

SNP and Micro Indel Detection with NextGENe™ Software

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

Introduction

SNP discovery and screening are important for disease discovery and treatment, such as asthma, addictions and cancer (1), as well as association studies (2). SNP detection can be determined by techniques such as Sanger sequencing, the many types of heteroduplex analysis (3), mass spectrometry (4) and microarrays (5). Each technique has its advantages and disadvantages.

With the advent of the next generation sequencing platforms, millions to hundreds of millions of the short sequence reads can be generated at a much higher rate, tremendously increasing the possibility of SNP detection. Illumina® Genome Analyzer utilizing the Solexa sequencing technology uses PCR on a surface, the Applied Biosystem SOLiD™ System uses emulsion PCR and sequencing by ligation, and the Genome Sequencer FLX System from Roche Applied Science utilizes the pyrosequencing technology developed by 454 Life Sciences. Each system has its disadvantages as compared to Sanger sequencing, such as the homopolymers for pyrosequencing and short reads generated by both the bridge amplification sequencing by synthesis and the sequencing by ligation methods. Error rates are higher in comparison, but these errors can statistically be overcome with increased coverage because of the large number of reads each system can generate.

The Condensation Assembly Tool is used to polish and lengthen the short sequence reads into fragment sizes that are more manageable and accurate. The short reads such as those from the Illumina Genome Analyzer System are often not unique within the genome being analyzed. By clustering similar reads containing a unique anchor sequence, data of adequate coverage are condensed, the short reads are lengthened and reads containing errors are filtered from the analysis.

The Alignment tools are designed to match the sequence reads to a user-defined annotated reference sequence. Multiple methods, including BLAST and SoftGenetics' alignment method, are available for aligning the reads to the reference. 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.

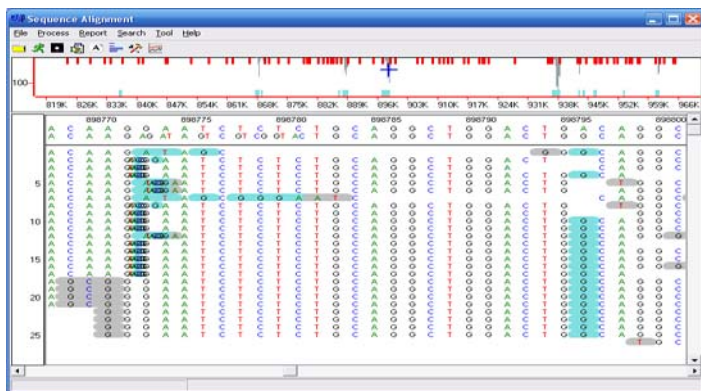


Figure 1: In the region of aligned sequence reads, mutation calls are highlighted in blue. A seven base pair insertion was found before position 898773 and a substitution is present at position 898796. The Whole Genome Pane is located at the top of the display – coverage is indicated by gray lines, red tick marks indicate the breakpoints between genes within reference, blue tick marks identify the location of SNPs.

Procedure

Condense and Lengthen Reads

1. Open the Condensation Assembly Tool and click on the Open Folder button.
 2. Click the Add button and choose the sample file.
 3. Set the Load Sample Section value to the number of reads analyzed simultaneously.
- NOTE:** For optimum speed, import a lower number of reads such as 1,500,000. If the data may contain SNPs at low frequency, then increase the number of reads analyzed simultaneously.
4. Click the Options button and set options accordingly.
 5. Click the Save button.
 6. When condensation is complete, a message will appear showing the start and end time of analysis. Click OK to view results.

NOTE: The Condensation Assembly Tool generates a condensed file that contains the elongated reads and an uncondensed file that contains all reads that were not used for elongation (often the reads containing many errors). Other files may also be created. Depending on the number of reads simultaneously analyzed, the condensed reads may be parsed into multiple files.

Align Reads and Detect SNPs

7. Open the Sequence Alignment Tool and select Load Data.
8. In GBK File field, select Open and choose the annotated sequence file(s) to use as the reference.
9. In Sample File field, select Open and choose the sample file(s) to be analyzed.
10. Choose Settings from the Process drop-down menu and adjust accordingly.
11. Close Settings and click the Run button. When analysis is complete, a message will appear showing the start and end time of analysis.

View alignments, open reports and export results.

Results

The Condensation Assembly 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 sequence reads of 35bps within this Solexa run contain only a 1% error rate in calls for the first 25 bases. Base calls towards the ends show an error rate closer to 5%. 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 Solexa reads than do base 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.

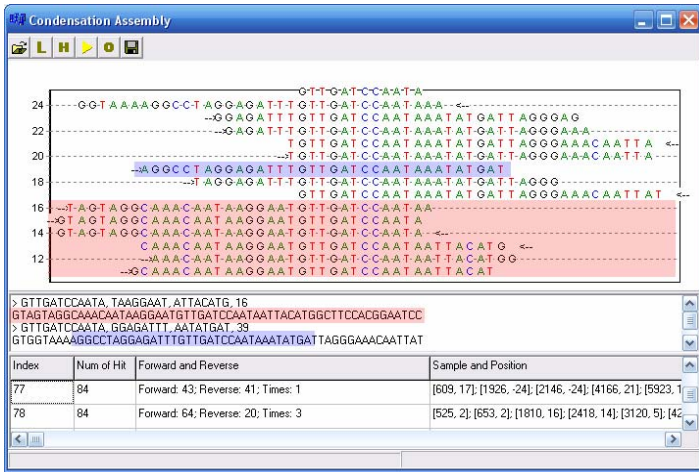


Figure 2: Of the 84 reads containing the anchor GTTGATCCAATA (Index 77), two groups of reads were identified using 55 of these reads. The red highlighted reads are members condensed into the first group. 39 reads contained identical shoulder sequences, allowing for the blue highlighted 35bp read to be 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 was used to align reads, determine coverage and detect mutations. As displayed in Figure 1, the reference that was used contains multiple annotated genes, separated by the red tick marks. The gray bars behind the red marks indicate coverage. Blue tick marks represent the SNPs and Indels that were detected by NextGENe. In the region of aligned sequence reads, an insertion was found before position 898773 and a substitution is present at position 898796.

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.

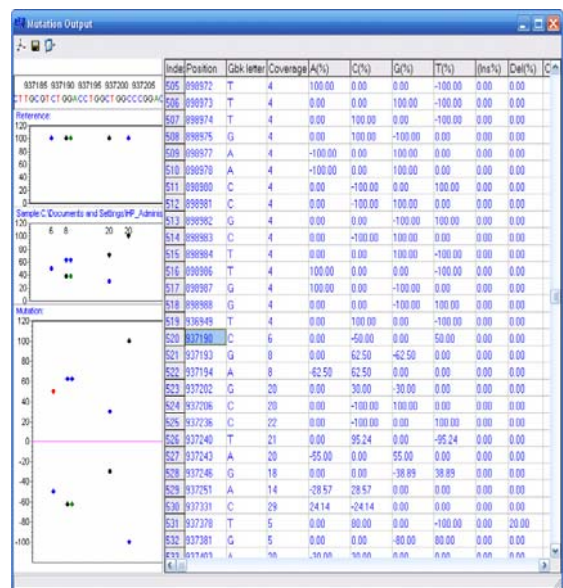


Figure 3: The Mutation Output shows a listing of all mutation calls. On the left is a graphical representation of the selected and adjacent positions. The top chart shows the reference nucleotide and expected percentage, the middle chart shows the percentage of coverage for all nucleotides at each position, 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 analysis of the Next Generation DNA sequencers. NextGENe is a versatile software package designed to support the Illumina Genome Analyzer, the Applied Biosystem SOLiD™ System and the Genome Sequencer FLX System from Roche Applied Science. This software package allows for easy and accurate identification of SNPs and micro Indels while reducing the number of false positives due to instrument error. The Sequence Alignment Tool can accept the original reads from the instrument in addition to the elongated reads from the Condensation Assembly Tool. The advantage to the files of condensed reads is that data has 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 about 1KB fragments 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. *Genome*. 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.