THE TRINITY PIPELINE

DE NOVO FULL-LENGTH ASSEMBLY AND DIFFERENTIAL ANALYSIS OF TRANSCRIPTOME FROM RNA-SEQ DATA IN MAIZE

*Nancy Wahl (nan241@tamu.edu), Seth C. Murray, Hong-Bin Zhang and Meiping Zhang
Department of Soil and Crop Sciences, Texas A&M University, College Station 77843-2474

Developed by
The Broad Institute and
Hebrew U. of
Jerusalem

Step 1
Generate FASTA sequence file in 3 stages:

Inchworm Chrysalis Butterfly

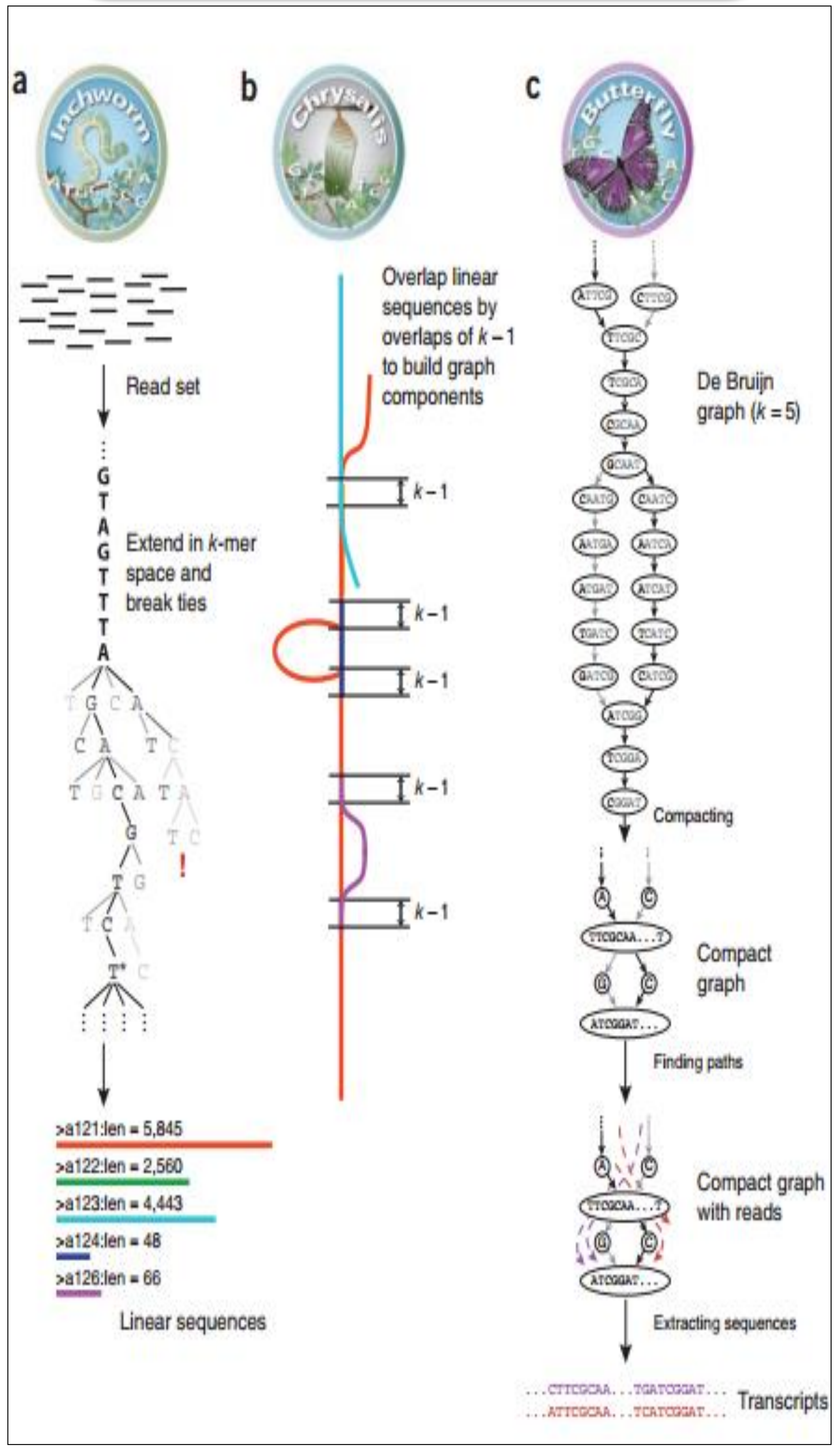
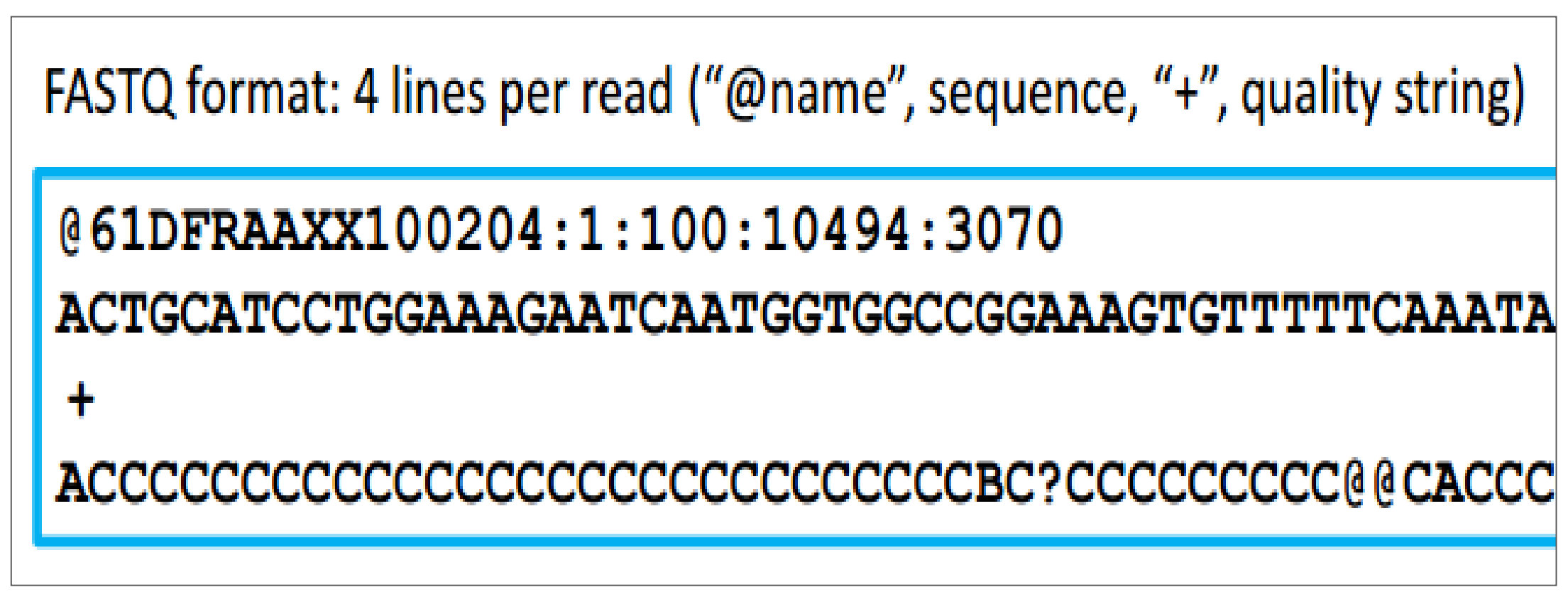


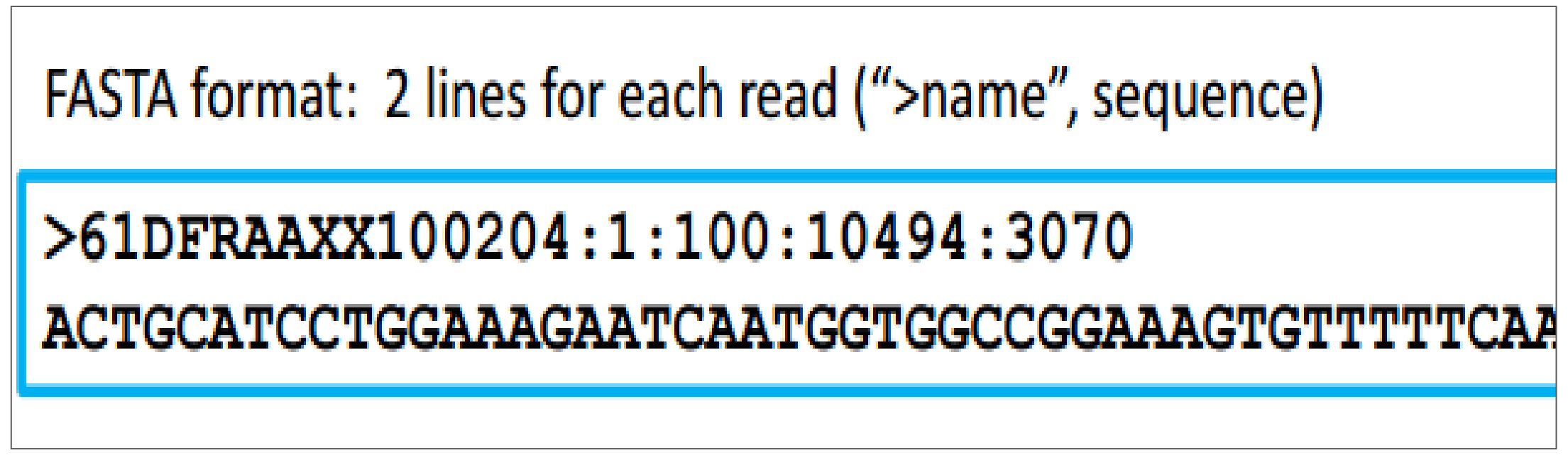
Figure 1. Stages of short read assembly: a. Inchworm assembles sequences into contigs. b. Chrysalis clusters these contigs that pertain to genes. c. Butterfly analyzes de Bruijn graphs from (b) in parallel in final assembly of transcripts.

Input: Millions of short paired end reads (sequenced from left and right) in fastq compressed files *.fq.gz that provide information on sequence and quality
Run time: 115 hours
Max Memory: 3814 MB **Max Threads:** 39
Key Script: Trinity --seqType fq --max_memory 50G --left <24 named *.fq.gz> --right <24 named *.fq.gz>
Output: Trinity.fasta file with assembled, labeled transcripts of genes and isoforms of genes.



FASTQ format: 4 lines per read (“@name”, sequence, “+”, quality string)

```
@61DFRAAXX100204:1:100:10494:3070
ACTGCATCCTGGAAAGAATCAATGGTGGCCGGAAAGTGTTTTTCAAATA
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCC
```



FASTA format: 2 lines for each read (“>name”, sequence)

```
>61DFRAAXX100204:1:100:10494:3070
ACTGCATCCTGGAAAGAATCAATGGTGGCCGGAAAGTGTTTTTCAA
```

REFERENCES

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011 May 15;29(7):644-52.
http://cbsu.tc.cornell.edu/lab/doc/Trinity_workshop_Part1.pdf

Bukowski, R and Sun, Q. *De novo transcriptome assembly using Trinity*.

ACKNOWLEDGEMENTS

USDA National Institute of Food and Agriculture Competitive Grant 2014-68004-21836
Key support provided by C. Michael Dickens of TAMU HPRC, Brian Haas of The Broad Inst. Of MIT and Harvard U., Mark Chapman of U. of Southhampton, U.K., Tiago Hori of The Center for Aquaculture Technologies and Ken Field of Bucknell U. are gratefully acknowledged.

INTRODUCTION

High-throughput sequencing and assembly of an organism’s genome (DNA sequence) and transcriptome (RNA transcripts) of its genes, which in turn code for proteins, requires **high-performance computing power.** This is especially true for reconstructing the entire DNA sequence of an unknown genome, or the full-length sequences of its transcriptome that lacks a reference genome, *de novo*. This study analyzes different RNA transcripts produced, which is termed “differential gene expression” (DE) in maize kernels treated with fungal spores using different methods at four different stages of kernel maturity. The maize genome has been fully sequenced, but due to significant numbers of small sequence variations among maize lines and unique experimental conditions, *de novo* sequencing was chosen.

Step 2
Create Gene Abundance Files

Input: Trinity.fasta, *.fq.gz files used as the reference for alignment based abundance estimation (counts of reads for each gene or isoform of gene)
Run time: 14.5 hours
Max Memory: 1193 MB **Max Threads:** 17
Modules Loaded: Trinity 2.2.0, RSEM 1.2.29, Bowtie2 2.2.9, SAMtools 1.3.
Key Script : align_estimate_abundance.pl --transcripts Trinity.fasta --seqType fq --left file1.fq.gz --right file2.fq.gz
Output: RSEM gene & isoform result files containing gene & isoform ids, sequence lengths and counts.

Step 3
Create Gene Count Matrix of Sequence Reads

Input: 24 RSEM gene result files
Run Time: About 60 seconds
Max Memory: 1841 MB **Max Threads:** 9
Modules Loaded: Trinity 2.2.0, RSEM 1.2.29, R_tamu 3.3.0.
Key Script: Abundance_estimtates_to_matrix.pl
--est_menthod RSEM (list of 24 RSEM.genes.results files)
Output: Matrix of gene counts in 152K rows of different gene sequences in 24 columns of sample libraries, matrix of normalized counts, and tables
of statistics, including library sizes (number of reads) and number of ‘genes’
expressed by at least 1 transcript per million in any one of the samples.

Step 4
Compare Replicates

Input: Matrix of gene counts, samples text file containing list of sample types alongside sample replicates.
Run Time: About 10 seconds
Max Memory: Range of 100 – 200 MB **Max Threads:** 7
Key Scripts: PtR --matrix (filename) -s samples.txt -- log2 (either 1. --compare_replicates , 2. --sample_cor_matrix or 3. --princ_comp 3)
Output: 1. Graphs of comparisons of replicates, 2. Heat map showing correlations among all replicates and 3. Principal component analysis (PCA) to show relationships among replicate samples.

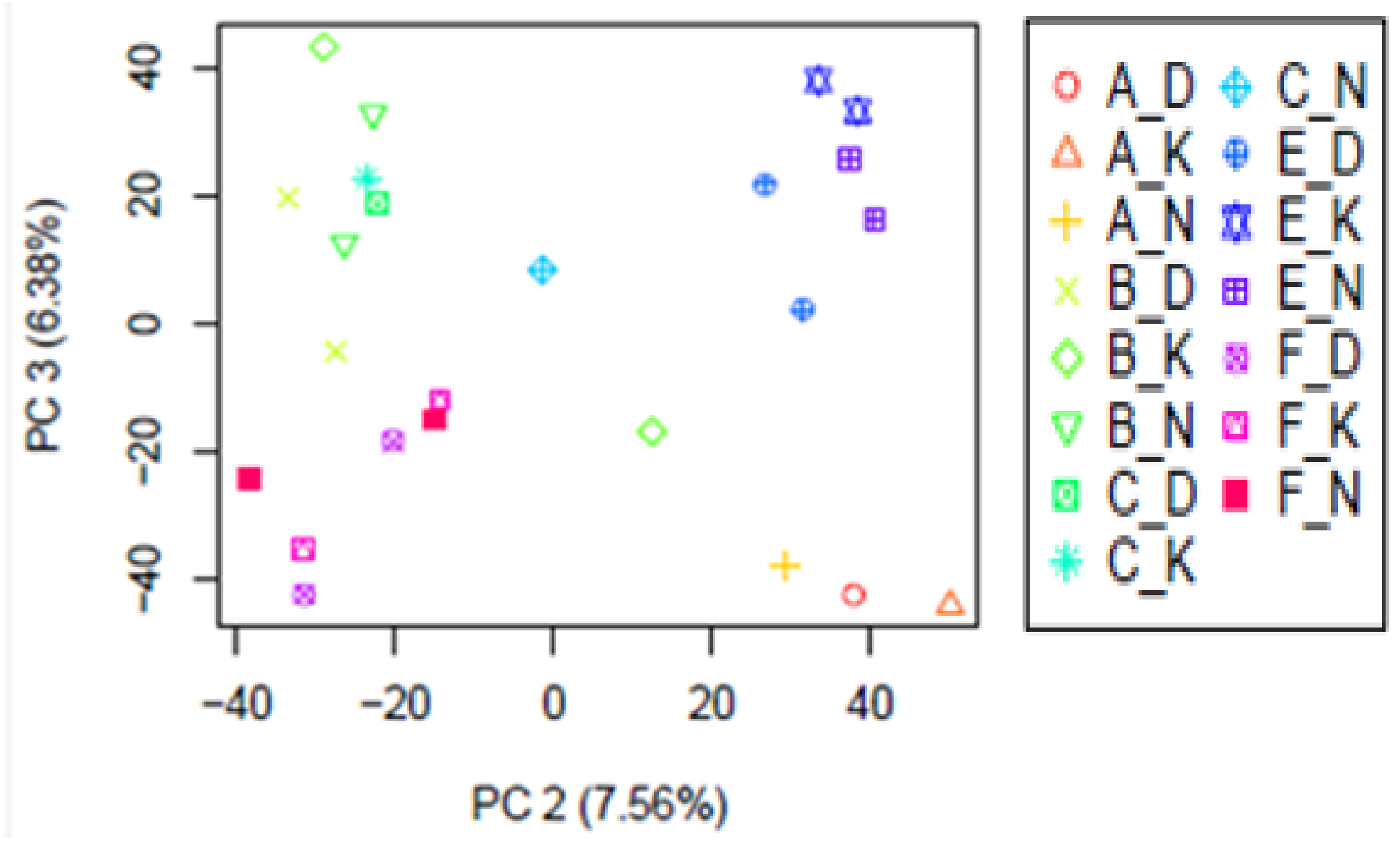


Figure 2. PCA of differential gene expression of maize kernels under different treatments. First letter refers to harvest dates, second letter refers to inoculation treatments of kernels. **A** – 6/08, **B** – 6/18, **E** – 6/26, **F** – 7/3. **D** – side needle, **K** – silk channel, **N**- none.

Step 7
Gene Ontology and Enrichment Analysis

Input: normalized and ‘trimmed’ expression matrix, comparison tables of DE genes obtained from output of Step 5. Optional: gene lengths text, GO annotation text
Run Time: About 4 -6 minutes
Max Memory: 800 MB **Max Threads:** 8
Key Scripts: analyze_diff_expr.pl --matrix <filename> --samples <string> --examine_GO_enrichment --GO_annots <string> --gene_lengths <string>
Output: Sample correlation heatmap, clustered heatmap of DE genes, tables of functional categories of up-regulated genes in defined comparisons (GO enrichment)

Table 1. (Below) Shows functional categories of genes differentially expressed in kernels inoculated with fungal spores vs non-inoculated at mid-maturity.

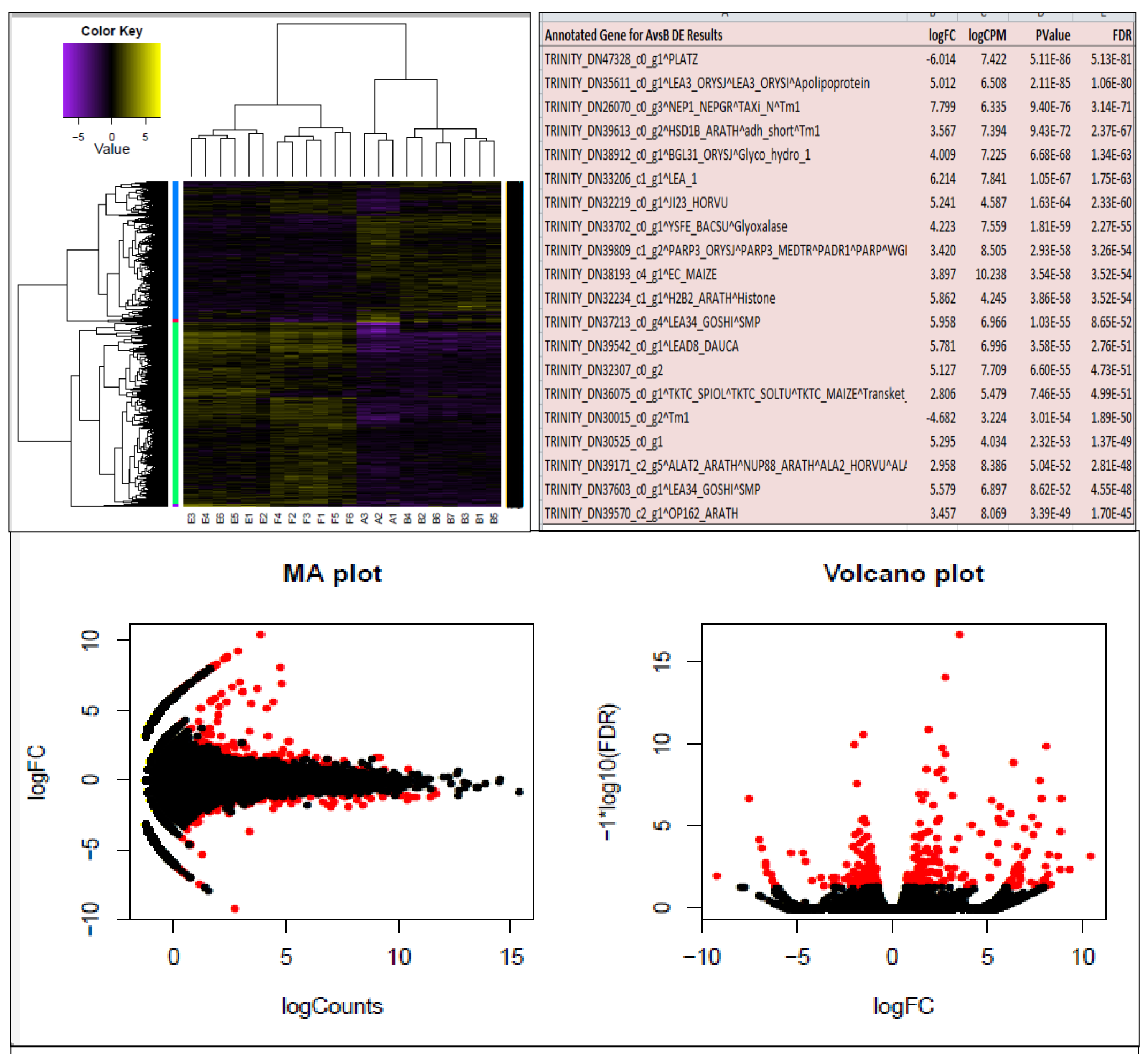
Gene Ontology	Over -	Number	Total Num	Functional Categories	Ontology
(GO) Categories	Rep	DE Genes	of Genes in		BP Biol Process
		in			
	Genes in	Category	Category		MF Molecluar Function
	Category				CC Cellular Component
GO:0009408	2.0E-06	4	340	response to heat	BP
GO:0009266	5.5E-05	4	792	response to temperature stimulus	BP
GO:0051731	8.8E-04	1	3	polynucleotide 5'-hydroxyl-kinase activity	MF
GO:1990534	9.0E-04	1	3	thermospermine oxidase activity	MF
GO:0003999	1.8E-03	1	6	adenine phosphoribosyltransferase activity	MF
GO:0006168	2.1E-03	1	7	adenine salvage	BP
GO:0043096	2.1E-03	1	7	purine nucleobase salvage	BP
GO:0006388	2.9E-03	1	10	tRNA splicing, via endonucleolytic cleavage and ligation	BP
GO:0050474	3.5E-03	1	12	(S)-norcoclaurine synthase activity	MF
GO:0046084	3.6E-03	1	12	adenine biosynthetic process	BP
GO:0016174	3.7E-03	1	13	NAD(P)H oxidase activity	MF
GO:0044209	3.8E-03	1	13	AMP salvage	BP
GO:0046083	3.8E-03	1	13	adenine metabolic process	BP
GO:0006379	4.3E-03	1	15	mRNA cleavage	BP
GO:0009628	4.5E-03	4	2540	response to abiotic stimulus	BP
GO:0007231	4.6E-03	1	16	osmosensory signaling pathway	BP
GO:0050664	5.5E-03	1	19	oxidoreductase activity, acting on NAD(P)H, oxygen as acceptor	MF
GO:0005849	5.5E-03	1	19	mRNA cleavage factor complex	CC

Step 6
Gene Annotation & Functional Analysis with Trinotate

TRINOTATE is a comprehensive annotation suite of programs designed to assign functional annotation to transcriptomes particularly de novo. It conducts homology searches to known sequence data (BLAST+/SwissProt) as applied to different protein domain predictions, and draws upon several annotation databases such as KEGG.
Input: Trinity.fasta file, Trinotate.sqlite database, conf.txt file
Run time: 76.5 hours
Max Memory: 4893 **Max Threads:** 28
Output: Annotation file on genes (or their isoforms)

Step 5
Analyze Differential Gene Expression

Input: Matrix of gene counts, method such as edgeR, samples text file, contrasts <string>
Run time: About 50 seconds
Max Memory: About 300 MB **Max Threads:** 7
Key Script: run_DE_analysis.pl --matrix <file name> --method edgeR --samples samples.txt --output <named directory>
Output: Comparison tables DE genes specified by contrast file with log fold changes, log counts per million (CPM) and FDR for each gene. Graphic displays described below.



Annotated Gene for AvsB DE Results	logFC	logCPM	PValue	FDR
TRINITY_DN47328_c0_g1^PLATZ	-6.014	7.422	5.11E-86	5.13E-81
TRINITY_DN35611_c0_g1^LEA3_ORYSJ^LEA3_ORYSI^Apolipoprotein	5.012	6.508	2.11E-85	1.06E-80
TRINITY_DN26070_c0_g3^NEP1_NEPGR^TAXi_N^Tm1	7.799	6.335	9.40E-76	3.14E-71
TRINITY_DN39613_c0_g2^HSD1B_ARATH^adh_short^Tm1	3.567	7.394	9.43E-72	2.37E-67
TRINITY_DN38912_c0_g1^BGL31_ORYSJ^Glyco_hydro_1	4.009	7.225	6.68E-68	1.34E-63
TRINITY_DN33206_c1_g1^LEA_1	6.214	7.841	1.05E-67	1.75E-63
TRINITY_DN32219_c0_g1^JI23_HORVU	5.241	4.587	1.63E-64	2.33E-60
TRINITY_DN33702_c0_g1^YSFE_BACSU^Glyoxalase	4.223	7.559	1.81E-59	2.27E-55
TRINITY_DN39809_c1_g2^PARP3_ORYSJ^PARP3_MEDTR^PADR1^PARP^WGI	3.420	8.505	2.93E-58	3.26E-54
TRINITY_DN38193_c4_g1^EC_MAIZE	3.897	10.238	3.54E-58	3.52E-54
TRINITY_DN32234_c1_g1^H2B2_ARATH^Histone	5.862	4.245	3.86E-58	3.52E-54
TRINITY_DN37213_c0_g4^LEA34_GOSHI^SMP	5.958	6.966	1.03E-55	8.65E-52
TRINITY_DN39542_c0_g1^LEAD8_DAUCA	5.781	6.996	3.58E-55	2.76E-51
TRINITY_DN32307_c0_g2	5.127	7.709	6.60E-55	4.73E-51
TRINITY_DN36075_c0_g1^TKTC_SPIOL^TKTC_SOLTU^TKTC_MAIZE^Transket	2.806	5.479	7.46E-55	4.99E-51
TRINITY_DN30015_c0_g2^Tm1	-4.682	3.224	3.01E-54	1.89E-50
TRINITY_DN30525_c0_g1	5.295	4.034	2.32E-53	1.37E-49
TRINITY_DN39171_c2_g5^ALAT2_ARATH^NUP88_ARATH^ALA2_HORVU^AL	2.958	8.386	5.04E-52	2.81E-48
TRINITY_DN37603_c0_g1^LEA34_GOSHI^SMP	5.579	6.897	8.62E-52	4.55E-48
TRINITY_DN39570_c2_g1^OP162_ARATH	3.457	8.069	3.39E-49	1.70E-45

Figure 3. Top left: Clustered heatmap of DE genes vs sample replicates. Top right: DE genes in one comparison. Bottom: MA plot of DE gene log counts and Volcano plots of log fold changes between un-inoculated maize kernels and side needle inoculated kernels of mid-maturity.