riboSeed: leveraging bacterial genomic architecture to assemble across ribosomal regions

Nick Waters

Bioinformatics Dept: February 2, 2017

National University of Ireland, Galway
James Hutton Institute, Dundee

Introduction

Background

My project: Comparitive Genomics of soil-Adapted E. coli

Given our 155 sequenced soil-adapted isolates, what can we learn about E. coli genomics?

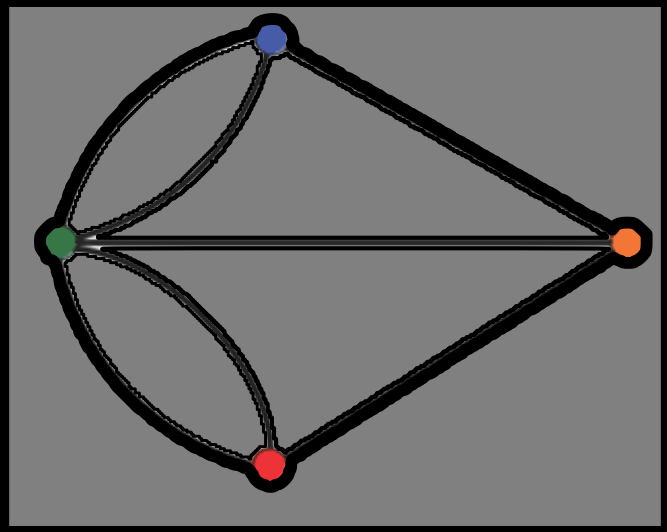
- Phylogeny
- Genomic Restructuring
- Virulence/AMR
- Detection

Short Ready Assembly: Background

Bridges of Königsberg

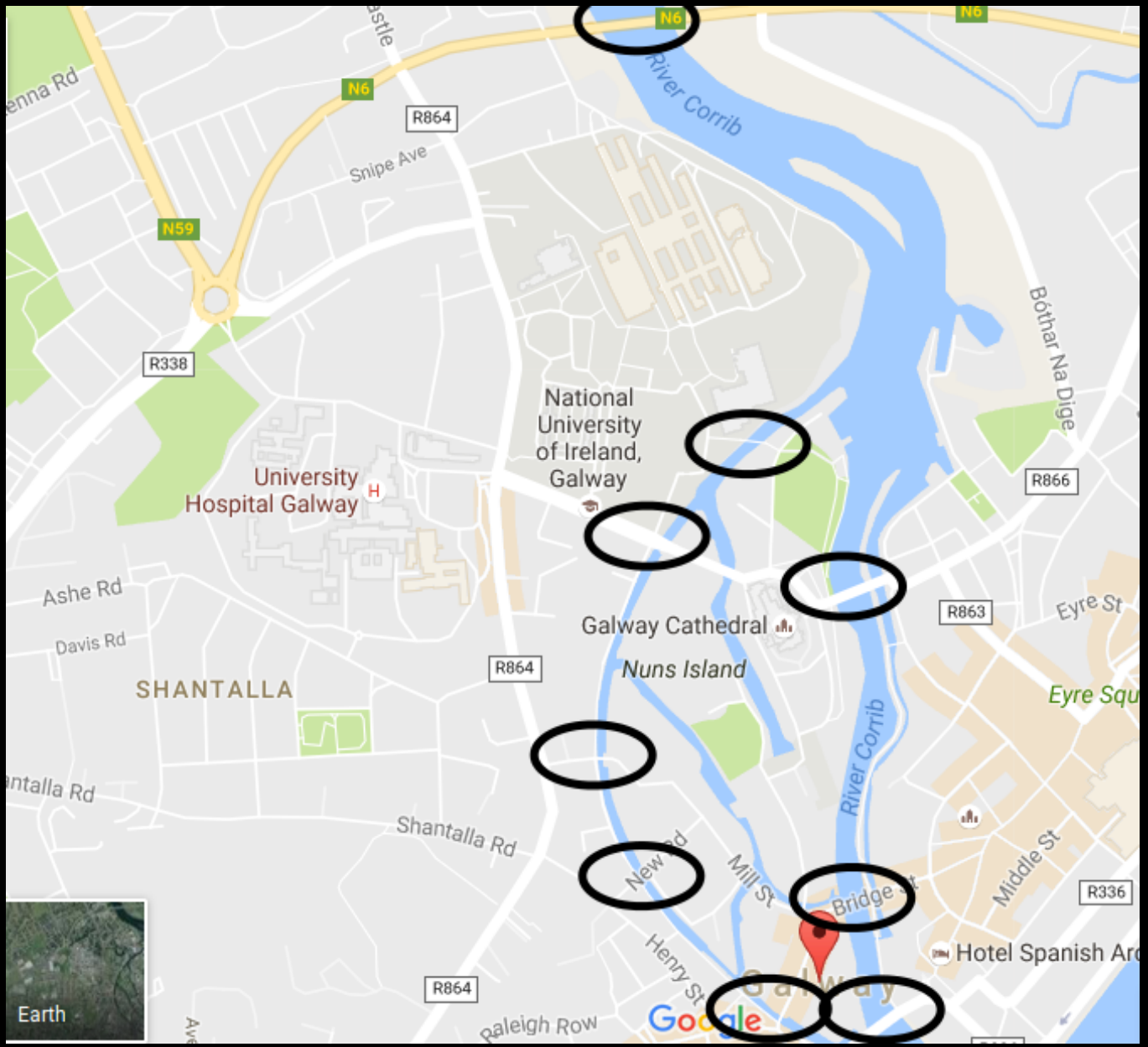
Source[Chaisson et al., 2015]

Bridges of Königsberg

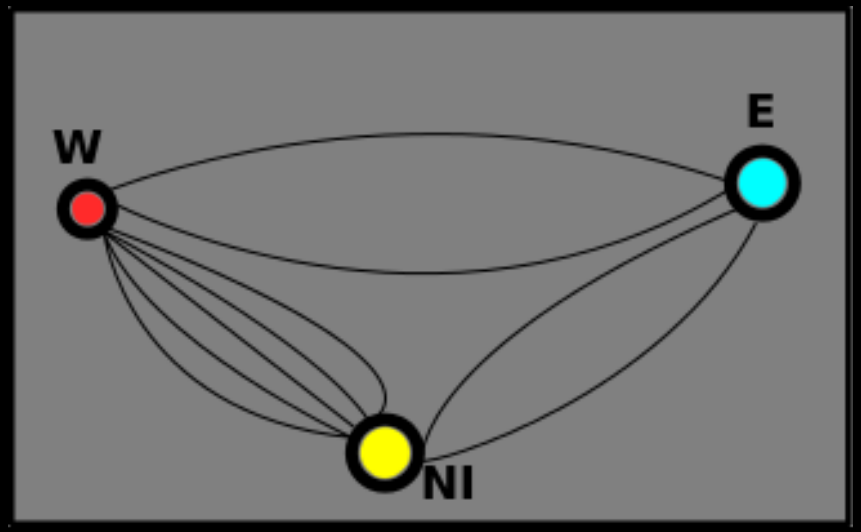


Source[Chaisson et al., 2015]

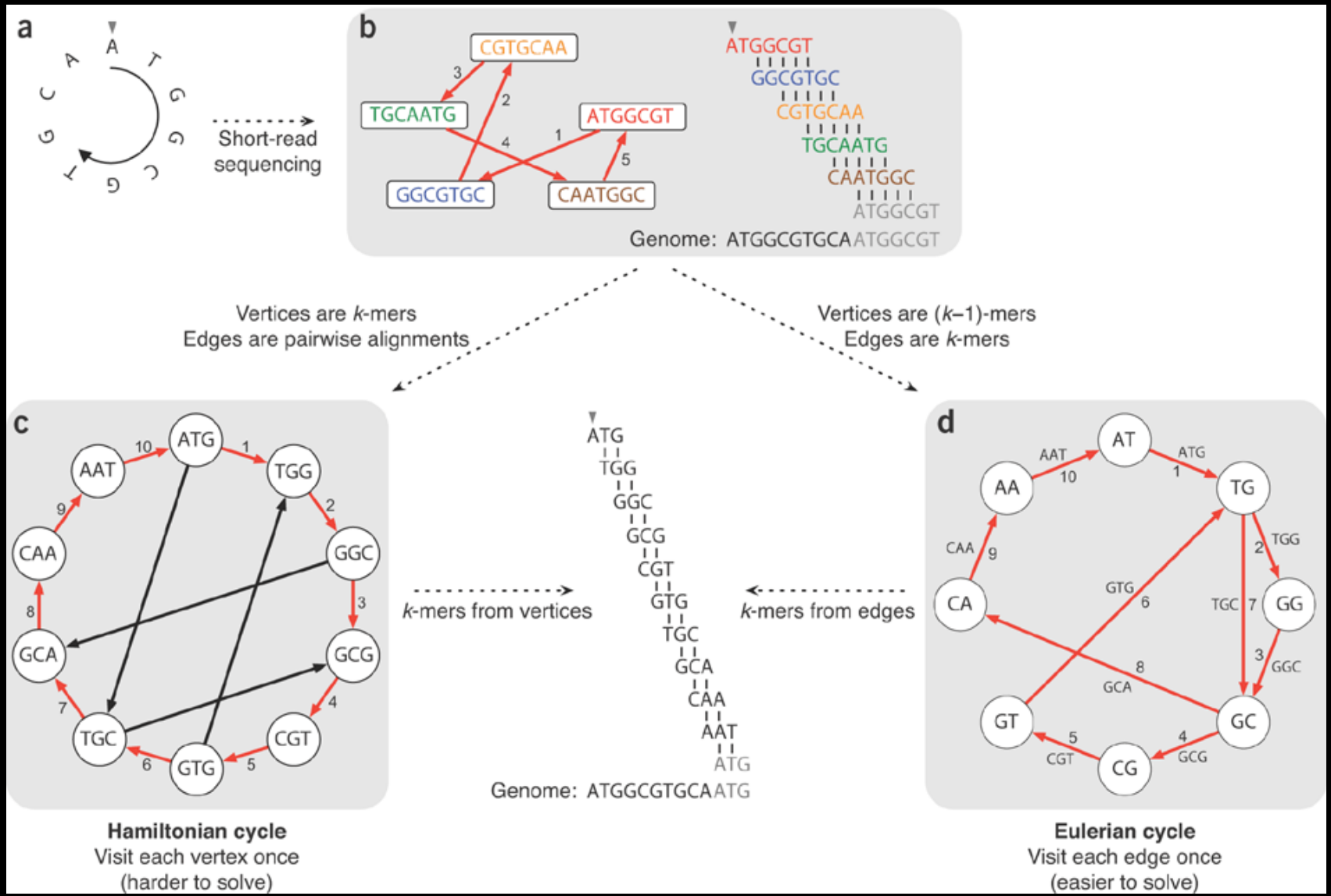
Bridges of Galway



Bridges of Galway, Simplified

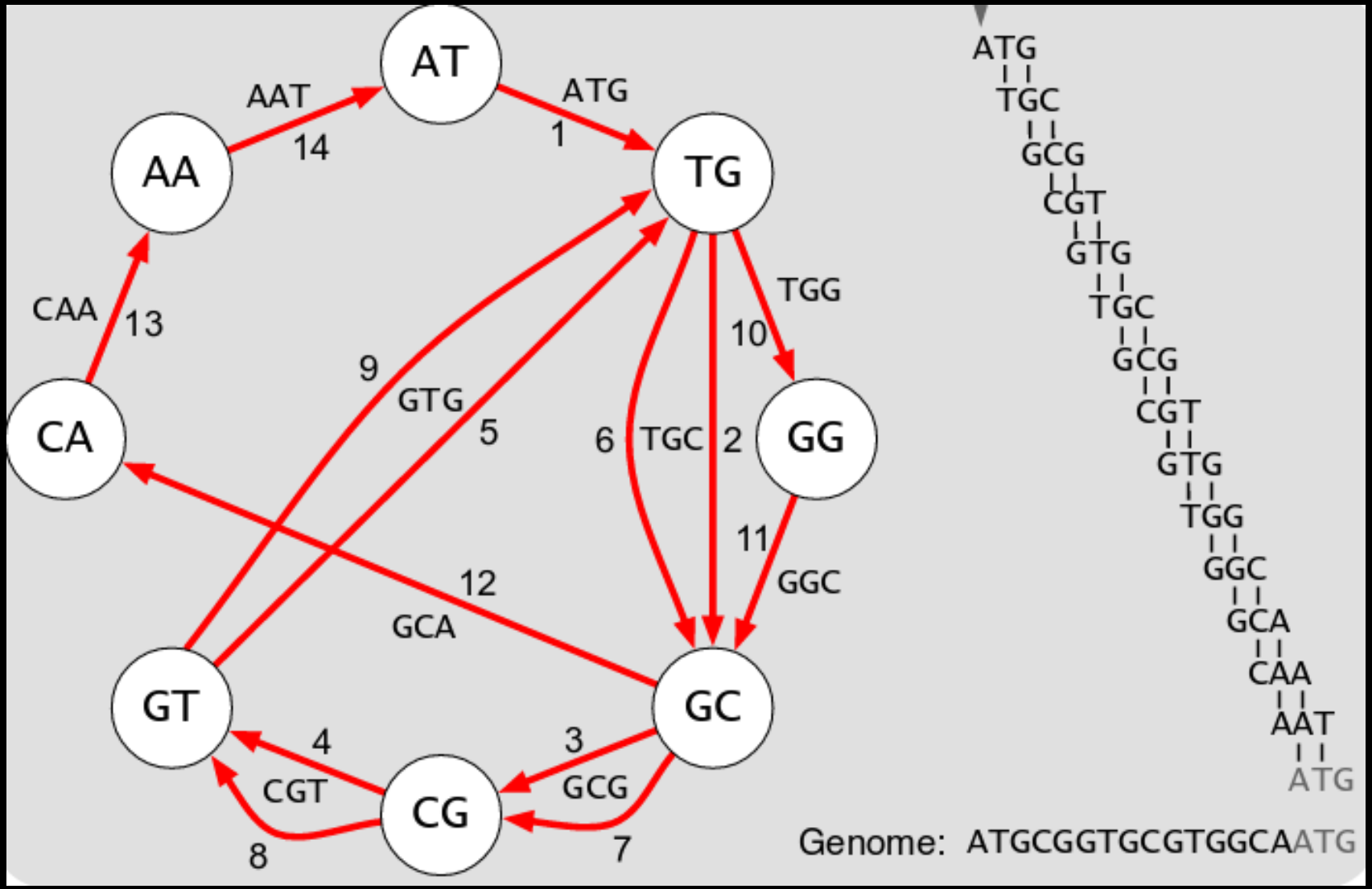


de Bruijn Graphs and Eulerian Paths



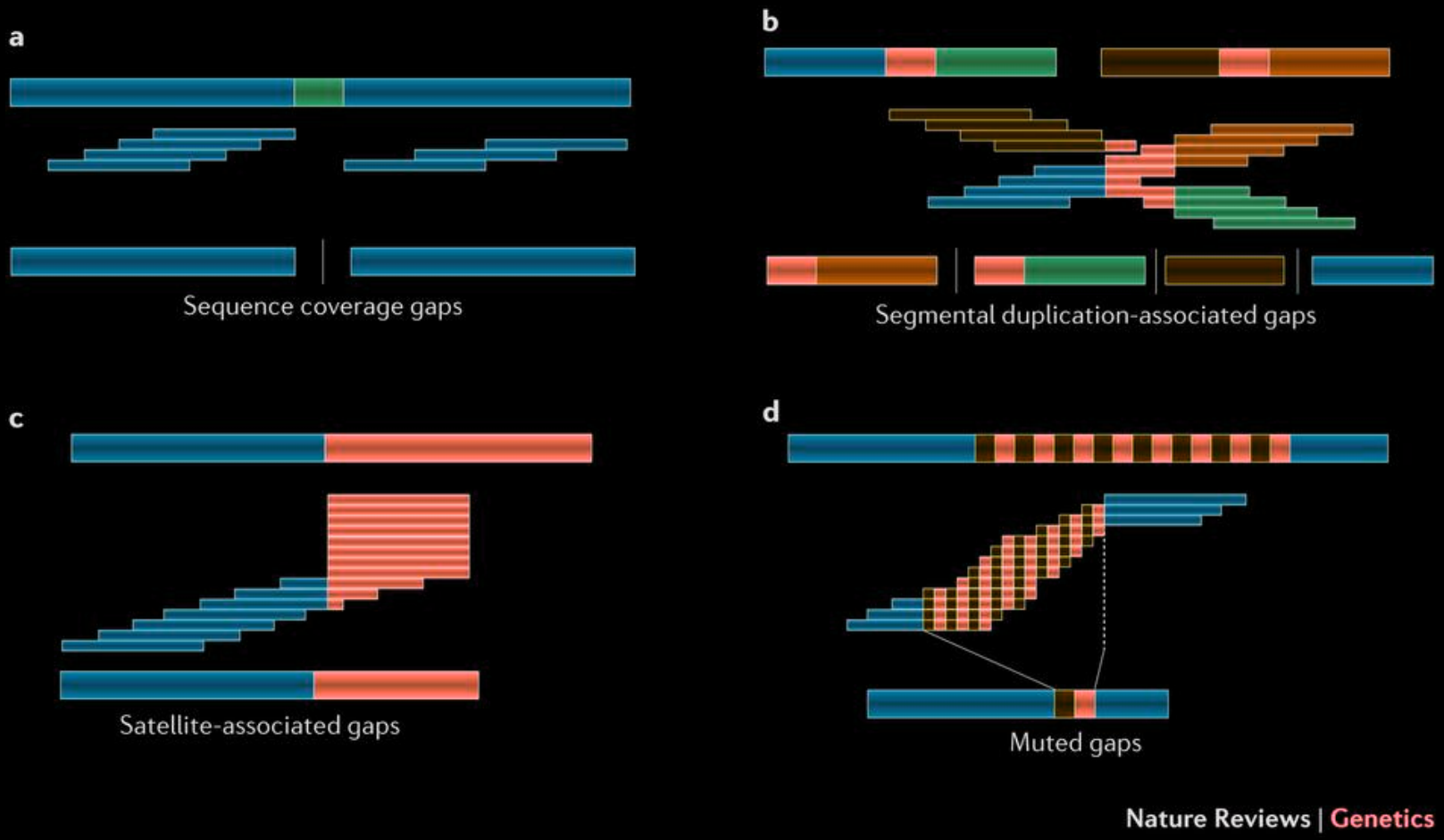
Source[Compeau et al., 2011]

Eulerian Paths with Repeats



Source[Compeau et al., 2011]

Problems



Source[Chaisson et al., 2015]

Source: T. Seemann

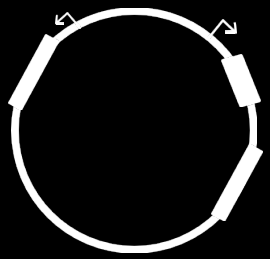
Problem Regions

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

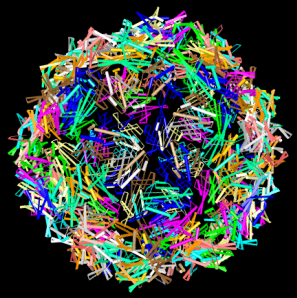
Problem Regions

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

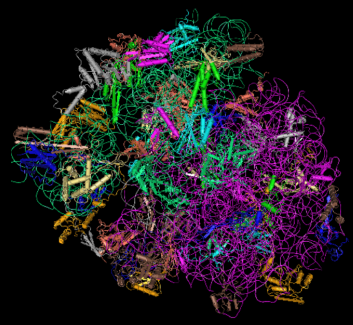
Transporters



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, availibility
reference assisted	easy to perform	not reliable

Possible Probability Levers

Law of the Probability Lever: Slight changes can make highly improbable events almost certain

Source: David Hard

Possible Probability Levers

Law of the Probability Lever: Slight changes can make highly improbable events almost certain

1. Within a taxanomic group, GC content is largely conserved (kmer strain typing, etc).

Source: David Hard

Possible Probability Levers

Law of the Probability Lever: Slight changes can make highly improbable events almost certain

1. Within a taxanomic group, GC content is largely conserved (kmer strain typing, etc).
2. Within a taxonomic group, genome size is largely conserved.

Source: David Hard

Possible Probability Levers

Law of the Probability Lever: Slight changes can make highly improbable events almost certain

1. Within a taxanomic group, GC content is largely conserved (kmer strain typing, etc).
2. Within a taxonomic group, genome size is largely conserved.
3. Bacterial genomes are dense.

Source: David Hard

Possible Probability Levers

Law of the Probability Lever: Slight changes can make highly improbable events almost certain

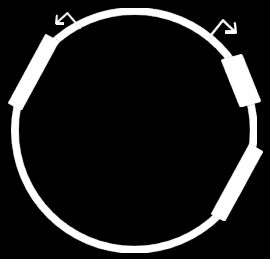
1. Within a taxanomic group, GC content is largely conserved (kmer strain typing, etc).
2. Within a taxonomic group, genome size is largely conserved.
3. Bacterial genomes are dense.
4. Nucleotide order is not random.

Source: David Hard

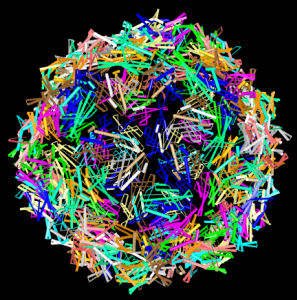
Problem Regions

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

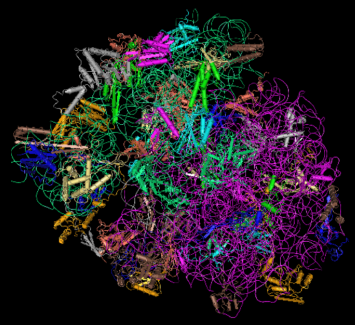
Transporters



Ω Plasmids



Prophages

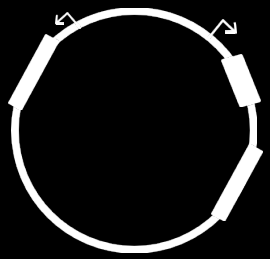


Ribosomes

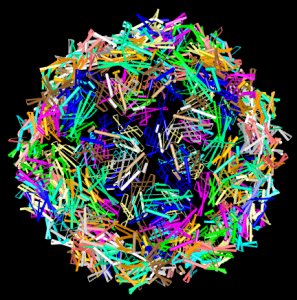
Problem Regions

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

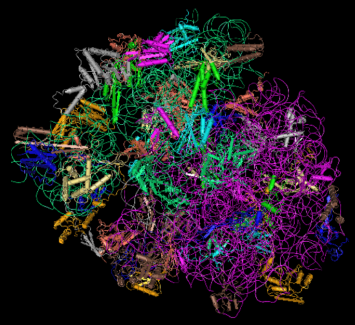
Transporters



Ω Plasmids



Prophages



Ribosomes

rDNA

rDNA: ribosomal DNA operon

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

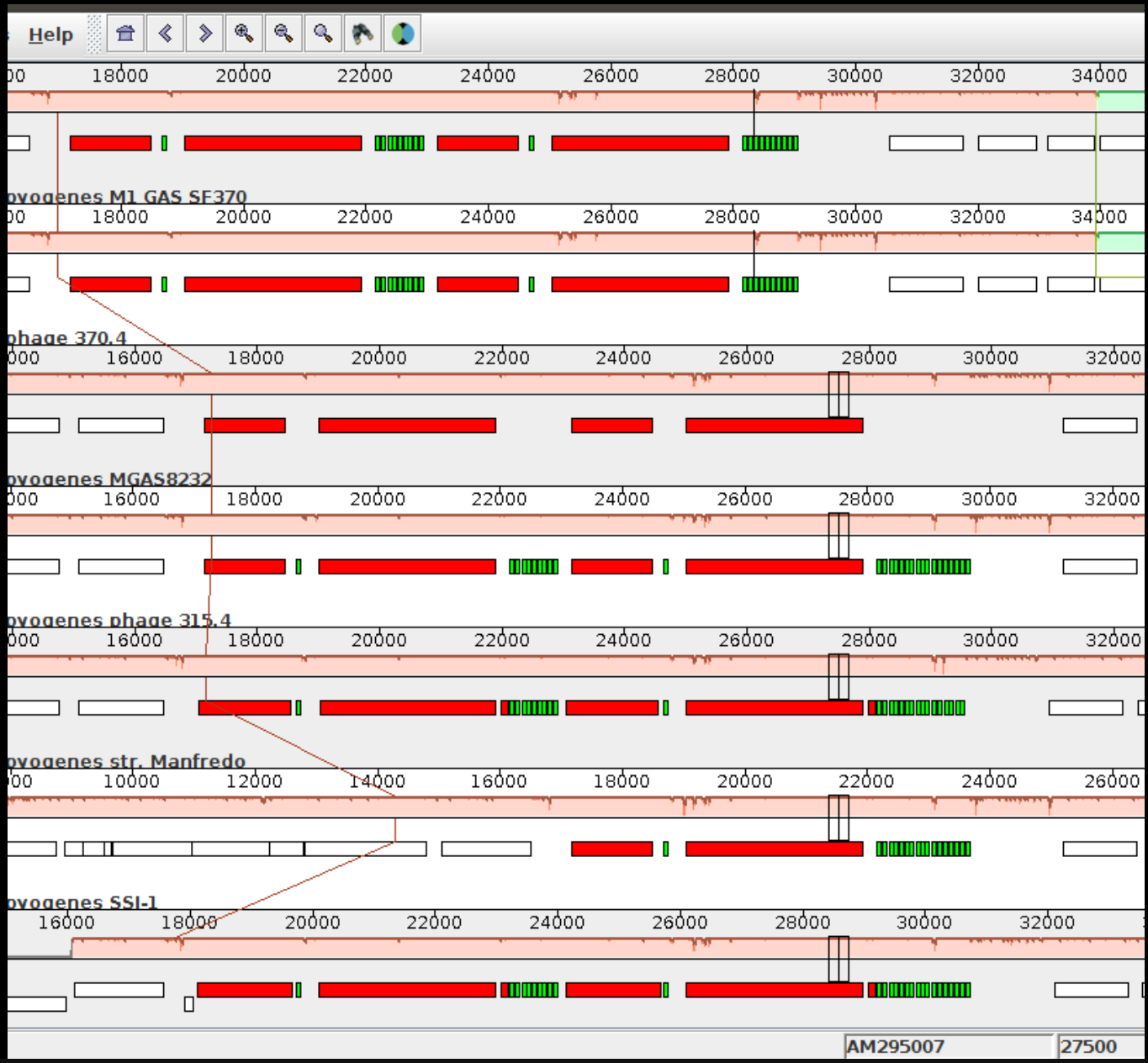
Hypotheses

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

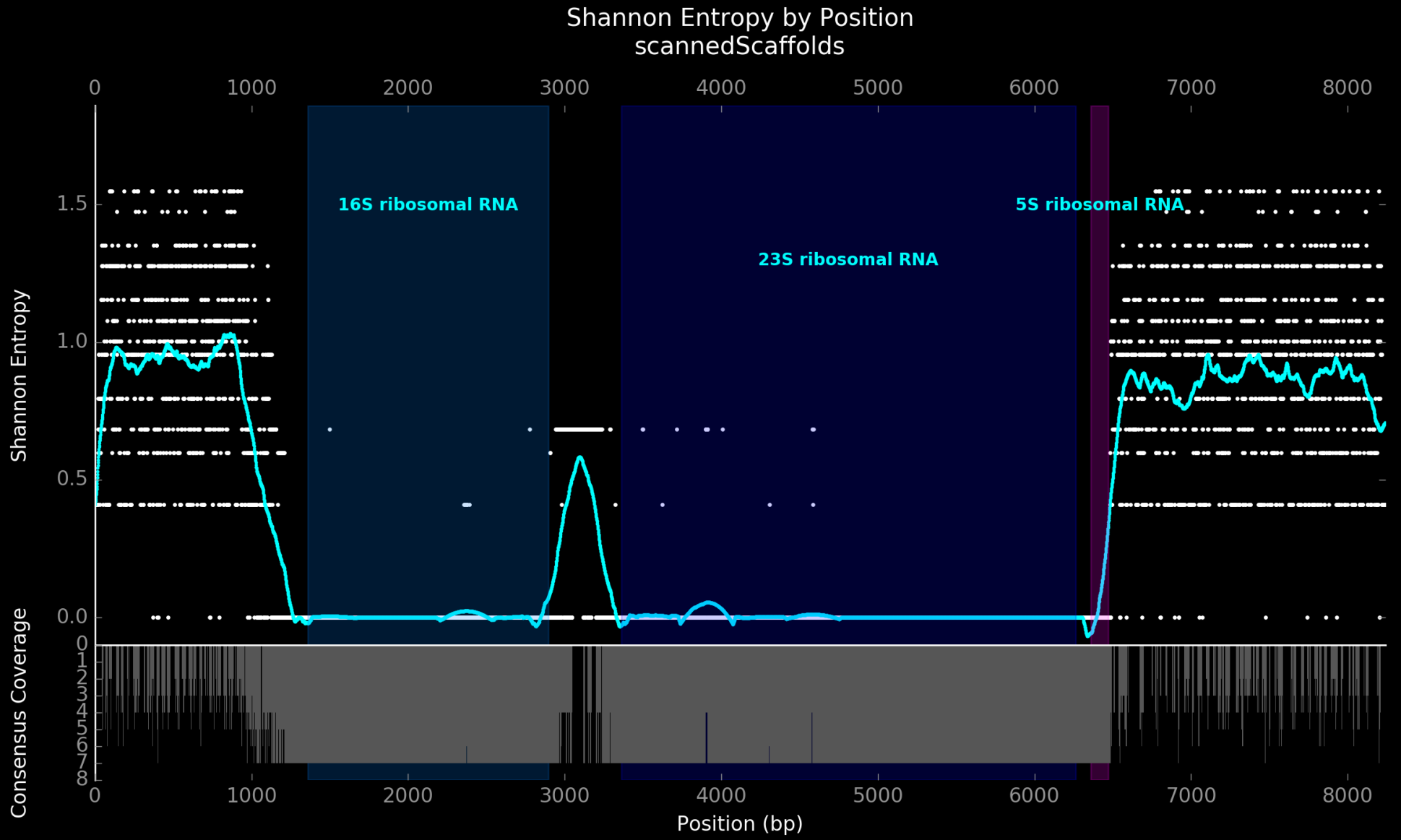
rDNA flanking regions are conserved conserved

rDNA flanking regions are conserved

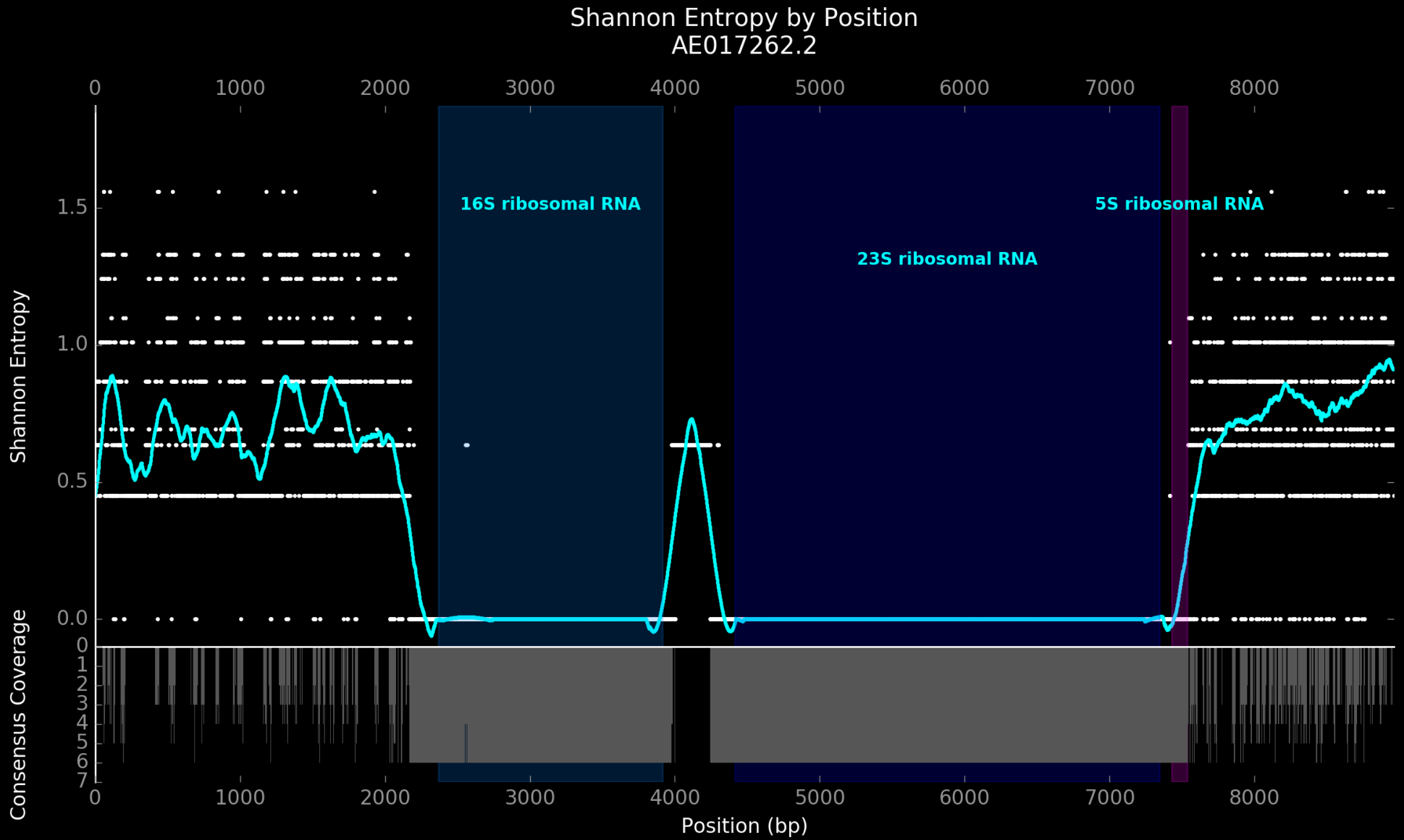


Hypothesis 2: Flanking Uniqueness

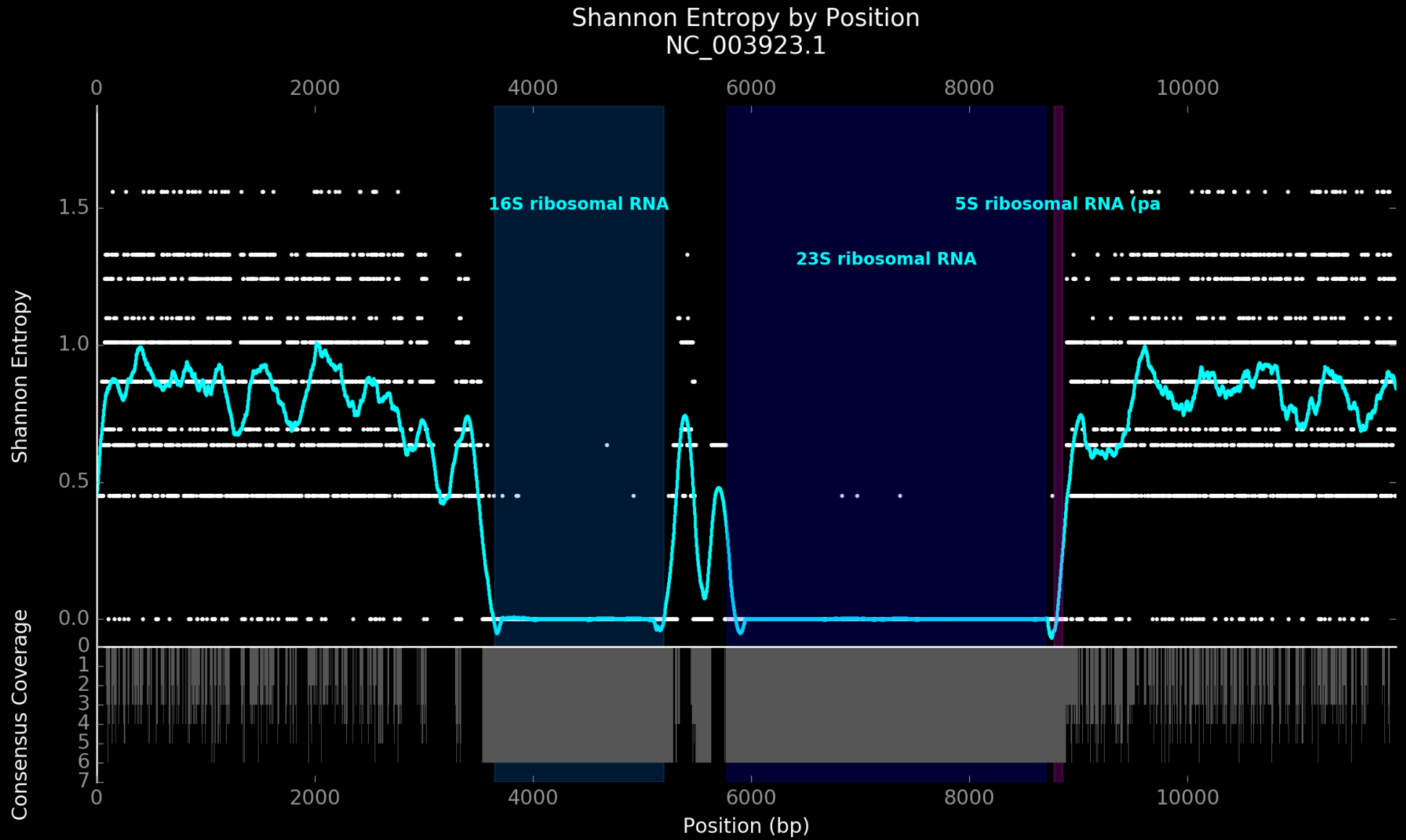
Flanking regions are unique within genome: *E. coli*



Flanking regions are unique within genome: *L. monocyt*



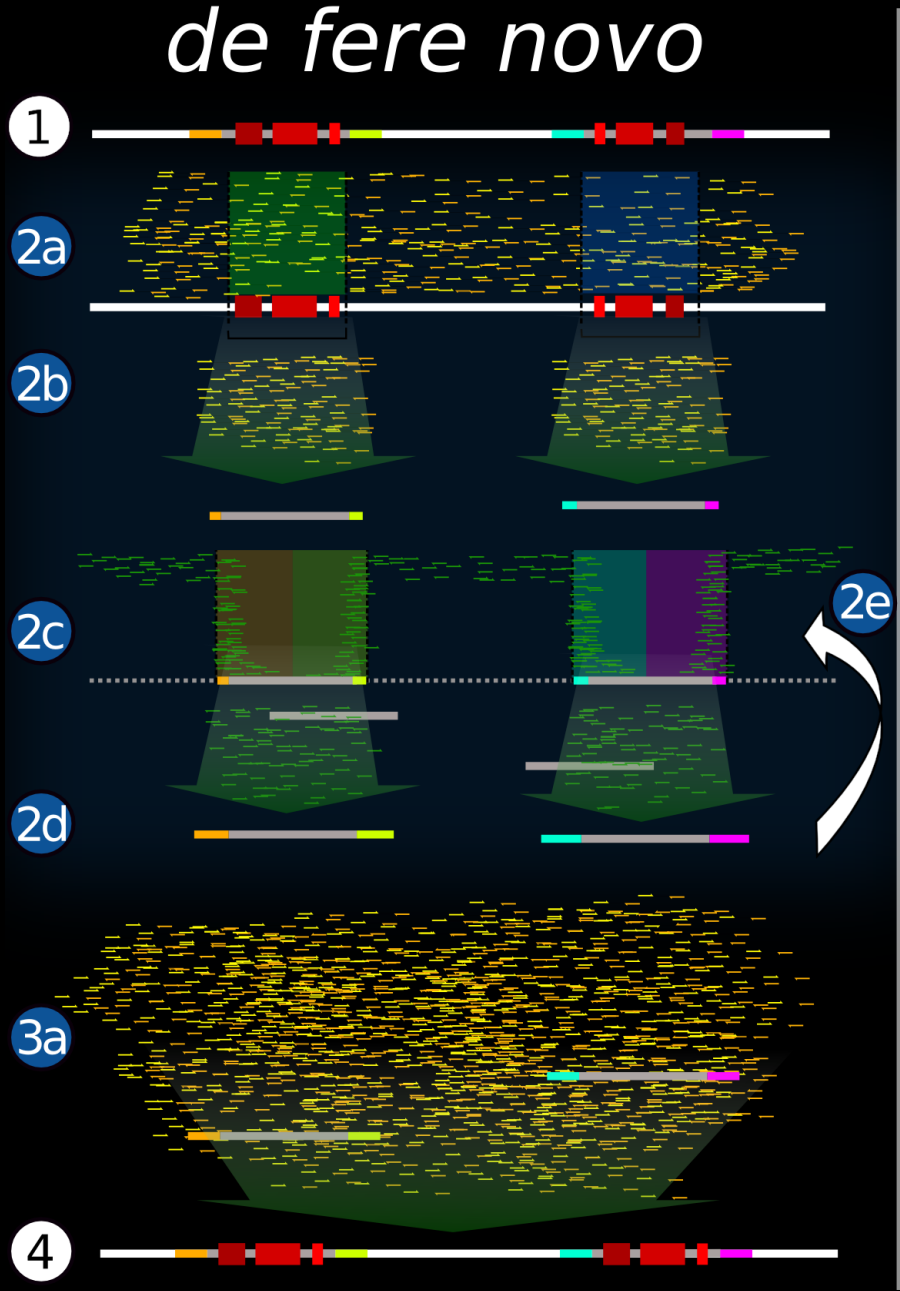
Flanking regions: *S. aureus*

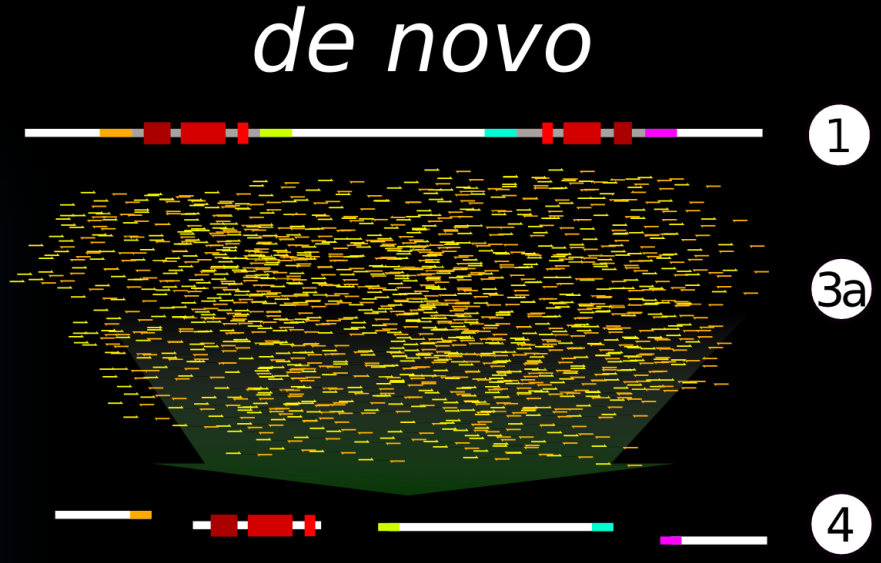


Hypothesis 3: Long Read Construction

riboSeed

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





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

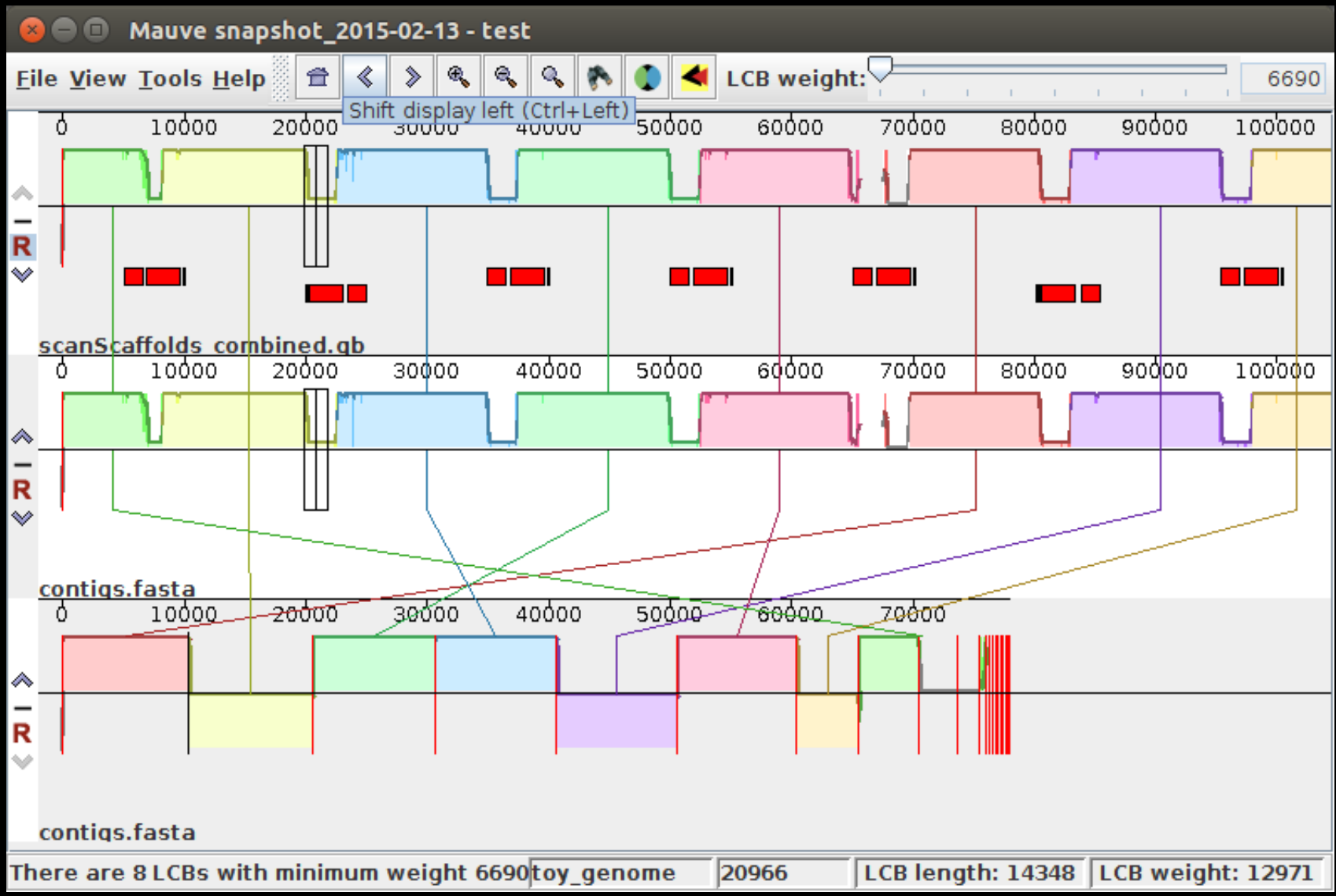
Genome
rDNA
N's
Paired Reads
Single Read
Unique Regions

Does it work?

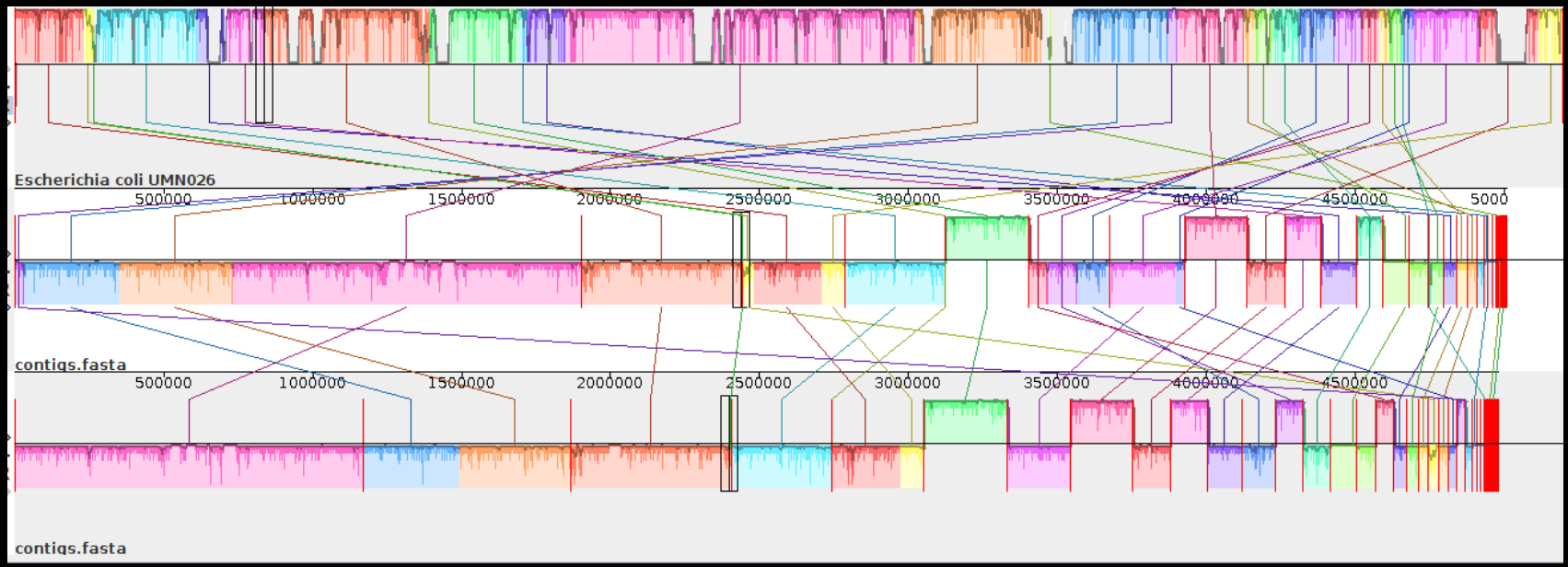
Benchmarking

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

Mauve Demo

Conclusions

Potential Downsides

1. Doesn't work for all datasets
   - coverage <10x
   - other repeated regions neighboring rDNA
2. Single problem/solution
3. Biased by reference choice

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 does’t work on in all cases, but rarely introduces new 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


Chaisson, M. J. P., Wilson, R. K., and Eichler, E. E. (2015).
Genetic variation and the de novo assembly of human genomes.
Nature Publishing Group, 16.


Compeau, P. E. C., Tesler, G., and Pevzner, P. A. (2011).
How to apply de Bruijn graphs to genome assembly.
Nature biotechnology, 29(11):987–991.

Acknowledgements

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

- Leighton Pritchard
- Ashleigh Holmnes

Acknowledgements

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

- Leighton Pritchard
- Ashleigh Holmnes

Questions?