

Deep Sequencing Analysis and Low Frequency SNP/Mutation Detection with NextGENe Software

Megan Manion, Kevin LeVan, Ni Shouyong and CS Jonathan Liu

Introduction

Next Generation sequencing platforms such as the Genome Sequencer FLX from Roche Applied Science (454 Sequencing), the SOLiD™ System from Applied Biosystems and the Illumina® Genome Analyzer have drastically reduced sequencing costs while producing data at an increased speed and quantity compared to Sanger methods. The large volume of data allows for the sequencing of genomes or genomic regions at very high coverage (5,000x – 20,000x). This high coverage makes it possible to detect low frequency mutations such as somatic mutations and rare variants, such as in virus infected samples. Yet, the high error rates of these technologies (1-3%) create difficulty in distinguishing instrument errors from true low frequency mutations. Because of this, unique software algorithms are required to produce accurate analysis of deep sequencing data.

The identification of low frequency variants is a valuable tool for developing treatments for individuals with various diseases, such as HIV (1) and cancer (2) by recognizing drug-resistant mutations prior to initiating therapies. Drug-resistant disease strains often exist at low abundance that cannot be detected using standard sequencing methods. The presence of these drug-resistant strains can then lead to the failure of drug therapies, so being able to identify these strains at a low threshold is an essential tool. This is also a valuable research tool for improving understanding of the genetics underlying diseases. Deep sequencing techniques can be used to compare case and control samples to detect possible disease-causing variants in pooled samples.

NextGENe software is able to effectively identify rare mutations from deep sequencing data by statistically polishing reads to correct instrument errors, lengthen reads and reduce data volume with its unique Condensation Tool™. After condensation, reads are accurately aligned to a reference sequence and base-calls that differ from the reference are evaluated to identify possible biases. This allows the software to distinguish between instrument errors and true low frequency variants. Replicate control samples (sequencing the same samples in 2 or more channels) can be used to evaluate the accuracy of mutation calls by evaluating the linearity of mutation calls and determining thresholds for false positives and false negatives.

Methodology

Condensation of Reads

NextGENe's unique Condensation Tool is used to correct instrumental errors, lengthen reads and reduce read count by taking advantage of the high coverage available in 2nd generation sequencing platforms. Because Next Generation sequencing systems produce data with varying characteristics that are used for numerous applications, NextGENe's Condensation Tool includes three different methods for reducing systematic errors. The Consolidation and Elongation methods both correct low frequency instrument errors and elongate reads. Elongation is able to maintain original read counts while Consolidation reduces read number by merging identical reads. The Elongation method is recommended for studies where paired reads are utilized such as de novo assembly and the detection of structural rearrangements. For deep sequencing projects, where high coverage is available, the Consolidation method can be used to facilitate computer processing. When Consolidation is used, all information about the original reads, including the number of reads used to produce each consolidated read, is maintained. Error Correction is another Condensation method which is designed to deal with low frequency instrument errors, especially homopolymer errors, for longer Roche/454 reads.

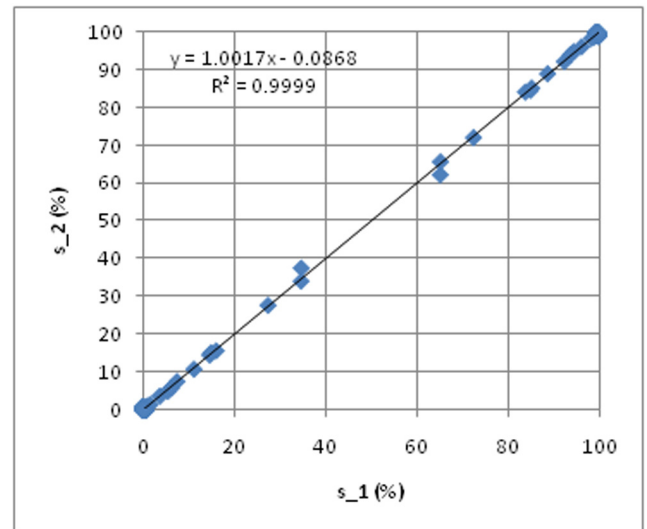


Figure 1: Linearity Plot for all mutations found in the replicate samples s_1 and s_2 from pooled samples of 364 patients, in 2 channels, is shown. Both samples were condensed and aligned to the reference and the mutation percentages using the number of original reads for all mutation calls are plotted. The mutation percentages are linear ($R^2 = 0.9999$) indicating the accuracy of the system and the software. All mutations greater than 1% were found in both samples.

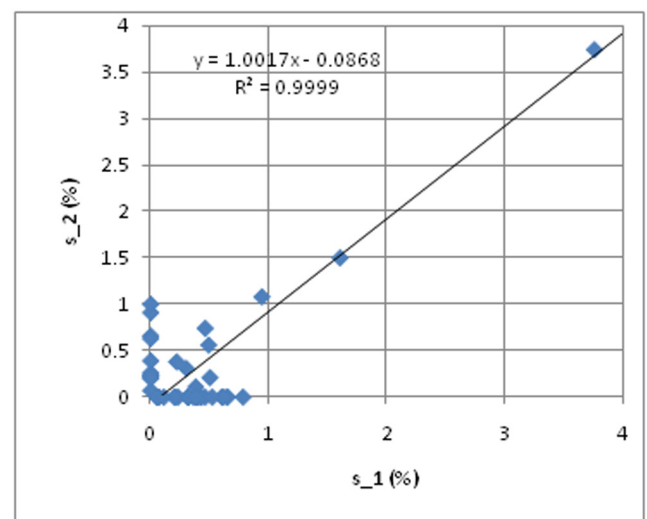


Figure 2: The mutation percentages at extremely low concentrations (0.1-1%) for replicate samples s_1 and s_2 showing false positives and false negatives. False negatives and false positives are observed at around 0.5% concentration; therefore we can consider that the sensitivity for this data is about 0.5%. The confidence interval of 95% is at 1% for SNP detection threshold.

The Condensation Tool works by clustering similar reads containing a unique, 12 bp anchor sequence and two flanking shoulder sequences. All reads containing the exact sequence is clustered together. The cluster of reads can be sorted by homologous shouldering nucleotides both upstream and downstream of the anchor sequence into groups of similar sequence. There may be multiple clusters for any anchor sequence. The consensus of these groups is often close to twice the original length of the reads which improves alignment accuracy and allows for the detection of indels. In creating the consensus sequence within a cluster, the 5' sequence is given a higher weight than that of the 3' end because of the difference in base call quality. The quality of the consensus sequences is significantly improved compared to the original reads allowing for more accurate alignments. The 2% instrument error in base calls can be reduced to 0.1% if the coverage is at least 20x. Reads containing irregular variations such as SNPs and indels are sorted into separate clusters to generate a different consensus sequence. The consensus sequence for a group can then be used in place of all the reads within the group, drastically reducing read count and making alignment much quicker than aligning each individual read.

Low frequency variants that occur in both 5' and 3' ends of reads and in both directions will be maintained because the Condensation Tool compares anchor sequences orthogonally. Each possible 12 bp anchor sequence is considered independently from the major allele. For this reason, each variant can form a cluster as long it is found in enough reads to meet the user-set Condensation settings coverage threshold. For deep sequencing studies with an average coverage of 20,000x, for example, a coverage threshold of 10 recommended. In this case, a variant must occur in at least 10 reads to be detected. Low frequency allele percentages are determined from original read counts, not condensed read counts.

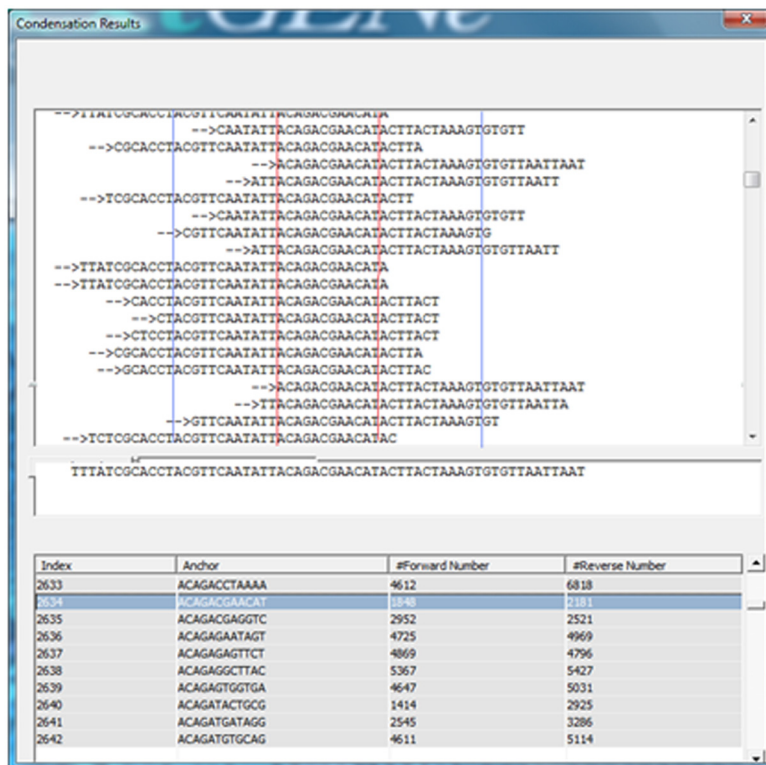


Figure 3a

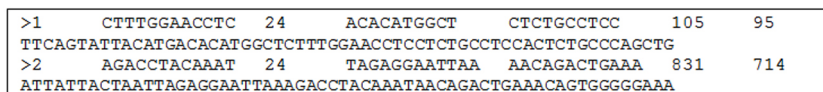


Figure 3b


Figure 3: a) Condensation clusters similar reads according to matching 12 bp anchor sequences and flanking shoulder sequences. It generates a consensus sequence with low frequency errors corrected and increased read length. b) Output consensus sequences are given read names that reflect the anchor sequence, shoulder sequences and counts of forward and reverse reads used.

The read name for each consensus sequence includes, from left to right, index number, anchor sequence, beginning position of anchor in the sequence, left shoulder sequence, right shoulder sequence, number of forward reads and number of reverse reads used to generate the consensus. In the Sequence Alignment Window, holding down the CTRL key over any aligned read shows the read's title information in a yellow box so original coverage information can be accessed easily.

Accurate Alignment and Mutation Calling

Following Condensation, sample reads are aligned to a reference sequence to identify variations. Users can specify both absolute and relative values for mismatch threshold. This improves the accuracy of alignments. For mutation calling, NextGENe uses several criteria to determine whether a variant is a true mutation or the result of sequencing bias. Sequencing errors often occur mainly in one direction, while true variants are found in both forward and reverse reads. NextGENe is able to recognize these variants that occur in one direction more than the other and does not call these mutations.

Procedure

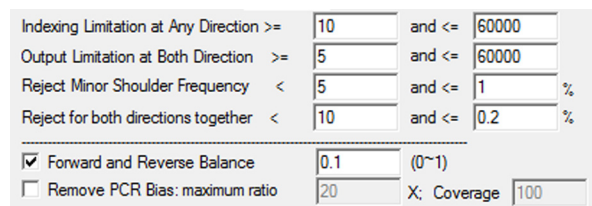
1. Open NextGENE's Run Wizard by clicking on the () icon in the main toolbar.
2. Select Instrument Type.
3. Select SNP/Indel Detection in Application.
4. Sequence Condensation and Sequence Alignment are automatically selected.
5. Click Next to load sample data.
6. Click Load to input sample file(s) in fasta format, or csfasta for SOLiD System data, and reference file(s) in fasta or GBK format.
 - a. If sample files are not in fasta format, use the Format Conversion Tool to convert.
 - b GBK files can be downloaded from the NCBI website.
7. Click Set to browse to appropriate output location.
8. Click Next to continue to Condensation and Alignment Settings.
 - a. In Condensation Settings, choose Consolidation as the Condensation Method.
 - b. Click on "Inspect Input Files" to allow software to determine recommended Condensation Advanced Settings or manually input information about the dataset.
 - c. Additional suggested settings for Condensation and Alignment are shown in Figures 4 and 5.
9. Choose appropriate settings and click Finish, Run NextGENE to begin processing the project.

While NextGENE is processing the project, a Run Log opens that will show the progress of the project. Once processing is complete, the results are shown automatically in the Sequence Alignment Window.

Settings

Condensation

Condensation Settings will vary depending on the dataset. Based on some general information about the sample and reference data, advanced settings are automatically recommended to best suit the data. After sample and reference files are loaded, clicking on "Inspect Input Files" allows software to scan files to automatically fill in this information and update advanced settings. This general information, as well as all advanced settings, can be changed manually as needed.



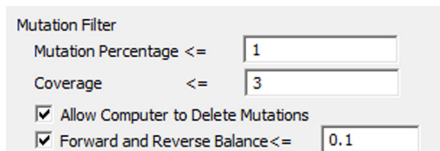
Indexing Limitation at Any Direction >=	10	and <=	60000
Output Limitation at Both Direction >=	5	and <=	60000
Reject Minor Shoulder Frequency <	5	and <=	1 %
Reject for both directions together <	10	and <=	0.2 %
<input checked="" type="checkbox"/> Forward and Reverse Balance	0.1	(0~1)	
<input type="checkbox"/> Remove PCR Bias: maximum ratio	20	X: Coverage	100

Figure 4: Suggested Condensation Settings are shown

These settings require an anchor sequence to occur in at least 10 total reads and at least 5 reads in each direction in order to be indexed. Selecting the "Forward and Reverse Balance" reduces false positives due to PCR errors by checking indexes for the number of forward oriented reads and the number of reverse oriented reads matching to each anchor. Some sequencing artifacts produce significant imbalances between the number of reads in each direction. This option excludes indices from being used in the index table if the ratio of the number of reads in either direction to the total number of reads is below a set threshold. For example, if an index contains 90 forward reads and 10 reverse reads, then the ratio is 0.1. With this option selected and set to 0.2, this index will be removed from the table of indices and no condensed read will be produced for this index. This eliminates interference from primer peaks that include sequences in only one direction.

Alignment

Alignment Settings will also vary according to the sample data, however the settings shown in Figure 5 should be used to detect low frequency mutations and reduce false positives.



Mutation Filter	
Mutation Percentage <=	1
Coverage <=	3
<input checked="" type="checkbox"/> Allow Computer to Delete Mutations	
<input checked="" type="checkbox"/> Forward and Reverse Balance <=	0.1

Figure 5: Suggested Alignment Settings are shown.

Selecting "Allow Computer to Delete Mutations" lets the software delete mutation calls that occur only in unreliable 3' ends of reads or in consensus reads that have low original coverage. When a mutation call is deleted, all information about the variation is maintained but it is not listed as a mutation. A list of deleted mutations can be viewed within the mutation report.

Selecting "Forward and Reverse Balance" requires that variations occur in both forward and reverse reads. If the ratio of forward reads (or reverse reads) that show a variation at the position to the total read count aligned at the position is less than the set threshold, the variation is not called as a mutation.

The Mutation Filter settings determine the requirements for variations to be called as mutations. The Mutation Percentage setting at 1 allows mutations to be called for variants that are found at frequencies as low as 1% of the total reads. The Coverage setting at 3 requires that at least 3 reads are aligned at a position for a mutation to be called.

Results

Aligning condensed reads to the reference solves many of the challenges in locating low frequency mutations. Following condensation, read length is increased from an average of 36 bp to an average of 60bp, allowing for the detection of indels and increasing alignment accuracy. Read count is reduced from 8,068,440 to 42,497 to improve the speed of the analysis. Also, false positives are drastically reduced minimizing low frequency errors that are called as mutations so that mutation calls reflect true variants. 4,183 mutation calls were made after aligning the raw reads which is reduced to just 77 mutation calls after aligning condensed reads.

Data Characteristics	Raw Data	After NextGENe Analysis
Read Length	Short (36bp)	Longer (60 bp)
Error Rate	High (1-2%)	Low (0.1%)
Low Frequency Allele Detection Quality	Poor	Accurate
Data Volume	Large(8M reads)	Manageable (0.04M reads)

Figure 6: NextGENe analysis improves the accuracy of low frequency mutation detection by increasing read length and reducing error rate. Speed of analysis is also increased by reducing data volume.

Reporting

NextGENe offers many reporting options that automatically and clearly describe project results. For SNP/Indel detection, the Mutation Report is available to provide information about each mutation call including the reference position, gene name, location relative to the gene, reference nucleotide, coverage and allele frequencies. Mutation call shows whether the mutation is considered heterozygous or homozygous. Amino Acid changes are shown for mutations within a coding region, and dbSNP identification is shown for known SNPs (shown in purple). The report can also be saved and exported.

Index	Reference Position	Gene	Segment Position	Reference Nucleotide	Coverage	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref	Mutation Cell	AminoAcid Change
1	72219	ABCC9	72219	G	55	45.45	0.00	54.55	0.00	0.00	0.00	dbSNP:616.c.2200G>AG	734V>IY	
2	141602	ACTC1	2296	C	291	0.00	53.95	0.00	46.05	0.00	0.00	c.286C>CT	90H>HY	
3	144579	ACTC1	5273	C	750	0.00	54.53	0.00	45.47	0.00	0.00	c.1092C>CT	364Y>YY	
4	199733	ACTN2	52796	G	467	0.00	50.11	49.89	0.00	0.00	0.00	dbSNP:228.IVS876-8G>CC		
5	220184	ACTN2	73247	C	585	0.00	56.92	0.00	43.08	0.00	0.00	c.2323C>CT	775H>HY	
6	256702	DES	2211	C	1279	0.00	0.00	0.00	100.00	0.00	0.00	dbSNP:105.c.828C>T	276D>D	
7	257059	DES	2569	G	1232	0.00	100.00	0.00	0.00	0.00	0.00	dbSNP:129.c.1014G>C	338L>L	
8	257535	DES	3044	G	1082	100.00	0.00	0.00	0.00	0.00	0.00	dbSNP:105.c.1104G>A	368A>A	

Figure 7: The mutation report provides detailed information about each mutation call.

Comparing different samples (case and control for example) or verifying replicate linearity is simple with the SNP Compare Tool that compares mutation calls made in two or more projects with the same reference. All positions where a mutation call was made in any of the projects is included in the report.

Index	Ref Position	Ref Base	S_1_cond_algn	Coverage	Mutation Call	S_2_cond_algn	Coverage	Mutation Call
1	94	T	15243		T>CT	15230		T>CT
2	417	C	13241		C>CG	13529		C>CG
3	418	A	14273		delAAG	14624		delAAG
4	419	A	14312		-	14846		-
5	420	G	14330		-	14855		-
6	1003	T	4344		T>CT	4573		T>CT
7	1418	C	5388		C>CT	5410		C>CT
8	1861	A	6237		delAAAA	6705		delAAAA
9	1862	A	6478		-	7065		-
10	1863	A	6756		-	7365		-
11	1864	A	6757		-	7361		-
12	1874	G	7212		G>AG	7763		G>AG

Figure 8: The SNP Compare Tool compares the mutation calls made in two different projects. If a mutation call is made in one project but not the other, the "Mutation Call" column is blank for the project that did not show the mutation. Additional columns including chromosomal location, gene location and allele frequencies can be shown when selected. Data shown is a selection of the SNP Compare Report produced for replicate samples and only basic information is displayed to illustrate linearity.

Discussion

NextGENe software is able to accurately detect low frequency mutations in massively parallel deep sequencing data by correcting instrument errors and applying unique mutation calling algorithms to distinguish instrument biases from true variants when aligning sample sequences to a reference. False positives can easily be discriminated from true mutations by comparing the mutant allele ratio and the normal allele ratio. The ratio of the mutant allele is defined as Ratio = # forward reads/total reads with mutant allele. If the two ratios are similar, it is likely to be a true mutation. If they differ by more than 2x, it is likely a false positive with a confidence greater than 95%. This new scoring system will be implemented in the next release of NextGENe software.

NextGENe also includes modules for applications such as *de novo* assembly, ChIP-Seq and transcriptome analysis, SAGE studies and small RNA detection and quantification.

References

1. Simen B B et al. 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes. The Journal of Infectious Diseases. 199: 693-701.
2. Thomas R K et al. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nature Medicine. 12: 852-855.

Trademarks are property of their respective owners.