ChIP-seq Datamining

Bong-Hyun Kim, Alexei Lobanov, Parthav Jailwala & Maggie Cam

CCBR

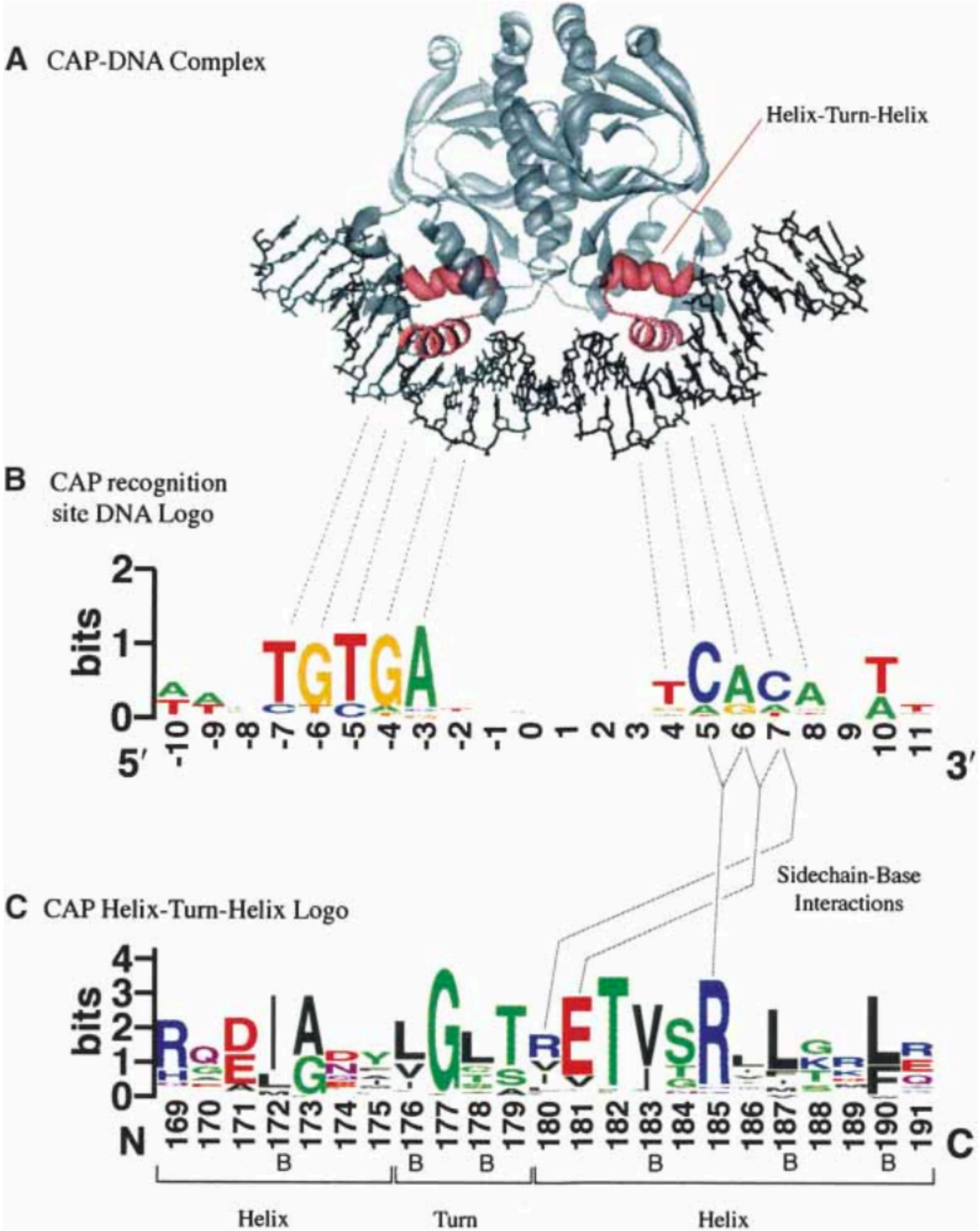
Contents

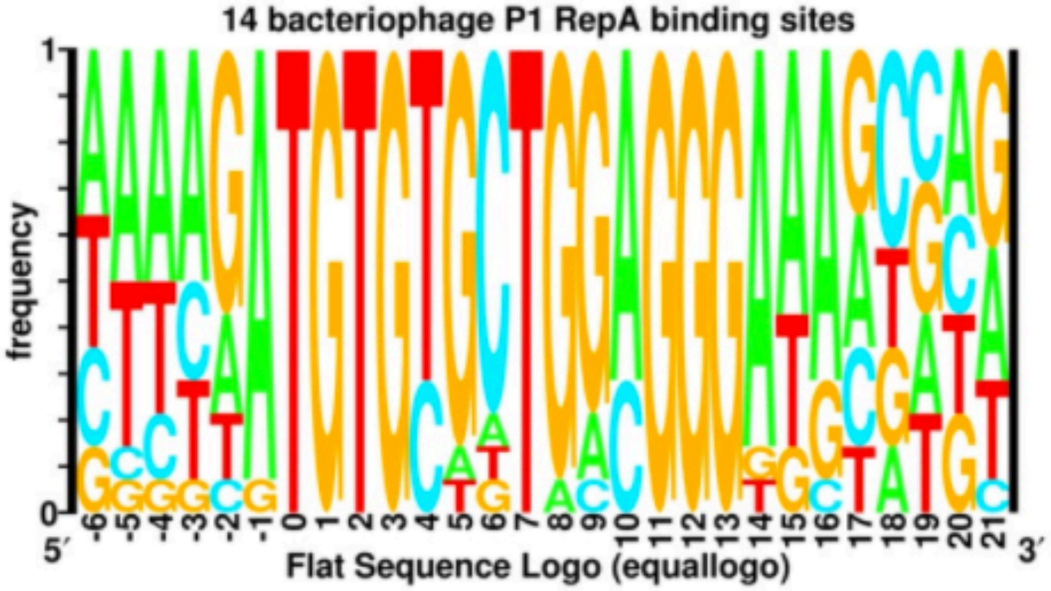
- Motif analysis
  - Motif databases
  - Motif analysis tools
  - http://ccg.vital-it.ch/chipseq/

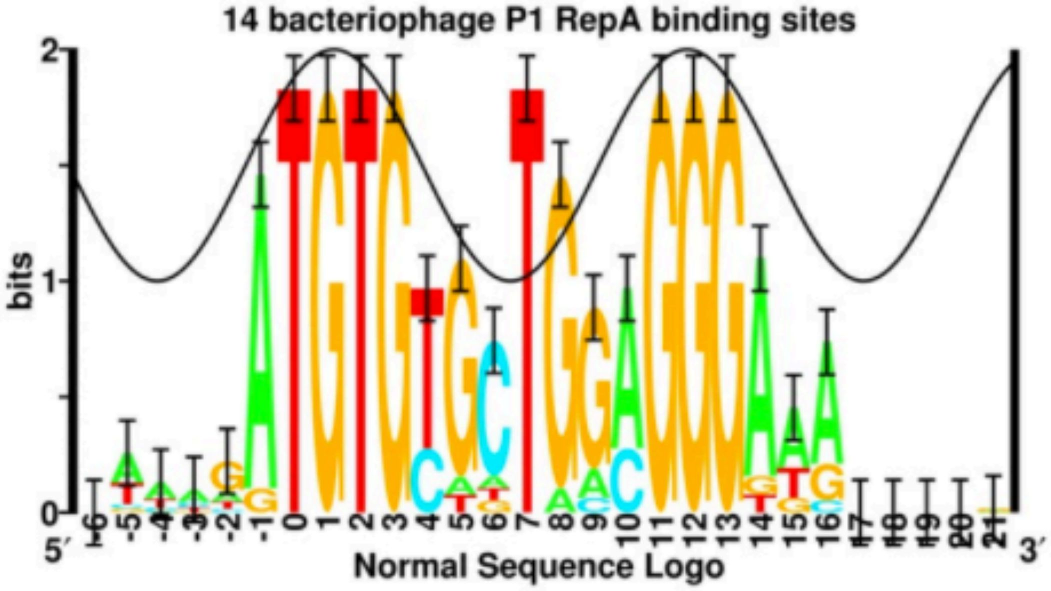
- ENCODE (ENCyclopedia Of DNA Elements) https://www.encodeproject.org/
- Mouse Encode & modENCODE
- Epigenome Roadmap

- Factorbook (http://www.factorbook.org/)
- RegulomeDB (http://regulomedb.org/)
- Cistrome (http://cistrome.org/Cistrome/Cistrome_Project.html)

When DNA meets a TF







$$H(l) = -\sum_{b=a}^{t} f(b,l)\log_2 f(b,l) \qquad \text{(bits per position)} \qquad (1)$$

where $H(l)$ is the uncertainty at position l, b is one of the bases (a, c, g, or t), and $f(b,l)$ is the frequency of base b at position l. Total information at the position is represented by the decrease in uncertainty as the binding site is located (or aligned):

$$R_{sequence}(l) = 2 - (H(l) + e(n)) \qquad \text{(bits per position)} \qquad (2)$$

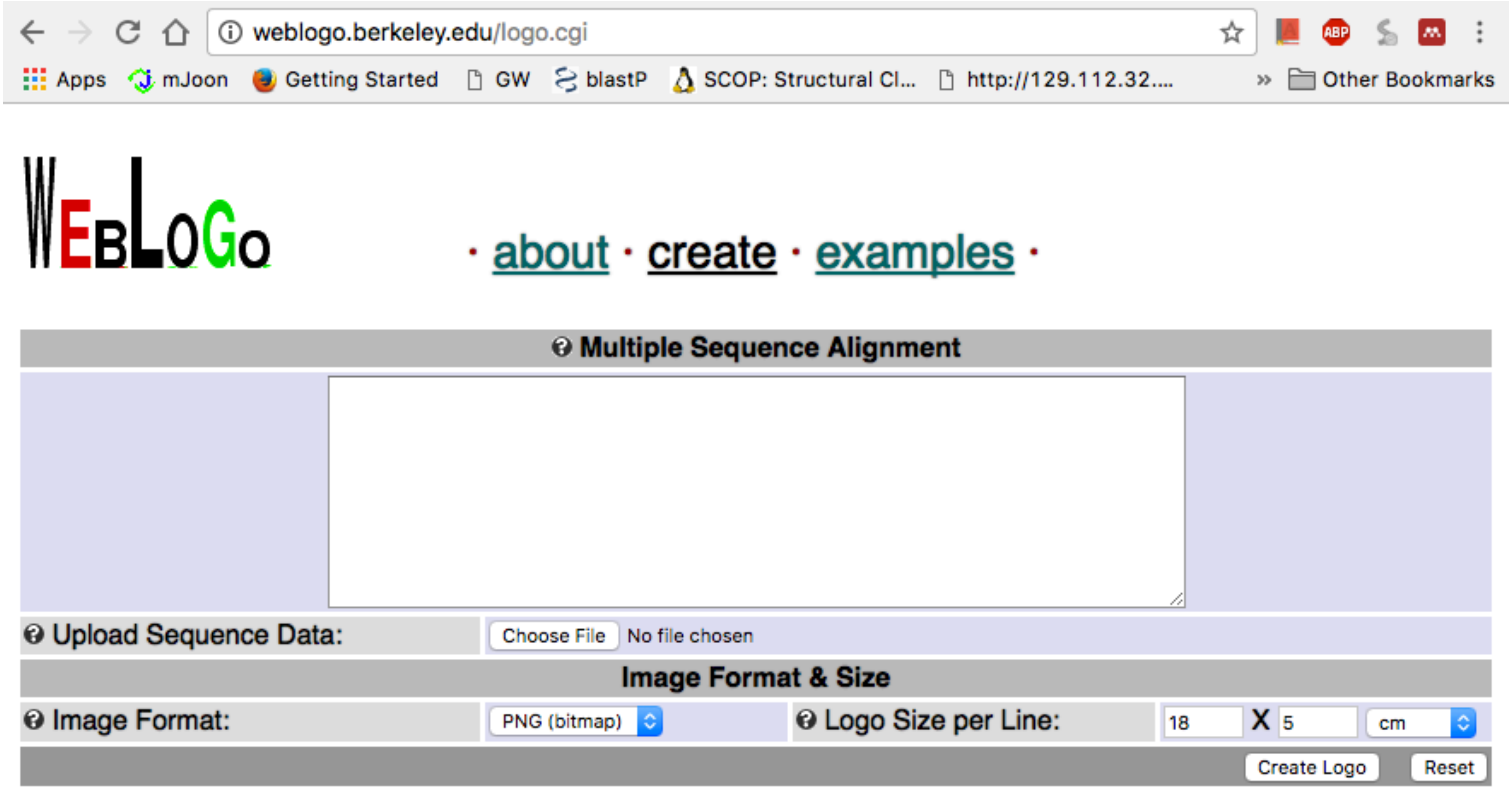
where $R_{sequence}(l)$ is the amount of information present at position l, 2 is the maximum uncertainty at any given position, and $e(n)$ is a correction factor required when one only has a few (n) sample sequences [9].

The entire set of $R_{sequence}(l)$ values forms a curve that represents the importance of various positions in the binding site [9, 10, 11]. The height of this curve is the height of the logo at that position. The size of each base printed in a logo is determined by multiplying the frequency of that base by the total information at that position:

$$\text{height of base } b \text{ at position } l = f(b,l)R_{sequence}(l). \qquad (3)$$

Version: You are using the JASPAR 2016 server: **jaspar.genereg.net/** . Previous stable JASPAR 2014 server: **jaspar2014.genereg.net**

The high-quality transcription factor binding profile database



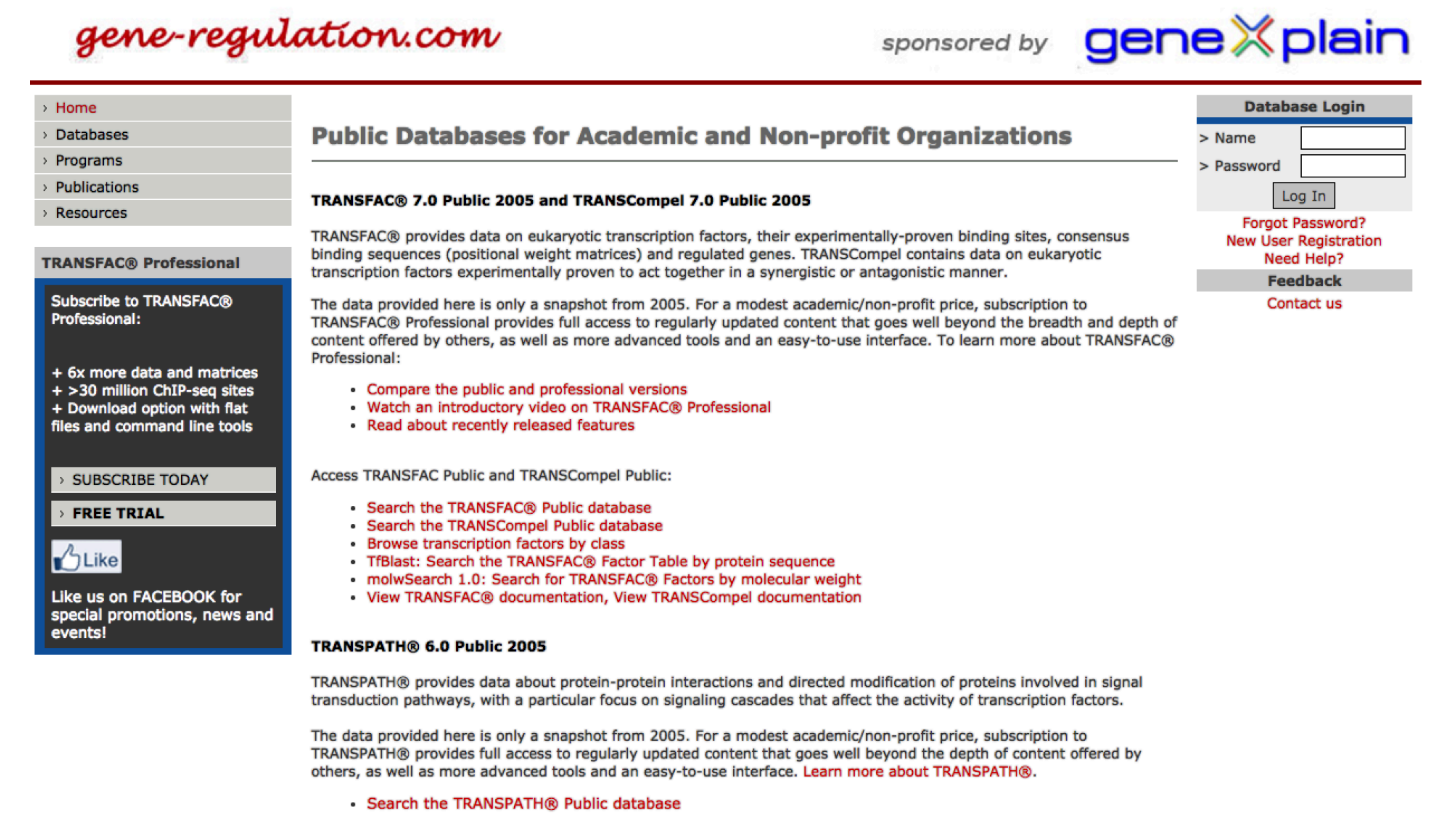
Multiple Sequence Alignment

Upload Sequence Data: Choose File No file chosen

Image Format & Size

Image Format: PNG (bitmap)

Logo Size per Line: 18 X 5 cm



TRANSFAC® Professional

Subscribe to TRANSFAC® Professional:

+ 6x more data and matrices
+ >30 million ChIP-seq sites
+ Download option with flat files and command line tools

Like us on FACEBOOK for special promotions, news and events!

Public Databases for Academic and Non-profit Organizations

TRANSFAC® 7.0 Public 2005 and TRANSCompel 7.0 Public 2005

TRANSFAC® provides data on eukaryotic transcription factors, their experimentally-proven binding sites, consensus binding sequences (positional weight matrices) and regulated genes. TRANSCompel contains data on eukaryotic transcription factors experimentally proven to act together in a synergistic or antagonistic manner.

The data provided here is only a snapshot from 2005. For a modest academic/non-profit price, subscription to TRANSFAC® Professional provides full access to regularly updated content that goes well beyond the breadth and depth of content offered by others, as well as more advanced tools and an easy-to-use interface. To learn more about TRANSFAC® Professional:

- Compare the public and professional versions
- Watch an introductory video on TRANSFAC® Professional
- Read about recently released features

Access TRANSFAC Public and TRANSCompel Public:

- Search the TRANSFAC® Public database
- Search the TRANSCompel Public database
- Browse transcription factors by class
- TfBlast: Search the TRANSFAC® Factor Table by protein sequence
- molwSearch 1.0: Search for TRANSFAC® Factors by molecular weight
- View TRANSFAC® documentation, View TRANSCompel documentation

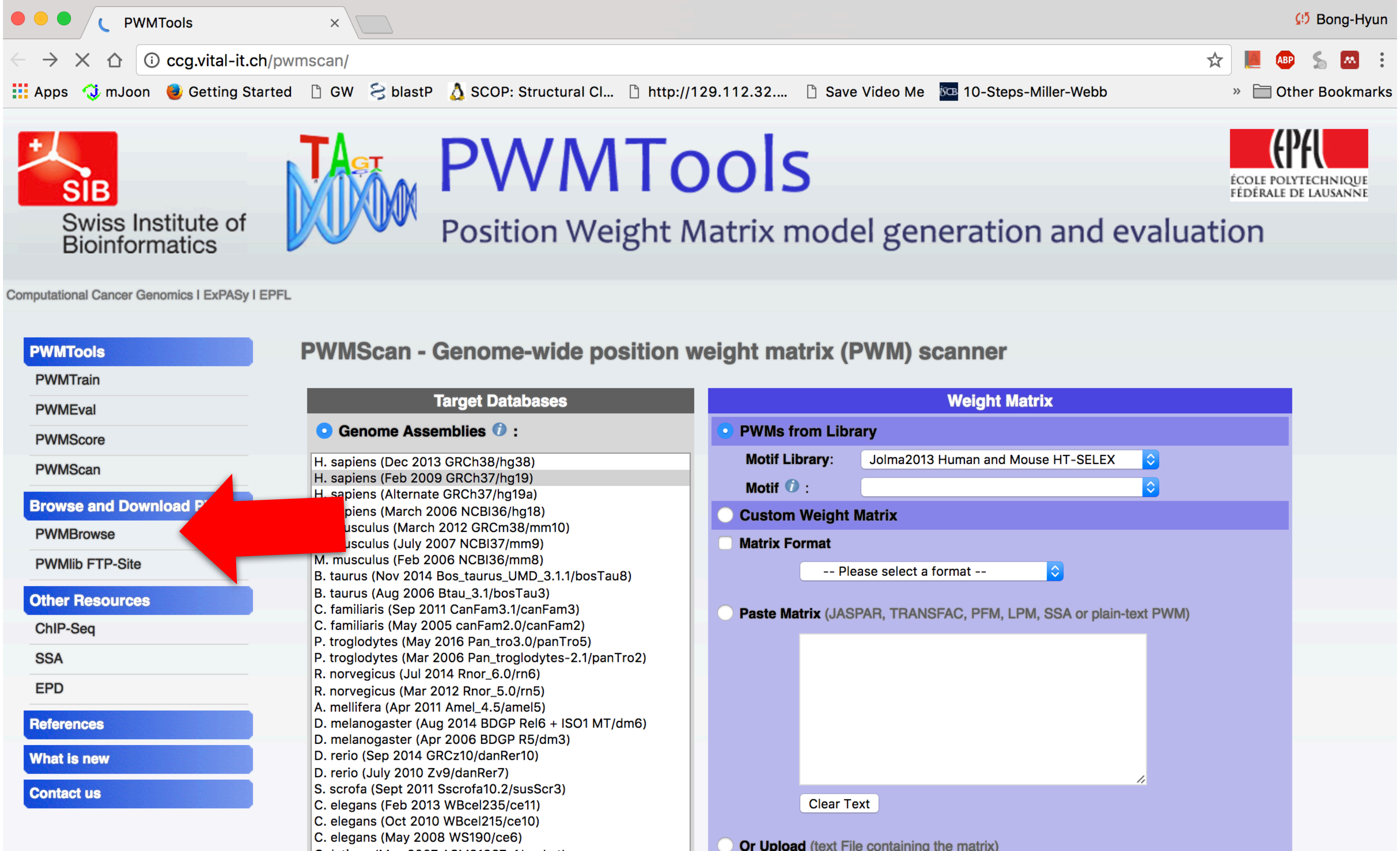
TRANSPATH® 6.0 Public 2005

TRANSPATH® provides data about protein-protein interactions and directed modification of proteins involved in signal transduction pathways, with a particular focus on signaling cascades that affect the activity of transcription factors.

The data provided here is only a snapshot from 2005. For a modest academic/non-profit price, subscription to TRANSPATH® provides full access to regularly updated content that goes well beyond the depth of content offered by others, as well as more advanced tools and an easy-to-use interface. Learn more about TRANSPATH®.

- Search the TRANSPATH® Public database

I have a motif. Where are the motifs in the genome.



PWMTools

Position Weight Matrix model generation and evaluation

Computational Cancer Genomics I ExPASy I EPFL

PWMTools
- PWMTrain
- PWMEval
- PWMScore
- PWMScan

Browse and Download P
- PWMBrowse
- PWMlib FTP-Site

Other Resources
- ChIP-Seq
- SSA
- EPD

References

What is new

Contact us

PWMScan - Genome-wide position weight matrix (PWM) scanner

Target Databases

● **Genome Assemblies** ⓘ :

H. sapiens (Dec 2013 GRCh38/hg38)
H. sapiens (Feb 2009 GRCh37/hg19)
H. sapiens (Alternate GRCh37/hg19a)
piens (March 2006 NCBI36/hg18)
usculus (March 2012 GRCm38/mm10)
usculus (July 2007 NCBI37/mm9)
M. musculus (Feb 2006 NCBI36/mm8)
B. taurus (Nov 2014 Bos_taurus_UMD_3.1.1/bosTau8)
B. taurus (Aug 2006 Btau_3.1/bosTau3)
C. familiaris (Sep 2011 CanFam3.1/canFam3)
C. familiaris (May 2005 canFam2.0/canFam2)
P. troglodytes (May 2016 Pan_tro3.0/panTro5)
P. troglodytes (Mar 2006 Pan_troglodytes-2.1/panTro2)
R. norvegicus (Jul 2014 Rnor_6.0/rn6)
R. norvegicus (Mar 2012 Rnor_5.0/rn5)
A. mellifera (Apr 2011 Amel_4.5/amel5)
D. melanogaster (Aug 2014 BDGP Rel6 + ISO1 MT/dm6)
D. melanogaster (Apr 2006 BDGP R5/dm3)
D. rerio (Sep 2014 GRCz10/danRer10)
D. rerio (July 2010 Zv9/danRer7)
S. scrofa (Sept 2011 Sscrofa10.2/susScr3)
C. elegans (Feb 2013 WBcel235/ce11)
C. elegans (Oct 2010 WBcel215/ce10)
C. elegans (May 2008 WS190/ce6)

Weight Matrix

● **PWMs from Library**

Motif Library: Jolma2013 Human and Mouse HT-SELEX

Motif ⓘ :

○ **Custom Weight Matrix**

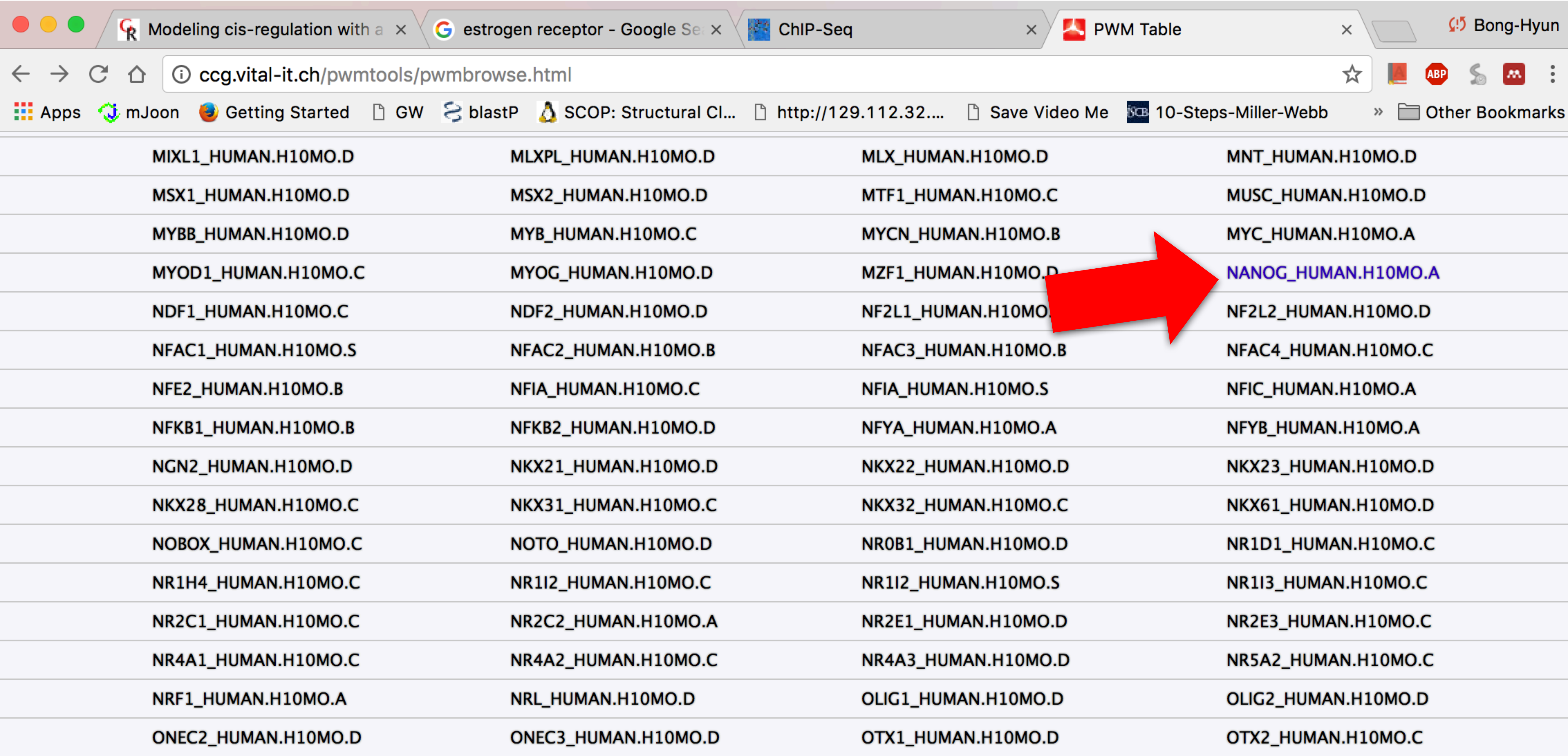
□ **Matrix Format**

-- Please select a format --

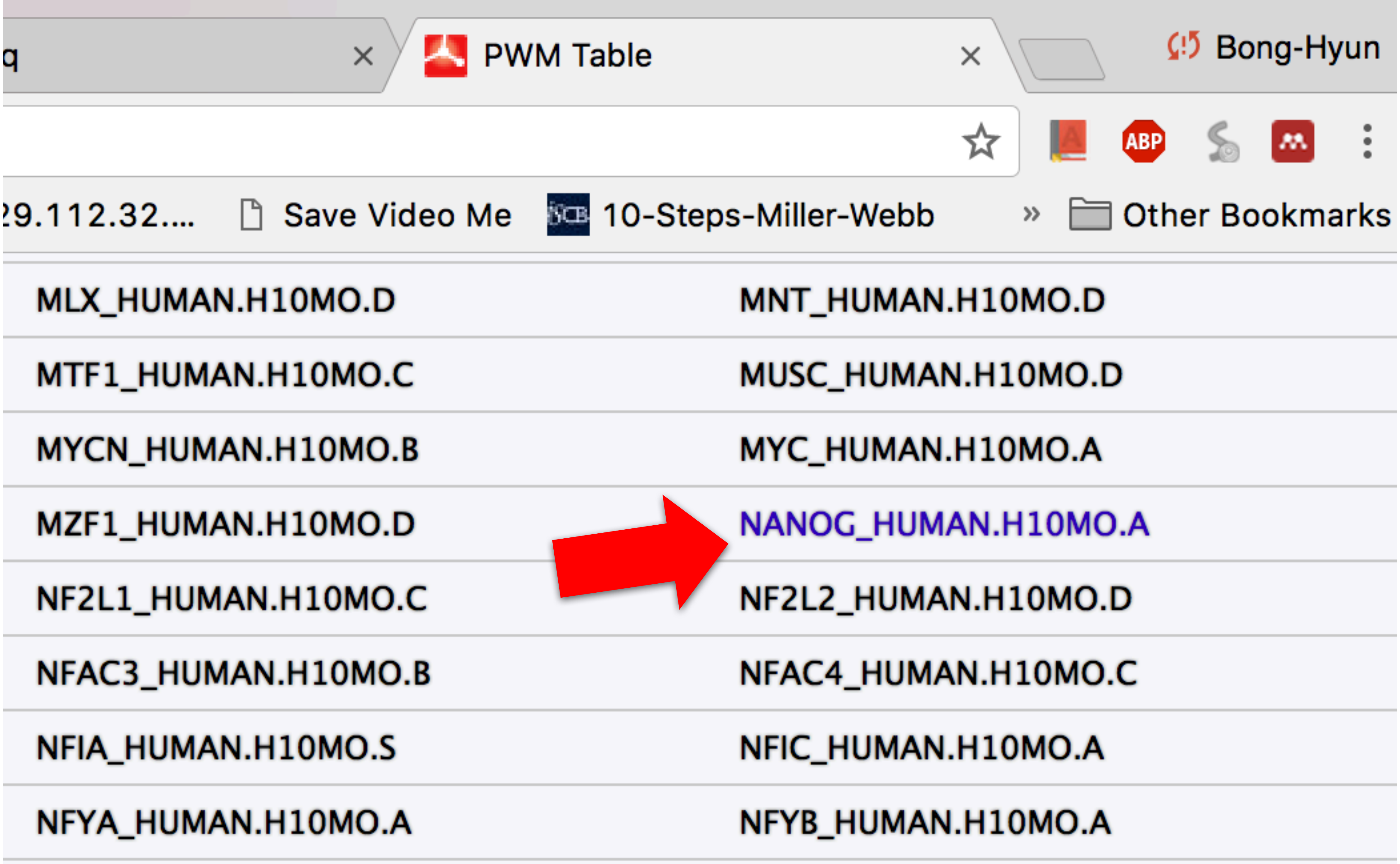
○ **Paste Matrix** (JASPAR, TRANSFAC, PFM, LPM, SSA or plain-text PWM)

Clear Text

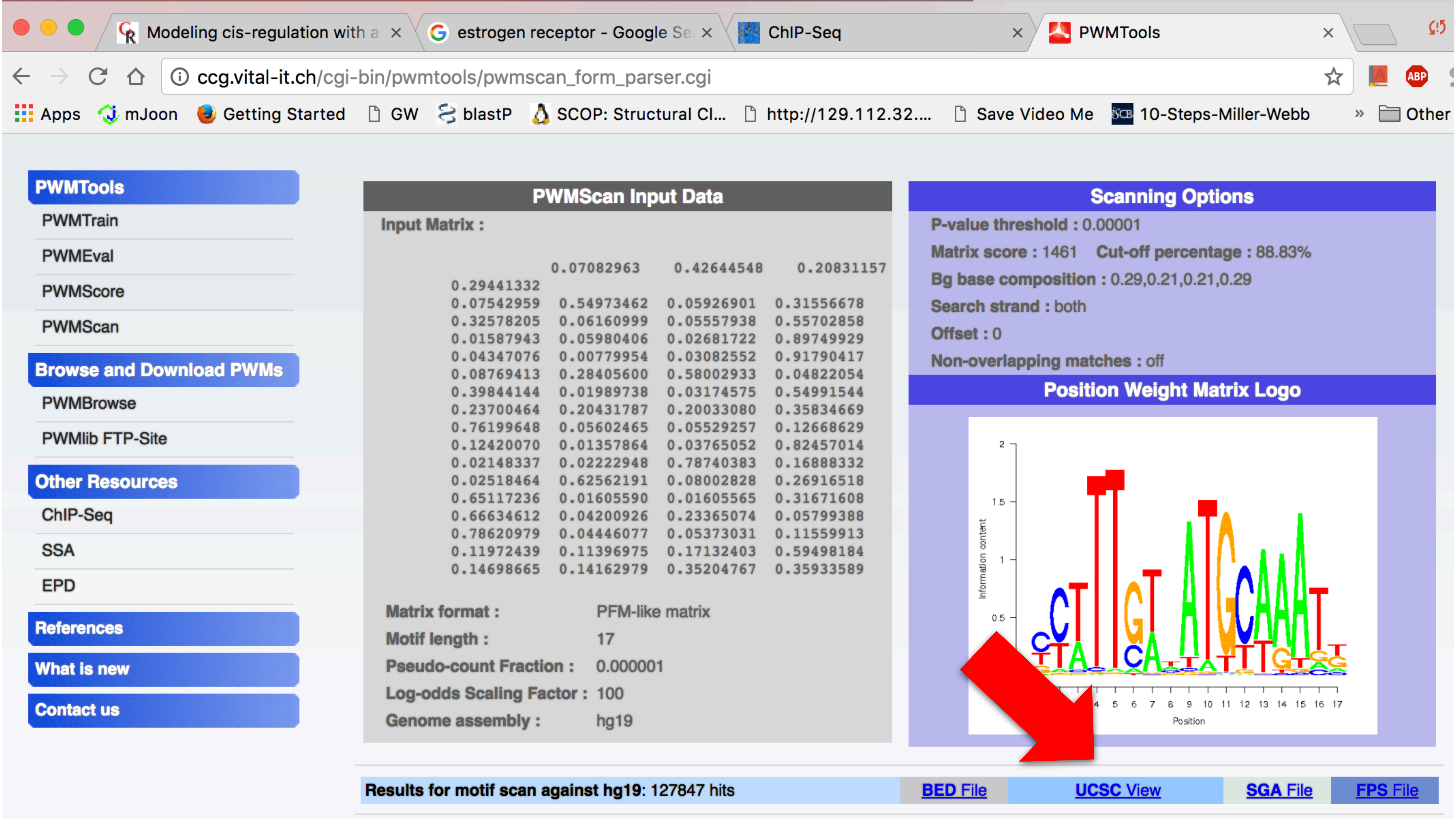
○ **Or Upload** (text File containing the matrix)



MIXL1_HUMAN.H10MO.D	MLXPL_HUMAN.H10MO.D	MLX_HUMAN.H10MO.D	MNT_HUMAN.H10MO.D
MSX1_HUMAN.H10MO.D	MSX2_HUMAN.H10MO.D	MTF1_HUMAN.H10MO.C	MUSC_HUMAN.H10MO.D
MYBB_HUMAN.H10MO.D	MYB_HUMAN.H10MO.C	MYCN_HUMAN.H10MO.B	MYC_HUMAN.H10MO.A
MYOD1_HUMAN.H10MO.C	MYOG_HUMAN.H10MO.D	MZF1_HUMAN.H10MO.D	NANOG_HUMAN.H10MO.A
NDF1_HUMAN.H10MO.C	NDF2_HUMAN.H10MO.D	NF2L1_HUMAN.H10MO.	NF2L2_HUMAN.H10MO.D
NFAC1_HUMAN.H10MO.S	NFAC2_HUMAN.H10MO.B	NFAC3_HUMAN.H10MO.B	NFAC4_HUMAN.H10MO.C
NFE2_HUMAN.H10MO.B	NFIA_HUMAN.H10MO.C	NFIA_HUMAN.H10MO.S	NFIC_HUMAN.H10MO.A
NFKB1_HUMAN.H10MO.B	NFKB2_HUMAN.H10MO.D	NFYA_HUMAN.H10MO.A	NFYB_HUMAN.H10MO.A
NGN2_HUMAN.H10MO.D	NKX21_HUMAN.H10MO.D	NKX22_HUMAN.H10MO.D	NKX23_HUMAN.H10MO.D
NKX28_HUMAN.H10MO.C	NKX31_HUMAN.H10MO.C	NKX32_HUMAN.H10MO.C	NKX61_HUMAN.H10MO.D
NOBOX_HUMAN.H10MO.C	NOTO_HUMAN.H10MO.D	NR0B1_HUMAN.H10MO.D	NR1D1_HUMAN.H10MO.C
NR1H4_HUMAN.H10MO.C	NR1I2_HUMAN.H10MO.C	NR1I2_HUMAN.H10MO.S	NR1I3_HUMAN.H10MO.C
NR2C1_HUMAN.H10MO.C	NR2C2_HUMAN.H10MO.A	NR2E1_HUMAN.H10MO.D	NR2E3_HUMAN.H10MO.C
NR4A1_HUMAN.H10MO.C	NR4A2_HUMAN.H10MO.C	NR4A3_HUMAN.H10MO.D	NR5A2_HUMAN.H10MO.C
NRF1_HUMAN.H10MO.A	NRL_HUMAN.H10MO.D	OLIG1_HUMAN.H10MO.D	OLIG2_HUMAN.H10MO.D
ONEC2_HUMAN.H10MO.D	ONEC3_HUMAN.H10MO.D	OTX1_HUMAN.H10MO.D	OTX2_HUMAN.H10MO.C



MLX_HUMAN.H10MO.D	MNT_HUMAN.H10MO.D
MTF1_HUMAN.H10MO.C	MUSC_HUMAN.H10MO.D
MYCN_HUMAN.H10MO.B	MYC_HUMAN.H10MO.A
MZF1_HUMAN.H10MO.D	NANOG_HUMAN.H10MO.A
NF2L1_HUMAN.H10MO.C	NF2L2_HUMAN.H10MO.D
NFAC3_HUMAN.H10MO.B	NFAC4_HUMAN.H10MO.C
NFIA_HUMAN.H10MO.S	NFIC_HUMAN.H10MO.A
NFYA_HUMAN.H10MO.A	NFYB_HUMAN.H10MO.A



PWMScan Input Data

Input Matrix :

```
                 0.07082963    0.42644548    0.20831157
0.29441332
0.07542959    0.54973462    0.05926901    0.31556678
0.32578205    0.06160999    0.05557938    0.55702858
0.01587943    0.05980406    0.02681722    0.89749929
0.04347076    0.00779954    0.03082552    0.91790417
0.08769413    0.28405600    0.58002933    0.04822054
0.39844144    0.01989738    0.03174575    0.54991544
0.23700464    0.20431787    0.20033080    0.35834669
0.76199648    0.05602465    0.05529257    0.12668629
0.12420070    0.01357864    0.03765052    0.82457014
0.02148337    0.02222948    0.78740383    0.16888332
0.02518464    0.62562191    0.08002828    0.26916518
0.65117236    0.01605590    0.01605565    0.31671608
0.66634612    0.04200926    0.23365074    0.05799388
0.78620979    0.04446077    0.05373031    0.11559913
0.11972439    0.11396975    0.17132403    0.59498184
0.14698665    0.14162979    0.35204767    0.35933589
```

Matrix format :	PFM-like matrix
Motif length :	17
Pseudo-count Fraction :	0.000001
Log-odds Scaling Factor :	100
Genome assembly :	hg19

Scanning Options

P-value threshold : 0.00001

Matrix score : 1461 **Cut-off percentage :** 88.83%

Bg base composition : 0.29,0.21,0.21,0.29

Search strand : both

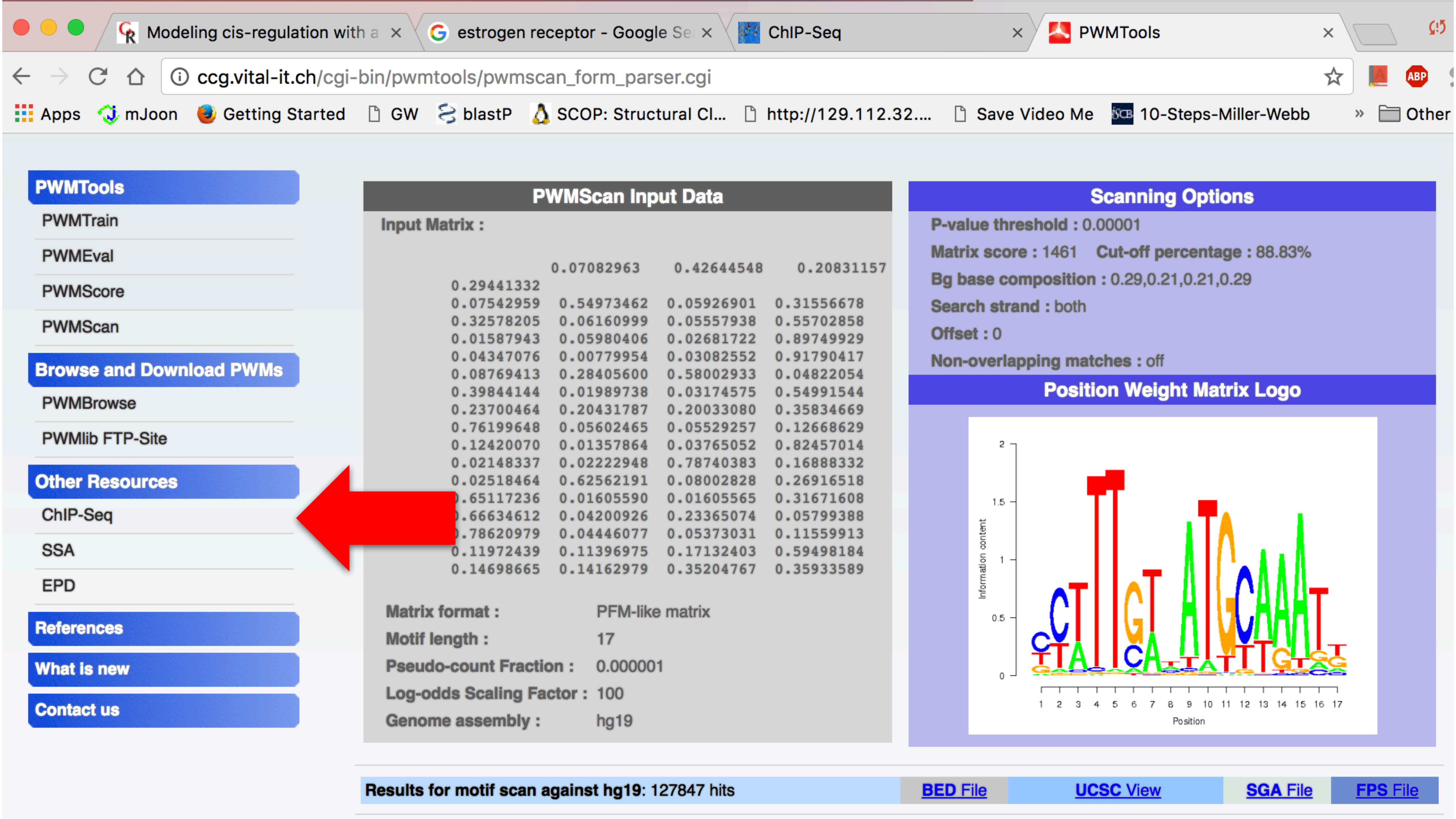
Offset : 0

Non-overlapping matches : off

Position Weight Matrix Logo

Results for motif scan against hg19: 127847 hits | BED File | UCSC View | SGA File | FPS File

Does my protein bind around TSS?



PWMScan Input Data

Input Matrix :

```
                0.07082963    0.42644548    0.20831157
  0.29441332
  0.07542959   0.54973462   0.05926901   0.31556678
  0.32578205   0.06160999   0.05557938   0.55702858
  0.01587943   0.05980406   0.02681722   0.89749929
  0.04347076   0.00779954   0.03082552   0.91790417
  0.08769413   0.28405600   0.58002933   0.04822054
  0.39844144   0.01989738   0.03174575   0.54991544
  0.23700464   0.20431787   0.20033080   0.35834669
  0.76199648   0.05602465   0.05529257   0.12668629
  0.12420070   0.01357864   0.03765052   0.82457014
  0.02148337   0.02222948   0.78740383   0.16888332
  0.02518464   0.62562191   0.08002828   0.26916518
   .65117236   0.01605590   0.01605565   0.31671608
   .66634612   0.04200926   0.23365074   0.05799388
   .78620979   0.04446077   0.05373031   0.11559913
  0.11972439   0.11396975   0.17132403   0.59498184
  0.14698665   0.14162979   0.35204767   0.35933589
```

Matrix format :	PFM-like matrix
Motif length :	17
Pseudo-count Fraction :	0.000001
Log-odds Scaling Factor :	100
Genome assembly :	hg19

Scanning Options

P-value threshold : 0.00001

Matrix score : 1461 **Cut-off percentage :** 88.83%

Bg base composition : 0.29,0.21,0.21,0.29

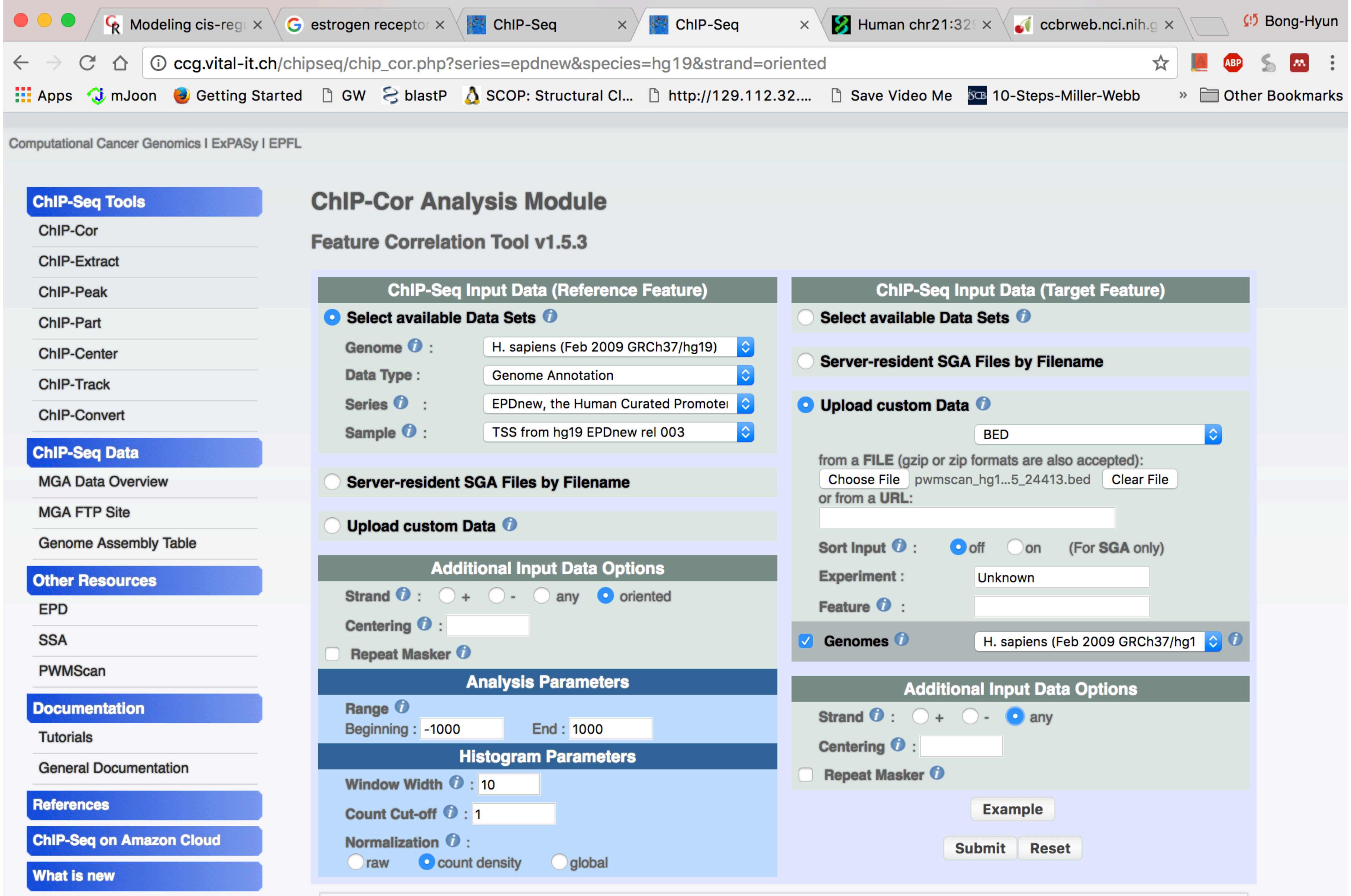
Search strand : both

Offset : 0

Non-overlapping matches : off

Position Weight Matrix Logo

Results for motif scan against hg19: 127847 hits



ccg.vital-it.ch/chipseq/chip_cor.php?series=epdnew&species=hg19&strand=oriented

Computational Cancer Genomics I ExPASy I EPFL

ChIP-Seq Tools
- ChIP-Cor
- ChIP-Extract
- ChIP-Peak
- ChIP-Part
- ChIP-Center
- ChIP-Track
- ChIP-Convert

ChIP-Seq Data
- MGA Data Overview
- MGA FTP Site
- Genome Assembly Table

Other Resources
- EPD
- SSA
- PWMScan

Documentation
- Tutorials
- General Documentation

References

ChIP-Seq on Amazon Cloud

What is new

ChIP-Cor Analysis Module

Feature Correlation Tool v1.5.3

ChIP-Seq Input Data (Reference Feature)

● Select available Data Sets

Genome : H. sapiens (Feb 2009 GRCh37/hg19)

Data Type : Genome Annotation

Series : EPDnew, the Human Curated Promoter

Sample : TSS from hg19 EPDnew rel 003

○ Server-resident SGA Files by Filename

○ Upload custom Data

Additional Input Data Options

Strand : ○ + ○ - ○ any ● oriented

Centering :

□ Repeat Masker

Analysis Parameters

Range

Beginning : -1000 End : 1000

Histogram Parameters

Window Width : 10

Count Cut-off : 1

Normalization :

○ raw ● count density ○ global

ChIP-Seq Input Data (Target Feature)

○ Select available Data Sets

○ Server-resident SGA Files by Filename

● Upload custom Data

BED

from a FILE (gzip or zip formats are also accepted):

Choose File pwmscan_hg1...5_24413.bed Clear File

or from a URL:

Sort Input : ● off ○ on (For SGA only)

Experiment : Unknown

Feature :

☑ Genomes H. sapiens (Feb 2009 GRCh37/hg1

Additional Input Data Options

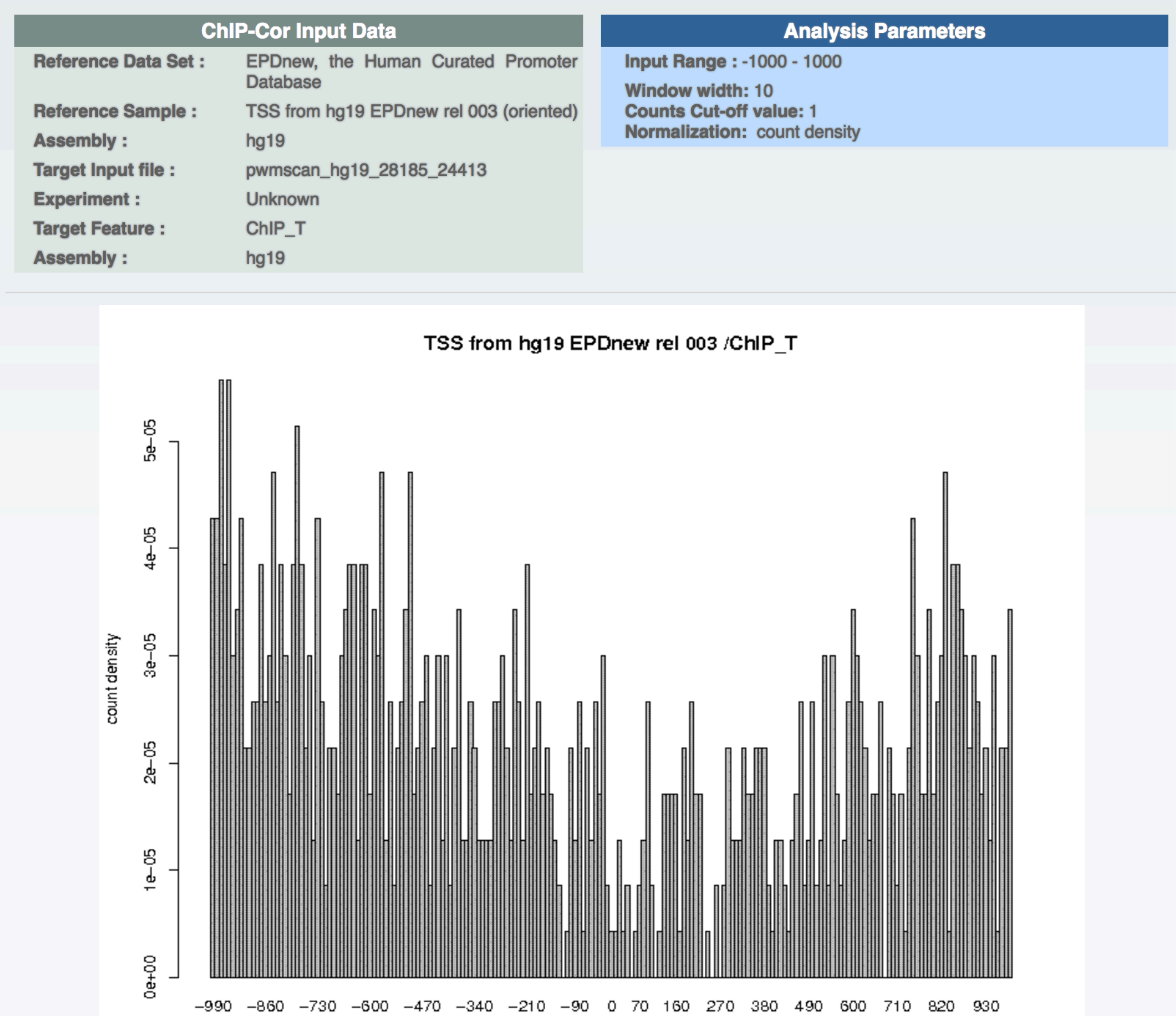
Strand : ○ + ○ - ● any

Centering :

□ Repeat Masker

Example

Submit Reset



ChIP-Cor Input Data

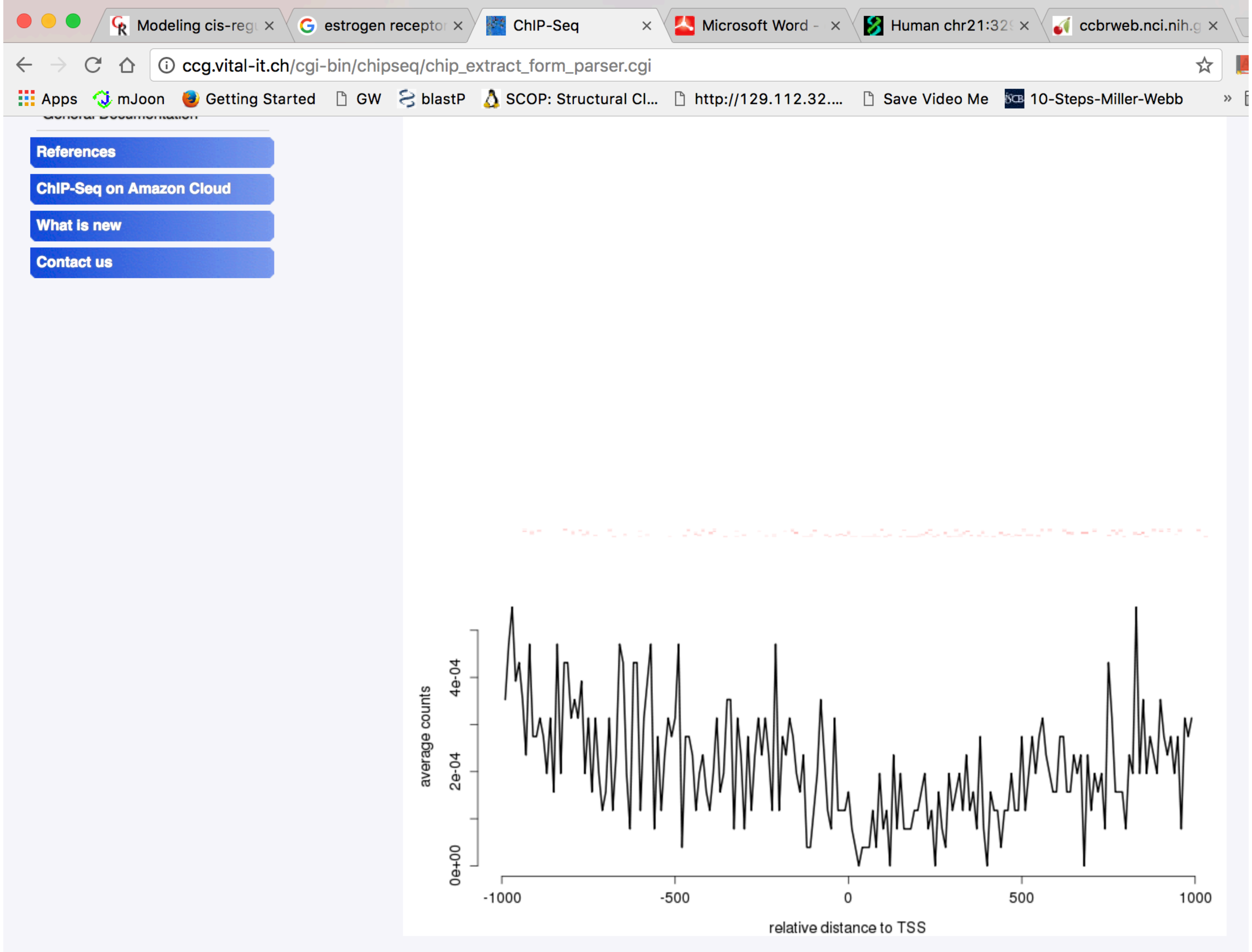
Reference Data Set :	EPDnew, the Human Curated Promoter Database
Reference Sample :	TSS from hg19 EPDnew rel 003 (oriented)
Assembly :	hg19
Target Input file :	pwmscan_hg19_28185_24413
Experiment :	Unknown
Target Feature :	ChIP_T
Assembly :	hg19

Analysis Parameters

Input Range : -1000 - 1000

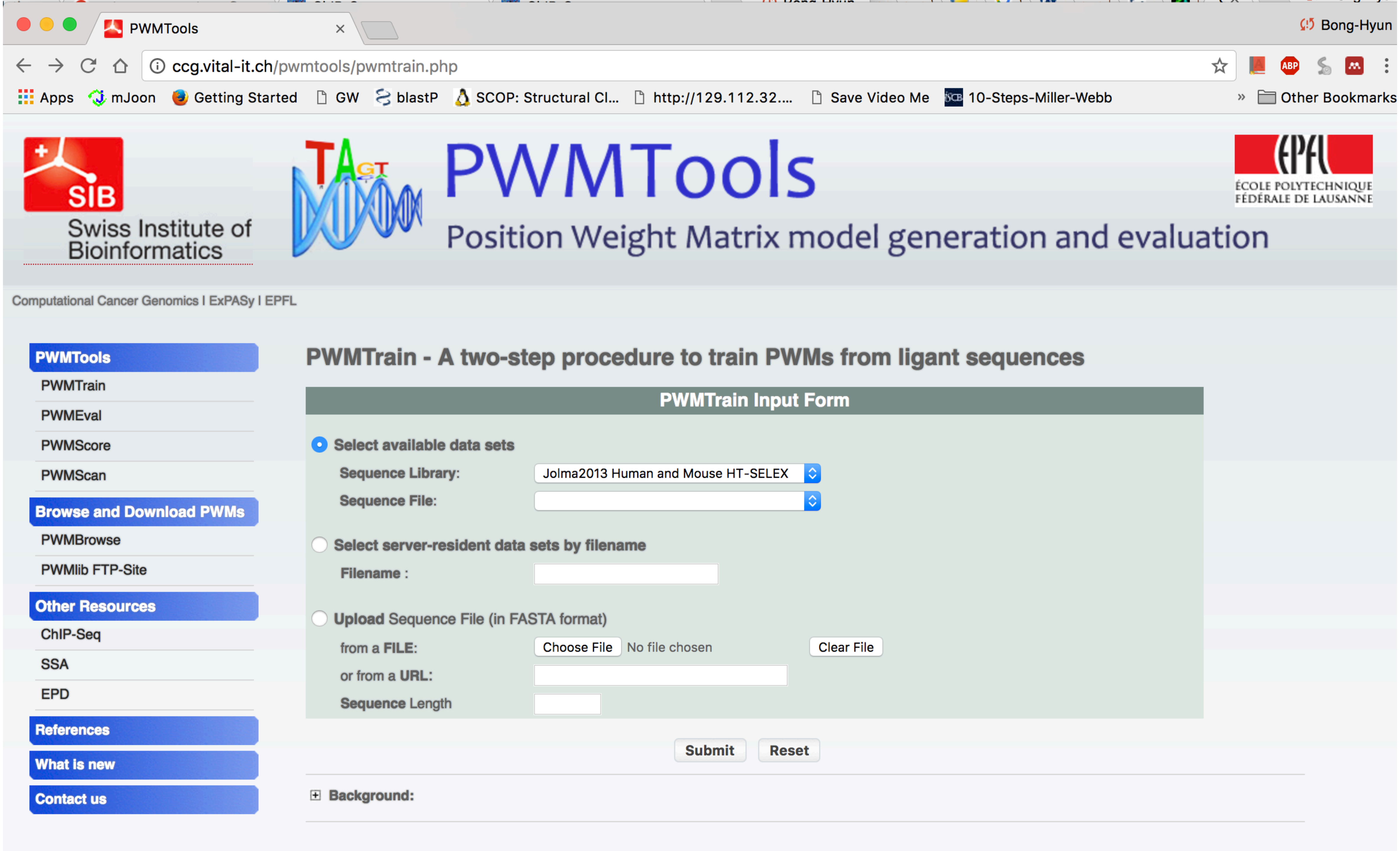
Window width: 10
Counts Cut-off value: 1
Normalization: count density

TSS from hg19 EPDnew rel 003 /ChIP_T



Modeling cis-reg
estrogen recepto
ChIP-Seq
Microsoft Word -
Human chr21:32
ccbrweb.nci.nih.g
ccg.vital-it.ch/cgi-bin/chipseq/chip_extract_form_parser.cgi
Apps
mJoon
Getting Started
GW
blastP
SCOP: Structural Cl...
http://129.112.32....
Save Video Me
10-Steps-Miller-Webb
General Documentation
References
ChIP-Seq on Amazon Cloud
What is new
Contact us
average counts
0e+00
2e-04
4e-04
-1000
-500
0
500
1000
relative distance to TSS

When you have sequences, and to find motifs



PWMTools
ccg.vital-it.ch/pwmtools/pwmtrain.php
Apps mJoon Getting Started GW blastP SCOP: Structural Cl... http://129.112.32.... Save Video Me 10-Steps-Miller-Webb Other Bookmarks
Bong-Hyun
Swiss Institute of Bioinformatics
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

PWMTools

Position Weight Matrix model generation and evaluation

Computational Cancer Genomics I ExPASy I EPFL

PWMTools
PWMTrain
PWMEval
PWMScore
PWMScan
Browse and Download PWMs
PWMBrowse
PWMlib FTP-Site
Other Resources
ChIP-Seq
SSA
EPD
References
What is new
Contact us

PWMTrain - A two-step procedure to train PWMs from ligant sequences

PWMTrain Input Form

● **Select available data sets**

Sequence Library:	Jolma2013 Human and Mouse HT-SELEX
Sequence File:	

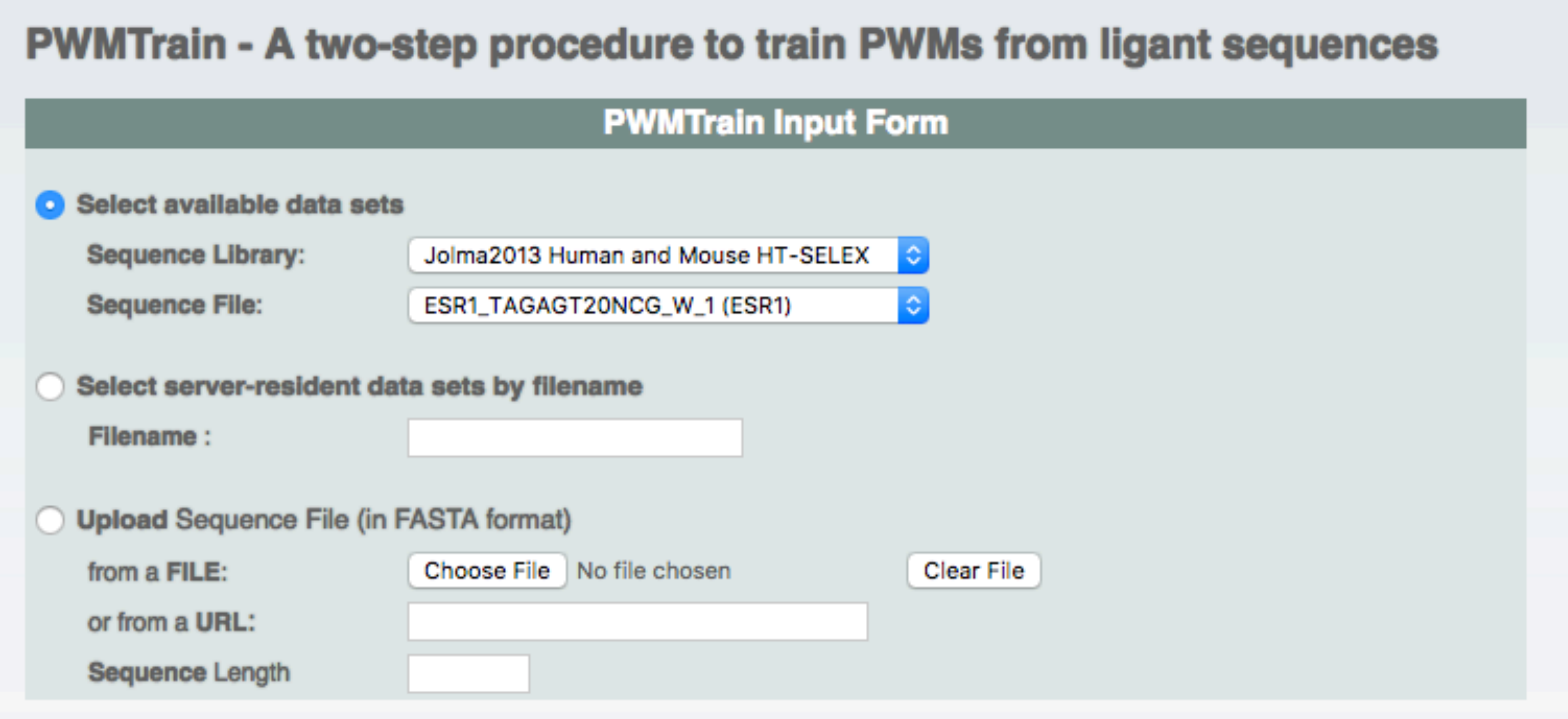
○ **Select server-resident data sets by filename**

Filename :

○ **Upload** Sequence File (in FASTA format)

from a **FILE:**	Choose File No file chosen Clear File
or from a **URL:**	
Sequence Length	

⊞ **Background:**



PWMTrain - A two-step procedure to train PWMs from ligant sequences

PWMTrain Input Form

● Select available data sets

Sequence Library: Jolma2013 Human and Mouse HT-SELEX

Sequence File: ESR1_TAGAGT20NCG_W_1 (ESR1)

○ Select server-resident data sets by filename

Filename :

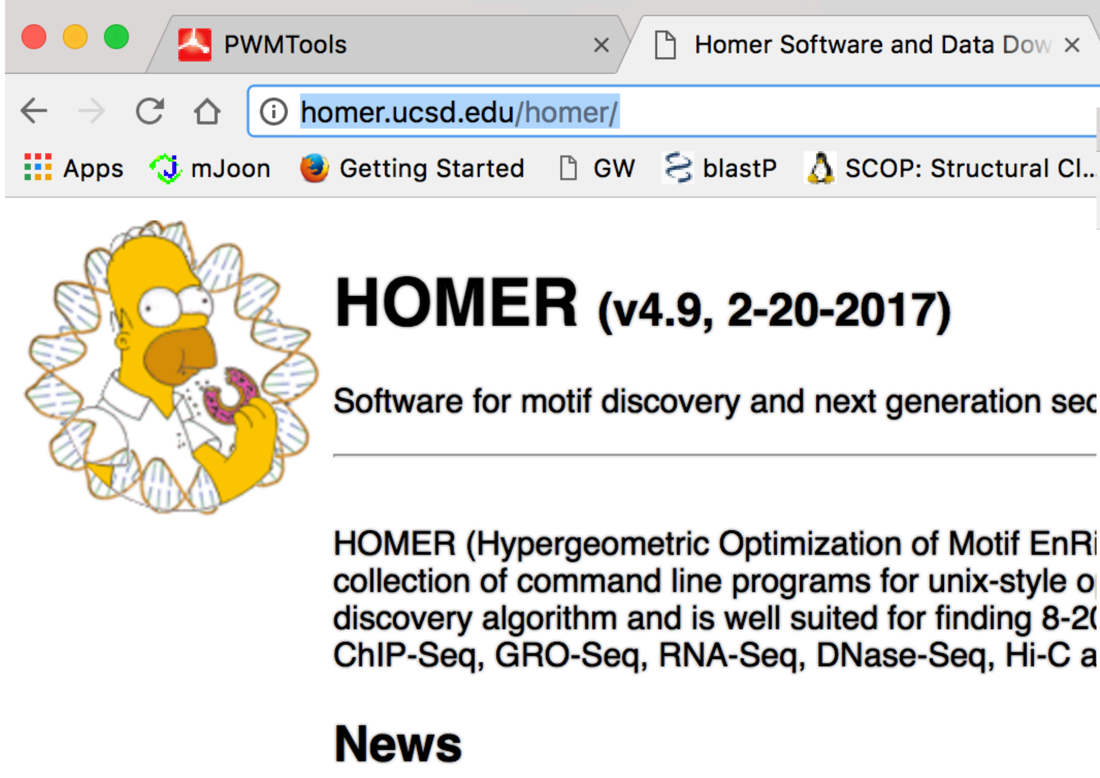
○ Upload Sequence File (in FASTA format)

from a FILE: Choose File No file chosen Clear File

or from a URL:

Sequence Length

Other motif related tools



PWMTools | Homer Software and Data Dow

homer.ucsd.edu/homer/

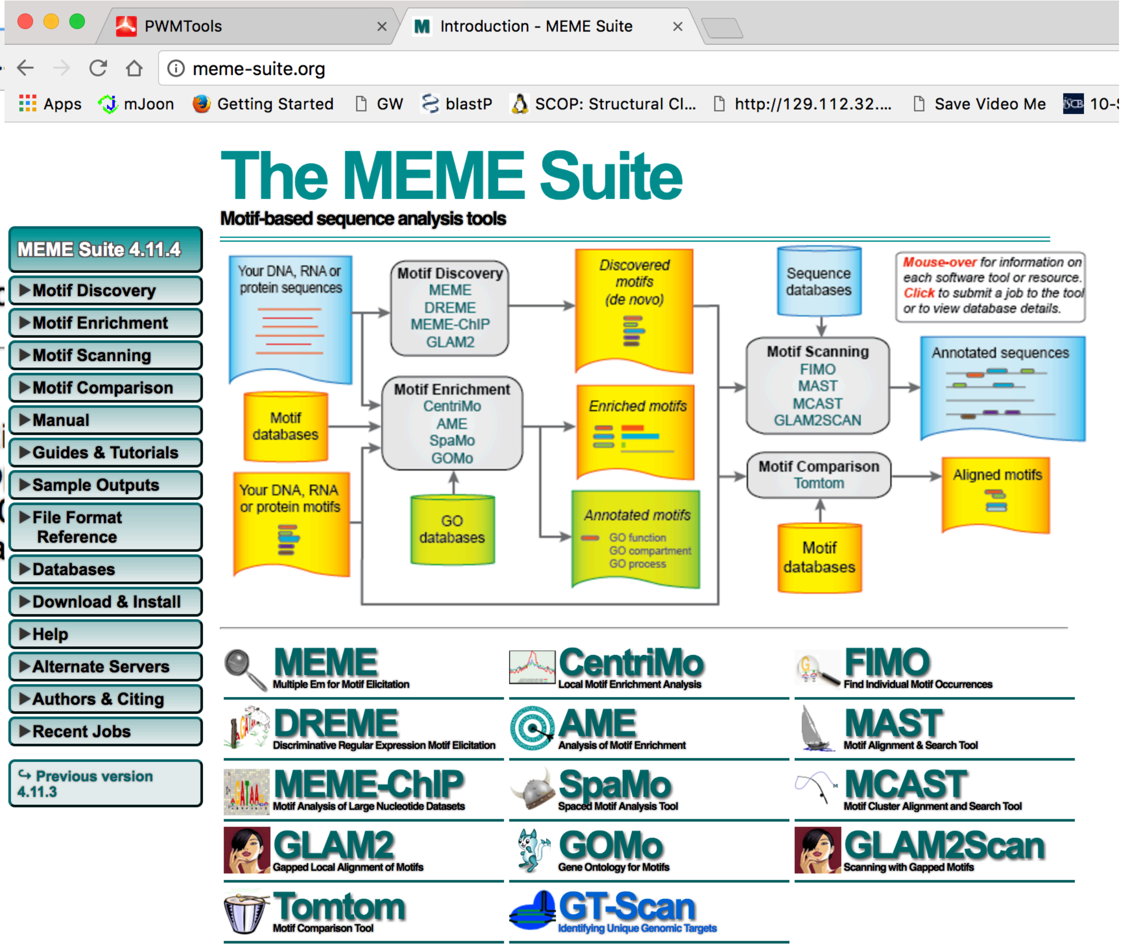
Apps | mJoon | Getting Started | GW | blastP | SCOP: Structural Cl...

HOMER (v4.9, 2-20-2017)

Software for motif discovery and next generation sec

HOMER (Hypergeometric Optimization of Motif EnRi
collection of command line programs for unix-style o
discovery algorithm and is well suited for finding 8-2(
ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C a

News



PWMTools | Introduction - MEME Suite

meme-suite.org

Apps | mJoon | Getting Started | GW | blastP | SCOP: Structural Cl... | http://129.112.32.... | Save Video Me | 10-

The MEME Suite

Motif-based sequence analysis tools

- MEME Suite 4.11.4
- ▶Motif Discovery
- ▶Motif Enrichment
- ▶Motif Scanning
- ▶Motif Comparison
- ▶Manual
- ▶Guides & Tutorials
- ▶Sample Outputs
- ▶File Format Reference
- ▶Databases
- ▶Download & Install
- ▶Help
- ▶Alternate Servers
- ▶Authors & Citing
- ▶Recent Jobs
- Previous version 4.11.3

Your DNA, RNA or protein sequences

Motif Discovery: MEME, DREME, MEME-ChIP, GLAM2

Discovered motifs (de novo)

Motif databases

Motif Enrichment: CentriMo, AME, SpaMo, GOMo

Enriched motifs

Your DNA, RNA or protein motifs

GO databases

Annotated motifs: GO function, GO compartment, GO process

Sequence databases

Motif Scanning: FIMO, MAST, MCAST, GLAM2SCAN

Annotated sequences

Motif Comparison: Tomtom

Aligned motifs

Motif databases

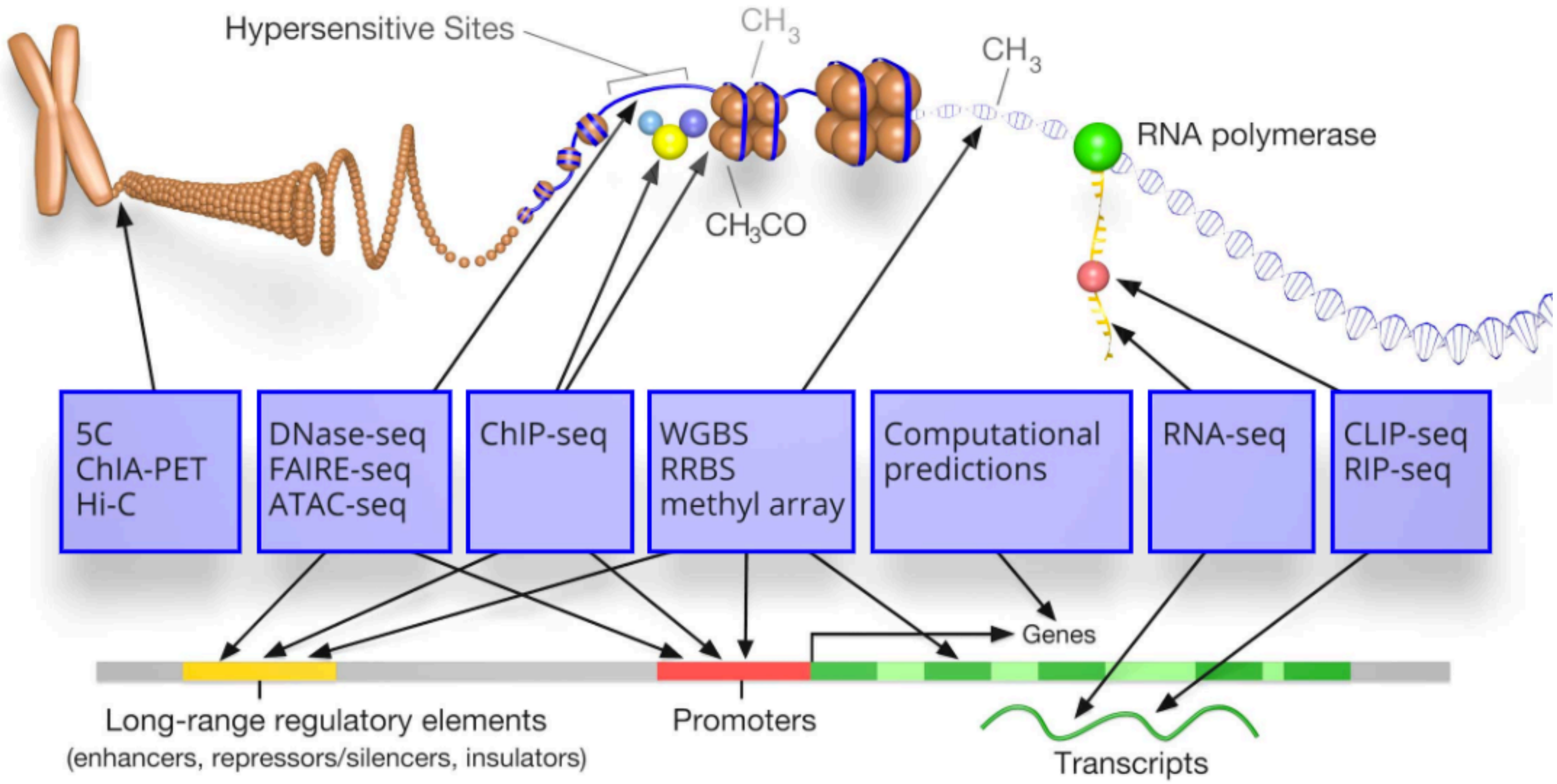
Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.

- MEME — Multiple Em for Motif Elicitation
- CentriMo — Local Motif Enrichment Analysis
- FIMO — Find Individual Motif Occurrences
- DREME — Discriminative Regular Expression Motif Elicitation
- AME — Analysis of Motif Enrichment
- MAST — Motif Alignment & Search Tool
- MEME-ChIP — Motif Analysis of Large Nucleotide Datasets
- SpaMo — Spaced Motif Analysis Tool
- MCAST — Motif Cluster Alignment and Search Tool
- GLAM2 — Gapped Local Alignment of Motifs
- GOMo — Gene Ontology for Motifs
- GLAM2Scan — Scanning with Gapped Motifs
- Tomtom — Motif Comparison Tool
- GT-Scan — Identifying Unique Genomic Targets

Observing all possible TF & DNA interaction (and something more)

https://www.encodeproject.org/

ENCODE: Encyclopedia of DNA Elements



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Goals of ENCODE

- Catalog the functional elements in model organisms such as human, mouse, fly and worm genomes
- Generate high quality data using high through-put pipelines
- Develop new technologies and analytical tools to generate, analyze and validate data
- Provide date and tools to the community in as useful form as possible

ENCODE project history

- ENCODE (Pilot phase) – 1% of human genome (2003-2007)
    - in selected cell lines
- ENCODE2 - Scale Up Phase I (2007-2012)
    - tier 1 & 2, common cell lines
- ENCODE3 –Production Phase (2012-2016)
    - tier 3 cell lines

- other ENCODE Projects:
    - Mouse ENCODE (2009-2012)
        - mouse cell line and tissue samples
    - modENCODE (2007-2012)
        - Fly tissue and Worm whole body samples

- Epigenome Roadmap Project
    - Human tissue samples
    - Raw and process data are now deposited in ENCODE DCC.

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

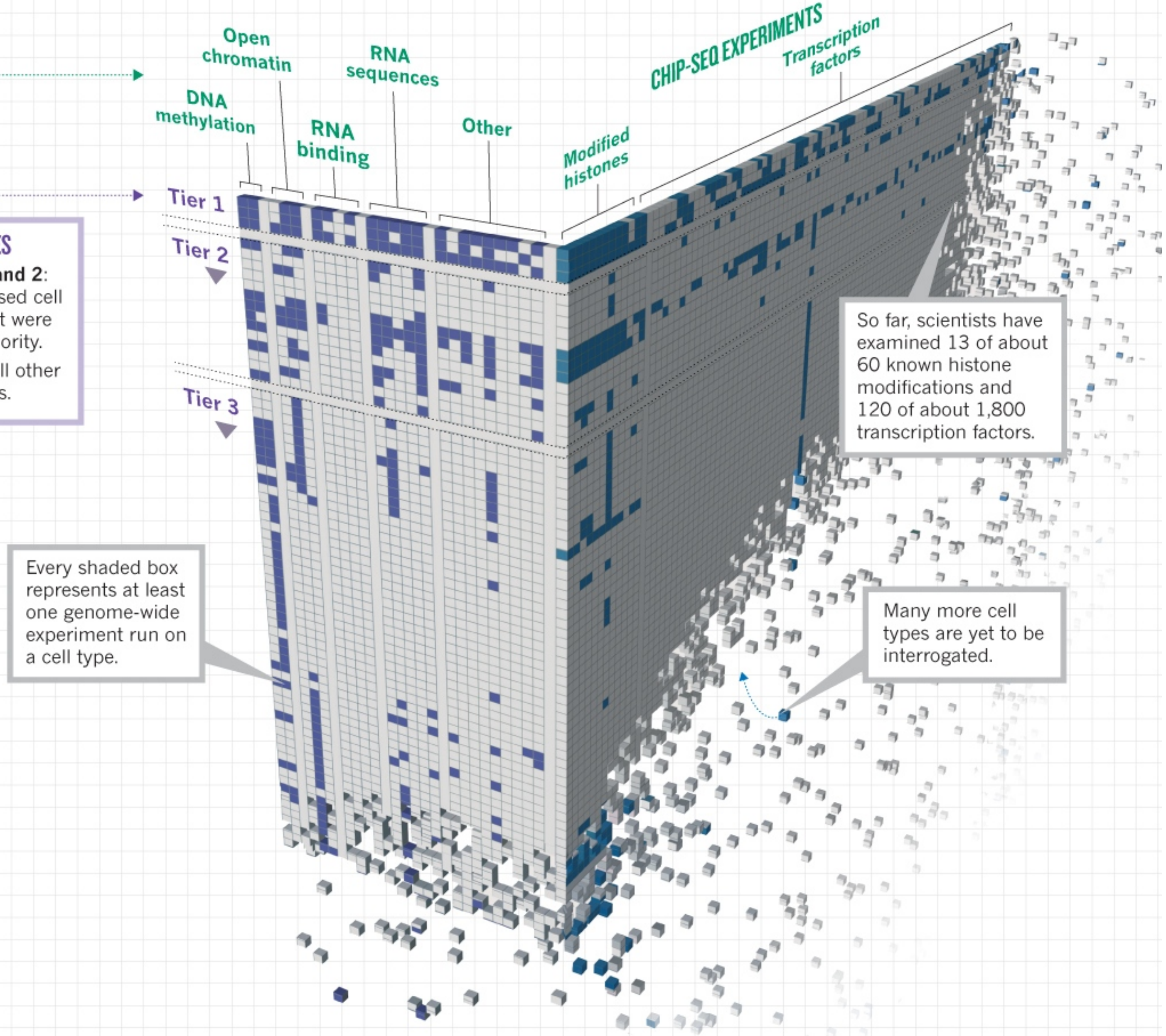
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.



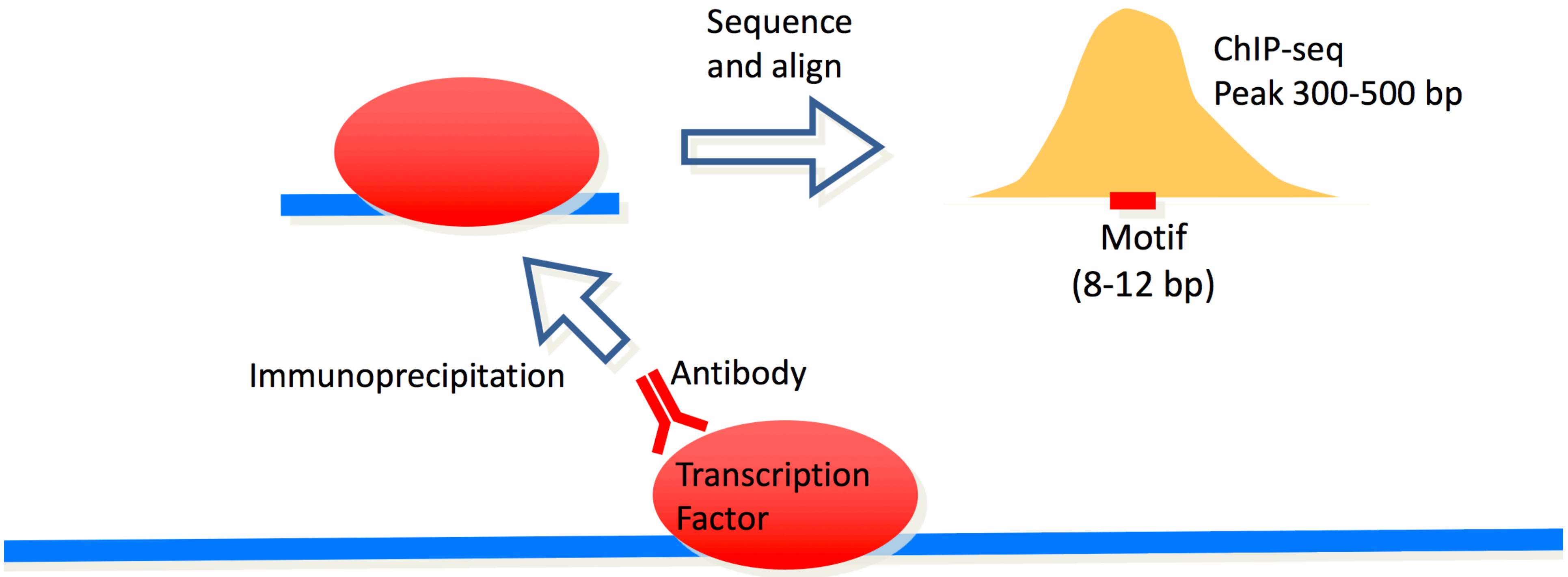
Every shaded box represents at least one genome-wide experiment run on a cell type.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Many more cell types are yet to be interrogated.

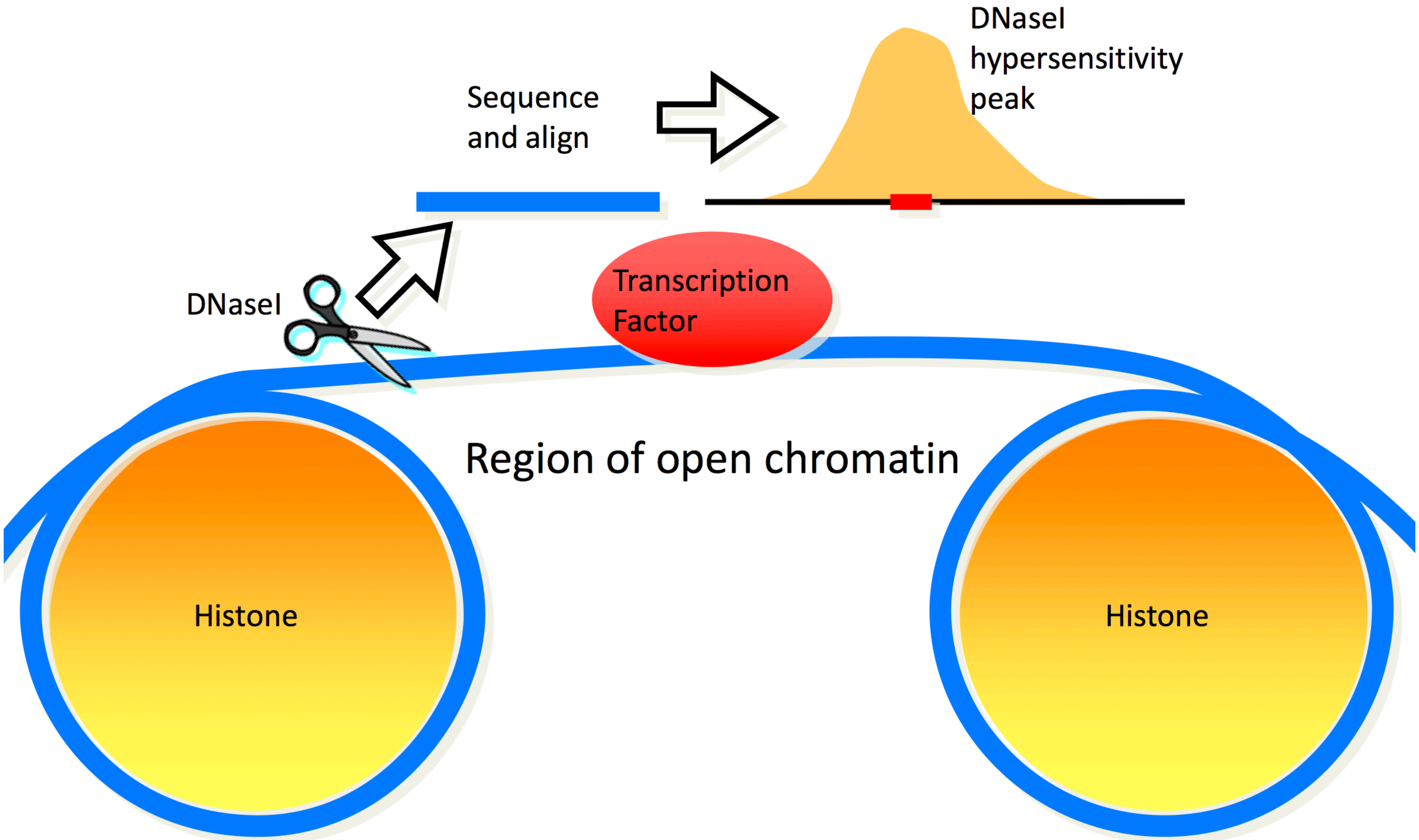
http://www.nature.com/news/encode-the-human-encyclopaedia-1.11312

Functional data: ChIP-seq

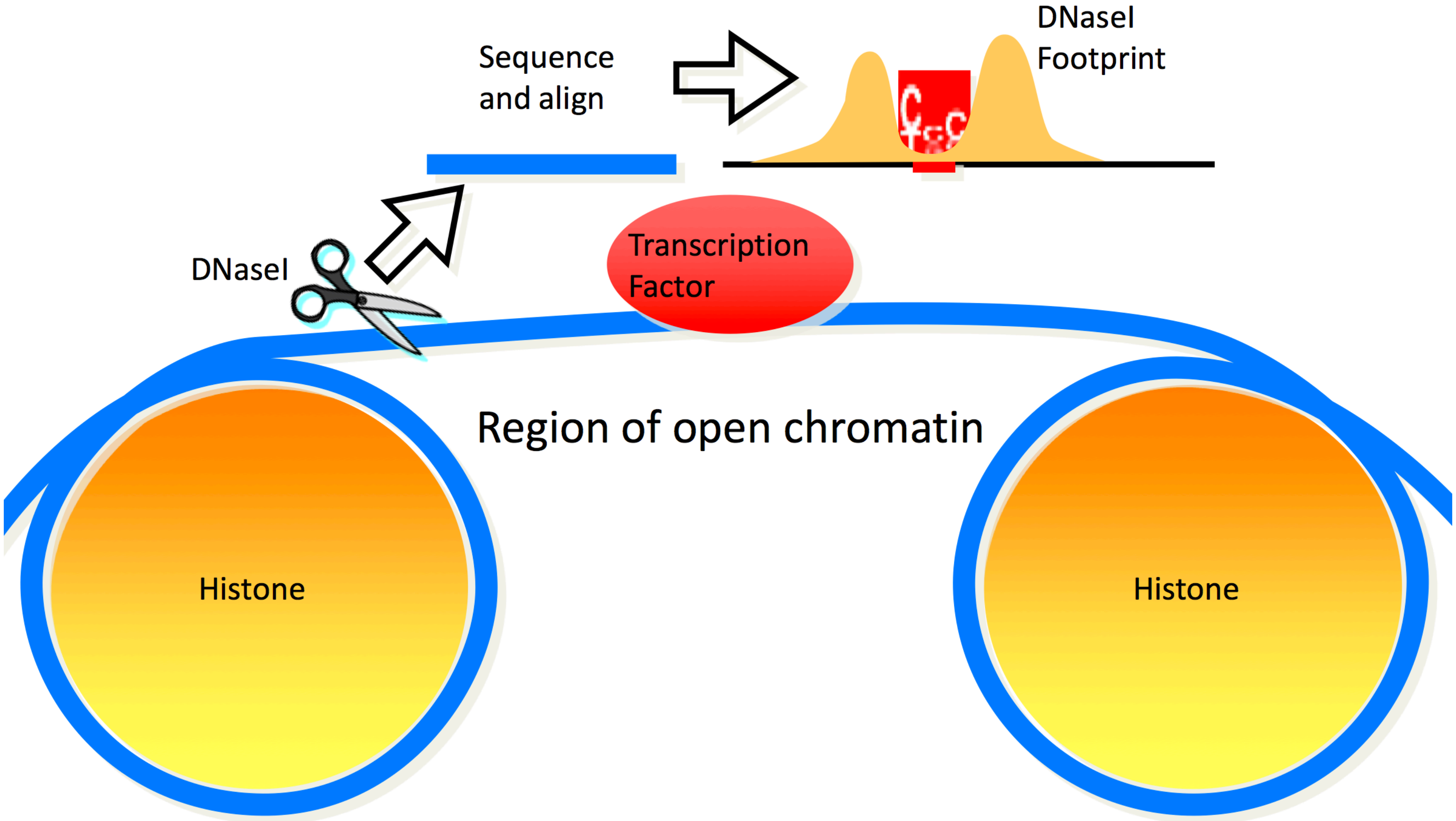


ChIP-exo
Histone Marks

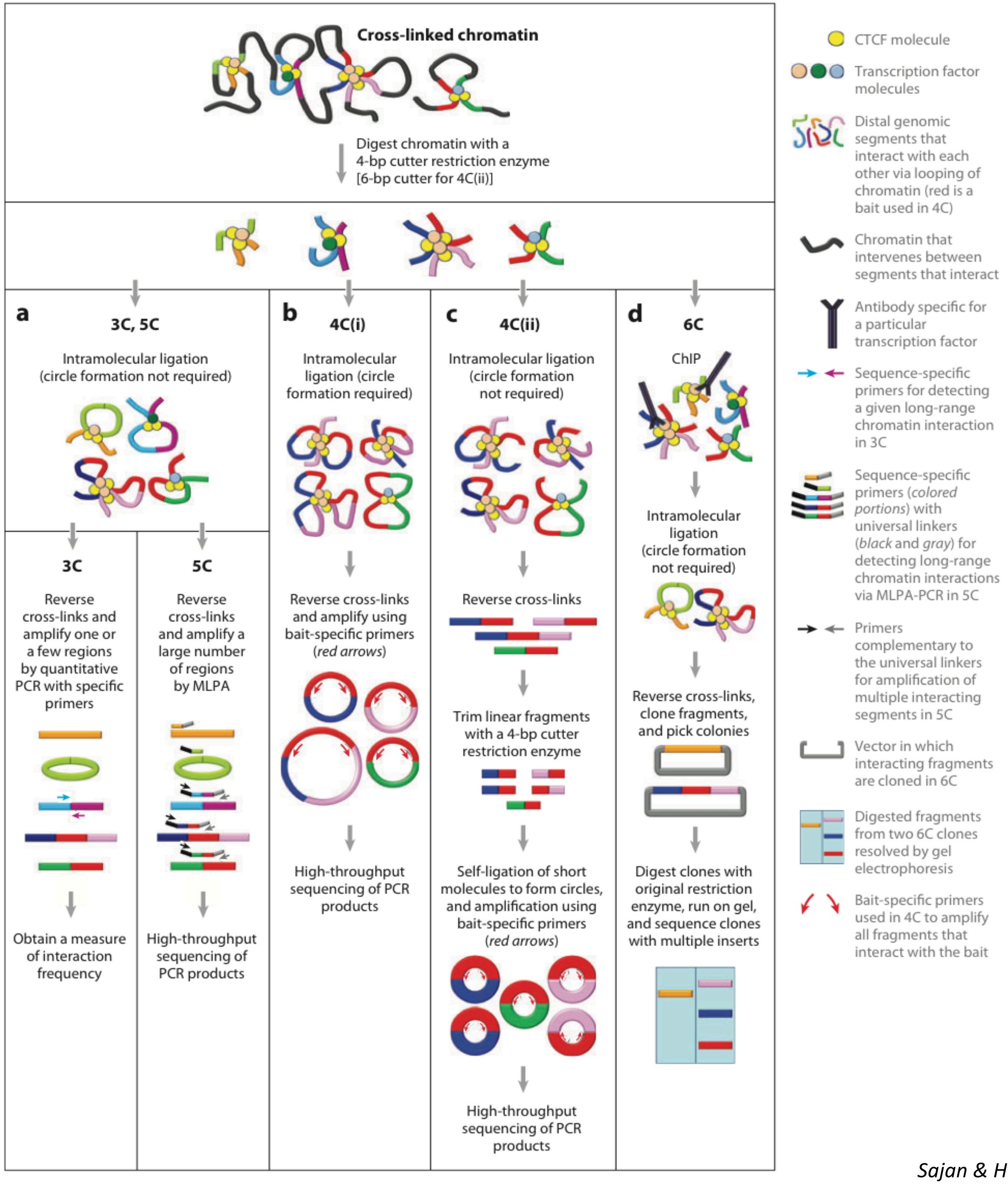
Functional data: DNase-seq



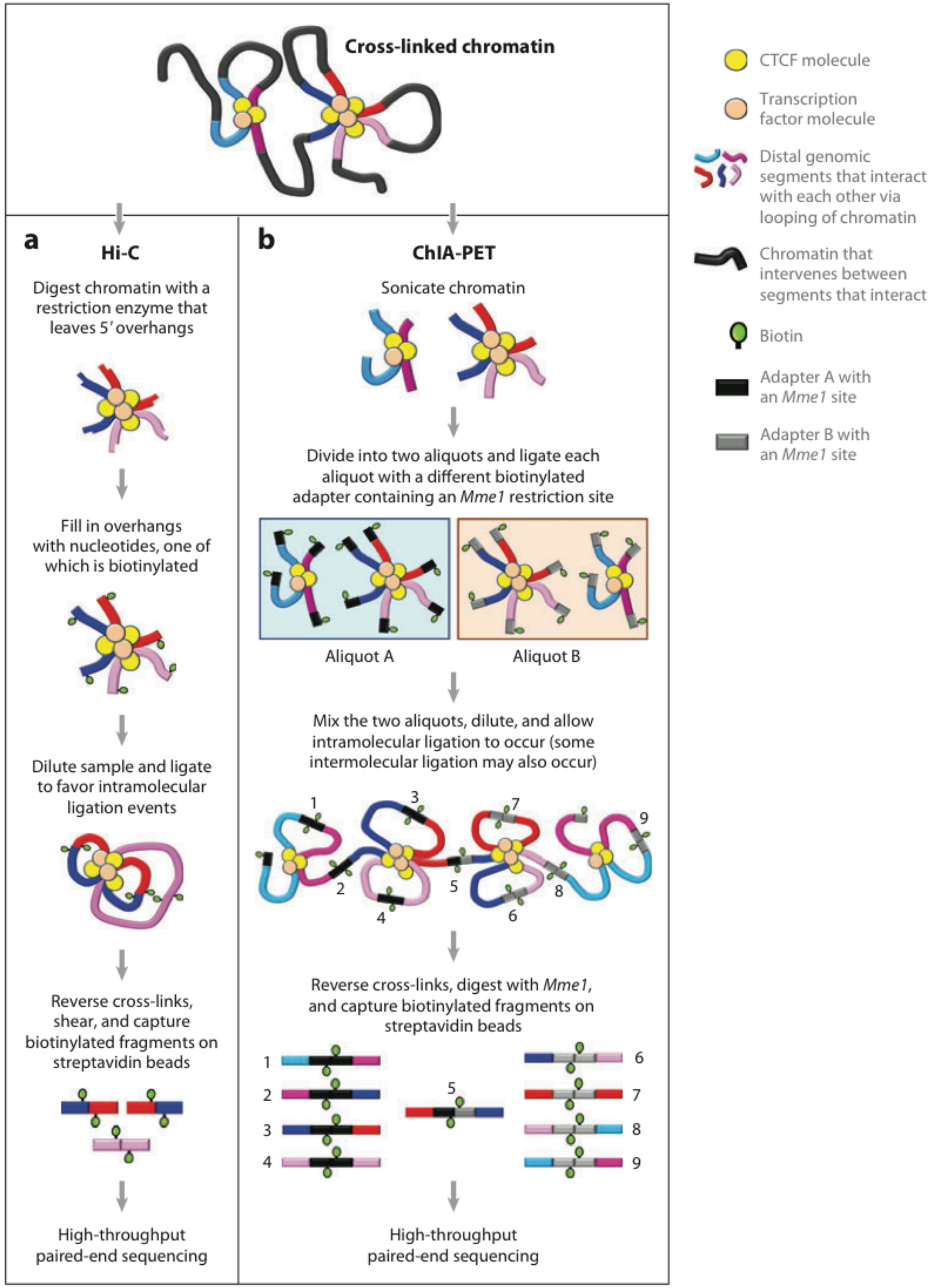
Functional data: DNase footprints



5C & and
long ts


Cross-linked chromatin
Digest chromatin with a 4-bp cutter restriction enzyme [6-bp cutter for 4C(ii)]
a
3C, 5C
Intramolecular ligation (circle formation not required)
3C
Reverse cross-links and amplify one or a few regions by quantitative PCR with specific primers
Obtain a measure of interaction frequency
5C
Reverse cross-links and amplify a large number of regions by MLPA
High-throughput sequencing of PCR products
b
4C(i)
Intramolecular ligation (circle formation required)
Reverse cross-links and amplify using bait-specific primers (red arrows)
High-throughput sequencing of PCR products
c
4C(ii)
Intramolecular ligation (circle formation not required)
Reverse cross-links
Trim linear fragments with a 4-bp cutter restriction enzyme
Self-ligation of short molecules to form circles, and amplification using bait-specific primers (red arrows)
High-throughput sequencing of PCR products
d
6C
ChIP
Intramolecular ligation (circle formation not required)
Reverse cross-links, clone fragments, and pick colonies
Digest clones with original restriction enzyme, run on gel, and sequence clones with multiple inserts
CTCF molecule
Transcription factor molecules
Distal genomic segments that interact with each other via looping of chromatin (red is a bait used in 4C)
Chromatin that intervenes between segments that interact
Antibody specific for a particular transcription factor
Sequence-specific primers for detecting a given long-range chromatin interaction in 3C
Sequence-specific primers (colored portions) with universal linkers (black and gray) for detecting long-range chromatin interactions via MLPA-PCR in 5C
Primers complementary to the universal linkers for amplification of multiple interacting segments in 5C
Vector in which interacting fragments are cloned in 6C
Digested fragments from two 6C clones resolved by gel electrophoresis
Bait-specific primers used in 4C to amplify all fragments that interact with the bait

Sajan & Hawkins (2012) Annu. Rev. Genomics Hum. Ger



Hi-C

Digest chromatin with a restriction enzyme that leaves 5′ overhangs

Fill in overhangs with nucleotides, one of which is biotinylated

Dilute sample and ligate to favor intramolecular ligation events

Reverse cross-links, shear, and capture biotinylated fragments on streptavidin beads

High-throughput paired-end sequencing

ChIA-PET

Sonicate chromatin

Divide into two aliquots and ligate each aliquot with a different biotinylated adapter containing an *Mme1* restriction site

Mix the two aliquots, dilute, and allow intramolecular ligation to occur (some intermolecular ligation may also occur)

Reverse cross-links, digest with *Mme1*, and capture biotinylated fragments on streptavidin beads

High-throughput paired-end sequencing

CTCF molecule

Transcription factor molecule

Distal genomic segments that interact with each other via looping of chromatin

Chromatin that intervenes between segments that interact

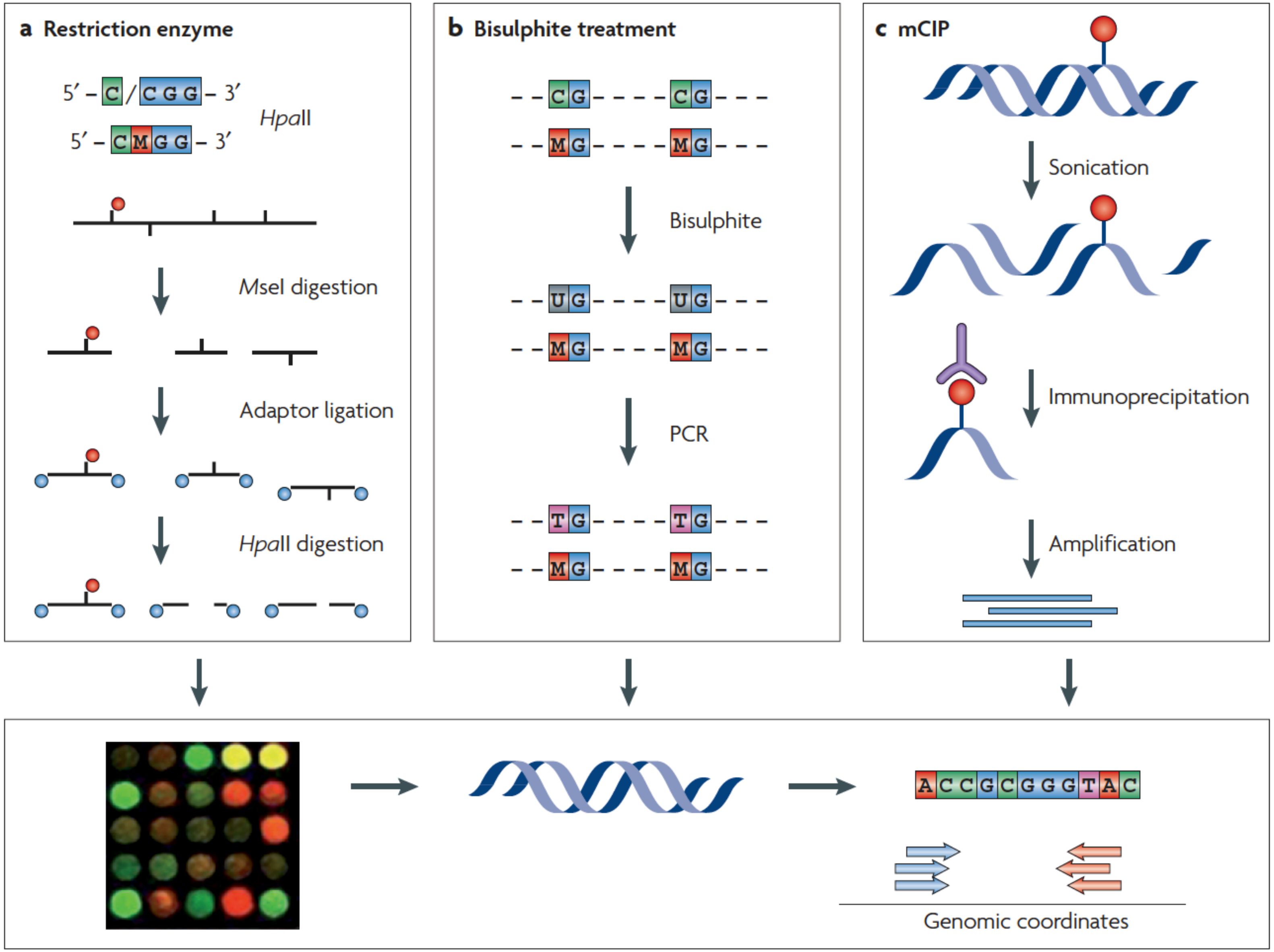
Biotin

Adapter A with an *Mme1* site

Adapter B with an *Mme1* site

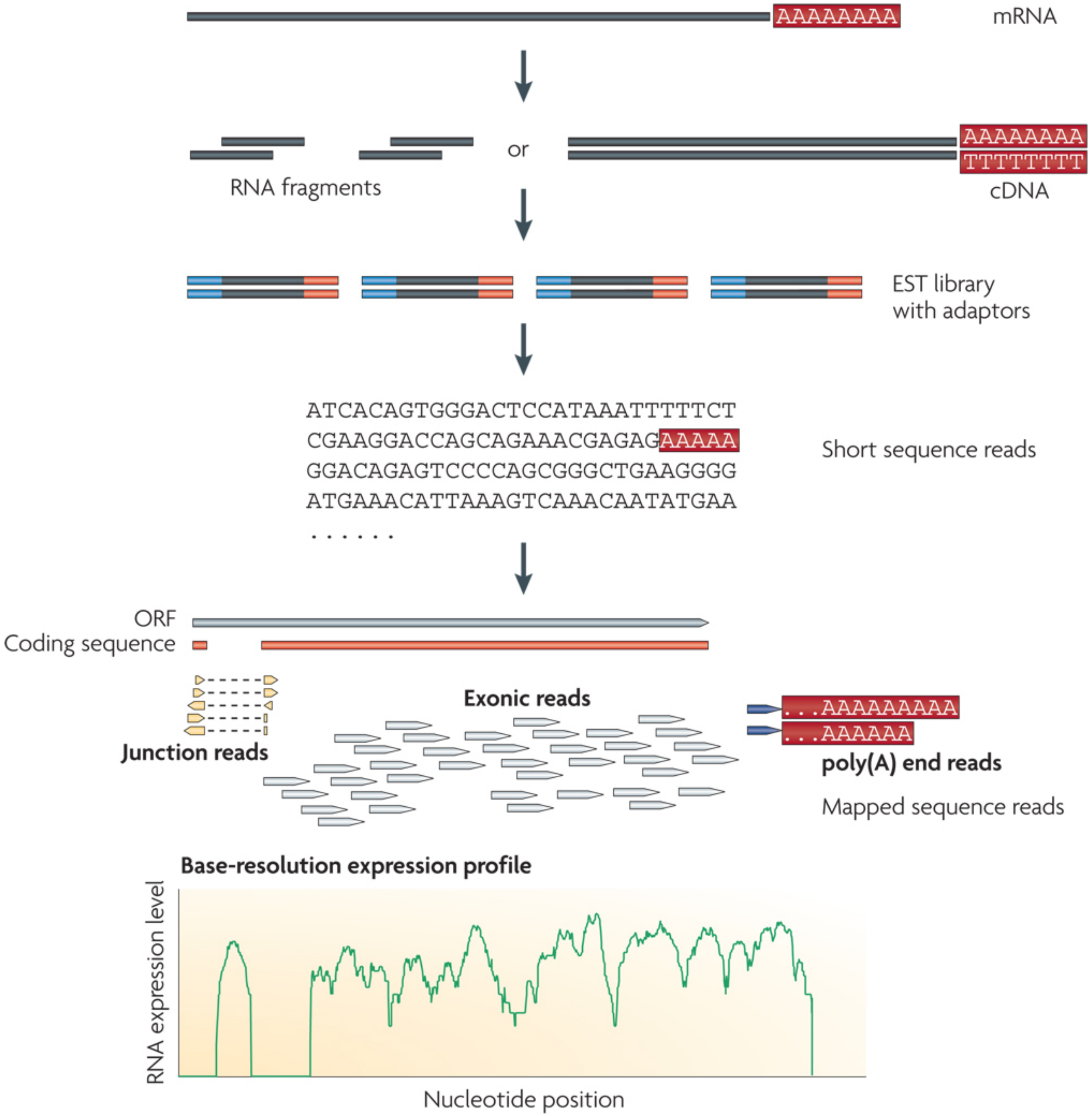
Sajan & Hawkins (2012) Annu. Rev. Genomics Hum. Gen

Measuring DNA methylation



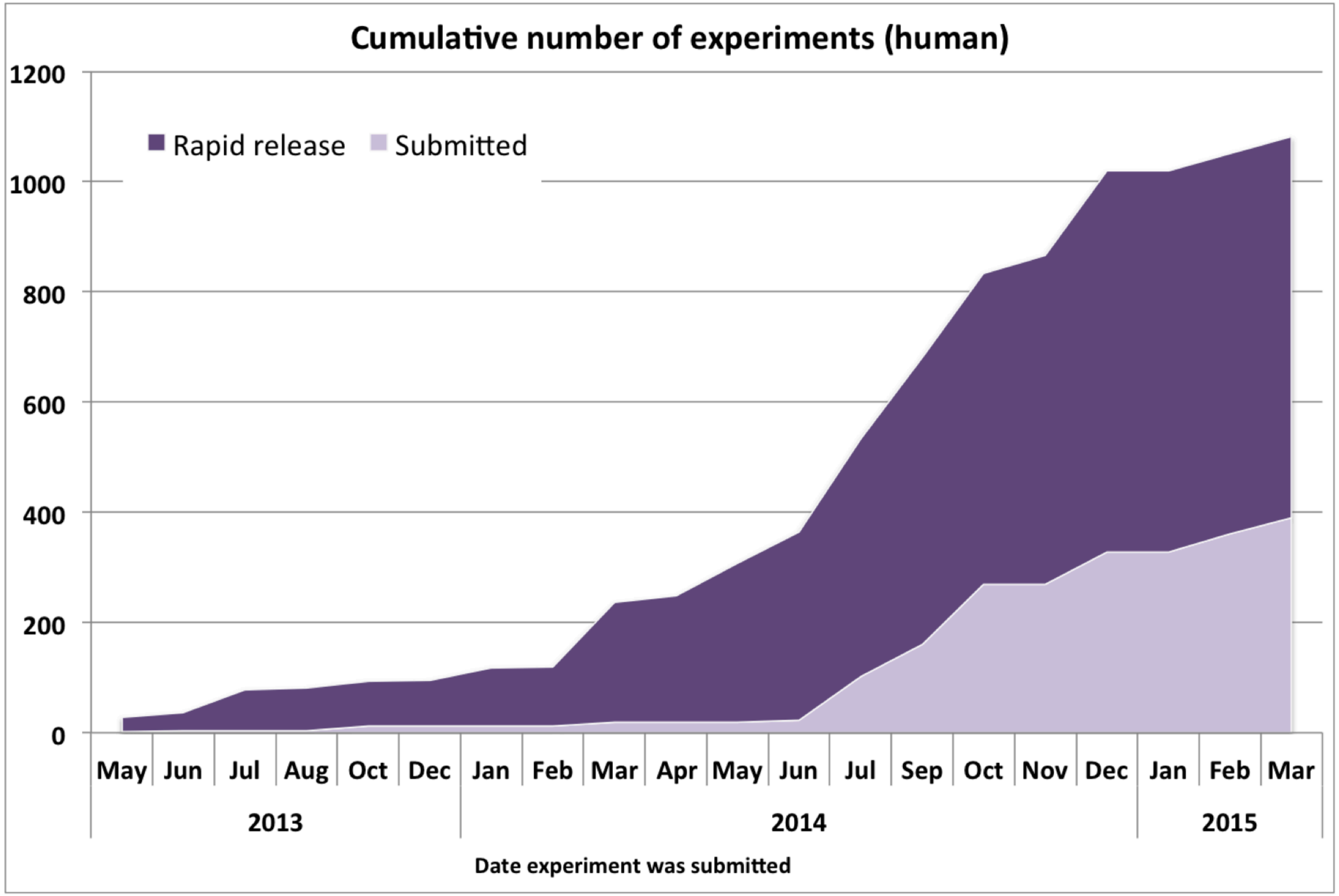
Schones & Zhao (2008) Nat. Rev. Genetics

RNA-seq



Wang, Z. et al. *RNA-Seq: a revolutionary tool for transcriptomics Nature Reviews Genetics (2009)*

Overview of Human Datasets

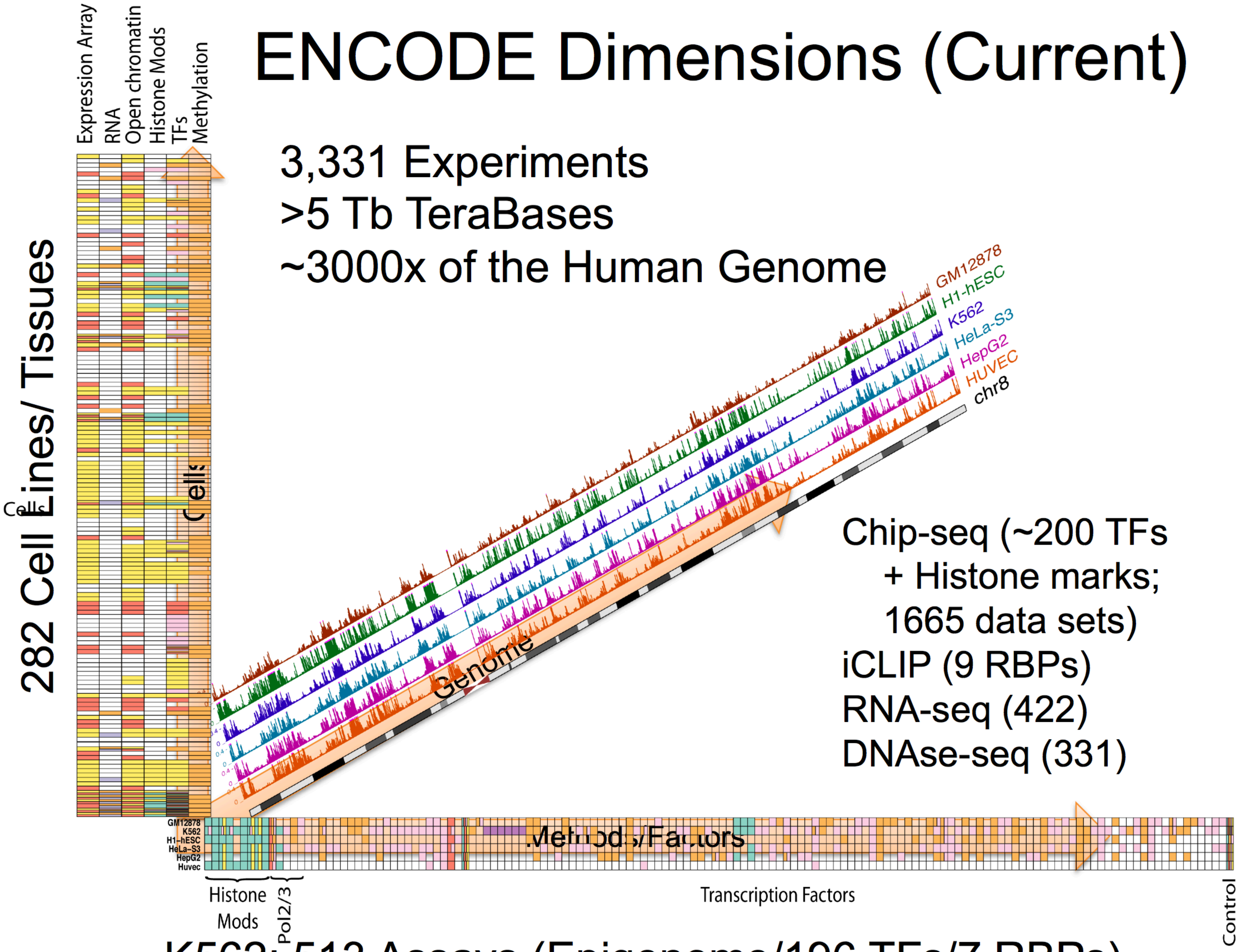


Human: 3,331 datasets submitted/released; 5,501 proposed; 8,832 total

https://www.genome.gov/pages/research/encode/ren_keystone_encode_workshop_2015.pdf

ENCODE Dimensions (Current)

3,331 Experiments
>5 Tb TeraBases
~3000x of the Human Genome



282 Cell Lines/ Tissues

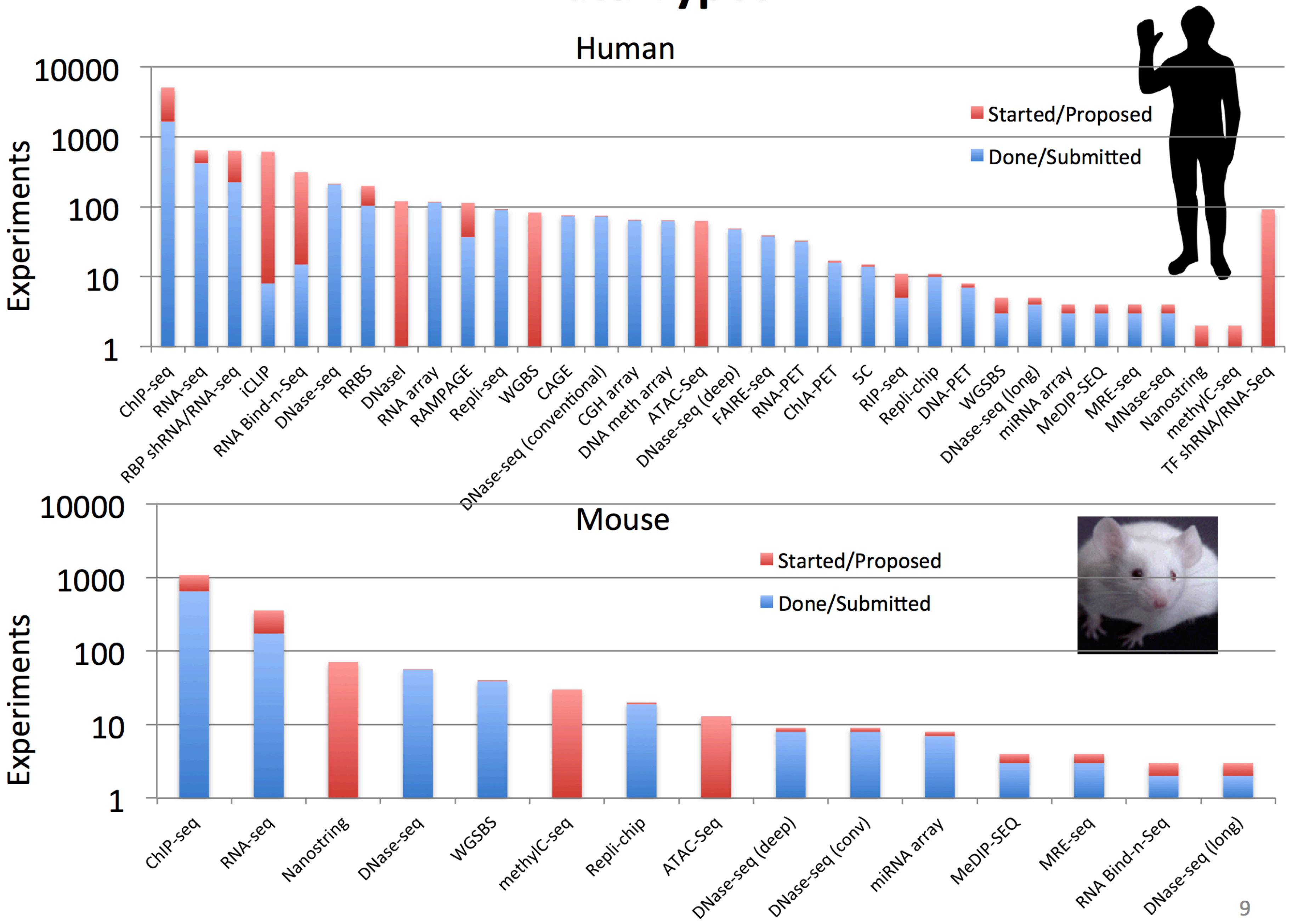
Chip-seq (~200 TFs
+ Histone marks;
1665 data sets)
iCLIP (9 RBPs)
RNA-seq (422)
DNAse-seq (331)

K562: 513 Assays (Epigenome/196 TFs/7 RBPs)

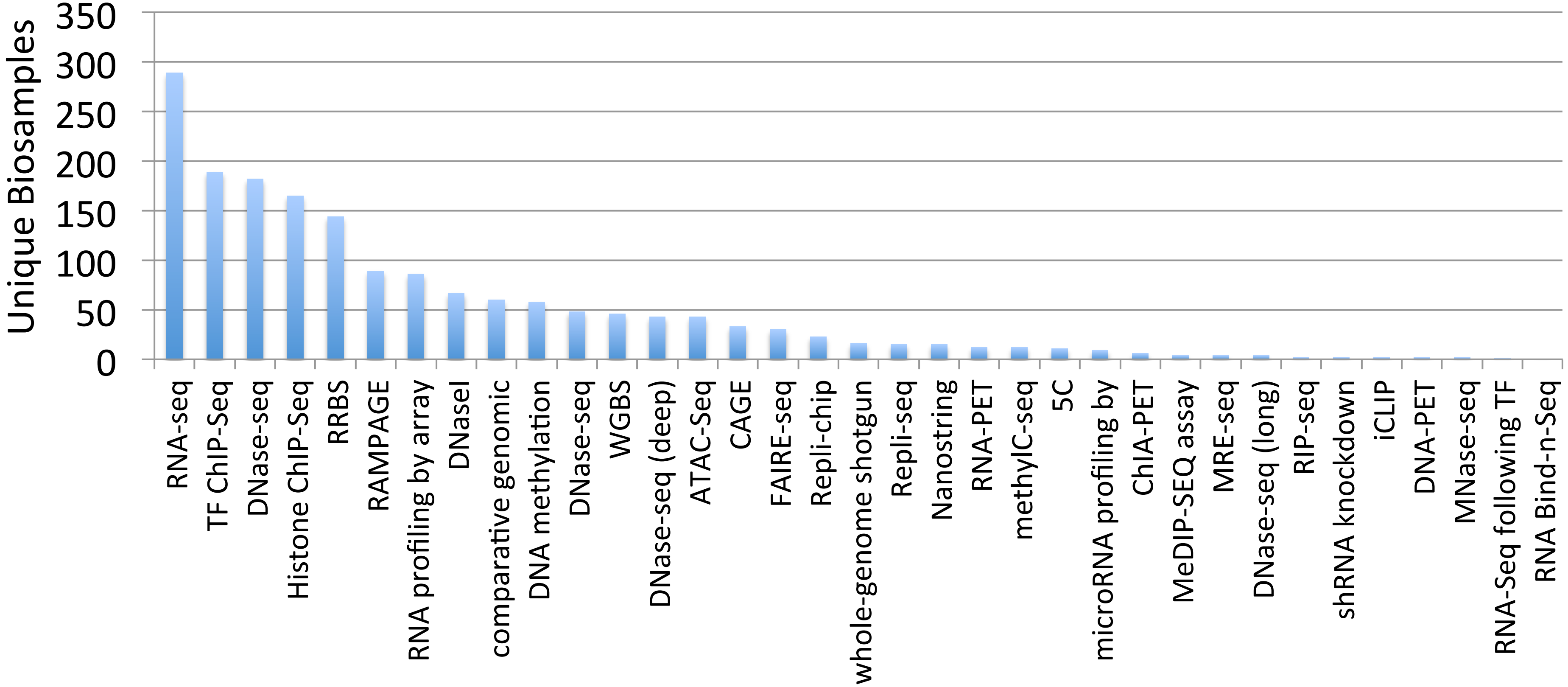
Sources of ENCODE data

- Immortalized cell lines
- Tissues
- Primary cells
- Stem cells
- In vitro differentiated cells
- Induced pluripotent stem cell line

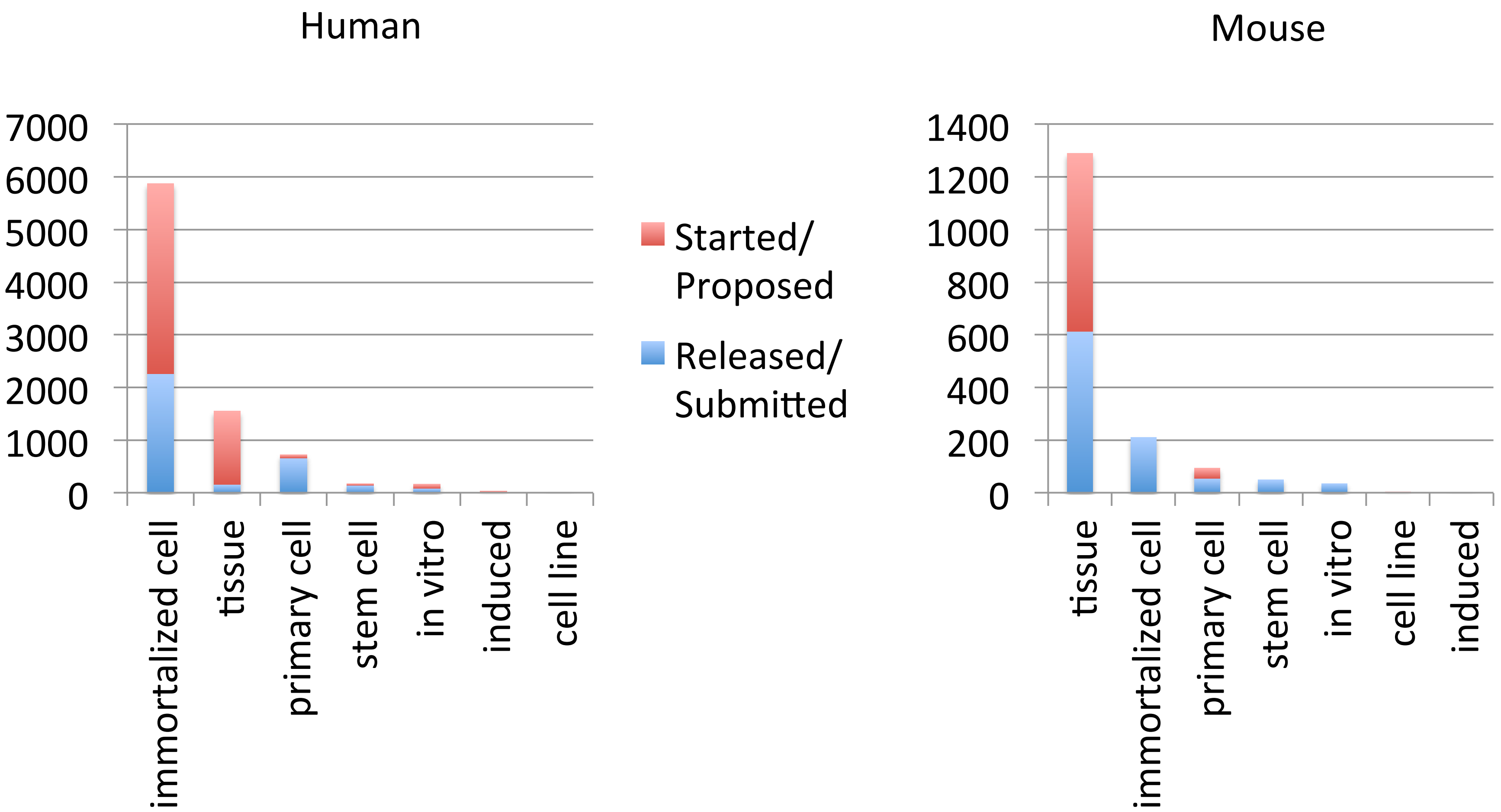
Data Types


Human
Experiments
10000
1000
100
10
1
🟥 Started/Proposed
🟦 Done/Submitted
ChIP-seq
RNA-seq
RBP shRNA/RNA-seq
iCLIP
RNA Bind-n-Seq
DNase-seq
RRBS
DNaseI
RNA array
RAMPAGE
Repli-seq
WGBS
CAGE
DNase-seq (conventional)
CGH array
DNA meth array
ATAC-Seq
DNase-seq (deep)
FAIRE-seq
RNA-PET
ChIA-PET
5C
RIP-seq
Repli-chip
DNA-PET
WGSBS
DNase-seq (long)
miRNA array
MeDIP-SEQ
MRE-seq
MNase-seq
Nanostring
methylC-seq
TF shRNA/RNA-Seq
Mouse
Experiments
10000
1000
100
10
1
🟥 Started/Proposed
🟦 Done/Submitted
ChIP-seq
RNA-seq
Nanostring
DNase-seq
WGSBS
methylC-seq
Repli-chip
ATAC-Seq
DNase-seq (deep)
DNase-seq (conv)
miRNA array
MeDIP-SEQ
MRE-seq
RNA Bind-n-Seq
DNase-seq (long)

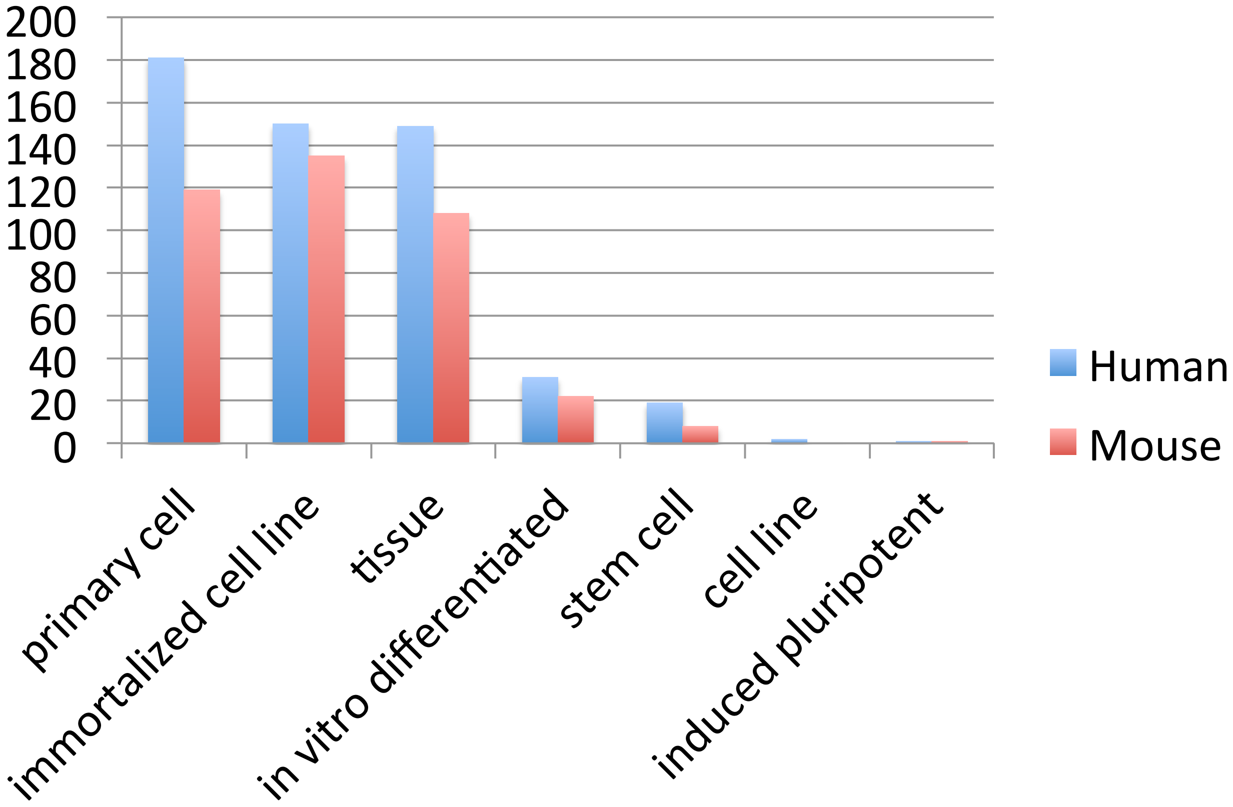
Some Assays Were Conducted Across a Broad Range of Biosamples



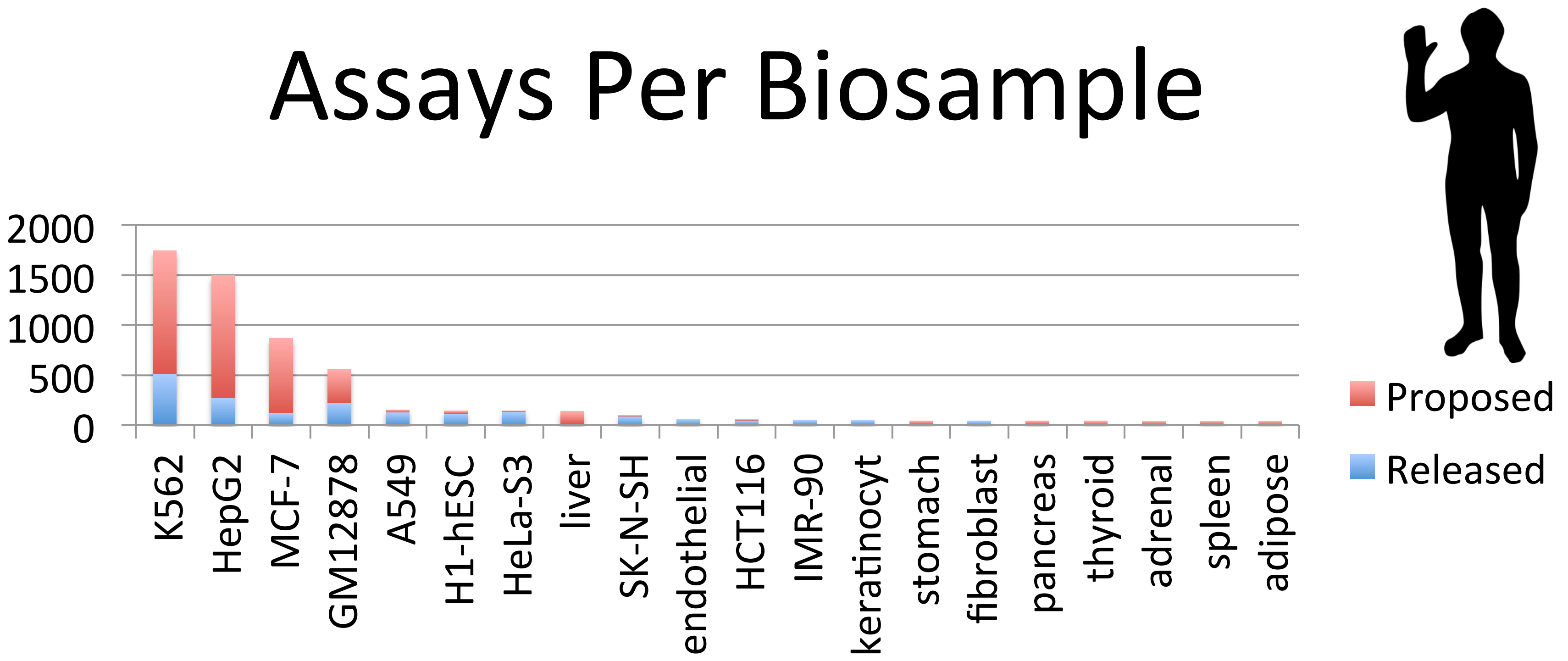
Number of Data Set Per Biosample Type

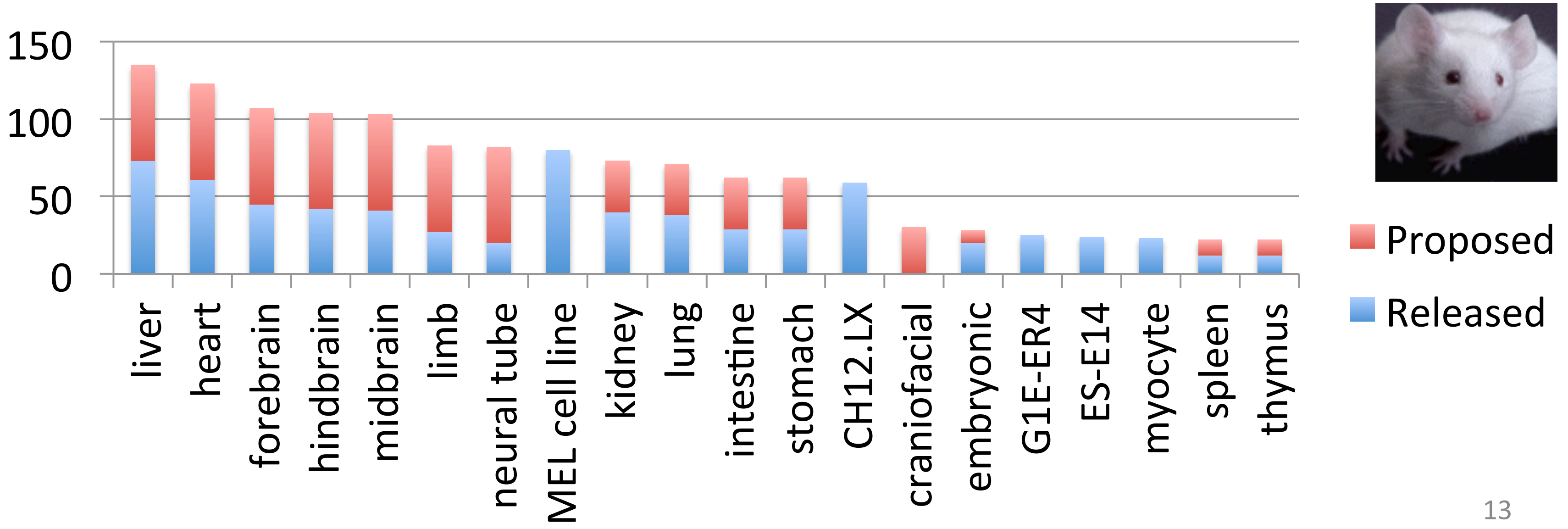


Unique Biosample Types

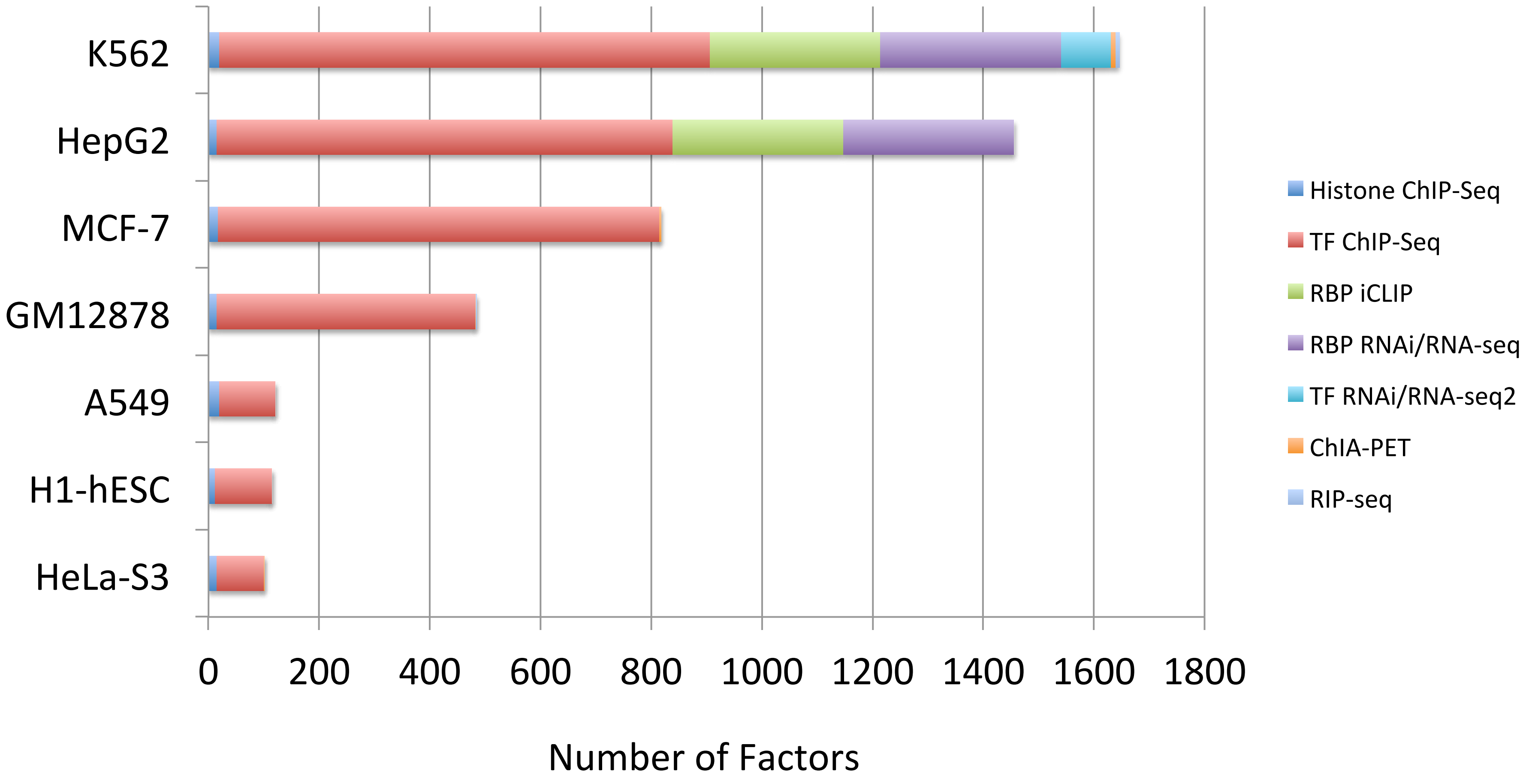


Assays Per Biosample





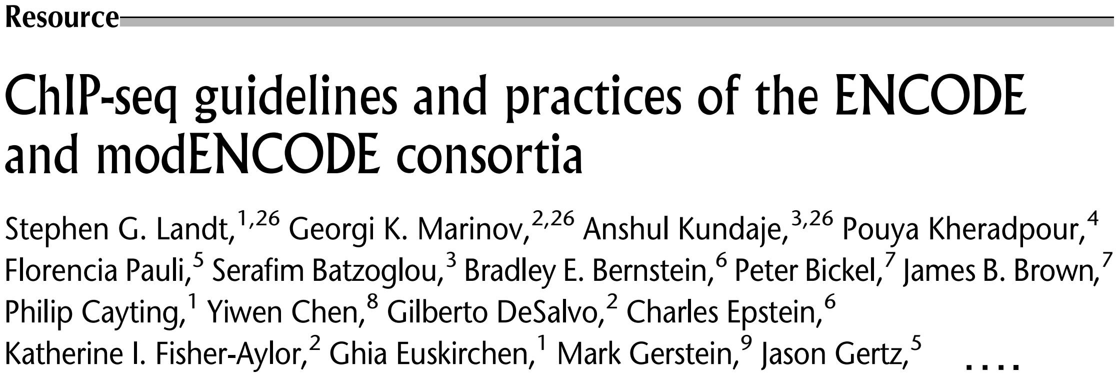
Deep Exploration of Factors



Established Standards For Community

- ChIP-Seq
- DNAaseHS
- RNA-Seq

Antibody characterizaiton, Biological replicates, QC measures



Resource

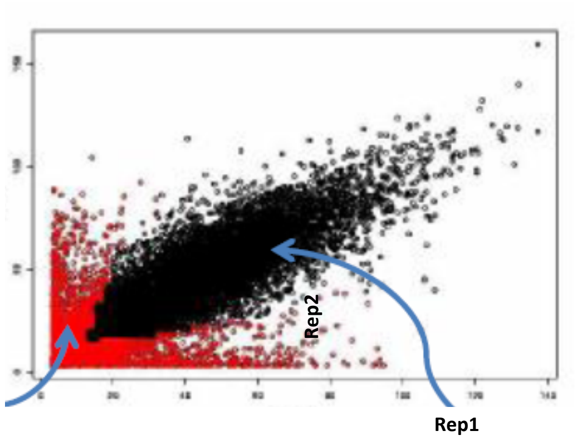
ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7] Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6] Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] ….

Genome Res. 2012

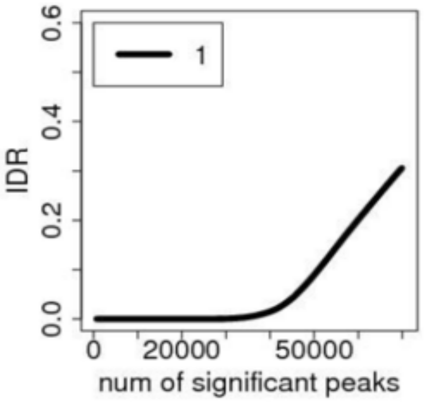
High Quality Data

- ≥ Two biological replicates
- Multiple quality control measures

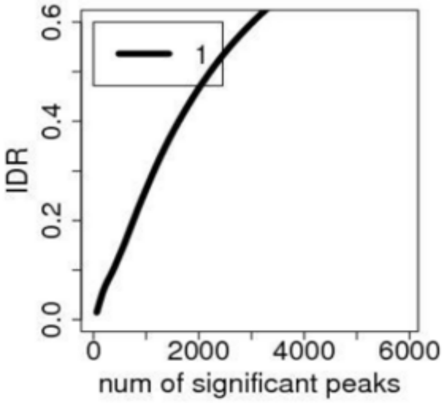


IDR Processing, QC and Blacklist Filtering

Good reproducibility



Poor reproducibility

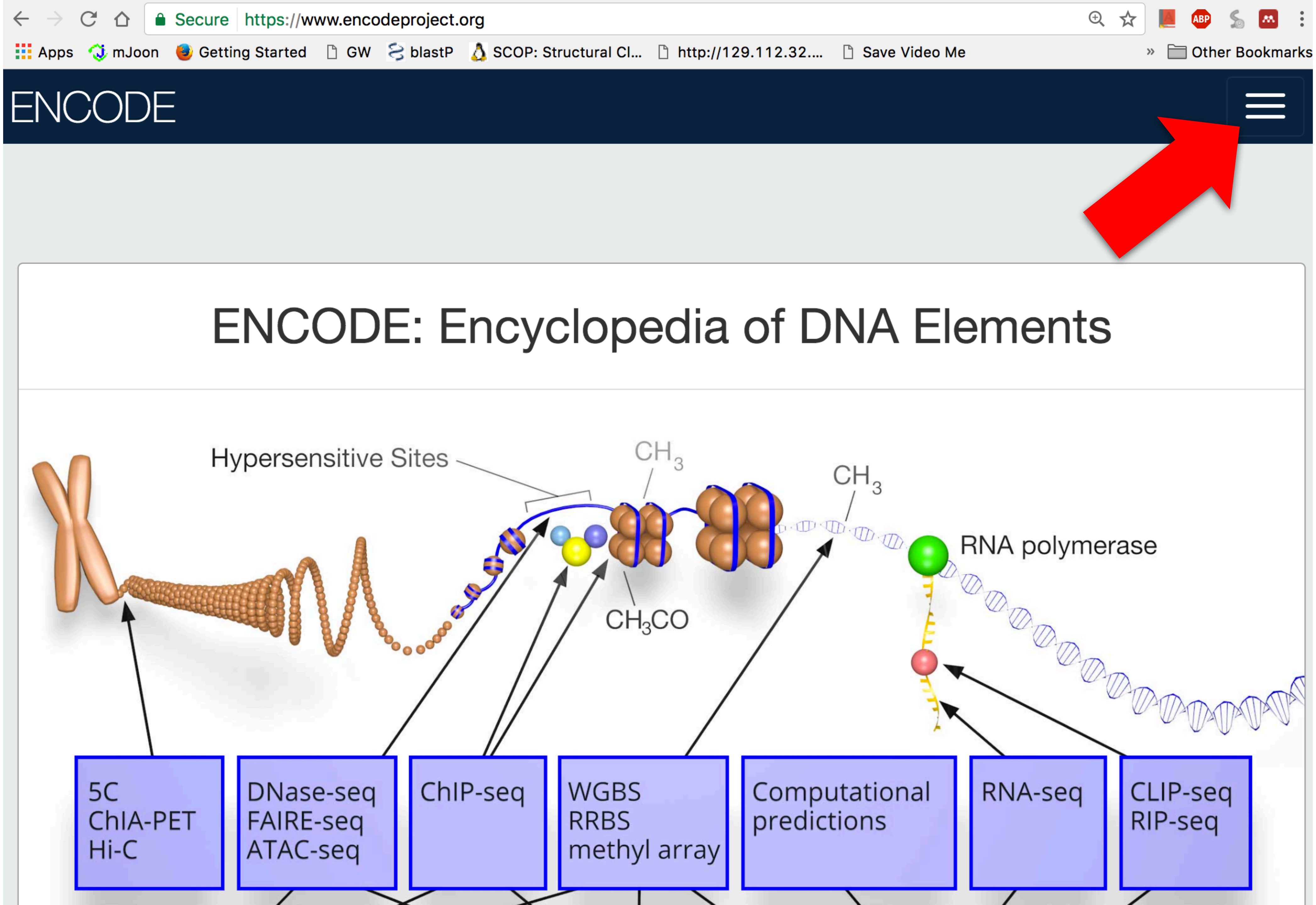


Major data types available in ENCODE: from raw data to analysis results

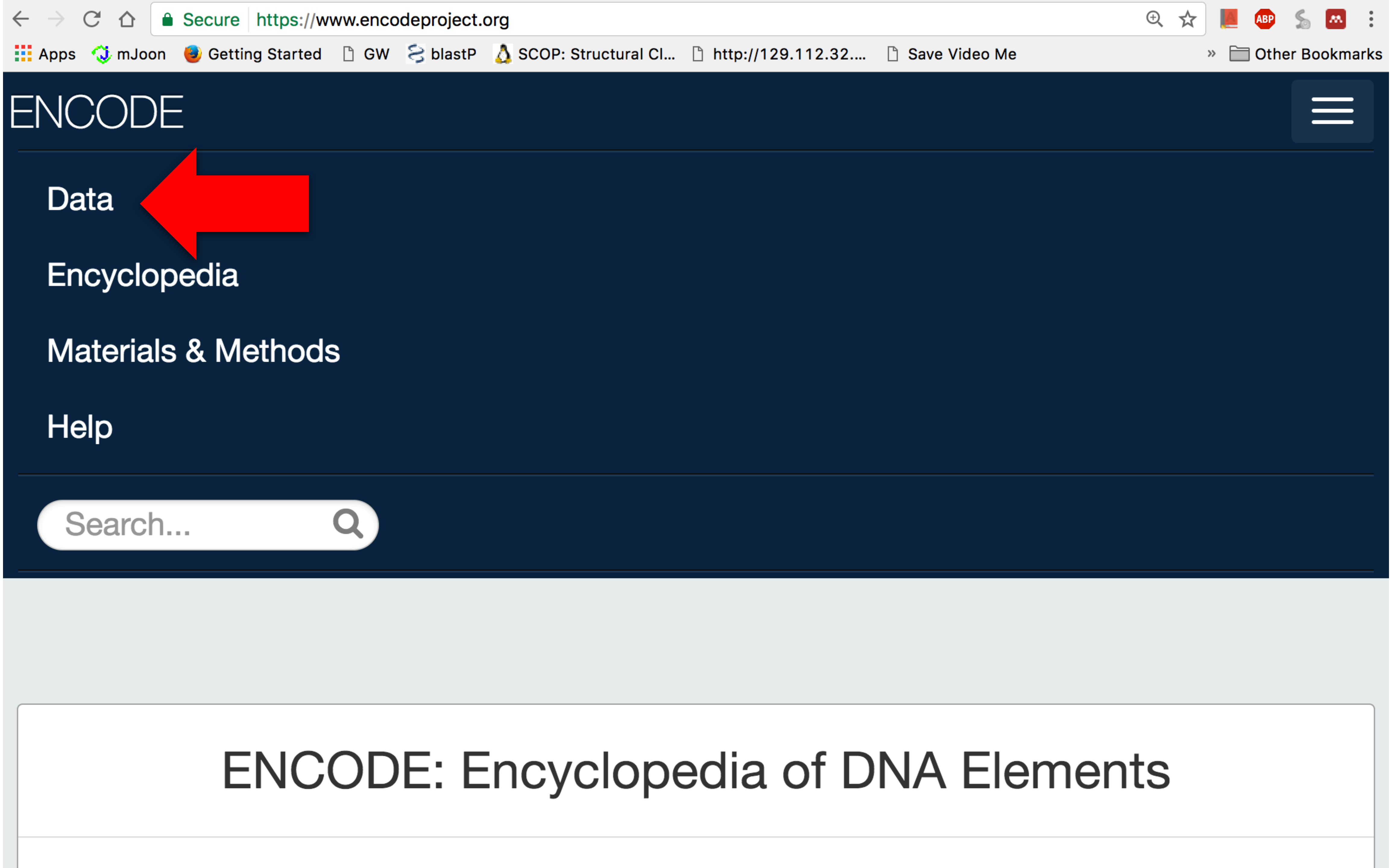
- Fastq
- BAM
- Peaks
- BigWig
- mRNA Expression profiles

What did we “learn” from ENCODE project?

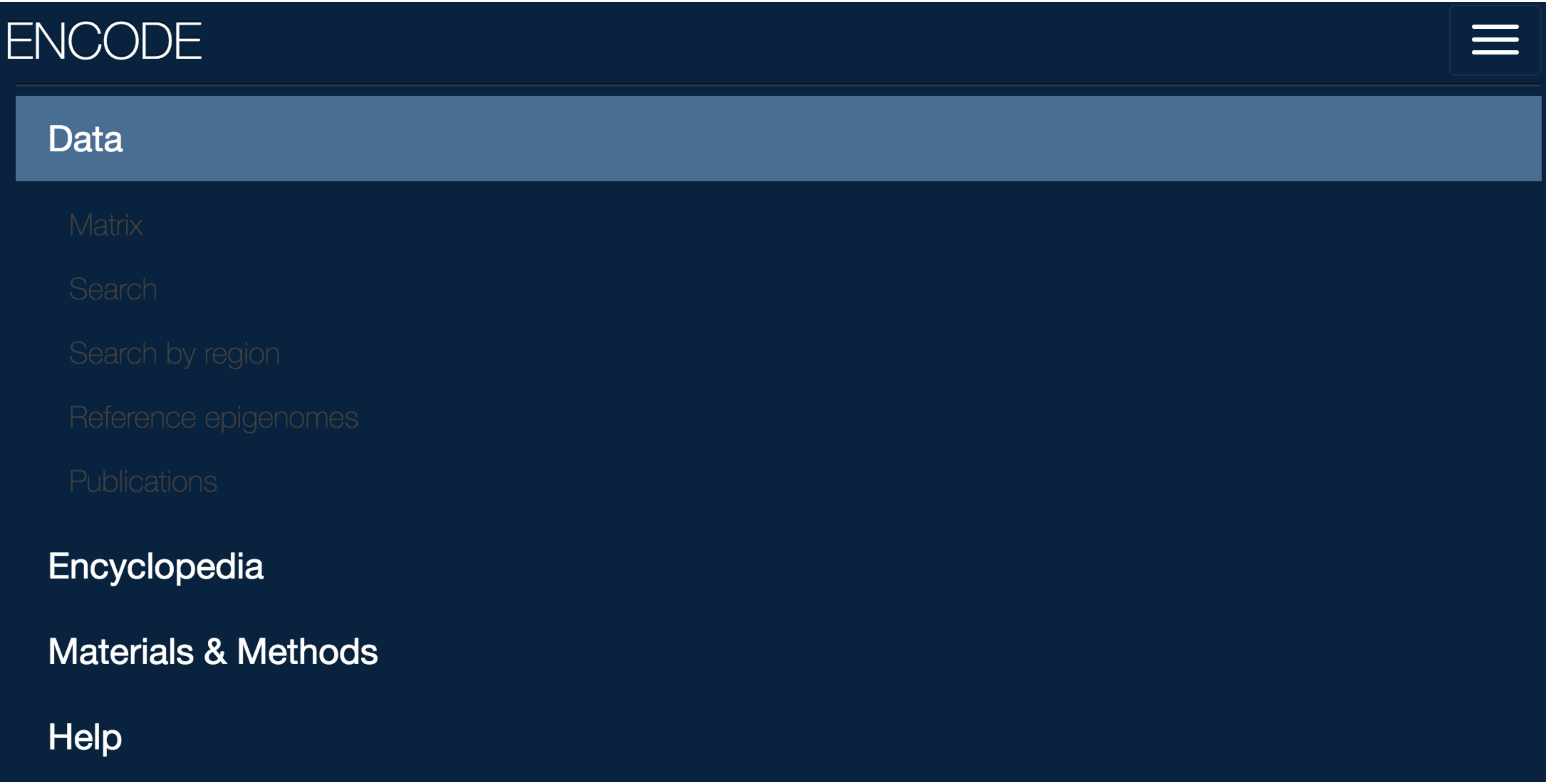
ENCODE Data Access

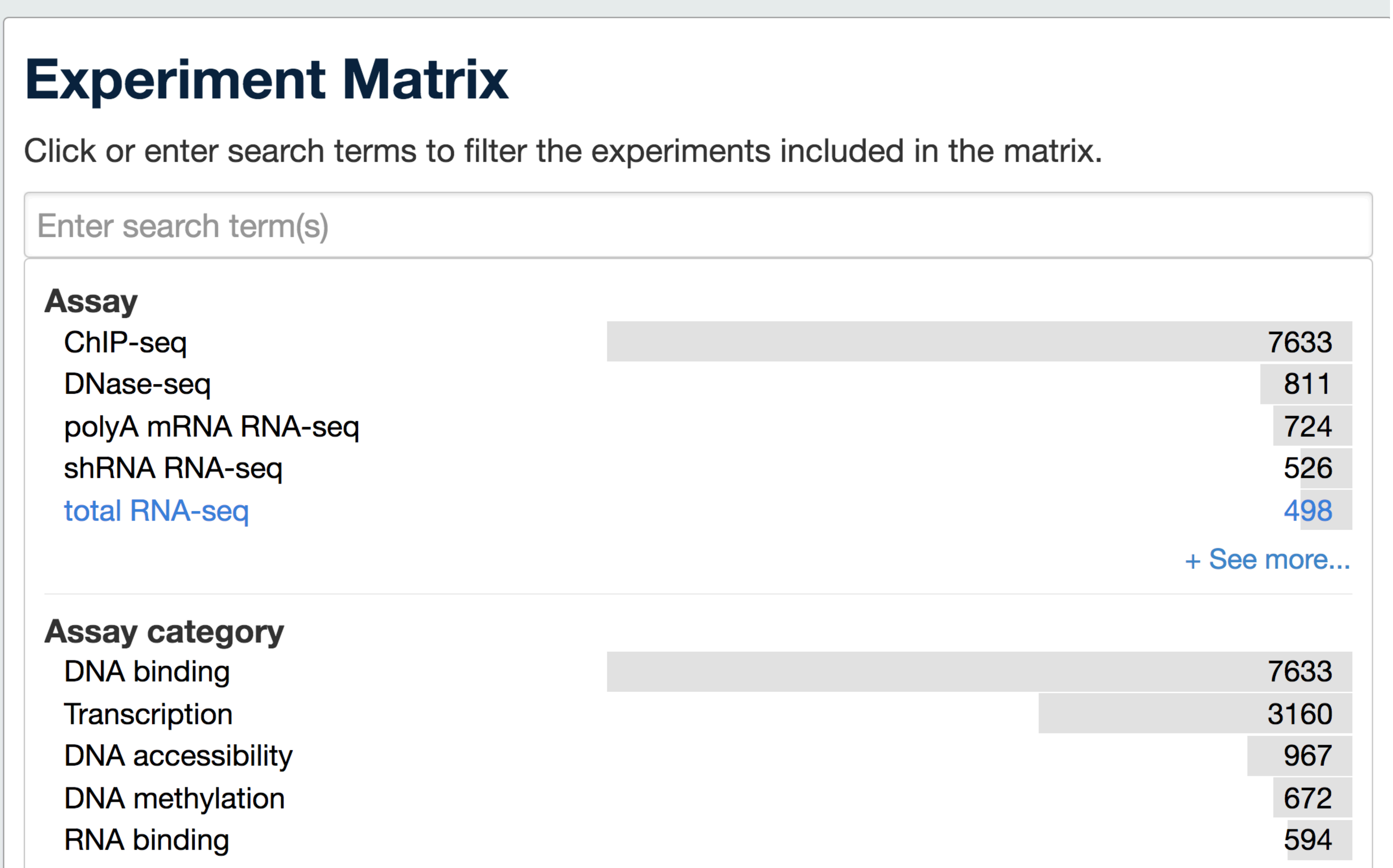


ENCODE: Encyclopedia of DNA Elements



ENCODE: Encyclopedia of DNA Elements


ENCODE
Data
Matrix
Search
Search by region
Reference epigenomes
Publications
Encyclopedia
Materials & Methods
Help



Experiment Matrix

Click or enter search terms to filter the experiments included in the matrix.

Enter search term(s)

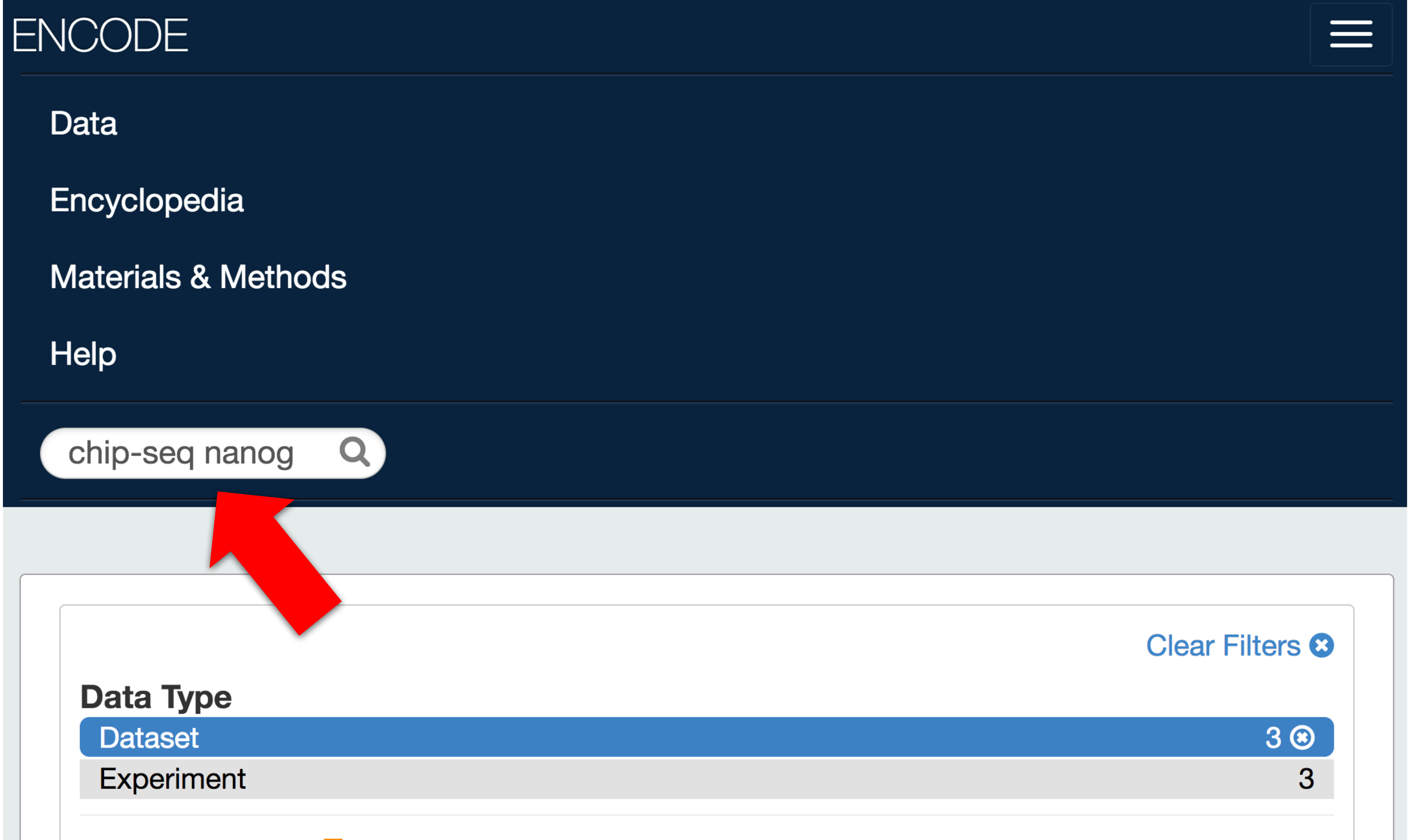
Assay

ChIP-seq	7633
DNase-seq	811
polyA mRNA RNA-seq	724
shRNA RNA-seq	526
total RNA-seq	498

+ See more...

Assay category

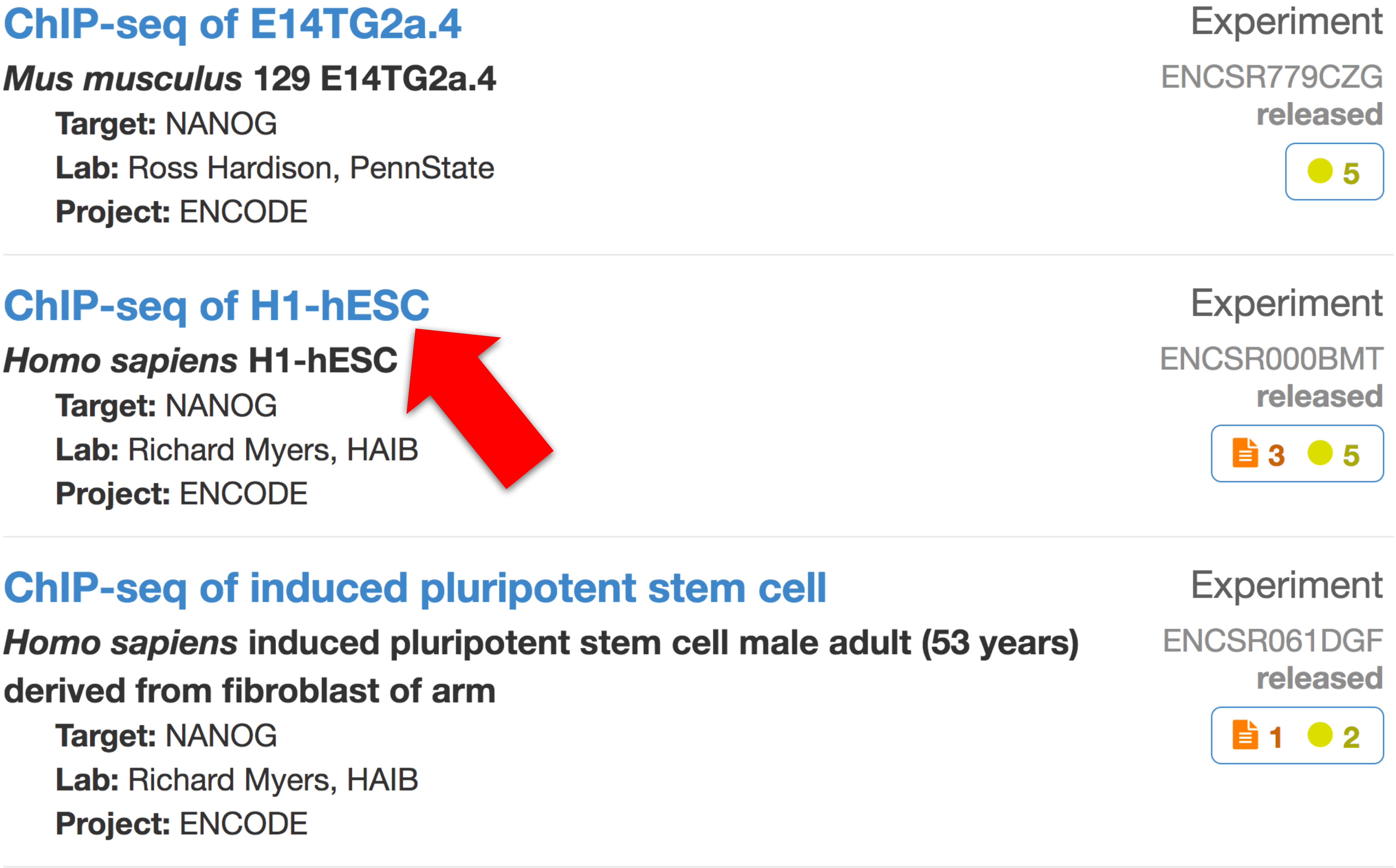
DNA binding	7633
Transcription	3160
DNA accessibility	967
DNA methylation	672
RNA binding	594



chip-seq nanog

Data Type

Dataset	3
Experiment	3



ChIP-seq of E14TG2a.4

Mus musculus **129 E14TG2a.4**

Target: NANOG
Lab: Ross Hardison, PennState
Project: ENCODE

Experiment
ENCSR779CZG
released
● 5

ChIP-seq of H1-hESC

Homo sapiens **H1-hESC**

Target: NANOG
Lab: Richard Myers, HAIB
Project: ENCODE

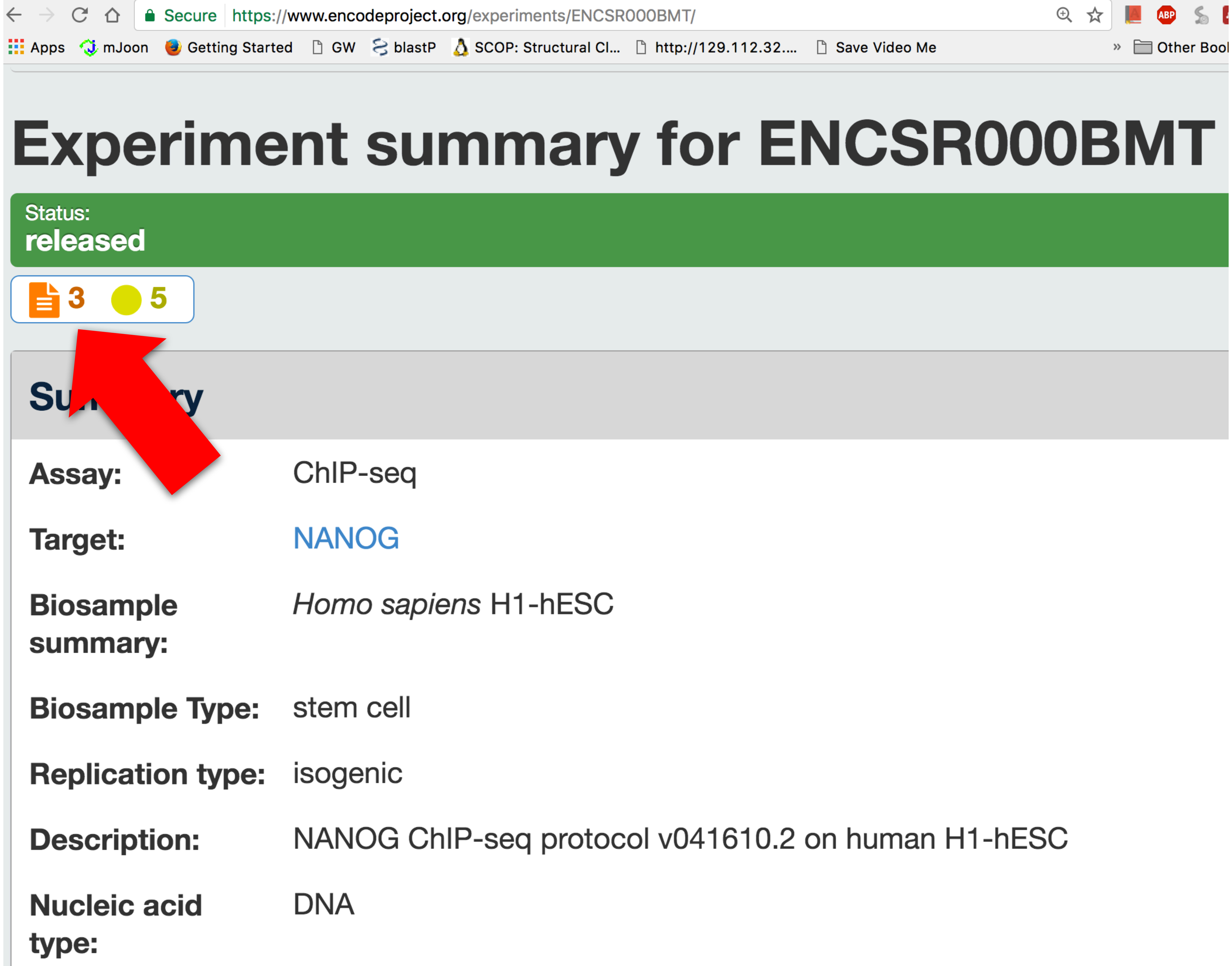
Experiment
ENCSR000BMT
released
📄 3 ● 5

ChIP-seq of induced pluripotent stem cell

Homo sapiens **induced pluripotent stem cell male adult (53 years) derived from fibroblast of arm**

Target: NANOG
Lab: Richard Myers, HAIB
Project: ENCODE

Experiment
ENCSR061DGF
released
📄 1 ● 2



Experiment summary for ENCSR000BMT

Status:
released

3 ● 5

Summary

Assay: ChIP-seq

Target: NANOG

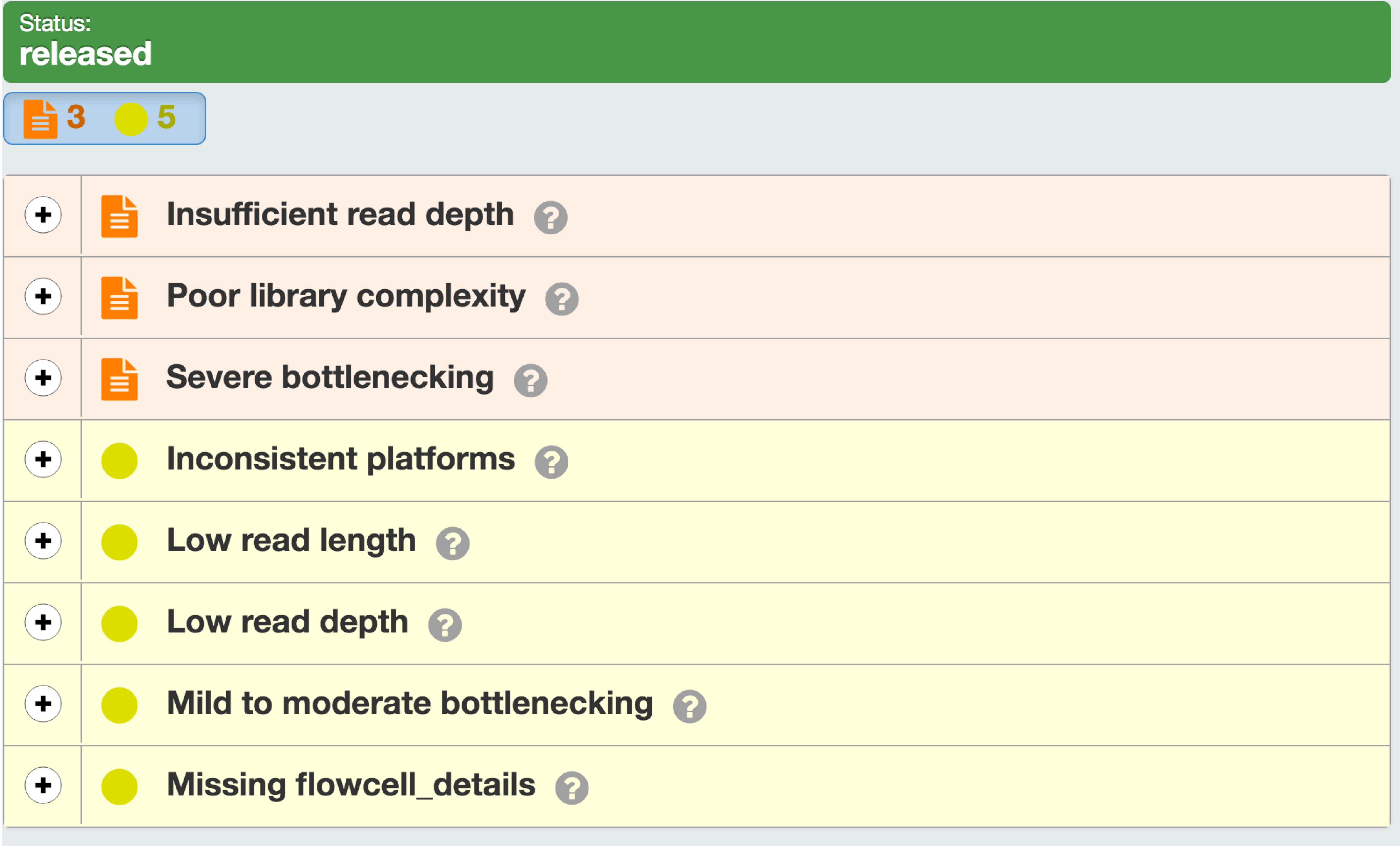
Biosample summary: *Homo sapiens* H1-hESC

Biosample Type: stem cell

Replication type: isogenic

Description: NANOG ChIP-seq protocol v041610.2 on human H1-hESC

Nucleic acid type: DNA

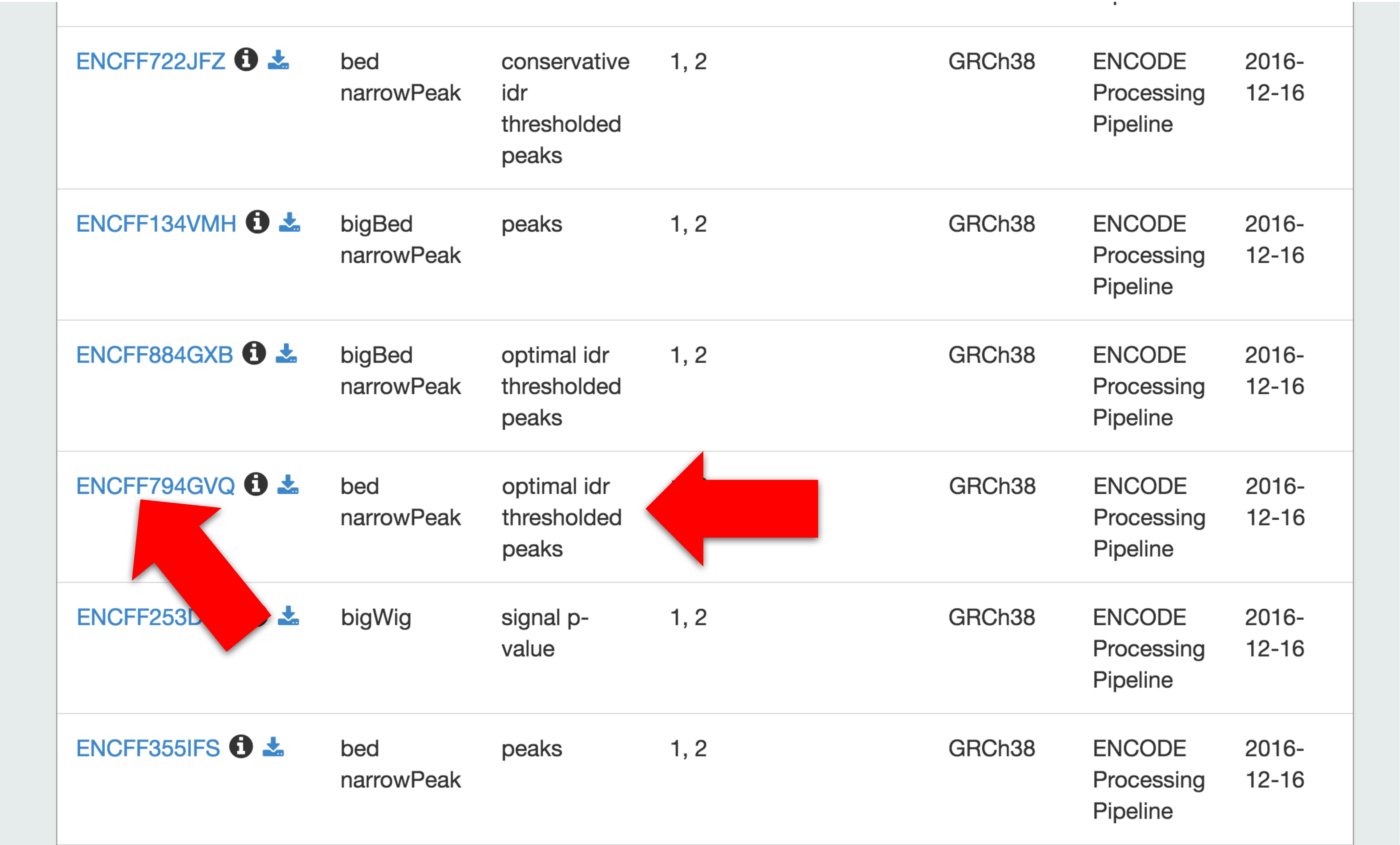


Status:
released

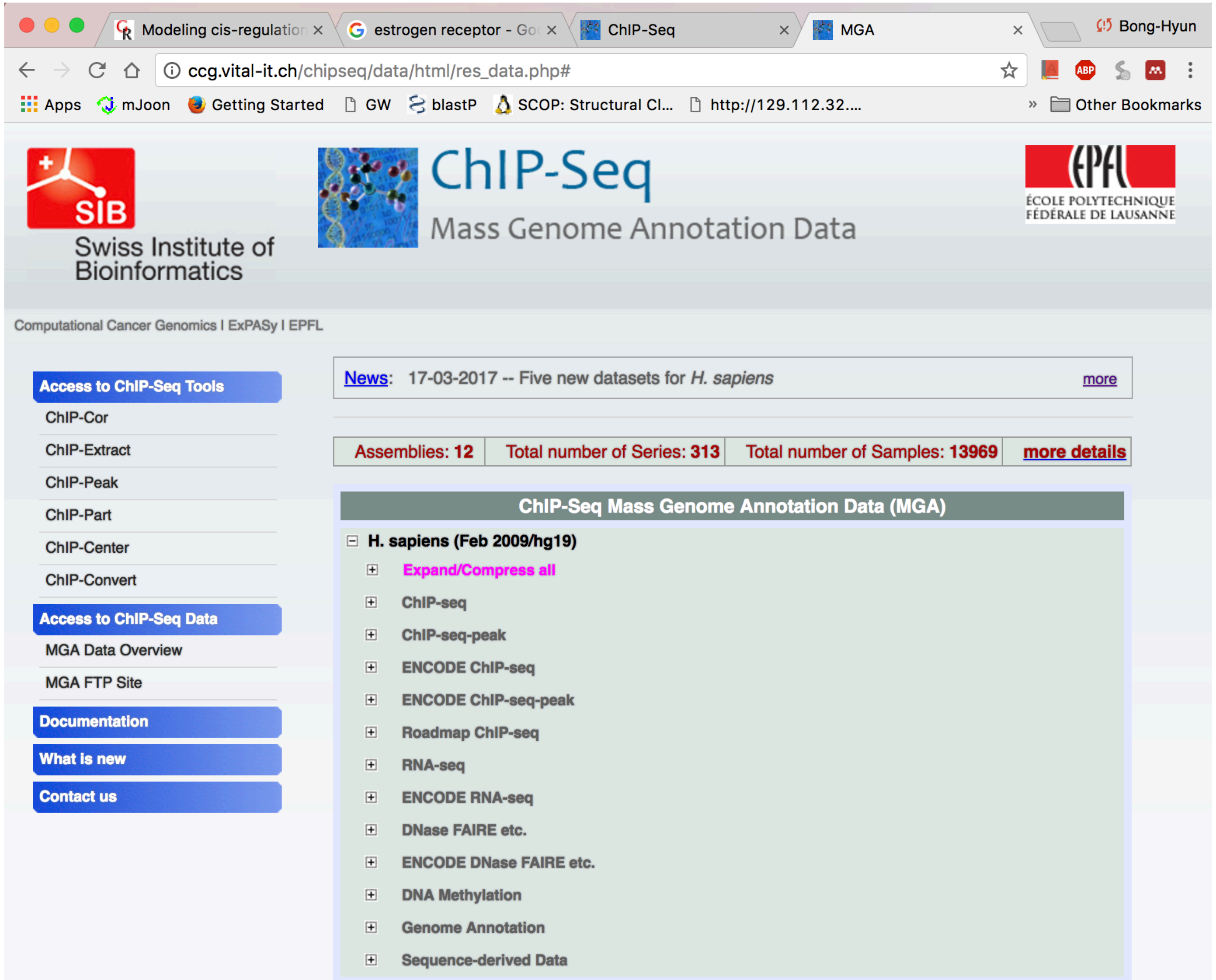
📄 3 ● 5

- ⊕ 📄 Insufficient read depth ?
- ⊕ 📄 Poor library complexity ?
- ⊕ 📄 Severe bottlenecking ?
- ⊕ ● Inconsistent platforms ?
- ⊕ ● Low read length ?
- ⊕ ● Low read depth ?
- ⊕ ● Mild to moderate bottlenecking ?
- ⊕ ● Missing flowcell_details ?

https://www.encodeproject.org/files/ENCFF794GVQ/@@download/ENCFF794GVQ.bed.gz



ENCFF722JFZ	bed narrowPeak	conservative idr thresholded peaks	1, 2	GRCh38	ENCODE Processing Pipeline	2016-12-16
ENCFF134VMH	bigBed narrowPeak	peaks	1, 2	GRCh38	ENCODE Processing Pipeline	2016-12-16
ENCFF884GXB	bigBed narrowPeak	optimal idr thresholded peaks	1, 2	GRCh38	ENCODE Processing Pipeline	2016-12-16
ENCFF794GVQ	bed narrowPeak	optimal idr thresholded peaks	[illegible]	GRCh38	ENCODE Processing Pipeline	2016-12-16
ENCFF253D	bigWig	signal p-value	1, 2	GRCh38	ENCODE Processing Pipeline	2016-12-16
ENCFF355IFS	bed narrowPeak	peaks	1, 2	GRCh38	ENCODE Processing Pipeline	2016-12-16



ChIP-Seq

Mass Genome Annotation Data

Computational Cancer Genomics I ExPASy I EPFL

News: 17-03-2017 -- Five new datasets for *H. sapiens* more

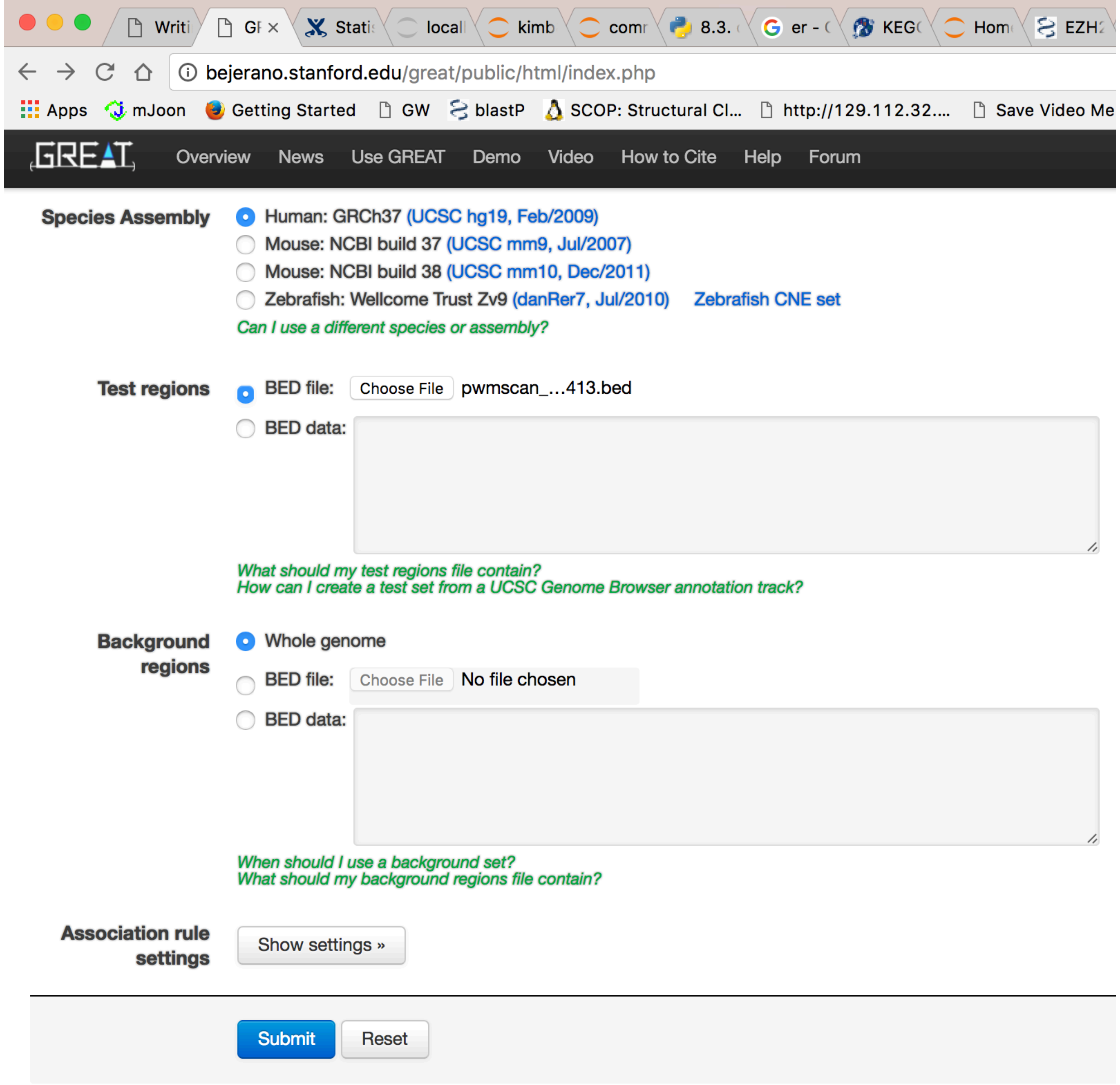
Assemblies: 12	Total number of Series: 313	Total number of Samples: 13969	more details

ChIP-Seq Mass Genome Annotation Data (MGA)

⊟ H. sapiens (Feb 2009/hg19)
- ⊞ Expand/Compress all
- ⊞ ChIP-seq
- ⊞ ChIP-seq-peak
- ⊞ ENCODE ChIP-seq
- ⊞ ENCODE ChIP-seq-peak
- ⊞ Roadmap ChIP-seq
- ⊞ RNA-seq
- ⊞ ENCODE RNA-seq
- ⊞ DNase FAIRE etc.
- ⊞ ENCODE DNase FAIRE etc.
- ⊞ DNA Methylation
- ⊞ Genome Annotation
- ⊞ Sequence-derived Data

GREAT

- Understanding the peaks



Species Assembly

● Human: GRCh37 (UCSC hg19, Feb/2009)
○ Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
○ Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
○ Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set

Can I use a different species or assembly?

Test regions

● BED file: Choose File pwmscan_...413.bed
○ BED data:

What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?

Background regions

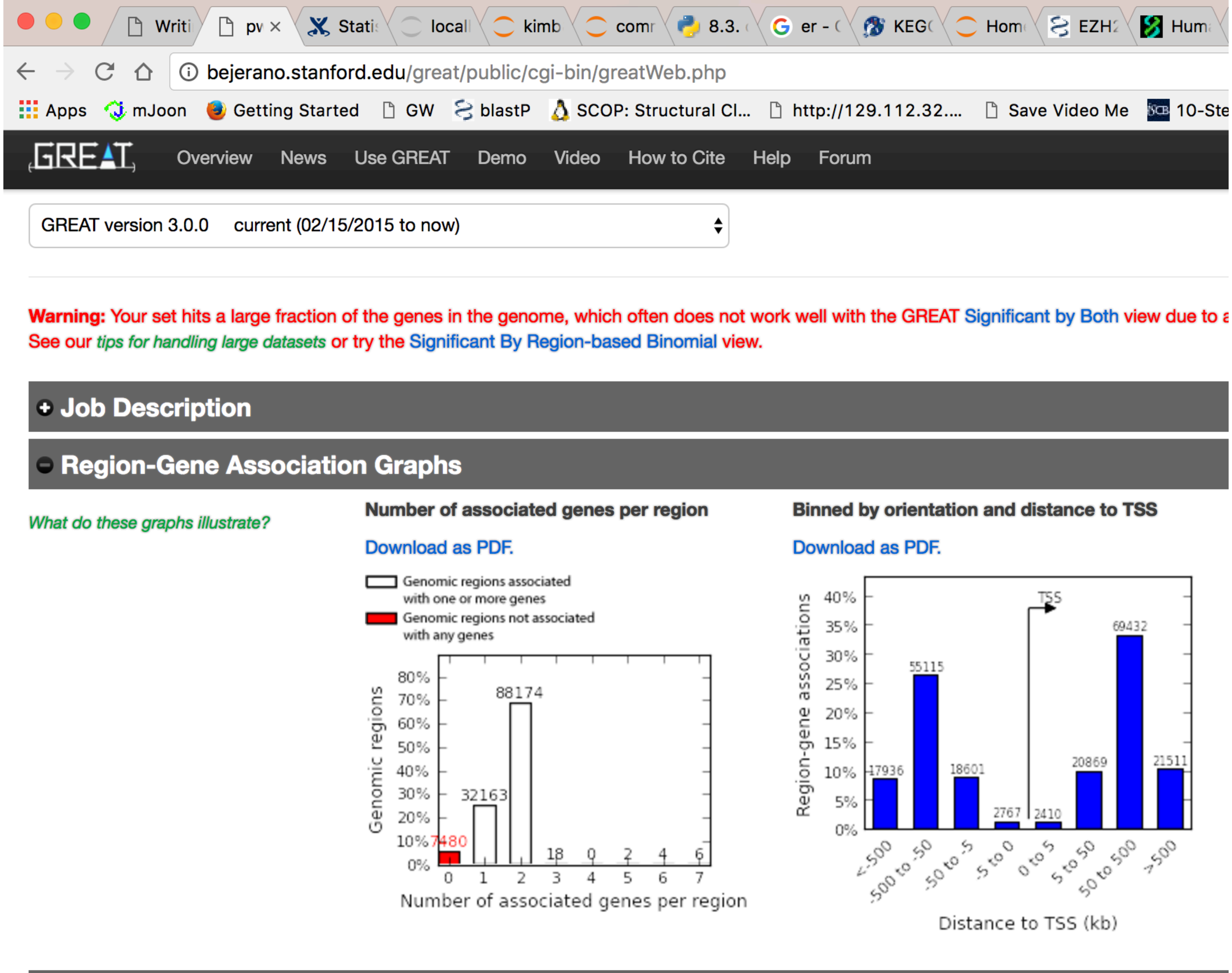
● Whole genome
○ BED file: Choose File No file chosen
○ BED data:

When should I use a background set?
What should my background regions file contain?

Association rule settings

Show settings »

Submit Reset



Warning: Your set hits a large fraction of the genes in the genome, which often does not work well with the GREAT Significant by Both view due to a
See our *tips for handling large datasets* or try the Significant By Region-based Binomial view.

Job Description

Region-Gene Association Graphs

What do these graphs illustrate?

Number of associated genes per region

Download as PDF.

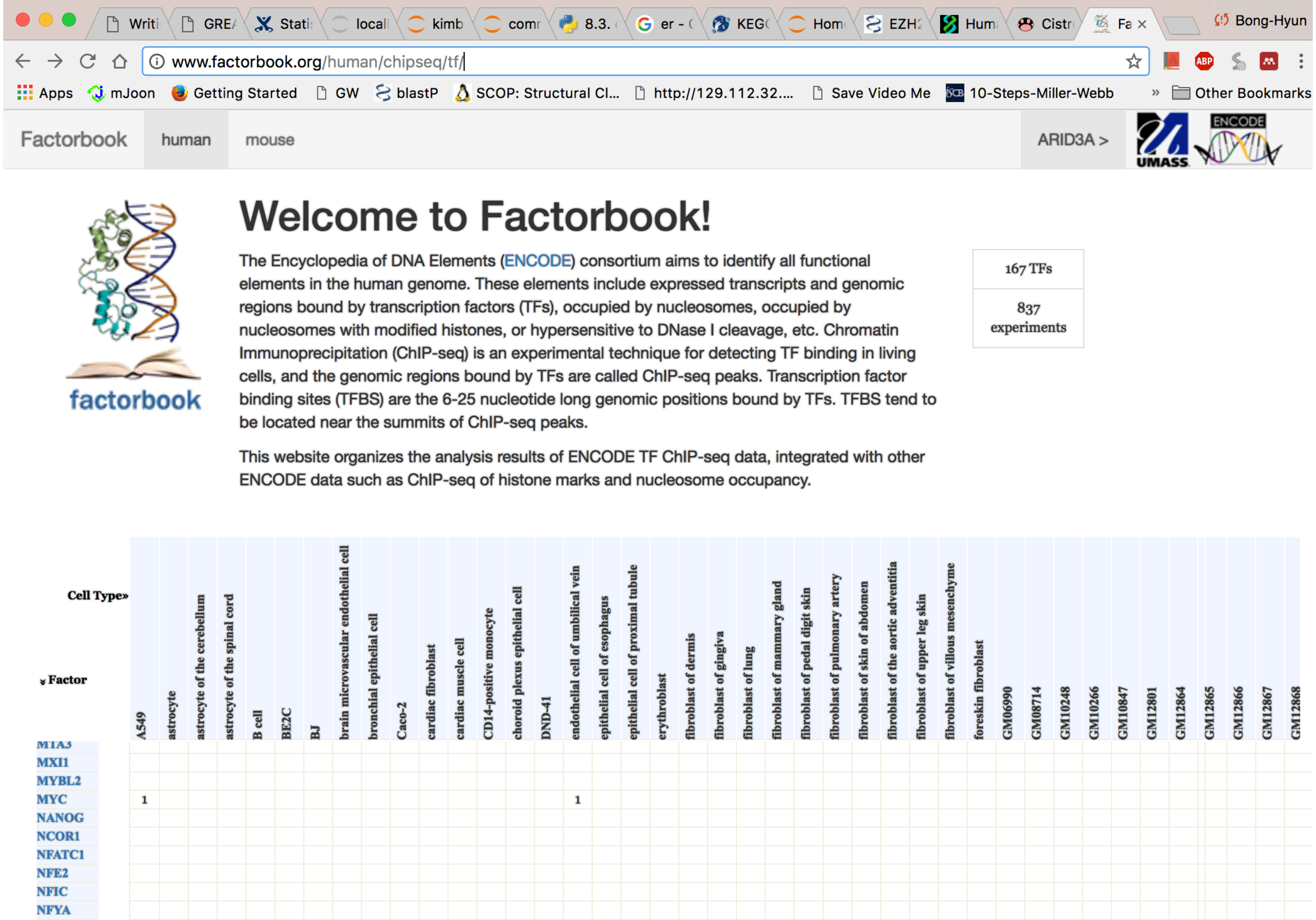
Genomic regions associated with one or more genes
Genomic regions not associated with any genes
Genomic regions
80%
70%
60%
50%
40%
30%
20%
10%
0%
7480
32163
88174
18
0
2
4
6
0
1
2
3
4
5
6
7
Number of associated genes per region

Binned by orientation and distance to TSS

Download as PDF.

But Cataloging is not enough!

Factorbook



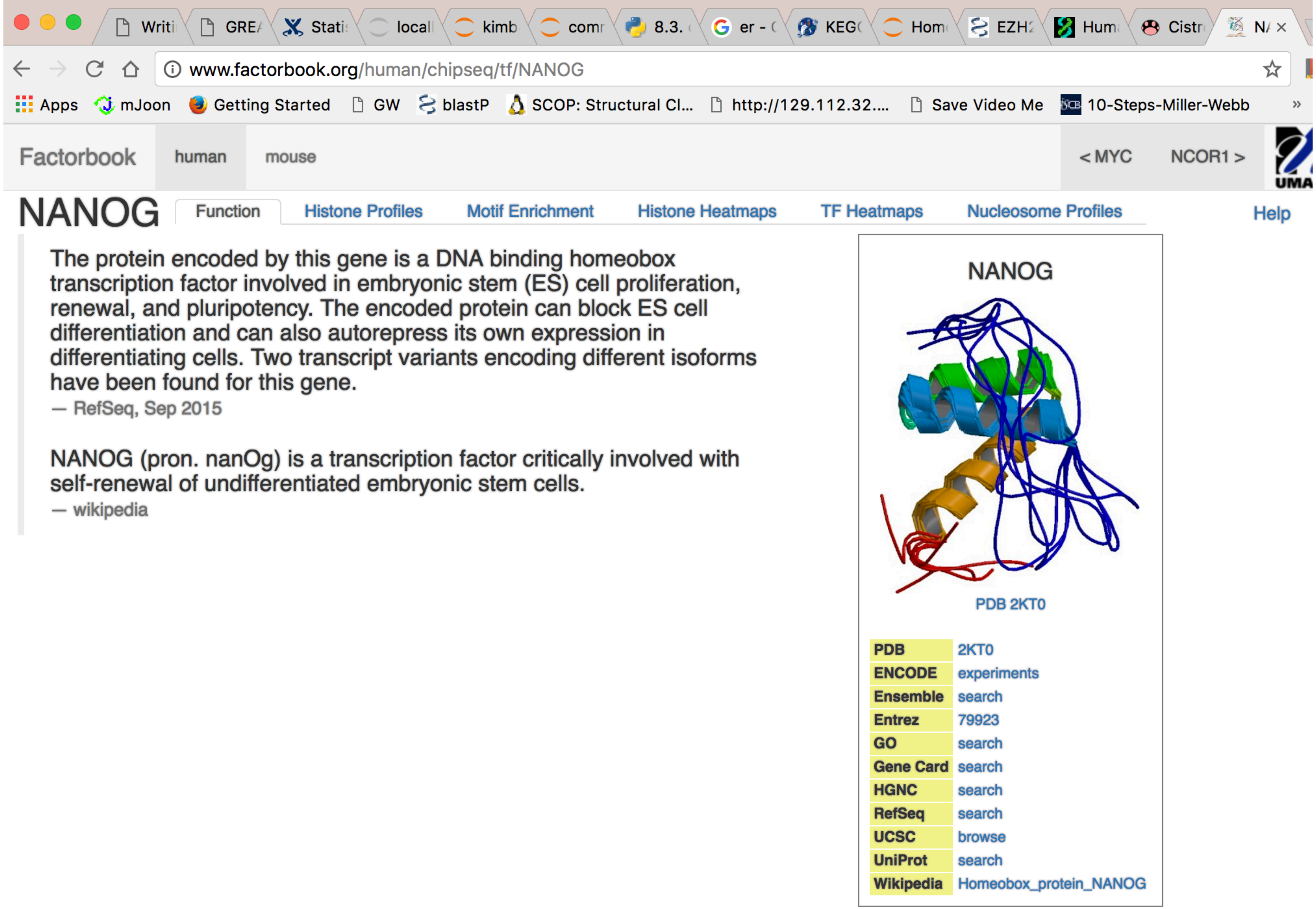
Welcome to Factorbook!

The Encyclopedia of DNA Elements (ENCODE) consortium aims to identify all functional elements in the human genome. These elements include expressed transcripts and genomic regions bound by transcription factors (TFs), occupied by nucleosomes, occupied by nucleosomes with modified histones, or hypersensitive to DNase I cleavage, etc. Chromatin Immunoprecipitation (ChIP-seq) is an experimental technique for detecting TF binding in living cells, and the genomic regions bound by TFs are called ChIP-seq peaks. Transcription factor binding sites (TFBS) are the 6-25 nucleotide long genomic positions bound by TFs. TFBS tend to be located near the summits of ChIP-seq peaks.

This website organizes the analysis results of ENCODE TF ChIP-seq data, integrated with other ENCODE data such as ChIP-seq of histone marks and nucleosome occupancy.

167 TFs
837 experiments

Cell Type» ⌄ Factor	A549	astrocyte	astrocyte of the cerebellum	astrocyte of the spinal cord	B cell	BE2C	BJ	brain microvascular endothelial cell	bronchial epithelial cell	Caco-2	cardiac fibroblast	cardiac muscle cell	CD14-positive monocyte	choroid plexus epithelial cell	DND-41	endothelial cell of umbilical vein	epithelial cell of esophagus	epithelial cell of proximal tubule	erythroblast	fibroblast of dermis	fibroblast of gingiva	fibroblast of lung	fibroblast of mammary gland	fibroblast of pedal digit skin	fibroblast of pulmonary artery	fibroblast of skin of abdomen	fibroblast of the aortic adventitia	fibroblast of upper leg skin	fibroblast of villous mesenchyme	foreskin fibroblast	GM06990	GM08714	GM10248	GM10266	GM10847	GM12801	GM12864	GM12865	GM12866	GM12867	GM12868
MTA3																																									
MXI1																																									
MYBL2																																									
MYC	1															1																									
NANOG																																									
NCOR1																																									
NFATC1																																									
NFE2																																									
NFIC																																									
NFYA																																									



NANOG

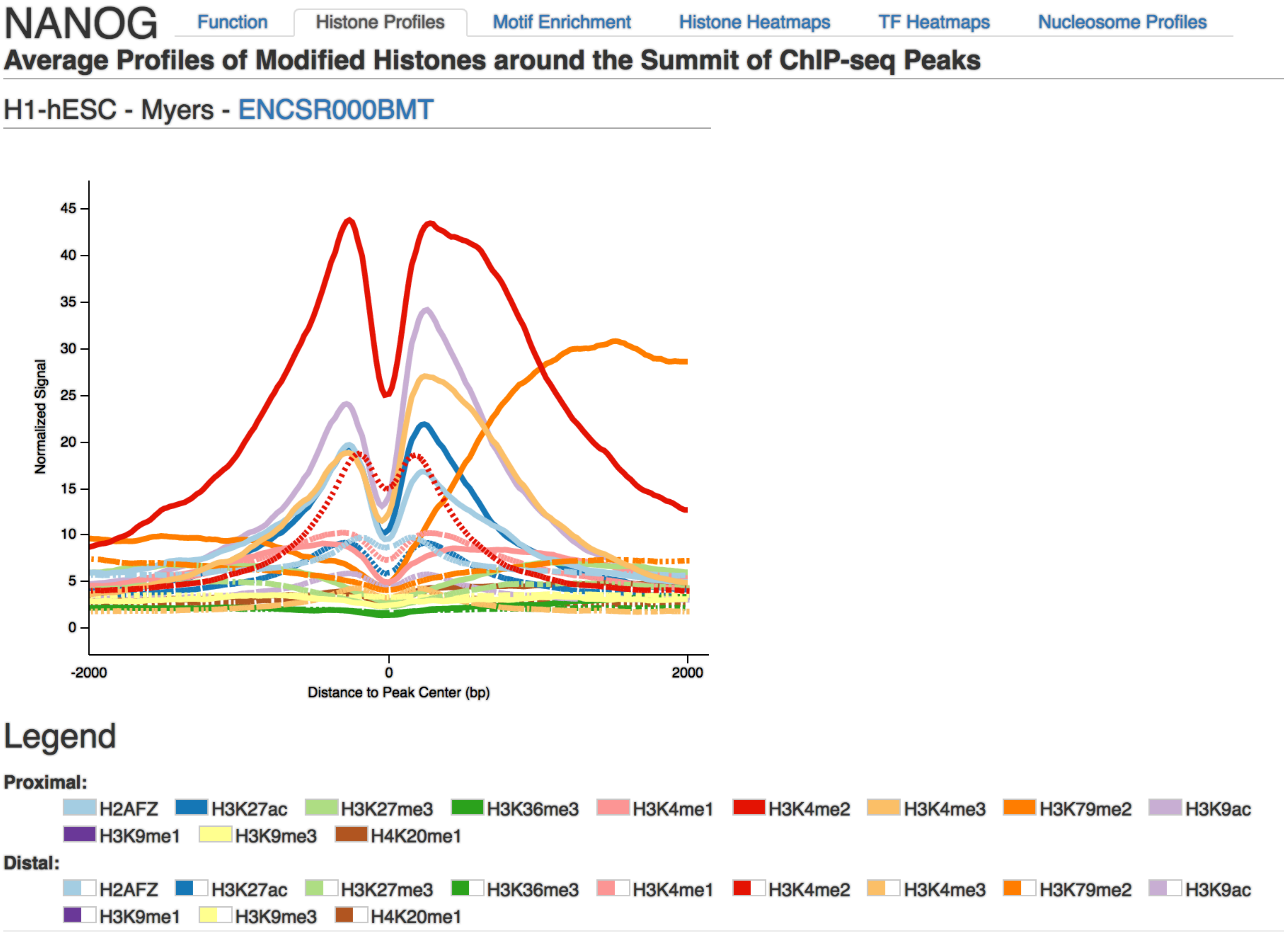
The protein encoded by this gene is a DNA binding homeobox transcription factor involved in embryonic stem (ES) cell proliferation, renewal, and pluripotency. The encoded protein can block ES cell differentiation and can also autorepress its own expression in differentiating cells. Two transcript variants encoding different isoforms have been found for this gene.
— RefSeq, Sep 2015

NANOG (pron. nanOg) is a transcription factor critically involved with self-renewal of undifferentiated embryonic stem cells.
— wikipedia

NANOG

PDB 2KT0

PDB	2KT0
ENCODE	experiments
Ensemble	search
Entrez	79923
GO	search
Gene Card	search
HGNC	search
RefSeq	search
UCSC	browse
UniProt	search
Wikipedia	Homeobox_protein_NANOG



NANOG

Function | Histone Profiles | Motif Enrichment | Histone Heatmaps | TF Heatmaps | Nucleosome Profiles

Average Profiles of Modified Histones around the Summit of ChIP-seq Peaks

H1-hESC - Myers - ENCSR000BMT

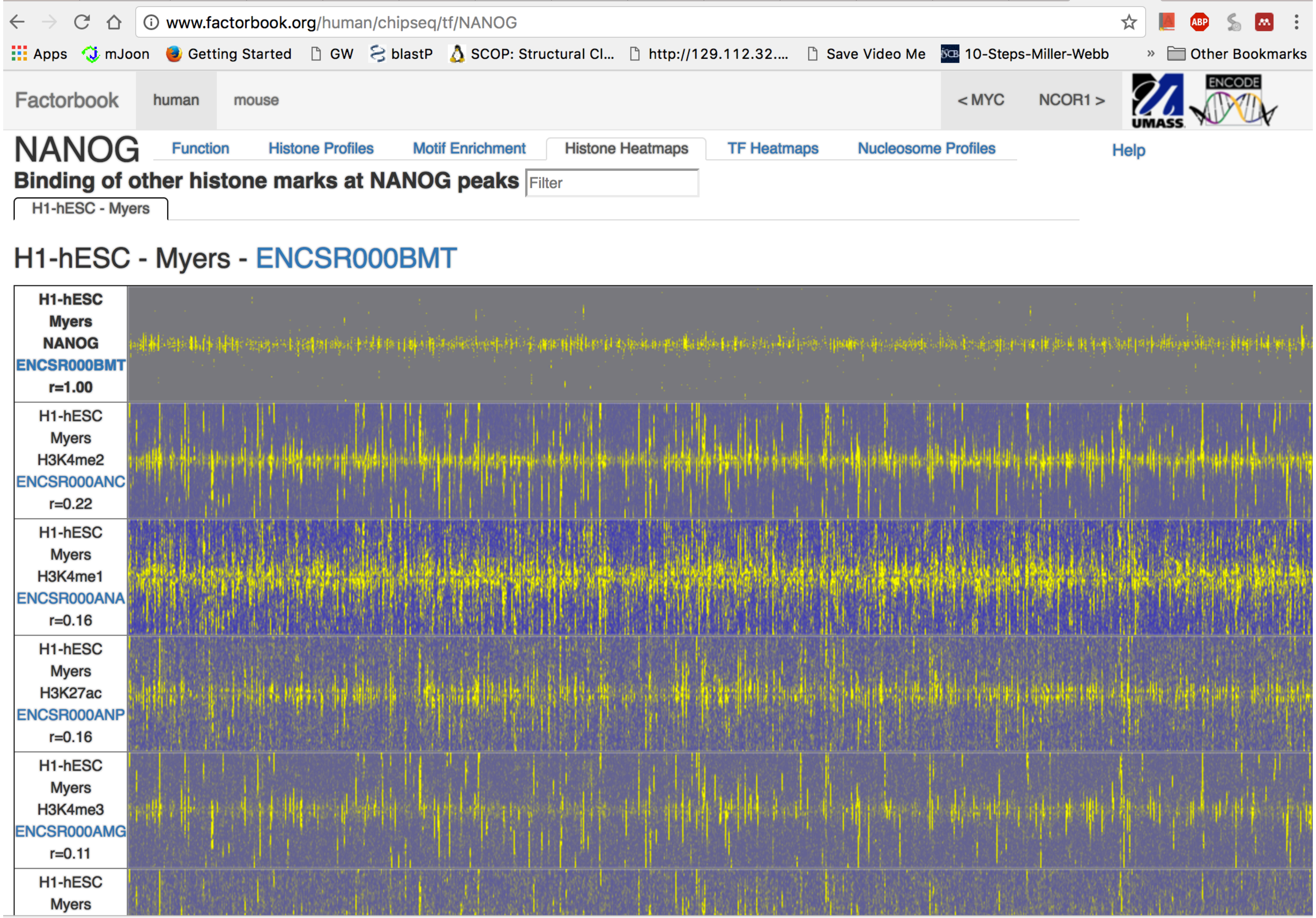
Legend

Proximal:

■ H2AFZ ■ H3K27ac ■ H3K27me3 ■ H3K36me3 ■ H3K4me1 ■ H3K4me2 ■ H3K4me3 ■ H3K79me2 ■ H3K9ac
■ H3K9me1 ■ H3K9me3 ■ H4K20me1

Distal:

◧ H2AFZ ◧ H3K27ac ◧ H3K27me3 ◧ H3K36me3 ◧ H3K4me1 ◧ H3K4me2 ◧ H3K4me3 ◧ H3K79me2 ◧ H3K9ac
◧ H3K9me1 ◧ H3K9me3 ◧ H4K20me1



NANOG

Binding of other histone marks at NANOG peaks

H1-hESC - Myers - ENCSR000BMT

H1-hESC
Myers
NANOG
ENCSR000BMT
r=1.00

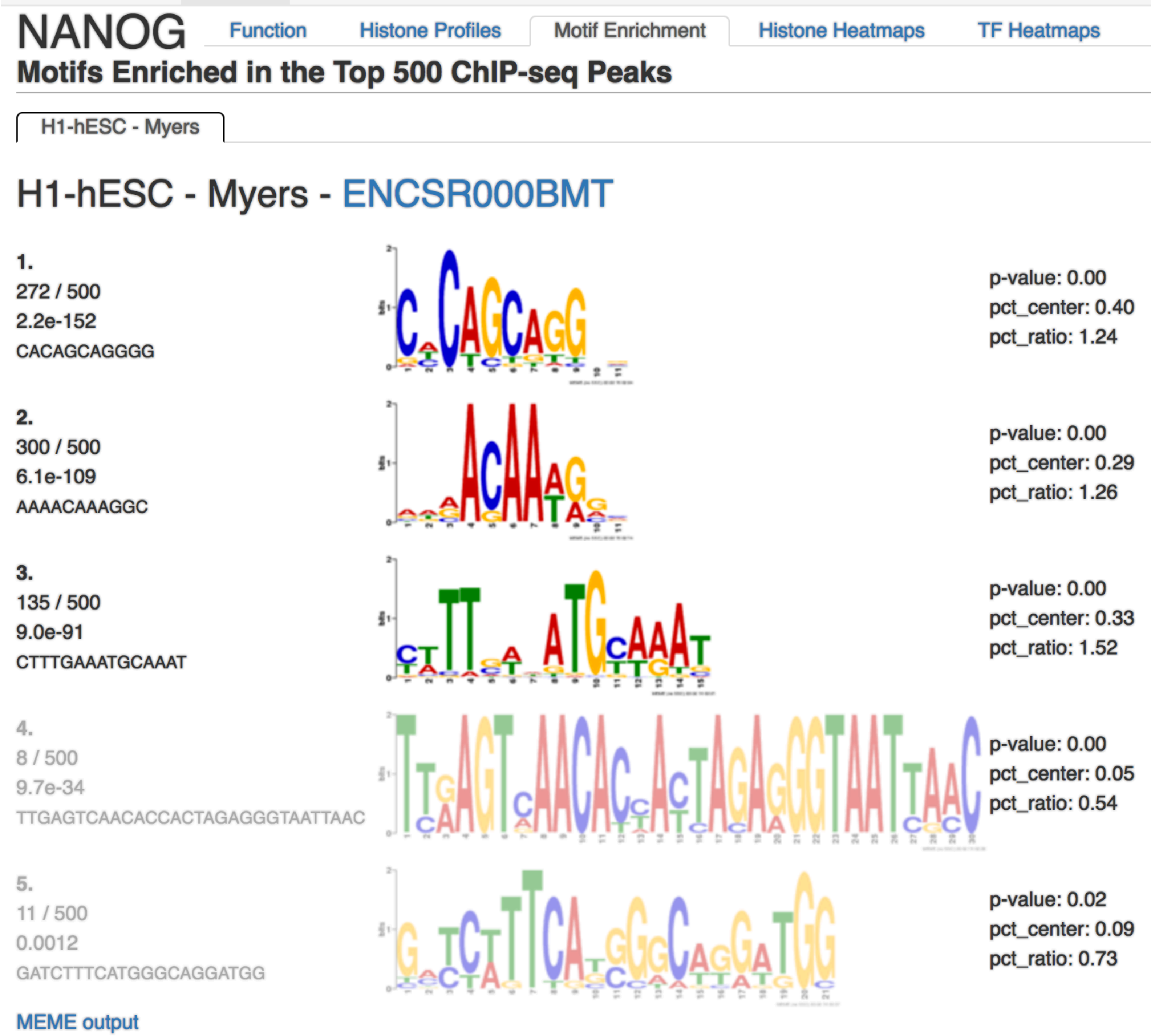
H1-hESC
Myers
H3K4me2
ENCSR000ANC
r=0.22

H1-hESC
Myers
H3K4me1
ENCSR000ANA
r=0.16

H1-hESC
Myers
H3K27ac
ENCSR000ANP
r=0.16

H1-hESC
Myers
H3K4me3
ENCSR000AMG
r=0.11

H1-hESC
Myers



NANOG

Motifs Enriched in the Top 500 ChIP-seq Peaks

H1-hESC - Myers - ENCSR000BMT

1.
272 / 500
2.2e-152
CACAGCAGGGG

p-value: 0.00
pct_center: 0.40
pct_ratio: 1.24

2.
300 / 500
6.1e-109
AAAACAAAGGC

p-value: 0.00
pct_center: 0.29
pct_ratio: 1.26

3.
135 / 500
9.0e-91
CTTTGAAATGCAAAT

p-value: 0.00
pct_center: 0.33
pct_ratio: 1.52

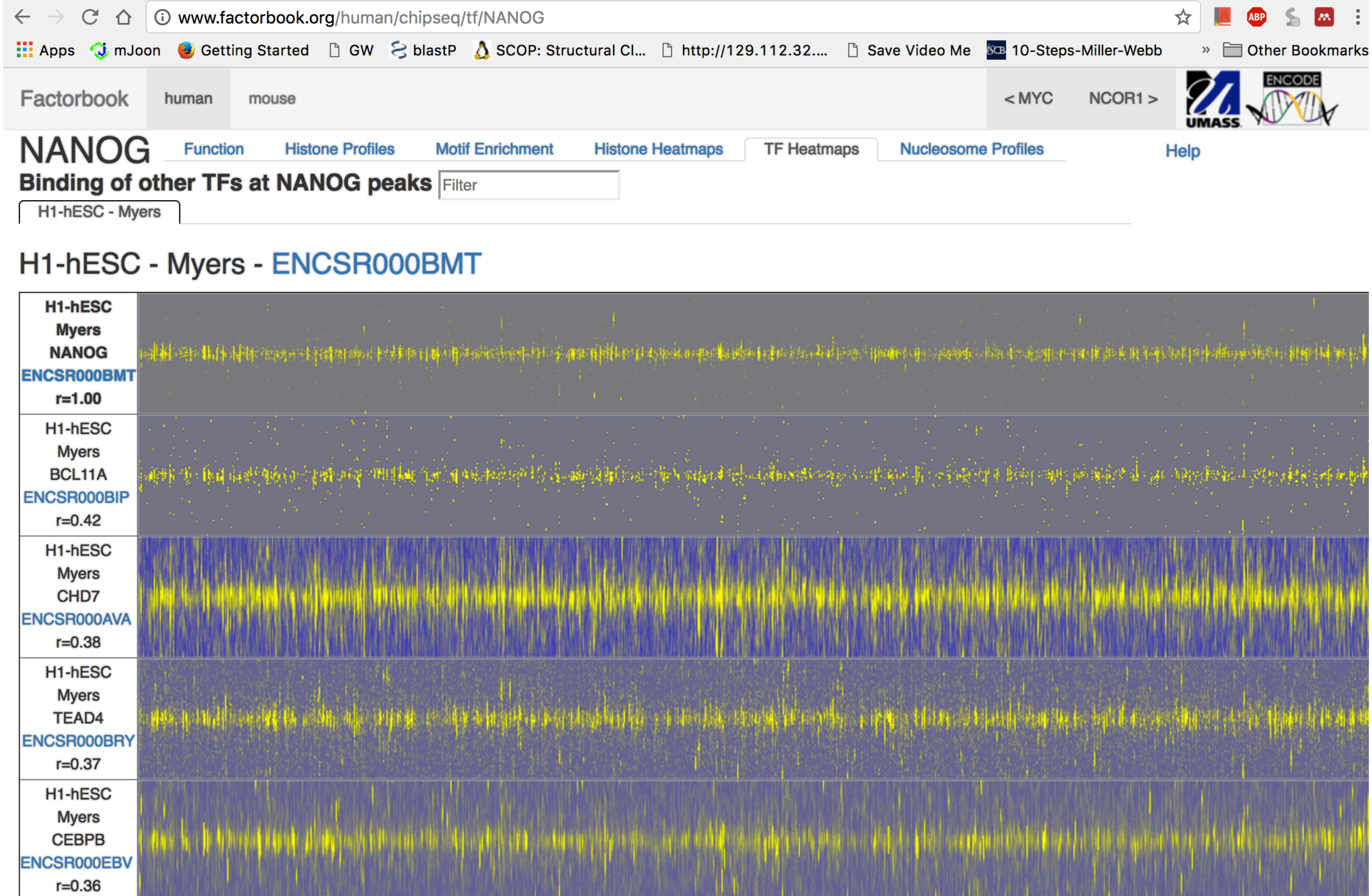
4.
8 / 500
9.7e-34
TTGAGTCAACACCACTAGAGGGTAATTAAC

p-value: 0.00
pct_center: 0.05
pct_ratio: 0.54

5.
11 / 500
0.0012
GATCTTTCATGGGCAGGATGG

p-value: 0.02
pct_center: 0.09
pct_ratio: 0.73

MEME output



NANOG

Binding of other TFs at NANOG peaks

H1-hESC - Myers - ENCSR000BMT

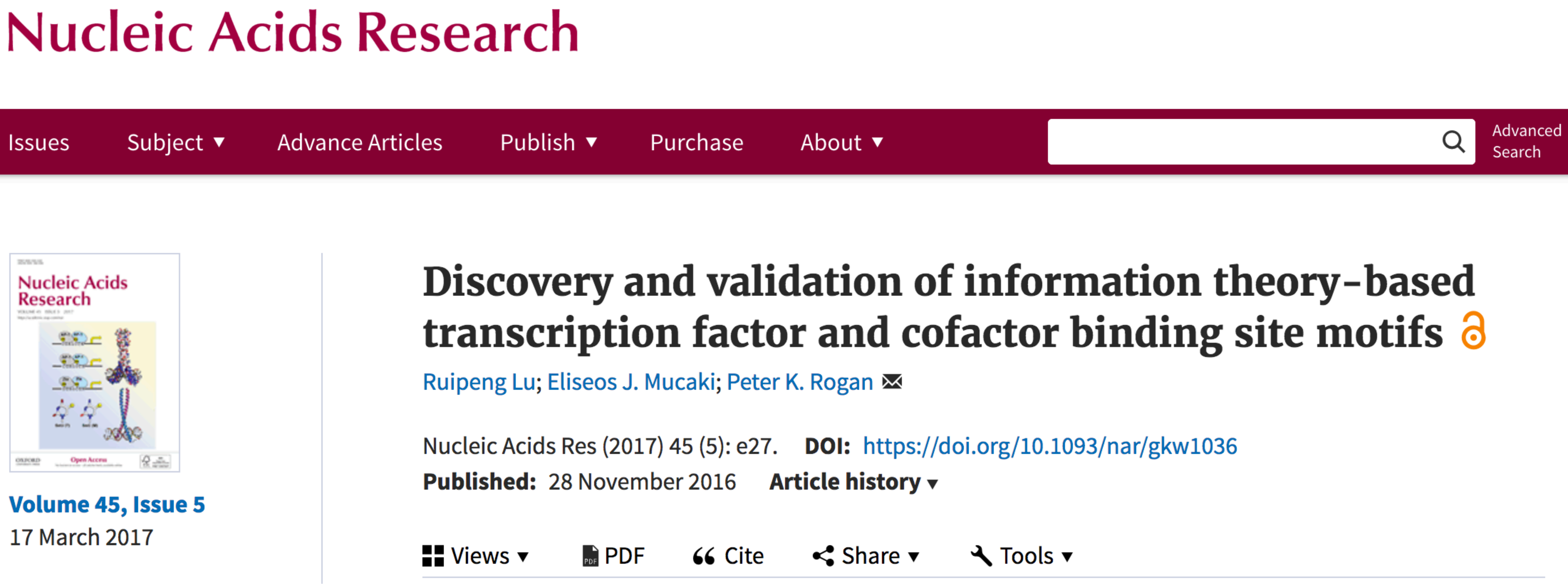
H1-hESC
Myers
NANOG
ENCSR000BMT
r=1.00

H1-hESC
Myers
BCL11A
ENCSR000BIP
r=0.42

H1-hESC
Myers
CHD7
ENCSR000AVA
r=0.38

H1-hESC
Myers
TEAD4
ENCSR000BRY
r=0.37

H1-hESC
Myers
CEBPB
ENCSR000EBV
r=0.36



Nucleic Acids Research

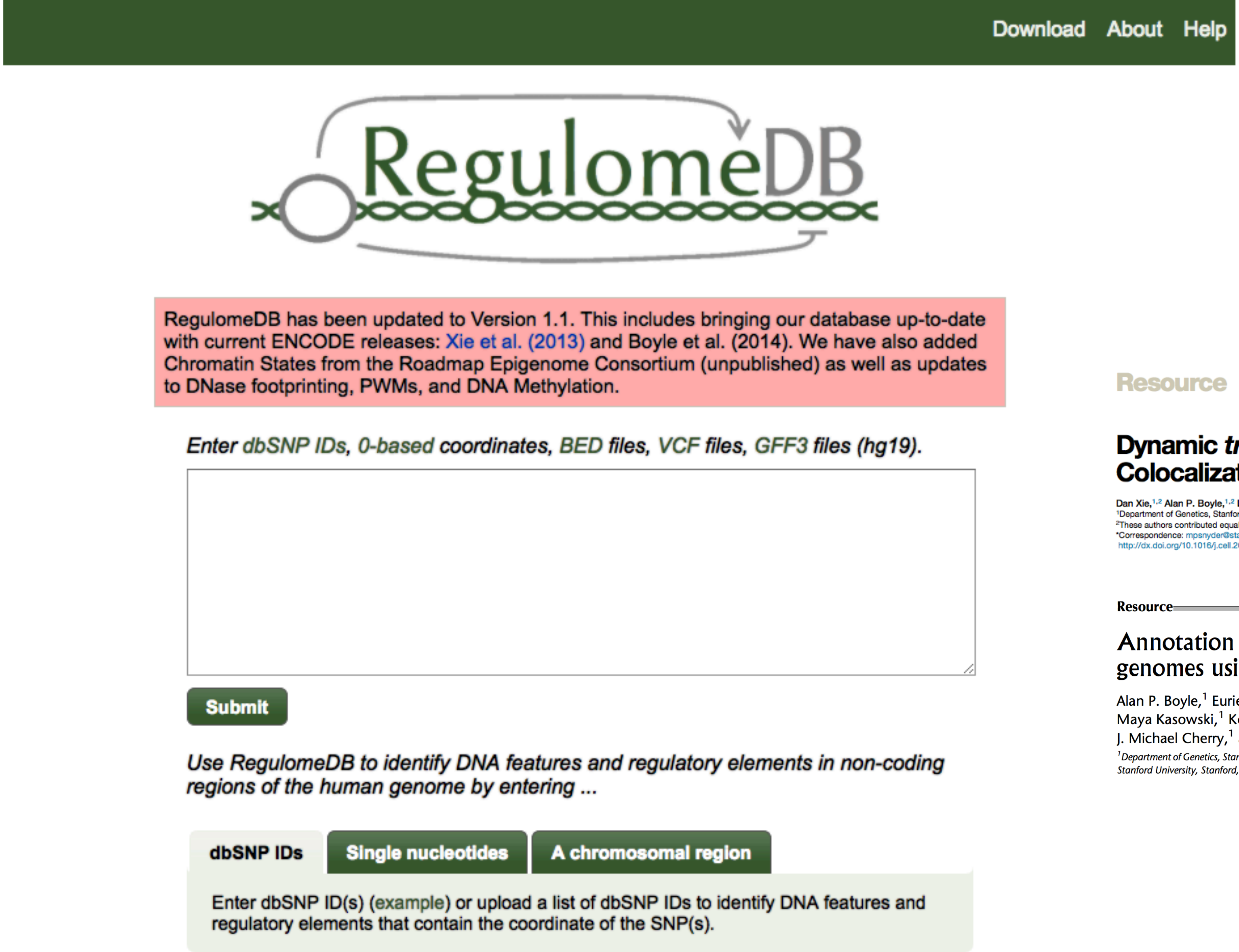
Volume 45, Issue 5
17 March 2017

Discovery and validation of information theory-based transcription factor and cofactor binding site motifs

Ruipeng Lu; Eliseos J. Mucaki; Peter K. Rogan

Nucleic Acids Res (2017) 45 (5): e27. **DOI:** https://doi.org/10.1093/nar/gkw1036

Published: 28 November 2016 **Article history** ▾

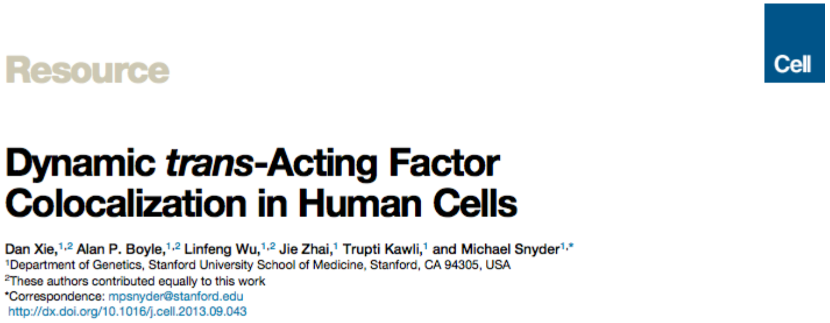


RegulomeDB has been updated to Version 1.1. This includes bringing our database up-to-date with current ENCODE releases: Xie et al. (2013) and Boyle et al. (2014). We have also added Chromatin States from the Roadmap Epigenome Consortium (unpublished) as well as updates to DNase footprinting, PWMs, and DNA Methylation.

Enter dbSNP IDs, 0-based coordinates, BED files, VCF files, GFF3 files (hg19).

Use RegulomeDB to identify DNA features and regulatory elements in non-coding regions of the human genome by entering ...

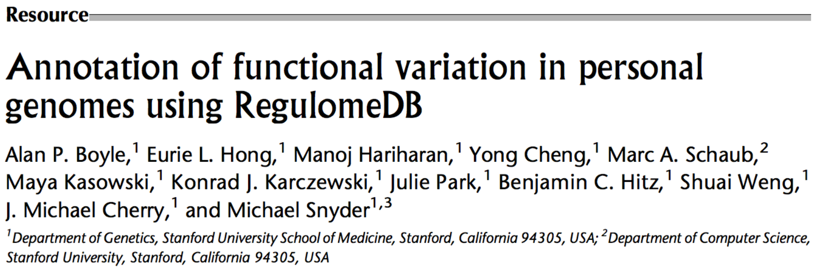
Enter dbSNP ID(s) (example) or upload a list of dbSNP IDs to identify DNA features and regulatory elements that contain the coordinate of the SNP(s).


Resource
Cell

Dynamic *trans*-Acting Factor Colocalization in Human Cells

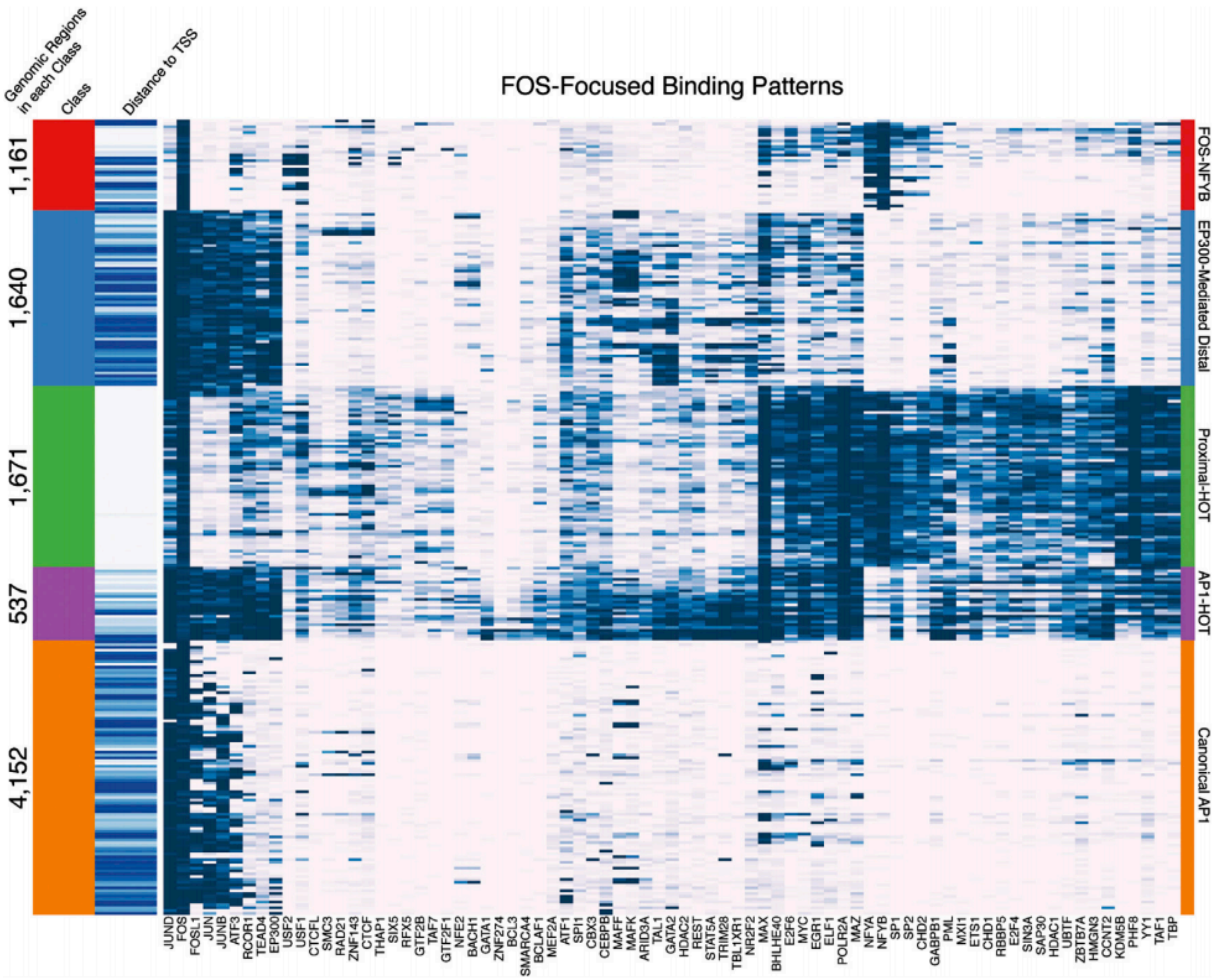
Dan Xie,[1,2] Alan P. Boyle,[1,2] Linfeng Wu,[1,2] Jie Zhai,[1] Trupti Kawli,[1] and Michael Snyder[1,*]
[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]These authors contributed equally to this work
*Correspondence: mpsnyder@stanford.edu

http://dx.doi.org/10.1016/j.cell.2013.09.043



Annotation of functional variation in personal genomes using RegulomeDB

Alan P. Boyle,[1] Eurie L. Hong,[1] Manoj Hariharan,[1] Yong Cheng,[1] Marc A. Schaub,[2] Maya Kasowski,[1] Konrad J. Karczewski,[1] Julie Park,[1] Benjamin C. Hitz,[1] Shuai Weng,[1] J. Michael Cherry,[1] and Michael Snyder[1,3]
[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; [2]Department of Computer Science, Stanford University, Stanford, California 94305, USA



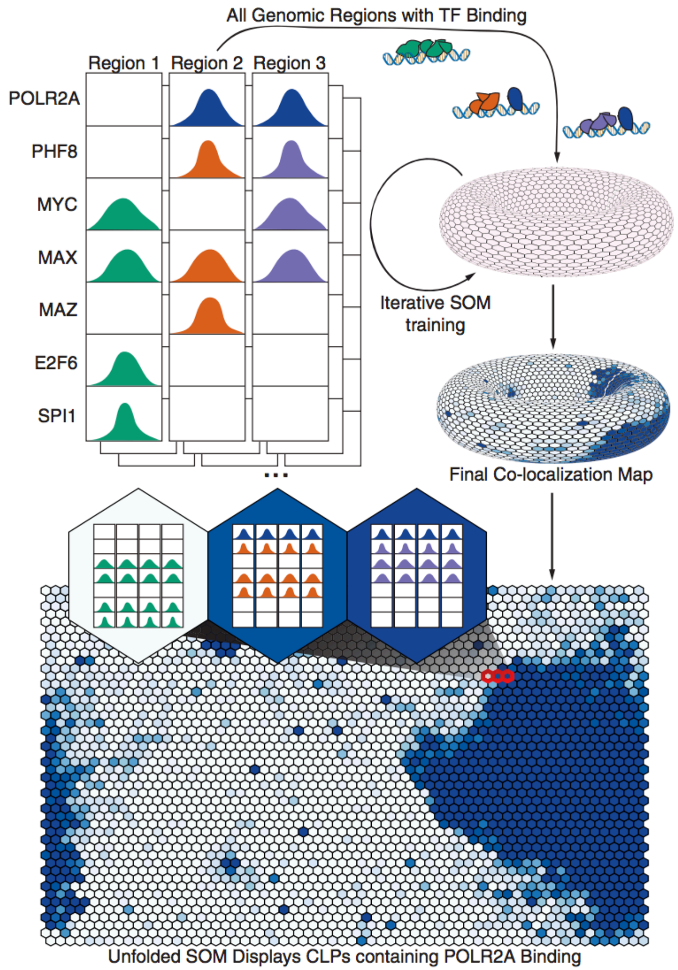
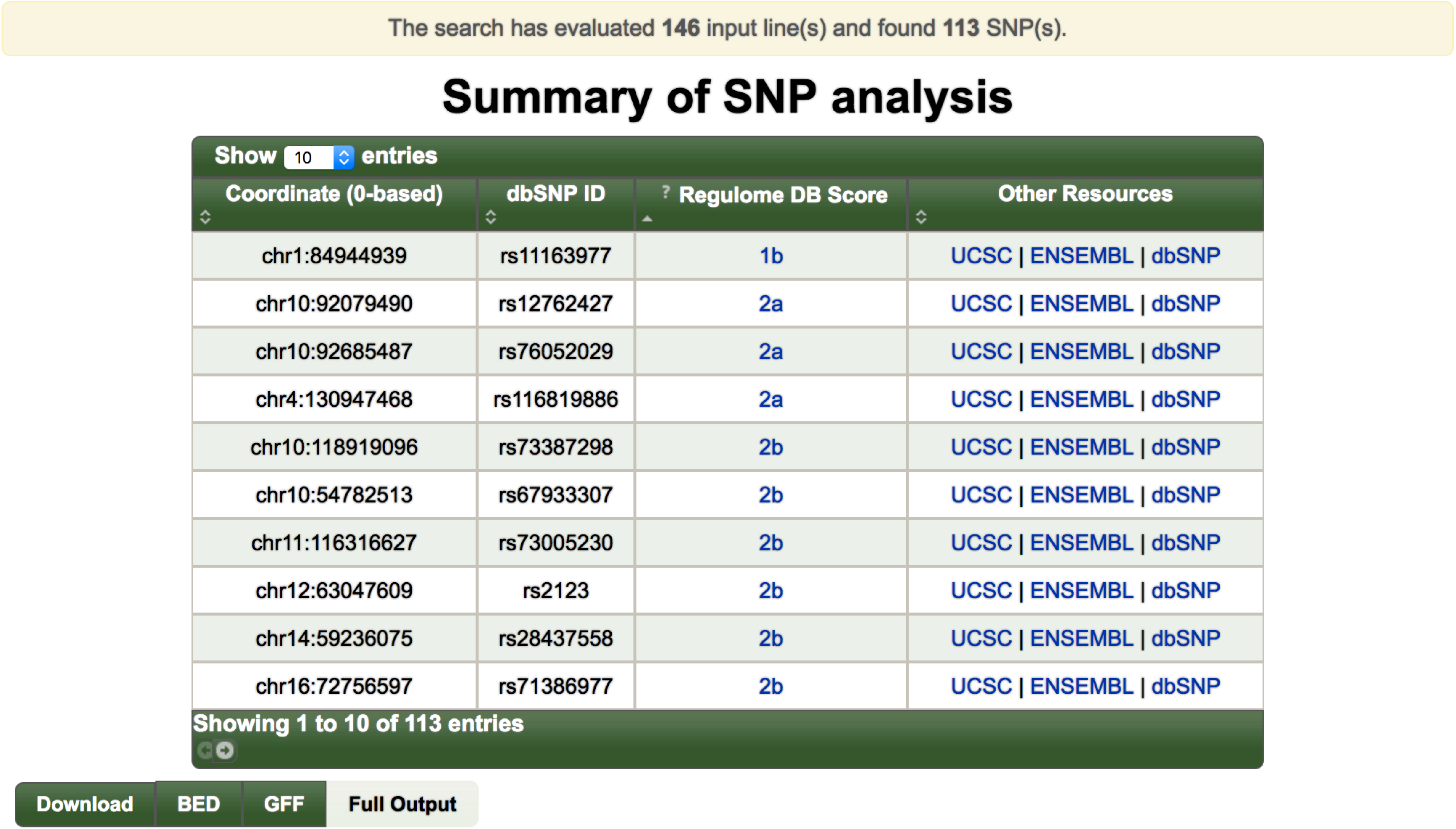

All Genomic Regions with TF Binding
Region 1
Region 2
Region 3
POLR2A
PHF8
MYC
MAX
MAZ
E2F6
SPI1
...
Iterative SOM training
Final Co-localization Map
Unfolded SOM Displays CLPs containing POLR2A Binding

Figure 2. FOS-Focused Binding Patterns

FOS containing colocalization patterns are clustered and shown as each row of a heatmap with blue indicating signal for each colocalized factor (columns). The FOS-focused colocalization patterns fall into five classes: FOS-NFYB, EP300-Mediated Distal, Proximal-HOT, AP1-HOT, and Canonical AP1, which are tagged with different colors. The number of genomic regions and distance to the closest TSS (white = proximal, blue = distal) for each class of colocalization pattern is shown on the left of the heatmap. See also Figure S2.



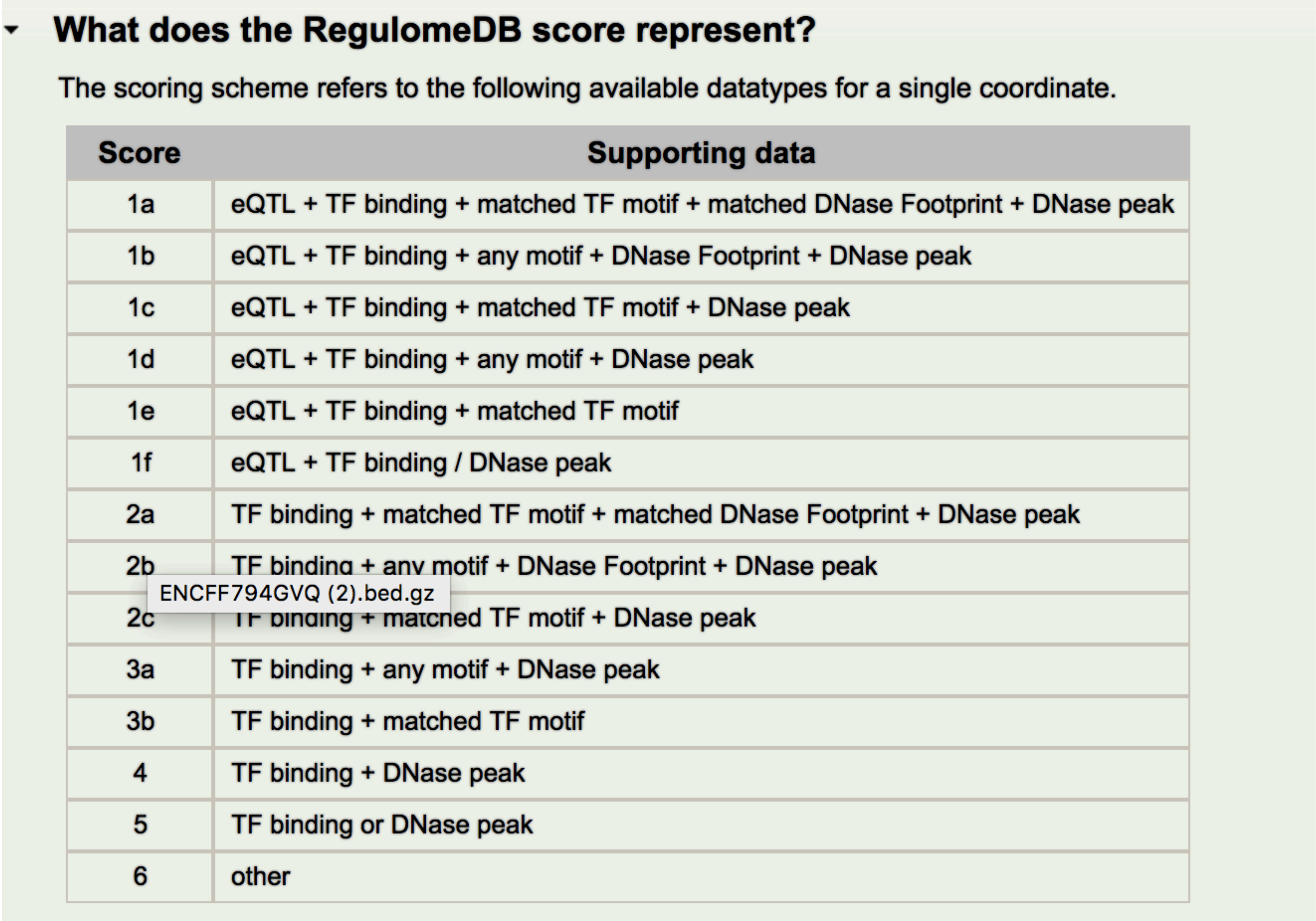
The search has evaluated **146** input line(s) and found **113** SNP(s).

Summary of SNP analysis

Show 10 entries

Coordinate (0-based)	dbSNP ID	? Regulome DB Score	Other Resources
chr1:84944939	rs11163977	1b	UCSC | ENSEMBL | dbSNP
chr10:92079490	rs12762427	2a	UCSC | ENSEMBL | dbSNP
chr10:92685487	rs76052029	2a	UCSC | ENSEMBL | dbSNP
chr4:130947468	rs116819886	2a	UCSC | ENSEMBL | dbSNP
chr10:118919096	rs73387298	2b	UCSC | ENSEMBL | dbSNP
chr10:54782513	rs67933307	2b	UCSC | ENSEMBL | dbSNP
chr11:116316627	rs73005230	2b	UCSC | ENSEMBL | dbSNP
chr12:63047609	rs2123	2b	UCSC | ENSEMBL | dbSNP
chr14:59236075	rs28437558	2b	UCSC | ENSEMBL | dbSNP
chr16:72756597	rs71386977	2b	UCSC | ENSEMBL | dbSNP

Showing 1 to 10 of 113 entries

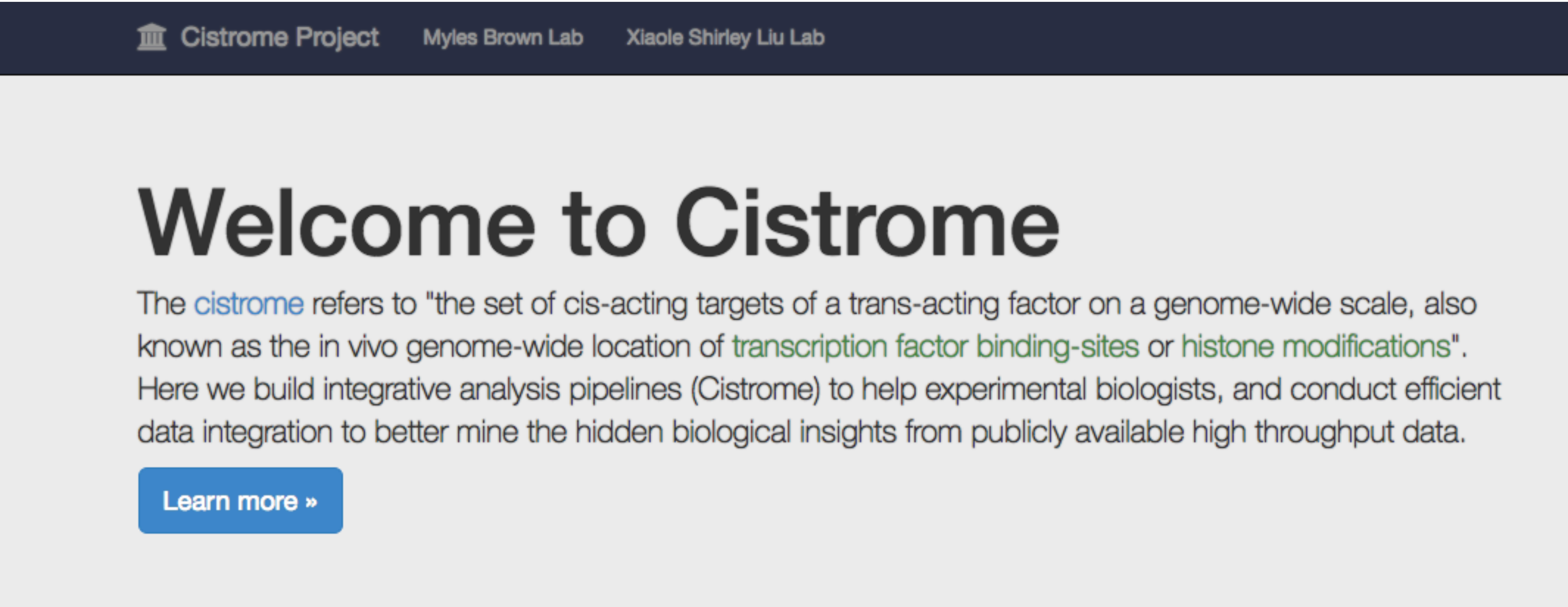


What does the RegulomeDB score represent?

The scoring scheme refers to the following available datatypes for a single coordinate.

Score	Supporting data
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	other

ENCFF794GVQ (2).bed.gz

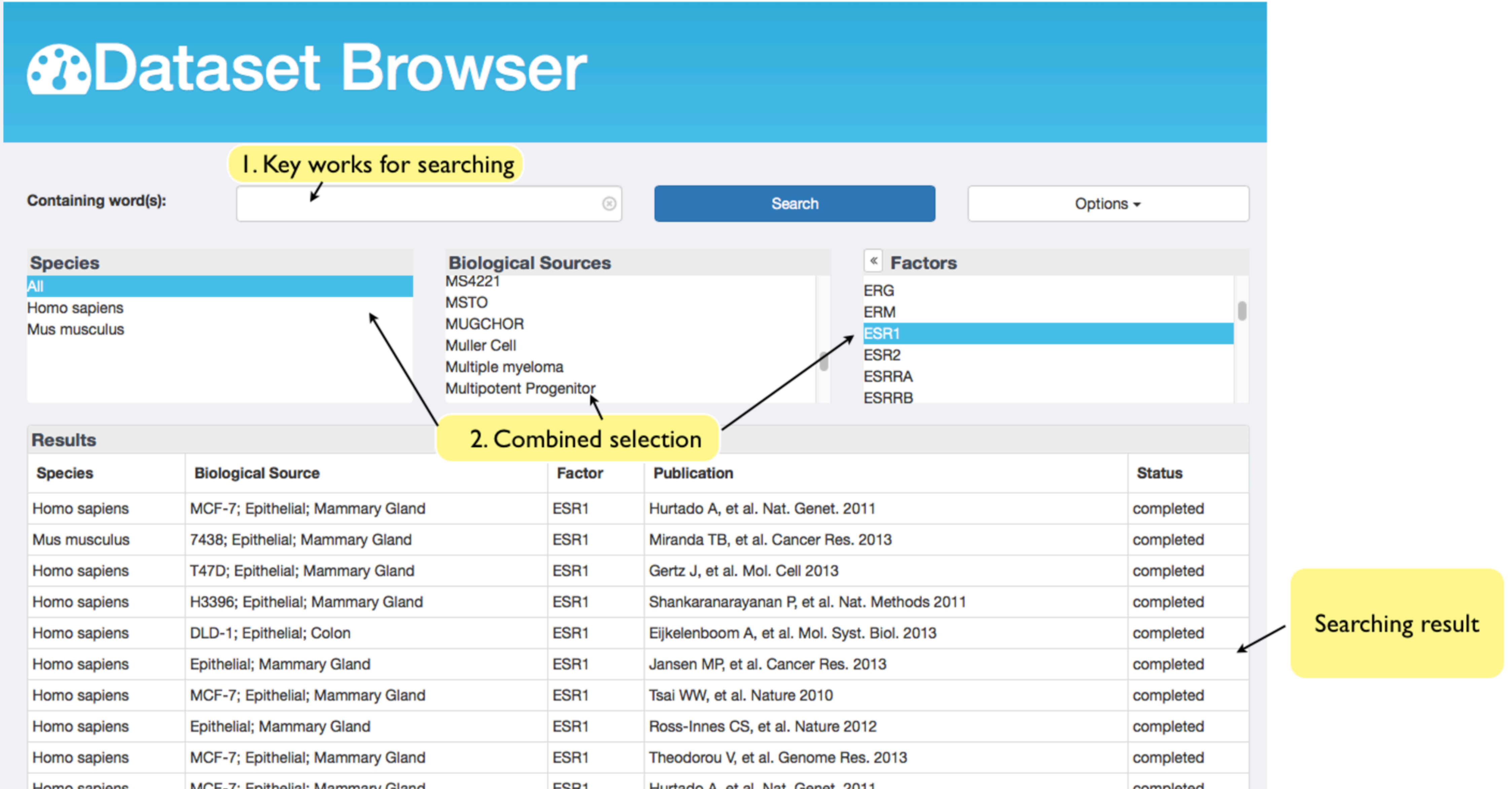


Welcome to Cistrome

The cistrome refers to "the set of cis-acting targets of a trans-acting factor on a genome-wide scale, also known as the in vivo genome-wide location of transcription factor binding-sites or histone modifications". Here we build integrative analysis pipelines (Cistrome) to help experimental biologists, and conduct efficient data integration to better mine the hidden biological insights from publicly available high throughput data.

1. Search your interesting assay, cell or tissue

Cistrome DB contains two options for searching the database,One is to select assay, species or biological sources, The other is based on advanced searching box. Searching result will list with a table of matched datasets. User can view detail data annotations; analysis result and QC metric by click dataset.



I. Key works for searching

2. Combined selection

Searching result

Species

All
Homo sapiens
Mus musculus

Biological Sources

MS4221
MSTO
MUGCHOR
Muller Cell
Multiple myeloma
Multipotent Progenitor

Factors

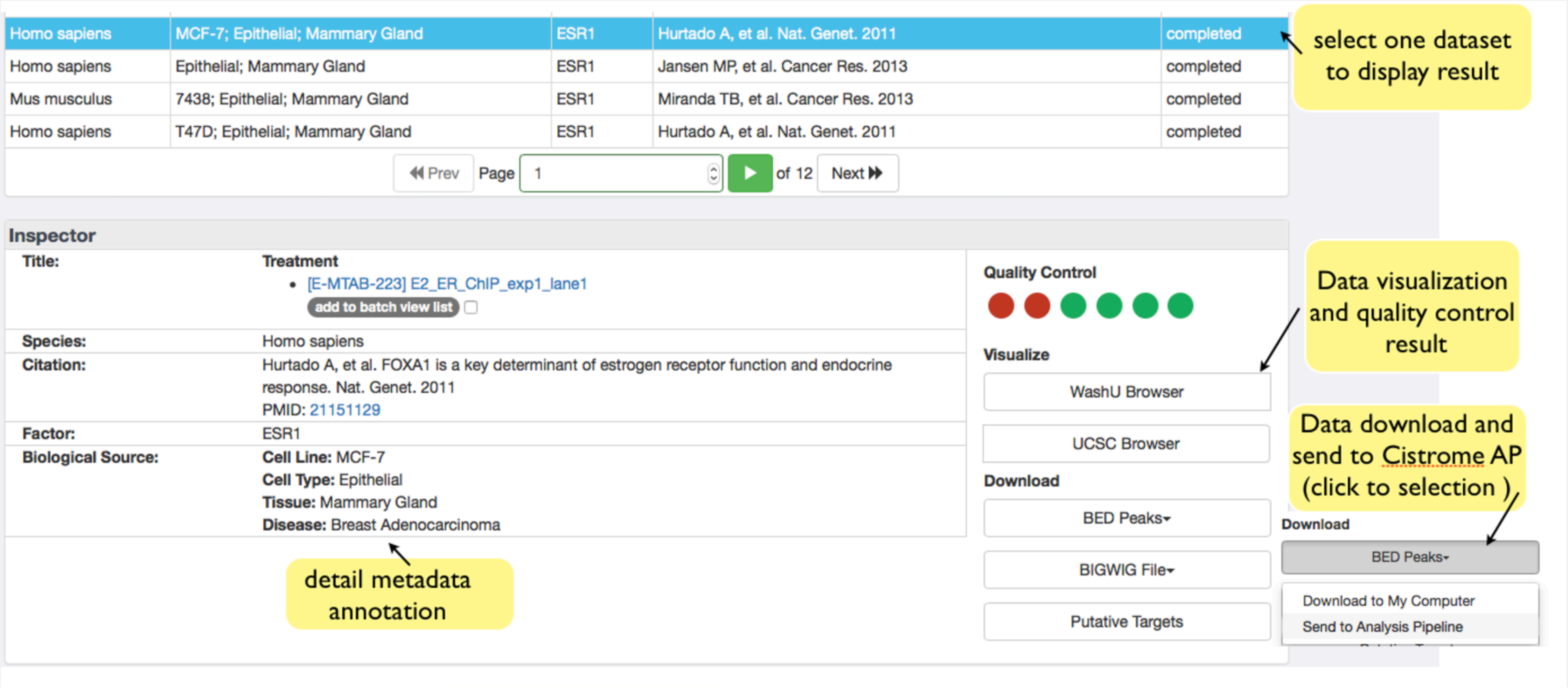
ERG
ERM
ESR1
ESR2
ESRRA
ESRRB

Results

Species	Biological Source	Factor	Publication	Status
Homo sapiens	MCF-7; Epithelial; Mammary Gland	ESR1	Hurtado A, et al. Nat. Genet. 2011	completed
Mus musculus	7438; Epithelial; Mammary Gland	ESR1	Miranda TB, et al. Cancer Res. 2013	completed
Homo sapiens	T47D; Epithelial; Mammary Gland	ESR1	Gertz J, et al. Mol. Cell 2013	completed
Homo sapiens	H3396; Epithelial; Mammary Gland	ESR1	Shankaranarayanan P, et al. Nat. Methods 2011	completed
Homo sapiens	DLD-1; Epithelial; Colon	ESR1	Eijkelenboom A, et al. Mol. Syst. Biol. 2013	completed
Homo sapiens	Epithelial; Mammary Gland	ESR1	Jansen MP, et al. Cancer Res. 2013	completed
Homo sapiens	MCF-7; Epithelial; Mammary Gland	ESR1	Tsai WW, et al. Nature 2010	completed
Homo sapiens	Epithelial; Mammary Gland	ESR1	Ross-Innes CS, et al. Nature 2012	completed
Homo sapiens	MCF-7; Epithelial; Mammary Gland	ESR1	Theodorou V, et al. Genome Res. 2013	completed
Homo sapiens	MCF-7; Epithelial; Mammary Gland	ESR1	Hurtado A, et al. Nat. Genet. 2011	completed

2. Result page for individual dataset

Each ChIP-seq and DNase-seq sample have a unique dataset ID, Cistrome DB comprises manually curated metadata annotations for each dataset, including species, factors, biological source, publication and process status. After clicking interested dataset, result page for individual sample will shows as follows. Result page contains detail metadata annotations, quality control report, analysis result and download section. User can also send data to our Cistrome analysis pipeline for subsequential analysis. Details explanation of QC sees in ChiLin document.



Homo sapiens	MCF-7; Epithelial; Mammary Gland	ESR1	Hurtado A, et al. Nat. Genet. 2011	completed
Homo sapiens	Epithelial; Mammary Gland	ESR1	Jansen MP, et al. Cancer Res. 2013	completed
Mus musculus	7438; Epithelial; Mammary Gland	ESR1	Miranda TB, et al. Cancer Res. 2013	completed
Homo sapiens	T47D; Epithelial; Mammary Gland	ESR1	Hurtado A, et al. Nat. Genet. 2011	completed

select one dataset to display result

Inspector

Title: **Treatment**

- [E-MTAB-223] E2_ER_ChIP_exp1_lane1

Species: Homo sapiens

Citation: Hurtado A, et al. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat. Genet. 2011
PMID: 21151129

Factor: ESR1

Biological Source: **Cell Line:** MCF-7
Cell Type: Epithelial
Tissue: Mammary Gland
Disease: Breast Adenocarcinoma

detail metadata annotation

Quality Control

● ● ● ● ● ●

Visualize

Download

Data visualization and quality control result

Data download and send to Cistrome AP (click to selection)

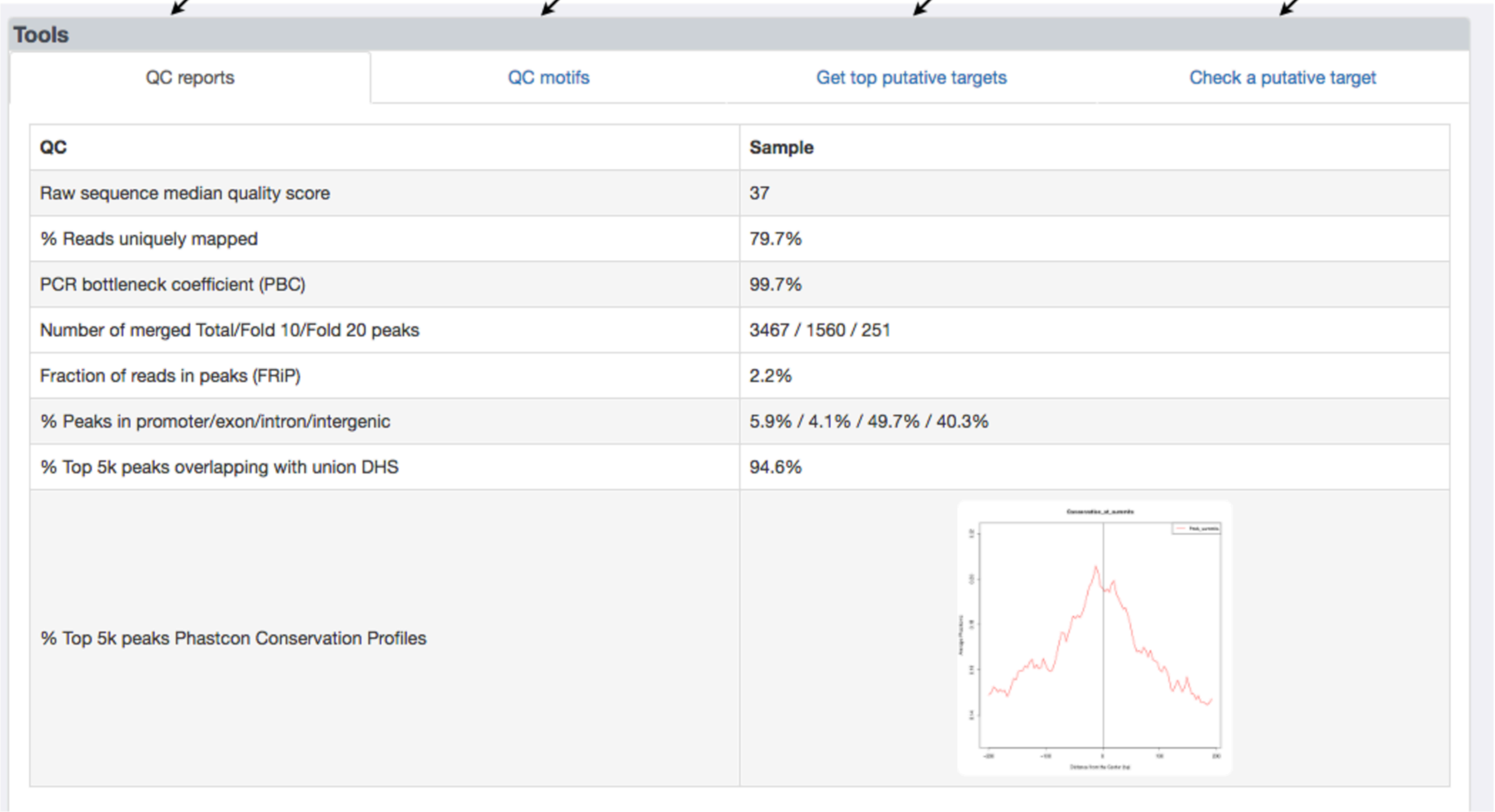
Download

QC report

motif analysis for TF ChIP-seq data

Putative targets (BETA result)

Searching putative target



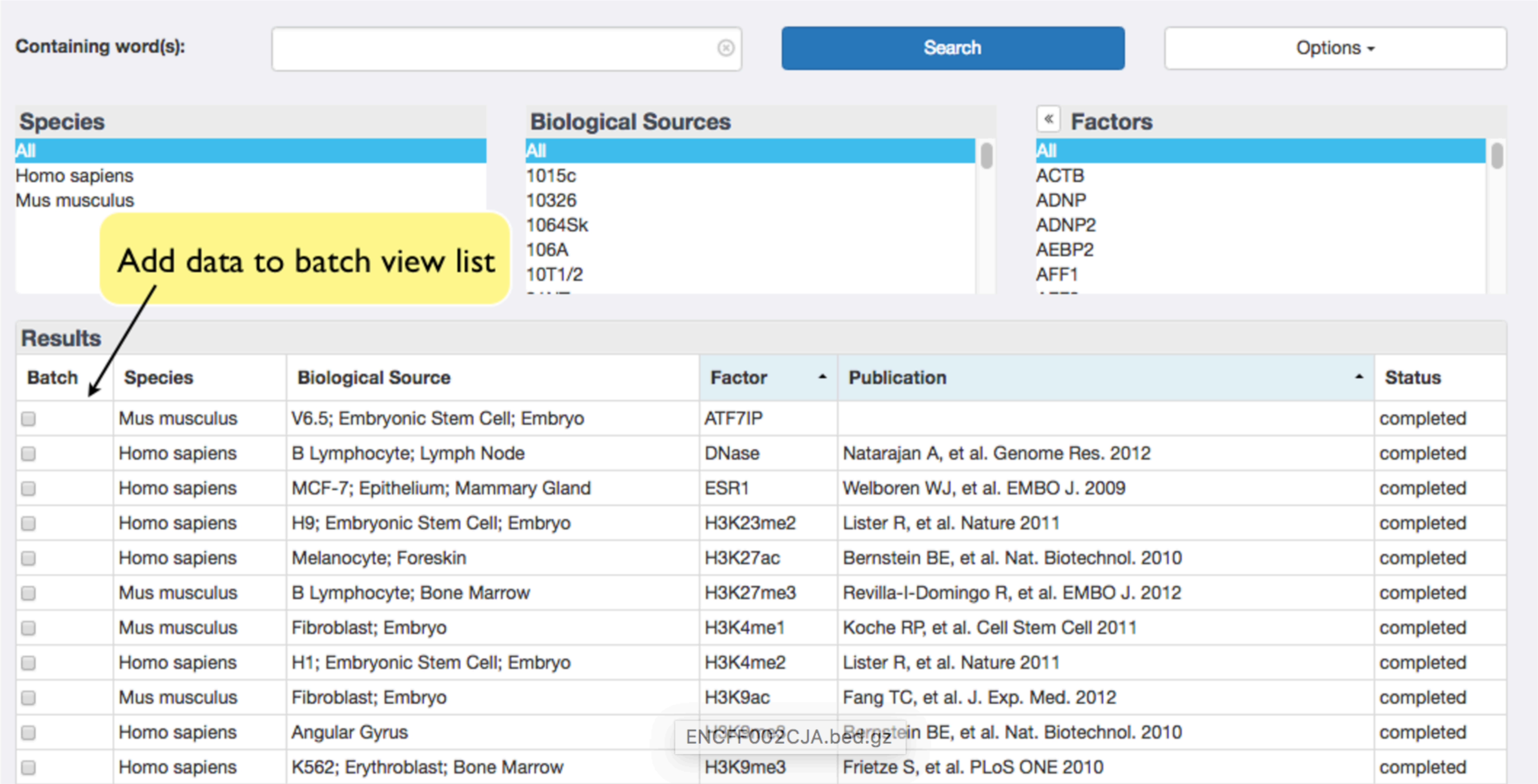
Tools

QC reports | QC motifs | Get top putative targets | Check a putative target

QC	Sample
Raw sequence median quality score	37
% Reads uniquely mapped	79.7%
PCR bottleneck coefficient (PBC)	99.7%
Number of merged Total/Fold 10/Fold 20 peaks	3467 / 1560 / 251
Fraction of reads in peaks (FRiP)	2.2%
% Peaks in promoter/exon/intron/intergenic	5.9% / 4.1% / 49.7% / 40.3%
% Top 5k peaks overlapping with union DHS	94.6%
% Top 5k peaks Phastcon Conservation Profiles	

3. Batch samples genome browser display

Besides metadata collection and data processing. CistromeDB also provide users batch data view function. After selected interested dataset, user can send data to genome browser for association study, such as co-factors, relationship between chromatin regulators and histone modifications.



Add data to batch view list

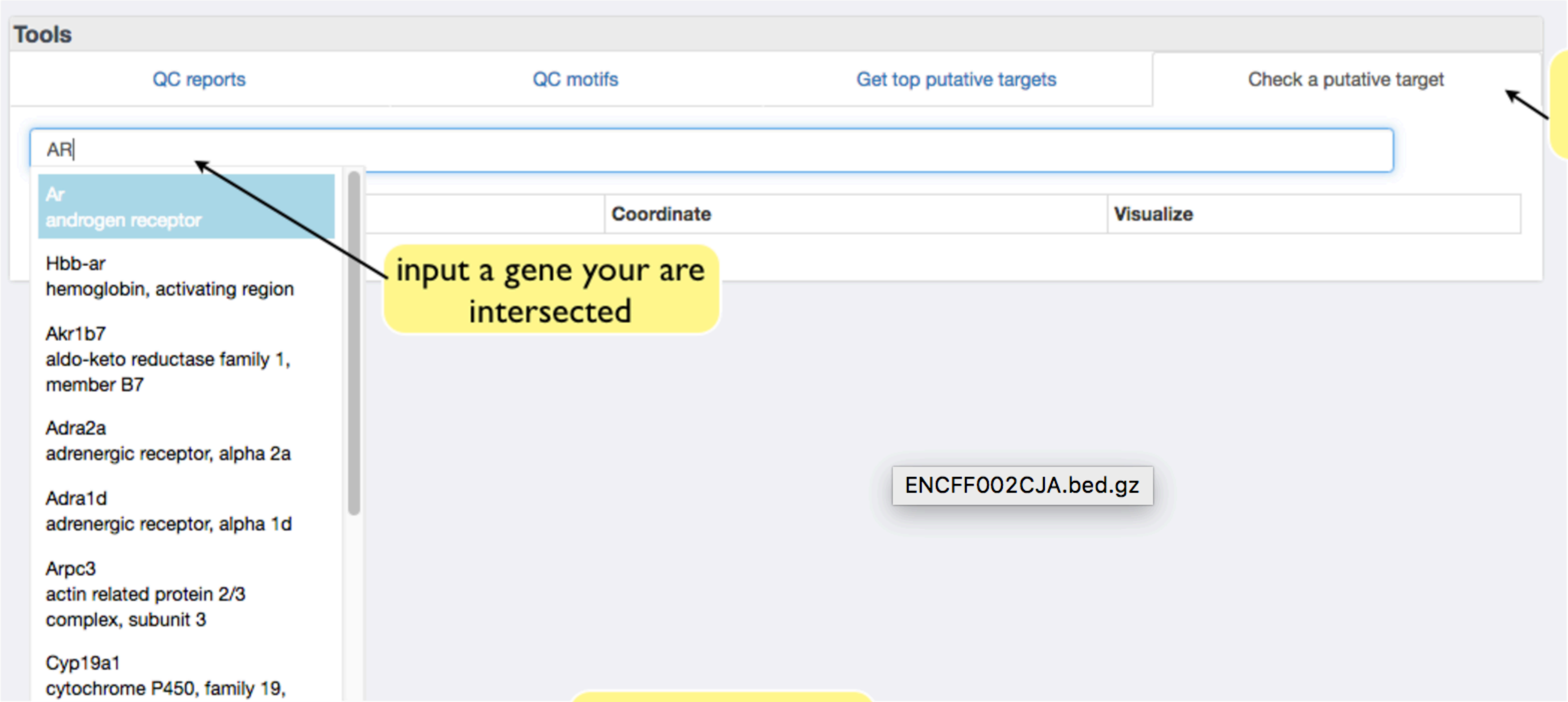
Results

Batch	Species	Biological Source	Factor	Publication	Status
□	Mus musculus	V6.5; Embryonic Stem Cell; Embryo	ATF7IP		completed
□	Homo sapiens	B Lymphocyte; Lymph Node	DNase	Natarajan A, et al. Genome Res. 2012	completed
□	Homo sapiens	MCF-7; Epithelium; Mammary Gland	ESR1	Welboren WJ, et al. EMBO J. 2009	completed
□	Homo sapiens	H9; Embryonic Stem Cell; Embryo	H3K23me2	Lister R, et al. Nature 2011	completed
□	Homo sapiens	Melanocyte; Foreskin	H3K27ac	Bernstein BE, et al. Nat. Biotechnol. 2010	completed
□	Mus musculus	B Lymphocyte; Bone Marrow	H3K27me3	Revilla-I-Domingo R, et al. EMBO J. 2012	completed
□	Mus musculus	Fibroblast; Embryo	H3K4me1	Koche RP, et al. Cell Stem Cell 2011	completed
□	Homo sapiens	H1; Embryonic Stem Cell; Embryo	H3K4me2	Lister R, et al. Nature 2011	completed
□	Mus musculus	Fibroblast; Embryo	H3K9ac	Fang TC, et al. J. Exp. Med. 2012	completed
□	Homo sapiens	Angular Gyrus	H3K9me3	Bernstein BE, et al. Nat. Biotechnol. 2010	completed
□	Homo sapiens	K562; Erythroblast; Bone Marrow	H3K9me3	Frietze S, et al. PLoS ONE 2010	completed

ENCFF002CJA.bed.gz

4. Get ChIP-seq putative targets and search interesting targets

To help user quickly locate putative targets. Cistrome DB provided two options for putative targets view. On the one hand, user can get whole list of putative targets by click "get top putative targets" menu. On the other hand, user can also enter the gene symbols to search intersected target.



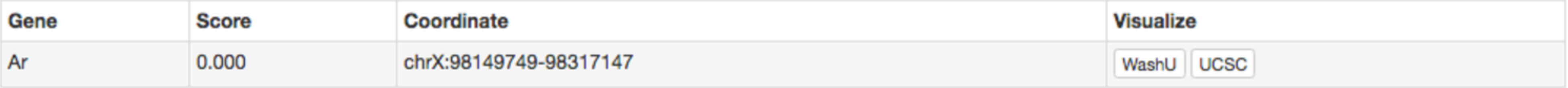
Tools
QC reports
QC motifs
Get top putative targets
Check a putative target
Coordinate
Visualize
Ar
androgen receptor
Hbb-ar
hemoglobin, activating region
Akr1b7
aldo-keto reductase family 1, member B7
Adra2a
adrenergic receptor, alpha 2a
Adra1d
adrenergic receptor, alpha 1d
Arpc3
actin related protein 2/3 complex, subunit 3
Cyp19a1
cytochrome P450, family 19,
ENCFF002CJA.bed.gz

AR

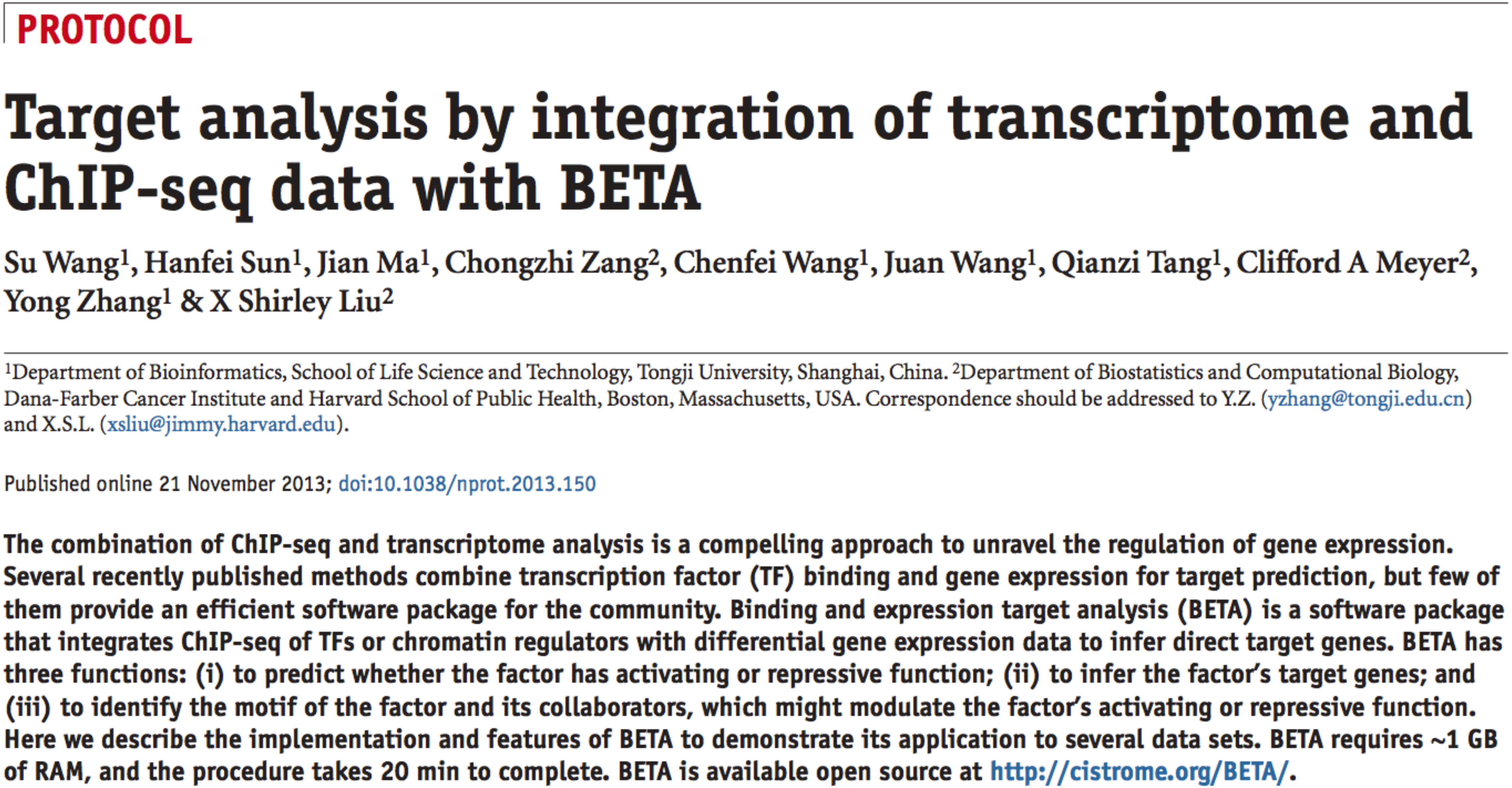
Check a putative target function

input a gene your are intersected

searching result



Gene	Score	Coordinate	Visualize
Ar	0.000	chrX:98149749-98317147	WashU UCSC



PROTOCOL

Target analysis by integration of transcriptome and ChIP-seq data with BETA

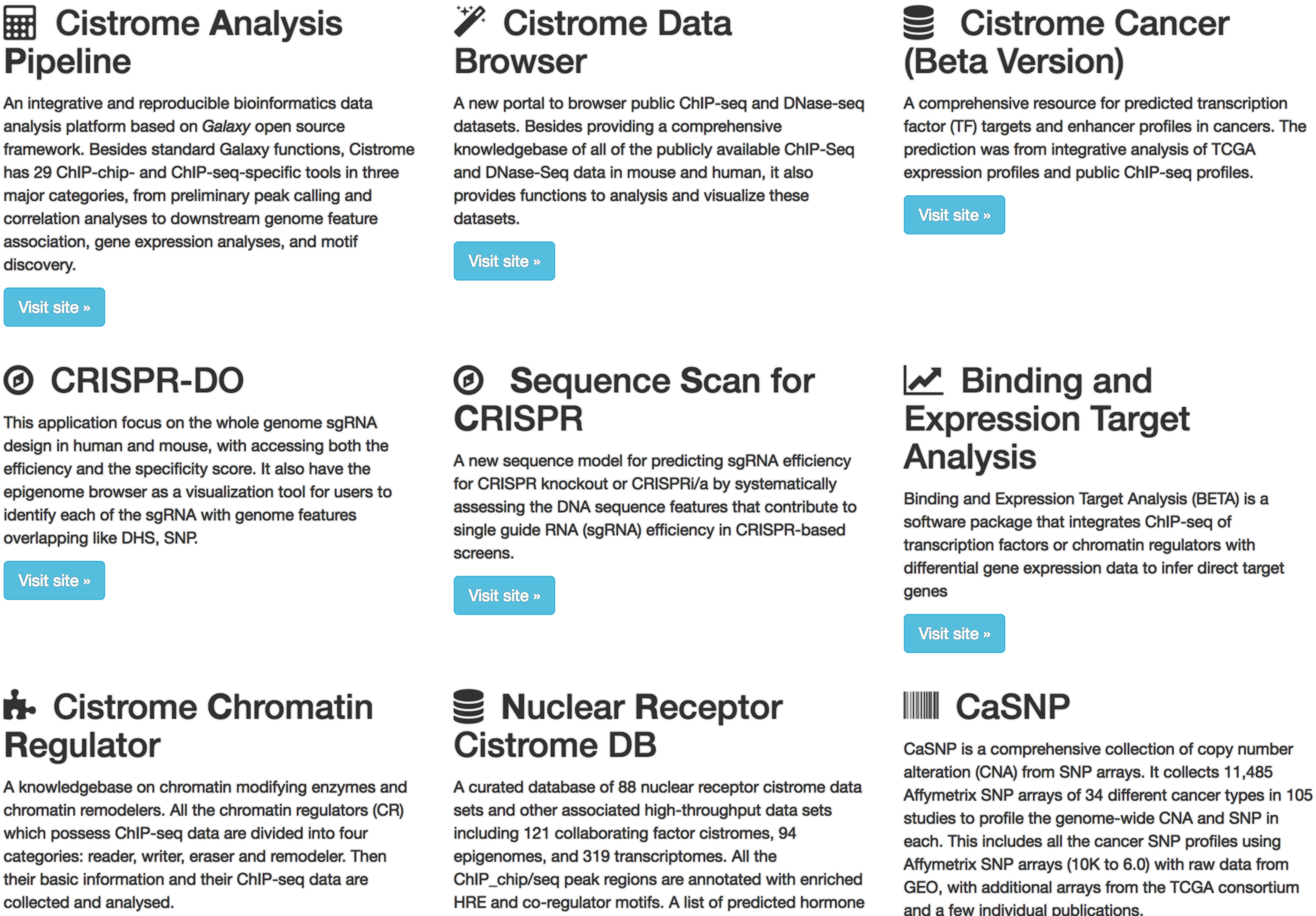
Su Wang[1], Hanfei Sun[1], Jian Ma[1], Chongzhi Zang[2], Chenfei Wang[1], Juan Wang[1], Qianzi Tang[1], Clifford A Meyer[2], Yong Zhang[1] & X Shirley Liu[2]

[1]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, China. [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to Y.Z. (yzhang@tongji.edu.cn) and X.S.L. (xsliu@jimmy.harvard.edu).

Published online 21 November 2013; doi:10.1038/nprot.2013.150

The combination of ChIP-seq and transcriptome analysis is a compelling approach to unravel the regulation of gene expression. Several recently published methods combine transcription factor (TF) binding and gene expression for target prediction, but few of them provide an efficient software package for the community. Binding and expression target analysis (BETA) is a software package that integrates ChIP-seq of TFs or chromatin regulators with differential gene expression data to infer direct target genes. BETA has three functions: (i) to predict whether the factor has activating or repressive function; (ii) to infer the factor's target genes; and (iii) to identify the motif of the factor and its collaborators, which might modulate the factor's activating or repressive function. Here we describe the implementation and features of BETA to demonstrate its application to several data sets. BETA requires ~1 GB of RAM, and the procedure takes 20 min to complete. BETA is available open source at http://cistrome.org/BETA/.

https://www.ncbi.nlm.nih.gov/pubmed/24263090



Cistrome Analysis Pipeline

An integrative and reproducible bioinformatics data analysis platform based on *Galaxy* open source framework. Besides standard Galaxy functions, Cistrome has 29 ChIP-chip- and ChIP-seq-specific tools in three major categories, from preliminary peak calling and correlation analyses to downstream genome feature association, gene expression analyses, and motif discovery.

Visit site »

Cistrome Data Browser

A new portal to browser public ChIP-seq and DNase-seq datasets. Besides providing a comprehensive knowledgebase of all of the publicly available ChIP-Seq and DNase-Seq data in mouse and human, it also provides functions to analysis and visualize these datasets.

Visit site »

Cistrome Cancer (Beta Version)

A comprehensive resource for predicted transcription factor (TF) targets and enhancer profiles in cancers. The prediction was from integrative analysis of TCGA expression profiles and public ChIP-seq profiles.

Visit site »

CRISPR-DO

This application focus on the whole genome sgRNA design in human and mouse, with accessing both the efficiency and the specificity score. It also have the epigenome browser as a visualization tool for users to identify each of the sgRNA with genome features overlapping like DHS, SNP.

Visit site »

Sequence Scan for CRISPR

A new sequence model for predicting sgRNA efficiency for CRISPR knockout or CRISPRi/a by systematically assessing the DNA sequence features that contribute to single guide RNA (sgRNA) efficiency in CRISPR-based screens.

Visit site »

Binding and Expression Target Analysis

Binding and Expression Target Analysis (BETA) is a software package that integrates ChIP-seq of transcription factors or chromatin regulators with differential gene expression data to infer direct target genes

Visit site »

Cistrome Chromatin Regulator

A knowledgebase on chromatin modifying enzymes and chromatin remodelers. All the chromatin regulators (CR) which possess ChIP-seq data are divided into four categories: reader, writer, eraser and remodeler. Then their basic information and their ChIP-seq data are collected and analysed.

Nuclear Receptor Cistrome DB

A curated database of 88 nuclear receptor cistrome data sets and other associated high-throughput data sets including 121 collaborating factor cistromes, 94 epigenomes, and 319 transcriptomes. All the ChIP_chip/seq peak regions are annotated with enriched HRE and co-regulator motifs. A list of predicted hormone

CaSNP

CaSNP is a comprehensive collection of copy number alteration (CNA) from SNP arrays. It collects 11,485 Affymetrix SNP arrays of 34 different cancer types in 105 studies to profile the genome-wide CNA and SNP in each. This includes all the cancer SNP profiles using Affymetrix SNP arrays (10K to 6.0) with raw data from GEO, with additional arrays from the TCGA consortium and a few individual publications.

Acknowledgement

- Peter Fizgerald
- Anand Merchand

- ENCODE Consortium
  - Bing Ren
  - Micheal Snyder
  - Anshul Kundaje