D882–D889 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 12 November 2019*
doi: 10.1093/nar/gkz1062

New developments on the Encyclopedia of DNA Elements (ENCODE) data portal

Yunhai Luo, Benjamin C. Hitz, Idan Gabdank, Jason A. Hilton, Meenakshi S. Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, Ulugbek K. Baymuradov, Keenan Graham, Casey Litton, Stuart R. Miyasato, J. Seth Strattan, Otto Jolanki, Jin-Wook Lee, Forrest Y. Tanaka, Philip Adenekan, Emma O'Neill and J. Michael Cherry*

Department of Genetics, Stanford University, Stanford, CA 94305-5477, USA

Received September 13, 2019; Revised October 18, 2019; Editorial Decision October 21, 2019; Accepted October 25, 2019

ABSTRACT

The Encyclopedia of DNA Elements (ENCODE) is an ongoing collaborative research project aimed at identifying all the functional elements in the human and mouse genomes. Data generated by the ENCODE consortium are freely accessible at the ENCODE portal (https://www.encodeproject.org/), which is developed and maintained by the ENCODE Data Coordinating Center (DCC). Since the initial portal release in 2013, the ENCODE DCC has updated the portal to make ENCODE data more findable, accessible, interoperable and reusable. Here, we report on recent updates, including new ENCODE data and assays, ENCODE uniform data processing pipelines, new visualization tools, a dataset cart feature, unrestricted public access to ENCODE data on the cloud (Amazon Web Services open data registry, https://registry.opendata.aws/encode-project/) and more comprehensive tutorials and documentation.

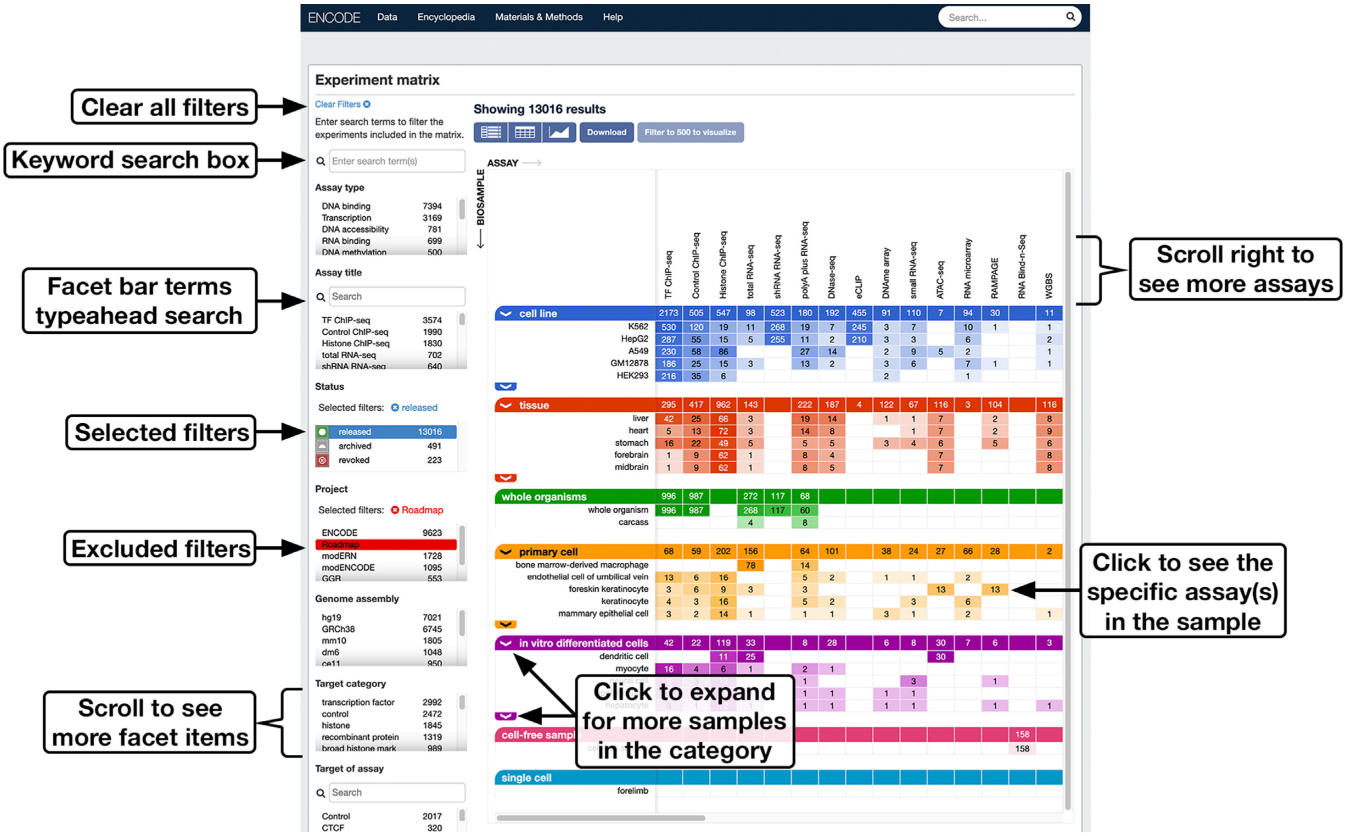
INTRODUCTION

Over 99% of the human genome was sequenced as a part of the Human Genome Project that was completed in April 2003 (1). Learning the function of genomic sequences remains a challenge and even today the biological purpose of a large fraction of the human genome is largely unknown. The Encyclopedia of DNA Elements (ENCODE) project is a public research consortium funded by the National Human Genome Research Institute initiated after the completion of the Human Genome Project in 2003 (2). The goal of the ENCODE project is to identify all the functional elements in the human and mouse genomes. All experimental metadata, raw data and analysis results generated by the ENCODE project are freely accessible to the scientific community on the ENCODE portal (https://www.encodeproject.org/), developed and maintained by the ENCODE Data Coordinating Center (DCC) (3–6). In addition to the data generated by the ENCODE consortium, the ENCODE portal also hosts data from modENCODE (7), modERN (8), The NIH Roadmap Epigenomics Consortium (9) and Genomics of Gene Regulation (https://www.genome.gov/27561317/genomics-of-gene-regulation/) projects as well as some datasets provided by other members of the scientific community. The ENCODE portal provides unrestricted access to more than 15 000 experimental results from more than 40 different categories of high-throughput sequencing (HTS) technologies in over 75 different cell and tissue types. The ENCODE portal hosts more than 640 terabytes of data files and this is expected to grow to over 1 petabyte by 2021.

The ENCODE DCC is a critical component of the ENCODE project, working to maximize the accessibility and utility of the data generated by the consortium. Using the data model developed by the DCC, ENCODE production labs format and submit to the ENCODE portal structured metadata describing their experiments and the data generated by the experiments. In addition to serving as the centralized data deposition repository of the ENCODE consortium, the DCC is also responsible for primary data analysis of key ENCODE assays. The DCC processes the raw data using the ENCODE uniform processing pipelines and submits processed analysis files to the portal (3,4,6). The DCC is also responsible for the documentation of a variety of metadata and data standards developed by the ENCODE consortium, as well as for the implementation of automated checks to validate data against these standards (6). Collectively, these efforts ensure that ENCODE data are findable, accessible, interoperable and reusable (10) and follow the original metadata organization principles (6).

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Email: cherry@stanford.edu
Database URL: www.encodeproject.org

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



ENCODE Data Encyclopedia Materials & Methods Help Search...

Experiment matrix

Clear Filters ⊗

Enter search terms to filter the experiments included in the matrix.

Enter search term(s)

Assay type

DNA binding	7394
Transcription	3169
DNA accessibility	781
RNA binding	699
DNA methylation	500

Assay title

Search

TF ChIP-seq	3574
Control ChIP-seq	1990
Histone ChIP-seq	1830
total RNA-seq	702
shRNA RNA-seq	640

Status

Selected filters: ⊗ released

released	13016
archived	491
revoked	223

Project

Selected filters: ⊗ Roadmap

ENCODE	9623
Roadmap	
modERN	1728
modENCODE	1095
GGR	553

Genome assembly

hg19	7021
GRCh38	6745
mm10	1805
dm6	1048
ce11	950

Target category

transcription factor	2992
control	2472
histone	1845
recombinant protein	1319
broad histone mark	989

Target of assay

Search

Control	2017
CTCF	320

Showing 13016 results

Download | Filter to 500 to visualize

ASSAY →

← BIOSAMPLE

	TF ChIP-seq	Control ChIP-seq	Histone ChIP-seq	total RNA-seq	shRNA RNA-seq	polyA plus RNA-seq	DNase-seq	eCLIP	DNAme array	small RNA-seq	ATAC-seq	RNA microarray	RAMPAGE	RNA Bind-n-Seq	WGBS
cell line	2173	505	547	98	523	180	192	455	91	110	7	94	30		11
K562	530	120	19	11	268	19	7	245	3	7		10	1		1
HepG2	287	55	15	5	255	11	2	210	3	3		6			2
A549	230	58	86			27	14		2	9	5	2			1
GM12878	186	25	15	3		13	2		3	6		7	1		1
HEK293	216	35	6						2			1			
tissue	295	417	962	143		222	187	4	122	67	116	3	104		116
liver	42	25	66	3		19	14		1	1	7		2		8
heart	5	13	72	3		14	8			1	7		2		9
stomach	16	22	49	5		5	5		3	4	6		5		6
forebrain	1	9	62	1		8	4				7				8
midbrain	1	9	62	1		8	5				7				8
whole organisms	996	987		272	117	68									
whole organism	996	987		268	117	60									
carcass				4		8									
primary cell	68	59	202	156		64	101		38	24	27	66	28		2
bone marrow-derived macrophage				78		14									
endothelial cell of umbilical vein	13	6	16			5	2		1	1		2			
foreskin keratinocyte	3	6	9	3		3					13		13		
keratinocyte	4	3	16			5	2			3		6			
mammary epithelial cell	3	2	14	1		1	1		3	1		2			1
in vitro differentiated cells	42	22	119	33		8	28		6	8	30	7	6		3
dendritic cell			11	25							30				
myocyte	16	4	6	1		2	1								
neural cell	9	3	12			1				3			1		
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]		1	1		1	1					
hepatocyte	3	1	12	1		1	1		1	1			1		1
cell-free sample														158	
cell-free sample														158	
single cell															
forelimb															

Clear all filters

Keyword search box

Facet bar terms typeahead search

Selected filters

Excluded filters

Scroll to see more facet items

Scroll right to see more assays

Click to see the specific assay(s) in the sample

Click to expand for more samples in the category

Figure 1. Experiment matrix. The ENCODE portal hosts data generated by more than 40 different biochemical assays, listed on the *x*-axis of the matrix. The *y*-axis lists various sample types represented on the portal. Each cell in the matrix indicates the number of experiments that are currently available on the portal with a particular sample type and the corresponding assay type. To the left of the matrix are faceted browsing interface bars that can be used to positively select (blue) and negate selection (red) of certain values of specific experimental metadata properties.

DATA ON THE ENCODE PORTAL

As an ongoing collaborative research project, both the size and the spectrum of the data available on the ENCODE portal continue to grow from year to year (3). More than 2000 experiments of various assay types were added to the portal in the past year, including experiments studying more than 30 new sample types (Supplementary Table S1). Data generated by new HTS assays and cutting-edge technologies have been made publicly available, including, but not limited to, single-nucleus ATAC-seq (11), icSHAPE (12) and long-read RNA-seq (13 Wyman, D. et al. bioRxiv, https://doi.org/10.1101/672931). To get a high-level overview of the available data on the portal, users can go to the experiment matrix page (https://www.encodeproject.org/matrix/?type=Experiment&status=released), which presents the experiments as a matrix with sample types listed as its *y*-axis and assay types as its *x*-axis. The matrix along with the faceted browsing interface on the left-hand side of the screen is designed to help users find experiments of interest (Figure 1).

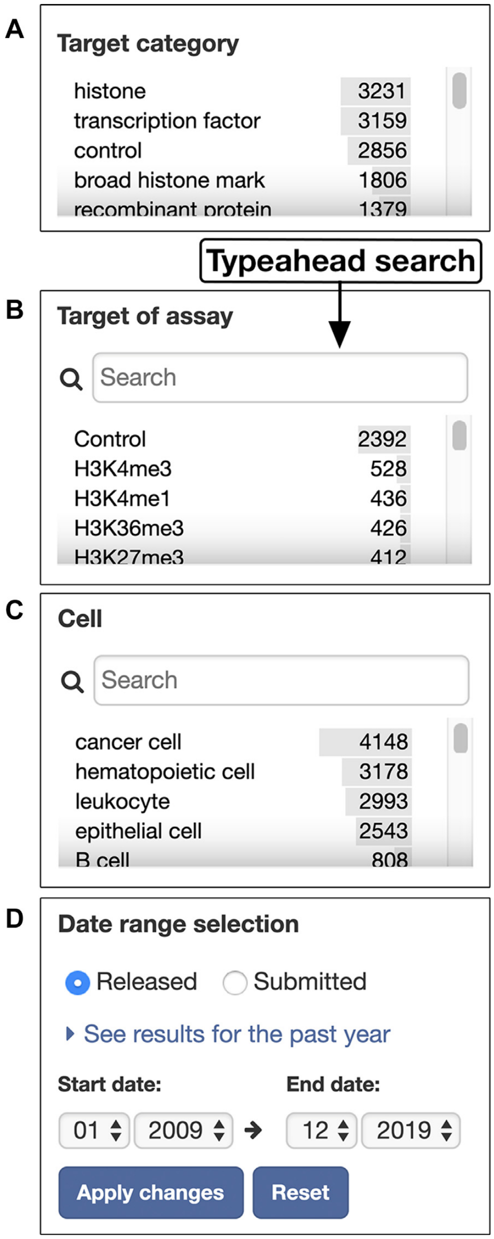
FACETED BROWSING INTERFACE

The faceted browsing interface was updated to include new facet bars: 'Cell', 'Target of assay' and 'Date range selection'. The 'Cell' facet bar helps to find experiments performed on specific cell type(s) and the 'Target of assay' facet bar helps to find experiments with a common target (e.g. transcription factor of interest). The 'Date range selection' facet bar allows the users to narrow down the time period in which experiments of interest became publicly available. Facet bars containing long lists of values (e.g. 'Cell' and 'Target of assay' facet bars) were added the 'typeahead' feature to facilitate the lookup for terms of interest (Figure 2).

Using data from NCBI Entrez Gene (https://www.ncbi.nlm.nih.gov/gene), HGNC (https://www.genenames.org/), MGNC (http://www.informatics.jax.org/mgihome/nomen/), FlyBase (https://flybase.org/), WormBase (https://www.wormbase.org) and the Gene Ontology Consortium (14–21), the DCC added Gene objects to the portal. Linkage of the target objects to the relevant gene objects allowed categorization of the targets based on the Gene Ontology annotations (for more details on the categorization method, visit https://www.encodeproject.org/target-categorization/). Target categories are listed under the 'Target category' facet bar (Figure 2A).

DATA PROCESSING

Processing the vast amount of data that are hosted on the ENCODE portal is challenging. The variation between experimental results coming from different labs and performed using similar, but not identical protocols requires



Target category

histone	3231
transcription factor	3159
control	2856
broad histone mark	1806
recombinant protein	1379

Target of assay

Control	2392
H3K4me3	528
H3K4me1	436
H3K36me3	426
H3K27me3	412

Cell

cancer cell	4148
hematopoietic cell	3178
leukocyte	2993
epithelial cell	2543
B cell	808

Date range selection

● Released ○ Submitted

Start date: 01 2009 → End date: 12 2019

Figure 2. Faceted browsing interface. Four new facet bars were added to the faceted browsing interface: (**A**) 'Target category' facet bar, (**B**) 'Target of assay' facet bar, (**C**) 'Cell' facet bar and (**D**) 'Date range selector' facet bar. The 'Cell' and 'Target of assay' facet bars that have long lists of values include the typeahead search feature to help users locate terms of interest more quickly.

robust, production-grade uniform processing pipelines that ensure successful analyses and generate comparable outputs. The DCC is tasked with the development, implementation and execution of the uniform processing pipelines that are first defined by collaborative consortium working groups. Centralized processing using uniform processing pipelines by the DCC generates comparable analysis results free of technical artifacts arising from methodological variability across labs and time. All of the pipelines that are developed by the DCC are open source and are maintained in the DCC GitHub repository (Supplementary Table S2, https://github.com/ENCODE-DCC). The pipelines that have been developed in the previous funding phase of ENCODE have an implementation on the DNAnexus cloud platform (https://dnanexus.com), making them accessible to the larger scientific community for unrestricted use (3). The pipelines the DCC is working on in the current funding phase of ENCODE are being developed in a new framework that involves modern technologies and platforms such as Docker (https://www.docker.com), Singularity (https://sylabs.io/, 22), WDL (https://doi.org/10.7490/f1000research.1114631.1), CircleCI (https://circleci.com) and automated testing. The DCC has established this pipeline development framework to ensure reproducibility, portability and robustness of the new uniform processing pipelines.

DATA VISUALIZATION

Data visualization is an essential step in the process of data analysis and interpretation. For the visualizable outputs of experiments on the ENCODE portal, the DCC has created tracks and track hubs, which can be visualized using the UCSC (https://genome.ucsc.edu/) and ENSEMBL (https://www.ensembl.org) genome browsers (23,24). These visualizations require redirection to the browser's portal and the DCC has found that user experience is compromised by the limited track-related metadata that can be passed to the genome browser websites. To improve the user experience and provide independent data visualization capability, the DCC has embedded a GPU accelerated genome browser powered by Valis (https://valis.bio) on the ENCODE portal (Figure 3). Experiment pages on the portal have been redesigned, and the tab 'Genome browser' has been added for the embedded genome browser. The new 'Genome browser' tab allows users to inspect visualizable processed files, e.g. bigWig and bigBed file formats, deriving from the experiment of interest.

Genome browsers such as UCSC, ENSEMBL and Valis are designed to visualize genomic data along one dimension, the genomic position axis. The results of experiments probing the three-dimensional organization of the genome are difficult to visualize with these tools. The ENCODE consortium has generated data from a set of Hi-C, ChIA-PET and 5C experiments that characterize the three-dimensional chromatin organization in different cells and tissues (5,25–29). To support visualization of Hi-C results, portal pages for these experiments were modified to include links, which can be found on the 'File details' tab, redirecting to the Juicebox visualization software web page (https://aidenlab.org/juicebox/) and to allow selection of the relevant files for visualization (30,31).

CART FOR CUSTOM COLLECTIONS OF DATASETS

The ENCODE portal has a free-text searching system and a faceted browsing interface that can be used to refine search queries on ENCODE metadata. Though these are powerful tools for finding datasets of interest, they do not sup-

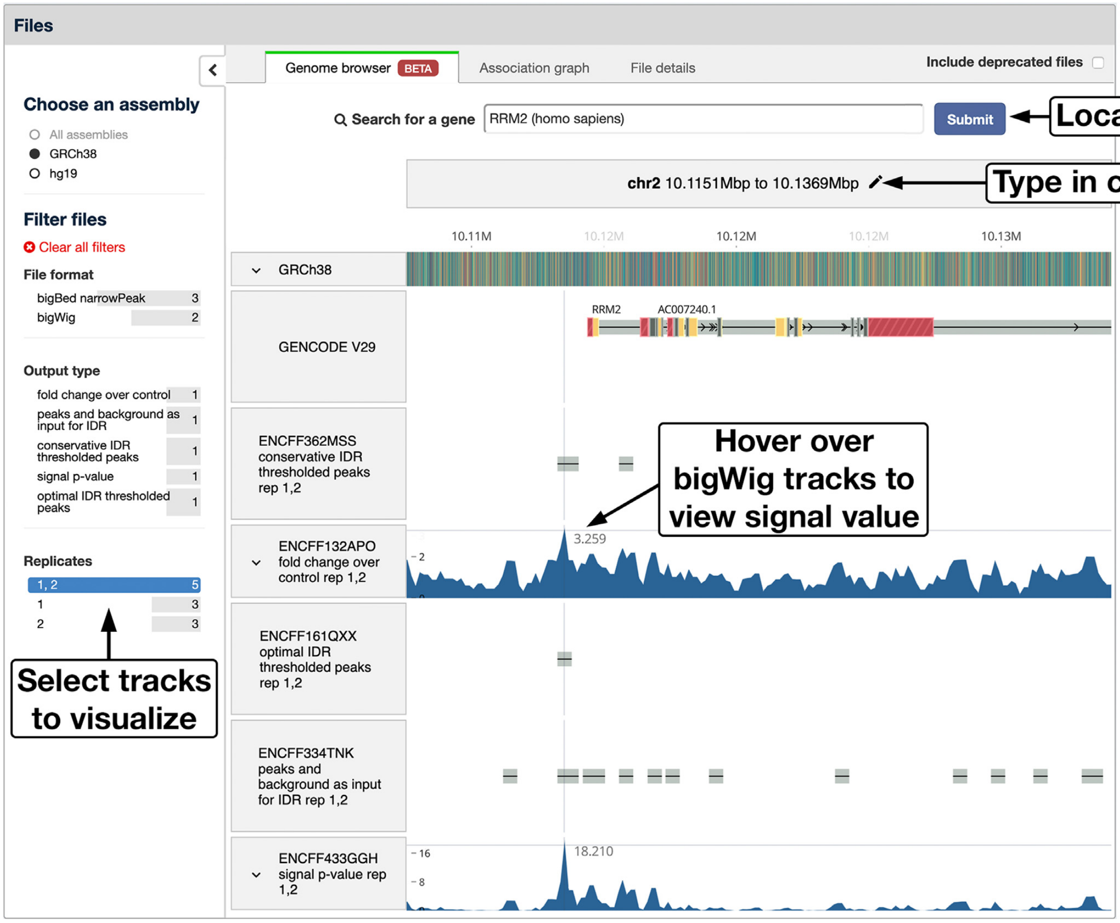

Files
Genome browser BETA
Association graph
File details
Include deprecated files □
Choose an assembly
○ All assemblies
● GRCh38
○ hg19
Filter files
Clear all filters
File format
bigBed narrowPeak 3
bigWig 2
Output type
fold change over control 1
peaks and background as input for IDR 1
conservative IDR thresholded peaks 1
signal p-value 1
optimal IDR thresholded peaks 1
Replicates
1, 2 5
1 3
2 3
Select tracks to visualize
Search for a gene
RRM2 (homo sapiens)
Submit
Locate by gene
chr2 10.1151Mbp to 10.1369Mbp
Type in coordinates
10.11M
10.12M
10.12M
10.12M
10.13M
GRCh38
GENCODE V29
RRM2
AC007240.1
ENCFF362MSS conservative IDR thresholded peaks rep 1,2
Hover over bigWig tracks to view signal value
ENCFF132APO fold change over control rep 1,2
3.259
2
ENCFF161QXX optimal IDR thresholded peaks rep 1,2
ENCFF334TNK peaks and background as input for IDR rep 1,2
ENCFF433GGH signal p-value rep 1,2
18.210
16
8

Figure 3. Valis genome browser embedded on the portal. Users can find their genomic region of interest by (i) searching for a specific gene, (ii) specifying genomic coordinates or (iii) using the mouse scrolling and zooming features directly to navigate the region. File selectors are available on the left, allowing users to select and visualize tracks for specific genome assembly, specific file format or output type, and files belonging to specific experimental replicate(s). This example can be found by clicking the 'Genome browser' tab on the experiment summary page (https://www.encodeproject.org/experiments/ENCSR807BGP/).

port the creation of a custom experimental grouping. To address this need, DCC developed the cart feature that allows users to add one or more experiments into a cart to create an arbitrary collection of experiments. After building a cart-based data collection, users can filter and download data files along with the associated metadata related to the experiments from the cart (Figure 4).

ENCODE AT AWS REGISTRY OF OPEN DATA

The ENCODE portal provides open and unrestricted access to experimental metadata as well as the data files associated with the experiments. The data files accessible through the portal can be directly downloaded or accessed on the Amazon cloud. ENCODE has recently become part of the AWS Registry of Open Data and ENCODE data files are accessible via a public AWS Simple Storage Service (S3) bucket. The bucket is fully accessible for local cloud computing on AWS products. Alternatively, bucket content can be transferred like any other file. Using ENCODE data with AWS services is highly performant when data come directly from S3. The link to the ENCODE AWS Registry of Open Data page that includes tutorials and examples on ENCODE data use (https://registry.opendata.aws/encode-project/) can be found under the 'Data' menu on the ENCODE portal top toolbar (Figure 5).

ENCODE PORTAL HELP AND DOCUMENTATION

All the experiments and their analyses on the ENCODE portal are represented using a structured metadata model that informs the interpretation of data in biological terms (3,4,6,32). The ENCODE portal includes new documentation and tutorials on how to use ENCODE metadata and data (https://www.encodeproject.org/help/getting-started/). In addition to FAQ and documentation web pages, a set of interactive tutorials was prepared using the WalkMe user onboarding platform (https://www.walkme.com/) and integrated on the portal web pages. Each tutorial guides users, step by step, toward a specific goal (Figure 5). The ENCODE portal also provides an API to facilitate programmatic access to metadata and data files. The API is documented using the Swagger platform with examples on the Swagger hub (https://app.swaggerhub.com/apis-docs/encodeproject/api/basic_search). The 'Help' rectangular widget located in the bottom-right corner of any page

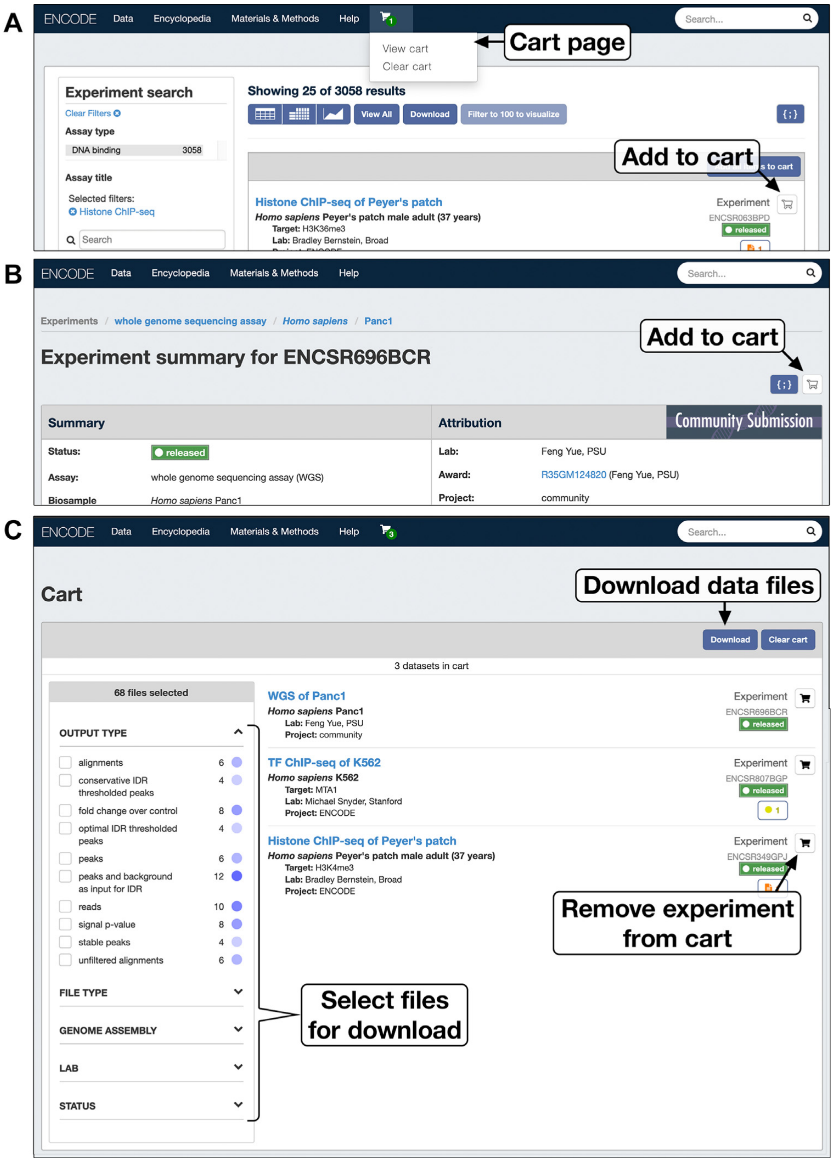


Figure 4. Cart for custom collections of datasets. Experiments can be added to the cart from the experiment search page **(A)** or by clicking on the cart icon on the experiment summary page **(B)**. Experiments that are no longer needed can be removed from the cart **(C)**. Files that belong to the experiments in the cart can be filtered using the file selectors on the left. For example, users can select only alignments mapped to the GRCh38 genome assembly. The cart allows users to download the data files satisfying specific criteria along with associated metadata. Currently, carts are saved per active browser session and will be erased after closing the web browser or refreshing the page.

of the ENCODE portal provides access to the list of various interactive tutorials, FAQ, SwaggerHub and documentation pages (Figure 5).

FUTURE DIRECTIONS

The ENCODE project is an ongoing collaborative research effort. For 16 years, ENCODE data have been used extensively by many researchers across the globe. The ENCODE consortium continues to provide unrestricted access to novel high-quality data for the scientific community. To more directly assess the function of ENCODE elements, new assays are underway, including massively parallel reporter assays (33), CRISPR screens (34) and STARR-seq (35). To increase the resolution of cell-type-specific states, new single-cell-based experimental assays are also underway. All of these new data will be made available through the ENCODE portal.

In addition to individual experiments and their metadata, the ENCODE portal also hosts results of integrative analysis such as the registry of cis-regulatory elements (manuscript submitted for publication) and other types of genome annotations computationally derived from experimental data coming from both the ENCODE consortium and the scientific community (https://www.encodeproject.org/matrix/?type=Annotation). The DCC will host and provide access to updated genome annotations generated by algorithms such as chromHMM (36) and Segway (37)

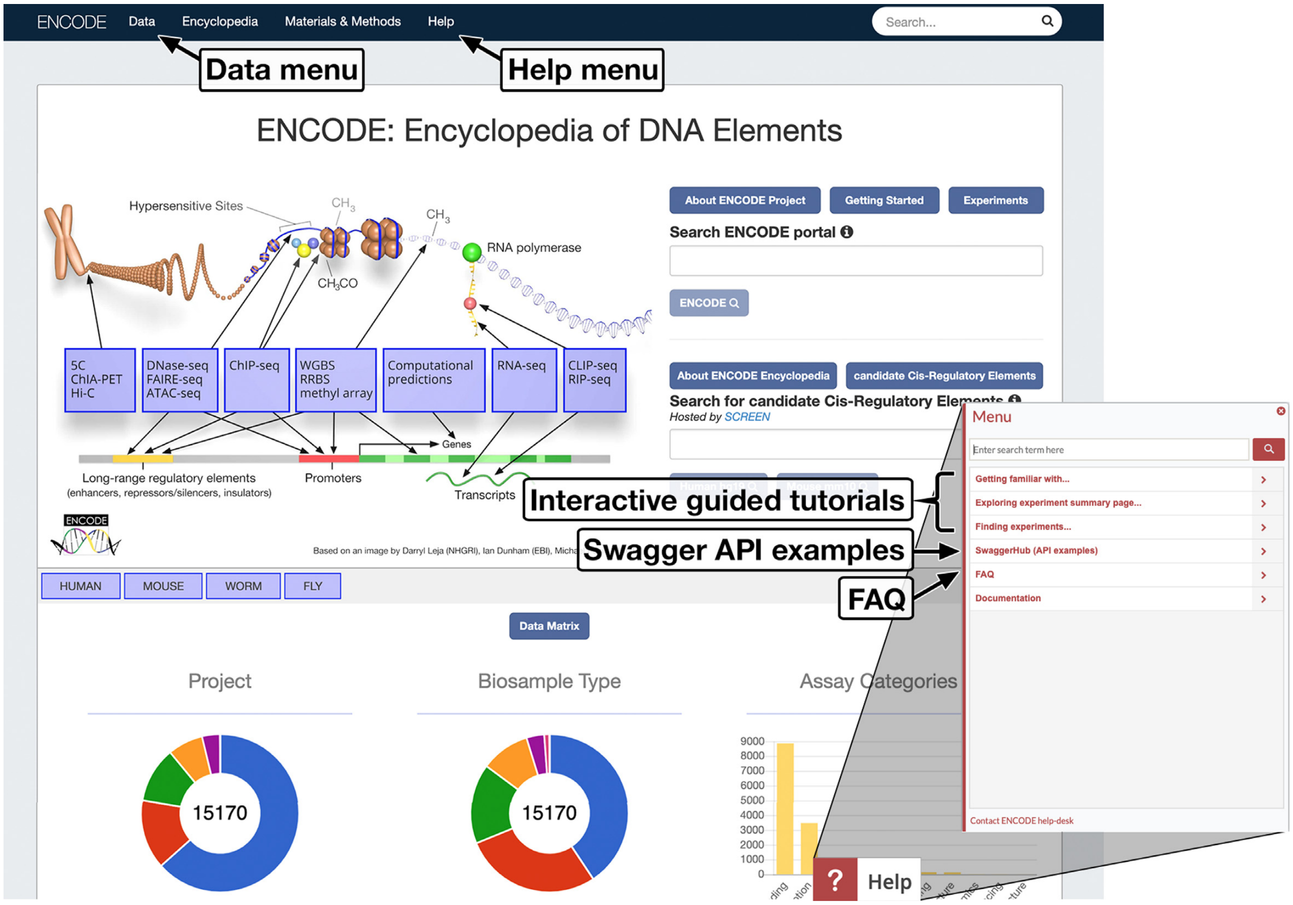


Figure 5. Tutorials, documentation and help on the ENCODE portal. Link to tutorials and examples for ENCODE AWS Registry of Open Data could be found under the 'Data' menu on the left-hand side of the top bar. Help information is accessible through the 'Help' menu on the top bar of every portal page. Interactive walk-through tutorials and additional help materials are available via the 'Help' widget visible in the bottom-right corner of the screen. The ENCODE DCC also supports users directly through the ENCODE help desk at encode-help@lists.stanford.edu.

using the rich and diverse data submitted to the ENCODE portal.

Genomic research is increasingly collaborative. The ENCODE consortium is an inherently collaborative project and extends this to collaborations with other consortia. For example, the ENCODE DCC shares epigenomic data with the International Human Epigenome Consortium (IHEC, http://ihec-epigenomes.org/). Together with IHEC, ENCODE initiated the EpiShare project, which is one of the driver projects of the Global Alliance for Genomics and Health (GA4GH, https://www.ga4gh.org/). The ENCODE portal infrastructure, metadata model and the uniform processing pipeline have been adopted by other projects and consortia. For example, the 4D Nucleome (4DN) (https://data.4dnucleome.org/) portal and Diabetes Epigenome Atlas (https://www.diabetesepigenome.org/) share the back-end SnoVault infrastructure developed and used for the ENCODE portal by the ENCODE DCC (38). IHEC and 4DN projects use uniform processing pipelines developed by the ENCODE DCC and the use of genome reference files required for data processing is being coordinated. The DCC continues to work closely with new consortia such as Human Cell Atlas and Human BioMolecular Atlas Program to create portable resources and interoperable data.

DATA AVAILABILITY

ENCODE is an open source project with all software available in the GitHub repository (https://github.com/ENCODE-DCC).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank past and current members of the ENCODE, modENCODE, modERN, GGR and REMC consortia in addition to the staff at the NCBI Entrez Gene, HGNC, MGNC, FlyBase, WormBase, IHEC, GA4GH, 4DN, T2D AMP Consortium, UCSC Genome, ENSEMBL and Valis browsers and the many talented graduate researchers, postdocs, bioinformaticians and ENCODE portal users for countless hours of data creation, site testing and feedback, without whom the ENCODE portal website would not be possible.

FUNDING

National Human Genome Research Institute, National Institutes of Health [U24 HG009397]. The content is solely theresponsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. The funders had no role in design, data processing, implementation, decision to publish, or preparation of the manuscript. Funding for open access charge: National Human Genome Research Institute, National Institutes of Health [U24 HG009397].

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M. and FitzHugh,W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. The ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
3. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
4. Sloan,C.A., Chan,E.T., Davidson,J.M., Malladi,V.S., Strattan,J.S., Hitz,B.C., Gabdank,I., Narayanan,A.K., Ho,M., Lee,B.T. *et al.* (2016) ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, **44**, D726–D732.
5. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
6. Hong,E.L., Sloan,C.A., Chan,E.T., Davidson,J.M., Malladi,V.S., Strattan,J.S., Hitz,B.C., Gabdank,I., Narayanan,A.K., Ho,M. *et al.* (2016) Principles of metadata organization at the ENCODE data coordination center. *Database*, **2016**, baw001.
7. Celniker,S.E., Dillon,L.A.L., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
8. Kudron,M.M., Victorsen,A., Gevirtzman,L., Hillier,L.W., Fisher,W.W., Vafeados,D., Kirkey,M., Hammonds,A.S., Gersch,J., Ammouri,H. *et al.* (2018) The ModERN Resource: genome-wide binding profiles for hundreds of *Drosophila* and *Caenorhabditis elegans* transcription factors. *Genetics*, **208**, 937–949.
9. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
10. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
11. Preissl,S., Fang,R., Huang,H., Zhao,Y., Raviram,R., Gorkin,D.U., Zhang,Y., Sos,B.C., Afzal,V., Dickel,D.E. *et al.* (2018) Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.*, **21**, 432–439.
12. Feng,C., Chan,D. and Spitale,R.C. (2017) Assaying RNA structure inside living cells with SHAPE. *Methods Mol. Biol.*, **1648**, 247–256.
13. Wyman,D. and Mortazavi,A. (2019) TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, **35**, 340–342.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
15. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **175**, 598–599.
16. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
17. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
18. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
19. Bult,C.J., Blake,J.A., Smith,C.L., Kadin,J.A., Richardson,J.E. and Mouse Genome Database Group (2019) Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.*, **47**, D801–D806.
20. Thurmond,J., Goodman,J.L., Strelets,V.B., Attrill,H., Gramates,L.S., Marygold,S.J., Matthews,B.B., Millburn,G., Antonazzo,G., Trovisco,V. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.
21. Lee,R.Y.N., Howe,K.L., Harris,T.W., Arnaboldi,V., Cain,S., Chan,J., Chen,W.J., Davis,P., Gao,S., Grove,C. *et al.* (2018) WormBase 2017: molting into a new stage. *Nucleic Acids Res.*, **46**, D869–D874.
22. Kurtzer,G.M., Sochat,V. and Bauer,M.W. (2017) Singularity: scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.
23. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
24. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
25. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
26. Rao,S.S.P., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
27. Sanborn,A.L., Rao,S.S.P., Huang,S.-C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6456–E6465.
28. Li,G., So,A.Y.-L., Sookram,R., Wong,S., Wang,J.K., Ouyang,Y., He,P., Su,Y., Casellas,R. and Baltimore,D. (2018) Epigenetic silencing of miR-125b is required for normal B-cell development. *Blood*, **131**, 1920–1930.
29. Vian,L., Pękowska,A., Rao,S.S.P., Kieffer-Kwon,K.-R., Jung,S., Baranello,L., Huang,S.-C., El Khattabi,L., Dose,M., Pruett,N. *et al.* (2018) The energetics and physiological impact of cohesin extrusion. *Cell*, **175**, 292–294.
30. Durand,N.C., Robinson,J.T., Shamim,M.S., Machol,I., Mesirov,J.P., Lander,E.S. and Aiden,E.L. (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.*, **3**, 99–101.
31. Robinson,J.T., Turner,D., Durand,N.C., Thorvaldsdóttir,H., Mesirov,J.P. and Aiden,E.L. (2018) Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.*, **6**, 256–258.
32. Malladi,V.S., Erickson,D.T., Podduturi,N.R., Rowe,L.D., Chan,E.T., Davidson,J.M., Hitz,B.C., Ho,M., Lee,B.T., Miyasato,S. *et al.* (2015) Ontology application and use at the ENCODE DCC. *Database*, **2015**, bav010.
33. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G. Jr, Kinney,J.B. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.
34. Shalem,O., Sanjana,N.E. and Zhang,F. (2015) High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.*, **16**, 299–311.
35. Arnold,C.D., Gerlach,D., Stelzer,C., Boryń,Ł.M., Rath,M. and Stark,A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
36. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.

37. Chan,R.C.W., Libbrecht,M.W., Roberts,E.G., Bilmes,J.A., Noble,W.S. and Hoffman,M.M. (2018) Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics*, **34**, 669–671.
38. Hitz,B.C., Rowe,L.D., Podduturi,N.R., Glick,D.I., Baymuradov,U.K., Malladi,V.S., Chan,E.T., Davidson,J.M., Gabdank,I., Narayana,A.K. *et al.* (2017) SnoVault and encodeD: a novel object-based storage system and applications to ENCODE metadata. *PLoS One*, **12**, e0175310.