DISCUSSION CLUB: CURRENT RESEARCH

Nick Waters

December 6, 2016

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

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

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

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

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

Background

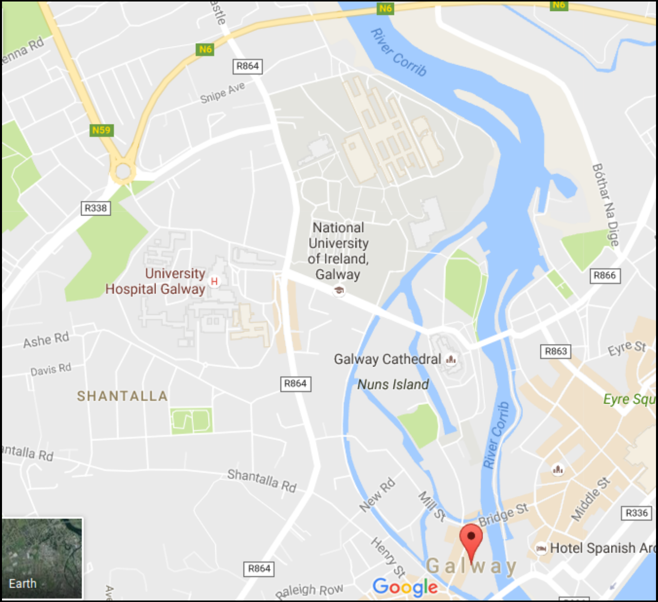
My project: Comparitive Genomics of soil-Adapted E. coli

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

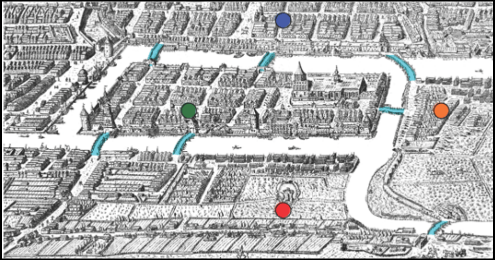
- Phylogeny
- Genomic Restructuring
- Virulence/AMR
- Detection

SHORT READY ASSEMBLY: BACKGROUND

Bridges of Galway

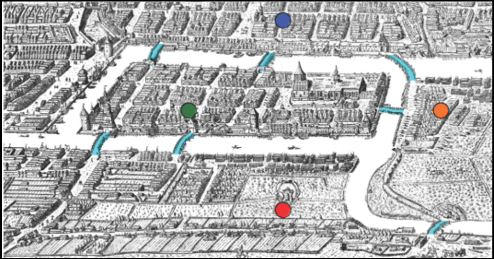


Bridges of Königsberg

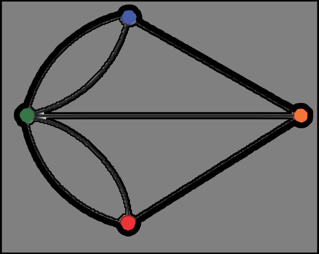


Source[Chaisson et al., 2015]

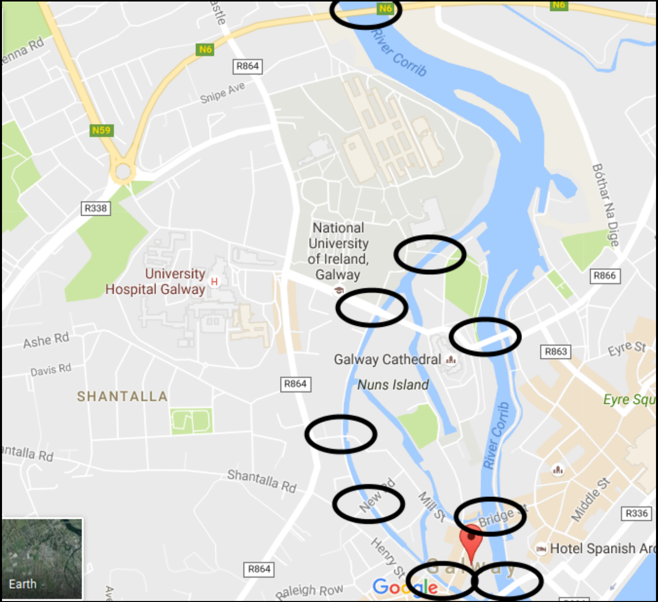
Bridges of Königsberg



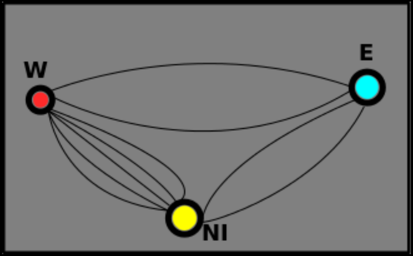
Source[Chaisson et al., 2015]



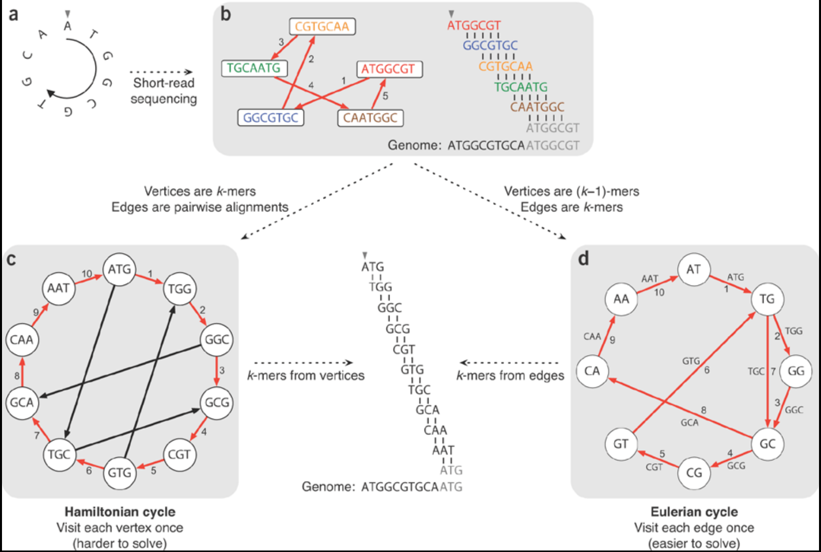
Bridges of Galway



Bridges of Galway, Simplified

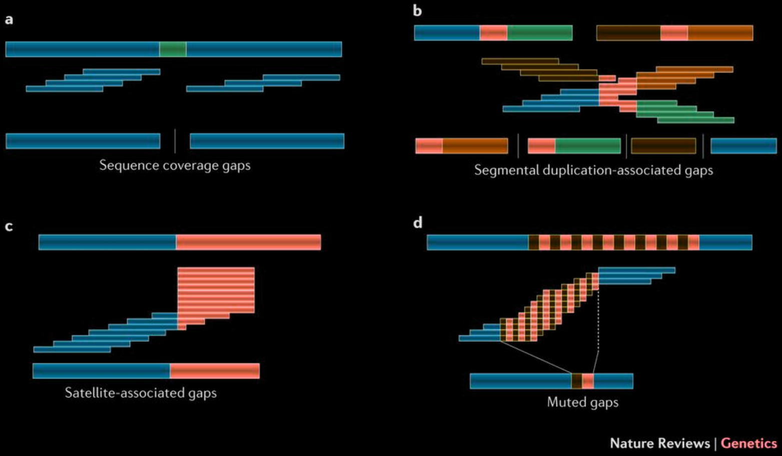


de Bruijn Graphs and Eulerian Paths



Source[Compeau et al., 2011]

Problems


a
Sequence coverage gaps
b
Segmental duplication-associated gaps
c
Satellite-associated gaps
d
Muted gaps
Nature Reviews | Genetics

Source[Chaisson et al., 2015]

Source: T. Seemann

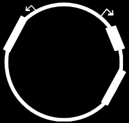
Problem Regions

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

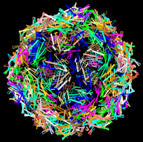
Problem Regions

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

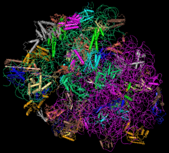
Transporters



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

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

Possible Solution

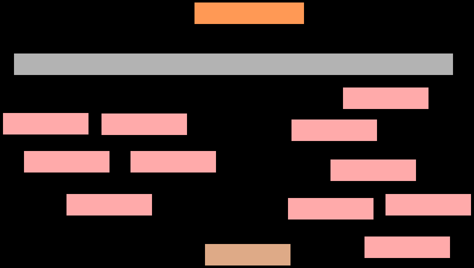
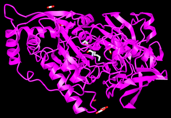


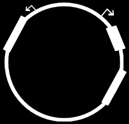
Figure: Bridge Reconstruction. Pink fragments are reads. Grey shows the gene of interest with interupted coverage. Orange fragemnt is a pseudoread generated from this situation under the hypothesis that the beige fragment exists but is underrepresented

Problem Regions

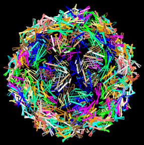
Repeated regions cannot be resolved with kmers shorter than the repeat!



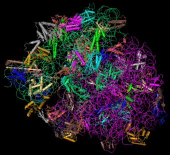
Transporters



Ω Plasmids



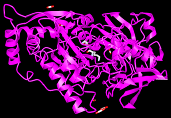
Prophages



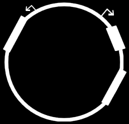
Ribosomes

Problem Regions

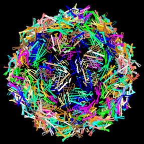
Repeated regions cannot be resolved with kmers shorter than the repeat!



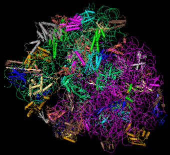
Transporters



Ω Plasmids



Prophages



Ribosomes

rDNA

rDNA: ribosomal DNA operon

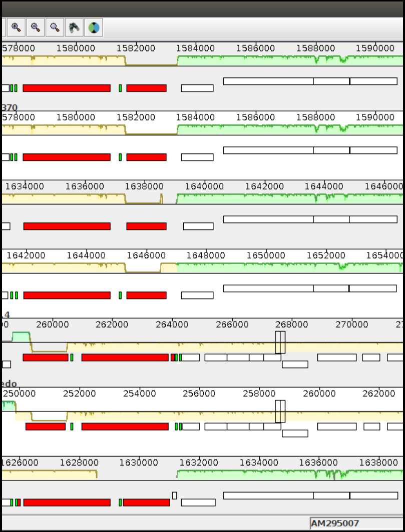
- Prokaryotes: 16S, 23S, 5S
- Conserved within taxa
- Repeated within the genome (1x to >14x)

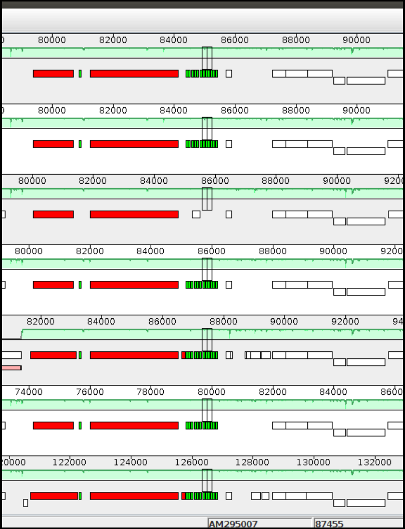
Hypotheses

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”

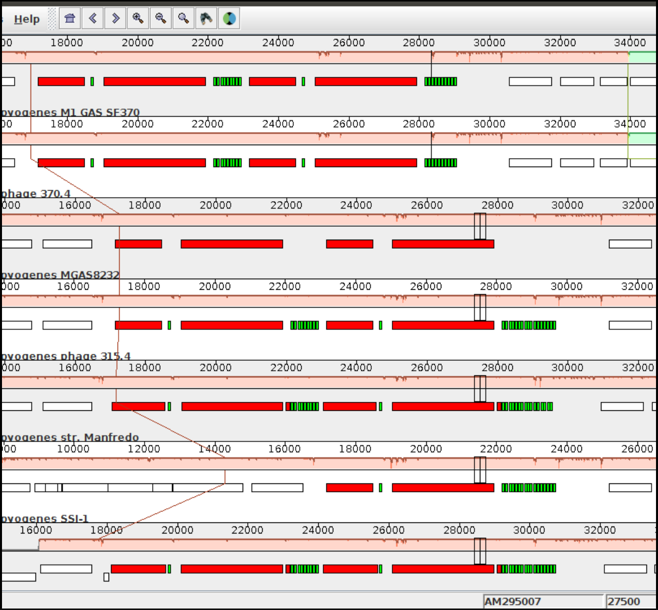
HYPOTHESIS 1: RIBOSOMAL OPERONS

rDNA flanking regions are conserved conserved



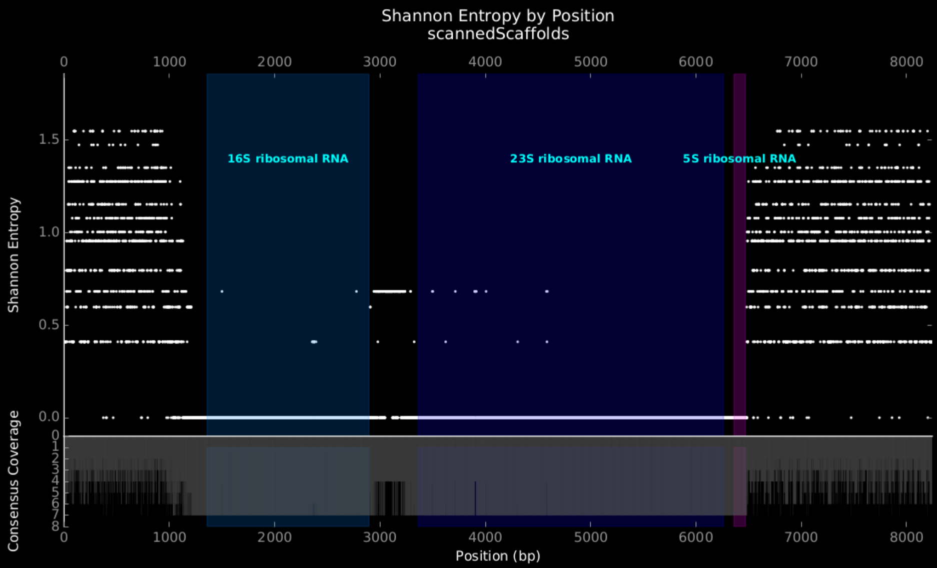


rDNA flanking regions are conserved



HYPOTHESIS 2: FLANKING UNIQUENESS

Flanking regions are unique within genome



HYPOTHESIS 3: LONG READ CONSTRUCTION

riboSeed

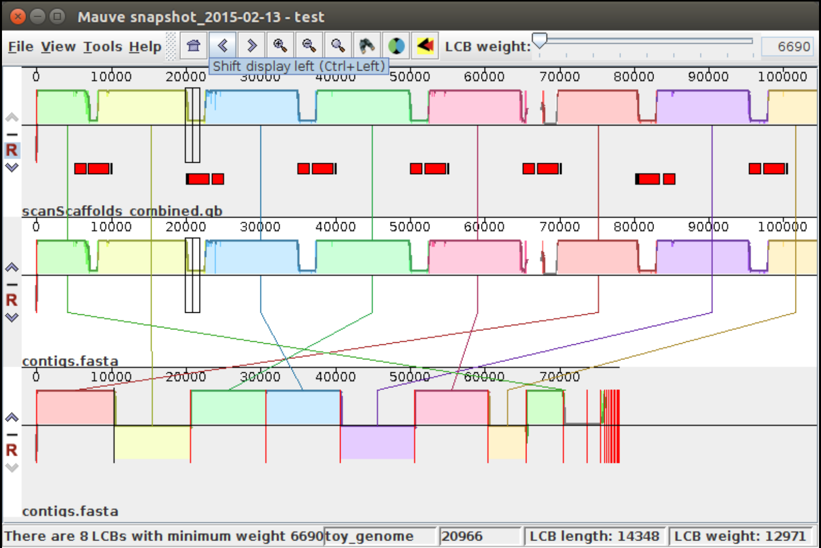
- Automated method for constructing select “long reads” from Illumina data
- Written in python3 and R, wrapping barrnap, SMALT, SPAdes, and samtools
- 5 stages:
  1. Identify rDNA clusters
  2. Extracts reads mapping to a cluster
  3. Assemble into long reads
  4. Repeat (3x default) to extend
  5. Submit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

DOES IT WORK?

Benchmarking

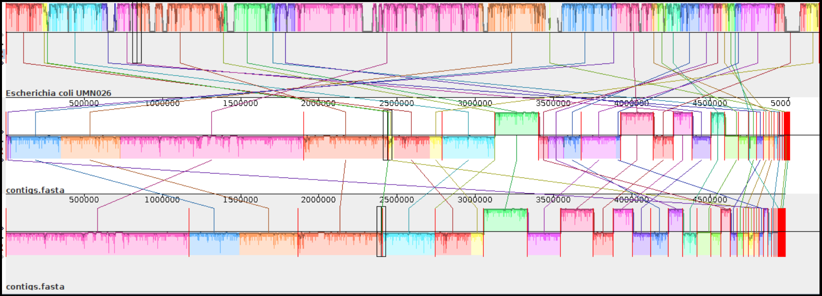
1. synthetic reads on synthetic genome (7 E. coli Sakai rDNAs separated by 6kb random sequence)
2. synthetic reads on real genome
3. short reads from hybrid assembly
4. GAGE-B datasets

Synthetic reads on synthetic genome



There are 8 LCBs with minimum weight 6690 | toy_genome | 20966 | LCB length: 14348 | LCB weight: 12971

Synthetic reads on real genome



Benchmarking with hybrid assembly

Mauve Demo

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 does’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

- 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

OÉ Gaillimh
NUI Galway

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

The James Hutton Institute

- Leighton Pritchard
- Ashleigh Holmnes

Acknowledgements

OÉ Gaillimh
NUI Galway

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

The James
Hutton
Institute

- Leighton Pritchard
- Ashleigh Holmnes

Questions?