Comprehensive functional annotation and comparative analysis of the duplicated genomes of salmonid fishes

On behalf of many collaborators

Professor Dan Macqueen. The Roslin Institute, University of Edinburgh

PLANT & ANIMAL GENOME CONFERENCE / PAG 30

THE PREMIER AND LARGEST AG-GENOMICS MEETING IN THE WORLD

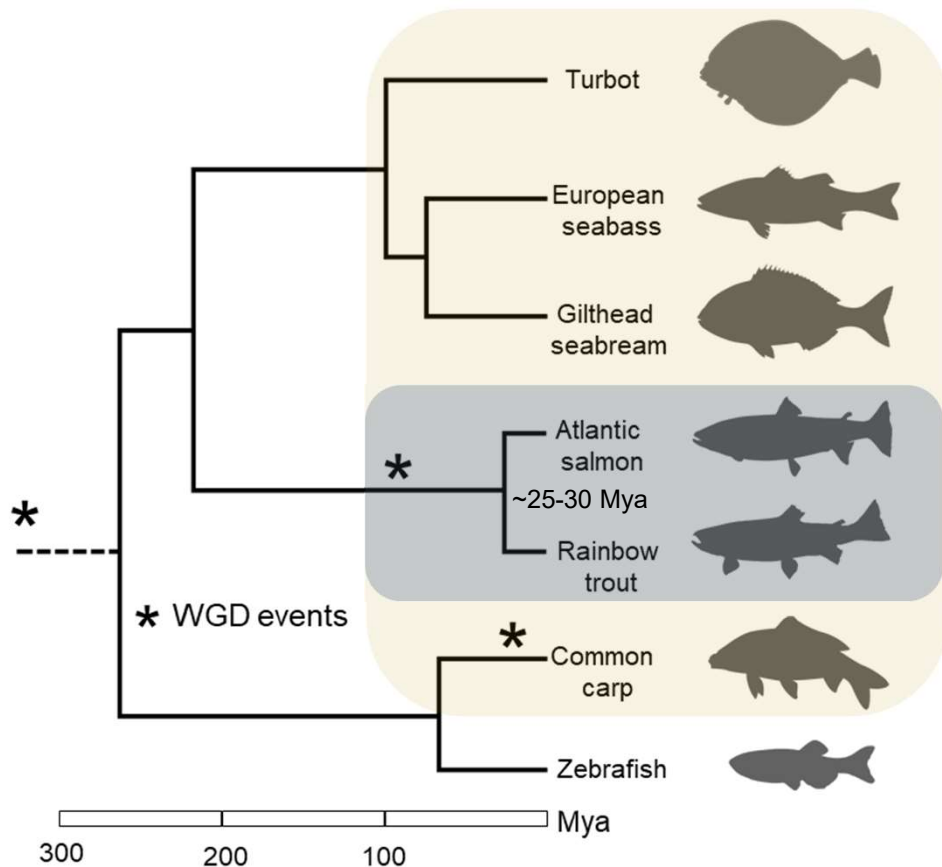**JANUARY 13-18, 2023**
TOWN & COUNTRY HOTEL
SAN DIEGO, CA

Functional Annotations of Animal Genomes (FAANG) Workshop – PAG30 – San Diego – Jan 13th 2023

FAANG
Functional Annotation of Animal Genomes

AQUA-FAANG in a slide

Genome-wide functional annotation maps for six commercially important fish species

- Funded at 6 million Euros by H2020
- May 2019 (--*Pandemic*--) to Oct 2023
- Coordinated by Sigbjørn Lien & Dan Macqueen
- One six EuroFAANG projects

Salmonids - 101

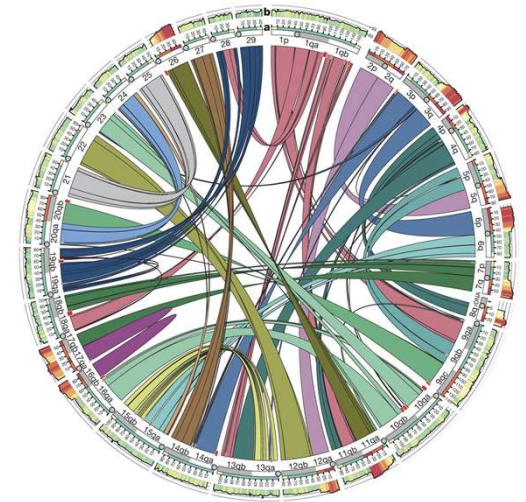**Important family of fishes with interesting genomes**

- Key species for aquaculture – Atlantic salmon & rainbow trout major farmed species in many countries

- Wild stocks have conservation value and cultural significance

Atlantic salmon

Rainbow trout

Salmonids - 101

Important family of fishes with interesting genomes

- Key species for aquaculture – Atlantic salmon & rainbow trout major farmed species in many countries

- Wild stocks have conservation value and cultural significance

- Lineage-specific whole genome duplication (WGD) - 100 MYA [1,2]
  - Pervasive retention of duplicated regions, with 60% of all genes found in duplicate pairs [3]
  - Genome evolution on same (complex) path as earlier key WGD events in vertebrate evolution [2,4]
  - Ideal system for understanding impacts of WGDs on gene expression and regulation evolution

[1] Macqueen and Johnston (2014) Proceedings B. 281: 20132881, [2] Gundappa et al. (2022). MBE. 39 msab310, [3] Lien et al. (2016) Nature 533: 200-5, [4] Robertson et al. (2017) Genome Biology. 18 1-14.

Lien et al. (2016)
Similar picture for any salmonid

AQUA-FAANG salmonid team

Roslin Institute, UK:
Dan Macqueen
Diego Perojil Morata
Pooran Dewari
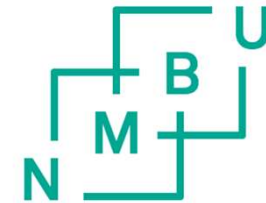Manu Kumar Gundappa
Richard Taylor

Imperial College, UK:
Boris Lenhard
Damir Baranašić

IBENS-ENS, France:
Hugues Roest Crollius
Camille Berthelot
Alexandra Louis
Elise Parey

Norwegian University of Life Sciences, Norway:

Sigbjørn Lien   Simen Sandve
Matthew Kent   Torgeir Hvidsten
Marie-Odile Baudement
Gareth Gillard
Tomasz Podgorniak
Lars Grønvold

INRAe, France:
Yann Guiguen
Audrey Laurent
Pierre Boudinot
Julien Bobe

EMBL-EBI:
Peter Harrison
Garth Ilsley
David Thybert

Salmonid dataset

Mission: Understand dynamics of genome regulation in light of ontogeny and evolution
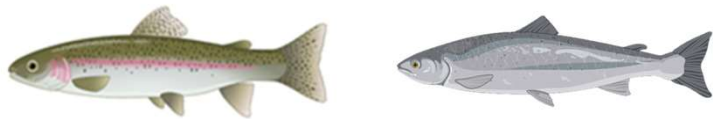
AQUA-FAANG

Embryogenesis samples:
- 14 stages (mRNA-Seq)
- 6 stages (ChIP-Seq and ATAC-Seq)
- μ-chipmentation (Diagenode)

Adult tissue samples for both sexes:
- Liver, brain, gill, intestine, muscle, head kidney, ovary and testis (mRNA-Seq, ATAC-Seq and ChIP-Seq)
- Taken independently at sexually immature and mature stages
- Classic ChIP approach

Salmonid dataset

Mission: Understand dynamics of genome regulation in light of ontogeny and evolution

Libraries sequenced per species:

- mRNA-Seq = 130
- small RNA-Seq = 70
- ATAC-Seq = 80
- ChIP-Seq = 270; histone marks H3K27ac, H3K4me1, H3K4me3, H3K27me3

AQUA-FAANG

FAANG
Functional Annotation of Animal Genomes

Data Portal

FAANG is the **F**unctional **A**nnotation of **AN**imal **Ge**nomes project. We are working to understand the genotype to phenotype link in domesticated animals.

This data portal will help find and browse FAANG's data. Let us know what you think at faang-dcc@ebi.ac.uk.

Data and metadata submitted to FAANG data portal

Data analysis - overview

Analyses run 'locally' using nf-core pipelines (complete) and separately up-taken by Ensembl to update gene annotations & produce novel regulatory annotations (in process)

- **mRNA-Seq nf-core v.3.4 using STAR (alignment) & RSEM (transcript quantification)**
  - RSEM deals effectively with reads multimapping across duplicated genes

- **ATAC-Seq nf-core v1.2.1, ChIP-Seq nf-core v.1.2.2**
  - Each using BWA (alignment) & MACS2 (peak calling)
  - Uniquely mapped reads captured peaks in even the most similar duplicated regions retained from salmonid WGD; multimapping data was noisy

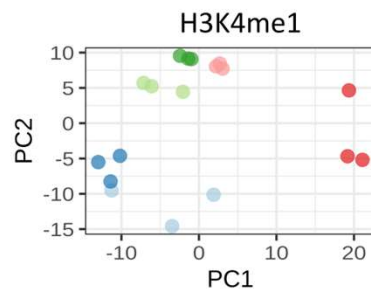# Data analysis - overview

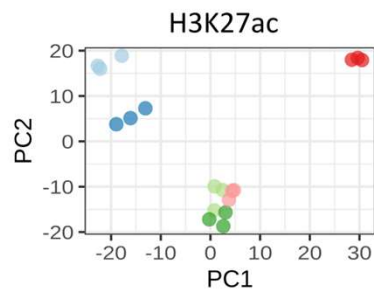## Data produced of high quality
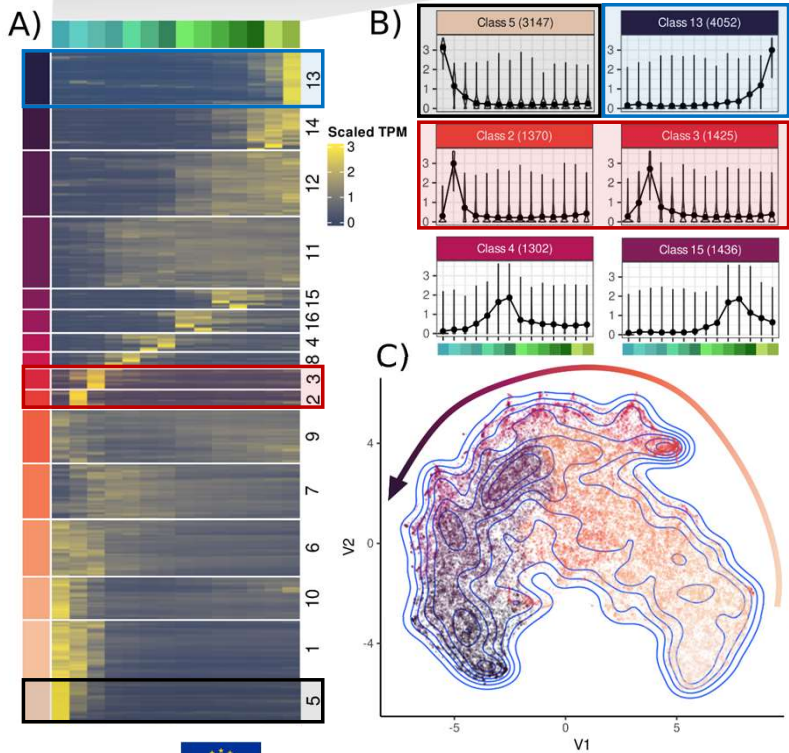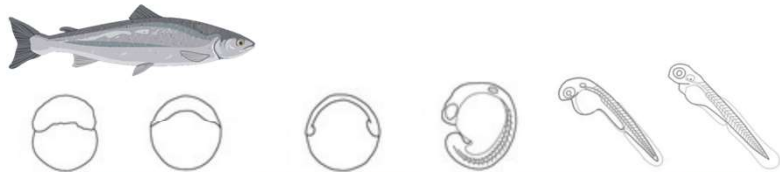
E.g. mRNA-Seq for adult tissues
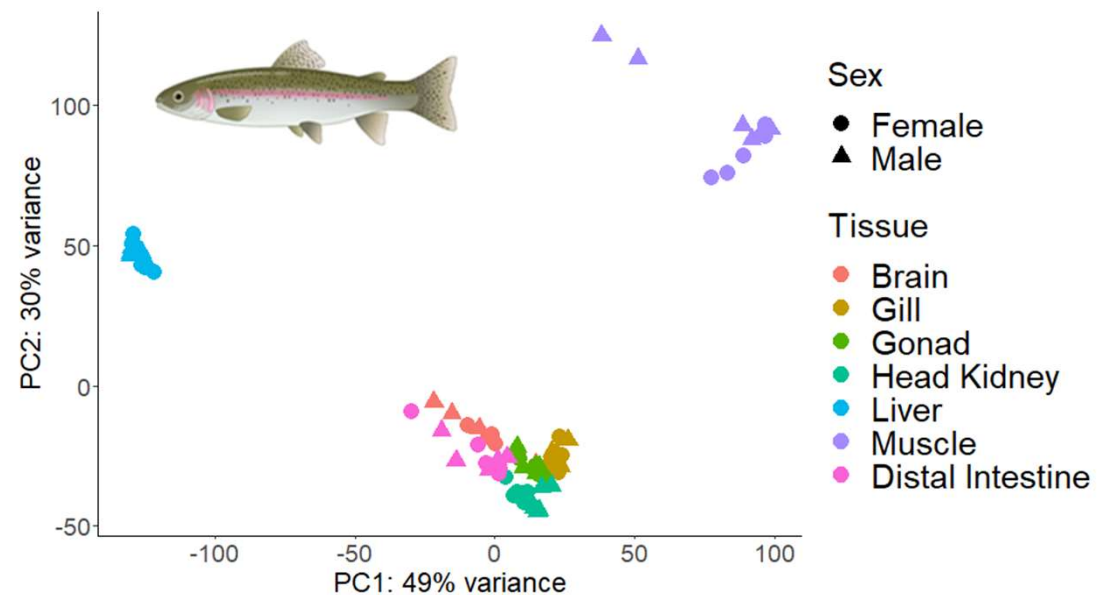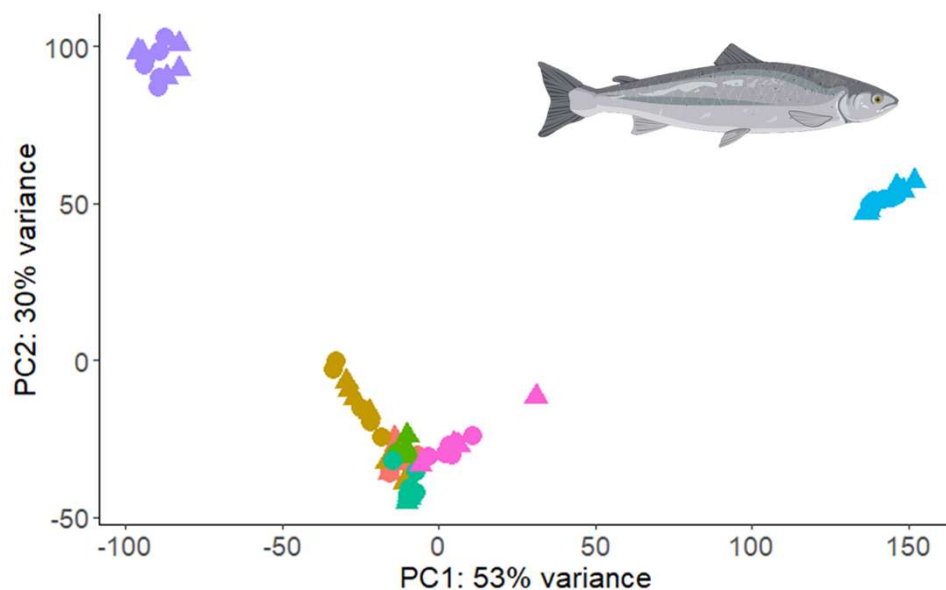
Data analysis - overview

Data produced of high quality

E.g. PCAs: ChIP-Seq peak data for embryo data
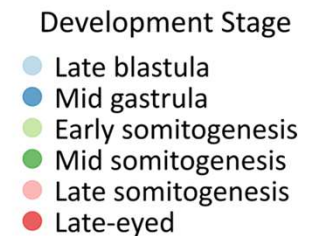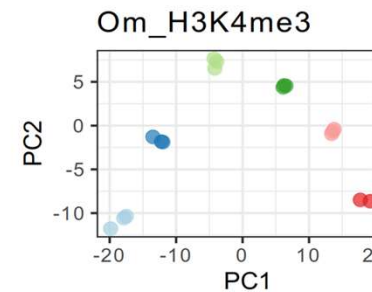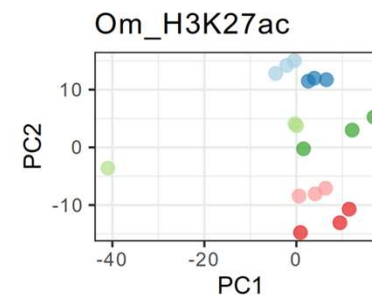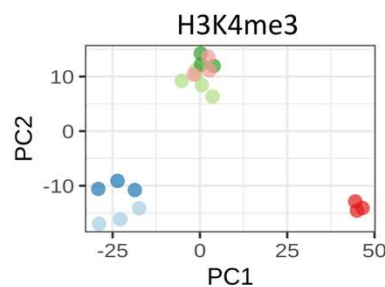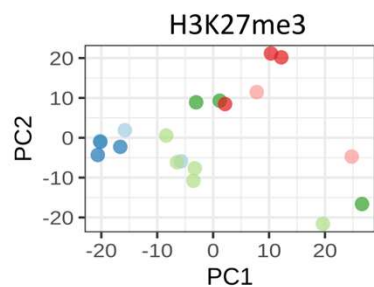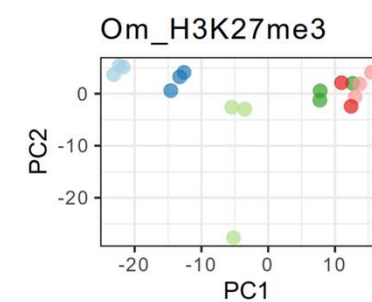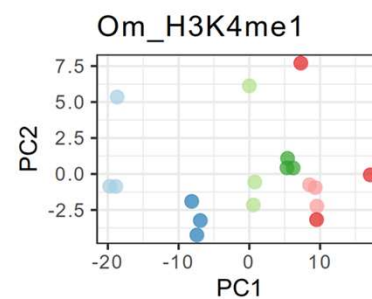
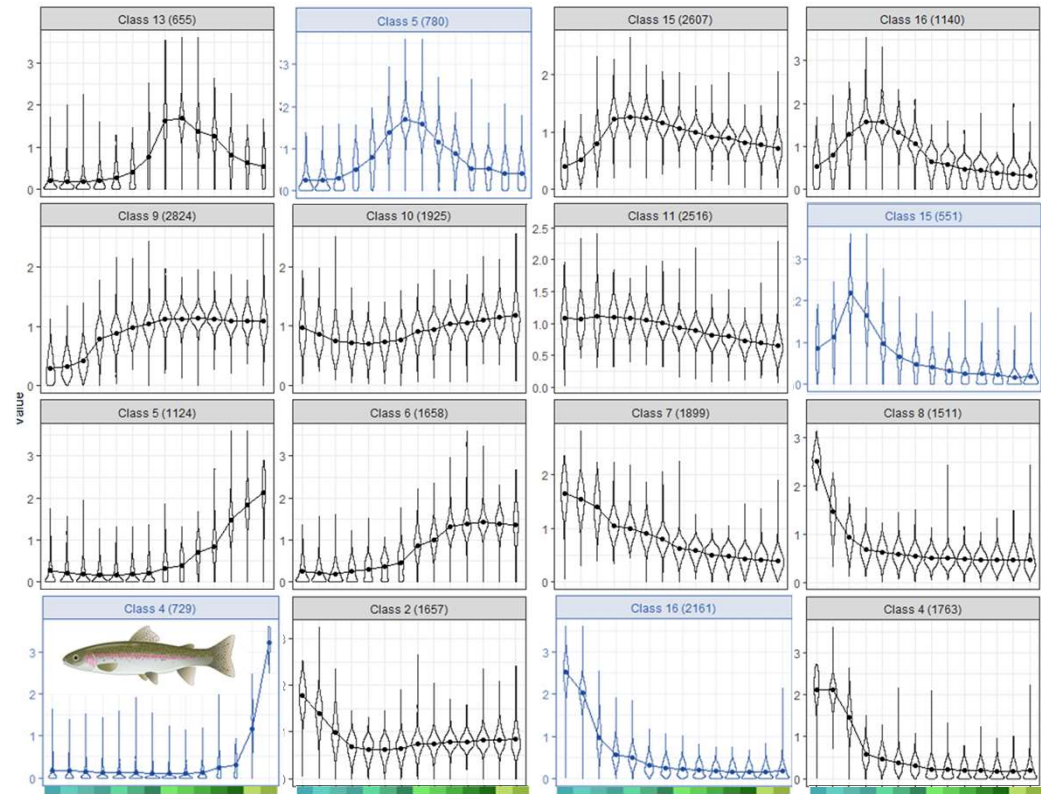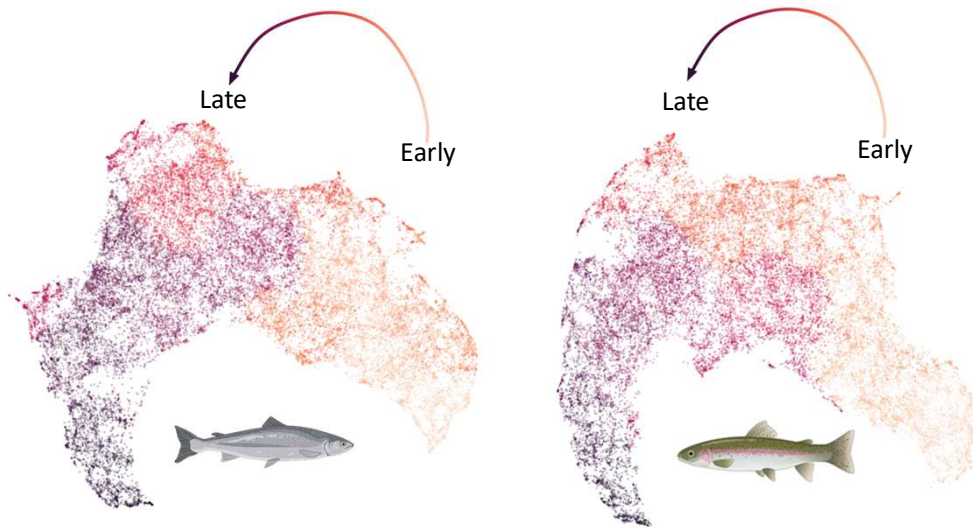# Gene expression across ontogeny

Self-organizing maps (SOM) for embryo mRNA-Seq data
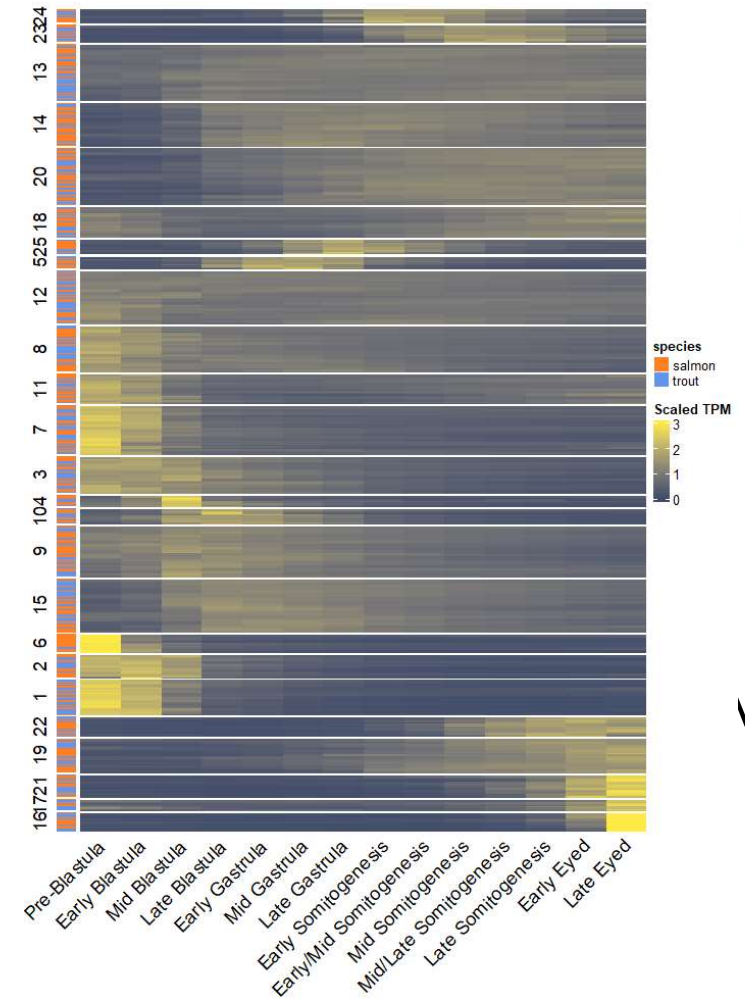
Similar SOM clusters observed across species
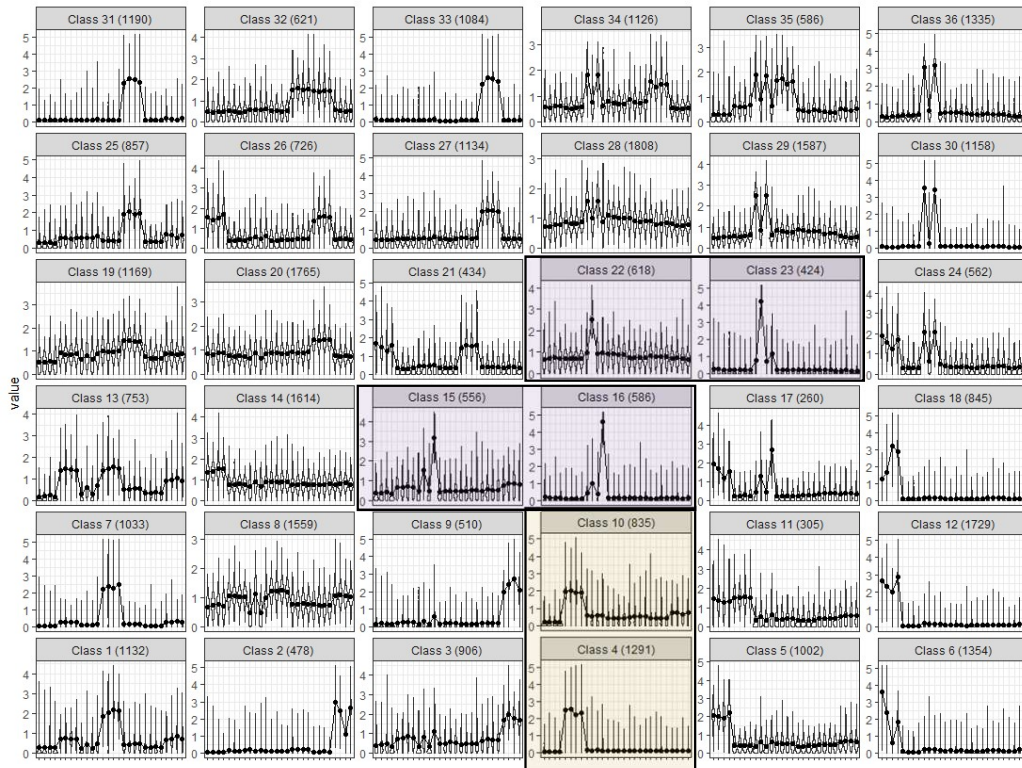
Gene expression across ontogeny

UMAP plots using matched data across species

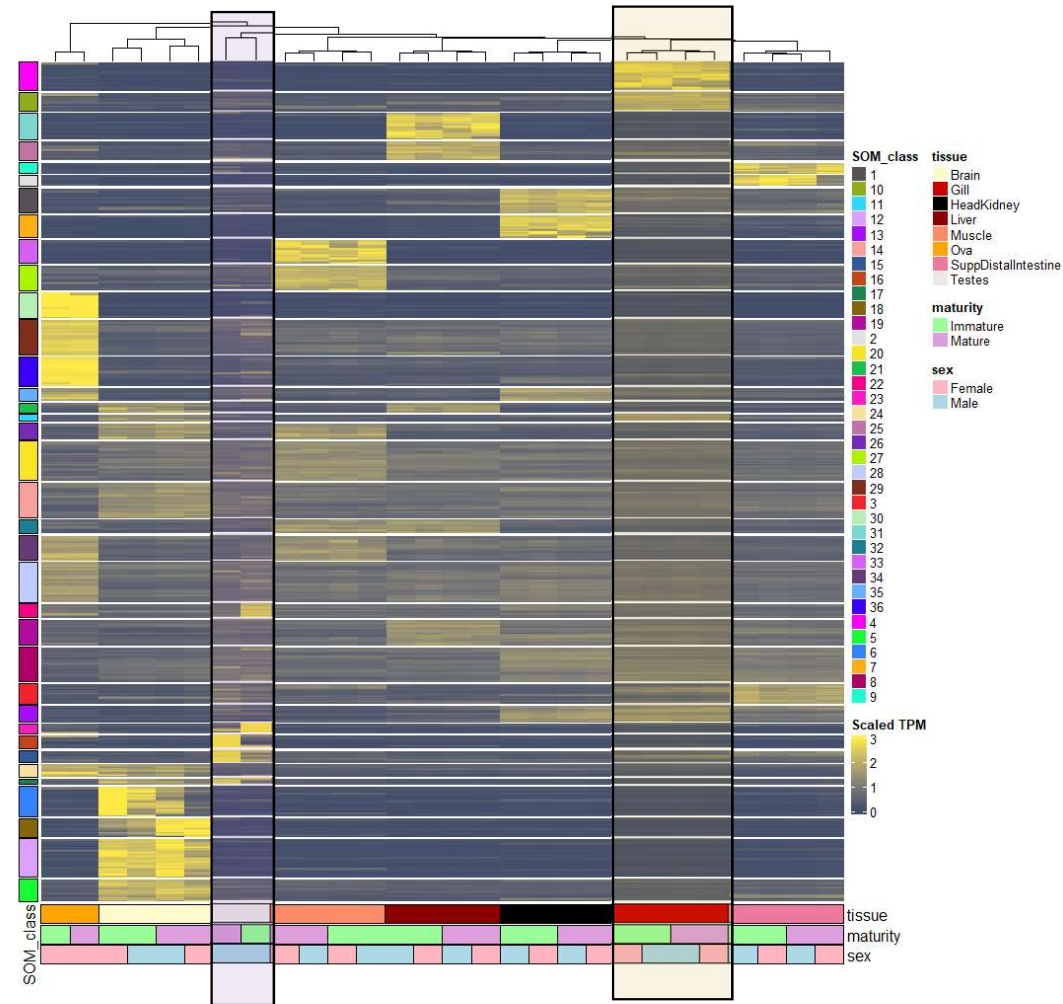- Common trajectory of gene expression during embryogenesis

Gene expression across ontogeny
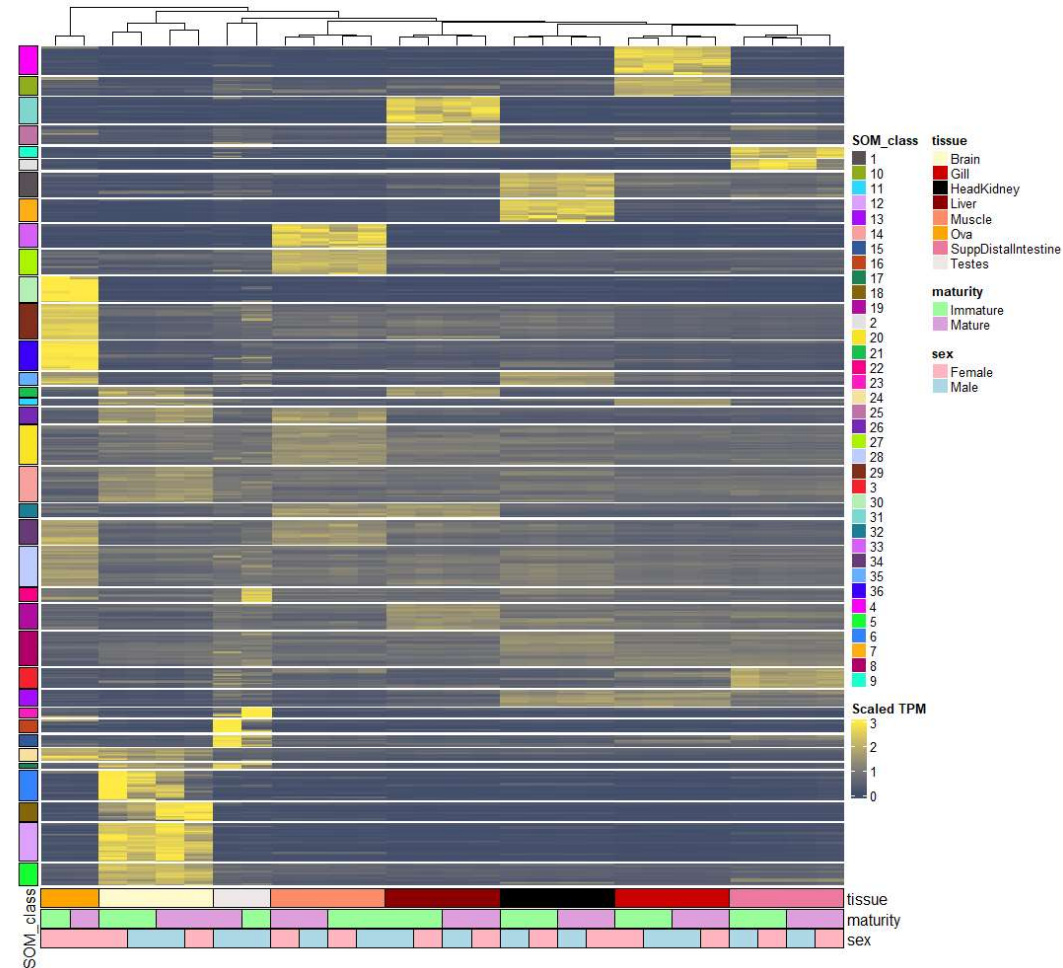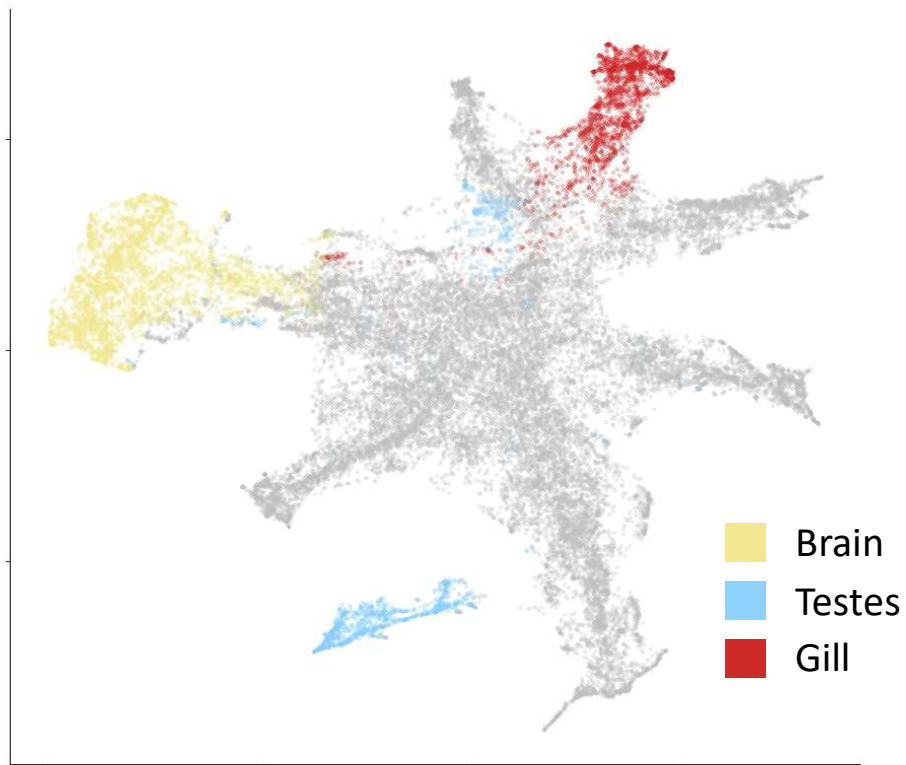
SOM of adult mRNA-Seq data

Tissue and stage expression modules

Gene expression across ontogeny
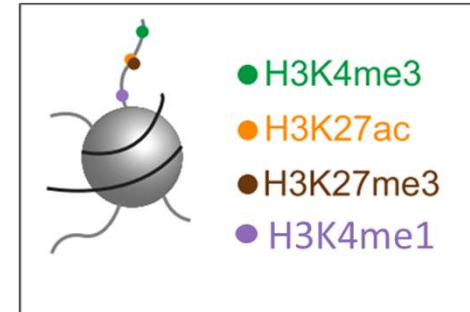
UMAP of adult tissue mRNA-Seq data

Brain
Testes
Gill

Alternative global visualization of expression dynamics
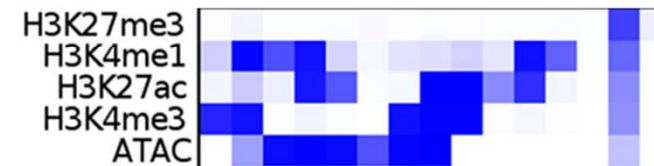
Gene regulation across ontogeny
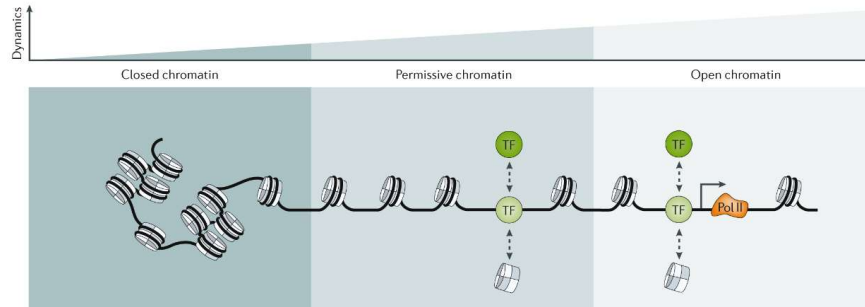
ChromHMM was used to predict and classify chromatin states
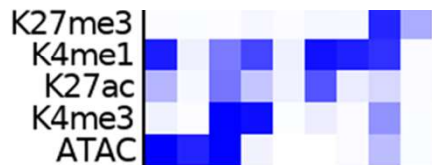
Embryo 10 state model

1: Quiescent
2- Repressed_polycomb
3- Bivalent_poised
4- Weak_Enhancer_Hetero
5- Enhancer_Hetero
6- Quiescent
7- Transcribed at gene
8- Active_Promoter
9- ATAC_island
10-Active_enhancer

Embryo & adult 15 state model

1- Repressed_polycomb
2- Bivalent/poised_TSS
3- Quiescent
4- Poised/inactive_enhancer
5- Active_enhancer
6- Weak_active_enhancer
7- Transcribed_without_ATAC???
8- Strong_active_promoter
9- Flanking_TSS
10- ATAC_island
11- Medium_active_enhancer
12- Strong_active_enhancer
13- Primed_Enhancer
14- Transcribed_at_gene
15- Flanking_TSS_downstream

Gene regulation across ontogeny

- Regulatory elements defined following Baranasic et al. 2022 (Nat. Genet.)

  - High-confidence ATAC-Seq peaks defined using the Irreproducibility Discovery Rate (IDR) across biological replicates & functionally annotated with ChromHMM predictions

High confidence ATAC-Seq peaks in embryos

Sample type	Average	S.D.
Embryos (all stages)	52,512	8,141
Brain (all)	146,152	16,960
Liver (all)	89,572	1,898
Muscle (all)	46,930	8,185
Gonad (all)	63,638	38,169

High-confidence ATAC peaks across samples (examples)
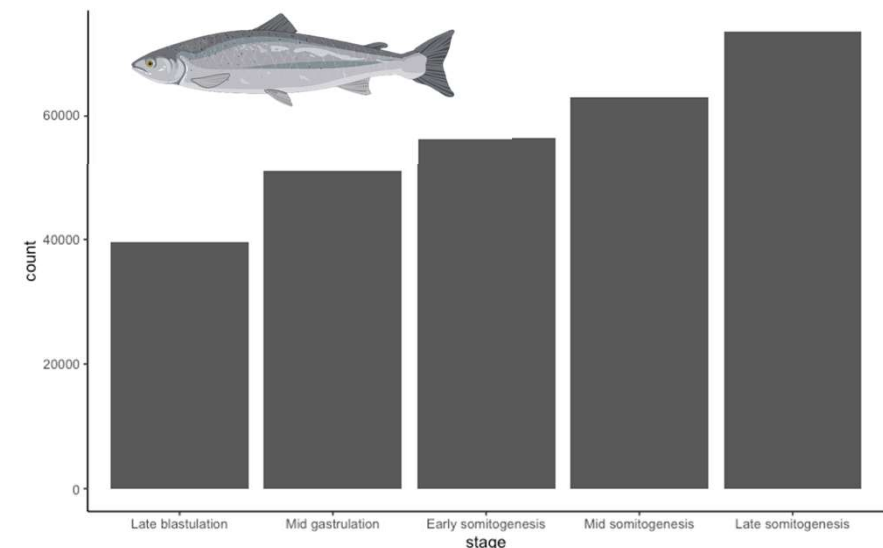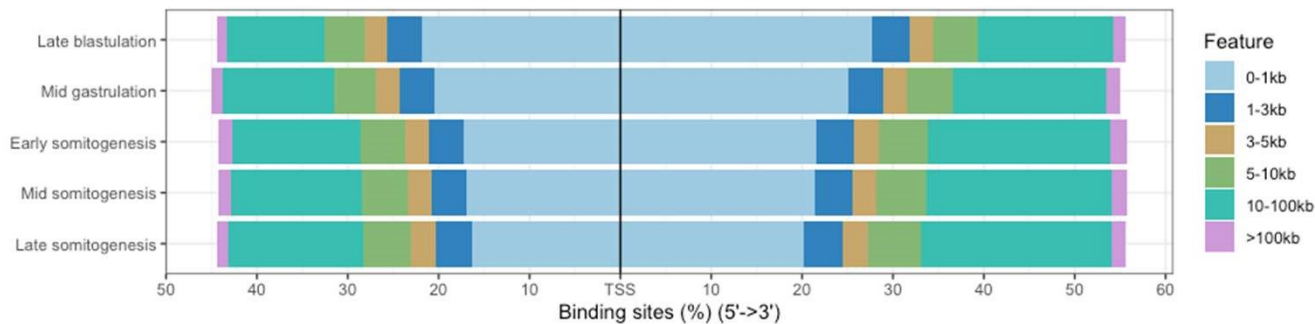
Gene regulation across ontogeny

Enrichment of transcription factor binding motifs in active promoters & enhancers (shown)

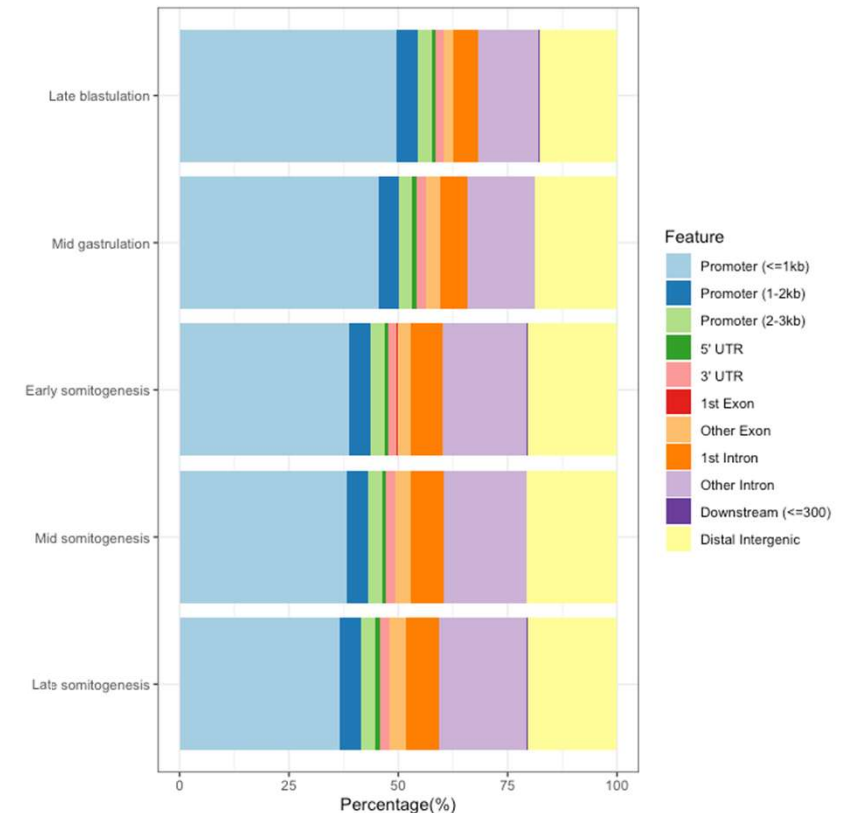- Expected enrichment across sample types
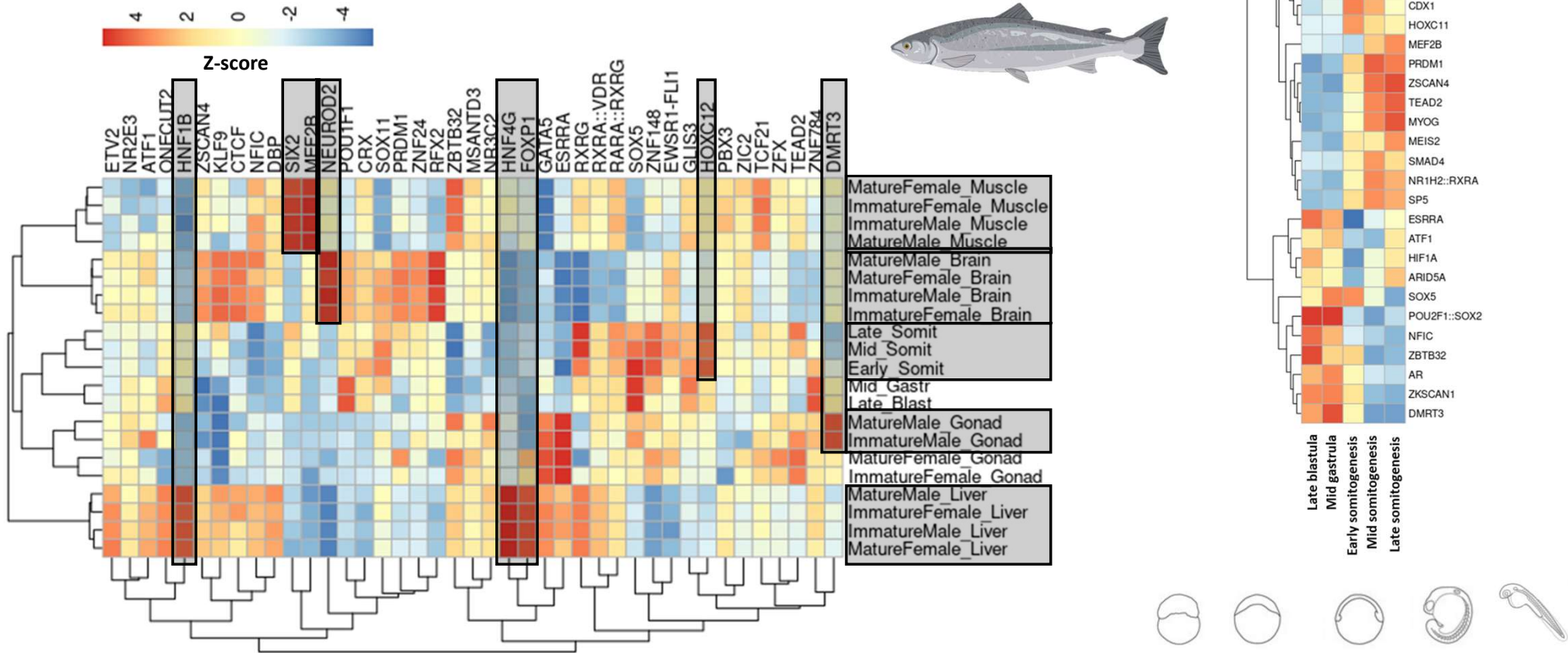
Gene regulation during embryogenesis

Distribution of transcription factor-binding loci relative to TSS

Chromatin progressively opens in promoter distal regions

- Early stages have more promoter & proximal promoter elements

- Later stages, in addition to having more ATAC peaks, have a greater proportion of distal elements – implying long-range regulation
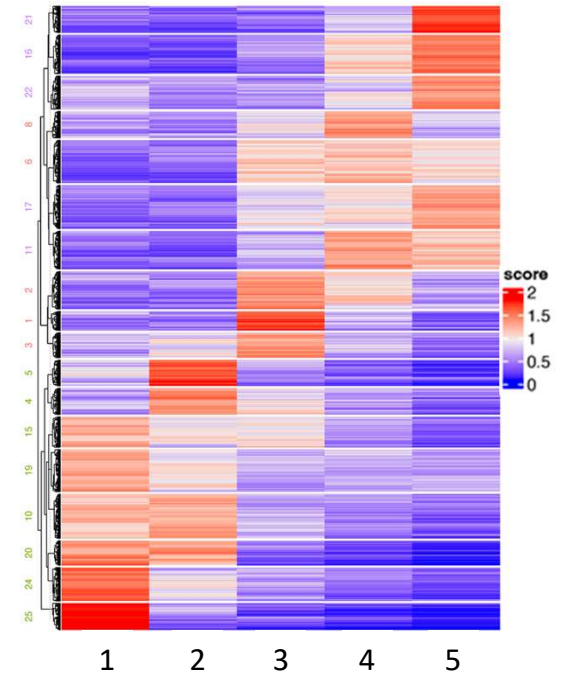
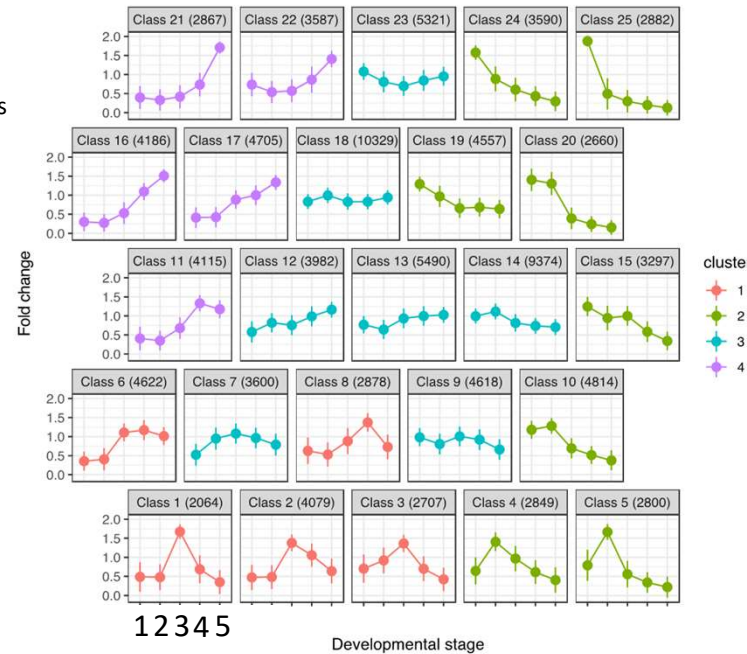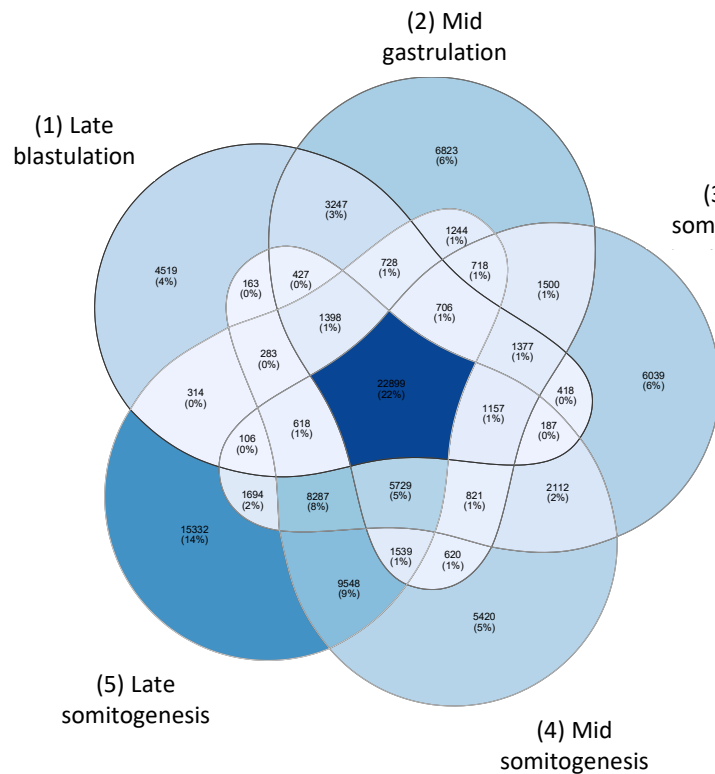- Similar to observations in zebrafish
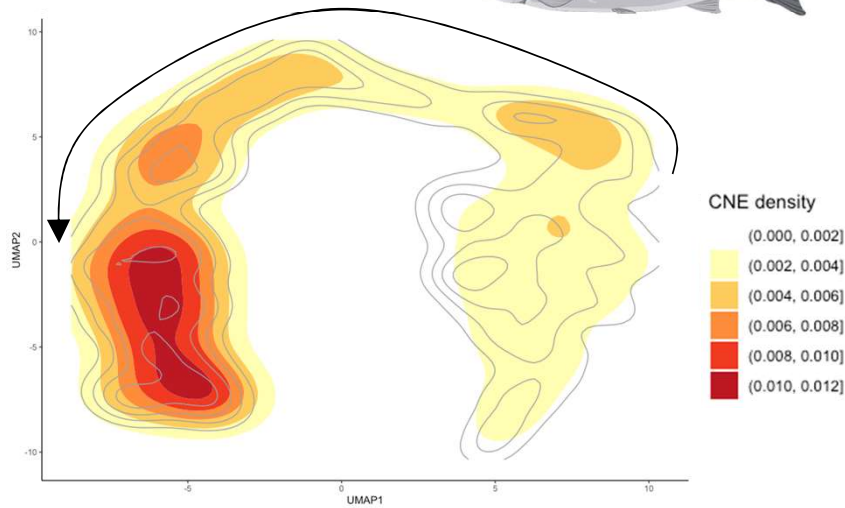
Gene regulation during embryogenesis

Large proportion of open chromatin regions common to all stages
- Most stage-specific peaks in late embryo stages
- Highly dynamic process – explored using SOM clustering of 105,973 peaks

Gene regulation during embryogenesis

UMAP dimensionality reduction captures dynamics of chromatin accessibility during embryonic development
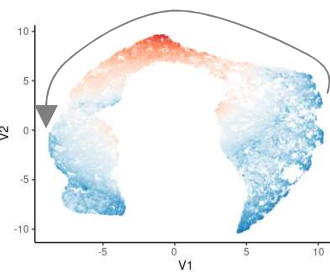
- Trajectory observed recaptures gene expression
- Pseudotime inferred using trajectory inference tool
- Similar results for rainbow trout

- Useful to visualize correlates of regulatory landscape
  - E.g. density of conserved non-coding elements
  - Greater evolutionary constraint on regulation at late stages of embryogenesis – *consistent with literature*
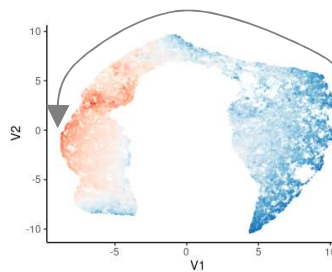
CNE density
- (0.000, 0.002]
- (0.002, 0.004]
- (0.004, 0.006]
- (0.006, 0.008]
- (0.008, 0.010]
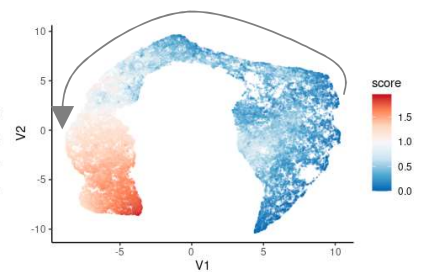- (0.010, 0.012]
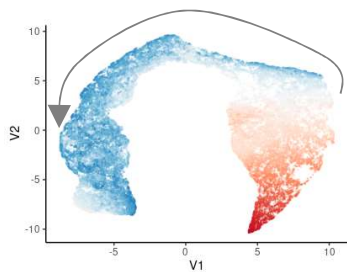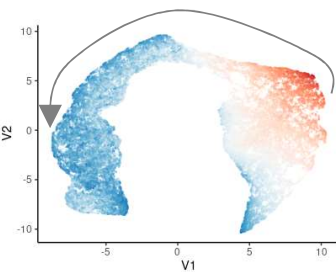
Blastulation (1) Gastrulation (2) Early somitogenesis (3) Mid somitogenesis (4) Late somitogenesis (5)

Gene expression evolution after WGD

Jenson-Shannon Distance (JSD)[1] used to compare expression divergence for 11,495 duplicate gene pairs retained from WGD

- Ranking of samples by average degree of conservation of duplicate gene expression

- Late embryonic stages show least divergence of duplicated gene expression

- Highest divergence - maternal mRNA to onset of zygotic genome activation in embryos and in liver & gonad of mature fish
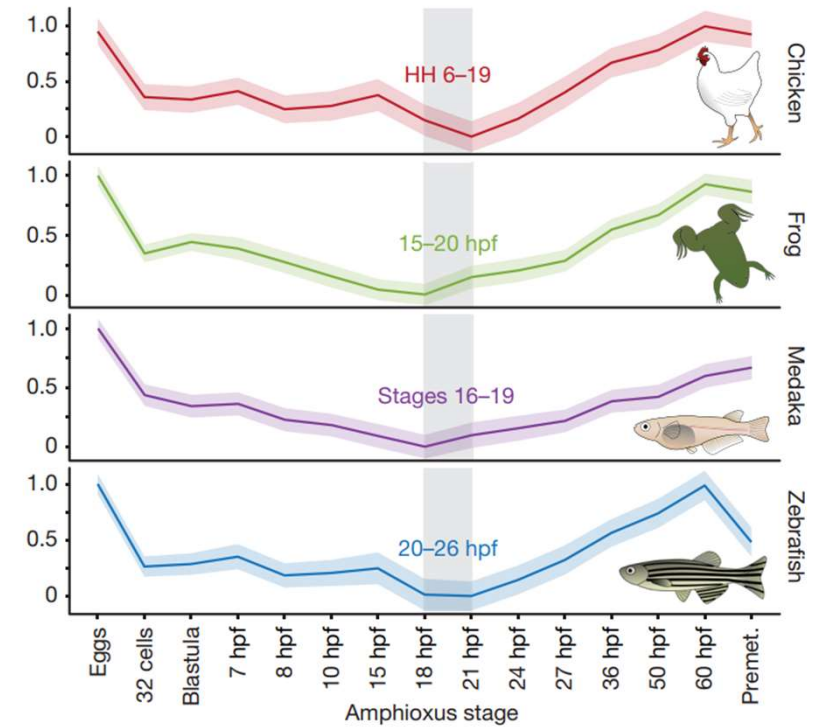
[1] Marlétaz et al. 2018. Nature. 564: 64–70.
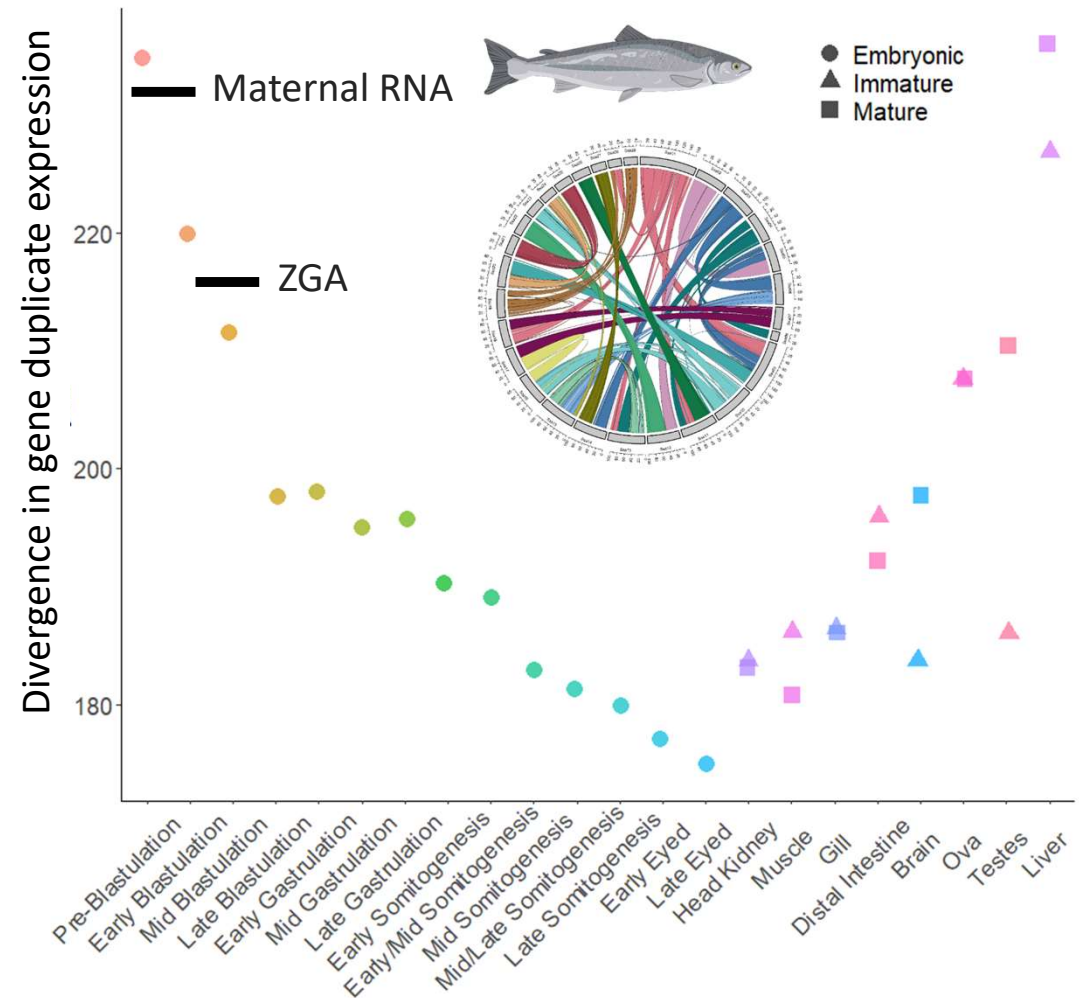
Gene expression evolution after WGD
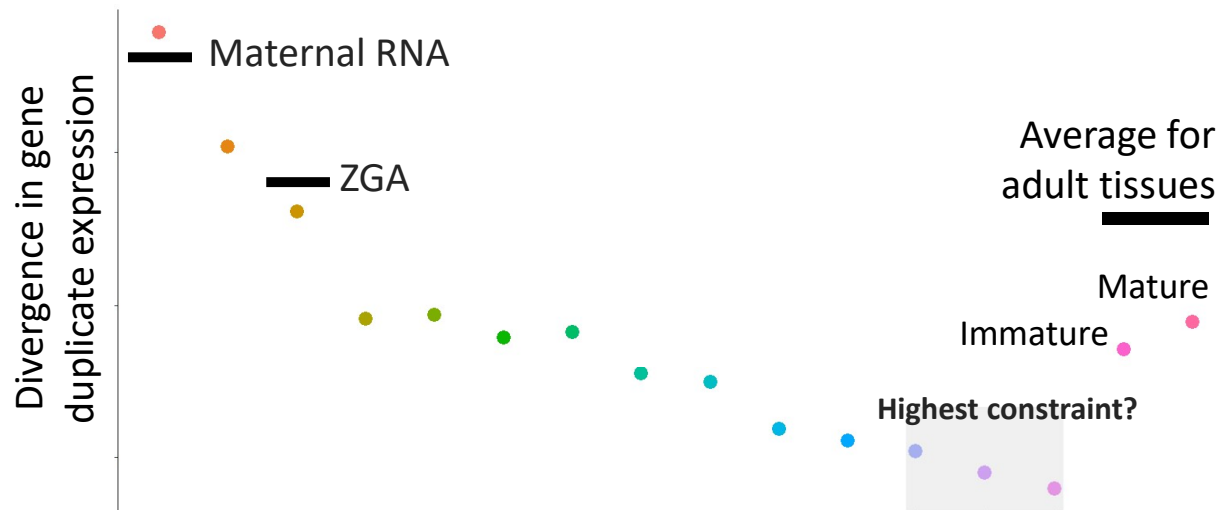
- Consistent with **hourglass model of development**, supported previously for orthologous transcriptome data

Marlétaz et al. 2018. Nature. 564: 64–70.

Acknowledgements

Roslin Institute, UK:
Dan Macqueen
Diego Perojil Morata
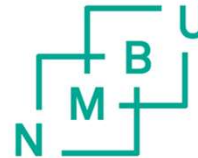Pooran Dewari
Manu Kumar Gundappa
Richard Taylor

Imperial College, UK:
Boris Lenhard
Damir Baranašić

IBENS-ENS, France:
Hugues Roest Crollius
Camille Berthelot
Alexandra Louis
Elise Parey

Norwegian University of Life Sciences, Norway:
Sigbjørn Lien   Simen Sandve
Matthew Kent   Torgeir Hvidsten
Marie-Odile Baudement
Gareth Gillard
Tomasz Podgorniak
Lars Grønvold

INRAe, France:
Yann Guiguen
Audrey Laurent
Pierre Boudinot
Julien Bobe

EMBL-EBI:
Peter Harrison
Garth Ilsley