Fostering data re-use with interactive visualisations of transcriptomics and epigenomics data

Guillaume Devailly

Fostering data re-use with interactive visualisations of transcriptomics and epigenomics data

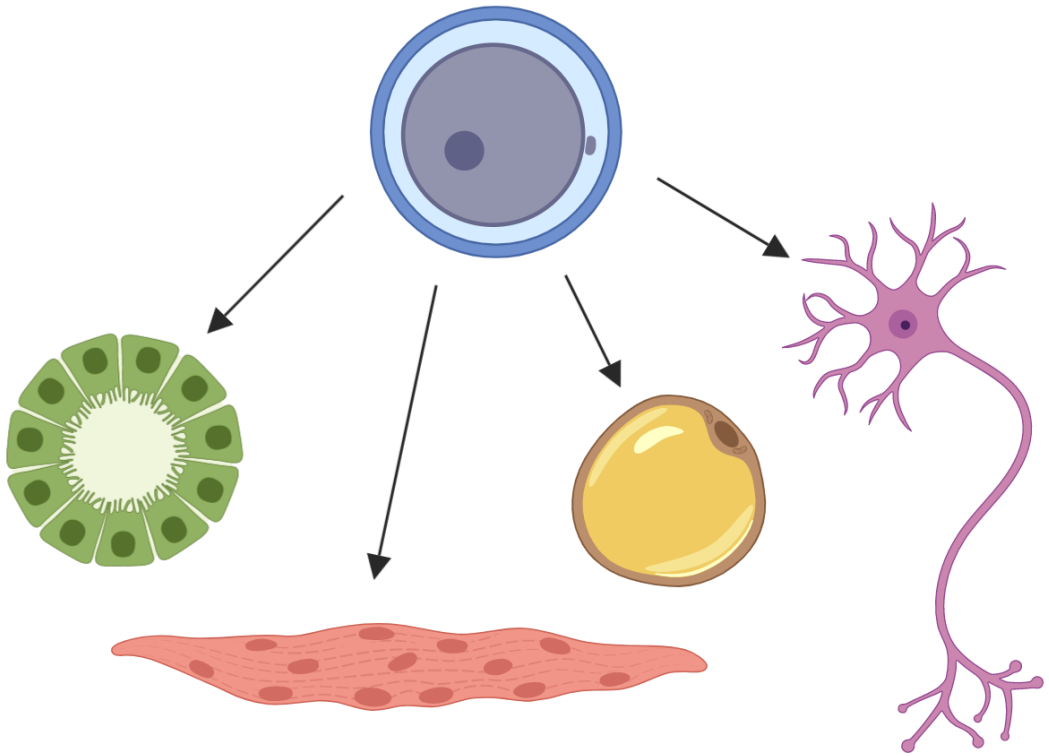
@G_Devailly

Guillaume Devailly

Séminaire MIAT

2020/02/07

One genome, many cell types



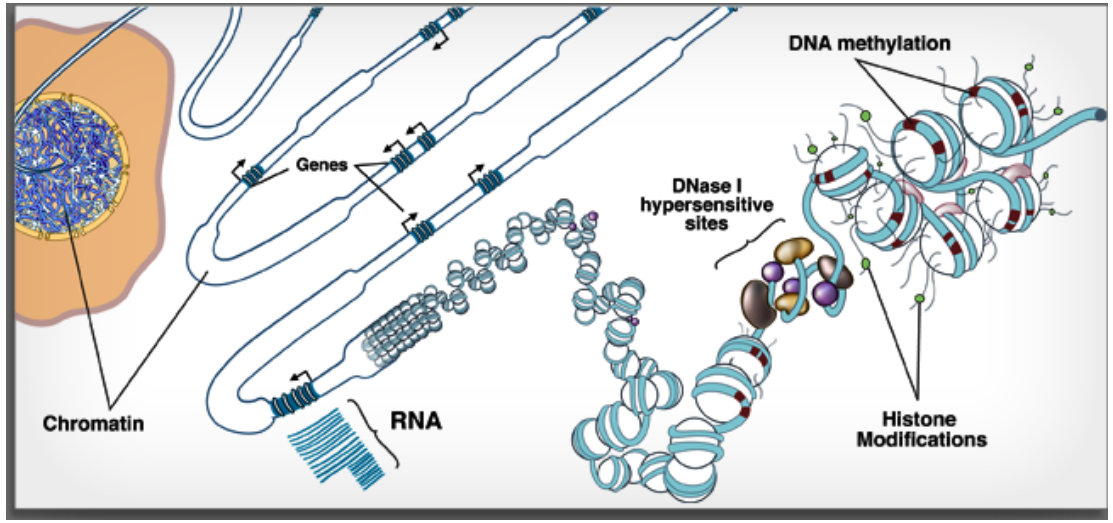
Differential gene expression =
cellular environment + epigenetics

INRAE GenPhySE

Transcriptomics & Epigenomics data

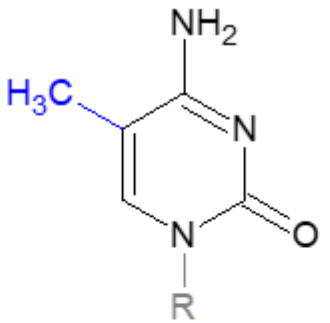
Each experiment is complex and costly, but many datasets and databases are available.

- Expression: RNA-seq
- Chromatin accessibility: DNAse1, ATAC-seq
- DNA methylation: WGBS
- ChIP-seq:
  - Transcription regulators
  - Histone variants
  - Histone modifications



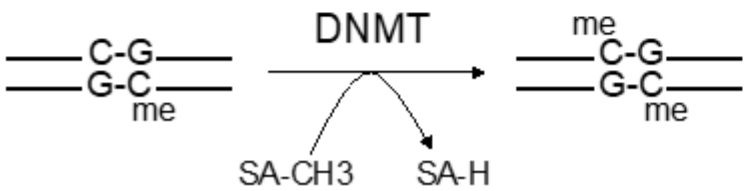
DNA methylation

Vertebrate DNA methyation

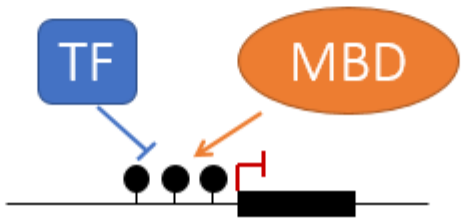


5-méthylcytosine
(R = H)

1- Write



2- Read

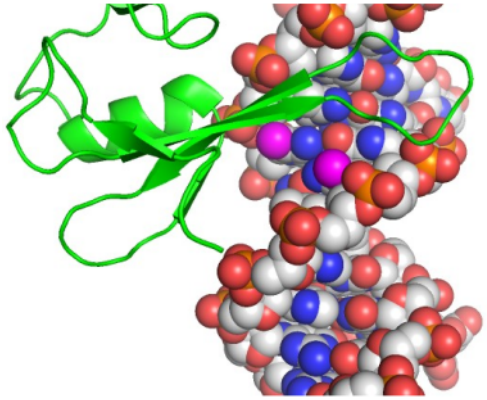


3- Erase

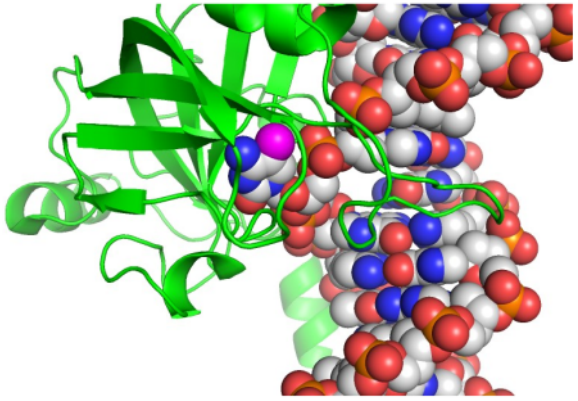
- 5hmC TET + BER
- BER / NER / MMR
- passive demethylation

DNA methylation

MBD2



UHRF1



Data re-use

The good

- Mostly FAIR data, good enough metadata

The bad

- Domain expertise, experimental artefacts
- Arguable arbitrary choices during analyses

Data re-use

The good

- Mostly FAIR data, good enough metadata

The bad

- Domain expertise, experimental artefacts,
- Discussable arbitrary choices during analyses

The ugly

- Fat data, too lazy to download it all

Raw sequencing data

Isogenic replicate	Library	Accession	File type	Run type	Read	Lab	Date added	File size	Audit status	File status
1	ENCLB597RSH	ENCFF164VNM	fastq	PE100nt	1	Michael Snyder, Stanford	2017-06-20	2.59 GB		released
		ENCFF825EGN	fastq	PE100nt	2	Michael Snyder, Stanford	2017-06-20	3.02 GB		released
2	ENCLB081NDC	ENCFF862VDD	fastq	PE100nt	1	Michael Snyder, Stanford	2017-06-20	3.06 GB		released
		ENCFF814SDG	fastq	PE100nt	2	Michael Snyder, Stanford	2017-06-20	3.53 GB		released

Processed data

Visualize	Accession	File type	Output type	Isogenic replicate	Mapped read length	Mapping assembly	Lab	Date added	File size
	ENCFF028TNY	bam	alignments	1	100	GRCh38	ENCODE Processing Pipeline	2017-06-23	4.18 GB
	ENCFF555KDG	bam	unfiltered alignments	1	100	GRCh38	ENCODE Processing Pipeline	2017-06-23	5.25 GB

1 experiment: 12.2 GB

INRAE GenPhySE

Genome browsers can help!

Few experiments at a time

Few genes at a time

INRAE GenPhySE

Interactive visualisations to foster data re-use

1. **Heat*seq: interactive correlation heatmaps of transcriptomics and epigenomics datasets**
2. PEREpigenomics: Profile Explorer of Roadmap Epigenomics data
3. VizFaDa: Visualisations of FAANG data

INRAE GenPhySE

Correlation heatmaps for transcriptomics and TF ChIP-seq datasets

Heat*seq workflow

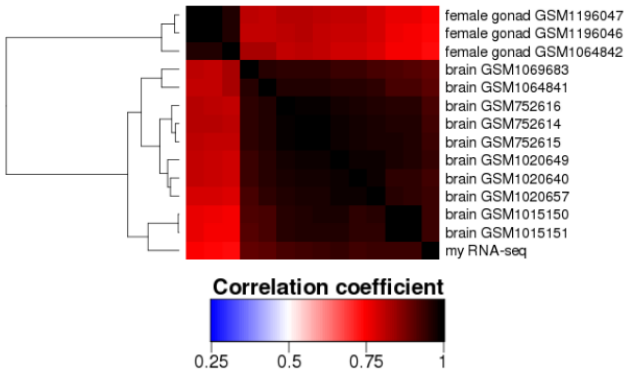
1- Visit website:
www.heatstarseq.roslin.ed.ac.uk

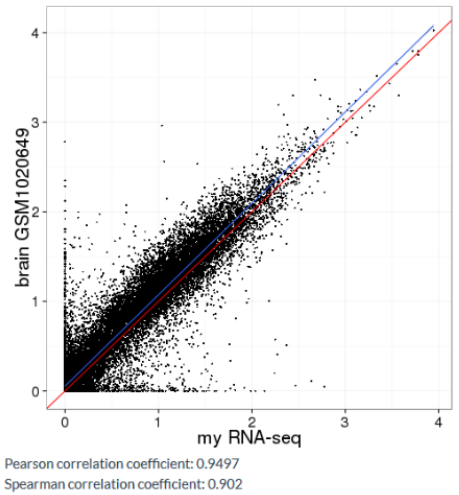
2- Select dataset

3-Upload processed data

4- Explore results

experiment	correlation
brain GSM1020649	0.9496700
Ammon's horn GSM759591	0.9445169
brain GSM1015150	0.9430056
brain GSM1015151	0.9426334
brain GSM1020657	0.9415478
Ammon's horn GSM759593	0.9414239
Ammon's horn GSM759589	0.9400874
Ammon's horn GSM759592	0.9383826
brain GSM752615	0.9375823





INRAE GenPhySE

HeatRNAseq workflow

Expression matrix

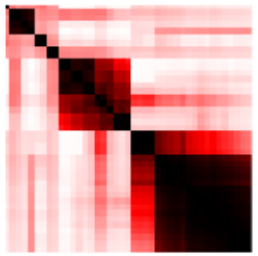
Gene name (~40.000)	Exp 1	Exp 2	Exp 3	...	Exp 77
ENSG00000000003	3.983	2.361	10.216	...	80.583
ENSG00000000005	0.071	0.260	0.000	...	3.329
ENSG00000000419	10.277	2.893	14.153	...	42.639
...	...	...	...	...	....
ENSG00000273493	0.000	0.000	0.000	...	0.000

Correlation matrix

Pearson's, after log10 scaling

\	Exp 1	Exp 2	Exp 3	...	Exp 77
Exp 1	1	0.942	0.938	...	0.663
Exp 2	0.942	1	0.917	...	0.680
Exp 3	0.938	0.917	1	...	0.706
...	...	...	...	1	...
Exp 77	0.663	0.680	0.706	...	1

Clustered heatmap



INRAE GenPhySE

HeatChIPseq workflow

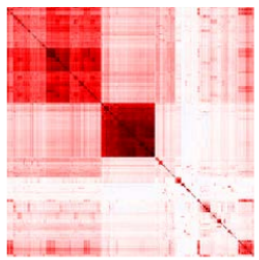
Binary peak matrix

Coordinates (~700.000 non overlapping regions)	Exp 1	Exp 2	Exp 3	...	Exp 690
chr1:10073-10413	F	T	F	...	F
chr1:16110-16390	F	F	T	...	T
chr1:29198-29688	F	F	F	...	F
...	...	...	...	...	....
chrY:28709160-28709494	T	T	F	...	F

Correlation matrix

\	Exp 1	Exp 2	Exp 3	...	Exp 690
Exp 1	1	0.059	0.786	...	0.035
Exp 2	0.059	1	0.058	...	0.118
Exp 3	0.786	0.058	1	...	0.047
...	...	...	...	1	...
Exp 690	0.035	0.118	0.047	...	1

Clustered heatmap



INRAE GenPhySE

Live Demo !

www.heatstarseq.roslin.ed.ac.uk

Correlations between TF ChIP-seq peaks and TSS list

Heat*seq conclusions & perspectives

- App: www.heatstarseq.roslin.ed.ac.uk
- Source code: github.com/gdevailly/HeatStarSeq_gh
- Publication: doi.org/10.1093/bioinformatics/btw407

Perspectives ?

- More datasets! (also, update the old ones...)
- More datatypes: Hitsone marks, gene lists, ...
- Multiple use files
- Gene name converter, liftover
- Datasets with more that 1000 experiments?

Heat*seq thanks:

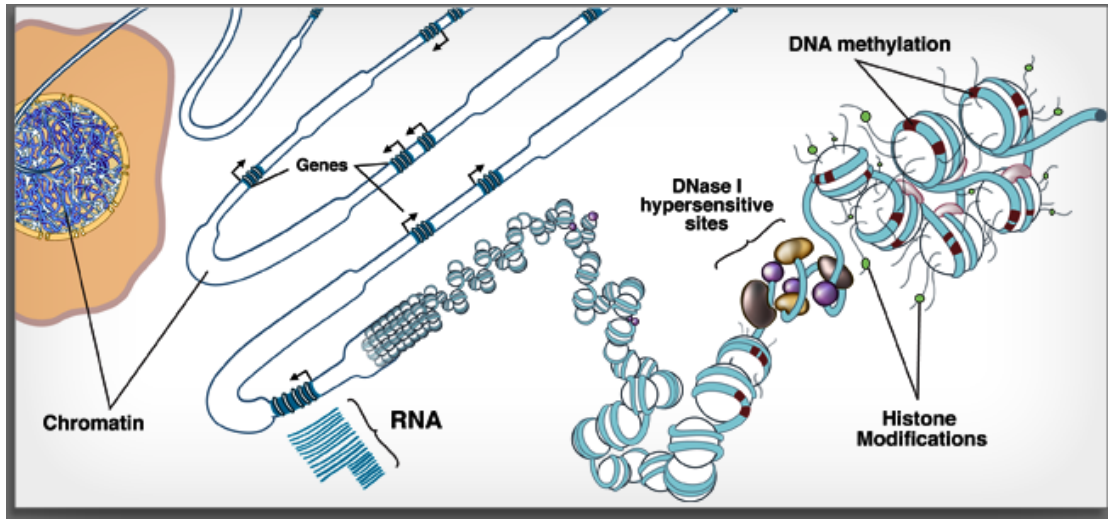
Interactive visualisations to foster data re-use

1. Heat*seq: interactive correlation heatmaps of transcriptomics and epigenomics datasets
2. **PEREpigenomics: Profile Explorer of Roadmap Epigenomics data**
3. VizFaDa: Visualisations of FAANG data

INRAE GenPhySE

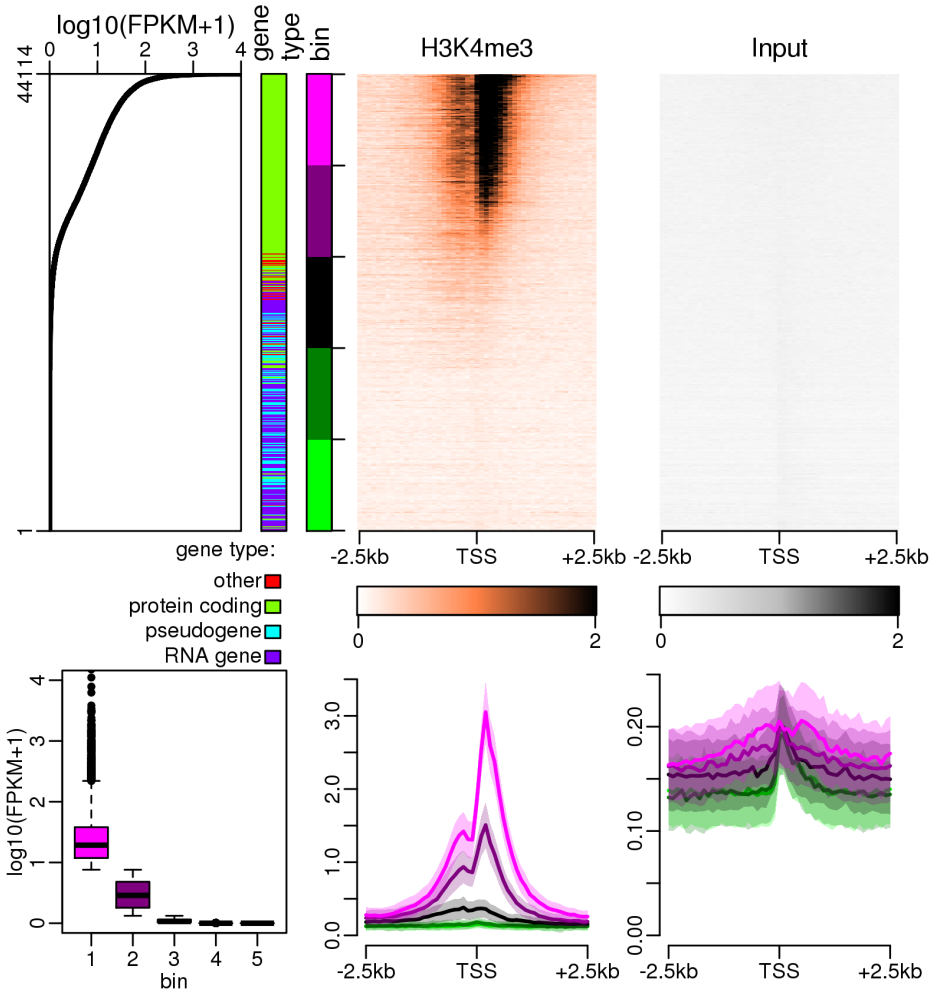
The Roadmap Epigenomics dataset

- RNA-seq
- DNAse1
- WGBS,
- 10 different histone methylations
- 17 different histone acetylations
- 33 human cell lines & tissues
- uniformly processed

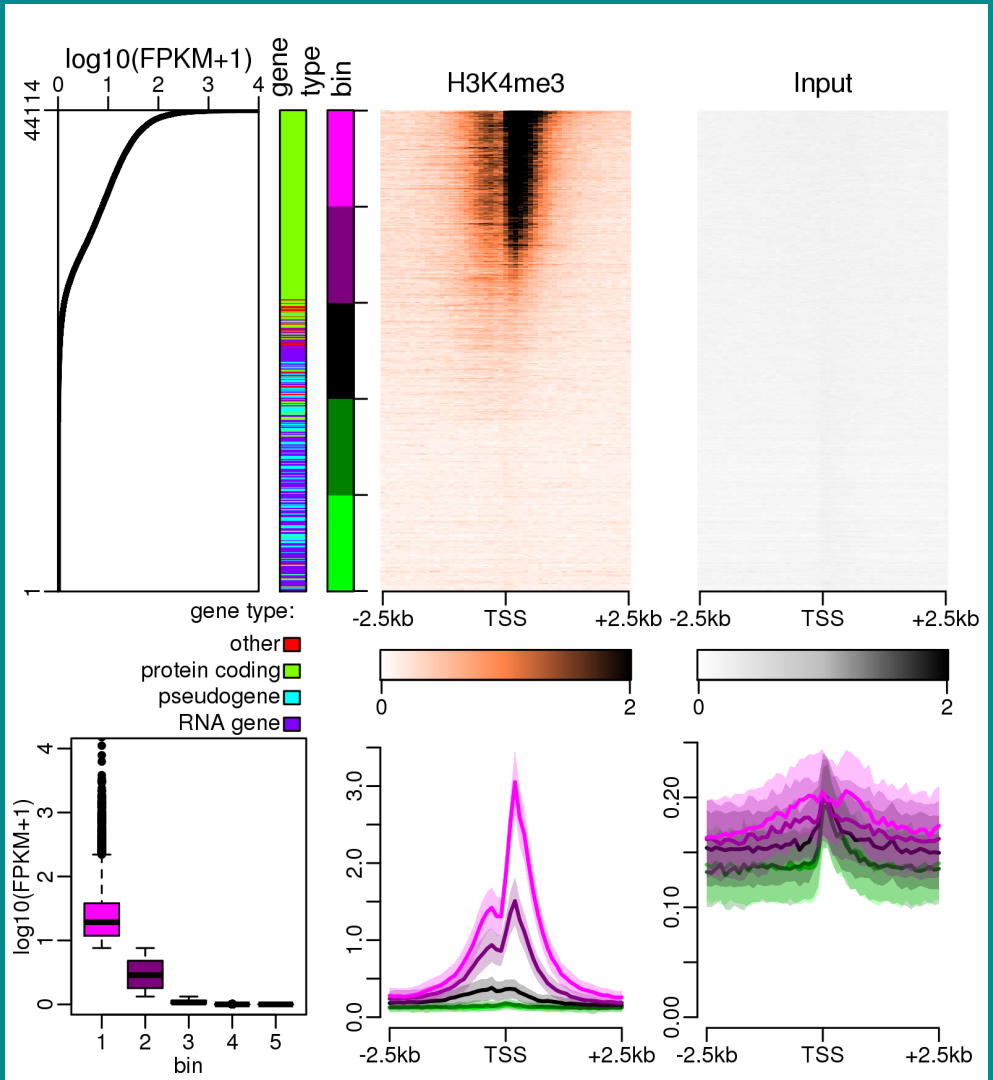


Objectives

- Visualisation of marks, sorted by **gene** transcription level:
  - at TSS (gene start)
  - at TTS (gene end)
- Visualisation of marks at **middle exons** starts, sorted by:
  - transcription level
  - inclusion ratio
- For **all** genes/exons in **each** cell type.
- For **each** gene/exon in **all** cell types.



H3K4me3, TSS, small intestine



What is a gene?

Version 22 (October 2014 freeze, GRCh38) - Ensembl 79, 80

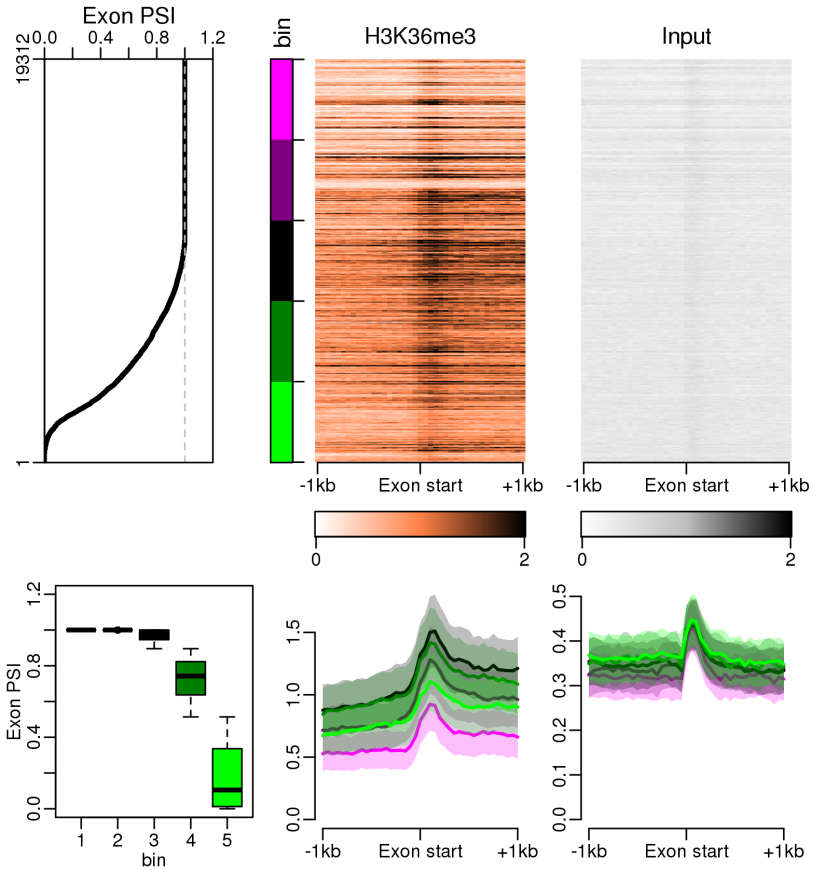
General stats

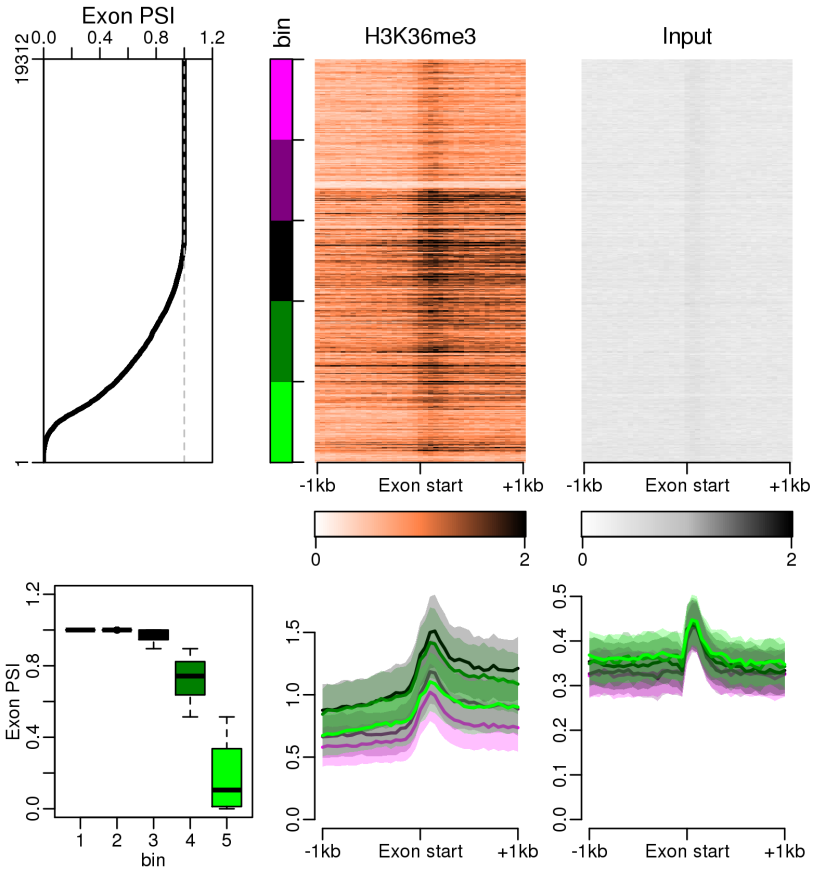
Total No of Genes	60483

Where is BRCA1's Transcription Start Site (TSS)?

More genes than pixels on the device!

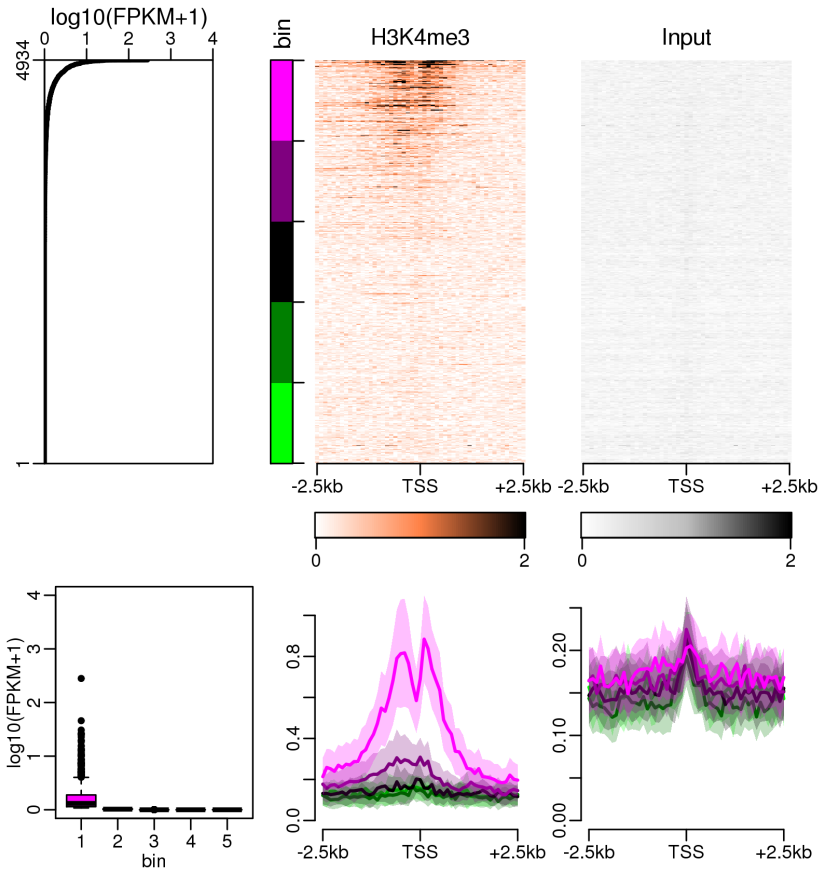
Ties shuffling



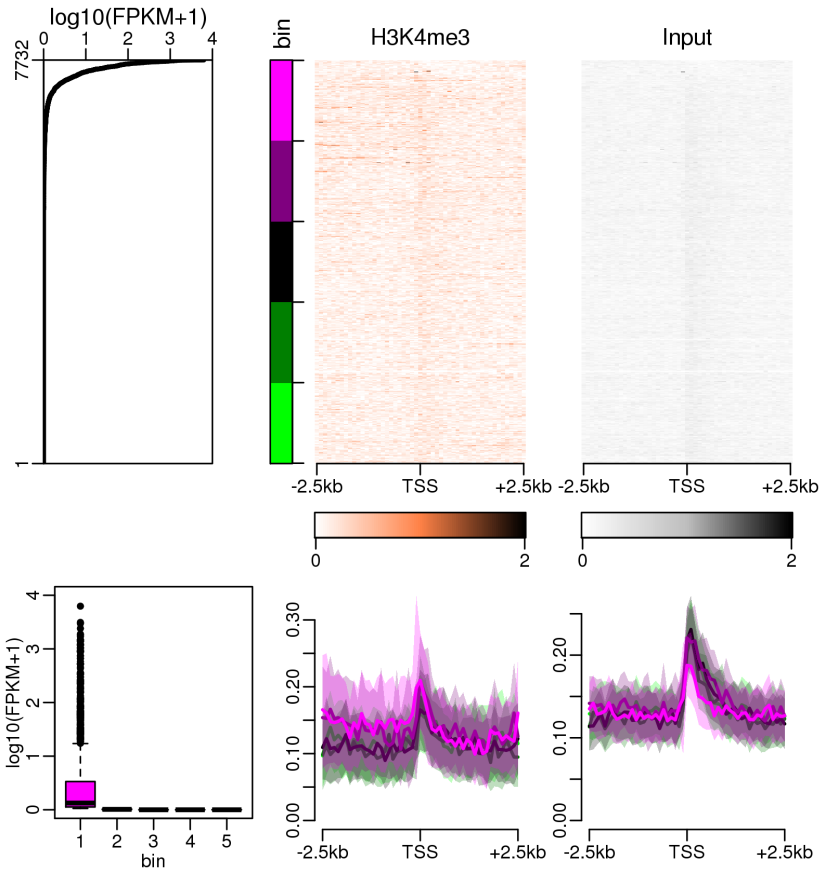


Different gene types, different associations

lincRNA

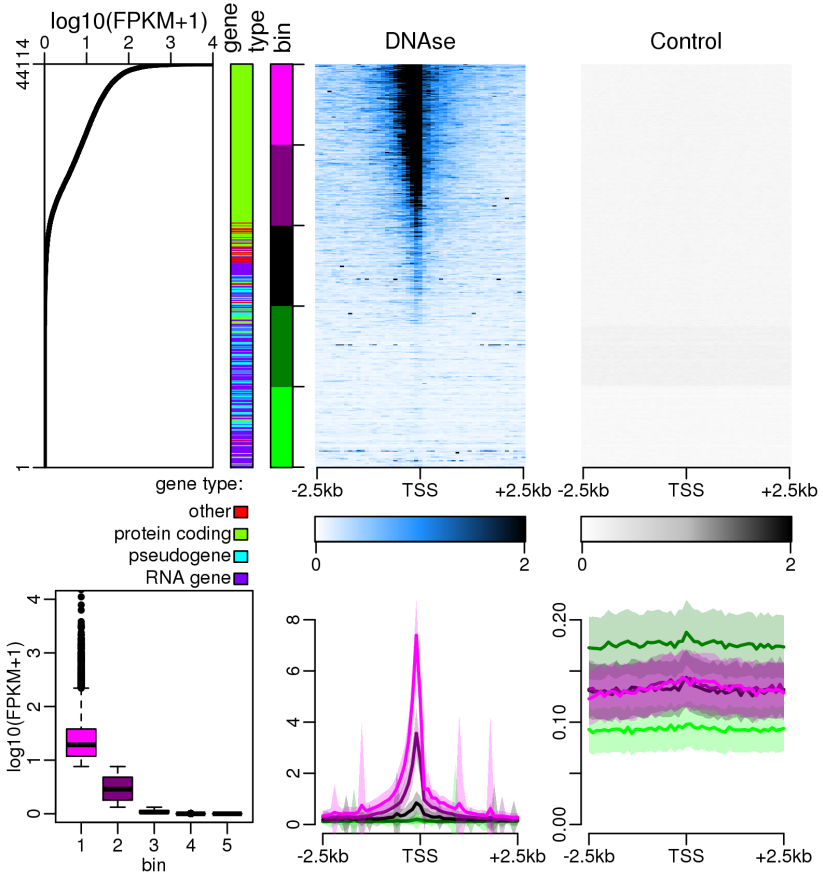


processed pseudogenes

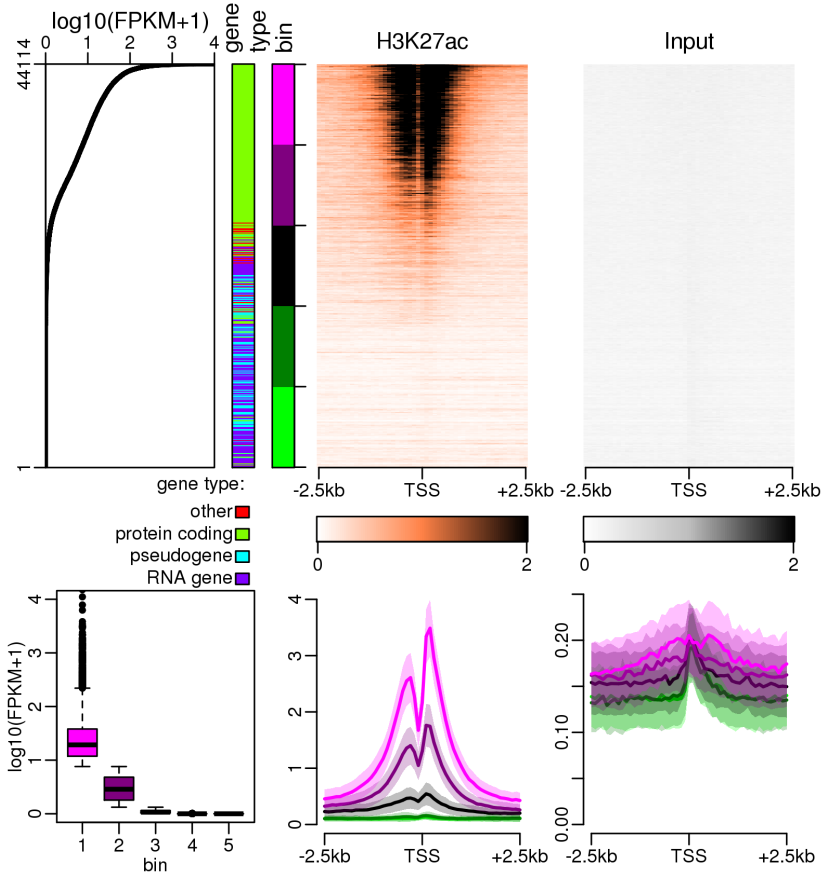


Position of signal

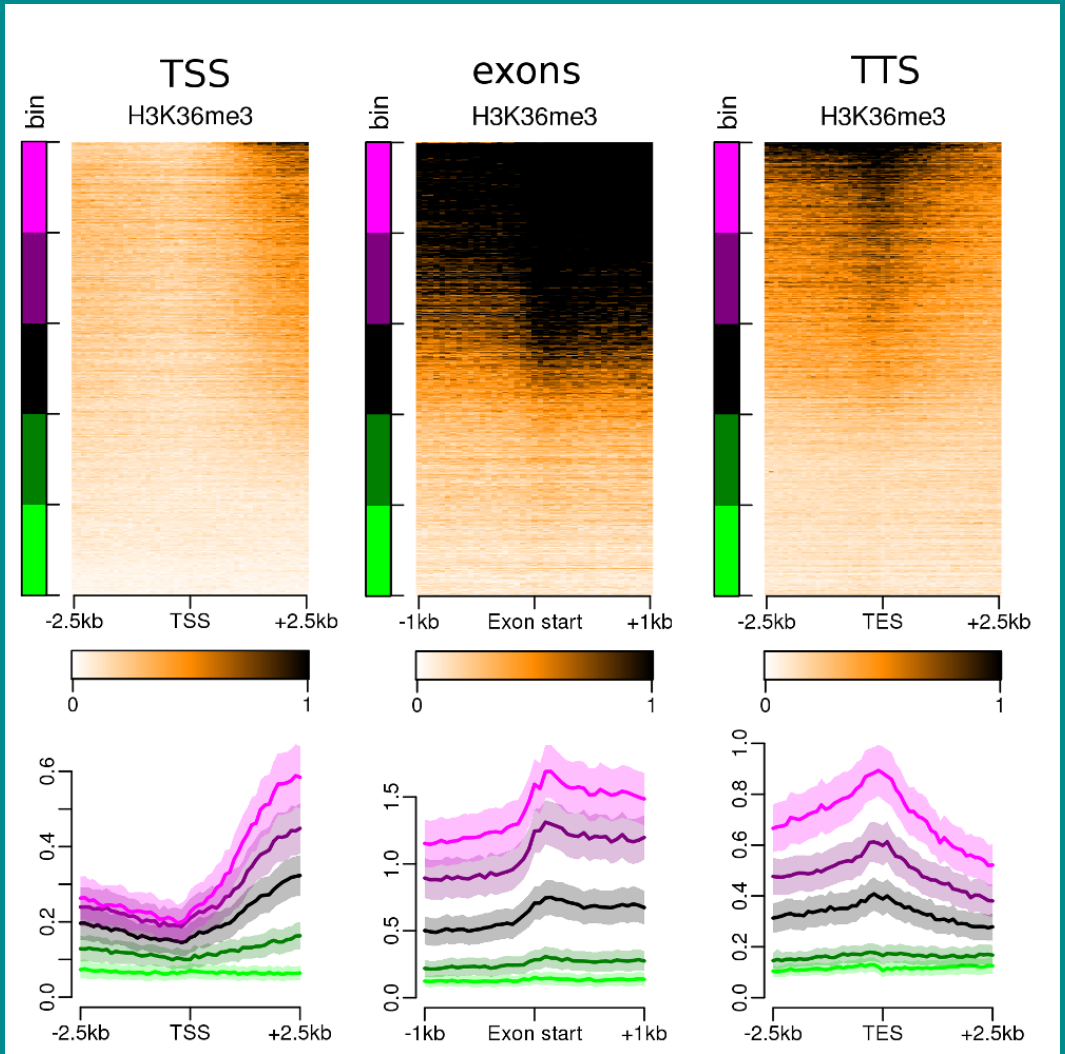
DNAse 1 (accessible chromatin)



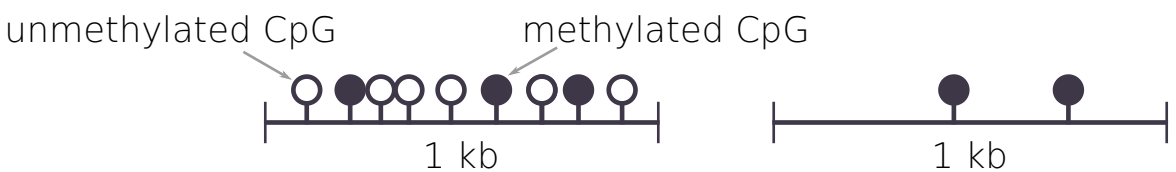
H3K27ac



Exonic mark: H3K36me3, foetal large intestine

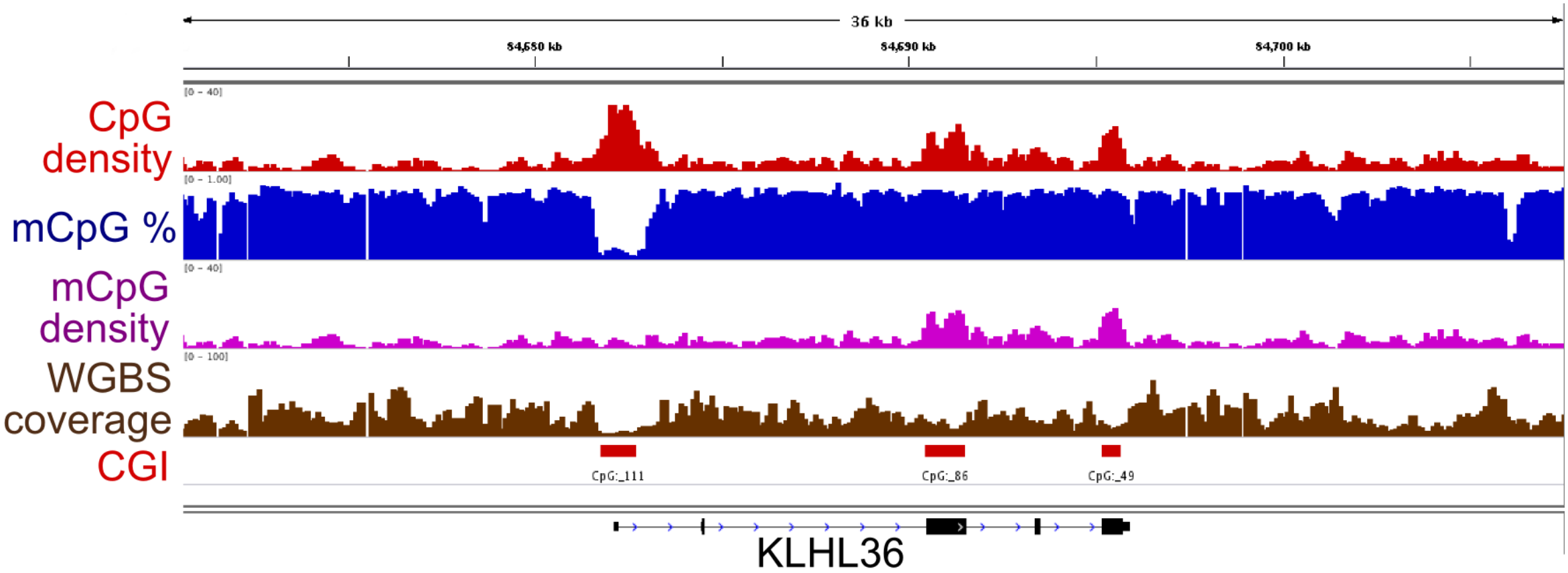


DNA methylation: ratio *and* density

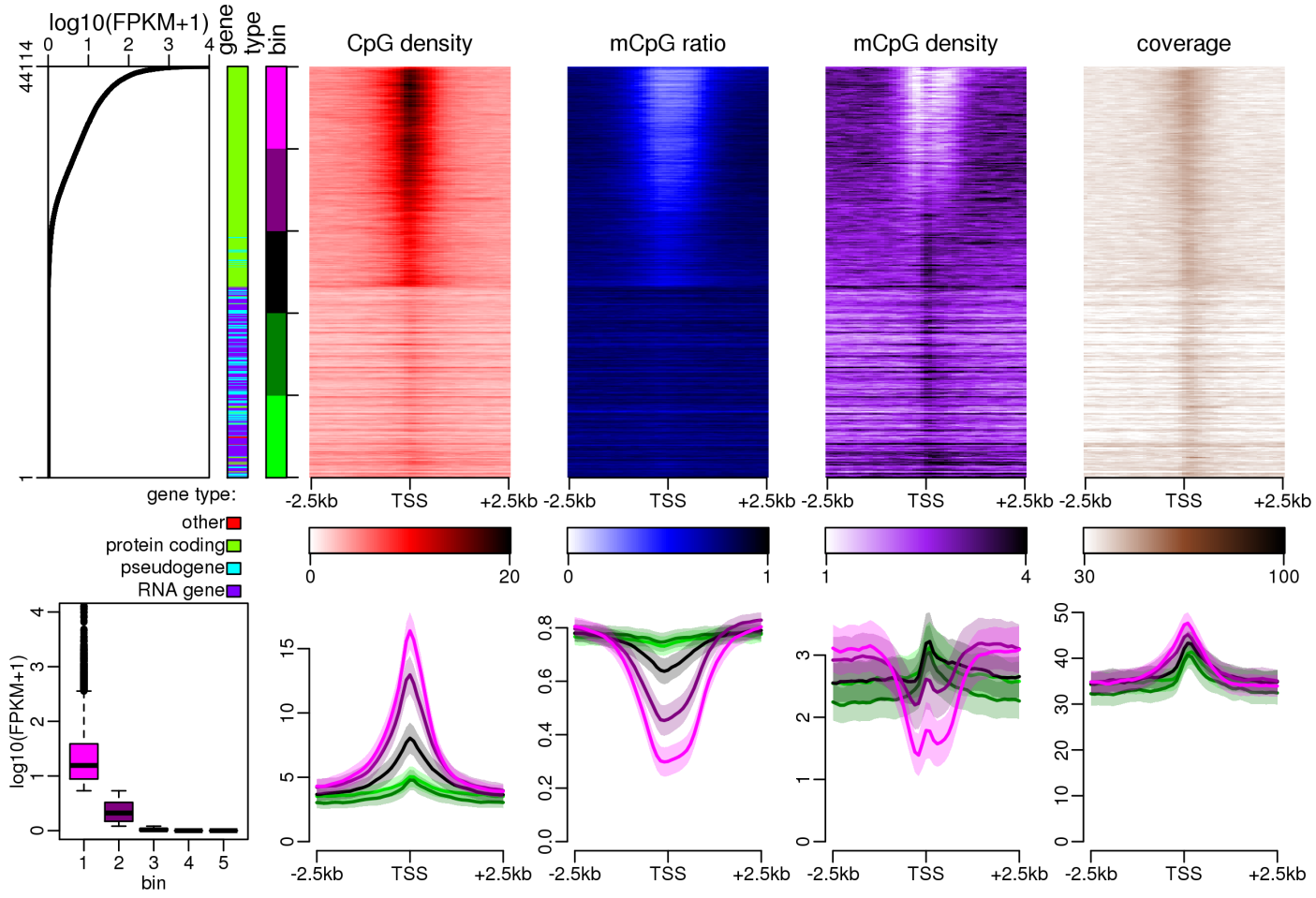


	CpG island	CpG poor region
CpG density	9 CpG/kb	2 CpG/kb
mCpG ratio	33%	100%
mCpG density	3 mCpG/kb	2 mCpG/kb

DNA methylation: ratio *and* density



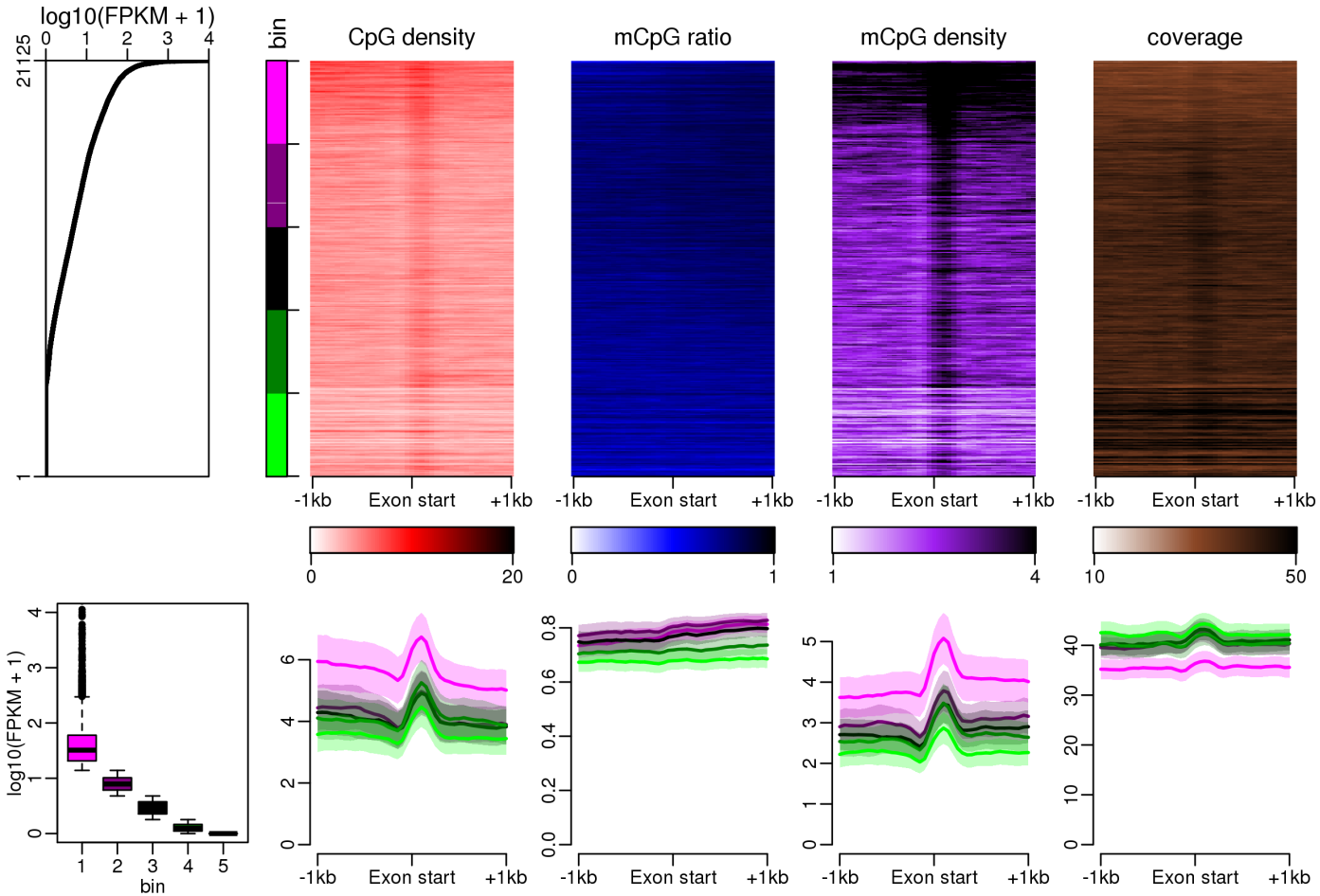
DNA methylation at TSS



WGBS, adult liver

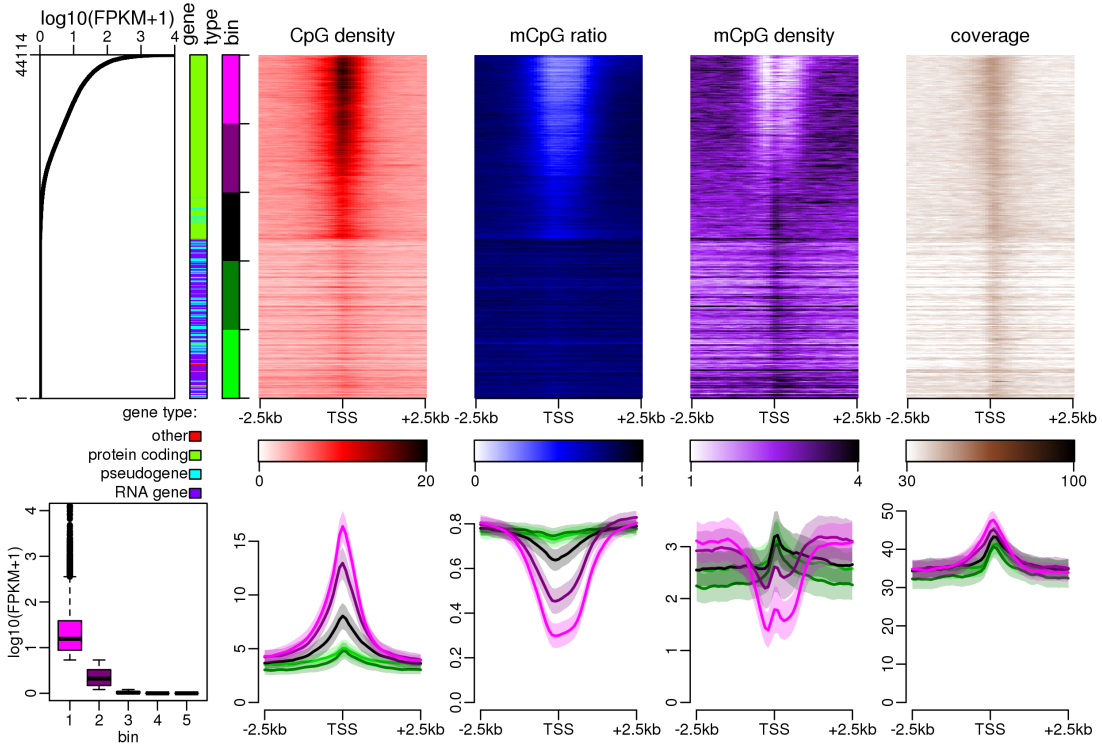
Exonic DNA methylation

WGBS, pancreas

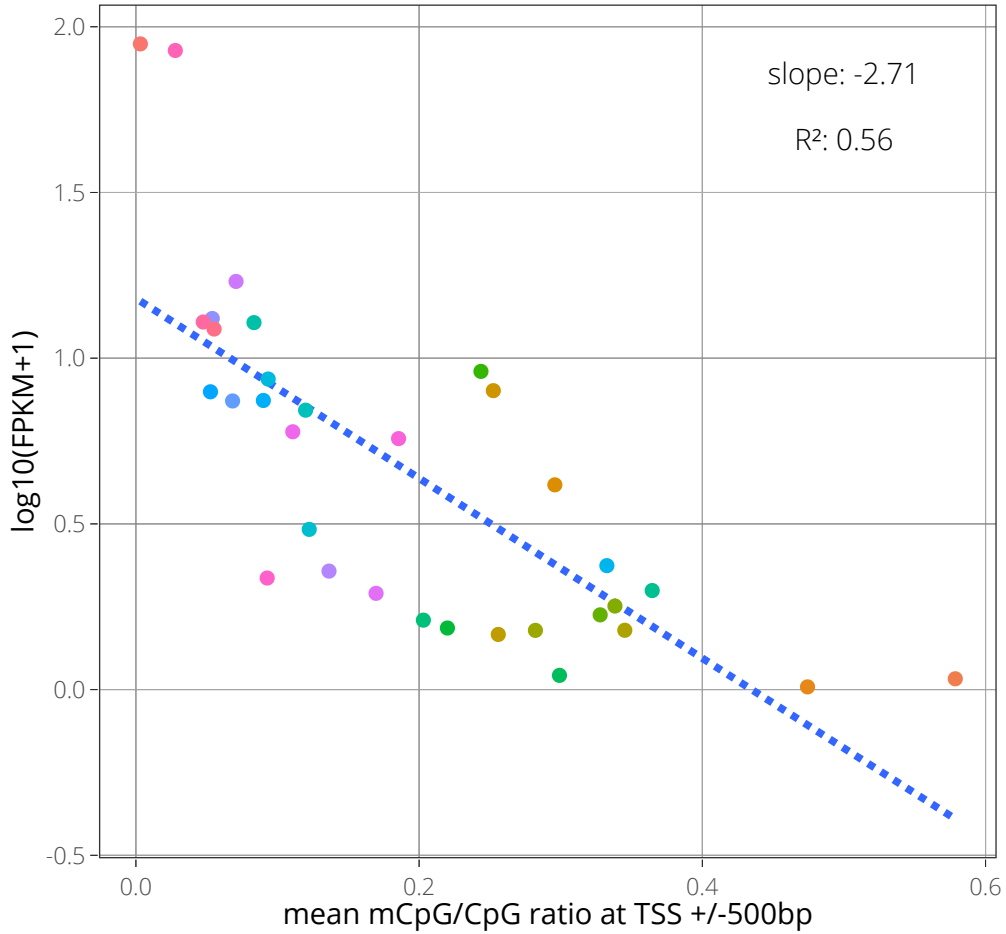


Cell by cell *vs* gene by gene

WGBS, adult liver



LYL1 (ENSG00000104903)



Repressing and activating marks

mCpG ratio

H3K4me3

Too many plots?

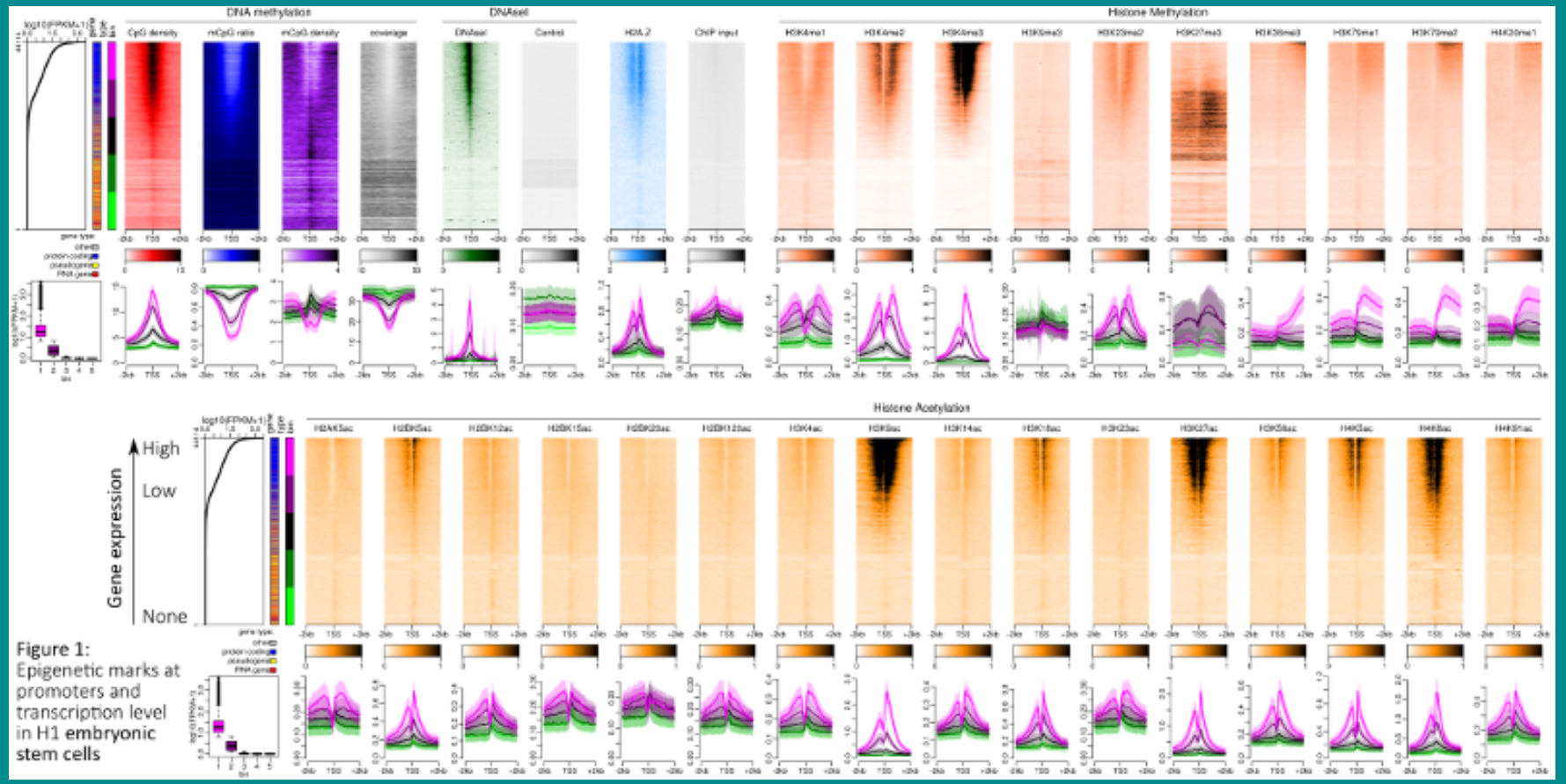
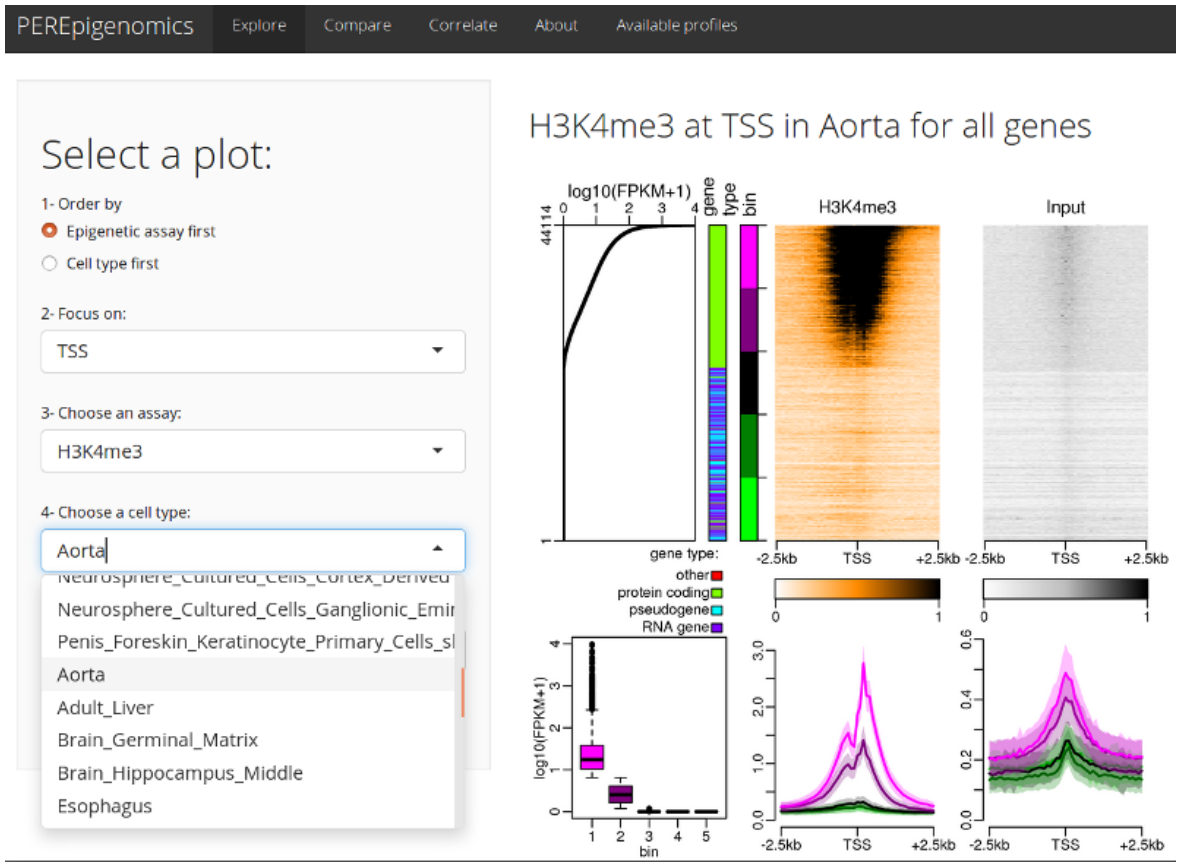


Figure 1:
Epigenetic marks at promoters and transcription level in H1 embryonic stem cells

PEREpigenomics

*P*rofile *E*xplorer of *R*oadmap *Epigenomic* data

joshiapps.cbu.uib.no/perepigenomics_app/



Summary of results

mark	cell type by cell type	gene by gene	center on TSS
WGBS	negative	negative	
DNAse	positive	positive	
H2A.Z	positive	balanced	
H3K4me1	positive	positive	
H3K4me2	positive	positive	
H3K4me3	positive	positive	
H3K9me3	unclear	unclear	no
H3K23me2	positive	NA	
H3K27me3	negative – variable	negative	
H3K36me3	positive	positive	no
H3K79me1	positive	positive	no
H3K79me2	positive	balanced	no

mark	cell type by cell type	gene by gene	center on TSS
H2AK5ac	positive	positive	
H2BK120ac	positive	positive	
H2BK12ac	positive	positive	
H2BK15ac	positive	balanced	
H2BK20ac	neutral	NA	
H2BK5ac	positive	positive	
H3K4ac	positive	positive	
H3K9ac	positive	positive	
H3K14ac	positive	positive	
H3K18ac	positive	positive	
H3K23ac	positive	positive	
H3K27ac	positive	positive	
H3K56ac	positive	NA	
H4K8ac	positive	positive	
H4K12ac	positive	NA	
H4K91ac	positive	positive	

Conclusions

PEREpigenomics offers interesting visualisations of epigenetic data gathered by Roadmap Epigenomics.

Perspectives

- documentation
- preprint
- develop similar approach for FAANG (Functional Annotation of the Animal Genomes) data

Interactive visualisations to foster data re-use

1. Heat*seq: interactive correlation heatmaps of transcriptomics and epigenomics datasets
2. PEREpigenomics: Profile Explorer of Roadmap Epigenomics data
3. **VizFaDa: Visualisations of FAANG data**

ANR Flash open sciences

VizFaDa: Visualisations of FAANG data

Objective: to provide interactive visualisations of FAANG data straight from the data portal

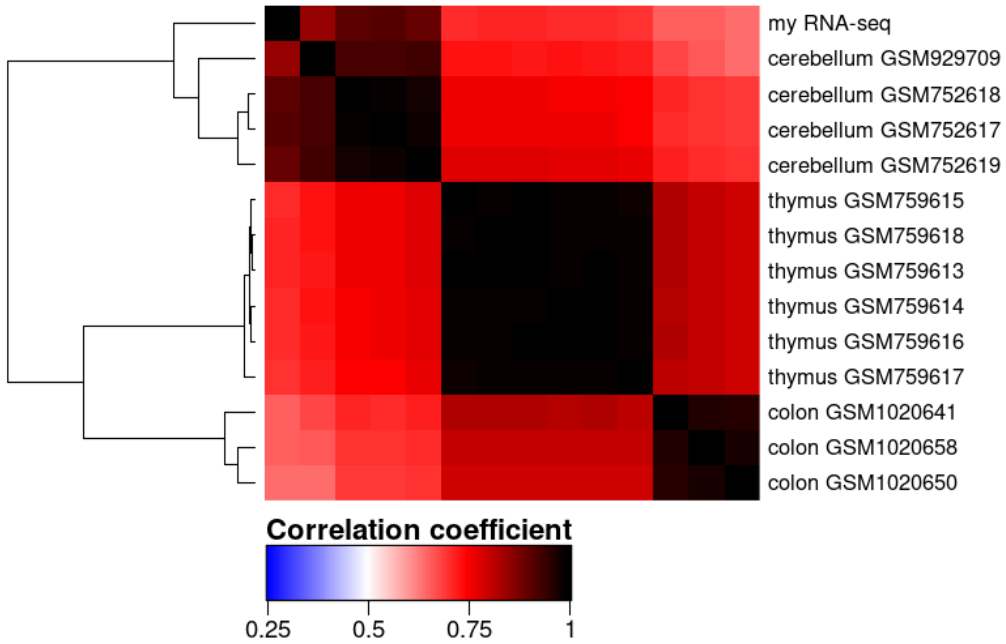
The 'ENCODE' of farm animal genomes

- open international consortium
- sharing data about gene expression and regulation
- metadata standards
- open data: data.faang.org

Expected results

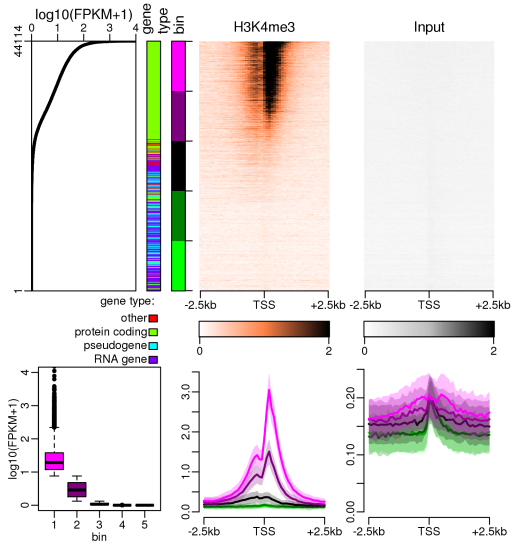
Correlation heatmaps

- broad view of the dataset
- quick comparison with user-provided data
- identification of outlier experiments



Stack profiles of epigenetic marks and gene expression level

- rich and informative visualisations
- can reveal complex biological or artefactual associations



People involved

We are recruiting (IE CDD 18m): genphyse.toulouse.inra.fr/job-offers

GenPhySe:

- **Guillaume Devailly**: data processing & visualisations
- **Sylvain Foissac**: scientific expertise, link with FAANG

Sigenae:

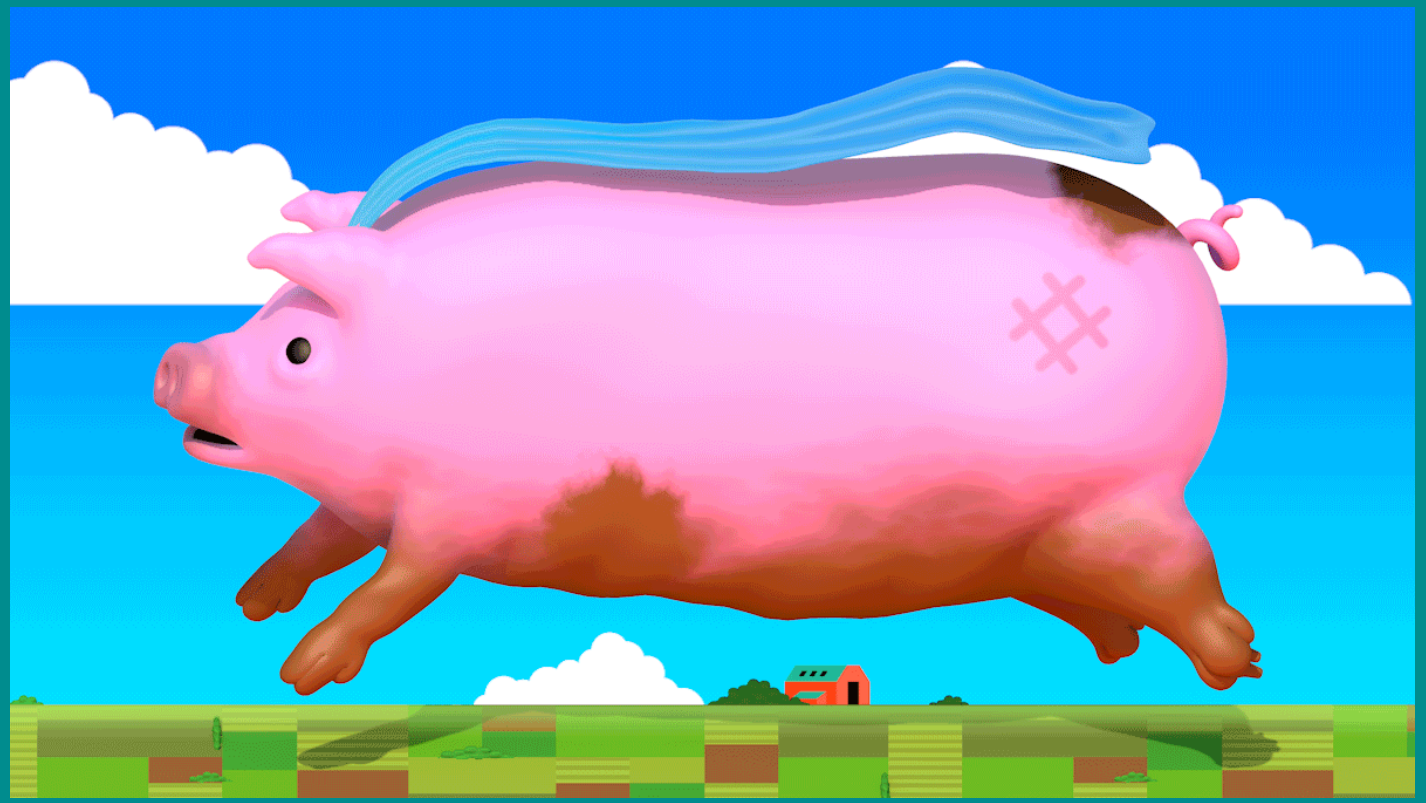
- **Philippe Bardou**: web application development
- **Patrice Dehais**: system administrator

EMBL-EBI:

- **Peter Harrison**: FAANG data coordination centre
- **Guy Cochrane**: FAANG data coordination centre

INRAE GenPhySE

Thank you for your attention!



www.bloomberg.com/news/features/2019-12-03/china-and-the-u-s-are-racing-to-create-a-super-pig

INRAE GenPhySE

Histone modification