

Web applications for human epigenomics

EPIGEN2018

2018-03-14

Guillaume Devailly

@G_Devailly



Slides available online. Feel free to share.

Why web applications?

- easy to access
- easy to use (?)
- fancy and attracting

→ **Allow users to do things they wouldn't have done otherwise**

Why web applications?

- easy to access
- easy to use (?)
- fancy and attracting

→ **Allow users to do things they wouldn't have done otherwise**

but...

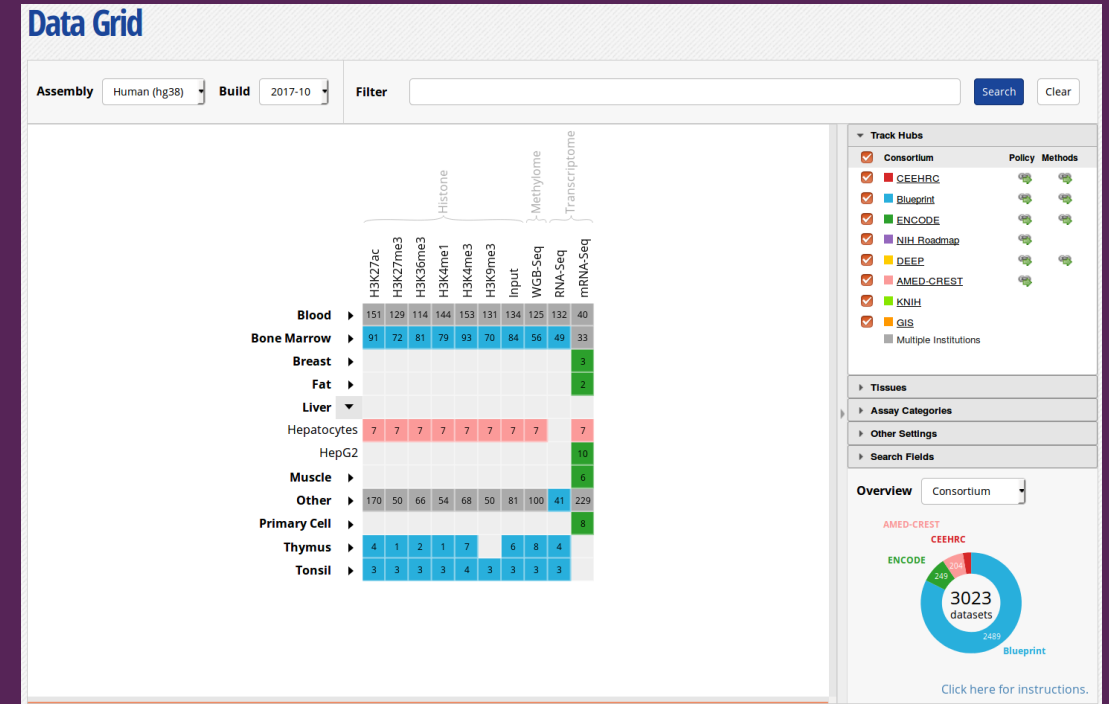
- someone has to develop and maintain them...
- (almost) no flexibility
- apps appear and disappear with a high turnover

Types of web-apps for epigenomics

1. Data portals
2. Genome browsers
3. Compare experiments
4. Analyse

Types of web-apps for epigenomics

1. Data portals
2. Genome browsers
3. Compare experiments
4. Analyse



Data portals

- An eagle-eye view of available data.
- Easier to use than SRA / GEO searches

Data portals

A precursor: ENCODE @ UCSC

ENCODE Encyclopedia of DNA Elements at UCSC 2003 - 2012

Human Data at UCSC
Downloads
Experiment Matrix
Search
Genome Browser (hg19)
Experiment List
Cell Types

Mouse Data at UCSC
Downloads
Experiment Matrix
Search
Genome Browser (mm9)
Experiment List
Cell Types

Metadata Terms
Registered Variables
Antibodies

About

The [Encyclopedia of DNA Elements](#) (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE results from 2007 and later are available from the ENCODE Project Portal, [encodeproject.org](#). This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. **All ENCODE data at UCSC are freely available for download and analysis.**

Explore ENCODE data (2003 - 2012) at UCSC

View ENCODE data (2003 - 2012) in the UCSC Genome Browser

unmaintained since 2012

Data portals

Modern ENCODE data portal

ENCODE Data Encyclopedia Materials & Methods Help Search...

ENCODE: Encyclopedia of DNA Elements

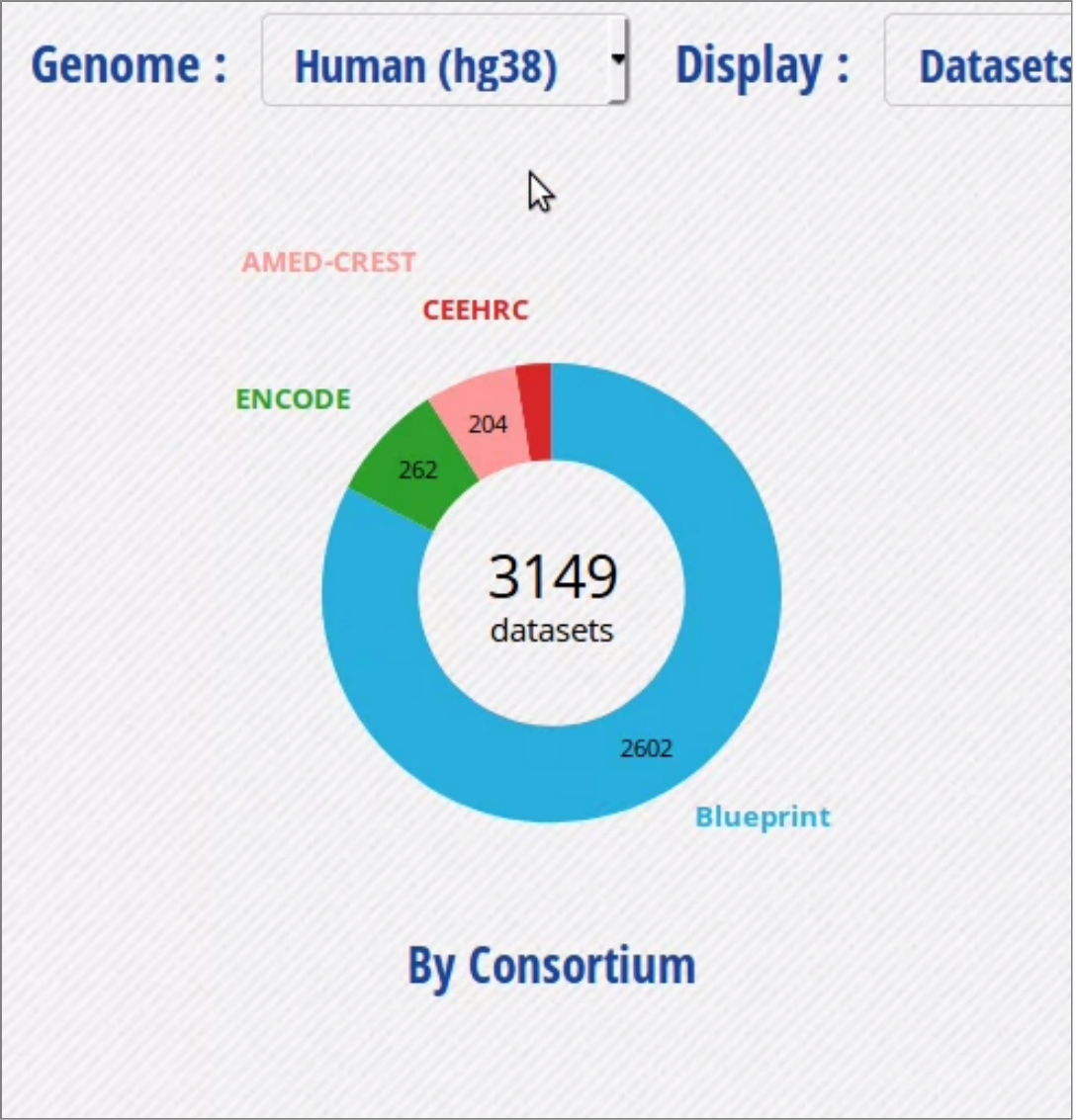
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[Get Started](#)

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

HUMAN MOUSE WORM FLY

International Human Epigenome Consortium (IHEC)



Data Grid

Assembly: Human (hg38) Build: 2017-10 Filter: Search Clear

	Histone					Methylation		Transcriptome		
	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K9me3	WGB-Seq	RNA-Seq	mRNA-Seq		
Blood	151	129	114	144	153	131	134	125	132	40
Bone Marrow	91	72	81	79	93	70	84	56	49	33
Breast										3
Fat										2
Liver										
Hepatocytes	7	7	7	7	7	7	7	7	7	7
HepG2										10
Muscle										6
Other	170	50	66	54	68	50	81	100	41	229
Primary Cell										8
Thymus	4	1	2	1	7		6	8	4	
Tonsil	3	3	3	3	4	3	3	3	3	3

Track Hubs

- Consortium
- CEEHRC
- Blueprint
- ENCODE
- NIH Roadmap
- DEEP
- AMED-CREST
- KNIH
- GIS
- Multiple Institutions

Tissues

Assay Categories

Other Settings

Search Fields

Overview Consortium

3023 datasets

[Click here for instructions.](#)

Data portals

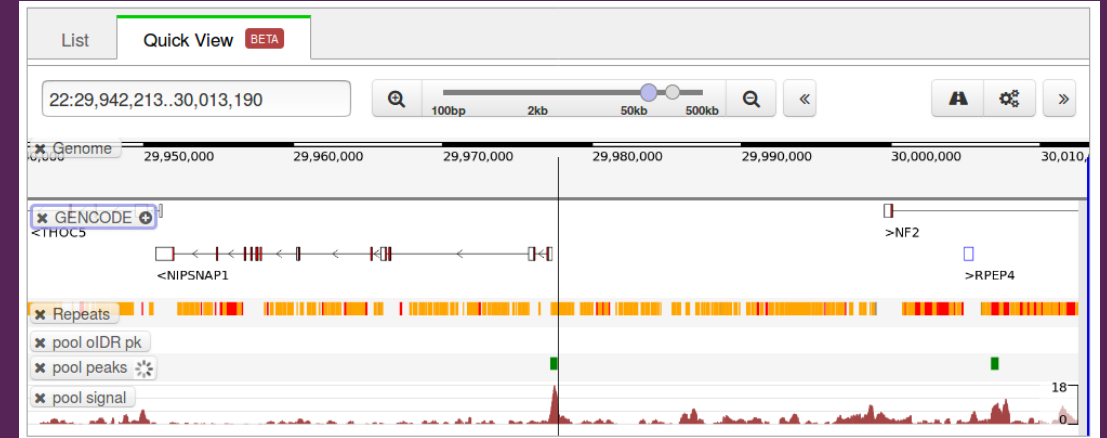
Portals browsing SRA/ENA public datasets:

- codex.stemcells.cam.ac.uk: TF & HisMod ChIP-seq, RNA-seq, DNase-seq in Haematopoietic Cells, or Embryonic Stem Cells
- cistrome.org: TF & HisMod & DNase/ATAC, all cell types & tissues
- ngs-qc.org: Almost everything (TF, HisMod, RNA, MeDP, ...), all cell types & tissues, assess dataset quality

→ Finding data is easy

Types of web-apps for epigenomics

1. Data portals
2. Genome browsers
3. Compare experiments
4. Analyse



Online Genome browsers

- quick look at the data
- looking at a few regions manually
- Sometimes ♥ *integrated to the Data portal* ♥



ENCODE quick view genome browser

ENCODE Data Encyclopedia Materials & Methods Help


EXPERIMENTS / [CHIP-SEQ](#) / [HOMO SAPIENS](#) / [HEK293](#)

Experiment summary for ENCSR348AGV

Supersedes ENCSR000EYD

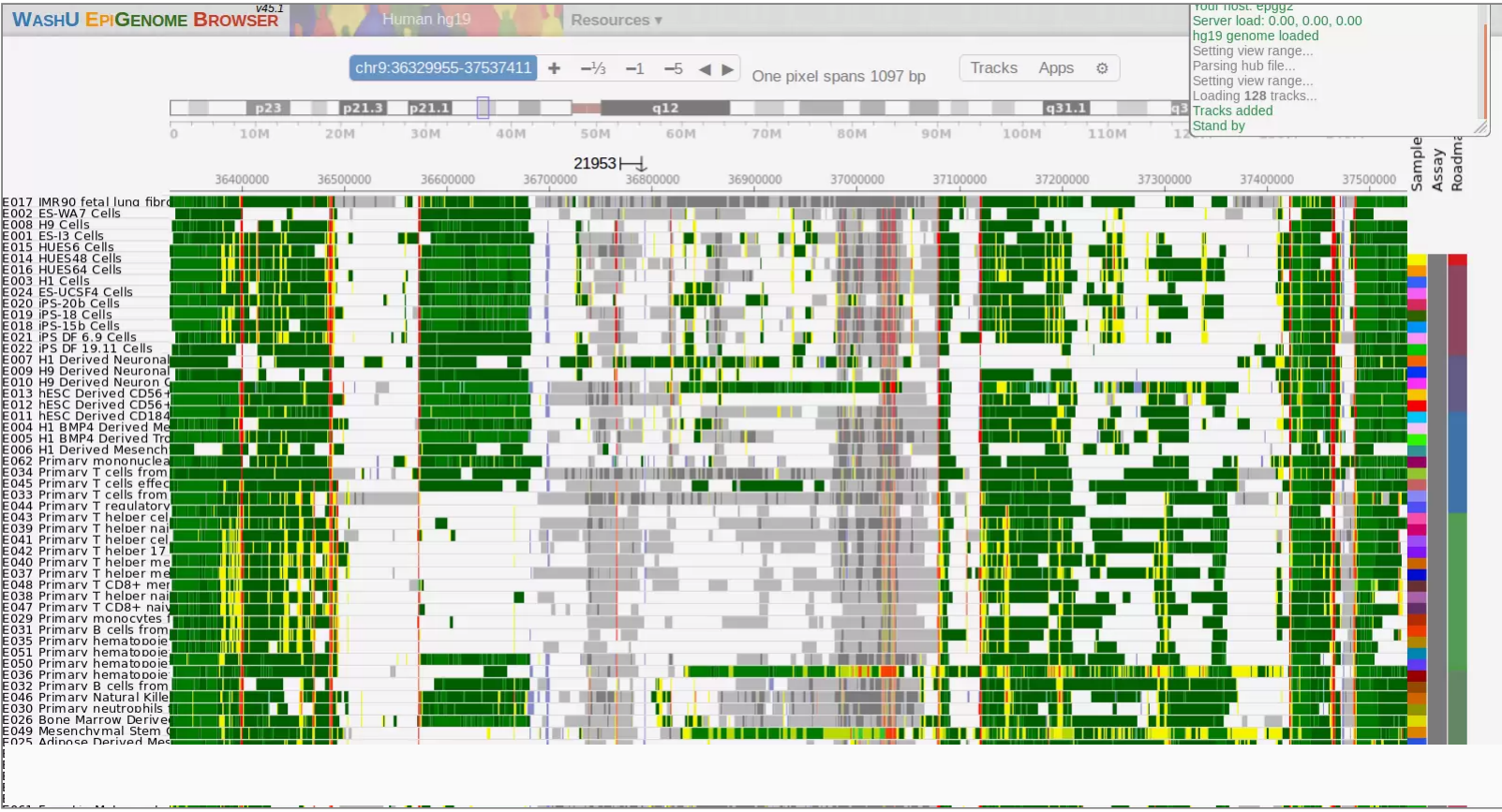
Status: released  1  3

Summary	Attribution
Assay: ChIP-seq	Lab: Peggy Farnham, USC
Target: SETDB1	Award: U54HG004558 (Michael Snyder, Stanford)
Biosample summary: <i>Homo sapiens</i> HEK293	Project: ENCODE
Biosample Type: immortalized cell line	Date released: June 1, 2017
Replication type: unreplicated	Submitter comment: Replaces ENCSR000EYD after it was discovered it was HEK293 and not U2OS
Description: SETDB1 ChIP-seq on human HEK293	
Nucleic acid type: DNA	
Lysis method: see document	
Extraction method: see document	
Fragmentation method: see document	
Size selection method: see document	



Roadmap *Chromatin states* browser

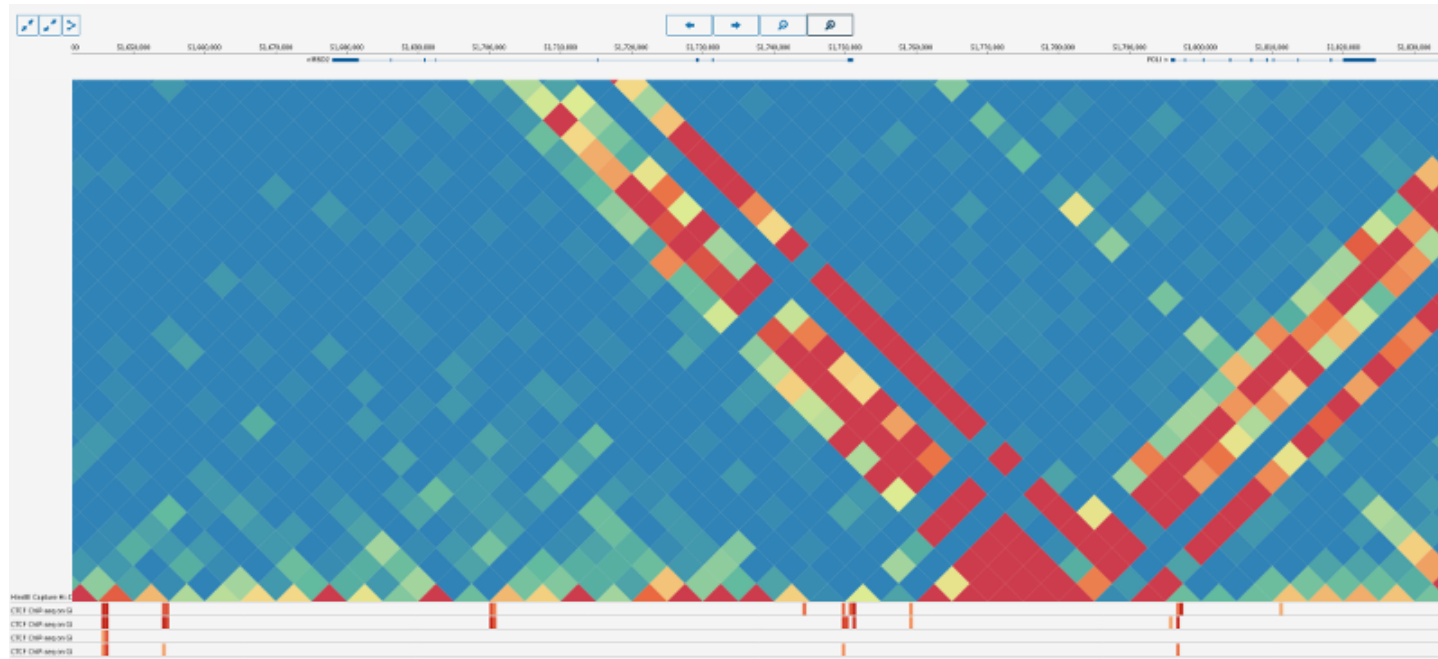
egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html



Uses the great *WahsU EpiGenome Browser*

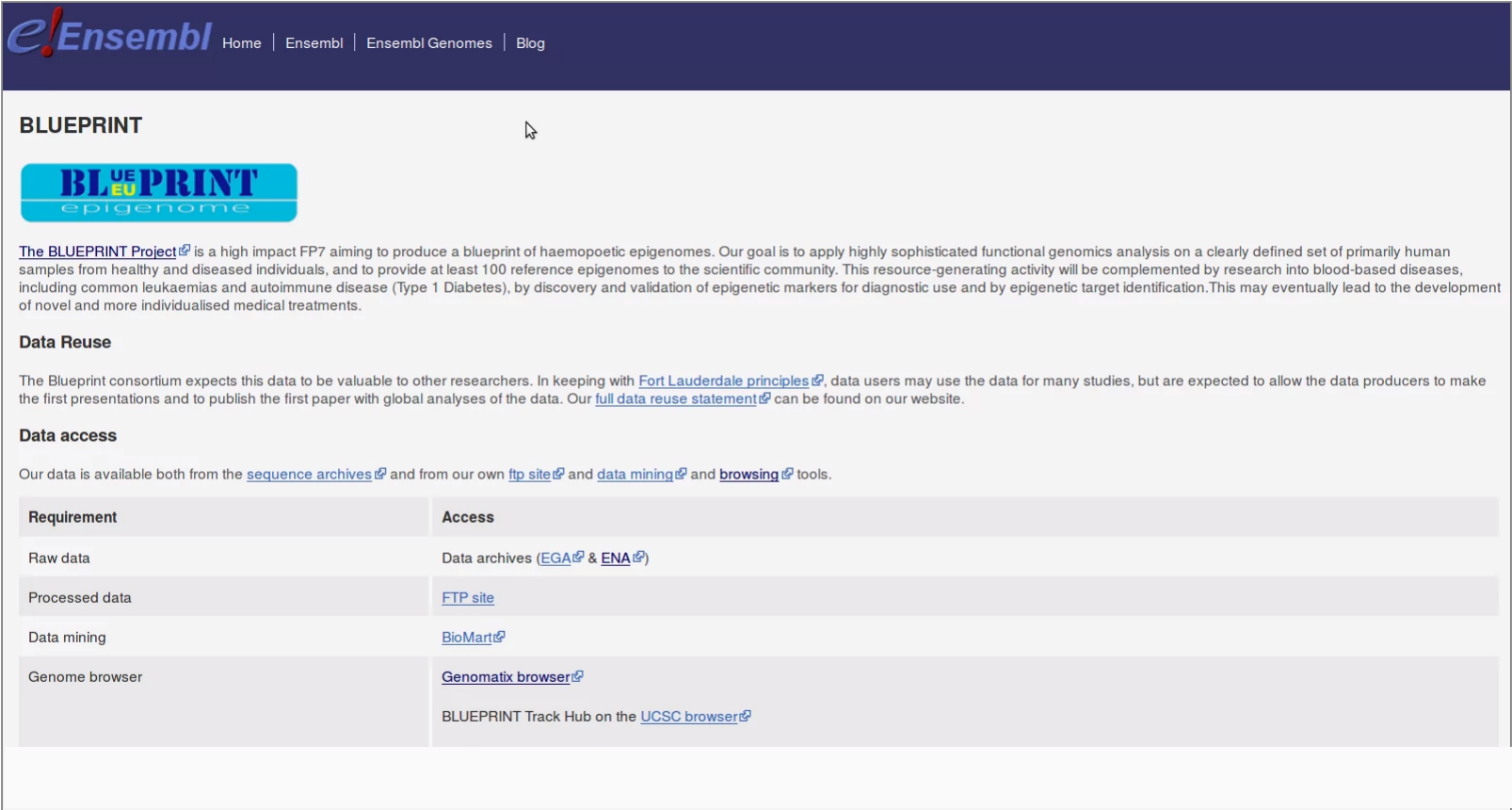
Other great options

- **Ensembl** regulatory build
- **Cistrome** send any tracks to the WashU EpiBrowser in a single click
- **QC Genomics** is developing **NAVi**




Most of the time...

Old & unfriendly or New & buggy ?



BLUEPRINT



The BLUEPRINT Project is a high impact FP7 aiming to produce a blueprint of haemopoietic epigenomes. Our goal is to apply highly sophisticated functional genomics analysis on a clearly defined set of primarily human samples from healthy and diseased individuals, and to provide at least 100 reference epigenomes to the scientific community. This resource-generating activity will be complemented by research into blood-based diseases, including common leukaemias and autoimmune disease (Type 1 Diabetes), by discovery and validation of epigenetic markers for diagnostic use and by epigenetic target identification. This may eventually lead to the development of novel and more individualised medical treatments.

Data Reuse

The Blueprint consortium expects this data to be valuable to other researchers. In keeping with [Fort Lauderdale principles](#), data users may use the data for many studies, but are expected to allow the data producers to make the first presentations and to publish the first paper with global analyses of the data. Our [full data reuse statement](#) can be found on our website.

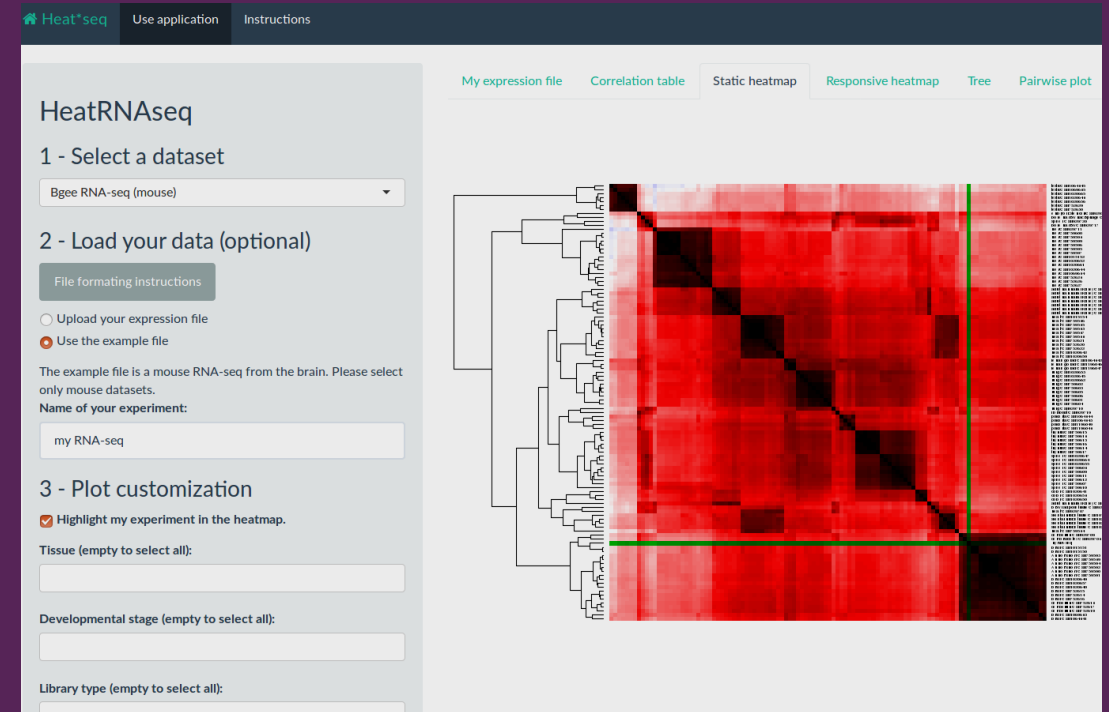
Data access

Our data is available both from the [sequence archives](#) and from our own [ftp site](#) and [data mining](#) and [browsing](#) tools.

Requirement	Access
Raw data	Data archives (EGA & ENA)
Processed data	FTP site
Data mining	BioMart
Genome browser	Genomatix browser BLUEPRINT Track Hub on the UCSC browser

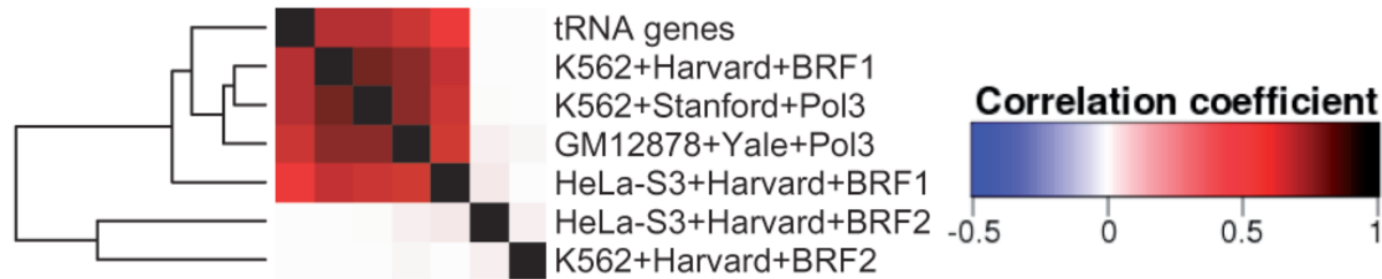
Types of web-apps for epigenomics

1. Data portals
2. Genome browsers
3. Compare experiments
4. Analyse



Clustered correlation heatmaps:

- Correlation of every pairs of experiment
- Clustering of the correlation matrix
- Two similar experiment will be close on the clustered correlation heatmap



Heat*seq: genome-wide comparison of NGS experiments

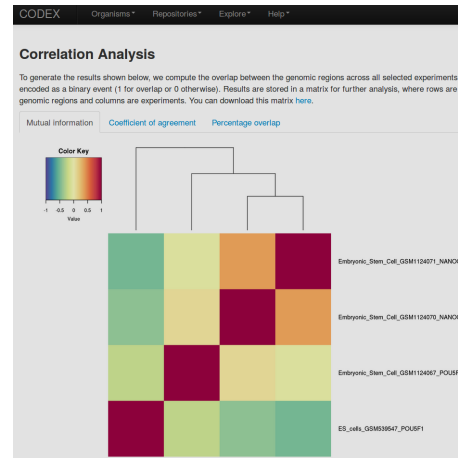
Example: ESR1 TF ChIP-seq from GEO, compared against all TF ChIP-seq from ENCODE

The screenshot displays the Heat*seq web application interface. The top navigation bar includes 'Heat*seq', 'Use application', and 'Instructions'. Below this, there are tabs for 'My peaks', 'Correlation table', 'Static heatmap', 'Responsive heatmap', 'Tree', 'Pairwise plot', and 'Samples metadata'. The main content area is divided into a left sidebar and a right main panel. The sidebar, titled 'HeatChIPseq', contains three sections: '1 - Select a dataset' with a dropdown menu showing 'ENCODE TFBS ChIP-seq (human, hg19)'; '2 - Load your data (optional)' with options to 'Upload your peak file' (selected) or 'Use the example file', a 'Choose a file:' section with a 'Browse...' button and 'No file selected' text, and a checkbox for 'My peak file contains a header.'; and '3 - Plot customization' with a checked checkbox for 'Highlight my experiment in the heatmap'. The main panel shows a heatmap with a dendrogram on the left and top. The heatmap is predominantly red, indicating high correlation, with a prominent diagonal line. The dendrogram on the left shows hierarchical clustering of the samples.

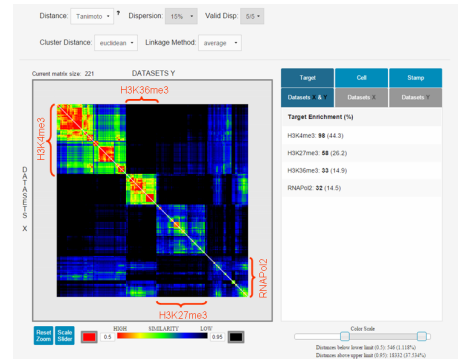
Input file: peak list, *.bed* format.

Alternatives:

CODEX Correlation analysis



QC Genomics: Comparator ngs-qc.org/qcgenomics/



Types of web-apps for epigenomics

1. Data portals
2. Genome browsers
3. Compare experiments
4. Analyse



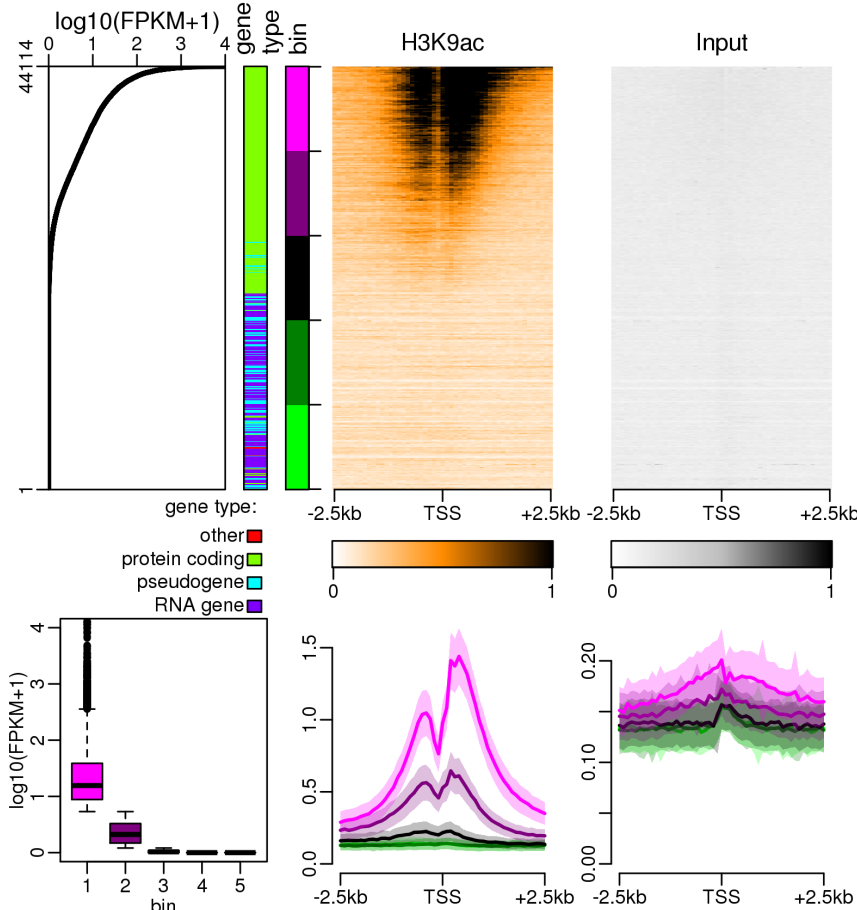
Profile Explorer for Roadmap Epigenomics

www.perepigenomics.roslin.ed.ac.uk

Using Roadmap Epigenomics data:

- Stack profiles
- Centered on a feature (here, TSS)
- Sorted according to expression (RNA-seq)

H3K9Ac, adult liver

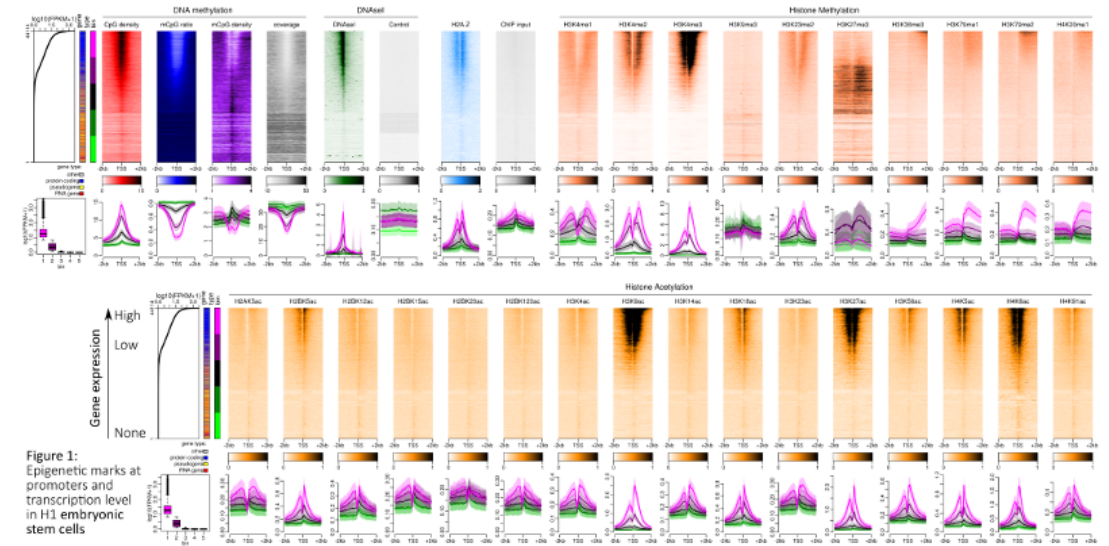


Profile Explorer for Roadmap Epigenomics

Pre-generated *stack profiles*:

- TSS, TES, middle exons
- By gene types (all, protein coding, lincRNA, pseudogenes, ...)
- WGBS, DNase1, H2AZ, 10 histone methylations, 16 Histone acetylations
- 30 cell types (cell lines + tissues)

→ 9 921 plots so far



Profile Explorer for Roadmap Epigenomics

PEREpigenomics Explore Compare Correlate About Available profiles

Select a plot:

- 1- Order by
 - Epigenetic assay first
 - Cell type first
- 2- Focus on:
 - TSS
- 3- Choose an assay:
 - DNase 1
- 4- Choose a cell type:
 - H1_Cell_Line
- 5- Choose a gene category:
 - all

[Download image](#)

DNase 1 at TSS in H1_Cell_Line for all genes

log10(FPKM+1) 44114 0 1 2 3 4

gene type bin

DNase Control

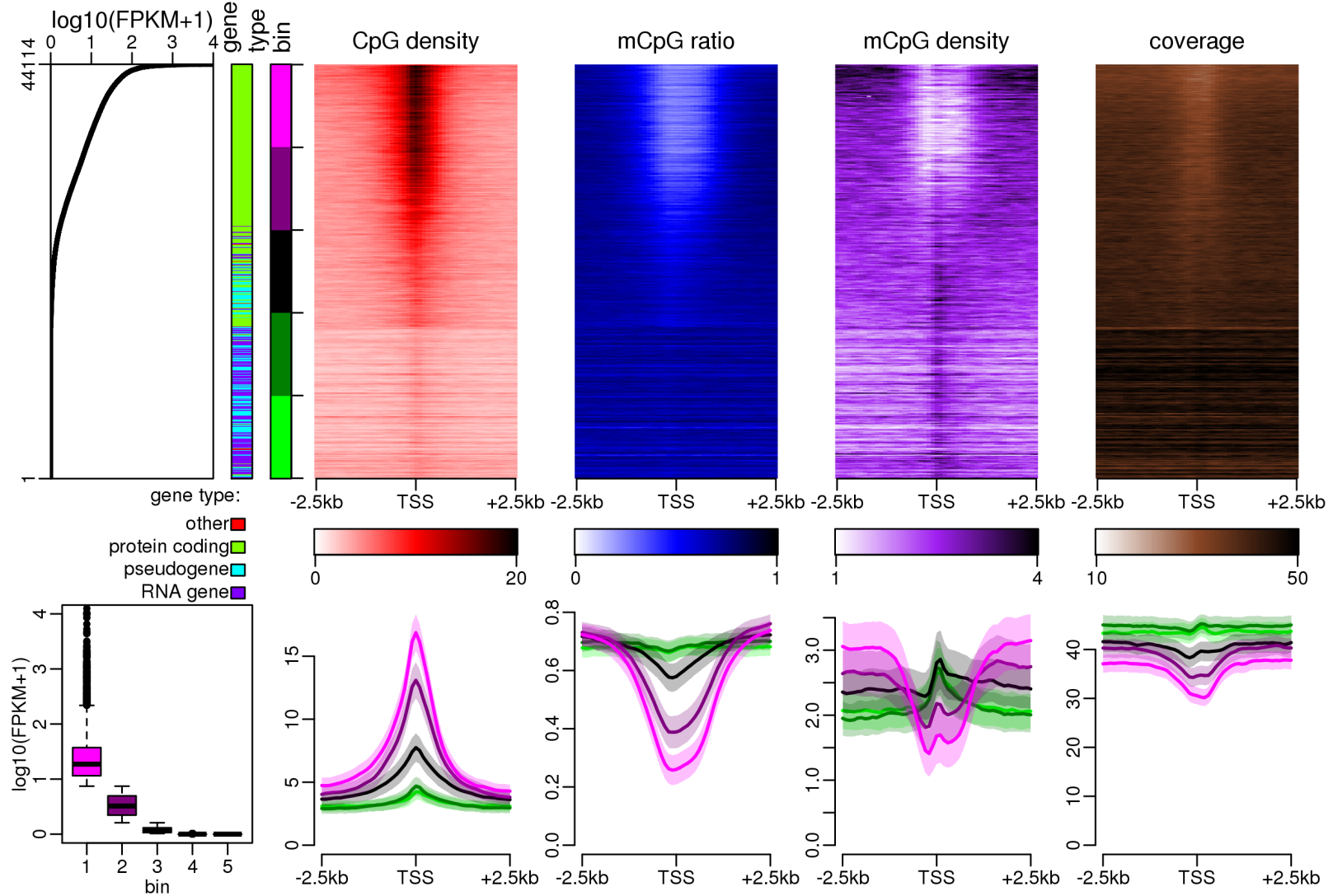
-2.5kb TSS +2.5kb -2.5kb TSS +2.5kb

gene type:
other
protein coding
pseudogene
RNA gene

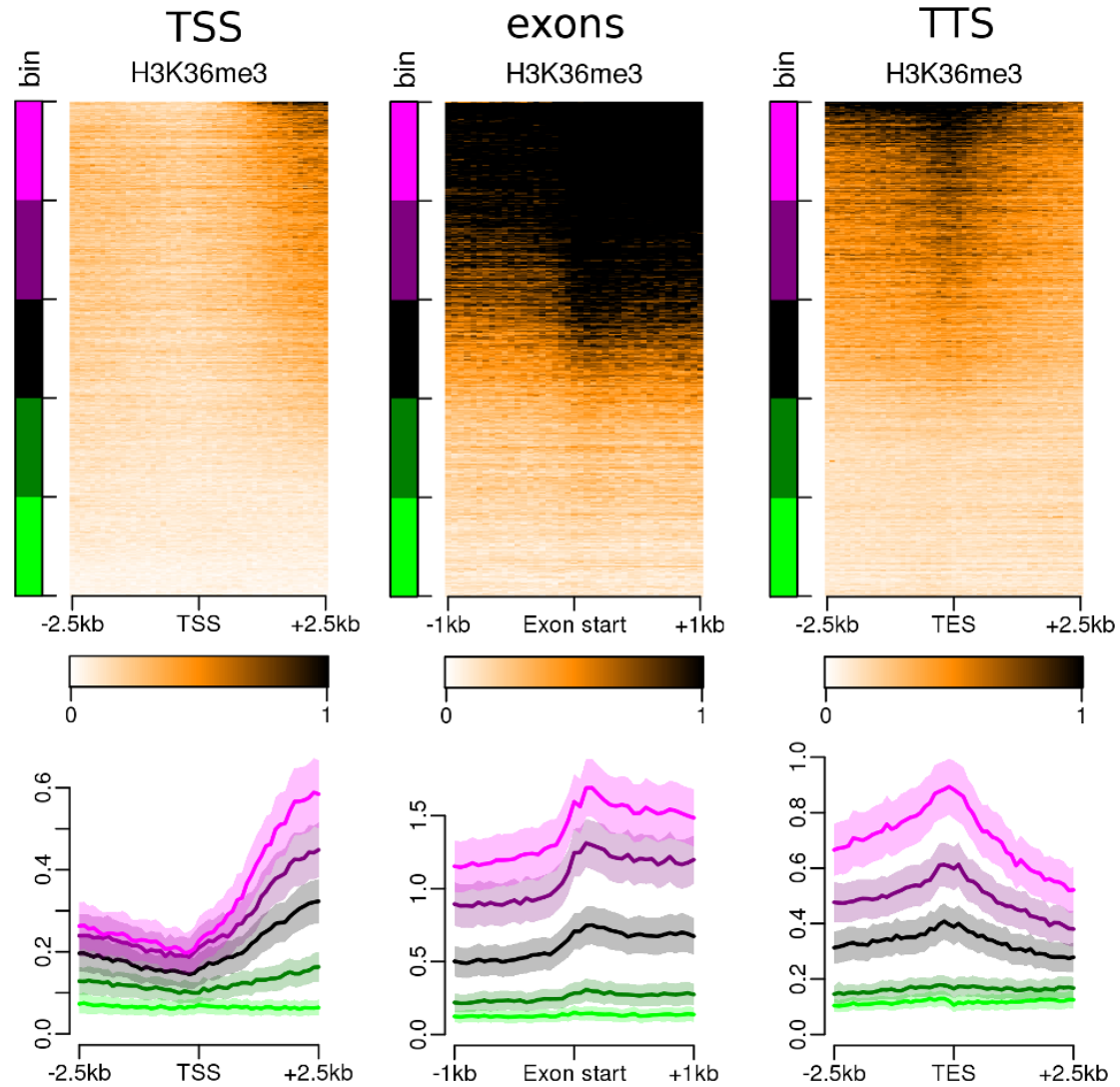
log10(FPKM+1)

0 0.05 0.10 0.15 0.20

WGBS, pancreas



H3K36me3, fetal large intestine



Alternatives

Tools from *Blueprint Epigenome*:

- *Blueprint* data analysis portal: blueprint-data.bsc.es/release_2016-08
- *DeepBlue* analysis server: deepblue.mpi-inf.mpg.de
- Dive (in early development): dive.mpi-inf.mpg.de

QC Genomics: ChromStater ngs-qc.org/qcgenomics/

Galaxy servers, notably:

- *deepTools*: deeptools.ie-freiburg.mpg.de
- *Cistrome*: cistrome.org/ap/root

Thanks

Anagha Joshi

Anna Mantsoki

Deepti Vipin

