

SNP and Micro Indel Detection in Illumina® Genome Analyzer Reads with NextGENe™ Software

Megan Manion, 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 for SNP detection. Illumina® Genome Analyzer utilizing the Solexa sequencing technology uses PCR on a surface to generate up to 200 million 75 bp reads per sequencing run. The Illumina Genome Analyzer has its advantages and disadvantages as compared to Sanger sequencing. The Illumina system is able to produce more data at a reduced time and cost, however error rates are higher and reads are shorter. Also, the volume of data produced creates a challenge for computational analysis.

SoftGenetics' Condensation Tool is used to polish and lengthen the short sequence reads into fragments of manageable size and improved accuracy. 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 BLAT 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: The top pane of the sequence alignment window shows the Whole Genome View, with gray regions indicating depth of coverage. In the region of aligned sequence reads, mutation calls of known SNPs are highlighted in purple, while novel SNPs are highlighted in blue. A known substitution is observed at position 7753. In the Whole Genome Pane coverage is indicated by gray regions. Purple tick marks indicate locations of known SNPs, blue tick marks identify the location of novel SNPs. The amino acid sequence is shown beneath nucleotide sequence and an amino acid change is shown for position 7753.

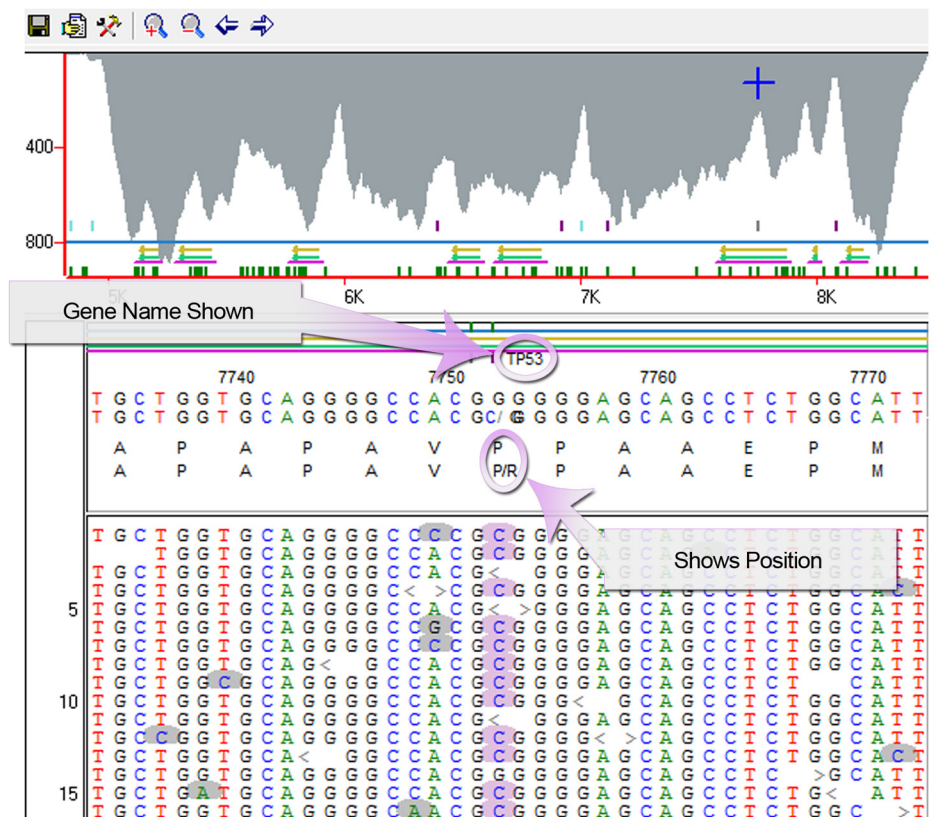



Figure 1

Procedure

1. Open NextGENe's Run Wizard by clicking on the  icon in the main toolbar.
2. Select Instrument Type.
3. Select SNP/Indel Discovery for Application Type.

Note: Sequence Condensation and Sequence Alignment will be selected automatically

4. Click Next to load data.
5. Click "Load" and browse to load sample data
Note: If sample file is not in fasta format, use format conversion tool to convert to fasta.
6. Click "Load" and browse to load reference file in fasta or GBK format.
7. Click "Set" and browse to specify output file location and file name.
8. Click Next to specify settings for Condensation (if applicable) and Alignment.
9. Click Finish.
10. Select Run NextGene to begin processing project.

Once NextGene has completed processing the project, the results will be displayed automatically 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 sequence reads of 35bps within this Illumina 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 Illumina 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: Condensation Results Display shows index table in bottom pane. Clicking on any index in the table displays the consensus sequence for that index above the index table with all the reads in the group shown in the upper pane lined up beneath the anchor sequence.

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.

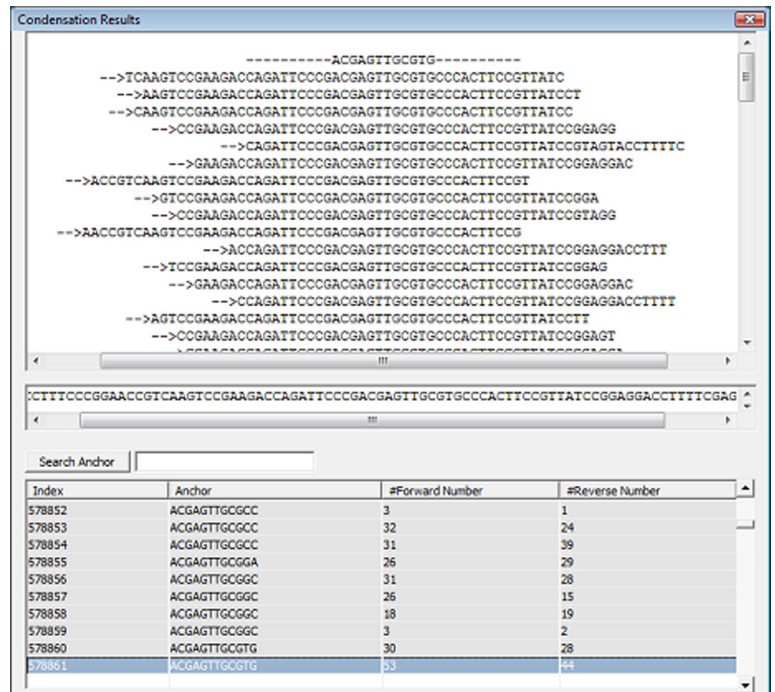


Figure 2

A Mutation 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, percentages for each allele found, percentages of reads containing indels, amino acid changes, gene and/or chromosome location and dbSNP identification. Several charts are displayed in the Mutation Report. The top chart shows the reference nucleotides and their expected percentages, the middle chart shows the percentages 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 observed at position 7753. The reference nucleotide is G while 73.13% of aligned read have a C at the position and the remaining 26.87% of aligned reads have a G at the position. The dbSNP identification is shown since this is a known SNP. The genotype is shown as CG indicating that the mutation is heterozygous.

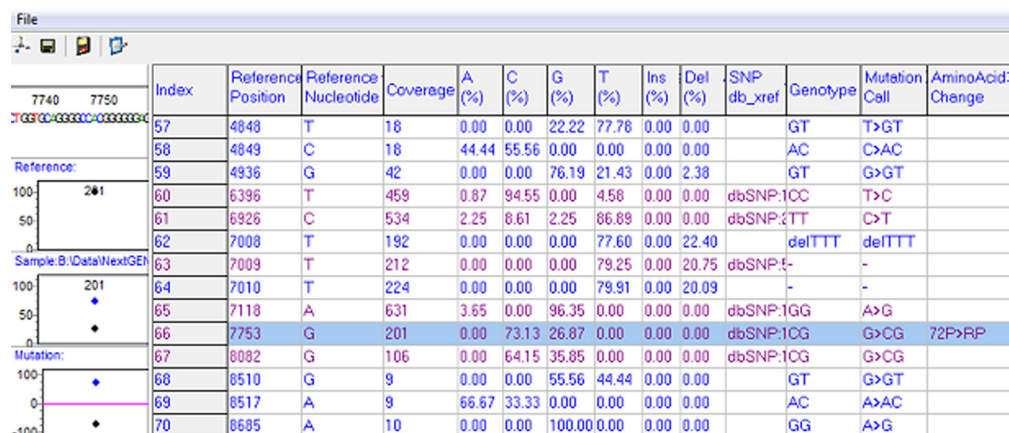


Figure 3

Figure 3: The Mutation Report shows a listing of all mutation calls. On the left is a graphical representation of the selected position. The top chart shows the reference nucleotide and expected percentage, the middle chart shows the percentage of coverage for all nucleotides, and the bottom chart shows the gain/loss of each allele. The Mutation Report provides information for each mutation including reference position, reference nucleotide, and coverage, observed percentages for each nucleotide and for insertions or deletions. dbSNP identification is provided for known SNPs (shown in purple) and double-clicking on the dbSNP ID links directly to the NCBI website. Genotype is included to indicate whether mutations are heterozygous or homozygous. Mutation calls show what type of mutation is observed and amino acid change is also shown for mutations in coding regions.

Discussion

Similar to SoftGenetics' Mutation Surveyor package for detection of variants in Sanger sequencing reads, NextGENe is a tool for analysis of Next Generation DNA sequencer. 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 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.

NextGENe software can be used for analysis of Next-Gen sequencer data for a variety of applications including de novo assembly, ChIP-Seq, Small RNA detection and quantification and Expression studies such as transcriptome studies, SAGE (Serial Analysis of Gene Expression) and DGE (Digital Gene Expression).

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 micro chip-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.