STUDYING PRIMING AND POISING OF CELLS

A Project Report

submitted by

ARIBA ANSARI

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

19 December 2022

THESIS CERTIFICATE

This is to certify that the thesis titled **STUDYING PRIMING AND POISING OF CELLS**, submitted by **Ansari Ariba Abdul Majeed**, to the INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI, for the award of the degree of **Masters of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr Vibhor kumar
Thesis Supervisor
Associate Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: December 2022

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Vibhor Kumar, for his constant support and mentoring. It would not have been possible without his efforts; he continually encouraged me to reach greater life goals and achieve better at work. Whenever I had stuck on a problem or needed assistance, he always made time for me, even on his busiest days. Because of his efforts and constant support, I could complete my work on time. I want to thank my Ph.D. mentors, Madhu Sharma, Shreya Mishra, and Neetesh Pandey, who has helped me through my thesis and were available to help at any time of the day whenever I was stuck with any problem. They supported me and guided me throughout my thesis. I would also like to thank my family for their care and constant encouragement.I would like to thank my younger and beloved sister Sameeha Ansari for always cheering up for me.

ABSTRACT

An increase in research and development of single-cell genomics led to various insights into cells and their function, it has become a primary focus of research leading to many great discoveries. RNA velocity is one such method obtained through single-cell genomics data that gives information on newly transcribed pre-mRNA and mature mRNA and distinguishes among them. It gives the ratio of spliced and unspliced mRNA, reveals the lineage relationships of a single cell, and predicts its future state on a time scale in a high-dimensional vector. One another great application of single-cell data is pathway enrichment analysis which gives the enriched biological pathways in a gene list. Here we aim to trace the lineage of single cells through RNA velocity and to find the pathways affecting the directionality of priming and poising of cells to get insights into the pathways activities. That is, which pathways are enriched during which lineage of cells will reveal which pathways are helping the cells to differentiate towards a particular lineage. Getting the relationship between the lineage of cells and the enriched pathways will be of great help in fields such as 3D bioprinting of organs and tissues, formation of organoids, and regenerative medicines.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A cell is a fundamental unit of life, and as we all know, there is increasing research on single-cell genomics and single-cell data, leading scientific research to great heights. As single-cell data gives us information about each cell and genes, it has become convenient to get insights into cells and their genetic information through next-generation sequencing. Sequencing of single cell data also provides us information about RNA seq, which can lead to gaining knowledge about RNA and its stage of transcription, One such method to find out the stage of mRNA is finding RNA velocity, which gives us information about spliced and unspliced RNA and ratio among them which has a great application such as when we extract cells from the body the cell dies. We only get to know the current stage of the cells. Still, what if we have to find out what was the stage of the cell in the past or what the cells would be in the future? Here, RNA velocity comes into the picture. It differentiates the newly transcribed mRNAs from mature RNA and gives us the stages of the cells from the past and what a cell will be in the future, which has led to some excellent research. If we apply this concept to human embryonic stem cells, we will know the early differentiation of cells and which cells and organs they will form in the future.

There is an increasing demand for organ transplantation, and at times it becomes difficult to get a donor, and many lives have been lost due to the unavailability of donors. In this study, we have tried to find a solution in the initial stage. By using the concept of RNA velocity and mapping the pathway activities which has been used in the formation of cells, we tried to find the correlation between the RNA velocity and Pathway activities. So that at each stage of cells, if we get to know its pathway activities, we can control the directionality and poising of cells, which will manipulate the formation of cells in the desired stage, which can be helpful in the formation of Organoids and 3D bioprinting of tissues and organs.

1.1 Single cell Genomics

The study of cell individuality using omics techniques is known as single-cell genomics. It is carried out to increase our fundamental understanding of illnesses and diagnostic methods. It helps to understand the diversity that lies within our cells. It uses next-generation sequencing techniques to obtain information from individual cells, which leads to a better understanding of the functions of individual cells. There are many sequencing methods for sequencing single-cell data and to retrive the information. RNA-seq is obtained by performing next-generation sequencing on single-cell data, which gives us information about the RNA in that single-cell data.

1.2 RNA Velocity

RNA velocity is a technique of forecasting future state of cells in a high dimensional vector high-dimensional on an hourly timeline.RNA velocity is calculated by using the RNA seq data. RNA velocity gives us a directionality of priming and poising of cells in pseudo time. It distinguishes the newly transcribed RNA from mature RNA, so that the information of the cell's state is recovered. That is, the cells were in which state and will be in which state in the furture. That helps us to understand the early development and differentiation of cells i.e., the time dependent derivative of mature mRNAs, which is useful in identifying the states of genes and the directionality of cell transitions.

1.2.1 Limitations of RNA Velocity

RNA velocity has some great applications, but on the other hand, it has some limitations, too. One of the major limitation is noise which should be eliminated to get smooth results. As of now, two models exist for estimating RNA velocity one is the steady state model, while the other is the dynamical model. The major limitation of the steady-state model is that it can lead to inaccuracies in velocity estimation due to its two major presumptions that the splicing rate is common across genes and that the sampled data has the least evidence of steady-state expression levels. Some more limitations of RNA velocity which are addressed are as follows[2]

- Mature cells simulation may lead to false projections due to the arrows provided

that cannot be seen in the ground truth.

- Some complex characteristics have been observed, such as during the transition towards different fates, a simulation up regulation and down regulation may lead to misleading patterns.
- Deflated curvature can be observed due to variable synthesis rates, such as curvature being inflated by slowly increasing splicing rates and decreasing degradation. In contrast, the curvature is flipped by slowly decreasing splicing rates and increasing degradation.
- In the current model of RNA velocity, the stochastics are ignored, which causes the heterogeneity in kinetic rates to remain unidentified.[2]

1.3 Pathways Enrichment Analysis

Enrichment analysis of pathways gives us insights into the biological pathways that are enriched in a list of genes. Comprehensive DNA, RNA, and protein quantification in biological materials is now standard practice. The generated data is accumulating at an exponential rate, and its analysis assists researchers in discovering novel biological functions, genotype-phenotype correlations, and disease mechanisms. However, many researchers face difficulties in analyzing and interpreting this data as the gene sets are quite massive. Analyses frequently yield huge lists of genes that require impractically extensive manual literature research to comprehend. Pathway enrichment analysis summarises the enormous gene list as a smaller list of more interpretable pathways.These pathway enrichment analyses of single-cell data give scores of how highly active the particular pathway is in each cell.

1.4 Problem Statement

Navigating developmental pathways that result in the differentiation of stem cells to desired lineages is necessary as at times, the differentiation of alternate pathways can lead to unwanted cells. Therefore, developmental roadmaps are essential for stem cell differentiation that will lead to the desired outcomes. Directing stem cells toward pure lineages at most branch points is necessary to prevent differentiation towards undesirable fates. A previously unknown process defined by HOPX expression in the formation

of the mesoderm, somite segmentation, was discovered by mapping changes in single-cell gene expression and sequential chromatin modifications.This roadmap is necessary as it can lead to the formation of desired organs and tissues, as the significant challenge during transplantation is the unavailability of donors and graft rejections.

1.5 Aim of our Thesis

Here we propose a method to map the lineages and pathways in embryonic stem cells, which will lead to the formation of Endoderm, Ectoderm, and Mesoderm. Further, the Mesoderm will lead to the formation of Human bone, Carlitage, Fibroblast, and Cardiomyocytes. We combined the RNA velocity with pathway scores to find the correlation between them, showing which pathways are highly correlated with the formation of respective cell lineages and their bifurcation. We aim to build a roadmap showing pathways used in the differentiation of cells from day 0 to day 1 and further on, this knowledge will lead to mapping the stem cell differentiation in the desired lineage and the formation of organoids or transplantable organs.

This roadmap formation allows the mapping of stem cell development to create transplantable human tissues, progenitors, and organs, which can further help in the 3D Bioprinting of transplantable Organs and Tissues.

1.6 Related Work

The differentiation of human embryonic stem cells can result in a diverse group of cells, by understanding the factors that regulate the bifurcation of embryonic lineages might effectively produce target cell types by blocking alternative fates[3]. At the time, several development pathways led to unwanted cell types; therefore, navigation of stem cell differentiation through roadmaps was required [1].

Kyle M.loh et.al developed a roadmap for endoderm differentiation, and the anterior-posterior patterning signals were examined. They initially found two prominent signaling pathways, BMP and WNT, for the anterior primitive streak, but after 24 hours, these two molecules inhibited endoderm and stimulated mesoderm. They found that primi-

tive streaks were initially formed with the help of FGF, Wnt, BMP, and TGF$/beta$. There were discoveries showing activin molecules, combined with BMP, PI3K, or FGF inhibitors, promote human embryonic stem cells toward definitive endoderm differentiation. However, some of these approaches continued to produce mixed lineage. The Definitive Endoderm differentiated from the primitive streak will form definitive endoderm through the anterior primitive streak, while the posterior primitive streak will lead to the formation of mesoderms.The Anterior primitive streak and the posterior primitive streak were induced on day 1 of human embryonic stem cells by the pathways such as BMP, Wnt, and FGF[3].

Alteration of the BMP pathway can lead to the formation of only the anterior primitive streak or posterior primitive streak. For example, the anterior primitive steak will be more prominent with the low endogenous BMP. In contrast, the posterior primitive streak becomes prominent with a higher level of BMP. However, Wnt and FGF were both equally required to form primitive streaks. It has been found that the BMP and Wnt have later resulted in the repression of definitive endoderm and induction of mesoderm. To repress mesoderm and to continue the formation of definitive endoderm, BMP signals were neutralized while the exogenous BMP was removed. There was some gene upregulation and downregulation were also involved. MESP1 is the gene that was downregulated, while FOX, SOX17, and HHEX were upregulated genes. Mesoderm formation was also blocked by the elimination of endogenous Wntβ signals. It has been concluded that the BMP and Wnt signal helps in the formation of mesoderm while the TGFβ and FGF help in the formation of endoderm[3].

A landscape of human mesoderm was mapped by Kyle M.loh et al., where the mesoderm development has been shown from pluripotent epiblast to anterior and mid primitive streak, which will further separate into paraxial and lateral mesoderm and other lineages. This will further form somites, and lateral mesoderm and somites will segregate into ventral somites forming bone and cartilage, dorsal somites forming skeletal muscles, and brown fats. In contrast, the lateral mesoderm will result in cardiomyocytes through cardiac mesoderm formation. For the mesoderm formation, the PI3K signals were blocked; on the other hand, the Wnt, FGF, and TGFβ signals were activated. For the formation of paraxial and lateral mesoderm, FGF and ERK signals were activated while the TGFβ signals were inhibited. While Wnt and BMP also play an essential role, the presence of exogenous BMP results in the formation of lateral mesoderm and

the suppression of paraxial mesoderm. At the same time, inhibition of BMP induces paraxial mesoderm and suppresses lateral mesoderm. On the other hand, activation of Wnt signals helps the formation of paraxial mesoderm and blocks lateral and cardiac mesoderm.[1]

On days 2-3 of human embryonic stem cell differentiation, WNT, and FGF inhibition downregulate the paraxial mesoderm and upregulate early somites' formation. For further development of early somites, signals such as BMP and TGFβ were also blocked. HH and WNT signals also play an essential role. If WNT were activated and HH was blocked, it inhibits the formation of dorsal somites and induces the formation of ventral somites. In contrast, the activation of WNT and inhibition of HH block the formation of ventral somites and specify dorsal somites. While the activation of WNT and HH together will lead to the formation of both ventral and dorsal somites. The WNT is also an important signal in the lateral mesoderm. The activation of WNT will form lateral mesoderm, which will further lead to the induction of limb markers such as PRRXI and HOXB5. On the other hand, the inhibition of WNT will suppress the posterior lateral mesoderm and the formation of cardiac mesoderm [1].

CHAPTER 2

DATASET

The data used for the experiment was single-cell RNA-seq data extracted from human embryogenesis stem cells. Human embryonic stem cells provide a cellular model to study the lineage and bifurcation of cells at the primary stage and help map cell differentiation. We have also used human pluripotent stem cells data to map endoderm progenitors' lineage. While the human embryonic stem cells data is for mesoderm progenitors.

2.1 Data Collection

We have validated our results using the two following datasets, which are collected from the GEO accession number in NCBI. GEO is a large genomic data repository.

1. **Human Embryonic Stem Cells (hESCs) for Endoderm Progenitors** - In this data Human embryonic stem cell entrance into endoderm progenitors was studied using a snapshot and temporal scRNA-seq of progenitor cells. This dataset is registered with the organization Morgridge Institute for Research. In this dataset, there are 1018 single cells from snapshot progenitors and 758 from time course profiling. The dataset is available with GEO Accession - GSE75748

   The dataset is available on the following link:

   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgiacc=GSE75748

2. **Human Embryonic Stem Cells (hESCs) for Mesoderm Progenitors** - H7 human embryonic stem cells (WiCell) and H7-derived downstream early mesoderm progenitors were grown in vitro. This dataset is registered with the organization Stanford University School of Medicine. It has Single cell Bulk cell RNA-seq data derived from in vitro-grown cells. The data is available with GEO Accession - GSE85066

   The dataset is available on the following link:

   https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE85066

2.2 Data Pre-Processing

2.2.1 Sequence Read Archive (SRA) Files

Sequence Read Archive (SRA) is a repository for DNA sequencing data. It mostly consists of short reads of length less than 1000 base pairs and was generated using High Throughout Sequencing.We have collected SRA files from both the dataset i.e., hESC for Endoderm progenitors (dataset 1) and hESC for mesoderm progenitors (dataset 2).

SRA files in Dataset 1

In dataset 1, we got 1432 SRA Files which was in all time point and consisted of a single human embryonic cell. The SRA Files were annotated with batch numbers. These batch numbers were distributed according to the time course of cells. The dataset was divided into six batches with a time course of 0hr, 12hr, 24hr, 36hr, 72hr, and 96hr.

SRA files in Dataset 2

In dataset 2, we have predefined batches, each representing cells stage and its time point. There were ten such batches divided as follows,

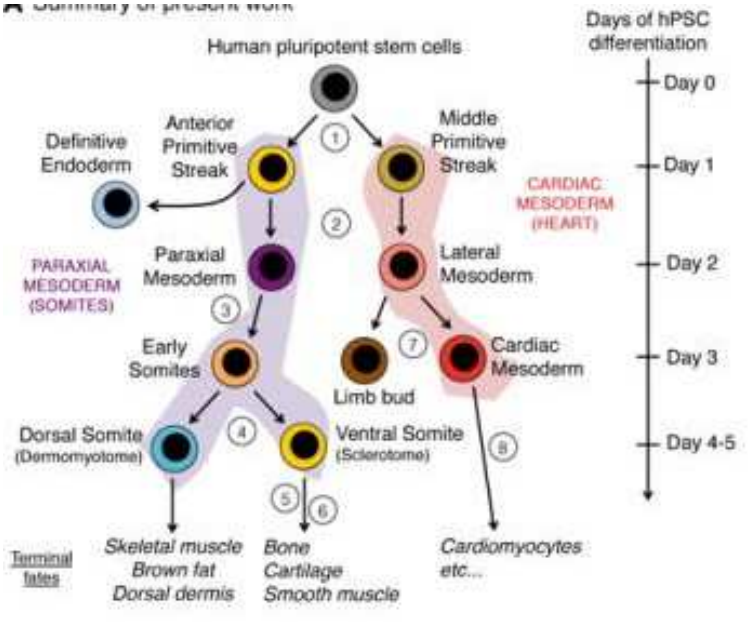

Figure 2.1: Human Pulripotent Stem Cells Differentiation[1]

Here in this dataset, day 0 is Human Pluripotent Stem Cells; further, on day 1 it is divided into groups of Anterior Primitive Streak and Middle Primitive Streak further, on day 2 the Anterior primitive streak is divided into definitive mesoderm and paraxial mesoderm while the mid primitive streak is differentiated into lateral mesoderm. On day 3 the paraxial mesoderm is differentiated into early somites, and the lateral mesoderm is differentiated into limb bud and cardiac mesoderm, which will further give rise to cardiomyocytes. On days 4-5, the early somites will differentiate into dorsal somites, which will further give rise to skeletal muscle, brown fat, and dorsal dermis, these early somites are also differentiated into ventral somites, which will further give rise to bone, cartilage, and smooth muscle cells.

SRA Toolkit

The SRA toolkit includes tools for converting data in SRA format, downloading data and retrieving SRA data in other formats. It also helps to download the SRA files in bulk. We downloaded the bulk SRA files using sratoolkit by providing a list of SRA accession numbers in text format.

2.2.2 FASTQ Files

FASTQ files store sequencing information in text format in a single letter code, which can be further helpful for Next Generation Sequencing (NGS). It also consists of quality scores based on the nucleotide sequence. The quality scores and the single letter code are encoded with ASCII characters. ASCII is an American Standard Code that is used for information exchange. The FASTQ files contain sequence information as follows.

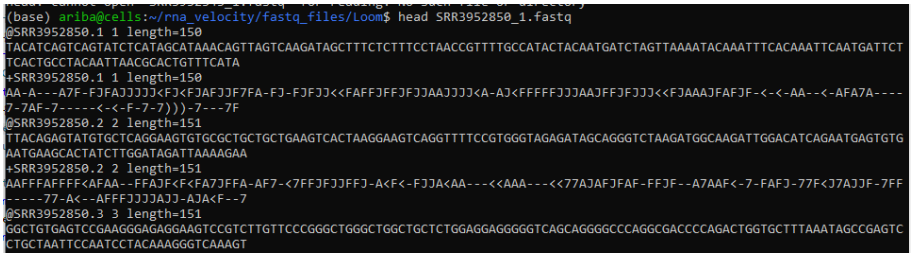

Figure 2.2: Snapshot of a FASTQ File

- It always starts with a @ symbol with a sequence identifier.

- The following line consists of sequences like A, T, G, C, and N.
- Then in the next line, there is a + sign with a separator that shows that the DNA sequence has ended.
- Then the last line consists of ASCII-encoded characters that indicate the quality score of the sequence.

SRA Toolkit

Here SRA toolkit is used to retrieve the sequence as FASTQ. Fastq-dump command in the SRA toolkit is used to convert the SRA files in both datasets into FASTQ files. The FASTQ file obtained here consists of paired-end reads. There's an option to retrieve a single FASTQ file or two FASTQ files, which will be annotated as 1 and 2. In the paired-end reads, one file contains information from the start of the sequence, while the second file consists of information from the end of the sequence, which is helpful in mapping or sequencing the reads through Next Generation Sequencing. As some tools require FASTQ paired-end reads for the sequencing.

2.2.3 Binary Alignment and Map (BAM) Files

A binary Alignment and Map (BAM) file is a binary version of the Sequence Alignment and Map (SAM) file that stores data about the sequence read alignments that are mapped against reference genomes, which is obtained through Next generation sequencing. BAM files are the same as SAM files. The only difference is the SAM files are in text format. Therefore, it is human readable. However, it is not readable by computers, so a BAM file is needed. BAM files store alignment information in binary format, which also takes less storage and is suitable for most analyzing software tools. Generation of BAM files is done in two steps: first, creating a reference index and second, mapping the reads against the reference index. After BAM files are generated, sorting bam files is also required as reads are random according to their position on the genome.

BAM files consist of two sections header and alignment.

- The header section contains some specific information, such as the length of the sequence, the name of the sample, and the alignment method used for the sequencing.

- The Alignment section consists of information on read sequence, chromosome number, start coordinate, quality of the reads, information on the alignment, and some custom tags such as Barcode tag (BC),Read Group (RG), single or paired read alignment quality tags (SM/AS), edit distance tag (NM), and Amplicon name tag (XN).

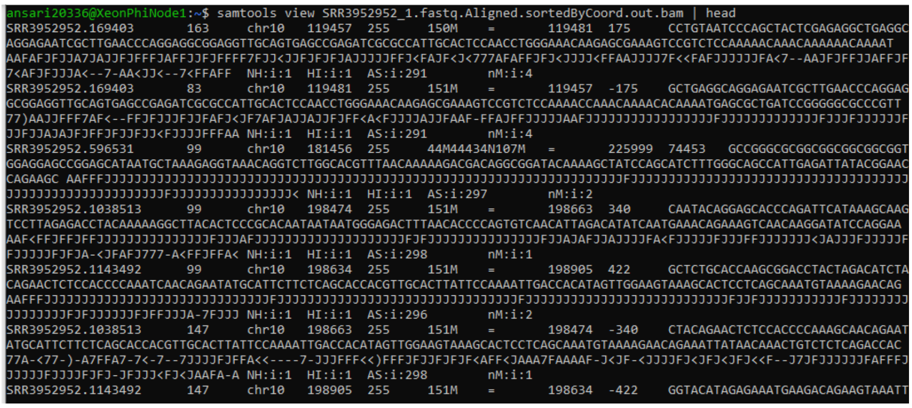

Figure 2.3: Snapshot of a BAM File which is sorted by coordinate

STAR Aligner

STAR (Spliced Transcripts Alignment to a Reference) Aligner is used to map the reads. STAR is a Next Generation Sequencing tool. Next-Generation sequencing(NGS) is a technique of parallel sequencing it can be used to find the nucleotide order in the given genome or the target position of DNA or RNA. NGS provides a rapid sequence of DNA that saves time and provides reliable results. STAR helps to map the reads with the help of an index file and gives efficient results. STAR works on two concepts searching the seeds and clustering, stitching, and scoring. Mapping with STAR consists of two steps Indexing and Mapping. Raw FASTQ reads are mapped using STAR Aligner with the help of reference genome index and give output as BAM files. These BAM files contain all the information on sequence alignment, chromosome number, and quality score of the sequence.

- **Indexing**
  Indexing provides a reference genome against which our raw FASTQ reads are sequenced. To make this index, we have used the tool STAR aligner in which we have provided a reference GRCh37(hg19) genome, which was in FASTA format. This FASTA file contains standard nucleotide sequence and chromosome information and is compatible with various bioinformatics tools. We have also provided an annotation file of GRCh37 (hg19) in GTF format, and this annotation file consists of functions, products, and locations of the genes. Here we

have created an index for the hg19 genome of homosapiens which has 7 regions with alternate loci.The chomosome names in both the files, genome sequence file and annotation file should match with each other. We have collected the Human genome reference FASTA file and annotation file from NCBI (National Centre for Biotechnology Information).

- Source for Human Reference Genome hg19 - https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
- Source for Genome Annotation file hg19 - https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

- **Mapping**
  After creating indices, mapping of the sequence is done using the tool STAR aligner. For mapping, we have provided the paired end FASTQ files, along with the reference index created, and we gave a command to sort the output BAM files, to sort it by coordinates. As we have 1432 files for hESC for Endoderm progenitors and 650 files for hESC for Mesoderm progenitors , we have applied a loop through all the FASTQ files, mapped it with STAR aligner, and got the output as sorted BAM files.

2.2.4 Loom Files

The loom file is a file that can be created and read in any language and is able to store large datasets with metadata for rows and columns. A Loom file is a very convenient file for single cell data consisting of large matrices. Loom files give efficient access to arbitrary rows and columns. It is based on an HDF5 concept. HGF5 files consist of groups and datasets; datasets are like matrices consisting of an array, while the groups are like folders that contain datasets.

2.2.5 Velocyto

Velocyto is a tool that is used to find and analyze RNA velocity. It Distinguishes spliced and unspliced RNA in single cell RNA data. Velocyto is a package available in both R and Python languages. However, only python codes were available to preprocess the data, such as converting BAM files into Loom files. We have used that to convert BAM files into Loom files. Each BAM file is converted into 1 Loom file, and then we merge all Loom files ofhESC for Endoderm progenitors into one loom file and hESC for Mesoderm progenitors into another Loom file. The loom files are combined through

the loompy.combine command in velocyto, after which we can find the RNA velocity and analyze it. To convert BAM into Loom, we also required a Genome annotation file and repeat masking genome file, which has locations of the genome sequence that are repeated. Repeat masking genome file helps find the repeated sequence's location. We have downloaded the repeat mask file for hg19 from the repeat masker website. While the genome annotation file consists of functions, products, and locations of the genes, We have downloaded the human genome annotation file for hg19 from NCBI.

- Source for repeat mask file of hg19 - https://www.repeatmasker.org/
- Source for Genome Annotation file hg19 - https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/

CHAPTER 3

METHODOLOGY

3.1 Finding RNA Velocity

There are two major modeling techniques available to calculate RNA velocity, and these techniques use kinetic expressions to find RNA velocity. The first proposed model was Velocyto based on steady-state approach, and later an extended version scVelo was proposed based on dynamical approaches.

3.1.1 Steady-State Approach(Velocyto)

The half-life of an mRNA is usually equivalent to the differentiation timeline of cells which takes place in a timeline of hours and days. The presence of newly transcribed and mature mRNA in abundance can be used to determine the rate of splicing and degradation. The steady state approach proposes a model that is dependent on transcriptional dynamics to compute the relationship between spliced and unspliced mRNA, which is time-dependent. Knowing the balance between the production of spliced and unspliced mRNA and the degradation of mRNA, the first time derivative is derived. This model is used when the α is kept as the constant transcription rate. When the rate of transcription is constant, stable states are reached asymptotically, and the spliced (s) and unspliced (u) molecules are kept in a steady state, which has a fixed-slope relationship [4].

$$u = s\gamma[4] \tag{3.1}$$

The regulatory features of gene,internal priming sites and it's number and the exonic intronic length ratio is taken into account while the formation of the equilibrium slope. It was widely shown that in most of the gene the steady-state behaviour was consistent with a single constant slope using a newly released compendium of mouse tissues[4].

The transcription rate increases during a dynamic process which leads to increasing of unspliced mRNA, which is followed by an increase in spliced mRNA until a new steady state is established. A decrease in transcription rate, on the other hand, causes a quick decrease in unspliced mRNA, which is followed by a decrease in spliced mRNA. During induction period of gene expression ,the unspliced mRNAs are present in excess at equilibrium rate, but during repression the converse is true . As a result, the balance of spliced and unspliced mRNA abundance is an indication of mature mRNA abundance in the future state, and consequently future state of the cell[4].

3.1.2 Dynamic Approach (scVelo)

This model based on a dynamical approach overcomes the shortcoming of the previous model as there was an error with velocity estimation as there can be a violation of mRNA levels with the steady state due to the assumption of shared rate of splicing and whole splicing rate observation. scVelo is introduced to overcome these restrictions by addressing the complete dynamics of transcription and of splicing kinetics by using a model based on dynamical likelihood [5].

Positive RNA velocity shows that a gene is upregulated, which occurs when cells have more unspliced mRNA for that gene. Negative velocity, however, shows down-regulation of gene. Forecasting an individual cell's future state can be possible by the combination of gene velocities [5].

scVelo is a dynamical model that addresses the whole gene-wise transcriptional dynamics. As a result, RNA velocity estimate is applied to transient systems and systems with heterogeneous subpopulation dynamics. In an efficient expectation-maximization (EM) framework,the transcription rate for gene specific reaction, splicing, and degradation are inferred. Also, there is an underlying latent time that is getting calculated, and it is a cell's internal clock which is shared by genes, which precisely defines the cell's location in the biological process. It also takes into consideration the direction of the motion and it's speed.In comparison to the steady-state model, the dynamical model produces velocity estimates that are more consistent, and properly distinguishes transcriptional stages. It yields really well information on the cell states of cycling [5].

3.2 Kinetics of RNA Velocity

The likelihood-based dynamical model expands the use of velocity estimation to the transient system. Here the kinetics are explained with the linear differential equation, which is completely decoupled and deterministic and has constant kinetic rates. The statistical efficacy of the approaches relies on the curvature in the phase picture because existing models struggle to tell if an up- or down-regulation is taking place when there isn't enough curvature. The ratio of spliced and unspliced primarily influences the phase portrait's overall curvature of divergence from the steady-state line [2].

The above figure demonstrates that a modest ratio would produce straight lines rather than an understandable curve. Statistical inference is restricted to genes where splicing is quicker or similar to degradation. The partially observer kinetics found in sub-populations where the genes have been upregulated only at the very end or down-regulated at the very beginning shows ambiguity and only reveals a straight line [2].

The ratio of splicing and degradation rate and the rate of transcription convergence are important factors for kinetic signals

$$\frac{du}{dt} = \alpha - \beta u, \frac{ds}{dt} = \sigma u - \gamma s[2] \tag{3.2}$$

where, β and σ is the splicing rate,for the sake of generality and to take into account technological factors like amplification biases, the splicing rate parameters β and σ are treated very differently [2].

There are residual with concavity and convexity for up-regulation and down-regulation which provides the kinetic signal.

$$C = \int_0^\infty r(t)ds(t) = \int_0^\infty r(t)\frac{ds}{dt}dt = \frac{1}{2}\frac{\beta}{\gamma+\beta}\frac{\alpha}{\beta}\frac{\delta\alpha}{\beta\gamma} = \frac{1}{2}\frac{\beta}{\gamma+\beta}s_{steady}u_{steady}[2] \tag{3.3}$$

3.3 scVelo on Human Embryonic Datasets

The estimation of RNA velocity is done on human embryonic datasets by using scVelo as it was an updated model overcoming the limitations of the previous model.

3.3.1 hESC with Endoderm Progenitors

The dataset of Human embryonic stem cells on Endoderm progenitors was used. After preprocessing, the data started to form the raw SRA files, we got the loom files, and after merging all the loom files, we got a single loom file with 43682 rows, 534 columns, and 4 layers. The loom files consist of the Accession column, Chromosome number, start and end of the sequence, and Gene column. The matrix row consists of cell names or IDs.Further will import the scVelo library and get the proportion of spliced and unspliced mRNAs in our data. We got 71 percent spliced and 29 percent unspliced RNAs in our data. We then filtered the genes with minimum shared counts and normalized the data. The minimum shared counts were set as 20.

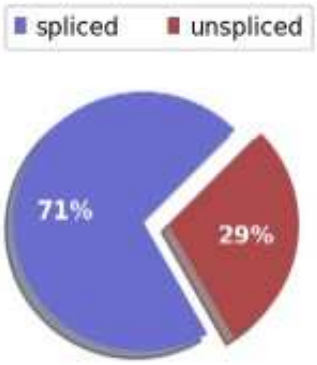

Figure 3.1: Spliced Unspliced Ratio in Endoderm Progenitors

Further, it extracted the highly variable genes and log normalized the data. Furthermore, KNN is applied with 30 neighbors, and keeping the pcs value also as 30. It computed the neighbors and saved the spliced and unspliced abundance in adata.layers. Further velocity of the data is calculated and projected using UMAP.

We got an Anndata object with multiple layers and a 534 x 1986 n obs x n var matrix. It has obs consisting of cells ID, initial spliced and unspliced size, counts, and velocity

self-transition. And has var, which consists of Accession, gene names, chromosome, strand and end, strand, gene counts, means, velocity, etc. It has layers that consist of ambiguous, a count matrix, spliced, unspliced, velocity, and velocity Variance. In this, the velocity layer will be used further as it contains a velocity matrix that consists of the velocity of every cell with respect to every gene. It is the matrix that has both positive and negative velocities. The positive means upregulated, while the negative me downregulated genes.

3.3.2 hESC with Mesoderm Progenitors

The Human embryonic stem cell data with Mesoderm Progenitors were pre-processed for raw SRA files and converted into loom files which were then merged into one file using velocyto. The merged loom file has 43683 rows and 453 columns. It also has four layers with chromosome number, its start and end, and accession and genes. As mentioned in Human Embryonic Stem cell data with endoderm progenitors here also, we have used scVelo and imported its library. And got the ratio of spliced and unspliced mRNAs. Here we got 74 percent spliced and 26 percent unspliced mRNAs. We further follow the same step as in hESC with Endoderm progenitors.

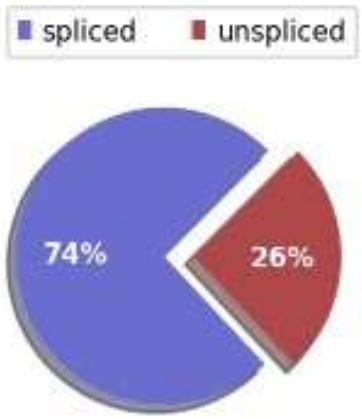

Figure 3.2: Spliced Unspliced Ratio in Mesoderm Progenitors

The Anndata object here is of size 453 x 1988. The obs data consist of initial spliced and unspliced size, n counts, root cells, endpoints, velocity pseudotime, velocity clusters, velocity length, etc. In var, it has velocity qreg ratio, spearman's score, velocity score, etc. The velocity matrix we need further is in layers object. This velocity matrix

is of 453 rows × 1988 columns containing the velocity of every cell with respect to genes with both positive and negative velocities.

3.4 Unipath

Unipath is used to analyze the single-cell transcriptomic and open chromatin data, which was challenging due to its advancements and novel applications. Unipath is used to represent single-cell data through pathways with gene enrichment scores. Unipath used a novel approach for analyzing single cell RNA-seq profiles to tackle the difficulty of single cell representation such as for single cell pathways and gene-set enrichment scores. This method scales the read counts across cells with a varied dropout rate and sequencing depth and not by using general methods such as Poisson distribution or negative Binomial distribution. Unipath uses a standard null model while analyzing single cell ATAC profiles to estimate the pathway enrichment scores, and the enhancers were identified by the global accessibility scores [6].

UniPath is a package in R programming which is used for converting single-cell transcriptome data to pathway scores. Single-cell data production is a highly complicated process, frequently resulting in batch effects or differences in the data. UniPath employs a technique that converts transcriptome profiles to route scores using a powerful statistical method and with great accuracy. UniPath is well-known for its ability to handle artifacts caused by sequencing depths and varied dropout rates. The input given to Unipath is in the p x q matrix, where p is gene names present as row names, and q are the columns in which there are cell IDs or names. The count data may be expressed in the following forms: UMI, RPKM, FPKM, and TPM counts. Also, the non-zero values in the matrix data of gene count follows the log-normal distribution while converting the gene count to pathway scores. Also, it is convertible to p Values. These p values of the gene set are combined using Brown's approach to reduce the covariation among the genes. We get the output as pathway scores as an m x n matrix consisting of pathway names in m as row names and n as column names with cell IDs or names. The pathway scores are calculated using the Monte-Carlo techniques that update the p values, and these updated p values are referred to as pathway scores [6].

3.4.1 Normalization of free Gene-set enrichment

It utilizes the logarithmic value of the expression in the form of FPKM, TPM, RPKM, and UMI-count of genes and each cell is treated individually while calculating the significance of pathway (gene-set) enrichment using scRNA-seq. As for normalization, it estimates the dispersion of a gene's tag count across several samples. It avoids producing artefacts, in contrast to previously published approaches. Due to the varying degree of noise and gene dropout rate among them, scaling and normalization between various cells might produce artefacts.UMI counts are independent of gene-length bais; therefore, Unipath considers it as an expression. A bimodal is created for log distribution i.e. gene expression. This bimodal consist of two-mode one represents genes with zero counts and the other with non-zero counts gene expression representing a normal distribution [6].

Probability Distribution Function for log value -

$$f(x) = p_0 I(x = 0) + (1 - p_0) N(x; \mu, \sigma)[6] \tag{3.4}$$

Here p_0 is a fraction of zero expression genes, including both true expressions and dropout, μ is the mean of non-zero expression in a log scale for every cell, σ is the standard deviation of non-zero expressions in a log scale for every cell, $N(x; \mu, \sigma)$ are the Gaussian probability distribution function for non-zero expression genes, $I(x = 0)$ is an indicator function. The mean and standard deviation is converted to a p-value by using the Gaussian distribution, and then the p-values are combined using the browns method [6].

The combined p-value with k genes having non-zero expression -

$$p_{combined} = 1.0 - \phi_{2f}(\frac{\psi}{c})[6] \tag{3.5}$$

In the following equation the P_i represents the P-value of log of gene i.ψ_{2f} is cumulative distribution function for chi-square distribution

The combined p-value is calculated for each gene set in every cell. A threshold of a minimum of five genes is kept so that the estimation will not be affected by one or

two genes. The combined p-values are also treated with a null-model which is based on permutation based testing, so that the effect of repeated hypothesis testing, promiscuously enriched gene sets, and housekeeping genes can be eliminated. Multiple cells from multiple cell types are chosen, and the scRNA-seq profiles with counts on the same gene list are downloaded in order to make a null model. Coefficient variation criteria are used to cluster the hierarchical genes, and the gene IDs with no symbols are dropped. Highly variable genes are selected and binned based on their mean values. The null model is designed to maintain heterogeneity, and the proportion of cells in the null model is used to calculate pathways adjusted p-value in the specific cell [6].

3.4.2 Unipath on Human Embryonic Datasets

We applied Unipath on both Human Embryonic Dataset datasets for Endoderm progenitor and mesoderm progenitors. We gave the count matrix as FPKM reads, which has gene names as the row names and cell IDs or cell names as the column names. Unipath has precompiled mouse and human null model data. It also has a pathway annotation file and a gene set marker file. We will first import the library and the human null model data. Then we imported symbol data known as c2.cp.v6.1.symbols for humans. Further will import our count matrix and set our row names as gene names and column names as cell IDs or names. Further will normalize the null data and the count matrix using a function binorm and save it in a new variable. Further will combine the c2.cp.v6.1.symbols, human null data, and normalized human null data by keeping the threshold as 2. we will combine the c2.cp.v6.1.symbols, count data, and normalized count data at threshold 2. The score is then calculated by adjusting both the combined data. We performed the same on both the datasets and got the pathway scores for both datasets.

3.5 Finding Correlation and Gene Markers

3.5.1 Correlation

Correlation is a statistical method to find the relationship between two or more variables. There are two methods to calculate the correlation coefficient. Pearson is also

known as the standard correlation coefficient, and the other is Spearman. Here, we have used the standard one to find the relationship between two columns. The correlation coefficient gives the statistical strength of the linear relationship between two variables and lies in the range of -1 to 1. The value 1 represents the positive correlation, 0 represents no correlation, and -1 means the inverse correlation. That is, when one rises, the other will decline. The p-value, which is obtained from the data sample's size and the coefficient's value, is used to determine the statistical significance of correlation coefficients produced through sampling. The standard correlation coefficient is calculated by dividing the covariance by the sum of the standard deviations of the two variables[7] .

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}[7] \tag{3.6}$$

where, ρ_{xy} is Correlation Coefficient,$Cov(x,y)$ represents Covariance of x and y,σ_x is Standard deviation of x, σ_y is standard deviation of y.

The standard deviation is a measurement of how far off the data are from the mean. The correlation coefficient quantifies the strength of the association between the two variables on a normalised scale from -1 to 1, whereas covariance indicates whether the two variables tend to move in the same direction[7].

The above equation can be elaborated as,

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}[7] \tag{3.7}$$

where,r represents correlation Coefficient and n is number of observations

3.5.2 Gene Markers

Gene markers are the known DNA sequence whose chromosome location is known. Chromosomes, genes, and genetic markers near one another are more likely to be inherited together. These markers can be detected by RNA sequencing or any other technique available to analyze the genome, which can further be used to create genetic maps of any organism. The identification and isolation of cells are possible due to cell markers. We have used stem cell markers in our study, and stem cell markers make it possible to identify neuronal, stromal, mesenchymal, and hematopoietic stem cells. One of the

fascinating areas of modern medicine is the study of stem cells because it can enable the creation of novel treatments through the repair, replacement, and regeneration of damaged or ill cells and tissues.

3.5.3 Finding correlation in Human embryonic dataset using gene markers

We have used human embryonic datasets for endoderm and mesoderm progenitors. We have calculated the velocity for both datasets with the dynamic approach and by using the tool scVelo, We got the velocity of each cell with respect to each gene. This velocity matrix have positive an negative values showing the up-regulation and down-regulation of genes. After finding the velocity for both the datasets. We mapped the cell's state and the markers used in that state to poise the cells towards the next lineage, then the mean of all cell markers in the velocity matrix was calculated concerning its cell's ID. Once we get the mean of all the markers will then calculate the correlation between the mean of cell makers and the pathway score, which will then give us the correlation and pathways activity, that is, which pathways are active in forming the lineage. Such as for Endoderm progenitors, the cells from 0hr, 12hr, 24hr, 36hr, 72hr, and 96hr represent a different stage of differentiation. On Day 0, there were human embryonic stem cells (hESC).

Further, on Day 1, it will differentiate into an Anterior primitive streak, and on day 2, Definitive Endoderm, and so on into an anterior foregut, posterior foregut, and midgut. We will take gene markers at each step and correlate them with pathway scores of the following stage. Likewise, We have done this for both datasets and explored the highly correlated pathways.

CHAPTER 4

EXPERIMENTATION

4.1 For hESC with Endoderm Progenitors

In this dataset, we have human embryonic stem cells at an early stage at 0 hr, which will differentiate into Definitive Endoderm, Ectoderm, and Mesoderm. Here we will calculate the correlation between the velocity score and the pathway scores of human embryonic stem cells, concluding which pathways show the higher correlation in the formation of which stage.

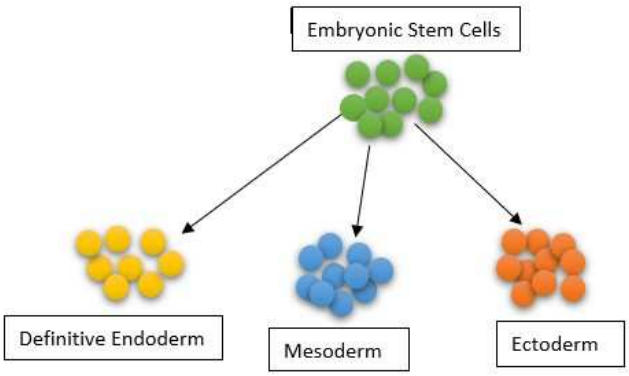

Figure 4.1: Human Embryonic stem cell Differentiation

4.1.1 Definitive Endoderm

For Human embryonic stem cells to show lineage towards definitive endoderm, we took the markers of definitive endoderm, calculated the mean of the velocities of these markers, and correlated it with the pathway scores of human embryonic stem cells at 0 hr. The markers considered were 'EOMES', 'FABP3', 'GATA4', 'HNF4', 'FOXA1', and 'MIXL1'. The mean of the velocities of these markers was calculated and correlated

with pathway scores using Person correlation. We have also considered some other markers of endoderm, such as 'CLDN', 'FOXP2', 'SOX7', and 'SOX10'. However, the best result obtained was from the previously mentioned markers.

4.1.2 Mesoderm

For Human embryonic stem cells to follow the mesoderm lineage, we have considered the mesoderm markers such as 'BMP2K', 'BMP3', 'BMP5', 'BMPER', 'BMP8B', 'GDF3', 'INHBA', 'NODAL', 'WNT' and 'TGFBI'. The mean of these markers was found, and the correlation with pathway scores was calculated. For experimentation purpose, we have considered some more markers such as 'ACVRC1','FGF7', 'FGF12', 'FGF17', 'GDF7', 'WNT2', 'WNT2B', 'WNT1' and 'WNT5A' along with the above-mentioned markers. However, the results with the markers mentioned above alone were more accurate.

4.1.3 Ectoderm

For the differentiation of human embryonic stem cells towards ectoderm, we have considered the following gene markers 'OTX2', 'PAX6', and 'PAX6-AS1'. The correlation was calculated by considering these markers. Also, some experimentation was done by considering other markers, such as 'BMP', ' FGF17', 'FGF13', 'FGF7', 'FGF12', 'PAX8' and 'PAX3' along with the markers mentioned above.

4.2 For hESC with Mesoderm Progenitors

We followed the development roadmap of Mesoderm stem cells for day 0 to day 6, at following each and every stage,

We followed the map shown in Fig 4.1, Starting from Day 0, by giving the accurate markers such as TGFβ and BMP and inhibiting FGF and WNT. It will be differentiated into two lineages, Anterior primitive streak, and Mid primitive streak, on Day 1.

Anterior primitive streak lineage The Anterior primitive streak, with the help

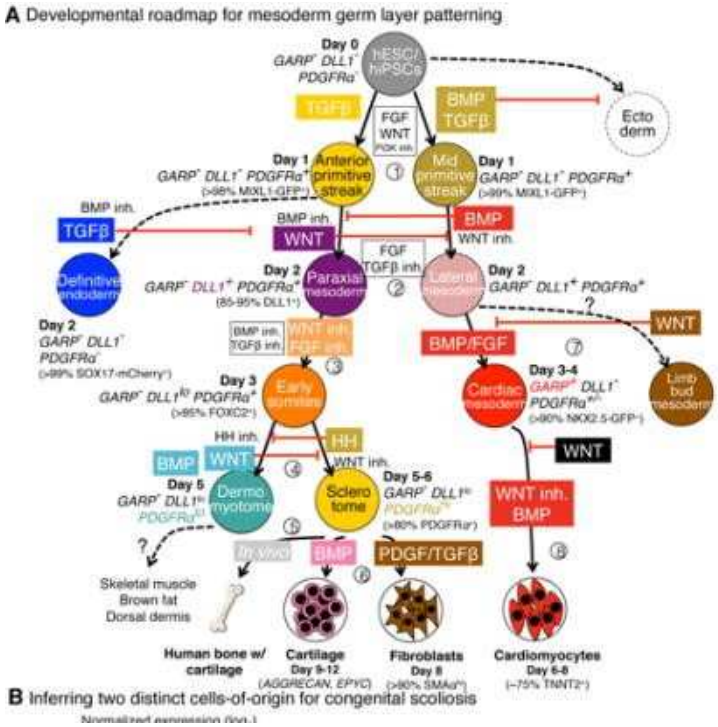

Figure 4.2: Roadmap of mesoderm stem cell differentiation [1]

of WNT and the inhibition of BMP and TGFβ, will form paraxial mesoderm and, on the other hand, with the help of TGFβ it can form definitive mesoderm and will follow its lineage. The paraxial mesoderm will further differentiate into Early somites by inhibiting BMP, WNT, FGF, and TGFβ. The early somites will differentiate into Dermomyotome by inhibiting HH and giving WNT and sclerotome by inhibiting WNT and giving HH. The Dermomyotome will further give rise to skeletal muscles, brown fat, and dorsal dermis, while the sclerotome will form Human bone. But, by providing BMP, it will form cartilage on Days 9 to 12. By supplying PDGF/TGFβ, it will develop fibroblast on Day 8.

Mid primitive streak lineage The Mid primitive streak will further form lateral mesoderm in the presence of BMP and inhibition of WNT and further form cardiac mesoderm in the presence of BMP/FGF and inhibition of WNT. On the other hand, in the presence of WNT, the lateral mesoderm will differentiate into limb bud mesoderm. In the presence of BMP and inhibition of WNT, the cardiac mesoderm will form cardiomyocytes on days 6-8.

4.2.1 Day 0 (hESC)

At Day 0, the cell was at a stage of human embryonic stem cells, which had been collected through Illumina Hiseq. From the velocity matrix calculated above, we will extract the cells at day 0, and the pathway scores of the same cells were extracted. Further will find the gene markers present in the data. For human embryonic stem cells to get differentiated into Anterior primitive streak, TGFβ is required. In our data, we got the 'TGFβI', 'TGFβR1', and 'TGFβ2' markers. We extracted them along the velocity scores, and their mean was calculated. These mean values are then saved in a single row variable and correlated with the pathway scores using Person correlation, and the values will be stored in a new list. This list is further converted into a data frame consisting of pathways and correlated scores. These correlated pathways will show which pathways are highly correlated and will lead to the formation of an Anterior primitive streak. The same procedure is repeated with human embryonic stem cell data but with different genetic markers such as BMP and TGFβ. The markers available in our data were 'BMP3', 'BMPR1B', 'BMPER', 'BMP1', 'TGFβI', 'TGFβR1', 'TGFβ2'. The mean of these markers was taken, and got the correlated pathways leading to the formation of Mid primitive streak.

4.2.2 Day 1

Anterior Primitive streak - The Anterior primitive streak will further differentiate into Definitive mesoderm and Paraxial mesoderm. Will extract the Anterior primitive streaks cells from the velocity matrix and the pathway scores of the same cells, the markers which will lead to the formation of definitive mesoderm is TGFβ and will also take PDGFRα as it is one of the markers for Definitive mesoderm and are present in our data. Along with PDGFRα, 'TGFβI' ,'TGFβR1', 'TGFβ2' markers were also taken and calculated the mean and correlated it with the pathways to find the highly correlated pathways which will lead to the formation of Definitive mesoderm. While for the formation of the Paraxial mesoderm, we have taken the WNT markers and PDGFRα as it is also one of the markers for Paraxial mesoderm. The WNT markers in our data are 'WNT2B', 'WNT5B', 'WNT3', and 'WNT5A'. The mean was taken, and the correlation was calculated.

Mid Primitive streak - For the formation of Mid primitive streak, both BMP and TGFβ markers were required. The markers we found in our data are 'BMP3', 'BMPR1B', 'BMPER', 'BMP1', 'TGFβI', 'TGFβR1', 'TGFβ2' and PDGFRα as PDGFRα is also a marker for the Mid primitive streak. The average of these markers was calculated, and the correlated pathways were obtained which will lead to the formation of Lateral mesoderm.

4.2.3 Day 2

Paraxial mesoderm - Paraxial mesoderm will lead to the formation of early somites. Here the gene markers extracted were PDGFRα and FOX, as these markers are present in Early somites, and the WNT, BMP, FGF, and TGFβ pathways were inhibited. The markers present in our datasets are PDGFRα, 'FOXA1', 'FOXL2NB', 'FOXI1', and 'FOXP2'. The mean of these markers was taken, and correlation was calculated concerning pathways leading to the formation of Early Somites.

Lateral mesoderm - Lateral mesoderm will lead to the formation of cardiac mesoderm and limb bud mesoderm. BMP, FGF, and PDGFRα markers were given for the formation of cardiac mesoderm. The marker spresent in our daatset are 'BMP3', 'BMPR1B', 'BMPER', 'BMP1', 'FGFR1', 'FGF13', 'FGFR2', 'FGF14', and 'FGF9'. The mean of these markers was calculated, and a correlation of pathways was found. While for the formation of Limb bud mesoderm will WNT markers were extracted, such as 'WNT2B', 'WNT5B', 'WNT3', and 'WNT5A', and the correlation with pathways was calculated.

4.2.4 Day 3-4

Early Somites - Early somites will bifurcate into two lineages Dermomyotome and Sclerotome. It will differentiate into Dermomyotome in the presence of WNT signals and inhibition of HH signals. The markers we took were 'WNT2B', 'WNT5B', 'WNT3', and 'WNT5A', and we calculated the correlation of pathways that will lead to the formation of Dermomyotome. While Sclerotome, we took markers such as 'HHIP', 'HHLA1', and 'PDGFRα' and calculated the correlation.

Cardiac mesoderm- The cardiac mesoderm in the presence of BMP will account for cardiomyocytes, the markers we provided were 'BMP3', 'BMPR1B', 'BMPER', BMP1' and 'TNNT2'. 'TNNT2' is a marker for cardiomyocytes. And the average velocity scores of these markers were calculated, and we got the correlation with pathways.

4.2.5 Day 5-6

Sclerotome - Sclerotome will form cartilage in the presence of BMP and fibroblast in the presence of PDGF/TGFβ. The BMP makers present in our data were 'BMP3', 'BMPR1B', 'BMPER', and BMP1' while the PDGF/TGFβ markers present in our data were 'PDGFβ', 'PDGFRα, 'TGFβI', 'TGFβR1' and 'TGFβ2'.

CHAPTER 5

Results

5.1 Results of Human Embryonic datasets with endoderm progenitors

In this dataset we took human wembryonic stem cells at day 0 and shown the lineage towards Definitive Endoderm, Mesoderm and Ectoderm with the help of specific markers.

5.1.1 Mesoderm

By correlation velocity of mesoderm markers in human embryonic stem cells with human embronic stem cells pathways, we got the following highly correlated pathwyas as mentioned in table 5.1

Notch pathway- Notch pathway lead the mesoderm cells to commits towards the cardiac mesoderm from human pulripotent stem cells hence notch pathway is of the prominent pathway in mesoderm cells.

Nodal Singlaing pathway - Nodal pathway is one of the essential pathways for mesoderm development. BMP and WNT signals upregulate the Nodal pathway. This pathway also helps to maintain the pluripotency of embryonic stem cells.

HIF1 pathway - HIF is the Hypoxia Induced Factor which has HIF1α and HIF1β. It helps in the enhancement of mesoderm cells and also helps in the differentiation of mesoderm into cardiac mesoderm.

Hedgehog pathways - We have for multiple hedgehog pathways. Hedgehog pathways are required for organ morphogenesis in all the germ layers.

Table 5.1: Highy correlated pathways for mesoderm

Pathways	*Correlation values*
PID NOTCH PATHWAY	0.24680
ST JNK MAPK PATHWAY	0.22181
WNT SIGNALING	0.19589
REACTOME SIGNALING BY NODAL	0.19256
NABA ECM AFFILIATED	0.18076
PID INTEGRIN A9B1 PATHWAY	0.17905
KEGG PPAR SIGNALING PATHWAY	0.17389
REACTOME GPCR LIGAND BINDING	0.16963
PID HIF1 TFPATHWAY	0.1692
ST ERK1 ERK2 MAPK PATHWAY	0.16706
REACTOME INTEGRIN ALPHAIIB BETA3 SIGNALING	0.16457
REACTOME SIGNALING BY GPCR	0.16317
PID HEDGEHOG 2PATHWAY	0.162269
KEGG HEDGEHOG SIGNALING PATHWAY	0.15939
ST WNT BETA CATENIN PATHWAY	0.15855
PID IL2 PI3K PATHWAY	0.15359
REACTOME SIGNALING BY BMP	0.15015
BIOCARTA ERK5 PATHWAY	0.148580
PID HEDGEHOG GLI PATHWAY	0.148564
ST WNT CA2 CYCLIC GMP PATHWAY	0.14632

WNT signaling pathways - WNT signaling pathway is one of the important pathways for mesoderm differentiation as it helps mesoderm to differentiate into definitive mesoderm, presomitic mesoderm, and cardiac mesoderm.

WNT Beta Catenin - WNT beta-catenin is also a very important pathway in the mesoderm, as it is an essential pathway in early mesoderm cells. It also leads to the activation of MAPK and FGF signals. It activates the important mesodermal genes required for mesoderm differentiation.

5.1.2 Endoderm

Here we have shown top 50 correlated pathways for endoderm. As the given data have endoderm progenitors so we got a good results for endoderm differentiation pathways.

	A	B	C
1	Pathways	Values	
2	BIOCARTA_TCAPOPTOSIS_PATHWAY	0.425081	
3	BIOCARTA_ERK_PATHWAY	0.393522	
4	PID_KIT_PATHWAY	0.378833	
5	REACTOME_CREATION_OF_C4_AND_C2_ACTIVATORS	0.372139	
6	ST_ADRENERGIC	0.352496	
7	PID_INSULIN_PATHWAY	0.332675	
8	BIOCARTA_MAPK_PATHWAY	0.33012	
9	ST_P38_MAPK_PATHWAY	0.329051	
10	REACTOME_EFFECTS_OF_PIP2_HYDROLYSIS	0.322282	
11	ST_T_CELL_SIGNAL_TRANSDUCTION	0.31912	
12	ST_WNT_CA2_CYCLIC_GMP_PATHWAY	0.317084	
13	BIOCARTA_EPHA4_PATHWAY	0.294273	
14	SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	0.283474	
15	BIOCARTA_RARRXR_PATHWAY	0.277642	
16	REACTOME_HORMONE_LIGAND_BINDING_RECEPTORS	0.277017	
17	PID_AP1_PATHWAY	0.267949	
18	REACTOME_GAB1_SIGNALOSOME	0.264901	
19	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULF	0.264728	
20	PID_S1P_S1P2_PATHWAY	0.263266	
21	KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	0.257922	
22	ST_G_ALPHA_I_PATHWAY	0.257723	
23	KEGG_CHEMOKINE_SIGNALING_PATHWAY	0.254697	
24	ST_GRANULE_CELL_SURVIVAL_PATHWAY	0.253712	
25	PID_NECTIN_PATHWAY	0.253297	
26	REACTOME_CLASS_C_3_METABOTROPIC_GLUTAMATE_PHEROMON	0.252628	
27	PID_FAK_PATHWAY	0.252563	

Figure 5.1: Highly correlated pathways for Endoderm

ERK Pathway - ERK pathways has been observed for suppressing the expression of pulripotency gens in order to induce the expression of endoderm [8]. ERK pathways

	A	B
28	BIOCARTA_IL17_PATHWAY	0.250293
29	KEGG_SULFUR_METABOLISM	0.242744
30	ST_JNK_MAPK_PATHWAY	0.241031
31	REACTOME_GLYCOSAMINOGLYCAN_METABOLISM	0.237572
32	PID_RETINOIC_ACID_PATHWAY	0.233569
33	REACTOME_SPHINGOLIPID_DE_NOVO_BIOSYNTHESIS	0.229971
34	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION_IN_THE_MEDIAI	0.228047
35	REACTOME_APOPTOTIC_CLEAVAGE_OF_CELLULAR_PROTEINS	0.226398
36	REACTOME_CHONDROITIN_SULFATE_DERMATAN_SULFATE_METAE	0.222734
37	BIOCARTA_THELPER_PATHWAY	0.220462
38	PID_MAPK_TRK_PATHWAY	0.220362
39	PID_AJDISS_2PATHWAY	0.218761
40	REACTOME_ACTIVATION_OF_RAC	0.217993
41	BIOCARTA_ASBCELL_PATHWAY	0.216594
42	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	0.216093
43	PID_IL1_PATHWAY	0.215974
44	REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMI	0.211495
45	REACTOME_SIGNAL_REGULATORY_PROTEIN_SIRP_FAMILY_INTERA	0.210424
46	BIOCARTA_PTEN_PATHWAY	0.210325
47	BIOCARTA_SPRY_PATHWAY	0.209125
48	ST_GAQ_PATHWAY	0.207983
49	BIOCARTA_CTCF_PATHWAY	0.206547
50	REACTOME_MAPK_TARGETS_NUCLEAR_EVENTS_MEDIATED_BY_MA	0.205999

Figure 5.2: Highly correlated pathways for Endoderm

also induces primitive endoderm differentiation[8].

MAPK pathway- MAPK helps in definitive endoderm formation through FGF signaling[9].

WNT pathways- WNT singling were found in early endoderm, also WNT helps in pattrening of intestinal endoderm.WNT signals are also gets activated in posterior endoderm[10]

GAB1 pathway and EPHA4 pathway - Both these pathways are required for the formation of endoderm [11].

S1P S1P2 pathway- Endodermal survival depends upon the S1P signaling molecules[12]

Chemokine pathway- Chemokine signals are present in endodermal cells and controls the directional migration [13].

Nectin pathway- Nectin is present in extracellular matrix and are glycoprotiens which are located on endodermal cells[11].

FAK pathway- The FAK is one of the highly activated pathway in enoderm. It has been found that the activation level of FAK were highest in endoderm.

Retinoic acid pathway - Retinoic acid is an essential signaling molecule which helps in the development of endoderm, it help to generate edodermal cells from embryonic stem cells [14].

RAC pathway- RAC is an important siganl is visceral endoderm, It helps the visceral endoderm to to move towards the anterior posterior body axis [15].

PTEN pathway- PTEN regulates multiple singaling pathways it help in morphogenesis of lung endoderm with the help of ERK pathway[16].

SPRY pathway - Spry-1 and Spry-2 are the two genes taht are expressed in pharyngeal enododerm.

5.1.3 Ectoderm

The highly correlated pathways for Ectoderm are as follows,

Epha4 pathway- Epha4 is expressed in paraxial an lateral ectoderm, the EphA4 signals has also been identified in midline ectoderm [17].

Cdc42 pathway- CDc42 simulations has been found in ectoderm Cdc42 signals also controls the cell morphogenesis of primary mesenchymal cells[18].

RAR and RXR pathways- RAR helps in the formation of both neural and non-neural ectoderm, they are heterodimers and bound to regulatory DNA.

Calcineurin pathway- Calcineurin is an important signals for neural differentiation of ectoderm. Calcineurin signals NFAT which is also am important factor for neural differentiation of ectoderm [19].

MAPK pathway and sulpher metobolism- MAPK is the mitogen activated protein kinase which help in endoderm neural development. As sulfur is an important component to get the essential nutrients in all organisms and it's development.

GAG and ERK pathway- GAG is found in ectoderm in higher percentage as precursor of sulphate gas [20].And ERK pathway is activated during the formation of en-

	REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY	0.268882
	PID_RXR_VDR_PATHWAY	0.25191
	WNT_SIGNALING	0.248319
	BIOCARTA_RARRXR_PATHWAY	0.2442
	REACTOME_CHONDROITIN_SULFATE_BIOSYNTHESIS	0.242994
)	REACTOME_GLYCOSAMINOGLYCAN_METABOLISM	0.242483
I	PID_ECADHERIN_NASCENT_AJ_PATHWAY	0.231422
2	BIOCARTA_TCAPOPTOSIS_PATHWAY	0.230592
3	PID_HNF3A_PATHWAY	0.230466
4	REACTOME_SIGNALING_BY_HIPPO	0.230014
5	BIOCARTA_CALCINEURIN_PATHWAY	0.227428
5	BIOCARTA_EPHA4_PATHWAY	0.225405
7	REACTOME_APOPTOTIC_CLEAVAGE_OF_CELLULAR_PROTEINS	0.224296
3	REACTOME_IL_7_SIGNALING	0.221502
9	REACTOME_SYNTHESIS_OF_PE	0.221151
)	PID_IL1_PATHWAY	0.21608
I	PID_CDC42_REG_PATHWAY	0.211772
2	REACTOME_CHONDROITIN_SULFATE_DERMATAN_SULFATE_METABOLISM	0.21052
3	KEGG_HOMOLOGOUS_RECOMBINATION	0.207087
4	PID_BMP_PATHWAY	0.206794
5	PID_NFAT_3PATHWAY	0.202622
5	REACTOME_SYNTHESIS_OF_PC	0.199448
7	REACTOME_CREATION_OF_C4_AND_C2_ACTIVATORS	0.199306
3	ST_P38_MAPK_PATHWAY	0.198851
9	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_HEPARAN_SULFATE	0.196616
)	PID_NFAT_TFPATHWAY	0.193835
I	KEGG_SULFUR_METABOLISM	0.192401
2	REACTOME_GLYCEROPHOSPHOLIPID_BIOSYNTHESIS	0.192217

Figure 5.3: Highly correlated pathways for Ectoderm

33	REACTOME_HS_GAG_BIOSYNTHESIS	0.186196
34	REACTOME_MYOGENESIS	0.182518
35	PID_HNF3B_PATHWAY	0.180334
36	PID_BETA_CATENIN_DEG_PATHWAY	0.180256
37	REACTOME_SPHINGOLIPID_DE_NOVO_BIOSYNTHESIS	0.179722
38	KEGG_GLYCOSAMINOGLYCAN_BIOSYNTHESIS_CHONDROITIN_SULFATE	0.176613
39	BIOCARTA_ERK_PATHWAY	0.175699
40	REACTOME_THE_NLRP3_INFLAMMASOME	0.174348
41	ST_WNT_CA2_CYCLIC_GMP_PATHWAY	0.172199
42	ST_STAT3_PATHWAY	0.16769
43	PID_RETINOIC_ACID_PATHWAY	0.165922

Figure 5.4: Highly correlated pathways for Ectoderm

doderm.

Beta catenin pathway- Beta catenin pathway is one of the most important pathway in the formation of ectoderm, It helps in defining enddoerm and endomesoderm [21].

WNT and BMP Pathways- WNT and BMP both helps in the formation of neural crest ectoderm, WNT6 signals are also present in ecotderm [22].

Retinol Acid pathway- Retinol acid is an essential pathway for ectoderm differentiation[14].

5.2 Results of Human Embryonic datasets with mesoderm progenitors

Here we followed the roadmap of mesoderm differentiation lineages. Following are the day wise results,

5.2.1 Day 0

Pathways highly correlated in the formtaion of Anterior Primitive streak and Mid primitive streak

Integrin Pathways-Integrin along with cadherin plays an important role in mesoderm differentiation and formation of primitive streak[19].

MAPK pathways- MAPK pathways along with FGF singaling is required for the differentiation of cells form epiblast to primitive streak of mesoderm [23]

Hedgehog pathway Hedgehog along with FGF signaling patterns the formation of anterior primitive streak [24].

RHO pathways -RHO pathways helps to control the directionality of mesoderm cells movements and leads to the formation of primitive streak [25].

p53 and PI3 Signaling Pathway -P53 activates WNT signals, this WNT signals lead to the induction of anterior primitive streak.[25].PI3 singalling helps in the differentiation of Anterior primitive streak [26].

	ID	values
2	PID_A6B1_A6B4_INTEGRIN_PATHWAY	0.296996
3	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.277272
4	KEGG_FOCAL_ADHESION	0.261207
5	NABA_ECM_REGULATORS	0.25065
6	KEGG_PATHWAYS_IN_CANCER	0.22851
7	PID_PI3KCI_PATHWAY	0.226306
8	KEGG_BLADDER_CANCER	0.224573
9	PID_A6B1_A6B4_INTEGRIN_PATHWAY	0.296996
10	PID_INTEGRIN1_PATHWAY	0.22185
11	REACTOME_SIGNALING_BY_ERBB4	0.211064
12	KEGG_P53_SIGNALING_PATHWAY	0.202604
13	PID_INTEGRIN_A9B1_PATHWAY	0.201182
14	REACTOME_GPCR_LIGAND_BINDING	0.200533
15	NABA_MATRISOME	0.198024
16	NABA_MATRISOME_ASSOCIATED	0.197362
17	KEGG_MAPK_SIGNALING_PATHWAY	0.197043
18	REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNA	0.19328
19	KEGG_ECM_RECEPTOR_INTERACTION	0.192479
20	REACTOME_GROWTH_HORMONE_RECEPTOR_SIGNALING	0.190764
21	REACTOME_INNATE_IMMUNE_SYSTEM	0.189219
22	REACTOME_ACTIVATED_NOTCH1_TRANSMITS_SIGNAL_TO_THE_NUCLEUS	0.188345
23	REACTOME_SIGNALING_BY_GPCR	0.187783

Figure 5.5: Highly correlated pathways for Anterior Primitive Streak

Notch siganiling pathway Notch siganling helps in the division of embryonic stem cells, and the cells in anterior primitive streak activates the Notch1 siganling molecules [27].

TGF Beta Signaling pathway- TGF beta siganling is required for the formation of primitive streak in vitro [28].

5.2.2 Day 1

On day 1 we have given Paraxial mesoderm markers to anetrior premitive streak and Lateral mesoderm markers to mid primitive streak, We got the following highly correlated pathways

Integrin and WNTβ Siganling pathway - Integrin and WNT Beta catenin becomes active in early mesoderm induction and helps in axial patterning.It also improves mesoderm commitmentof pathways [29].

BMP singaling pathway -BMP induce the development of paraxial mesoderm forming Bmp-4 molecules. In paraxial mesoderm It also helps in Vasculogenesis [30].

WNT siganling pathwayWNT siganal helps in the formation of paraxial meso-

	A	B
1	ID	values
2	PID_A6B1_A6B4_INTEGRIN_PATHWAY	0.265501
3	NABA_ECM_REGULATORS	0.263438
4	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.255107
5	REACTOME_SIGNALING_BY_ERBB4	0.251468
6	KEGG_FOCAL_ADHESION	0.243951
7	KEGG_P53_SIGNALING_PATHWAY	0.2356
8	NABA_MATRISOME_ASSOCIATED	0.22641
9	NABA_MATRISOME	0.225242
10	PID_BMP_PATHWAY	0.214744
11	PID_INTEGRIN_A9B1_PATHWAY	0.211628
12	NABA_ECM_GLYCOPROTEINS	0.209745
13	NABA_CORE_MATRISOME	0.209552
14	REACTOME_GPCR_LIGAND_BINDING	0.206583
15	KEGG_ERBB_SIGNALING_PATHWAY	0.206111
16	KEGG_TGF_BETA_SIGNALING_PATHWAY	0.205875
17	KEGG_PATHWAYS_IN_CANCER	0.202278
18	PID_INTEGRIN1_PATHWAY	0.200479
19	PID_LYMPH_ANGIOGENESIS_PATHWAY	0.198377
20	KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAM	0.197616
21	REACTOME_SIGNALING_BY_GPCR	0.195107
22	REACTOME_NFKB_AND_MAP_KINASES_ACTIVATIO	0.194982
23	KEGG_ECM_RECEPTOR_INTERACTION	0.194078
24	REACTOME_GPCR_DOWNSTREAM_SIGNALING	0.191084
25	KEGG_MAPK_SIGNALING_PATHWAY	0.187662
26	KEGG_BASAL_CELL_CARCINOMA	0.184594
27	REACTOME_INNATE_IMMUNE_SYSTEM	0.179827
28	REACTOME_GROWTH_HORMONE_RECEPTOR_SIGN	0.1793

hesc correlation pathways 2

Figure 5.6: Highly correlated pathways for Mid Primitive Streak

	ID	values
2	REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	0.365877
3	ST_WNT_BETA_CATENIN_PATHWAY	0.307573
4	REACTOME_HEMOSTASIS	0.276236
5	REACTOME_REGULATION_OF_INSULIN_SECRETION_BY_GLUCAGC	0.263338
6	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	0.256198
7	PID_MYC_REPRESS_PATHWAY	0.254087
8	REACTOME_FACTORS_INVOLVED_IN_MEGAKARYOCYTE_DEVELOF	0.246511
9	PID_BMP_PATHWAY	0.229823
10	REACTOME_GPCR_LIGAND_BINDING	0.228726
11	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	0.214419
12	REACTOME_CRMPS_IN_SEMA3A_SIGNALING	0.214374
13	KEGG_WNT_SIGNALING_PATHWAY	0.198432
14	WNT_SIGNALING	0.197796
15	PID_ILK_PATHWAY	0.197258
16	REACTOME_THE_ROLE_OF_NEF_IN_HIV1_REPLICATION_AND_DIS	0.184761
17	KEGG_MELANOGENESIS	0.178195
18	NABA_SECRETED_FACTORS	0.176089
19	REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_	0.175108
20	REACTOME_DNA_REPLICATION	0.170633
21	ST_GA13_PATHWAY	0.163504
22	REACTOME_GROWTH_HORMONE_RECEPTOR_SIGNALING	0.162281
23	BIOCARTA_P38MAPK_PATHWAY	0.160636
24	REACTOME_SPHINGOLIPID_METABOLISM	0.160436
25	REACTOME_SIGNALING_BY_NOTCH	0.159942
26	REACTOME_KINESINS	0.156595

Figure 5.7: Highly correlated pathways for Paraxial mesoderm

	A	B	C
1	PID_INTEGRIN_A9B1_PATHWAY	0.355893	
2	REACTOME_INTEGRATION_OF_ENERGY_METABOLISM	0.312073	
3	KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	0.288363	
4	KEGG_CALCIUM_SIGNALING_PATHWAY	0.284147	
5	KEGG_FRUCTOSE_AND_MANNOSE_METABOLISM	0.262135	
6	REACTOME_SIGNALING_BY_FGFR	0.261756	
7	SIG_CHEMOTAXIS	0.259831	
8	KEGG_LONG_TERM_DEPRESSION	0.255245	
9	KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	0.240606	
10	REACTOME_GPCR_DOWNSTREAM_SIGNALING	0.234762	
11	REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES	0.231319	
12	REACTOME_PLATELET_HOMEOSTASIS	0.225636	
13	KEGG_PROSTATE_CANCER	0.21109	
14	REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	0.206762	
15	REACTOME_S_PHASE	0.203055	
16	PID_P53_DOWNSTREAM_PATHWAY	0.202664	
17	REACTOME_OPIOID_SIGNALLING	0.20239	
18	REACTOME_GLYCOLYSIS	0.19949	
19	REACTOME_GASTRIN_CREB_SIGNALLING_PATHWAY_VIA_PKC_AND_MAPK	0.197492	
20	REACTOME_REGULATION_OF_INSULIN_SECRETION	0.197049	
21	BIOCARTA_GLYCOLYSIS_PATHWAY	0.197044	
22	REACTOME_DOWNSTREAM_SIGNALING_OF_ACTIVATED_FGFR	0.196752	
23	REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	0.196025	
24	REACTOME_MUSCLE_CONTRACTION	0.193614	
25	REACTOME_G_ALPHA_Q_SIGNALLING_EVENTS	0.189568	
26	KEGG_ENDOCYTOSIS	0.189454	
27	PID_AVB3_OPN_PATHWAY	0.18868	
28	PID_GLYPICAN_1PATHWAY	0.188659	

Figure 5.8: Highly correlated pathways for Lateral mesoderm

derm, it influence Hox gene expression which is important in the formation of paraxial mesoderm [31].

Notch Siganling Pathway Notch siganling pathway are present in paraxial mesoderm which will further helps in formation of somites [32].

MAPK pathway MAPK and ERK helps further in formation of somites and are present in mesoderm [33].

FGFR singaling pathway - FGFR siganling pathway helps in the morphogenesis of mesoderm and further differentiation of latera mesoderm.FGFR also induces signals that helps in fate patterning in mesoderm [34].

Integrin and chemotaxis pathway -Integrin helps to regulate XGIPC and helps to initiate gastrulation in mesoderm cells. while chemotaxis helps in migration of mesoderm cells during gastrulation [19].

Opoid pathway -Opioid is a growth factor that involves in the development of embryonic stem cells and it fates such as mesoderm, endoderm and ectoderm [35].

5.2.3 Day 2

Here we took early somites markers along with paraxial mesoderm and got the following pathways,

ERBB siganling pathways Erbb4 expressed in anterior somites and early somites, it also appears in dermal mesenchyme cells this are above the smoites.

PDGF Siganling pathways PDGF Signals lead to the activation of PI3 and it also controls mesoderm differentiation and it's migration it also controls the directional information [36].

As in the dataset pathways such as WNT, BMP, FGF and TGFβ were inhibited for the formation of early somites, therefore we didn't took these markers and as a results we got only limited pathways.

	A	B	C
1	Pathways	Values	
2	REACTOME_SIGNALING_BY_ERBB2	0.605074	
3	REACTOME_DOWNSTREAM_SIGNALING_OF_ACTIVATED_FGFR	0.409503	
4	KEGG_PURINE_METABOLISM	0.353635	
5	KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	0.3455	
6	REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	0.321993	
7	REACTOME_DOWNSTREAM_SIGNAL_TRANSDUCTION	0.309255	
8	REACTOME_SIGNALING_BY_ERBB4	0.276903	
9	BIOCARTA_KERATINOCYTE_PATHWAY	0.254807	
10	KEGG_GLYCOLYSIS_GLUCONEOGENESIS	0.232792	
11	PID_SYNDECAN_2_PATHWAY	0.22352	
12	PID_ENDOTHELIN_PATHWAY	0.210032	
13	KEGG_DILATED_CARDIOMYOPATHY	0.185954	
14	KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	0.18407	
15	KEGG_P53_SIGNALING_PATHWAY	0.162311	
16	REACTOME_RECRUITMENT_OF_MITOTIC_CENTROSOME_PROTEINS_AND_COMPLEX	0.162152	
17	REACTOME_LIPOPROTEIN_METABOLISM	0.157624	
18	KEGG_MTOR_SIGNALING_PATHWAY	0.15585	
19	REACTOME_HEPARAN_SULFATE_HEPARIN_HS_GAG_METABOLISM	0.152487	
20	REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_	0.132052	
21	KEGG_ERBB_SIGNALING_PATHWAY	0.124047	
22	REACTOME_GLYCOSAMINOGLYCAN_METABOLISM	0.122573	
23	REACTOME_SIGNALING_BY_PDGF	0.119942	
24	KEGG_PATHWAYS_IN_CANCER	0.11745	
25	KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.110658	
26	NABA_ECM_GLYCOPROTEINS	0.108455	

Figure 5.9: Highly correlated pathways for Early somites

5.2.4 Day 3-4

Early somites will lead to the formation of Dermomyotome and Sclerotome. We got teh following correlated pathways for Dermomyotome (dorsal-somites) and Sclerotome (ventral-somites),

5.2.5 Day 4-5

The Sclerotome will further differntiate into cartilage and fibroblast, we got the following highly correlated pathways.

	A	B
1	Pathways	values
2	REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	0.289294
3	bIOCARTA_CARDIACEGF_PATHWAY	0.28916
4	PID_BMP_PATHWAY	0.246576
5	REACTOME_VOLTAGE_GATED_POTASSIUM_CHANNELS	0.240371
6	KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLOR	0.233474
7	KEGG_HEDGEHOG_SIGNALING_PATHWAY	0.232423
8	REACTOME_TRANSPORT_OF_INORGANIC_CATIONS_ANIONS_	0.224643
9	REACTOME_DEVELOPMENTAL_BIOLOGY	0.223296
10	KEGG_LYSOSOME	0.220891
11	REACTOME_SIGNALING_BY_GPCR	0.19606
12	REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX	0.193377
13	REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLE	0.19144
14	REACTOME_GPCR_LIGAND_BINDING	0.189354
15	REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	0.179386
16	KEGG_STEROID_HORMONE_BIOSYNTHESIS	0.17286
17	WNT_SIGNALING	0.171234
18	PID_CASPASE_PATHWAY	0.170978
19	KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.167217
20	REACTOME_TRANSCRIPTIONAL_REGULATION_OF_WHITE_ADIF	0.163655
21	KEGG_CELL_ADHESION_MOLECULES_CAMS	0.159757
22	PID_AP1_PATHWAY	0.1484
23	REACTOME_SPHINGOLIPID_METABOLISM	0.140819

Figure 5.10: Highly correlated pathways for Dermomyotome

	A	B	C
1	pathways	values	
2	REACTOME_VOLTAGE_GATED_POTASSIUM_CHANNELS	0.356679	
3	REACTOME_METABOLISM_OF_NUCLEOTIDES	0.319012	
4	PID_LKB1_PATHWAY	0.317492	
5	REACTOME_ASPARAGINE_N_LINKED_GLYCOSYLATION	0.301733	
6	REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_	0.300221	
7	KEGG_TYPE_II_DIABETES_MELLITUS	0.281408	
8	REACTOME_NEUROTRANSMITTER_RECEPTOR_BINDING_AND_DOWNSTREAM_	0.267152	
9	PID_HDAC_CLASSI_PATHWAY	0.250574	
10	REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	0.24907	
11	REACTOME_NEURONAL_SYSTEM	0.208271	
12	REACTOME_ACTIVATED_TLR4_SIGNALLING	0.206239	
13	REACTOME_CELL_CELL_JUNCTION_ORGANIZATION	0.206185	
14	REACTOME_CELL_CELL_COMMUNICATION	0.203308	
15	PID_ARF6_PATHWAY	0.200473	
16	ST_FAS_SIGNALING_PATHWAY	0.194096	
17	REACTOME_INTEGRATION_OF_ENERGY_METABOLISM	0.192574	
18	KEGG_INSULIN_SIGNALING_PATHWAY	0.18971	
19	PID_BMP_PATHWAY	0.188532	
20	KEGG_GNRH_SIGNALING_PATHWAY	0.184604	
21	REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	0.182072	
22	REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	0.180829	
23	REACTOME_CELL_JUNCTION_ORGANIZATION	0.17824	
24	REACTOME_INNATE_IMMUNE_SYSTEM	0.17609	
25	REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIG	0.174828	
26	PID_CXCR4_PATHWAY	0.171176	
27	REACTOME_GENERIC_TRANSCRIPTION_PATHWAY	0.169425	

Figure 5.11: Highly correlated pathways for Sclerotome

	A	B
1	Pathways	values
2	KEGG_HYPERTROPHIC_CARDIOMYOPATHY_HCM	0.22472
3	REACTOME_ASPARAGINE_N_LINKED_GLYCOSYLATION	0.222959
4	KEGG_PURINE_METABOLISM	0.205189
5	KEGG_DILATED_CARDIOMYOPATHY	0.201529
6	REACTOME_GABA_RECEPTOR_ACTIVATION	0.200259
7	KEGG_CALCIUM_SIGNALING_PATHWAY	0.197548
8	KEGG_GLIOMA	0.193768
9	REACTOME_POST_TRANSLATIONAL_PROTEIN_MODIFICATION	0.192607
10	REACTOME_METABOLISM_OF_CARBOHYDRATES	0.181858
11	REACTOME_METABOLISM_OF_NUCLEOTIDES	0.174819
12	REACTOME_PLATELET_HOMEOSTASIS	0.169666
13	PID_ECADHERIN_STABILIZATION_PATHWAY	0.157234
14	KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	0.14386
15	PID_E2F_PATHWAY	0.136979
16	REACTOME_GLYCOSAMINOGLYCAN_METABOLISM	0.13687
17	REACTOME_HEPARAN_SULFATE_HEPARIN_HS_GAG_METABOLISM	0.136424
18	KEGG_PANCREATIC_CANCER	0.135526
19	KEGG_ADHERENS_JUNCTION	0.13395
20	REACTOME_SEMAPHORIN_INTERACTIONS	0.129971
21	BIOCARTA_MAPK_PATHWAY	0.129889
22	KEGG_COLORECTAL_CANCER	0.127621
23	BIOCARTA_ALK_PATHWAY	0.122117
24	REACTOME_PRE_NOTCH_EXPRESSION_AND_PROCESSING	0.121688
25	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	0.121546
26	KEGG_PYRIMIDINE_METABOLISM	0.119588
27	KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	0.11279

Figure 5.12: Highly correlated pathways in Cartilage formation

Pathways	values
PID_TCPTP_PATHWAY	0.34592
ST_INTEGRIN_SIGNALING_PATHWAY	0.335996
REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE	0.329024
PID_ENDOTHELIN_PATHWAY	0.314004
PID_LYSOPHOSPHOLIPID_PATHWAY	0.303224
KEGG_JAK_STAT_SIGNALING_PATHWAY	0.295662
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	0.289573
KEGG_OOCYTE_MEIOSIS	0.270559
PID_CDC42_PATHWAY	0.270493
REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	0.25838
REACTOME_FATTY_ACID_TRIACYLGLYCEROL_AND_KETONE_BODY_METABOLISM	0.254179
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	0.254089
REACTOME_INTERFERON_GAMMA_SIGNALING	0.247582
REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	0.245688
REACTOME_TRANSCRIPTIONAL_REGULATION_OF_WHITE_ADIPOCYTE_DIFFERENTI	0.242275
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	0.241498
PID_HEDGEHOG_2PATHWAY	0.240238
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS	0.229253
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	0.225316
REACTOME_DOWNSTREAM_SIGNALING_OF_ACTIVATED_FGFR	0.220769

Figure 5.13: Highly correlated pathways in Fibroblast formation

5.3 Applying Baysien Network

We then applied baysien network on human embryonic stem cell with endoderm progenitors database. And we have found some pathways which are directly associated with mesoderm, ectoderm and endoderm.Here Hill climbing alogorithm is used with 1001 iterations to get the baysien network.

For mesoderm we got the ST WNT BETA CATENIN PATHWAY directly associated with mesoderm, which is an essential pathway for mesoderm development It also leads to the activation of MAPK and FGF signals. Other pathways such as KEGG PPAR SIGNALING PATHWAY AND PID NOTCH PATHWAY were also found to be associated with ST WNT BETA CATENIN PATHWAY and indirecty with mesoderm.

For ectoderm we got the BIOCARTA CALCINEURIN PATHWAY directly associated with ectoderm, BIOCARTA CALCINEURIN PATHWAY signals NFAT which helps in ectoderm differentiation [19]. We also found other pathways associating indirectly with endoderm, such as PID RXR VDR PATHWAY, RECTOME HS GAG BIOSYNTHESIS, PID RETINOIC ACID PATHWAY etc.

For endoderm we got the KEGG CHEMOKINE SIGNALING PATHWAY directly

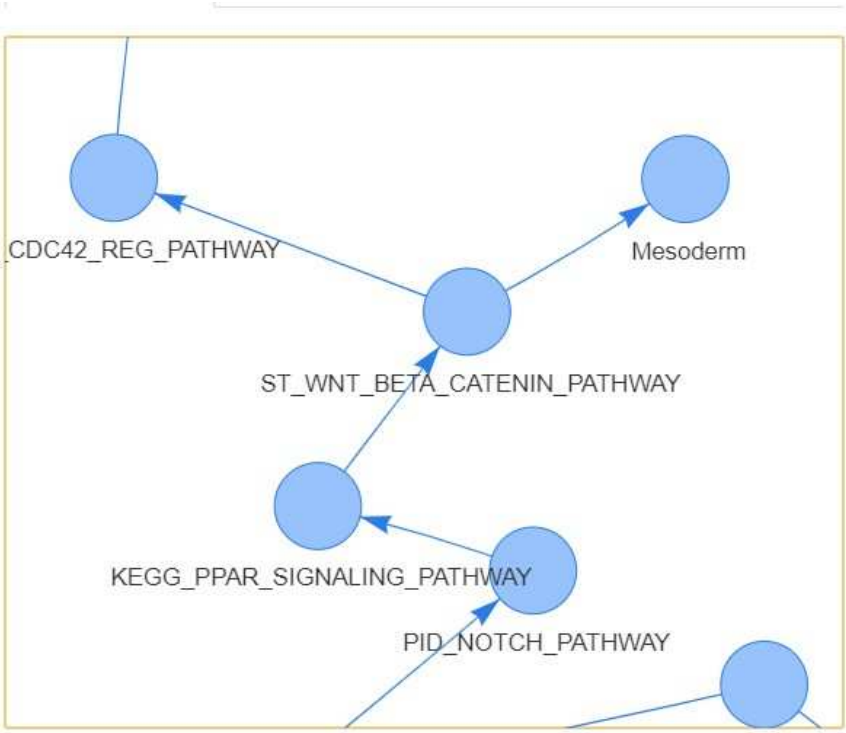

Figure 5.14: Pathways Associated with Mesoderm

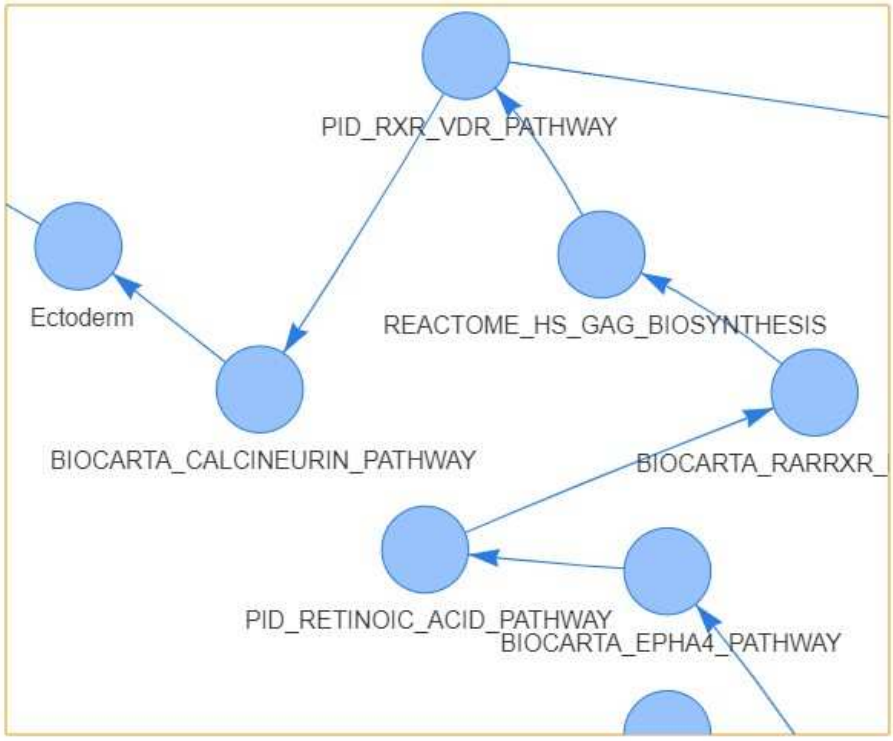

Figure 5.15: Pathways Associated with Ectoderm

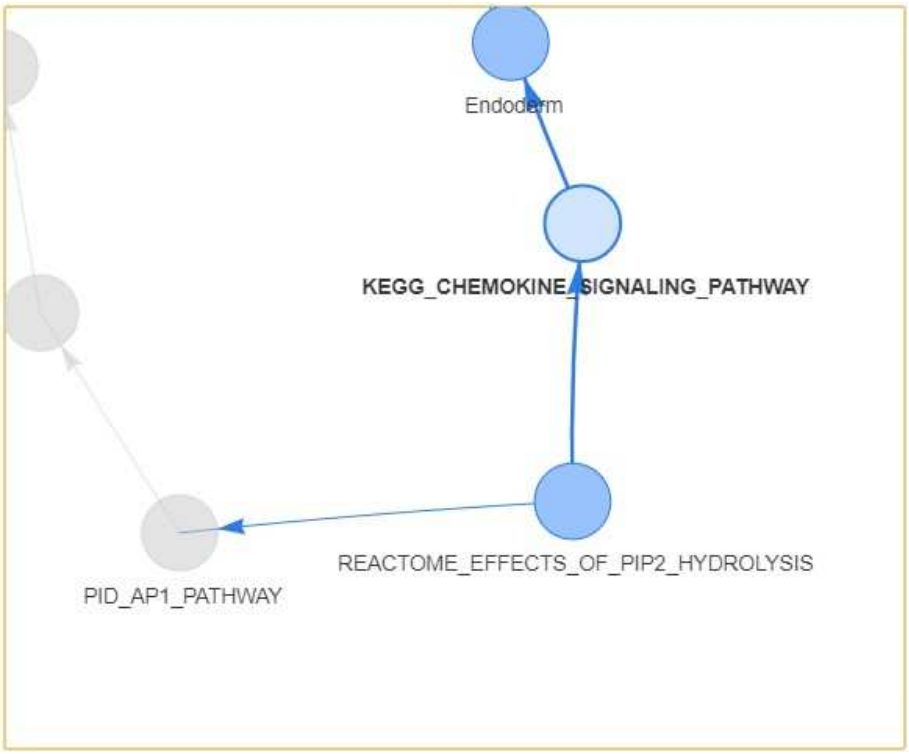

Figure 5.16: Pathways Associated with Endoderm

associated with endoderm, KEGG CHEMOKINE SIGNALING PATHWAY helps in the directional migration of endoderm [13]. Some other pathways such as RECTOME EFFECTS OF PIP2 HYDROLYSIS and PID AP1 PATHWAY were also found.

CHAPTER 6

Conclusion and Future scope

6.0.1 Conclusion

Here we conclude that in some of the lineage we have got an absolutely great result, by getting the maximum correlated pathways which was involved in the formation of that particular state. Here we state a unique in scilico method to find the pathways enriched in a particular lineage or which pathways will be essential to form a particular lineage. We have found pathways directing the cells towards a particular lineage. Such that in human embryonic stem cells with endoderm progenitors we have seen the human embryonic stem cells at Day 0 with endoderm markers has shown a great result and maximum number of highly correlated pathways helping the formation of endoderm cells. We have seen prominant pathways such as WNT, ERK , Nectin etc.

In the second dataset we have got a highly correlated pathways during intial differentiation of embryonic stem cells into primitive streak, we have got essential pathways like BMP, WNT, Nodal, Notch, Hedgehog etc. It shows that by providing accurate markers in silico we can find proficient pathways involved in the formation of a particular cell types.

6.0.2 Future scope and challenges

By having the knowledge of pathways involved in the formation of particular lineage we can direct desired cell formation in vitro by supplying the essential pathways and inhibiting the unwanted pathways. Further it can also had a great application in formation or organoids by directing the useful pathways. However the major challenge is this directing of cells through pathways sometime may lead to unwanted cell growth, to overcome this is one of the major challenge.

REFERENCES

[1] K. M. Loh, A. Chen, P. W. Koh, T. Z. Deng, R. Sinha, J. M. Tsai, A. A. Barkal, K. Y. Shen, R. Jain, R. M. Morganti, N. Shyh-Chang, N. B. Fernhoff, B. M. George, G. Wernig, R. E. Salomon, Z. Chen, H. Vogel, J. A. Epstein, A. Kundaje, W. S. Talbot, P. A. Beachy, L. T. Ang, and I. L. Weissman, "Mapping the pairwise choices leading from pluripotency to human bone, heart, and other mesoderm cell types," vol. 166, no. 2, pp. 451–467. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867416307401

[2] V. Bergen, R. A. Soldatov, P. V. Kharchenko, and F. J. Theis, "RNA velocity—current challenges and future perspectives," vol. 17, no. 8. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.15252/msb.202110282

[3] K. Loh, L. Ang, J. Zhang, V. Kumar, J. Ang, J. Auyeong, K. Lee, S. Choo, C. Lim, M. Nichane, J. Tan, M. Noghabi, L. Azzola, E. Ng, J. Durruthy-Durruthy, V. Sebastiano, L. Poellinger, A. Elefanty, E. Stanley, Q. Chen, S. Prabhakar, I. Weissman, and B. Lim, "Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations," vol. 14, no. 2, pp. 237–252. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1934590913005560

[4] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," vol. 560, no. 7719, pp. 494–498. [Online]. Available: http://www.nature.com/articles/s41586-018-0414-6

[5] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing RNA velocity to transient cell states through dynamical modeling," vol. 38, no. 12, pp. 1408–1414. [Online]. Available: https://www.nature.com/articles/s41587-020-0591-3

[6] S. Chawla, S. Samydurai, S. L. Kong, Z. Wu, Z. Wang, W. L. Tam, D. Sengupta, and V. Kumar, "UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles," vol. 49, no. 3, pp. e13–e13. [Online]. Available: https://academic.oup.com/nar/article/49/3/e13/6020193

[7] R. Taylor, "Interpretation of the correlation coefficient: A basic review," *Journal of Diagnostic Medical Sonography*, vol. 6, no. 1, pp. 35–39, 1990. [Online]. Available: https://doi.org/10.1177/875647939000600106

[8] S. Karaçali, "Human embryonic stem cell n-glycan features relevant to pluripotency," vol. 40, pp. 1050–1058. [Online]. Available: https://journals.tubitak.gov.tr/biology/vol40/iss5/10

[9] L. Sui, J. K. Mfopou, M. Geens, K. Sermon, and L. Bouwens, "FGF signaling via MAPK is required early and improves activin a-induced definitive endoderm formation from human embryonic stem cells," vol. 426, no. 3, pp. 380–385.

[10] R. I. Sherwood, R. Maehr, E. O. Mazzoni, and D. A. Melton, "Wnt signaling specifies and patterns intestinal endoderm," vol. 128, no. 7, pp. 387–400.

[11] F. Zito, E. Nakano, S. Sciarrino, and V. Matranga, "Regulative specification of ectoderm in skeleton disrupted sea urchin embryos treated with monoclonal antibody to pl-nectin," vol. 42, no. 5, pp. 499–506.

[12] H. Fukui, K. Terai, H. Nakajima, A. Chiba, S. Fukuhara, and N. Mochizuki, "S1p-yap1 signaling regulates endoderm formation required for cardiac precursor cell migration in zebrafish," vol. 31, no. 1, pp. 128–136.

[13] S. Nair and T. F. Schilling, "Chemokine signaling controls endodermal migration during zebrafish gastrulation," vol. 322, no. 5898, pp. 89–92. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2770598/

[14] G. M. Kelly and T. A. Drysdale, "Retinoic acid and the development of the endoderm," vol. 3, no. 2, pp. 25–56, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2221-3759/3/2/25

[15] I. Migeotte, J. Grego-Bessa, and K. V. Anderson, "Rac1 mediates morphogenetic responses to intercellular signals in the gastrulating mouse embryo," vol. 138, no. 14, pp. 3011–3020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119308/

[16] Y. Xing, R. Wang, C. Li, and P. Minoo, "PTEN regulates lung endodermal morphogenesis through MEK/ERK pathway," vol. 408, no. 1, pp. 56–65.

[17] C. Schmidt, B. Christ, M. Maden, B. Brand-Saberi, and K. Patel, "Regulation of epha4 expression in paraxial and lateral plate mesoderm by ectoderm-derived signals," vol. 220, no. 4, pp. 377–386.

[18] S. P. Sepúlveda-Ramírez, L. Toledo-Jacobo, J. H. Henson, and C. B. Shuster, "Cdc42 controls primary mesenchyme cell morphogenesis in the sea urchin embryo," vol. 437, no. 2, pp. 140–151. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5973877/

[19] A. Cho, Y. Tang, J. Davila, S. Deng, L. Chen, E. Miller, M. Wernig, and I. A. Graef, "Calcineurin signaling regulates neural induction through antagonizing the BMP pathway," vol. 82, no. 1, pp. 109–124. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011666/

[20] J. E. Pintar, "Distribution and synthesis of glycosaminoglycans during quail neural crest morphogenesis," vol. 67, no. 2, pp. 444–464. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0012160678902117

[21] S. Darras, J. Gerhart, M. Terasaki, M. Kirschner, and C. J. Lowe, "-catenin specifies the endomesoderm and defines the posterior organizer of the hemichordate saccoglossus kowalevskii," vol. 138, no. 5, pp. 959–970. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035098/

[22] M. I. García-Castro, C. Marcelle, and M. Bronner-Fraser, "Ectodermal wnt function as a neural crest inducer," vol. 297, no. 5582, pp. 848–851.

[23] K. M. Hardy, T. A. Yatskievych, J. Konieczka, A. S. Bobbs, and P. B. Antin, "FGF signalling through RAS/MAPK and PI3k pathways regulates cell movement and gene expression in the chicken primitive streak without affecting e-cadherin expression," vol. 11, p. 20. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3071786/

[24] A. Guzzetta, M. Koska, M. Rowton, K. R. Sullivan, J. Jacobs-Li, J. Kweon, H. Hidalgo, H. Eckart, A. D. Hoffmann, R. Back, S. Lozano, A. M. Moon, A. Basu, M. Bressan, S. Pott, and I. P. Moskowitz, "Hedgehog–FGF signaling axis patterns anterior mesoderm during gastrulation," vol. 117, no. 27, pp. 15 712–15 723, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1914167117

[25] V. Stankova, N. Tsikolia, and C. Viebahn, "Rho kinase activity controls directional cell movements during primitive streak formation in the rabbit embryo," vol. 142, no. 1, pp. 92–98. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299133/

[26] S. N. Villegas, M. Rothová, M. E. Barrios-Llerena, M. Pulina, A.-K. Hadjantonakis, T. Le Bihan, S. Astrof, and J. M. Brickman, "PI3k/akt1 signalling specifies foregut precursors by generating regionalized extra-cellular matrix," vol. 2, p. e00806. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871052/

[27] M. B. Favarolo and S. L. López, "Notch signaling in the division of germ layers in bilaterian embryos," vol. 154, pp. 122–144. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925477318300650

[28] P. Gadue, T. L. Huber, P. J. Paddison, and G. M. Keller, "Wnt and TGF-signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells," vol. 103, no. 45, pp. 16 806–16 811, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/full/10.1073/pnas.0603916103

[29] A. Schohl and F. Fagotto, "A role for maternal -catenin in early mesoderm induction in xenopus," vol. 22, no. 13, pp. 3303–3313. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165652/

[30] R. G. James and T. M. Schultheiss, "Bmp signaling promotes intermediate mesoderm gene expression in a dose-dependent, cell-autonomous and translation-dependent manner," vol. 288, no. 1, pp. 113–125.

[31] M. Hofmann, K. Schuster-Gossler, M. Watabe-Rudolph, A. Aulehla, B. G. Herrmann, and A. Gossler, "WNT signaling, in synergy with t/TBX6, controls notch signaling by regulating dll1 expression in the presomitic mesoderm of mouse embryos," vol. 18, no. 22, pp. 2712–2717. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC528888/

[32] Further development 17.1: Notch signaling and somite formation - developmental biology 12e student resources - learning link. [Online]. Available: https://learninglink.oup.com/access/content/barresi-12e-student-resources/barresi-12e-further-development-17-1-notch-signaling-and-somite-formation

[33] M.-C. Delfini, J. Dubrulle, P. Malapert, J. Chal, and O. Pourquié, "Control of the segmentation process by graded MAPK/ERK activation in the chick embryo," vol. 102, no. 32, pp. 11 343–11 348. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1183560/

[34] T.-P. Fan, H.-C. Ting, J.-K. Yu, and Y.-H. Su, "Reiterative use of FGF signaling in mesoderm development during embryogenesis and metamorphosis in the hemichordate ptychodera flava," vol. 18, p. 120. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6091094/

[35] M. Oginuma, P. Moncuquet, F. Xiong, E. Karoly, J. Chal, K. Guevorkian, and O. Pourquié, "A gradient of glycolytic activity coordinates FGF and wnt signaling during elongation of the body axis in amniote embryos," vol. 40, no. 4, pp. 342–353.e10. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5403012/

[36] X. Yang, H. Chrisman, and C. J. Weijer, "PDGF signalling controls the migration of mesoderm cells during chick gastrulation by regulating n-cadherin expression," vol. 135, no. 21, pp. 3521–3530.