Open Data Resources for Human Genomics Research

A/Prof Jason Wong
Head, Computational Cancer Genomics
School of Biomedical Sciences
University of Hong Kong

Dr Rebecca Poulos
NHMRC Early Career Fellow
Children's Medical Research Institute
University of Sydney

Part One — Human genomics data resources — 1:30pm – 3pm

- UCSC Genome Browser + Exercise
- gnomAd + Exercise
- GTEx + Exercise
- ENCODE database + Exercise
- Exercise: Putting it all together

Part Two — Cancer genomics data resources — 3:30pm – 5pm

- Brief introduction to TCGA
- cBioPortal
- Genomic Data Commons
- Xena Browser
- Cancer Genomics Cloud

Part One

Human genomics data resources

1:30pm – 3pm

- UCSC Genome Browser

  Visualise and download genomic datasets.

- gnomAd

  Find data on single nucleotide polymorphisms (SNPs)

- GTEx

  Get data for tissue-specific gene expression and regulation

- ENCODE database

  Identify functional elements in the human genome

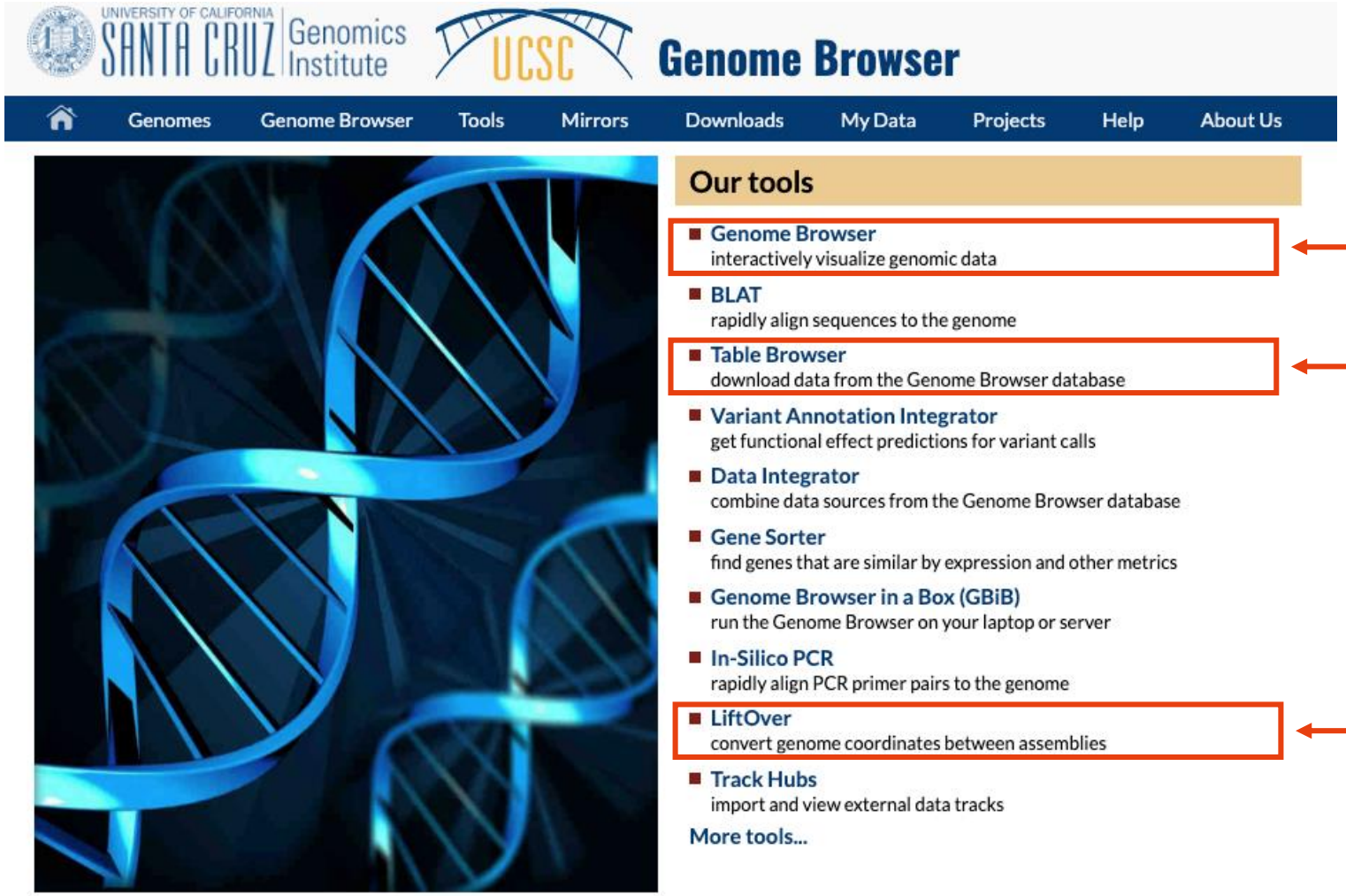
Each section will be accompanied by an exercise to help you test your skills, plus an exercise at the end to put all your knowledge together.

UCSC Genome Browser

Visualise and download genomic datasets.

UCSC Genome Browser

https://genome.ucsc.edu/



Visualise genomic data

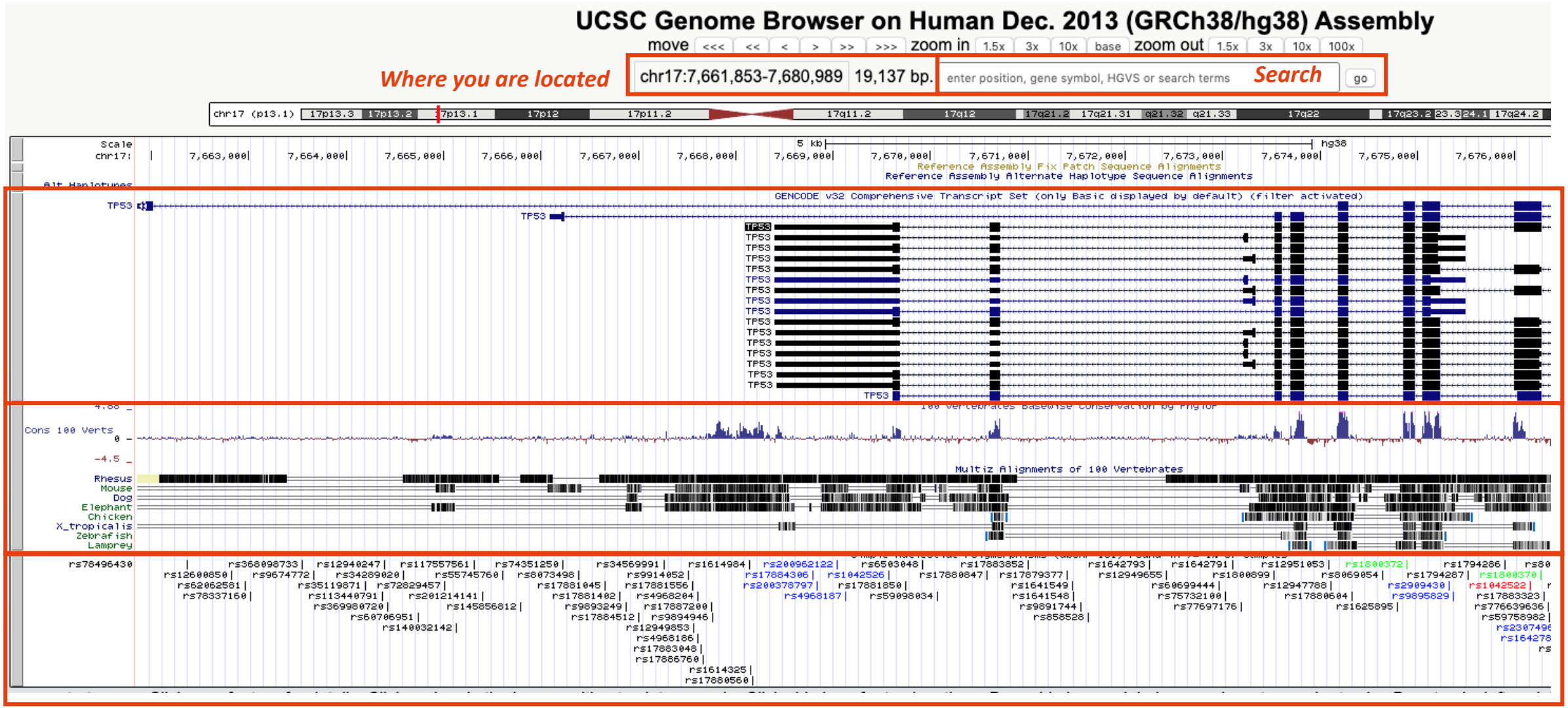
Download genomic data

Convert genomic data from one reference genome to another

UCSC Genome Browser

link

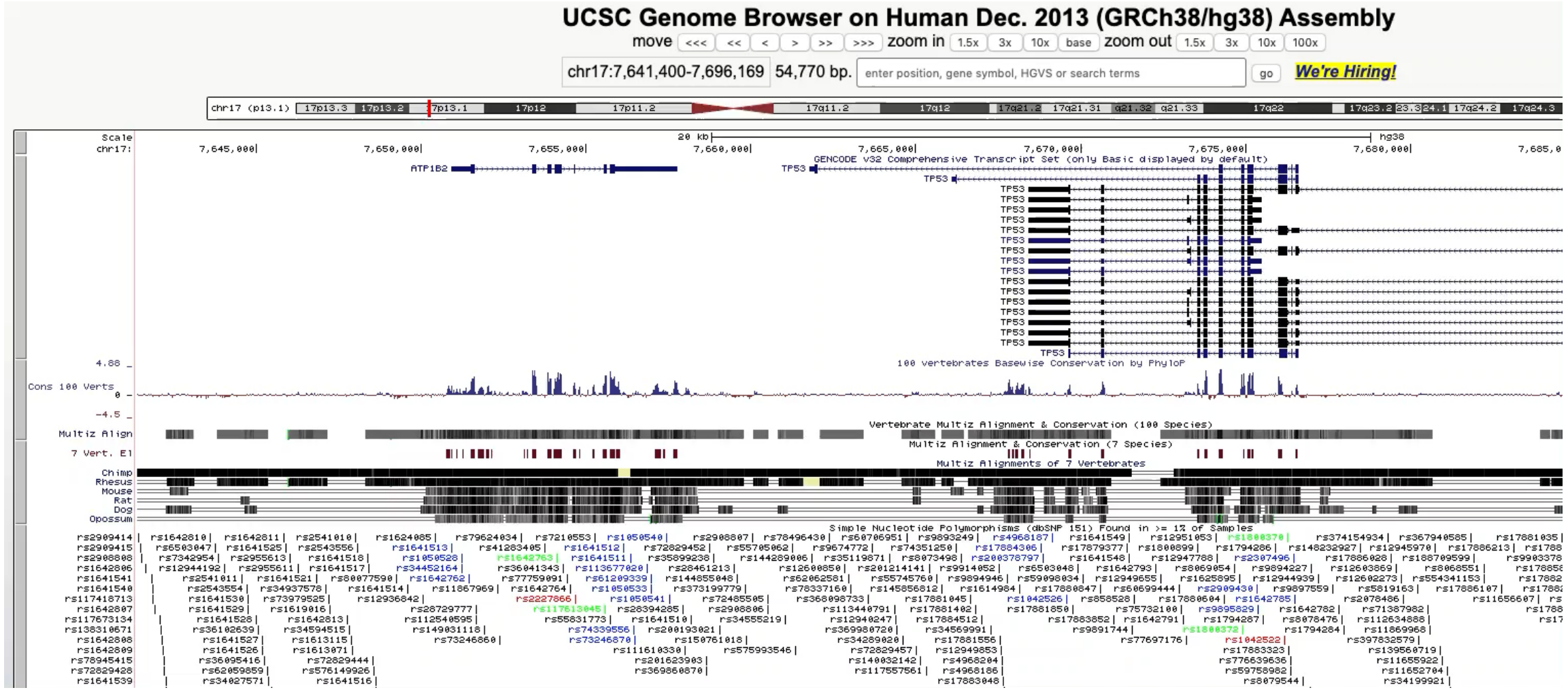
Genome Browser



UCSC Genome Browser

link

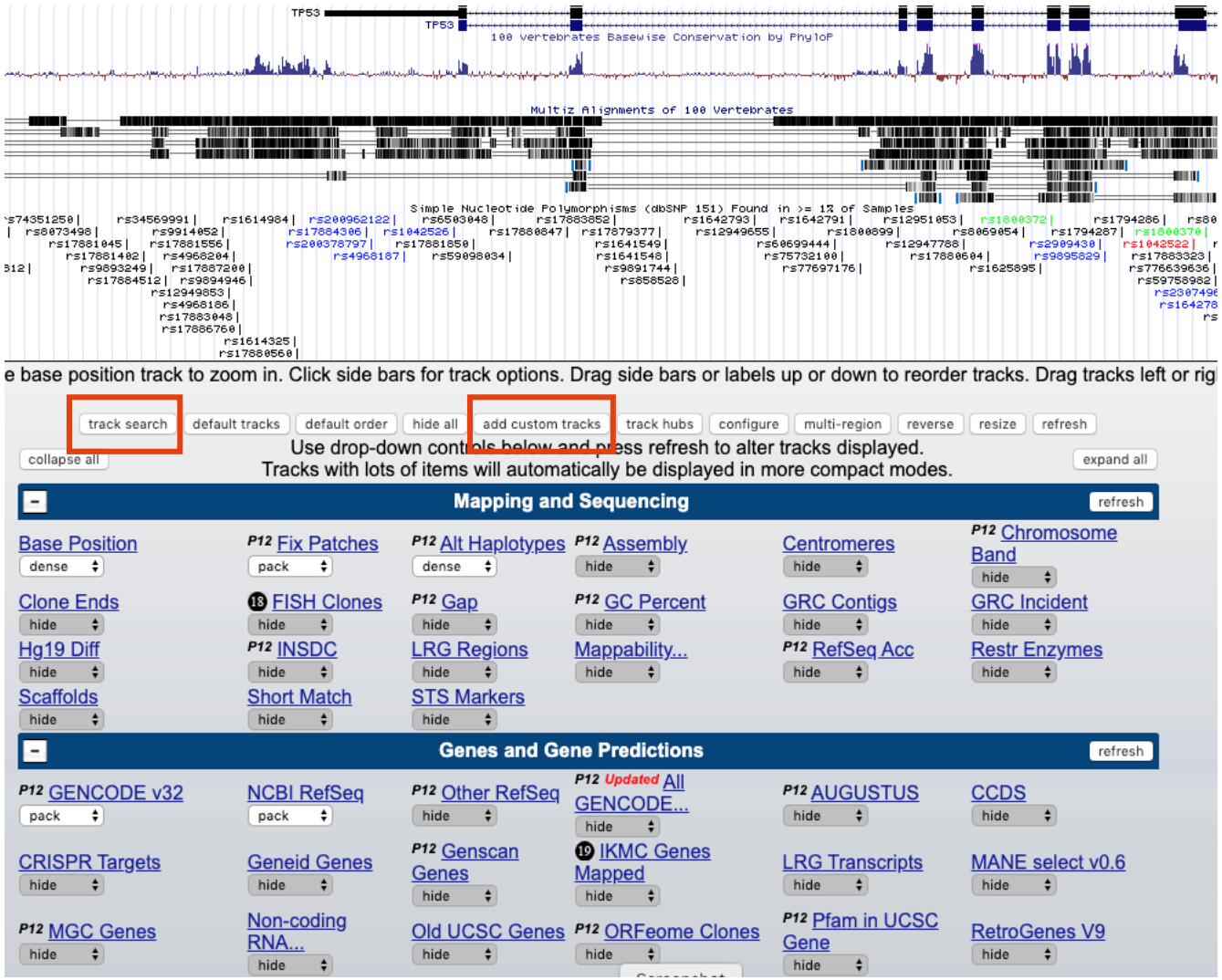
Genome Browser



UCSC Genome Browser

link

Genome Browser



Add other tracks

Select common tracks to load

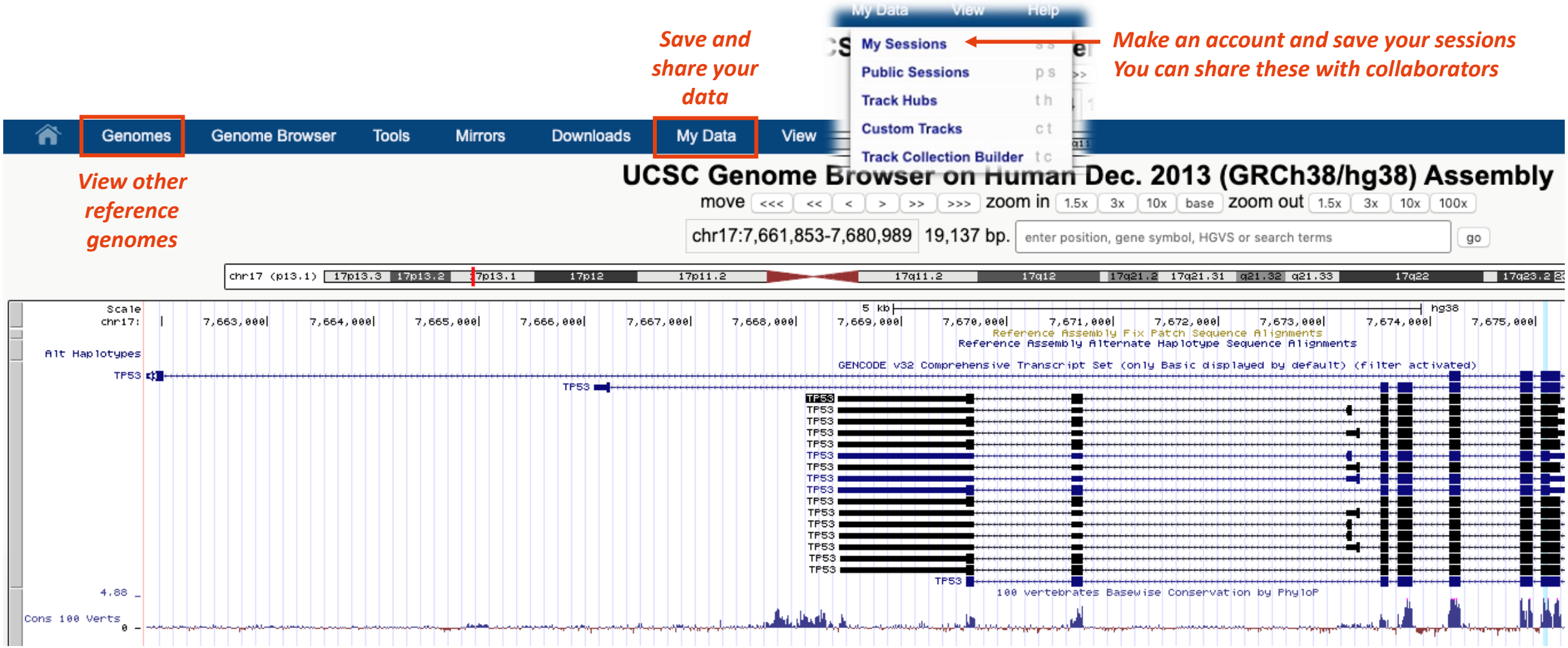
Add custom tracks from your own data

UCSC Genome Browser

Save and share your data

Make an account and save your sessions
You can share these with collaborators

View other reference genomes

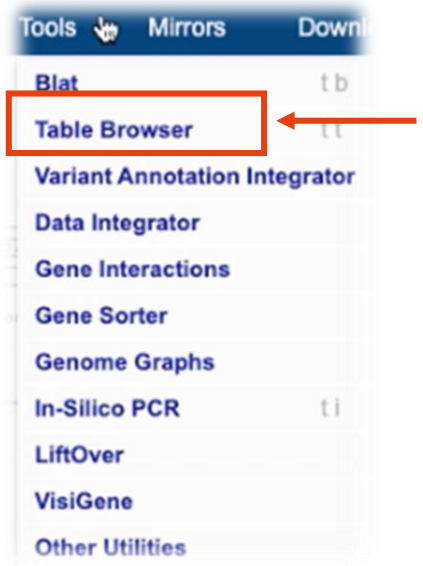


UCSC Genome Browser

link

Table Browser

Download genomic datasets



UCSC Genome Browser

link

LiftOver

Convert data from one reference genome to another

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Projects | Help | About Us

Lift Genome Annotations

This tool converts genome coordinates and genome annotation files between assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selected from the pull-down menus, a direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 2004, to Mouse, Feb. 2006, and then from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome: Human | Original Assembly: Dec. 2013 (GRCh38/hg38)

New Genome: Human | New Assembly: Feb. 2009 (GRCh37/hg19)

Minimum ratio of bases that must remap: 0.95

BED 4 to BED 6 Options

Allow multiple output regions:

Minimum hit size in query: 0

Minimum chain size in target: 0

BED 12 Options

Min ratio of alignment blocks or exons that must map: 1

If thickStart/thickEnd is not mapped, use the closest mapped base:

Paste in data (BED or chrN:start-end formats):

Submit

Clear

Or upload data from a file (BED or chrN:start-end in plain text format):

Choose file | No file chosen | Submit File

UCSC Genome Browser

Exercise…

What is the location and nucleotide of the most frequent mutation arising in the KRAS gene in the TCGA cancer cohort?

Chromosome 12 at 25,245,350. C > T

Challenge…

What amino acid is present at this location in mouse? *(HINT: look at Conservation)*

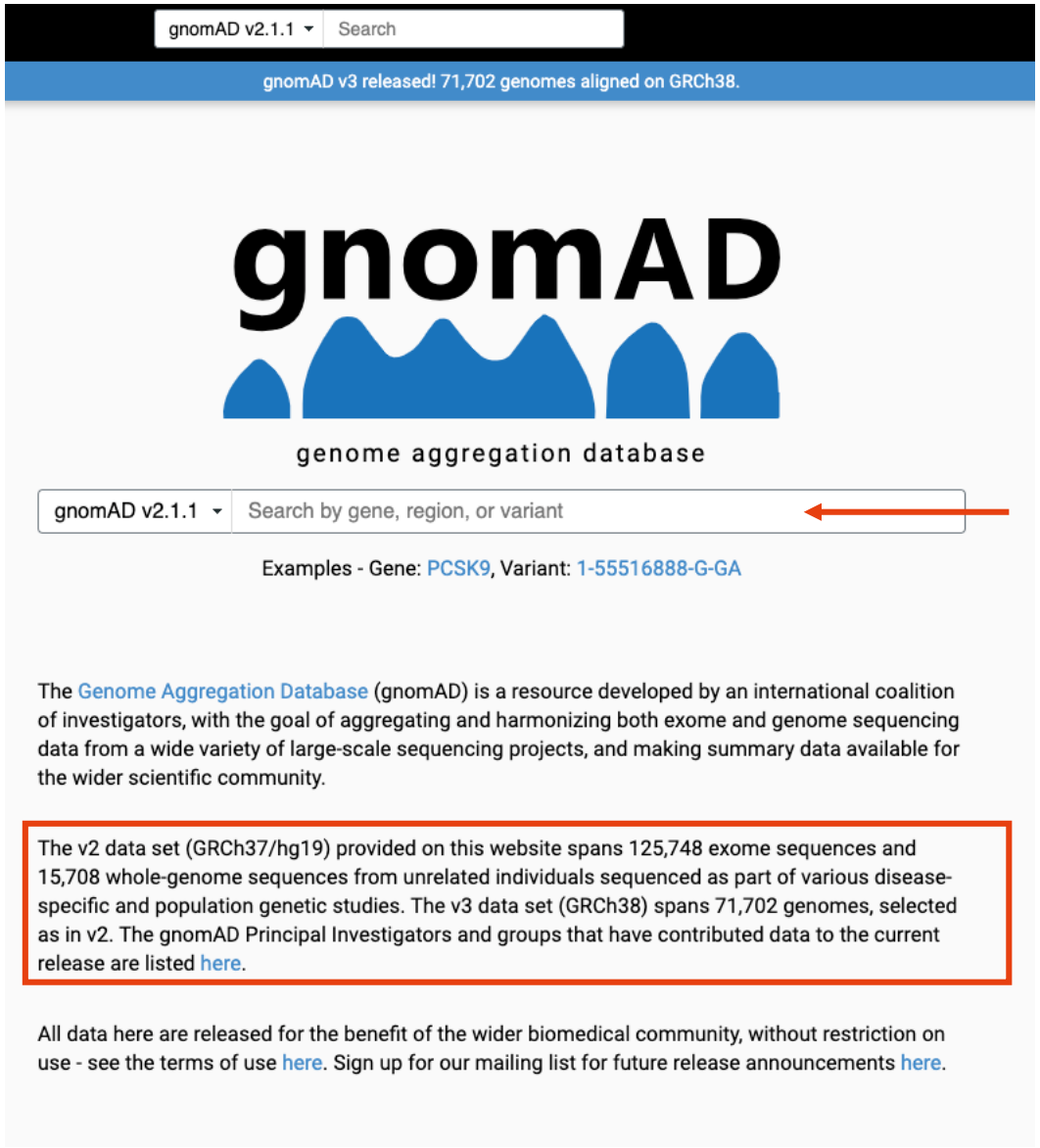
G / Glycine

Genome Aggregation Database (gnomAd)

Find data on single nucleotide polymorphisms (SNPs)

gnomAD

link

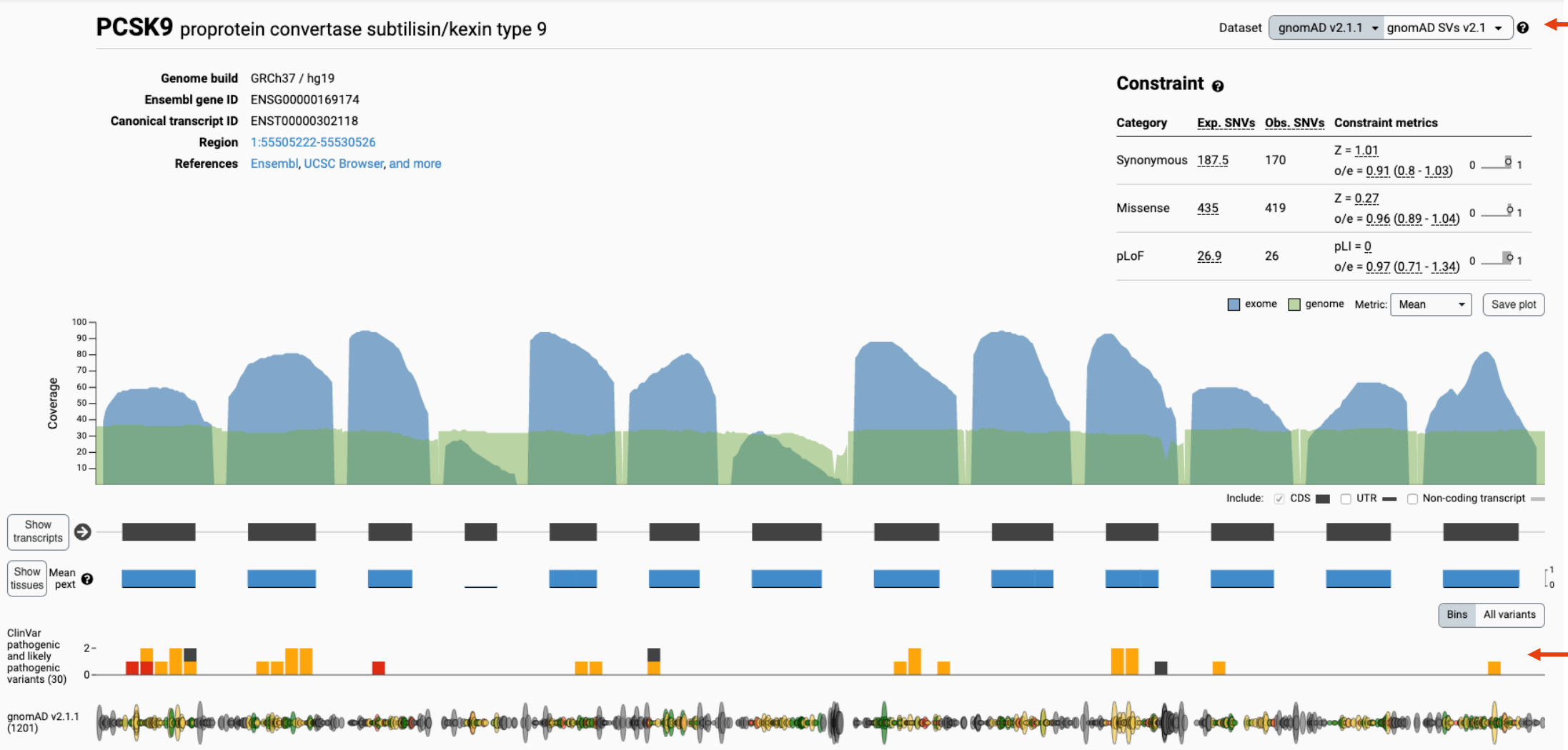


Search for a gene or variant

gnomAD

link

Swap to structural variants



Find pathogenic variants

gnomAD

link

Insertion: 1-55516888-G-GA (GRCh37)

Dataset: gnomAD v2.1.1

View allele frequencies

	Exomes	Genomes	Total
Filter	Pass	Pass	
Allele Count	121	156	277
Allele Number	128230	31382	159612
Allele Frequency	0.0009436	0.004971	0.001735
Popmax Filtering AF (95% confidence)	0.01338	0.01542	
Number of homozygotes	2	0	2

This variant is multiallelic. Other alt alleles are:

- 1-55516888-G-A

References

- dbSNP (rs527413419)
- UCSC

View in UCSC

Report

- Report this variant
- Request additional information

Annotations

This variant falls on 4 transcripts in 1 gene.

frameshift

- PCSK9
  - ENST00000452118
    HGVSp: p.Gly183GlufsTer23
    pLoF: High-confidence

intron

- PCSK9
  - ENST00000302118 *
  - ENST00000543384

non coding transcript exon

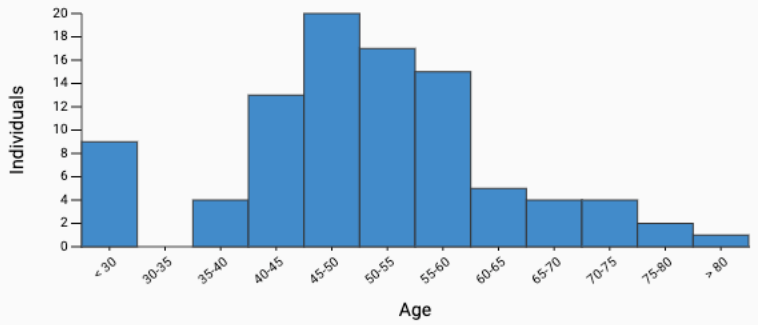
- PCSK9
  - ENST00000490692

See breakdown of population frequencies

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	251	14784	2	0.01698
Latino	18	25174	0	0.0007150
Other	3	5074	0	0.0005912
European (non-Finnish)	5	62976	0	0.00007940
Ashkenazi Jewish	0	8378	0	0.000
East Asian	0	11986	0	0.000

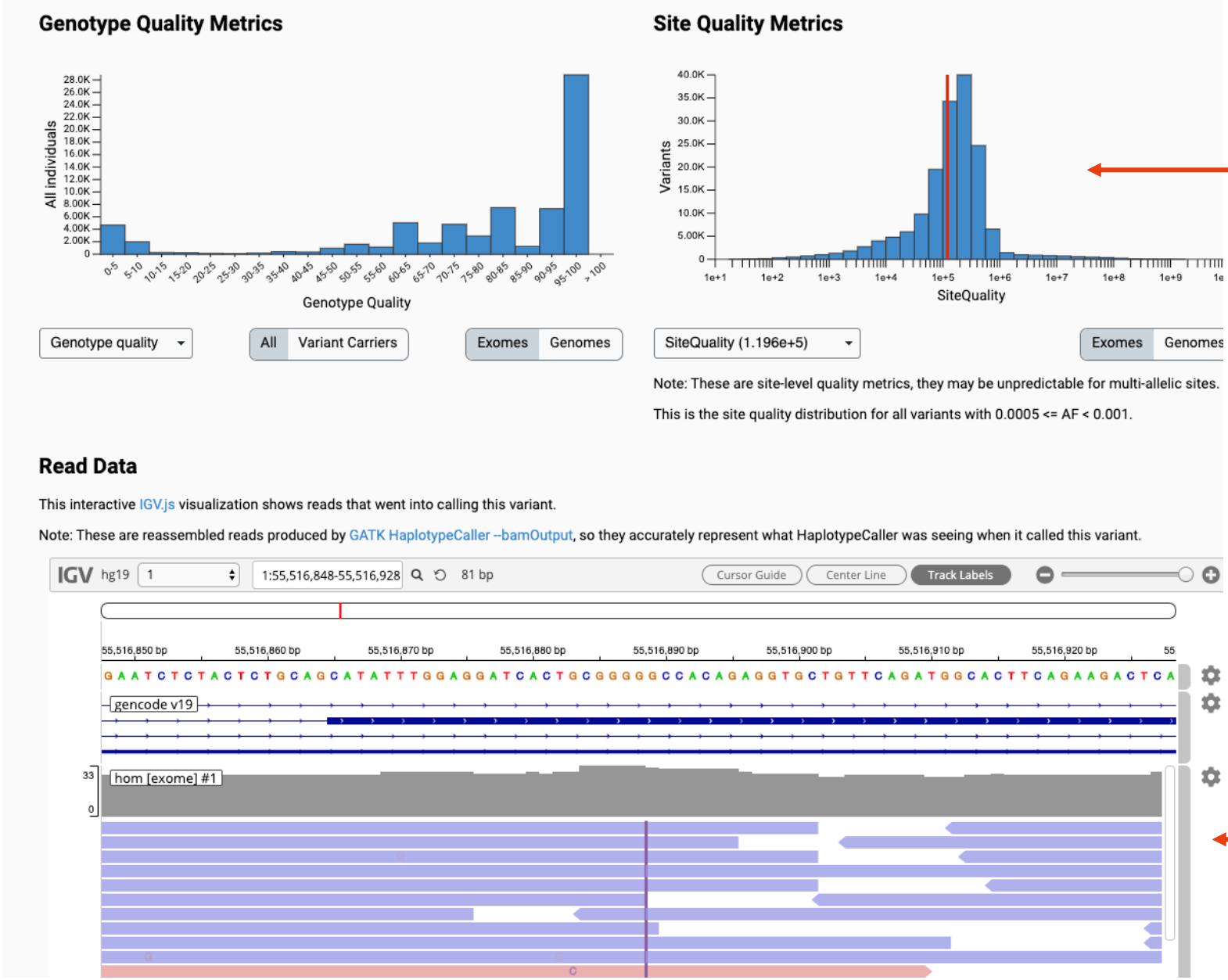
Age Distribution



See age breakdown

gnomAD

link



Quality metrics

Visualise raw data

gnomAD

Exercise…

What is the size of the structural variant in the BRAF gene with the highest allele frequency?

515 bp

Challenge…

In what population (ie ethnicity) does this variant have the highest allele frequency?

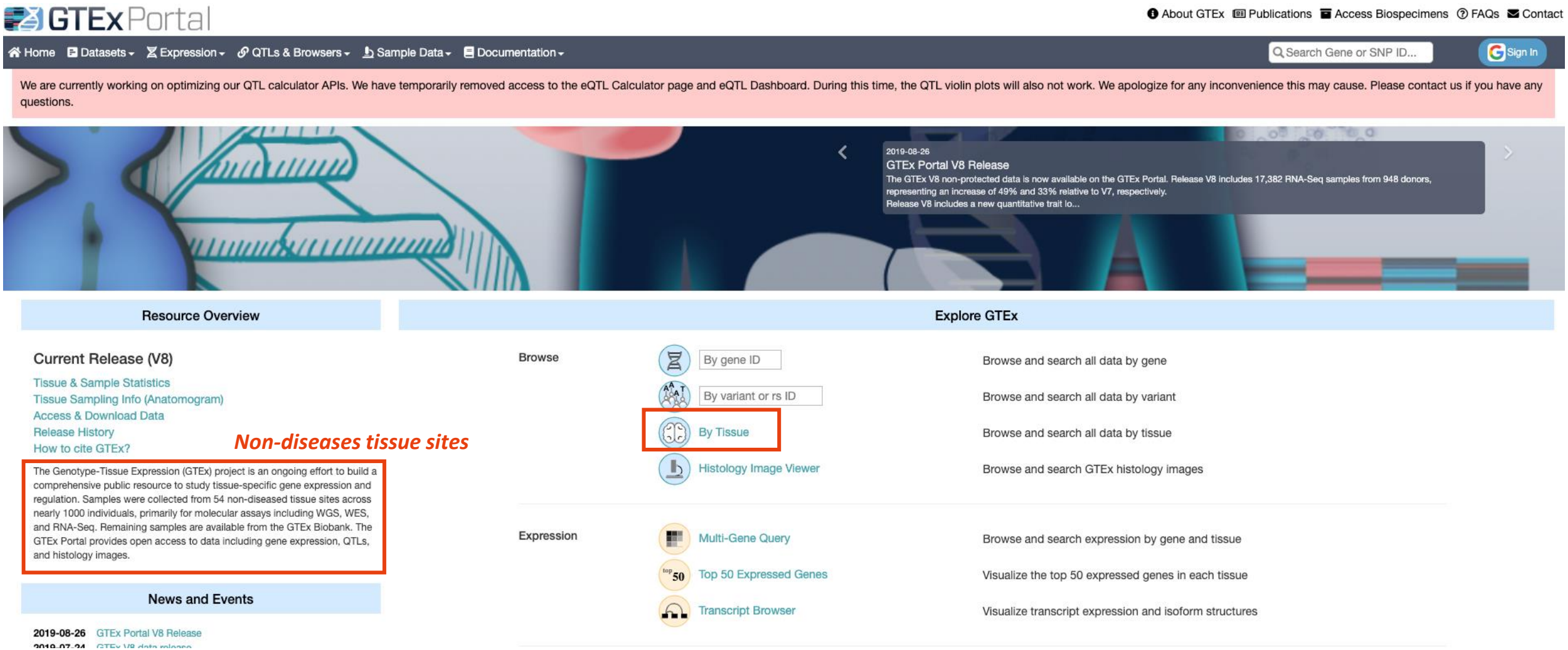
East Asian

Genotype-Tissue Expression (GTEx) Portal

Get data for tissue-specific gene expression and regulation

GTEx Portal

https://gtexportal.org/home/



GTEx Portal

link

Tissue Page

Colon - Sigmoid

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

Colon - Sigmoid (n=318) Change Tissue

Tissue Information

Colon - Sigmoid	Counts
Total samples	373
Total samples with donor genotype	318
Number of significant eGenes	10550
Number of significant sGenes	3269

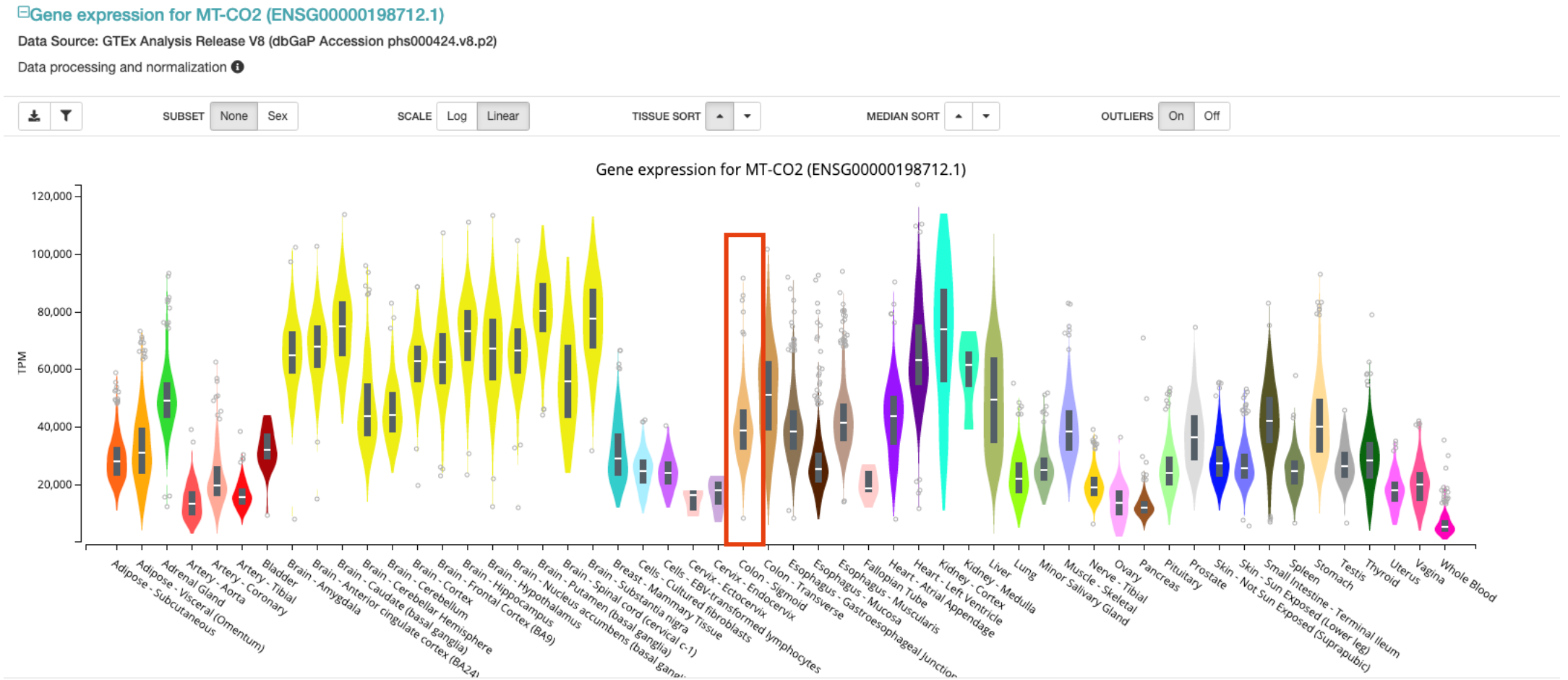
Top 100 Expressed Genes in Colon - Sigmoid

Copy CSV Include Mitochondrial Genes Exclude Mitochondrial Genes

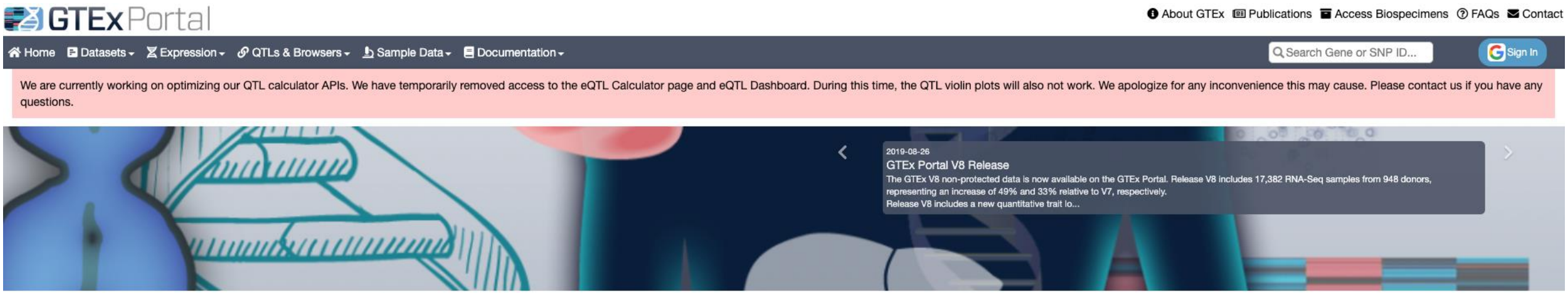
Search: Show 10 entries

Gencode Id	Gene Symbol	Median TPM
ENSG00000198712.1	MT-CO2	38803.7
ENSG00000198886.2	MT-ND4	36183.8
ENSG00000198899.2	MT-ATP6	35550.3
ENSG00000198938.2	MT-CO3	33373.4
ENSG00000198804.2	MT-CO1	27547.2

GTEx Portal


Gene expression for MT-CO2 (ENSG00000198712.1)
Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)
Data processing and normalization
SUBSET None Sex
SCALE Log Linear
TISSUE SORT
MEDIAN SORT
OUTLIERS On Off
Gene expression for MT-CO2 (ENSG00000198712.1)
TPM
120,000
100,000
80,000
60,000
40,000
20,000
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - Cultured fibroblasts
Cells - EBV-transformed lymphocytes
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Kidney - Medulla
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

GTEx Portal


GTEx Portal
About GTEx
Publications
Access Biospecimens
FAQs
Contact
Home
Datasets
Expression
QTLs & Browsers
Sample Data
Documentation
Search Gene or SNP ID...
Sign In
We are currently working on optimizing our QTL calculator APIs. We have temporarily removed access to the eQTL Calculator page and eQTL Dashboard. During this time, the QTL violin plots will also not work. We apologize for any inconvenience this may cause. Please contact us if you have any questions.
2019-08-26
GTEx Portal V8 Release
The GTEx V8 non-protected data is now available on the GTEx Portal. Release V8 includes 17,382 RNA-Seq samples from 948 donors, representing an increase of 49% and 33% relative to V7, respectively.
Release V8 includes a new quantitative trait lo...
Resource Overview
Current Release (V8)
Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?
The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.
News and Events
2019-08-26 GTEx Portal V8 Release
Explore GTEx
Browse
By gene ID
Browse and search all data by gene
By variant or rs ID
Browse and search all data by variant
By Tissue
Browse and search all data by tissue
Histology Image Viewer
Browse and search GTEx histology images
Expression
Multi-Gene Query
Browse and search expression by gene and tissue
Top 50 Expressed Genes
Visualize the top 50 expressed genes in each tissue
Transcript Browser
Visualize transcript expression and isoform structures

GTEx Portal

KLK3 – Prostate Specific Antigen (PSA)

Gene Symbol	Gencode ID	Entrez Gene ID	Location	Gene Description	Actions
KLK3	ENSG00000142515.14	354	chr19:50854915-50860764:+	kallikrein related peptidase 3 [Source:HGNC Symbol;Acc:HGNC:6364]	IGV eQTL Browser, Ensembl, UCSC

Showing 1 to 1 of 1 entries

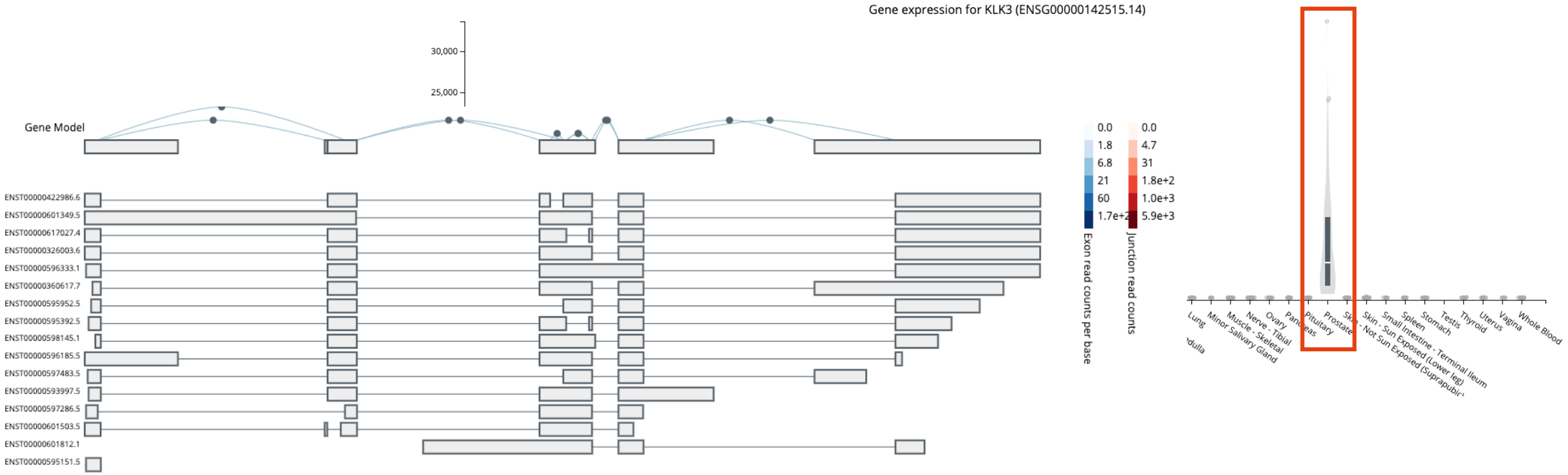
First Previous 1 Next Last

Gene expression for KLK3 (ENSG00000142515.14)

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

Data processing and normalization

SUBSET None Sex SCALE Log Linear TISSUE SORT MEDIAN SORT OUTLIERS On Off



GTEx Portal

Exercise…

Which cell type has the highest median expression of the PTEN gene?

EBV-transformed lymphocytes

Challenge…

Which PTEN exon has the highest median read count per base in this cell type?

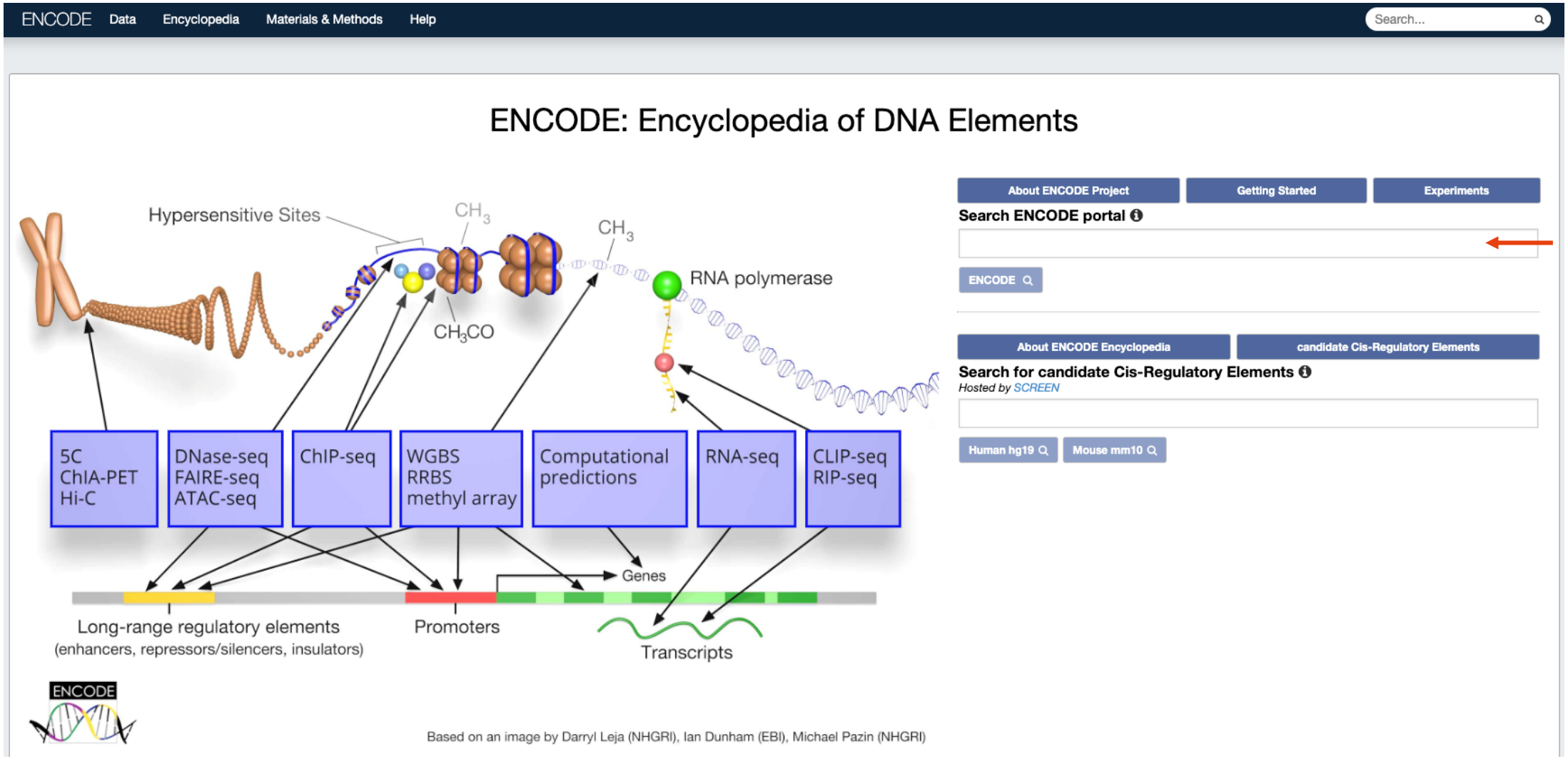
Exon 8

Encyclopedia of DNA Elements (ENCODE)

Identify functional elements in the human genome

ENCODE

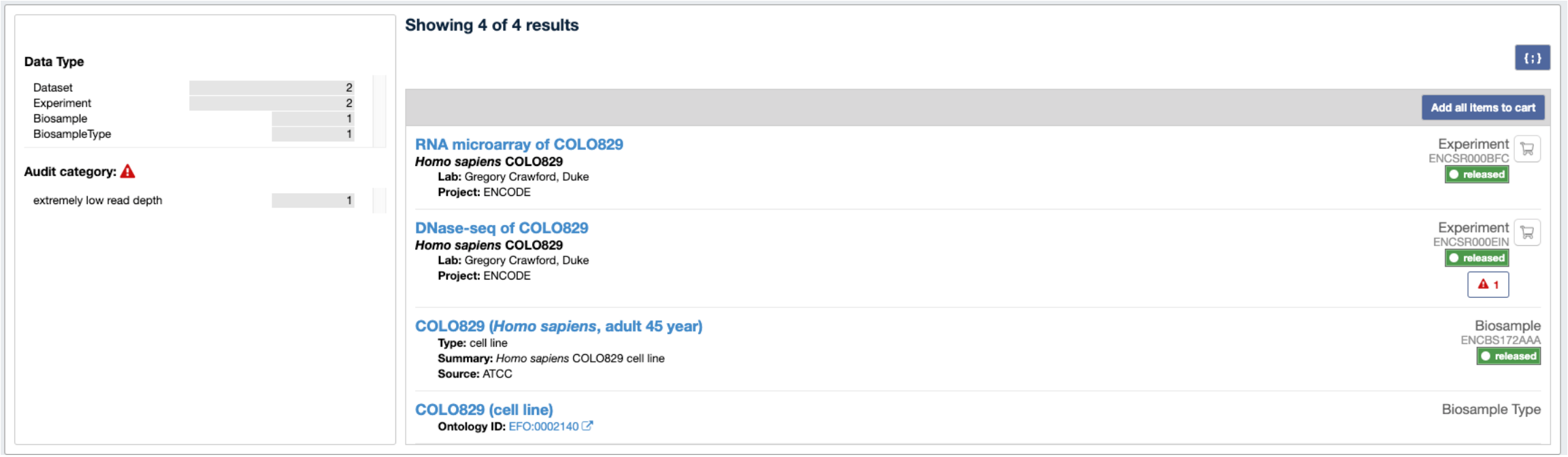
https://www.encodeproject.org/



Search bar

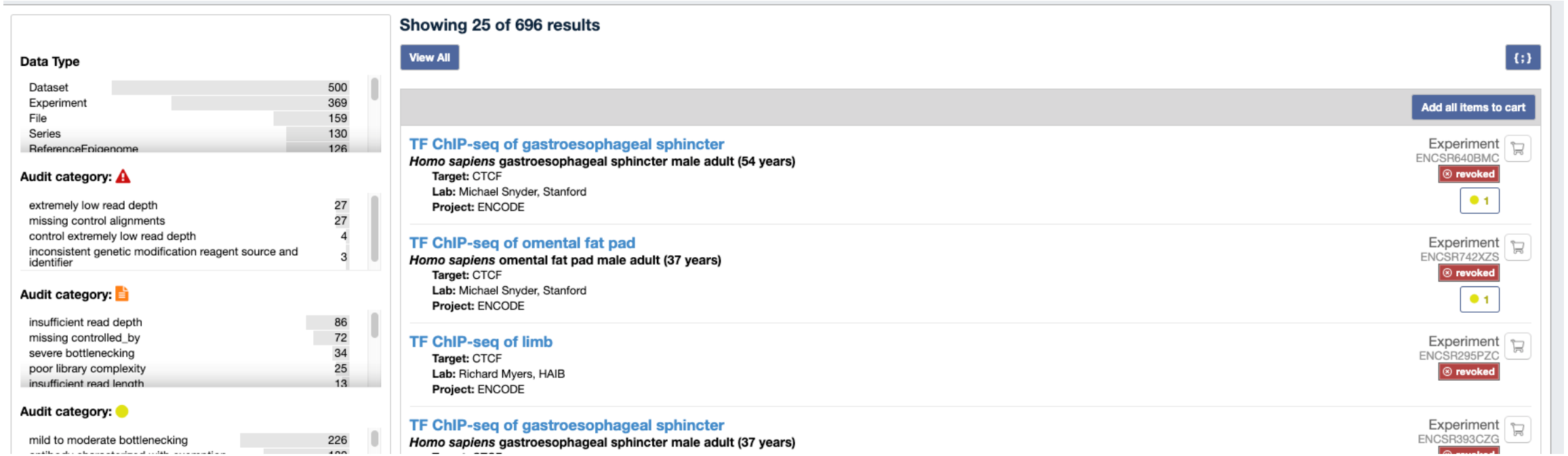
ENCODE

Search results for "COLO829" cell line

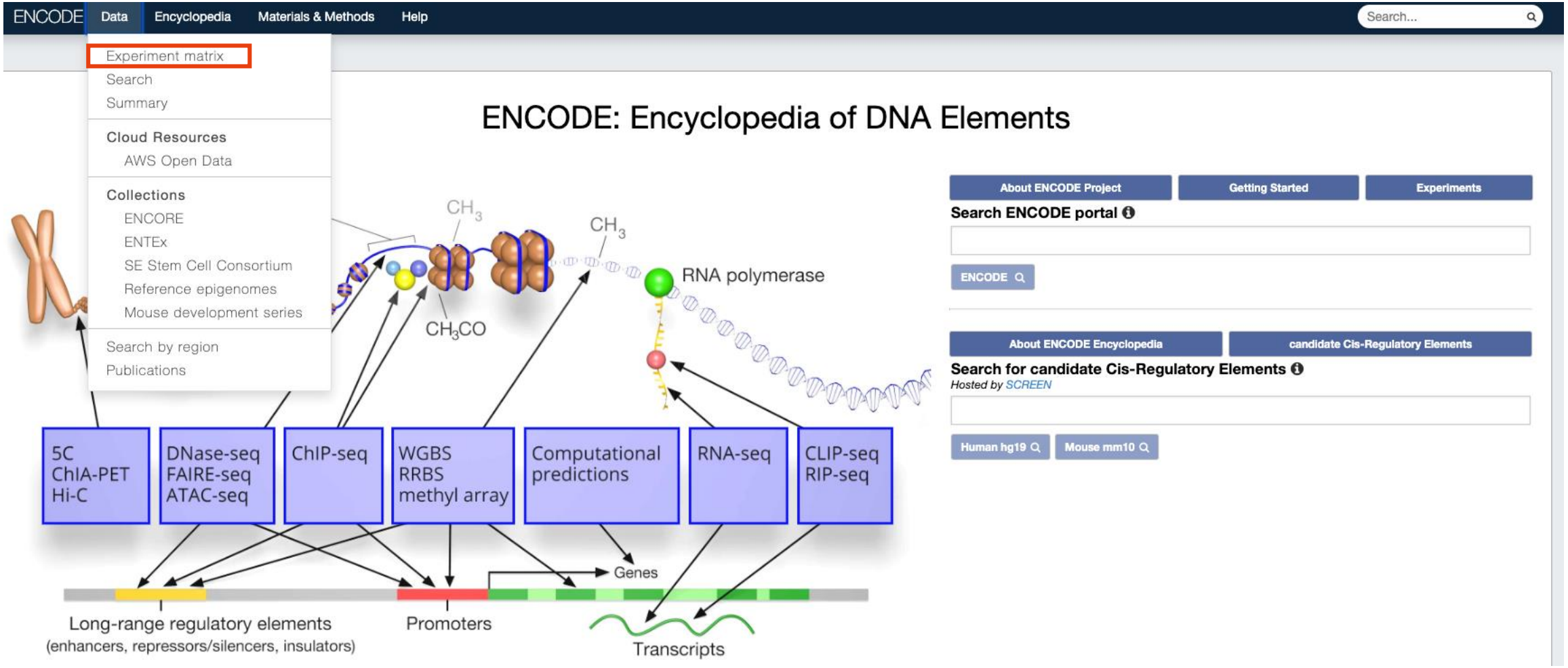


ENCODE

Search results for "CTCF ChIP-seq"

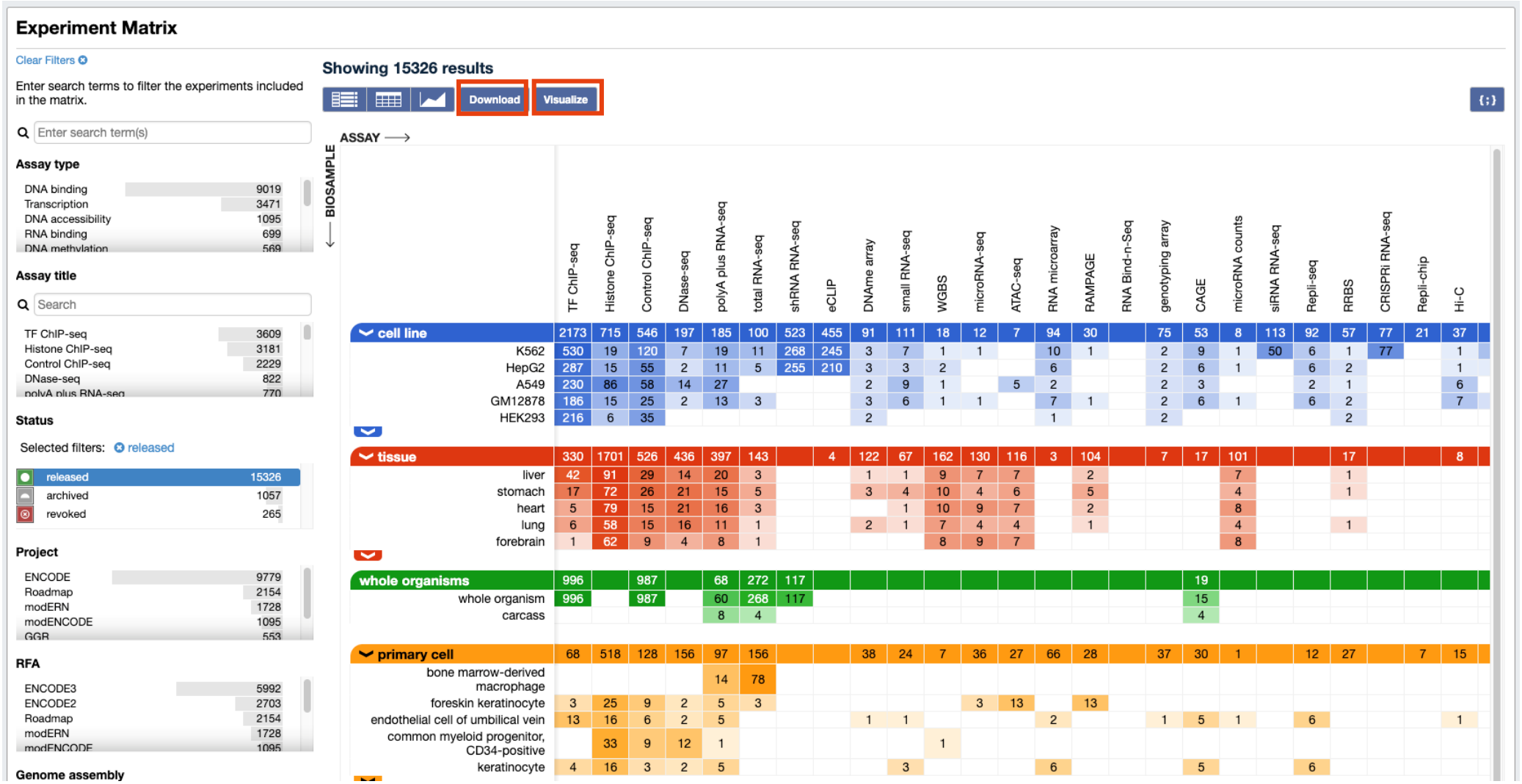


ENCODE

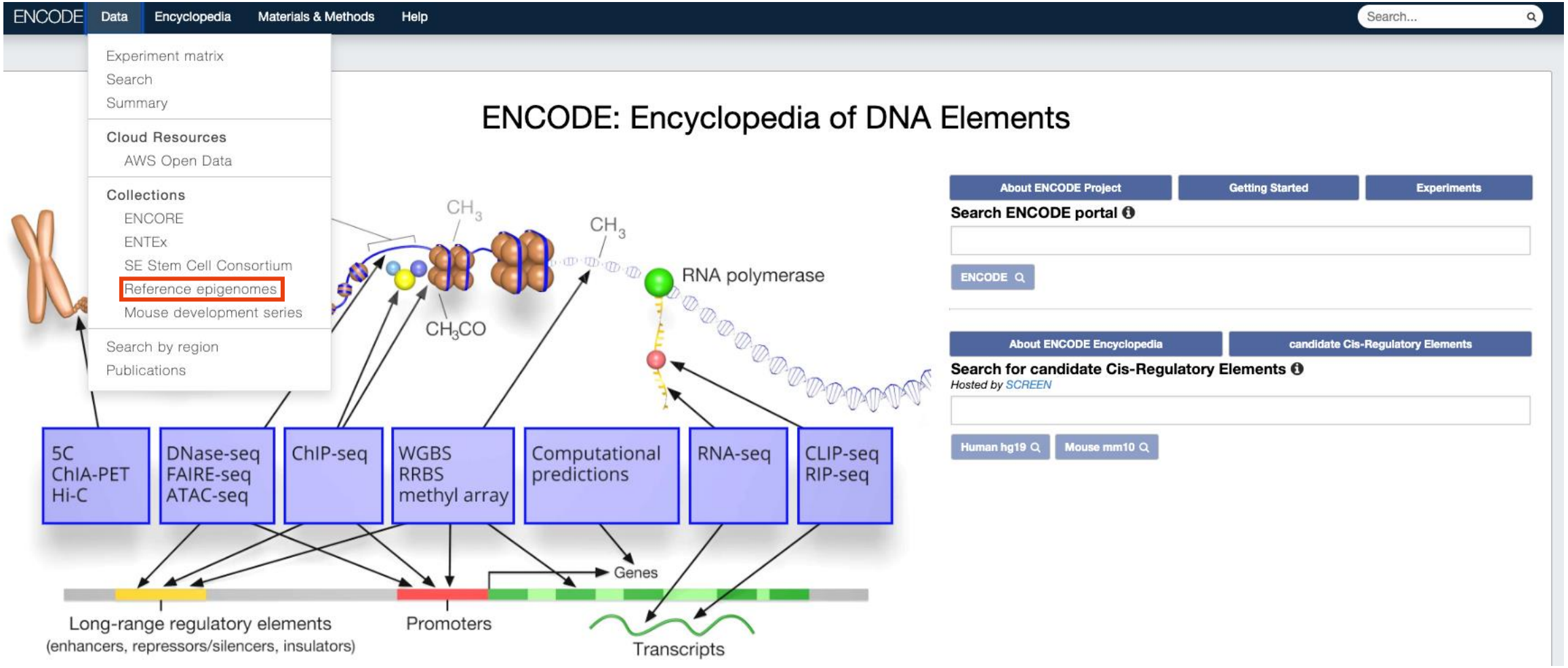

ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Experiment matrix
Search
Summary
Cloud Resources
AWS Open Data
Collections
ENCORE
ENTEx
SE Stem Cell Consortium
Reference epigenomes
Mouse development series
Search by region
Publications
ENCODE: Encyclopedia of DNA Elements
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational
predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
About ENCODE Project
Getting Started
Experiments
Search ENCODE portal
ENCODE
About ENCODE Encyclopedia
candidate Cis-Regulatory Elements
Search for candidate Cis-Regulatory Elements
Hosted by SCREEN
Human hg19
Mouse mm10

ENCODE

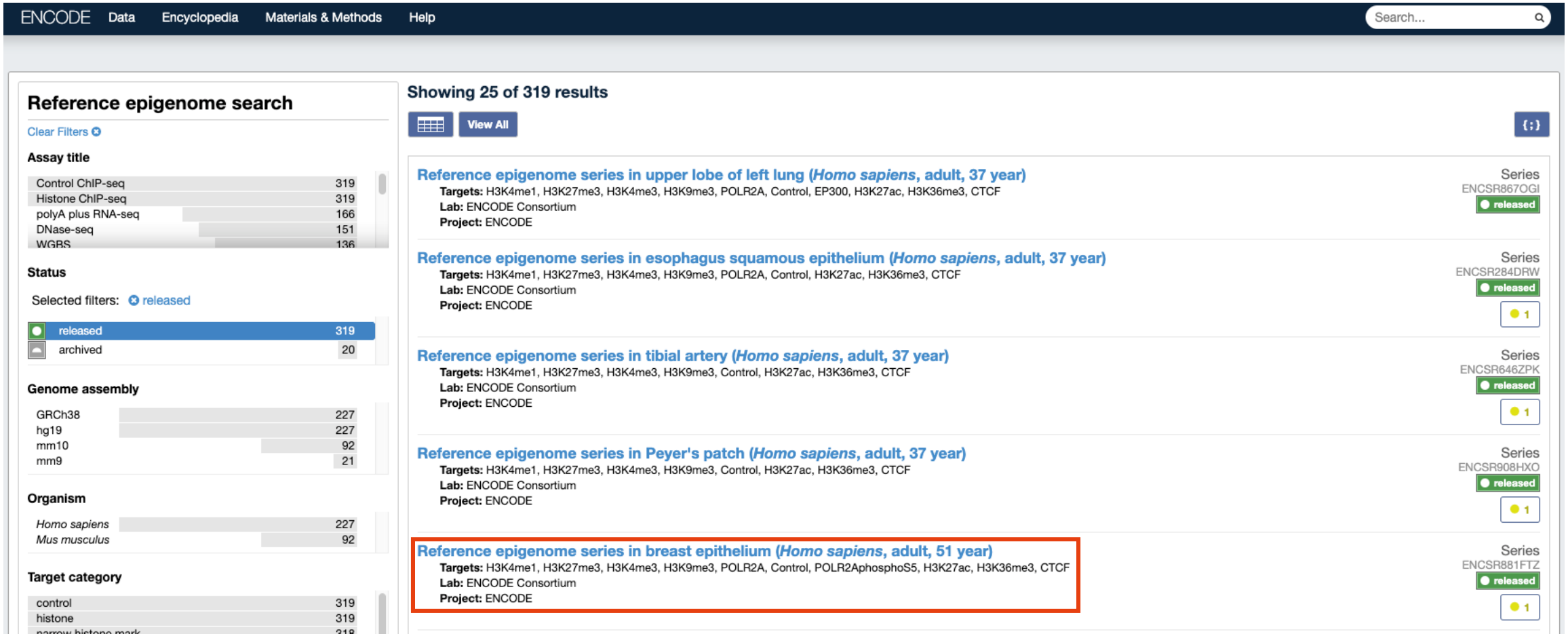
link


Experiment Matrix
Clear Filters
Enter search terms to filter the experiments included in the matrix.
Enter search term(s)
Assay type
DNA binding 9019
Transcription 3471
DNA accessibility 1095
RNA binding 699
DNA methylation 569
Assay title
Search
TF ChIP-seq 3609
Histone ChIP-seq 3181
Control ChIP-seq 2229
DNase-seq 822
polyA plus RNA-seq 770
Status
Selected filters: released
released 15326
archived 1057
revoked 265
Project
ENCODE 9779
Roadmap 2154
modERN 1728
modENCODE 1095
GGR 553
RFA
ENCODE3 5992
ENCODE2 2703
Roadmap 2154
modERN 1728
modENCODE 1095
Genome assembly
Showing 15326 results
Download
Visualize
ASSAY
BIOSAMPLE
TF ChIP-seq
Histone ChIP-seq
Control ChIP-seq
DNase-seq
polyA plus RNA-seq
total RNA-seq
shRNA RNA-seq
eCLIP
DNAme array
small RNA-seq
WGBS
microRNA-seq
ATAC-seq
RNA microarray
RAMPAGE
RNA Bind-n-Seq
genotyping array
CAGE
microRNA counts
siRNA RNA-seq
Repli-seq
RRBS
CRISPRi RNA-seq
Repli-chip
Hi-C
cell line 2173 715 546 197 185 100 523 455 91 111 18 12 7 94 30 75 53 8 113 92 57 77 21 37
K562 530 19 120 7 19 11 268 245 3 7 1 1 10 1 2 9 1 50 6 1 77 1
HepG2 287 15 55 2 11 5 255 210 3 3 2 6 2 6 1 6 2 1
A549 230 86 58 14 27 2 9 1 5 2 2 3 2 1 6
GM12878 186 15 25 2 13 3 3 6 1 1 7 1 2 6 1 6 2 7
HEK293 216 6 35 2 1 2 2
tissue 330 1701 526 436 397 143 4 122 67 162 130 116 3 104 7 17 101 17 8
liver 42 91 29 14 20 3 1 1 9 7 7 2 7 1
stomach 17 72 26 21 15 5 3 4 10 4 6 5 4 1
heart 5 79 15 21 16 3 1 10 9 7 2 8
lung 6 58 15 16 11 1 2 1 7 4 4 1 4 1
forebrain 1 62 9 4 8 1 8 9 7 8
whole organisms 996 987 68 272 117 19
whole organism 996 987 60 268 117 15
carcass 8 4 4
primary cell 68 518 128 156 97 156 38 24 7 36 27 66 28 37 30 1 12 27 7 15
bone marrow-derived macrophage 14 78
foreskin keratinocyte 3 25 9 2 5 3 3 13 13
endothelial cell of umbilical vein 13 16 6 2 5 1 1 2 1 5 1 6 1
common myeloid progenitor, CD34-positive 33 9 12 1 1
keratinocyte 4 16 3 2 5 3 6 5 6

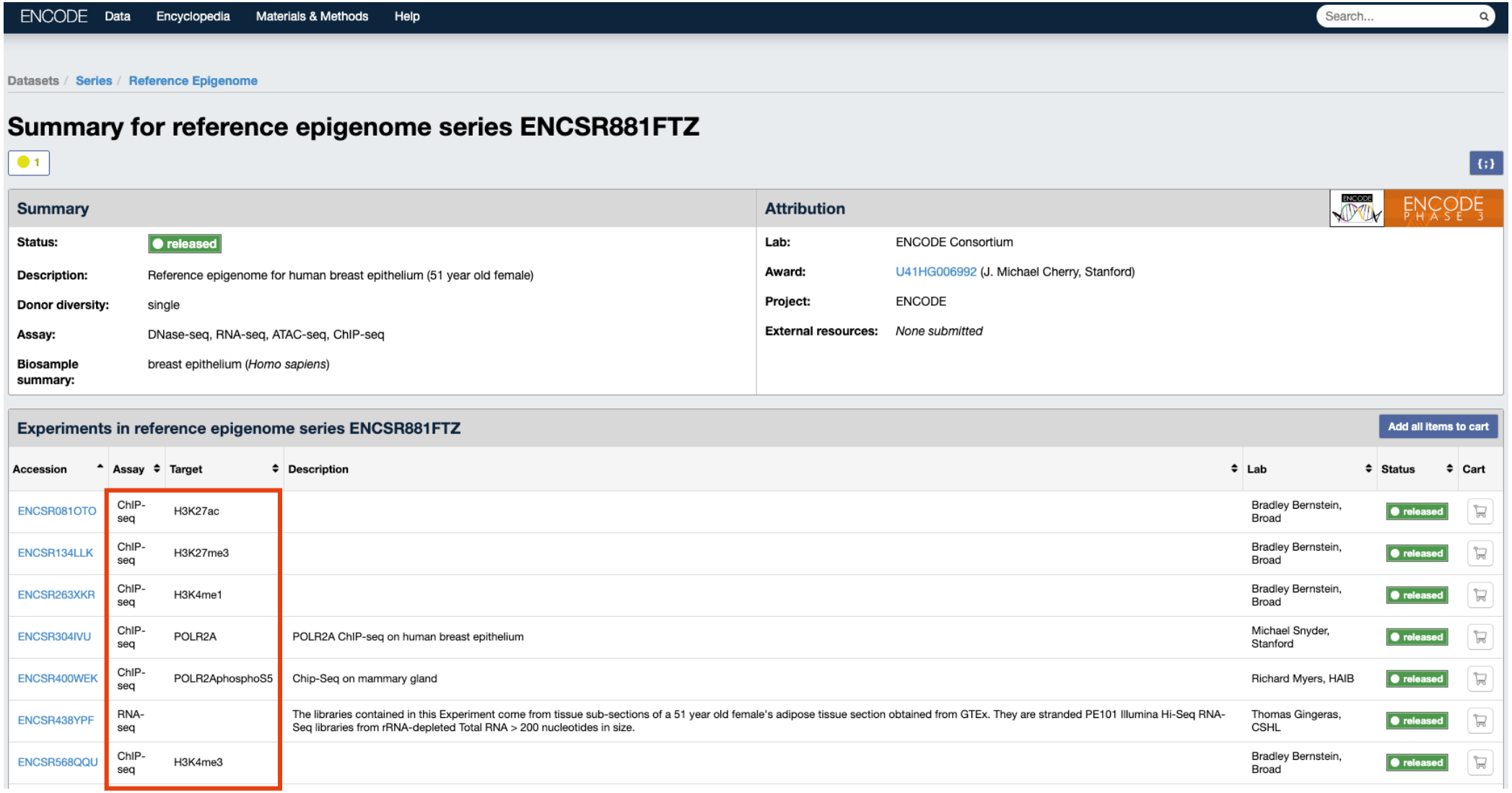
ENCODE


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Experiment matrix
Search
Summary
Cloud Resources
AWS Open Data
Collections
ENCORE
ENTEx
SE Stem Cell Consortium
Reference epigenomes
Mouse development series
Search by region
Publications
ENCODE: Encyclopedia of DNA Elements
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational
predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
About ENCODE Project
Getting Started
Experiments
Search ENCODE portal
ENCODE
About ENCODE Encyclopedia
candidate Cis-Regulatory Elements
Search for candidate Cis-Regulatory Elements
Hosted by SCREEN
Human hg19
Mouse mm10

ENCODE


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Search...
Reference epigenome search
Clear Filters
Assay title
Control ChIP-seq 319
Histone ChIP-seq 319
polyA plus RNA-seq 166
DNase-seq 151
WGBS 136
Status
Selected filters: released
released 319
archived 20
Genome assembly
GRCh38 227
hg19 227
mm10 92
mm9 21
Organism
Homo sapiens 227
Mus musculus 92
Target category
control 319
histone 319
Showing 25 of 319 results
View All
Reference epigenome series in upper lobe of left lung (Homo sapiens, adult, 37 year)
Targets: H3K4me1, H3K27me3, H3K4me3, H3K9me3, POLR2A, Control, EP300, H3K27ac, H3K36me3, CTCF
Lab: ENCODE Consortium
Project: ENCODE
Series
ENCSR867OGI
released
Reference epigenome series in esophagus squamous epithelium (Homo sapiens, adult, 37 year)
Targets: H3K4me1, H3K27me3, H3K4me3, H3K9me3, POLR2A, Control, H3K27ac, H3K36me3, CTCF
Lab: ENCODE Consortium
Project: ENCODE
Series
ENCSR284DRW
released
1
Reference epigenome series in tibial artery (Homo sapiens, adult, 37 year)
Targets: H3K4me1, H3K27me3, H3K4me3, H3K9me3, Control, H3K27ac, H3K36me3, CTCF
Lab: ENCODE Consortium
Project: ENCODE
Series
ENCSR646ZPK
released
1
Reference epigenome series in Peyer's patch (Homo sapiens, adult, 37 year)
Targets: H3K4me1, H3K27me3, H3K4me3, H3K9me3, Control, H3K27ac, H3K36me3, CTCF
Lab: ENCODE Consortium
Project: ENCODE
Series
ENCSR908HXO
released
1
Reference epigenome series in breast epithelium (Homo sapiens, adult, 51 year)
Targets: H3K4me1, H3K27me3, H3K4me3, H3K9me3, POLR2A, Control, POLR2AphosphoS5, H3K27ac, H3K36me3, CTCF
Lab: ENCODE Consortium
Project: ENCODE
Series
ENCSR881FTZ
released
1

ENCODE



ENCODE Data Encyclopedia Materials & Methods Help Search...

Datasets / Series / Reference Epigenome

Summary for reference epigenome series ENCSR881FTZ

1

Summary

Status:	released
Description:	Reference epigenome for human breast epithelium (51 year old female)
Donor diversity:	single
Assay:	DNase-seq, RNA-seq, ATAC-seq, ChIP-seq
Biosample summary:	breast epithelium (*Homo sapiens*)

Attribution

ENCODE PHASE 3

Lab:	ENCODE Consortium
Award:	U41HG006992 (J. Michael Cherry, Stanford)
Project:	ENCODE
External resources:	*None submitted*

Experiments in reference epigenome series ENCSR881FTZ

Add all items to cart

Accession	Assay	Target	Description	Lab	Status	Cart
ENCSR081OTO	ChIP-seq	H3K27ac		Bradley Bernstein, Broad	released	
ENCSR134LLK	ChIP-seq	H3K27me3		Bradley Bernstein, Broad	released	
ENCSR263XKR	ChIP-seq	H3K4me1		Bradley Bernstein, Broad	released	
ENCSR304IVU	ChIP-seq	POLR2A	POLR2A ChIP-seq on human breast epithelium	Michael Snyder, Stanford	released	
ENCSR400WEK	ChIP-seq	POLR2AphosphoS5	Chip-Seq on mammary gland	Richard Myers, HAIB	released	
ENCSR438YPF	RNA-seq		The libraries contained in this Experiment come from tissue sub-sections of a 51 year old female's adipose tissue section obtained from GTEx. They are stranded PE101 Illumina Hi-Seq RNA-Seq libraries from rRNA-depleted Total RNA > 200 nucleotides in size.	Thomas Gingeras, CSHL	released	
ENCSR568QQU	ChIP-seq	H3K4me3		Bradley Bernstein, Broad	released	

ENCODE

Exercise…

What is the experiment number (ie ENCXXXXXXXX) for WGBS in the K562 cell line?

ENCSR765JPC

Challenge…

What is the command you can use to download these data?

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

Exercise: Putting it all together

Exercise: Putting it all together

1. **Go to the region chr5:112,838,930-112,838,940 in hg38. What gene spans this region?**

   APC

2. **What is the methylation level (hint: use WGBS data) of this site in human sigmoid colon tissue? (You will need to use the ENCODE database).**

   Between 90-100%

3. ***How many samples in the TCGA cancer cohort are mutated at codon R114?***

   24

4. ***How many missense SNPs are present in this gene according to the gnomAD browser?***

   1,436

5. ***In which organ does this gene have the highest median expression? (Try to answer this question without leaving the UCSC genome browser).***

   Brain

See you back again at 3:30pm!

Part Two

Cancer genomics data resources

3:30pm – 5pm

- Brief introduction to TCGA
- cBioPortal
- Genomic Data Commons
- Xena Browser
- Cancer Genomics Cloud