Breakout Session 2: Track A

SRA RNA-seq Precomputed Alignments and Gene Expression Counts

Dr. Kim Pruitt (Moderator)
Acting Director, NCBI, NIH/NLM

SRA RNA-seq precomputed alignments and gene expression counts

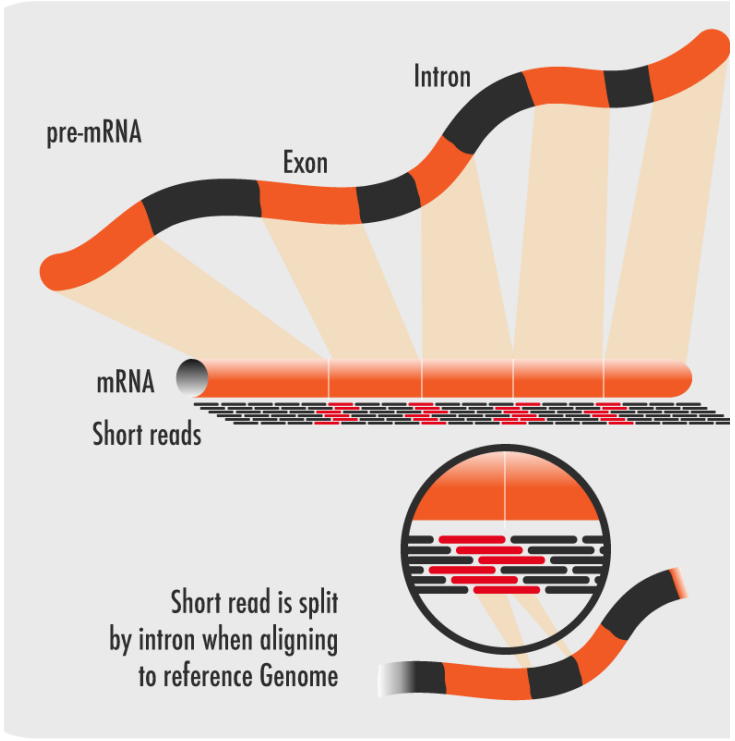
- RNA-seq Value
- Challenges
- NCBI Cloud Pipeline
- Lessons Learned
- Future Directions

Kim Pruitt, PhD; Acting Director, NCBI

ODSS Cloud Supplement PI meeting, January 18, 2024

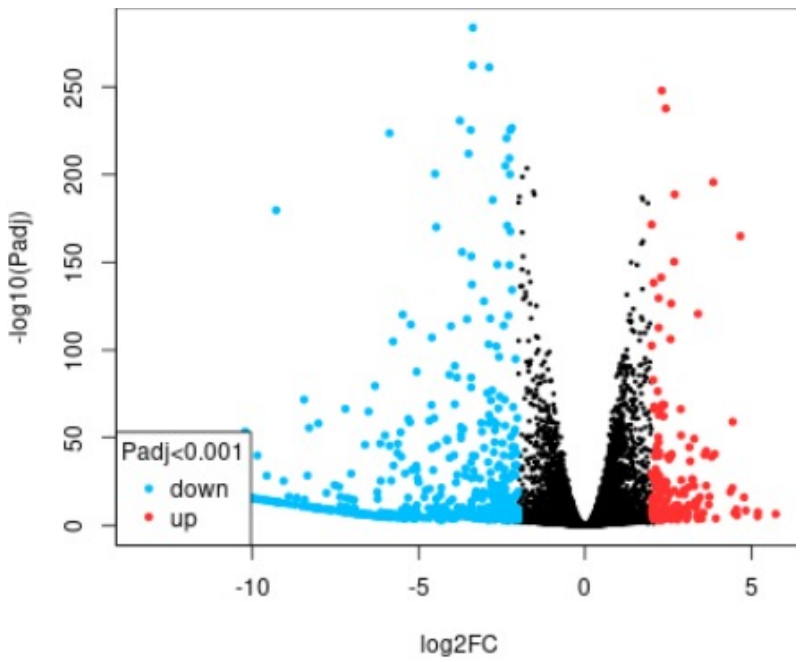
U.S. National Library of Medicine
National Center for Biotechnology Information

RNA-seq has revolutionized biomedical science

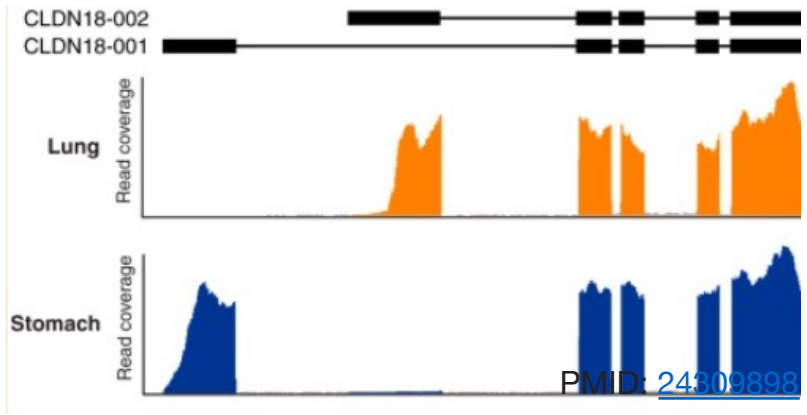


Credit: Technology Networks

RNA transformed into quantifiable short DNA sequence reads



Genomic-scale gene expression



Refined genome annotation

RNA-seq data pose challenges for data analysis and re-use

- SRA data files are large (1.5 GB average)
- Requires sufficient computing storage and power
- Requires specialized computing knowledge
- Pre-computed expression analysis results are more FAIR

NCBI's Cloud-based RNA-seq pipeline:

- ✓ Aligns SRA public human ~1.5Gb scale RNA-seq runs to genome assembly GRCh38.p13
- ✓ Analysis carried out on GCP
- ✓ Produces gene-level counts for each run in a ~560 Kb small file (*2600-fold reduction in data size*)
- ✓ Count data accessible from GEO
- ○ Count data will be given SRA unique identifiers
- ○ Count data will be delivered from the cloud, estimated Q3FY24

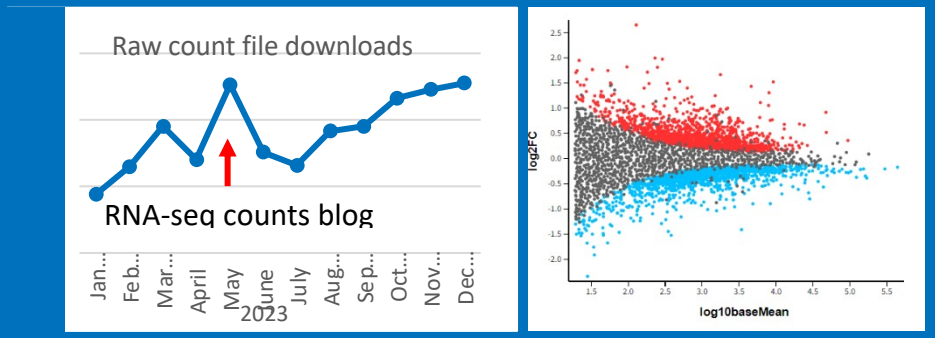
NCBI RNA-seq cloud analysis pipeline: Output to date

- 1.9 million human runs processed so far
- 1.3 million human runs passed 50% alignment threshold
- Each run processed with the pipeline:
  - Costs $0.16
  - Takes a few minutes
- ~25K GEO studies have counts available for download or analysis in GEO2R

GEO ACCESS

1. Access/analyze using GEO2R Tools
2. Download for local analysis

~3 fold-increase in file downloads from January – December 2023



RNA-seq in the cloud: lessons learned

- NCBI workforce development
- Quality control metrics design (will be released on Cloud)
- Data management (metadata, stable IDs, attribution, status, deployment)
- Advanced planning for access and search
- Learned how to better use existing data model for data delivery and provided feedback for making the model more robust

RNA-seq in the cloud: Future directions

- Release data on SRA cloud
- Pipeline will be updated to use SRA Lite format (will reduce cost by 84%, to about $.02 per run)
- Processing of 1.9 million mouse RNA-seq runs
- Improve UX with download button from GEO
- Raise awareness of data availability
- Acquire feedback with user research
- Publishing pipeline

Acknowledgements

Kim Pruitt, Ph.D.

Valerie Schneider, Ph.D.

Ilene Mizrachi, Ph.D.

Rodney Brister, Ph.D.

Ryan Connor, Ph.D.

Tanya Barrett, Ph.D.

Emily Clough, Ph.D.

Andrey Kochergin

Lukas Wagner, Ph.D.

Vadim Zalunin

Alexandra Soboleva

Maxim Tomashevsky

Naigong Zhang, Ph.D.

Nadezhda Serova

Corinne Matti

Funding:

- National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
- ODSS Co-funding

NIH National Library of Medicine
National Center for Biotechnology Information