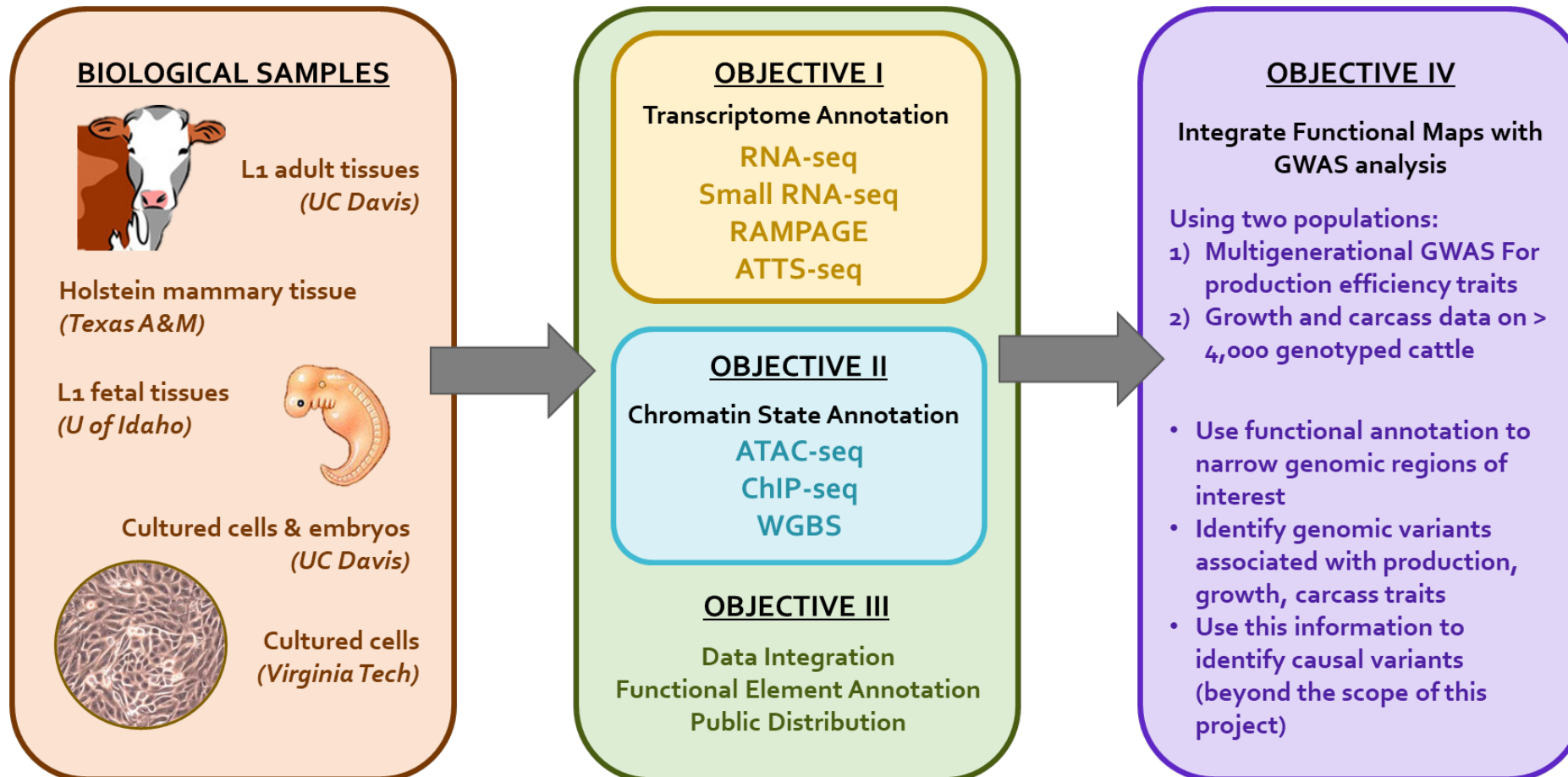


# Functional Annotation of the Bovine Genome

USDA-NIFA-AFRI GRANT NO: 2018-67015-27500

Huaijun Zhou, Pablo Ross, Stephanie McKay, Monique Rijnkels, Brenda Murdoch, Tim Smith, Clare Gill, Zhihua Jiang, James Reecy, James Koltes, Wansheng Liu, Honglin Jiang, Juan Mendrano, Angela Canovas, Graham Plastow

University of California Davis, University of Vermont, Texas A&M University, University of Idaho, USDA-ARS, Washington State University, Iowa State University, Pennsylvania State University, Virginia Tech, University of Guelph, University of Alberta.



# Acknowledgements



Sarah Corum, Gonzalo Rincon, Avinash Baktula, Donald Nkrumah



Shally Xu, Zhangyuan Pan, Dailu Guan, Ying Wang, Claire Prowse-Wilkins



Kim Davenport, Gabrielle Becker, Morgan Stegemiller, Denise Konetchy



Jannifer Michal, Xiaohui Zhang, Yunqi Zhang, Hongyang Wang



Pengcheng Lyu



Kristen Kuhn



Chandlar Kern, Wenzhi Ma

Hamid Beiki



Guosong Wang

# Samples collected and epigenomic data generated

- 40 adult tissues from L1 Hereford line
- 8 fetal tissues from L1 Hereford line
- 4 primary cell lines (Pre-Myocytes, Myocytes, Pre-Adipocytes, Adipocytes)
- 5 stages of Holstein mammary gland development

## ASSAYS-BY-SEQUENCE

### Expressed regions

<b>RNA-seq</b>	Large transcripts expression - variants
<b>smRNA-seq</b>	Small transcript expression
<b>RAMPAGE</b>	Transcription start sites – promoter activity
<b>WTTS-seq</b>	Transcription termination sites

### Chromatin states

	WGBS	DNA methylation
	ATAC-seq	Open Chromatin profiling
<b>ChIP-seq</b>	H3K4me3	Active promoters
	H3K27me3	Polycomb repression
	H3K4me1	Active enhancers
	H3K27ac	Enhancers and promoters
	CTCF	Insulators and promoters
	H3K9me3	Heterochromatin
	H3K36me3	Active gene bodies

## Tissues analyzed for RNA-seq and small RNA-seq

Tissue	Tissue	Tissue	Tissue
abomasum	esophagus	tongue	ovary
adipose	fetal brain	jejunum	pituitary
mammary gland (adult)	fetal gonad	kidney cortex	placental caruncle
bladder	fetal heart	kidney	placental cotyledon
bone marrow	fetal kidney	mammary gland (latepregnant)	reticulum
brain (frontal cortex)	fetal liver	longissimus dorsi muscle	rumen
cecum	fetal muscle	lung	skin
cerebellum	fetal spleen	lymph nodes	spleen
colon	fetal umbilicalcord	mammary gland (virgin)	testis
duodenum	follicular cells	mammary gland (midpregnant)	thymus
mammary gland (earlylactating)	gall bladder	myoblast	uterine endometrium
epididimus caput	Ileum	omasum	

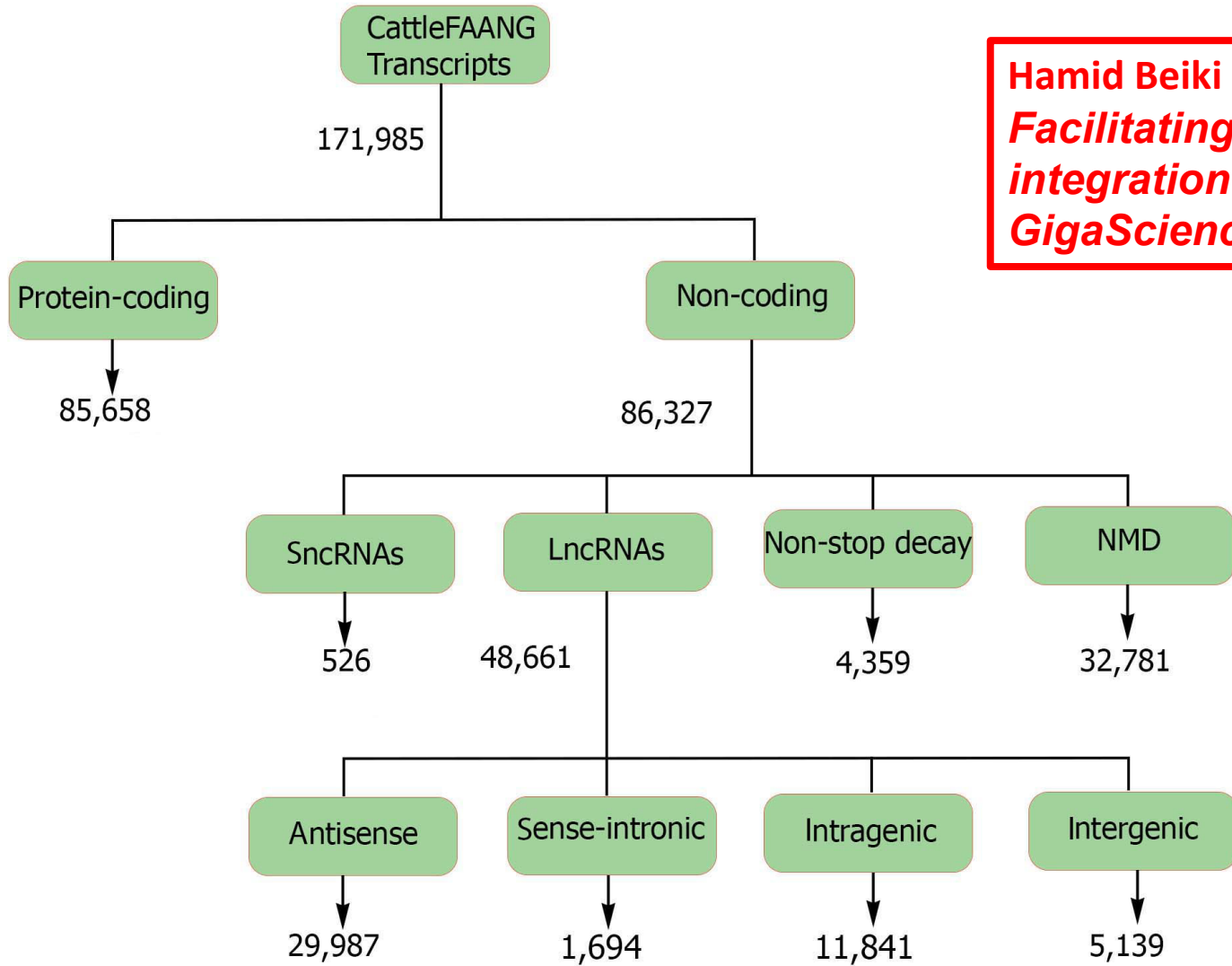
Adult and fetal tissues from cattle closely related to Dominette

# Summary of expressed transcripts/genes

Feature	Annotation <sup>1</sup>		
	Current project	<u>Ensembl</u> (Release 2021-03)	NCBI (Release 106)
Number of genes	35,150 (21,193)	27,607 (21,880)	35,143 (21,355)
Number of transcripts	171,985 (85,658)	43,984 (37,538)	83,195 (47,280)
Number of spliced transcripts	130,531	37,299	73,423
Number of transcripts per gene	4.9	1.5	2.3
Median number of 5' UTRs per gene	2	1	1
Median number of 3' UTRs per gene	1	1	1

<sup>1</sup>Numbers in parentheses indicate the number of protein-coding genes/transcripts.

# Classification of the predicted transcripts into different biotypes



Hamid Beiki et al.

*Facilitating Functional genomics of cattle through integration of multi-omics data*  
*GigaScience, accepted*

# Genes/Transcripts expressions across tissues between adult and fetal stages

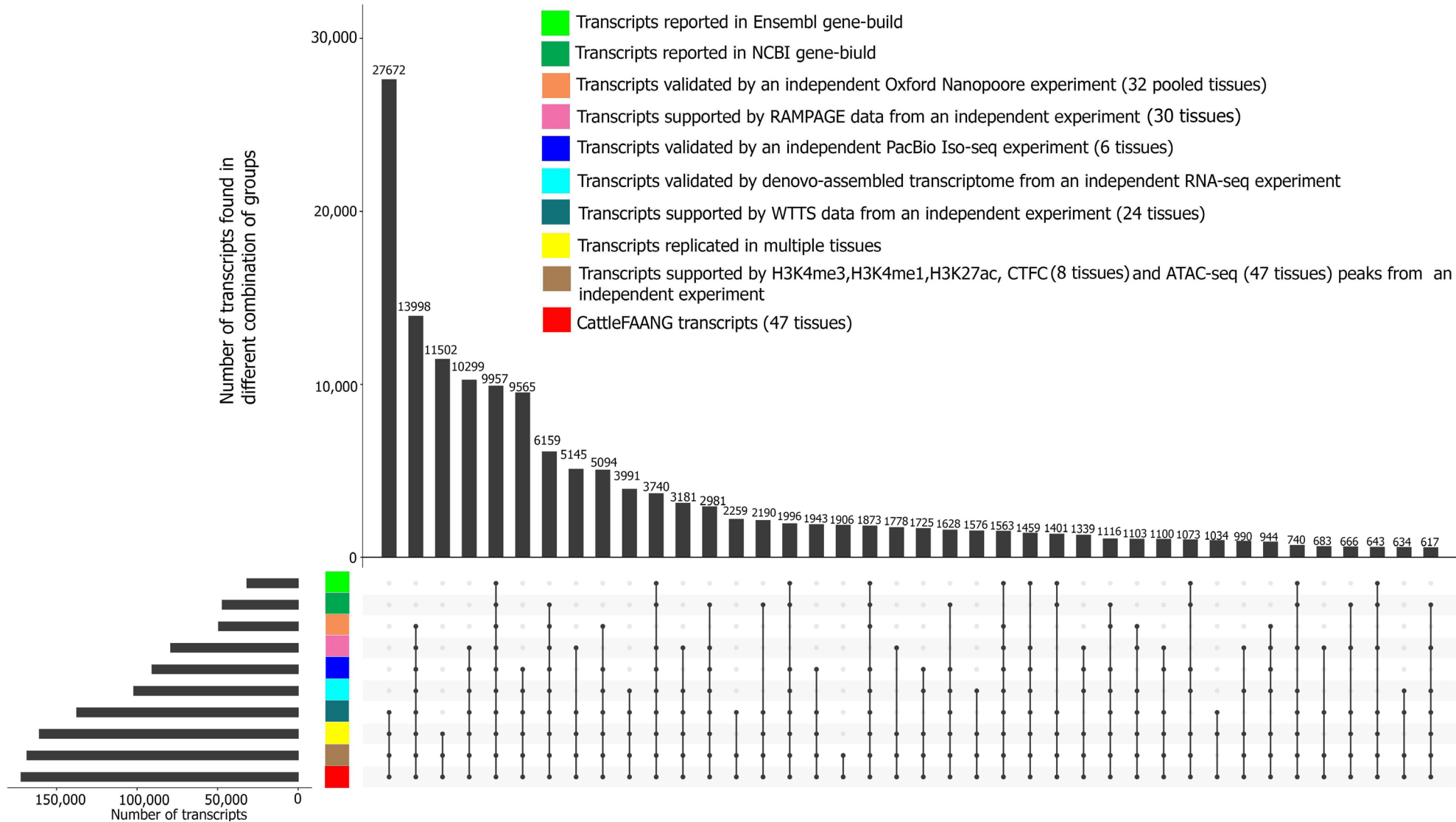
- Fetal testis showed the highest number of **expressed genes** while **fetal brain and muscle** tissues showed the highest number and percentage of **non-coding genes**
- Fetal tissues had a significantly higher rate of **alternative splicing** events than **adult tissues**
- Fetal tissues had significantly higher proportions of **unique non-coding** transcripts compared to **protein-coding** transcripts than adult tissues
- 106 **non-coding genes** identified in **fetal tissues** that switched to **protein-coding genes** in their matched **adult tissues**

# Tissue specificity

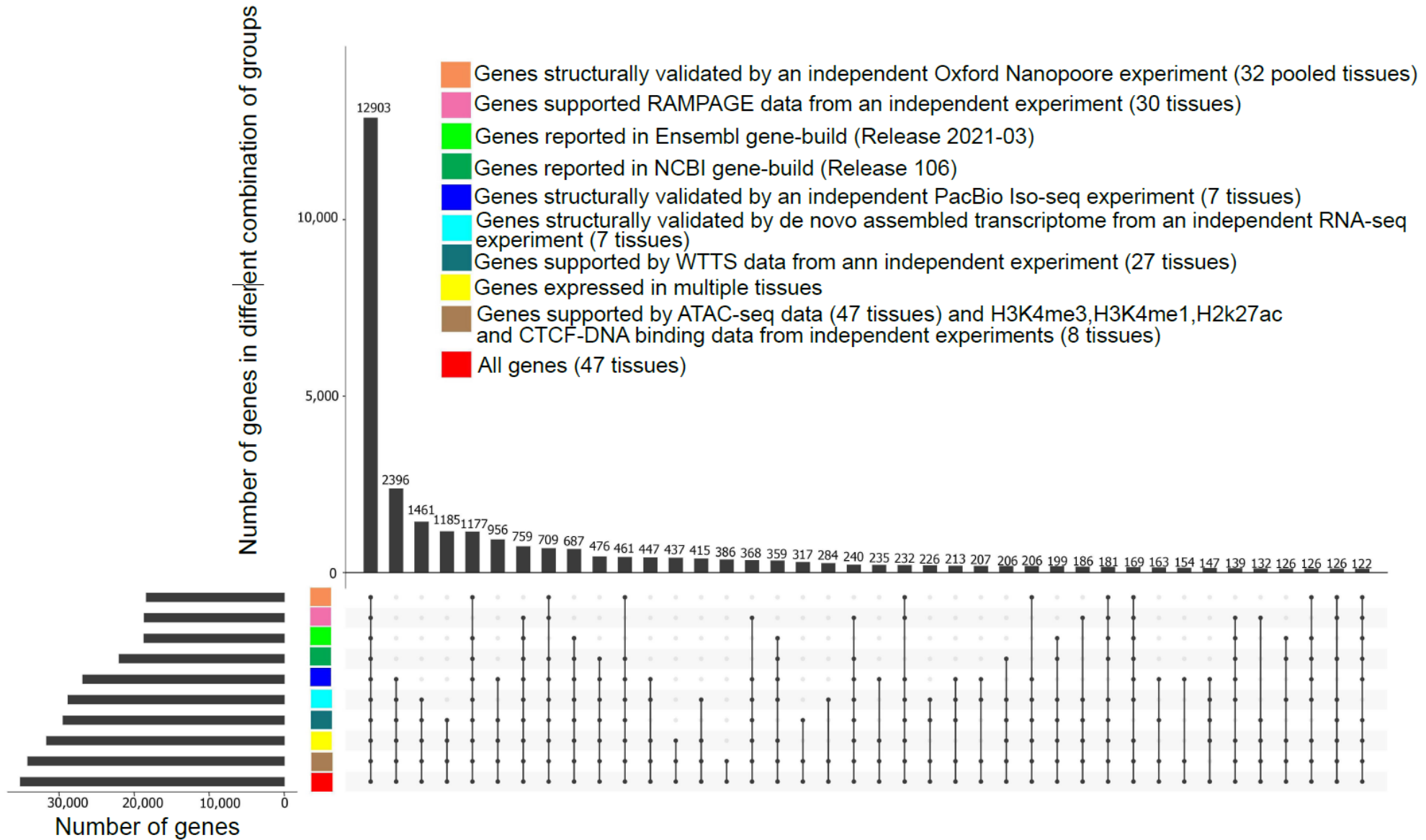
- **Nine** percent of all genes (3,174) and transcripts (15,562) were only expressed in **a single tissue**
- The majority of tissue-specific genes (75%) and transcripts (84%) were **un-annotated**
- **Testis and thymus** had the highest number of tissue-specific genes and transcripts
- As we expected, the expression level of **tissue-specific genes** and transcripts was significantly lower than that of their **non-tissue-specific** counterparts



# Validation of predicted transcripts using independent data from different technologies



# Validation of predicted genes using independent data from different technologies

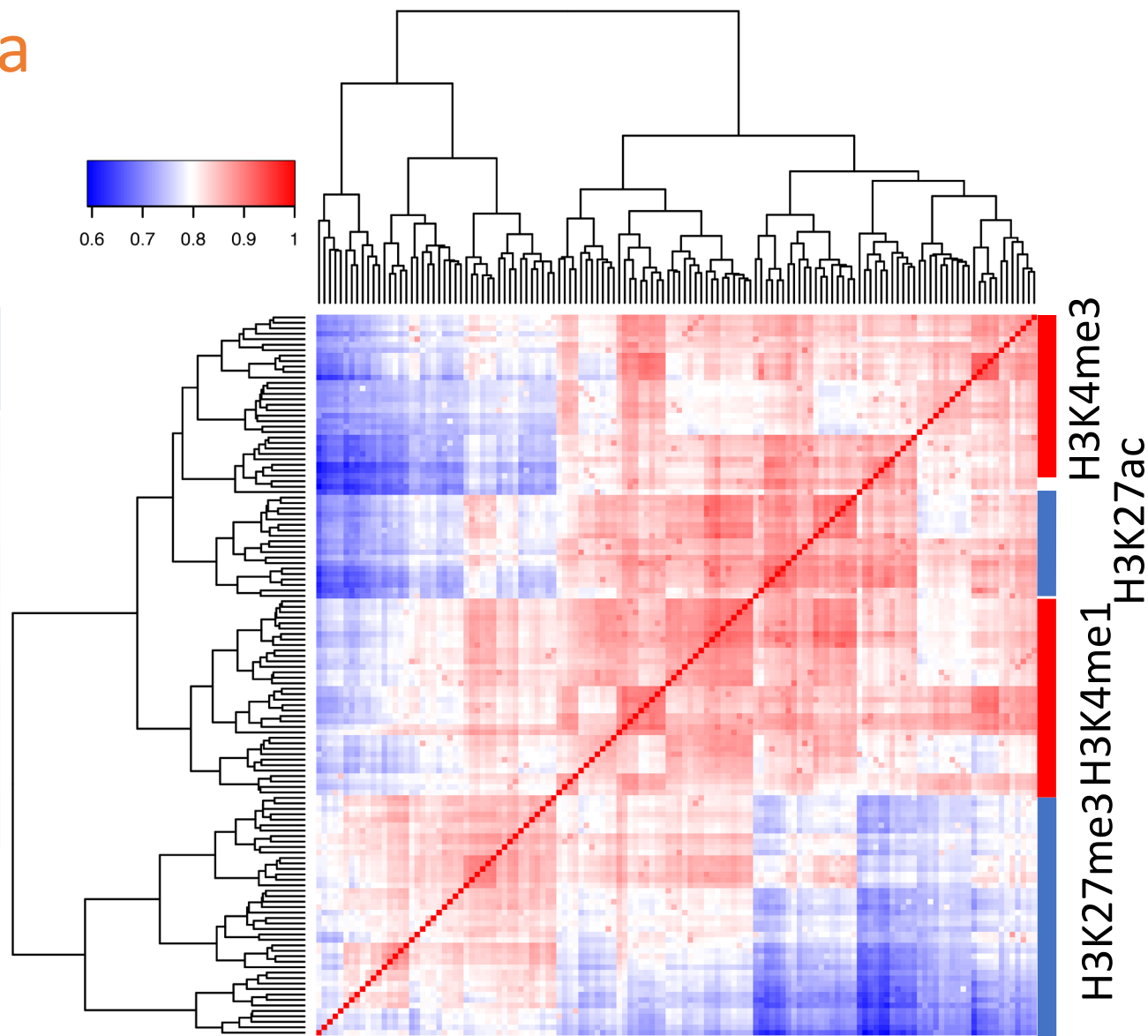


# Generation of epigenomic data

- Histone marks (ChIP-seq)

	Adult tissue	Fetal tissue	Cell line	Total
CTCF	52	20	8	80
H3K27ac	52	20	8	80
H3K27me3	46	20	8	74
H3K4me1	52	20	8	80
H3K4me3	52	20	8	80
H3K36me3	29	20	8	57
H3K9me3	38	20	8	66

- WGBS: 95
- ATAC-seq: 114
- WTTS:76



# Functionally annotate epigenomic elements of the bovine genome

Build a map of of regulatory elements by integrating:

a) Project data

- 27 Adult tissues (Herefords)
- 5 Fetal tissues (Herefords)
- Mammary gland (Holstein 5 stages)
- 4 Cell lines

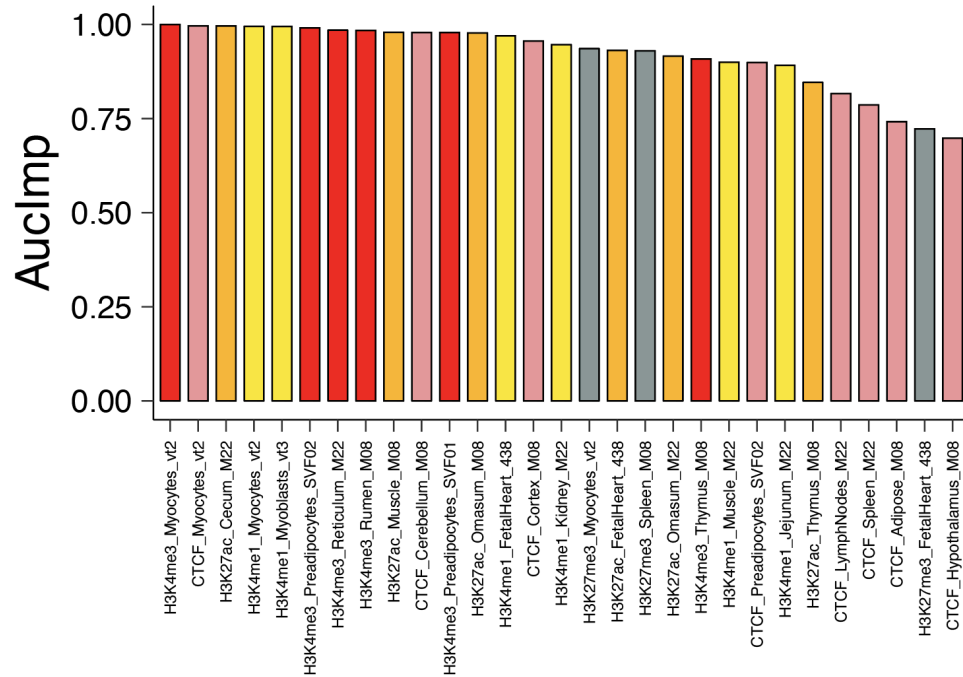
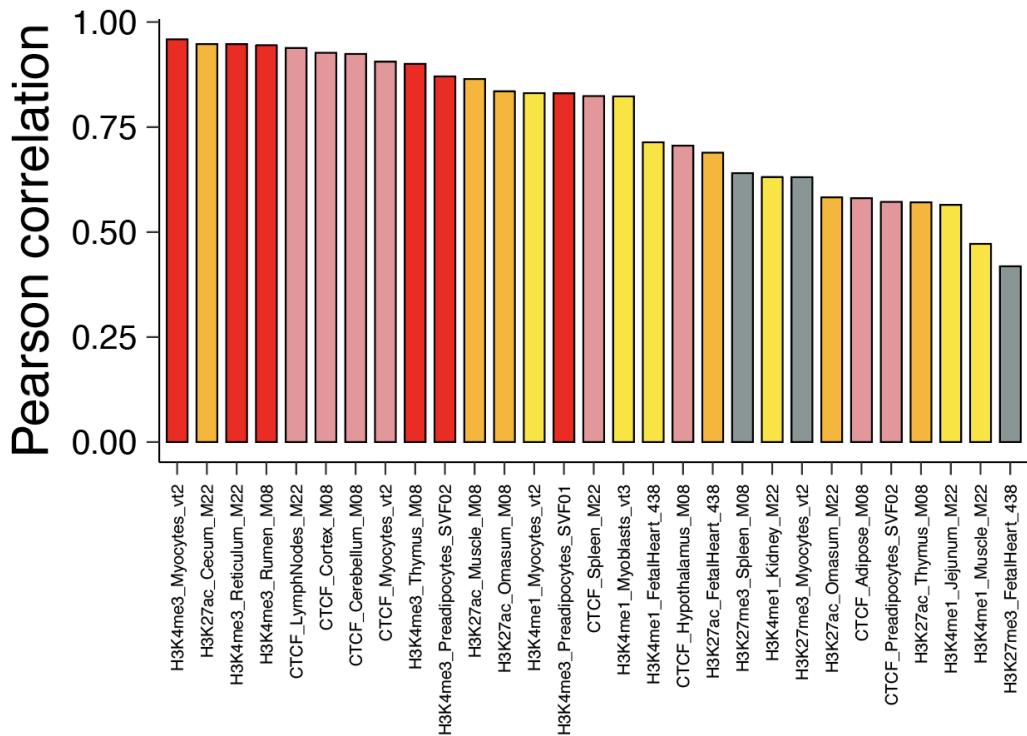
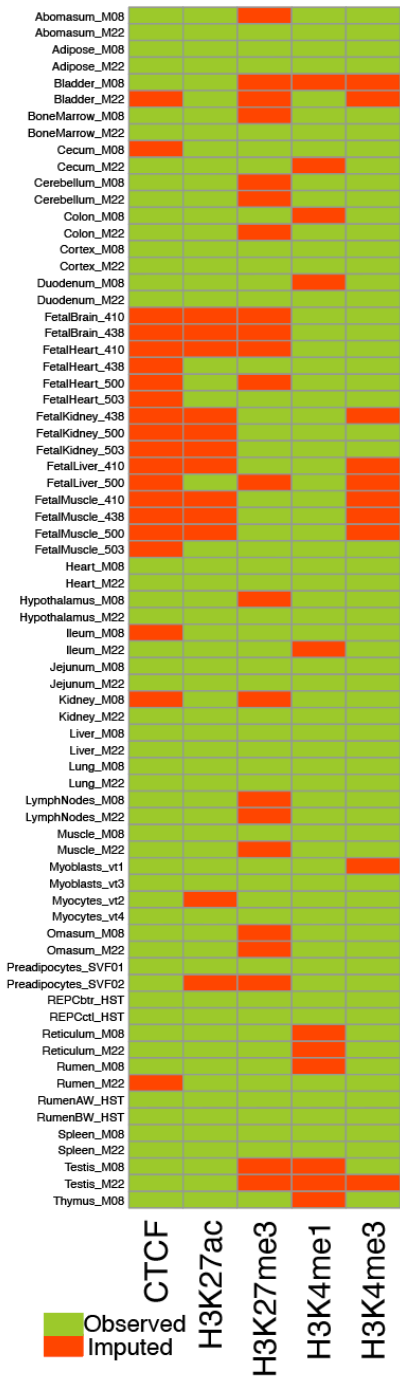
b) Public data (Holstein) ←

- 23 Adult tissues
- 4 Fetal tissues

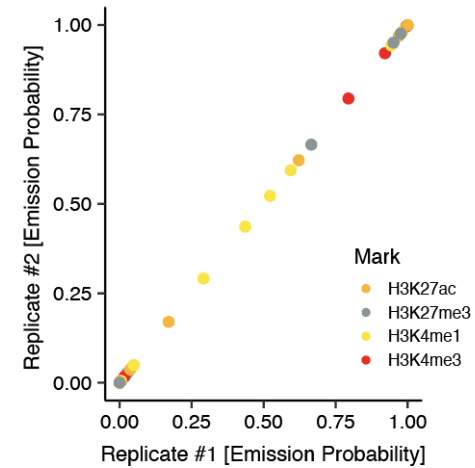
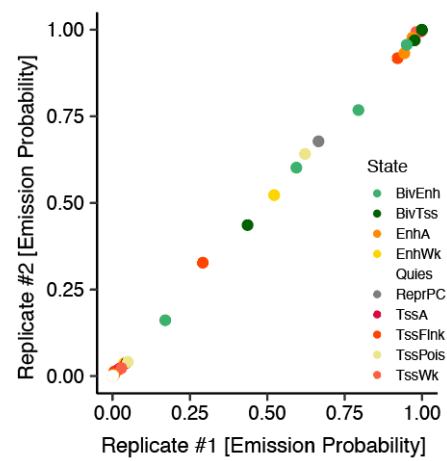
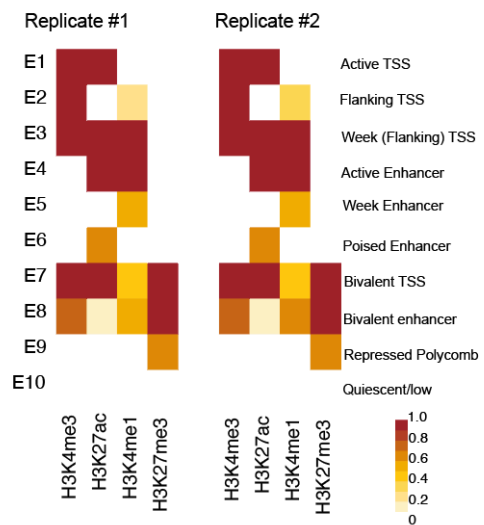
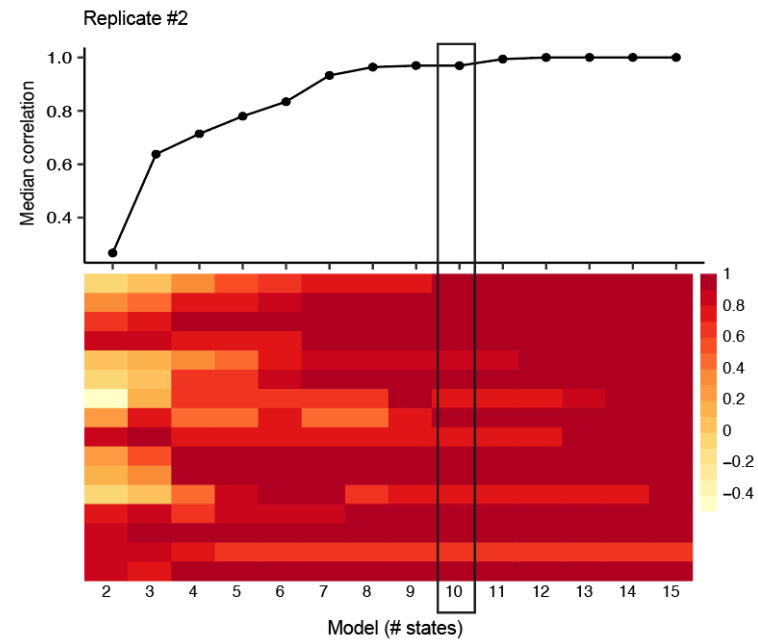
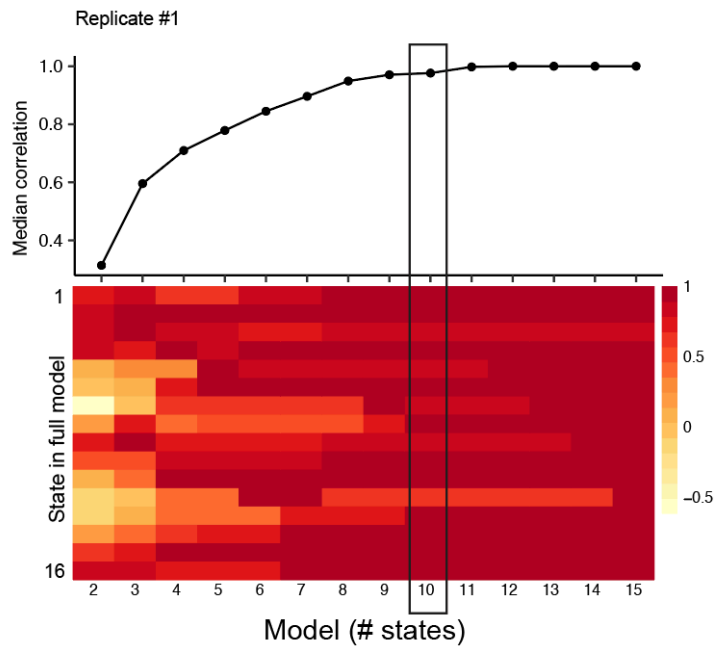
	Adult tissue	Fetal tissue	Cell line	Total
CTCF	47	8	1	56
H3K27ac	47	8	1	56
H3K27me3	47	6	1	56
H3K4me1	51	8	1	60
H3K4me3	51	8	1	60

→ **A total of 46 adult tissues, 6 fetal tissues, and 5 cell lines**

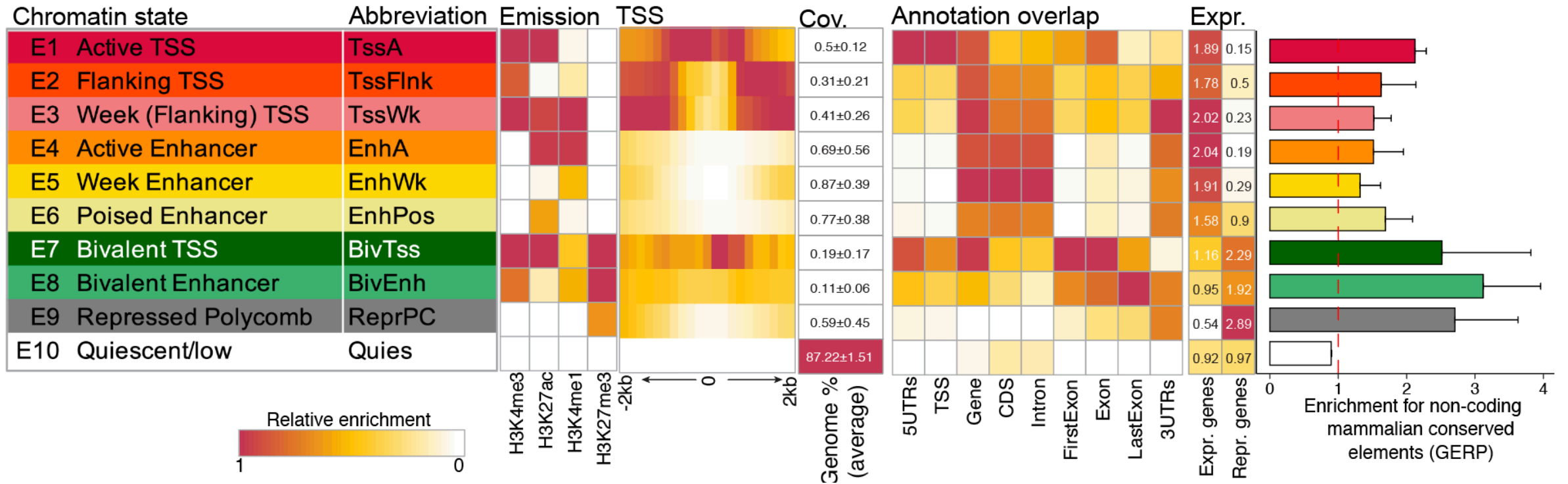
# Imputation of missing ChIP-seq marks



# Chromatin state model optimization

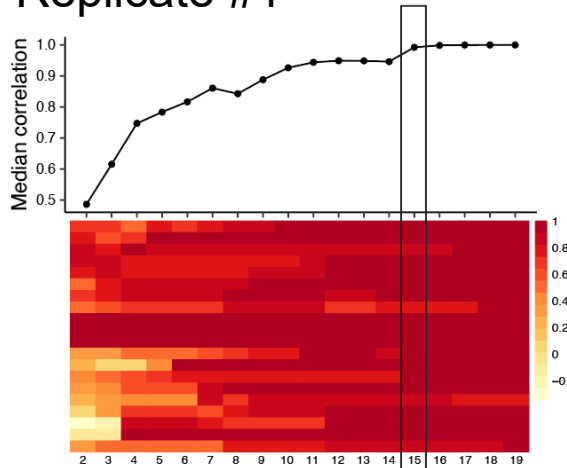


# 'Core' 10-state model

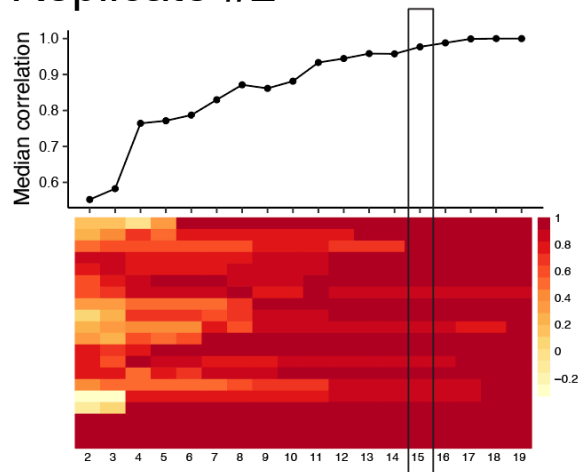


# 'Expanded' 15-state model

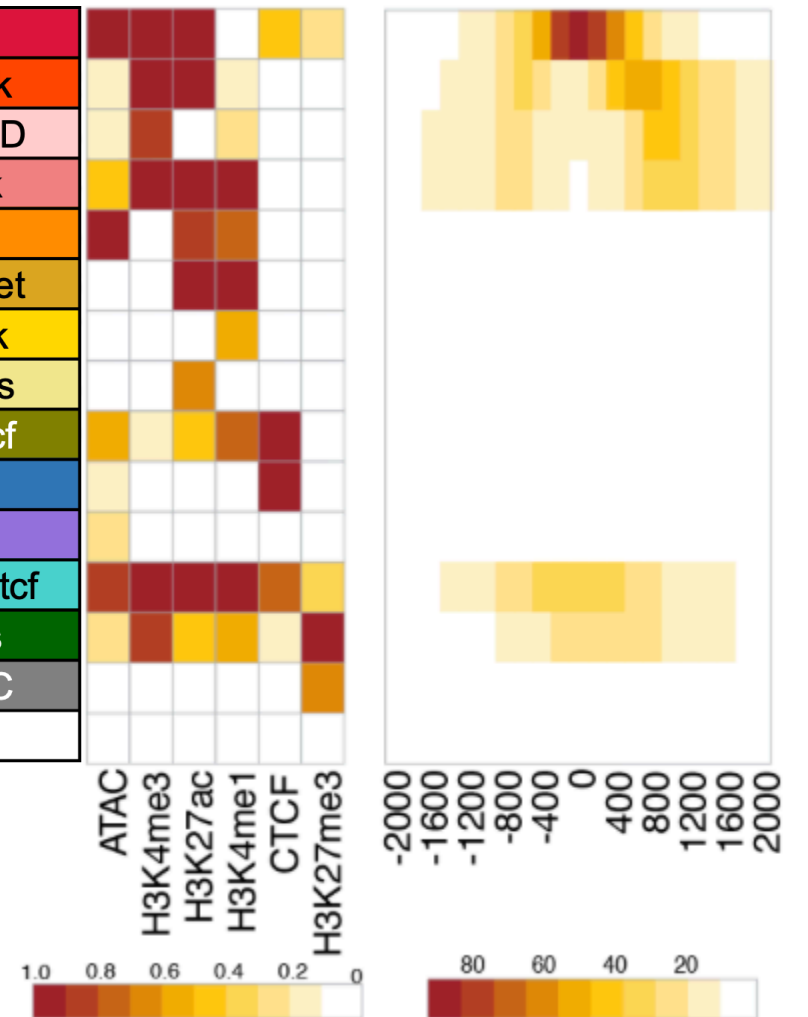
Replicate #1



Replicate #2

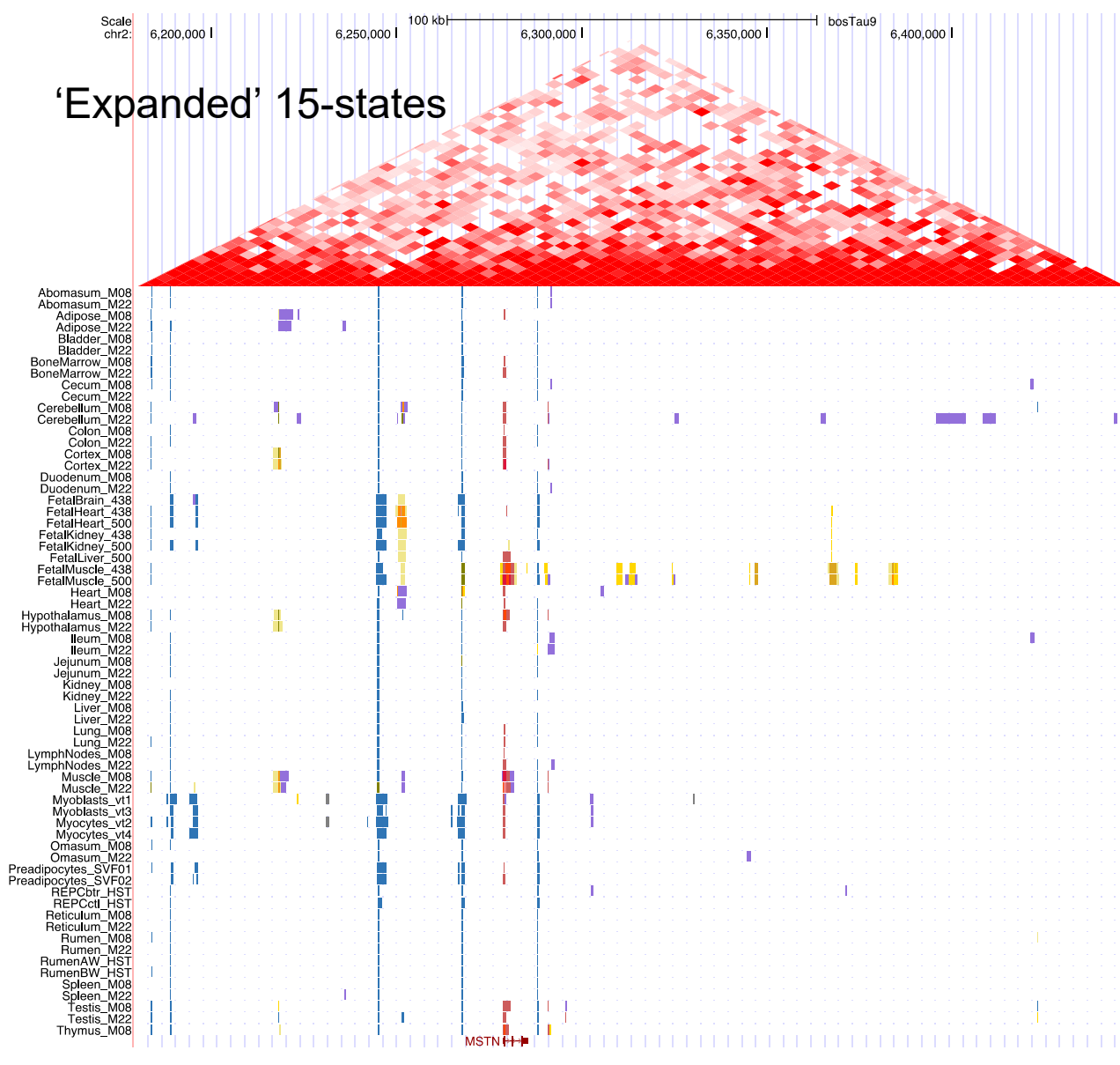
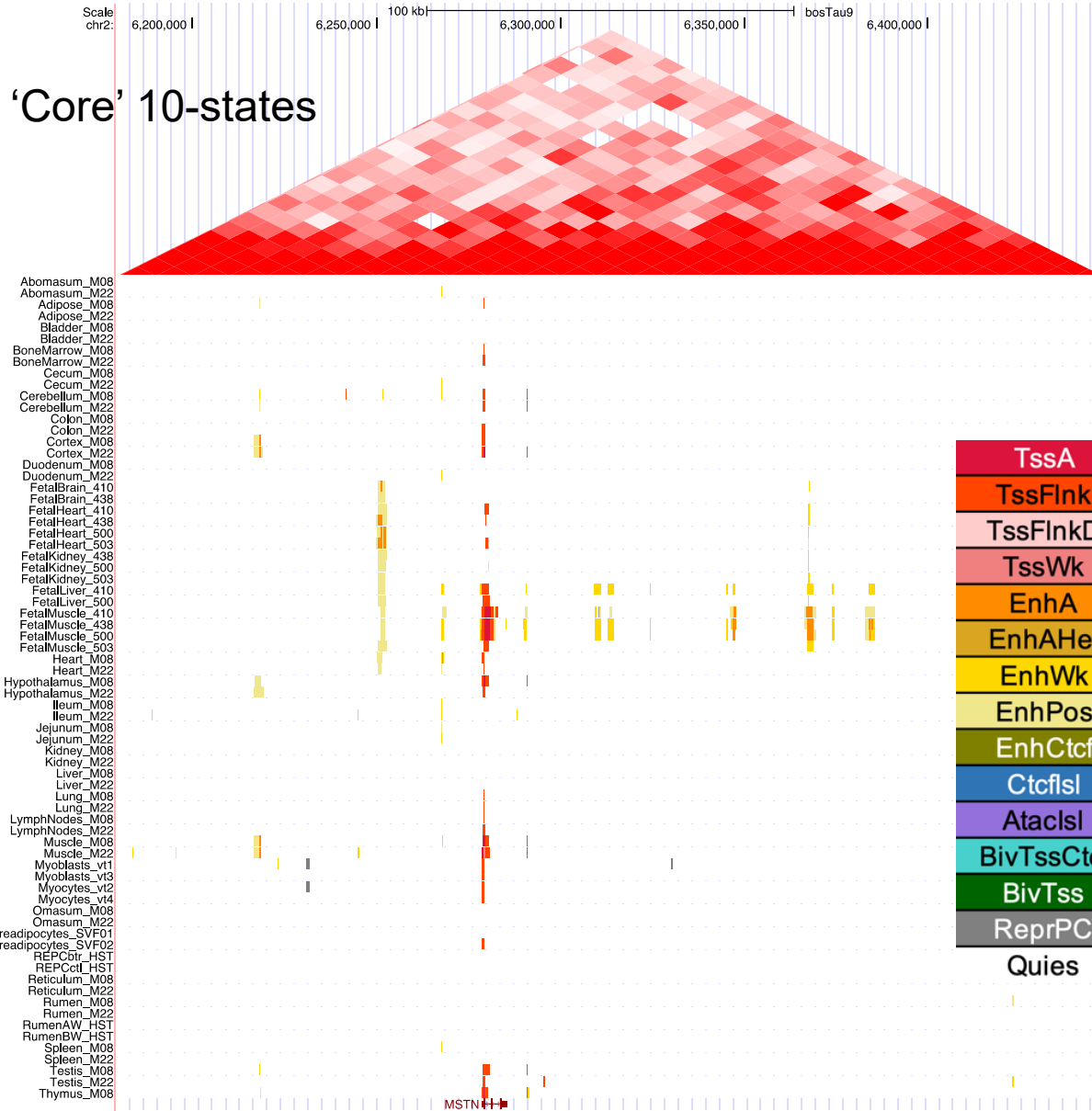


E1	Active TSS	TssA
E2	Flanking TSS	TssFlnk
E3	Flanking TSS downstream	TssFlnkD
E4	Week (Flanking) TSS	TssWk
E5	Active Enhancer	EnhA
E6	Active Enhancer without ATAC (hetero)	EnhAHet
E7	Week Enhancer	EnhWk
E8	Poised Enhancer	EnhPos
E9	Enhancer with CTCF	EnhCtcf
E10	CTCF island	CtcfIsl
E11	ATAC island	AtacIsl
E12	Bivalent TSS with CTCF	BivTssCtcf
E13	Bivalent TSS	BivTss
E14	Repressed Polycomb	ReprPC
E15	Quiescent/low	Quies

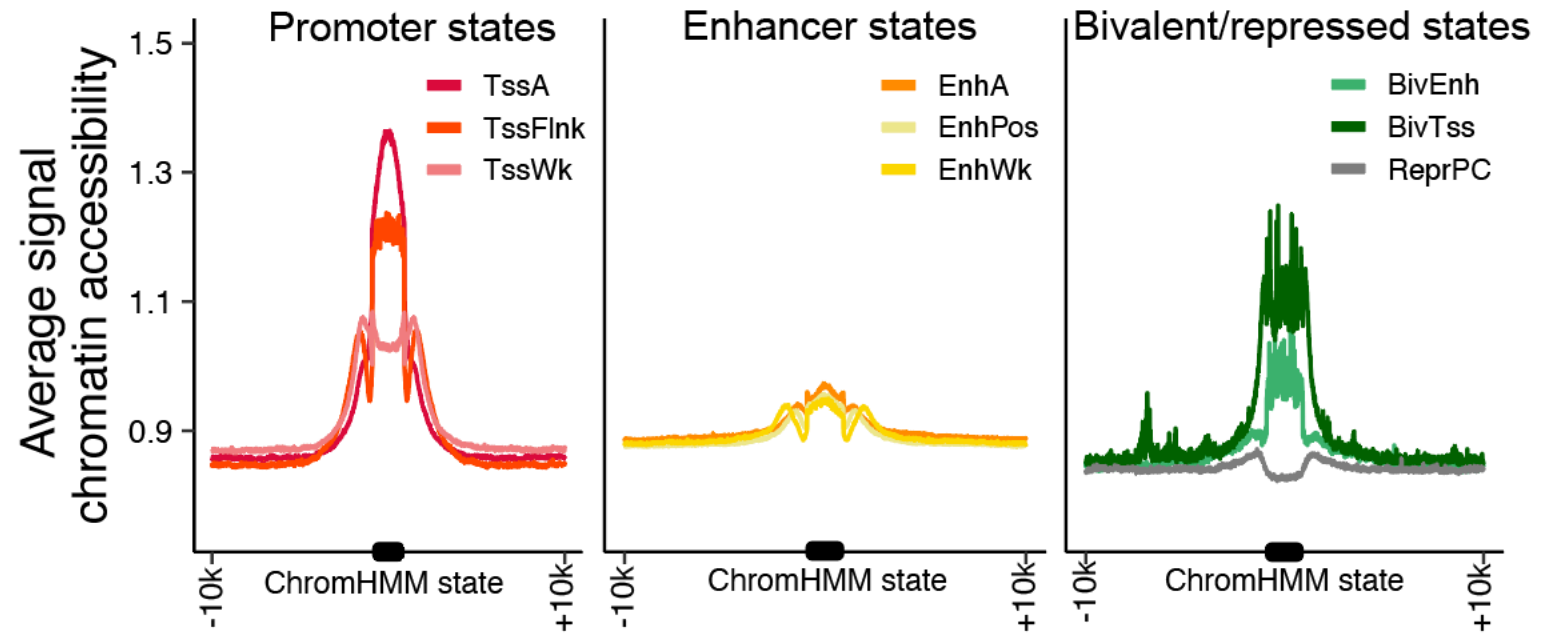
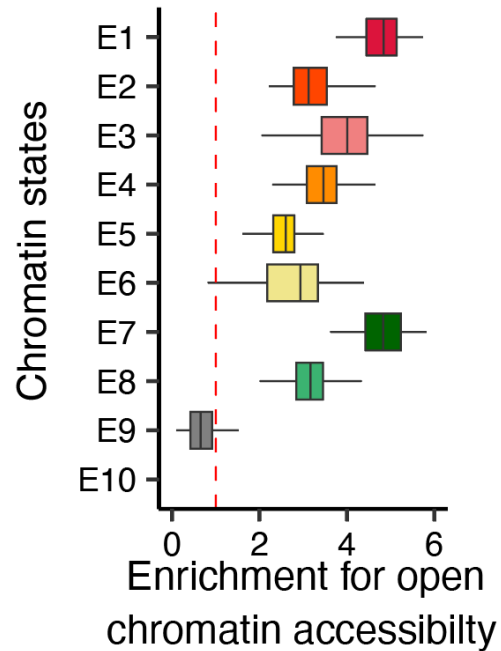




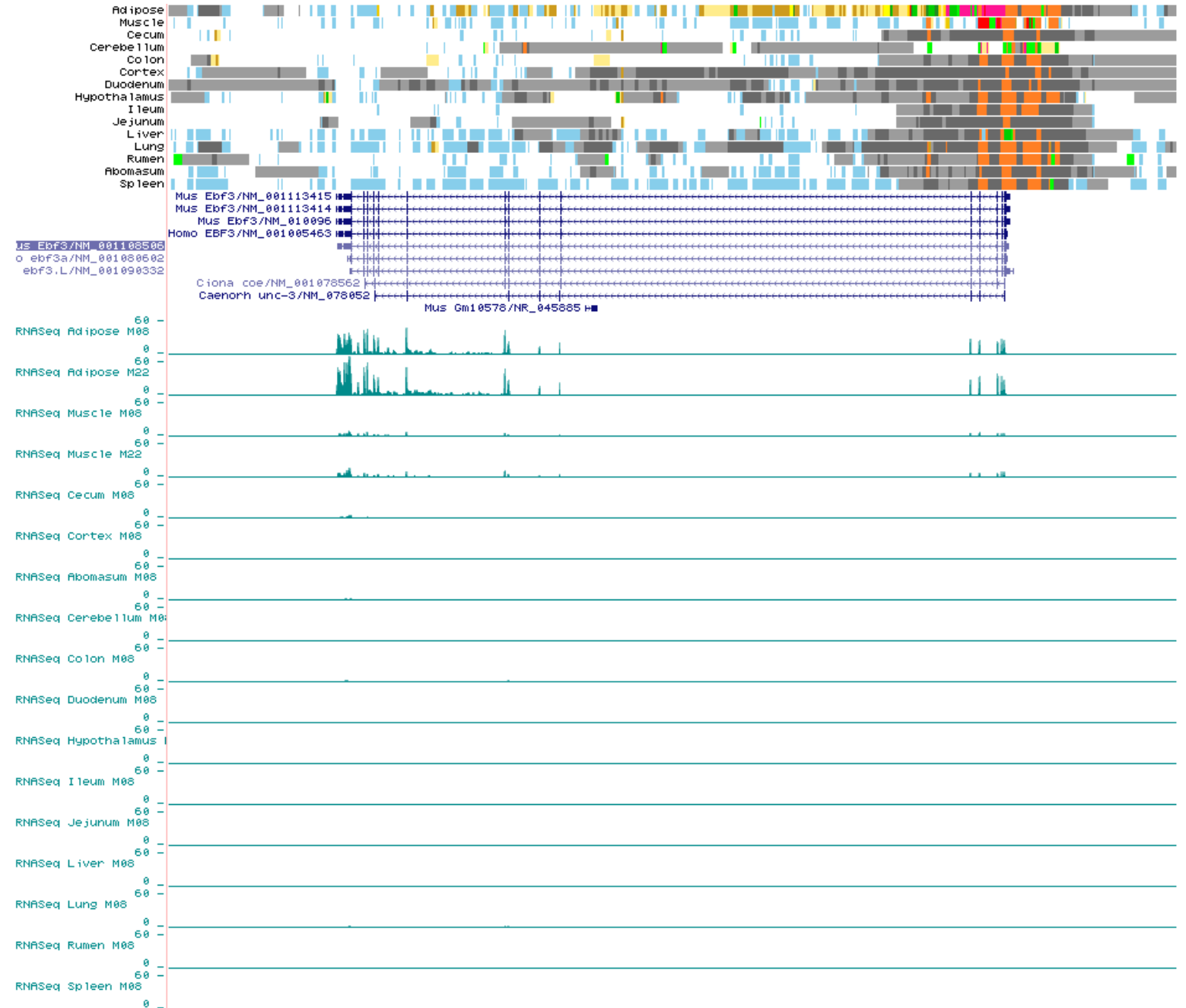
# Chromatin states at the *MSTN* locus



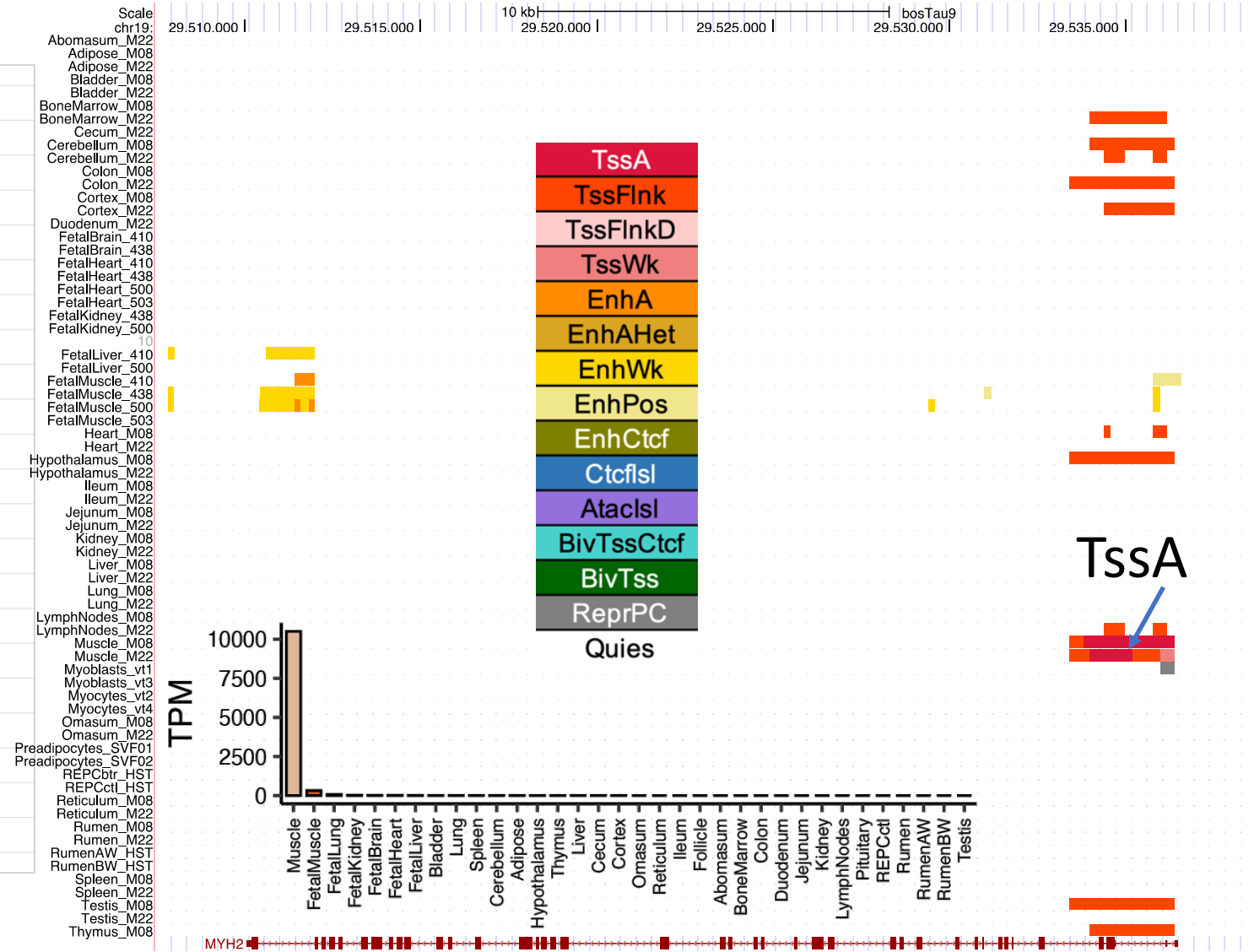
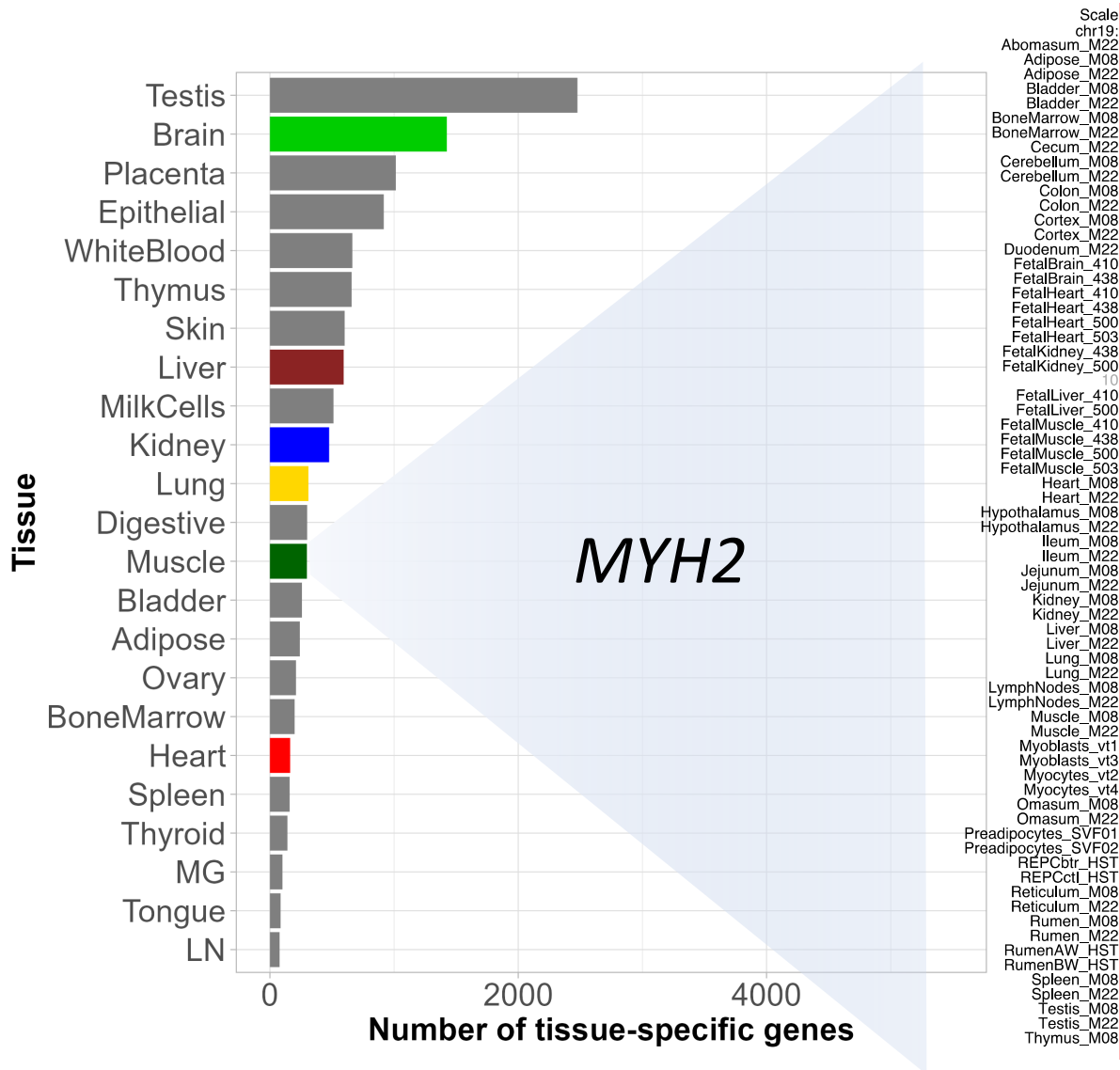
# Open chromatin accessibility regions are indicative of active regulatory elements



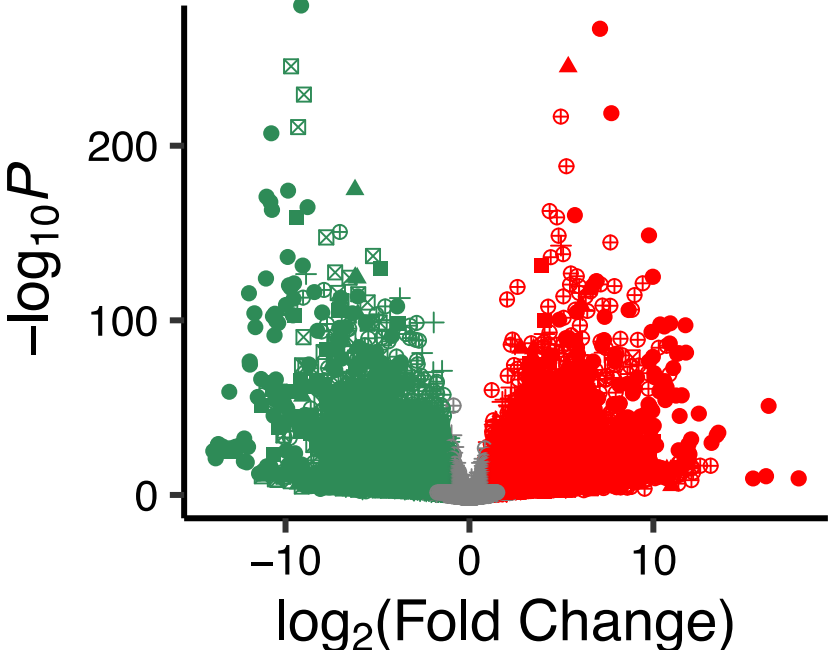
- *EBF3* specifically expressed in adipose tissue.
- Adipose have many tissue-specific enhancers



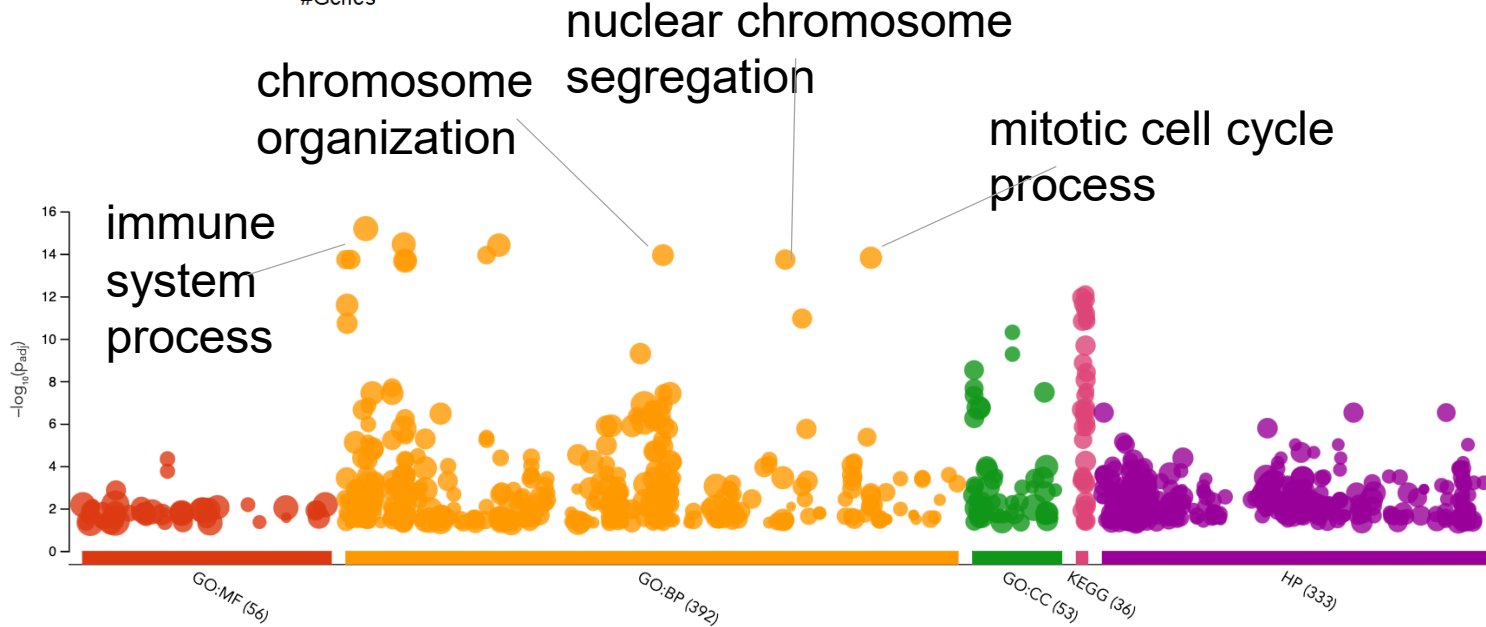
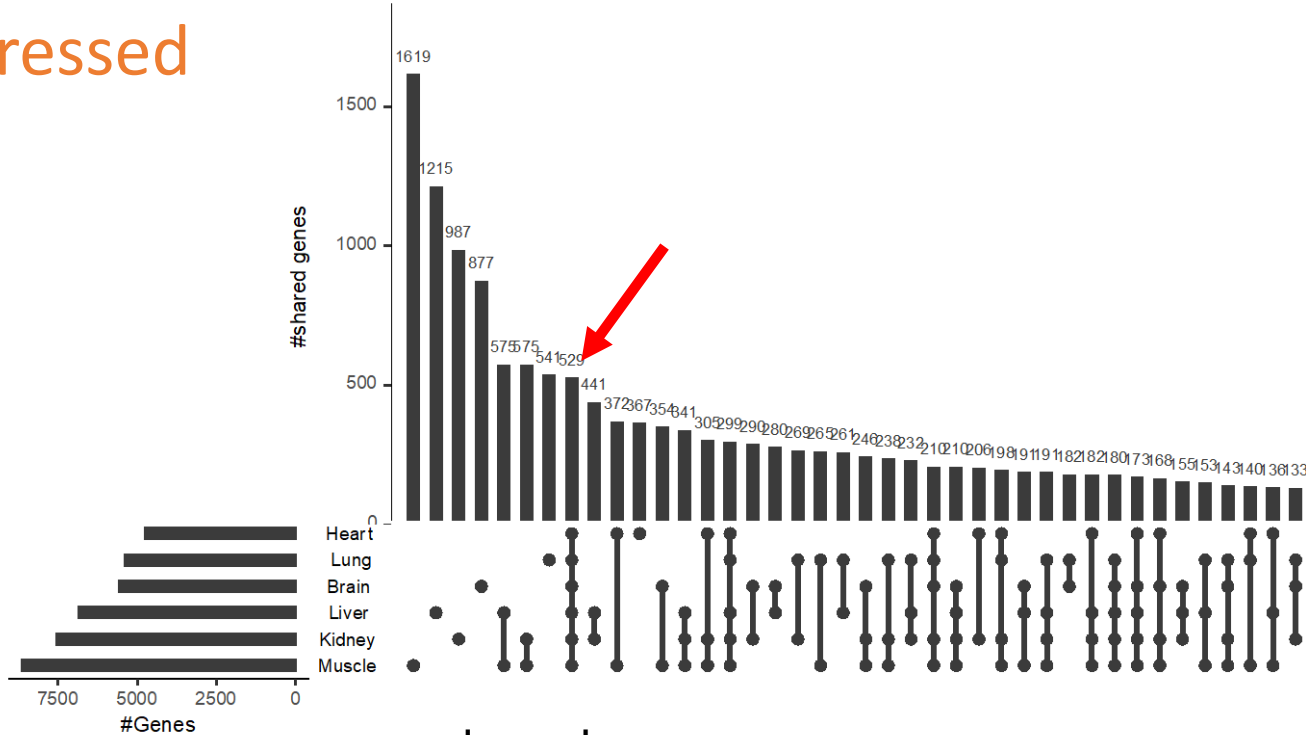
# Tissue-specific gene expression contributed by tissue-specific regulators



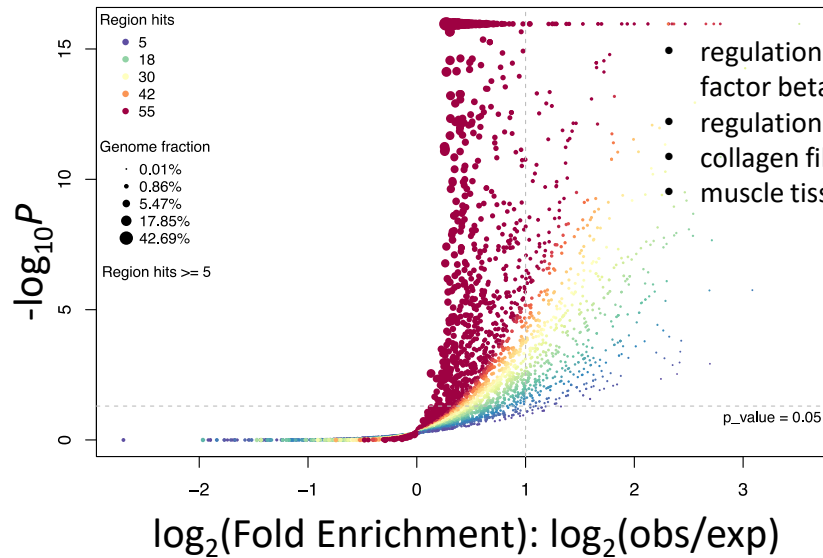
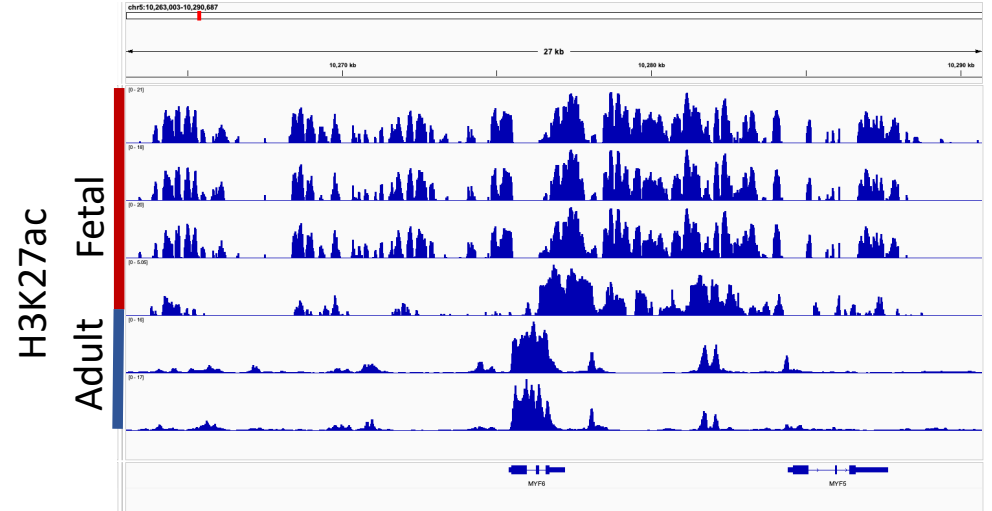
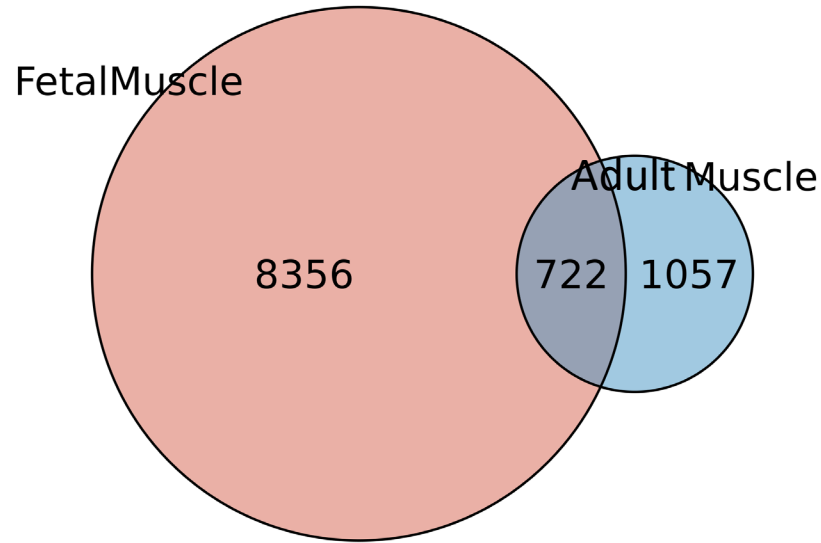
# Thousands of genes differentially expressed between fetal and adult tissues



- + Brain
- ▲ Heart
- Kidney
- Liver
- ⊠ Lung
- ⊙ Muscle
- Not signif
- Signif & downregulated
- Signif & upregulated

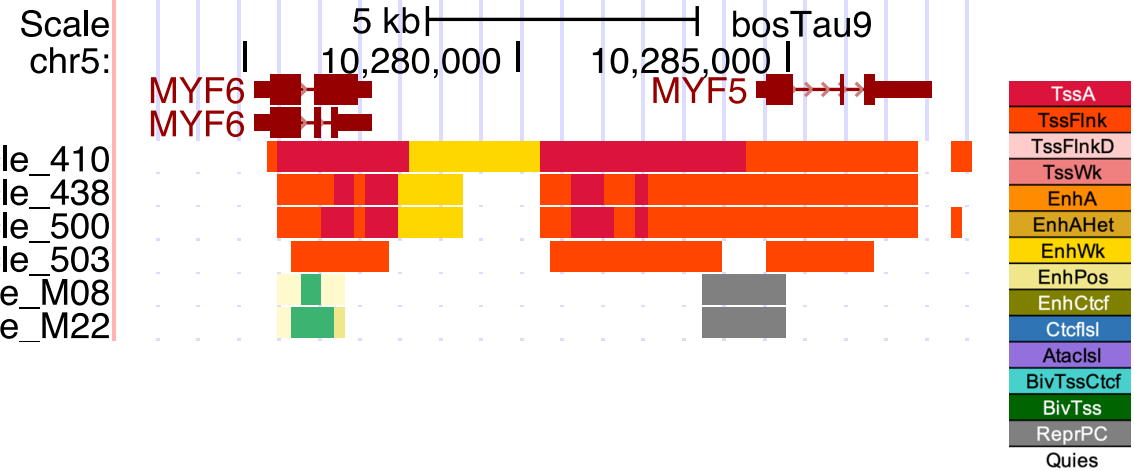


# Differential active enhancer (E4) between fetal and adult muscle

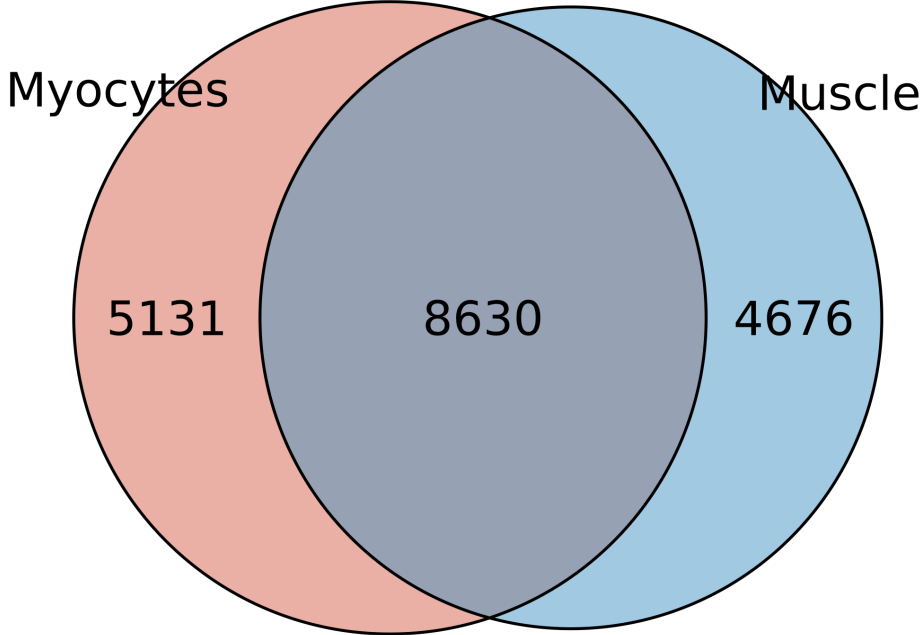


- regulation of transforming growth factor beta production
- regulation of myelination
- collagen fibril organization
- muscle tissue development

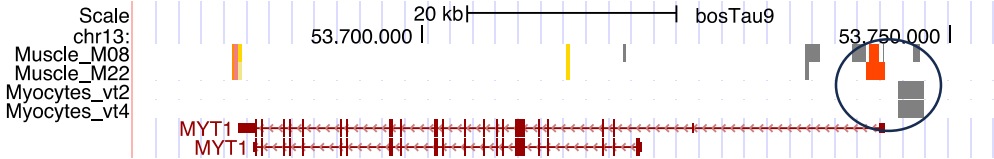
FetalMuscle\_410  
 FetalMuscle\_438  
 FetalMuscle\_500  
 FetalMuscle\_503  
 Muscle\_M08  
 Muscle\_M22



# Primary cell lines vs bulk tissues

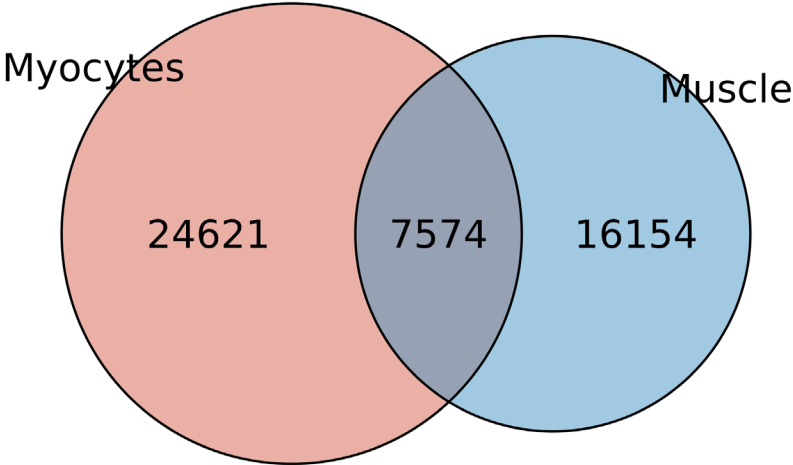


## Active promoter (E1)



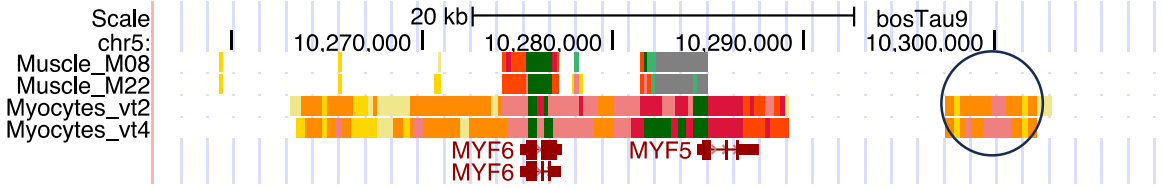
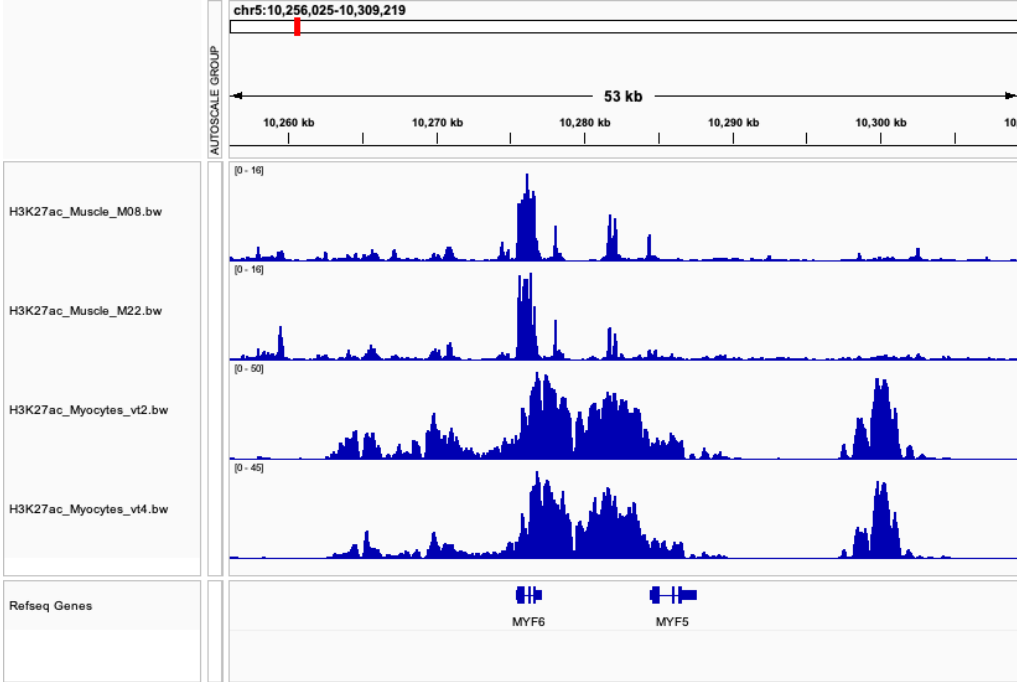
- TssA
- TssFlnk
- TssFlnkD
- TssWk
- EnhA
- EnhAHet
- EnhWk
- EnhPos
- EnhCtcf
- CtcfIsl
- AtacIsl
- BivTssCtcf
- BivTss
- ReprPC
- Quies

# Primary cell lines vs bulk tissues



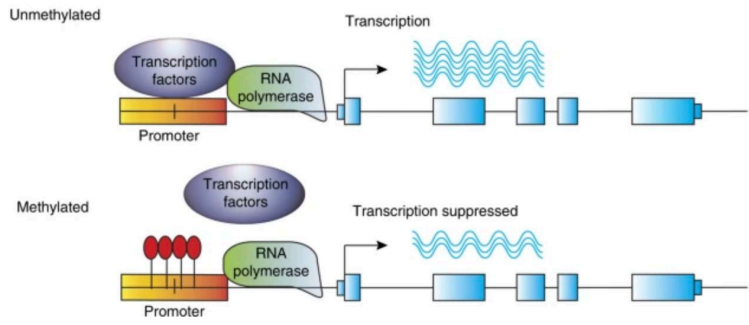
- TssA
- TssFlnk
- TssFlnkD
- TssWk
- EnhA
- EnhAHet
- EnhWk
- EnhPos
- EnhCtcf
- CtcfIsl
- AtacIsl
- BivTssCtcf
- BivTss
- ReprPC
- Quies

## Active enhancer (E4)

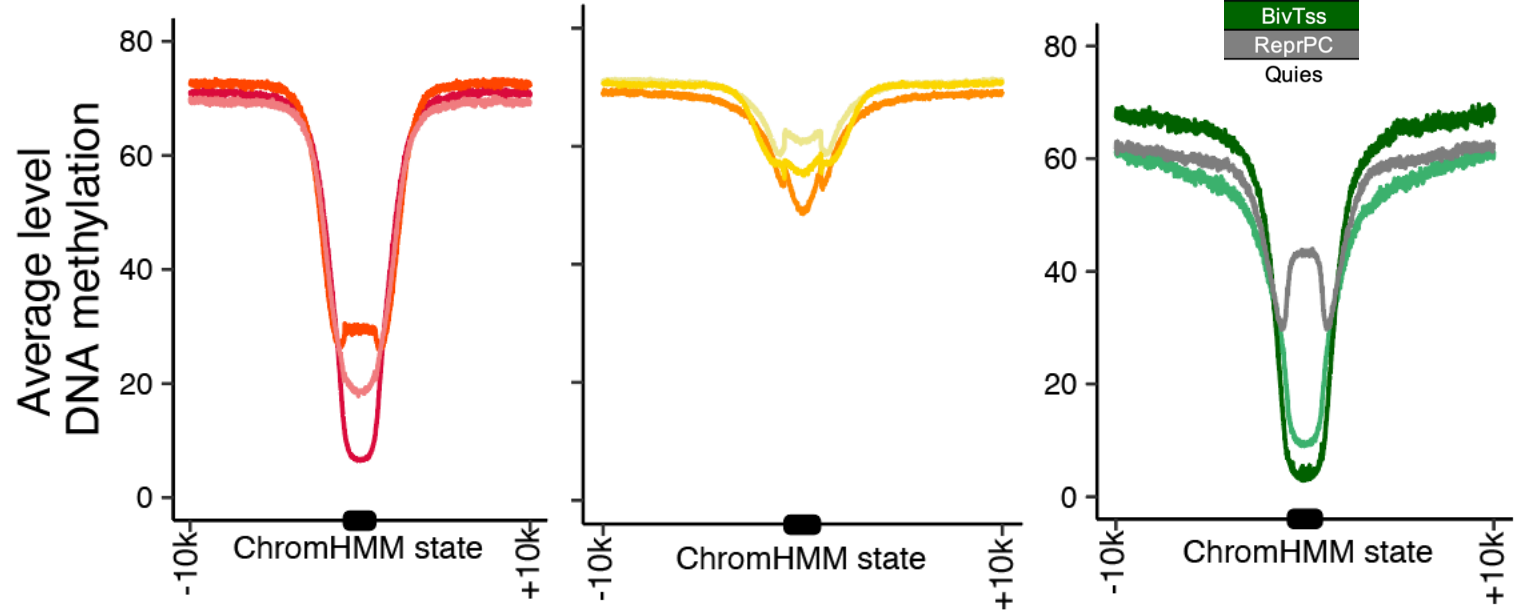
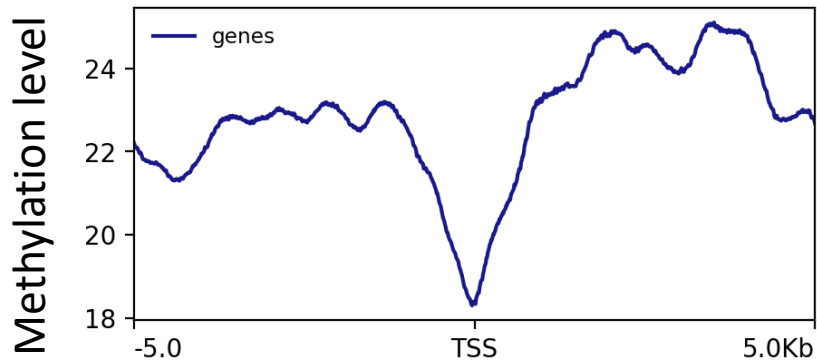




# DNA methylation and chromatin states



Wu et al. 2018

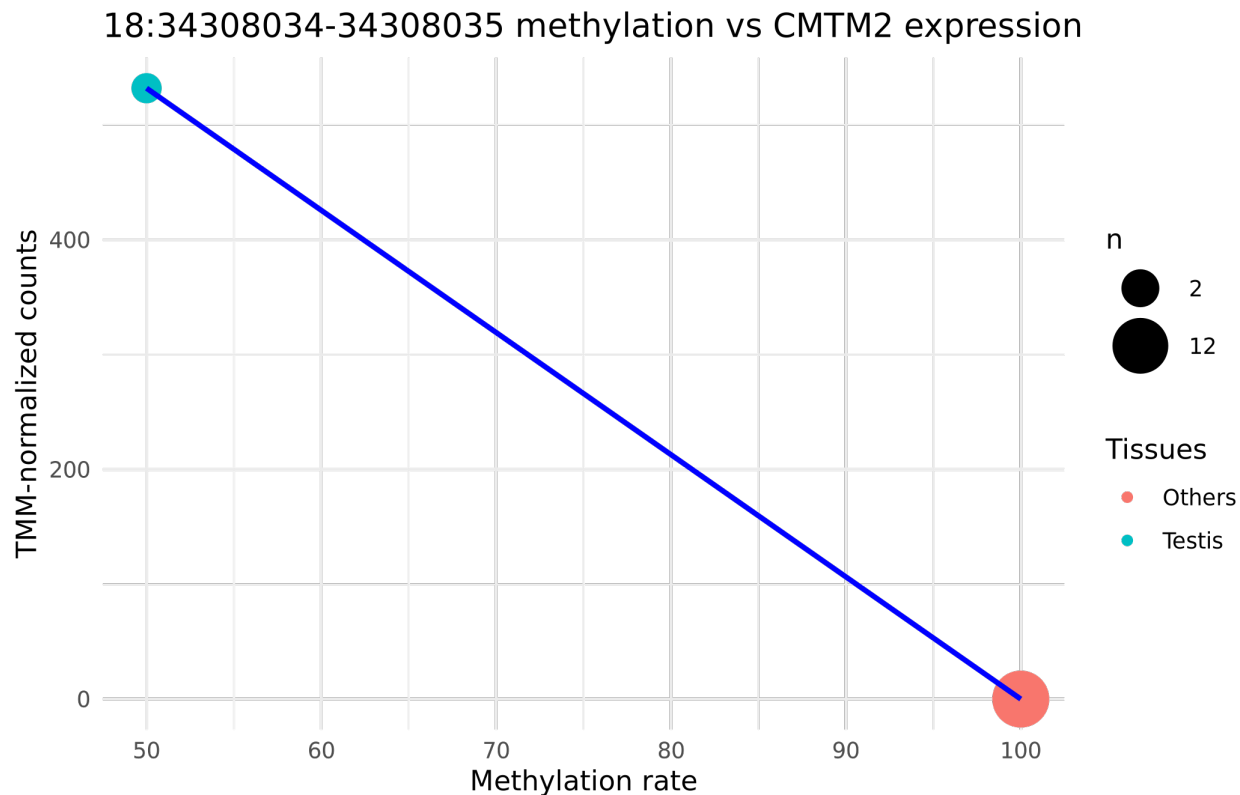


TssA
TssFlnk
TssFlnkD
TssWk
EnhA
EnhAHet
EnhWk
EnhPos
EnhCtcf
CtcfIsl
AtacIsl
BivTssCtcf
BivTss
ReprPC
Quies

- Promoter-like states show lower methylation levels, confirming the well-known negative correlation between promoter methylation and gene expression

# Differentially methylated regions (DMRs) across tissues

- Identified 208,665 differentially methylated regions (DMRs) across tissues, predicted to affect 1,080,550 motifs



**Two cytosines** within a DMR for Testis correlated ( $\rho = -1$ ) with *CMTM2* → spermatogenesis and reproduction<sup>1</sup>

<sup>1</sup>10.1016/j.theriogenology.2019.07.026

# Number of DMRs and Potential motif disruption by promoter methylation

