Report of Bioinformatics

Coupling Structure Dynamics with Translation Regulation

Group 2

Instructor: Lu Zhi
Instructing TA: Liu Xiaofan
Group Members: Chen Yini, Li Na, Xing Yicai, Li Chenxi

Thesis contents

Report of Bioinformatics: Coupling Structure Dynamics with Translation Regulation

Chen Yini, Li Na, Xing Yicai, Li Chenxi, Tsinghua university, 2021

Abstract

Based on SHAPE-map sequencing data and Ribo-seq sequencing data, we studied possible posttranscriptional regulation, structural changes and translation efficiency of Arabidopsis thaliana before and after stimulation. Through various existing tools and statistical analysis methods, we linked the information of differential splicing, differential expression, structural changes and differential translation in the two cases, and found that the secondary structure of RNA had a certain correlation with differential splicing, and it would affect the transcription efficiency. These secondary structures that affect gene expression have a certain motif.

Background

1.1 Post-transcriptional regulation of RNA

Eukaryotes, including human and Arabidopsis, have post transcriptional regulation, a regulation of gene expression at RNA level(figure 1). It occurs between the transcription and translation stages of gene expression.[1] The process includes: capping, RNA scattering, addition of poly (a) tail, RNA editing, mRNA stability, nuclear export. RNA splicing is one of the steps. RNA splicing is mainly to remove introns, which are transcribed into non coding regions of RNA, so that mRNA can produce proteins. The cell binds to both sides of the intron through the splicing body, circulates the intron into a circle, and then cuts it. The two ends of the exon are then linked together.

1.2 Probing RNA secondary structure

Dimethyl sulfate (DMS) probe combined with next-generation sequencing (NGS) mutation spectrum analysis (DMS-MaPseq) is a new method to reveal the whole genome or targeted RNA structure. In this method, RNA was modified by dimethylsulfate (DMS), which can accurately reflect the real folding state of RNA in living cells.(figure 2)

In 2018, Zhang's research group proposed an improved DM-MaPseq method for plant materials, which optimized the DMS processing conditions and simplified the library preparation process.[2]

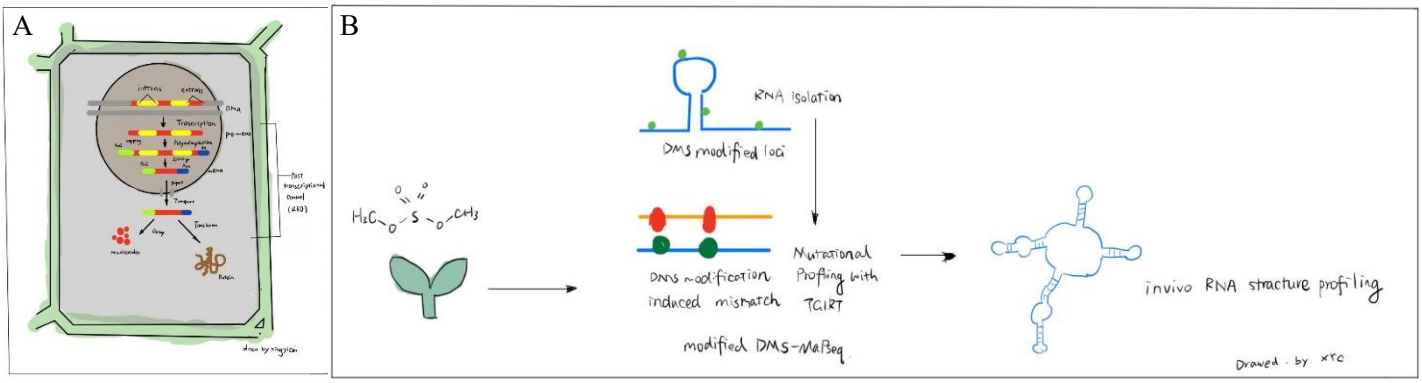


Figure1:A The process includes: capping, RNA scattering, addition of poly (a) tail, RNA editing, mRNA stability, nuclear export. RNA splicing is one of the steps .**B** zhang's experiment results.

However, DMS still have some defect. The probe called 2'hydroxyacylation (SHAPE) have better effect for DMS only probes the structural information of As (adenylic acid) and Cs (cytosine).Whereas information on the pairing status of Us (uracil) and Gs (guanylate) is missing.

Reagents such as n-methylisatinic anhydride (NMTA) and 1-methyl-7-nitroisatinic anhydride (1M7) react with the 2 '- hydroxyl group of RNA to form adducts on the 2' - hydroxyl group of the RNA backbone. Compared with chemicals used in other RNA probing techniques, these reagents have the advantage of not being substantially biased toward base identity, while remaining very sensitive to conformational dynamics. Constrained nucleotides, usually by base pairing, show less adduct formation than unpaired nucleotides. Adduct formation was quantified for each nucleotide in a given RNA by extending the complementary DNA primer with reverse transcriptase and comparing the resulting fragment with that from an unmodified control. This way unpaired RNAs are detected, allowing inference of RNA secondary structures.[3,4]

This experiment relied mainly on SHAPE-MaP.

1.3 RNA secondary structure and expression

1.3.1 Steam-loop structure of mRNA

RNA molecules have two properties in eukaryotic cells: they have a natural tendency to form highly stable secondary and tertiary structures in vitro and in vivo [5,6], and the second is that alterations in these stereo structures represent a well-known regulatory mechanism for many RNA cellular processes.

Stem-loop intramolecular base pairing is a common pattern that generally occurs on single stranded RNA. This structure is also known as a hairpin or hairpin loop. The resulting structure is a key component of many RNA secondary structures. As an important secondary structure of RNA, it can guide the folding of RNA, protect the structural stability of messenger RNA, and provide recognition sites for RNA binding proteins.

1.3.2 Molecular biological evidence

Many studies have shown that the secondary structure of 5’untranslated region (5’UTR) of RNA usually reduces the efficiency of translation initiation, thus reducing the total protein production. However, there are few studies on the extent to which the secondary structure of CDs and 3′ UTR affects the protein production.[7,8]

Studies have shown that RNA secondary structures reduce elongation because ribosomes have to unravel every structure they encounter during translation. Therefore, in highly translated mRNA, the intensity of mRNA secondary structure is considered to be decreased. There was a positive correlation between mRNA folding intensity and protein abundance.[9]

The secondary structure of mRNA shortens the distance of ribosome through the dynamic change of folding strength. It is worth noting that when adjacent ribosomes are close to each other, the secondary structure of mRNA between them disappears, and the use of codons determines the elongation. More importantly, in highly translated mRNAs, the interaction of mRNA secondary structure and codon usage leads to a shorter ribosomal distance in the structural region, thus eliminating the structure in the process of translation and leading to high elongation.

1.3.3 Recent hypothesis

A study this year a team found that RNA secondary structure can regulate splicing, so it has a strong role in gene regulation. The most complete list of conserved complementary regions (PCCR) in human protein coding genes is proposed. PCCRs tend to occur in introns, inhibit inserted exons, and block hidden and inactive splice sites. The double stranded structure of PCCRs is supported by the decreased nucleotide accessibility of icSHAPE, the high abundance of RNA editing sites and the frequent occurrence of forked eCLIP peaks. The response of introns containing PCCRs to RNAPII moderation showed an obvious splicing pattern, which indicated that splicing was widely influenced by CO transcriptional RNA folding. The enrichment of the 3’- end in PCCRs raises an interesting hypothesis that the coupling between RNA folding and splicing can mediate the co transcriptional inhibition of pre mRNA premature cleavage and polyadenylation.[10]

Data

2.1 Molecular biological preparation

2.1.1Arabidopsis thaliana

Arabidopsis thaliana, a small flowering plant native to Eurasia and Africa with a relatively short life cycle, is a popular model organism in plant biology and genetics.[11]

The most commonly used mutant lines are *ler* (Landsberg erecta) and *col* or *Columbia*[11] other mutant lines less cited in the scientific literature are *WS*, or *wassilewskija*, *C24*, *CVI,* or Cape Verde Islands, a group of closely related accessions named *Col-0, CoL-I. Col-0* is an Arabidopsis ecotype and a direct descendant of *Col-1* donated via AIS.

2.1.2 Genetic treatment

The experimental group receiving external stimuli and the control group receiving no external stimuli were set up, respectively. SHAPE map experiments as well as control experiments without chemical modification as well as ribo-SEQ experiments were subsequently performed on experimental and control plants. SHAPE map uses NAI as a chemical modifier, whereas control experiments were treated with the addition of an equal volume of DMSO solution. Three biological replicates were performed for each set of experiments.

2.2 Raw data

2.2.1 Content

The data of this experiment come from teachers and do not need us to do wet experiment operation. Three sections mainly include transcriptome data (RNA-seq raw data), ribosome footprinting analysis data (Ribo-seq raw data) and secondary structure data(SHAPE-seq raw data). Each file includes samples from six plants.

Primary versus secondary structure sequencing files of RNA do not have to be described in excess. Ribo-seq, also known as ribosome foot-printing, uses specialized messenger RNA (mRNA) sequencing to determine which mRNAs are being actively translated, based on the finding that mRNAs within ribosomes can be isolated by nucleases that degrade unprotected mRNA regions. This technique analyzes the regions in which mRNAs are translated into proteins, as well as the level of translation in each region, to provide insight into overall gene expression.[13]

2.2.2 Data format

RNA-seq and ribo-seq files were generated using BAM format. BAM file is a compressed binary version of the SAM file used to represent alignment sequences up to 128 MB. Bam and Sam formats are designed to contain the same information. Sam format is easier to read and easier to handle by traditional text-based processing procedures. The BAM format provides binary versions of most of the same data and is designed to be reasonably well compressed.[14]

The SHAPE-seq raw data were generated using fastq format. The details can be refer to 3.1.

Technology

This part is mainly to explore and learn some of the main software used in this experiment and its main underlying principles, which is used to supplement the extra-curricular learning of bioinformatics

3.1 Fastq storage format

The FASTA format is a format of stored sequences that can store either nucleic acid sequences (DNA / RNA) or amino acid sequences of proteins (amino acid sequences, referred to as AA sequences), which are mainly divided into 2 parts. The first part is a line starting with " > ", stored mainly descriptive information of the sequence, and the second part is the sequence part, in a certain format. The format was originally developed by the *Sanger Institute of the Wellcome foundation* to integrate FASTA format sequences and their quality data. At present, fastq format has become the de facto standard for preserving high-throughput sequencing results.[15]

3.2 Bowtie algorithm

Bowtie is an ultrafast, memory efficient alignment program for aligning short DNA sequence reads to large genomes. For the genome, burrows Wheeler indexing allows bowtie to align more than 25 million reads per CPU hour with a memory footprint of ~ 1.3 GB. [16]. It is a software package commonly used for bioinformatic

sequence alignment and sequence analysis. [17] Its algorithm, for the BWT algorithm, was written in the C ++ language.

3.3 STAR

The splicing transcription alignment reference (star) software of RNA SEQ alignment algorithm is a bioinformatics software commonly used for splicing analysis. It was originally developed to align large (> 80 billon reads) coding transcriptome RNA SEQ datasets. Star is more than 50 times faster than other comparators in mapping speed, and improves the alignment sensitivity and accuracy. In addition to unbiased ab initio detection of typical junctions, star can also detect atypical splicing and chimeric (fusion) transcripts, and can map full-length RNA sequences. Using roche454 to sequence the RT-PCR amplicons, dobin laboratory verified 1960 new splicing links between genes, with a success rate of 80-90%, which confirmed the high precision of star mapping strategy. [18]

3.4 ShapeMapper

ShapeMapper automatically calculates RNA structure probing reactivity through mutation profiling (MAP) experiment. The chemical adducts on RNA are detected as internal mutations of cDNA by reverse transcription and read out by large-scale parallel sequencing.[19,20] ShapeMapper integrates careful processing of all sequence changes caused by adducts, sequence variation correction, basic call quality filters and quality control warnings, and now can accurately identify RNA adduct sites through careful manual analysis of electrophoresis data (the previous highest precision standard).[21]

3.5 FeatureCounts

FeatureCounts is a reading summary program, which is suitable for calculating the reading data generated by RNA or genomic DNA sequencing experiments. FeatureCounts implements efficient chromosome hashing and feature block techniques. It is much faster than existing methods (an order of magnitude for gene level abstracts) and requires less computer memory. It can be used for single ended or paired end reading and provides a wide range of options for different sequencing applications.[22]

3.6 htseq-count

Htseq is a python library for analyzing high-throughput sequencing (HTS) data. Htseq provides parsers for many common data formats in HTS projects, as well as classes that represent data, such as genome coordinates, sequences, sequencing reads, alignments, gene model information and variant calls, and provides data structures that allow queries through genome coordinates. Htseq-count is a tool developed with htseq. It preprocesses RNA SEQ data by calculating the overlapped reading with genes for differential expression analysis.[23]

3.7 Xtail

Xtail is the method published by Professor Yang of our university in 2016.[24]

This is an analysis pipeline tailored for ribosome profiling data that can comprehensively and accurately identify differentially translated genes in pairwise comparisons. The close regulation of the mRNA translation process is key to the precise control of protein abundance and quality. Ribosome profiling, a combination of ribosome footprinting and RNA deep sequencing, has been used in numerous studies to quantify genome-wide mRNA translation. Applied to both simulated and real datasets, xtail has high sensitivity and minimal false positive rate, outperforming existing methods in quantifying the accuracy of differential translation. Through published datasets of ribosome profiling, Yang using xtail revealed not only biologically meaningful differentially translated genes, but also perturbed mTOR signaling in human cancer cells and interferons in human primary macrophages- γ （IFN- γ ） New events in differential translation- γ) treatment in theirs paper. This demonstrates the value of xtail in providing new insights into the molecular mechanisms involved in translational dysregulation.

3.8 DESeq

This method was first proposed by Simon Anders in 2010. This method is based on negative binomial distribution. It connects the variance and mean value through local regression, and realizes deseq as R / Bioconductor package. So far, this article has been cited more than 10,000 times.[25]

High throughput sequencing analysis such as RNA SEQ,

chip SEQ or bar code counting provides quantitative readings in the form of counting data. In order to correctly infer the difference signal in these data and have good statistical ability, it is necessary to estimate the data variability and appropriate error model in the whole dynamic range.

3.9 EdgeR

EdgeR is a Bioconductor package for detecting differential expression of copy count data. It was first proposed in 2009 by the laboratory of Davis J. McCarthy, will this article has been cited more than 20,000 times. He used an over dispersed Poisson model to account for biological and technical variability. Empirical Bayes methods were used to moderate the degree of transcript overdispersion and improve the reliability of inference. This method can be used as long as at least one phenotype or experimental condition has been replicated, even with the lowest level of replication. In addition to sequencing data, this software may have other applications such as proteome peptide count data. Digital gene expression (DGE) technology surpasses microarray technology in many functional genomics applications. A fundamental data analysis task, especially for studies of gene expression, involves determining whether there is evidence that counts of transcripts or exons differ significantly across experimental conditions.[26]

Experiment

For the detailed process and specific code of this part, please refer to the attachment "script.pdf".

4.1 RNA-seq analysis

4.1.1 Mapping

First of all, we use fastqc to check the quality of the data, and then use fastp default parameters to automatically carry out all-round quality control of the data. Use bowtie to compare the fastp quality control file obtained in the previous step to the rRNA index, and remove the part compared to the rRNA index, so as to get the file without rRNA reads. Finally, genome index was constructed by using *Arabidopsis* genome sequence file and reference genome annotation file.

We use Star software to compare the results. In order to prevent star from sorting in the sort process, we do not sort in the star process, but use samtools to sort separately. We use samtools to sort by TAG value, and then use samtools index to index the sorted sequence. Then the comparison result file after mapping and sorting index by samtools is used as input, and the number of reads is calculated by featurecounts software.

4.1.2 Data analysis

EdgR and DEseq2 were used for differential expression analysis. **EdgR:** Construct the DDS matrix, normalize the original DDS, normalize the normalization coefficient sizefactor, estimate the degree of gene dispersion, carry out statistical test, obtain the analysis results and analyze the differences. **DEseq2:** Firstly, we construct the DEGlist object, then filter out the low counts data, and use TMM algorithm to standardize DEGlist to get the difference analysis results.

We use the comparison result file after mapping and sorted and indexed by samtools as the input file for subsequent analysis. We use rmats to calculate the splitting event, filter the rmats script, extract the different splicing genes of different splicing events after filtering, and download them from the server. Then we go to *David functional annotation tool website* for analysis. Finally, the differential splicing events are visualized by using the software *rmats2sashimiplot*.

4.2 Ribo-seq analysis

4.2.1 Mapping

Just like RNA mapping, we first check the quality of the data, and use fastp default parameters to automatically conduct all-round quality control of the data.

Then use bowtie to compare the fastp quality control file obtained in the previous step to the rRNA index, and remove the part compared to the rRNA index, so as to get the file without rRNA reads file. We use Star software to compare the results and use samtools to sort by TAG value, and then use samtools index to index the sorted sequence. Then we use the comparison result file after mapping and sorted and indexed by samtools as input. Use htseq count software to calculate the number of reads

4.2.2 Data analysis

We use metaplots in ribo-code to obtain 3-nt periodicity and ORF analysis reports. Then we use *xtail*'s ribo-seq based count matrix and RNA-seq's count matrix to calculate the differential translation efficiency. For the convenience of comparison, RNA-seq only counts the reads of CDs region. The R package *xtail* is used to analyze the translation differences.

4.3 Structure analysis

We took the integrated all transcript results calculated by shapemapper to calculate the resulting structurally altered regions. We calculated the hit level, looked at the distribution of the hit level, and, after determining the threshold, computed the normalization factor, computed the activity preprocessing and normalization, merging the structurally changed regions.

4.4 Data intergration

4.4.1 Transcripts abundant changes and translation efficiency changes analysis

We represent the relationship between pre - and post light changes in transcript abundance and the extent of pre - and post light changes in translational efficiency (TE) by scatter plots.

4.4.2 Transcription structure changes and translation efficiency changes analysis

Here we utilized hypothesis testing to examine whether the degree of structural alteration affects the degree of translational efficiency TE changes. We define the structural alteration genes as hit level >1 and have structure changed region (|delta Gini index|>0.1). Structure up-regulation when delta Gini index > 0.1 and structure down-regulation when delta Gini index < - 0.1.

4.4.3 Transcription structure changes and translation efficiency changes analysis

We enriched sequence motifs for 3'UTR and 5'UTR of all genes with TE up and TE down regulated, respectively. All 3'UTR / 5'UTR fastq files were extracted, and the TE up / down regulated genes were listed, in merge_5utr.fasta and merge_3utr.fasta) were screened, and the resulting 5'UTR, 3'UTR sequence information for genes with TE changes was obtained.

Finally, motif analysis was performed using meme online. Motif information on the input sequence was predicted using meme's motif discovery tool

Results

5.1 Data matrix preparation

5.1.1 RNA-seq analysis

After completing the reads processing, removal RNA and mapping work of the samples, mapped reads can be obtained and quality control related plots drawn (**figure 2**), and the RNA-seq reads count matrix can be calculated.

gene_id	CD1_1	CD1_2	CD1_3
AT1G01010	174	176	154
AT1G01020	439	481	524
AT1G01030	2	1	1
AT1G03987	41	47	45
		…	
ATCG01310	1	1	0

Chart 1. RNA-seq reads count matrix (27628 lines)

5.1.2 Ribo-seq analysis

After completing the reads processing, removal RNA and mapping work of the samples, mapped reads can be obtained and quality control related plots drawn (**figure 3**), and the RNA-seq reads count matrix can be calculated.

gene_id	CD1_1	CD1_2	CD1_3
AT1G01010	258	166	157
AT1G01020	47	3	18
AT1G01030	30	0	12
AT1G01040	323	273	165
		…	
ATCG01410	0	4	0

Chart 1. RNA-seq reads count matrix

5.1.3 SHAPE-seq analysis

The mean reactivity values for AT1G09530.3 in a given sample were **0.00205639**.

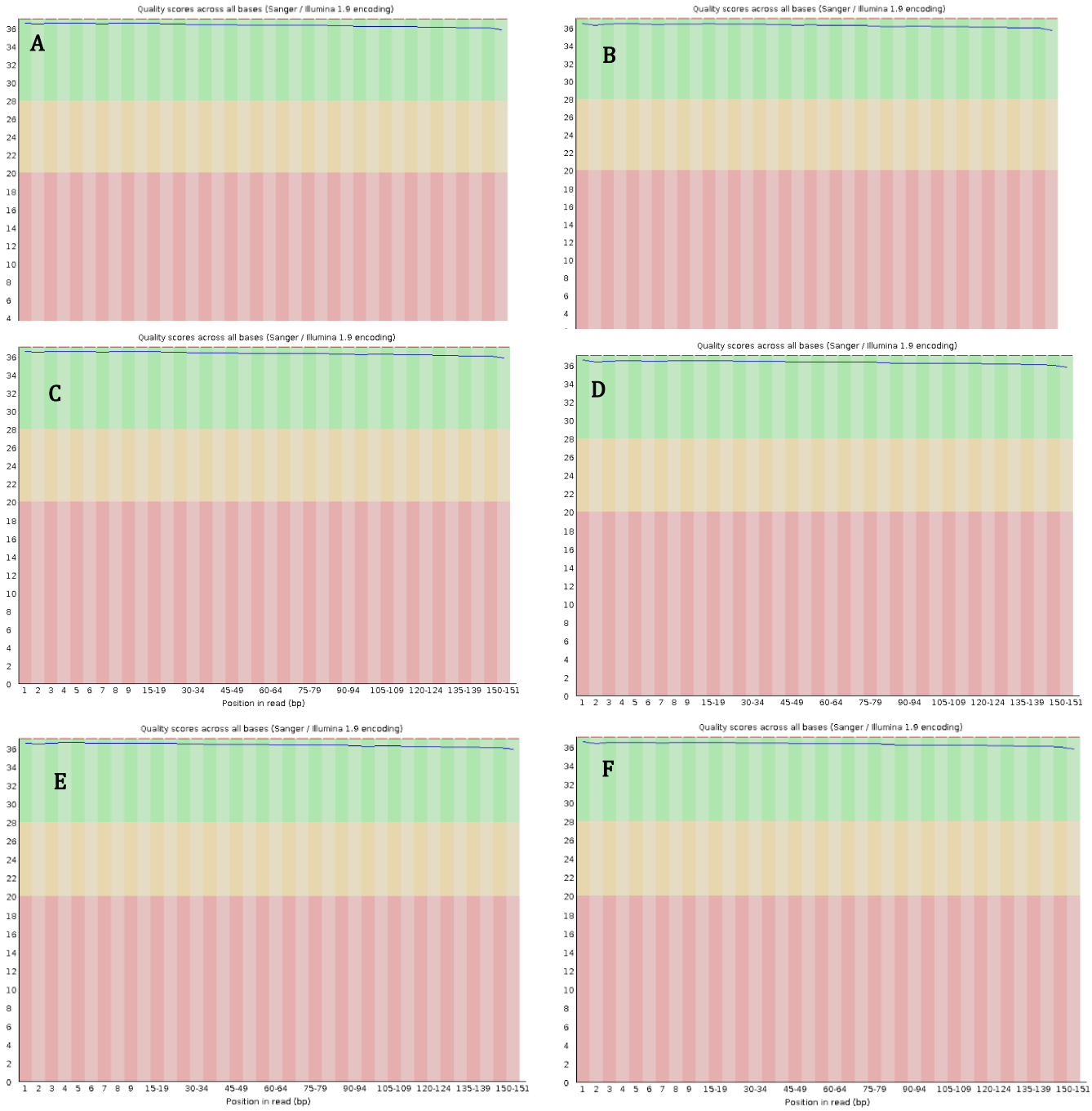


Figure 2. Per base sequence quality.(by fastqc, RNA-seq)
The horizontal axis in the figure is from base 1 to base 151 of the sequencing sequence, and the vertical axis is the quality score. The red line in the figure represents the median, and the thin blue line is the line of the mean value for each position. (A) Per base sequence quality of the data in CD1-1.clean.1(+); (B) CD1-1.clean.2(-); (C) CD1-2.clean.1; (D) CD1-2.clean.2; (E) CD1-3.clean.1; (F) CD1-3.clean.2.

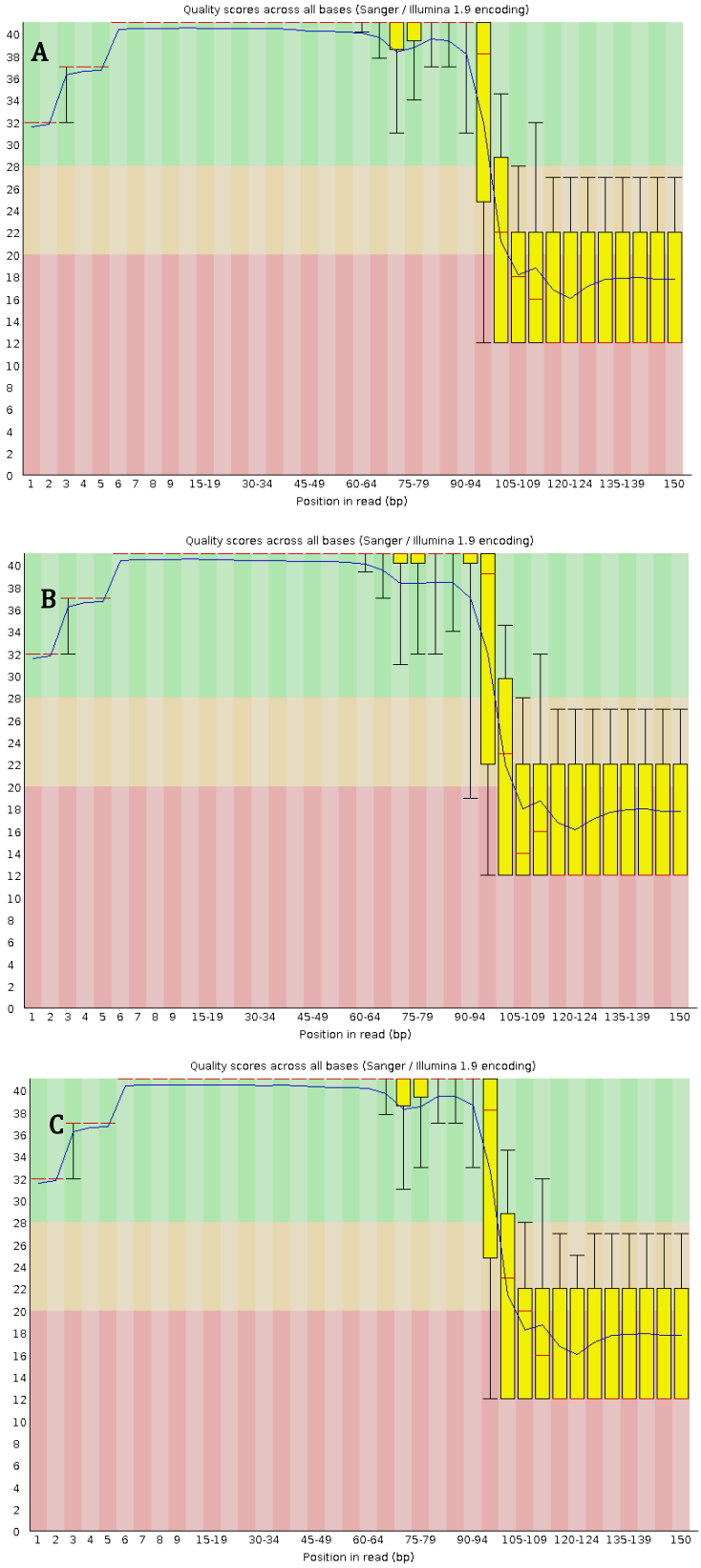


Figure 3. Per base sequence quality. (by fastqc, ribo-seq)
The horizontal axis in the figure is from base 1 to base 151 of the sequencing sequence, and the vertical axis is the quality score. The red line in the figure represents the median, and the thin blue line is the line of the mean value for each position. (A) Per base sequence quality of the data in CR1-1.; (B) CR1-2.; (C) CR1-3;.

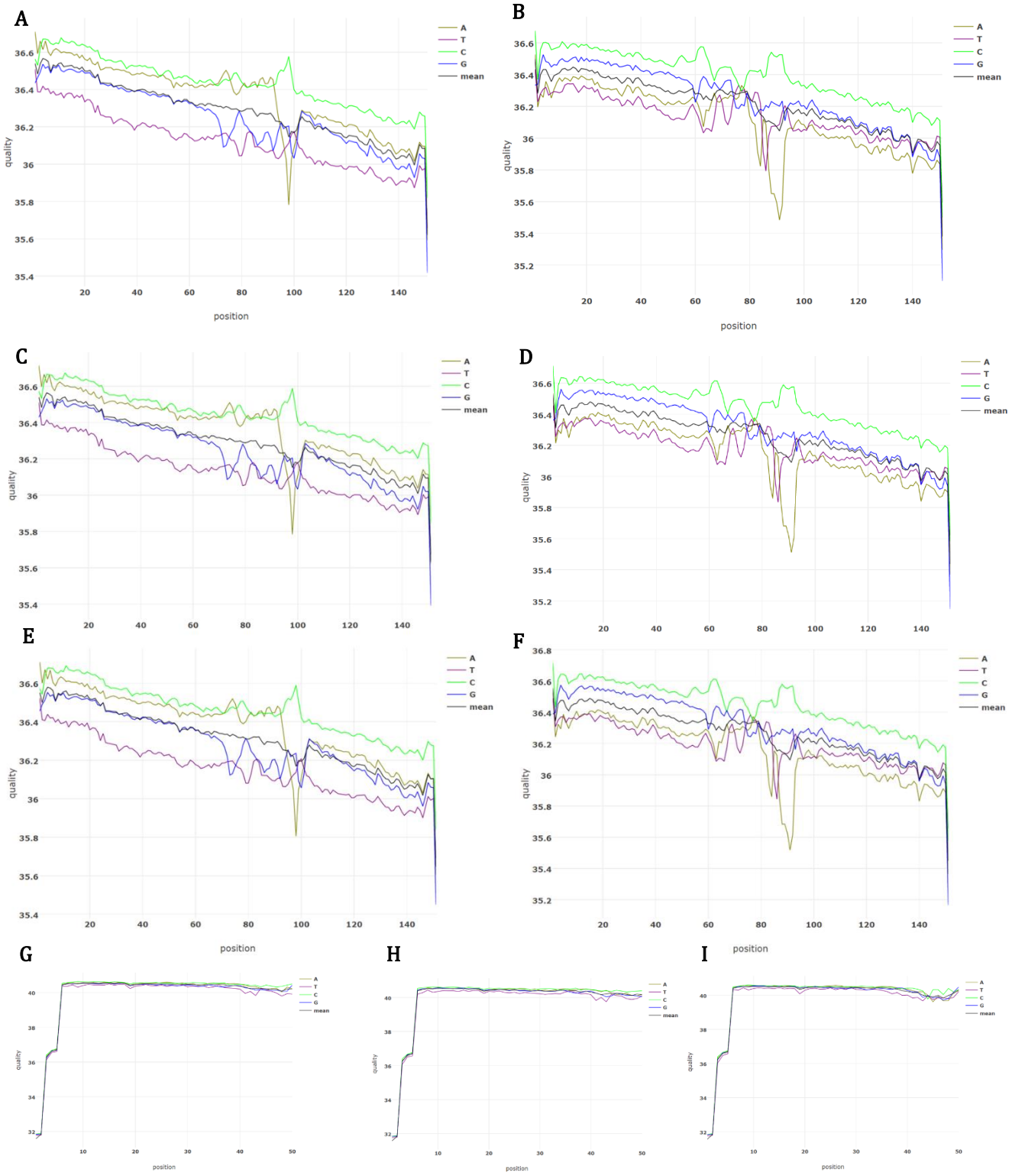


Figure 4: **reads quality**
Generally speaking, the base mass distribution at different positions should be more than 30 with less fluctuation, which is a good data. Q20 and q30 represent the percentage of a certain base quality value in the total base number, which is similar to the qualified rate of products. Different quality standards will produce different qualified rates. The higher the standard is, the better the quality is, and the less qualified products will be; The higher the pass rate is, the more qualified data will be. Generally speaking, for the second generation sequencing, it is better to achieve more than 95% of the bases of Q20 (the worst is not less than 90%) and more than 85% of the bases of q30 (the worst is not less than 80%). RNA-seq:(A) Per base sequence quality of the data in CD1-1.clean.1(+); (B) CD1-1.clean.2(-); (C) CD1-2.clean.1; (D) CD1-2.clean.2; (E) CD1-3.clean.1; (F) CD1-3.clean.2.ribo-seq: (H) Per base sequence quality of the data in CR1-1.; (I) CR1-2.; (J) CR1-3;.

5.2 GO/KEGG pathway analysis

5.2.1 Pathway map of differentially expressed genes

In the stimulated Arabidopsis thaliana, the expression of some genes was down regulated (**figure5, G**), while another part was up-regulated (**figure5, F)**. According to the existing knowledge of plant physiology, different stimuli will respond to different responses. Therefore, we can infer the stimulation of Arabidopsis through these differentially expressed genes. This part will be mainly described in the first section of the discussion part.

Genes whose expression was upregulated mainly included Taurine and hypotaurine metabolism, phenylpropanoid biosynthesis, cyanoamino acid metabolism.

Genes whose expression was downregulated mainly included squalene synthase, PAI1, circadian rhythm, flavonoid biosynthesis, diterpenoid biosynthesis, glutathione metabolism (glutathione S-transferase tau 9/GSTU9, spermidine synthase 3/SPDS3), sesquiterpenoid and triterpenoid biosynthesis, flavone and flavonol biosynthesis and cutin, suberin wax biosynthesis.

5.2.2 Pathway map of differential splicing genes

Differentially spliced genes mainly include glycosylphosphatidylinostitol(GPI)-anchor biosynthesis, plant-pathogen interaction, spliceosome and SNARE interaction in vesicular transport, fatty acid degeneration(ACX6 and KAT5) .(**figure 5, D, E, F**)

5.2.3 Pathway map of differentially translated genes

The upregulated genes mainly included tryptophan metabolism, phenylpropanoid biosynthesis and N-Glycan biosynthesis. (**figure5, B**)

The down regulated genes (**figure5, A**) mainly included plant-pathogen interaction, trehalose-6-phosphate synthase(TPS1), nitrogen metabolism, cyanoamino acid metabolism, glutathione metabolism (glutathione S-transferase tau 9/GSTU9, spermidine synthase 3/SPDS3).

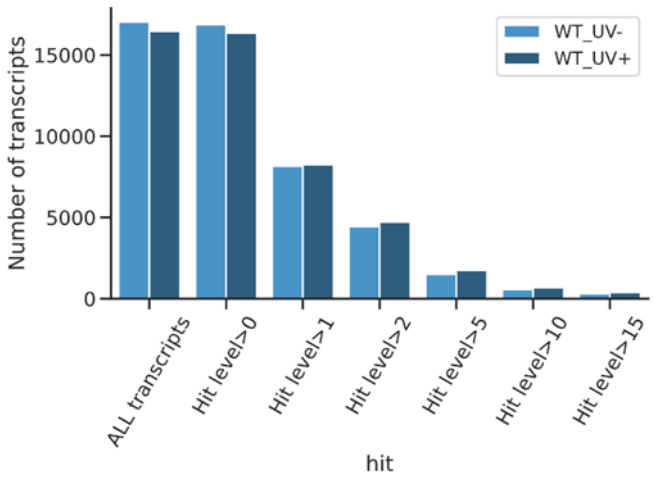
Some of the gene in this result maybe presence in 2 distinct pathways, final results are therefore we pooled.

5.3 SHAPE-seq analysis

In this part we obtained a trend plot of transcripts as a function of hit level (lower panel).

apendix: summary_result _merge_50_1. CSV: Result for transcripts with hit level greater than 0.

summary_result_merge_50_1_New.csv:Result for transcri - pts with hit level greater than 2.



The transcripts used for normalization were “AT3G41768.1”, “ATMG01390.1”, “AT3G06355.1”.

5.4 Data intergration

5.4.1 Relationship between TE and RNA seq

See figure 6, A.Fold change is a measure that describes how much a quantity changes between the original and subsequent measurements. It is defined as the ratio between two quantities; For quantities a and B, then the folding change of B relative to a is B / A.

5.4.2 The relationship between the degree of structural change and the degree of TE change in translation efficiency

See figure 6, B.

```
shape up,TE down
5.387478374073877e-23
shape down,TE up
0.7150732345900125
```

5.4.3 Motif analysis

See figure 6, C-F. In the results, we select according to evaluate, and the motif found by sites has appeared in several structural change areas, so we can calculate the coverage rate of the motif. We choose the motif with a higher coverage rate, where we take more than 4%.

A

Publist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Plant-pathogen interaction	RT		16	1.7	6.9E-5	5.3E-3
	KEGG_PATHWAY	Starch and sucrose metabolism	RT		11	1.2	4.0E-3	1.5E-1
	KEGG_PATHWAY	Nitrogen metabolism	RT		6	0.6	7.9E-3	2.0E-1
	KEGG_PATHWAY	Cyanoamino acid metabolism	RT		6	0.6	3.3E-2	6.3E-1
	KEGG_PATHWAY	Glutathione metabolism	RT		7	0.7	6.0E-2	8.0E-1
	KEGG_PATHWAY	Biotin metabolism	RT		3	0.3	6.4E-2	8.0E-1
	KEGG_PATHWAY	Arginine and proline metabolism	RT		5	0.5	7.4E-2	8.0E-1

B

Publist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Ribosome	RT		27	5.0	4.7E-7	3.8E-5
	KEGG_PATHWAY	Tryptophan metabolism	RT		6	1.1	5.2E-3	2.1E-1
	KEGG_PATHWAY	Phenylpropanoid biosynthesis	RT		8	1.5	9.3E-2	1.0E0
	KEGG_PATHWAY	N-Glycan biosynthesis	RT		4	0.7	9.4E-2	1.0E0

C

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Plant-pathogen interaction	RT		16	1.7	6.9E-5	5.3E-3
KEGG_PATHWAY	Starch and sucrose metabolism	RT		11	1.2	4.0E-3	1.5E-1
KEGG_PATHWAY	Nitrogen metabolism	RT		6	0.6	7.9E-3	2.0E-1
KEGG_PATHWAY	Cyanoamino acid metabolism	RT		6	0.6	3.3E-2	6.3E-1
KEGG_PATHWAY	Glutathione metabolism	RT		7	0.7	6.0E-2	8.0E-1
KEGG_PATHWAY	Biotin metabolism	RT		3	0.3	6.4E-2	8.0E-1
KEGG_PATHWAY	Arginine and proline metabolism	RT		5	0.5	7.4E-2	8.0E-1

D

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	RT		3	2.2	3.2E-3	4.8E-2

E

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Plant-pathogen interaction	RT		6	1.6	2.7E-2	1.0E0
KEGG_PATHWAY	Spliceosome	RT		6	1.6	5.4E-2	1.0E0
KEGG_PATHWAY	SNARE interactions in vesicular transport	RT		3	0.8	9.4E-2	1.0E0

F

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Fatty acid metabolism	RT		3	2.5	3.8E-2	1.0E0
KEGG_PATHWAY	Metabolic pathways	RT		13	11.0	6.8E-2	1.0E0

G

Term	RT	Genes	Count	%	P-Value	Benjamini
Biosynthesis of secondary metabolites	RT		40	8.6	3.4E-6	1.3E-4
Circadian rhythm - plant	RT		8	1.7	4.1E-6	1.3E-4
Flavonoid biosynthesis	RT		5	1.1	6.0E-4	1.3E-2
Diterpenoid biosynthesis	RT		4	0.9	8.2E-3	9.7E-2
Glutathione metabolism	RT		7	1.5	8.6E-3	9.7E-2
Sesquiterpenoid and triterpenoid biosynthesis	RT		4	0.9	9.2E-3	9.7E-2
Flavone and flavonol biosynthesis	RT		2	0.4	5.7E-2	5.1E-1
Cutin, suberine and wax biosynthesis	RT		3	0.6	9.4E-2	7.0E-1

H

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
KEGG_PATHWAY	Taurine and hypotaurine metabolism	RT		3	1.2	4.1E-3	7.8E-2
KEGG_PATHWAY	Phenylpropanoid biosynthesis	RT		6	2.4	4.2E-3	7.8E-2
KEGG_PATHWAY	Biosynthesis of secondary metabolites	RT		14	5.7	1.5E-2	1.9E-1
KEGG_PATHWAY	Cyanoamino acid metabolism	RT		3	1.2	6.5E-2	6.0E-1

Figure 5: **A** . Functions corresponding to genes with altered transcription efficiency in the stimulated versus unstimulated groups in Arabidopsis， **B**. Functions corresponding to genes whose translation efficiency was upregulated in the stimulated versus the unstimulated group in Arabidopsis. **C.** Functions corresponding to genes whose translation efficiency was downregulated in the stimulated versus the unstimulated group in Arabidopsis. **D.** Functions corresponding to differentially spliced genes in stimulated versus unstimulated groups in Arabidopsis (A3SS). **E**.Functions corresponding to differentially spliced genes in stimulated versus unstimulated groups in Arabidopsis(RI). **F.** Functions corresponding to differentially spliced genes in stimulated versus unstimulated groups in Arabidopsis(SE).**G.** Functions corresponding to down regulated genes expressed in stimulated versus unstimulated groups in Arabidopsis. **H**. Functions corresponding to up-regulated genes in stimulated versus unstimulated groups in Arabidopsis.

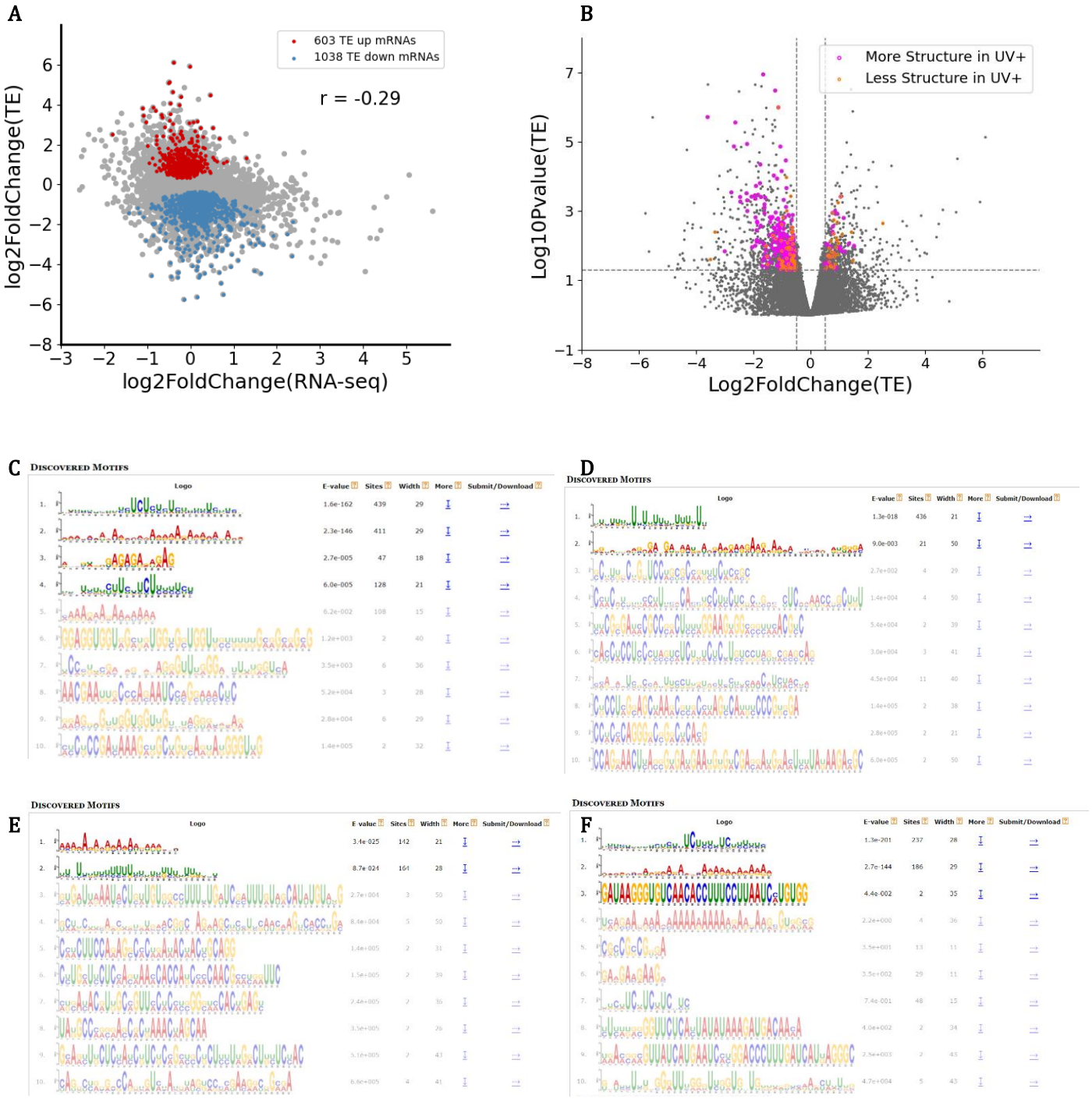


Figure 6: **Data intergration**
(A) Plot of log2foldchange (TE) and log2foldchange (RNA-seq) (B) The relationship between the degree of structural change and the degree of te change in translation efficiency; motif analysis: (C) merge_TE_down_5UTR.fasta; (D) merge_TE_down_3UTR.fasta; (E) merge_TE_up_3UTR.fasta; (F) merge_TE_up_5UTR.fasta

Discussion ■

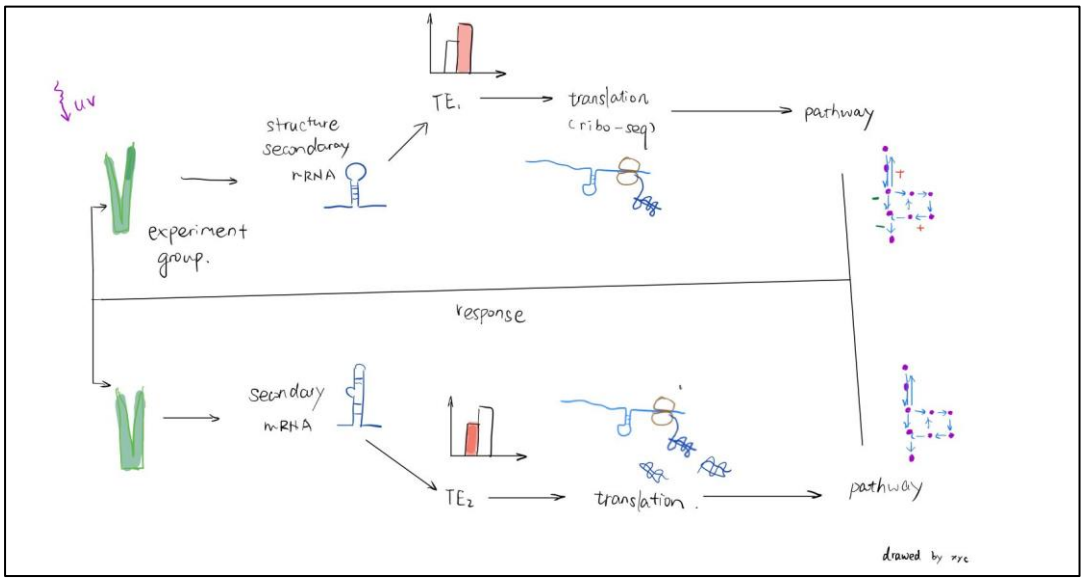


Figure 7: **main processes of the experiment**
This experiment is based on SHAPE-MAP sequencing data and Ribo-seq sequencing data to determine the possible post transcriptional regulation, structural changes and the resulting translation efficiency of Arabidopsis thaliana before and after stimulation. Through various existing tools and statistical analysis methods, we hope to link differential splicing, differential expression, structural changes and other information under the two conditions with differential translation, hoping to find the results with biological significance, such as establishing a dynamic model of RNA secondary structure closely related to external stimuli, and exploring the relationship between structural unfolding dynamics and translation efficiency

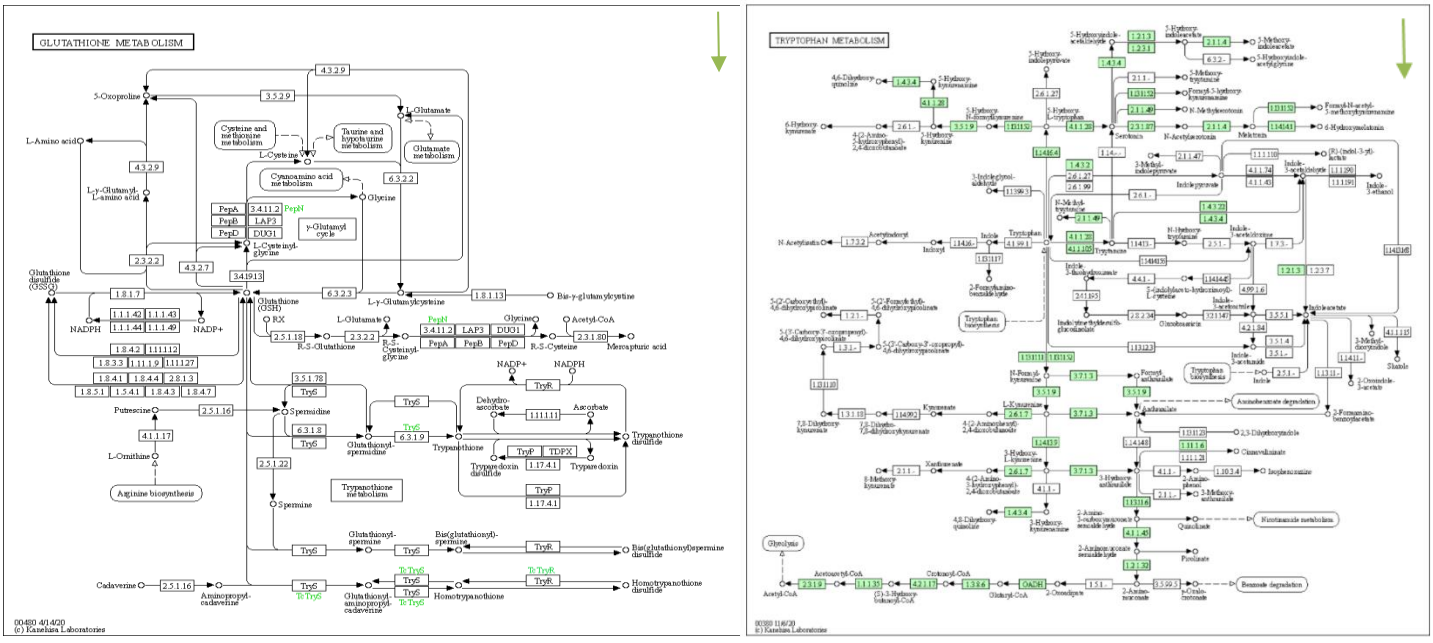


Figure 8: **glutathione metabolism and tryptophan metabolism** glutathione and tryptophan metabolism are required for immunity during the hypersensitive response of Arabidopsis thaliana to hemiascomycetes.

6.1Response of Arabidopsis thaliana in experimental groups to UV environment

From the pathway analysis data, we can know that both glutathione metabolism and tryptophan metabolism were down regulated in the treated plants. Studies have shown that glutathione and tryptophan metabolism are required for immunity during the hypersensitive response of Arabidopsis thaliana to hemiascomycetes.[27] γ - Glutamylcysteine synthetase GSH1 and tryptophan (TRP) metabolism(figure 8), which are essential for glutathione biosynthesis, contribute to HR (a strong immune response found in plants) and block the development of fungal pathogens with a semi trophic infection pattern.

In other words, UV light irradiation may eventually lead to the weakening of plant immune response（It's a correlation, not necessarily a direct causal relationship). Another evidence is that plant-pathogen interaction is weaken in the experiment group. A research results showed that UV irradiation could induce the antibacterial activity of Arabidopsis thaliana. Pathogenic bacteria in plants are usually susceptible, and the key role of UV induced DNA damage is pointed out. They also suggest that UV treatment can bypass the identification requirements for microorganisms. Parasite molecules activate immune response through Arabidopsis protein. [28]

6.2 Hit level and mutation rate of RNA

SHAPE map structural analysis read out by massively parallel sequencing provides a valuable tool for structural interrogation of RNA at the single nucleotide level.[29] Several other methods with similar targets have been developed. To compare the read depth requirements of SHAPE map (and its mutation spectrum reads) with other methods, we calculated the 'hit level'. The hit level metric quantifies the total background subtracted signal for each transcript nucleotide:

$$hit\ level = \frac{total\ events_S - \frac{read\ depth_S}{read\ depth_B} \times total\ events_B}{transcript\ length}$$

Since mutation counts in shape map are directly proportional to read depth, we estimated the relationship between sequencing read depth and hit level by dividing the observed hit level by the median read depth under the experimental conditions.

High resolution RNA structure detection and modeling require the detection of most or all RNA under high hit rate. Detecting a single region with a low hit rate, even if the overall average hit rate is $\geqslant$ 1, may contain significant errors. In pars experiment, assuming that the background of digestion data is zero, the minimum threshold of each transcript nucleotide is 1, and the average read stop is 5,50, corresponding to hit level 1. Similarly, a report describing chemical detection of DMS, structure SEQ, uses a similar threshold of $\geqslant$ 10 stops per a or C nucleotide; This is equivalent to an estimated hit level (according to our definition) of 0.2, assuming that signal:background ratio= 1.7 (estimated from extended data figure 1, panel D in Ref. 10), half of the nucleotides of all transcripts were a or C. The creator of dms-seq11 needs at least 15 reads per a or C on average.

Reactive Nan nucleotides with an apparent mutation rate above 0.02 in any untreated sample were excluded from analysis. These artifacts were identified as regions of at least 10 nucleotides where three or more of 10 nucleotides showed mutation rates above 0.03 in the absence of Nai treatment, or modified mutation rates above 0.1 under any condition, and were also excluded from analysis. Nucleotides with a read depth below 100 under any condition were also excluded from analysis. Reactivity normalization reactivities were normalized in each probing condition to the average of the 92-98 percentile reactivities of nucleotides from rRNA that were sequenced to high depth and showed large hit levels (hit level > 0) in the experimental probing condition.

6.3 Data intergration

The results showed that 603 translatable fragments were up-regulated and 1038 down regulated in tens of thousands of structural genes. These results were obtained through strict statistical tests and controlled trials.

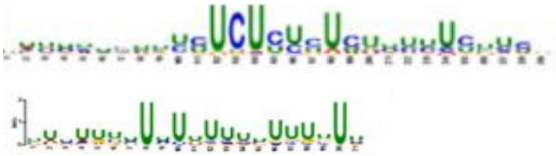
We can see that **(figure 6.A)** there is a weak correlation between the abundance of transcripts and translation efficiency. The translation efficiency of some genes has improved a lot, but the abundance of transcripts is still not

high, which indicates that the translation efficiency has nothing to do with the quantity of mRNA. This suggests that the secondary structure of RNA may play a greater role in the change of TE.

On the contrary, we can see from this picture **(figure 6.B)** whether the degree of structural change affects the efficiency of translation. More structure can lead to lower translation efficiency. The degree of structural change is strongly related to the efficiency of translation. The secondary structure of mRNA of most genes with different translation efficiency has changed greatly.

Finally, we obtained the most conserved primary structure motif among these secondary structure altered mRNA, which has the strongest correlation with the secondary structure. We chose different motifs of up-regulated and down regulated genes in the two site (3'UTR&5'UTR).

We can see some features from this motif.



These two motifs are down regulated transcriptional genes. Many U repeats can be seen in the motif of 3'UTR. Studies have shown that cell proteins can bind to the poly (U) region of RNA3 'untranslated region, which may slow down the efficiency of translation. [29]

It may also be related to the formation of secondary structure. At present, the research of protein from primary structure to secondary structure is relatively mature, and RNA may have similar characteristics, which may need the comprehensive consideration of topology, information technology and thermodynamics.

Conclusion

In this experiment, we stimulated Arabidopsis thaliana in the experimental group, and by analyzing the results of RNA sequencing, ribosomal sequencing and RNA secondary structure sequencing, we got the conclusion that RNA secondary structure will affect the transcription efficiency.

This experiment mainly focuses on high-throughput data analysis. What we get is a macroscopic and general experimental result. I think its theoretical significance is far more than its biological significance. In the follow-up of this experiment, I put forward two tentative directions.

The first is to study the central principle of a single gene after stimulation. We can study the sequence changes, mRNA abundance and secondary structure changes of a single gene after stimulation, the processes involved in the expression of the gene (e.g. with trans acting factors), and the role of the expressed products of the gene in specific biochemical metabolic pathways.

Another idea is that we can see whether the motif is species-specific and whether there is a certain rule in the up-regulation and down-regulation of various genes, so as to find the bottom logic that the secondary structure of mRNA affects the efficiency of translation.

The secondary structure may affect translation through its own steric hindrance, or it may interact with other macromolecules. These are very important biological problems that we can't see in our experimental data.

References

1. Franks A, Airoldi E, Slavov N (May 2017). "Post-transcriptional regulation across human tissues". PLoS Computational Biology. 13 (5): e1005535.
2. Zhiye Wang, Meiyue Wang, Tian Wang, Yijing Zhang, Xiuren Zhang, Genome-wide probing RNA structure with the modified DMS-MaPseq in Arabidopsis, Methods, Volume 155, 2019, Pages 30-40, ISSN 1046-2023.
3. Low JT, Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods. 2010 Oct;52(2):150-8.
4. Mortimer SA, Weeks KM (2007). "A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry". J Am Chem Soc. 129 (14): 4144–45.
5. Brion, P., and E. Westhof. 1997. Hierarchy and dynamics of RNA folding. Annu. Rev. Biophys. Biomol. Struct. 26:113-137.
6. Conn, G. L., and D. E. Draper. 1998. RNA structure. Curr. Opin. Struct. Biol. 8:278-285.
7. Y. Ding et al., In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505, 696–700 (2014).
8. Y. Wan et al., Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505, 706–709 (2014)
9. Mao Y, Liu H, Liu Y, Tao S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. Nucleic Acids Res. 2014;42(8):4813-4822.
10. Kalmykova, S., Kalinina, M., Denisov, S. et al. Conserved long-range base pairings are associated with pre-mRNA processing of human genes. Nat Commun 12, 2300 (2021).
11. Hoffmann MH (2002). "Biogeography of Arabidopsis thaliana (L.) Heynh. (Brassicaceae)". Journal of Biogeography. 29: 125–134.
12. "Eurasian Arabidopsis Stock Centre (uNASC)". arabidopsis.info.
13. Ingolia NT (March 2014). "Ribosome profiling: new views of translation, from single codons to genome scale". Nature Reviews. Genetics. 15 (3): 205–13.
14. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. (2009). "The Sequence Alignment/Map format and SAMtools" (PDF). Bioinformatics. 25 (16): 2078–2079.
15. Cock, P. J. A.; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2009, 38 (6): 1767–1771.
16. Langmead, B., Trapnell, C., Pop, M. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009).
17. "Bowtie 2: fast and sensitive read alignment". bowtie-bio.sourceforge.net. Retrieved 2021-03-28
18. Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics, Volume 29, Issue 1, January 2013, Pages 15–21
19. Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods. 2014, 11(9):959-65.
20. Busan S, Weeks KM. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA. 2018, 24(2):143-148.
21. Busan S, Weeks KM. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA. 2018 Feb;24(2):143-148.
22. Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics, Volume 30, Issue 7, 1 April 2014, Pages 923–930
23. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.
24. Xiao, Z., Zou, Q., Liu, Y. et al. Genome-wide assessment of differential translations with ribosome profiling data. Nat Commun 7, 11194 (2016).
25. Anders, S., Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106 (2010).
26. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140.
27. Plant immunity during HR to hemibiotrophs;Kei

Hiruma, Satoshi Fukunaga, Paweł Bednarek, Mariola Piślewska-Bednarek, Satoshi Watanabe, Yoshihiro Narusaka, Ken Shirasu, Yoshitaka TakanoProceedings of the National Academy of Sciences Jun 2013, 110 (23) 9589-9594

28. Kunz BA, Dando PK, Grice DM, Mohr PG, Schenk PM, Cahill DM. UV-induced DNA damage promotes resistance to the biotrophic pathogen Hyaloperonospora parasitica in Arabidopsis. Plant Physiol.
29. Siegfried NA, Busan S, Rice GM, Nelson JA, Weeks KM. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods. 2014;11(9): 959-965.
30. Luo G. Cellular proteins bind to the poly(U) tract of the 3' untranslated region of hepatitis C virus RNA genome. Virology. 1999 Mar 30;256(1):105-18. doi: 10.1006/viro.1999.9639. PMID: 10087231.