HIGH SCHOOL EDITION
Journal of Student Research

Characterizing Pathogenic Enhancer Activity at Single-cell Resolution

Arul Loomba[1] and Sohum Mehta[1]

[1]Rancho Cucamonga High School

ABSTRACT

The ZRS enhancer, a regulatory sequence found in numerous organisms, plays an important role in early embryonic limb development. ZRS controls the expression of the Sonic Hedgehog gene (Shh), and therefore early limb development in an organism as Shh has been shown to control the width of the limb bud by stimulating mesenchyme cell proliferation due to its ability to regulate the anterior-posterior length of the apical ectodermal ridge. Several transcription factors, acting as repressors or activators of the Shh gene, coordinate this limb development process in tandem with the ZRS enhancer. While the significance of normal ZRS activity is evident, this study looks deeper into the effects of pathogenic changes to the ZRS enhancer and the development of associated limb disorders such as preaxial polydactyly (PPD) by focusing on several aspects of ZRS regulation and its relation to Shh expression. This was accomplished by characterizing the expression of Shh and mCherry, an introduced luminescence gene regulated by ZRS, through single-cell RNA sequenced cells from a developing limb bud of a mouse embryo. Additionally, this study characterized specific transcription factors as potential repressors or activators of ZRS by determining TF enrichment or depletion in highly expressive Shh and mCherry cells. Classifying such TFs is vital in identifying the regulatory elements that control the formation of limb malformation disorders such as preaxial polydactyly and aid in the development of future therapeutic interventions.

1 Introduction

1.1 Enhancers

Enhancers are a type of cis-regulatory sequences (cis-rs) which control the spatial and temporal aspects of gene expression [7, 14]. Enhancers increase gene expression by regulating promoters independent of their location and orientation [13]. Promoters function as a binding site for RNA Polymerase II, enhancers, and other regulatory proteins (Figure 1. Once bound, Pol II transcribes genetic DNA into RNA, a form which can be detected by scRNA techniques. Most 1 enhancers and promoters are found together in topological associated domain (TAD) and are DNA regions that have separate environments or separate enhancers, promoters, and TFs controlling them [4]. TADs play an important role in increasing the efficiency of gene regulation. Transcription for most genes are only effective when enhancers interact with promoters. The average enhancer is not long, mapping close to 20-50 kb in vertebrates and about 10 in D. Melanogaster [8]. Fully understanding how enhancers boost transcription is a biological challenge with far-reaching implications in the field of genetics and evolution.

Enhancer specificity is mediated by the binding of proteins known as transcription factors. Most enhancers have specific binding sites or recognition motifs for these TFs [8, 14]. Both an enhancer's affinity for a TF and the number of binding sites available effects TF occupancy [9-10]. At certain developmental stages, the binding of multiple specific transcription factors (including repressors or activators), either dependent on the DNA sequence or intrinsic affinity, allows for activation or repression of a gene in terms of its expression (Figure 1). Combinatorial binding of TFs can occur at different stages of development, allowing many TFs to occupy multiple enhancers with a correlating motif [20]. TFs may not be distinct for all genes, therefore making the main cause of differential gene regulation the specific combination of TFs at an enhancer which was investigated within this study.

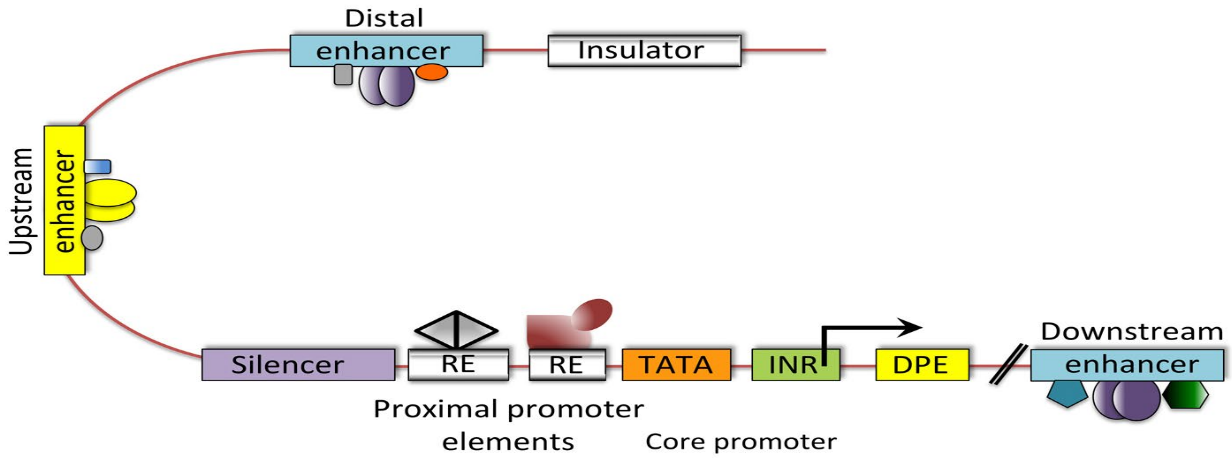

Figure 1: Transcription factors interact with a distal enhancer to promote or repress gene activity. Together, these components of the transcription initiation complex regulate gene expression. [16]

1.2 Normal and pathogenic limb development

The central theme of this paper is to draw a connection between enhancer function, transcription factor enrichment/depletion, and phenotypic limb changes. The limb acts as a model for vertebrate development, with precise control of complex genomic pathways required for proper limb formation [8, 17, 20]. Enhancers control the spatiotemporal aspects of genes and proteins thereby making them a crucial part of limb developmental. DNA comparisons between different animals have proved the existence of similar proteins which regulate limb development. The activity of such proteins can help determine the placement of limb axes and digits during embryonic development. One protein responsible for axis formation in young animal embryos is the Sonic Hedgehog (Shh) protein, named after a mutation in D. melanogaster. Shh plays an important role in a limb-bud regulatory region of vertebrates which is known as the zone of polarizing activity (ZPA). The ZPA region, a specialized mesenchyme cluster, is located directly in the posterior portion of the limb. The presence of Shh in the ZPA regulates the development of the limb along the anterior-posterior axis (running from digit I to V) [7] (Figure 2). Furthermore, Shh expression is regulated by the ZRS enhancer where normal ZRS regulation results in the asymmetrical expression of Shh in the posterior of a developing limb bud which is essential for proper limb development. However, experimentation of the ZRS enhancer sequence from different genomes have resulted in many identifiable phenotypic variations among transgenic organisms, implying the importance the ZRS enhancer has upon the early limb formation phases in the development of an organism.

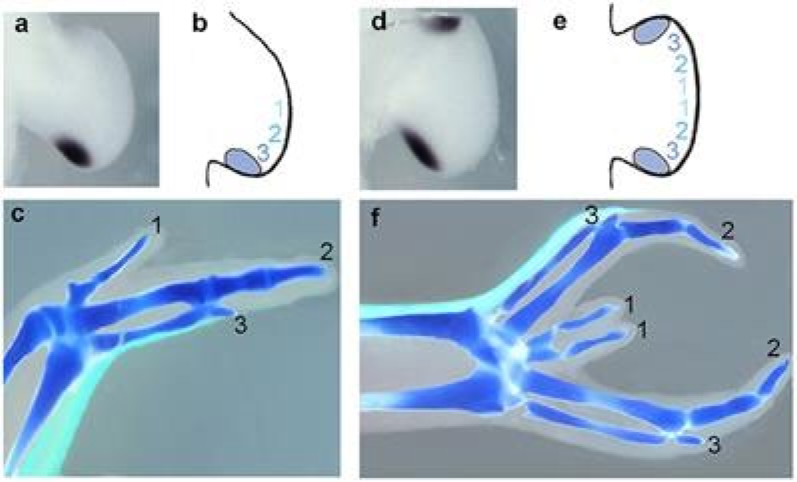

Figure 2: Normal ZRS enhancer activity results in normal Shh expression in the posterior limb bud. This also results in normal digit development (bottom left). A single mutation in ZRS can result in ectopic Shh expression and polydactyly (bottom right) [22]

Limb development is an easily modifiable element of an organism's growth. A limb forms from the collective activity of the genome and develops from the production of essential proteins like Shh. Mutations in the genome can lead to a change in gene regulation and overall protein production. Limb malformations and changes in digits primarily result from mutations in enhancers. Most enhancers reside in non-coding DNA (ncDNA), places where mRNA is not transcribed. Regarding congenital transformations, mutations and changes often happen in ncDNA rather than genes. Studies of transgenic mouse embryos with mutations in non-coding areas such as introns and repetitive sequences have resulted in organisms with bodily modifications, including the wrong placement of digits. Since 90% of enhancers have their sequence in ncDNA, the latter has proven to be an important source of genomic pathogenic activity. In most cases, changes to enhancer regions are less identifiable due to enhancer redundancy [6]. However, despite the presence of some redundant buffers, not all enhancers can remain non-pathogenic, especially those that lack redundancy like ZRS. This enhancer regulates the expression of Shh, an important protein for limb axis development, in the posterior mesenchyme of the zone of polarizing activity (ZPA) [3, 7, 14]. More than 21 human mutations have been identified in the ZRS enhancer which results in congenital disorders, especially preaxial polydactyly (PPD) and extra digits [3-4, 5]. PPD is a common hereditary limb malformation in the hands or feet (Figure 2). This study identified how pathogenic enhancer ability, or point-mutations in enhancers, can result in a congenital malformation like PPD [10] (Figure 2). One commonly studied mutation of ZRS is a small point mutation called the Cuban mutation which results in the ectopic activity of Shh. Mice embryos with the "Cuban" variant, similar to a mutation found in humans, were found to exhibit strong enhancer activity in the anterior limb bud. Instead of exclusive expression in the posterior of the developing limb, Shh is also expressed in the anterior portion of the limb. Consequently, ZRS directly alters the expression of Shh, giving it a tight control over posterior anterior axis formation. Research on the ZRS enhancer has also provided insight beyond the enhancer's pathogenic activity, edging into evolutionary history. For example, the evolution of tetrapods from lungfish is consistent with novel enhancers like ZRS, which regulate limb outgrowth, and proteins like Shh and fibroblast growth factors which determine limb axes. Modifications of DNA and chromatin structure can thus cause morphological shifts for many species

1.3 Genomic structuring

Before enhancers interact with a promoter, they are bound by activator or repressor proteins called specific transcription factors. These TFs interact with a mediator complex, which then recruits pol II and general transcription factors to begin transcribing a gene. This transcription initiation complex is essential to robust and specific gene regulation [12]. Enhancers can also be found within introns (non-coding DNA) and at far distances. The ZRS enhancer, for example, is found in the intron of an unrelated Lmbr1 gene [8], located 1 million nucleotide bases upstream from the Shh promoter. Despite this distance, ZRS can proactively interact with its promoter and transcription factors. The most common way for an enhancer to interact with its promoter is through looping [3]. The looping model proposes that two physically distinct sites interact with one another through the extrusion of DNA. While this model proves efficient for most genes, mutations in enhancers can lead to changes in loop formation. For example, a mutation in a TF motif region can cause increased binding of a TF, and therefore an enhancement to gene activity, which can prove to be beneficial or harmful. In the case of the ZRS enhancer, this type of mutation can cause critical changes, such as shifts in spatiotemporal activity of Shh. This pathogenic activity is noticeable when small mutations were inserted into the enhancer sequence or when the entire enhancer was deleted, resulting in faulty regulation and ectopic expression of Sonic Hedgehog in both cases [19, 7, 9]. Moreover, supporting this theory, is previous research in which mice without the ZRS enhancer form without limbs [18]. Without any regulation of Shh, the limb fails to develop beyond its bud. This odd case, and many undiscovered others, adds to the ambiguity of enhancer specificity and control.

This paper intends to determine the cell types in the developing mouse limb bud and the spatially specific expression of Shh and mCherry genes in these cell types. A list of 292 unique transcription factors was used to determine TF enrichment in Shh and mCherry cells. The list was shortened to only the TFs significant to the regulation of the two genes and characterize the enriched TFs and depleted TFs as potential activators or repressors. While this list is important, it fails to accurately identify the TFs impacted by a mutation in ZRS. mCherry was used as a detector of modified TFs. mCherry, a gene misregulated by ZRS due to mutation, is an appropriate target for determining the TFs potentially responsible for preaxial polydactyly.

2 Methods

2.1 Data collection

Single-celled RNA sequencing data was a major component of this research. Cells from mouse-limb buds were used as a repository of information and were sequenced to create a cell matrix. This matrix highlights the 10,000+ genes in the mouse genome and the differential expression of said genes in different cells. To create and log this data effectively, a drop-sequencing technique was used. 10x genomics is a drop-sequencing technique in which bar-coded beads are combined with cells and reagents to form small droplets. These droplets are isolated and eventually form an emulsion of droplets. 10X genomics is more effective than other sequencing techniques because beads and cells are not loaded at small concentrations, and more than 90% of the droplets contain only 1 bead. Once the emulsion is formed, reverse transcription of the cellular DNA occurs within the droplets. Once all droplets have been reverse-transcribed from RNA to DNA, PCR is used to amplify the DNA and construct a genetic library - all in a single tube. 10X genomics, therefore, acts as a quick way to do scRNA research in a small space that can hold more than 100,000 cells.

2.2 Computational analysis

For the purpose of this research, scRNA-seq data was already collected from a mouse limb bud. To further analyze this data and convert it into something readable, the data was cleaned through various steps of quality control.

scRNASeq initially returned the data in a 31054 (genes) x 19835 (cells) matrix; this data was in a general matrix which needed further scaling and normalization.

First, several RStudio packages including the Seurat package were imported. From the Seurat package, specialized for data control and analysis, the Read10X() function was used to return a unique molecular identified (UMI) count matrix of the initial cells. Inside this matrix, the values represent the number of molecules, creating a data-set with 600 million elements. Next, the CreateSeuratObject() function was used to create a container, matr, to hold the data and for further analysis. The purpose of quality control is to select cells for further analysis. Filtered cells include low-quality cells, cell doublets, dying cells, and those with an excess of mitochondrial gene expression. For the latter, cells that express a large number of mitochondrial genes generally are less representative of pathogenic genome activity.

The data was filtered through several computational steps. Initially, genes with unique features were isolated using the subset() function. These unique features were based upon featureRNA levels above and below a certain point - 200 and 5000. Additionally, cells were filtered according to their mitochondrial counts: those with greater than 10% mitochondrial DNA would be considered dying and useless cells, unnecessary for analysis. A visualization of a violin plot (Figure 3) helped quantify the parameters needed to narrow the data.

After removing some of the unwanted cells - those dying or highly variableand features from the dataset, the next step was narrow it even further; the data was normalized. This was done using the NormalizeData() function. This function balanced the amount of feature expression in all cells, allowing similar cells to have similar unique feature values; to make this more effective, the sequenced Feature RNA was scaled by a factor of 10,000. Both these functions helped in quality control as they grouped cells which seemed different at a low scale, but were fairly similar at a larger scale. Moreover, it removed the cells which were still different at a high scale - insignificant and variable. Next, 2000 of the most highly variable genes were isolated for future analysis. From these highly variant features, a plot displaying the standardized variance of the top 2000 genes was created. (Figure 4).

The next step was linear transformation. The ScaleData() function performed the linear transformation and scaled the matrix containing the 2000 variable genes. This function shifted the expression of each gene, making the mean expression across cells 0, and the variance across cells to be 1. This would give equal weight to further analysis and would prevent highly-expressed genes from dominating. Both a scale and linear dimensional reduction on the data were performed. A Principal Component Analysis (PCA) technique on the variable features allowed for gene grouping (about 50 dimensions). Each cluster holds positive and negative markers, and can then be visualized using several 6 techniques like dimensional loading, dot plots, and heat maps. Overall, dimensional reduction techniques further compressed the data into a readable and viewable form.

Each Principal Component represents a meta-feature of data that clusters cells based on their PCA scores. A PC essentially combines information across correlated feature sets, making the top PCs a robust compression of the dataset. For this project, 100 components were chosen. The JackStraw() function completed this step as it permuted a subset of the data and create a null distribution of feature scores. Simply, the most ’significant’ PCs with many low p-value features were considered better for the study.

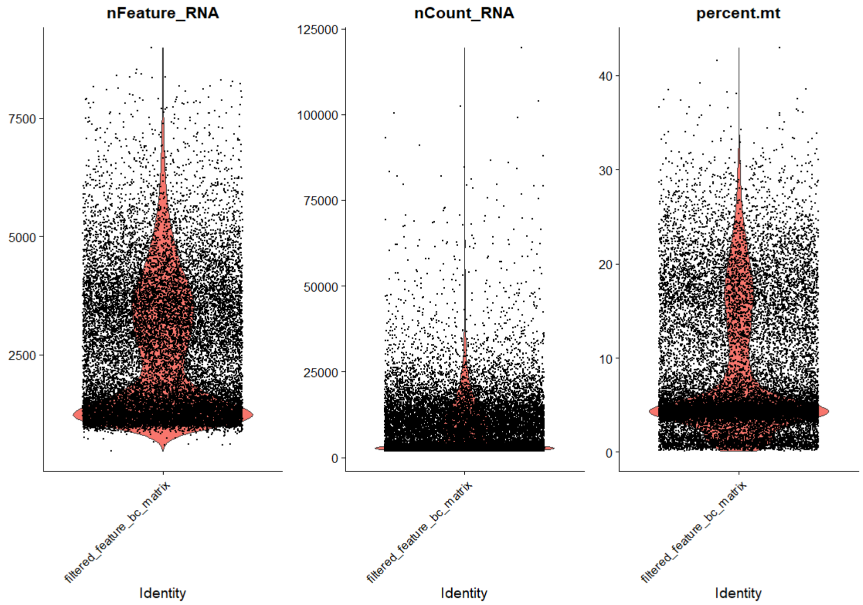

Figure 3: This plot is created using the VlnPlot() function, creating an initial distribution that compares several features. The plot on the right compares feature RNA (x-axis) to the number of cells in which the feature RNA is expressed. The central plot shows all the RNA detected and the number of cells it is observed in. The left-most plot shows the percent of mitochondrial RNA detected in each cell.

2.3 Constructing the cell map

The first step to constructing a cell map was to group the PCs into clusters. Based on gene expression levels and feature-RNA dimensional reduction 12 significant clusters were determined in the mouse-limb bud. Later these 12 clusters were grouped to become 12 cell clusters or cell lineages in the developing limb. While there were 12 cell clusters, many of the cell types overlapped (discussed later in results). Using the most significant genes for each cluster was important in naming the clusters.

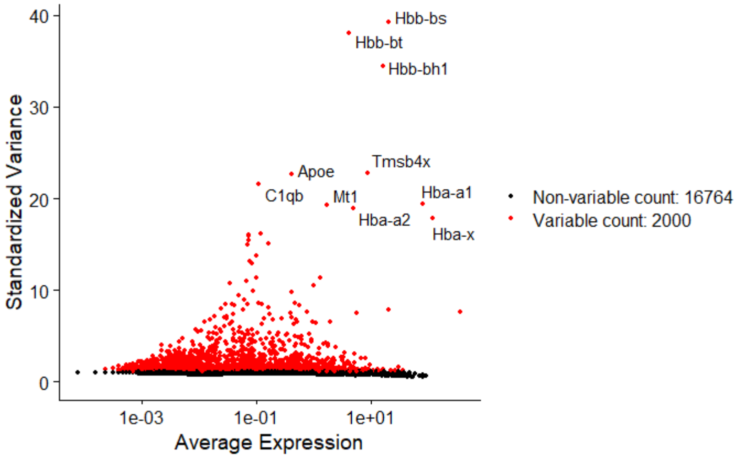

Figure 4: Isolating the 2000 most highly variable genes of the 18000 identifiable genes. Labeled variable counts are the top ten most highly variable genes and their standardized variance.

The FindMarkers() function was used to find the most significant genes in each cluster. The grouped genes were then arranged according to a value called avg log2FC which compares the expression of a gene within a certain area to outside it (higher value = more significance). This value, along with the list of most significant genes, helped in creating a new grouping of genes. These groupings and previous publications about specific genes were used to determine the cell names. However, due to the young age of the embryo, 11.5 days, many of the cell clusters were similar in differentiation and cell lineage, like mesenchymal cells; most of the cell clusters were undifferentiated. As such, cluster 0, which had varying levels of expression for many genes, was unidentifiable by manual means and a computational program. Once the primary cell clusters were determined, dimensional heat maps like the one seen in Figure 5 were used to confirm the guesses; most were confirmed.

2.4 Visualizing mCherry and Shh expression

In order to create accurate visualizations of mCherry and Shh expression and localization, a threshold was set to more accurately classify and label cells that expressed mCherry and Shh. Cells that expressed a count value greater than equal to 2 were visualized while cells that only expressed Shh and mCherry once were discarded in hopes of removing cells that did not really express mCherry and/or Shh. This was done to get a better understanding of which cell types, and inherently cell clusters, exhibited the greatest expression of Shh and mCherry which would be an important part in finding the pathogenic mCherry markers. For both cell types, cells with a gene expression count of 2 or greater were significant. To let this happen, a Count Matrix was created containing all cells expressing either gene (repeated twice for Shh and mCherry). This count matrix was then grouped according to count, with the variable like "mCherry(or Shh) count >= 2".

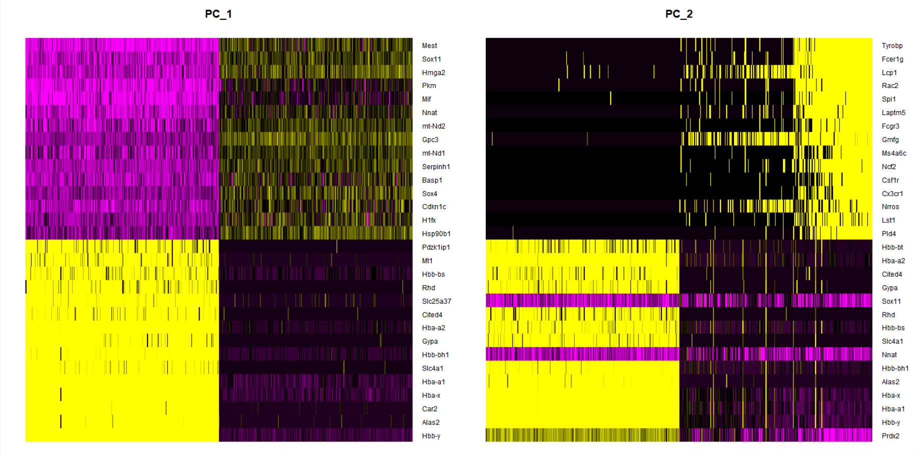

Figure 5: Shows PC clusters, 1 and 2, in which the expression levels of the highly variable genes are compared to the cells. The bright yellow color suggest high expression, while purple and black colors suggests little to no expression.

A feature plot was used to visualize the cells, either with a count greater than 2 or less than 2. Figure 7 represents the expression of mCherry across the 12 cell clusters with the specified count. Figure 6, shows a similar plot but represents the expression of Shh in cells at a count greater than or equal to 2. Both these figures were created using the DimensionalPlot() function in Seurat and were grouped according to RNA counts of specific genes. These plots were useful for future comparison of TF expression in Shh and mCherry cells (Figures 12b, 13b, and 14b). The indexes containing the two types of counts were "mCherry 2" and "Shh 2".

2.5 Utilizing transcription factor data

One of the final steps to characterize the potential enriched and depleted transcription factors was to utilize the original TF matrix. At the beginning of this study, about 292 potential TFs were identified to be present in the mouse limb bud; the goal was to narrow this list. The “mCherry 2” and “Shh 2” indexes were used as parameters for the FindMarkers() function. Another main parameter for this function was Avg log2FC. A higher value correlates with a more unique expression of a gene, making it more significant to specific cells and clusters. The base threshold for most functions was a value of .25.

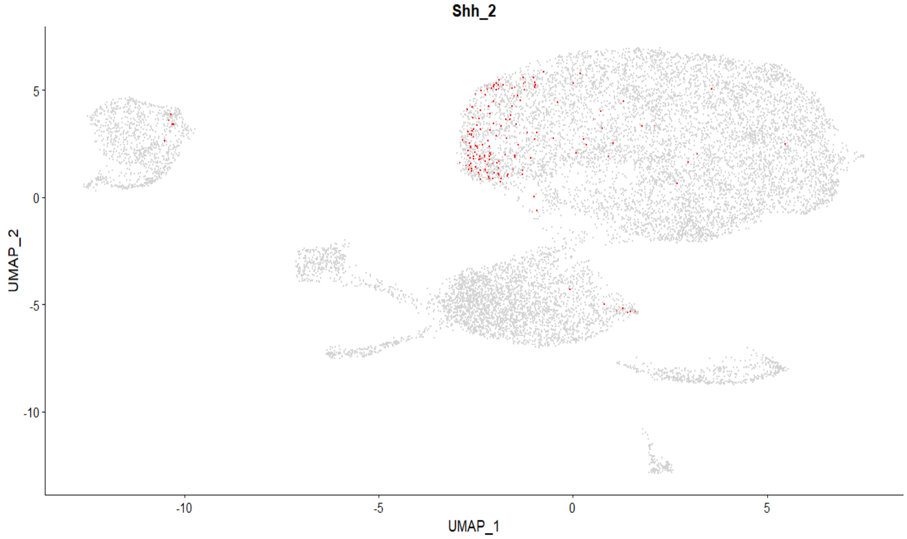

Figure 6: This plot shows the Shh and non-Shh cells in all 12 clusters. Each cell is classified as an Shh cell when it has an Shh expression count of over two, Shh count¿=2, and is a non-Shh cell for anything below that, Shh count¡2. This plot can be used to isolate TFs.

In the FindMarkers() function, the TF genes were used as features and were grouped according to their avg log2FC values. The first analysis found the enriched TFs for Shh and mCherry, and the second found the depleted TFs for Shh and mCherry cells. 19 TFs were enriched in Shh cells (Table 1) and 12 were enriched specifically in mCherry cells (Table 3). After switching a few parameters in the FindMarkers() function, including count values, 26 TFs were depleted at a Avg log2FC threshold of .25 in Shh cells. In mCherry cells, about 13 TFs were depleted (Table 4). A detailed analysis of these enriched and depleted transcription factors was a core result.

Once finalizing the list of enriched and depleted TFs in normally regulated Shh and mCherry cells, the pathogenic and misregulated cells from the data became the main focus. In only cells misregulated by the mutated ZRS were pathogenic mCherry cells. The parameters for identifying these cells were three genes only expressed in the posterior of a normal embryo: Shh, Hoxd13, and Hand2. Any expression of these genes in mCherry cells would be normal. The cells which failed to express any of the three genes, however, would be pathogenic, since they would likely not be in the posterior. Isolating these cells would allow for the FindMarkers() function to discover the enriched TFs, and overall characterize them as genes that change due to ZRS mutation.

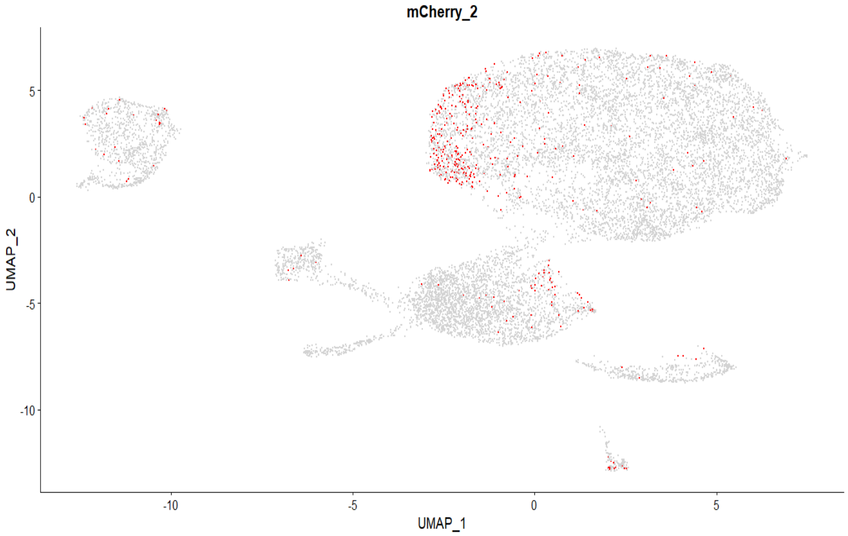

Figure 7: This plot shows the expression of mCherry at a count ¿= 2. Those displaying these numbers were classified as mCherry cells, and those having a count lower than 2, or mCherry count¡2, were non-mCherry cells. Similar to the Shh graph, these cells were used for transcription factor comparison and evaluation.

Pathogenic mCherry cell isolation was completed in a similar way to how all mCherry cells were identified. First, the ”mCherry 2” index was converted into a readable matrix using the subset() and idents() functions. These functions took the matrix including the 7 mesenchymal clusters and “mCherry count >= 2” as parameters. An intersection analysis was performed after creating the matrix of cells expressing mCherry - intersect() function.

The intersect() function required the cell indexes for three different genes: mCherry cells lacking Shh; mCherry cells lacking Hoxd13 ; and mCherry cells lacking Hand2. Similar to the steps for finding the count matrix of “mCherry 2”, performing the which() function on the count matrices isolated the cell index numbers. With the requirement fulfilled, the intersect() function resulted in a cell list containing only the similar cell index numbers of the three which functions.

The variable containing the indexes of the pathogenic mCherry cells was named FinalPathogenicCells; this marked the final step before discovering the 11 enriched TFs within them. The enriched TFs were found using FindMarkers(), with noexpression cells (those not expressing the 3 genes) as the cells to examine and the 292 TFs as the features. This function returned a list of 13 enriched TFs with an Avg log2FC threshold of .25.

3 Results

3.1 Shh and mCherry expression in 12 cell clusters

To create a more accurate analysis of the 12 cell clusters in the mouse embryo, dimensional heat maps similar to that in Figure 5 were used and allowed for the development of a UMAP with the unique cell types for all clusters (Figure 8). The 12 cell clusters were primarily composed of mesenchyme cells, including anterior/posterior mesenchyme and proximal limb mesenchyme. Analysis of the heatmap also identified Chondrocyte precursors, Erythrocytes, Vascular cells, Muscle precursors, Ectoderm, and Non-erythroid blood as cell types. Many cell clusters showed similarities between their most highly expressed genes and were identified as similar cells type, with mesenchyme as the most common. This signified that pluripotent stem cells were initially the most common, and many would eventually diversify to other cell lineages.

As seen in the UMAP in Figure 8 5 of the 12 clusters are separated from the other 7, including erythrocytes and non-erythroid blood which are large distances from other cell types. One significant correlation between the cluster location in the UMAP and their identification, was that cells separated from connected clusters were more accurately identified; this is likely because cells located further apart with little connection to other clusters are more differentiated. For example, in the case of erythrocytes, hemoglobin (Hb) was found to be the most unique gene to the cluster. Since Hb genes are most highly expressed in oxygenated red blood cells, the determination of cell type became easier. In contrast, clusters that were grouped closer or even together were of a similar and less differentiated, making them harder to identify as separate types. About 7 clusters, shown in Figure 8 display this distribution.

The 12 cluster UMAP was an important step towards creating a visual of mCherry and Shh expression the clusters (Figure 6 and Figure 7). These two figures grouped the expression of these two genes into the anterior/posterior mesoderm cells. In regards to the importance of mCherry, it functions as a luminscence gene. In data collection from live organisms, mCherry introduced with ZRS can be used as a flourescent marker of Shh activity in normal vs pathogenic regulation. Since both genes are regulated by the ZRS enhancer upon introduction, mutations in ZRS can show misregulation of Shh in the lens of a luminescence protein. In the case of this study, since the expression of mCherry and Shh is different, the ZRS enhancer is likely mutated for mCherry resulting in ectopic expression. Overall, this ectopic expression serves as a beacon of ex vitro Shh misexpression without truly harming the model organism.

The 12 cell clusters in the developing limb of a mouse embryo were the first results of this study. Using this foundation, the expression levels of mCherry and Shh could be compared to particular cell types and clusters (Figure 9). Within the 12 cell clusters, as mentioned above, Shh and mCherry were expressed in similar areas, supporting the idea of similar ZRS regulation. One surprising observation came from comparing the expression plots of mCherry and Shh: the two genes were regulated by different TFs. While they were expressed in some similar regions, like Cluster 1, mCherry was also expressed in Cluster 0 (Figure 9). Differential expression of the two genes may have something to do with the specific misregulation of mCherry due to a mutation in ZRS. However, some similarities do remain between Shh and mCherry regulation, specifically because both are most expressed in one cluster, in the mesoderm/mesenchyme.

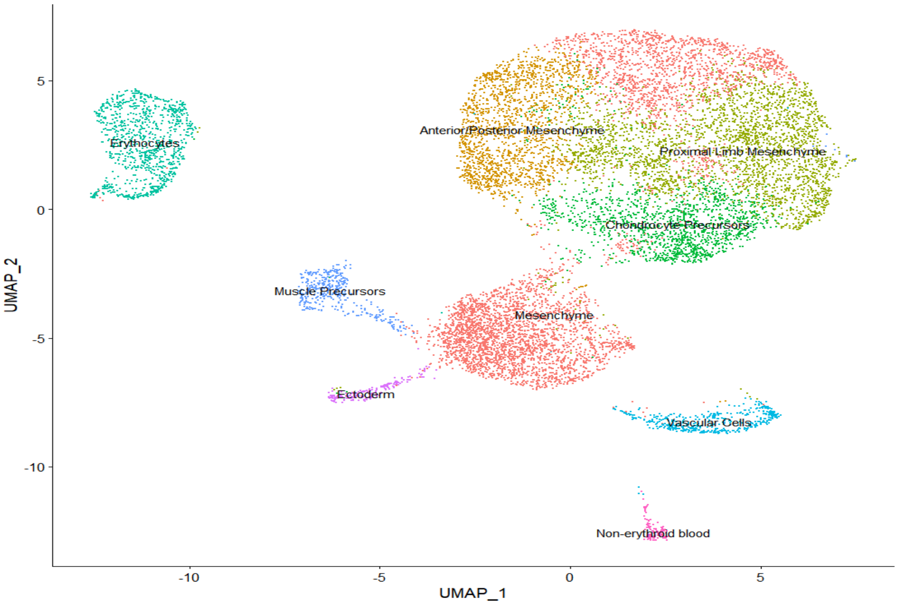

Figure 8: This image represents the fundamental map of the 12 cell clusters in the developing mouse limb. These clusters were labeled using significant genes. 7 of the clusters, represented by the big group in the top right and cluster 0 in the middle, are mesenchymal cell types. Mesenchymal cell types have yet to differentiate, making them more significant and similar for analysis [1]

Additionally, comparing the feature plots proved that Shh has more specific expression than mCherry. Even within a single cluster, the anterior/posterior mesenchyme, Shh is only expressed at a low rate. As the data reflects normal ZRS activity, Shh can be deemed to be effective in allowing normal limb development at lower concentrations. Any change to this concentration, either through upregulation or de-regulation, can result in phenotypic changes and limb disorders like PPD. Finally, after identifying the location of Shh and mCherry expression, the TFs involved in normal ZRS regulation were characterized.

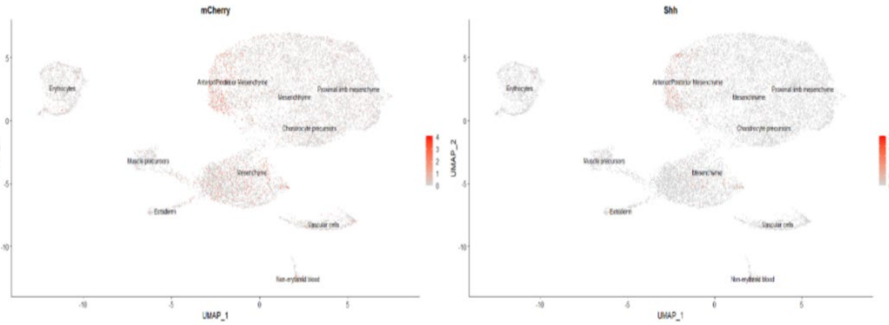

Figure 9: The left figure displays mCherry expression in the 12 cluster UMAP. The mCherry gene is expressed in Cluster 1 and Cluster 0. The image on the right displays Shh expression. Shh, in contrast to mCherry is primarily expressed in Cluster 1. This comparison concludes that Shh has a more specific expression.

3.2 Transcription factor identification, enrichment, and depletion

To determine the specific TFs and characterize them as activators or repressors, 292 transcription factors were isolated - active in the mouse limb bud. Using the UMAP in Figure 8, the enrichment of TFs in the 12 cell clusters was visualized. The focus for this step, specifically, was to determine which TFs were expressed in Shh cells and mCherry cells, or those expressed in the anterior/posterior mesenchyme. The cells used for this comparison were identified using Shh count¿=2 and mCherry count¿=2 (Figures 6 7). The TFs normally enriched in Shh cells, can potentially be activators of the ZRS enhancer. The TFs depleted in Shh cells were classified as repressors, factors that prevent Shh expression outside of a certain region; this allows the formation of an asymmetrical Shh gradient.

There were 292 unique transcription factors expressed primarily in the 12 cell clusters. However, since only TF enrichment in 7 mesenchymal clusters was observed, the list needed further cleaning. After performing the first FindMarkers analysis, 75 of the 292 TFs were found to be expressed in the 7 mesenchymal clusters. While these 75 may have been expressed in other clusters, they were also enriched in at least 1 of the 7 mesenchymal cells. After narrowing 14 down the TFs to 75, determining the enriched TFs in the cell cluster in which Shh was expressed was next. This specified the list to only include the TFs enriched in cluster 1, or the Anterior/Posterior mesenchyme. This analysis discovered that only about 25 TFs from the original 292 were expressed in the Anterior/Posterior. After visualizing some of these TFs (Figure 10, like the gene Lhx2, it was clear that most of the 25 TFs were still expressed in other clusters, at a high or low level, but were only grouped because they were also expressed in cluster 1. Figure 10 shows the expression of a TF expressed in Shh cells. The dark dots represent the TF expression in specific cells, while the coloring behind the dots represents the percentage of cells expressing a specific gene. Lhx2 is expressed in a higher percentage of Shh cells than non-Shh cells, making it a potential activator.

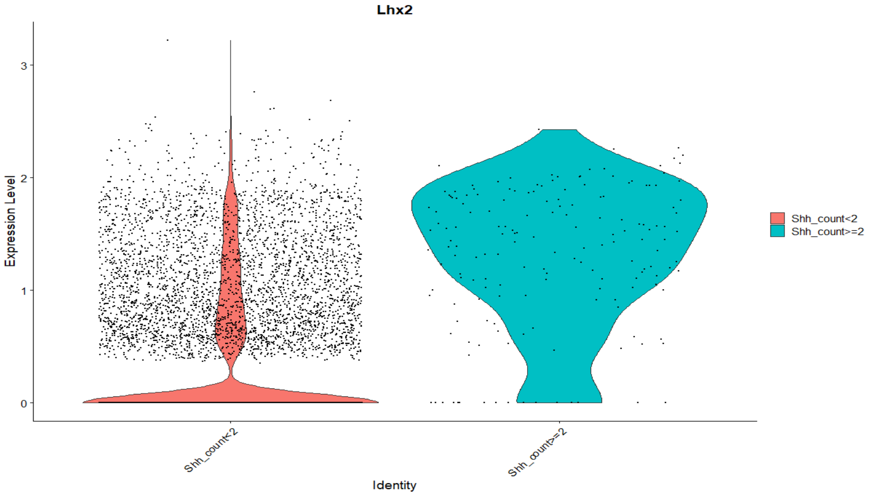

Figure 10: This violin plot displays the expression of Lhx2, one of the 25 genes expressed in the Anterior/Posterior mesenchyme. While the gene is still expressed in other clusters, it has a very high enrichment in cluster 1, making it a potential activator of ZRS.

While this method was able to determine the 25 TFs expressed in the Anterior/Posterior, a separate statistical technique was required to isolate the TFs in only Shh cells. Running the code created a list of cells expressing Shh with more than 2 counts of RNA (Figure 6). Using this list as a group, and the 292 TFs as the features, about 19/292 TFs expressed in the Shh cells were isolated. Again, a recurring observation with these TFs, was that many of them were enriched in Shh-expressing cells, but also in other clusters, or non-Shh cells. However, before using another, more specific statistical analysis, the TFs into a list (Table 1). Besides calculating the number of TFs enriched in Shh cells, the depleted TFs were identified. Of the 72 present in the 7 mesenchymal clusters, only about 26 were not expressed at all in Shh cells, classifying them as depleted.

Table 1: This table lists the 5 most enriched TFs in Shh cells. The original number of enriched TFs was 19, and their significance is determined based on each TFs avg log2FC score. This score compares the expression of specific genes inside and outside of a specific area of interest, and helps identify the most unique genes.

	myAUC	avg diff	power	avg loc2FC	pct.1	pct.2
Lhx2	0.843	0.8700344	0.686	1.2551943	0.905	0.358
Msx1	0.831	0.8510525	0.662	1.2278092	0.959	0.453
Tbx2	0.860	0.8081063	0.720	1.1658510	0.858	0.168
Hoxa13	0.780	0.5453664	0.560	0.7867975	0.763	0.205
Tfap2b	0.694	0.4660573	0.388	0.6723786	0.497	0.106

There is a large potential difference between enriched and depleted transcription factors. The latter, can be classified as potential repressors for a specific enhancer in their area of depletion. Enriched TFs in an area of interest, would classify as potential activators for a certain enhancer. Looking at the ZRS enhancer, the 19 TFs enriched in Shh cells can potentially be activators, because in their presence, Shh is transcribed and expressed. On the other hand, those depleted from these Shh cells, do not play a role in activating the ZRS enhancer. Additionally, besides the anterior/posterior mesenchyme cluster, all other clusters lack Shh expression, and therefore some of the depleted TFs must be repressing expression of the gene. Of the 72 TFs expressed in the mesenchymal cell types, 26 TFs had no expression in Shh cells, but had expression in some of the other 6 clusters (Table 2).

Table 2: This table shows the attributes and names the 5 most significant TFs found from the 26 depleted TFs from Shh cells. These 5 TFs are the most likely to act as potential repressors of the ZRS. Like the previous and future tables, the significance of these genes is ordered by their avg log2FC score.

	myAUC	avg diff	power	avg loc2FC	pct.1	pct.2
Sox9	0.669	0.7129625	0.338	1.0285875	0.574	0.325
Ebf1	0.618	0.6078095	0.236	0.8768837	0.449	0.278
Meis2	0.593	0.5838668	0.186	0.8423417	0.412	0.296
Shox2	0.596	0.5208733	0.192	0.7514614	0.713	0.757
Pbx1	0.604	0.4595937	0.208	0.6630536	0.660	0.675

The accuracy of determining potential activators and repressors was low simply by creating a tables; some of the TFs needed to be visualized. This was to determine their expression in the mesenchymal clusters, and to make the analysis more accurate. If one of these 26 were expressed primarily in one cluster or even two, they would likely play smaller role in repressing ZRS. Similarly, if a TF had high enrichment in Shh cells along with other mesenchymal clusters, it would play a smaller role as an activator. However, if some of the TFs had uniform expression levels in the 6 other mesenchymal clusters, they served as a more accurate representative of a repressor. One example of a TF exhibiting repressive properties was Sox9 which was depleted in Shh cells, but also similarly expressed in other clusters (Figure 11).

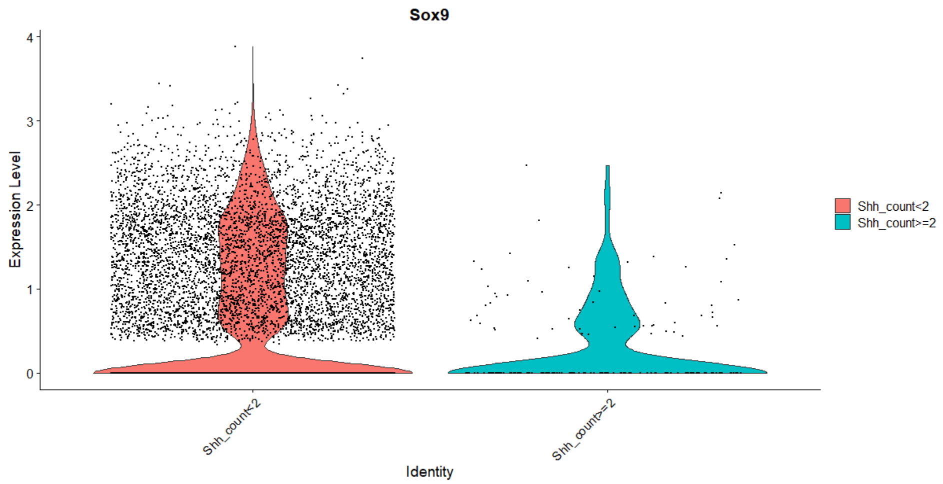

Figure 11: In contrast to the Lhx2 gene, Sox9 is mostly depleted in Shh cells. Therefore, a larger percentage of non-Shh cells express the Sox9 TF, as represented by the thick orange gradient behind the black cell dots. Moreover, while Sox9 is somewhat expressed in Shh cells, it is only at a low level.

Once the enriched and depleted TFs were fully classified and grouped, a similar procedure was performed for mCherry cells. In these cells, 12 of the 72 TFs in the 7 clusters were enriched in mCherry cells (Table 3). Additionally, there were about 13 TFs depleted specifically from the mCherry cells (Figure 4). Unsurprisingly, many of the TFs enriched and depleted were similar between mCherry and Shh cells, likely because they were expressed in similar clusters 17 and cells, and were also regulated by the same enhancer (Figures 6 and 7). There were, however, some differences in TF numbers, as Shh cells had 19 and 26 TFs enriched and depleted respectively, while mCherry had 12 and 13.

Table 3: There were a total of 12 enriched TFs present in mCherry cells, and this table displays the top 5. These 5 TFs are the most likely to play an activator role in mCherry-expressing cells. An avg log2FC value of .25 was used.

	myAUC	avg diff	power	avg loc2FC	pct.1	pct.2
Msx1	0.689	0.5872118	0.378	0.8471676	0.754	0.440
Lhx2	0.688	0.4999622	0.376	0.7212929	0.681	0.344
Hoxd13	0.680	0.4569968	0.360	0.6593070	0.593	0.249
Tfap2b	0.634	0.4045428	0.268	0.5836319	0.360	0.095
Hoxd12	0.657	0.3803993	0.314	0.5488002	0.523	0.215

Table 4: This table displays the top 5 of the 13 most depleted TFs from mCherry cells. These TFs are the most likely to play a repressive role towards mCherry expression. The avg log2FC value was set at a base of .25.

	myAUC	avg diff	power	avg loc2FC	pct.1	pct.2
Shox2	0.606	0.4123065	0.212	0.5948326	0.718	0.662
Sox9	0.585	0.3847544	0.170	0.5550833	0.576	0.494
Ebf1	0.554	0.3485868	0.108	0.5029044	0.450	0.396
Meis2	0.569	0.3259508	0.138	0.4702477	0.417	0.316
Pbx1	0.578	0.2864946	0.156	0.4133243	0.662	0.637

Despite finding the enriched and depleted TFs in mCherry and Shh cells, as Shh was not misregulated, visualizations were primarily concerned with those TFs. These visualizations helped validate the results, as a comparison of TFs expression to its luminescent expression in the developing mouse limb was performed.

The 2 most significant TFs enriched in Shh cells were visualized: Lhx2 and Msx1. One of the most depleted TFs was also visualized and compared: Ebf1. Since the misregulated gene in the data was mCherry, forming visualizations on its enriched and depleted TFs was avoided.

The Lhx2 transcription factor had was the most enriched in Shh cells compared to others cells (classified using the avg log2FC), with a value of 1.26. The first visualization, displayed in Figure 13, showed that the general expression of Lhx2 was in the same region as where Shh was expressed, in the posterior of the mouse limb bud. However, since this visualization emitted direction and orientation, already researched and recorded images of Lhx2 expression in a mouse limb bud was used. Figure 12 shows an image of Lhx2 expression in a young embryo. Lhx2 presence, indicated by a purplish luminescence at the edge of the limb verified the hypothesis that Lhx2 is expressed in Shh cells, and therefore a potential activator of ZRS.

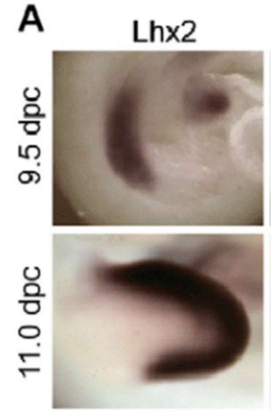

Figure 12: Expression of Lhx2 in a mouse limb bud [23]

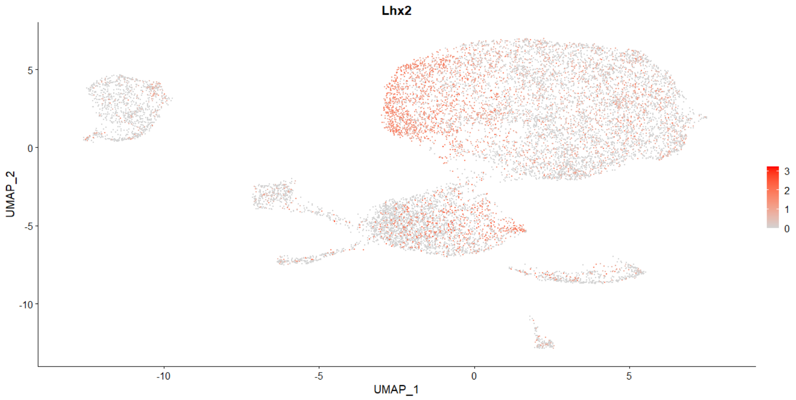

Figure 13: Expression levels of Lhx2 across cell clusters.

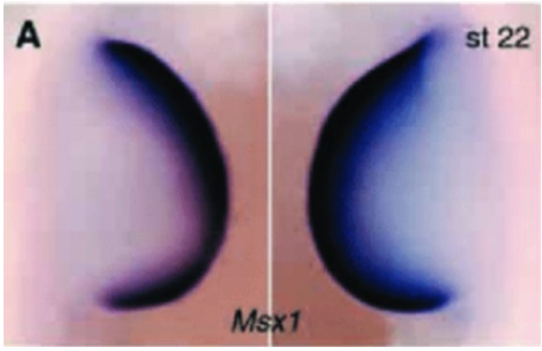

Figure 14: Expression of Msx1 in young mouse limb bud [15]

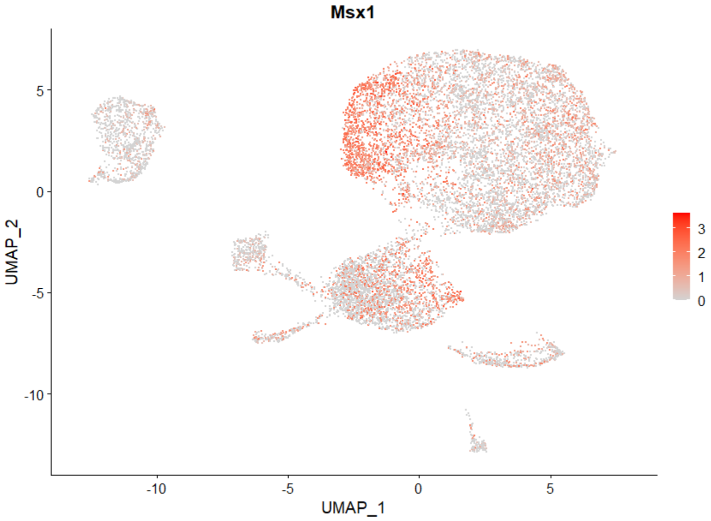

Figure 15: Msx1 expression levels.

The next TF enriched the most in Shh cells was Msx1, which had a Avg log2FC value of 1.22, not much lower than the enrichment of Lhx2. Similar to the visualizations of Lhx1, Figures 14 and 15 represents the comparison of the analysis in part B to the already completed analysis in part A. Again, since the Msx1 gene is expressed primarily in the posterior of the developing limb bud 19 and in cells also expressing Shh, these top 2 enriched TFs support the results.

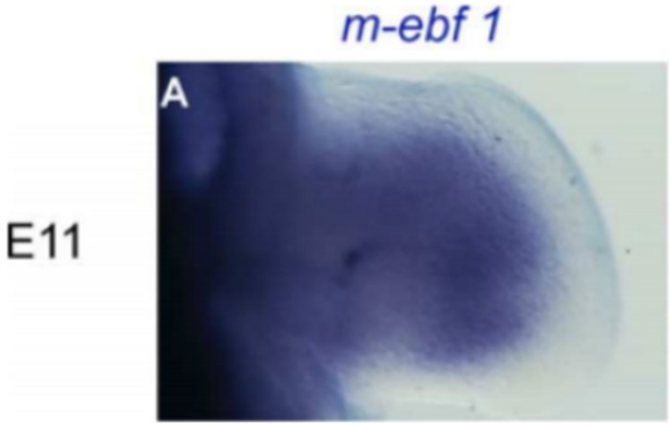

Figure 16: Expression of potential repressor Ebf1 in mouse limb bud [11]

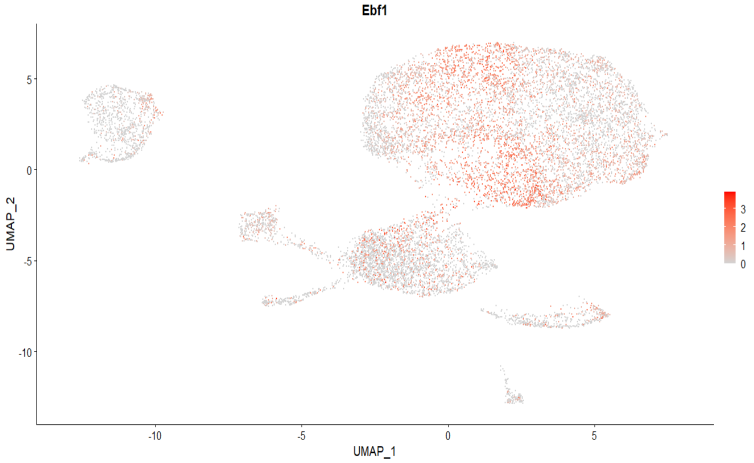

Figure 17: Ebf1 expression levels.

A similar examination of the top depleted genes of Shh was performed to further support the findings. One of these TFs, Ebf1, had an avg log2FC value of about .88 in the non-Shh cells, marking it as a depleted TF. In contrast to the two depleted genes, Ebf1 shows little expression in Shh cells, and considerable expression outside them (Figure 17). The visual in Figure 16 also shows a heavy expression of Ebf1 outside of the posterior, in the center of the limb. Both these parts A and B in this case add evidence to the classification of Ebx1 as a potential repressor.

3.3 Transcription factors in pathogenic mCherry cells

Table 5: The table displays the top five most enriched TFs out of the 13 TFs identified to be present in pathogenic mCherry cells, or those which are misregulated by a mutated ZRS.

	myAUC	avg diff	power	avg loc2FC	pct.1	pct.2
Irx3	0.566	0.2974739	0.132	0.4291642	0.236	0.118
Hoxc10	0.576	0.2753341	0.152	0.3972231	0.664	0.652
Meis2	0.530	0.2590548	0.060	0.3737371	0.336	0.312
Prrx2	0.569	0.2445884	0.138	0.3528665	0.764	0.860
Hoxd9	0.551	0.1731790	0.102	0.2498444	0.355	0.284

The final step towards determining the transcription factors behind preaxial polydactyly, was to isolate the misregulated mCherry cells and the enriched TFs within. About 110 cells in the mesenchymal clusters had pathogenic mCherry expression, making them the final focus of the study. Once the FindMarkers() was used, about 13 transcription factors were identified to be enriched particularly in the misregulated areas. Compared to the list of TF enrichment in Shh cells and normal mCherry cells, the TF list of pathogenic mCherry cells was completely different (Table 5).

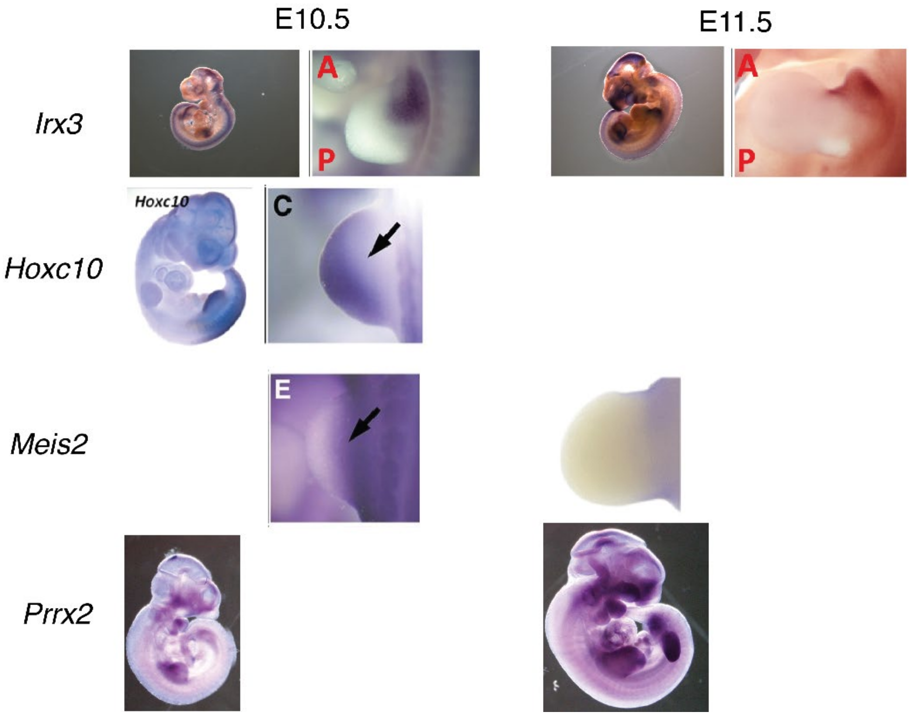

Figure 18: This image shows the expression of the top 4 genes enriched in pathogenic mCherry cells. The image comes from an analysis on a developing mouse limb of young age; any dark-colored luminescence represents gene expression in the anterior of the limb [21, 2, 24]. The two times shown in the images are day 10.5 and day 11.5.

Therefore, the 13 transcription factors which were found enriched in the pathogenic mCherry cells were classified as candidate TFs that cause pathogenic misexpression. This hypothesis was further supported by images from different data bases and publications, including the one located in Figure 18, which contains the limb expression of the top 4 TFs.

4 Discussion

Enhancers like ZRS are crucial to the proper formation of an organism's phenotype. Enhancers themselves are unable to have a large impact, and they depend 21 on other genomic elements like transcription factors. The specificity of many transcription factors allows unique gene expression patterns, making normal expression important. Most TFs have a chemical sequence that matches directly with their motif on an enhancer. In normal conditions, enhancers can regulate and bind with TFs with great efficiency. For certain genes like Shh, which plays a role in limb development, transcription factor activation and repression are important.

Activators allow expression of Shh in specific cell types and locations, like in the posterior mesenchyme of the developing limb. Repressors, play the opposite role and prevent the expression of Shh in cells outside of the posterior limb. Similarly, the enriched and depleted TFs which played a role in this study, work in different locations to create a Shh expression gradient - allows for proper limb formation.

Through the greater understanding gained from this study, it is clear that a tiny mutation in the ZRS enhancer can impact the way specific enriched and depleted transcription factors bind to ZRS. Due to modified binding, the gene expression gradient of Shh can also change, resulting in ectopic expression of the gene in the anterior of the limb. In this research, normal potential activators and repressors of ZRS can be determined by observing their expression in the 7 isolated mesenchymal clusters. Potential activators are often active only within specific areas, like Shh cells were in this study. Characterization can be supported by looking at the depletion of TFs within the 7 clusters. If a TF has little expression in the Shh cell area, but high and symmetrical expression across other clusters, its role as a repressor would be confirmed. Comparing such images was done in Figures 12b, 13b, and 14b.

The results worked similarly in this study: a list of enriched and depleted TFs was created, and each was characterized as a potential activator or repressor. The more enriched TFs, or activators, were more likely to be changed and impacted by pathogenic enhancer activity than those which were depleted; a lack of TF presence in a cell may suggest depletion and overall presence suggests a lack of TF activity and interaction in gene regulation.

This observation is important because of its relationship to phenotypic change or PPD. A mutated ZRS enhancer can change the way TFs interact and bind to the enhancer sequence. The final results, displaying the enriched TFs of the misregulated mCherry cells, supported this conclusion. The TFs enriched in the pathogenic mCherry cells were different from the enriched TFs in normal Shh and mCherry cells. Because there was a large change in enrichment, the 13 transcription factors, identified, including the four shown in Figure 18, are the most probable candidates for TFs that change due to pathogenic enhancer behavior.

The results of this study outline a clear direction towards preventing common congenital malformations like PPD. Understanding where the mutations occur and how they shift TF motif recognition, can allow for great progress. In a broader biological scope, the mutations which cause disorders like PPD lead to severe phenotypic changes, and therefore, play a large role in the evolution of more complex limb structures.

Research using single-celled resolution allowed for a closer look into how Shh is misregulated outside of the labratory, through accurate visuals brought from mCherry, a harmless luminescent protein. mCherry was introduced into the mouse embryo with a mutated ZRS enhancer, meaning to replicate a natural ZRS mutation and change in Shh expression. By using an enhancer that properly regulated Shh but misregulated mCherry, this study was able visualize, using many figures, the changes Shh expression undergoes due to natural mutations.

However, one critical aspect of this study was the pure correlative approach taken. The transcription factors identified in both normal Shh and mCherry cells as well as pathogenic mCherry cells, was purely based on RNA transcript presence. Since enhancers have specific motifs which match specific transcription factors, the list of transcription factors isolated in results are only potential activators and repressors; a more specific analysis is required to determine and characterize the specific attributes of each TF.

Therefore, as a way to make the TF list more accurate, several approaches can be taken. One such approach is determining the exact motif sequence of the mutated and normal ZRS enhancer. In this data, only one mutation has been introduced into the ZRS enhancer, meaning that only one of the 13 enriched TFs in mCherry cells are changed, and therefore cause mCherry misexpression. The key to discovering what TF is involved and whether it is a activator or repressor, relies on its ability to bind to the ZRS motif. For an activator, a motif is created by the single mutation, allowing for an activator to be misexpressed and allow for ectopic expression. On the other hand, if a mutation causes a TF to lose the ability to bind to a motif, then that TF would be a repressor because its absence causes increased gene expression.

This study can likely be taken further by identifying normal ZRS motifs and TFs which act as activators/repressors, and also identifying which motif and TF changes due to pathogenic ZRS activity; only 1 of the 13 cause PPD.

The approach used in this research also holds some importance. As this analysis allowed for the identification of TFs which can potentially lead to congenital malformations, a similar approach can be used. For more lethal disorders and malformations, such as those affecting the brain, studies of transcription factor activity and change due to pathogenic enhancer activity can have a large influence.

5 Acknowledgements

We would like to thank professors Evgeny Kvon, Ph.D., and Qing Nie from the University of California Irvine (UCI) for their continuous support and mentorship throughout our research. Their guidance provided a clear path to our research and was fundamental to the success of this research project. Additionally, we would like to acknowledge our teaching assistant Yutong Sha, Ph.D. candidate at UCI, for her help and constant advice. Finally, we would like to thank the directors of the UCI Math MathExpLR and MathBioU program for providing the wonderful opportunity to conduct meaningful research with PIs from the University of California Irvine.

References

[1] Arnold I. Caplan. “Mesenchymal stem cells”. In: Journal of Orthopaedic Research 9.5 (Sept. 1991), pp. 641–650. doi: 10.1002/jor.1100090504.

[2] Katherine Q Chen et al. “Development of the Proximal-Anterior Skeletal Elements in the Mouse Hindlimb Is Regulated by a Transcriptional and Signaling Network Controlled by Sall4”. In: Genetics 215.1 (May 1, 2020), pp. 129–141. doi: 10.1534/genetics.120.303069.

[3] Eileen E. M. Furlong and Michael Levine. “Developmental enhancers and chromosome topology”. In: Science 361.6409 (Sept. 28, 2018), pp. 1341– 1345. doi: 10.1126/science.aau0320. (Visited on 06/30/2021).

[4] Linh Huynh and Fereydoun Hormozdiari. “TAD fusion score: discovery and ranking the contribution of deletions to genome structure”. In: Genome Biology 20.1 (Mar. 21, 2019), p. 60. doi: 10.1186/s13059-019-1666-7.

[5] Evgeny Z. Kvon et al. “Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants”. In: Cell 180.6 (Mar. 2020), 1262–1271.e15. doi: 10.1016/j.cell.2020.02.031.

[6] Evgeny Z. Kvon et al. “Enhancer redundancy in development and disease”. In: Nature Reviews Genetics 22.5 (May 2021), pp. 324–336. doi: 10.1038/ s41576-020-00311-x.

[7] L. A. Lettice. “A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly”. In: Human Molecular Genetics 12.14 (July 15, 2003), pp. 1725–1735. doi: 10.1093/hmg/ddg180. (Visited on 06/30/2021).

[8] Hannah K. Long, Sara L. Prescott, and Joanna Wysocka. “Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution”. In: Cell 167.5 (Nov. 2016), pp. 1170–1187. doi: 10.1016/j.cell.2016. 09.018. (Visited on 06/30/2021).

[9] Dar´ıo G. Lupi´a˜nez et al. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions”. In: Cell 161.5 (May 2015), pp. 1012–1025. doi: 10.1016/j.cell.2015.04.004.

[10] S. Malik. “Polydactyly: phenotypes, genetics and classification”. In: Clinical Genetics 85.3 (Mar. 2014), pp. 203–212. doi: 10.1111/cge.12276.

[11] S´ebastien Mella et al. “Expression patterns of the coe/ebf transcription factor genes during chicken and mouse limb development”. In: Gene Expression Patterns 4.5 (Sept. 2004), pp. 537–542. doi: 10.1016/j.modgep.2004. 02.005. 24

[12] Yanling Peng and Yubo Zhang. “Enhancer and super-enhancer: Positive regulators in gene transcription”. In: Animal Models and Experimental Medicine 1.3 (Sept. 2018), pp. 169–179. doi: 10 . 1002 / ame2 . 12032. (Visited on 06/30/2021).

[13] Len A. Pennacchio et al. “Enhancers: five essential questions”. In: Nature Reviews Genetics 14.4 (Apr. 2013), pp. 288–295. doi: 10.1038/nrg3458.

[14] Florence Petit, Karen E. Sears, and Nadav Ahituv. “Limb development: a paradigm of gene regulation”. In: Nature Reviews Genetics 18.4 (Apr. 2017), pp. 245–258. doi: 10.1038/nrg.2016.167. (Visited on 06/30/2021).

[15] Sandrine Pizette and L Niswander. “BMPs negatively regulate structure and function of limb apical ectodermal ridge”. In: Development (Cambridge, England) 126 (Mar. 1999), pp. 883–94.

[16] Valerie Reinke. “Transcriptional regulation of gene expression in C. elegans”. In: WormBook (June 4, 2013), pp. 1–31. doi: 10.1895/wormbook.1.45.2.

[17] Michael I. Robson, Alessa R. Ringel, and Stefan Mundlos. “Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D”. In: Molecular Cell 74.6 (June 2019), pp. 1110–1122. doi: 10 . 1016 / j . molcel.2019.05.032. (Visited on 06/30/2021).

[18] Tomoko Sagai et al. “Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb”. In: Development 132.4 (Feb. 15, 2005), pp. 797–803. doi: 10.1242/dev.01613.

[19] Stefan Schoenfelder and Peter Fraser. “Long-range enhancer–promoter contacts in gene expression control”. In: Nature Reviews Genetics 20.8 (Aug. 2019), pp. 437–455. doi: 10.1038/s41576-019-0128-0. (Visited on 06/30/2021).

[20] Francois Spitz and Eileen E. M. Furlong. “Transcription factors: from enhancer binding to developmental control”. In: Nature Reviews Genetics 13.9 (Sept. 2012), pp. 613–626. doi: 10 . 1038 / nrg3207. (Visited on 06/30/2021).

[21] Leila Taher et al. “Global Gene Expression Analysis of Murine Limb Development”. In: PLoS ONE 6.12 (Dec. 9, 2011). Ed. by Costanza Emanueli, e28358. doi: 10.1371/journal.pone.0028358.

[22] Cheryll Tickle and Matthew Towers. “Sonic Hedgehog Signaling in Limb Development”. In: Frontiers in Cell and Developmental Biology 5 (Feb. 28, 2017). doi: 10.3389/fcell.2017.00014.

[23] Itai Tzchori et al. “LIM homeobox transcription factors integrate signaling events that control three-dimensional limb patterning and growth”. In: Development 136.8 (Apr. 15, 2009), pp. 1375–1385. doi: 10.1242/dev. 026476.

[24] Shigetoshi Yokoyama et al. “Analysis of transcription factors expressed at the anterior mouse limb bud”. In: PLOS ONE 12.5 (May 3, 2017). Ed. by Michael Schubert, e0175673. doi: 10.1371/journal.pone.0175673.