Valorisations of epigenomic and transcriptomic data

Guillaume Devailly

GenEpi@GenPhySE

INRA
SCIENCE & IMPACT

Valorisations of epigenomic and transcriptomic data

Guillaume Devailly

@G_Devailly

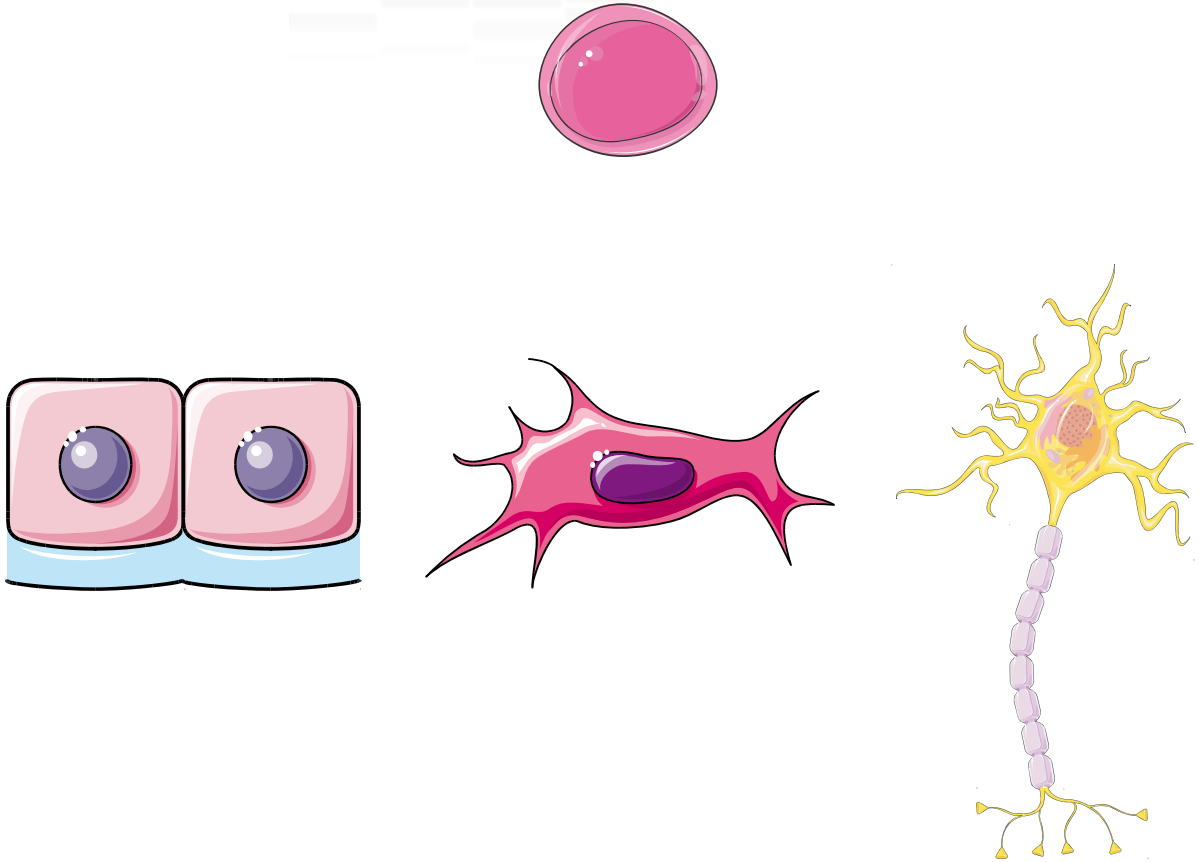
2017-12-19

Summary

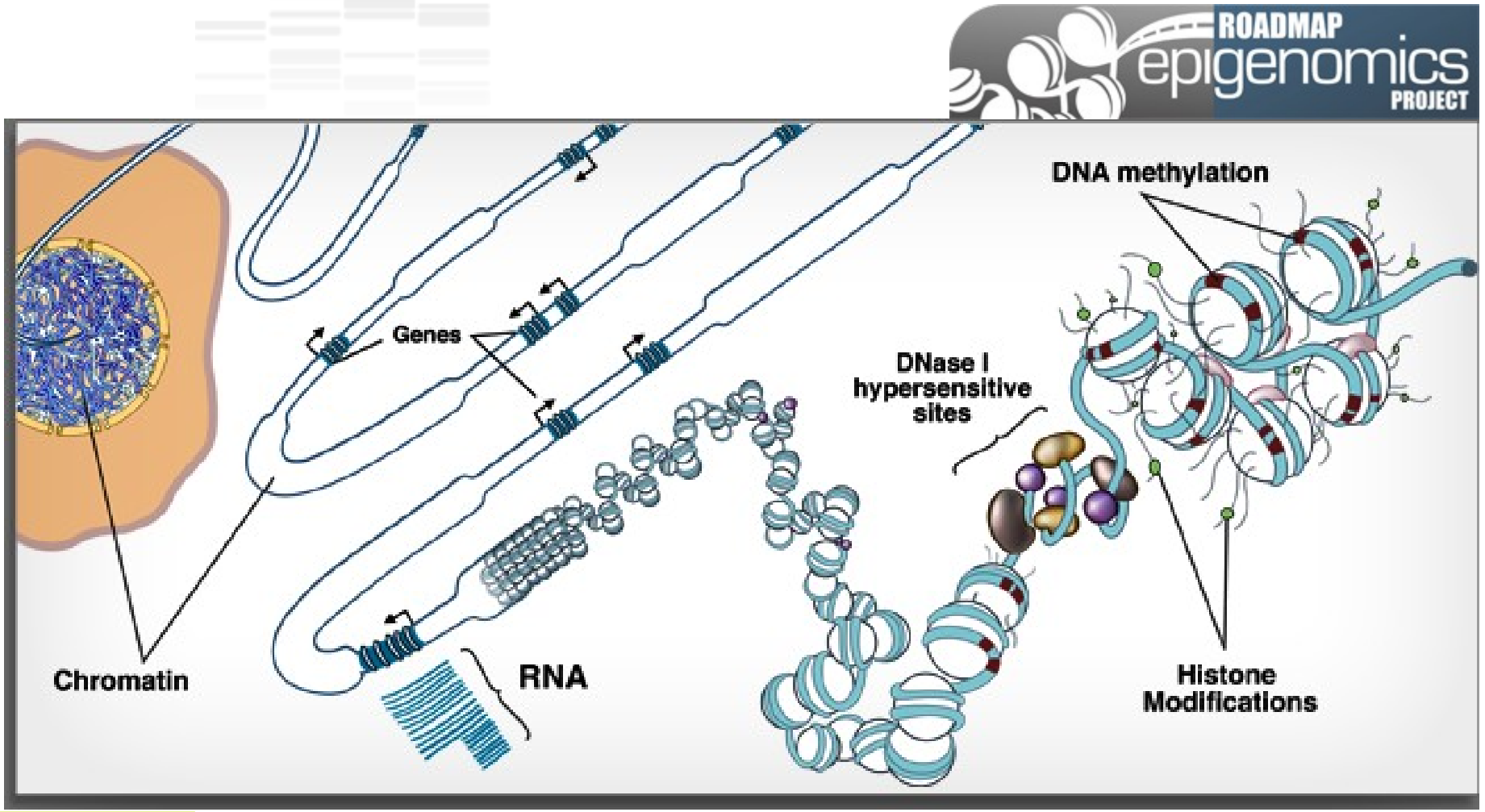
Transcriptionnal regulation: a 2-slides introduction

- MBD2, a major DNA methylation reader
- Heat*seq web app:

  http://www.heatstarseq.roslin.ed.ac.uk/
- Functionnal annotation transfert using co-expression networks
- Epigenomics and the human transcript diversity



- Same genome
- Distinct cellular phenotypes
- Differential gene expression
- Cellular environment + epigenetics

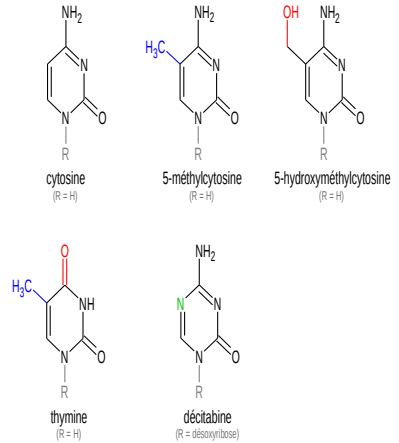

ROADMAP
epigenomics
PROJECT
DNA methylation
Genes
DNase I
hypersensitive
sites
Chromatin
RNA
Histone
Modifications

MBD2 binding dynamics during oncogenic transformation

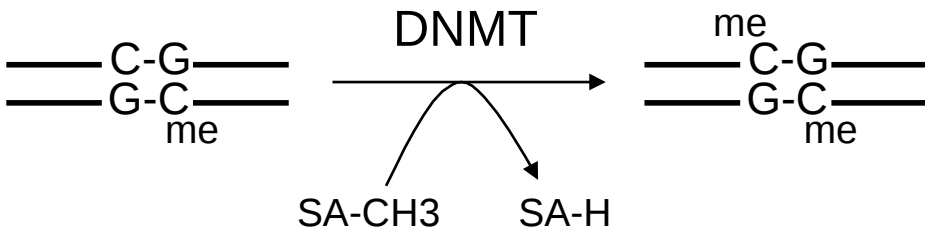
PhD at the Cancer Research Center of Lyon
Supervisor: Robert Dante
Patrick Mehlen's team

Vertebrate DNA methyation

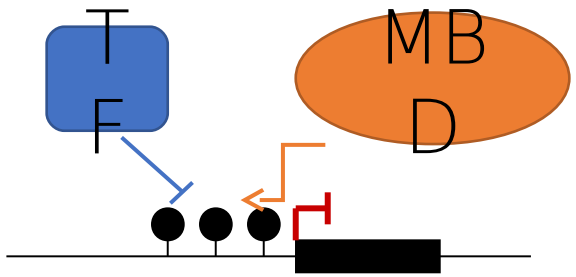
A prototipical epigenetics mark



1- Write



2- Read



3- Erase

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

MBD family

MECP2

MBD 1

MBD2

MBD 4

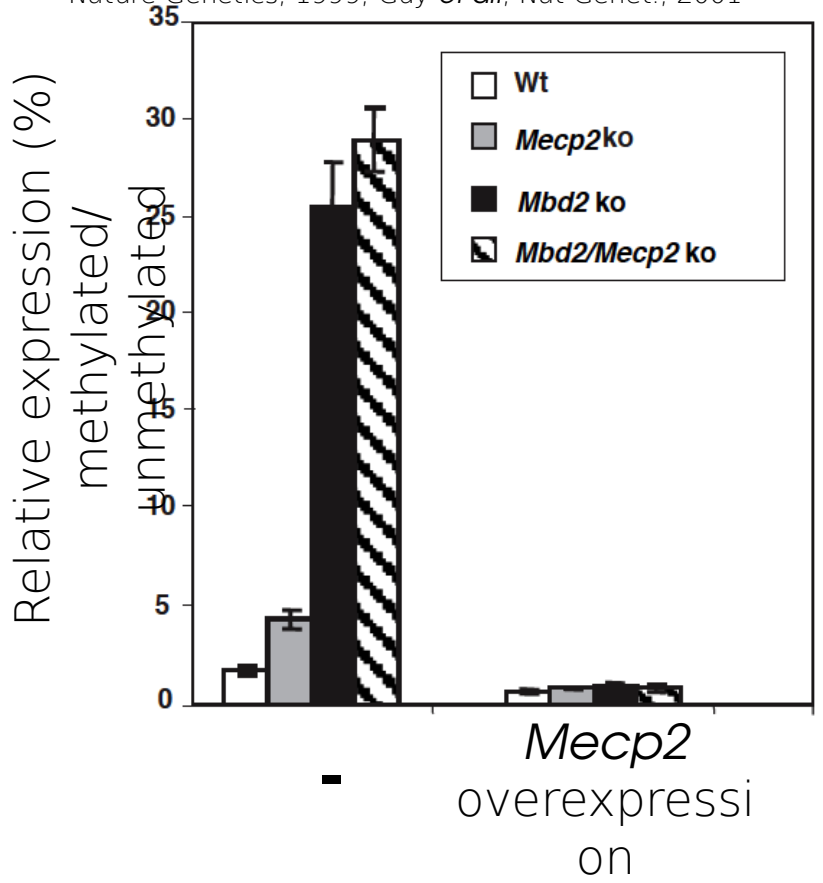
UHRF family

UHRF 1

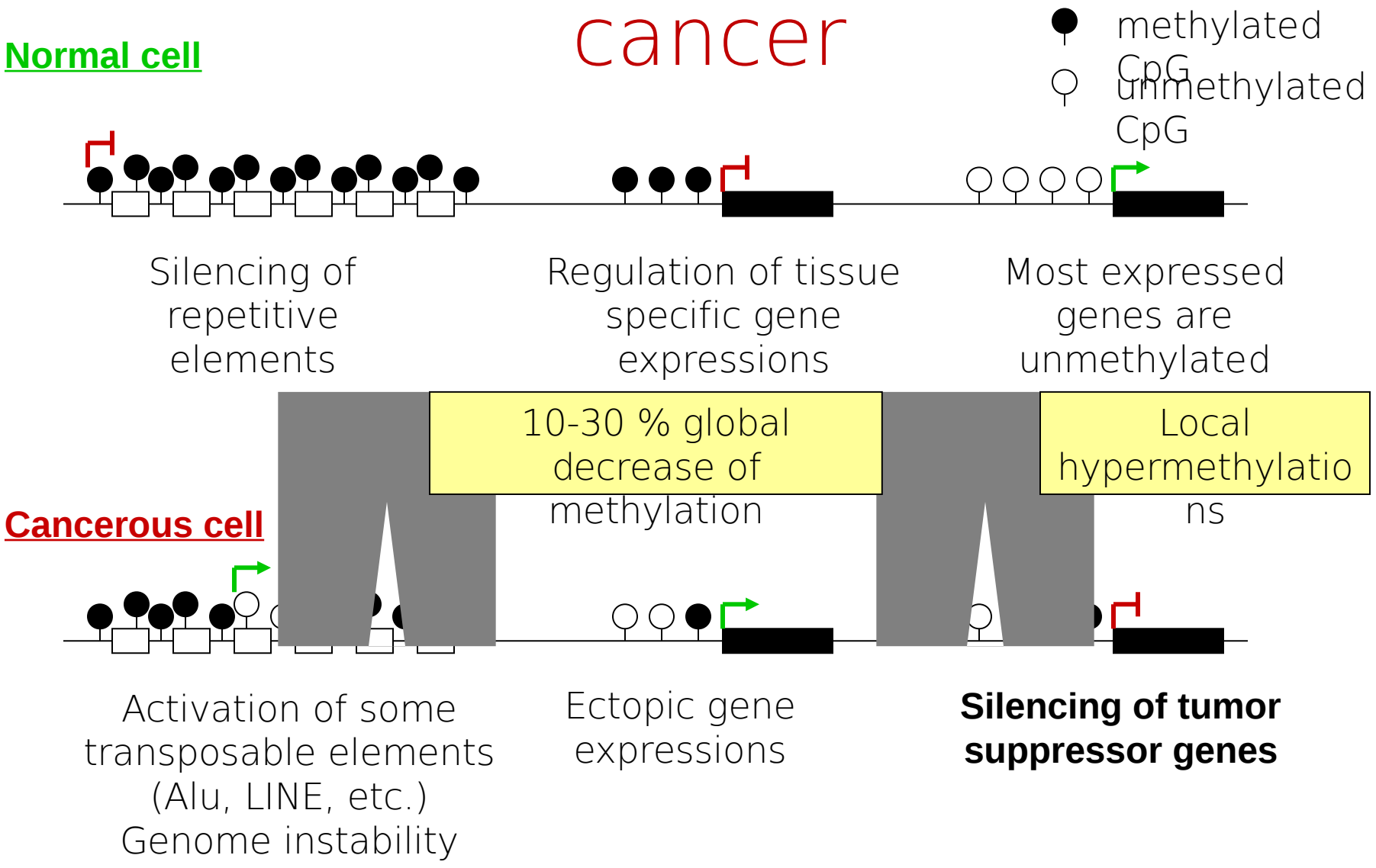
UHRF 2

MBD2 represses transcription of methylated plasmids.

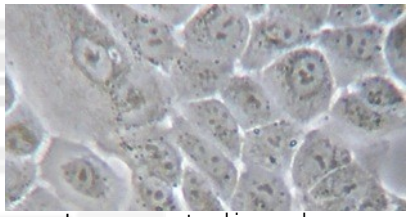
Ng *et al.*, Nature Genetics, 1999, Guy *et al.*, Nat Genet., 2001



DNA methylation alterations in cancer



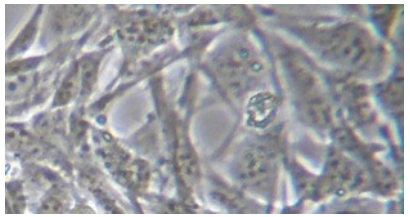
HMEC-hTERT



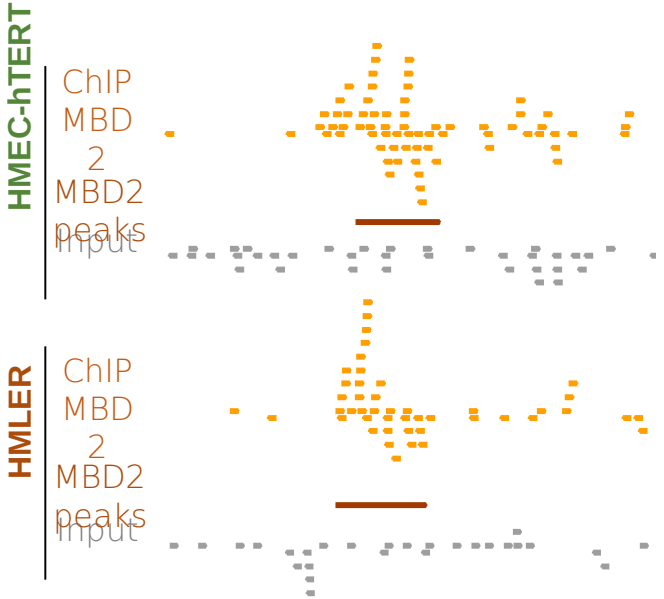
Immortalized
Non tumorigenic

T/t SV40
HRAS G12V

HMLER



Tumorigenic



ChIP-seq:

Endogenous proteins
Pool of 5 experiments

Input
MBD2 ChIP
MeDP (MethylMiner, invitrogen)

RNA-seq:

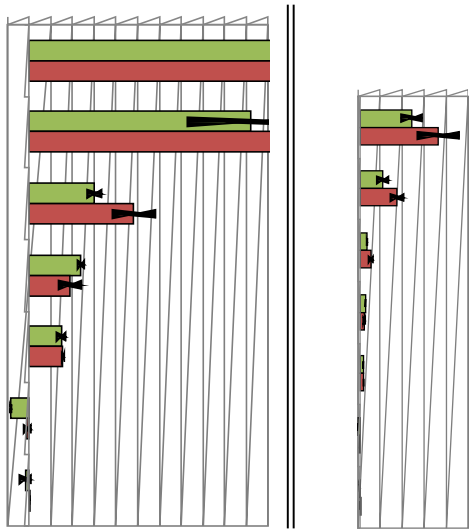
RNA extract from untreated cells or cells treated with **si*MBD2*** or **DAC**

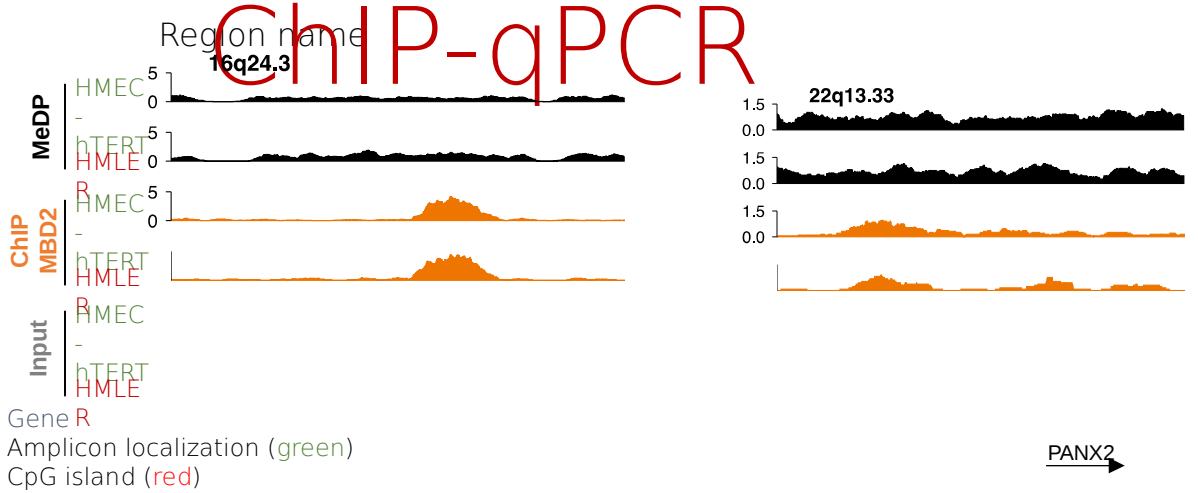
High throughput sequencing

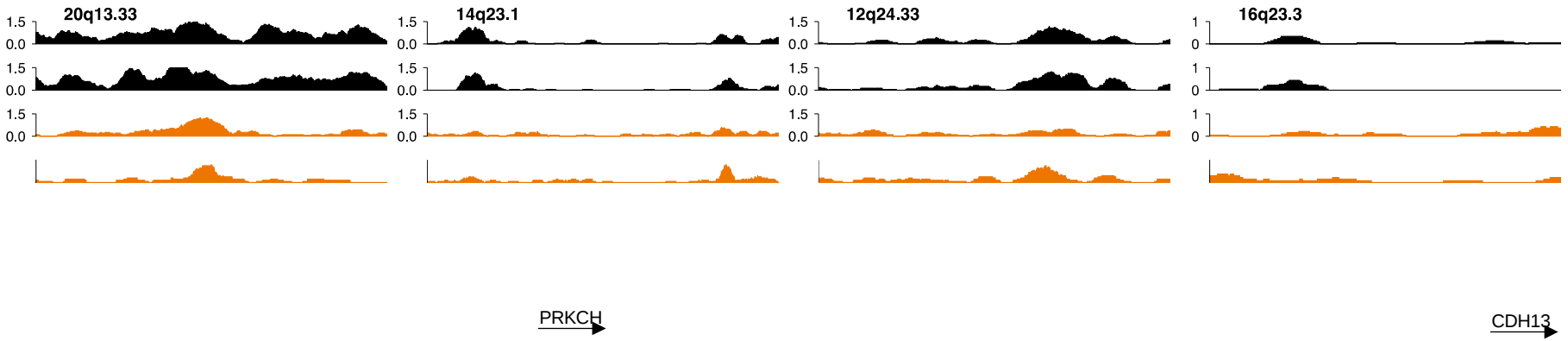
30 - 50 millions of 50 bp reads per conditions
Illumina Hi$eq 2000

华大基因
BGI
ProfileXpert - LCMT
MGX

MBD2 ChIPseq validation 1/3: ChIP-qPCR

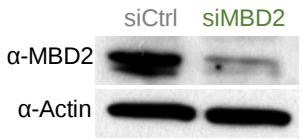


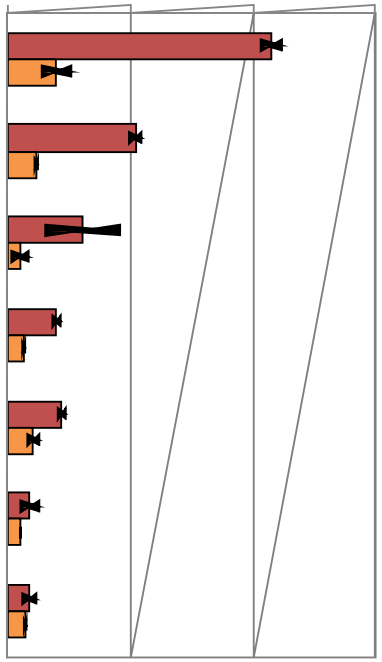


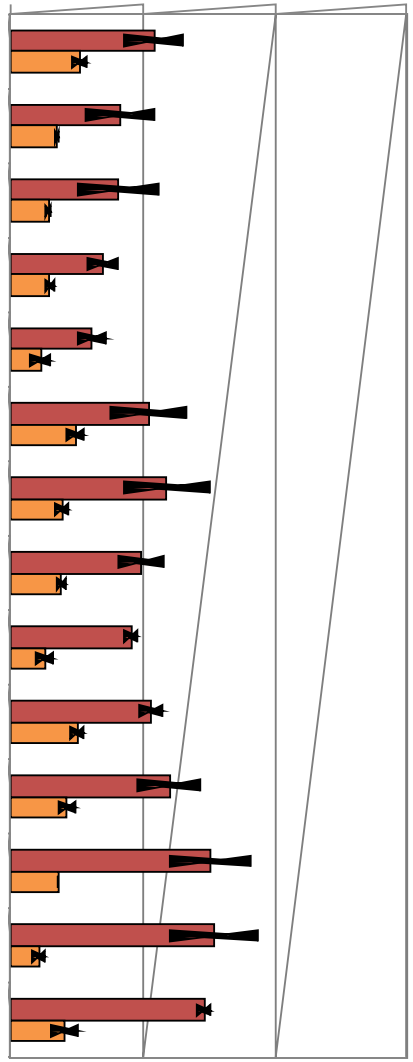


MBD2 ChIPseq validation 2/3:

MBD2 siRNA

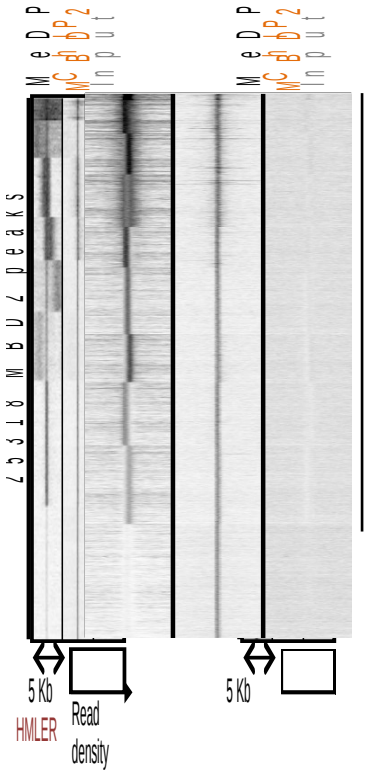






Endogenous MBD2 binds methylated DNA *in vivo*

Almost all MBD2 binding sites are methylated

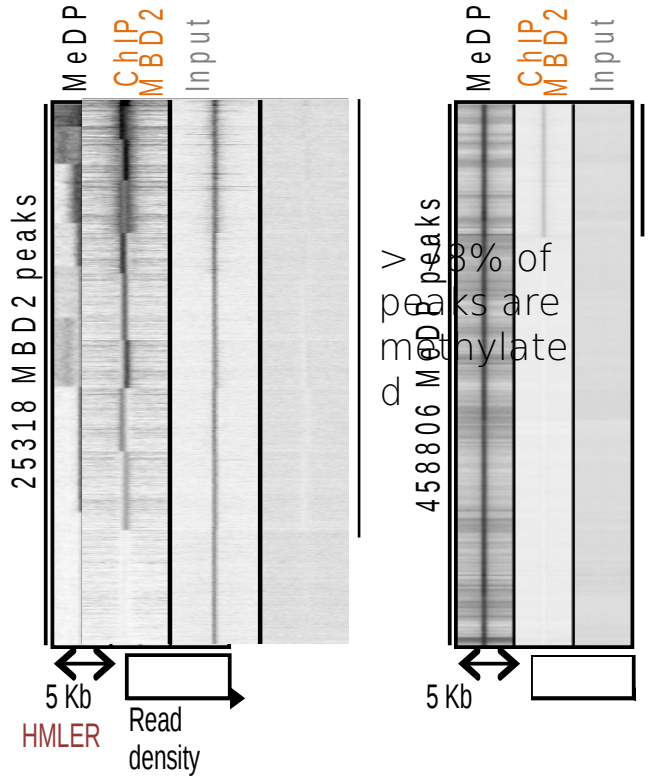


> 78% of peaks are methylated

Endogenous MBD2 binds methylated DNA *in vivo*

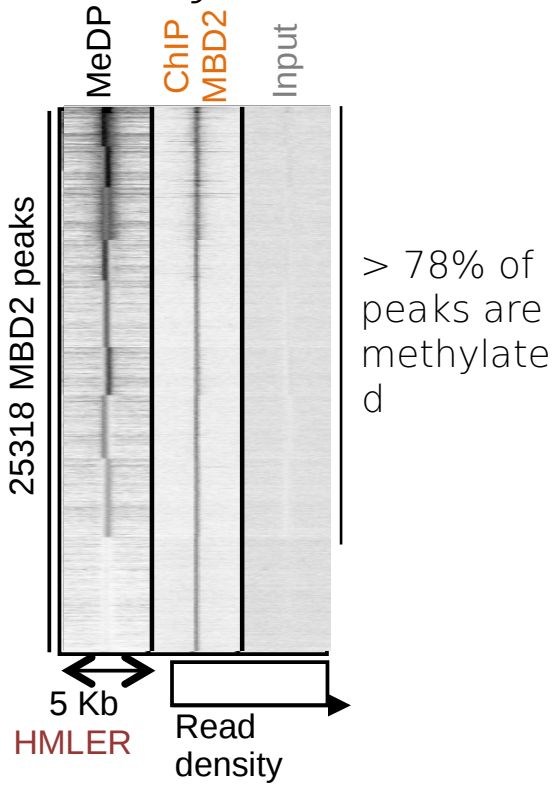
Almost all MBD2 binding sites are methylated

1/4th of methylated DNA regions are bound by endogenous MBD2

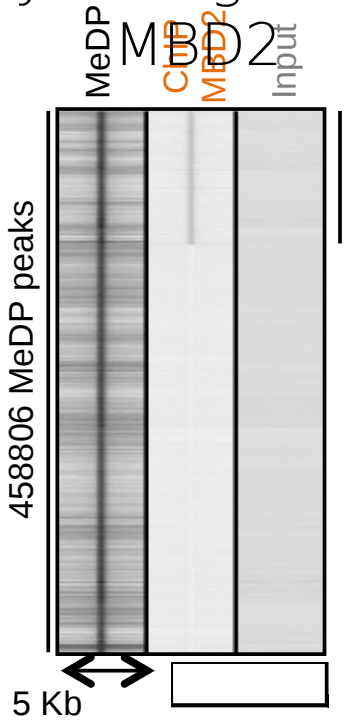


Endogenous MBD2 binds methylated DNA *in vivo*

Almost all MBD2 binding sites are methylated

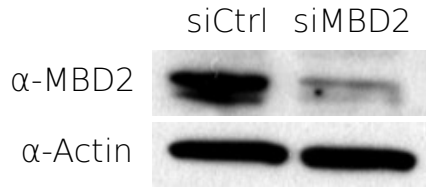


1/4th of methylated DNA regions are bound by endogenous MBD2

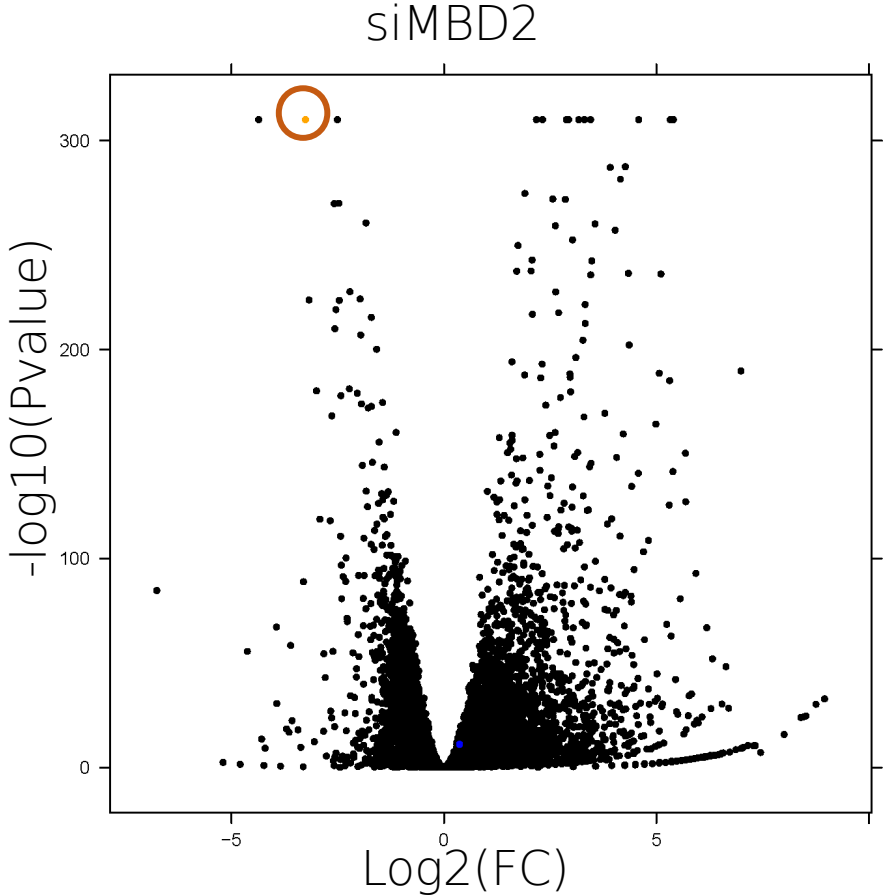


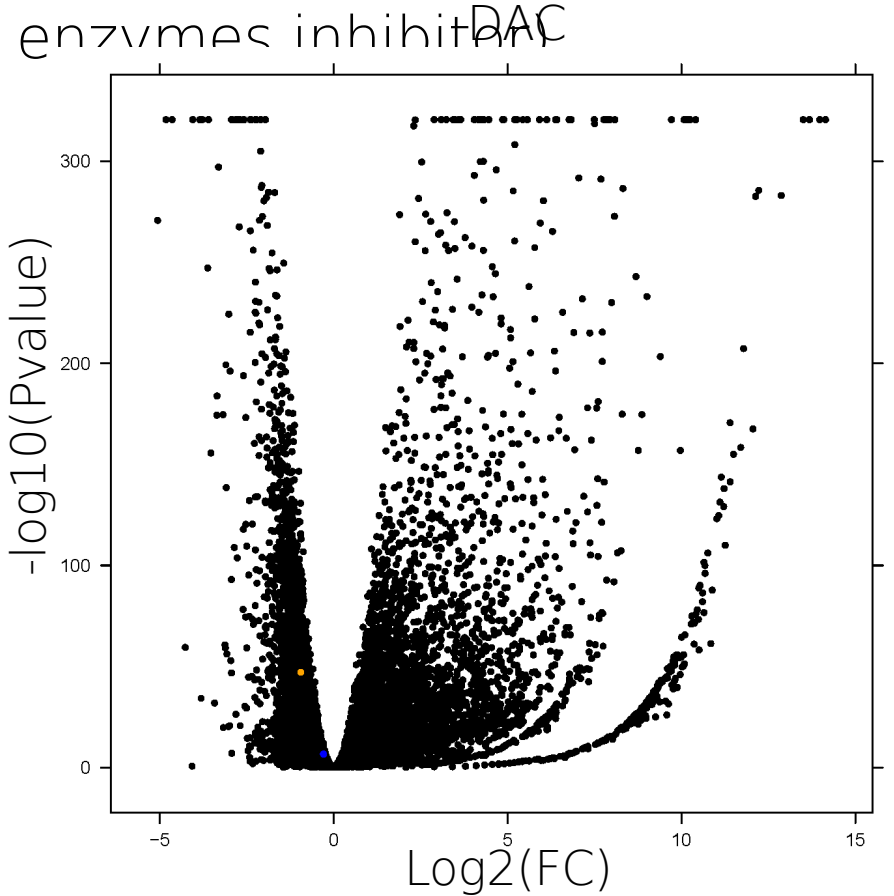
MBD2 binding on CpG island depends on their methylation levels

Endogenous MBD2 is a DNA methylation dependent transcriptional repressor

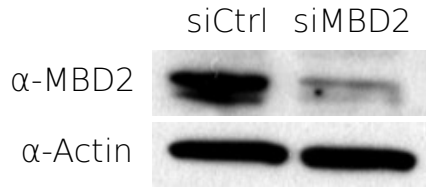


RNAseq after depletion of MBD2 by siRNA,
or after DAC (5-aza-deoxycytidine) treatment (DNA methylation enzymes inhibitor)





MBD2 is a DNA methylation dependent transcriptional repressor

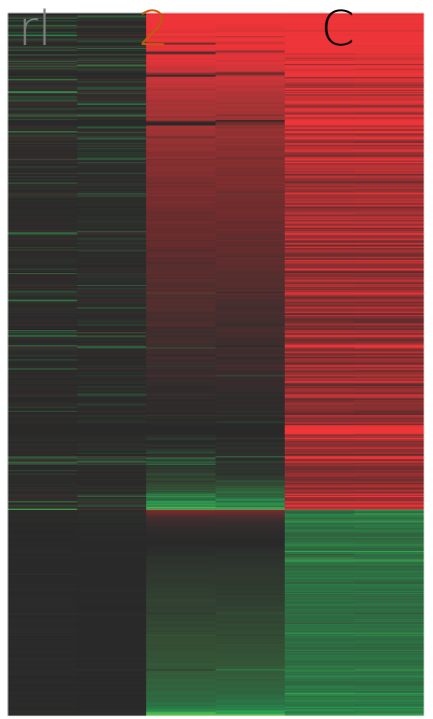


RNAseq after depletion of MBD2 by siRNA,
or after DAC (5-aza-deoxycytidine) treatment (DNA methylation enzymes inhibitor)

2164 genes up by siMBD2

70% also up by DAC

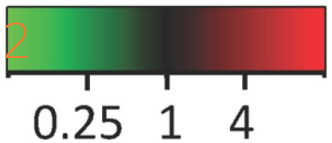
667 genes down by siMBD2
40% also up down DAC



3510 genes up by DAC

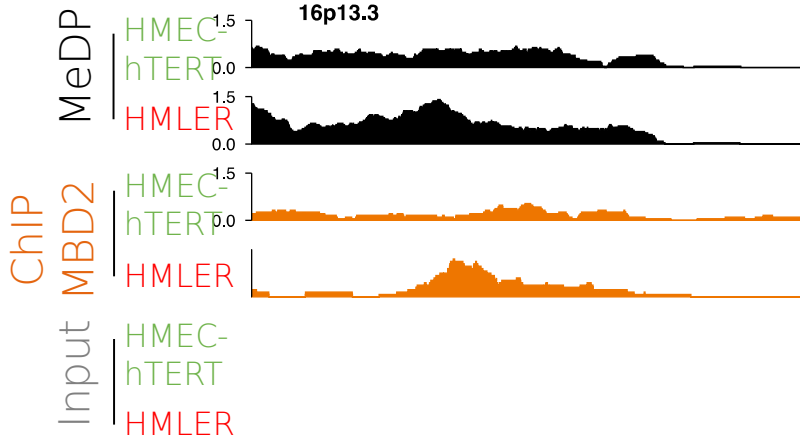
43% also up by siMBD2

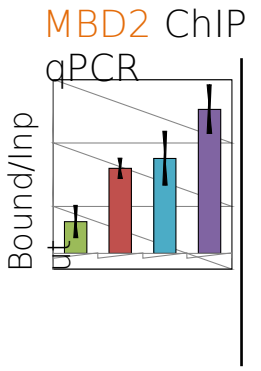
1459 genes down by DAC
18% also up down siMBD2

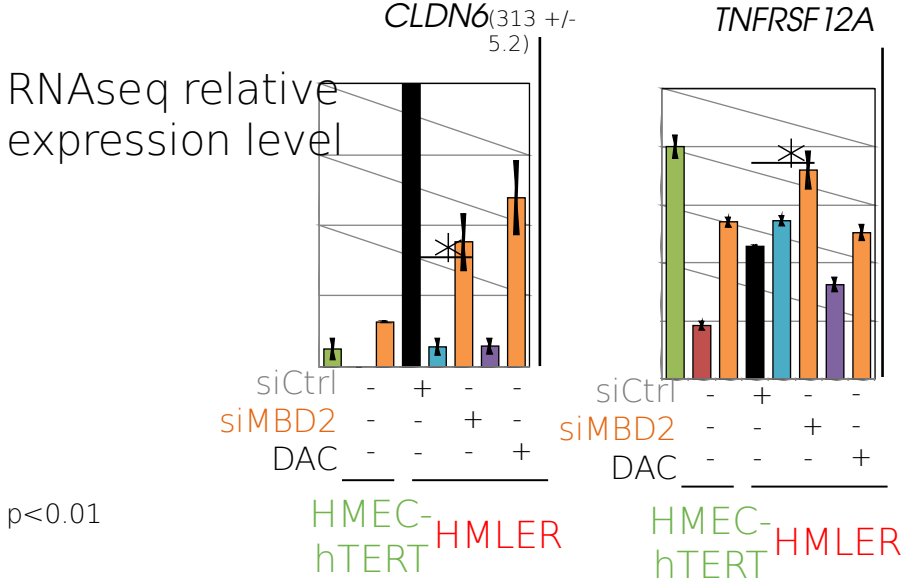


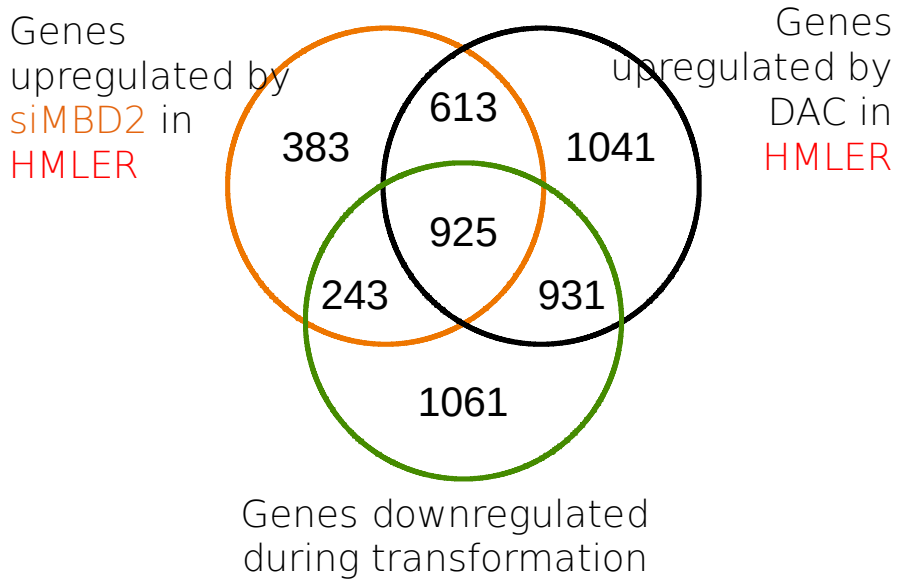
HMLER

Redistribution of MBD2 during oncogenic transformation: example of *CLDN6/TNFRSF12A*



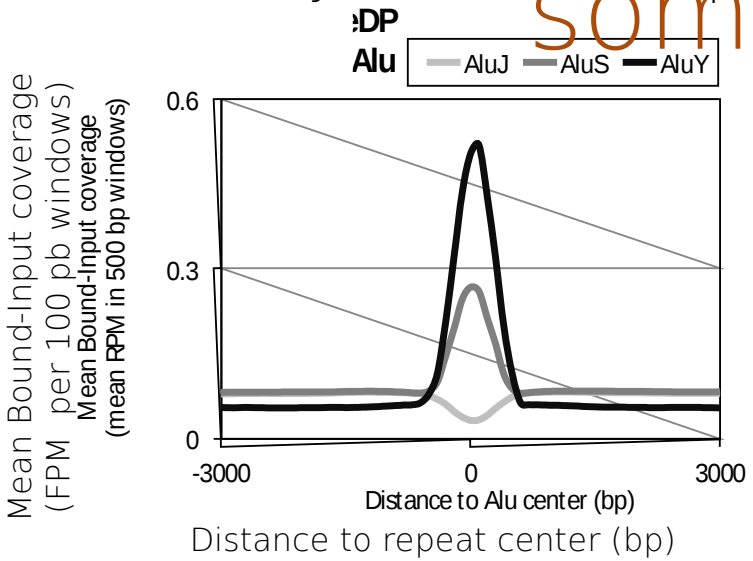




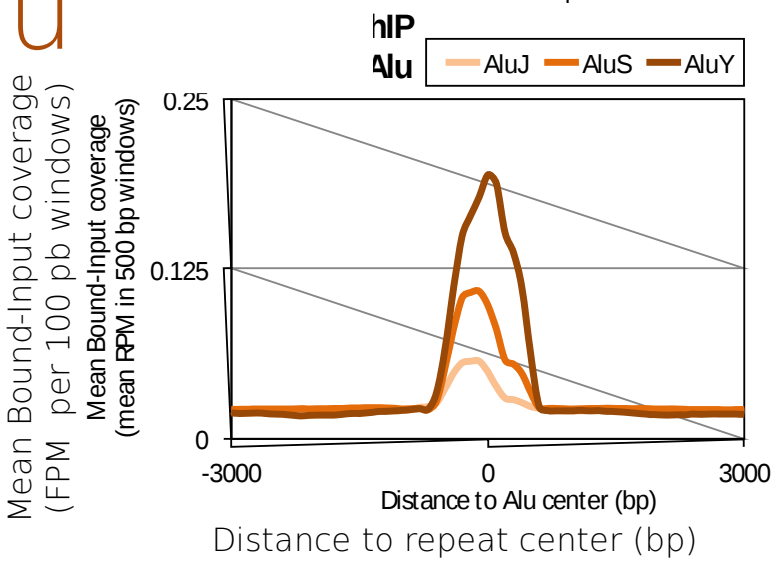


MBD2 represses transcription of some AluY

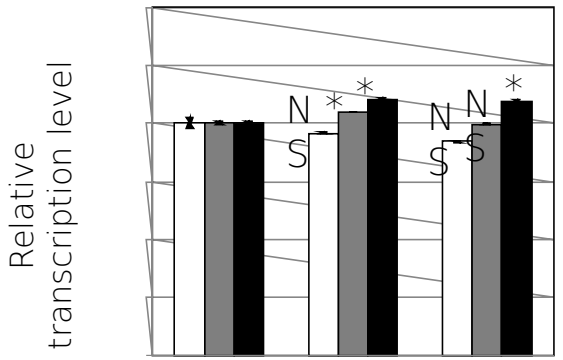
DNA methylation (MeDPseq)



MBD2 ChIPseq



Transcription level (RNAseq)



Conclusions & Perspectives

- **MBD2** is a major reader of DNA methylation, at least in studied cell lines.
- **MBD2** is redistributed during oncogenic transformation. This redistribution plays role in the transcriptome modifications of transformed cell lines.
- What are the mechanism implicated in the redistribution of MBD2?
- Does targeting MBD2 reduces the transformed phenotype?

http://ngs-qc.org/navi/index.php

Accession	Organism	Sample	Experiment	Mapped Reads	QC stamp
GSM1544114	Homo sapiens	N/A	MBD2 ChIP-seq	57,009,909	BBB
GSM1544111	Homo sapiens	N/A	MBD2 ChIP-seq	57,767,880	BBB
GSM972973	Mus musculus	ESC, stem cell	MBD2 ChIP-seq	28,211,050	CCB
GSM2534656	Homo sapiens	immortalized cell	MBD2 ChIP-seq	23,401,218	CCC
GSM1322266	Homo sapiens	N/A	MBD2 ChIP-seq	32,869,475	CCC
GSM972994	Mus musculus	ESC, stem cell	MBD2 ChIP-seq	19,671,342	CDD
GSM2527610	Homo sapiens	K-562	MBD2 ChIP-seq	23,414,725	CDD
GSM1388122	Homo sapiens	N/A	MBD2 ChIP-seq	27,212,212	DCC
GSM972978	Mus musculus	N/A	MBD2 ChIP-seq	25,225,988	DCC
GSM2527611	Homo sapiens	K-562	MBD2 ChIP-seq	27,605,362	DCC
GSM1077269	Mus musculus	ESC, stem cell	MBD2 ChIP-seq	27,360,560	DDC
GSM1006707	Homo sapiens	HeLa	MBD2-V5 ChIP-seq	10,943,628	DDD

Thanks

Robert Dante
Mélodie Grandin
Pauline Mathot
Véronique Corset
Catherine Guy
Solène Le Guervenel
Duygu Ozmadenci
Laury Perriaud
Anne-Pierre Morel
Patrick Mehlen

Collaborators

CNRS UMR7216, Université Paris 7

Pierre-Antoine Defossez
Olivier Kirsh
Benoit Miotto
Audrey Roussel-Gervais

ROSLIN

Heat*seq

Compare your HTS experiment with

RNA-seq
ChIP-seq
CAGE

datasets in

Guillaume Devailly
Anagha Joshi's group, The Roslin Institute, University of Edinburgh
@G_Devailly

THE UNIVERSITY of EDINBURGH

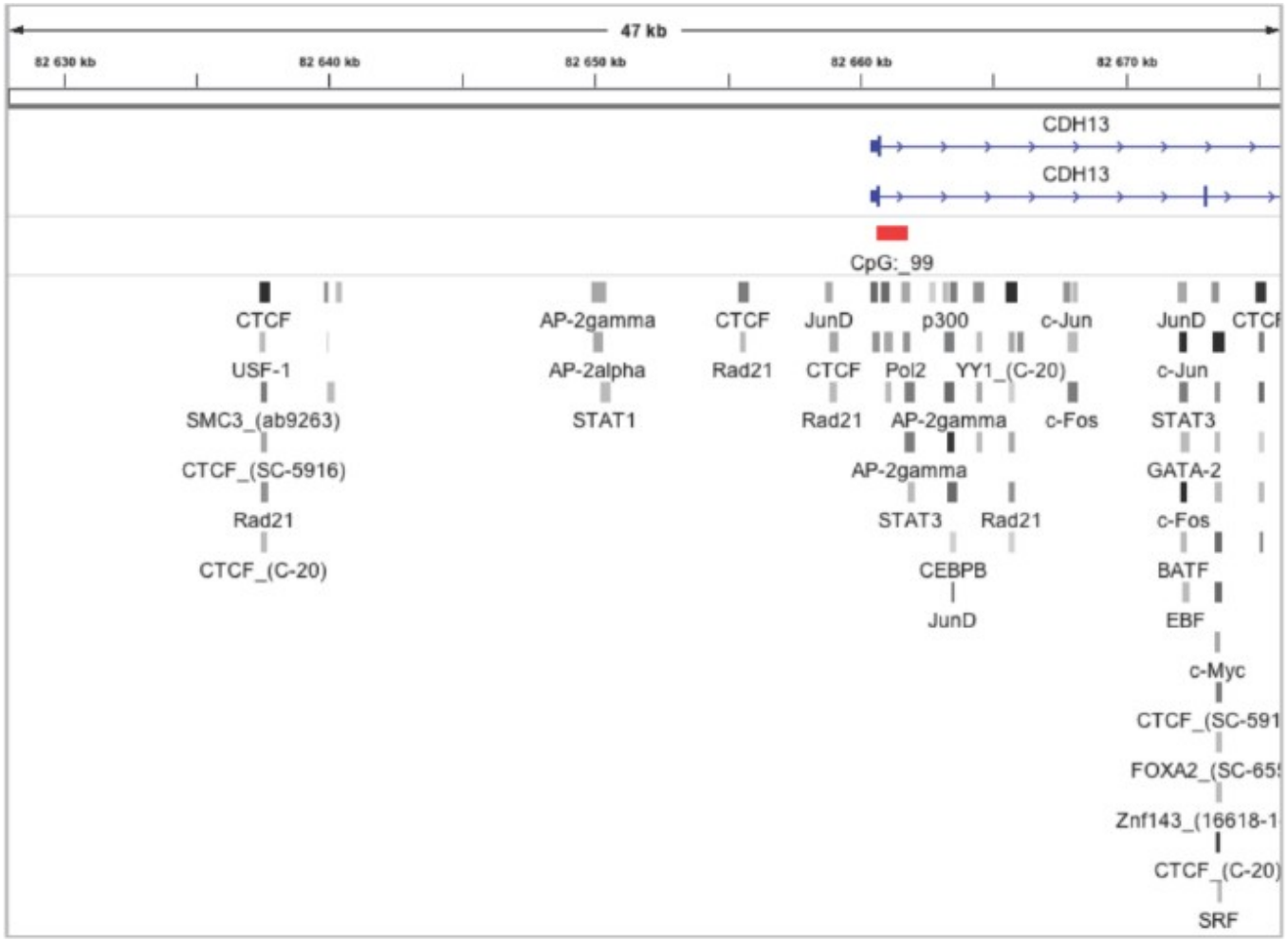
BBSRC

Heat*Seq

HTS is cheap

Big public datasets

Genome-wide comparison is challenging



Workflow

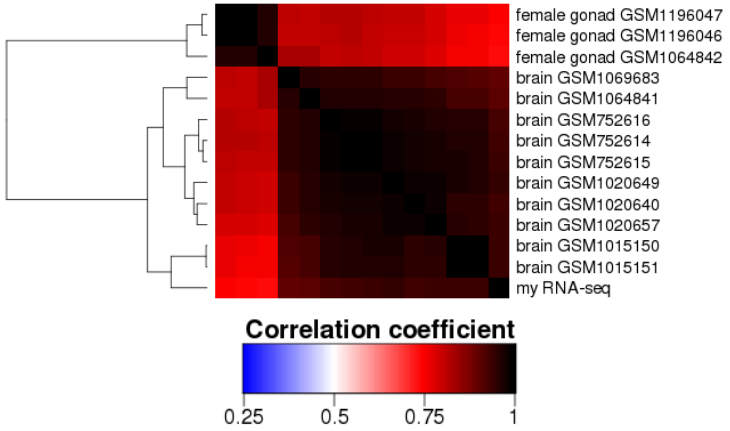
1- Visit website:
ww.heatstarseq.roslin.ed.ac.uk

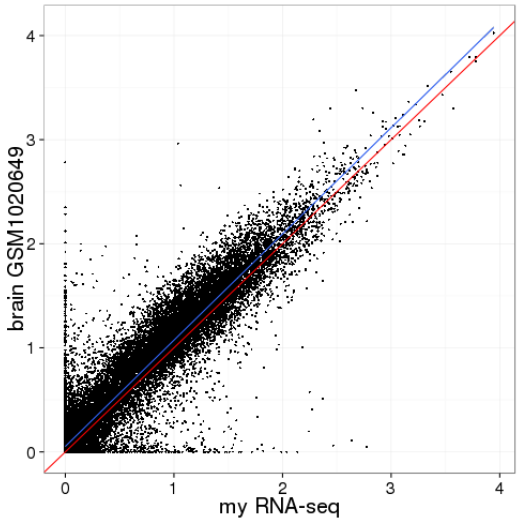
2- Select dataset

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
brain GSM1020640	0.9374279





Pearson correlation coefficient: 0.9497
Spearman correlation coefficient: 0.902

Datasets

Assay	Dataset	Organism	Number of experiments
RNA-seq	Bgee	human	77
		mouse	109
	Blueprint epigenome	human	163
	ENCODE	human	302
		mouse	192
	Roadmap Epigenomics	human	57
	Flybase	drosophila	124
	GTEx	human	8555
TF ChIP-seq	ENCODE	human	690
		mouse	156
	CODEX	human	238
		mouse	651
	modENCODE	drosophila	85

HeatRNAseq

Expression matrix

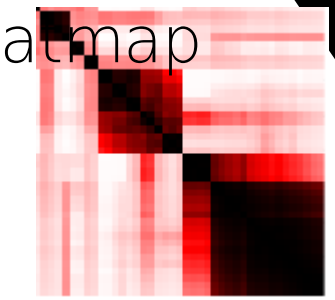
Gene name (~40.000)	Exp 1	Exp 2	Exp 3	…	Exp 77
ENSG00000000003	3.983	2.361	10.216	…	80.583
ENSG00000000005	0.071	0.260	0.000	…	3.329
ENSG00000000419	10.277	2.893	14.153	…	42.639
…	…	…	…	…	
ENSG00000273493	0.000	0.000	0.000		

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	…
	0.663	0.680	0.706	…	1

Clustered heatmap



HeatChIPseq

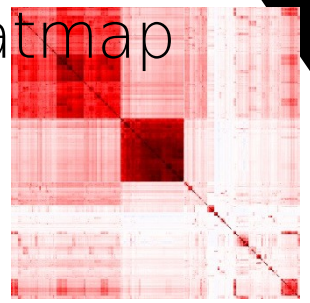
Binary peak matrix

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	…	…

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	…
…	0.035	0.118	0.047	…	1

Clustered heatmap



File format

Tab-delimited text file

RNA-seq

HeatRNAseq

geneID	tpm
ENSG00000134046	120.12
ENSG00000141644	0
ENSG00000169057	85.24
ENSG00000174282	0.54
ENSG00000187098	42
...	...

Ensembl or Flybase gene id

ChIP-seq

HeatChIPseq

CAGE

HeatCAGEseq

File format

Tab-delimited text file

RNA-seq

HeatRNAseq

ChIP-seq

HeatChIPseq

CAGE

HeatCAGEseq

chr	start	end
chr1	125423	125891
chr1	8545032	8546254
chr4	4523698	4524785
chr12	854120	854870
chrX	2458750	2459872
...	...	...

Same genome version than the dataset

File format

Tab-delimited text file

RNA-seq

HeatRNAseq

ChIP-seq

HeatChIPseq

CAGE

HeatCAGEseq

chr	start	end	name	rpm	strand
chr2	25486325	25486487	CAGEpeak_1	458.12	+
chr6	5896321	5896380	CAGEpeak_2	25.03	+
chr6	223541	223602	CAGEpeak_3	1.23	-
chr17	5012035	5012100	CAGEpeak_4	45.3	+
chr21	960032	960098	CAGEpeak_5	8.70	-
...	...	...	...	...	...

Implementation

Source code:
https://github.com/gdevailly/HeatStarSeq_gh

plotly

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

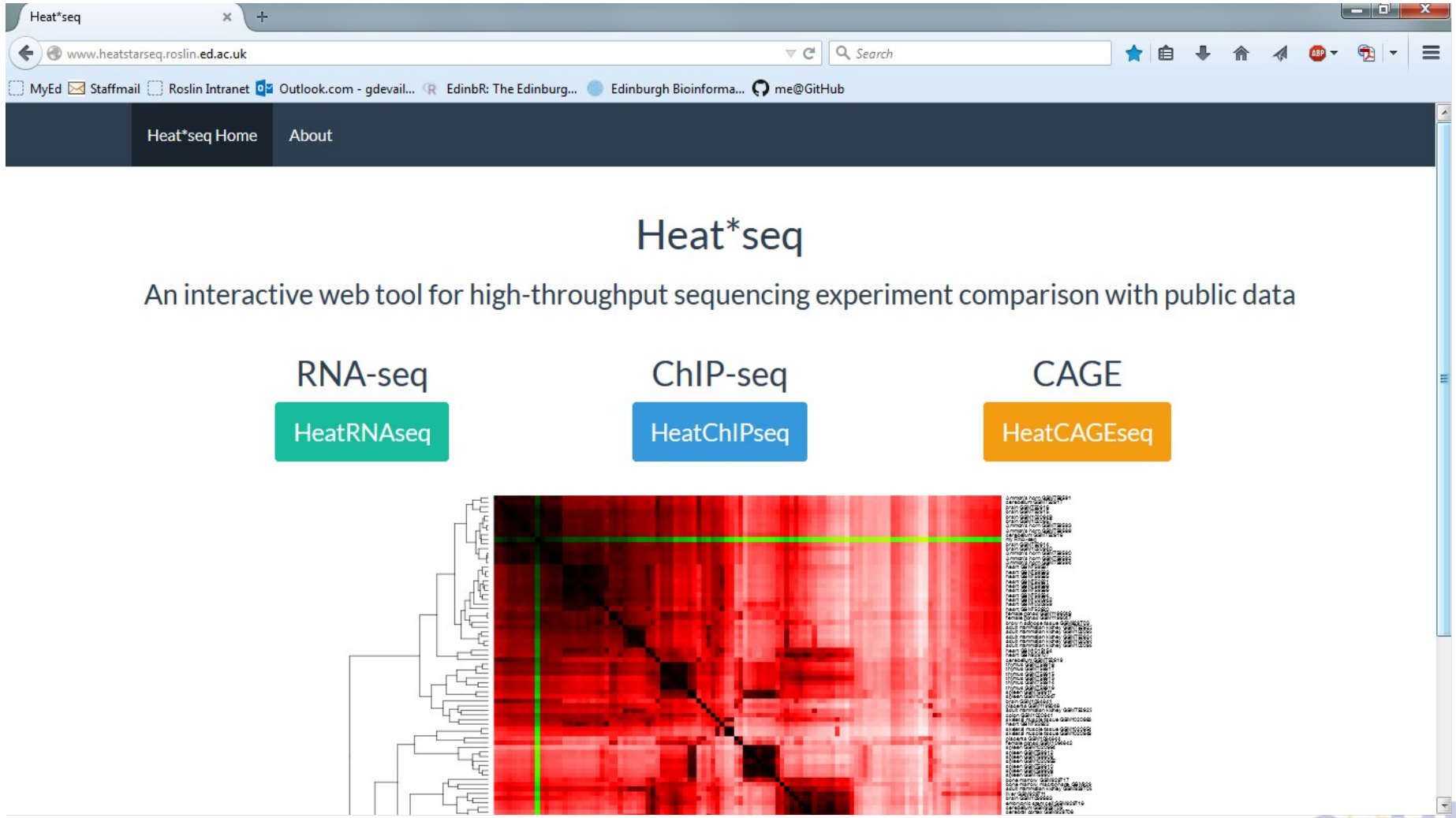
BBSRC

Icons:

Live demonstration

www.heatstarseq.roslin.ed.ac.uk

If live demo fails…



HeatRNAseq

1 - Select a dataset

Bgee RNA-seq (mouse)

2 - Load your data (optional)

File formating instructions

- (•) Upload your expression file
- ( ) Use the example file

Choose a file:

Browse... No file selected.

☑ **The expression file contains a header.**

Name of your experiment:

my RNA-seq

3 - Plot customization

☑ **Highlight my experiment in the heatmap.**

Tissue (empty to select all):

Developmental stage (empty to select all):

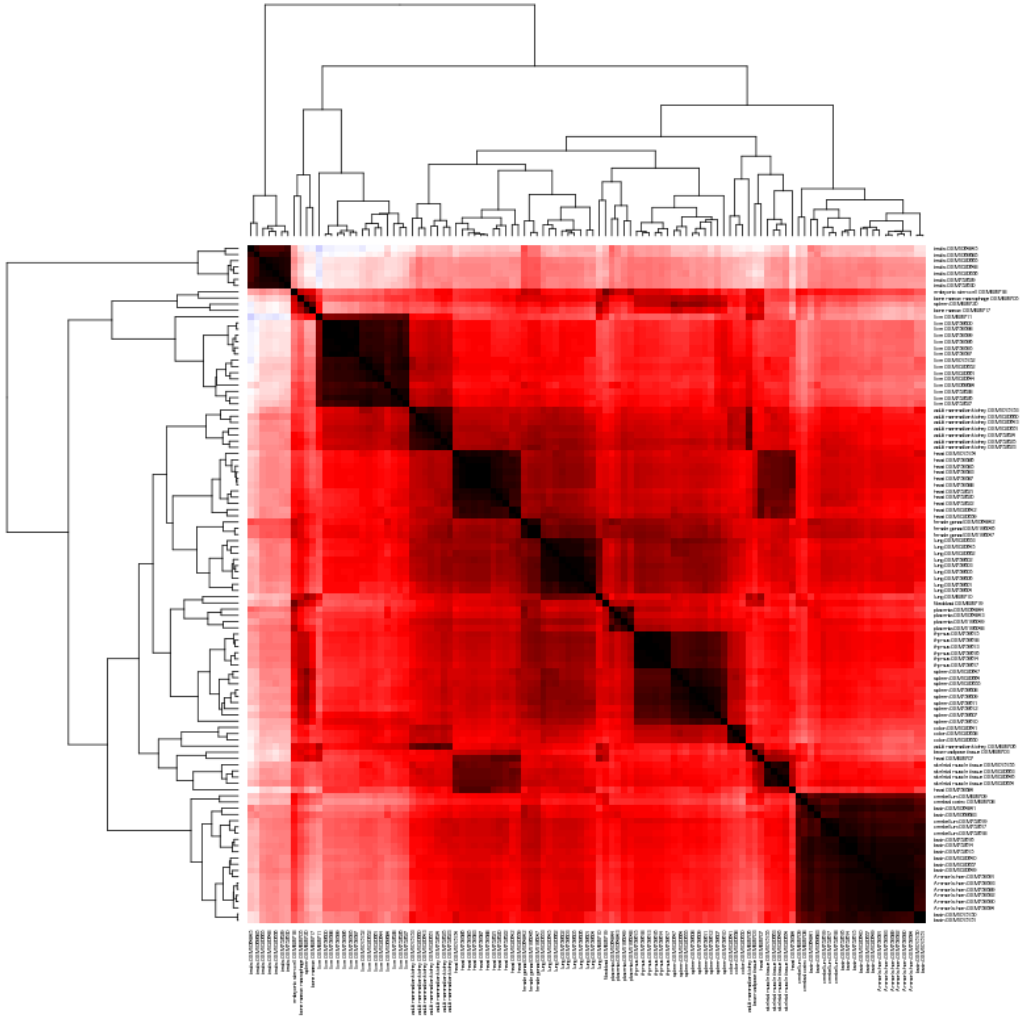
Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options

My expression file | Correlation table | Static heatmap | Responsive heatmap | Tree | Pairwise plot | Samples metadata



Heat*seq | Use application | Instructions

My expression file | Correla

Samples metadata

HeatRNAseq

1 - Select a dataset

Bgee RNA-seq (mouse)

2 - Load your data (optional)

File formating instructions

- Upload your expression file
- Use the example file

Choose a file:

Browse... No file selected.

The expression file contains a header.

Name of your experiment:

my RNA-seq

3 - Plot customization

Highlight my experiment in the heatmap.

Tissue (empty to select all):

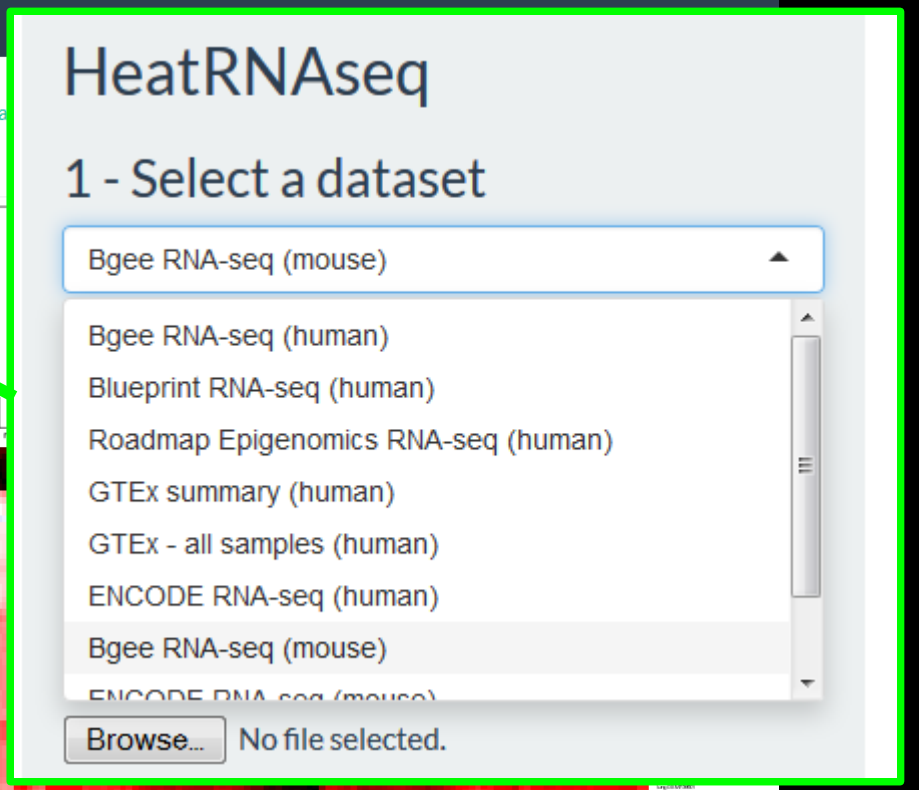
Developmental stage (empty to select all):

Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options



HeatRNAseq

1 - Select a dataset

Bgee RNA-seq (mouse)

2 - Load your data (optional)

File formating instructions

- (•) Upload your expression file
- ( ) Use the example file

Choose a file:

Browse... No file selected.

☑ **The expression file contains a header.**

Name of your experiment:

my RNA-seq

3 - Plot customization

☑ **Highlight my experiment in the heatmap.**

Tissue (empty to select all):

Developmental stage (empty to select all):

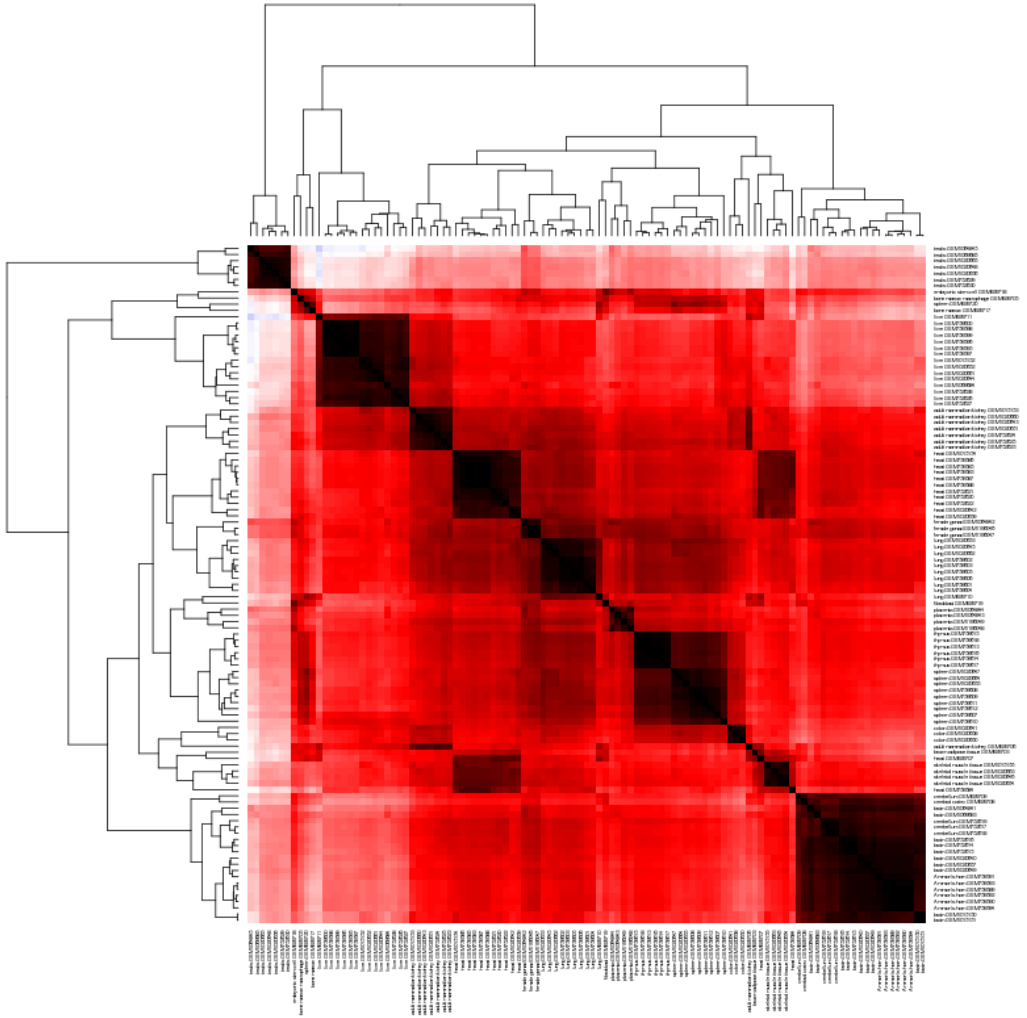
Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options

My expression file | Correlation table | Static heatmap | Responsive heatmap | Tree | Pairwise plot | Samples metadata



HeatRNAseq

1 - Select a dataset

Bgee RNA-seq (mouse)

2 - Load your data (optional)

File formating instructions

- (•) Upload your expression file
- ( ) Use the example file

Choose a file:

Browse... mK3_1.txt

Upload complete

☑ **The expression file contains a header.**

Name of your experiment:

kidney mk3 cell line

3 - Plot customization

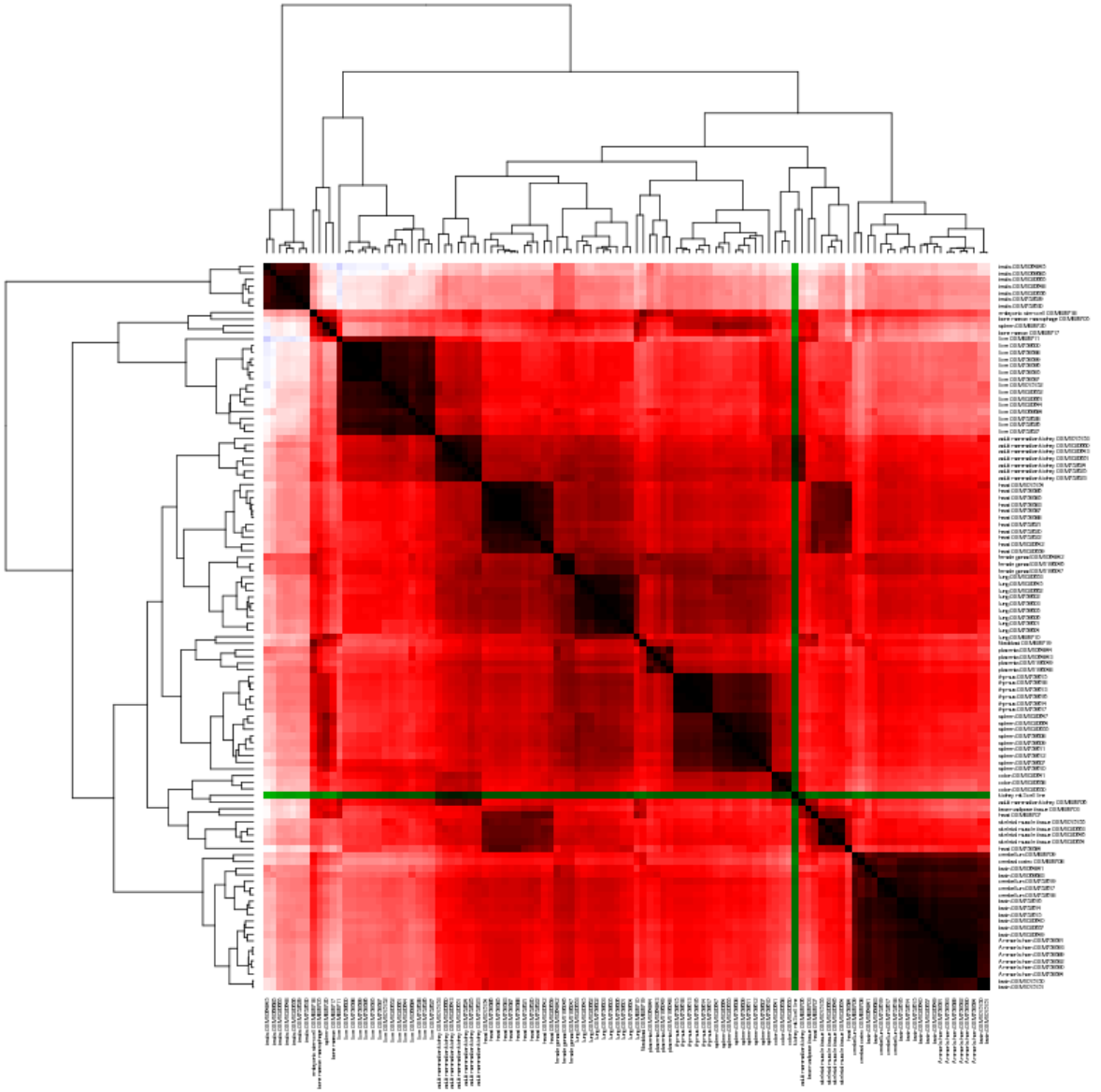
☑ **Highlight my experiment in the heatmap.**

Tissue (empty to select all):

Developmental stage (empty to select all):

Library type (empty to select all):

My expression file | Correlation table | Static heatmap | Responsive heatmap | Tree | Pairwise plot | Samples metadata



3 - Plot customization

☑ Highlight my experiment in the heatmap.

Tissue (empty to select all):

Developmental stage (empty to select all):

Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options

Label size:

◉ Automatic

○ Adjust manually

Show dendrogram(s)?

both

Show labels?

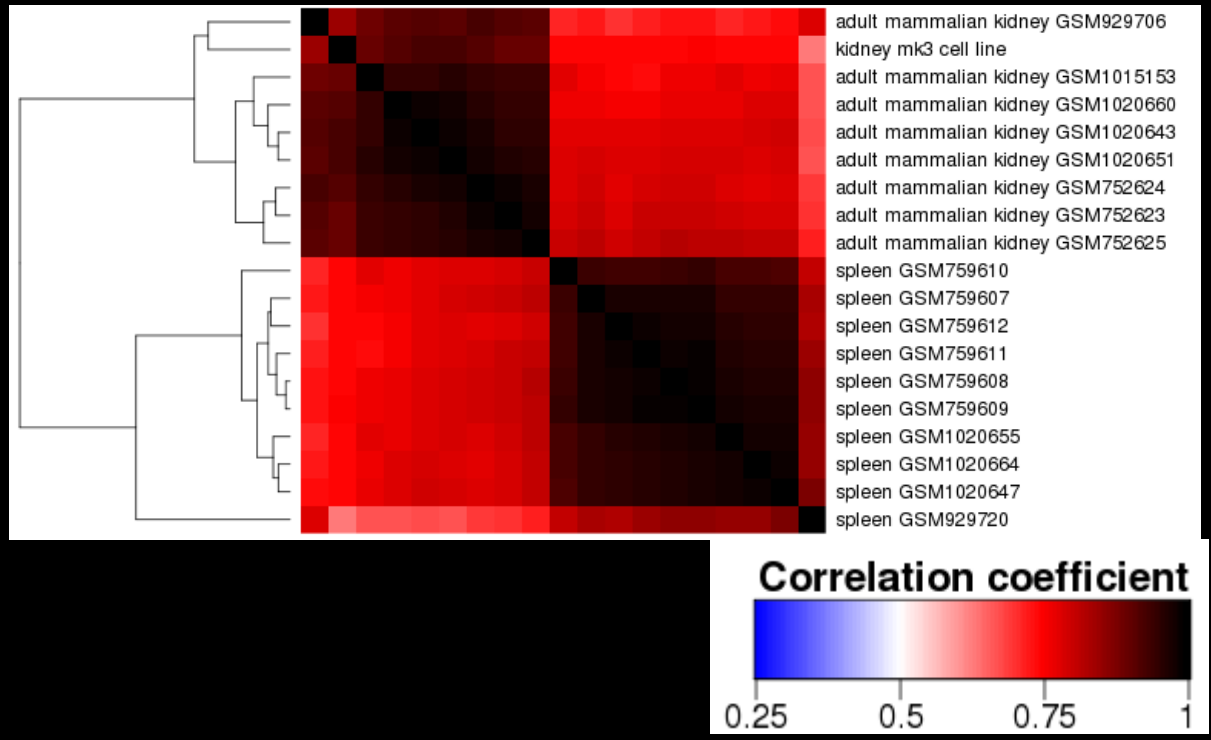
both

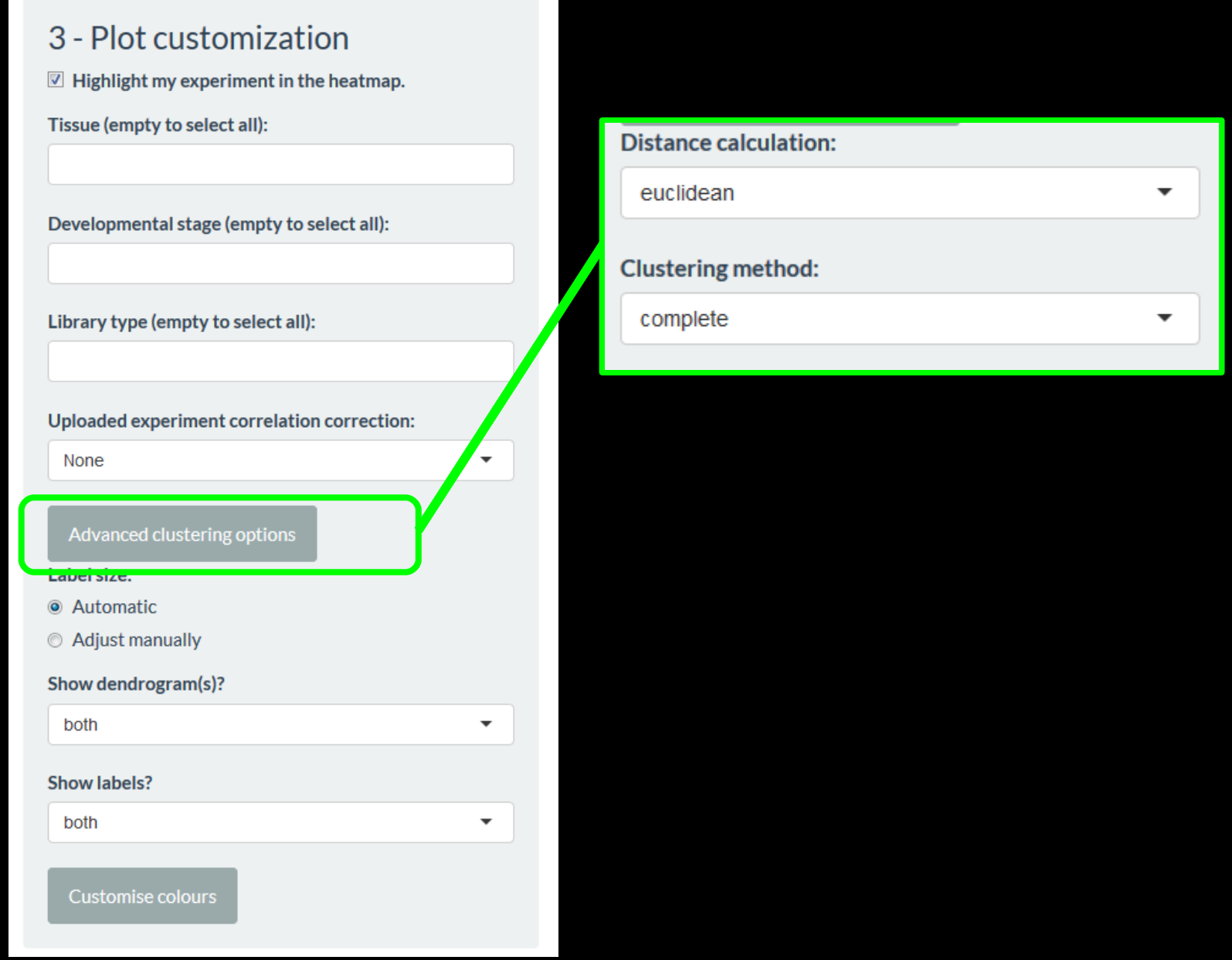
Customise colours

Tissue (empty to select all):

adult mammalian kidney b

brain
bone marrow
brown adipose tissue
bone marrow macrophage
cerebellum
fibroblast
cerebral cortex
embryonic stem cell




3 - Plot customization
Highlight my experiment in the heatmap.
Tissue (empty to select all):
Developmental stage (empty to select all):
Library type (empty to select all):
Uploaded experiment correlation correction:
None
Advanced clustering options
Automatic
Adjust manually
Show dendrogram(s)?
both
Show labels?
both
Customise colours
Distance calculation:
euclidean
Clustering method:
complete

3 - Plot customization

☑ Highlight my experiment in the heatmap.

Tissue (empty to select all):

Developmental stage (empty to select all):

Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options

Label size:

◉ Automatic

○ Adjust manually

Show dendrogram(s)?

both

Show labels?

both

Customise colours

Customise colours

Colour 1:

#FFFFFF

Value 1:

-1 0.7

-1 -0.83 -0.66 -0.49 -0.32 -0.15 0.02 0.19 0.36 0.53 0.7

Colour 2:

#FFFFFF

Value 2:

0.75 0.8

0.75 0.76 0.76 0.77 0.77 0.78 0.78 0.79 0.79 0.8 0.8

Colour 3:

#FF8800

Value 3:

0.85 0.9

0.85 0.85 0.86 0.86 0.87 0.88 0.88 0.89 0.89 0.9 0.9

Colour 4:

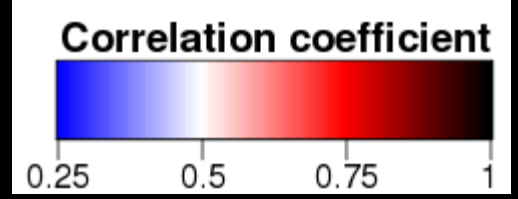
#000000

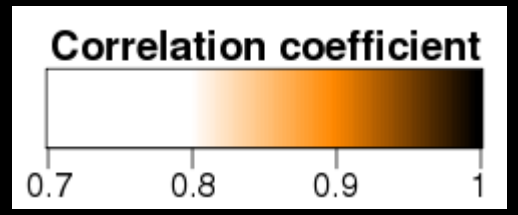
Value 4:

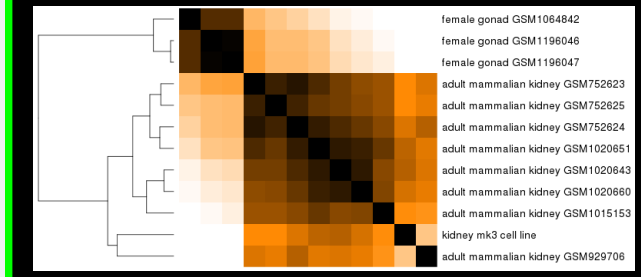
0.95 1

0.95 0.95 0.96 0.96 0.97 0.97 0.98 0.98 0.99 0.99 1

Apply colour changes







File formating instructions

Upload your expression file

Use the example file

Choose a file:

Browse... No file selected.

The expression file contains a header.

Name of your experiment:

my RNA-seq

3 - Plot customization

Highlight my experiment in the heatmap.

Tissue (empty to select all):

Developmental stage (empty to select all):

Library type (empty to select all):

Uploaded experiment correlation correction:

None

Advanced clustering options

Label size:

Automatic

Adjust manually

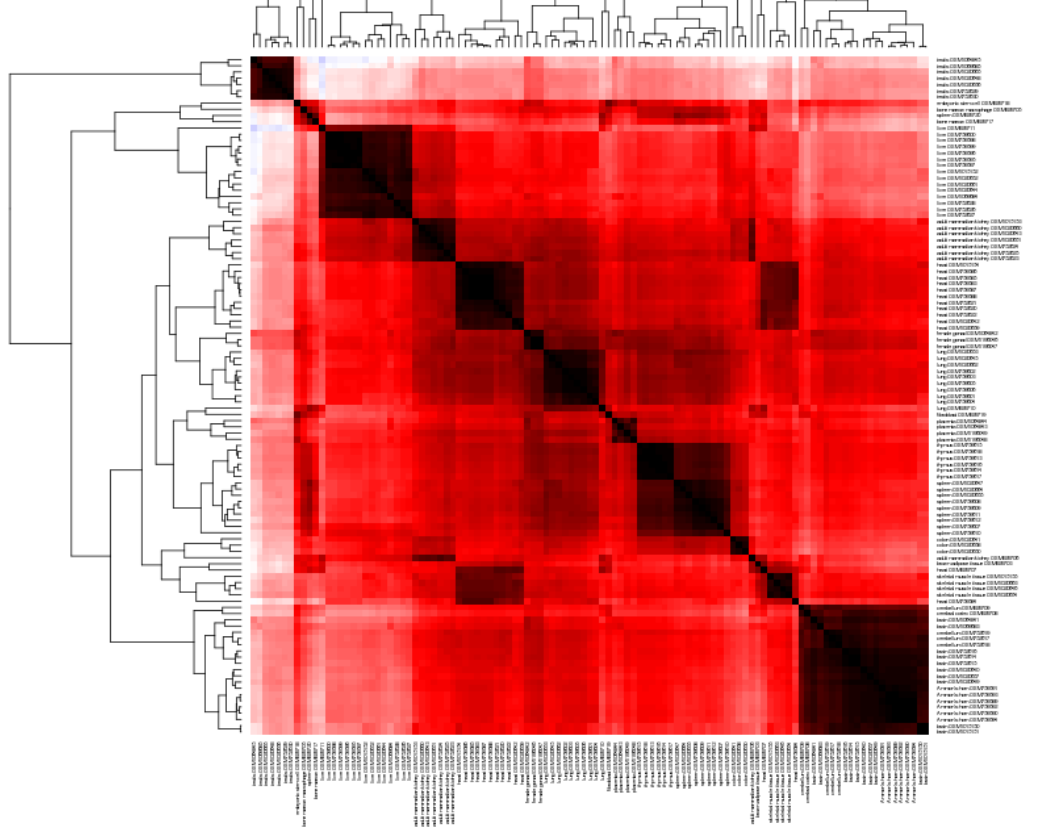
Show dendrogram(s)?

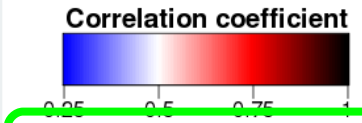
both

Show labels?

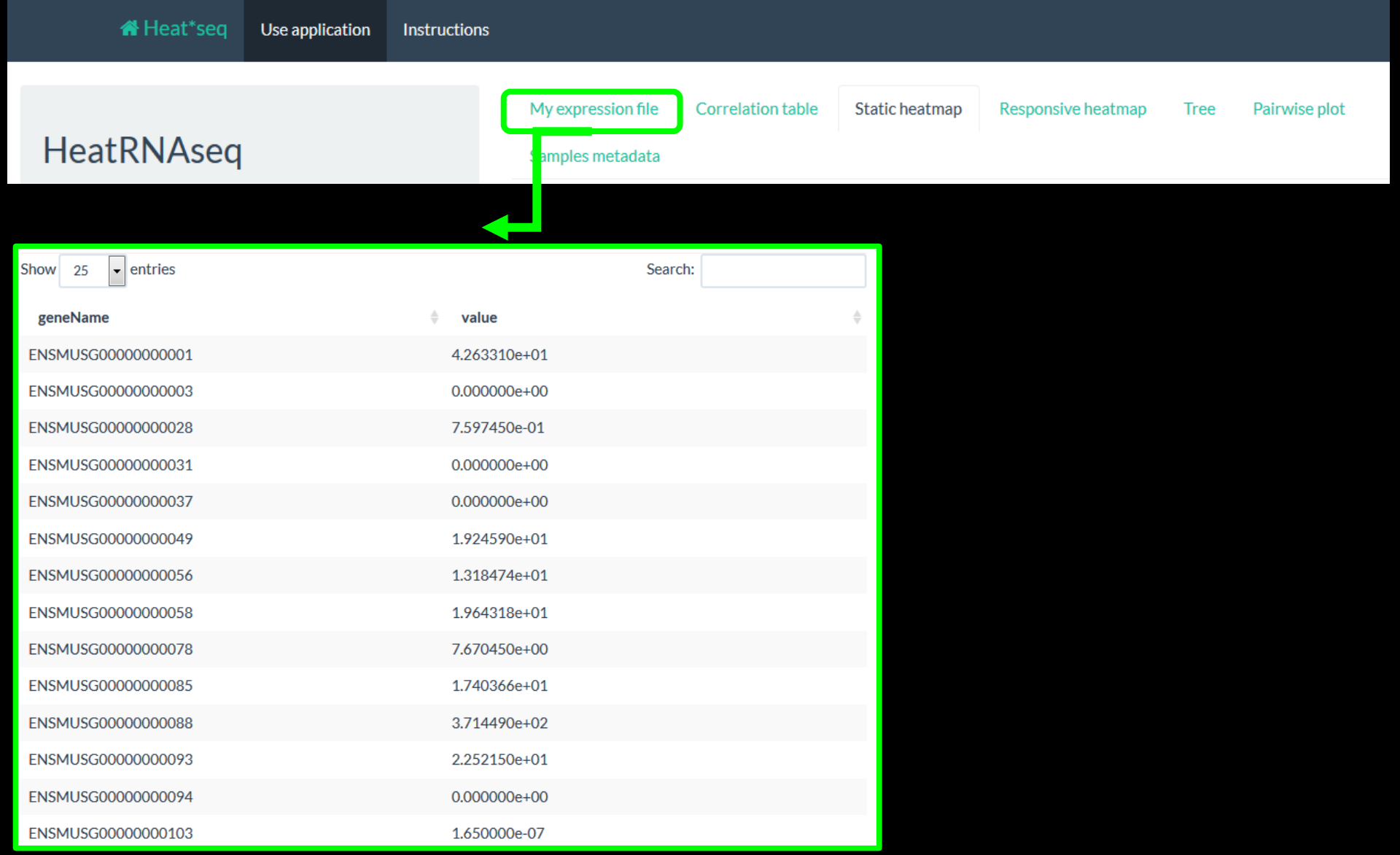
both

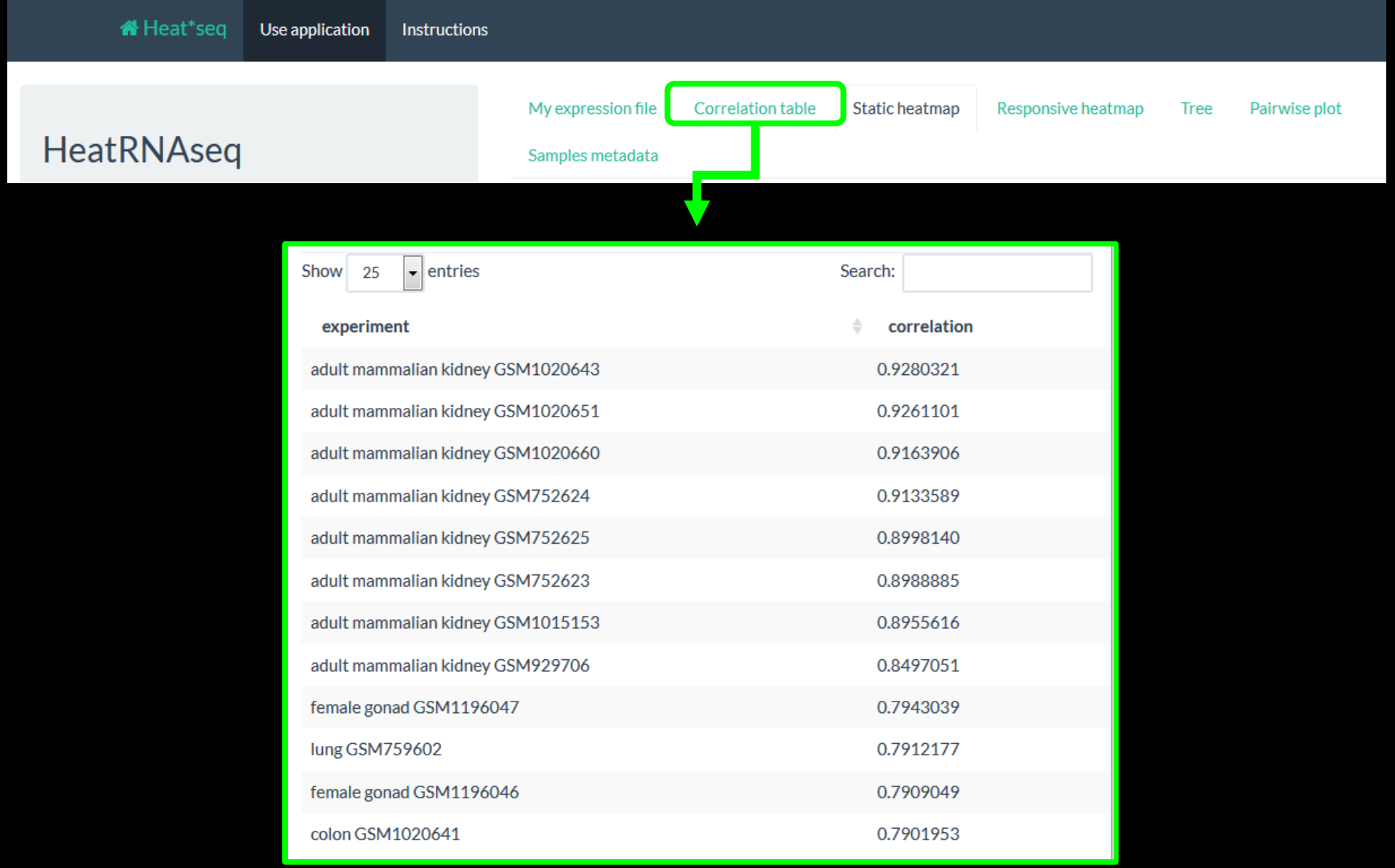
Customise colours





Save as png | Save as pdf | Save as svg | Export data as tab delimited .txt


Heat*seq
Use application
Instructions
HeatRNAseq
My expression file
Correlation table
Static heatmap
Responsive heatmap
Tree
Pairwise plot
Samples metadata
Show 25 entries
Search:
geneName
value
ENSMUSG00000000001 4.263310e+01
ENSMUSG00000000003 0.000000e+00
ENSMUSG00000000028 7.597450e-01
ENSMUSG00000000031 0.000000e+00
ENSMUSG00000000037 0.000000e+00
ENSMUSG00000000049 1.924590e+01
ENSMUSG00000000056 1.318474e+01
ENSMUSG00000000058 1.964318e+01
ENSMUSG00000000078 7.670450e+00
ENSMUSG00000000085 1.740366e+01
ENSMUSG00000000088 3.714490e+02
ENSMUSG00000000093 2.252150e+01
ENSMUSG00000000094 0.000000e+00
ENSMUSG00000000103 1.650000e-07


Heat*seq
Use application
Instructions
HeatRNAseq
My expression file
Correlation table
Static heatmap
Responsive heatmap
Tree
Pairwise plot
Samples metadata
Show 25 entries
Search:
experiment
correlation
adult mammalian kidney GSM1020643 0.9280321
adult mammalian kidney GSM1020651 0.9261101
adult mammalian kidney GSM1020660 0.9163906
adult mammalian kidney GSM752624 0.9133589
adult mammalian kidney GSM752625 0.8998140
adult mammalian kidney GSM752623 0.8988885
adult mammalian kidney GSM1015153 0.8955616
adult mammalian kidney GSM929706 0.8497051
female gonad GSM1196047 0.7943039
lung GSM759602 0.7912177
female gonad GSM1196046 0.7909049
colon GSM1020641 0.7901953

Mouse hovering widget

Zoom by drag and drop

Send plot to plotly and share with the world

plotly

HeatRNAseq

Show 25 entries Search:

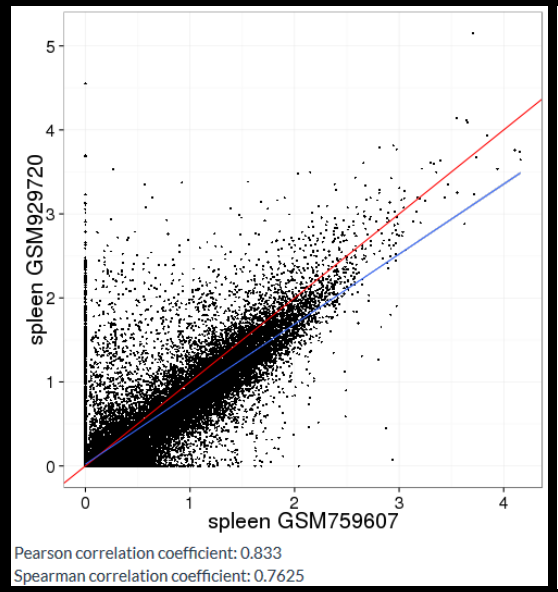
geoAccession	tissue	name	stage	libraryType	url
GSM752614	brain	brain GSM752614	post-juvenile adult stage	single	link
GSM752615	brain	brain GSM752615	post-juvenile adult stage	single	link
GSM752616	brain	brain GSM752616	post-juvenile adult stage	single	link
GSM752617	cerebellum	cerebellum GSM752617	post-juvenile adult stage	single	link
GSM752618	cerebellum	cerebellum GSM752618	post-juvenile adult stage	single	link
GSM752619	cerebellum	cerebellum GSM752619	post-juvenile adult stage	single	link
GSM752620	heart	heart GSM752620	post-juvenile adult stage	single	link
GSM752621	heart	heart GSM752621	post-juvenile adult stage	single	link
GSM752622	heart	heart GSM752622	post-juvenile adult stage	single	link

Heat*seq | Use application | Instructions

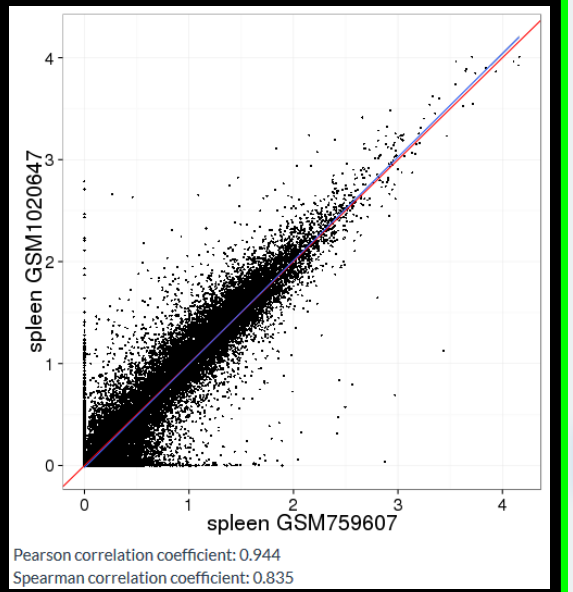
HeatRNAseq

My expression file | Correlation table | Static heatmap | Responsive heatmap | Tree | Pairwise plot

Samples metadata



Pearson correlation coefficient: 0.833
Spearman correlation coefficient: 0.7625



Pearson correlation coefficient: 0.944
Spearman correlation coefficient: 0.835

3 - Plot customization

Experiment 1:

adult mammalian kidney GSM1015153

Experiment 2:

adult mammalian kidney GSM1020643

Plot type:

XY

Data scaling:

log10(e + 1)

☑ Add regression line (blue)

☑ Add guide line (red)

plotly

Conclusions

- A lightweight web application
- Quick comparison of RNA-seq, ChIP-seq and CAGE experiments
- First step to guide you before in depth analysis

What's next

- Other datasets (suggestions?)
- Other data types (Histones, DNAse1, Gene Lists/Ontologies)
- Multiple user files
- Gene name converter / *liftover* integration

Functionnal annotation transfert using co-expression networks

Pía Francesca Loren Reyes, Tom Michoel, Anagha Joshi, Guillaume Devailly

INRA SCIENCE & IMPACT

@G_Devailly

.048

2017-12-19

- Plenty of *omic* data
- Not enough functional annotations :(
- Can we use *omic* data to improve functional annotations?

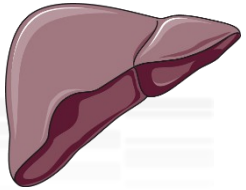
- Functional annotations are often inferred from other species

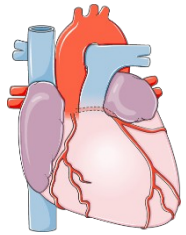
- This lead to over-annotations mistakes, notably in the cases of one-to-many or many-to-many homology groups

Hypothesis: if two orthologs are in the **same co-expression cluster** in different species, then functional annotation are likely to be transferable.

And co-expression cluster are easy to build from public *omic* data! \o/

Public microarray expression data





Mous

Series ID	n
total	*920*
GSE50789	96
GSE9630	59
GSE55756	47
GSE63027	39
GSE51885	27
GSE38067	24
…	…

Rat

Series ID	n
total	*620*
GSE13270	101
GSE59495	90
GSE24104	47
GSE5509	40
GSE23748	30
GSE27625	30
…	…

Mous

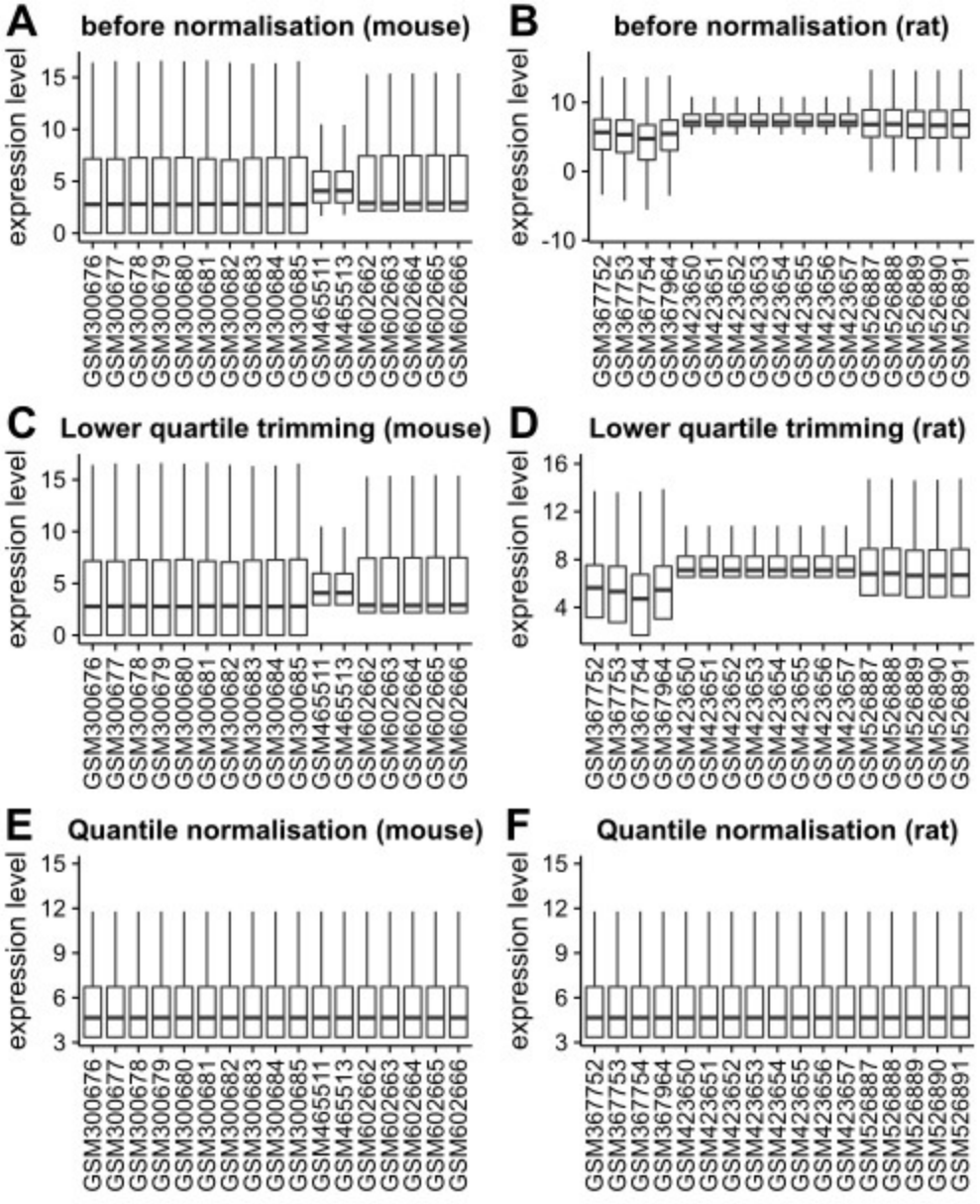
Series ID	n
total	*248*
GSE1479	36
GSE3530	36
GSE7487	24
GSE5500	21
GSE7605	18
GSE3440	15
…	…

Rat

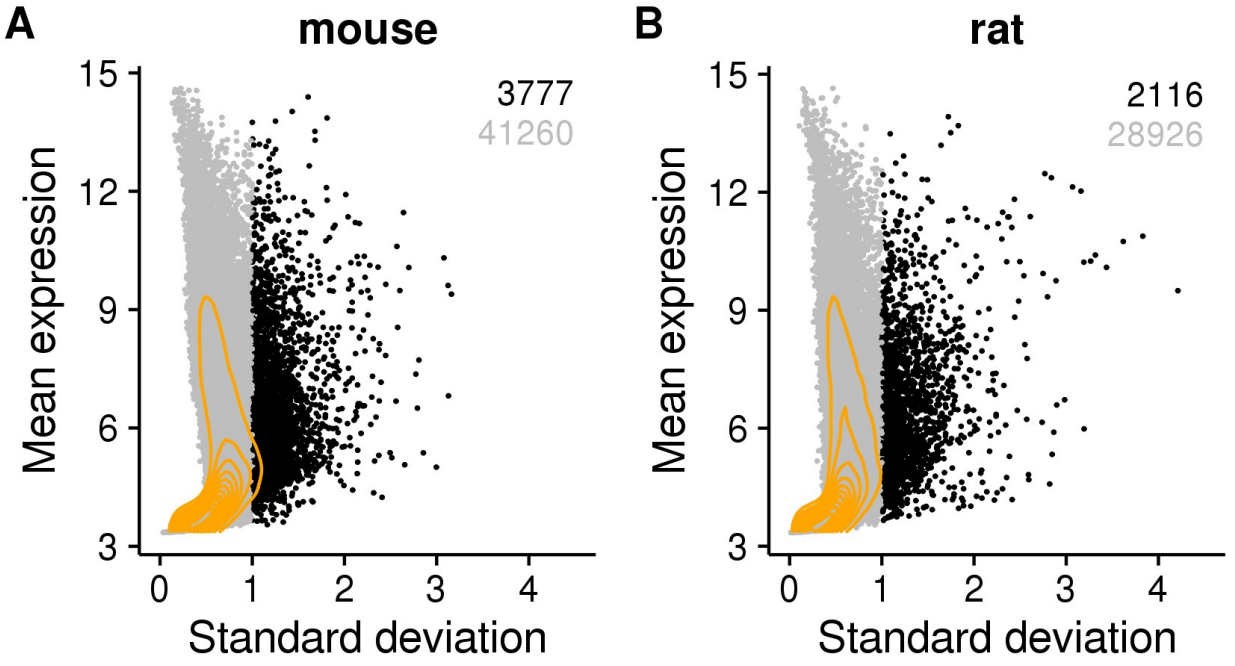
Series ID	n
total	*1202*
GSE57822	433
GSE57800	429
GSE19290	82
GSE6104	45
GSE7999	30
GSE11851	20
…	…

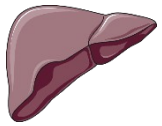
Affymetrix Mouse Genome 430 2.0
Affymetrix Rat Genome 230 2.0

Data normalization

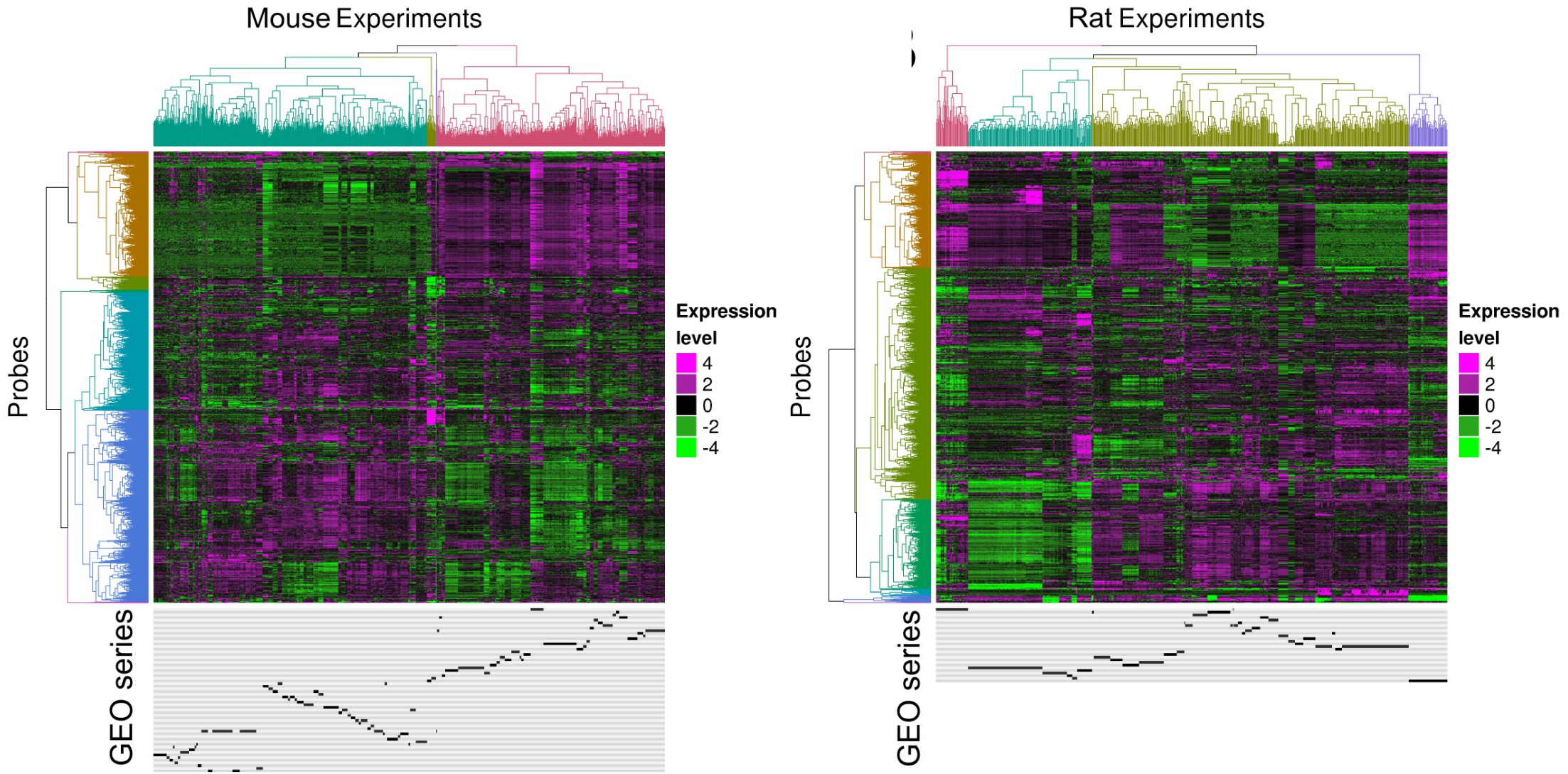


Identification of variable probes

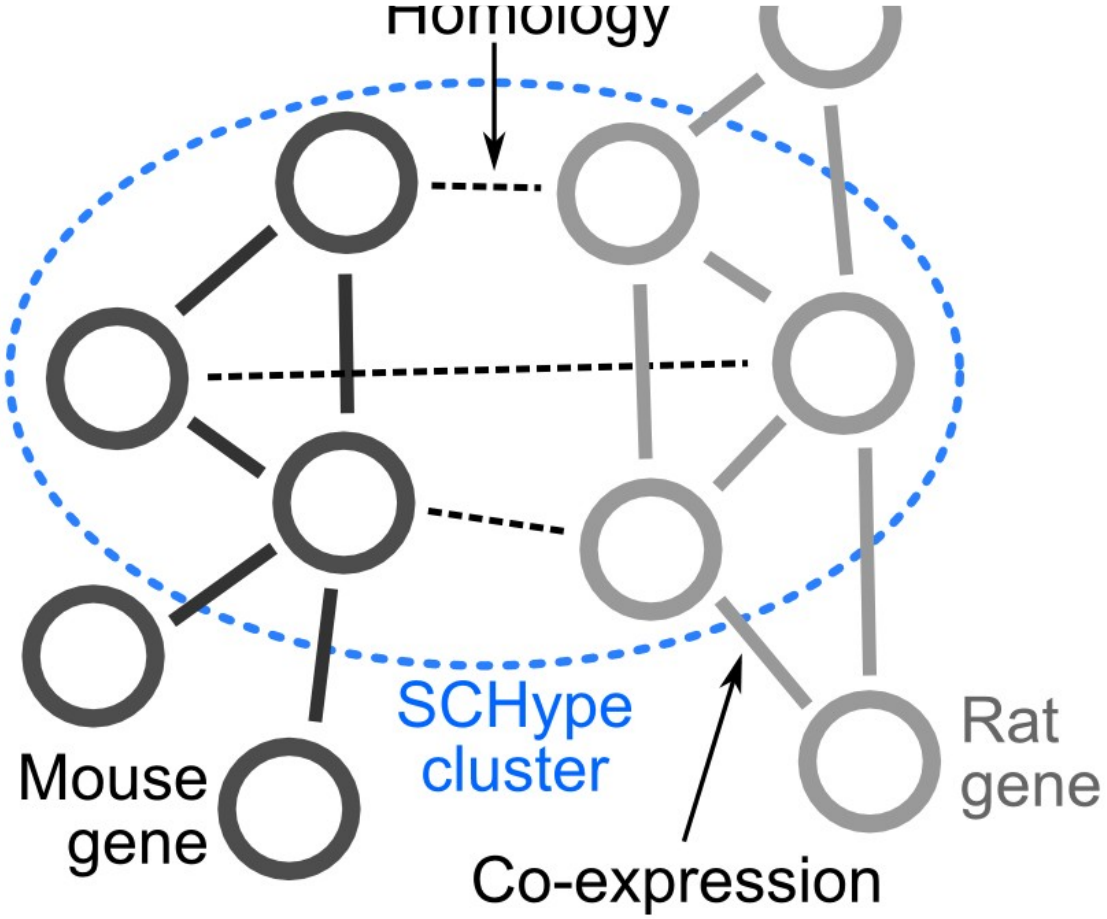


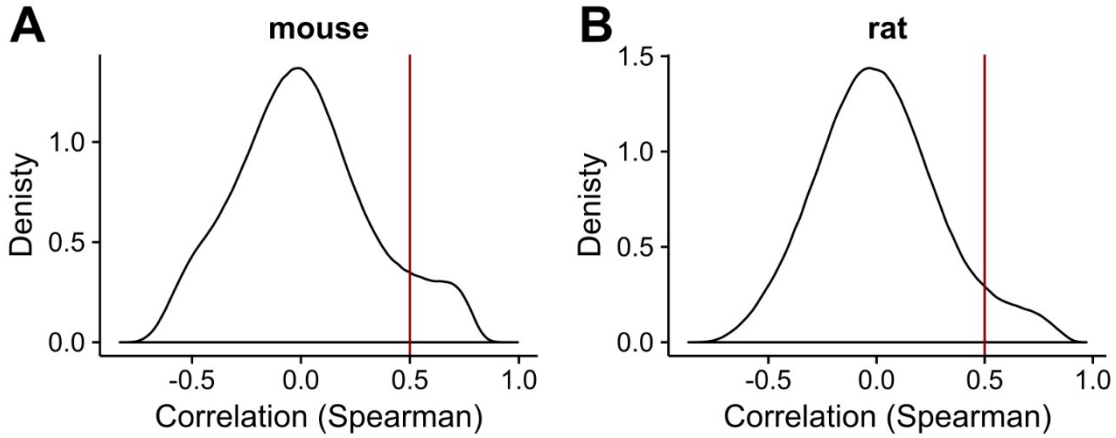


Species	Category	Term	Gene	FE	P-value
Mouse	Reactome	Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE)	11	4.78	4.29E-02
		Activation of gene expression by SREBF (SREBP)	15	4.34	5.18E-03
		Regulation of cholesterol biosynthesis by SREBP (SREBF)	17	3.94	4.36E-03
		Cytochrome P450 - arranged by substrate type	27	2.72	7.78E-03
		Phase 1 - Functionalization of compounds	37	2.55	7.58E-04
	GO slim BP	fatty acid metabolic process	52	2.26	2.95E-05
		steroid metabolic process	50	2.18	1.31E-04
Rat	Reactome	Synthesis of bile acids and bile salts via 24-hydroxycholesterol	7	8.63	2.95E-02
		Endosomal/Vacuolar pathway	10	7.93	1.15E-03
		Striated Muscle Contraction	11	6.78	1.48E-03
		ER-Phagosome pathway	10	6.53	6.29E-03
		Activation of gene expression by SREBF (SREBP)	10	6.53	6.29E-03
		Antigen Presentation: Folding, assembly and peptide loading of class I MHC	13	6.27	3.84E-04
		Regulation of cholesterol biosynthesis by SREBP (SREBF)	10	5.55	2.51E-02
		Biological oxidations	25	2.95	3.58E-03
		Metabolism of lipids and lipoproteins	68	2.13	1.15E-05
	GO slim BP	response to biotic stimulus	12	4.16	1.12E-02
		fatty acid metabolic process	22	2.52	2.52E-02


Mouse Experiments
Probes
GEO series
Expression level
4
2
0
-2
-4
Rat Experiments
Probes
GEO series
Expression level
4
2
0
-2
-4

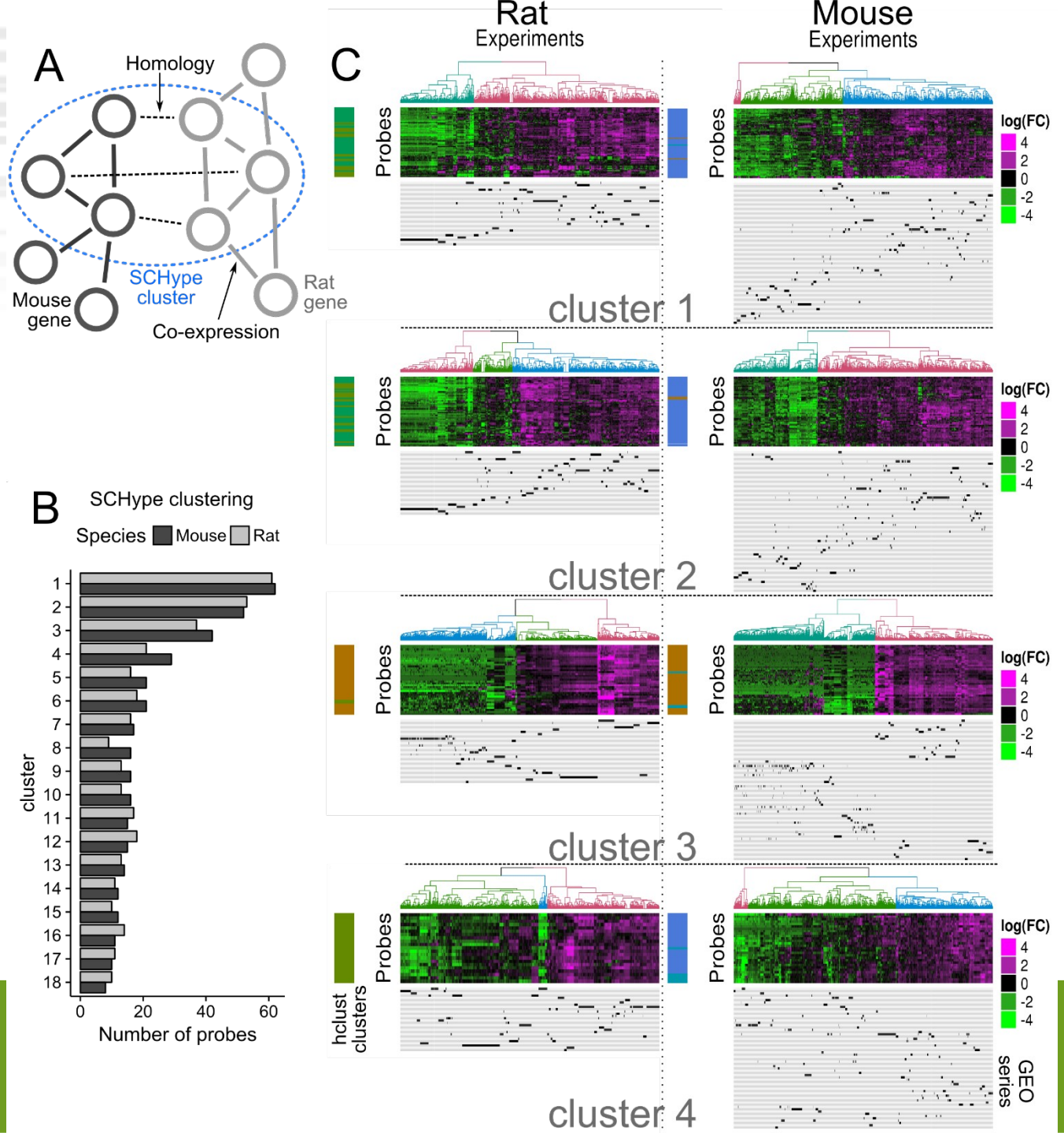
SCHype hyper-graph clustering tool





NCBI

HomoloGene


A
Homology
SCHype cluster
Mouse gene
Rat gene
Co-expression
B
SCHype clustering
Species Mouse Rat
cluster
Number of probes
C
Rat Experiments
Mouse Experiments
Probes
log(FC)
cluster 1
cluster 2
cluster 3
cluster 4
hclust clusters
GEO series

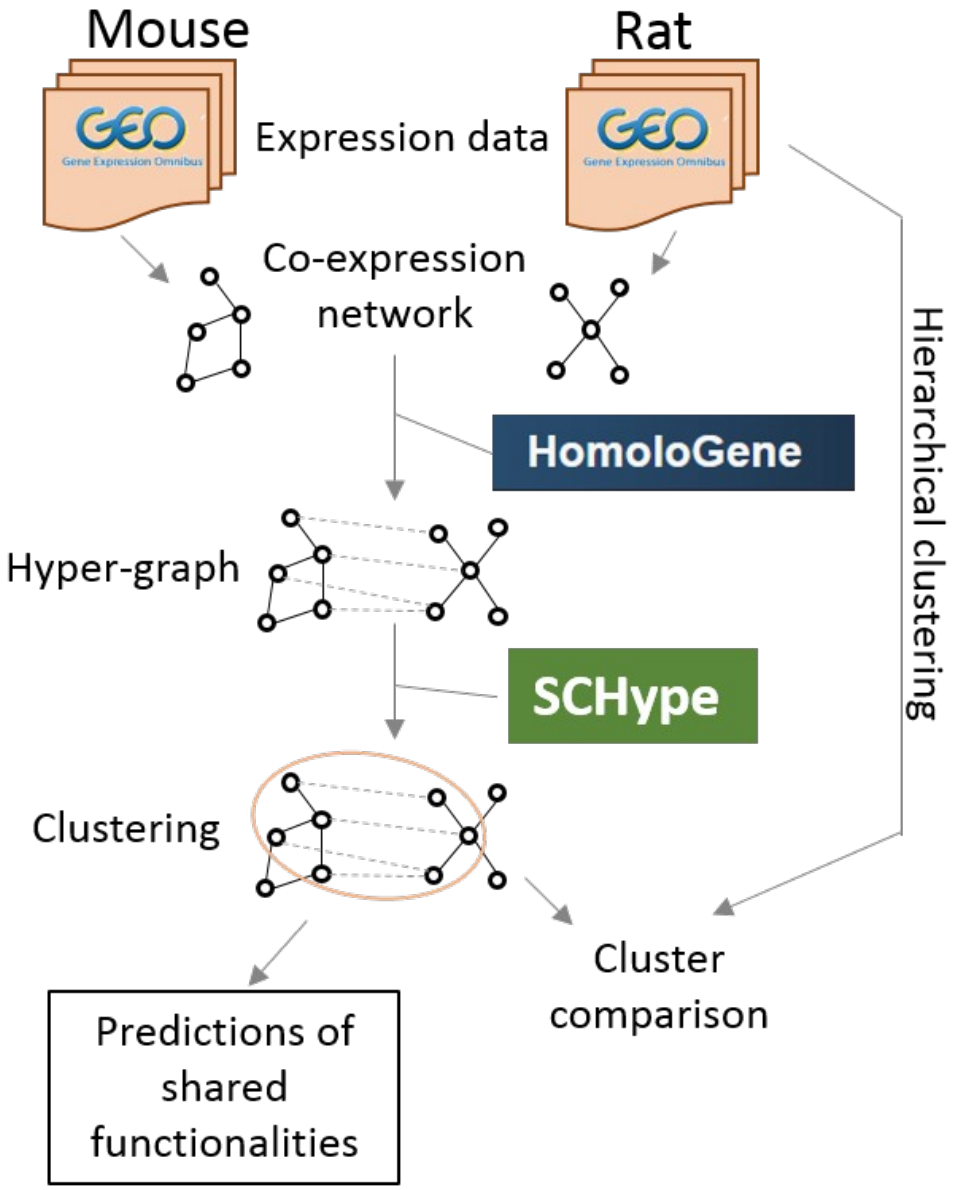
HomoloGene

Homology group	Species	Gene name	SCHype cluster	
137229			cluster 69	
	mouse	*Anp32a*	✓	
	rat	*Anp32a*	✓	
	rat	*LOC100909983*		
68982			cluster 7	cluster 30
	mouse	*Ccnb1*	✓	✓
	mouse	*Gm5593*		
	rat	*Ccnb1*	✓	✓
10699			cluster 2	cluster 118
	mouse	*Cd248*	✓	✓
	rat	*Cd248*	✓	✓
	rat	*LOC100911932*		
	rat	*LOC100911882*		
3938			cluster 1	
	mouse	*Ppp1r3c*	✓	
	rat	*Ppp1r3c*	✓	
	rat	*LOC100910671*		
14108			cluster 2	
	mouse	*Rasl10b*	✓	
	rat	*Rasl10b*	✓	
	rat	*LOC100912246*		

INRA SCIENCE & IMPACT

Homology group	Species	Gene name	SCHype cluster		
128630			cluster 9	cluster 12	cluster 45
	mouse	*Ceacam1*	✓		
	mouse	*Ceacam2*	✓	✓	✓
	rat	*Ceacam1*	✓	✓	✓
11456			cluster 5		
	mouse	*Elovl6*	✓		
	rat	*Elovl6*	✓		
	rat	*LOC102549542*	✓		
20277			cluster 35		
	mouse	*Rrm2*	✓		
	rat	*Rrm2*	✓		
	rat	*LOC100359539*	✓		
55991			cluster 1	cluster 119	
	mouse	*Tmed2*	✓	✓	
	mouse	*Gm21540*	✓	✓	
	rat	*Tmed2*	✓	✓	
11890			cluster 10	cluster 43	cluster 81
	mouse	*Tnks2*	✓	✓	✓
	rat	*LOC100910717*	✓	✓	✓
	rat	*Tnks2*	✓	✓	✓

Homology group	Species	Gene name	SCHype cluster
117948			cluster 102
	mouse	*Cyp2c38*	✓
	mouse	*Cyp2c29*	
	mouse	*Cyp2c39*	
	rat	*Cyp2c7*	✓
104115			cluster 33
	mouse	*Hsd3b5*	✓
	mouse	*Gm10681*	
	mouse	*Hsd3b4*	
	mouse	*Gm4450*	
	rat	*Hsd3b5*	✓
	rat	*LOC100911116*	✓
137425			cluster 2
	mouse	*Lce3c*	✓
	rat	*LOC100361951*	✓
	rat	*LOC100911982*	✓
	rat	*Lce3d*	
129514			cluster 17
	mouse	*Rdh9*	✓
	mouse	*Rdh1*	
	mouse	*Rdh16*	
	mouse	*Rdh19*	
	mouse	*BC089597*	
	rat	*Rdh16*	✓
	rat	*LOC100365958*	✓



Conclusions & perspectives

- Provide evidence for resolving functional annotation transfer in 28 complex homology groups (>100 genes)
- Using free data!

- Different array, RNA-seq
- More tissues!
- More species!
- Still need functional annotation in at least one species

Epigenomics and the human transcript diversity

using Roadmap Epigenomics data

@G_Devailly

.061

2017-12-19

- Epigenetic marks role at TSS and enhancer is (quite) well characterized
- Putative roles in TES, exon recognition, cryptic TSS repression?
- Plenty of data generated by Roadmaps Epigenomics (+ENCODE)
- RNA-seq, WGBS, DNAse1, plenty of HisMod in > 30 cell types

What’s a gene?

GENCODE

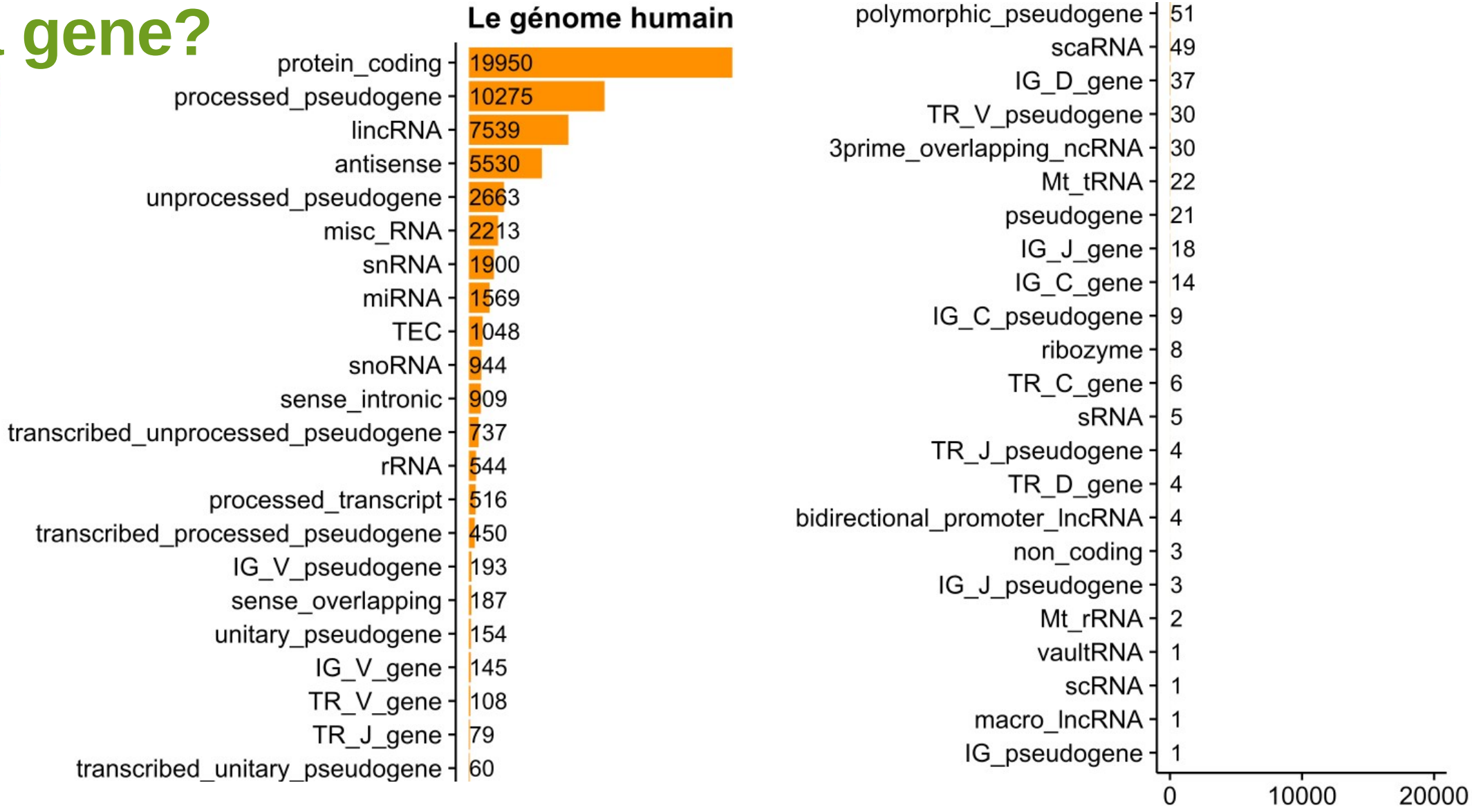
Version 27 (January 2017 freeze, GRCh38) - Ensembl 90

General stats

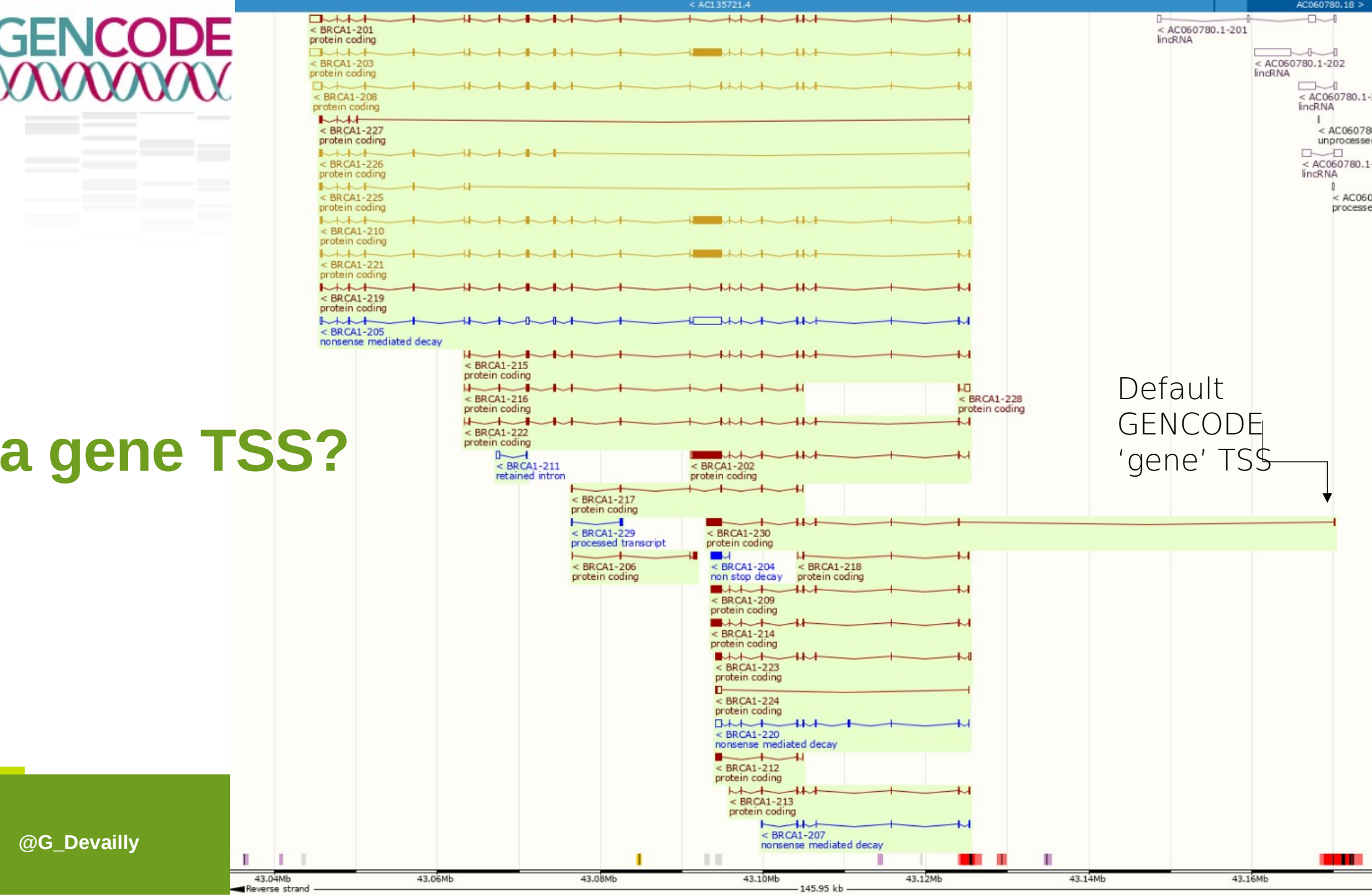
Total No of Genes	58288

What's a gene?

GENCODE

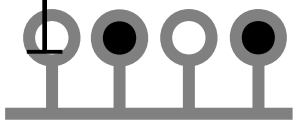
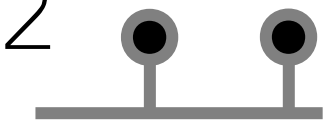


What's a gene TSS?



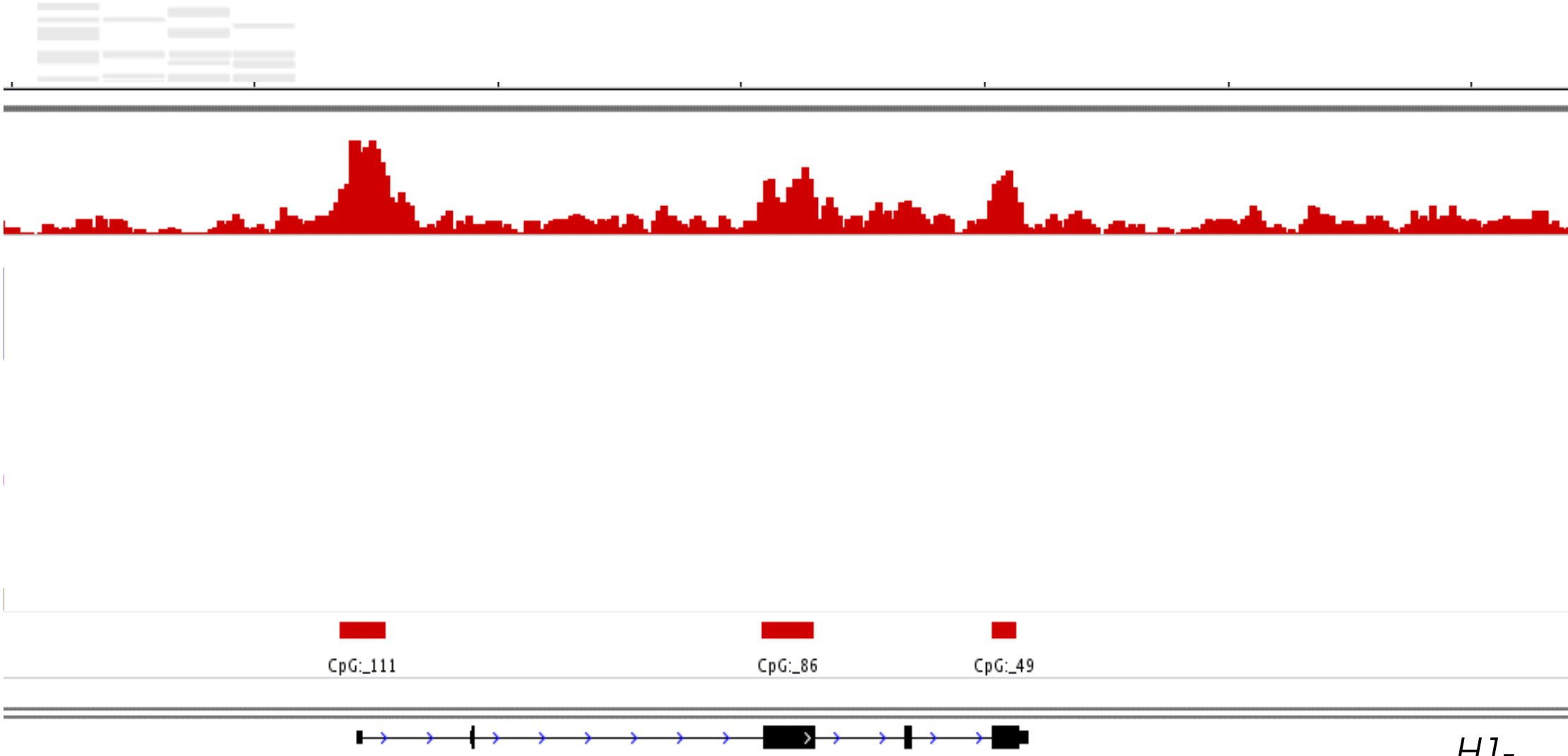
@G_Devailly

Which metric for DNA methylation?

	Region 1	Region 2
		
CpG density	4	2
mCpG/CpG density	0.5	1
mCpG density	2	2

mCpG density main driver of methylation reader?

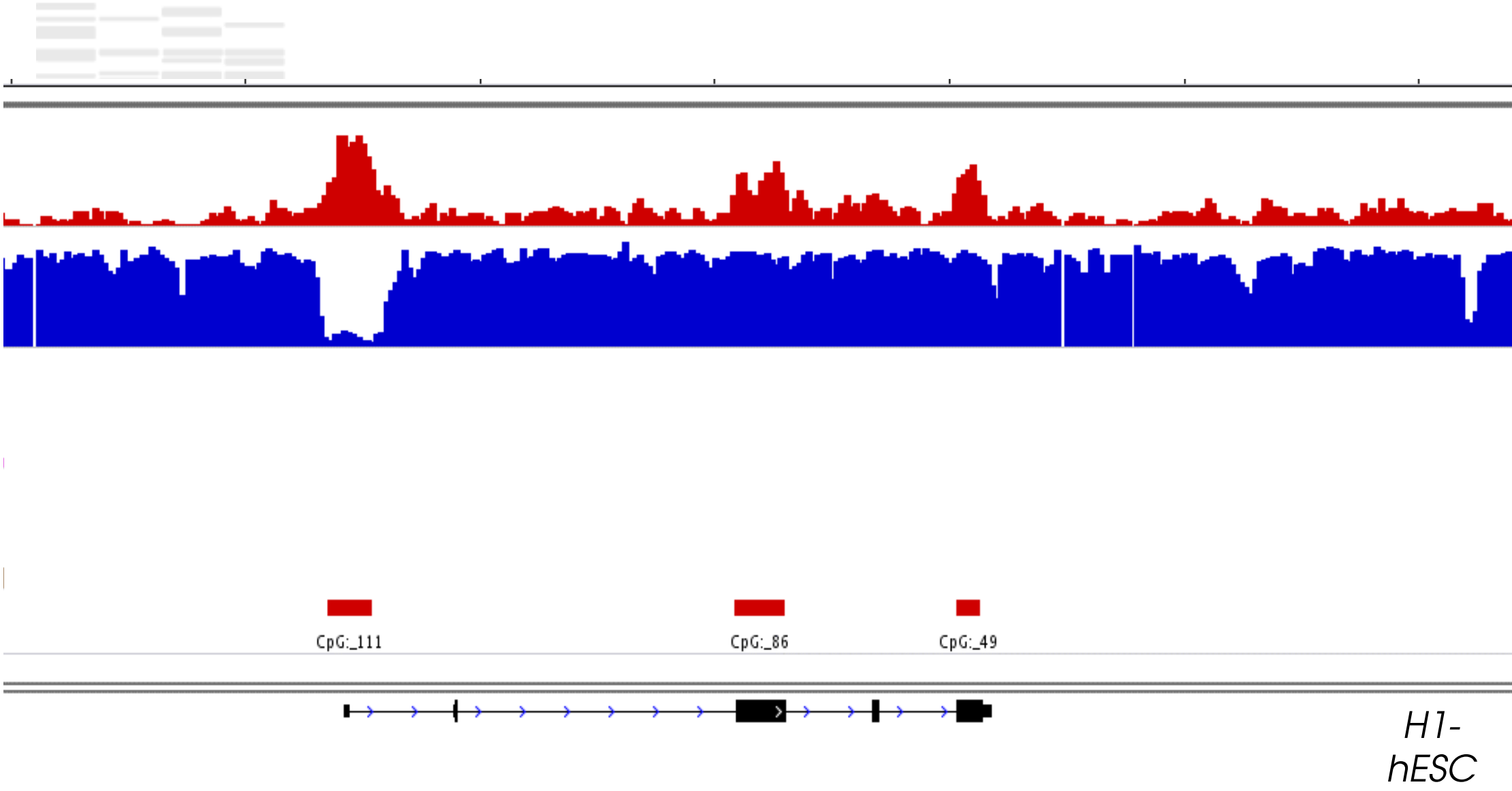
CpG density



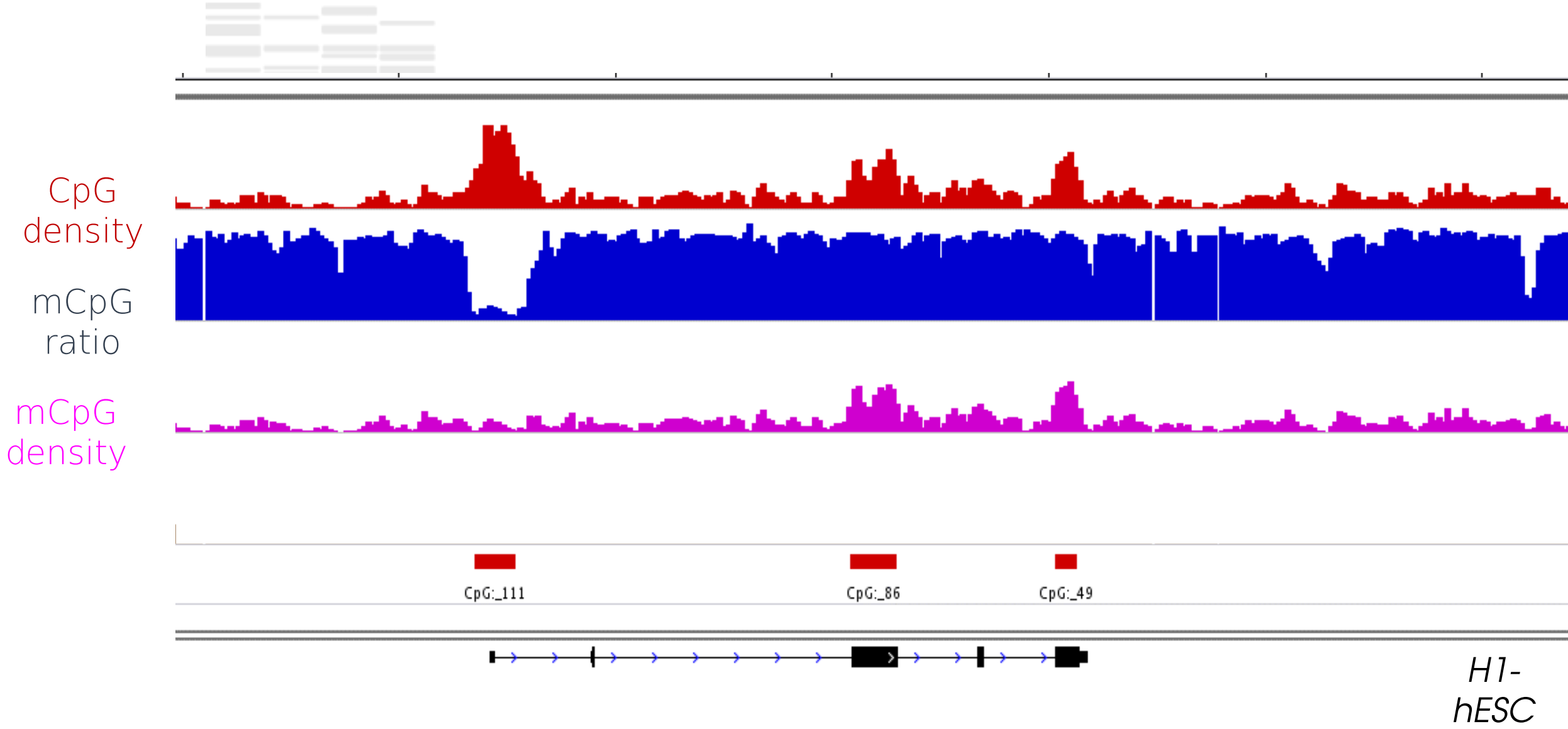
H1-hESC

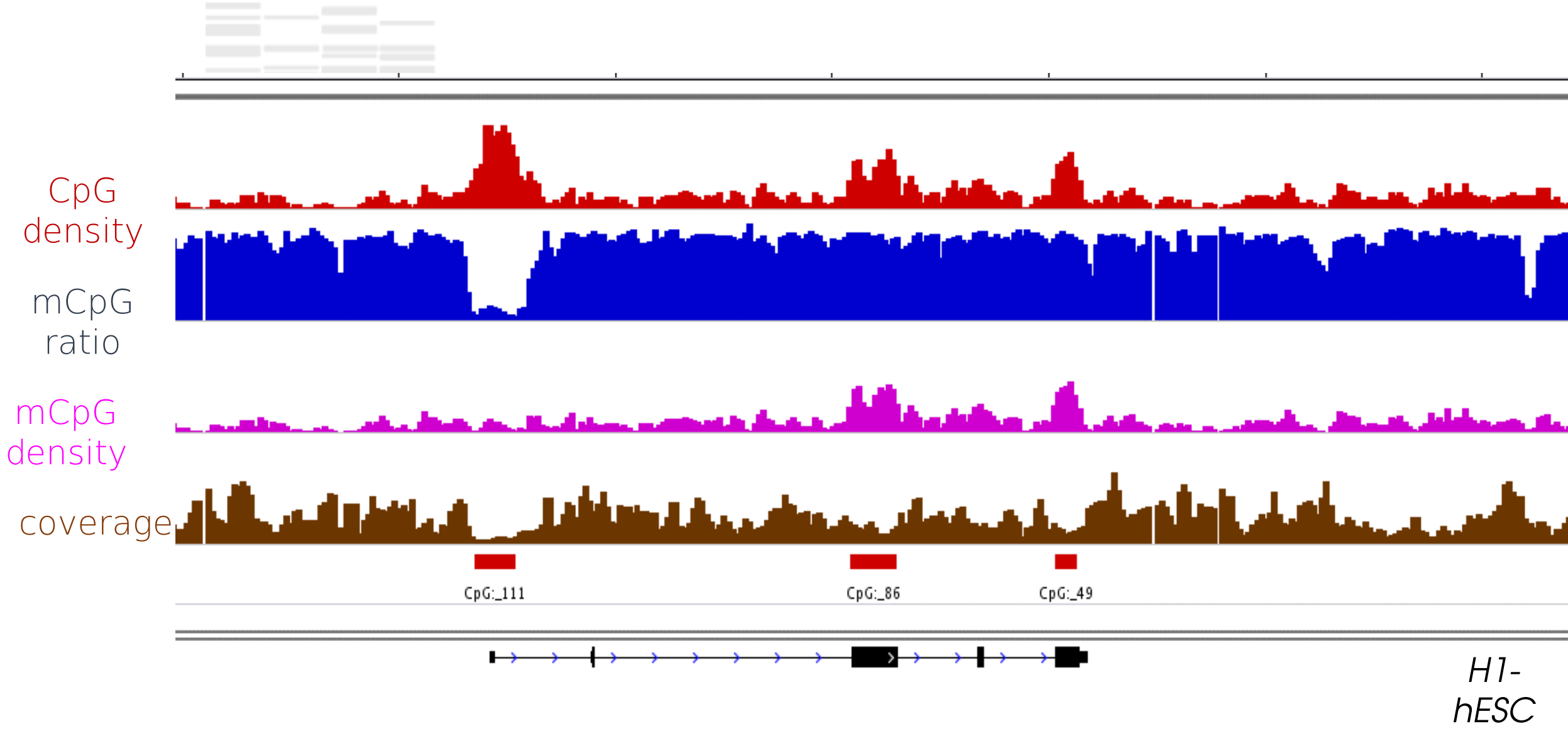
CpG density

mCpG ratio



H1-hESC


CpG density
mCpG ratio
mCpG density
CpG:_111
CpG:_86
CpG:_49
H1-hESC

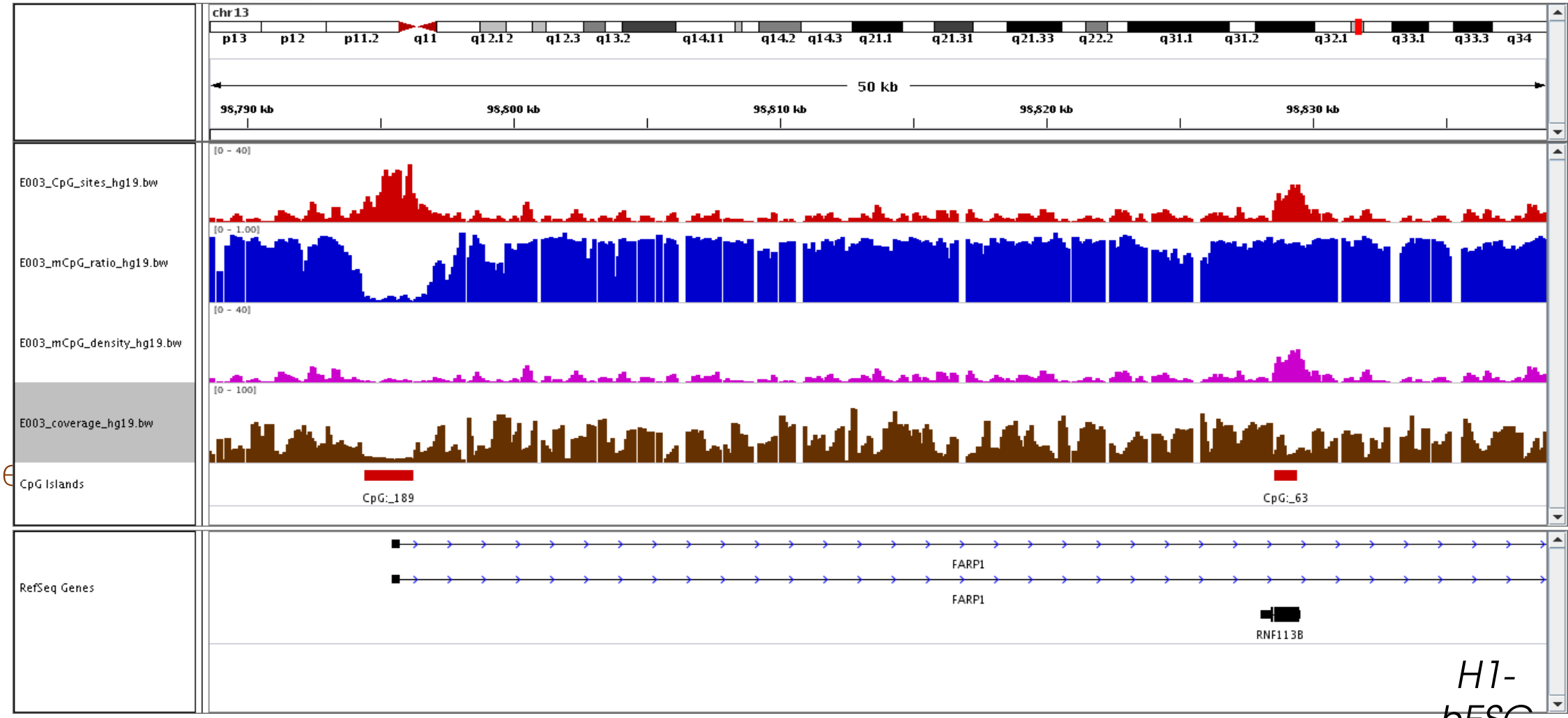

CpG density
mCpG ratio
mCpG density
coverage
CpG:_111
CpG:_86
CpG:_49
H1-hESC

CpG density

mCpG ratio

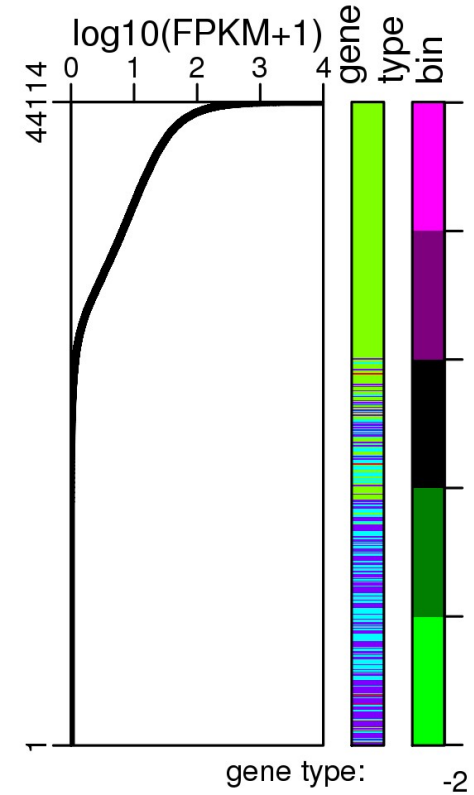
mCpG density

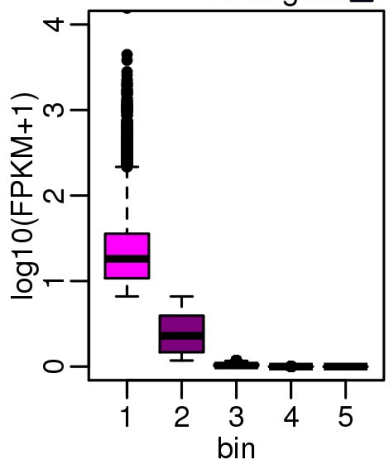
coverage



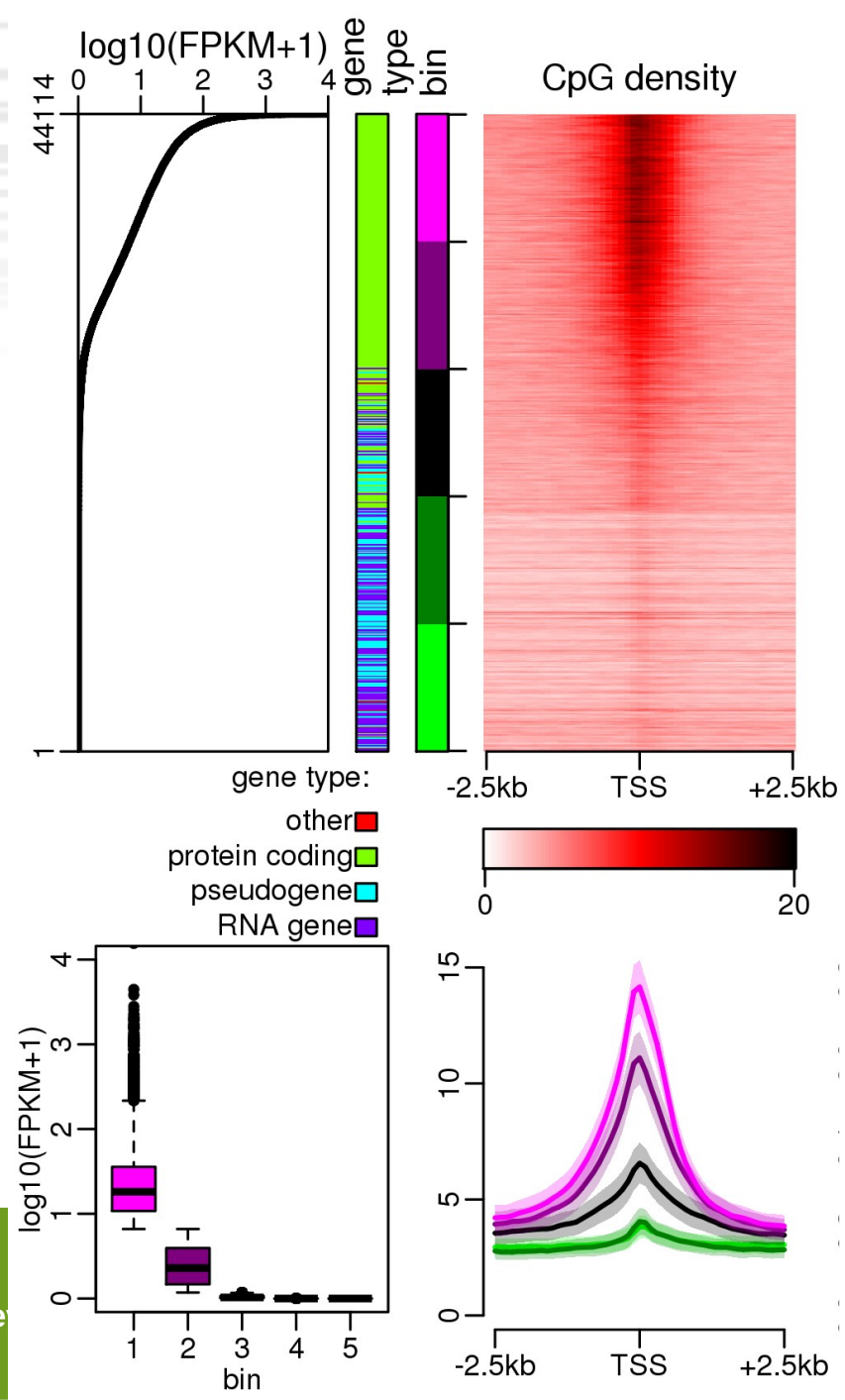
H1-hESC

H1-hESC

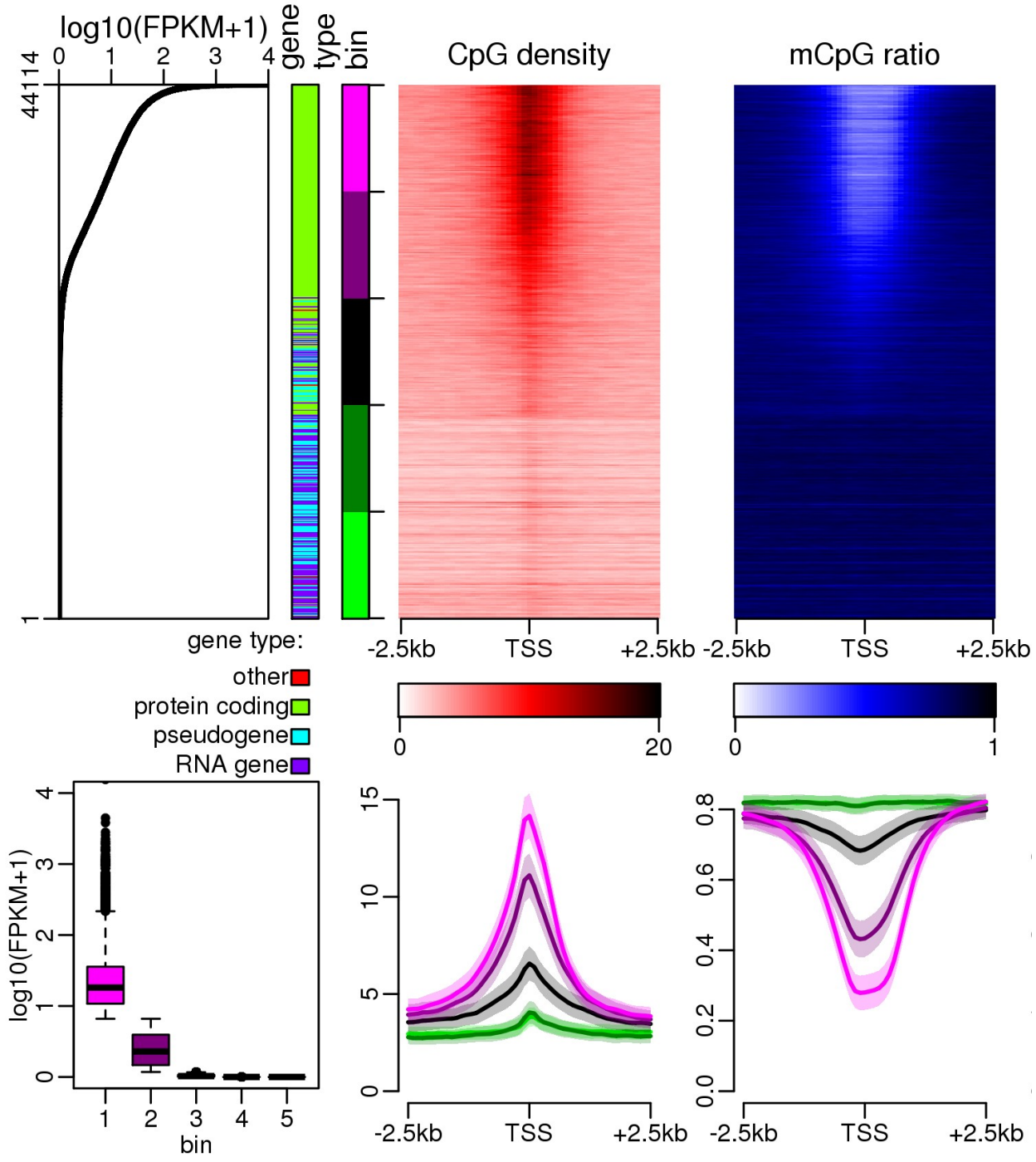




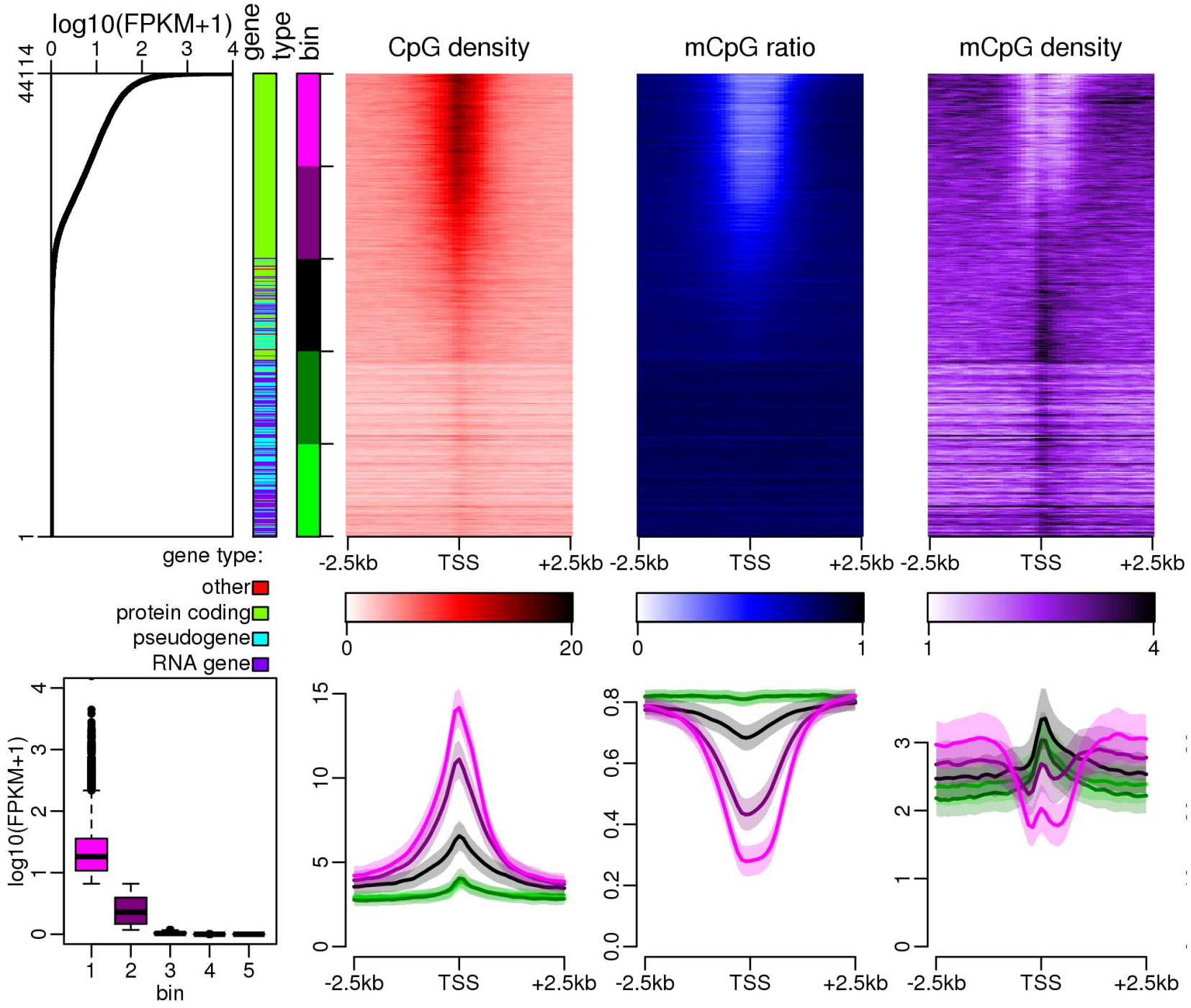
H1-hESC


log10(FPKM+1)
0
1
2
3
4
44114
1
gene type bin
CpG density
-2.5kb
TSS
+2.5kb
gene type:
other
protein coding
pseudogene
RNA gene
0
20
log10(FPKM+1)
0
1
2
3
4
1
2
3
4
5
bin
0
5
10
15
-2.5kb
TSS
+2.5kb

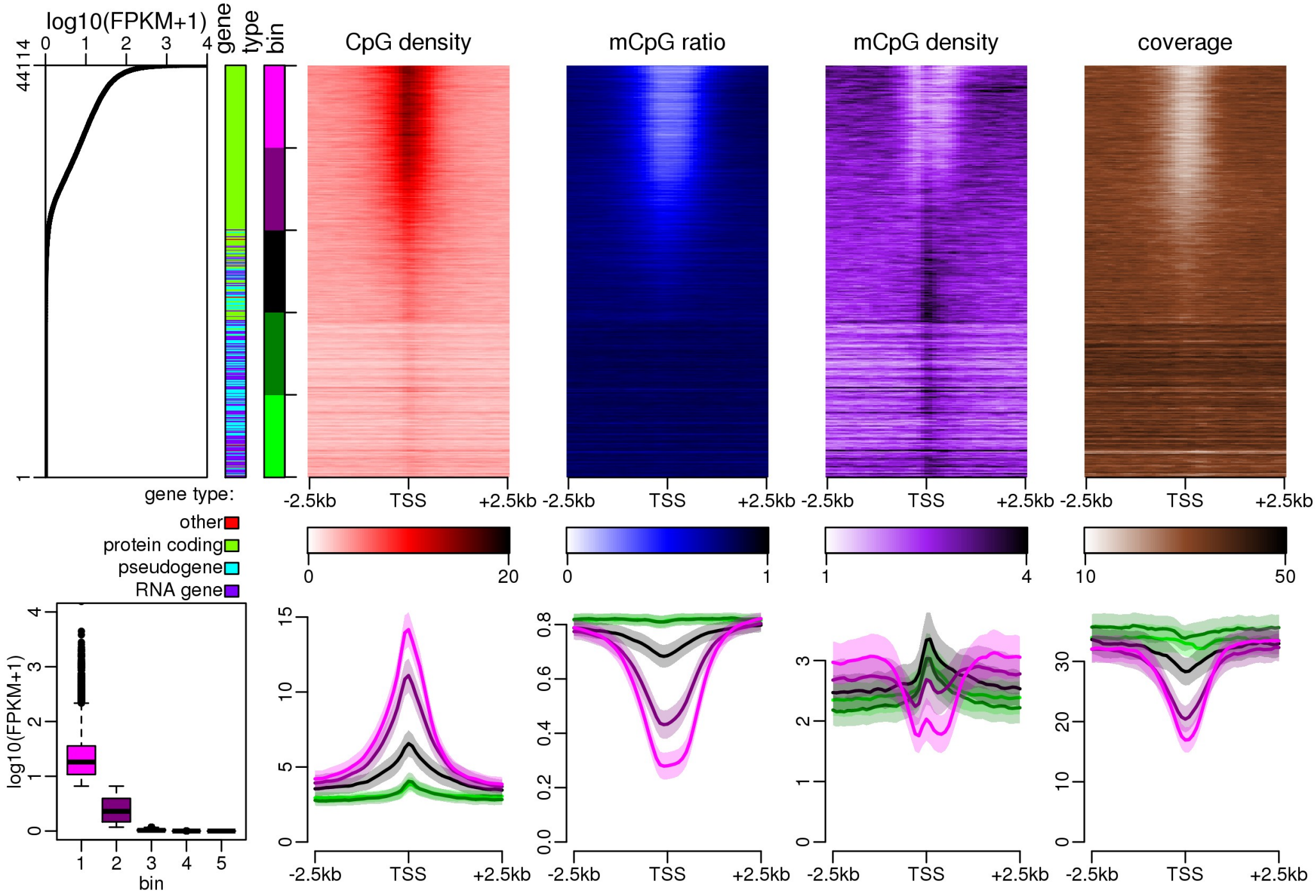
H1-hESC



H1-hESC



H1-hESC



- Features x marks x cell types x gene types
- Combinatorial explosion: >5000

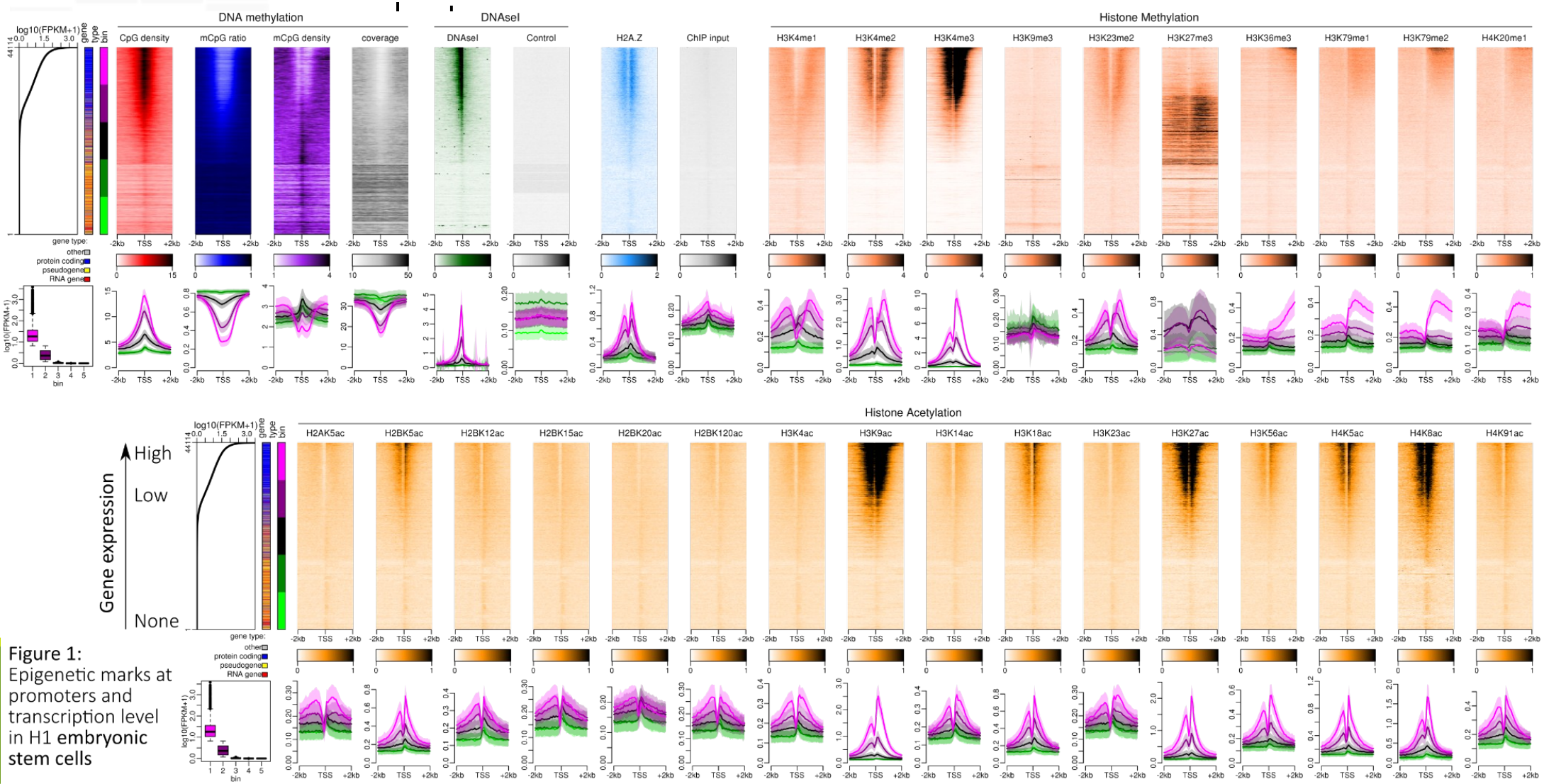
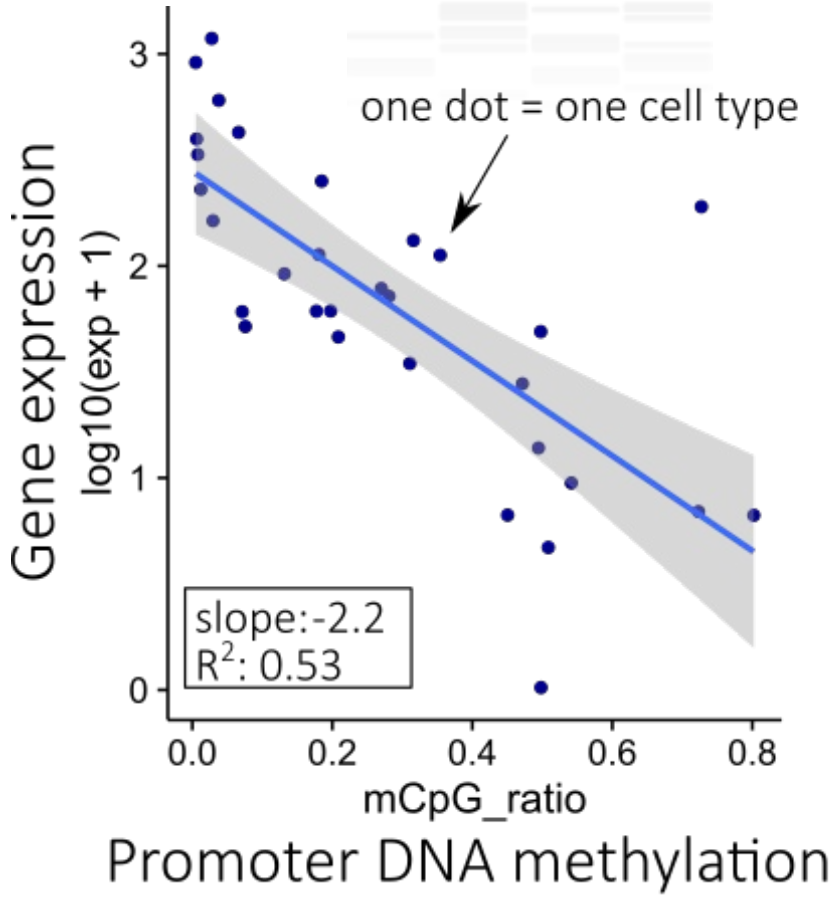
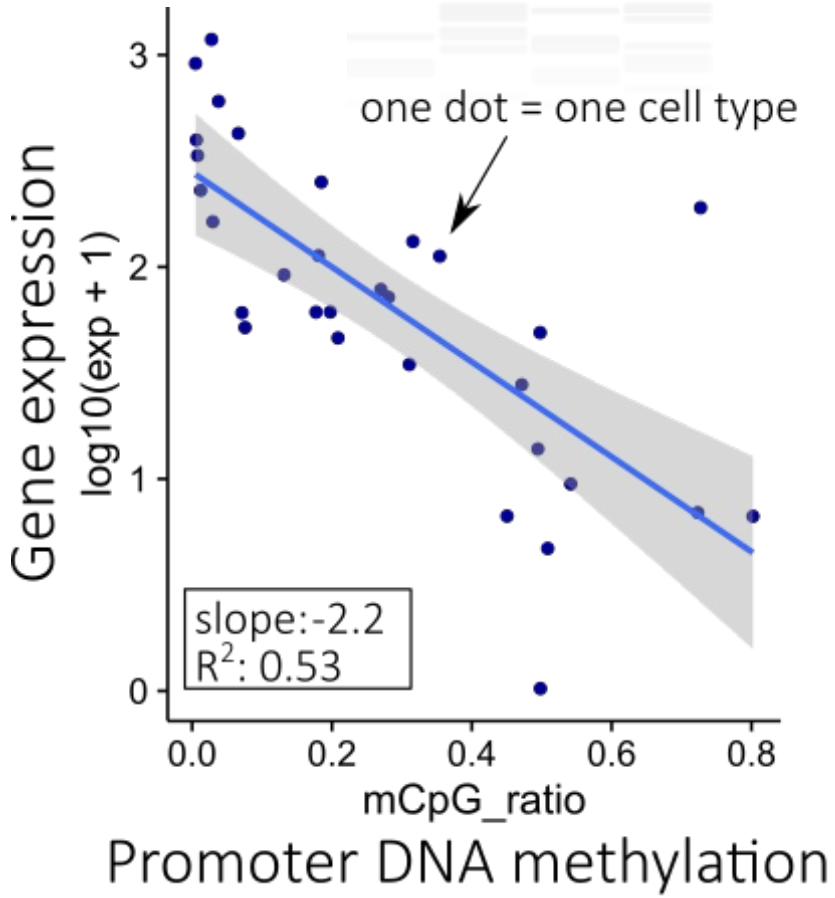


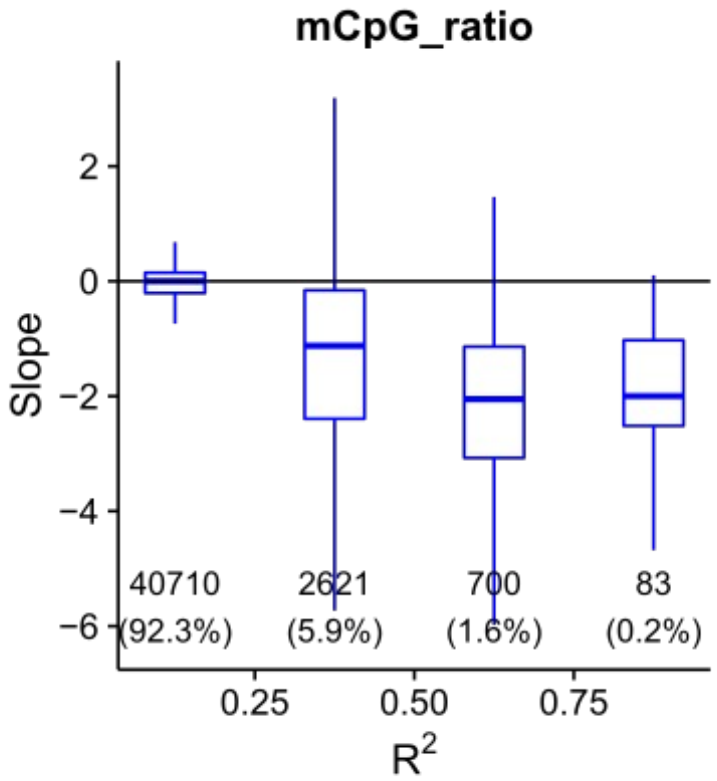
Figure 1:
Epigenetic marks at promoters and transcription level in H1 **embryonic stem cells**

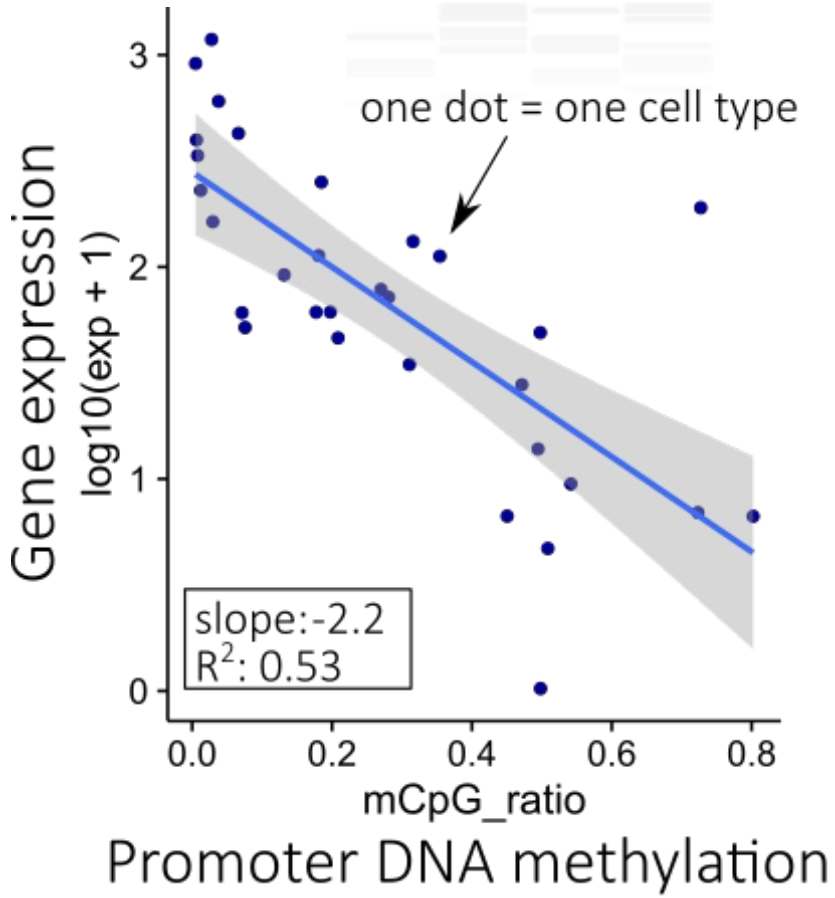
INRA SCIENCE & IMPACT

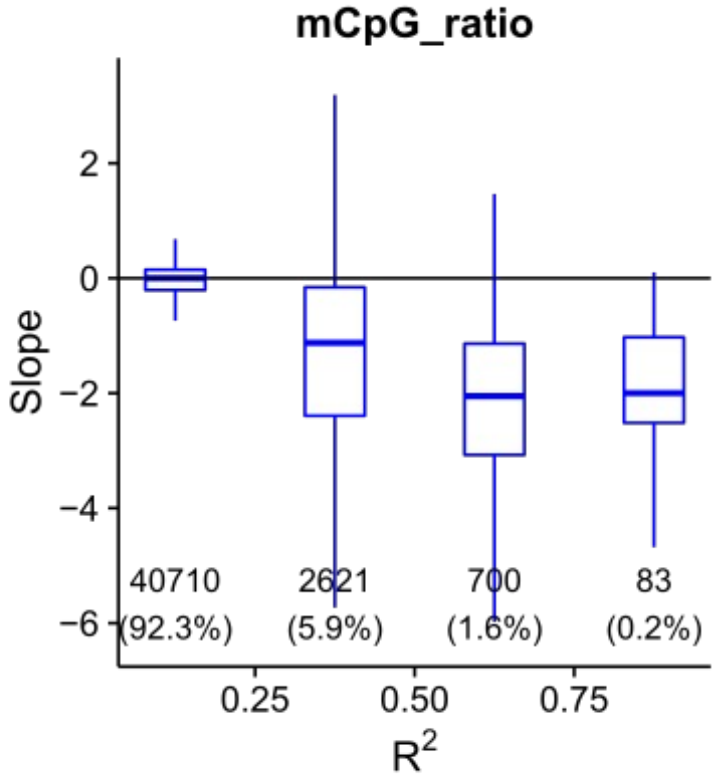
GJA1 (ENSG00000152661)

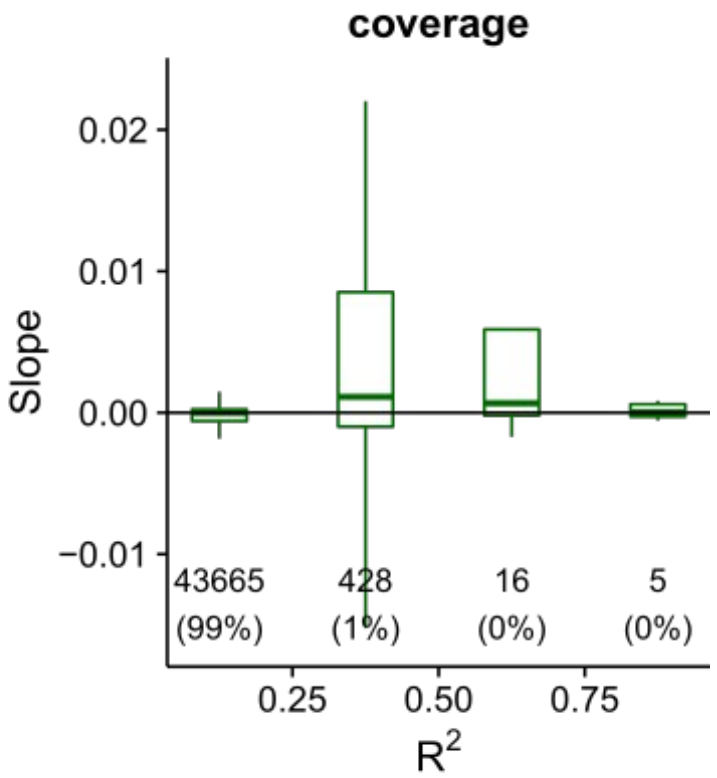


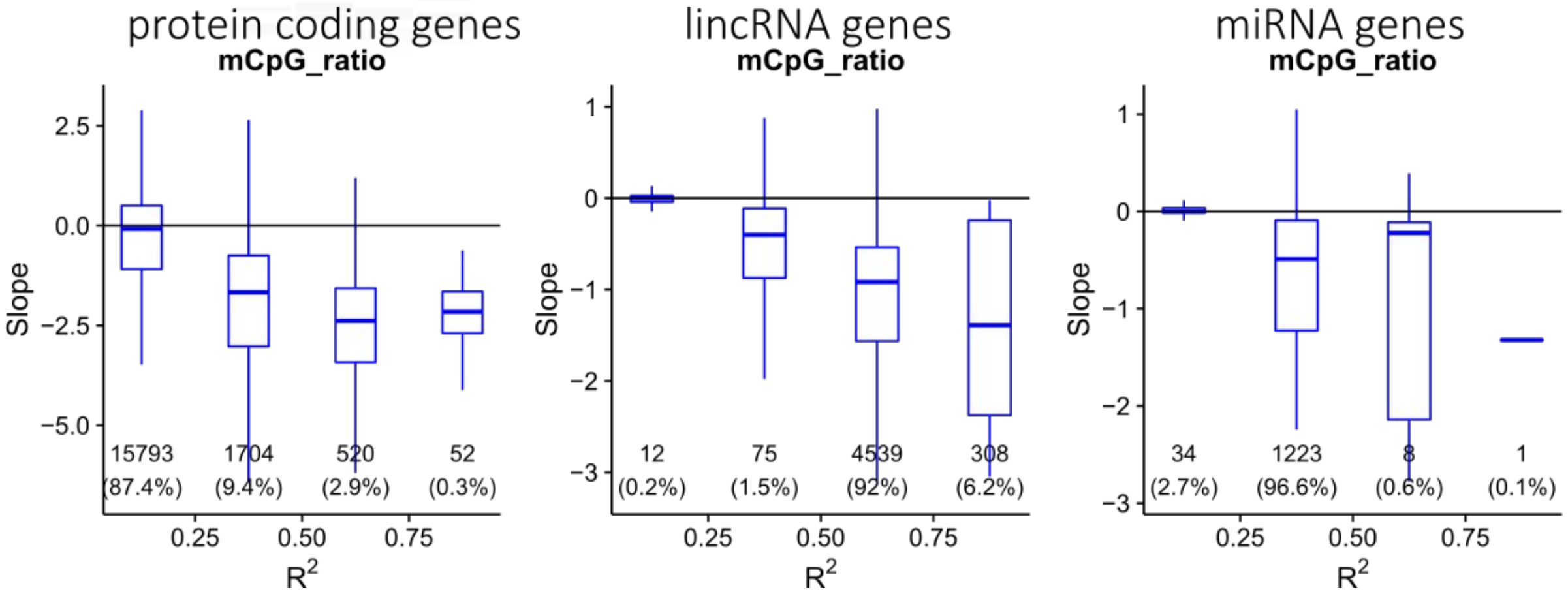

one dot = one cell type
Gene expression
log10(exp + 1)
slope:-2.2
R^2: 0.53
mCpG_ratio
Promoter DNA methylation


mCpG_ratio
Slope
40710
(92.3%)
2621
(5.9%)
700
(1.6%)
83
(0.2%)
R^2

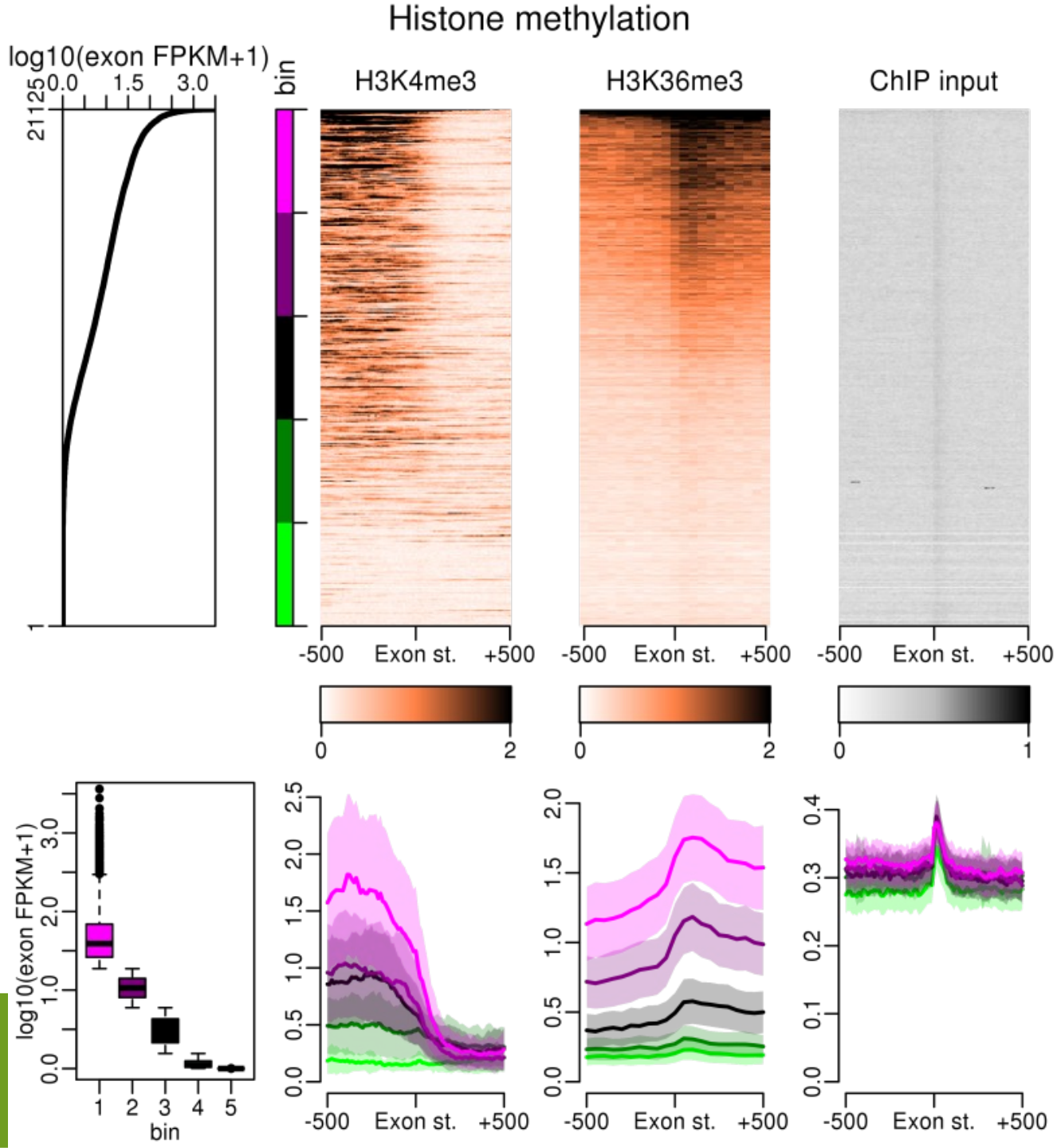

one dot = one cell type
slope:-2.2
R^2: 0.53
Gene expression
log10(exp + 1)
mCpG_ratio
Promoter DNA methylation


mCpG_ratio
Slope
40710
(92.3%)
2621
(5.9%)
700
(1.6%)
83
(0.2%)
R^2

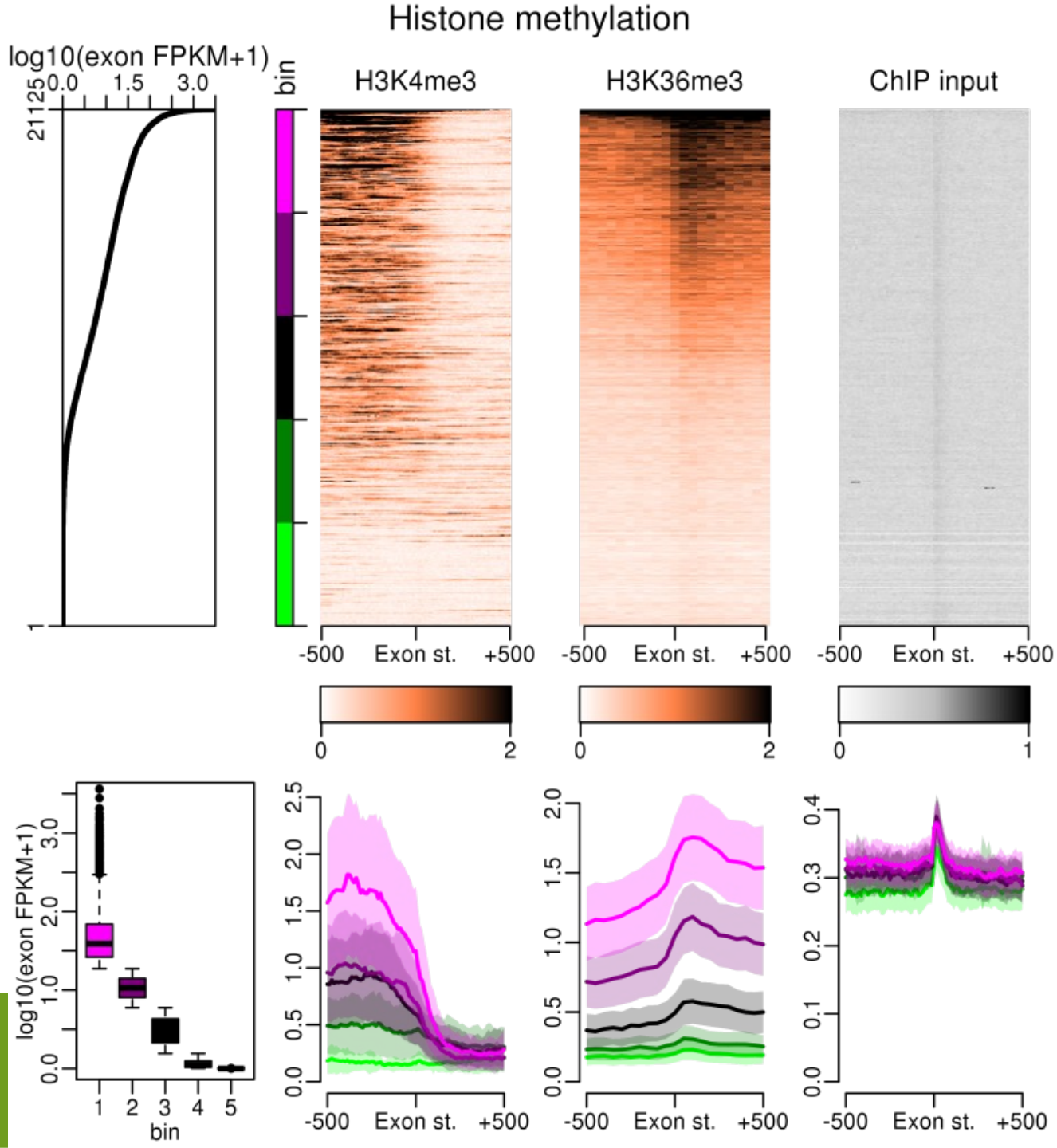

coverage
Slope
43665
(99%)
428
(1%)
16
(0%)
5
(0%)
R^2


protein coding genes
mCpG_ratio
Slope
2.5
0.0
−2.5
−5.0
15793
(87.4%)
1704
(9.4%)
520
(2.9%)
52
(0.3%)
0.25
0.50
0.75
R2
lincRNA genes
mCpG_ratio
Slope
1
0
−1
−2
−3
12
(0.2%)
75
(1.5%)
4539
(92%)
308
(6.2%)
0.25
0.50
0.75
R2
miRNA genes
mCpG_ratio
Slope
1
0
−1
−2
−3
34
(2.7%)
1223
(96.6%)
8
(0.6%)
1
(0.1%)
0.25
0.50
0.75
R2

Epigenetic marks at intron/exon boundaries?



Epigenetic marks at intron/exon boundaries?



Perspectives

- Epigenetic marks and **alternative** splicing?
- Gene by gene tool in webapp
- Polish, write and submit

Thanks

Illustrations: smart.servier.com, BY

ROSLIN


Anagha Joshi


Anna Mantsoki


Barry Horne


Tom Michoel


Angeles Arzalluz-Luque


Deepti Vipin


Pia Francesca Loren Reyes

BBSRC

CASCADE-FELLOWS

GenEpi@GenPhySE

Epigenetics and heritability

- missing heritability
- Improving livestock epigenome

Non-heritable epigenetics

- Early prediction?

FAANG
Functional Annotation of Animal Genomes

- Variant annotation

(physiology and nutrition)