Accuracy of *de novo* assembly of Sanger trace data: SeqMan Pro versus three alternative pipelines

Introduction

DNASTAR provides two software applications for assembling DNA sequence fragments: SeqMan NGen and SeqMan Pro. SeqMan NGen is used to assemble Next Generation Sequencing (NGS) data, while SeqMan Pro is primarily used to assemble Sanger ABI trace data.

A previous white paper described SeqMan Pro's "trace evidence" method of consensus calling, including its proprietary quality scoring system for trace data.

This paper discusses the results of *de novo* assembly trials designed to test the accuracy of SeqMan Pro's assembler, versus those used in three competing software applications:

- Geneious 10.1.3
- Sequencher DNA Sequence Analysis Software 5.4.6
- CLC Bio Genomics Workbench 10.01

Two data sets were tested, and consisted of Sanger ABI reads from *E. coli* and from a *Shigella* plasmid. In both cases, SeqMan Pro's "trace evidence" method of consensus calling generated the most accurate consensus. In addition, SeqMan Pro assembled more reads than any of the other three applications.

Testing Procedure

All work was performed on the same 64-bit Windows 7 machine. The testing procedure consisted of two main steps:

1) *De novo* assemble the sets of Sanger ABI reads using each of the four software applications. Default consensus-calling settings were used for all four assembly applications. In the case of SeqMan Pro, these defaults included use of the "Pro" assembler and the "trace evidence" method of consensus calling. If available in a given application, both quality trimming of sequences and automatic removal of Janus vector were requested prior to assembly.

2) Perform a pairwise alignment in order to determine the number of mismatches between the calculated consensus and the published reference sequence. This step utilized the EMBOSS program "Stretcher," based on the Needleman-Wunsch alignment algorithm.

Data

Two data sets were assembled using each of the four applications:

- The *"E. coli"* data set consisted of 498 files in Sanger .abi trace data format. After *de novo* assembly in each of the applications, Stretcher was used to align the resulting consensus sequence(s) against a 39,929 bp reference genome fragment from *E. coli* K-12 MG1655 (Blattner FR *et al*.,1997).

- The *"Shigella"* data set consisted of 540 files in Sanger .abi trace data format. After *de novo* assembly in each of the applications, Stretcher was used to align the resulting consensus sequence(s) against a 23,555 bp reference genome fragment from *Shigella flexneri* plasmid pWR501 (Wei J *et al.* 2003; Venkatesan MM *et al.* 2001).

Results

Tables 1 and 2 show the accuracy results and other statistical metrics for each of the two data sets.

Table 1. Accuracy results for the *E. coli* data set

Assembler	# Contigs	# Errors[1]	# Reads Assembled[1]	Contig Length[1]	% of Ref Covered[1,2]
DNASTAR SeqMan Pro 14.1	**1**	**20**	**498**	**39,772**	**99.61**
Geneious 10.1.3	1	65	498	39,593	99.16
Sequencher DNA Sequence Analysis Software 5.4.6	2	365	497	41,067	102.85
CLC Bio Genomics Workbench 10.01	13	184	495	43,978	110.14

Table 2. Accuracy results for the *Shigella* data set

Assembler	# Contigs	# Errors[1]	# Reads Assembled[1]	Contig Length[1]	% of Ref Covered[1,2]
DNASTAR SeqMan Pro 14.1	**1**	**10**	**540**	**23,547**	**99.97**
Geneious 10.1.3	1	17	539	23,518	99.84
Sequencher DNA Sequence Analysis Software 5.4.6	1	42	532	23,561	100.03
CLC Bio Genomics Workbench 10.01	9	192	533	27,076	114.95

[1] Column contains summations for all contigs represented by that row.

[2] Due to overlap (from multiple contigs and/or the circular nature of the *Shigella* plasmid), it was possible for “% of Reference Covered” to exceed 100%.

Conclusion

The results shown in Tables 1 and 2 demonstrate that, of the four products tested, SeqMan Pro's default assembly method gave the highest accuracy when assembling Sanger ABI trace data.

Compared to the other three applications, SeqMan Pro 14.1 created single contigs in both tests (as did Geneious), made the fewest errors, and had the greatest number of reads incorporated into the assemblies. It also created the most complete coverage of the target sequence without introducing circular redundancy.

References

1) Blattner FR *et al*. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. Science 05 Sep 1997: Vol. 277, Issue 5331, pp. 1453-1462. DOI: 10.1126/ Science.277.5331.1453.

2) Wei J *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect Immun 71(5): 2775-2786. [PubMed] [full text] [abstract]

3) Venkatesan MM *et al.* (2001) Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. Infect Immun 69(5): 3271-3285. [PubMed] [full text] [abstract]

4) Myers E and Miller W, "Optimal Alignments in Linear Space," CABIOS 4, 1 (1988), 11-17. [EMBOSS Stretcher, PubMed]

[This white paper was last updated on May 24, 2017.]