Fpkm_trackingToGct Documentation

Description: Takes a Cufflinks FPKM_tracking file and converts it to GCT format.

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

Summary

This module converts Cufflinks FPKM_tracking format files to GCT format, which can be used as input by many other GenePattern modules. The Fpkm_trackingToGct module extracts expression data from a FPKM_tracking file and writes it to a GCT file, a tab-delimited file format used to describe expression datasets that is commonly used in GenePattern and other tools.

The FPKM_tracking file format is a tab-delimited format produced by some of the Cufflinks GenePattern modules (Cufflinks and Cufflinks.cuffdiff). In the FPKM_tracking file, RNA-seq expression values are in units of **F**ragments **P**er **K**ilobase of exon model per **M**illion mapped fragments (FPKM). The text elements in the FPKM_tracking file used for the row names and descriptions in the GCT file can be selected from: Tracking ID, Nearest Reference Transcript ID, Gene Name, TSS ID, and Locus.

Any non-numeric values (such as NaN) in FPKM columns will be converted to NA. These can be filtered from the GCT in a post-processing step using the RemoveMissingValues module, or imputed using the ImputeMissingValuesKNN module.

Links

Cufflinks documentation: http://cufflinks.cbcb.umd.edu/

Parameters

Name	Description
input file	The FPKM_tracking file from a Cufflinks module to be converted.
row labels	The column of identifiers in the file that will be used for row names in the GCT file. Options include: • Tracking ID (default) • Nearest Reference Transcript ID • Gene Name • TSS ID • Locus

row descriptions	The column of identifiers in the file that will be used for row descriptions in the GCT file. Options include: • Locus (default) • Tracking ID • Nearest Reference Transcript ID • Gene Name • TSS ID • Class Code
filter rows	If set to *no*, all rows in the input will appear in the output file. If set to *yes*, rows whose row label is specified as "-" in the input file will not appear in the output file. Default: yes.
output prefix	A label that will be used to name the GCT output file.

Input Files

1. <filename>.fpkm_tracking

   This FPKM_tracking file contains RNA-seq expression values as output by the Cufflinks or Cufflinks.cuffdiff GenePattern modules.

Output Files

1. <output.prefix>.gct

   This GCT file contains the expression values and the selected output column values, row names, and row descriptions from the FPKM_tracking file.

Platform Dependencies

Module type:	RNA-seq
CPU type:	any
OS:	any
Language:	Perl

GenePattern Module Version Notes

Date	Version	Description
09/28/12	2	Added data columns parameter and fixed error with specifying non default values for row labels and row descriptions
4/9/2013	3	Bug fix for the first column of the fpkm_tracking file always included as a sample column in the output file
5/2/2013	4	Restricted output to only the FPKM values.
5/10/2013	5	Updated to output scientific values and 0s as output by Cufflinks