riboSeed: leveraging genomic architecture to assemble across ribosomal regions

Microsoc Postgrad Seminar Series

Nick Waters

May 12, 2017

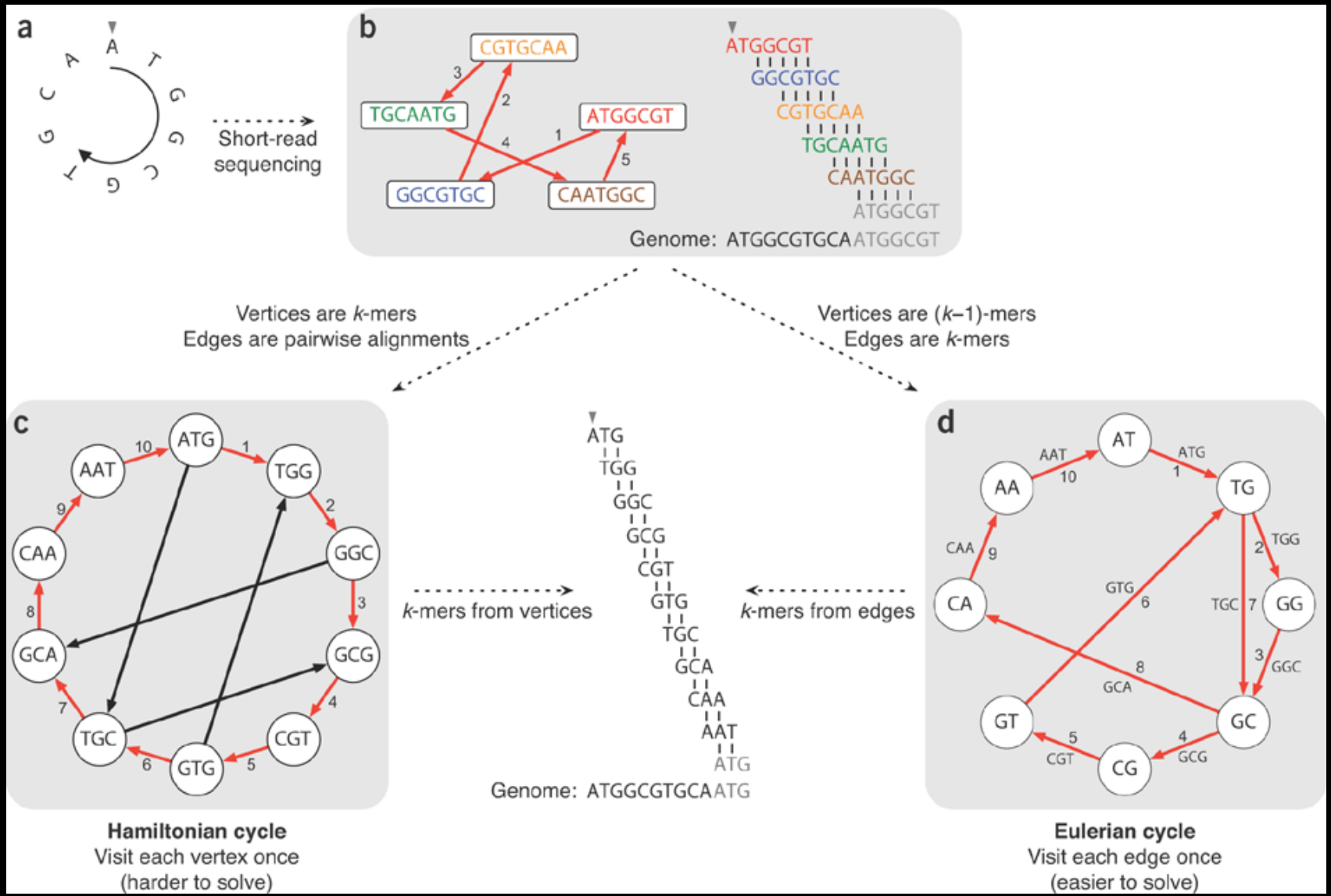
Department of Microbiology
School of Natural Sciences
National University of Ireland Galway

Outline

- De Bruijn Graphs and short read assembly
- Shortcoming of Short Read Assembly
- Genome Finishing Strategies
- riboSeed
- Challenges

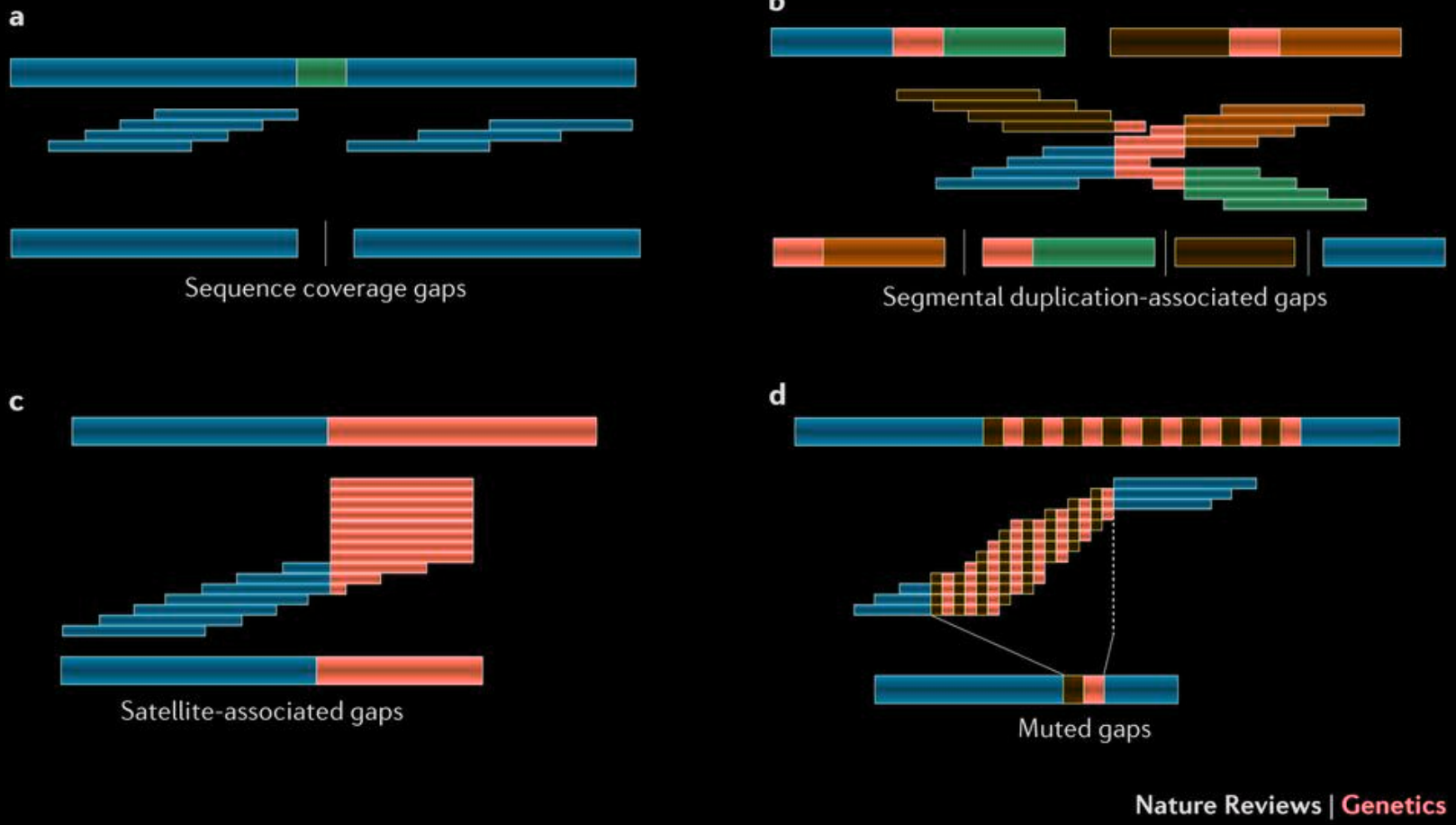
Short Ready Assembly: Background

de Bruijn Graphs and Eulerian Paths



Source: Compeau2011

Problems



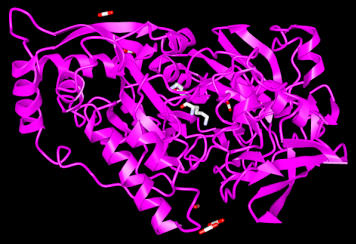
Source:Chaisson2015

Problem Regions

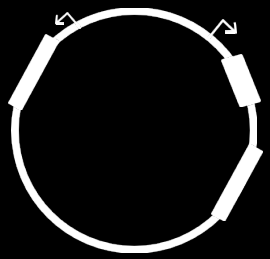
Repeated regions cannot be resolved with kmers shorter than the repeat!

Problem Regions

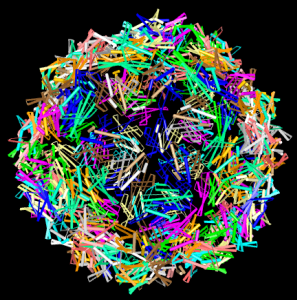
Repeated regions cannot be resolved with kmers shorter than the repeat!



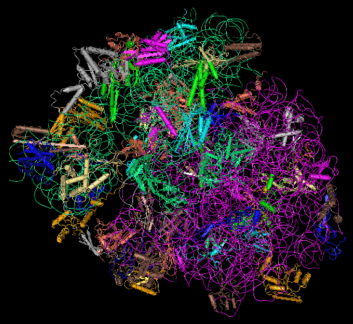
Transporters



Ω Plasmids



Prophages



Ribosomes

Is it Hopeless?

Genome Finishing

method	benefits	drawbacks
PCR + Sanger	it works	its difficult
re-sequencing	improve coverage	issues with repeats
long reads	solves repeats	cost, availability
reference assisted	easy to perform	not reliable
genome finishing tools	many options	many options

Genome Finishing Tools

Tool	Reference	Method Summary
GapFiller	[Boetzer2012]	utilize paired end reads
GapCloser/IMAGE	[Luo2012], [Tsai2010]	iteratively maps reads to contigs
CloG	[Yang2011]	uses trimmed de novo contigs in hybrid assembly
FGap	[Piro2014,Guizelini2016]	uses BLAST to find potential gap closures
GFinisher	[Guizelini2016]	uses GC-skew to refine assemblies
GapFiller	[Nadalin2012]	uses a hash-based method to produce “long-reads”
CONTIGuator	[Galardini2011]	generates a contig map and PCR primer sets to validate in the lab
Konnector	[Vandervalk2015]	uses a Bloom filter representation of a de Bruijn graph

Possible Solution

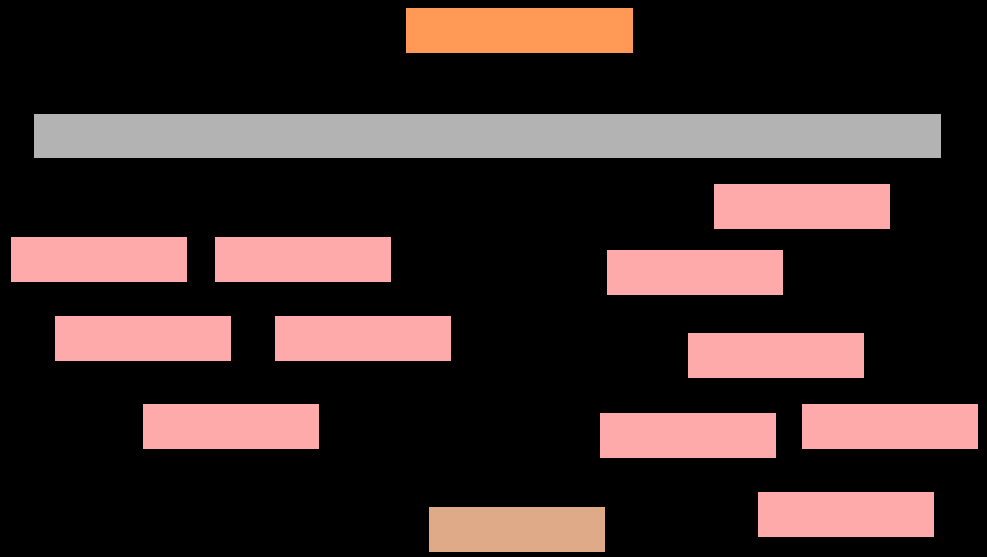
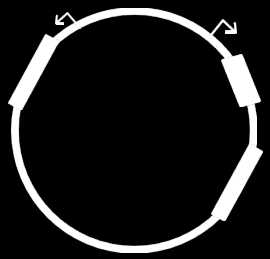


Figure: Bridge Reconstruction. Pink fragments are reads. Grey shows the gene of interest with interupted coverage. Orange fragemnt is a pseudoread generated from this situation under the hypothesis that the beige fragment exists but is underrepresented

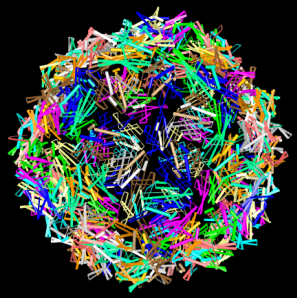
Problem Regions

Repeated regions cannot be resolved with kmers shorter than the repeat!

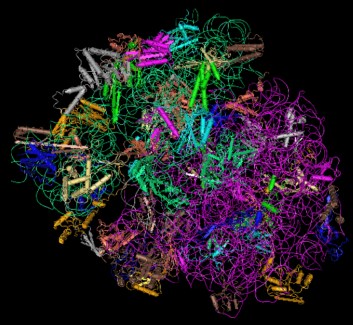
Transporters



Ω Plasmids



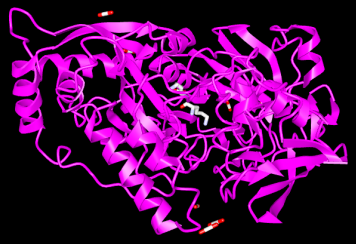
Prophages



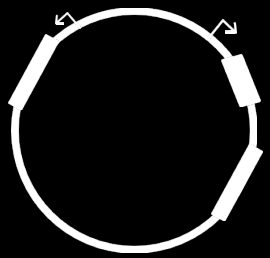
Ribosomes

Problem Regions

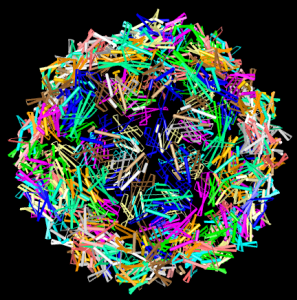
Repeated regions cannot be resolved with kmers shorter than the repeat!



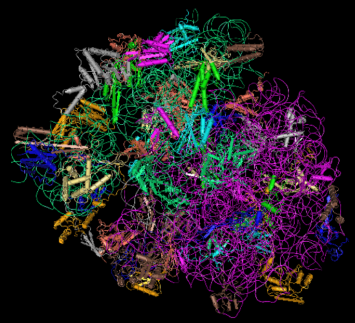
Transporters



Ω Plasmids



Prophages



Ribosomes

rDNA

rDNA: ribosomal DNA operon

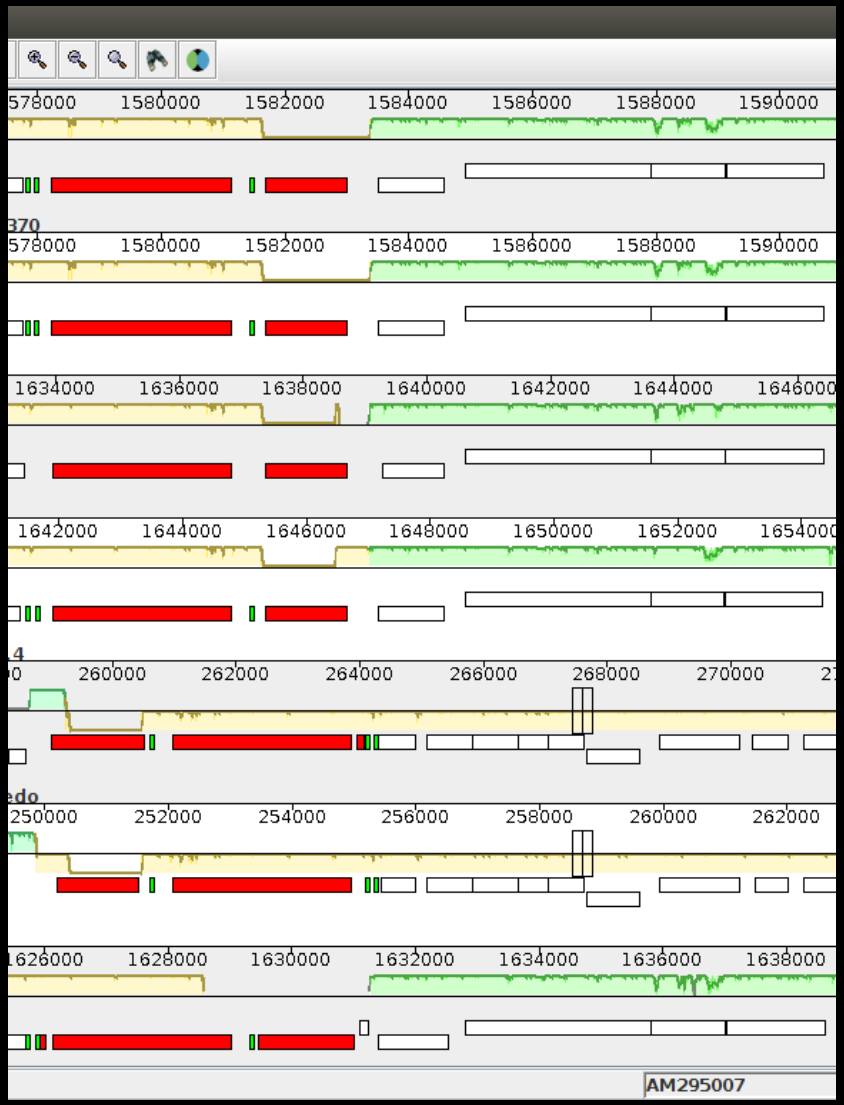
- Prokaryotes: 16S, 23S, 5S
- Conserved within taxa
- Repeated within the genome (1x to >14x)

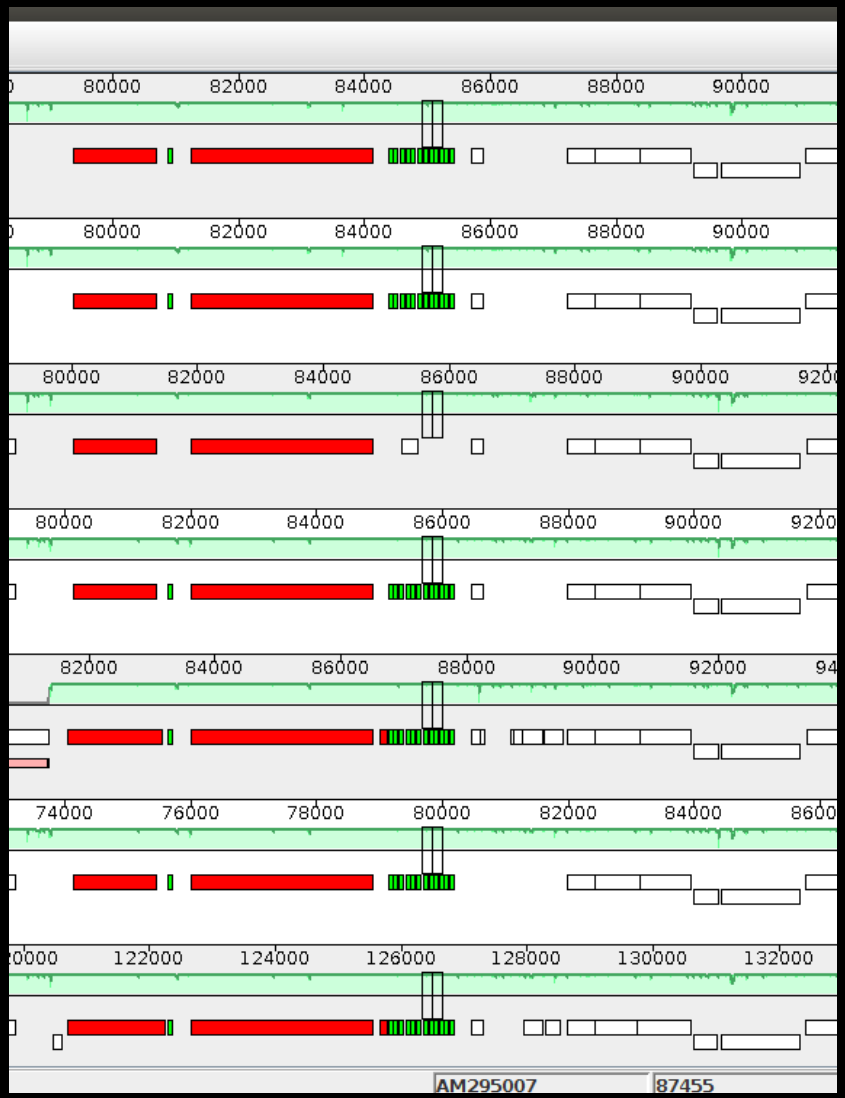
Hypotheses

1. Since the rDNA structure is conserved within taxa, rDNA flanking regions may be conserved
2. Regions flanking the rDNA region will be unique within genomes
3. If flanking regions are unique, they can be used to build “long reads”

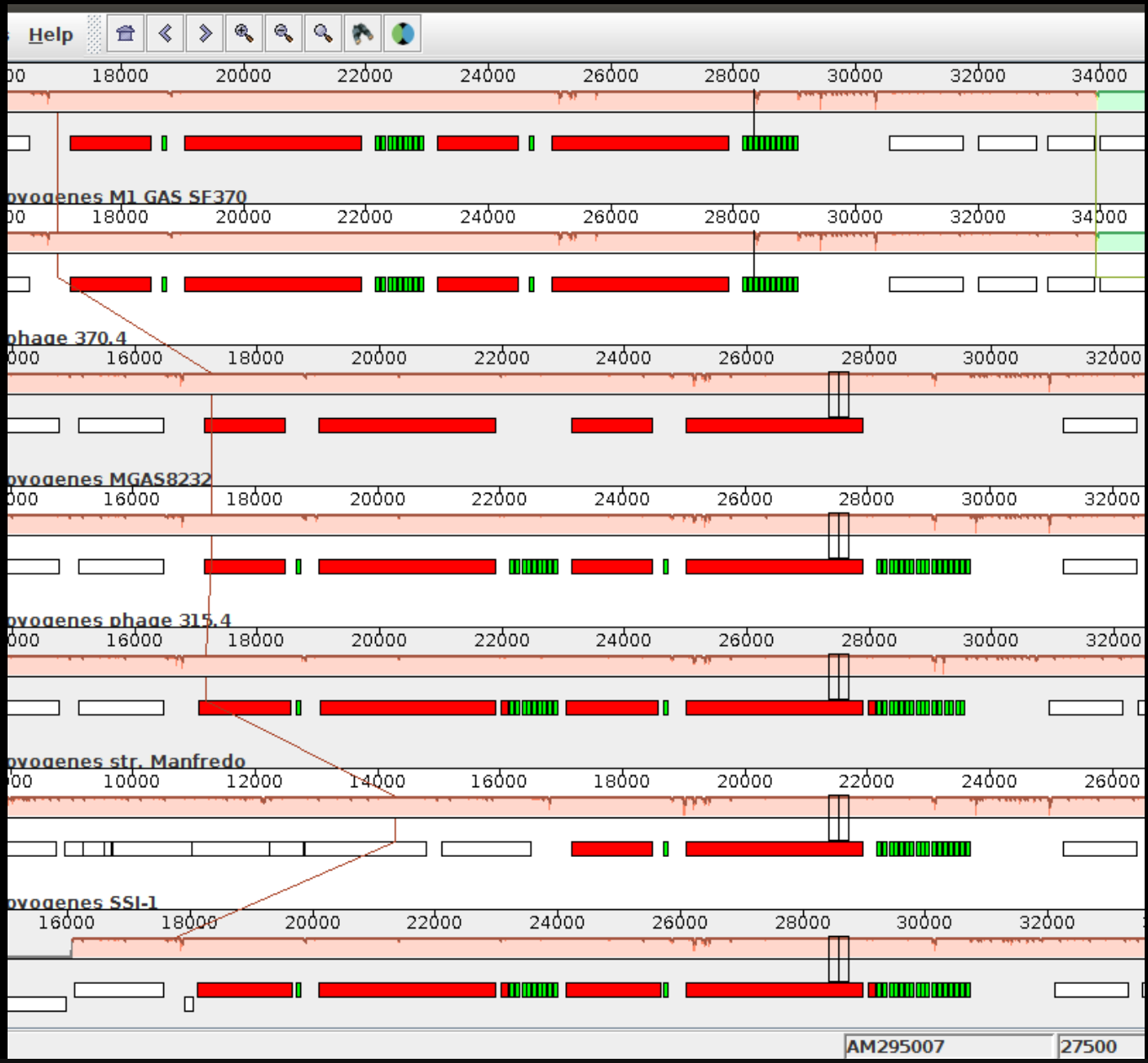
Hypothesis 1: ribosomal Operons

rDNA flanking regions are conserved conserved



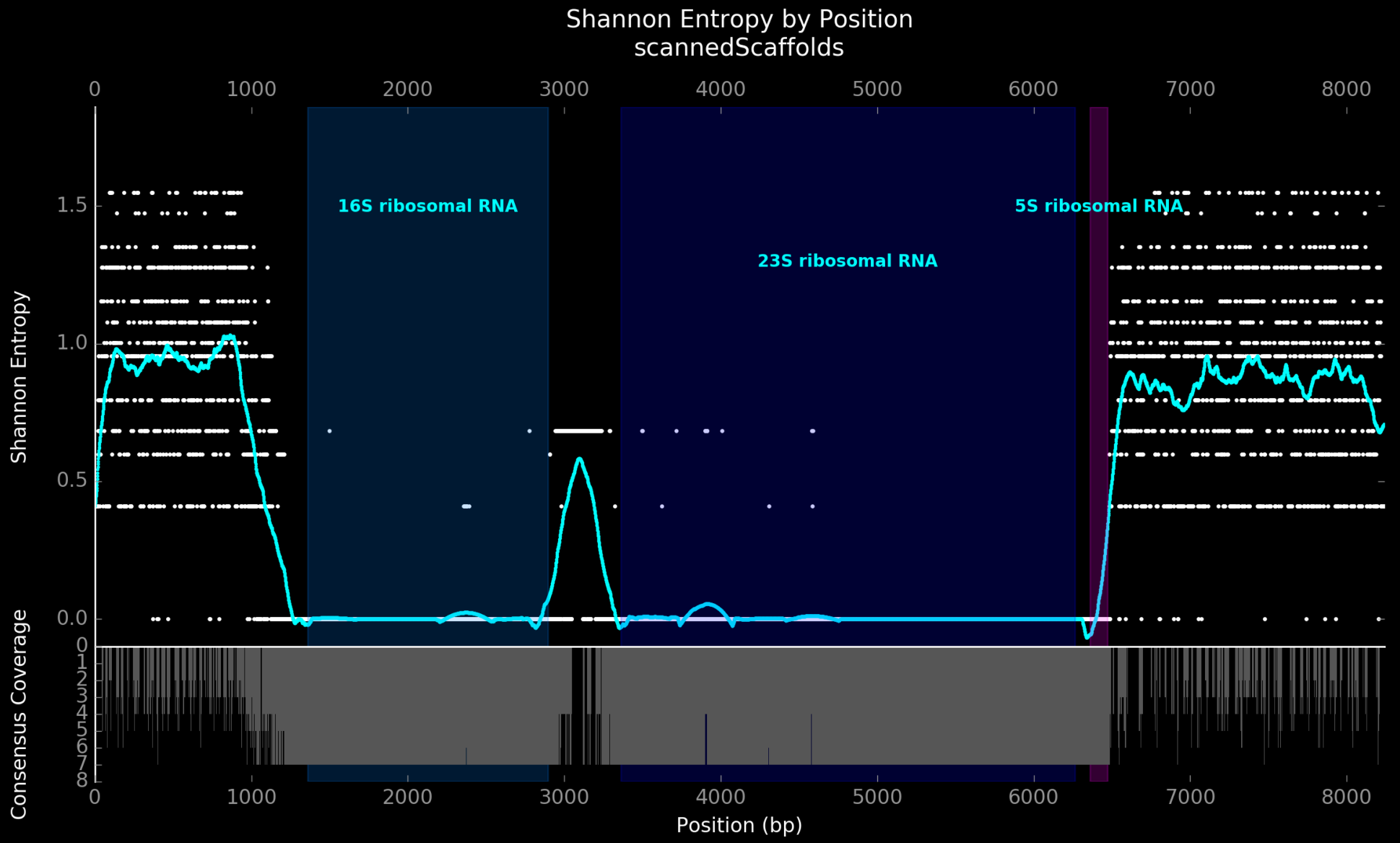


rDNA flanking regions are conserved

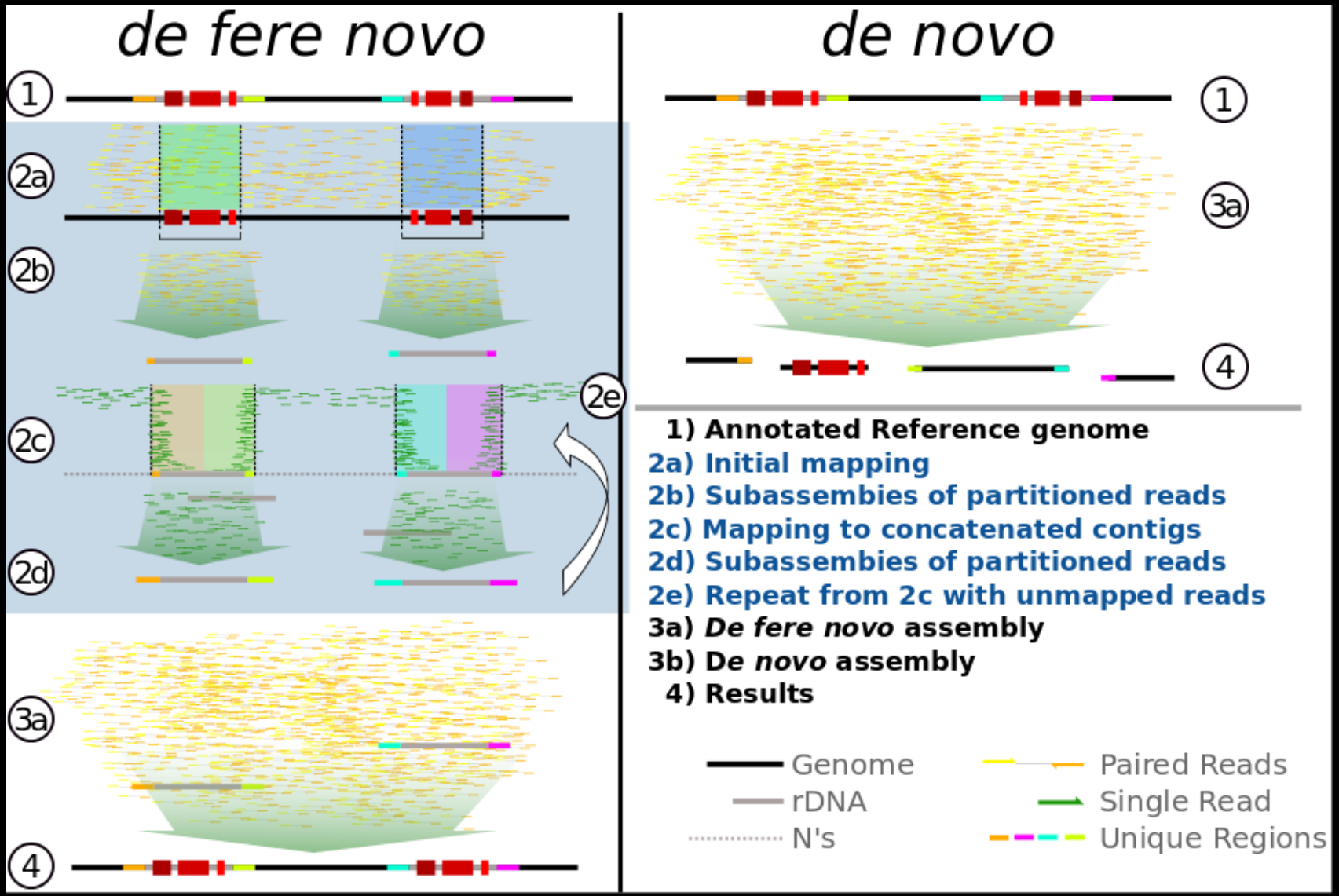


Hypothesis 2: Flanking Uniqueness

Flanking regions are unique within genome



Hypothesis 3: Long Read Construction


de fere novo
de novo
1
2a
2b
2c
2d
2e
3a
4
1) Annotated Reference genome
2a) Initial mapping
2b) Subassembies of partitioned reads
2c) Mapping to concatenated contigs
2d) Subassembies of partitioned reads
2e) Repeat from 2c with unmapped reads
3a) De fere novo assembly
3b) De novo assembly
4) Results
Genome
rDNA
N's
Paired Reads
Single Read
Unique Regions

riboSeed

- Automated method for constructing select “long reads” from Illumina data
- Written in python3 and R, wrapping barrnap, SMALT, SPAdes, and samtools
- 5 stages:
    1. Identify rDNA clusters
    2. Extracts reads mapping to a cluster
    3. Assemble into long reads
    4. Repeat (3x default) to extend
    5. Submit rDNA long reads to de novo assembly

riboSeed v0.3.06

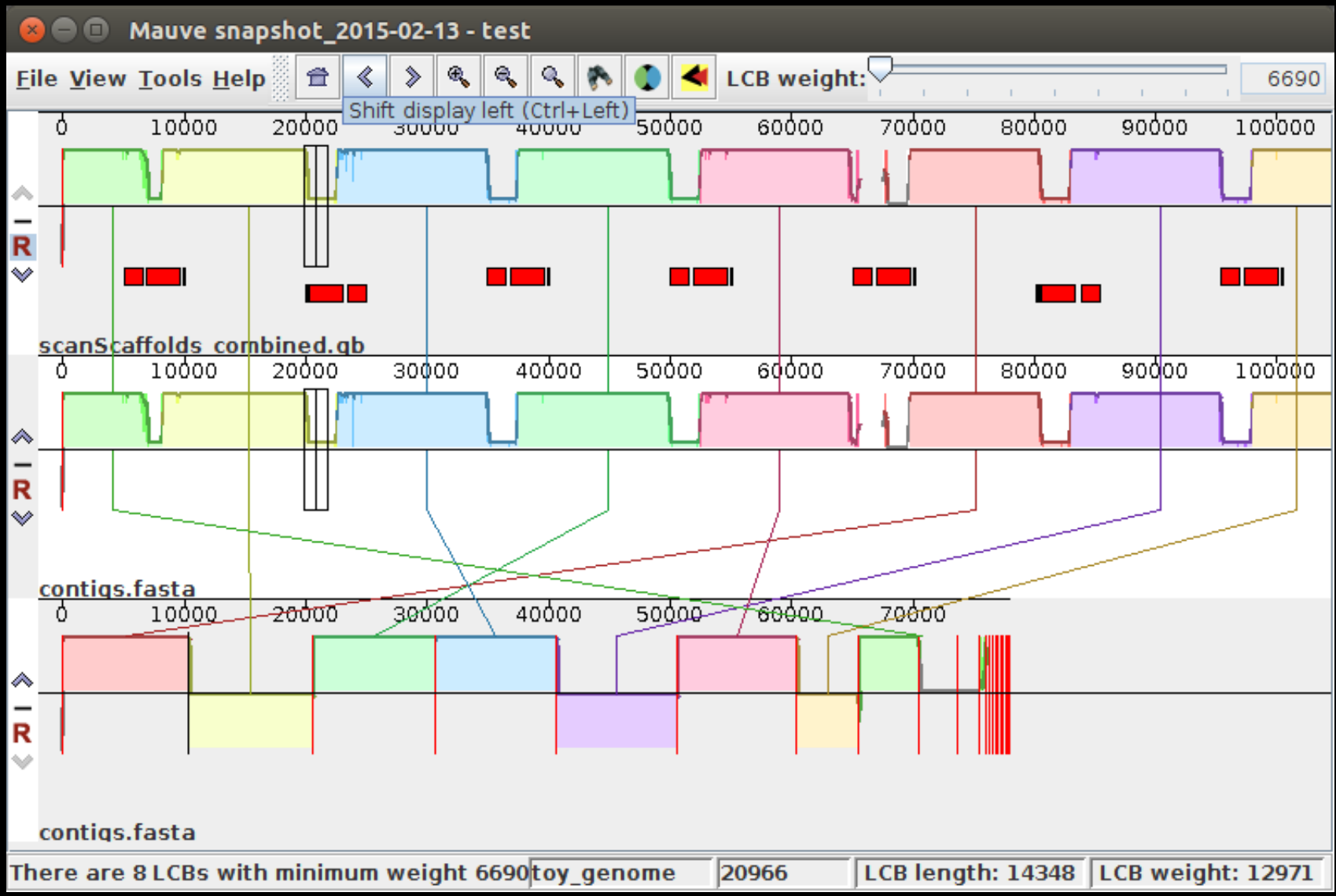
- Adjusts “trustworthiness” of long reads based on mapping
- Deals with rDNAs near origin
- Includes depth-of-coverage tool

Does it work?

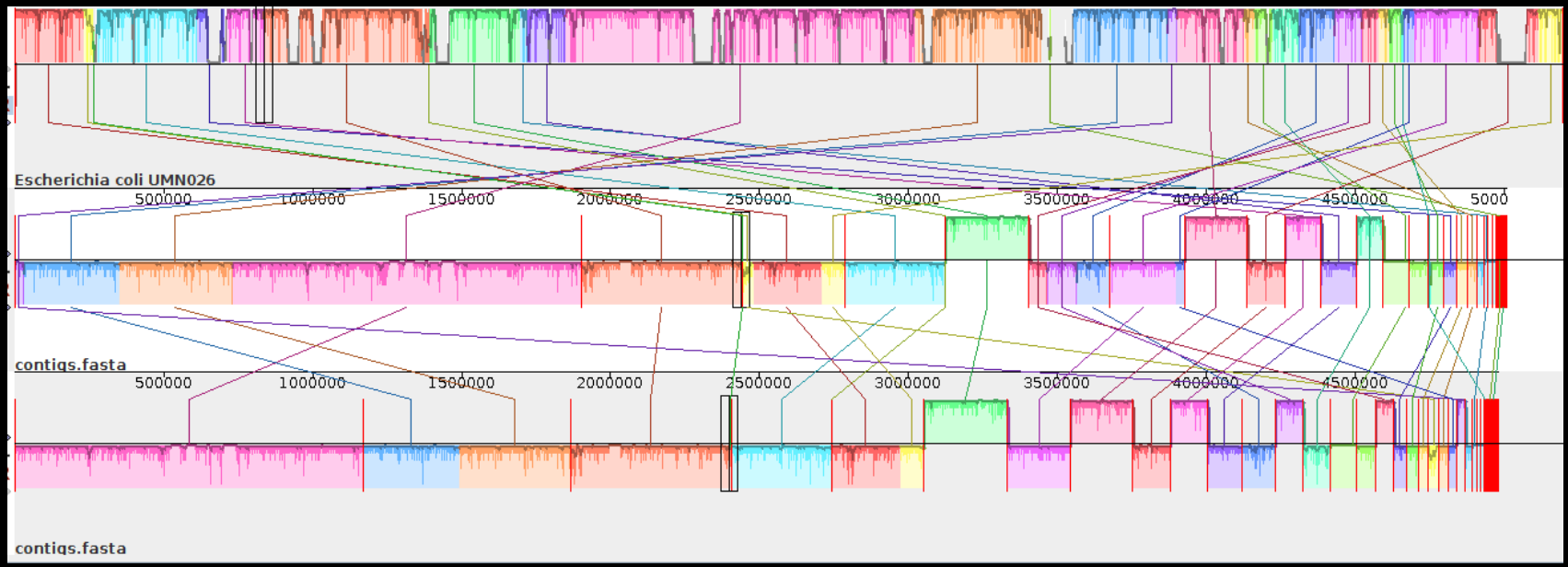
Benchmarking

1. synthetic reads on synthetic genome (7 E. coli Sakai rDNAs separated by 6kb random sequence)
2. synthetic reads on real genome
3. short reads from hybrid assembly
4. GAGE-B datasets

Synthetic reads on synthetic genome



Synthetic reads on real genome

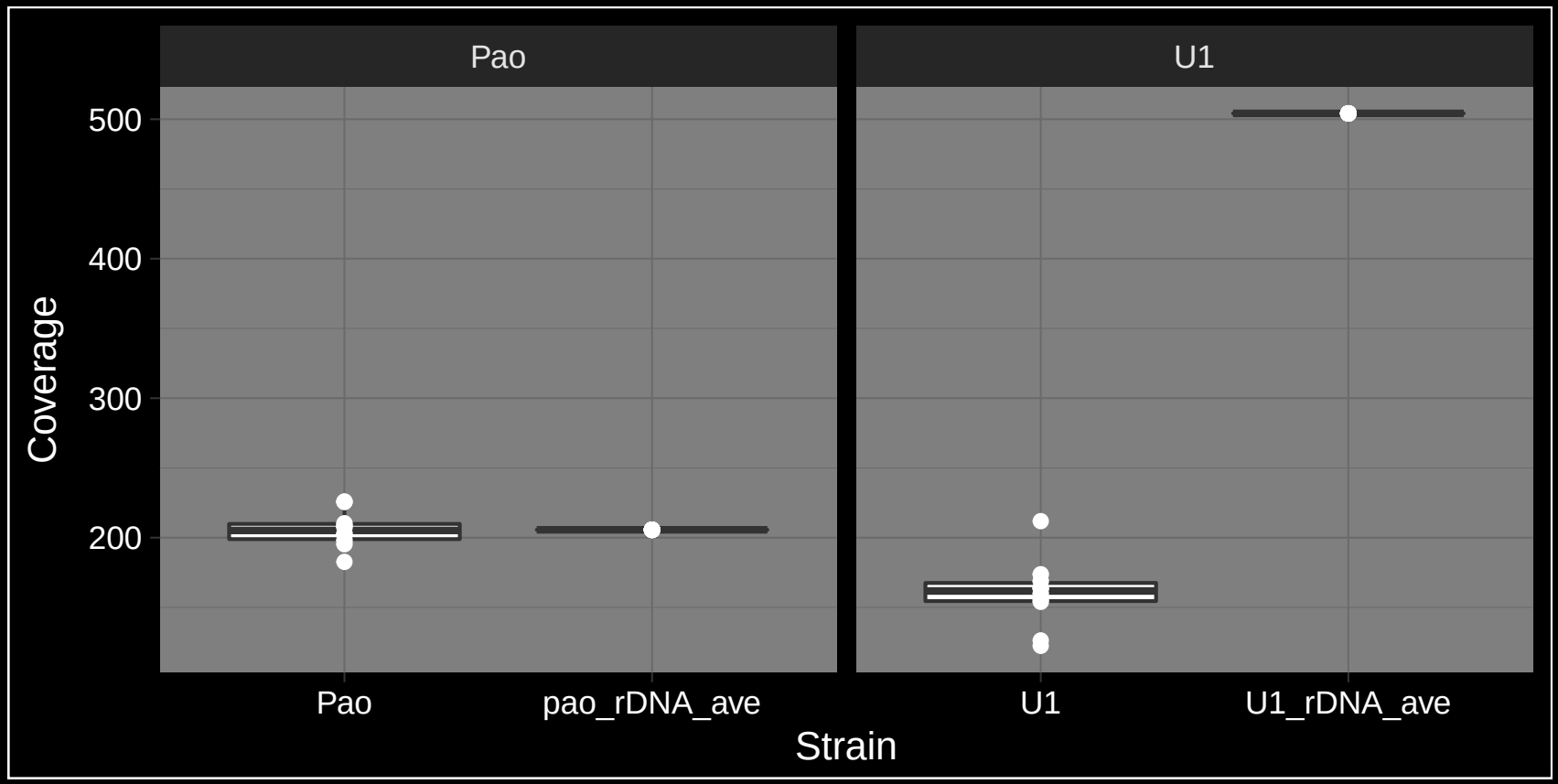


Benchmarking with hybrid assembly

- Successfull assembly across all 4 rDNAs
- Highly accurate reconstruction (3 SNPs per 8000bp)

Challenges

Reference Reliability



Other Challenges

- Data Availability
- Cross Platform Performance
- Recruiting Beta Testing

Conclusions

Potential Downsides

1. Unpredictable
2. Single problem/solution
3. Biased by reference

Summary

- The architecture of bacterial genomes can aid assembly

Summary

- The architecture of bacterial genomes can aid assembly
- rDNA flanking regions are unique within a genome

Summary

- The architecture of bacterial genomes can aid assembly
- rDNA flanking regions are unique within a genome
- riboSeed improves assemblies at best

Summary

- The architecture of bacterial genomes can aid assembly
- rDNA flanking regions are unique within a genome
- riboSeed improves assemblies at best
- riboSeed doesn't work on in all cases, but rarely introduces errors

Next Steps

- Benchmark against GAGE-B
- Benchmark against more hybrid assembly studies
- Find early indicator
- Apply to fungal genomic
- Apply to other conserved regions

Acknowledgments

- Fiona Brennan
- Florence Abram
- Matthias Waibel
- Camilla Thorn
- Stephen Nolan

- Leighton Pritchard
- Ashleigh Holmnes

Acknowledgments

- Fiona Brennan
- Florence Abram
- Matthias Waibel
- Camilla Thorn
- Stephen Nolan

- Leighton Pritchard
- Ashleigh Holmnes

Questions?