Scaffold-builder for Combining De Novo and Reference-guided Assembly

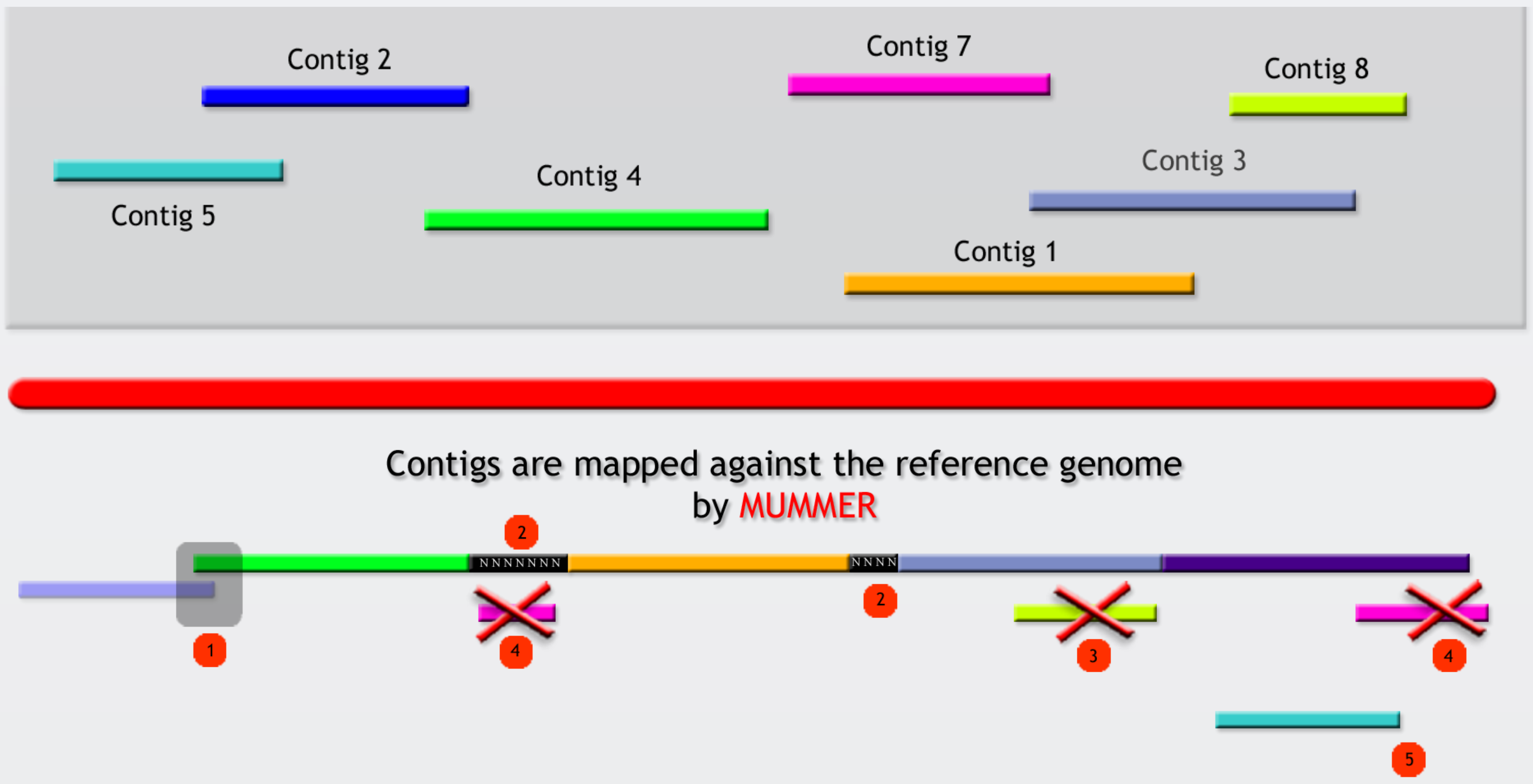
Genivaldo G. Z. Silva, Bas E. Dutilh, T. David Matthews, Keri Elkins, Elizabeth A. Dinsdale and Robert A. Edwards.

Summary

The abundance of repeat elements in genomes can impede the assembly of a single sequence. The tool **scaffold_builder** was designed to generate scaffolds (super contigs of sequences joined by N-bases) using the homology provided by a closely related reference sequence.

Methods

Scaffold_builder is an advanced wrapper for Nucmer, written in Python. The Figure below illustrates how **scaffold_builder** resolves several situations that may arise when mapping contigs to the reference genome.



1 **Overlap:** align the overlaps using Needleman-Wunsch's algorithm.

2 **Filling the gaps:** fill the gaps with N in the regions without a contig mapping.

3 **Overlapping contig sub region:** the contig is ignored because it maps in a location where was occupied by another contig with a longer hit.

4 **Ambiguous mapping:** contigs ignored in scaffolding because they mapped to more than one location on the reference.

5 **Contig not mapped:** contigs ignored in scaffolding because they were not mapped to the reference.

Results

The application was evaluated using simulated pyrosequencing reads of the three bacterial genomes, and two newly sequenced genomes. As shown in the Table below, **scaffold_builder** decreases the number of contigs by ~62% while increasing their average length by ~200%.

Conclusions

Scaffold_builder helps to create longer sequences during genome assembly. It allows the user to combine the strengths of *de novo* assembly with the structure provided by a closely related reference.

	Escherichia coli 042	Lactobacillus salivarius UCC118	S. enterica subsp. Enterica sv Typhi P-stx-12	S. typhimurium SDT1291	S. typhimurium G455
	Simulated data			Real data	
Number of sequencing reads	400,000	400,000	400,000	341,126	388,386
Average Number of sequences (Assembly)	75.1	41.1	62.1	259.0	159.0
Average Length (Assembly)	69,443.8	43,439.7	75,805.8	18,591	30,383
Average Number of sequences (Scaffold)	43.8	14.0	24.1	74.0	50.0
Average Length (Scaffold)	112,014.1	203,819.5	179,623	63,123	94,698

Web-based version and Code:

http://edwards.sdsu.edu/scaffold_builder