BroadE Workshop: Genome Assembly

March 20th, 2013

BROAD INSTITUTE

Introduction & Logistics

De-Bruijn Graph Interactive Problem (45 minutes)

Assembly Theory Lecture (45 minutes)

Break (10-15 minutes)

Assembly in Practice Lecture (30 minutes)

Assembly Analysis Lecture (45 minutes)

Break (10-15 minutes)

Assembly Analysis Interactive Problem (45 minutes)

Instructors

Sante Gnerre

Sante has been working on assemblers for more than XII years, first as part of David Jaffe's group developing ARACHNE and ALLPATHS-LG, then as part of the Genome Assembly & Analysis Group (GAAG) working on reference-assisted assembler technology. He is now part of the BTL working on furthering assembly and novel new technologies.

Aaron Berlin

Aaron has been analyzing assemblies for 6 years. As part of GAAG, he specialized in analysis and assembly with new sequencing technologies and assemblies of large vertebrate genomes. Aaron is now part of the BTL, still keeping up on the cutting edge sequencing technologies

Sean Sykes

Sean has been analyzing assemblies for 7 years. As part of GAAG, Sean was lead on assembling an amazing amount of reference bacteria as part of the Human Microbiome project. Sean now leads the team that builds our GAEMR assembly analysis software and maintains our high-throughput assembly analysis pipelines.

De-Bruijn Graph Assembly Exercise

Workshop Overview

1. Assembly Theory
   - WGS Assembly Primer
   - Sanger Read Assemblers
   - New Technologies
   - Short Read Assemblers
2. Assembly in Practice
   - What makes a good assembly?
   - How do genome and sequencing issues impact assembly?
3. Assembly Analysis
   - Contiguity
   - Completeness
   - Correctness
   - Putting It All Together

Workshop Overview

1. Assembly Theory
   - WGS Assembly Primer
   - Sanger Read Assemblers
   - New Technologies
   - Short Read Assemblers
2. Assembly in Practice
   - What makes a good assembly?
   - How do genome and sequencing issues impact assembly?
3. Assembly Analysis
   - Contiguity
   - Completeness
   - Correctness
   - Putting It All Together

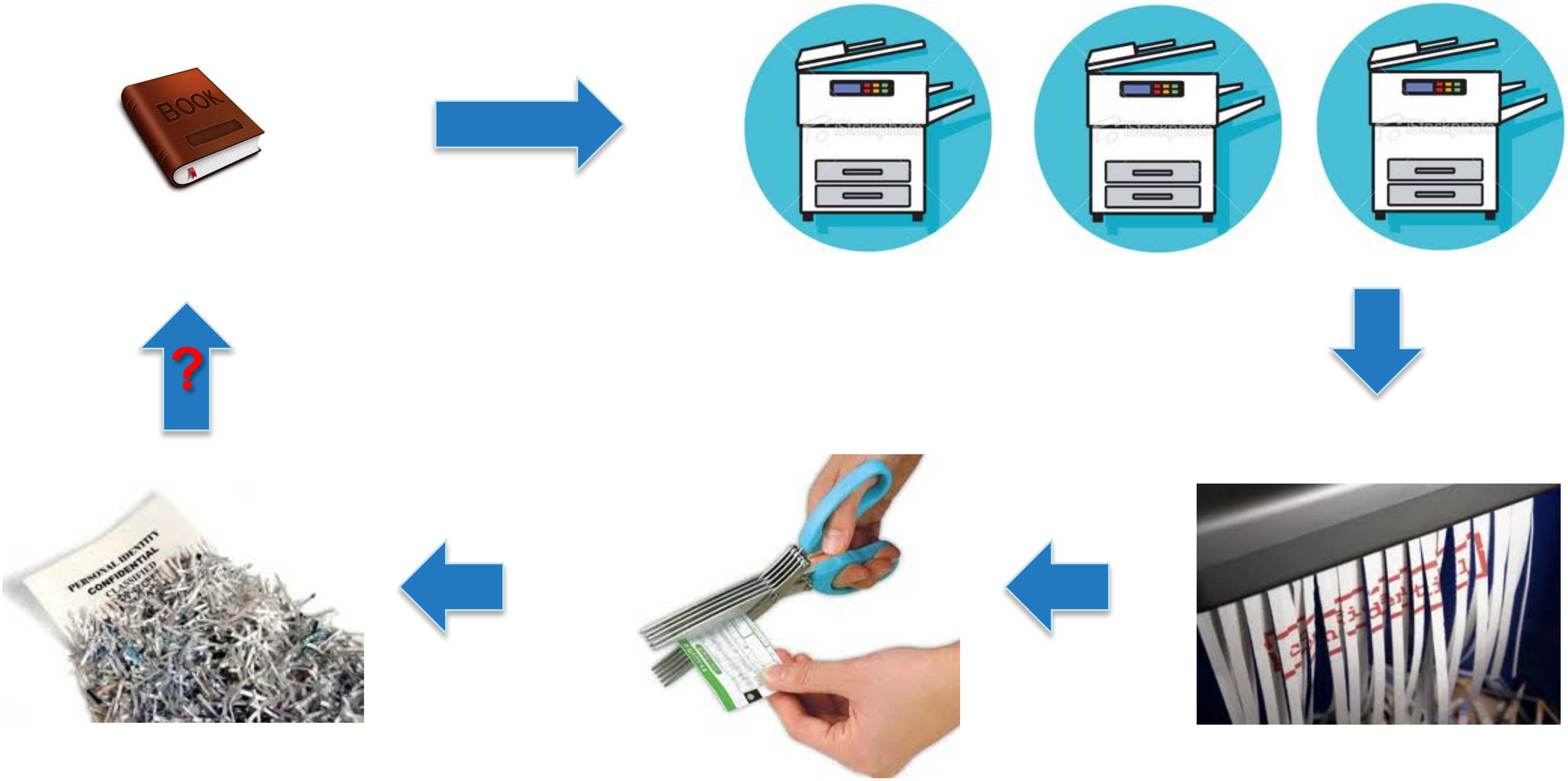
Assembly Theory Overview

- WGS Assembly Primer
- Sanger Read Assemblers
- New Technologies
- Short Read Assemblers

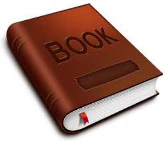
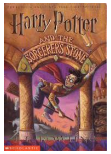
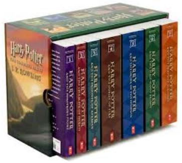
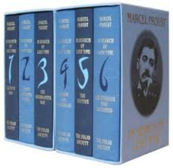
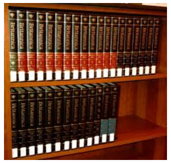
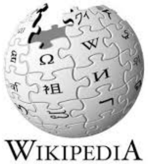
Assembly Theory Overview

- WGS Assembly Primer
- Sanger Read Assemblers
- New Technologies
- Short Read Assemblers

What is WGS Assembly?



What is WGS Assembly?

	Estimated genome size
	350 Kb
	4.9 Mb
	5.4 Mb
	180 Mb
	11.2 Gb

Really, What is WGS Assembly?

- Read sequences of C, G, T, As from a given organism
- We do not know where each sequence comes from
- Length varies
- Quality varies
- Enough sequences to cover DNA many times (coverage)
- Automatic, relatively inexpensive lab process
- Very hard algorithmically

Why is it a Hard Problem?

Challenges

Polymorphism
Repeats

Sequencing Errors
Bias
Contamination

Engineering

Genomes Are Not Really Random

- Polymorphism
  - Humans are diploid (23 homologous pairs)
  - Reads from homologous regions may differ
- Repetitiveness
  - SINEs = Short INterspersed Elements
    - Usually ~500 b in length
    - About 1.5M in the human genome
  - LINEs = Long INterspersed Elements
    - Usually ~1 Kb in length
    - About 0.5M in the human genome
  - Large repeats, segmental duplications…
    - 40 Kb and more!

Sequencing is Not Perfect

- Sequencing Errors
  - Base accuracy varies - Phred scores
  - Logarithmically linked to probability of error
    - Q10: P[wrong base call] = 1 in 10
    - Q20: P[wrong base call] = 1 in 100
    - Q30: P[wrong base call] = 1 in 1,000
    - Q40: P[wrong base call] = 1 in 10,000
    - Q50: P[wrong base call] = 1 in 100,000
  - Q50 is considered very good
  - Thousands of errors for mammalian genomes!
- Cloning bias
  - Some regions not represented, some over-represented
  - Not truly random

Assembly Theory Overview

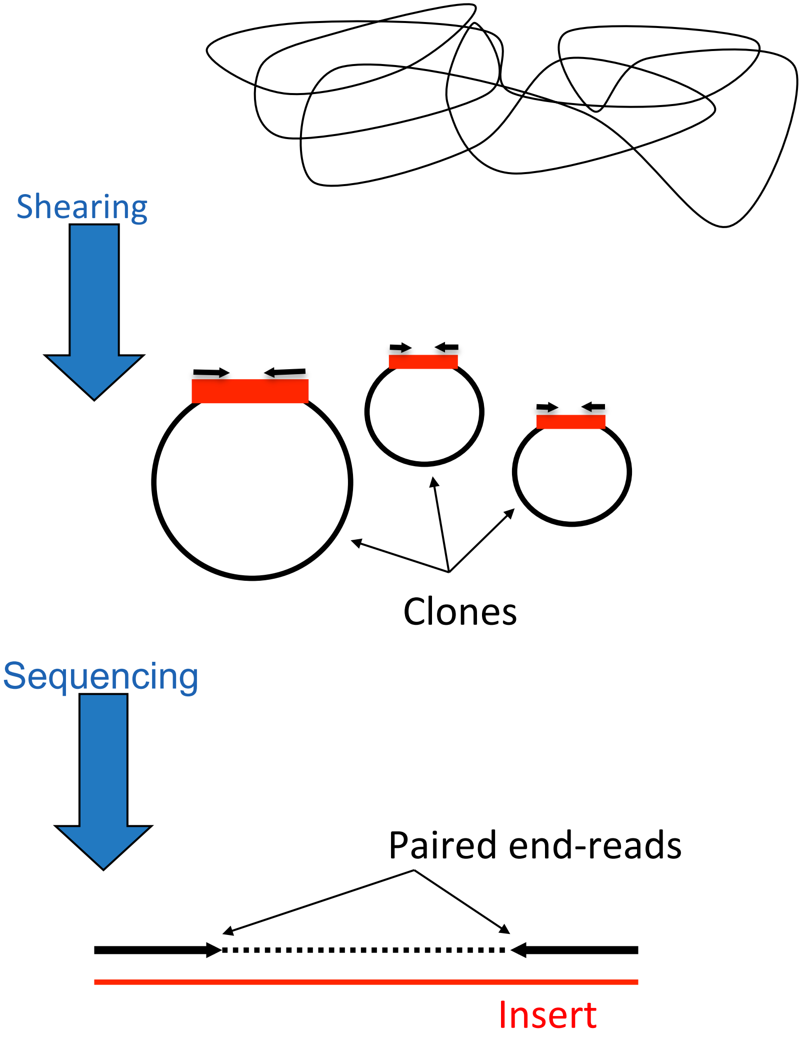
- WGS Assembly Primer
- Sanger Read Assemblers
- New Technologies
- Short Read Assemblers

Sanger Sequencing

Genomic DNA

Sequenced libraries

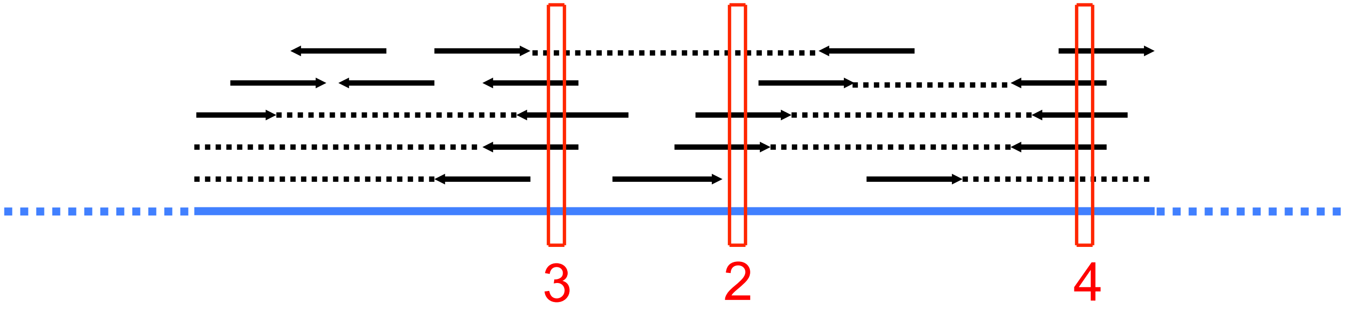
Paired end-reads



Sanger Data

- Inserts
  - Different library sizes (4Kb, 10Kb, 40Kb)
  - Sometimes, BAC ends (200Kb)
  - Length of inserts is known *probabilistically*
  - Some *chimerism* is expected
- Reads
  - Each comes with its own Phred scores
  - Average read length: 750 b
  - Average total coverage: 7X (it varies)

More About Coverage

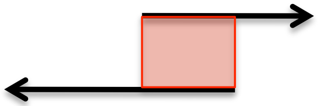


- Depth of coverage: how many *reads* on average cover any given base of the sequenced genome

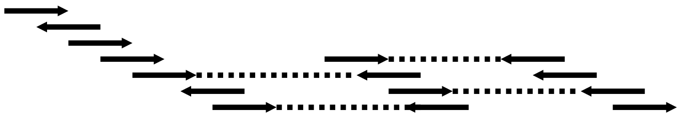
- It depends on the estimated genome size

The ARACHNE Assembler

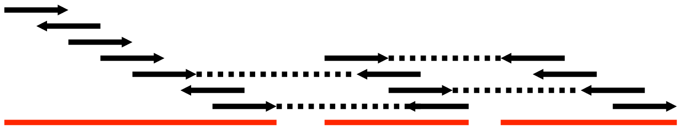
1. Find all read-read overlaps



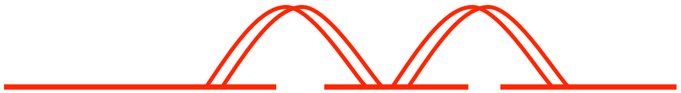
2. Layout



3. Consensus



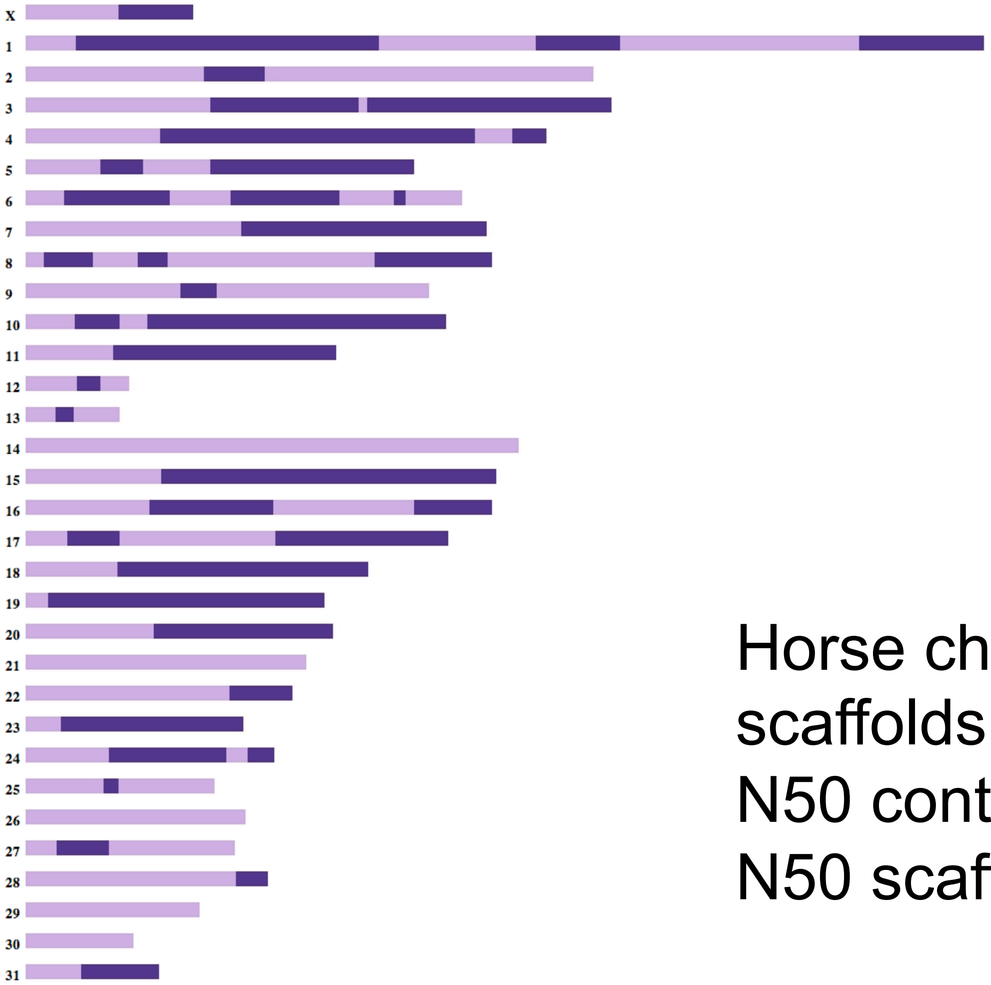
4. Scaffolds



Finding Read-Read Alignments is the Key

- If we had all and only the “true” aligns
  - The problem would be trivial
  - We could get a perfect answer
- Missing aligns
  - Sequencing errors
  - Short aligns are not detected
- Wrong aligns
  - Sequencing errors
  - Repeats

The Horse Genome



Horse chromosomes colored by scaffolds

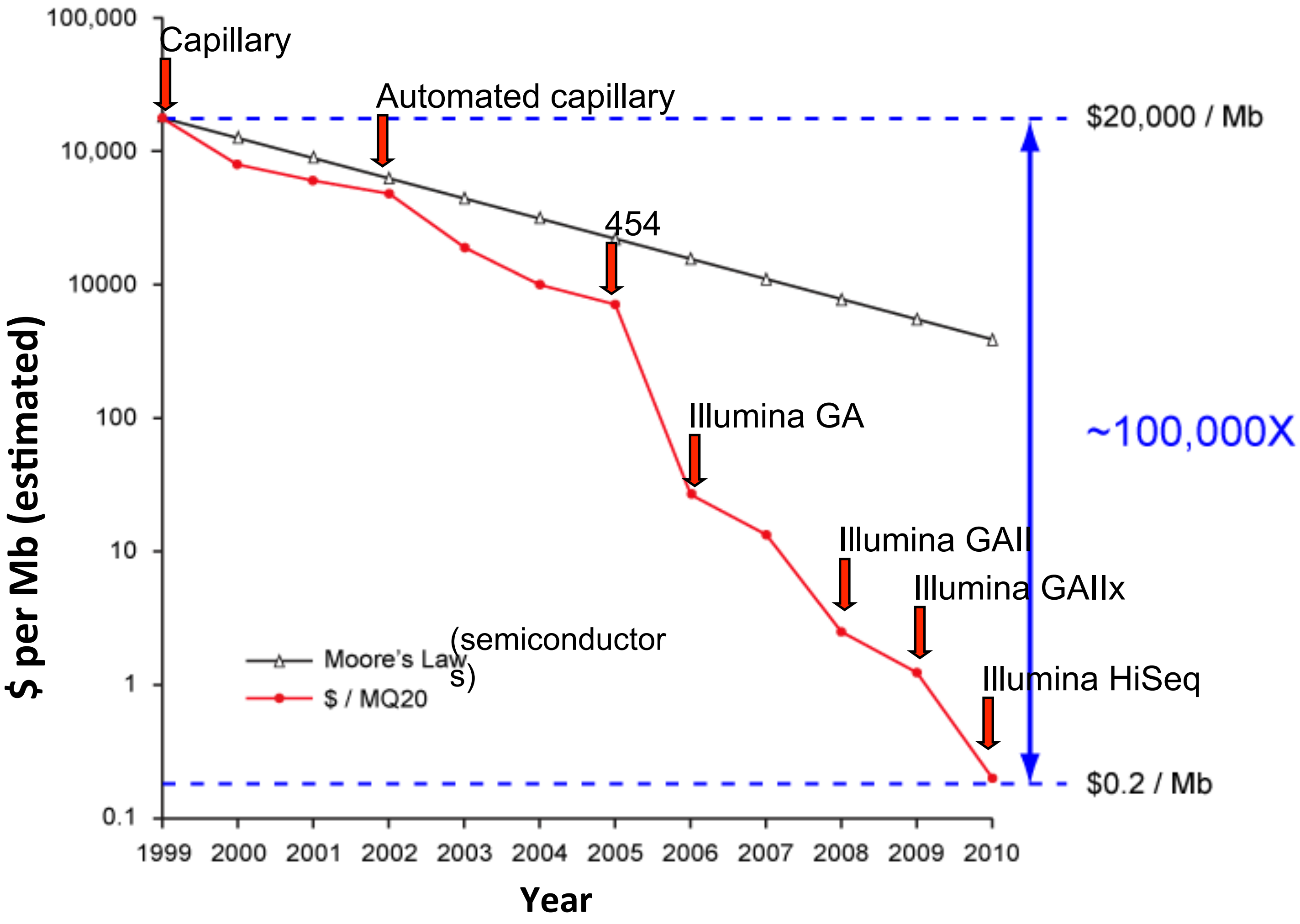
N50 contig size: 116 Kb

N50 scaffold size: 29 Mb

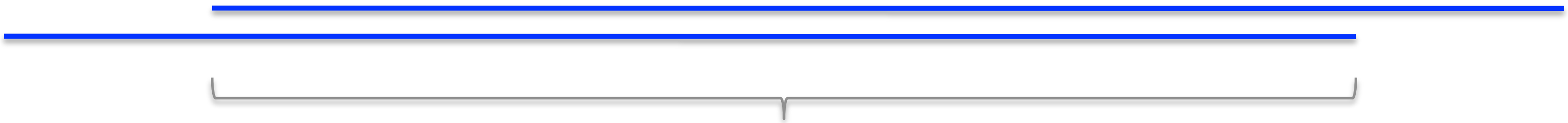
Assembly Theory Overview

- WGS Assembly Primer
- Sanger Read Assemblers
- New Technologies
- Short Read Assemblers

Sequencing Cost is Dropping Fast…

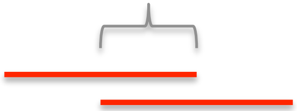


…But Reads Are Much Shorter



Sanger reads average length: ~750 b (long overlaps)

New sequencing technologies reads average length: ~100 b (short overlaps)



Assembly Theory Overview

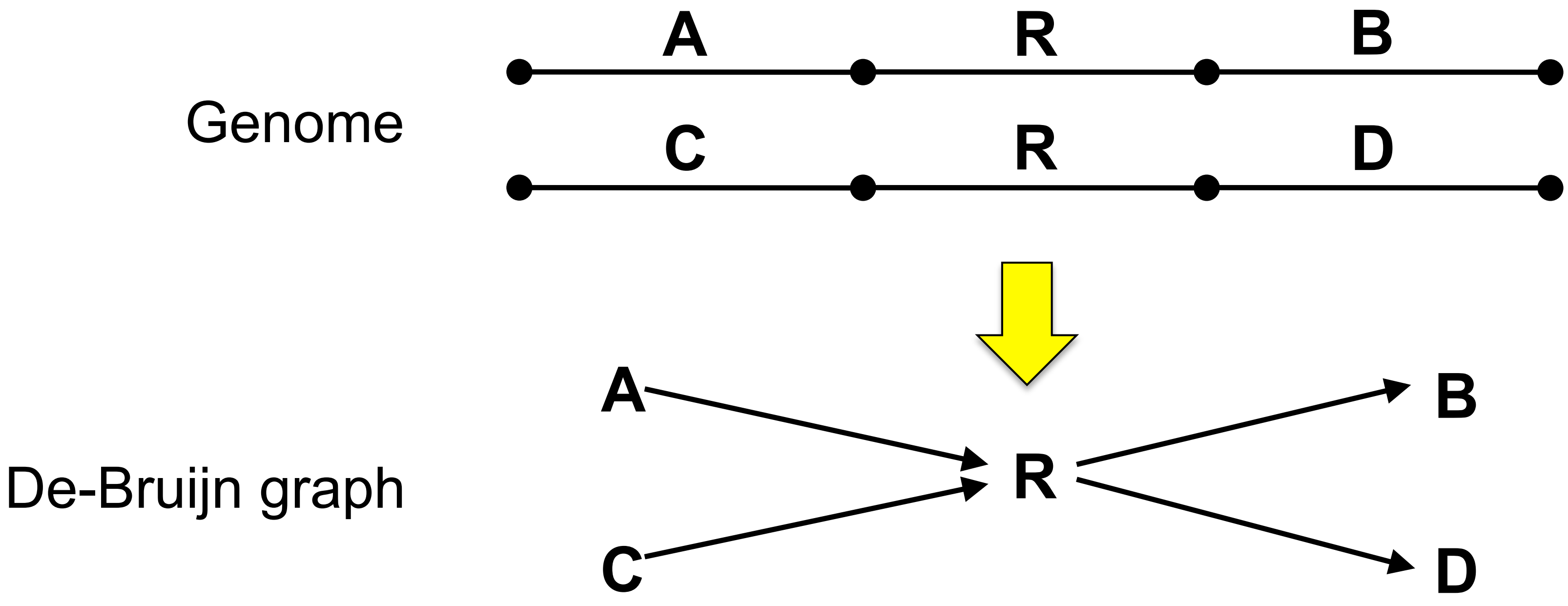
- WGS Assembly Primer
- Sanger Read Assemblers
- New Technologies
- Short Read Assemblers

Shorter Reads, So What?

- Compensate length with coverage
  - It is still vastly cheaper
  - Billions of reads in input
- Need to find a way to compress data
- De Bruijn graph

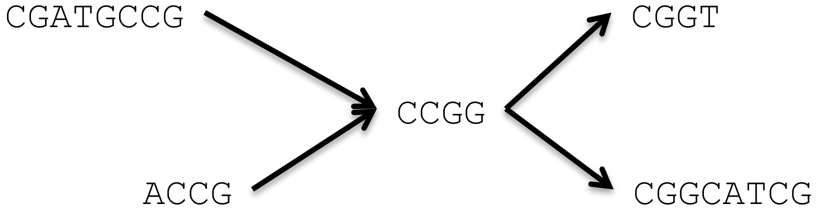
De-Bruijn Graph

- A Mathematical way to compress genomic data
- It depends on a chosen k-mer size (k)
- In brief:
  - Squeeze together perfect repeats of size ≥ k
  - Build a directed graph (edges are perfect k-1 overlaps)

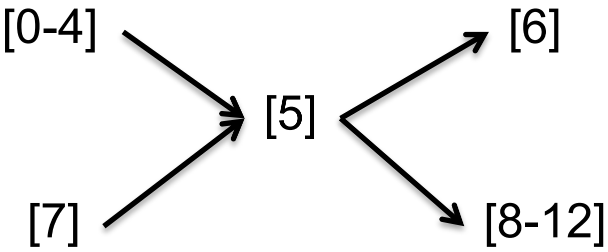


De-Bruijn Graph by Example

```
                  CGATGCCGGT
k-mer 0 →  CGAT
k-mer 1 →   CATG
k-mer 2 →    ATGC
…                 …
```



k-mer size = 4

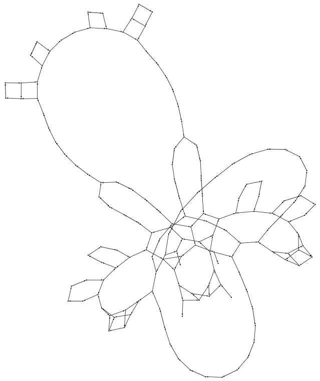
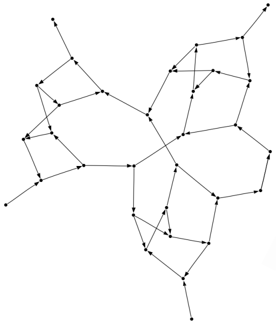
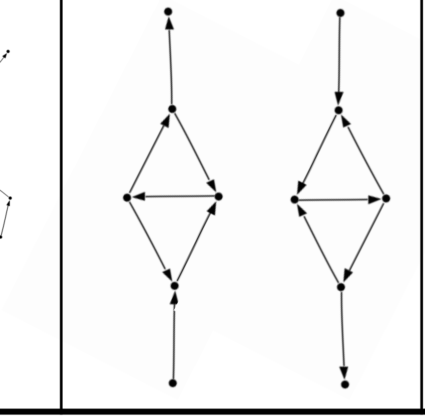
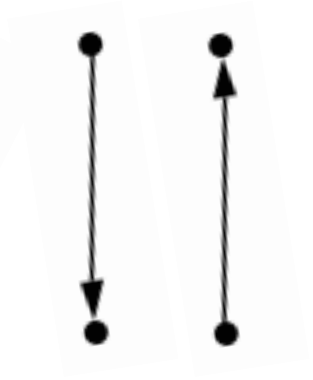


k-mer size = 8

CGATGCCGGTACCGGCATCG

[0-2]   [3-5]

Larger K is Better

C. jejuni – 2 Mb				
K	100	1,000	2,000	10,000
edges	236	44	14	2
graph				

Building De-Bruijn Graphs with Reads

- If
  - Reads are perfect (no errors)
  - Coverage is perfect (no cloning bias “holes”)
  - We know the “true” (haploid) genome
- And if
  - B1 := de-Bruijn graph built from the genome
  - B2 := de-Bruijn graph built from the reads
- Then
  - B1 = B2

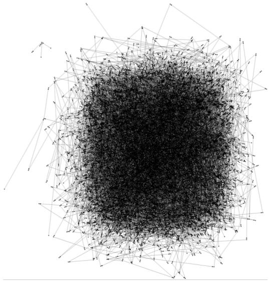
Using Reads: In the Real World

- Reads have errors
  - Must error correct reads first
  - The graph would explode otherwise
- Usually deal with polyploid genomes
  - Diploid differences appears as “bubbles”
- Sequencing bias
  - It causes loss of connectivity in the graph

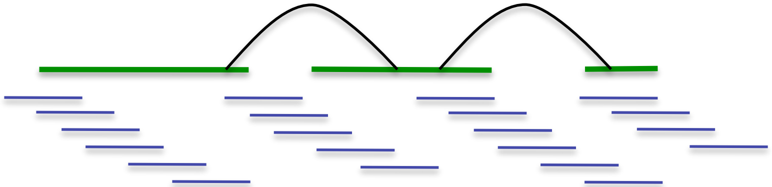
Illumina Data

- Inserts
  - Different library sizes (frags, jumps, …)
  - Length of inserts is known *probabilistically*
  - Chimerism expected (jumps, long-jumps)
- Reads
  - Each comes with its own Phred scores
  - Read length: ***101 b***
  - Average coverage: about ***100X***

ALLPATHS-LG

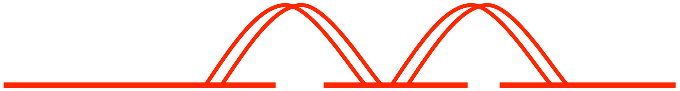


1. Error correct, and build unipath graph



2. Localize using jump reads

3. Build contigs

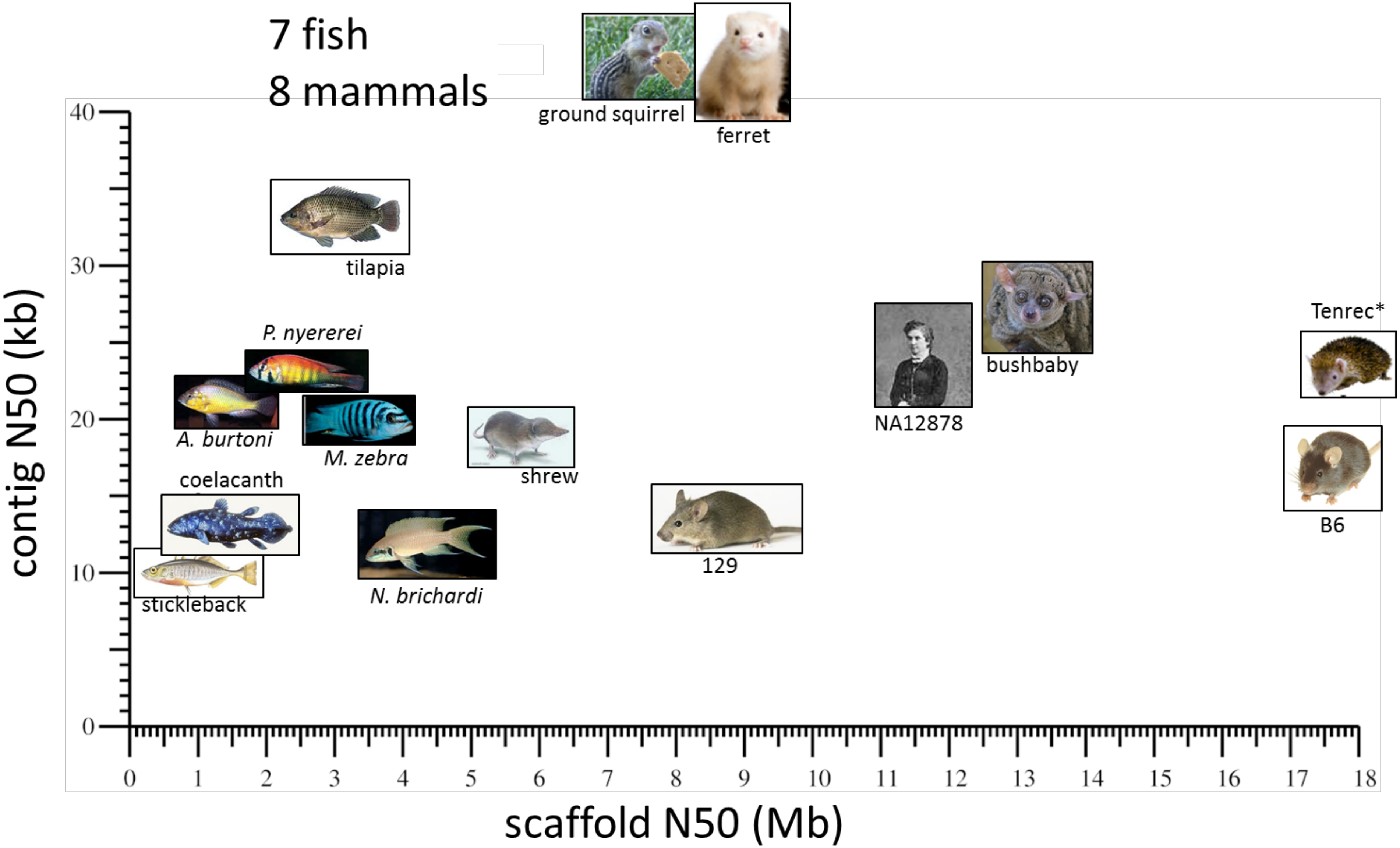


4. Join contigs in scaffolds

It Works on Small and Large Genomes

7 fish

8 mammals



Assembly: Still an Open Problem

- By and large, it works, but caveat emptor!
- Some genomes are hard, or impossible
  - Large nuclear size
  - Very polymorphic
  - Too repetitive
- Assessing assemblies is difficult
  - What to expect in output?
  - How to find problems?
  - How to compare different assemblies?

Questions?

Break Time!

Please Enjoy a Short Break!

Workshop Overview

1. Assembly Theory
   - WGS Assembly Primer
   - Sanger Read Assemblers
   - New Technologies
   - Short Read Assemblers
2. Assembly in Practice
   - “Good” Assemblies
   - Limitations of a good assembly
   - Where is my Gene?
3. Assembly Analysis
   - Contiguity
   - Completeness
   - Correctness
   - Putting It All Together

Assembly in Practice Overview

- “Good” Assemblies
- Limitations of Good Assemblies
- Where is my Gene?

Assembly in Practice Overview

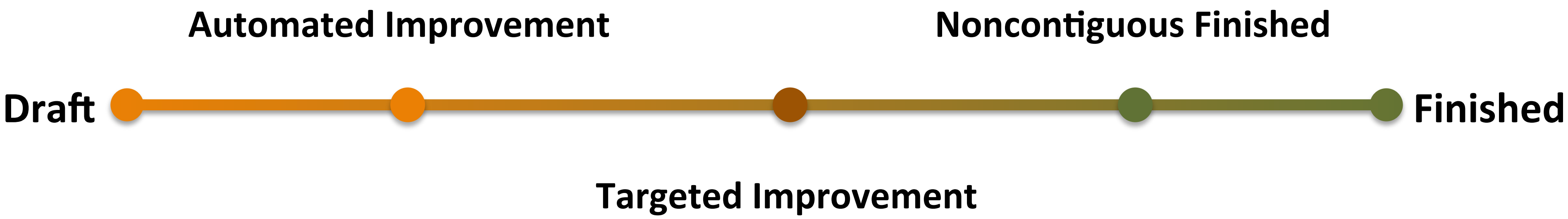
- “Good” Assemblies
- Limitations of Good Assemblies
- Where is my Gene?

Everyone wants one

but hard to define

User Defines a “Good” Assembly

- Depends on the goals and purpose
  - Contigs vs. scaffolds
  - Base quality
  - Repeat content
- Spectrum of assembly products



Automated Improvement

Noncontiguous Finished

Draft

Finished

Targeted Improvement

Assembly in Practice Overview

- "Good" Assemblies
- Limitations of Good Assemblies
- Where is my Gene?

Limitations to a “Good” Assembly

Challenges

Polymorphism

Repeats

Sequencing Errors

Bias

Contamination

Engineering

Effect of Polymorphism

Challenges

Polymorphism

Repeats

Sequencing Errors

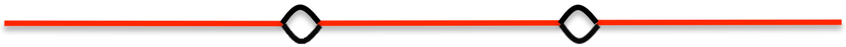
Bias

Contamination

Engineering

- Polymorphism creates local complexity in the graph
- This can lead to:
  - Inability to simplify the graph (contig breaks)
  - Incorrectly simplifying the graph (misassemblies)
- Result: Gaps, Small Contigs, Misassemblies

```
            CGAWGCCGGT
k-mer 0 →   CGAA
k-mer 1 →    CAAG
k-mer 2 →     ATGC
k-mer 3 →      TGCC
```



Effect of Genomic Repeats

Challenges

Polymorphism

Repeats

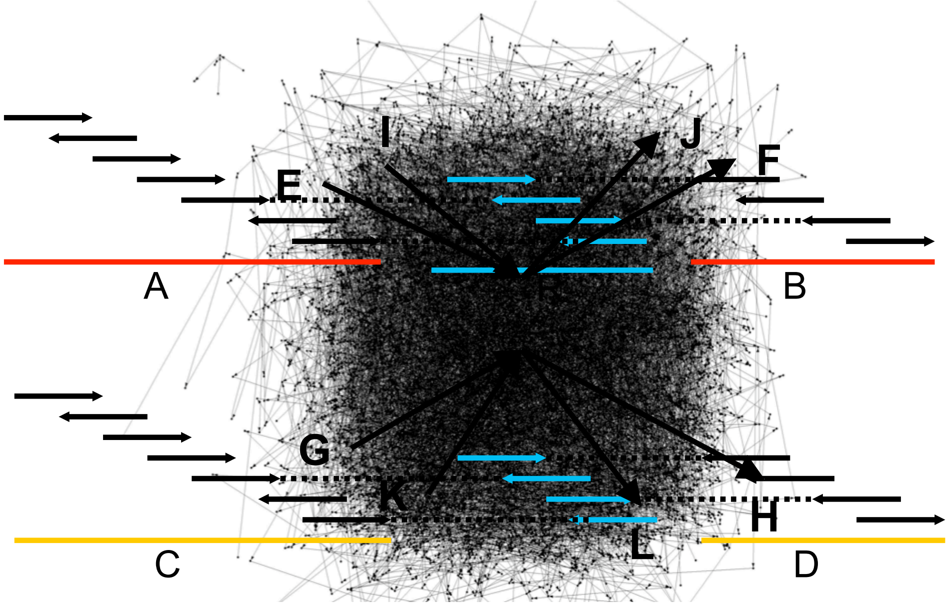
Errors

Bias

Contamination

Engineering

- Create a tangled graph
- Read pairs can help to untangle
  - Span across repeats
  - Reach in from unique on each side
- Result: Collapsed Repeats, Misassemblies, Gaps



Effect of Sequencing Errors

Assembly Challenges

Polymorphism

Repeats

Sequencing Errors

Bias

Contamination

Engineering

- Random errors can be corrected
- Systematic errors can accumulate
  - Looks like polymorphism in the assembly
- Result: Gaps, Consensus Errors

```
Genome:      CGATGCCGGT
k-mer 0 →    CGAA
k-mer 1 →     CAAG
k-mer 2 →      ATGC
k-mer 3 →       TGCC
Consensus:   CGAAGCCGGT
```

Effect of Sequencing Bias

Assembly Challenges

Polymorphism

Repeats

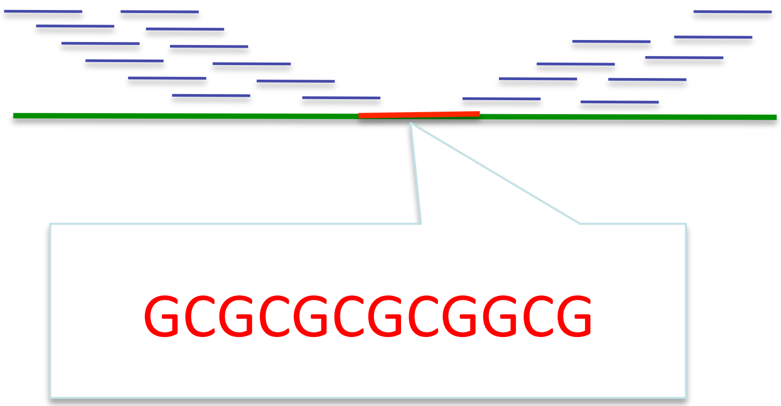
Errors

Bias

Contamination

Engineering

- Certain patterns of DNA can be recalcitrant to Illumina sequencing
  - coverage will drop close to zero
- Some library preparation techniques unevenly amplify DNA,
  - Areas of very low and very high coverage
- Result: Gaps



Effect of Contamination

Assembly Challenges

Polymorphism

Repeats

Errors

Bias

Contamination

Engineering

- Contamination does not typically affect the building of your assembly
  - Reduces true input coverage
  - Causes problems when using the assembly
- Can enter at any stage of the process
  - Commonly due to inefficient DNA extraction
- Result: More contigs, larger assembly size

Effect of Compute Limitations

Assembly Challenges

Polymorphism

Repeats

Errors

Bias

Contamination

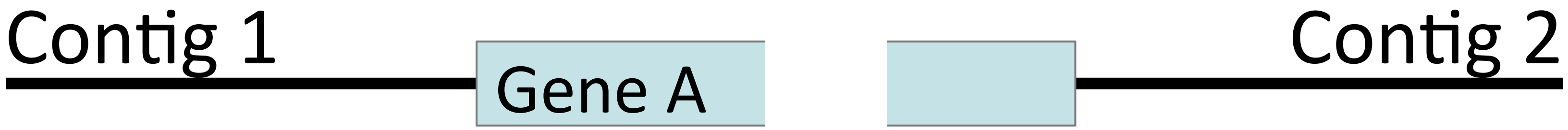
Engineering

- Problem is too complex
- Common reasons for assemblies crashing:
  - Reads have too many errors
    - Error correction is too complex
  - Genome is too repetitive
    - Assembler will stall sorting out repeats
  - Genome is too large, or you have too much data
    - Machine to runs out of memory
- Result: No Assembly

Assembly in Practice Overview

- “Good” Assemblies
- Limitations of Good Assemblies
- Where is my Gene?

Gene Broken


Contig 1

Gene A

Contig 2

Why Is Gene Broken?

- Repeats

Contiguity Problem

- Contamination
- Bias
- Data quality

Gene Missing

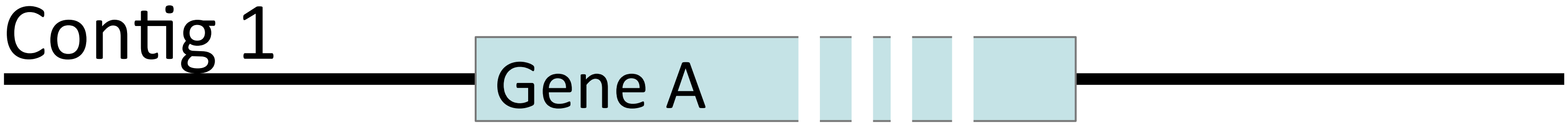
Contig 1

Why Is Gene Missing?

- True deletion
- Misassembly

Completeness Problem

Gene Differs


Contig 1

Gene A

Why Does Gene Differ?

- True variation

Correctness Problem

- Data quality
- Misassembly

Questions?

Break Time!

Please Enjoy a Short Break!

Workshop Overview

1. Assembly Theory
   - WGS Assembly Primer
   - Sanger Read Assemblers
   - New Technologies
   - Short Read Assemblers
2. Assembly in Practice
   - “Good” Assemblies
   - Limitations of a good assembly
3. Assembly Analysis
   - Contiguity
   - Completeness
   - Correctness
   - Putting It All Together

Assembly Analysis Overview

- Source of Problems
- How to Identify Problems
- Putting the Pieces Together

Source of Assembly Issues

Challenges

Polymorphism

Repeats

Sequencing Errors

Bias

Contamination

Engineering

How To Identify Problems

- Contiguity

  “Long contigs and scaffolds”

- Completeness

  “Minimal missing sequence”

- Correctness

  “Few assembly errors”

How To Identify Problems

- Contiguity

  “Long contigs and scaffolds”

- Completeness

  “Minimal missing sequence”

- Correctness

  “Few assembly errors”

Contiguity Questions

- “Why is my gene broken”?
- How many pieces?
- How large are the pieces?
- In line with expectations?
- Phenotypes indicate potential problems?

Contiguity Analysis

- Total Number
- Total Size
- N50 Size
- Ungapped vs. gapped size

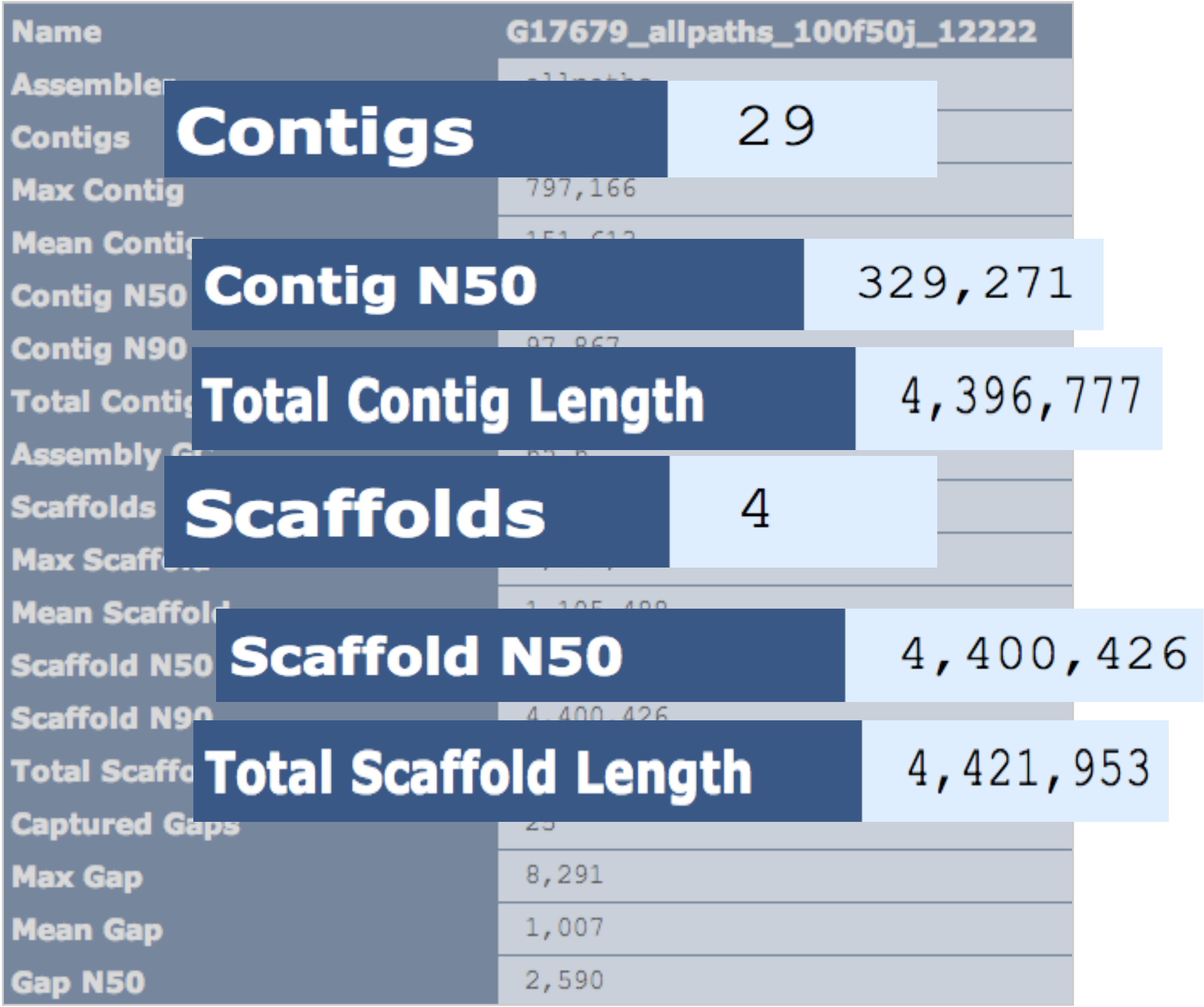
N50 Size Calculation

- Length-weighted median
- Sort sizes from largest to smallest
- Sum sizes to get total length
- Find contig size where sum >= ½ assembly size

Sizes
1,000
1,500
3,000
4,000
1,000
1,000
500

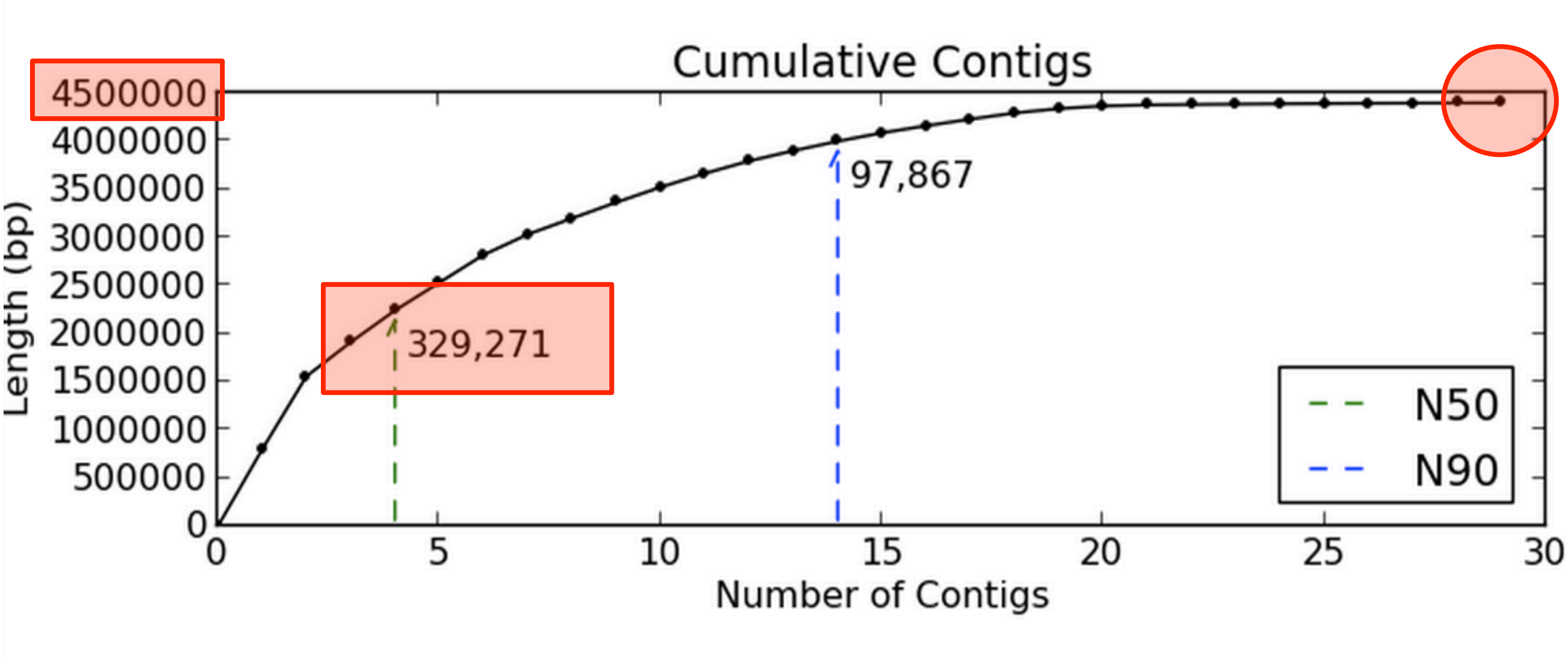
N50 Size

Contiguity Stats Table



Name	G17679_allpaths_100f50j_12222
Assembler	allpaths
Contigs	29
Max Contig	797,166
Mean Contig	151,613
Contig N50	329,271
Contig N90	97,867
Total Contig Length	4,396,777
Assembly GC	[illegible]
Scaffolds	4
Max Scaffold	[illegible]
Mean Scaffold	1,105,488
Scaffold N50	4,400,426
Scaffold N90	4,400,426
Total Scaffold Length	4,421,953
Captured Gaps	25
Max Gap	8,291
Mean Gap	1,007
Gap N50	2,590

Cumulative Sizes Plots



How To Identify Problems

- Contiguity

  “Long contigs and scaffolds”

- Completeness

  “Minimal missing sequence”

- Correctness

  “Few assembly errors”

Completeness Questions

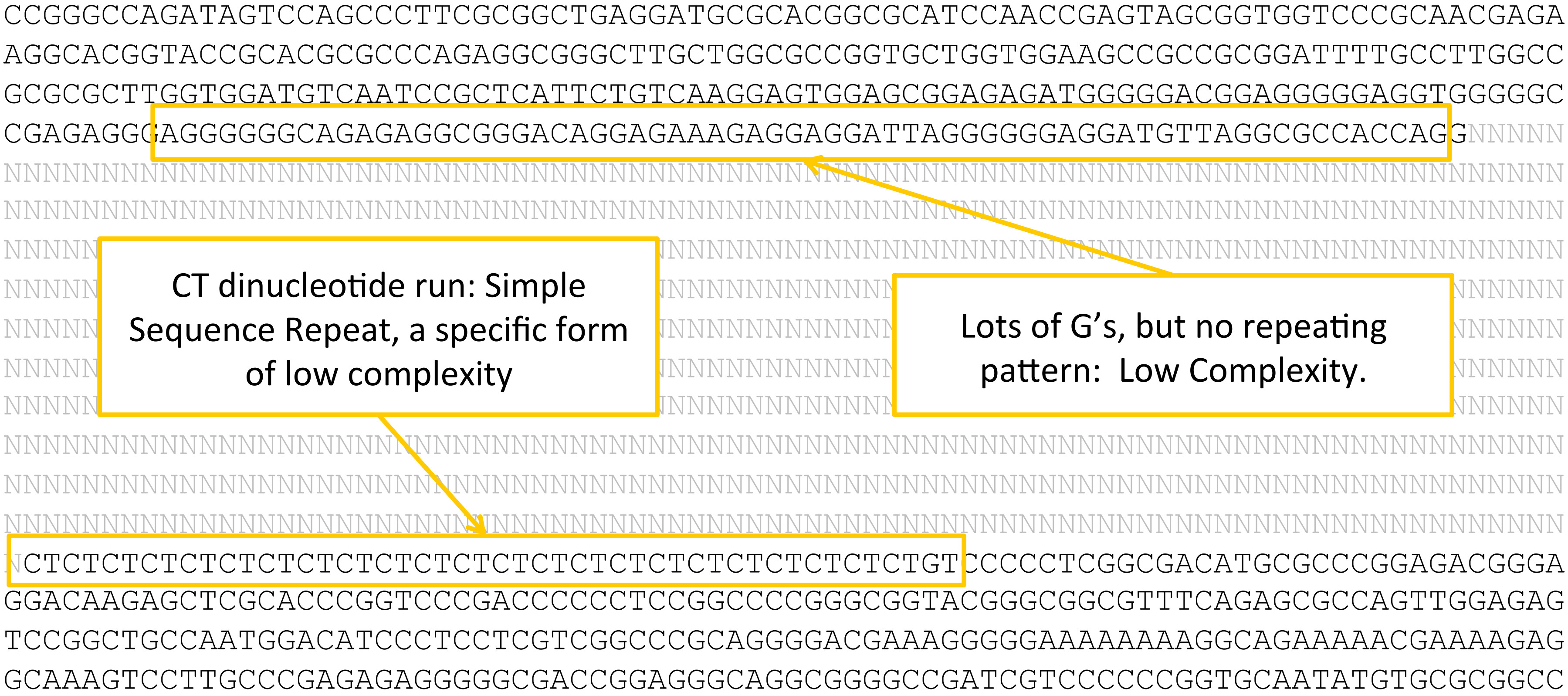
- “Why is my gene missing?”
- Any missing information?
- Have we used read data effectively?

Completeness Analysis

- Gap end sequence
- Read pair mapping
- Reference covered

Gap End Analysis

Example Gap Flanked by Low Complexity

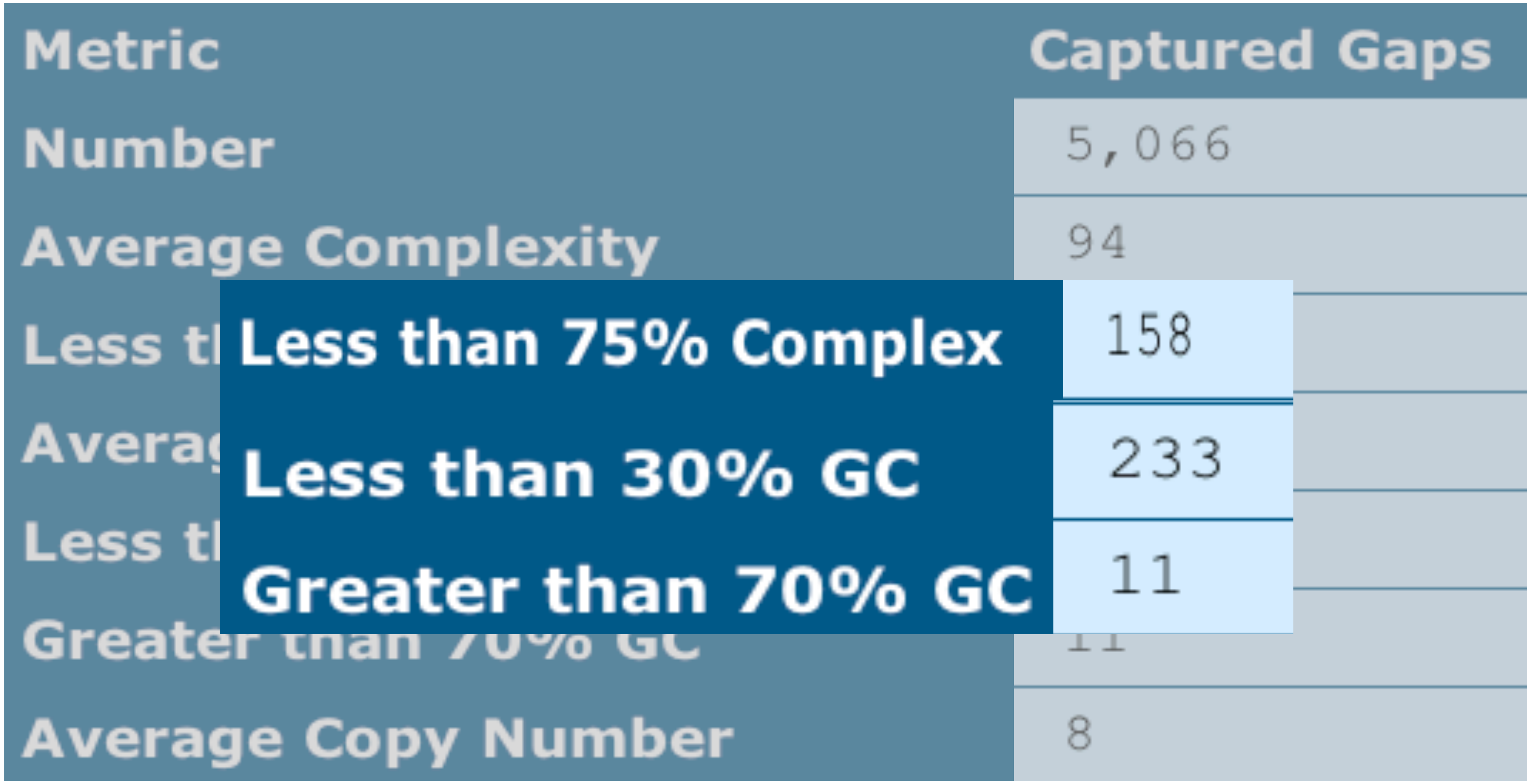


```
CCGGGCCAGATAGTCCAGCCCTTCGCGGCTGAGGATGCGCACGGCGCATCCAACCGAGTAGCGGTGGTCCCGCAACGAGA
AGGCACGGTACCGCACGCGCCCAGAGGCGGGCTTGCTGGCGCCGGTGCTGGTGGAAGCCGCCGCGGATTTTGCCTTGGCC
GCGCGCTTGGTGGATGTCAATCCGCTCATTCTGTCAAGGAGTGGAGCGGAGAGATGGGGGACGGAGGGGGAGGTGGGGGC
CGAGAGGGAGGGGGGCAGAGAGGCGGGACAGGAGAAAGAGGAGGATTAGGGGGGAGGATGTTAGGCGCCACCAGGNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNN                            NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNN                            NNNNNNNNNNNN                             NNNNNN
NNNNN                            NNNNNNNNNNNN                             NNNNNN
NNNNN                            NNNNNNNNNNNN                             NNNNNN
NNNNN                            NNNNNNNNNNNN                             NNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTCCCCCTCGGCGACATGCGCCCGGAGACGGA
GGACAAGAGCTCGCACCCGGTCCCGACCCCCCTCCGGCCCCGGGCGGTACGGGCGGCGTTTCAGAGCGCCAGTTGGAGAG
TCCGGCTGCCAATGGACATCCCTCCTCGTCGGCCCGCAGGGGACGAAAGGGGGAAAAAAAAGGCAGAAAAACGAAAAGAG
GCAAAGTCCTTGCCCGAGAGAGGGGGCGACCGGAGGGCAGGCGGGGCCGATCGTCCCCCCGGTGCAATATGTGCGCGGCC
```

CT dinucleotide run: Simple Sequence Repeat, a specific form of low complexity

Lots of G's, but no repeating pattern: Low Complexity.

Gap End Analysis

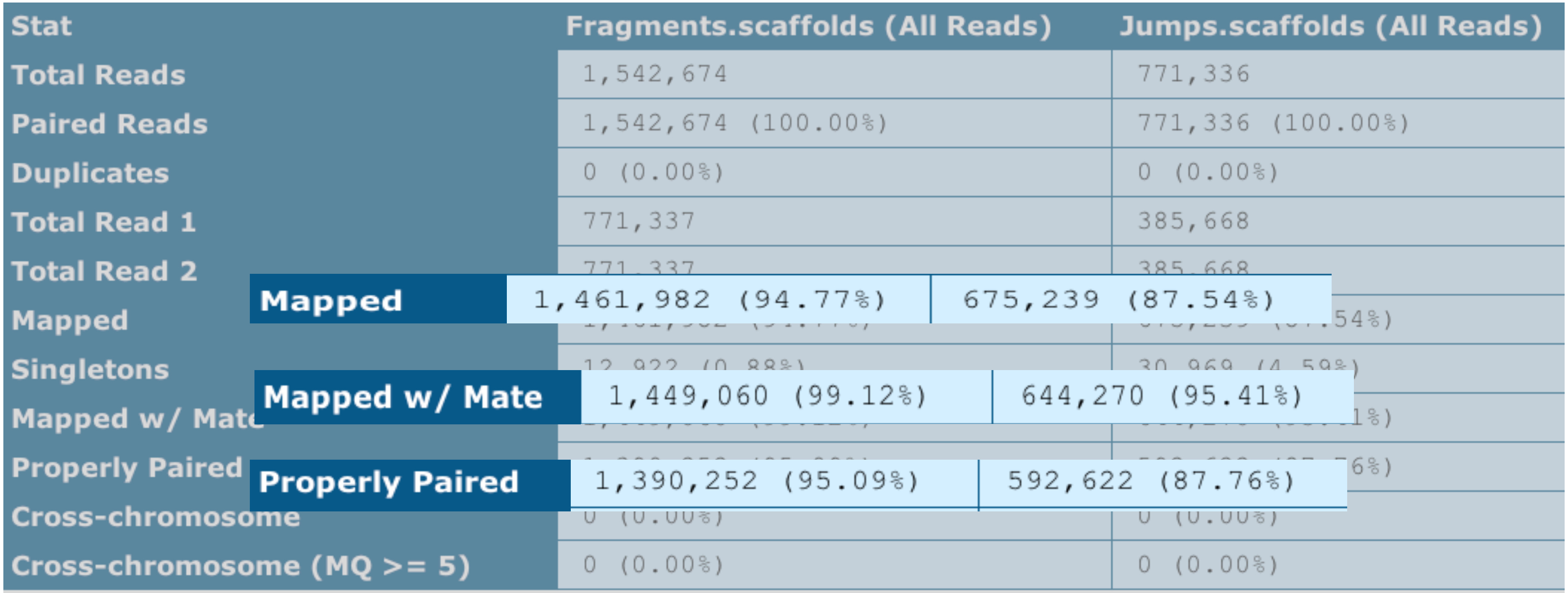


Metric	Captured Gaps
Number	5,066
Average Complexity	94
Less th	
Averag	
Less th	
Greater than 70% GC	11
Average Copy Number	8

Less than 75% Complex	158
Less than 30% GC	233
Greater than 70% GC	11

Read Mapping Stats

- Align read data back to scaffolds using BWA
- Using samtools, report alignment stats



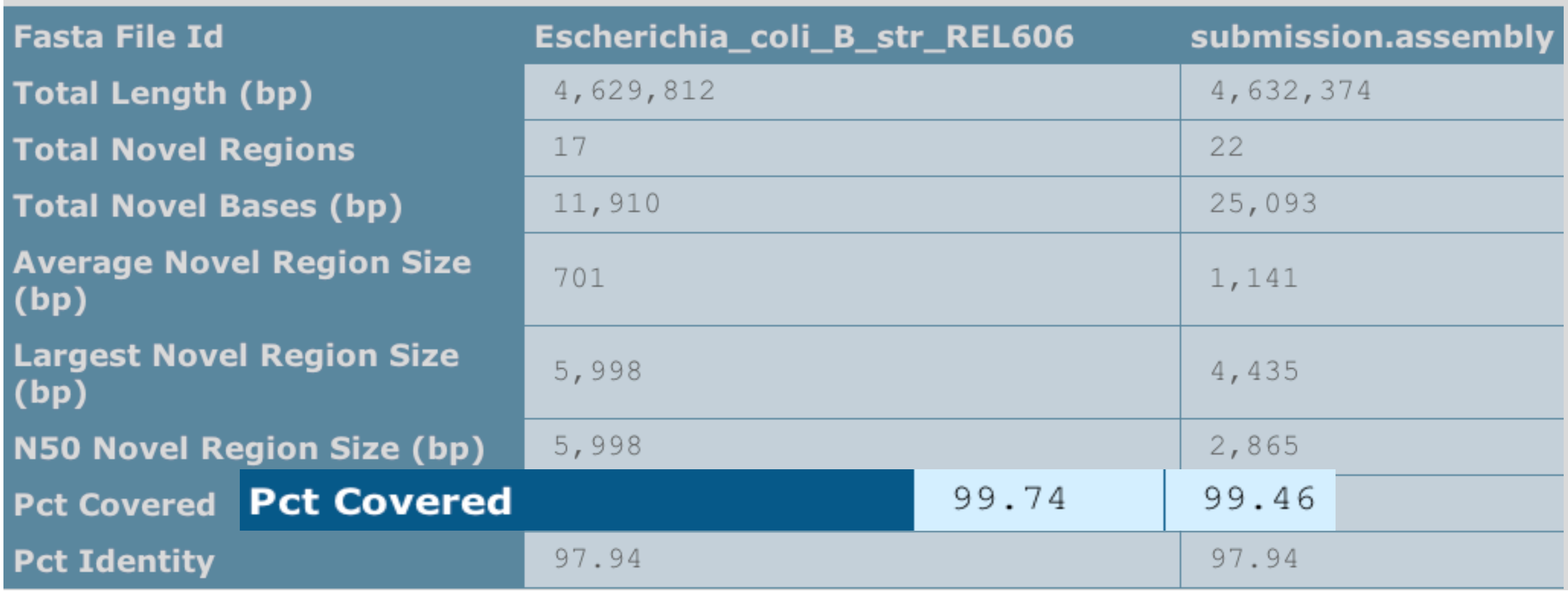
Stat	Fragments.scaffolds (All Reads)	Jumps.scaffolds (All Reads)
Total Reads	1,542,674	771,336
Paired Reads	1,542,674 (100.00%)	771,336 (100.00%)
Duplicates	0 (0.00%)	0 (0.00%)
Total Read 1	771,337	385,668
Total Read 2	771,337	385,668
Mapped	1,461,982 (94.77%)	675,239 (87.54%)
Singletons	12,922 (0.88%)	30,969 (4.59%)
Mapped w/ Mate	1,449,060 (99.12%)	644,270 (95.41%)
Properly Paired	1,390,252 (95.09%)	592,622 (87.76%)
Cross-chromosome	0 (0.00%)	0 (0.00%)
Cross-chromosome (MQ >= 5)	0 (0.00%)	0 (0.00%)

Mapped	1,461,982 (94.77%)	675,239 (87.54%)
Mapped w/ Mate	1,449,060 (99.12%)	644,270 (95.41%)
Properly Paired	1,390,252 (95.09%)	592,622 (87.76%)

http://samtools.sourceforge.net/

Comparison To Reference

- Nucmer for global alignment
- Parse coords output for coverage information



Fasta File Id	Escherichia_coli_B_str_REL606	submission.assembly
Total Length (bp)	4,629,812	4,632,374
Total Novel Regions	17	22
Total Novel Bases (bp)	11,910	25,093
Average Novel Region Size (bp)	701	1,141
Largest Novel Region Size (bp)	5,998	4,435
N50 Novel Region Size (bp)	5,998	2,865
Pct Covered	99.74	99.46
Pct Identity	97.94	97.94

mummer.sourceforge.net

How To Identify Problems

- Contiguity

  “Long contigs and scaffolds”

- Completeness

  “Minimal missing sequence”

- Correctness

  “Few assembly errors”

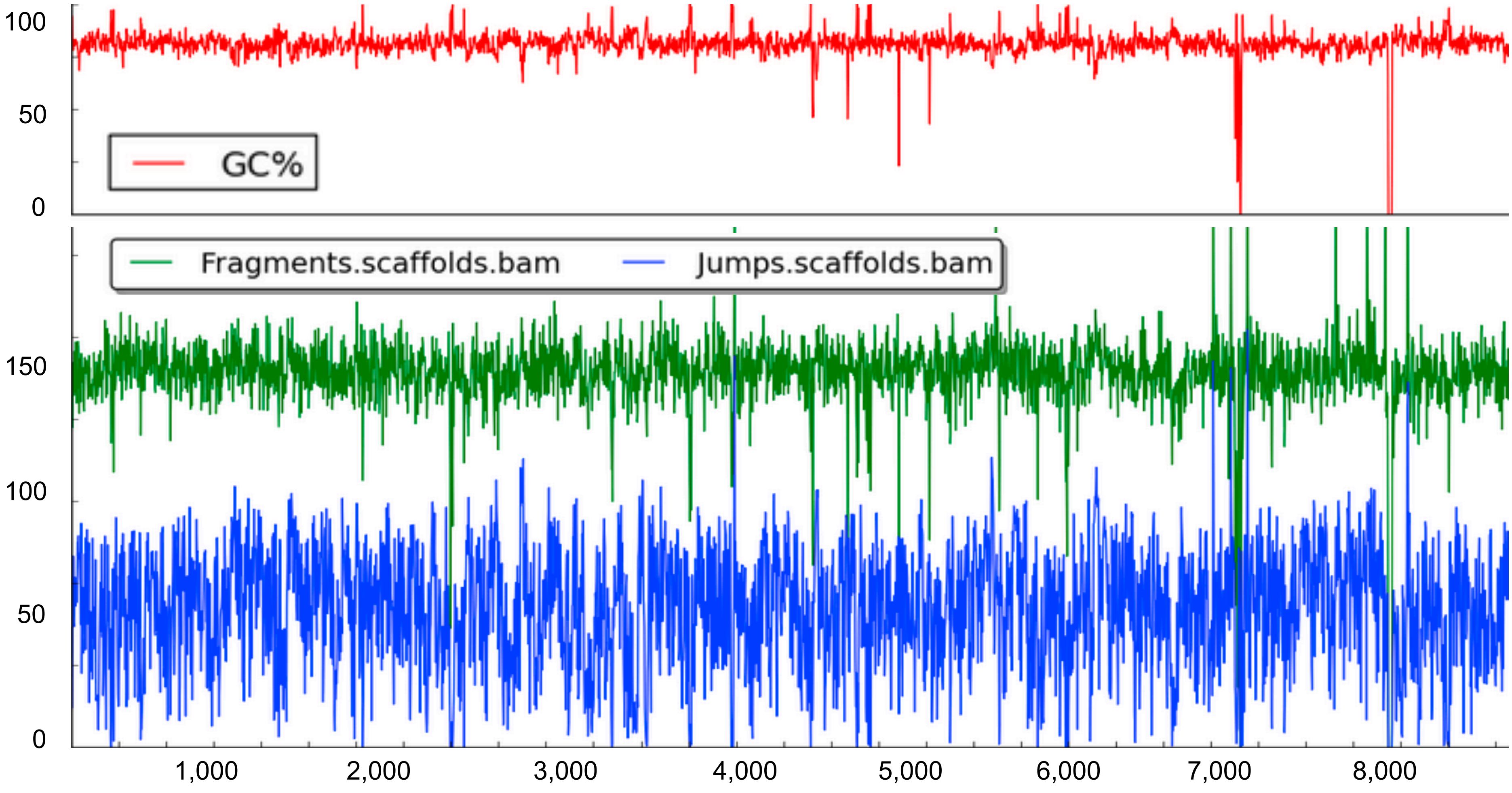
Correctness Questions

- “Why does my gene look different?”
- Do the read data look consistent?
- Does this assembly match what we expect?

Correctness Analysis

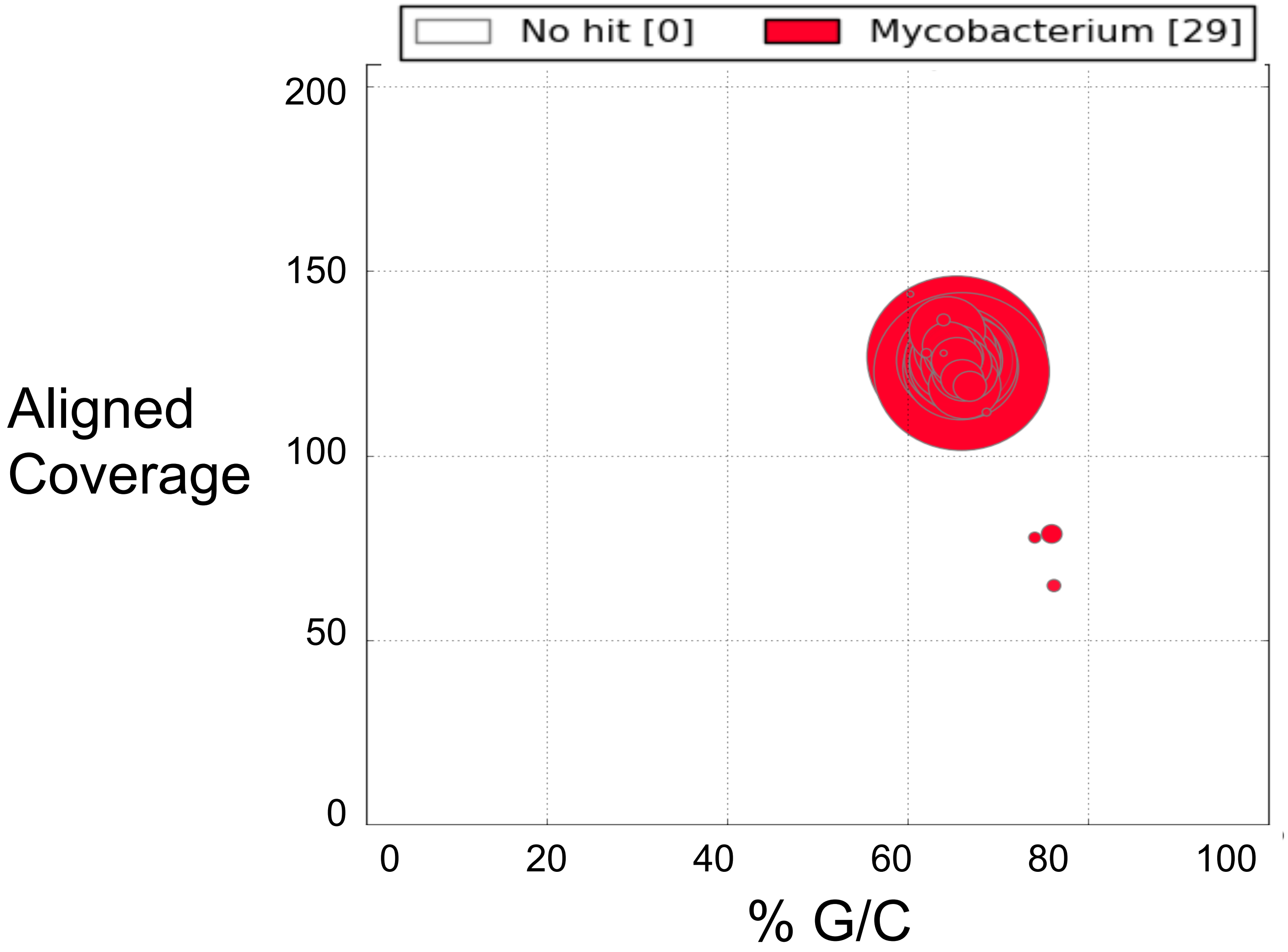
- Read coverage along assembly
- BLAST taxonomic classification
- Alignment to reference
- External genomic information

Read Coverage Along Assembly

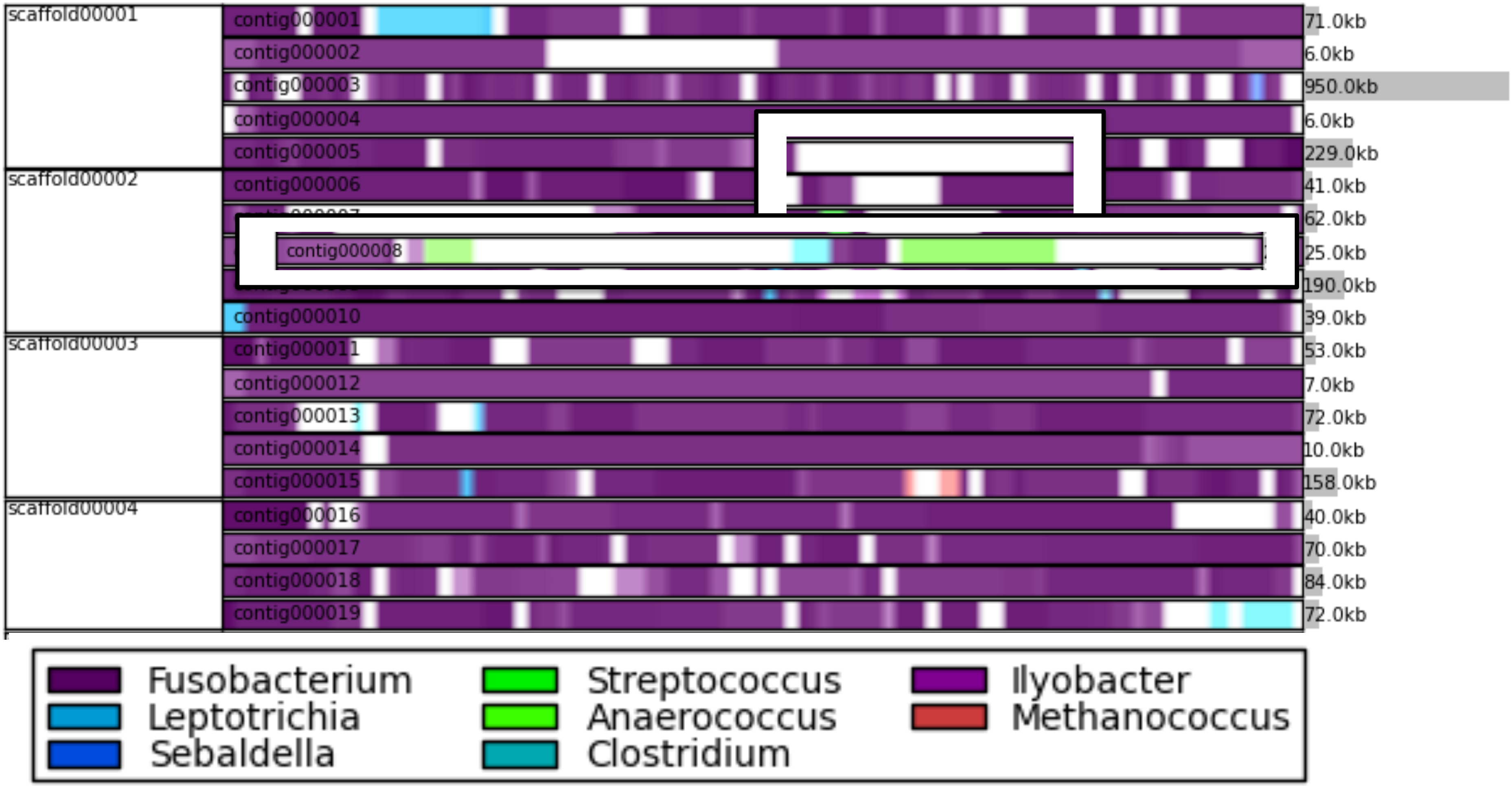


Scaffold00001

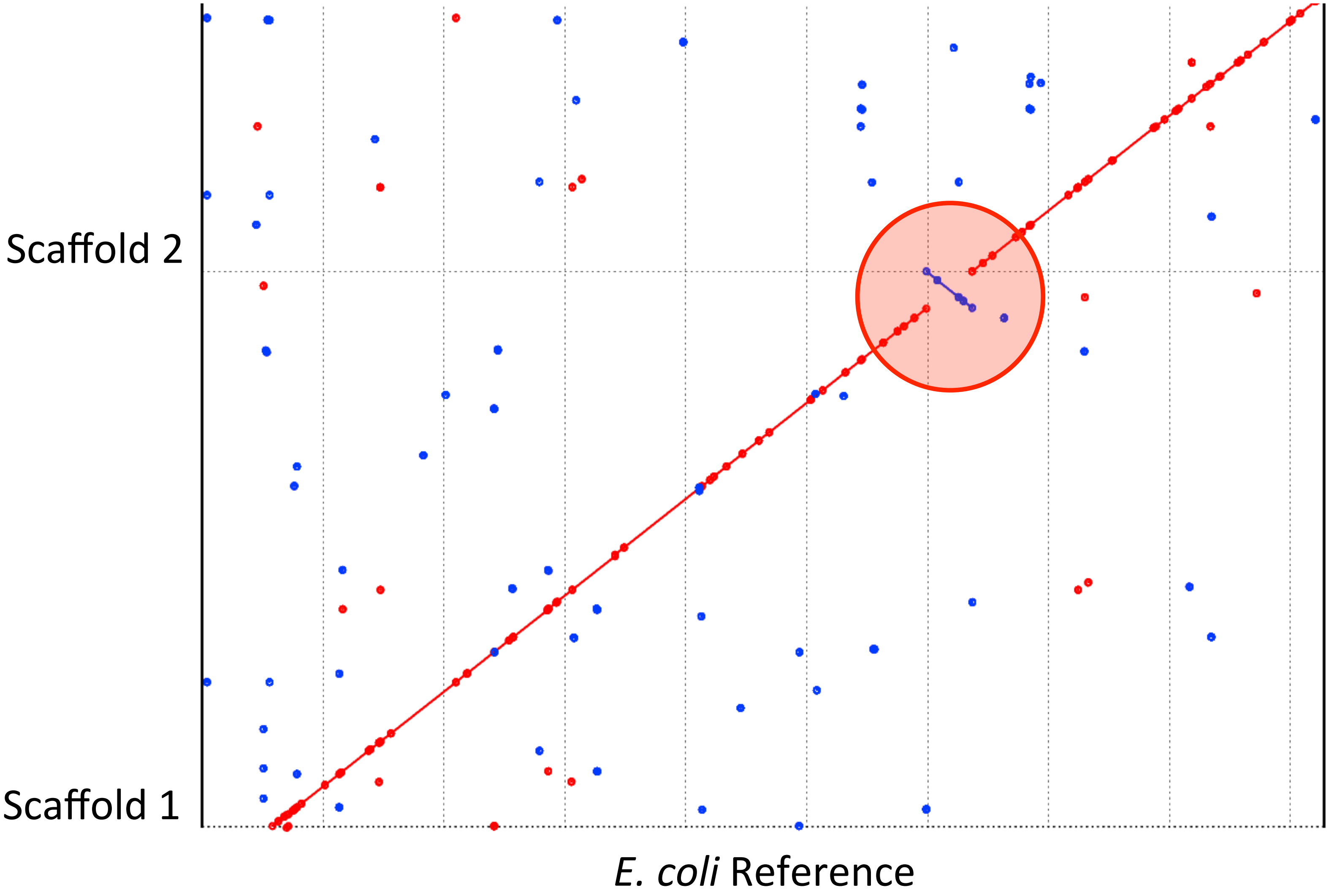
BLAST Bubbles



BLAST Heatmap

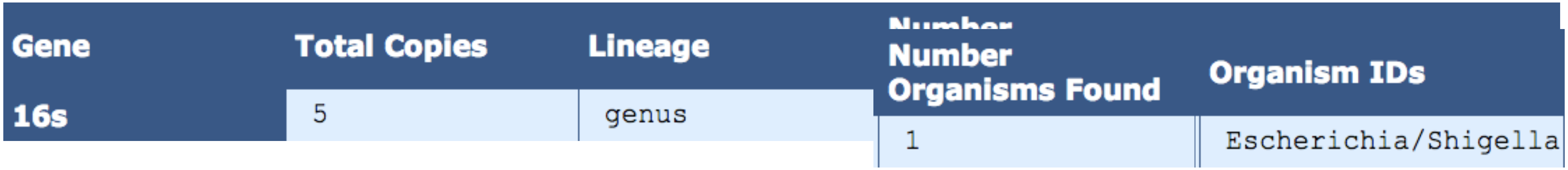


Alignment To Reference



mummer.sourceforge.net

16s Analysis Stats



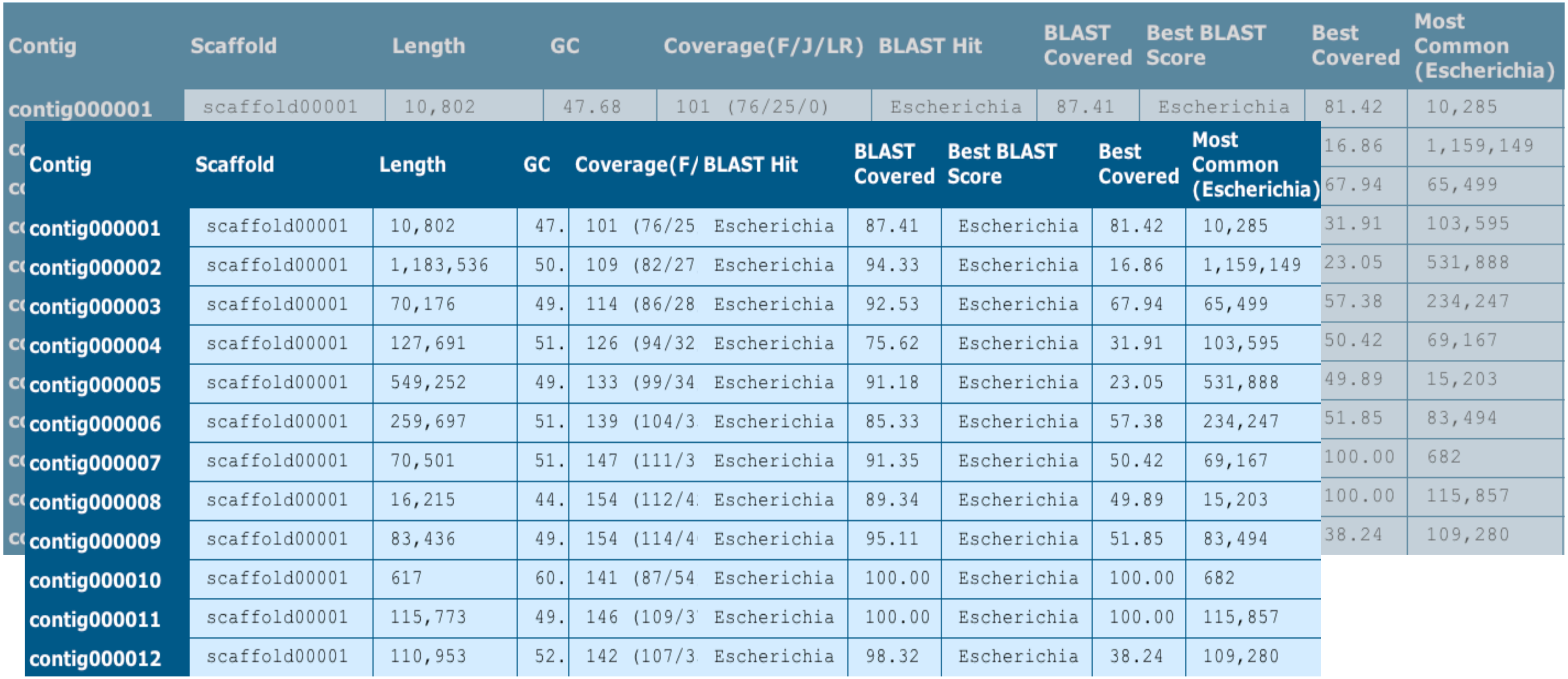
Gene	Total Copies	Lineage	Number Organisms Found	Organism IDs
16s	5	genus	1	Escherichia/Shigella

http://www.cbs.dtu.dk/services/RNAmmer
http://rdp.cme.msu.edu

Putting The Pieces Together

- Key concepts do not exist in a vacuum
- Analysis blurs these main concepts
- Metrics define course of action
- Not a standard process

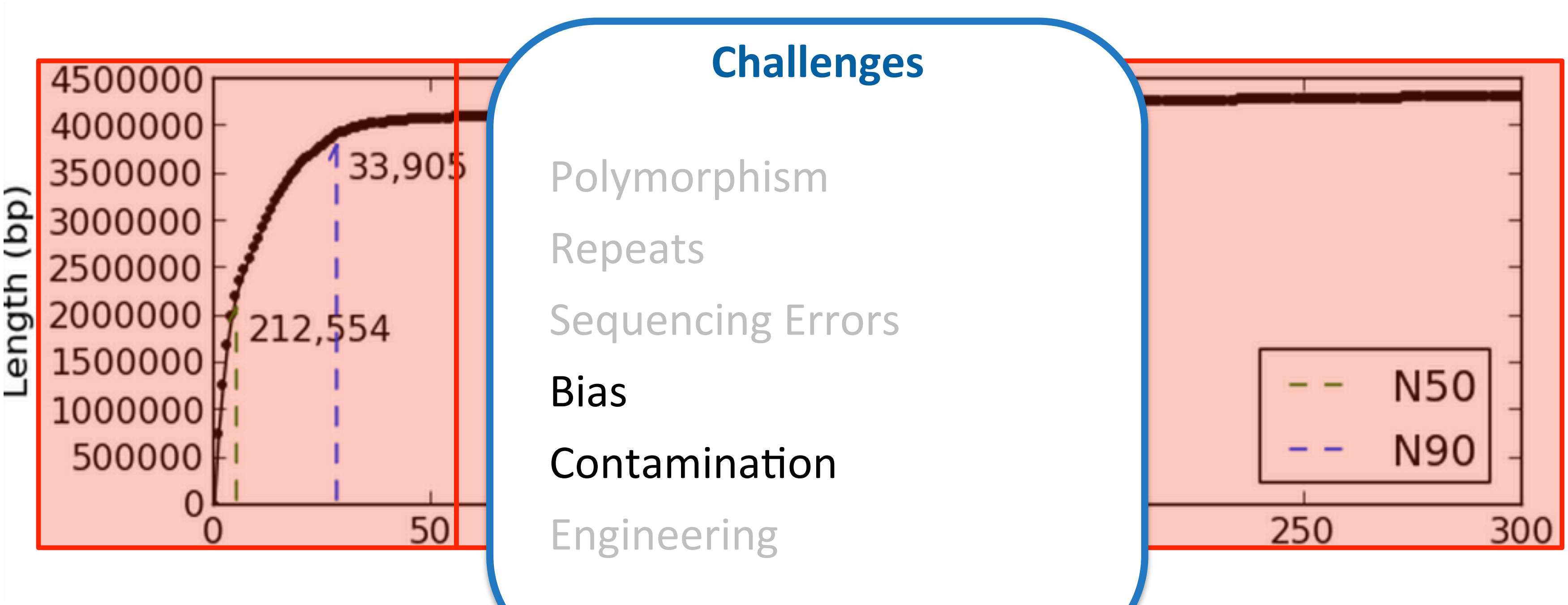
Contig Details



Contig	Scaffold	Length	GC	Coverage(F/J/LR)	BLAST Hit	BLAST Covered	Best BLAST Score	Best Covered	Most Common (Escherichia)
contig000001	scaffold00001	10,802	47.68	101 (76/25/0)	Escherichia	87.41	Escherichia	81.42	10,285
co								16.86	1,159,149
co								67.94	65,499
co								31.91	103,595
co								23.05	531,888
co								57.38	234,247
co								50.42	69,167
co								49.89	15,203
co								51.85	83,494
co								100.00	682
co								100.00	115,857
co								38.24	109,280

Contig	Scaffold	Length	GC	Coverage(F/	BLAST Hit	BLAST Covered	Best BLAST Score	Best Covered	Most Common (Escherichia)
contig000001	scaffold00001	10,802	47.	101 (76/25	Escherichia	87.41	Escherichia	81.42	10,285
contig000002	scaffold00001	1,183,536	50.	109 (82/27	Escherichia	94.33	Escherichia	16.86	1,159,149
contig000003	scaffold00001	70,176	49.	114 (86/28	Escherichia	92.53	Escherichia	67.94	65,499
contig000004	scaffold00001	127,691	51.	126 (94/32	Escherichia	75.62	Escherichia	31.91	103,595
contig000005	scaffold00001	549,252	49.	133 (99/34	Escherichia	91.18	Escherichia	23.05	531,888
contig000006	scaffold00001	259,697	51.	139 (104/3	Escherichia	85.33	Escherichia	57.38	234,247
contig000007	scaffold00001	70,501	51.	147 (111/3	Escherichia	91.35	Escherichia	50.42	69,167
contig000008	scaffold00001	16,215	44.	154 (112/4	Escherichia	89.34	Escherichia	49.89	15,203
contig000009	scaffold00001	83,436	49.	154 (114/4	Escherichia	95.11	Escherichia	51.85	83,494
contig000010	scaffold00001	617	60.	141 (87/54	Escherichia	100.00	Escherichia	100.00	682
contig000011	scaffold00001	115,773	49.	146 (109/3	Escherichia	100.00	Escherichia	100.00	115,857
contig000012	scaffold00001	110,953	52.	142 (107/3	Escherichia	98.32	Escherichia	38.24	109,280

Contiguity Potential Problems



Challenges

Polymorphism

Repeats

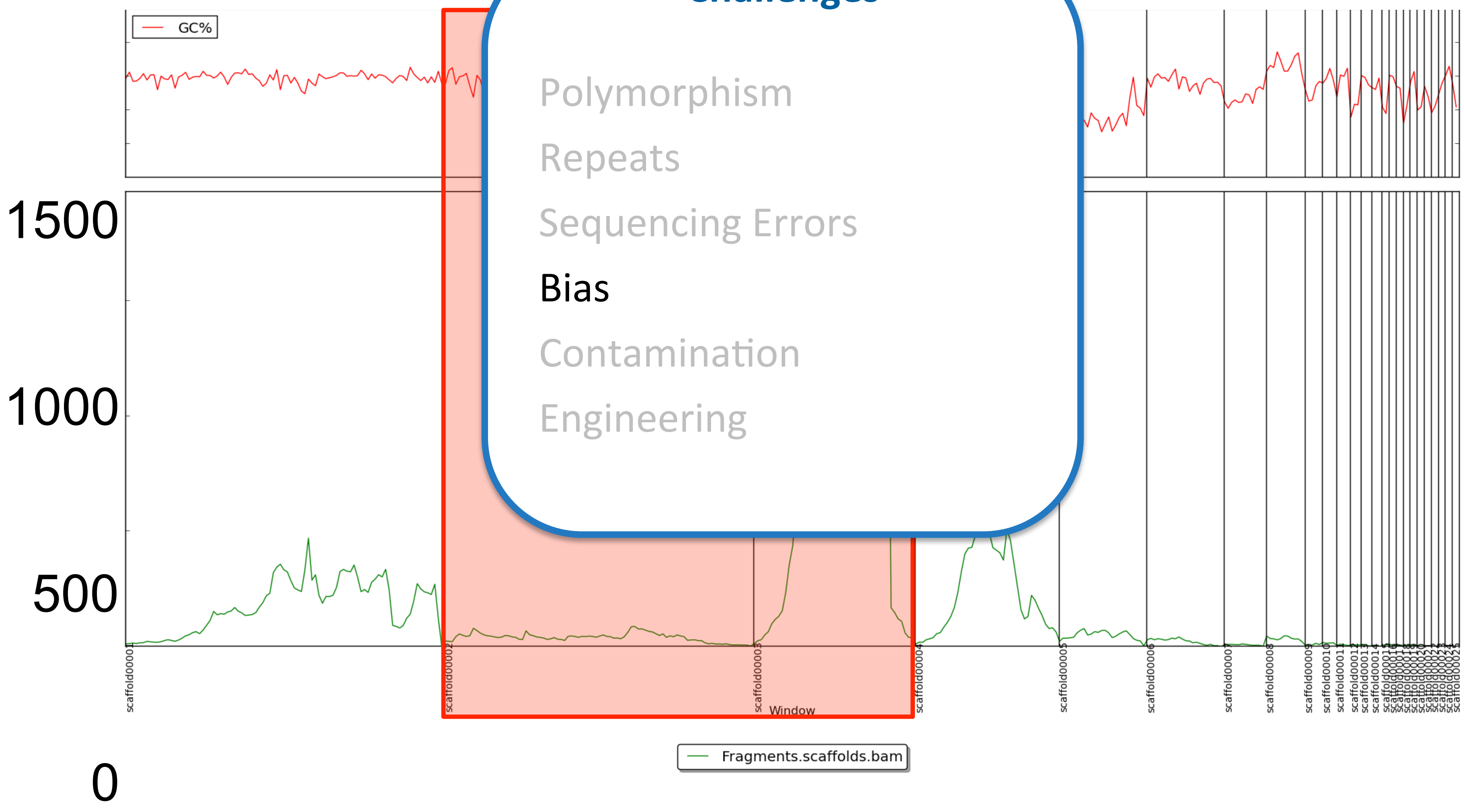
Sequencing Errors

Bias

Contamination

Engineering

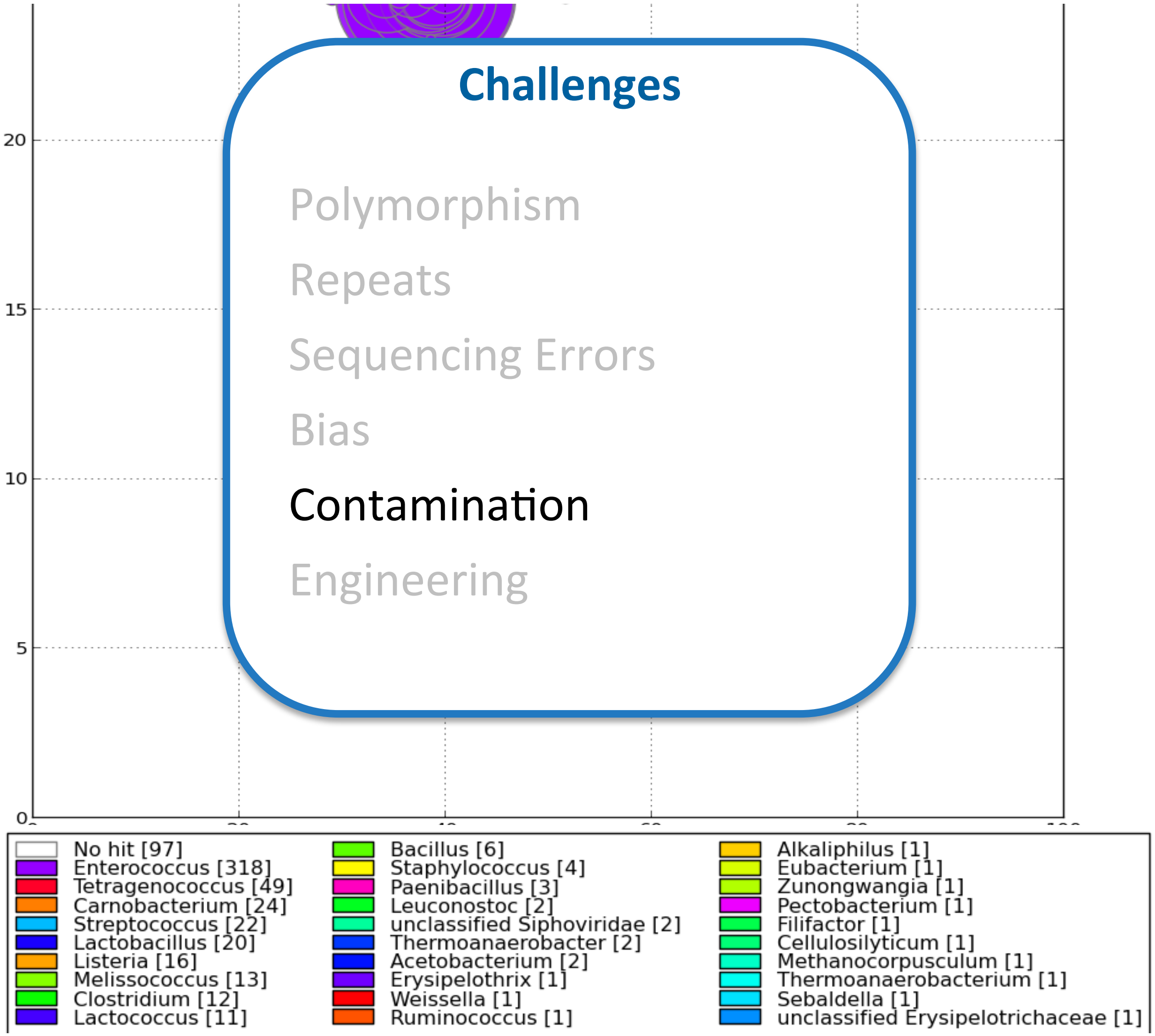
Read Coverage Problems



Challenges

- Polymorphism
- Repeats
- Sequencing Errors
- Bias
- Contamination
- Engineering

BLAST Bubble Problems



Challenges

Polymorphism

Repeats

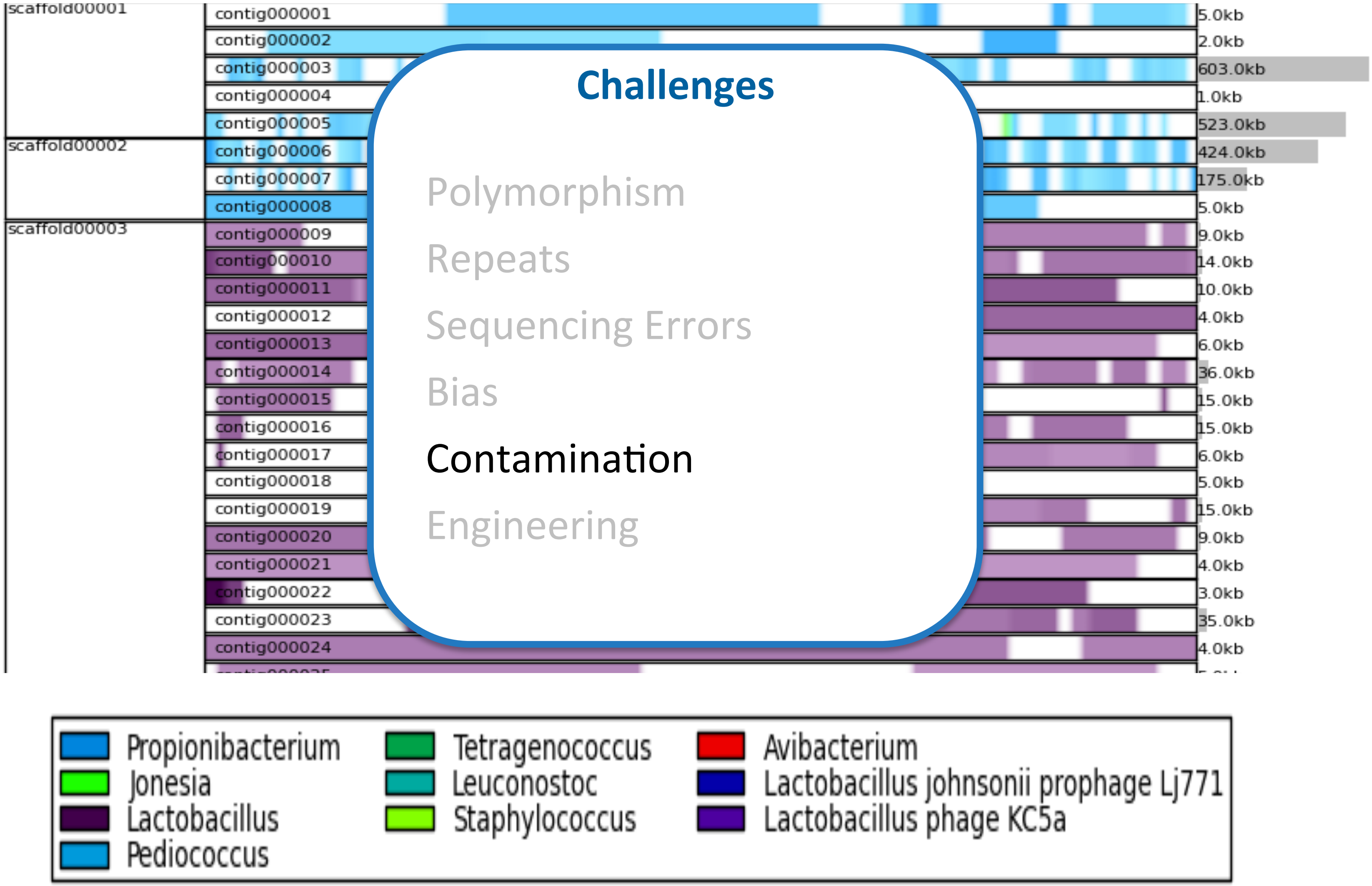
Sequencing Errors

Bias

Contamination

Engineering

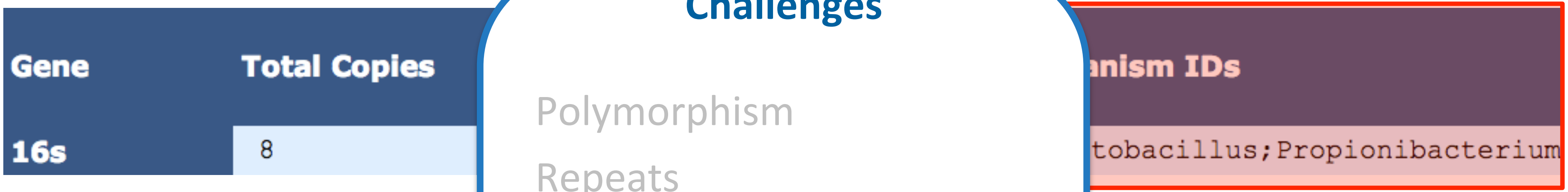
BLAST Heatmap Problems



Challenges

- Polymorphism
- Repeats
- Sequencing Errors
- Bias
- Contamination
- Engineering

16s Analysis Problems

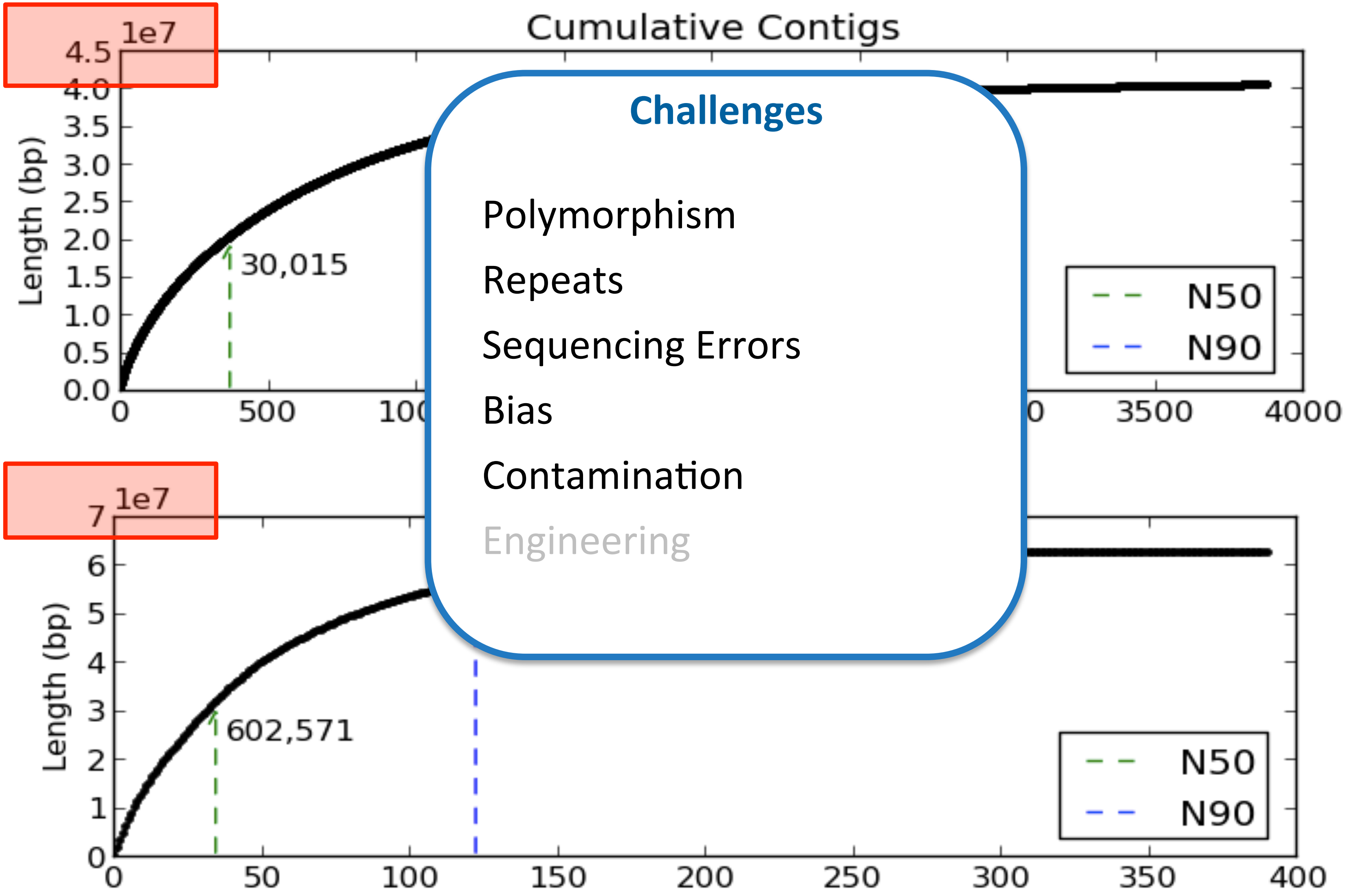


Gene	Total Copies	anism IDs
16s	8	tobacillus;Propionibacterium

Challenges

- Polymorphism
- Repeats
- Sequencing Errors
- Bias
- Contamination
- Engineering

Contiguity Potential Problems



Challenges

Polymorphism

Repeats

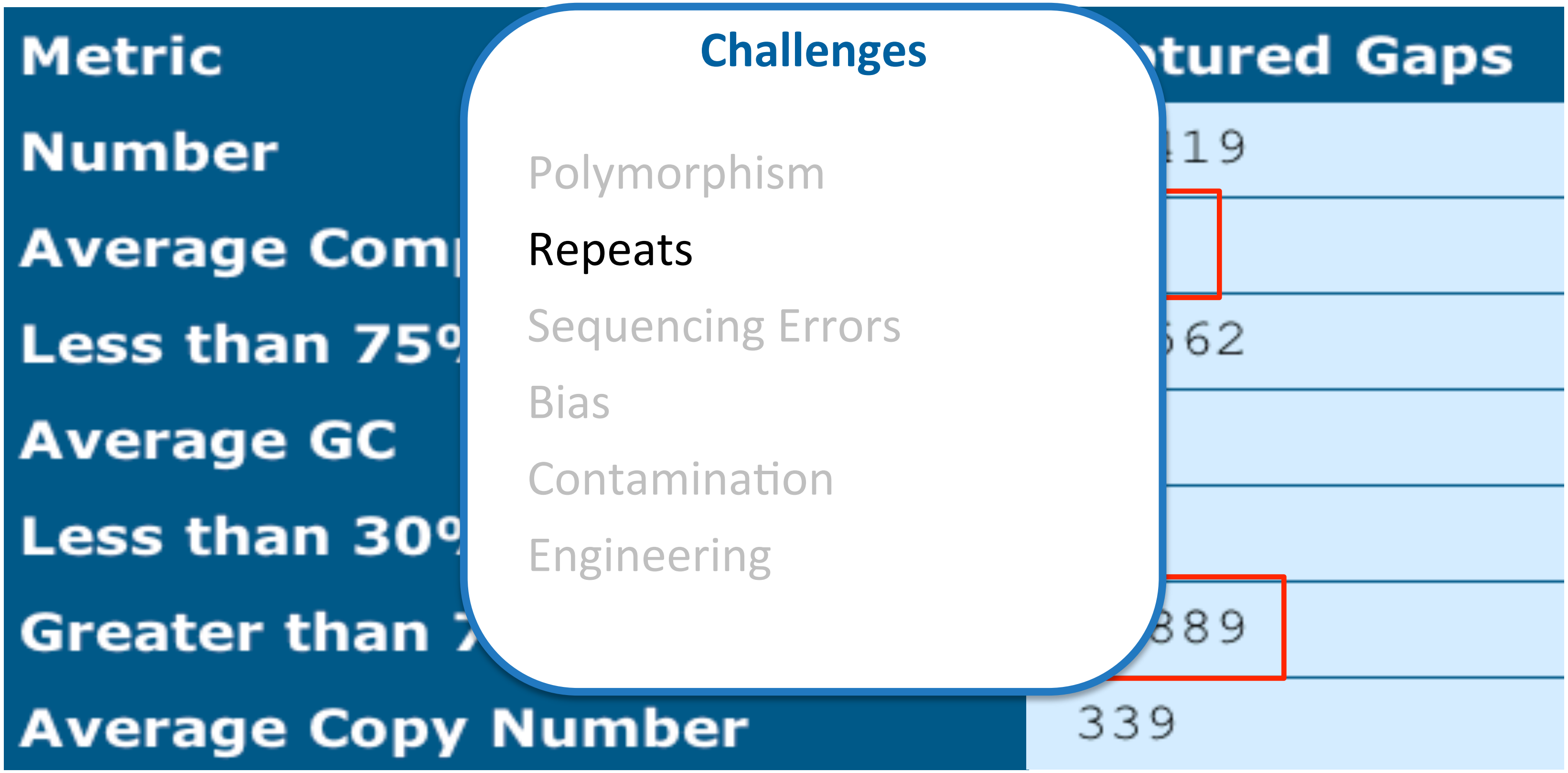
Sequencing Errors

Bias

Contamination

Engineering

Gap End Potential Problems

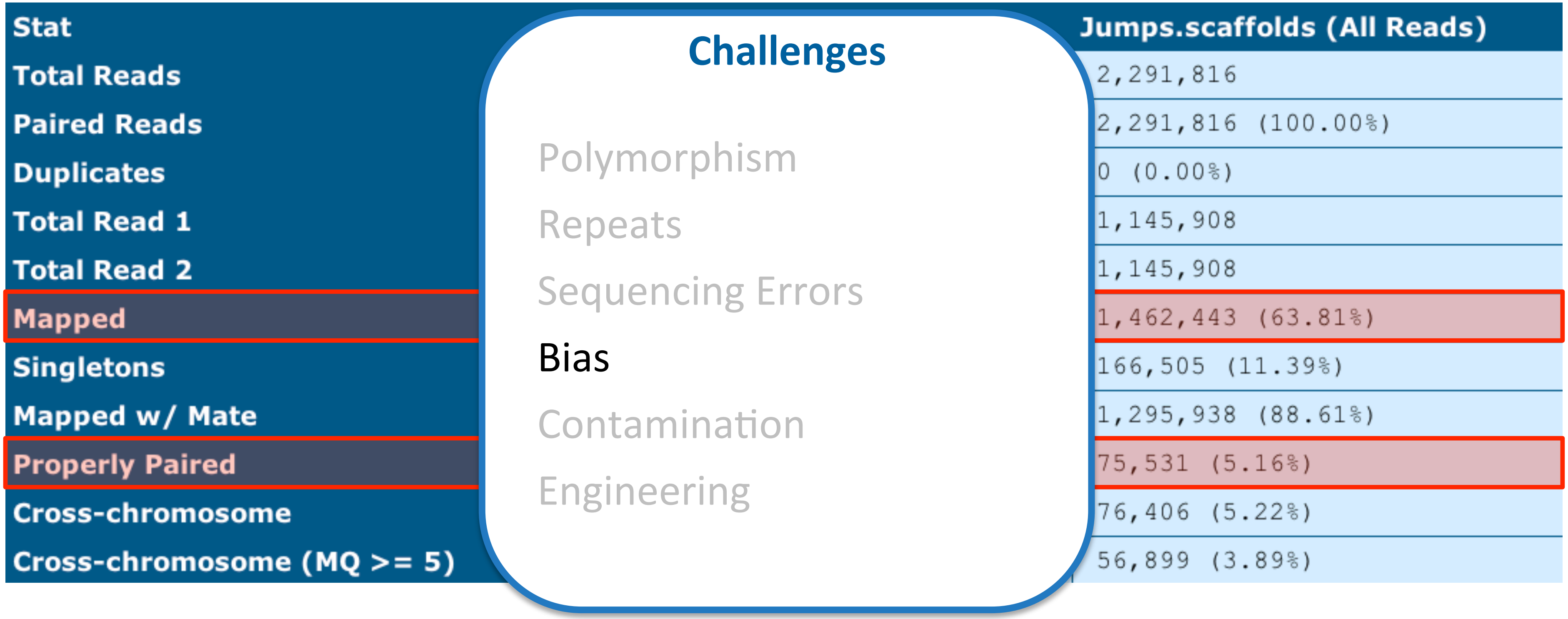


Metric	tured Gaps
Number	119
Average Com	
Less than 75	62
Average GC	
Less than 30	
Greater than 7	889
Average Copy Number	339

Challenges

- Polymorphism
- Repeats
- Sequencing Errors
- Bias
- Contamination
- Engineering

Read Mapping Potential Problems



Stat	Jumps.scaffolds (All Reads)
Total Reads	2,291,816
Paired Reads	2,291,816 (100.00%)
Duplicates	0 (0.00%)
Total Read 1	1,145,908
Total Read 2	1,145,908
Mapped	1,462,443 (63.81%)
Singletons	166,505 (11.39%)
Mapped w/ Mate	1,295,938 (88.61%)
Properly Paired	75,531 (5.16%)
Cross-chromosome	76,406 (5.22%)
Cross-chromosome (MQ >= 5)	56,899 (3.89%)

Challenges

Polymorphism

Repeats

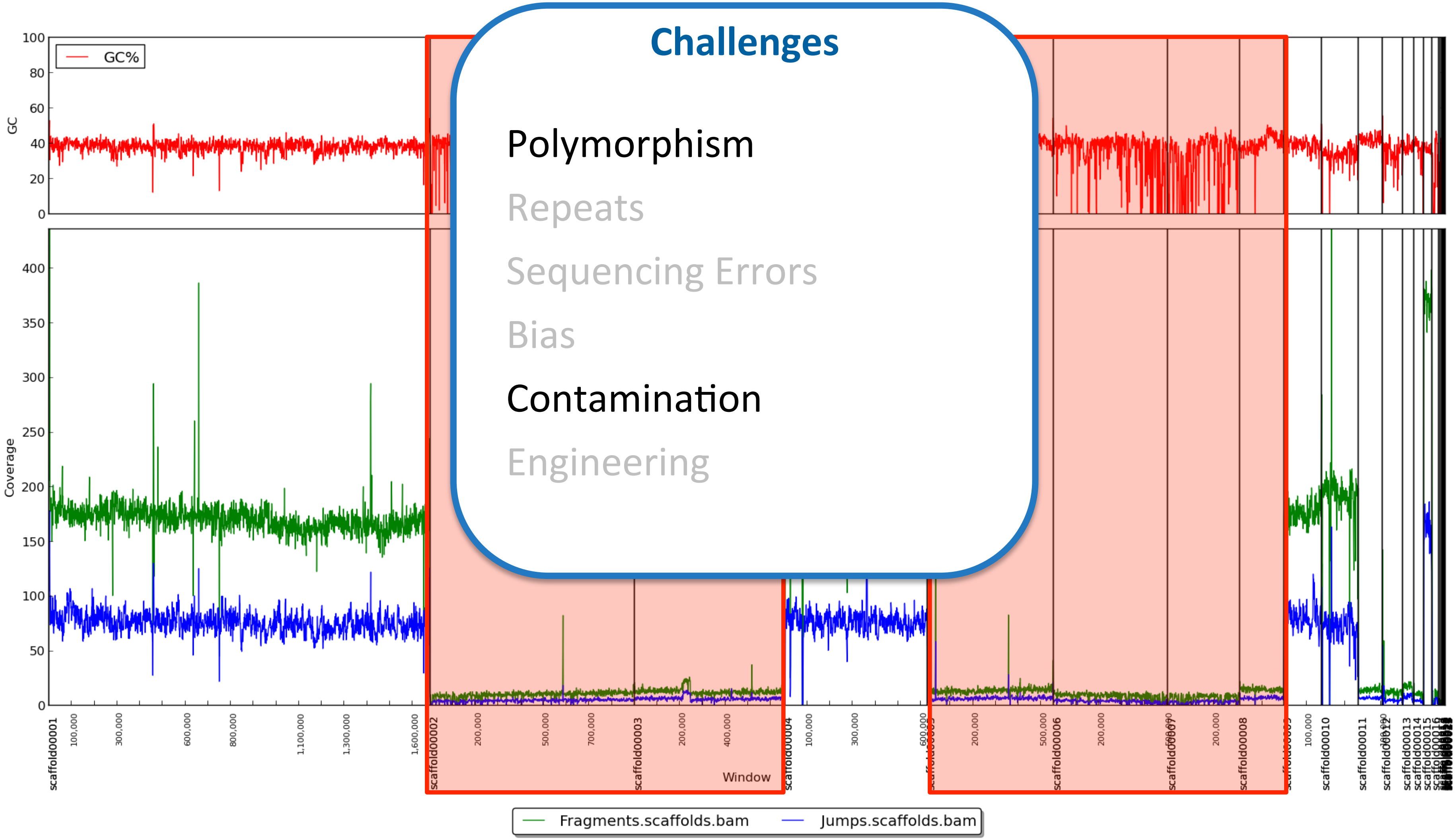
Sequencing Errors

Bias

Contamination

Engineering

Read Coverage Problems



Challenges

Polymorphism

Repeats

Sequencing Errors

Bias

Contamination

Engineering

Questions?

The Three C's

- Contiguity

  "Long contigs and scaffolds"

- Completeness

  "Minimal missing sequence"

- Correctness

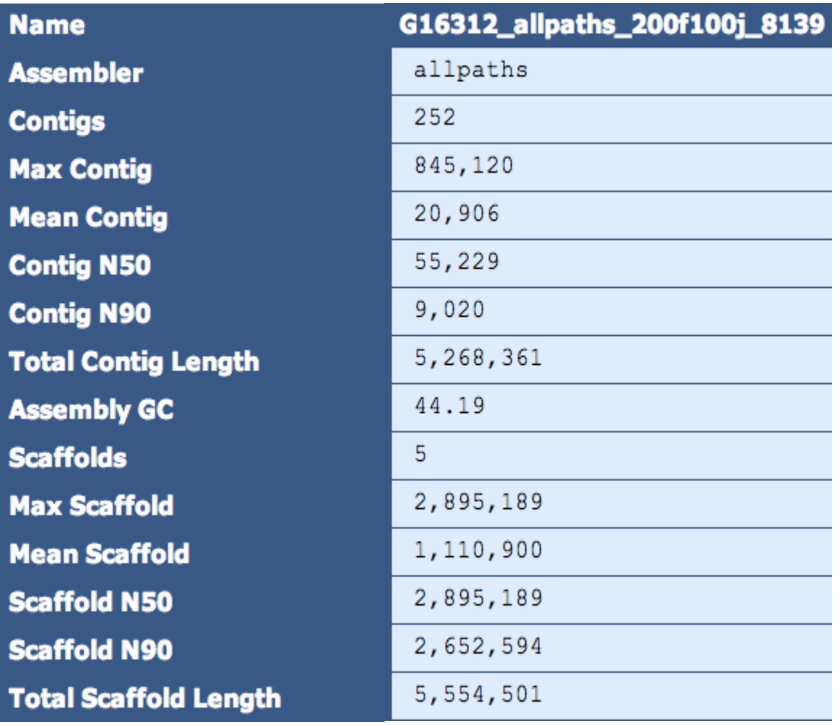
  "Few assembly errors"

Assembly Analysis Exercise #1

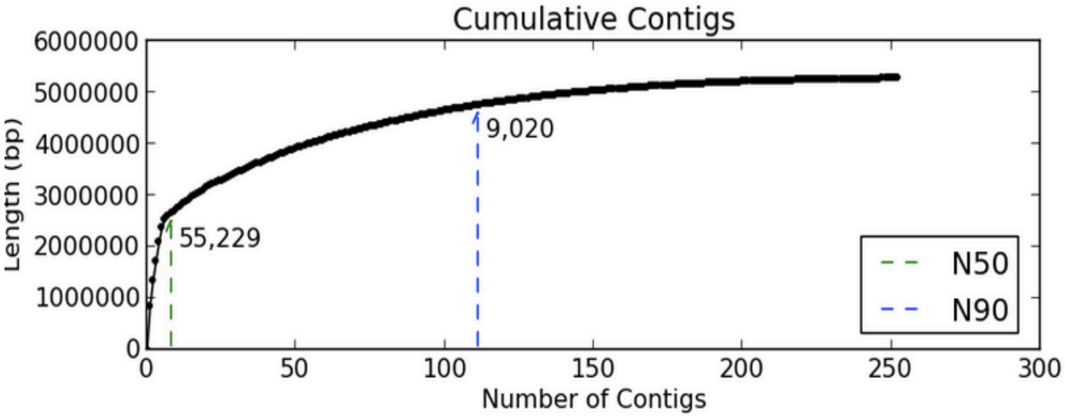
Background: You have created an assembly for a bacterial organism, *Treponema*, that has a genome estimated to be 2.5 Mb in size, with no known reference. From prior experience, you use 200x fragment read coverage and 100x jump read coverage, and you anticipate that assemblies of this organism will be in the 10 scaffold with 100 contigs range and total size very close to the estimate.

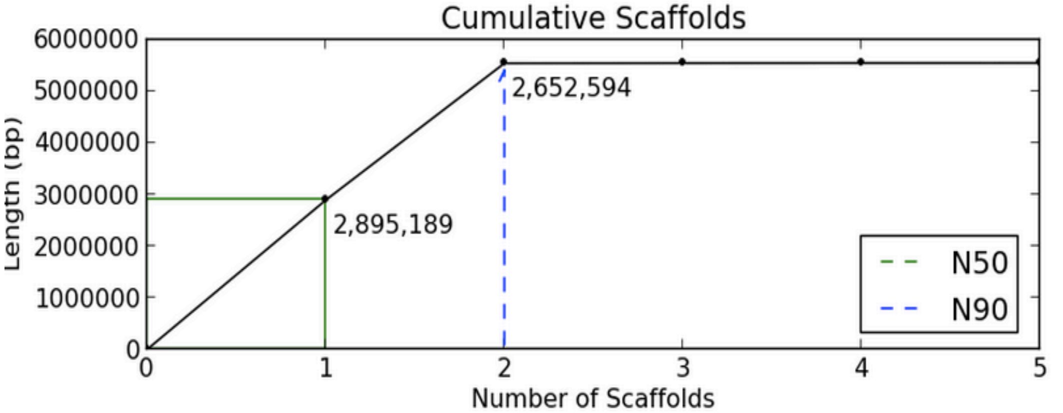
Assembly Analysis Exercise #1

A.) Contiguity stats can quickly highlight issues which may be present. What stands out when looking at the table and/or chart below?



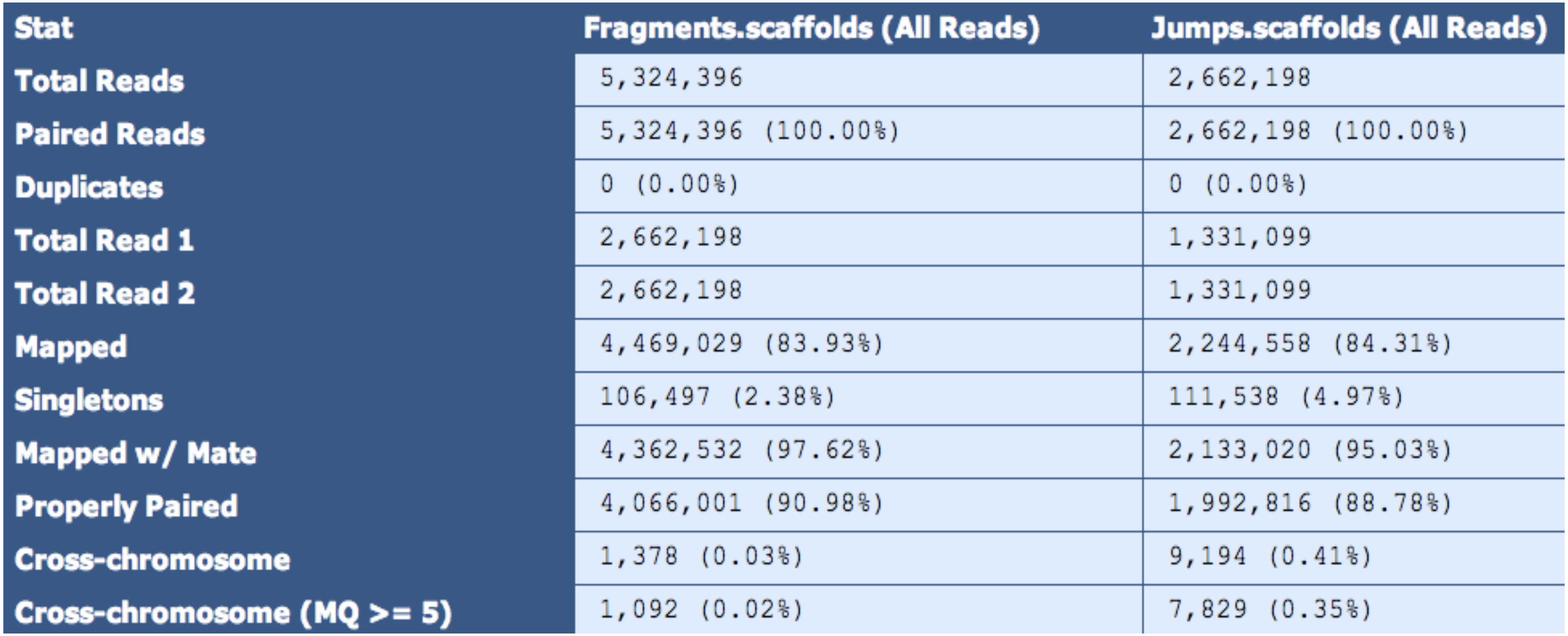
Name	G16312_allpaths_200f100j_8139
Assembler	allpaths
Contigs	252
Max Contig	845,120
Mean Contig	20,906
Contig N50	55,229
Contig N90	9,020
Total Contig Length	5,268,361
Assembly GC	44.19
Scaffolds	5
Max Scaffold	2,895,189
Mean Scaffold	1,110,900
Scaffold N50	2,895,189
Scaffold N90	2,652,594
Total Scaffold Length	5,554,501





Assembly Analysis Exercise #1

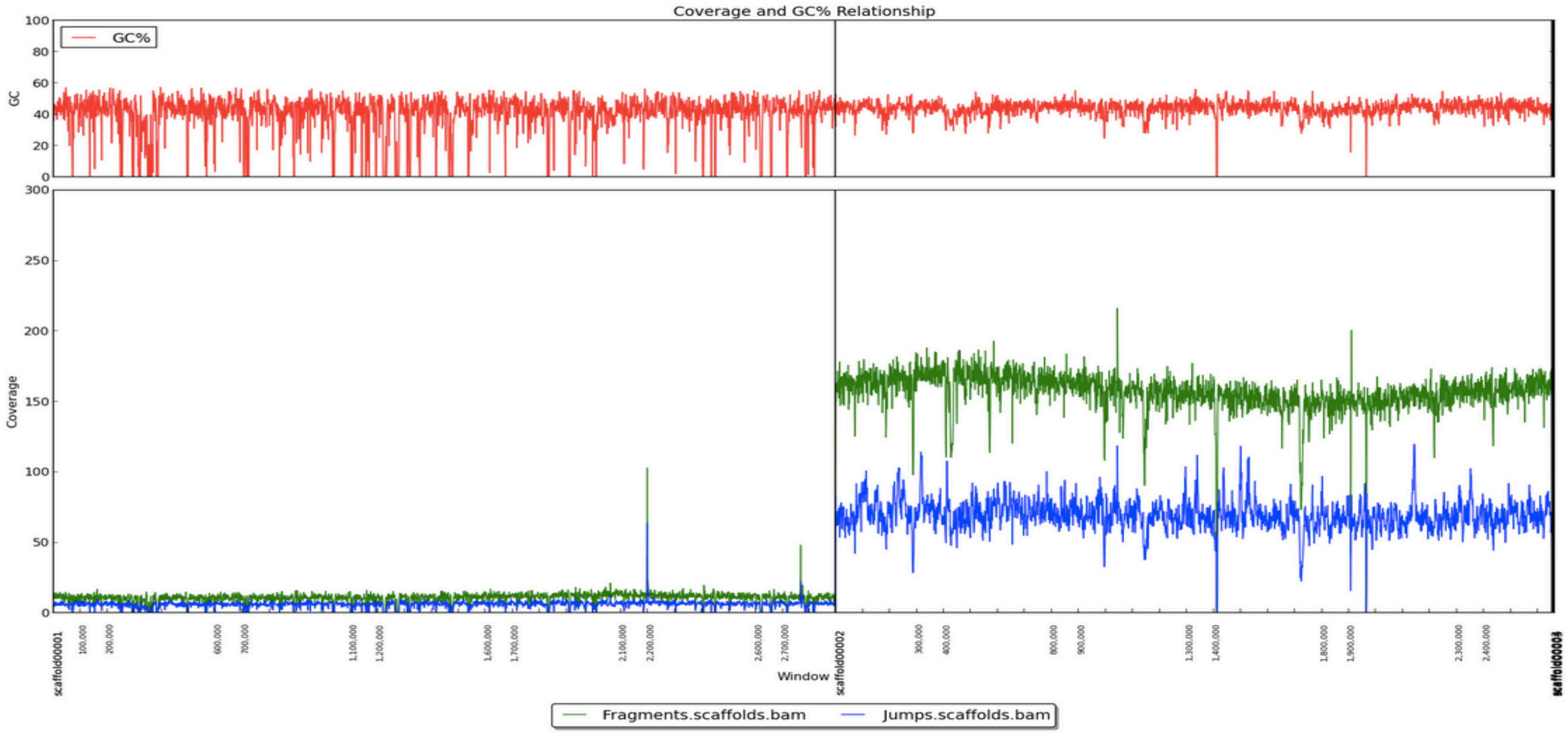
B.) Next, we want to look at how well our read data maps back to the assembly to look for any problems. What can we learn from the table below? Have we used our read data effectively?



Stat	Fragments.scaffolds (All Reads)	Jumps.scaffolds (All Reads)
Total Reads	5,324,396	2,662,198
Paired Reads	5,324,396 (100.00%)	2,662,198 (100.00%)
Duplicates	0 (0.00%)	0 (0.00%)
Total Read 1	2,662,198	1,331,099
Total Read 2	2,662,198	1,331,099
Mapped	4,469,029 (83.93%)	2,244,558 (84.31%)
Singletons	106,497 (2.38%)	111,538 (4.97%)
Mapped w/ Mate	4,362,532 (97.62%)	2,133,020 (95.03%)
Properly Paired	4,066,001 (90.98%)	1,992,816 (88.78%)
Cross-chromosome	1,378 (0.03%)	9,194 (0.41%)
Cross-chromosome (MQ >= 5)	1,092 (0.02%)	7,829 (0.35%)

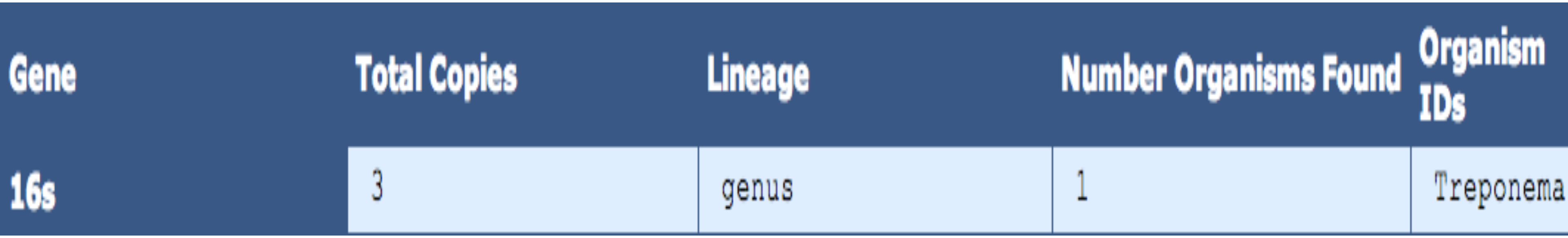
Assembly Analysis Exercise #1

C.) Now that we've seen how our read data was used in the assembly, we should investigate the read coverage along our assembly. What information can you quickly learn from the chart below?



Assembly Analysis Exercise #1

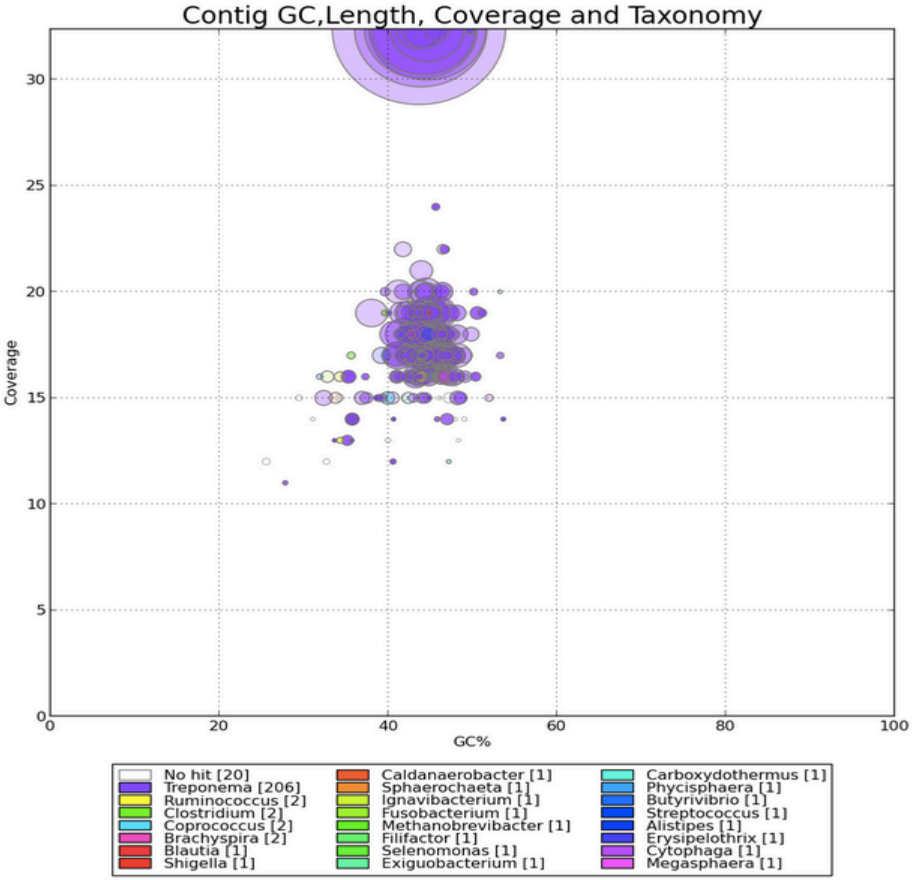
D.) Since this organism is bacterial, we can look to our 16s analysis to see if there any inconsistencies in our assembly. Are there any indications here about possible assembly problems?



Gene	Total Copies	Lineage	Number Organisms Found	Organism IDs
16s	3	genus	1	Treponema

Assembly Analysis Exercise #1

E.) BLAST taxonomy information can help determine contamination. Does the plot below indicate the presence or absence of contamination?



Assembly Analysis Exercise #1

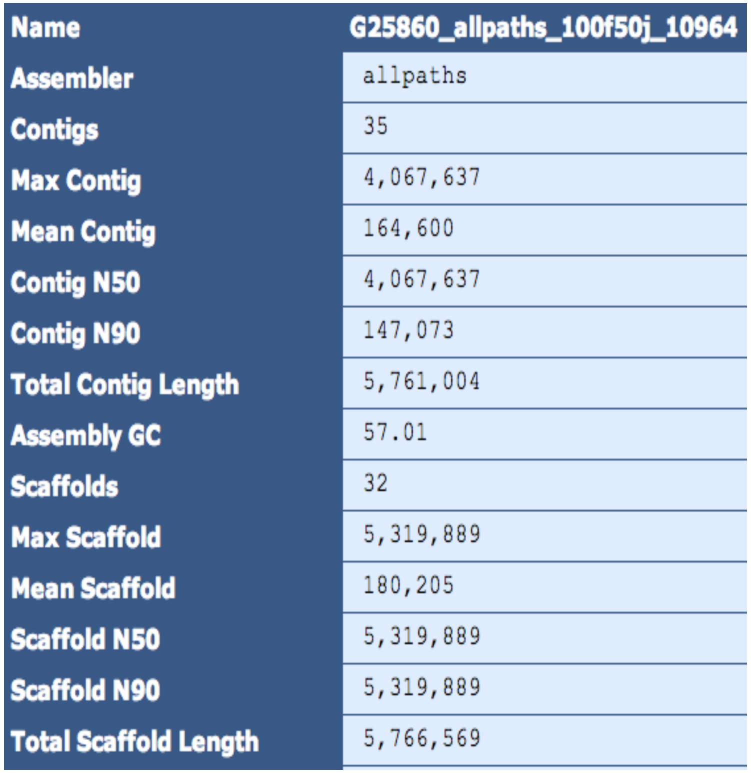
F.) Is there a problem with this assembly? If so, what do you think is the issue? If you are unsure, what other questions could you ask about the data?

Assembly Analysis Exercise #2

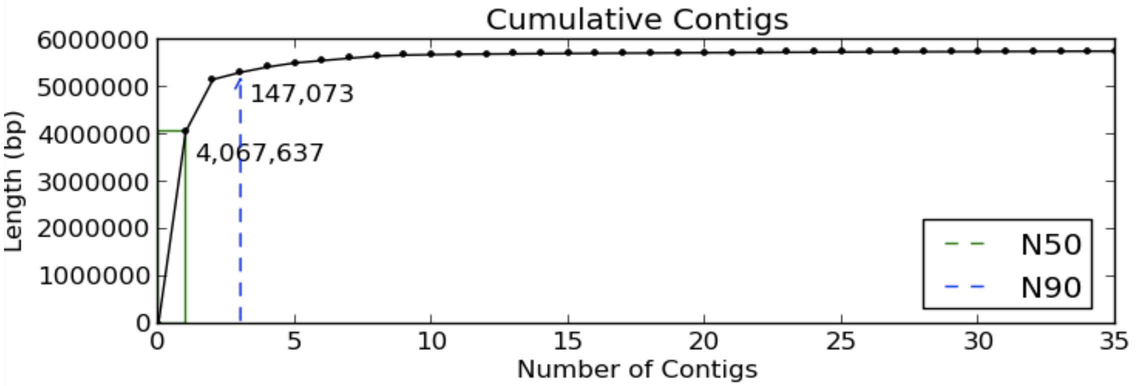
Background: You have created an assembly for a bacterial organism, *Klebsiella*, that has a genome size estimated to be in the range of 5.5 - 6.5 Mb. Previous *Klebsiella* assemblies have assembled together in the range 3-10 scaffolds and 23-71 contigs. The researcher states that sometimes *Klebsiella* strains have non-chromosomal (plasmid) sequences.

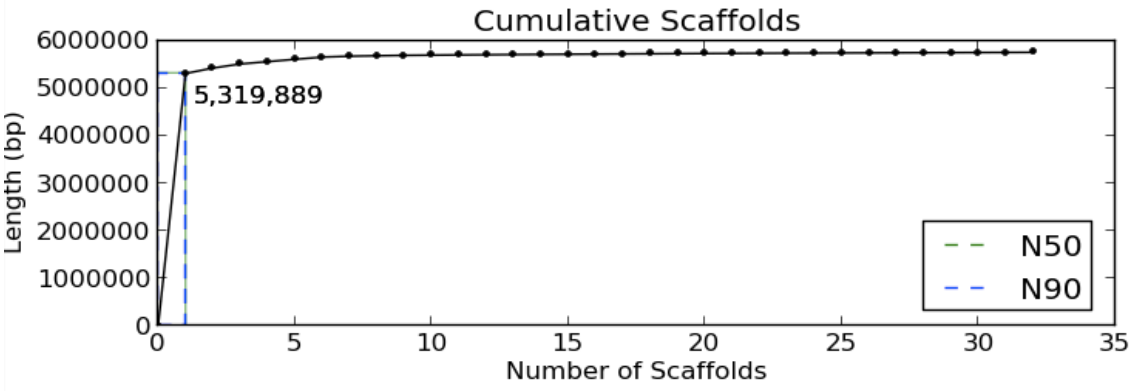
Assembly Analysis Exercise #2

A.) Contiguity stats can quickly highlight issues which may be present. What stands out when looking at the table and/or charts below?



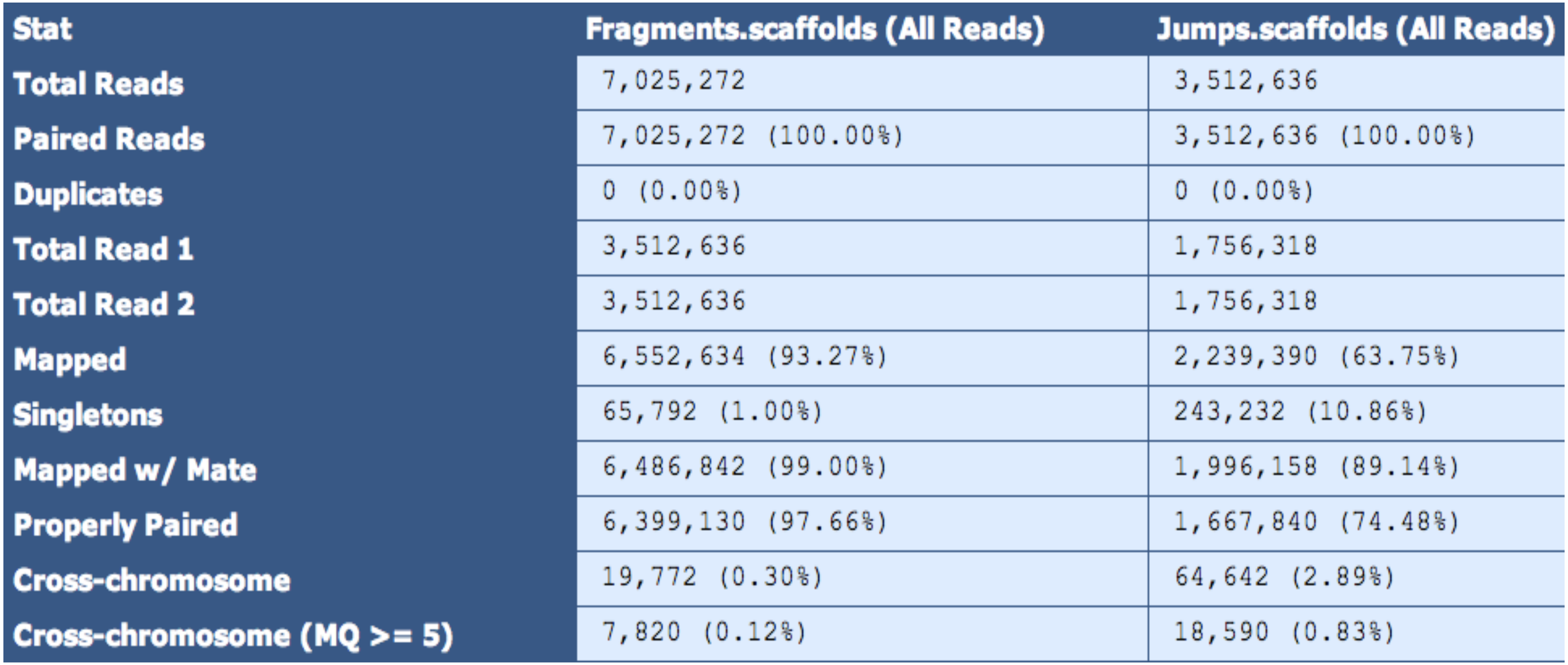
Name	G25860_allpaths_100f50j_10964
Assembler	allpaths
Contigs	35
Max Contig	4,067,637
Mean Contig	164,600
Contig N50	4,067,637
Contig N90	147,073
Total Contig Length	5,761,004
Assembly GC	57.01
Scaffolds	32
Max Scaffold	5,319,889
Mean Scaffold	180,205
Scaffold N50	5,319,889
Scaffold N90	5,319,889
Total Scaffold Length	5,766,569





Assembly Analysis Exercise #2

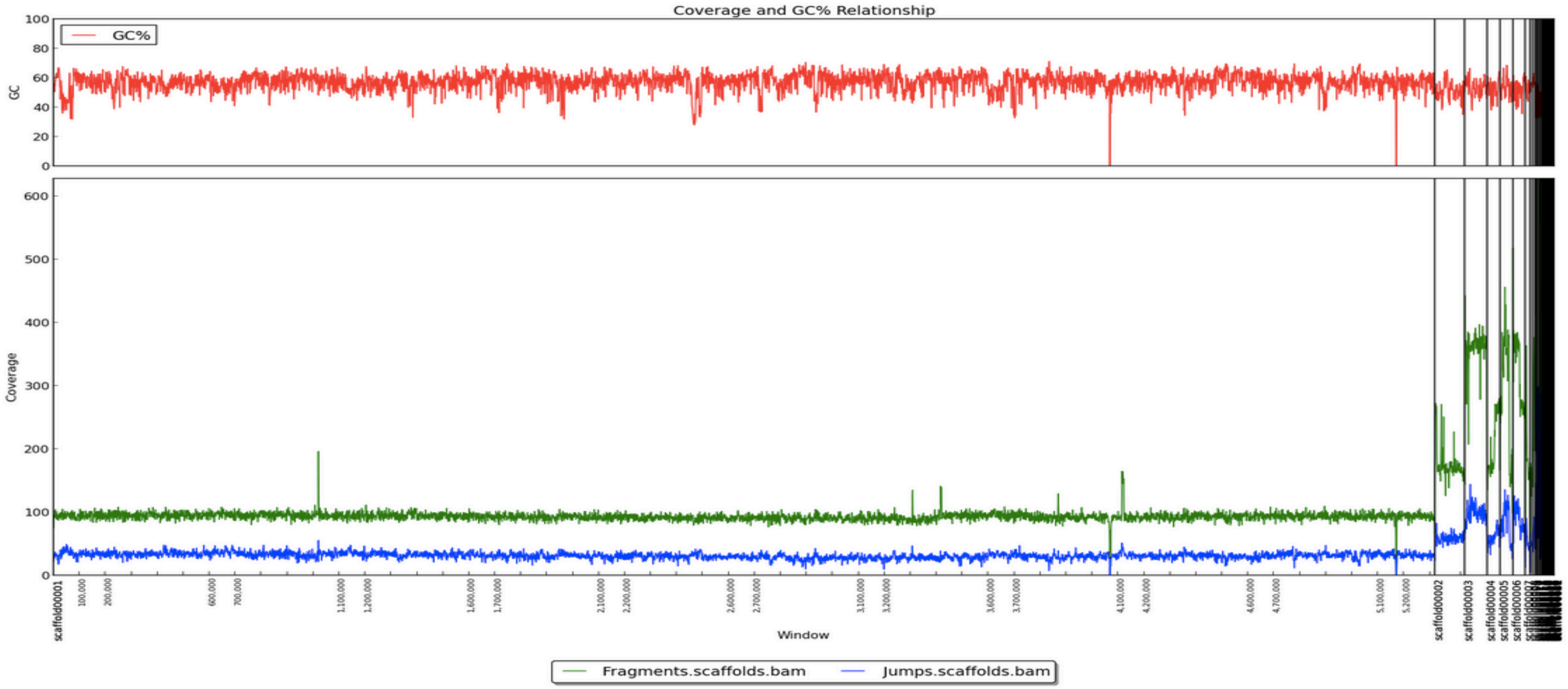
B.) Next, we want to look at how well our read data maps back to the assembly to look for any problems. What can we learn from the table below? Have we used our read data effectively?



Stat	Fragments.scaffolds (All Reads)	Jumps.scaffolds (All Reads)
Total Reads	7,025,272	3,512,636
Paired Reads	7,025,272 (100.00%)	3,512,636 (100.00%)
Duplicates	0 (0.00%)	0 (0.00%)
Total Read 1	3,512,636	1,756,318
Total Read 2	3,512,636	1,756,318
Mapped	6,552,634 (93.27%)	2,239,390 (63.75%)
Singletons	65,792 (1.00%)	243,232 (10.86%)
Mapped w/ Mate	6,486,842 (99.00%)	1,996,158 (89.14%)
Properly Paired	6,399,130 (97.66%)	1,667,840 (74.48%)
Cross-chromosome	19,772 (0.30%)	64,642 (2.89%)
Cross-chromosome (MQ >= 5)	7,820 (0.12%)	18,590 (0.83%)

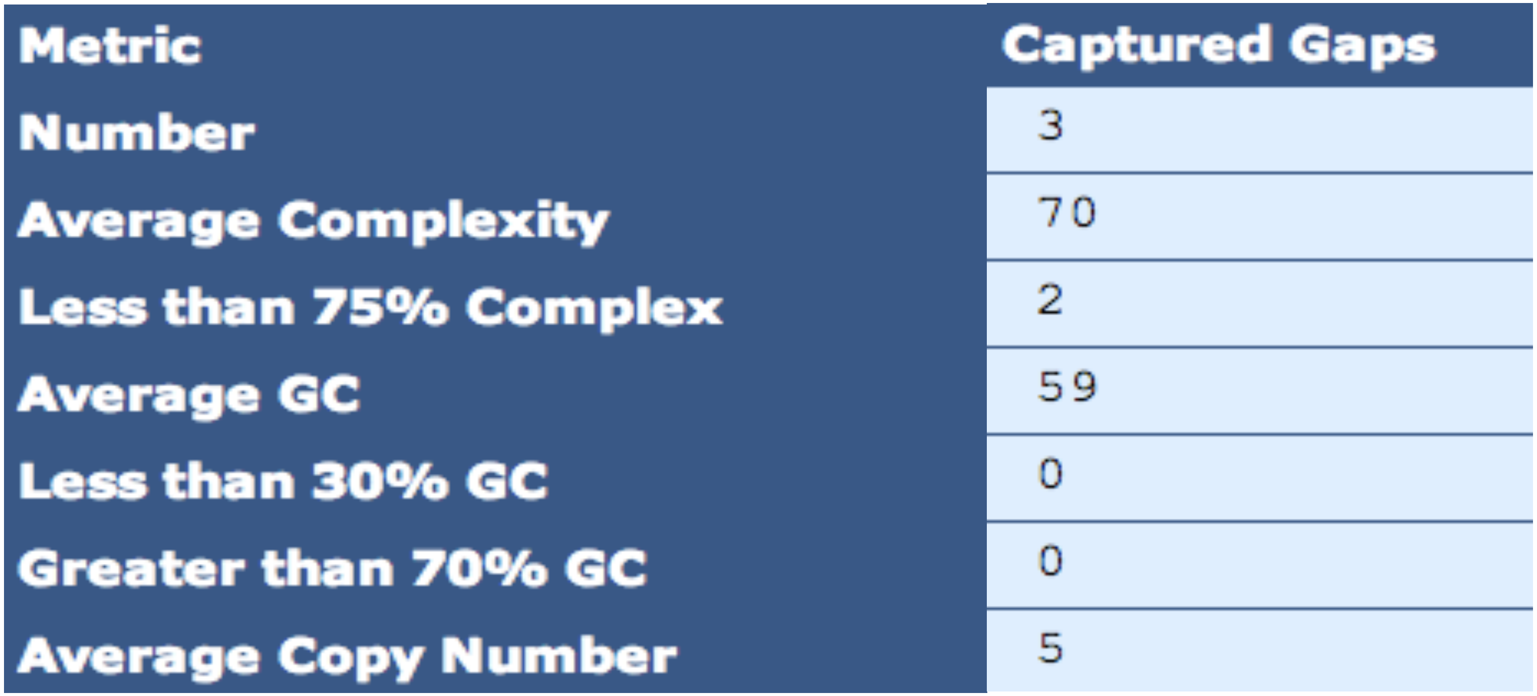
Assembly Analysis Exercise #2

C.) Now that we've seen how our read data was used in the assembly, we should investigate the GC content and read coverage along our assembly. What information can you quickly learn from the chart below?



Assembly Analysis Exercise #2

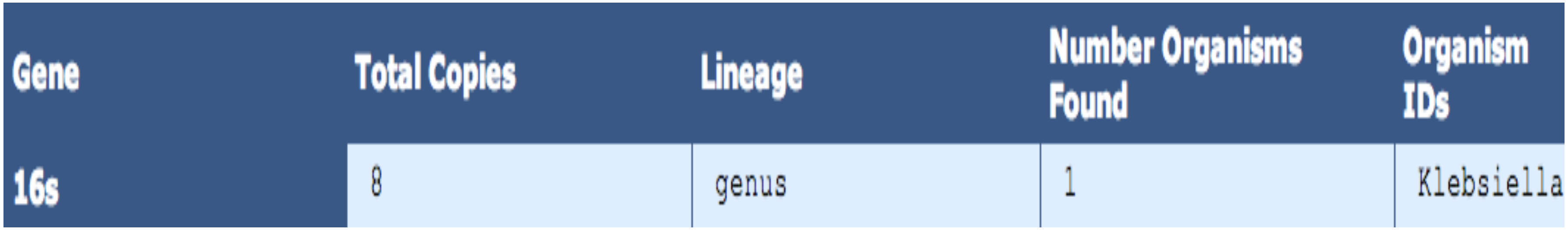
D.) A look at the contig sequence leading into gaps can provide insight into dis-contiguity. Does the sequence at the ends of contigs help explain the fragmentation of the assembly?



Metric	Captured Gaps
Number	3
Average Complexity	70
Less than 75% Complex	2
Average GC	59
Less than 30% GC	0
Greater than 70% GC	0
Average Copy Number	5

Assembly Analysis Exercise #2

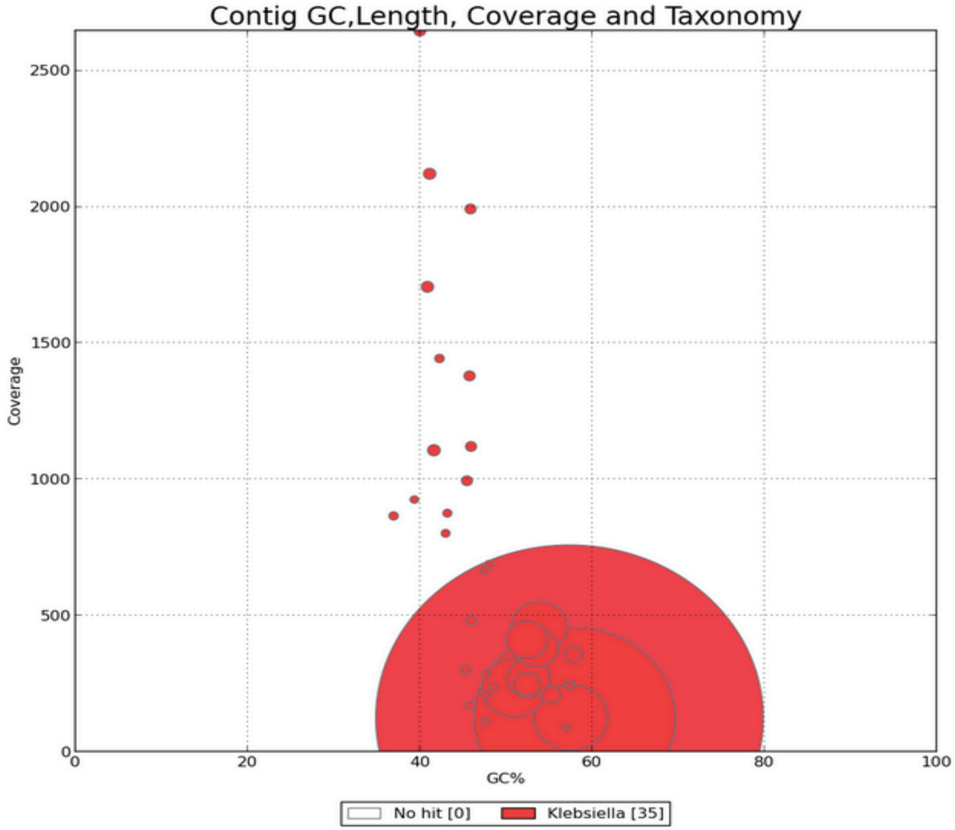
E.) Since this organism is bacterial, we can look to our 16s analysis to see if there any inconsistencies in our assembly. Are there any indications here about possible assembly problems?



Gene	Total Copies	Lineage	Number Organisms Found	Organism IDs
16s	8	genus	1	Klebsiella

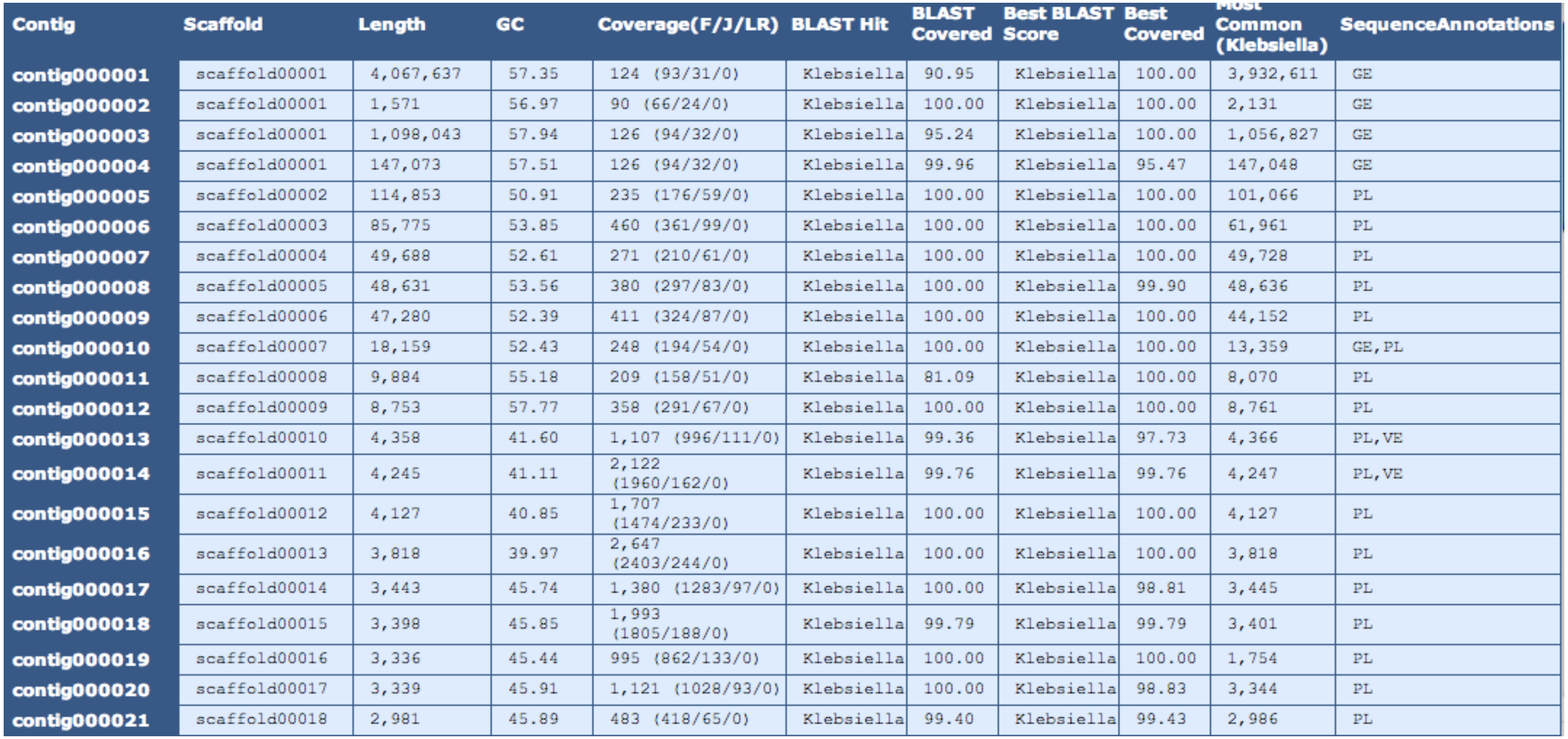
Assembly Analysis Exercise #2

F.) BLAST taxonomy information can help determine contamination. Does the plot below indicate the presence or absence of contamination?



Assembly Analysis Exercise #2

G.) Further NCBI blast information is available in supplemental tables. Hits are characterized in the "SequenceAnnotations" column (GE=genomic; VE=vector; PL=plasmid). Is there additional blast and taxonomic information to provide insight into the nature of the assembly?



Contig	Scaffold	Length	GC	Coverage(F/J/LR)	BLAST Hit	BLAST Covered	Best BLAST Score	Best Covered	Most Common (Klebsiella)	SequenceAnnotations
contig000001	scaffold00001	4,067,637	57.35	124 (93/31/0)	Klebsiella	90.95	Klebsiella	100.00	3,932,611	GE
contig000002	scaffold00001	1,571	56.97	90 (66/24/0)	Klebsiella	100.00	Klebsiella	100.00	2,131	GE
contig000003	scaffold00001	1,098,043	57.94	126 (94/32/0)	Klebsiella	95.24	Klebsiella	100.00	1,056,827	GE
contig000004	scaffold00001	147,073	57.51	126 (94/32/0)	Klebsiella	99.96	Klebsiella	95.47	147,048	GE
contig000005	scaffold00002	114,853	50.91	235 (176/59/0)	Klebsiella	100.00	Klebsiella	100.00	101,066	PL
contig000006	scaffold00003	85,775	53.85	460 (361/99/0)	Klebsiella	100.00	Klebsiella	100.00	61,961	PL
contig000007	scaffold00004	49,688	52.61	271 (210/61/0)	Klebsiella	100.00	Klebsiella	100.00	49,728	PL
contig000008	scaffold00005	48,631	53.56	380 (297/83/0)	Klebsiella	100.00	Klebsiella	99.90	48,636	PL
contig000009	scaffold00006	47,280	52.39	411 (324/87/0)	Klebsiella	100.00	Klebsiella	100.00	44,152	PL
contig000010	scaffold00007	18,159	52.43	248 (194/54/0)	Klebsiella	100.00	Klebsiella	100.00	13,359	GE,PL
contig000011	scaffold00008	9,884	55.18	209 (158/51/0)	Klebsiella	81.09	Klebsiella	100.00	8,070	PL
contig000012	scaffold00009	8,753	57.77	358 (291/67/0)	Klebsiella	100.00	Klebsiella	100.00	8,761	PL
contig000013	scaffold00010	4,358	41.60	1,107 (996/111/0)	Klebsiella	99.36	Klebsiella	97.73	4,366	PL,VE
contig000014	scaffold00011	4,245	41.11	2,122 (1960/162/0)	Klebsiella	99.76	Klebsiella	99.76	4,247	PL,VE
contig000015	scaffold00012	4,127	40.85	1,707 (1474/233/0)	Klebsiella	100.00	Klebsiella	100.00	4,127	PL
contig000016	scaffold00013	3,818	39.97	2,647 (2403/244/0)	Klebsiella	100.00	Klebsiella	100.00	3,818	PL
contig000017	scaffold00014	3,443	45.74	1,380 (1283/97/0)	Klebsiella	100.00	Klebsiella	98.81	3,445	PL
contig000018	scaffold00015	3,398	45.85	1,993 (1805/188/0)	Klebsiella	99.79	Klebsiella	99.79	3,401	PL
contig000019	scaffold00016	3,336	45.44	995 (862/133/0)	Klebsiella	100.00	Klebsiella	100.00	1,754	PL
contig000020	scaffold00017	3,339	45.91	1,121 (1028/93/0)	Klebsiella	100.00	Klebsiella	98.83	3,344	PL
contig000021	scaffold00018	2,981	45.89	483 (418/65/0)	Klebsiella	99.40	Klebsiella	99.43	2,986	PL

Assembly Analysis Exercise #2

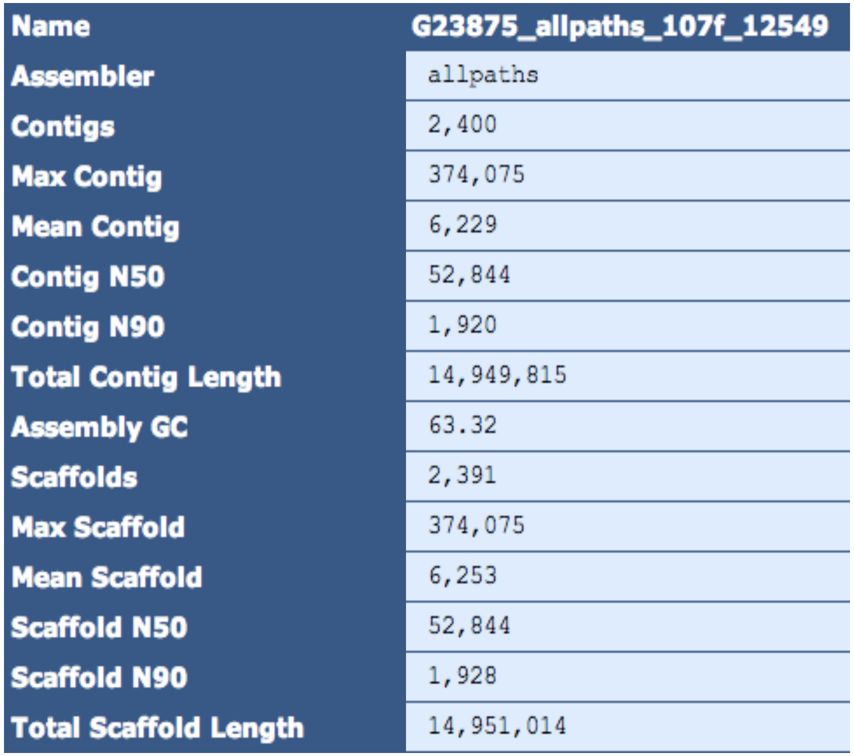
H.) Is there a problem with this assembly? If so, what do you think is the issue? If you are unsure, what other questions could you ask about the data?

Assembly Analysis Exercise #3

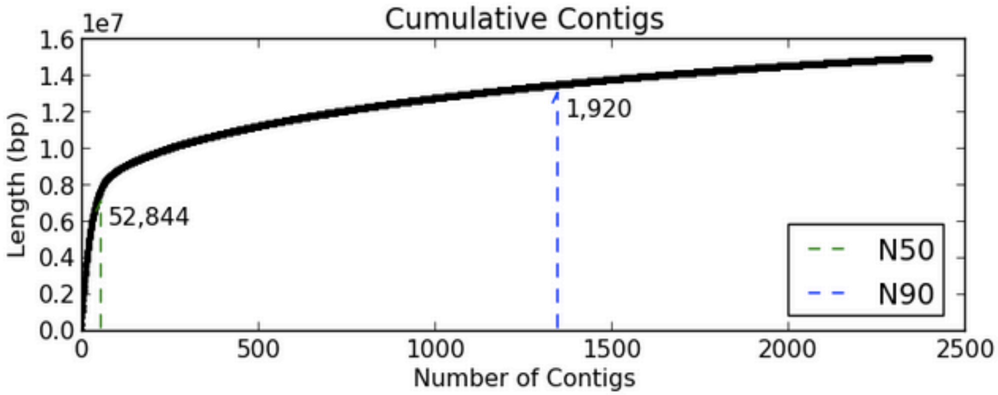
Background: Several attempts were made to assemble the genome of a sample presumed to be *Brucella ovis*, with estimated genome size of 3.2 Mb and expected GC of 56%. Only fragment read library was available and based on previous experience 100x coverage with similar genomes produced good assemblies with 20-40 scaffolds and N50 sizes of ~250 kb.

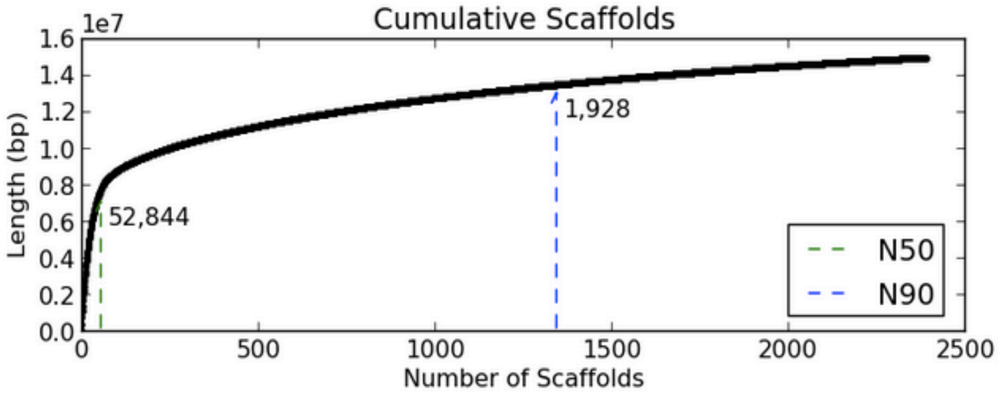
Assembly Analysis Exercise #3

A.) Contiguity stats can quickly highlight issues which may be present. What stands out when looking at the table and/or charts below?



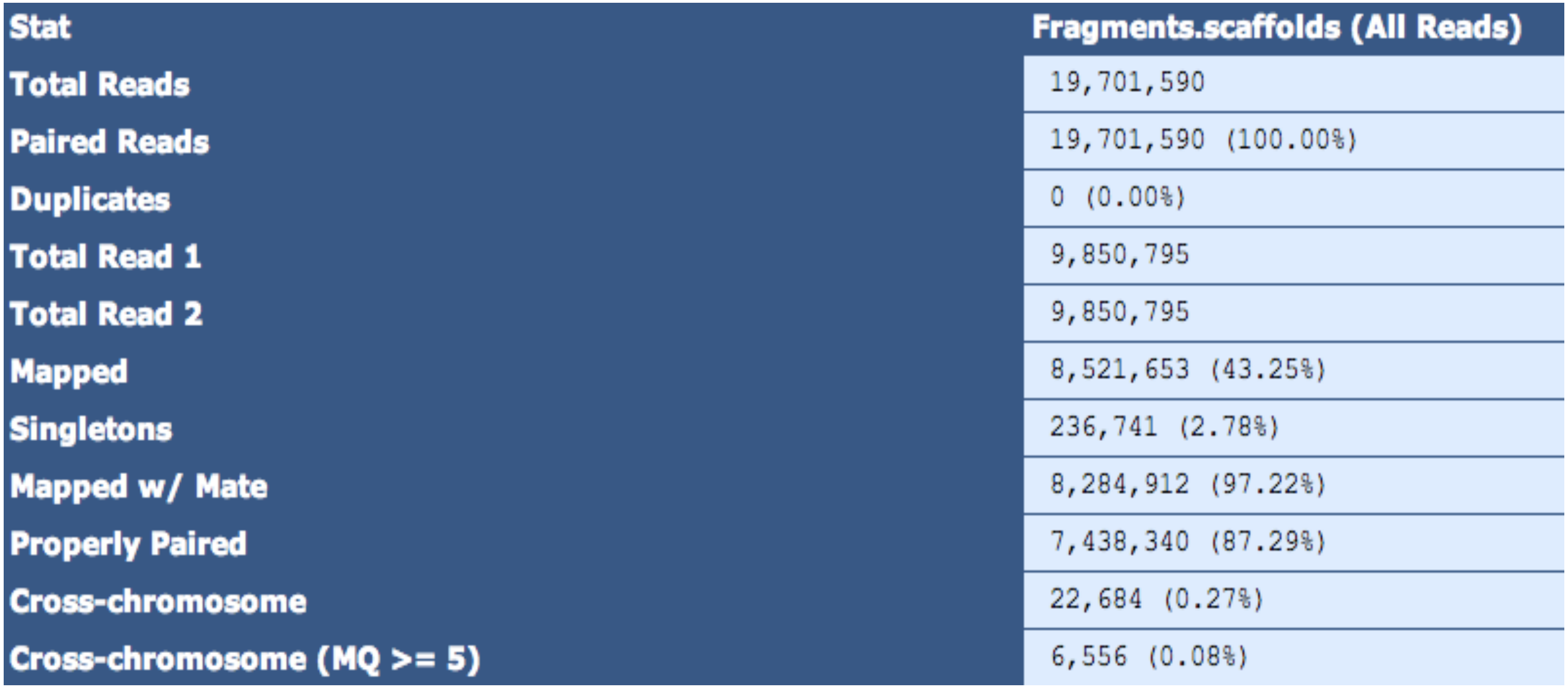
Name	G23875_allpaths_107f_12549
Assembler	allpaths
Contigs	2,400
Max Contig	374,075
Mean Contig	6,229
Contig N50	52,844
Contig N90	1,920
Total Contig Length	14,949,815
Assembly GC	63.32
Scaffolds	2,391
Max Scaffold	374,075
Mean Scaffold	6,253
Scaffold N50	52,844
Scaffold N90	1,928
Total Scaffold Length	14,951,014





Assembly Analysis Exercise #3

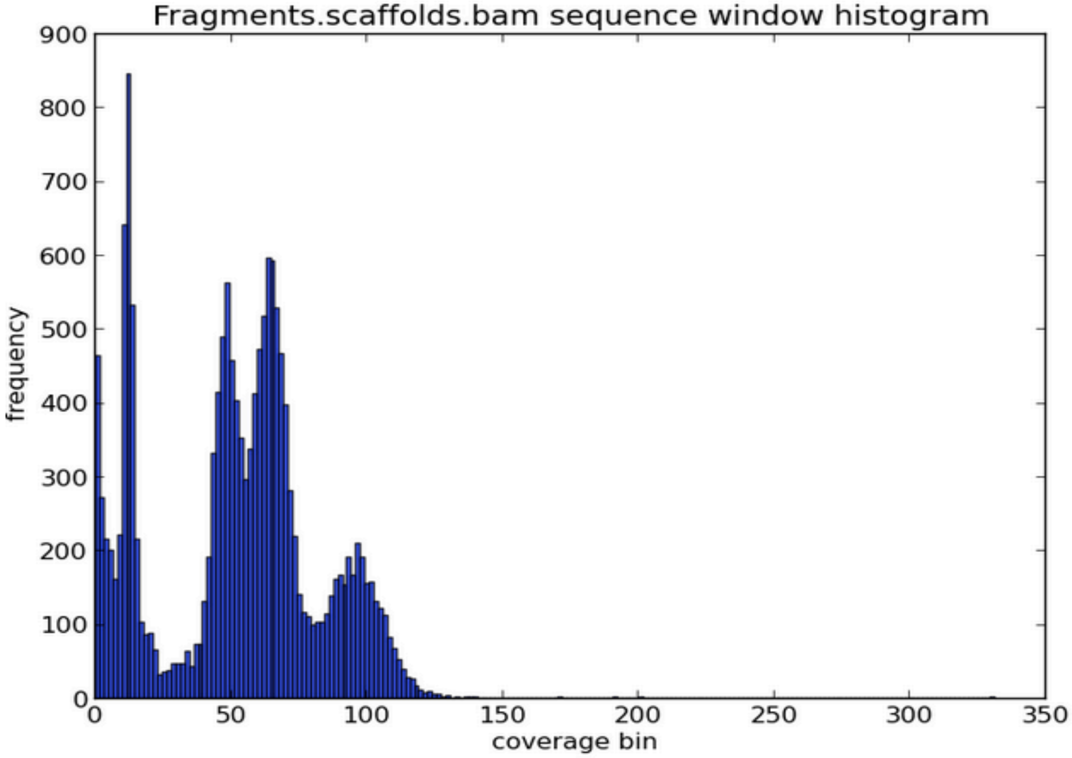
B.) Next, we want to look at how well our read data maps back to the assembly to look for any problems. What can we learn from the table below? Have we used our read data effectively?



Stat	Fragments.scaffolds (All Reads)
Total Reads	19,701,590
Paired Reads	19,701,590 (100.00%)
Duplicates	0 (0.00%)
Total Read 1	9,850,795
Total Read 2	9,850,795
Mapped	8,521,653 (43.25%)
Singletons	236,741 (2.78%)
Mapped w/ Mate	8,284,912 (97.22%)
Properly Paired	7,438,340 (87.29%)
Cross-chromosome	22,684 (0.27%)
Cross-chromosome (MQ >= 5)	6,556 (0.08%)

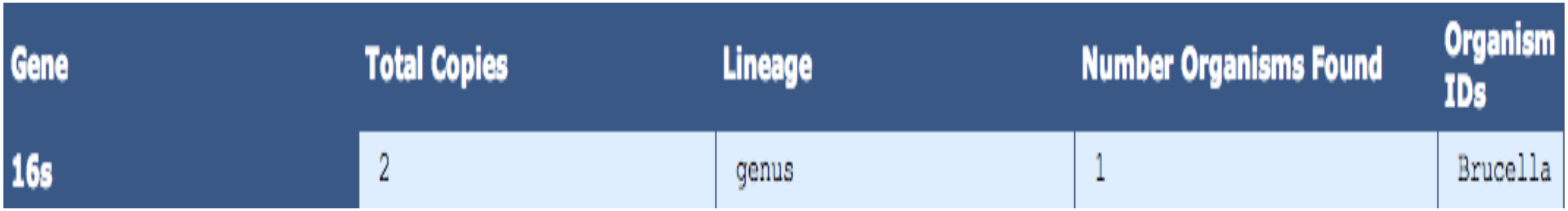
Assembly Analysis Exercise #3

C.) Now that we’ve seen how our read data was used in the assembly, we should investigate the read coverage in our assembly. Since there are so many scaffolds, looking at coverage along the reference becomes difficult. In this histogram, we count up the coverage at each base, and then plot the totals at each coverage. We expect a bell-shaped curve. What information can you quickly learn from the chart below?



Assembly Analysis Exercise #3

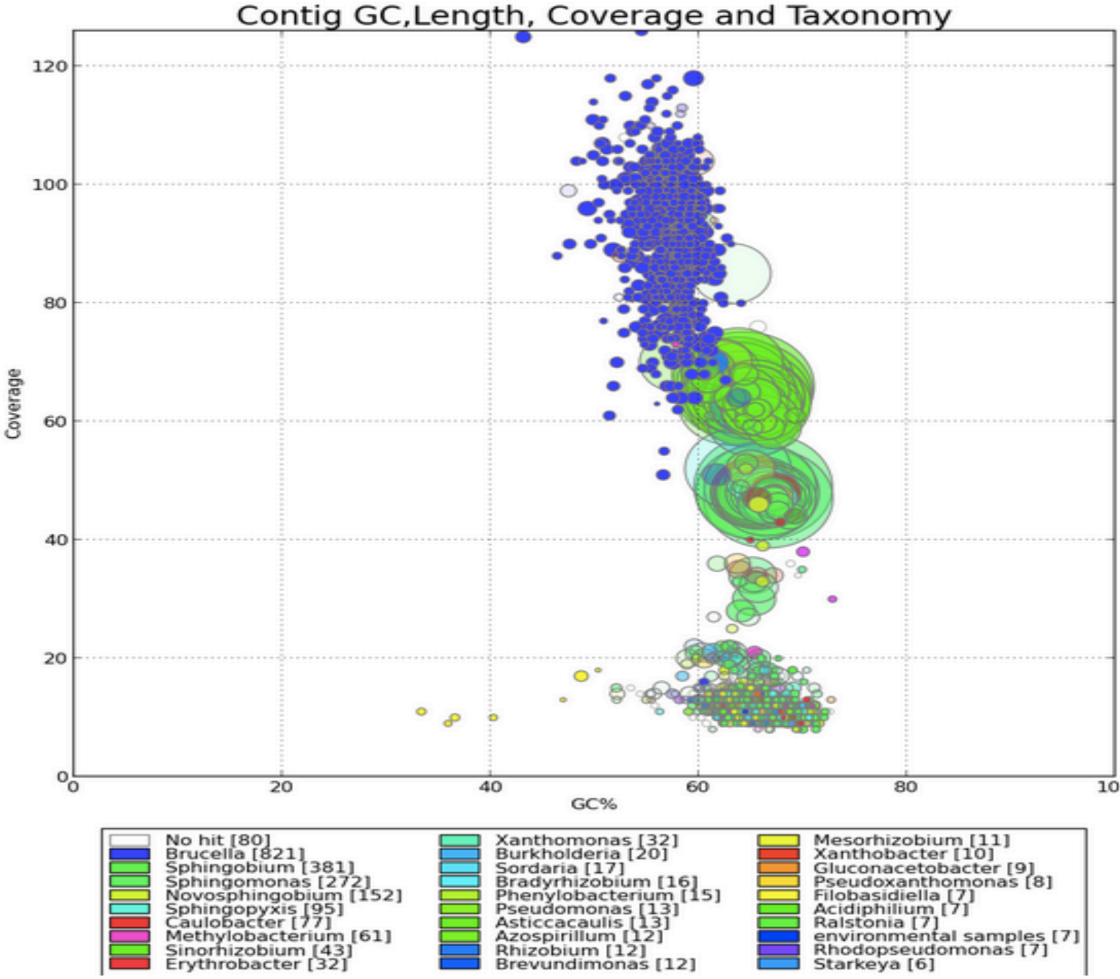
D.) Since this organism is bacterial, we can look to our 16s analysis to see if there any inconsistencies in our assembly. Are there any indications here about possible assembly problems?



Gene	Total Copies	Lineage	Number Organisms Found	Organism IDs
16s	2	genus	1	Brucella

Assembly Analysis Exercise #3

E.) BLAST taxonomy information can help determine contamination. Does the plot below indicate the presence or absence of contamination?



Assembly Analysis Exercise #3

F.) Is there a problem with this assembly? If so, what do you think is the issue? If you are unsure, what other questions could you ask about the data?

Assembly Analysis Summary

- There are many reasons for a bad assembly
- Key metrics define assembly quality
- Metrics aid in diagnosing potential issues

Assembly Analysis At The Broad

- GAEMR software package
  - http://www.broadinstitute.org/software/gaemr/
  - Python
  - Comprehensive
  - Modular

Questions?

Power of Multiple Assemblies

- Why do they help?
- Same project
  - Options Testing
  - Contamination
  - Misassembly
- Between projects
  - Sanity check metrics

Why do multiple assemblies help?

- Stochastic process
  - Small changes to input creates different results
- Many varying factors
  - Input coverage
  - Input libraries
  - Assembler Options

Multiple assemblies of the same project

- Testing options
- Impact of coverage
- Contamination detection
  - Coverage levels can reduce or remove contamination
- Misassembly verses Rearrangement
  - Reproducibility

Multiple assemblies compared between projects

- Range of metrics
  - Locate outliers
  - Contamination
  - Sequencing Bias
  - Poor library construction