Functional annotation of the chicken genome

FAANG Workshop - PAG 2024

Yvonne Drechsler, Ph.D.
Professor
College of Veterinary Medicine
Western University of Health Sciences

R. David Hawkins, Ph.D.
Professor
DOM & Genome Sciences
University of Washington

FAANG
Functional Annotation of Animal Genomes

Chicken Genome Annotation

- Transcriptomics
  - Identify mRNA and lncRNAs expressed
  - determine isoforms
  - any putatively novel genes
  - Allele-specific expression (ASE)

- Epigenomics
  - Identify candidate *cis*-regulatory elements.
  - DNA methylation (WGBS)
    - DMRs for key comparisons
    - Allele-specific DNAme (ASM)
  - Chromatin dynamics
    - CUT&Tag for 5 histone modifications
      - QC greatly improved
    - ATAC-seq

Transcriptome & Epigenomic annotation of *cis*-regulatory elements

Targeted Tissues

- Immune cells
- Female reproductive tissues
- Large intestine
- Tissue-resident immune cells
- Immune-related tissues
- Muscle tissues
- Total: 20 cells/tissues
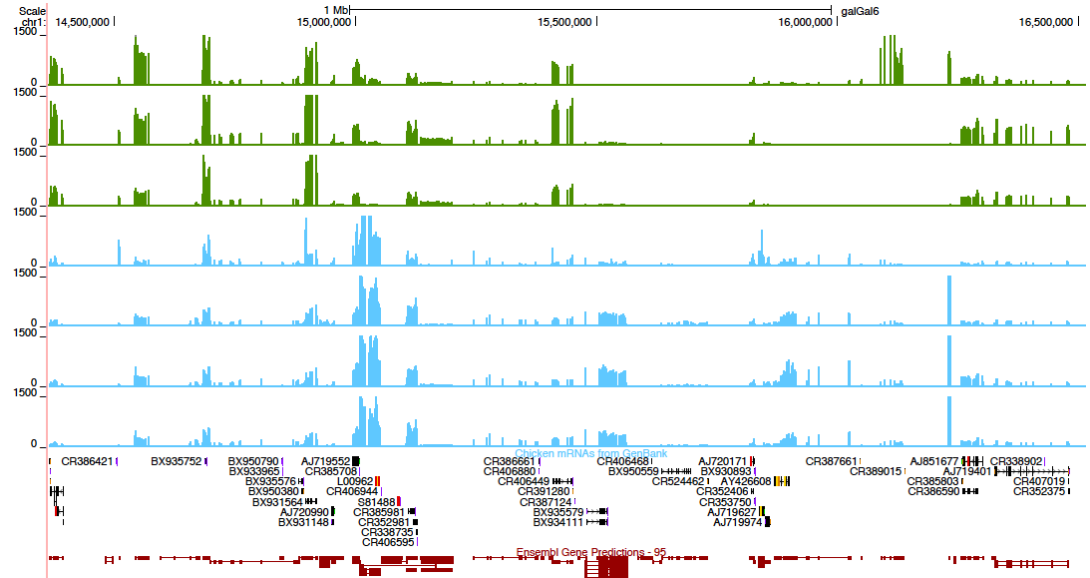 - Fell short of every single assay in every single cell/tissue.

Assays
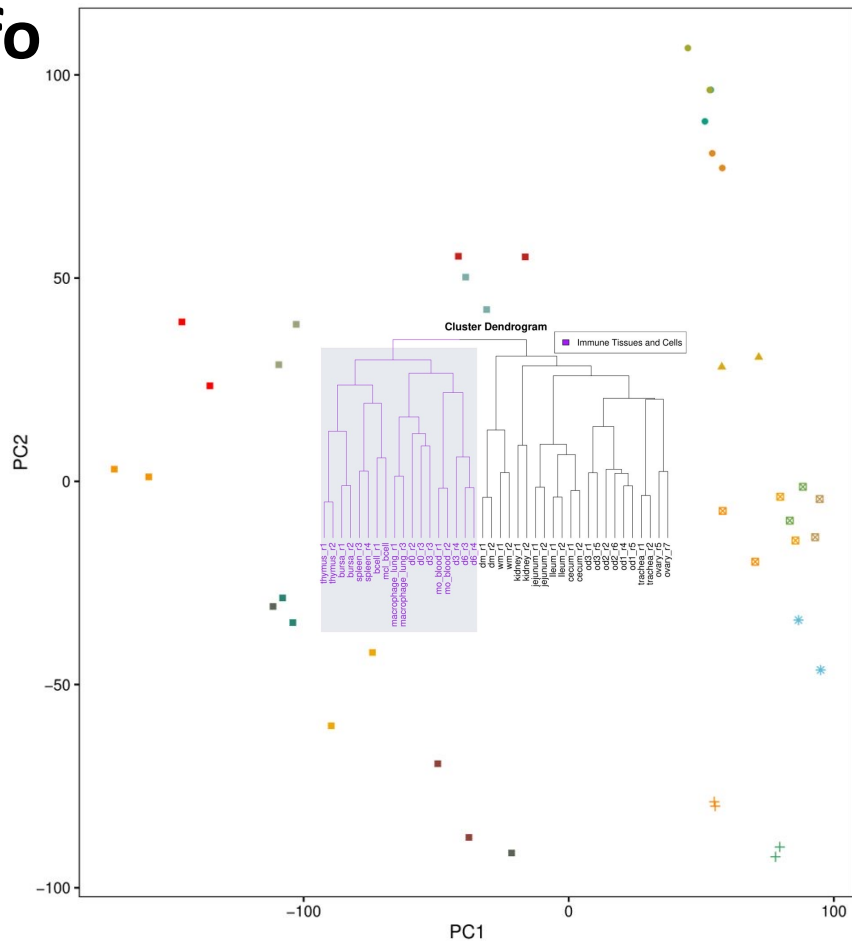
- RNA-seq
- ATAC-seq
- ChIP-seq/CUT&Tag
- WGBS

Data production line: ADOL line 6x7 F1.

Transcriptome

- ribo-minus RNA-seq
- PE150
- Targeted 50M read pairs
  - Median: 39.5M
- Quantitate RNA biotypes
- Identify isoforms
- Identify novel transcripts
- Identify novel genes
- Determine allele-specific expression (ASE)
- DEGs

Overbey et al., 2021; Oliveira de Lima, unpublished

RNA-seq: Sample Info

NCBI vs. Ensembl

NCBI

| 116596 Transcripts |
| 29088 Loci |

Missed exons: 0/265223 (0.0%)
Novel exons: 33257/337145 (9.9%)
Missed introns:191/228760 (0.1%)
Novel introns: 17454/264532 (6.6%)
Missed loci: 0/24223 (0.0%)
Novel loci:8658/29088   (29.8%)

Ensembl

| 119884 Transcripts |
| 34310 Loci |

Missed exons: 0/310680 (0.0%)
Novel exons: 42323/395311 (10.7%)
Missed introns: 344/254734 (0.1%)
Novel introns: 20325/300331 (6.8%)
Missed loci: 0/29507   (0.0%)
Novel loci: 10657/34310   (31.1%)

Transcripts/isoforms/novel genes

NCBI basic stats (bGalGal1.mat.broiler.GRCg7b)

- 116596 Transcripts
  - 95,189 protein-coding
  - 14,520 ncRNAs
  - 1,766 small RNAs
 - 5,121 novel genes at unannotated loci
- 29088 loci

73,772 annotated NCBI transcripts

5,121 novel genes

code

= c i j k m n o p s u x y
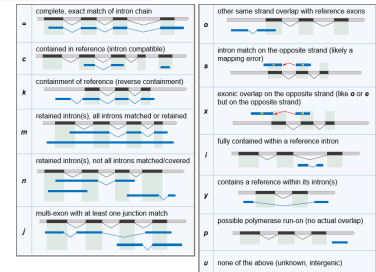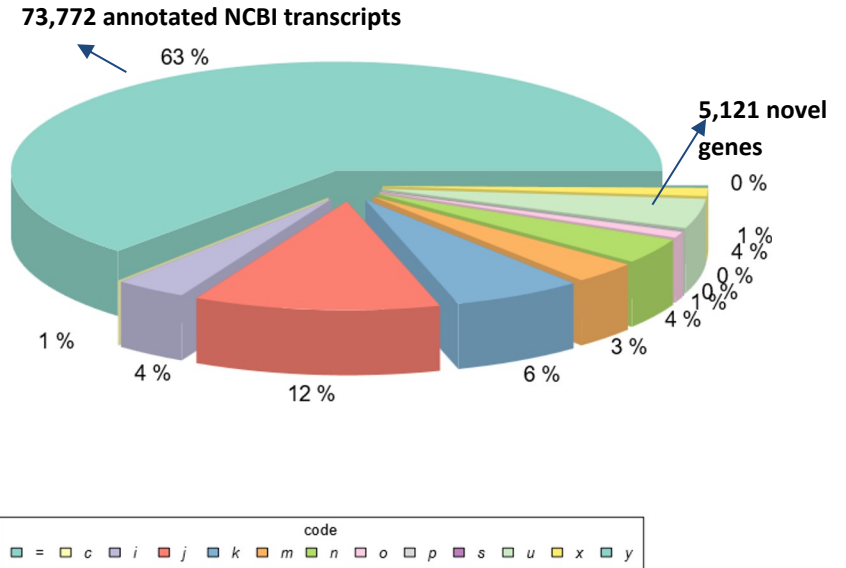
Gene expression

Total of transcript isoforms expressed by tissue/cell type

Tissues	Expressed Genes
Jejunum	44109
Monocyte (blood)	44615
Spleen T Cells	44645
B.muscle[2]	45710
Thymus	45950
Bursa	48341
B cell	49083
T.muscle[1]	47730
Proximal Cecum	50086
Magnum (od1)	49460
Macrophage (Kidney)	44252
Shell Gland (od3)	50685
Macrophage (lung)	52051
Isthmus (od2)	53050
Ileum	56351
Macrophage (spleen)	58605
Ovary	65539
Macrophage (d3)[4]	65018
Macrophage (d0)[3]	68666
Macrophage (d6)[5]	73554

[1]Iliotibialis major; [2]Pectoralis major; [3]macrophage differentiation
[4]macrophage differentiation at day 3; [5]macrophage differentiati
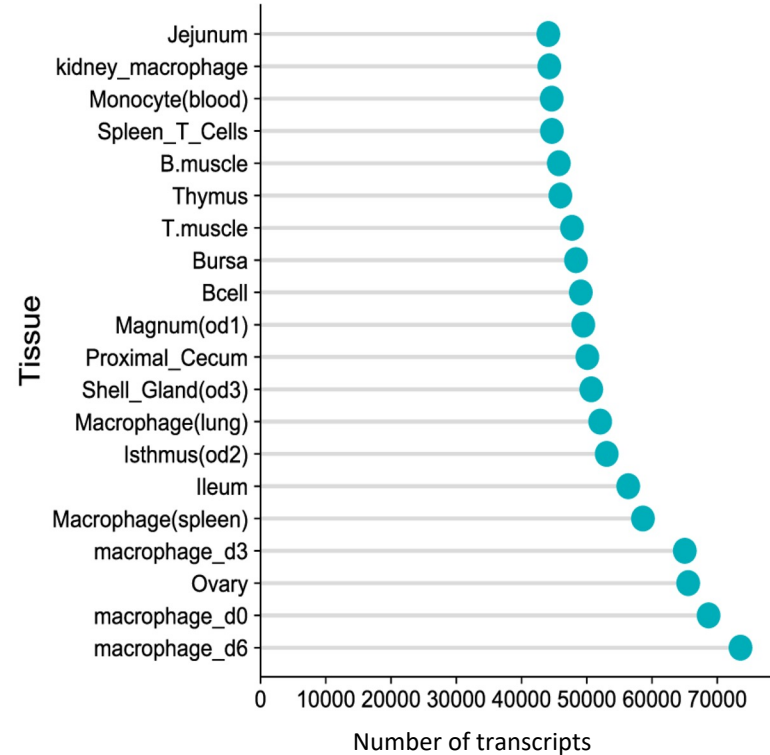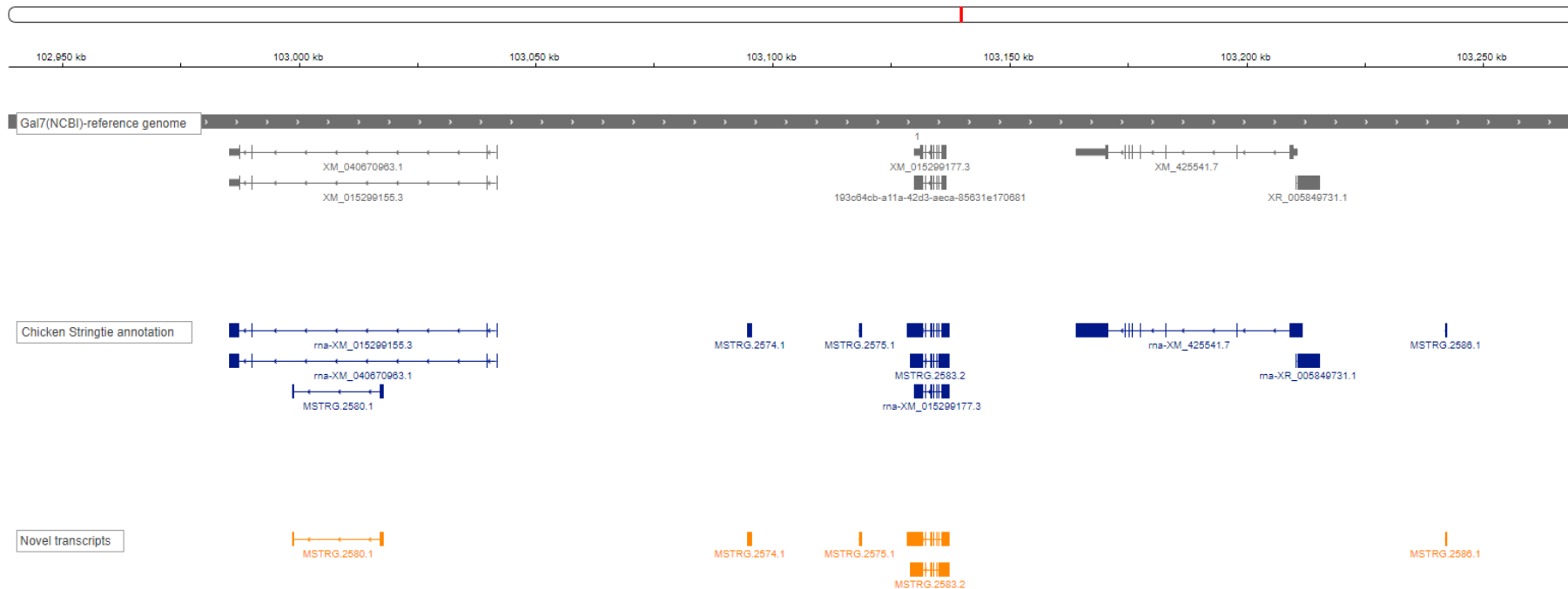
Novel Isoforms
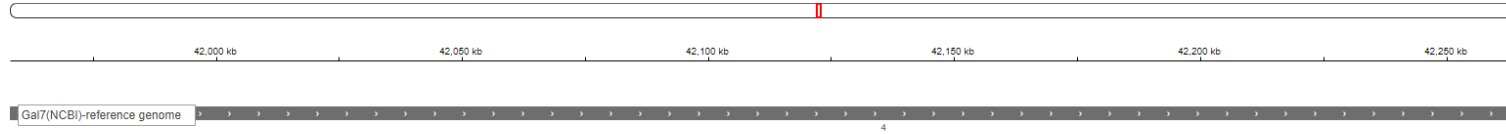
Novel Genes

Transcriptome

- Currently mapped to NCBI & Ensembl transcriptomes.

- StringTie for transcriptome assembly to find novel isoforms and new genes.
  - 30% of transcripts are novel isoforms of known genes
  - 4% of transcripts are novel genes
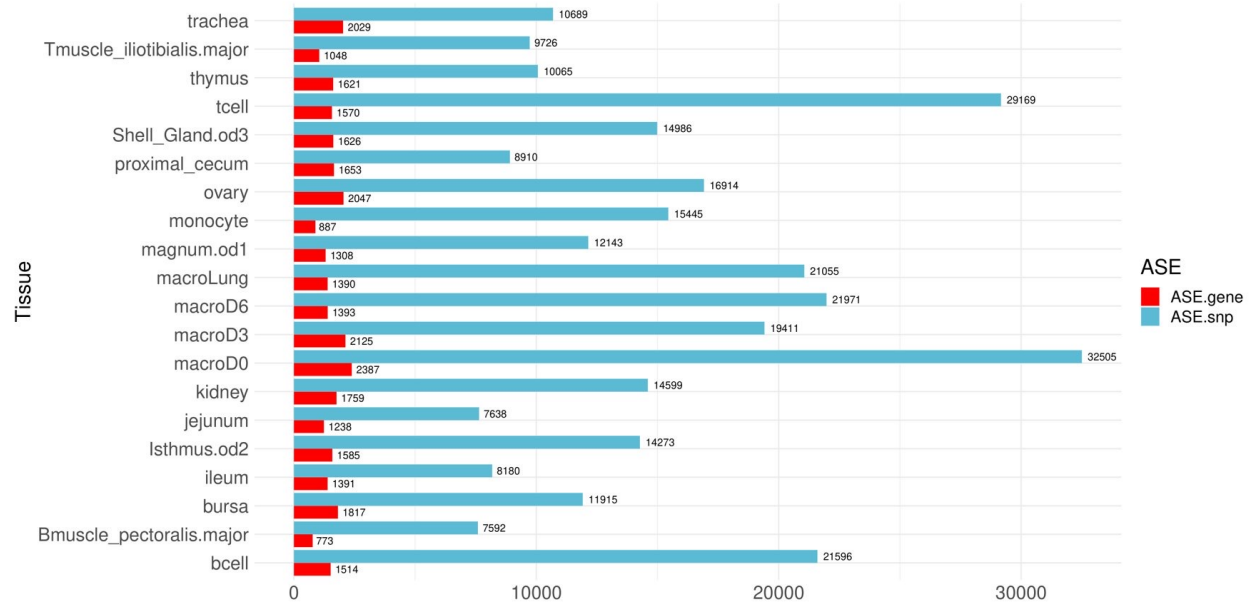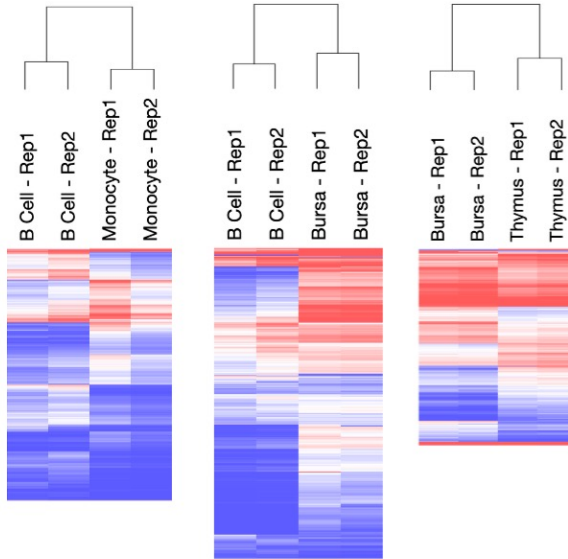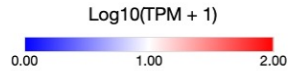   - 22% predicted as mRNAs
    - 78% predicted lncRNAs
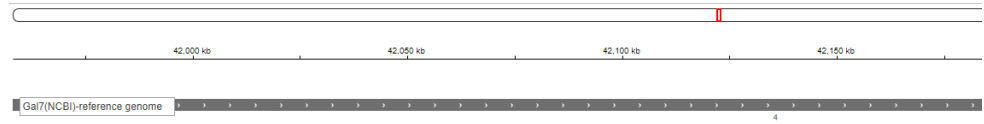
Allele-specific expression

- Allele-specific expression, ASE, is the unequal expression between the two alleles.
  - One allele is not necessarily "unexpressed"

- Can be influenced by genetic or epigenetic factors.
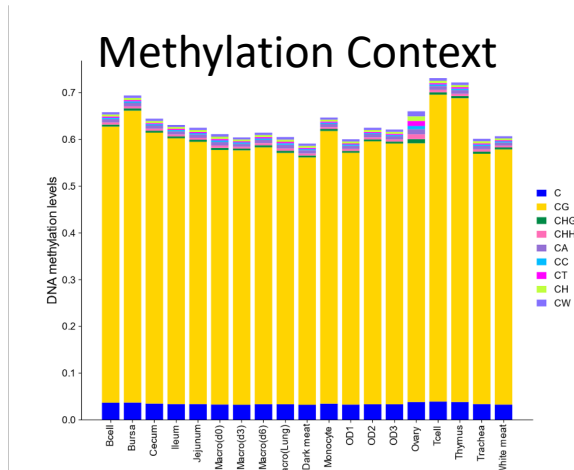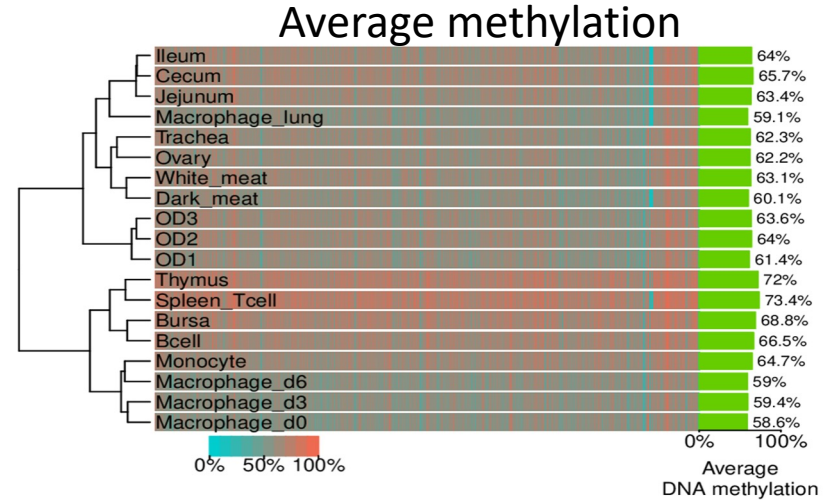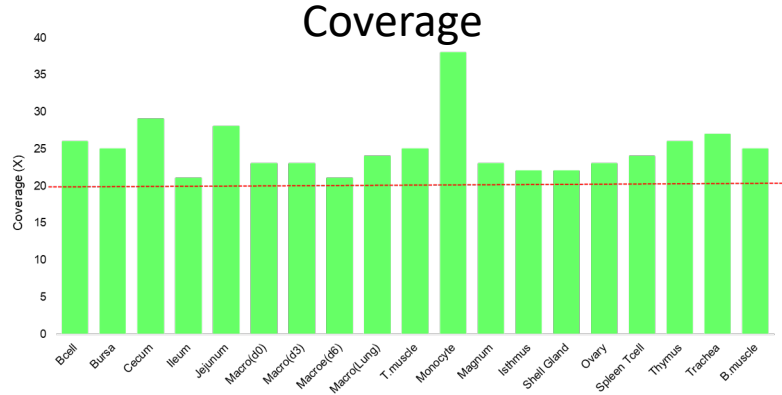 - regulatory variants
  - allelic DNA methylation
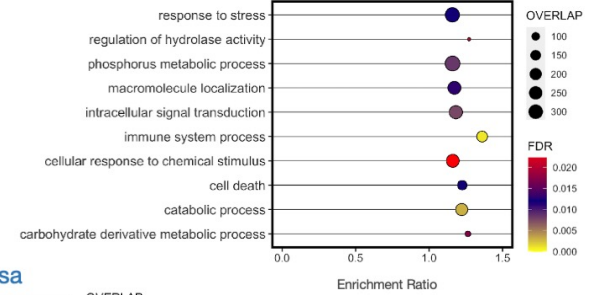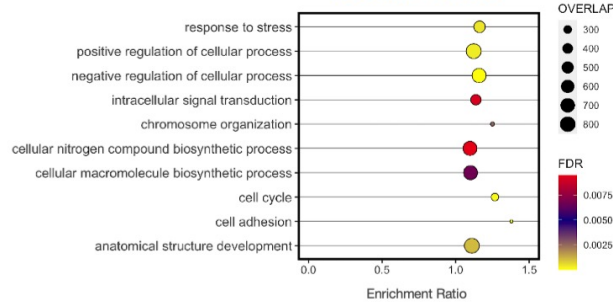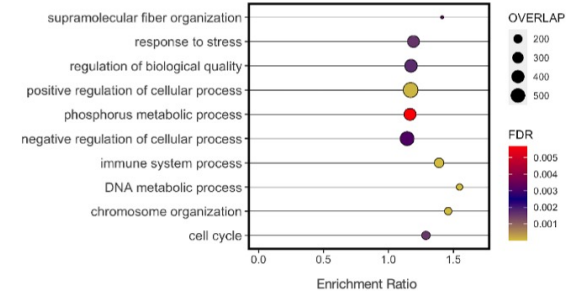
WGBS

- Targeted 20X coverage.
- Differential methylation analysis for key comparisons.
- Determine allele-specific methylated regions.
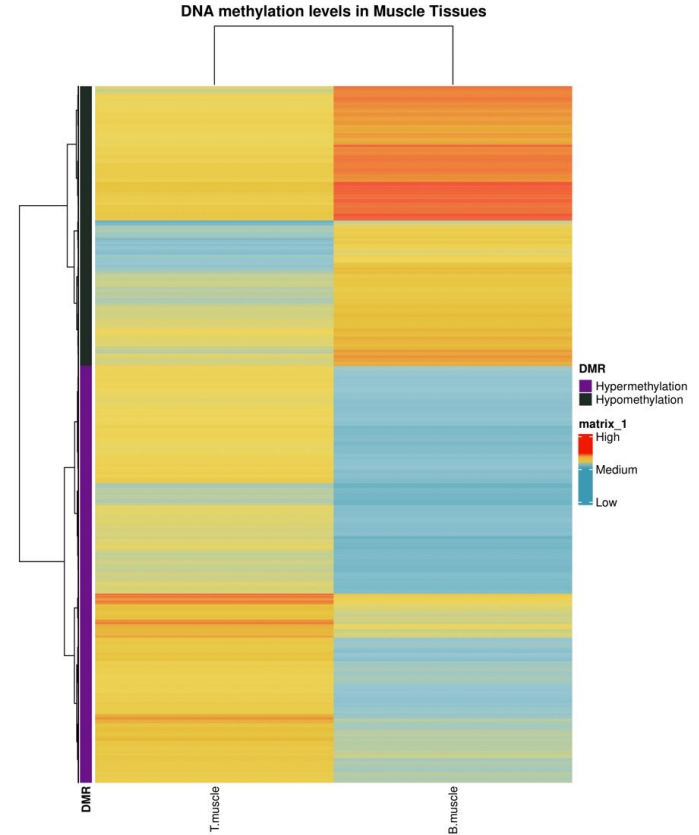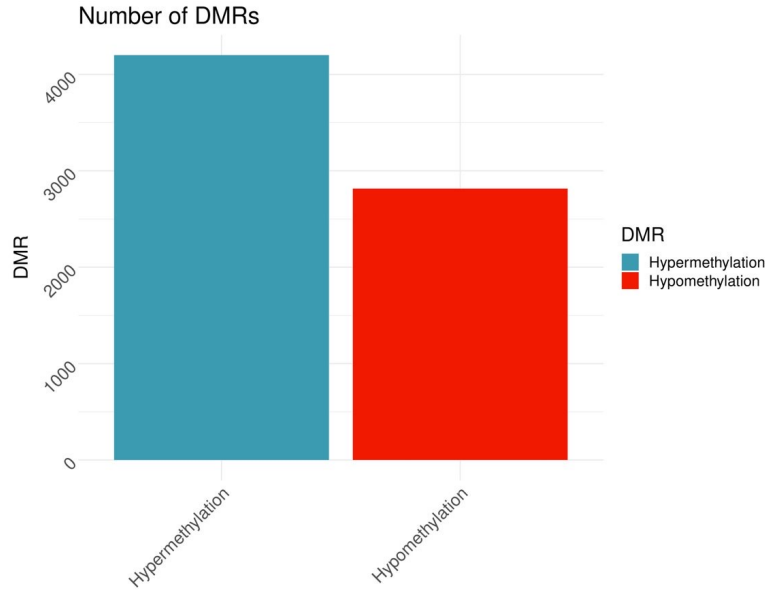- Integration w/ other data ongoing.

DNA methylation (WGBS)

Coverage

Methylation Context

Average methylation
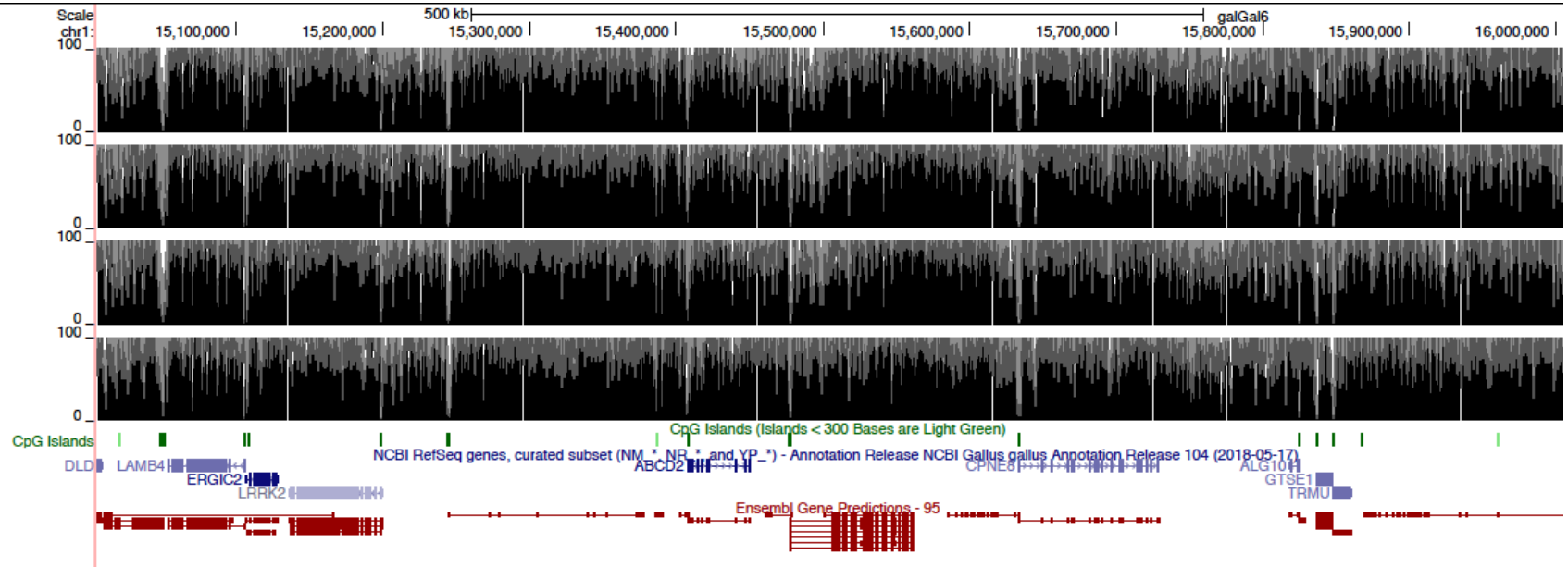
Bisulfite conversion: >99%

Mapping efficiency: ~80.0%

Differentially Methylated Regions (DMRs)

Breast vs Thigh Muscle

DMRs with pvalue <= 0.05 and Δmc > = 0.2
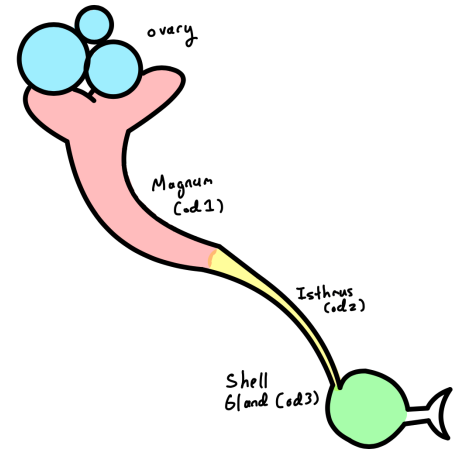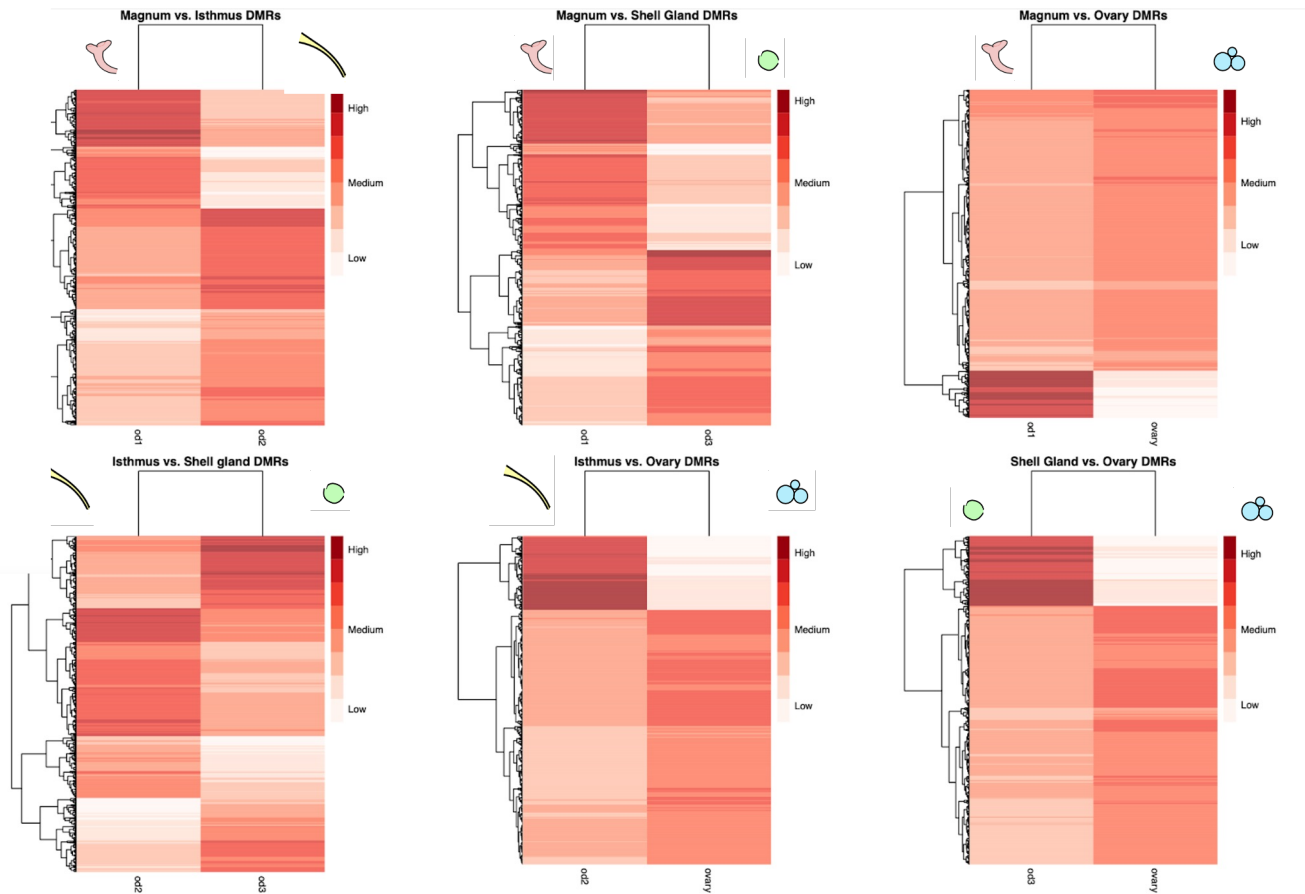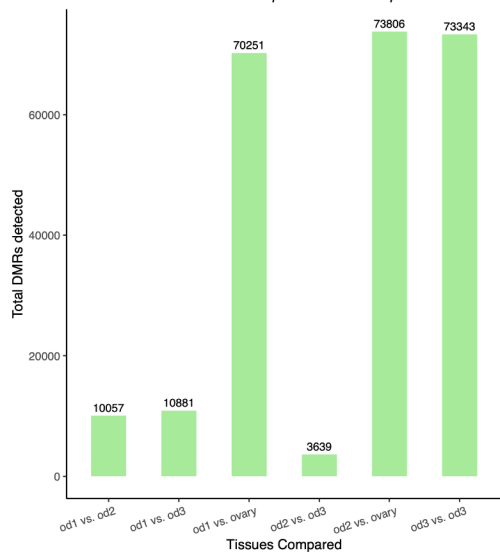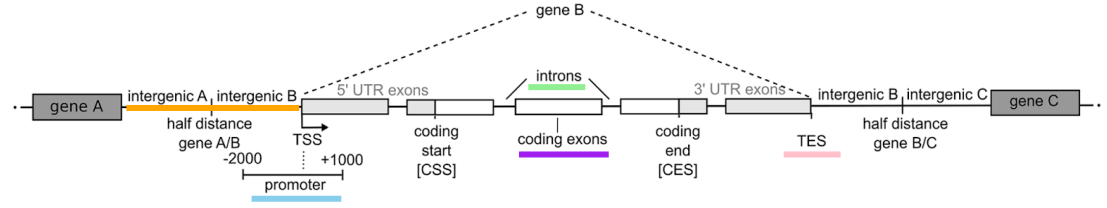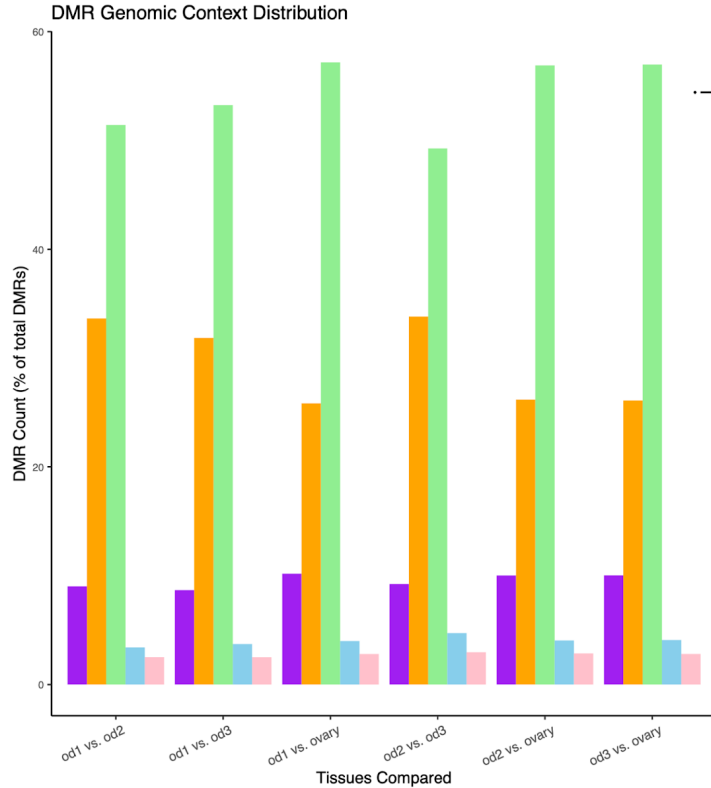
DMRs of the reproductive tract
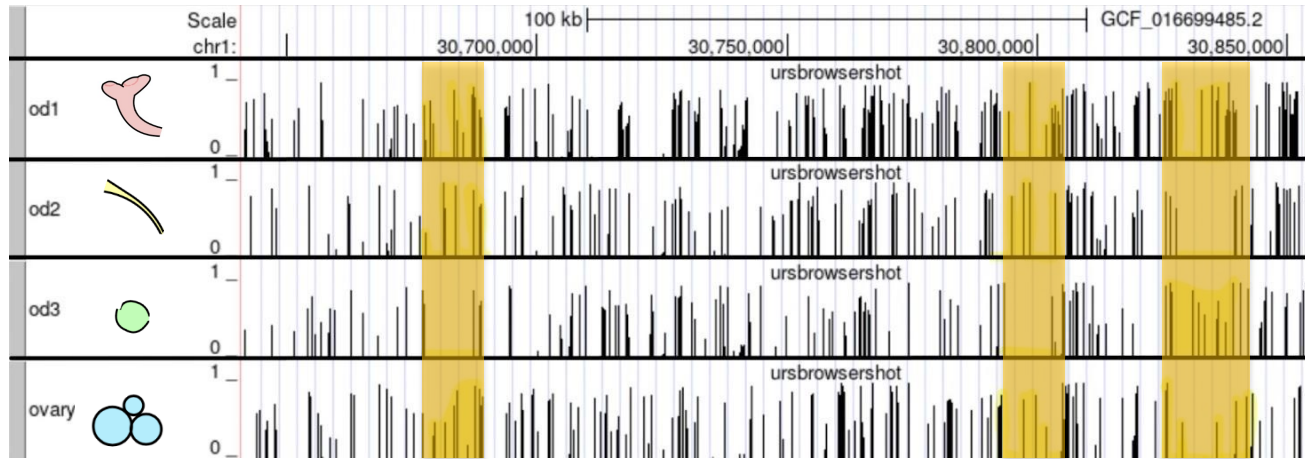
DMRs of the reproductive tract
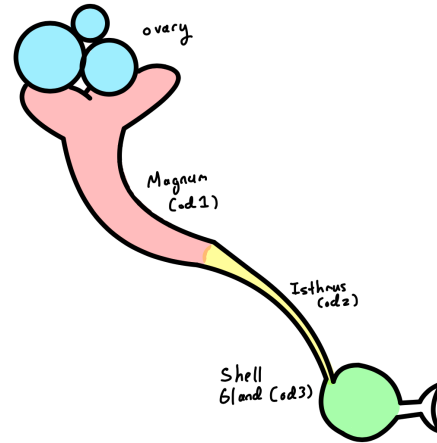

Total DMRs detected per Pairwise Comparison

- # of DMRs = a proxy for dissimilarity b/w tissues

- Ovary had the highest DMR count = most different (in function / structure) from other tissues
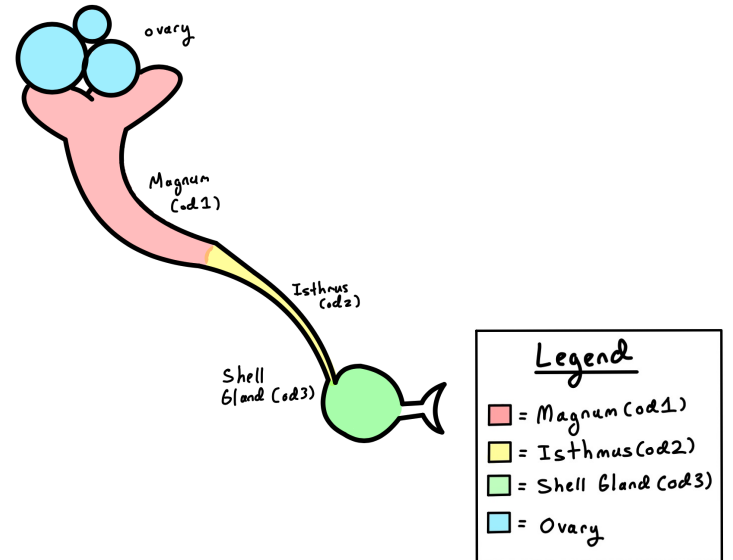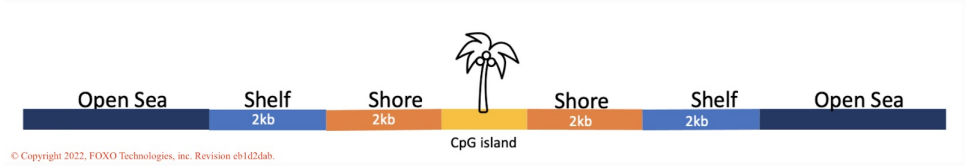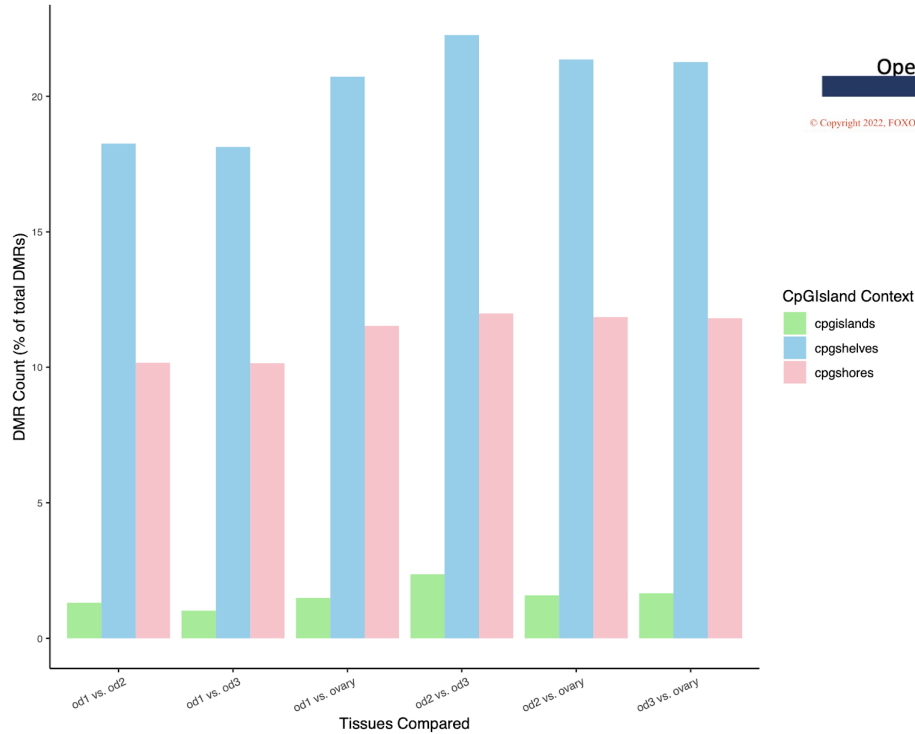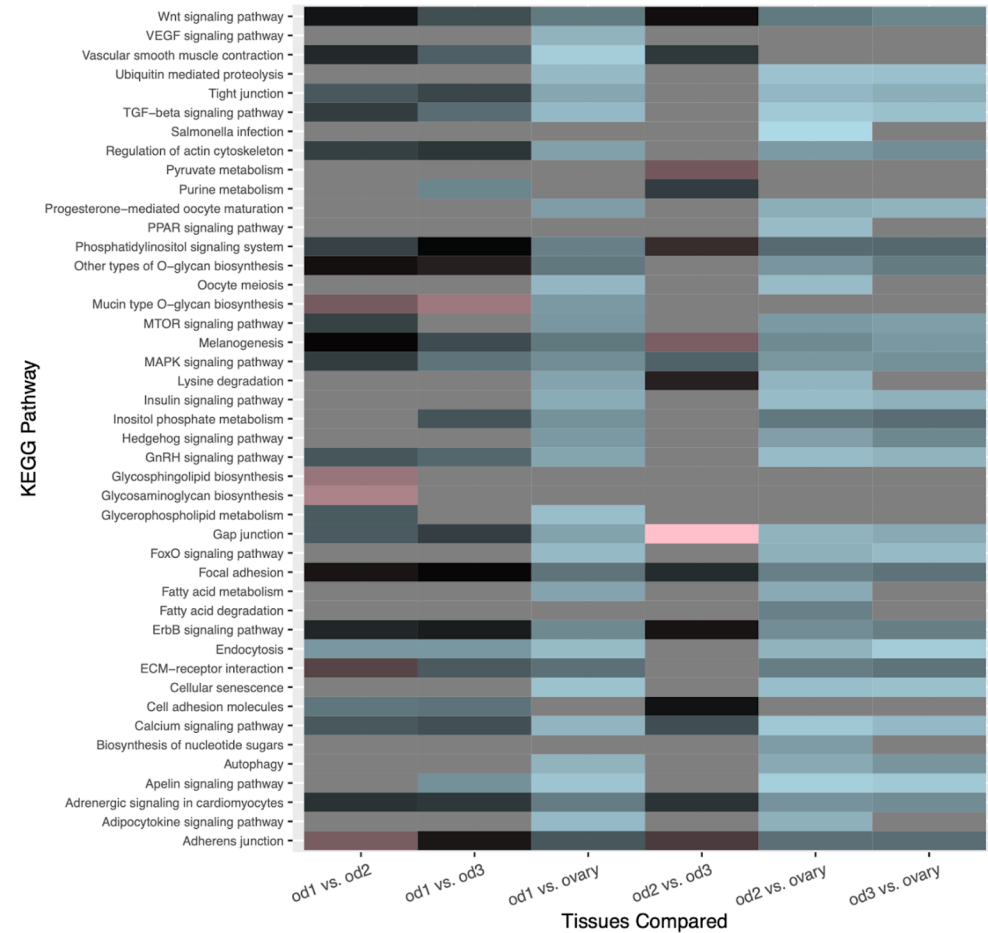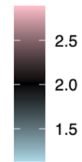
DMR Genomic Context
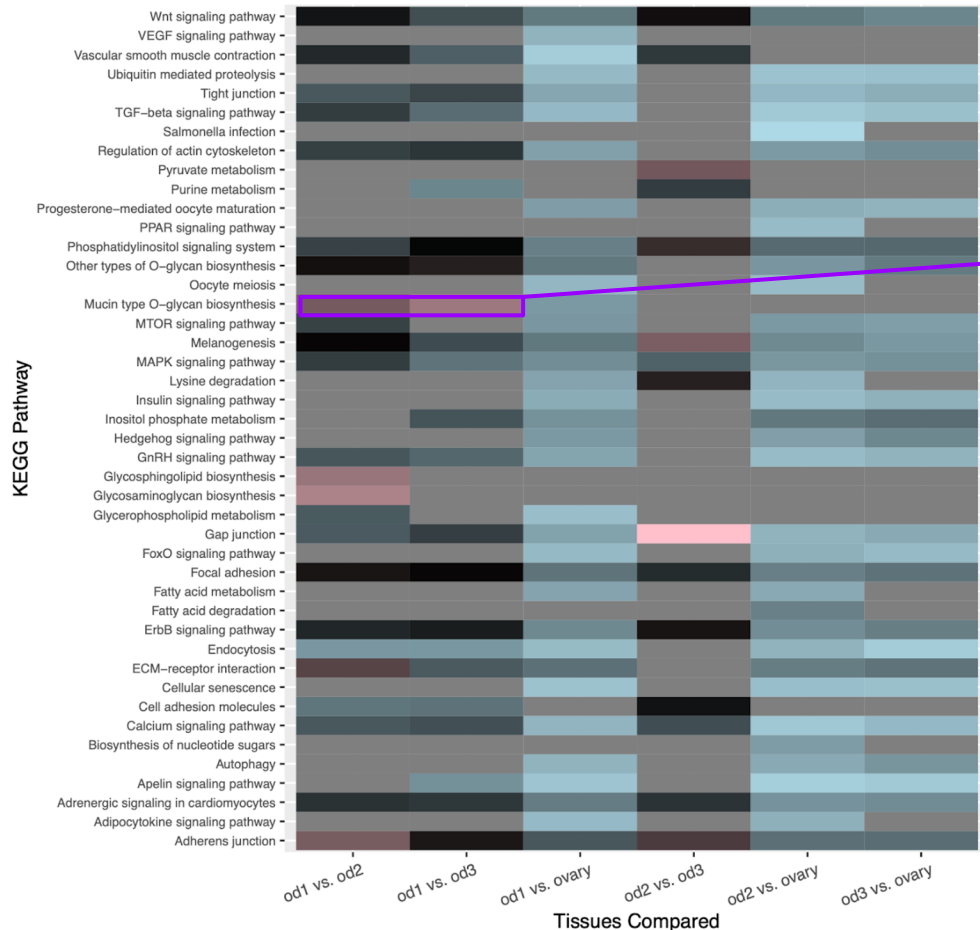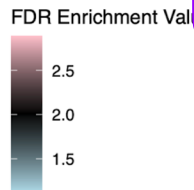
DMR CpG Island Context

DMR KEGG Pathways (FDR Enrichment-Based)

Enrichment analysis

DMR KEGG Pathways (FDR Enrichment-Based)

Enrichment analysis

Enrichment Analysis: Mucin type O-glycan Biosynthesis Pathway

When compared to od2 & od3, od1 DMRs were highly enriched in Mucin type O-glycan Biosynth. genes.

ovary

Magnum (od1)

Isthmus (od2)

Shell Gland (od3)

thick white

thin white

chalazae

germinal disc

yolk

LM

membrane

egg shell

cuticula

air cell

Jalili-Firoozinezhad, Sasan, et al. "Chicken Egg White: Hatching of a New Old Biomaterial." Science Direct, 11 Aug. 2020, www.sciencedirect.com/science/article/pii/S1369702120302170.

Lysozyme

Ovomucoid

Ovalbumin

Conalbumin

Ovomucin

Mucin type O-glycan Biosynthesis

Glycosylated Ovomucin

Provides gel-forming & emulsifying properties

Viscosity prevents bacterial penetration

Antiviral activities [bovine rotavirus, hen newcastle disease virus, influenza virus]

Antimicrobial (antiadhesive) [E.coli O157:H7]

Anti-cancer activities (mech. needs to be further studied)
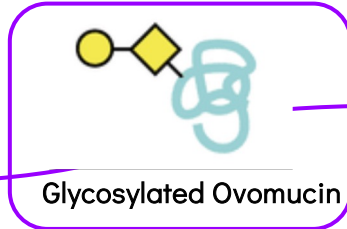
Appropriate epigenetic regulation (methylation)

Quality eggs / offspring
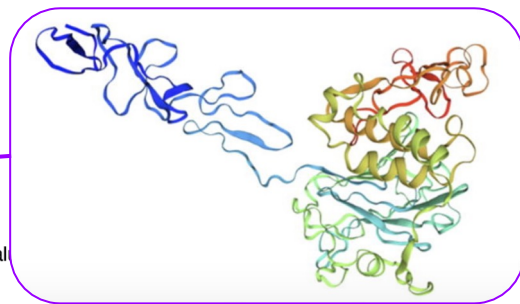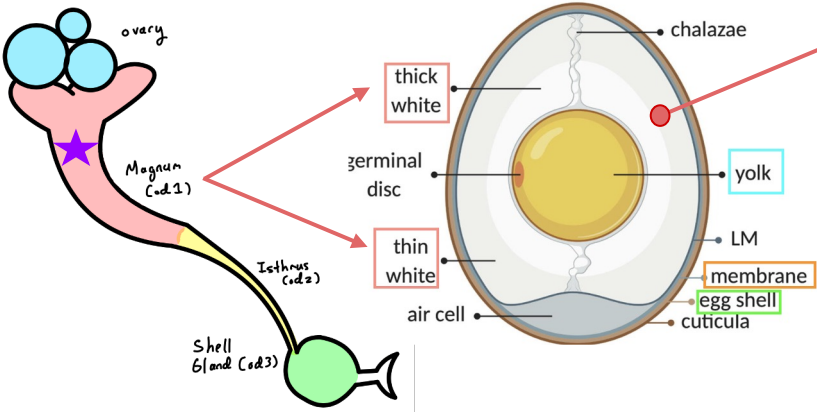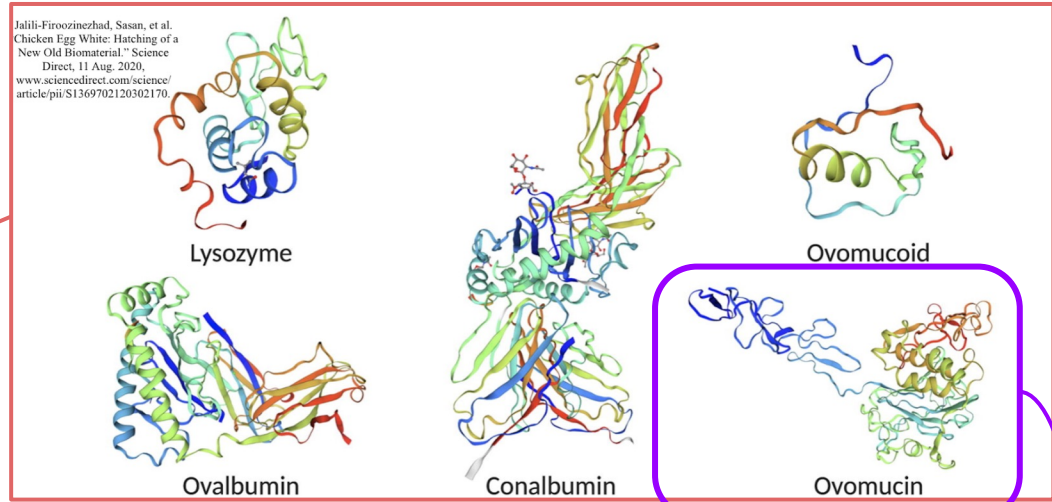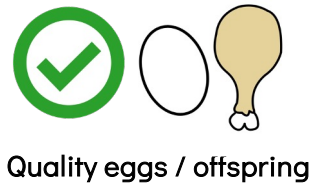
DMR KEGG Pathways (FDR Enrichment-Based)
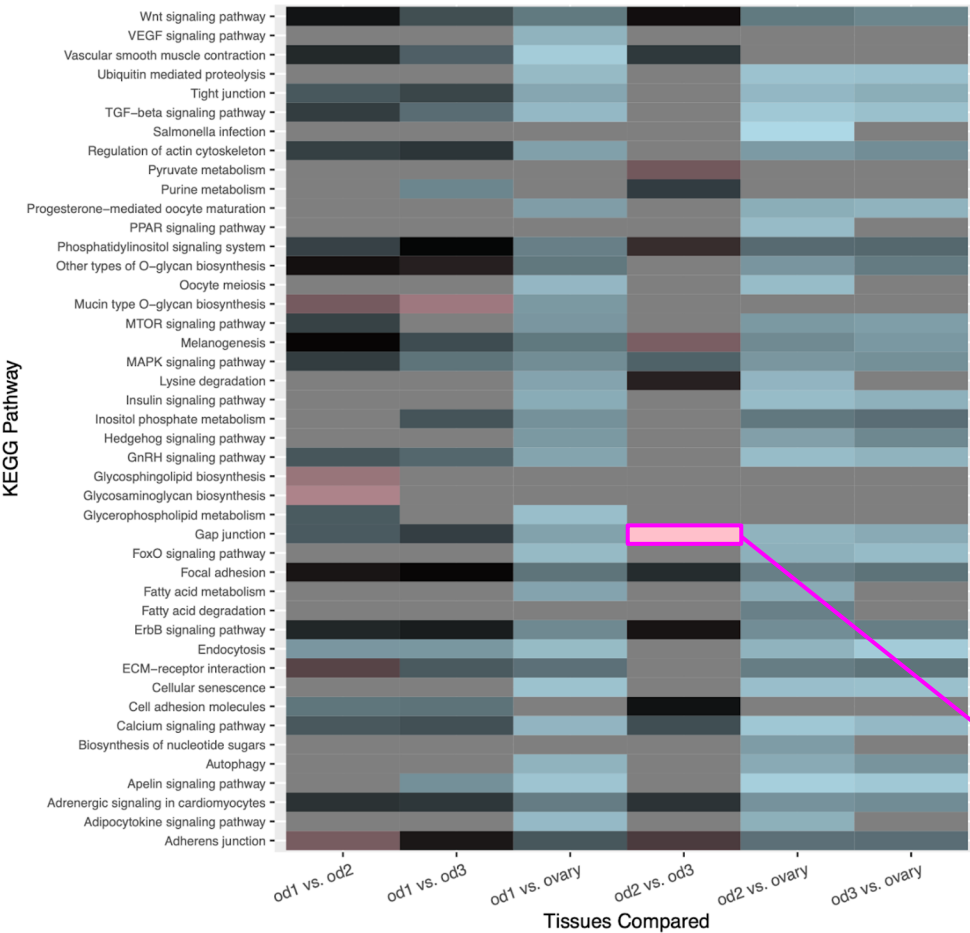
Enrichment analysis

FDR Enrichment Value

Figure 8-13b part 2 Biological Science, 2/e   © 2005 Pearson Pren

Gap junctions

Membrane proteins from adjacent cells line up to form a channel

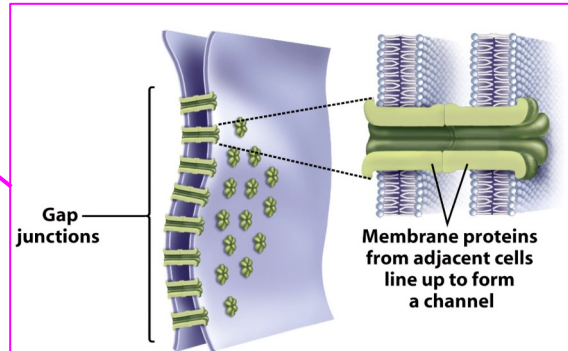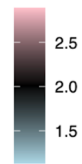# Enrichment Analysis: Gap Junction Pathway

Od2 vs. od3 DMRs were highly enriched in Gap Junction genes

Appropriate epigenetic regulation (methylation)

Quality eggs / offspring

Gap junctions create gaps that connect animal cells.

Gap junctions

Membrane proteins from adjacent cells line up to form a channel

Figure 8-13b part 2 Biological Science, 2/e

© 2005 Pearson Prentice Hall, Inc.

ovary

Magnum (od1)

Isthmus (od2)

Shell Gland (od3)

chalazae

thick white

germinal disc

yolk

thin white

LM

air cell

membrane

egg shell

cuticula
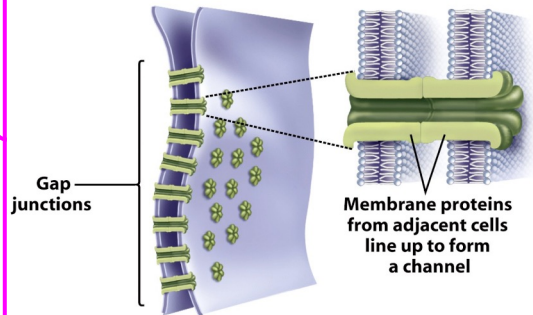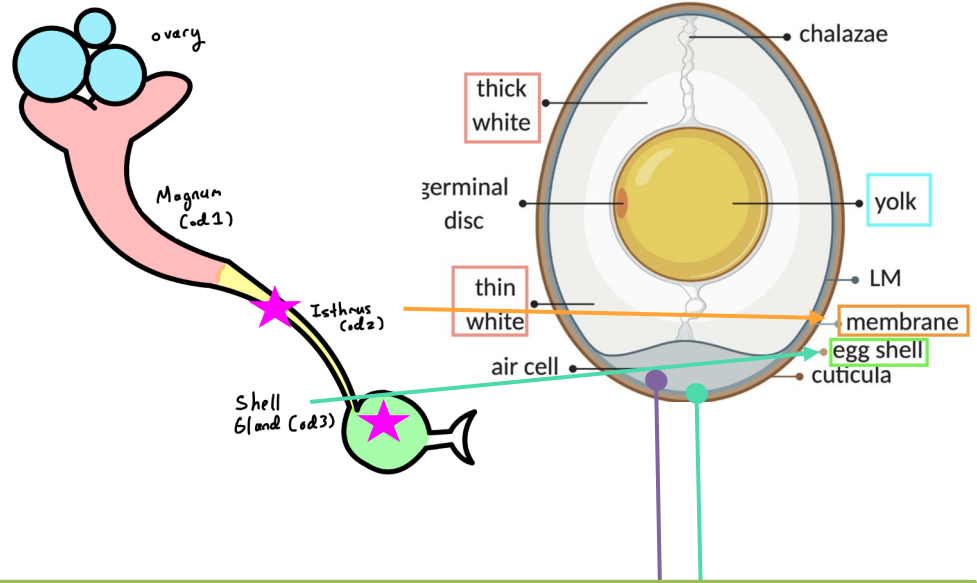
Egg shell / shell membrane formation requires intercellular communication in order to coordinate and properly time CaCO3 deposition + specialized protein secretion

Gap Junction

# Differentially Methylated Regions (DMRs)

Key pairwise comparisons

P-value <= 0.05
Δmc > = 0.2

Allele-specific methylation

- Allele-specific methylated regions (ASM/ASR) exhibit asymmetry in DNA methylation levels.

- Can result in ASE.

- Can be associated with trait or disease loci.
  - Largely driven by genetic variation.

ASM Distribution by Chromosome
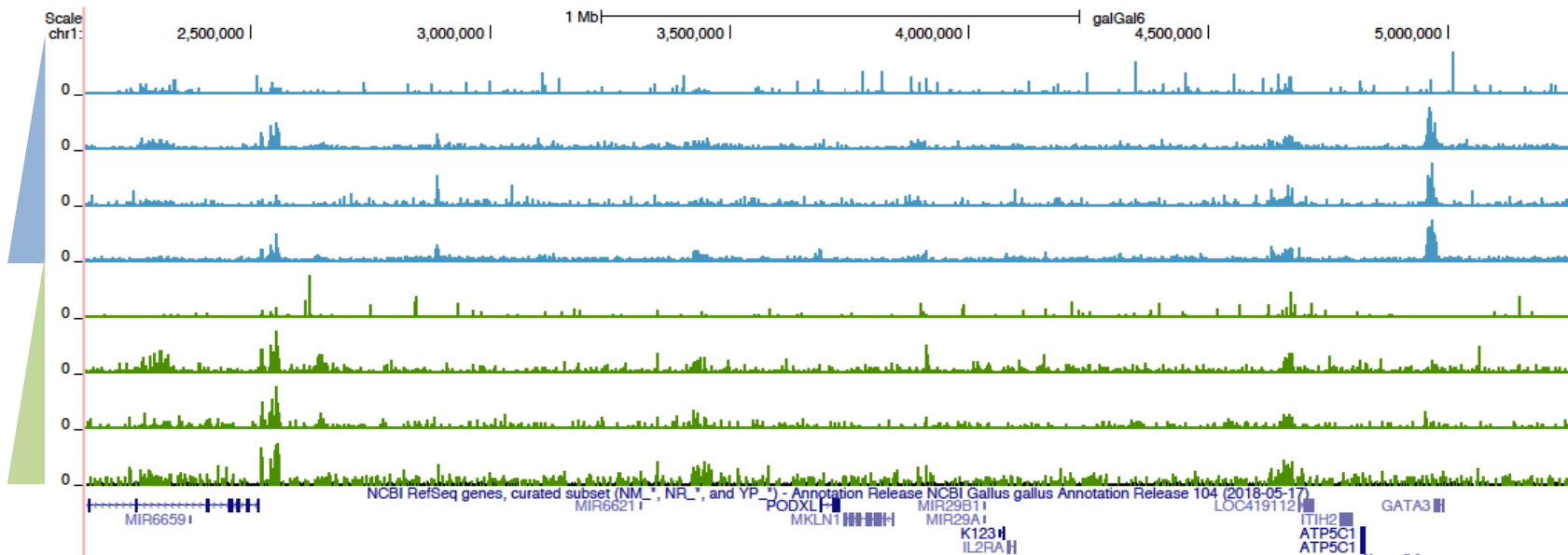

ASM distribution by Tissue
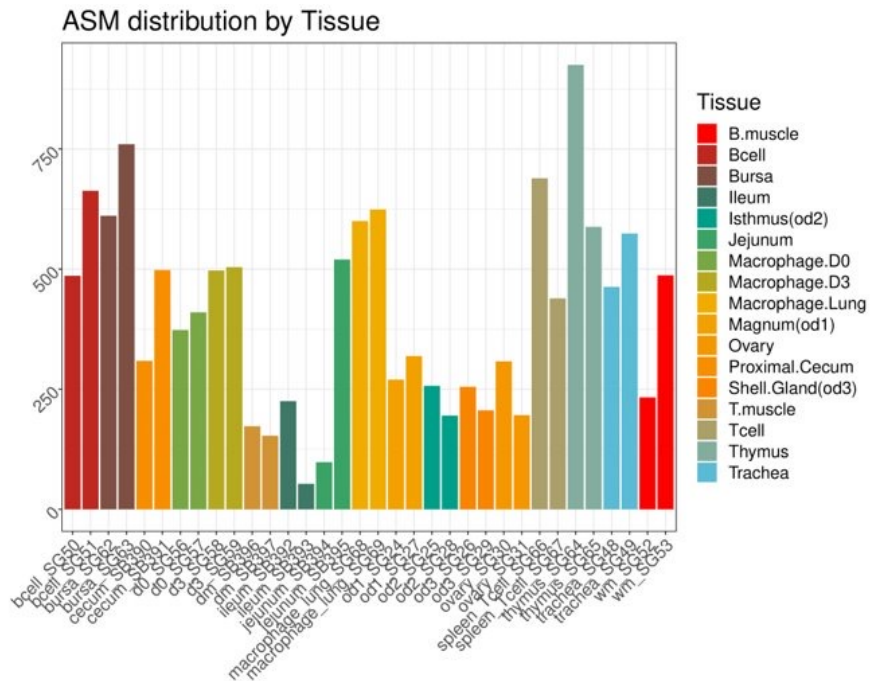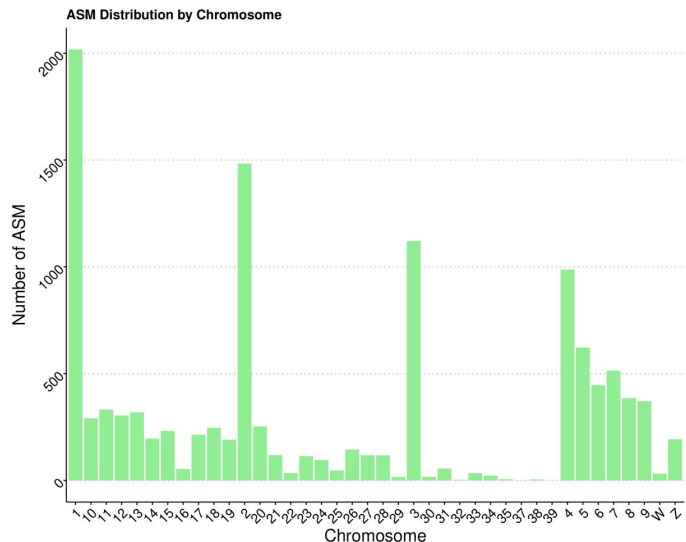
Epigenomic annotation of *cis*-regulatory elements
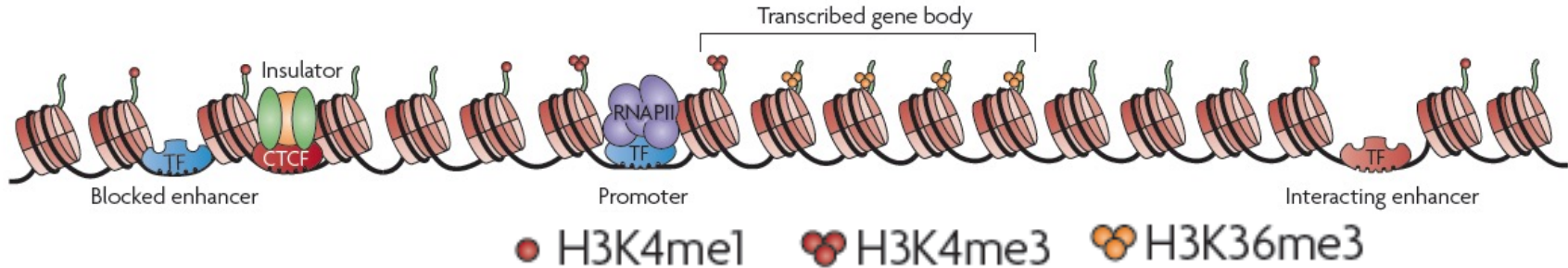
- H3K4me3:   novel & alternative promoters
- H3K4me1:   enhancer elements
- CTCF: insulators, architectural elements
- H3K27ac: indicative of active promoters/enhancers
- H3K27me3:   repressed, bivalent promoters
- H3K9me3: heterochromatin
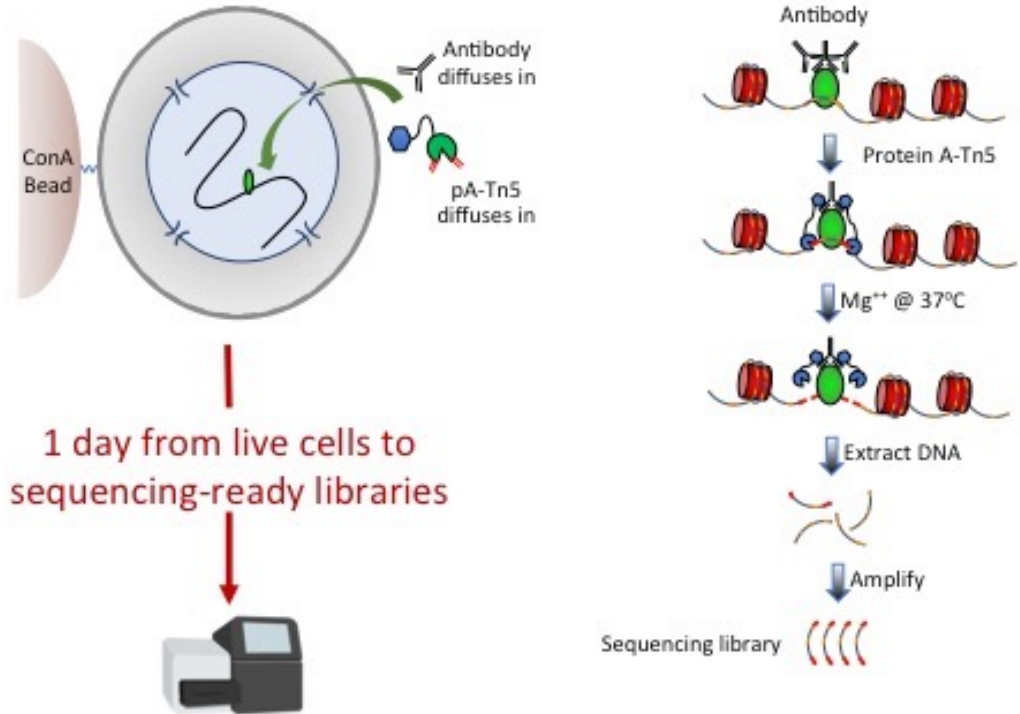
ChIP-seq (re)optimization

ChIP-seq

H3K4me3
H3K4me1
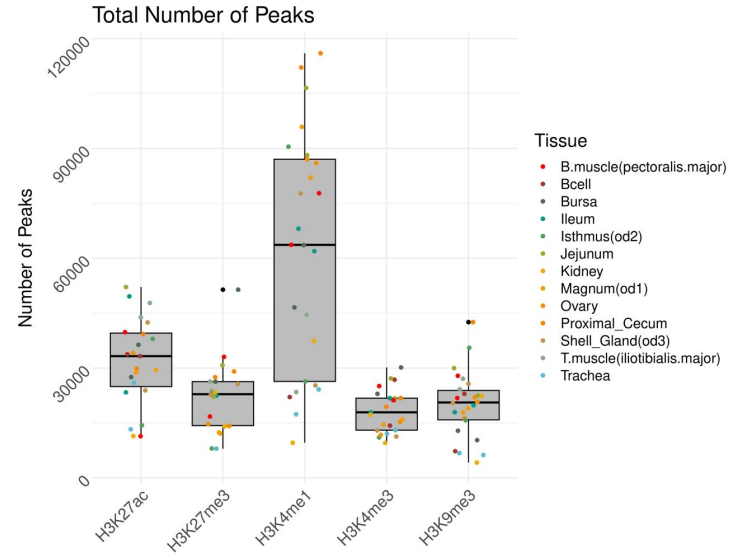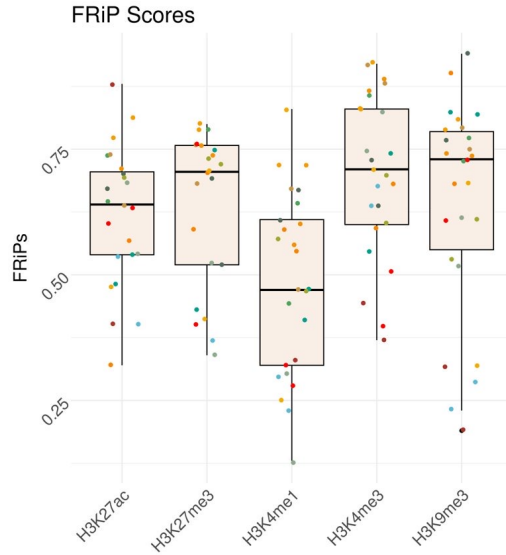H3K27ac
H3K27me3
H3K9me3
CTCF

ATAC-seq

Optimized tissue dissociation
Nuclei isolation

CUT&Tag (Cleavage Under Targets & Tagmentation)

Kaya-Okur, 2019

CUT&Tag: mapping histone marks

- Median reads per replicate: 6.3 M
- Median alignment rate per replicate : 93.4%
- Median peaks per replicate: 28000
- Median FRiPs score: 0.7

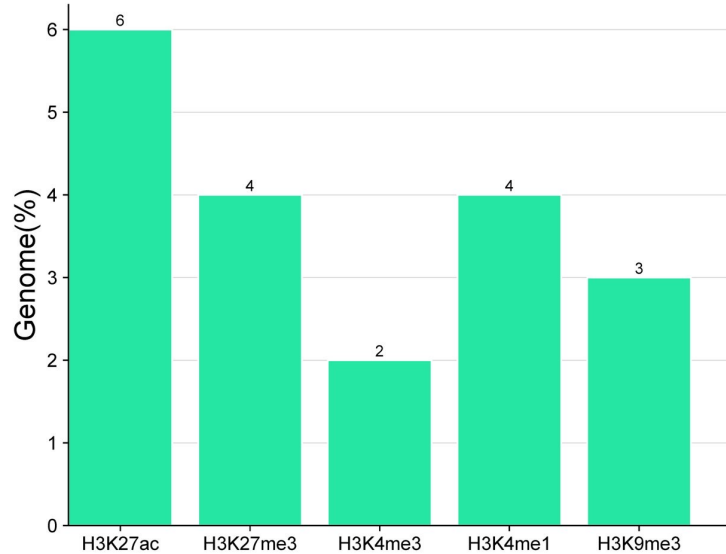
FRiP Scores

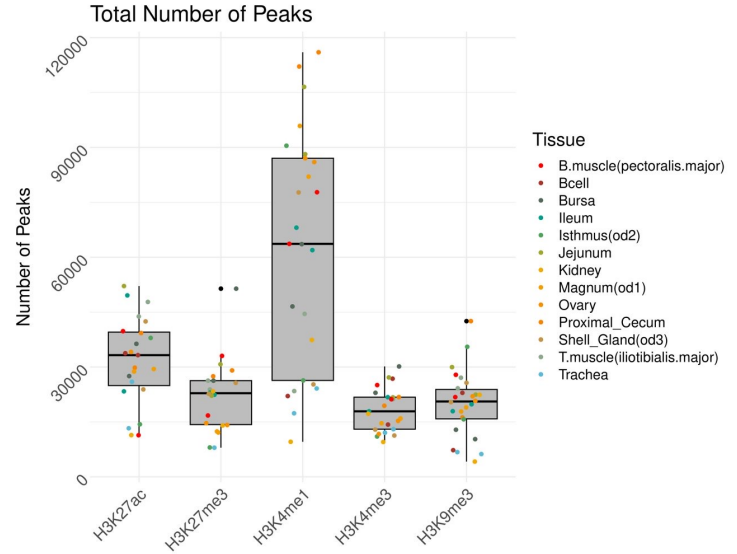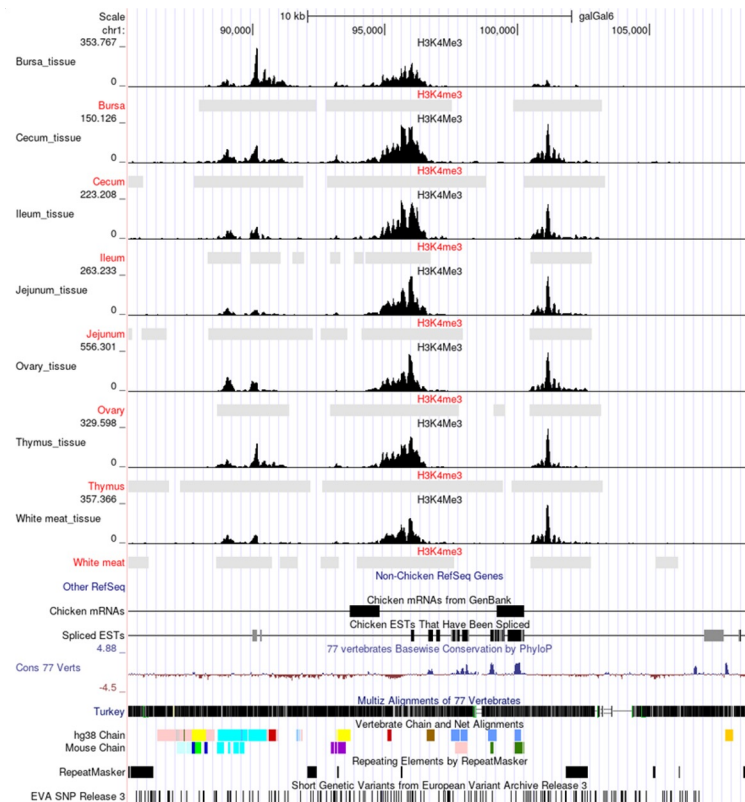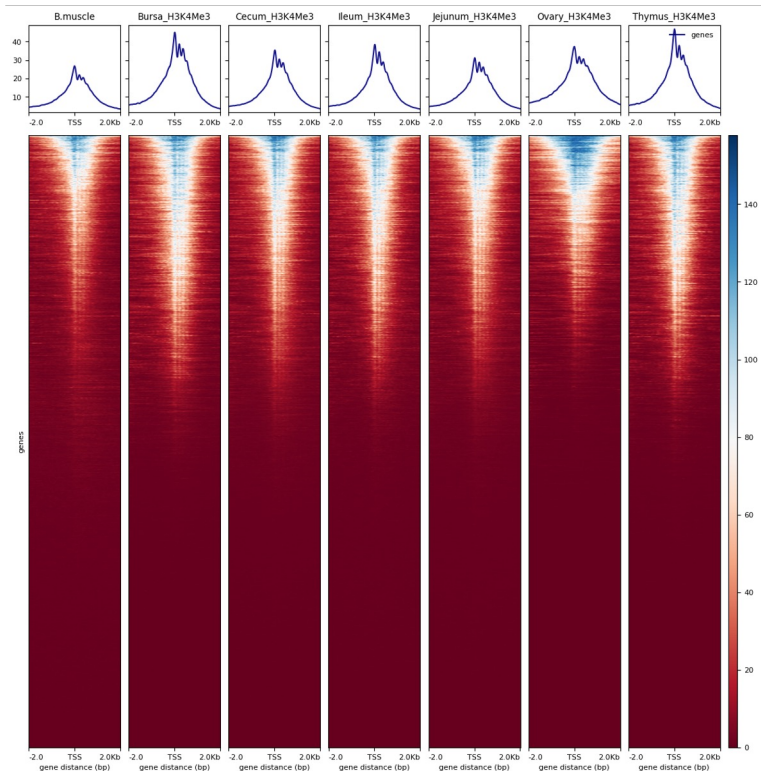Total Number of Peaks

CUT&Tag: mapping histone marks

- Median reads per replicate: 6.3 M
- Median alignment rate per replicate : 93.4%
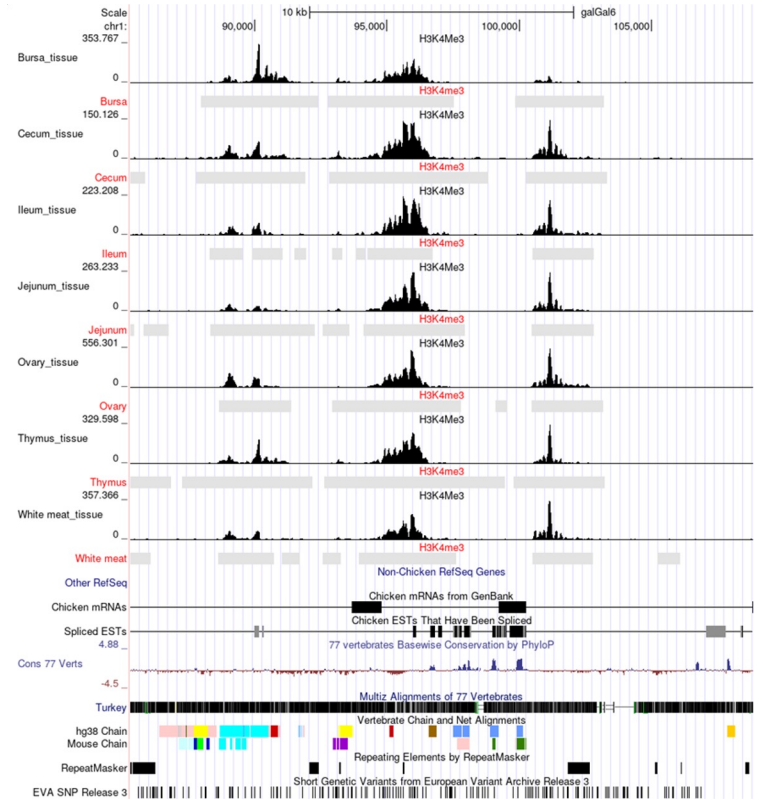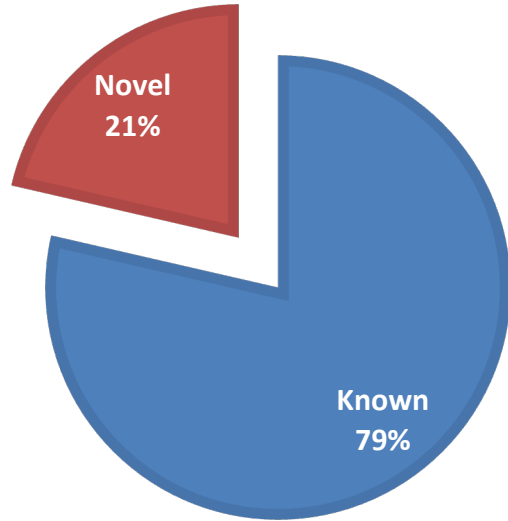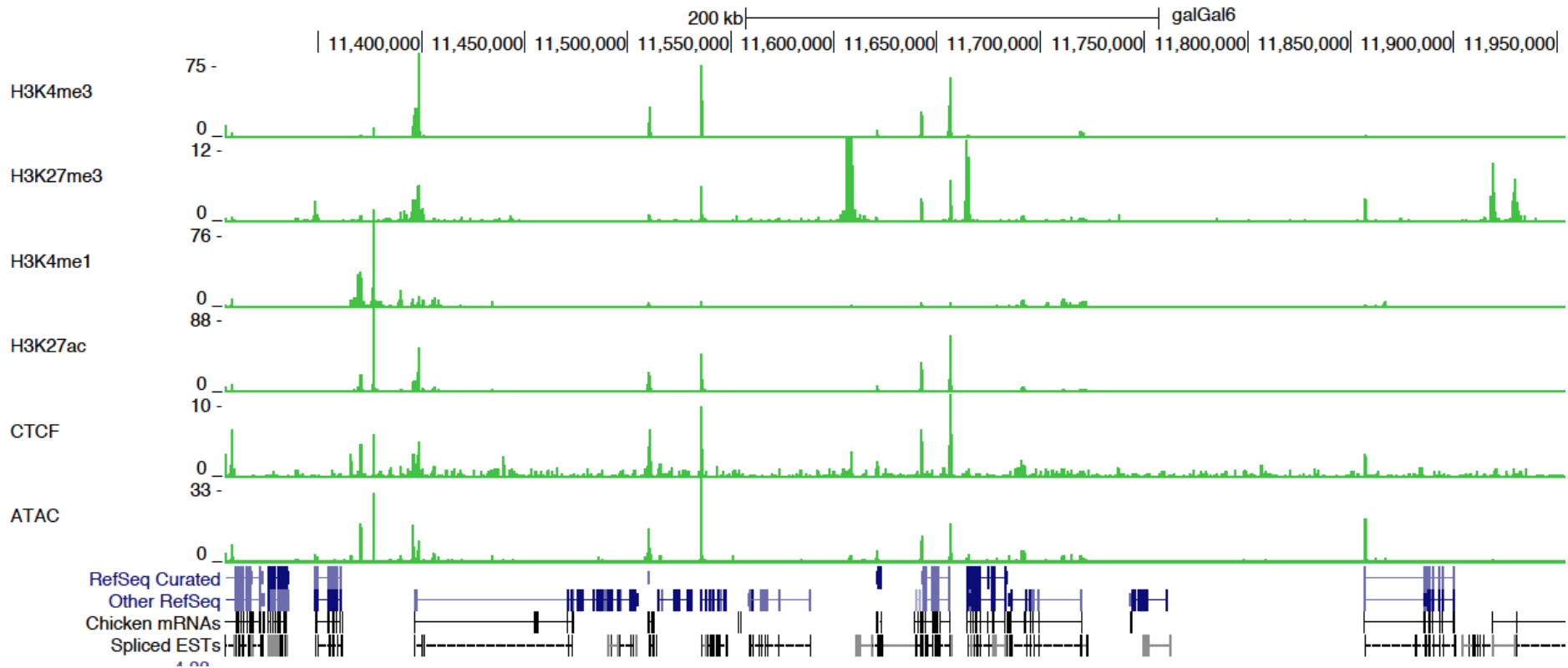- Median peaks per replicate: 28000
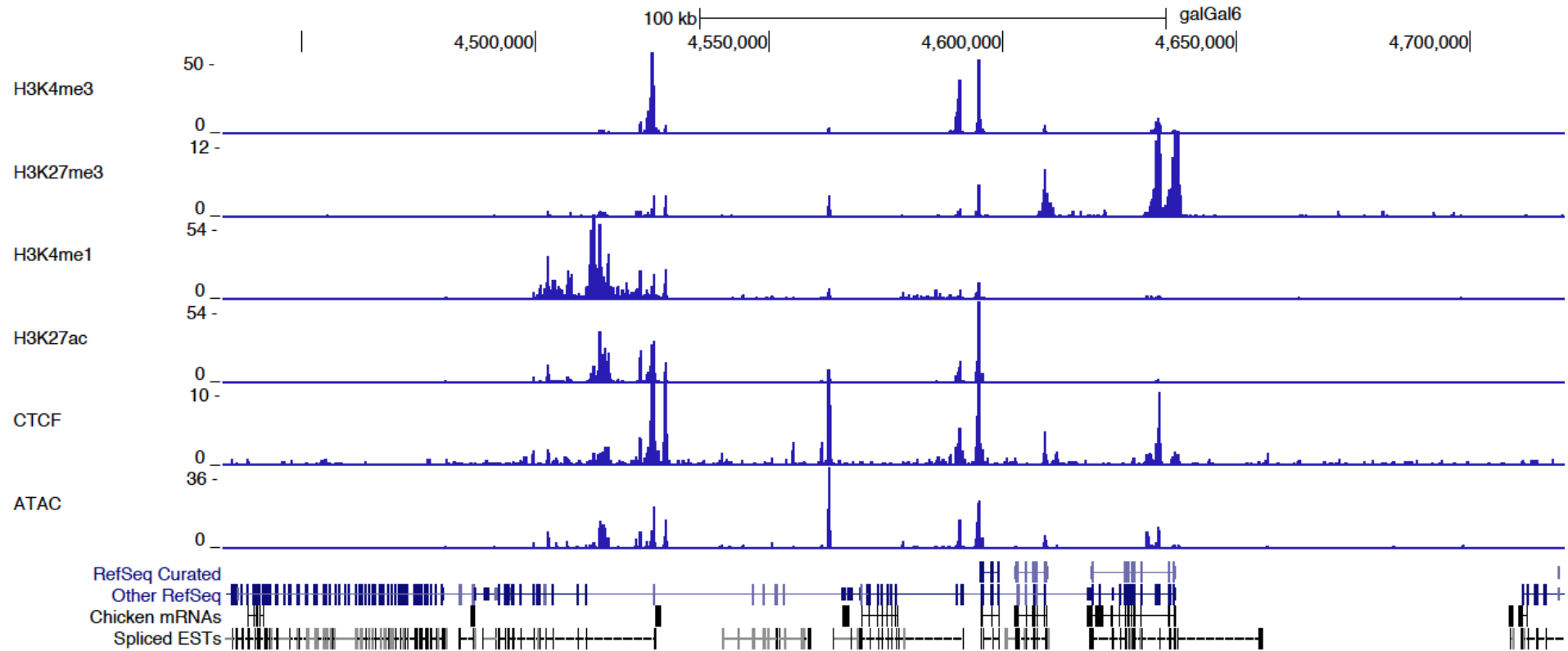- Median FRiPs score: 0.7

CUT&Tag-H3K4me3

cCRE chromatin states

- Categorize cCREs for each cell/tissue type.
  - Define active vs poised regulatory elements

cCRE chromatin states

chromHMM annotations

- Categorize cCREs for each cell/tissue type.
  - Define active vs poised regulatory elements

Bursa

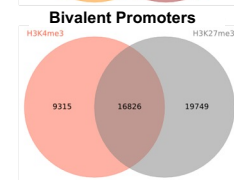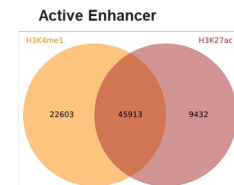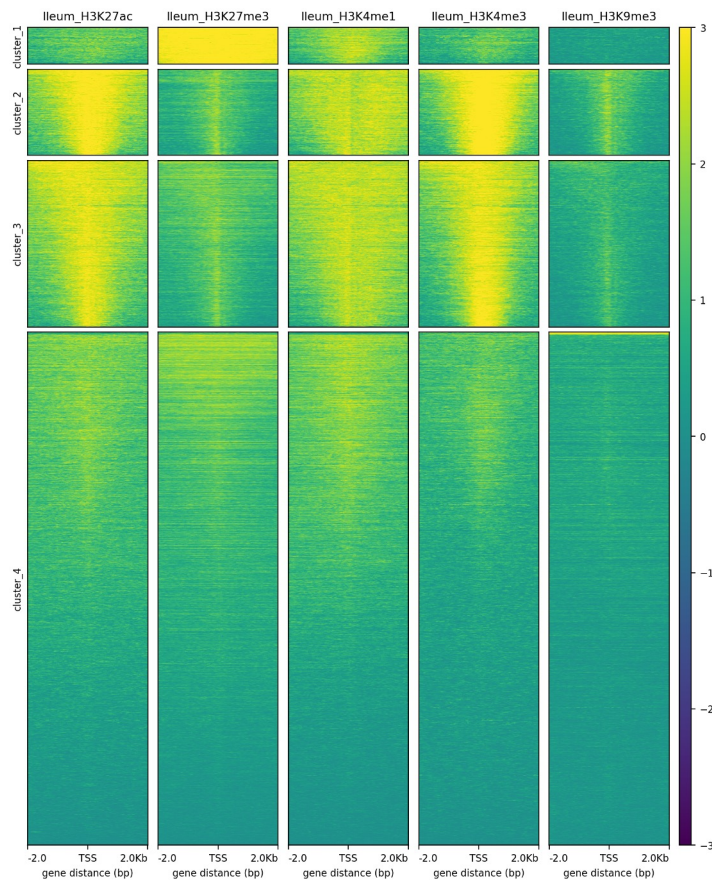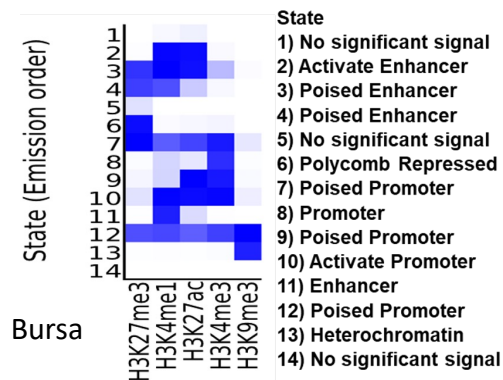| State	H3K27me3	H3K4me1	H3K27ac	H3K4me3	H3K9me3
1) No significant signal	0.002	0.027	0.056	0.001	0.000
2) Activate Enhancer	0.007	0.974	0.955	0.023	0.001
3) Poised Enhancer	0.799	0.998	0.940	0.273	0.013
4) Poised Enhancer	0.757	0.691	0.208	0.028	0.002
5) No significant signal	0.120	0.001	0.002	0.000	0.001
6) Polycomb Repressed	0.954	0.015	0.030	0.044	0.003
7) Poised Promoter	0.995	0.636	0.735	0.902	0.093
8) Promoter	0.015	0.163	0.091	0.822	0.009
9) Poised Promoter	0.039	0.166	0.995	0.904	0.021
10) Activate Promoter	0.115	0.984	0.957	0.978	0.067
11) Enhancer	0.004	0.878	0.143	0.003	0.001
12) Poised Promoter	0.689	0.720	0.627	0.746	0.994
13) Heterochromatin	0.007	0.007	0.006	0.013	0.875
14) No significant signal	0.001	0.000	0.000	0.000	0.001
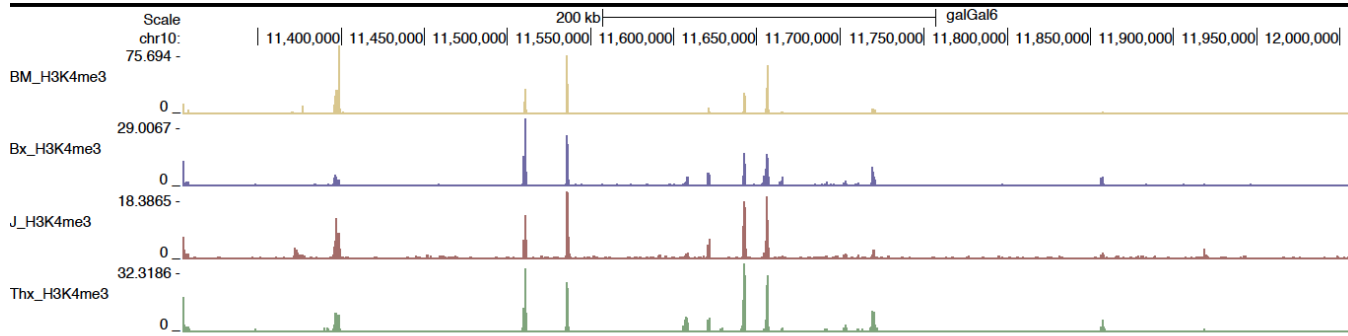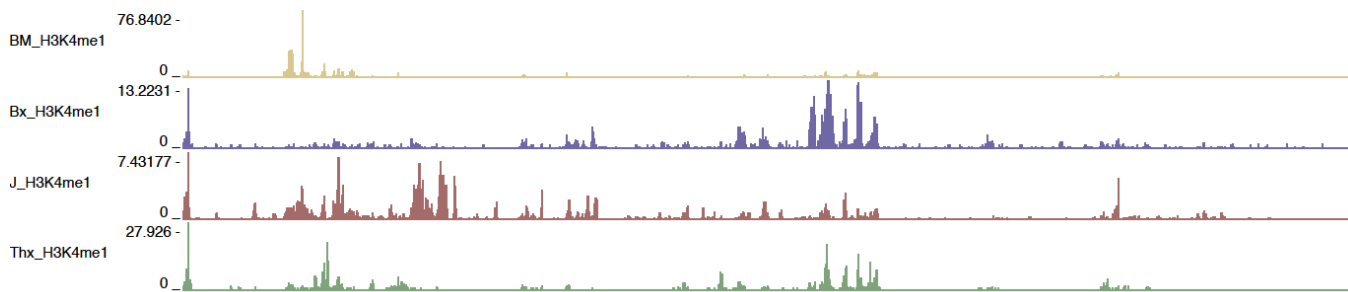
Candidate CREs

ATAC-seq

- Last batch in seq queue.
- Optimizations improved quality.
  - FRiP scores: ~0.4 - 0.7

ChIP-seq

- Tissue dissoiciation optimization.
- CUT&Tag improved data quality.
  - FRiP scores: 0.5 – 0.8
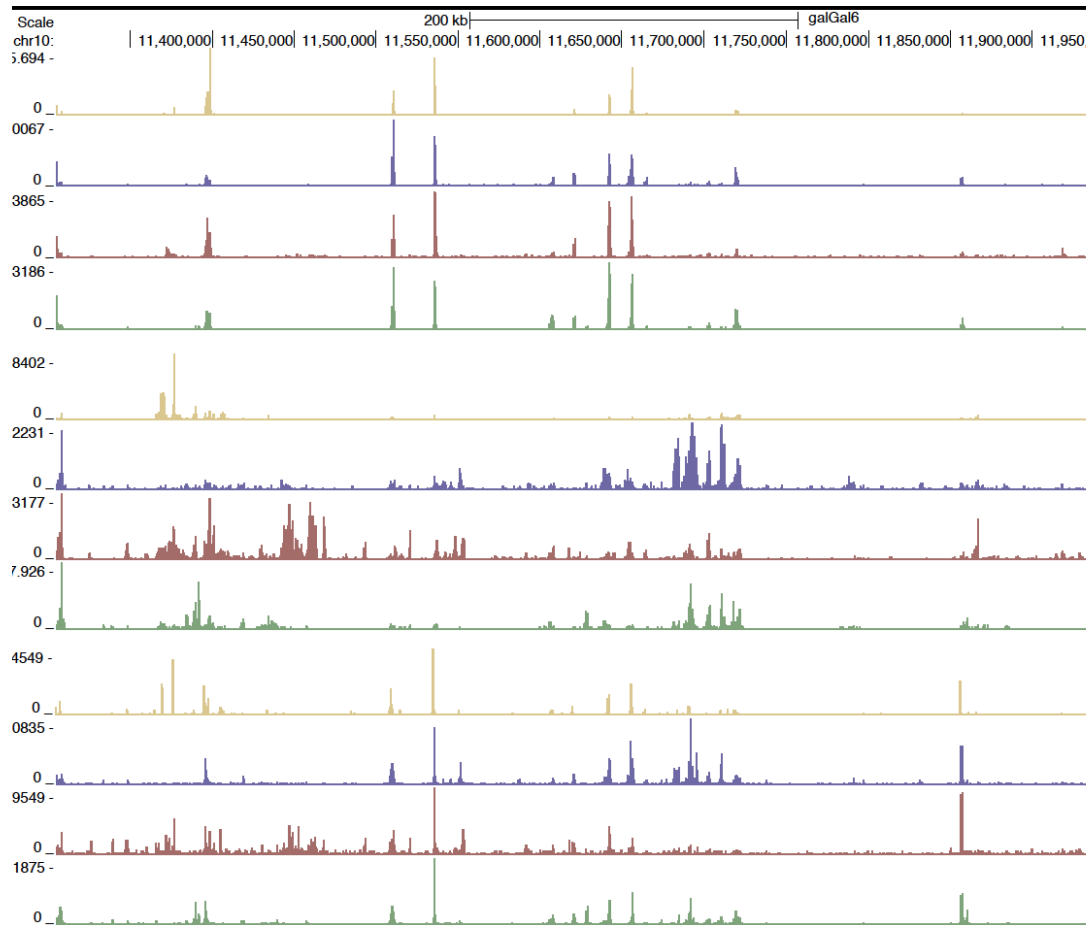
Wrapping Up

- Transcriptomes completed
  - Isoform and novel gene annotation
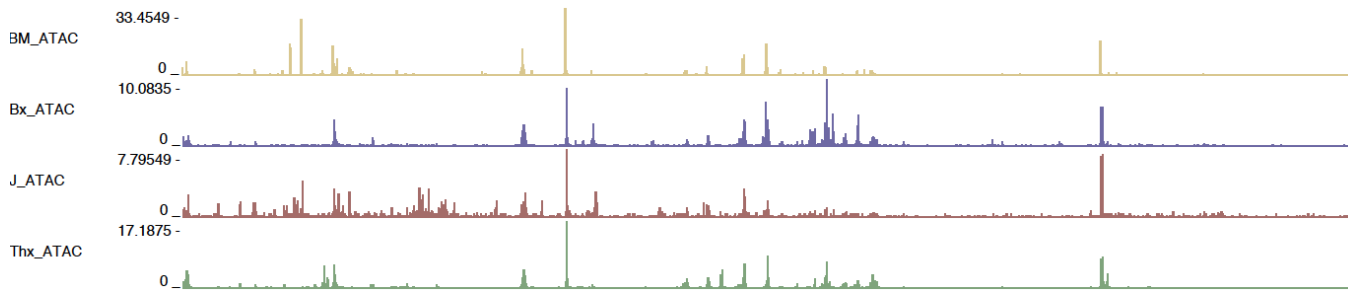 - ASE analysis
  - DEGs for key comparison

- Chromatin mapping in final stages
  - A few samples still lacking some C/T replicates
  - Last batch (hopefully) of ATAC in seq queue
  - Some cell-specific analyses ongoing

- DNA methylation largely completed
  - DMRs for key comparisons
  - Allele-specific methylation
  - Integration with expression ongoing

Challenges

- Pandemic
  - Greatly delayed data production

- Chromatin mapping
  - Even started with mouse ENCODE tissue protocol for ChIP-seq
  - Optimization of tissue dissociations
  - CUT&Tag greatly improved data quality
    - CTCF still remains most variable
  - Optimization of ATAC-seq

- So much exciting data that analysis is still ongoing
  - Hopefully integrative analysis goes on for years!

Looking ahead

- More data integration

- Everything is moving to single-cell
    - In essence, two data modalities

- Bulk maps are still useful
    - Allow for multiple data modalities to be generated and integrated

- Utilizing current maps
    - Functional validation
    - Regulatory variants identification and validation
    - Mapping regulatory loops
    - Comparative analyses across species

- Challenges to systems to find "response elements"
    - Going beyond the static state

- How can others best utilize the current/forthcoming maps?

    - How to make the data more consumable for the community?
        - What type of files, format or interfaces are needed/usable?

        - How to flag QC for data in portal?

        - Regulatory element coordinate system?

Acknowledgements