

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

July 2012

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

Introduction

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. Figure 1 shows the detection of a large deletion.

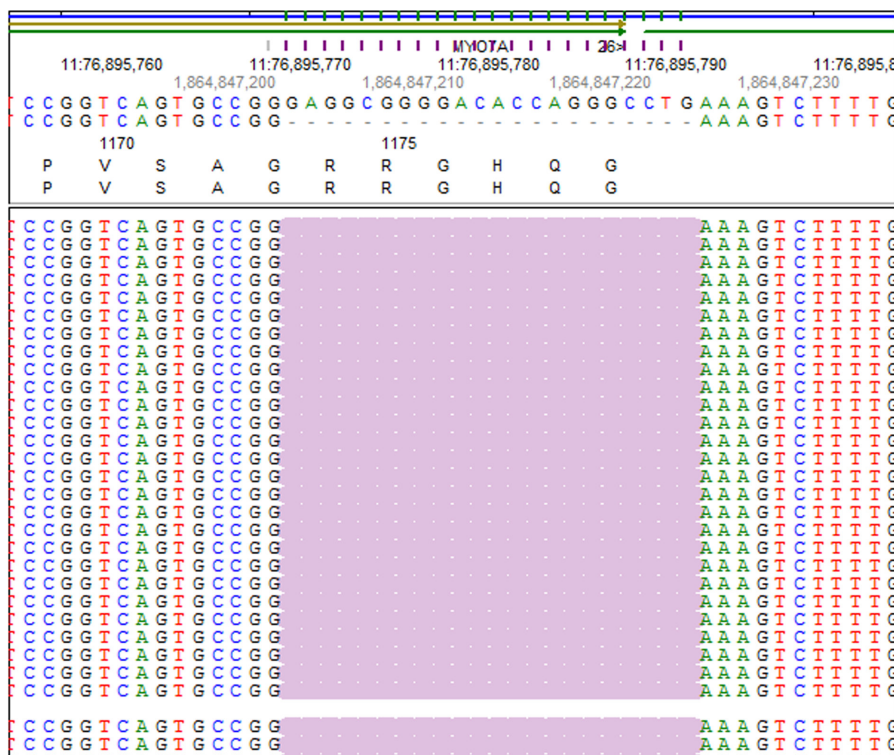


Figure 1: A 22bp deletion (rs111033223) found in the MYO7A gene in an Inherited Disease Panel sample

Procedure

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:

C05-401: Cancer Panel sample that was barcoded and run on a 314 chip

B26-204: Inherited Disease Panel sample run on a 316 chip

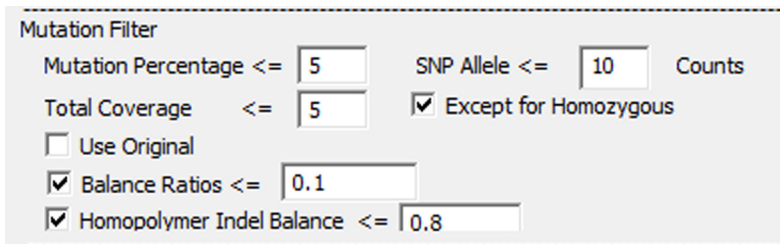
FLO-528: Comprehensive Cancer Panel sample run on a 318 chip

Format Conversion

- As with all NextGENe projects, the raw data (FASTQ or SFF in this case) is first converted to FASTA format
- Basecall quality scores are used to filter and trim the raw data
- Default settings were used in this analysis- reads are trimmed at the 3' end if two or more consecutive bases (not counting homopolymers) are found with a score ≤ 12 . Reads are rejected if they are less than 25 bp after any trimming.

Alignment to the human genome

- Alignment is performed against a pre-index human genome reference in order to avoid false positives caused by nonspecific amplification of untargeted regions
- First reads are mapped if they match the reference with no mismatches. Then the remaining reads are aligned using a seeded alignment algorithm. The “Detect Large Indels” option performs a secondary alignment step to improve indel detection.
- Mismatches are called as variants in the mutation report if they pass the mutation filter. Mutation filter settings will vary depending on the desired sensitivity and specificity. Figure 2 shows the settings used for the Comprehensive Cancer Panel project in this analysis.
- By default two filters are enabled- variants with balance ratios < 0.1 and small homopolymer indels with F/R balance < 0.8 are removed from the final report in order to reduce the number of false positives. By default, the Balance Ratios filter requires that at least 10% of reads with the mutant allele occur in each direction and that the proportion of reads in each direction is at least 10% of the proportion for the reference allele. Setting the Homopolymer Indel Balance filter to 1 will remove all small homopolymer indels.



Mutation Filter

Mutation Percentage <= 5 SNP Allele <= 10 Counts

Total Coverage <= 5 Except for Homozygous

Use Original

Balance Ratios <= 0.1

Homopolymer Indel Balance <= 0.8

Figure 2: Mutation filter settings used in this analysis

Project review

- The mutation report, expression report, and coverage curve reports are filtered by loading a BED file specifying amplicon or targeted loci regions. The latter two reports provide information about coverage in the regions of interest. The mutation report has many additional filtering options.

Results

Table 1 lists the results of format conversion. Over 95% of the original reads and over 94% of the original bases were kept for each sample. Table 2 lists alignment results- around 95% of reads were successfully aligned, and about 95% of those reads were aligned in the amplicon regions.

	C05-401	B26-204	FLO-528
Total Reads	65,143	3,820,981	5,885,326
% Kept Reads	98.27%	98.25%	95.94%
Total Bases	5,248,517	580,620,343	669,753,923
% Kept Bases	97.86%	94.15%	94.73%

Table 1: Format Conversion Results

	C05-401	B26-204	FLO-528
Processing Time	8 min, 26 sec	41 min, 40 sec	50 min, 20 sec
Aligned Reads	62,263	3,672,236	5,405,895
% Aligned	97.26%	97.82%	95.74%
Reads on Target	60,967	3,619,112	4,946,254
% on Target	97.92%	98.55%	91.50%
Average Coverage in Amplicons	216.30	339.41	319.32

Table 2: Alignment Results - datasets were run on an 8-core laptop with 8 GB of RAM

Mutation Calling results are summarized in table 3. The results were very concordant with mutation calls provided by Ion Torrent using the Torrent Suite software. More strict filtering could improve the specificity even further, especially for small homopolymer indels.

	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

Table 3: Mutation Calling Results. A multiple-base deletion is counted as multiple deletions when totaling the number of Indels.

Figures 3, 4, and 5 show examples of a mutation found in each project. The Comprehensive Cancer Panel project was an 80:20 mixture of two 1000 genomes samples (NA12878 and NA19240). The mutation in figure 5 is heterozygous in NA19240 and not present in the other sample, so the frequency is expected to be close to 10%. It was detected in 5.71% of the reads at that position.

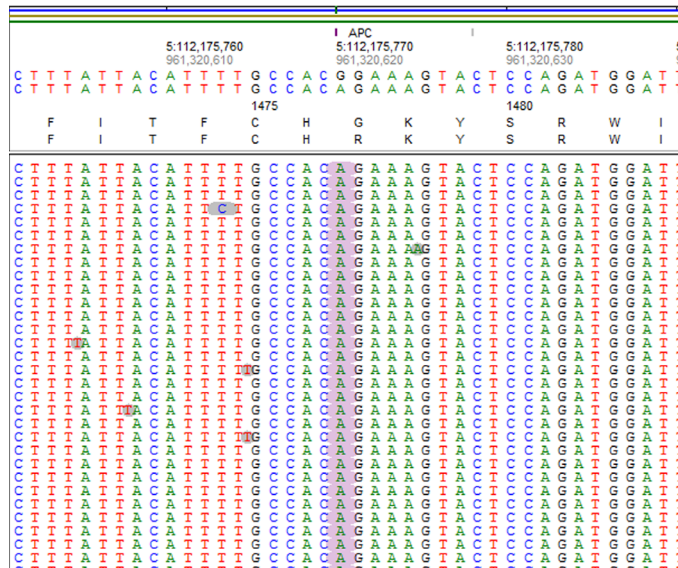


Figure 3: A homozygous mutation (rs41115) found in the APC gene in the C05-401 sample.

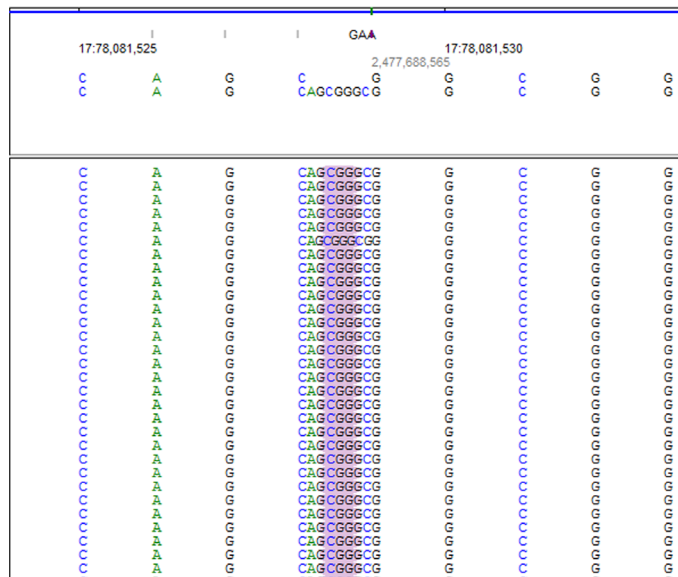


Figure 4: A 7bp homozygous insertion (rs35373675) found in the GAA gene in the Inherited Disease Panel sample (B26-204)

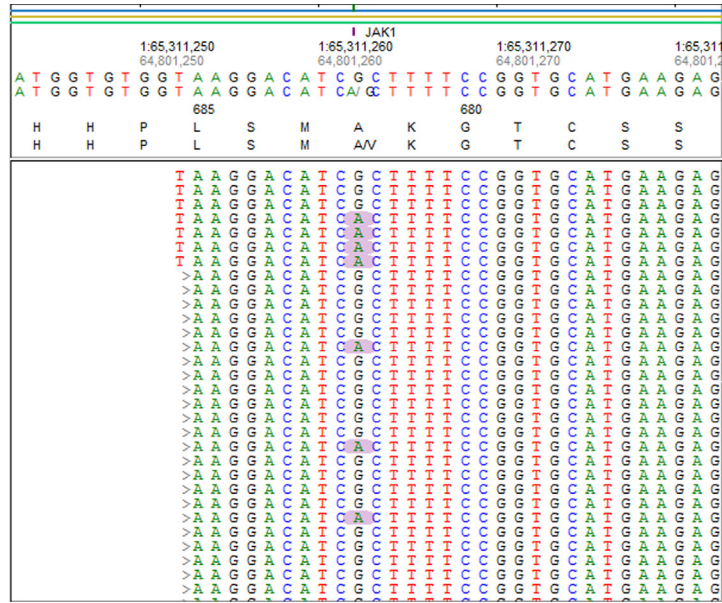


Figure 5: A low frequency (5.71%) variant (rs2230587) found in the JAK1 gene in the Comprehensive Cancer Panel sample (FLO-528)

Discussion

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. The coverage curve report is very useful for measuring potential loss of sensitivity (figure 6) and the coverage curve report summary (figure 7) is useful for visualizing coverage of the targeted regions.

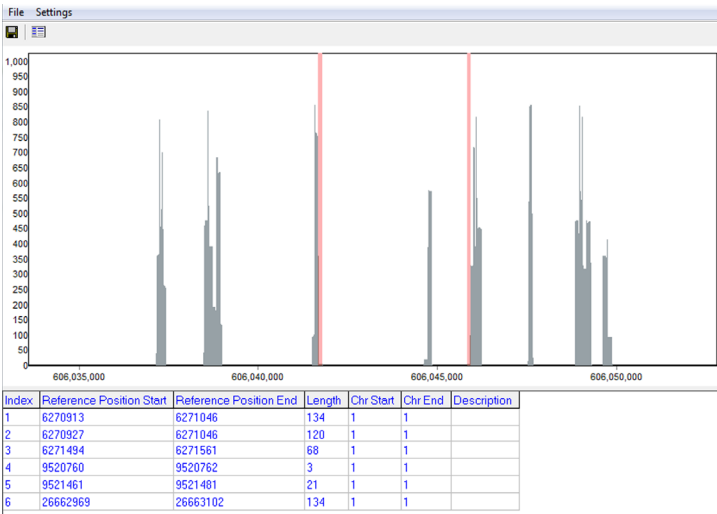


Figure 6: The coverage curve report for the Comprehensive Cancer Panel project

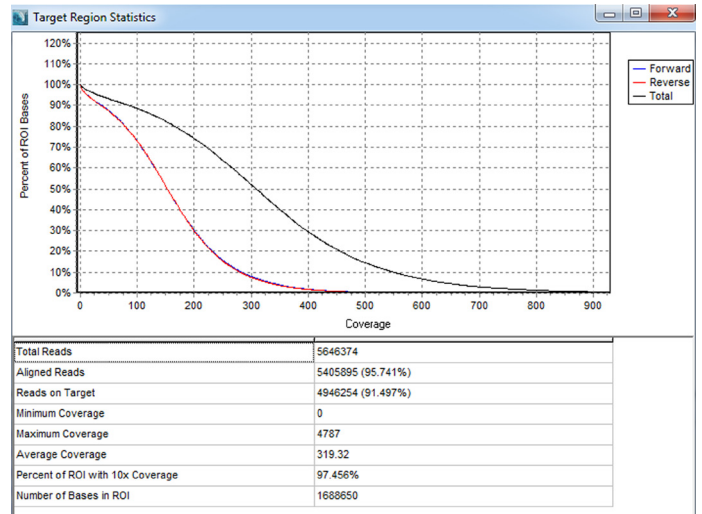


Figure 7: The coverage curve report summary for the Comprehensive Cancer Panel project

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. After alignment several projects can be compared and filtered against one another. Compound heterozygous, shared/different, low coverage, and Mendelian inheritance filtering are all possible (figure 8).

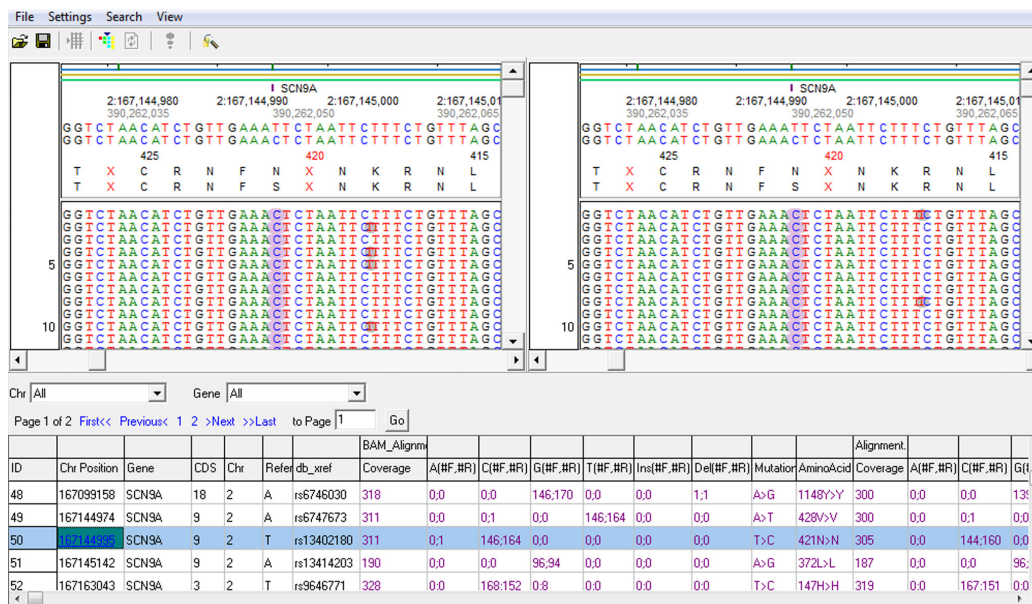


Figure 8: A comparison between the TMAP alignment (imported from a BAM file using NextGENe’s BAM Reader tool) and the NextGENe alignment for the Inherited Disease Panel sample

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. Both databases can be imported using the “Track Manager” tool in NextGENe and then queried from the Viewer.

Acknowledgements

We would like to thank Life Technologies for supplying the AmpliSeq data used in this analysis.

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).