Estimating RNA Velocity and Pathway Activity of Single cell RNA and finding correlation between them.

Ariba Ansari MT20336
Rajat Talukdar MT20343
Nitesh Narwade MT20329
Prabhat Singh MT20341
Group number - 8

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Introduction

>Understanding Cell's fate is of prime importance in cellular process, as it affects all aspects of its behaviour.

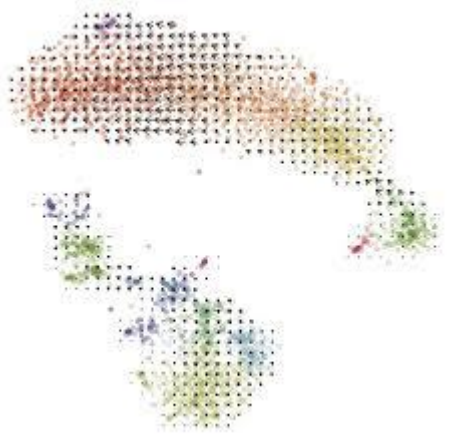
>These behaviour defines cell's morphology,migratory status and proliferation associated with its differentiation state.

>One such way to determine the state of an individual cell is by measuring it RNA abundance and its Dynamics.

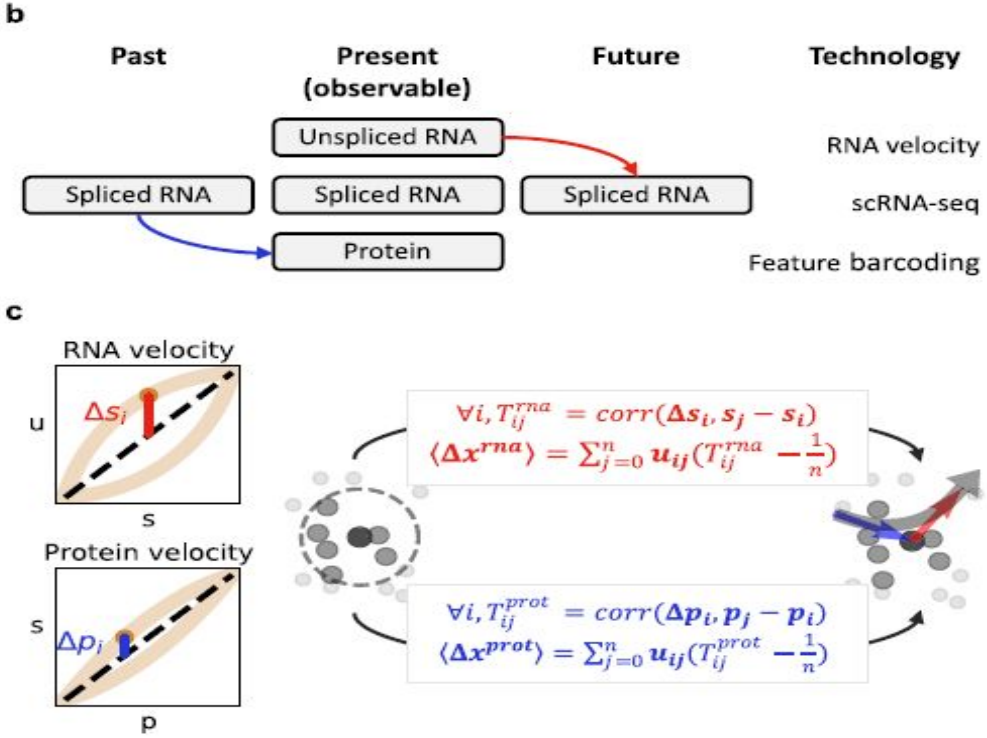
>This can be done by estimating its RNA velocity which is time derivative of gene expression state and it predicts the cell's fate in the timescale of hours, by distinguishing between spliced and unspliced mRNA's.

>RNA velocity helps in analysis of cellular dynamics and developmental lineages.

> Similarly, Unipath a novel method used to predict temporal order of single cells using pathway enrichment scores helps in getting correct order of cells and analyzing its heterogeneity.



Introduction



Model for transcriptional dynamics for quantification of time dependent relationship between pre and matured mRNA, in which the RNA velocity is estimated by the balance between production of spliced mRNA from unspliced mRNA, and the mRNA degradation.

Objective

>In our project we have estimated the RNA velocity from single-cell RNA-seq data followed by using unipath to get pathway enrichment score and finally tried to correlate it with the RNA velocity to get meaningful information.

Tools Used

>SRA Toolkit

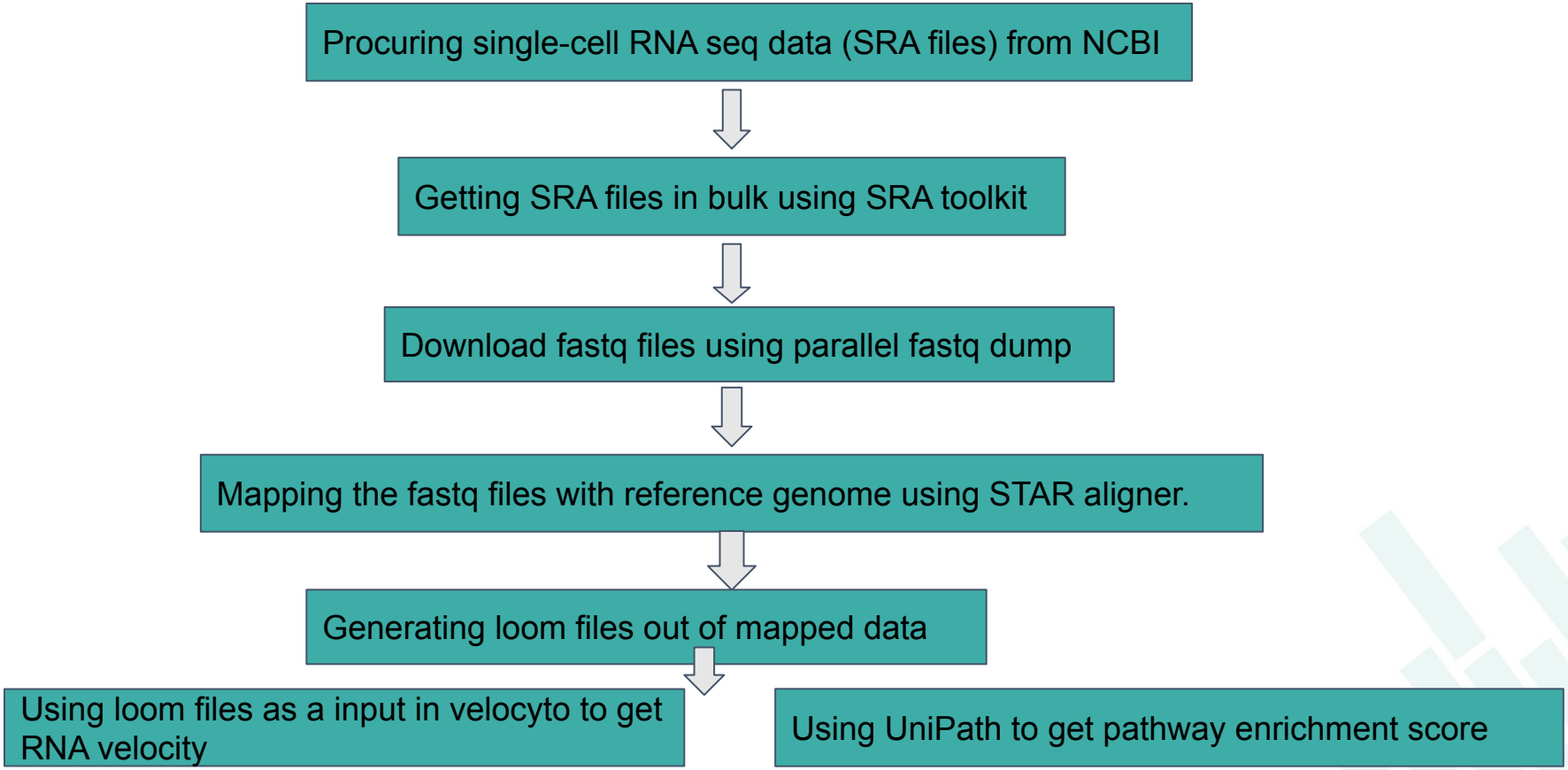
>Fastq-dump

>STAR aligner

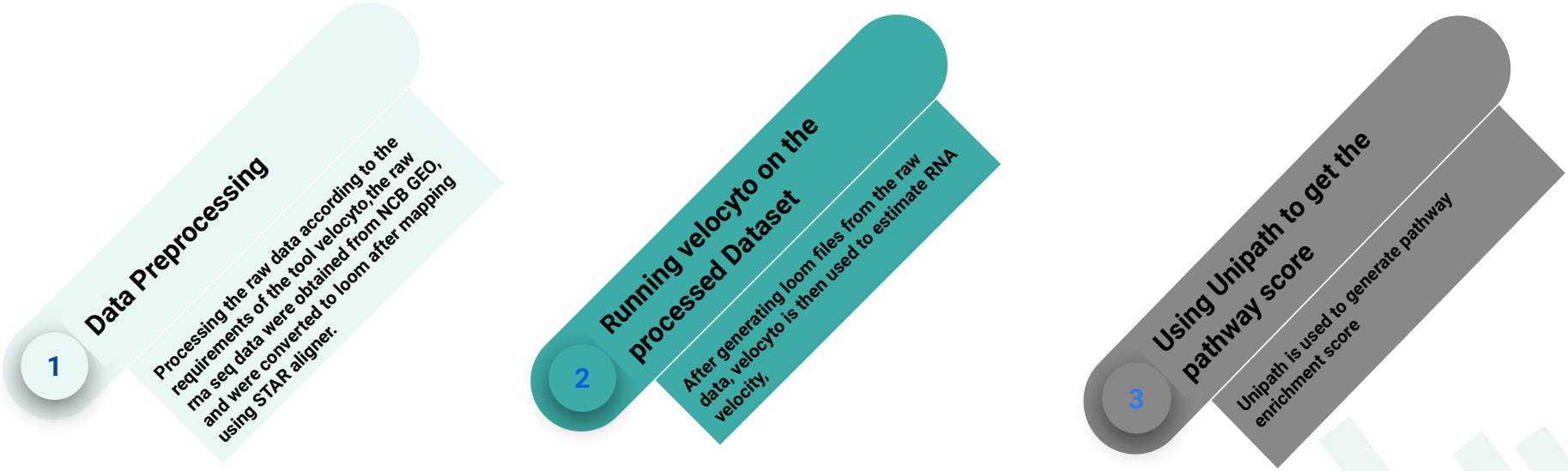
>UniPath

>Velocyto

Workflow

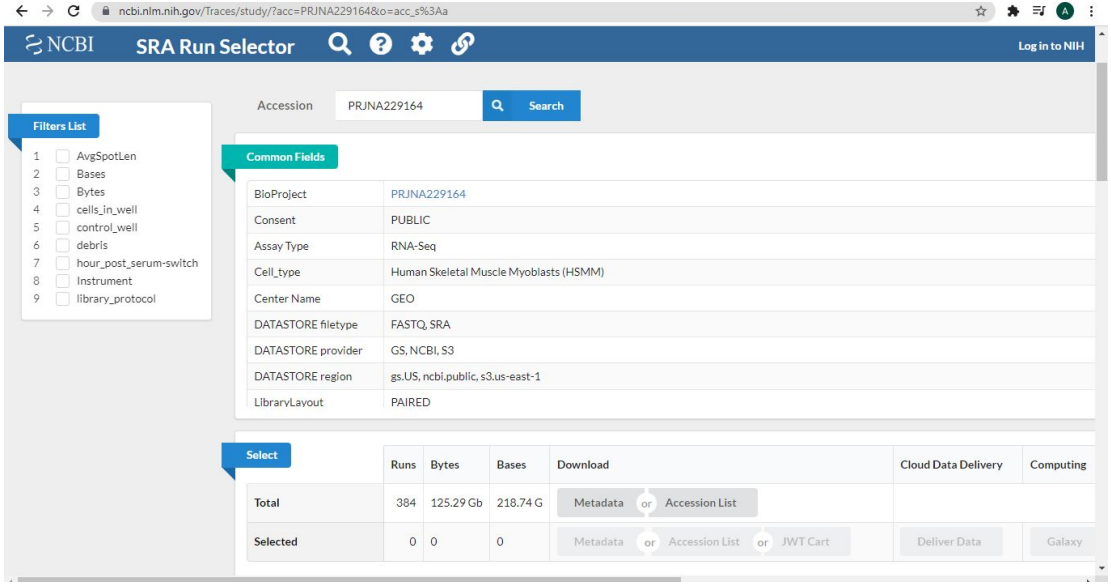

Procuring single-cell RNA seq data (SRA files) from NCBI
Getting SRA files in bulk using SRA toolkit
Download fastq files using parallel fastq dump
Mapping the fastq files with reference genome using STAR aligner.
Generating loom files out of mapped data
Using loom files as a input in velocyto to get RNA velocity
Using UniPath to get pathway enrichment score

Methodology



Dataset Description

- Here we have used myoblast differentiation data
- The raw single cell RNA-seq reads ,Sequence Read Archive(SRA) were downloaded from Genome expression omnibus (GSE52529).
- To Download SRA files in bulk SRA Tool Kit was used.



Dataset Description

For Batch Downloading of SRA files

esearch -db sra -query SRP033135 | efetch --format runinfo | cut -d ',' -f 1 | grep SRR|parallel -j 4 "prefetch {}"

For converting SRA to FASTQ Files

parallel-fastq-dump --sra-id SRR1033234 SRR1033235 SRR1033236 SRR1033237 SRR1033238 SRR1033239 SRR1033240 SRR1033241--threads 10 --outdir out/ --split-files --gzip

Dataset Pre-processing

- **Fastq files were then mapped with the reference genome hg19 using STAR aligner.**

```
# !/bin/bash


index=/storage/vibhor/Mtech/Ariba/genome_dir/genomehg19_index

FILES=/storage/vibhor/Mtech/Ariba/FastqFiles/*_1.fastq
OUTPUT=/storage/vibhor/Mtech/Ariba/BamFiles

for f in $FILES
do

 #  echo = $f
   base=$(basename $f.)

#    echo = $base
    echo = $f ${f%_1.fastq}_2.fastq

   ./STAR --runThreadN 10 --genomeDir $index --readFilesIn  $f ${f%_1.fastq}_2.fastq\
        --outSAMtype BAM SortedByCoordinate \
         --quantMode GeneCounts --outFileNamePrefix $OUTPUT/$base
done
```

Dataset Pre-processing

- **The output Bam files then converted into Loom files**
- **Here hg19 annotations and hg19 repeat mask file has been used**

```
# !/bin/bash


GTF=home/ansari20336/anaconda3/bin/hg19.ncbiRefSeq.gtf

FILES=/storage/vibhor/Mtech/Ariba/BamFiles/*.bam
OUTPUT=/storage/vibhor/Mtech/Ariba/Loom


for f in $FILES
do
   echo = $f
  ./velocyto run -c -U -o $OUTPUT -m hg19_rmsk.gtf $f /home/ansari20336/anaconda3/bin/hg19.ncbiRefSeq.gtf
done


echo "done!"
```

Dataset Pre-processing

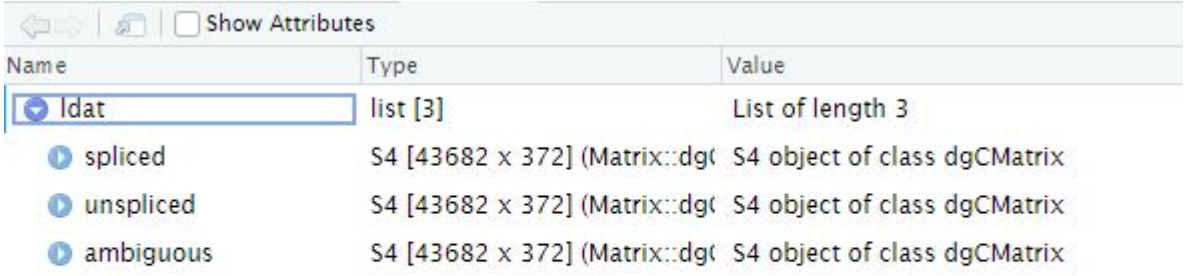
- The Output Loom files then merged into one Loom file using velocyto

Out[5]: 43682 rows, 372 columns, 4 layers
(showing up to 10x10)
merge.loom
name: *20211206T064750.782616Z*
name: *3.0.0*
name: *0.17.17*
name: *Default*

						CellID	SRR1022054Aligned_5EFVZ:SRR1022054Aligned.sortedByCoord.out.bam	SRR1033003_1
Accession	**Chromosome**	**End**	**Gene**	**Start**	**Strand**			
WASH7P	1	29370	WASH7P	14362	-		1.0	
MIR6859-1	1	17436	MIR6859-1	17369	-		0.0	
FAM138A	1	36081	FAM138A	34611	-		0.0	
SEPTIN14P18	1	129225	SEPTIN14P18	126642	-		0.0	
LOC729737	1	140566	LOC729737	134773	-		0.0	
RNU6-1100P	1	157887	RNU6-1100P	157784	-		0.0	
RPL23AP21	1	228787	RPL23AP21	228262	-		0.0	
CICP7	1	332282	CICP7	328518	-		0.0	
WBP1LP7	1	379573	WBP1LP7	379067	-		0.0	
LOC101928626	1	564389	LOC101928626	562760	-		0.0	
...	...	...	...	...	...	...	...	...

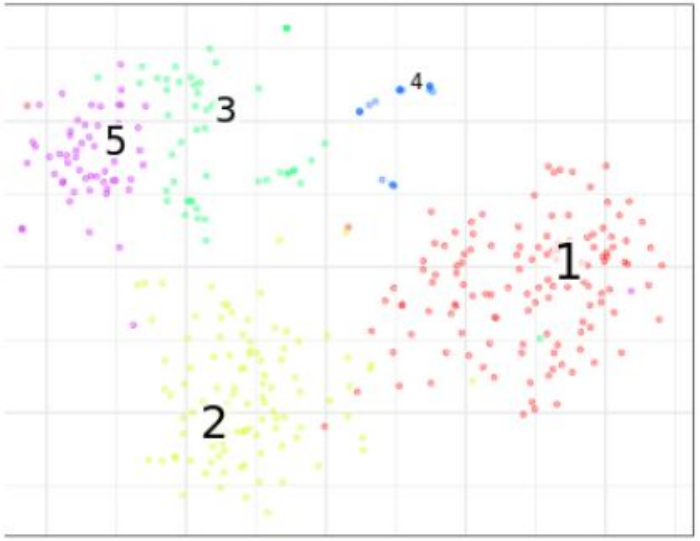
Data Analysis

- The merge Loom file is further analyzed using velocyto tool
- We got the Spliced and Unspliced RNA in the loom file

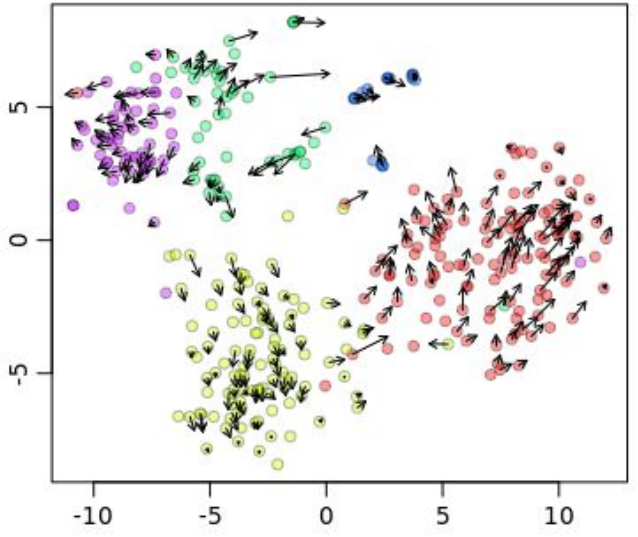


Data Analysis

- Using Velocyto.R we have got the following clusters

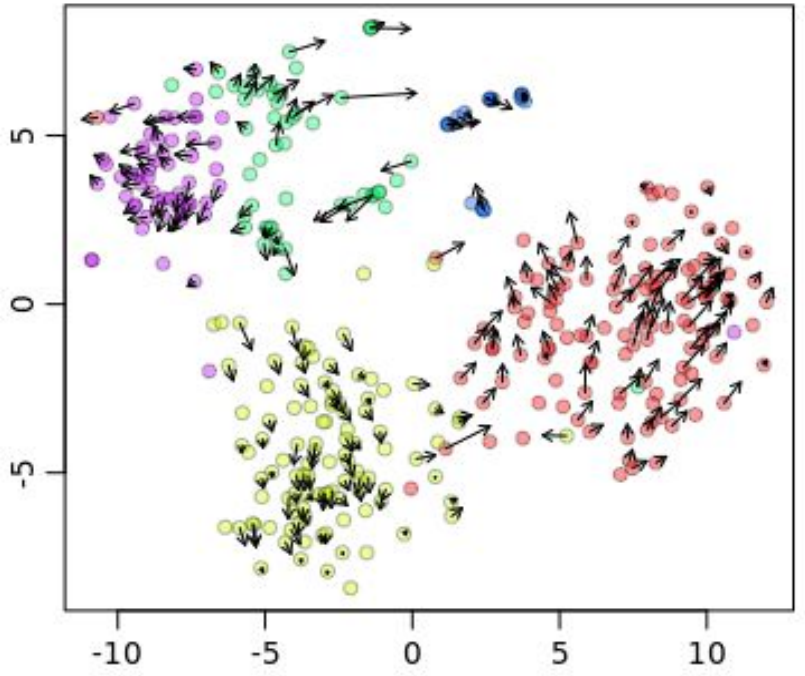


1- actively proliferating cells such as CDK1
2-differentiating myoblast
3- contaminating interstitial mesenchymal cells
4- an unknown cluster
5- muscle differentiation such as MYOG



Data Analysis

- Final velocity plot
- The arrows are in little unsynchronized manner as it is an early differentiation stage



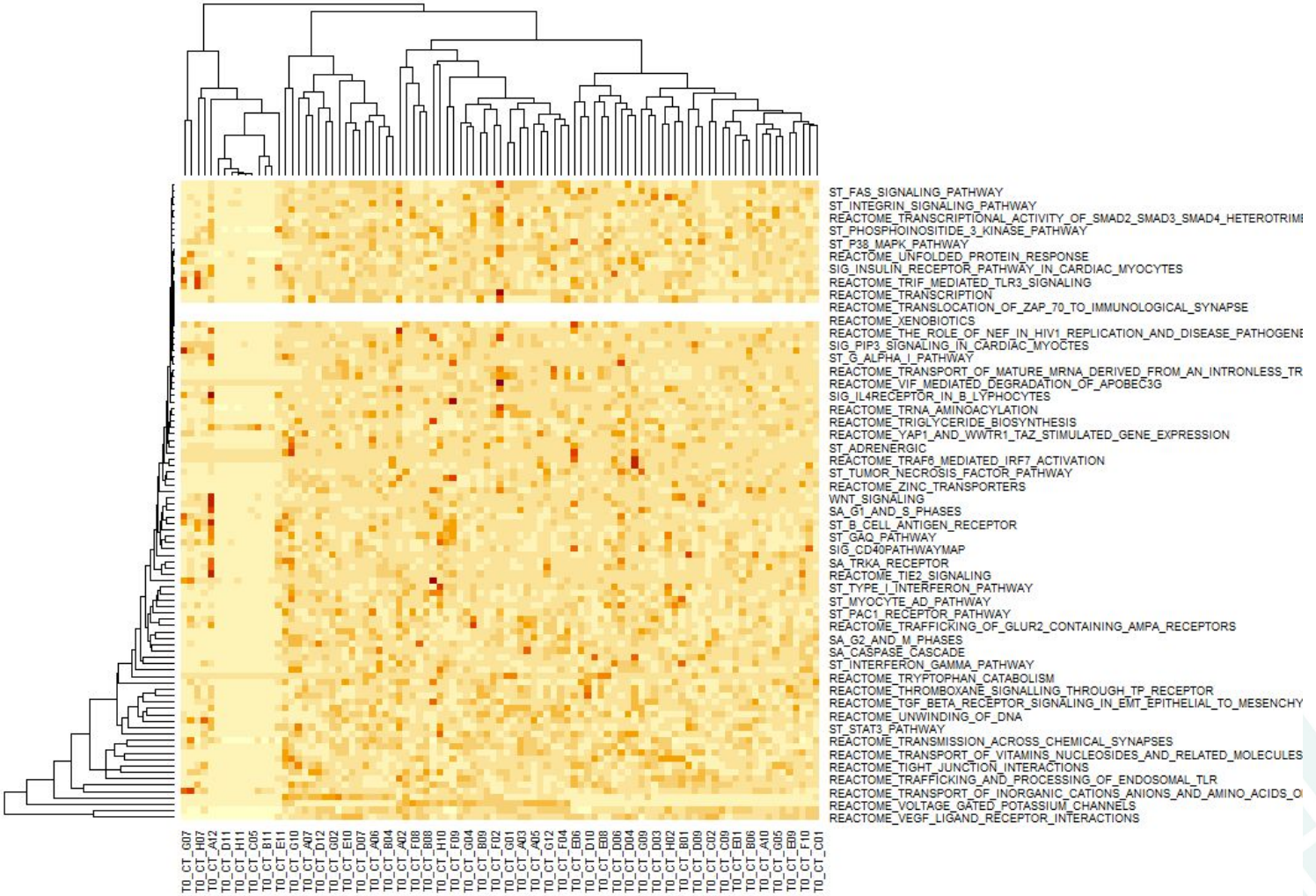
Data Analysis

- Now we have analysed the myoblast differentiation raw data count using UniPath
- Using UniPath we have got the pathway score for each stage of cell differentiation
- Here we have 4 stages-
- T0, T24, T48, T72
- We have plotted the top 50 pathway scores of each stage on a heat map

T0

Top 50

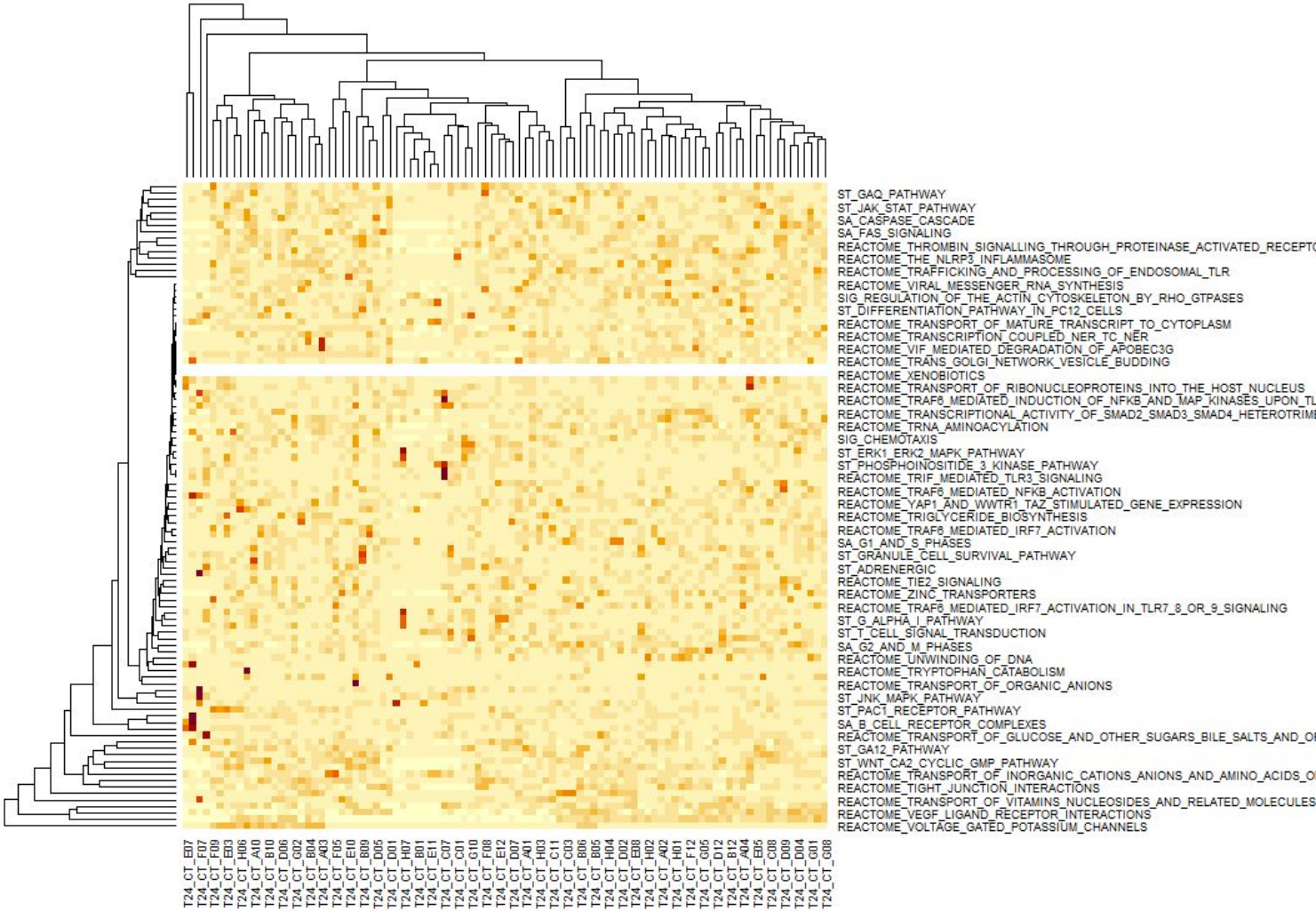
Pathways



T24

Top 50

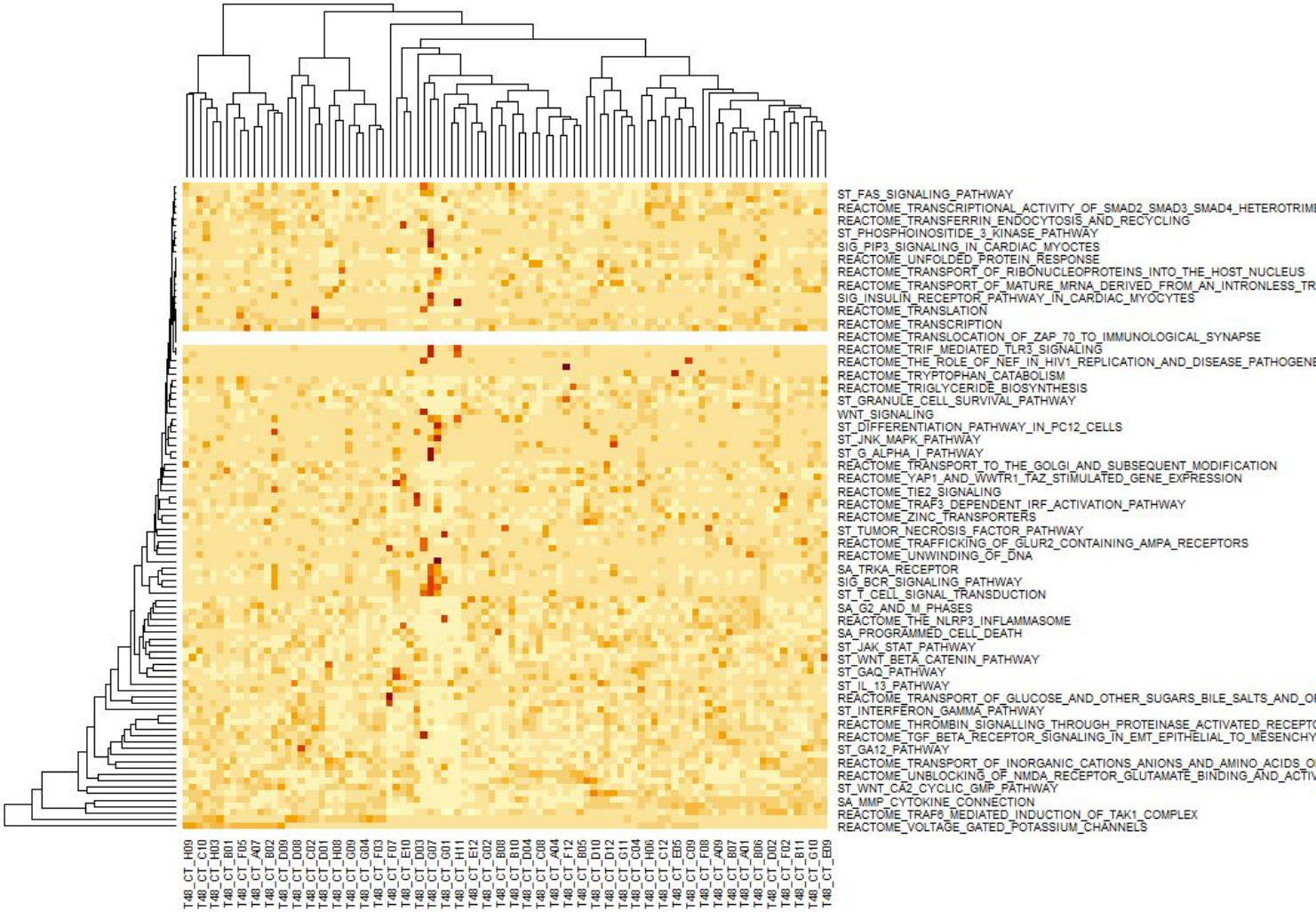
Pathways



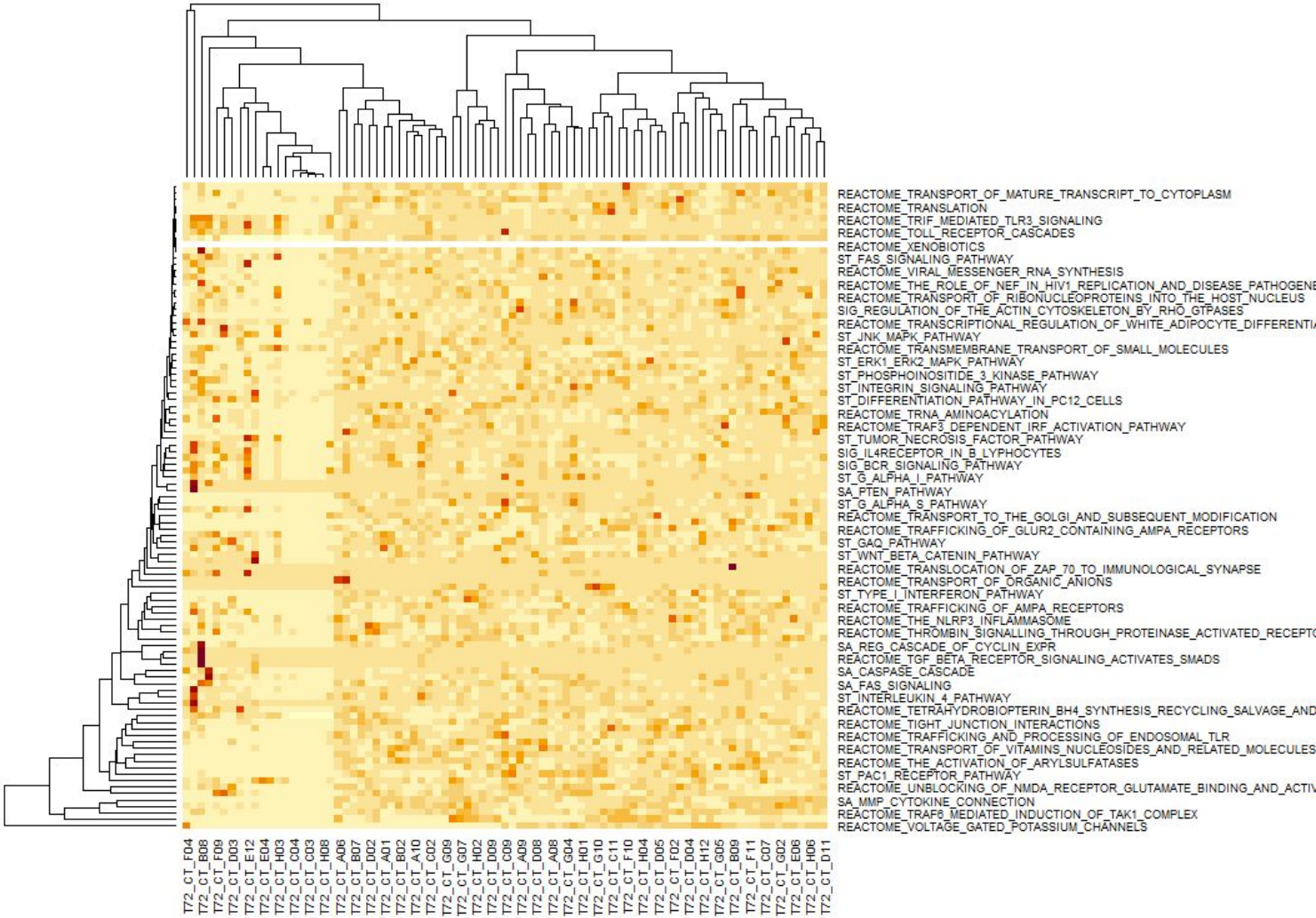
T48

Top 50

Pathways



T72 Top 50 Pathways



- Here we can see some important pathways which helps in myoblast differentiation
- Fas signaling pathway which is a programmed cell death pathway has been seen in all the 4 stages T0,T24,T48 and T72
- T0 where the cell are at active proliferation stage we can observe the active pathways such as -
- Integrin signaling pathway,Rho GTPase,p38 MAPK pathway where all these pathways are related to cell cell communication
- T24 where cell are differentiating myoblast the active pathways are
- Gaq pathway,JAK/STAT pathway,NLP3 Inflammasome which leads to cell growth
- T48 and T72 stage will lead to the formation of muscle cells the active pathways are
- Activity of SMAD,JNK pathway,TLR3 signaling,Transport of mature transcript to cytoplasm
- These pathways are leading to transcription and cell differentiation

Conclusion

Hence by observing the pathway activities and cells fate by RNA velocity we can conclude that all the pathways activity are related to cell differentiation which is cells fate.

And by knowing this pathway activity we can predict any disease state of the cell if there will be any unknown pathway active during differentiation of cells.

Also we have used myoblast cells which leads to the formation of muscle cells we can predict the diabetic cells by keeping a track of insulin pathway

Bibliography

https://github.com/alexdobin/STAR
https://gist.github.com/ipurusho/f6a6e53e0aa798c44e09c87bdc8b74fd
https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html
https://sydney-informatics-hub.github.io/training-RNAseq/02-BuildAGenomeIndex/index.html
https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

THANK YOU!