Sequence Assembly and Mutation Detection with Mutation Surveyor®

Kevin LeVan, Jonathan Liu, Softgenetics LLC

Introduction

Within the past few decades, shotgun sequencing has been demonstrated to be a powerful tool in sequencing the genomes of several organisms (1). Due to limitations of only several hundred consecutive bases with adequate resolution, techniques needed to be developed to piece together billions of bases into consecutive order. Next Generation Sequencing technologies make sequencing faster and generate billions of short sequences. The genomes from many organisms including *Drosophila melanogaster, Escherichia coli*, mouse and of course human have been fully sequenced and assembled and are available in multiple databases for use by the public. However, many researchers are not working with sequences that have been assembled and published.

For the genetic analysts working with sequences that do not have genomic references, Mutation Surveyor is a powerful tool capable of taking a group of sequences and assembling them into one long reference file. Just as the shotgun sequencing method when used to sequence the human genome required these short sequences to be assembled, many other projects require the same functionality. Mutation Surveyor is capable of assembling sequences of many formats including *ab1, *abi and *scf. The resulting text file containing a single nucleotide text string can then be utilized directly by the software as the reference file or annotated similarly to a GenBank file for identification of the coding sequence, variations and more.

In addition to its short Sequence Assembly Tool, Mutation Surveyor software is a powerful genetic analysis package capable of comparing samples to reference sequences and easily finding nucleotide and amino acid variations. All types of single nucleotide polymorphisms (SNPs) including homozygous substitutions, heterozygous substitutions with as little as 5-10% contribution from the minor allele and insertions and deletions (indels) can be detected with the click of only a couple buttons. Mutation Surveyor software provides an accuracy >99% in detecting homozygous and heterozygous SNPs (2), when both forward and reverse traces are of Phred 20 quality. In addition, the complex heterozygous indels are deconvoluted from each other yielding two clean traces for easy identification of nucleotide variations.
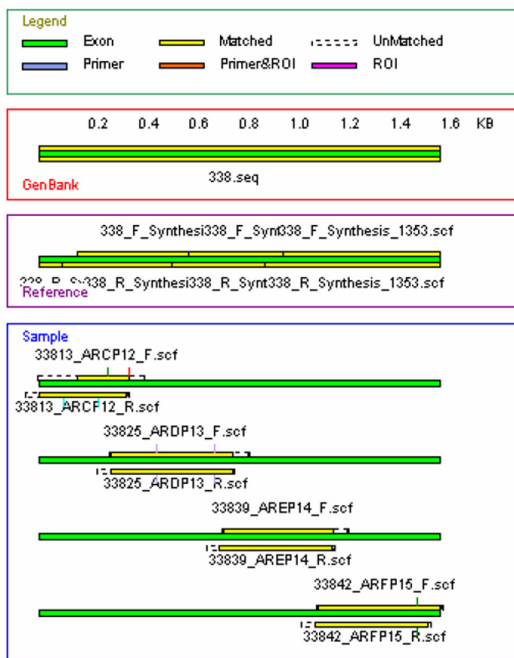

Figure 1: *Samples traces overlap at ends and align. These samples were assembled to create a single reference *seq file shown in GenBank window. Within the Sample window the traces are shown aligned to the reference.*

Procedure

Assemble Sequences

1. From the Mutation Surveyor Tools menu open Sequence Assembler.
2. Load overlapping forward directional sequences to be assembled by selecting the Add button.
3. Choose a file name and location for the assembled sequence and press Save to generate the *seq file with single nucleotide string.
4. Load *seq file into the Seq File Editor or Advanced GBK File Editor for further annotation (optional).

Data Entry (open files window)

1. Add 1- or 2-directional sample traces into the Sample Files field.
2. Add assembled sequence as *seq or *gbk file into the GenBank Sequence File field.

Data Analysis

1. Select OK and press Run.
2. After processing the data, a mutation report is displayed listing all variations found between the samples and references.
3. Choose between several reporting options including tables and graphical reports.
4. Double-clicking on table cells will activate the Graphic Analysis Display where the variations can be reviewed and edited.

Results

Mutation Surveyor software will assemble overlapping sequences into one sequence text string. As shown in **Figure 1** the sample traces used to generate the assembled sequence are aligned properly to the assembled sequence.

Mutation Surveyor software has tools to assist in improving the usefulness of the assembled sequence file including defining the coding sequence, setting primer sites, generating regions of interest, and selecting known variations. In **Figure 2** the assembled sequence file was annotated to include the coding sequences, regions of interest surrounding each exon and known variations. The *seq file can now be saved as a *gbk file and opened in the GenBank Sequence File field when opening data to process.
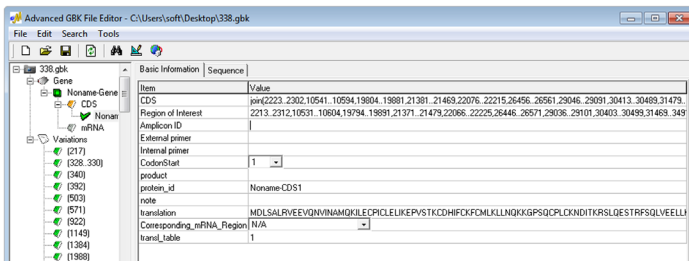
Figure 2: The assembled sequence is opened in the Advanced GBK File Editor, exon starting and ending nucleotides are defined in the CDS section, known variations can be added, and when refreshed, the amino acid sequence is displayed. This can be saved as a *gbk file and used as the GenBank reference.

Mutation Surveyor software aligns sample traces to the assembled reference trace allowing for easy determination of genetic variations. **Figure 3** shows a heterozygous mutation C>CT, the amino acid position 156 is normally Alanine and because of this variation the amino acid can be either an Alanine or a Valine.
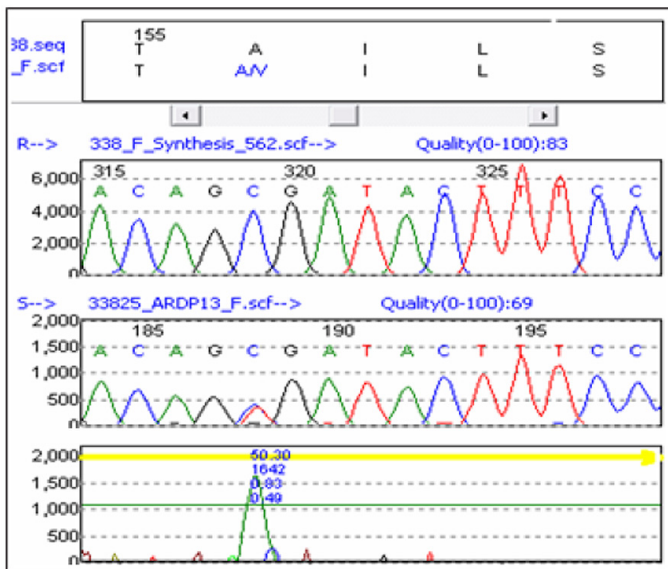

Figure 3: The assembled *seq file was annotated to include the exons, allowing Mutation Surveyor to display the amino acid sequence (at top). The reference sequence (R) shows the correct nucleotide sequence generated by the assembled traces allowing the visualization of the sample trace (S) showing a heterozygous nucleotide change from C>CT. The bottom trace shows the correlation between the reference and sample trace, sharp peaks indicate variations.

Discussion

Due to the need to assemble short sequences into one single nucleotide text string with techniques such as shotgun sequencing, researchers require software to make this possible. SoftGenetics' Mutation Surveyor software is capable of assembling sequences, annotating these reference files similarly to GenBank files, and using these assembled files as references to determine variations within other samples. In addition to assembling nucleotide text strings, a tool is available to generate a trace file from this text string.

Mutation Surveyor is a robust software package for finding nucleotide variations by not relying solely on the trace's base call. Through the use of anti-correlation technology and a unique physical comparison of the migration time for reference and sample traces, higher accuracy of calling is attained. This process also allows for the detection of heterozygous insertions and deletions while eliminating false calls caused by text-based comparison and alignment.

Figure 4: The Advanced 2-directional report shows forward and reverse sequences in adjacent rows with the same background shading, contigs separated by row containing summary information, and mutation information for each trace identifying nucleotide and amino acid positions and calls. Additional color-coding of mutation text identifies confidence; a pink background represents an amino acid change.

There are many reporting options available throughout Mutation Surveyor (Figure 4). Tables can be generated listing all or specific types of detected variations that list details about the traces and each mutation. Various nomenclatures are available to display mutations based on the Human Genome Variation Society (HGVS) standards or genomic numbering. And reports can be generated including pictures of the variations. Color-coding is a useful feature integrated throughout Mutation Surveyor. Font colors distinguish mutations of high and low confidence, confirmed and negative SNPs, while background colors assist in identifying missense mutations, reported variations and more.

Mutation Surveyor is a powerful tool used by thousands of researchers and clinicians across the world. With its ability to assemble traces and identify variations with high sensitivity and accuracy, Mutation Surveyor is the right tool for many sequencing applications.

Notes

Some software packages that are capable of assembling traces are Sequencher™ from Gene Code, Ann Arbor, Michigan; SeqScape® from Applied Biosystems Inc., Foster City, CA; Phred, Phrap and Consed from the University of Washington, Seattle; inSNP, novoSNP, seeSNP, spotSNP, Codon Code Aligner, PolyBayer, and Paracel Agent. Mutation Surveyor is a very robust software package capable of detecting mutations with high sensitivity, separate frame shift indels and also assemble traces into a single sequence.

References

1. J. Malek, et al. 2004. Protein interaction mapping on a functional shotgun sequence of Rickettsia sibirica. Nucleic Acids Research. 32: 1059-1064.
2. Haines, et al. 2005. Complement factor H variant increases the risk of age related macular degeneration. Science. 308: 419-21.

SOFTGENETICS®
Software PowerTools for Genetic Analysis

MUTATION Surveyor®
DNA Variant Analysis Software