Targeted Sequence Analysis with NextGENe Software

December 2010

John McGuigan, Megan Manion, Kevin LeVan, CS Jonathan Liu

Introduction

NextGENe Software is able to analyze targeted sequencing data from massively parallel platforms such as the Illumina GA & HiSeq systems, Roche/454 GS FLX, FLX Titanium & Junior, Applied Biosystems' SOLiD System, and the Ion Personal Genome Machine from Ion Torrent. Targeted sequencing using Agilent SureSelect, NimbleGen, Raindance, Fluidigm, or other technologies can all be processed with NextGENe. Sample data can be aligned to target sequences only or aligned to the whole reference genome. When data is aligned to the reference genome, a BED file listing the targeted regions can be loaded to create reports specific to the target regions only.

Analysis results are shown in the NextGENe Viewer which provides detailed visualization not available in other commercial programs like Lasergene's SeqMan, DNASTAR's NGEN & CLC Bio or in open-source tools like Bowtie, MAQ, BWA, or SOAP. The viewer displays aligned results with detected mutations highlighted and full annotation of genes, coding regions, amino acid sequences and reported mutations.

There are several unique properties of targeted sequencing data that must be considered in order to obtain the most useful results. Read directionality, duplicate read removal, specificity, and uniformity. An initial alignment of the data can be useful for assessing these issues for a given sequencing project. The expression report and mutation report can be filtered using a bed file in order to save coverage and variant information for the regions of interest. After the initial alignment, settings can be optimized in order to obtain the best results. The data used in this analysis was provided by Applied Biosystems. It is Agilent SureSelect capture data sequenced on a SOLiD sequencing system.

Procedure

1. Convert the data to FASTA format if it isn't already using the format conversion tool
    a. The reads can also be filtered and trimmed using quality scores
2. Perform an alignment to the whole-genome reference
3. Observe directionality, look for duplicate reads, and determine specificity and uniformity.
4. Create a new alignment to improve the results by making necessary settings adjustments
    a. Use the optional Condensation tool
    b. Remove duplicate reads if needed
    c. Align to gbk files for faster alignment or custom annotation
        i. If capture specificity is low, the minimum matching requirement must be increased

Alignment to the whole-genome reference for the initial alignment is recommended in order to prevent alignment of non-specifically targeted reads from aligning in targeted regions.

Read Directionality

Some targeted sequencing technologies are designed to capture genomic DNA in only one direction in order to maximize the amount of sequence that can be obtained in a single experiment. After an initial alignment the read distribution can be viewed with NextGENe's Distribution Report (figure 1).

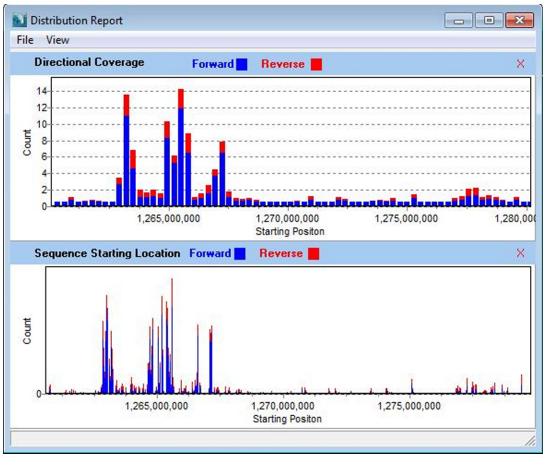


Figure 1: Agilent SureSelect Targeted Sequencing. Most of the reads in this region are aligned in the forward direction.

If directional bias exists, changes must be made to the advanced settings in the optional condensation tool if it is used. As seen in figure 2- the values of the settings highlighted in red should be changed to -1. This will remove requirements that reads are present in both directions and allow one-directional data to condense properly.

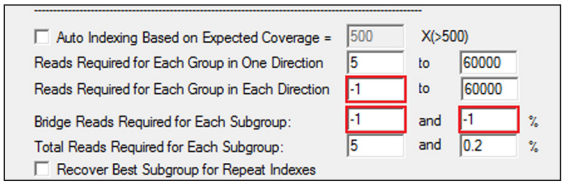


Figure 2: Advanced Condensation Tool Settings

Additionally, when the project is opened in the NextGENe Viewer the Read Balance score and Allele Balance score should be ignored in the total score calculations. This is because they can introduce penalties for variants found in regions where reads are aligned in one direction.

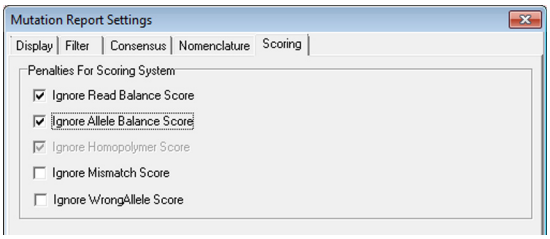


Figure 3: Scoring settings changes

Duplicate Read Removal

All current targeted sequencing technologies use PCR which may introduce biases due to differential amplification of separate alleles. This bias can be more pronounced when a PCR step is used after the captured sequence has been fragmented and sequencing adapters added. The solution is to keep a single copy of reads that have duplicates. This ensures that only one copy of each captured fragment is used to calculate variant allele percentages. If this step is necessary it can be easily and quickly accomplished with NextGENe's sequence operation tool (figure 4).

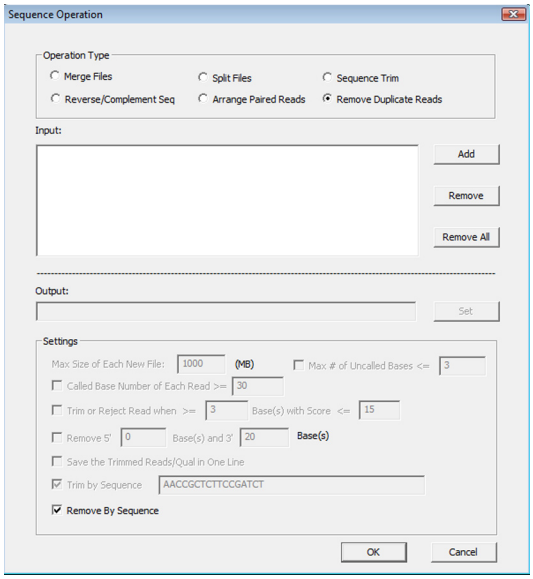


Figure 4: The Sequence Operation tool can be used to remove duplicate reads

Specificity

Current targeted sequencing technologies are not always accurate when selecting sequences. Sometimes sequences similar to the regions of interest are captured inadvertently. It is important to assess how much of the sequence data aligns within the regions of interest. To do this from the NextGENe Viewer:

1. In the NextGENe Viewer open "Expression Report" from the "Reports" menu
2. Select "Input Region of Interest", load a .bed file (figure 5), and click Ok
    a. The BED file should not contain any overlapping regions because reads may be counted twice. If two regions overlap they should be merged into one interval.
3. Sum the number of reads from each segment (shown in the "Read Counts" column) in order to get the total number of reads aligned in the ROI
4. Divide this number by the total number of aligned reads in order to find the specificity.

Uniformity

The coverage of different captured regions can vary by several orders of magnitude. Low uniformity of coverage will result in a wider range of mutation scores. A mutation score of 16 is perfect for 100x coverage, but does not lend much confidence if the coverage is 1000x since the maximum score is 24. Some regions may have little or no coverage, so additional experiments may be required.

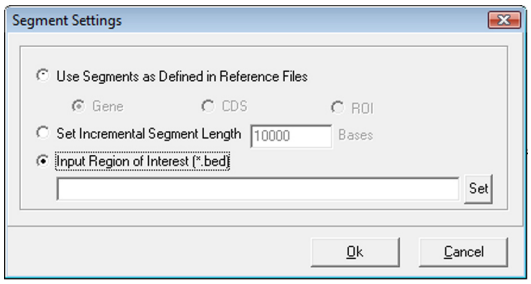
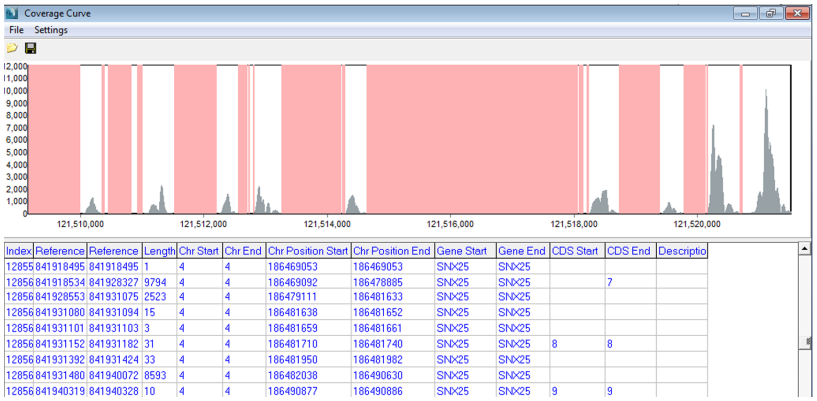


Figure 5: A BED file can be used to filter several of the reports provided by the NextGENe Viewer

Results

There were 72,793,142 total 50 bp reads and 51,839,963 (71.2%) were successfully converted. The initial alignment resulted in 45,960,788 reads (88.7%) being mapped. As seen in figure 1, a directional bias was observed. Duplicate reads did not appear to be a problem with this dataset.

A bed file listing the targeted regions was provided with the sample data and was used to generate an expression report from the initial alignment. 17,001,056 reads (33%) were mapped within the regions of interest and 21,966,350 (47.8%) were mapped within 20 bp of the region of interest. The average coverage within each targeted exon was 40.5 but it ranged from 0 to over 7000. 82.6% of targeted exons had at least 10x coverage and 62.8% had at least 20x coverage. Figure 6 shows the Coverage Curve report, which highlights and lists regions with less than a specified coverage level. This is can be useful for finding missed or low coverage regions of interest.



Index	Reference	Reference	Length	Chr Start	Chr End	Chr Position Start	Chr Position End	Gene Start	Gene End	CDS Start	CDS End	Descriptio
128558	41918495	841918495	1	4	4	186469053	186469053	SNX25	SNX25			
128560	41918534	841928327	9794	4	4	186469092	186478885	SNX25	SNX25		7	
128562	41928553	841931075	2523	4	4	186479111	186481633	SNX25	SNX25			
128566	41931080	841931094	15	4	4	186481638	186481652	SNX25	SNX25			
128568	41931101	841931103	3	4	4	186481659	186481661	SNX25	SNX25			
128570	41931152	841931182	31	4	4	186481710	186481740	SNX25	SNX25	8	8	
128572	41931392	841931424	33	4	4	186481950	186481982	SNX25	SNX25			
128574	41931480	841940072	8593	4	4	186482038	186490630	SNX25	SNX25			
128576	41940319	841940328	10	4	4	186490877	186490886	SNX25	SNX25	9	9	

Figure 6: The Coverage Curve report

The variant report for the regions of interest was saved by selecting "Output the Points of Interest" from the "Mutation Report" menu and inputting the same BED file used in the expression report. 8,786 variants were found in the ROI including a deletion of TAAC in the second exon of the L1TD1 gene on chromosome 1 (figure 7).

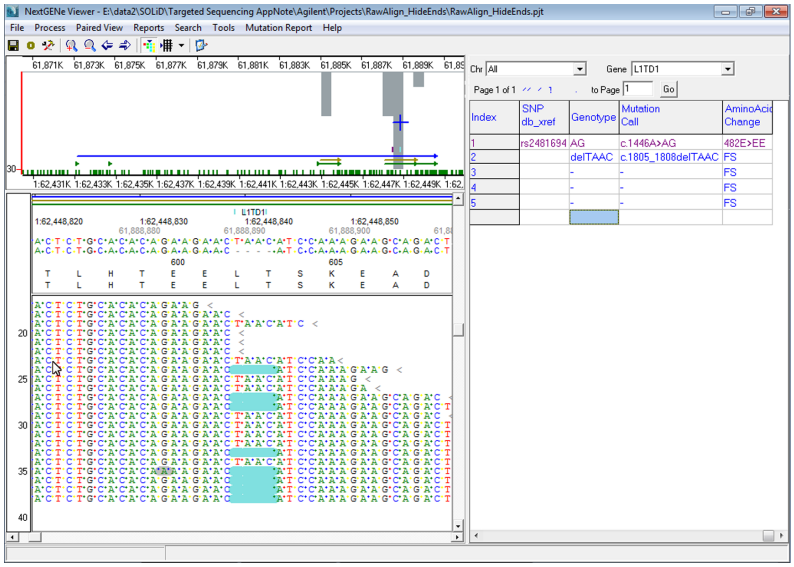


Figure 7: A 4-base deletion found in the L1TD1 gene

Discussion

Further analysis could be performed on this dataset. The Condensation tool can be used to increase read length, correct sequencing errors, and reduce the total read count as long as the settings are adjusted (as described above) to account for the directional bias. The data could also be aligned to a collection of gbk files instead of the whole genome reference as long as the minimum matching percentage was increased. This would help prevent nonspecifically captured sequence from aligning within the targeted exons.

There are several additional options if paired-end data is used. For libraries with overlapping paired reads, the Paired-End Merge tool can combine overlapping pairs from the fastq files before format conversion. This results in highly accurate long single reads, increasing the accuracy of alignment. If the gap size is relatively small it may be possible to elongated the reads with the condensation tool so that they overlap enough to be merged.

NextGENe is a powerful tool for variant analysis using sequence data produced by massively parallel genome analyzers. It has several tools that are useful for processing targeted sequence data including the ability to filter reports with BED files and remove duplicate reads with the sequence operation tool. NextGENe is a versatile software package designed for the new era of high through-put genome sequencing, supporting Illumina GA & HiSeq systems, Roche/454 GS FLX, FLX Titanium & Junior and Applied Biosystems' SOLiD System. Some of NextGENe's advantages include accurate results for indel and SNP detection with whole genome alignment and an easy-to-use graphical user interface.

NextGENe includes software applications for a variety of application types including expression studies like Digital Gene Expression, RNA-Seq analysis, microRNA studies and SAGE, as well as de novo assembly, SNP and indel detection, and ChIP-Seq.