RNA-Seq Gene Expression Analysis at the Joint Genome Institute

Vasanth Singan (vrsingan@lbl.gov), Anna Lipzen, James Han and Erika Lindquist

DOE-Joint Genome Institute, 2800 Mitchell Dr., Walnut Creek, CA

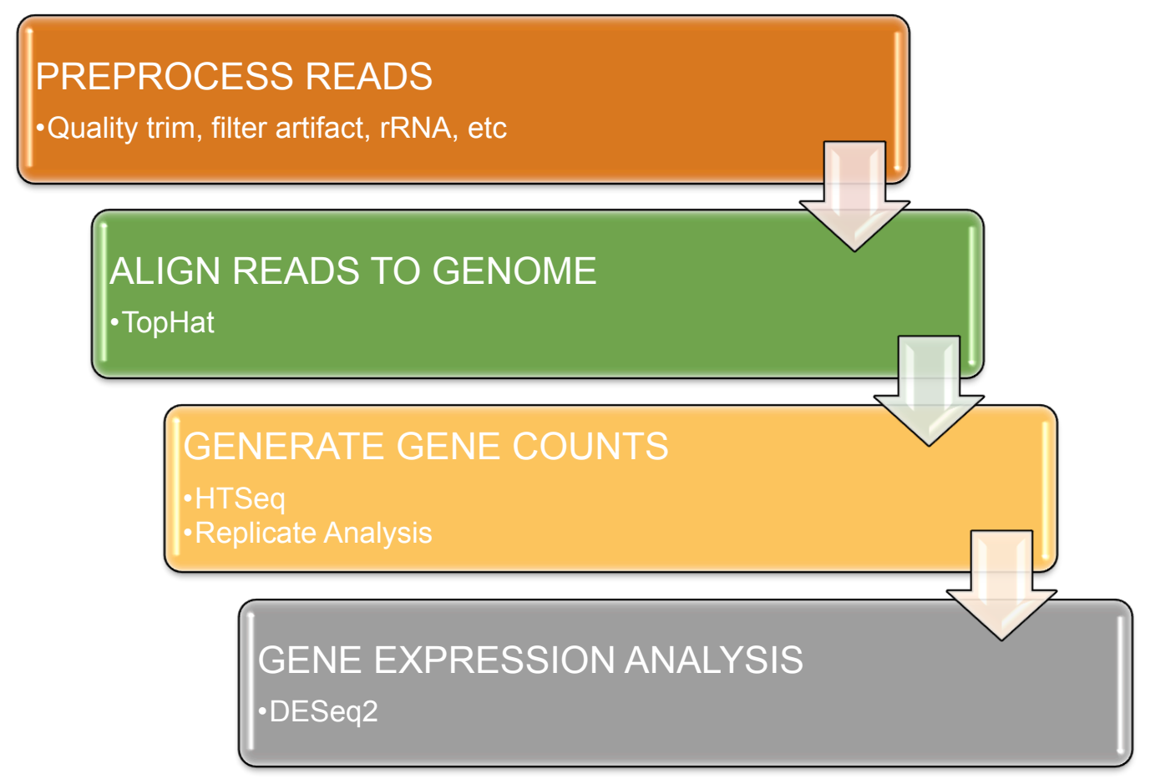
INTRODUCTION

Genes expressed in-tissues, levels of expression and comparative expression between different experimental conditions can be characterized by measurement of mRNA levels using RNA-Seq (Transcriptome sequencing). RNA-Seq uses Next-Generation Sequencing (NGS) to identify the presence and to quantify expressed genes. At the JGI, RNA from different experimental conditions provided by the users are sequenced using the RNA-seq technology and gene-wise expression levels are provided.

Several open source Differential Gene Expression (DGE) tools (Cufflinks, DESeq and EdgeR) were analyzed in an effort to improve the RNA-Seq gene expression analysis pipeline for Eukaryotic projects. The outcome of the analyses, current and new deliverables to users and the schematic of the analysis pipeline are covered in this poster.

RNA-Seq PIPELINE AT THE JGI

The RNA-Seq pipeline at the JGI combines a variety of tools to generate gene counts and call differentially expressed genes (see Figure 1). Reads from sequencers are preprocessed to perform a variety of tasks including quality trimming, filtering artifacts and removal of rRNA. Reads are then aligned to a reference genome using a splice-aware aligner (TopHat[1]). HTSeq[5] is then used to generate counts. The pipeline then calls differential expression based on the counts using DESeq2[3].



PREPROCESS READS
- Quality trim, filter artifact, rRNA, etc

ALIGN READS TO GENOME
- TopHat

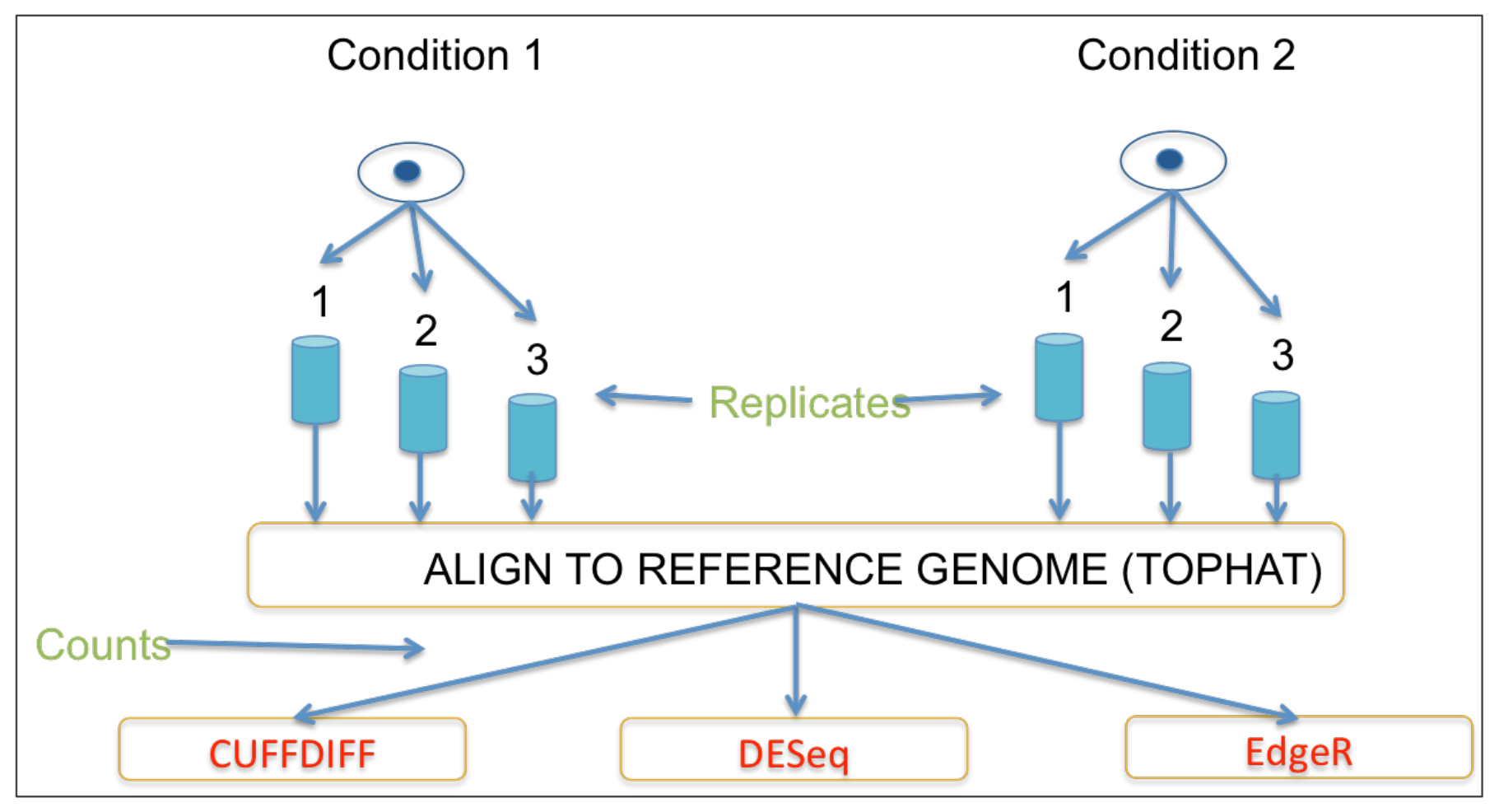
GENERATE GENE COUNTS
- HTSeq
- Replicate Analysis

GENE EXPRESSION ANALYSIS
- DESeq2

Figure 1: RNA-Seq DGE pipeline at the JGI

EXPERIMENTAL WORKFLOW (Tool Comparison)

The choice of tools for the DGE pipeline at JGI was decided based on an experimental workflow (Figure 2) comparing Cuffdiff[2], EdgeR[4] and DESeq[3]. To test performance of the three DGE tools, samples from two different conditions with three replicates each were aligned to the Genome using TopHat[1] and the counts were provided to the tool for differential gene calling. Significant differentially expressed genes, as identified by the tools were compared and investigated.



Condition 1

Condition 2

Replicates

ALIGN TO REFERENCE GENOME (TOPHAT)

Counts

CUFFDIFF

DESeq

EdgeR

Figure 2: Experimental Workflow

CHOICE OF TOOLS

Based on the experimental workflow, DGE tools were compared for a variety of fungal and plant samples.

DETECTING DIFFERENTIALLY EXPRESSED GENES

Genes called differentially expressed by the three tools were compared (Figure 3). In all tested cases, Cuffdiff called more genes as differentially expressed. DESeq was the most conservative.

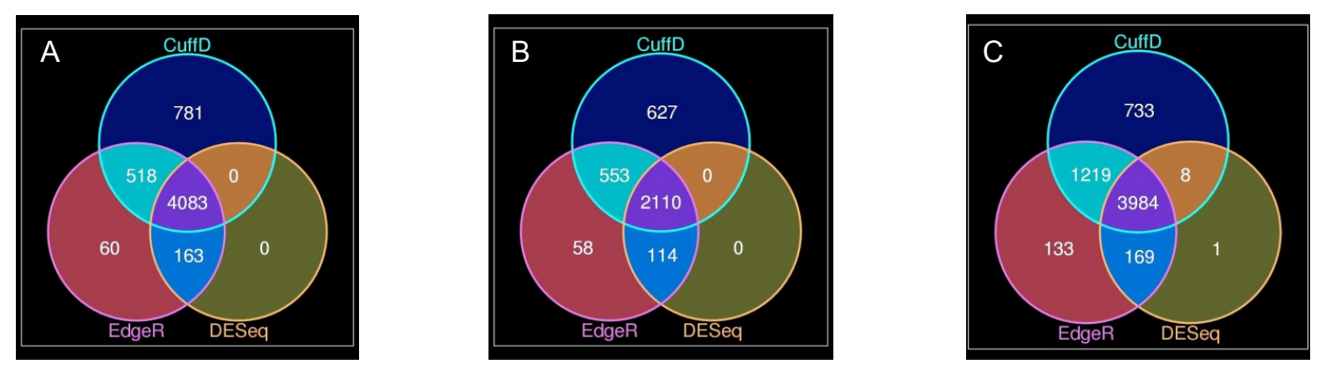


Figure 3: Venn Diagrams of significant differentially expressed gene identified by the 3 tools. A) *Dalidinia sp.* B) *Hypoxylon sp.* C) *Acidomycetes sp.*

P-VALUE THRESHOLD

In order to compare the 3 tools, fpkm vs fpkm plots were generated with each dot representing a gene and the color representing the tool(s) that identified the gene as differentially expressed (Figure 4). The Cuffdiff p-value calculation appeared to be different when compared to the other techniques. In the *Daldinia sp.* example provided below, at a threshold of p<0.05, Cuffdiff found 781 more genes than the other tools as evident in the (normalized gene count) fpkm vs fpkm plot (Figure 4A – CuffD only). Adjusting the threshold to p<0.01 eliminated over 95% of these genes (Figure 4B). This trend was consistent across fungal and plant genomes. In cases where fpkm of genes was relatively low in one of the conditions, DESeq and EdgeR identified these genes as differentially expressed (Figure 4 – orange dots) whereas Cuffdiff did not.

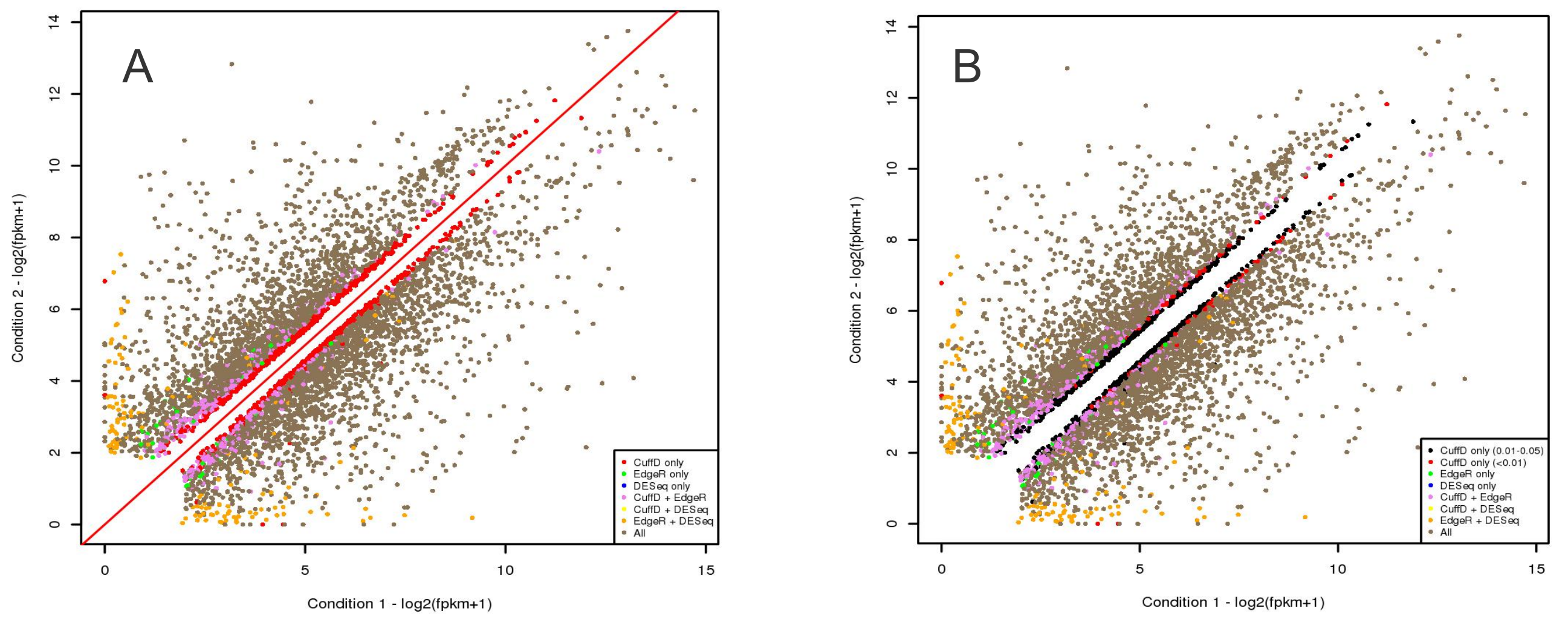


Figure 4: A) *Daldinia sp.* fpkm vs fpkm plot with Cuffdiff p<0.05. B) *Daldinia sp.* fpkm vs fpkm plot with Cuffdiff p<0.01 and p<0.05.

REPLICATE VARIANCE

To test the effect of replicate variance, the fpkm vs fpkm plot was generated with the size of the dots based on %CV (Coefficient of Variance) between replicates (larger dot = higher %CV) (Figure 5). In cases where genes are differentially expressed but are expressed in both conditions, Cuffdiff and EdgeR identified some genes as differentially expressed even in cases with high %CV (Fig 5 black arrows). DESeq appeared to handle the replicate variability better for DGE analysis.

ON/OFF GENES

In cases where genes were ON in one condition and OFF in another, both DeSeq and EdgeR identified the genes as differentially expressed whereas Cuffdiff did not. In the *Oryza sativa* example provided below, about 4% of the genes were ON and OFF between conditions. Cuffdiff did not call any of these genes as differentially expressed (For example see Figure 5 – blue arrows). ON/OFF genes have high %CV (Figure 5 – larger dots with blue arrows) due to low fpkm in one of the conditions but are still significant. This trend was consistent across all experiments.

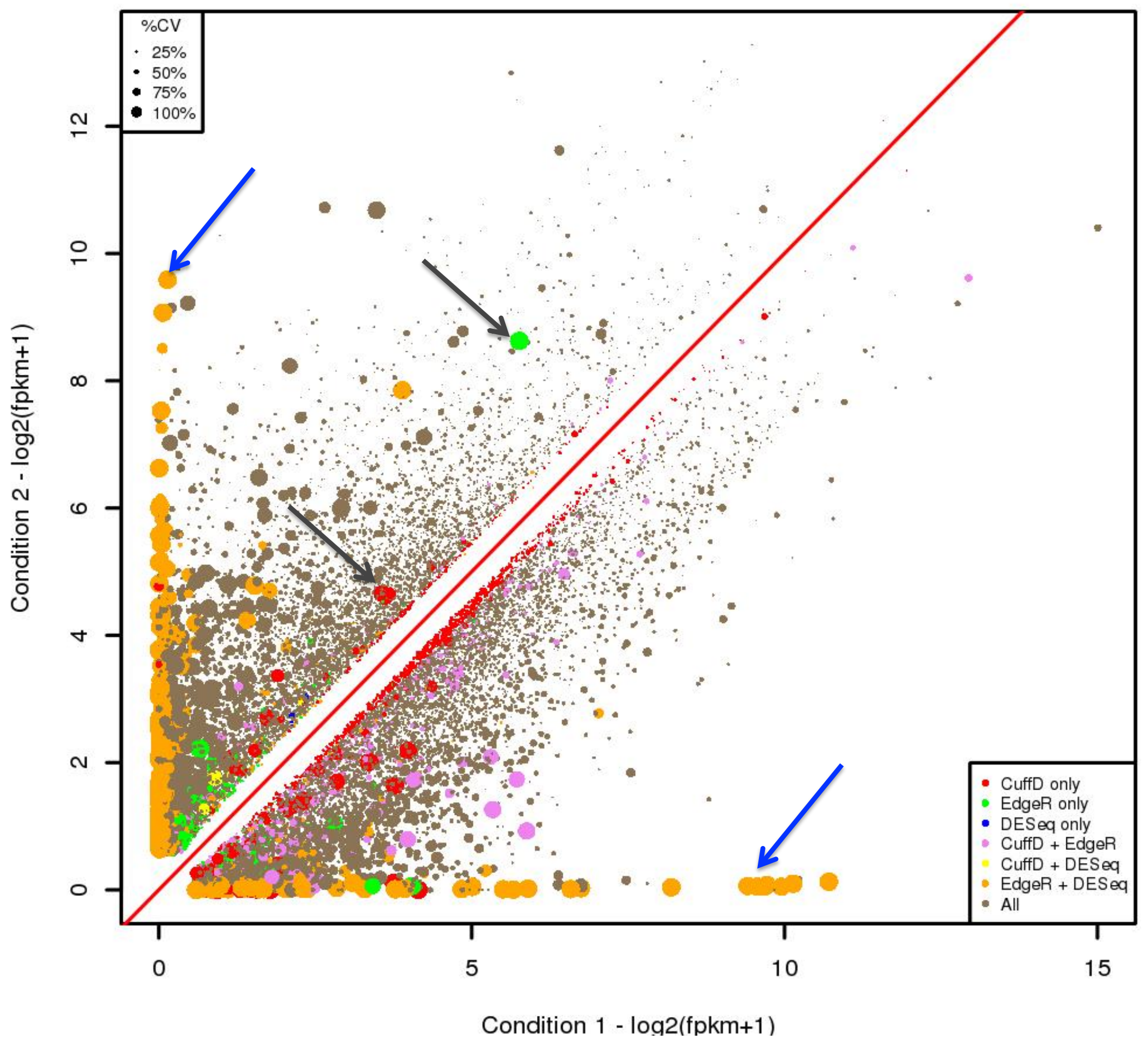


Figure 4: *Oryza sativa* fpkm vs fpkm of differentially expressed genes identified by the 3 tools and sized by CV% between replicates.

Based on the above tests, DESeq was selected as the tool of choice for the pipeline.

LIBRARY QC

Replicate analysis is done using Pearson Correlation (PC) of fragment counts for each pair of libraries. A textfile containing the correlations in a matrix format (**replicate_analysis.txt**) is provided as part of the deliverables. Additionally, a heatmap visualization of correlations (**replicate_analysis_heatmap.pdf**) grouped by replicates (white box) is provided (see Figure 6). Poorly correlated replicates can be removed from downstream DGE analysis.

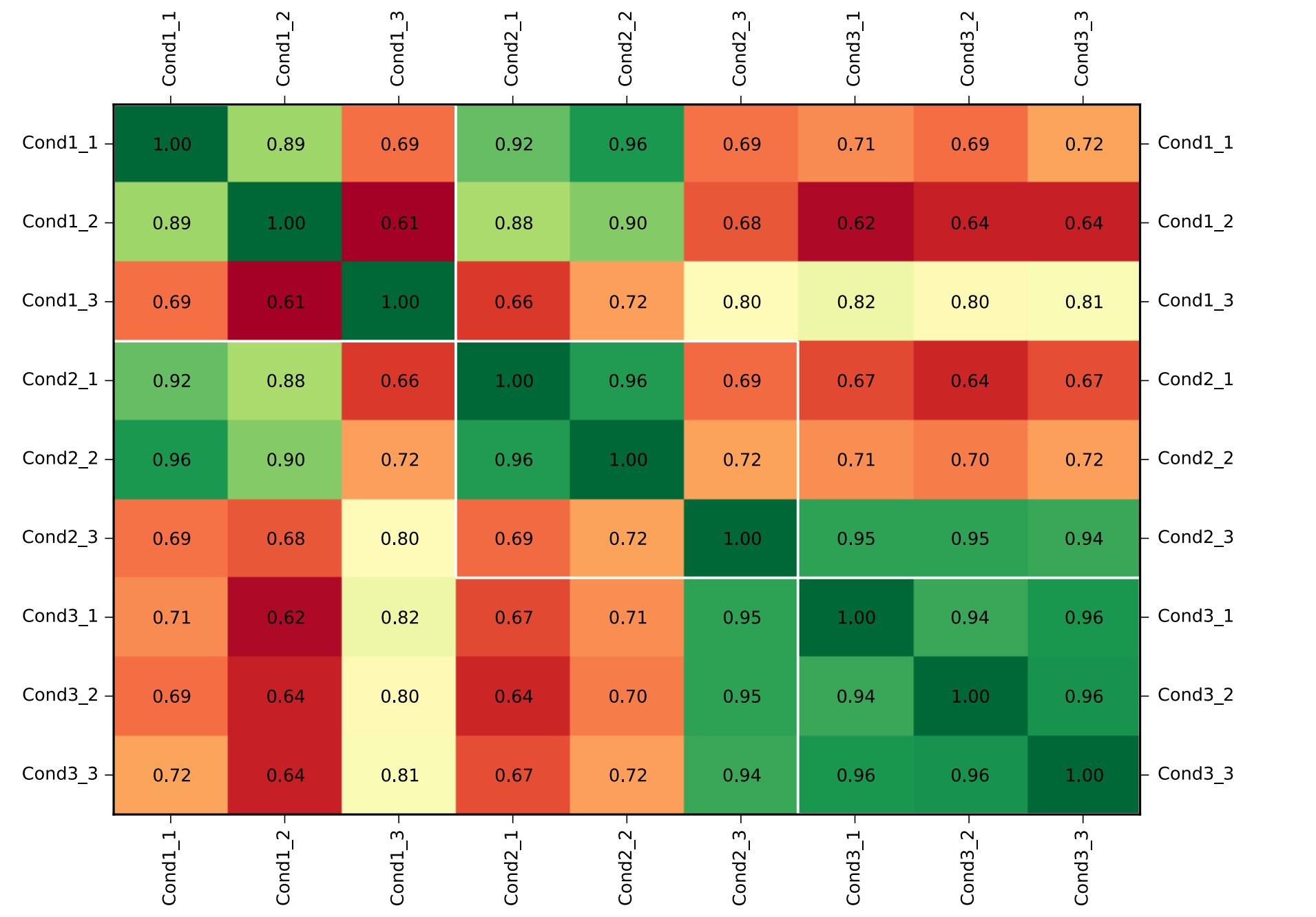


Figure 6: replicate_analysis_heatmap.pdf

DELIVERABLES

The following are lists of deliverables provided by the RNA-seq pipeline.

Fragment counts generated using HTSeq are provided as a text file (**counts.txt**). Figure 7 shows an example of the counts file. Fragment counts are provided for each library (replicate).

GeneID	LIB1	LIB2	LIB3
Gene 1	0	0	1
Gene 2	28	146	220
:	:	:	:
:	:	:	:

Figure 7: counts.txt (fragment counts)

Differential Gene Expression analysis is performed using DESeq2. The replicate correlation information is used to determine which of the replicates to include in the DGE analysis. Outliers with a low correlation may be excluded from the DGE analysis because they may bias the results by introducing artificially high noise. A summary table (**DGE_summary.txt**) containing the log2 Fold Change, adjusted p-value (padj) and a Boolean significance value based on threshold of padj at 0.05 for all pairs of conditions is provided (Figure 8).

GeneID	cond_1vs2 log2FoldChange	cond_1vs2 padj	Cond_1vs2 significant	cond_2vs3 log2FoldChange	Cond_2vs3 padj	cond_2vs3 significant
Gene 1	-16	NA	NA	1	NA	NA
Gene 2	-0.48	0.48	FALSE	-0.15	0.88	FALSE
Gene 3	-1.8	0.0034	TRUE	1.3	0.1	FALSE
:	:	:	:	:	:	:
:	:	:	:	:	:	:

Figure 8: DGE_summary.txt (pairwise DGE analysis results)

REFERENCES

1. Trapnell C, Pachter L, Salzberg SL. **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* doi:10.1093/bioinformatics/btp120
2. Trapnell C, Hendrickson D,Sauvageau S, Goff L, Rinn JL, Pachter L **Differential analysis of gene regulation at transcript resolution with RNA-seq** *Nature Biotechnology* doi:10.1038/nbt.2450
3. Anders S, Huber W **Differential expression analysis for sequence count data.** *Genome Biology* 2010;11(10):R106. doi: 10.1186/gb-2010-11-10-r106. Epub 2010 Oct 27.
4. Robinson MD, McCarthy DJ, Smyth GK **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11
5. Anders S, Pyl PT, Huber W **HTSeq – A python framework to work with high-throughput sequencing data.** *bioRxiv* preprint (2014), doi: 10.1101/002824

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.