Gene expression matrix

The ENCODE gene expression matrix is obtained by collecting into a single file the gene quantification files produced by the ENCODE3 long RNA-seq pipeline. The matrix is created using in-house scripts, and provided in both TSV and JSON formats. The current version of the matrix includes TPM and FPKM values for all genes with no additional filtering. Values from each biological replicate of the long RNA-seq experiment are preserved. Different matrices are created for different genome assemblies, annotation versions and type of experimental assay.

Experiments are grouped into different expression matrices according to the following criteria:

1. Species
    a. *H. sapiens* (human)
    b. *M. musculus* (mouse)
2. Combination of reference genome assembly and reference annotation version, *e.g.*:
    a. hg19 and GENCODE v.19 (human)
    b. GRCh38 and GENCODE v.24 (human)
    c. mm10 and GENCODE v.M4 (mouse)
3. Assay type
    a. RNA-seq (RNA-seq, polyA mRNA RNA-seq, polyA depleted RNA-seq)
    b. RBP disruption (CRISPR genome editing followed by RNA-seq , shRNA knockdown followed by RNA-seq)
    c. single cell isolation followed by RNA-seq

Tabular formatted matrix.

The tabular formatted matrix (tsv) contains the gene names and corresponding Ensembl IDs, as rows and all replicates of the same experiment as columns. The cells of the matrix contains both TPM and FPKM values. Column identifiers in the header contain the experiment identifier, followed by underscore ("_") and the bio-replicate number.

Example for two bio-replicates:
ExperimentId_1,2, where
"ExperimentId" - is the replicate.experiment.accession id, *e.g.* ENCSR000AAA
"1" - replicate.biological_replicate_number is 1
"2" - replicate.biological_replicate_number is 2

Note: For the single cell experiments column identifiers in the header contain the experiment identifier and the technical replicate number. Example: *ENCSR673UIY_1,2,3,4,5,6,7,8,9,*

The values in the cell contains two strings, one for TPM values and another for FPKM values, separated by underscore ("_"). Each string contains values for each replicate, separated by colon (":").

Format for two bioreplicates:
TPM1:TMP2_FPKM1:FPKM2, where
“TPM1” is the gene’s TPM in the ExperimentId, biological replicate 1
“TPM2” is the gene’s TPM in the ExperimentId, biological replicate 2
“FPKM1” is the gene’s FPKM in the ExperimentId, biological replicate 1
“FPKM2” is the gene’s FPKM in the ExperimentId, biological replicate 2

Example

gene_id	gene_name	ENCSR000AAA_1,2,
ENSG00000000419.8	DPM1	9.02:3.91_22.27:17.71

JSON formatted matrix

The JSON formatted file contains information stored as a JSON array of objects containing information about genes (gene name, gene Ensembl ID) and gene expression values across the samples. Gene expression values are stored again in a JSON array of objects named “`expression_values`”. Bio-replicate expression values are stored in the same element of the gene expression array.

Format for two bioreplicates:

[
{ "gene_name": "GENE1", "ensembl_id": "ENSG0000000...",
expression_values: [
{ "dataset": "ExperimentId", "rep1_tpm:": 0.00, "rep2_tpm”: 0.00, "rep1_fpkm": 0.00, "rep2_fpkm": 0.00}, #end of ExperimentId record
{"dataset": "ExperimentId_2", ...}
] } #end of GENE1,
...
] , where

“ExperimentId” - is replicate.experiment.accession id, *e.g.* ENCSR000AAA
rep1_tpm - is the TPM value in the ExperimentId, biological replicate 1 for the gene GENE1
rep2_tpm - is the TPM value in the ExperimentId, biological replicate 2 for the gene GENE1
rep1_fpkm - is the FPKM value in the ExperimentId, biological replicate 1 for the gene GENE1
rep2_fpkm - is the FPKM value in the ExperimentId, biological replicate 2 for the gene GENE1

example

[...
{ "gene_name": "DPM1", "ensembl_id": "ENSG00000000419.8", "expression_values": [{ "dataset": "ENCSR000AAA", "rep1_tpm": 9.02, "rep2_tpm": 3.91, "rep1_fpkm": 22.27, "rep2_fpkm": 17.71 },
...
]

Contact

Questions, comments and requests should be addressed to Anna Vlasova (anna.vlasova@crg.eu) or Julien Lagarde (julien.lagarde@crg.eu)