Finding lncRNAs in bone marrow and fetal liver erythroid progenitor cells in mice

by

Alec G. Garza-Galindo

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
June 6, 2014

Signature redacted

Certified by .....
Harvey F. Lodish
Professor of Biology
Thesis Supervisor

Signature redacted

Accepted by .....
Albert R. Meyer
Chairman, Masters of Engineering Thesis Committee

Finding lncRNAs in bone marrow and fetal liver erythroid progenitor cells in mice

by

Alec G. Garza-Galindo

Submitted to the Department of Electrical Engineering and Computer Science
on June 6, 2014, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Electrical Engineering and Computer Science

Abstract

Red blood cell development is crucial to the survival of all mammals and occurs primarily in the liver during embryogenesis and then in the bone marrow during adulthood. In spite of the different microenvironments of the liver and bone marrow, current research shows that the majority of protein-coding genes important to erythropoiesis are expressed during both the fetal and adult developmental stage and the difference between red blood cells from the two stages remains unknown. Differences in the expression of long non-coding genes, which are more tissue and developmental stage specific than coding genes, may play a role in the production of red blood cells in the two tissues but also have not been studied. In this paper, we analyze RNA-seq, ChIP-seq, and DNase-seq experimental data in order to shed light on the differences and similarities in gene expression of both coding and non-coding genes, chromatin markings and promoter and enhancer activity, and transcription factor motifs between fetal and adult red blood cell progenitors. We find that solute carriers make up many of the differentially expressed genes, most notably Glut1, and that there is no clear difference in chromatin markings near transcription start sites of differentially expressed genes. Further work will analyze enriched transcription factor motifs inside DNase-I footprints generated in this work to determine which transcription factors may be responsible for differential gene expression.

Thesis Supervisor: Harvey F. Lodish
Title: Professor of Biology

Acknowledgments

This thesis would not have been possible without the support of a great number of friends and family.

I would like to thank my supervisor Juan Alvarez for his support and mentorship during my thesis project. The project would not have happened without him, and I would not have any idea how to go about computational biology research without his help.

I would also like to thank Bingbing Yuan of the bioinformatics department for putting together pipelines to expedite the process of pretty much every aspect of this thesis. A big thanks also goes out to Wenqian Hu for conducting the RNA-seq experiments from which this thesis analyzes data.

Another thank you to Havey Lodish and all of Lodish lab for having me join their lab and for the support for seeing me through lab meetings and deadlines.

To my brothers at the Massachusetts Gamma Phi Delta Theta and my friends on the cheerleading team here at MIT, thank you for being there for me and laughing with me through my years here at school. You all are some of my best friends, and I hope to stay close to you for the rest of my life.

And finally, a thank you to my family. Without your love and support, I would not be where I am today. Please take care of yourselves so I can still see you in another twenty years!

Contents

List of Figures

Chapter 1

Erythropoiesis and Long non-coding RNA Background

Red blood cells are the most common type of blood cell and the principal means of delivering oxygen to tissues through the circulatory system. The development of red blood cells is known as erythropoiesis and is vital to the survival of all mammals and occurs throughout their lifetimes. Hematopoietic stem cells in the fetal liver and adult bone marrow undergo cell lineage specification, proliferation, and differentiation in order to form mature erythrocytes[19]. Dysregulation of erythropoiesis can lead to diseases such as anemia and leukemia, and identifying the transcriptional networks and regulatory circuitry involved in this process can lend insight into the development of treatments for those diseases[5,23].

Long non-coding RNAs have been identified as regulators of erythropoiesis, and previous works have shown several lncRNAs that are crucial to mature erythrocyte formation. In this chapter, we give background on erythropoiesis and lncRNAS to motivate their study in the role of red blood development.

1.1 Red Blood Cell Development

In mammals, the blood is continuously replenished from stem cells located in the liver in the fetal developmental stage and in the bone marrow in the adult developmental

stage. These stem cells all derive from a single type of multipotent hematopoietic stem cell (HSC) which in turn differentiates into the more restricted myeloid and lymphoid progenitor stem cells. The body can exquisitely control the proliferation and differentiation of precursor cells for each blood-cell lineage through cytokines such as erythropoietin which promotes the formation of erythrocytes and G-CSF which promotes the formation of macrophages[30].

The earliest committed erythroid progenitor from the hematopoietic lineages is the burst-forming unit-erythroid (BFU-E). Early BFU-E cells divide and differentiate through the mature BFU-E stage into rapidly dividing colony-forming unit-erythroid (CFU-E cells). CFU-E progenitors divide 3 to 5 times over 2 to 3 days as they differentiate and undergo changes such as a reduction in cell cells, chromatin condensation, hemoglobinization, and eventual enucleation to form mature erythrocytes[19].

As previously mentioned, cytokines can regulate erythropoiesis, and many other proteins, RNAs, transcription factors, and chromatin modifications do as well. The transcription factors GATA1 and TAL1 bind upstream of many important erythroid genes and promote their expression[19]. GATA1 binding tends to occur at distal enhancer elements marked with the chromatin modification H3K4me1, and several lncRNAs are targeted by enhancers bound by GATA1, implicating them in erythropoiesis. In spite of this, the contribution of lncRNAs to the regulation of erythropoiesis remains largely unknown, although several lncRNAs have been shown to be crucial to the formation of mature erythrocytes.

1.2 lncRNA Background

The central dogma of gene expression is that DNA is transcribed into messenger RNA which is then translated into an amino acid sequence for protein synthesis. However, extensive RNA transcripts that do not code for proteins have been discovered, showing the centrality of RNA in gene regulation[36]. The noncoding genes which code for these RNA can produce small regulatory RNAs such as ribosomal RNA, transfer RNA, small nucleolar RNA, and microRNAs as well as longer transcripts such as lncRNAs.

lncRNAs are RNA transcripts longer than 200 nucleotides that do not encode for a protein. lncRNAs are similar to messenger RNAs which code for proteins: most are Pol II transcripts that undergo splicing (removal of introns), 5′ capping, polyadenylation, and are regulated by specific transcription factors and chromatin modifiers[6,16]. However, they mainly reside in the nucleus and rarely engage translating ribosomes in the cytoplasm, unlike mRNA[6,8]. The versatility of RNA has allowed lncRNAs to be implicated in various processes: recruitment and regulation of transcription factors and chromatin modifiers, chromosomal domain organization, and mRNA splicing modulation, translation, and degradation[7,12,27,32,34]. LncRNAs also have properties such as moderate sequence conservation between that of introns and exons of coding genes across mammals, greater tissue-specificity than coding genes, lower expression levels compared to coding genes, smaller size than coding genes, and multi-exonic structure with alternative splicing[4,16,17]. Many lncRNAs are differentially expressed across tissues, developmental stages, and physiological states and some have specialized functions in response to developmental or physiological cues[2,4,6,9,13,15,28,40].

Due to their evolutionary conservation and expression patterns, lncRNAs have potentially interesting and cell-specific biological functions. Indeed, several lncRNAs have been functionally profiled and associated with cellular processes such as apoptosis, stem cell proliferation and lineage commitment, and dosage compensation, and imprinting[39]. Dozens of lncRNAs expressed in mouse embryonic stem cells were found to be necessary for maintaining pluripotency[18]. For blood cells, one lncRNA termed lincRNA erythroid prosurvival (lincRNA-EPS) identified and profiled in previous work was found to rescue erythroid precursors from apoptosis in the absence of Epo through ectopic expression, and inhibition of lincRNA-EPS blocked proliferation of erythroid precursors and promoted apoptosis[20].

Because of their tissue specificity, roles in development, and their differential expression under different physiologic conditions, lncRNAs are potentially important to the shift of erythropoiesis from the fetal liver to the adult bone marrow. Additionally, several lncRNAs have already been implicated in regulating erythropoiesis. In this paper, we identify lncRNAs from the mouse fetal liver and adult bone marrow

in order to find those differentially expressed between the two developmental stages. Profiling the function of such lncRNAs may reveal insights into the reasons behind erythropoiesis moving and lead to greater understanding of erythropoiesis in general.

Chapter 2

Methods

In order to find lncRNA transcripts differentially expressed between mouse fetal liver (FL) and adult bone marrow erythroid-committed cells (BM), we first had to assemble the transcriptomes of erythroid progenitor cells from both tissues. Assembling the transcriptome of a cell involves sequencing all of the RNA messages being transcribed in a cell, mapping them back to the genome from which they came, and determining the structure and expression level of genes from the mapped reads.

The mouse July 2007 (NCBI37/mm9) genome assembly was used throughout the study. ENSEMBL transcript structures and annotations were obtained from ENSEMBL version 67 (`http://useast.ensembl.org/info/data/ftp/`). RefSeq and UCSC transcript structures were downloaded from the UCSC genome browser (December, 2013). RNA-seq mapped reads from CSHL, PSU, UW, and LICR for different tissues and developmental stages were downloaded from the Expression and Regulation tracks of the UCSC genome browser (January, 2014).

2.1 RNA-Seq Methodology

This section describes the process of preparing cells for RNA sequencing, the sequencing library preparation and method used, and the assembly of the transcriptome of mouse fetal liver and adult bone marrow erythroid progenitor cells. Additionally, we analyze known genes from ENSEMBL in order to verify the correctness of our assem-

bly. The process of gathering mouse fetal liver cells and preparing RNA sequencing libraries for them was done previously in the lab and is described below.

Mouse embryonic day 14.5 (E14.5) fetal liver cells were separated into TER119+ and TER119- fractions via magnetic-assisted sorting cell sorting. Total RNA was isolated from these cells using QIAGEN miRNeasy Kit according to the manufacturer's instructions. Ribosomal RNA was depleted from 4 ug total RNA using the Ribo-Zero Gold Kit from Epicentre. Strand-specific sequencing libraries were generated following a previously described protocol from the total RNA (TER119+/- cells) or from the poly(A)+ and poly(A)- fractions (TER119+ cells)[3]. The latter fractions were separated using the Solexa kit (Illumina) according the manufacturer's instructions. cDNA fragments of 400-600 bp from these libraries were selected by gel purification and then sequenced on an Illumina HiSeq2000 sequencer. The resulting directional 100 base pair paired-end reads were quality-checked with FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). A similar protocol was followed to obtain mouse adult bone marrow erythroid progenitor cell paired-end reads.

Reads were mapped to mm9 using TopHat v2.0.8[37]. The bone marrow reads were mapped using default parameters and the options `-p 4 --solexa1.3-quals --library-type fr-firststrand --min-anchor 5 -r 200`. Insert sizes were collected for the BM mapping using Picard tools (`http://picard.sourceforge.net/`) and TopHat was rerun with default parameters and the options `-p 4 --solexa1.3-quals --library-type fr-firststrand --min-anchor 5 -r 187 --mate-std-dev 95`. The junctions found by TopHat from the FL mapping (gathered through the same process as BM mappings) were merged with the junctions from the BM mapping and TopHat was run again on the BM reads using default parameters and the options `-p 4 --no-novel-juncs --solexa1.3-quals --library-type fr-firststrand --min-anchor 5 -r 187 --mate-std-dev 95` and on FL reads using default parameters and the options `-p 4 --no-novel-juncs --solexa1.3-quals --library-type fr-firststrand --min-anchor 5 -r 150 --mate-std-dev 56` as well as the `-j` option set to the merged junctions for both BM and FL read mappings.

Accepted hits from BM and FL reads were assembled using Cufflinks v2.1.1[38] in order to construct transcript and gene models and estimate gene expression using FPKM (fragments per kilobase of exon per million mapped reads). Cufflinks was run with default parameters and the options `-p 4 --min-frags-per-transfrag 0` and the `-M` option (mask option) set to a GTF of mouse rRNA from ENSEMBL for BM and FL accepted hits to assemble their transcript structures *de novo*. The assemblies produced by Cufflinks were merged with Cuffmerge and then Cufflinks was rerun with the merged assemblies as a guide. An ENSEMBL transcript models GTF was also used as a guide for Cufflinks assembly for both BM and FL accepted hits via the `-G` option.

Looking at those differentially expressed (¿ 4x expression in one tissue relative to the other) we found 1,435 BM over-expressed genes and 278 FL over-expressed genes. In order to determine the function and type of genes being differentially expressed, we used DAVID to find the gene ontology terms most enriched in each set of differentially expressed genes[21,22]. Only GO Biological Process and Molecular Function terms (GOTERM_BP_FAT and GOTERM_MF_FAT) were considered. To identify the most significant non-redundant GO terms in a gene list, we grouped annotation terms into non-redundant clusters using the Functional Annotation Clustering tool, and then selected the most significant term in each of the top clusters ranked by their enrichment score. Only GO terms enriched in our lists with a Benjamini-Hochberg adjusted p-value ¡0.05 are reported. Figure 2-1 shows the most highly enriched GO terms in the BM over-expressed gene list and Figure 2-2 shows the most highly enriched GO terms in the FL over-expressed gene list. Unfortunately, the terms most highly enriched in the BM over-expressed gene list are not related to erythropoietic cell function and are most concerted with the immune system. This suggests that some cells from another hematopoietic lineage, probably the lymphoid linage, were present in the preparation of the RNA-seq libraries for BM and sequenced.

In light of the potential contaminants, we followed a process similar to one described previously[21]. We eliminated sentinel contaminants from our list of genes. Additionally, we used CSHL, PSU, UW, and LICR mapped RNA-seq reads from

Figure 2-1: BM unregulated genes GO term enrichment.

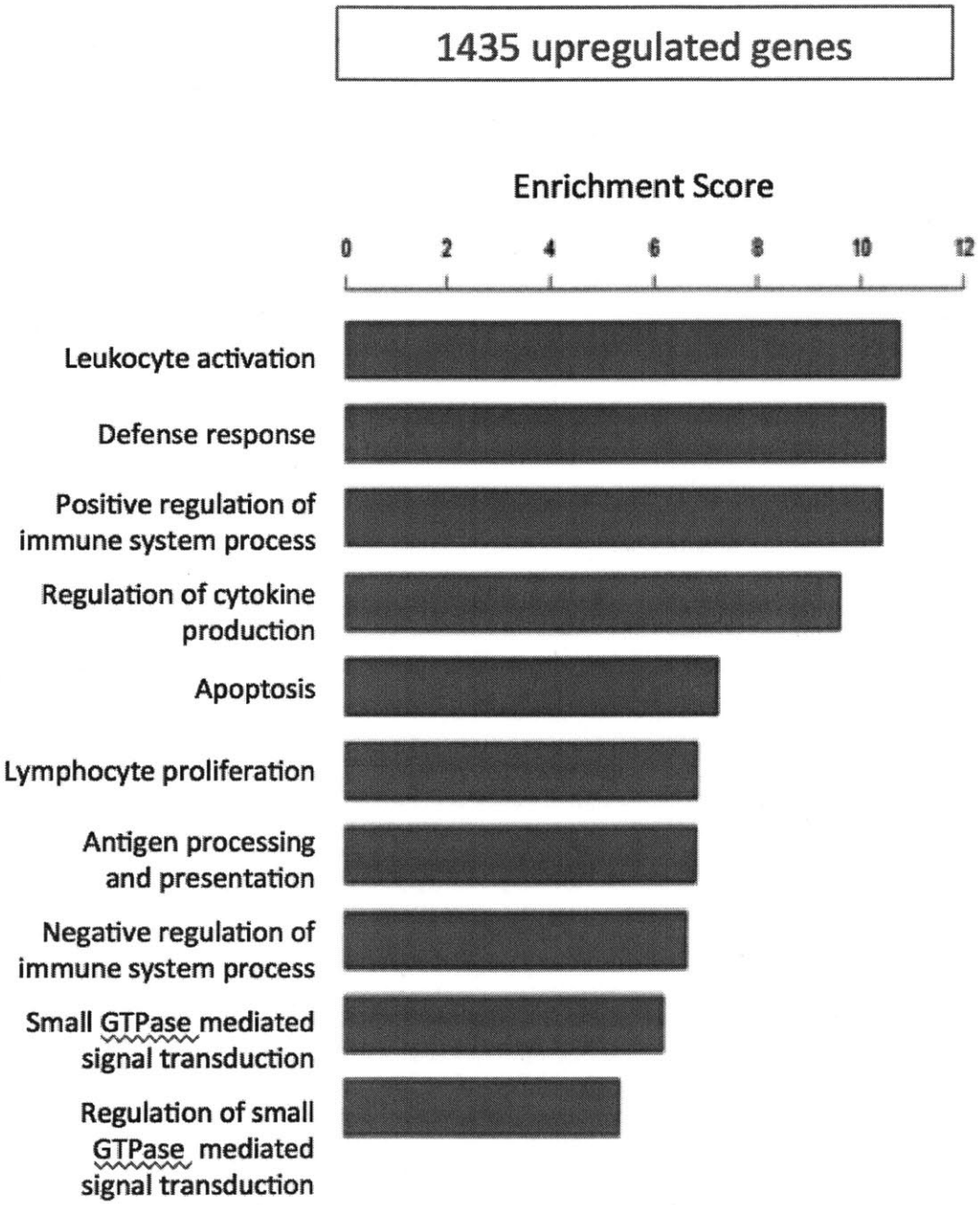
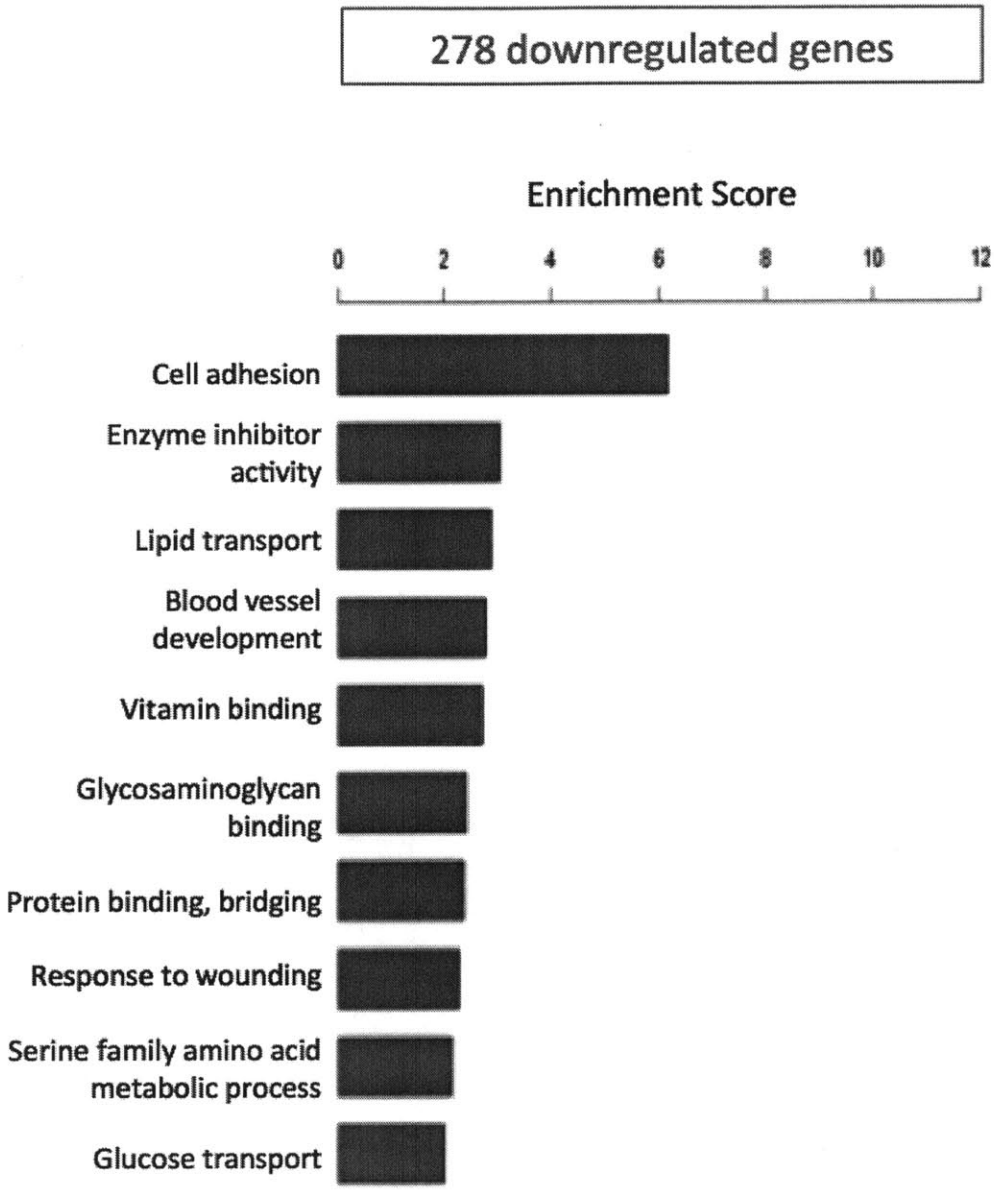


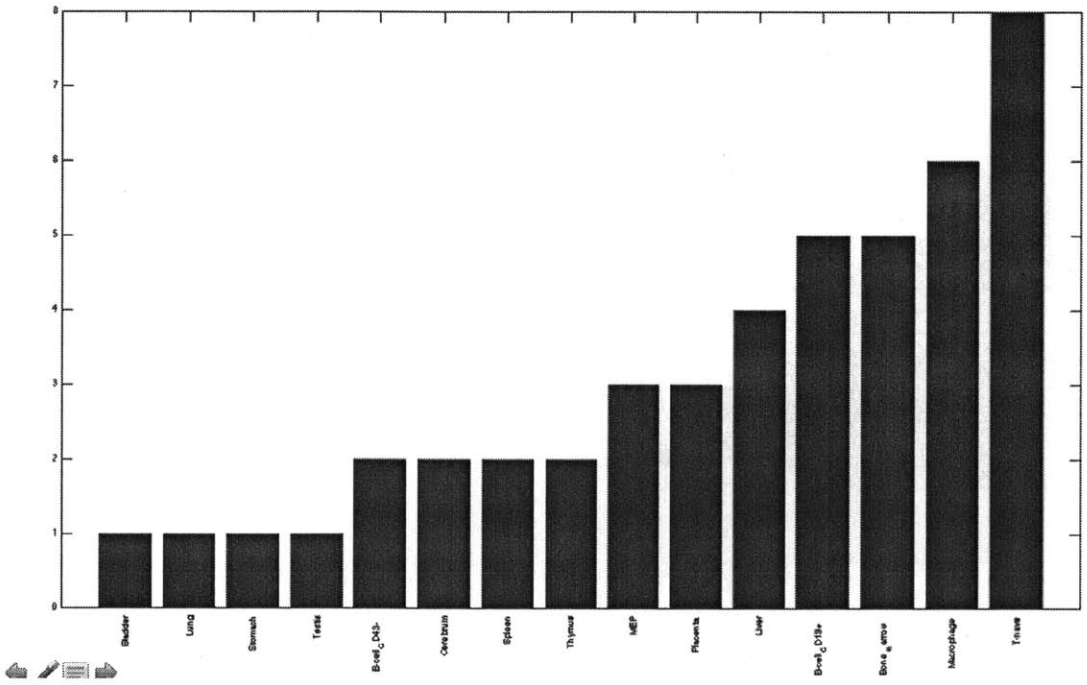
Figure 2-2: FL unregulated genes GO term enrichment.



the UCSC browser's expression and regulation track to estimate ENSEMBL gene expression levels in 31 different tissues in order to determine genes most specific to contaminant tissues. We chose placenta, adult liver, T-naive, B-cell, and macrophage

as contaminant tissues and calculated tissue specificity scores for each tissue using a metic previously described[4]. Approximately 3,500 genes were removed from consideration for being most specific to a contaminating tissue. Contaminating tissues were chosen by analyzing from which tissues sentinel contaminants and probable contaminants came, and figures 2-3 and 2-4 contain histograms of to which tissue each sentinel and probable contaminant is most specific.

Figure 2-3: Histogram of sentinel contaminants and their tissue of maximum specificity. T-naive, macrophage, bone marrow, B-cell, liver, and placenta are the tissues where the most sentinel contaminants are most specific.

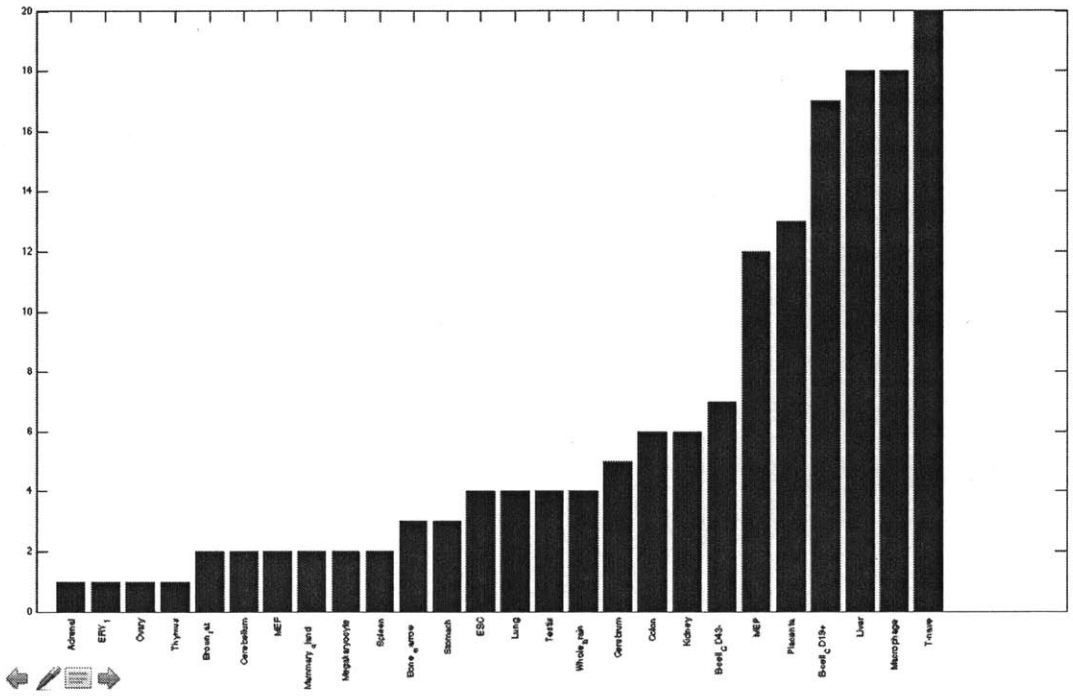


2.2 lncRNA Filtering Pipeline

After assembling transcripts *de novo* using Cufflinks, we implemented the following process in order to identify potential lncRNAs:

1. Transcripts shorter than 200 nucleotides were excluded per the definition of a lncRNA.

Figure 2-4: Histogram of probable contaminants and their tissue of maximum specificity. T-naive, macrophage, liver, B-cell, and placenta are the tissues where the most probable contaminants are most specific.



2. Transcripts with a single exon were discarded as well as those with no coverage greater than or equal to 3 in either BM or FL to ensure that the transcripts did not result from strange behavior during assembly.

3. Transcripts with known protein coding domains were removed by retrieving the longest ORF for each transcript in all three possible frames using the Sixpack tool from EMBOSS, and then used HMMER3 to query the Pfam A and Pfam B databases (downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/ on Nov 2012) with default parameters[11,35]. Any transcript with a significant Pfam hit (E-value ¿ 0.001) was discarded.

4. Repeat-masked transcripts were also blasted against the human, rat and mouse RefSeq databases separately using Blastx[14]. Transcripts mapping to any of the three protein databases with an E-value < 0.0001 were removed.

5. We used PhyloCSF to filter out transcripts under evolutionary pressure to preserve synonymous amino acid codons[29]. The PhyloCSF score of a given transcript indicates how much more probable its alignment across 29 mammalian genomes is under a model of protein-coding sequence evolution than under a non-coding model. We calculated PhyloCSF scores with removeRefGaps –frames=3 –orf=ATGStop parameters and discarded any transcript with a score >100, which corresponds to a 9.3% false negative rate and a 9.7% false positive rate using RefSeq mRNAs and RefSeq lincRNAs as reference.

6. We used the Coding Potential Calculator to exclude transcripts with characteristic coding features independent of their conservation[25]. The CPC score of a given transcript indicates its distance to a classification as protein-coding based on significant similarity to sequence features of known protein-coding transcripts learned via support vector machine learning. We calculated CPC scores using default parameters and discarded any transcript with a CPC score > 0, which corresponds to coding or weakly coding classifications.

7. We used BEDTools to intersect our *de novo* transcript models with transcript models from the RefSeq, UCSC and Ensembl databases, and discarded any transcript overlapping at least 1 bp in the same strand with any known mRNA exon.

Applying this process to the transcripts assembled *de novo* from Cufflinks, we identified 1,243 potential lncRNA genes and 1,398 potential lncRNA transcripts. In order to filter out potential contaminants from the potential lncRNAs transcripts identified, we used a similar procedure as that for filtering out contaminants from known ENSEMBL genes. Using CSHL, PSU, UW, and LICR mapped RNA-seq reads from the UCSC browser's expression and regulation track, we assembled transcripts using Cufflinks and a GTF of our potential lncRNAs as a guide for assembly. Again, transcripts most specific to placenta, adult liver, T-naive, B-cell, and macrophage were discarded, resulting in 1165 potential non-contaminant lncRNA transcripts.

2.3 DNase-seq and ChIP-seq Datasets, Peak and Footprint Calling Methodology

DNase-seq data for fetal liver was used from DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington, GEO accession GSE37074. Bone marrow DNase-seq data as well as ChIP-seq data for H3K4me3, H3K4me1, H3K27me3, and RNAP2 was used from[26] GEO accession number GSE27921. Fetal liver ChIP-seq data for H3K4me2, H3K4me3, H3K27me3, and RNAPII was used from[41] GEO accession number GSE32111. Peaks were called for ChIP-seq data using MACS[42] with $p < 0.01$ and DNase-seq data using Peakranger[10] $q < 0.01$.

DNase-I footprints were called within Peakranger peaks using an algorithm adapted from[33]. A per-nucleotide DNase-I cleavage score was generated by counting the number of 5' ends of DNase-seq data falling at that nucleotide. Left and right regions of 3-10 nucleotides around a center region of 6-40 nucleotides were analyzed and called footprints if the average cleavage score in the center region was smaller than the average cleavage score in the left and right regions. Then, we calculated a footprint occupancy score (FOS) defined as (C+1)/L + (C+1)/R where C represents the average cleavage in the center region, and L and R represent the average cleavage in the left and right flanking regions, respectively. We generated all possible footprints within each peak and then iteratively chose the footprint with the smallest FOS and removed all other footprints within that peak with center regions that overlapped the minimum FOS footprint's center region until no more footprints were left within that peak that overlapped a selected footprint. To estimate FDR, we randomly shuffled the total cleavage score inside a single peak and called footprints on the shuffled data for each dataset. At a FOS score cutoff of 0.97, the number of footprints called in the randomly shuffled datasets divided by the number of footprints called in the actual datasets first reached less than 0.01 and was used as a threshold for FOS. The code used to generate footprints can be made available on request.

Chapter 3

Conclusions and Additional Work

3.1 Discussion of Results

3.2 Discussion of RNA-seq Analysis

In order to characterize the difference between erythroid progenitor cells in the fetal liver and adult bone marrow, we analyzed RNA-seq data from Ter119+ cells from both tissues. Total RNA was isolated in both RNA datasets. Using the pipeline in figure 3.2 we assembled gene models and transcript expression estimates for known ENSEMBL protein-coding genes as well as lncRNA genes. Identifying lncRNA transcripts followed a protocol done before in the lab and is detailed in the methods section. In order to focus on reliably transcribed genes in both tissues, we only include genes with an FPKM (fragments per kilobase of exon per million mapped reads) of 0.1 or greater in both fetal liver and adult bone marrow and an FPKM of 1 or greater in either fetal liver or bone marrow, leaving 10,965 genes in total. Of the 10,965 genes analyzed, 1,435 had expression levels four-fold greater in bone marrow over fetal liver, and 278 had expression level four-fold greater in fetal liver than adult bone marrow. Figure 3-2 contains a plot of the log of expression level of the 10,965 genes in bone marrow versus fetal liver. With an R-squared of 0.65, the correlation of gene expression between erythroid cells in the two tissues is very good. For 489 known erythropoiesis-induced mRNAs such as Gata1 and Tal1, the expression level

correlates even better between the tissues Figure 3-3 contains a plot of the log of expression level of 489 known erythropoiesis-induced mRNAs in bone marrow versus fetal liver, with an R-squared of 0.86[1]. Based off of this analysis, it appears that the gene expression amongst reliably-transcribed protein-coding genes and known erythropoiesis-induced genes are similar between the two tissues as expected.

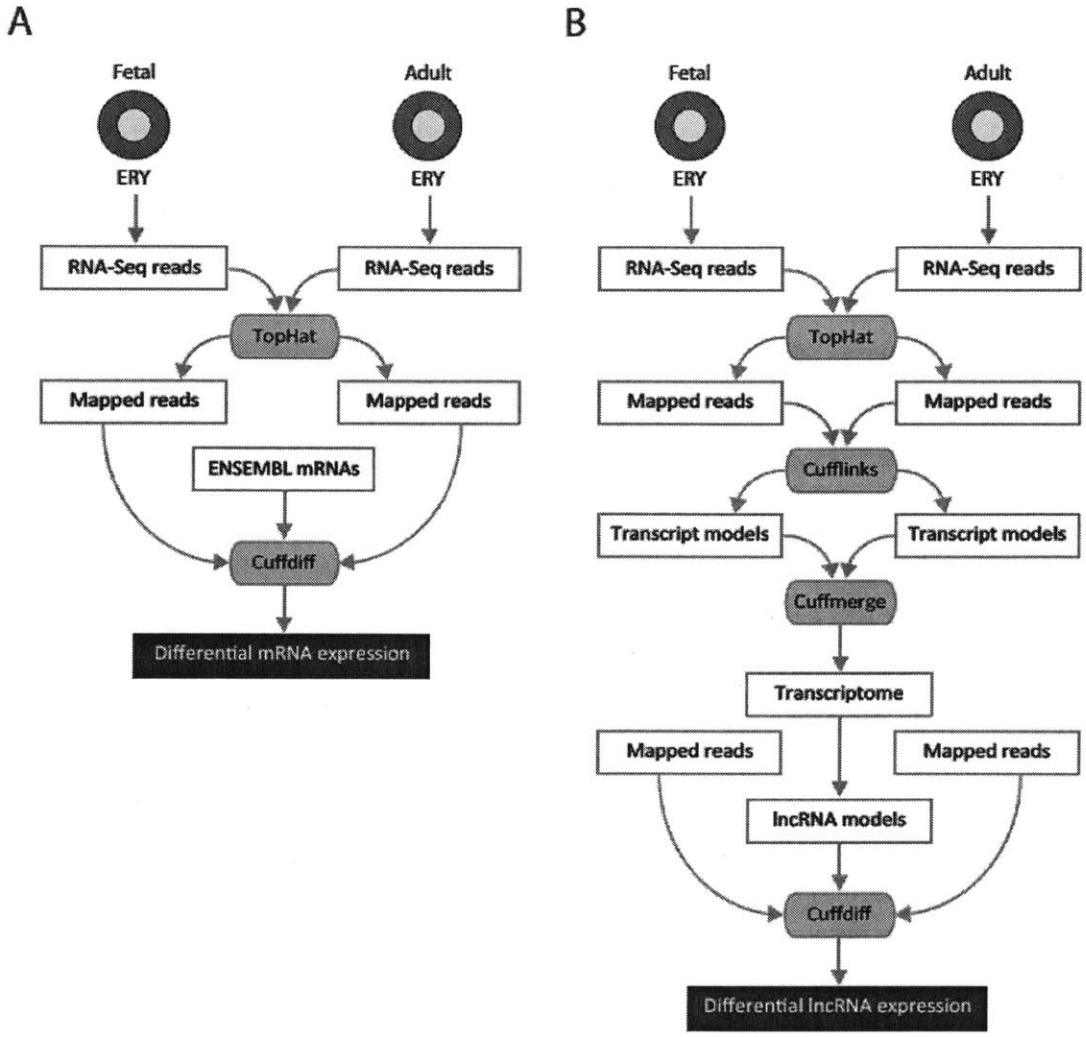


Figure 3-1: Pipeline for RNA-seq analysis.

A. ENSEMBL mRNAs transcript assembly process. RNA-seq read were mapped to the mm9 genome using TopHat and then transcript models and expression levels were generated using Cuffidff with a GTF of known ENSEMBL mRNAs as a guide for mapping to identify differentially expressed genes. B. de-novo lncRNA transcript assembly process. RNA-seq reads were mapped to the mm9 genome using TopHat as before. Cufflinks was then used on mapped reads in order to form transcript models which were then merged using Cuffmerge to form the entire transcriptome for bone marrow and fetal liver erythroid progenitor cells. lncRNA models were selected from the transcriptome and used as a guide for Cuffdiff to estimate differential gene expression.

Unfortunately, most of the genes that are enriched in bone marrow relative to fetal liver are known to be immune cell or macrophage-specific. For example, the genes Irf4, Ifitm6, and Ltf are among the most enriched in bone marrow and all are involved in immune system-related functions such as interferon regulation and antimicrobial activity. Indeed, the three most highly enriched gene-ontology terms for the 1,435 unregulated bone marrow genes are leukocyte activation, defense response, and positive regulation of immune system process. Similar contamination occurs in the fetal liver enriched cells by different tissues. For example, the gene Afp is known to be produced in the yolk sac and liver during fetal development in huge quantities, implicating it as a contaminant as well. Analysis of 59 sentinel contaminant genes from a paper doing a similar study showed that the contaminants were mostly from T-naive cells, macrophages, bone marrow, B-cells, liver, and placenta as expected. Including 264 probable contaminant genes whose expression is highly correlated with sentinel contaminants from the same study, we found that the probable contaminants also come from the same tissues. To filter out genes potentially from contaminating tissues in our own datasets, we chose to analyze expression levels of ENSEMBL genes across 30 tissues including our RNA-seq data from bone marrow and fetal liver and to remove genes from our analysis when the gene has highest expression across the tissues in a contaminating tissue (T-naive cell, macrophage, bone marrow, B-cell, liver, and placenta). This resulted in approximately 3,500 genes removed from our analysis and significantly changed the number and type of genes differentially expressed. Figure 3-4 contains two heatmaps: one of tissue expression level in differentially expressed genes before removal of contaminants and the other post-removal of contaminants. The considerable overlap of expression between bone marrow and macrophage has been reduced, leaving the largest overlap of gene expression as those genes that are highly expressed in bone marrow and fetal liver erythroid progenitors.

Of the approximately 9,000 genes left, 275 (3%) were significantly differentially expressed more highly in bone marrow and 113 (1%) more highly in fetal liver. The top enriched genes minus contaminants in bone marrow were Adpgk, Serpinb1a, Plbd1, Chi3l1, Rsad2, and Camp. The top enriched genes minus contaminants in

Figure 3-2: Scatter plot of log_{10} values of fetal liver FPKM versus bone marrow FPKM for genes with FPKM values over 0.1 in either bone marrow and fetal liver and over 1 in either bone marrow or fetal liver (R-squared 0.65). 10,965 genes are plotted in total with 1,435 genes (shown in red) having four-fold higher expression in bone marrow over fetal liver and 278 (shown in blue) having four-fold enrichment in fetal liver over bone marrow.

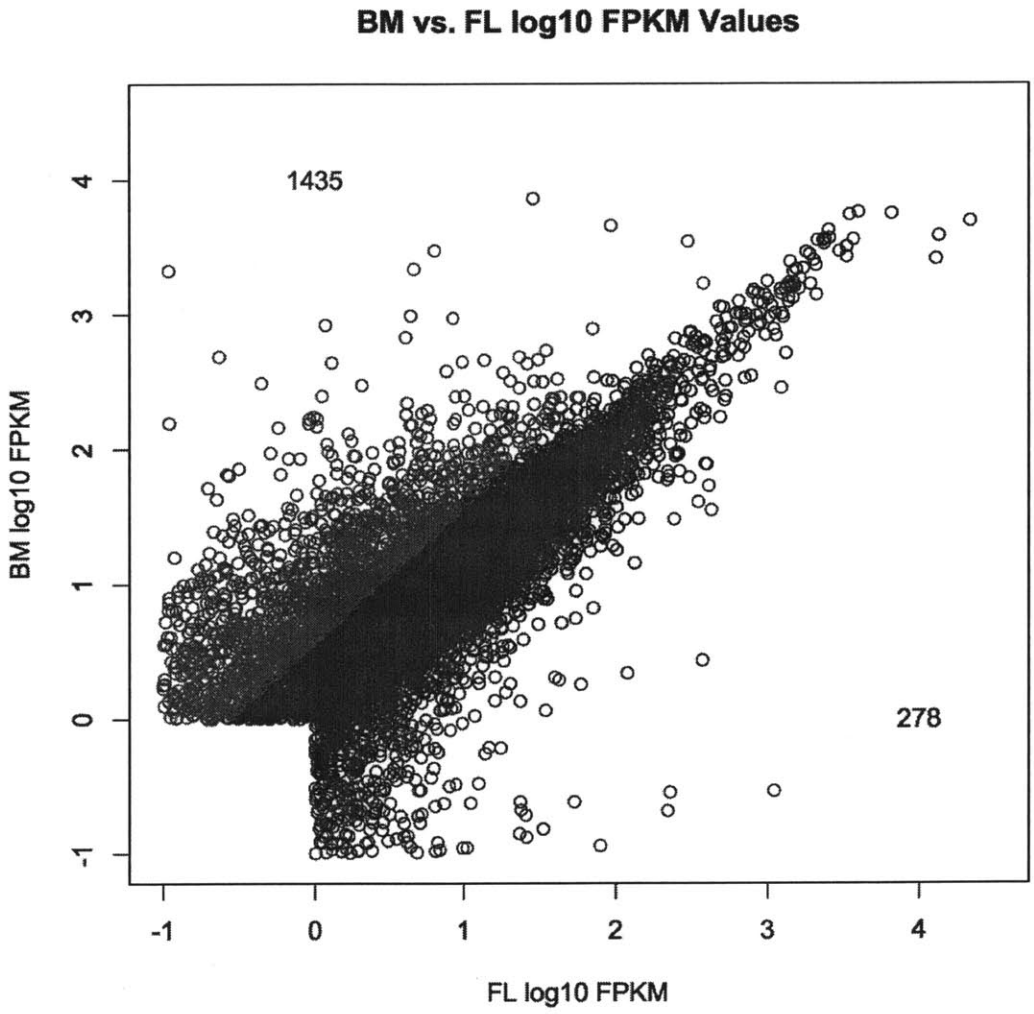
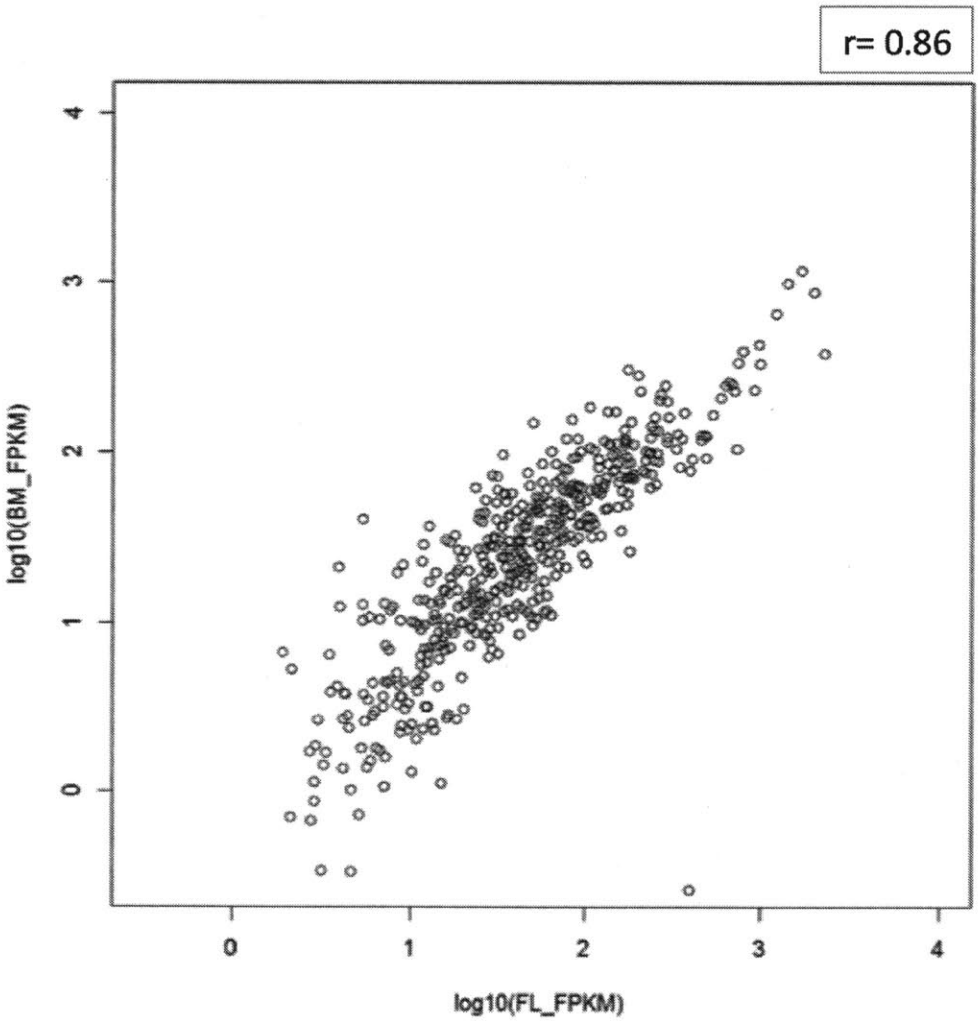


Figure 3-3: Scatter plot of log_{10} values of fetal liver FPKM versus bone marrow FPKM for 489 genes known to be induced during erythropoiesis (R-squared 0.86).



fetal liver were Hif3a, Podxl, Pacsin3, Slc30a10, Scd2, and Slc2a1. Interestingly, 152 of the total genes differentially expressed are solute carriers, with 72 high in adult bone marrow and 80 high in fetal liver. Transcription factors also made up 19 of the differentially expressed genes, with 11 high in adult bone marrow and 8 high in fetal liver. Perhaps due to the anaerobic environment of the fetal liver, different solute carriers are necessary to ensure survival of erythroid progenitors. Slc2a1 (Glut1) is necessary for glucose transport and is widely distributed in fetal tissues and Hif3a is induced during hypoxia, perhaps indicating a larger need for glucose for fetal liver erythroid progenitors as both genes are highly expressed in fetal liver and not bone marrow. Grb10 is a protein that interacts with insulin receptors and insulin-like growth-factor receptors and is also high in fetal liver and not bone marrow, again implicating some connection between glucose metabolism as a difference between the erythroid progenitor cells. Unfortunately, some genes suspected to be contaminants such as Afp survived our filtering of contaminants, so genes from contaminating tissues may still be represented.

We identified 1,335 potential lncRNA genes in our RNA-seq fetal liver and bone marrow datasets and performed filtering as with mRNAs by removing those lncRNAs from consideration that had expression level highest in contaminating tissues, leaving 519 lncRNAs. Of these 519 lncRNA genes, 84 (16%) were labeled as significantly differentially expressed and high in bone marrow and 26 (5%) significantly differentially expressed and high in fetal liver. Consistent with earlier findings surrounding lncRNAs, those identified by us in our RNA-seq datasets are more tissue-specific than coding genes. Figure 3-5 contains a heat map of lncRNA and mRNA differentially expressed genes minus contaminants which shows that even among differentially expressed genes, lncRNA genes appear more tissue specific than coding genes.

In summary from our RNA-seq experiment analysis, the gene expression profiles of significantly expressed coding genes in fetal liver and bone marrow erythroid progenitors is highly correlated and even more highly correlated amongst known erythropoiesis-induced mRNAs. Unfortunately, contamination by non-erythroid cells means that some filtering is necessary to remove potentially contaminating tran-

Figure 3-4: Heatmaps of differentially expressed gene expression levels in different tissues pre- and post-contaminant removal. Left: pre-contaminant removal. There is significant overlap of genes that have high expression in bone marrow erythroid progenitors and macrophage pre-contaminant removal. Right: post-contaminant removal. The overlap between highly expressed genes in bone marrow erythroid progenitors and macrophage has been reduced, leaving more genes that are high in bone marrow erythroid progenitors or fetal liver erythroid progenitors.

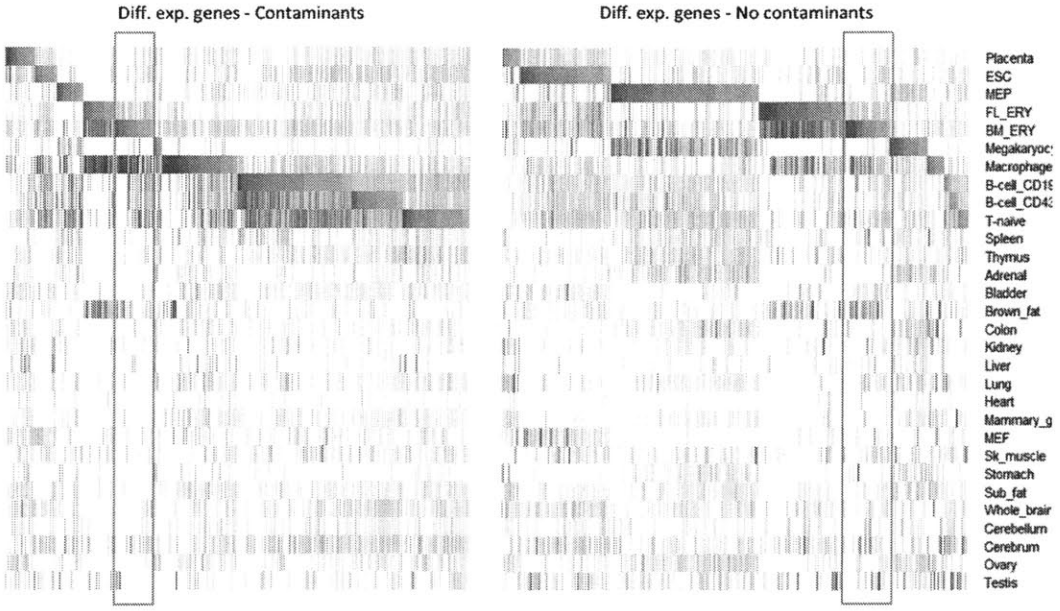
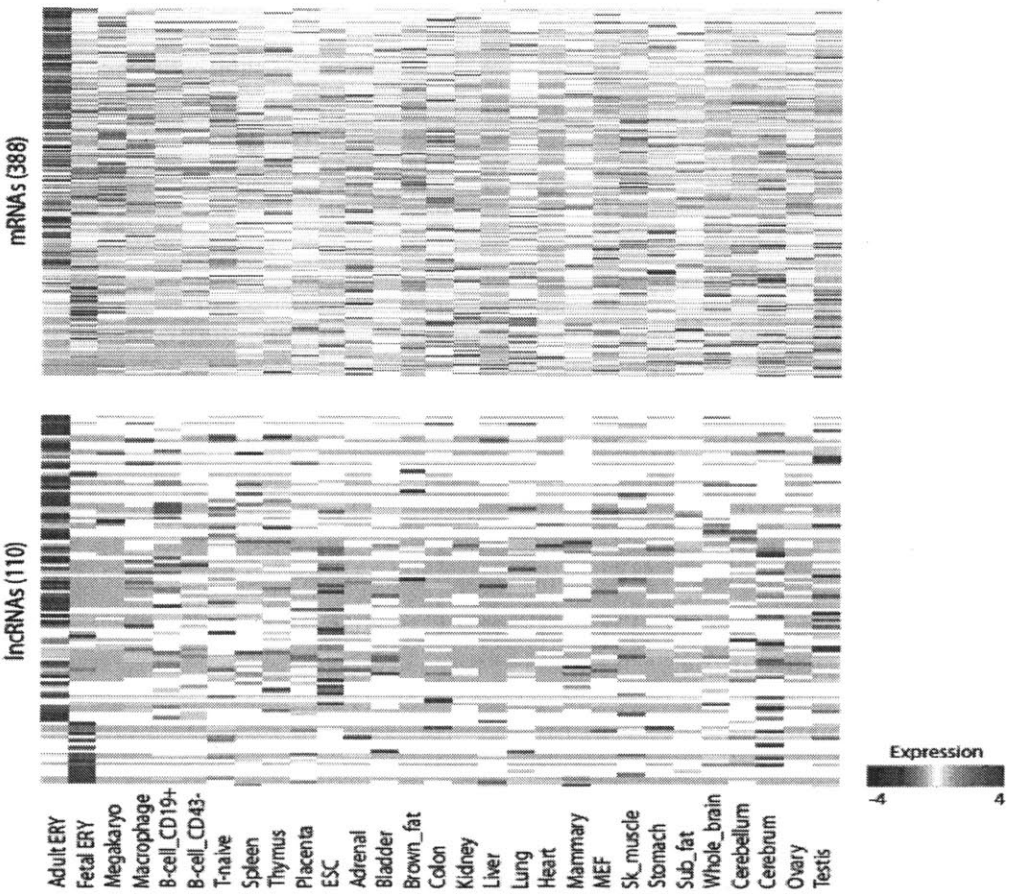


Figure 3-5: Top: Heatmap of 388 differentially expressed mRNAs and Z-score of their expression level in different tissues. Bottom: Heatmap of 110 differentially expressed lncRNAs and Z-score of their expression level in different tissues. Differentially expressed lncRNA genes appear to be more specific to bone marrow or fetal liver erythroid progenitors than differentially expressed mRNA genes.



scripts. Filtering by expression level in known contaminating tissue and cell lines proved useful for eliminating many contaminating transcripts, but was not effective in removing all suspected contaminants. Future work will analyze DNase-I hypersensitivity (DHS) sites at the promoters of differentially expressed genes in bone marrow and fetal liver in order to determine if there is evidence that the differentially expressed genes are really being expressed as the chromatin in their promoters is open in erythroid progenitor cells and not a contaminating cell. For example, the gene Afp (fetal liver contaminant) lacks a DHS site within +/- 2kb of its transcription start site, indicating that it may not actually be transcribed in fetal liver erythroid progenitors. Additionally the gene Irf4 (bone marrow contaminant) lacks a DHS site within +/- 2kb of its transcription start site. It will be useful for other RNA-seq experiments in the future to combine their results with DNaseI-seq experiments in order to determine more reliably if largely expressed genes are in actuality coming from the cell type desired. The analysis so far has found that many solute carriers are differentially expressed between fetal liver and bone marrow, most interestingly Slc2a1 (Glut1). lncRNA genes show more differential expression compared to mRNAs between fetal liver and bone marrow and more tissue specificity than mRNAs, consistent with previous findings about lncRNAs.

3.3 Discussion of ChIP-seq Analysis

In order to determine if differences in promoter or enhancer elements were driving differential expression of genes between fetal liver and adult bone marrow, we analyzed ChIP-seq experiments of fetal liver and adult bone marrow, including ChIP-seq performed for H3K3me3, H3K3me1, and H3K27ac chromatin modifications, TAL1, GATA1, CTCF, RNA polymerase II, and NEF2. Figure 3-6 contains plots of read density near transcription start sites for ChIP-seq and DNase-seq experiments for differentially expressed genes enriched in bone marrow. Chromatin markings for active promoters and enhancers are higher near transcription start sites in bone marrow for bone marrow enriched genes, as expected. The genes ALDH1A1, CTSE, and RSAD2

are all enriched in bone marrow relative to fetal liver. Figures 3-7, 3-8, and 3-9 contain images from the UCSC browser near the genes Aldh1a1, Ctse, and Rsad2, respectively. Tracks from bone marrow-related data are displayed in blue while tracks from fetal liver data are displayed in red. Gray boxes represent areas around promoters or enhancers that have similar chromatin markings between bone marrow and fetal liver, blue boxes surround bone marrow promoters or enhancers with active chromatin markings that do not appear in fetal liver, and red boxes surround fetal liver promoters or enhancers with active chromatin markings that do not appear in bone marrow. The first four tracks at the top are RNA-seq read density from bone marrow positive strand and negative strand and fetal liver positive strand and negative strand, respectively. The order of pairs of tracks after RNA-seq read density with continuous signal, from top to bottom, are DNase-hypersensitivity, Gata1, H3K4me3, H3K4me1, H3K27ac, and H3K27me3. All three genes show H3K4me3 enrichment near their promoters in bone marrow that are lacking in fetal liver. CTSE and RSAS2 additionally show H3K27ac enrichment near their promoters. Interestingly, CTSE appears to have a shorter isoform that is only expressed in bone marrow by RNA-seq read density and chromatin markings.

Figure 3-10 contains plots of read density near transcription start sites for ChIP-seq and DNase-seq experiments for differentially expressed genes enriched in fetal liver. H3K4me3 ChIP-seq read density shows the greatest difference near transcription start sites for fetal liver enriched differentially expressed genes, while DNase-seq and ChIP-seq data for H3K4me1 and H3K27ac do not appear significantly different between bone marrow and fetal liver. The genes Slc2a1 (Glut1), Mt1 (metallothionein 1), and Mt2 (metallothionein 2) are differentially expressed and enriched in fetal liver. Figures 3-11 and 3-12 contain images from the UCSC browser near the genes Slc2a1, and Mt1 and Mt2, respectively. Mt1 and Mt2 have enrichment of H3K4me3 H3K27ac at proximal enhancers in fetal liver that are lacking in bone marrow, suggesting active enhancers are responsible for their differential expression. Slc2a1 contains enrichment for DNase-I, H3K4me3, H3K4me1, and H3K27ac near its promoter and proximal enhancers that are lacking in bone marrow, suggesting that

Figure 3-6: Top row: Heatmap of DNase Hypersensitivity, H3K4me3, H3K4me1, and H3K27ac read density around transcription start site for 275 adult bone marrow-high differentially expressed genes. Bottom row: Graph of DNase Hypersensitivity, H3K4me3, H3K4me1, and H3K27ac read density around transcription start site for 275 adult bone marrow-high differentially expressed genes with fetal liver in green and adult bone marrow in orange. Active chromatin markings are present in higher density in bone marrow.

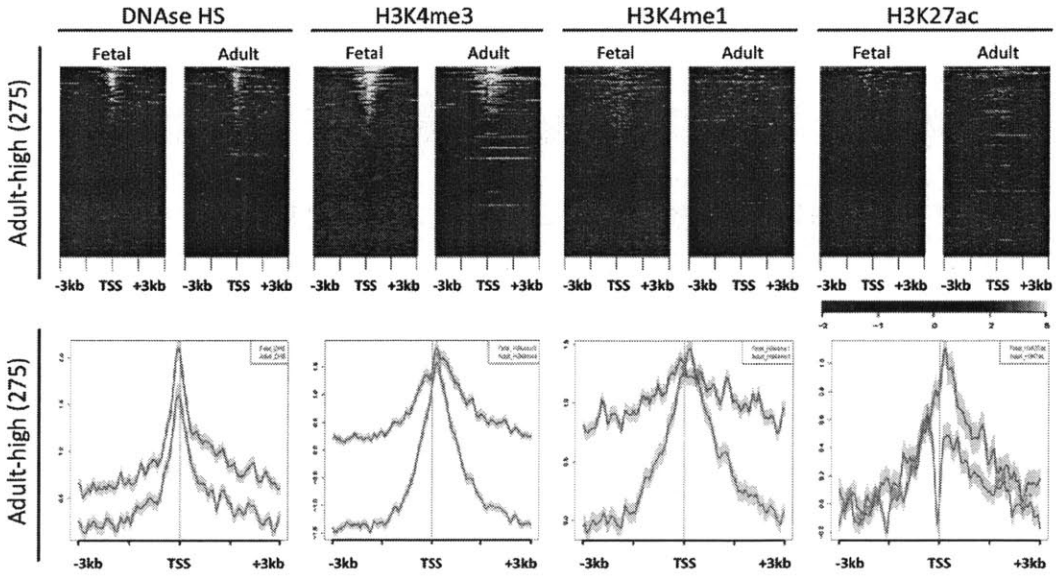


Figure 3-7: Aldh1a1 shows active promoter chromatin markings in bone marrow and not in fetal liver. DNase-hypersensitivity, H3K4me3, and H3K4me1, and H3K27ac enrichment appears at Aldh1a1's promoter in bone marrow and not in fetal liver.

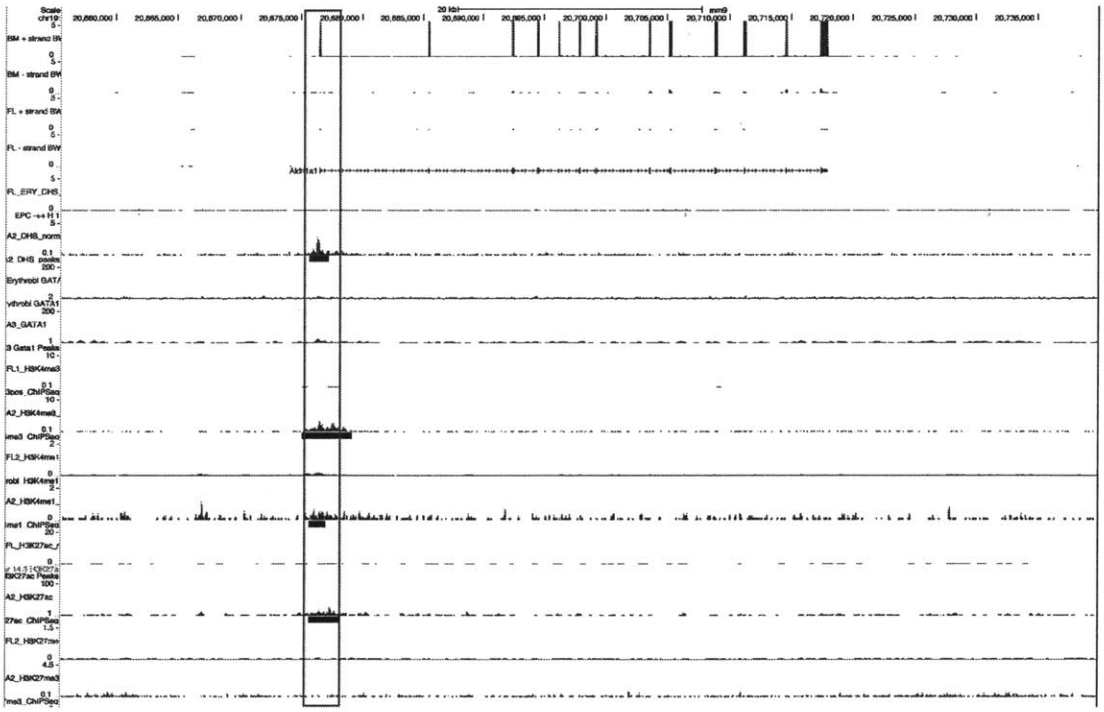


Figure 3-8: Ctse shows active enhancer chromatin markings (H3K27ac) upstream of its transcription start site and active promoter chromatin markings (H3K4me3, H3K4me1, and H3K27ac) at the promoter of a shorter isoform that may only be expressed in bone marrow. Chromatin markings at the promoter of the longer isoforms appear similar between bone marrow and fetal liver.

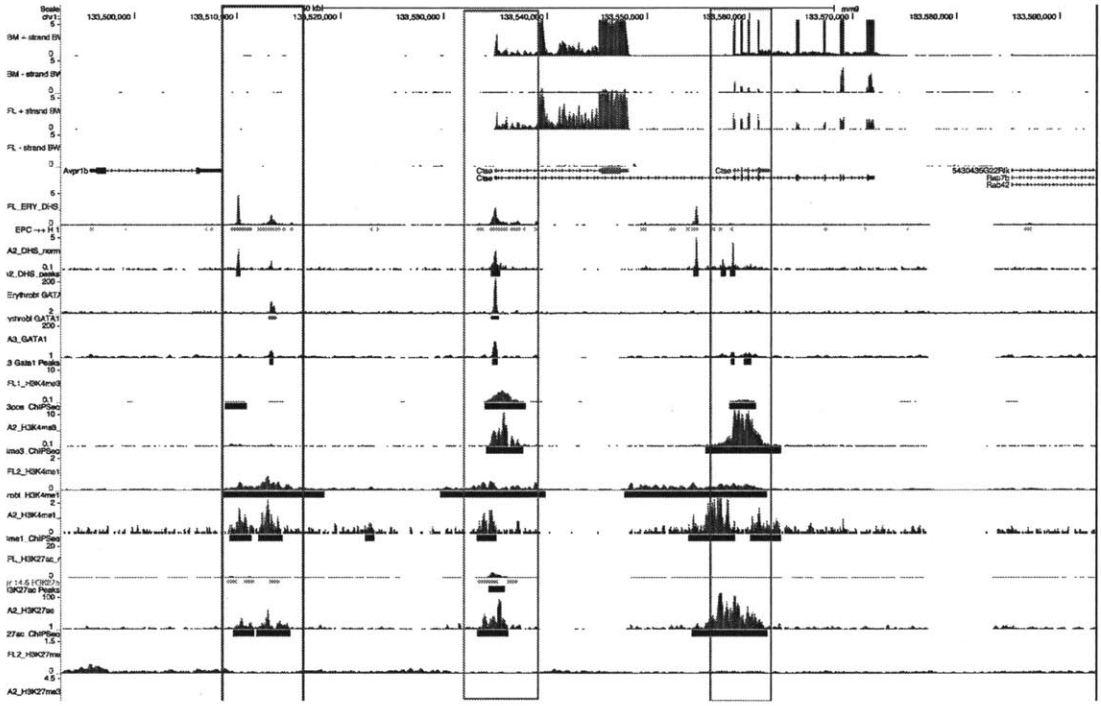


Figure 3-9: Rsad2 shows active chromatin markings (H3K4me3 and H3K27ac) at its promoter in bone marrow that are lacking in fetal liver.

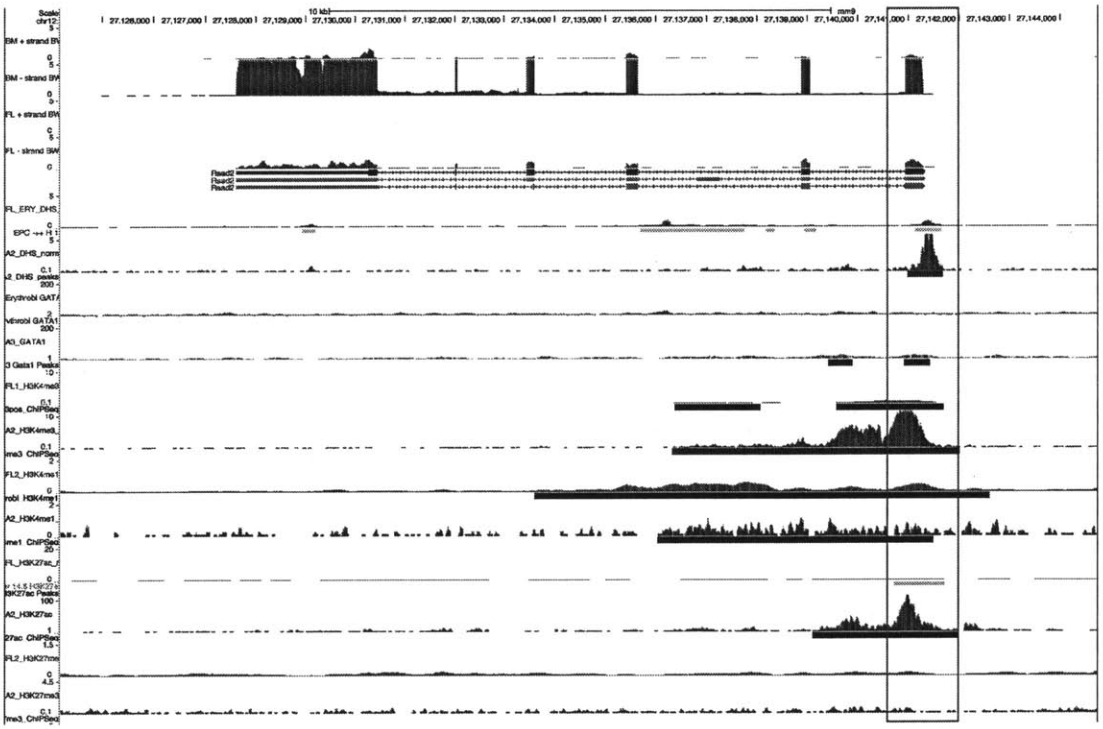
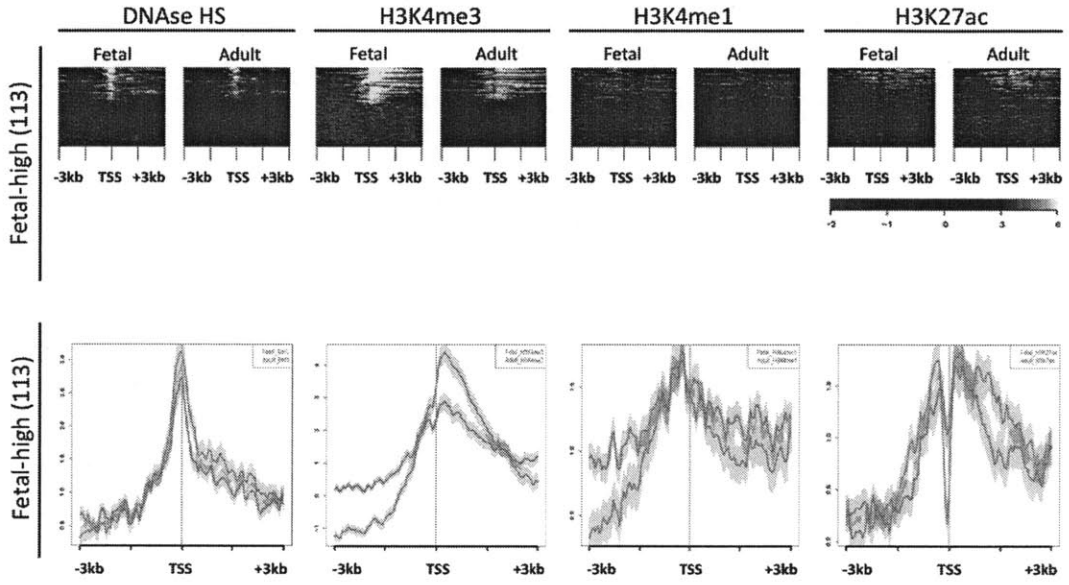


Figure 3-10: Top row: Heatmap of DNase Hypersensitivity, H3K4me3, H3K4me1, and H3K27ac read density around transcription start site for 113 fetal liver-high differentially expressed genes. Bottom row: Graph of DNase Hypersensitivity, H3K4me3, H3K4me1, and H3K27ac read density around transcription start site for 113 fetal liver-high differentially expressed genes with fetal liver in green and adult bone marrow in orange. Active chromatin markings (H3K4me3 and H3K27ac) are present at a higher density near transcription start sites of fetal-high genes in fetal liver than in bone marrow.



active promoters and enhancers are responsible for Slc2a1's differential expression. Interestingly, Gata1 shows enrichment at proximal enhancers upstream of Slc2a1's transcription start site.

The known genes induced during erythropoiesis-induced genes Gata1 and Band3 show similar chromatin enrichment at their promoters and proximal enhancers. Figures 3-13 and 3-14 contain images from the UCSC browser near the genes Gata1 and Band3, respectively. Gata1 has enrichment for H3K4me3, H3K27ac, and Gata1 at its promoter and proximal enhancer in both bone marrow and fetal liver. Band3 also has enrichment for H3K4me3, H3K27ac, and Gata1 at its promoter and proximal enhancer in both bone marrow and fetal liver. Another known erythropoiesis-induced gene Tal1 also has similar enrichment for chromatin marks indicative of active pro-

Figure 3-11: Slc2a1 (Glut1) shows active chromatin markings at promoters and many upstream enhancers as well as Gata1 binding at some upstream enhancers in fetal liver which are completely lacking in bone marrow.

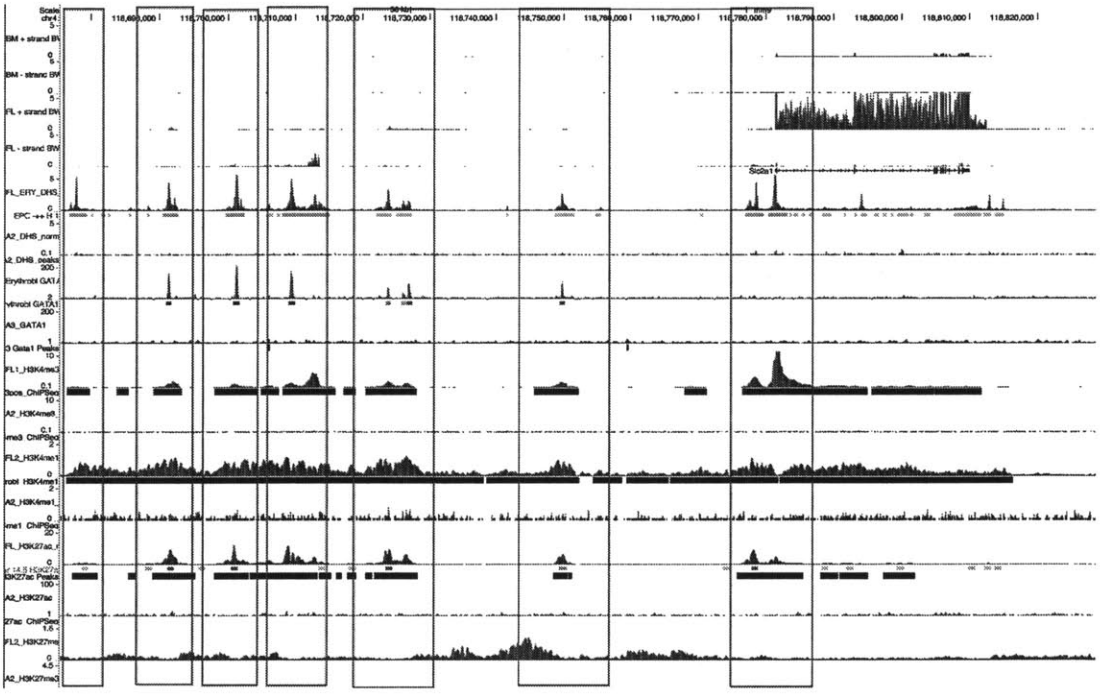


Figure 3-12: Mt1 shows active chromatin markings at downstream enhancers and Mt2 shows active chromatin markings in upstream enhancers that are present in fetal liver yet lacking in bone marrow. Active promoter chromatin markings are present in both Mt1 and Mt2 in both bone marrow and fetal liver.

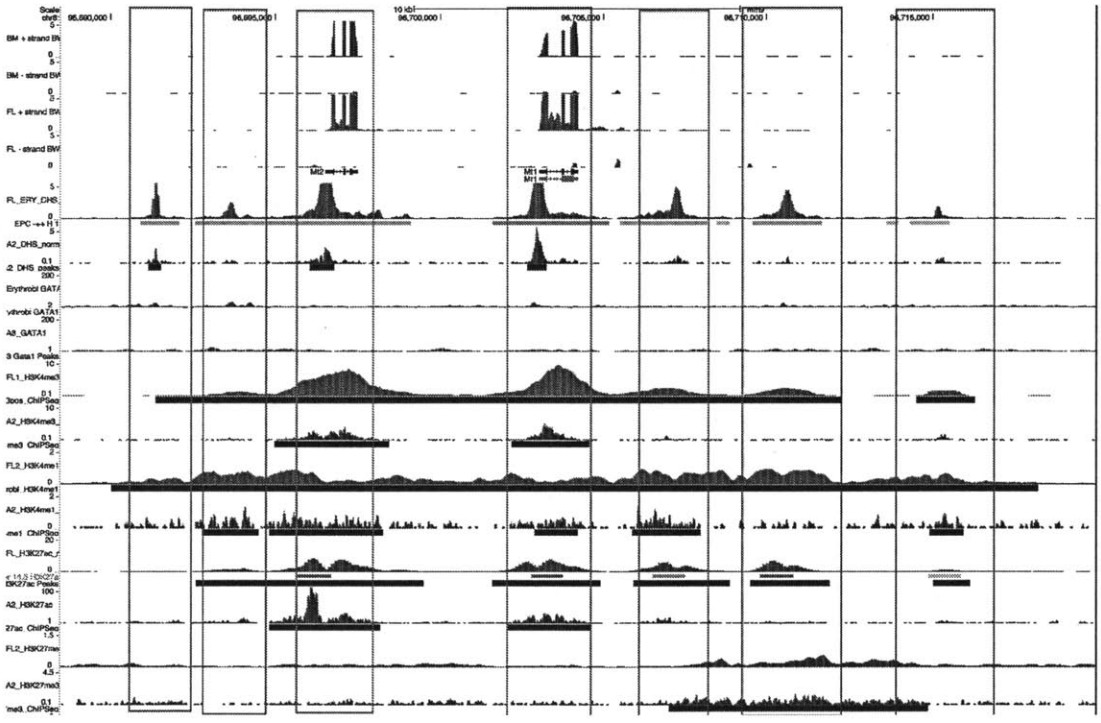
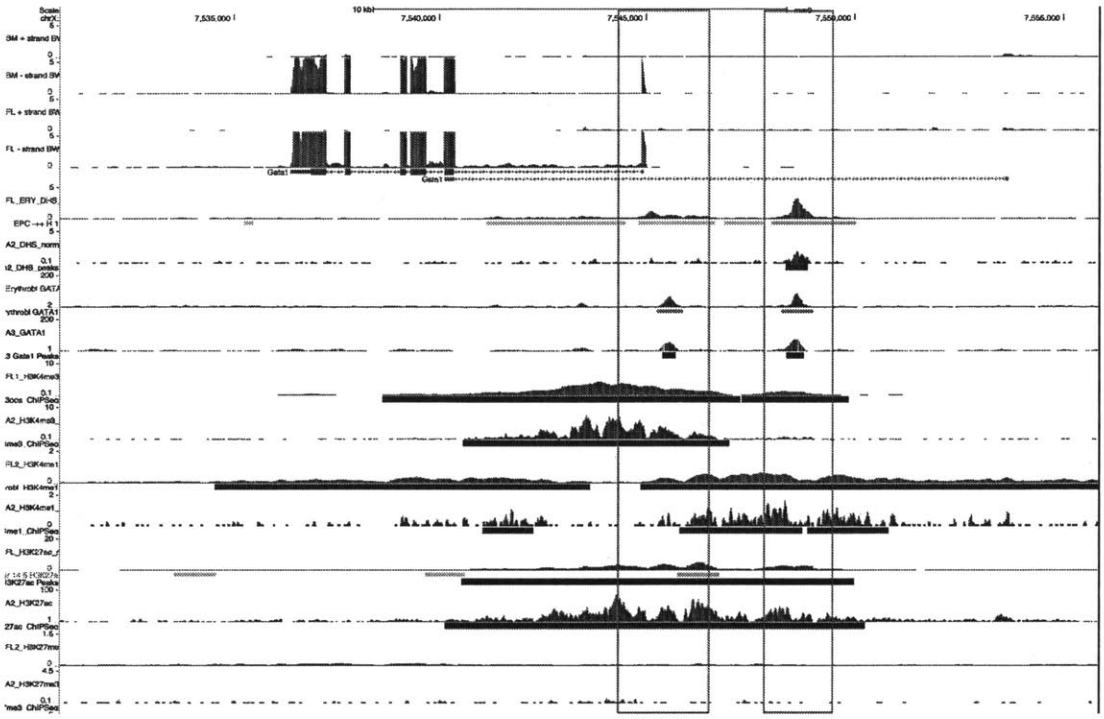


Figure 3-13: Gata1 shows similar chromatin markings between bone marrow and fetal liver at its promoter and upstream enhancer.



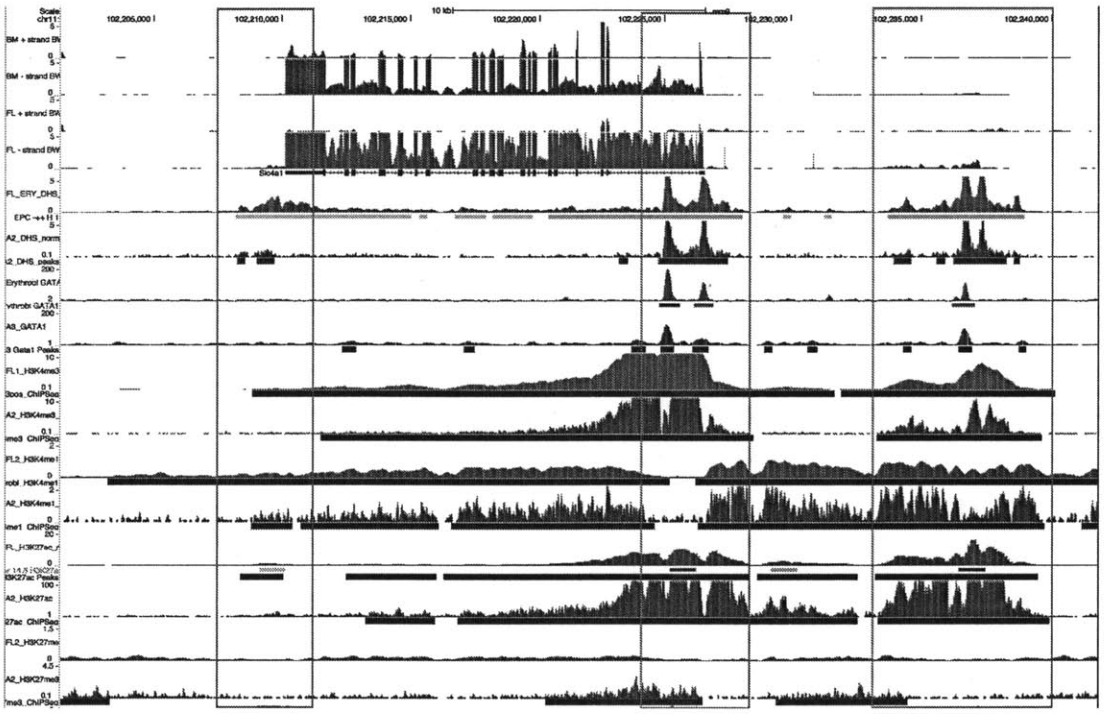
moters and proximal enhancers.

In summary, differentially expressed genes enriched in bone marrow show enrichment for chromatin marks indicative of active promoters near their transcription start sites, H3K4me3 and H3K27ac, at higher levels than fetal liver, implicating promoters drivers of differential expression. Those genes differentially expressed and enriched in fetal liver show enrichment for chromatin marks indicative of active promoters, H3K4me3, at higher levels than bone marrow, again implicating promoters in differential expression. Genes involved in erythropoiesis such as Gata1, Band3, and Tal1 show similar enrichment for chromatin marks indicative of active promoters.

3.4 Discussion of DNase-seq Analysis

DNase-I hypersensitivity sites are places where DNase-I can cleave DNA due to open chromatin. We analyzed DNase-seq I data of fetal liver and adult bone marrow ery-

Figure 3-14: Slc4a1 (Band3) shows active chromatin markings at its promoter, upstream enhancer, and downstream enhancer that are similar between bone marrow and fetal liver.



throid progenitor cells in order to determine areas of DNase-I hypersensitivity near promoters and enhancers of differentially expressed genes. We called peaks or DNase-I hypersensitivity sites using the software Peakranger and found ares within sites with enrichment for JASPAR and TRANSFAC motifs. Further work will analyze enriched motifs in order to determine if there are specific transcription factors driving differential expression of genes between bone marrow and fetal liver. In order to reduce the search space and increase specificity of search for enriched motifs, we analyzed DNase-I footprints within DNase-I hypersensitivity sites as well. Locations where transcription factors bind DNA block DNase-I from cleaving DNA leaving areas of low cleavage inside sites between sites of high cleavage. Figure 3-15 shows an example of a DNase-I footprint near the promoter of the human gene MTPN. Interestingly, the area within the DNase-I footprint shows enrichment for the transcription factor NRF1 as well as conservation among vertebrates by PhyloP scores. We adapted a DNase-I footprint finding algorithm used to find footprints for human cells for mouse. In short, we calculated a per-nucleotide cleavage score by taking the 5' end of each DNase-seq read and found areas within DNase-I hypersensitivity sites where the average cleavage score was high in 3-10 base left and right flanking regions and high in 6-40 base center regions. Although DNase-I footprints make up only 12% of DNase-I hypersensitivity sites in the fetal liver datasets, 53% of enriched JASPAR motifs fall within them. Figure ?? contains an image from the UCSC browser near Slc2a1 which demonstrates that the promoter of Slc2a1 has a called DNase-I footprint with significant enrichment for a number of transcription factors: Ctcf, Egr1, Klf4, Klf5, Sp1, Sp2, and Zfx. Figure 3-16 contains an image from the UCSC browser near Hif3a, another gene differentially expressed and enriched in fetal liver which also exhibits DNase-I footprints overlapping an enriched Ctcf motif. Hif3a additionally contains DNase-I footprints and enriched motifs in its promoter and proximal enhancers. Interestingly, Grb10 (differentially expressed and enriched in fetal liver) contains a DNase-I footprint as well which overlaps an enriched Ctcf motif near a proximal enhancer. Where the enriched motif occurs in all three genes (Slc2a, Hif3a, and Grb10) also contains enrichment for Ctcf by ChIP-seq in fetal liver. While study of more differentially ex-

pressed genes is necessary, preliminary analysis of three differentially expressed genes enriched in fetal liver suggest that while the repressive chromatin mark H3K27me3 is located near all three genes, Ctcf may act as an insulator to prevent transcription silencing of these genes.

3.5 Further Work

Continuing work will analyze enriched motifs more throughly at DNase-I hypersensitivity sites for genes differentially expressed between bone marrow and fetal liver in order to gain insight into transcription factors that may be driving differential expression. Further work will also attempt to filter contaminating transcripts by alternative methods, rather than our expression level highest in a contaminating tissue approach, such as by finding evidence for open chromatin near the promoter of genes in our cells. As highly expressed contaminating transcripts would show up in RNA-seq data as being differentially expressed because the contaminating tissues are different between fetal liver and adult bone marrow, other methods that are sensitive instead to the number of cells rather than transcript abundance such as DNase-seq would be more useful in determining whether or not a transcript is actually expressed in the desired cell line.

3.6 Miscellaneous Work

In addition to the work described here, a website was developed to make it easy to identify potential lncRNAs involved in erythropoiesis. Previous work in our group catalogued the repertoire of lncRNAs active during erythropoiesis, including over 100 previously unannotated lncRNA genes that are often erythroid-restricted. They comprehensively characterized these RNAs by their tissue specificity, expression patters, chromatin state, and TF binding *in vivo*, and integrated these features to select candidates for functional studies. The table located at `http://lodishlab.wi.mit.edu/data/lncRNAs/` contains a table of the lncRNAs identified in the previous work[1]. The

Figure 3-15: Figure 1 (a) from[33]. DNase-seq cleavage exhibits a footprint where there is an enriched motif for NRF1 and evidence of NRF1 binding by ChIP-seq in human K562 cells near the gene MTPN's promoter. The footprint also shows vertebrate conversion by phyloP.

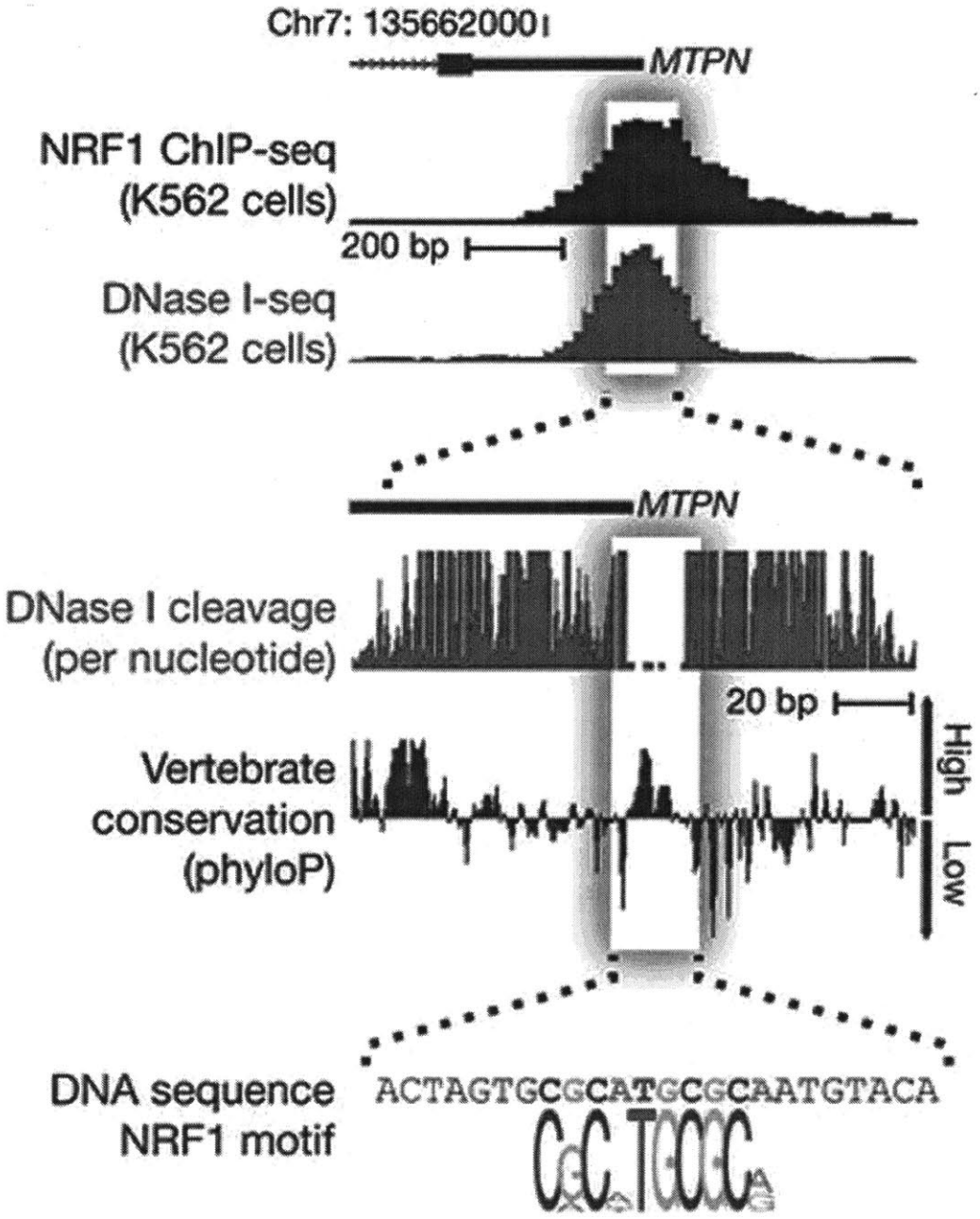
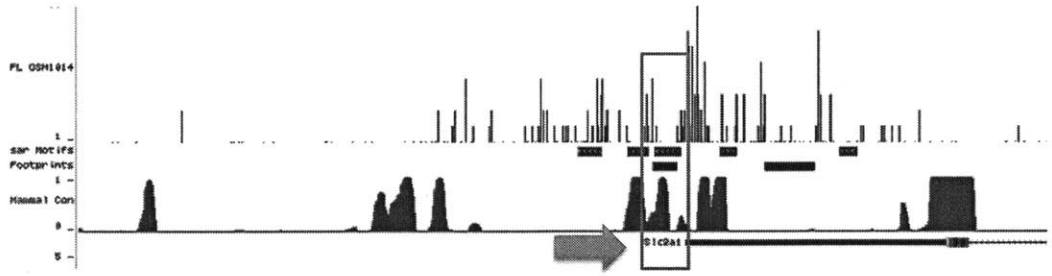


Figure 3-16: Slc2a1 (Glut1) shows a DNase-I footprint overlapping enriched motifs and significant mammal conservation at its promoter. The tracks are DNase-I cleavage, JASPAR motifs, DNase-I footprints, and mammal conservation. The enriched motifs are for the transcription factors Ctcf, Egr1, Klf4, Klf5, Sp1, Sp2, and Zfx[31].



website contains data indicating each lncRNA's locus, if it has CAGE (Cap analysis gene expression) tags, an RPOL II (RNA polymerase II) peak, an H3K4me3 peak, an H3K35me3 peak, an H3K4me1 peak, an H3K27ac peak, a GATA1 peak, a TAL1 peak, a KFL1 peak, and the erythroid tissue specificity score for each lncRNA gene.

Bibliography

[1] Alvarez-Dominguez, J. R., Hu, W., Yuan, B., Shi, J., Park, S. S., Gromatzky, A. A., Oudenaarden, A. v., and Lodish, H. F. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. *Blood*, 123(4):570–581, 2014.

[2] Anguera, M. C., Ma, W., Clift, D., Namekawa, S., Kelleher, R. J., and Lee, J. T. Tsx produces a long noncoding RNA and has general functions in the germline, stem cells, and brain. *PLoS Genet.*, 7(9):e1002248, 2011.

[3] Borodina, T., Adjaye, J., and Sultan, M. A strand-specific library preparation protocol for RNA sequencing. *Meth. Enzymol.*, 500:79–98, 2011.

[4] Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, 25(18):1915–1927, 2011.

[5] Cantor, A. B. and Orkin, S. H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*, 21(21):3368–3376, 2002.

[6] Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigo, R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9):1775–1789, 2012.

[7] Dethoff, E. A., Chugh, J., Mustoe, A. M., and Al-Hashimi, H. M. Functional complexity and regulation through RNA dynamics. *Nature*, 482(7385):322–330, 2012.

[8] Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E.,

Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L. H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T. R. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[9] Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P., and Kohtz, J. D. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.*, 20(11):1470–1484, 2006.

[10] Feng, X., Grossman, R., and Stein, L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics*, 12:139, 2011.

[11] Finn, R. D., Clements, J., and Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, 39(Web Server issue):29–37, 2011.

[12] Geisler, S. and Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, 14(11):699–712, 2013.

[13] Gibb, E. A., Vucic, E. A., Enfield, K. S., Stewart, G. L., Lonergan, K. M., Kennett, J. Y., Becker-Santos, D. D., MacAulay, C. E., Lam, S., Brown, C. J., and Lam, W. L. Human cancer long non-coding RNA transcriptomes. *PLoS ONE*, 6(10):e25915, 2011.

[14] Gish, W. and States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.*, 3(3):266–272, 1993.

[15] Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., Zornig, M., MacLeod, A. R., Spector, D. L., and Diederichs, S. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.*, 73(3):1180–1189, 2013.

[16] Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227, 2009.

[17] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28(5):503–510, 2010.

[18] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28(5):503–510, 2010.

[19] Hattangadi, S. M., Wong, P., Zhang, L., Flygare, J., and Lodish, H. F. From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood*, 118(24):6258–6268, 2011.

[20] Hu, W., Yuan, B., Flygare, J., and Lodish, H. F. Long noncoding RNA-mediated anti-apoptotic activity in murine erythroid terminal differentiation. *Genes Dev.*, 25(24):2573–2578, 2011.

[21] Huang, d. a. W., Sherman, B. T., and Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, 2009.

[22] Huang, d. a. W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[23] Kerenyi, M. A. and Orkin, S. H. Networking erythropoiesis. *J. Exp. Med.*, 207(12):2537–2541, 2010.

[24] Kingsley, P. D., Greenfest-Allen, E., Frame, J. M., Bushnell, T. P., Malik, J., McGrath, K. E., Stoeckert, C. J., and Palis, J. Ontogeny of erythroid gene expression. *Blood*, 121(6):e5–e13, 2013.

[25] Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., and Gao, G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, 35(Web Server issue):W345–349, 2007.

[26] Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., McGowan, S. J., De Gobbi, M., Hosseini, M., Vernimmen, D., Brown, J. M., Gray, N. E., Collavin, L., Gibbons, R. J., Flint, J., Taylor, S., Buckle, V. J., Milne, T. A., Wood, W. G., and Higgs, D. R. Intragenic enhancers act as alternative promoters. *Mol. Cell*, 45(4):447–458, 2012.

[27] Kretz, M., Webster, D. E., Flockhart, R. J., Lee, C. S., Zehnder, A., Lopez-Pajares, V., Qu, K., Zheng, G. X., Chow, J., Kim, G. E., Rinn, J. L., Chang,

H. Y., Siprashvili, Z., and Khavari, P. A. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev.*, 26(4):338–343, 2012.

[28] Li, L., Liu, B., Wapinski, O. L., Tsai, M. C., Qu, K., Zhang, J., Carlson, J. C., Lin, M., Fang, F., Gupta, R. A., Helms, J. A., and Chang, H. Y. Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Rep*, 5(1):3–12, 2013.

[29] Lin, M. F., Jungreis, I., and Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–282, 2011.

[30] Lodish, H., Berk, A., Kaiser, C., Krieger, M., Bretscher, A., Ploegh, H., Amon, A., and Scott, M. *Molecular Cell Biology*. Molecular Cell Biology. W. H. Freeman, 2012. ISBN 9781429234139.

[31] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 42(Database issue):D142–147, 2014.

[32] Mercer, T. R., Dinger, M. E., and Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, 10(3):155–159, 2009.

[33] Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. A., Groudine, M., Kaul, R., and Stamatoyannopoulos, J. A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

[34] Ponting, C. P., Oliver, P. L., and Reik, W. Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–641, 2009.

[35] Rice, P., Longden, I., and Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16(6):276–277, 2000.

[36] Rinn, J. L. and Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.*, 81:145–166, 2012.

[37] Trapnell, C., Pachter, L., and Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[38] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, 2010.

[39] Wang, K. C. and Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, 43(6):904–914, 2011.

[40] Wapinski, O. and Chang, H. Y. Long noncoding RNAs and human disease. *Trends Cell Biol.*, 21(6):354–361, 2011.

[41] Wong, P., Hattangadi, S. M., Cheng, A. W., Frampton, G. M., Young, R. A., and Lodish, H. F. Gene induction and repression during terminal erythropoiesis are mediated by distinct epigenetic changes. *Blood*, 118(16):e128–138, 2011.

[42] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.