GLADSTONE BIOINFORMATICS CORE

Overview of results provided for an RNA-seq analysis
Report of differentially expressed genes

The most detailed result of any differential expression analysis is a table of gene-level (or other feature, e.g. isoforms or exons) counts, which reports: 1) Differential expression P-value, 2) False-discovery-rate (FDR) adjusted P-value, 3) Fold change, 4) Raw counts for each feature, 5) Normalized counts for each feature, in CPM or FPKM as applicable.
Dendrogram of samples clustered by gene expression (log2 FPKM)

Here, we have 12 samples—three groups of four, shown as color-coded groups on the dendrogram. (Each endpoint of the dendrogram is a single sample). In this specific case, each of the 3 larger sample groups is divided into two sub-groups based on a specific treatment type.

This figure can be useful for finding outliers with technical problems or for finding samples that cluster together in unusual ways; for example, it would allow us to detect whether samples were grouped by the date each sample was collected, rather than the treatment condition.

“GapMap” overview heatmap of the same data from the dendrogram

(Gaps sizes are proportional to branch length on the dendrogram.)
Assessment of gene body coverage

In this example, the technology involved sequences almost exclusively at the 3' end. This is a good example of highly-concordant replicates; no obvious problems are visible here.

Summary QC metrics generated using **FastQC**

- Quality by position
- Base calls by position
- GC Content
- Duplication rate
- Over-represented sequences by position in read
- Splicing analysis using **RSeQC**
Overview of a subset of genes

Heatmap of genes with high variance (or high fold change between samples)

A heatmap is often a straightforward way of getting an overview of change in gene expression (or potentially lack thereof) between samples. Clustering here is done by both genes (rows) and samples (columns).

Sometimes this can be used to identify particularly unexpected samples.
Cell / Tissue type predictions

Cell type / tissue type predictions for each sample, generated using *AltAnalyze*

Cell types are shown in a hierarchical tree view.

Specific marker genes are used to predict cell type / associated tissue type for each sample.

- Astrocytes
- Oligodendrocyte Progenitor
- Ovaries
- Pituitary Gland
- Astrocyte Progenitor
- Oligodendrocytes
- Placenta Day 9.5
- Prostate
- Spinal Cord
- Macrophages
- Skeletal Muscle
- Urinary Bladder
- Lungs
- Mammary Glands
- Pluripotent Stem Cells
- Umbilical Cord
- Brown Adipose
- Embryo Day 11.5
- Embryo
- Embryo Day 9.5
- Heart
- Adrenal Gland
RNA-seq pipeline

The current RNA-seq pipeline:

1. **FASTQ**
   - Raw input files

2. **FILTERING**
   - FASTQ-MCF (optional)

3. **ALIGNMENT**
   - GTF (gene annotation)
   - FASTA (protein data)
   - Choose:
     - TOPHAT
     - BOWTIE
     - STAR

4. **BAM**
   - Alignment

5. **DUPILCUTES**
   - PCR removal
   - Duplicate

6. **COUNTING**
   - SAM/BAM Feature Counts

7. **NORMALIZATION**
   - Pos. Controls
   - RUVSeq (R)

8. **DIFFERENTIAL GEN EXPRESSION**
   - Choose:
     - Edger (R)
     - DESeq2 (R)
     - CuffDiff

9. **SUMMARIZATION**
   - PCA Plot
   - Dendograms
   - Lineage Profiling (AltAnalyze)
   - Sample Comparisons

10. **VISUALIZATION OF INTERESTING GENES**
    - Heatmap / MA Plots (R)
    - Pathways (AltAnalyze)

**KEY TO SYMBOLS**

- **INPUT**: This is a generic input file, e.g., FASTQ or bam
- **SEQUENCE PROGRAM**: Free or open-source programs
- **SOFTWARE PROGRAM**: Experiment-specific programs
- **DATA**: Access to raw data from a user experiment

**DIFFERENTIAL GEN EXPRESSION TOOLS**

- DESeq1 (R)
- EdgeR (R)
- Subread
- Subread

**FILTERING TOOLS**

- MA Plots (R)
- DUPLICATES
- DESeq1 (R)
- EdgeR (R)
- Remove

**QUALITY CONTROL TOOLS**

- FASTQ-MCF (optional)
- MA Plots (R)
- Quality statistics

**SUMMARIZATION TOOLS**

- Edit plots
- DeSeq2 plots
- CuffDiff plots
- AltAnalyze

**VISUALIZATION TOOLS**

- Heatmap
- MA plot
- Pathways

**DATA SOURCES**

- Reference genome
- Gene annotation
- Quality control
- Processed data
Sample correlations

Correlation between all pairs of samples

Here we see that within-group correlations are extremely high (Pearson’s R reported in each cell in the lower-left), and between-group correlations show a substantial number of differentially expressed genes (the off-diagonal points). Each point is one gene. Scatterplot axes are log2(FPKM + 1).
PCA Plots

Percent variance explained by each component

PCA plots can show unexpected sources of variance / technical error. Normally, we expect samples to visually cluster by an experimental condition (or biological classification).

PCA plots for various components

Right: typically observed results if sample are primarily distinguished by some underlying biological variability.

Below: a case that may warrant further investigation of the outlying samples, unless this is expected based on the nature of the experiment.
By default, we generate and temporarily host your aligned reads as tracks that can be directly viewed in any web browser. Aligned reads can also be viewed locally on your own computer using IGV from the Broad Institute.

We can provide you with all necessary files (BAM files and BigWig tracks) for re-hosting on your own web space.

Strand-specific reads can be viewed with their original orientation.

An example of "wiggle" tracks viewed on the UCSC Genome Browser.
Custom Figures / downstream analysis

These are not part of our standard analysis, but should give some idea of additional types of figures and downstream analysis that we can perform.

Differential expression overlap: 28,815 transcript clusters in total. Numbers below are the number of genes that were differentially expressed at a cutoff of $P \leq 0.05$ in the given comparison.

Pathways from WikiPathways, with graph nodes colored by user-specifiable parameters.

Custom analysis results.

Gene networks visualized using Cytoscape.

Additional examples of custom figures.

Venn diagrams showing gene overlap between conditions.