NextGENe™
by SoftGenetics
February 2009

SNP and Small Indel Discovery with SOLiD™ System Sequence Reads Using NextGENe™ Software and Condensation Tool™

Megan Manion, 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 employs the sequencing 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. However, the volume of data produced creates a challenge for computational analysis.

There are two major challenges for variant detection with 2nd generation 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 instrument errors that are generated by massively parallel sequencers. The SOLiD System is able to produce more data at a reduced time and cost, however error rates are higher and reads are shorter as compared with Sanger Sequencing.

SoftGenetics' novel, patent-pending Condensation Tool is used to polish and lengthen the short sequence reads into fragments of improved accuracy and manageable size. Short reads such as those from the SOLiD 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 errors are statistically 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.

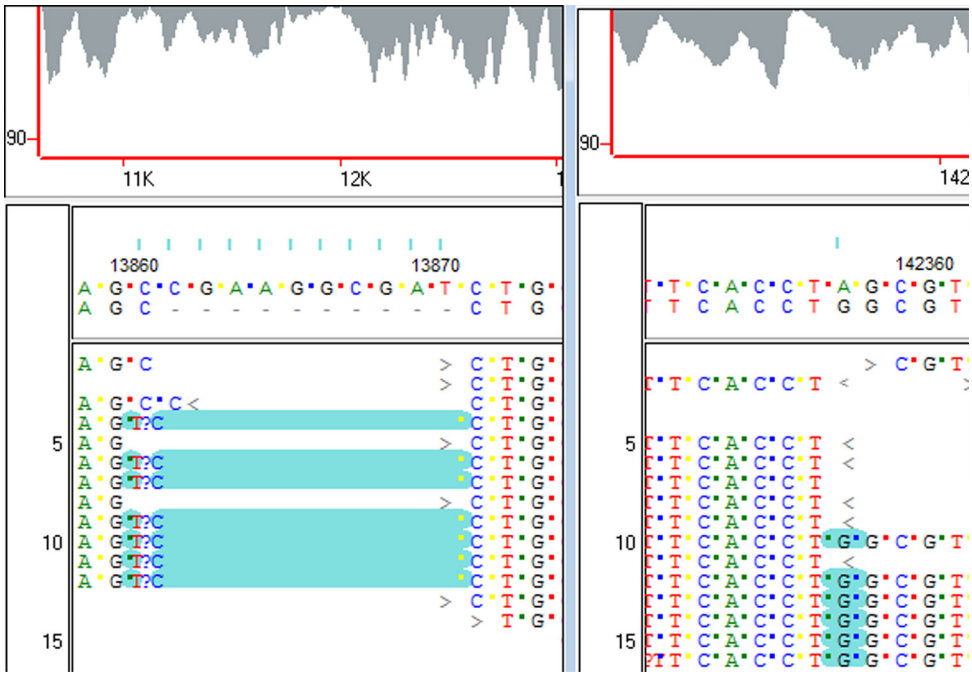


Figure 1: The Sequence Alignment View shows reads in color-space and base-space. Mutation calls are highlighted in blue. NextGENe accurately indentifies substitutions as well as indels. A ten-base deletion of CGAAGGCGAT is called at position 138612 – 13871. A heterozygous substitution was found at position 142348. The Whole Genome Pane is located at the top of the display – coverage is indicated by gray regions and blue tick marks identify the location of SNPs.

Figure 1

Methodology

The short reads generated by 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 Sequence Alignment Tool can be used to accurately align to a reference sequence.

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 clusters into multiple groups. The consensus sequence in each group is obtained from the short reads. By using the consensus, random low-frequency sequencing errors are corrected. 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 System produces reads in color-space, NextGENe shows both color-space and base-space information for aligned reads. 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

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 error rates within the individual SOLiD System sequence reads increases towards the ends of the reads. Therefore, the software assumes that the accuracy at the 5' end of reads is more reliable. 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. NextGENe is able to detect mutations with 99% accuracy. Increasing the read length also allows for the detection of Indels up to 30 bps that are not detectable in short reads alone.

Figure 2

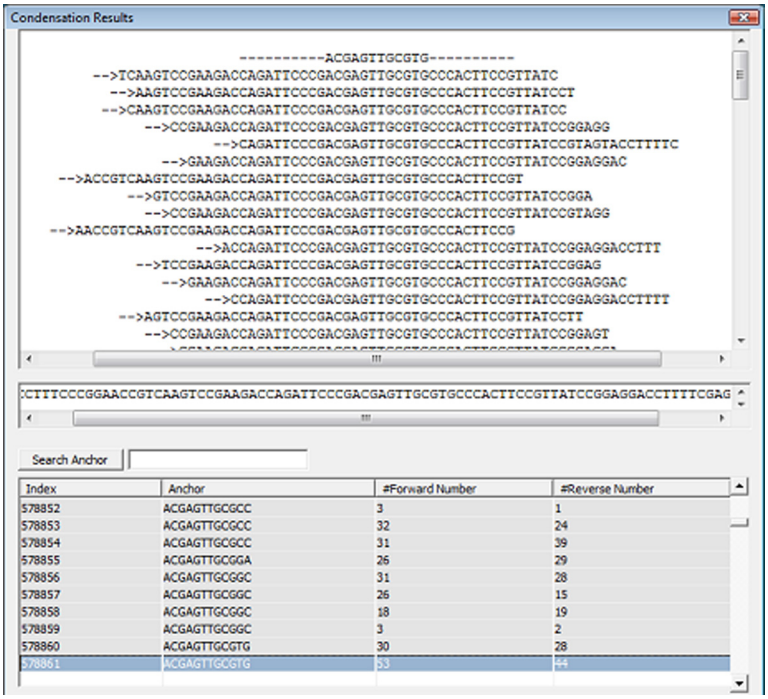


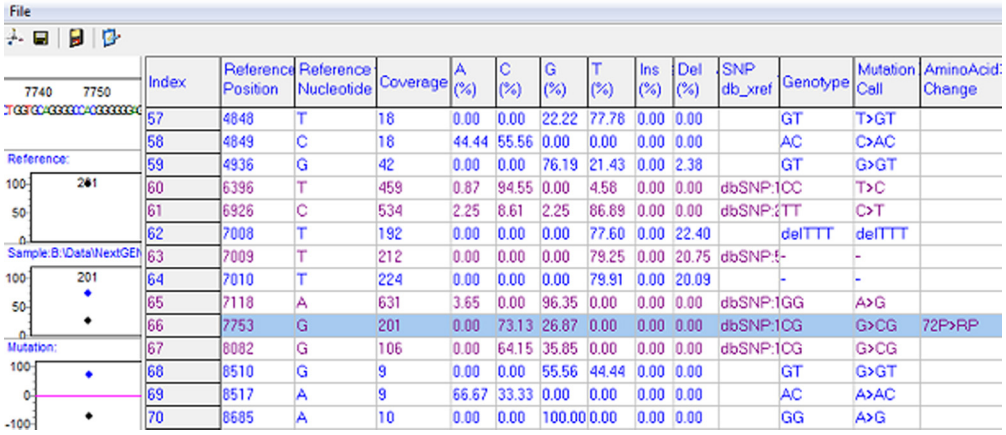
Figure 2: Condensation Results Display shows index table in bottom pane. Clicking on any index in the table displays the consensus sequence for that index above the index table with all the reads in the group shown in the upper pane lined up beneath the anchor sequence.

SoftGenetics LLC 100 Oakwood Ave, Suite 350 State College, PA 16803 USA
Phone: 814/237/9340 Fax 814/237/9343
www.softgenetics.com email: info@softgenetics.com

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 low frequency errors, novel mutation call and reported mutation calls respectively.

A Mutation Report is generated for each alignment project, 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 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 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.



Index	Reference Position	Reference Nucleotide	Coverage	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref	Genotype	Mutation Call	AminoAcid Change
57	4848	T	18	0.00	0.00	22.22	77.78	0.00	0.00		GT	T>GT	
58	4849	C	18	44.44	55.56	0.00	0.00	0.00	0.00		AC	C>AC	
59	4936	G	42	0.00	0.00	76.19	21.43	0.00	2.38		GT	G>GT	
60	6396	T	459	0.87	94.55	0.00	4.58	0.00	0.00	dbSNP:1	CC	T>C	
61	6926	C	534	2.25	8.61	2.25	86.89	0.00	0.00	dbSNP:2	TT	C>T	
62	7008	T	192	0.00	0.00	0.00	77.60	0.00	22.40		delTTT	delTTT	
63	7009	T	212	0.00	0.00	0.00	79.25	0.00	20.75	dbSNP:5	-	-	
64	7010	T	224	0.00	0.00	0.00	79.91	0.00	20.09		-	-	
65	7118	A	631	3.65	0.00	96.35	0.00	0.00	0.00	dbSNP:1	GG	A>G	
66	7753	G	201	0.00	73.13	26.87	0.00	0.00	0.00	dbSNP:1	CG	G>CG	72P>RP
67	8082	G	106	0.00	64.15	35.85	0.00	0.00	0.00	dbSNP:1	CG	G>CG	
68	8510	G	9	0.00	0.00	55.56	44.44	0.00	0.00		GT	G>GT	
69	8517	A	9	66.67	33.33	0.00	0.00	0.00	0.00		AC	A>AC	
70	8685	A	10	0.00	0.00	100.00	0.00	0.00	0.00		GG	A>G	

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 in 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 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.

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.