The Effects of Hdac6-/- on Hematopoietic Stem and Progenitor Cell Homeostasis in Mouse Bone Marrow

Shane Schroeder and William Gregor

Introduction

We performed RNA-seq analysis using data from hematopoietic stem/progenitor cells (HSC and HPC respectively) in the bone marrow of Hdac6$^{-/-}$ mice. HDAC6 is a histone deacetylase which deacetylates histones, tubulin, and cytosolic enzymes. Changes in protein acetylation lead to an increase or decrease in activity depending on the effects of the structural change induced by the addition or removal of the functional group. A cytosolic enzyme of interest that is deacetylated by HDAC6 is isocitrate dehydrogenase 1 (IDH1). IDH1 converts isocitrate into α-ketoglutarate (αKG) and is shown to have decreased activity when it is acetylated. The acetylation of this enzyme is crucial for controlling DNA methylation since αKG is a cofactor for the ten eleven translocation (TET) enzyme family of demethylases. These enzymes convert 5-methylcytosine (5mC) into the demethylation intermediate 5-hydroxymethylcytosine (5hmC). Proper control of TET activity is especially important in stem cells which have highly dynamic expression patterns to regulate their replication and differentiation. Dysregulation of the process by which HSCs and HPCs differentiate into various blood components is linked to a number of blood and immune related pathologies. For this reason, it is important to understand the consequences of reduced IDH1 activity due to the absence of HDAC6 deacetylation on the gene expression of HSCs and HPCs.

Methods

We got our experimental data from NCBI using the SRA Toolkit and *M. musculus* Refseq data from ENSEMBL. We then used various other programs and commands to process and analyze the data. All of the analyses and computations were done on the Redhawk high-performance computing cluster. The specific software and versions used are sratoolkit.3.0.0, parallel-fastq-dump, fastqc_v0.11.9, Trimmomatic-0.36, kallisto-0.48.0, IGV_Linux_2.15.4_WithJava, R-4.1.2, rsem-1.3.3, and bowtie-2-2.4.5.

To analyze our experimental data we used two RNA-seq analysis pipelines. These were the RSEM-EBSEQ and Kallisto-DESeq2 pipelines from Tutorials 12 and 13 on Canvas, respectively.

Kallisto-DESeq2

The Kallisto-DESeq2 pipeline consisted of several Jupyter notebooks that guide the users all the way through analysis starting from importing data using the SRA Toolkit. The crucial analysis steps of this pipeline were trimming, pseudoalingment, feature count extraction, and Differential Gene Expression analysis. For trimming our reads we used Trimmomatic as follows:

```
java -jar trimmomatic-0.36.jar SE -threads 24 \
  SRR123.fastq SRR123_trimmed.fastq \
```

```
SLIDINGWINDOW:4:20 MINLEN:75
```

The Java program uses `SE` as the first parameter, indicating single-end mode which was a source of error in our experiment. `SLIDINGWINDOW:4:20 MINLEN:75` indicates that the reads were scanned with a window size of 4 and if the average quality score of the window is below 20, that read will be trimmed. `MINLEN:75` specifies that any reads under 75 bases will be trimmed.

We then used the *M. musculus* transcriptome and annotation files from release 97 along with our trimmed reads to achieve pseudoalingment using Kallisto. We used Kallisto as follows:

```
kallisto quant --single --threads=24 \
     --index=Mus_musculus.GRCm38_index --bootstrap-samples=25 \
     --fragment-length=200 --sd=20 --output-dir=$output \
     --genomebam --gtf=Mus_musculus.GRCm38.97.chr.gtf.gz \
     --chromosomes=mouse_chromosomes.tsv \
```

Pseudoalignment differs from normal alignment because pseudoalignment aims to "determine, for each read, not where in each transcript it aligns, but rather which transcripts it is compatible with" (Chan, 2015). It achieves this by using Transcriptome de Bruijn Graphs (T-DBGs). Typically de Bruijn graphs are build using k-mers from the reads, however T-DBGs are build from the k-mers from the transcriptome. This achieves faster computation but less accuracy than normal alignment and quantification. The accuracy however is still very similar to RSEM.

We then used the Rsubread package for R to calculate the expression counts for our reads using the BAM files generated from Kallisto. We used it in R as follows:

```
featureCounts(files={bam files},annot.ext={GTF annotation},
isGTFAnnotationFile=TRUE,GTF.featureType="exon",GTF.attrType="gene_id,
nthreads=24)
```

This Rsubread function will generate expression counts based on genes, up to the exon level, not isoform level. This will output a CSV file that contains the expression counts for each of our samples for all the genes annotated in the GTF file.

The final step of this pipeline is to use the feature expression counts to run differential expression analysis. We used DESeq2 which is another package in R and it was used as follows:

```
dds <- DESeqDataSetFromMatrix(countData=countData,
                              colData=metaData,
                              design=~Treatment, tidy = TRUE)
dds <- DESeq(dds)
res <- results(dds)
res <- res[order(res$padj),]
```

The first line generates a DESeq Data Object which is then analyzed using the DESeq2 function. Then we used the results function to get our results table and ordered it by `padj` which stands for adjusted p-value. This is how we collected our 6 most differentially expressed genes.

RSEM-EBSEQ

The RSEM-EBSEQ pipeline consisted of a batch script that was submitted to a compute node. This script is specifically designed for this project and is not general purpose. The crucial analysis steps of this pipeline were trimming, alignment and feature count extraction, and Differential Gene and Transcript Expression analysis. Data was treated as paired-end for all programs in this pipeline. When extracting the FASTQ data from the SRA data we made sure to split the files for each end of the read. Trimming the reads was done slightly differently now that we had separate files for each sample. We used Trimmomatic again as follows:

```
java -jar trimmomatic-0.36.jar PE -threads 24 \
      SRR123_1.fastq SRR123_2.fastq \
      SRR123_1.trimmed.fastq SRR123_1.trimmedOrphan.fastq \
      SRR123_2.trimmed.fastq SRR123_2.trimmedOrphan.fastq \
      SLIDINGWINDOW:4:20 MINLEN:75
```

In this pipeline we used Trimmomatic with the `PE` parameter indicated paired-end reads. As a result we needed to specify each read and separate output files for trimming results. We used the same sliding window and minimum length options to have consistency between the two pipelines.

This pipeline does alignment and feature count extraction using one program and is not done in a separate step. This was done using RSEM.

Finally the feature counts were analyzed with EBSEQ. This will analyze all of the expression count data across the samples. We used an additional program from RSEM that takes the results from EBSEQ and with 95% confidence extracts the genes that were significantly differentially expressed.

Results

From each pipeline we selected the top 6 most differentially expressed genes identified. The Kallisto-DESeq2 pipeline yielded these six genes:

Gene ID	Gene Name
ENSMUSG00000110469	Gm10358
ENSMUSG00000107383	Gm4366
ENSMUSG00000057657	Rps18-ps3
ENSMUSG00000044424	Gm9493
ENSMUSG00000114547	Gm3226
ENSMUSG00000084149	Gm12944

The RSEM-EBSEQ pipeline yielded these six genes:

Gene ID	Gene name

ENSMUSG00000094194	Ighv5-16
ENSMUSG00000076614	Ighg1
ENSMUSG00000076534	Igkv12-89
ENSMUSG00000076535	Igkv1-88
ENSMUSG00000076596	Igkv3-10
ENSMUSG00000076577	Igkv8-30

The pipelines identified completely different genes of interest. Between the two results, even the genes that were identified were of completely different function. Kallisto-DESeq2 identified mostly predicted genes whereas RSEM-EBSEQ identified only Immunoglobulin genes.

Conclusion

The biggest problem with our RNA-sequence analysis was with the Kallisto-DESeq2 pipeline. This pipeline does not properly handle paired-end reads and instead will treat all FASTQ data as single-end reads. To improve our analysis we suggest updates to the Kallisto-DESeq2 pipeline to handle paired end reads. The RSEM-EBSEQ pipeline used in this project could also be improved to take command line arguments to specify the BioProject number and be automated for any set of data.

Due to the reduced credibility of the Kallisto-DESeq2 results, the data generated from the RSEM-EBSEQ analysis was used to draw biologically significant conclusions. There was significant upregulation of the immunoglobulin genes and 60 genes related to hematopoietic differentiation were identified to be upregulated. This pattern of increased expression levels, while not compatible with the regulatory pattern described in the introduction, suggests increased HSC and HPC differentiation rates. This increase in differentiation is associated with decreased stem cell replication and can therefore lead to leukemia as the number of undifferentiated HSCs and HPCs decrease. Further investigation of the regulatory mechanism is necessary to understand how decreased αKG production would result in so many instances of increased gene expression. This conclusion seems contradictory because IDH1 would not be deacetylated which would cause permanently low enzyme activity, thus resulting in decreased levels of αKG which would cause decreased TET DNA demethylase activity(Raineri & Mellor, 2018).

One major issue with the biological analysis was the lack of information in the project abstract on the gene regulatory pathway that is affected by the knockout. A considerable amount of time was spent trying to piece together the details of how the deacetylation of IDH1 by HDAC6 results in a change in gene expression. I was under the impression that the majority of acetylations caused increased enzymatic activity and I neglected to verify this information until late in the project. Upon finding that acetylation decreases IDH1 activity, I could no longer use my initial logic to justify the changes in gene expression(Weeks et al., 2021). Given that a large number of genes appeared to have increased expression, including all of the immunoglobulin genes, it was difficult to understand how a knockout which would decrease demethylation would cause fold increases in gene expression.

References

Chan, F. C. (2015, September 2). How "Pseudoalignments" Work in kallisto. How "pseudoalignments" work in Kallisto. Retrieved May 5, 2023, from https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

Ma, L., Tang, Q., Gao, X., Lee, J., Lei, R., Suzuki, M., Zheng, D., Ito, K., Frenette, P. S., & Dawlaty, M. M. (2022, March 4). Tet-mediated DNA demethylation regulates specification of hematopoietic stem and progenitor cells during mammalian embryogenesis. Science advances. Retrieved May 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8890710/

Raineri, S., & Mellor, J. (2018, October 23). idh1: Linking metabolism and epigenetics. Frontiers in genetics. Retrieved May 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6206167/

Tekpli, X., Urbanucci, A., Hashim, A., Vågbø, C., Lyle, R., Kringen, M., Staff, A., Dybedal, I., Mills, I., Klungland, A., & Staerk, J. (2016, May 31). Changes of 5-hydroxymethylcytosine distribution during myeloid and lymphoid differentiation of CD34+ cells. Epigenetics and Chromatin. Retrieved May 5, 2023, from https://ora.ox.ac.uk/objects/uuid:7900c90f-0188-4ae1-b84c-79eebc481a31

Weeks, J., Strom, A. I., Widjaja, V., Alexander, S., Pucher, D. K., & Sohl, C. D. (2021). Evaluating mechanisms of IDH1 regulation through site-specific acetylation mimics. Biomolecules, 11(5), 740. https://doi.org/10.3390/biom11050740

Zhang Y;Kwon S;Yamaguchi T;Cubizolles F;Rousseaux S;Kneissel M;Cao C;Li N;Cheng HL;Chua K;Lombard D;Mizeracki A;Matthias G;Alt FW;Khochbin S;Matthias P; (n.d.). Mice lacking histone deacetylase 6 have hyperacetylated tubulin but are viable and develop normally. Molecular and cellular biology. Retrieved May 5, 2023, from https://pubmed.ncbi.nlm.nih.gov/18180281/