

RNAseq Differential Gene Expression Analysis Report

Customer Name:

Institute/Company:

Project:

NGS Data: IlluminaHiSeq2500 2x126bp PE

Bioinformatics Service: Differential gene expression analysis

Sample Species:

Number of Samples:

Date:

Otogenetics Contact: Bioinformatics

Phone: (855) OTOGENE (686-4363)

Email: bioinfor@otogenetics.com

1. Description of Workflow

Total RNA was submitted to Otogenetics Corporation (Atlanta, GA USA) for RNA-Seq assays. Briefly, the integrity and purity of total RNA were assessed using Agilent Bioanalyzer and OD260/280 using Nanodrop. 1-2 µg of cDNA was generated using Clontech Smart cDNA kit (Clontech Laboratories, Inc., Mountain View, CA USA, catalog# 634925) from 100ng of total RNA. cDNA was fragmented using Covaris (Covaris, Inc., Woburn, MA USA), profiled using Agilent Bioanalyzer, and subjected to Illumina library preparation using NEBNext reagents (New England Biolabs, Ipswich, MA USA, catalog# E6040).

1.1 Illumina RNA-Seq sample preparation workflow

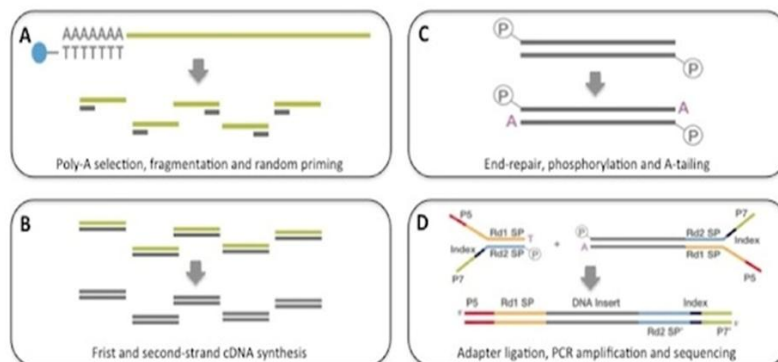


Figure 1.1 RNA sample preparation. A. mRNAs are purified using Poly(A) selection from total RNA sample, and then fragmented. **B.** First strand of cDNA is synthesized using random priming, followed by the synthesis of the second strand of cDNA. **C.** The resulting double-strand cDNA from step B is end repaired, phosphorylated and A-tailed. **D.** Adapter ligation and PCR amplification are performed, the library is ready for clustering and sequencing.

2. Raw Data Overview

Table 2.1. Quality control. Data summary of generated reads.

Lane	Barcode	Oto#	Species	Allowance Gbp	Actual Size Gbp	Reads	Sufficient	Coverage	Efficiency(%)	Mapped Percent	Duplicates(%)	Intergenic In Mapped(%)	Fast Qc Report	Quality	Q30 Pct	Duplicates Pct
HS-835_RN1	ID42	Ot##	Human	5.50	5.16	40,931,484	Yes	63.0	65.7%	99.8%	9.59%		fastq1 fastq2	pass	98.1%; 95.7%	30.07%; 29.08%
HS-835_RN1	ID43	Ot##	Human	5.50	5.94	47,145,336	Yes	71.4	64.9%	99.8%	9.59%		fastq1 fastq2	pass	98.1%; 95.2%	31.36%; 30.27%
HS-835_RN1	ID45	Ot##	Human	5.50	5.42	43,024,450	Yes	66.4	65.9%	99.9%	9.85%		fastq1 fastq2	pass	98.1%; 95.9%	32.46%; 31.49%
HS-835_RN1	ID46	Ot##	Human	5.50	5.21	41,374,002	Yes	63.9	66.3%	99.8%	9.22%		fastq1 fastq2	pass	98.1%; 95.6%	29.85%; 28.63%
HS-835_RN1	ID47	Ot##	Human	5.50	5.14	40,758,108	Yes	62.9	66.0%	99.8%	9.90%		fastq1 fastq2	pass	98.2%; 95.7%	29.8%; 28.98%

The quality, quantity and size distribution of the Illumina libraries were determined using an Agilent Bioanalyzer 2100. The libraries were then submitted for Illumina HiSeq2500 sequencing according to the standard operation. Paired-end 90-100 nucleotide (nt) reads were generated and checked for data quality using FASTQC (Babraham Institute, Cambridge, UK). After achieving optimum QC results, samples were analyzed.

3. Bioinformatics analysis workflow

To analyze the data we used the following pipeline.

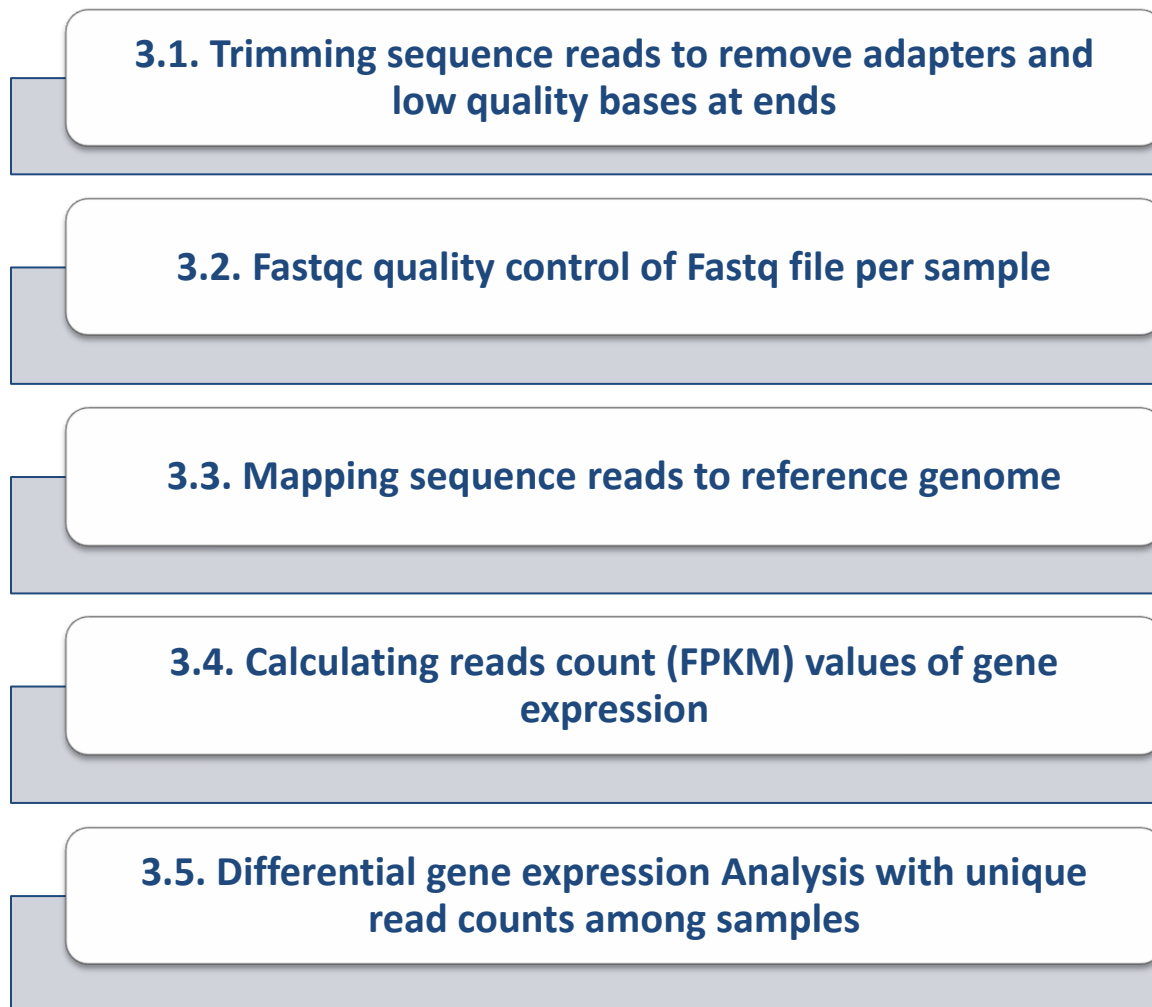


Figure 3.1. RNA-seq workflow data analysis.

3.1. Trimming sequence reads to remove adapters and low quality bases at ends.

Poor quality or technical sequences can affect the downstream analysis and data interpretation, which lead to inaccurate results. To assess quality of raw sequenced data, we used FastQC before and after trimming the adapters.

Sequence reads were trimmed to remove possible adapter sequences and nucleotides with poor quality (error rate < 0.05) at the end. After trimming, sequence reads shorter than 30 nucleotides were discarded.

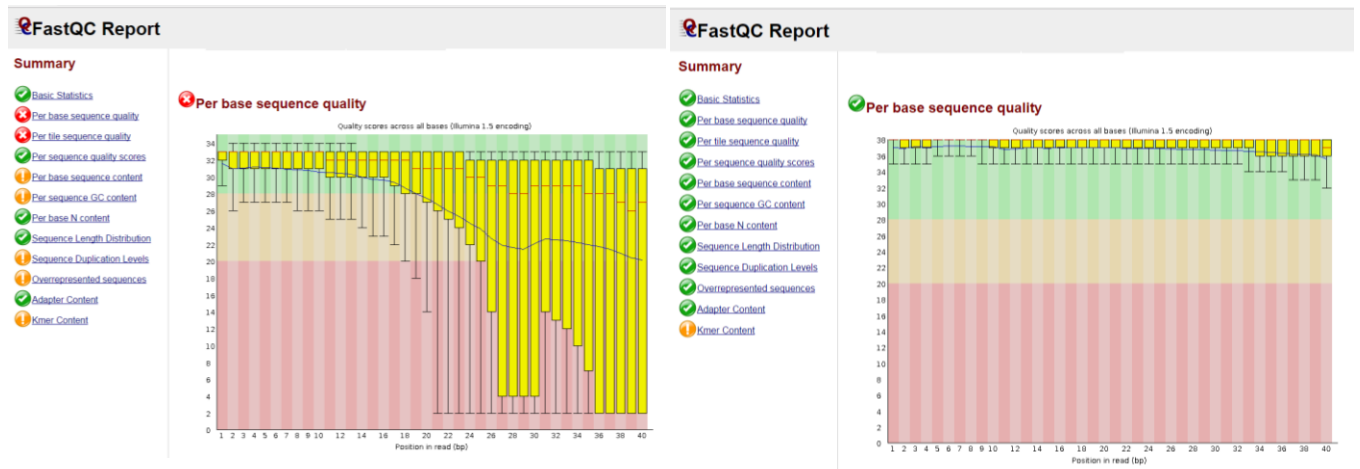


Figure 3. 2. Data quality control. Before trimming (left) and after trimming (right).

You will see data after trimming quality control in your folder under “fastq” with an extension of .fastqc.gz. (There are two per sample, left and right reads).

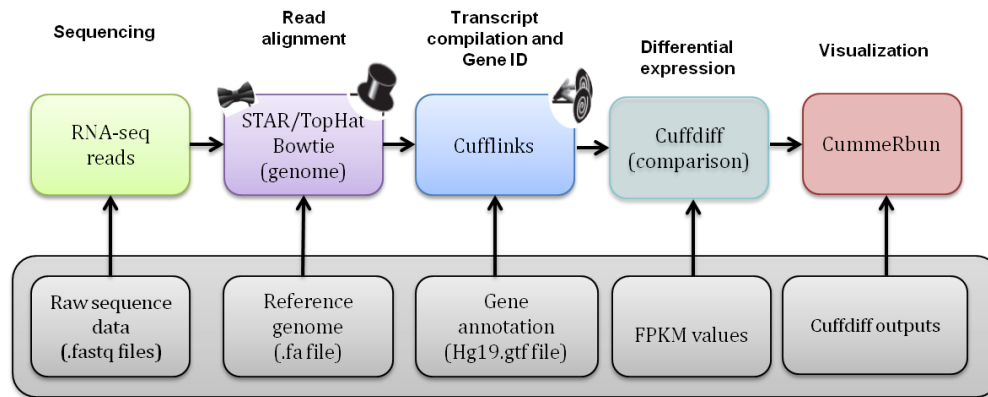


Figure 3.3. Data analysis pipeline. Pipeline use to perform the RNA-seq data analysis.

3.2 Mapping sequence reads to reference genome

The short reads were then mapped to a reference genome assembly to discover their locations with respect to that reference using HiSat a popular spliced aligner for RNA-sequence (RNA-seq) experiments, from the tuxedo protocol.

The results of the mapping can be found in the file: mapping/yoursample.bam

The table below shows the summary of the mapping for the samples. Summary Stats of Mapping (MappingStats.log.gz)

Table 3.1. Summary of the mapping for the samples.

Started job on	X/XX/2017 12:40:11 AM
Started mapping on	X/XX/2017 12:40:21 AM
Finished on	X/XX/2017 1:00:16 AM
Mapping speed, Million of reads per hour	149.47
Number of input reads	49615041
Average input read length	252
UNIQUE READS:	
Uniquely mapped reads number	40503208
Uniquely mapped reads %	81.63%
Average mapped length	239.96
Number of splices: Total	37682359
Number of splices: Annotated (sjdb)	37093908
Number of splices: GT/AG	37253900
Number of splices: GC/AG	288473
Number of splices: AT/AC	29937
Number of splices: Non-canonical	110049
Mismatch rate per base, %	0.41%
Deletion rate per base	0.01%
Deletion average length	1.62
Insertion rate per base	0.01%
Insertion average length	1.53
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	3016636
% of reads mapped to multiple loci	6.08%
Number of reads mapped to too many loci	20654
% of reads mapped to too many loci	0.04%
UNMAPPED READS:	
% of reads unmapped: too many mismatch	0.00%
% of reads unmapped: too short	12.22%
% of reads unmapped: other	0.03%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

4. Calculating reads count (FPKM) values of gene expression

The mapping bam files are then imported to the “Cufflinks” software from the tuxedo protocol, and FPKM values are calculated. You can find these results in the “Cufflinks” directory of your results.

Your results will have:

4.1. Transcriptome assembly: transcripts.gtf

This GTF file contains Cufflinks’ assembled isoforms. The first 7 columns are standard GTF, and the last column contains attributes, some of which are also standardized (“gene_id”, and “transcript_id”).

4.2. Transcript-level expression: isoforms.fpkm_tracking

This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format.

4.3. Gene-level expression: genes.fpkm_tracking

This file contains the estimated gene-level expression values in the generic FPKM Tracking Format.

For more information about the file format visit: <http://cole-trapnell-lab.github.io/cufflinks/cufflinks/>

5. Differential gene expression analysis with unique read counts among samples.

Identification of differentially expressed genes/transcripts using Cuffdiff: Cufflinks includes “Cuffdiff,” which is a program used to find significant changes in transcript expression.

Cuffdiff: uses all the bam files from tophat output and compares (differential) across samples or group of samples and will generate similar files with suffix .diff ; you can use reported fold change information for comparing samples.

Cuffdiff like Cufflinks calculates the FPKM of each transcript, primary transcript, and gene in each sample. FPKMs are computed by summing the FPKMs of transcripts in each primary transcript group or gene group.

The results are output in FPKM tracking files in the format described here:

5.1. FPKM tracking files: There are four FPKM tracking files.

Sum the FPKMs of transcripts in each primary transcript group or gene group.

isoforms.fpk_tracking: Transcript FPKMs

genes.fpk_tracking: Gene FPKMs

cds.fpk_tracking: Coding sequence FPKMs.

tss_groups.fpk_tracking: Primary transcript FPKMs.

5.2. Count tracking files: There are four Count tracking files.

Cuffdiff estimates the number of fragments that originated from each transcript.

isoforms.count_tracking: Transcript counts

genes.count_tracking: Gene counts.

cds.count_tracking: Coding sequence counts.

tss_groups.count_tracking: Primary transcript counts.

5.3. Read group tracking files: There are four read group tracking files.

Cuffdiff calculates the expression and fragment count for each transcript, primary transcript, and gene in each replicate.

isoforms.read_group_tracking: Transcript read group tracking

genes.read_group_tracking: Gene read group tracking.

cds.read_group_tracking: Coding sequence FPKMs.

tss_groups.read_group_tracking: Primary transcript FPKMs.

5.4. Differential expression tests: Four files are created.

Tab delimited file lists the results of differential expression testing between samples

isoform_exp.diff: Transcript-level differential expression.

gene_exp.diff: Gene-level differential expression.

tss_group_exp.diff: Primary transcript differential expression.

cds_exp.diff: Coding sequence differential expression.

For more information about the output files format visit: <http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/>

6. Data visualization via CummeRbund

This is an R package designed to help with the visualization of the large amount Cuffdiff RNA-Seq outputs. Figures reproduce from CummeRbund are showed in figure 6.1.

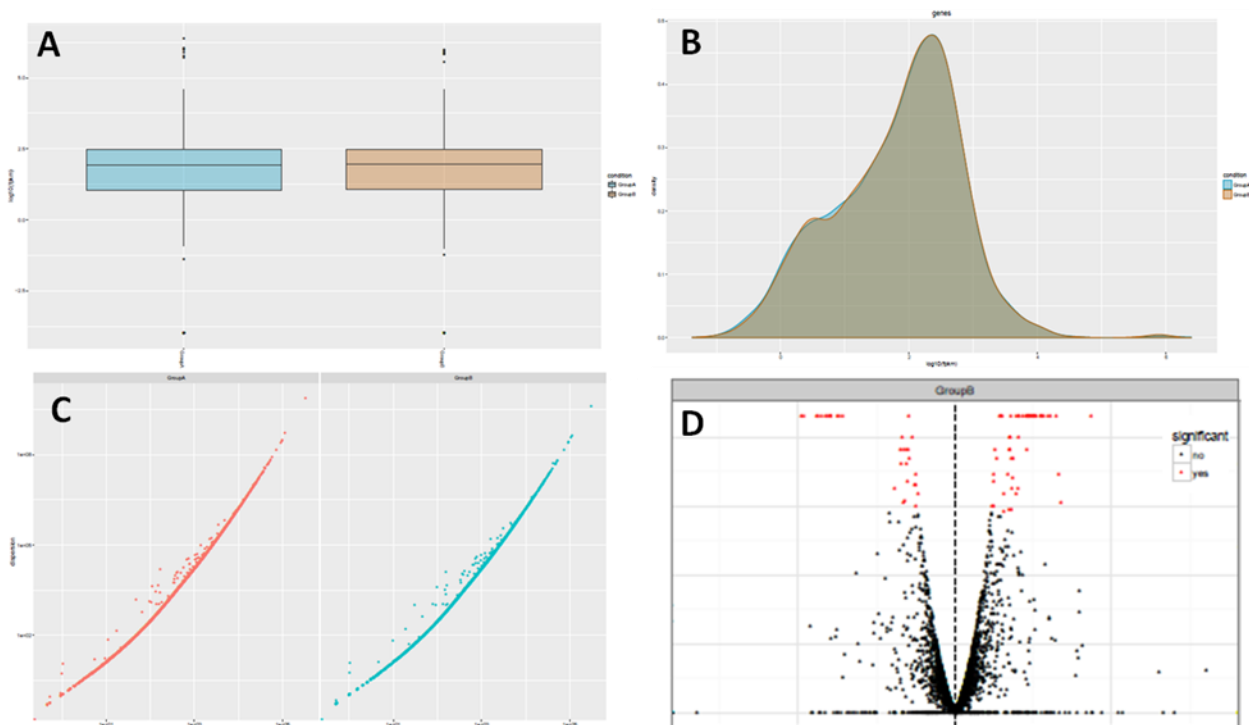


Figure 6.1. Visualization of the results cummeRbund. **A.** Box plots of the groups. **B.** Expression level distribution for all genes in the experimental conditions. **C.** Scatter plots highlight general similarities and specific outliers between conditions. **D.** Volcano plots reveal genes, transcripts, TSS groups or CDS groups that differ significantly between the pairs of conditions.

6. Deliverables

Raw fastq and fastqc files: Fastq dataset

Yoursamplename_R1_001.fastq.gz

Yoursamplename_R1_001.fastq.gz.md5.txt

Yoursamplename_R1_001_fastqc.zip

Yoursamplename_R2_001.fastq.gz

Yoursamplename_R2_001.fastq.gz.md5.txt

Yoursamplename_R2_001_fastqc.zip

Yoursamplename_SampleSheet.csv

Mapping

Yoursamplename.genome.bam

Yoursamplename.genome.deduplicated.bai

Yoursamplename.genome.deduplicated.bam

Yoursamplename.genome.deduplicated.RNAseqmetric

Yoursamplename.genome.duplication_metrics

Yoursamplename.MappingStats.log.gz

Yoursamplename.SJ.out.tsv.gz

Cufflinks

Yoursamplename.genes.fpkms_tracking

Yoursamplename.isoforms.fpkms_tracking

Yoursamplename.transcripts.gtf.gz

skipped.gtf

transcripts.gtf.gz

cuffdiff/

Each comparison data analysis output tables: GroupA_Vs_GroupB/

bias_params.info

cds.count_tracking

cds.diff

cds.fpkms_tracking

cds.read_group_tracking

cds_exp.diff

gene_exp.diff

genes.count_tracking

genes.fpkms_tracking

genes.read_group_tracking

isoform_exp.diff

isoforms.count_tracking

isoforms.fpkm_tracking

isoforms.read_group_tracking

promoters.diff

read_groups.info

run.info

splicing.diff

tss_group_exp.diff

tss_groups.count_tracking

tss_groups.fpkm_tracking

tss_groups.read_group_tracking

var_model.info

vennt-report.html

CummeRbund figures

Box.pdf

Density.pdf

Dispersion.pdf

FPKM.pdf

FPKM_replicates

Volcano.pdf

Overall data analysis report.

References

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29:15-21.

Trapnell C, Roberts A, Goff L, Perte G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;7:562–578.

Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25:1105-1111.