ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

RNA-seq analysis

Analysis in R

Alexey Sergushichev

2021-08-26, Tomsk

Overview of the day

- RNA-seq quantification – from raw data to an expression table
- (**RNA-seq analysis in R** – from an expression table to pathway analysis)
- Visual gene expression analysis in Phantasus

- Materials and slides are available at Google Drive
- Dockerfile and the scripts are available at https://github.com/ctlab/sysbio-training/tree/master/tomsk-scs-2021

Prepare

- Go to https://ctlab.itmo.ru/rstudio-sbNN/
- login: student
- password: sysbiopass

- Open the project from the previous module
- Open do_deseq2.R

Export expression values

- Run steps 0 & 1
- Export:
  - counts.txt
  - es.gct

ExpressionSet

- A single place to store both expression data and metadata
- exprs(es) – expression matrix
- pData(es) or phenoData(es) – sample metadata
- fData(es) or featureData(es) – gene metadata
- experimentData(es) – experiment metadata

Org.db packages

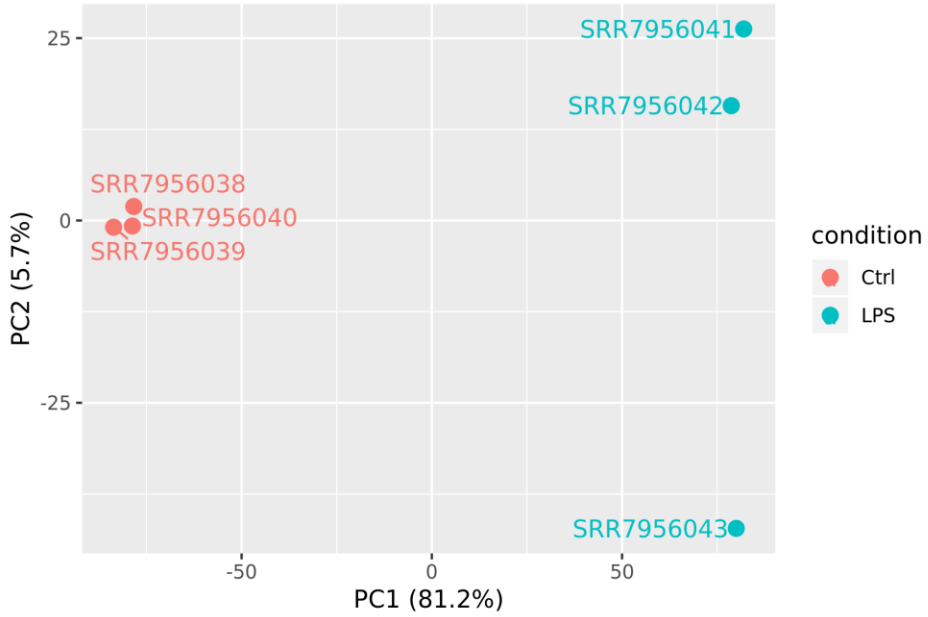
- Org.Mm.eg.db, Org.Hs.eg.db, …
- Contain gene annotation data for an organism
- Functions:
  - mapIds()
  - columns()
  - keys()
  - select()

Normalization

- Run step 2
- Possible normalization for RNA-seq:
  - log2 + quantile
  - divide by median expression
  - DESeq2::getVarianceStabilizedData
  - DESeq2::rlog
  - log2 + limma::voom
- The saved gct file can be opened in Phantasus

PCA plot

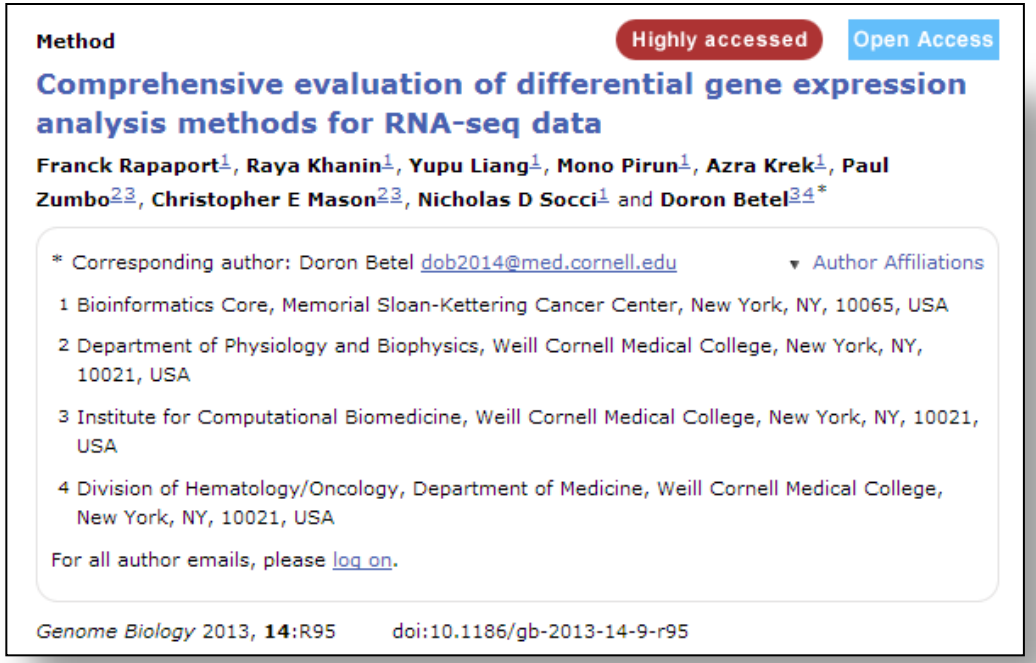
- Run step 3



Differential expression for RNA-seq

- **DESeq2**
- EdgeR
- limma+voom normalization
- kallisto/sleuth
- …

- Run step 4



Method

Highly accessed

Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], **Raya Khanin**[1], **Yupu Liang**[1], **Mono Pirun**[1], **Azra Krek**[1], **Paul Zumbo**[2,3], **Christopher E Mason**[2,3], **Nicholas D Socci**[1] and **Doron Betel**[3,4]*

* Corresponding author: Doron Betel dob2014@med.cornell.edu

▾ Author Affiliations

1 Bioinformatics Core, Memorial Sloan-Kettering Cancer Center, New York, NY, 10065, USA

2 Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, 10021, USA

3 Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, 10021, USA

4 Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, New York, NY, 10021, USA

For all author emails, please log on.

Genome Biology 2013, **14**:R95 doi:10.1186/gb-2013-14-9-r95

Pathway databases

- msigdbr
- reactome.db with fgsea::reactomePathways()
- a gmt file with fgsea::gmtPathways()
- KEGG pathways via KEGGREST
- Enrichr pathways http://amp.pharm.mssm.edu/Enrichr/#stats
- Gene Ontology via gage or Org.db packages

Pathway analysis

- fgsea
- DOSE/clusterProfiler:
  - fgsea-based and hypergeometric
- limma:
  - camera
  - roast
- gage

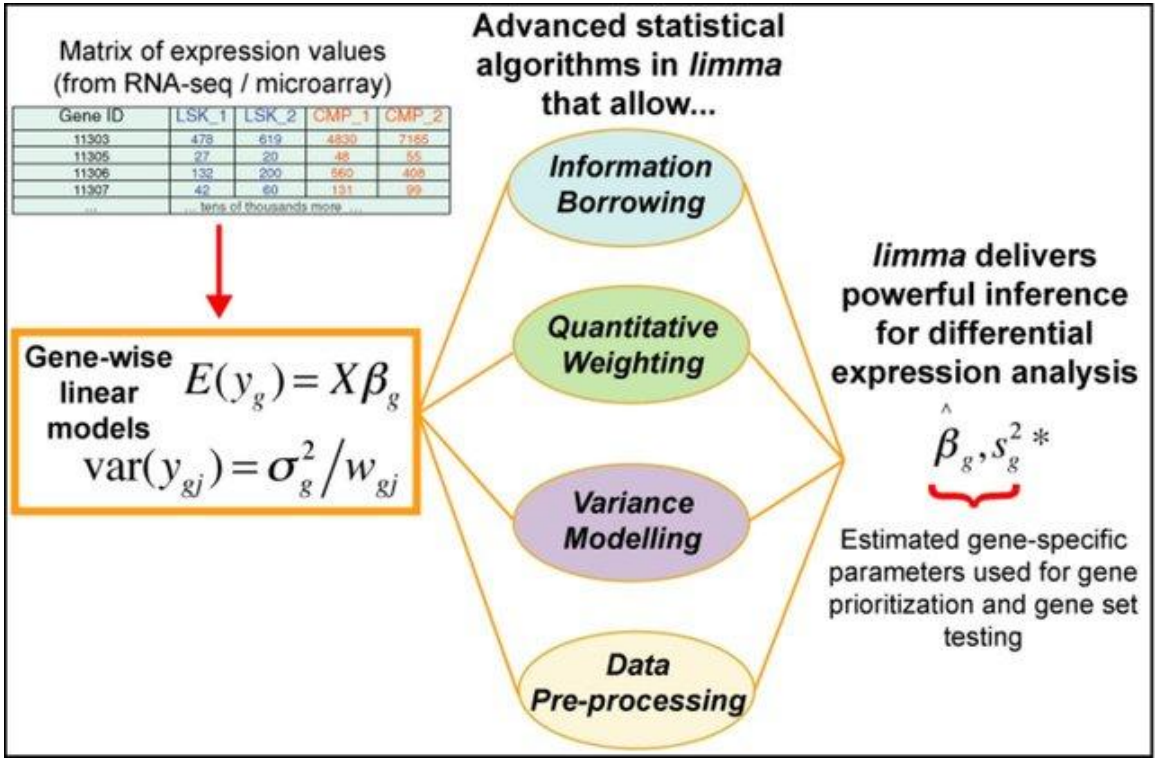
- Run step 5

GEOquery

- Works only for microarrays
- Not for all arrays there is “annotated” (i.e. curated) annotation
- RNA-seq datasets result in an empty matrix
  - Data can be loaded from ARCHS4 file

- Open do_limma.R
- Run everything

limma



Smyth group, NAR 2015

Exercises

- Plot PCA
- Add batch information to the design, calculate differential expression. Are the results differ?
- Do pathway analysis with fgsea
- Do pathway analysis with camera()
  - compare results to fgsea

Summary

- There are several common RNA-seq pipelines: alignment-based and kallisto-like
- Multiple tools for downstream analysis
- Visualize and QC your data