Report for Ribo-Seq Analysis for Arabidopsis Thaliana

PhD Ornela Maloku

December 2020

1 Introduction

Arabidopsis Thaliana belongs to the angiosperm family with its small nuclear genome around (125 Mb). It has a very short generation time compared to many other plants (6–8 weeks) and this makes it a good candidate to be studied over other plant species, it is a self-fertilizing plant, with a diploid chromosome number of 10 (five pairs) and it produces a large number of seeds each generation making it easy for genetic screens and variant analysis. Thanks to these advantages, Arabidopsis is a good model for the investigation of the translational regulation of the circardian clock in plants. The main topic of this project is the study and analysis of 36 Arabidopsis Thaliana samples obtained through Ribo-Seq experiments (also known as ribosome footprinting). All the calculations computed up to now are carried out in order to understand which are the profiles for which the positional information is coherent.

The figure below describes the experimental design carried out.

2 Processing of Data

2.1 Datasets Acquisition of raw fastq files

From Ribo-Seq experiments 36 samples in total were sequenced in outsourcing. In order to take in consideration the circardia clock variable, 12 Time

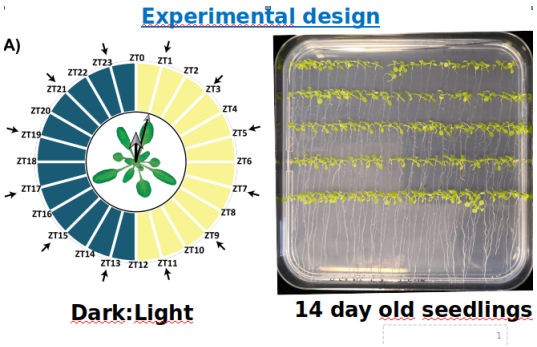


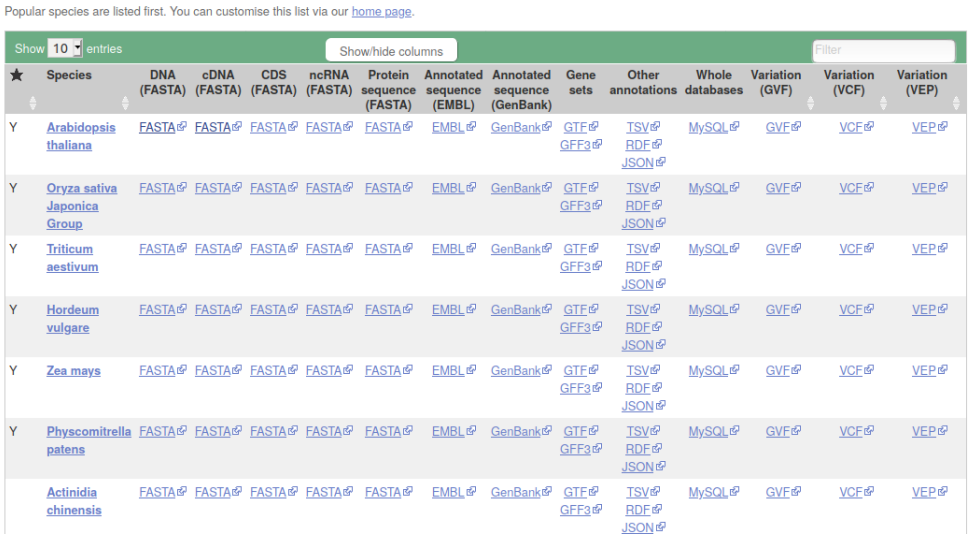
Figure 1: Experimental design for Arabidopsis Thaliana samples

Points (TP) were considered and for each Time Point 3 replicates fro each sample: TP1-1,TP1-2,TP1-3,TP3-1,TP3-2,TP3-3,TP5-1,TP5-2,TP5-3,TP7-1,TP7-2,TP7-3,TP9-1,TP9-2,TP9-3,TP11-1,TP11-2,TP11-3,TP13-1,TP13-2,TP13-3,TP15-1,TP15-2,TP15-3,TP17-1,TP17-2,TP17-3,TP19-1,TP19-2,TP19-3,TP21-1,TP21-2,TP21-3,TP23-1,TP23-2,TP23-3.
The files were acquired in fastq format.

2.2 Upstream pre-processing of raw data

The first step of the pre-processing is the trimming of the adapter sequences with CUTADAPT 3.1 (`https://cutadapt.readthedocs.io/en/stable/guide.html`).
For the 36 samples the adapter sequence to trim was TGGAATTCTCGGGTGCCAAGG.
In order to align Ribo-Seq reads onto the reference genome we used HISAT2 2.20 (`http://daehwankimlab.github.io/hisat2/`). Instead of using the genome as a reference (Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz), we used the transcriptome Arabidopsis_thaliana.TAIR10.cdna.all.fa.gz. The FASTA file were downloaded from (`https://plants.ensembl.org/index.html`). In the picture below you can see how the database EnsemblPlants is constructed.

To map the reads and to assemble and quantify the transcripts represented by those reads,the hisat-build function was used. The function takes in input the FASTA file Arabidopsis_thaliana.TAIR10.cdna.all.fa.gz and generates a set of files (1.ht2, .2.ht2, .3.ht2, .4.ht2, .5.ht2, .6.ht2, .7.ht2.) which build the index file for the alignment step.



Popular species are listed first. You can customise this list via our home page.

Show 10 entries | Show/hide columns | Filter

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)
Y	Arabidopsis thaliana	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Oryza sativa Japonica Group	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Triticum aestivum	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Hordeum vulgare	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Zea mays	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
Y	Physcomitrella patens	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP
	Actinidia chinensis	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP

Figure 2: Database from EnsemblPlants

The hisat2 function takes in input the index file generated by hisat2-build and the fastq sample to be analyzed. For each sample it gives in output a SAM file (Sequence Alignment Map). In order to obtain sorted BAM files we used the function samtools sort from SAMtools (`https://en.wikipedia.org/wiki/SAMtools`) to convert SAM files in BAM files.
Summarizing all the steps: from 36 raw fastq samples we obtained 36 SAM files and 36 sorted BAM files.

2.3 Production of absolute coverage matrix and relative coverage matrix

In order to obtain the coverage matrix all the BAM files were converted into BED files by bedtools (`https://bedtools.readthedocs.io/en/latest/`). The script used in R to generate the coverage matrix from the bed file is Sample_matrix_generator.R. This script generates a coverage matrix with the ORFs names in the rows and the position of the single nucleotide in the column.

In the picture below you can see the a matrix coverage generated by the Sample_matrix_generator.R script.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31
2	AT1G01010.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	AT1G01030.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	AT1G01030.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	AT1G01040.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	AT1G01040.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	AT1G01050.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	AT1G01050.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	AT1G01060.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	AT1G01060.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	AT1G01060.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	AT1G01060.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	AT1G01060.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	AT1G01060.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	AT1G01060.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	AT1G01060.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	AT1G01070.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	AT1G01070.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	AT1G01080.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	AT1G01080.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	AT1G01080.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	AT1G01090.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	AT1G01100.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	AT1G01100.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	AT1G01100.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	AT1G01100.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	AT1G01110.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

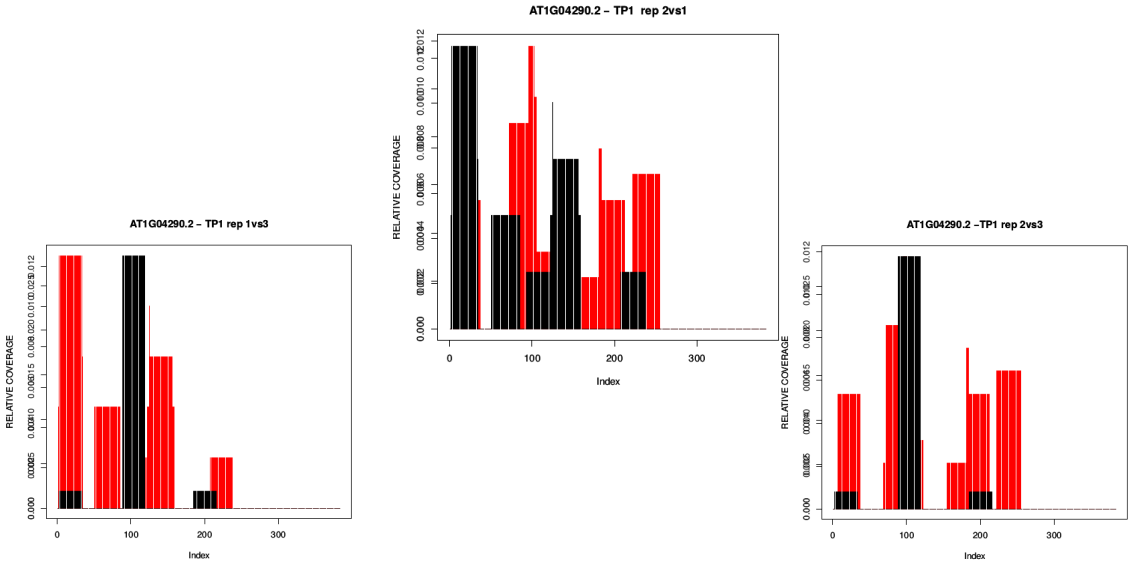
Figure 3: Picture of a matrix coverage for Time point 1 replicate 1. ORFs in the rows and nucleotide position in columns

3 Preliminary results

In order to better understand the differences between replicates for each Time point 1 we selected 100 genes with a high coverage, (not all 100 have a high coverage as we expected), did correlation analysis between replicates for each gene and the calculation of the total coverage for that gene divided by the transcript length in order to understand the total coverage regardless of the length of the transcript. In the subsections below some transcripts results are described, with different relative coverages and correlations results.

3.1 Plots of the relative coverages vs the nucleotide position and Pearson's Correlation analysis for AT1G04290.2

In this subsection some plots of two transcripts are shown. The plots refers to data obtained from Time point 1 samples. The script use to compute the correlation analysis in correlation.R.



The transcript taken in consideration is AT1G04290.2 As it possible to see from the coverage plots some differences appear from rep 1vs3, rep 2vs1 and rep 3vs3 for the sample TP1.(you should put the attention in the index position x-axis and not y). In order to have more information between replicates, we carried out Pearson's correlation analyses.

As you can see from coverage plots and scatterplots obtained from Pearson's correlation analyses there are differences among replicates: rep 1v3 has correlation coefficient r = 0.0085 , rep 1vs2 with r = 0.16 and rep 2vs3 with the highest correlation coefficient r = 0.42. Furthermore the results obtained calculating the total coverage divided by the transcript length is : rep 1 1.039062 , rep 2 2.427083 and for rep 3 1.375. Here below you can find the results of another transcript: AT1G07600.1, which has a better relative coverage and higher correlation coefficient between replicates compared to the transcript AT1G04290.2.

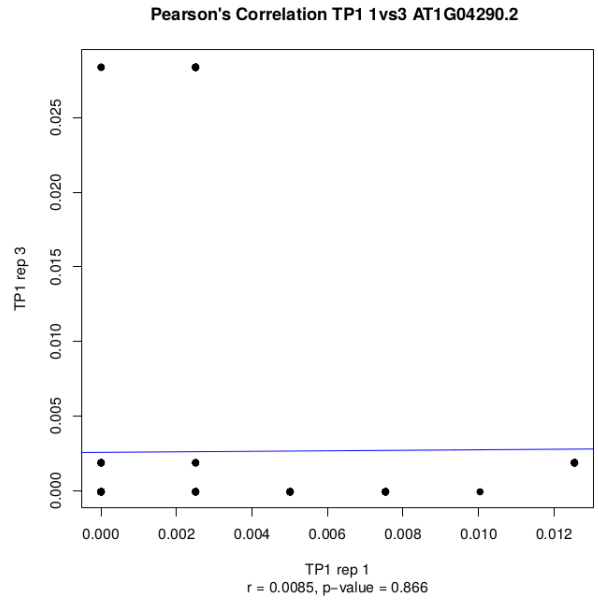
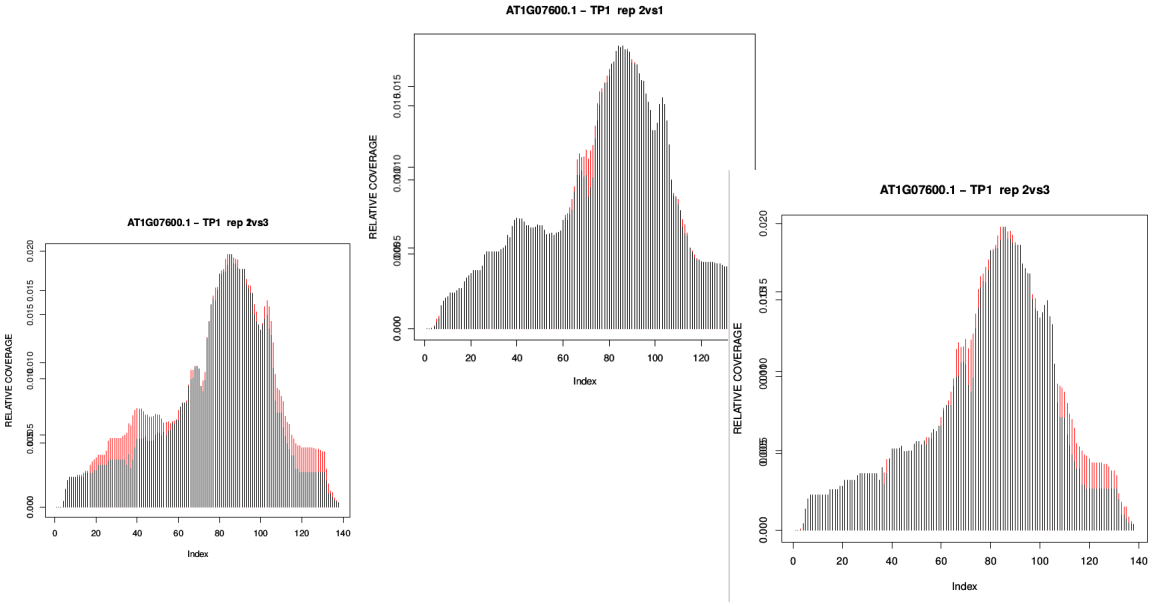


Figure 4: Scatterplot and Pearson's correlation analyses for AT1G04290.2 TP1 rep 1vs3

3.2 Plots of the relative coverages vs the nucleotide position and Pearson's Correlation analysis for AT1G07600.1



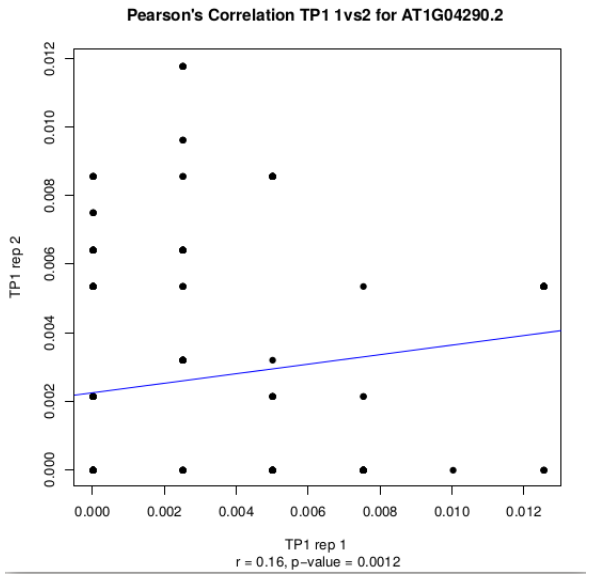


Figure 5: Scatterplot and Pearson's correlation analyses for AT1G04290.2 TP1 rep 1vs2

What can be seen is the similarity among relative coverage plots, differently from the coverage plots obtained from AT1G04290.2

As it is possible to see from the scatterplots of the ORF AT1G07600.1, the coefficient correlations are higher than those found in AT1G04290.2. For rep 1v3 r = 0.99, for rep 1vs2 r =0.98 and fro rep 2vs3 r = 0.98. In addition the results obtained from the calculation of the total coverage divided by the transcript length are : for rep 1 214, fore rep 2 270.3768 and for rep 3 332.4058

4 Comments

From the calculation carried out up to now it was possible to observe that normally higher coverage means higher correlation (not always) and, therefore, to select the most reproducible profiles we will use a ranking based on both coverage and correlation results. from the dirst table in section 5 (first 14 ORfs) and the table in section 6 (first 14 ORFs) it is possible to notice that most the transcripts with high correlation coefficients have also a high ratio of total_coverage / transcript_length, except for transcript such as

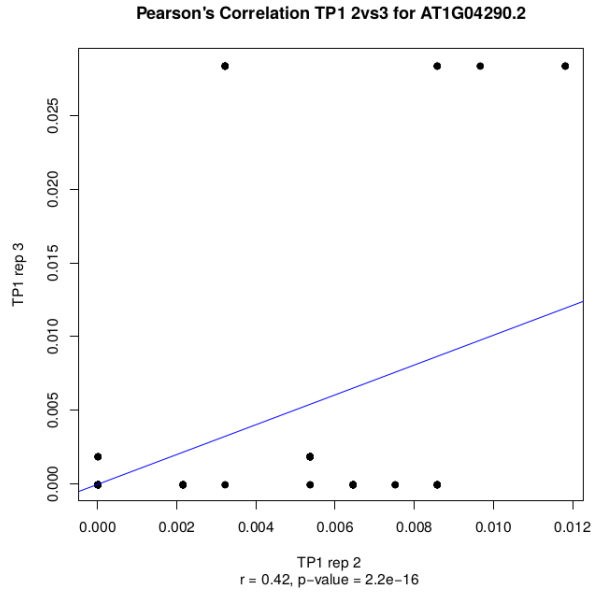


Figure 6: Scatterplot and Pearson's correlation analyses for AT1G04290.2 TP1 rep 2vs3

AT1G09800.1 and AT1G09800.2 correlations but not high relative coverages (look at the tables).

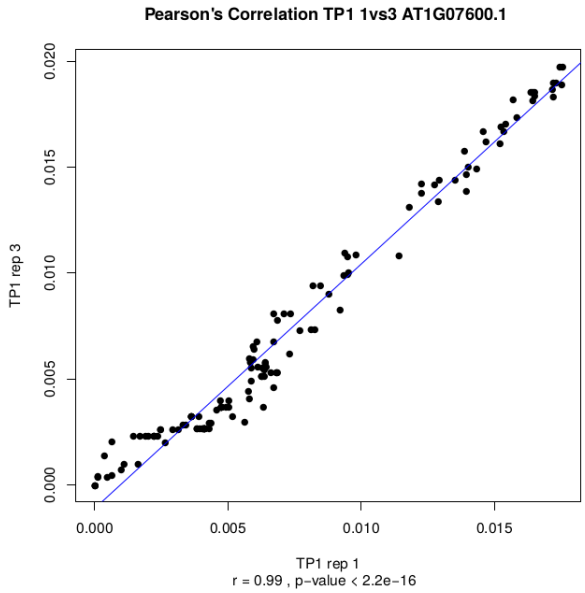


Figure 7: Scatterplot and Pearson's correlation analyses for AT1G07600.1 TP1 rep 1vs3

5 Summary Table for the 100 ORFs

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1vs2 Code	rep 1vs3	rep 2vs3
AT1G07600.1	0.9782345	0.9871618	0.9843594
AT1G09800.1	0.9833288	0.9801196	0.9795593
AT1G09800.2	0.9835631	0.9803233	0.9783562
AT1G53541.1	0.9950662	0.9984103	0.9900715
AT1G55673.1	0.9349811	0.8910161	0.9760975
AT3G13857.1	0.994706	0.9965271	0.999506
AT3G49920.2	0.9997458	0.9993935	0.9985879
AT4G09320.1	0.8796609	0.7369294	0.7123674
AT5G59613.1	0.9201291	0.8577776	0.6633689
ATCG00020.1	0.9755895	0.9817863	0.9592844
ATCG00070.1	0.9405155	0.9384468	0.8680467
ATCG00080.1	0.9351464	0.9511234	0.9016879
ATCG00120.1	0.9264573	0.894858	0.9226401
ATCG00140.1	0.8585744	0.9777155	0.8293562

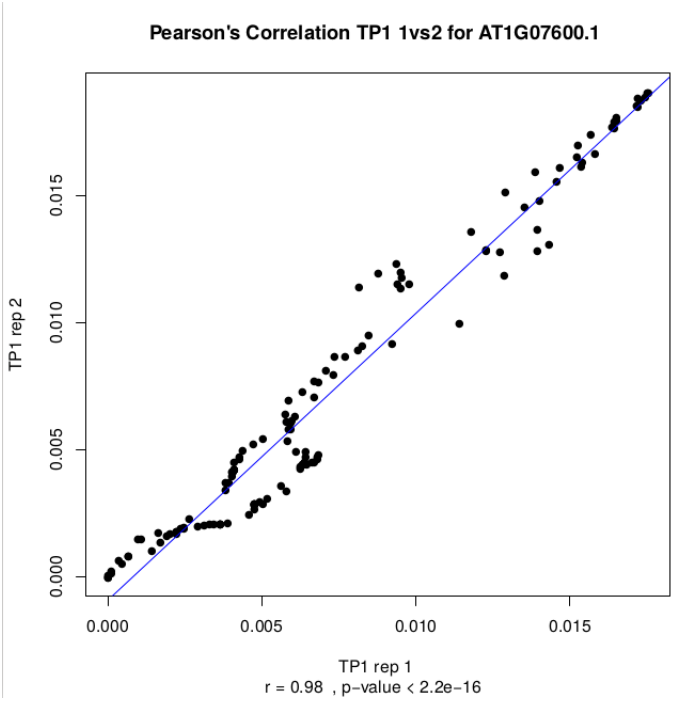


Figure 8: Scatterplot and Pearson's correlation analyses for AT1G07600.1 TP1 rep 1vs2

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1vs2 Code	rep 1vs3	rep 2vs3
ATCG00210.1	0.9459145	0.9837845	0.9759091
ATCG00220.1	0.9964742	0.9746892	0.9791441
ATCG00340.1	0.9451678	0.9615384	0.9277673
ATCG00350.1	0.9600231	0.9517733	0.9390848
ATCG00470.1	0.9637587	0.9658835	0.9258683
ATCG00490.1	0.9541093	0.9719953	0.9582485
ATCG00520.1	0.7252346	0.7054191	0.721893
ATCG00550.1	0.9698525	0.9872503	0.9798797
ATCG00560.1	0.9858116	0.958706	0.9899947
ATCG00570.1	0.9944103	0.9837862	0.9857812
ATCG00600.1	0.9492368	0.8891024	0.9179491
ATCG00650.1	0.8481483	0.860967	0.7282736
ATCG00660.1	0.8272129	0.7886692	0.7245616

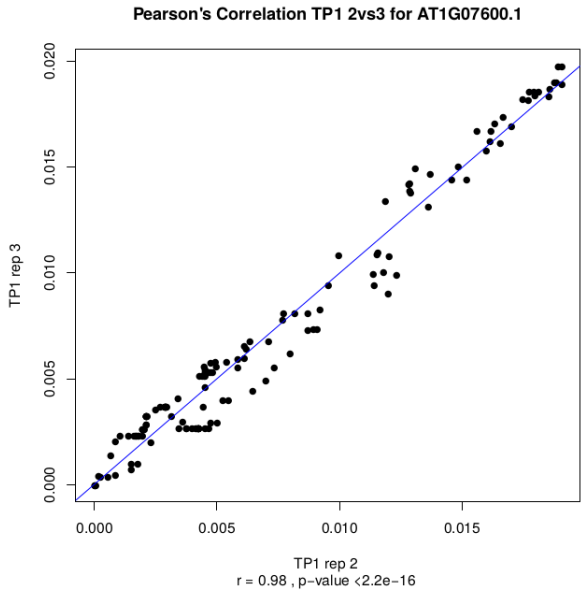


Figure 9: Scatterplot and Pearson's correlation analyses for AT1G07600.1 TP1 rep 2vs3

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
ATCG00680.1	0.9524347	0.9778689	0.9109693
ATCG00690.1	0.9899309	0.9960905	0.9815598
ATCG00720.1	0.9532643	0.9458089	0.9307036
ATCG00730.1	0.9724487	0.9703169	0.9741233
ATCG00760.1	0.6717877	0.2362805	0.2059832
ATCG00900.1	0.9704931	0.9672175	0.9535953
ATCG01240.1	0.9704931	0.9672175	0.9535953
ATMG01120.1	0.586791	0.9238498	0.7636374
ATMG01190.1	0.9378849	0.9628283	0.9086002

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1vs2 Code	rep 1vs3	rep 2vs3
AT2G41105.1	0.9382517	-0.1353951	-0.1574317
AT2G41100.2	0.7026138	0.6326103	0.5970612
AT2G42130.3	0.4860657	0.3296707	0.2778763
AT2G42130.5	0.5212842	0.3682946	0.3116968
AT2G42690.1	0.4155831	0.4270805	0.3245948
AT2G43100.1	0.6296907	0.4539188	0.4252924
AT2G43590.1	0.4046109	0.0573626	0.273754
AT2G44350.2	0.6403793	0.3698657	0.1980952
AT3G02870.2	0.3071559	0.3978649	0.2026351
AT3G22120.1	0.9085493	0.7344701	0.7546498
AT3G44310.3	0.9246538	0.8954354	0.8774787
AT3G45050.1	0.4323513	0.03918308	0.3427215
AT3G45050.2	0.4174323	0.02554599	0.3296404
AT3G45050.3	0.4381752	0.04628171	0.3468034
AT3G45050.4	0.4122539	0.02046674	0.3253962
AT3G54500.1	0.3469385	0.4160851	0.3870005

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1vs2 Code	rep 1vs3	rep 2vs3
AT3G55400.2	0.3688431	0.07983475	0.1568523
AT3G55850.2	0.9991016	0.9996162	0.9990187
AT3G55850.3	0.9991458	0.9996568	0.9991735
AT3G55850.4	0.9991216	0.9996354	0.999115
AT3G56310.2	0.4765446	0.5489358	0.5406405
AT3G57520.2	0.4830004	0.3771317	0.3669914
AT3G57520.3	0.4854871	0.4374985	0.4224039
AT3G57770.1	0.9562386	0.9930371	0.9569783
AT3G59410.3	0.9248161	0.8756108	0.9236214
AT4G22485.1	0.8178777	0.6065737	0.6673626
AT4G22520.1	0.4346354	-0.04123621	0.1146863
AT4G28755.1	0.9700321	0.9565663	0.9286071
AT4G29810.3	0.4471732	0.1539229	0.4788206
AT4G34050.2	0.7947452	0.6274912	0.8861327
AT5G03280.1	0.2779384	0.2246053	0.4022873
AT5G26000.2	0.937243	0.920053	0.9104639

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1vs2 Code	rep 1vs3	rep 2vs3
AT5G03240.2	0.8787724	0.7092396	0.6327312
AT5G13630.2	0.8864468	0.8488483	0.8641365
AT5G24930.1	0.6343677	0.5752556	0.405814
AT5G26000.1	0.9367364	0.921297	0.9124741
AT5G26742.3	0.8241427	0.7509631	0.8144447
AT5G36170.2	0.5430403	0.2911528	0.2339622
AT5G36170.4	0.5521581	0.29394	0.2425487
AT5G40400.2	0.9588395	0.9817328	0.9601588
AT5G42890.1	0.5517928	0.3853661	0.5105825
AT5G50950.1	0.271689	0.1544282	0.09659548
AT5G50950.4	0.3239874	0.1690028	0.1470133
AT1G05870.12	0.4440592	0.2684446	0.4635513
AT1G07320.2	0.8983535	0.8694005	0.82678
AT1G06680.2	0.9615652	0.9247685	0.9172959
AT1G03150.1	0.104773	0.2826901	-0.005809842
AT1G01100.1	0.9677989	0.7826819	0.8404301
AT1G04290.2	0.1637726	0.008584815	0.4288032
AT1G48860.2	0.5282851	0.5602539	0.4940424
AT1G53541.1	0.9950662	0.9984103	0.9900715
AT1G55140.2	0.5271701	0.7061018	0.5103548
AT1G55673.1	0.9349811	0.8910161	0.9760975
AT1G55673.1	0.9349811	0.8910161	0.9760975
AT1G58983.1	0.8537198	0.2971607	0.1128908
AT1G62515.1	0.3366875	0.3381441	0.2794562
AT1G79850.1	0.638102	0.638102	0.6567817
AT1G79850.1	0.8508563	0.638102	0.6567817
AT2G01021.1	0.990747	0.9905263	0.9974016
AT2G03140.8	0.09390592	-0.02597176	-0.02687407
AT2G04170.2	0.1375134	0.3008315	0.5707911
AT2G04270.3	0.08201327	-0.02109957	0.1363805
AT2G05380.2	0.9855249	0.9704587	0.9680205
AT2G07707.1	0.7913244	0.8018006	0.8373346
AT2G07741.1	0.8372214	0.8804926	0.8499462
AT2G11890.1	0.6176145	0.5634936	0.7292272
AT2G20410.1	0.9884911	0.9843935	0.9911937

6 Top 14 ORFs ratio total_cov/transcript_length for TP1

Pearson's Correlation Analysis from 100 ORFs selected from Time point 1 sample			
gene name	rep 1	rep 2	rep 3
AT1G07600.1	214	270.3768	332.4058
AT1G09800.1	1.822222	3.060969	4.139601
AT1G09800.2	2.15557	3.531312	5.056691
AT1G53541.1	115.7297	39.66667	192.7207
AT1G55673.1	38.47599	51.96868	64.29436
AT3G13857.1	358.3305	282.954	524.0795
AT3G49920.2	1.236942	0.8198335	0.34595
AT4G09320.1	44.93556	55.25111	70.83778
AT5G59613.1	4.240654	5.580607	8.247664
ATCG00020.1	4453.401	4666.079	6630.265
ATCG00070.1	319.5	336.0699	741.3065
ATCG00080.1	289.9009	328.9189	533.1982
ATCG00120.1	179.8425	227.9436	326.0814
ATCG00140.1	902	1001.5	1399.049

All the scripts used up to now are in the github repository `github.com/Ornela88/Arabidopsis_Riboseq`. All the scripts will be uploaded in this repository.