ScripturePipeline Documentation

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

Contact: GenePattern Team, gp-help@broadinstitute.org

Summary

The ScripturePipeline takes unaligned reads in FASTA or FASTQ format and aligns them, sorts them, and reconstructs a mammalian transcriptome using the following modules:

- TopHat is a fast splice junction mapper for RNA-seq reads. It aligns RNA-seq reads to mammalian-sized genomes and then analyzes the mapping results to identify splice junctions between exons. The software is optimized for reads 75bp or longer.
- SortSam sorts a SAM or BAM file, in this case BAM, and outputs the 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
prebuilt.bowtie.index	An indexed genome. A number of pre-built indexes are available: • *A. thaliana* • *B. taurus* • *E. coli* • *C. elegans,* WormBase, WS200 • *H. sapiens*, UCSC hg19 • *H. sapiens*, UCSC hg18 • *M. musculus*, UCSC mm9 • *M. musculus*, UCSC mm8 • *M. musculus*, NCBI 37 • *S. cerevisiae* If this list does not include the genome the user requires, an indexed genome can be generated using Bowtie.indexer. **Either** a prebuilt **or** a custom Bowtie index **must** be specified.

custom.bowtie.index	A ZIP archive containing Bowtie index files. **Either** a prebuilt **or** a custom Bowtie index **must** be specified.
reads.pair.1 (required)	Unpaired reads file or first mate for paired reads. This can be a file in FASTA or FASTQ format, a ZIP archive containing FASTA or FASTQ files, or a directory that is accessible to the GenePattern server containing FASTA or FASTQ files. For more information on the FASTA format, see the NIH description here: http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml. For more information on the FASTQ format, see the specification here: http://nar.oxfordjournals.org/content/early/2009/12/16/nar.gkp1137.full.
reads.pair.2 (optional)	Second mate for paired reads. This can be a file in FASTA or FASTQ format, a ZIP archive containing FASTA or FASTQ files.
mate.inner.dist (optional)	The expected mean inner distance between mate pairs. For example, for paired-end runs with fragments selected at 300 bp, where each end is 50 bp, you should set this to be 200. Default: 50
mate.std.dev (optional)	The standard deviation for the distribution on inner distances between mate pairs. This does not have to be specified for paired end reads.
library.type (optional)	Library type for strand specific reads. Options include: • Standard Illumina • dUTP, NSR, NNSR • Ligation, Standard SOLiD
integer.quals (optional)	Quality values are space-delimited integer values; this becomes the default when you select *Yes* for *colorspace reads.* Default: No
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 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)