Amazon Web Services (AWS) Machines: Linux R

All of the following was done for exploring/learning Bioinformatics in AWS Cloud.

AMI Image Instance chosen from many options   
[A] (Ubuntu Linux, t2.xlarge, 4 vCPUs 2.3 GHz, 16 GB Memory, EBS Volume SSD 40 GB)  
Config Security Groups, Key Pairs, IAM, SSH connect to AMI Instances

In each of the AWS provisioning scenarios; successful setup, configuration Enables   
- RStudio (Local PC Web Browser) <== HTTP ==> RStudio Server (EC2 AMI)  
- Bioinformatics workflows using Bioconductor Packages.

EC2 AMI:   
 <https://www.bioconductor.org/help/bioconductor-cloud-ami/#ami_ids>   
 Bioconductor Community maintained Public AMI image  
 AMI Image: [A]   
 Includes: Bioconductor distribution (R Packages, RStudio Server, Git)

EC2 Image Builder:   
 Service enables custom image builds  
 Bioinformatics Developers choose build components.  
 AMI Image: [A]   
 Includes: RStudio Server, Bioconductor Packages, Git;   
 And other components of choice; e.g., Tensorflow,

EC2 Container Service (ECS):   
 <https://docs.aws.amazon.com/AmazonECS/latest/developerguide/launch_container_instance.html>   
 Service supports Docker containers  
 Bioconductor Community maintains Docker image builds.  
 AMI Image: [A]   
 Public AMI ECS-Optimized; with Docker Engine, ECS container agent, etc.

Docker Image1: RStudio Server, Bioconductor Packages, Git;   
 <https://hub.docker.com/r/bioconductor/bioconductor_docker>   
 Docker ImageN: And other Docker images of choice; e.g., Tensorflow,   
 Enables: Docker on AMI <== Docker Pull/Push ==> DockerHub (Bioconductor.org image)

Elastic Container Registry (ECR):   
 Docker container registry service  
 Docker container images build components  
 Public Linux repositories 823 (12/14/2020)  
 Airflow, Tensorflow, Spark,   
   
With particular interest in   
Next Generation Sequencing (NGS) RNA-Seq Differential Expression

Statistical Analysis, Machine Learning

Exploratory Data Analysis

Data Mining

[GitHub Pages] (<https://github.com/william-p-kahley/williamkahley.github.io/William.P.Kahley.CovLet.121420.htm>)

**[S1]:** **Special effort;**

NGS RNA-Seq bioinformatics tools, workflows

"differential expression"  
"exploratory data analysis"

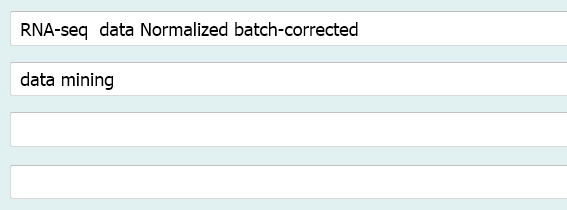
I am especially attracted to the aspects of workflows following the pre-processing, normalization (Quantification of Mapped Reads, Differential Gene Expression).

The Application of Inferential Statistics, Machine Learning (PCA, Feature/Dimensional Reduction, Clustering).

**Google Search**

[**enabling cross study analysis of rna-sequencing data**](https://www.google.com/searchlr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ)

<https://www.google.com/searchlr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ>

  
<https://www.google.com/search?as_q=RNA-seq++data+Normalized+batch-corrected+&as_epq=data+mining+&as_oq=&as_eq=&as_nlo=&as_nhi=&lr=&cr=&as_qdr=all&as_sitesearch=&as_occt=any&safe=images&as_filetype=&tbs=>

**Keywords:**

RNA-seq

data mining

cross study analysis

data normalization

batch-corrected RNA-seq

|  |
| --- |
| **[S1]:** **Special study of;   NGS RNA-Seq: Bioinformatics tools, Workflows** |
| **Relevant journal publications**   |  | | --- | | **RNA-seq workflow: gene-level exploratory analysis and differential expression**  2015 Oct 14  <http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html> | | Abstract  Here we walk through an end-to-end gene-level RNA-Seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were aligned to the reference genome, and prepare a count matrix which tallies the number of RNA-seq reads/fragments within each gene for each sample. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.  Keywords  RNA-seq, differential expression, gene expression, Bioconductor, statistical analysis, high-throughput sequencing, visualization, genomics | |  |  |  | | --- | | **RNA-seq: Basic Bioinformatics Analysis**  Curr Protoc Mol Biol. 2018 Oct  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6168365/> | | The workflow includes three parts:  (a) mapping sequencing reads to a reference genome or transcriptome;  (b) quantifying expression levels of individual genes and transcripts;  (c) identifying specific genes and transcripts that are differentially expressed between samples.  Alignment/Mapping:  the sequence of each read to a reference genome, annotation of genes  - **STAR** alignment tool  Quality Assessment:  After mapping reads to the genome, it is important to survey the quality of the RNA-seq data  - **Picard**  - **SAMTools**  Quantification of Mapped Reads:  - **HTseq** to quantify sequencing reads mapped to each gene  (a) identify genes that are differentially expressed between conditions (sample groups),  (b) derive gene expression values for each individual transcript  Approaches for normalization include  CPM (counts per million reads),  RPKM (reads per kilobase per million reads),  FPKM (fragments per kilobase per million reads),  TPM (transcripts per million reads).  Differential Gene Expression Analysis:  **EdgeR** (Robinson et al., 2010)  **DESeq2** (Love et al. 2014)  Download and install required tools:  STAR: <https://github.com/alexdobin/STAR>  Picard: <https://broadinstitute.github.io/picard/>  HTseq: <https://htseq.readthedocs.io/en/release_0.9.1/install.html>  R: <https://www.r-project.org> |  |  | | --- | | **ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases**  BMC Genomics 2014 volume 15, Article number: 284  415 Citations  <https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-284> | |  |  |  | | --- | | **pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components**  BMC Bioinformatics 2019 volume 20, Article number: 331  17 Citations  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2879-1> | | **pcaExplorer**  <http://bioconductor.org/packages/release/bioc/html/pcaExplorer.html>  **pcaExplorer User Guide**  27 October 2020  <http://bioconductor.org/packages/release/bioc/vignettes/pcaExplorer/inst/doc/pcaExplorer.html>  **Up and running with pcaExplorer**  27 October 2020  <http://bioconductor.org/packages/release/bioc/vignettes/pcaExplorer/inst/doc/upandrunning.html>  data component (count matrix, experimental data, dds object, annotation) |      |  | | --- | | **A Beginner’s Guide to Analysis of RNA Sequencing Data**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6096346/>  Am J Respir Cell Mol Biol. 2018 Aug | | **Abstract:**  A general understanding of the principles underlying each step of RNA-seq data analysis allows investigators without a background in programming and bioinformatics to critically analyze their own datasets as well as published data. Our goals in the present review are to break down the steps of a typical RNA-seq analysis and to highlight the pitfalls and checkpoints along the way that are vital for bench scientists and biomedical researchers performing experiments that use RNA-seq.  **Keywords:**  RNA sequencing, transcriptomics, bioinformatics, data analysis  **Associated Data:**  The RNA-seq data reported in this article has been deposited in NCBI’s Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE116583.  **Experimental Design and Approach:**  A major goal of RNA-seq analysis is to identify differentially expressed and coregulated genes and to infer biological meaning for further studies. Source material can be cells cultured in vitro, whole-tissue homogenates, or sorted cells. The ability to interpret findings depends on appropriate experimental design, implementation of controls, and correct analysis. Every effort should be made to minimize batch effect, because small and uncontrolled changes in an environment can result in identification of differentially expressed genes (DEGs) unrelated to the designed experiment. Sources of batch effect can occur during the experiment, during the RNA library preparation, or during the sequencing run and include but are not limited to those listed in Table 1. Once a well-designed and controlled experiment is performed, a structured approach to the dataset allows for quality control followed by unbiased analysis of the data. In the present analysis, we use an approach that includes setting low count filtering, establishing a noise threshold, checking for potential outliers, running appropriate statistical tests to identify DEGs, clustering of genes by expression pattern, and testing for gene ontology (GO) enrichment. For each of these analysis components, we aim to highlight important checkpoints and quality controls that will streamline and strengthen data analysis, avoid bias, and allow investigators to maximally use their datasets.  **Clustering**  The two most common clustering methods used for RNA-seq data analysis are hierarchical and k-means clustering (see Clustering box). The most common form of hierarchical clustering is a bottom-up agglomerative approach that organizes the data into a tree structure without user input by starting with each data point as its own cluster and iteratively combining them into larger clusters or “clades.” In contrast, k-means clustering requires the investigator to define the number of clusters (k) a priori, and data are then sorted into the cluster with the nearest mean. It is possible to assess a range of k-values to decide how to best capture the trends. In addition, various tools such as Silhouette exist to help the investigator determine the ideal k-value, but some subjectivity remains (21). By adjusting the k, the investigator may set the degree of granularity they would like to achieve with the data. For either approach, the user must specify the distance metric by which data points are considered similar. Typically, Pearson’s correlation is used, and this is generally the default in software designed for RNA-seq analysis. Both approaches are widely used, and both aid the investigator in identifying groups of genes that display similar expression patterns, allowing for further downstream analyses. The clusters can then be used as input for an analysis of functional enrichment (see next section).  Why do we use clustering on RNA-seq data?  Clustering of RNA-seq data may be used to identify patterns of gene expression by grouping genes based on their distance in an unsupervised manner. Clustering RNA-seq data is used as an exploratory tool that allows the user to organize and visualize relationships between groups of genes, and to select certain genes for further consideration.  **Hierarchical clustering**  The most commonly used hierarchical clustering approach is a form of agglomerative, or bottom-up, clustering that iteratively merges clusters (originally consisting of individual data points) into larger clusters or “clades”.  **K-means clustering**  Data points are iteratively partitioned into clusters based on the minimum distance to the cluster mean. The number of clusters (k) is set by the investigator. |      |  | | --- | | **Review of RNA-Seq Data Analysis Tools**  February 17, 2016  <https://rna-seqblog.com/review-of-rna-seq-data-analysis-tools/> |      |  | | --- | | **Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function**  BMC Bioinformatics volume 17, Article number: 459 (2016)  7 Citations  <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1349-2> |      |  | | --- | | **GEO2RNAseq: An easy-to-use R pipeline for complete pre-processing of RNA-seq data**  September 16, 2019 bioRxiv Preprint  <https://www.biorxiv.org/content/10.1101/771063v1.full> |      |  | | --- | | **RNAseq data analysis in R - Notebook**  <http://monashbioinformaticsplatform.github.io/RNAseq-DE-analysis-with-R/RNAseq_DE_analysis_with_R.html> | | Install and load packages  Mapping reads to a reference genome  Count reads for each feature  QC and stats  Differential Expression  Gene Annotation  Gene Set Enrichment |      |  | | --- | | **Introduction to differential gene expression analysis using RNA-seq**  September 2015 updated November 14, 2019  <https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf> | |

|  |
| --- |
| **[S1]:** **Special effort;   NGS RNA-Seq: Statistical Methods, ML, Data Mining** |
| **Journal Publications:**   |  | | --- | |  | |  | |  |      |  | | --- | | **Google**  **enabling cross-study analysis of RNA-Sequencing data**  [https://www.google.com/search?lr=&as\_qdr=all&sxsrf=AleKk00QSjfjp\_4g\_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ](https://www.google.com/search?lr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ)  **Google**  **RNA-Seq data mining** | | **A survey of best practices for RNA-seq data analysis**  Genome Biology volume 17, Article number: 13 (2016)  769 Citations  <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8>  **Massive mining of publicly available RNA-seq data from human and mouse**  Nature Communications volume 9, Article number: 1366 (2018)  96 Citations  <https://www.nature.com/articles/s41467-018-03751-6>  **Omics Playground: Explore Omics Data Freely**  <https://omicsplayground.readthedocs.io/en/latest/index.html>  Omics Playground is a comprehensive self-service platform platform for visualization, analytics and exploration of Big Omics Data. It allows users to apply a multitude of state-of-the-art analysis tools to their own data to explore and discover underlying biology in a short time.  The platform offers a unique combination of features that distinguishes it from the other analytics platforms currently available. We believe that data preprocessing (primary analysis) and statistical testing (secondary analysis) are now well established, and the most challenging task is currently data interpretation (tertiary analysis) that often takes the longest time but where actual insights can be gained. Therefore, Omics Playground focuses strongly on tertiary analysis while providing good support for secondary analysis.  Reanalyzing Public Datasets  To illustrate the use case of the Omics Playground, we reanalyzed different types of publics datasets, including microarray, bulk RNA-seq, single-cell RNA-seq and proteomic datasets to recapitulate the results.  <https://omicsplayground.readthedocs.io/en/latest/examples/examples.html> |  |  | | --- | | **comparing rna-seq datasets** | | **Broad Institute Gene Set Enrichment Analysis (GSEA)**  <https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Using_RNA-seq_Datasets_with_GSEA>  **BioStars**  **Question: RNA-seq data comparison across experiments**  <https://www.biostars.org/p/117451/>  **BioStars**  **Question: Comparing similarity of RNAseq datasets**  <https://www.biostars.org/p/334235/>  **Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations**  Scientific Reports volume 8, Article number: 11226 (2018) Cite this article  7 Citations  <https://www.nature.com/articles/s41598-018-29506-3> |   **Books:**  **NCBI Books Advanced Search**  <https://www.ncbi.nlm.nih.gov/books/advanced/>   |  | | --- | | **Computational Biology**  Brisbane (AU): Codon Publications; 2019 Nov 21.  ISBN-13: 978-0-9944381-9-5  <https://www.ncbi.nlm.nih.gov/books/NBK550339/>  Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)  <https://creativecommons.org/licenses/by-nc/4.0/> | | Chapter 4 Biological Sequence Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550342/>  Chapter 5 Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550343/>  Chapter 6 Statistical Methods for RNA Sequencing Data Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550334/> | |

|  |
| --- |
| **[K5]:** **Knowledge of Publicly Available Genomic Data Resources, Viewers** |
| **Genome Data Viewer (GDV) <--> GEO <--> Integrative Genomics Viewer (IGV)**  NGS data deposited in the GEO database can be visualized through the genome data viewer function  To check the quality of raw sequence data in the FASTQ format,  NGS data (FASTQ files)   |  | | --- | |  | | **RNA-seq workflow: gene-level exploratory analysis and differential expression**  <http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html> | |