Genome Studies of Gene Expression and alternative splicing during iPSC Skeletal Muscle Induction and Differentiation

by

Yibo Wu

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Bioinformatics and Computational Biology

May 2019

APPROVED:

Zheyang Wu

Dr. Zheyang Wu, Advisor

DKorkin

Dr. Dimitry Korkin, Head of Department

Abstract

Facioscapulohumeral muscular dystrophy(FSHD) is a disorder characterized by muscle weakness and wasting (atrophy). This disease is typically inherited as autosomal dominant and has a complex genetic and epigenetic etiology. Our collaborator had differentiated healthy human pluripotent stem cells(iPSC) into skeletal muscles and exploited ISO-Seq to explore cell gene expression and transcript alternative splicing usage profile during 8 differentiation stages. Later, stage specific gene differential expression, transcript alternative splicing, gene ontology and novel gene/transcript were analysed to characterize the feature of each stage during the differentiation. In terms of expressed genes with more than or equal to 5 transcripts, each stage had shown their own stage specific features. About transcripts, iPS, S1, ADM.D0, ADM.D4 have about 30% to 40% more total transcripts than the rest 4 stages. 4 kinds of alternative splicing events are generally distributed and S2 stage has the least alternative splicing events potentially due to technical reasons. As for gene differential expressions, ADM.D4 has considerable amount of differential expressed genes with 5 other stages and it has minor difference with ISM.D4 and S3 stages(they are all myotubes cells). The gene ontology analysis is performed according to the results of previous step, stage specific GO terms are revealed.

Acknowledgements

Here I would like to thank everyone that has ever helped me in the past two years, especially my instructors and classmates. During the time working on this project my instructor, Dr. Zheyang Wu provided lots of valuable suggestions on how to do a scientific research and gave me many chances to learn practical techniques, I would like to express my appreciation to him. My team members Siqin Li and Ruosi Zhang has share much useful advice with me, in the mean time we became good friends, the friendship will sure continue.

This project is a collaboration work between WPI and Wellstone Muscular Dystrophy Program at University of Massachusetts Medical School. I would like to thank Dr. Charlies Emerson for providing such a nice platform to me to learn and grow, and also thank Dr. Oliver King's advice on Iso-Seq analyzing strategies, Dr. Dongsheng Guo's sharing on FSHD and iPSCs .

Finally, I would like to give my gratitude to all faculty of WPI BCB program. Thank everyone for the days we spent together in WPI and thank everyone for making BCB a better program.

Contents

List of Figures

List of Tables

1 Introduction

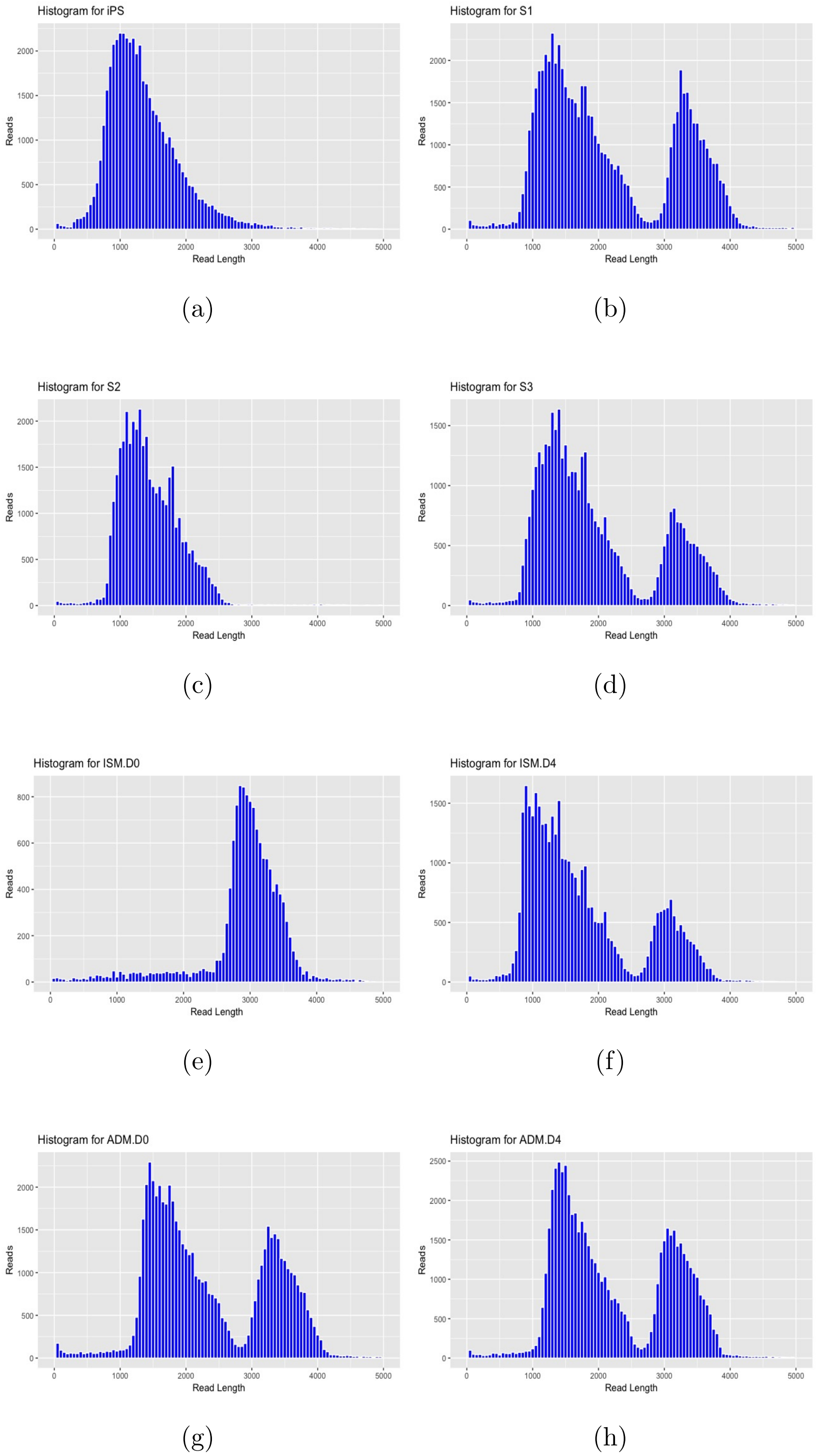
1.1 Induced pluripotent Stem Cells

Under certain conditions, adult cells can be genetically reprogrammed into induced pluripotent stem cells(iPSCs). Due to their remarkably similarity to embryonic stem cells in many key aspects, iPSCs have the potential to become effective tools to understand and model diseases and deliver cell-replacement therapy to support regenerative medicine [1].

Facioscapulohumeral muscular dystrophy is typically inherited as autosomal dominant and has a complex genetic and epigenetic etiology [2], characterized by muscle weakness and wasting [3]. In our study, human iPS cells were differentiates into muscle cells, 8 samples from 3 cell lines are collected. In the first cell line, iPS cells were obtained at day 0. After 5 days in S1 medium, iPS cells were developped into skeletal muscle progenitor cells(stage S1), then 4 days in S2 medium, skeletal muscle myoblasts(S2 stage) were induced. At day 7, S2 stage cells differentiated as myotubes(S3 stage). Additionally, iPS derived secondary myoblasts(ISM.D0) were differentiated into myotubes(ISM.D4) 4 days later. There is also another parental adult line, adult myoblasts(ADM.D0) differentiated into adult myotubes(ADM.D4) at day 4. Generally speaking we have 3 cell lines: iPS to S1 to S2 to S3; ISM.D0 to ISM.D4 and ADM.D0 to ADM.D4

1.2 Iso-Seq analysis of RNA expression

The Pacific Biosciences(PacBio) transcript Sequencing(ISO-Seq) method employs long read to sequence transcript transcripts from the 5‘ end to their poly-A tails [4]. This new technique can reduce the effort and error during reconstructing and inferencing short reads. In preivious researches, ISO-Seq has been used to analyze full-length splice transcripts in human organs and embryonic stem cells, indicating that even in highly characterized transcriptome like human, the identification of genes and splice transcripts is far from complete [5]. Here, ISO-Seq was performed to characterize the stage specific RNA expression profile. The sequencing work was conducted by Umass Medical Sequencing Core. They had also classified and clustered the circular consensus(CCS) reads following the PacBio ISO-Seq analysis application work flow [6] and produced high quality transcript sequence files. Except for iPS stage, the sequencing data of each stage consists of 2 parts: 1-3kb, over 3kb. The over 3kb part data of S2 stage was dropped because of containing lots of mitochondrial sequences and caused much trouble during the ISO-Seq analyze step(Figure 1). Sequence data of iPS stage is mainly distributed in 1-3kb part, this could be resulted by Iso-Seq technique or sample preparation reasons. During our analysis, the 1-3kb and over 3kb data were first combined into one file, then I aligned the transcripts to reference genome using GMAP/GSNAP(Genomic Mapping and Alignment Program for mRNA and EST Sequences, and Genomic Short-read Nucleotide Alignment Program) [7]. According to the sam file alignment, transcript sequences were collapsed into final set of unique, full-length, high-quality transcripts following ToFU(transcript transcripts: Full-length and Unassembled) [8] pipeline. After that, basic statistical summary of the data was made by Python, alternative splicing events distribution was counted by SpliceGrapher [9]. Gene differential expression analysis performed by R package edgeR [10] revealed some relations between stages, gene ontology analysis of differential expressed genes were performed by R package clusterProfiler [11]. Finally novel genes and novel transcripts were obtained by comparing with reference genome using IGV [12].



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 1: Read length Distribution.

2 ISO-Seq analysis of gene expression during iPSC induction and differentiation

2.1 Pipeline

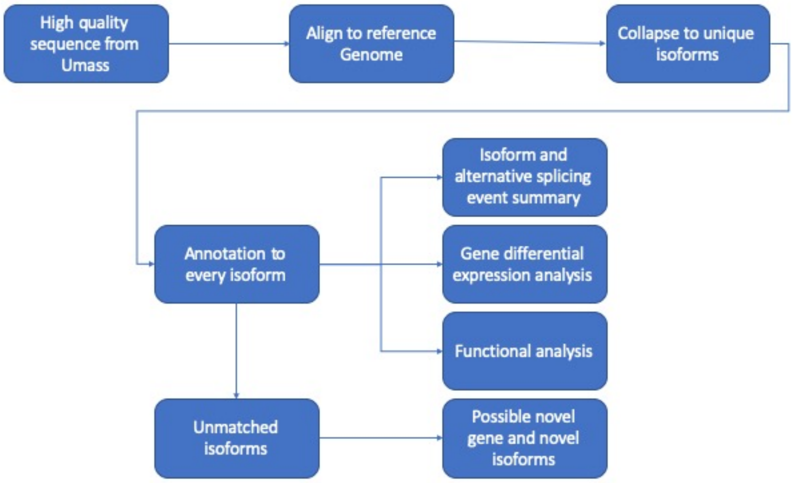

High quality sequence from Umass
Align to reference Genome
Collapse to unique isoforms
Isoform and alternative splicing event summary
Annotation to every isoform
Gene differential expression analysis
Functional analysis
Unmatched isoforms
Possible novel gene and novel isoforms

Figure 2: Pipeline of the analysis

The high quality transcript sequence data was obtained from Umass Medical Sequencing Core, I aligned the fasta file to reference genome using GMAP(parameter setting: $-f\ samse\ \ -n\ 0\ \ -t\ 16\ \ --cross-species\ \ --max-intronlength-ends\ 200000\ \ -z\ sense\_force$). Cupcake Tofu collapsed all redundant reads into unique transcripts and annotated all transcripts according to reference genome. Unmatched isforms are considered as possible novel transcripts, matched transcripts are used to perform gene expression analysis and functional analysis.

Our data was provided by researchers from Umass medical. They classified and clustered the raw data from Iso-Seq platform and generated high quality transcript sequence data. In the high quality data, each sequence is assumed to be full-length, supported by 2 or more full length reads and have a predicted accuracy over 99% by default. Because of the natural 5' degradation in rranscripts and clustering algorithm trade off between sensitivity and specificity, it is possible that some identical or redundant transcripts still exist [8]. I used Cupcake ToFU to collapse redundant transcripts to obtain unique transcripts. With errors and redundant eliminated, read numbers in our data also dropped(Table1).

Then every unique transcript was annotated using Genecode v19 human Genome data. Annotated transcripts are used to explore differential expressed genes between stages using R package edgeR. Functional analysis conducted by R package clusterProfiler also revealed the functions and pathways related to these differential expressed genes. transcripts that can not be annotated are considered as possible novel gene or transcripts, part of them were validated by IGV visualization.

step	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0	ADM.D4
high quality data	28282	34282	17654	19942	17174	17847	42037	31022
unique transcripts	23199	27497	14580	16564	14360	14895	27528	25422
matched transcripts	22450	26866	14292	16387	14197	14670	26677	24667
unmatched transcripts	749	631	288	177	163	225	851	755

Table 1: Read numbers during each step

The high quality data still contains some redundant transcript. After collapsing, redundant is eliminated, unique transcripts are obtained, which means every read is a unique transcript.

2.2 Gene annotation and transcript analysis

High quality transcripts were aligned to Gencode v19 human genome using GMAP [7]. Since Clustering algorithm would balance its sensitivity and specificity, it is possible that some high quality sequences represent identical or redundant transcripts, Cupcake ToFU pipeline [8] was used to collapse identical transcripts and obtain final set of unique, full-length, high quality transcripts, after this step multiple reads could be collapsed into one transcript. The annotation files were produced by Cupcake Annotation comparing against Genecode v19 gene model.

2.2.1 Transcript counts distribution

R was used for the statistical analysis and plotting.From the bar plot, ISM.D0 stages has the lowest number of transcripts 14197, which is similar to ISM.D4, S2 and S3 stage, but much lower than that of other stages. S1 stages has the highest number of transcripts, 26866. There is no significant a distinct gap between number of transcripts during ISM differentiation, and adult myoblast differentiation. Although over 3kb part is missing in S2 stage, significant decreasing can be observed during mononucleated myocytes differentiation(S1 to S3).

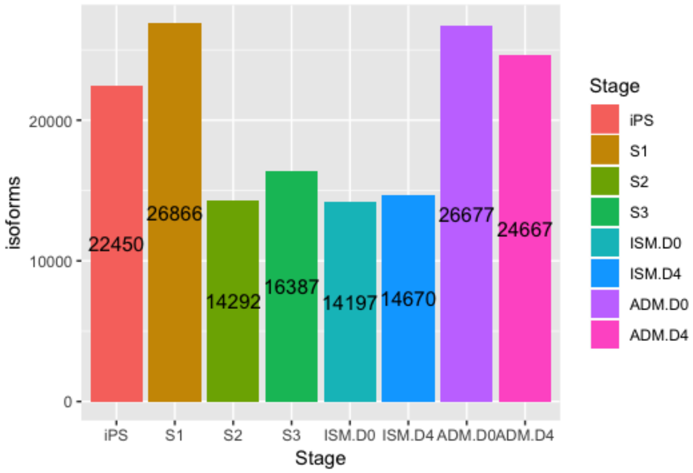


Figure 3: transcript numbers distribution

2.2.2 Alternative splicing distribution

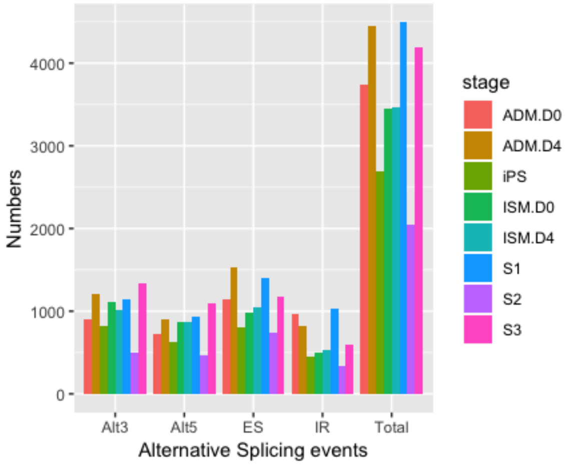


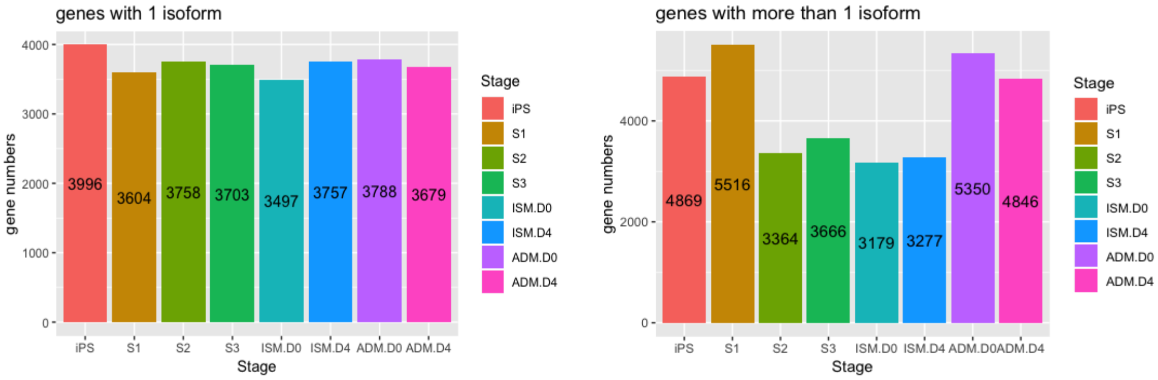
Figure 4: Alternative Splicing events distribution

4 kinds of alternative splicing events are counted in our study: Alt3, 3‘ alternative splicing; Alt5, 5‘ alternative splicing; ES, exon skipping; IR, intron retention

With alternative splicing(AS), a gene can be transcripted into different transcripts, with SpliceGrapher, 4 types of AS events are counted in our data: Alt5: 5′ alternative splicing; Alt3: 3′ alternative splicing; ES: exon skipping; IR: intron retention. From the figure above, the distribution of AS events are pretty average across different stages(S2 stage contains less AS events might be resulted from its missing data). 2 ISM stages are extremely similar to each other. For each stages, their genes with most AS events are also genes with most transcripts, detail information could be found in table 4 to table 11. However, the alternative splicing event numbers are significantly smaller than transcript numbers(Figure 1), since I didn't compare our data against alternative splicing reference data, much alternative splicing information can not be obtained by comparing against reference genome, the result is reasonable.

2.2.3 Gene counts distribution

Since Cupcake Annotation already mapped every transcript to a reference transcript and reference gene, the gene counts distribution can be obtained directly(Figure 5). Comparing to total transcript counts distribution(Figure 3), over 80% of transcripts are generated by about 50% genes, these genes are potentially to play an important role in stage-wise biological functions. S1 and ADM.D0 stage has the largest expressed gene numbers. Significant differences can be observed between secondary myoblasts differentiation and adult myoblasts differentiation processes(ISM.D0 to ISM.D4 and ADM.D0 to ADM.D4), while both secondary myoblasts differentiation and adult myoblasts differentiation didn't show much gene number differences within the process. In total, 1966 genes are expressed in all 8 stages.



(a)

(b)

Figure 5: Number of genes distribution during different stages.

(a)genes with 1 transcript. (b) genes with more than 1 transcript. Most transcripts are expressed by a small number of genes.

After the collapse step in Cupcake ToFU, multiple reads could be collapsed into one transcript, here we didn't consider about the read number for every gene, simply regard every observed gene as expressed, and counted shared gene numbers between stages. From the 2 tables below(Table 2 and Table 3), the absolute number of share genes in all 8 stages data are quite similar with each other. The proportion of shared genes among every 2 stage group is about 50% (proportion of shared genes = Intersect(genes in stage A, genes in stage B)/Union(genes in stage A, genes in stage B)). Theoretically, the gene number and transcript numbers may be strongly influenced by the total number of reads sequenced even we have obtained the unique transcript data, which means the gene numbers expressed in samples may be differ due to experiment technical reasons .

But when I narrow the range of genes, things changed a little bit. When I only compare genes with over 1 transcript, the proportion of share genes dropped significantly(Table 3), when genes with over or equal to 5 transcripts were compared, the 8 stages showed their stage features(Table 4). The proportion of shared genes in genes with over of equal to 5 transcripts is relatively small, which suggest that most of the genes are related to stage wise functions.

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	6009						
S2	4834	5086					
S3	4845	5143	4633				
ISM.D0	4544	4820	4551	4504			
ISM.D4	4635	4823	4587	4819	4504		
ADM.D0	5232	5790	4687	4971	4549	4695	
ADM.D4	5031	5511	4548	4912	4415	4708	5897

Table 2: Intersection of expressed genes across different stages(absolute value)

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	54.41%						
S2	46.58%	51.15%					
S3	46.33%	50.92%	48.62%				
ISM.D0	44.86%	49.31%	50.53%	49.03%			
ISM.D4	44.56%	48.43%	49.13%	51.49%	50.23%		
ADM.D0	46.19%	51.18%	45.55%	49.66%	45.32%	48.09%	
ADM.D4	45.55%	52.03%	45.88%	48.25%	45.54%	46.43%	54.88%

Table 3: Intersection of expressed genes across different stages(proportion)

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	46.99%						
S2	34.58%	38.56%					
S3	33.40%	38.62%	35.90%				
ISM.D0	33.36%	36.74%	38.39%	37.64%			
ISM.D4	31.98%	35.21%	37.07%	41.72%	39.74%		
ADM.D0	37.11%	45.31%	33.28%	37.94%	34.80%	34.42%	
ADM.D4	35.50%	42.57%	33.39%	40.12%	34.15%	38.05%	49.66%

Table 4: Intersection of expressed genes(>1 transcript) across different stages(proportion)

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	32.36%						
S2	20.98%	19.80%					
S3	18.01%	21.02%	24.67%				
ISM.D0	19.75%	22.12%	30.92%	29.73%			
ISM.D4	17.78%	18.27%	26.47%	35.23%	30.56%		
ADM.D0	20.49%	28.76%	15.47%	22.37%	19.86%	17.62%	
ADM.D4	20.60%	26.19%	15.93%	25.60%	19.63%	23.04%	36.96%

Table 5: Intersection of expressed genes(>=5 transcripts) across different stages(proportion)

The transcript numbers could vary between different genes. In the same stage, some genes might contains hundreds of transcripts, while some only contain 1(Figure 5)Genes with extremely large transcript numbers must be closely related to stage-wise functions even though gene numbers could be influenced by other factors. Top 10 genes ranked by transcript numbers of each stage are listed below. In generally, some genes are highly expressed in many cell stages like collagen genes, heterogeneous nuclear ribonucleoprotein K gene, myosin genes, insulin-like growth factor gene, pyruvate kinase gene and titin genes, etc. Some genes are only observed highly expressed in certain stages, like TERF1(iPS), CSDE1(S1), TUBB(S2), SPARC(S3), ANXA2(ISM.D0), MEG3(ISM.D4), SULF1(ADM.D0), they obviously have distinct stage feature.

gene	transcripts	Description
TERF1	29	telomeric repeat binding factor 1
JARID2	27	jumonji, AT rich interactive domain 2
NAP1L1	25	nucleosome assembly protein 1-like 1
KPNB1	24	karyopherin (importin) beta 1
SNHG14	24	small nucleolar RNA host gene 14
PABPC1	23	poly(A) binding protein, cytoplasmic 1
HNRNPK	22	heterogeneous nuclear ribonucleoprotein K
HSP90AA1	21	heat shock protein 90kDa alpha (cytosolic), class A member 1
HNRNPC	20	heterogeneous nuclear ribonucleoprotein C (C1/C2)
BPTF	20	bromodomain PHD finger transcription factor

Table 6: List of top ranked genes in iPS stage

gene	transcripts	Description
COL11A1	40	collagen, type XI, alpha 1
COL1A2	36	collagen, type I, alpha 2
CSDE1	32	cold shock domain containing E1, RNA-binding
HMGA2	30	high mobility group AT-hook 2
HNRNPK	30	heterogeneous nuclear ribonucleoprotein K
KPNB1	29	karyopherin (importin) beta 1
MYH9	28	myosin, heavy chain 9, non-muscle
PABPC1	28	poly(A) binding protein, cytoplasmic 1
SEC31A	25	SEC31 homolog A (S. cerevisiae)
SEPT11	24	septin 11

Table 7: List of top ranked genes in S1 stage

gene	transcripts	Description
ACTG1	17	actin, gamma 1
MAP1B	16	microtubule-associated protein 1B
TUBB	14	tubulin, beta class I
HSP90AA1	14	heat shock protein 90kDa alpha (cytosolic), class A member 1
IGF2	14	insulin-like growth factor 2
HSPD1	14	heat shock 60kDa protein 1
HNRNPK	13	heterogeneous nuclear ribonucleoprotein K
PKM	13	pyruvate kinase, muscle
HNRNPA2B1	13	heterogeneous nuclear ribonucleoprotein A2/B1
MEG3	13	maternally expressed 3 (non-protein coding)

Table 8: List of top ranked genes in S2 stage

gene	transcripts	Description
COL3A1	70	collagen, type III, alpha 1
COL1A2	63	collagen, type I, alpha 2
COL1A1	43	collagen, type I, alpha 1
MYH3	33	myosin, heavy chain 3, skeletal muscle, embryonic
H19	32	H19, imprinted maternally expressed transcript (non-protein coding)
COL5A2	30	collagen, type V, alpha 2
SPARC	28	secreted protein, acidic, cysteine-rich (osteonectin)
COL4A2	28	collagen, type VI, alpha 2
COL4A1	27	collagen, type VI, alpha 1
IGF2	26	insulin-like growth factor 2

Table 9: List of top ranked genes in S3 stage

gene	transcripts	Description
ANXA2	58	annexin A2
COL1A1	53	collagen, type I, alpha 1
H19	48	H19, imprinted maternally expressed transcript (non-protein coding)
ACTG1	46	actin, gamma 1
PKM	46	pyruvate kinase, muscle
COL3A1	45	collagen, type III, alpha 1
IGF2	44	insulin-like growth factor 2 (somatomedin A)
TPM1	44	tropomyosin 1 (alpha)
HNRNPK	42	heterogeneous nuclear ribonucleoprotein K
ITGB1	42	integrin, beta 1

Table 10: List of top ranked genes in ISM.D0 stage

gene	transcripts	Description
TTN	38	titin
COL1A1	35	collagen, type I, alpha 1
MYH3	33	myosin, heavy chain 3, skeletal muscle, embryonic
COL3A1	32	collagen, type III, alpha 1
IGF2	28	insulin-like growth factor 2 (somatomedin A)
COL1A2	26	collagen, type I, alpha 2
MEG3	26	maternally expressed 3 (non-protein coding)
COL4A1	22	collagen, type IV, alpha 1
FN1	21	fibronectin 1
PALLD	20	palladin, cytoskeletal associated protein

Table 11: List of top ranked genes in ISM.D4 stage

gene	transcripts	Description
TTN	294	titin
FN1	262	fibronectin 1
COL1A2	131	collagen, type I, alpha 2
PALLD	95	palladin, cytoskeletal associated protein
DST	92	dystonin
COL1A1	85	collagen, type I, alpha 1
NEB	80	nebulin
SULF1	76	sulfatase 1
COL3A1	75	collagen, type III, alpha 1
MEF2C	75	myocyte enhancer factor 2C

Table 12: List of top ranked genes in ADM.D0 stage

gene	transcripts	Description
TTN	259	titin
NEB	74	nebulin
COL1A2	69	collagen, type I, alpha 2
FN1	67	fibronectin 1
MEF2C	61	myocyte enhancer factor 2C
DST	56	dystonin
PALLD	54	palladin, cytoskeletal associated protein
MYH3	52	myosin, heavy chain 3, skeletal muscle, embryonic
MYH8	44	myosin, heavy chain 8, skeletal muscle, perinatal
COL1A1	43	collagen, type I, alpha 1

Table 13: List of top ranked genes in ADM.D4 stage

2.3 Differential expression analysis

In order to explore the insights of relations between different stages, R package EdgeR was used to comparing the expressing level of each gene in different stages.Here, 1 times fold change is used to characterize the expression level change, False Discovery Rate(FDR) is used to indicate the reliability of the results. By program default, $|\log_2 fold| > 1$ and FDR < 0.05 was set as threshold. The $|\log_2 fold|$ represent the gene differential expressed level between 2 stages, a positive or negative number means up regulating or down regulating. The default settings of the program would produce reliable differential expressed genes between stages [10]. Here I have listed all of the differential expressed genes when other stages are compared to one certain stage.

2.3.1 iPS

stage	genes
S1	/
S2	/
S3	COL3A1, COL1A2, COL1A1, MYH3, H19, COL5A2, IGF2, POSTN, COL5A1
ISM.D0	H19
ISM.D4	MYH3, COL3A1, COL1A1, IGF2, TTN, JARID2, PALLD, TNNT2
ADM.D0	FN1, SULF1, RUNX1, COL3A1, PALLD, COL1A2, ADAMTSL1, COL1A1, COL6A3, ZEB1, ELN, ATP2B1,,ESRG, FBN1, ADAM9, CBS
ADM.D4	TTN, NEB, MEF2C, MYH3, PALLD, MYH8, COL3A1, POSTN, FN1, COL1A2, ACTN2, DMD, RUNX1, NCAM1, COL1A1, SGCD, COL5A2, FBN1, SULF1, DST, ESRG, COL6A2, ARPP21, ZEB1

Table 14: differential expressed genes for iPS stage over other stages

2.3.2 S1

stage	genes
iPS	/
S2	/
S3	MYH3, H19, TTN, COL3A1, POSTN, IGF2
ISM.D0	H19
ISM.D4	TTN, MYH3, IGF2
ADM.D0	FN1, TTN, SULF1, RUNX1, ELN
ADM.D4	TTN, NEB, MYH3, MYH8, NCAM1, ACTN2, POSTN, SGCD, DLG2, MEF2C, F13A1, LDB3, DCLK1, BIN1, LMO7, SULF1, RUNX1, LIN28A, CADM2, COL6A2, ARPP21, MYH10, SMC4, ITGA7, TNNT2, DST, FN1

Table 15: differential expressed genes for S1 stage over other stages

2.3.3 S2

stage	genes
iPS	/
S1	/
S3	COL3A1, MYH3, COL1A1, TTN
ISM.D0	/
ISM.D4	TTN, MYH3
ADM.D0	FN1
ADM.D4	TTN, NEB, MEF2C, MYH8, POSTN, MYH3, NCAM1, PALLD, DCLK1, ACTN2, FN1, SGCD, DLG2, SULF1, LIMCH1,F13A1, LDB3

Table 16: differential expressed genes for S2 stage over other stages

2.3.4 S3

stage	genes
iPS	COL3A1, COL1A2, COL1A1, MYH3, H19, COL5A2, IGF2, POSTN, COL5A1
S1	MYH3, H19, TTN, COL3A1, POSTN, IGF2
S2	COL3A1, MYH3, COL1A1, TTN
ISM.D0	MYH3
ISM.D4	/
ADM.D0	MYH3, POSTN
ADM.D4	TTN

Table 17: differential expressed genes for S3 stage over other stages

2.3.5 ISM.D0

stage	genes
iPS	H19
S1	H19
S2	/
S3	MYH3
ISM.D4	TTN, MYH3
ADM.D0	/
ADM.D4	TTN, NEB, MEF2C, MYH3, NCAM1, MYH8, DCLK1, ACTN2, POSTN, DMD, DLG2

Table 18: differential expressed genes for ISM.D0 stage over other stages

2.3.6 ISM.D4

stage	genes
iPS	MYH3, COL3A1, COL1A1, IGF2, TTN, JARID2, PALLD, TNNT2
S1	TTN, MYH3, IGF2
S2	TTN, MYH3
S3	/
ISM.D0	TTN, MYH3
ADM.D0	MYH3
ADM.D4	/

Table 19: differential expressed genes for ISM.D4 stage over other stages

2.3.7 ADM.D0

stage	genes
iPS	FN1, SULF1, RUNX1, COL3A1, PALLD, COL1A2, ADAMTSL1, COL1A1, COL6A3, ZEB1, ELN, ATP2B1,,ESRG, FBN1, ADAM9, CBS
S1	FN1, TTN, SULF1, RUNX1, ELN
S2	FN1
S3	MYH3, POSTN
ISM.D0	/
ISM.D4	MYH3
ADM.D4	TTN, MYH3, POSTN, NEB, MYH8

Table 20: differential expressed genes for ADM.D0 stage over other stages

2.3.8 ADM.D4

stage	genes
iPS	TTN, NEB, MEF2C, MYH3, PALLD, MYH8, COL3A1, POSTN, FN1, COL1A2, ACTN2, DMD, RUNX1, NCAM1, COL1A1, SGCD, COL5A2, FBN1, SULF1, DST, ESRG, COL6A2, ARPP21, ZEB1
S1	TTN, NEB, MYH3, MYH8, NCAM1, ACTN2, POSTN, SGCD, DLG2, MEF2C, F13A1, LDB3, DCLK1, BIN1, LMO7, SULF1, RUNX1, LIN28A, CADM2, COL6A2, ARPP21, MYH10, SMC4, ITGA7, TNNT2, DST, FN1
S2	TTN, NEB, MEF2C, MYH8, POSTN, MYH3, NCAM1, PALLD, DCLK1, ACTN2, FN1, SGCD, DLG2, SULF1, LIMCH1,F13A1, LDB3
S3	TTN
ISM.D0	TTN, NEB, MEF2C, MYH3, NCAM1, MYH8, DCLK1, ACTN2, POSTN, DMD, DLG2
ISM.D4	/
ADM.D0	TTN, MYH3, POSTN, NEB, MYH8

Table 21: differential expressed genes for ADM.D4 stage over other stages

From the results, several interesting place can be noticed. iPS, S1, S2 and ISM.D0 have little differential expressed genes. Since ISM.D0 and S2 are all myoblasts, S2 is developed from S1 and iPS, the result might indicate that HG19 gene distinguishes ISM cell line from iPS cell line; in iPS and S stages, cells are very similar to ISM cell line.

iPS cells have significantly more differential expressed genes with later stage cells then with early stage cells. During the process of differentiation, genes with stage specific functions will be turned on, thus the differential expressed genes between iPS stages are potentially connected to certain differentiation function.

S3, ISM.D4 and ADM.D4 stage have little differential expressed genes. Being myotubes could explain their little differences. Also S2, ISM.D0, ADM.D0 have little differential expressed genes could be explained by the same reason. Generally, a huge difference can be observed between ADM cell line and other cell lines, these genes are very likely to be related with adult cell differentiation.

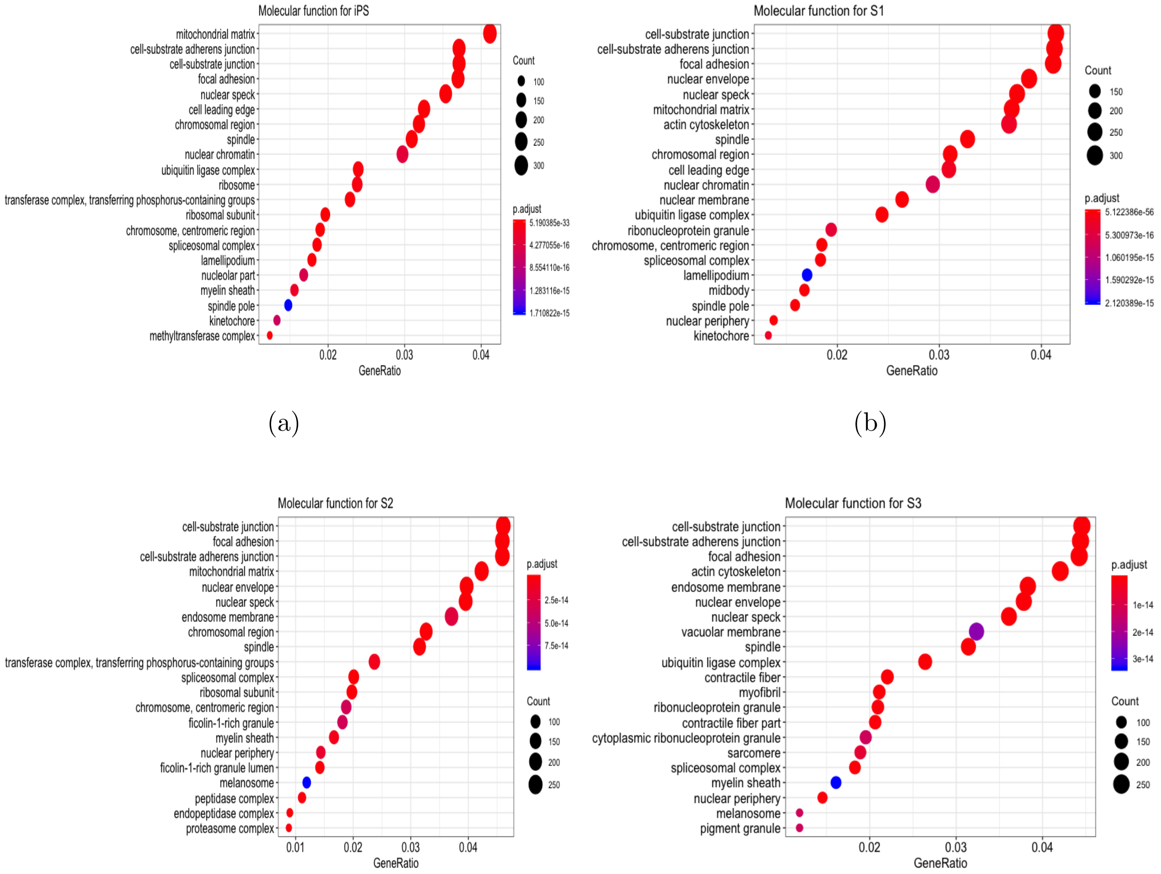
The number of differential expressed genes is much smaller than total gene numbers in our data, by exploring into the results, I noticed that, most differential expressed genes are eliminated because of high FDR value. Since the 8 datasets in our data are all unique ones, lacking of replica might be the reason that the result has a high FDR value. In further analysis we plan to change the threshold of FDR value so that we could obtain more gene symbols.

As a conclusion, the gene expression differences in cells from similar tissues(both myoblasts or myotubes) are less then cells from same cell lines. For example the differential expressed genes between ADM.D4 vs ISM.D4 group is more then ADM.D4 vs ADM.D0 group or ISM.D4 vs ISM.D0 group. As the start point of the differentiation process, iPS stage mostly express basic household genes while later stages would have more stage specific genes expressed. But further analysis also should be performed, a small number of genes might not tell much about stage wise functions.

2.4 Functional analysis

R package clusterProfiler was used to perform the Gene Ontology(GO) and Kyoto Encyclopedia of Genes and Genomes(KEGG) pathway analysis. Information for Molecular Function(the specific activity that gene products play a role in), Cellular Component(the specific place in a cell where a gene product is located), Biological Process(biological activity which a group of genes or gene products participate in) is obtained using default parameter. Detailed GO information is available in supplementary information. Here terms with high gene counts are presented in dot charts. Top 21 process ranked by gene ratio(number of target gene in term/number of target gene) are shown in the dot chart. the p value indicated the significance of enrichment analysis, usually the enrichment is significant when $p < 0.05$.

2.4.1 Molecular function



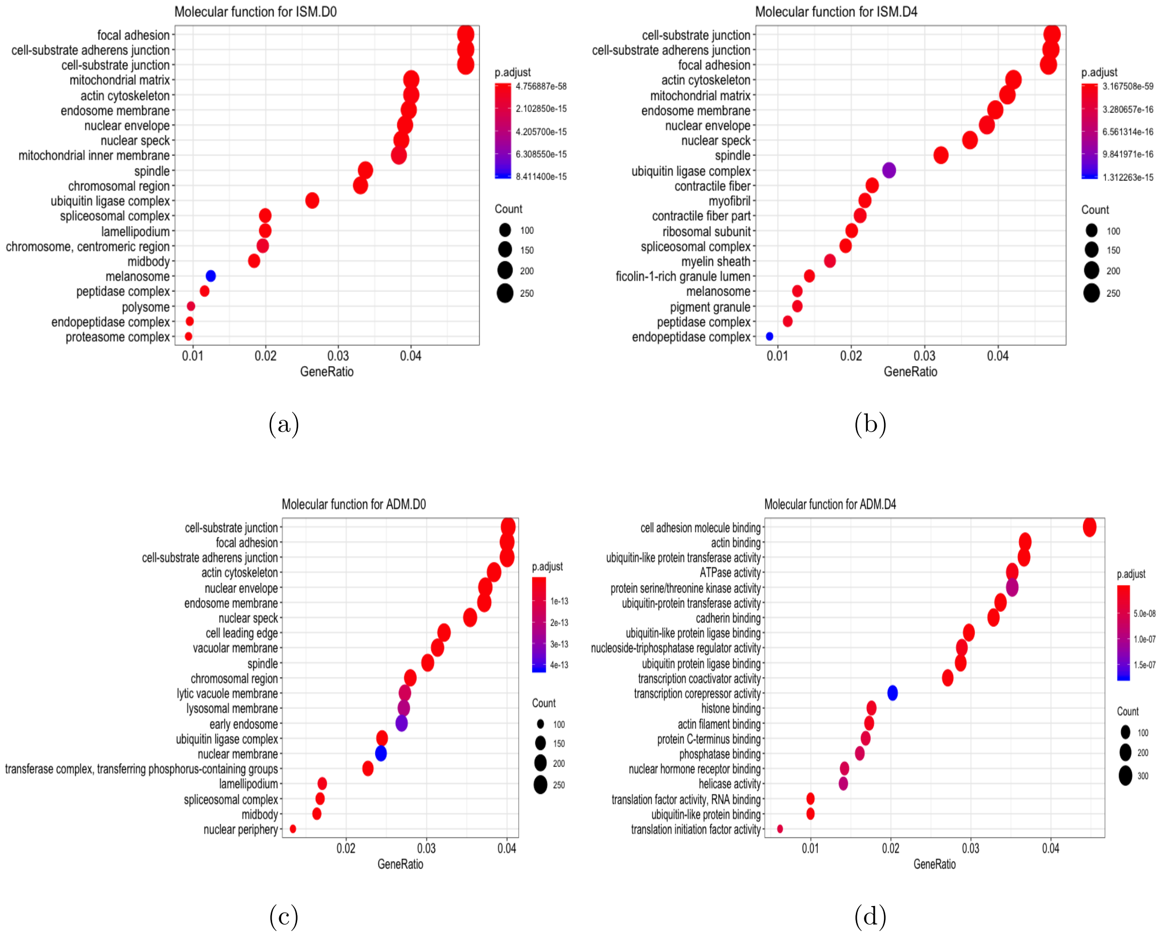
(a)

(b)

(c)

(d)

Figure 6: Gene Ontology analysis in terms of molecular function in each stage.



(a)

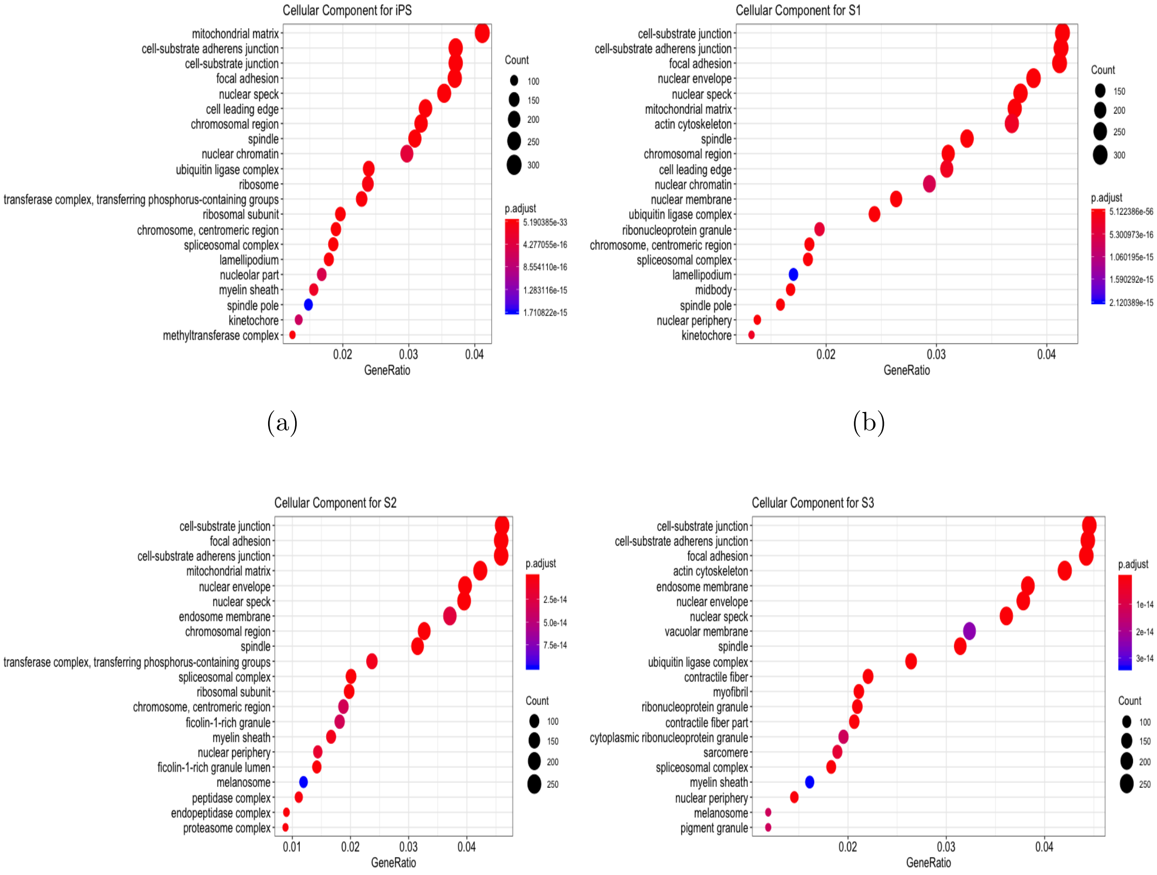
(b)

(c)

(d)

Figure 7: Gene Ontology analysis in terms of molecular function in each stage.

2.4.2 Cellular Component



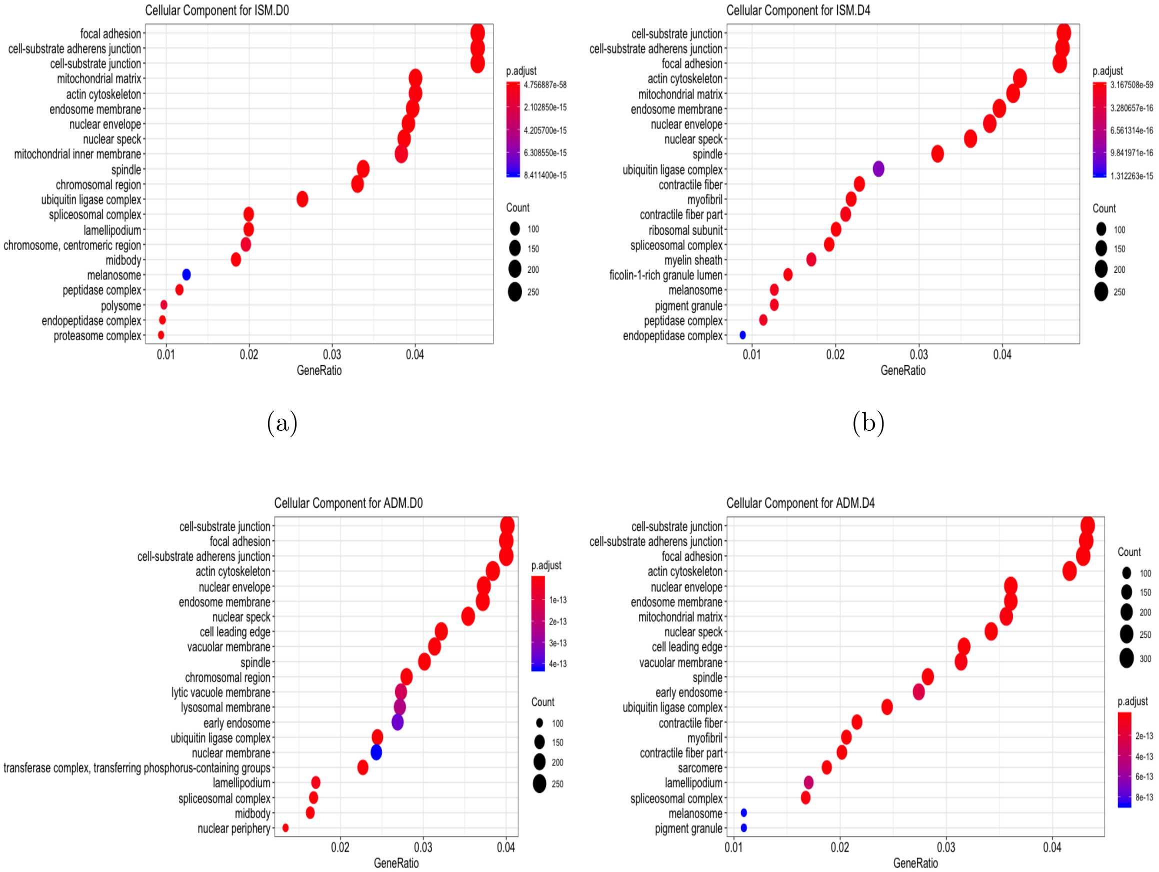
(a)

(b)

(c)

(d)

Figure 8: Gene Ontology analysis in terms of cellular component in each stage.



(a)

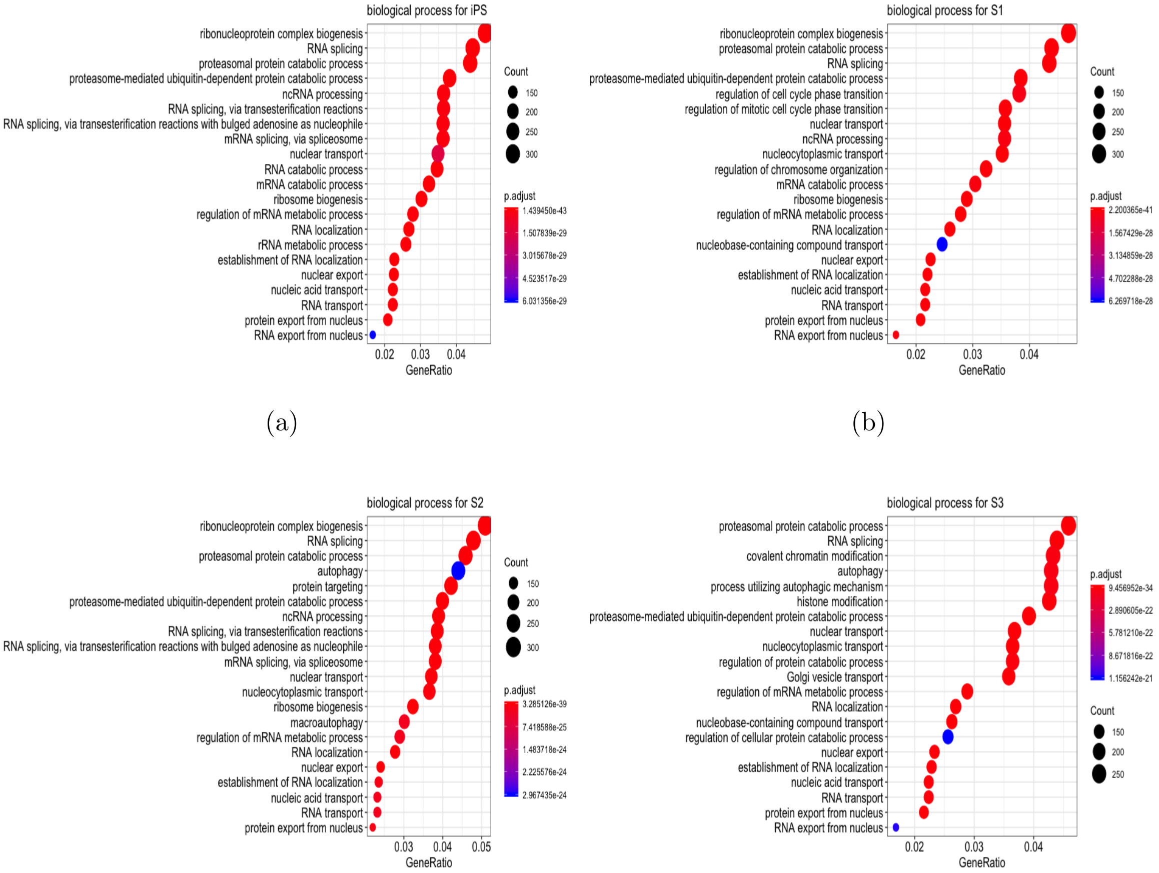
(b)

(c)

(d)

Figure 9: Gene Ontology analysis in terms of cellular component in each stage.

2.4.3 Biological Process



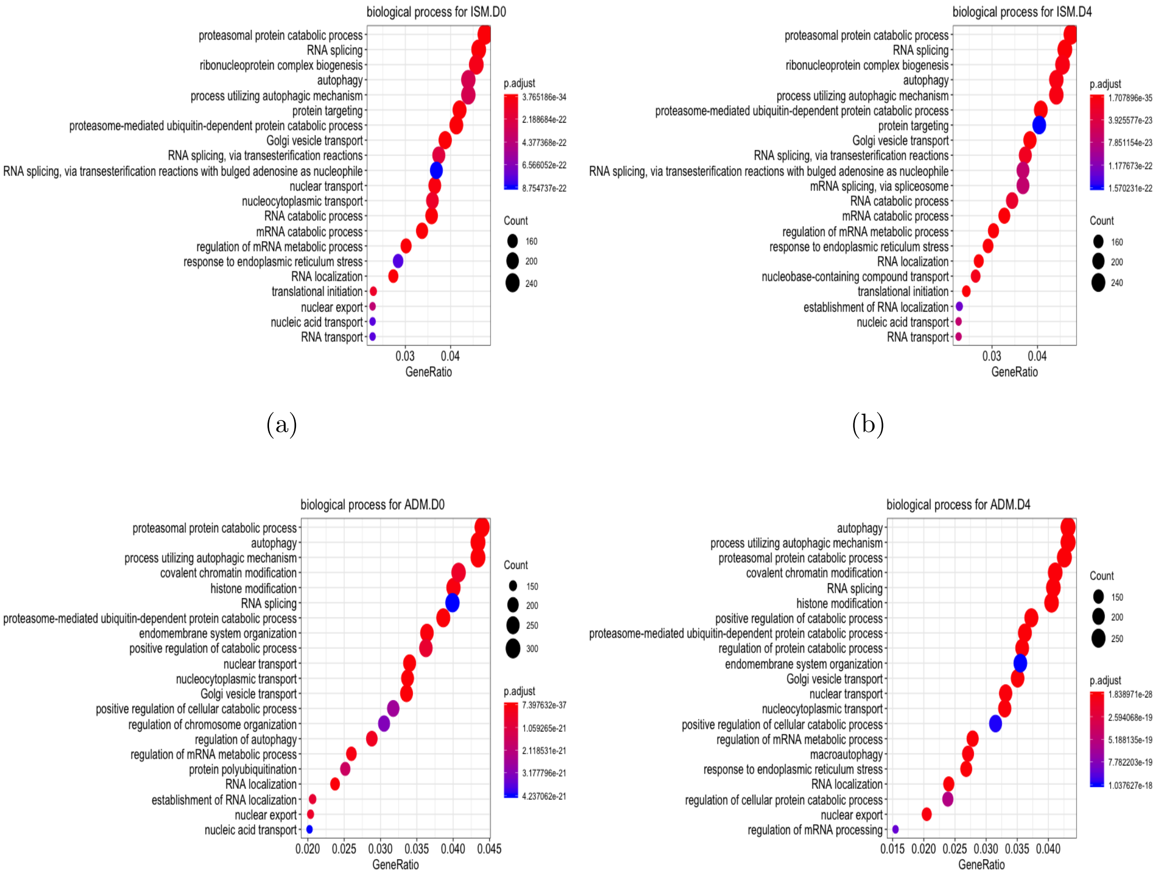
(a)

(b)

(c)

(d)

Figure 10: Gene Ontology analysis in terms of biological process in each stage.



(a)

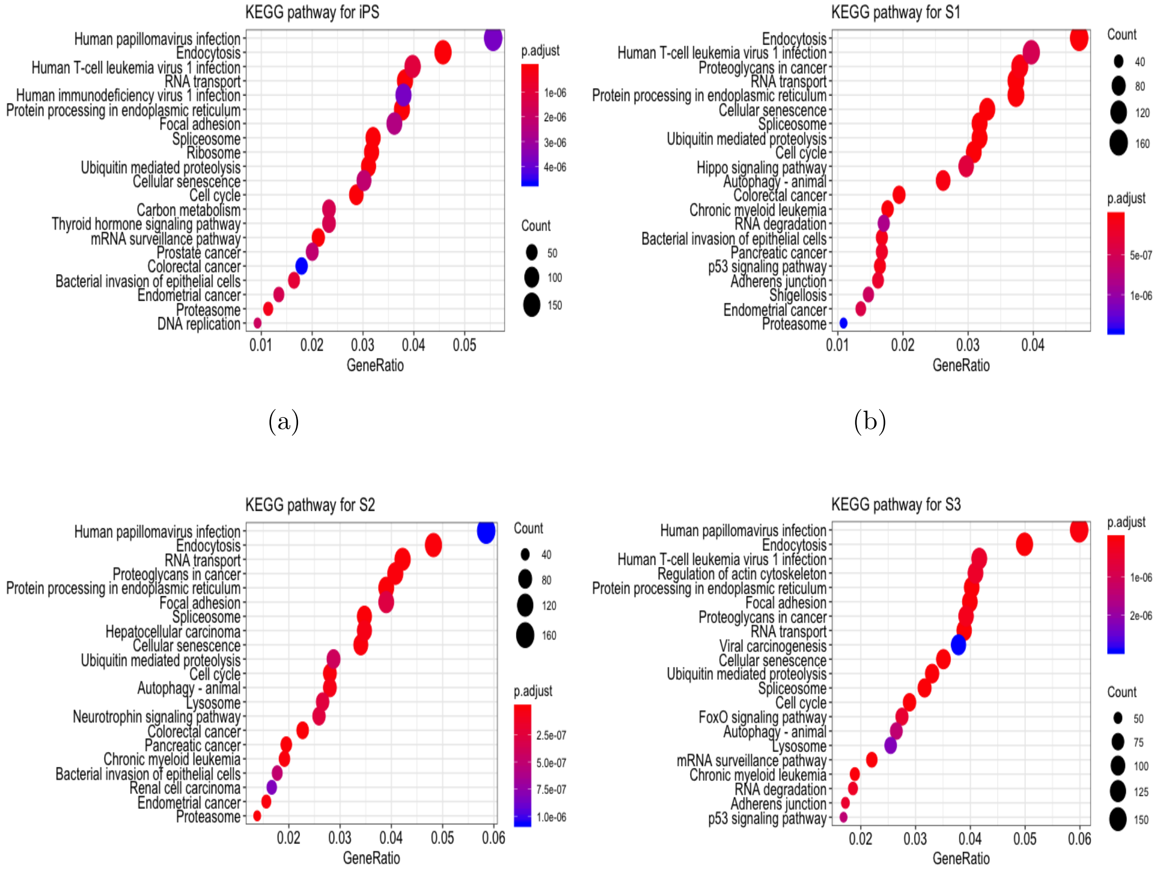
(b)

(c)

(d)

Figure 11: Gene Ontology analysis in terms of biological process in each stage.

2.4.4 KEGG analysis



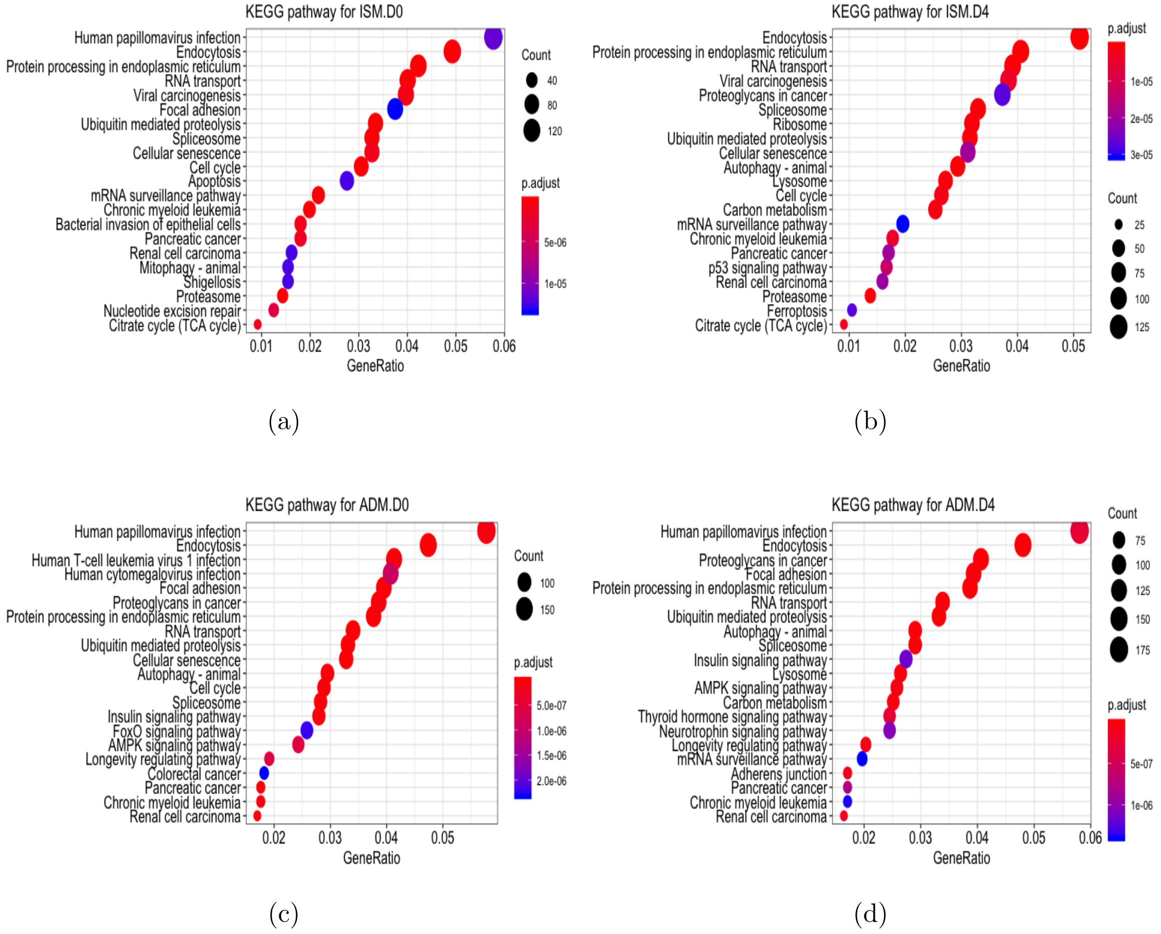
(a)

(b)

(c)

(d)

Figure 12: Gene Ontology analysis in terms of KEGG pathway in each stage.



(a)

(b)

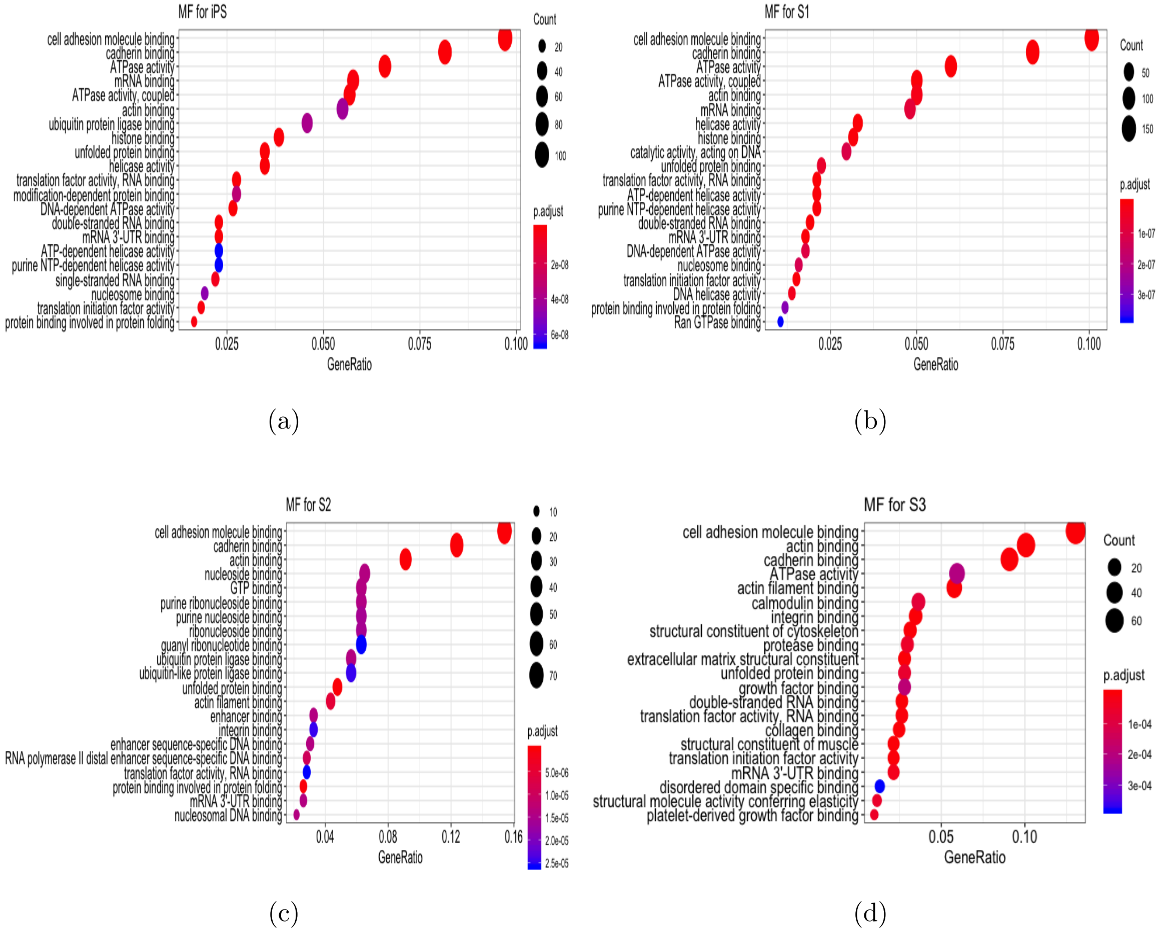
(c)

(d)

Figure 13: Gene Ontology analysis in terms of KEGG pathway in each stage.

The ontology analysis of all genes had included too many functional terms, since the analysis only concerns about the gene symbol. From previous results I noticed that for genes with more than 5 transcripts, every stage has expressed their own features, so I decided to perform a functional analysis in genes with over or equal to 5 transcripts. The result is listed below.

2.4.5 Molecular function(over 5 transcripts)



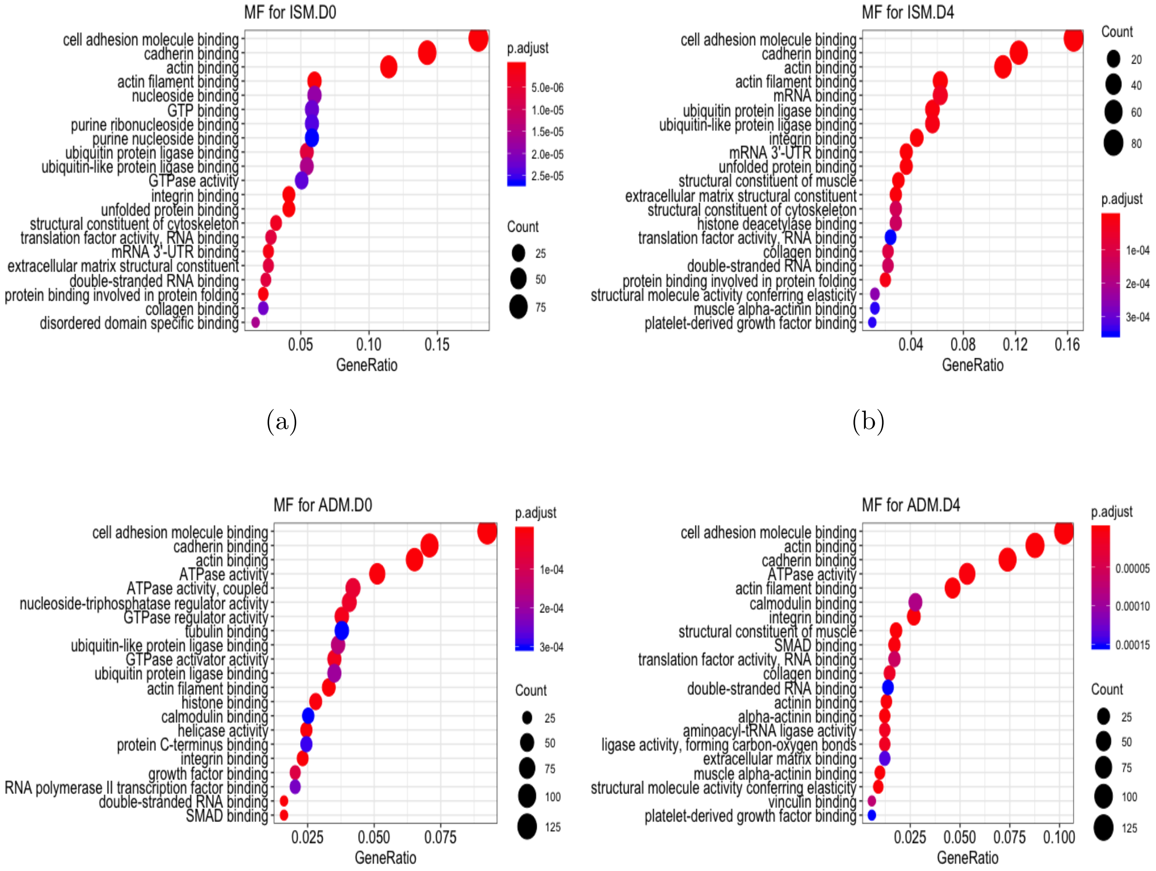
(a)

(b)

(c)

(d)

Figure 14: Gene Ontology analysis in terms of molecular function in each stage.



(a)

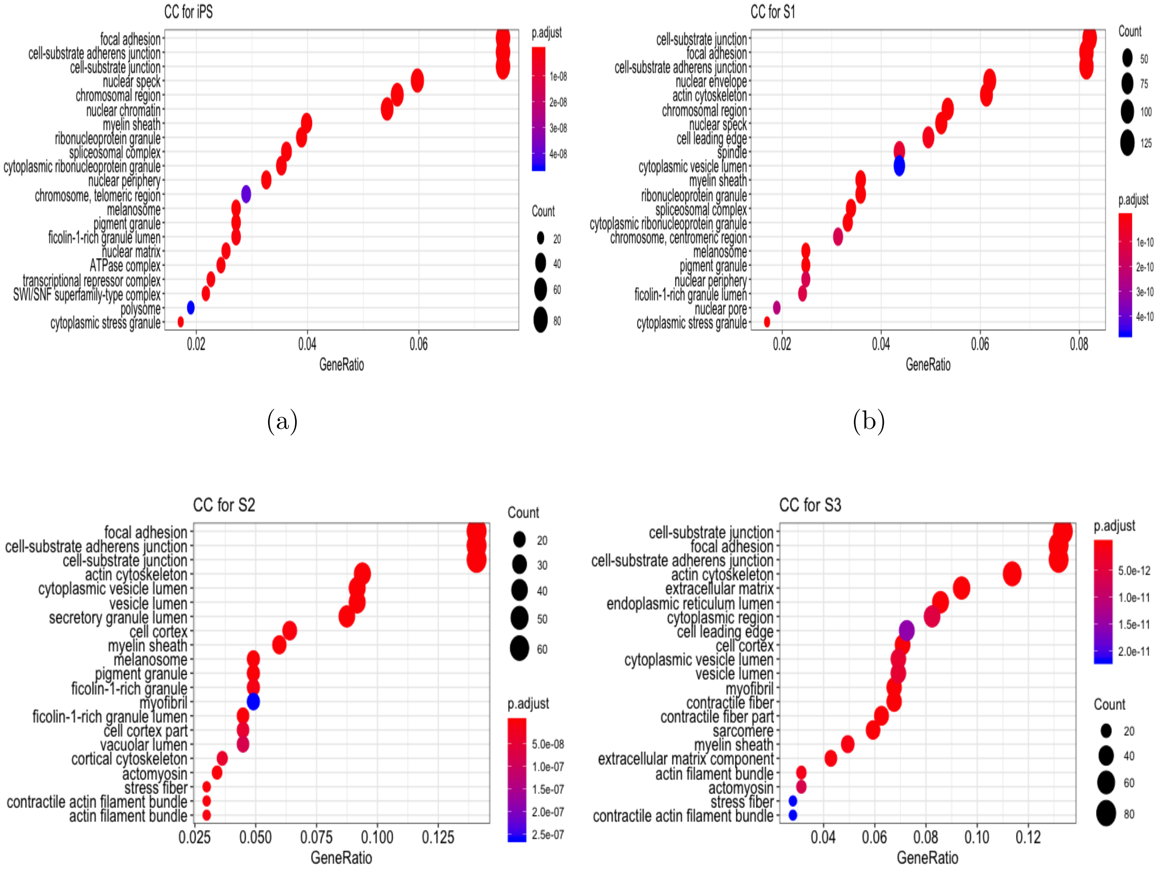
(b)

(c)

(d)

Figure 15: Gene Ontology analysis in terms of molecular function in each stage.

2.4.6 Cellular Component(over 5 transcripts)



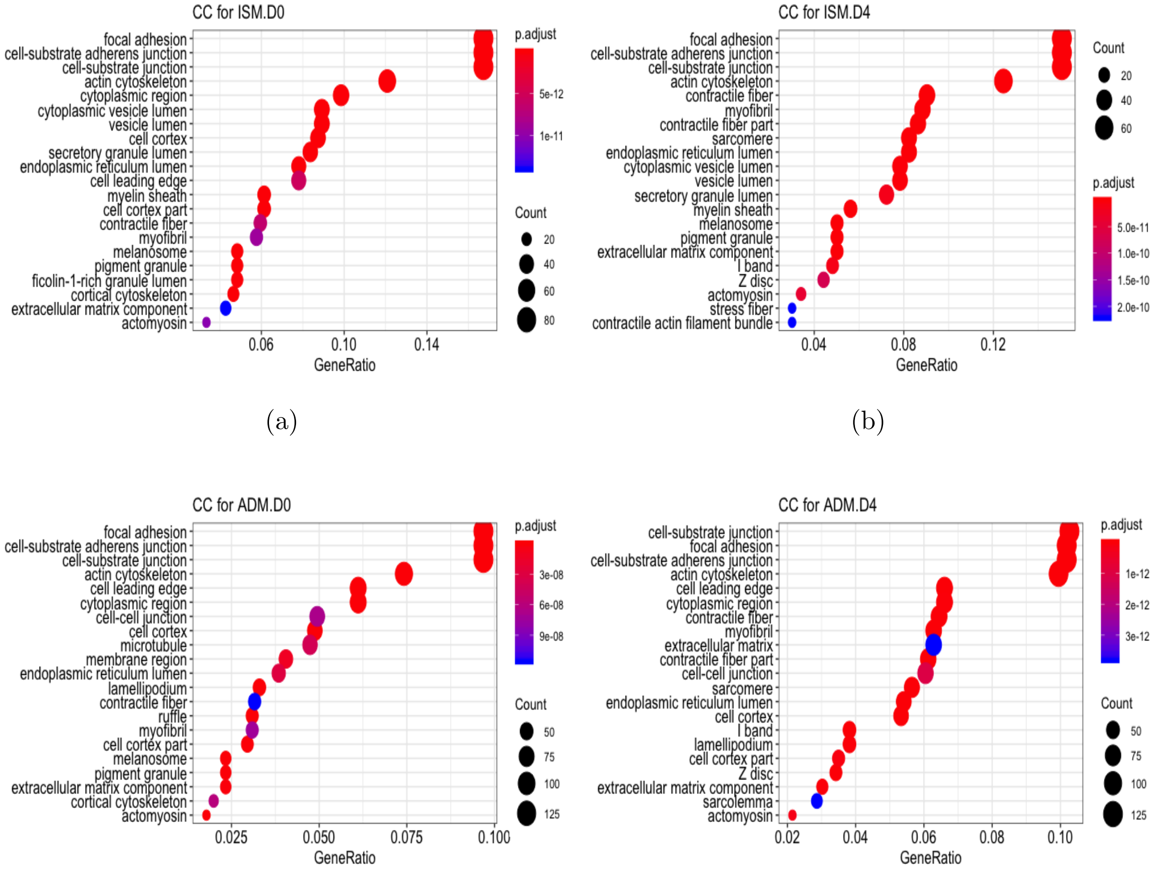
(a)

(b)

(c)

(d)

Figure 16: Gene Ontology analysis in terms of cellular component in each stage.



(a)

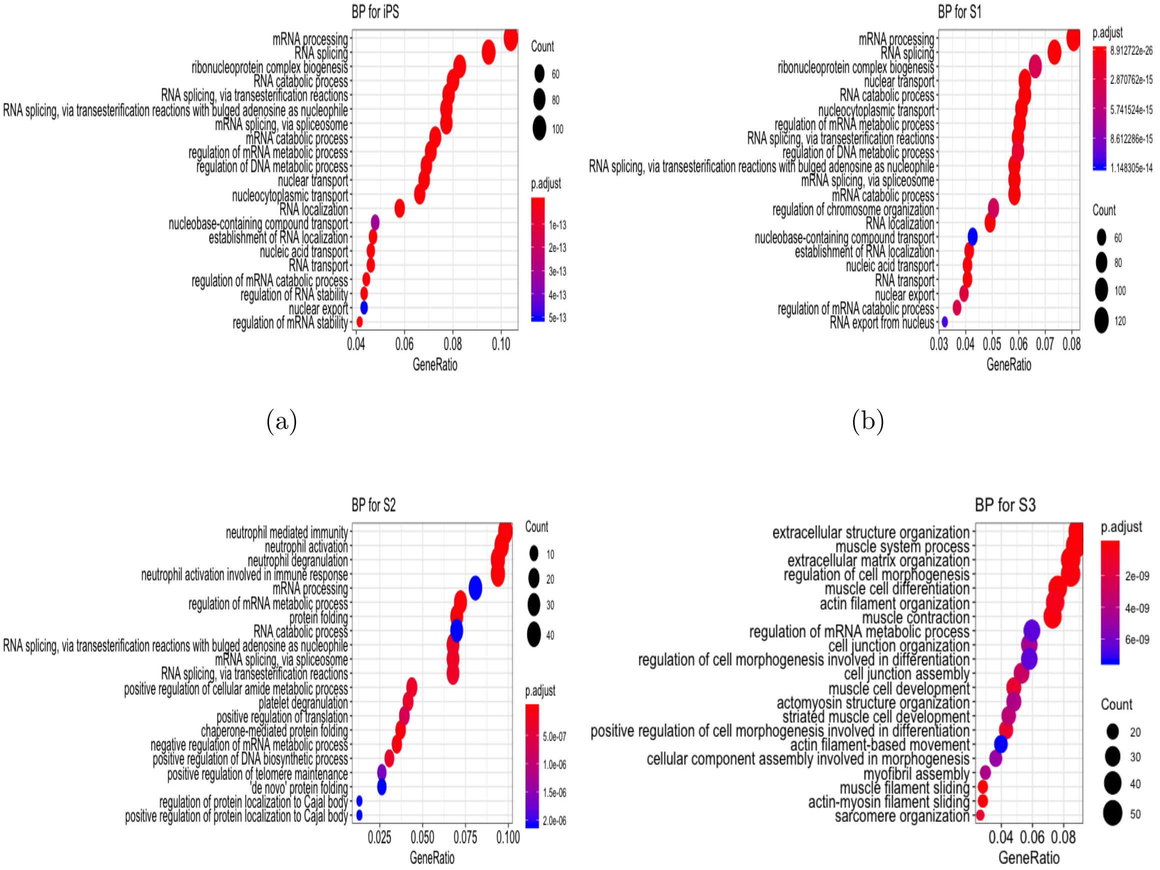
(b)

(c)

(d)

Figure 17: Gene Ontology analysis in terms of cellular component in each stage.

2.4.7 Biological Process(over 5 transcripts)



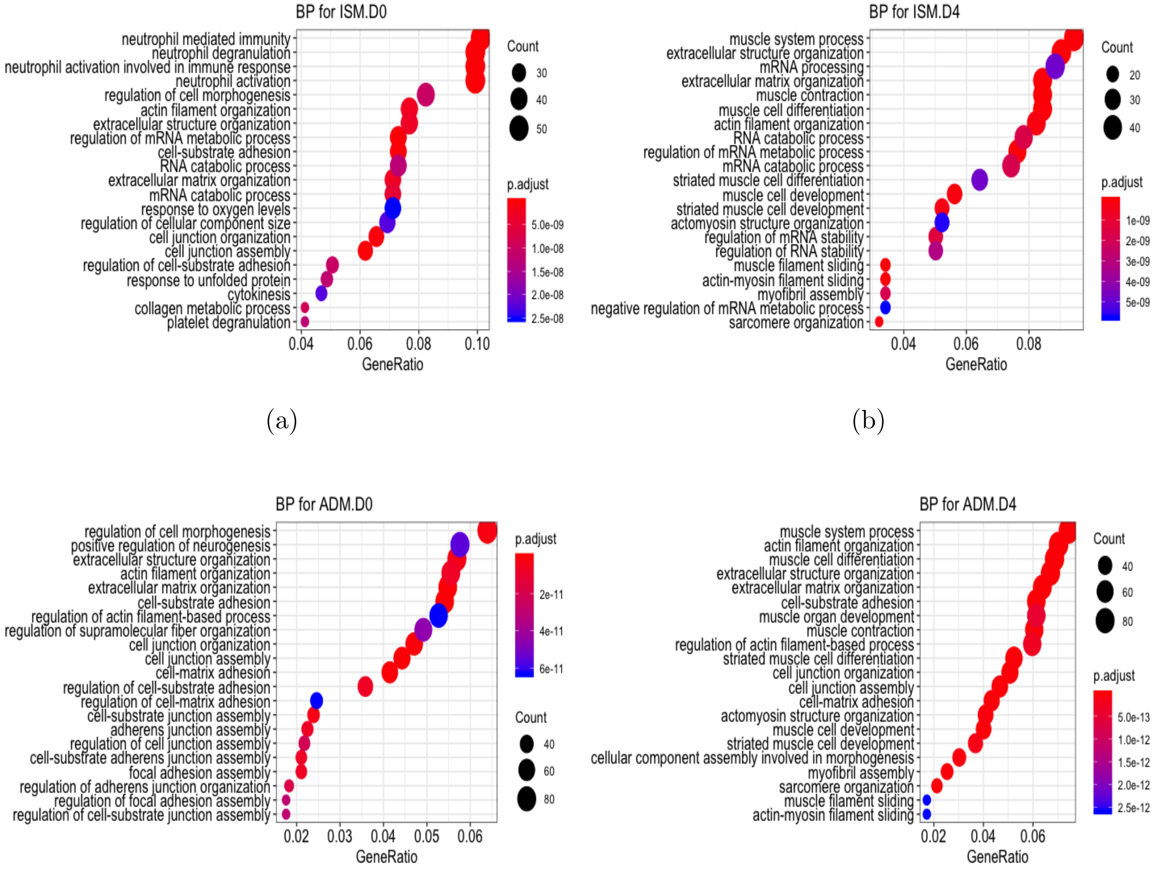
(a)

(b)

(c)

(d)

Figure 18: Gene Ontology analysis in terms of biological process in each stage.



(a)

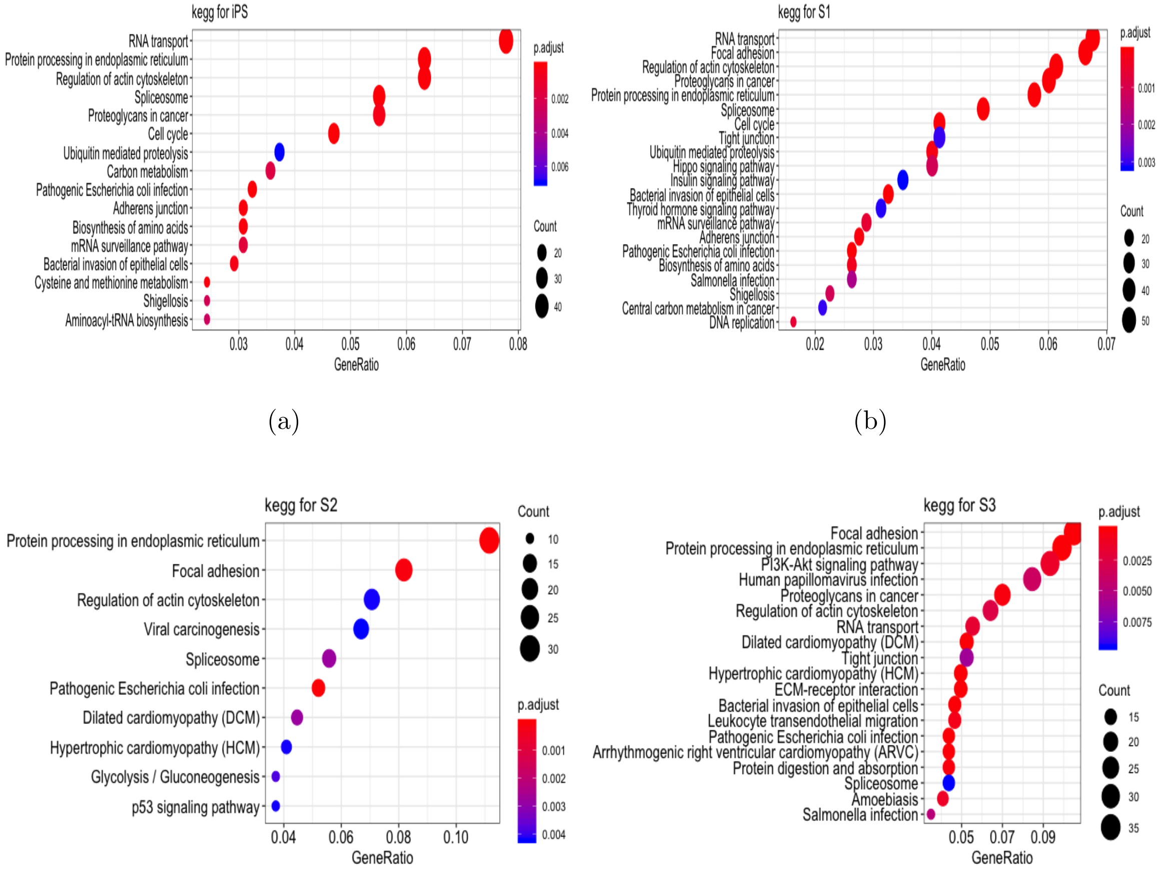
(b)

(c)

(d)

Figure 19: Gene Ontology analysis in terms of biological process in each stage.

2.4.8 KEGG analysis(over 5 transcripts)



(a)

(b)

(c)

(d)

Figure 20: Gene Ontology analysis in terms of KEGG pathway in each stage.

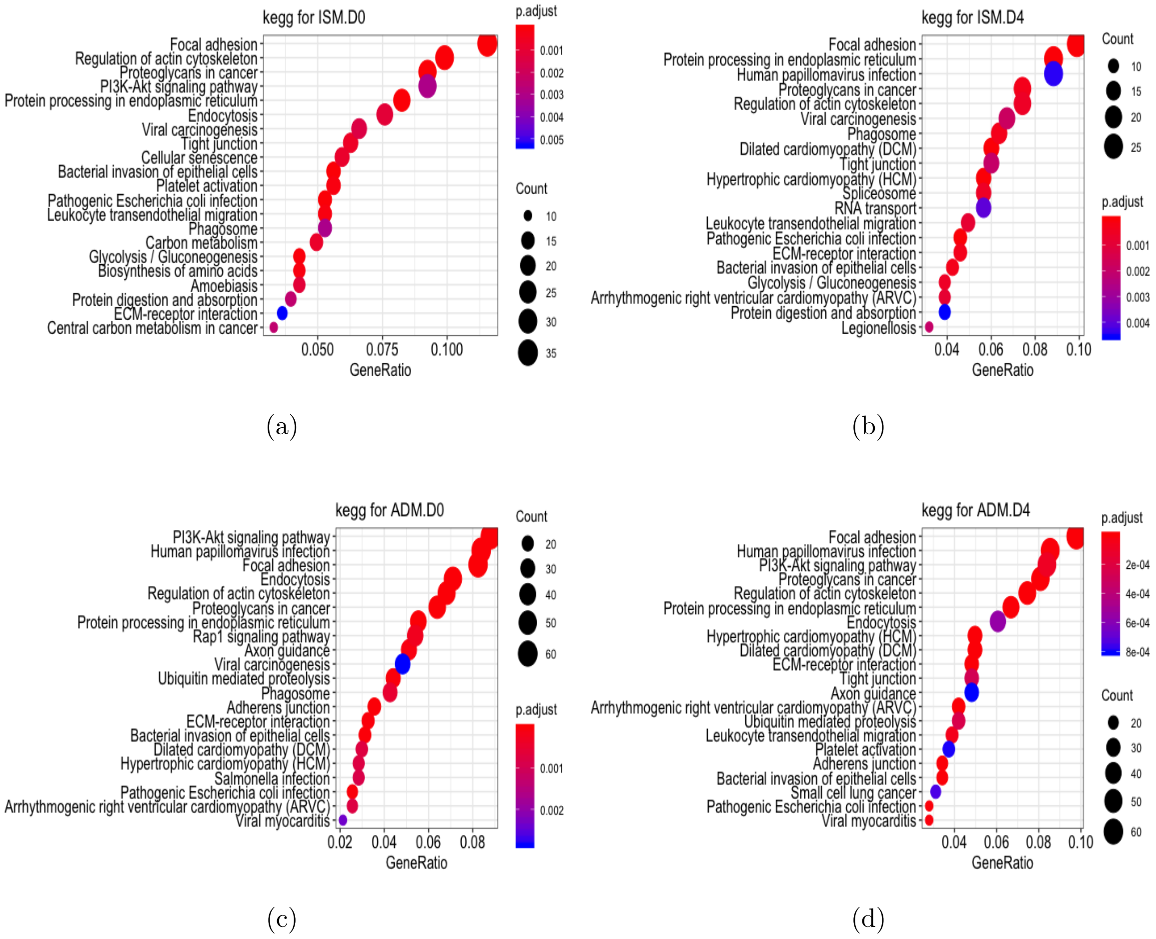


Figure 21: Gene Ontology analysis in terms of KEGG pathway in each stage.

2.4.9 Stage wise funcitonal analysis

Simply listing all of the results is pretty hard for us to find out the relations under different stages, according to the differential expressed gene analysis from 2.2, I have compared gene ontology terms between iPS stage and ADM.D4 stage. Only 3 gene symbols are listed in the table for each term, detailed information can be found in supplementary information. The terms listed below could be connected to development of adult myotubes, much differential expressed genes between iPS and ADM.D4 from 2.2 are also related to these terms.

ID	Term	Gene symbol
GO:0048193	Golgi vesicle transport	TAPBP/VPS52/CUX1
GO:0006914	autophagy	SEC22B/MTM1/MTOR
GO:0061919	process utilizing autophagic mechanism	SEC22B/MTM1/MTOR
GO:0034976	response to endoplasmic reticulum stress	HSPA1A/FLOT1/UBE2G2
GO:0042176	regulation of protein catabolic process	HSPA1A/FLNA/SEC22B
GO:0010256	endomembrane system organization	FLOT1/PDE4DIP/TARDBP
GO:0009896	positive regulation of catabolic process	HSPA1A/SEC22B/DVL1
GO:0016236	macroautophagy	SEC22B/MTM1/MTOR
GO:0016570	histone modification	NELFE/RING1/BAZ1B

Table 22: Biological Process for ADM.D4 over iPS

ID	Term	Gene Symbol
GO:0015629	actin cytoskeleton	FLOT1/FLNA/ESPN
GO:0044440	endosomal part	HLA-A/HLA-E/HLA-B
GO:0005635	nuclear envelope	ABCF1/TUBB/TRIM27
GO:0010008	endosome membrane	HLA-A/HLA-E/HLA-B
GO:0005774	vacuolar membrane	FLOT1/GNB1/MTOR
GO:0000151	ubiquitin ligase complex	HSPA1A/RING1/UBE2J2
GO:0043292	contractile fiber	SMN2/FLNA/PDE4DIP
GO:0030016	myofibril	SMN2/FLNA/PDE4DIP
GO:0044449	contractile fiber part	SMN2/FLNA/MTM1
GO:0030017	sarcomere	SMN2/FLNA/MTM1
GO:0042470	melanosome	FLOT1/SEC22B/SLC2A1
GO:0048770	pigment granule	FLOT1/SEC22B/SLC2A1

Table 23: Cellular Component for ADM.D4 over iPS

ID	Term	Gene Symbol
GO:0003779	actin binding	FLNA/ESPN/KLHL21
GO:0060589	nucleoside-triphosphatase regulator activity	DNAJC7/GDI1/SRGAP2B
GO:0051015	actin filament binding	FLNA/ESPN/CAPZB
GO:0019902	phosphatase binding	PPP1R11/RCAN3/RPA2
GO:0035257	nuclear hormone receptor binding	CDK7/ZNHIT3/PADI2
GO:0008135	translation factor activity, RNA binding	ABCF1/EIF2D/EIF4G3
GO:0032182	ubiquitin-like protein binding	DDI2/FAF1/SPRTN
GO:0003743	translation initiation factor activity	EIF2D/EIF4G3/EIF3I

Table 24: Molecular Function for ADM.D4 over iPS

ID	Term
hsa05205	Proteoglycans in cancer
hsa04140	Autophagy - animal
hsa04910	Insulin signaling pathway
hsa04142	Lysosome
hsa04152	AMPK signaling pathway
hsa04722	Neurotrophin signaling pathway
hsa04211	Longevity regulating pathway
hsa04520	Adherens junction
hsa05212	Pancreatic cancer
hsa05220	Chronic myeloid leukemia
hsa05211	Renal cell carcinoma

Table 25: KEGG pathway for ADM.D4 over iPS

2.5 Possible Novel genes and transcripts

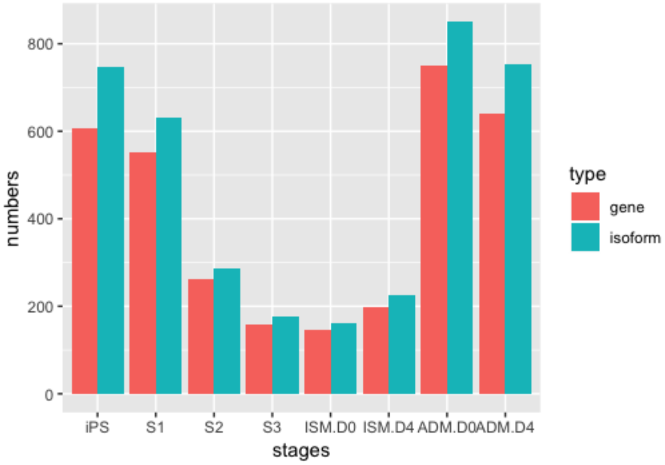


Figure 22: unmatched genes and transcripts

According to the description of ToFU pipeline, for each input sequence, NA will be produced if reference can not be found in reference gene model during the annotation step [8]. Natural we could assume these sequences that can not be matched to be novel transcripts. The potential novel transcript numbers are produced by: $total\ transcript\ numbers - matched\ transcript\ numbers$, the potential novel gene numbers are produced by: $total\ gene\ numbers - matched\ gene\ numbers$. IGV sequence visualization was used to check the authenticity of the novel transcripts preliminary, several transcripts were visualized and proved to be not overlapping with any reference genemodel area.

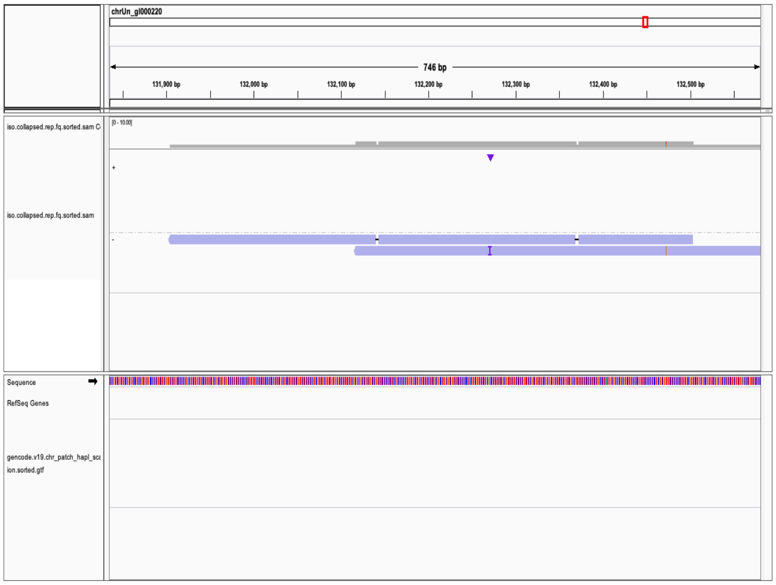


Figure 23: visualization of one novel gene region

Figure 13 is a visualization of 2 reads from iPS stage, $c22873/f1p0/596 \mid GL000220.1 : 132118 - 132868(-)$ and $c22883/f1p1/742 \mid GL000220.1 : 132118 - 132868(-)$. After the previous error correction and collapse step, all redundant reads are collapsed into one unique transcript. Here, 2 reads are 2 transcripts, they are not overlapped with any transcript or gene with Genecode v19 reference gene model, they are likely novel transcripts and from a novel gene.

3 Discussion

3.1 Conclusions

In this study, we characterized the transcriptome information for human skeletal muscle cell in different stages. The transcripts and alternative splicing events distribution are investigated across stages. Based on a comprehensive analysis of differentiated cells, the expression and functional information obtained in this study revealed the process of human skeletal muscle cell differentiation. The data will provide a genomic reference for further skeletal muscle cell differentiation or FSHD research using iPS cells.

There are also some limitations of our work. The 8 datasets in our data are all unique ones, which means the features shown in our data could possibly be coincidences. No replica also brought lots of trouble when I was performing the gene differential expression analysis, because high FDR value, lots of results were eliminated.

Also our data are simply collected from healthy people, with no data from patient we can only explore the gene expression features during skeletal muscle cell differentiation, while the genes and pathways information related to FSHD can not be obtained.

And the largest challenge in this study is lacking resource for Iso-Seq analysis. Iso-Seq is a relatively new technique, it is kind of advanced than other RNA-Seq techniques in many aspects, which also means we are not able to find many publications related to Iso-Seq anlysis. Before I was running the analysis following Cupcake ToFU pipeline, 4 different kinds pipeline had been tried and abandoned, two of them requires short read sequences(IDP and SpliceGrapher). Pipeline SQANTI can not even work on their own tutorial data, and pipeline TAPIS contains lots of coding error in their python scripts and spent me lots of time fixing the script bugs(the fixed scripts is available in splimentary information and TAPIS website). When TAPIS was finally managed to work, I noticed that it would alter the strand information of the sequence for some unknown reason. These situations all tell that this is a field that few people are working on.

3.2 Future work

Lacking experimental group is the major limitation of this study. Our sequencing data are all from healthy people. Without an experimental group, we can only acquire cell differentiation related information, by comparing healthy data with patient data we could easily locate genes or pathways underlying FSHD.

More replica could also be used in further study. Since read numbers and gene numbers can vary from sample to sample due to technique reasons, it is important to set several replicas to reduce the bias. In our data set, the over 3k part data of S2 stage is dropped because of containing much mitochondrial sequence, this brought more uncontrolled bias to our analysis. Solutions for such circumstances should be discussed to ensure to correctness of the analysis.

At last, The analysis revealed the gene expression and gene ontology differences between stages and cell lines, hypothesis were made based on these results. However, more biological experiments should be performed to validate the basic results and hypothesis.

References

[1] Charles A. Goldthwaite. The promise of induced pluripotent stem cells (ipscs). https://stemcells.nih.gov/info/Regenerative_Medicine/2006chapter10.htm.

[2] DeSimone AM, Pakula A, Lek A, and Emerson CP Jr. Facioscapulohumeral muscular dystrophy. *Compr Physiol.*, pages 1229–1279, sep 2017.

[3] National Institute of Health. Facioscapulohumeral muscular dystrophy. https://ghr.nlm.nih.gov/condition/facioscapulohumeral-muscular-dystrophy#.

[4] Michelle N. Vierra, Sarah B. Kingan, Elizabeth Tseng, Ting Hon, William J. Rowell, Jacquelyn Mountcastle, Olivier Fedrigo, Erich D. Jarvis, and Jonas Korlach. From rna to full-length transcripts: The pacbio iso-seq method for transcriptome analysis and genome annotation. https://www.pacb.com/wp-content/uploads/Vierra-G10K-2017-From-RNA-to-Full-Length-Transcripts-1.pdf.

[5] Salah E. Abdel-Ghany, Michael Hamilton2, Jennifer L. Jacobi, Peter Ngam, Nicholas Devitt, Faye Schilkey, Asa Ben-Hur, and Anireddy S.N. Reddy. A survey of the sorghum transcriptome using single-molecule long reads. *nature communication*, jun 2016.

[6] Pacific Biosciences. Iso-seq analysis(isoform sequencing). https://www.pacb.com/videos/tutorial-iso-seq-analysis-application/.

[7] Thomas D. Wu and Colin K. Watanabe. Gmap: a genomic mapping and alignment program for mrna and est sequences. *Bioinformatics*, 2005.

[8] Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, and et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mrna sequencing. *PLoS One*, jul 2015.

[9] M. F. Rogers, J. Thomas, A. S. N. Reddy, and A. Ben-Hur. Splicegrapher: Detecting patterns of alternative splicing from rna-seq data in the context of gene models and est data. *Genome Biology*, 2012.

[10] Robinson MD, McCarthy DJ, and Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010.

[11] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 2016.

[12] James T. Robinson, Helga Thorvaldsdttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, may 2011.