REPORT

weSEQ.iT

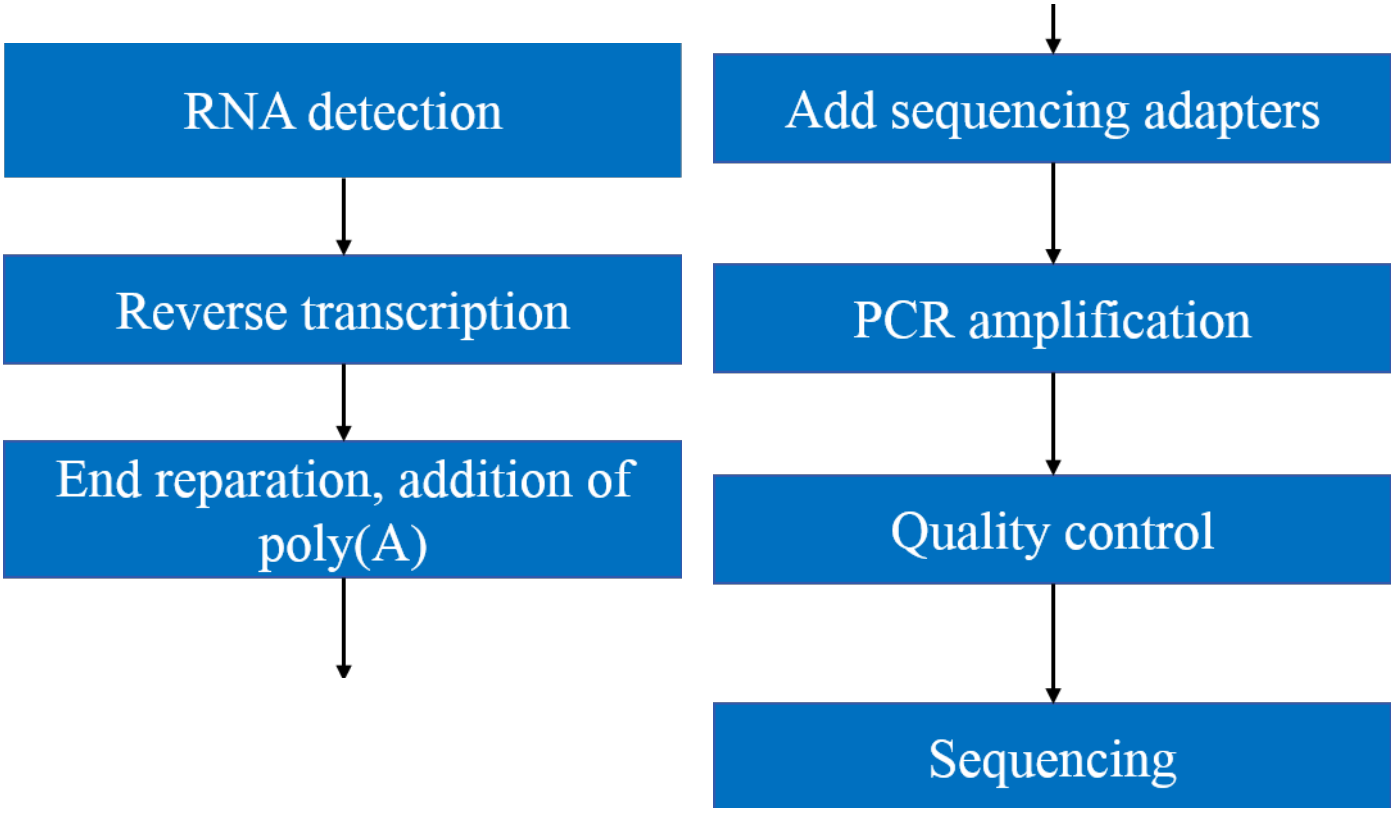
PROJECT INFORMATION

Client:
Institute:
Project: Library Preparation & mRNA Sequencing
NGS Data: Illumina NovaSeq 6000; PE150
Bioinformatics Service: yes
Number of Samples: DEMO
Date:

Results

1. Experimental workflow

A total of 6 were processed for transcriptome sequencing, generating 46.33Gb Clean Data. At least 7.53Gb clean data were generated for each sample with minimum 94.62% of clean data achieved quality score of Q30. Clean reads of each sample were mapped to specified reference genome. Mapping ratio ranged from 96.39% to 97.51%. Prediction of alternative splicing, gene structure optimization analysis and novel gene discovery was processed on top of mapping results, during which 1,345 were discovered and 519 novel genes were annotated with a putative function. The expression of genes was quantified and differentially expressed genes were identified based on their expression. These DEGs were further processed for functional annotation and enrichment analysis.
As shown in the following figure, the workflow of mRNA sequencing includes sample preparation, library construction, library quality control and sequencing.



2. RNA Quality Assessment

Purity, concentration and integrity of RNA sample were examined by NanoDrop, Qubit 2.0, Agilent 2100, etc. Only RNA with good quality could move on to following procedures.

2.1 Library Construction

Qualified RNA were processed for library construction. The procedures are described as follow:

(1) mRNA was isolated by Oligo(dT)-attached magnetic beads.

(2) mRNA was then randomly fragmented in fragmentation buffer.

(3) First-strand cDNA was synthesized with fragmented mRNA as template and random hexamers as primers, followed by second-strand synthesis with addition of PCR buffer, dNTPs, RNase H and DNA polymerase I. Purification of cDNA was processed with AMPure XP beads.

(4) Double-strand cDNA was subjected to end repair. Adenosine was added to the end and ligated to adapters. AMPure XP beads were applied here to select fragments within size range of 300-400 bp.

(5) cDNA library was obtained by certain rounds of PCR on cDNA fragments generated from step

2.2 Library Quality Control

In order to ensure the quality of library, Qubit 2.0 and Agilent 2100 were used to examine the concentration of cDNA and insert size. Q-PCR was processed to obtain a more accurate library concentration. Library with concentration larger than 2 nM is acceptable.

2.3 Sequencing

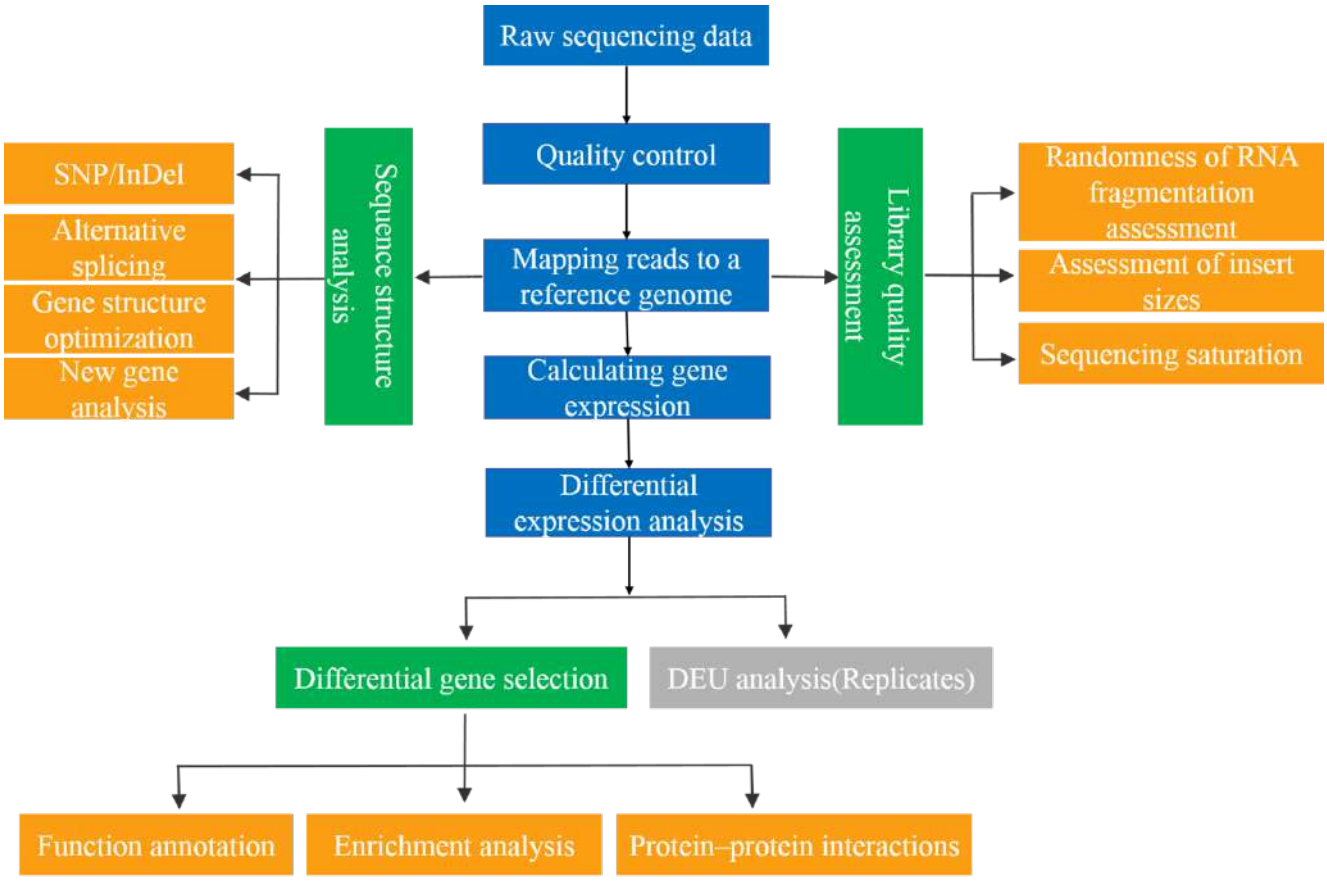
The qualified library was pooled based on pre-designed target data volume and then sequenced on Illumina sequencing platform.

3. Bioinformatics Analysis

3.1 Summary of Bioinformatics Analysis

Clean data with high quality was obtained by filtering Raw data, which removes adapter sequence and reads with low quality. These clean data were further mapped to pre-defined reference genome generating mapped data. Assessment on insert size and sequencing randomness were processed on mapped data as library quality control. Basic analysis on mapped data included gene expression quantification, alternative splicing analysis, novel genes prediction and genes structure optimization.

RNA sequencing bioinformatics pipeline was shown below:



3.2 Data Quality Control

Based on sequencing-by-synthesis (Sequencing By Synthesis, SBS) technology, cDNA libraries were sequenced on Illumina high-throughput platform, generating significant amounts of high-quality data known as raw data. Raw data was saved in FASTQ format. Each sample has two FASTQ file, containing cDNA reads measured at both ends respectively.

Following figure shows a demo FASTQ file.

```
@HWI-7001455:172:H98E5ADXX:1:1105:13734:97199 1:N:0:TAAGGC
CTGACATACGATGTTGTTGACGGGTCATCATCTTCTTCAACGGTCCCAGTCTTGAAACCAACAATGTCAATAGGACAAATACTGAGCAAGAAGCTACTTCT
+
BBCFFFFFHGFFHHIIJIJJJJJJJHIHIJJIJJJJJJJJIJJJJJJJIIJIIIIIIJIIJIGHHHFDFFFFFEECEDDDDDDDDCDDDDDDDCDDDDDDDD
@HWI-7001455:172:H98E5ADXX:1:1105:13927:97189 1:N:0:TAAGGC
GGGACTTCTCCAGTGGAGTCGTAGCAGCTTAATCCCGTCGATTTATACGCCAATCAACAAAACCCAATTCTCTATCGCCGCCTGTGTCATCGAGCGGAACC
+
CCCFFFFFFHHHHHFGIIJGHHHIIIJJJJJJJJJJJJJJJJIJIJJJJJIJJJJJJJJJIJHHGHFFFFEEEEEEDDDDDDDDDDDDDDDEDDDDDDDDDD
@HWI-7001455:172:H98E5ADXX:1:1105:14131:97087 1:N:0:TAAGGC
TGTCACTTCATCGATGACGTCCTTAGAGAGGGTGCTCAAGAATGATACAACAGATGAAGAAAATGCATCAAGCCGTAAAAGAAAAGATGTCATTGTAAGAG
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJIJJJIIJJIJJJJJJJJJGJJJGIIHIIJJJJIIJJJJJIJIHGHHHFEDDDEEDDDDDDDDCDEFEDCDEDDD
@HWI-7001455:172:H98E5ADXX:1:1105:14089:97119 1:N:0:TAAGGC
ATTTGATCCCTTCCTCTACCGTCTTGAATTTCATCAGTGACATGACTGGTCCAAAGATTTCATCTTGGTATATCTTCATATCCTCAGTGACATCTGCGAAT
+
CCCFFFFFHHHHHJJJJJJJJIJJJJJJJJJJJJJJIJJJJJJJJJJJJJJJJJJJHIJJJGIJJJJJJFHIIJJJJJIJJHHHHHHFFFFFFFEEEEDBDD
@HWI-7001455:172:H98E5ADXX:1:1105:14060:97189 1:N:0:TAAGGC
TGATAGCAAAGCCGGCTCTGTCTTCACTGCTCTGGGATCTGATGGAGGGGTGAAGCTATACTGGGGCATAGCTGCTGATGGATCTGTTGTTATATCAGATG
+
FFFFFFFFHGHHHJJJJJJJJJJIJGIJJJJIIIJGHIGHIAFGEFHIJJAGHHFHHFHDFFFFFBDDCDDDDDDDDDDDDCCCCDCDDCDDEEEDEEDDD
```

Figure. Fastq format

Note: A FASTQ file normally contains four lines:
The first line begins with @ and is followed by sequence ID and an optional description.
The second line is a series of single letters representing sequence.
The third line begins with + and optional description.
The last line is the corresponding quality value of the bases in the second line. The length of this line should be exactly the same as Line 2. Base quality score is calculated as ASCII-33.

3.2.1 Sequencing bases quality score

Quality Score or Q-score represents the probability of an incorrect base. This Phred quality score is defined as following equation [1]:

$$Q = -10 * log_{10}P$$

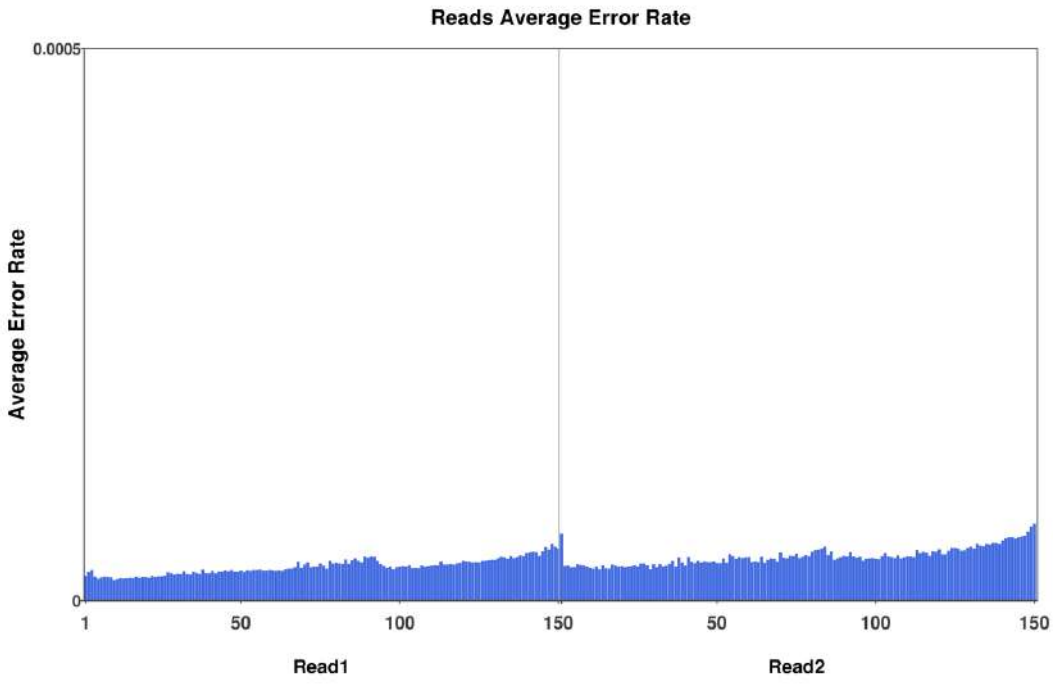
In the equation, P stands for the base calling error probabilities. Following table shows the relations between quality score and base calling accuracy:

Table. Quality score and base calling accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accurancy
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%

Base call with higher Q-scores are believed to be more reliable and less likely to be error base. For example, Q20 is equivalent to the probability of one incorrect base call in 100 times.

Distribution of error rate along reads were shown in the following figures.

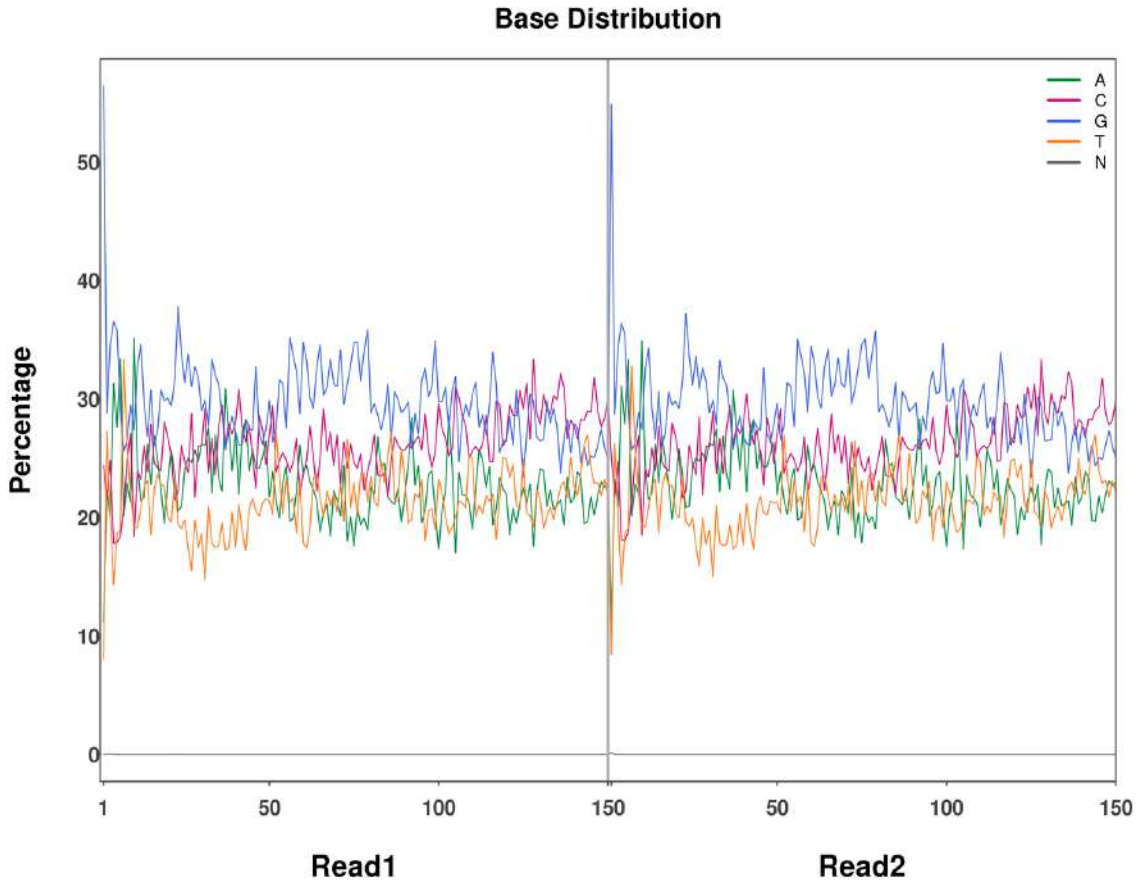


Note: X-axis: Position on the reads. Y-axis: Average error rate on corresponding position.

Rate of error basing calling is influenced by the instrument, reagents, samples, etc. It is commonly found in Illumina platform that the rate slowly climbs along the reading of sequence due to the consumption of reagents. The high error rate at first six bases of reads are normally caused by inefficient binding between random hexamer primers and RNA templates.

3.2.2 Nucleotide Distribution on Reads

Nucleotide distribution test is designed to detect separation of AT and GC. In theory, according to the cDNA random fragmentation process and complementary base pairing principle, the frequency of A, T, G, C should be the same and steady along the reads. However, practically, fluctuations at 5'-end are commonly seen due to certain bias in binding of random haxemer primers and templates.



Note: X-axis: Position on reads. Y-axis: Percentage of certain nucleotide on corresponding position.

3.2.3 Sequencing quality control

It is crucial to ensure the quality of the reads before moving onto following analysis. Raw data contains useless data such as primers, adapters, etc., which need to be removed before analysis. Procedures for data quality control were listed as follow:
(1) Trim adapter contaminations

(2) Remove nucleotides with low Quality-score.

Data processed by above steps is named "Clean data". Clean data was provided in FASTQ format.

3.2.4 Sequencing data statistics

Statistics of sequencing data was provided in the following table.

Table. Sequencing data Statistics

Samples	Clean reads	Clean bases	GC Content	%≥Q30
N1	25,506,178	7,595,723,240	55.90%	95.11%
N2	26,279,690	7,830,705,610	55.79%	95.23%
N3	25,218,061	7,527,096,708	55.63%	94.86%
T1	25,823,411	7,698,882,934	55.15%	94.75%

Note:
(1)Samples: Sample name;
(2)Clean reads: Counts of clean PE reads;
(3)Clean bases: total base number of Clean Data;
(4)GC content: Percentage of G,C in clean data.
(5)≥Q30%: Percentage of bases with Q-score no less than Q30.

After quality control of sequencing data, 46.33Gb Clean Data were obtained and and more than 94.62% of bases in each sample had a Q-score no less than Q30.

3.3 Data alignment to reference genome

Reference genome was pre-defined for the analysis. The download address is: http://asia.ensembl.org/Mus_musculus/Info/Index.
HISAT2 [2] is a highly efficient system for mapping RNA-seq reads, which is a more advanced version of TopHat2/Bowtie2.HISAT2 uses a Burrows-Wheeler Transform and Ferragina-Manzini（FM）index based search. HISAT2 uses one global graph FM index (GFM) to represent general population, as well as small indexes (local indexes) combined with several alignment strategies in order to achieve more efficient alignment.
StringTie [3] was applied to assemble the mapped reads. The algorithm is established based on optimality theory. It utilizes a novel network flow algorithm as well as an optional de novo assembly step to assemble and quantify transcripts representing multiple spliced variants for each gene locus.
The workflow of analysis was shown in the figure below.

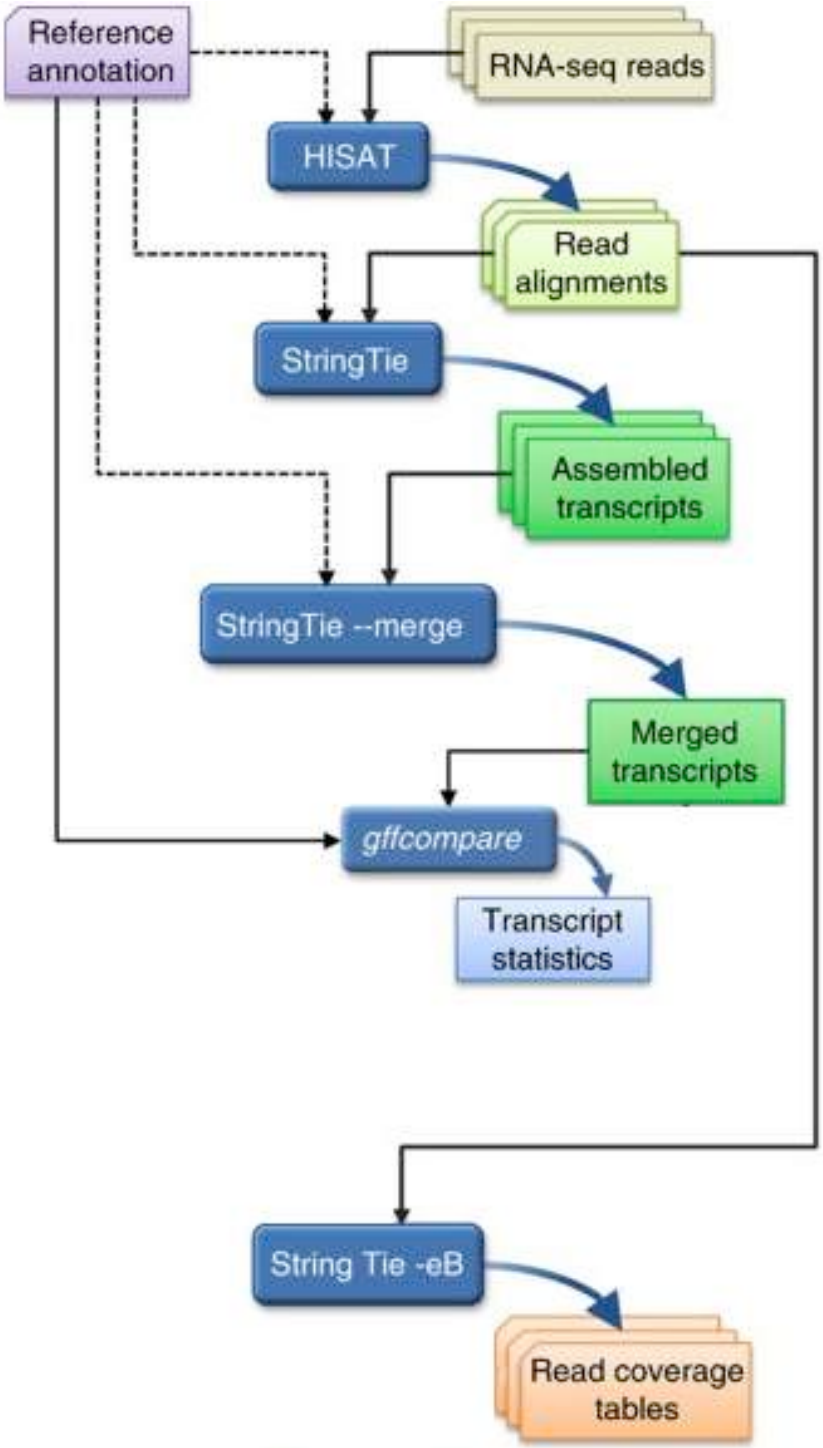


Figure. Schematic flow of HISAT2

3.3.1 Mapping Statistics

Mapping ratio refers to the percentage of Mapped Reads in Clean Reads, which indicates the utilization of RNA data. Besides the influence of sequencing data quality, mapping ratio is also affected by the quality of reference genome assembly, biological classification relation between sequenced sample and reference subspecies. Mapping ratio is an important parameter to examine if reference genome is suitable for following bioinformatic analysis.

Statistics on sequencing data yield for each sample is shown in the following table

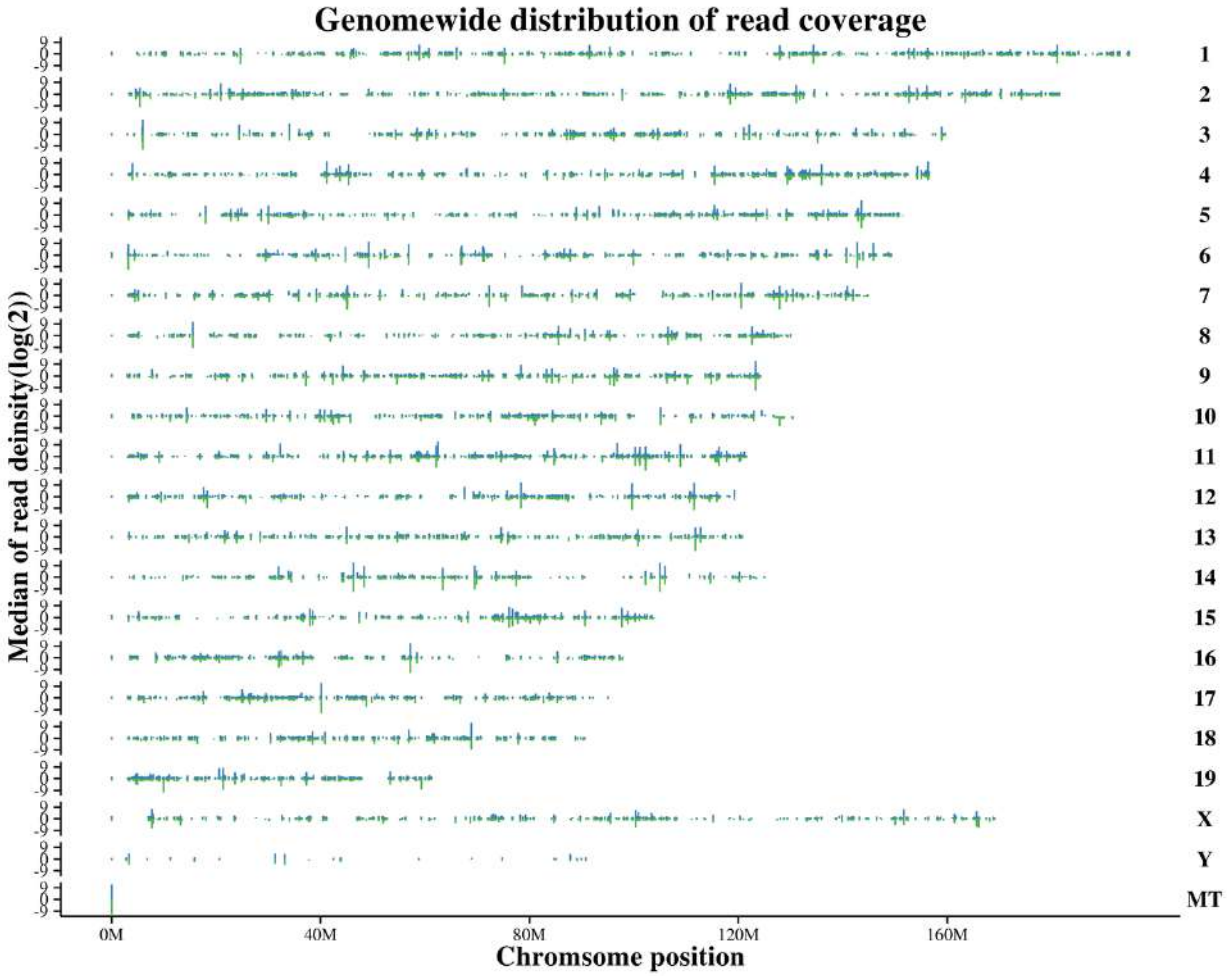
Sample	Total Reads	Mapped Reads	Uniq Mapped Reads	Multiple Map Reads	Reads Map to '+'	Reads Map to '-'
N1	51,012,356	49,626,531 (97.28%)	22,543,577 (44.19%)	27,082,954 (53.09%)	53,107,106 (104.11%)	53,194,824 (104.28%)
N2	52,559,380	51,009,386 (97.05%)	21,836,396 (41.55%)	29,172,990 (55.50%)	57,933,979 (110.23%)	57,968,149 (110.29%)
N3	50,436,122	48,997,675 (97.15%)	22,730,355 (45.07%)	26,267,320 (52.08%)	52,657,620 (104.40%)	52,644,752 (104.38%)
T1	51,646,822	49,780,793 (96.39%)	23,899,221 (46.27%)	25,881,572 (50.11%)	55,468,537 (107.40%)	55,548,200 (107.55%)
T2	52,386,734	50,839,244 (97.05%)	34,627,422 (66.10%)	16,211,822 (30.95%)	40,610,272 (77.52%)	40,588,815 (77.48%)
T3	52,500,208	51,191,870 (97.51%)	24,053,761 (45.82%)	27,138,109 (51.69%)	57,077,040 (108.72%)	57,069,675 (108.70%)

Note: Sample: sample ID in system;
Total Reads: Counts of Clean Reads, counted as single end;
Mapped Reads: Counts of mapped reads and the proportion of that in clean data;
Uniq Mapped Reads: Counts of reads mapped to a unique position on reference genome and proportion of that in clean data;
Multiple Mapped Reads: Counts of reads mapped to multiple positions on reference genome and proportion of that in clean data;
Reads Map to '+': Counts of reads mapped to the sense chain and the proportion of that in clean data;
Reads Map to '-': Counts of reads mapped to antisense chain and proportion of that in clean data.

The mapping ratio of each sample against reference genome ranged from 96.39% to 97.51%.

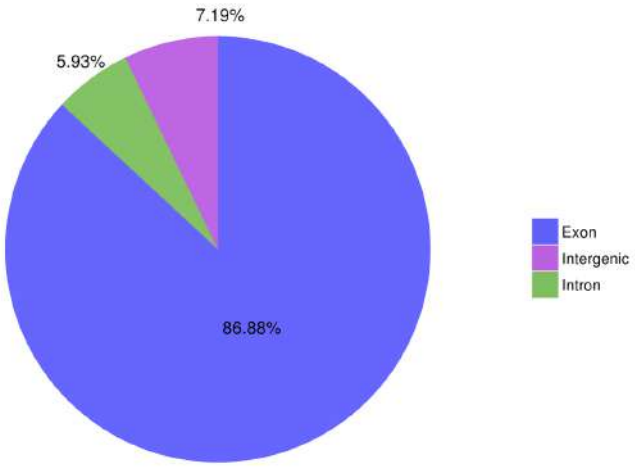
3.3.2 Summary on Mapping

Distribution of coverage depth on reference genome was plotted based on the position of each mapped read on different chromosomes.



Note: X-axis: Position on chromosome; Y-axis: Log2 of coverage depth (coverage depth was defined as reads counted within a chromosome window of 10 kb in length); Blue represents + strand and green represents - strand.

By summarizing the number of reads mapped to different regions of genes on reference genome, i.e. exons, introns and intergenic regions, the distribution pie chart of mapped reads on different gene regions were generated, which was shown below.



Note: Genome was divided into exon, intron and intergenic regions, which are colored differently. The size of each area indicates the proportion of that in total mapped reads.

In theory, reads originated from mature mRNA should be aligned to exons. The ones mapped to introns may come from RNA precursor or intron retention(alternative splicing events). The ones aligned to intergenic regions may due to imperfect annotation of genome.

3.3.3 Visualization of Mapping

Integrative Genomics Viewer (IGV) is recommended for visualization of mapping output (BAM format) and annotation file of reference genome. Following information can be obtained through IGV.

(1) It is able to present positions of one or more reads on reference genome at different scales, including reads distribution on each chromosome and that on exons, introns, splice junction regions, intergenic regions ,etc.

(2) It is able to present the abundance of reads on different regions at difference scales, which indicates transcription level of each region.

(3) It is able to present annotation of genes and splicing isoforms

(4) It is able to present other annotation information.

(5) It is able to download annotation information either from a remote server or load that from local.

Figure. Demo of IGV browser interface

3.4 Library Quality Control

It is crucial to assure a library of good quality in order to obtain a better output of RNA sequencing. To ensure the quality of library, following three quality examinations were performed on RNA library.

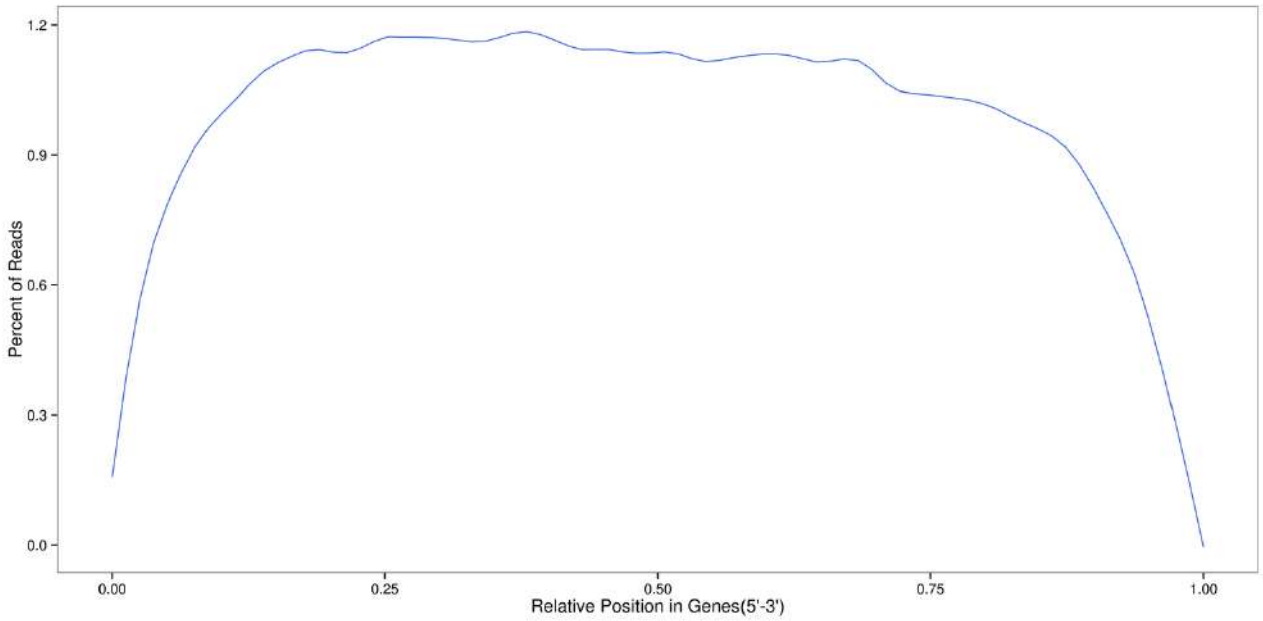
(1) Fragment randomness and degradation of RNA sample was estimated by checking the distribution of mapped reads on genome

(2) Length dispersion was examined by the length distribution of inserts

(3) Sufficiency of library Volume (or mapped reads) was examined by generating saturation curve between sampled mapped reads against genes identified within certain expression accuracy.

3.4.1 mRNA Fragmentation Randomness Check

Ideally, we expect the reads generated by sequencing to cover mRNA evenly, which strongly counts on a higher randomness mRNA fragmentation. The randomness mRNA fragmentation is largely guaranteed by a sufficient amount of sample, proper method and time on fragmentation, etc.
The randomness of mRNA fragmentation is examined by the distribution of mapped reads on each mRNA. The bases on the mRNA won't be read during sequencing if the mRNA is heavily degraded, i.e. there will be no reads mapped to the region. Therefore, the degradation of RNA sample can also be checked by distribution of mapped reads on transcripts. Following figure shown the distribution of mapped reads on transcripts.



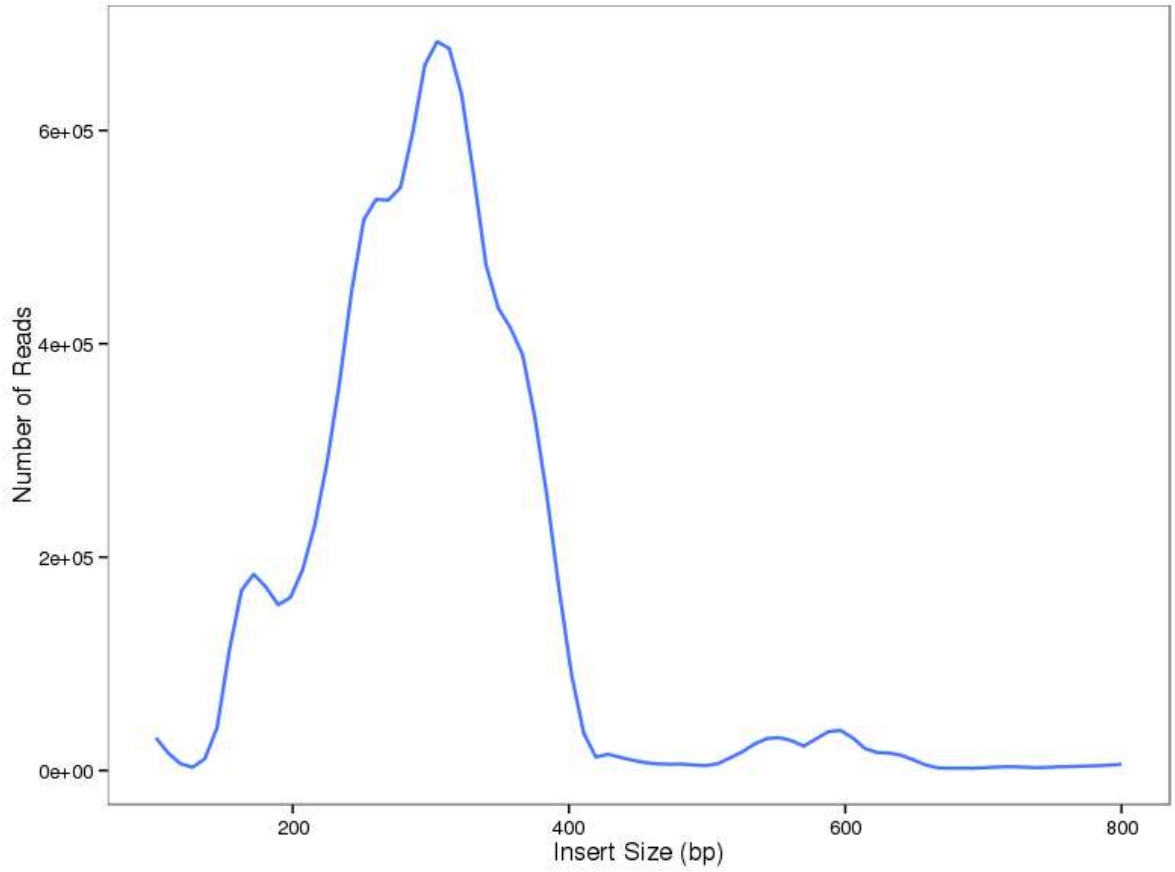
Note: X-axis: Normalized mRNA position; Y-axis: Percentage of reads mapped to corresponding region in total mapped reads. Since the length of mRNAs differs from each other, all mRNAs were divided into 100 parts in order to count the mapped reads in each part. The figure shows the sum of the percentage on all mRNAs.

3.4.2 Length Distribution of Inserts

The dispersion of insert length is an important parameter representing the quality of library construction, especially in purification by magnetic beads. The size of inserts were counted as the distance between the start and end point on reference genome in paired-end reads mapping.

In majority of eukaryotic genome, the DNA coding region is not continuous, i.e. the exons are divided by introns. However, in RNA sequencing, mature mRNA without introns is sequenced. In this case, when the reads cover the region cross introns, the distance between start and end point of reads on reference genome will be larger than the insert size. Therefore, these reads may form several small peaks in the distribution curve.

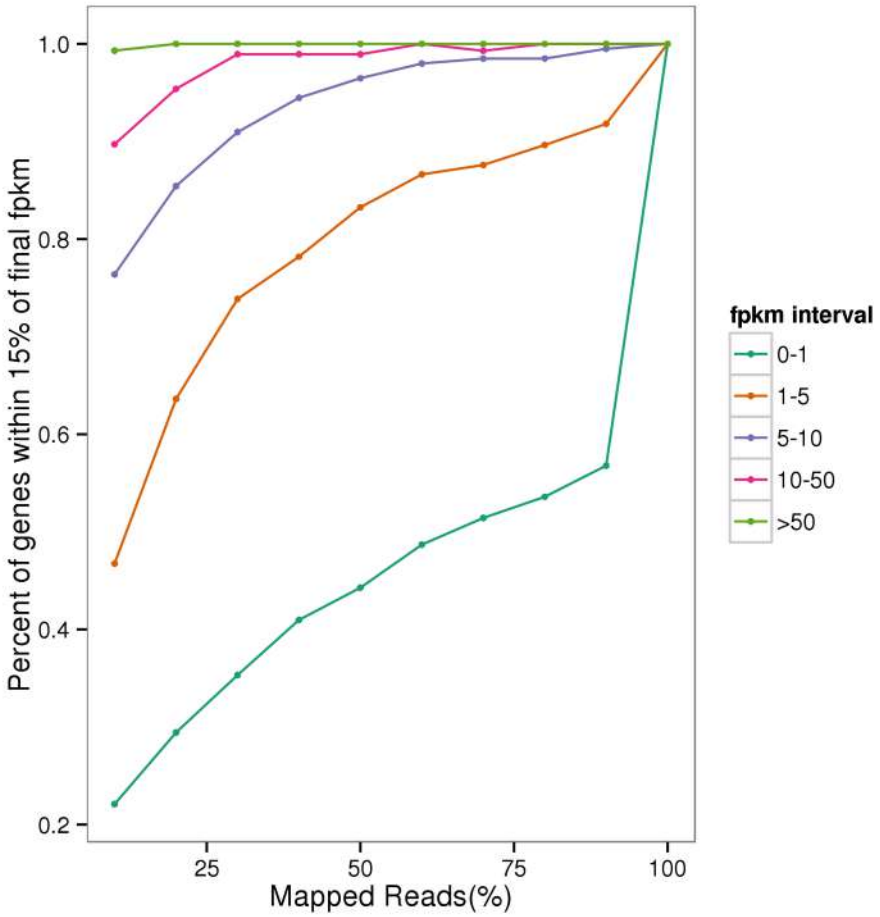
Insert length distribution of each sample was shown in the figures below.



Note: X-axis: Insert size (bp), which stands for the distance between the start and end point on reference genome in paired-end reads mapping, ranging from 0 to 800 bp. Y-axis: Number of inserts.

3.4.3 Saturation Test on RNA Sequencing Data

In order to ensure the sufficiency of data volume, saturation in gene recognition against data volume needs to be checked. Since there are only limited number of genes in a species as well as limited transcripts at certain point, the number of genes recognized will gradually reach saturation along with the increase of data size. Genes with higher expression are more likely to be identified and quantified. Therefore, larger data volume is required to quantify low abundant genes.
The saturation of mapped data on genes at different expression level can be mimicked by checking the increase in number of genes identified with the increase in mapped data size. The saturation curve was shown below.



Note: Mapped reads were divided into 10 fractions. By counting the number of genes identified at different expression level along with the increase of reads, the saturation curve was generated. X-axis: Percentage of sampled reads in total mapped reads; Y-axis: Percentage of genes identified in total genes at different expression level with error within 15% fpkm. The lines closer to 1.0 on Y-axis indicates a more saturated situation. Lines with different colors represents genes with different expression level.

3.5 SNP/InDel analysis

SNP (Single Nucleotide Polymorphisms) is defined as a genetic marker formed by substitution of a single nucleotide on genome, which occurs quite frequently on a genome. The discovery of potential SNP sites is mainly relying on mapping of sequences obtained against reference genome. Here, GATK [4] was employed to identify the single-base mismatches as potential SNP site. Base on the position of SNPs, functional effects of the SNPs can be predicted, such as if an SNP can affect gene expression level or protein production.

InDel (insertion-deletion) is defined as insertion or deletion of bases on a genome comparing with reference genome. InDel can occur as one base or several bases. The identification of InDel was also processed by GATK. InDel is not as common as SNP, however, it can also lead to changes in gene functions, for example the InDels on coding region, which causes frame shift. The criterion for GATK to recognize SNP/InDel are listed below.

(1) Less than 3 continuous single nucleotide mismatch within 35 bp ;

(2) SNP Quality score generated by GATK is larger than 2.0.

Above criterion were applied in the analysis of all samples to harvest reliable SNP sites.

SnpEff [5] is a software designed for SNP/InDel annotation and functional effects prediction. Combining the location information of SNP on reference genome and corresponding locations of genes, the position of SNP on gene regions (intergenic region, gene region, CDS ,etc.) can be obtained as an important clue for functional effects (synonymous or non-synonymous mutation) prediction.

Since mRNA will go through several maturation processes including adding capping, adding Poly(A) tail, alternative splicing and some will also go through RNA editing, during which, single nucleotide substitution, insertion or deletion are produced. However, these polymorphisms are different from the inherent ones on genome, which can not be distinguished by mapping. Therefore, SNPs predicted in RNA_seq may contain those generated by RNA editing.

SNP/InDel Sites info

final.indel.anno.gatk.all.list

final.snp.anno.gatk.all.list

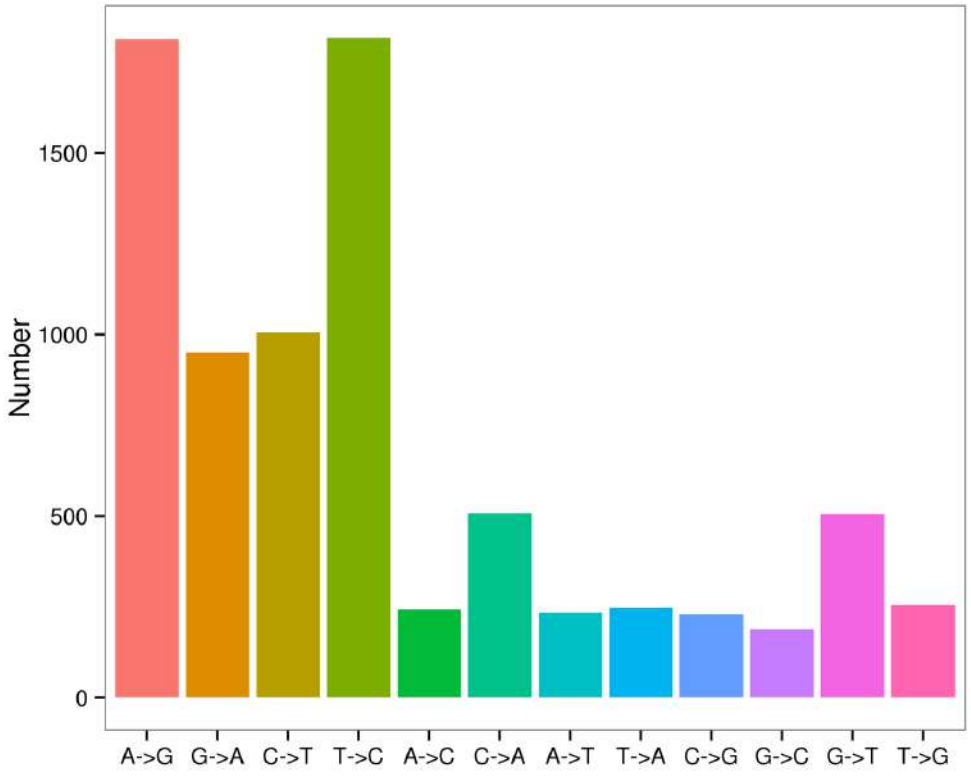
3.5.1 Statistics of SNP sites

SNP can be generally divided into two types: Transition and Transversion, based on nucleotide substitutions. Based on the allele number on SNP site, i.e. the number of different nucleotide on the site, SNPs can be divided into homozygous SNP and heterozygous SNP. The percentage of heterozygous SNP may differ from different species. Number of SNPs, substitution type and heterozygosity were summarized in the table below.

Sample	SNP Number	Genic SNP	Intergenic SNP	Transition	Transversion	Heterozygosity
N1	1,284	686	598	75.62%	24.38%	17.76%
N2	1,352	706	646	77.37%	22.63%	18.86%
N3	1,814	998	816	73.70%	26.30%	19.29%
T1	2,623	1,449	1,174	72.51%	27.49%	20.13%
T2	5,611	3,463	2,148	70.08%	29.92%	33.65%
T3	1,524	773	751	77.43%	22.57%	21.39%

Note: Sample: Sample ID in system;
SNP Number: Total number of SNP;
Genic SNP: Number of SNP in genic region;
Intergenic SNP: Number of SNP in intergenic SNP;
Transition: Percentage of transition SNP in total SNP;
Transversion: Percentage of transversion SNP in total SNP;
Heterozygosity: Percentage of heterozygous SNP in total SNP.

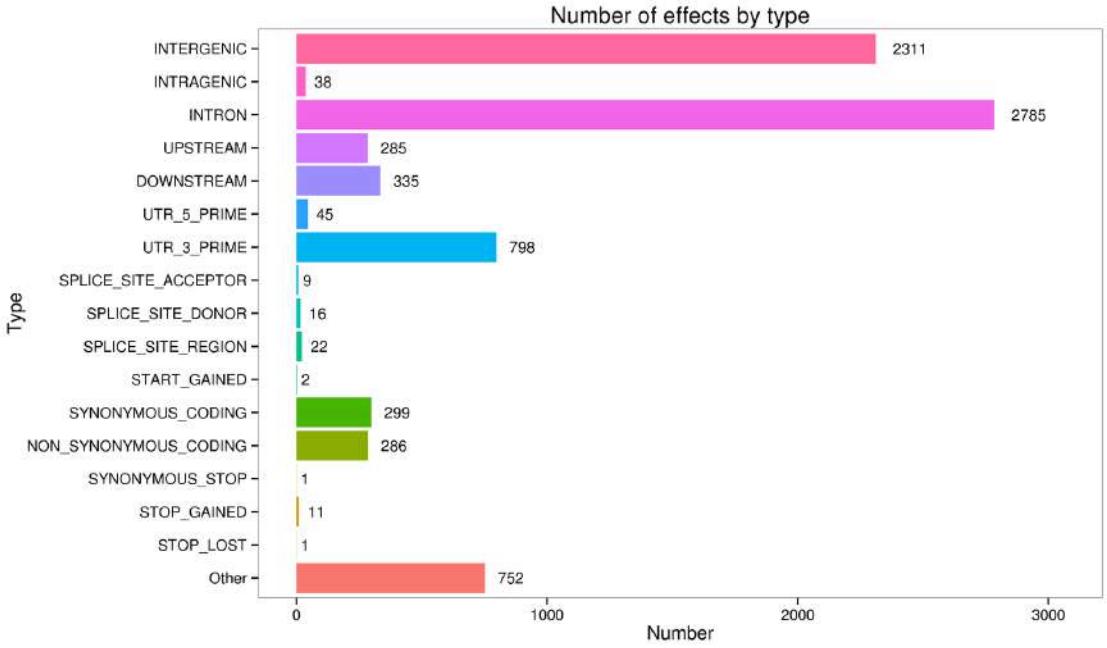
Statistics of SNP type was shown in the figure below.



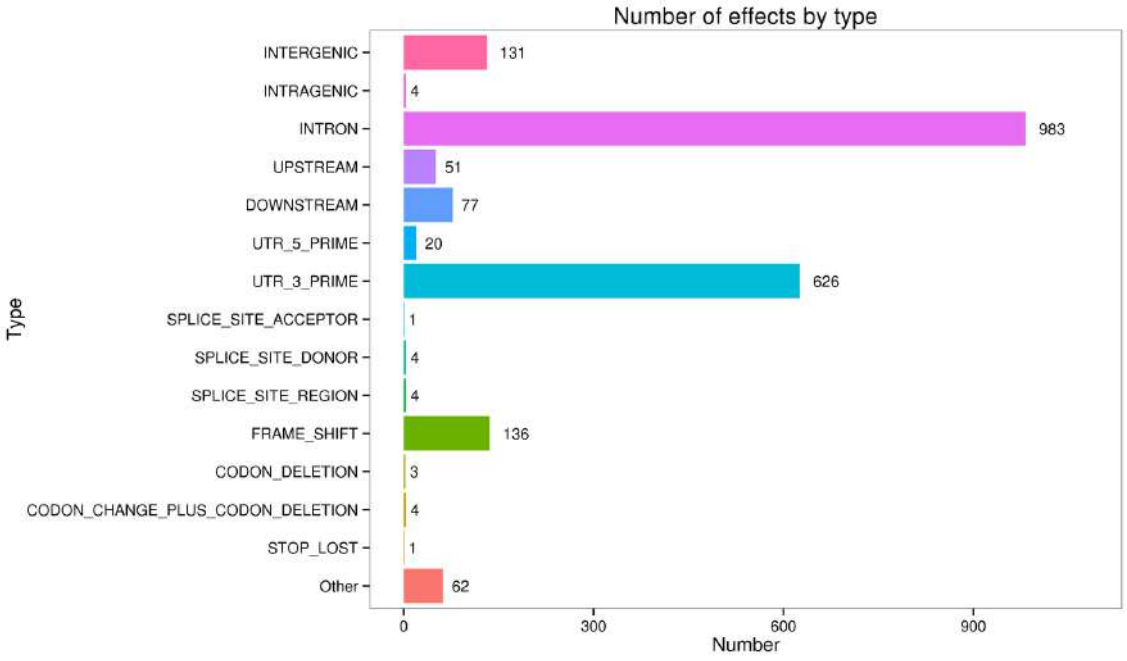
Note: X-axis: SNP type; Y-axis: Number of SNPs

3.5.2 Gene SNP Density Distribution

The density of gene SNP was defined as the ratio of SNP number on the gene over the gene length. Distribution of gene density was generated by counting SNP density of all genes.
Distribution of gene SNP density was shown in the figure below.



Note: Y-axis: Region and types of SNP; X-axis: Number of SNP.



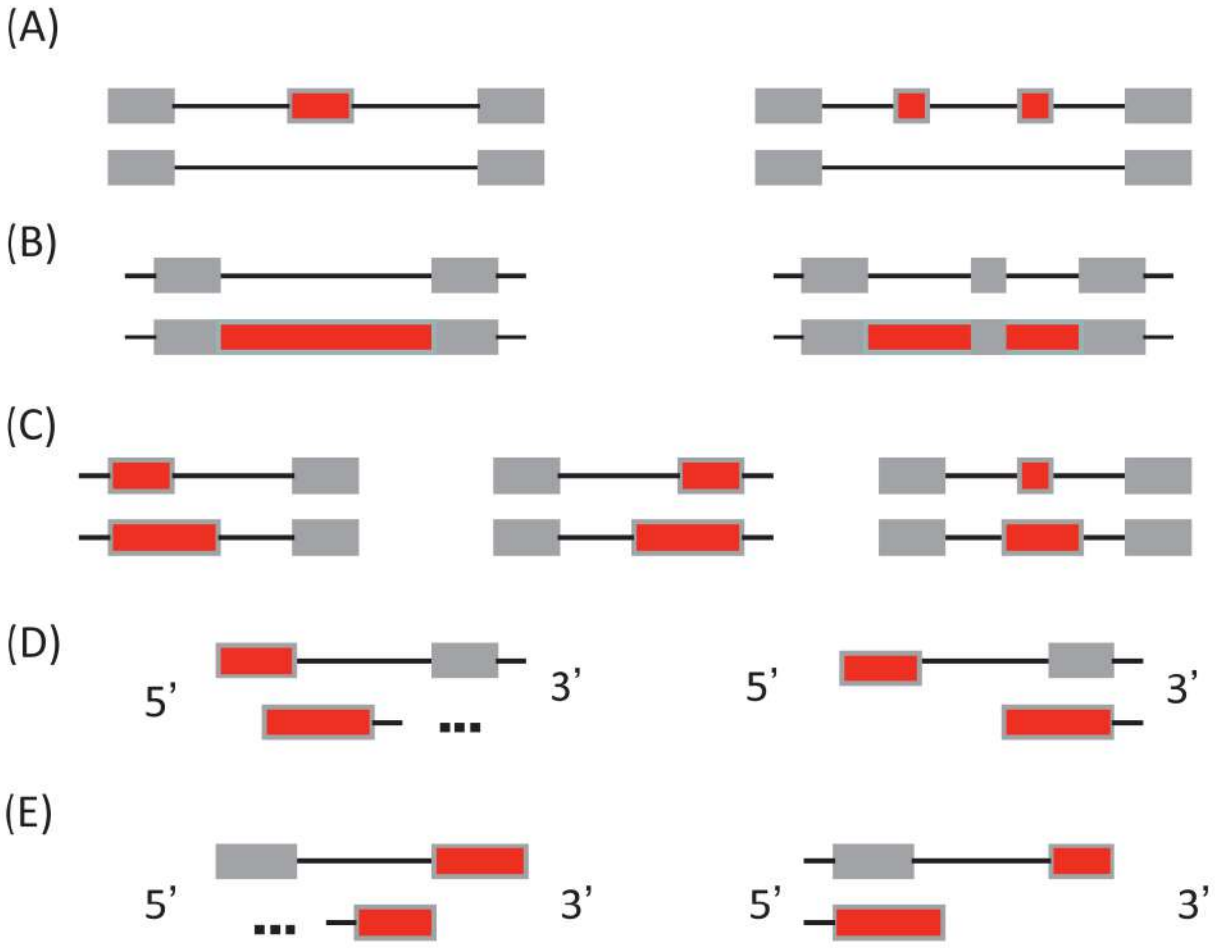
Note: Y-axis: Regions or types of InDel; X-axis: Number of InDel

3.6 Alternative splicing prediction

In gene expression processes, particular exons on pre-mRNA may be included in or excluded from the final, resulting in different versions of mature mRNA. In this case, multiple proteins with different structures and biological functions can be translated from these alternatively spliced mRNA originated from the same gene. The process described here is named alternative splicing.

StringTie [3] was applied to assemble the mapped reads generated by Hisat2. ASprofile [5] was employed to predict alternative splicing events in each samples and sorting them into 12 types. Typical alternative splicing scheme were shown in the figure below.

Figure. Typical alternative splicing event



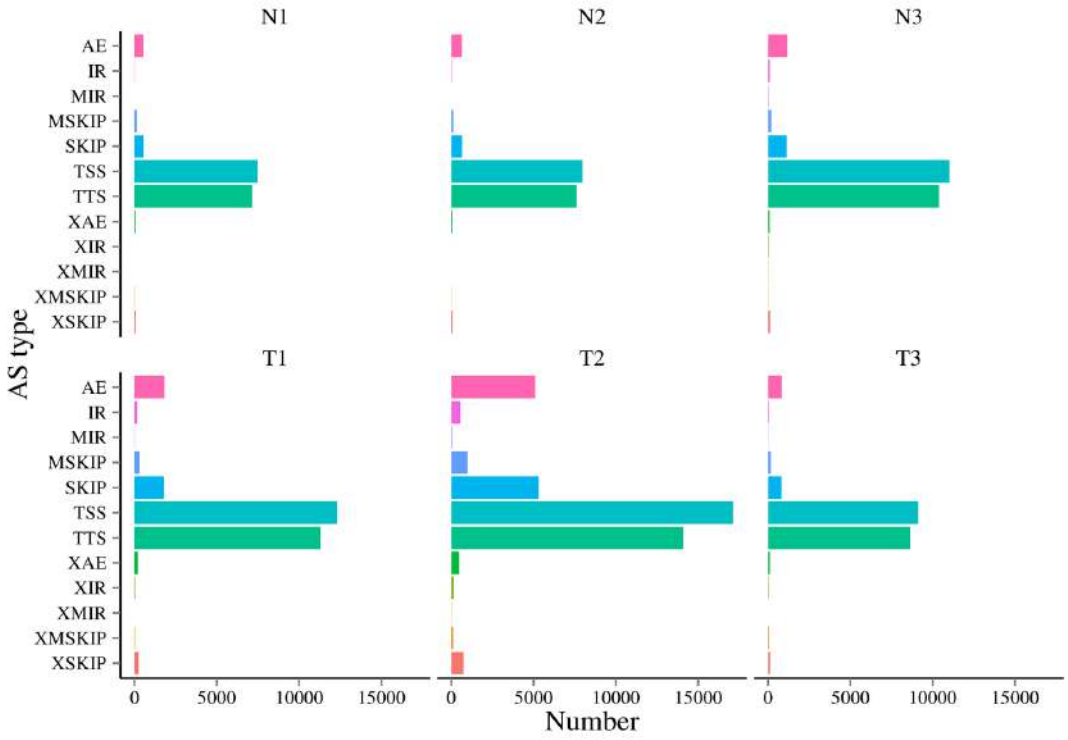
Note: (A) Skipped exon and Multi-exon SKIP; (B) Intron retention and Multi-Intron retention; (C) Alternative exon; (D) transcription start site; (E) transcription terminal site, the red was the type of alternative splicing event.

Alternative splicing events were classified into 12 types in ASProfile tool which were shown below:

(1) TSS: Alternative 5' first exon (transcription start site) the first exon splicing;
(2) TTS: Alternative 3' last exon (transcription terminal site) the last exon splicing;
(3) SKIP: Skipped exon(SKIP_ON,SKIP_OFF pair) single exon skipping;
(4) XSKIP: Approximate SKIP (XSKIP_ON,XSKIP_OFF pair) single exon skipping (fuzzy boundary);
(5) MSKIP: Multi-exon SKIP (MSKIP_ON,MSKIP_OFF pair) multi-exon skipping;
(6) XMSKIP: Approximate MSKIP (XMSKIP_ON,XMSKIP_OFF pair) multi-exon skipping (fuzzy boundary);
(7) IR: Intron retention (IR_ON, IR_OFF pair) single intron retention;
(8) XIR: Approximate IR (XIR_ON,XIR_OFF pair) single intron retention (fuzzy boundary);
(9) MIR: Multi-IR (MIR_ON, MIR_OFF pair) multi-intron retention;
(10) XMIR: Approximate MIR (XMIR_ON, XMIR_OFF pair) multi-intron retention (fuzzy boundary);
(11)AE: Alternative exon ends (5', 3', or both);
(12) XAE: Approximate AE variable 5' or 3' end (fuzzy boundary);

3.6.1 Statistics of Alternative Splicing Events

Statistics of predicted alternative splicing events were shown in the figures below.



Note: X-axis: Number of transcripts in specific alternative splicing type; Y-axis: 12 alternative splicing types.

3.6.2 Alternative splicing pattern

List of alternative splicing (AS) pattern
N1.AS.list
N2.AS.list
N3.AS.list
T1.AS.list
T2.AS.list
T3.AS.list

event_id	event_type	gene_id	Symbol	chrom	event_start	event_end	event_pattern	strand
1,000,001	TSS	ENSMUSG00000003134	Tbc1d8	1	39,445,833	39,445,951	39,445,833	-
1,000,002	TTS	ENSMUSG00000003134	Tbc1d8	1	39,411,292	39,411,867	39,411,867	-
1,000,003	TSS	ENSMUSG00000003135	Cnot11	1	39,577,383	39,577,504	39,577,504	+
1,000,004	TTS	ENSMUSG00000003135	Cnot11	1	39,581,481	39,581,578	39,581,481	+
1,000,005	TSS	ENSMUSG00000003458	Ncstn	1	171,910,168	171,910,317	171,910,168	-
1,000,006	TTS	ENSMUSG00000003458	Ncstn	1	171,893,762	171,894,363	171,894,363	-
1,000,007	TSS	ENSMUSG00000003464	Pex19	1	171,954,322	171,954,415	171,954,415	+
1,000,008	TTS	ENSMUSG00000003464	Pex19	1	171,961,774	171,962,850	171,961,774	+

Note: event_id: ID for AS;
event_type: Type of AS;
gene_id: Gene ID;
symbol: Gene symbol ;
chrom: Chromosome ID;
event_start: Starting position of AS;
event_end: Ending position of AS;
event_pattern: Pattern of AS;
strand: +/- strand.

3.7 Gene structure optimization

The accuracy of gene annotation gained from reference genome could be limited by the software, quality of data, etc. Therefore, it is necessary to process gene structure optimization on annotated genes. During this process, if continuous mapped reads were found outside the boundaries of original genes, the boundary of a gene may be corrected by extending untranslated region (UTR) to upper and down stream. In this project, 1,342 genes were optimized, which were listed in the following table.

Mus_musculus.geneStructure.optimize.xls

Note: GeneID: Gene ID;
Locus: Gene locus (chromosome ID: starting position-ending position);
Strand: +/- strand;
Site: Site of optimization (on 3' UTR or 5'UTR); OriginalRegion: Starting and ending position of original annotated genes;
OptimizedRegion: Starting and ending position of optimized genes.

3.8 Novel Gene Analysis

3.8.1 Novel gene discovery

In order to optimize the annotation information of a genome, discovery of novel transcripts and genes was achieved by StringTie on base of reference genome. The mapped reads were assembled and compared with original annotations of the genome. The transcript regions without annotation obtained by above processes are defined as novel transcripts. Excluding short transcripts(coding peptides with less than 50 amino acids) or those containing only one exons, 1,345 novel genes were discovered in this project. GFF file of novel genes was shown below.

Mus_musculus.newGene_final.filtered.gff

Note: #Seq_ID: Chromosome ID;
Source: Source of annotation (normally StringTie);
Type: Annotation Feature;
Start/End: Starting and ending position of the feature;
Score: Confidence of the annotation ("." represents a null value);
Strand: +/- strand of the feature;
Phase: phase of CDS feature (only available for CDS); be either "0", "1" or "2"; "." indicates not available;
Attributes: All the other information pertaining to this feature

.
Besides supplementary information in genome annotation, FASTA file of novel gene sequences were provided, as shown in the following documents.

Mus_musculus.newGene.longest_transcript.fa

3.8.2 Functional annotation of novel genes

Novel genes were annotated by DIAMOND [8] against databases including NR [9], Swiss-Prot [10], COG [11], KOG [12] and KEGG [13]. KEGG Orthology of novel genes were obtained by above processes. GO [14] Orthology of novel genes were obtained by the underlying software InterProScan [15] basic on the InterPro database. The amino acid sequences of novel genes were blasted against Pfam [16] database by HMMER [17] to gain the annotation information.

Summary of annotated novel genes by each database were shown in the table below.

Annotated databases	New Gene Number
COG	**25**
GO	**365**
KEGG	**331**
KOG	**104**
Pfam	**305**
Swiss-Prot	**344**
TrEMBL	**451**
eggNOG	**370**
nr	**506**
All	**519**

Note: Annotated databases: Database applied; New Gene Number: Number of annotated genes in specific database.

3.9 Gene Expression Quantification

3.9.1 Gene expression quantification

The number of fragments from a transcript is affected by sequencing data volume (or number of mapped reads), length of the transcript, expression level of transcripts. In order to reveal the expression level of each transcript more accurately, the number of mapped reads needs to be normalized by the length of its transcripts. FPKM(Fragments Per Kilobase of transcript per Million fragments mapped) was applied to measure the expression level of a gene or transcript by StringTie using maximum flow algorithm. The equation for FPKM is shown below.

$$FPKM = \frac{cDNAFragments}{MappedFragments(Millions) * TranscriptLength(kb)}$$

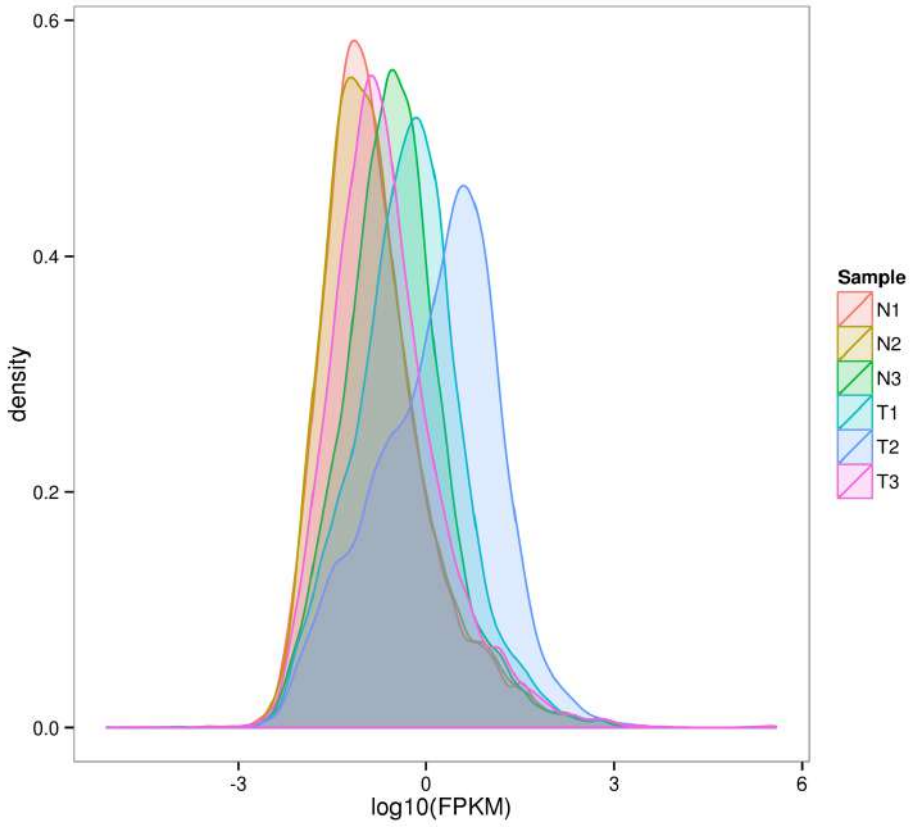
In the equation, cDNA Fragments represents the number of PE reads mapped to the specific transcript; Mapped Fragments (Millions) is the number of all mapped reads, which is counted as 10^6; Transcript Length(kb) is the length of transcript in unit of 10^3 b.

All_gene_fpkm.list

Note: #ID: gene ID; Values in the rest columns: FPKM value of the specific gene in each sample.

3.9.2 Distribution of gene expression

RNA-Seq is able to achieve highly-sensitive quantification of gene expression. Generally, a detectable trancriptome expression (FPKM) is ranging from 10^(-2) to 10^4 [19].



Note: Curves with different colors represent different samples; X-axis: log10(FPKM); Y-axis: Probability density.

3.10 Differential Expression Analysis

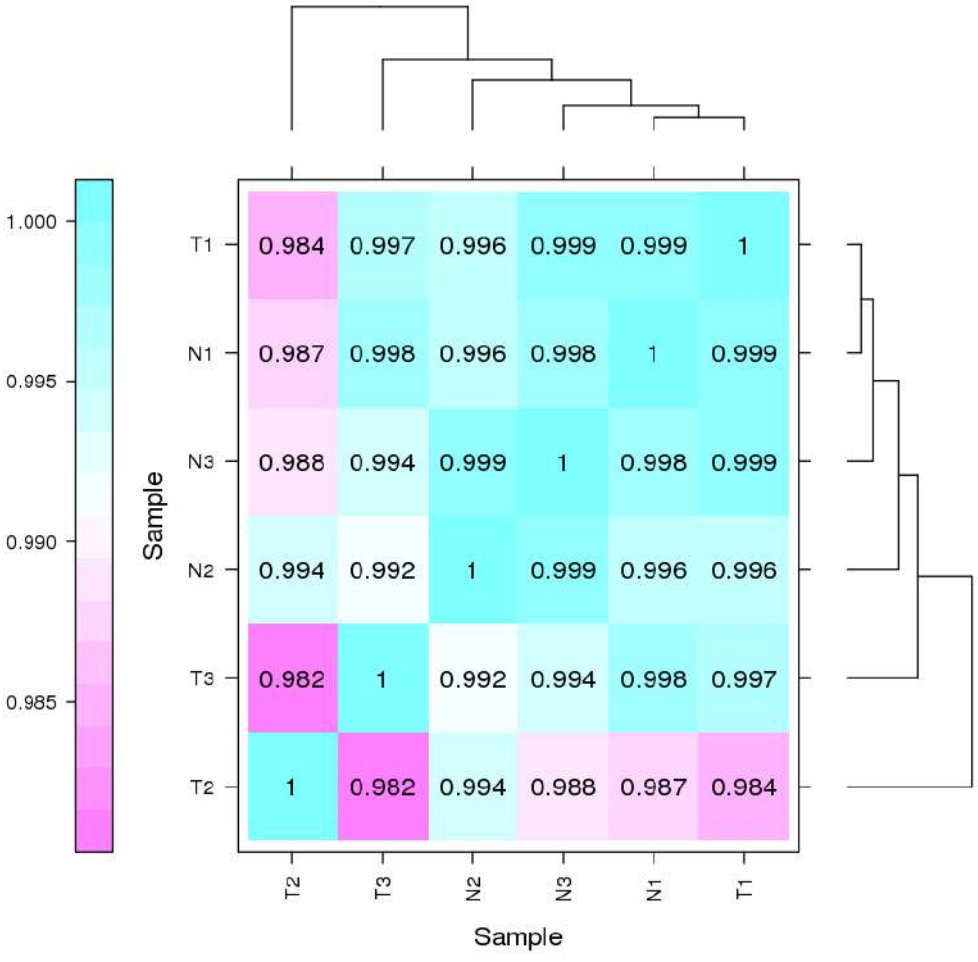
The expression of a gene can be influenced by both external stimuli and internal environment, which is highly temporal-specific and tissue-specific. The genes expressed significantly different under different conditions, such as treatment vs control, wild type vs mutants, different time points, different tissue, etc., are defined as Differentially Expressed Genes (DEG). Similarly, transcripts with significantly different expression level are named Differentially Expressed Transcript (DET). The collection of genes acquired in differential expression analysis is defined as DEG set. In result files, the gene sets were named as "A_VS_B" to specify the comparing pair. Normally, "A" represents control group, wild type or former time point. "B" normally represents corresponding treated group, mutant or later time point. The genes with a higher expression level in B than A are defined as up-regulated genes. The ones with lower expression level in B are defined as down-regulated genes. Therefore, up-reg and down-reg are relative definitions, which relies on the order of A and B.

3.10.1 Correlation assessment of biological replicates

It has been widely proven that gene expression level fluctuates among individuals differently (known as biological variability) [20] [21], which can not be eliminated via RNA sequencing, qPCR or microarray. In order to identify genes with true differential expression between groups, biological variability should be taken into consideration [22]. To date, one of the most commonly used and effective method to distinguish random fluctuation and real difference is to design biological replicates. The reliability of differential expression analysis is largely depending on the quality and the number of replicates. Therefore, in projects with biological replicates, it is crucial to ensure the reproducibility of the replicates by correlation analysis. In addition, correlation analysis could also help screening for abnormal samples.

Pearson correlation coefficient R (Pearson's Correlation Coefficient) was applied in this project to evaluate reproducibility of biological replicates [23]. A closer R2 value to 1 indicates better reproducibility between the two samples. We committed that all biological replicates will be processed by the same technician in the same batch including RNA extraction and library construction. The libraries will be sequenced in the same run on the same lane. We will also perform in depth analysis on abnormal samples. Basing on the outputs, we will discuss with our clients and make final decision on whether the abnormal sample should be removed in downstream analysis.

Correlations between samples were shown below.



3.10.2 Differentially expressed genes

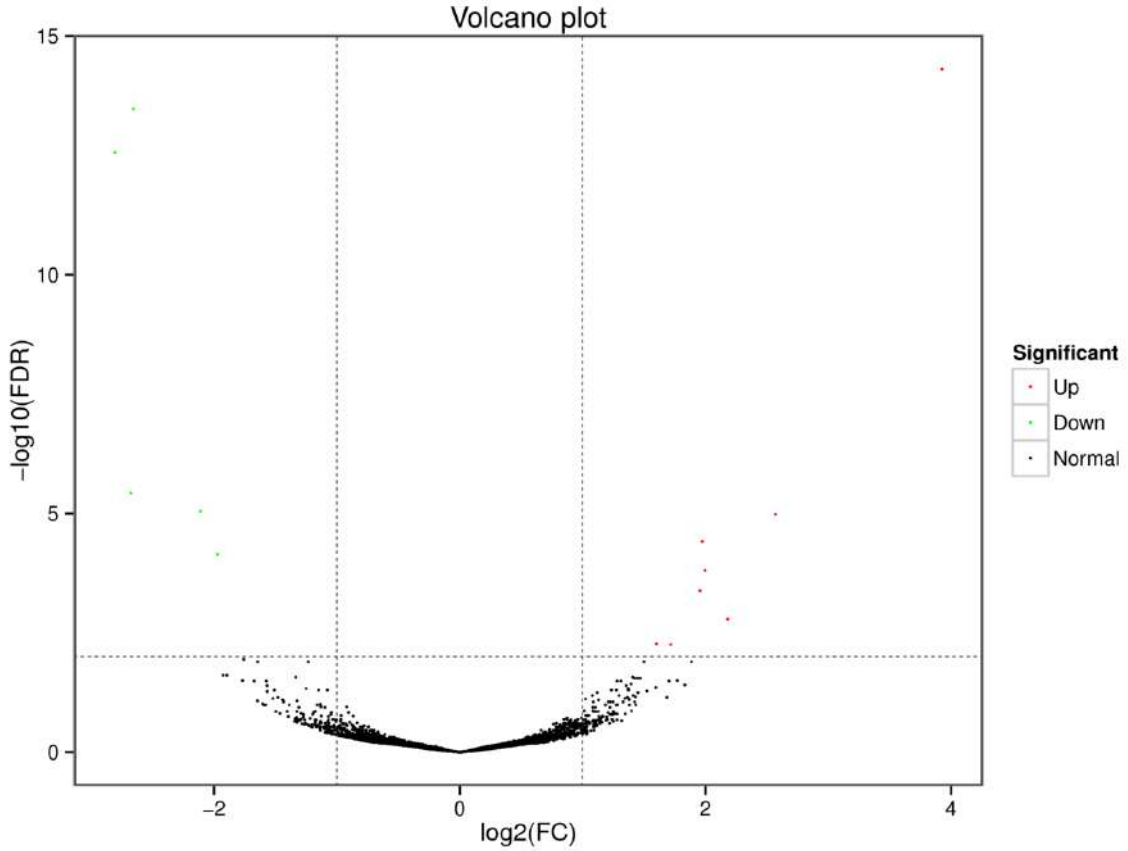
For experiments with biological replicates, differential expression analysis is processed by DESeq2 [24]. For projects without biological replicates, edgeR [25] is applied.

Criteria for differentially expressed genes was set as Fold Change(FC)≥2 and FDR<0.01. Fold change(FC) refers to the ratio of gene expression in two samples. False Discovery Rate (FDR) refers to adjusted p-value, which is used to measure significancy of difference.
Differential expression analysis output

N1_N2_N3_vs_T1_T2_T3.DEG.final.xls

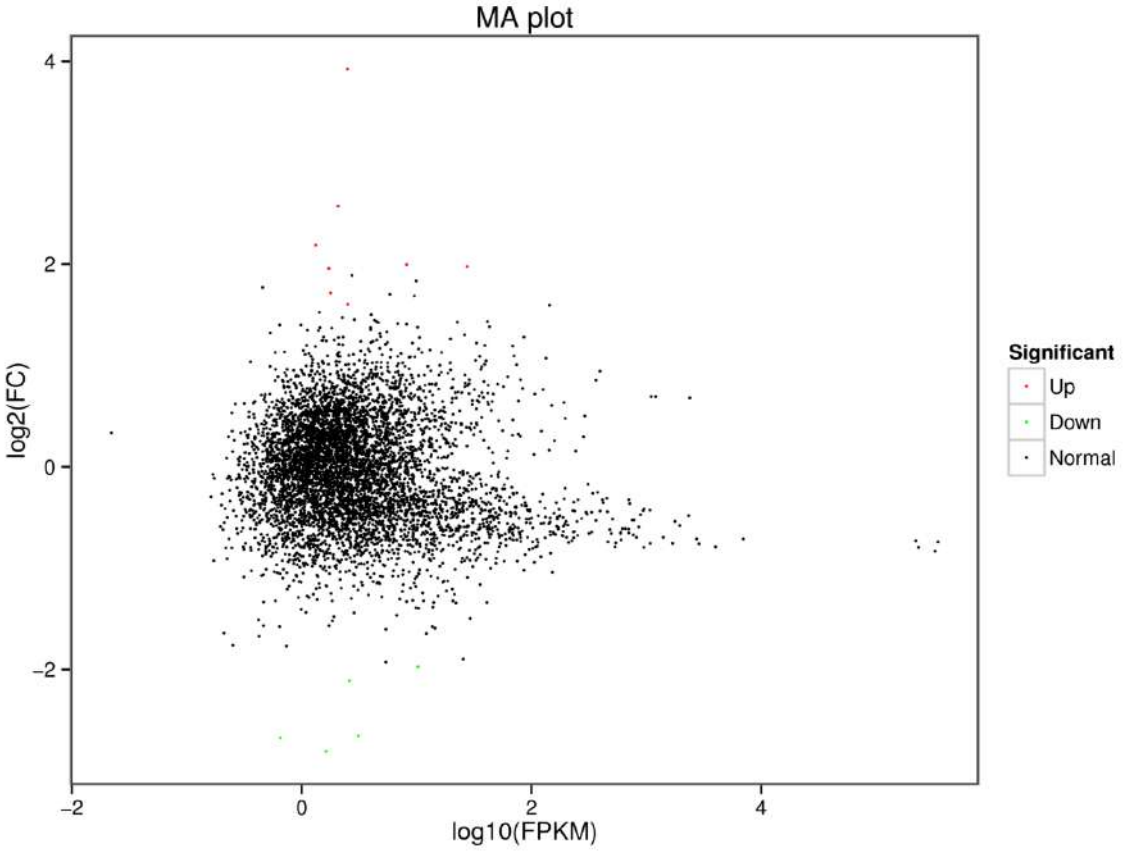
Note: ID: Gene ID; *_Count: Gene expression(reads count) in corresponding sample; *_FPKM: Gene expression(FPKM) in corresponding sample; FDR: False Discovery Rate; log2FC: Fold change normalized by log2; regulated: Up or down regulated.

Volcano plot is able to directly present difference in gene expression between two samples and statistical significancy of the difference. Volcano plots of two samples were shown below.



Note:
In volcano plot, each dot represents a gene. X-axis: log2Fold change of expression; Y-axis: -log10(FDR) or -log10(P-value). Dots farther to y=0 represent genes with large difference in expression between two samples. Dots farther to x=0 represents genes of which the difference is more reliable. Green dots are down-regulated genes, while red dots are up-regulated ones and black dots are genes without significant difference.

A plots shows the overall distribution of gene expression and fold change of expression level between two samples. MA plot of differentially expressed genes were shown in figures below.



Note: In MA plot, each dot represents a single gene.
X-axis: A value, i.e. log2 (FPKM);
Y-axis: M value, i.e. log2(FC);
The dots coloured in red and green stand for significant up-regulated and down-regulated genes respectively. Black dots stand for the genes without significant difference in expression between two samples.

3.10.3 Statistics on DEGs

Differential expressed genes identified in all groups were shown below.

DEG Set	DEG Number	up-regulated	down-regulated
N1_N2_N3_vs_T1_T2_T3	**13**	**8**	**5**

Note: DEG Set: Comparing sample pair; DEG Number: Number of differentially expressed genes; up-regulated: Number of up-regulated genes; down-regulated: Number of down-regulated genes.

3.11 Functional annotation of DEGs and enrichment analysis

The DEGs identified in differential expression analysis were annotated. The summary of annotations was shown in the table below.

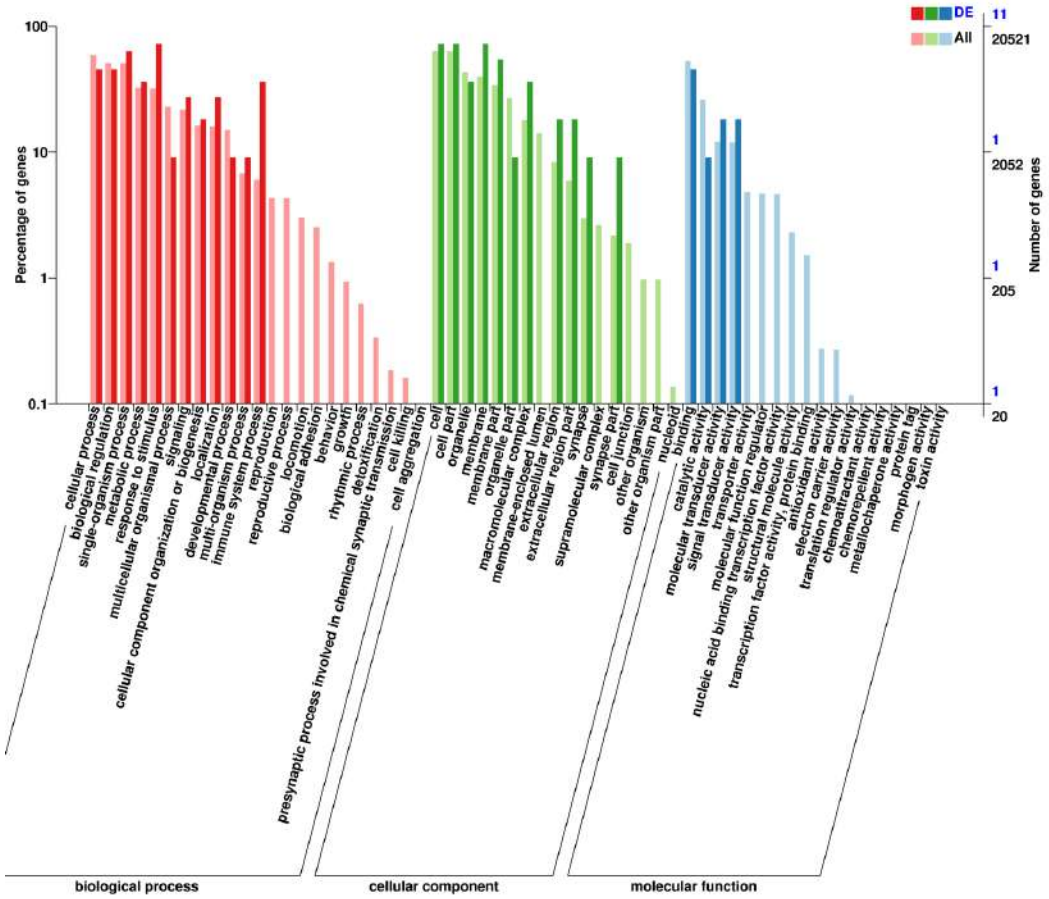
DEG Set	Total	COG	GO	KEGG	KOG	NR	Pfam	Swiss-Prot	eggNOG
N1_N2_N3_vs_T1_T2_T3	12	2	11	10	4	12	11	11	11

Note: DEG Set: Group set of DEG analysis; Total: Number of annotated DEGs; The rest columns are the numbers of annotated DEGs in corresponding database.

3.11.1 GO analysis on DEGs

GO (Gene Ontology) database is a structured biological annotation system established in 2000 containing a standard vocabulary of gene and gene products functions. GO annotation system is a directed acyclic graph containing three main branches: Biological Process, Molecular Function and Cellular Component.

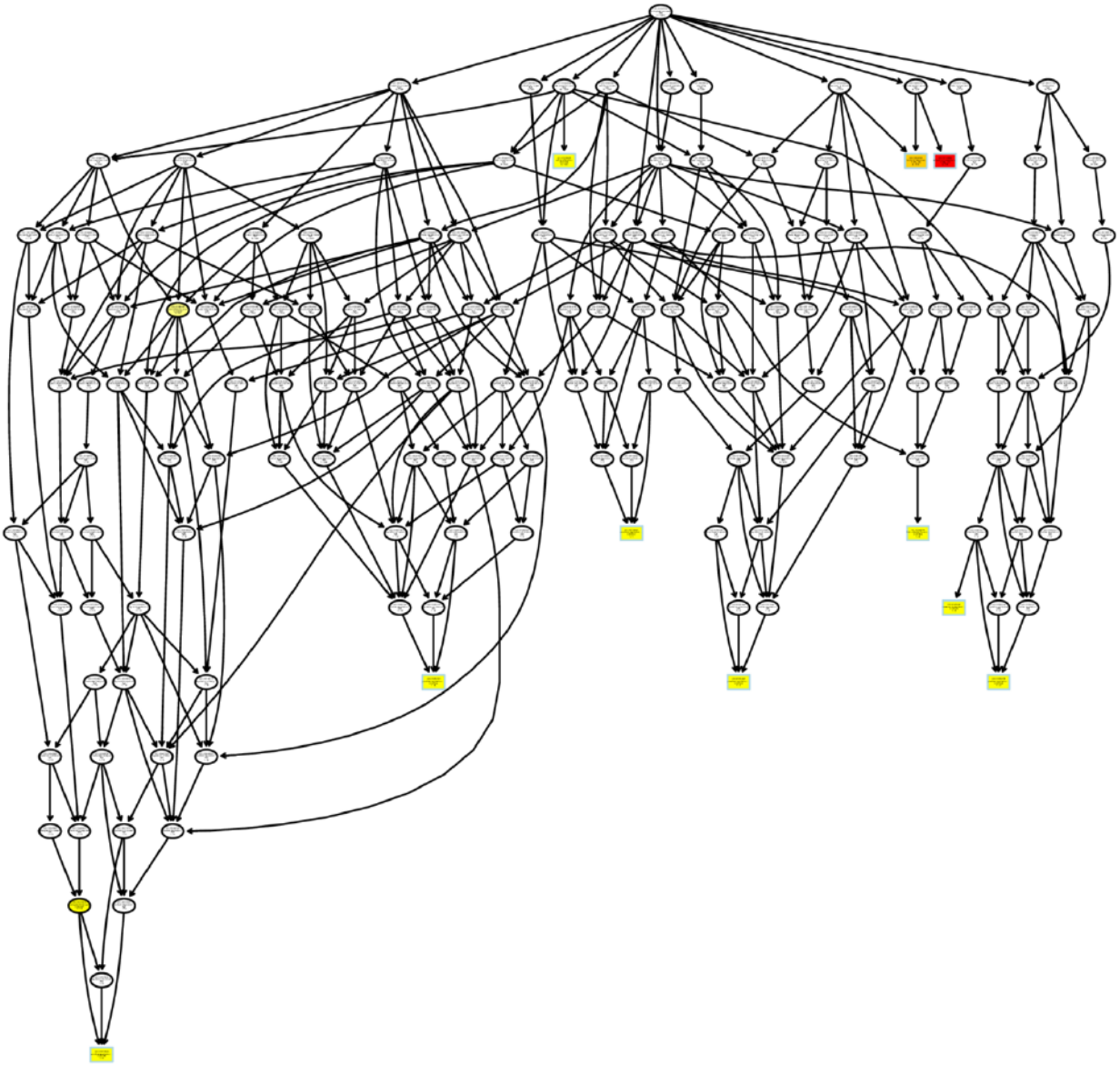
GO classification of DEGs between samples was shown in the following figures.

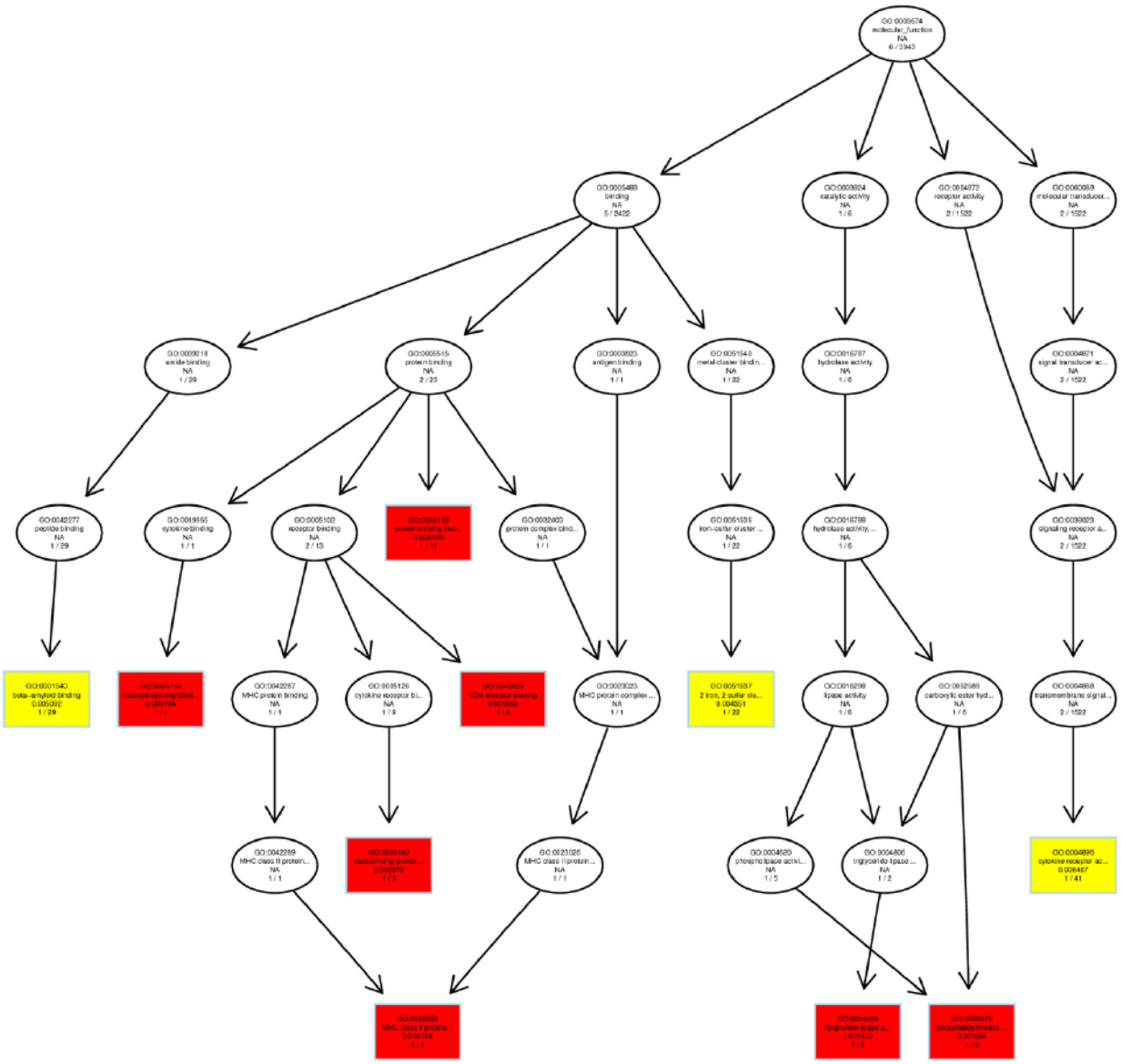


Note: X-axis: Go terms and classifications; Y-axis: Number of DEGs(genes) annotated to the term(right) and percentage of that in all DEGs(genes) (Left).This figure shows the GO enrichment in DEGs and in all genes, which indicates the importance of a specific GO term in DEGs and all genes respectively. The terms with two bars significantly different from each other can be picked up as potential targets for further analysis on functions, since these GO terms are enriched differently between DEGs-based and all-gene-based enrichment.

3.11.2 GO enrichment analysis on DEGs

DEGs were then subjected to functional enrichment analysis and the enriched GO terms and corresponding inclusion relationships were shown in the directed acyclic graph. In the figure, the direction of arrows represents inclusion relations between terms, i.e. the nodes are more specific than their upper nodes. Directed acyclic graphs of DEGs generated by topGO [26]were shown below.



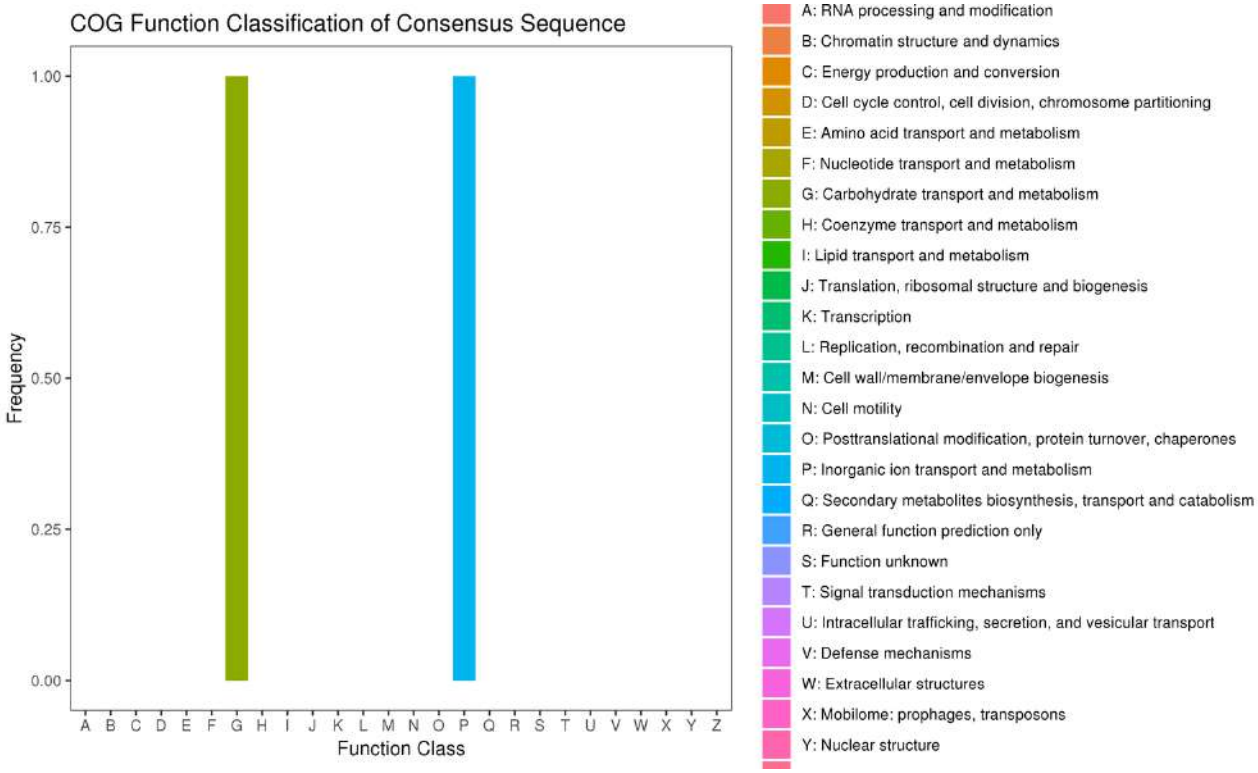


Note: The most significantly enriched 10 terms were shown in cubes. Each box or node contains a description of GO term and significancy value of enrichment. The color represents the significancy, where a darker color indicates a more significant enrichment.

3.11.3 COG classification on DEGs

COG（Cluster of Orthologous Groups of proteins is a database collecting phylogenetic classification of proteins, which can be provide orthologous classification information of gene products.

Summary of COG classifications on DEGs were shown in the figures below.

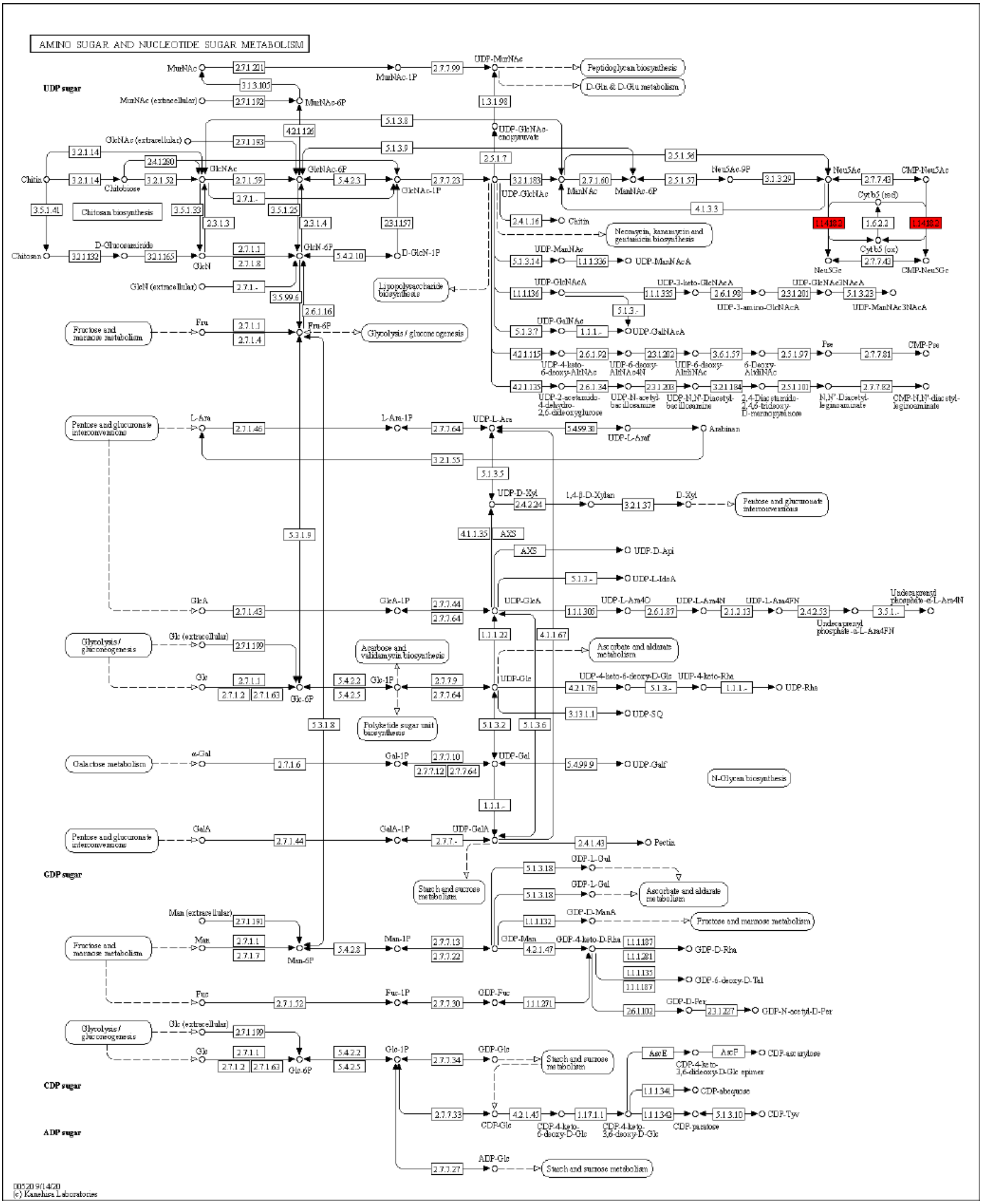


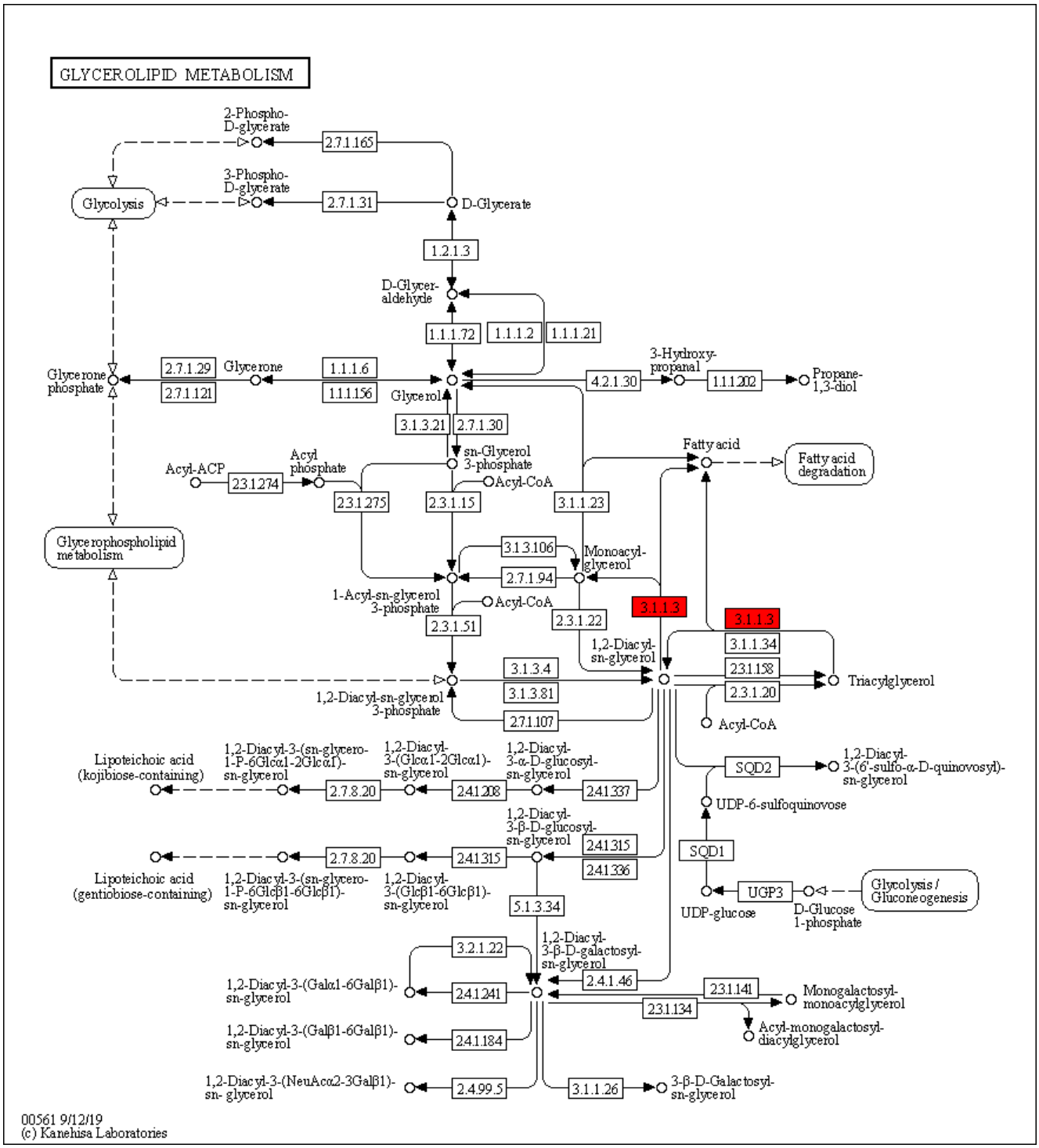
Note: X-axis: COG classification terms; Y-axis: Number of genes in the term. In the different functional classes, the number of genes reflects the preference of gene functions in different experimental groups, such as metabolic function or physiological bias, etc. These can be explained based on specific research subjects.

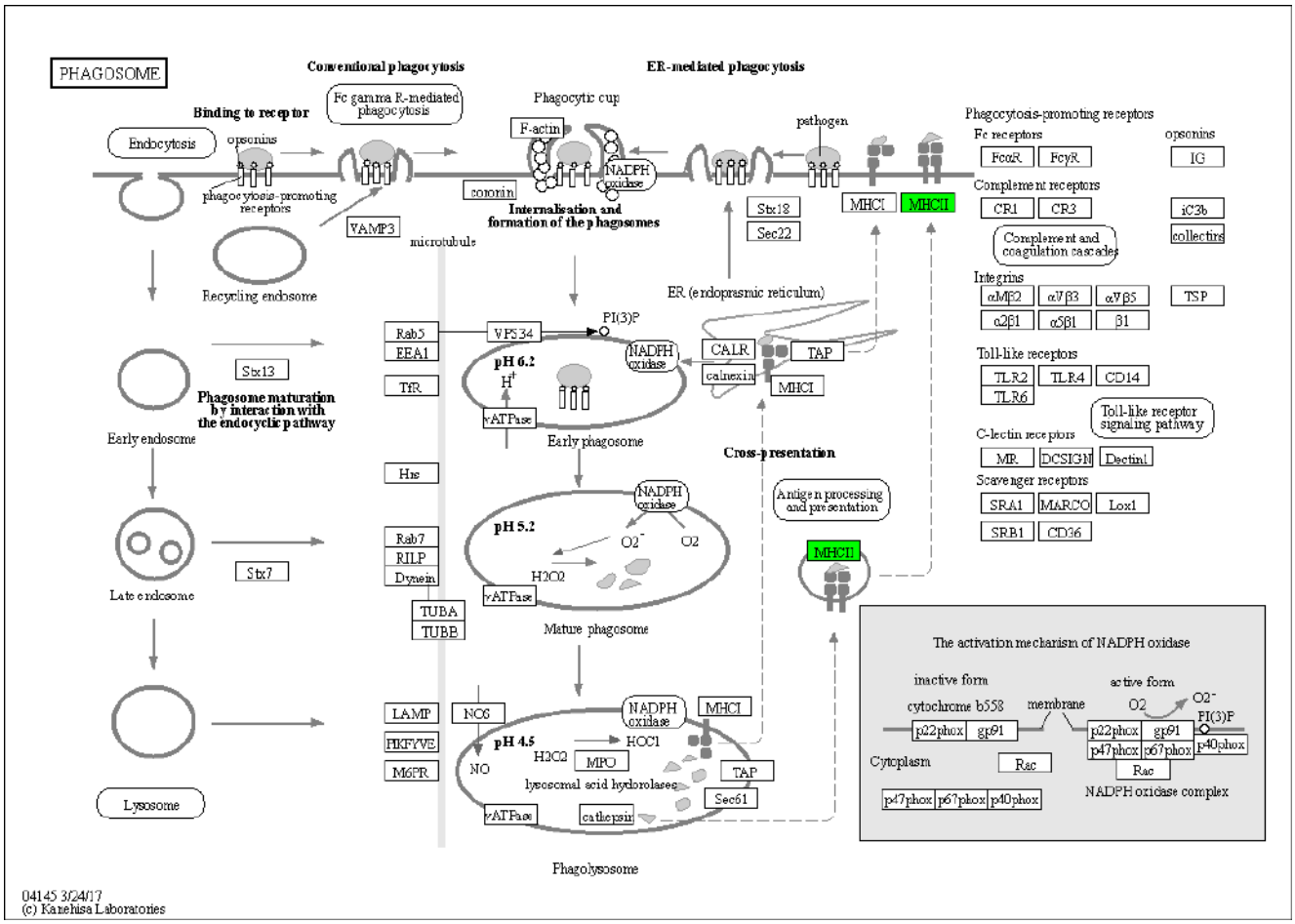
3.11.4 KEGG annotation of differentially expressed genes

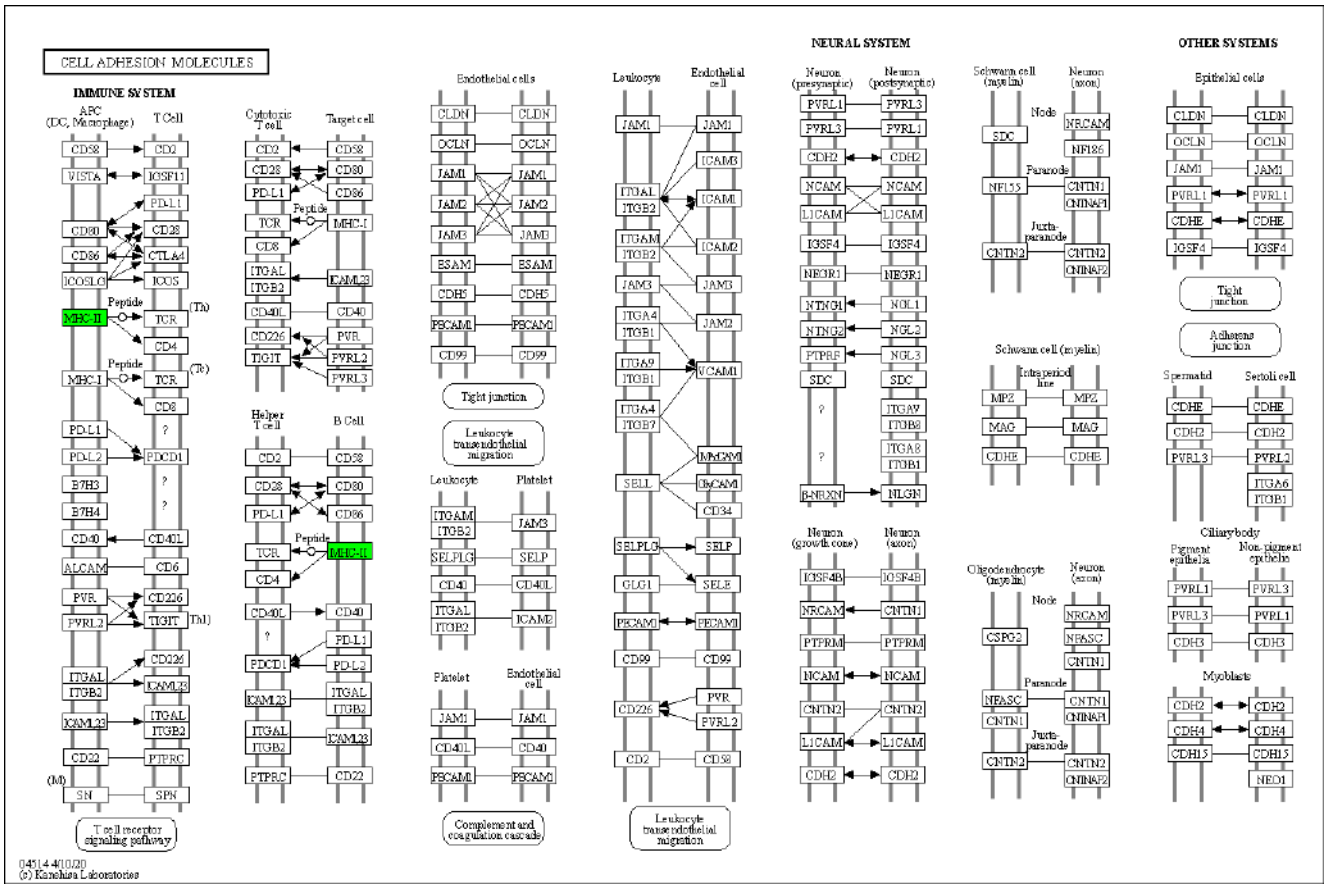
In biological organisms, series of gene products are working synergistically to perform biological functions, which is so called pathway. Annotating genes within pathway networks could largely benefit further analysis on biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes) is one of the major databases on pathways, including metabolic pathways of carbohydrates, nucleotides, amino acids and biological degradation of organics. Besides metabolic pathways, KEGG contains comprehensive description on enzymes involved in the pathways, including amino acid sequences, links to PDB database, etc.

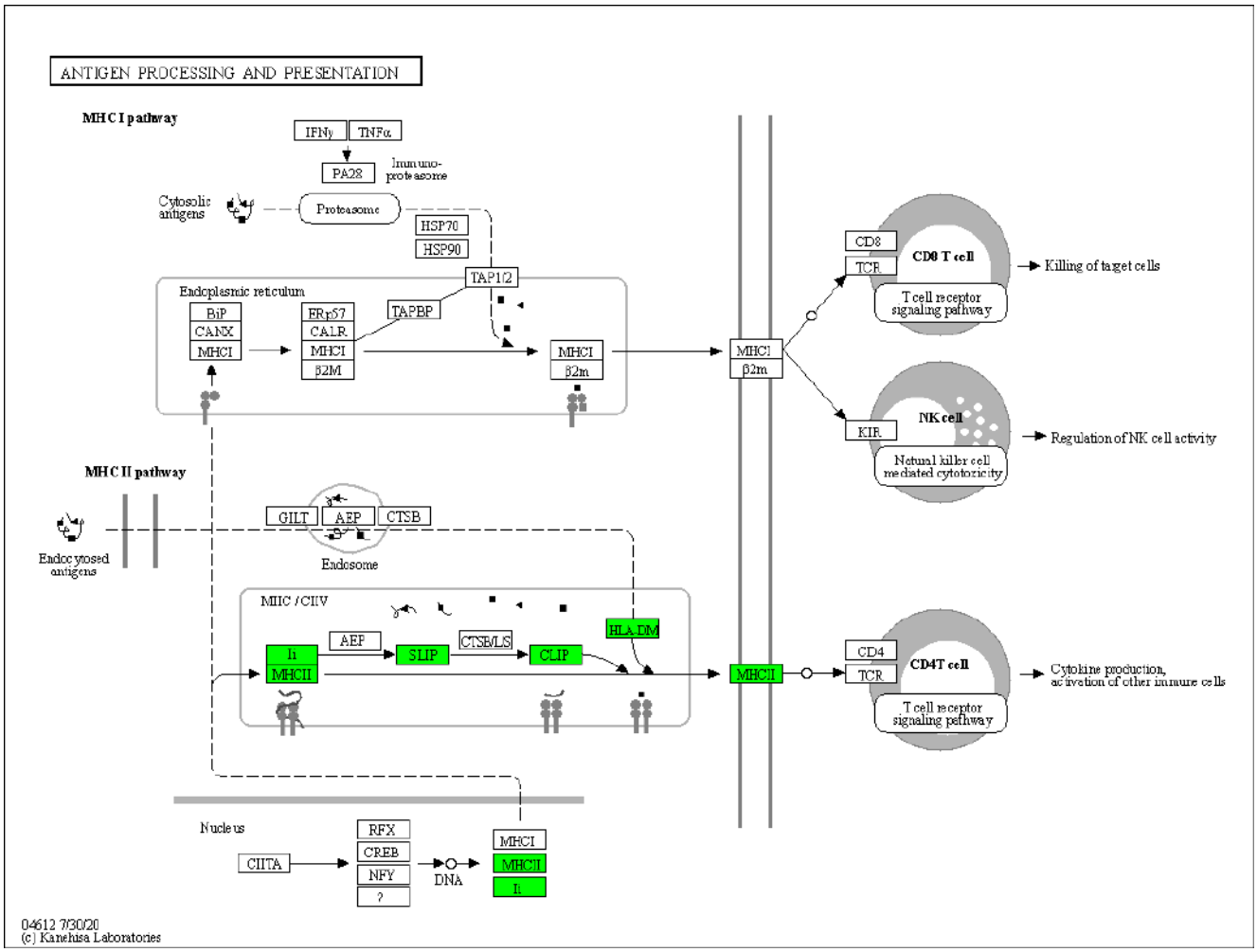
KEGG pathway annotation on DEGs were shown in the following figure.

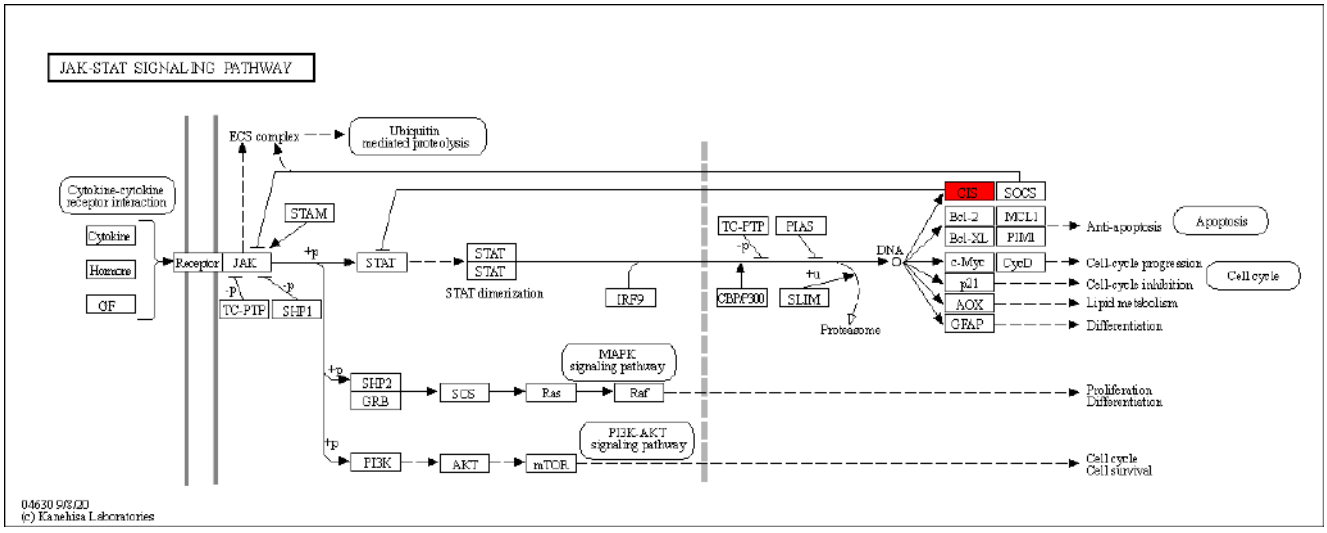

AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM
UDP sugar
GDP sugar
CDP sugar
ADP sugar
MurNAc
MurNAc-1P
UDP-MurNAc
Peptidoglycan biosynthesis
D-Gln & D-Glu metabolism
MurNAc (extracellular)
MurNAc-6P
GlcNAc (extracellular)
UDP-GlcNAc-enolpyruvate
Chitin
Chitobiose
GlcNAc
GlcNAc-6P
GlcNAc-1P
UDP-GlcNAc
ManNAc
ManNAc-6P
Neu5Ac-9P
Neu5Ac
CMP-Neu5Ac
Cytb5 (red)
Cytb5 (ox)
Neu5Gc
CMP-Neu5Gc
Chitosan biosynthesis
Chitosan
D-Glucosaminide
GlcN
GlcN-6P
D-GlcN-1P
GlcN (extracellular)
Neomycin, kanamycin and gentamicin biosynthesis
UDP-ManNAc
UDP-ManNAcA
UDP-GlcNAcA
UDP-3-keto-GlcNAcA
UDP-3-amino-GlcNAcA
UDP-GlcNAc3NAcA
UDP-ManNAc3NAcA
Lipopolysaccharide biosynthesis
UDP-GalNAc
Fructose and mannose metabolism
Fru
Fru-6P
Glycolysis / gluconeogenesis
Pse
CMP-Pse
Pentose and glucuronate interconversions
L-Ara
L-Ara-1P
UDP-L-Ara
UDP-L-Araf
Arabinan
UDP-D-Xyl
1,4-β-D-Xylan
D-Xyl
UDP-D-Api
UDP-L-IdoA
GlcA
GlcA-1P
UDP-GlcA
UDP-L-Ara4O
UDP-L-Ara4N
UDP-L-Ara4FN
Glc (extracellular)
Acarbose and validamycin biosynthesis
Ascorbate and aldarate metabolism
Glc
Glc-6P
Glc-1P
UDP-Glc
UDP-4-keto-6-deoxy-D-Glc
UDP-4-keto-Rha
UDP-Rha
UDP-SQ
Polyketide sugar unit biosynthesis
Galactose metabolism
α-Gal
Gal-1P
UDP-Gal
UDP-Galf
N-Glycan biosynthesis
GalA
GalA-1P
UDP-GalA
Pectin
GDP-L-Gul
GDP-L-Gal
Starch and sucrose metabolism
GDP-D-ManA
Man (extracellular)
Man
Man-6P
Man-1P
GDP-Man
GDP-4-keto-D-Rha
GDP-D-Rha
GDP-6-deoxy-D-Tal
Fuc
Fuc-1P
GDP-Fuc
GDP-D-Per
GDP-N-acetyl-D-Per
GDP-Glc
CDP-ascarylose
CDP-abequose
CDP-Tyv
CDP-Glc
CDP-4-keto-6-deoxy-D-Glc
CDP-4-keto-3,6-dideoxy-D-Glc
CDP-paratose
ADP-Glc
1.14.18.2
00520 9/14/20
(c) Kanehisa Laboratories

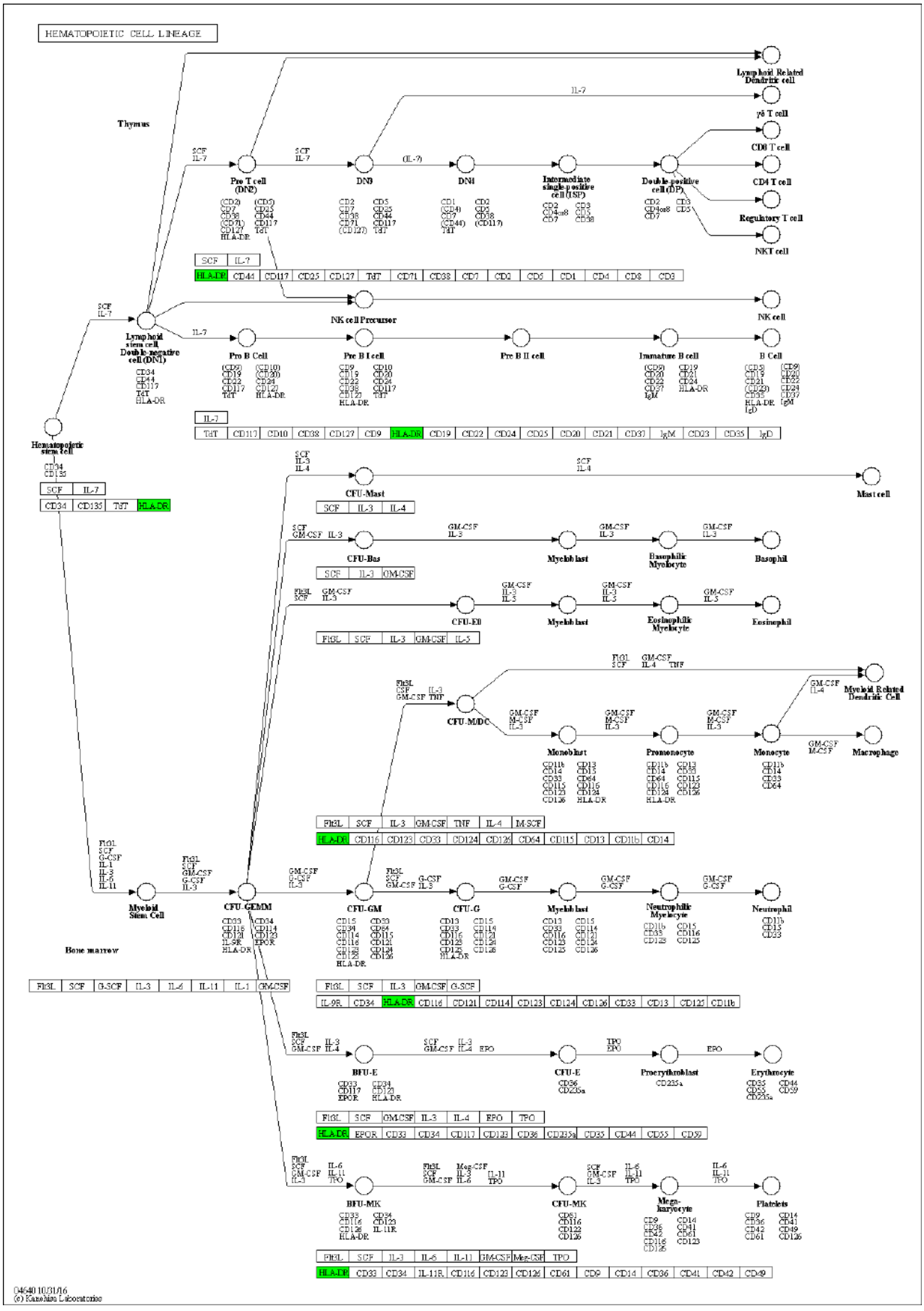

GLYCEROLIPID METABOLISM
2-Phospho-D-glycerate
2.7.1.165
3-Phospho-D-glycerate
Glycolysis
2.7.1.31
D-Glycerate
1.2.1.3
D-Glycer-aldehyde
1.1.1.72
1.1.1.2
1.1.1.21
Glycerone phosphate
2.7.1.29
2.7.1.121
Glycerone
1.1.1.6
1.1.1.156
Glycerol
4.2.1.30
3-Hydroxy-propanal
1.1.1.202
Propane-1,3-diol
3.1.3.21
2.7.1.30
Acyl-ACP
23.1.274
Acyl phosphate
sn-Glycerol 3-phosphate
Acyl-CoA
Fatty acid
Fatty acid degradation
2.3.1.275
2.3.1.15
3.1.1.23
Glycerophospholipid metabolism
3.1.3.106
Monoacyl-glycerol
2.7.1.94
1-Acyl-sn-glycerol 3-phosphate
Acyl-CoA
2.3.1.51
2.3.1.22
3.1.1.3
3.1.1.3
3.1.1.34
1,2-Diacyl-sn-glycerol
3.1.3.4
23.1.158
Triacylglycerol
1,2-Diacyl-sn-glycerol 3-phosphate
3.1.3.81
2.3.1.20
2.7.1.107
Acyl-CoA
Lipoteichoic acid (kojibiose-containing)
1,2-Diacyl-3-(sn-glycero-1-P-6Glcα1-2Glcα1)-sn-glycerol
2.7.8.20
1,2-Diacyl-3-(Glcα1-2Glcα1)-sn-glycerol
2.4.1.208
1,2-Diacyl-3-α-D-glucosyl-sn-glycerol
2.4.1.337
SQD2
1,2-Diacyl-3-(6'-sulfo-α-D-quinovosyl)-sn-glycerol
UDP-6-sulfoquinovose
1,2-Diacyl-3-β-D-glucosyl-sn-glycerol
2.4.1.315
2.4.1.336
SQD1
Lipoteichoic acid (gentiobiose-containing)
1,2-Diacyl-3-(sn-glycero-1-P-6Glcβ1-6Glcβ1)-sn-glycerol
1,2-Diacyl-3-(Glcβ1-6Glcβ1)-sn-glycerol
5.1.3.34
UGP3
Glycolysis / Gluconeogenesis
UDP-glucose
D-Glucose 1-phosphate
1,2-Diacyl-3-β-D-galactosyl-sn-glycerol
3.2.1.22
2.4.1.46
1,2-Diacyl-3-(Galα1-6Galβ1)-sn-glycerol
2.4.1.241
23.1.141
Monogalactosyl-monoacylglycerol
23.1.134
Acyl-monogalactosyl-diacylglycerol
1,2-Diacyl-3-(Galβ1-6Galβ1)-sn-glycerol
2.4.1.184
1,2-Diacyl-3-(NeuAcα2-3Galβ1)-sn- glycerol
2.4.99.5
3.1.1.26
3-β-D-Galactosyl-sn-glycerol
00561 9/12/19
(c) Kanehisa Laboratories

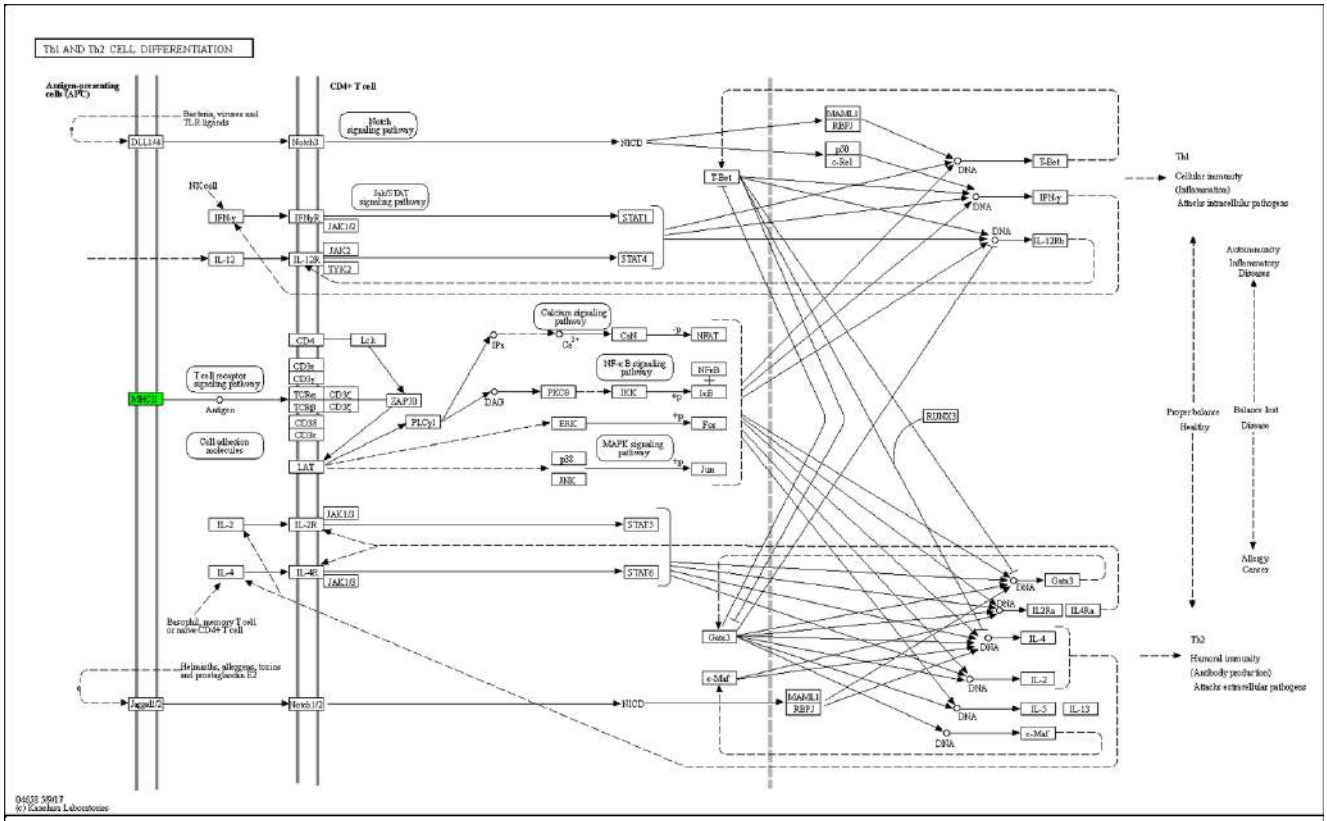

PHAGOSOME
Conventional phagocytosis
ER-mediated phagocytosis
Binding to receptor
Fc gamma R-mediated phagocytosis
Phagocytic cup
F-actin
pathogen
Phagocytosis-promoting receptors
Fc receptors
FcαR
FcγR
opsonins
IG
Endocytosis
opsonins
phagocytosis-promoting receptors
NADPH oxidase
Complement receptors
CR1
CR3
iC3b
coronin
Internalisation and formation of the phagosomes
Stx18
Sec22
MHCI
MHCII
Complement and coagulation cascades
collectins
VAMP3
microtubule
Integrins
αMβ2
αVβ3
αVβ5
TSP
α2β1
α5β1
β1
Recycling endosome
ER (endoplasmic reticulum)
PI(3)P
Rab5
EEA1
VPS34
CALR
TAP
Toll-like receptors
TLR2
TLR4
CD14
TLR6
Stx13
pH 6.2
H+
calnexin
MHCI
Toll-like receptor signaling pathway
Phagosome maturation by interaction with the endocytic pathway
TfR
vATPase
C-lectin receptors
MR
DCSIGN
Dectin1
Early endosome
Early phagosome
Cross-presentation
Scavenger receptors
Hrs
NADPH oxidase
Antigen processing and presentation
SRA1
MARCO
Lox1
pH 5.2
O2-
O2
MHCII
SRB1
CD36
Rab7
RILP
Dynein
H2O2
Stx7
vATPase
Late endosome
TUBA
TUBB
Mature phagosome
The activation mechanism of NADPH oxidase
inactive form
active form
cytochrome b558
membrane
O2
O2-
PI(3)P
LAMP
NC6
NADPH oxidase
MHCI
p22phox
gp91
p22phox
gp91
pH 4.5
HOCl
p47phox
p67phox
p40phox
H2O2
MPO
NO
Cytoplasm
Rac
Rac
lysosomal acid hydrolases
TAP
p47phox
p67phox
p40phox
NADPH oxidase complex
Sec61
vATPase
cathepsin
Lysosome
Phagolysosome
04145 3/24/17
(c) Kanehisa Laboratories

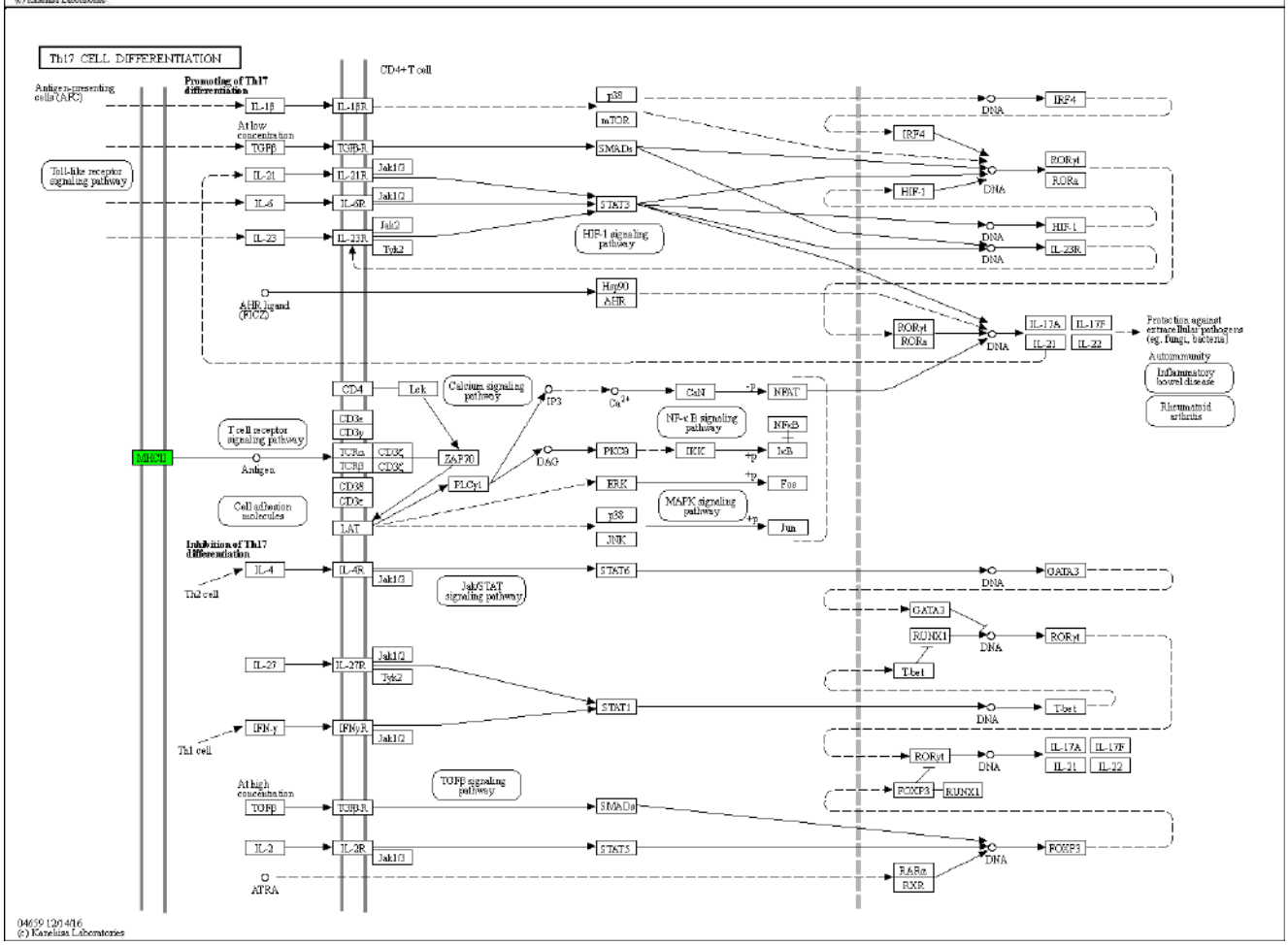

CELL ADHESION MOLECULES
IMMUNE SYSTEM
NEURAL SYSTEM
OTHER SYSTEMS
T cell receptor signaling pathway
Tight junction
Leukocyte transendothelial migration
Complement and coagulation cascade
Adherens junction
(c) Kanehisa Laboratories

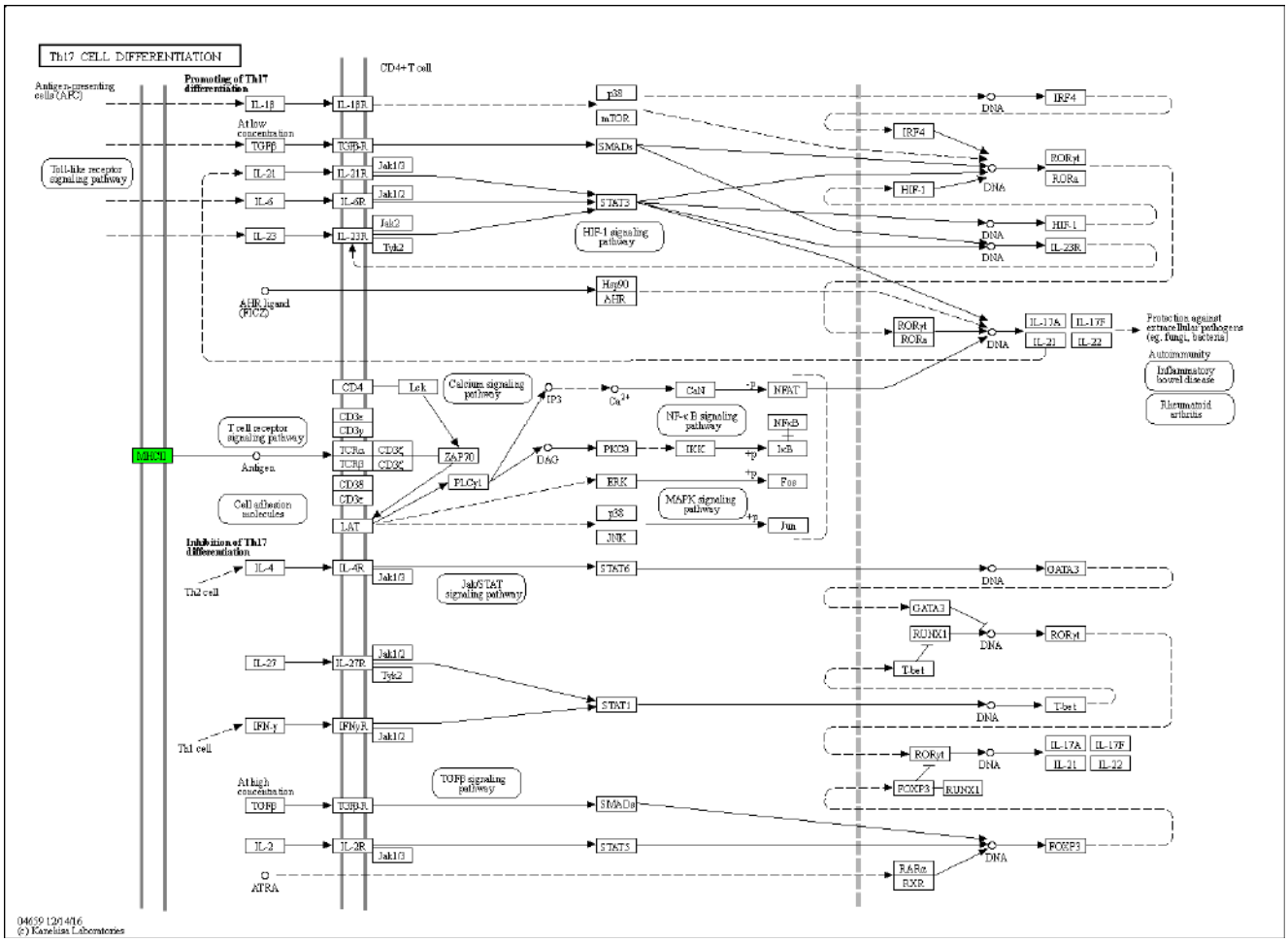

ANTIGEN PROCESSING AND PRESENTATION
MHC I pathway
IFNγ
TNFα
PA28
Immuno-proteasome
Cytosolic antigens
Proteasome
HSP70
HSP90
TAP1/2
Endoplasmic reticulum
BiP
CANX
MHCI
ERp57
CALR
MHCI
β2M
TAPBP
MHCI
β2m
MHCI
β2m
CD8
TCR
CD8 T cell
T cell receptor signaling pathway
Killing of target cells
KIR
NK cell
Natural killer cell mediated cytotoxicity
Regulation of NK cell activity
MHC II pathway
Endocytosed antigens
GILT
AEP
CTSB
Endosome
MIIC / CIIV
Ii
MHCII
AEP
SLIP
CTSB/LS
CLIP
HLA-DM
MHCII
CD4
TCR
CD4T cell
T cell receptor signaling pathway
Cytokine production, activation of other immune cells
Nucleus
CIITA
RFX
CREB
NFY
?
DNA
MHCI
MHCII
Ii
04612 7/30/20
(c) Kanehisa Laboratories

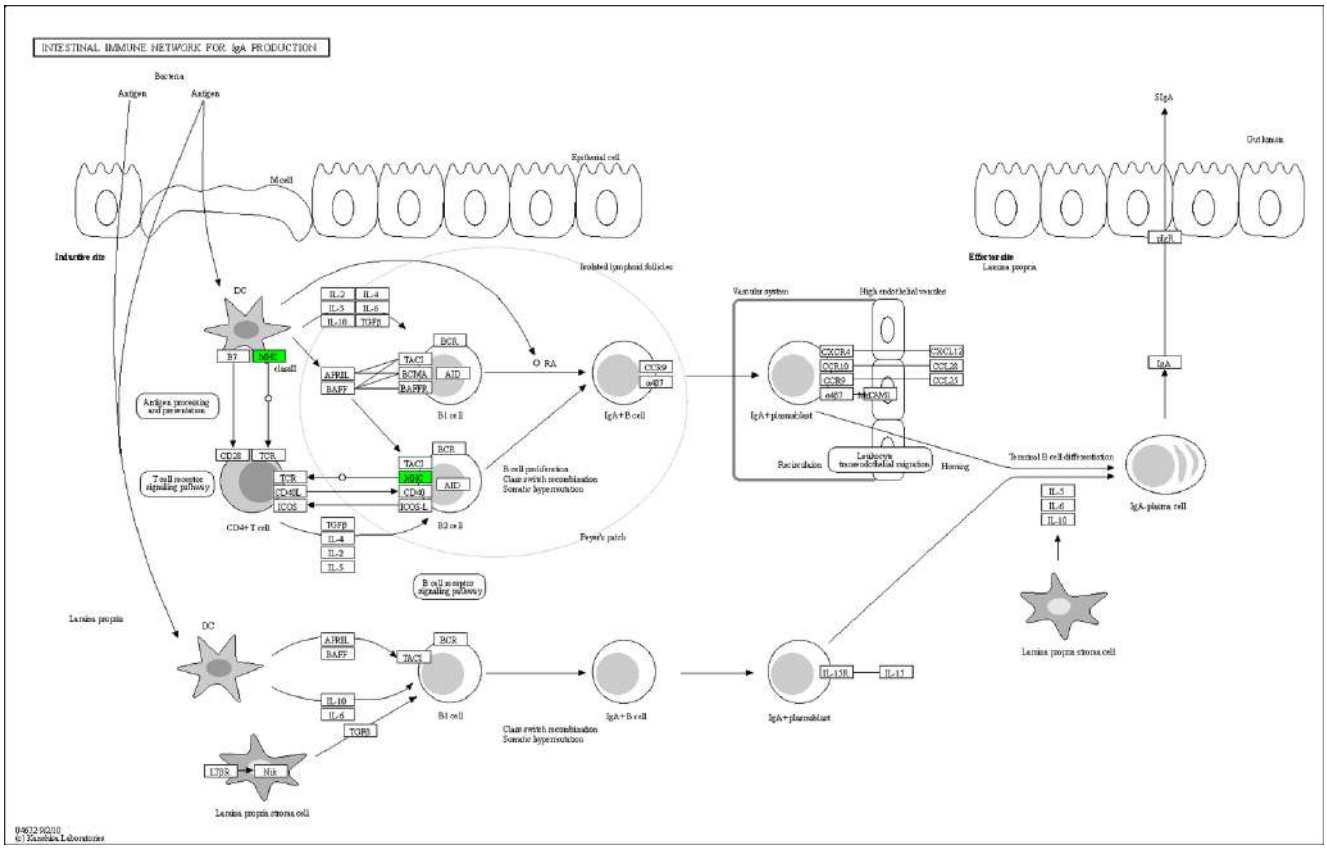

JAK-STAT SIGNALING PATHWAY
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
Cytokine
Hormone
GF
STAM
Receptor
JAK
P
STAT
STAT
STAT
STAT dimerization
IRF9
TC-PTP
SHP1
TC-PTP
PIAS
CBP/p300
SLIM
P
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
c-Myc
CycD
p21
AOX
GFAP
Anti-apoptosis
Apoptosis
Cell-cycle progression
Cell-cycle inhibition
Lipid metabolism
Differentiation
Cell cycle
SHP2
GRB
SOS
Ras
Raf
MAPK signaling pathway
Proliferation
Differentiation
PI3K
AKT
mTOR
PI3K-AKT signaling pathway
Cell cycle
Cell survival
04630 9/3/20
(c) Kanehisa Laboratories

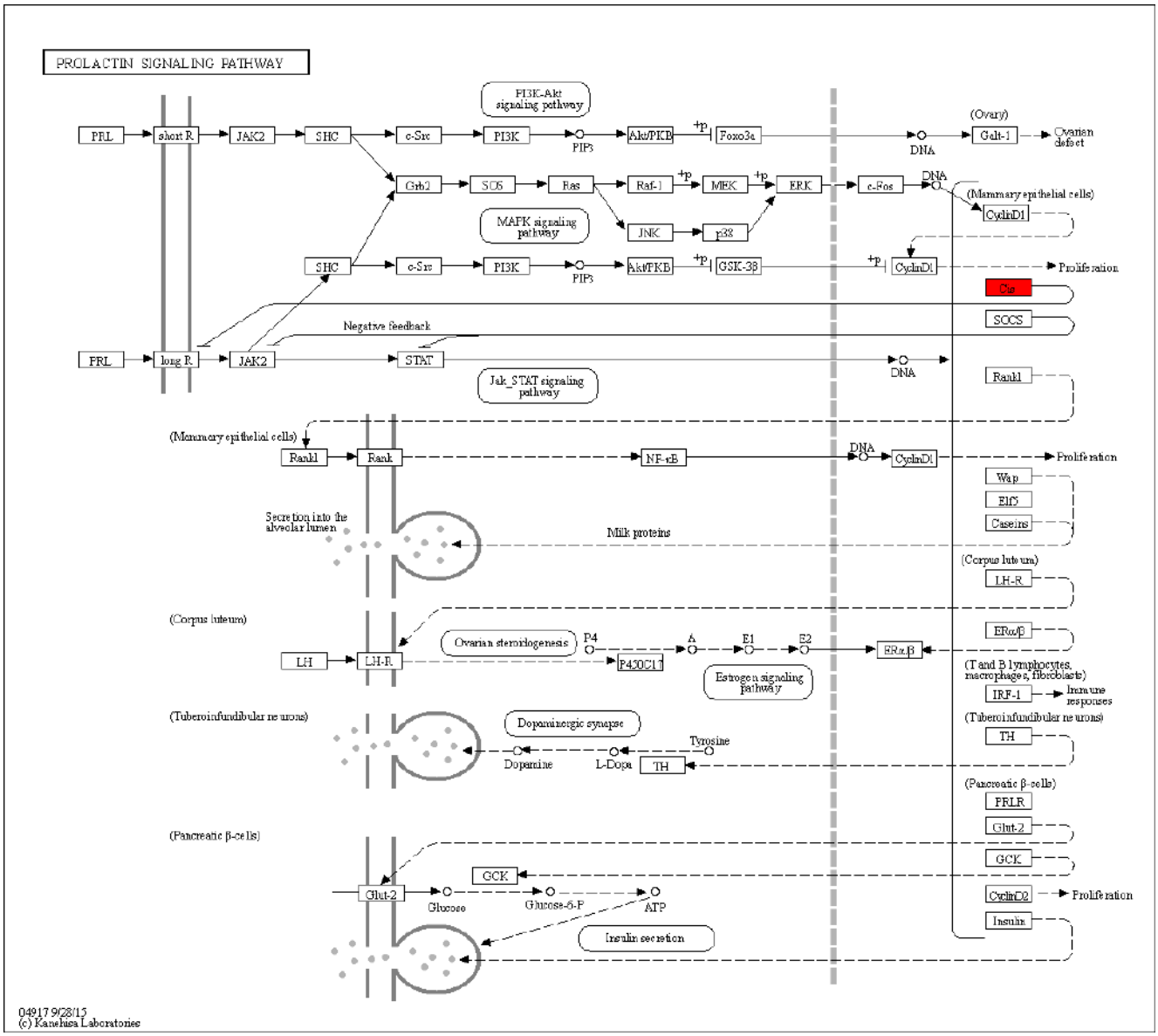

HEMATOPOIETIC CELL LINEAGE
Thymus
Lymphoid Related Dendritic cell
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
NK cell Precursor
NK cell
Lymphoid stem cell Double-negative cell (DN1)
Pro B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
Hematopoietic stem cell
CFU-Mast
Mast cell
CFU-Bas
Myeloblast
Basophilic Myelocyte
Basophil
CFU-Eo
Eosinophilic Myelocyte
Eosinophil
CFU-M/DC
Monoblast
Promonocyte
Monocyte
Myeloid Related Dendritic Cell
Macrophage
Myeloid Stem Cell
CFU-GEMM
CFU-GM
CFU-G
Neutrophilic Myelocyte
Neutrophil
Bone marrow
BFU-E
CFU-E
Proerythroblast
Erythrocyte
BFU-MK
CFU-MK
Megakaryocyte
Platelets
04640 10/31/16
(c) Kanehisa Laboratories

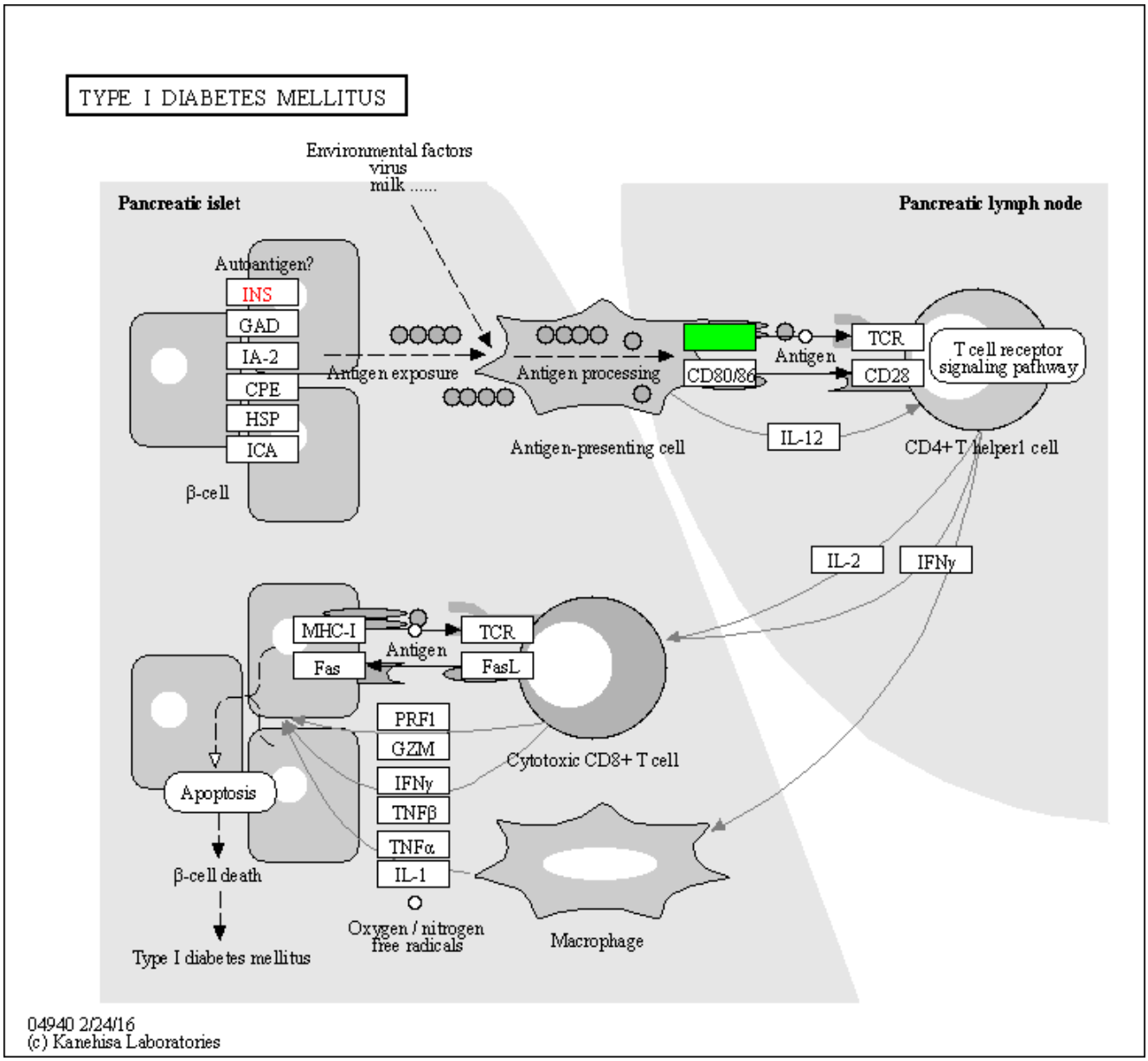

Th1 AND Th2 CELL DIFFERENTIATION
Antigen-presenting cell (APC)
CD4+ T cell
Bacteria, viruses and TLR ligands
Notch signaling pathway
DLL1/4
Notch1
NICD
MAML1
RBPJ
p20
c-Rel
T-Bet
DNA
IFN-γ
IL-12Rb
NK cell
IFNγ
IFNγR
JAK1/2
Jak/STAT signaling pathway
STAT1
IL-12
IL-12R
JAK2
TYK2
STAT4
Th1
Cellular immunity (Inflammation)
Attacks intracellular pathogens
Autoimmunity Inflammatory Diseases
Proper balance Healthy
Balance lost Disease
Allergy Cancer
CD4
Lck
Calcium signaling pathway
IP3
Ca2+
CaN
NFAT
NF-κB signaling pathway
NFκB
IκB
PKCθ
IKK
ZAP70
PLCγ1
DAG
ERK
Fos
MAPK signaling pathway
p38
JNK
Jun
T cell receptor signaling pathway
MHC II
Antigen
CD3ε
CD3γ
TCRα
TCRβ
CD3ζ
CD3δ
CD3ε
Cell adhesion molecules
LAT
RUNX3
IL-2
IL-2R
JAK1/3
STAT5
IL-4
IL-4R
JAK1/3
STAT6
Basophil, memory T cell or naive CD4+ T cell
Helminths, allergens, toxins and prostaglandin E2
Jagged1/2
Notch1/2
Gata3
IL2Rα
IL4Rα
IL-4
IL-2
IL-5
IL-13
c-Maf
Th2
Humoral immunity (Antibody production)
Attacks extracellular pathogens
04658 5/9/17
(c) Kanehisa Laboratories

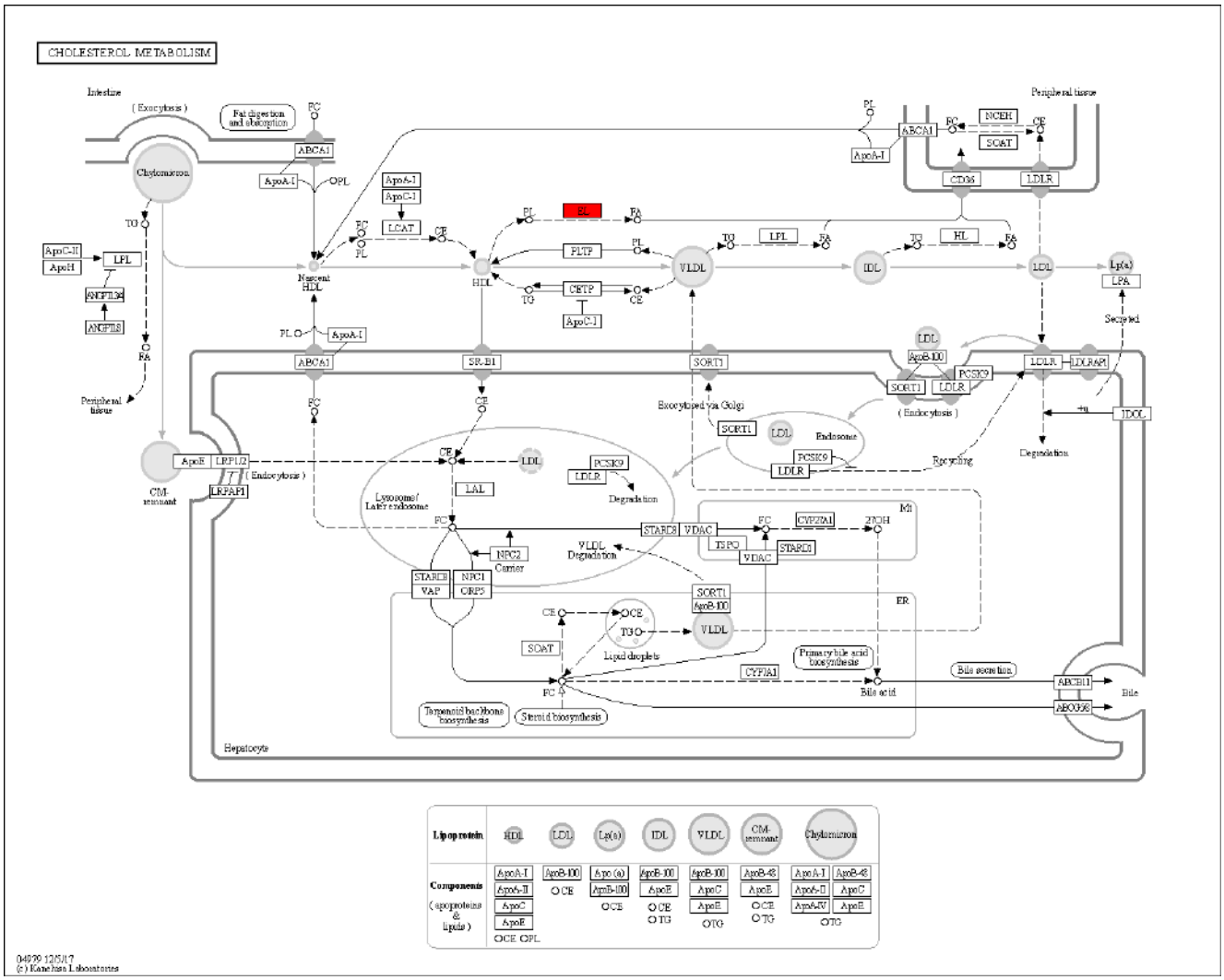

Th17 CELL DIFFERENTIATION
CD4+ T cell
Antigen-presenting cells (APC)
Promoting of Th17 differentiation
IL-1β
IL-1βR
At low concentration
TGFβ
TGFβ-R
Toll-like receptor signaling pathway
IL-21
IL-21R
Jak1/3
IL-6
IL-6R
Jak1/2
IL-23
IL-23R
Jak2
Tyk2
p38
mTOR
SMADs
STAT3
HIF-1 signaling pathway
Hsp90
AHR
AHR ligand (FICZ)
IRF4
RORγt
RORα
HIF-1
IL-23R
IL-17A
IL-17F
IL-21
IL-22
Protection against extracellular pathogens (eg. Fungi, bacteria)
Autoimmunity
Inflammatory bowel disease
Rheumatoid arthritis
CD4
Lck
Calcium signaling pathway
IP3
Ca2+
CaN
NFAT
NF-κB signaling pathway
NFκB
IκB
PKCθ
IKK
ZAP70
PLCγ1
DAG
ERK
Fos
MAPK signaling pathway
p38
JNK
Jun
T cell receptor signaling pathway
MHC II
Antigen
CD3ε
CD3γ
TCRα
TCRβ
CD3ζ
CD3δ
CD3ε
Cell adhesion molecules
LAT
Inhibition of Th17 differentiation
IL-4
IL-4R
Jak1/3
Th2 cell
Jak/STAT signaling pathway
STAT6
GATA3
IL-27
IL-27R
Jak1/2
Tyk2
RUNX1
T-bet
STAT1
IFN-γ
IFNγR
Th1 cell
At high concentration
TGFβ signaling pathway
TGFβ
TGFβ-R
SMADs
IL-2
IL-2R
Jak1/3
STAT5
FOXP3
ATRA
RARα
RXR
04659 12/14/16
(c) Kanehisa Laboratories

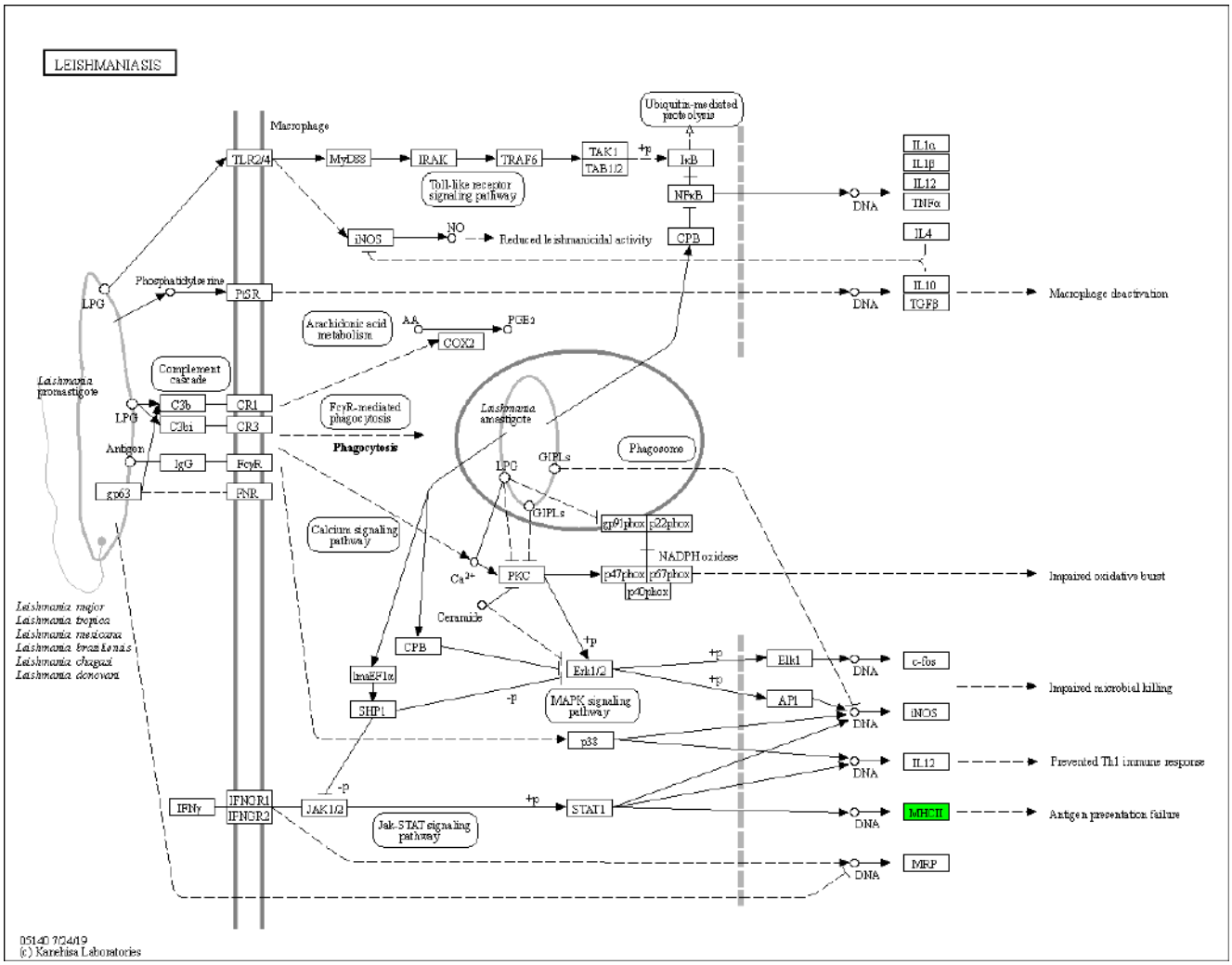

Th17 CELL DIFFERENTIATION
CD4+T cell
Antigen-presenting cells (APC)
Promoting of Th17 differentiation
IL-1β
IL-1R
At low concentration
TGFβ
TGFβ-R
Toll-like receptor signaling pathway
IL-21
IL-21R
IL-6
IL-6R
IL-23
IL-23R
Jak1/3
Jak1/2
Jak2
Tyk2
STAT3
p38
mTOR
SMADs
HIF-1 signaling pathway
Hsp90
AHR
AHR ligand (FICZ)
IRF4
RORγt
RORα
HIF-1
DNA
IL-17A
IL-17F
IL-21
IL-22
Protecting against extracellular pathogens (eg. fungi, bacteria)
Autoimmunity
Inflammatory bowel disease
Rheumatoid arthritis
CD4
Lck
Calcium signaling pathway
IP3
Ca2+
CaN
NFAT
NF-κB signaling pathway
NFκB
IκB
T cell receptor signaling pathway
MHCII
Antigen
CD3ε
CD3γ
TCRα
TCRβ
CD3ζ
CD3ζ
ZAP70
PLCγ1
DAG
PKCθ
IKK
ERK
MAPK signaling pathway
Fos
Jun
p38
JNK
Cell adhesion molecules
LAT
Inhibition of Th17 differentiation
IL-4
IL-4R
Th2 cell
Jak1/3
JaK/STAT signaling pathway
STAT6
GATA3
RUNX1
Tbet
IL-27
IL-27R
Jak1/2
Tyk2
STAT1
IFNγ
IFNγR
Th1 cell
At high concentration
TGFβ signaling pathway
SMADs
IL-2
IL-2R
STAT5
FOXP3
ATRA
RARα
RXR
04659 12/14/16
(c) Kanehisa Laboratories

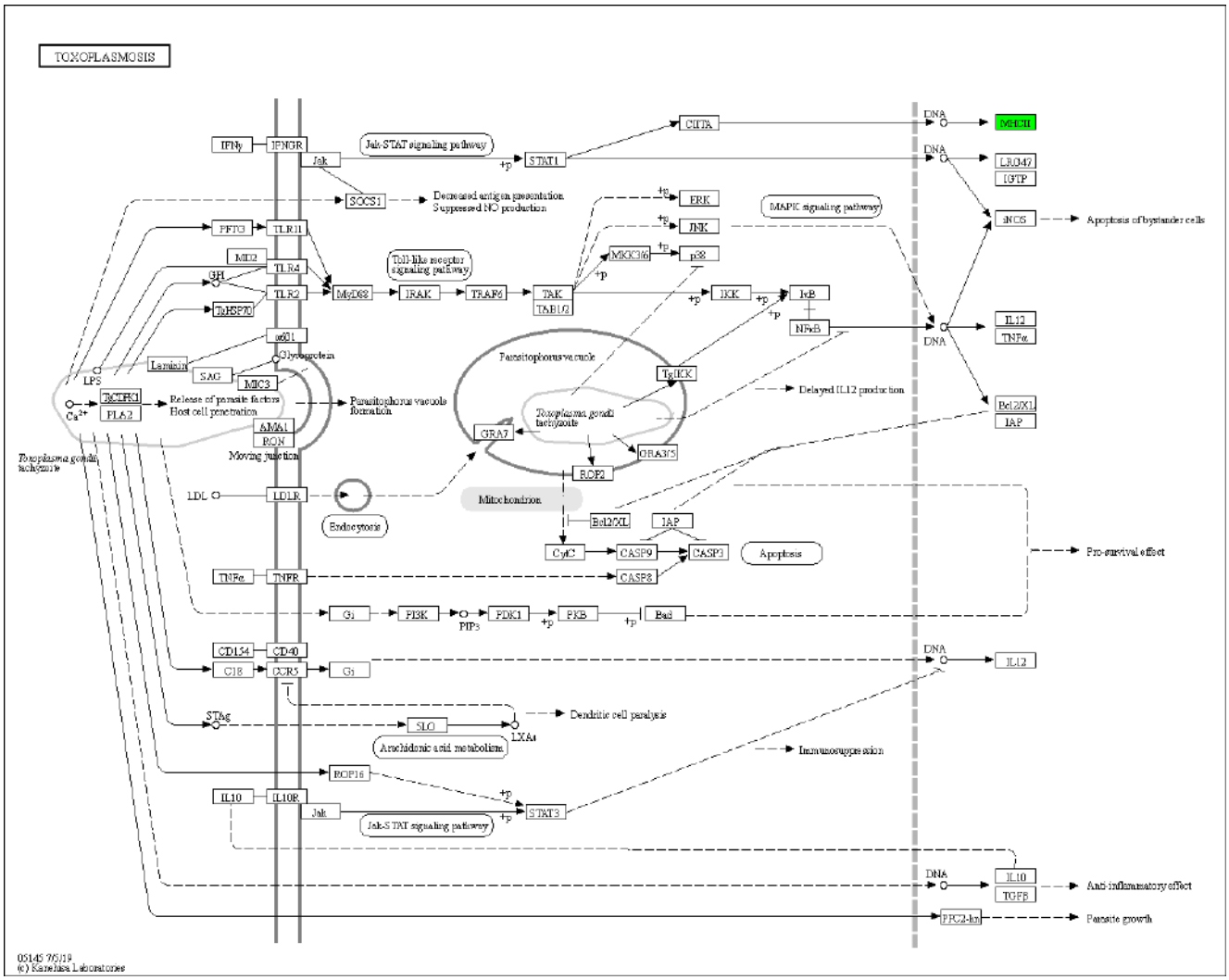

INTESTINAL IMMUNE NETWORK FOR IgA PRODUCTION
Bacteria
Antigen
Antigen
Epithelial cell
M cell
Inductive site
DC
Isolated lymphoid follicles
IL-2
IL-4
IL-5
IL-6
IL-10
TGFB
B7
MHC classII
APRIL
BAFF
TACI
BCR
BCMA
AID
BAFFR
B1 cell
RA
CCL28
IgA+B cell
Antigen processing and presentation
CD28
TCR
T cell receptor signaling pathway
MHC
CD40
CD40L
ICOS
ICOS-L
CD4+ T cell
B2 cell
Peyer's patch
B cell receptor signaling pathway
Lamina propria
Vascular system
High endothelial venules
CXCR4
CCR10
CCR9
CXCL12
CCL28
CCL25
IgA+plasmablast
Recirculation
Leukocyte transendothelial migration
Homing
Terminal B cell differentiation
IgA-plasma cell
SIgA
Gut lumen
pIgR
IgA
Effector site
Lamina propria
IL-5
IL-6
IL-10
Lamina propria stroma cell
IL-15R
IL-15
LTBR
NIK
Lamina propria stroma cell
04672 9/2/10
© Kanehisa Laboratories

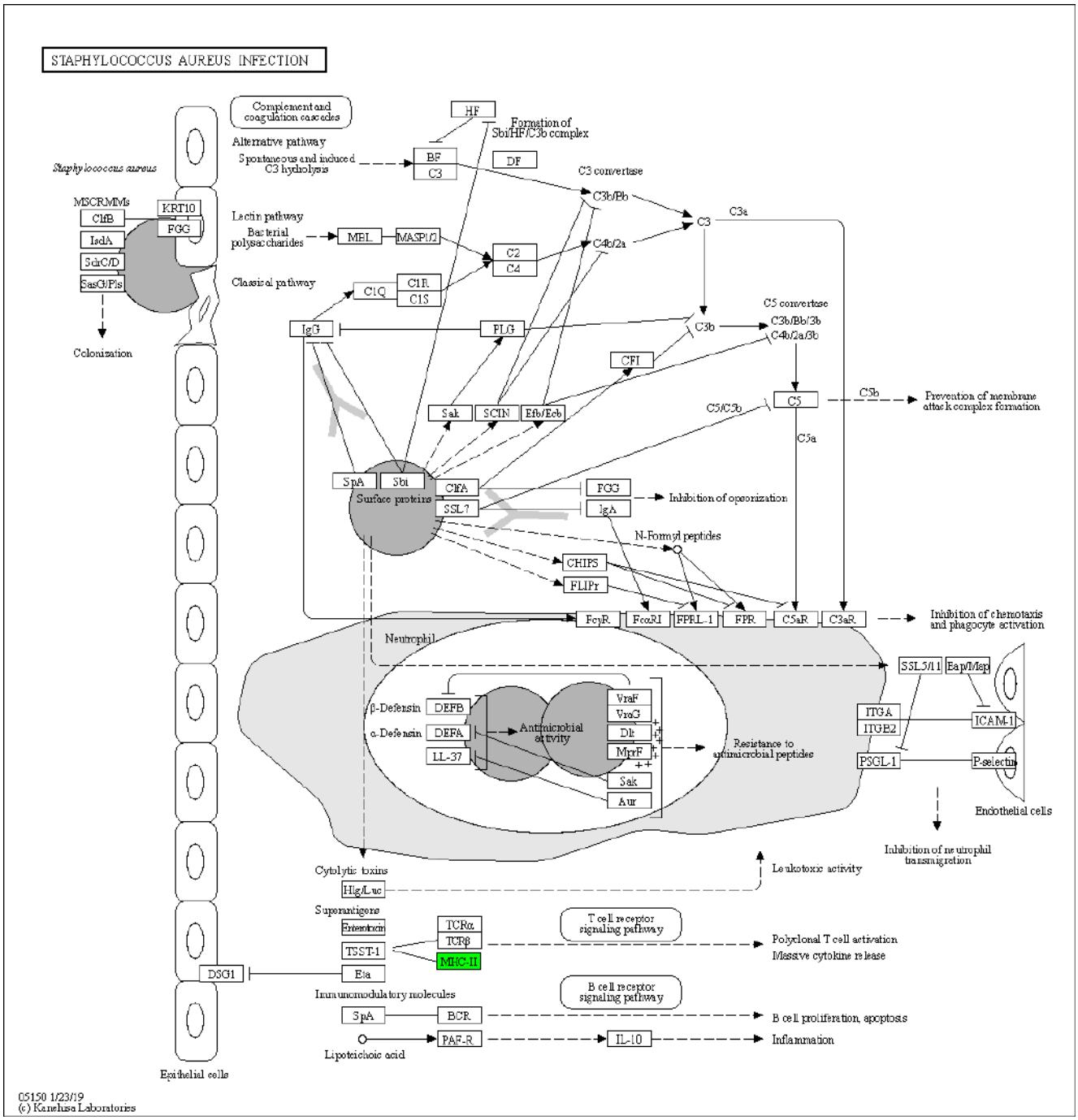

PROLACTIN SIGNALING PATHWAY
PI3K-Akt signaling pathway
MAPK signaling pathway
Jak-STAT signaling pathway
Estrogen signaling pathway
Dopaminergic synapse
Insulin secretion
PRL
short R
long R
JAK2
SHC
Grb2
SOS
Ras
Raf-1
MEK
ERK
JNK
p38
c-Src
PI3K
PIP3
Akt/PKB
Foxo3a
GSK-3β
STAT
c-Fos
CyclinD1
Cis
SOCS
Rankl
Rank
NF-κB
CyclinD1
Wap
Elf5
Caseins
Milk proteins
LH
LH-R
Ovarian steroidogenesis
P4
A
E1
E2
P450C17
ERα/β
IRF-1
Immune responses
TH
Tyrosine
Dopamine
L-Dopa
PRLR
Glut-2
GCK
Glucose
Glucose-6-P
ATP
CyclinD2
Insulin
Proliferation
(Ovary)
Galt-1
Ovarian defect
(Mammary epithelial cells)
(Corpus luteum)
(Tuberoinfundibular neurons)
(Pancreatic β-cells)
(T and B lymphocytes, macrophages, fibroblasts)
Negative feedback
DNA
04917 9/28/15
(c) Kanehisa Laboratories

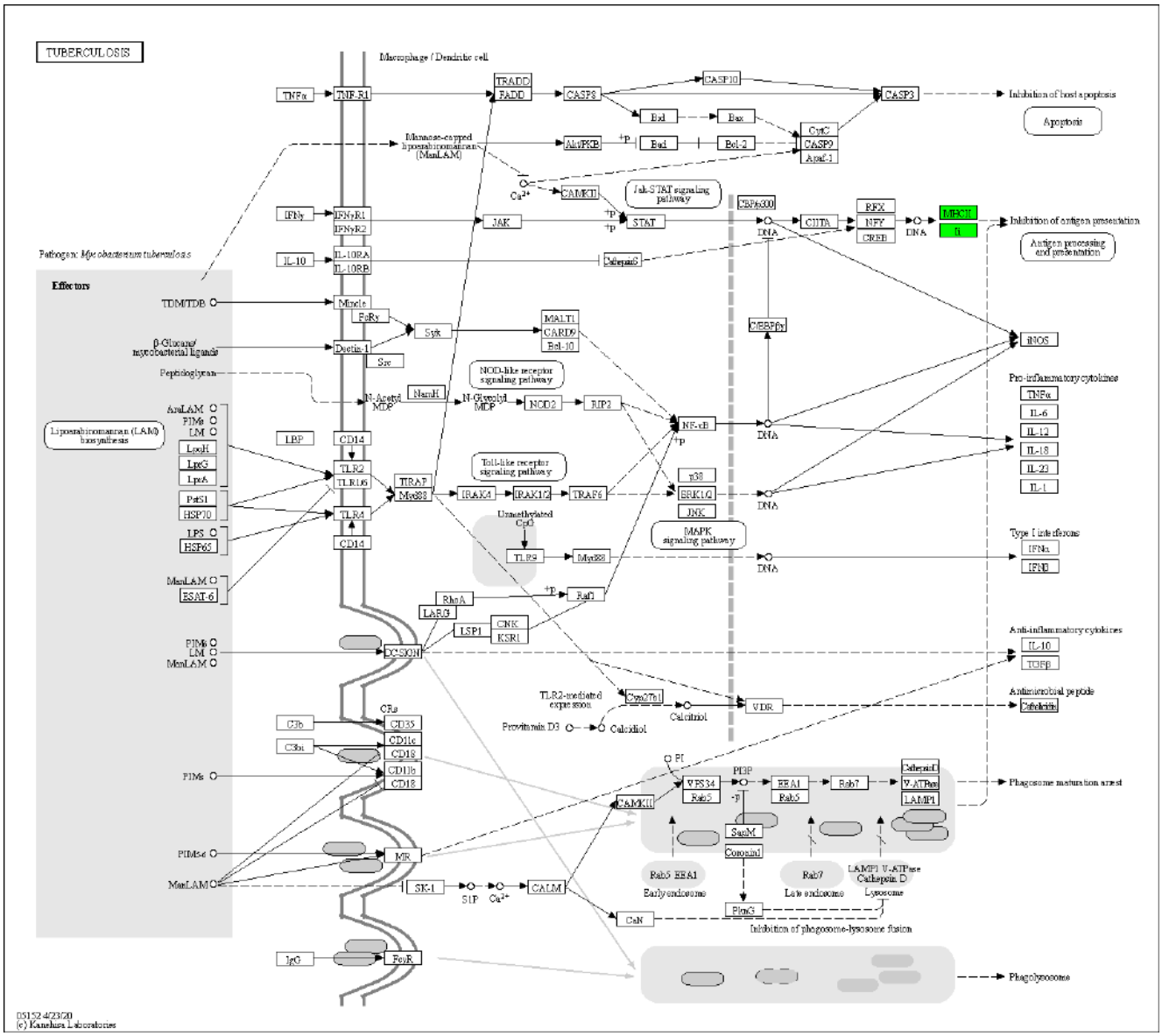

TYPE I DIABETES MELLITUS
Environmental factors
virus
milk .....
Pancreatic islet
Pancreatic lymph node
Autoantigen?
INS
GAD
IA-2
CPE
HSP
ICA
β-cell
Antigen exposure
Antigen processing
Antigen-presenting cell
CD80/86
Antigen
TCR
CD28
T cell receptor signaling pathway
IL-12
CD4+ T helper1 cell
IL-2
IFNγ
MHC-I
Antigen
TCR
Fas
FasL
PRF1
GZM
IFNγ
TNFβ
TNFα
IL-1
Cytotoxic CD8+ T cell
Apoptosis
β-cell death
Type I diabetes mellitus
Oxygen / nitrogen
free radicals
Macrophage
04940 2/24/16
(c) Kanehisa Laboratories

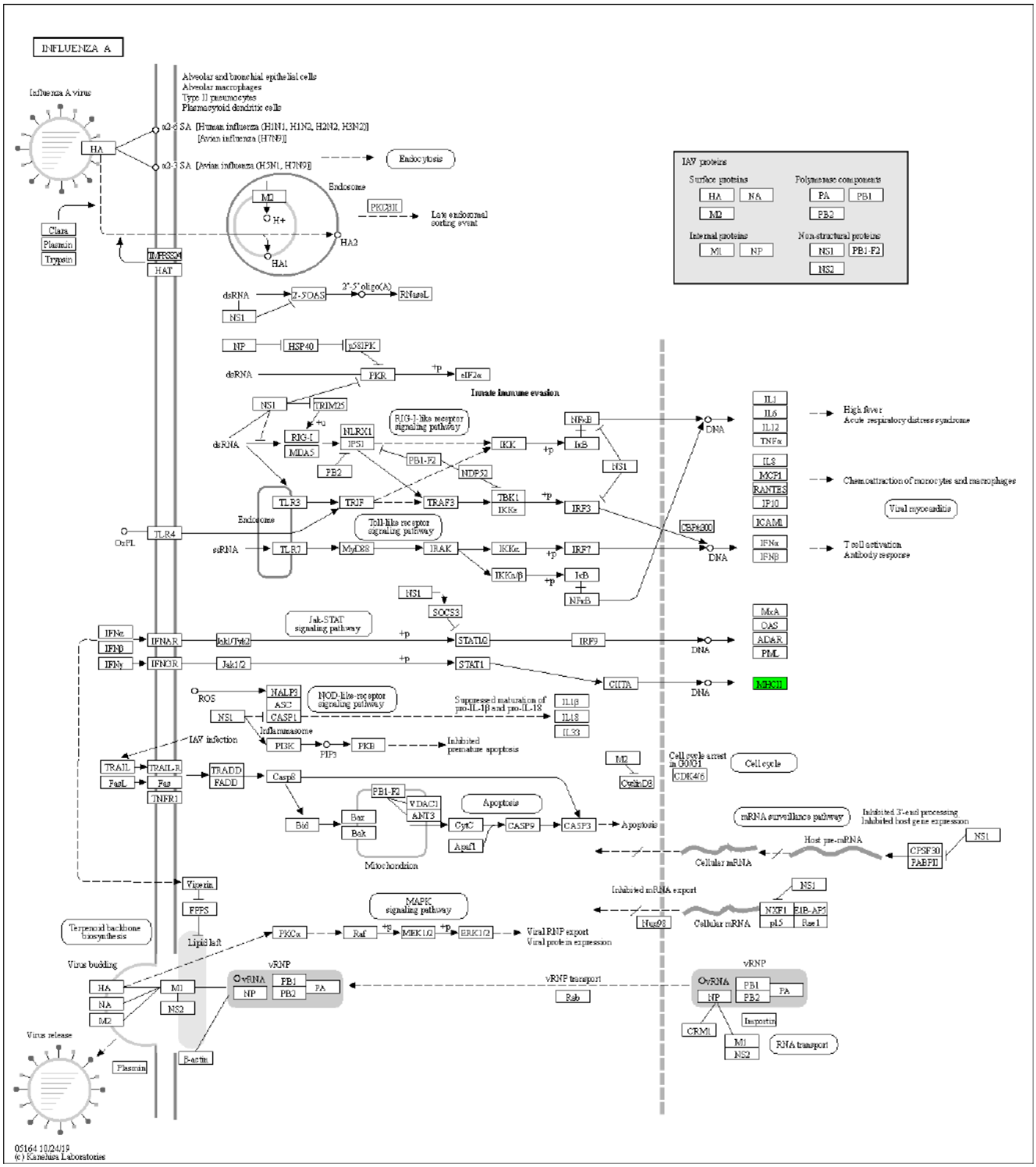

CHOLESTEROL METABOLISM
Intestine
Peripheral tissue
Hepatocyte
Chylomicron
Nascent HDL
HDL
VLDL
IDL
LDL
Lp(a)
CM remnant
Bile
Bile acid
Lipid droplets
Endosome
Lysosome / Late endosome
ER
Golgi
Mt
Lipoprotein
Components (apoproteins & lipids)

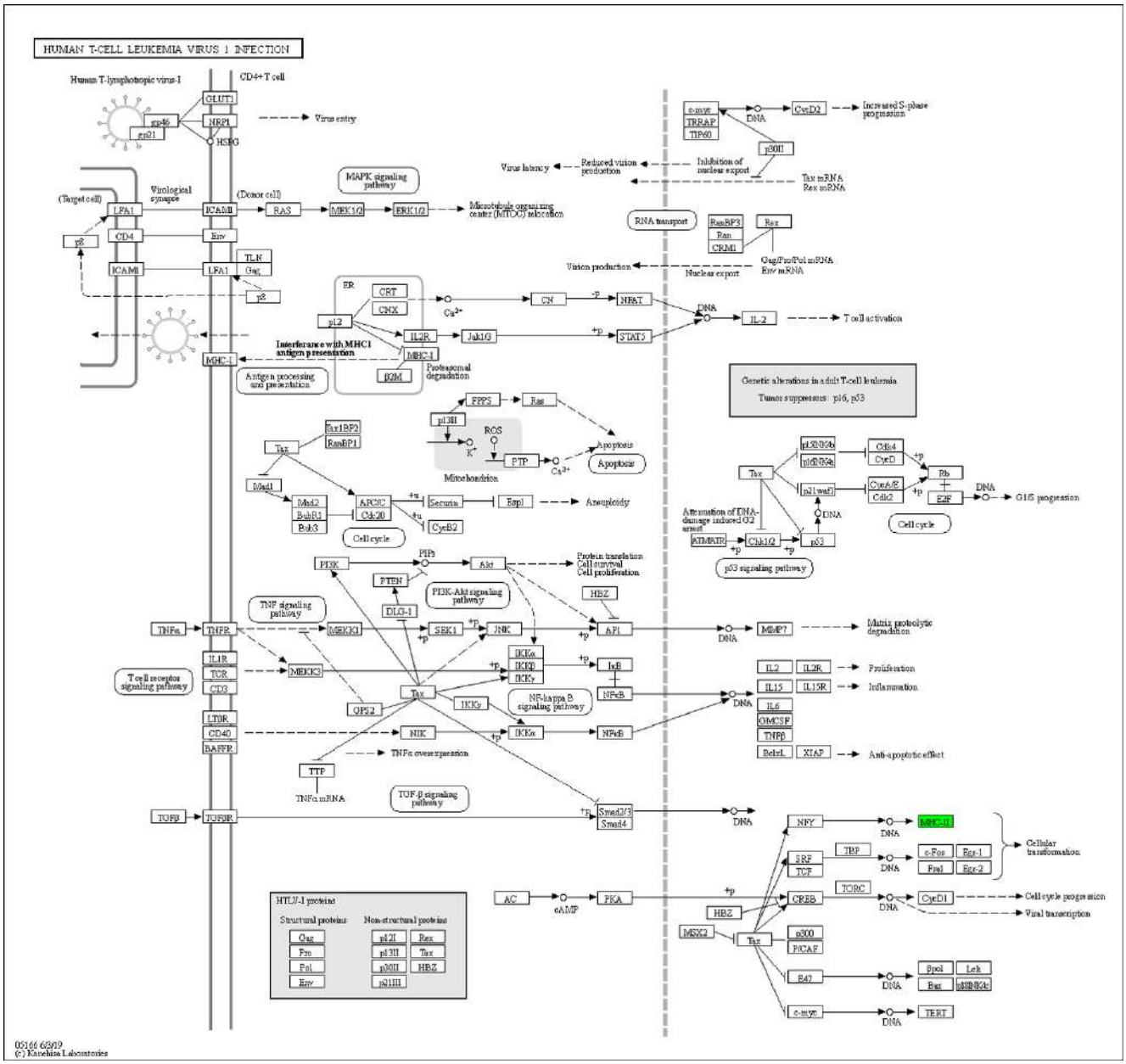

LEISHMANIASIS
Macrophage
TLR2/4
MyD88
IRAK
TRAF6
TAK1
TAB1/2
Toll-like receptor signaling pathway
Ubiquitin-mediated proteolysis
IκB
NFκB
DNA
IL1α
IL1β
IL12
TNFα
IL4
IL10
TGFβ
Macrophage deactivation
iNOS
NO
Reduced leishmanicidal activity
CPB
LPG
Phosphatidylserine
PSR
Arachidonic acid metabolism
AA
PGE2
COX2
Complement cascade
Leishmania promastigote
C3b
C3bi
CR1
CR3
FcγR-mediated phagocytosis
Phagocytosis
Leishmania amastigote
Antigen
IgG
FcγR
gp63
PNR
LPG
GIPLs
Phagosome
Calcium signaling pathway
gp91phox
p22phox
NADPH oxidase
Ca2+
PKC
p47phox
p67phox
p40phox
Impaired oxidative burst
Leishmania major
Leishmania tropica
Leishmania mexicana
Leishmania braziliensis
Leishmania chagasi
Leishmania donovani
Ceramide
CPB
eIF1α
SHP1
Erk1/2
MAPK signaling pathway
Elk1
AP1
c-fos
Impaired microbial killing
iNOS
p38
IL12
Prevented Th1 immune response
IFNγ
IFNGR1
IFNGR2
JAK1/2
STAT1
Jak-STAT signaling pathway
MHC II
Antigen presentation failure
MRP
05140 7/24/19
(c) Kanehisa Laboratories

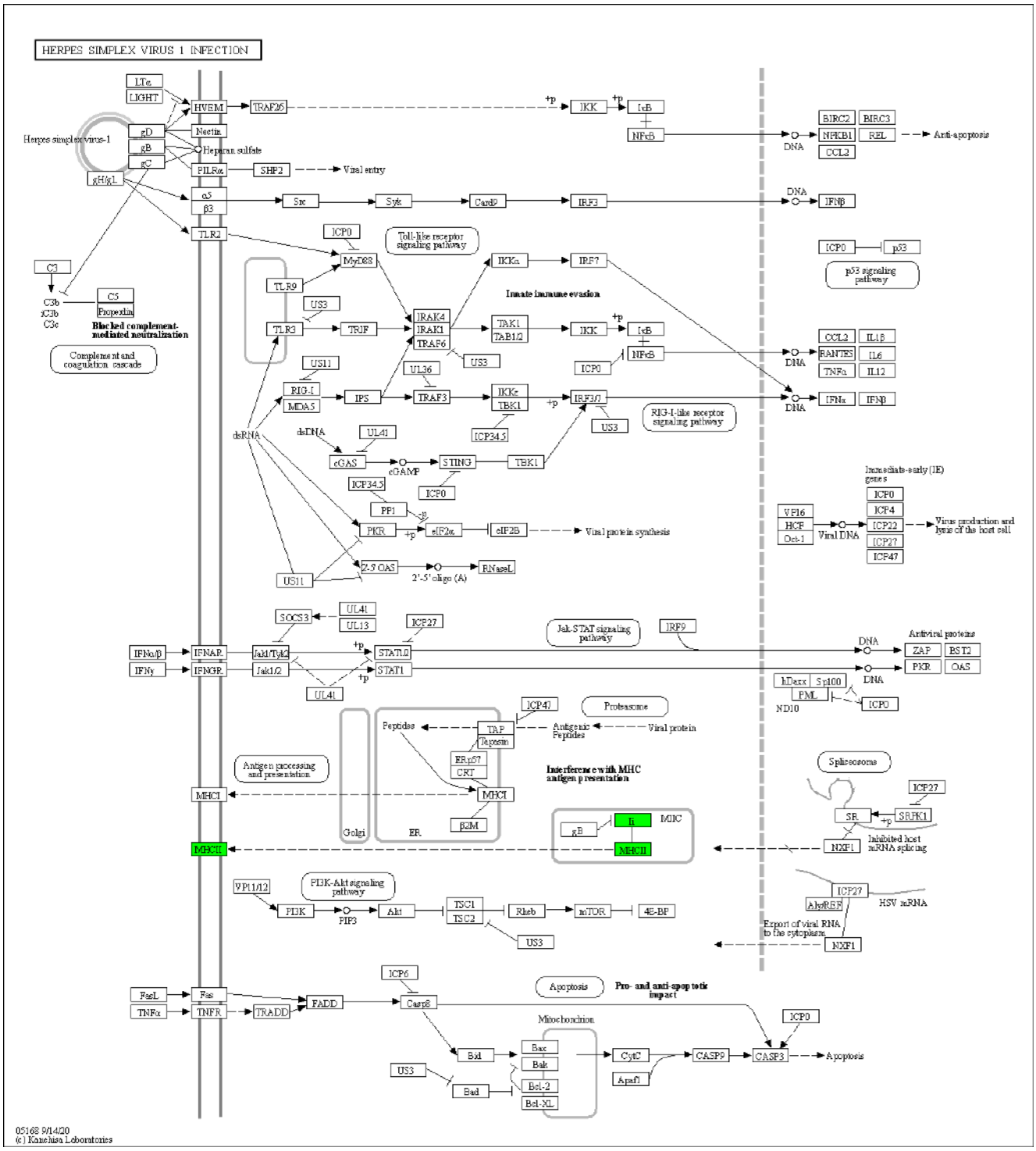

TOXOPLASMOSIS
IFNγ
IFNGR
Jak
Jak-STAT signaling pathway
STAT1
CIITA
DNA
MHCII
IL10/47
IGTP
SOCS1
Decreased antigen presentation
Suppressed NO production
ERK
JNK
MAPK signaling pathway
iNOS
Apoptosis of bystander cells
PFTG
TLR11
MD2
TLR4
GPI
TLR2
TgHSP70
Toll-like receptor signaling pathway
MyD88
IRAK
TRAF6
TAK
TAB1/2
MKK3/6
p38
IKK
IκB
NFκB
IL12
TNFα
Bcl2/X1
IAP
α6β1
Glycoprotein
LPS
Laminin
SAG
MIC3
Parasitophorous vacuole
Toxoplasma gondii tachyzoite
TgIEK
Delayed IL12 production
GRA7
ORA3/5
ROP2
Ca2+
BCLIFK1
FLA2
Release of parasite factors
Host cell penetration
Parasitophorous vacuole formation
AMA1
RON
Moving junction
LDL
LDLR
Endocytosis
Mitochondrion
Bcl2/XL
IAP
CytC
CASP9
CASP3
Apoptosis
Pro-survival effect
TNFα
TNFR
CASP8
G1
PI3K
PIP3
PDK1
PKB
Bad
CD154
CD40
C18
CCR5
G1
IL12
STAg
SLO
LXAs
Dendritic cell paralysis
Arachidonic acid metabolism
Immunosuppression
ROP16
IL10
IL10R
Jak
STAT3
Jak-STAT signaling pathway
IL10
TGFβ
Anti-inflammatory effect
PPC2-kn
Parasite growth
05145 7/5/19
(c) Kanehisa Laboratories

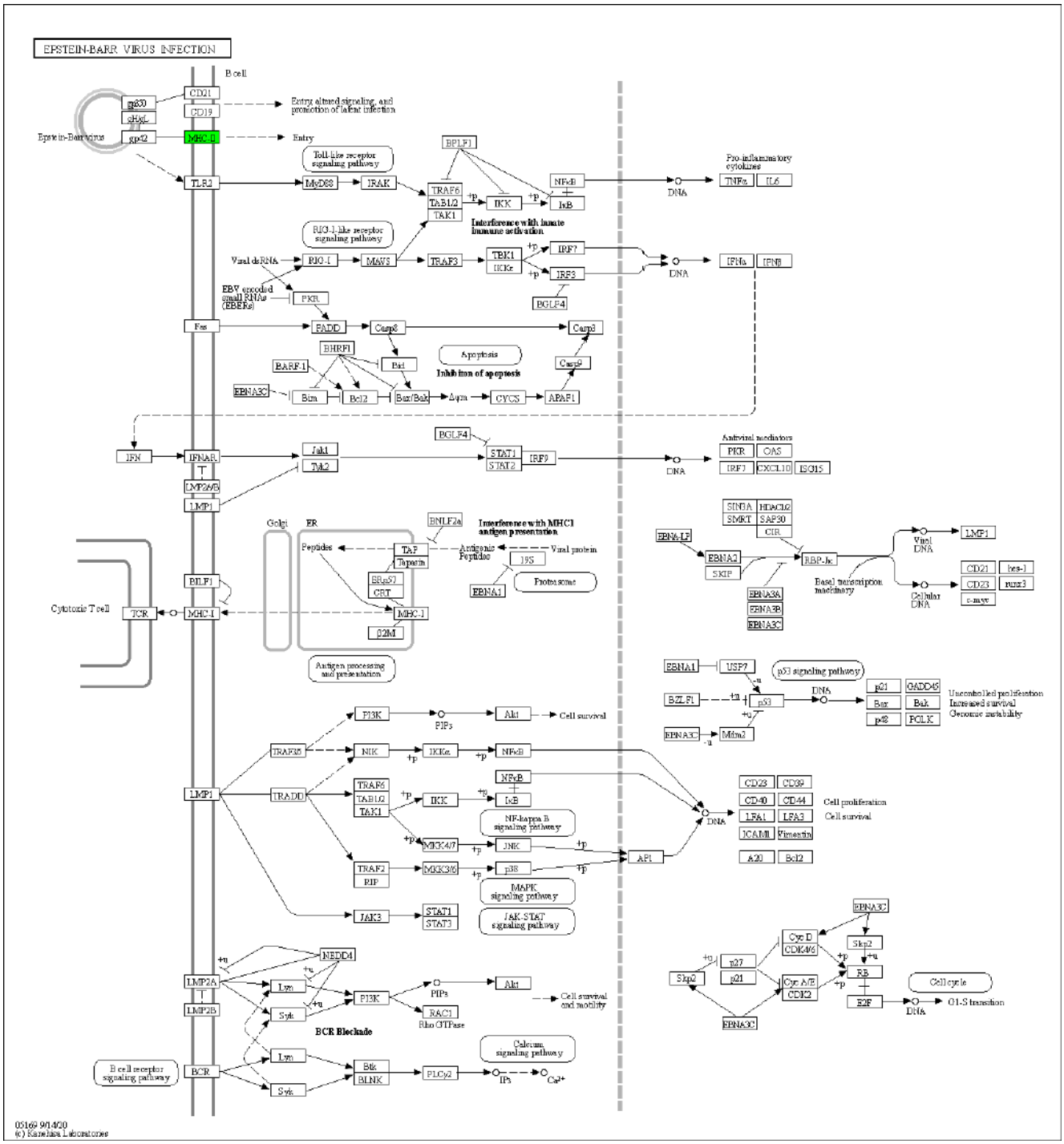

STAPHYLOCOCCUS AUREUS INFECTION
Complement and coagulation cascades
Alternative pathway
Spontaneous and induced C3 hydrolysis
Staphylococcus aureus
MSCRAMMs
ClfB
IsdA
SdrC/D
SasG/Pls
KRT10
FGG
Colonization
Lectin pathway
Bacterial polysaccharides
MBL
MASP1/2
Classical pathway
C1Q
C1R
C1S
IgG
HF
Formation of SbiHF/C3b complex
BF
C3
DF
C3 convertase
C3b/Bb
C3
C3a
C2
C4
C4b/2a
C5 convertase
C3b/Bb/3b
C4b/2a/3b
C3b
PLG
CFI
Sak
SCIN
Efb/Ecb
C5/C3b
C5
C5b
Prevention of membrane attack complex formation
C5a
SpA
Sbi
Surface proteins
ClfA
SSL7
FCG
IgA
Inhibition of opsonization
N-Formyl peptides
CHIPS
FLIPr
FcγR
FcαRI
FPRL-1
FPR
C5aR
C3aR
Inhibition of chemotaxis and phagocyte activation
Neutrophil
β-Defensin
α-Defensin
DEFB
DEFA
LL-37
Antimicrobial activity
VraF
VraG
Dlt
MprF
Sak
Aur
Resistance to antimicrobial peptides
SSL5/11
EapM/Map
ITGA
ITGB2
PSGL-1
ICAM-1
P-selectin
Endothelial cells
Inhibition of neutrophil transmigration
Cytolytic toxins
Hlg/Luc
Leukotoxic activity
Superantigens
Enterotoxin
TSST-1
Eta
DSG1
TCRα
TCRβ
MHC-II
T cell receptor signaling pathway
Polyclonal T cell activation
Massive cytokine release
Immunomodulatory molecules
SpA
BCR
B cell receptor signaling pathway
B cell proliferation, apoptosis
PAF-R
IL-10
Inflammation
Lipoteichoic acid
Epithelial cells
05150 1/23/19
(c) Kanehisa Laboratories

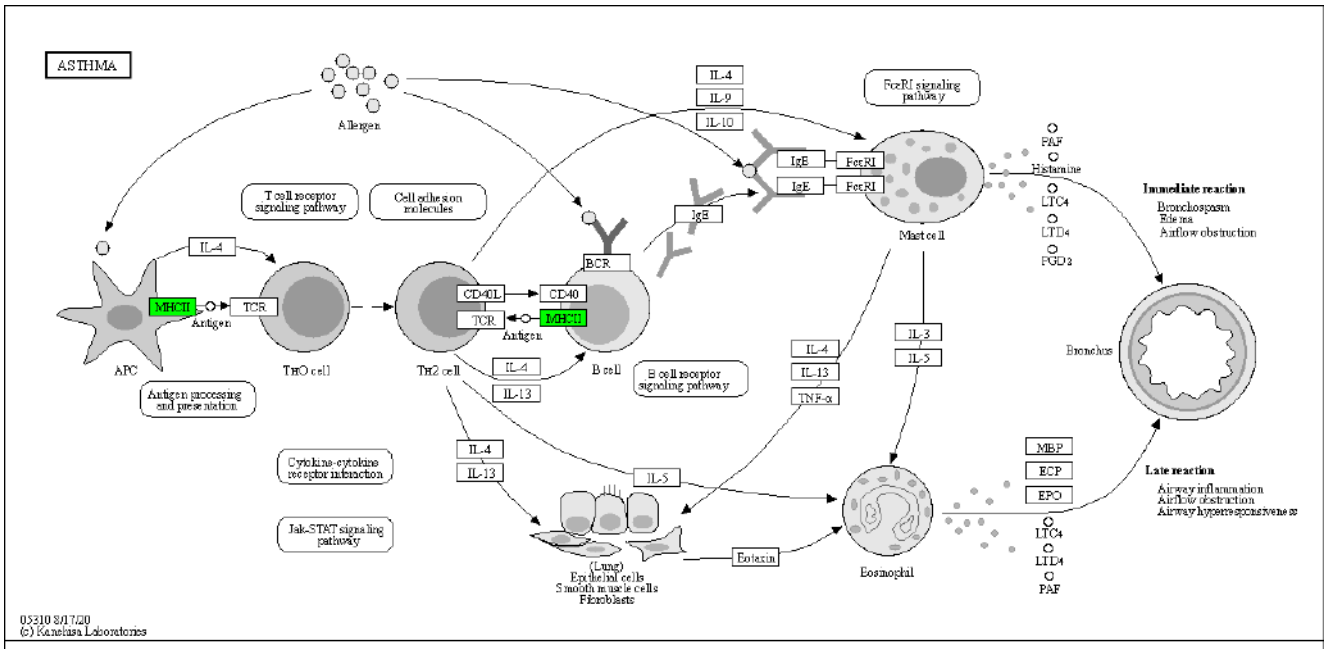

TUBERCULOSIS
Macrophage / Dendritic cell
Pathogen: Mycobacterium tuberculosis
Effectors
TNFα
TNF-R1
TRADD
FADD
CASP8
CASP10
CASP3
Inhibition of host apoptosis
Apoptosis
Bad
Bax
Bcl-2
CytC
CASP9
Apaf-1
AktPKB
Mannose-capped lipoarabinomannan (ManLAM)
Ca2+
CAMKII
Jak-STAT signaling pathway
IFNγ
IFNγR1
IFNγR2
JAK
STAT
CBP/p300
CIITA
RFX
NFY
CREB
MHCII
Ii
Inhibition of antigen presentation
Antigen processing and presentation
IL-10
IL-10RA
IL-10RB
Cathepsin S
TDM/TDB
Mincle
FcRγ
Syk
MALT1
CARD9
Bcl-10
C/EBPβγ
iNOS
β-Glucans/ mycobacterial ligands
Dectin-1
Src
Peptidoglycan
NOD-like receptor signaling pathway
N-Acetyl MDP
NanH
N-Glycolyl MDP
NOD2
RIP2
NF-κB
Pro-inflammatory cytokines
TNFα
IL-6
IL-12
IL-18
IL-23
IL-1
AraLAM
PIMs
LM
Lipoarabinomannan (LAM) biosynthesis
LBP
CD14
LpqH
LprG
LprA
PstS1
HSP70
TLR2
TLR1/6
TLR4
TRAF
Myd88
Toll-like receptor signaling pathway
IRAK4
IRAK1/2
TRAF6
p38
ERK1/2
JNK
MAPK signaling pathway
LPS
HSP65
Unmethylated CpG
TLR9
Myd88
DNA
Type I interferons
IFNα
IFNβ
ManLAM
ESAT-6
RhoA
LARG
LSP1
CNK
KSR1
Raf1
PIMs
LM
ManLAM
ICOSLG
Anti-inflammatory cytokines
IL-10
TGFβ
TLR2-mediated expression
Cyp27b1
VDR
Antimicrobial peptide
Cathelicidin
Provitamin D3
Calcidiol
Calcitriol
CRs
C3b
C3bi
CD35
CD16
CD18
CD11b
CD18
PIMs
PI
PI3P
VPS34
Rab5
EEA1
Rab5
Rab7
Cathepsin L
V-ATPase
LAMP1
Phagosome maturation arrest
CAMKII
SapM
Coronin1
PIMs
MR
ManLAM
SK-1
S1P
Ca2+
CALM
Rab5 EEA1
Early endosome
Rab7
Late endosome
LAMP1 V-ATPase Cathepsin D
Lysosome
CaN
PtpS
Inhibition of phagosome-lysosome fusion
IgG
FcγR
Phagolysosome
05152 4/23/20
(c) Kanehisa Laboratories

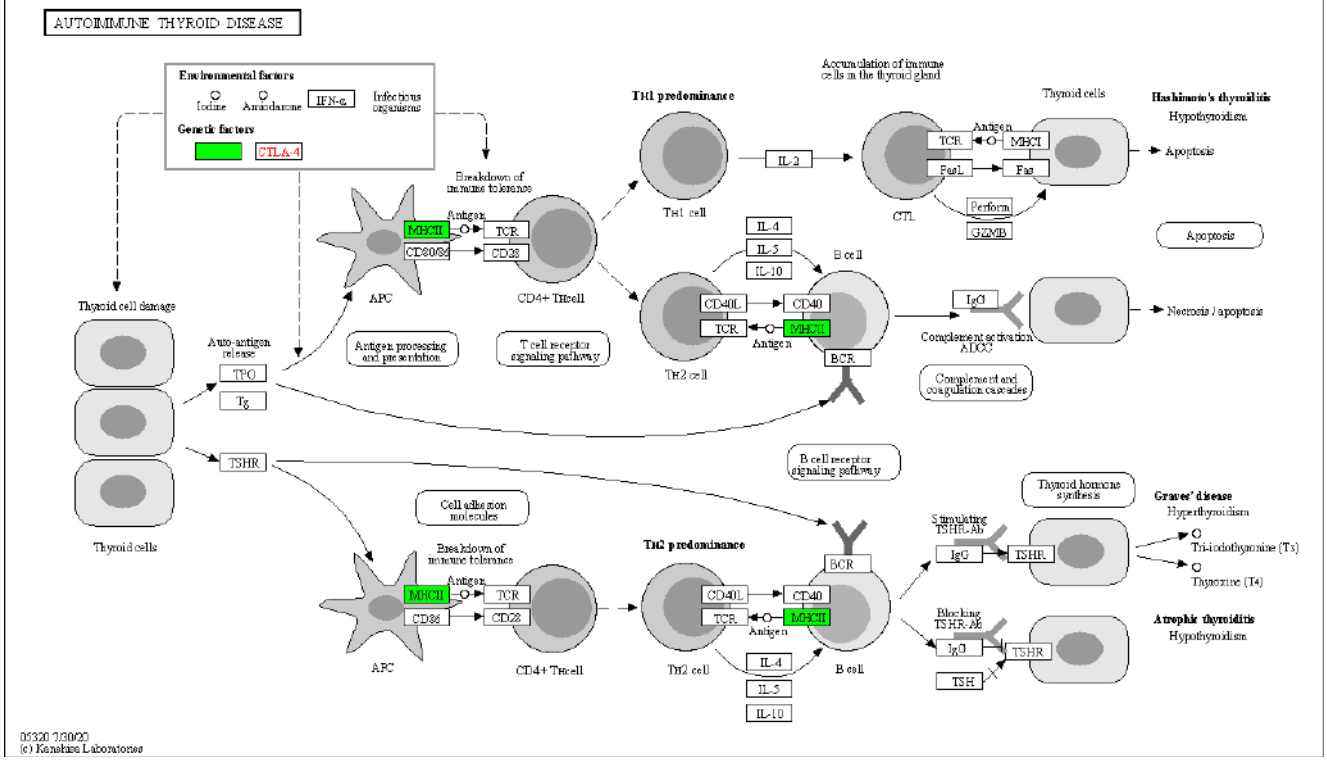

INFLUENZA A
Influenza A virus
Alveolar and bronchial epithelial cells
Alveolar macrophages
Type II pneumocytes
Plasmacytoid dendritic cells
HA
Endocytosis
Endosome
M2
H+
HA2
HA1
Clara
Plasmin
Trypsin
TMPRSS4
HAT
PKCδII
Late endosomal sorting event
IAV proteins
Surface proteins
HA
NA
M2
Polymerase components
PA
PB1
PB2
Internal proteins
M1
NP
Non-structural proteins
NS1
PB1-F2
NS2
dsRNA
2-5OAS
2',5'-oligo(A)
RNaseL
NS1
NF
HSP40
p58IPK
PKR
eIF2α
Innate immune evasion
TRIM25
RIG-I-like receptor signaling pathway
RIG-I
MDA5
NLRX1
IPS1
PB1-F2
PB2
NDP52
IKK
IκB
NFκB
NS1
DNA
IL1
IL6
IL12
TNFα
High fever
Acute respiratory distress syndrome
IL8
MCP1
RANTES
IP10
ICAM1
Chemoattraction of monocytes and macrophages
Viral myocarditis
TLR3
TRIF
TRAF3
TBK1
IKKε
IRF3
CBP/p300
IFNα
IFNβ
T cell activation
Antibody response
Endosome
CpG DNA
TLR4
ssRNA
TLR7
MyD88
IRAK
IKKε
IRF7
IKKα/β
IκB
NFκB
NS1
SOCS3
Jak-STAT signaling pathway
IFNα
IFNβ
IFNλ
IFNAR
Jak1/Tyk2
STAT1/2
IRF9
IFNGR
Jak1/2
STAT1
CIITA
MHC II
MxA
OAS
ADAR
PML
ROS
NALP3
ASC
CASP1
NOD-like receptor signaling pathway
Inflammasome
Suppressed maturation of pro-IL-1β and pro-IL-18
IL1β
IL18
IL33
NS1
IAV infection
PI3K
PIP3
PKB
Inhibited premature apoptosis
M2
Cell cycle arrest in G0/G1
CDK4/6
CyclinD3
Cell cycle
TRAIL
FasL
TRAIL-R
Fas
TNFR1
TRADD
FADD
Casp8
PB1-F2
VDAC1
ANT3
Bid
Bax
Bak
Apoptosis
CytC
Apaf1
CASP9
CASP3
Apoptosis
mRNA surveillance pathway
Inhibited 3'-end processing
Inhibited host gene expression
Host pre-mRNA
CPSF30
PABPII
NS1
Cellular mRNA
Mitochondrion
Viperin
FPPS
Terpenoid backbone biosynthesis
Lipid raft
MAPK signaling pathway
PKCα
Raf
MEK1/2
ERK1/2
Viral RNP export
Viral protein expression
Inhibited mRNA export
NS1
NXF1
E1B-AP5
p15
Rae1
Nup98
Cellular mRNA
Virus budding
HA
NA
M2
M1
NS2
vRNP
vRNA
NP
PB1
PB2
PA
vRNP transport
Rab
vRNP
vRNA
NP
PB1
PB2
PA
Importin
CRM1
M1
NS2
RNA transport
Virus release
Plasmin
β-actin
05164 10/24/19
(c) Kanehisa Laboratories

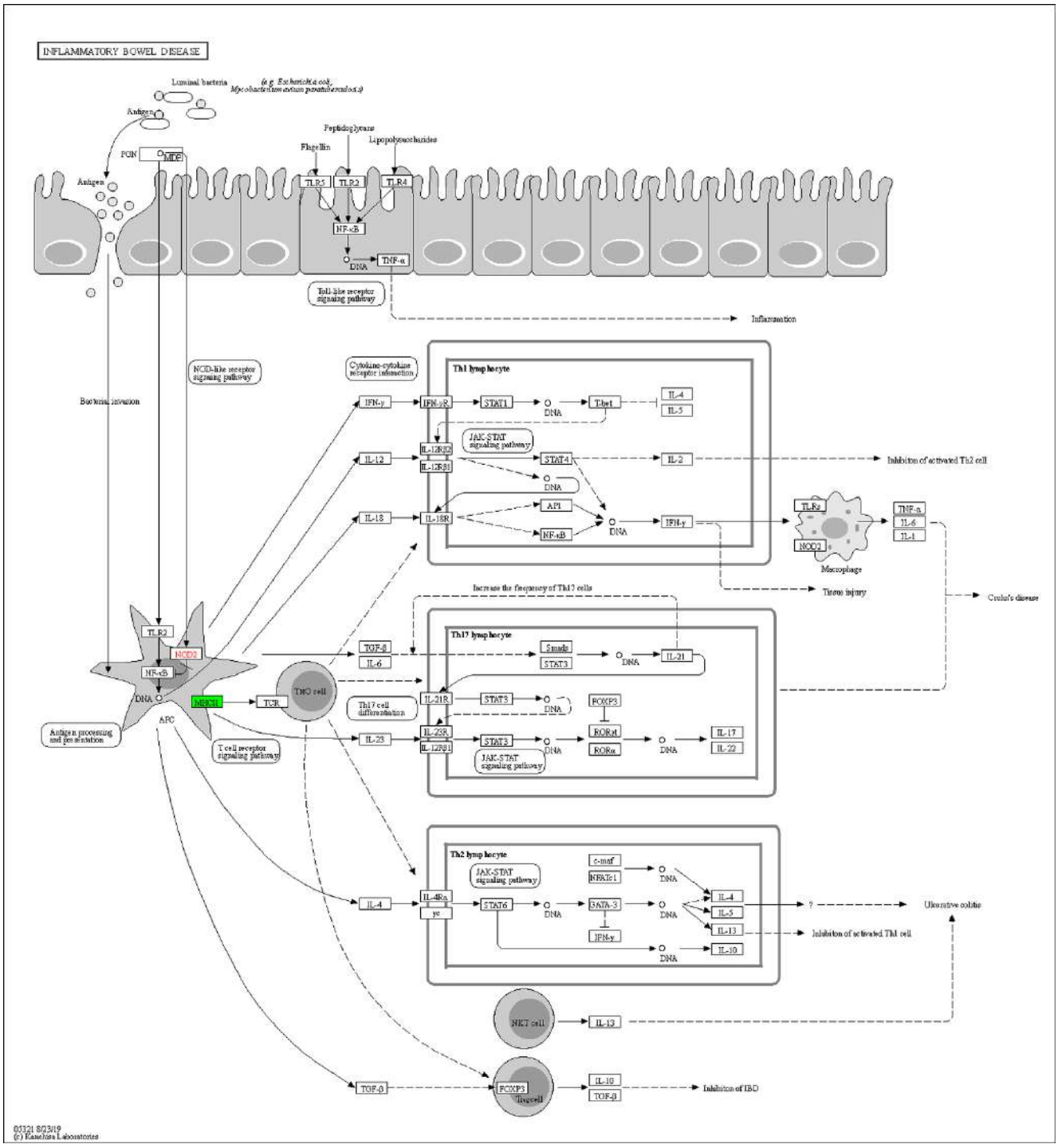

HUMAN T-CELL LEUKEMIA VIRUS 1 INFECTION
Human T-lymphotropic virus-I
CD4+ T cell
Virus entry
Virological synapse
(Target cell)
(Donor cell)
MAPK signaling pathway
RAS
MEK1/2
ERK1/2
Microtubule organizing center (MTOC) formation
Virus latency
Reduced virion production
Inhibition of nuclear export
Tax mRNA
Rex mRNA
Increased S-phase progression
RNA transport
Virion production
Nuclear export
Gag/Pro/Pol mRNA
Env mRNA
Interference with MHC1 antigen presentation
Antigen processing and presentation
Proteasomal degradation
T cell activation
Genetic alterations in adult T-cell leukemia
Tumor suppressors: p16, p53
Apoptosis
Mitochondrion
Aneuploidy
Cell cycle
G1/S progression
p53 signaling pathway
Protein translation
Cell survival
Cell proliferation
PI3K-Akt signaling pathway
TNF signaling pathway
T cell receptor signaling pathway
NF-kappa B signaling pathway
Proliferation
Inflammation
Anti-apoptotic effect
TNFα overexpression
TGF-β signaling pathway
Matrix proteolytic degradation
Cellular transformation
Cell cycle progression
Viral transcription
HTLV-I proteins
Structural proteins
Non-structural proteins
Gag
Pro
Pol
Env
p12I
p13II
p30II
p21III
Rex
Tax
HBZ
05166 6/8/19
(c) Kanehisa Laboratories

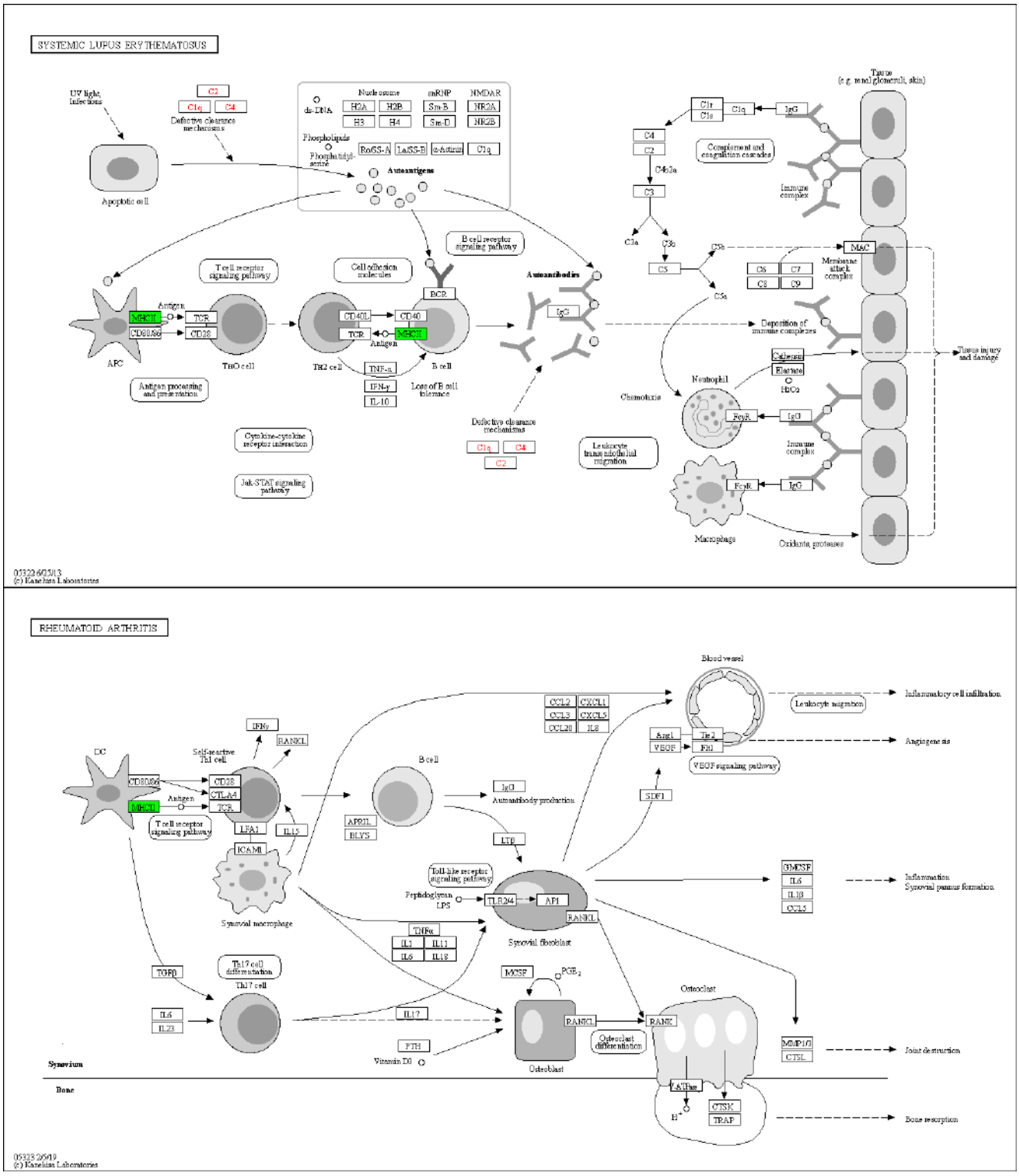

HERPES SIMPLEX VIRUS 1 INFECTION
Herpes simplex virus-1
Blocked complement-mediated neutralization
Complement and coagulation cascade
Toll-like receptor signaling pathway
Innate immune evasion
RIG-I-like receptor signaling pathway
Jak-STAT signaling pathway
Antigen processing and presentation
Interference with MHC antigen presentation
PI3K-Akt signaling pathway
Apoptosis
Pro- and anti-apoptotic impact
p53 signaling pathway
Spliceosome
Proteasome
Mitochondrion
Golgi
ER
Viral protein synthesis
Viral DNA
Virus production and lysis of the host cell
Inhibited host mRNA splicing
HSV mRNA
Export of viral RNA to the cytoplasm
Apoptosis
Anti-apoptosis
Viral entry
05168 9/14/20
(c) Kanehisa Laboratories

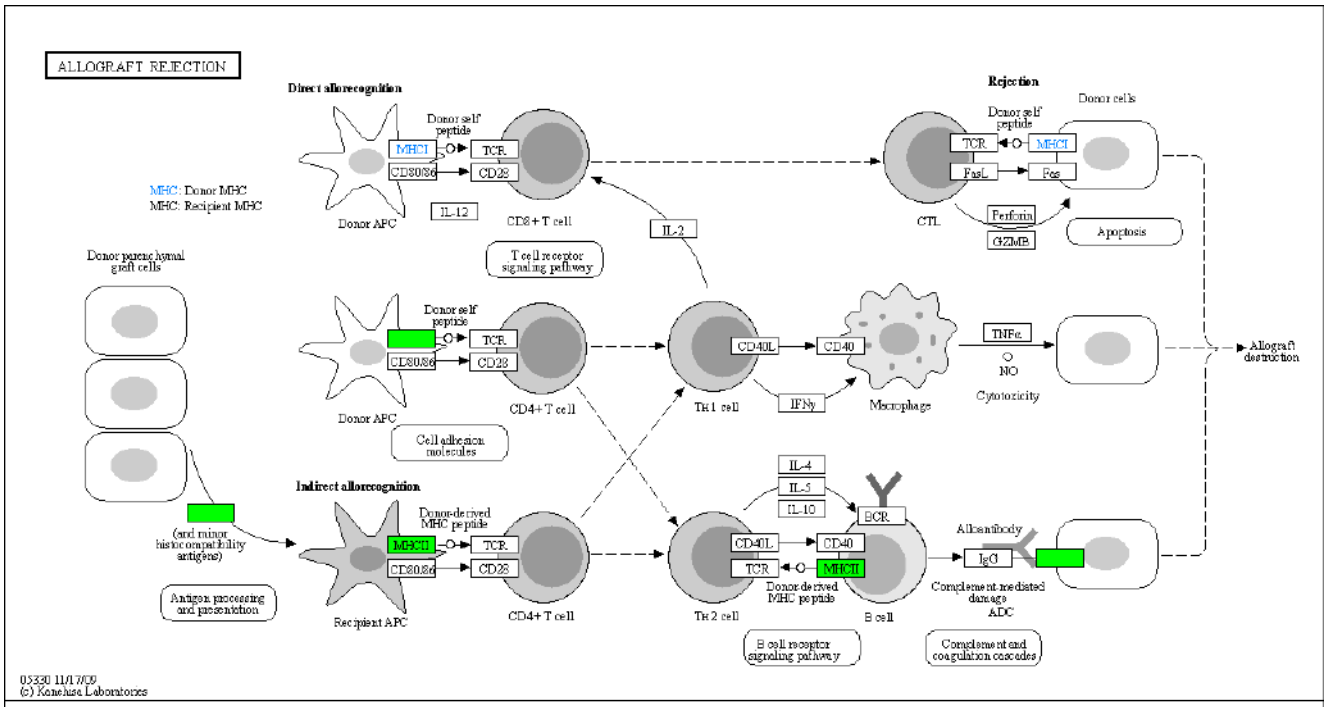

EPSTEIN-BARR VIRUS INFECTION
B cell
Epstein-Barr virus
Entry
Entry, altered signaling, and promotion of latent infection
Toll-like receptor signaling pathway
RIG-I-like receptor signaling pathway
Pro-inflammatory cytokines
Interference with innate immune activation
Viral dsRNA
EBV encoded small RNAs (EBERs)
Apoptosis
Inhibition of apoptosis
Antiviral mediators
Golgi
ER
Interference with MHC I antigen presentation
Peptides
Antigenic Peptides
Viral protein
Proteasome
Cytotoxic T cell
Antigen processing and presentation
Cell survival
Cell proliferation
Cell survival
NF-kappa B signaling pathway
MAPK signaling pathway
JAK-STAT signaling pathway
BCR Blockade
Cell survival and mobility
Calcium signaling pathway
B cell receptor signaling pathway
Basal transcription machinery
Viral DNA
Cellular DNA
Uncontrolled proliferation
Increased survival
Genome instability
p53 signaling pathway
Cell cycle
G1-S transition
05169 9/1/400
(c) Kanehisa Laboratories

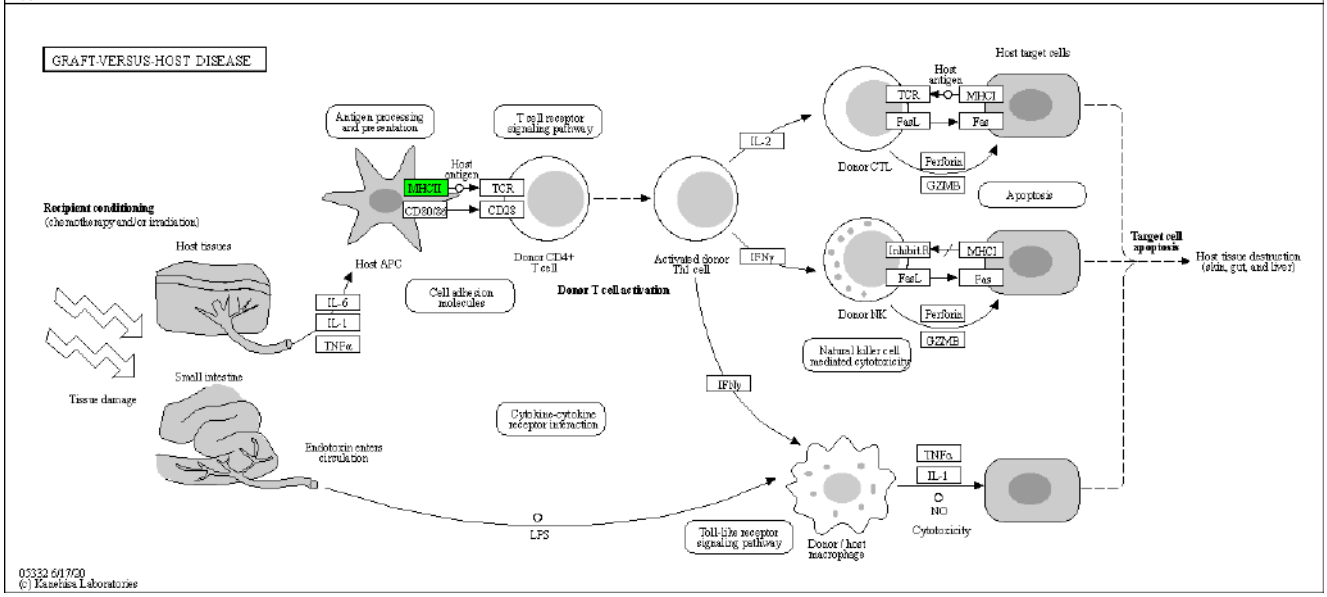

ASTHMA
Allergen
IL-4
IL-9
IL-10
FcεRI signaling pathway
IgE
FcεRI
Mast cell
PAF
Histamine
LTC4
LTD4
PGD2
Immediate reaction
Bronchospasm
Edema
Airflow obstruction
T cell receptor signaling pathway
Cell adhesion molecules
MHC II
TCR
Antigen
APC
TH0 cell
TH2 cell
CD40L
CD40
BCR
B cell
B cell receptor signaling pathway
IL-13
IL-3
IL-5
TNF-α
Bronchus
Antigen processing and presentation
Cytokine-cytokine receptor interaction
Jak-STAT signaling pathway
Eotaxin
(Lung) Epithelial cells Smooth muscle cells Fibroblasts
Eosinophil
MBP
ECP
EPO
LTC4
LTD4
PAF
Late reaction
Airway inflammation
Airflow destruction
Airway hyperresponsiveness
05310 8/17/20
(c) Kanehisa Laboratories


AUTOIMMUNE THYROID DISEASE
Environmental factors
Iodine
Amiodarone
IFN-α
Infectious organisms
Genetic factors
CTLA-4
Breakdown of immune tolerance
Antigen
MHC II
CD80/86
TCR
CD28
APC
CD4+ TH cell
TH1 predominance
TH1 cell
IL-2
Accumulation of immune cells in the thyroid gland
Thyroid cells
Hashimoto's thyroiditis
Hypothyroidism
CTL
FasL
Fas
Perforin
GZMB
Apoptosis
IL-4
IL-5
IL-10
B cell
CD40L
CD40
TH2 cell
IgG
Complement activation ADCC
Necrosis / apoptosis
Complement and coagulation cascades
Thyroid cell damage
Auto-antigen release
TPO
Tg
TSHR
Antigen processing and presentation
T cell receptor signaling pathway
B cell receptor signaling pathway
Cell adhesion molecules
TH2 predominance
Thyroid hormone synthesis
Graves' disease
Hyperthyroidism
Stimulating TSHR-Ab
Tri-iodothyronine (T3)
Thyroxine (T4)
Blocking TSHR-Ab
Atrophic thyroiditis
Hypothyroidism
TSH
05320 7/30/20
(c) Kanehisa Laboratories


INFLAMMATORY BOWEL DISEASE
Luminal bacteria
Antigen
Peptidoglycans
Flagellin
Lipopolysaccharides
PGN
MDP
TLR5
TLR2
TLR4
NF-κB
DNA
TNF-α
Toll-like receptor signaling pathway
Inflammation
NOD-like receptor signaling pathway
Bacterial invasion
Cytokine-cytokine receptor interaction
Th1 lymphocyte
IFN-γ
IFN-γR
STAT1
Tbet
IL-4
IL-5
JAK-STAT signaling pathway
IL-12
IL-12R
STAT4
IL-2
Inhibition of activated Th2 cell
IL-18
IL-18R
AP1
NF-κB
IFN-γ
TLRs
NOD2
TNF-α
IL-6
IL-1
Macrophage
Tissue injury
Crohn's disease
Increase the frequency of Th17 cells
TLR2
NOD2
NF-κB
DNA
MHCII
TCR
APC
Antigen processing and presentation
T cell receptor signaling pathway
DN0 cell
TGF-β
IL-6
Th17 lymphocyte
Smads
STAT3
DNA
IL-21
Th17 cell differentiation
IL-21R
STAT3
FOXP3
IL-23
IL-23R
IL-12Rβ1
STAT3
JAK-STAT signaling pathway
RORγt
RORα
IL-17
IL-22
Th2 lymphocyte
JAK-STAT signaling pathway
c-maf
NFATc1
IL-4
IL-4Rα
γc
STAT6
GATA-3
IFN-γ
IL-4
IL-5
IL-13
IL-10
?
Ulcerative colitis
Inhibition of activated Th1 cell
NKT cell
IL-13
TGF-β
FOXP3
IL-10
TGF-β
Inhibition of IBD
05121 8/23/19
(c) Kanehisa Laboratories


SYSTEMIC LUPUS ERYTHEMATOSUS
UV light, Infections
Apoptotic cell
Defective clearance mechanisms
Nucleosome
dsDNA
H2A
H2B
H3
H4
snRNP
Sm-B
Sm-D
NMDAR
NR2A
NR2B
Phospholipids
Phosphatidyl-serine
Ro6S-A
LaSS-B
α-Actinin
C1q
Autoantigens
B cell receptor signaling pathway
T cell receptor signaling pathway
Cell adhesion molecules
Autoantibodies
Antigen
MHCII
TCR
CD80/86
CD28
APC
ThO cell
CD40L
CD40
Th2 cell
BCR
IgG
B cell
TNF-α
IFN-γ
IL-10
Loss of B cell tolerance
Antigen processing and presentation
Cytokine-cytokine receptor interaction
Jak-STAT signaling pathway
Defective clearance mechanisms
Leukocyte transendothelial migration
Tissue (e.g. renal glomeruli, skin)
C1r
C1s
Complement and coagulation cascades
Immune complex
C4
C2
C4b2a
C3
C3a
C3b
C5
C5a
C5b
MAC
C6
C7
C8
C9
Membrane attack complex
Deposition of immune complexes
Tissue injury and damage
Chemotaxis
Neutrophil
Cathepsin
Elastase
H2O2
FcγR
Macrophage
Oxidants, proteases
01322 6/23/13
(c) Kanehisa Laboratories
RHEUMATOID ARTHRITIS
Blood vessel
Leukocyte migration
Inflammatory cell infiltration
Angiogenesis
VEGF signaling pathway
CCL2
CXCL1
CCL3
CXCL5
CCL20
IL8
Th2
FGF
VEGF
Angl
SDF1
DC
CD80/86
MHCII
Antigen
T cell receptor signaling pathway
Self-reactive Th1 cell
CD28
CTLA4
TCR
LFA1
IL15
IC6MI
IFNγ
RANKL
Synovial macrophage
B cell
APRIL
BLYS
IgG
Autoantibody production
LTB
Toll-like receptor signaling pathway
Peptidoglycan
LPS
TLR2/4
AP1
RANKL
Synovial fibroblast
GMCSF
IL6
IL1β
CCL5
Inflammation
Synovial pannus formation
TNFα
IL1
IL11
IL6
IL18
TGFβ
Th17 cell differentiation
Th17 cell
IL6
IL23
IL17
M-CSF
PGE2
Osteoclast
RANKL
RANK
Osteoclast differentiation
FTH
Vitamin D3
Osteoblast
MMP13
CTSL
Joint destruction
Synovium
Bone
LAMP
CTSK
TRAP
Bone resorption
05323 2/9/19
(c) Kanehisa Laboratories


ALLOGRAFT REJECTION
Direct allorecognition
MHC: Donor MHC
MHC: Recipient MHC
Donor self peptide
MHCI
TCR
CD8αβ
CD28
Donor APC
IL-12
CD8+ T cell
T cell receptor signaling pathway
IL-2
Rejection
Donor cells
Donor self peptide
TCR
MHCI
FasL
Fas
CTL
Perforin
GZMB
Apoptosis
Donor parenchymal graft cells
Donor self peptide
TCR
CD8αβ
CD28
Donor APC
CD4+ T cell
Cell adhesion molecules
CD40L
CD40
Th1 cell
IFNγ
Macrophage
TNFα
NO
Cytotoxicity
Allograft destruction
Indirect allorecognition
(and minor histocompatibility antigens)
Antigen processing and presentation
Donor-derived MHC peptide
MHCII
TCR
CD8αβ
CD28
Recipient APC
CD4+ T cell
IL-4
IL-5
IL-10
CD40L
CD40
TCR
MHCII
Donor derived MHC peptide
Th2 cell
BCR
B cell
Alloantibody
IgG
Complement-mediated damage
ADC
B cell receptor signaling pathway
Complement and coagulation cascades
05330 11/17/09
(c) Kanehisa Laboratories


GRAFT-VERSUS-HOST DISEASE
Host target cells
Host antigen
TCR
MHCI
FasL
Fas
Donor CTL
Perforin
GZMB
Apoptosis
Antigen processing and presentation
T cell receptor signaling pathway
Host antigen
MHCII
TCR
CD8αβ
CD28
Host APC
Donor CD4+ T cell
IL-2
Activated donor Th0 cell
IFNγ
Recipient conditioning
(chemotherapy and/or irradiation)
Host tissues
IL-6
IL-1
TNFα
Cell adhesion molecules
Donor T cell activation
Inhibit R
MHCI
FasL
Fas
Donor NK
Perforin
GZMB
Natural killer cell mediated cytotoxicity
Target cell apoptosis
Host tissue destruction (skin, gut, and liver)
Tissue damage
Small intestine
Endotoxin enters circulation
Cytokine-cytokine receptor interaction
IFNγ
LPS
Toll-like receptor signaling pathway
Donor / host macrophage
TNFα
IL-1
NO
Cytotoxicity
05332 6/17/09
(c) Kanehisa Laboratories

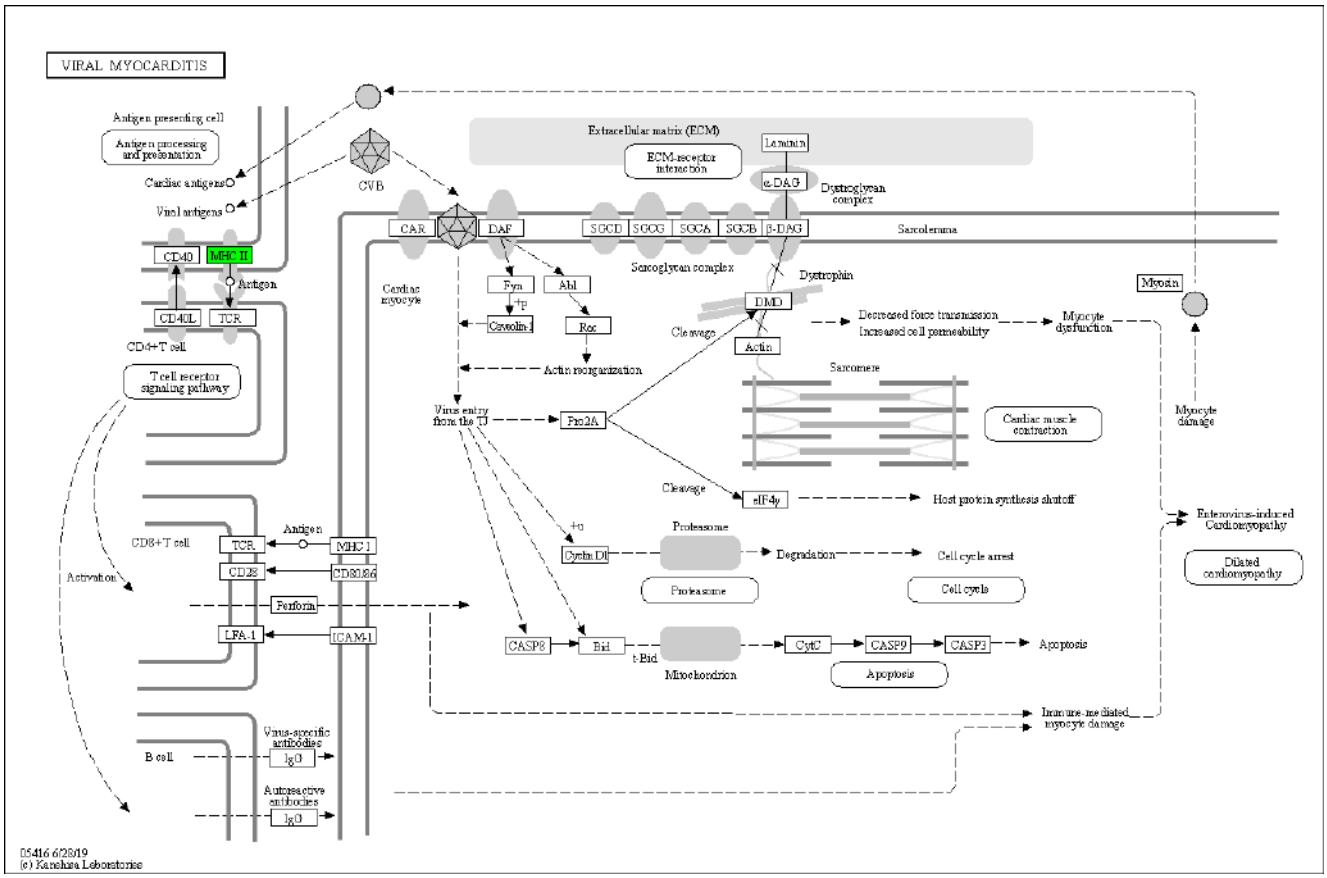
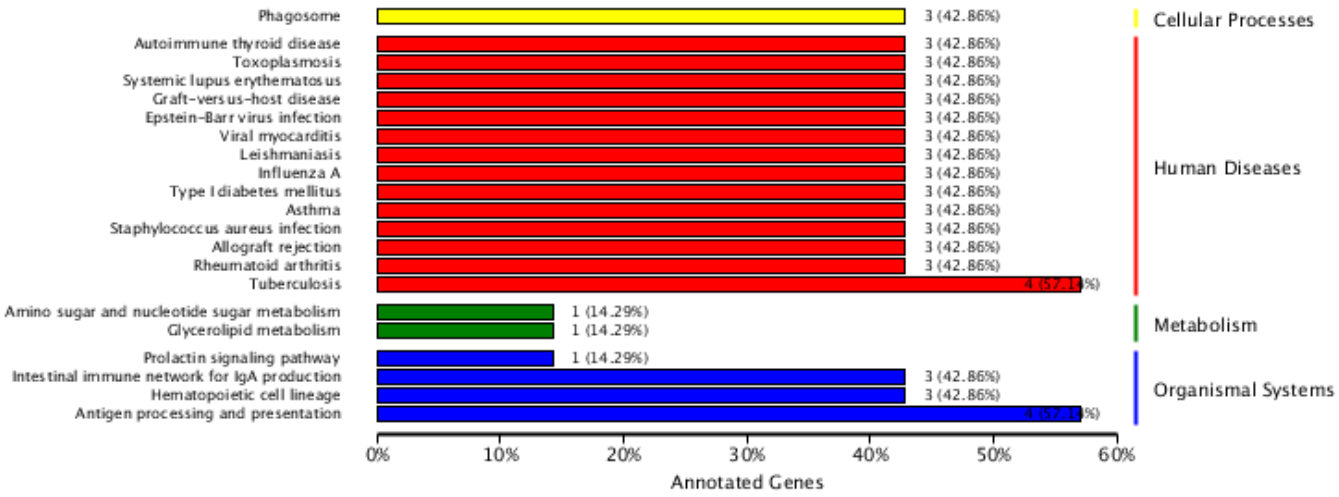


Figure. Demo of the KEGG annotation on DEGs

Note: Relative to control group, the nodes colored in red represent the enzymes related to up-regulated genes and the green ones represent that of down-regulated genes.
Blue ones represent enzymes related to both up and down-regulated genes. The number in the box stands for EC number. The pathway consists of many complex biochemical reactions involving multiple enzymes. The DEGs annotated to the pathway were colored on the figure. Researchers can pick pathways of their own interest basing on the highlighted pathways and research subjects for further analysis and interpretation.

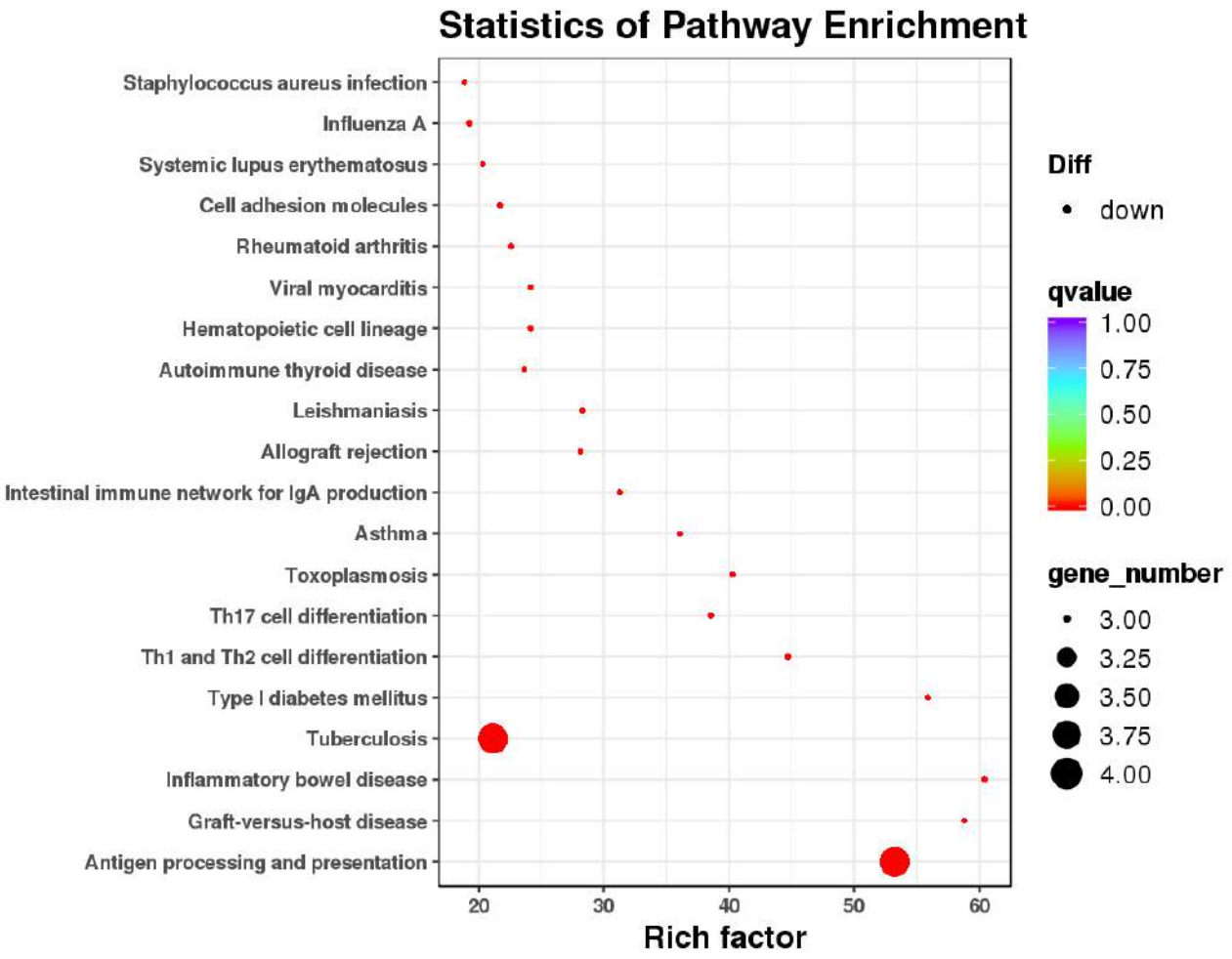
The KEGG annotations of DEGs were classified according to the type of pathways. Detailed classification was shown in the following figure.



Note: Y-axis: KEGG pathway terms; X-axis: Number and the percentage of genes annotated to the KEGG pathway.

3.11.5 KEGG pathway enrichment analysis on DEGs

In this session, we examined if the pathways are over-presented with DEGs. Enrichment factors and fisher test were applied in the determination of enrichment degree and significancy of the pathway. Enrichment of DEGs in KEGG pathways are shown in the figures below. Top 20 enriched pathways (with smallest Q-value) were shown.



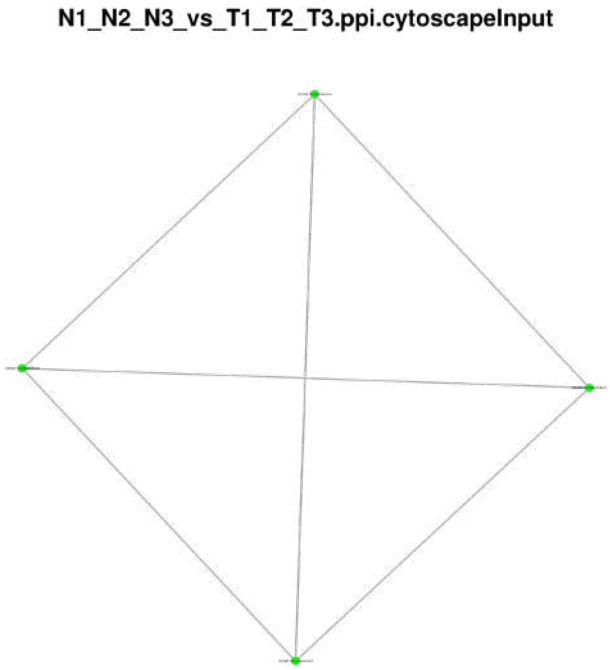
Note: Each dot represents a KEGG pathway. Y-axis: Pathway; X-axis: Enrichment factor. Enrichment factor is calculated as "Enrichment factor=(Ratio of DEGs annotated to the term over all DEGs)/(Ratio of genes annotated to the term over all genes)"
A larger enrichment factor indicates a more significant enrichment of the pathway.
The color of the dots stands for q-value (adjusted p-value). The smaller the q-value is, the more significant or reliable the enrichment is.
The size of the dots represents the number of DEGs enriched in this pathway. The larger the dot is, the more genes it contains.

In this figure, the dots closer to lower right area are more reliable in differential analysis. Top 20 enriched pathways (with smallest Q-value) were shown.

3.11.6 Protein-protein interaction network of DEGs

STRING [27] is a database containing information of predicted and proved protein-protein interactions (PPI) of a collection of species. The interactions refer to both direct physical interactions and indirect functional interactions. The PPI network was built based on the DEGs generated in the differential expression analysis and existing information on interactions in database. For the species included in the database, the interactions of targeted genes can be extracted directly from the database for network construction. For species couldn't be found in STRING, homologous proteins were used for network construction. The PPT networks can be visualized by Cytoscape [28].

PPI networks of DEGs visualized in Cytoscape were shown as below.



Note: Each node in the figure represents a protein. The edge between nodes represents interactions.
The size of the nodes represents their degree, i.e. the number of interactions linked to them. The larger the nodes are, the more interactions they are involved in.
The color of the node is related to the clustering coefficient. With spectrum from green to red, the clustering coefficient increases. A higher clustering coefficient (red nodes) indicates a better connectivity of the node to surrounding nodes.
The thickness of the edge between two nodes represents the strength of interactions. The thicker the edge is, the stronger the interaction is.
Nodes without connections means there are no PPT found in the analysis.

3.11.7 Transcriptional factor annotation

Regulations at transcription level is a crucial step in gene expression regulation. Transcription factor (TF) specifically binds to sequences in upper-stream of a gene to regulate transcription of the gene. A number of biological functions are regulated by altering specific TF to stimulate or inhibit the expression of corresponding genes. Therefore, it is necessary to annotate TFs of differentially expressed genes between groups. TFs of DEGs were identified and annotation based on AminalTFDB database.

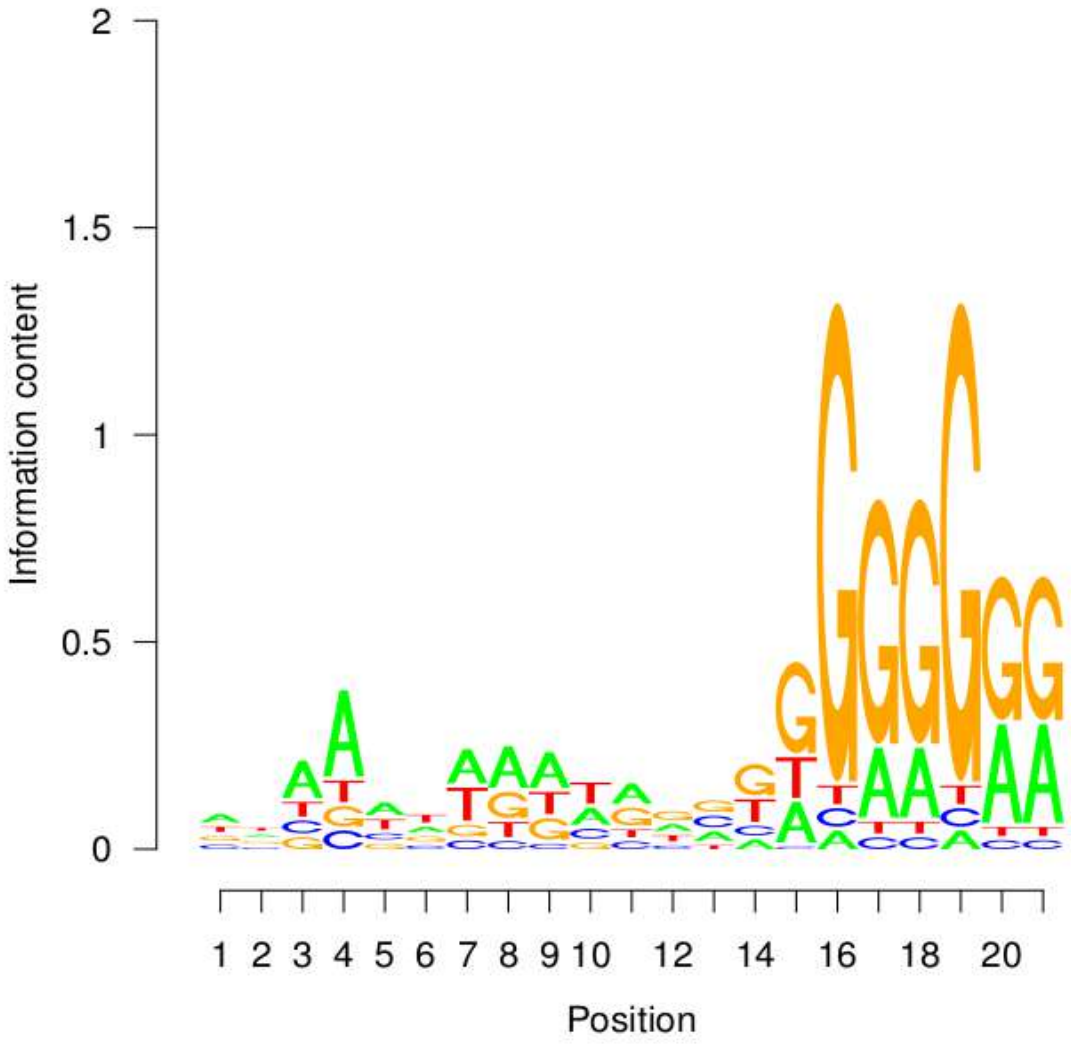
3.11.8 Prediction of transcription factor binding sites

Transcription factor binding site (TFBS) refers to DNA fragments where transcription factors bind to. The length of TFBS ranges from 5 to 20 bp. Normally, a transcription factor regulates several genes simultaneously. Its bind sites on different genes share conserved region, however, are not exactly the same. In current analysis, R package TFBStools [33] is applied to predict TFBS in promoter regions of DEGs (Promoter region of a gene is defined as 1 kb upper-stream of the gene.). JASPAR [32] database (http://jaspar.genereg.net/) is used here as reference motif database. Output of prediction was listed in the table below.

Model_id	seqname	Symbol	start	end	score	strand	frame	TF	class	sequence	Pvalue
MA0004.1	ENSMUSG00000000001	Gnai3	776	781	1	-	.	Arnt	Basic helix-loop-helix factors (bHLH)	CACGTG	0
MA0006.1	ENSMUSG00000000001	Gnai3	606	611	0.99	-	.	Ahr::Arnt	Basic helix-loop-helix factors (bHLH)	CGCGTG	0.00024
MA0009.1	ENSMUSG00000000001	Gnai3	362	372	0.9	-	.	T	T-Box factors	CTAGGTGTAAT	1.04904174804688e-05
MA0027.1	ENSMUSG00000000001	Gnai3	770	780	0.91	-	.	En1	Homeo domain factors	ACGTGGTGTGC	0.00071
MA0035.1.1	ENSMUSG00000000001	Gnai3	249	254	0.95	-	.	Gata1	Other C4 zinc finger-type factors	TGATAG	0.0032

Note: Model_id: TFBS motif ID; seqname: Gene name; start: Starting position; end: Ending position; score: Score stands for the possibility of binding between TF and TFBS; strand: direction of strand; frame : .; TF: Transcription factor ID; class: Annotation of TF; sequence: TFBS sequence; Pvalue: P-value.

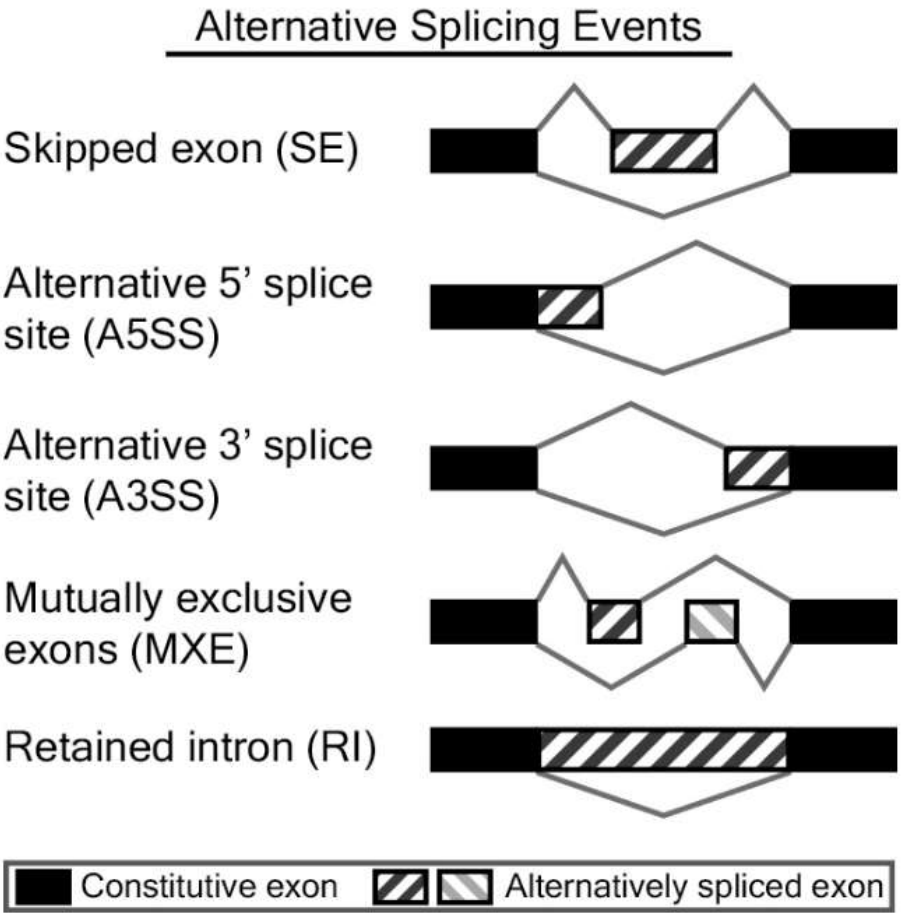
TFBS sequence features were shown in the figure below.



Note: X-axis: Relative position of base in motif; Y-axis: Conservation of the base on the position. The height of the signal stands for relative frequency of corresponding base at the position.

3.12 Differential alternative splicing analysis

Differential alternative splicing analysis was processed by rMATS [31]. The number of reads that uniquely mapped to the transcript (the exon inclusion isoform or the exon skipping isoform) is defined as inclusion level of alternative splicing. The rMATS statistical calculates the p-value between IncLevel (Inclusion level) of two groups of samples by likelihood-ratio test. The p-values were then corrected by Benjamini Hochberg to get FDR value. In current analysis, the default threshold for rMATS screening is $|\Delta\psi| > c$ (c=0.0001); i.e. P-value between mean ψ values of two samples larger than the threshold c. rMATS can identify following 5 alternative splicing events: exon skipping (SE), alternative 5' splicing junction (A5SS), alternative 3' splicing junction (A3SS), mutually exclusive exon (MXE) and intron retention (RI).



rMATS software analysis results:

N1_N2_N3_vs_T1_T2_T3.A3SS.MATS.JC.xls
N1_N2_N3_vs_T1_T2_T3.A5SS.MATS.JC.xls
N1_N2_N3_vs_T1_T2_T3.MXE.MATS.JC.xls
N1_N2_N3_vs_T1_T2_T3.RI.MATS.JC.xls
N1_N2_N3_vs_T1_T2_T3.SE.MATS.JC.xls

Note: GeneID: Gene ID; geneSymbol: Gene symbol; chr: chromosome No. Strand: +/-strand; exonStart_0base: Starting position of exon (from 0); exonEnd: Ending position of exon; upstreamES: Starting position of upper-stream exon; upstreamEE: Ending position of upper-stream exon; downstreamES: Starting position of down-stream exon; downstreamEE: Ending position of downstream exon; (Other types of alternative splicing may have some different columns); IJC_SAMPLE_1：counts of inclusion junction counts in SAMPLE_1, replicates are divided by ","; SJC_SAMPLE_1： counts of skipping junction in SAMPLE_1, replicates are divided by ","; IJC_SAMPLE_2 counts of inclusion junction in SAMPLE_2, replicates are divided by ","; SJC_SAMPLE_2：counts of skipping junction in SAMPLE_2; IncFormLen： Valid length of inclusion form; SkipFormLen: Valid length of skipping form; P-Value: Significancy in alternative splicing events between two samples; FDR: FDR value; IncLevel1: Inclusion level of samples in group SAMPLE_1 (replicates are divided by ","); IncLevel2: Inclusion level of samples in group SAMPLE_2 (replicates are divided by ","); IncLevelDifference： average(IncLevel1) – average (IncLevel2).

Statistics of differential alternative splicing events:

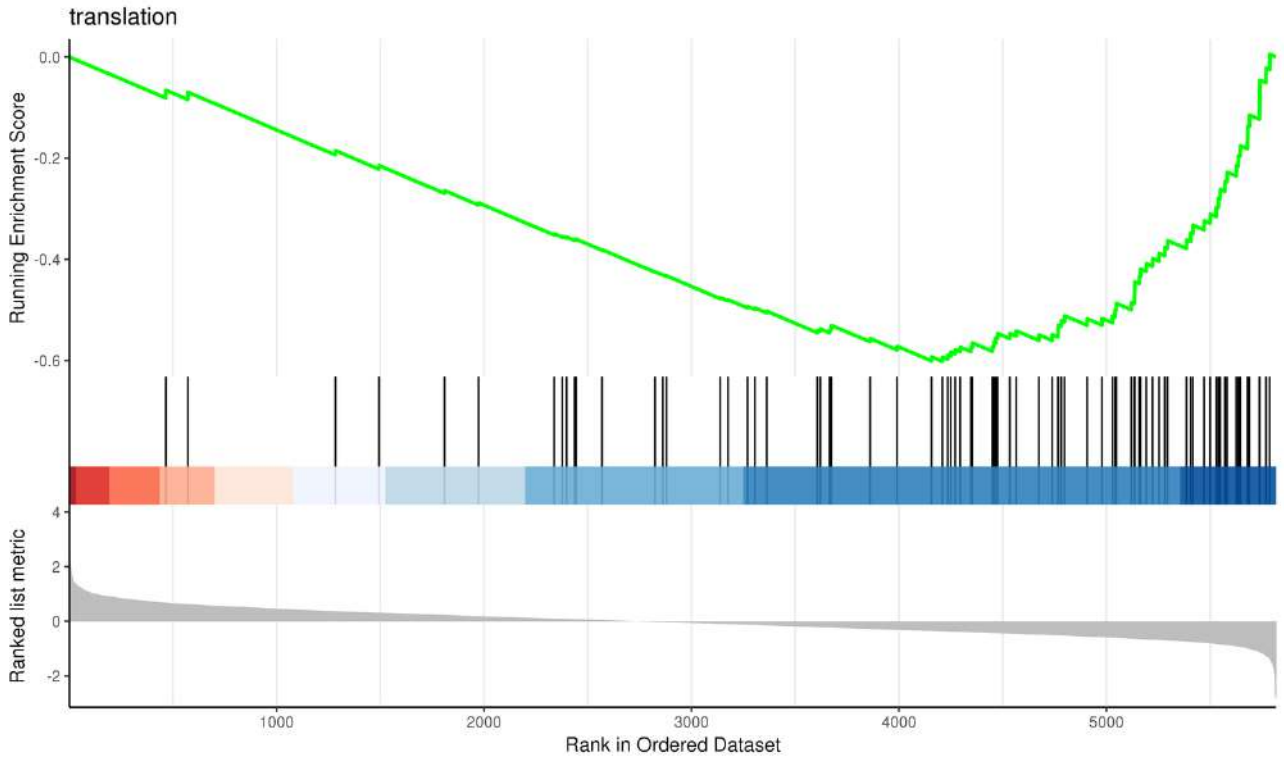
Table. Statistics of differential alternative splicing events

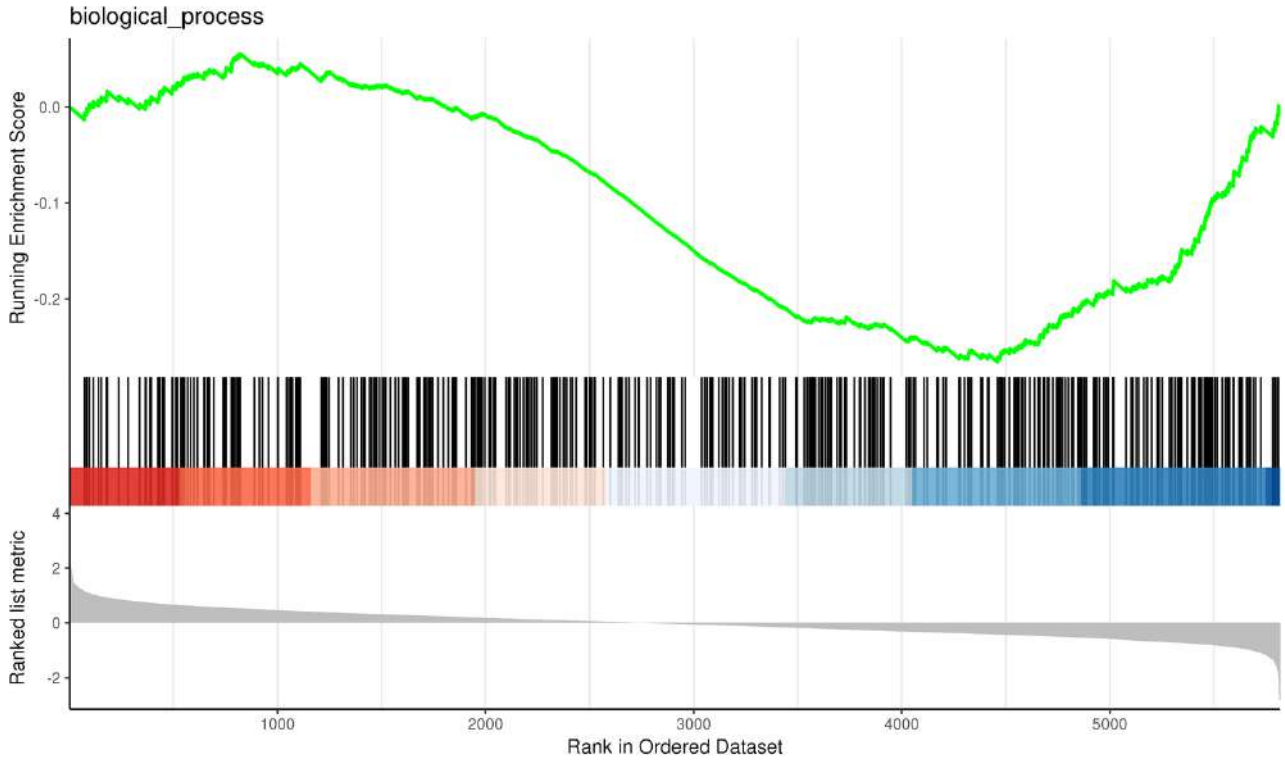
DEG Set	A3SS	A5SS	MXE	RI	SE
N1_N2_N3_vs_T1_T2_T3	**1,428**	**745**	**1,944**	**338**	**15,208**

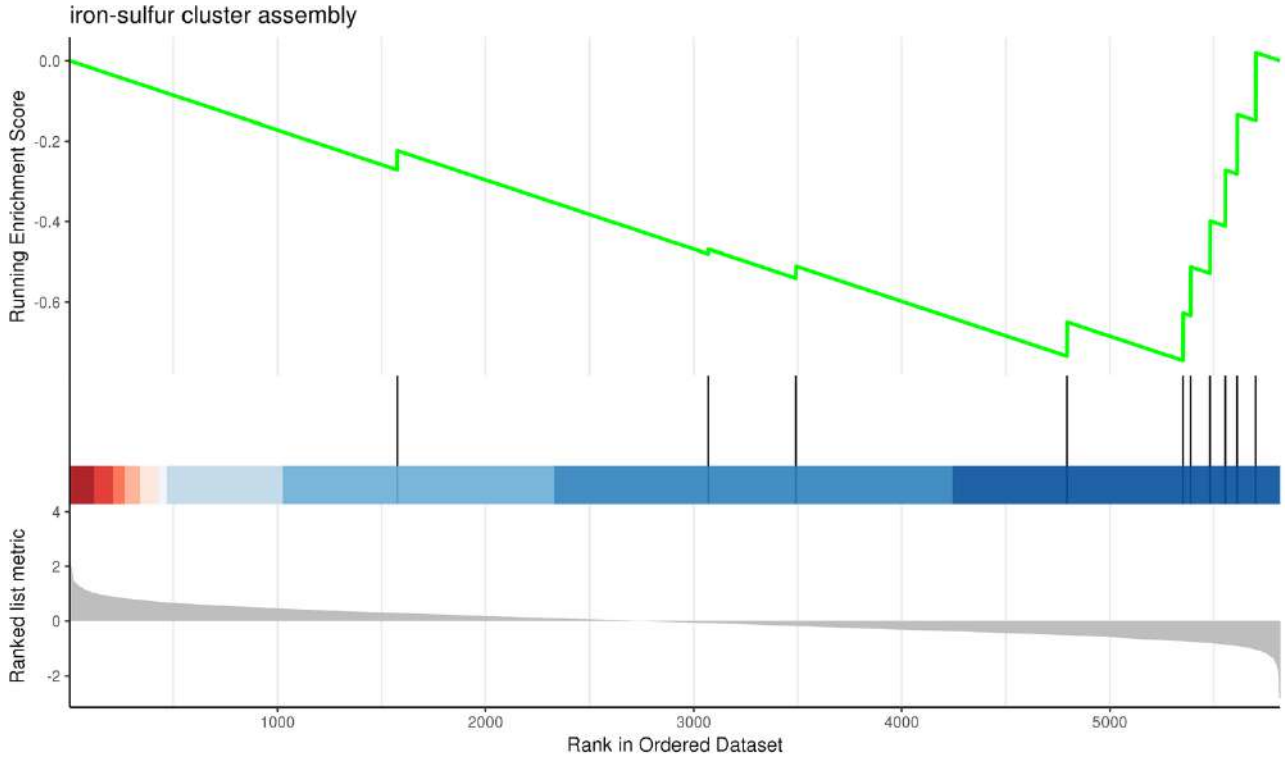
Note: DEG set: Name of DEG set; The rest columns: number of DEGs in corresponding alternative splicing events; A3SS: Alternative 3' splice junction; A5SS: Alternative 5' splice junction; MXE: Mutually exclusive exons; RI: Intron retention; SE: Exon skipping.

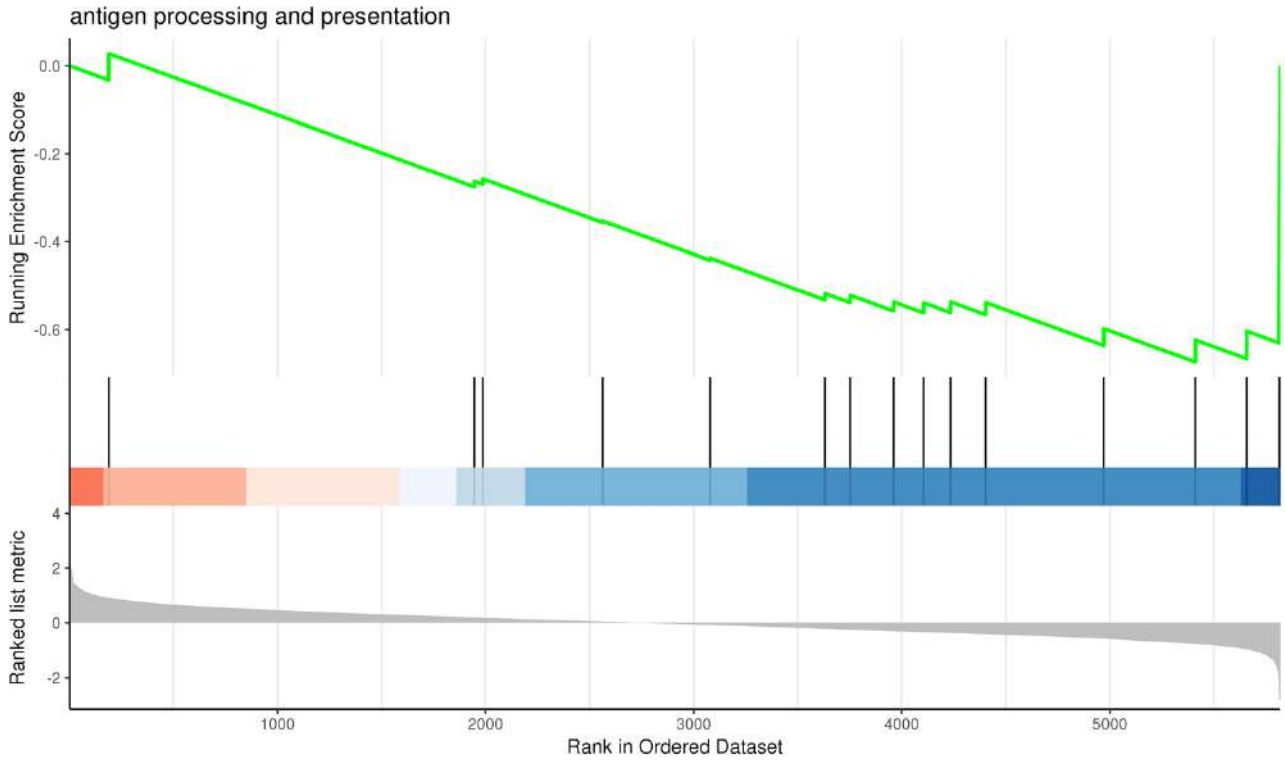
3.13 GSEA analysis

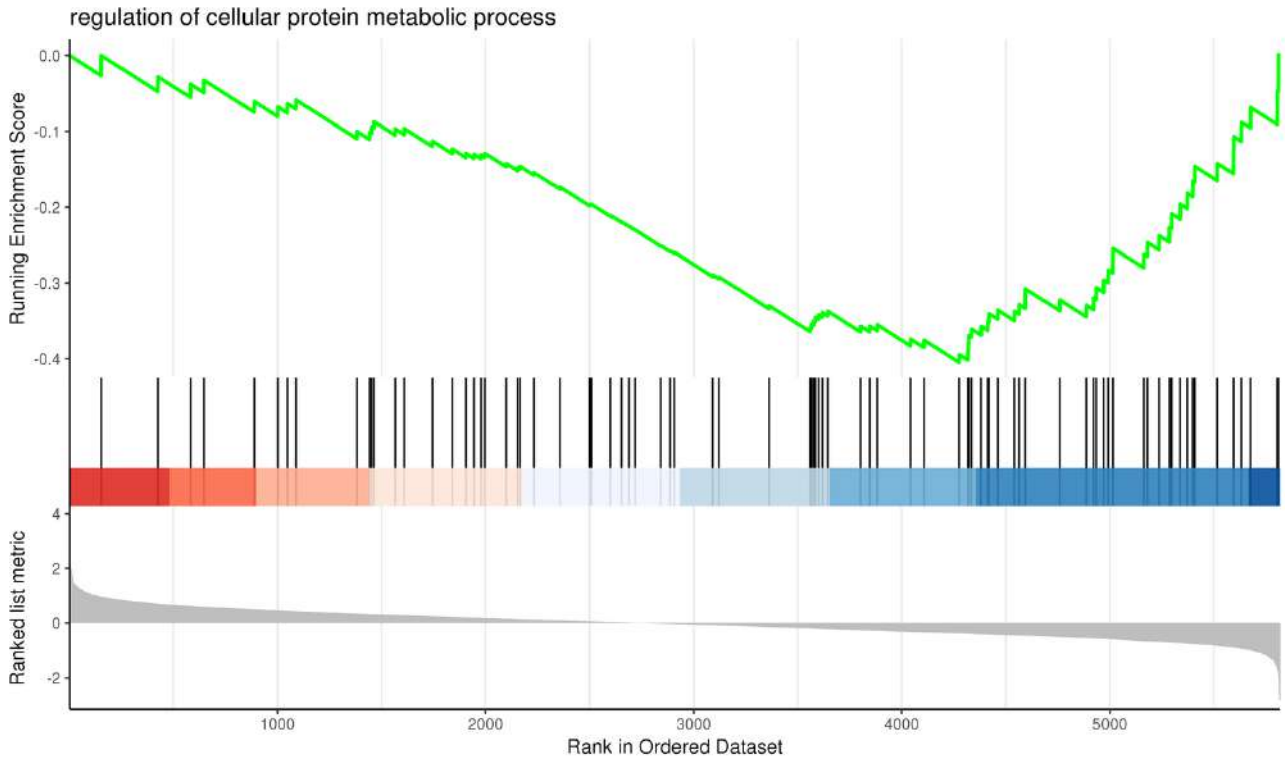
Gene Set Enrichment Analysis(GSEA) [32] was processed on all genes based on expression level. Normally, differential expression analysis only focus on up- or down-regulated genes with statistical significancy. However, this may mask the genes, which are altered slightly without significancy but play vital role in biological functions. Without setting threshold on fold change and significancy, GSEA is able to detect weak alterations in gene expression. In this analysis, genes sets of KEGG pathway and GO terms on BP, CC, MF were employed as gene sets of interest. Genes of each group were used as background gene set. Enriched gene sets were identified as p-value<0.001 and FDR<0.05.

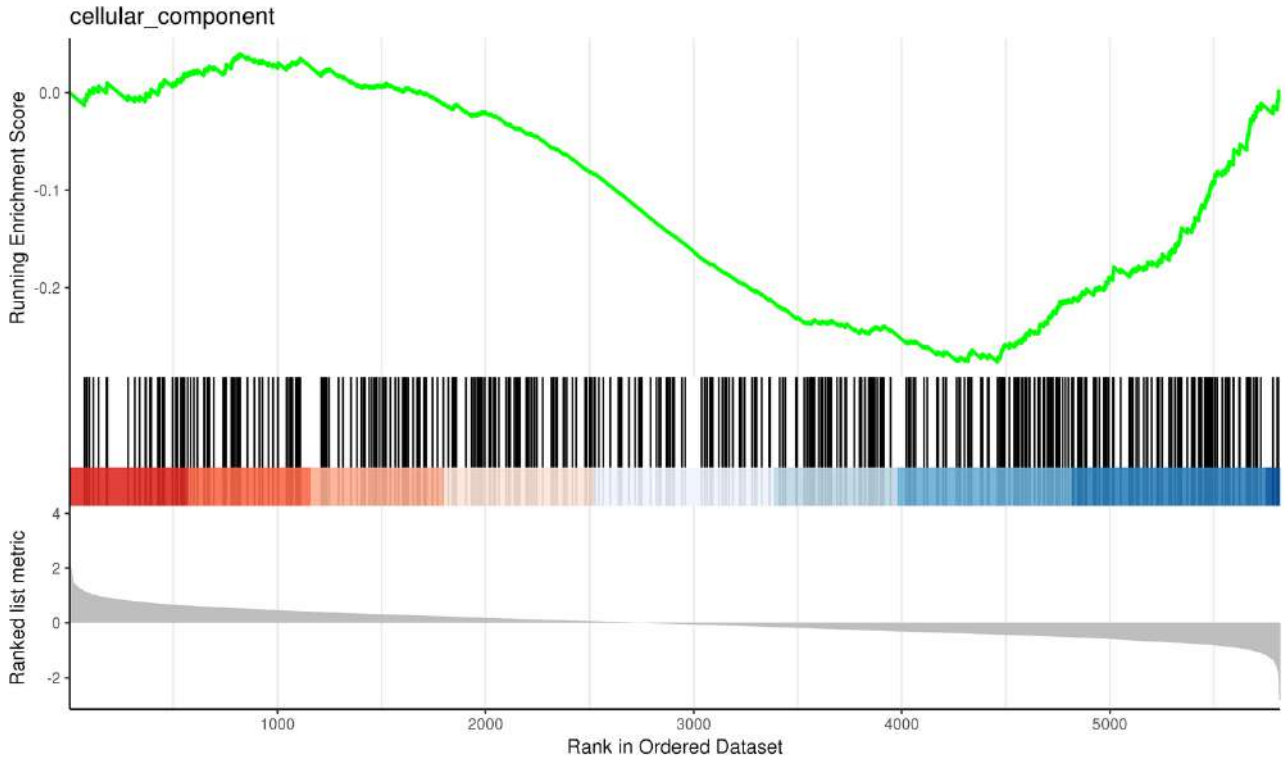


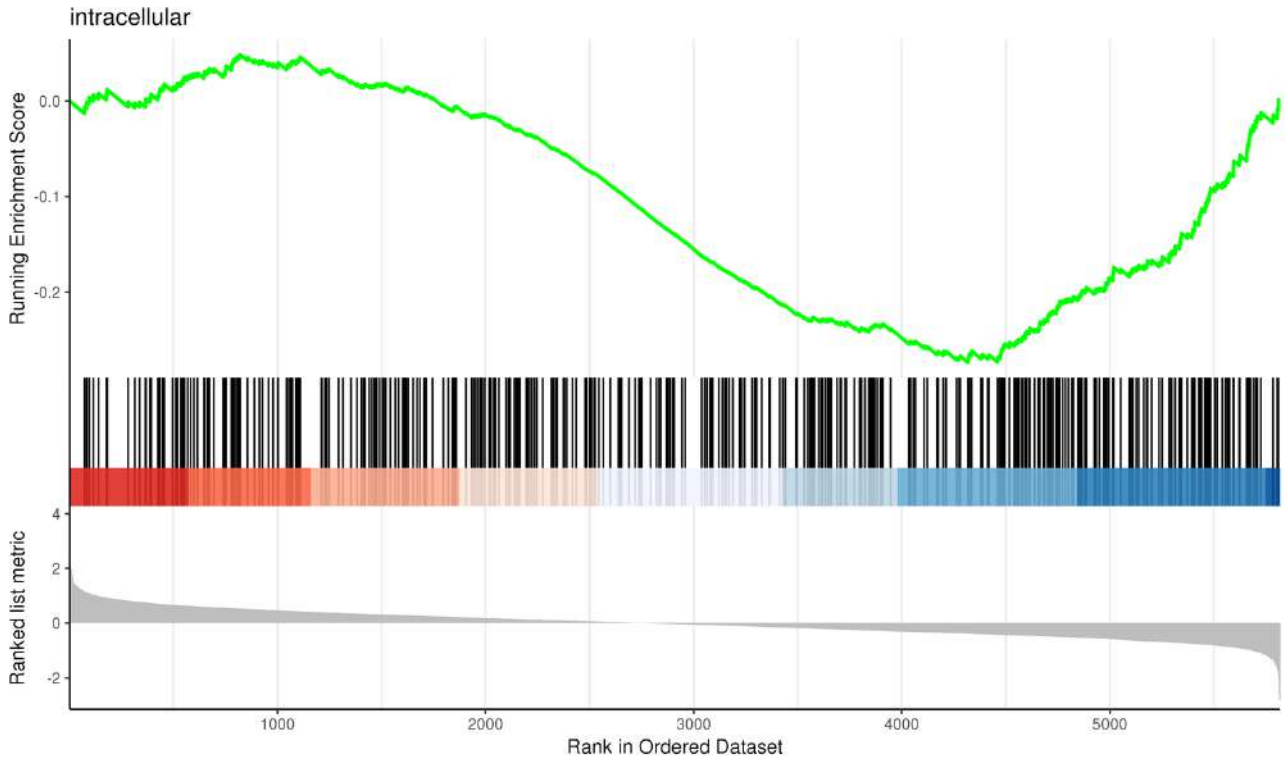

biological_process
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

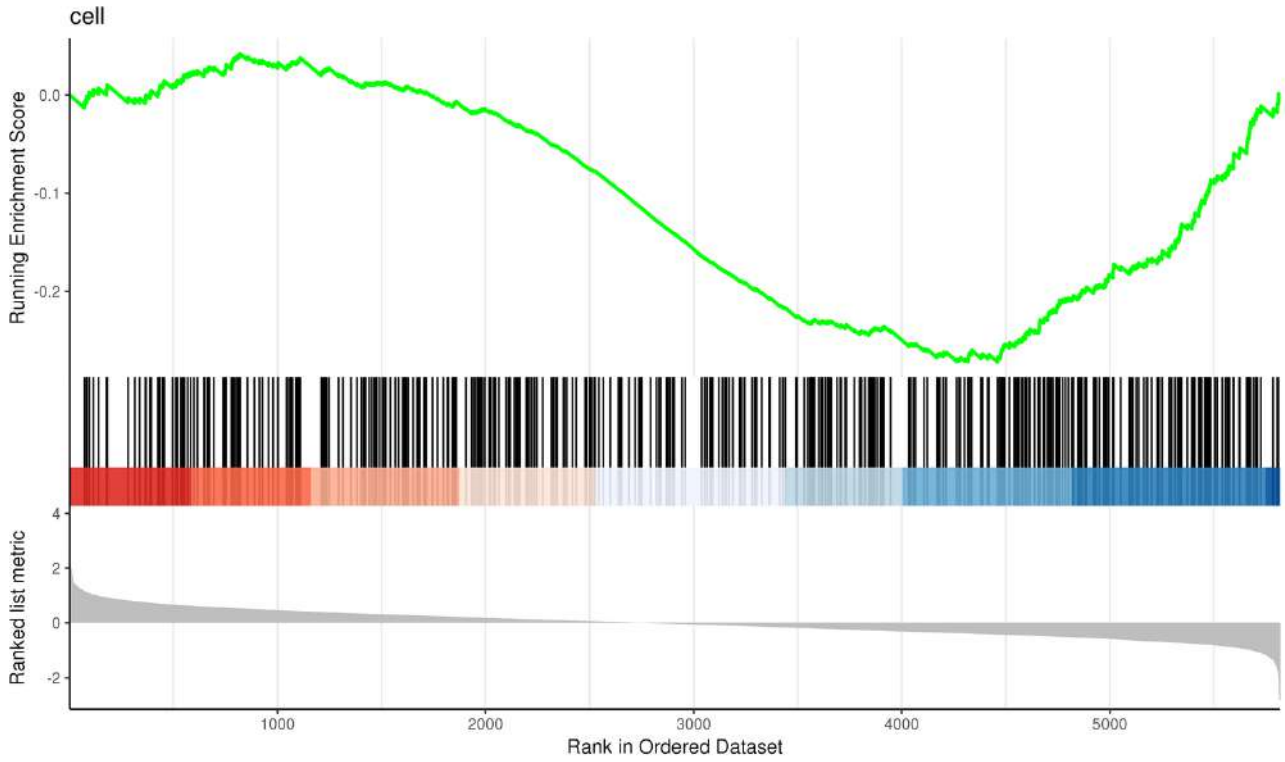

iron-sulfur cluster assembly
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

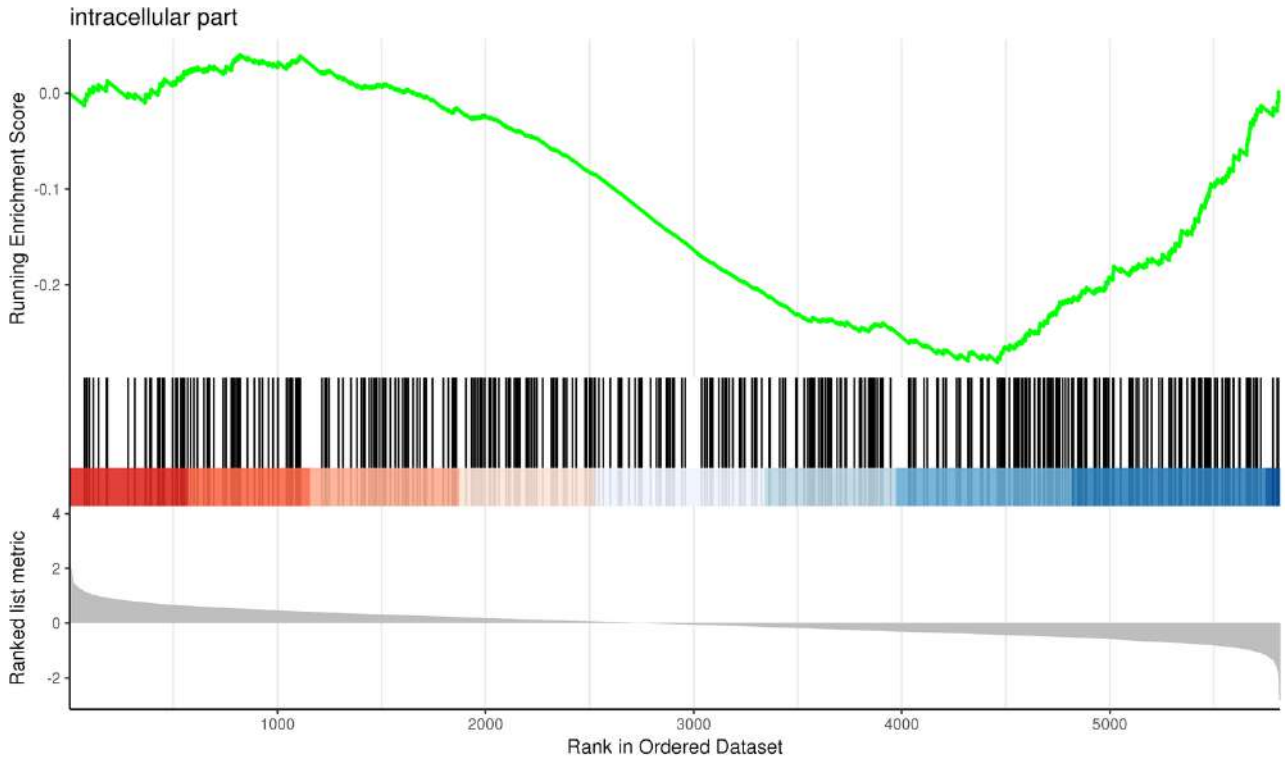

antigen processing and presentation
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

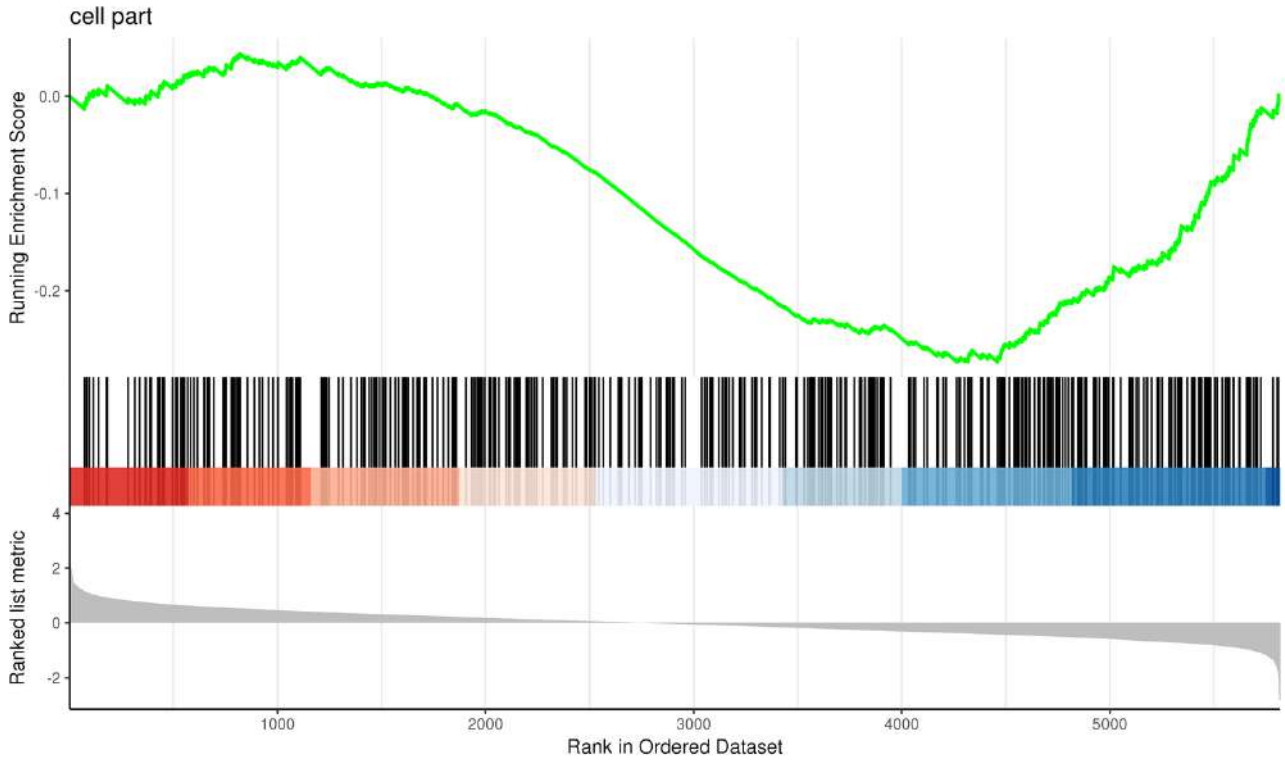

regulation of cellular protein metabolic process
Running Enrichment Score
0.0
-0.1
-0.2
-0.3
-0.4
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

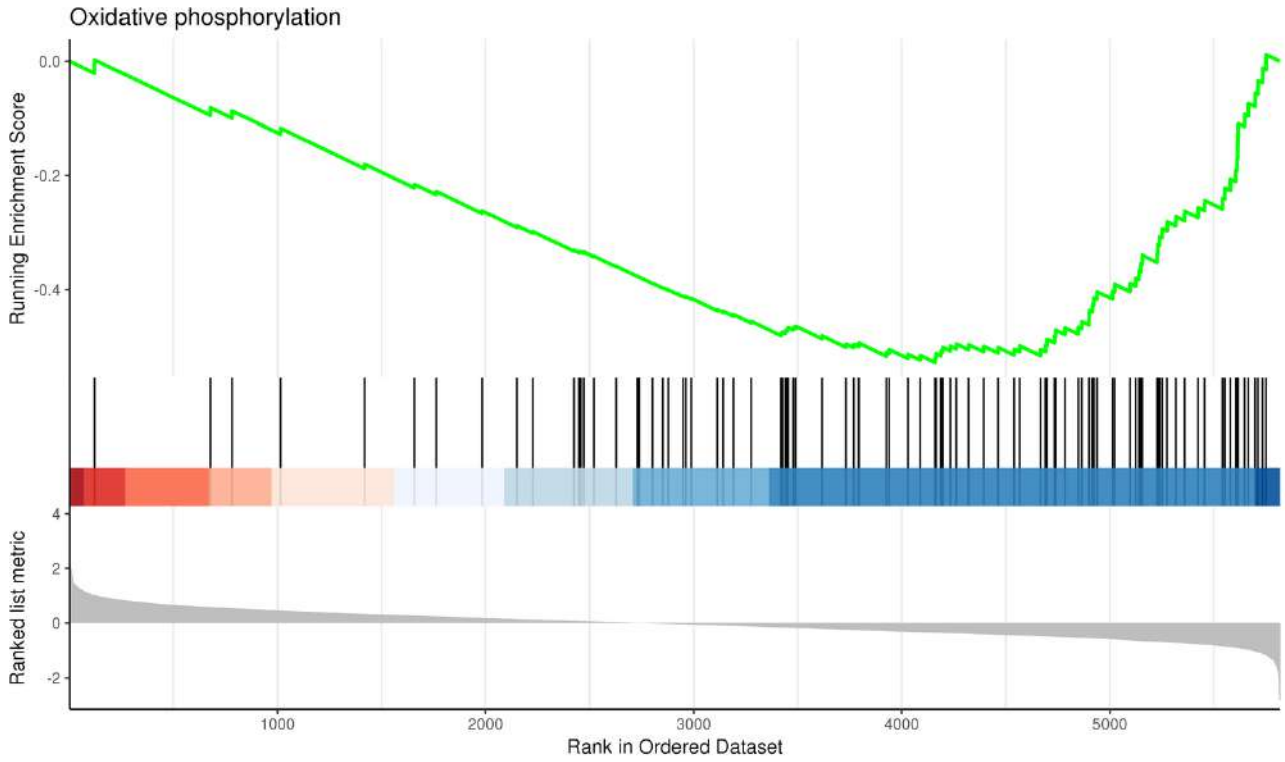

cellular_component
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

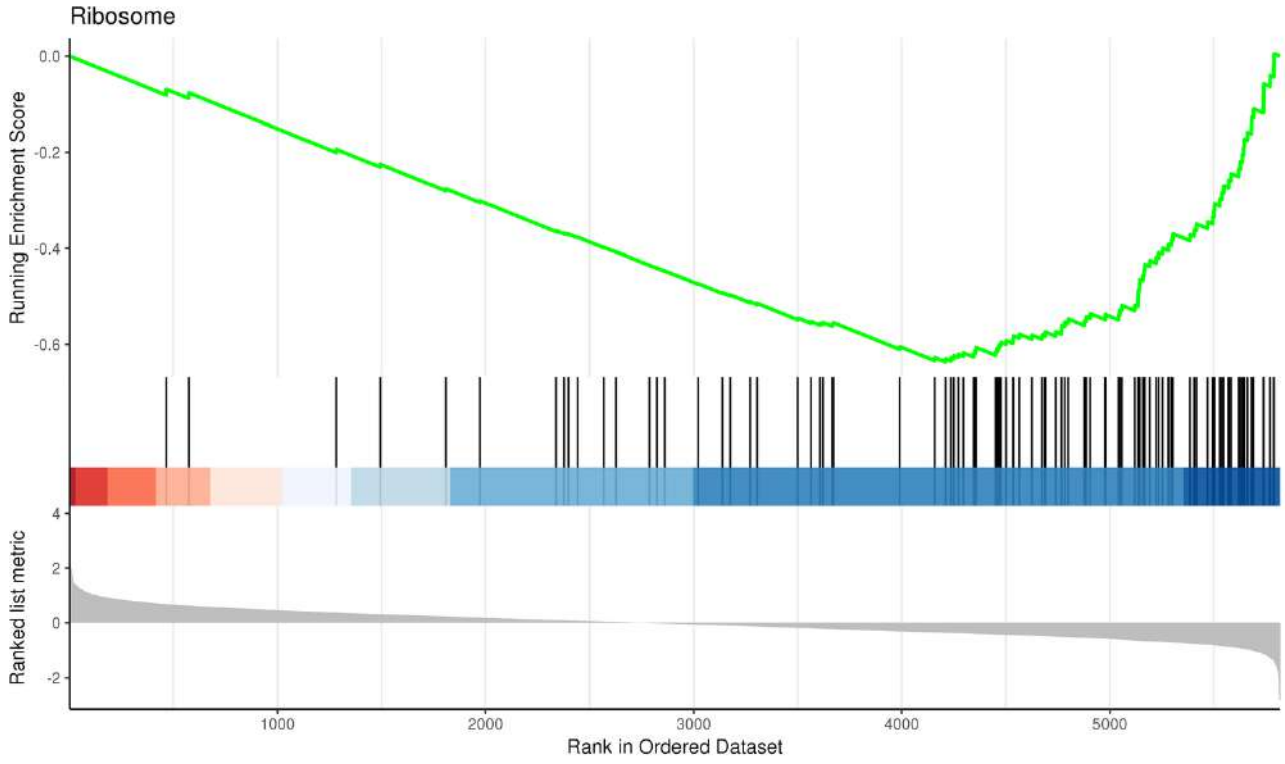

intracellular
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

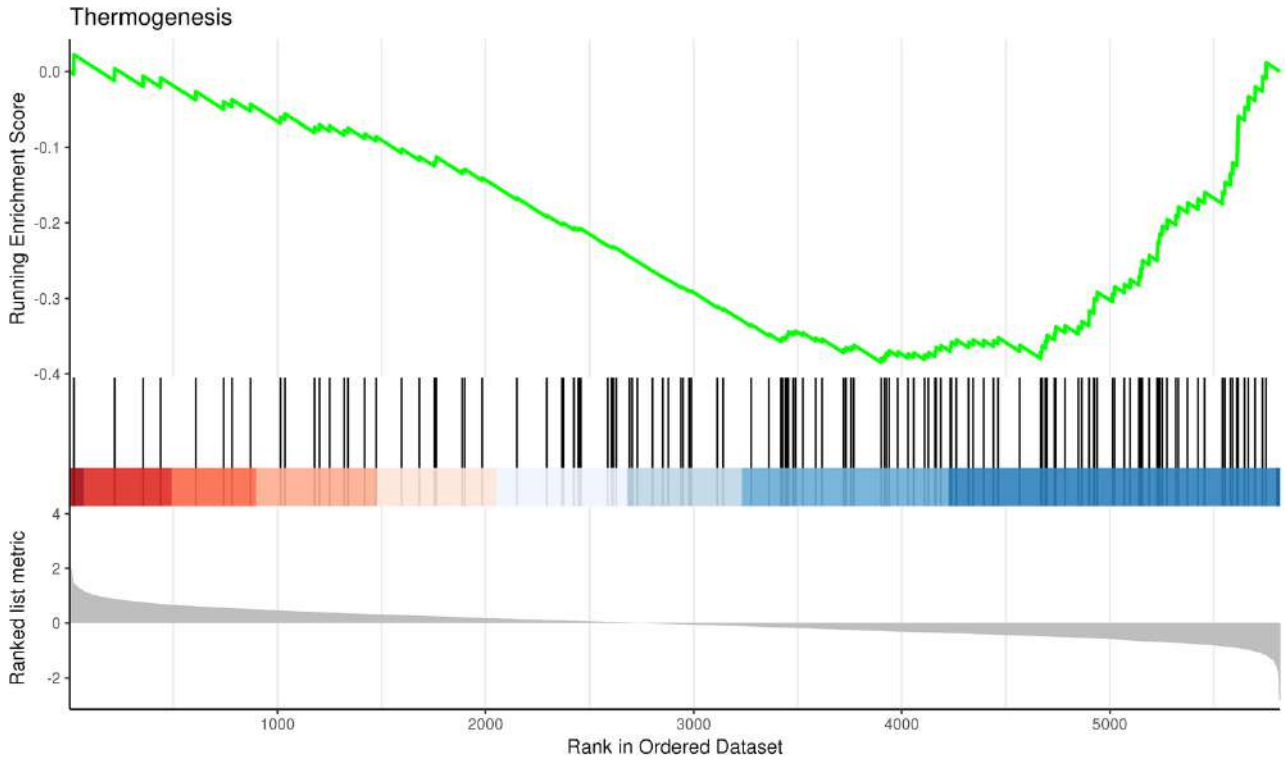

cell
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

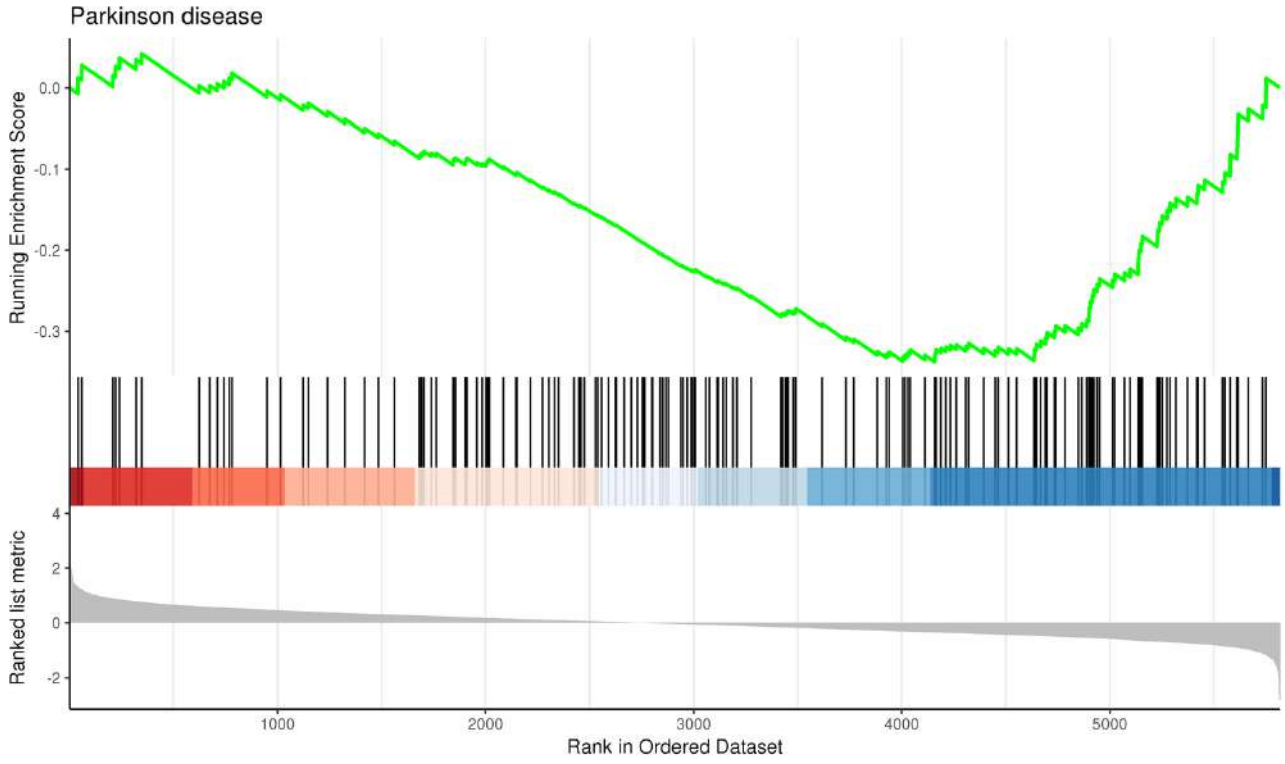

intracellular part
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

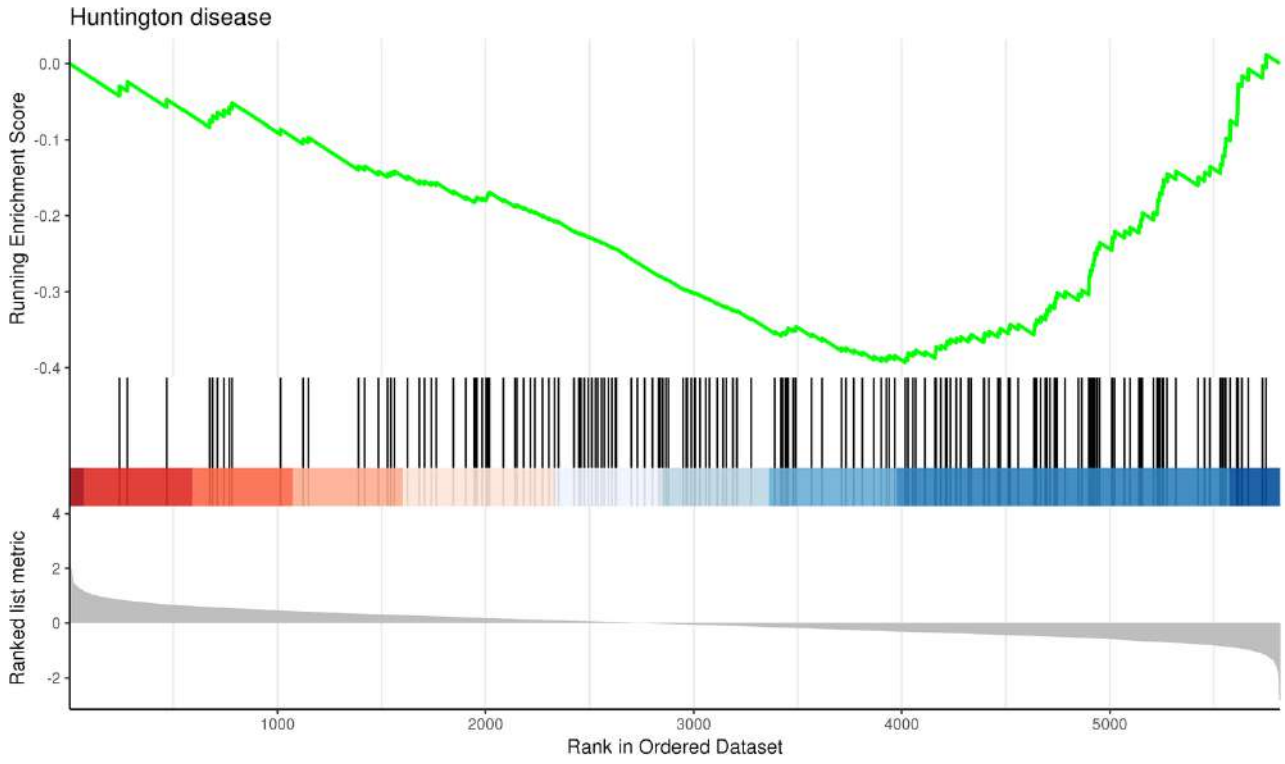

cell part
Running Enrichment Score
0.0
-0.1
-0.2
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

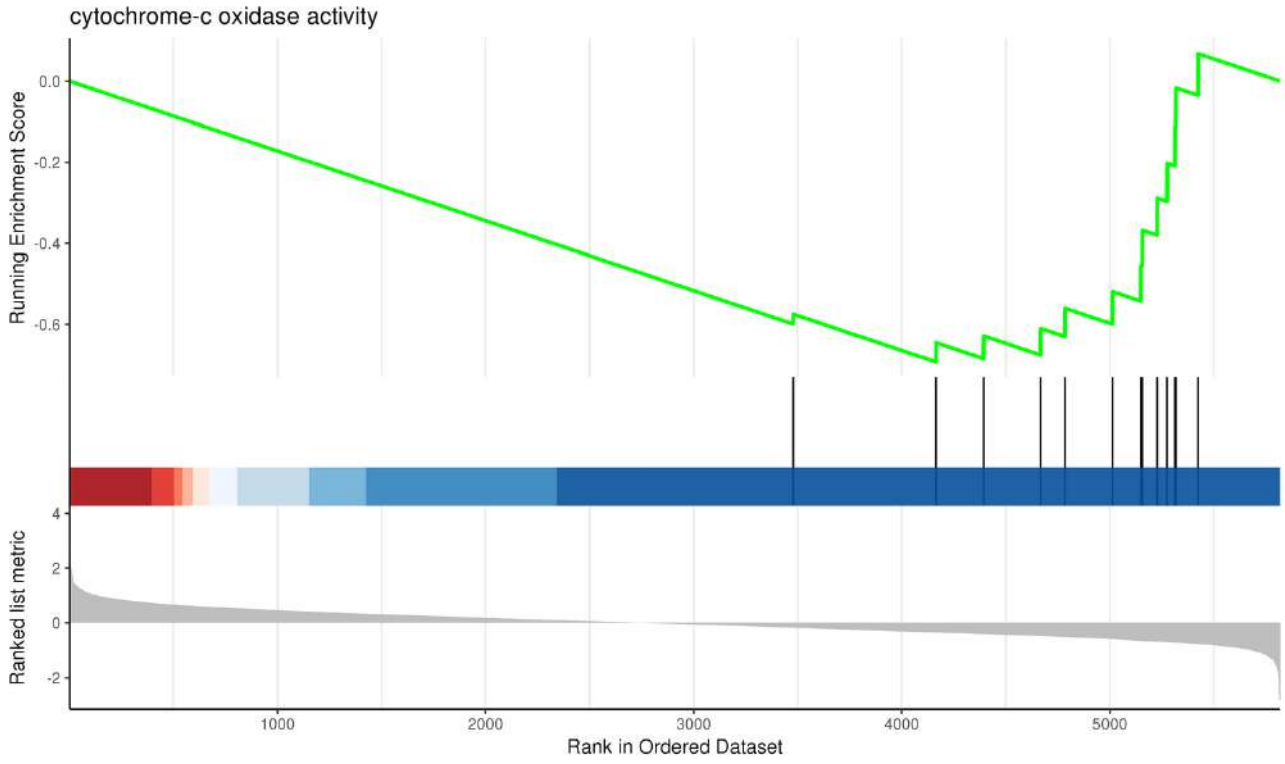

Oxidative phosphorylation
Running Enrichment Score
0.0
-0.2
-0.4
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

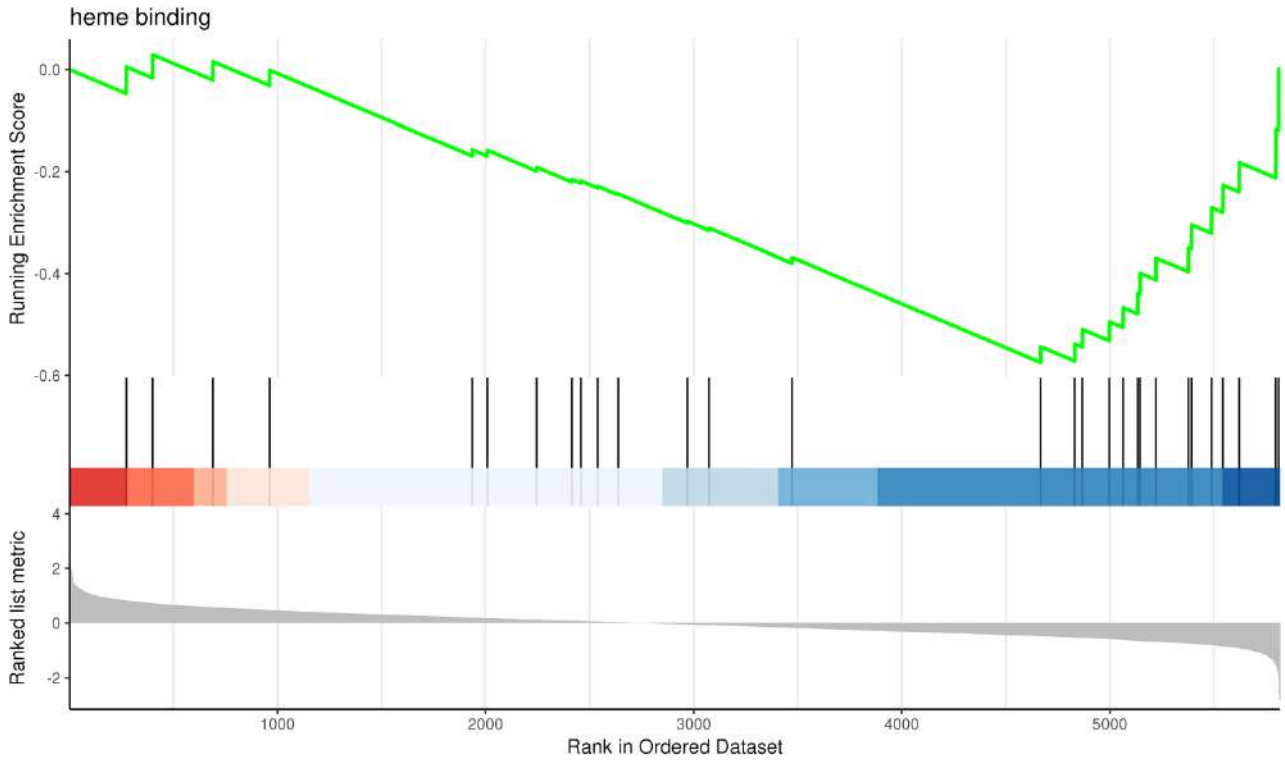

Ribosome
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset

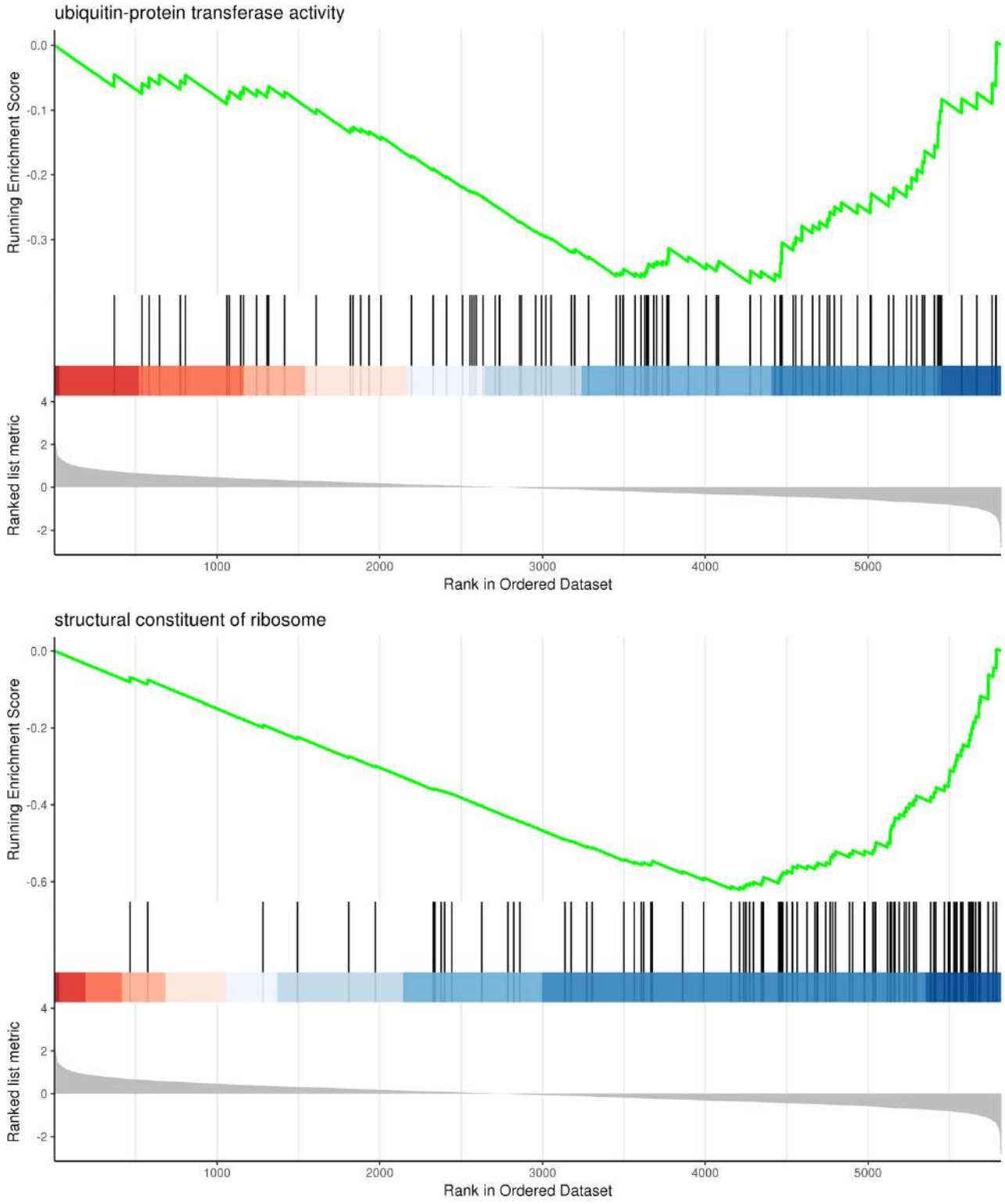

Thermogenesis
Running Enrichment Score
0.0
-0.1
-0.2
-0.3
-0.4
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset


Parkinson disease
Running Enrichment Score
0.0
-0.1
-0.2
-0.3
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset


Huntington disease
Running Enrichment Score
0.0
-0.1
-0.2
-0.3
-0.4
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset


cytochrome-c oxidase activity
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset


heme binding
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
Ranked list metric
4
2
0
-2
1000
2000
3000
4000
5000
Rank in Ordered Dataset



Note: In the upper figure, X-axis: Position of gene set after ordering; Y-axis: Enrichment score; The lines on the top represent genes in the gene set. Green curve shows the enrichment score of each gene set across positions. In the lower figure, X-axis: Position of gene set after ordering. Y-axis: Score. Each line represents a gene in gene set. The length of lines indicates corresponding score.

3.14 DEU analysis

Differential Exon Usage (DEU) analysis aims at revealing differentially expressed genes at exon level. For experiments with biological replicates, DEXSeq is employed in DEU analysis. DEXSeq [29] identifies differentially expressed gene by use of negative binomial generalized linear models (GLM). Threshold for differential expression was set as FDR<0.01.

Outputs of DEU analysis:

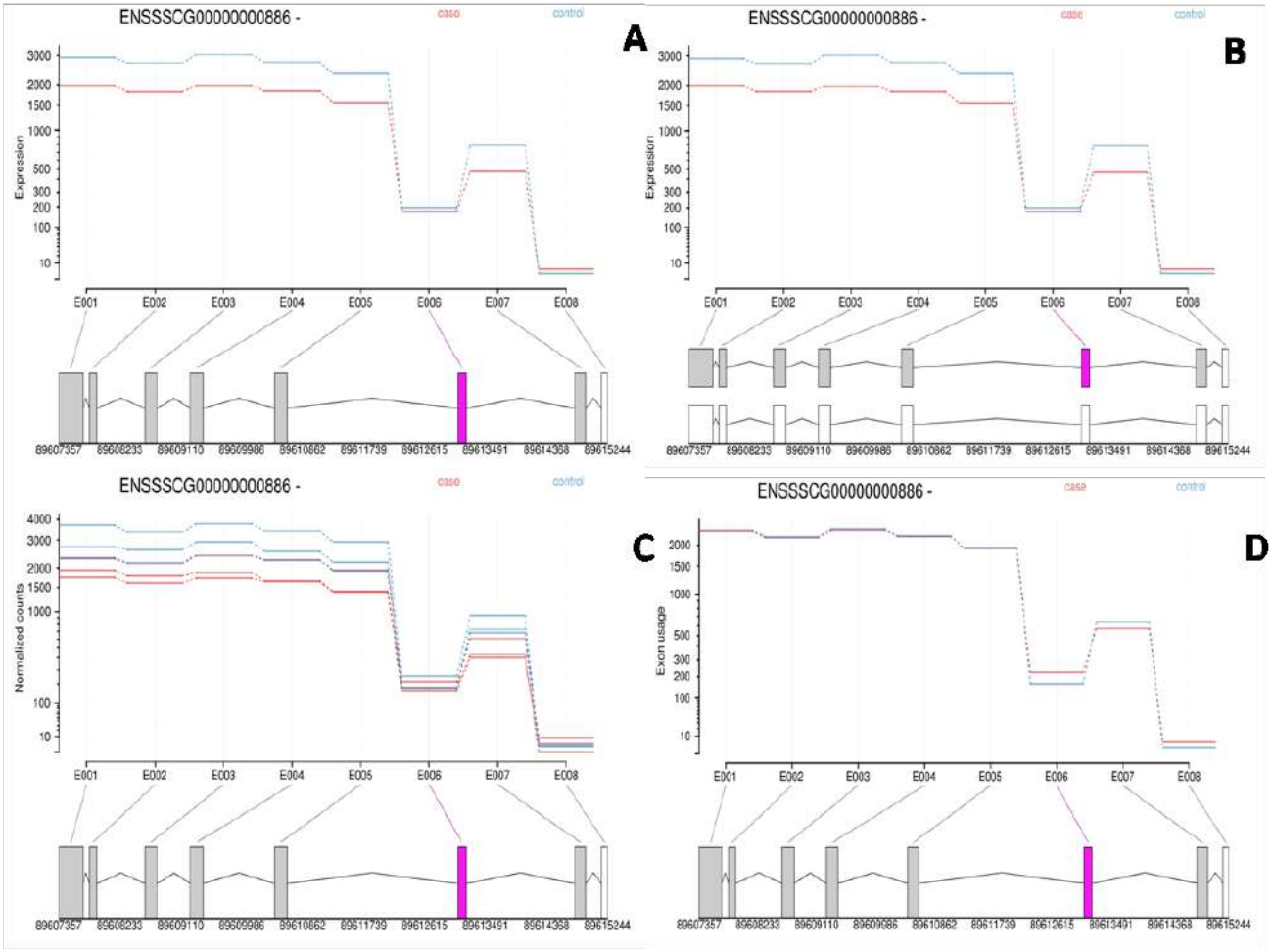
N1_N2_N3_vs_T1_T2_T3DEU.final.xls

Example: Outputs of DEU analysis

geneID	Symbol	Exon ID	log2(FC)	p value	FDR
ENSMUSG00000026028	Trak2	E004	0.17	6.6744729097847e-06	0.0083
ENSMUSG00000026280	Atg4b	E022	2.71	7.63813212697088e-06	0.0089
ENSMUSG00000026532	Spta1	E040	-1.1	6.47621408337145e-06	0.0083
ENSMUSG00000038733	Wdr26	E015	0.19	5.10274976013689e-06	0.0072
ENSMUSG00000037395	Rcor3	E011	-15	1.69642841193394e-06	0.0036
ENSMUSG00000047648	Fbxo30	E003	0.21	5.0724601362087e-09	3.18369680228261e-05
ENSMUSG00000020125	Elane	E003	-0.85	8.84503076243205e-07	0.0024
ENSMUSG00000035242	Oaz1	E004	0.054	8.60886397370575e-06	0.0096
ENSMUSG00000035242	Oaz1	E011	0.55	5.74589919366954e-06	0.0077

Note: geneID: Gene ID;
exonID: Exon ID;
Log2(FC): log2(Fold change);
pvalue: Significancy of difference;
FDR: False discovery rate.

Demo figures of DEU were as follows:



Note: (A) Fitted expression. The plot represents the expression estimates from a call to testForDEU. Shown in red is the exon that showed significant differential exon usage.(B) Transcripts. As in Figure A, but including the annotated transcript models.(C) Normalized counts. As in Figure A, with normalized count values of each exon in each of the samples. (D) Fitted splicing. The plot represents the estimated effects, as in Figure A, but after subtraction of overall changes in gene expression.

DEU outputs in html version

N1_N2_N3_vs_T1_T2_T3testForDEU.html

3.15 Gene fusion analysis

Gene fusion refers to end-to-end hybridization of coding regions of two or more genes. These genes form a chimeric gene that shares a same regulatory sequences including promoter, enhancer, RBS, terminator, etc.) The gene product of fusion genes is named fusion protein. Candidate fusion genes is selected by Fusionmap, which mapped pair end sequences in transcriptome sequencing data to reference genome. False positive recognitions were removed by blasting against Nt or other database.

Note: Red line represents gene fusion occurred on the same chromosome. Green line represents gene fusion occurred across chromosomes.

Statistics of gene fusion events

References

1. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research. 1998,8 (3): 175-185.
2. Kim D, Langmead B, Salzberg S L. HISAT: a fast spliced aligner with low memory requirements[J]. Nature methods, 2015, 12(4): 357-360.
3. Pertea M, Pertea G M, Antonescu C M, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads[J]. Nature biotechnology, 2015, 33(3): 290-295.
4. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010, 20(9): 1297-1303.
5. Cingolani P, Platts A, Wang L L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012, 6(2): 80-92.
6. Trapnell C, Williams BA, Pertea G, Mortazavi A, et al. Transcript assembly and quantification by RNA Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology. 2010, 28(5):511-515.
7. Florea L, Song L, Salzberg S L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Research, 2013, 2:188.
8. Buchfink B, Xie C, Huson DH, "Fast and sensitive protein alignment using DIAMOND", Nature Methods 12, 59-60 (2015)
9. Deng YY, Li JQ, Wu SF, Zhu YP, et al. Integrated nr Database in Protein Annotation System and Its Localization. Computer Engineering. 2006, 32(5):71-74.
10. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the universal protein knowledgebase. Nucleic acids research. 2004, 32: D115-D119.
11. Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000, 25(1): 25-29.
12. Tatusov RL, Galperin MY, Natale D A. The COG database: a tool for genome scale analysis of protein functions and evolution. Nucleic Acids Research. 2000, 28(1):33-36.
13. Koonin EV, Fedorova ND, Jackson JD, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome biology. 2004, 5(2): R7.
14. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic acids research. 2013: gkt1223.
15. Kanehisa M, Goto S, Kawashima S, Okuno Y, et al. The KEGG resource for deciphering the genome. Nucleic Acids Research. 2004, 32:D277-D280.
16. Jones P, Binns D, Chang H Y, et al. InterProScan 5: genome-scale protein function classification[J]. Bioinformatics, 2014, 30(9): 1236-1240.
17. Eddy S R. Profile hidden Markov models. Bioinformatics, 1998, 14(9): 755-763.
18. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009, 25(8): 1026-1032.
19. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. Nature. 2012, 489(7414): 101-108.

20. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002, 297:1183-1186.
21. Kasper D. Hansen, Zhijin Wu, et al. Sequencing technology does not eliminate biological variability. Nature Biotechnology. 2011, 572-573.
22. Robasky K, Lewis NE, Church GM. The role of replicates for error mitigation in next-generation sequencing. Nature Reviews Genetics. 2013.
23. Insights into the correlation between Physiological changes in and seed development of tartary buckwheat (Fagopyrum tataricum Gaertn.). BMC Genomics. 2018 Aug 31;19(1):648.
24. Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550. doi: 10.1186/s13059-014-0550-8.
25. Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140
26. Alexa A, Rahnenfuhrer J. topGO: enrichment analysis for gene ontology. R package version 2.8, 2010.
27. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research. 2013, 41: D808-D815.
28. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003, 13(11): 2498-2504.
29. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome research, 2012, 22(10): 2008-2017.
30. Shen S., Park JW., Lu ZX., Lin L., Henry MD., Wu YN., Zhou Q., Xing Y.(2014) rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.PNAS, 111(51):E5593-601. doi: 10.1073/pnas.1419161111
31. Khan A, Oriol Fornés, Stigliani A, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework[J]. Nucleic Acids Research, 2017, 77(21):e43.
32. Ge T, Boris L. TFBSTools: an R/bioconductor package for transcription factor binding site analysis:[J]. Bioinformatics, 2016, 32(10):1555-1556.
33. Nicolle R, Radvanyi F, Elati M. COREGNET: reconstruction and integrated analysis of co-regulatory networks[J]. Bioinformatics, 2015:btv305.

Appendix

Appendix1: Software list

Table. Software list

Tools	Description	Link
HISAT2	A spliced read mapper for RNA-Seq	http://ccb.jhu.edu/software/hisat2/index.shtml
StringTie	Transcript assembly for RNA-Seq	https://ccb.jhu.edu/software/stringtie/index.shtml
ASprofile	ASprofile is a suite of programs for extracting, quantifying and	http://ccb.jhu.edu/software/ASprofile/

Tools	Description	Link
	comparing alternative splicing (AS) events from RNA-seq data	
BLAST	Basic Local Alignment Search Tool	http://blast.ncbi.nlm.nih.gov/Blast.cgi
DESeq	An R package for RNA-Seq Differential Expression Analysis based on a model using the negative binomial distribution	http://www.bioconductor.org/packages/release/bioc/html/DESeq.html
EBSeq	An R package for RNA-Seq Differential Expression Analysis based on Bayesian approach	https://www.biostat.wisc.edu/~kendzior/EBSEQ/
Cytoscape	An open source software platform for visualizing complex networks	http://www.cytoscape.org/
topGO	An R package for gene ontology enrichment analysis	#
rMATs	MATS is a computational tool to detect differential alternative splicing events from RNA-Seq data.	http://rnaseq-mats.sourceforge.net/
TFBStools	An R package for the analysis and manipulation of transcription factor binding sites.	http://www.bioconductor.org/packages/release/bioc/html/TFBSTools.html

Note: # represents no links for software at the third column. The bioinformatic analysis softwares not given in the report are developed by us, and are not shown with the table.

Appendix2: Database list

Table. Database table

Database	Description	Homepage
NR	non-redundant protein sequence database	ftp://ftp.ncbi.nih.gov/blast/db/
Swiss-Prot	A manually annotated, non-redundant protein sequence database	http://www.uniprot.org/
GO	Gene Ontology database	http://www.geneontology.org/
COG	The database of Clusters of Orthologous Groups of proteins	http://www.ncbi.nlm.nih.gov/COG/
KOG	The database of Clusters of Protein homology	http://www.ncbi.nlm.nih.gov/KOG/
Pfam	The database of Homologous protein family	http://pfam.xfam.org/
KEGG	The database of Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://www.string-db.org/
Ensembl	Database Sscrofa10.2 download from	http://asia.ensembl.org/index.html
Cosmic	COSMIC, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.	https://cancer.sanger.ac.uk/cosmic
JASPAR	Database of transcription factor binding profiles	http://jaspar.genereg.net/

Appendix3: Nucleic acid coding list

Table. Nucleic acid coding table

Nucleic Acid Code	Meaning	Mnemonic
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
R	A or G	puRine
Y	C, T or U	pYrimidines
K	G, T or U	bases which are Ketones
M	A or C	bases with aMino groups
S	C or G	Strong interaction
W	A, T or U	Weak interaction
B	not A (i.e. C, G, T or U)	Bcomes after A
D	not C (i.e. A, G, T or U)	Dcomes after C
H	not G (i.e., A, C, T or U)	Hcomes after G
V	neither T nor U (i.e. A, C or G)	Vcomes after U
N	A C G T U	Nucleic acid

Appendix4: Description on annotation databases

Table. Description on annotation databases

Database name	Database description
NR database	Non-redundant protein database in NCBI, including Swissprot, PIR (Protein Information Resource), PRF (Protein Research Foundation), PDB (Protein Data Bank) protein database and CDS from GenBank and RefSeq
Swissprot database	A database maintained by EBI (European Bioinformatics Institute) containing a collated database of protein annotation information with relevant references and high credibility
COG database	A database for homologous classification of gene products. It is an early database for the identification of orthologous genes, which is obtained by comparing a large number of protein sequences of various organisms.
KOG database	For eukaryotes, homologous genes from different species are divided into different Ortholog clusters based on gene orthologous relationships and evolutionary relationships. Currently, KOG has 4852 classifications. Genes from the same Ortholog have the same function, so that functional annotations can be directly inherited to other members of the same KOG cluster.
Pfam database	The most comprehensive classification system for protein domain annotations. Proteins are composed of domains, and the protein sequences of each particular domain are somewhat conserved. Pfam divides the protein domain into different protein families, and establishes an HMM statistical model of the amino acid sequence of each family through alignment of protein sequences.
GO database	The internationally standardized gene function classification system provides a dynamically updated standard vocabulary to fully describe the functional properties of genes and gene products in organisms. There are three main categories of the database, namely molecular function, cellular component and biological process, each describing the molecular function that the gene product may perform, and the cellular environment and Participation in biological processes. The most basic concept in the GO database is Term, each entry has a Term name, such as "cell", "fibroblast growth factor receptor binding" or "signal transduction", with a unique number, like GO:nnnnnnn
KEGG database	A database that systematically analyzes the metabolic pathways of gene products in cells and the function of these gene products. It integrates data on genomics, chemical molecules, and biochemical systems, including PATHWAY, DRUG, DISEASE, GENES, and GENOME. Using this database helps to study the genes and their expressions as a whole network.

Note: # represents no links for software at the third column. The bioinformatic analysis software not given in the report are developed by us, and are not shown with the table.

Appendix5: Materials and methods

1 Sample collection and preparation

1.1 RNA quantification and qualification

RNA concentration and purity was measured using NanoDrop 2000(Thermo Fisher Scientific, Wilmington, DE). RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

1.2 Library preparation for Transcriptome sequencing

A total amount of 1 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer（5X）. First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase. Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H . Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of preferentially 240 bp in length, the library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 μl USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95°C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. At last, PCR products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

1.3 Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina platform and paired-end reads were generated.

2 Data analysis

2.1 Quality control

Raw data (raw reads) of fastq format were firstly processed through in-house perl scripts. In this step, clean data(clean reads) were obtained by removing reads containing adapter, reads containing ploy-N and low quality reads from raw data. At the same time, Q20, Q30, GC-content and sequence

duplication level of the clean data were calculated. All the downstream analyses were based on clean data with high quality.

2.2 Comparative analysis

The adaptor sequences and low-quality sequence reads were removed from the data sets. Raw sequences were transformed into clean reads after data processing. These clean reads were then mapped to the reference genome sequence. Only reads with a perfect match or one mismatch were further analyzed and annotated based on the reference genome. Hisat2 tools soft were used to map with reference genome.

2.3 Gene functional annotation

Gene function was annotated based on the following databases: Nr (NCBI non-redundant protein sequences) ; Nt (NCBI non-redundant nucleotide sequences) ; Pfam (Protein family) ; KOG/COG (Clusters of Orthologous Groups of proteins) ; Swiss-Prot (A manually annotated and reviewed protein sequence database) ; KO (KEGG Ortholog database) ; GO (Gene Ontology).

2.4 SNP calling

Picard - tools v1.41 and samtools v0.1.18 were used to sort, remove duplicated reads and merge the bam alignment results of each sample. GATK2 or Samtools software was used to perform SNP calling. Raw vcf files were filtered with GATK standard filter method and other parameters (clusterWindowSize: 10; MQ0 >= 4 and (MQ0/(1.0*DP)) > 0.1; QUAL < 10; QUAL < 30.0 or QD < 5.0 or HRun > 5), and only SNPs with distance > 5 were retained.

2.5 Quantification of gene expression levels

Quantification of gene expression levels. Gene expression levels were estimated by fragments per kilobase of transcript per million fragments mapped. The formula is shown as follow:

$$FPKM = \frac{cDNAFragments}{MappedFragments(Millions) * TranscriptLength(kb)}$$

2.6 Differential expression analysis

For the samples with biological replicates:
Differential expression analysis of two conditions/groups was performed using the DESeq2. DESeq2 provide statistical routines for determining differential expression in digital gene expression data using a model based on the negative binomial distribution. The resulting P values were adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate. Genes with an adjusted P-value < 0.01 found by DESeq2 were assigned as differentially expressed.
For the samples without biological replicates:

Differential expression analysis of two samples was performed using the edgeR. The FDR < 0.01 & Fold Change≥2 was set as the threshold for significantly differential expression.

2.7 GO enrichment analysis

Gene Ontology (GO) enrichment analysis of the differentially expressed genes (DEGs) was implemented by the GOseq R packages based Wallenius non-central hyper-geometric distribution (Young et al, 2010), which can adjust for gene length bias in DEGs.

2.8 KEGG pathway enrichment analysis

KEGG (Kanehisa et al., 2008) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (http://www.genome.jp/kegg/). We used KOBAS (Mao et al., 2005) software to test the statistical enrichment of differential expression genes in KEGG pathways.

2.9 PPI (Protein Protein Interaction)

The sequences of the DEGs was blast (blastx) to the genome of a related species (the protein protein interaction of which exists in the STRING database: http://string-db.org/) to get the predicted PPI of these DEGs. Then the PPI of these DEGs were visualized in Cytoscape (Shannon et al, 2003).

References

1. AltschulEwing B, Hillier L, Wendl S F, Madden T L, Schaffer A A, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402. (BLAST)
2. Anders S, Huber W. (2010).Differential expression analysis for sequence count data. Genome Biology,doi:10.1186/gb-2010-11-10-r106. (DESeq)
3. Finn R D, Tate J, Mistry J, et al. (2008). The Pfam protein families database. Nucleic Acids Res 36, D281-D288. (Pfam)
4. Gotz S, Garcia-Gomez J M, Terol J, et al. (2008).High-throughput functional annotation and data mining with the Blast2GO suite.Nucleic Acids Research 36, 3420-3435. (BLAST2go)
5. Mao X, Cai T, Olyarchuk J G, et al. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.Bioinformatics 21, 3787-3793. (KOBAS)
6. McKenna A, Hanna M, Banks E, et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303. (GATK)
7. Li H, Handsaker B,Wysoker A. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009,25 (16): 2078-2079.(Samtools)
8. Kanehisa M, Araki M, Goto S, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids research 36:D480-D484. (KEGG)

9. Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550. doi: 10.1186/s13059-014-0550-8.(DESeq2)
10. Robinson M D, McCarthy D J, Smyth G K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.(edgeR)
11. Young M D, Wakefield M J, Smyth G K, et al. (2010).Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, doi:10.1186/gb-2010-11-2-r14. (GOseq)
12. Shannon P, Markiel A, Ozier O, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498-2504. (Cytoscape)