Processing Ion AmpliSeq™ Data using NextGENe® Software v2.3.0

McGuigan J.R., Manion M.R., Liu C.S.
SoftGenetics LLC, 100 Oakwood Ave Suite 350, State College, PA 16803, USA

Abstract

The Ion AmpliSeq™ Panels use highly multiplexed PCR in order to generate thousands of amplicons for targeted sequencing. These amplicons are sequenced on the Ion PGM™, allowing for rapid turnaround and low cost. NextGENe® software provides an easy-to-use and completely customized analysis of the results, including detection of novel variants or alleles, alleles found at lower frequencies (less than 5%), or alleles in regions with lower coverage. NextGENe includes many useful features, such as quality control reports, functional prediction scoring, and advanced project comparison. Recently released NextGENe v2.3.0 incorporates improved quality filtering technology (keeping more of the original data) and enhanced alignment algorithms to provide higher accuracy, and lower false positives in homopolymer regions.

Methods

Datasets can be processed in minutes on a desktop computer running a 64-bit Windows operating system. All steps are performed with an easy-to-use point-and-click interface with no scripting required. Three datasets were processed in this analysis:

Sample	Panel	Chip
C05-401	Cancer Panel	314 (Barcoded)
B26-204	Inherited Disease Panel	316
FLO-528	Comprehensive Cancer Panel	318

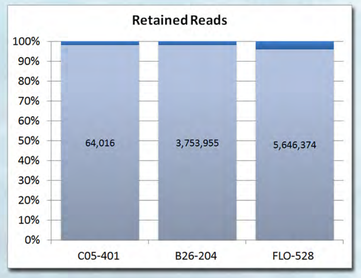
Processing includes:

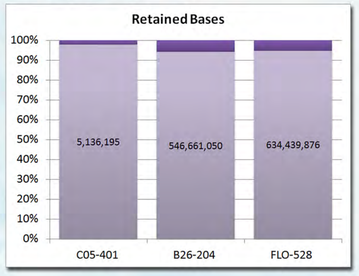
1. Format Conversion and Quality Filtering
2. Alignment and Mutation Calling
3. Quality Assessment

After calling mutations, some functional prediction information is available from the dbNSFP 1.x database (1). This includes PolyPhen-2, SIFT, MutationTaster, LRT, and PhyloP in addition to 1000 genomes frequencies. The Sanger COSMIC (Catalog of Somatic Mutations in Cancer) database (2) is also available, with COSMIC IDs reported for coding and noncoding variants in the database.

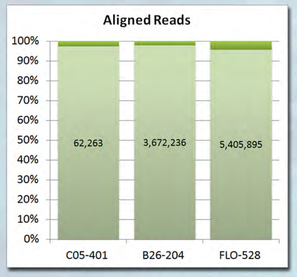
Results

Format Conversion and Quality Filtering



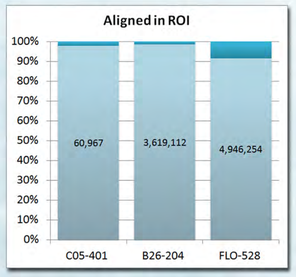


Alignment Results



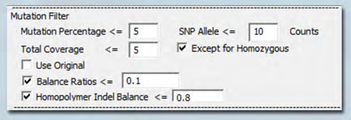
Processing Time	
C05-401	8 min, 26 sec
B26-204	41 min, 40 sec
FLO-528	50 min, 20 sec

Avg Coverage in ROI	
C05-401	216.3
B26-204	339.41
FLO-528	319.32



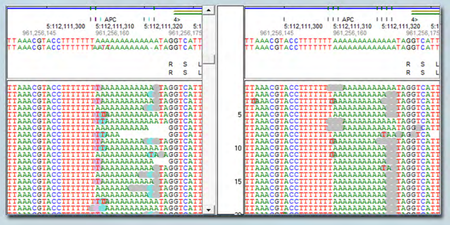
Mutation Calling Results

Filtering Settings

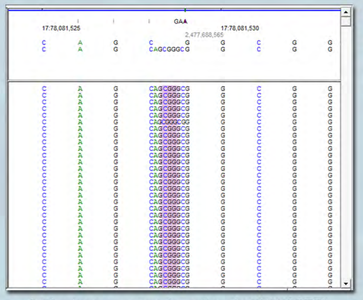


	C05-401	B26-204	FLO-528
Substitutions	7	893	1627
Known (dbSNP)	5	880	1431
Also Called by Ion Torrent	7	885	1416
Indels	11	605	1214

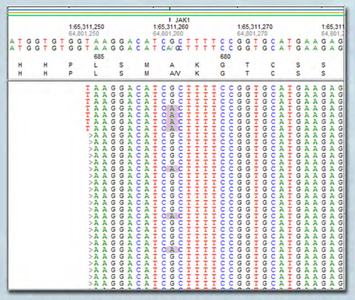
The mutation report was also filtered to exclude any mutations outside the targeted regions.



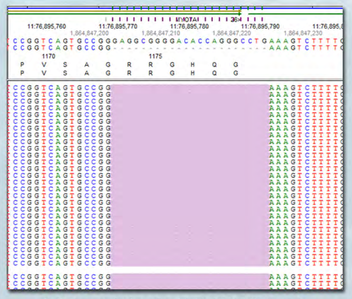
An alignment comparison between version 2.2.1 (left) and 2.3.0 (right) in an area with conse=cutive long homopolymers. The improved alignment algorithms reduce false positives in regions like this.



A 7bp homozygous insertion (rs35373675) found in the GAA gene in the Inherited Disease Panel sample (B26-204)



A low frequency (5.71%) variant (rs2230587) found in the JAK1 gene in the Comprehensive Cancer Panel sample (FLO-528)



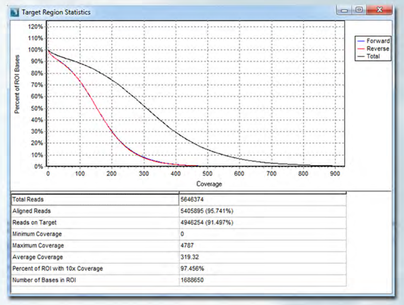
A 22bp deletion (rs111033223) found in the MYOA7 gene in the Inherited Disease Panel sample (B26-204)

Discussion

NextGENe's whole genome alignment algorithm has three steps- Match perfect reads, match reads with some number of mismatches, and finally a seeded alignment. Shorter amplicons will benefit from the second step because there may not be enough bases on either side of a mutation to align perfectly matching seeds. Longer seed sizes will improve alignment specificity, and fewer seeds will improve speed.

When choosing alignment settings, it is important to consider the expected results. Increasing the minimum depth of coverage will reduce the number of false positives (even at lower mutation frequencies), but it may also decrease sensitivity. Setting a minimum number of mutant allele reads will allow for detection of low frequency variants in high coverage regions without allowing low frequency false positives to be called in low coverage regions.

All homopolymer indels can be ignored if the forward/reverse balance filter is set to 1. In NextGENe v2.3.1 the balance filters also include frequency cutoffs, so that the filter isn't applied to high-frequency variants.



The coverage curve report summary for the Comprehensive Cancer Panel project. This report is very useful for measuring potential loss of sensitivity and for visualizing coverage of the targeted regions.

Conclusion

NextGENe provides and easy-to-use, fast, and accurate method for analyzing the result of Ion Torrent AmpliSeq panels. The latest release version (2.3.0) has several improvements to make analysis even easier and more accurate.

References

[1] Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation 32, 894-899 (2011).
[2] Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 39, D945-D950 (2010).

NextGENe®
2nd Generation Sequence Analysis Software

SOFTGENETICS®
Software PowerTools for Genetic Analysis
www.softgenetics.com