Ion TorrentTM RNA-Seq Analysis with NextGENe[®] Software v2.30 and above

March 2012

John McGuigan, Lisa Wang, CS Jonathan Liu

Introduction

Ion TorrentTM semiconductor sequencing technology is fundamentally different from most other sequencing. It is a "post-light" technology because it doesn't depend on detection of light emission from nucleotide incorporation. Instead, it uses a silicon chip containing millions of individual pH meters. Its flow-based approach detects pH changes caused by release of hydrogen ion during incorporation of unmodified nucleotides in DNA replication. The rapidly improving throughput of these sequencing chips including the soon-to-be-released Ion ProtonTM sequencer makes it very useful for transcriptome studies.

Due to reference sequence difficulties associated with alternative splicing and fusion genes, alignment of RNA-seq data is more challenging than alignment of DNA sequences. Short reads- especially those that fall within large exons- are able to align normally since they will generally match the reference with very few mismatches. Reads that span an exon-exon junction are more difficult because they must be split at the correct position and each part of the read must align correctly. Fusion genes provide even more of a challenge because the partial reads can align almost anywhere in the genome.

Different solutions to these challenges have been implemented in various software packages. Q-PALMA uses a machine learning algorithm and training datasets in order to identify splice junctions [1]. SuperSplat divides sequence reads at multiple positions and tries to find mapping sites where the sub-reads are separated by an intron in a certain size range[2]. TopHat is a software package that first finds potential exons based on coverage and then finds splice sites and links using canonical splice site sequence information [3]. NextGENe uses a novel algorithm to correctly align reads belonging to annotated and novel transcripts while providing the added benefit of a highly graphical interface that doesn't require use of scripting or the command line. Analysis can be performed on a desktop PC in just a few hours without any training datasets or pre-filtering of the reads.

NextGENe's unique RNA-Seq alignment algorithm and viewer are greatly improved starting in v2.30. The update provides better results (more aligned reads) and more intuitive display of splicing variants. The improvements are demonstrated here with an Ion Torrent 318 chip RNA-Seq dataset (C18-199) which is available on the Ion Torrent Dev community website. The improved transcript view can be seen in figure 1.



Figure 1: The new RNA-Seq Transcript View. The purple links and exons are known, the blue links are novel, and the orange exons are alternative splicing.



SoftGenetics LLC 100 Oakwood Ave. Suite 350 State College, PA 16803 USA Phone: 814/237/9340 Fax 814/237/9343 www.softgenetics.com email: info@softgenetics.com



Methodology

NextGENe's approach takes advantage of previously-known isoform splice sites that still allows for detection of novel transcripts. The RNA-Seq application has a few main steps:

- 1. Align reads to the genome- either GBK files (<250 MB genomes) or a pre-indexed reference
 - a. If a pre-indexed reference is not available, it can be generated with the "Build Index for WGA" tool
 - b. Alignment to a pre-indexed reference will first try to match the read perfectly, then break the read up into seeds. Alignment to GBK files will utilize all 12 bp seeds for alignment.
- 2. Identify area of coverage that represent potential exons
- 3. Compare potential exons to annotated transcripts and build a reference of transcripts
- 4. Align the original reads to the transcript reference to ensure the best alignment
- 5. Convert the aligned read positions to genomic coordinates
- 6. Identify the transcripts that were found along with any splicing variants
- 7. Perform mutation detection for SNPs and short indels
- Paired data is used to create more links and to identify distant regions that may have spliced together.

The single-strand option can resolve strand-specific sequences.

Procedure

- 1. The format conversion tool is used to filter and trim the raw data based on the quality score information. The FASTQ file for this run was processed with the default settings for Ion Torrent.
- The converted FASTA files had the adapter sequence (TCACCGACTGCCCATCTGAGACTGCCAAGG) trimmed using the sequence operation tool (figure 2).
- 3. The transcriptome tool is run.
 - a. "Transcriptome" is selected for application type and the "Alternative Splicing" box is checked. If this box is not checked a normal (no read splitting) alignment will be performed.
 - b. The converted data files are added.
 - c. The reference is selected.
 - d. The output directory is set.
 - e. The options are adjusted (figure 3)..
 - i. Lowering the average (expected) coverage will decrease the coverage threshold for calling transcripts.
 - ii. The settings for seed alignment (seed size, step, and number of allowed ambiguous alignments) and settings for the alignment and mutation filters can be adjusted.
- 4. The project automatically opens in the NextGENe Viewer

peration Type		
C Merge Files	C Split Files	Sequence Trim
C Arrange Paired Reads	C Remove Duplicate Reads	C Reverse/Complement Seq
ut:		
		Add
		Remove
		Remove A
put:		
		C . 1
		Set
attings		
ettings	1 (MB) E Maria	
Max Size of Each New File:	1 (MB) V Max #	of Uncalled Bases <= 3
Max Size of Each New File:	1 (MB)	of Uncalled Bases <= 3
Max Size of Each New File: Called Base Number of Eac Trim or Reject Read when	1 (MB)	of Uncalled Bases <= 3
ettings Max Size of Each New File: ✓ Called Base Number of Eac Trim or Reject Read when Remove 5' 2 8	1 (MB) $r r$ Max # h Read >= 50 50 >= 3 Base(s) with Score sase(s) and 3' 2	of Uncalled Bases <= 3
Max Size of Each New File: Called Base Number of Eac Trim or Reject Read when Remove 5' 2 B Save the Trimmed Reads/v	I (MB) I Max # h Read >= 50	of Uncalled Bases <= 3
attings Max Size of Each New File: If Called Base Number of Each Trim or Reject Read when Remove 5' Save the Trimmed Reads/k If this Save the Save the the Trimmed Reads/k	1 (MB)	of Uncalled Bases <= 3







	Chan Designations
	Show Project Log>:
	Analysis Options
	Auto Detect PE Library Size Library Size: Min 200 Max 500
p	Paired Reads
polication	Parameters for Alternative Splicing Analysis
pplication	Average Coverage 10
	, Remove Non-Linked Exons
	Seed Length 21
oad Data	Move Step 4
	✓ Ignore Fusions Between Similar Genes
	Rigorous Fusion Detection
ndensation	
	Allowable Ambiguous Number 50
	Range[10,50]
ssembly	Parameters for New Gene Detection
	Exon Size: Min 12 Max 10000 Average Coverage 10
lignment	Intron Size: Min 6 Max 200000 Donor-Acceptor GT AG
	Parameters for Hash-Table Alignment
	Matching Requirement: >= 30 Bases and >= 85 %
	Allow Ambiguous Mapping
	Remove Ambiguously Mapped Read
	Parameters for Mutation Detection
	Mutation Percentage <= 20 SNP Allele <= 3 Counts
	Total Coverage <= 5
	Default Settions Save Settions Load Settions

Figure 3: Transcriptome Alignment Settings

Results

5,193,130 of 5,384,713 reads were converted successfully (96.44%). Only 466 reads needed to have the adapter sequence removed. Alignment results are summarized in table 1.

Reads Matched	4,125,288
% Reads Matched	79.44%
Processing Time	4 hours, 25 min
Mutations Found	41,223
Substitutions	8,683 (21.1%)
In dbSNP 135	7,264 (17.6%)

Table 1: Alignment Results (run on a 16-core computer, 16 GB of RAM used)

Discussion

When the project is first opened in the viewer, the mutation report view is shown. The mutation report details all of the called variants, and the expression report (figure 4) is also available- it lists several different expression measurements for the sample. Most of the genes with the highest expression (sorted by RPKM) were part of the mitochondrial genome.

gment Settings	Expression	Report											-
Regions	File Settings												
 Use Segments as Defined in Reference Files 													
G Gene C mBNA C CDS	Software	NextGENe V 2.3.0	1										
	Project Name	C18-199.pjt											
C Continuous mRNA C Continuous CDS	Date/Time	3/23/2012 1:31:2											
C ROI	Total Reads	5193130											
C. Set Incompatel Segment Length Isoppo	Matched Reads	4125288											
Set inclemental segment Length 10000 Bases	Instrument												
C Input Region of Interest (*.bed)	Application	Transcriptome											
Set													
	Index	Chr	Chr Position	Chr Position	Gene	Length	Min Counts	Max Counts	Average Cou	Read Counts	Forward Rea	▼ RPKM	RPK .
imits	1	15	83205504	83209207	RPS17L; ·	485	1	7195	2637.00	30812	0	12335.1239	5.9825
Limit to First 200 bp	2	м	7586	8269	C0X2; +	684	1159	9083	4811.90	38654	38588	10972.4524	7.5052
	3	м	8366	8572	ATP6; +	207	1901	4367	3129.30	11271	11261	10572.0077	2.1884
Limit to Last 200 bp	4	м	10760	12137	ND4; +	1378	190	8640	3595.00	67648	67512	9531.7163	13.1347
	5	м	5904	7445	C0X1;+	1542	357	13239	3759.00	74532	74494	9384.7729	14.4713
Ok Cancel	6	12	6643657	6647535	GAPDH; +	1310	3	7538	2464.00	54541	54541	8083.8299	10.5898

Figure 4: Expression report settings and the expression report for this sample sorted by RPKM





In the transcript view it is possible to view the discovered exons and links. The NextGENe viewer can toggle between the mutation detection view and the transcript view with the push of a button:

- Link Types links connect locations where a read was split for alignment
 - Purple = Known (annotated) link
 - Blue = Novel link (doesn't match any annotated links)
 - Black = Fusion link (links together two different genes)
- Node Types nodes are regions (often exons) where reads aligned (or didn't align, in the case of deletions).
 - Purple = Known exon
 - Blue = Novel exon in an unannotated region (possible new gene)
 - Red = Intron Retention or Insertion
 - Pink = Exon Skipping
 - Orange = Alternate Splice Site (exon is longer or shorter than annotated)
 - Gray = Alternative Start/Stop (the first or last exon in a gene is alternatively spliced at the edge of the gene) or anti-sense transcription for single-strand sequencing

The transcript report has many detailed, flexible filters (figure 5). When a reference is built from gbk files, the known annotation can be even more detailed- ncRNAs, mRNA, tRNA, etc are annotated separately instead of in an inclusive "Annotated RNA" group.

M Transcript Report Settings		- 0	×
Filters Columns			
Record Type		1	
Link Type	Sequence Type		
Vovel	Annotated RNA		
🔽 Known	✓ Unannotated		
✓ Fusion	✓ Intron Retention		
Single Strand Type	✓ Insertion		
Sense Transcription	Exon Skipping		
Antisense Transcription	Alt Splice		
	Alt Transcript Start/Stop		
Filter Min Number of Novel Link 1 Min Number of Fusion 0 Min PE Number of Fusion 0 Min Coverage of Alt Splice	Min Coverage of Unannotated Min Coverage of Intron Retention Min Coverage of Intron Retention Min Coverage of Insertion Min Coverage of All Start/Stop 1		
Save Settings	Load Settings Default	Cancel	

Figure 5: Transcript Report Filters

Many genes were found to have normal splicing, including the highly expressed mitochondrial genes which are not spliced (figure 6) and the mitochondrial ribosomal protein L9 (MRPL9) which is a reverse-strand gene with 7 exons (figure 7).



Figure 6: Expression of the Mitochondrial Genes





SoftGenetics LLC 100 Oakwood Ave. Suite 350 State College, PA 16803 USA Phone: 814/237/9340 Fax 814/237/9343 www.softgenetics.com email: info@softgenetics.com



SOFTGENETICS® Software PowerTools for Genetic Analysis NextGENe was able to align reads in unannotated regions to detect possible new genes (figure 8).



Figure 8

The scale of the local coverage graph can be adjusted, allowing for investigation of low-frequency alternative splicing, such as in the HNRNPA1 gene (figure 9). In this gene the alignment indicated that some transcripts expressed less than 5% of the time had alternatively spliced exons (orange).



Figure 9

Alternative transcript start and stop sites are reported and highlighted in gray. One example is shown in figure 10. Premature stop sites are often caused by low coverage of the 5' end of the gene.











Figure 11 shows an example of intron retention and partial intron retention (called as "insertion") in the STRA13 gene.



Figure 12 shows an example of exon skipping. The skipped exon is highlighted in pink, which the previously unannotated link over the exon is marked as novel (blue).



References

References

- 1. Fabio De Bona et al., "Optimal spliced alignments of short sequence reads," Bioinformatics 24, no. 16 (2008): i174 -i180.
- 2. Douglas W. Bryant et al., "Supersplat-spliced RNA-seq alignment," Bioinformatics 26, no. 12 (June 15, 2010): 1500 -1505.
- 3. Cole Trapnell, Lior Pachter, and Steven L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," Bioinformatics 25, no. 9 (May 1, 2009): 1105 -1111.



