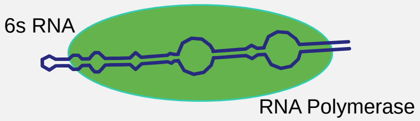
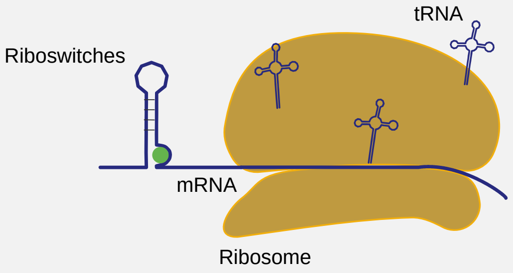
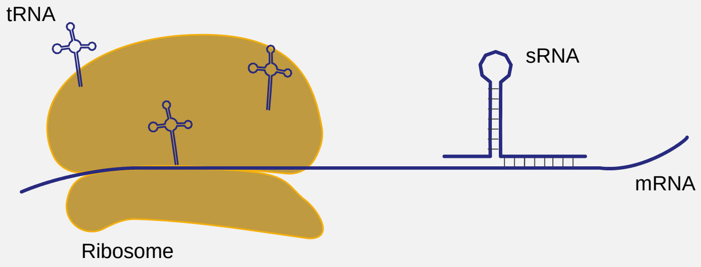


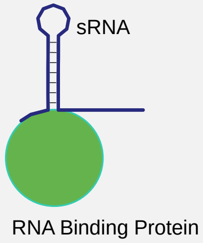


COMPARATIVE SRNA ANALYSIS

Thomas Nicholson

University of Otago

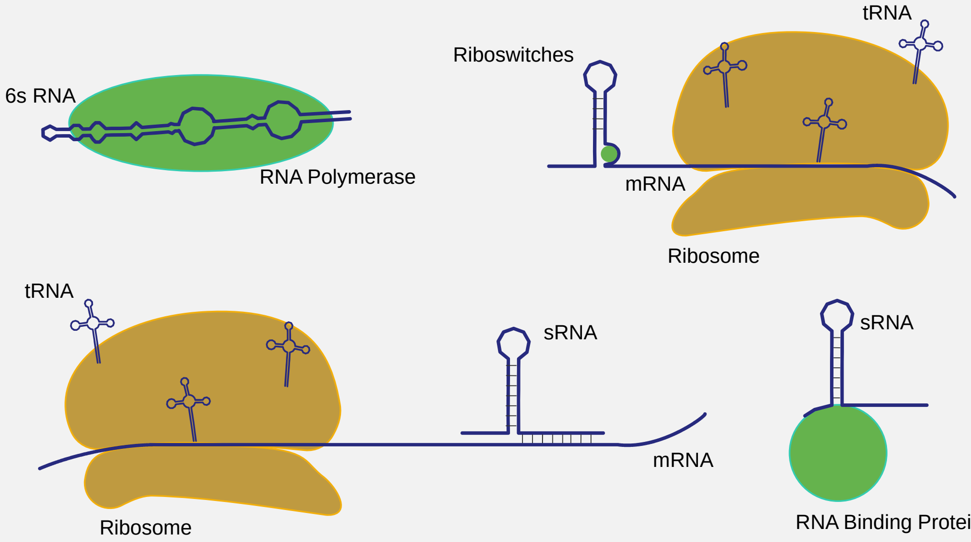


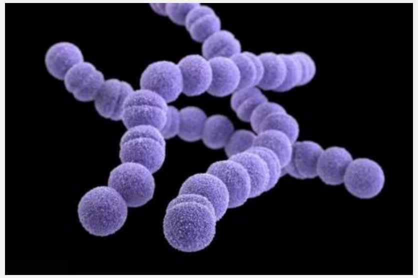


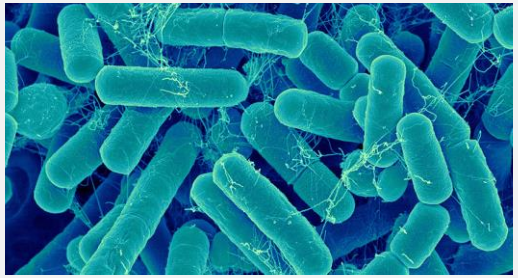
OVERVIEW

Bacterial ncRNAs are critical for a wide range of cell functions

- Transcription/translation
  - rRNA, tRNA, 6sRNA etc.
- Antiviral response
  - CRISPR-cas
- Regulation
  - Riboswitches, sRNAs binding to mRNA etc.
- Virulence

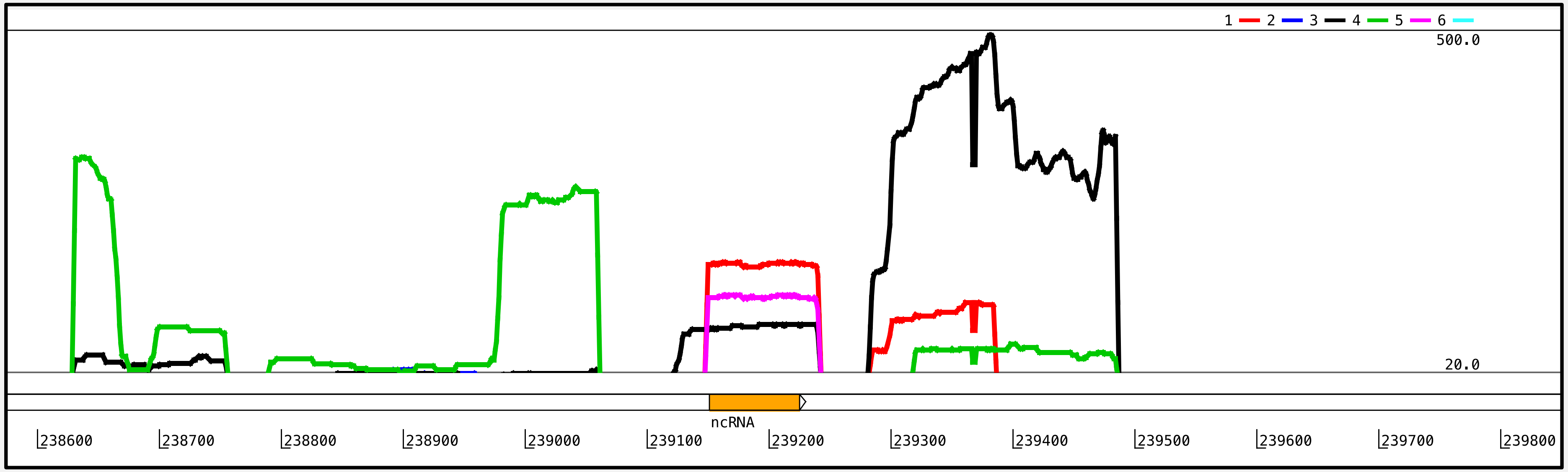




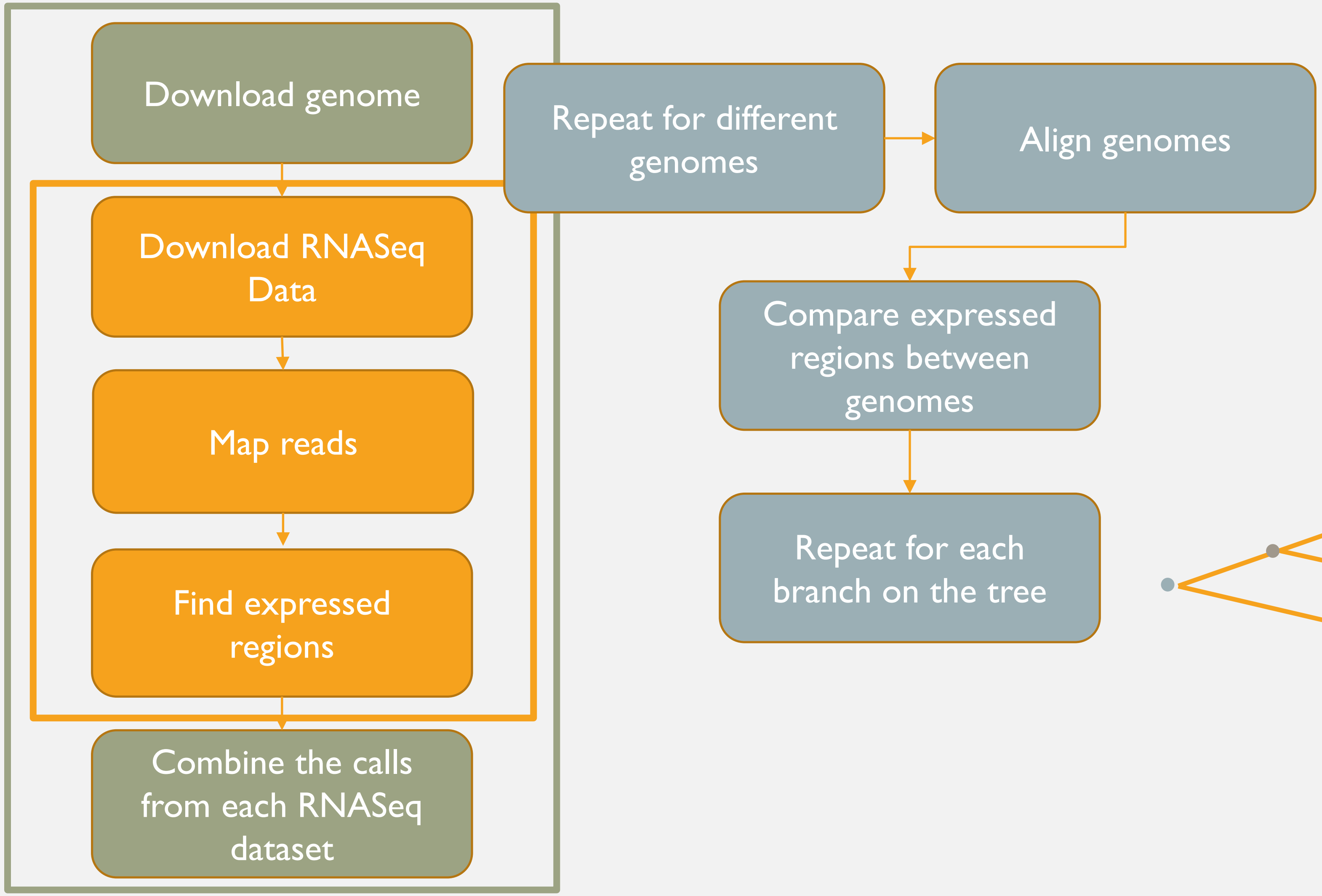


IDENTIFYING SRNAS

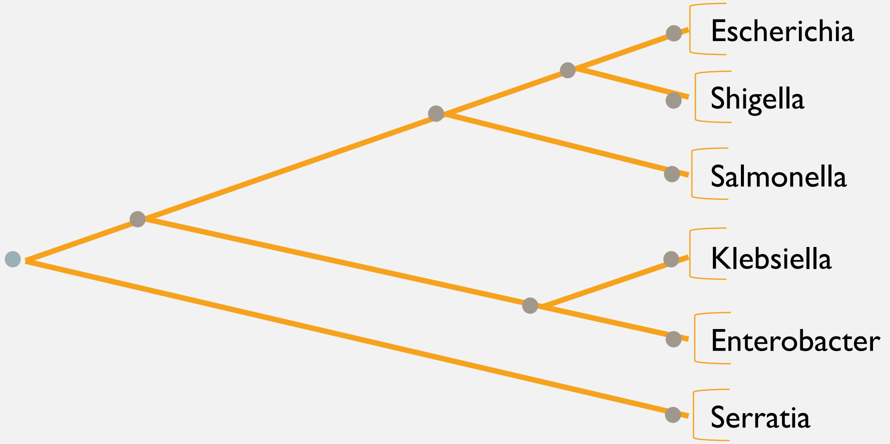
- Small non-coding RNAs (sRNA)
- Searching for similar sequences to known RNA families
  - Does not require expression
  - Requires the RNA family to be known
- Using RNASeq data to find regions that are expressed
  - Will find novel RNAs
  - Conditions of the experiment may change expression



METHOD

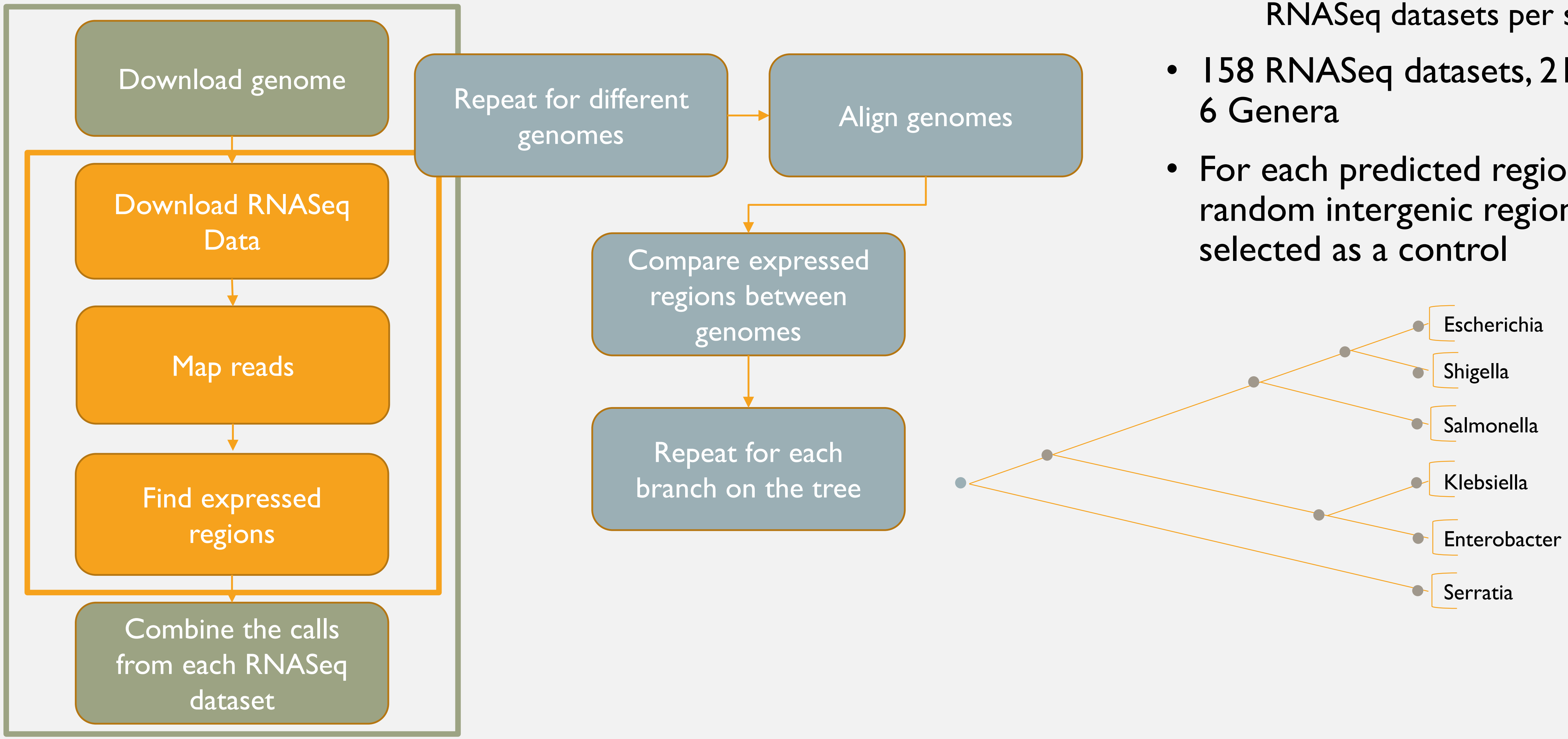

Download genome
Download RNASeq Data
Map reads
Find expressed regions
Combine the calls from each RNASeq dataset
Repeat for different genomes
Align genomes
Compare expressed regions between genomes
Repeat for each branch on the tree

- Selected data from available RNASeq datasets
  - Used a clade that provided multiple strains with at least 5 RNASeq datasets per strain
- 158 RNASeq datasets, 21 Strains, 6 Genera
- For each predicted region, a random intergenic region was selected as a control



METHOD

- Selected data from available RNASeq datasets
  - Used a clade that provided multiple strains with at least 5 RNASeq datasets per strain
- 158 RNASeq datasets, 21 Strains, 6 Genera
- For each predicted region, a random intergenic region was selected as a control



Download genome

Download RNASeq Data

Map reads

Find expressed regions

Combine the calls from each RNASeq dataset

Repeat for different genomes

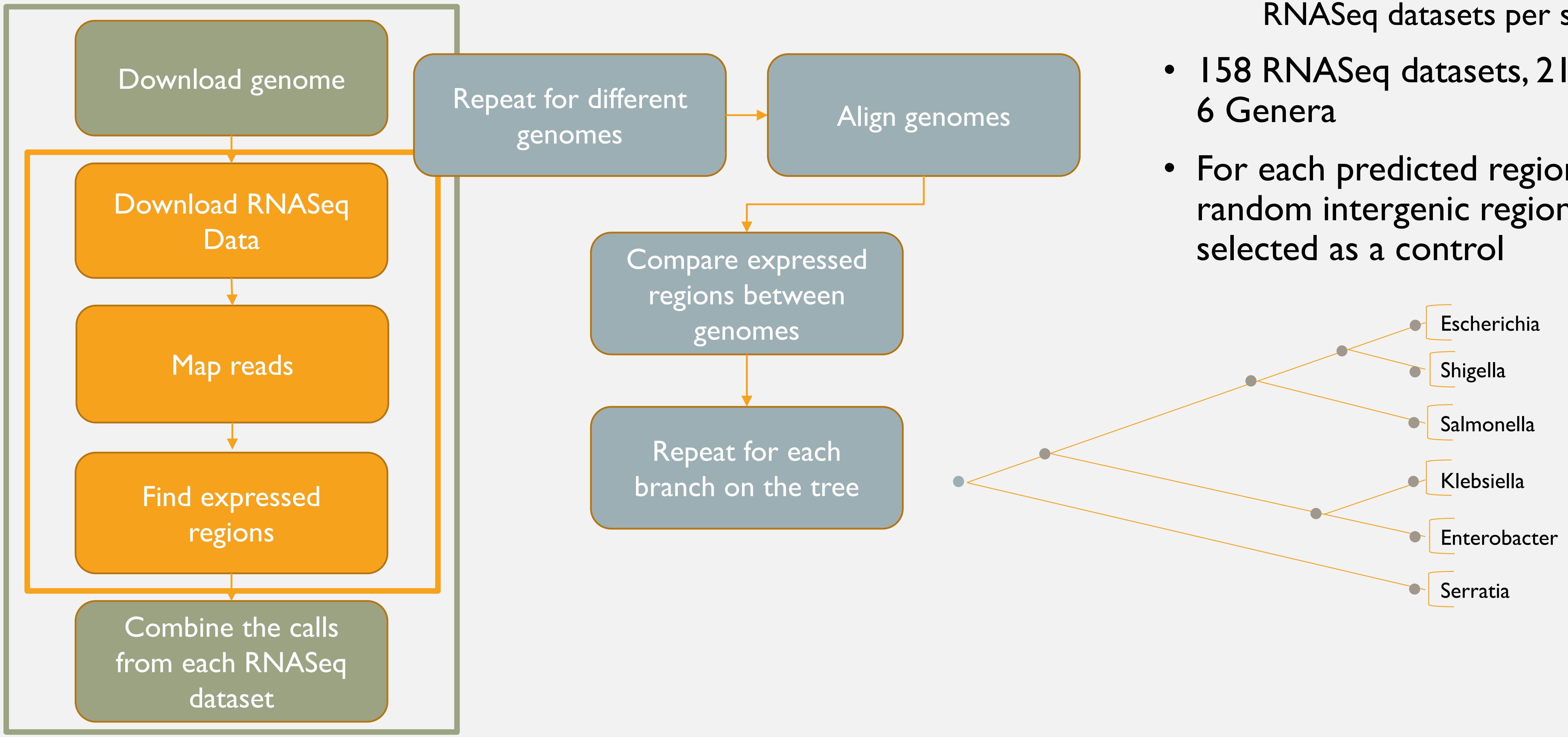
Align genomes

Compare expressed regions between genomes

Repeat for each branch on the tree

METHOD

- Selected data from available RNASeq datasets
  - Used a clade that provided multiple strains with at least 5 RNASeq datasets per strain
- 158 RNASeq datasets, 21 Strains, 6 Genera
- For each predicted region, a random intergenic region was selected as a control



Download genome

Download RNASeq Data

Map reads

Find expressed regions

Combine the calls from each RNASeq dataset

Repeat for different genomes

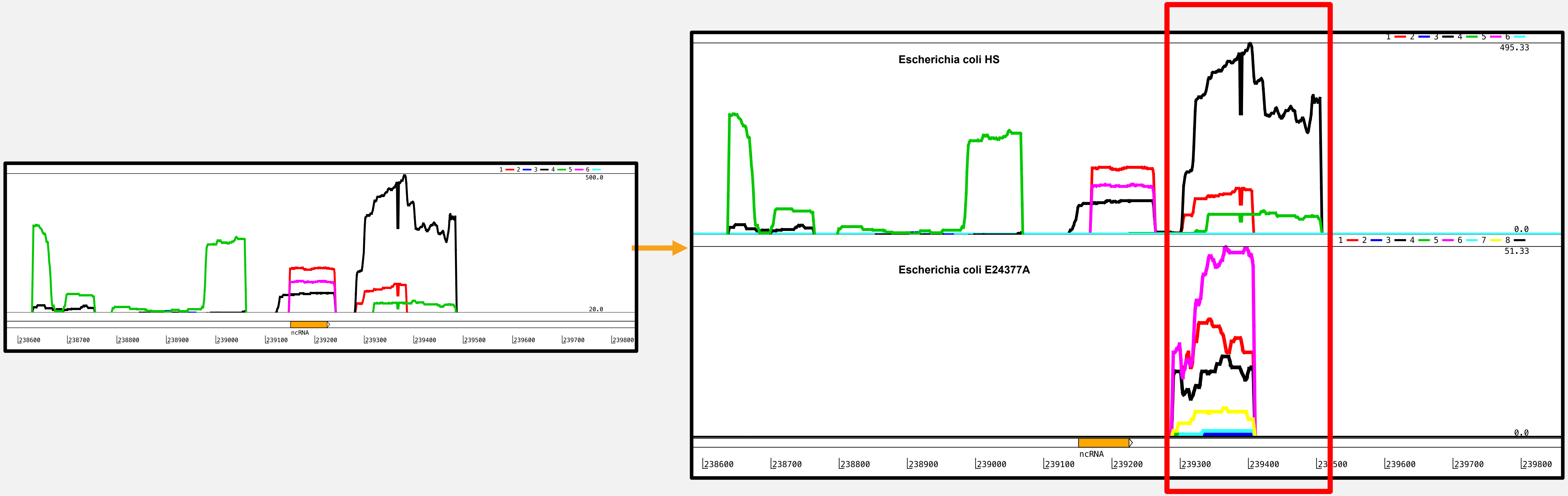
Align genomes

Compare expressed regions between genomes

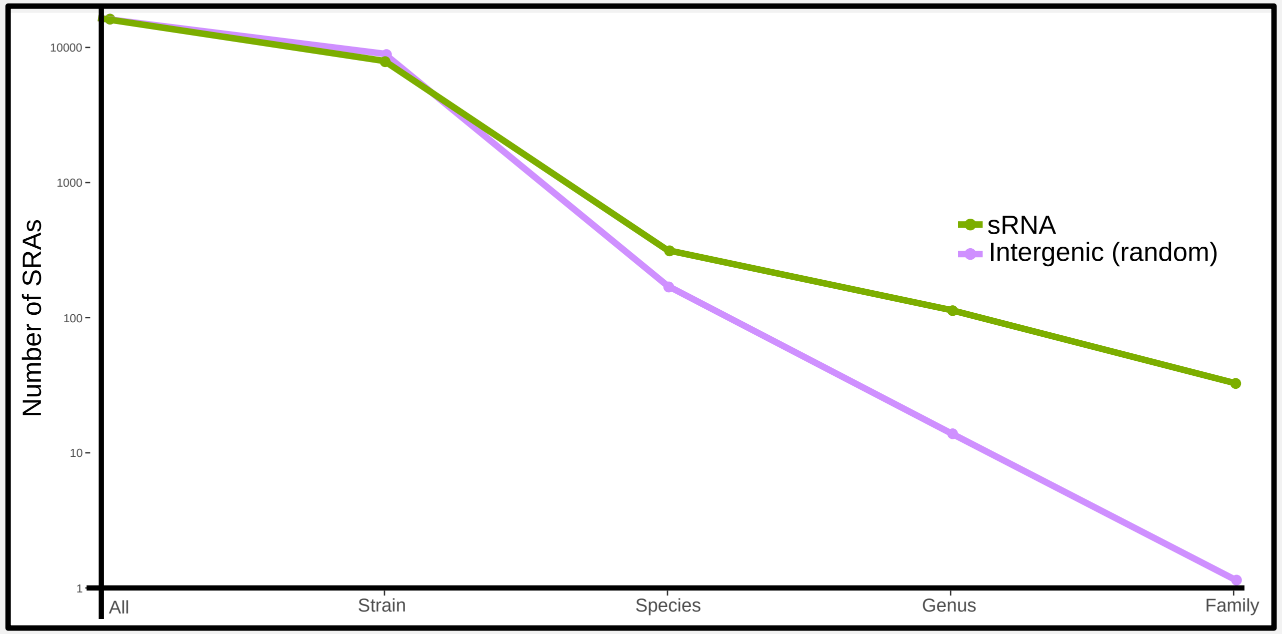
Repeat for each branch on the tree

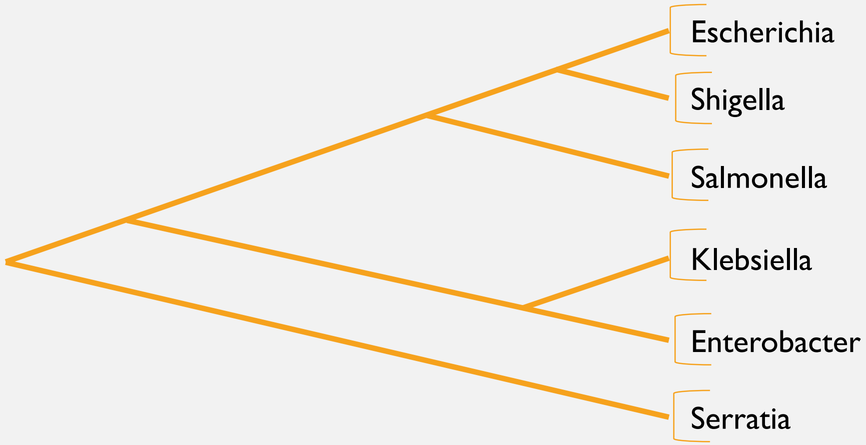
IDENTIFYING SRNAS

- Identify regions of expression in multiple species
- Align genomes
- Compare regions of expression and look for conservation of expression



RESULTS

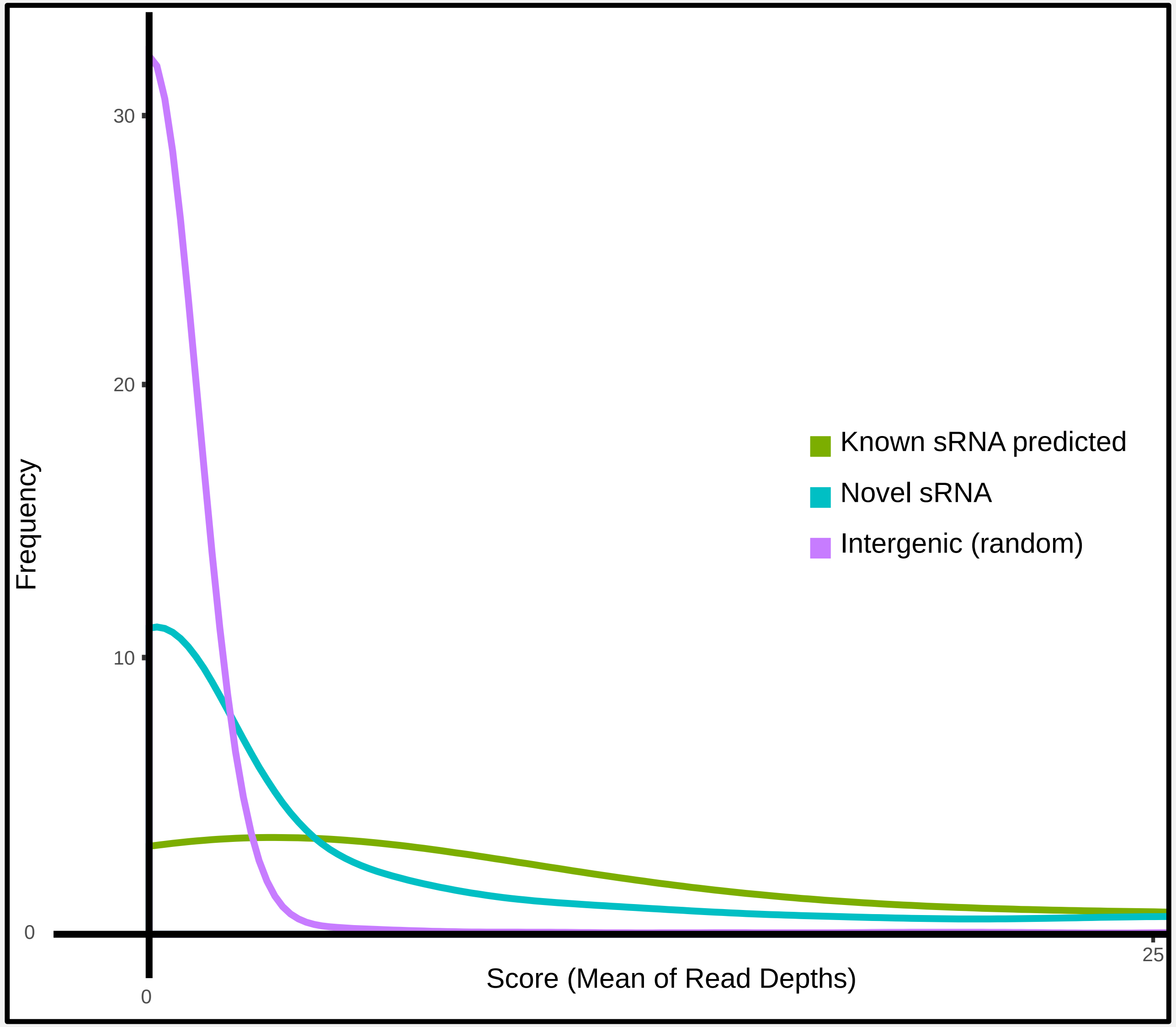


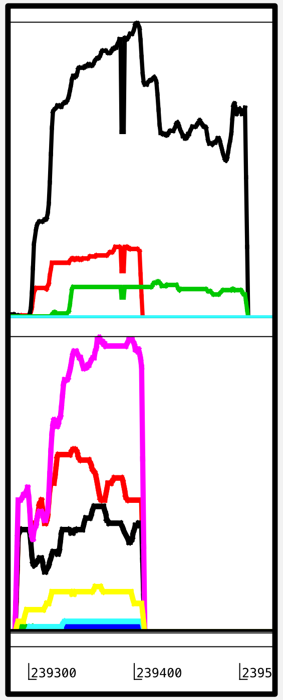


- 6,984 putative sRNAs
  - 332.6 sRNAs per strain
- 10,786 known ncRNAs
  - Previously annotated
  - Predicted with rFAM models
- 514 conserved sRNAs retained
  - 65 sRNAs conserved across different families

RESULTS

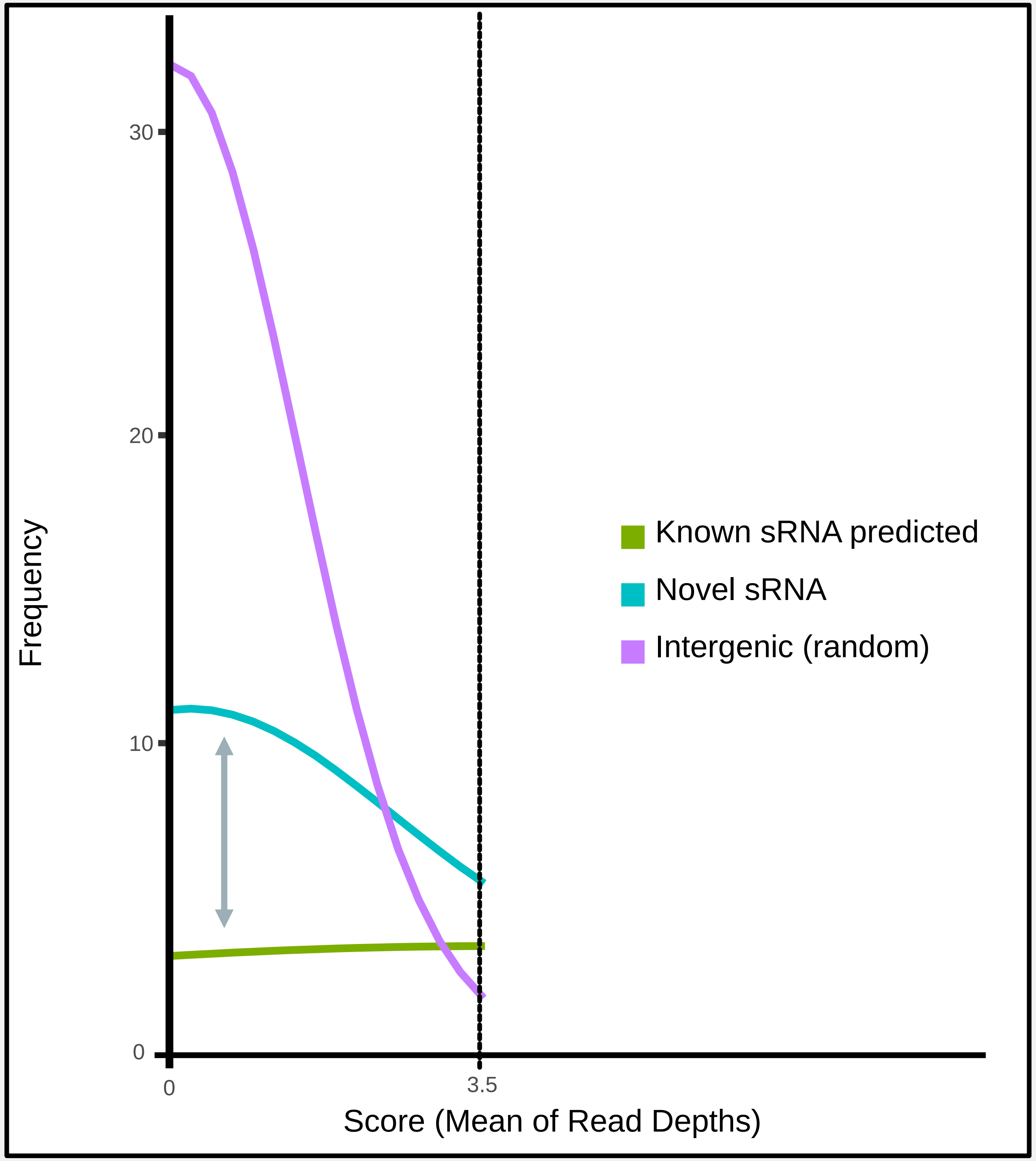
- Used known sRNAs, conserved expressed regions and randomly selected intergenic regions
- Calculated the mean of the read depths for each accession
  - Mean along an sRNA for individual RNASeq dataset
  - Average of all the individual RNASeq dataset means





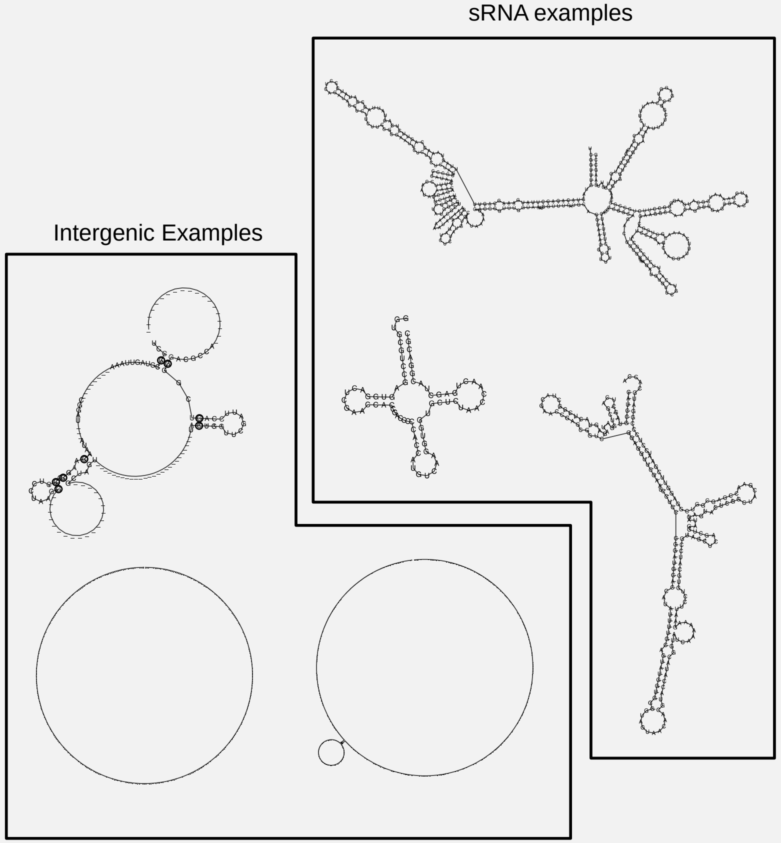
SIGNAL OR NOISE?

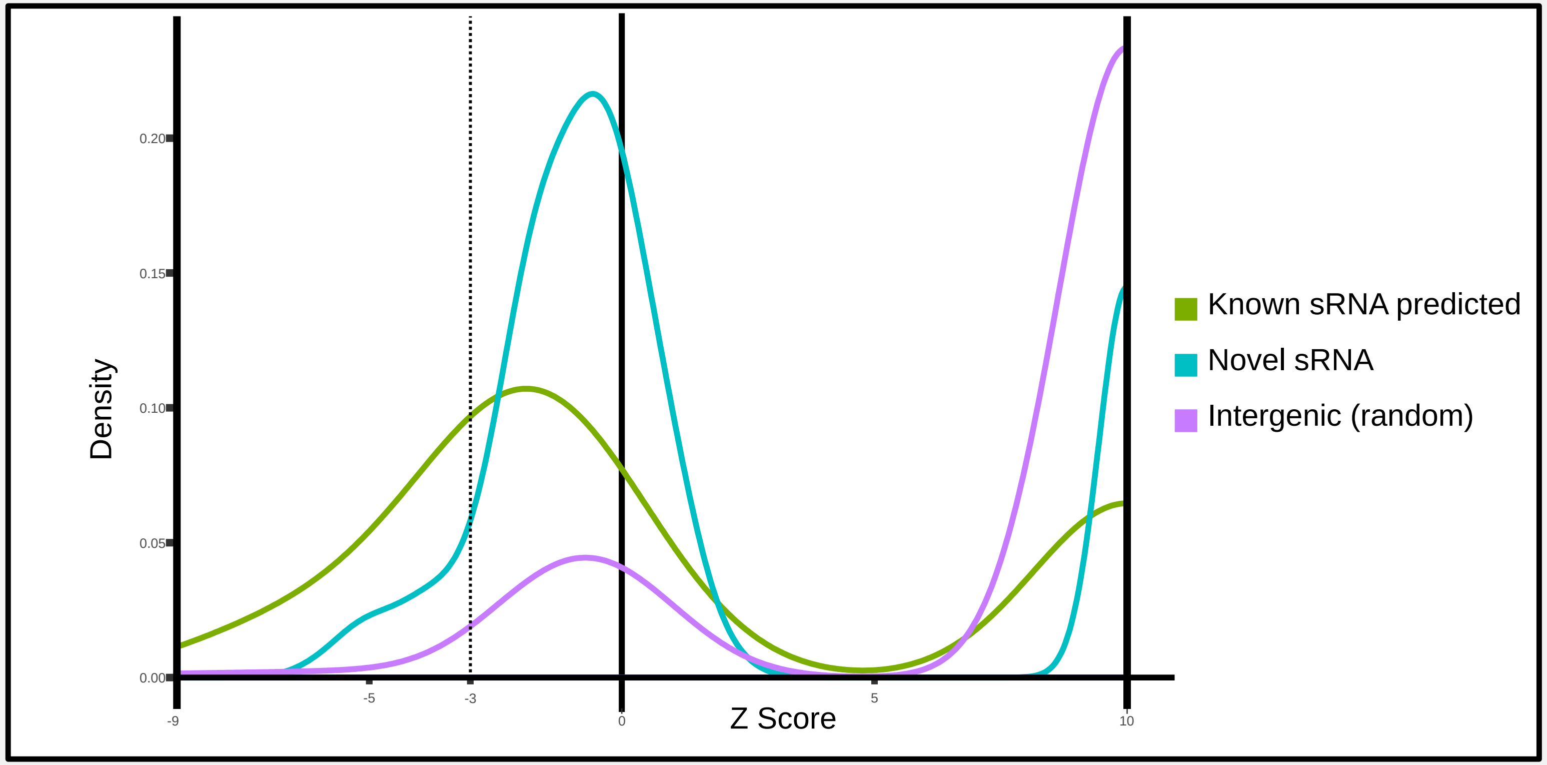
- A score of < 3.5 includes 95% of the random data
- 18.4% of the known sRNAs are in the same region
- 49.2% of the novel sRNAs in the the same region



SECONDARY STRUCTURE

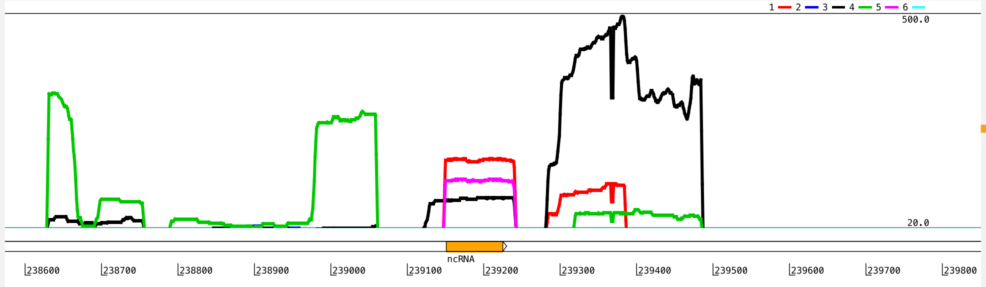
- Sequences folding with significant z-score
  - 19.9% of known sRNAs
  - 11.49% of conserved sRNAs
  - 7.62% of non-conserved sRNAs
  - 1.24% of intergenic (random) sequences

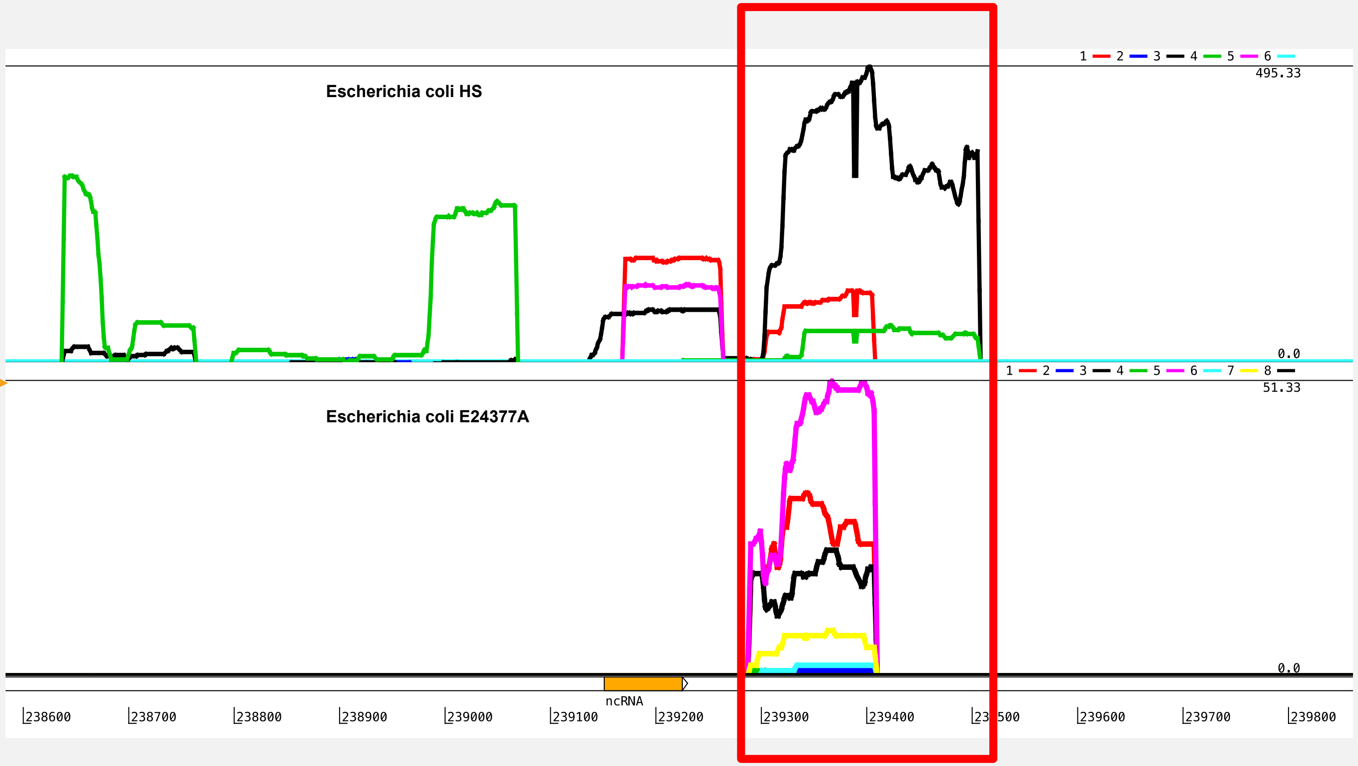




CONCLUSION

- Predicting sRNAs for single genomes can be difficult
- 30% of the predicted regions appear to be noise
- Using a comparative approach can help improve the signal to noise
- There is a need for RNASeq data targeting a wider range of bacteria





ACKNOWLEDGMENTS

- Paul Gardner, Beth Jose and the Gardner lab.
- Funding: Bio-protection CoRE