Processing CFTR Amplicon Data from the Ion PGM™ using NextGENe® Software

October 2011

John McGuigan, Megan Manion, Shouyong Ni, Sean Liu, C.S. Jonathan Liu

Introduction

Amplicon sequencing is a quick and highly accurate method used to screen for known mutations as well as novel variants. This is often used for genes known to be associated with diseases, such as BRCA1/BRCA2 and CFTR. Current next-generation sequencers do have some drawbacks. They require a large initial investment in the sequencing hardware, which can be difficult for smaller labs. A single run can take several days, making it too slow for many clinical applications. Semiconductor sequencing- developed by Ion Torrent, a part of Life Technologies- is a novel sequencing method that is better suited for this type of analysis. Instead of an optical system measuring fluorescence it uses millions of pH meters to monitor hydrogen ions released during DNA replication. The Ion Torrent PGM™ offers a quick and relatively inexpensive sequencing workflow while NextGENe® software provides quick, accurate, and easy-to-use analysis of PGM™ Sequencer data.

NextGENe makes analysis of amplicon data very fast and very easy. Using a simple point-and-click interface, a typical amplicon dataset from the Ion PGM can be analyzed in just a few minutes. The default alignment settings are widely applicable, but with some adjustments and filtering, it is often possible to increase sensitivity and specificity. Alignment can be performed against FASTA files listing the sequence of the amplicons, or against a GenBank file of the gene or genes of interest. The GenBank file will allow annotation to be displayed in the viewer (figure 1).

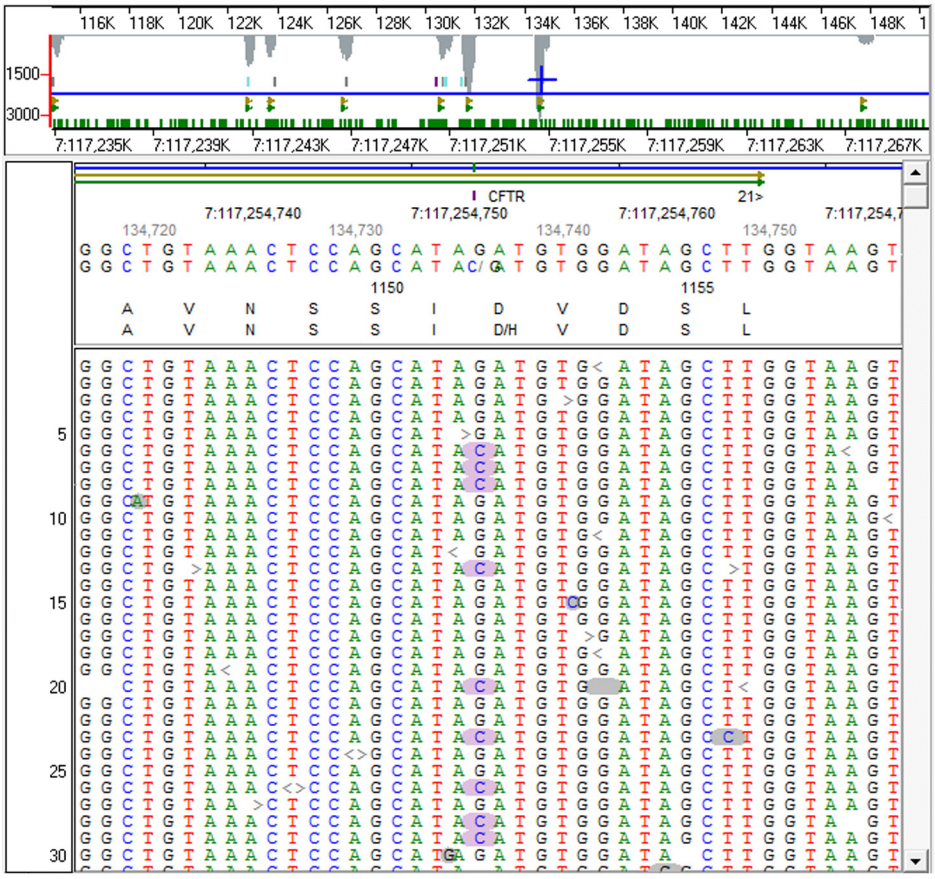


Figure 1 - A known mutation (rs75541969) found in CFTR exon 21 in sample GM14323 (41.84% C>CG)

Mutation calling can be limited to certain regions of the GenBank files by specifying regions of interest in the file itself or by loading a BED file.

Procedure

1. Format conversion - FASTQ or SFF files are converted to FASTA format while performing some quality filtering and trimming.
2. Barcode Sorting (if needed) is performed using the barcode sorting tool
3. Alignment is performed. Using the default settings provides great sensitivity, but adjustment of some alignment settings can improve detection of indels and low frequency variants. The number of false positives can also be reduced.
   a. "Detect Large Indels" and "Rigorous Alignment"
      i. These options add extra processing steps to improve alignment accuracy, especially for indels. The extra processing time is insignificant for smaller projects, such as runs from an Ion PGM™.
   b. "Delete Small Homopolymer Indels with F/R Balance <= 0.25"
      i. This option will remove indel calls in homopolymer regions if the ratio of forward and reverse reads with the indel is less than 0.25.
      ii. Increasing the ratio threshold may remove more false positives but it may also remove some real mutations.
      iii. If homopolymer indels are not being considered for the final report, this option can be set to 1.0 in order to remove all small homopolymer indels.
   c. Mutation filter settings
      i. Adjusting the minimum coverage, mutant allele frequency, and mutant allele count can be very effective at limiting false positives.
      ii. Higher coverage samples should use an increased coverage filter.
      iii. Lowering these settings significantly will introduce false positives due to sequencing error.
      iv. Disabling the "except homozygous" option will ignore mutations below the coverage filter even if they are at 100% frequency.
4. The project is examined using the NextGENe Viewer
   a. The mutation scores can be used to filter out potential false positives. There is a coverage score and several penalty scores which are multiplied together to get the overall score.
   b. The mutations can be filtered based on the overall score or any of the sub-scores.
   c. The Variant Comparison Tool allows several projects to be visualized side-by-side.

Results

Five sets of CFTR amplicon data (4 samples and one control) were analyzed using NextGENe software. On average each sample had 328,000 reads, and 90% were kept after quality filtering and trimming. The samples were aligned to a FASTA file reference of amplicon sequences and the CFTR GenBank file from NCBI (NC_000007.13). Each alignment took less than one minute to run on a typical laptop computer. After an alignment with default settings, some parameters were adjusted to improve the specificity without lowering the sensitivity. Some score filtering was applied to the second set of alignments in order to improve the results even further. Results are summarized in table 1.

		Called Mutations (validated)		
Sample	# of Validated Mutations	Default Settings	Adjusted settings	After Filtering
GM07552	23	33 (20)	32 (22)	21 (21)
GM12785	12	28 (12)	24 (12)	12 (12)
GM12960	2	17 (2)	15 (2)	2 (2)
GM13423	2	18 (2)	13 (2)	2 (2)
NA12878	0	19 (0)	14 (0)	0 (0)

Table 1 – Mutation Calling Results

Discussion

The adjusted alignment settings were:

- Enable 'Rigorous Alignment' and 'Detect Large Indels'
- Enable 'Delete Small Homopolymer Indels if F/R Balance <= 0.25"
- Adjust the mutation filter settings to 20x coverage, 15% frequency, and 5x SNP count

The adjusted alignment settings slightly improved sensitivity in some samples. This can be attributed to better indel detection provided by the "rigorous alignment" and "detect large indels". The specificity was slightly improved in the second set of alignments, usually due to the removal of false positive indels by the "Delete small homopolymer indels with F/R < 0.25" option.

The filtering settings were:

- Hompolymer score <= 0.9
- Total score <= 5.0

The homopolymer score penalizes indels occurring in homopolymer regions. After filtering, 4 of the 5 samples had full sensitivity and specificity- the only detected mutations were validated by sanger sequencing. The 5th sample had 100% specificity, but missed two validated mutations. One of these mutations was a deletion in a long homopolymer, making it difficult to distinguish from error. This mutation is called when the score filtering is not used. The remaining mutation was a 2 bp deletion. On manual examination it was found to be present at a frequency (4.6%) below the cutoff value (15%).

Figure 2 shows a comparison of all 5 samples. At this position a 2 bp deletion (blue highlight) was found in the first two samples. In the global view at the top of the screen it is possible to see the depth of coverage of several amplicons in each sample separated by red lines.

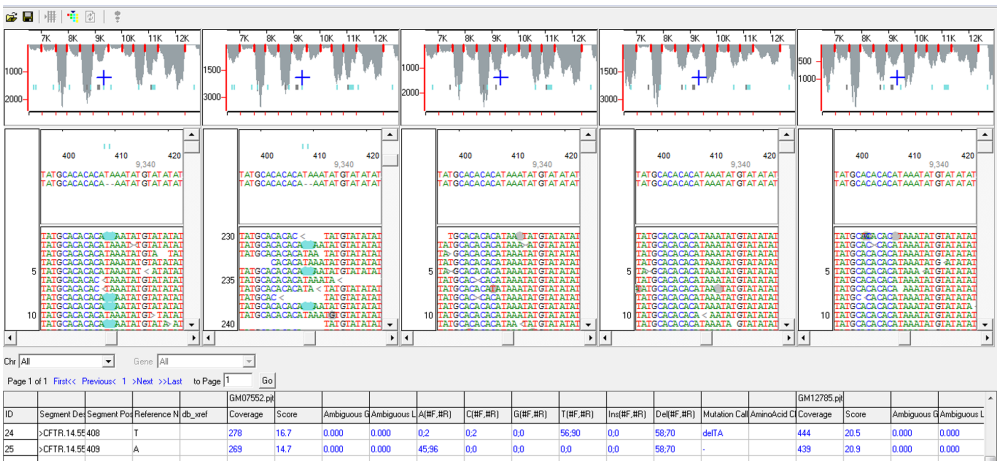
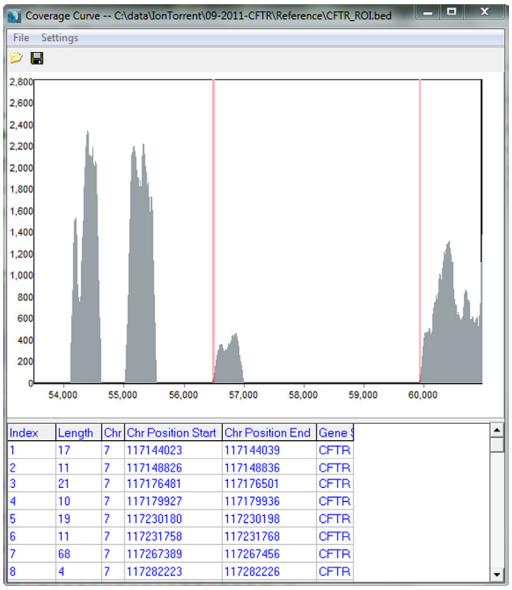


Figure 2 - Comparison of all 5 projects

The coverage curve report can be used to highlight regions below a specified level of coverage (figure 3). This is very useful for detecting failed amplicons or regions with a lack of sequence coverage. When used with gbk files, a BED file can be loaded to specify the regions of interest.



Index	Length	Chr	Chr Position Start	Chr Position End	Gene
1	17	7	117144023	117144039	CFTR
2	11	7	117148826	117148836	CFTR
3	21	7	117176481	117176501	CFTR
4	10	7	117179927	117179936	CFTR
5	19	7	117230180	117230198	CFTR
6	11	7	117231758	117231768	CFTR
7	68	7	117267389	117267456	CFTR
8	4	7	117282223	117282226	CFTR

Figure 3 - Coverage Curve Report

NextGENe's module for Ion PGM amplicon sequencing analysis outputs quick and accurate results in just a few seconds. In addition to its accuracy and speed, NextGENe also has a variety of tools that can be used for quality control. Once the sequence data is aligned, the Variant Comparison tool can be used to compare samples against one another for discovery of novel variants or to confirm known variations. All of NextGENe's reports are highly flexible and can be exported as tab-delimited text files.

Acknowledgements

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