Methodology article

Highly accessed Open Access

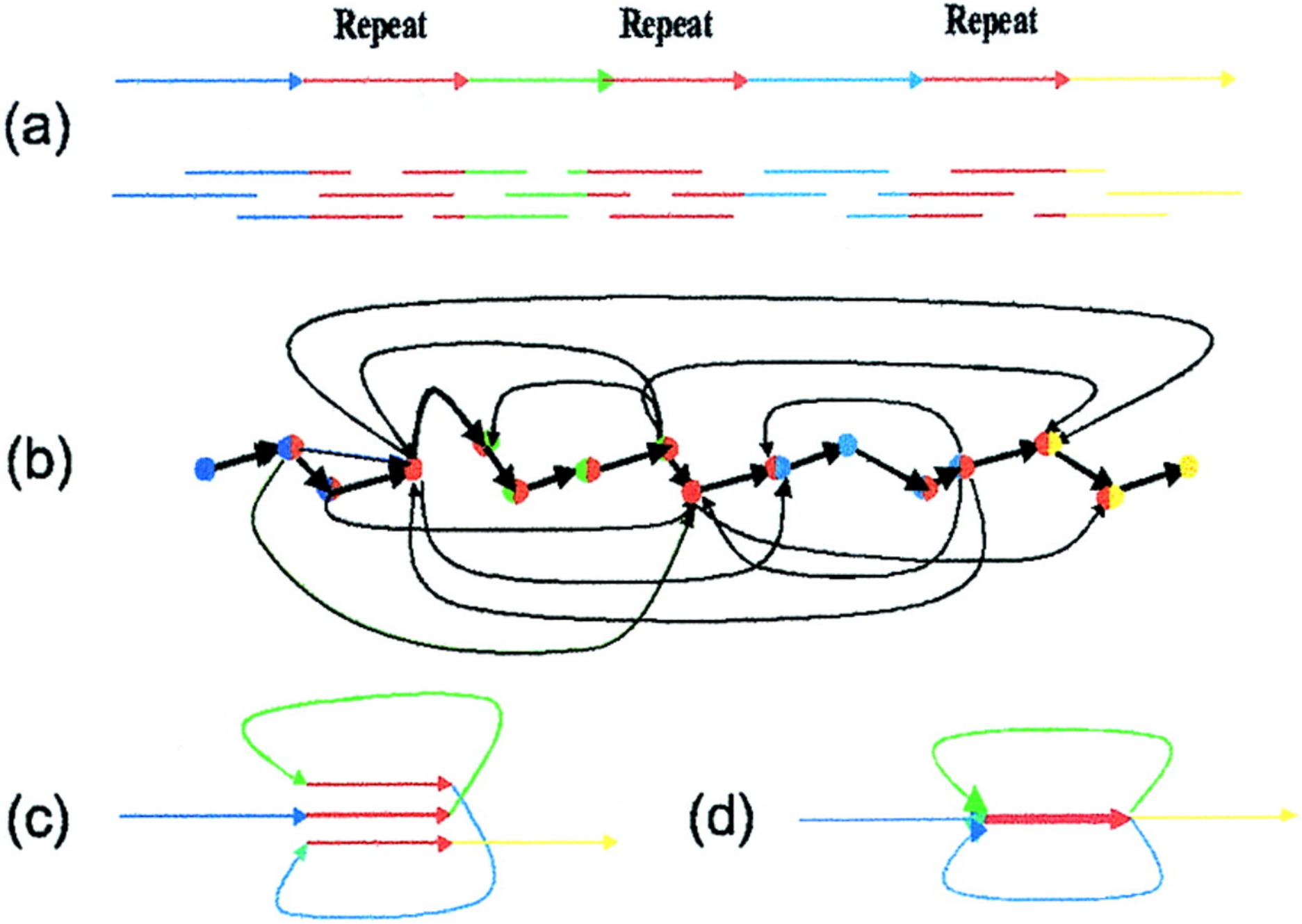
AGORA: Assembly Guided by Optical Restriction Alignment

Henry C Lin[1], Steve Goldstein[2,3,4], Lee Mendelowitz[1,5], Shiguo Zhou[2,3,4], Joshua Wetzel[6], David C Schwartz[2,3,4] and Mihai Pop[1]*

BMC Bioinformatics. 2012 Aug 2;13:189.

JC 22.04.2013

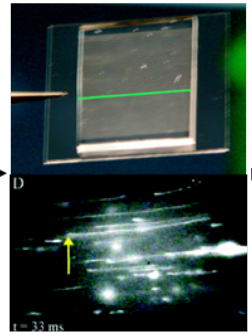
De Bruijn graph



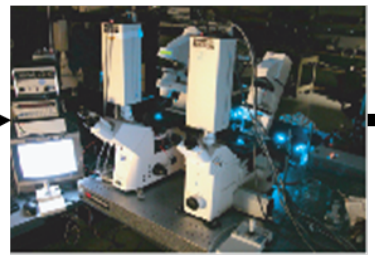
The Optical Mapping

Genomic DNA

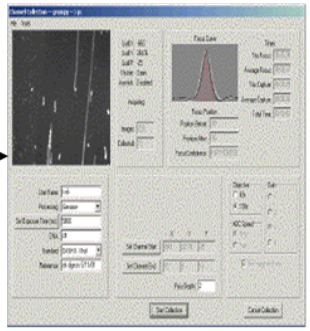
μ-fluidics



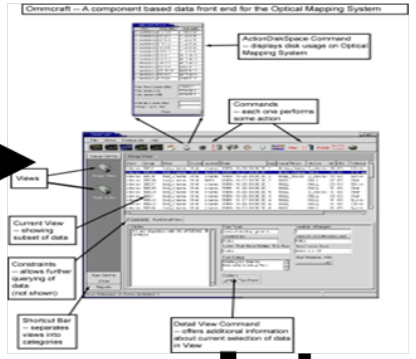
Genome Zephyr



AutoCollect



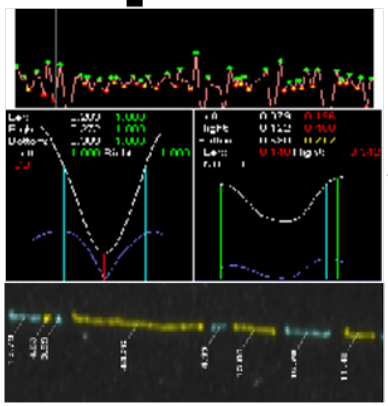
Software and System

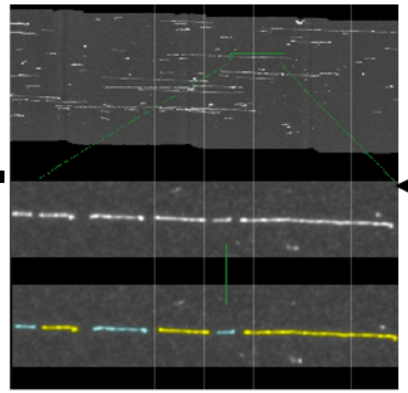


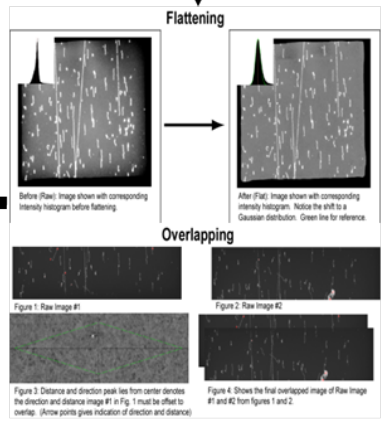
Integrated

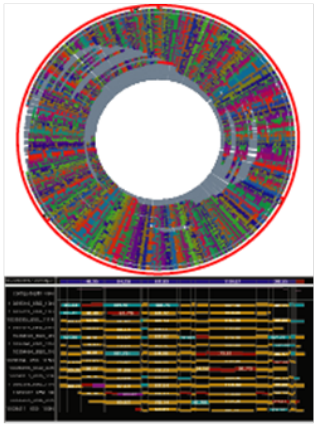
Optical Mapping System

High-Throughput









Single Molecule Map Construction by "PathFinder"

Flatten and Overlap

Map Assembler

Aim of the study

- Develop new algorithm to implement optical map data within the popular de Bruijn graph assembly paradigm

- How optical mapping error and the choice of restriction enzyme affect the quality of final assembly

- How new algorithm works with an experimentally determined optical map

Experimental data

- 369 sequenced bacterial genomes
- Created error-free de Bruijn graphs from complete genome sequences and simulated optical maps
- Additional data from published optical map of *Y. pestis KIM* genome (bacteria causing "Black Death" plague)

Error-free de Bruijn graph of a genome sequence of order k is identical to the de Bruijn graph constructed from a collection of error-free sequence reads where every k-mer in the genome is covered by at least one read

Simulating errors in optical maps

- Computed *in silico* optical map for each genome, then added errors

- Three types of errors of optical mapping:
  - Fragment sizing errors (L, U)
  - Small fragments missing (μ)
  - Restriction site errors

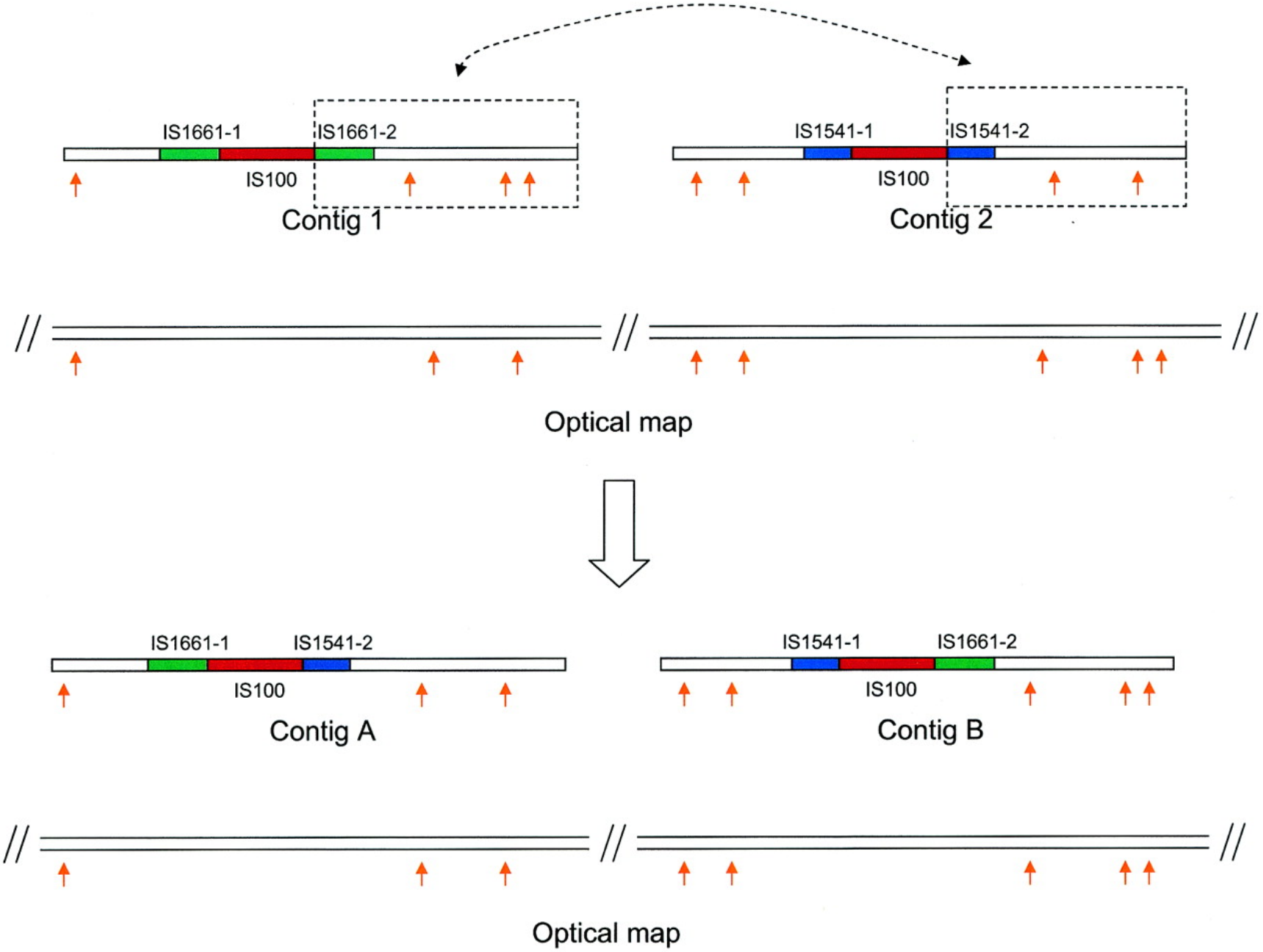
- Three levels of noise (used for genomes):
  - Low error ($\alpha = 1.01$, $\beta = 100$ bp, and $\mu = 0$)
  - Medium error ($\alpha = 1.05$, $\beta = 1000$bp, and $\mu = 1000$bp)
  - High error ($\alpha=1.10$, $\beta=2000$ bp, and $\mu=2000$ bp)

  L(lower) = max($S/\alpha - \beta$, μ) and U(upper) = $\alpha S + \beta$, where S is actual fragment size

AGORA algorithm

- Optical map information eliminates alternate paths from *in silico* map of the sequence corresponding to a partially completed path
  - Identify “landmark” edges within de Bruijn graph which only match at one location in the genome optical map
  - Parallel edges with greater than 99% sequence similarity were collapsed, as long as the difference in the sequences did not create or remove any restriction sites
  - Use depth first search (DFS) for paths between “landmark” edges
  - When partial path agrees with optical map, proceed with the DFS
  - Otherwise backtrack and proceed along different path until path to next “landmark” matching optical map is found

Sequence correction via optical mapping



Zhou S et al., Appl Environ Microbiol. 2002 Dec;68(12):6321-31.

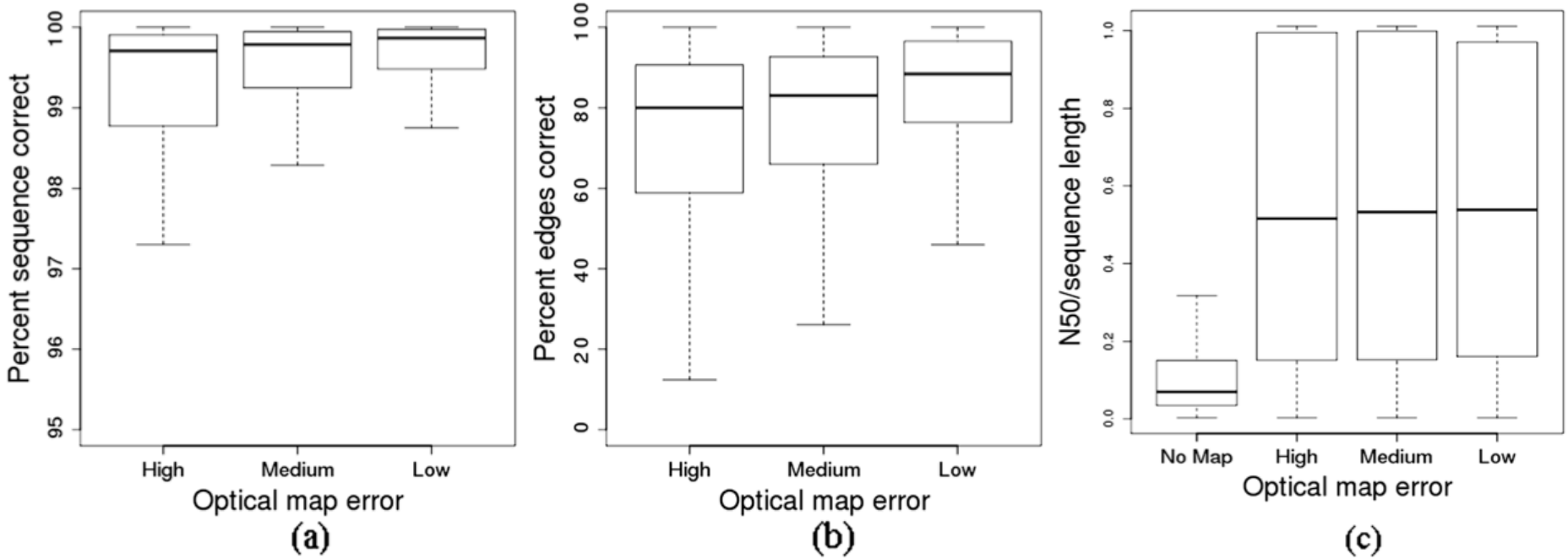
Assembly with simulated optical maps

Table 1 Statistics of the de Bruijn graphs and optical maps used in our simulations

	Min	Median	Mean	Max
Nodes	1	35	63.63	1,023
Edges	3	110	324.4	14,251
N50 Size (kbp)	14	212.1	419.2	3,587
Genome Length (Mbp)	0.34	2.91	3.2	9.14
Restriction Sites	6	334	491.7	9,668

The de Bruijn graphs for 369 bacterial genomes were generated with k-mer size 100 from the known sequences from [34] (without errors and without bubble collapsing), and the N50 size was computed for each genome, treating each edge in the de Bruijn graph as a separate contig. The row "Restriction Sites" refers to the number of cuts within the genome when using the restriction enzyme BamHI.

Assembly with optical map errors



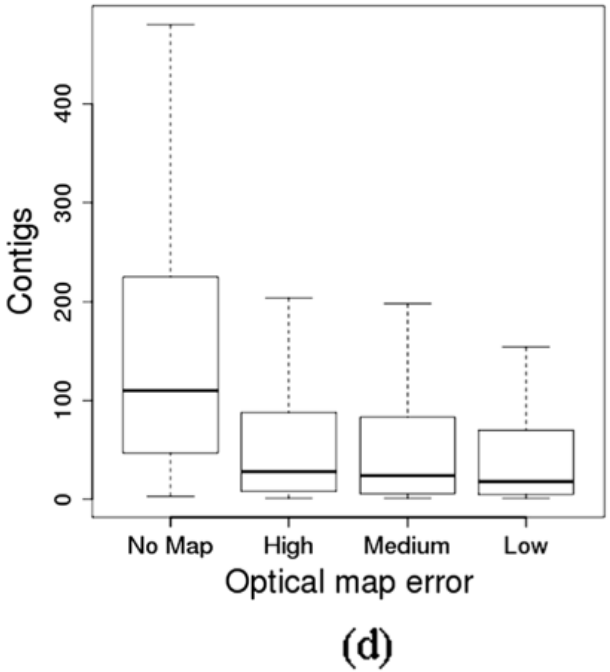


Figure 1 Assessment of the quality of bacterial genome assemblies. Measurements of the quality of assemblies produced by our algorithm on 369 bacterial genomes under three different optical map error rates. In each boxplot, we extend the whiskers beyond the upper and lower quartiles for 1.5 times the interquartile range, and omit outliers beyond the whiskers. **(a)** Sequence correctness of the assemblies, measuring the percent of the genome that was correctly assembled. (76, 72, and 65 outliers are not shown the high, medium, and low error bars, respectively.) Over 3 4 of the genomes are assembled with greater than 98% sequence correctness, even in the high error setting. **(b)** Percent of edges assembled in the correct order by our algorithm on the 369 genomes, over three error rates. The percent of edges correct is generally lower than the sequence correctness percentages, but the difference is mostly due to short edges misplaced by the algorithm. (26, 33, and 37 outliers are not shown in the high, medium, and low error settings, respectively.) **(c)** N50 size of the final contigs produced by our algorithm, after breaking genomic segments at assembly errors, normalized by genome size. (54 outliers were omitted from the first bar, measuring assembly without a map.) **(d)** Number of contigs that would be produced with no optical map (and only the de Bruijn graph), and with optical maps simulated with three different levels of noise. (We omit 42, 53, 50, and 45 outliers in the no map, high error, medium error, and low error settings, respectively.) We see a substantial improvement in both the final number of contigs and the final N50 size, when given an optical map with any one of the three error rates.

N50 improvement

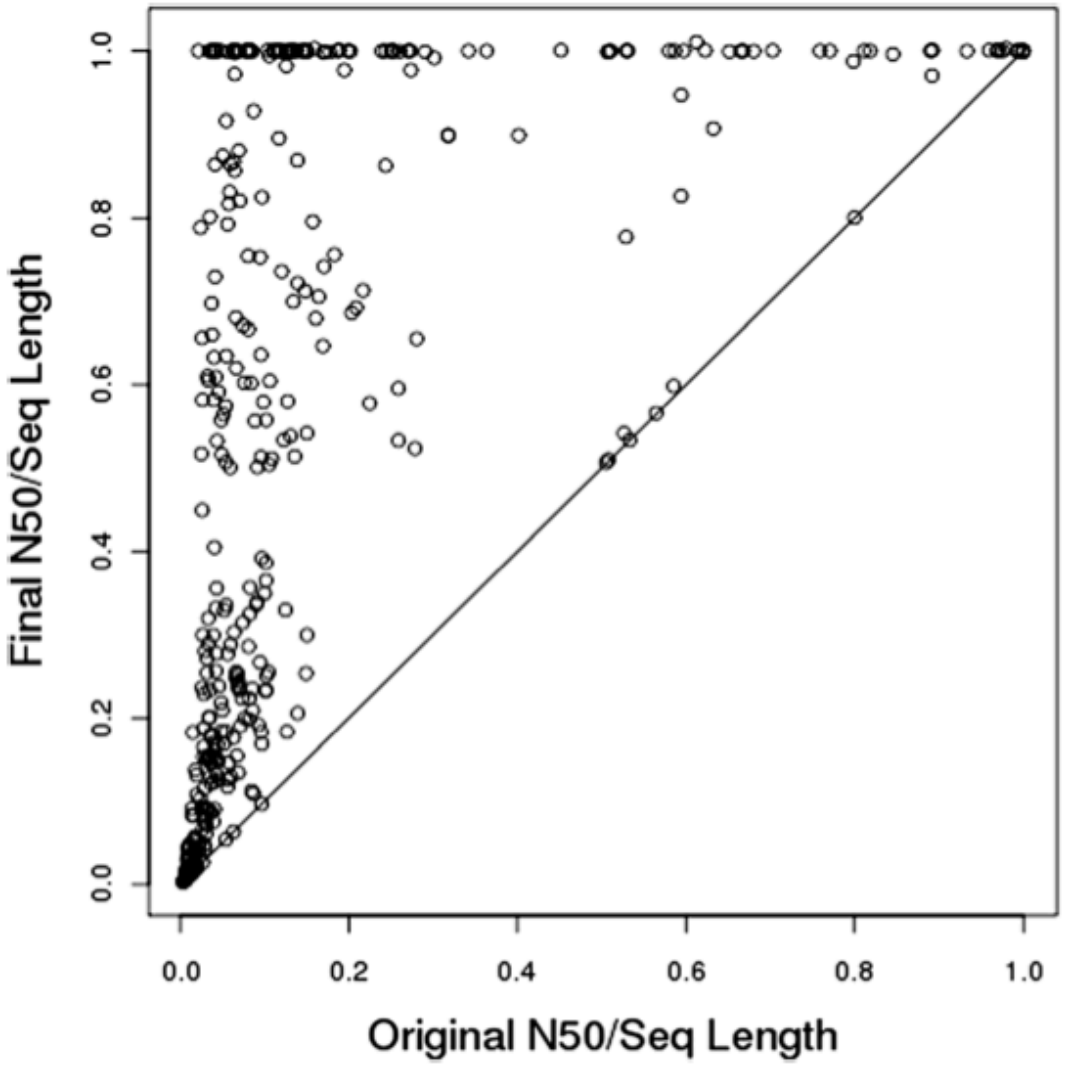


Figure 2 Improvement in normalized N50 size after assembly. For each of our 369 bacterial genomes, we plot the initial normalized N50 size (x axis) relative to the normalized N50 size after assembly (y axis) when provided a simulated optical map with the medium error rate, as described in the Methods section. The N50 sizes are normalized by dividing by the genome length. Most genomes exhibit substantial improvement in the normalized N50 size with the exception of complex genomes (with low initial normalized N50 size), and some simple genomes (with initial N50 size already close to the entire genome size).

Edge length affecting assembly quality

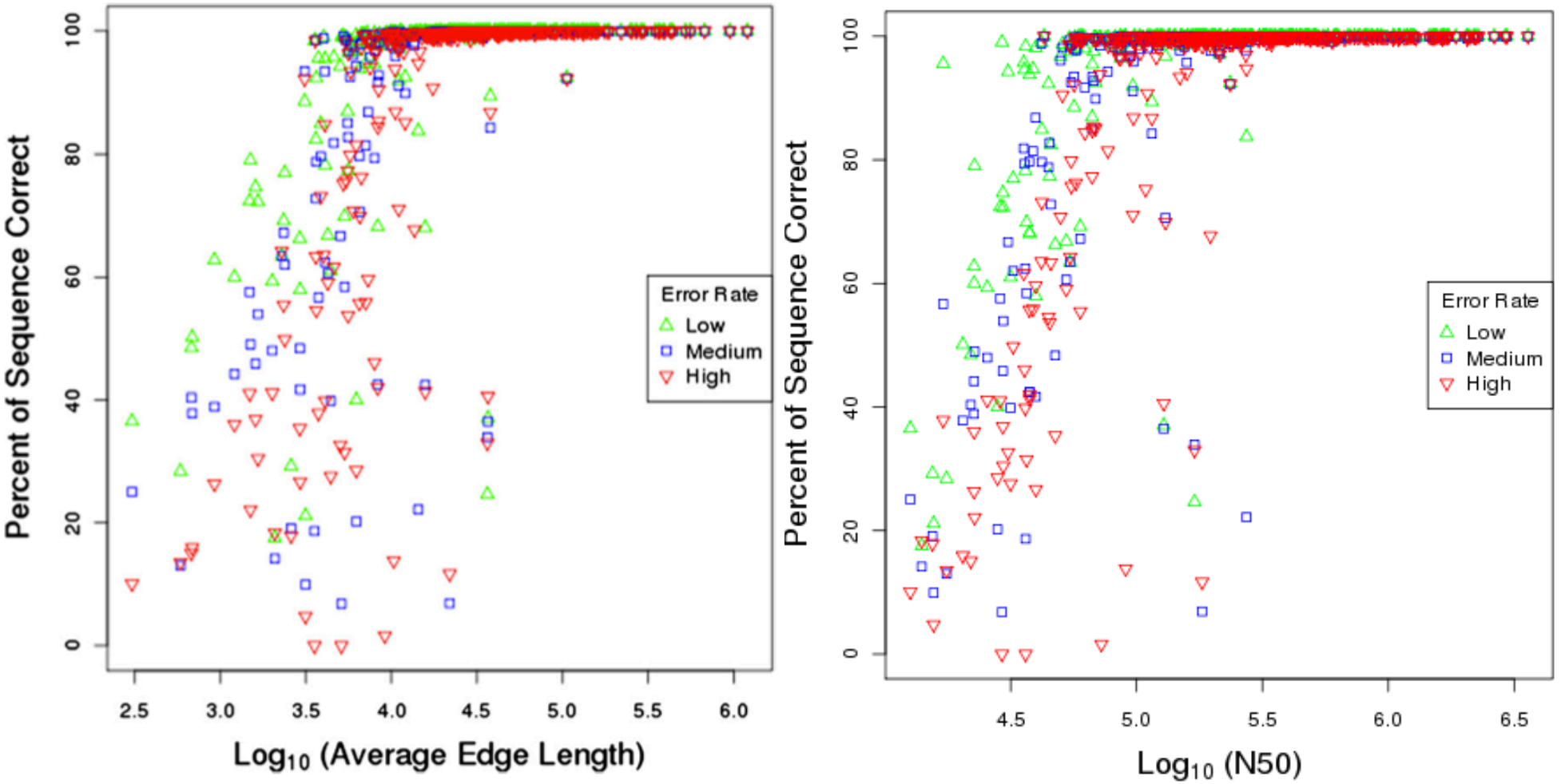


Figure 3 Impact of average edge length of de Bruijn graph on sequence correctness of assembly. A plot showing the sequence correctness of 369 bacterial genome assemblies versus the average edge length of their starting de Bruijn graphs, under three different optical map error rates. Genomes with average edge length greater than 10 kbp are generally assembled with near perfect correctness over all three error rates, while the results are mixed for genomes with shorter average edge lengths. For genomes with average edge length below 10 kbp, correctness may improve by as much as 40% when moving from the high error to low error setting, highlighting the potential benefits of more accurate mapping technologies.

Assembly with *Y. pestis* real map

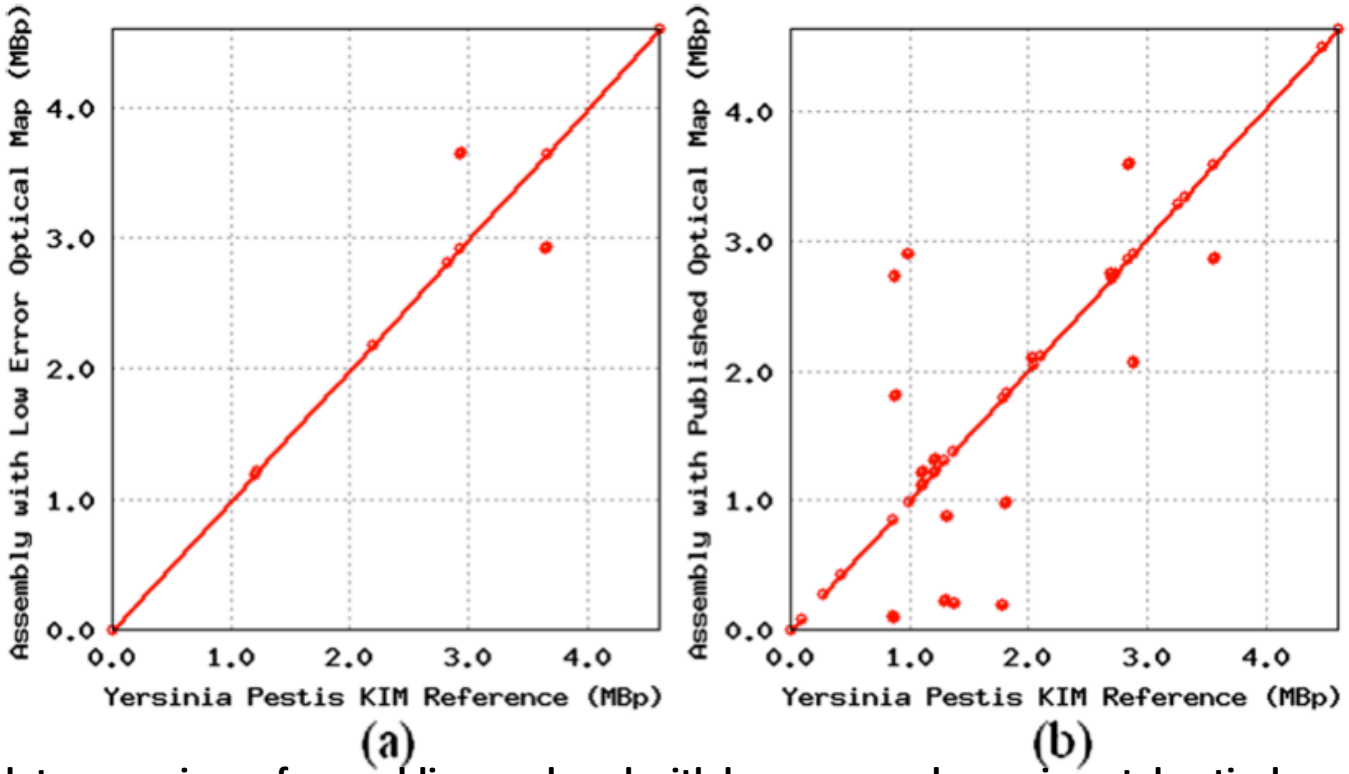


Figure 4 Mummerplot comparison of assemblies produced with low error and experimental optical map. Two dot plots generated by Mummerplot comparing the known genome sequence of Y. pestis KIM to the sequence assembly produced by our algorithm, when given an optical map with low error added **(a)**, and when using the experimental optical map from [13] **(b)**.

Table 2 Statistics on the assembly of *Y. Pestis* KIM with optical maps of different error rates

	Sequence Correct	Edges Correct	Landmarks	Final Contigs	N50 Size
Low Error	99.13%	192/199	64	12	1,190,834
Med Error	97.57%	188/199	38	20	905,369
High Error	90.52%	169/199	25	81	776,452
Map from [13]	86.74%	149/199	26	80	405,321

A summary of the results of our algorithm on assembling *Y. Pestis* KIM, when given a de Bruijn graph with k-mer size 500, and a simulated optimal or experimental optical map from [13].

Optimal restriction enzyme

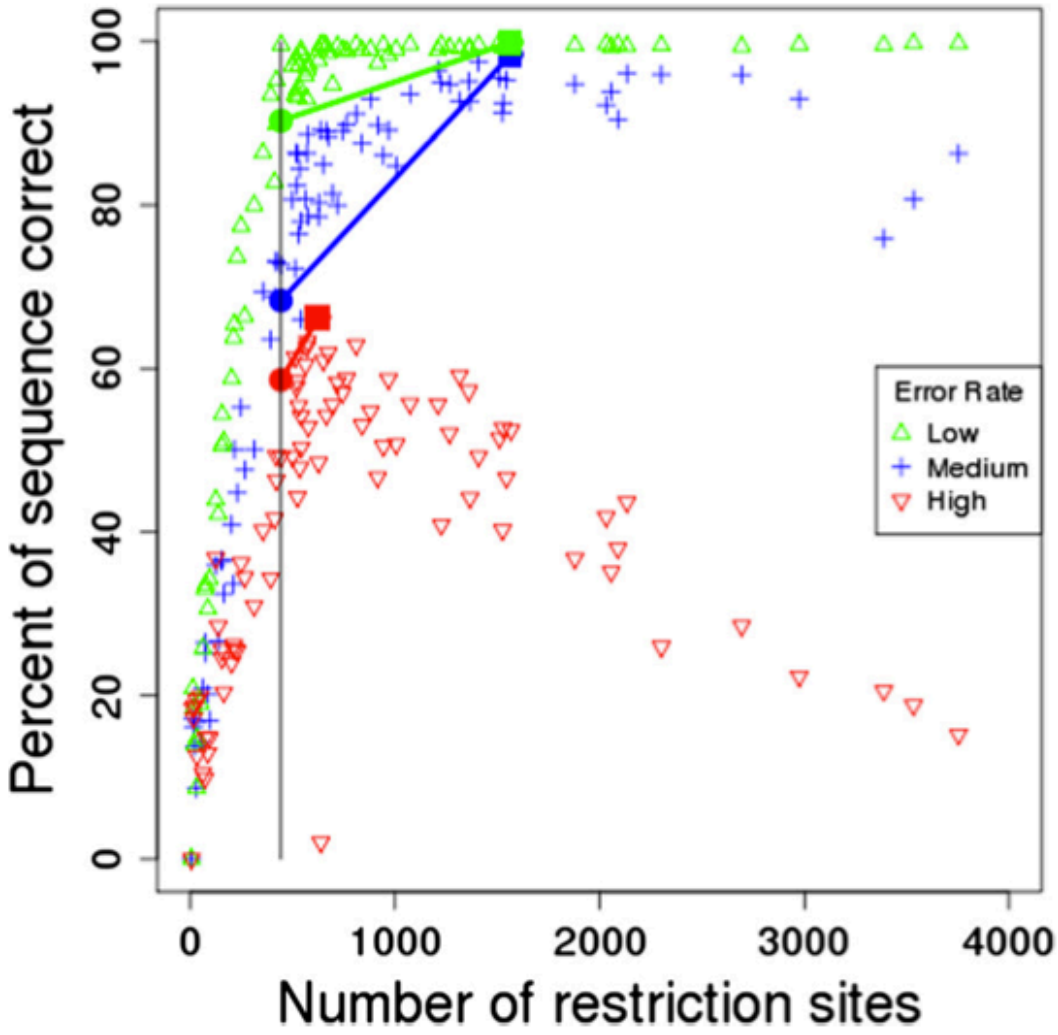


Figure 5 Impact of restriction enzyme choice on assembly quality. The choice of restriction enzymes can impact the correctness of the assembly. Each point represents the sequence correctness of an assembly of Y. pestis KIM when given a de Bruijn graph of k-mer size 100 and an optical map of low, medium, or high error rate. The vertical line in the picture indicates the number of restriction sites for the enzyme PvuII used to construct the experimental optical map of this genome, and the colored circles represent the correctness that can be achieved under the three error rates for the PvuII enzyme. The red, blue, and green filled squares to the right of the vertical line, indicate an improvementof between 7.7% and 30.1% in the final sequence correctness that can be achieved when choosing a better restriction enzyme in the high, medium, and low error settings, respectively.

Conclusions

- Implemented AGORA algorithm that uses optical restriction maps in improved genome assembly
- With optical maps, over ¾ of tested bacterial genomes were assembled with over 98% accuracy
- N50 size improved by a factor between 6.4 and 18.9. Contig number reduced by a factor of between 6.67 and 10.74
- Choice of restriction enzyme significantly affects assembly quality
- Only tested with error-free assembly data and need to be adapted with real sequencing outcome

De Bruijn graph

- Graph **nodes** correspond to k-mers (sequences of length k) and **edges** correspond to (k+1)-mers
- Edge joins two nodes if one of the nodes is a prefix of the edge and other is a suffix
- Node is created for each k-mer in the set of reads and an edge for each (k+1)-mer
- *"Chinese postman path"* – path through de Bruijn graph that visits all edges at least once (which represents the true genome sequence)
- Implementations: Velvet, SOAP, ALLPATHS, ABySS