De Novo Sanger Data Assembly in SeqMan Ultra

The subject of this paper is SeqMan Ultra's "Classic" assembler, the one used for the vast majority of Sanger assemblies in Lasergene. This assembler was developed originally for SeqMan Ultra's pre-2020 predecessor, SeqMan Pro, and is extremely fast and accurate. For de novo assemblies, this assembler uses a proprietary consensus calling method, "Trace Evidence," which has proven superiority over the commonly used "Majority" method.

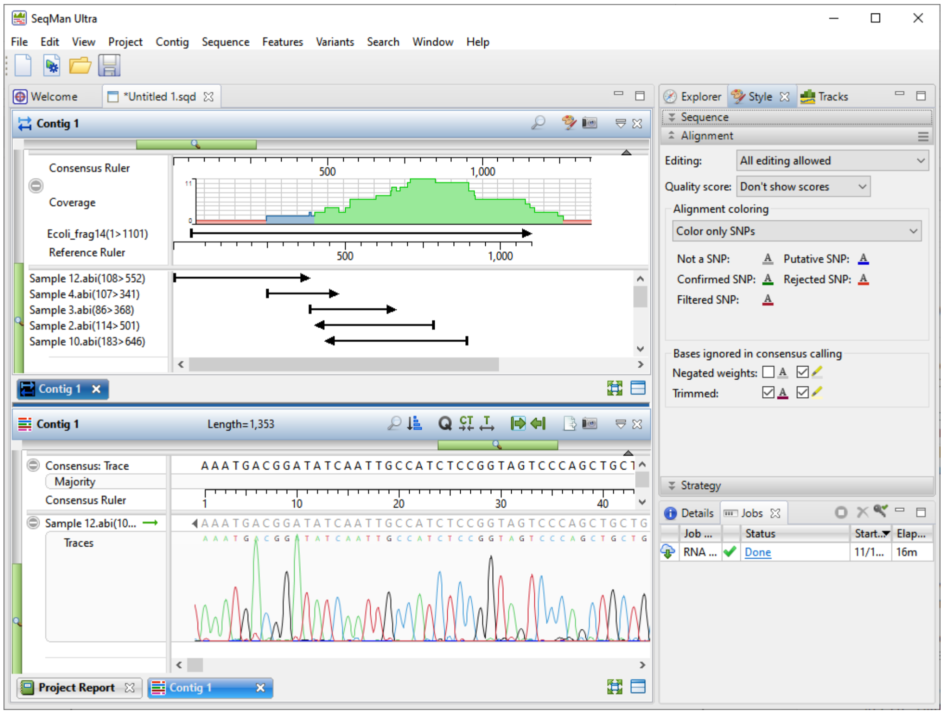
The first part of this paper discusses the results of de novo assembly trials designed to test the accuracy of SeqMan Ultra's assembly algorithm versus those used in three competing software applications. The second part provides background on how the "Classic" assembly algorithm's "Trace Evidence" method calculates and uses quality scores for exceptional accuracy.

Part I. De Novo Sanger Assembly Accuracy: Seqman Ultra Vs. Three Alternative Pipelines

Introduction

This part of the paper discusses the results of *de novo* assembly trials designed to test assembly accuracy in four competing applications: SeqMan Ultra (Figure 1), Geneious, Sequencher DNA Sequence Analysis Software, and CLC Bio Genomics Workbench.

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



Editing: All editing allowed

Quality score: Don't show scores

Alignment coloring: Color only SNPs

Figure 1. A completed Sanger assembly being analyzed in SeqMan Ultra

Testing Procedure

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

- *De novo* assemble the sets of Sanger ABI reads using each of the four applications using their default consensus-calling settings. For SeqMan Ultra, the default consensus calling method is "trace evidence," which is described in detail in Part II of this paper. If available in a given application, both quality trimming of sequences and automatic removal of Janus vector were requested prior to assembly.
- Perform a pairwise alignment 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 2 and 3 (next page) show the accuracy results and other statistical metrics for each of the two data sets.

Part I Conclusion

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

Compared to the other three applications, SeqMan Ultra's algorithm 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.

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 Ultra[3]	**1**	**20**	**498**	**39,772**	**99.61**
Geneious[3]	1	65	498	39,593	99.16
Sequencher DNA Sequence Analysis Software[3]	2	365	497	41,067	102.85
CLC Bio Genomics Workbench[3]	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 Ultra[3]	**1**	**10**	**540**	**23,547**	**99.97**
Geneious[3]	1	17	539	23,518	99.84
Sequencher DNA Sequence Analysis Software[3]	1	42	532	23,561	100.03
CLC Bio Genomics Workbench[3]	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%.

[3] SeqMan Pro 14.0 (SeqMan Ultra uses the identical algorithm), Geneious 10.1.3, Sequencher DNA Sequence Analysis Software 5.4.6, CLC Bio Genomics Workbench 10.01

Part II. Algorithms used in SeqMan Ultra for *de novo* Sanger assembly

This part of the paper describes the “Trace Evidence” consensus calling method used in SeqMan Ultra’s “Classic” assembly algorithm.

SeqMan Ultra’s Trace Evidence consensus calling method

In 1999, DNASTAR software developer Carolyn Allex published a doctoral thesis (Allex CF, 1999), in which she compared several algorithmic methods for consensus calling, including the “Majority” method and her own “Trace Evidence” method (Allex CF *et al.* 1997). The latter method was a novel approach for generating quality scores and consensus calling based on geometry and quality of peaks in the trace data.

Allex’s analysis indicated that Trace Evidence had significantly better consensus calling accuracy than Majority, even when many of the individual bases had been called incorrectly. Trace Evidence was also more likely than Majority to make the correct call when the base of the well-defined (true) peak was hidden below a high-intensity valley. By contrast, Majority methods often incorrectly called the base that was associated with the valley.

DNASTAR implemented the Trace Evidence algorithm into SeqMan Ultra, and later, into its successor, SeqMan Ultra. The Majority method is now recommended only when data consist of text sequences rather than fluorescence trace data.

In addition, the quality scoring algorithm that was developed for use with the Trace Evidence method is now also used in SeqMan Ultra for SNP calling and quality-based sequence end trimming.

Quality score calculations

When ABI trace data are used in an assembly, SeqMan Ultra analyzes the shape and intensity of peaks to calculate quality scores (Q), and averaged quality scores (Q/n). In quality score calculations:

- Taller, sharper peaks receive higher scores than less distinct peaks. The heights of any underlying peaks are subtracted from the highest peak's score during the calculation.
- The further a peak is from the location at which the base was called, the lower the quality score.

The trace data for a DNA sequence comprises four sets of traces—one each for *A*, *C*, *G*, and *T*. Each trace contains a sequence of intensity values that can be plotted to form a graphical display of trace data. The portions of the four traces associated with a single base call each contain about ten to twelve data points. Only the trace from which the base call is derived is used to calculate a quality score (e.g., if the base call is a *T*, only the *T* trace is analyzed to calculate a quality score). Figure 2 shows data plotted for five base calls. The data points associated with the center base—a *T*—are marked with black dots.

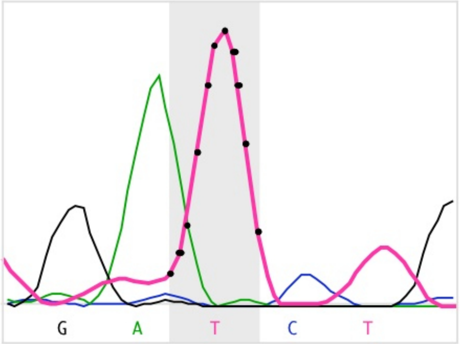


Figure 2. Data plotted for five base calls

SeqMan Ultra calculates each of the peaks in the trace data. A *peak* is defined as trace data that exhibits negative curvature. Slope is used to differentiate between three kinds of peaks: *strong*, *medium*, and *weak*. *Strong* peaks exhibit a change in the sign of the slope, *medium* peaks contain a shoulder with a slope of zero, and *weak* peaks have neither a change in sign nor a shoulder.

If the trace data for a base call do not contain a peak, its quality score is zero. Figure 3 contains examples of the three kinds of peaks for the highlighted *T* base.

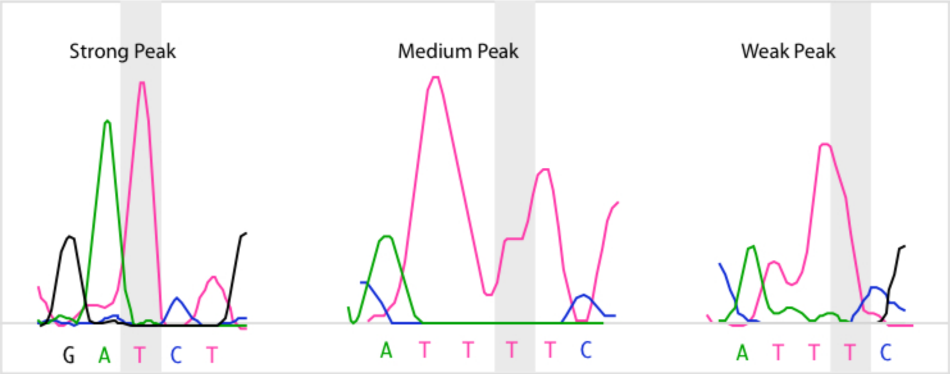


Figure 3. Examples of the three kinds of peaks for the highlighted 'T' base

Quality score calculations take into account several parameters:

- Three extreme intensity points: *5Min* (5' minimum), *3Min* (3' minimum), and *Max*. *5Min* and *3Min* (examples shown in Figures 4-7) are the intensity values to either side of the base call that are the minimum values of the data for that base. If a run of identical base calls occurs, then the minimums are taken from either side of the homopolymeric run. *Max* is the intensity value of the peak.

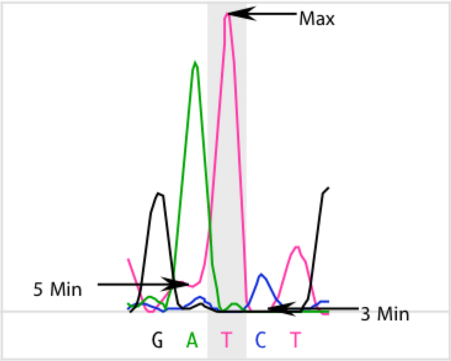


Figure 4. Intensity point example A

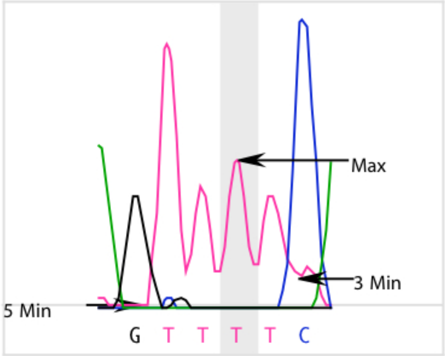


Figure 5. Intensity point example B

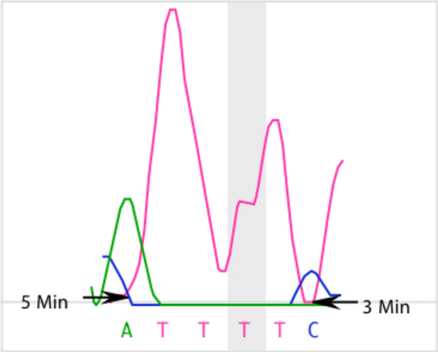


Figure 6. Intensity point example C

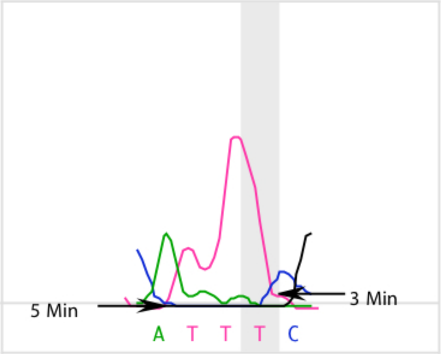


Figure 7. Intensity point example D

- Each quality score calculation includes division by the maximum intensity of all four traces for an entire sequence. This assigns higher scores to higher peaks. In this example, an *A* peak has the highest intensity value. Its intensity value, *MaxTrace* (Figure 8) is used in the quality score calculation for all bases in the sequence.

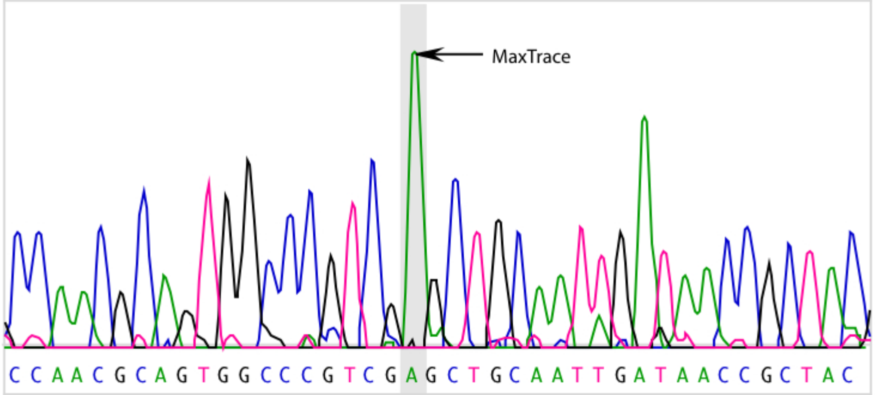


Figure 8. MaxTrace example

- Trace data files identify the point in the trace data where the base was called, or "distance weight." In high quality data, this usually coincides with the point where SeqMan Ultra detects a peak. In poorer quality data, the peak can be offset significantly. Each quality score is adjusted to reflect the distance from the detected peak to the point where the base was called (Figure 9).

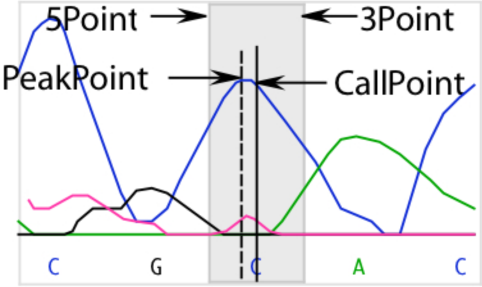


Figure 9. Example of peak offset

The fraction of the number of points in the offset to the total number of points is the *Dist* weight used in the quality score calculation. *Dist* is calculated as follows (Equation 1).

$$Dist = \frac{\text{abs}\left(PeakPoint - CallPoint\right) + 1}{\left(3Point - 5Point\right) + 1}$$

Equation 1. Calculation of *Dist*

End-trimming based on averaged quality scores

SeqMan Ultra uses averaged quality scores (*Q/n*) to identify regions of poor-quality data at the end of sequences. Averaged quality scores are calculated as the average of the quality scores, *Q*, over a window of 21 bases. The average score is assigned to the base in the center of the window. Averaging the scores smooths out the quality scores and quantifies the general quality of data in a region. To perform quality end-trimming, a threshold is set and the longest sequence of bases with all *Q/n* meeting the threshold is identified. Below-threshold ends to either side of the high-quality region of the sequence are trimmed off before assembly.

In Figure 10, the quality scores, *Q*, and averaged quality scores, *Q/n*, are graphed for the 5' end of a 794 base pair sequence. A dashed horizontal line marks the quality end-trimming threshold. The average scores, *Q/n*, are compared to the threshold and the first 14 bases are trimmed from the 5' end of this sequence.

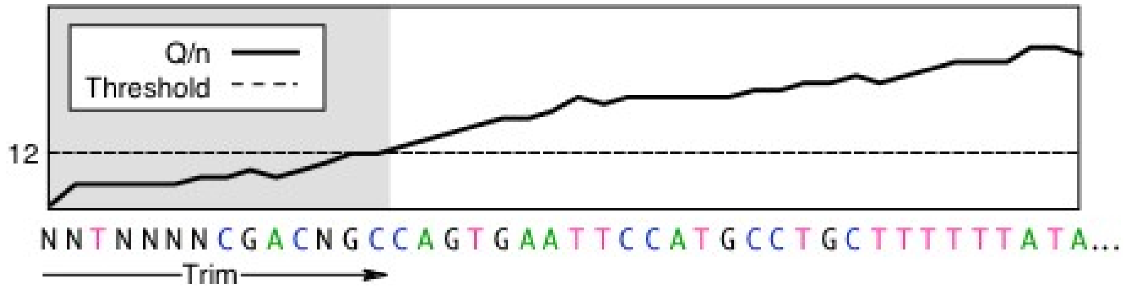


Figure 10. Quality score graph for the 5' end of a 794-base pair sequence

Poor quality data on the ends of sequences often contain miscalled bases that produce mismatches in alignments with other sequences. If the number of mismatches is high enough that SeqMan Ultra's Minimum Match Percentage threshold is not met, sequences will not be assembled in the same contig. Trimming the poor quality data from the ends of sequences allows better and more complete assembly.

Figure 11 compares *Q* to *Q/n* scores for the 5’ end of a sequence.

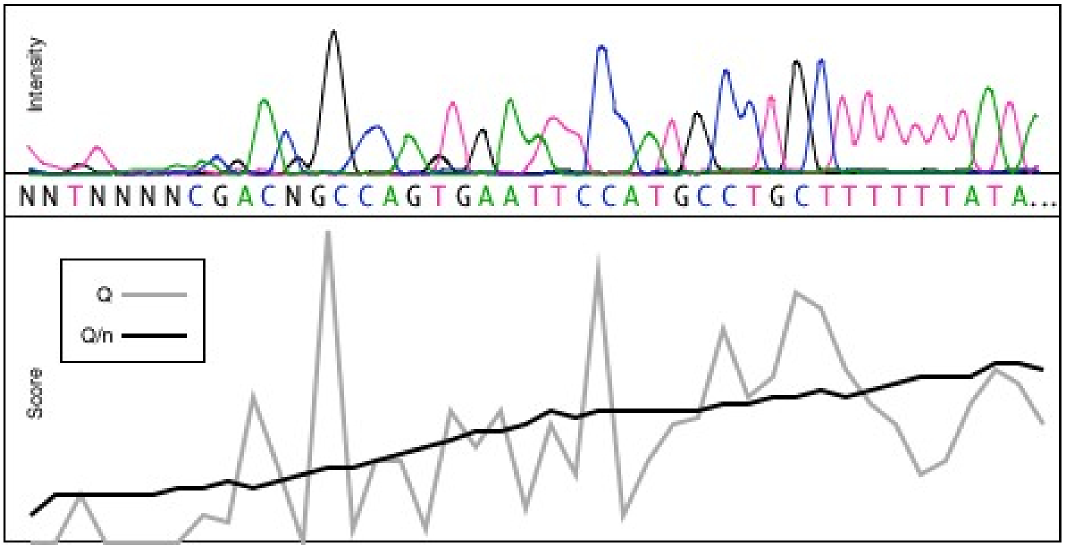


Figure 11. Comparison of Q to Q/n scores for the 5’ end of a sequence

SNP calling using neighborhood quality scores

SeqMan Ultra provides the option to use a neighborhood quality score threshold when identifying SNPs. This threshold can be changed by opening SeqMan Ultra’s SNP Discovery Parameters dialog and defining a new ‘Neighborhood Window’ value. By default, the Neighborhood Window value is zero, meaning that the neighborhood quality score threshold is not used.

A neighborhood quality score is equal to the lowest quality score of any of the bases in the defined window surrounding a SNP base. The size of the window can be adjusted by editing the Neighborhood Window value. For example, if the Neighborhood Window value is set to 5, then the 5 bases upstream and the 5 bases downstream from the SNP base will be considered. If the specified window contains one or more mismatches to the reference sequence, the putative SNP will be rejected.

Unless you have specific thresholds you would like to use, you may wish to start with the Q-Score Threshold values from Altshuler *et al*. (2000):

- Minimum Score at SNP: 20
- Minimum Neighborhood Score: 15
- Neighborhood Window: 5

Benchmark testing for Trace Evidence vs. Majority in SeqMan Ultra

In October 2016, we used SeqMan Pro version 14 to perform a benchmark test comparing errors resulting from the Majority and Trace Evidence methods of consensus calling. Though results were not re-run using SeqMan Ultra 17, we expect that application would produce identical results, as the algorithm has not changed.

SeqMan Pro was used to automatically remove Janus vector and then *de novo* assemble 498 *.abi* trace files from *E. coli* using the "Classic" assembler. Assembly took only a few seconds. The consensus was calculated twice, once using the Majority method, and once using Trace Evidence. The resulting consensus sequences were then exported and saved.

Next, SeqMan Pro was used to assemble the consensus sequences against a 39,774 bp reference genome fragment from *E. coli* K12 MG1655 (Blattner FR *et al*.,1997). A SNP (variant) report was generated automatically. Since the reference genome should theoretically be an exact match to all four consensus sequences, any observed differences were presumed to be errors, rather than true SNPs. Table 3 lists the number of errors found in each of the consensus sequences.

Table 3: Number of consensus calling errors using combinations of two assembly methods and two consensus calling algorithms in SeqMan Ultra

Number of Errors	
Majority Method	Trace Evidence Method
153	11

Conclusion for Part II

These data show that the Trace Evidence method produced far fewer consensus calling errors than the Majority method, corroborating the findings of Allex CF (1999).

Resources and Free Trial Software

To try SeqMan Ultra and see the results of its quality-score based algorithms, download and install a free trial of Lasergene.

For help getting started with SeqMan Ultra, consult our easy-to-understand SeqMan Ultra User Guide, which also features written tutorials with sample data. You can also get friendly, helpful support from fellow scientists via support@dnastar.com or by calling one of the numbers in the footer of this document.

References

- Allex, C.F., Baldwin, S.F., Shavlik, J.W., and Blattner, F.R. (1996). Improving the quality of automatic DNA sequence assembly using fluorescent trace-data classifications. *Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology*, 3-14. St. Louis, MO. Menlo Park, CA. *ISMB-96 Proceedings*, AAAI Press.

- Allex, C.F., Baldwin, S.F., Shavlik, J.W., and Blattner, F.R. (1997). Increasing consensus accuracy in DNA fragment assemblies by incorporating fluorescent trace representations. *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology*, 3-14. Halkidiki, Greece. Menlo Park, CA. *ISMB-97 Proceedings*, AAAI Press.

- Allex, C.F., Shavlik, J.W., and Blattner, F.R. (1999). Neural network input representations that produce accurate consensus sequences from DNA fragment assemblies. *BioInformatics* 15(9):723-728.

- Allex CF (1999). Computational Methods for Fast and Accurate DNA Fragment

Assembly (Doctoral thesis). Department of Computer Sciences, University of Wisconsin-Madison.

- Altshuler *et al*. (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513-516.
- 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.
- Myers E and Miller W, "Optimal Alignments in Linear Space," CABIOS 4, 1 (1988), 11-17. [EMBOSS Stretcher, PubMed]
- 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]
- 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]