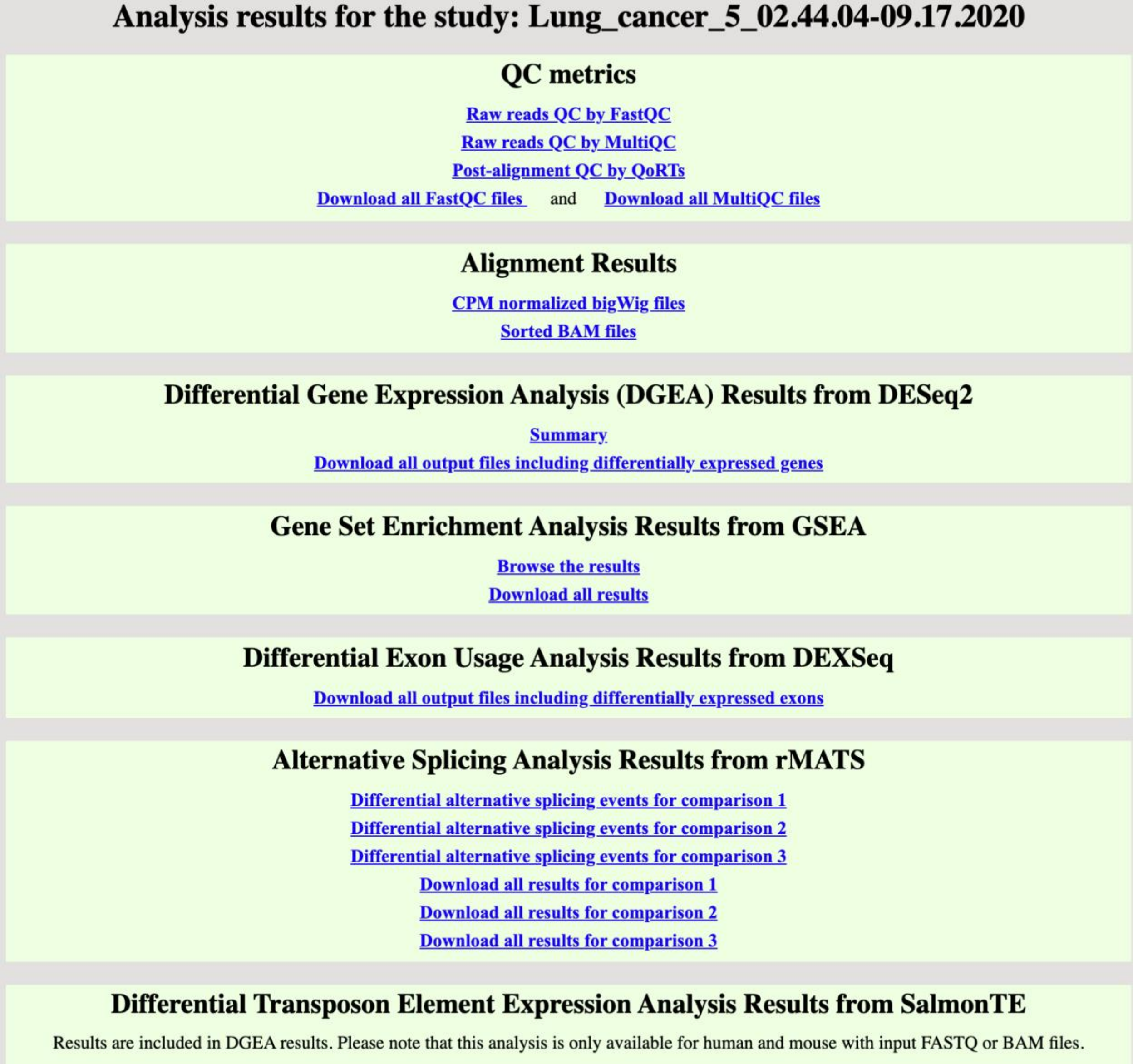
Description of outputs from analyzing the example lung_cancer dataset used in the case study

Table of Contents

Link to example results:

https://mccb.umassmed.edu/OneStopRNAseq/ViewResults.php?user=116&study=Lung_cancer_5_02.44.04-09.17.2020&study_id=604

A screenshot of analysis results index page:



Analysis results for the study: Lung_cancer_5_02.44.04-09.17.2020

QC metrics

Raw reads QC by FastQC
Raw reads QC by MultiQC
Post-alignment QC by QoRTs
Download all FastQC files and Download all MultiQC files

Alignment Results

CPM normalized bigWig files
Sorted BAM files

Differential Gene Expression Analysis (DGEA) Results from DESeq2

Summary
Download all output files including differentially expressed genes

Gene Set Enrichment Analysis Results from GSEA

Browse the results
Download all results

Differential Exon Usage Analysis Results from DEXSeq

Download all output files including differentially expressed exons

Alternative Splicing Analysis Results from rMATS

Differential alternative splicing events for comparison 1
Differential alternative splicing events for comparison 2
Differential alternative splicing events for comparison 3
Download all results for comparison 1
Download all results for comparison 2
Download all results for comparison 3

Differential Transposon Element Expression Analysis Results from SalmonTE

Results are included in DGEA results. Please note that this analysis is only available for human and mouse with input FASTQ or BAM files.

QC metrics:

Raw reads QC by FastQC and MultiQC:

A MultiQC report shows the FastQC metrics for all FASTQ files. Here is an official guide to understand FastQC report: https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

Post-alignment QC by QoRTs:

QoRTs post-alignment summary plots display various characteristics of each sequenced library such as Insert size distribution, gene-body coverage, clipping, deletion, insertion, splice junction rate for each read position, nucleotide composition for each read cycle, and Phread score for each read cycle. Please refer to http://hartleys.github.io/QoRTs/doc/QoRTs-vignette.pdf to understand each plot in details.

Alignment Results:

CPM normalized bigWig files:

These are normalized smoothed read coverage files in bigWig format (https://software.broadinstitute.org/software/igv/bigwig). CPM is calculated by deepTools, with binSize of 10bp. Only uniquely mapped reads (MQ>20) are included if users selected 'Include only uniquely mapped reads' when submitting the job. Otherwise, all aligned reads are included. These files are relatively small in size and can be used to visualize read coverage along the genome annotation, alternative splicing events, and differential exon usage via trackViewer (https://www.bioconductor.org/packages/release/bioc/html/trackViewer.html), UCSC Genome Browser (https://genome.ucsc.edu/goldenPath/help/bigWig.html), Integrated Genome Browser (IGV, https://software.broadinstitute.org/software/igv/home), or other genome browsers. In addition, they can be used to create interactive figures—that is, figures one can easily customize the features of by clicking, dragging, and typing using trackViewer (https://www.bioconductor.org/packages/release/bioc/html/trackViewer.html)

Sorted BAM files:

These are coordinate-sorted alignment files in base-pair resolution in BAM file format (https://samtools.github.io/hts-specs/SAMv1.pdf). These files are much larger than bigWig files and can be visualized using trackViewer or IGV but not by UCSC Genome Browser. The advantage of these files is that users can visualize the nucleotide composition, the Phred quality score, and detailed mapping status for each read/fragment in base-pair resolution. Therefore, these files are very useful for detailed quality assessment and for downstream analysis such as differential gene expression and alternative splicing analyses. These files are also useful for creating Shashimi plot via IGV and for visualizing alternative splicing evens and differential exon usage (https://software.broadinstitute.org/software/igv/Sashimi). In addition to sorted BAM files, OneStopRNAseq also provide corresponding index files (*.bai) for users' convenience.

Differential Gene Expression Analysis (DGEA) Results from DESeq2:

Summary Report: DESeq2.html

This is a html page showing read count distribution, dispersion plot, experimental design, number of genes significant. It also shows the model design, code, and package versions.

Result tables:

{contrast_name}.deseq2.xlsx:

This is an Excel table containing LFC_raw, LFC_shrunken, FDR (padj), and TPM expression value for all genes. Please note that LFC_shrunken is recommended over LFC_raw for ranking and visualization, as it assigns lower LFC_shrunken for noisy low count genes. Please see DESeq2 vignette for more details: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

{contrast_name}.sig.FDR.*.LFC.*.xlsx:
This is an Excel table containing only the significantly differentially expressed genes with FDR < MAX_FDR, and LFC > MIN_LFC.

TPM.xlsx:
This is an Excel table containing the TPM (Transcripts Per Kilobase Million) normalized expression values. We recommend users use TPM for between sample/gene comparisons. To find how it's difference from FPKM/RPKM, please refer to https://rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

FPKM.xlsx:
This is an Excel table containing the FPKM (Fragments Per Kilobase Million) normalized expression values. When single-end RNAseq data was analyzed, it contains the RPKM (Reads Per Kilobase Million) values. FPKM/RPKM accounts for normalization of sequencing depth and gene length. However, it can't be used for between sample comparisons. It is provided for users to make within sample comparisons. For more information, please refer to https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

DESeq2NormalizedCounts.xlsx:
This is an Excel table containing the DESeq2 normalized expression values used for differential gene expression analysis.

COUNTS.xlsx:
This is an Excel table containing the raw read count for each sample and each gene. This is for users that need to perform further count based analysis by themselves.

Plots:
PCA plot: sample_PCA.labeled.pdf/sample_PCA.pdf
PCA plots show the clustering of samples in a 2D space of PC1 and PC2. This is very helpful for finding sample outliers and spot potential problem with sample preparation and sequencing. It is also very helpful in showing the clustering of samples and spotting batch effects. Please note that x-axis (PC1) and y-axis (PC2) are of different scale. The PCA was performed on variance stabilizing transformation (VST) processed expression values.

Heatmap: sample_poisson_distance.pdf
A heatmap shows the overall similarity between samples. This plot is based on the Poisson distance calculated using the PoissonDistance function, implemented in the PoiCalClu package (Written, 2011) with the original count matrix. This plot helps users to assess which sample are similar to each other and whether the sample clustering fits the expection.

Volcano-plots:

{contrast_name}.LFC_shrunken.pdf:
A volcano-plot shows the significant genes in red (if there are any) with LFC_shrunken in x-axis and -log10(FDR) in y-axis. This plot is the recommended volcano-plot over the one below which uses LFC_raw in x-axis.

{contrast_name}.LFC_raw.pdf:
A volcano-plot shows LFC_raw in x-axis and and -log10(FDR) in y-axis. This plot is for users who are interested in visualizing the un-shrunken noise associated with lowly expressed genes. When lots of genes of interest are lowly expressed, please consider generating libraries with more input RNA amount and more sequencing depth.

Z-score heatmaps:

{contast_name}.heatmap.pdf:
A Z-score heatmap shows the expression pattern of significantly differentially expressed genes with colors representing Z-scores standardized among samples within each gene. For example, the sample with the lowest expression value among all samples for a gene will have the lowest negative Z-score for the corresponding gene, the sample with the mean expression value will have the Z-score of zero.

{contast_name}.heatmap.v2.pdf:
This plot is similar to the above heatmap except that samples are also hierarchically clustered.

{contast_name}.heatmap.gene_class.xlsx:
This is an Excel table listing the genes in each class (class I, class II) clustered in the above Z-score heatmap.

Gene Set Enrichment Analysis Results from GSEA

Users can browse GSEA analysis results by drilling down into individual subfolders named as the contrast name (groupA_vs_groupB), then reviewing the index.html file in each subfolder named as the corresponding gene set (e.g. c1-c7, h).

A detailed description of gene sets collected by MSigDB can be found at https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp

Please refer the GSEA Report section at https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html on how to interprete the GSEA analysis results.

Differential Exon Usage Analysis Results from DEXSeq

Count files:

These txt format files are read counts for each exon. They are intermediate files for DEXSeq.

Additional output:

DEXSeq_{contrast_name}.xlsx:

An Excel table containing the gene symbol, FDR (padj), log2 fold change of all exons. The log2 fold change shows the difference in exon usage due to the condition, i.e. GROUP_ID, changes. It is the difference between the expression trend in this exon and the overall expression trend of the gene. When no batch effect is present, the model is ~ sample + exon + condition:exon, and condition:exon was tested. When batch effect is present, the full model is ~sample + exon + batch:exon + condition:exon, the reduced model is ~ sample + exon + batch:exon.

DEXSeq_{contrast_name}.sig.xlsx:

An Excel table containing only the significant exons.

{contrast_name}.{gene_symbol}_top_{i}_normalized_counts.pdf:

A DEXSeq plot showing the normalized count for each sample in the context of all transcript isoforms annotated.

{contrast_name}.{gene_symbol}_top_{i}_relative_exon_usage.pdf:

A DEXSeq plot showing the fitted expression value for each group.

Alternative Splicing Analysis Results from rMATS

Please refer to rMATS official documentation for a detailed explanation:
https://github.com/Xinglab/rmats-turbo/blob/v4.1.0/README.md
SE: skipped exon
MXE: mutually exclusive exons
A3SS: alternative 3’ splice sites
A5SS: alternative 5’ splice sites
RI: retained intron

Results_JunctionCountsAndExonCountsBased:
Final output including both reads that span junctions defined by rmats (Junction Counts) and reads that do not cross an exon boundary (Exon Counts)

Results_JunctionCountsBased:
Final output including only reads that span junctions defined by rmats (Junction Counts)

fromGTF:
All identified alternative splicing (AS) events derived from GTF and RNA