

SNP and Indel Detection for Pyrosequencing Reads from the Genome Sequencer FLX System with NextGENe Software

Megan Manion, Kevin LeVan, Shouyong Ni, CS 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 (6). The Genome Sequencer FLX System from Roche Applied Science utilizes the pyrosequencing technology (sequencing-by-synthesis) developed by 454 Life Sciences. Error rates for this method are higher when compared to Sanger sequencing methods, particularly homopolymer indel errors. However, the production of 400-600 million high-quality, filter-passed bases per run with the GS FLX Titanium series provides increased coverage which allows software to statistically overcome these errors.

NextGENe's SNP detection application for data from the Genome Sequencer FLX System is able to effectively identify single nucleotide polymorphisms (SNPs) by accurately aligning sample reads with a reference. NextGENe's software is unique in that it allows for a relative amount of base pair mismatch between sample reads and the reference in addition to an absolute number. This allows NextGENe to accurately align reads with long indels and identify them as mutations.

The Alignment tool of NextGENe is designed to match the sequence reads to a user-defined annotated reference sequence. Multiple methods are available for aligning the reads to the reference. Once the reads have been aligned, SNPs and Indels are highlighted for quick identification. The Sequence Alignment Tool also provides information about amino acid changes, exon-intron boundaries, copy numbers, and methylation sites. Interactive reports displaying the variations and statistics can be produced and exported. Variations identified by the software can be linked directly to NCBI dbSNP database.

Procedure

1. Open NextGENe's Run Wizard
2. Select Roche/454 under Instrument Type
3. Select SNP/Indel Discovery under Application Type
 - a. Under Steps, Sequence Alignment is selected automatically
4. Click Next to select Sample and Reference Files
5. If sample file is not in fasta format, click "Format Conversion" button.
 - a. Select Roche 454 file format (SFF) to convert from
6. Load Reference File in fasta or GBK format
7. Specify Output Field
8. Click Next to proceed to Alignment Settings

Alignment Settings

NextGENe uses an alignment method that is similar to BLAT methodology to align samples to a reference. The reference file is first divided into an index table. Every 12 bases of each sequence read is aligned to this table. The positions of alignment between reads and reference are determined; and alignment is evaluated linearly. If they are in a line, the sample sequence can be aligned to the reference at the target positions. There may be some jumps in the line due to true or false positive indels. However, true indels can be distinguished from false positives because true indels will result in alignment jumps in both directions, while false positives result in alignment jumps in only one direction. A uniqueness score is established for each index using $1/\sqrt{N}$, where N is the frequency at which the index appears in the reference. A total uniqueness score can be determined for a read aligning to a location of the reference by summing all of the scores for the indices. When a read does not align exactly with a reference location, due to SNP or Indel, the software will choose the location with the best uniqueness score. This alignment method allows for proper matching of reads with snp/indel errors.

Users can set both an absolute matching criterion (Matching Base Number \geq "X") and a relative matching criterion (Matching Base Percentage \geq "Y") as requirements for alignment. For reads that align perfectly at more than one location, ambiguous alignments can be allowed by selecting "Allow Ambiguous." Alternatively, when this option is not selected, the read will be arbitrarily assigned to the first matching region.

Results

Once NextGENe processes a project, the results are automatically shown in a sequence alignment window. This window offers interactive reports displaying the variations and statistics that are produced as well as exporting options. The display shows the reference sequence, aligned sequence reads, breakpoints between genes, coverage, and other biological information for the position. The Mutation Report shows a list of mutations found, the biological information relative to the segment containing the variant, coverage, directional information, allele ratios, dbSNP identification and amino acid changes. The projects and results can be saved for further analysis.

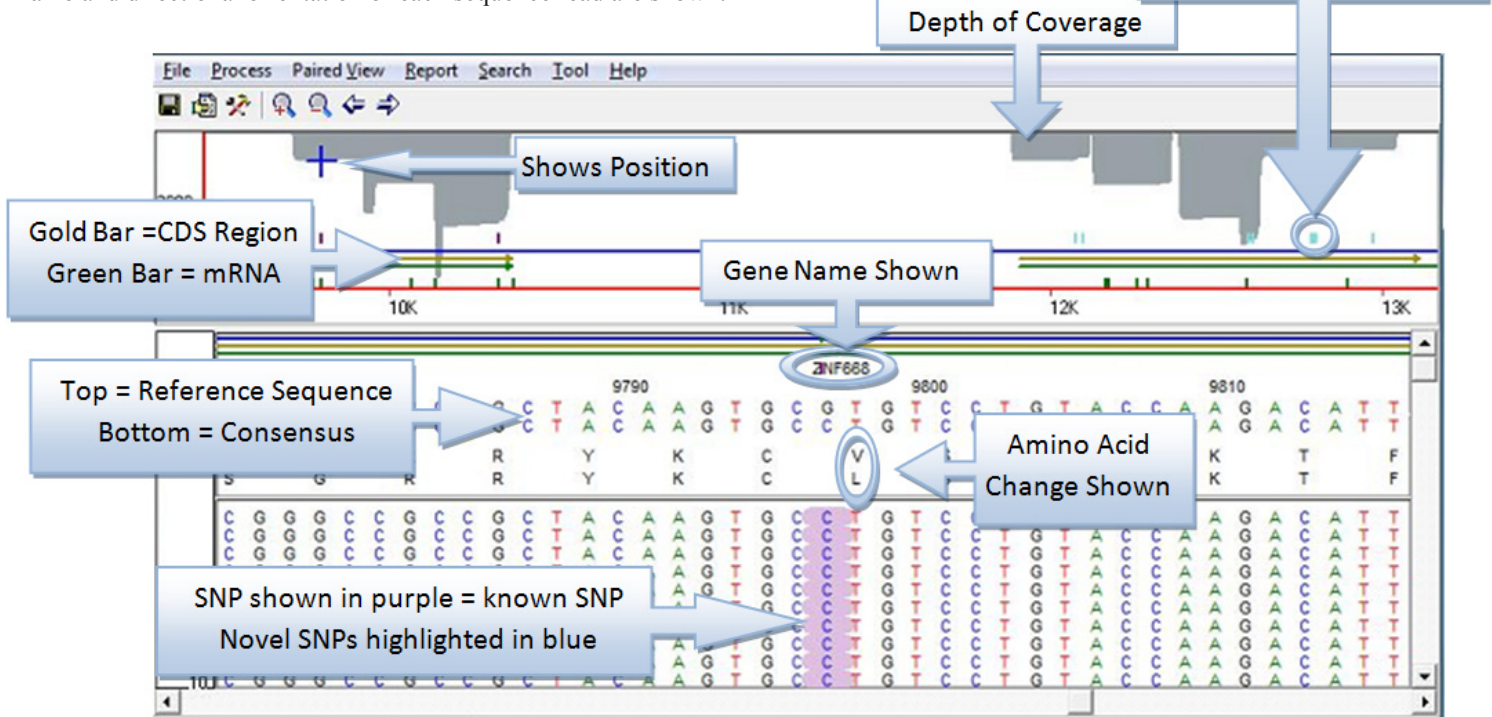
Main View

Whole Genome Viewer

The Whole Genome Viewer is the upper pane within the display that shows a global view of the project. The whole reference is contained within this viewer, in addition to segment breakpoints and the biological annotations for each, coverage, mutation calls, exon-intron boundaries, amino acid sequences, mRNA and CDS region, and current position of reads in the Alignment Viewer. Navigation within the Whole Genome Viewer can be done with a few simple mouse and keyboard hotkeys. The left-mouse button can be used to zoom in or out. To zoom in, draw a box from the upper left corner of the window towards the lower right corner, forming the box around the area to be zoomed in on. Similarly, to zoom out, draw a box from the lower right corner of the window towards the upper left corner. The right-mouse key can be used to scroll horizontally. Also, holding down the Ctrl Key while cursor is in the Whole Genome Viewer will display the comments for each genome, gene, transcript, or other segment, depending on how the reference is organized.

Alignment Viewer

The Alignment Viewer contains a view of all the reads as they align to the reference sequence. Discrepancies between the reference and sample sequences are highlighted – gray positions are variations not reported as mutation calls because they were deleted or fall below a set threshold, often instrumental error, blue positions are novel mutation calls, and purple positions are previously reported variations. Zooming and scrolling can be done similar to the method described in previous section. Holding down the Ctrl key is again used to display information. In this viewer, the name and directional orientation of each sequence read are shown.



Reports

The sequence alignment tool also provides access to several reports that contain detailed information about the possible SNPs and how well the sequence aligns with the reference.

The mutation report shows a list of all locations that have been identified as mutations. For each location, information such as the position, reference nucleotide, percentages for each possible nucleotide as well as insertion and deletion percentages are given. Users can choose which columns of information to have shown in the report. Reviewers can add, delete and confirm mutation calls through this mutation report. The mutation report also shows charts which provide a graphical representation of bases present at positions where mutations are identified.

Index	Reference Position	Reference Nucleotide	Coverage	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref	AminoAcid Change
1	9797	G	1243	0.00	95.82	4.10	0.00	0.00	0.08	dbSNP:2032917	25V>L
2	10330	C	3758	0.32	66.76	0.03	32.73	0.00	0.16	dbSNP:2303223	202G>GG
3	12079	T	1140	0.00	0.44	0.79	98.68	37.98	0.09		FS
4	12101	G	1126	0.18	0.27	99.38	0.18	24.96	0.00		FS
5	12600	A	4515	97.48	0.11	2.30	0.02	23.01	0.09		FS
6	12611	T	4476	0.04	1.16	1.74	97.01	26.36	0.04		FS
7	12792	G	1084	0.18	26.20	73.25	0.00	0.00	0.37		512K>NK
8	12793	C	1084	0.00	74.26	25.74	0.00	0.00	0.00		513P>PA
9	12799	C	1083	0.18	70.54	0.46	0.00	0.00	28.81		FS
10	12804	T	1076	0.09	0.65	0.00	66.45	0.00	32.81		FS
11	12977	G	831	0.24	0.48	99.28	0.00	21.90	0.00		FS

Figure 2: The mutation report is shown which lists all SNPs identified by NextGENe, with known SNPs shown in purple and all novel SNPs shown in blue. Mutations are listed by position in the reference. Users can select how much information is shown in the report. In this example, the reference nucleotide, coverage, percentage of all nucleotides observed, percentages for insertions and deletions, and amino acid changes are provided for all mutations. FS is used to indicate frameshift mutations. For known mutations dbSNP identification is also provided.

Users can determine how information is displayed in the mutation report by clicking on the settings icon (⚙) in the mutation report toolbar. Settings can be adjusted to determine what information is included in the report and how information is displayed.

Filter Mutations

In CDSs Only +/- 0

In ROI Only

Display Setting

Added Automatically

Deleted

Confirmed

Added Manually

Negative

Figure 3

Figure 3: The following settings can be used to determine how mutations are displayed in the mutation report. Mutations can be shown for all locations, in CDS region only (with a set number of flanking base pairs) or in ROI region only. ROI indicates region of interest as determined by user. Refer to “Advanced GBK Editor” section of NextGENe manual for more information on designating a region of interest. Users can also select what types of mutations are listed in the report. For the settings shown below, the report will contain mutations that were identified automatically by the software, those confirmed by the user, those manually added by the user, and negative mutations, but will not include mutations deleted by the user.

Sequence Direction

Non-direction

With-direction


Figure 4

Figure 4: Mutation Report Settings can be selected so that the report will include information about the direction of reads. Including this information can reflect whether mutations are valid SNPs or indels or if they are directional biases.

Figure 5: For the highlighted position in the mutation report shown below, a deletion is observed in forward reads only. This indicates this is a directional bias and probable false positive. Points such as this can be automatically filtered from the results.

Index	Coverage	A (#F,#R)	C (#F,#R)	G (#F,#R)	T (#F,#R)	Ins (#F,#R)	Del (#F,#R)	f %
48	72	6,8	1,0	0,0	0,0	0,0	35,22	1
49	72	1,5	0,0	1,0	0,0	0,0	40,25	8
50	72	1,0	40,31	0,0	0,0	27,10	0,0	1
51	62	11,27	0,0	0,0	3,0	0,0	21,0	6
52	47	0,0	16,10	0,0	13,8	0,0	0,0	0
53	30	0,0	6,4	0,0	9,10	0,0	1,0	0
54	32	8,14	0,0	3,0	0,0	0,0	5,2	6

Figure 5

The Consensus Sequence for SNP locations can be exported from the mutation report by clicking on the  icon. The full consensus sequence can be exported or the consensus sequence for only covered bases or uncovered bases can be exported individually.

Output Consensus Sequence	
Homozygote(0.0%-100.0%)	<input type="text" value="75.000"/> %
IUPAC Heterozygote(0.0%-100.0%)	<input type="text" value="25.000"/> %
Homozygote Indel(20.00%-100%)	<input type="text" value="30.000"/> %
Output SNP Consensus Sequence	
before and after SNP	<input type="text" value="50"/>

Figure 6: Choose the following settings for the consensus sequence. In the consensus sequence, a location will be shown as heterozygous (both alleles shown, separated by a slash, e.g. G/C) when both alleles occur above a user-set heterozygote percentage. In this example, a location will be shown as heterozygous when both alleles appear in at least 25% of reads. Also, the number of base pairs before and after the SNP to be included in the consensus sequence can be set.

The Matched/Unmatched report provides information about which reads aligned to the reference and which did not. Other reports are available that provide statistical information about the alignment. Upon the completion of each alignment project, a file is created with the name *statinfo.txt which details information such as the total number of reads matched to the reference, number of covered bases, average coverage and average length.

Results

Similar to SoftGenetics' Mutation Surveyor package for detection of variants in Sanger sequencing reads, NextGENe is a tool for analysis of data from the Next Generation DNA sequencers. The software is designed to support data from the Genome Sequencer FLX System from Roche Applied Science, with specialized tools for handling various application types. 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.

NextGENe is able to accurately distinguish between false positives and true SNPs or indels by comparing sequence alignment in forward and reverse directions. Since true variations appear in both forward and reverse reads, variations that appear in only reads of one direction can be identified as false positives. The Sequence Alignment Tool's Mutation Report can be set to show allele ratios with direction information so that users can manually delete mutation calls that appear in only one direction. Alternatively, mutation settings can be chosen to allow the software to automatically delete mutation calls that occur in only one direction. This allows for much greater accuracy for identifying SNPs and indels.

NextGENe can also be used with 454 data for expression studies including SAGE, microRNA and Transcriptome analyses, and ChIP-Seq. Previously existing libraries can be used as references for alignment, expression reports can be generated, and peak-finding can be done for studies including ChIP-Seq.

In addition, NextGENe contains the 454 Error Correction tool, designed to specifically handle errors common to the 454 Sequencing method. This tool, which is particularly useful for de novo assembly of short reads, is located in the Sequence Alignment Tool window. For datasets that contain Multiplex Identifiers (MIDs), or barcodes, NextGENe is also capable of parsing results according to each sample-specific MID to identify variations to each sample within a pool.

Acknowledgements

We would like to thank Dr. Phillip Buckhaults from the University of South Carolina for providing data and software requests and Dr. Lan-Szu Chou from ARUP laboratory for his suggestions.

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 microchipbased 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.
6. Brockman W et al. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*. 18:763-770.

Trademarks are property of their respective owners