High Throughput Sequencing Analysis Report

CEL-Seq – *C. elegans*

Date: October-2018
PI: Prof. Ehud Cohen
Date of run: 16.08.2018
Initial Analysis: Dr. Efrat Kligun

Contents:

1. Experiment description

24 Samples of *C.elegans* were sequenced and analyzed in this project:

- N2: wild-type worms (untreated, treated) – 4 biological replicates (8 samples)
- Strain CL2006: Nematodes that express Aβ in their body wall muscles under the regulation of the *unc-54* promoter (untreated, treated) – 4 biological replicates (8 samples)
- Strain AM140: Animals that express polyQ35-YFP in their muscles under the regulation of the *unc-54* promoter (untreated, treated) – 4 biological replicates (8 samples)

The aim of the analysis is to characterize gene differentially respond to the proteotoxic challenges of Aβ and polyQ35-YFP.

2. Run info and statistics

Sequencer: Illumina HiSeq 2500
Recipe: PE: 15 (CEL-Seq barcode and UMI), 50 (data read) and barcode read
Raw data quality values: phred+33
Preparation: CEL-Seq preparation protocol

During CEL-Seq sample preparation protocol, each sample is marked with a primer containing a unique barcode. In this experiment each initial RNA sample was barcoded using a pool of 3 different CEL-Seq primers. Using several primers to create technical replicates for each sample reduces the possibility of technical variability caused by differences in barcode efficiency.

Table 1: Demultiplexing results

Sample ID	# Reads	% From total	Per sample # Reads	Per sample % from total
N2_EV_1_sample_0001	2,018,995	1.62		
N2_EV_1_sample_0002	3,128,502	2.51	8,203,848	6.58
N2_EV_1_sample_0003	3,056,351	2.45		
N2_tor2_1_sample_0004	1,203,305	0.97		
N2_tor2_1_sample_0005	2,414,827	1.94	6,258,466	5.02
N2_tor2_1_sample_0006	2,640,334	2.12		
N2_EV_2_sample_0007	1,135,075	0.91		
N2_EV_2_sample_0008	1,266,563	1.02	3,594,907	2.88
N2_EV_2_sample_0009	1,193,269	0.96		
N2_tor2_2_sample_0010	1,532,022	1.23		
N2_tor2_2_sample_0011	1,473,945	1.18	5,638,279	4.52
N2_tor2_2_sample_0012	2,632,312	2.11		
N2_EV_3_sample_0013	1,937,762	1.55		
N2_EV_3_sample_0014	3,288,458	2.64	8,774,647	7.04
N2_EV_3_sample_0015	3,548,427	2.85		
N2_tor2_3_sample_0016	1,695,266	1.36		
N2_tor2_3_sample_0017	2,674,595	2.15	6,738,313	5.41
N2_tor2_3_sample_0018	2,368,452	1.90		
N2_EV_4_sample_0019	1,235,860	0.99		
N2_EV_4_sample_0020	2,147,857	1.72	5,121,678	4.11
N2_EV_4_sample_0021	1,737,961	1.39		
N2_tor2_4_sample_0022	2,079,218	1.67		
N2_tor2_4_sample_0023	3,361,581	2.70	8,565,521	6.87
N2_tor2_4_sample_0024	3,124,722	2.51		

AM140_EV_1_sample_0025	1,777,050	1.43		
AM140_EV_1_sample_0026	1,217,295	0.98	4,605,514	3.69
AM140_EV_1_sample_0027	1,611,169	1.29		
AM140_tor2_1_sample_0028	1,751,251	1.40		
AM140_tor2_1_sample_0029	1,723,401	1.38	5,353,686	4.29
AM140_tor2_1_sample_0030	1,879,034	1.51		
AM140_EV_2_sample_0031	589,662	0.47		
AM140_EV_2_sample_0032	1,363,903	1.09	3,171,274	2.54
AM140_EV_2_sample_0033	1,217,709	0.98		
AM140_tor2_2_sample_0034	1,487,188	1.19		
AM140_tor2_2_sample_0035	1,778,703	1.43	4,673,863	3.75
AM140_tor2_2_sample_0036	1,407,972	1.13		
AM140_EV_3_sample_0037	4,941	0.00		
AM140_EV_3_sample_0038	74,634	0.06	89,438	0.07
AM140_EV_3_sample_0039	9,863	0.01		
AM140_tor2_3_sample_0040	1,893,167	1.52		
AM140_tor2_3_sample_0041	1,949,629	1.56	6,099,414	4.89
AM140_tor2_3_sample_0042	2,256,618	1.81		
AM140_EV_4_sample_0043	1,415,194	1.14		
AM140_EV_4_sample_0044	1,616,907	1.30	4,730,150	3.79
AM140_EV_4_sample_0045	1,698,049	1.36		
AM140_tor2_4_sample_0046	1,160,765	0.93		
AM140_tor2_4_sample_0047	1,227,623	0.98	3,762,558	3.02
AM140_tor2_4_sample_0048	1,374,170	1.10		
CL2006_EV_1_sample_0049	1,403,443	1.13		
CL2006_EV_1_sample_0050	1,319,049	1.06	4,112,521	3.30
CL2006_EV_1_sample_0051	1,390,029	1.12		
CL2006_tor2_1_sample_0052	1,699,216	1.36		
CL2006_tor2_1_sample_0053	1,283,724	1.03	4,312,911	3.46
CL2006_tor2_1_sample_0054	1,329,971	1.07		
CL2006_EV_2_sample_0055	1,200,661	0.96		
CL2006_EV_2_sample_0056	304,091	0.24	2,718,462	2.18
CL2006_EV_2_sample_0057	1,213,710	0.97		
CL2006_tor2_2_sample_0058	639,072	0.51		
CL2006_tor2_2_sample_0059	493,368	0.40	1,775,664	1.42
CL2006_tor2_2_sample_0060	643,224	0.52		
CL2006_EV_3_sample_0061	2,469,634	1.98		
CL2006_EV_3_sample_0062	1,541,983	1.24	5,833,769	4.68
CL2006_EV_3_sample_0063	1,822,152	1.46		
CL2006_tor2_3_sample_0064	1,983,264	1.59		
CL2006_tor2_3_sample_0065	2,422,687	1.94	7,707,577	6.18
CL2006_tor2_3_sample_0066	3,301,626	2.65		

CL2006_EV_4_sample_0067	1,733,526	1.39		
CL2006_EV_4_sample_0068	869,485	0.70	4,203,706	3.37
CL2006_EV_4_sample_0069	1,600,695	1.28		
CL2006_tor2_4_sample_0070	1,538,648	1.23		
CL2006_tor2_4_sample_0071	2,315,581	1.86	5,164,264	4.14
CL2006_tor2_4_sample_0072	1,310,035	1.05		
Undetermined Indices			3,358,046	2.69

Quality control and trimming:

Adapter and quality trimming were performed on the data in order to optimize mapping accuracy.

The minimal quality threshold was set to 20 (phred scale) and the minimal length required for each read after trimming was set to 25. Two sequences were set for removal:

1. "Small RNA 5' adapter" – GATCGTCGGACT.
2. Poly A – during sample preparation, a poly T sequence is used in order to capture the 3' end of the transcripts. For short fragments, parts of the read might contain residues of the poly-A tail.

Table 2: Trimming statistics – Quality and adapter trimming:

Sample	Total reads	Reads containing adapter (%)	Trimmed bases (%)	Reads removed (%)	Reads removed (#)
AM140_EV_1	4,605,514	1.5	0.3	0.1	5,780
AM140_EV_2	3,171,274	1.5	0.3	0.1	4,197
AM140_EV_3	89,438	1.6	0.4	0.2	205
AM140_EV_4	4,730,150	1.5	0.3	0.1	6,162
AM140_tor2_1	5,353,686	1.5	0.3	0.1	7,065
AM140_tor2_2	4,673,863	1.5	0.3	0.1	6,335
AM140_tor2_3	6,099,414	1.5	0.3	0.1	8,414
AM140_tor2_4	3,762,558	1.5	0.3	0.1	4,851
CL2006_EV_1	4,112,521	1.4	0.3	0.1	5,147
CL2006_EV_2	2,718,462	1.5	0.3	0.1	3,613
CL2006_EV_3	5,833,769	1.5	0.3	0.1	7,454
CL2006_EV_4	4,203,706	1.5	0.3	0.1	5,114
CL2006_tor2_1	4,312,911	1.5	0.3	0.1	5,645
CL2006_tor2_2	1,775,664	1.5	0.3	0.1	2,293
CL2006_tor2_3	7,707,577	1.5	0.3	0.1	10,332
CL2006_tor2_4	5,164,264	1.5	0.3	0.1	7,225
N2_EV_1	8,203,848	1.5	0.3	0.1	10,407
N2_EV_2	3,594,907	1.5	0.3	0.1	4,519
N2_EV_3	8,774,647	1.5	0.3	0.1	11,186
N2_EV_4	5,121,678	1.5	0.3	0.1	6,700

N2_tor2_1	6,258,466	1.5	0.3	0.1	9,368
N2_tor2_2	5,638,279	1.5	0.3	0.1	7,583
N2_tor2_3	6,738,313	1.5	0.3	0.1	8,593
N2_tor2_4	8,565,521	1.5	0.3	0.1	11,014

Table 2: Trimming statistics – poly-A trimming:

Sample	Total reads (polyA trimming)	Reads containing polyA (%)	Trimmed bases (%)	Reads removed (%)	Reads removed (#)	Reads removed total (#)	Reads removed total (%)
AM140_EV_1	4,599,734	6.7	1.9	0.5	24,226	30,006	0.65
AM140_EV_2	3,167,077	6.8	1.9	0.5	15,933	20,130	0.63
AM140_EV_3	89,233	6.8	1.9	0.5	429	634	0.71
AM140_EV_4	4,723,988	6.7	1.9	0.5	23,535	29,697	0.63
AM140_tor2_1	5,346,621	6.8	1.9	0.5	27,446	34,511	0.64
AM140_tor2_2	4,667,528	7	1.9	0.5	23,903	30,238	0.65
AM140_tor2_3	6,091,000	7.1	1.9	0.5	30,763	39,177	0.64
AM140_tor2_4	3,757,707	6.8	1.9	0.5	18,428	23,279	0.62
CL2006_EV_1	4,107,374	6.8	1.9	0.5	20,210	25,357	0.62
CL2006_EV_2	2,714,849	6.7	1.9	0.5	14,682	18,295	0.67
CL2006_EV_3	5,826,315	6.8	1.9	0.5	29,044	36,498	0.63
CL2006_EV_4	4,198,592	6.8	1.9	0.5	21,270	26,384	0.63
CL2006_tor2_1	4,307,266	6.7	1.9	0.5	21,795	27,440	0.64
CL2006_tor2_2	1,773,371	6.9	1.9	0.5	8,845	11,138	0.63
CL2006_tor2_3	7,697,245	6.6	1.8	0.4	34,488	44,820	0.58
CL2006_tor2_4	5,157,039	6.7	1.9	0.5	25,124	32,349	0.63
N2_EV_1	8,193,441	6.8	1.9	0.5	38,733	49,140	0.60
N2_EV_2	3,590,388	6.9	2	0.5	18,825	23,344	0.65
N2_EV_3	8,763,461	6.9	1.9	0.5	41,782	52,968	0.60
N2_EV_4	5,114,978	6.9	1.9	0.5	25,966	32,666	0.64
N2_tor2_1	6,249,098	6.8	1.9	0.5	30,562	39,930	0.64
N2_tor2_2	5,630,696	6.9	1.9	0.5	28,538	36,121	0.64
N2_tor2_3	6,729,720	6.9	1.9	0.5	32,919	41,512	0.62
N2_tor2_4	8,554,507	6.8	1.9	0.5	40,198	51,212	0.60

*The two last columns summarize the amount and percentage of reads removed during the trimming procedure (due to low quality or presence of adapter or poly-A sequences).

The quality of the sequenced data (all sequenced bases in all reads) before and after trimming, as well as read length distributions after trimming, were evaluated using FASTQC. The quality scores are presented as phred values ($10\log_{10}P(\textit{base call is wrong})$), i.e. values higher than 30 indicate a probability of less than 10^{-3} of an incorrect base call.

Figure 1: Read length distribution after trimming

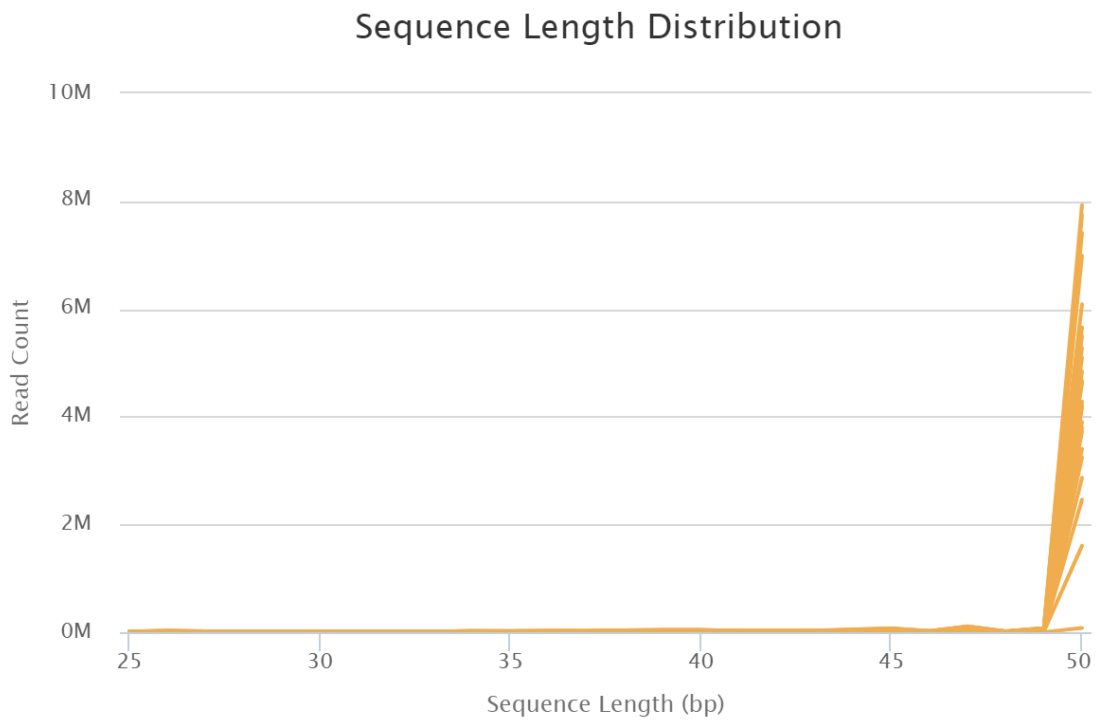


Figure 2: Per-base quality scores before trimming

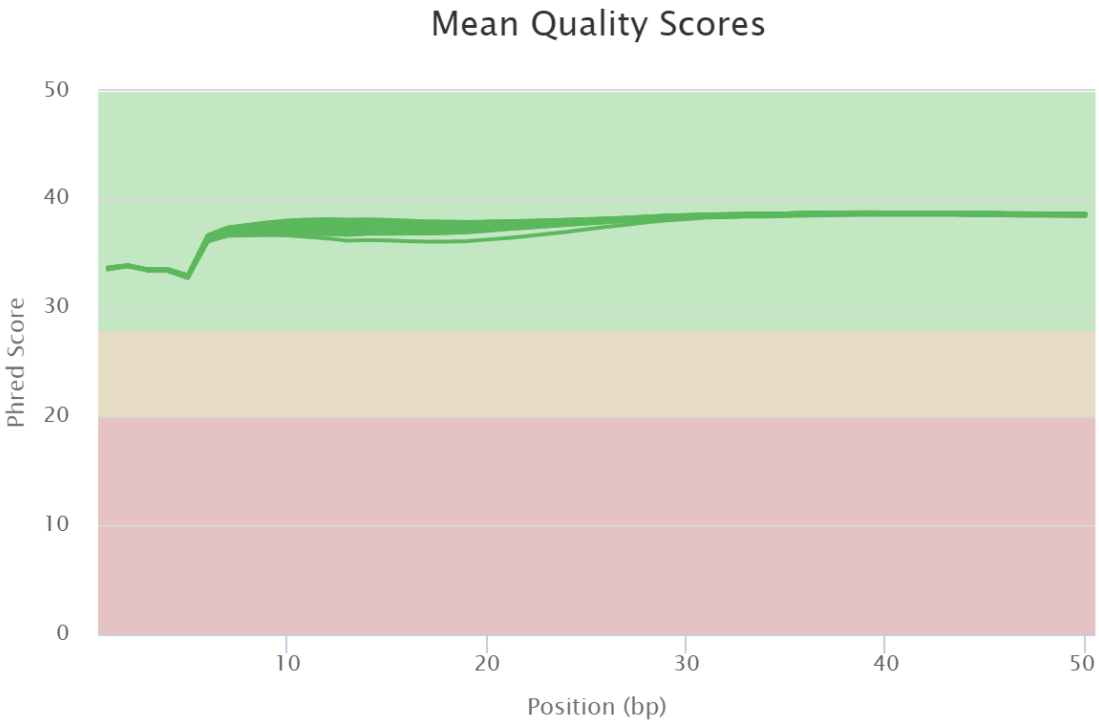
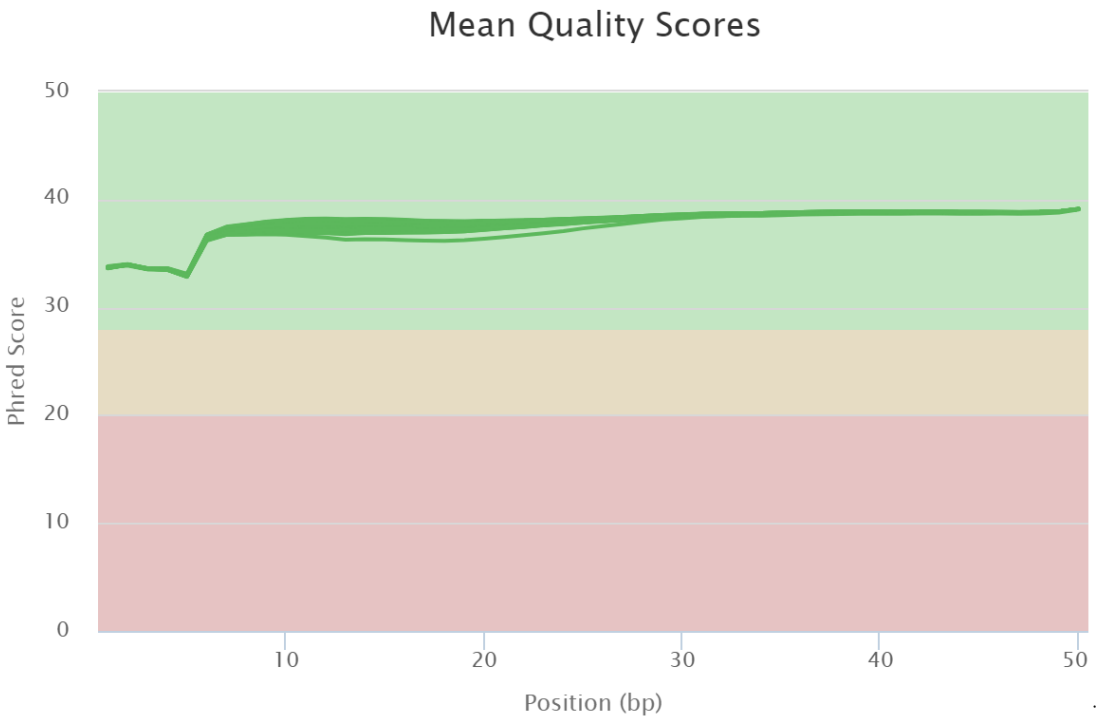


Figure 3: Per-base quality scores after trimming



Conclusions from the statistical data

The per-base quality scores indicate that the sequencing was of high quality. Only a small percentage of reads were discarded due to trimming and the peak of the length distribution is at the original length of 51.

3. Analysis Pipeline

Software and applications used:

Library quality control: FASTQC version 0.11.5
Quality and adapter trimming: trim_galore (uses cutadapt version 1.10)
Mapping: Tophat2 version 2.1.0 (uses Bowtie2 version 2.2.6)
Gene counting: HTseq-count version 0.6.1
Normalization and differential expression analysis: DESeq2 R package version 1.18.1

Input files

The reference genome was downloaded from Ensembl (iGenomes site). The link is:
http://support.illumina.com/sequencing/sequencing_software/igenome.ilmn
The version used is Ensembl: WBcel235
Counting was performed using an annotation file downloaded from Ensembl at the following link: ftp://ftp.ensembl.org/pub/release-94/gtf/caenorhabditis_elegans/

A. Mapping

Alignment details (Tophat2):

Tophat2 is a program designed specifically for alignment of RNA-Seq data to a chosen reference genome. The software uses the high-throughput, short read aligner 'Bowtie2' and analyzes the mapping results to identify splice junctions between exons.
Tophat2 divides the reads into shorter segments and generates a database of possible splice junctions (built using both the input data of the current project as well as a given annotation file specifying the coordinates of all known exons of the organism). Input reads are first mapped to the generated database of possible splice junctions (transcriptome) and the unmapped reads from this procedure are than mapped to the entire genome.
TopHat2 parameters were specified for identification of alignments with a maximum of 3 mismatches from the reference sequence per read, and a maximum of 3 mismatches for a segment of a read. The maximum and minimum intron sizes were set to 25 and 101000, respectively (according to the literature).

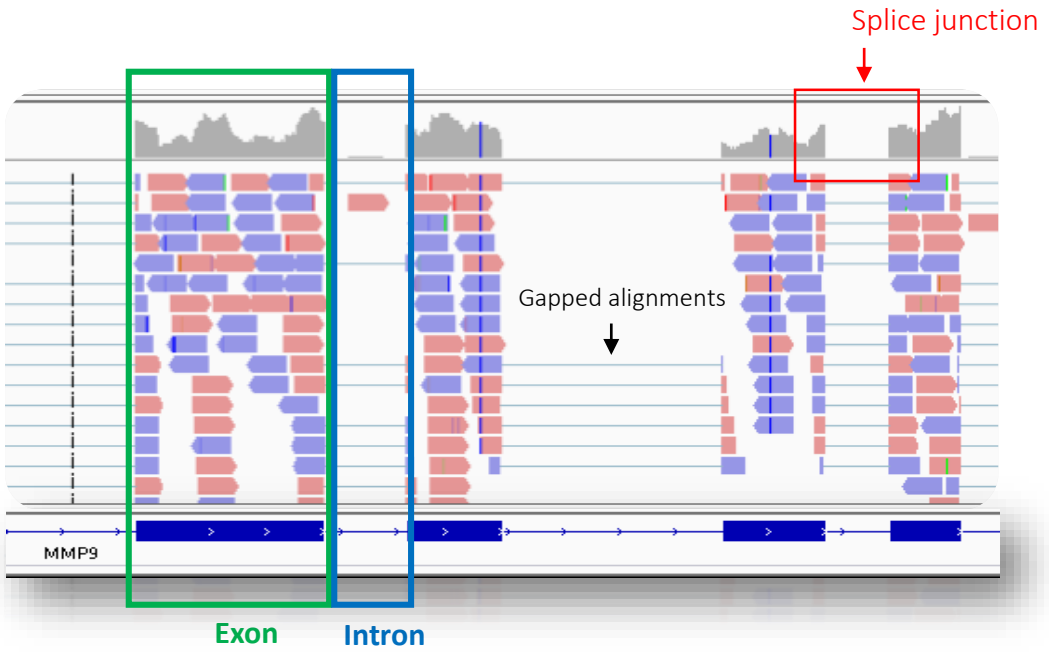


Figure 4: Illustration of RNA-Seq read alignments using 'Tophat' – A graphic representation of the mapping patterns of RNA-Seq reads to a reference genome. The figure above reveals the distribution of reads that are mapped to exons (outlined in green), a splice junction (outlined in red), and introns (outlined in blue). The figure allows the user to see that the majority of reads map to exons, slightly fewer reads map in a gapped manner to the pictured splice junction, and—consistent with the nature of RNA library preparation preference for mature mRNAs—almost no reads map to introns.

Table 3: Mapping statistics

Sample	Total reads after trimming (#)	Uniquely mapped (#)	Uniquely mapped (%)	Unmapped (%)	Multi mapped (%)	Gapped uniquely mapped (%)
AM140_EV_1	4,575,508	4,439,609	97.03	2.14	0.83	1.35
AM140_EV_2	3,151,144	3,059,103	97.08	2.04	0.88	1.39
AM140_EV_3	88,804	84,372	95.01	4.28	0.71	1.26
AM140_EV_4	4,700,453	4,555,391	96.91	1.97	1.11	1.47
AM140_tor2_1	5,319,175	5,161,762	97.04	2.05	0.91	1.31
AM140_tor2_2	4,643,625	4,501,574	96.94	2.17	0.89	1.24
AM140_tor2_3	6,060,237	5,880,399	97.03	2.11	0.85	1.22
AM140_tor2_4	3,739,279	3,627,258	97.00	2.03	0.96	1.38
CL2006_EV_1	4,087,164	3,972,277	97.19	2.06	0.75	1.38
CL2006_EV_2	2,700,167	2,615,348	96.86	2.02	1.12	1.38
CL2006_EV_3	5,797,271	5,636,846	97.23	2.02	0.74	1.28
CL2006_EV_4	4,177,322	4,054,124	97.05	2.05	0.90	1.41
CL2006_tor2_1	4,285,471	4,165,151	97.19	2.04	0.77	1.35
CL2006_tor2_2	1,764,526	1,711,222	96.98	2.20	0.82	1.33
CL2006_tor2_3	7,662,757	7,457,550	97.32	1.90	0.78	1.37
CL2006_tor2_4	5,131,915	4,989,559	97.23	1.98	0.80	1.40
N2_EV_1	8,154,708	7,937,174	97.33	1.99	0.68	1.21
N2_EV_2	3,571,563	3,467,340	97.08	2.17	0.75	1.21
N2_EV_3	8,721,679	8,493,049	97.38	1.90	0.72	1.17
N2_EV_4	5,089,012	4,944,881	97.17	1.99	0.84	1.24
N2_tor2_1	6,218,536	6,048,517	97.27	2.01	0.73	1.23
N2_tor2_2	5,602,158	5,443,475	97.17	2.01	0.83	1.26
N2_tor2_3	6,696,801	6,509,284	97.20	2.04	0.76	1.16
N2_tor2_4	8,514,309	8,274,719	97.19	1.91	0.9	1.26

- **Uniquely mapped** – Reads aligned with high confidence to a single genomic location with up to 3 mismatches. Only the uniquely mapped reads are used for further analysis.
- **Unmapped** – Reads for which no alignment was found to the reference genome.

- **Multi-mapped** – Reads mapped to more than one possible location in the genome. These reads are not used in the analysis.
- **Gapped uniquely mapping** – Reads aligned uniquely to a splice junction (Fig. 4)

B. Raw gene counts:

Uniquely mapped reads are assigned to genes based on the gene annotation file. The 'HTSeq-count' package was used to obtain gene counts with the "union" mode option and the parameter '--stranded' was set to 'yes' according to the sample preparation kit specifications. For more information, please see the HTSeq-count documentation: http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Table 4: Read assignments according to gene annotations

Sample	Uniquely mapped reads (#)	Counted reads (#)	Counted reads (%)	Ambiguous reads (%)	No feature reads (%)
AM140_EV_1	4,439,609	4,284,583	96.51	0.65	2.84
AM140_EV_2	3,059,103	2,941,300	96.15	0.69	3.17
AM140_EV_3	84,372	81,580	96.69	0.72	2.59
AM140_EV_4	4,555,391	4,387,607	96.32	0.66	3.02
AM140_tor2_1	5,161,762	4,985,562	96.59	0.64	2.77
AM140_tor2_2	4,501,574	4,352,043	96.68	0.64	2.68
AM140_tor2_3	5,880,399	5,697,375	96.89	0.65	2.47
AM140_tor2_4	3,627,258	3,508,056	96.71	0.63	2.66
CL2006_EV_1	3,972,277	3,837,561	96.61	0.63	2.76
CL2006_EV_2	2,615,348	2,521,828	96.42	0.62	2.96
CL2006_EV_3	5,636,846	5,470,270	97.04	0.62	2.33
CL2006_EV_4	4,054,124	3,923,233	96.77	0.63	2.59
CL2006_tor2_1	4,165,151	4,025,094	96.64	0.63	2.74
CL2006_tor2_2	1,711,222	1,658,937	96.94	0.58	2.47
CL2006_tor2_3	7,457,550	7,241,339	97.1	0.64	2.25
CL2006_tor2_4	4,989,559	4,829,632	96.79	0.63	2.57
N2_EV_1	7,937,174	7,709,717	97.13	0.63	2.23
N2_EV_2	3,467,340	3,365,567	97.06	0.67	2.27
N2_EV_3	8,493,049	8,262,397	97.28	0.62	2.09
N2_EV_4	4,944,881	4,783,138	96.73	0.66	2.61
N2_tor2_1	6,048,517	5,857,415	96.84	0.67	2.49
N2_tor2_2	5,443,475	5,273,466	96.88	0.63	2.49
N2_tor2_3	6,509,284	6,319,702	97.09	0.65	2.26
N2_tor2_4	8,274,719	8,027,620	97.01	0.66	2.32

- **Counted reads** – uniquely mapped reads assigned to annotated genes.
- **No feature –** reads that could not be assigned to any annotated gene.
- **Ambiguous –** reads that can be assigned to more than one annotated gene, and are therefore not counted to any gene.

C. Gene counts normalization:

Normalization of raw counts is an important step which brings all samples to a common scale and enables comparing expression levels between different biological conditions.

The need for normalizations may arise due to several possible reasons. First being technical deviations, such as different library sizes (total number of reads) when sequencing several samples together on a lane. Another possible cause is the biological condition, for example a certain treatment may cause a significant up-regulation of a small subset of genes, and thus affecting the overall reads distribution among genes.

Normalization is conducted using 'DESeq2' R package (which is also used for the differential expression analysis), during the process a normalization factor (size factor) is calculated for each sample and the raw counts are divided by this factor.

For details see the DESEq2 documentation:

http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

D. Replicates evaluation

In order to analyze the similarity between the suggested replicates and explore the relations between the samples, heatmaps and PCA plots were generated. Principal component Analysis (PCA) generates a plot that span(s) the samples in 2D plane by their first two principal components. This plot is useful for visualizing the overall effect of experimental covariates and batch effects.

Fig. 5: Heatmap of Euclidean distances

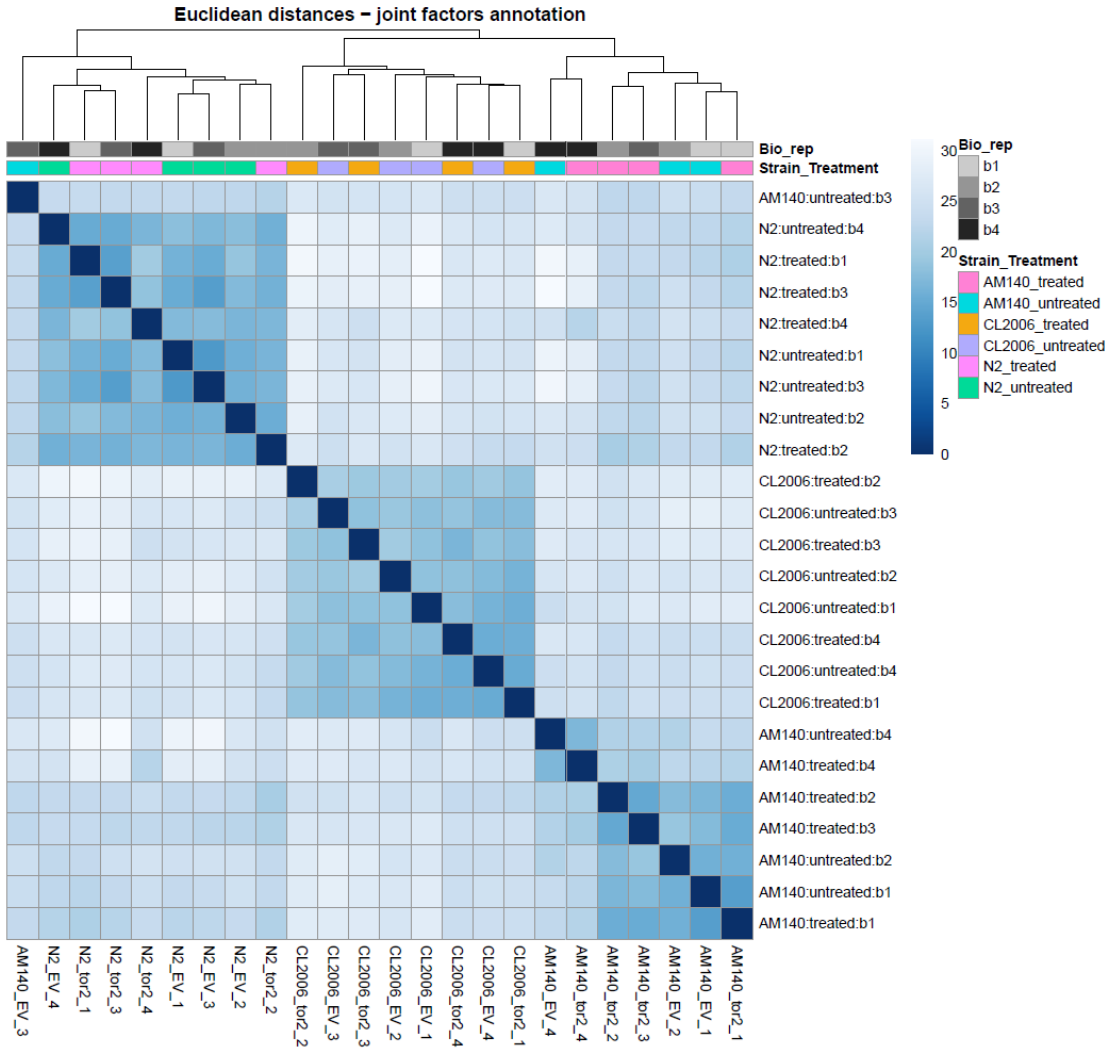
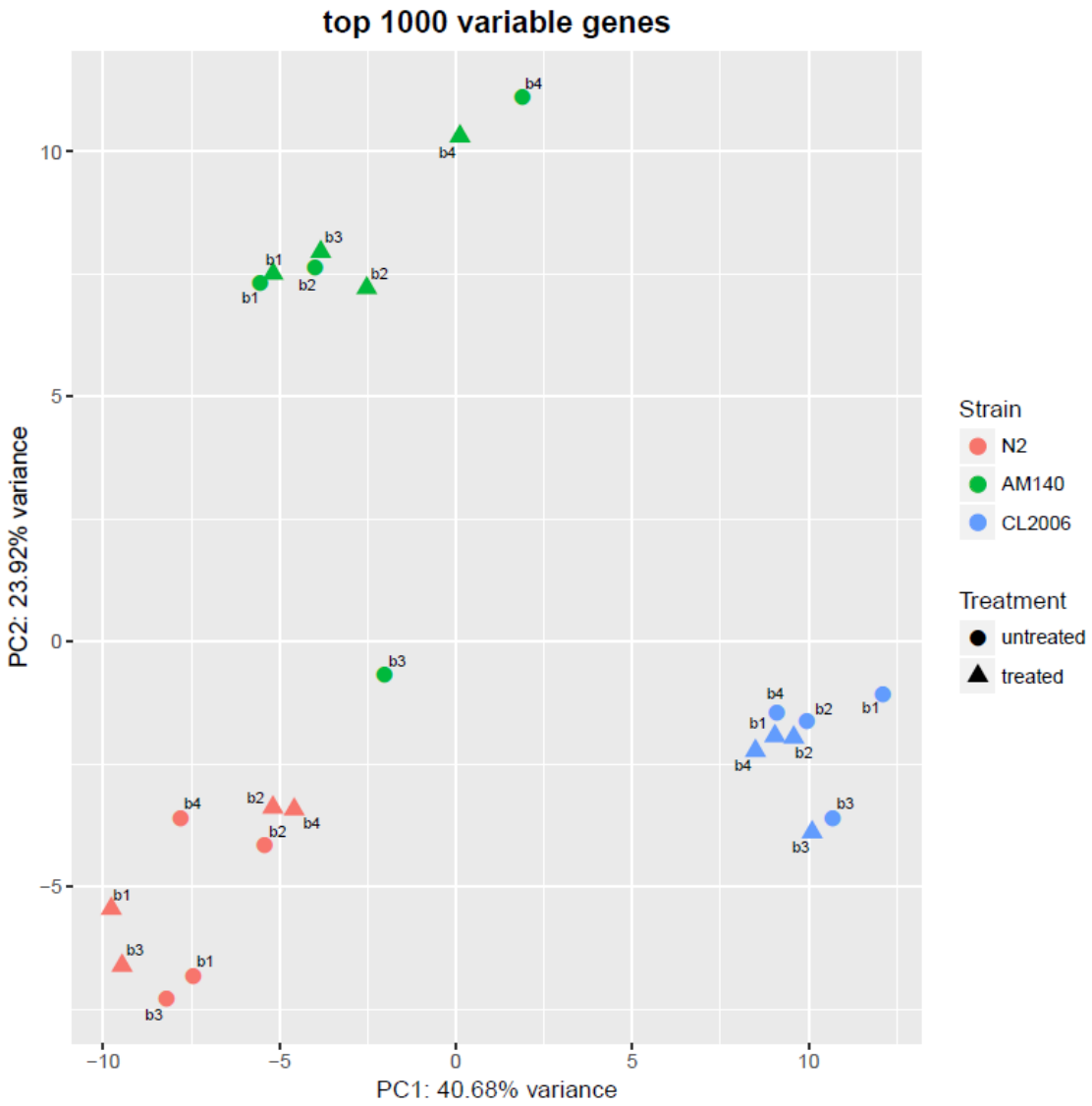


Fig. 6: PCA plot



As can be seen from the plots above, the third replicate of AM140 (AM140_EV_3) is an outlier sample – thus it was removed from the analysis.

Fig. 7: Heatmap of Euclidean distances (without sample: AM140_EV_3)

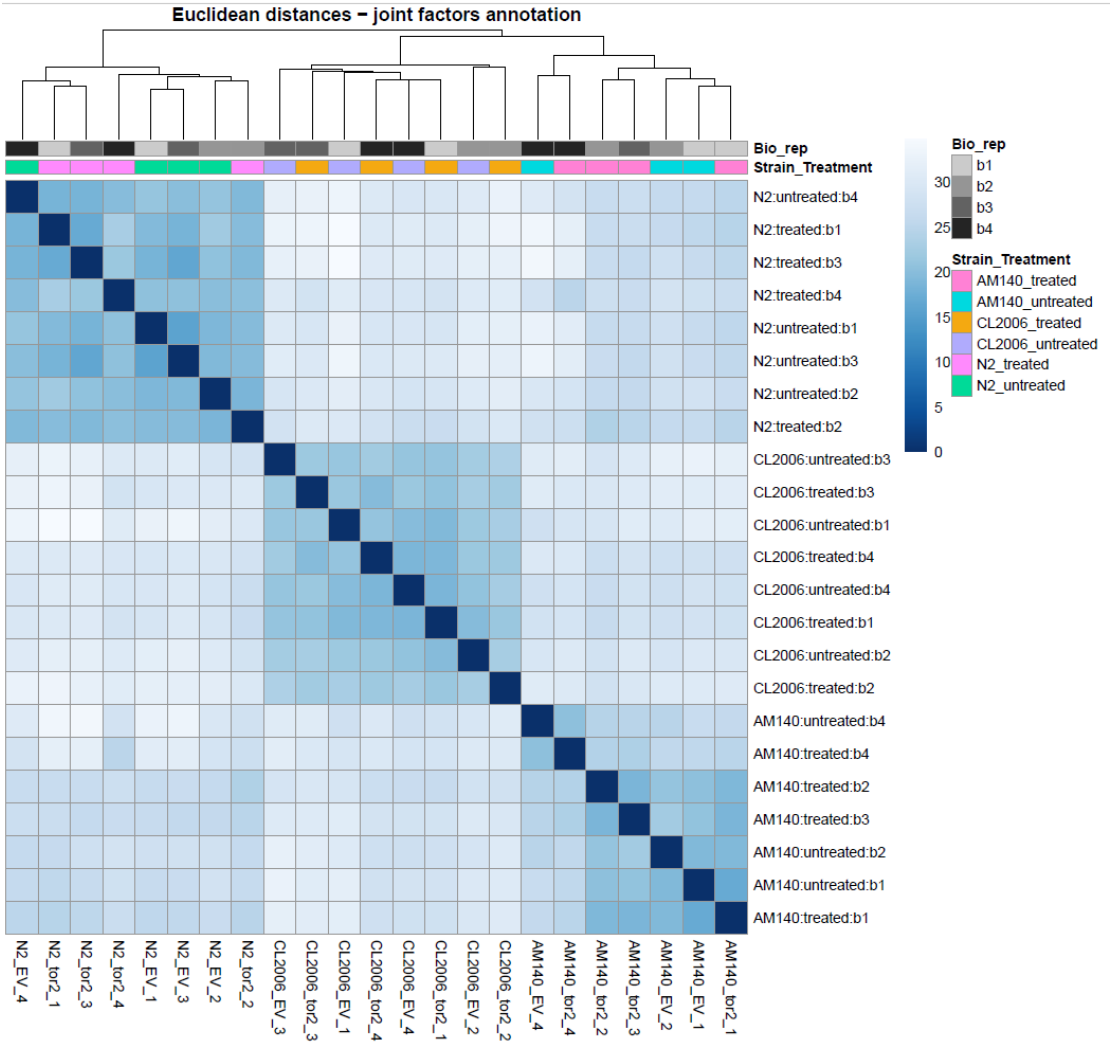
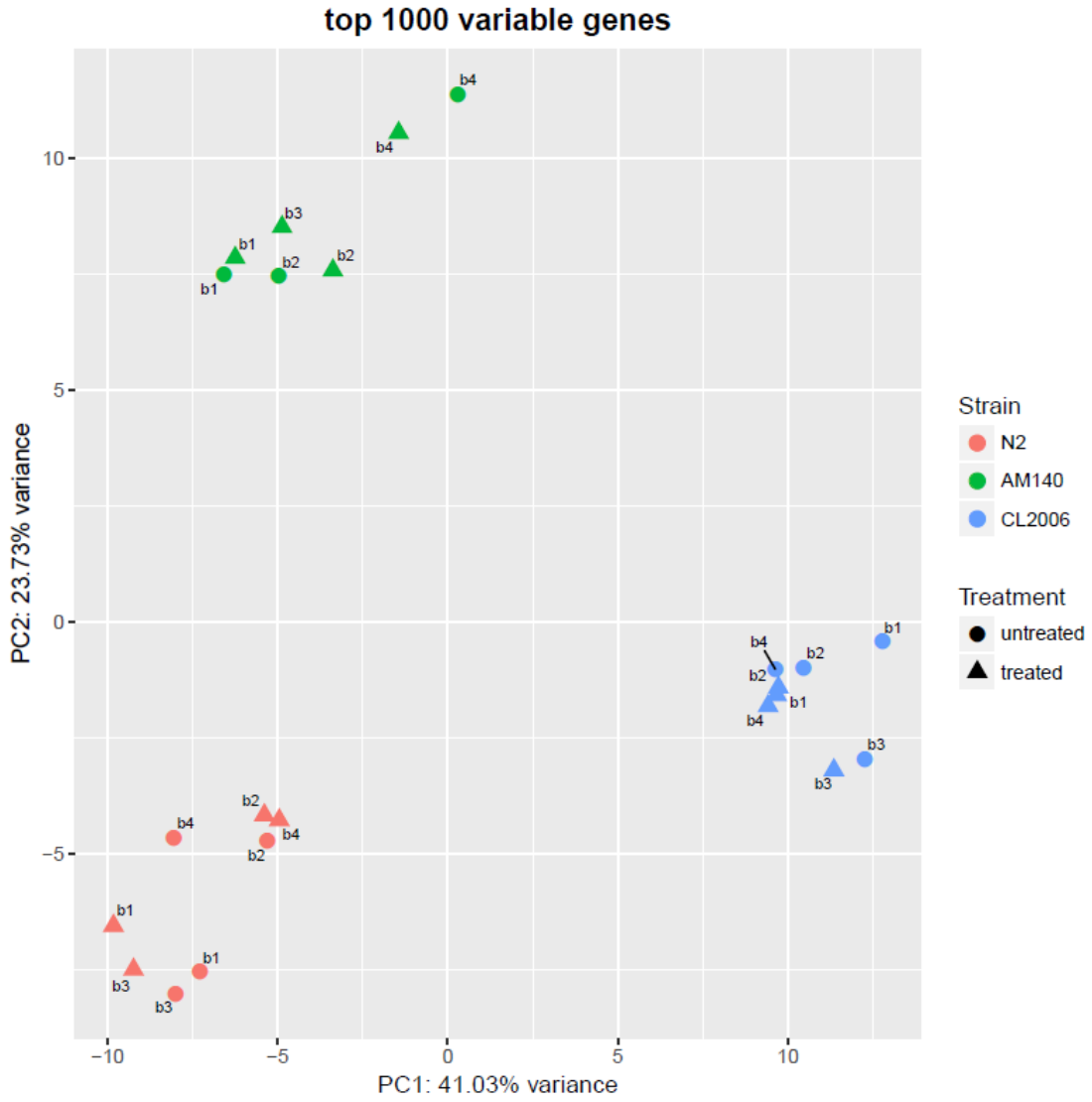


Fig. 8: PCA plot – (without sample: AM140_EV_3)



The plots above suggest generally clustering of the three strains and no apparent clustering of the treated and untreated groups.

E. Differential gene expression:

The differential expression analysis was conducted using 'DESeq2' R package. DESeq2 generates a matrix of normalized counts and performs statistical tests to determine whether genes are differentially expressed between pairs of samples or combinations of factors.

In this analysis we have used the Wald hypothesis test provided by DESeq2.

Wald is used to compare between two specific conditions (such as treatment vs control, this is pairwise comparison, while different response comparison is ratio of ratio comparison, for example, $\frac{\frac{AM140\_treated}{AM140\_untreated}}{\frac{N2\_treated}{N2\_untreated}}$).

Table 5: Summary of differential expression analysis

Comparison	Total # Genes	All Zero	Low Counts	Tested	Significant	
					UP	Down
untreated_CL2006_vs_N2	46,778	28,724	8778	9276	977	964
untreated_AM140_vs_N2			9954	8100	522	618
untreated_CL2006_vs_AM140			8,715	9,339	359	346
N2_treated_vs_untreated			9,254	8,800	0	0
AM140_treated_vs_untreated			8,862	9,192	0	2
CL2006_treated_vs_untreated			8,709	9,345	0	1
AM140_vs_N2_response			8,724	9,330	0	
CL2006_vs_N2_response			8,551	9,503	1	
CL2006_vs_AM140_response			8,361	9,693	0	

- **Total number of gens:** All the genes present in the used annotation file
- **All zero:** Genes that were not detected in any of the samples.
- **Low counts:** Genes which did not pass a minimal detection threshold (As explained in the 'final analysis results' section).
- **Tested:** This is the number of genes that entered the statistical analysis and received a final adjusted p-value.
- **Significantly Differentially Expressed Gene:** Significantly differentially expressed genes, according to a threshold of 0.05 on the adjusted p-value. (separated to up and down regulated according to the sign on the log2 fold change)

4. Analysis Result files:

1. Images of the heatmaps, PCA plots.
2. **Normalized counts** tables containing the normalized counts of each sample.
3. **Gene expression analysis results** tables that contain the following columns:

   Gene ID, gene name, gene biotype and chromosomal position for each gene
   Base mean - mean of the normalized counts across all samples
   For each comparison:
   Log2 fold change - log2 of the fold change between the pair of samples
   Signed fold change – Fold change – up regulated will have (+) sign and down-regulate (-) sign.
   p-value – p-value for differential expression
   padj – the Benjamini & Hochberg adjusted p values (corrected for multiple testing)
   Flag – can be either of the following:
   - *Tested – the statistical test was conducted, the log2 fold change and the adjusted p-value are valid*
   - *Low count – indicates one of the following cases:*
     - *The gene was filtered by the software DESeq2 due to low expression levels. The automatic filtration threshold is set by the software in order to maximize the number of significantly DE genes. In this case the adjusted p-value will be set to "NA"*
     - *The average expression of each of the replicates' groups was below 5*

     Note: When dealing with low expression levels, especially ones that are very close to zero, the statistical information might be misleading. We strongly recommend filtering the genes which were marked as "Low counts" from the table, or approach them with suspicion and use the normalized counts in order to verify that a result is indeed significant
   - *All zero – the counts of this gene are 0 in all samples. Such genes are not tested for differential expression and their* log2 fold change, p-value and adjusted p-value are all 'NA'

   Significantly DE – Significantly differentially expressed according to a threshold of 0.05 on the adjusted p-value - can be either "Up" or "Down" or "No".
   Columns of normalized counts grouped per condition

Notes:

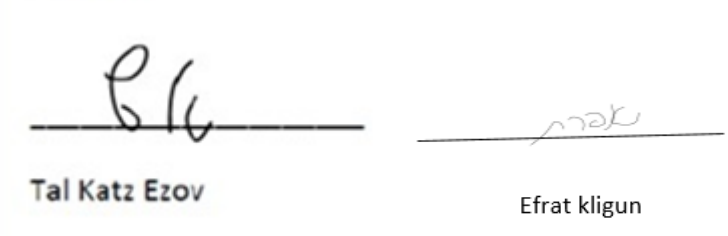
- The p-values and/or fold change values results for particular genes are set to 'NA' for either one of the following reasons:
  1. If log2 fold change, p-value and adjusted p-value are all 'NA' it indicates that all samples have zero counts in that gene (marked as "All_Zero" in the flag column).
  2. If only the adjusted p-value is 'NA' it indicates that automatic independent filtering was applied. The threshold for filtration is set automatically by the software for each comparison, so that the number of significantly DEGs will be maximized.

4. Files that allow visualization of the reference and alignment results files using Genome View software (http://www.genomeview.org/) or IGV (Integrative Genomics Viewer), available at www.broadinstitute.org/igv/.

You receive bam files (alignment files) and their index file (bai). When uploading the bam file to the IGV, the index file must be present in the same directory.

Visualization may be used to explore coverage profiles of genes of interest and detection of changes between samples.

Signatures:



Tal Katz Ezov

Efrat kligun

We appreciate acknowledgement of TGC sequencing and bioinformatics services in the publication of results obtained using our services by including the following sentence:

Sequencing, quality control, and differential expression analyses were conducted by the **"Technion Genome Center"**.