RNA-Seq Analysis with NextGENe Software

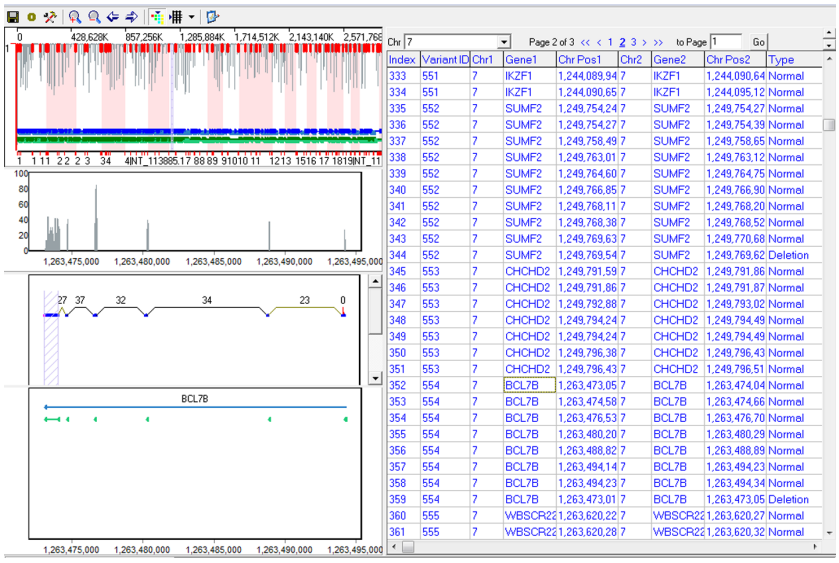
February 2011

John McGuigan, Yaping You, CS Jonathan Liu

Introduction

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.



Index	Variant ID	Chr1	Gene1	Chr Pos1	Chr2	Gene2	Chr Pos2	Type
333	551	7	IKZF1	1,244,089,94	7	IKZF1	1,244,090,64	Normal
334	551	7	IKZF1	1,244,090,65	7	IKZF1	1,244,095,12	Normal
335	552	7	SUMF2	1,249,754,24	7	SUMF2	1,249,754,27	Normal
336	552	7	SUMF2	1,249,754,27	7	SUMF2	1,249,754,39	Normal
337	552	7	SUMF2	1,249,758,49	7	SUMF2	1,249,758,65	Normal
338	552	7	SUMF2	1,249,763,01	7	SUMF2	1,249,763,12	Normal
339	552	7	SUMF2	1,249,764,60	7	SUMF2	1,249,764,75	Normal
340	552	7	SUMF2	1,249,766,85	7	SUMF2	1,249,766,90	Normal
341	552	7	SUMF2	1,249,768,11	7	SUMF2	1,249,768,20	Normal
342	552	7	SUMF2	1,249,768,38	7	SUMF2	1,249,768,52	Normal
343	552	7	SUMF2	1,249,769,63	7	SUMF2	1,249,770,68	Normal
344	552	7	SUMF2	1,249,769,54	7	SUMF2	1,249,769,62	Deletion
345	553	7	CHCHD2	1,249,791,59	7	CHCHD2	1,249,791,86	Normal
346	553	7	CHCHD2	1,249,791,86	7	CHCHD2	1,249,791,87	Normal
347	553	7	CHCHD2	1,249,792,88	7	CHCHD2	1,249,793,02	Normal
348	553	7	CHCHD2	1,249,794,24	7	CHCHD2	1,249,794,49	Normal
349	553	7	CHCHD2	1,249,794,24	7	CHCHD2	1,249,794,49	Normal
350	553	7	CHCHD2	1,249,796,38	7	CHCHD2	1,249,796,43	Normal
351	553	7	CHCHD2	1,249,796,43	7	CHCHD2	1,249,796,51	Normal
352	554	7	BCL7B	1,263,473,05	7	BCL7B	1,263,474,04	Normal
353	554	7	BCL7B	1,263,474,58	7	BCL7B	1,263,474,66	Normal
354	554	7	BCL7B	1,263,476,53	7	BCL7B	1,263,476,70	Normal
355	554	7	BCL7B	1,263,480,20	7	BCL7B	1,263,480,29	Normal
356	554	7	BCL7B	1,263,488,82	7	BCL7B	1,263,488,89	Normal
357	554	7	BCL7B	1,263,494,14	7	BCL7B	1,263,494,23	Normal
358	554	7	BCL7B	1,263,494,23	7	BCL7B	1,263,494,34	Normal
359	554	7	BCL7B	1,263,473,01	7	BCL7B	1,263,473,05	Deletion
360	555	7	WBSCR22	1,263,620,22	7	WBSCR22	1,263,620,27	Normal
361	555	7	WBSCR22	1,263,620,28	7	WBSCR22	1,263,620,32	Normal

The transcript variant view of the NextGENe Viewer

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 pre-indexed reference using NextGENe's Whole Genome Alignment method
    a. Save alignments when a read matches the reference perfectly
    b. Break the unmatched reads into seeds of a specified size
    c. Match the seeds and extend the alignment where matching positions are found
    d. Ignore seeds that map to more than a specified number of sites
2. Align remaining seeds using an exon junction reference
    a. Use the alignment information to break the reads and align them to the whole-genome reference
3. Mark covered regions (potential exons) and record the IDs of reads aligned in those regions
4. Create links between regions when the same read is partially mapped in both regions
5. Compare discovered transcripts to annotated transcripts, marking any insertions, deletions, or fusions
6. Align the original reads to the discovered transcripts to ensure the best alignment and re-call the transcripts based on the aligned reads
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.

Two projects are output- a normal variant detection project and a transcript viewer project which shows detected transcripts and reports normal exons, insertions, deletions, novel transcripts, and fusions rather than SNPs and small indels which can be found in the regular project file.

Procedure

Paired-end Human RNA-Seq data sequenced on an Illumina Genome Analyzer was downloaded from the NCBI Sequence Read Archive (SRX011551) and used in this analysis.

1. The pre-indexed whole-genome reference provided by SoftGenetics (multiple species and builds are available) contains all annotated transcripts in order to be most effective. One of these genomes should be used with the query genome annotation tool (found under the tools menu) in order to save the annotation information (figure 1). This step only needs to be performed once, but can be repeated if the database changes. The database name must also be correct and can be set on the "DataBase" tab of the options window found under the "Process" menu (figure 2)

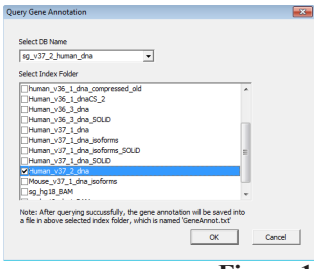


Figure 1

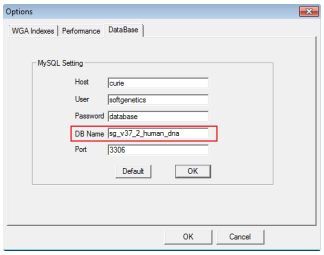


Figure 2

2. The format conversion tool is used to filter and trim the data based on the quality score information. It outputs the data in .fasta format (.csfasta for SOLiD data). The "minimum called base number" should be set to at least 50- this will remove reads that were trimmed too short to be effective in this analysis.

3. The transcriptome tool is run (figure 3).
   a. The converted data files are added.
   b. The reference is selected.
   c. The output directory is set.
   d. The options are adjusted.
      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. Two projects are generated- a SNP/Indel detection project and a transcript analysis project. Both can be opened in the NextGENe viewer for review.

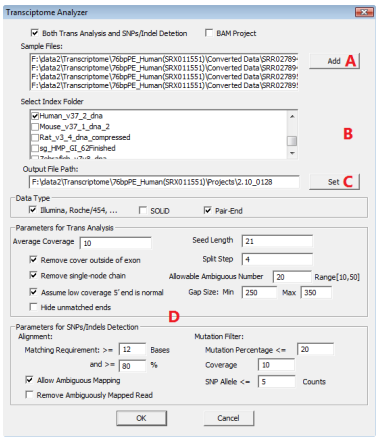


Figure 3

Results

35,022,710 of 40,497,204 reads were converted successfully (86.48%). Processing took approximately 3 hours and 45 minutes. At the end of the analysis 20,822,825 reads were used (59.5%). The results of mutation and transcript variant detection are summarized in table 1.

Variant	Number Identified
Fusions	2
Unannotated exons	443
Alternate Splice Sites	1,293
Exon Skipping	1,074
Intron Retention	43
Normally expressed exons	22,263
Total SNPs and Indels	19,397
High Confidence SNPs	4,558
Substitutions	15,517
SNPs and Indels in dbSNP	2,713
SNPs and Indels in CDS	4,052

Table 1 - Variant Detection Results

Table 1

Discussion

The transcript and mutation reports can be seen in figures 4 and 5. The transcript report is showing all of the results for the ILK gene on chromosome 11 while the mutation report is showing some of the results for the RGS1 gene. The transcript report gives information about detected exons- location, coverage/link number (average coverage in an exon or the number of links in a fusion), type of variant (insertion, deletion, new, or normal exon), function (UTR, CDS, or unknown), location type (alternative splice site, exon skipping, etc), and isoform/protein information. The mutation report is highly customizable with many different filter and display options.

Index	Variant	Chr1	Gene1	Chr Pos1	Chr2	Gene2	Chr Pos2	Type	Coverage/Link	Function	Location	Isoform	Protein
10627	820	11	ILK	1,752,530,178	11	ILK	1,752,530,266	Normal	0	UTR		NM_001014794.1	NP_00101479
10628	820	11	ILK	1,752,530,670	11	ILK	1,752,530,804	Normal	9	CDS/UTR		NM_001014794.1	NP_00101479
10629	820	11	ILK	1,752,534,490	11	ILK	1,752,534,655	Normal	13	CDS		NM_001014794.1	NP_00101479
10630	820	11	ILK	1,752,534,838	11	ILK	1,752,534,933	Normal	23	CDS		NM_001014794.1	NP_00101479
10631	820	11	ILK	1,752,535,138	11	ILK	1,752,535,412	Normal	21	CDS		NM_001014794.1	NP_00101479
10632	820	11	ILK	1,752,535,516	11	ILK	1,752,535,601	Normal	12	CDS		NM_001014794.1	NP_00101479
10633	820	11	ILK	1,752,535,744	11	ILK	1,752,535,853	Normal	16	CDS		NM_001014794.1	NP_00101479
10634	820	11	ILK	1,752,535,957	11	ILK	1,752,536,084	Normal	21	CDS		NM_001014794.1	NP_00101479
10635	820	11	ILK	1,752,536,169	11	ILK	1,752,536,290	Normal	22	CDS		NM_001014794.1	NP_00101479
10636	820	11	ILK	1,752,536,381	11	ILK	1,752,536,460	Normal	16	CDS		NM_001014794.1	NP_00101479
10637	820	11	ILK	1,752,536,593	11	ILK	1,752,536,723	Normal	17	CDS		NM_001014794.1	NP_00101479
10638	820	11	ILK	1,752,536,907	11	ILK	1,752,537,287	Normal	18	CDS/UTR		NM_001014794.1	NP_00101479
10639	820	11	ILK	1,752,537,287	11	ILK	1,752,537,315	Deletion	0	UTR	Alternative splice	NM_001014794.1	NP_00101479
10640	820	11	ILK	1,752,535,234	11	ILK	1,752,535,329	Insertion	0	New	Alternative splice	NM_001014794.1	NP_00101479

Figure 4

Index	Gene	CDS	Chr	Ref Nuc	Coverage	Score	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref	Mutation Call
1	RGS1		1	G	2575	27.2	99.88	0.00	0.04	0.04	0.00	0.04	rs7535818	IVS137+40G>A
2	RGS1		1	A	11027	21.3	8.58	0.80	0.02	69.75	0.00	20.86		IVS138-2delA/IVS138-2A>T
3	RGS1		1	G	12001	26.4	0.19	0.09	7.38	72.82	0.00	19.52		IVS138-1G>T
4	RGS1		1	A	4125	23.4	38.79	61.14	0.02	0.05	0.00	0.00		IVS219-1A>AC
5	RGS1		1	G	16041	22.2	0.34	77.55	21.80	0.29	0.00	0.02		IVS280+1G>CG
6	RGS1		1	A	4453	29.1	52.08	0.04	47.81	0.07	0.00	0.00	rs12136860	IVS280+177A>AG
7	RGS1		1	G	49079	24.7	92.52	0.09	7.33	0.05	0.00	0.00		IVS444+1G>A
8	RGS1		1	G	47349	20.1	0.16	92.22	7.54	0.07	0.00	0.00		IVS444+1G>C
9	RGS1		1	G	46464	20.1	92.19	0.03	7.57	0.12	0.00	0.09		IVS444+5G>A
10	RGS1		1	T	8624	31.4	0.17	55.00	0.02	44.78	0.00	0.02	rs2760532	IVS444+276T>CT
11	RGS1		1	C	2465	27.1	0.00	44.06	0.00	55.94	0.00	0.00	rs2760533	IVS445-226C>CT
12	RGS1		1	T	2284	26.8	0.00	53.63	0.00	45.53	0.00	0.83	rs2816306	IVS445-222T>CT
13	RGS1		1	G	2349	26.9	0.04	0.04	56.53	43.38	0.00	0.00	rs2816307	IVS445-38G>GT
14	RGS1		1	T	113881	40.4	0.07	0.06	49.50	50.37	0.00	0.00	rs13232291	IVS630+91T>GT
15	RGS1		1	G	118401	40.5	0.12	47.20	52.48	0.11	0.00	0.10	rs2816308	IVS630+107G>CG

Figure 5

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®
Next Generation Sequencing Software

Figure 6 shows the transcript view for the ILK gene including an insertion (maroon) and an alternative splice site at the 3' end of the gene (pink). The link between the first and second exon is a virtual link because coverage was too low at the 5' end of the gene. RNA-Seq often shows a bias for higher coverage at the 3' end. Figure 7 shows one of two detected fusions.

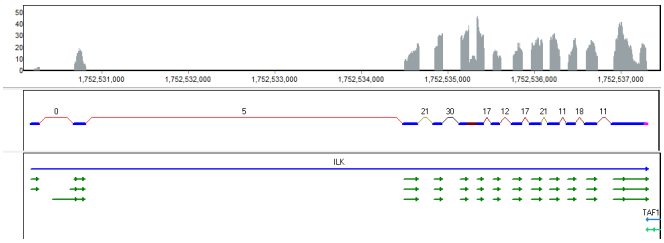


Figure 6

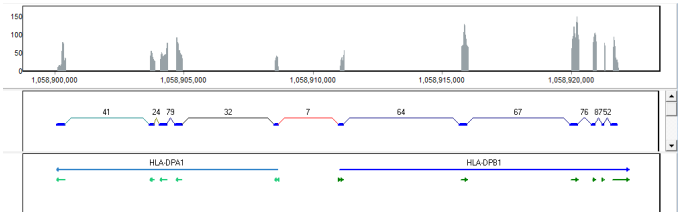


Figure 7

As seen in figure 8, expression levels are reported in NextGENe's expression report, which can be accessed from the SNP/Indel project. In this example expression levels were calculated on a per-gene basis and sorted by normalized expression levels (RPKM). The expression comparison report can be used to compare expression levels in multiple projects.

SOLiD data is always processed in colorspace as seen in figure 9. The ">" symbol indicates the 5' end of a read and the symbol is red when the read is split.

Several enhancements are planned for the near future including:

- Combined transcript and SNP/Indel View
- Improved fusion and multiple-isoform discovery based on detected mutations
- Integration of the tool into NextGENe's project wizard
- Addition of a single-strand sequencing analysis option

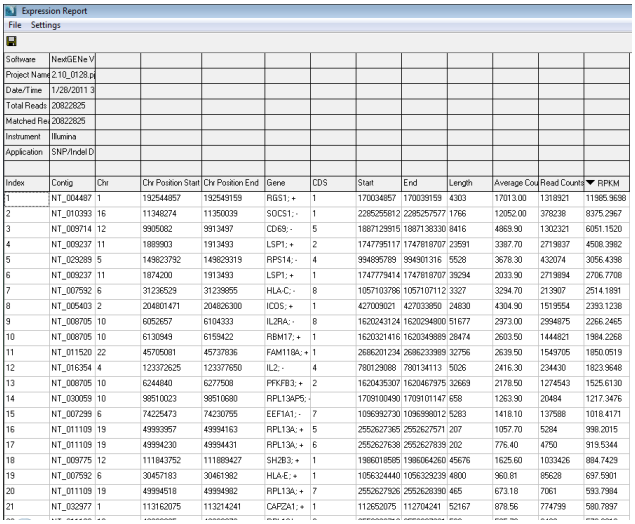


Figure 8

Figure 9

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.