NextGENe™ by SoftGenetics

September 2008

SNP and Small Indel Discovery in SOLiD™ System Sequence Reads

Kevin LeVan, Shouyong Ni, Jin Yu, Sean Liu, Jacie Wu and ChangSheng Jonathan Liu

Introduction

Single Nucleotide Polymorphism (SNP) discovery and screening are important for disease discovery and treatment, such as asthma, addictions and cancer **(1)**, as well as association studies **(2)**. Some of the more mature methods for SNP detection include techniques such as Sanger sequencing, the many types of heteroduplex analysis **(3)**, mass spectrometry **(4)** and microarrays **(5)**. The Applied Biosystems SOLiD™ System implores the sequence by ligation technique to generate several gigabases of short sequence reads in a single run, a tremendous increase in output for fast and reliable detection of SNPs. Error rates are higher in comparison to those of Sanger sequencing reads, but the sequence by ligation technique takes advantage of a two base encoding scheme to help identify several of these errors.

Two major challenges for variant detection with short reads, the first is identifying the correct alignment of each short read to the correct location of the genome and the second is distinguishing the true SNPs and Indels from the false positives that are generated at a higher rate by these genome analyzers.

Next*GENe* is a software package designed to import the millions of short reads generated by the SOLiD system and reliably detect SNPs and Indels. Included in this package are tools to remove and clean up low quality reads, a tool designed to condense the short sequence reads by simultaneously correcting errors and lengthening reads, a Sequence Alignment tool designed to align and display the sequence reads, and several additional features for creating and exporting results.

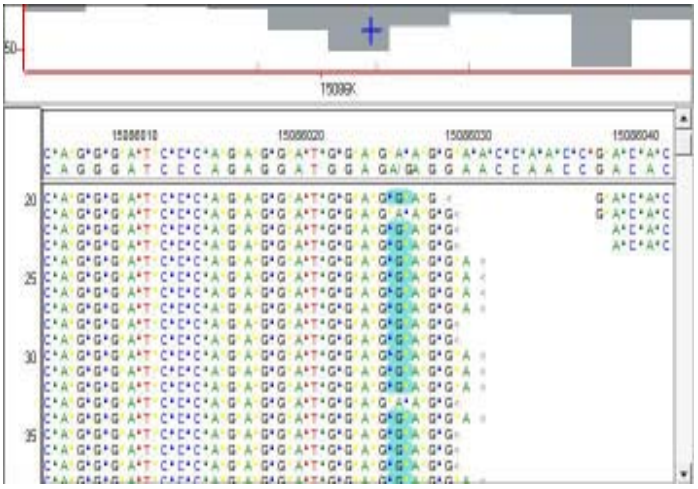

Figure 1: The Sequence Alignment View shows reads in color-space and base-space. Mutation calls are highlighted in blue. A heterozygous substitution was found at position 15086027. The Whole Genome Pane is located at the top of the display – coverage is indicated by gray lines and blue tick marks identify the location of SNPs.

Methodology

The 25-35 bp reads generated the SOLiD system are often not unique within the genome being analyzed. By clustering similar reads containing a unique anchor sequence, the Condensation Tool statistically polishes data of adequate coverage – the short reads are lengthened and reads containing errors are filtered from the analysis or corrected. Once the dataset has been cleaned to remove low quality reads and ends, the remainder of the process, starting with Condensation, is fully automated via use of a Run Wizard that guides you through the project configuration.

Through Condensation, all reads with the same 12 bp anchor sequence are collected into a cluster. The two shoulder sequences on either side of the anchor sequence are used to sort the short reads into multiple groups. The consensus sequence in each group is obtained from the short reads. By using the consensus, random sequencing errors are corrected. The ending bases are removed from the consensus when the base is covered by only one sequence read or there is inconsistency between multiple reads. The 5' sequence has higher weight than that of 3' end because of its higher quality. With 50x coverage within one group, confidence of the condensed sequence is about 99.8%.

The Sequence Alignment tool is designed to match the sequence reads to a user-defined annotated reference sequence. Because the SOLiD instrument produces reads in color-space, NextGENe shows both color-space and base-space. Once the reads have been aligned, SNPs and Indels are highlighted for quick identification. Interactive reports displaying the variations and statistics can be produced and exported.

Procedure

Sample File Preparation

NextGENe can remove low quality reads and trim low quality ends from reads. The first color call represents the color change between the last base of the primer and the first base of the sequence. This position doesn't represent the base call in the genome and is used only for translational purposes between color space and base space; so this position is not used.

1. Choose Format Conversion from NextGENe's Tools menu to remove low quality calls.
    a. Choose the SOLiD SystemCSFASTA (Color) File Format Type.
    b. Add the CSFASTA and QUAL files generated for one SOLiD analysis to the Input field.
    c. Browse to an Output Path location to save resultant CSFASTA file. Filename is automatically appended with "_converted".
    d. Add checkmark to desired settings for removing low quality calls. A suggested start would be to remove reads with a Median Score below 13 and to trim when three consecutive bases are below 10.
    e. Click OK. The Format Conversion window will close when resultant file is created.
2. Choose Sequence Operation from NextGENe's Tools menu to remove first base from each read.
    a. Choose the Sequence Trim Operation Type.
    b. Add the CSFASTA file produced by the Format Conversion Tool to the Input field. If no removal of low quality reads is necessary, load the original CSFASTA file.
    c. Browse to an Output Path location to save resultant CSFASTA file. Filename is automatically appended with "_trimmed".
    d. Add checkmark to the Remove Setting and type 1 in First Bases and type 0 in Last Bases.
    e. Click OK. The Sequence Operation window will close when resultant file is created.

NOTE: The Sequence Trim operation removes bases from the CSFASTA file while leaving the QUAL file unmodified. Therefore, the positional information between the two files is no longer available.

Project Configuration

3. Open NextGENe's Run Wizard through the Process menu.
4. From the Application window, select SOLiD for Instrument Type and *SNP/Indel Discovery* for Application Type. This enables the Condensation and Alignment steps of NextGENe. Click Next.
5. From the Load Data Window, click Load button next to Sample Files field to add the sample file that has low quality bases removed and first base trimmed from each read.
   NOTE: Sample size is limited to 3 million reads or 200 megabytes with a 32-bit Windows® system. Input size increases to 10 million reads with a 64-bit Windows system containing a quad processor and 8GB RAM.
6. Click the Load button next to the Reference Files field to add the reference file.
7. Set an Output Path location to save the assembled project files and click Next.
8. Configure Condensation cycles. Set the number of cycles to 1 and click Set. Default settings are ideal for 100X coverage.
9. Click finish and choose to Run NextGENe.
10. The Running Log shows when the project has completed. The resultant project contains several files, including a statistics file. Results are opened in the Sequence Alignment Tool.

Results

The Condensation Tool clusters the reads with the same anchor sequence, groups them by identical shoulders, and generates a consensus sequence for each group. This process increases the length of the short reads in addition to filtering out the errors with base calling.
The error rates within the individual SOLiD sequence reads increase toward the ends of the reads. Therefore, the software assumes the accuracy at 5' end of reads is more reliable. Reads that are oriented in the forward direction for a particular anchor sequence are more reliable upstream of the anchor (left side), and reads that are reverse complemented for the anchor are more reliable downstream of the anchor (right side). Utilizing this information, the reads in Figure 2 were lengthened from 35bp to 58bp, more than a 1.65 -fold increase. The consensus sequence errors are reduced significantly, far below 0.5%.

Because Single Nucleotide Substitutions occur more consistently for a given position within sequence reads than do color call differences due to instrument error, multiple groups of reads will be created for each of the SNPs and will not get filtered out like reads with errors. By increasing the read length, Indels previously at the ends of reads will be centralized, giving a higher accuracy with mutation calls.

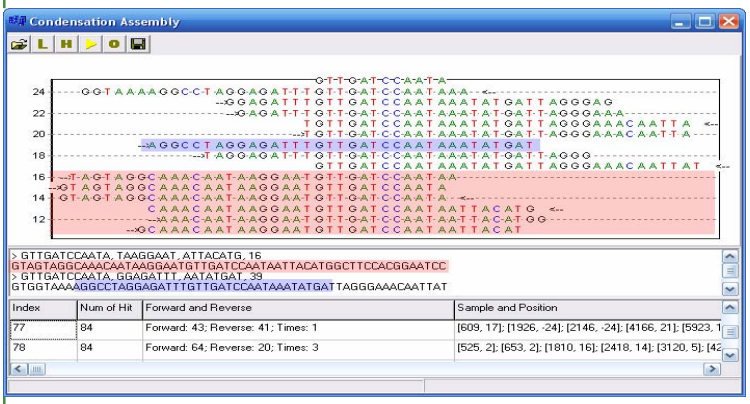

Figure 2: Of the 84 reads containing the anchor GTTGATCCAATA (Index 77), two groups of reads were identified using 55 of these reads. The red highlighted reads are members condensed into the first group. 39 reads contained identical shoulder sequences, allowing for the blue highlighted 35bp read to be condensed with others and generate a single read of 58bp. The other 29 reads contain multiple sequencing errors or match more appropriately to other indexes (not shown).

After the reads were statistically polished – many of the errors have been removed and reads were lengthened – the Sequence Alignment Tool displays the aligned reads, determines allele frequencies and assigns mutations calls as shown in Figure 1. Both color-space and base-space are shown simultaneously, discrepancies between samples and reference are shaded with a gray, blue or purple background for error, unreported mutation call and reported mutation calls respectively.

A Mutation Output report was generated for the run, showing a list of all variations marked as mutation calls. Calls can be manually reviewed, and this report allows for calls to be edited, deleted or added. Options are available for customizing the view of this information, in addition to further filtering. The calls within this report are organized by position within the reference, and each line contains the position within reference, the reference nucleotide, coverage, relative gain/loss for each allele, percentages of reads containing Indels and any additional comments.

Several charts are displayed in the Mutation Output report. The top chart shows the reference nucleotides and their expected percentages, the middle chart shows the percentage of coverage for all nucleotides at each position, and the bottom chart shows the gain/loss of each allele. In Figure 3, a mixture of alleles is shown for the region between 937190 and 937202 bases where several heterozygous substitutions can be viewed, all close to 50% contribution. In addition, NextGENe is displaying a homozygous substitution at base 937206.

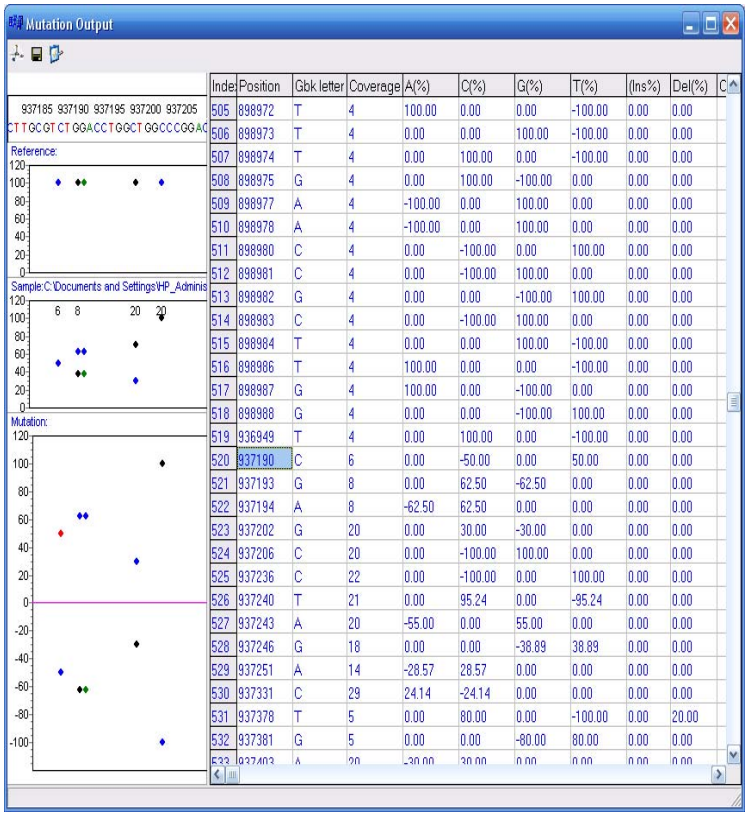

Figure 3: The Mutation Output shows a listing of all mutation calls. On the left is a graphical representation of the selected and adjacent positions. The top chart shows the reference nucleotide and expected percentage, the middle chart shows the percentage of coverage for all nucleotides at each position, and the bottom chart shows the gain/loss of each allele.

Discussion

Similar to SoftGenetics' Mutation Surveyor package for detection of variants in Sanger sequencing reads, NextGENe is a tool for SNP and Indel discovery in sequence reads generated by the SOLiD System. This software package allows for easy and accurate identification of SNPs and micro Indels while reducing the number of false positives due to misaligned reads and instrument error. The Sequence Alignment Tool can accept the original short sequence color-space reads from the instrument in addition to the elongated reads generated by the Condensation Tool, making this an easy tool for validating its functionality. The advantage to the condensed color-space reads is that reads have been trimmed, errors were filtered out, and the sequences have been elongated.

In addition, NextGENe can be used for expression studies including SAGE, microRNA and Transcriptome analyses as well as *de novo* assembly of short reads. This package is capable of assembling 36 bps short reads to contigs that end with repeat sequences.

References

1. K Giacomini et al. 2007. The pharmacogenetics research network: from SNP discovery to clinical drug response. Clinical Pharmacology & Therapeutics. 81: 328-345.
2. P. Ng, S. Henikoff. 2002. Accounting for human polymorphisms predicted to affect protein function. Genome Research. 12: 436-446.
3. H. Tian et al. 2000. Rapid detection of deletion, insertion and substitution mutations via heteroduplex analysis using capillary- and microchip-based electrophoresis. Genome Research. 10: 1403-1413.
4. K. Mohlke et al. 2002. High throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. PNAS. 99: 16928-16933.
5. M. Raitio et al. 2001. Y-chromosomal SNPs in Finno-Ugric-speaking populations analyzed by minisequencing on microarrays. Genome Research. 11: 471-482.

SoftGenetics LLC 200 Innovation Blvd. Suite 235 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com