AMIA 2016 Annual Symposium Workshop

RNAseq Data Analysis and Clinical Applications

Riyue Bao, Ph.D. The University of Chicago
rbao@bsd.uchicago.edu

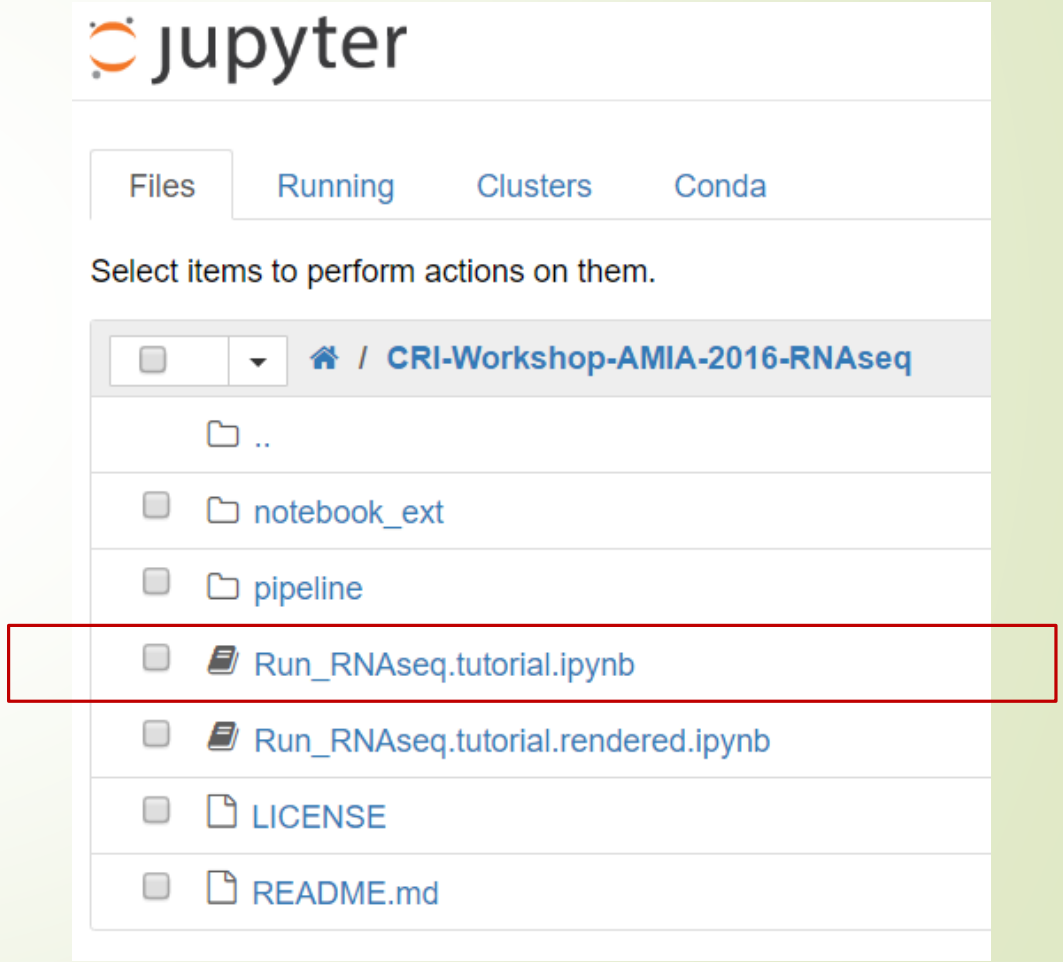
Objective

- Introduction to RNAseq technology and clinical application
- How to perform RNAseq analysis: From raw data to differentially expressed genes & pathways <u>hands-on</u>
  - *Dataset: two groups (PRDM11 KO vs WT, human U2932 cells), 6 samples*
- How to associate gene expression data with clinical outcome: survival analysis <u>hands-on</u>
  - *Dataset: The Cancer Genome Atlas (TCGA), ovarian cancer, ~600 primary tumors*

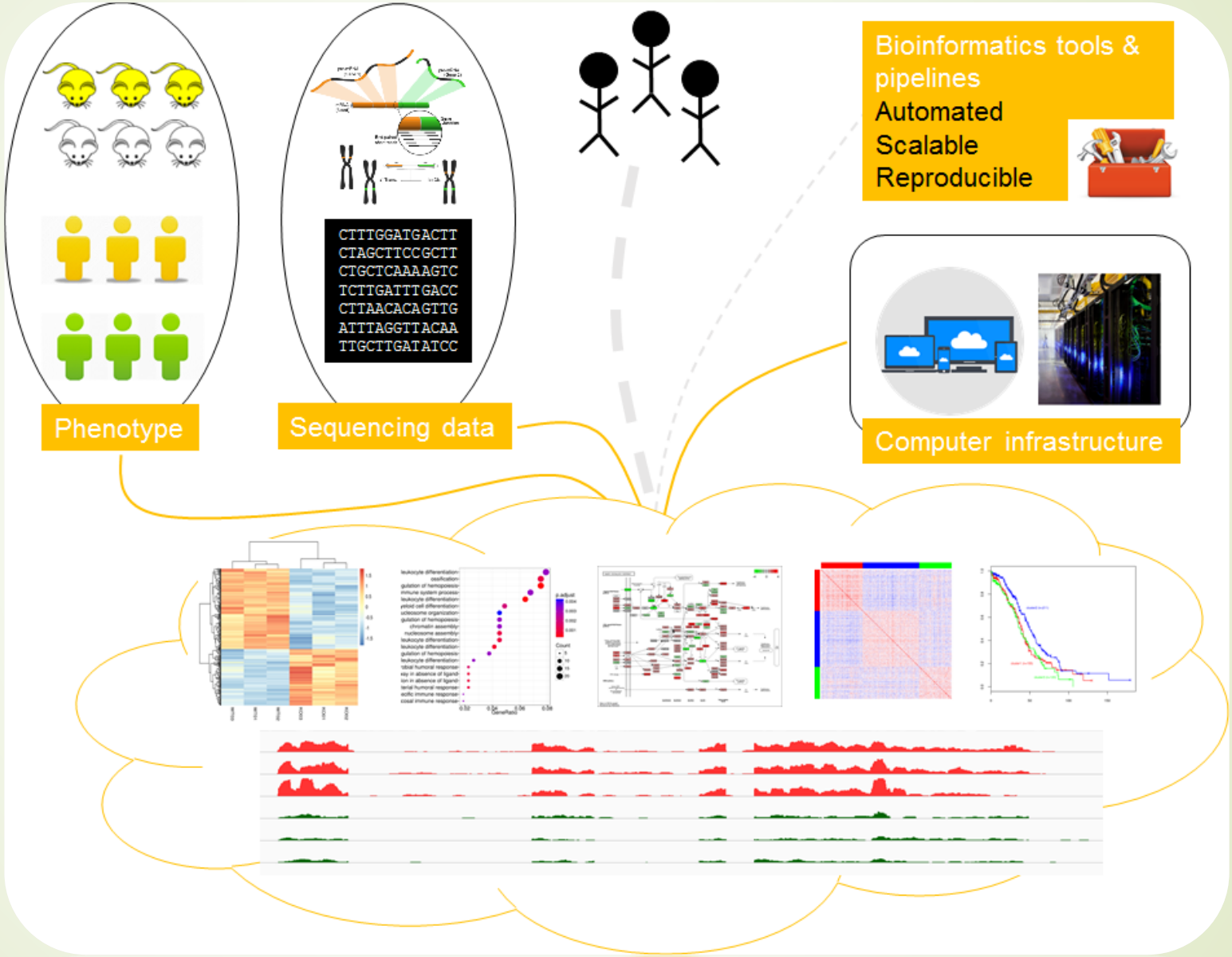
PRDM11 = PR/SET domain 11

Workshop materials

- **AWS-EC2 cloud**
- GitHub
  - https://github.com/cribioinfo
- This lecture note contains the same contents as the notebook. In addition, the notebook also contains hands-on materials
  - **Run_RNAseq.tutorial.ipynb** (in directory *CRI-Workshop-AMIA-2016-RNAseq*)



https://<IP>:8888/notebooks/CRI-Workshop-AMIA-2016-RNAseq/Run_RNAseq.tutorial.ipynb


Bioinformatics tools & pipelines
Automated
Scalable
Reproducible
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
Phenotype
Sequencing data
Computer infrastructure

Biological and clinical questions

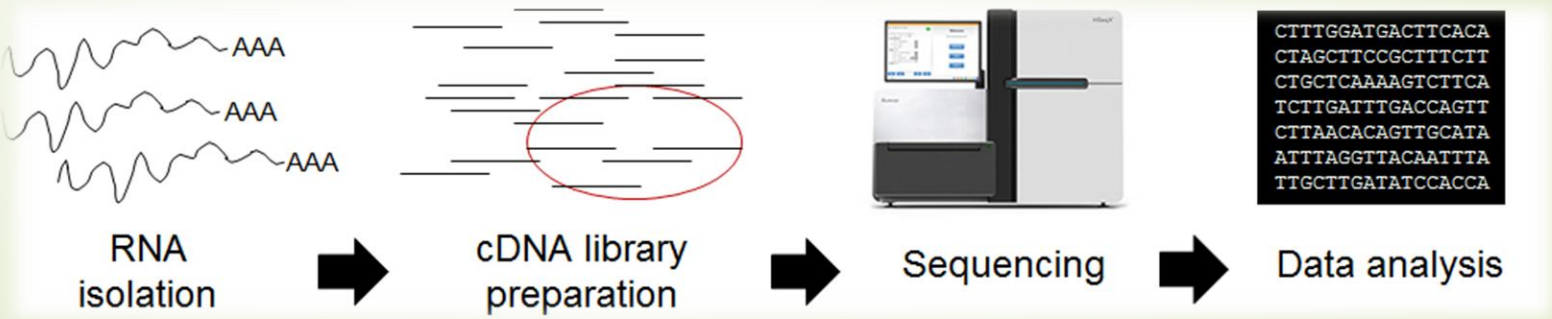
- I am interested in studying transcriptional landscape shift before and after drug treatment in cell lines
- I want to identify which pathways are affected after knocking down my favorite gene in mice
- I have expression data of clinical isolates collected at various time points, when patient's response changed. Why?
- I have a cohort of patients and want to discover which gene signature predicts patient's response to treatment
- I want to detect gene fusions, expressed mutations, and disrupted isoforms in tumors that may be related to disease

... and more!

What is RNAseq?

High-throughput sequencing of RNA: Profile, identify or assemble transcripts

- Detect gene expression changes between conditions
- Identify novel splice sites / exons, mutations, fusion genes, etc.
- Broad detection range, high sensitivity, low requirement of RNA amount
- Available for all species (reference genome is optional): reference genome-guided alignment or *de novo* assembly



Experimental design: Biological replicates

Include biological replicates for **increased discovery power** and reduced false positives/negatives!

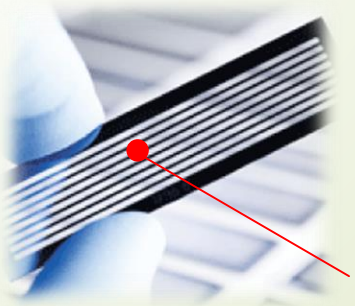
Include biological replicates for every condition!

- 3+ for cell lines
- 5+ for inbred lines
- 20+ for human samples

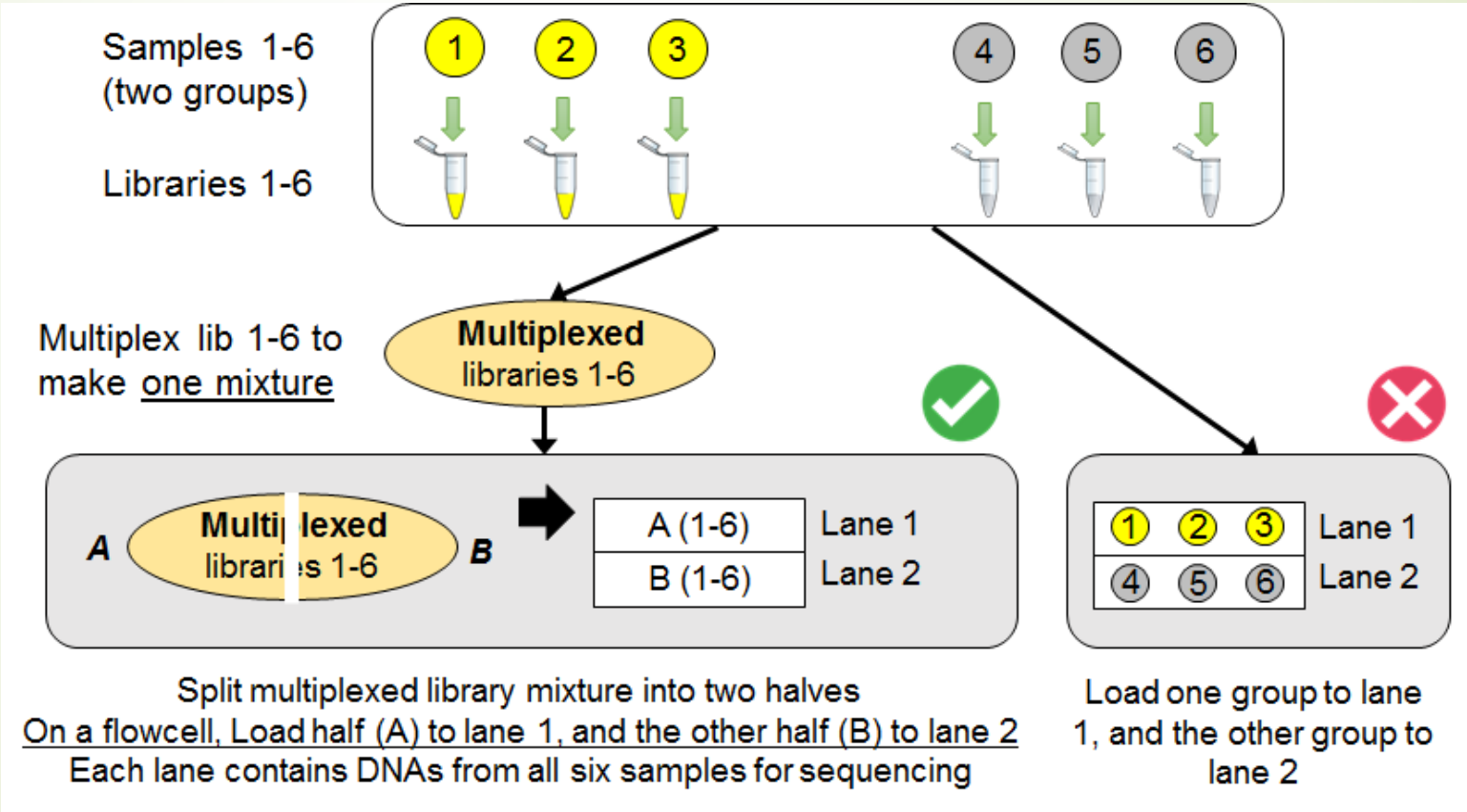
Experimental design: Multiplexing and Randomization

- **Multiplexing**: simultaneously measures multiple libraries in one sequencing lane. Unique barcodes are added to label DNA molecules from each library

- **Randomization**: Avoid loading samples from the same biological group in the same sequencing lane. Minimizes technical bias and lane-specific effects.

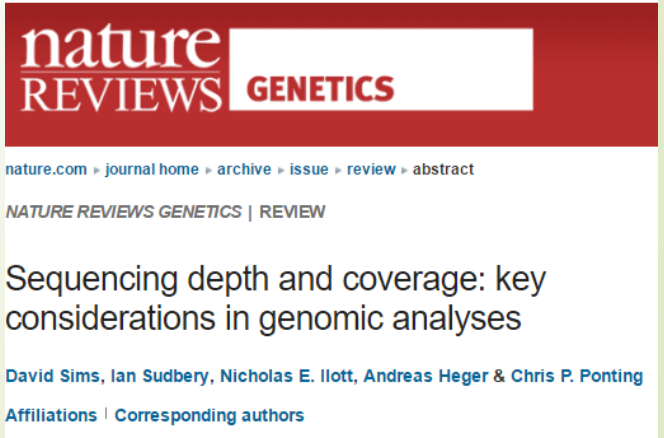


Experimental Design: Multiplexing and Randomization



Challenges and limitations

- Relatively poor RNA quality for tumor FFPE samples
- Contamination from adjacent normal tissue
- Still more expensive than targeted-panel sequencing such as NanoString
- 40 million mapped reads are usually sufficient for gene profiling, but > 80 million are required to detect bottom 1% lowly expressed genes


nature REVIEWS GENETICS

nature.com ▸ journal home ▸ archive ▸ issue ▸ review ▸ abstract

NATURE REVIEWS GENETICS | REVIEW

Sequencing depth and coverage: key considerations in genomic analyses

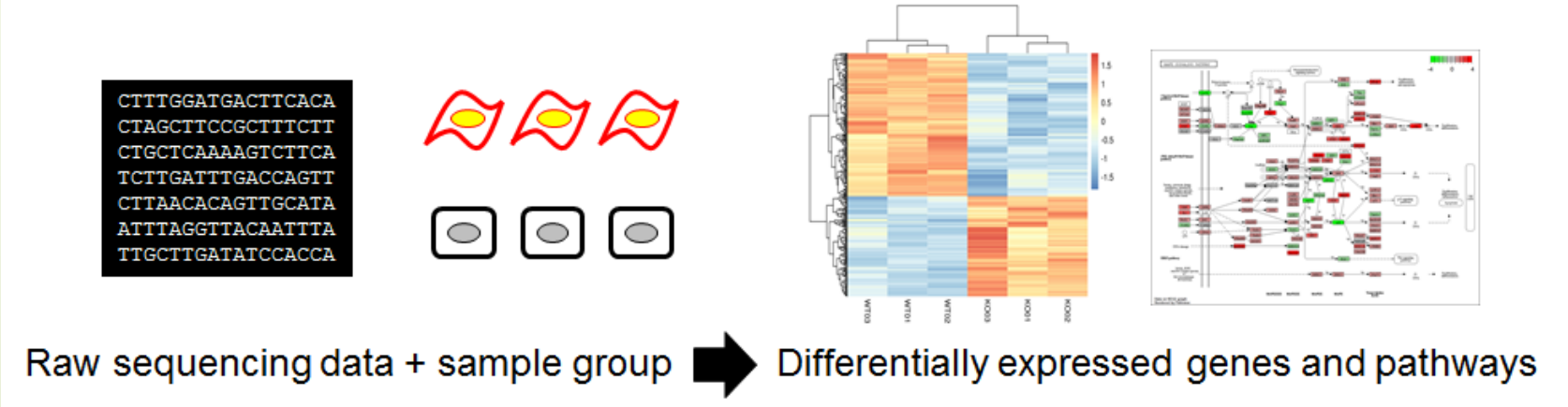
David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger & Chris P. Ponting

Affiliations | Corresponding authors

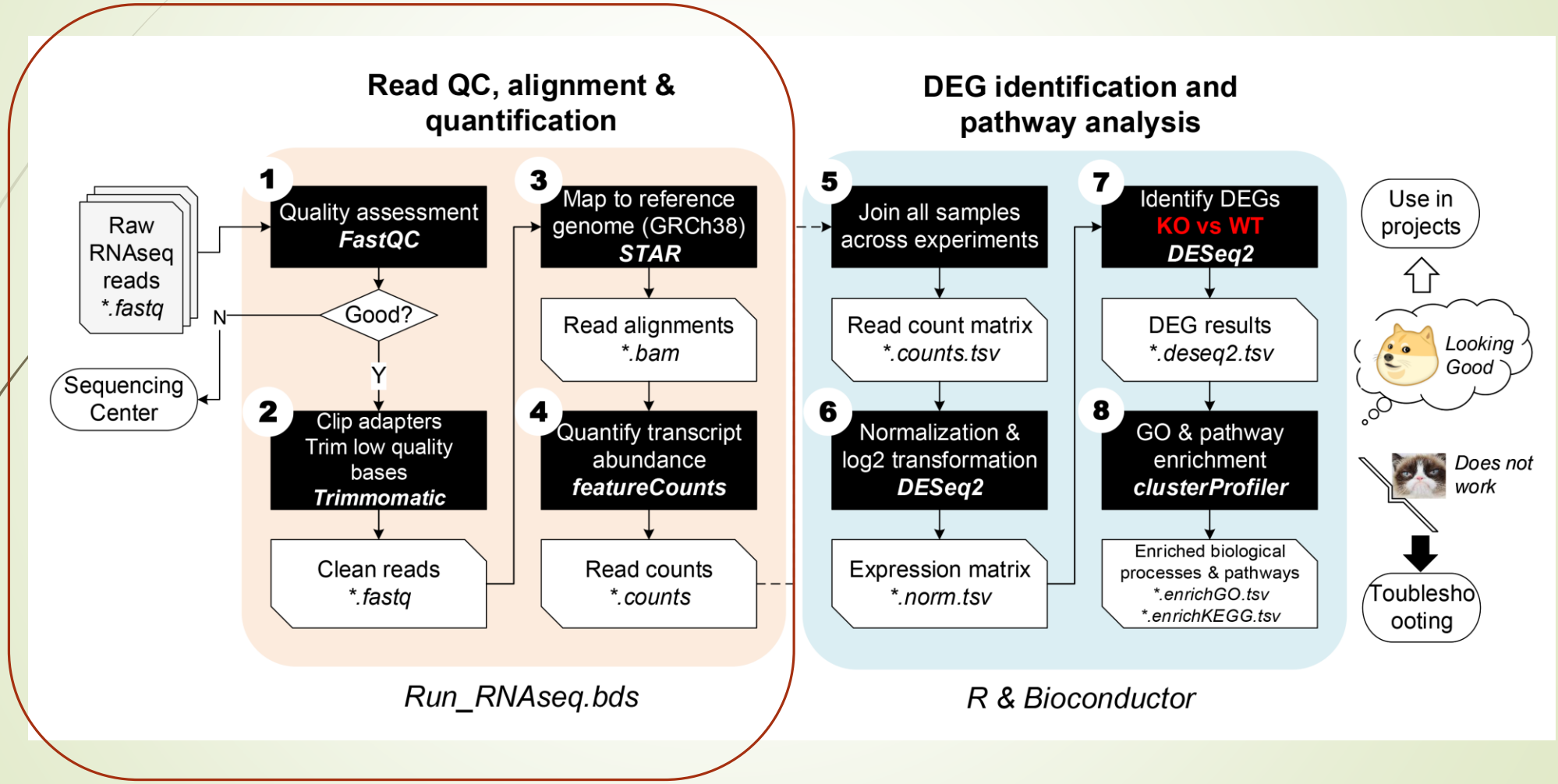
How to perform RNAseq analysis

The good-practice analysis protocol takes 8 major steps.

- **01-04**: From raw sequencing to transcript quantification is automated through BigDataScript (BDS) pipeline
- **05-08**: DEG and pathway analysis will be practiced interactively to better interpret the results.



How to perform RNAseq analysis



01-02: Quality assessment and preprocessing of raw sequencing reads

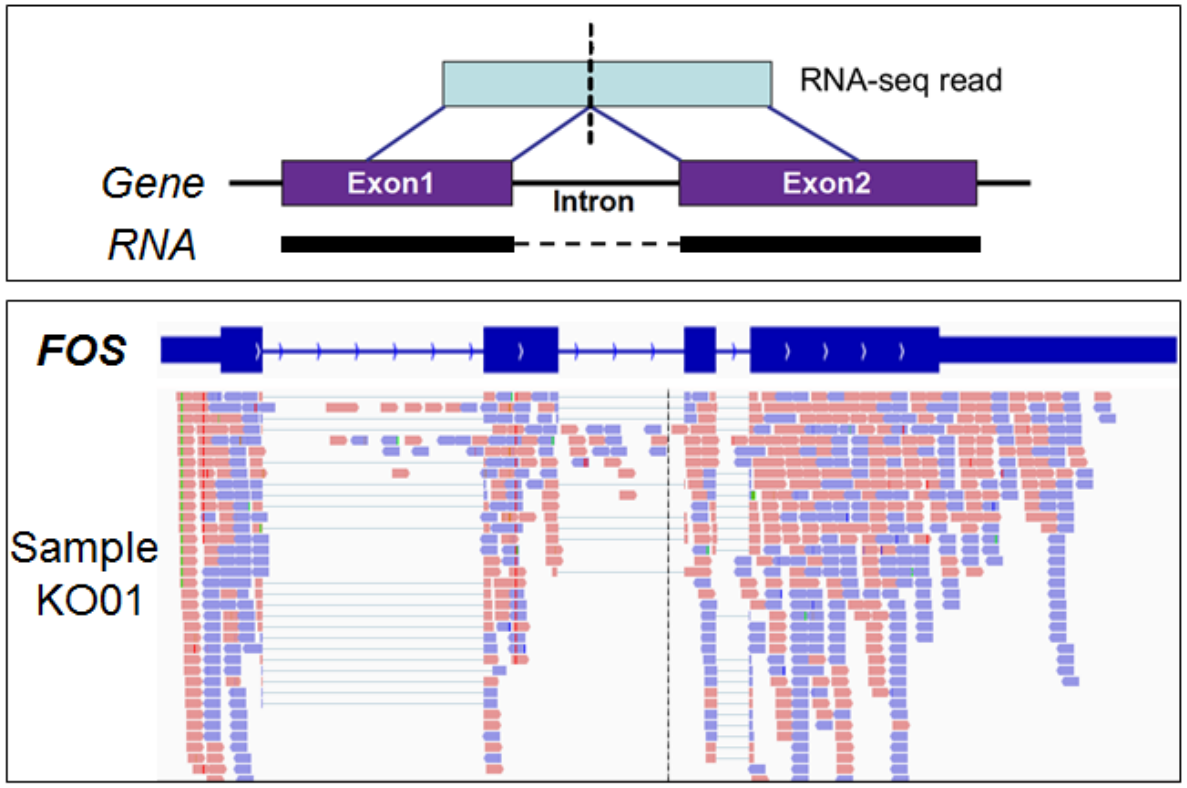
- **Raw sequencing reads** are stored in FastQ format (e.g. *KO01.fastq.gz*), where each read is presented by 4 lines

```
Sequence header --  @SRR1205282.43583628
Sequence string --  GACTATCTTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAA
 Quality header --  +
 Quality string --  CCCFFFFFHHHHHIIIIIIIHHIIIIIIIIIIIIIIIIIIIIIIIIIII
```

- QC produces reports that help you evaluate if a sequencing run is successful and if reads are of high quality (e.g. *MultiQC*)
- Preprocess reads to improve mapping rate and accuracy
  - Trim low-quality bases, clip adapters, etc.
  - Avoid over-trimming in RNAseq!

03-04: Map reads to reference genome and quantification of transcript abundance

- Read mapping identifies the location in the genome where a sequencing read comes from
- **Splice-aware aligner** (e.g. *STAR*)



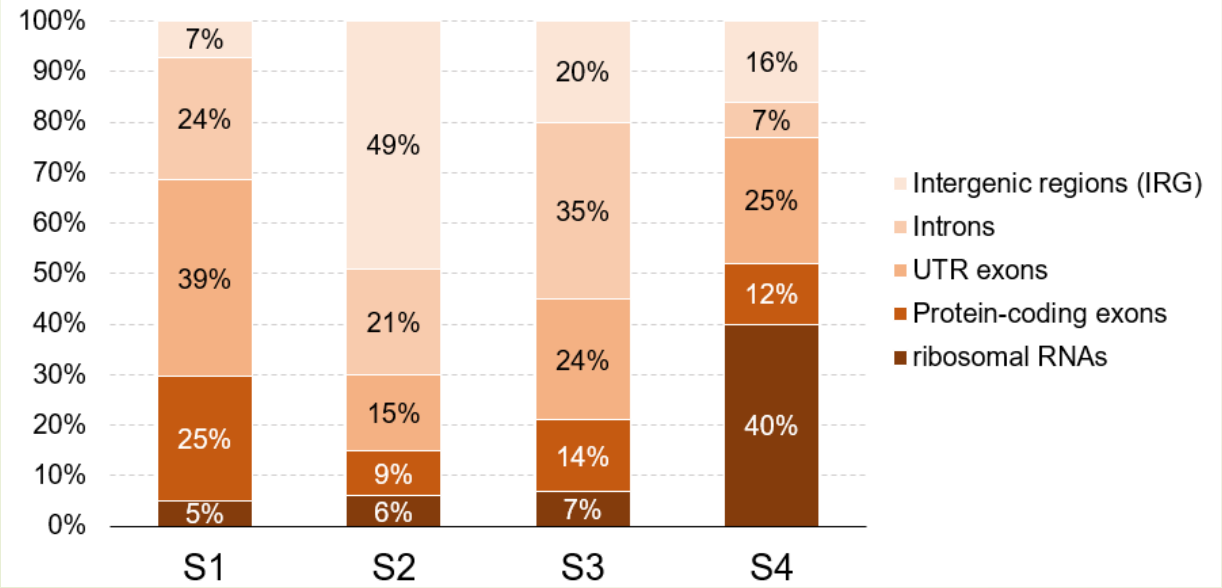
780 reads map to *FOS* gene

Each horizontal bar represents one read. Red/blue indicates reads aligned to plus/minus strand on the genome, respectively.

RNAseq metrics

Collect metrics to evaluate RNA sample quality and identify potential problems

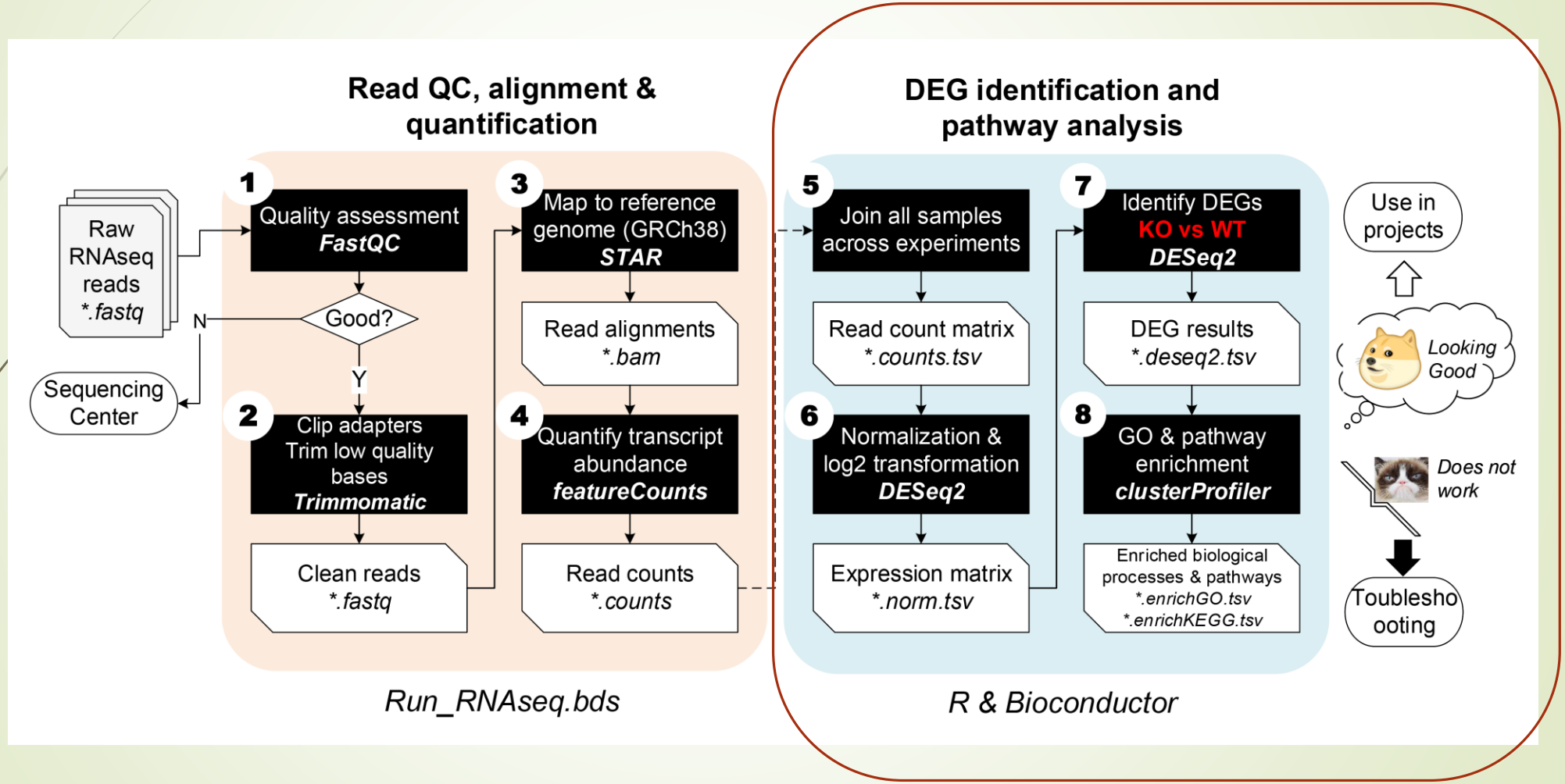
- Is there high-level genomic DNA contamination?
- Was ribosome RNA successfully depleted during library prep?



Q1: Which sample (S1-4) has the most severe genomic DNA contamination?
Hint: higher percentage of intergenic reads indicates more severe DNA contamination in RNA samples

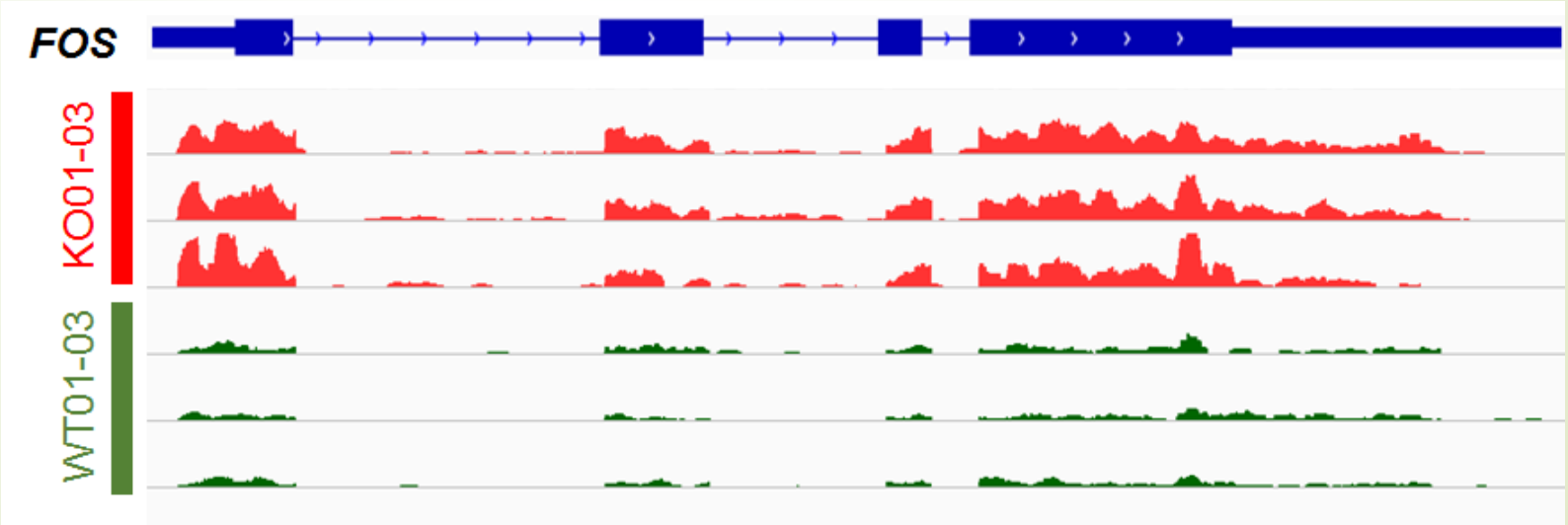
Q2: Which sample (S1-4) has the least efficient depletion of ribosome RNAs?
Hint: rRNAs account for > 80% of the whole transcriptome. If not removed, the majority of the sequencing reads will be derived from rRNA

How to perform RNAseq analysis



05-08: Identify differentially expressed genes and pathways

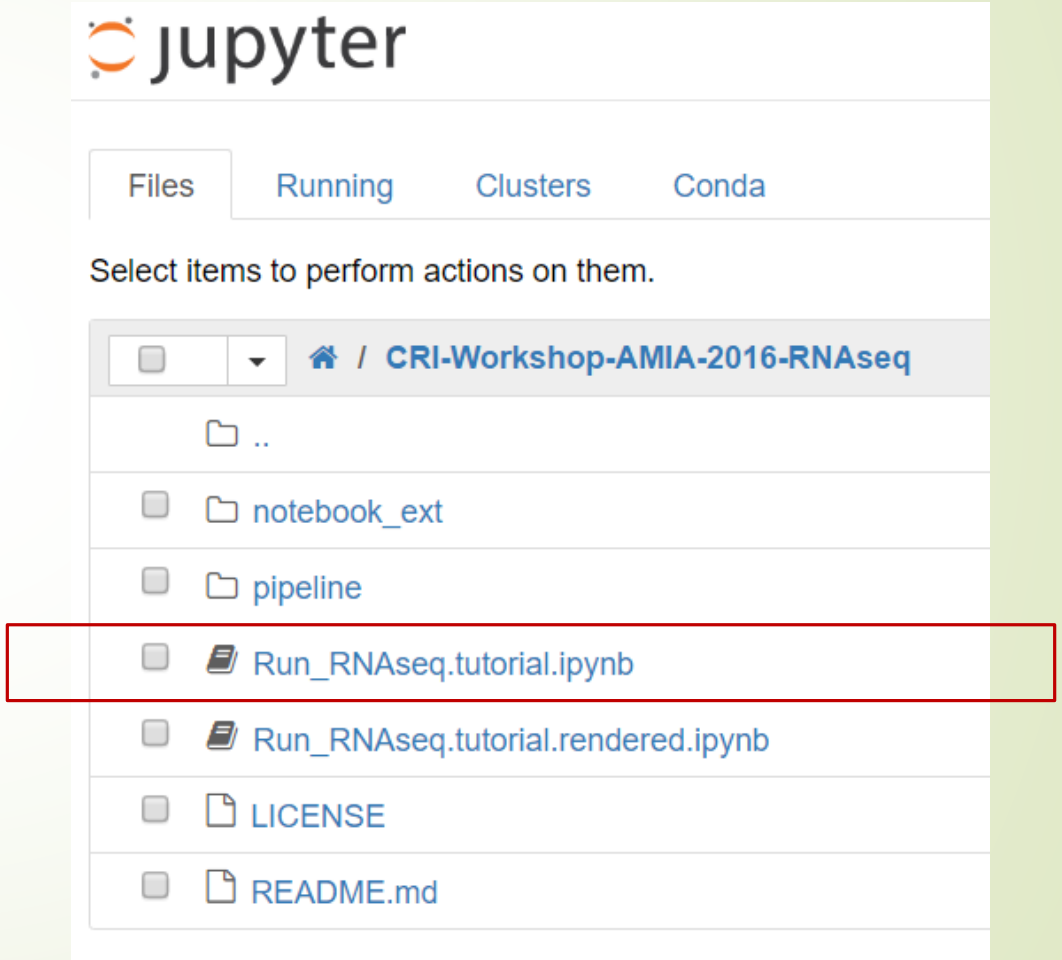
- After steps 01-04, we have generated read alignment and counts for every annotated gene on the genome
- The next step is to utilize the read counts data to detect DEGs
- For example, if we visualize *FOS* gene across 6 samples in genome browser



FOS = Fos proto-oncogene, AP-1 transcription factor subunit

Hands-on practice START

- **Open your notebook on the AWS machine**



https://<IP>:8888/notebooks/CRI-Workshop-AMIA-2016-RNAseq/Run_RNAseq.tutorial.ipynb