Scripture Documentation

Description:	Method for transcriptome reconstruction that relies on RNA-seq reads and an assembled genome to build a transcriptome *ab initio.*
Author:	Manuel Garber and Mitchell Guttman, mgarber@broadinstitute.org
Module Version:	1
Contact:	Chris Williams, gp-help@broadinstitute.org

Summary

Scripture is a comprehensive method for *ab initio* reconstruction of the transcriptome of a mammalian cell that uses gapped alignments of reads across splice junctions and reconstructs reads into statistically significant transcript structures.

Scripture makes use of aligned reads to the genome. The quality of transcriptome reconstruction is highly dependent on the quality of the aligner used. It is critical to use a spliced aligner, which is an aligner that can map reads across exon-exon junctions. Scripture is agnostic to the exact aligner used. However, testing has shown that TopHat works well in providing input for the Scripture module.

For a full explanation of the Scripture algorithm, see the following reference.

References

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*. 2010;28:503-510. (http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1633.html)

Parameters

Name	**Description**
alignment.file (required)	Alignment file in SAM or BAM format, sorted by start position.

alignment. index.file (required)	Index file in SAI or BAI format. The index file must have the same root name as the SAM/BAM alignment file. For example, if the alignment file is foo.sam, the index file must be named foo.sai or foo.sam.sai. If the alignment file is foo.bam, the alignment file must be named foo.bai or foo.bam.bai. When specifying this parameter as a file path rather than uploading the file, the index file should be in the same location as the alignment file. Otherwise, the index file and alignment file will be copied to the same temporary directory, which make take several minutes for large files.
chromosome. size.file (required)	A two-column, tab-separated file which lists the chromosome name followed by the chromosome size. Each chromosome should appear on a separate line.
chromosome (required)	The selected chromosome. For example, chr19.
chromosome. sequence.file (required)	The chromosome sequence in FASTA format.
output.prefix (required)	A label that will be used to name output files.

Output Files

1. <output.prefix>.bed

   This is a BED file for all reconstructed transcripts. For more information about the BED file format, see the UCSC FAQ: http://genome.ucsc.edu/FAQ/FAQformat.html

2. <output.prefix>.enrichment.gct

   This GCT file contains the enrichment score for each transcript. The enrichment score is the ratio of the observed number of reads to the expected number of reads for transcript length.

3. <output.prefix>.totalreads.gct

   This GCT file contains the total number of reads across each transcript.

4. <output.prefix>.readsperbase.gct

   This GCT file contains the mean number of reads per base for each transcript.

5. <output.prefix>.rpkm.gct

   This GCT file contains the reads per kilobase of exon model per million mapped reads (RPKM) value for each transcript.

6. <output.prefix>.segments

   This file contains all the data in the above five output files, in addition to 4 additional values for each transcript: the FWER-corrected p-value for the observed read count across the transcript; lambda, the number of reads per base across transcript genomic loci rather than spliced transcript; the transcript length; and the nominal p-value for the observed read count across the transcript.

7. introns.bed

   This is BED file contains the coordinates of all introns.

8. <output.prefix>.segments.dot

   This file is the transcript graph constructed by Scripture. This file is in DOT format; for more information, see the DOT specification: http://www.graphviz.org/pdf/dotguide.pdf. This file can be used to visualize the transcript graph in GraphViz (http://www.graphviz.org).

Platform Dependencies

Module type:	RNA-seq
CPU type:	any
OS:	Macintosh, Linux
Language:	Java (minimum version 1.5)