RNAseq gene and transcript expression matrices

The ENCODE gene expression matrices are obtained by collecting into a single file the gene quantification files produced by the ENCODE3 long RNA-seq pipeline. The matrices are created using in-house scripts, and provided in tabular separated TSV format. The current version of 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.

Header 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 (":").

Cell format example 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

Contact

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