PERTRAN: Genome-guided RNA-seq Read Assembler

Shengqiang Shu[1], David M. Goodstein[1], & Dan Rokhsar[1]

[1]Lawrence Berkeley National Laboratory/DOE Joint Genome Institute, Walnut Creek, California 94598, USA.

**To whom correspondence should be addressed*: S. Shu (sqshu@lbl.gov @lbl.gov)

October 28, 2013

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

PERTRAN: Genome-guided RNA-seq Read Assembler

Shengqiang Shu

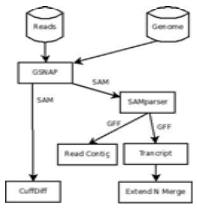
David M. Goodstein and Daniel Rohksar

Abstract

As short RNA-seq reads become a standard, affordable input to any genome annotation project, a sensitive and accurate transcript assembler is an essential part of any gene prediction system. PERTRAN is a pipeline for assembling transcripts from RNA-seq reads which demonstrates higher sensitivity, with fewer fused exons (in most cases), and faster run times compared to other TOPHAT/CUFFLINKS and genome-guided Trinity. PERTRAN shows slightly lower specificity with increased gene fusions in some cases, discussed below. SAM files generated from PERTRAN can be used to compute expression level by cuffdiff and result is comparable to that from TOPHAT.

Introduction

The ability to produce comprehensive RNA-seq data at low cost has greatly benefitted gene structure prediction, facilitating the prediction of UTR and alternative splice isoforms, compared to cases where prediction was dominated by protein homology and relatively sparse EST sets. The accurate and time-efficient construction of RNA-seq based transcript assemblies is an absolute requirement for taking advantage of RNA-seq data in gene prediction, and is an area of active research, with several programs already available. The JGI Plant Program is responsible for the initial and subsequent re-annotation of several large plant genomes, and has been focusing on improving the performance of reference-genome-guided RNA-seq transcript assemblers that are generally better than *de novo* assemblers (Lu, 2013). Using TOPHAT/CUFFLINKS (Trapnell, 2010) in gene annotation can be a challenge due to lower sensitivity and fused genes (Brian Haas, personal communication). Genome-guided Trinity assembler in beta (Haas, 2013) was available near the end of this study. Here we report a pipeline (work flow below) implemented in PERL consisting of 3 scripts: GSNAP SAM parser, transcript assembler and reassembler, and a read aligner, GSNAP (Wu, 2010), which is the best RNA-seq read aligner (Grant, 2011) in the absence of transcriptome assemblies. Parallelization was achieved by splitting a FASTAQ file into multiple GSNAP jobs while trimming off very low quality bases and assembling transcripts in many overlapping segments per chromosome. Computes are managed by our in-house compute management system that can run on several types of commodity clusters and high-core-number servers.



Pipeline Work Flow

Contact: S. Shu (sqshu@lbl.gov)

Methods

Pipeline is depicted in a chart in the lower left. Reads are first aligned to reference genome by GSNAP.

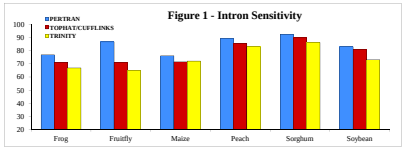
Each mapped read is represented by its genomic position, with pairing position and/or with splice position(s), if any, in GFF2 format (using gap_next/gap_prev or spliced_next attribute for pairing or splice respectively). The number of mapped reads is recorded as expression level for each genomic position. Gap position and number of pairs are recorded if distance between 2 paired reads is bigger than a threshold (gap length threshold), normally read pair inner distance average plus 1 variance, and so are introns from splice-aware-aligned reads (spliced reads). Any continuous block with expression across a chromosome is a raw read contig. The raw contig is strandless unless reads are sequenced with a strand protocol. A contig could be split into multiple contigs if expression level in a gap(s) or intron(s) suggests intron(s) in it. Each position expression of a raw contig is deducted from the average expression in gap(s) or intron(s) and a continuous block of expression level below zero is a putative intron. Splice sites between any 2 contigs of inner distance <= maximum intron length (a user parameter) with the number of gaps/introns higher than a threshold (default: 3 gaps or 1 spliced read for seeding) are computed using position weight matrix either from priori spliced site sequences or canonical splice empirical distribution (donor: 0.993 GT, 0.006 GC, 0.001 AT, acceptor: 0.999 AG, 0.001 AC). Splice site score is computed from the position weight matrix and the score is penalized by its position relative to contig edge with +9 from left side (or -9 from right side) as center with no penalty. Score is set to an arbitrary high number if splice site is supported by spliced reads. An intron is made from 2 sensical splice sites with the highest score total for any given 2 contig pair and additional intron(s) if supported by spliced reads. For >1 introns per contig pair to be retained, all introns must be supported by spliced reads. Transcript graphs with exons as nodes and introns as edges are made. Each transcript graph with alternative path as alternative transcript is trimmed if not meeting minimum of minor transcript ratio to major (default 10%) in the number of spliced reads, in the number of gaps with minimum of 0.1% of overall spliced reads per site in the transcript or minimum spliced read threshold, or in overall expression level with spliced reads with minimum spliced read threshold. Overlapping introns on opposite strands of lower number of spliced reads or gaps are removed unless the number of spliced reads exceeds a threshold (default is number of spliced reads for seeding plus 2). Two terminal overlapping exons on opposite strand are shrunk if expression level drops by >2X in the middle so 2 exons would not overlap using the lowest level as the split point.

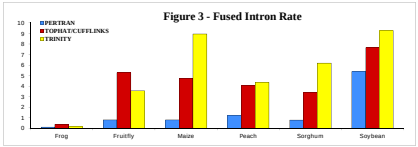
For computation speed, overlapping chromosome segments (default 10M BP with 1K BP overlapping) are used in transcript computation resulting in a few overlapping partial transcripts. Overlapping transcripts on the same strand are merged if their ends are different and their overlapping exons match. Terminal exons are extended if supported by expression using raw contigs and the extension wouldn't run into another assembly. Furthermore, terminal exons are shrunk from either end until base expression level exceeds exon average expression level times half of minor transcript ratio (default is 5% of average expression level).

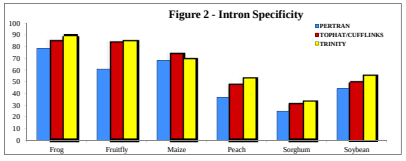
Comparison data were obtained by comparing computed transcripts with a reference gene set that had no input from RNA-seq. A spliced gene is a gene with spliced CDS and is recovered if one of spliced CDS transcripts in a gene is 100% matched in both CDS exons and intron(s) with one computed transcript. Fused genes are transcripts overlapping with >1 genes and aren't penalized for spliced gene sensitivity. A fused exon is a computed exon overlapping with >1 exons in the reference gene set that does not have such an exon in alternative transcripts.

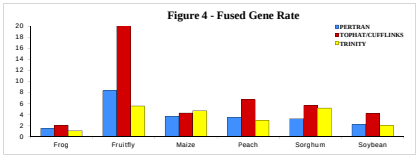
Results

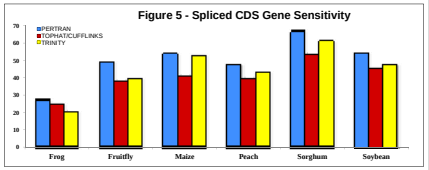
PERTRAN is more sensitive in recovering introns (Fig. 1) with a fewer fused exons (Fig. 3) than other 2, more so compared to Trinity, while intron specificity (Fig. 2) is lower with slightly higher fused gene rate in some cases (Fig. 4). Sensitivity in recovering spliced genes is higher than other 2 (Fig. 5) while specificity in this category is comparable among three (Fig. 6). PERTRAN has the shortest completion time by far (Fig. 7). Expression level in FPKM by cuffdiff (Trapnell, 2010) from GSNAP SAMs is highly correlated with that from TOPHAT for all soybean tissues except for one tissue where there are a few outliers, most of which are due to reads found only by GSNAP and one only by TOPHAT (Fig. 8, showing only 2 tissues).

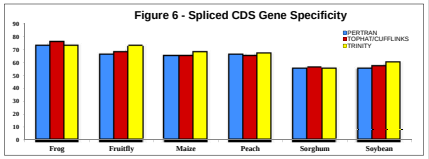


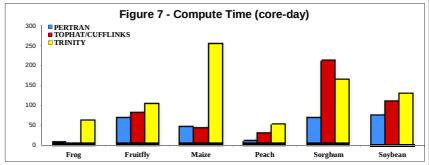


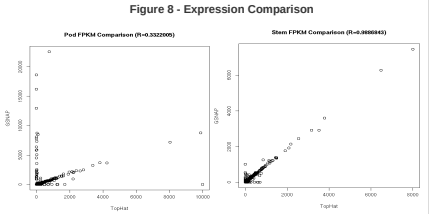












References

Hass, BJ. (2013) http://trinityrnaseq.sourceforge.net/genome_guided_trinity.html

Grant, GR, Farkas, MH, Pizarro, AD, Lahens, NF, Schug, J, Brunk, BP, Stoeckert, CJ, Hogenesch, JB, Pierce, EA. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**:2518-28

Lu, B, Zheng, Z, Shi, T. (2013) Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci China Life Sci*, 56:143-155

Trapnell C, Williams, BA, Pertea, G, Mortazavi, AM, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, B, Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**:511-515

Wu, TD, and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics,* **26**:873-881

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.