

# Online Repositories

---

RNA-seq data analysis

**Paulo Czarnewski** | 30-Nov-2020

# ENSEMBL

### 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

### Variation 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

### 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](#)

### Search

for

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

### 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

#### Other reference assemblies

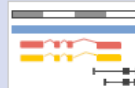
- NCBIM37 (Ensembl release 67)

#### Other strains

This species has data on 15 additional strains. [View list of strains](#)



View karyotype



Example region

### 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

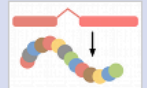
Download FASTA files for genes, cDNAs, ncRNA, proteins

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*

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<sup>1</sup>, Sara M. Parigi<sup>1</sup>, Chiara Sorini<sup>1</sup>, Oscar E. Diaz<sup>1</sup>, Srustidhar Das<sup>1</sup>, Nicola Gagliani<sup>1,2,3</sup> & Eduardo J. Villablanca<sup>1,3</sup>

### Reuse public data

**Table 1 Publicly available human data sets used in this paper**

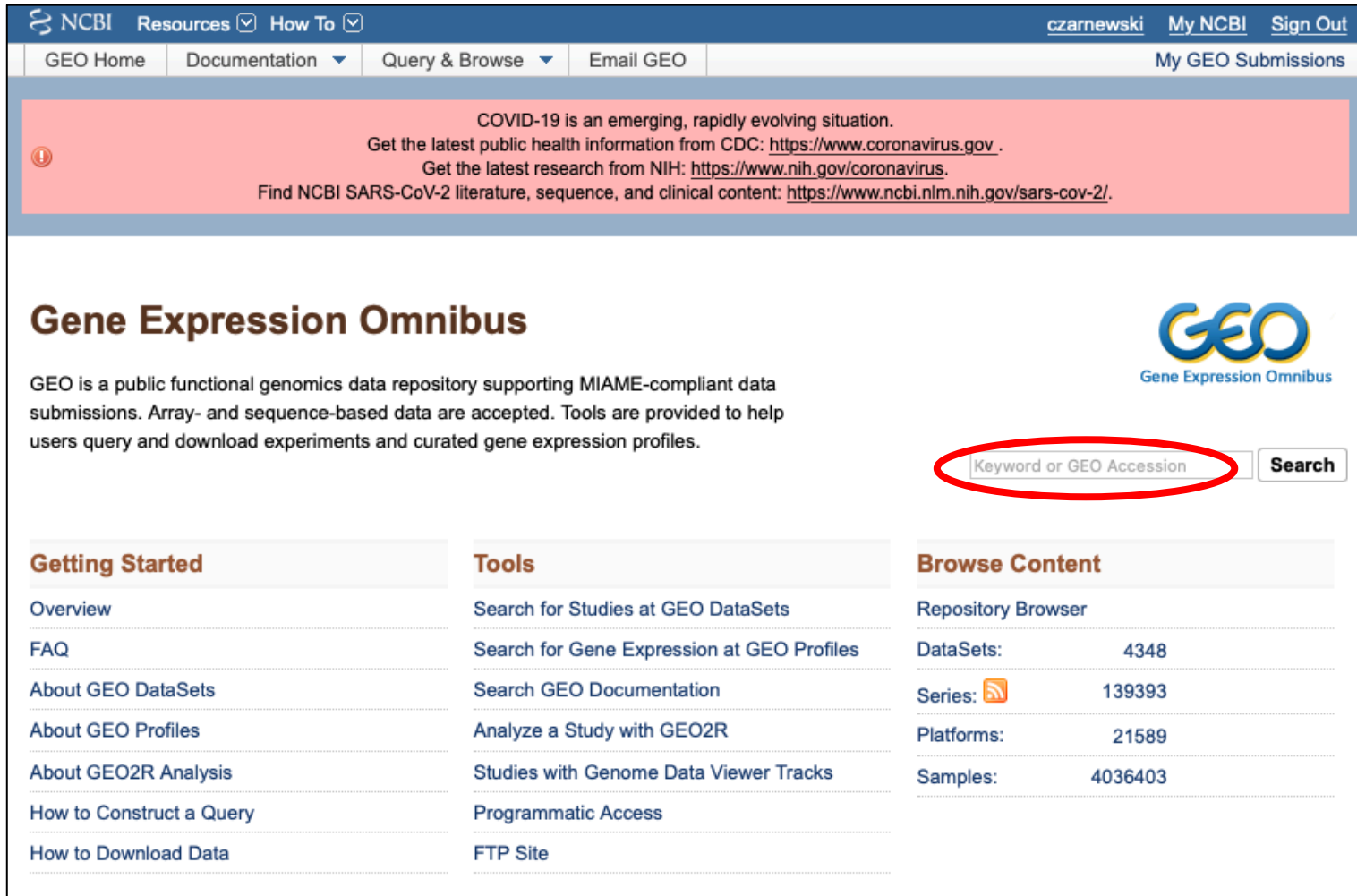
Data set ID	Total	Responders	Nonresponders	Ref.
Infliximab:				
GSE12251	23	11	12	13
GSE73661	23	15	8	15
GSE23597	32	7	25	14
GSE16879	24	16	8	12
Sum	102	49	53	
Vedolizumab:				
GSE73661	37	23	14	15
Pediatric UC:				
GSE109142	206	105	101	33

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 accession number **GSE131032**.

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



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 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.

**GEO**  
Gene Expression Omnibus

Keyword or GEO Accession

### Getting Started


- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

### Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

### Browse Content

Repository Browser

DataSets:	4348
Series: 	139393
Platforms:	21589
Samples:	4036403

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

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:  Format:  Amount:  GEO accession:

**Series GSE131032**  [Query DataSets for GSE131032](#)

Status	Public on May 11, 2019
Title	Time-series reveals processes underlying colon inflammation and repair
Organism	<a href="#">Mus musculus</a>
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)	<a href="#">Czarnewski P</a> , <a href="#">Villablanca EJ</a>
Citation(s)	Czarnewski P, Parigi SM, Sorini C, Diaz OE et al. Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification. <i>Nat Commun</i> 2019 Jun 28;10(1):2892. PMID: <a href="#">31253778</a>
NIH grant(s)	<input type="button" value="Add grant"/>



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	Tomtebodavägen 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
<a href="#">SOFT formatted family file(s)</a>	SOFT <a href="#">?</a>
<a href="#">MINiML formatted family file(s)</a>	MINiML <a href="#">?</a>
<a href="#">Series Matrix File(s)</a>	TXT <a href="#">?</a>

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

NCBI > GEO > **Accession Display** [?](#) Contact: [czarnewski](#) [?](#) | [My submissions](#) [?](#) | [Sign Out](#) [?](#)

**GEO help:** Mouse over screen elements for information.

Scope:  Format:  Amount:  GEO accession:

**Sample GSM3760139**  [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: HLNYPHBCXX

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
<a href="#">GSM3760139_KI_PC1606_01.tsv.gz</a>	1.1 Mb	<a href="#">(ftp)</a> <a href="#">(http)</a>	TSV

[SRA Run Selector](#) [?](#)

Metadata info

Raw counts

# SRA

## *Sequence Read Archive*

NCBI Resources How To czarnewski My NCBI Sign Out

SRA SRA **SRP197582** Search 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: Filters: [Manage Filters](#)

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 >>

- GSM3760164: DSSd14\_3; Mus musculus; RNA-Seq**  
1. 1 ILLUMINA (Illumina HiSeq 2500) run: 26.7M spots, 1.3G bases, 636.1Mb downloads  
Accession: SRX5818211
- GSM3760163: DSSd14\_2; Mus musculus; RNA-Seq**  
2. 1 ILLUMINA (Illumina HiSeq 2500) run: 25.2M spots, 1.3G bases, 599.7Mb downloads  
Accession: SRX5818210
- GSM3760162: DSSd14\_1; Mus musculus; RNA-Seq**  
3. 1 ILLUMINA (Illumina HiSeq 2500) run: 26.3M spots, 1.3G bases, 627.5Mb downloads  
Accession: SRX5818209
- GSM3760161: DSSd12\_3; Mus musculus; RNA-Seq**  
4. 1 ILLUMINA (Illumina HiSeq 2500) run: 27.8M spots, 1.4G bases, 665.1Mb downloads  
Accession: SRX5818208
- GSM3760160: DSSd12\_2; Mus musculus; RNA-Seq**

**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]

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)

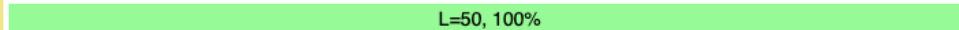
[Change accession...](#)

**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:



[Legend](#)

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
<a href="#">SRX5818211</a>		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE	<input type="button" value="BLAST"/>

Biosample	Sample Description	Organism	Links
<a href="#">SAMN11619097</a> (SRS4746980)		<a href="#">Mus musculus</a>	<a href="#">PRJNA542350 [Time-series reveals processes underlying colon inflammation and repair]</a>

Bioproject	SRA Study	Title
<a href="#">PRJNA542350</a>	<a href="#">SRP197582</a>	Time-series reveals processes underlying colon inflammation and repair

[Show abstract](#)

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.

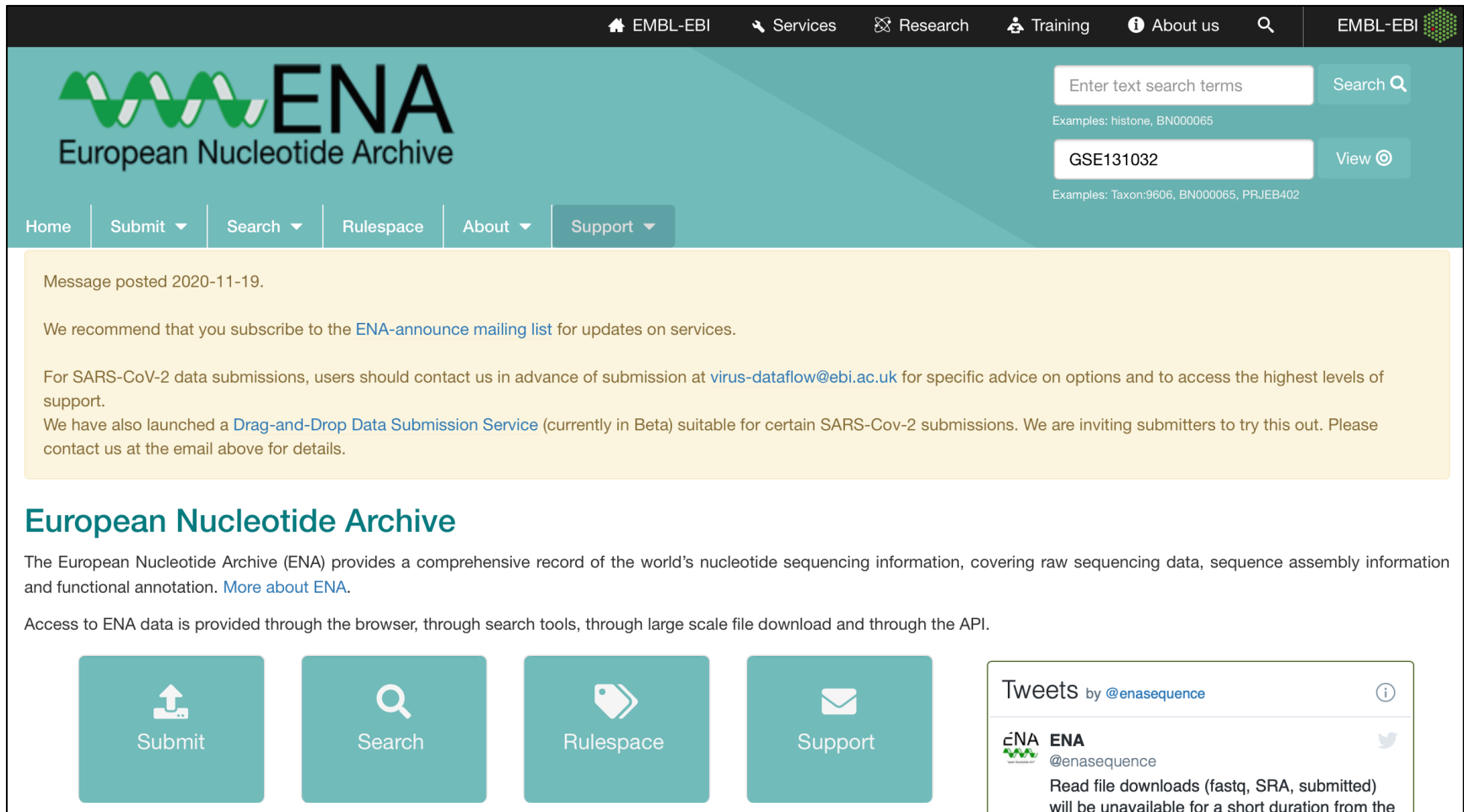
Type	Size	Location	Name	Free Egress	Access Type
run	651,383 Kb	NCBI	<a href="https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos2/sra-pub-run-15/SRR9041115/SRR9041115.1">https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos2/sra-pub-run-15/SRR9041115/SRR9041115.1</a>	worldwide	anonymous
		AWS	s3://sra-pub-run-5/SRR9041115/SRR9041115.1	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*

Similar to SRA, but in Europe.



The screenshot shows the ENA website interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, and About us. The main header features the ENA logo and a search bar with the text "Enter text search terms" and a "Search" button. Below the search bar, there are examples of search terms: "histone, BN000065" and "GSE131032". A "View" button is also present. A secondary navigation bar includes links for Home, Submit, Search, Rulespace, About, and Support. A yellow banner contains a message posted on 2020-11-19, recommending users subscribe to the ENA-announce mailing list and providing contact information for SARS-CoV-2 data submissions. Below the banner, the heading "European Nucleotide Archive" is followed by a paragraph describing the archive's comprehensive record of nucleotide sequencing information. A section titled "Access to ENA data" lists four methods: browser, search tools, large scale file download, and API. At the bottom, there are four teal buttons labeled "Submit", "Search", "Rulespace", and "Support". On the right side, there is a "Tweets by @enasequence" widget showing a tweet from ENA (@enasequence) about read file downloads being unavailable for a short duration.



ENA is also linked to samples deposited in SRA.

The screenshot shows the ENA website interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, About us, and a search icon. Below this is a teal header with the ENA logo and a search bar containing 'GSE131032'. A secondary search bar labeled 'Enter accession' is also present. Below the header is a navigation menu with links for Home, Submit, Search, Rulespace, About, and Support. The main content area is titled 'Text Search' and includes a sub-header 'Search term:' with a text input field containing 'GSE131032' and a green 'Search' button. Below the search bar, the results are displayed under the heading 'Search results for GSE131032'. The results are categorized into 'Study' and 'Study (Sequence)'. The 'Study' result is for SRP197582, and the 'Study (Sequence)' result is for PRJNA542350. Both results have a description: 'Time-series reveals processes underlying colon inflammation and repair'. At the bottom left, there is a logo for 'elixir Core Data Resource' and text stating 'The European Nucleotide Archive (ENA) is part of the ELIXIR infrastructure' and 'The ENA is an ELIXIR Core Data Resource. Learn more >'. The footer of the page contains the number '17'.

## Depositing your data

## 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 | 30-Nov-2020**