Technical Brief: *Processing Ion Torrent data with RTG Investigator*

October 13th, 2011

Introduction

The Ion Personal Genome Machine (PGM)™ sequencer from Life Technologies Corporation offers a fast, cost-effective alternative for genomic research and diagnostic applications. Based on semiconductor sequencing, the Ion Torrent platform promises the fastest time to results through a unique combination of simplicity, scalability and speed.

The Ion Torrent platform with the latest 316 chip generates over 100Mb of data with read lengths beyond 150bp. The 318 chip is anticipated to provide a ten-fold increase in throughput in the same time frame with raw base accuracy over 99%. The throughput, read length and error profile characteristics of Ion Torrent data will demand high performance analysis tools to match diverse biological application requirements.

Independently developed, RTG Investigator sequence analysis software offers comprehensive variant detection and metagenomic analysis pipelines for large-scale Illumina, Complete Genomics and Roche 454 data sets, and now supports Ion Torrent. The RTG software comes with a rich set of production level tools, including novel metagenomics and variant detection modules.

RTG Investigator proves fastest to results in the most demanding downstream genomic analyses. The RTG gapped aligner is designed for high-throughput mapping of NGS reads and performs with exceptional error tolerance even at read lengths exceeding 150bp. RTG Investigator pipelines integrate the many functions required for delivery of comprehensive, accurate results into simple, consistent commands. These pipelines are easy to install and operate compared to open source alternatives.

We analyzed the performance of RTG Investigator on Ion Torrent human amplicon and bacterial genome datasets. The human amplicon dataset was specific to the cystic fibrosis transmembrane conductance regulator (CFTR) gene. The bacterial genome dataset was sample TY2482 of Shiga-toxin-producing E. coli O104:H4, from the outbreak in Germany that started May 2011.

Figure 1 shows a typical alignment score distribution for mappings of Ion Torrent data (B15-410). In this dataset there are 2,761,903 reads of which RTG Investigator mapped 2,650,746 (94.9%) of them. Here a score of 0 is a perfect alignment, and the even score peaks are indicative of the larger penalty associated with inserts and deletes (+2 INDELs versus +1 for substitutions). Overall there are a high number of low alignment score mappings, which in turn gives better data for downstream processing.

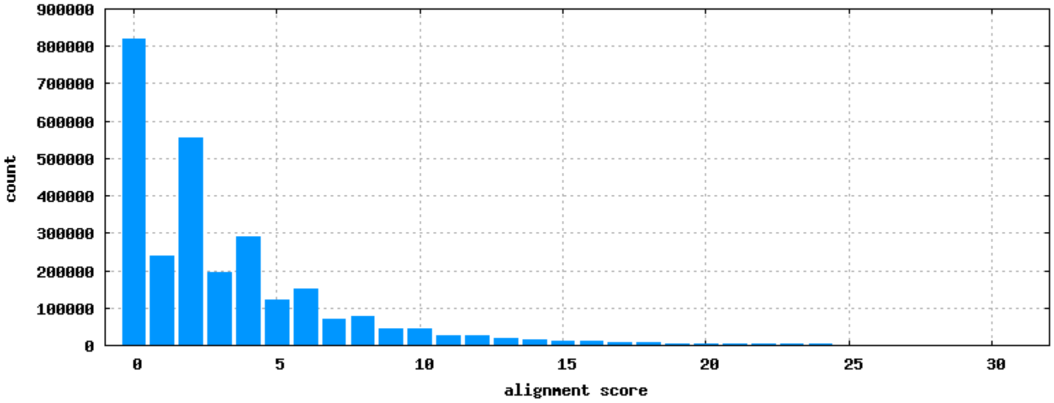


Figure 1 - RTG `map` alignment score distribution for Ion Torrent B15-410 dataset.

Variant Analysis of CFTR Amplicon

Method

The following procedure applies the RTG Investigator variant detection pipeline to the processing of CFTR Amplicon read data from Ion Torrent. The procedure assumes that barcode information has been removed from reads before processing.

1. Trim low quality parts of reads. The RTG format command applies quality filtering and trimming FASTQ, FASTA, or SAM/BAM during the conversion to RTG SDF format.
2. Align reads against a genome reference sequence. A specified SAM file read-group header line informs the aligner (`map`) to adjust alignment penalties to suit Ion Torrent reads error characteristics.
3. Call variants. The SNP/INDEL caller (`snp`) takes the mappings and corresponding read quality recalibration information from the mapper to update its Bayesian priors to target both the Ion Torrent read characteristics and the reference genome type.

Figure 2 shows the read length distribution of the B15-410 dataset before and after quality trimming is applied. The profile is typical of the trimming that occurs, where the peak of the curve drops and is pushed to the left. The third line on the curve indicates how well the trimmed reads are mapped as it closely shadows the trimmed read curve.

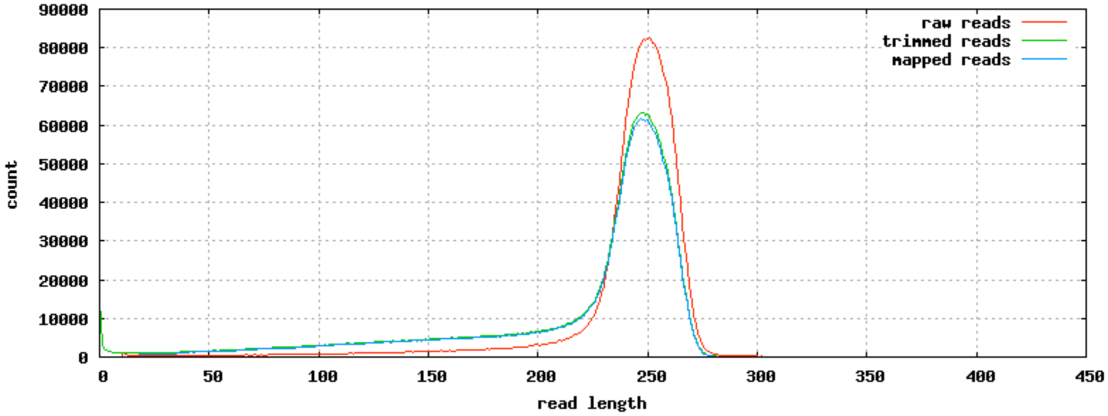


Figure 2 – Read length distribution for B15-410 dataset, before and after RTG quality trimming.

Results

As an example of a useful variant detection application we replicated the results produced in the Ion Torrent Targeted Resequencing Guide Application Note [2]. Two sets of CFTR amplicon data (one sample and one control) were analyzed using RTG Investigator. After read trimming with a quality score <15, between 16% and 22% of the read residues were removed across the 3 samples. When mapped these reads shows similar mapping coverage characteristics to Ion Torrent's TMAP and its soft clipping approach. Figures 3 and 4 show the read mapping distribution across the CFTR.21.420s region for each dataset when using TMAP with soft clipping and RTG `map` with quality clipping.

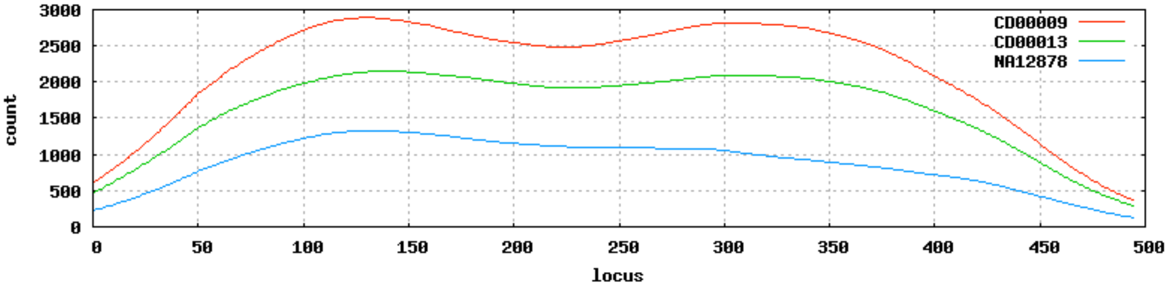


Figure 3 - TMAP read mapping coverage for 3 CFTR samples over the CFTR.21.420s region.

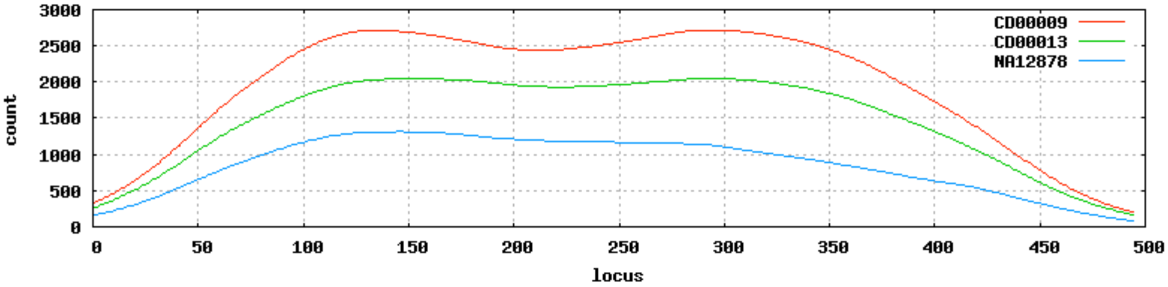


Figure 4 - RTG `map` coverage for 3 CFTR samples over the CFTR.21.420s region.

Table 1 shows counts of the reads mapped and software run times for the two amplicon samples when mapped with Ion Torrent's TMAP and RTG map. In each case RTG map is four times faster than TMAP. Though mapping percentages are lower for RTG, the effective coverage is comparable to TMAP because RTG maps the full length of the reads after clipping.

		TMAP		RTG map	
Sample	Reads	Mapped	Time(s)	Mapped	Time(s)
CD00009	409712	219274	47.2	167452	11.1
NA12878	384880	213392	53.2	149216	11.0

Table 1 - Read mapping counts and run times for TMAP and RTG `map` on 3 amplicon datasets.

After mapping, each amplicon sample is run through the RTG `snp` caller. Table 2 shows RTG Investigator calls compared to known calls. For the NA12878 and CD00009 samples all calls match those detected by the TMAP/samtools/bcftools pipeline. Although the last variant in the RTG CD00009 list was not detected at the same location, it was detected close by. Further investigation showed that the region around the 'GATT' MNP in the reference had 16 copies of GATT, and the RTG snp caller had chosen a different variant of it.

rsID	Primer Name	Ref	PGM Genotype	RTG Genotype	
		NA12878	CD00009	NA12878	CD00009
rs34855237	CFTR.10.80s.1	A	G	A	G
rs213950	CFTR.10.80s.1	G	A	G	A
rs4148711	CFTR.12.570s.1	T	T	T	T
rs11978434	CFTR.13.120s.1	T	T	T	T
rs1042077	CFTR.14.550s.1	T	T	T	T
rs4148712	CFTR.14.550s.1	AT	AT	AT	AT
hg19 chr7:117251899	CFTR.19.560s.1	TT	TT/-	TT	TT/-
rs213989	CFTR.2.520s.1	C	C	C	A/C
rs214164	CFTR.21.420s.1	G	G	G	G
rs4727855	CFTR.25.190s.1	G	G	G	G
rs35516286	CFTR.3.70s.1	T	T	T	T
rs1429566	CFTR.3.70s.1	G	A/G	G	A/G
rs34159932	CFTR.5.520s.1	G	A/G	G	A/G
rs1800503	CFTR6.210s.1	C	C/T	C	C/T
rs67140043	CFTR6.210s.1	GATT	GATT/-	GATT	?

Table 2 - Variant calls for CFTR amplicon samples.

Species Identification of Shiga-Toxin-Producing E. Coli O104:H4

Method

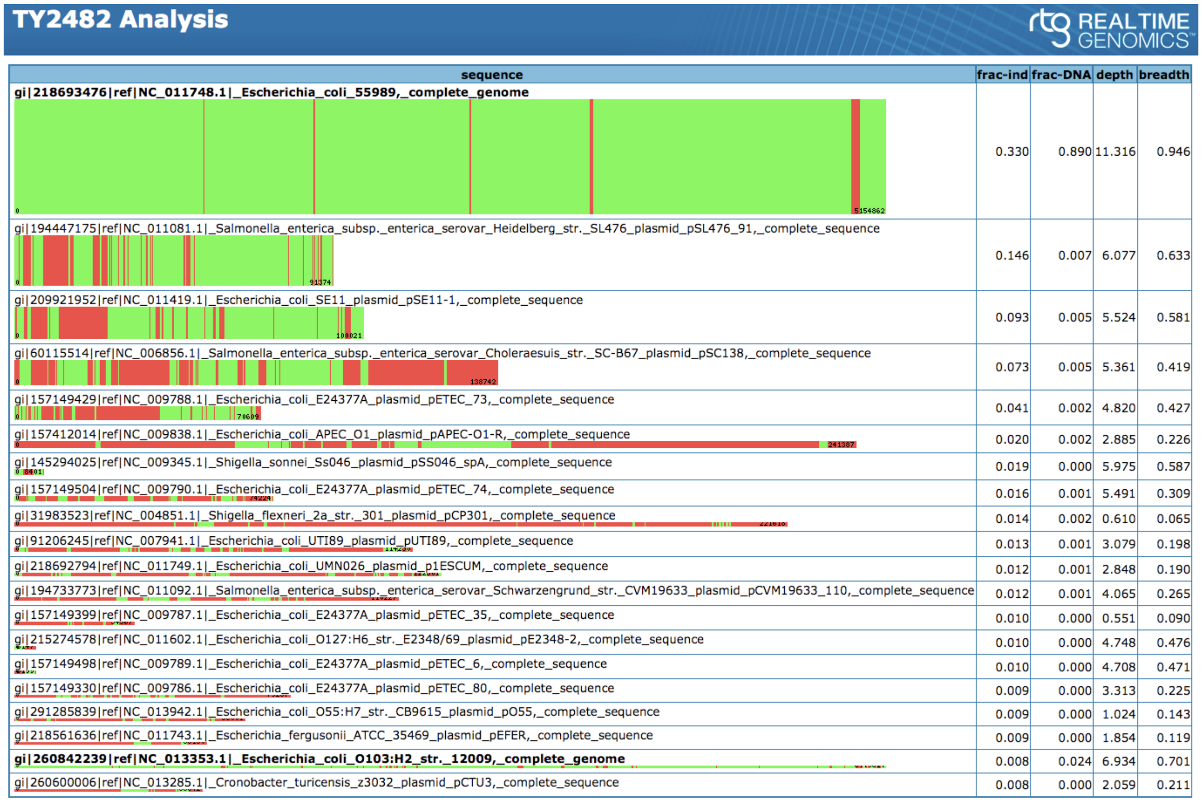
The following procedure applies elements of the RTG Investigator metagenomic analysis pipeline to process Ion Torrent PGM™ reads to examine sample DNA composition. This procedure also assumes that barcode information has been removed from the reads.

1. Trim low quality parts of reads. FASTQ, FASTA, or SAM/BAM files are converted to RTG SDF format and quality filtering and trimming is applied.
2. Align reads against a bacterial reference sequence database.
3. Run RTG `species` tool on mapped reads.

Results

First, sequence data from *E. coli* sample TY2482 were aligned against a bacterial sequence database containing 1,300 species (with a total of 2,500 sequences, contigs and plasmids), and then the RTG `species` module was applied to the alignment results. Figure 5 shows a visualization of the outputs from the RTG Investigator species module. In the first column, the height of each bar represents depth of coverage; green areas show sequence that is covered and red areas indicate a lack of coverage. Data for these bar plots was generated using RTG coverage.

The second column presents estimated abundance of each species/sequence as a fraction of sequences that have a hit, whereas the third column gives the amount of DNA in the sample coming from that sequence. The last two columns give more traditional mapping coverage depth and breadth counts for each sequence.



TY2482 Analysis

<table>
<tr><th>sequence</th><th>frac-ind</th><th>frac-DNA</th><th>depth</th><th>breadth</th></tr>
<tr><td><b>gi|218693476|ref|NC_011748.1|_Escherichia_coli_55989,_complete_genome</b></td><td>0.330</td><td>0.890</td><td>11.316</td><td>0.946</td></tr>
<tr><td>gi|194447175|ref|NC_011081.1|_Salmonella_enterica_subsp._enterica_serovar_Heidelberg_str._SL476_plasmid_pSL476_91,_complete_sequence</td><td>0.146</td><td>0.007</td><td>6.077</td><td>0.633</td></tr>
<tr><td>gi|209921952|ref|NC_011419.1|_Escherichia_coli_SE11_plasmid_pSE11-1,_complete_sequence</td><td>0.093</td><td>0.005</td><td>5.524</td><td>0.581</td></tr>
<tr><td>gi|60115514|ref|NC_006856.1|_Salmonella_enterica_subsp._enterica_serovar_Choleraesuis_str._SC-B67_plasmid_pSC138,_complete_sequence</td><td>0.073</td><td>0.005</td><td>5.361</td><td>0.419</td></tr>
<tr><td>gi|157149429|ref|NC_009788.1|_Escherichia_coli_E24377A_plasmid_pETEC_73,_complete_sequence</td><td>0.041</td><td>0.002</td><td>4.820</td><td>0.427</td></tr>
<tr><td>gi|157412014|ref|NC_009838.1|_Escherichia_coli_APEC_O1_plasmid_pAPEC-O1-R,_complete_sequence</td><td>0.020</td><td>0.002</td><td>2.885</td><td>0.226</td></tr>
<tr><td>gi|145294025|ref|NC_009345.1|_Shigella_sonnei_Ss046_plasmid_pSS046_spA,_complete_sequence</td><td>0.019</td><td>0.000</td><td>5.975</td><td>0.587</td></tr>
<tr><td>gi|157149504|ref|NC_009790.1|_Escherichia_coli_E24377A_plasmid_pETEC_74,_complete_sequence</td><td>0.016</td><td>0.001</td><td>5.491</td><td>0.309</td></tr>
<tr><td>gi|31983523|ref|NC_004851.1|_Shigella_flexneri_2a_str._301_plasmid_pCP301,_complete_sequence</td><td>0.014</td><td>0.002</td><td>0.610</td><td>0.065</td></tr>
<tr><td>gi|91206245|ref|NC_007941.1|_Escherichia_coli_UTI89_plasmid_pUTI89,_complete_sequence</td><td>0.013</td><td>0.001</td><td>3.079</td><td>0.198</td></tr>
<tr><td>gi|218692794|ref|NC_011749.1|_Escherichia_coli_UMN026_plasmid_p1ESCUM,_complete_sequence</td><td>0.012</td><td>0.001</td><td>2.848</td><td>0.190</td></tr>
<tr><td>gi|194733773|ref|NC_011092.1|_Salmonella_enterica_subsp._enterica_serovar_Schwarzengrund_str._CVM19633_plasmid_pCVM19633_110,_complete_sequence</td><td>0.012</td><td>0.001</td><td>4.065</td><td>0.265</td></tr>
<tr><td>gi|157149399|ref|NC_009787.1|_Escherichia_coli_E24377A_plasmid_pETEC_35,_complete_sequence</td><td>0.010</td><td>0.000</td><td>0.551</td><td>0.090</td></tr>
<tr><td>gi|215274578|ref|NC_011602.1|_Escherichia_coli_O127:H6_str._E2348/69_plasmid_pE2348-2,_complete_sequence</td><td>0.010</td><td>0.000</td><td>4.748</td><td>0.476</td></tr>
<tr><td>gi|157149498|ref|NC_009789.1|_Escherichia_coli_E24377A_plasmid_pETEC_6,_complete_sequence</td><td>0.010</td><td>0.000</td><td>4.708</td><td>0.471</td></tr>
<tr><td>gi|157149330|ref|NC_009786.1|_Escherichia_coli_E24377A_plasmid_pETEC_80,_complete_sequence</td><td>0.009</td><td>0.000</td><td>3.313</td><td>0.225</td></tr>
<tr><td>gi|291285839|ref|NC_013942.1|_Escherichia_coli_O55:H7_str._CB9615_plasmid_pO55,_complete_sequence</td><td>0.009</td><td>0.000</td><td>1.024</td><td>0.143</td></tr>
<tr><td>gi|218561636|ref|NC_011743.1|_Escherichia_fergusonii_ATCC_35469_plasmid_pEFER,_complete_sequence</td><td>0.009</td><td>0.000</td><td>1.854</td><td>0.119</td></tr>
<tr><td><b>gi|260842239|ref|NC_013353.1|_Escherichia_coli_O103:H2_str._12009,_complete_genome</b></td><td>0.008</td><td>0.024</td><td>6.934</td><td>0.701</td></tr>
<tr><td>gi|260600006|ref|NC_013285.1|_Cronobacter_turicensis_z3032_plasmid_pCTU3,_complete_sequence</td><td>0.008</td><td>0.000</td><td>2.059</td><td>0.211</td></tr>
</table>

Figure 5 – Analysis of the novel Ion Torrent TY2482 *E. coli* sample shows a close relationship to E. coli 55989 and presence of highly related plasmid sequences.

As identified in other other published analyses, the sequenced strain, *E.coli* TY2482, is most similar to *E.coli* 55989. The composition analysis also identified several related plasmid DNA sequences, the major component being highly similar to pEC_Bactec [3].

Discussion

For deployment, RTG Investigator is installed as a single Java executable on a standard Linux-based computing environment. The software is available free for individual use, and can be downloaded directly from the Real Time Genomics website.

Updates to the models for read error characteristics enabled RTG Investigator to process Ion Torrent sequence data quickly and accurately. Native support for the Ion Torrent data type allowed use of the existing variant detection and metagenomic analysis pipelines deployed for Illumina or Roche 454-style reads.

For example, use of the RTG Investigator variant detection analysis pipeline on CFTR amplicon data validated results previously reported by Ion Torrent. Similarly, application of the RTG `species` module in the metagenomic analysis pipeline presented insight into potential sequence component inheritance using a relatively simple and straight forward bioinformatics technique.

Acknowledgements

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

References

1. Rohde H, Qin J, Cui Y, et al. Opensource genomic analysis of Shiga-toxin–producing *E. coli* O104:H4. N Engl J Med 2011. DOI: 10.1056/NEJMoa1107643.
2. Life Technologies [2011] Ion Torrent Resequencing Giude, Application Note.
3. Smet A, Van Nieuwerburgh F, Vandekerckhove TT, et al. Complete nucleotide sequence of CTX-M-15-plasmids from clinical Escherichia coli isolates: insertional events of transposons and insertion sequences. PLoS ONE 2010;5(6):e11202.