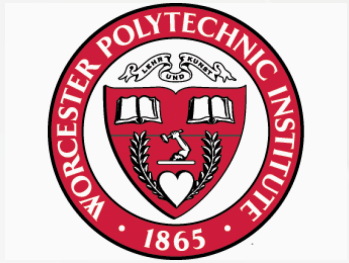


Genome Studies of Gene Expression and Alternative Splicing of iPSC Skeletal Muscle Induction and Differentiation

Speaker: Yibo Wu

Content

01 Background

02 Methods & Results

03 Discussion

Part 01
Background

Facioscapulohumeral Muscular Dystrophy(FSHD)

- Muscle weakness and wasting
- Autosomal dominant
- Decreased facial expression
- Trouble walking
- Difficulty standing up

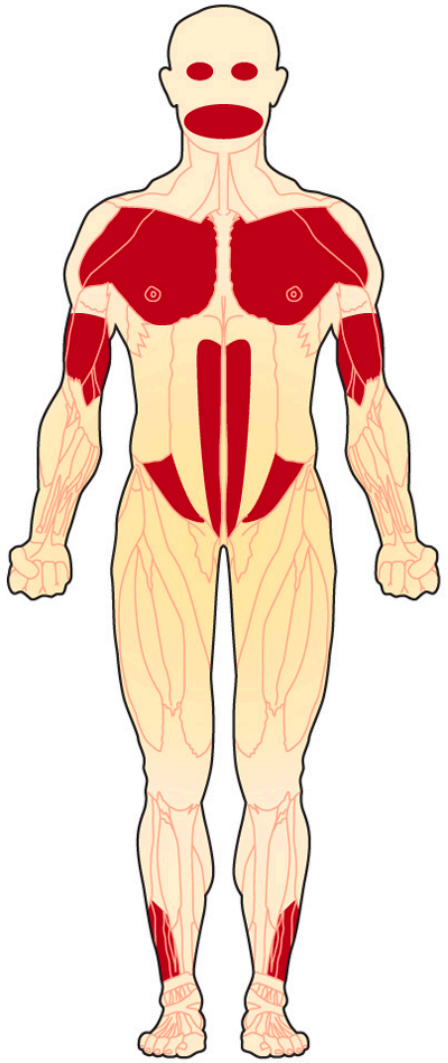
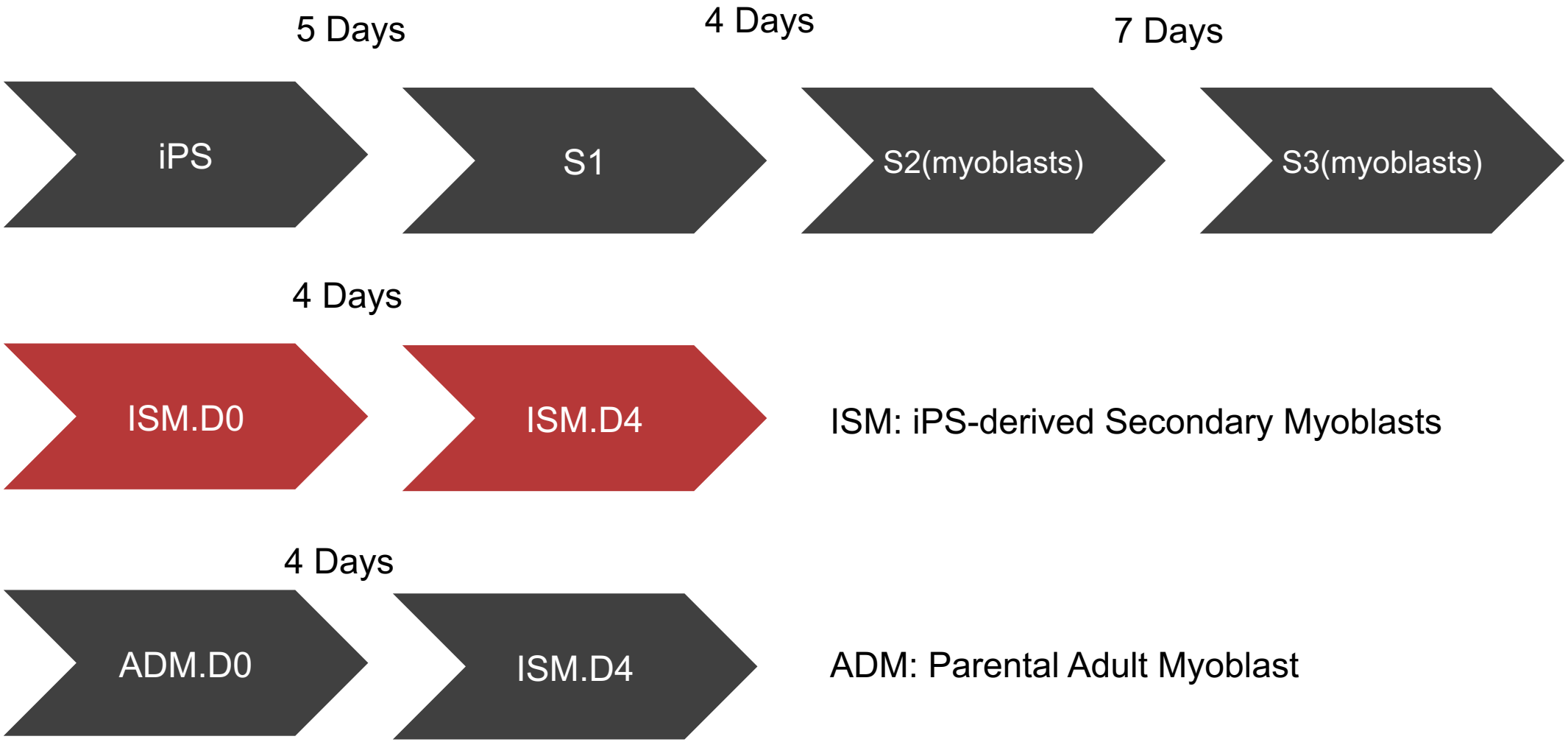


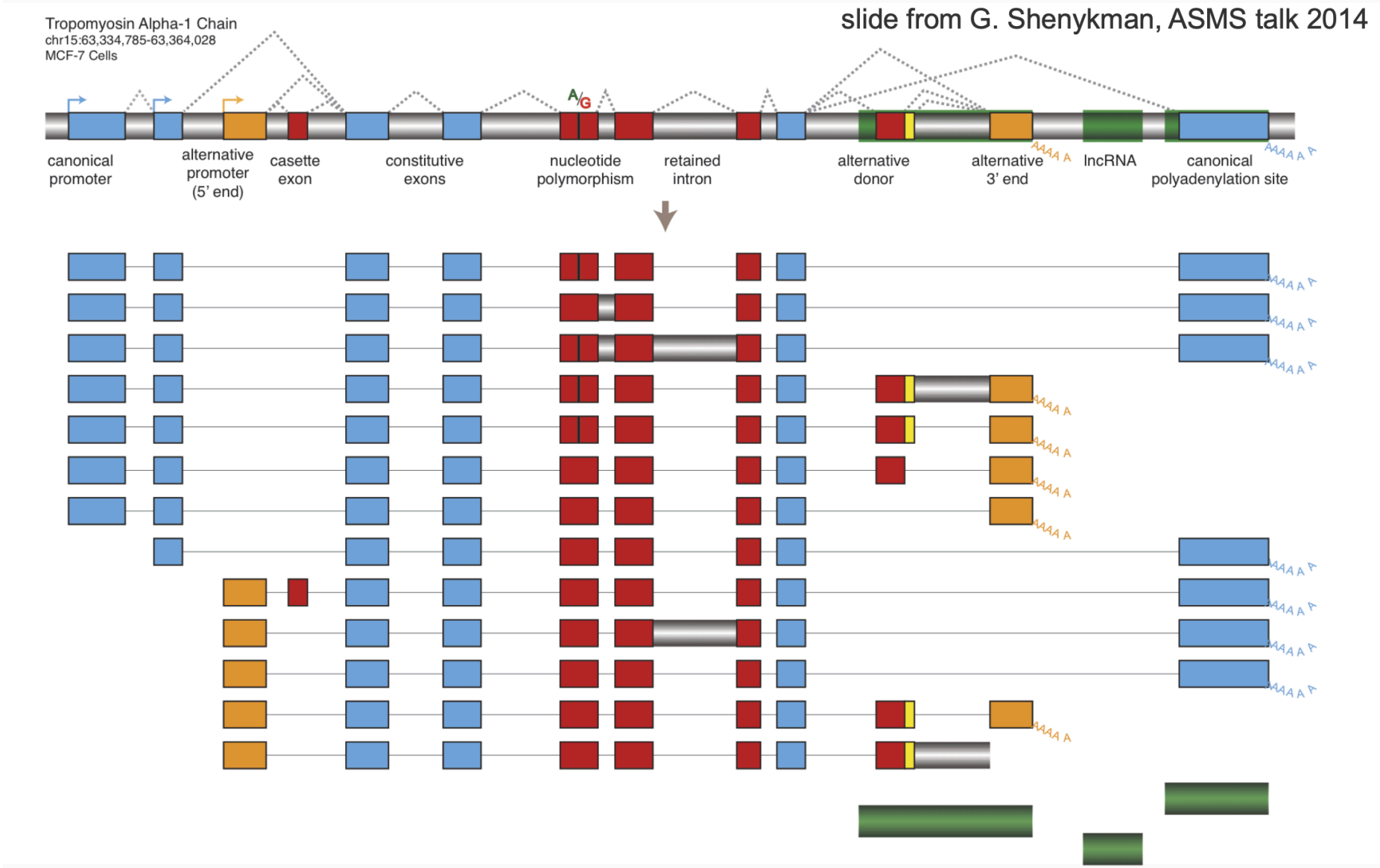
Figure source:
https://www.google.com/search?q=FSHD&rlz=1C5CHFA_enUS794US794&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiB7bvekMDiAhVKMqwKHYKNAasQ_AUIDigB&biw=1440&bih=733#imgrc=T-5jAIlgggL9bM:

Induced Pluripotent Stem Cells(iPSC)



Iso-Seq

slide from G. Shenykman, ASMS talk 2014

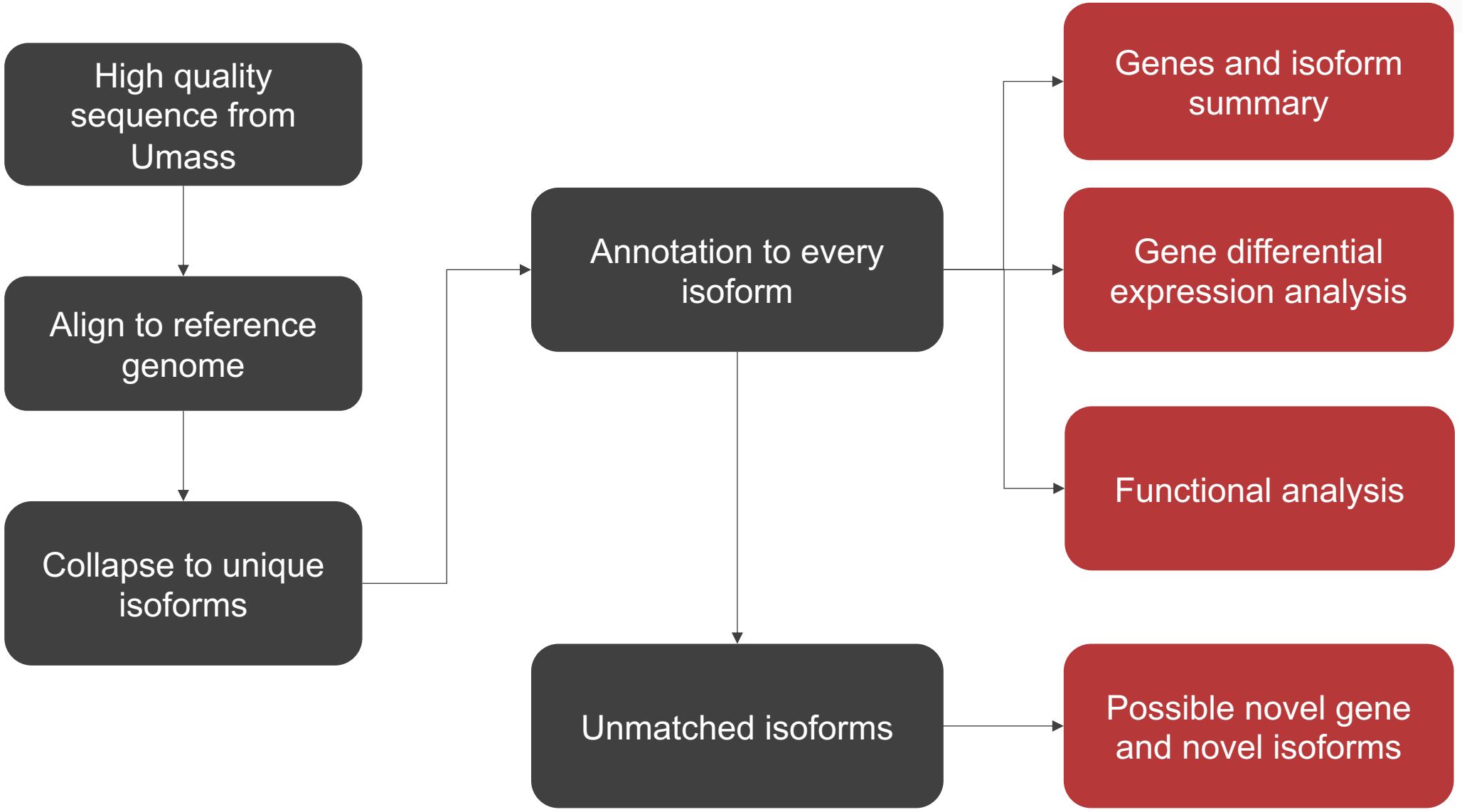


Iso-Seq

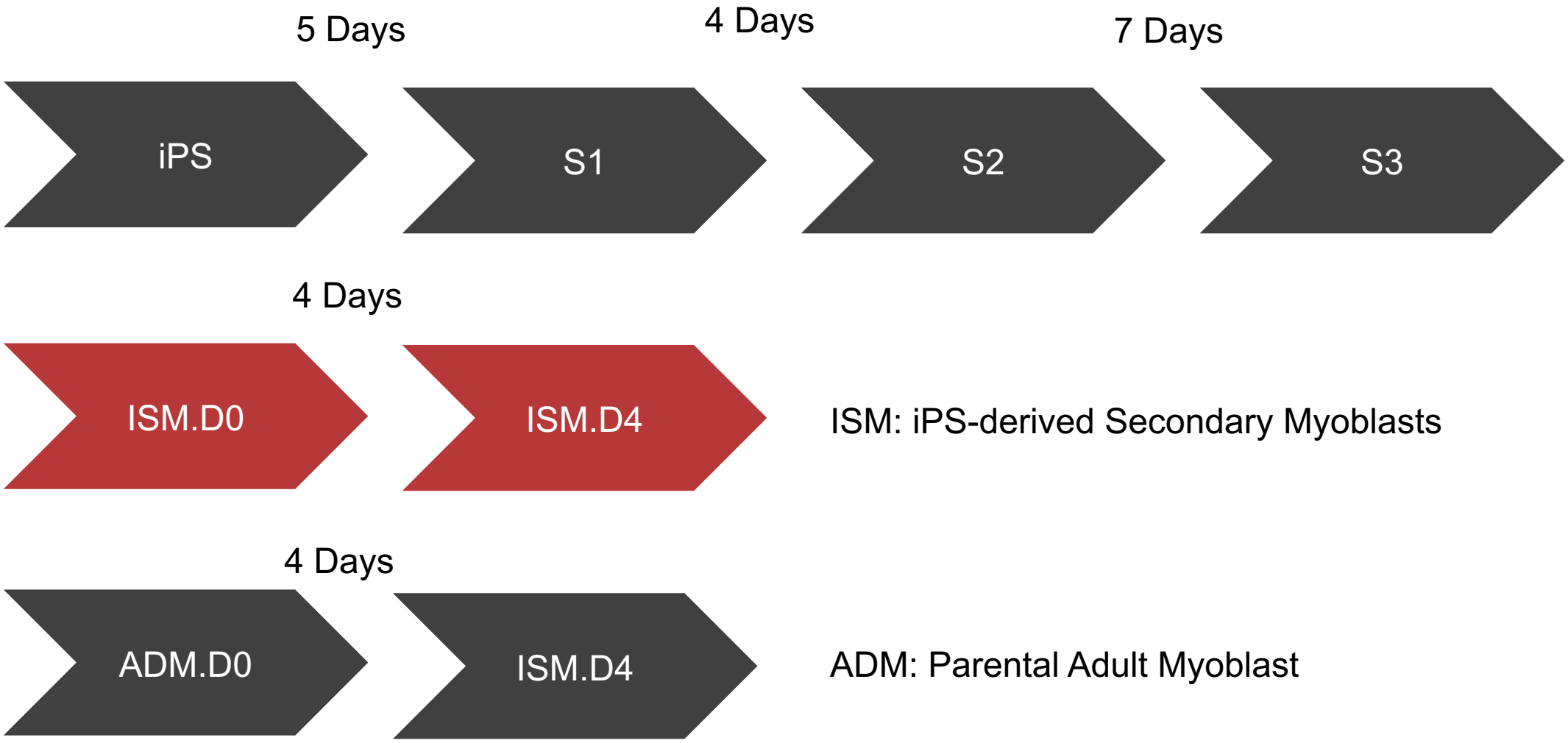
- Employs long read
  - Avoid errors during short reads assembly
  - One read = one transcript
  - Full-length(have both 5` end primer and poly A)
  - High quality(predicted accuracy >=99%)
  - Supported by 2 or more full length reads

Part 02
Methods

Methods

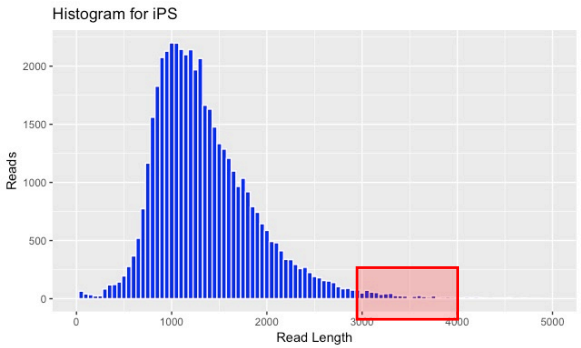

High quality sequence from Umass
Align to reference genome
Collapse to unique isoforms
Annotation to every isoform
Unmatched isoforms
Genes and isoform summary
Gene differential expression analysis
Functional analysis
Possible novel gene and novel isoforms

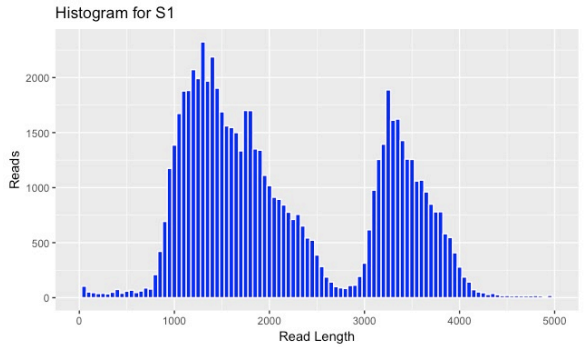
Data summary

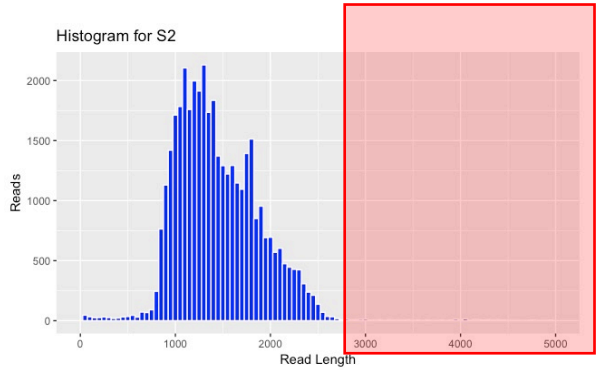


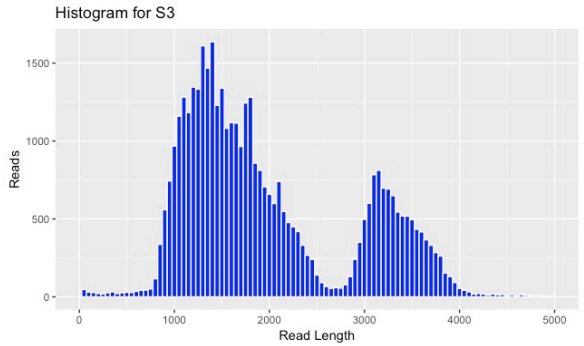
Data summary

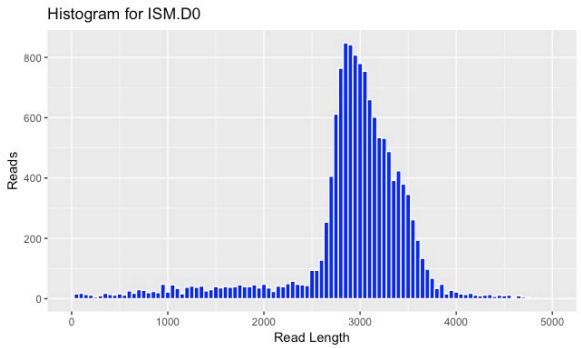
Tool: samtools & ggplot2

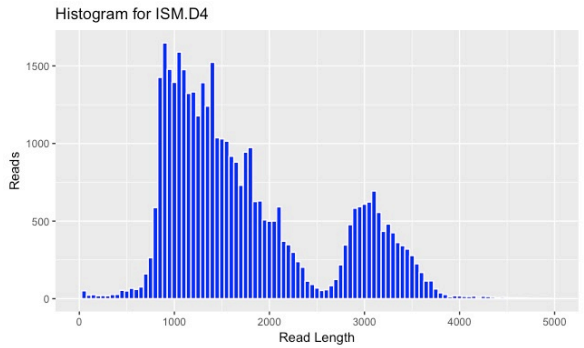


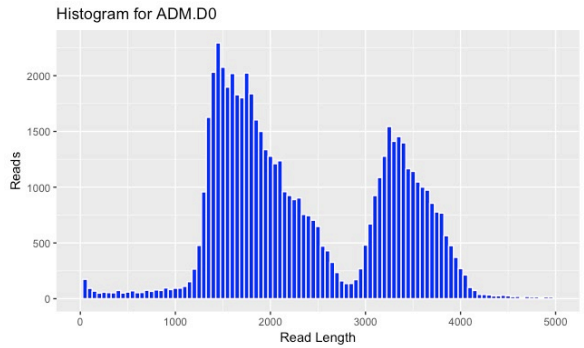


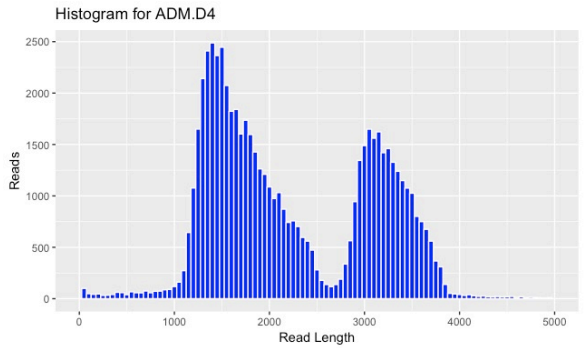












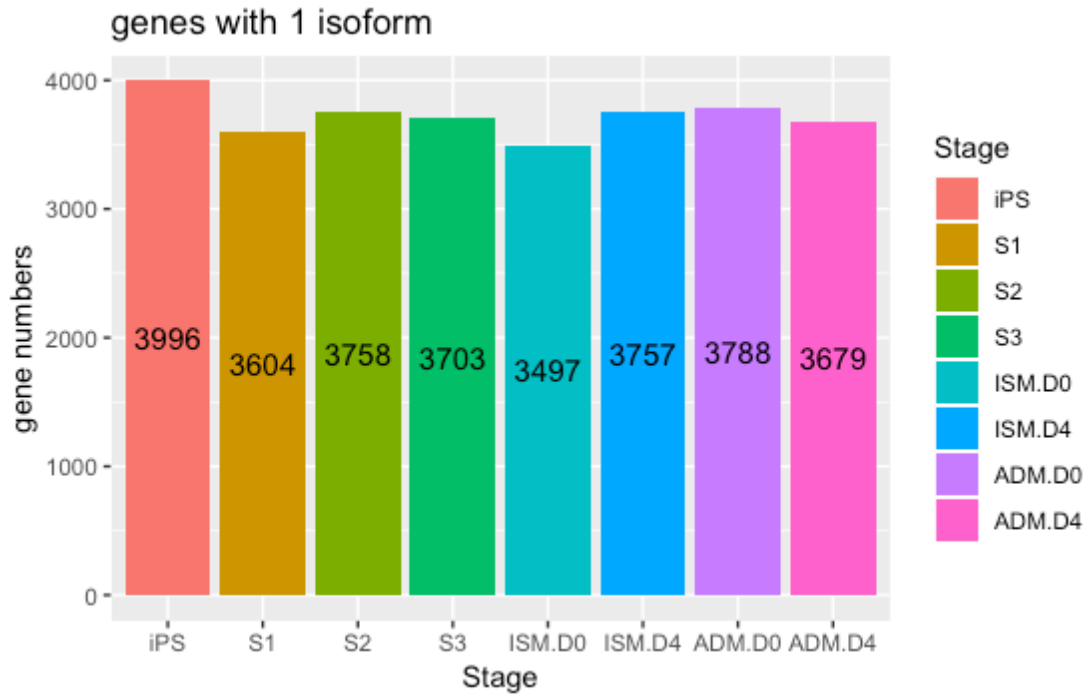
Data summary

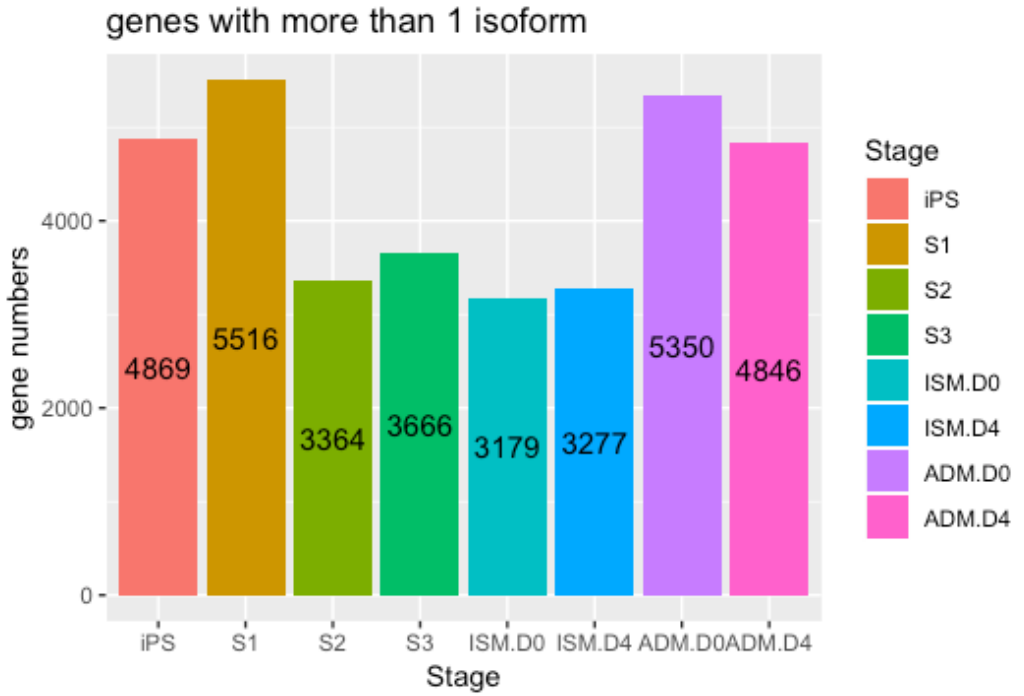
Step	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0	ADM.D4
High Quality isoforms	28282	34282	17654	19942	17174	17847	42037	31022
Unique isoforms	23199	27497	14580	16564	14360	14895	27528	25422
Matched isoforms	22450	26866	14292	16387	14197	14670	26677	24667
Unmatched isoforms	749	631	288	177	163	225	851	755

Unique isoforms: one read = one isoform = one alternative splicing event

Reference genome: Genecode v19

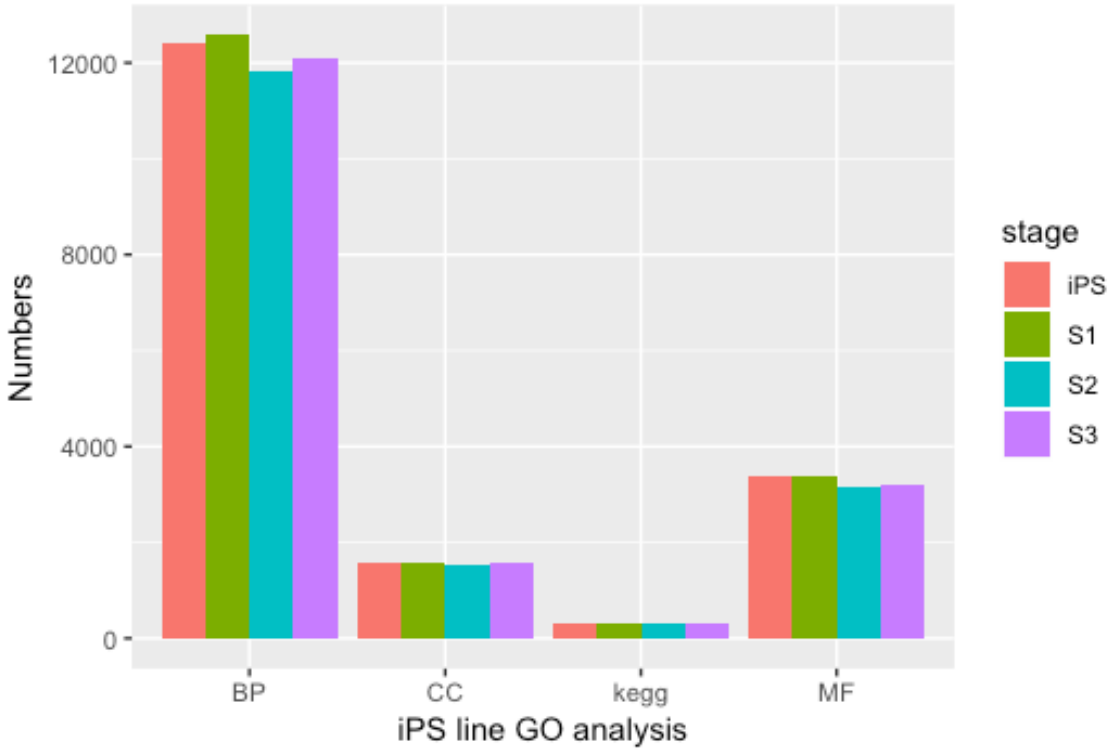
Data summary

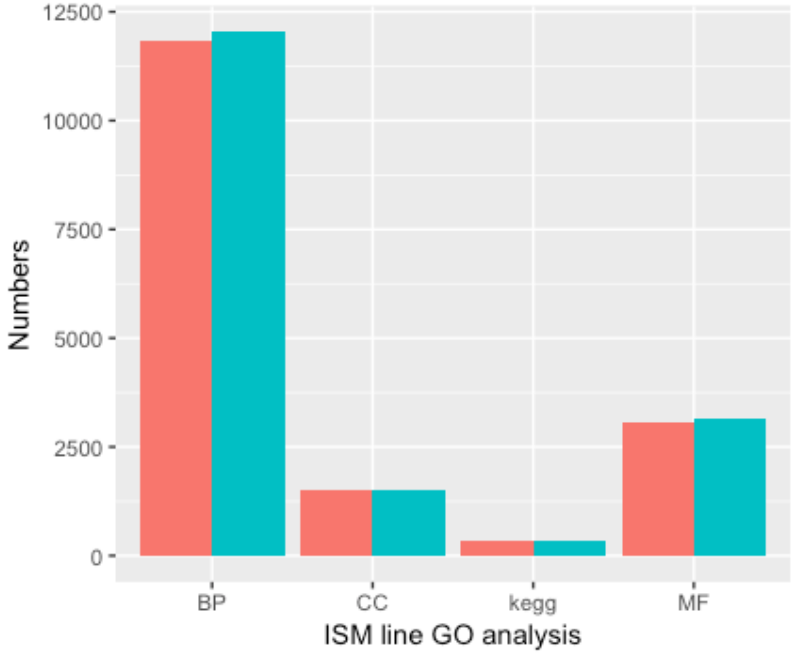


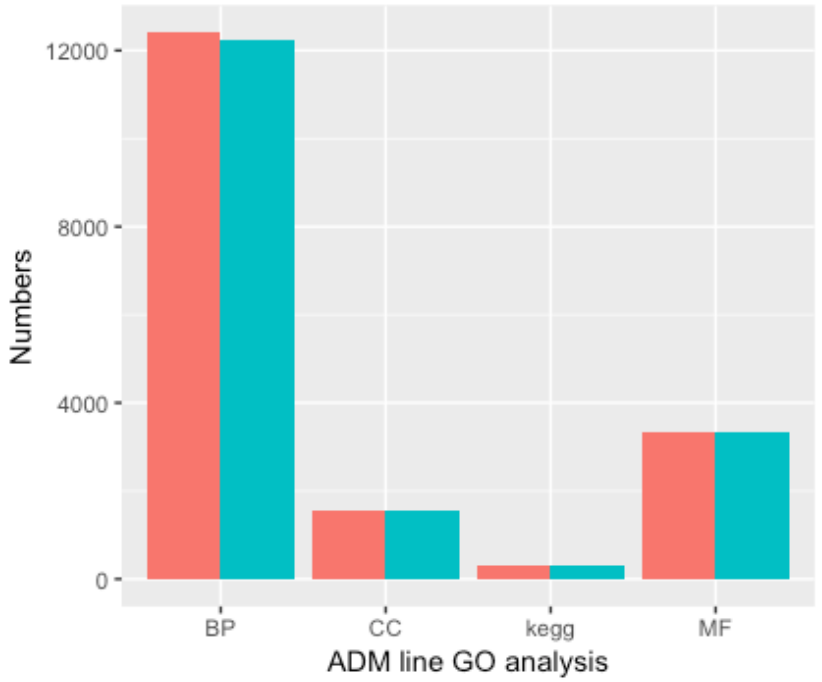


stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0	ADM.D4
Isoform numbers	18454	23262	10534	12684	10700	10913	22889	20988

Functional analysis







Differential expression analysis

Fold change: how much the expression level changed.

False Discovery Rate(FDR): reliability of the result.

$|log_2 \text{fold}| > 1$
FDR < 0.05

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	\						
S2	\	\					
S3	9	6	4				
ISM.D0	1	1	\	1			
ISM.D4	8	3	2	\	2		
ADM.D0	15	5	1	2	\	1	
ADM.D4	24	27	17	1	11	\	5

Differential expression analysis

Fold c

False

	logFC	logCPM	PValue	FDR
COL11A1	3.995970	9.815278	7.558856e-06	0.1200422
ESRG	-7.460962	8.969640	3.288446e-05	0.1740794
LIMCH1	7.412118	9.057600	3.288446e-05	0.1740794
SLIT2	7.348373	9.004297	4.932668e-05	0.1958393
NAV3	7.138266	8.831066	1.138974e-04	0.3617609
GRID2	-7.048651	8.624644	2.729433e-04	0.6192303
MME	6.979005	8.702447	2.729433e-04	0.6192303
ARID5B	6.892257	8.633434	4.289109e-04	0.8514417
COL3A1	6.799957	8.560856	6.812114e-04	1.0000000
TMEM132B	-6.729445	8.371245	1.094067e-03	1.0000000
ANXA1	6.701346	8.484324	1.094067e-03	1.0000000

stage							
S1							
S2							
S3							
ISM.D							
ISM.D							
ADM.D							
ADM.D4	24	27	17	1	11	\	5

1
5

Shared genes between stages

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%

$$\text{Proportion} = \frac{intersect(genes\ in\ stage\ A\ and\ stage\ B)}{union(genes\ in\ stage\ A\ and\ stage\ B)}$$

1966 genes are shared by all 8 stages

Shared genes between stages

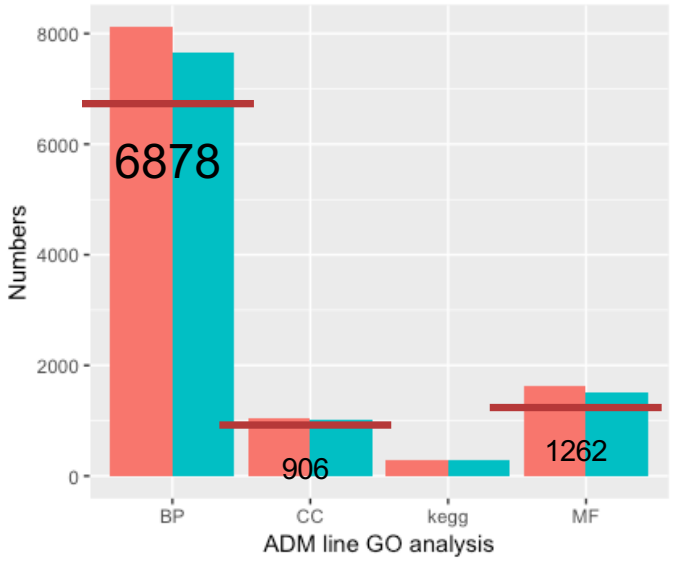
stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	46.99%						
S2	34.58%	38.56%			Genes with over 1 isoform		
S3	33.40%	38.62%	35.90%				
ISM.D0	33.36%	36.74%	38.39%	37.64%			
ISM.D4	31.98%	35.21%	37.07%	41.72%	39.74%		
ADM.D0	37.11%	45.31%	33.28%	37.94%	34.80%	34.42%	
ADM.D4	35.50%	42.57%	33.39%	40.12%	34.15%	38.05%	49.66%

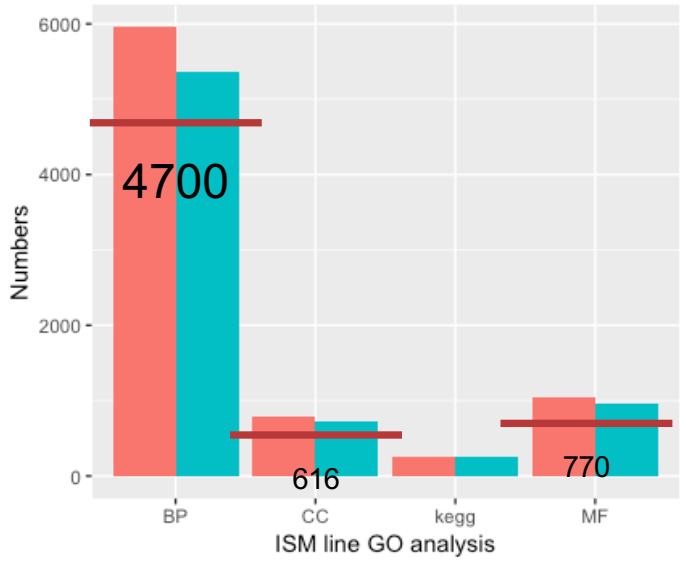
$$\text{Proportion} = \frac{intersect(genes\ in\ stage\ A\ and\ stage\ B)}{union(genes\ in\ stage\ A\ and\ stage\ B)}$$

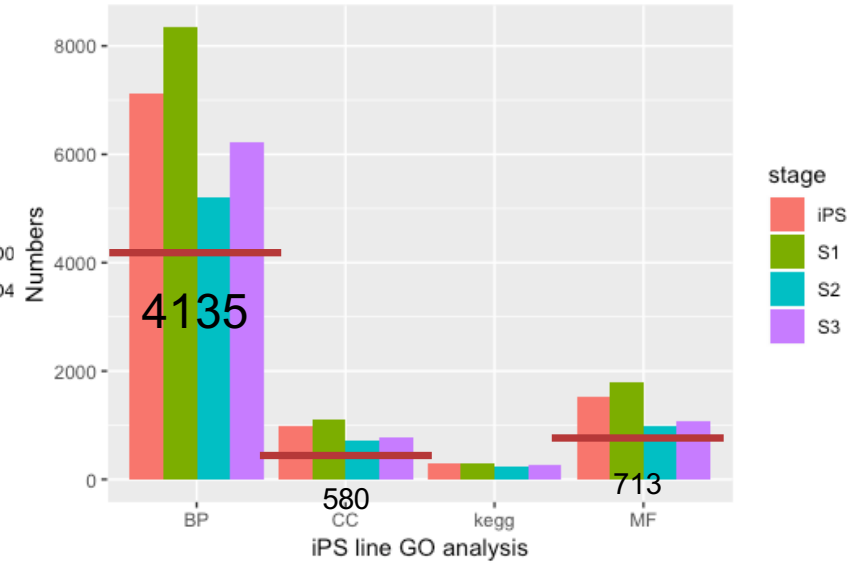
Shared genes between stages

stage	iPS	S1	S2	S3	ISM.D0	ISM.D4	ADM.D0
S1	32.36%						
S2	20.98%	19.80%			Genes with over 5 isoform		
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%

$$\text{Proportion} = \frac{intersect(genes\ in\ stage\ A\ and\ stage\ B)}{union(genes\ in\ stage\ A\ and\ stage\ B)}$$

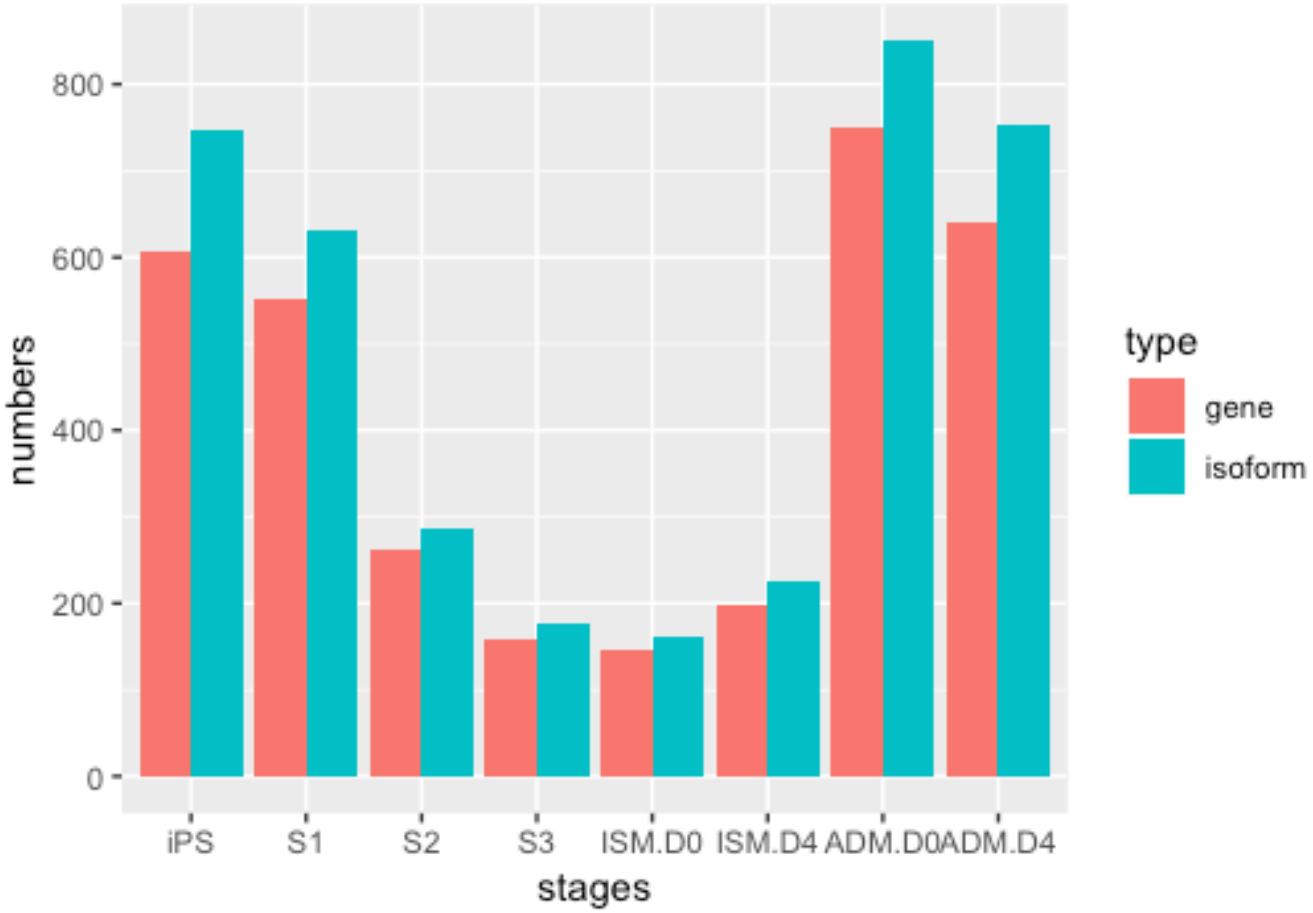




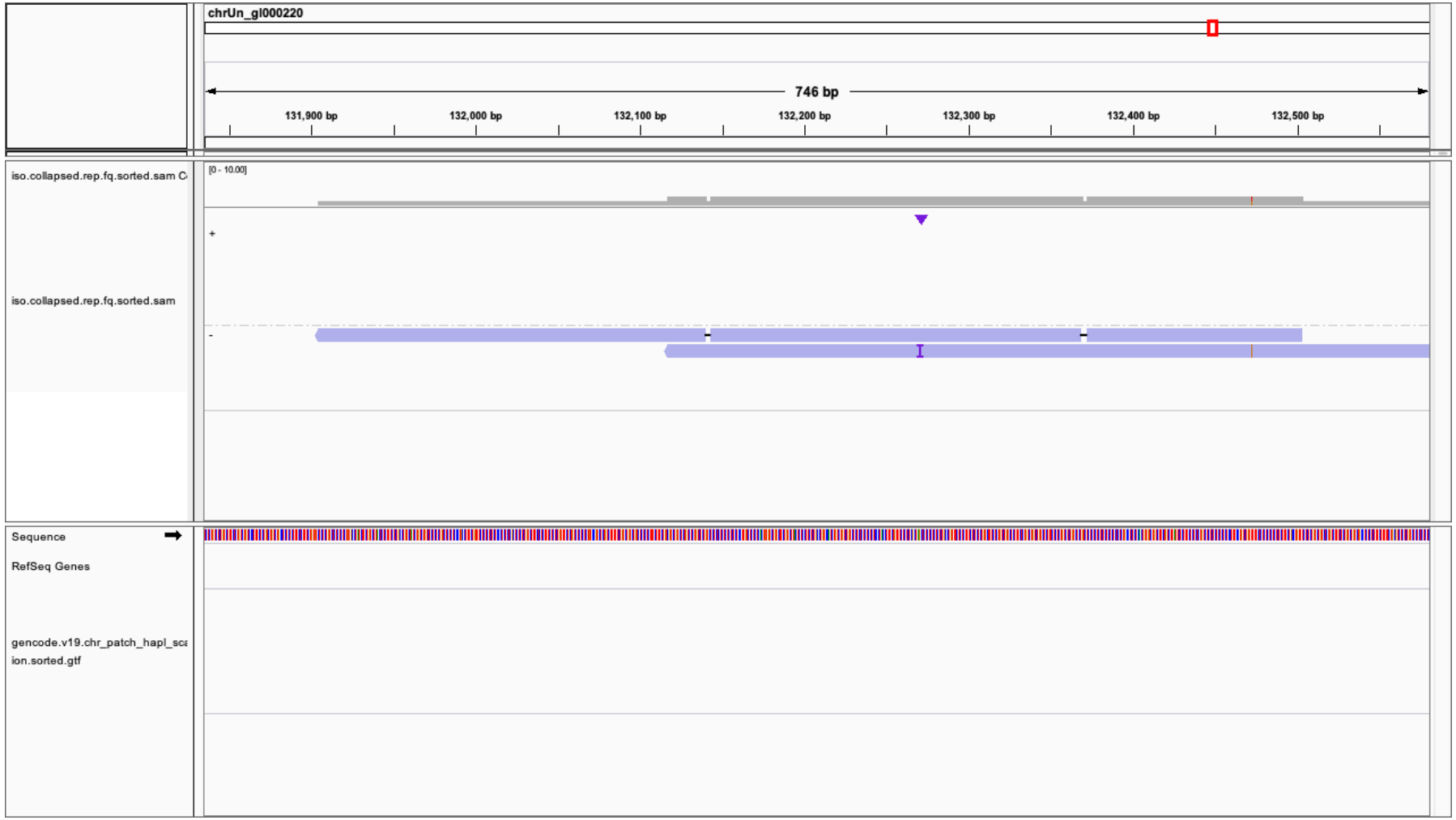


Functional analysis for active genes

Possible novel genes and isoforms



Possible novel genes and isoforms



c22873/f 1p0/596 | GL000220.1 : 132118 − 132868(−)
c22883/f 1p1/742 | GL000220.1 : 132118 − 132868(−)

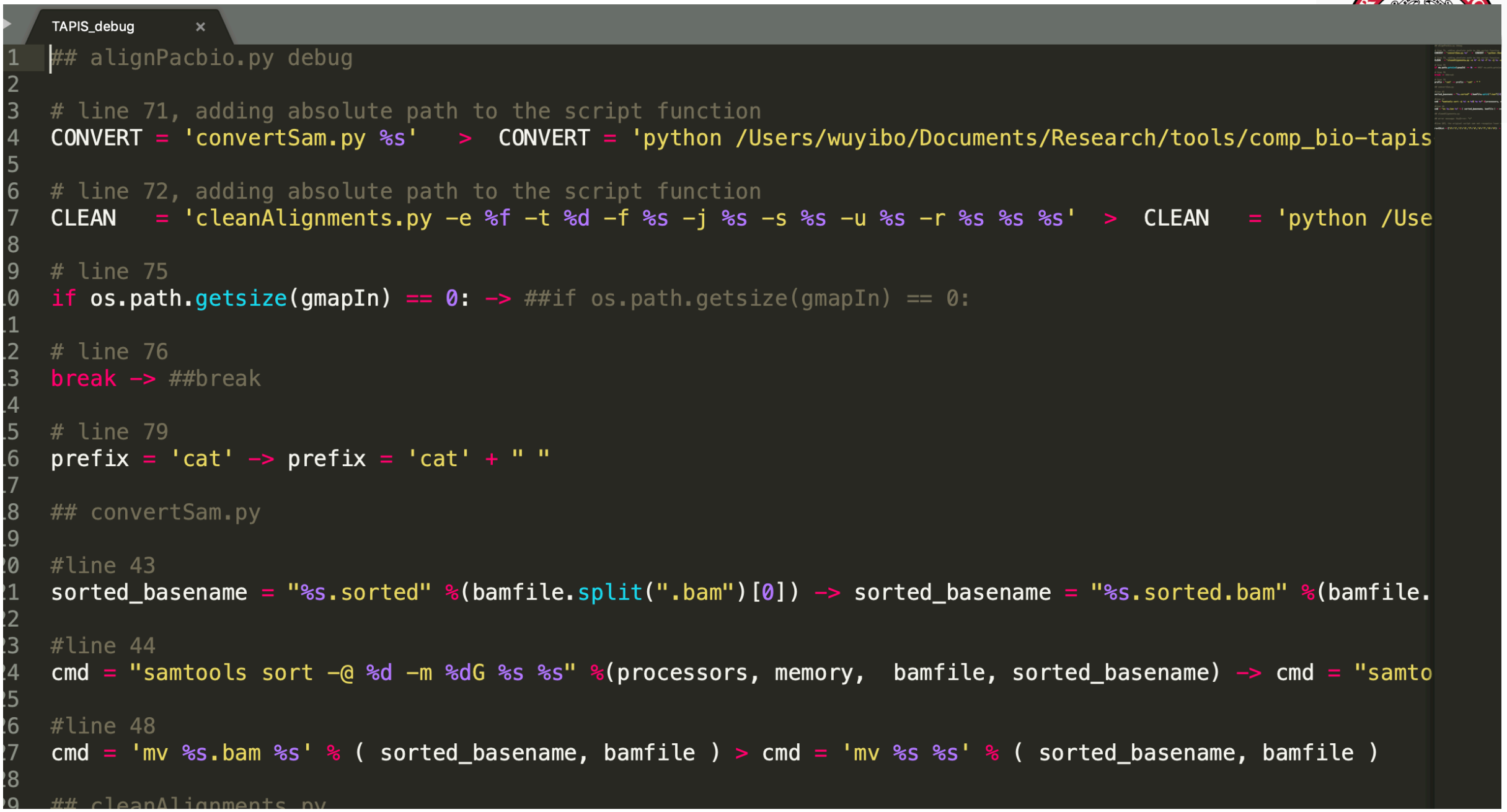
Part 03
Discussion

Discussion

- Accurate
- More information
- Lacking replica
- Lacking patient sample
- Lacking resource for Iso-Seq data analysis
- Further experiment to prove the hypothesis

Discussion

software	problems
IDP	Require short reads
SpliceGrapher	Require short reads
SQANTI	Can not even work on their tutorial data
TAPIS	Too many coding errors Alter the strand information of sequences


TAPIS_debug
alignPacbio.py debug
# line 71, adding absolute path to the script function
CONVERT = 'convertSam.py %s' > CONVERT = 'python /Users/wuyibo/Documents/Research/tools/comp_bio-tapis
# line 72, adding absolute path to the script function
CLEAN = 'cleanAlignments.py -e %f -t %d -f %s -j %s -s %s -u %s -r %s %s %s' > CLEAN = 'python /Use
# line 75
if os.path.getsize(gmapIn) == 0: -> ##if os.path.getsize(gmapIn) == 0:
# line 76
break -> ##break
# line 79
prefix = 'cat' -> prefix = 'cat' + " "
convertSam.py
#line 43
sorted_basename = "%s.sorted" %(bamfile.split(".bam")[0]) -> sorted_basename = "%s.sorted.bam" %(bamfile.
#line 44
cmd = "samtools sort -@ %d -m %dG %s %s" %(processors, memory, bamfile, sorted_basename) -> cmd = "samto
#line 48
cmd = 'mv %s.bam %s' % (sorted_basename, bamfile) > cmd = 'mv %s %s' % (sorted_basename, bamfile)
cleanAlignments.py

Hello,

When running alignPacBio.py as tutorial recommended, there are some fatal error. And I listed them as follows.

Traceback (most recent call last): File "./software/python2/bin/cleanAlignments.py", line 81, in <module> bamfile = pysam.Samfile(args.bam_input , 'rb') File "pysam/libcalignmentfile.pyx", line 444, in ysam.libcalignmentfile.AlignmentFile.**cinit** File "pysam/libcalignmentfile.pyx", line 664, in pysam.libcalignmentfile.AlignmentFile._open ValueError: file has no sequences defined (mode='rb') - is it SAM/BAM format? Consider opening with check_sq=False Traceback (most recent call last): File "./software/python2/bin/alignPacBio.py", line 75, in <module> if os.path.getsize(gmapIn) == 0: File "./software/python2/lib/python2.7/genericpath.py", line 57, in getsize return os.stat(filename).st_size OSError: [Errno 2] No such file or directory: './temp_dir/Homo_sapiens-GM12878_fixed_r1.fa'

Besides, I wonder can fastq file can be the input of alignPacBio.py or only fasta file can be accepted.

I appreciate any reply for this message

Comments (1)

YW Yibo Wu

Hi Xiaoyu Zhan,

I met this problem too, this problem is caused by some commands in the convertSam.py file, please try:

line 43: sorted_basename = "%s.sorted" %(bamfile.split(".bam")[0])

change to: sorted_basename = "%s.sorted.bam" %(bamfile.split(".bam")[0])

line 44: cmd = "samtools sort -@ %d -m %dG %s %s" %(processors, memory, bamfile, sorted_basename)

change to: cmd = "samtools sort -@ %d -m %dG %s > %s" %(processors, memory, bamfile, sorted_basename)

line 48: cmd = 'mv %s.bam %s' % (sorted_basename, bamfile)

change to: cmd = 'mv %s %s' % (sorted_basename, bamfile)

I think this would work.

Good luck!

Yibo

Edit • Delete • 2019-01-10

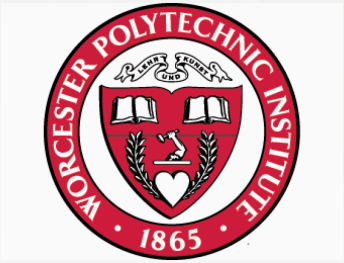
Assignee
Type
Priority
Status
Votes
Watchers

Acknowledgement

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 Dys- trophy 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.



Thanks!

Speaker: Yibo Wu