RNA-Seq Data Analysis

14th & 15th June 2021

Helen Lockstone

Head of Bioinformatics and Training

Bioinformatics Core, Wellcome Centre for Human Genetics

Assistant Instructors: Irina Chelysheva and Srinivasa Rao

RNA-Seq: Workshop aims

- Hands-on practice with the key steps in analysing RNA-Seq data
- Cover some key concepts and use a real dataset to find differentially expressed genes
- Highlight critical steps for performing analysis correctly

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_June2021

RNA-Seq: Workshop format

- Interactive guided demonstration of 3 practical tutorials covering key steps in the analysis process:
  1. Data exploration and quality control (QC)
  2. Data formatting and pre-processing
  3. Data analysis to find differentially expressed genes

- Feel free to ask questions in the chat or directly as we go along (including R/code issues – we will resolve if possible or help during breaks or after the session)

RNA-Seq: Early processing steps

- Brief re-cap of steps up to our starting point, a gene count table
  - RNA sample preparation
  - Sequencing – FASTQ files
  - Mapping to reference genome (splice-aware aligners)*
  - Counting reads mapping to annotated genes (features)

- Gene count table
  - Large fraction of genes with zero or very low counts
  - Extra sequencing does *not* help pick up low-expressed genes in a global transcriptome approach
  - Many factors influence counts

*computationally intensive step, usually requiring cluster resources

RNA-Seq: Data normalisation

wellcome centre human genetics

Raw gene counts are influenced by:

- Sequencing depth - varies by sample
- Gene length
- Amplification/sequencing biases – GC content
- Positional biases – preferential locations for RNA fragmentation
- Also by sample composition, which is less obvious – described in the TMM normalisation performed by edgeR

Count summaries accounting for various factors

- RPKM (reads per kilobase per million reads
- FPKM (fragments per kilobase per million reads)
- CPM (counts per million reads)
- TPM (transcripts per million reads)

RNA-Seq: Data quality

- Vitally important to plot data as first step - from a few simple plots gain important information about the characteristics of a dataset
- Use to design the appropriate analysis
- Little, if any, is evident from the count tables alone

- From QC plots, the analyst can assess:
  - overall quality of data
  - failed samples and potential outliers
  - clustering patterns
  - factors influencing gene expression profiles
  - potentially mis-labelled or swapped samples (independent verification needed though)

wellcome centre human genetics

Practical Tutorial I

Data Exploration and Quality Checks

RNA-Seq: Data exploration and QC

Look at the html version of Practical Tutorial 1 at the link below:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_June2021/Practical_tutorial_1/

From the same link, you can download the 'RNA_workshop_June2021_Practical1_QC.rmd' file, which can be opened in your R session to run the commands interactively

wellcome centre human genetics

Practical Tutorial II

Data Preprocessing

RNA-Seq: Data pre-processing

Practical tutorial 2 covers the initial steps of reading data into R, formatting, and pre-processing steps of normalisation and filtering:

https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_June2021/Practical_tutorial_2

View with the html version and load the *Rmarkdown* (.rmd) version into Rstudio to run the commands