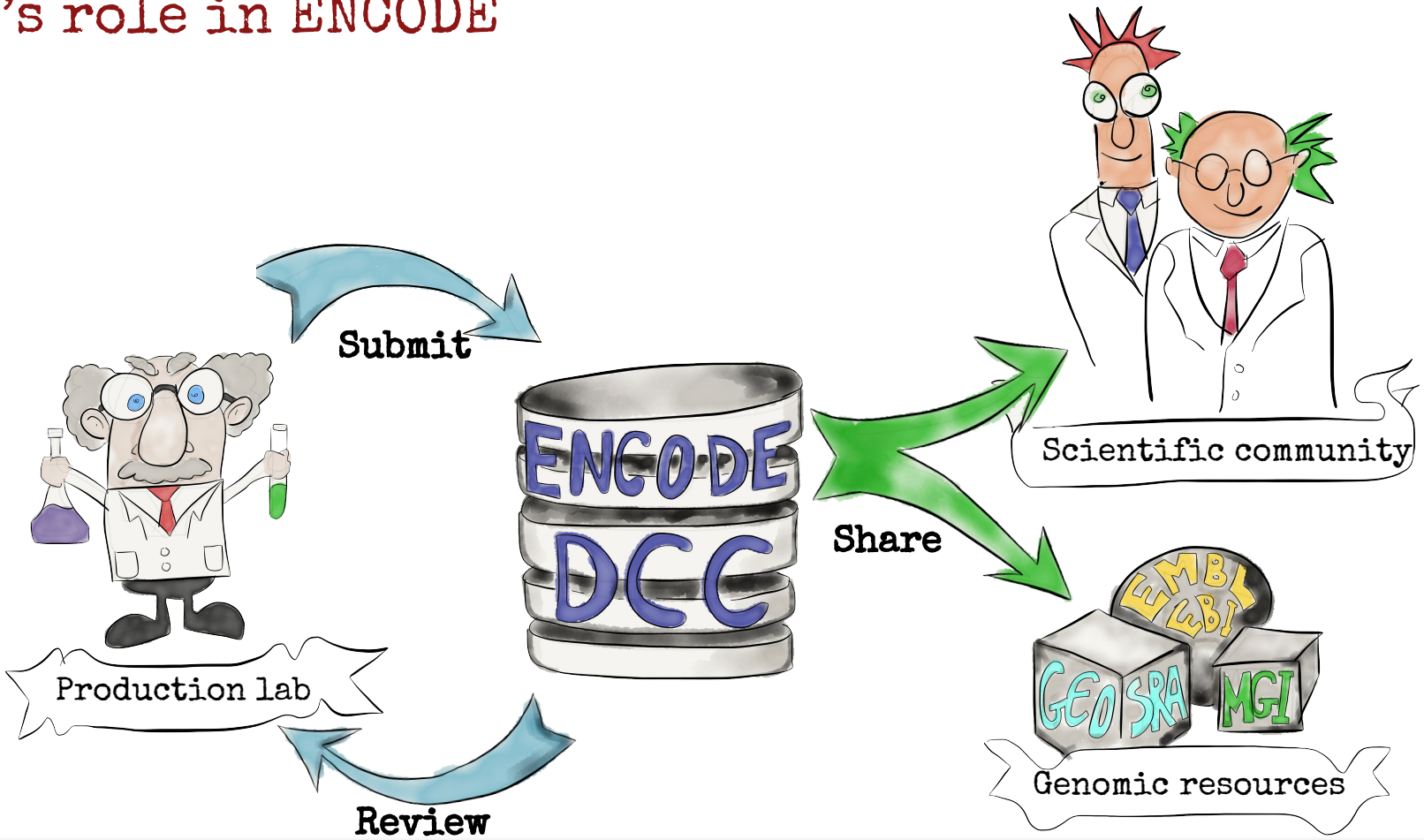
Intro to the ENCODE portal

Jason Hilton, PhD

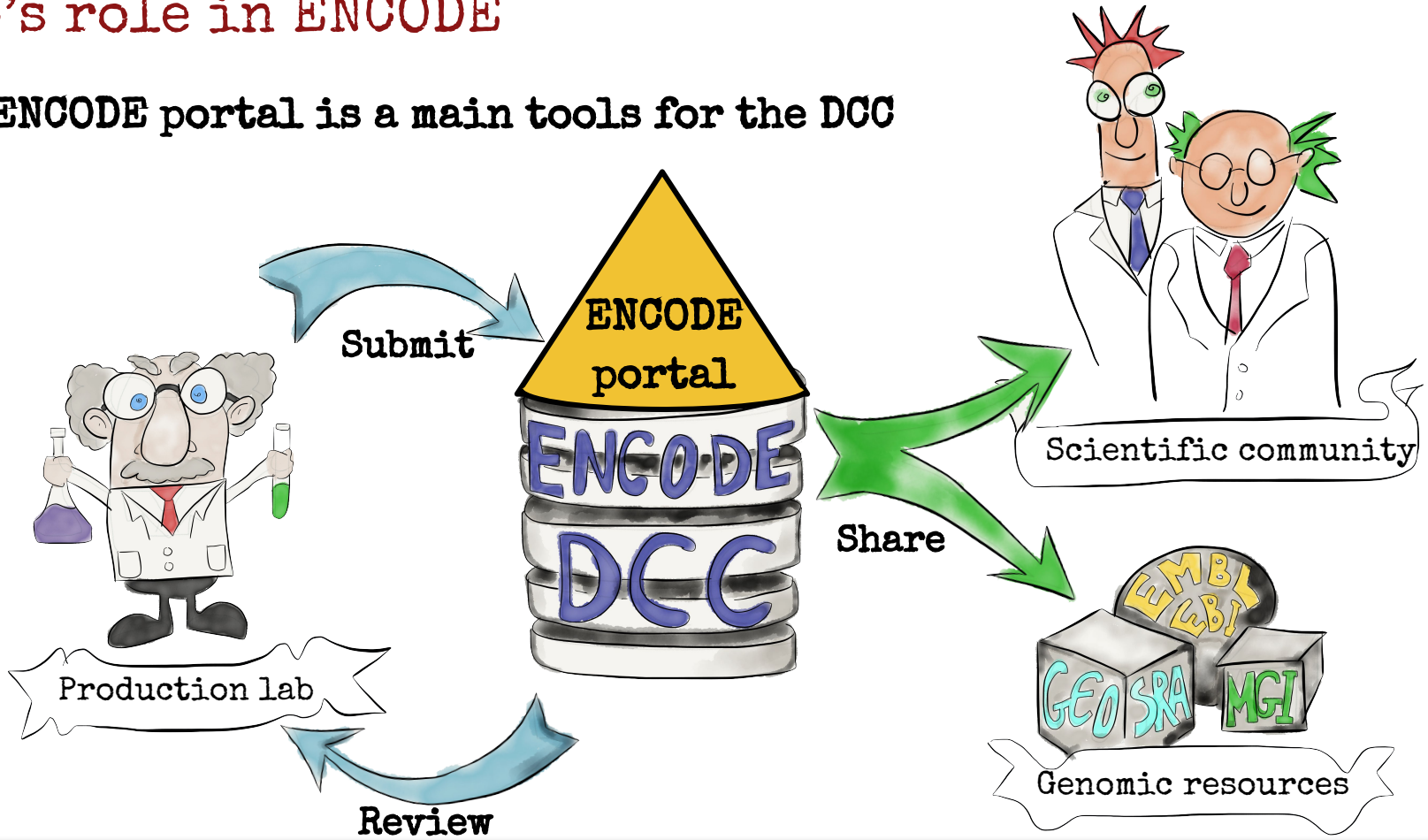
ENCODE Data Coordination Center

DCC's role in ENCODE



DCC's role in ENCODE

The ENCODE portal is a main tools for the DCC



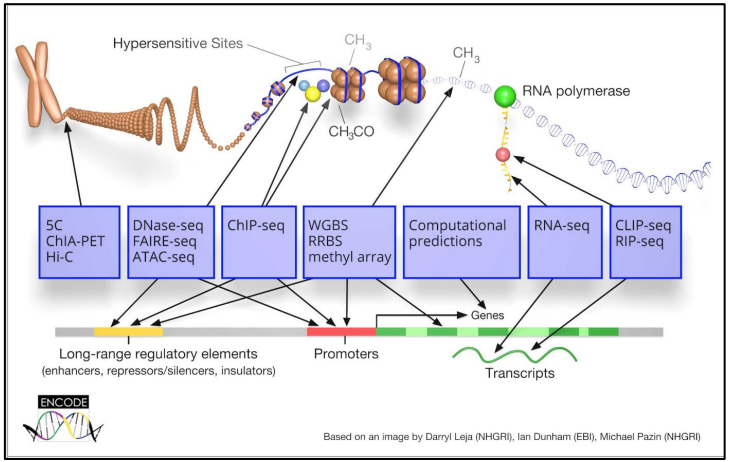
ENCODE portal

encodeproject.org

15K experiments

46 assay "flavors"

600TB data files



Genomics of Gene Regulation | ROADMAP epigenomics PROJECT | modERN | modENCODE

ENCODE portal

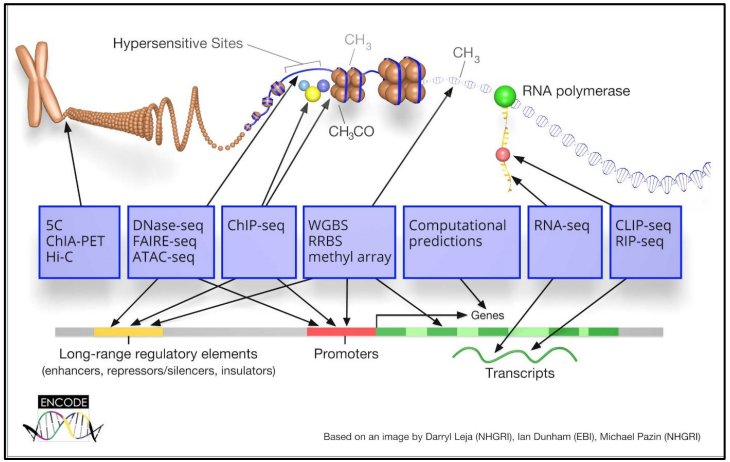
encodeproject.org

15K experiments

46 assay "flavors"

600TB data files

PRIMARY ORGANIZATIONAL UNIT ON ENCODE PORTAL



Genomics of Gene Regulation | ROADMAP epigenomics PROJECT | modERN | modENCODE

Outline

An ENCODE experiment

Browse & Search experiments

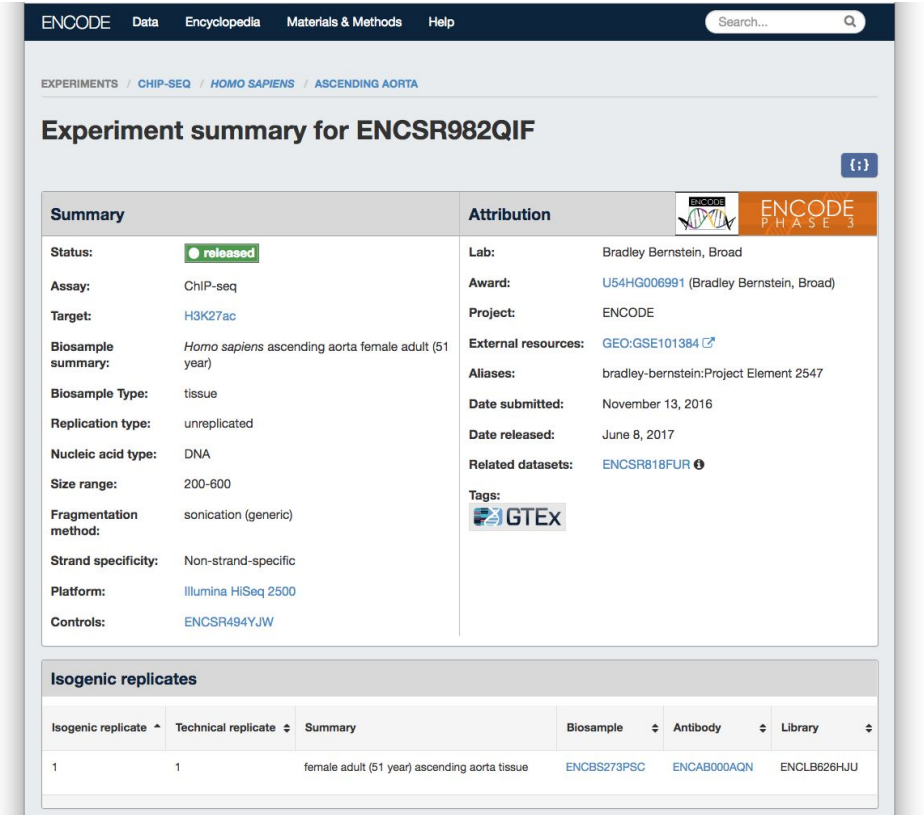
Visualize & Download files from many experiments

Searching by a region of interest

Access the ENCODE portal programmatically

An ENCODE experiment
https://www.encodeproject.org/ENCSR982QIF

The experiment page



ENCODE Data Encyclopedia Materials & Methods Help Search...

EXPERIMENTS / CHIP-SEQ / *HOMO SAPIENS* / ASCENDING AORTA

Experiment summary for ENCSR982QIF

Summary

Status:	released
Assay:	ChIP-seq
Target:	H3K27ac
Biosample summary:	*Homo sapiens* ascending aorta female adult (51 year)
Biosample Type:	tissue
Replication type:	unreplicated
Nucleic acid type:	DNA
Size range:	200-600
Fragmentation method:	sonication (generic)
Strand specificity:	Non-strand-specific
Platform:	Illumina HiSeq 2500
Controls:	ENCSR494YJW

Attribution

ENCODE PHASE 3

Lab:	Bradley Bernstein, Broad
Award:	U54HG006991 (Bradley Bernstein, Broad)
Project:	ENCODE
External resources:	GEO:GSE101384
Aliases:	bradley-bernstein:Project Element 2547
Date submitted:	November 13, 2016
Date released:	June 8, 2017
Related datasets:	ENCSR818FUR
Tags:	GTEx

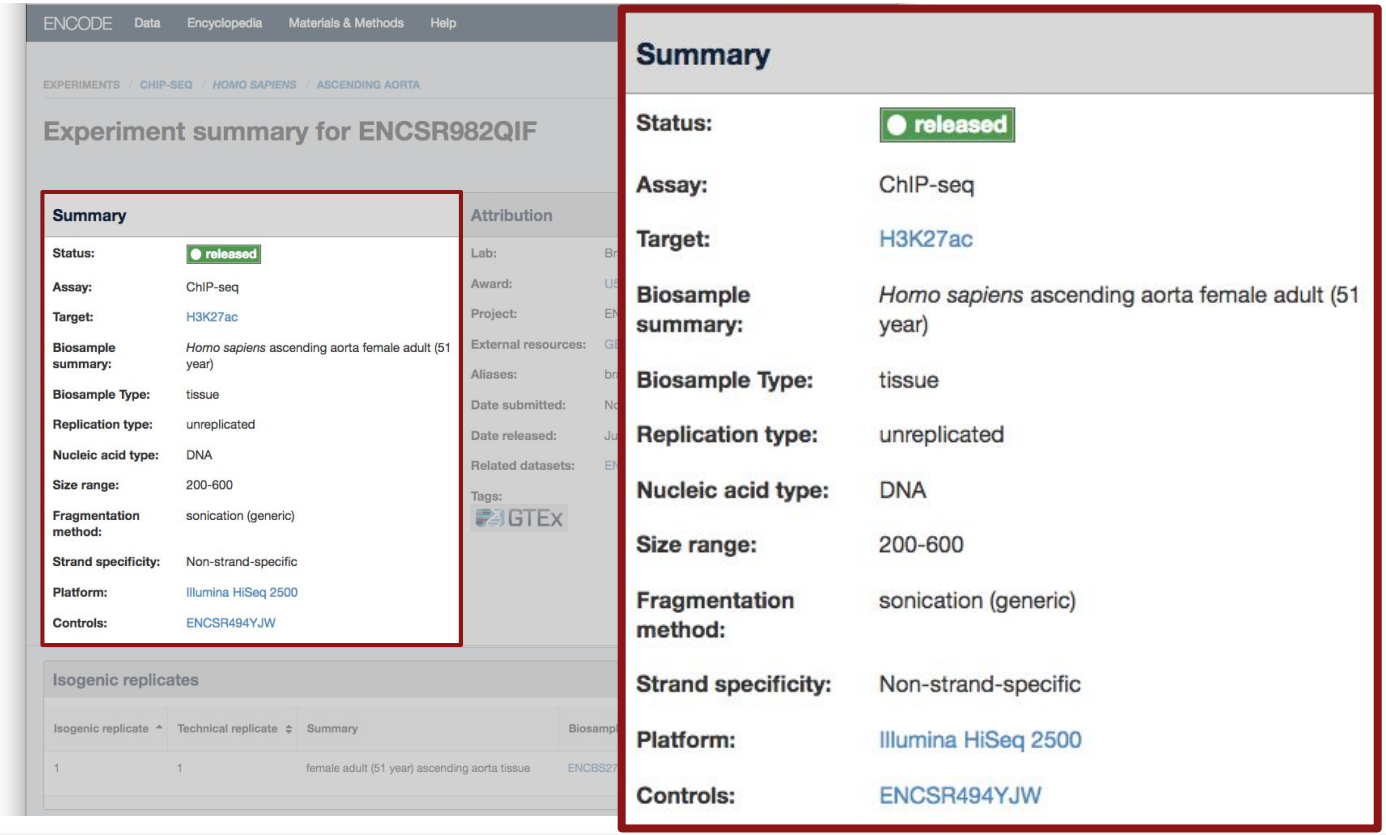
Isogenic replicates

Isogenic replicate	Technical replicate	Summary	Biosample	Antibody	Library
1	1	female adult (51 year) ascending aorta tissue	ENCBS273PSC	ENCAB000AQN	ENCLB626HJU

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

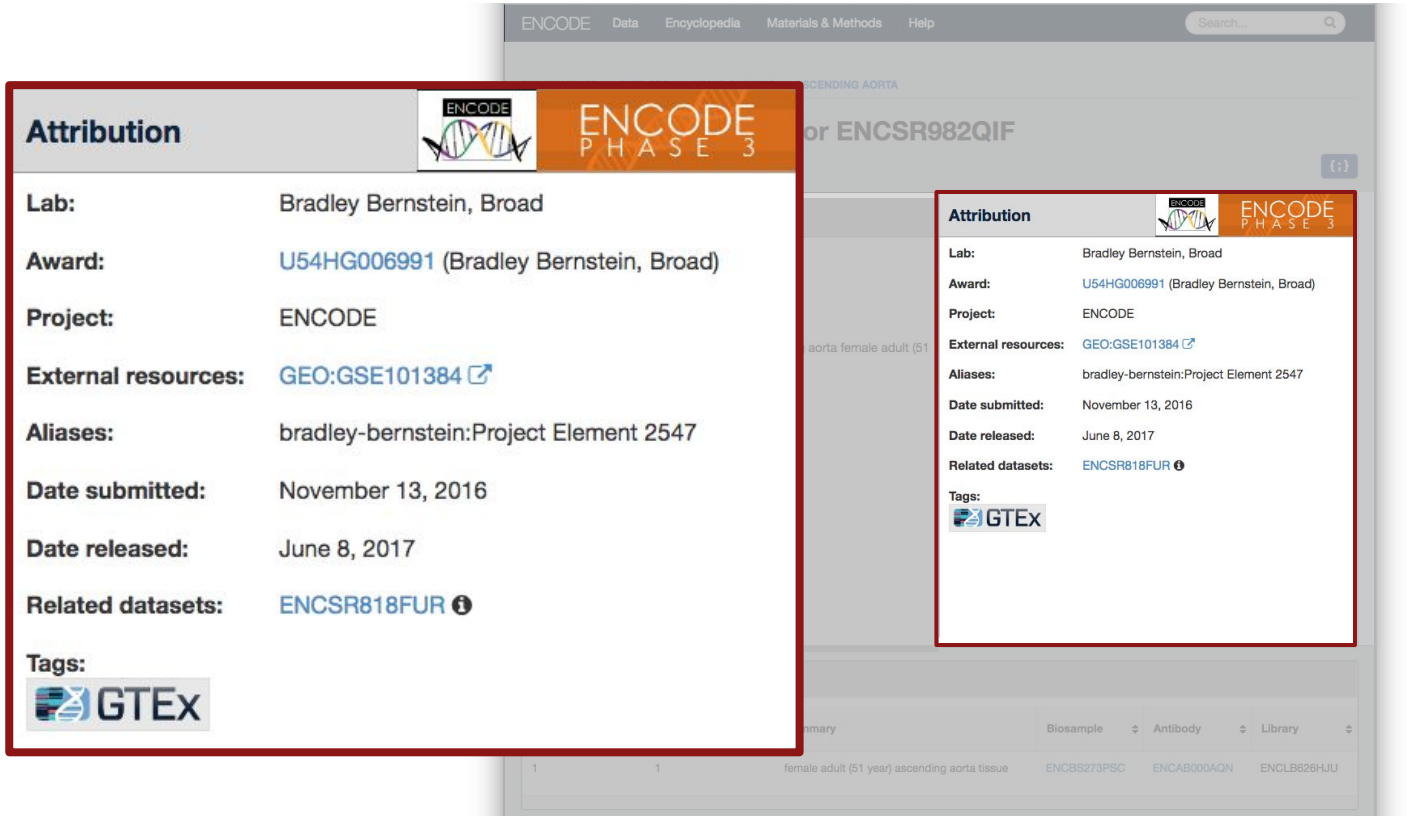
What experiment was done



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

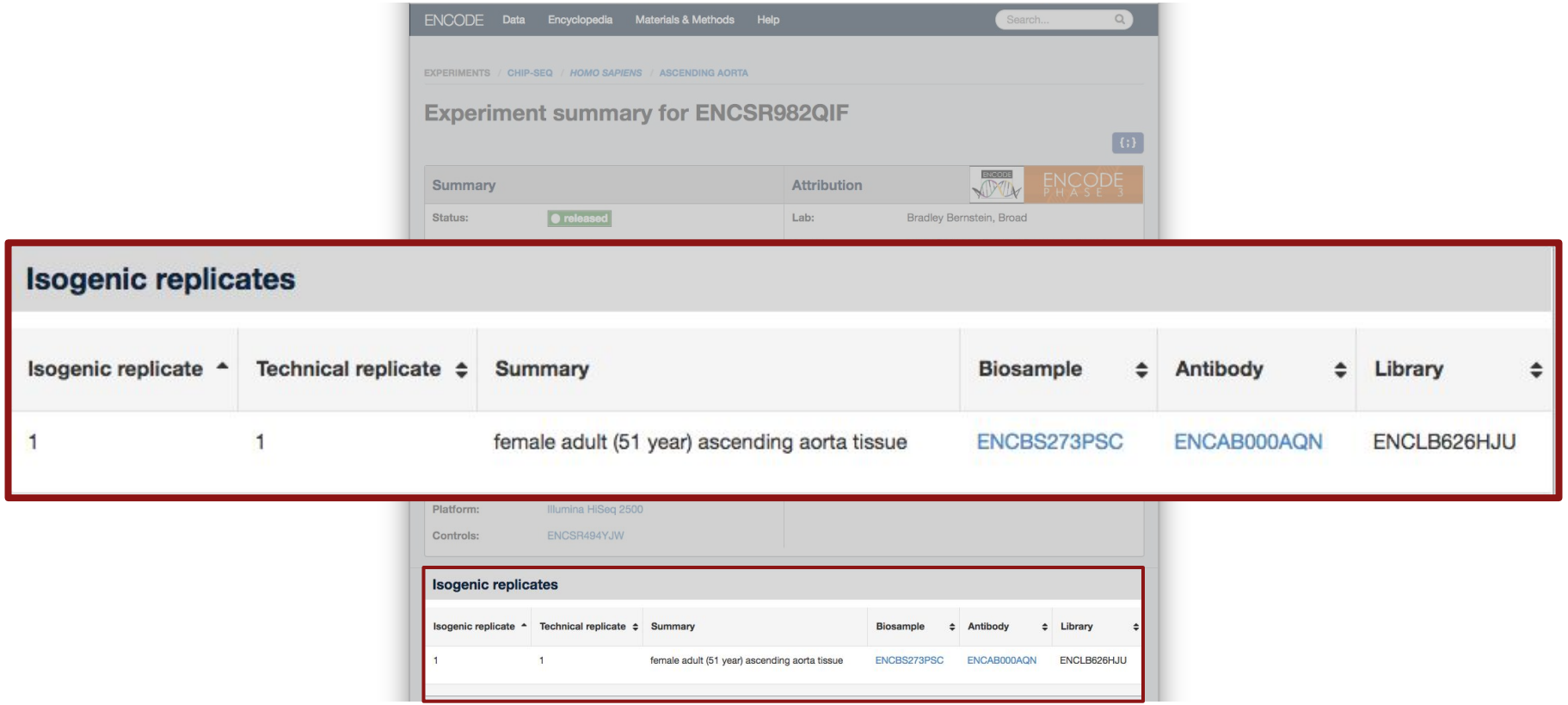
Experiment grouping



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

Experimental entities



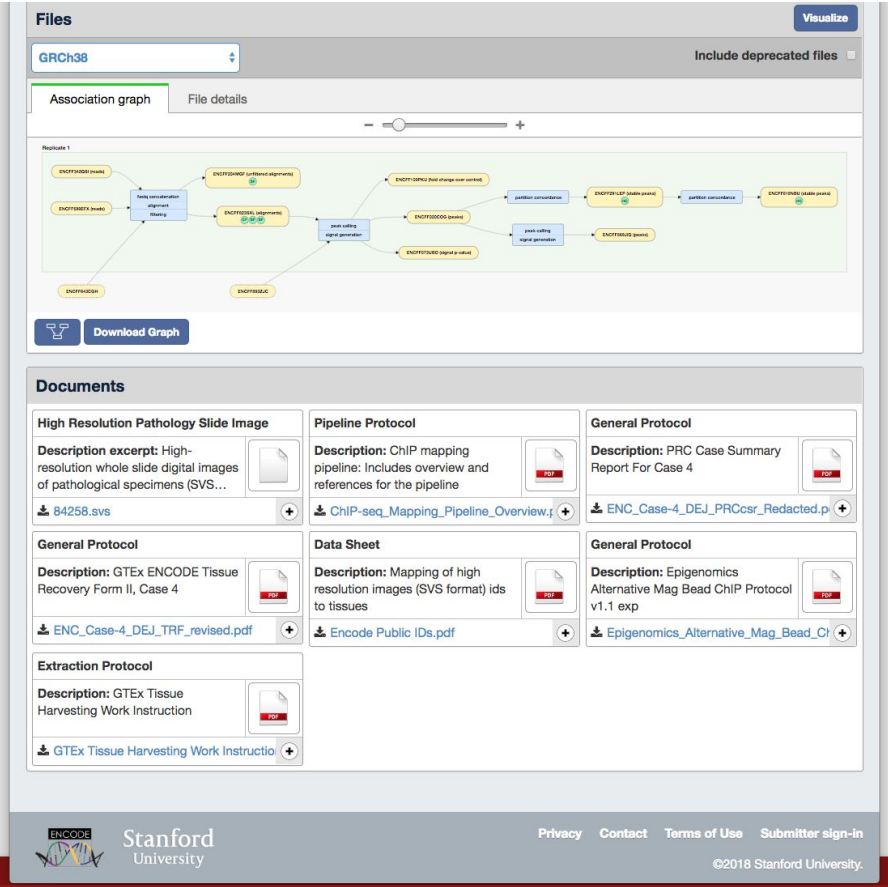
Isogenic replicates

Isogenic replicate	Technical replicate	Summary	Biosample	Antibody	Library
1	1	female adult (51 year) ascending aorta tissue	ENCBS273PSC	ENCAB000AQN	ENCLB626HJU

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

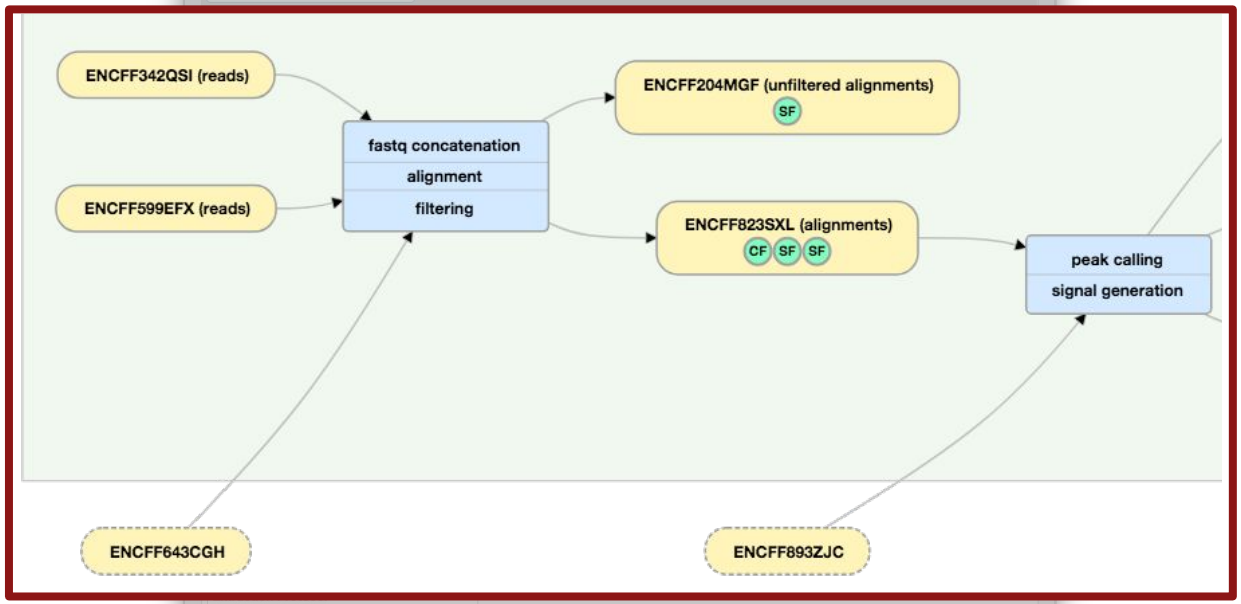
Scrolling down...



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

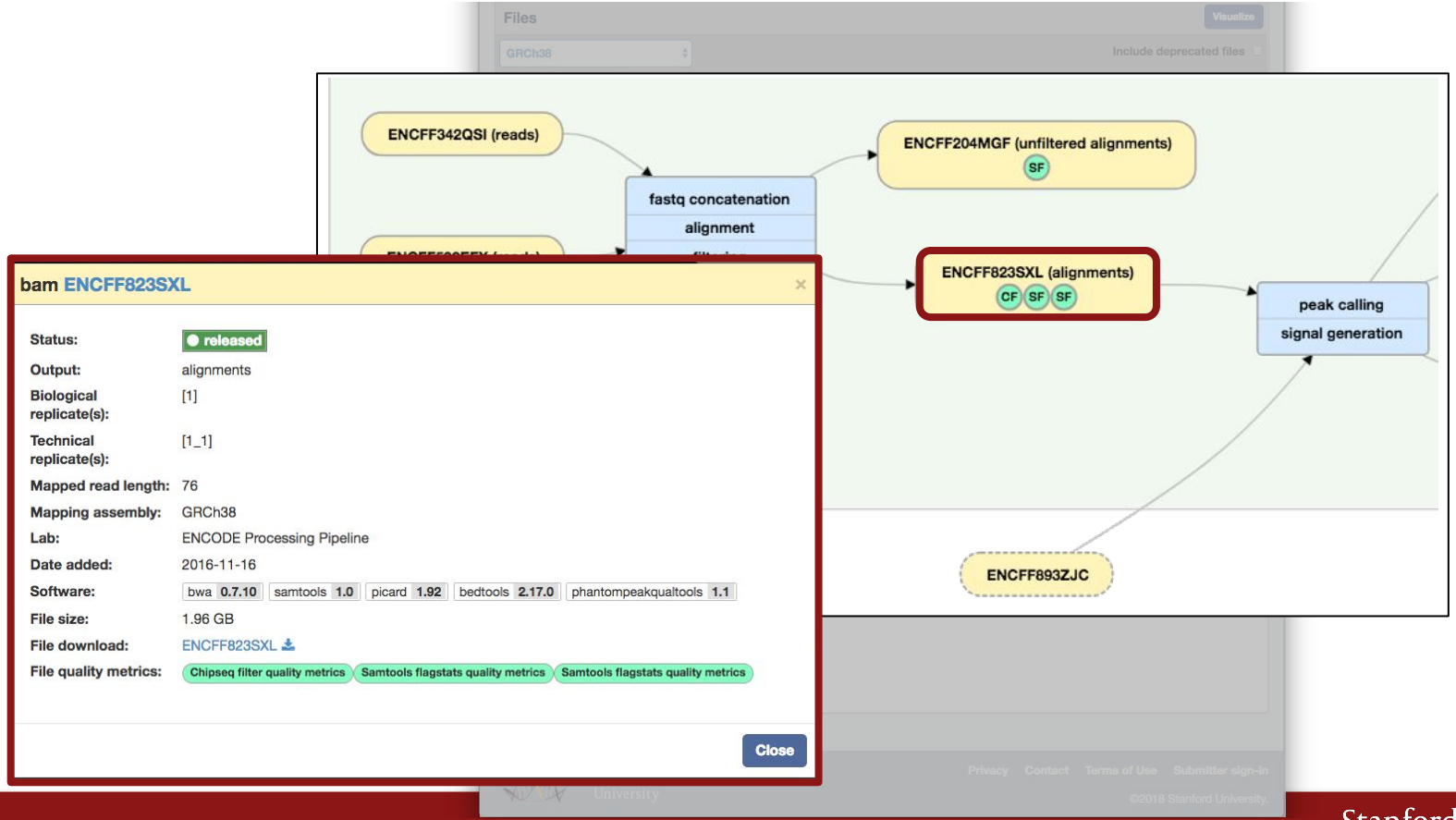
Data provenance



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

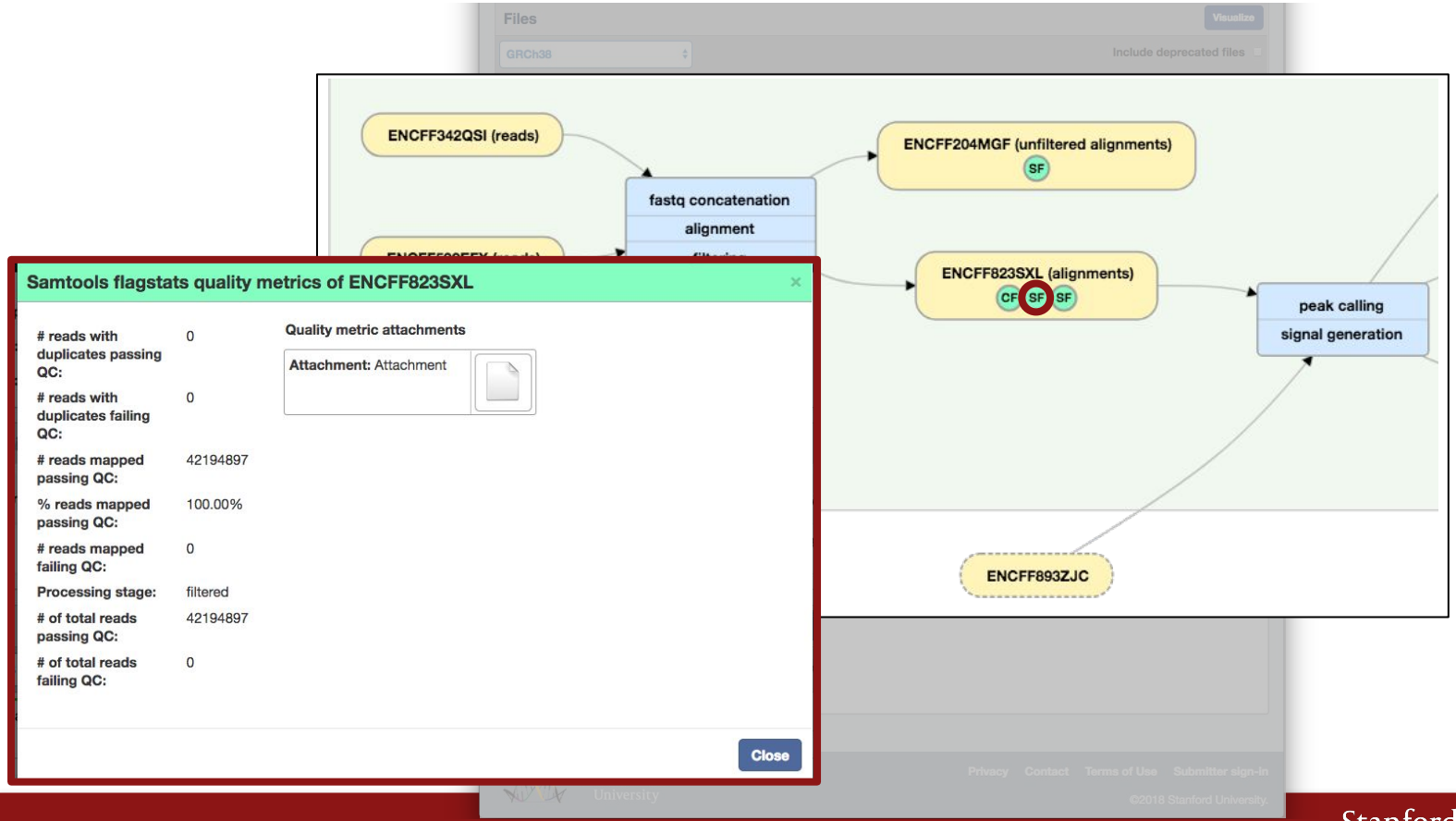
File details



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

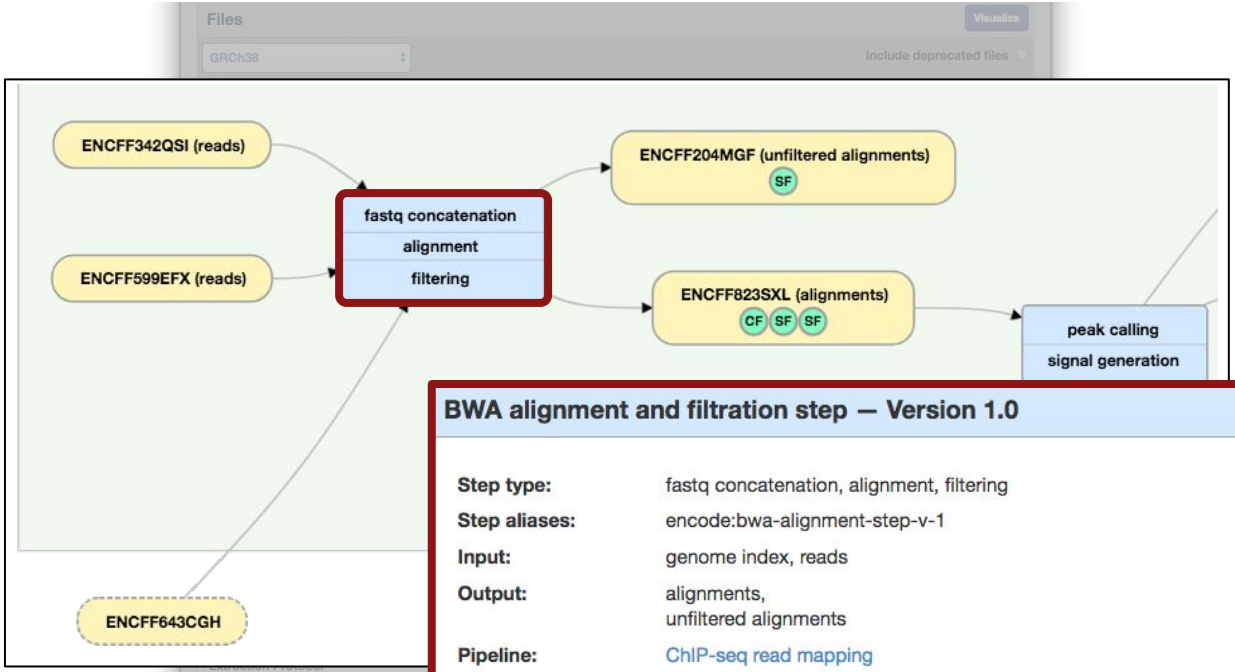
QC metric bubbles



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

Pipeline step metadata



BWA alignment and filtration step — Version 1.0

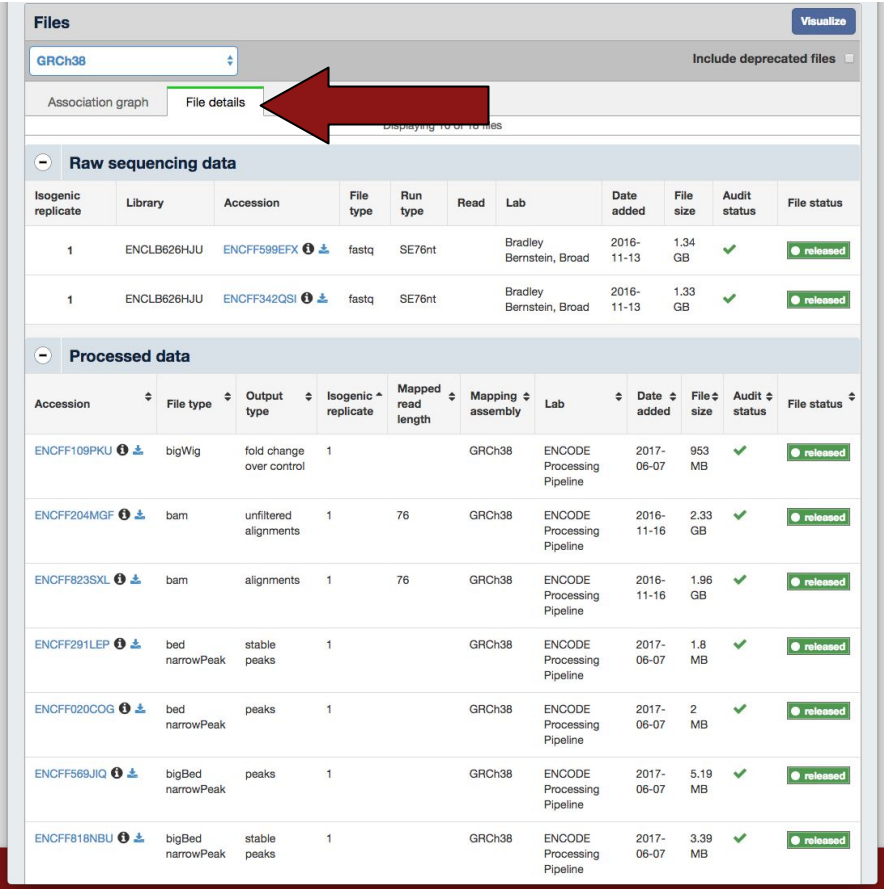
Step type:	fastq concatenation, alignment, filtering
Step aliases:	encode:bwa-alignment-step-v-1
Input:	genome index, reads
Output:	alignments, unfiltered alignments
Pipeline:	ChIP-seq read mapping
Software:	bwa 0.7.10 · samtools 1.0 · picard 1.92 · bedtools 2.17.0 · phantompeakqualtools 1.1

Close

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

File graph <-> table



Files

Visualize

GRCh38

Include deprecated files

Association graph | File details

Displaying 10 of 10 files

Raw sequencing data

Isogenic replicate	Library	Accession	File type	Run type	Read	Lab	Date added	File size	Audit status	File status
1	ENCLB626HJU	ENCFF599EFX	fastq	SE76nt		Bradley Bernstein, Broad	2016-11-13	1.34 GB	✔	released
1	ENCLB626HJU	ENCFF342QSI	fastq	SE76nt		Bradley Bernstein, Broad	2016-11-13	1.33 GB	✔	released

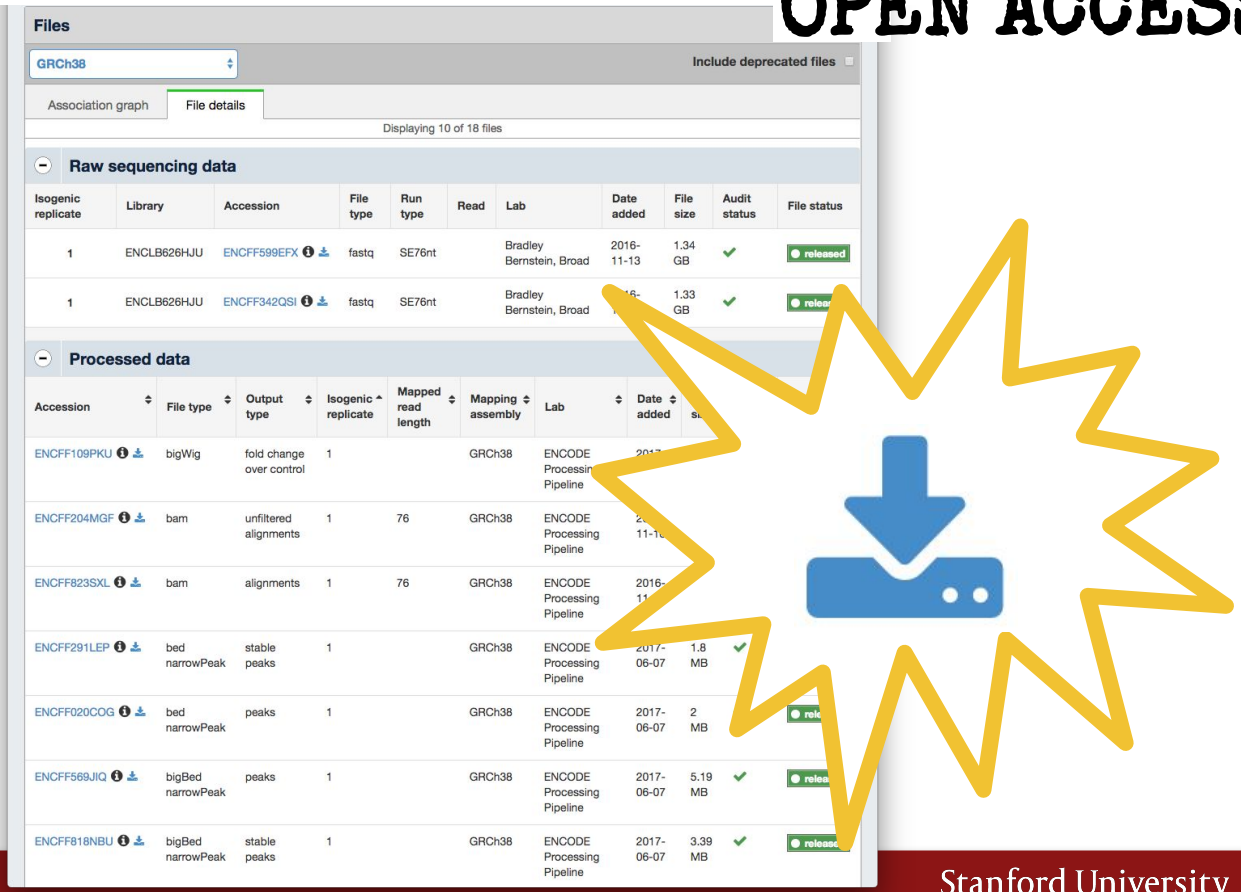
Processed data

Accession	File type	Output type	Isogenic replicate	Mapped read length	Mapping assembly	Lab	Date added	File size	Audit status	File status
ENCFF109PKU	bigWig	fold change over control	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	953 MB	✔	released
ENCFF204MGF	bam	unfiltered alignments	1	76	GRCh38	ENCODE Processing Pipeline	2016-11-16	2.33 GB	✔	released
ENCFF823SXL	bam	alignments	1	76	GRCh38	ENCODE Processing Pipeline	2016-11-16	1.96 GB	✔	released
ENCFF291LEP	bed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	1.8 MB	✔	released
ENCFF020COG	bed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	2 MB	✔	released
ENCFF569JIQ	bigBed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	5.19 MB	✔	released
ENCFF818NBU	bigBed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	3.39 MB	✔	released

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

ENCODE IS ALL OPEN ACCESS!!!!



Files

GRCh38

Include deprecated files

Association graph | File details

Displaying 10 of 18 files

Raw sequencing data

Isogenic replicate	Library	Accession	File type	Run type	Read	Lab	Date added	File size	Audit status	File status
1	ENCLB626HJU	ENCFF599EFX	fastq	SE76nt		Bradley Bernstein, Broad	2016-11-13	1.34 GB	✔	released
1	ENCLB626HJU	ENCFF342QSI	fastq	SE76nt		Bradley Bernstein, Broad	2016-	1.33 GB	✔	released

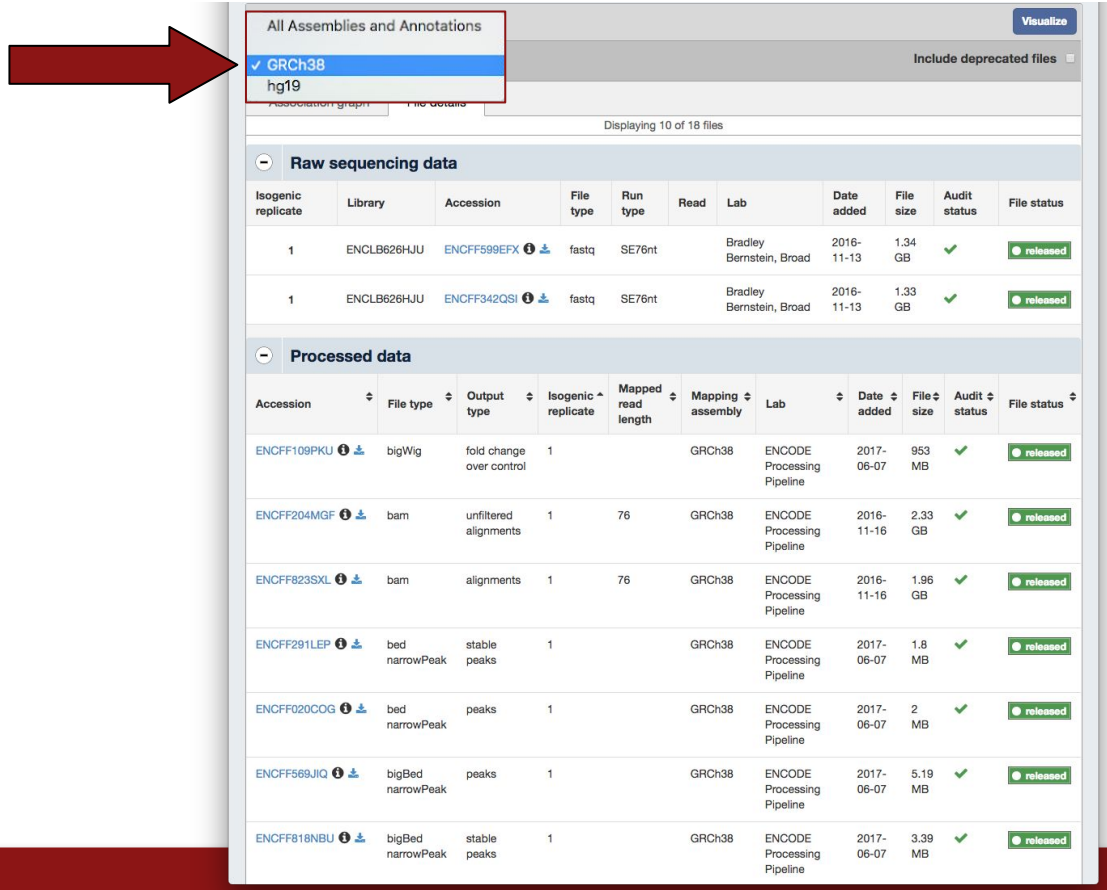
Processed data

Accession	File type	Output type	Isogenic replicate	Mapped read length	Mapping assembly	Lab	Date added			
ENCFF109PKU	bigWig	fold change over control	1		GRCh38	ENCODE Processing Pipeline	2017-			
ENCFF204MGF	bam	unfiltered alignments	1	76	GRCh38	ENCODE Processing Pipeline	2017-11-1			
ENCFF823SXL	bam	alignments	1	76	GRCh38	ENCODE Processing Pipeline	2016-11			
ENCFF291LEP	bed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	1.8 MB	✔	
ENCFF020COG	bed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	2 MB		released
ENCFF569JIQ	bigBed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	5.19 MB	✔	released
ENCFF818NBU	bigBed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	3.39 MB	✔	released

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

Assembly drop-down



Raw sequencing data

Isogenic replicate	Library	Accession	File type	Run type	Read	Lab	Date added	File size	Audit status	File status
1	ENCLB626HJU	ENCFF599EFX	fastq	SE76nt		Bradley Bernstein, Broad	2016-11-13	1.34 GB	✔	released
1	ENCLB626HJU	ENCFF342QSI	fastq	SE76nt		Bradley Bernstein, Broad	2016-11-13	1.33 GB	✔	released

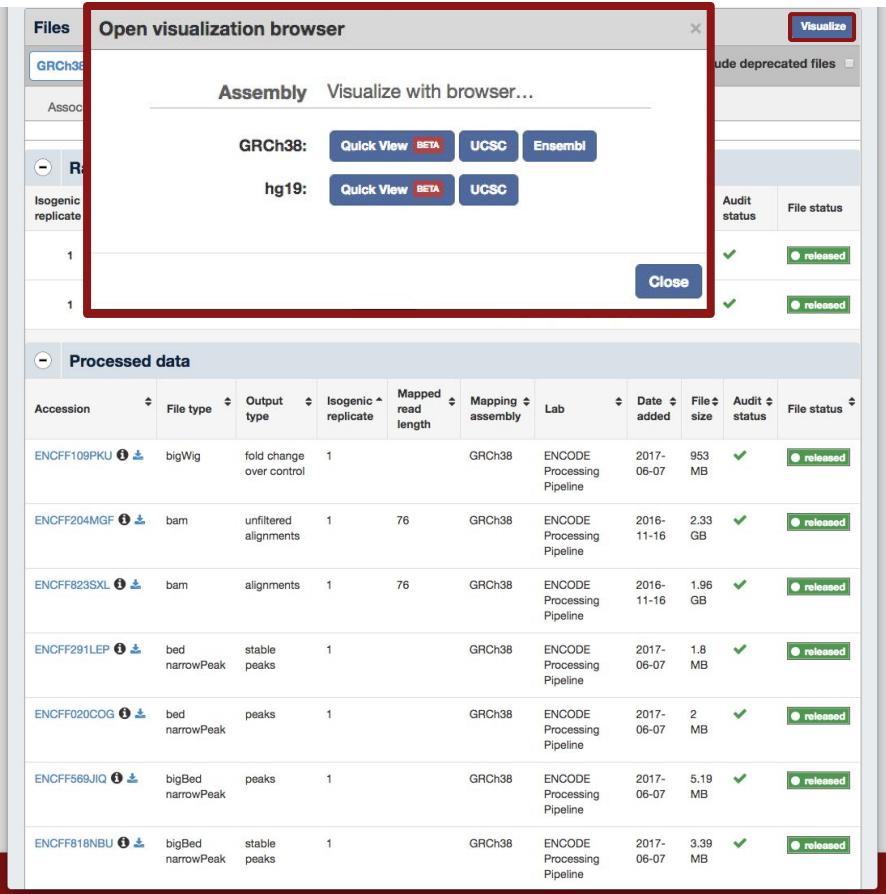
Processed data

Accession	File type	Output type	Isogenic replicate	Mapped read length	Mapping assembly	Lab	Date added	File size	Audit status	File status
ENCFF109PKU	bigWig	fold change over control	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	953 MB	✔	released
ENCFF204MGF	bam	unfiltered alignments	1	76	GRCh38	ENCODE Processing Pipeline	2016-11-16	2.33 GB	✔	released
ENCFF823SXL	bam	alignments	1	76	GRCh38	ENCODE Processing Pipeline	2016-11-16	1.96 GB	✔	released
ENCFF291LEP	bed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	1.8 MB	✔	released
ENCFF020COG	bed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	2 MB	✔	released
ENCFF569JIQ	bigBed narrowPeak	peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	5.19 MB	✔	released
ENCFF818NBU	bigBed narrowPeak	stable peaks	1		GRCh38	ENCODE Processing Pipeline	2017-06-07	3.39 MB	✔	released

An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

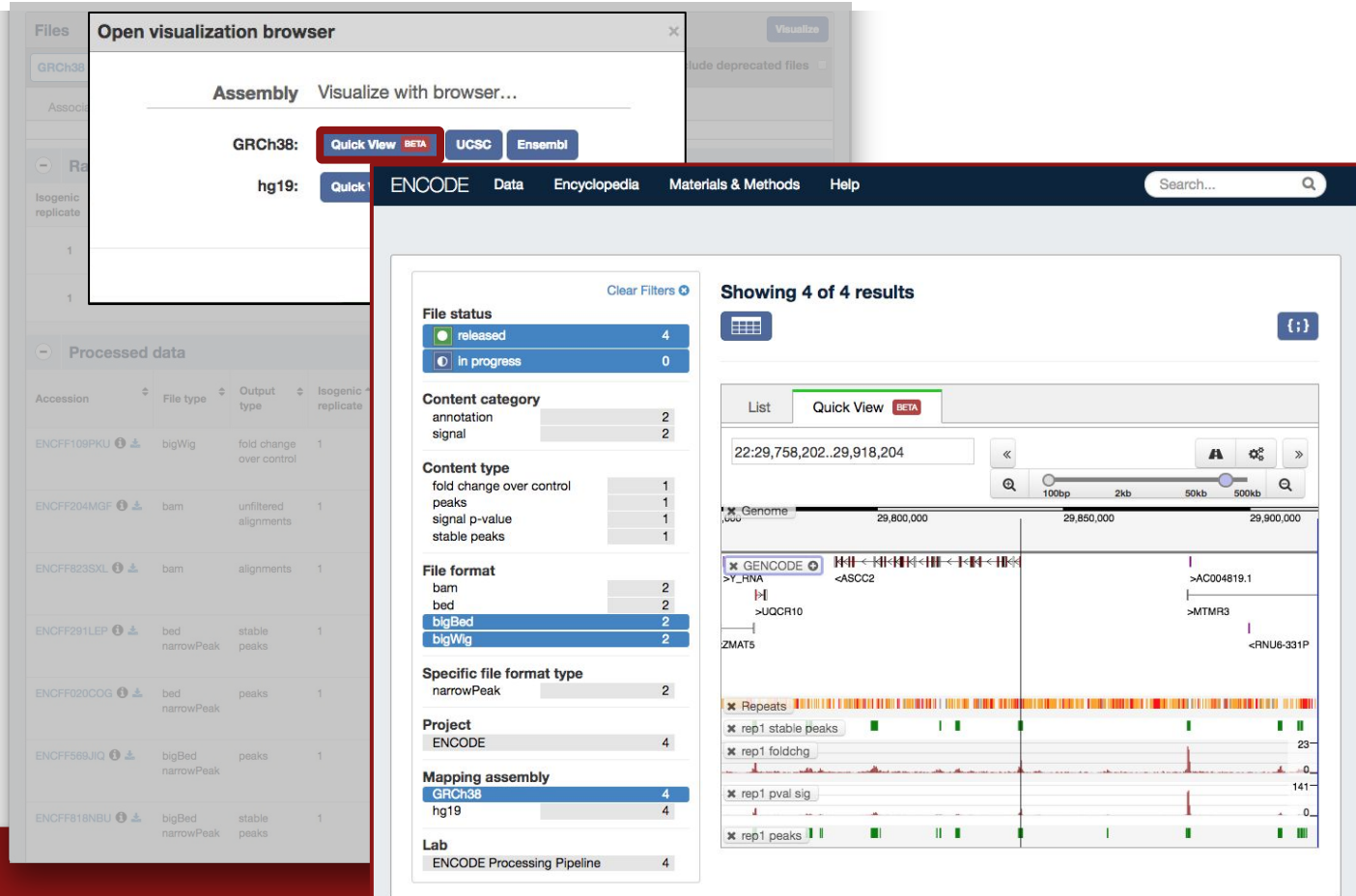
Visualize options



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

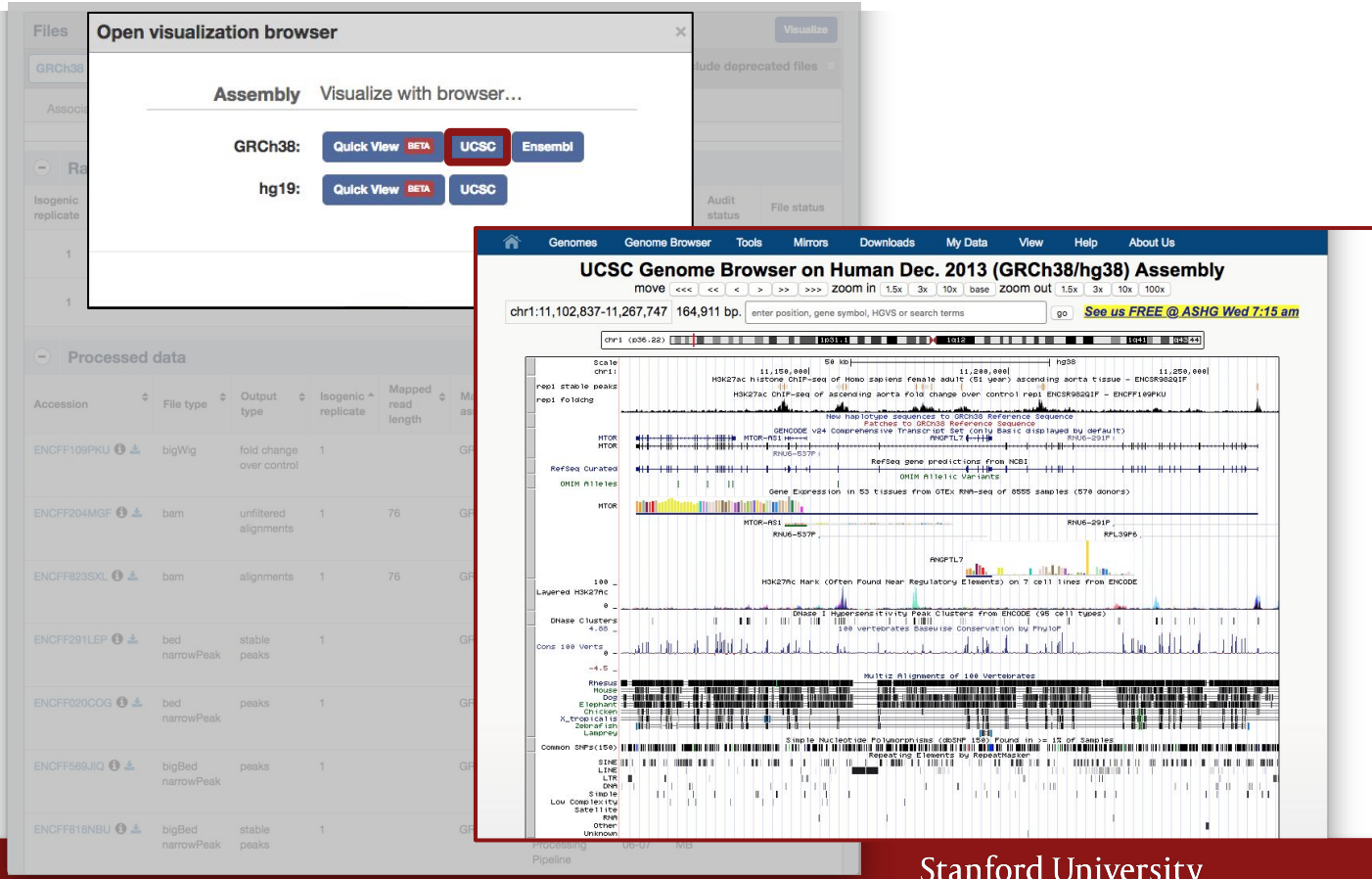
Quick viz on ENCODE portal



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

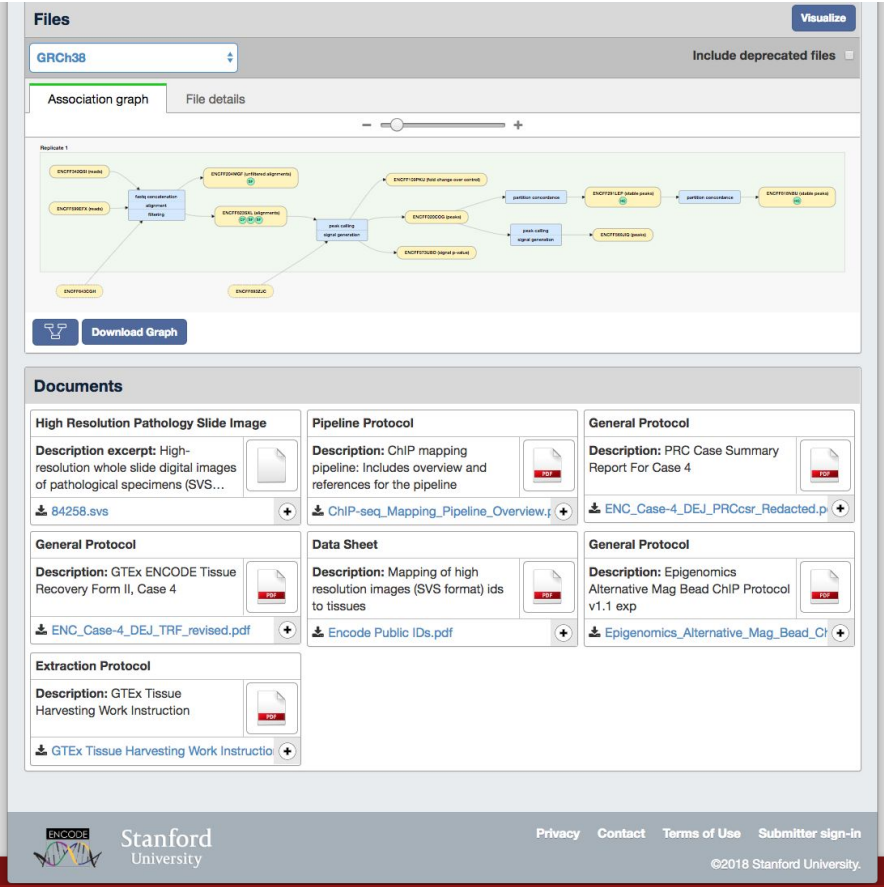
Viz at UCSC/Ensembl



An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF

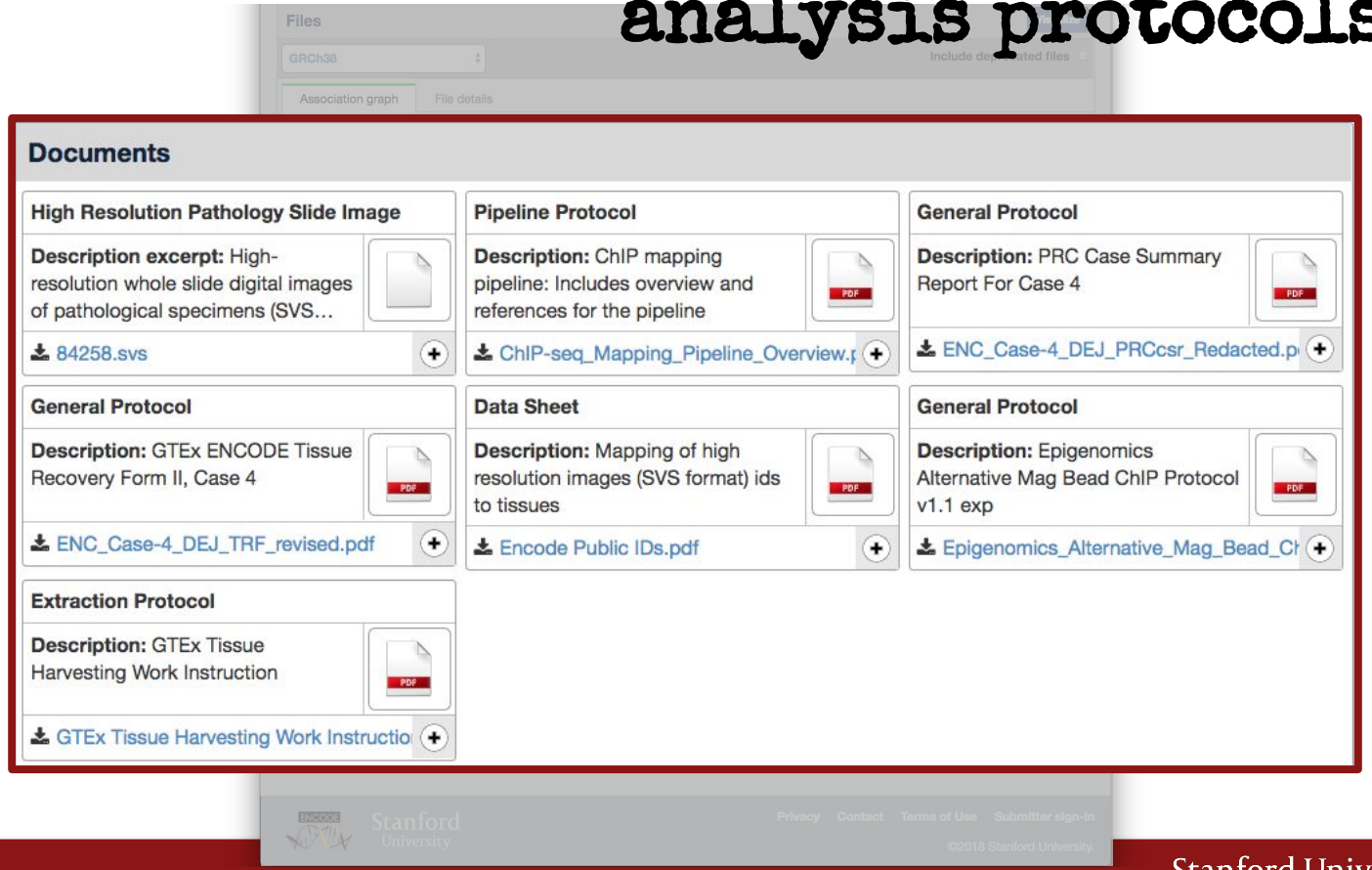
Supplementary docs



An ENCODE experiment

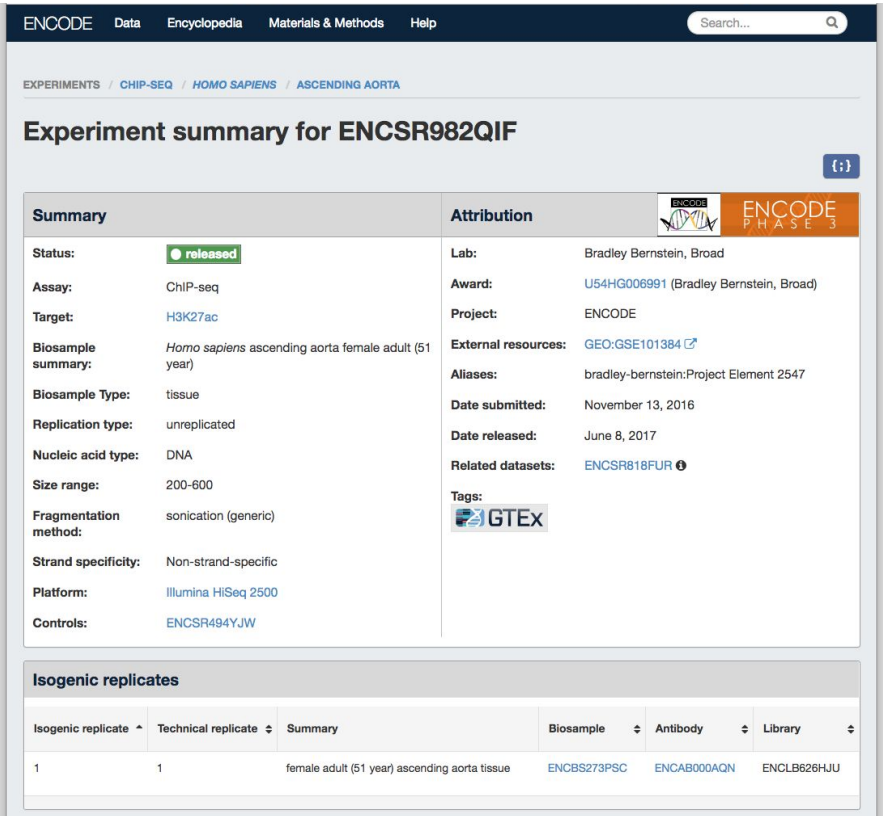
https://www.encodeproject.org/ENCSR982QIF

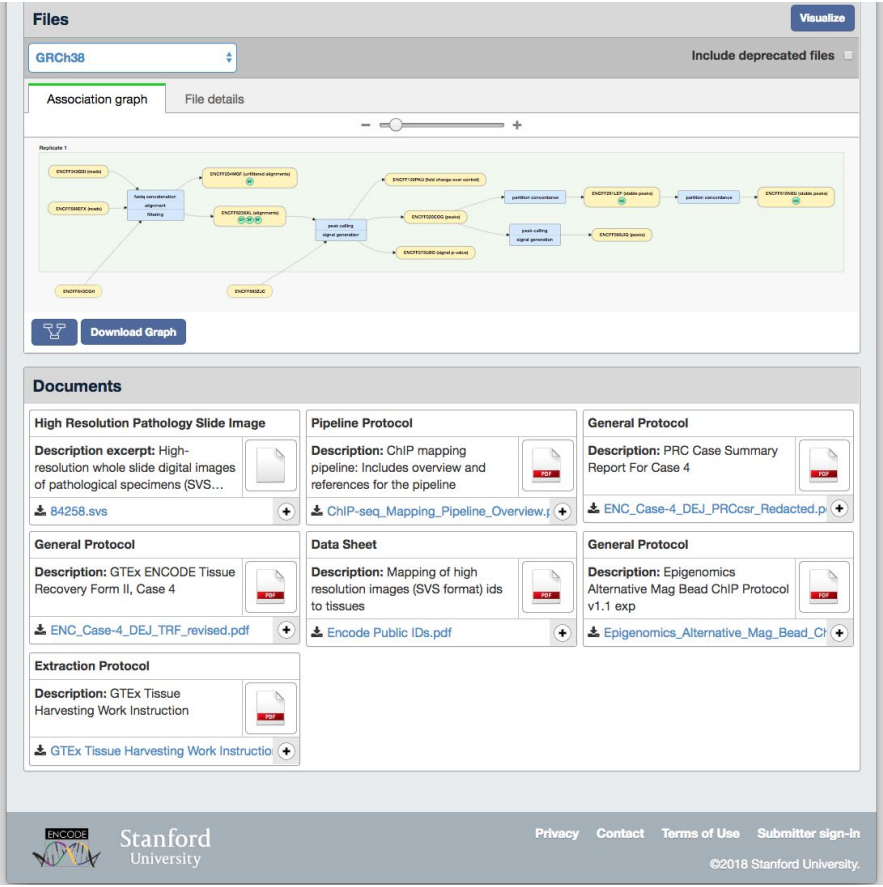
Growth, library prep, analysis protocols, etc.



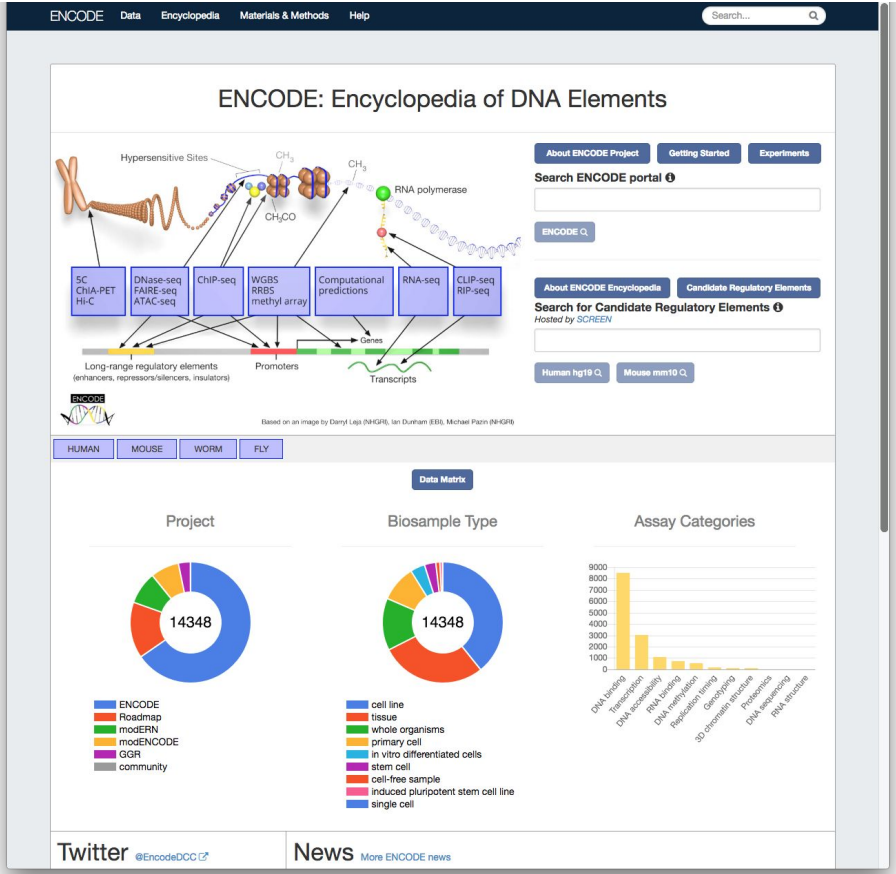
An ENCODE experiment

https://www.encodeproject.org/ENCSR982QIF



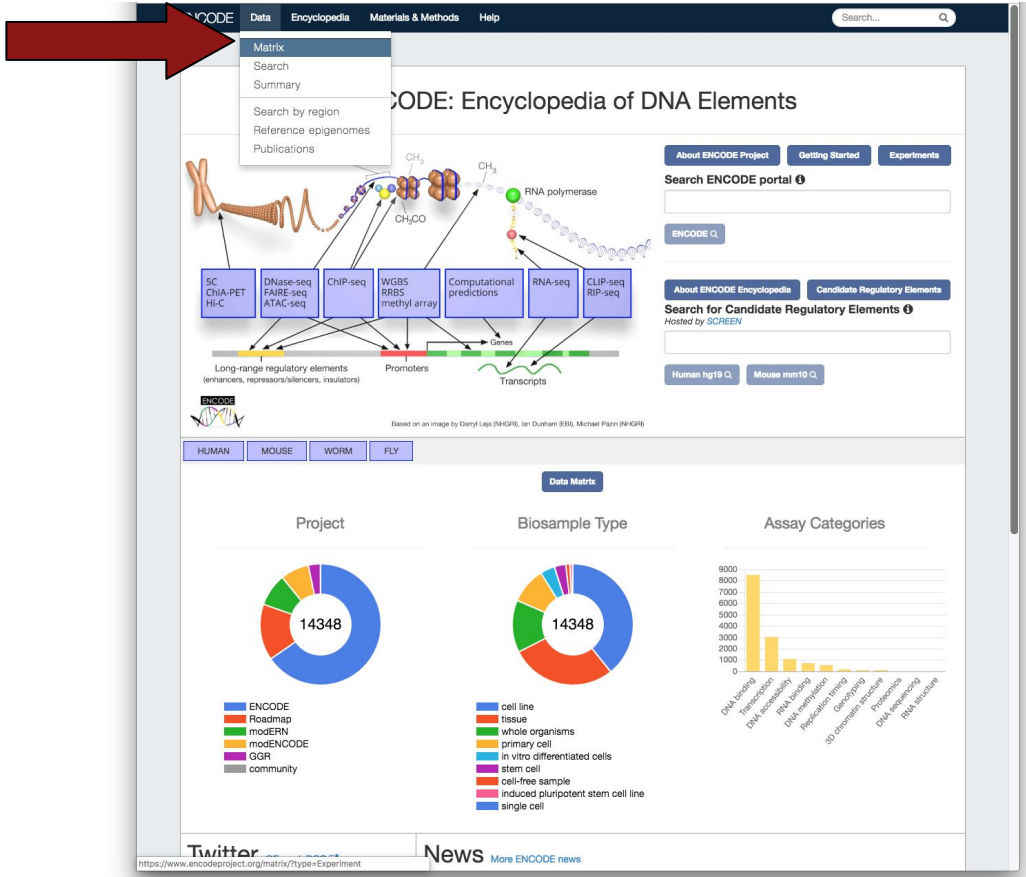


Browsing the ENCODE portal



Browsing the ENCODE portal

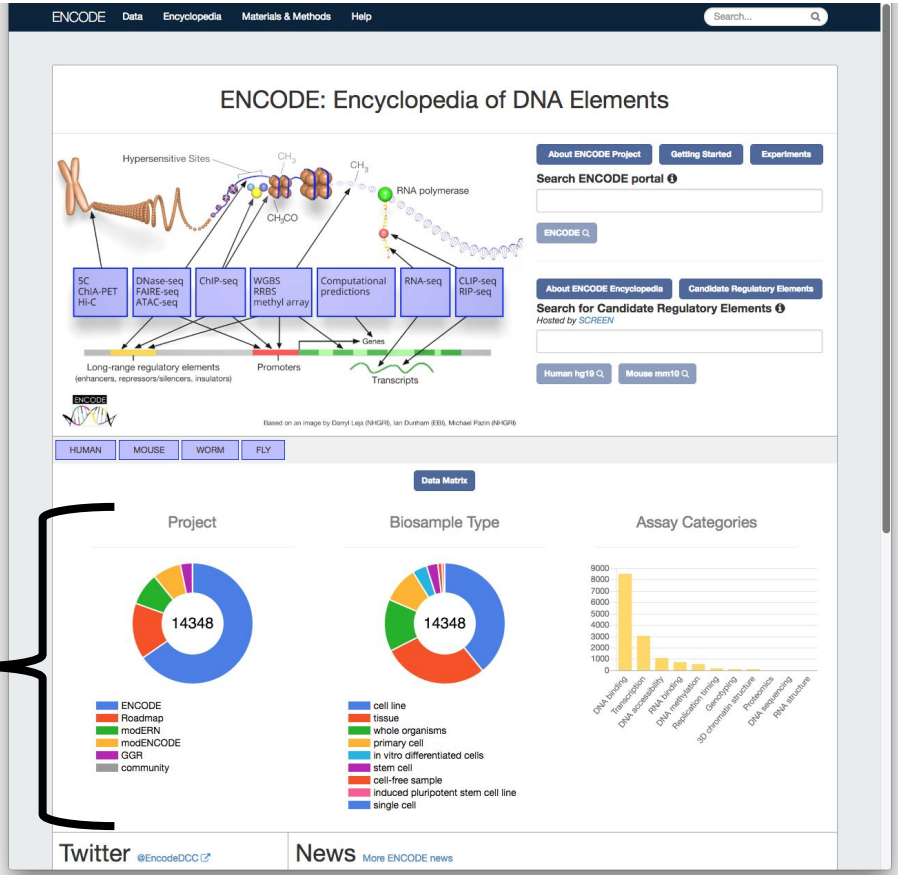
Jumping in: matrix



Browsing the ENCODE portal

Jumping in: matrix

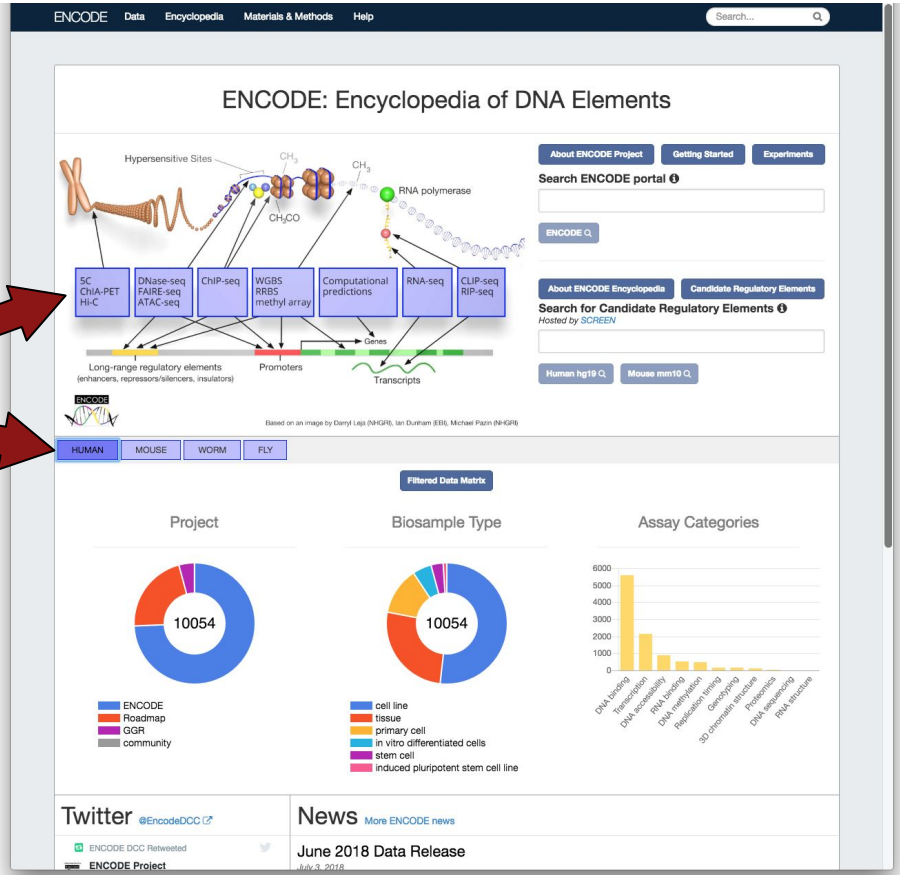
Jump to filtered matrix



Browsing the ENCODE portal

Jumping in: matrix

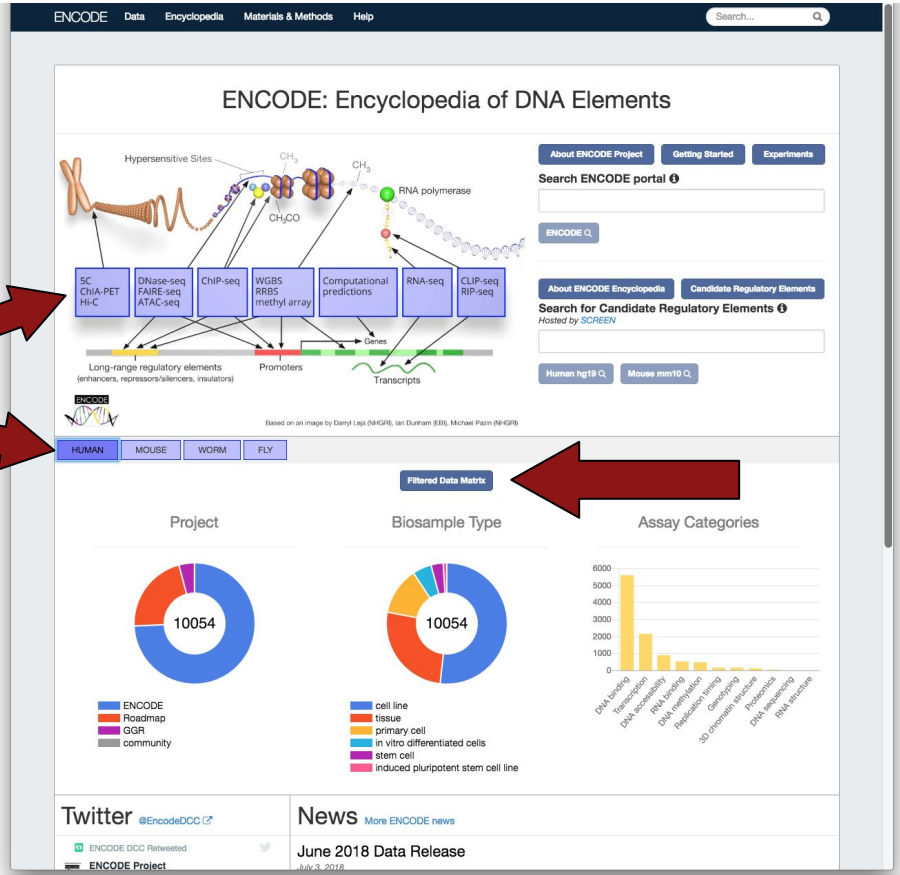
Filter by assay category & organism



Browsing the ENCODE portal

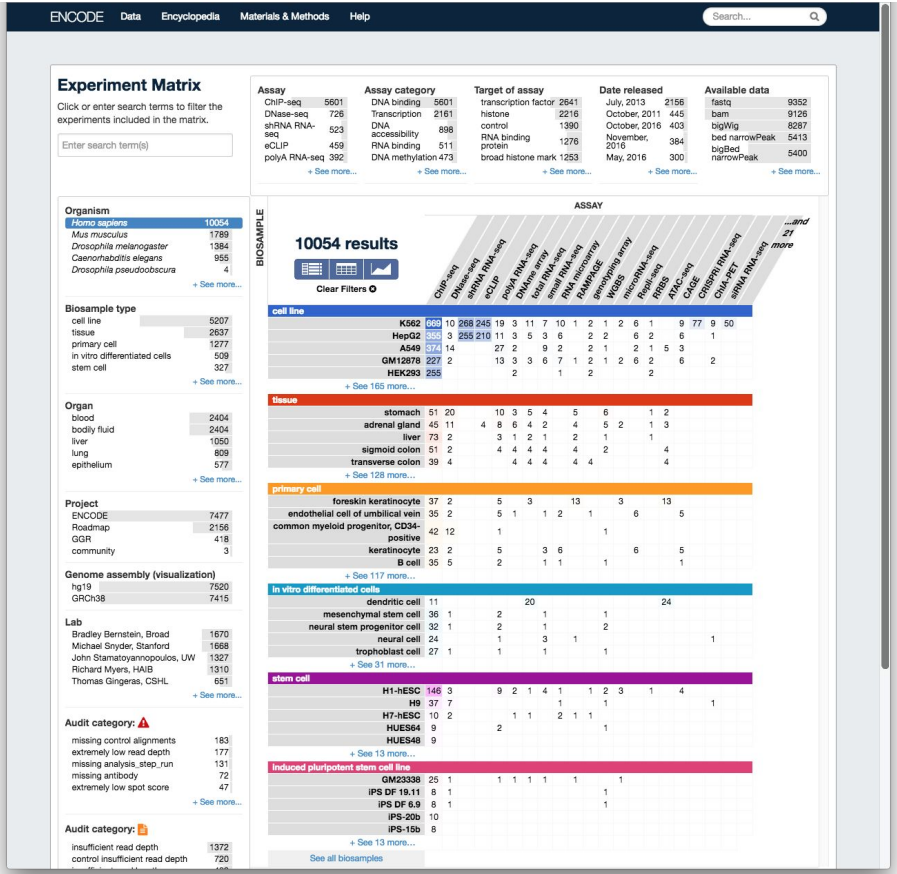
Jumping in: matrix

Filter by assay category & organsim



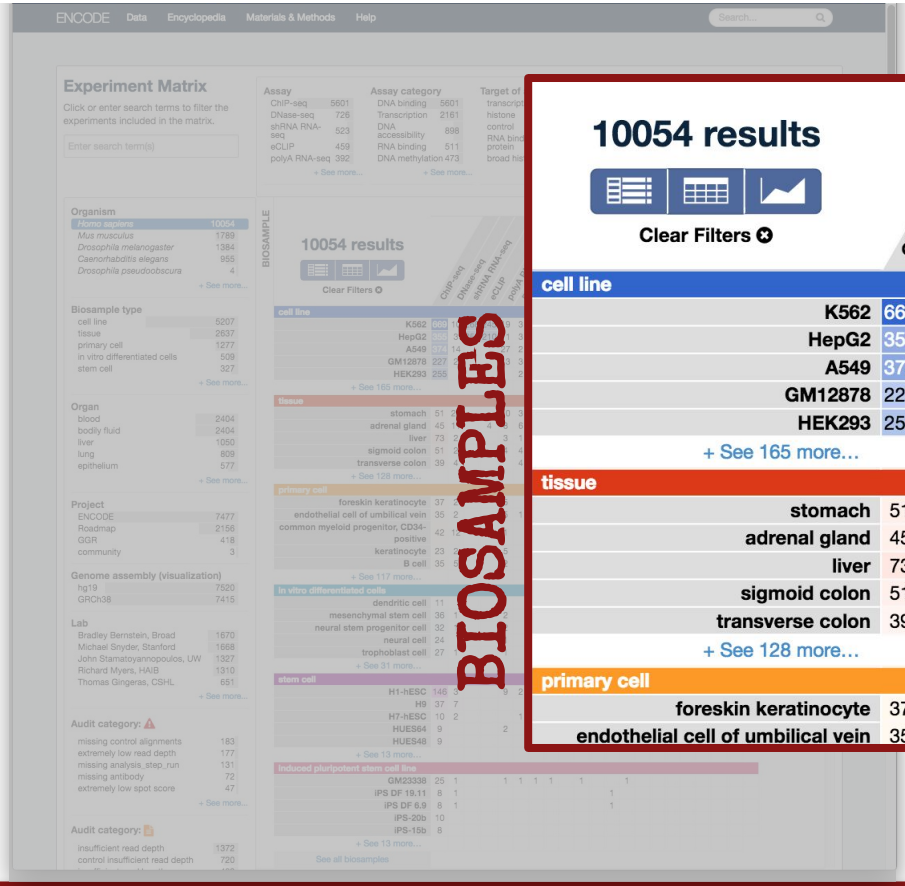
Browsing the ENCODE portal

Experiment matrix



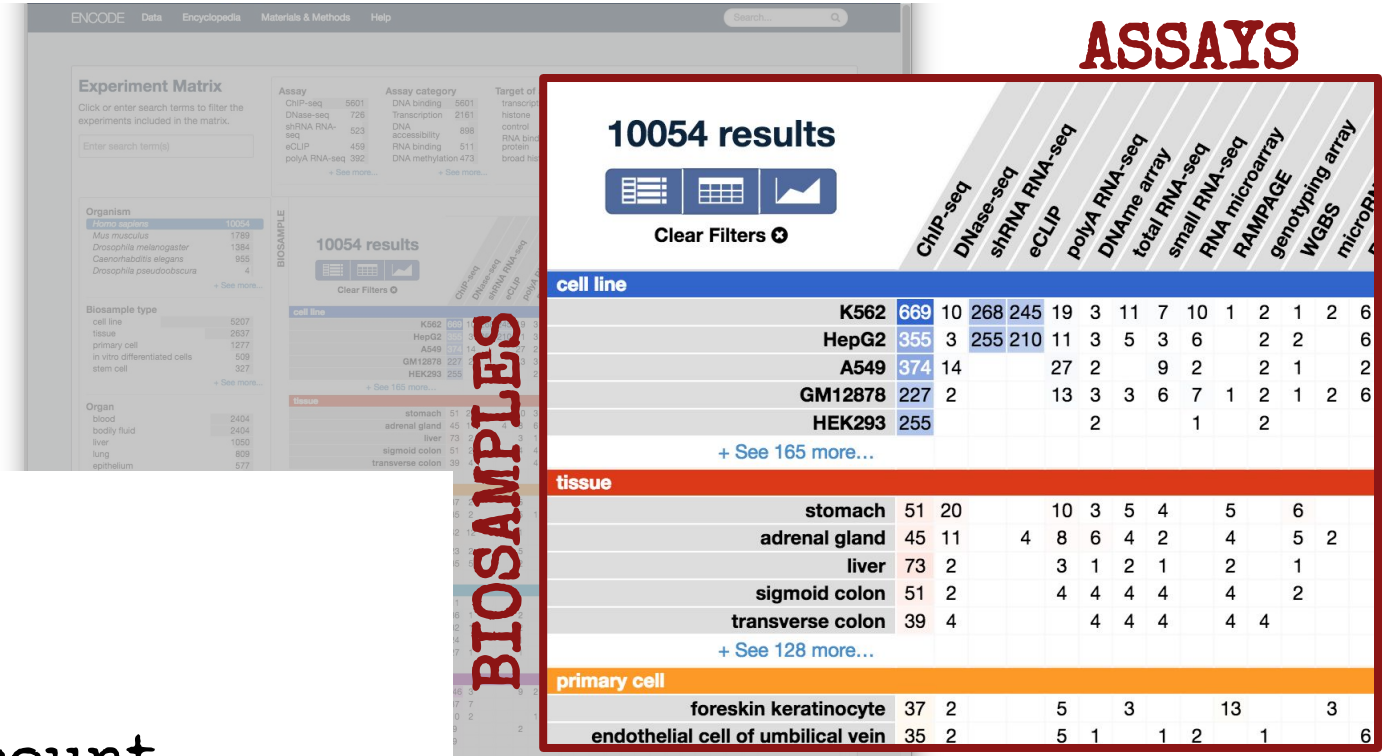
Browsing the ENCODE portal

Experiment matrix



Browsing the ENCODE portal

Experiment matrix



Select:

- Assay
- Biosample
- Experiment count

-> narrowed Search results

Browsing the ENCODE portal

Facets

Assay category

DNA binding	5601
Transcription	2161
DNA accessibility	898
RNA binding	511
DNA methylation	473

+ See more...

Target of assay

transcription factor	2641
histone	2216
control	1390
RNA binding protein	1276
broad histone mark	1253

+ See more...

Organism

Homo sapiens	10054
Mus musculus	1789
Drosophila melanogaster	1384
Caenorhabditis elegans	955
Drosophila pseudoobscura	4

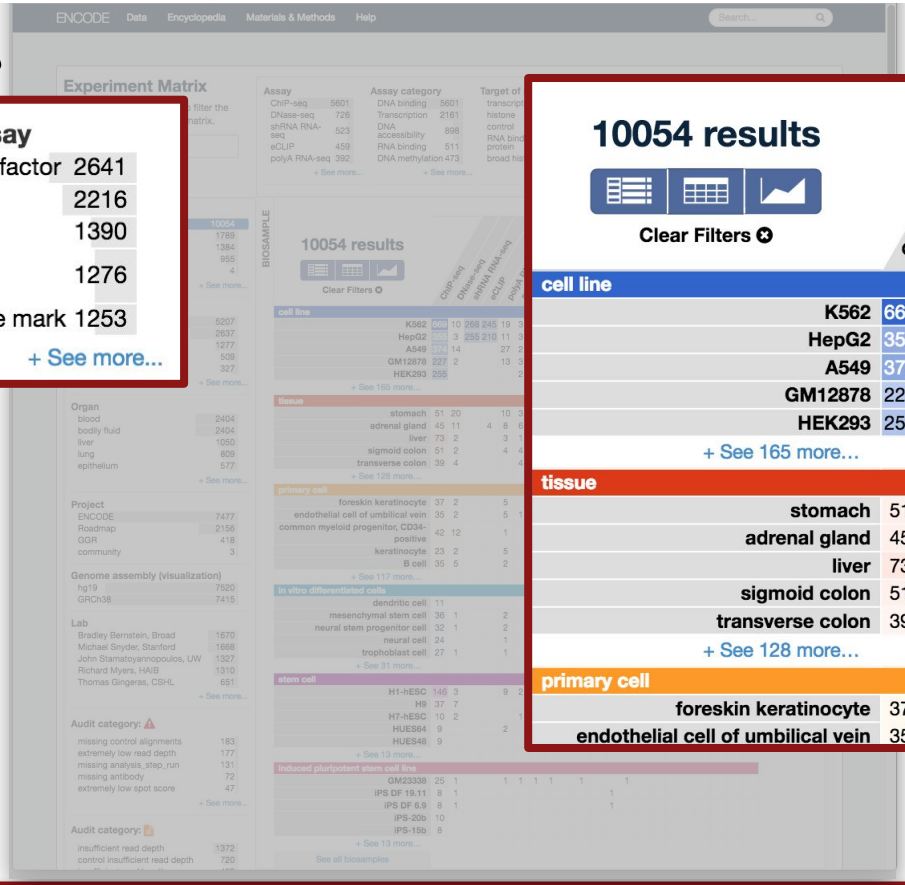
+ See more...

Organ

blood	2404
bodily fluid	2404
liver	1050
lung	809
epithelium	577

+ See more...

Experiment matrix



10054 results

Clear Filters

	ChIP-seq	DNase-seq	shRNA RNA-seq	eCLIP	polyA RNA-seq	DNAme array	total RNA-seq	small RNA-seq	RNA microarray	RAMPAGE	genotyping array	WGBS	microRNA
cell line													
K562	669	10	268	245	19	3	11	7	10	1	2	1	2	6
HepG2	355	3	255	210	11	3	5	3	6		2	2		6
A549	374	14			27	2		9	2		2	1		2
GM12878	227	2			13	3	3	6	7	1	2	1	2	6
HEK293	255					2			1		2			
+ See 165 more...														
tissue														
stomach	51	20			10	3	5	4		5		6		
adrenal gland	45	11		4	8	6	4	2		4		5	2	
liver	73	2			3	1	2	1		2		1		
sigmoid colon	51	2			4	4	4	4		4		2		
transverse colon	39	4				4	4	4		4	4			
+ See 128 more...														
primary cell														
foreskin keratinocyte	37	2			5		3			13			3	
endothelial cell of umbilical vein	35	2			5	1		1	2		1			6

Browsing the ENCODE portal

Experiment matrix

Facets

Assay category

DNA binding	545
Transcription	127
DNA accessibility	76
DNA methylation	32
Genotyping	11

+ See more...

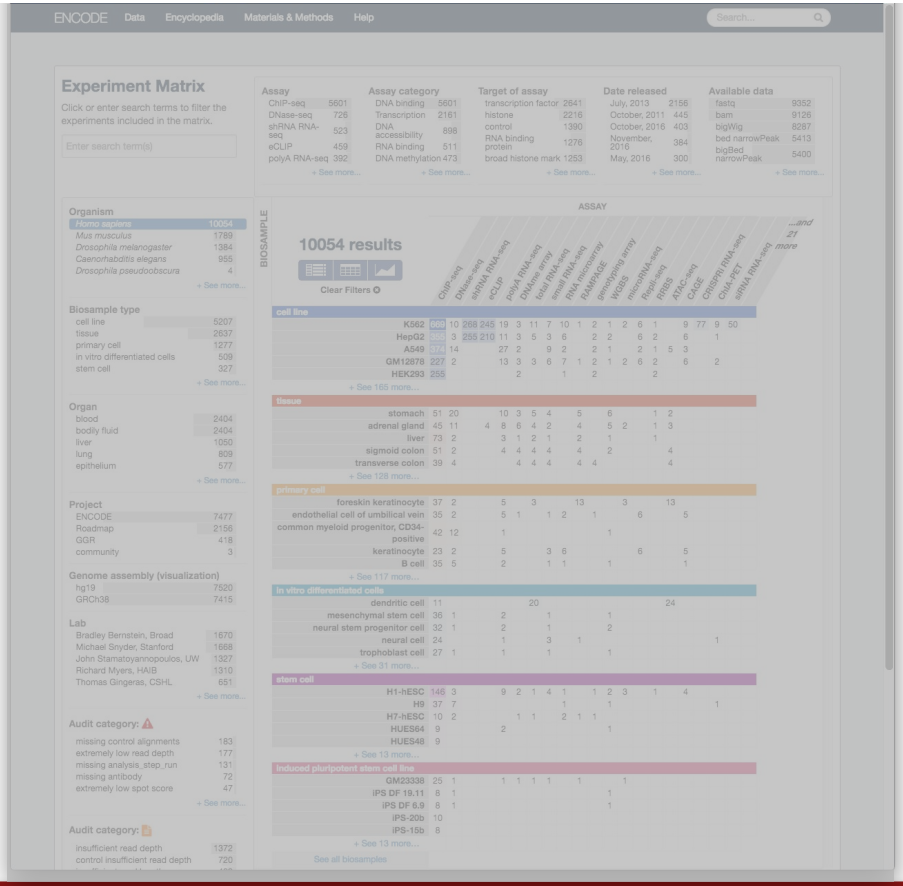
Organism

Homo sapiens	10054
Mus musculus	1789
Drosophila melanogaster	1384
Caenorhabditis elegans	955
Drosophila pseudoobscura	4

+ See more...

Organ

blood	585
bodily fluid	585
liver	337
epithelium	181
skin of body	154
lung	127
connective tissue	121



Browsing the ENCODE portal

Experiment matrix

Assay category

DNA binding	545
Transcription	127
DNA accessibility	76
DNA methylation	32
Genotyping	11

+ See more...

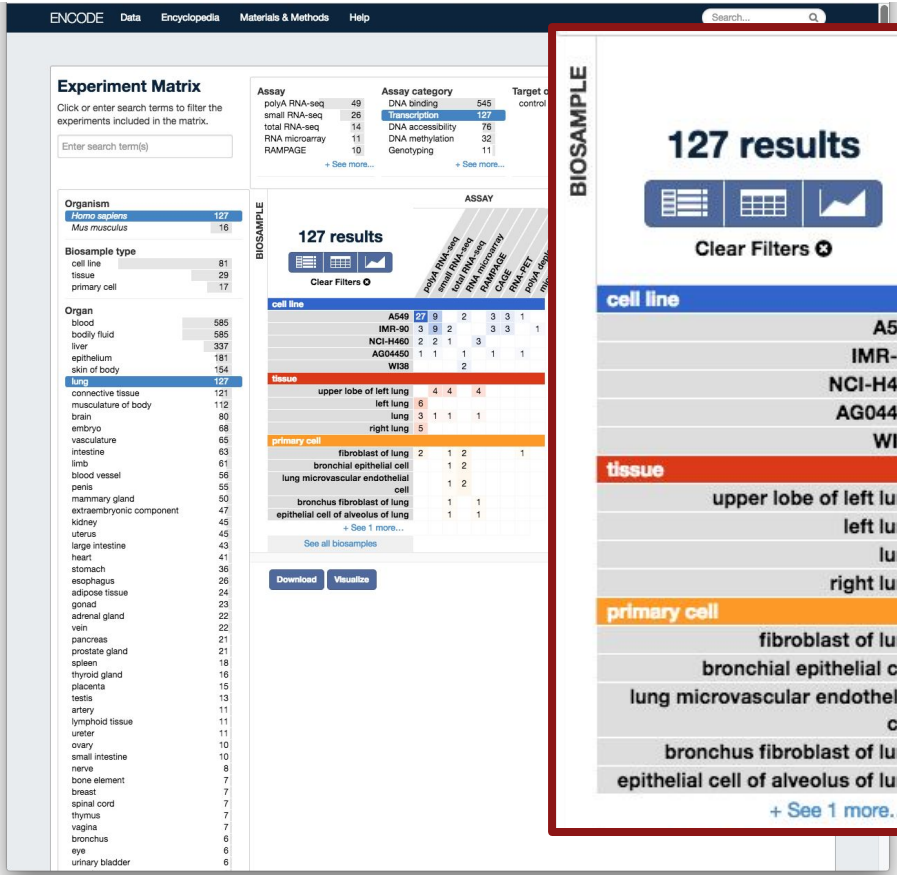
Organism

Homo sapiens	10054
Mus musculus	1789
Drosophila melanogaster	1384
Caenorhabditis elegans	955
Drosophila pseudoobscura	4

+ See more...

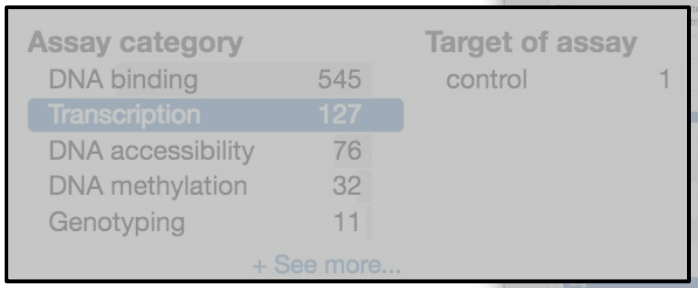
Organ

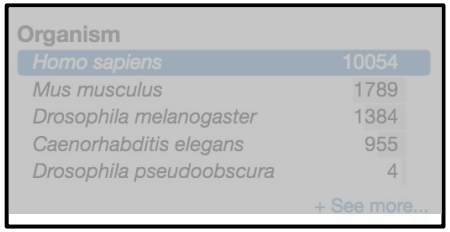
blood	585
bodily fluid	585
liver	337
epithelium	181
skin of body	154
lung	127
connective tissue	121

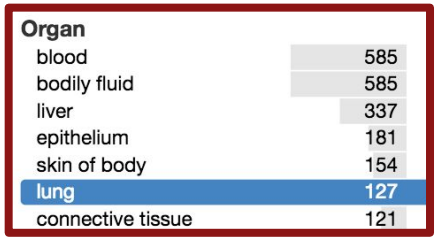


Browsing the ENCODE portal

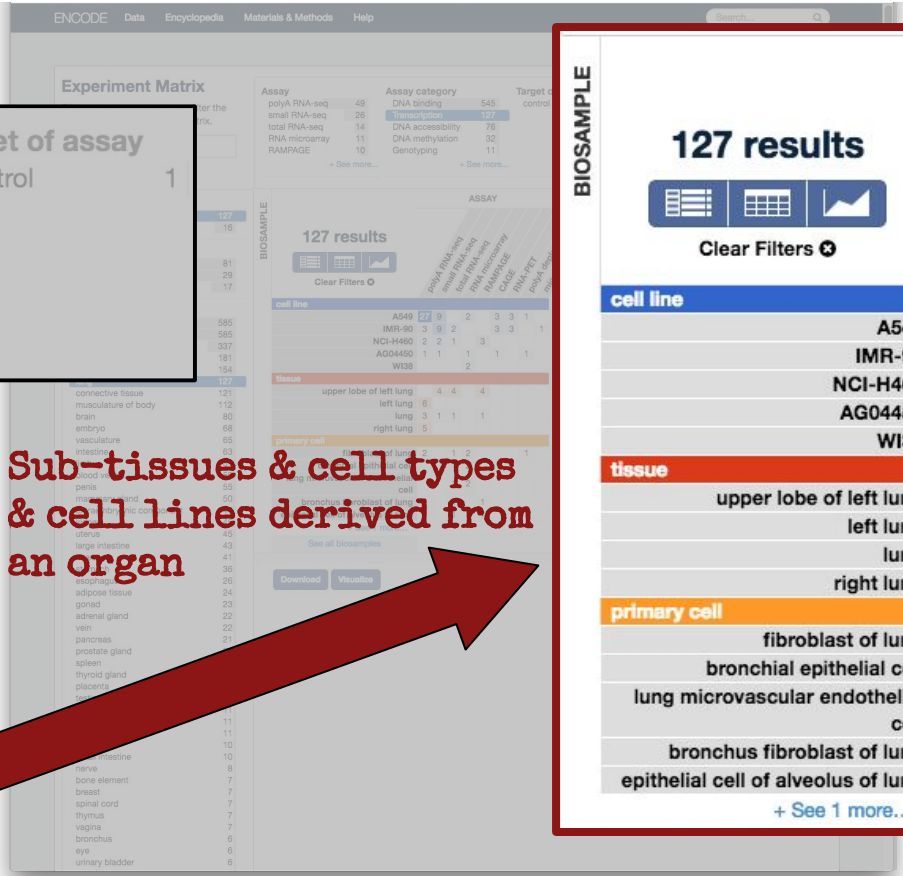
Use of ontology

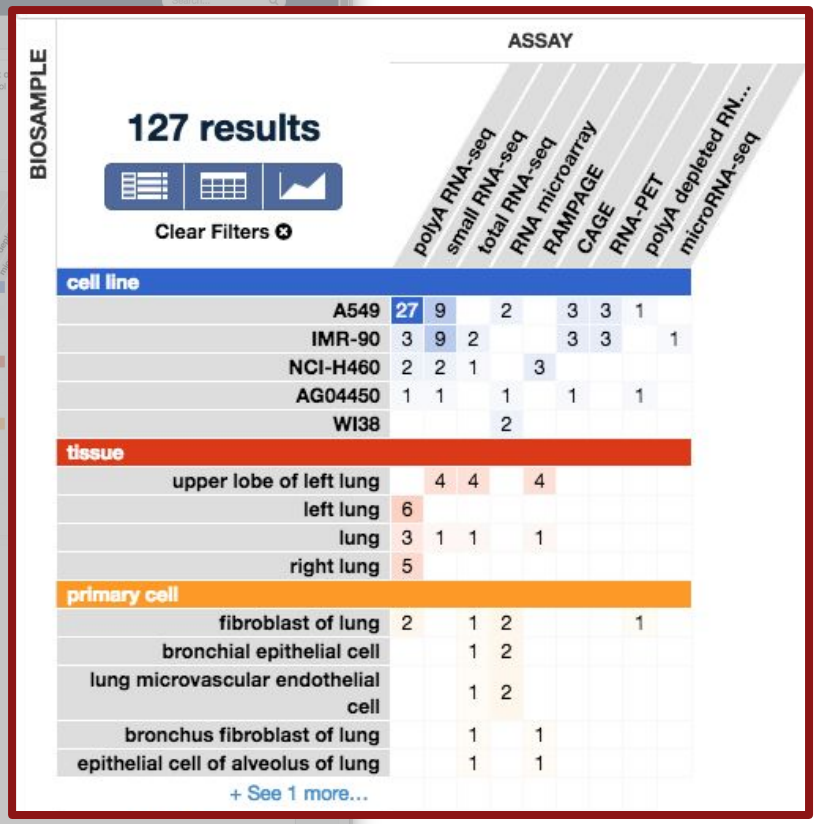






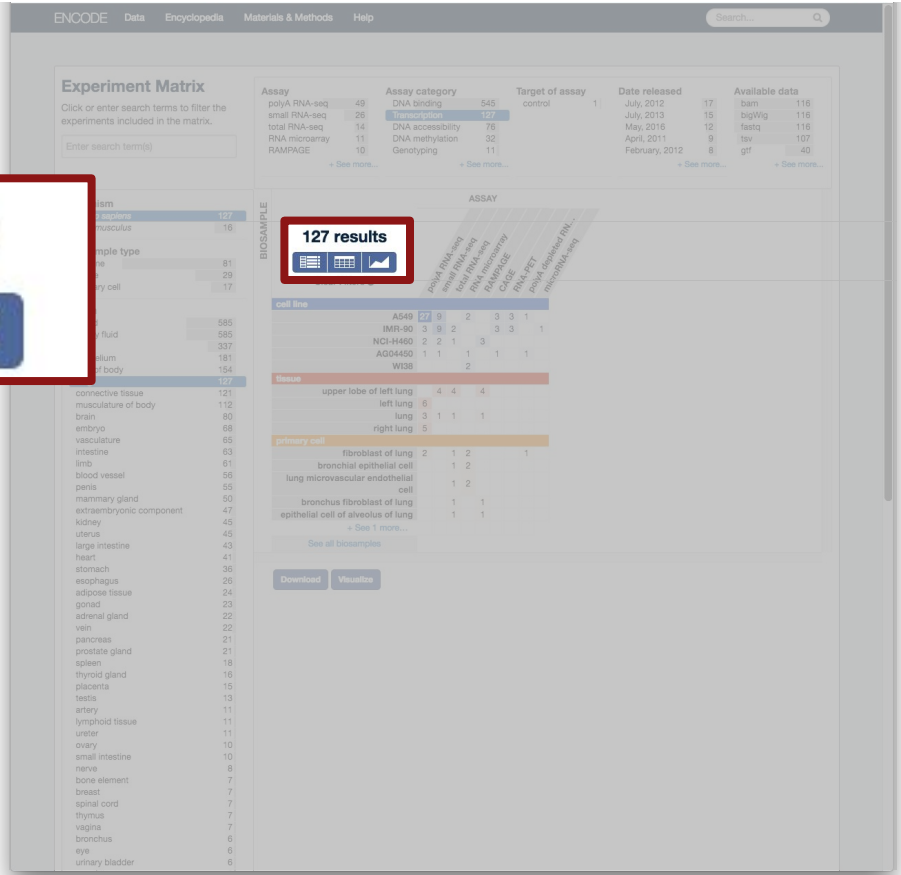
Sub-tissues & cell types & cell lines derived from an organ





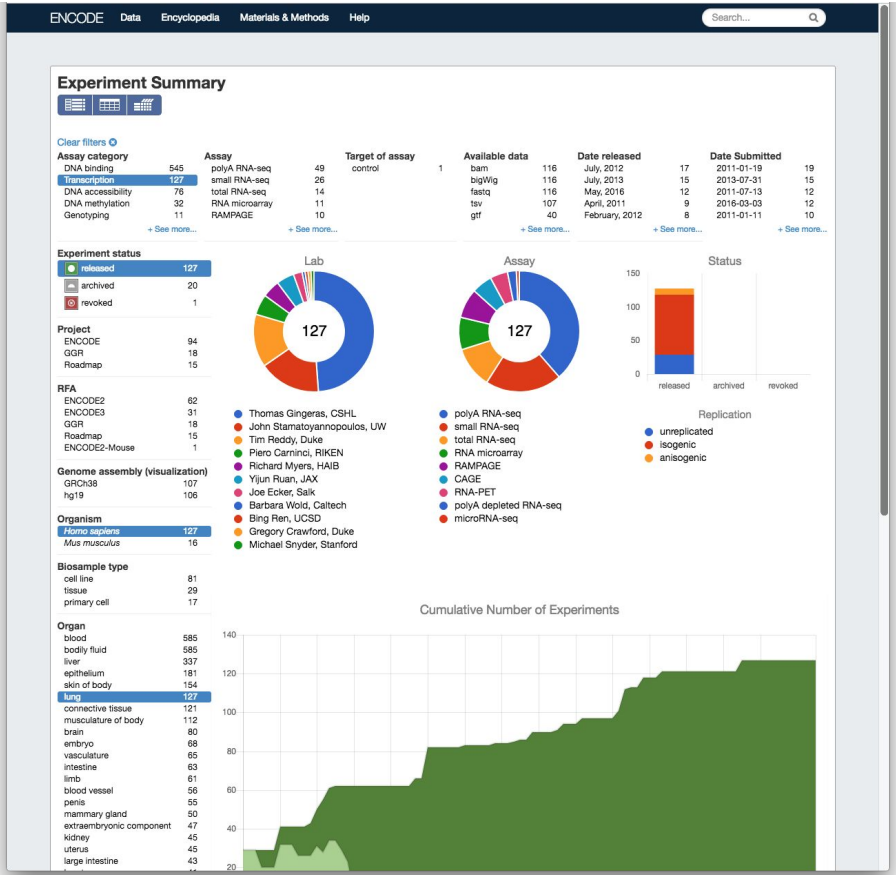
Browsing the ENCODE portal

Different views



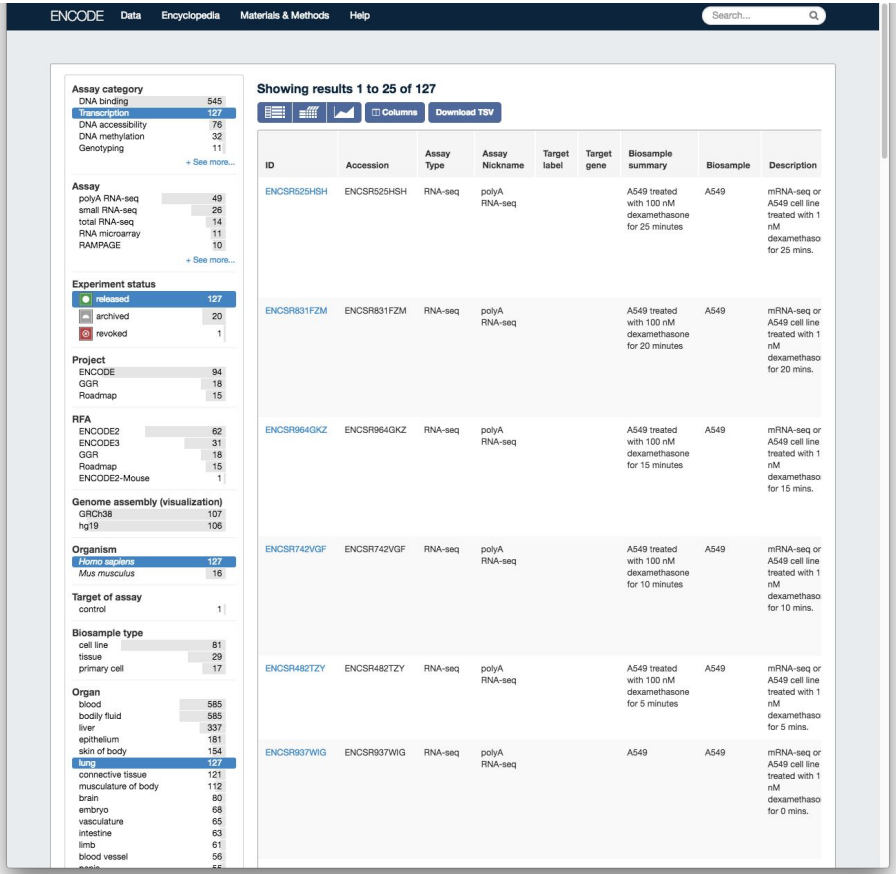
Browsing the ENCODE portal

Summary view



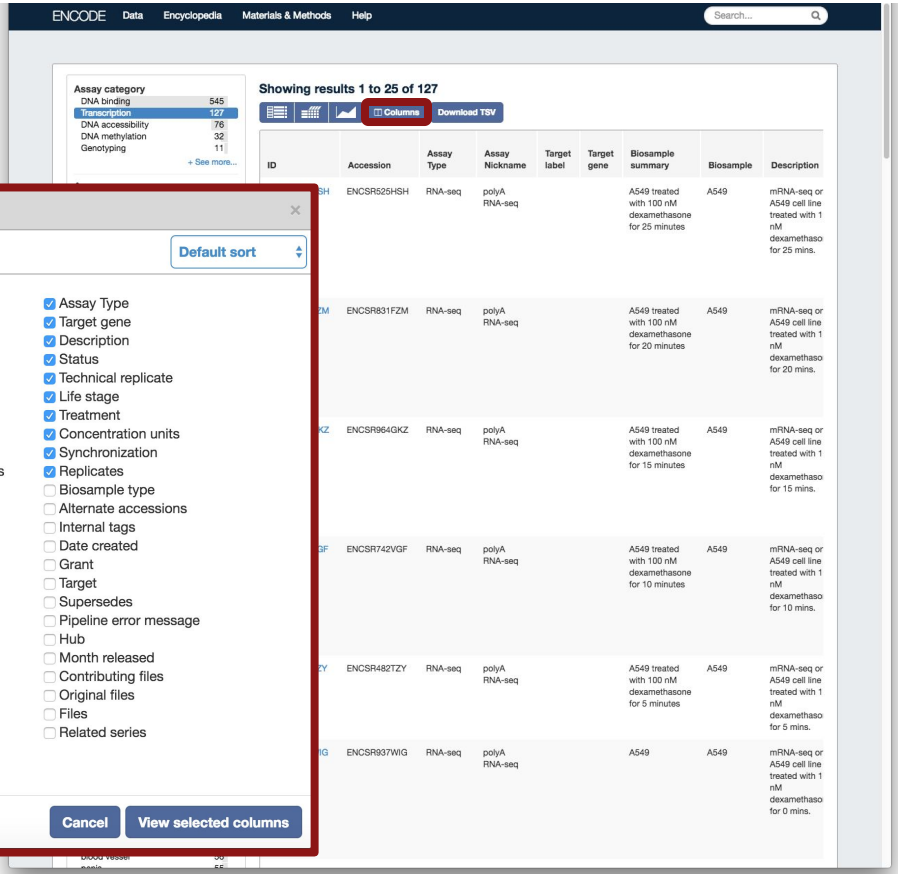
Browsing the ENCODE portal

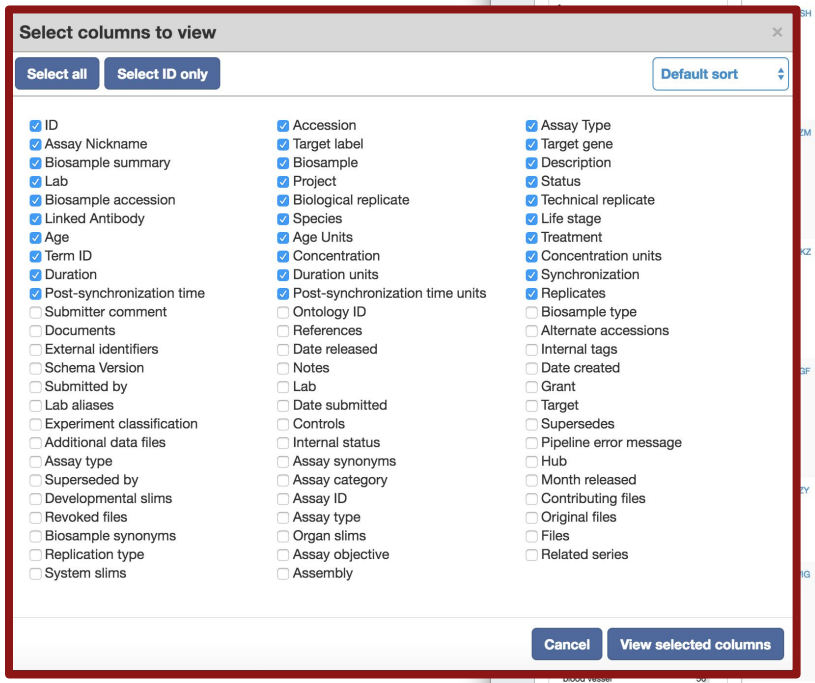
Report view



Browsing the ENCODE portal

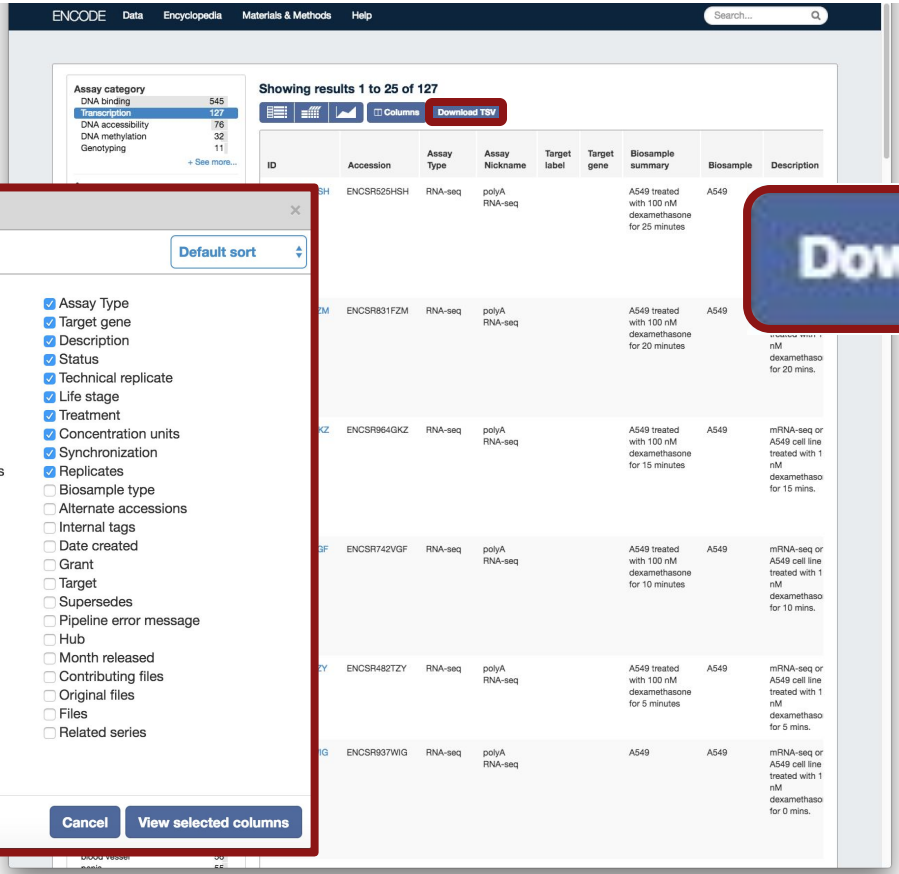
Report view

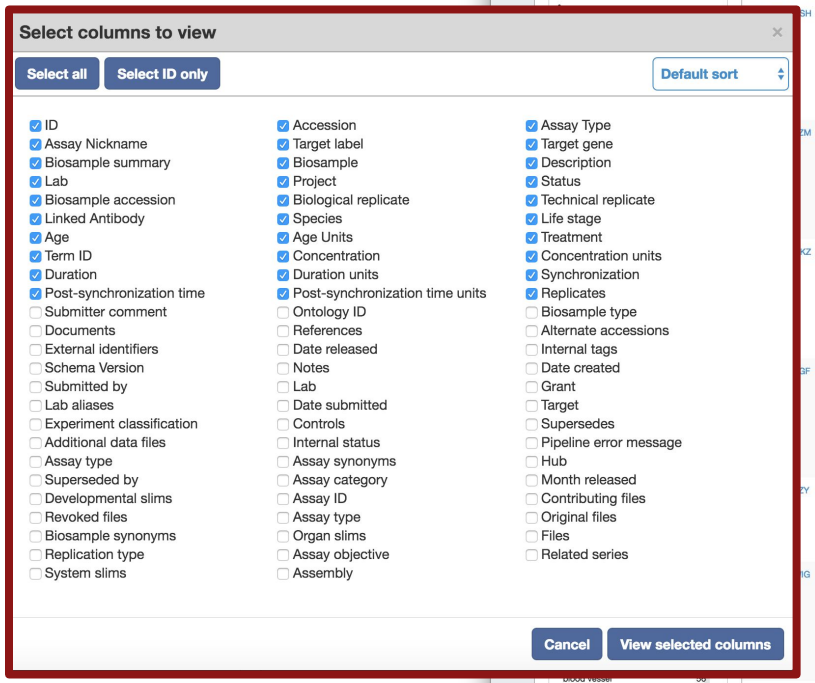


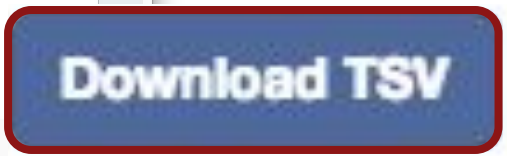


Browsing the ENCODE portal

Report view

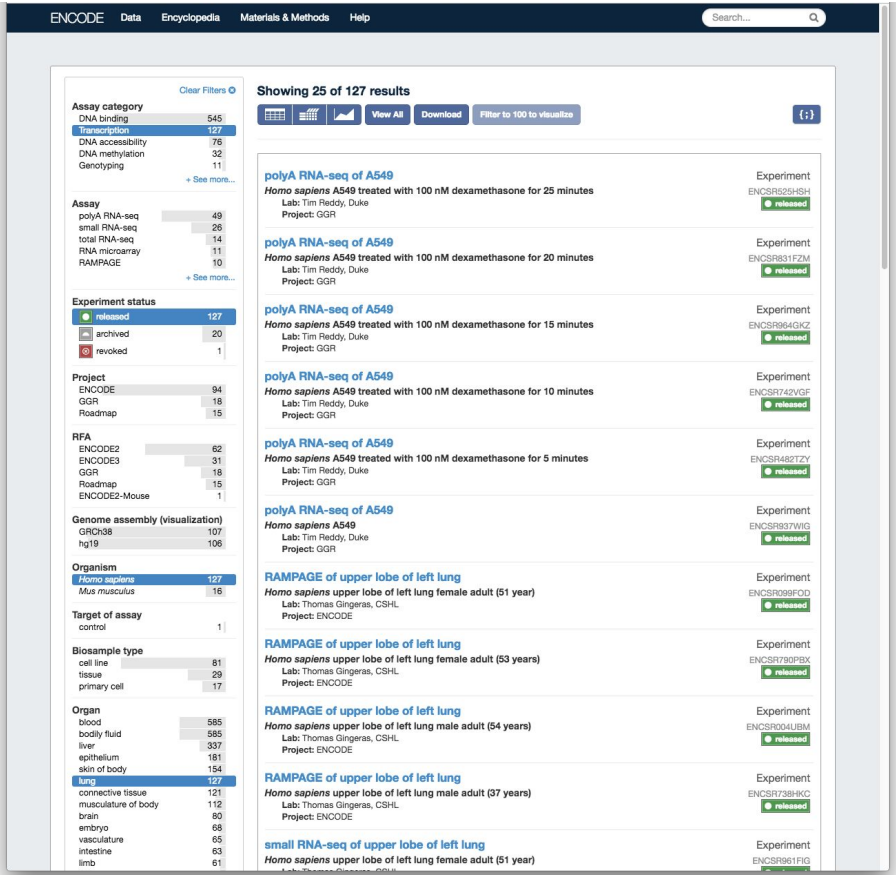






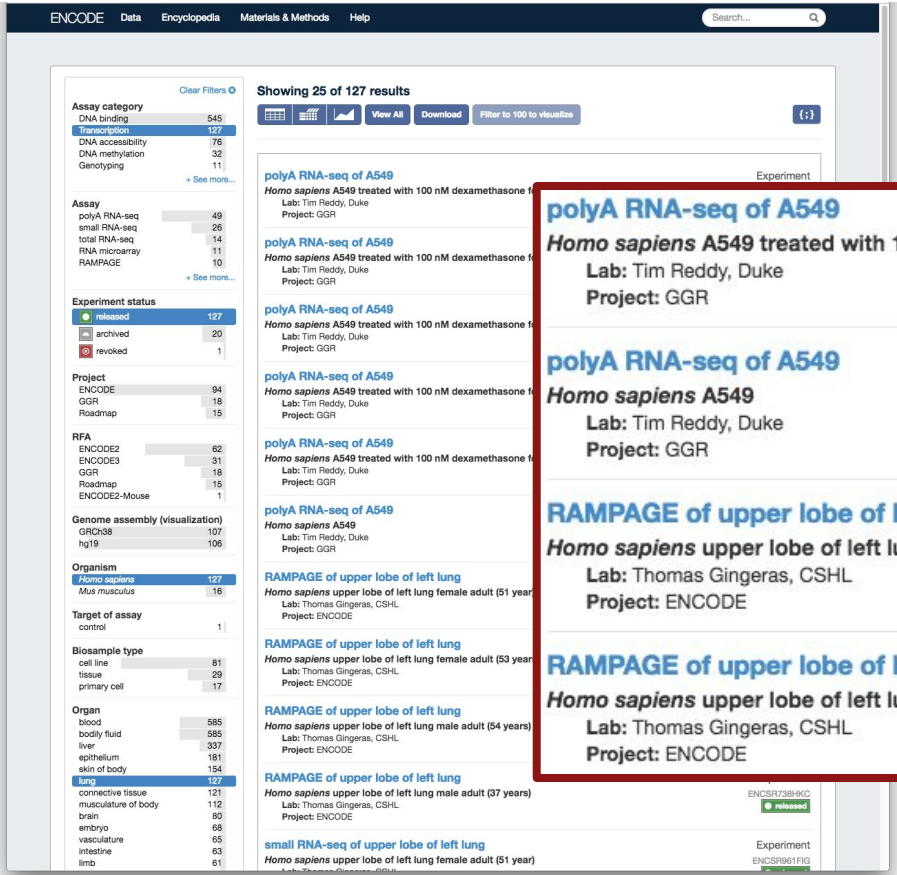
Browsing the ENCODE portal

Search view



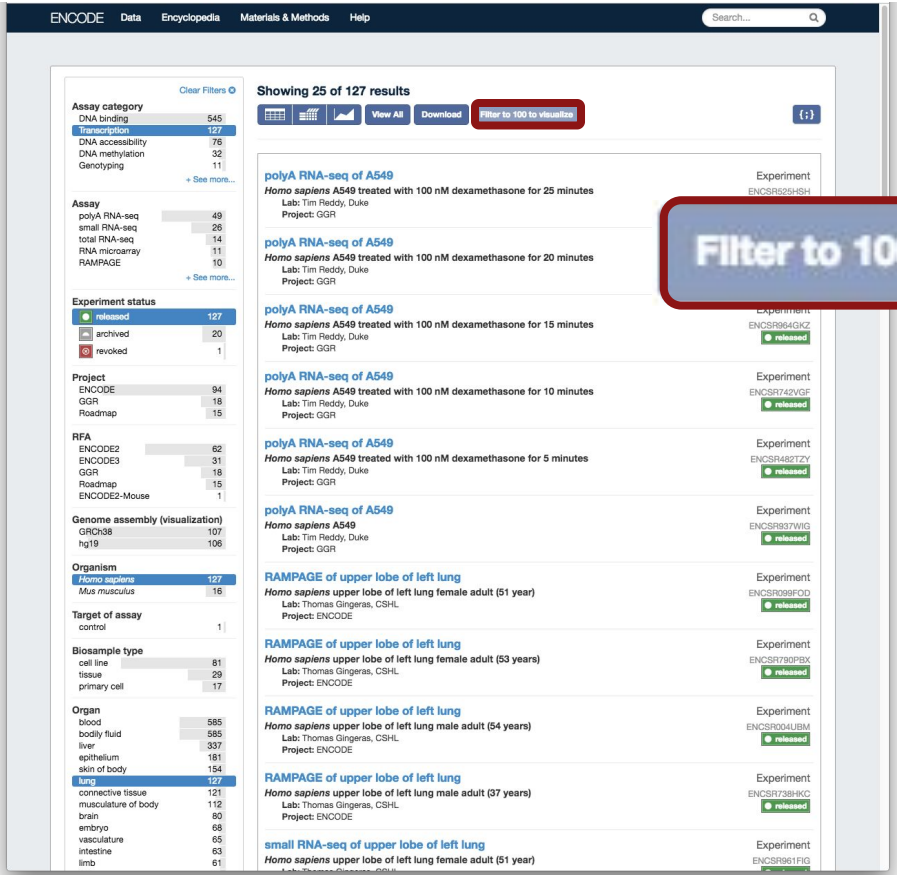
Browsing the ENCODE portal

Search view



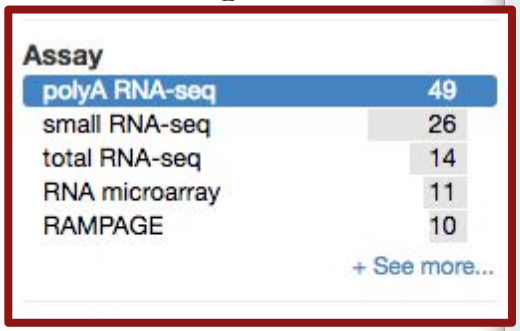
Visualizing data

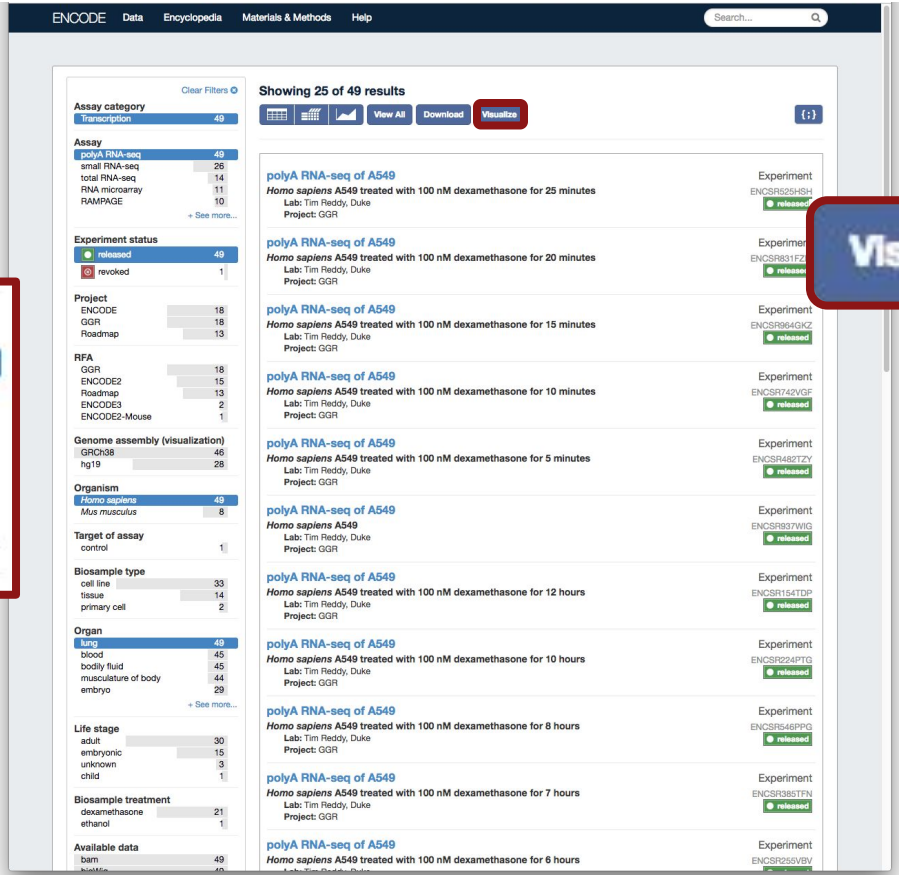
Search view



Visualizing data

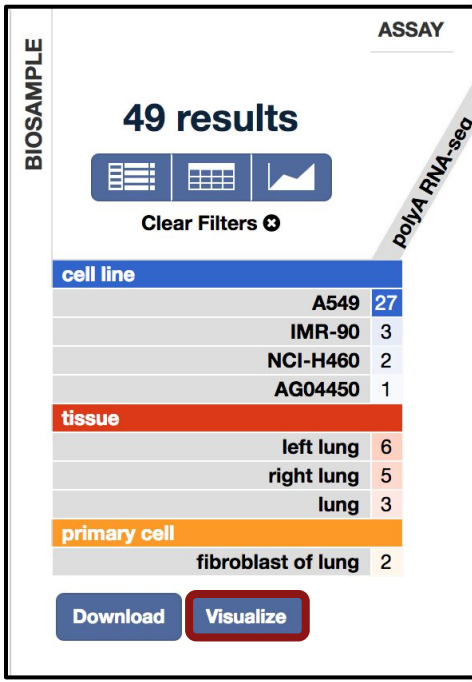
Filter further to ≤100 experiments

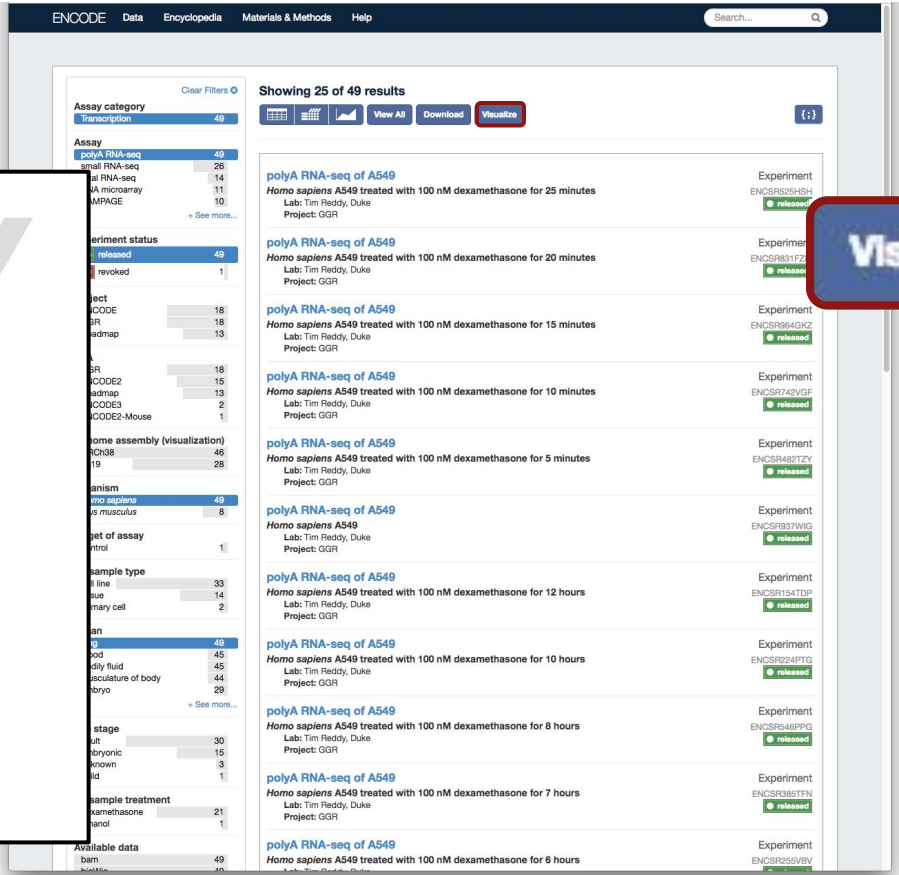




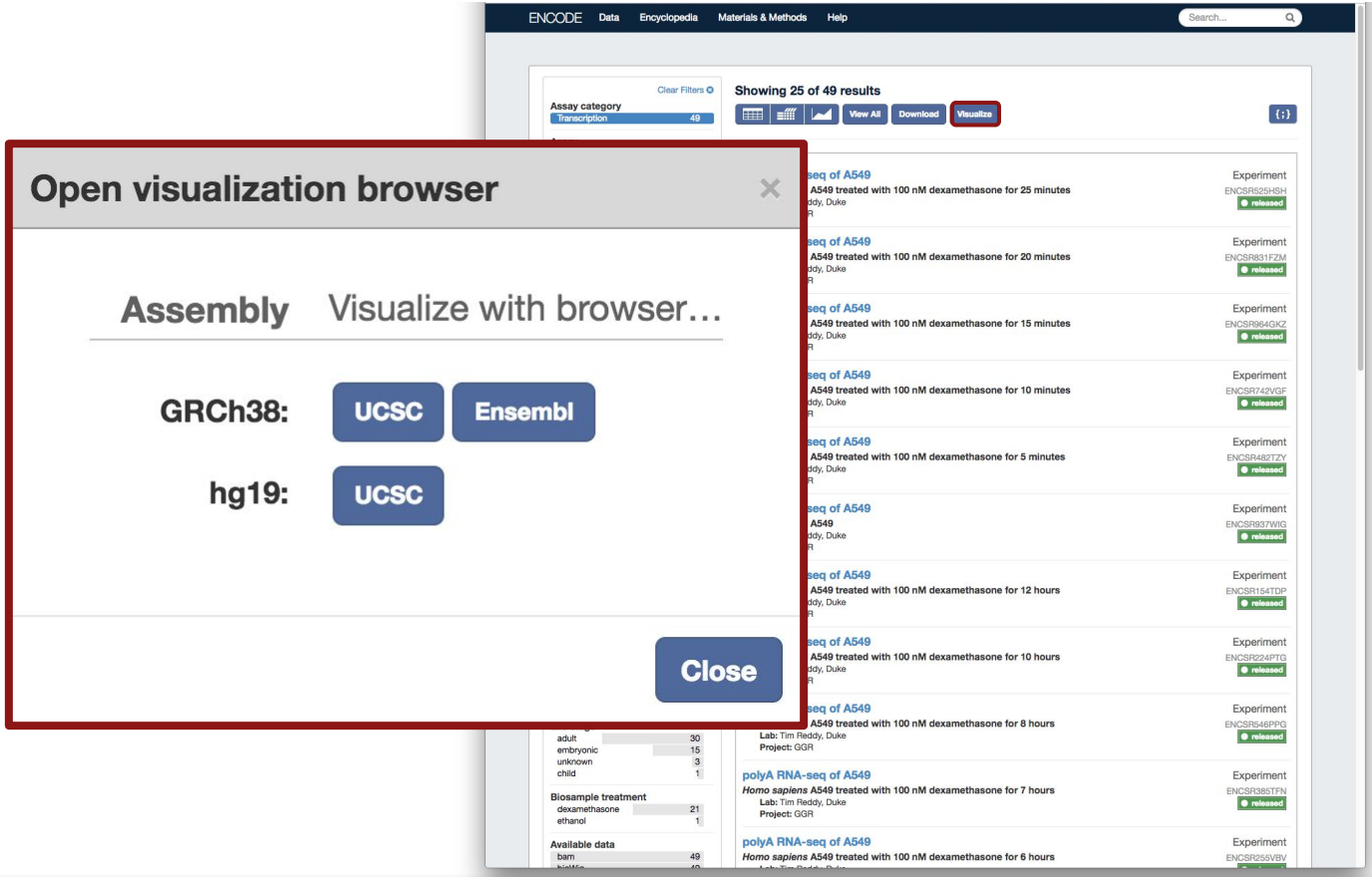
Visualizing data

Also available on matrix view

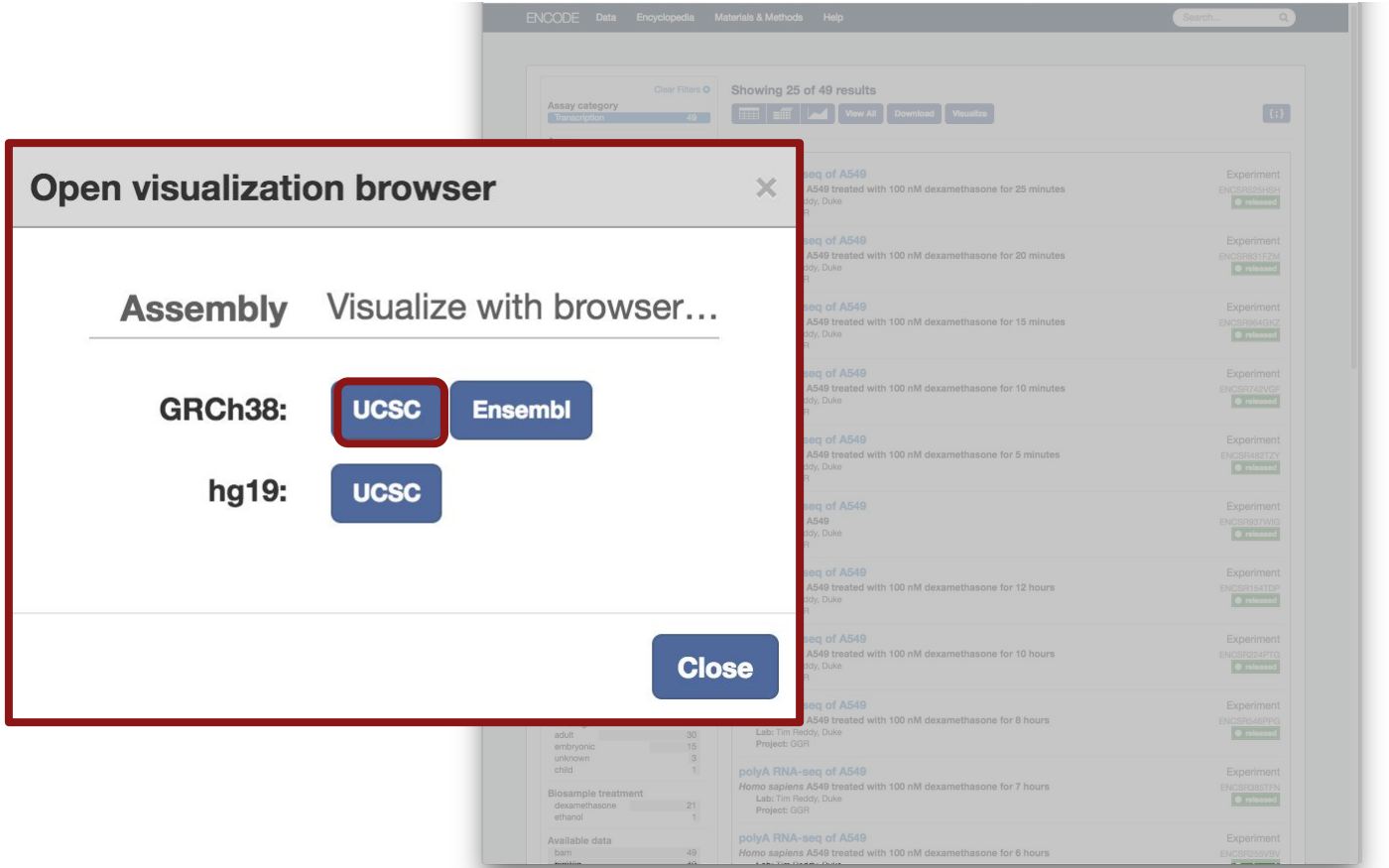


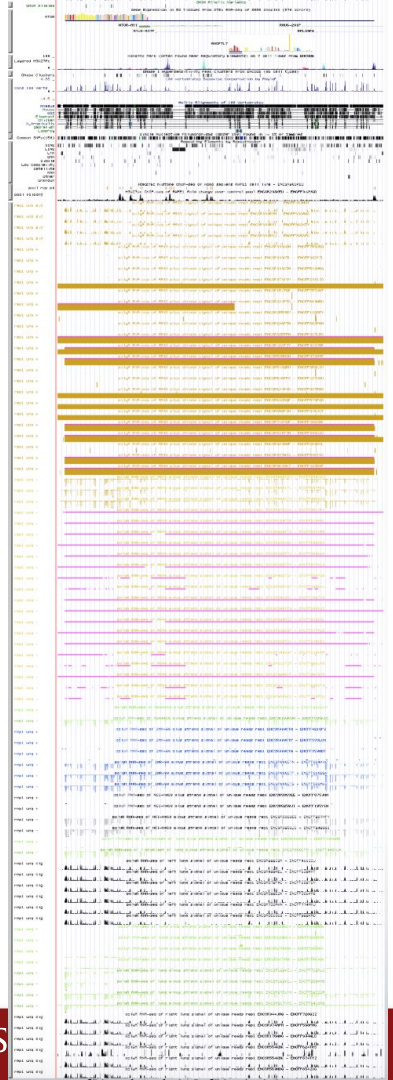


Visualizing data



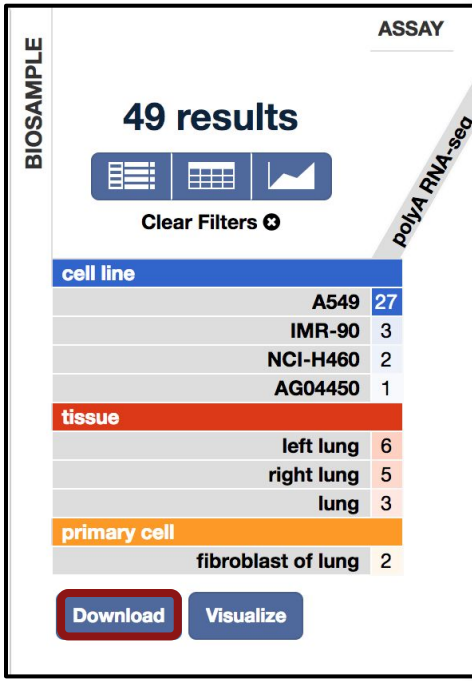
Visualizing data

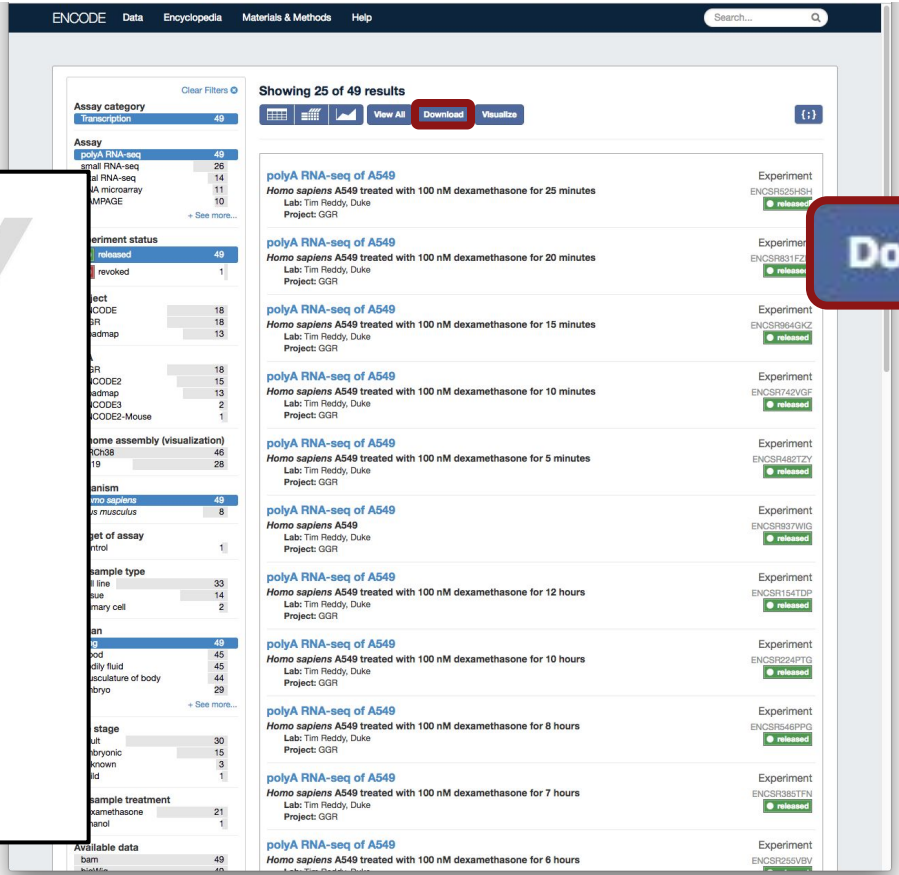




Batch file download

Also available on matrix view

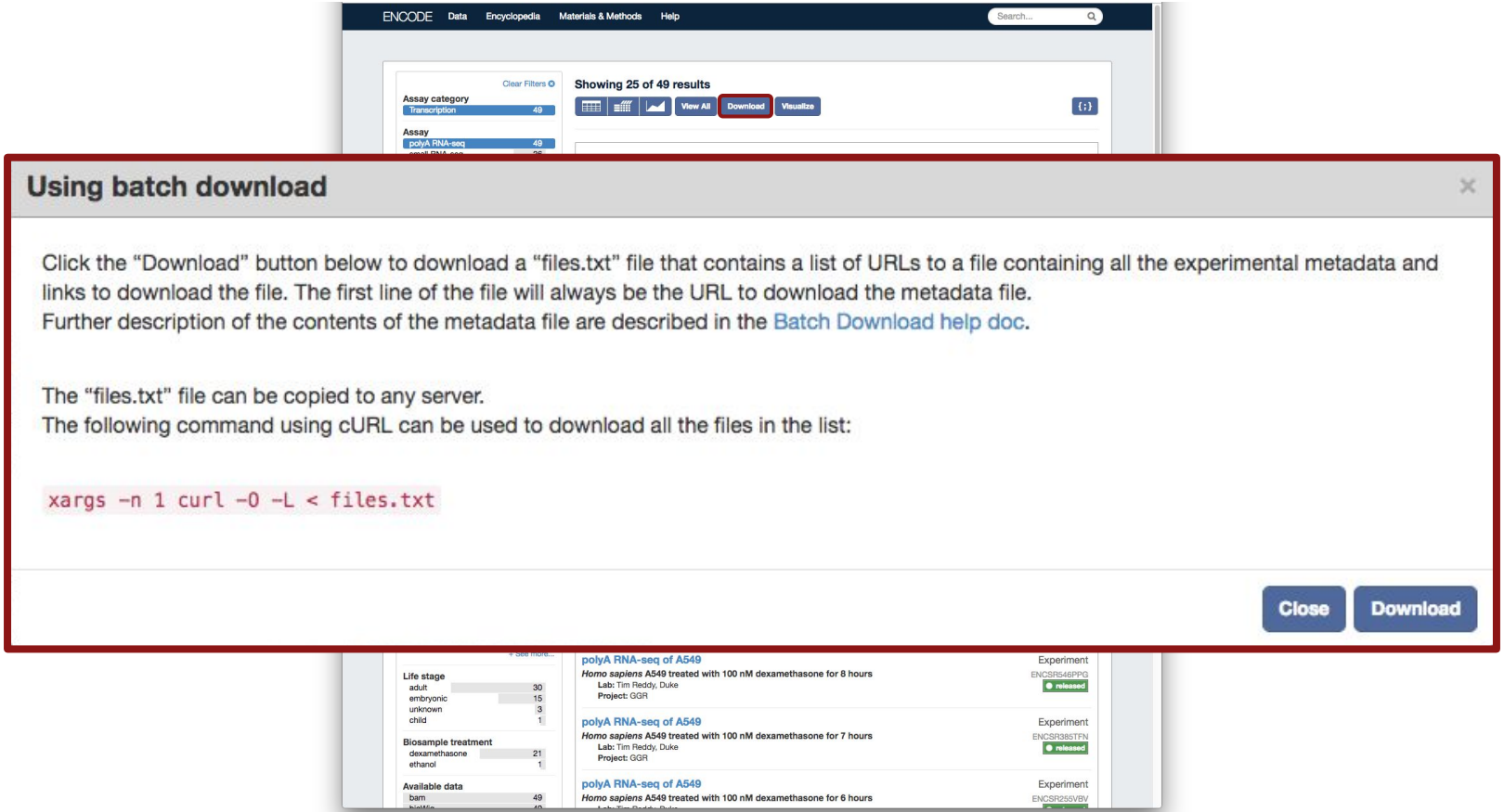




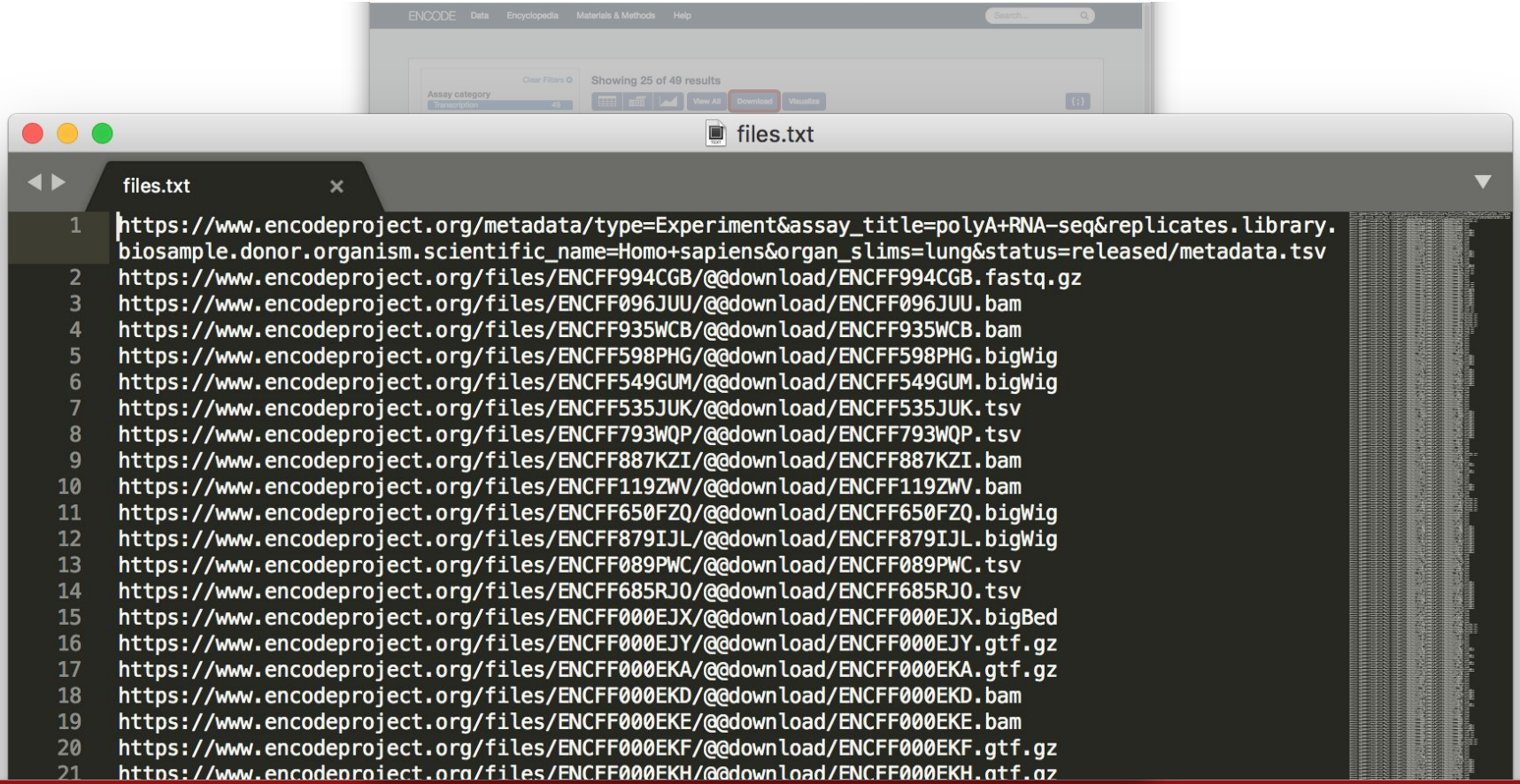
Download

No exp limit!

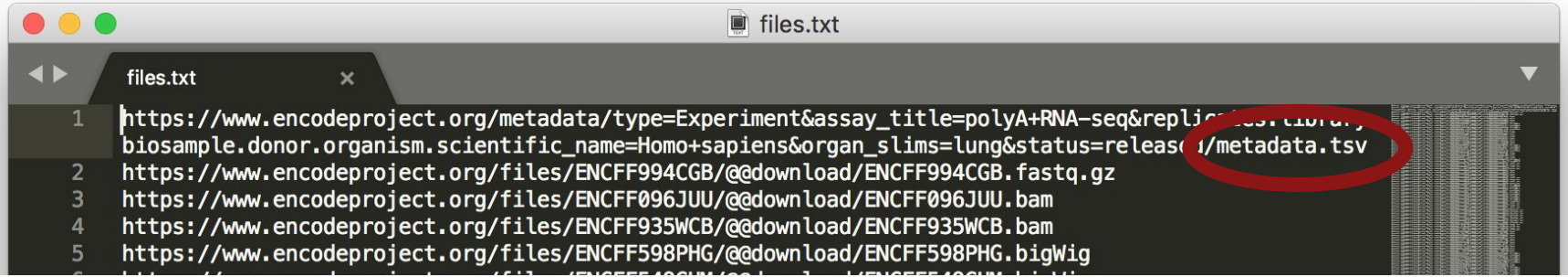
Batch file download



Batch file download



Batch file download



metadata.tsv

one line for each data file in files.txt

file format, output type, read info, assembly

also assay, biosample, library specifics

Batch file download

To download all files, run command...

```
xargs -n 1 curl -O -L < files.txt
```

Batch file download

To download all files, run command...

```
xargs -n 1 curl -O -L < files.txt
```

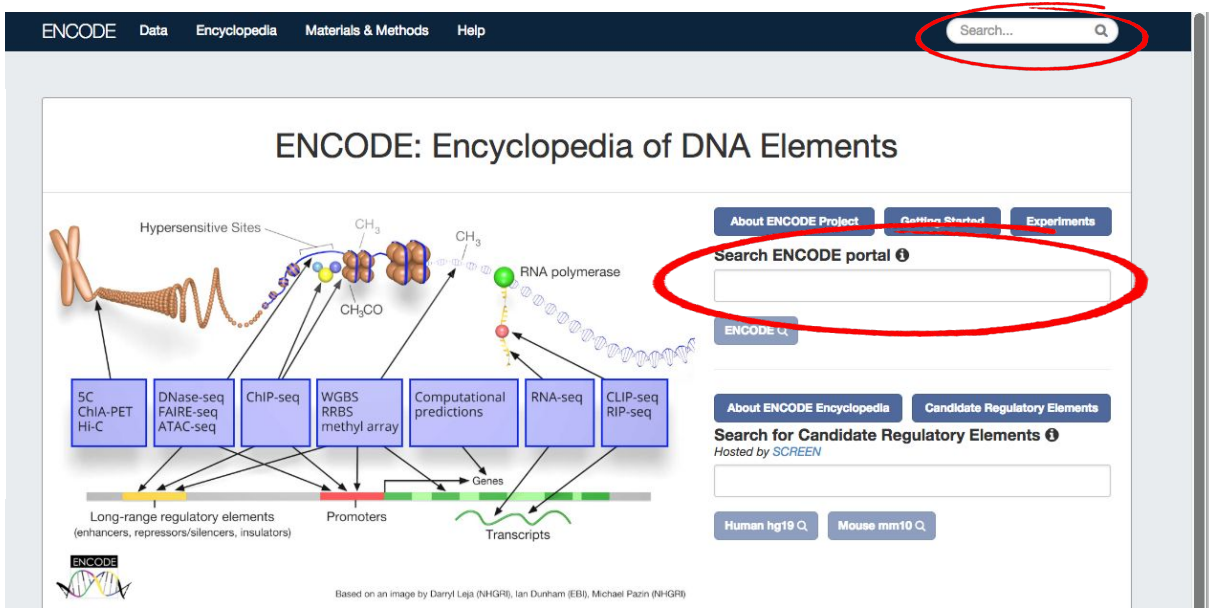
Or copy/paste metadata.tsv link into browser
filter metadata.tsv,
collect download links (column AK in metadata.tsv),
put in new file to run command

```
xargs -n 1 curl -O -L < filtered_files.txt
```

Searching the ENCODE portal

Keyword Search

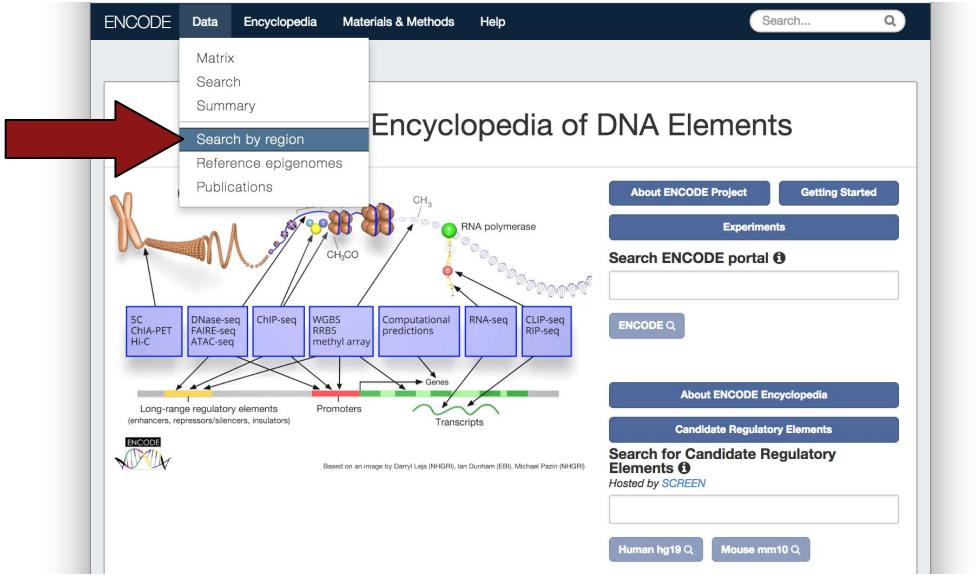
"liver"
"CTCF"
"gtex"
...



Searches the whole portal
(not just Experiments)

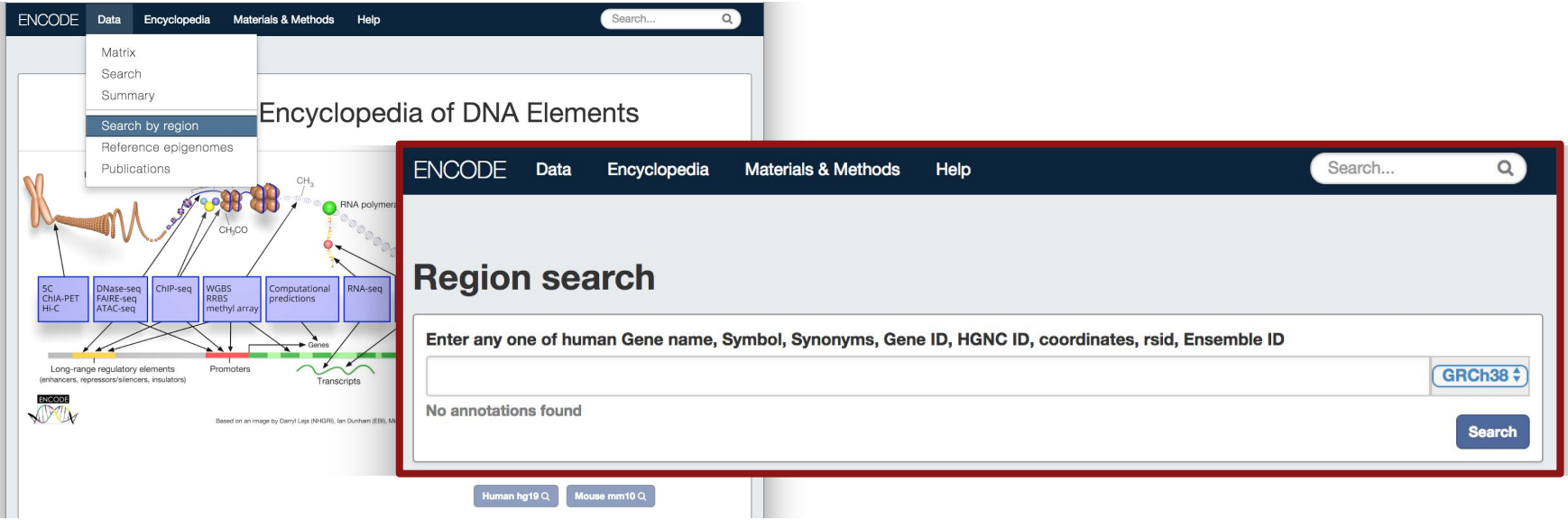
Searching the ENCODE portal

Region search



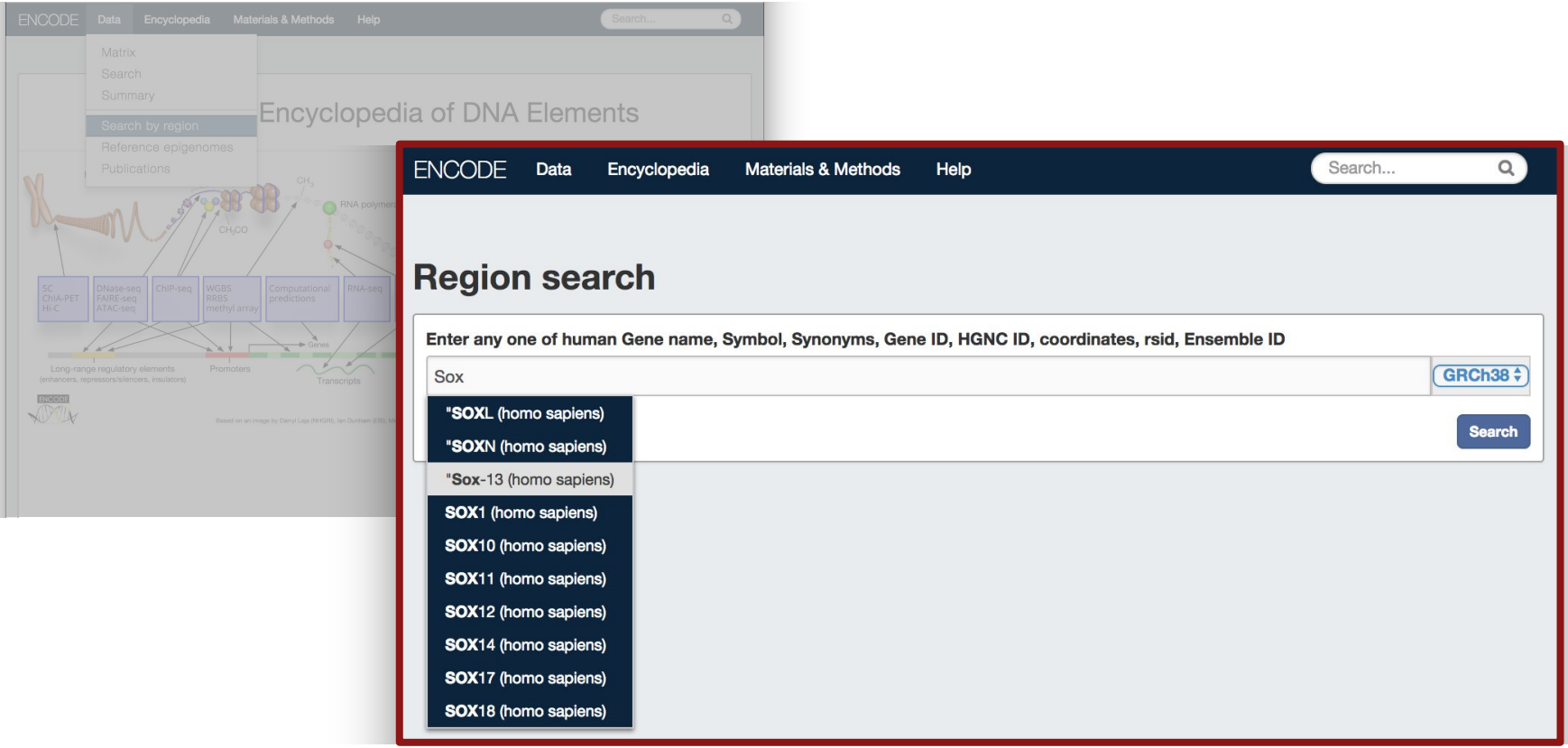
Browsing the ENCODE portal

Region search



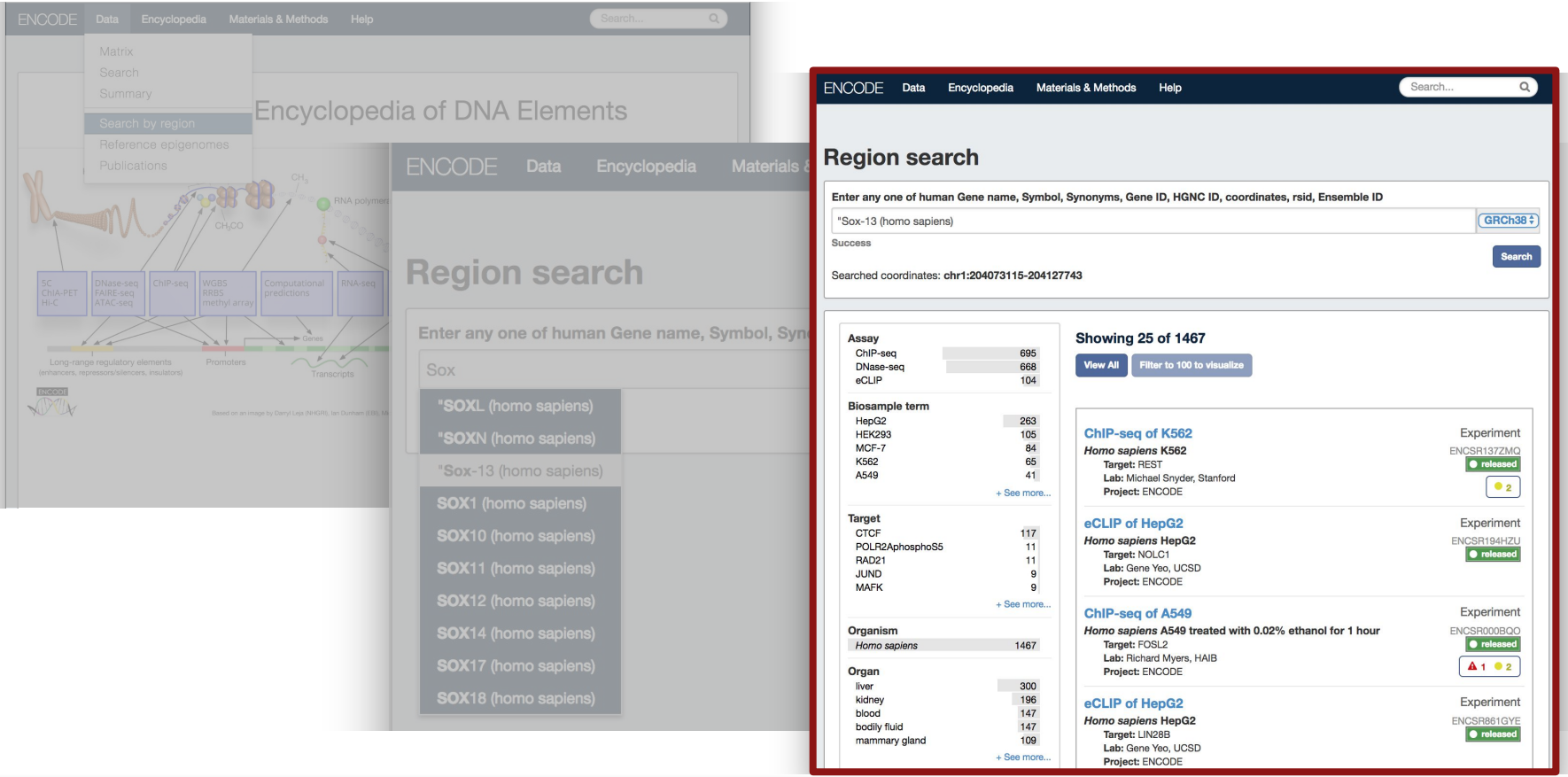
Browsing the ENCODE portal

Region search



Browsing the ENCODE portal

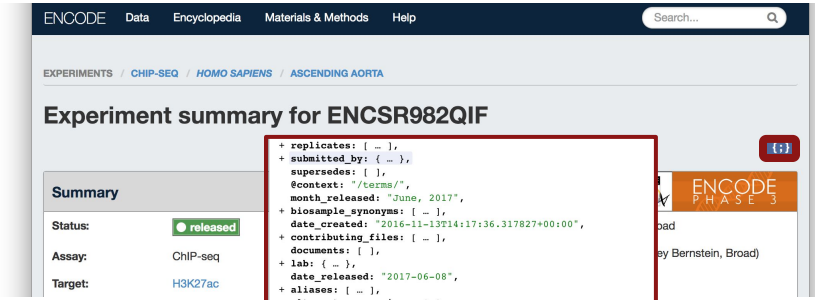
Region search

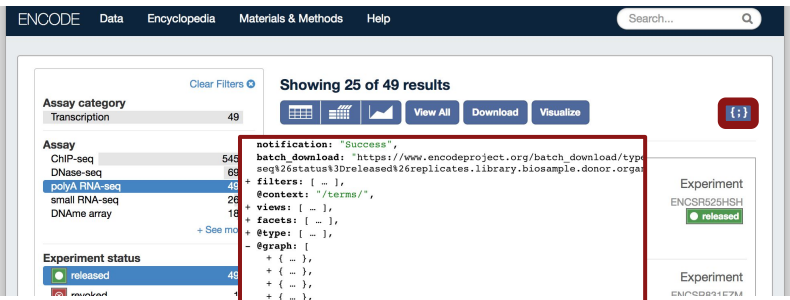


Programmatic access

JSON objects

All objects on the portal are json format, click the icon to view json (w/ a json plug-in)



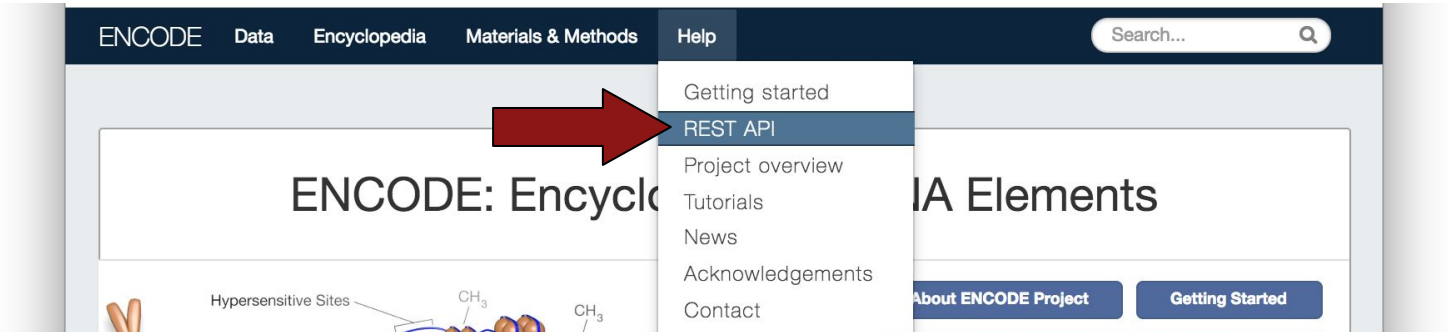



```
+ replicates: [ … ],
+ submitted_by: { … },
  supersedes: [ ],
  @context: "/terms/",
  month_released: "June, 2017",
+ biosample_synonyms: [ … ],
  date_created: "2016-11-13T14:17:36.317827+00:00",
+ contributing_files: [ … ],
  documents: [ ],
+ lab: { … },
  date_released: "2017-06-08",
+ aliases: [ … ],
  alternate_accessions: [ ],
  biosample_summary: "ascending aorta female adult (51 year)",
+ visualize: { … },
  biosample_type: "tissue",
  accession: "ENCSR982QIF",
+ type_slims: [ … ],
+ related_series: [ … ],
+ files: [ … ],
  related_files: [ ],
+ assay_slims: [ … ],
  schema_version: "21",
  audit: { },
  hub: "/experiments/ENCSR982QIF/@@hub/hub.txt",
  assay_term_id: "OBI:0000716",
  description: "",
+ possible_controls: [ … ],
+ target: { … },
+ @type: [ … ],
  revoked_files: [ ],
+ system_slims: [ … ],
+ experiment_classification: [ … ],
+ original_files: [ … ],
  uuid: "d0032903-956c-4ff8-8c8a-38cbbdaeb598",
  internal_status: "pipeline completed",
  biosample_term_name: "ascending aorta",
  date_submitted: "2016-11-13",
  @id: "/experiments/ENCSR982QIF/",
+ dbxrefs: [ … ]
}
```

```
  notification: "Success",
  batch_download: "https://www.encodeproject.org/batch_download/type
  seq%26status%3Dreleased%26replicates.library.biosample.donor.orga
+ filters: [ … ],
  @context: "/terms/",
+ views: [ … ],
+ facets: [ … ],
+ @type: [ … ],
- @graph: [
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … },
    + { … }
  ],
  all: "/?type=Experiment&assay_title=polyA+RNA-
  seq&status=released&replicates.library.biosample.donor.organism.s
+ sort: { … },
  total: 49,
+ visualize_batch: { … },
+ columns: { … },
  @id: "/search/?type=Experiment&assay_title=polyA+RNA-
```

Programmatic access

Tutorials



ENCODE DCC

encode-help@lists.stanford.edu
@EncodeDCC
https://github.com/ENCODE-DCC

ASHG posters

414 RegulomeDB
1510 ENCODE portal
1598 ENCODE pipelines
1987 Roadmap/Reference Epigenomes

Mike

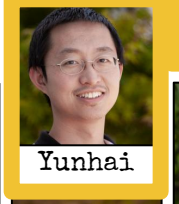
Idan

Ben

Seth

Jin

Jason


Yunhai

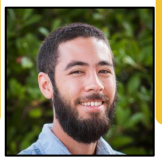
Weiwei

Casey

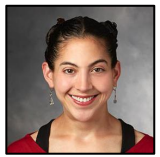
Forrest

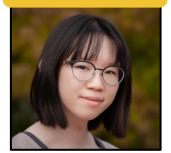
Bek

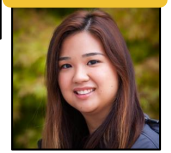
Paul


Zack

Otto


Vanessa


Jennifer


Khine

Emma

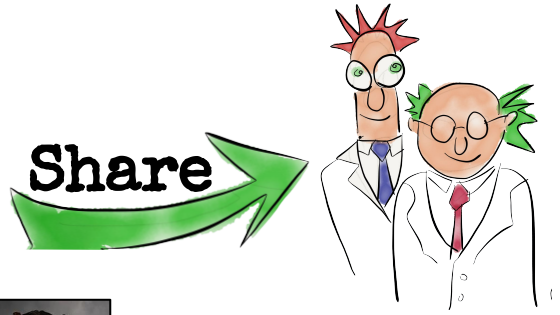
Phil

Keenan

Data

Pipelines

Software



NIH U24 HG009397