BROAD INSTITUTE

Mapping Bacterial Transcriptomes from RNA-Seq Expression Data

Andrew Baxter

Broad Institute of MIT and Harvard

The importance of mapping bacterial transcriptomes

Defining transcribed regions and how they change in different environments gives us clues about bacterial physiology and evolution

Using RNA-Seq to map bacterial transcriptomes

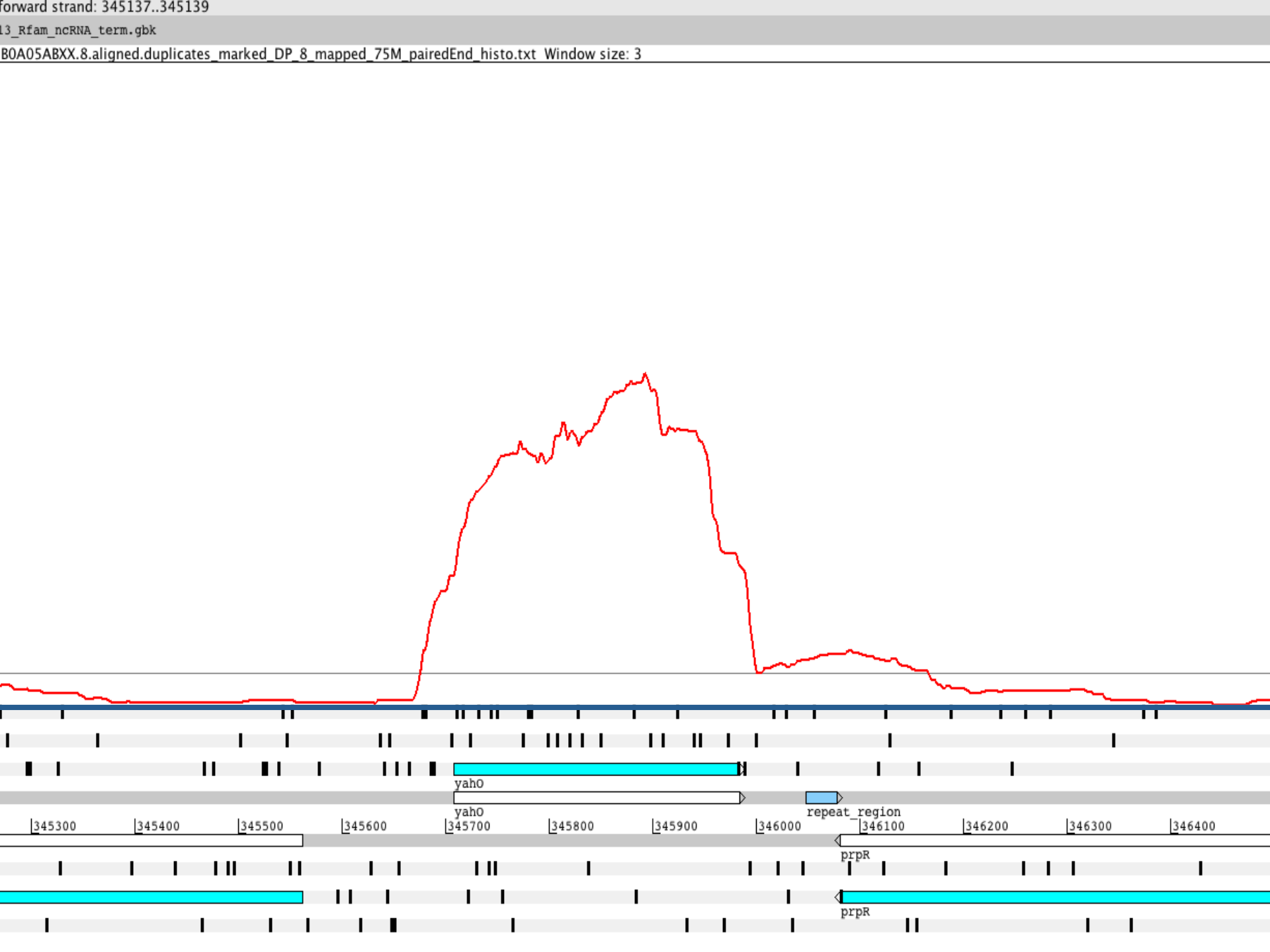
RNA-Seq

- Convert RNA into cDNA libraries
- Generates sequencing "reads"
- Reads are aligned to a reference genome (when available) with nucleotide resolution
- Number of reads in region correlates to expression of RNA from region

forward strand: 345137..345139

13_Rfam_ncRNA_term.gbk

B0A05ABXX.8.aligned.duplicates_marked_DP_8_mapped_75M_pairedEnd_histo.txt Window size: 3

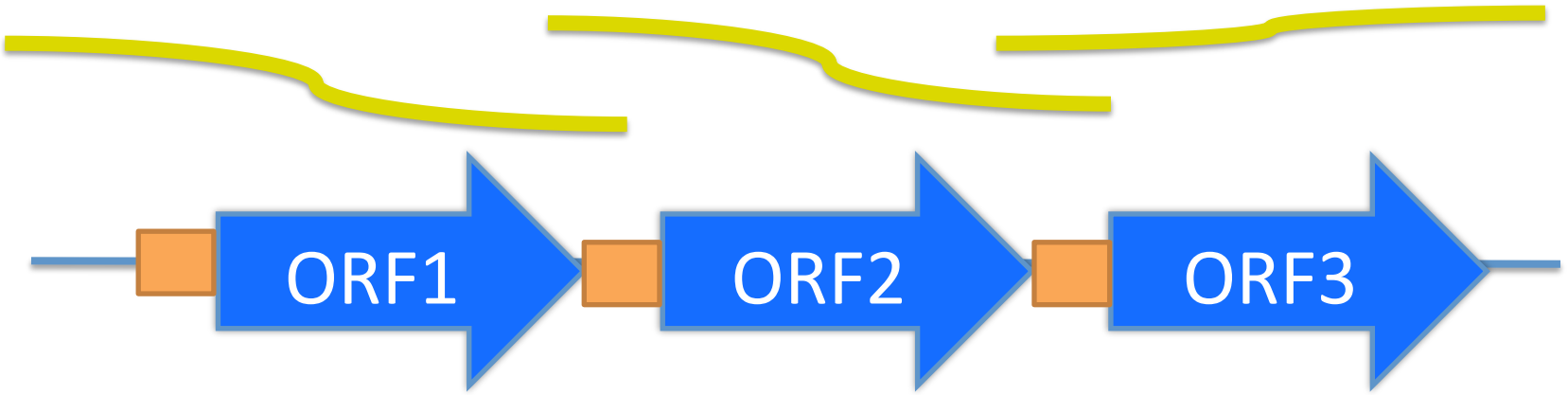


The challenges of mapping bacterial transcripts by RNA-Seq

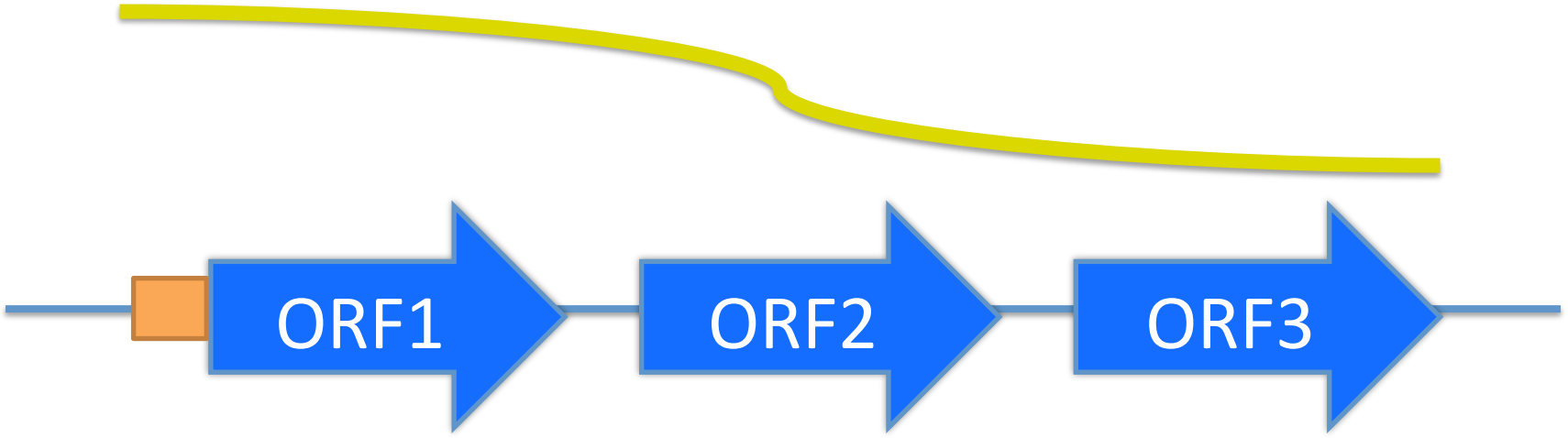
- Biases introduced during cDNA library construction
- Complexity of bacterial transcriptomes

The complexity of bacterial transcriptomes

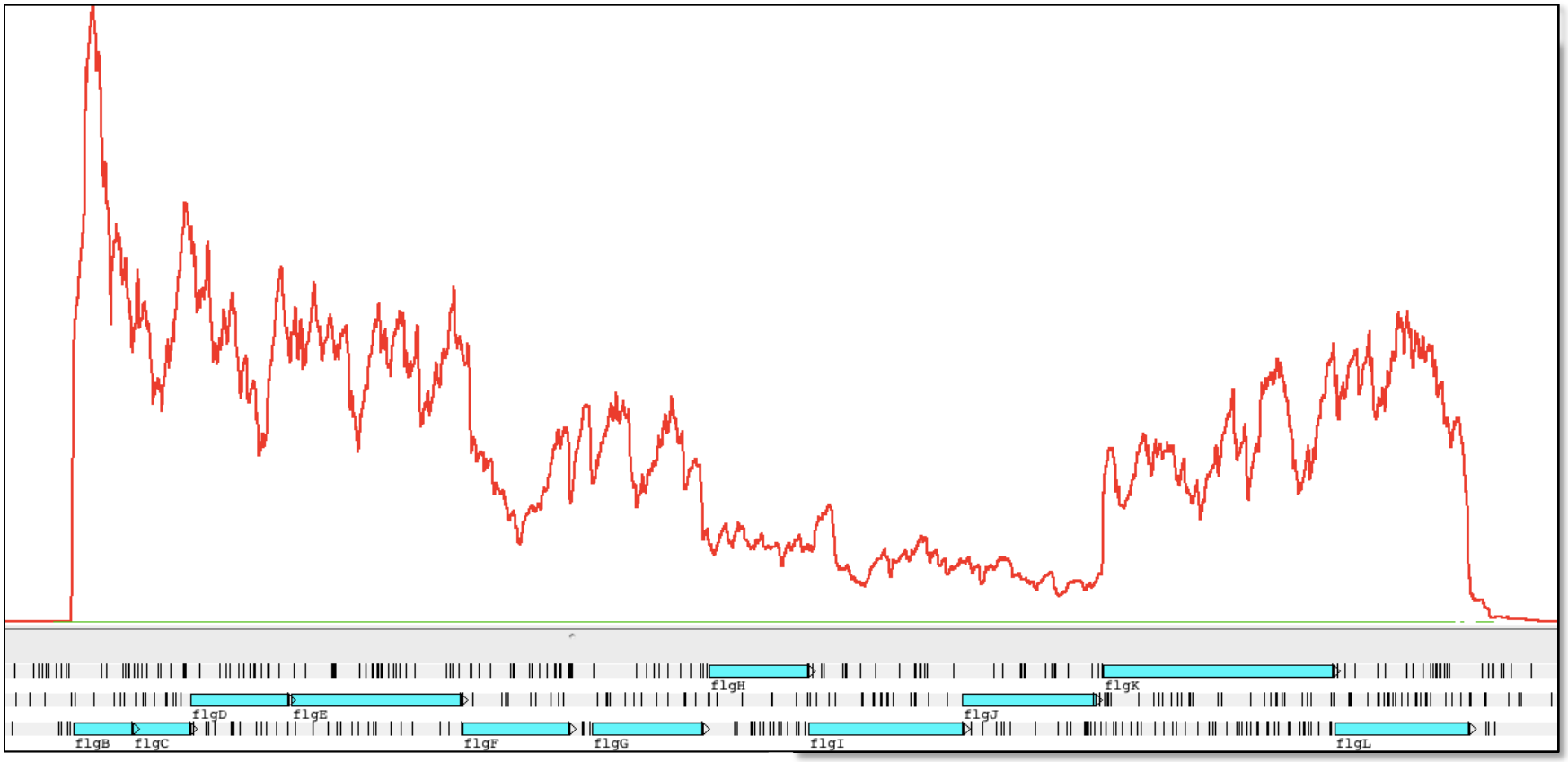
Many adjacent genes in bacteria are independently transcribed…



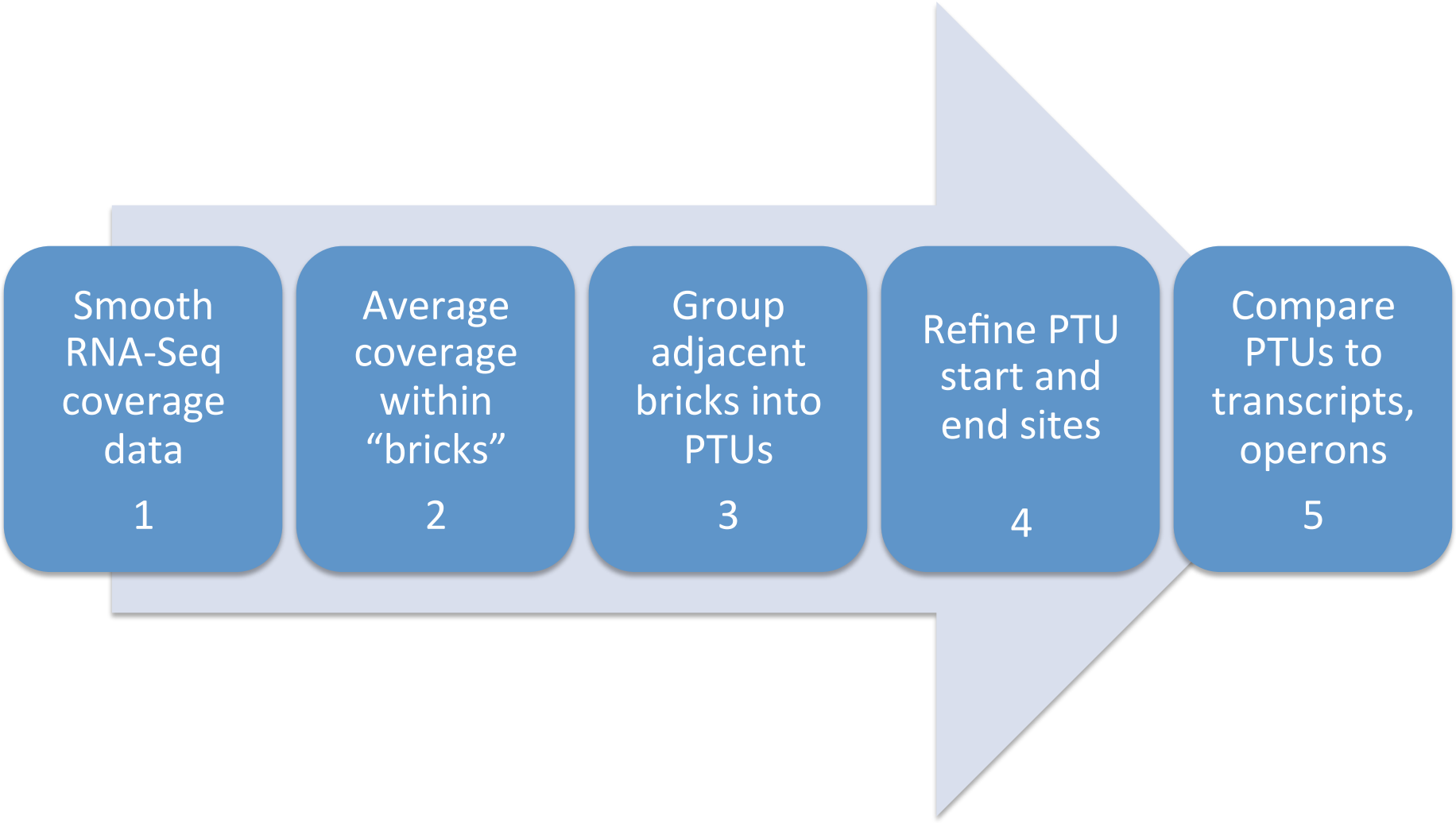
Many others are coordinately transcribed as part of a single mRNA (operons)



The challenges of mapping bacterial transcripts by RNA-Seq: an example



Developing an algorithm for using RNA-Seq data to define *E. coli* putative transcription units (PTUs)



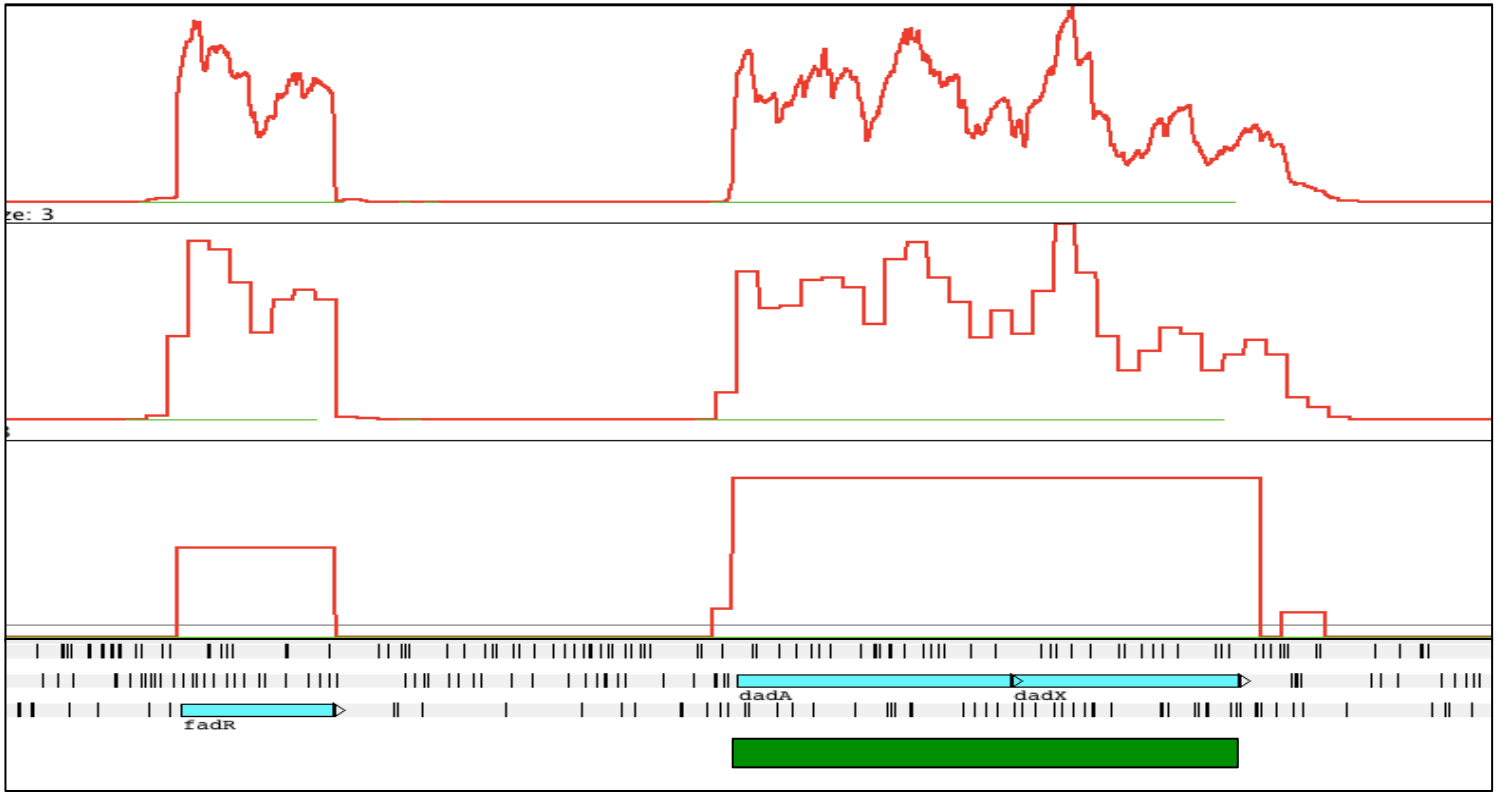
The algorithm in action

Known transcription units identified, the algorithm worked

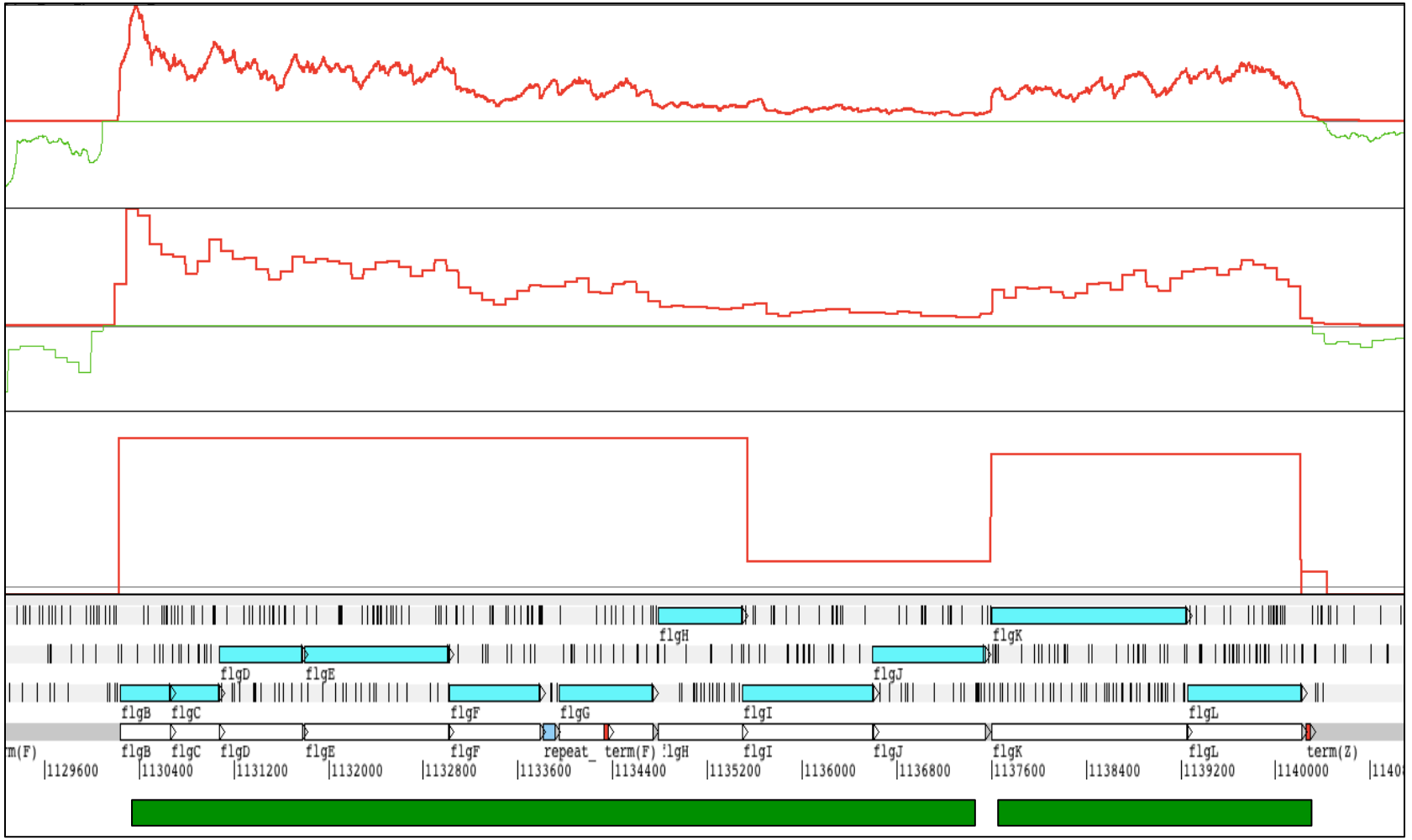
Raw data

Data smoothed, converted to bricks

Bricks grouped, boundaries refined



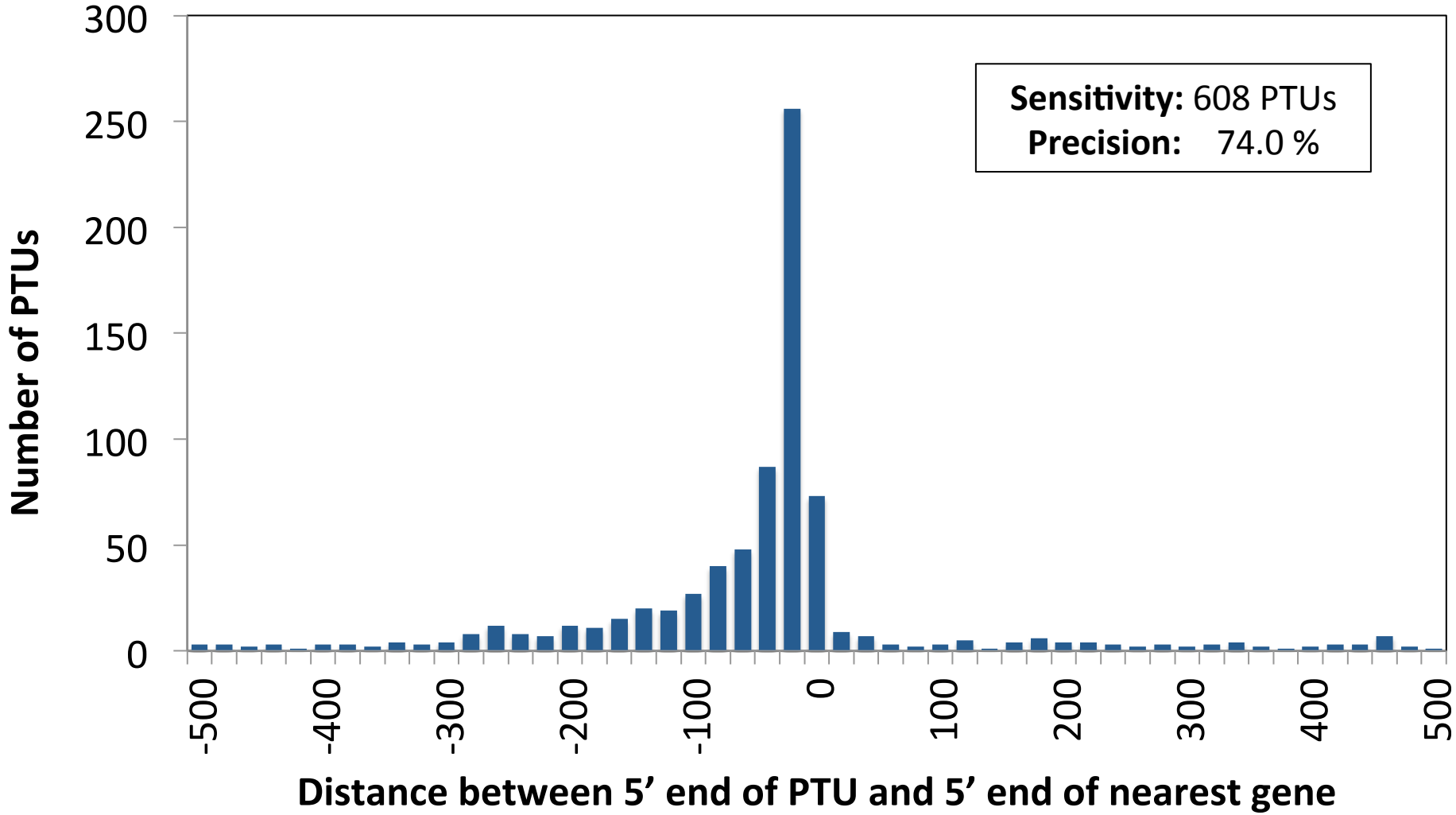
Operon missed, so close yet so far…



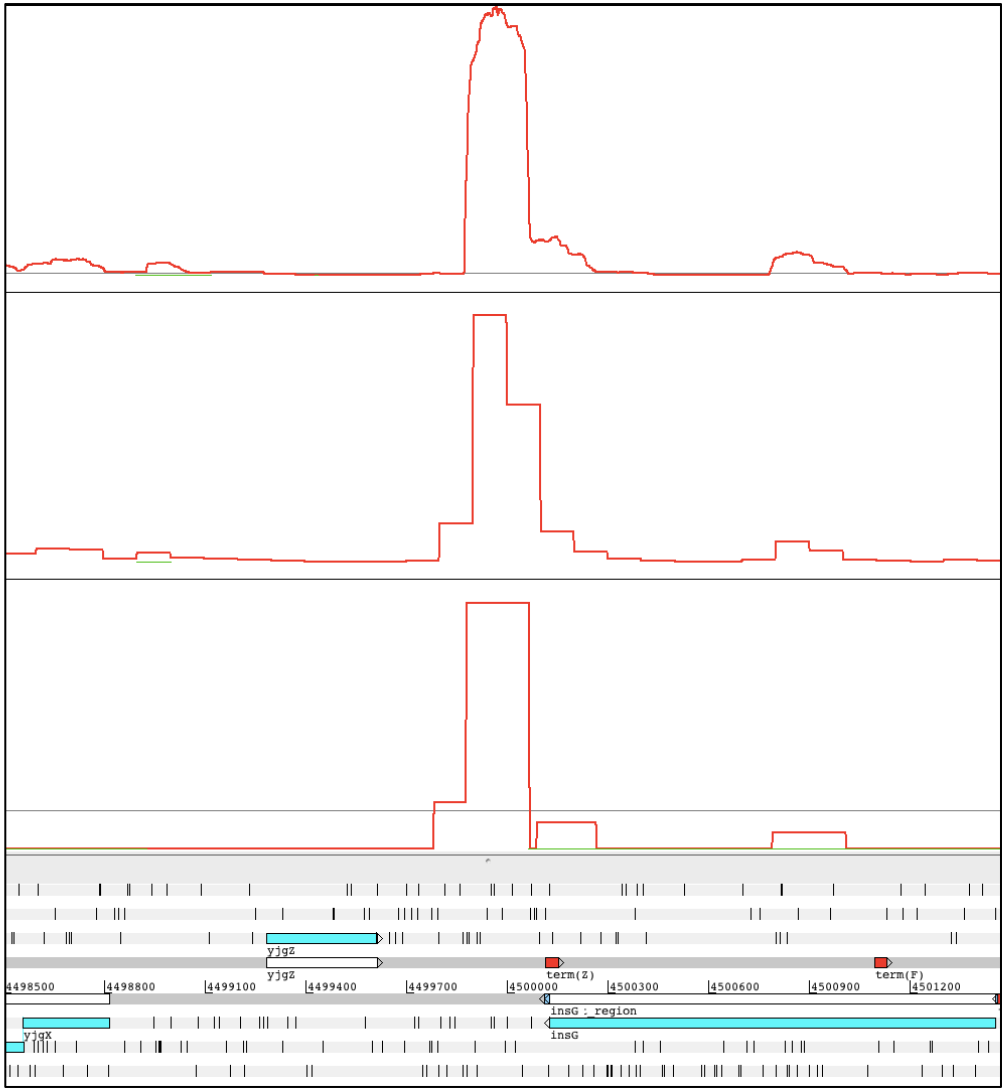
How do we systematically measure the performance of our algorithm?

- Success of the algorithm is measured by sensitivity and precision.
- **Sensitivity** - the total number of PTUs whose 5' end is between 0 and 200bp upstream of the nearest annotated gene
- **Precision** – percent of total number of PTUs identified whose 5' end is between 0 and 200bp upstream of the nearest annotated gene

Measuring the success of our algorithm



Candidates for previously unknown transcripts: an example



Conclusions

- We have developed an algorithm for mapping bacterial transcriptomes with RNA-Seq data
- The algorithm successfully identified many previously annotated transcription units, failed to identify others, and identified numerous novel PTUs

Going Forward

- Further fine tuning of the algorithm for higher precision and sensitivity
- Testing how decreasing the depth of RNA-Seq data effects the algorithm's performance
- Testing the algorithm on RNA-Seq data from different bacteria

BROAD INSTITUTE

Acknowledgments

Jonathan Livny
Brian Haas
Melissa Chin
Moran Yassour

Special Thanks to
SRPG
Bruce Birren
Eboney Smith
Brandon Ogbunugafor
Francie Latour