txiimport.DESeq2.Normalize Documentation

Description: Imports RNA-seq quantifications using the tximport package and outputs files for downstream analysis.

Author(s): Charlotte Soneson (tximport), Mike Love (DESeq2). Wrapped as a module by Anthony S. Castanza, Mesirov Lab, UCSD School of Medicine.

Contact: genepattern.org/help

Summary: Imports RNA-seq quantifications using the tximport package and outputs a DESeq2 "normalized counts" file in .GCT format suitable for use with GSEA, and/or a TPM .GCT suitable for ssGSEA.
If a Sample Info file is provided containing assignments of samples to phenotypes, gene level differential expression will be calculated using DESeq2 for the two phenotypes.

Parameters:

Name	Description
Quantifications	Output files from the respective pipeline. Files must be named with the sample ID and the appropriate extension for the quantification tool. Examples: ".genes.results" files output from RSEM: Sample_1.genes.results ".quant.sf" or ".quant.sf.gz" files output from Salmon/Sailfish: Sample_1.quant.sf/Sample_1.quant.sf.gz ".abundance.h5" or ".abundance.tsv" files output from Kallisto: Sample_1.abundance.h5/Sample_1.abundance.h5.gz /Sample_1.abundance.tsv/Sample_1.abundance.tsv.gz
Sample Info	Optional tab delimited text file containing Sample ID to Phenotype mappings. Sample IDs should match exactly the ID used in the input file (without the file extension). If provided, this file will be used to compute deferentially expressed genes using DESeq2. Covariates can be provided as additional "Factor" columns. The first factor column (the second column in the table) will always be used as the primary factor for the differential expression analysis. Column headers are recommended with the first column being titled SampleID, and the remaining columns being named by the factor represented (genotype, sex, etc.)
Quant Type	The specific quantification tool that generated the data to be imported.
Transcriptome Database	For Salmon/Sailfish/Kallisto the transcriptome used to

	quantify the data must be provided in GTF/GFF3 format. The transcriptome may be provided as a gzip (.gz). This information is ignored for importing RSEM data, but can still be supplied if annotating genes if Annotate DEGs is TRUE.
Output Normalized Counts	If TRUE, Compute count normalization with DESeq2 and output the normalized matrix. This is the ideal output for compatibility with GSEA.
Output TPM	If TRUE, Transcripts per million (TPM) quantifications will be extracted and output. This is the ideal output for compatibility with ssGSEA.
output file base	The base name of the output file(s). File extensions will be added automatically.

Advanced Parameters:

Name	Description
Reverse Sign	If computing differentially expressed genes, by default, Log2(FC) will be calculated for the first Factor level vs the second Factor level (Phenotype_A vs. Phenotype_B). Setting Reverse.Sign = TRUE will reverse the direction of the calculation and Log2(FC) will be calculated for the second Factor level vs the first Factor level (Phenotype_B vs. Phenotype_A).
Annotate DEGs	If a Sample Info file and a Transcriptome Database have been provided, setting Annotate DEGs to TRUE will attempt to add gene symbols to the computed differential expression analysis. Note: This does not affect the .GCT files which will still output only the underlying Gene IDs.
Split Identifiers	Some pipelines produce a gene label that combines a reference ID and a Gene Symbol joined by an underscore (e.g. ENSG00000141510_TP53). **TRUE** will attempt to split these into the reference ID and the Gene Symbol. The reference ID will be used as the NAME column of the GCT and the Gene Symbol will be used as the Description. **FALSE** will leave the identifiers as-is.
Min Count Filter	Filtering parameter to remove low/non-expressed genes, by default genes with >=1 count across all samples are kept. Setting Min Count Filter = 0 disables filtering entirely.
random seed	Integer used to configure the workspace. Setting this to a constant ensures uniformity for any operations that use random numbers.

Output File(s):

Name	Description
<output file base>.Normalized.Counts.gct	If Output Normalized Counts is TRUE, this file contains the DESeq2 Normalized counts in GCT format.
<output file base>.TPM.gct	If Output TPM is TRUE, this file contains the TPM normalization extracted from the input data in GCT format.
<output file base>.cls	If a Sample Info file is provided the first factor level will

	be used to construct a categorical CLS file suitable for use with GSEA.
<output file base>.Differential.Expression.txt	If a Sample Info File is provided this result will contain the DESeq2 differential expression analysis results. By default, Log2(FC) will be calculated for the first listed factor level vs the second listed factor level.

Module Language: R
Source Repository: https://github.com/genepattern/tximport.DESeq2.Normalize/releases/tag/v1
Docker image: jupyter/datascience-notebook:r-4.0.3

Version	Comment
1	Initial release.

Citation(s):

Soneson C, Love MI, Robinson MD (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research*, **4**. doi: 10.12688/f1000research.7563.1.

Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550. doi: 10.1186/s13059-014-0550-8.