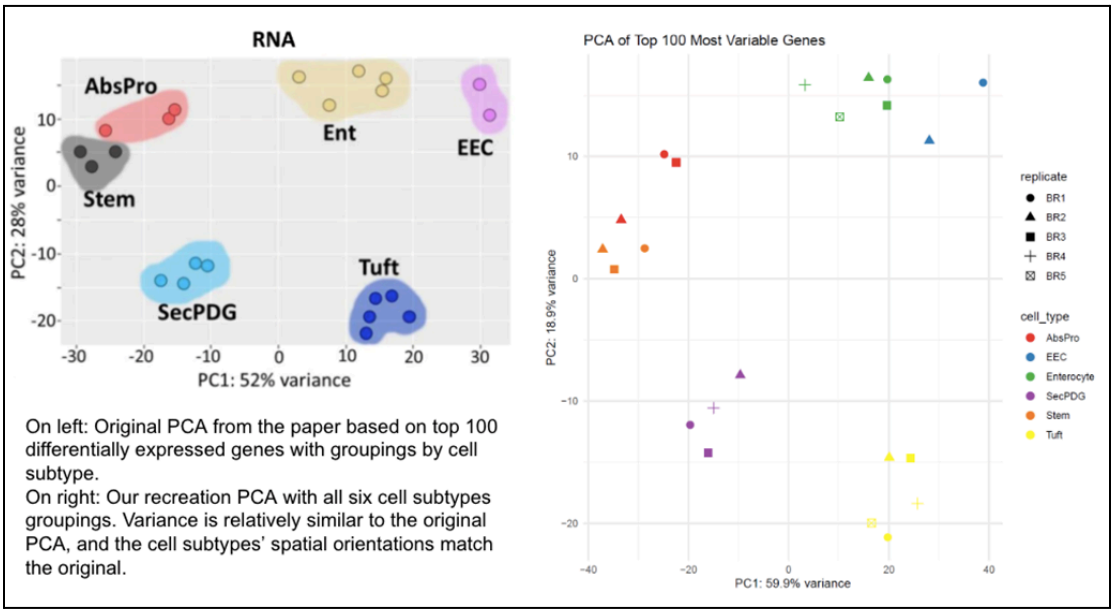
Example of MulitQC report for Kate's max read length change. Top: default parameter. Bottom: max read set to 85

Differential Expression Analysis:

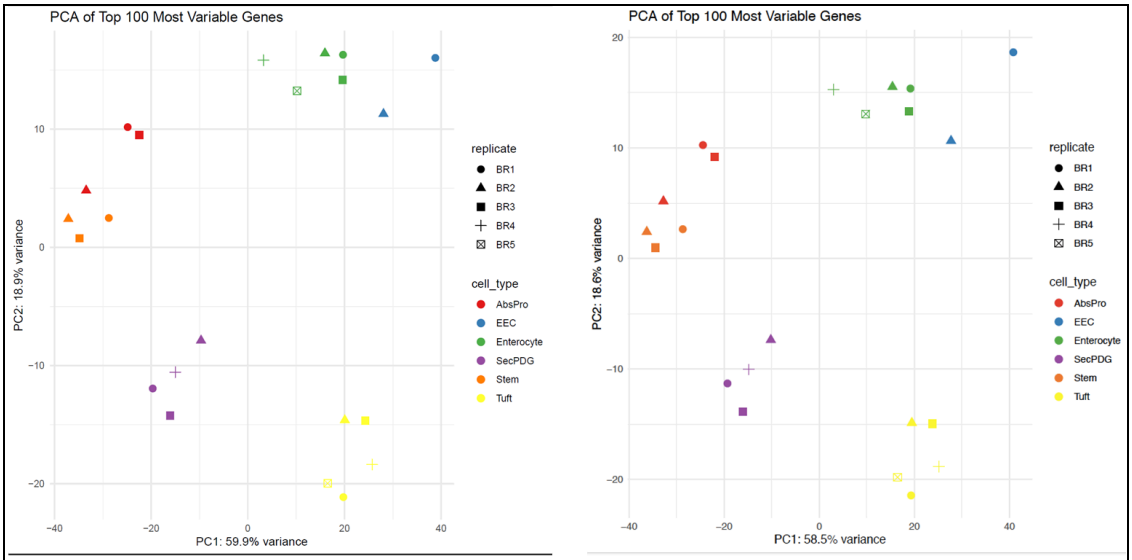
PCA

PCA maps generated using default parameters matched closely with what the authors published in the reference paper. There were changes with variance explained percentages and some of the grouping was slightly different. We attribute these differences to different versions of DESeq2 used in the paper versus what we used in our analysis (reference paper DESeq2 = 1.16.1. Our version = 1.48.0). The other area that can explain the difference we see comes from how the authors defined biological replicates as well as how the authors filtered genes after STAR. In total there were 53 files for this paper, however only 22 biologic replicates were used in the paper (stem = 3, AbsPro = 3, SecPDG = 4, tuft = 5, Ent = 5, and EEC = 2). Eventually, we

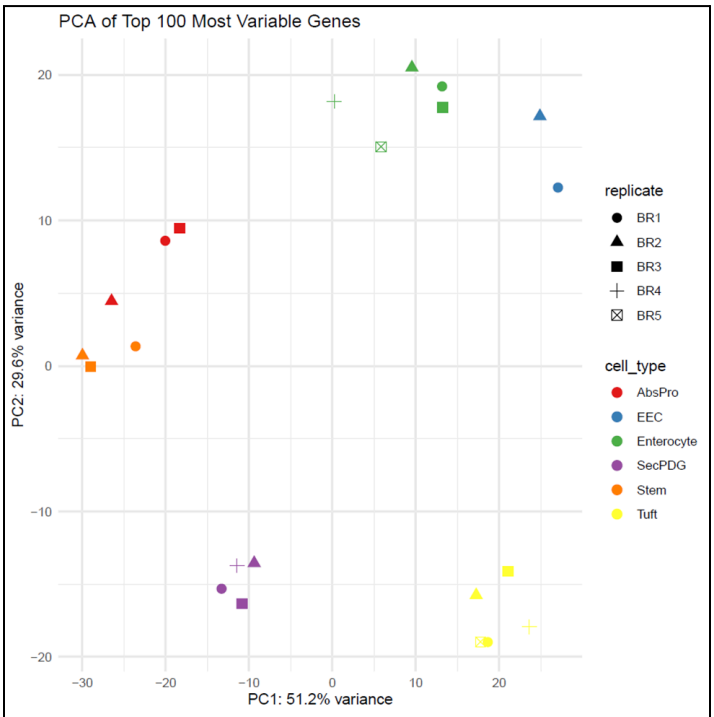
found that each of the 22 biologic replicates had differing numbers of technical replicates within each biological replicate. The authors did not directly address how technical replicates were handled in their work. For our analysis, we kept one of the technical replicates that had the highest counts for each biological replicate. If the authors used all 53 samples in their analysis versus our 22 then this could explain the slight difference we see in the PCA. In addition to replicates, we also found that at some point during the author's pipeline they specifically filtered out protein coding genes for DESeq2 analysis. The raw counts from their dataset maps to over 70,000 genes, however their final published DESeq2 results show a little over 20,000 genes - indicating that the genes were parsed at some point that was not addressed in the text. We made this discovery after initializing our individual data analysis. To confirm that our results using the full geneset still matched what the authors published, we filtered protein coding genes from our dataset and made the PCA plot again. After filtering, the number of genes in our dataset was reduced down to 21,748 genes which matched closely with the authors 24,421. Our results from this filtered set were similar to our original PCA plots using all values. This gave us confidence to keep and move forward with our analysis and we chose to keep the full gene list for our analysis. Comparing PCA plots generated by Alex and Kate using default parameters show similar plots with slight differences to the variance explained percentages on the x- and y-axis as well as grouping for EEC. We were careful to run our analysis in a similar manner, however the difference in grouping on the PCA could potentially come from the slight differences in variance explained that was calculated in R.



On left: Original PCA from the paper based on top 100 differentially expressed genes with groupings by cell subtype.
On right: Our recreation PCA with all six cell subtypes groupings. Variance is relatively similar to the original PCA, and the cell subtypes' spatial orientations match the original.



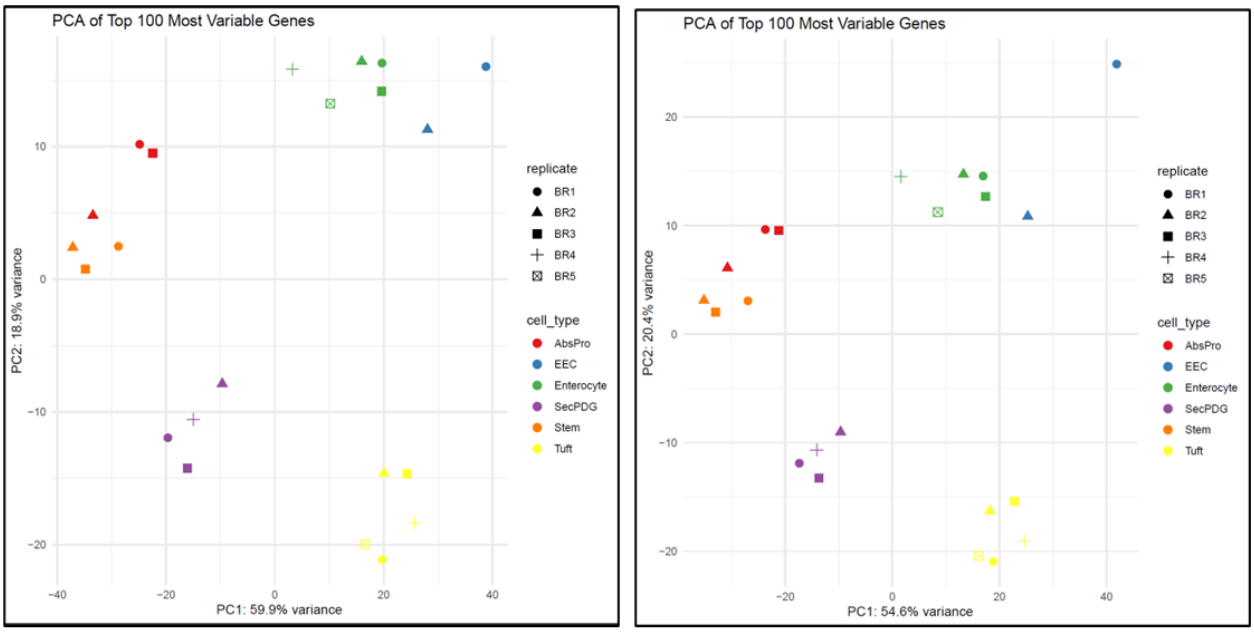
Left: Alex's PCA plot showing top 100 genes using default parameters. Right: Kate's PCA plot showing top 100 genes using default parameters.



PCA plot generated after we filter only protein coding genes from the dataset

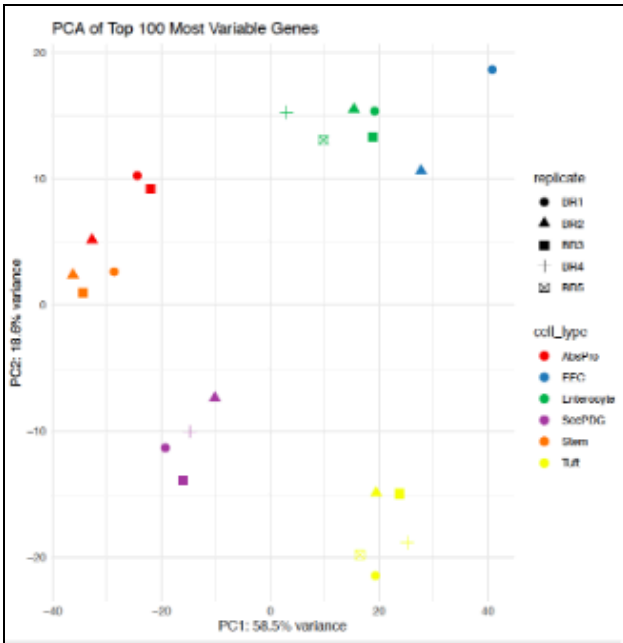
Sliding Window Parameter Change. Changing the sliding window from 1 to 2 showed modest changes to the grouping of the samples in the PCA plot. For secPDG cells the clusters became

tighter. Indicating that for secPDG the larger window improved the biologic signal potentially due to a reduction in noise when averaging bases. On the other hand, for the EEC population, one of the biologic replicates no longer clusters with the other biologic replicate. This could be due to the averaging losing single bases resolution that is important for this cell type. The paper discussed how the EEC population was the most difficult to sort and enrich for this analysis due to low cell numbers. They had to use multiple mice to get sufficient cells for RNAseq. It could be that the increased window is exaggerating inherent sample to sample variability due to the difficulty in enriching this population.

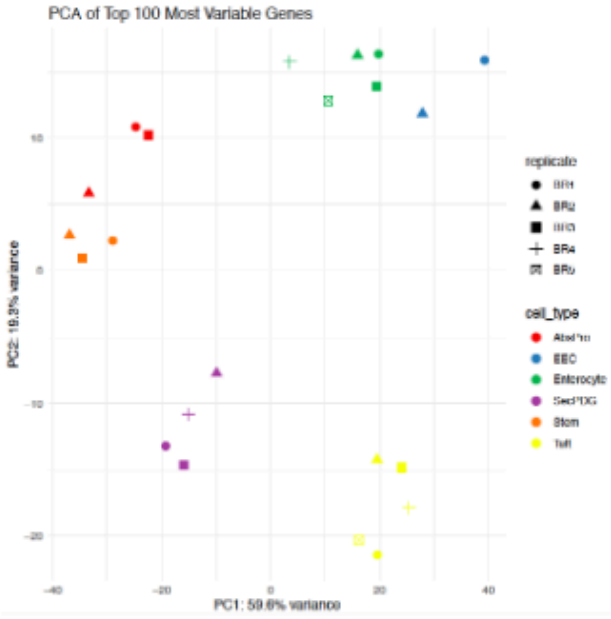


Left: PCA at default parameter of 1. Right: PCA with sliding window adjusted to 2.

Max Read Length Parameter Change. The second parameter we adjusted for the PCA plot was maximum read length, which we set to 85 bp. This changed the plot very slightly, most noticeably by shifting the two EEC samples closer together. Overall, the minimal changes suggested the original full-length reads had good integrity, but the trimming did improve clustering for the enteroendocrine cell subtype.



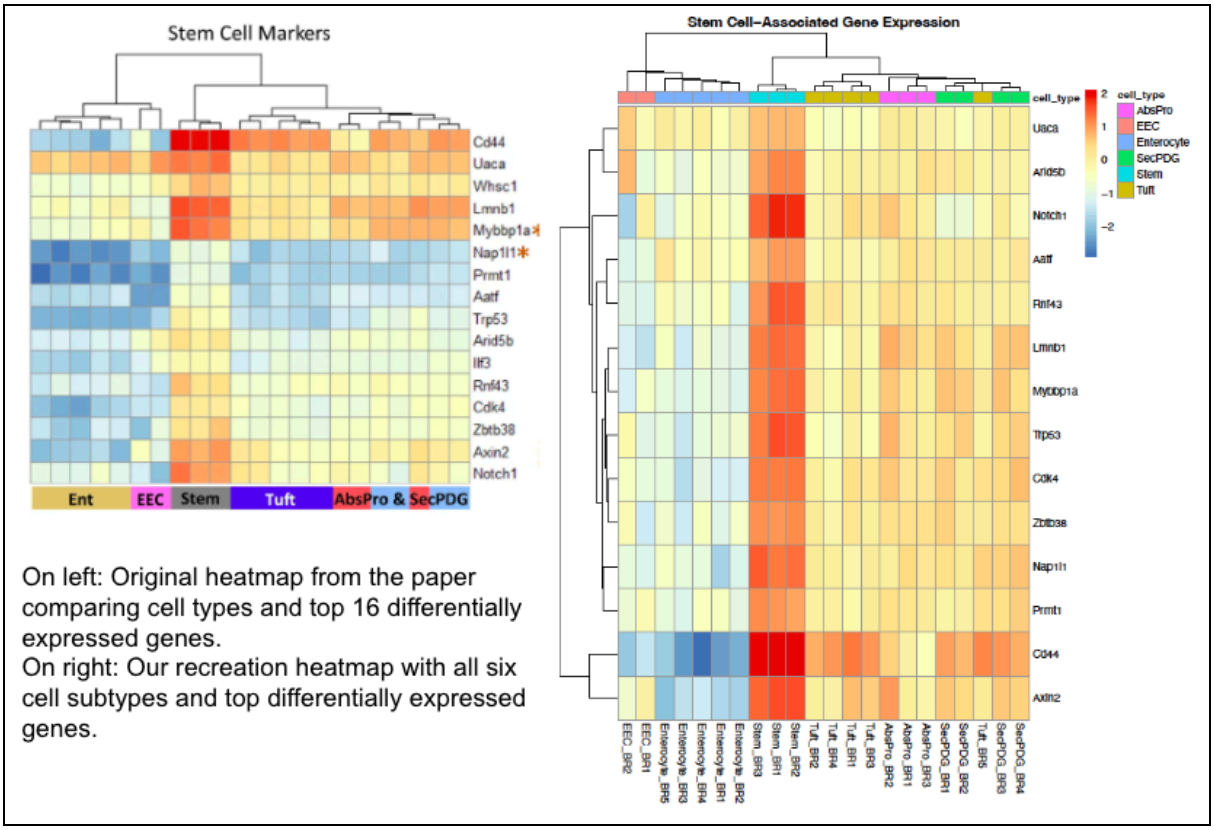
PCA of top 100 differentially expressed genes with default parameters.



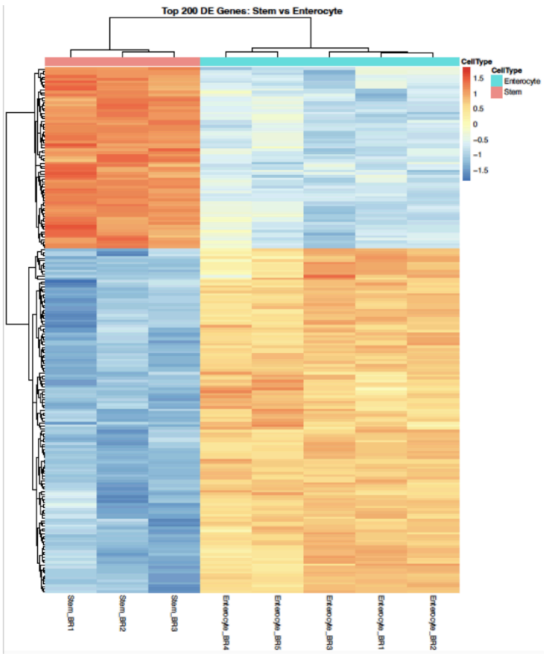
PCA of top 100 differentially expressed genes with max read length set to 85.

Heatmap

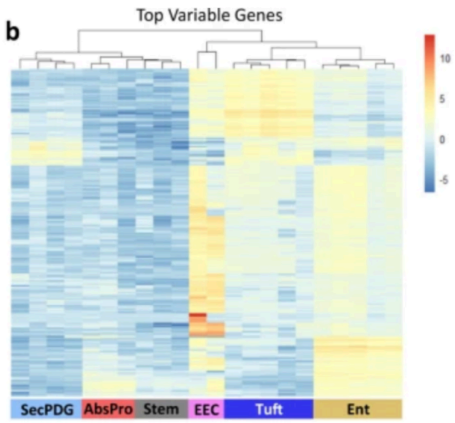
There were three heatmaps in this paper, each segregated by cell type. The first looked at the top 200 differentially expressed genes in all six cell subpopulations (secretory cells, absorptive progenitors, stem cells, enteroendocrine, tuft, and enterocytes); this was autoscaled with unsupervised clustering. Then, the top 16 genes expressed in stem cells were selected and depicted across all six cell subtypes again. We recreated this heatmap to the best of our ability below. Next, the researchers selected the genes that were most changed in the non-stem cell populations of secretory cells and absorptive progenitors. These 17 genes are depicted in stem cells, SecPDG, and AbsPro in the final heatmap. For our purposes, we chose to recreate the heatmap comparing the top 200 differentially expressed genes, but focusing on stem cells and enterocyte populations. We have one Stem vs Ent heatmap that was clustered according to rows and columns, and one clustered only by row; this clustering variation accounts for differences between the default parameter heatmaps. Focusing only on the Stem and Ent columns from the paper's heatmap, our heatmaps can be considered satisfactory recreations.



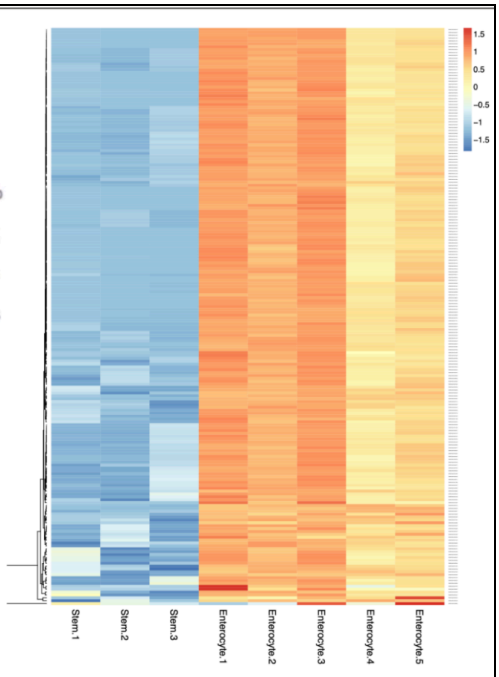
On left: Original heatmap from the paper comparing cell types and top 16 differentially expressed genes.
On right: Our recreation heatmap with all six cell subtypes and top differentially expressed genes.



Stem vs Ent heatmap of top 200 DE genes clustered by rows and cols.

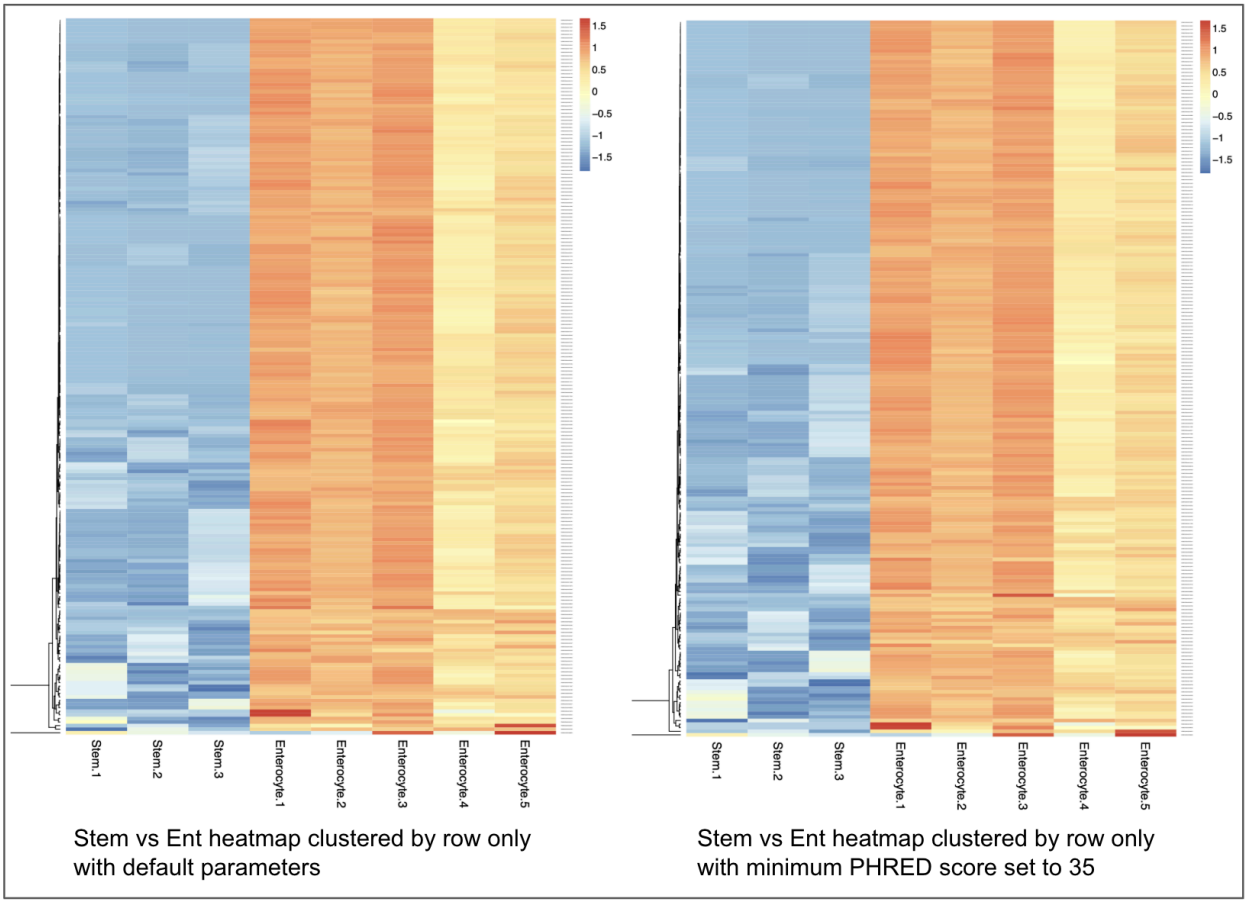


Original heatmap of top 200 DE genes from paper.



Stem vs Ent heatmap of top 200 DE genes clustered by rows only.

Min PHRED Score Parameter Change. Increasing the minimum PHRED score from 20 to 35 showed very minor changes in the heatmap of the top 200 differentially expressed genes in stem cells vs enterocytes. Since the parameter change eliminates only lower quality reads, the top 200 differentially expressed genes remain biologically consistent across both pipelines. The only difference we see is brighter expression in some areas of the modified PHRED score heatmap due to the elimination of lower quality reads and therefore statistical noise. This is a satisfactory result, as the goal of increasing the minimum PHRED score parameter is to refine results rather than drastically alter them.

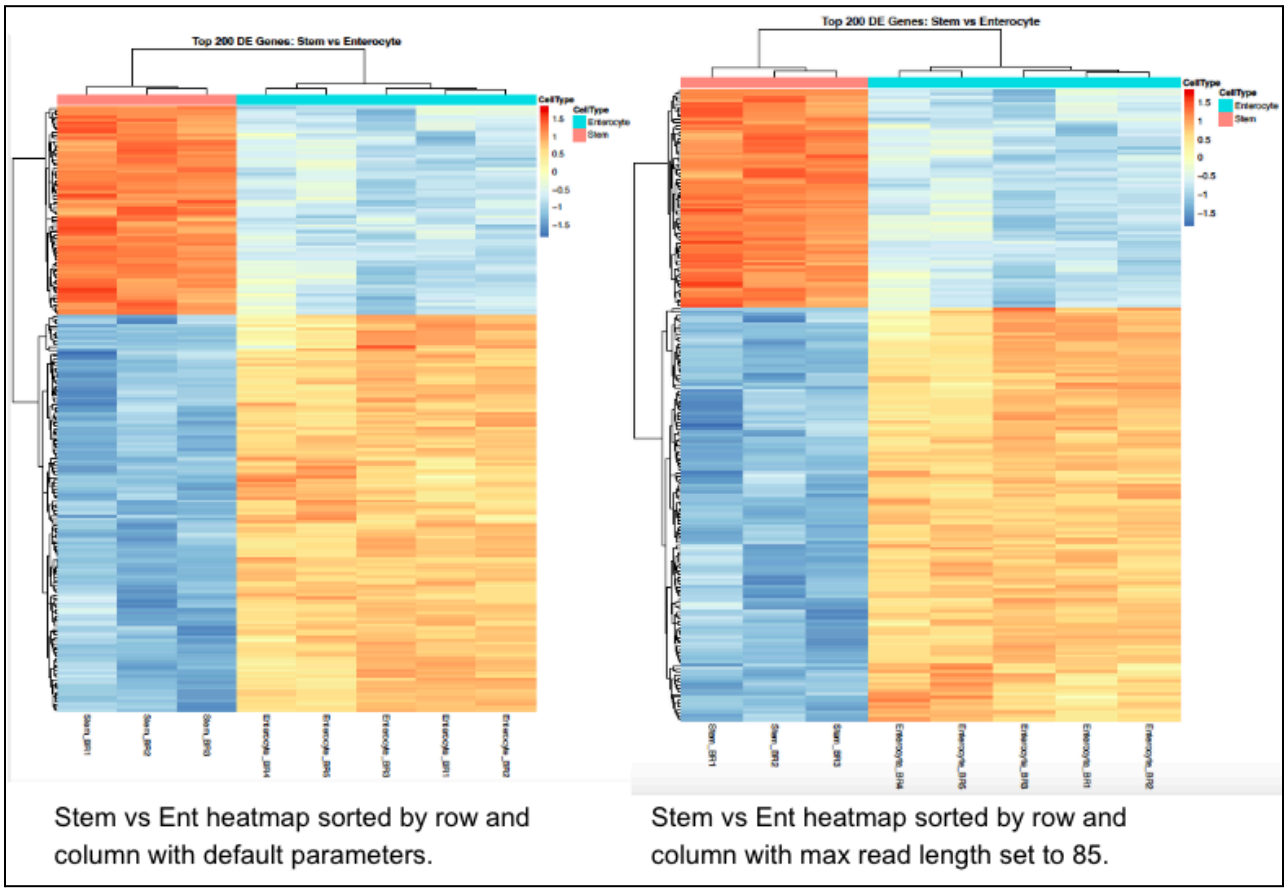


Stem vs Ent heatmap clustered by row only with default parameters

Stem vs Ent heatmap clustered by row only with minimum PHRED score set to 35

Max Read Length Parameter Change. Setting the maximum read length to 85 base pairs did not obviously change the results in our heatmap for top 200 differentially expressed genes in stem cells vs enterocytes. Because the original data was already clean and reliable, the majority of clustering and overall patterns remained the same in the modified max length heatmap. However, the more stringent parameter did likely reduce noise and increase integrity of our

results, and there are some stripes in the heatmap that have shifted around to reflect the more defined clustering.



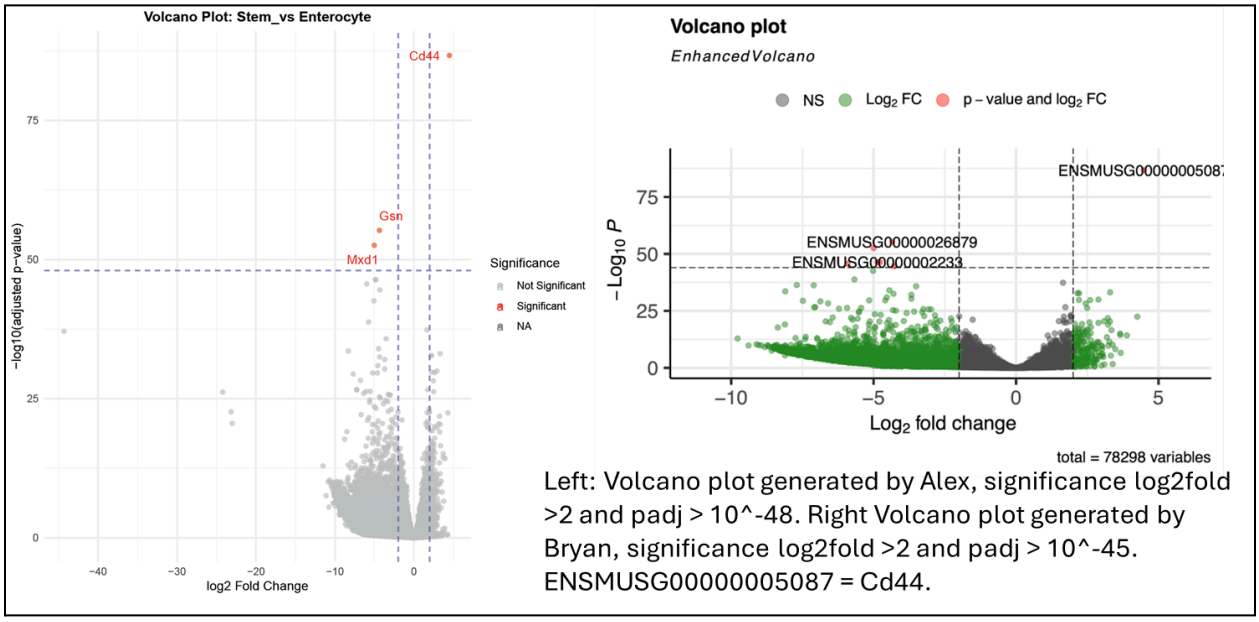
Stem vs Ent heatmap sorted by row and column with default parameters.

Stem vs Ent heatmap sorted by row and column with max read length set to 85.

Volcano Plot

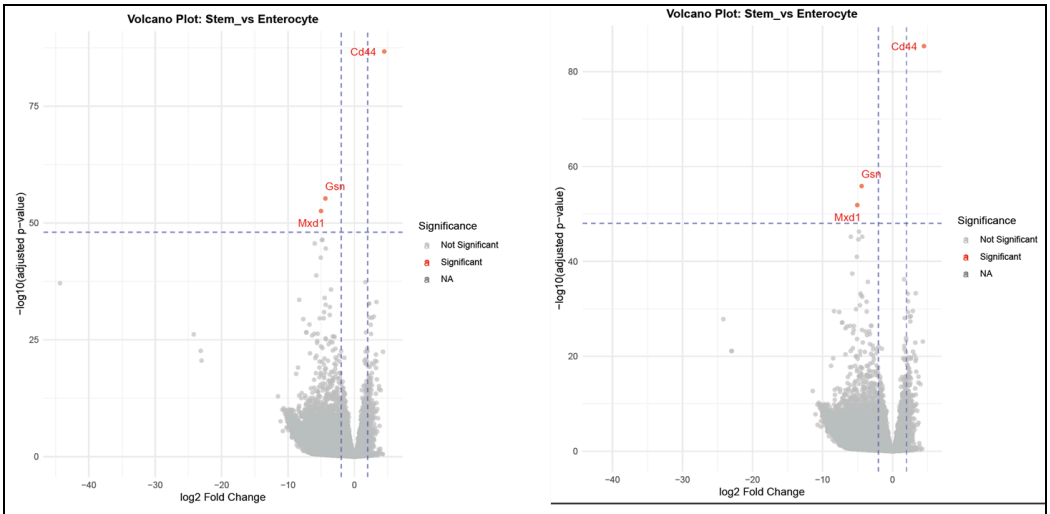
The authors did not publish any volcano plots with this paper, however our in-group comparisons of the volcano plots showed good agreement. We identified the same down and upregulated genes using similar criteria for significance when comparing stem to enterocytes. Importantly, when comparing stem cells to enterocytes *Cd44* is the most upregulated gene. Given that the authors used FACS targeting *Cd44* to sort stem cells, we expect that this gene should be highly enriched. This is what our volcano plots show. We set a stringent threshold for our volcano plots (log2fold > 2 and padj < 10^-48 for Alex's plot and 10^-45 for Bryan's plot). Genes that passed this threshold are biologically relevant for the comparison between stem cells and enterocytes. For example, *Mxd1* was one of the top genes we identified as

downregulated. *Mxd1* is a competitive antagonist for the transcription factor *Myc*.[4] *Myc* is important for intestinal stem cells to maintain their fate.[3] Since this gene is enriched in enterocytes it could potentially antagonize *Myc* to prevent enterocytes from transcribing pro-stem genes.



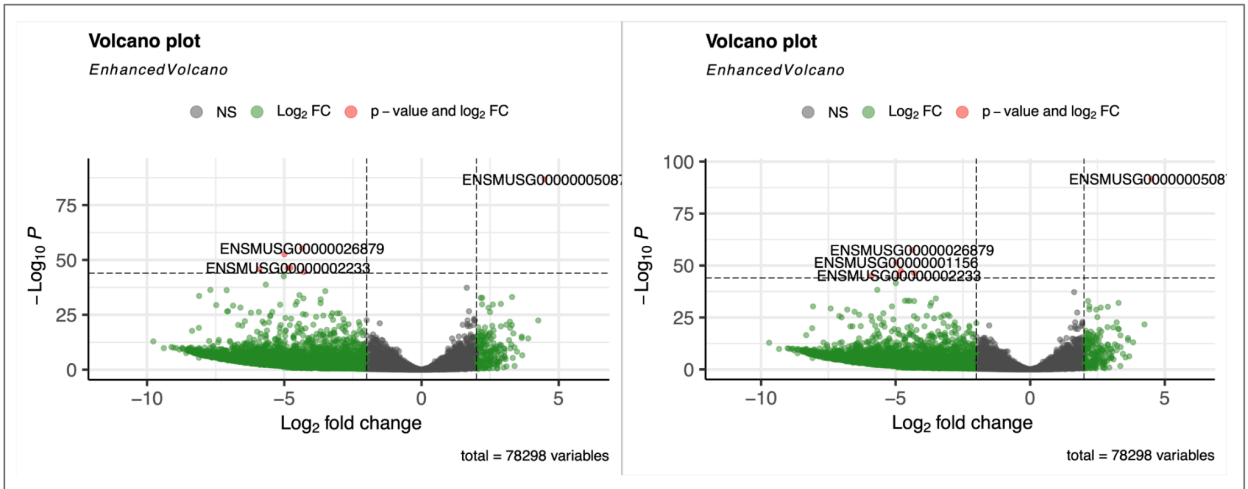
Left: Volcano plot generated by Alex, significance log2fold >2 and padj > 10^-48. Right Volcano plot generated by Bryan, significance log2fold >2 and padj > 10^-45. ENSMUSG00000005087 = Cd44.

Sliding Window Parameter Change. Increasing the window size from 1 to 2 showed minimal effects on the volcano plots. There are slight differences in the overall shape of the data for genes deemed not significant, however genes identified as significant remained so with the parameter change. This is likely due to our stringent threshold for significance.



Left: Volcano plot with default sliding window of 1. Right: Volcano plot with sliding window set to 2.

Min PHRED Score Parameter Change. Increasing the minimum PHRED score from 20 to 35 showed moderate effects on significant genes in the volcano plot. The most significant gene, *Cd44*, saw a significant decrease in its adjusted p-value. Log2FC values remained consistent for significant genes across both plots. The decrease in adjusted p-value can be attributed to the high quality filtering strengthening true biological signals by removing noise, while the consistent Log2FC is further proof that the parameter change did not alter any true signals. In short, the volcano plot with minimum PHRED score increased to 35 showed similar results with greater statistical confidence.



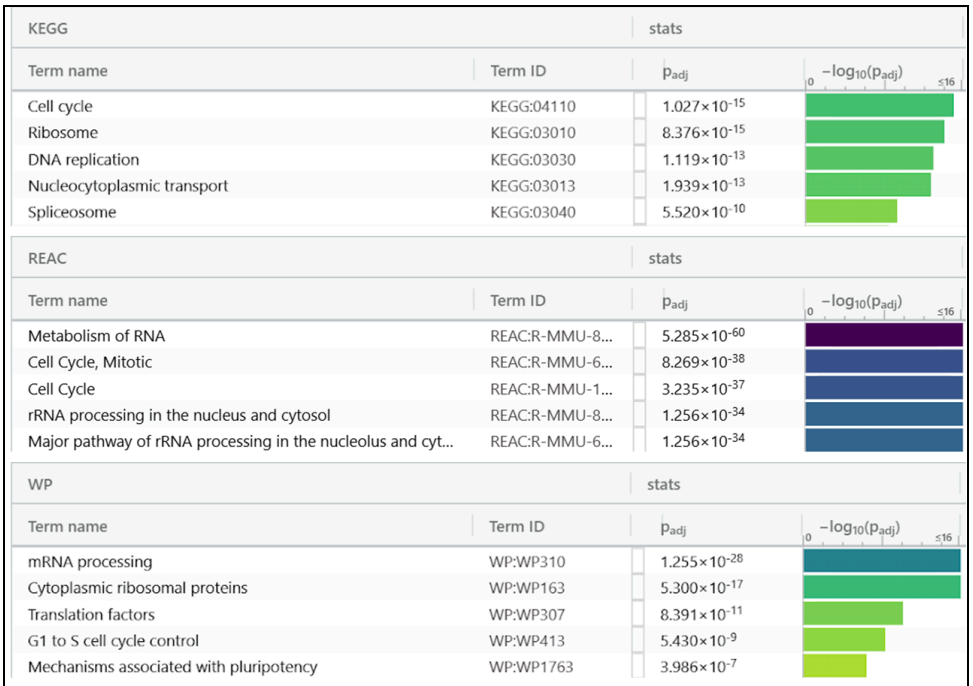
Volcano Plot with default parameters

Volcano Plot with minimum PHRED score set to 35

Functional Analysis:

G:profiler

Using default HPC parameters and the significance threshold outlined in the paper of padj < 0.01 and > 50 mean normalized counts, we ran g:profiler to look at KEEG, Reactome, and WikiPathways for both up- and downregulated genes identified comparing stem cells to enterocytes. Significance in g:profiler was set to benjamini-hochberg FDR <0.05. Both group members had good agreement in g:profiler results. Importantly, the pathways downregulated in this comparison indicated that known pathways for enterocytes were enriched, such as focal adhesions.[1] In addition to our separate g:profiler analyses we also endeavored to recreate GO pathway analysis from our reference paper. In this analysis, the authors used Panther classification system for 107 most upregulated genes in all cells except stem cells. We were able to somewhat recapitulate the author's results using Panther, however the text lacked detail on specifically how they assessed what qualified as top genes in this data. Our results were similar for Panther Protein Class indicating that Transporter was the top term. For molecular function and cellular component we got similar results to the paper, such as Cytoplasm, however for these two categories our results failed the FDR <0.05 test. We attribute these differences to the lack of detail regarding how to determine the top genes to be used as well as how to appropriately use the Panther classification system.

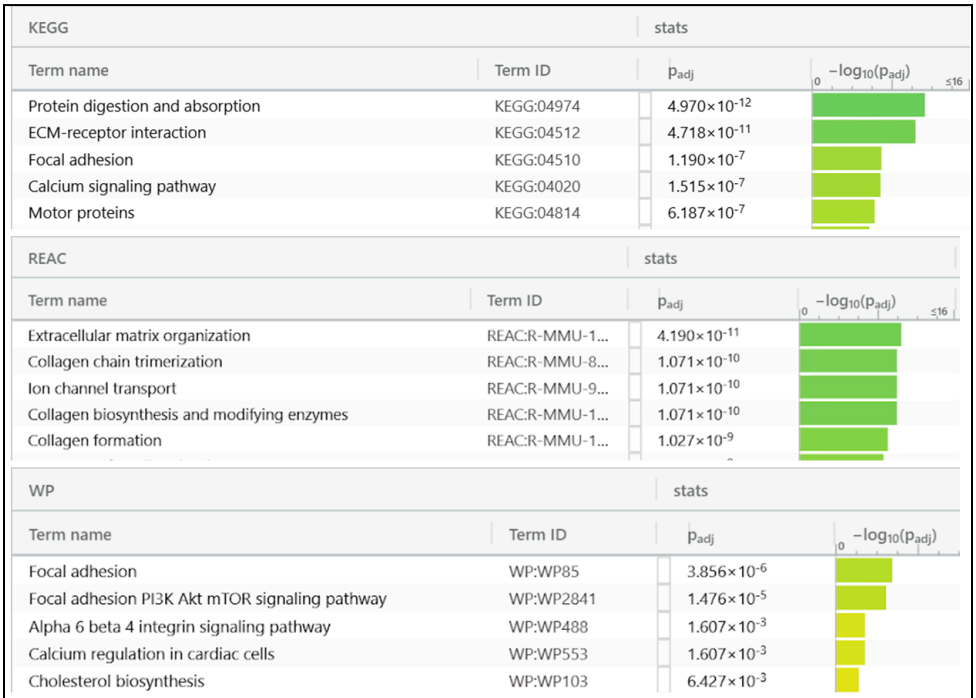


KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Cell cycle	KEGG:04110	1.027×10^{-15}	
Ribosome	KEGG:03010	8.376×10^{-15}	
DNA replication	KEGG:03030	1.119×10^{-13}	
Nucleocytoplasmic transport	KEGG:03013	1.939×10^{-13}	
Spliceosome	KEGG:03040	5.520×10^{-10}	

REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Metabolism of RNA	REAC:R-MMU-8...	5.285×10^{-60}	
Cell Cycle, Mitotic	REAC:R-MMU-6...	8.269×10^{-38}	
Cell Cycle	REAC:R-MMU-1...	3.235×10^{-37}	
rRNA processing in the nucleus and cytosol	REAC:R-MMU-8...	1.256×10^{-34}	
Major pathway of rRNA processing in the nucleolus and cyt...	REAC:R-MMU-6...	1.256×10^{-34}	

WP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
mRNA processing	WP:WP310	1.255×10^{-28}	
Cytoplasmic ribosomal proteins	WP:WP163	5.300×10^{-17}	
Translation factors	WP:WP307	8.391×10^{-11}	
G1 to S cell cycle control	WP:WP413	5.430×10^{-9}	
Mechanisms associated with pluripotency	WP:WP1763	3.986×10^{-7}	

Alex's top gene set enrichment from g:profiler for upregulated genes comparing stem versus enterocyte



KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Protein digestion and absorption	KEGG:04974	4.970×10^{-12}	
ECM-receptor interaction	KEGG:04512	4.718×10^{-11}	
Focal adhesion	KEGG:04510	1.190×10^{-7}	
Calcium signaling pathway	KEGG:04020	1.515×10^{-7}	
Motor proteins	KEGG:04814	6.187×10^{-7}	

REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Extracellular matrix organization	REAC:R-MMU-1...	4.190×10^{-11}	
Collagen chain trimerization	REAC:R-MMU-8...	1.071×10^{-10}	
Ion channel transport	REAC:R-MMU-9...	1.071×10^{-10}	
Collagen biosynthesis and modifying enzymes	REAC:R-MMU-1...	1.071×10^{-10}	
Collagen formation	REAC:R-MMU-1...	1.027×10^{-9}	

WP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Focal adhesion	WP:WP85	3.856×10^{-6}	
Focal adhesion PI3K Akt mTOR signaling pathway	WP:WP2841	1.476×10^{-5}	
Alpha 6 beta 4 integrin signaling pathway	WP:WP488	1.607×10^{-3}	
Calcium regulation in cardiac cells	WP:WP553	1.607×10^{-3}	
Cholesterol biosynthesis	WP:WP103	6.427×10^{-3}	

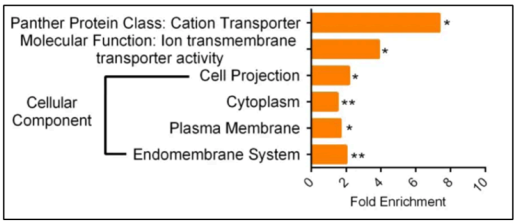
Alex's top gene set enrichment from g:profiler for downregulated genes comparing stem versus enterocyte

KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Cell cycle	KEGG:04110	1.027×10^{-15}	
Ribosome	KEGG:03010	8.376×10^{-15}	
DNA replication	KEGG:03030	1.119×10^{-13}	
Nucleocytoplasmic transport	KEGG:03013	1.939×10^{-13}	
Spliceosome	KEGG:03040	5.520×10^{-10}	
REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Metabolism of RNA	REAC:R-MMU...	5.285×10^{-60}	
Cell Cycle, Mitotic	REAC:R-MMU...	8.269×10^{-38}	
Cell Cycle	REAC:R-MMU...	3.235×10^{-37}	
rRNA processing in the nucleus and cytosol	REAC:R-MMU...	1.256×10^{-34}	
Major pathway of rRNA processing in the nucleolus and...	REAC:R-MMU...	1.256×10^{-34}	
WP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
mRNA processing	WP:WP310	1.255×10^{-28}	
Cytoplasmic ribosomal proteins	WP:WP163	5.300×10^{-17}	
Translation factors	WP:WP307	8.391×10^{-11}	
G1 to S cell cycle control	WP:WP413	5.430×10^{-9}	
Mechanisms associated with pluripotency	WP:WP1763	3.986×10^{-7}	

Kate's top 5 enriched gene sets from g:profiler for upregulated genes with default parameters.

KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Protein digestion and absorption	KEGG:04974	4.970×10^{-12}	
ECM-receptor interaction	KEGG:04512	4.718×10^{-11}	
Focal adhesion	KEGG:04510	1.190×10^{-7}	
Calcium signaling pathway	KEGG:04020	1.515×10^{-7}	
Motor proteins	KEGG:04814	6.187×10^{-7}	
REAC		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Extracellular matrix organization	REAC:R-MMU...	4.190×10^{-11}	
Collagen chain trimerization	REAC:R-MMU...	1.071×10^{-10}	
Ion channel transport	REAC:R-MMU...	1.071×10^{-10}	
Collagen biosynthesis and modifying enzymes	REAC:R-MMU...	1.071×10^{-10}	
Collagen formation	REAC:R-MMU...	1.027×10^{-9}	
WP		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Focal adhesion	WP:WP85	3.856×10^{-6}	
Focal adhesion PI3K Akt mTOR signaling pathway	WP:WP2841	1.476×10^{-5}	
Alpha 6 beta 4 integrin signaling pathway	WP:WP488	1.607×10^{-3}	
Calcium regulation in cardiac cells	WP:WP553	1.607×10^{-3}	
Cholesterol biosynthesis	WP:WP103	6.427×10^{-3}	

Kate's top 5 enriched gene sets from g:profiler for downregulated genes with default parameters.



	Mus musculus (REF)		Client Text Box Input (▽ Hierarchy NEW!)				
PANTHER Protein Class	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
transporter	1044	15	4.88	3.08	+	9.95E-05	1.96E-02

Analysis Type:	PANTHER Overrepresentation Test (Released 20240807)							
Annotation Version and Release Date:	PANTHER version 19.0 Released 2024-06-20							
Analyzed List:	Client Text Box Input (Mus musculus)							
Reference List:	Mus musculus (all genes in database)							
Test Type:	FISHER							
Correction:	FDR							
PANTHER GO-Slim Molecular Function	Mus musc	Client Text	Client Text	Client Text	Client Text	Client Text	Client Text	Client Text Box Input (FDR)
transition metal ion binding (GO:0046914)	93	4	0.43	+	9.21	9.53E-04	1.92E-01	
inorganic anion transmembrane transporter activity (GO:0015103)	94	3	0.44	+	6.83	9.76E-03	2.46E-01	
channel activity (GO:0015267)	385	6	1.8	+	3.34	9.43E-03	2.48E-01	

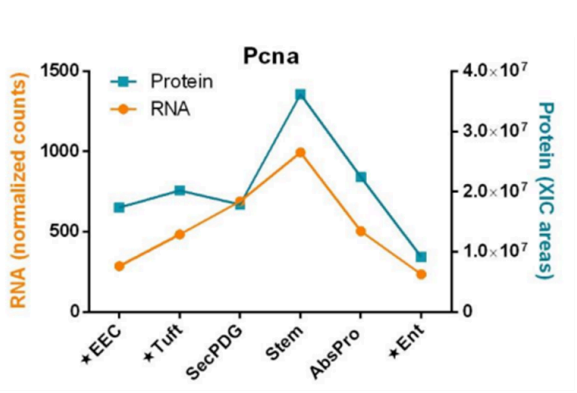
Displaying all results; click here to display only significant results

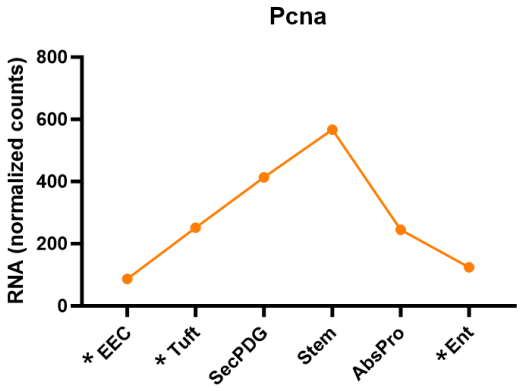
	Mus musculus (REF)	Client Text Box Input (Hierarchy) NEW!					
PANTHER GO-Slim Cellular Component	#	#	expected	Fold Enrichment	+/-	raw P value	▲ FDR
membrane	4688	39	21.90	1.78	+	1.38E-04	6.73E-02
exocytic vesicle	98	4	.46	8.74	+	1.16E-03	1.42E-01
secretory vesicle	166	5	.78	6.45	+	1.10E-03	1.79E-01
neuron projection	385	7	1.80	3.89	+	2.21E-03	1.80E-01
cytoplasm	5502	40	25.70	1.56	+	1.86E-03	1.82E-01
Golgi membrane	88	4	.41	9.73	+	7.75E-04	1.89E-01
transport vesicle	128	4	.60	6.69	+	3.07E-03	2.15E-01
endomembrane system	1426	15	6.66	2.25	+	3.53E-03	2.16E-01
plasma membrane bounded cell projection	667	9	3.12	2.89	+	4.03E-03	2.19E-01
cell projection	690	9	3.22	2.79	+	5.02E-03	2.45E-01

Left: Panther classification from the paper.
Right: Our Panther analysis for Protein Class, Molecular Function, and Cellular Component

Validation of functional analysis: To further validate that our pathway enrichment results are consistent with what was published in the paper we reproduced two figures from the paper related to specific pathways that appeared in our g:profiler results. The first is based on the upregulation of cell cycle genes in stem cells when compared to enterocytes. The authors of the paper stated, *"RNA markers of proliferation (Mki67, Pcna, and Mcms) are highest in stem cells, but interestingly, their protein products are readily detectable in differentiated cells, thus highlighting inconsistencies between mRNA and protein biomarkers of proliferation (Supplementary Fig. 18)."* Based on this text, we pulled expression data from DESeq2 for the gene *Pcna* as well as significance for Pcna comparing stem to each of the other cell types. The significance data for stem versus enterocyte is the same data that was used in our pathway enrichment analysis and the comparisons to other cell types was done in an identical manner to the enterocyte comparison. If our data is matching the paper we should be able to reproduce the figure as well as the significance. The figure below shows that we achieved both of these goals indicating that our data does match pathway analysis from the paper with significance defined as padj < 0.01 and > 50 mean normalized counts. It is important to note that the scaling on the y-axis for our version of the figure is different from the original. We suspect this is due to our handling of technical replicates in our analysis compared to the paper (first discussed in the PCA section of the report). In our analysis we only kept one technical replicate from the RNAseq runs because we were unsure as to how the authors handled these replicates - nowhere in their text do they indicate how technical replicates were addressed. However, if the authors merged these replicates together before downstream analysis this could lead to the overall larger

normalized counts as we see when comparing their figure to ours. Despite this our figure still recapitulates the original with good accuracy.

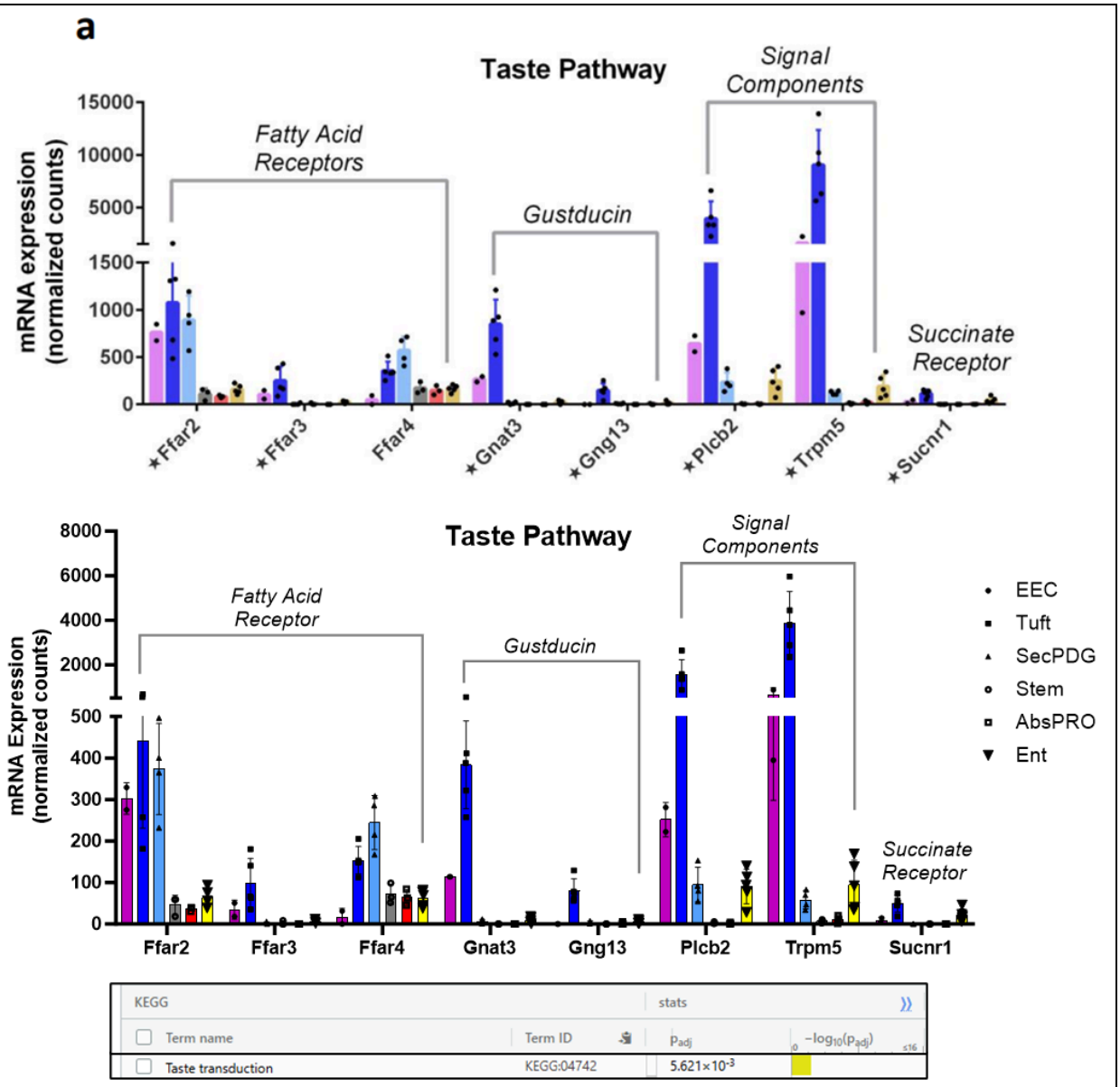




gene_name	comparison	log2FoldChange	padj	significant	regulation	
PCNA	Stem_vs AbsPro	1.212	0.02188	Not Significant		
PCNA	Stem_vs EEC	2.633	3.41E-09	Significant	Upregulated	*
PCNA	Stem_vs Enterocyte	2.182	1.71E-10	Significant	Upregulated	*
PCNA	Stem_vs SecPDG	0.454	0.38229	Not Significant		
PCNA	Stem_vs Tuft	1.168	0.00126	Significant	Upregulated	*

Top left: Original figure from the paper (Supplementary Figure 18). Top Right: Our reproduction of the figure showing normalized counts for Pcna, stars on x-axis indicate significance when comparing stem cells to each other cell type.
Bottom: Summary table of values generated comparing stem to each other cell type that were used in functional analysis for Pcna. Stars match significance in our graphs

The second analysis we did to validate our pathway enrichment data was done comparing stem cells to tuft cells. This comparison is different from our main comparison of stem versus enterocyte, however we chose this because the authors indicated that the Taste pathway is enriched in tuft cells when compared to stem cells. We did this comparison of stem versus tuft in the same manner as our main comparison of stem versus enterocyte with significance for the genes included in the pathway analysis as padj < 0.01 and >50 mean normalized counts. We were able to reproduce the figure with good accuracy - the only exception is that the scaling on the y-axis is different as discussed above. In conjunction with reproducing the counts for the Taste pathway we also used g:profiler to show that our data comparing stem to tuft cells does show enrichment for the taste pathway for downregulated genes as expected. This comparison was done in an identical manner as to our main comparison of stem versus enterocyte. The results indicate that our pathway enrichment for the main comparison is sound.



Top: original figure from the paper (supplementary figure 10).
Middle: Our recreation of the figure for the taste pathway.
Bottom: Snapshot of KEGG pathway enrichment comparing stem to tuft cells showing that the taste pathway is enriched with our data.

Sliding Window Parameter Change. As with our other analyses in this report, increasing the sliding window size from 1 to 2 showed minimal effects on g:profiler results. When comparing stem versus enterocytes at default parameters, most enriched KEGG pathways that were upregulated had some relation to the cell cycle or protein synthesis such as DNA replication, cell cycle, and ribosome. The increased window slightly changed the order of the most enriched

pathways. With default parameters, cell cycle is the top hit followed by ribosome for KEGG. However, with the increased window this is flipped and ribosome is the top hit followed by cell cycle. These are minor changes and the overall interpretation of the data is similar. This indicates that despite the change in trimming, top pathways are still being detected accurately.

KEGG

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
Ribosome	KEGG:03010	2.457×10^{-15}	
Cell cycle	KEGG:04110	1.180×10^{-14}	
Nucleocytoplasmic transport	KEGG:03013	3.376×10^{-14}	
DNA replication	KEGG:03030	7.315×10^{-14}	
Spliceosome	KEGG:03040	4.321×10^{-10}	

REAC

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
Metabolism of RNA	REAC:R-MMU-8...	1.522×10^{-61}	
Cell Cycle, Mitotic	REAC:R-MMU-6...	5.792×10^{-36}	
Cell Cycle	REAC:R-MMU-1...	5.792×10^{-36}	
rRNA processing	REAC:R-MMU-7...	1.415×10^{-35}	
Major pathway of rRNA processing in the nucleolus and cyt...	REAC:R-MMU-6...	1.415×10^{-35}	

WP

Term name	Term ID	p_{adj}	stats: $-\log_{10}(p_{adj})$ (0 – ≤16)
mRNA processing	WP:WP310	5.205×10^{-28}	
Cytoplasmic ribosomal proteins	WP:WP163	5.912×10^{-19}	
Translation factors	WP:WP307	8.391×10^{-11}	
G1 to S cell cycle control	WP:WP413	8.297×10^{-10}	
Mechanisms associated with pluripotency	WP:WP1763	1.587×10^{-7}	

Alex's top gene set enrichment from g:profiler for upregulated genes comparing stem versus enterocyte with sliding window changed to 2

KEGG		stats
Term name	Term ID	p_{adj}
Protein digestion and absorption	KEGG:04974	8.199×10^{-9}
Calcium signaling pathway	KEGG:04020	9.293×10^{-8}
ECM-receptor interaction	KEGG:04512	9.293×10^{-8}
Cholinergic synapse	KEGG:04725	1.236×10^{-7}
Circadian entrainment	KEGG:04713	7.347×10^{-7}

REAC		stats
Term name	Term ID	p_{adj}
Ion channel transport	REAC:R-MMU-9...	5.004×10^{-11}
Transport of small molecules	REAC:R-MMU-3...	2.382×10^{-8}
Collagen chain trimerization	REAC:R-MMU-8...	6.372×10^{-8}
Extracellular matrix organization	REAC:R-MMU-1...	1.566×10^{-7}
Collagen biosynthesis and modifying enzymes	REAC:R-MMU-1...	1.829×10^{-7}

WP		stats
Term name	Term ID	p_{adj}
Focal adhesion	WP:WP85	2.797×10^{-5}
Focal adhesion PI3K Akt mTOR signaling pathway	WP:WP2841	2.531×10^{-4}
Calcium regulation in cardiac cells	WP:WP553	4.940×10^{-4}
Alpha 6 beta 4 integrin signaling pathway	WP:WP488	5.609×10^{-4}
Prostaglandin synthesis and regulation	WP:WP374	3.416×10^{-3}

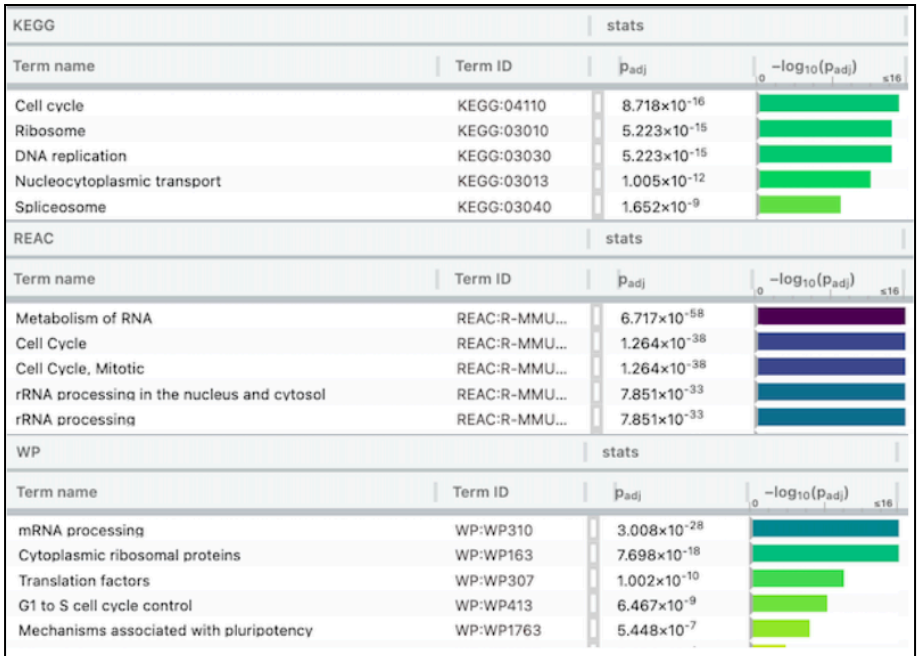
Alex's top gene set enrichment from g:profiler for downregulated genes comparing stem versus enterocyte with sliding window changed to 2

KEGG		stats
Term name	Term ID	p_{adj}
Cell cycle	KEGG:04110	1.027×10^{-15}
Ribosome	KEGG:03010	8.376×10^{-15}
DNA replication	KEGG:03030	1.119×10^{-13}
Nucleocytoplasmic transport	KEGG:03013	1.939×10^{-13}
Spliceosome	KEGG:03040	5.520×10^{-10}
mRNA surveillance pathway	KEGG:03015	4.283×10^{-9}
Ribosome biogenesis in eukaryotes	KEGG:03008	2.411×10^{-8}
Nucleotide excision repair	KEGG:03420	6.486×10^{-7}
Mismatch repair	KEGG:03430	6.486×10^{-7}
Coronavirus disease - COVID-19	KEGG:05171	1.603×10^{-6}
Viral carcinogenesis	KEGG:05203	1.603×10^{-6}
ATP-dependent chromatin remodeling	KEGG:03082	1.641×10^{-6}
Systemic lupus erythematosus	KEGG:05322	2.246×10^{-6}
Base excision repair	KEGG:03410	2.823×10^{-6}
Alcoholism	KEGG:05034	4.001×10^{-6}
Neutrophil extracellular trap formation	KEGG:04613	1.799×10^{-5}
Aminoacyl-tRNA biosynthesis	KEGG:00970	1.717×10^{-4}
Progesterone-mediated oocyte maturation	KEGG:04914	1.224×10^{-3}
Oocyte meiosis	KEGG:04114	2.673×10^{-3}
Glutathione metabolism	KEGG:00480	4.381×10^{-3}
Polycomb repressive complex	KEGG:03083	8.753×10^{-3}
One carbon pool by folate	KEGG:00670	8.939×10^{-3}
Biosynthesis of amino acids	KEGG:01230	9.829×10^{-3}
Cysteine and methionine metabolism	KEGG:00270	1.976×10^{-2}
Hepatocellular carcinoma	KEGG:05225	1.976×10^{-2}
Cellular senescence	KEGG:04218	2.011×10^{-2}
Transcriptional misregulation in cancer	KEGG:05202	2.219×10^{-2}
Carbon metabolism	KEGG:01200	2.870×10^{-2}
Drug metabolism - other enzymes	KEGG:00983	3.726×10^{-2}
RNA polymerase	KEGG:03020	4.546×10^{-2}
Amyotrophic lateral sclerosis	KEGG:05014	4.651×10^{-2}

KEGG		stats
Term name	Term ID	p_{adj}
Ribosome	KEGG:03010	2.457×10^{-15}
Cell cycle	KEGG:04110	1.180×10^{-14}
Nucleocytoplasmic transport	KEGG:03013	3.376×10^{-14}
DNA replication	KEGG:03030	7.315×10^{-14}
Spliceosome	KEGG:03040	4.321×10^{-10}
Ribosome biogenesis in eukaryotes	KEGG:03008	2.379×10^{-8}
mRNA surveillance pathway	KEGG:03015	7.447×10^{-8}
Nucleotide excision repair	KEGG:03420	6.019×10^{-7}
Mismatch repair	KEGG:03430	6.019×10^{-7}
Coronavirus disease - COVID-19	KEGG:05171	1.447×10^{-6}
ATP-dependent chromatin remodeling	KEGG:03082	1.538×10^{-6}
Systemic lupus erythematosus	KEGG:05322	6.969×10^{-6}
Base excision repair	KEGG:03410	1.698×10^{-5}
Viral carcinogenesis	KEGG:05203	2.103×10^{-5}
Alcoholism	KEGG:05034	2.640×10^{-5}
Neutrophil extracellular trap formation	KEGG:04613	1.070×10^{-4}
Aminoacyl-tRNA biosynthesis	KEGG:00970	1.591×10^{-4}
Polycomb repressive complex	KEGG:03083	9.537×10^{-4}
Biosynthesis of amino acids	KEGG:01230	3.560×10^{-3}
Glutathione metabolism	KEGG:00480	4.063×10^{-3}
Cysteine and methionine metabolism	KEGG:00270	6.323×10^{-3}
One carbon pool by folate	KEGG:00670	8.201×10^{-3}
Progesterone-mediated oocyte maturation	KEGG:04914	8.201×10^{-3}
RNA polymerase	KEGG:03020	1.242×10^{-2}
Oocyte meiosis	KEGG:04114	1.242×10^{-2}
Amyotrophic lateral sclerosis	KEGG:05014	1.641×10^{-2}
Hepatocellular carcinoma	KEGG:05225	1.641×10^{-2}
Transcriptional misregulation in cancer	KEGG:05202	1.930×10^{-2}
Drug metabolism - other enzymes	KEGG:00983	3.372×10^{-2}
Cellular senescence	KEGG:04218	3.372×10^{-2}
Biosynthesis of cofactors	KEGG:01240	4.245×10^{-2}

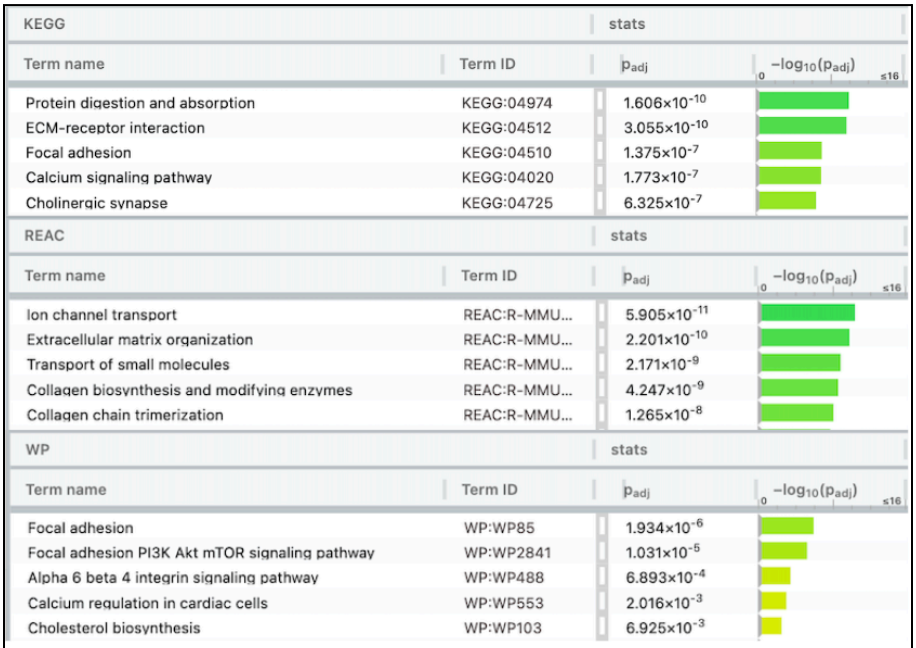
Left: KEGG g:profiler results for upregulated genes in stem versus enterocyte comparison with default parameters. Right: KEEG g:profiler results with increased sliding window size to 2.

Max Read Length Parameter Change. Modifying the maximum read length by capping at 85 bp slightly changed our g:profiler data. Within each of the gene groupings (KEGG, Reactome, or WikiPathways), the main pathways did not change, but the Padj score tended to shift slightly when max read length was modified. However, the Padj tenth-power values were the same, so the overall significance was not largely impacted. Modifying max read length increased our confidence in our initial results because the main pathways were consistently up or downregulated.



KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Cell cycle	KEGG:04110	8.718×10^{-16}	
Ribosome	KEGG:03010	5.223×10^{-15}	
DNA replication	KEGG:03030	5.223×10^{-15}	
Nucleocytoplasmic transport	KEGG:03013	1.005×10^{-12}	
Spliceosome	KEGG:03040	1.652×10^{-9}	
REAC		**stats**	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Metabolism of RNA	REAC:R-MMU...	6.717×10^{-58}	
Cell Cycle	REAC:R-MMU...	1.264×10^{-38}	
Cell Cycle, Mitotic	REAC:R-MMU...	1.264×10^{-38}	
rRNA processing in the nucleus and cytosol	REAC:R-MMU...	7.851×10^{-33}	
rRNA processing	REAC:R-MMU...	7.851×10^{-33}	
WP		**stats**	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
mRNA processing	WP:WP310	3.008×10^{-28}	
Cytoplasmic ribosomal proteins	WP:WP163	7.698×10^{-18}	
Translation factors	WP:WP307	1.002×10^{-10}	
G1 to S cell cycle control	WP:WP413	6.467×10^{-9}	
Mechanisms associated with pluripotency	WP:WP1763	5.448×10^{-7}	

Top 5 g:profiler enriched gene sets for upregulated genes with modified max read length.



KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Protein digestion and absorption	KEGG:04974	1.606×10^{-10}	
ECM-receptor interaction	KEGG:04512	3.055×10^{-10}	
Focal adhesion	KEGG:04510	1.375×10^{-7}	
Calcium signaling pathway	KEGG:04020	1.773×10^{-7}	
Cholinergic synapse	KEGG:04725	6.325×10^{-7}	
REAC		**stats**	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Ion channel transport	REAC:R-MMU...	5.905×10^{-11}	
Extracellular matrix organization	REAC:R-MMU...	2.201×10^{-10}	
Transport of small molecules	REAC:R-MMU...	2.171×10^{-9}	
Collagen biosynthesis and modifying enzymes	REAC:R-MMU...	4.247×10^{-9}	
Collagen chain trimerization	REAC:R-MMU...	1.265×10^{-8}	
WP		**stats**	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Focal adhesion	WP:WP85	1.934×10^{-6}	
Focal adhesion PI3K Akt mTOR signaling pathway	WP:WP2841	1.031×10^{-5}	
Alpha 6 beta 4 integrin signaling pathway	WP:WP488	6.893×10^{-4}	
Calcium regulation in cardiac cells	WP:WP553	2.016×10^{-3}	
Cholesterol biosynthesis	WP:WP103	6.925×10^{-3}	

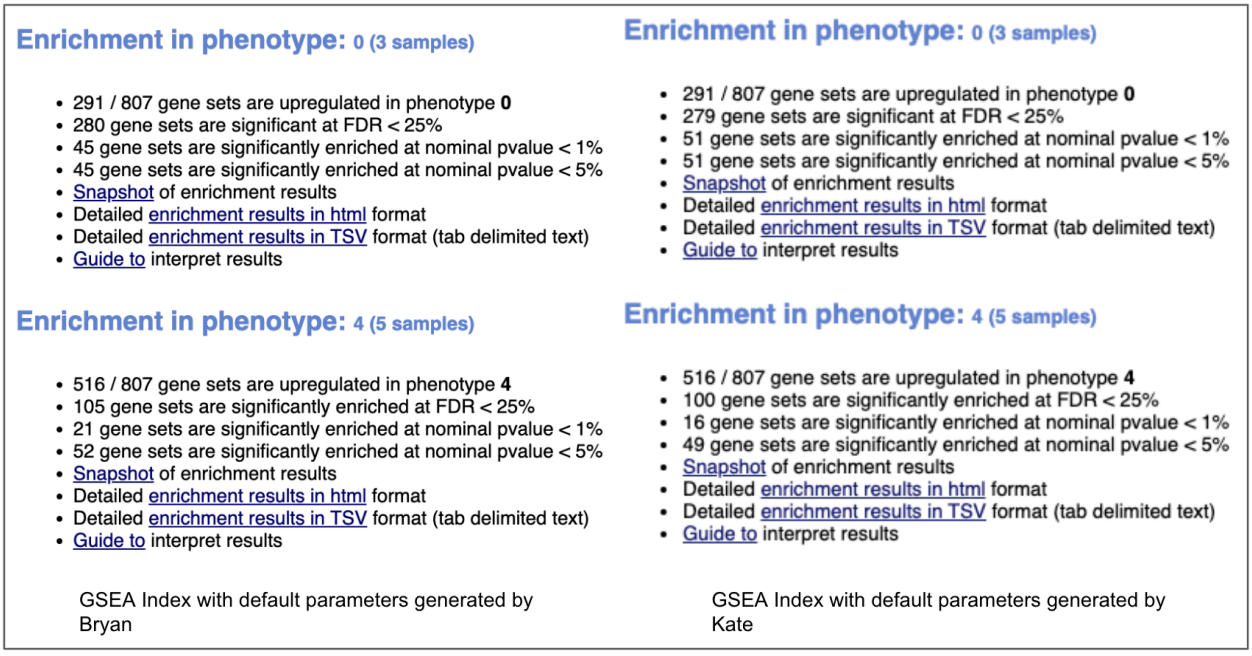
Top 5 g:profiler enriched gene sets for downregulated genes with modified max read length.


Results from gprofiler for downregulated genes with modified max read length.
WP
stats
Term name
Term ID
Focal adhesion
WP:WP85
Focal adhesion PI3K Akt mTOR signaling pathway
WP:WP2841
Alpha 6 beta 4 integrin signaling pathway
WP:WP488
Calcium regulation in cardiac cells
WP:WP553
Cholesterol biosynthesis
WP:WP103
Prostaglandin synthesis and regulation
WP:WP374
Alzheimer 39 s disease
WP:WP2075
Myometrial relaxation and contraction pathways
WP:WP385
Splicing factor NOVA regulated synaptic proteins
WP:WP1983


Results from gprofiler for upregulated genes with modified max read length.
KEGG
stats
Term name
Term ID
Cell cycle
KEGG:04110
Ribosome
KEGG:03010
DNA replication
KEGG:03030
Nucleocytoplasmic transport
KEGG:03013
Spliceosome
KEGG:03040
mRNA surveillance pathway
KEGG:03015
Mismatch repair
KEGG:03430
Ribosome biogenesis in eukaryotes
KEGG:03008
Nucleotide excision repair
KEGG:03420
ATP-dependent chromatin remodeling
KEGG:03082
Viral carcinogenesis
KEGG:05203
Coronavirus disease - COVID-19
KEGG:05171
Systemic lupus erythematosus
KEGG:05322
Base excision repair
KEGG:03410
Alcoholism
KEGG:05034
Neutrophil extracellular trap formation
KEGG:04613
Aminoacyl-tRNA biosynthesis
KEGG:00970
Progesterone-mediated oocyte maturation
KEGG:04914
Oocyte meiosis
KEGG:04114
Glutathione metabolism
KEGG:00480
Prion disease
KEGG:05020
Pyrimidine metabolism
KEGG:00240
One carbon pool by folate
KEGG:00670
Biosynthesis of amino acids
KEGG:01230
RNA polymerase
KEGG:03020
Hepatocellular carcinoma
KEGG:05225
Cellular senescence
KEGG:04218
Homologous recombination
KEGG:03440
Amyotrophic lateral sclerosis
KEGG:05014
Drug metabolism - other enzymes
KEGG:00983
Transcriptional misregulation in cancer
KEGG:05202
Cysteine and methionine metabolism
KEGG:00270
WP
stats
Term name
Term ID
Show evidence codes
mRNA processing
WP:WP310
Cytoplasmic ribosomal proteins
WP:WP163
Translation factors
WP:WP307
G1 to S cell cycle control
WP:WP413
Mechanisms associated with pluripotency
WP:WP1763
Mismatch repair
WP:WP1257
Amino acid metabolism
WP:WP662
Homologous recombination
WP:WP1258
mir 193a and MVP in colon cancer metastasis
WP:WP3979

GSEA

There was no direct comparison we could make to the reference paper for GSEA, however, we were able to separately generate two GSEA runs with default parameters that resemble each other as much as possible. Both runs showed 807 enriched gene sets: 291 upregulated in stem cells, and 516 upregulated in enterocytes. The ranked lists of enriched gene sets in enterocytes were very similar, with the top ranked gene set in both runs being CLASS I PEROXISOMAL MEMBRANE PROTEIN IMPORT. The ranked lists of enriched gene sets in stem cells showed far less consistency between runs. It is important to note that GSEA uses permutation testing to generate p-values and normalized enrichment scores (NES). If a seed is not set for permutation testing, each run of GSEA will likely run different permutations, slightly altering the final results. This fact explains the minor differences between runs. The greater discrepancy between the ranked lists of enriched gene sets in stem cells can be attributed to stem cells' relatively low RNA count. Lower sample size, where sample size in this case is RNA count, causes weaker statistical power and therefore greater variance between different runs of permutation testing within GSEA.



Enrichment in phenotype: 0 (3 samples)

- 291 / 807 gene sets are upregulated in phenotype **0**
- 280 gene sets are significant at FDR < 25%
- 45 gene sets are significantly enriched at nominal pvalue < 1%
- 45 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 516 / 807 gene sets are upregulated in phenotype **4**
- 105 gene sets are significantly enriched at FDR < 25%
- 21 gene sets are significantly enriched at nominal pvalue < 1%
- 52 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA Index with default parameters generated by Bryan

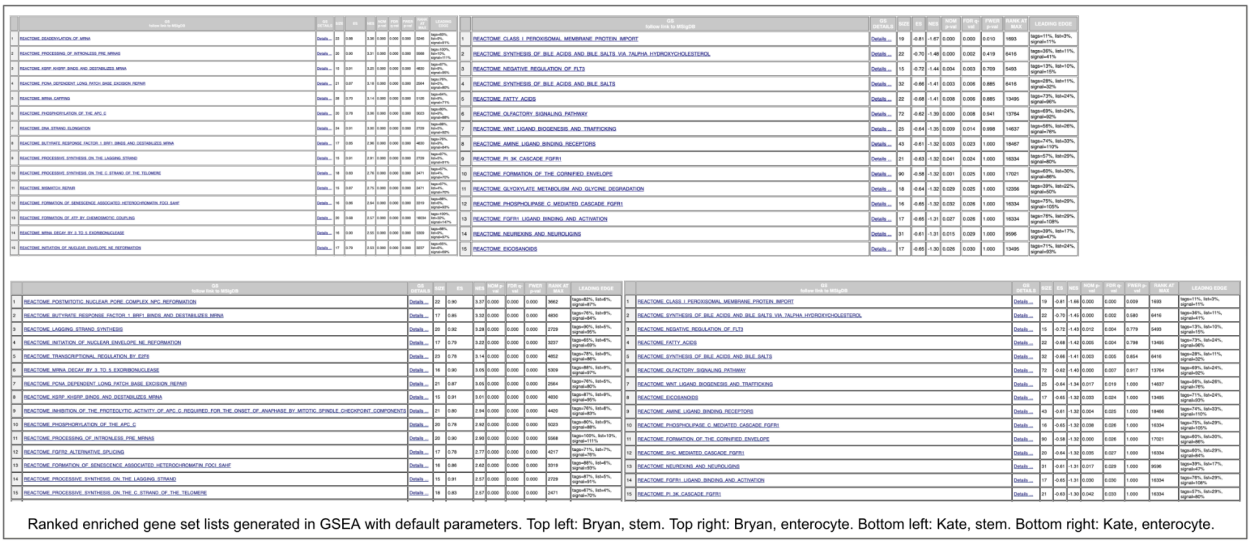
Enrichment in phenotype: 0 (3 samples)

- 291 / 807 gene sets are upregulated in phenotype **0**
- 279 gene sets are significant at FDR < 25%
- 51 gene sets are significantly enriched at nominal pvalue < 1%
- 51 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 516 / 807 gene sets are upregulated in phenotype **4**
- 100 gene sets are significantly enriched at FDR < 25%
- 16 gene sets are significantly enriched at nominal pvalue < 1%
- 49 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA Index with default parameters generated by Kate



Ranked enriched gene set lists generated in GSEA with default parameters. Top left: Bryan, stem. Top right: Bryan, enterocyte. Bottom left: Kate, stem. Bottom right: Kate, enterocyte.

Min PHRED Score Parameter Change. Increasing the minimum PHRED score from 20 to 35 exhibited moderate impacts on both GSEA generated index files and ranked enriched gene set lists. The number of gene sets enriched in stem cells decreased from 291 to 287, while the number of significantly enriched gene sets in stem cells increased from 45 to 53. The number of gene sets enriched in enterocytes increased from 516 to 520, while the number of significantly enriched gene sets in enterocytes remained about the same. The changes across runs were relatively small due to the high quality of the data, and only a small fraction of reads being filtered out by the increased PHRED threshold. The decrease in enriched gene sets in stem cells and increase in enterocytes can be attributed to each cell's RNA content. The low RNA content of stem cells leads to more low quality associated reads that get filtered out by the stricter PHRED threshold, while the high RNA content of enterocytes benefits from the reduced noise after stricter filtering. This filtering of lower quality reads associated with stem cells explains the increase in significantly enriched gene sets in stem cells. Further, results showed consistency between ranked enriched gene set lists for enterocytes, and relatively high discrepancy between ranked enriched gene set lists for stem cells. The top two enriched gene sets in enterocytes remained the same. These were: CLASS I PEROXISOMAL MEMBRANE PROTEIN IMPORT and SYNTHESIS OF BILE ACIDS AND BILE SALTS VIA 7ALPHA HYDROXYCHOLESTEROL. These results can also be attributed to the RNA content of stem cells and enterocytes, and the discarding of more stem cell associated reads because of their

relatively low RNA count compared to enterocytes.

Enrichment in phenotype: 0 (3 samples)

- 291 / 807 gene sets are upregulated in phenotype **0**
- 280 gene sets are significant at FDR < 25%
- 45 gene sets are significantly enriched at nominal pvalue < 1%
- 45 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 516 / 807 gene sets are upregulated in phenotype **4**
- 105 gene sets are significantly enriched at FDR < 25%
- 21 gene sets are significantly enriched at nominal pvalue < 1%
- 52 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA Index with default parameters

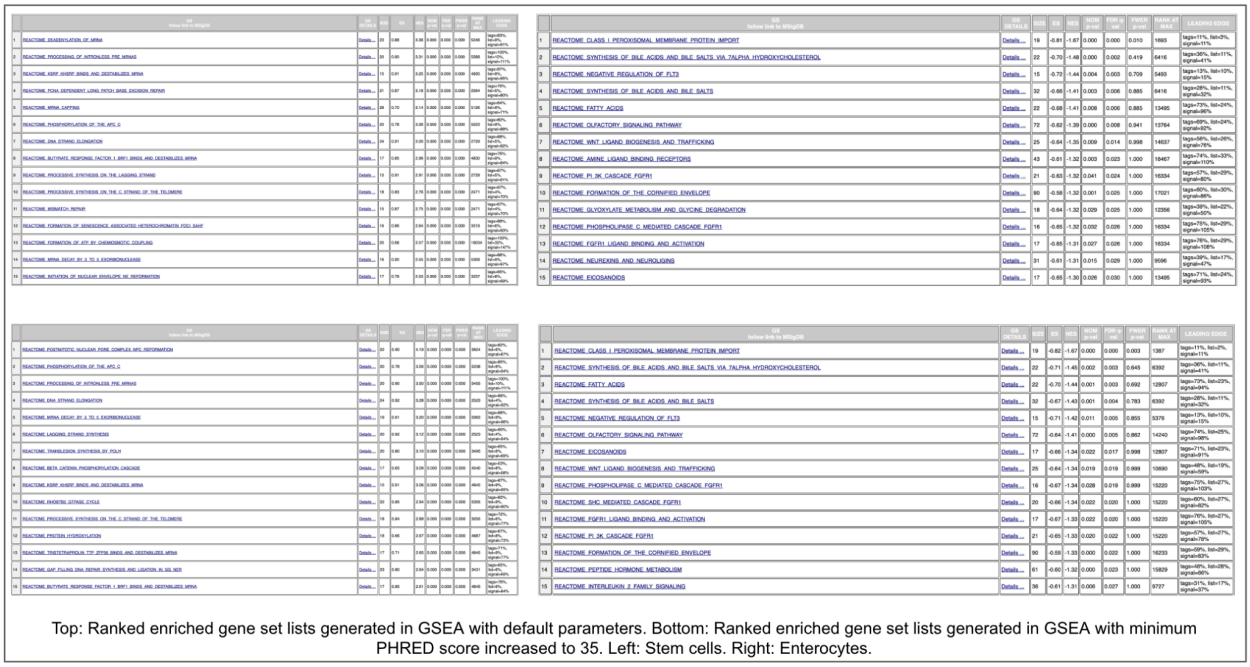
Enrichment in phenotype: 0 (3 samples)

- 287 / 807 gene sets are upregulated in phenotype **0**
- 277 gene sets are significant at FDR < 25%
- 53 gene sets are significantly enriched at nominal pvalue < 1%
- 53 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 520 / 807 gene sets are upregulated in phenotype **4**
- 118 gene sets are significantly enriched at FDR < 25%
- 19 gene sets are significantly enriched at nominal pvalue < 1%
- 52 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA Index with minimum PHRED score set to 35



	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_DEADENYLATION_OF_MRNA	Details ...	23	0.88	3.06	0.000	0.000	0.000	5348	tags=65%, list=9%, signal=71%
2	REACTOME_PROCESSING_OF_INTRONLESS_PRE_MRNAS	Details ...	20	0.90	3.01	0.000	0.000	0.000	5966	tags=75%, list=11%, signal=79%
3	REACTOME_KSRP_KHSRP_BINDS_AND_DESTABILIZES_MRNA	Details ...	15	0.91	3.00	0.000	0.000	0.000	4808	tags=67%, list=9%, signal=73%
4	REACTOME_POLH_DEPENDENT_LONG_PATCH_BASE_EXCISION_REPAIR	Details ...	21	0.87	3.19	0.000	0.000	0.000	6944	tags=67%, list=12%, signal=76%
5	REACTOME_MRNA_CAPPING	Details ...	28	0.70	3.14	0.000	0.000	0.000	5106	tags=46%, list=9%, signal=51%
6	REACTOME_PHOSPHORYLATION_OF_THE_APC_C	Details ...	20	0.78	3.08	0.000	0.000	0.000	5920	tags=60%, list=10%, signal=66%
7	REACTOME_DNA_STRAND_ELONGATION	Details ...	34	0.91	3.05	0.000	0.000	0.000	2702	tags=68%, list=5%, signal=72%
8	REACTOME_BUTYRATE_RESPONSE_FACTOR_1_BRF1_BINDS_AND_DESTABILIZES_MRNA	Details ...	17	0.85	2.99	0.000	0.000	0.000	4830	tags=76%, list=9%, signal=84%
9	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_LAGGING_STRAND	Details ...	15	0.91	2.91	0.000	0.000	0.000	2722	tags=67%, list=5%, signal=70%
10	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_C_STRAND_OF_THE_TELOMERE	Details ...	18	0.83	2.79	0.000	0.000	0.000	2972	tags=67%, list=5%, signal=70%
11	REACTOME_MISMATCH_REPAIR	Details ...	15	0.87	2.79	0.000	0.000	0.000	2972	tags=67%, list=5%, signal=70%
12	REACTOME_FORMATION_OF_SENESCENCE_ASSOCIATED_HETEROCHROMATIN_FOCI_SAHF	Details ...	16	0.86	2.84	0.000	0.000	0.000	5172	tags=69%, list=9%, signal=76%
13	REACTOME_FORMATION_OF_ATP_BY_CHEMIOSMOTIC_COUPLING	Details ...	20	0.68	2.57	0.000	0.000	0.000	19034	tags=100%, list=33%, signal=147%
14	REACTOME_MRNA_DECAY_BY_3_TO_5_EXORIBONUCLEASE	Details ...	16	0.80	2.55	0.000	0.000	0.000	5006	tags=88%, list=9%, signal=96%
15	REACTOME_INITIATION_OF_NUCLEAR_ENVELOPE_NE_REFORMATION	Details ...	17	0.79	2.65	0.000	0.000	0.000	5337	tags=65%, list=9%, signal=72%

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_CLASS_I_PEROXISOMAL_MEMBRANE_PROTEIN_IMPORT	Details ...	19	-0.81	-1.67	0.000	0.000	0.010	1693	tags=11%, list=3%, signal=11%
2	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	Details ...	22	-0.70	-1.48	0.000	0.002	0.419	6416	tags=36%, list=11%, signal=41%
3	REACTOME_NEGATIVE_REGULATION_OF_FLT3	Details ...	15	-0.72	-1.44	0.004	0.003	0.709	5493	tags=13%, list=10%, signal=15%
4	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	Details ...	32	-0.66	-1.41	0.003	0.006	0.885	6416	tags=28%, list=11%, signal=32%
5	REACTOME_FATTY_ACIDS	Details ...	22	-0.68	-1.41	0.008	0.006	0.885	13495	tags=73%, list=24%, signal=96%
6	REACTOME_OLFACTORY_SIGNALING_PATHWAY	Details ...	72	-0.62	-1.39	0.000	0.008	0.941	13764	tags=69%, list=24%, signal=92%
7	REACTOME_WNT_LIGAND_BIOGENESIS_AND_TRAFFICKING	Details ...	25	-0.64	-1.35	0.009	0.014	0.998	14637	tags=68%, list=26%, signal=76%
8	REACTOME_AMINE_LIGAND_BINDING_RECEPTORS	Details ...	43	-0.61	-1.32	0.003	0.023	1.000	18467	tags=74%, list=33%, signal=110%
9	REACTOME_PI_3K_CASCADE_FGFR1	Details ...	21	-0.63	-1.32	0.041	0.024	1.000	16334	tags=57%, list=29%, signal=60%
10	REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	Details ...	90	-0.58	-1.32	0.001	0.025	1.000	17021	tags=60%, list=30%, signal=88%
11	REACTOME_GLYOXYLATE_METABOLISM_AND_GLYCINE_DEGRADATION	Details ...	18	-0.64	-1.32	0.029	0.025	1.000	12956	tags=39%, list=22%, signal=50%
12	REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE_FGFR1	Details ...	16	-0.65	-1.32	0.032	0.026	1.000	16334	tags=75%, list=29%, signal=105%
13	REACTOME_FGFR1_LIGAND_BINDING_AND_ACTIVATION	Details ...	17	-0.65	-1.31	0.027	0.026	1.000	16334	tags=76%, list=29%, signal=108%
14	REACTOME_NEUREXINS_AND_NEUROLIGINS	Details ...	31	-0.61	-1.31	0.015	0.029	1.000	9596	tags=39%, list=17%, signal=47%
15	REACTOME_EICOSANOIDS	Details ...	17	-0.65	-1.30	0.026	0.030	1.000	13495	tags=71%, list=24%, signal=93%

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_POSTMITOTIC_NUCLEAR_PORE_COMPLEX_NPC_REFORMATION	Details ...	22	0.92	3.18	0.000	0.000	0.000	3804	tags=45%, list=6%, signal=48%
2	REACTOME_PHOSPHORYLATION_OF_THE_APC_C	Details ...	20	0.78	3.05	0.000	0.000	0.000	5238	tags=60%, list=9%, signal=65%
3	REACTOME_PROCESSING_OF_INTRONLESS_PRE_MRNAS	Details ...	20	0.90	3.00	0.000	0.000	0.000	5400	tags=70%, list=10%, signal=77%
4	REACTOME_DNA_STRAND_ELONGATION	Details ...	34	0.92	3.08	0.000	0.000	0.000	2522	tags=68%, list=5%, signal=72%
5	REACTOME_MRNA_DECAY_BY_3_TO_5_EXORIBONUCLEASE	Details ...	16	0.91	3.00	0.000	0.000	0.000	5966	tags=69%, list=11%, signal=77%
6	REACTOME_LAGGING_STRAND_SYNTHESIS	Details ...	26	0.92	3.10	0.000	0.000	0.000	2522	tags=46%, list=5%, signal=48%
7	REACTOME_TRANSLESION_SYNTHESIS_BY_POLH	Details ...	20	0.90	3.10	0.000	0.000	0.000	3440	tags=65%, list=6%, signal=69%
8	REACTOME_BETA_CATENIN_PHOSPHORYLATION_CASCADE	Details ...	17	0.90	3.09	0.000	0.000	0.000	4040	tags=76%, list=8%, signal=82%
9	REACTOME_KSRP_KHSRP_BINDS_AND_DESTABILIZES_MRNA	Details ...	15	0.91	3.00	0.000	0.000	0.000	4840	tags=67%, list=9%, signal=73%
10	REACTOME_RNASTBS_STRAND_CYCLE	Details ...	22	0.90	2.99	0.000	0.000	0.000	5006	tags=82%, list=9%, signal=90%
11	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_C_STRAND_OF_THE_TELOMERE	Details ...	18	0.84	2.89	0.000	0.000	0.000	3966	tags=67%, list=7%, signal=71%
12	REACTOME_PROTEIN_HYDROXYLATION	Details ...	18	0.90	2.97	0.000	0.000	0.000	4867	tags=78%, list=9%, signal=85%
13	REACTOME_TRETETRAPROLIN_TTP_ZFP36_BINDS_AND_DESTABILIZES_MRNA	Details ...	17	0.71	2.93	0.000	0.000	0.000	6648	tags=76%, list=12%, signal=87%
14	REACTOME_GAP_FILLING_DNA_REPAIR_SYNTHESIS_AND_LIGATION_IN_GG_NER	Details ...	33	0.80	2.94	0.000	0.000	0.000	3431	tags=61%, list=6%, signal=65%
15	REACTOME_BUTYRATE_RESPONSE_FACTOR_1_BRF1_BINDS_AND_DESTABILIZES_MRNA	Details ...	17	0.85	2.91	0.000	0.000	0.000	4840	tags=76%, list=9%, signal=84%

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_CLASS_I_PEROXISOMAL_MEMBRANE_PROTEIN_IMPORT	Details ...	19	-0.82	-1.67	0.000	0.000	0.003	1387	tags=11%, list=2%, signal=11%
2	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	Details ...	22	-0.71	-1.45	0.002	0.003	0.645	6392	tags=36%, list=11%, signal=41%
3	REACTOME_FATTY_ACIDS	Details ...	22	-0.70	-1.44	0.001	0.003	0.692	12807	tags=73%, list=23%, signal=94%
4	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	Details ...	32	-0.67	-1.43	0.001	0.004	0.783	6392	tags=28%, list=11%, signal=32%
5	REACTOME_NEGATIVE_REGULATION_OF_FLT3	Details ...	15	-0.71	-1.42	0.011	0.005	0.855	5376	tags=13%, list=10%, signal=15%
6	REACTOME_OLFACTORY_SIGNALING_PATHWAY	Details ...	72	-0.64	-1.41	0.000	0.005	0.862	14240	tags=74%, list=25%, signal=98%
7	REACTOME_EICOSANOIDS	Details ...	17	-0.66	-1.34	0.022	0.017	0.998	12807	tags=71%, list=23%, signal=91%
8	REACTOME_WNT_LIGAND_BIOGENESIS_AND_TRAFFICKING	Details ...	25	-0.64	-1.34	0.019	0.019	0.999	10680	tags=48%, list=19%, signal=59%
9	REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE_FGFR1	Details ...	16	-0.67	-1.34	0.028	0.019	0.999	15220	tags=75%, list=27%, signal=103%
10	REACTOME_SHC_MEDIATED_CASCADE_FGFR1	Details ...	20	-0.66	-1.34	0.022	0.020	1.000	15220	tags=60%, list=27%, signal=83%
11	REACTOME_FGFR1_LIGAND_BINDING_AND_ACTIVATION	Details ...	17	-0.67	-1.33	0.022	0.020	1.000	15220	tags=76%, list=27%, signal=100%
12	REACTOME_PI_3K_CASCADE_FGFR1	Details ...	21	-0.65	-1.33	0.020	0.022	1.000	15220	tags=67%, list=27%, signal=78%
13	REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	Details ...	90	-0.59	-1.33	0.000	0.022	1.000	16233	tags=59%, list=29%, signal=82%
14	REACTOME_PEPTIDE_HORMONE_METABOLISM	Details ...	61	-0.60	-1.32	0.000	0.023	1.000	15829	tags=48%, list=28%, signal=66%
15	REACTOME_INTERLEUKIN_2_FAMILY_SIGNALING	Details ...	36	-0.61	-1.31	0.008	0.027	1.000	9707	tags=31%, list=17%, signal=37%

Top: Ranked enriched gene set lists generated in GSEA with default parameters. Bottom: Ranked enriched gene set lists generated in GSEA with minimum PHRED score increased to 35. Left: Stem cells. Right: Enterocytes.

Max Read Length Parameter Change. Modifying max read length by capping at 85 bp slightly lowered the number of gene sets enriched in GSEA for stem cells (291 in default to 288), but increased in enterocytes (516 in default to 519). This is possibly because there is more mRNA in enterocytes so the natural robustness is higher than in stem cells, which have less mRNA.

Thus, increasing stringency can reduce the number of enriched gene sets in stem cells because the stem cells are more susceptible to changes in parameters. We would also expect some changes in GSEA after capping max read length because trimming some reads impacted alignment and thus expression levels, and these subtle changes can shift genes around on the rank list. GSEA is based on this rank list so even small changes may show up in the gene sets that were enriched. For example, in the stem cell GSEA data below, “postmitotic nuclear pore complex” is most enriched before and after modifying maximum read length, but the second most enriched gene set is different (“butyrate response factor” vs “PCNA dependent base excision repair”). Although the order changes, other enriched gene sets like “lagging strand synthesis” and “formation of senescence-associated heterochromatin foci” are in the top 15 for both GSEA results, which further confirms the raw data had high integrity.

GSEA results summary with default parameters (stem cell on top, enterocyte below)

Enrichment in phenotype: 0 (3 samples)

- 291 / 807 gene sets are upregulated in phenotype **0**
- 279 gene sets are significant at FDR < 25%
- 51 gene sets are significantly enriched at nominal pvalue < 1%
- 51 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 516 / 807 gene sets are upregulated in phenotype **4**
- 100 gene sets are significantly enriched at FDR < 25%
- 16 gene sets are significantly enriched at nominal pvalue < 1%
- 49 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA results summary with modified max read length (stem cell on top, enterocyte below)

Enrichment in phenotype: 0 (3 samples)

- 288 / 807 gene sets are upregulated in phenotype **0**
- 267 gene sets are significant at FDR < 25%
- 46 gene sets are significantly enriched at nominal pvalue < 1%
- 46 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: 4 (5 samples)

- 519 / 807 gene sets are upregulated in phenotype **4**
- 103 gene sets are significantly enriched at FDR < 25%
- 20 gene sets are significantly enriched at nominal pvalue < 1%
- 52 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

GSEA results for Stem Cell

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_POSTMITOTIC_NUCLEAR_PORE_COMPLEX_NPC_REFORMATION	Details ...	22	0.90	3.37	0.000	0.000	0.000	3662	tags=82%, list=6%, signal=87%
2	REACTOME_BUTYRATE_RESPONSE_FACTOR_1_BRF1_BINDS_AND_DESTABILIZES_MRNA	Details ...	17	0.85	3.32	0.000	0.000	0.000	4830	tags=76%, list=9%, signal=84%
3	REACTOME_LAGGING_STRAND_SYNTHESIS	Details ...	20	0.92	3.28	0.000	0.000	0.000	2729	tags=90%, list=5%, signal=95%
4	REACTOME_INITIATION_OF_NUCLEAR_ENVELOPE_NE_REFORMATION	Details ...	17	0.79	3.22	0.000	0.000	0.000	3237	tags=65%, list=6%, signal=69%
5	REACTOME_TRANSCRIPTIONAL_REGULATION_BY_E2F6	Details ...	23	0.78	3.14	0.000	0.000	0.000	4852	tags=78%, list=9%, signal=86%
6	REACTOME_MRNA_DECAY_BY_3_TO_5_EXORIBONUCLEASE	Details ...	16	0.90	3.05	0.000	0.000	0.000	5309	tags=88%, list=9%, signal=97%
7	REACTOME_PCNA_DEPENDENT_LONG_PATCH_BASE_EXCISION_REPAIR	Details ...	21	0.87	3.05	0.000	0.000	0.000	2564	tags=76%, list=5%, signal=80%
8	REACTOME_KSRP_KHSRP_BINDS_AND_DESTABILIZES_MRNA	Details ...	15	0.91	3.01	0.000	0.000	0.000	4830	tags=87%, list=9%, signal=95%
9	REACTOME_INHIBITION_OF_THE_PROTEOLYTIC_ACTIVITY_OF_APC_C_REQUIRED_FOR_THE_ONSET_OF_ANAPHASE_BY_MITOTIC_SPINDLE_CHECKPOINT_COMPONENTS	Details ...	21	0.80	2.94	0.000	0.000	0.000	4420	tags=76%, list=8%, signal=83%
10	REACTOME_PHOSPHORYLATION_OF_THE_APC_C	Details ...	20	0.78	2.92	0.000	0.000	0.000	5023	tags=80%, list=9%, signal=88%
11	REACTOME_PROCESSING_OF_INTRONLESS_PRE_MRNAS	Details ...	20	0.90	2.90	0.000	0.000	0.000	5568	tags=100%, list=10%, signal=111%
12	REACTOME_FGFR2_ALTERNATIVE_SPLICING	Details ...	17	0.78	2.77	0.000	0.000	0.000	4217	tags=71%, list=7%, signal=76%
13	REACTOME_FORMATION_OF_SENESCENCE_ASSOCIATED_HETEROCHROMATIN_FOCI_SAHF	Details ...	16	0.86	2.62	0.000	0.000	0.000	3319	tags=88%, list=6%, signal=93%
14	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_LAGGING_STRAND	Details ...	15	0.91	2.57	0.000	0.000	0.000	2729	tags=87%, list=5%, signal=91%
15	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_C_STRAND_OF_THE_TELOMERE	Details ...	18	0.83	2.57	0.000	0.000	0.000	2471	tags=67%, list=4%, signal=70%

GSEA results for Stem Cell with modified max read length

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_POSTMITOTIC_NUCLEAR_PORE_COMPLEX_NPC_REFORMATION	Details ...	22	0.89	3.59	0.000	0.000	0.000	3777	tags=82%, list=7%, signal=88%
2	REACTOME_PCNA_DEPENDENT_LONG_PATCH_BASE_EXCISION_REPAIR	Details ...	21	0.87	3.42	0.000	0.000	0.000	2589	tags=76%, list=5%, signal=80%
3	REACTOME_RHOBTB2_GTPASE_CYCLE	Details ...	22	0.85	3.25	0.000	0.000	0.000	5305	tags=82%, list=9%, signal=90%
4	REACTOME_INHIBITION_OF_THE_PROTEOLYTIC_ACTIVITY_OF_APC_C_REQUIRED_FOR_THE_ONSET_OF_ANAPHASE_BY_MITOTIC_SPINDLE_CHECKPOINT_COMPONENTS	Details ...	21	0.80	3.09	0.000	0.000	0.000	5006	tags=81%, list=9%, signal=89%
5	REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_LAGGING_STRAND	Details ...	15	0.91	3.09	0.000	0.000	0.000	2767	tags=87%, list=5%, signal=91%
6	REACTOME_LAGGING_STRAND_SYNTHESIS	Details ...	20	0.92	3.05	0.000	0.000	0.000	2767	tags=90%, list=5%, signal=95%
7	REACTOME_MISMATCH_REPAIR	Details ...	15	0.87	2.91	0.000	0.000	0.000	2333	tags=67%, list=4%, signal=70%
8	REACTOME_FORMATION_OF_SENESCENCE_ASSOCIATED_HETEROCHROMATIN_FOCI_SAHF	Details ...	16	0.86	2.89	0.000	0.000	0.000	3374	tags=94%, list=6%, signal=100%
9	REACTOME_KSRP_KHSRP_BINDS_AND_DESTABILIZES_MRNA	Details ...	15	0.91	2.83	0.000	0.000	0.000	4852	tags=87%, list=9%, signal=95%
10	REACTOME_MRNA_DECAY_BY_3_TO_5_EXORIBONUCLEASE	Details ...	16	0.90	2.83	0.000	0.000	0.000	5319	tags=88%, list=9%, signal=97%
11	REACTOME_PROTEIN_HYDROXYLATION	Details ...	18	0.74	2.72	0.000	0.000	0.000	4749	tags=67%, list=8%, signal=73%
12	REACTOME_TRISTETRAPROLIN_TTP_ZFP36_BINDS_AND_DESTABILIZES_MRNA	Details ...	17	0.71	2.68	0.000	0.000	0.000	4852	tags=71%, list=9%, signal=77%
13	REACTOME_INITIATION_OF_NUCLEAR_ENVELOPE_NE_REFORMATION	Details ...	17	0.79	2.66	0.000	0.000	0.000	3243	tags=65%, list=6%, signal=69%
14	REACTOME_CONVERSION_FROM_APC_C_CDC20_TO_APC_C_CDH1_IN_LATE_ANAPHASE	Details ...	20	0.73	2.54	0.000	0.000	0.000	5006	tags=75%, list=9%, signal=82%
15	REACTOME_TELOMERE_EXTENSION_BY_TELOMERASE	Details ...	20	0.71	2.52	0.000	0.000	0.000	3826	tags=65%, list=7%, signal=70%

GSEA results for Enterocyte (default parameters)

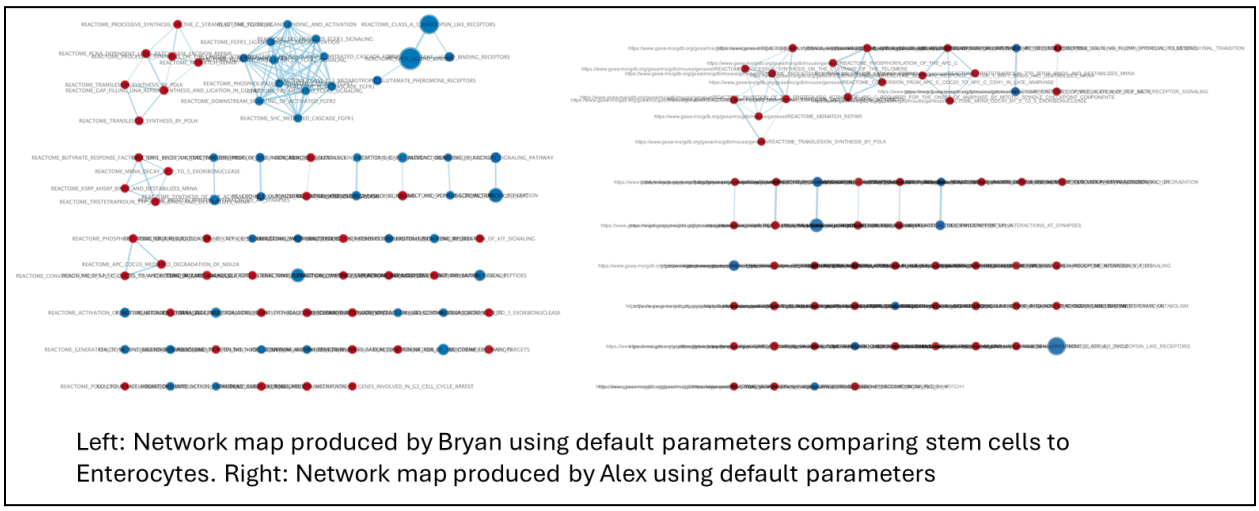
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_CLASS_I_PEROXISOMAL_MEMBRANE_PROTEIN_IMPORT	Details ...	19	-0.81	-1.66	0.000	0.000	0.009	1693	tags=11%, list=3%, signal=11%
2	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	Details ...	22	-0.70	-1.45	0.000	0.002	0.580	6416	tags=36%, list=11%, signal=41%
3	REACTOME_NEGATIVE_REGULATION_OF_FLT3	Details ...	15	-0.72	-1.43	0.012	0.004	0.779	5493	tags=13%, list=10%, signal=15%
4	REACTOME_FATTY_ACIDS	Details ...	22	-0.68	-1.42	0.005	0.004	0.798	13495	tags=73%, list=24%, signal=96%
5	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	Details ...	32	-0.66	-1.41	0.003	0.005	0.854	6416	tags=28%, list=11%, signal=32%
6	REACTOME_OLFACTORY_SIGNALING_PATHWAY	Details ...	72	-0.62	-1.40	0.000	0.007	0.917	13764	tags=69%, list=24%, signal=92%
7	REACTOME_WNT_LIGAND_BIOGENESIS_AND_TRAFFICKING	Details ...	25	-0.64	-1.34	0.017	0.019	1.000	14637	tags=56%, list=26%, signal=76%
8	REACTOME_EICOSANOIDS	Details ...	17	-0.65	-1.32	0.033	0.024	1.000	13496	tags=71%, list=24%, signal=93%
9	REACTOME_AMINE_LIGAND_BINDING_RECEPTORS	Details ...	43	-0.61	-1.32	0.004	0.025	1.000	18466	tags=74%, list=33%, signal=110%
10	REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE_FGFR1	Details ...	16	-0.65	-1.32	0.038	0.026	1.000	16334	tags=75%, list=29%, signal=105%
11	REACTOME_FORMATION_OF_THE_CORNIFIED_ENVELOPE	Details ...	90	-0.58	-1.32	0.000	0.026	1.000	17021	tags=80%, list=30%, signal=96%
12	REACTOME_SHC_MEDIATED_CASCADE_FGFR1	Details ...	20	-0.64	-1.32	0.035	0.027	1.000	16334	tags=60%, list=29%, signal=84%
13	REACTOME_NEUREXINS_AND_NEUROLIGINS	Details ...	31	-0.61	-1.31	0.017	0.029	1.000	9596	tags=39%, list=17%, signal=47%
14	REACTOME_FGFR1_LIGAND_BINDING_AND_ACTIVATION	Details ...	17	-0.65	-1.31	0.030	0.030	1.000	16334	tags=76%, list=29%, signal=108%
15	REACTOME_PI_3K_CASCADE_FGFR1	Details ...	21	-0.63	-1.30	0.042	0.033	1.000	16334	tags=57%, list=29%, signal=80%

GSEA results for Enterocyte with modified max read length

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	REACTOME_CLASS_I_PEROXISOMAL_MEMBRANE_PROTEIN_IMPORT	Details ...	19	-0.82	-1.67	0.000	0.000	0.005	1685	tags=11%, list=3%, signal=11%
2	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	Details ...	22	-0.70	-1.44	0.001	0.003	0.655	6372	tags=36%, list=11%, signal=41%
3	REACTOME_FATTY_ACIDS	Details ...	22	-0.68	-1.42	0.001	0.005	0.796	13408	tags=73%, list=24%, signal=95%
4	REACTOME_NEGATIVE_REGULATION_OF_FLT3	Details ...	15	-0.71	-1.41	0.014	0.006	0.877	5432	tags=13%, list=10%, signal=15%
5	REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS	Details ...	32	-0.66	-1.41	0.003	0.006	0.894	6372	tags=28%, list=11%, signal=32%
6	REACTOME_OLFACTORY_SIGNALING_PATHWAY	Details ...	72	-0.62	-1.38	0.000	0.010	0.973	13728	tags=69%, list=24%, signal=92%
7	REACTOME_PI_3K_CASCADE_FGFR1	Details ...	21	-0.66	-1.36	0.012	0.013	0.994	15122	tags=57%, list=27%, signal=78%
8	REACTOME_FGFR1_LIGAND_BINDING_AND_ACTIVATION	Details ...	17	-0.67	-1.35	0.029	0.015	0.996	15122	tags=76%, list=27%, signal=104%
9	REACTOME_SHC_MEDIATED_CASCADE_FGFR1	Details ...	20	-0.66	-1.35	0.012	0.015	0.996	15122	tags=60%, list=27%, signal=82%
10	REACTOME_PHOSPHOLIPASE_C_MEDIATED_CASCADE_FGFR1	Details ...	16	-0.67	-1.34	0.037	0.017	0.999	15122	tags=75%, list=27%, signal=102%
11	REACTOME_WNT_LIGAND_BIOGENESIS_AND_TRAFFICKING	Details ...	25	-0.64	-1.34	0.017	0.019	1.000	14540	tags=56%, list=26%, signal=75%
12	REACTOME_PI_3K_CASCADE_FGFR2	Details ...	22	-0.63	-1.32	0.029	0.025	1.000	15122	tags=55%, list=27%, signal=74%
13	REACTOME_EICOSANOIDS	Details ...	17	-0.65	-1.31	0.042	0.026	1.000	13408	tags=71%, list=24%, signal=93%
14	REACTOME_FRS_MEDIATED_FGFR1_SIGNALING	Details ...	22	-0.63	-1.31	0.019	0.026	1.000	15122	tags=55%, list=27%, signal=74%
15	REACTOME_AMINE_LIGAND_BINDING_RECEPTORS	Details ...	43	-0.60	-1.31	0.002	0.026	1.000	18383	tags=74%, list=33%, signal=110%

Enrichment Mapping:

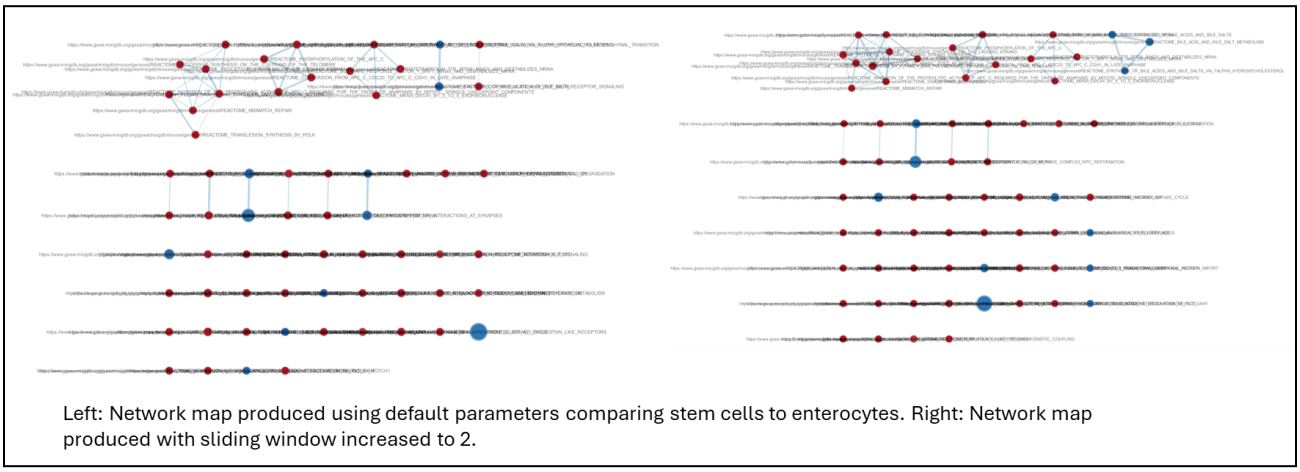
There was no direct comparison we could make to the reference paper for enrichment mapping, however, each group member was able to generate enrichment maps that somewhat resemble each other. We attribute the differences between runs to how GSEA runs the initial analysis that we then used to make enrichment maps in Cytoscape. Due to the use of permutation testing and random seeding in GSEA, we would expect differences in which enriched pathways are most significant. This will affect the downstream enrichment map visualization. However, the in-group comparison between network maps are still similar with nodes correlating between separate analyses. For instance, each member using default parameters produced an enrichment map with seven upregulated nodes connected with one of the nodes labeled as REACTOME_TRANSLATION_SYNTHEIS_BY_POLK.



Left: Network map produced by Bryan using default parameters comparing stem cells to Enterocytes. Right: Network map produced by Alex using default parameters

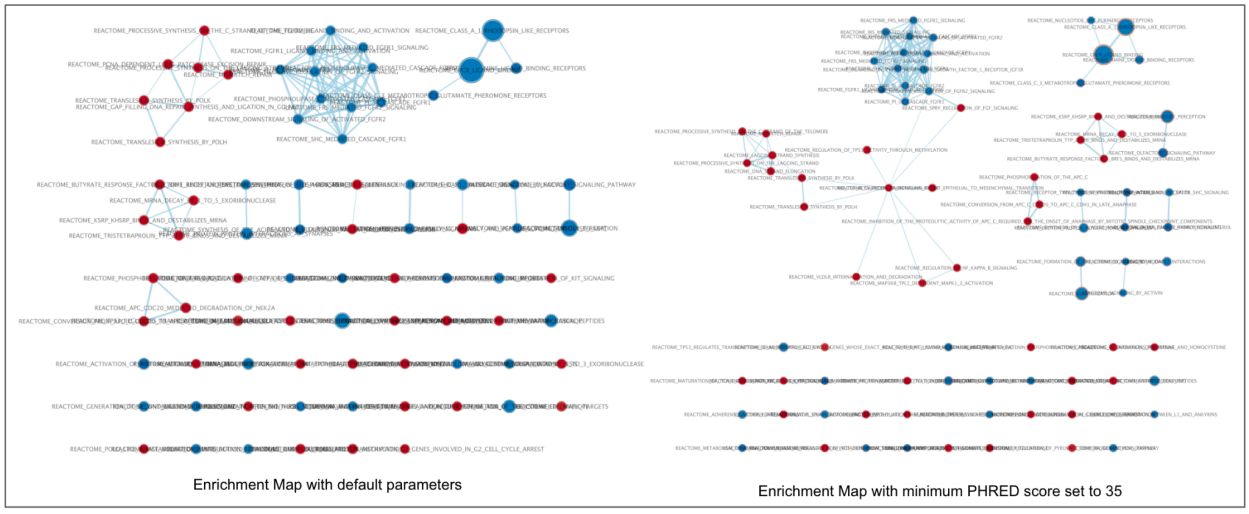
Sliding Window Parameter Change. Adjusting the sliding window during trimming from 1 to 2 slightly altered the final network maps. However, they remained similar overall. In general, the increased window produced more connected nodes with common nodes shared between the two analyses. For instance, with default parameters the left of the network map contains seven nodes connected that are upregulated with one of the nodes being REACTOME_MISMATCH_REPAIR. With the increased window, this portion of the network map increased from seven nodes to nine while still retaining the REACTOME_MISMATCH_REPAIR

node. These results are expected since the increased window should trim less reads, and retain more, leading to an increased number of nodes in our downstream enrichment maps.



Left: Network map produced using default parameters comparing stem cells to enterocytes. Right: Network map produced with sliding window increased to 2.

Min PHRED Score Parameter Change. Increasing the minimum PHRED score parameter from 20 to 35 showed minor, yet significant differences between enrichment maps. The map with default parameters exhibited two major clusters — one of gene sets upregulated in stem cells, and one of gene sets upregulated in enterocytes. In this map, the clusters were lone standing, and not connected to any other clusters or gene sets. In the enrichment map with the minimum PHRED score parameter increased to 35, many previously unconnected gene sets became intertwined, and the two major clusters became indirectly connected through this new network. This increase in connections is likely due to the elimination of lower quality reads. With the removal of lower quality reads, previously misrepresented or ambiguous connections due to noise now show clearer, more reliable associations.



Enrichment Map with default parameters

Enrichment Map with minimum PHRED score set to 35

Methods:

Raw RNAseq data for our report was generated from SRA#SRP242699 (GSE GSE143915). Raw reads were downloaded from the European Nucleotide Archive and stored on The University of Arizona's HPC. Slurm batch jobs were used to run Fastp (v0.23.2), FastQC (v0.11.9), STAR (v2.7.3a), and MultiQC (v1.20) using the HPC. Parameter changes were done at the level of Fastp and consisted of - changing the default minimum PHRED score to 35, increasing the sliding window to 2, and decreasing the maximum read length to 85. STAR annotations were done using mouse genome gencode.vM36.primary_assembly.annotation.gtf. Featurecounts (v2.0.3) was performed on STAR output .bam files for both default and modified parameters. Featurecounts output .txt files were loaded into R (v4.5) and downstream analysis was performed in RStudio (Build 563 2024.12.1). Gene IDs were converted into gene names using biomaRt (v2.64.0) in Rstudio using mmusculus_gene_ensembl dataset. DESeq2 (v1.48.0) analysis was performed in Rstudio using default parameters. Significance was determined using the reference papers published values of padi < 0.01 and > 50 mean counts. rlog (blind = false) was used to transform DESeq2 data for visualization. PCA plots were generated from rlog data using Rstudio's plotPCA function. PCA plots were generated using top 100 genes based on variance. Heatmaps were generated using Rstudio's pheatmap function using rlog transformed data. Volcano plots were generated in Rstudio using enhancedvolcano plot function with rlog transformed data. Significance for volcano plots was set to log2fold change >2 and padj <10^-48 for Alex or <10^-45 for Bryan.

All functional analyses were performed comparing stem cells to enterocytes and significance was determined using padj <0.01 and > 50 normalized count. G:profiler web interface (version *e112_eg59_p19_25aa4782*, database updated on 03/02/2025) with g:GOSt was used to assess pathway enrichment for KEEG, Reactome, and WikiPathways with significance defined as Benjamini-Hochberg FDR <0.05. Panther (v19.0) web interface was used for Panther pathways classification. Gene set enrichment analysis was performed using GSEA (v4.4.0) on rlog transformed data using m2.cp.reactome.v2024.1. Enrichment maps were generated in Cytoscape (v3.10.3) by directly exporting GSEA data into Cytoscape using GSEA's enrichment map visualization tool.

The final SRR accession numbers for each biological replicate used in our analysis are SRR10913313 (Stem, BR#1), SRR10913315 (Stem, BR#2), SRR10913317 (Stem, BR#3), SRR10913319 (AbsPro, BR#1), SRR10913321 (AbsPro, BR#2), SRR10913323 (AbsPro, BR#3), SRR10913325 (SecPDG, BR#1), SRR10913328 (SecPDG, BR#2), SRR10913331 (SecPDG, BR#3), SRR10913334 (SecPDG, BR#4), SRR10913337 (Tuft, BR#1), SRR10913340 (Tuft, BR#2), SRR10913343 (Tuft, BR#3), SRR10913346 (Tuft, BR#4), SRR10913349 (Tuft, BR#5), SRR10913352 (Enterocyte, BR#1), SRR10913354 (Enterocyte, BR#2), SRR10913356 (Enterocyte, BR#3), SRR10913358 (Enterocyte, BR#4), SRR10913360 (Enterocyte, BR#5), SRR10913362 (EEC, BR#1), SRR10913364 (EEC, BR#2).

AI was used to assist in generating code for HPC data processing and for Rstudio analysis. AI models used were ChatGPT-4, Claude 3.7 Sonnet, and Cursor (v0.46.6).

References:

1. Ashton, G. H., Morton, J. P., Myant, K., Phesse, T. J., Ridgway, R. A., Marsh, V., Wilkins, J. A., Athineos, D., Muncan, V., Kemp, R., Neufeld, K., Clevers, H., Brunton, V., Winton, D. J., Wang, X., Sears, R. C., Clarke, A. R., Frame, M. C., & Sansom, O. J. (2010). Focal Adhesion Kinase Is Required for Intestinal Regeneration and Tumorigenesis Downstream of Wnt/c-Myc Signaling. *Developmental Cell*, *19*(2), 259–269. https://doi.org/10.1016/j.devcel.2010.07.015

2. Liu, Y., & Chen, Y.-G. (2020). Intestinal epithelial plasticity and regeneration via cell dedifferentiation. *Cell Regeneration*, *9*(1). https://doi.org/10.1186/s13619-020-00053-5

3. Pond, K. W., Morris, J. M., Alkhimenok, O., Varghese, R. P., Cabel, C. R., Ellis, N. A., Chakrabarti, J., Zavros, Y., Merchant, J. L., Thorne, C. A., & Paek, A. L. (2022). Live-cell imaging in human colonic monolayers reveals ERK waves limit the stem cell compartment to maintain epithelial homeostasis. *eLife*, *11*. https://doi.org/10.7554/elife.78837

4. Raffeiner, P., Hart, J. R., García-Caballero, D., Bar-Peled, L., Weinberg, M. S., & Vogt, P. K. (2020). An MXD1-derived repressor peptide identifies noncoding mediators of MYC-driven cell proliferation. *Proceedings of the National Academy of Sciences*, *117*(12), 6571–6579. https://doi.org/10.1073/pnas.1921786117

5. VanDussen, K. L., Carulli, A. J., Keeley, T. M., Patel, S. R., Puthoff, B. J., Magness, S. T., Tran, I. T., Maillard, I., Siebel, C., Kolterud, Å., Grosse, A. S., Gumucio, D. L., Ernst, S. A., Tsai, Y.-H., Dempsey, P. J., & Samuelson, L. C. (2012). Notch signaling modulates proliferation and differentiation of intestinal crypt base columnar stem cells. *Development*, *139*(3), 488–497. https://doi.org/10.1242/dev.070763

6. Zhao, R., & Michor, F. (2013). Patterns of Proliferative Activity in the Colonic Crypt Determine Crypt Stability and Rates of Somatic Evolution. *PLoS Computational Biology*, *9*(6), e1003082. https://doi.org/10.1371/journal.pcbi.1003082

7. Gehart, H., & Clevers, H. (2019). Tales from the crypt: New insights into intestinal stem cells. Nature Reviews Gastroenterology & Hepatology, 16(1), 19–34. https://doi.org/10.1038/s41575-018-0081-y