EMBL-EBI

RNA-seq Analysis API – version 1.3[1]
31st July 2017

A simple RESTful API to access analysis results of all public RNA-seq data for 291 species in European Nucleotide Archive.
N.B. Changes from version 1.2 to 1.3 are highlighted in blue.

Authors: Robert Petryszak[1,*], Nuno A. Fonseca[1], Anja Füllgrabe[1], Laura Huerta[1], Maria Keays[1], Y. Amy Tang[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, EMBL-EBI, Hinxton, UK
* Contact: rnaseq@ebi.ac.uk

This document describes the RESTful API that was developed to provide easy access to the results of analysis of public RNA-seq data in European Nucleotide Archive (ENA). The analysis of each sequencing run was performed by the EMBL-EBI's Gene Expression Team using the iRAP pipeline. Firstly quality-filtered reads were aligned to the latest genome reference from Ensembl via TopHat 2 (and STAR for large genomes, e.g. wheat), then the resulting BAM file was converted to CRAM format. Finally expression of genes and exons in the corresponding Ensembl GTF file was quantified using HTSeq and DEXSeq respectively.

We have extended the iRAP pipeline to analyse public RNA-seq data in the most 'RNA-seq data-rich' organisms present in ENA. To date, data in 291 organisms have been analysed, including:

25 in Ensembl
44 in Ensembl Plants
87 in Ensembl Fungi
4 in Ensembl Metazoa
17 in Ensembl Protists
114 in WormBase ParaSite

The pipeline analyses sequencing runs as soon as they become public in ENA, with the results available via the RESTful API shortly after. The annotation of the sequencing metadata to Experimental Factor Ontology (EFO) is performed at scale for each new release of EFO, via a tool called Zooma. The Zooma knowledgebase from which annotations are derived is based on the manual curation of ENA's sequencing metadata in ArrayExpress and Expression Atlas, performed by the curators in the Gene Expression Team. If you have any questions, problems using the API or would like us to add to the analysis new organisms of interest please contact the email address above.

[1] If you use this data in your research or service, please reference this publication. Thank You

This API has also been incorporated into BioServices Python Package and CPAN Perl Package .

Analysis Results Per Run

Format

Item	Description
URL PATTERN	http://www.ebi.ac.uk/fg/rnaseq/api//FORMAT/MAPPING_QUALITY/getRun...
FORMAT	tsv or json
MAPPING_QUALITY	Minimum percentage of reads mapped to genome reference
ORGANISM	See Ensembl, Plants, Fungi, Metazoa, Protists and WormBase ParaSite
CONDITION	Check if term exists in EFO, e.g. cancer

Example Calls to Retrieve Individual Run Data

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/70/getRunsByOrganism/oryza_longistaminata
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/70/getRunsByOrganismCondition/homo_sapiens/central nervous system
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/90/getRunsByStudy/SRP033494
http://www.ebi.ac.uk/fg/rnaseq/api/json/70/getRun/SRR1042759

Returned Fields

Field	Description
ASSEMBLY_USED	Genome reference assembly name
BIOREP_ID	ENA Run ID or a unique label for technical replicates in RUN_IDS
ENA_LAST_UPDATED	Date ENA record for any RUN_IDs was last updated
CRAM_LOCATION	FTP location of the CRAM file
BEDGRAPH_LOCATION	FTP location of the bedGraph file
BIGWIG_LOCATION	FTP location of the BigWig file
LAST_PROCESSED_DATE	Date any RUN_IDs were last analysed
ORGANISM	Organism of samples in SAMPLE_IDS
MAPPING_QUALITY	Percentage of reads mapped to the genome reference
REFERENCE_ORGANISM	Genome reference organism
RUN_IDS	List of ENA Run ID's corresponding to BIOREP_ID
SAMPLE_ATTRIBUTE_TYPE	Matched sample attribute type
SAMPLE_ATTRIBUTE_VALUE	Matched sample attribute value
SAMPLE_IDS	BioSamples DB ID's corresponding to BIOREP_ID
STATUS	Processing status in our analysis pipeline
STUDY_ID	ENA Study ID

Analysis Results Per Study

Format

Example Calls to Retrieve Individual Run and Study Data

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getStudiesByOrganism/arabidopsis_thaliana
http://www.ebi.ac.uk/fg/rnaseq/api/json/getStudy/SRP033494

Returned Fields

Field	Description
ASSEMBLY_USED	Genome reference assembly name
GENES_FPKM_COUNTS_FTP_LOCATION	FTP location of gene FPKM counts
GENES_TPM_COUNTS_FTP_LOCATION	FTP location of gene TPM counts
GENES_RAW_COUNTS_FTP_LOCATION	FTP location of gene RAW counts
EXONS_FPKM_COUNTS_FTP_LOCATION	FTP location of exon FPKM counts
EXONS_TPM_COUNTS_FTP_LOCATION	FTP location of exon FPKM counts
EXONS_RAW_COUNTS_FTP_LOCATION	FTP location of exon RAW counts
GTF_USED	GTF file used in expression quantification
LAST_PROCESSED_DATE	Date the run(s) were last analysed
ORGANISM	Organism studied in STUDY_ID
REFERENCE_ORGANISM	Genome reference organism
SOFTWARE_VERSIONS_FTP_LOCATION	FTP location of pipeline tools info
STATUS	Processing status
STUDY_ID	ENA Study ID

Sample Attributes Per Run

Format

Item	Description
URL PATTERN	http://www.ebi.ac.uk/fg/rnaseq/api/FORMAT/getSampleAttributes...
FORMAT	tsv or json

Example Calls to Retrieve Individual Run Data

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByRun/SRR805786
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesPerRunByStudy/SRP020492
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesCoverageByStudy/SRP020492

Example Call to Retrieve Distinct Sample Attributes Across All Runs

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributes

Returned Fields

Field	Description
EFO_URL	URL of EFO term matching VALUE
RUN_ID	ENA Run ID
STUDY_ID	ENA Study ID
TYPE	Sample Attribute Type
VALUE	Sample Attribute Value
NUM_OF_RUNS	Number of runs annotated with TYPE/VALUE
PCT_OF_ALL_RUNS	Runs annotated with TYPE/VALUE, as a percentage of all runs
SAMPLE_IDS	BioSamples DB ID's corresponding to BIOREP_ID

Baseline Expression Per Gene - for Tissue, Cell Type, Developmental Stage, Sex and Strain

Format

Item	Description
http://www.ebi.ac.uk/fg/rnaseq/api	http://www.ebi.ac.uk/fg/rnaseq/api/FORMAT/MIN_NUMBER_OF_RUNS/getExpression…
FORMAT	tsv or json
MIN_NUMBER_OF_RUNS	Reported expression is a median of expressions (TPM) across all runs corresponding to a given condition. This filter excludes conditions with less that the specified minimum number or runs.
ORGANISM	Species of the gene symbol provided ('**any**' for all species)
GENE_SYMBOL	Gene symbol in ORGANISM to select expression of

Example Calls to Retrieve Baseline Expression Per Gene

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/50/getExpression/homo_sapiens/REG1B
http://www.ebi.ac.uk/fg/rnaseq/api/json/0/getExpression/oryza_sativa/BURP7
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/10/getExpression/any/ALB
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/0/getExpression/ENSG00000172023

Example Call to Retrieve All Organisms with Expression Data

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getExpressionOrganisms

Returned Fields

Field	Description
GENE_ID	Ensembl gene identifier
ORGANISM	GENE_ID's species
MEDIAN EXPRESSION	Median expression value for GENE_ID, aggregated across expressions (TPM) in all sequencing runs corresponding to the reported condition (i.e. tissue, cell type, developmental stage, sex and strain - see below)
COEFFICIENT_OF_VARIATION	Measure of dispersion of individual runs' expressions in around the expression mean across all runs. It is calculated as: (standard deviation) / mean. The lower its value, the more consistent the expression is across multiple runs.
NUM_OF_RUNS	Number of runs corresponding to the reported condition
ORGANISM_PART	Tissue (NA if no value available)
CELL_TYPE	Cell type (ditto)
DEVELOPMENTAL_STAGE	Developmental stage (ditto)
SEX	Sex (ditto)
STRAIN	Strain (NA if no value available or not applicable)
ALL_SAMPLE_ATTRIBUTES	The API link to display all sample attributes associated with runs aggregated for the reported condition, e.g. http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByCondition/1178
REFERENCE_SOURCE	The source of the genome reference used in the analysis (C.f. Ensembl, Plants, Fungi, Metazoa, Protists and WormBase ParaSite)

Mapping Quality Statistics Across All Organisms

Format

Item	Description
URL	http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getOrganismsMappingQuality
FORMAT	tsv or json

Returned Fields

Field	Description
ORGANISM	Organism
MEAN_MAPPING_QUALITY	Average mapping quality across all analysed runs for the organism
STDDEV_MAPPING_QUALITY	Standard deviation of mapping quality across all analysed runs for the organism

Full Mapping Statistics Per Run or Study

Format

Item	Description
URL	http://www.ebi.ac.uk/fg/rnaseq/api/FORMAT/getMappingStatisticsBy…
FORMAT	tsv or json

Example Calls

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getMappingStatisticsByRun/ERR1680082
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getMappingStatisticsByStudy/ERP004375

Returned Fields

Field	Description
STUDY_ID	ENA Study ID
RUN_ID	ENA Run ID
ALL_ENTRIES	Number of entries (reads or alignments) in BAM
VALID_ENTRIES	Number of valid entries in BAM (BAM has a flag to say that an entry is valid or not)
DUPLICATE ALIGNMENTS	Number of duplicate alignments (in RNA-seq it should always be 0, or there's an error)
ALL_ALIGNMENTS	Number of alignments
SPLICED_ALIGNMENTS	Number of alignments that are spliced (NB. a multi-mapped read may result in more than one spliced alignment)
READS_SPLICED	Number of reads that have a splice alignment (read: they must have aligned to make it possible to say if the alignment is spliced or not)
PAIRED	Number of paired entries (with both mates present) in bam file, both aligned and unaligned
PAIRED_MAPPED	Number of paired entries (with both mates present) in bam file, aligned only
PAIRED_MAPPED_MATE1	Number of alignments in which mate 1 is mapped
PAIRED_MAPPED_MATE2	Number of alignments in which mate 2 is mapped
READS_UNMAPPED	Number of reads unmapped
READS_MAPPED	Number of reads mapped
ALIGNMENTS_WITH_0MISMATCH	Number of alignments with number of mismatches = 0
ALIGNMENTS_WITH_1MISMATCH	Number of alignments with number of mismatches = 1
ALIGNMENTS_WITH_2MISMATCH	Number of alignments with number of mismatches = 2
ALIGNMENTS_WITH_GE3MISMATCH	Number of alignments with number of mismatches >= 3
ALIGNMENTS_ON_PLUS_STRAND	Number alignments on the plus strand
ALIGNMENTS_ON_MINUS_STRAND	Number alignments on the minus strand
UNIQUELY_MAPPED_READS	Number of uniquely mapped reads
MULTIMAP_READS	Number of reads that map to more than one locus

READS_ALIGNED_TO_1LOCUS	Number of reads that align uniquely (i.e. number of multimaps = 1)
READS_ALIGNED_TO_GE2_LOCI	Number of reads that align to 2 or more loci
READS_ALIGNED_TO_GE10_LOCI	Number of reads that align to 10 or more loci
READS_ALIGNED_TO_GE20_LOCI	Number of reads that align to 20 or more loci
ALIGNMENTS_WITH_1SPLICEJNCT	Number of alignments that have 1 splice junction
ALIGNMENTS_WITH_2SPLICEJNCTS	Number of alignments that have 2 splice junctions
ALIGNMENTS_WITH_3SPLICEJNCTS	Number of alignments that have 3 splice junctions
ALIGNMENTS_WITH_GE4SPLICEJNTS	Number of alignments that have 4 or more splice junctions
COVERAGE	Coverage or depth of the sequencing run - defined as average (read length * the number of reads mapped) / total length of genes in the run's species

Full FASTQ Information Per Run or Study

Format

Item	Description
URL	http://www.ebi.ac.uk/fg/rnaseq/api/FORMAT/getFastqInfoBy…
FORMAT	tsv or json

Example Calls

URL
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getFastqInfoByRun/ERR1680082
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getFastqInfoByStudy/ERP004375

Returned Fields

Field	Description
STUDY_ID	ENA Study ID
RUN_ID	ENA Run ID
NUM_READS_1	Number of reads in mate 1 of paired-end, or in single-end, library
NUM_READS_2	Number of reads in mate 2 of paired-end library
READ_LENGTH_1	Read length in mate 1 of paired-end, or in single-end, library
READ_LENGTH_2	Read length in mate 2 of paired-end library
READ_LENGTH	Read length in single-end library, or across two mates of a paired-ed library
QUALITY_RANGE_MIN	Minimum FASTQ quality encoding value
QUALITY_RANGE_MAX	Maximum FASTQ quality encoding value
QUALITY_ENCODING	FASTQ quality encoding value chosen based on QUALITY_RANGE_MIN and QUALITY_RANGE_MAX
INSERT_SIZE	Insert size used for mapping the paired-end library
STANDARD_DEV	Standard deviation used for mapping the paired-end library

Acknowledgements

The initial work to develop this API and perform the RNA-seq was funded by the BBSRC, for which we express our gratitude. We would also like to thank the Non-vertebrate Genomics Team for obtaining the funding and their work on displaying the resulting CRAM files in Ensembl Plants track hubs; to the European Nucleotide Archive Team for facilitating access to the raw RNA-seq data; and to the Samples, Phenotypes and Ontologies Team for the provision of tools for retrieval of the sequencing metadata from BioSamples database and up-to-date annotation of sequencing meta-data to Experimental Factor Ontology. Finally, a big thank you is due to members of the Gene Expression Team without whom none of this would have been possible.