The Question

How to make ribosome profiling data reliable, comparable & easier to analyse?

04 March 2020

Flic Anderson
The Wallace Lab, University of Edinburgh

The Answer

RiboViz?

The Longer Answer

- What is RiboViz
- Ribosome Profiling Process
  - What can the data answer?
  - Methods
  - Problems
- What does RiboViz Do?
  - Workflow
  - Outputs
- Challenges:
  - User Experience
  - Testing Reliability
- Summary: Back to The Question

What is RiboViz

RiboViz:

software pipeline to enable comparative analyses of ribosome-profiling datasets to allow researchers to identify patterns of transcriptional and translational regulation across different organisms and conditions

Carja, O., Xing, T., Wallace, E.W.J. et al. riboviz: analysis and visualization of ribosome profiling datasets. BMC Bioinformatics 18, 461 (2017). https://doi.org/10.1186/s12859-017-1873-8

Thanks / Acknowledgements

- BBSRC-NSF funded project
- Collaborative project:
  - Edward Wallace: *The Wallace Lab,* The University of Edinburgh.
  - Premal Shah, John Favate, Tongji Xing: *The Shah Lab,* Rutgers University.
  - Liana Lareau, Amanda Mok: *The Lareau Lab,* University of California, Berkeley.
  - Kostas Kavousannakis, Mike Jackson: *EPCC,* The University of Edinburgh.
  - Oana Carja, Joshua Plotkin: The University of Pennsylvania

Ribosome Profiling

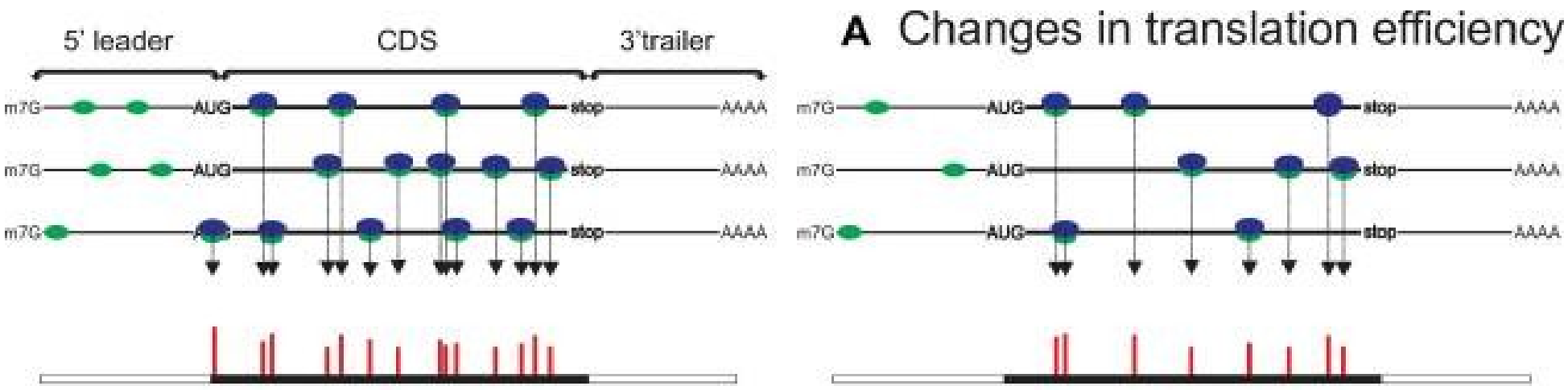
- Technique exploits ribosome-protected sections of mRNA: snapshot of active translation in vivo
- *Showing active translation allows direct comparison of translational efficiency and gene expression between conditions*
- Allows us to address key questions about gene regulation in a range of organisms

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009). "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling". Science. 324 (5924): 218–23. doi:10.1126/science.1168978

What Will It Answer?

- Aiming to gather robust comparable data to test hypotheses
  - Translational efficiency differences (changes in gene expression)
  - Transcription start sites
  - Ribosome occupancy per codon

Translational Efficiency



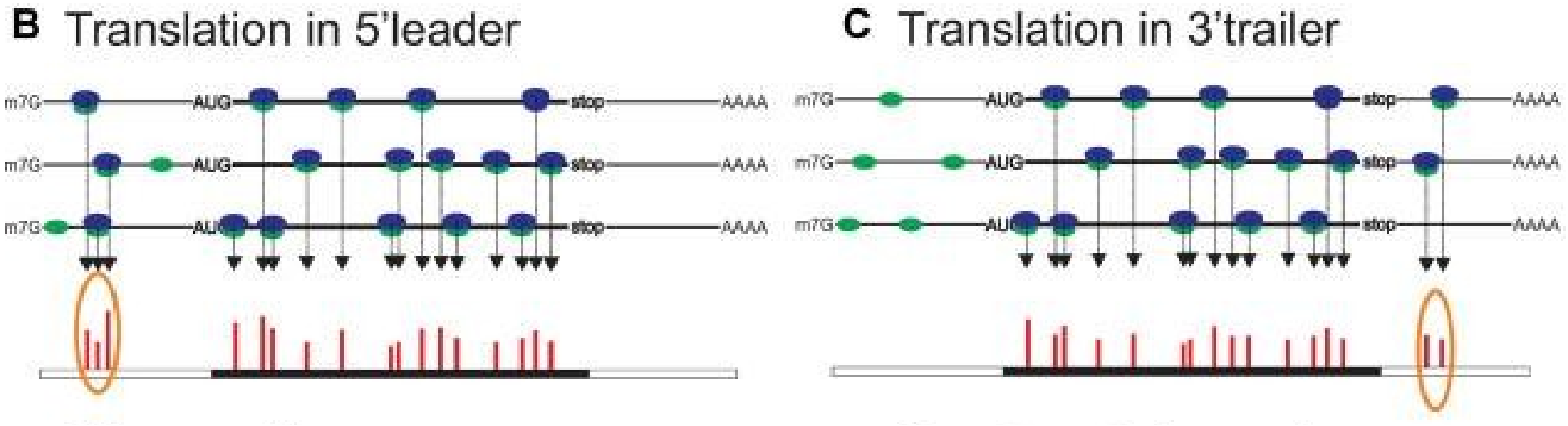
Control

Changed Conditions

A. Changes in translation efficiency (TE), represented here by decrease of TE between Control & Changed Conditions

Andreev et al. (2017). "Insights into the mechanisms of eukaryotic translation gained with ribosome profiling". Nucleic Acids Research. 45 (2): 513–526. doi:10.1093/nar/gkw1190

5’ & 3’ Investigations

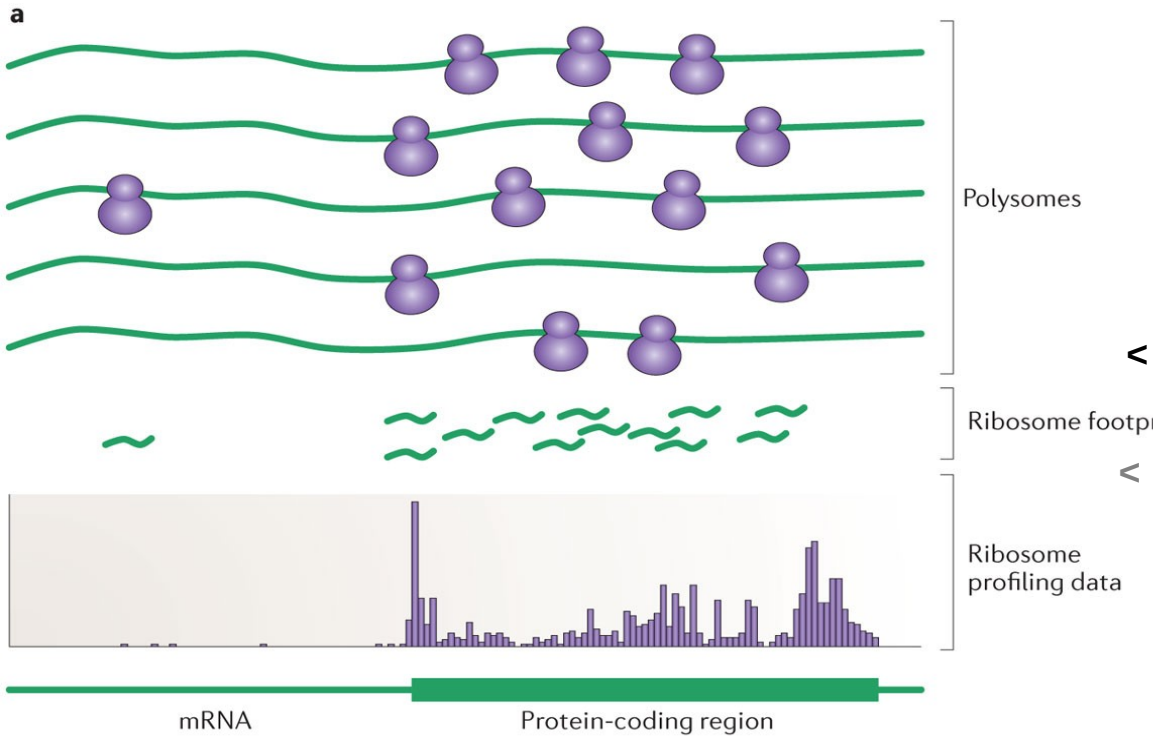


B. Translation in 5′ leader

C. Presence of ribosomes in 3′ trailer

Andreev et al. (2017). "Insights into the mechanisms of eukaryotic translation gained with ribosome profiling". Nucleic Acids Research. 45 (2): 513–526. doi:10.1093/nar/gkw1190

Polysomes to Footprints



< Experimental Stages

< Processing & Analysis

Ingolia (2014). "Ribosome profiling: new views of translation, from single codons to genome scale". Nature Reviews. Genetics. 15 (3): 205–13. doi:10.1038/nrg3645

Polysomes to Footprints

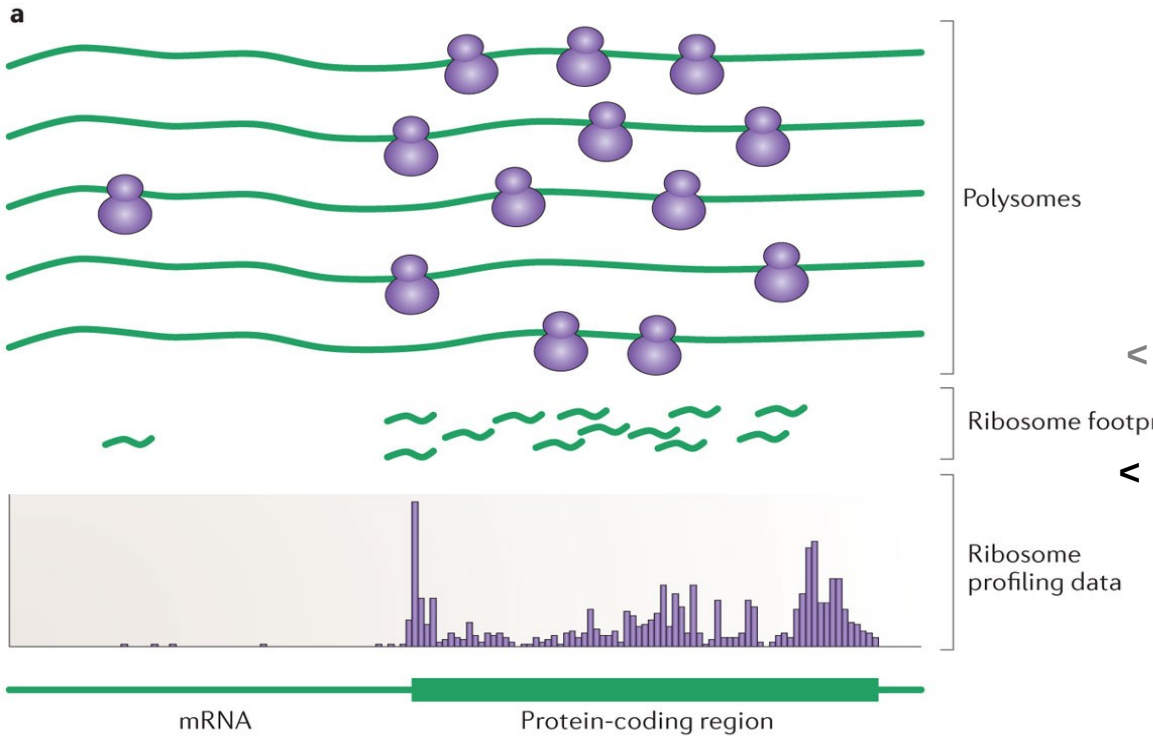
- Lyse the cells to get at the mRNA molecules bound to ribosomes.
- 'Stop' the translation process: e.g. with cycloheximide or other means
- Digest the non-protected RNA using a nuclease
- Strip away the ribosomes and proteins
- Size-select for these previously 'masked' fragments of mRNA
- Add adapters
- Reverse-transcribe to complimentary DNA
- Amplify
- Sequence
- *… PARTY?**

  (NO! Not yet.)

Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009). "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling". Science. 324 (5924): 218–23. doi:10.1126/science.1168978

* *Note: not included in original methods of Ingolia et al.*

Footprints to Profiling Data



< Experimental Stages

< Processing & Analysis

Ingolia (2014). "Ribosome profiling: new views of translation, from single codons to genome scale". Nature Reviews. Genetics. 15 (3): 205–13. doi:10.1038/nrg3645

Footprints to Ribosome Profiling Data

- **Processing**: *lots of steps*
  - Removing adapter sequences
  - Remove UMIs (Unique Molecular Identifiers) & barcodes if present
  - Demultiplex / Deduplicate reads if required
  - Need to filter out contaminant reads
  - Align reads to transcriptome

- **Analysis**: *more steps*
  - Analyse & quantify data
  - Create outputs (including for quality-control, further analysis

Footprints to Profiling Data: Problems

- **Processing**: *lots of steps & lots of problems*
  - Different tools
  - Compatability of inputs/outputs
  - Poor testing / no checks on reliability
  - Tools not updated, necessary features may not be added

- **Analysis**: *more steps & even more problems*
  - Data analysis methods not transparent (black box software)
  - Software versions
  - Unhelpful documentation
  - Re-running datasets can be difficult
  - Output file chaos

Methods Discussions

- Methods used in experimental and processing/analysis steps can impact results considerably:
  - e.g. Debate on experimental protocol aspects such as use of cycloheximide (Duncan & Mata 2017)
- Processing & analysis steps haven’t been widely examined and discussed;
- Potential for obtaining unreliable results
  - inconsistencies between methods
  - lack of transparency
  - unknown compatibility between inputs/outputs at key steps
- Leads to difficulties in reproducing results

Duncan & Mata (2017). “Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe”. Sci Rep 7, 10331. https://doi.org/10.1038/s41598-017-10650-1

What does RiboViz do?

- Software pipeline that uses a combination of local Python and R scripts and third-party components
- Riboviz takes:
  - one configuration file with the key parameters
  - organism-specific input files
- Builds script from these parameters & runs the pipeline, recording your workflow exactly!
- Ensures inputs/outputs at each step are compatible
- Delivers outputs with provenance information

RiboViz Workflow: Inputs

Organism Specific

Transcript Sequences
.fasta

Genome / Transcriptome Features
.gff

Contaminant Sequences (rRNA)
.fasta

(Additional Organism-Specific Data)

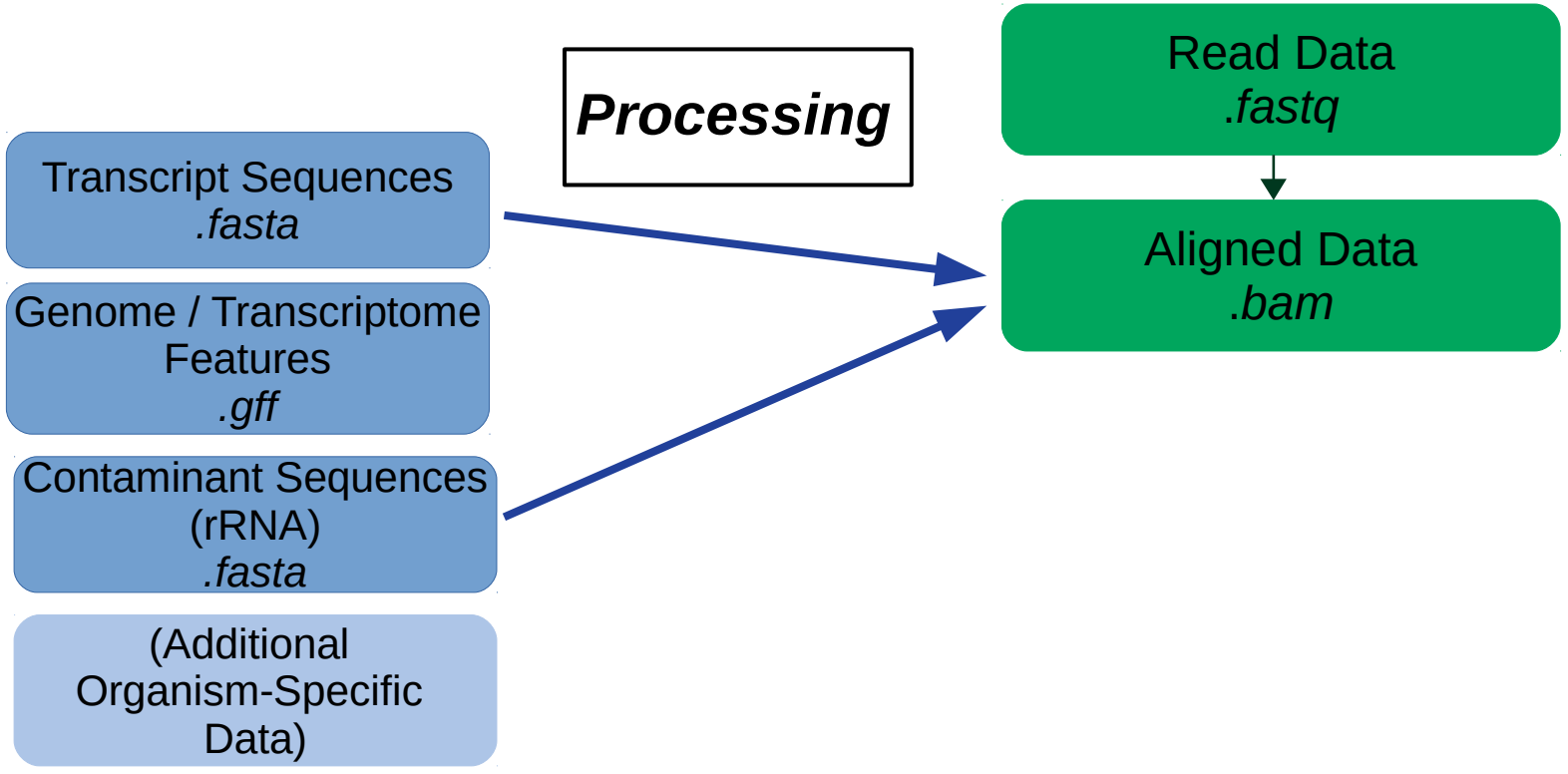
Sample Specific

Read Data
.fastq

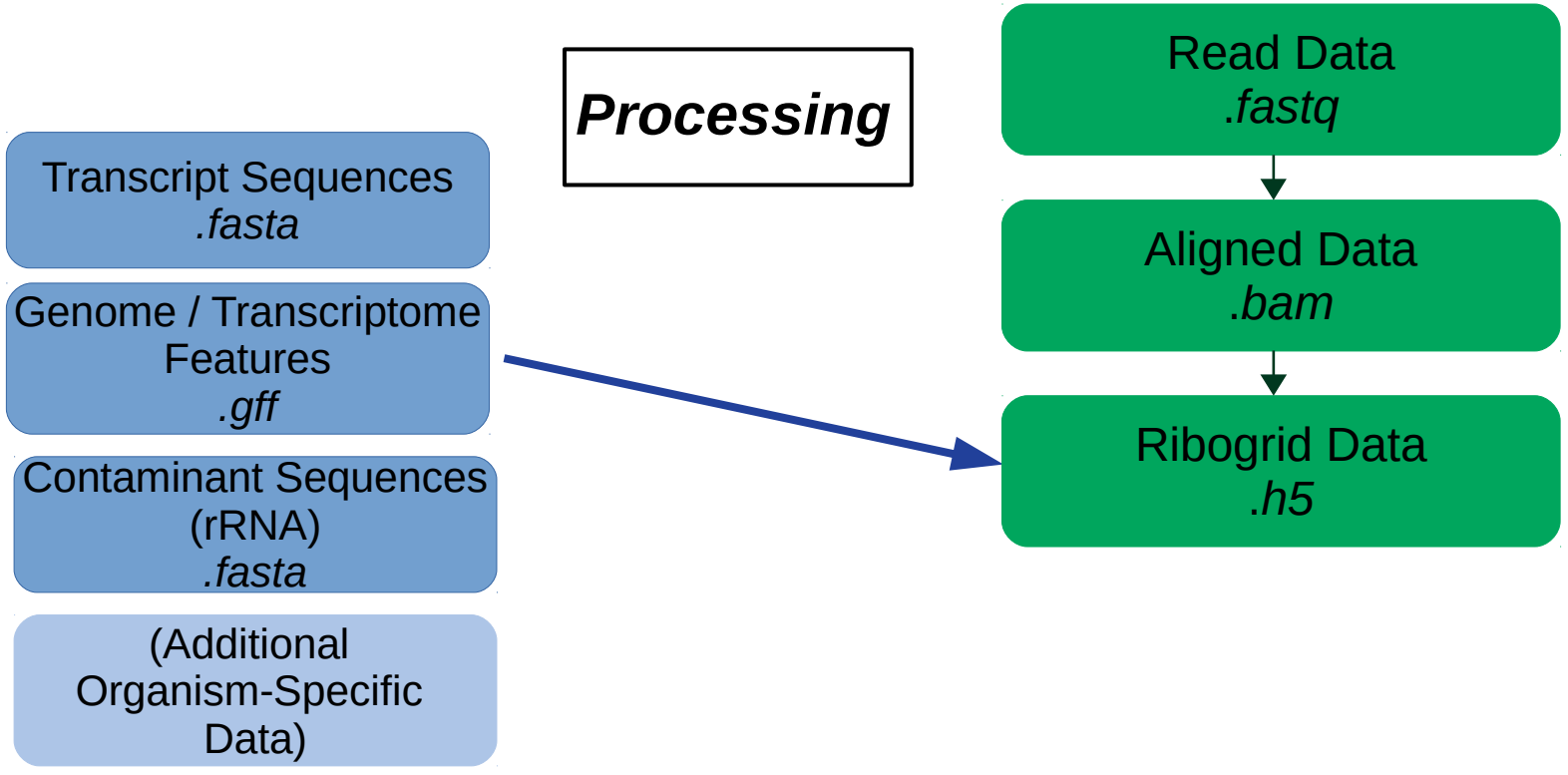
Configuration File
.yaml

Configuration File lists all files & parameters needed to run RiboViz

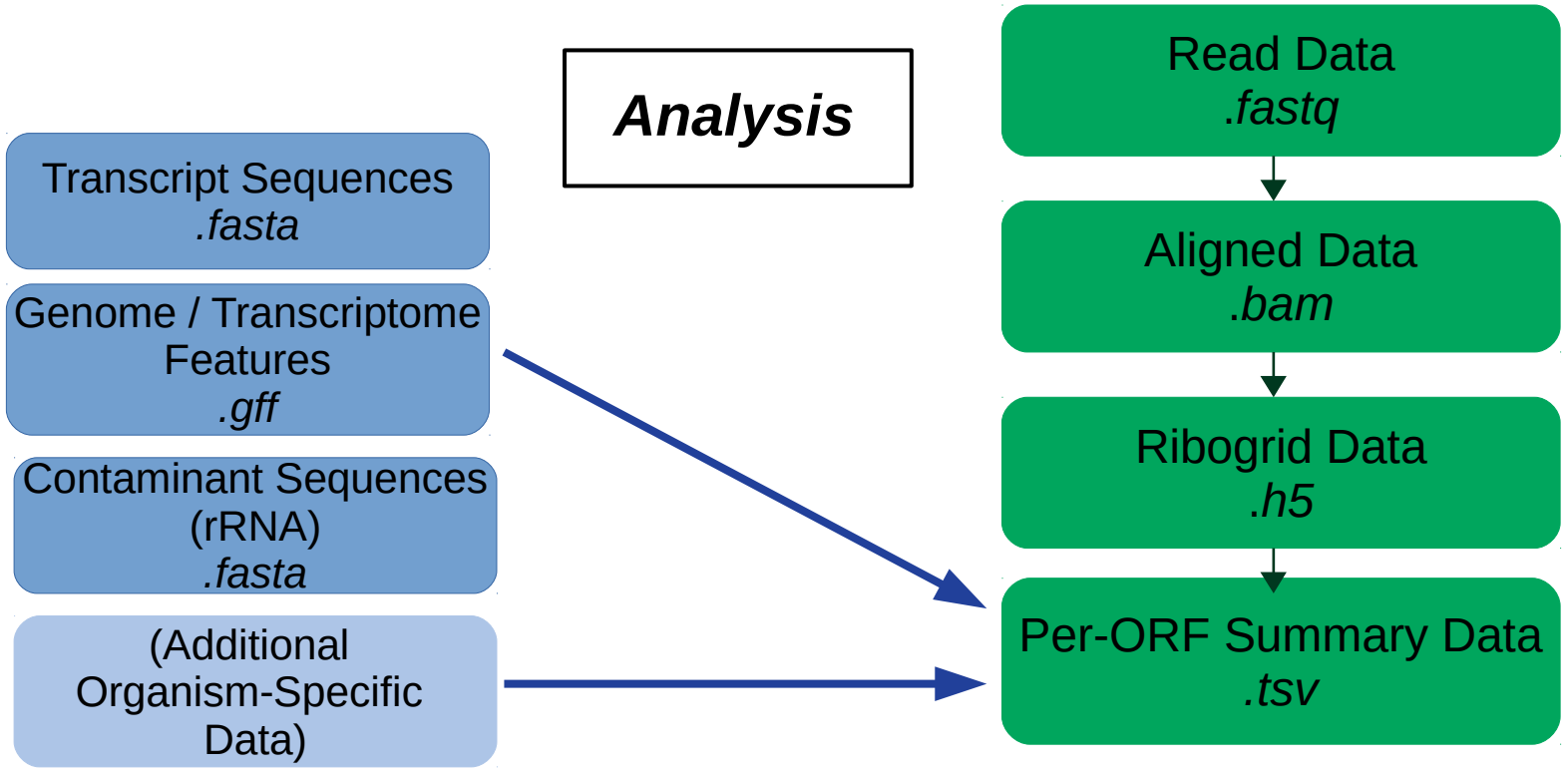
RiboViz Workflow



RiboViz Workflow



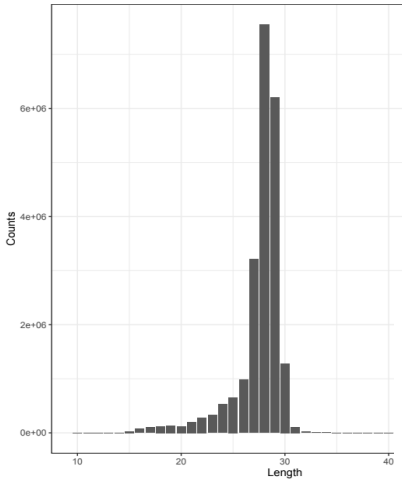
RiboViz Workflow

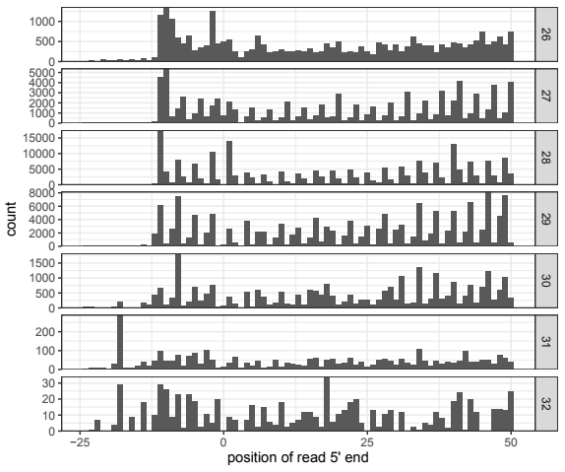


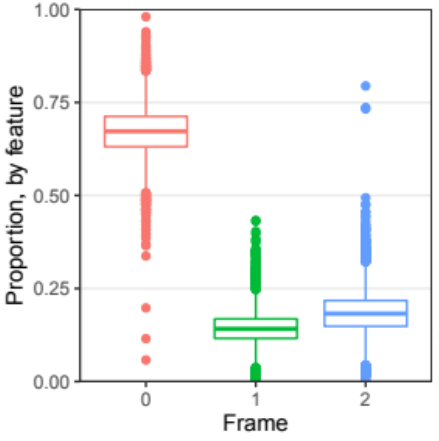
Analysis of Ribosome Profiling Data

- Looking for 3-NT periodicity: ribosomes moving along transcript 1 codon at a time
- Most reads map to coding regions (98.8% in Ingolia et al 2009)
- Reasonable read-lengths (e.g. know should look for appx ~28-30NT)
- Looking for most reads to be in one frame

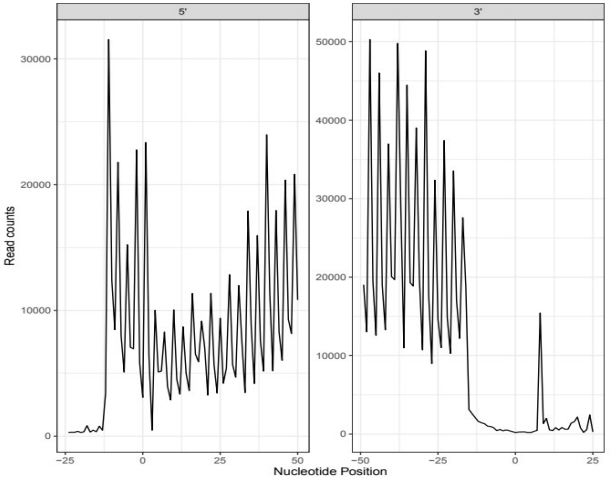
How does Riboviz show dataset is reliable?





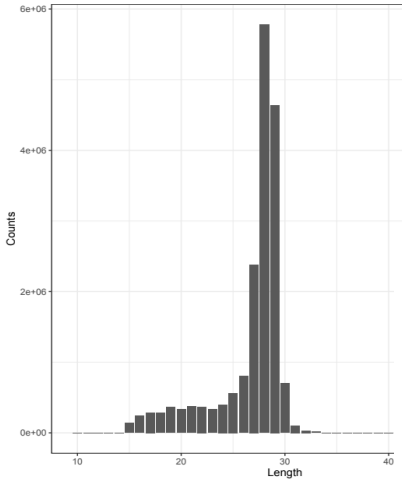


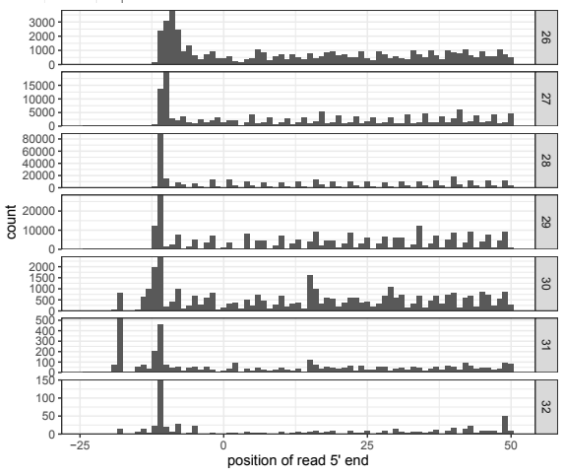
G-Sc_2014: WTnone

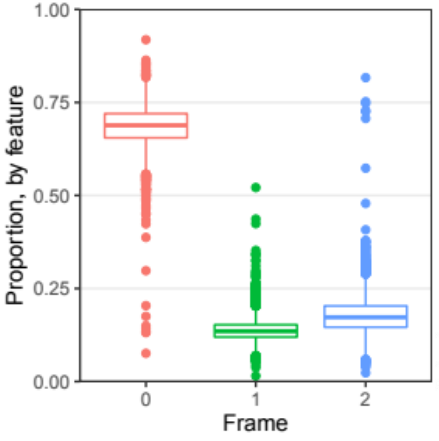


Data: Guydosh & Green (2014). “Dom34 rescues ribosomes in 3' untranslated regions.” Cell. 156 (5): 950-62. doi: 10.1016/j.cell.2014.02.006.

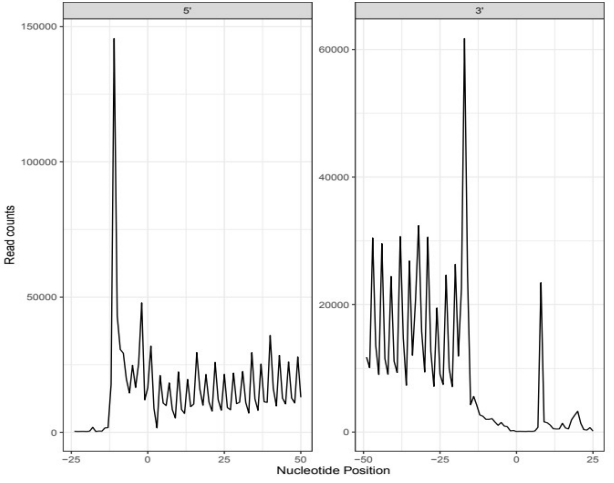
How does Riboviz show dataset is reliable?







G-Sc_2014: WT3AT



Data: Guydosh & Green (2014). “Dom34 rescues ribosomes in 3' untranslated regions.” Cell. 156 (5): 950-62. doi: 10.1016/j.cell.2014.02.006.

Good… How could I make it better?

- Comparable plots (& underlying data) but difficult to ‘really’ compare
  - Currently different files in different folders;
  - Plotted at different scales
- Going to redesign output format: .html output with clearer plotting
- Test analysis functions in the code to ensure they’re reliable
- Requires refactoring & developing RiboViz analysis code & documentation… (Ongoing!)

Challenges: User Experience (UX)

- **Users:**
  - Build user base → supporting lab members / collaborators to use RiboViz; identify more potential users
  - What do users might need / want? → need to survey the ribosome profiling community
  - ‘Onboarding’ → creating ‘oven-ready’ example datasets in separate repository as learning aid

- **Support & Documentation:**
  - Existing documentation suitable? → ‘test’ documentation with new users & improve based on their experiences
  - Improve documentation of outputs → will improve documentation around analysis stage, including interpretation
  - Workshop at Translation UK, June 2020 → will create teaching materials to support the workshop

- **v2.0-beta:**
  - Coming soon! More easily discoverable & ready for testing...

Challenges: Testing Reliability

- **General Testing:**
  - Developing features in RiboViz → ongoing software testing, bug fixes, code reviews
  - Running new datasets
    - Multiple datasets run so far… *S. cerevisiae* & *C. albicans*
    - Wallace Lab Ribosome Profiling Data? TBC!
    - In March/April: Mouse data (Rutgers), Drosophila data (Leeds)
- **Methods Testing:**
  - Are analysis methods working? → better testing in the analysis section
  - Expected outputs? → working with Amanda Mok in the Lareau Lab to run RiboViz on statistically ‘realistic’ simulated ribosome profiling data

Back to The Question

The Question:

How to make ribosome profiling data reliable, comparable & easier to analyse?

The Answer?

RiboViz answers the key problems in ribosome profiling data processing & analysis...

… And the development team are working to make it more reliable and friendly to use, with easily comparable outputs.

… And when ribosome profiling data is **reliable, comparable & easy to analyse**?

?

… And when ribosome profiling data is **reliable, comparable & easy to analyse**?

… And when ribosome profiling data is **reliable, comparable & easy to analyse**?

Questions?

EXTRA SLIDES

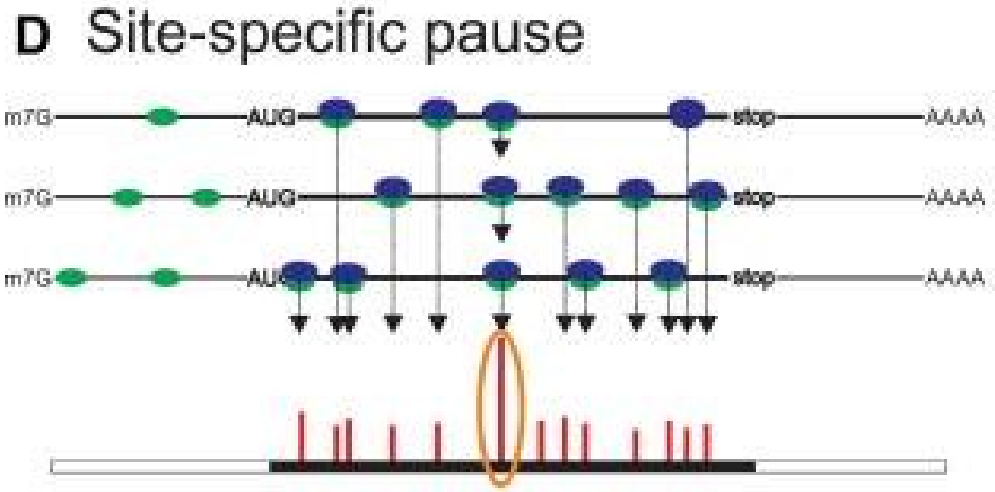
Run Times

- Test Vignette: 2 downsampled samples (94MB total)… 2m

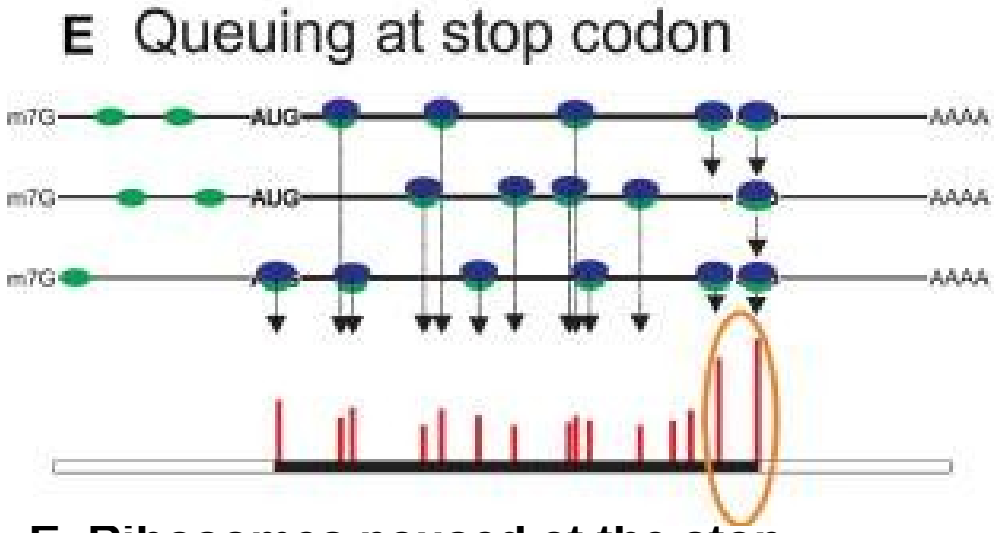
On Ubuntu 18.04 Linux, using 8 processors:

- G-Sc_2014: 8 samples (13.3GB total)… ~4hr
- M-Ca_2014: 3 samples (17GB total)... ~3hr 10m
- W-Sc_2016: 1 sample with UMIs (3.7GB)… ~4hr 30m

Pausing & Queuing



D. Site specific pause originating from ribosomes stalled within acORF (annotated coding ORF) at a specific location.



E. Ribosomes paused at the stop codon and queued upstream ribosomes

Andreev et al. (2017). "Insights into the mechanisms of eukaryotic translation gained with ribosome profiling". Nucleic Acids Research. 45 (2): 513–526. doi:10.1093/nar/gkw1190

RiboViz Technical Bits

- Bash, Python & R co-ordinating multiple externally- & internally- developed tools.
  - hisat2-build: build rRNA and ORF indices.
  - cutadapt: cut adapters.
  - hisat2: align reads.
  - riboviz.tools.trim_5p_mismatch: trim 5' mismatches from reads and remove reads with more than a set number of mismatches (local script, in riboviz/tools/).
  - umi_tools (extract, dedup, group): extract barcodes and UMIs, deduplicate reads and group reads.
  - riboviz.tools.demultiplex_fastq: demultiplex multiplexed files (local script, in riboviz/tools/).
  - samtools (view, sort, index): convert SAM files to BAM files and index.
  - bedtools (genomecov): export transcriptome coverage as bedgraphs.
  - bam_to_h5.R: convert BAM to compressed H5 format (local script, in rscripts/)
  - generate_stats_figs.R: generate summary statistics, analyses plots and QC plots (local script, in rscripts/)
  - collate_tpms.R: collate TPMs across samples (local script, in rscripts/)
  - riboviz.tools.count_reads: count the number of reads (sequences) processed by specific stages of the workflow (local script, in riboviz/tools/).