**Data Analysis, Modeling, Visualization (Software Tools Overview)**

|  |
| --- |
| **Categorized Bioinformatics Software Tools** <https://bioinformaticshome.com/tools/tools-main.html> |
| |  | | --- | | [**RNA-seq Pre-analysis Tools**](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html) <https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html>  Pre-analysis quality control of raw reads includes assessment of tolerable GC and k-mer contents, removal of sequence adaptors, PCR artifacts, and contaminations. The assessment of duplicates and sequencing errors. In addition, sequencing quality tends to decrease towards the 3' end of the reads; Thus, the reads must be trimmed to remove the low-quality ends. [Data Quality Assessment](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html)   * [Filtering](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Filtering) * [Trimming](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Trimming) * [Filtering and Trimming](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Filtering-and-Trimming) * [Reporting/Visualization](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Reporting-Visualization) * [Other Pre-analysis RNA-seq Tools](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Other) | | **FASTX-Toolkit** <https://bioinformaticshome.com/tools/rna-seq/descriptions/FASTX-Toolkit.html> **FastQC** <https://bioinformaticshome.com/tools/rna-seq/descriptions/FastQC.html> |  |  | | --- | | **RNA-seq Core Analysis Tools**  <https://bioinformaticshome.com/tools/rna-seq/core-analysis.html> | | **1. Transcriptome Profiling**  1.1 Read mapping or assembly  1.1.1 De novo (reference free) transcriptome assembly  1.1.1.1 Unstranded  1.1.1.2 Stranded  1.1.1.3 Quality Control  1.1.2 Mapping to a reference genome or transcriptome  1.1.2.1 Splice Aware  1.1.2.2 Splice unaware  1.1.2.3 Quality Control  1.2 Expression Quantification  1.2.1 Union-exon Based  1.2.2 Transcript Based  1.2.3 Bacterial genome  **2. Differential Expression Analysis**  2.1 Pre-processing DEA  2.2 Parametric  2.3 Non-parametric  2.4 Power analysis  **3. Functional Profiling**  3.1 Enrichment Analysis (GSEA), annotation, other  3.2 Comparison with Genome |  |  | | --- | | **Whole Genome Assembly (WGA) Analysis Tools** - Software and Resources <https://bioinformaticshome.com/tools/wga/wga.html> | | GAML  GAML is a tool for genome assembly based on maximum likelihood. It implements a probabilistic model to take into account sequencing error rates, insert lengths and other characteristics to produce a final genome assembly. This tool can work on sequenced data generated from multiple sequencing platforms (e.g. Illumina, 454, PacBio).  Operation: Genome assembly  Software interface: Command-line user interface  Language: -  Operating system: Linux  License: Not stated  Cost: Free  <https://bioinformaticshome.com/tools/wga/descriptions/GAML.html>  GAML: genome assembly by maximum likelihood  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4454275/> | |

|  |
| --- |
| **Google Searches** |
| **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> | |

|  |
| --- |
| **Genomic Data Viewers**  (navigation, search, content interpretation, data accession) |
| |  | | --- | | **NCBI Genome Data Viewer (GDV)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.38>  The NCBI Genome Data Viewer (GDV) is a genome browser supporting the exploration and analysis of eukaryotic RefSeq genome assemblies. Genome Data Viewer is also used by different NCBI resources, such as GEO and dbGaP, to display datasets associated with specified experiments or samples in a genome browser context. | | **Tutorials**  **NCBI Genome Data Viewer (Tutorial Page Functionality)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/#LAYOUT>  **NCBI Genome Data Viewer (Tutorial 11 videos Last updated Sep 23, 2020)**  <https://www.youtube.com/playlist?list=PLH-TjWpFfWruHgL0WRzZfQwp-MWzhIj16> |  |  | | --- | | **NCBI Gene Expression Omnibus (GEO)** <https://www.ncbi.nlm.nih.gov/geo/>  GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. | | **GEO DataSets (GDS)**  <https://www.ncbi.nlm.nih.gov/gds/> This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.  **About GEO DataSets** <https://www.ncbi.nlm.nih.gov/geo/info/datasets.html>  **GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geoprofiles/> This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics.  **Querying GEO DataSets and GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html> | | **Related Publications**  **The Gene Expression Omnibus database Methods Mol Biol. 2016; 1418: 93–110.**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944384/>  **NCBI GEO: mining millions of expression profiles database and tools Nucleic Acids Res. 2005 Jan 1**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539976/> |  |  | | --- | | **IGV (Integrative Genomics Viewer)**  <http://broadinstitute.org/software/igv/>  **IGV-Web application**  <https://igv.org/app> | | **IGV User Guide**  <http://broadinstitute.org/software/igv/UserGuide>  **Integrative Genomics Viewer**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346182/>  Video: <https://www.youtube.com/channel/UCb5W5WqauDOwubZHb-IA_rA>  Play All: <https://www.youtube.com/watch?v=sFeK25K5PE&list=PLSplvWwdPpSrhPn3V2iuPUzyxVIDYZ1xS> |  |  | | --- | | **Genome Analysis Toolkit (GATK)** Java framework  <https://gatk.broadinstitute.org/>  <https://gatk.broadinstitute.org/hc/en-us>  <https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1> A large Java library for variant analysis, discovery and genotyping,  powerful processing engine and high-performance computing features make it capable of taking on projects of any size. | | **GATK Best Practices Workflow for DNA-Seq**  <https://bioinformaticsworkbook.org/dataAnalysis/VariantCalling/gatk-dnaseq-best-practices-workflow.html#gsc.tab=0>  **GATK Getting Started**  <https://gatk.broadinstitute.org/hc/en-us/categories/360002302312>  Best Practices Workflows, Tutorials, Computing Platforms  **GATK Technical Documentation**  <https://gatk.broadinstitute.org/hc/en-us/categories/360002310591>  Troubleshooting, Glossary, Algorithms  **GATK Community Topics**  <https://gatk.broadinstitute.org/hc/en-us/community/topics>  Browse community discussions  **GATK / broadinstitute / gatk**  GitHub: <https://github.com/broadinstitute/gatk/releases>  Docker image: <https://hub.docker.com/r/broadinstitute/gatk/>  **Terra Support Quickstart New users overview**  Bioinformatics in the cloud on Terra  <https://support.terra.bio/hc/en-us/articles/360022714931-Bioinformatics-in-the-cloud-on-Terra> |   **NCBI Sequence Read Archive (SRA)**  <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?  **SRA Explorer**  tool aims to make datasets within the Sequence Read Archive more accessible.  <https://sra-explorer.info/> |

|  |
| --- |
| **1000 Genomes Project** |
| |  | | --- | | **1000 Genomes Project FAQ**  If you have any other questions you can’t find the answer to please email [info@1000genomes.org](mailto:info@1000genomes.org)  <https://www.internationalgenome.org/faq>  **IGSR: The International Genome Sample Resource**  Open human variation data  <https://www.internationalgenome.org/home>  **1000 Genomes 30x on GRCh38**  <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>  **1000 Genomes Browser (Phase 3)**  <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>  **Ensembl**  <http://uswest.ensembl.org/Homo_sapiens/Info/Index> |  |  | | --- | | **Data Portal (Search by Population, Technology, Data Collection)**  <https://www.internationalgenome.org/data-portal/sample>  Technology  Exome  Low coverage WGS  Integrated variant call sets  File Format  FASTQ  VCF | |