

ScripturePrealigned Documentation

Description: Performs *ab initio* transcriptome reconstruction starting from unsorted, aligned reads. Uses the SortSam and Scripture modules.

Contact: Chris Williams, gp-help@broadinstitute.org

Summary

The ScripturePrealigned is set up to take reads in SAM or BAM format.

The aligned reads are sorted by chromosome and start position, indexed, and then used to reconstruct a mammalian transcriptome using the following modules:

- SortSam sorts a SAM or BAM file and outputs a sorted BAM file and an index BAI file.
- Scripture is a comprehensive method for *ab initio* reconstruction of mammalian transcriptomes. This module uses gapped alignments of reads across splice junctions to reconstruct statistically significant transcript structures.

Parameters

Name	Description
input.file (required)	Aligned paired or unpaired reads in SAM or BAM format.
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>.readspibase.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 RPKM value for each transcript. The RPKM is the number of reads per kilobase of exon model per million mapped reads.
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>).

Example Data

See the Scripture walkthrough example:

http://www.broadinstitute.org/software/scripture/Walkthrough_example

Platform Dependencies

Module type:	Pipeline
CPU type:	any
OS:	Macintosh, Linux
Language:	Perl, C++, Java (minimum version 1.6)