JAGuaR: Junction Alignments to Genome for Repositioning of RNA-seq Reads

(RNA-Seq-Repos)

Yaron S. Butterfield, Richard Corbett, Nina Thiessen, An He, Inanc Birol, Steve JM Jones, Marco A. Marra

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

JAGuaR is a Python package for the alignment of sequence reads from whole transcriptome shotgun sequencing. A transcriptome reference sequence is built based on a transcript model from Ensembl or other sources. Reads of various sizes from 25mer to 100mers or higher can be accurately aligned to this reference using BWA and are split over as many exons as needed. These are subsequently repositioned to genome coordinates. This is a novel approach as compared to traditional alignment based methods of the aligning to a genome reference concatenated with exon-exon junction sequences (Fig 1).

As read lengths get longer, it becomes less feasible to align to a reference built from one exon-exon junction region. If a read has a length that is greater than the size of an exon, it is not possible to create a unique exon-exon junction sequence. Typically, clipping of a sequence read that aligns to this location is required, resulting in less accurate coverage calculations and potentially less reads to contribute to SNVs and indels (Fig 2 and 3). JAGuaR provides a reference built on as many exon-exon junctions that are required. This provides the context for a read to completely cover the exons that are smaller than the read length and split into two other exons, one on each end.

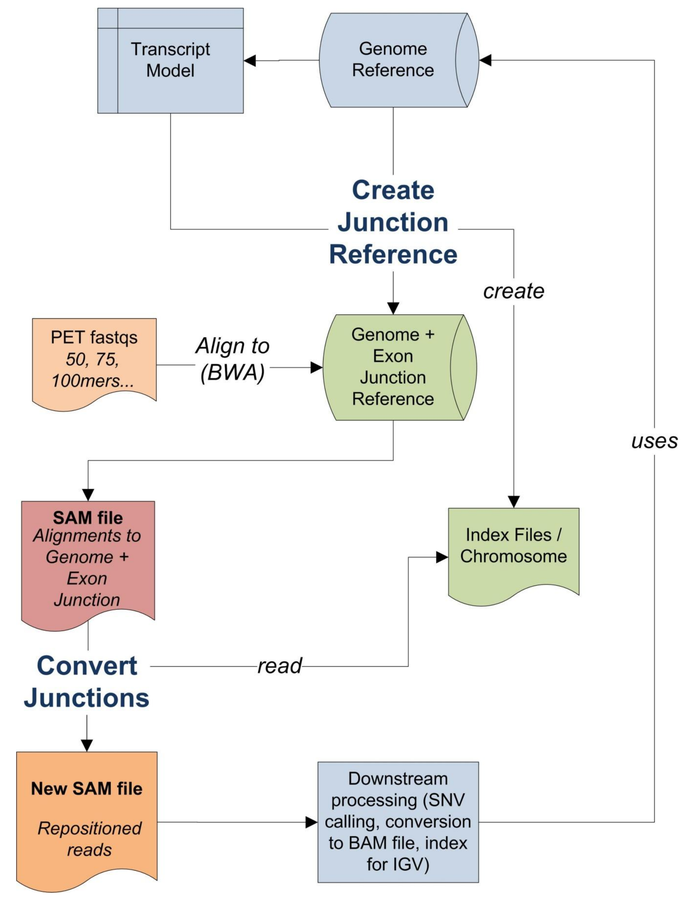

Fig 1. Process Flow. JAGuaR first requires the reference genome of interest and a transcript model (see Table 1 for format) in order to build the reference of the genome sequence and exon junctions. This step is run once and is required for each length of sequence reads. After reads are aligned to this reference with BWA (Li et. al.) resulting in a SAM file, JAGuar is used to convert the exon junction aligned reads to genome coordinates.

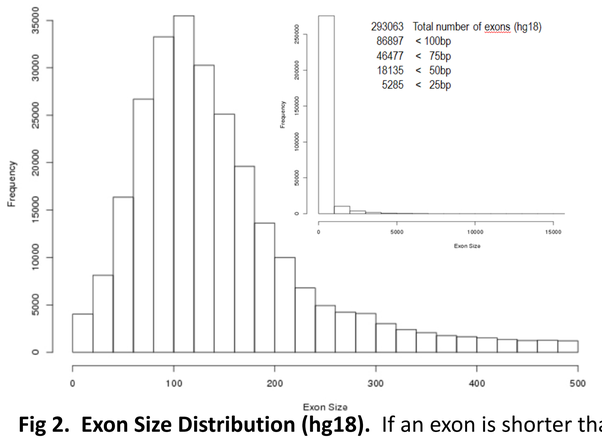

Fig 2. Exon Size Distribution (hg18). If an exon is shorter than the length of the sequence read, it is impossible to create a unique junction reference of two exon junctions and therefore less reads align unambiguously. As read lengths get larger, this becomes more of an issue. There are over 85,000 exons that are smaller than 100bp. In these cases, a junction reference that is built to span exons less than the size of the read length, results in more accurate alignments.

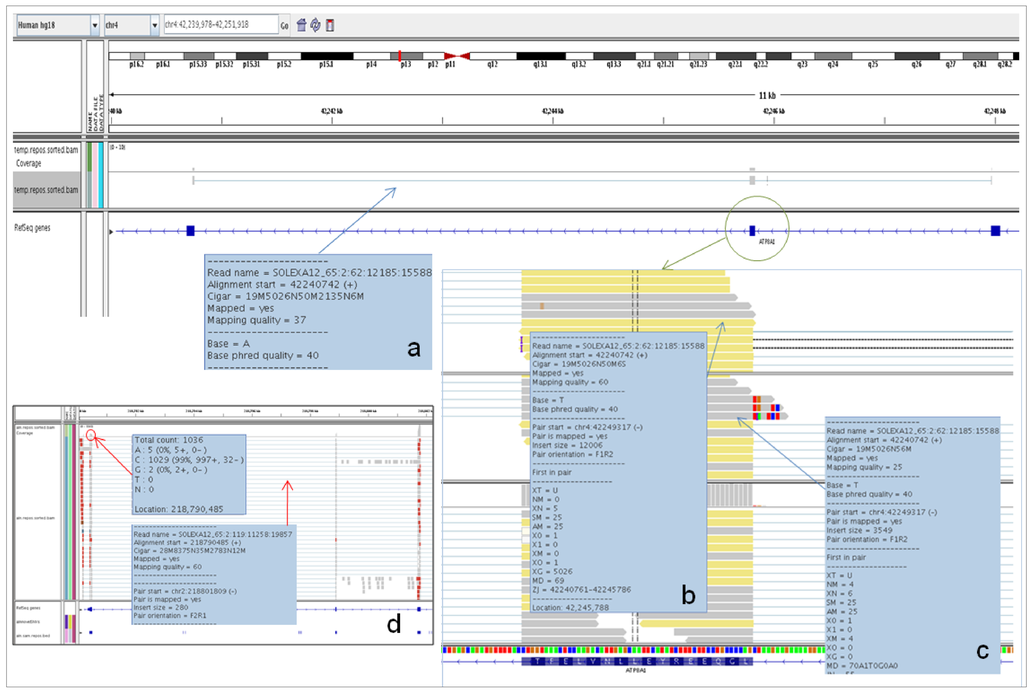

Fig 3. Comparison of exon-exon junction methods. IGV (Robinson et. al) view comparing 75bp sequence alignments to a gene with an exon that is 50bp. When the junction reference incorporates information from three exons in this case using JAGuaR, the read is cleanly aligned (a) as evident in the view and in the CIGAR string. In a two exon junction model, even though the first part of the read covers 19bp of the first exon, the remaining is still greater than the size of the adjoining 50bp exon shown here. The end can either be clipped, eliminating coverage information and possibly SNV evidence (b) or will have non-matching bases potentially resulting in a false SNV (c). Alignments of 75bp reads to a JAGuaR built reference spanning an even smaller exon of 35bp (d) are split correctly across three exons. The read information for one read is shown out of a total of 1036 reads that have covered this location.

Results:

Tool	Read Length	dbSNP129	SNVs called	Concordant SNVs	Fraction	Library
Jaquar (w/ BWA)	50	12908101	115012	99062	0.8613	A00123
GSNAP	50	12908101	148406	127858	0.8615	A00123
TopHat	50	12908101	131956	92828	0.8615	A00123
Jaquar (w/ BWA)	75	12908101	82234	69294	0.8426	HS2937
GSNAP	75	12908101	110691	81952	0.7404	HS2937
TopHat	75	12908101	98765	58979	0.5972	HS2937
Jaquar (w/ BWA)	100	12908101	174859	119072	0.6810	A05010
GSNAP	100	12908101	166645	104689	0.6282	A05010
TopHat	100	12908101	159172	106636	0.6699	A05010

Table 3. Comparison to other methods. SNV results from alignments using this method have been compared to other methods (Trapnell et. al., Wu et. al.) and match or exceed concordance to dbSNP. In the case of the 100bp read length alignment, while concordances are fairly similar, the JAGuaR respositioned alignment file resulted in more concordant SNVs.

Transcript ID	Chr	Exon Start	Exon End	Rank	Strand
ENST0385499	7	157956020	157956126	1	1
---	---	---	---	---	---
t1	1	10002981	10003083	1	+
t1	1	10009696	10010032	2	+
t2	1	10003486	10003573	1	+
t2	1	10032076	10032246	2	+
t2	1	10035650	10035833	3	+
t2	1	10041089	10041228	4	+

Table 1. Example Transcript Model. A text file is required for creating the junction reference. Top (of dashed line): one line based on Ensembl which can be generated for your genome version of interest from BioMart. Bottom: Custom model can be based on a combination of RefSeq, Ensembl and UCFC known genes or novel exons. There is one line per exon for all transcripts.

Total reads	99976	
Exon	70843	70.9%
Intron	9568	9.6%
Intergenic	8009	8.0%
Unaligned	11556	11.6%
Repositioned	22564	22.6%
Exon/Intron Ratio	7.4042	

Table 2. Alignment summary. Results such as identification of intron and exon reads are included in the output in addition to the repositioned SAM file.

Further development:

Currently the reference and format required for TCGA is being implemented. Another important parameter especially in the context of clinical sequencing is the length of time repositioning takes. We are testing and looking at ways of improving this metric. JAGuaR meets the standards, guidelines and best practices for RNA-Seq as set by the ENCODE Consortium (V1.0 – June 2011).

References:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Thomas D. Wu and Serban Nacu
Fast and SNP-tolerant detection of complex variants and splicing in short reads
Bioinformatics 2010 26:873-881

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics doi:10.1093/bioinformatics/btp120

Availability:

JAGuaR

http://www.bcgsc.ca/platform/bioinfo/software/jaguar

ybutterf@bcgsc.ca

Acknowledgements:

The project described was supported by Award Number U24CA143866 from the National Cancer Institute.

We thank all involved with TCGA data production and analysis at BCGSC, and TCGA research network.