Online Repositories

RNA-seq data analysis

Paulo Czarnewski
https://czarnewski.github.io/czarnewski/index.html

ENSEMBL

ENSEMBL

NBIS SciLifeLab

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species...

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 101 (August 2020)

- Update to human GENCODE 35
- New population frequency data from the Gambian Genome Variation Project
- New genomes: 8 mammals, 10 birds, 1 reptile, 1 amphibian, 4 fish
- New sheep reference genome

More release news on our blog

Search

All species for

Go

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

All genomes

-- Select a species --

Pig breeds

Pig reference genome and 12 additional breeds

View full list of all species

Favourite genomes

Human
GRCh38.p13

Still using GRCh37?

Mouse
GRCm38.p6

Zebrafish
GRCz11

Other news from our blog

- 20 Nov 2020: Training in the Time of Pandemic
- 09 Nov 2020: Job: Genome Annotator (Regulation)
- 02 Nov 2020: Job: Outreach Officer

ENSEMBL

NBIS SciLifeLab

Login/Register

e!Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search Mouse...

Mouse (GRCm38.p6)

Search Mouse (Mus musculus)

Search all categories | Search... | Go

e.g. Cntnap1 or 4:136366473-136547301 or rs27096498 or adipocyte

Genome assembly: GRCm38.p6 (GCA_000001635.8)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCm38 coordinates
- Display your data in Ensembl

View karyotype

Example region

Other reference assemblies

- NCBIM37 (Ensembl release 67)

Other strains

This species has data on 15 additional strains. View list of strains

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

Example variant

Example

Regulation

GEO

Gene Expression Omnibus

GEO: Gene Expression Omnibus

NBIS SciLifeLab

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-10769-x OPEN

Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification

Paulo Czarnewski[1], Sara M. Parigi[1], Chiara Sorini[1], Oscar E. Diaz[1], Srustidhar Das[1], Nicola Gagliani[1,2,3] & Eduardo J. Villablanca[1,3]

Reuse public data

Table 1 Publicly available human data sets used in this paper

Data set ID	Total	Responders	Nonresponders	Ref.
Infliximab:				
GSE12251	23	11	12	13
GSE73661	23	15	8	15
GSE23597	32	7	25	14
GSE16879	24	16	8	12
Sum	102	49	53	
Vedolizumab:				
GSE73661	37	23	14	15
Pediatric UC:				
GSE109142	206	105	101	33

Data sets used for the classification of ulcerative colitis molecular profiles. Only the number of patients used for analysis are shown (inflamed mucosa before receiving any therapy)

Deposited new data to public

Data availability

All the raw data generated in this study were deposited at the Gene Expression Omnibus under assession number GSE131032.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131032

GEO: Gene Expression Omnibus

NBIS SciLifeLab

NCBI Resources How To czarnewski My NCBI Sign Out

GEO Home | Documentation | Query & Browse | Email GEO | My GEO Submissions

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Gene Expression Omnibus

GEO Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession | Search

Getting Started	Tools	Browse Content	
Overview	Search for Studies at GEO DataSets	Repository Browser	
FAQ	Search for Gene Expression at GEO Profiles	DataSets:	4348
About GEO DataSets	Search GEO Documentation	Series:	139393
About GEO Profiles	Analyze a Study with GEO2R	Platforms:	21589
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples:	4036403
How to Construct a Query	Programmatic Access		
How to Download Data	FTP Site		

https://www.ncbi.nlm.nih.gov/geo/

GEO: Gene Expression Omnibus

Release day (usually after acceptance letter from journal)

Dataset description

Experimental design

Linked publication

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Contact: czarnewski | My submissions | Sign Out

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE131032 | GO

Series GSE131032 | UPDATE | Query DataSets for GSE131032

Status	Public on May 11, 2019
Title	Time-series reveals processes underlying colon inflammation and repair
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	To elucidated through an unbiased manner which genes and pathways are differentially regulated during mouse colonic inflammation followed by a tissue regeneration phase. In particular, we took advantage of the widely used dextran sodium sulfate (DSS)-induced model of colitis. This model is one of the few characterized by a phase of damage followed by a phase of regeneration. Therefore, this model gave the possibility to identify also sets of genes essential in the regeneration phase, a key step towards the resolution of the inflammation. In short, mice were exposed to DSS in the drinking water for 7 days, then allowed to recover for the following 7 days. During this period, we collected colonic tissue samples every second day to then be analyzed by RNA sequencing (RNA-seq). Next, we performed a RNA-seq analysis from colonic samples throughout the experiment and computed differentially expressed genes (DEGs) taking the complete kinetics of expression into consideration for p-value estimation using EdgeR.
Overall design	C57BL/6J female mice were treated with 2.5% DSS in order to induce colinic inflammation. 2-3 animals were sacrificed at different time points when the colonic tissue was collected.
Contributor(s)	Czarnewski P, Villablanca EJ
Citation(s)	Czarnewski P, Parigi SM, Sorini C, Diaz OE et al. Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification. *Nat Commun* 2019 Jun 28;10(1):2892. PMID: 31253778
NIH grant(s)	Add grant

GEO: Gene Expression Omnibus

NBIS SciLifeLab

Technology used

Individual sample files (raw counts)

SRA accession

Metadata file

Processed count files

Submission date	May 10, 2019
Last update date	Jul 29, 2019
Contact name	Paulo Victor Czarnewski Barenco
Organization name	Stockholm University
Department	NBIS
Street address	Tomteboda vägen 23
City	Stockholm
ZIP/Postal code	171 65
Country	Sweden

Platforms (1) GPL17021 Illumina HiSeq 2500 (Mus musculus)

Samples (26) GSM3760139 DSSd00_1
More... GSM3760140 DSSd00_2
GSM3760141 DSSd00_3

Relations

BioProject PRJNA542350
SRA SRP197582

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE131032_RAW.tar	27.8 Mb	(http)(custom)	TAR (of TSV)
GSE131032_kallisto_counts.csv.gz	897.1 Kb	(ftp)(http)	CSV
GSE131032_log2_counts_per_million.csv.gz	3.1 Mb	(ftp)(http)	CSV

GEO: Gene Expression Omnibus

NBIS SciLifeLab

NCBI > GEO > **Accession Display** Contact: czarnewski | My submissions | Sign Out

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM3760139 GO

Sample GSM3760139 UPDATE Query DataSets for GSM3760139

Status	Public on May 11, 2019
Title	DSSd00_1
Sample type	SRA
Source name	Colon_DSS_day0_untreated
Organism	Mus musculus
Characteristics	strain: C57BL/6J Sex: Female age_at_dss_day0: 9 weeks old day_of_dss: 00 replicate: 1 group: day00 tissue: Proximal Colon cage: A flowcell: HLNYHBCXX

Metadata info

Platform ID	GPL17021
Series (1)	GSE131032 Time-series reveals processes underlying colon inflammation and repair

Relations

BioSample	SAMN11619125
SRA	SRX5818186

Supplementary file	Size	Download	File type/resource
GSM3760139_KI_PC1606_01.tsv.gz	1.1 Mb	(ftp)(http)	TSV

SRA Run Selector

Raw counts

SRA

Sequence Read Archive

SRA: Sequence Read Archive

NBIS SciLifeLab

NCBI Resources How To czarnewski My NCBI Sign Out

SRA

SRA SRP197582 Search

Create alert Advanced Help

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Access
Public (26)

Source
RNA (26)

Library Layout
single (26)

Platform
Illumina (26)

Strategy
other (26)

Data in Cloud
GS (26)
S3 (26)

File Type
fastq (26)

Clear all

Show additional filters

Summary 20 per page Send to:

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 26

<< First < Prev Page 1 of 2 Next > Last >>

1. GSM3760164: DSSd14_3; Mus musculus; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 26.7M spots, 1.3G bases, 636.1Mb downloads
   Accession: SRX5818211
2. GSM3760163: DSSd14_2; Mus musculus; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 25.2M spots, 1.3G bases, 599.7Mb downloads
   Accession: SRX5818210
3. GSM3760162: DSSd14_1; Mus musculus; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 26.3M spots, 1.3G bases, 627.5Mb downloads
   Accession: SRX5818209
4. GSM3760161: DSSd12_3; Mus musculus; RNA-Seq
   1 ILLUMINA (Illumina HiSeq 2500) run: 27.8M spots, 1.4G bases, 665.1Mb downloads
   Accession: SRX5818208

GSM3760160: DSSd12_2; Mus musculus; RNA-Seq

Filters: Manage Filters

Search in related databases

Database	Access		all
	public	controlled	
BioSample			
BioProject			
dbGaP			
GEO Datasets	1		1

Find related data

Database: Select

Find items

Search details

SRP197582[All Fields]

SRA: Sequence Read Archive

NBIS SciLifeLab

NCBI Site map All databases Search

List of Sequence Read Archive Studies

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

GSM3760164: DSSd14_3; Mus musculus; RNA-Seq (SRR9041115)

Change accession...

Metadata | Analysis | Reads | Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR9041115	26.7M	1.3Gbp	667.0M	51.2%	2019-05-13	public

Quality graph (bigger)

This run has 1 read per spot:

L=50, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX5818211		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	BLAST

Biosample	Sample Description	Organism	Links
SAMN11619097 (SRS4746980)		Mus musculus	PRJNA542350 [Time-series reveals processes underlying colon inflammation and repair]

Bioproject	SRA Study	Title
PRJNA542350	SRP197582	Time-series reveals processes underlying colon inflammation and repair

Show abstract

SRA: Sequence Read Archive

NBIS SciLifeLab

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

GSM3760164: DSSd14_3; Mus musculus; RNA-Seq (SRR9041115)

Change accession...

Metadata | Analysis | Reads | Data access

SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Egress	Access Type
run	651,383 Kb	NCBI	https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos2/sra-pub-run-15/SRR9041115/SRR9041115.1	worldwide	anonymous
		AWS	s3://sra-pub-run-5/SRR9041115/SRR9041115.1	s3.us-east-1	aws identity
		GCP	gs://sra-pub-run-3/SRR9041115/SRR9041115.1	gs.US	gcp identity

Egress and Access: what does it mean?

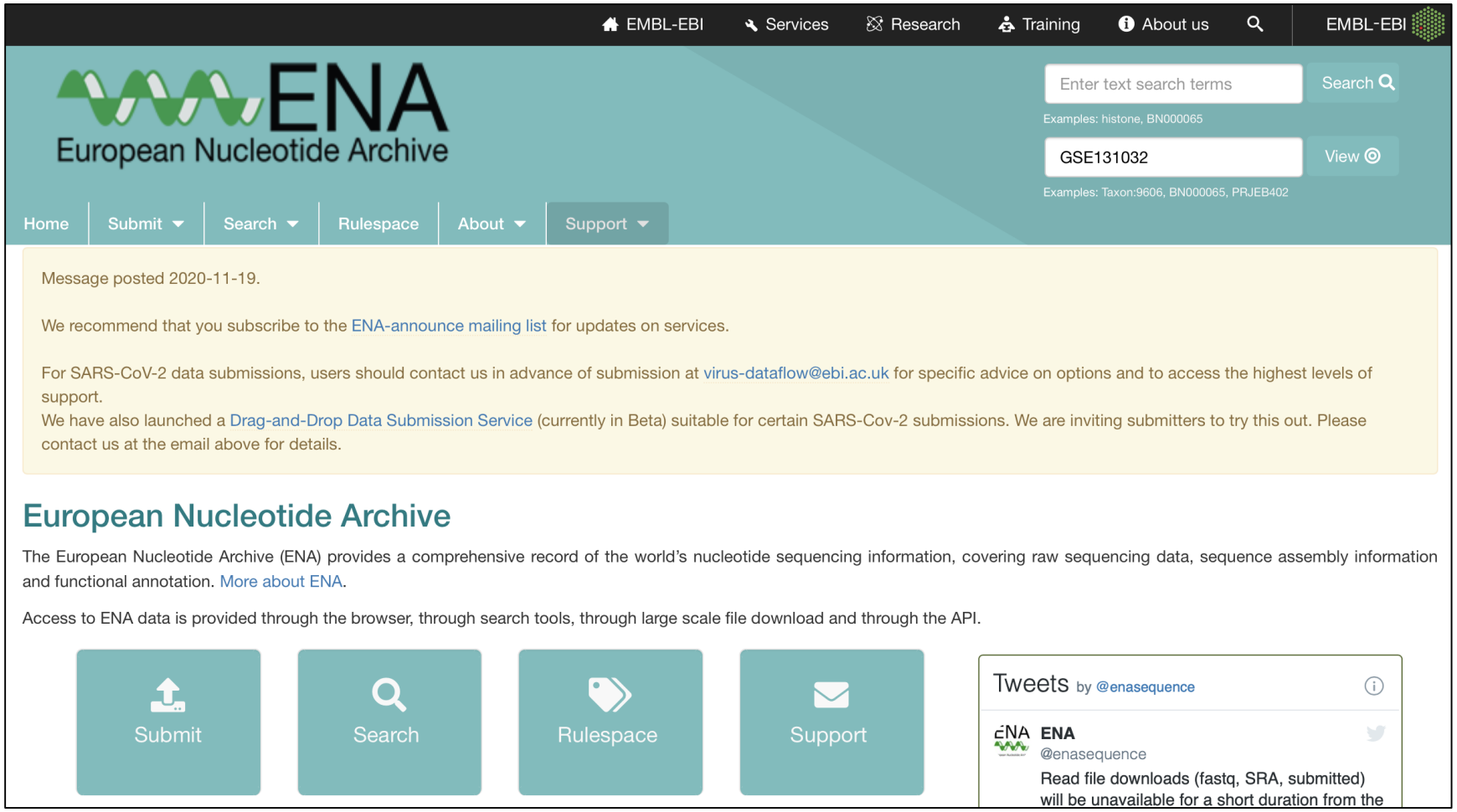
Why is SRA data in the cloud?

ENA

European Nucleotide Archive

ENA: European Nucleotide Archive

Similar to SRA, but in Europe.



ENA: European Nucleotide Archive

ENA is also linked to samples deposited in SRA.

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

ENA
European Nucleotide Archive

GSE131032 Search
Examples: histone, BN000065
Enter accession View
Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Text Search

Uses EBI Search to perform a free text search across ENA data. For more detailed usage please refer to the help & documentation section.

Search term: GSE131032 Search

Search results for GSE131032

- **Study**
  - Study (1)
  - Study (Sequence) (1)

Study	
SRP197582	Time-series reveals processes underlying colon inflammation and repair

Study (Sequence)	
PRJNA542350	Time-series reveals processes underlying colon inflammation and repair

Powered by EBI Search

The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure
The ENA is an ELIXIR Core Data Resource. Learn more ›

Depositing your data

Depositing your data

NBIS SciLifeLab

What:

All raw sequencing data, metadata and any additional processed counts/data/information.

Why:

To allow others and your-future-self to reproduce your results and re-use your data.

When:

- You can submit your data to GEO before submitting the manuscript. The data can remain private for a maximum of **3 years**.
- Once the manuscript is finally accepted, you can release it to the public.

Where:

For **non-human** RNA-seq samples:

- Submit everything to GEO, raw FASTQ files, metadata and processed count matrices

For **human** RNA-seq samples:

- Send email to GEO about sending human samples
- Submit raw FASTQ and metadata files to ENA (with access restrictions)
- Submit processed count matrices and metadata (without patient information) to GEO.

Thank you. Questions?

Paulo Czarnewski