RNASeq (bulk) data analysis

Updated 10/08/2019

Please email us at EICC@emory.edu with any questions or comments regarding data analysis.
For sequencing data acquisition please contact Emory Integrated Genomics Core(EIGC@emory.edu).

1. Description of Steps

All RNA-Seq projects start with the following steps:

Quality control and alignment of FASTQ files

Samples are sequenced on an Illumina NovaSeq6000 system. Sequences are quality-checked using FastQC for completeness, depth and read quality. Sequences are aligned to the HG38 reference genome using STAR aligner[1]. Gene quantification is done using HTSeq-count[2]. If you are specifically looking for long noncoding RNAs (lncRNAs), we use annotations from GENCODE lncRNAs in alignment (STAR) and gene quantification (HTSeq-count) steps[3].

Filtering and normalization

Genes with low counts are removed in packages which rely on the Negative Binomial (NB) model like edgeR and DESeq2 though each package differs in how they filter and normalize results. Filtering lowly expressed genes removes statistical background noise, improves the algorithm's ability to determine differential expression which provides increased statistical power. Normalization is similar to averaging out the data so that if for some reason one sample (or a few samples) was sequenced more deeply than another, all samples will be comparable to one another.

Differential Expression analysis

All of our RNA-Seq differential expression analysis is performed in R. However, the specific package we use for any given project will depend on the goals of your project. Results are expected to be consistent across all packages but because they all use slightly different algorithms, one package may be better if you are doing exploratory analysis where you want to find as many genes as possible without concern for the possibility of false positives and another may be better if you have an idea of what you are expecting to see and want to be stricter about what you count as significant.

We are happy to work with you to determine the best package to use to meet your experimental goals. As of this writing we are familiar with edgeR, DESeq2, and baySeq and we are open to learning new packages as a project may require it. All packages generally follow these analytical steps and we will be sure to provide you a description of the specific steps we use for your data throughout the process. **For detailed walkthroughs of the code used for differential expression analysis, please see our GitHub page:**
https://github.com/EmoryIntegratedComputationalCore/Methods

Packages available for RNA-Seq analysis differ from one another primarily in how they call genes "significantly differentially expressed". Even within packages there are often several options available. No matter what package or method we use, the goals of your project will determine which method is most suitable and we will also be clear about which method we used to determine differential expression along with our reasoning for the choice, any relevant assumptions, and any limitations of the method. Briefly, each package determines differential expression in the following ways:

baySeq

baySeq uses empirical Bayesian inference to determine the likelihood that genes in compared samples are differentially expressed. baySeq seeks to improve accuracy in DE estimation over other popular packages by using the underlying structure of the data itself rather than assuming a particular distribution. baySeq normalizes the results by scaling them by library size as part of the analysis. baySeq does show improved performance in the case of more complex study designs (i.e. multiple group comparisons) and in studies with large numbers of libraries compared to other popular packages[4].

DESeq2

DESeq2 assumes the NB model which mathematically accounts for the fact that we are assessing gene counts and we are assuming that most genes we are comparing between the groups will not be differentially expressed [5]. DESeq2 offers users the option to filter raw counts and the need for this will vary by project. DESeq2 primarily uses a process called independent filtering on normalized counts as part of the analysis. As part of analysis, a process called independent filtering removes lowly expressed genes based on the mean of normalized gene counts as part of the analysis process. This provides the highest number of genes with raw p-values below the specified adjusted p/FDR cut-off to improve the ability to detect DE genes if they are present. Finally, DESeq2 uses the Wald test for significance which tests the hypothesis that there are no differences between the groups. This test allows for simple pairwise comparisons and more complex comparisons which account for covariates like sex, age, ethnicity, or other variables of interest. This method is best for exploratory projects where you do not have specific genes in mind you are looking to find.

edgeR

Similarly, to DESeq2, edgeR also uses the NB distribution to determine the likelihood of differential expression between two or more groups to account for the fact that we are assessing gene counts and we are assuming that most genes we are comparing between the groups will not be differentially expressed [6]. Count data from HTSeq-count will be transformed to CPM and filtered before analysis to remove lowly expressed genes. Next, we normalize the data to account for differences in library sizes between samples with a method called TMM normalization. edgeR filters lowly expressed genes using counts-per-million (CPM) in order to account for differences in library size between samples. As part of the analysis, edgeR converts raw counts to log transformed CPM with no user input required. For graphing heat maps however, the user must first transform counts into CPM and log transform them to produce meaningful results. edgeR's recommended method for differential expression analysis is the quasi-likelihood F-test. The simple null hypothesis for this test is that there is no difference between the control and the treatment samples. This test also allows for the incorporation of variables like sex, age, ethnicity, or other variables of interest into analysis. This method is best for projects where you have a list of genes in mind that you are expecting to find.

Gene ontology, pathway, and set analyses

baySeq and DESeq2 do not offer direct methods for pathway analysis but provided there are significantly differentially expressed genes between the experimental groups of interest, edgeR uses either gene ontology (GO) or KEGG pathway enrichment analysis to determine which GO pathways are most strongly or most weakly

represented in the results. Both types of analyses use the NCBI RefSeq annotation and require the Entrez Gene Identifier (ID) for each gene. edgeR also uses Gene Set Testing which tests the hypothesis that the majority of genes identified as DE are indeed differentially expressed across a comparison of interest such as sex, ethnicity or any other covariate[6].

2. Example Methods Sections

Below are examples of Methods sections for manuscripts using these methodologies

Sequencing, alignment, and quantification

Samples are sequenced on an Illumina NovaSeq6000 system. Sequences are quality-checked using FastQC for completeness, depth and read quality. Sequences are aligned to the HG38 reference genome using STAR aligner[1]. Gene quantification is done using HTSeq-count[2].

Differential Expression

DESeq2, edgeR and baySeq determine differentially expressed genes between at least two experimental groups [3–5]. Genes with low counts are filtered by mean normalized counts in DESeq2 and by expression of the counts per million (CPM) in edgeR. Raw p-values generated by both packages are transformed using the Benjamini-Hochberg correction to account for multiple hypothesis testing. Genes considered significantly differentially expressed are those with adjusted p-values, also called False Discovery Rates (or FDR) which is controlled at a specific threshold. baySeq uses the Bayesian empirical approach to estimate a posteriori probability of each set of models, which defines differential expression patterns for each tuple. Genes considered significantly differentially expressed are those with high posterior likelihoods and adjusted p-values, also called False Discovery Rates (or FDR) which fall below a specific threshold. Graphs are generated using package-specific recommended data transformations.

The choice of differential expression analysis tool, edgeR, DESeq2, baySeq, or any other package will be project-specific and we will work with you to choose the best tool for your project.

Gene Set Enrichment Analysis (GSEA)

To characterize differentially expressed genes, a rotation based gene set enrichment analysis (GSEA) will be performed. A gene set from the Molecular Signatures Database will be used as input. GSEA generates enriched pathways ordered by enrichment score (ES) of the genes based on a weighted Kolmogorov–Smirnov (K–S) statistical test, which measures the difference between the number of genes in a given dataset and the number of occurrences of genes observed in a pathway. Raw p-values generated by this test are transformed using the Benjamini-Hochberg correction to account for multiple hypothesis testing. Gene sets considered to be enriched are those with adjusted p-values, also called False Discovery Rates (or FDR) which fall below a specific threshold [6].

Gene ontology (GO) and pathway analysis

Alternatively, the identified differentially expressed genes will be used in pathway analysis in edgeR. Given a list of differentially expressed genes, edgeR uses the Entrez Gene Identifier (ID) and species to perform a linear model analysis and return a list of enriched pathways. P-values generated by this test are evaluated for over-representation. Pathways considered to be most enriched are those which meet the criteria of either the Fischer's exact test or the Wallenius' noncentral hypergeometric distribution [4].

3. Required Files:

1. Raw data files (FASTQ)
2. Spreadsheet with data about samples (which are cases/controls, sex, age, ethnicity, or any other covariate of interest etc.)
3. Organism (Human/Mouse/Rat etc..)

4. Deliverables:

1. A power point presentation including all agreed upon analyses and results with interpretations
2. Excel spreadsheets of raw counts and complete differential expression analysis, pathway analysis, and gene set enrichment analysis results as agreed upon
3. Image files of selection of agreed upon number of graphs/plots with interpretations
   -gene specific box plots, volcano plots, PCA plots, MA plots, heat maps etc.

5. References

1. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635
2. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btu638
3. Hardcastle, T. J. & Kelly, K. A. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-422
4. Robinson, M., McCarthy, D. & Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, (2010).
5. Love, M. I., Anders, S. & Huber, W. *Differential analysis of count data - the DESeq2 package*. *Genome Biology* (2014). doi:110.1186/s13059-014-0550-8
6. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0506580102