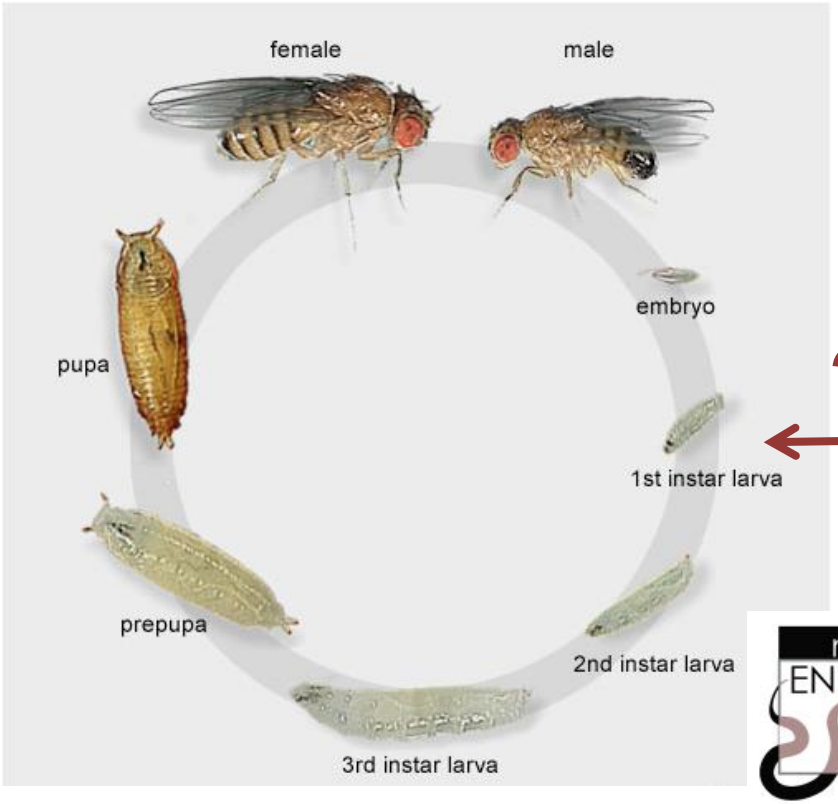
Testing-based Measures for Comparing Genomic Samples

Jingyi Jessica Li

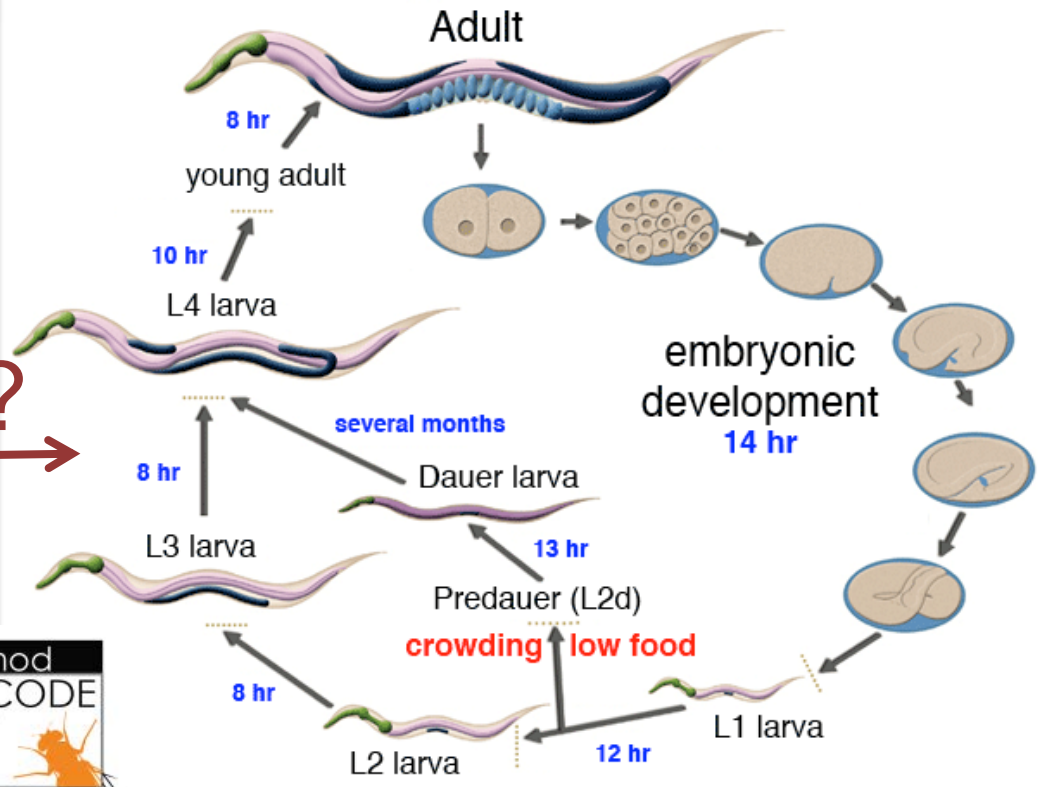
Assistant Professor
Department of Statistic & Department of Human Genetics
University of California, Los Angeles

Motivating example 1

D. melanogaster



C. elegans



?

mod ENCODE

RNA-seq data

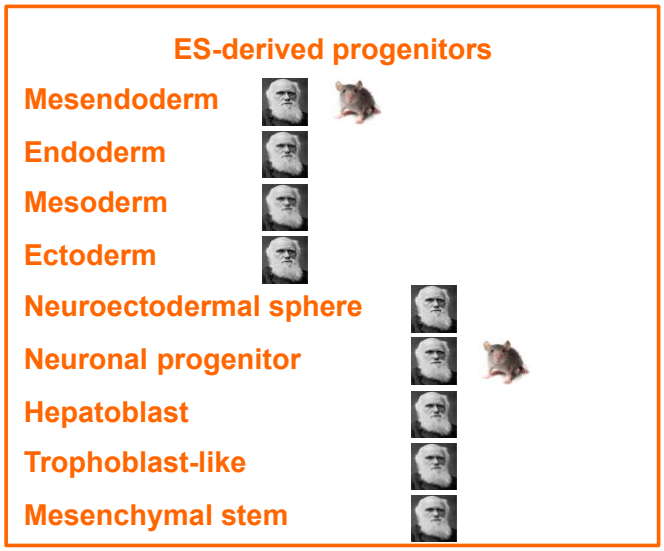
FlyMove (Weigmann et al. 2003)

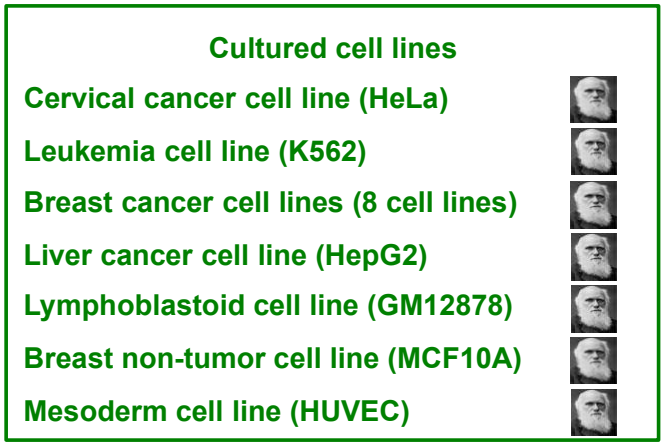
Wormatlas (Altun 2002-2012)

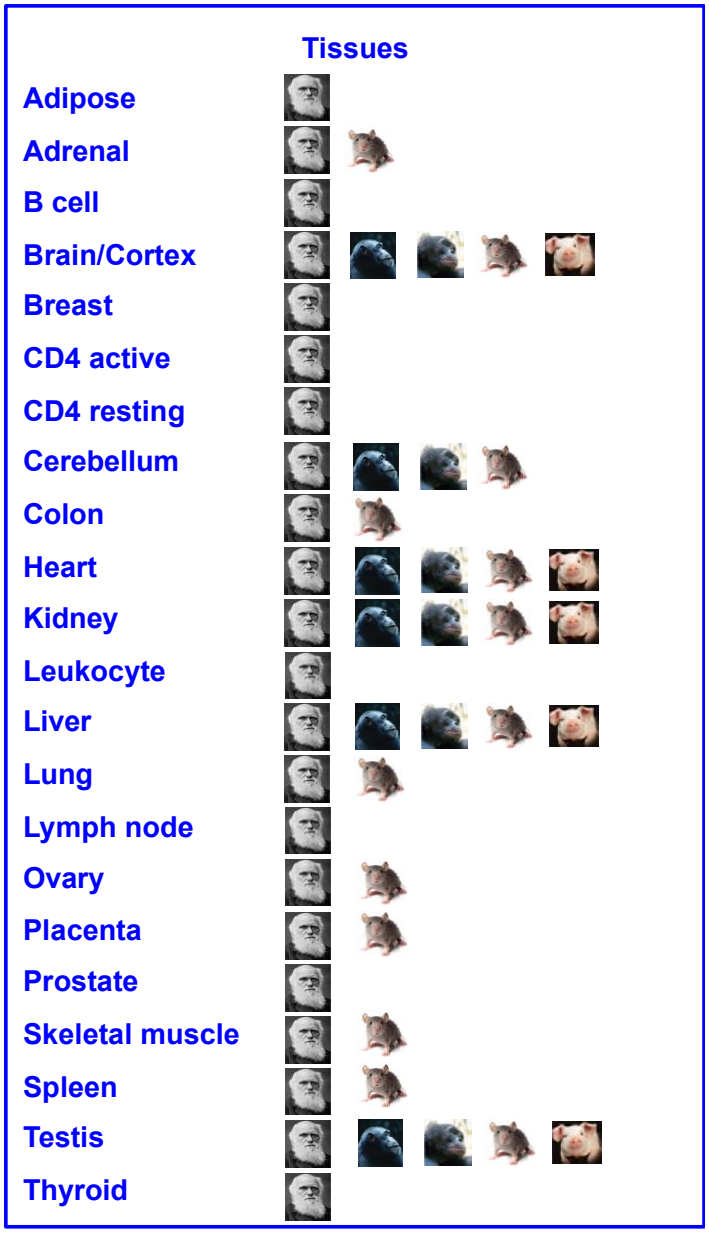
Motivating example 2

RNA-seq data

courtesy to Yu-Cheng T. Yang (Tsinghua University)

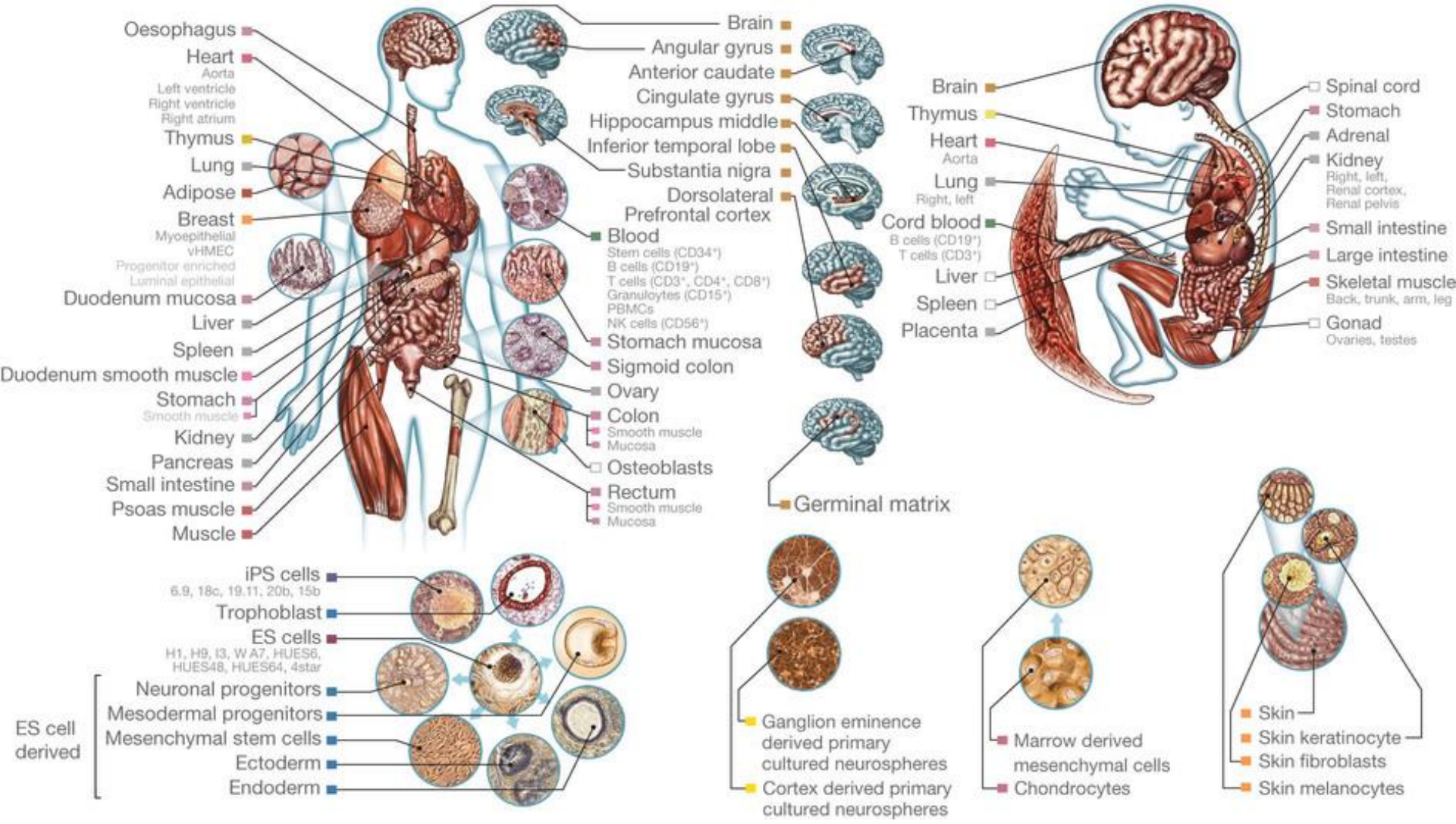






ChIP-seq data

Motivating example 3



Kundaje et al. “Integrative analysis of 111 reference human epigenomes”. *Nature* (2015).

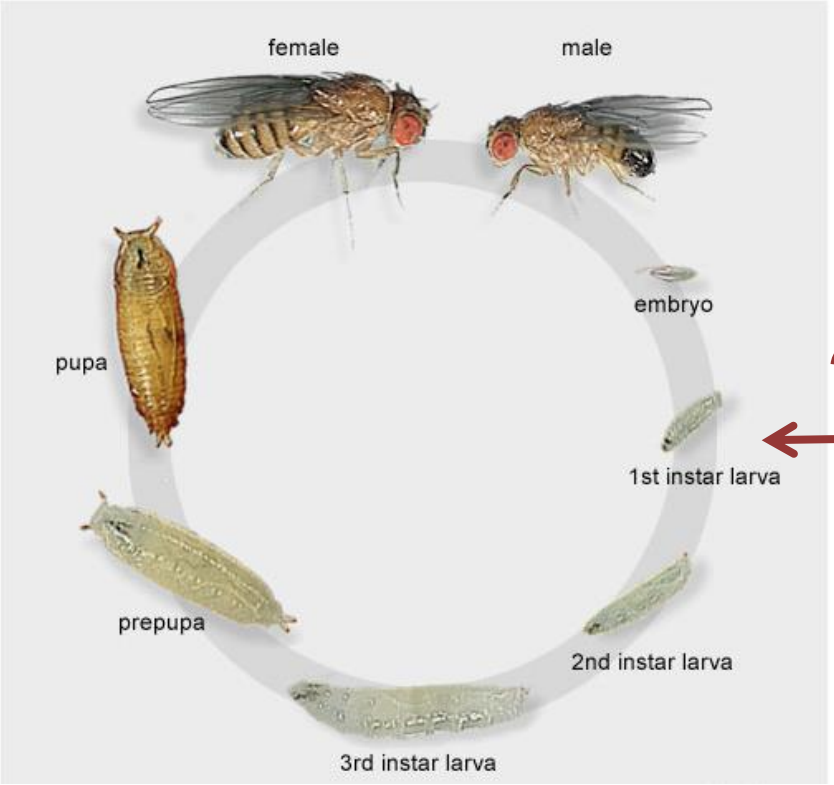
How to compare genomic samples ?

- Transcriptome Overlap Measure (TROM)
- Epigenome Overlap Measure (EPOM)

Part I
Transcriptome Overlap Measure (TROM)

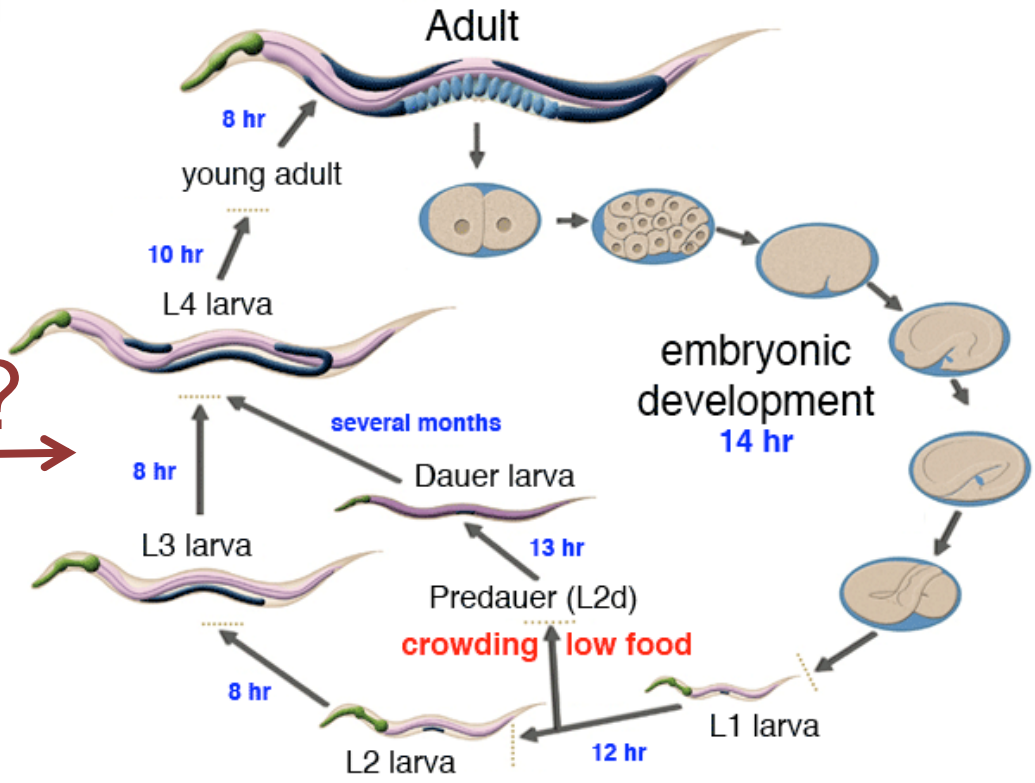
Motivating example 1

D. melanogaster



?

C. elegans

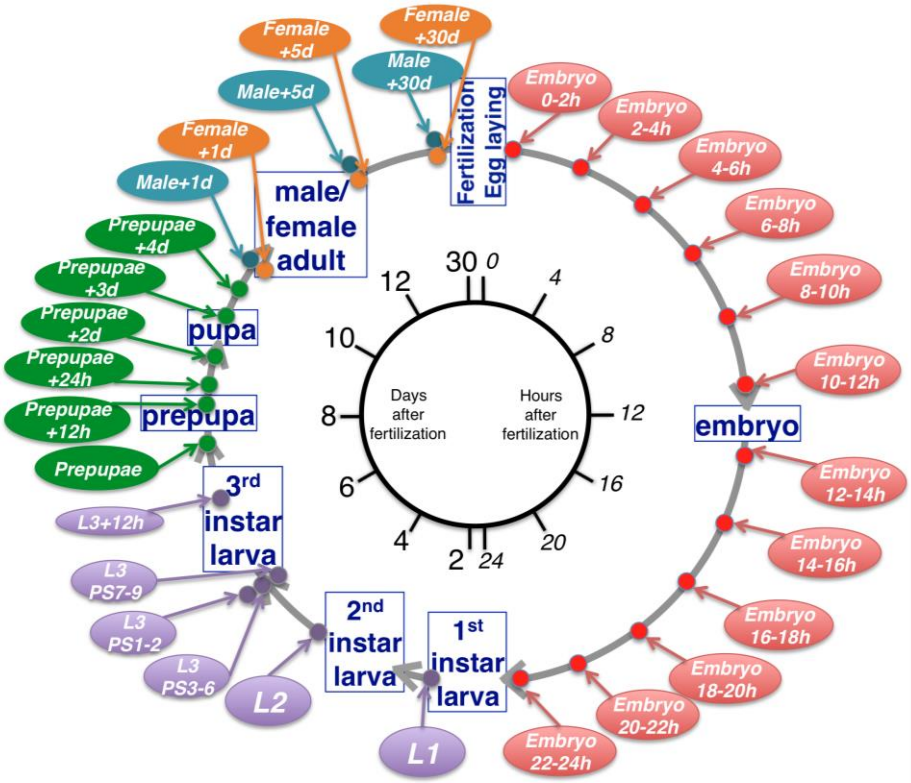


modENCODE timecourse RNA-Seq data

D. melanogaster

(Labs: Celniker, Graveley)

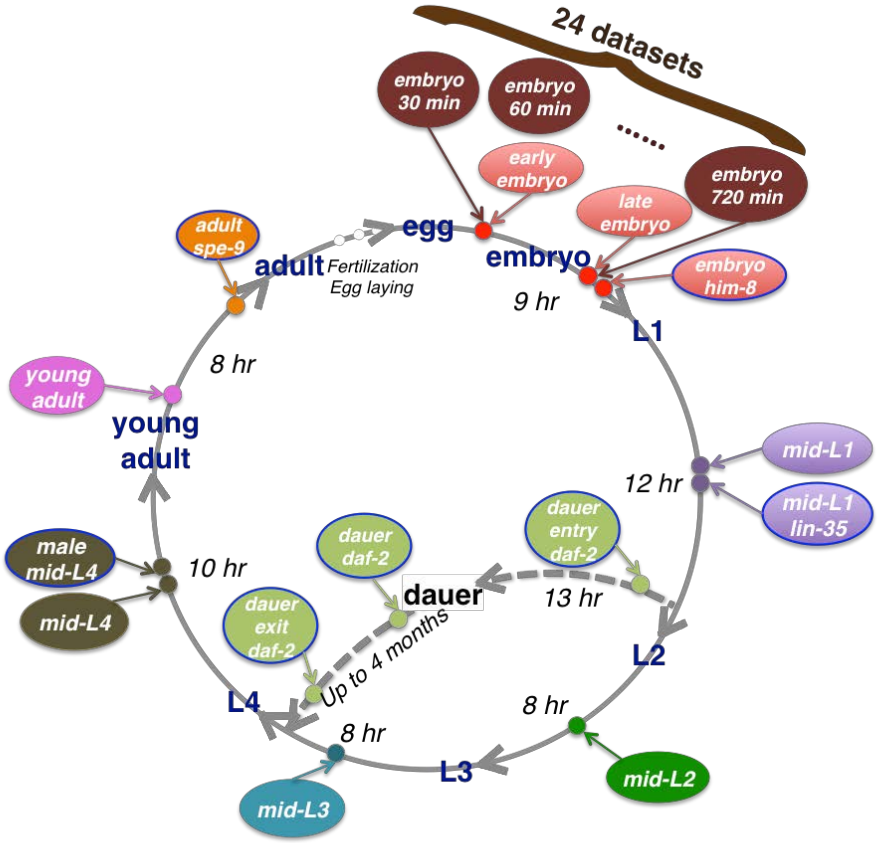
modENCODE consortium, *Science*, 2010



C. elegans

(Lab: Waterston)

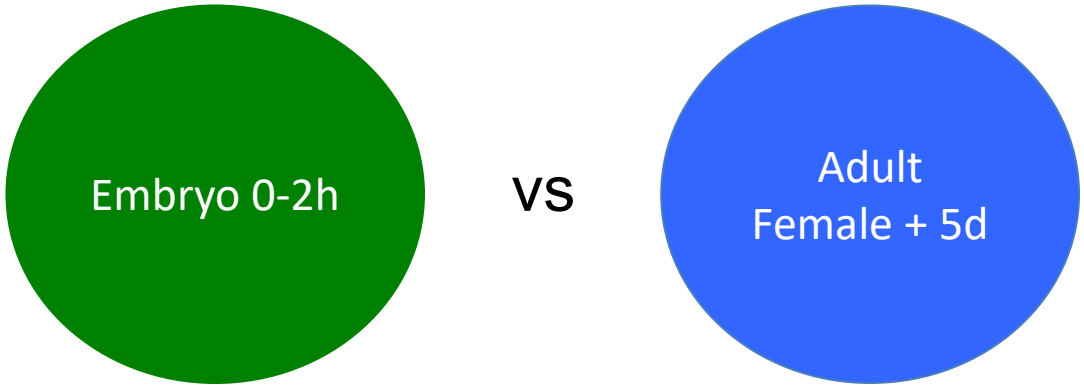
Gerstein et al, *Science*, 2010



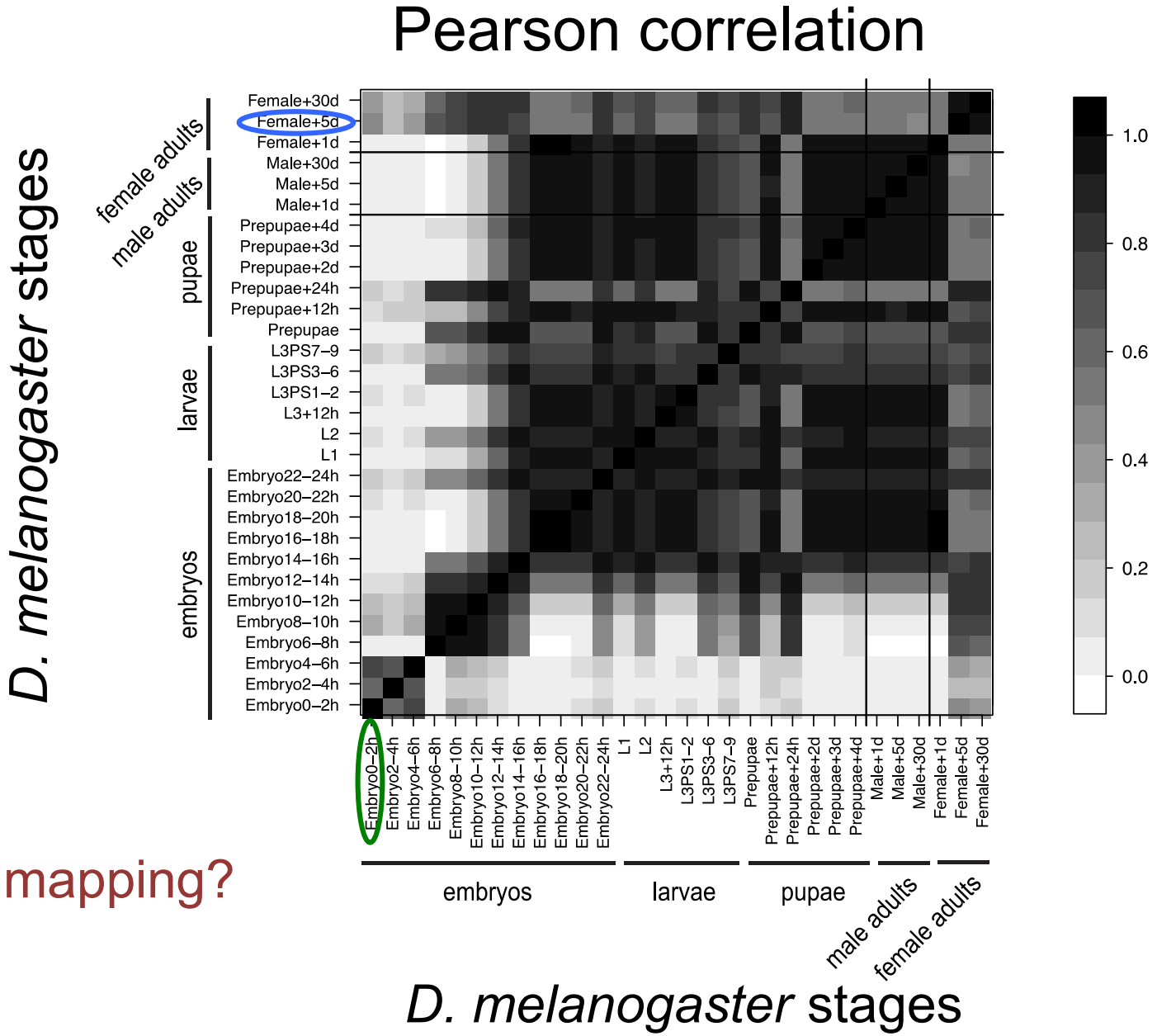
Question 1

- Within each species, is there a clear mapping between its developmental stages in terms of gene expression?

- Example: *D. melanogaster* stages

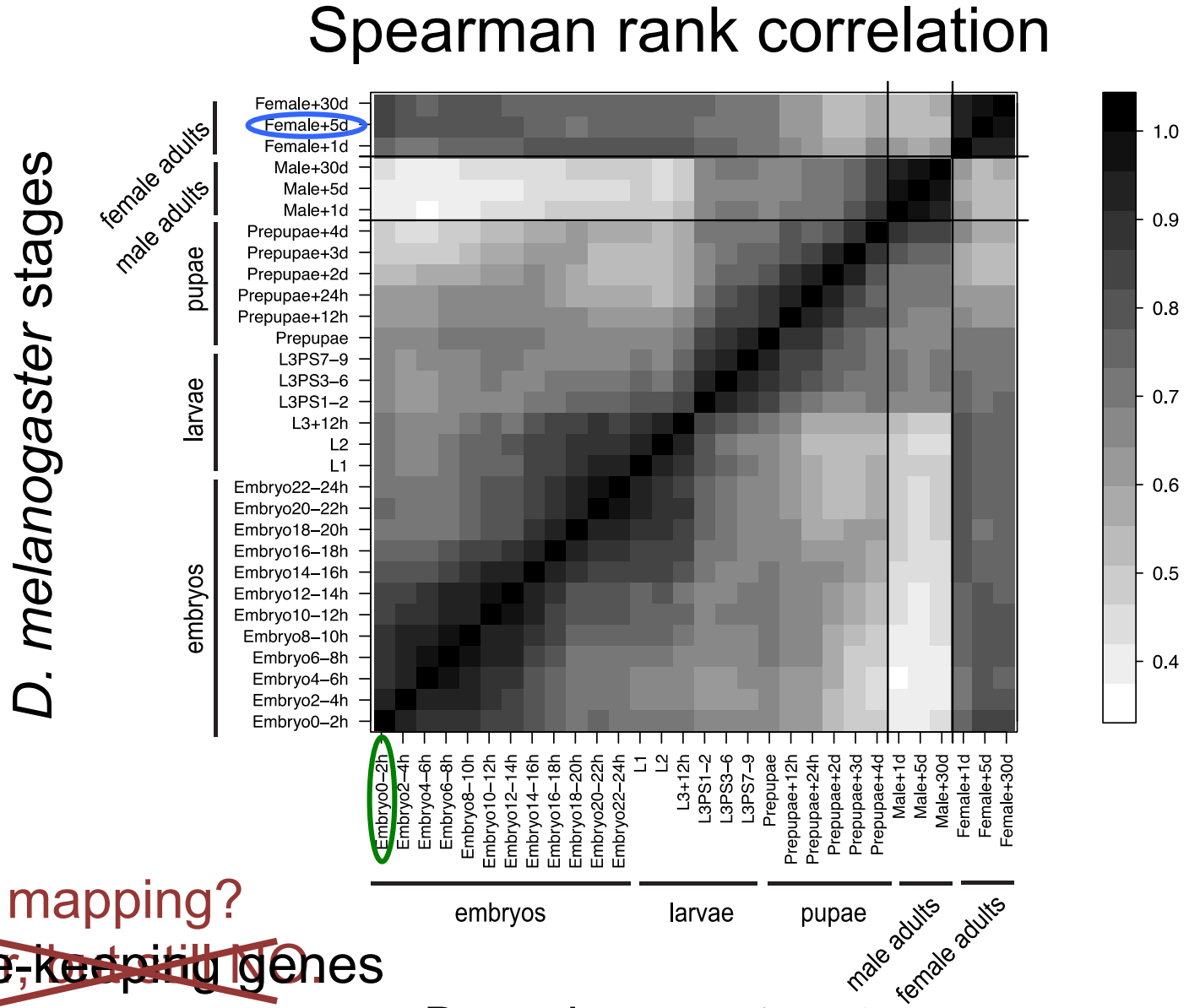


Quick answer: correlation analysis?



clear mapping?
NO.

Quick answer: correlation analysis?



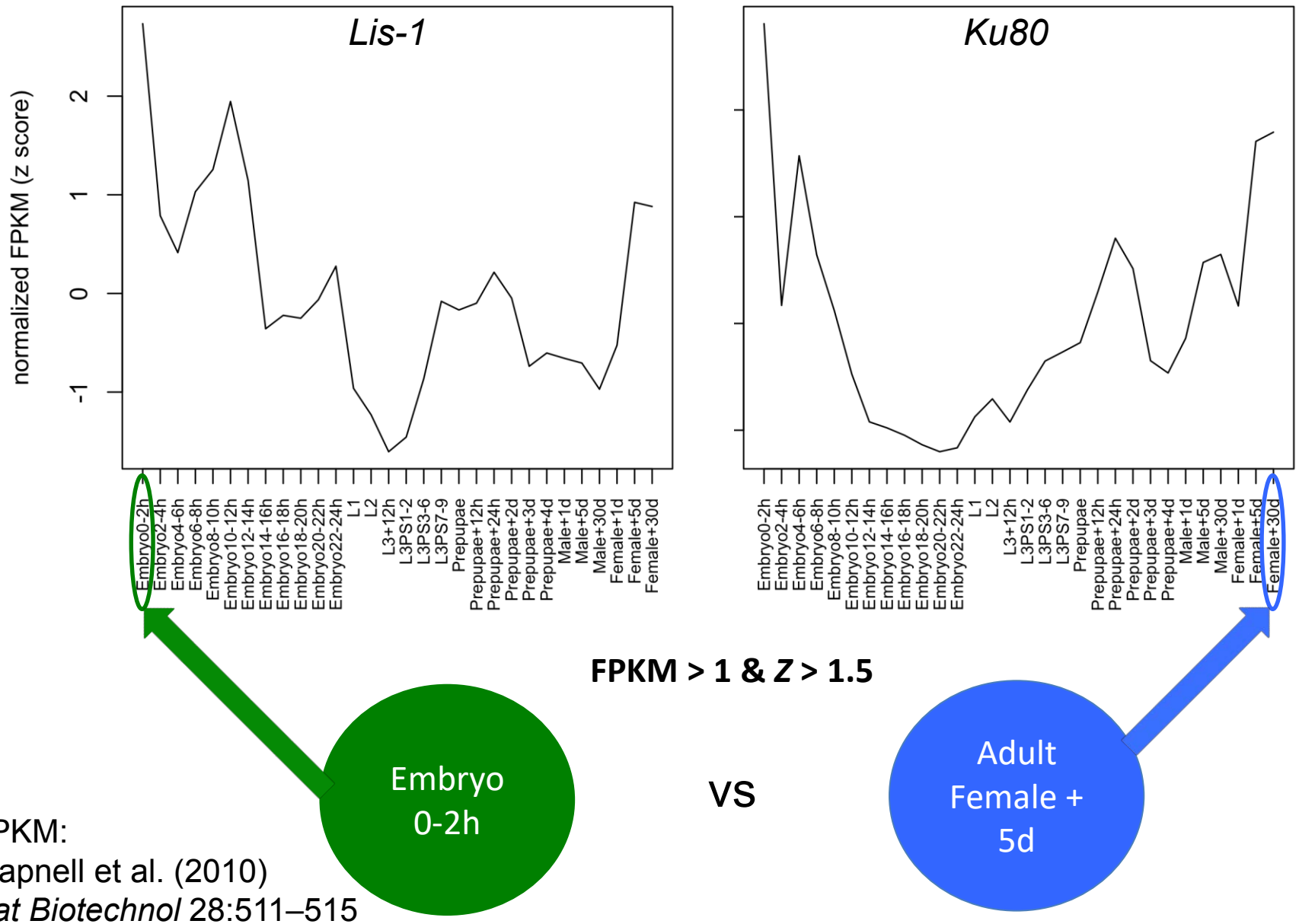
Why?

Clear mapping?

~~Better, but still NO.~~

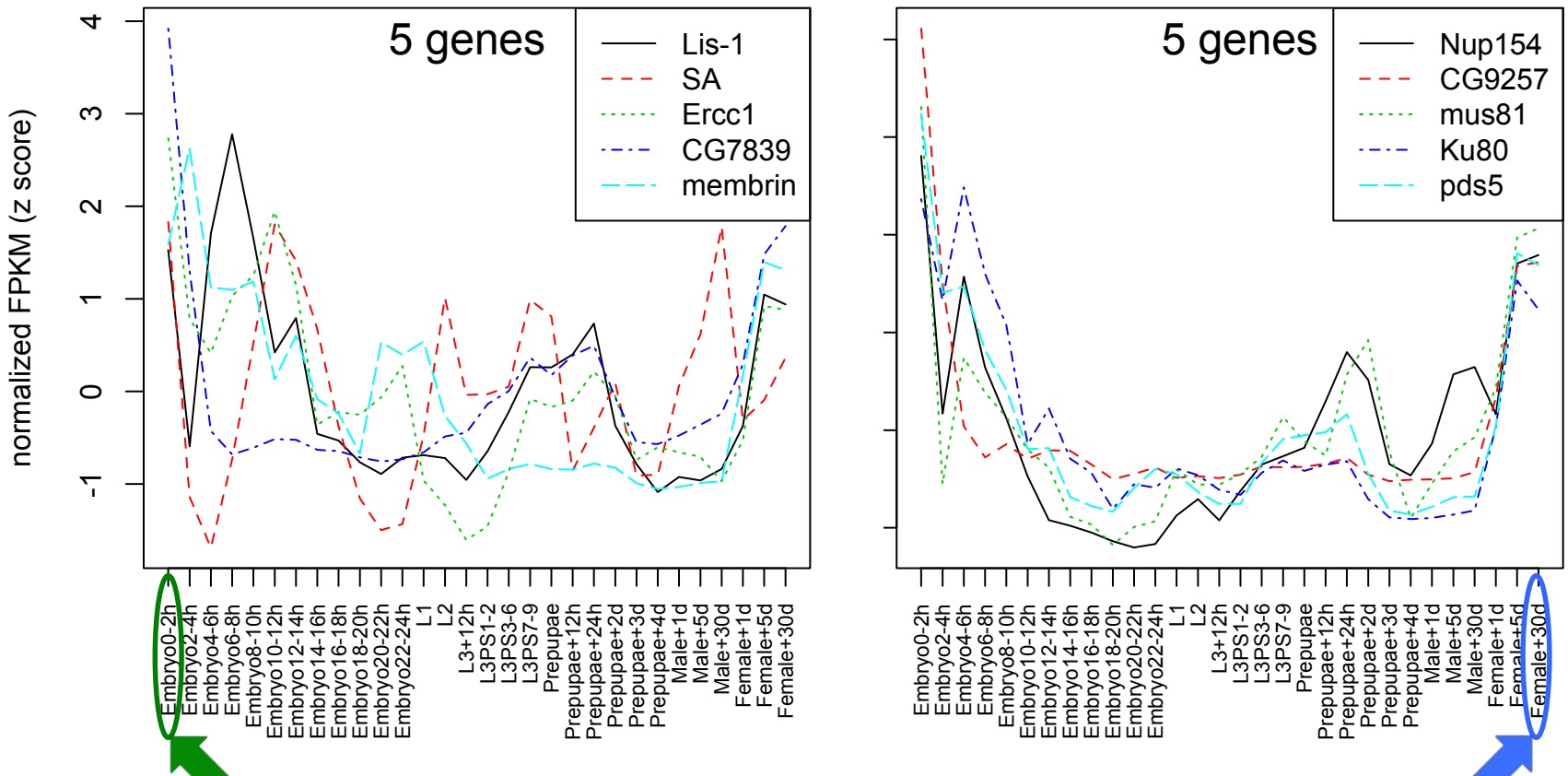
House-keeping genes

Stage **associated** genes



FPKM:
Trapnell et al. (2010)
Nat Biotechnol 28:511–515

Stage **associated** genes



FPKM > 1 & *Z* > 1.5

Embryo 0-2h

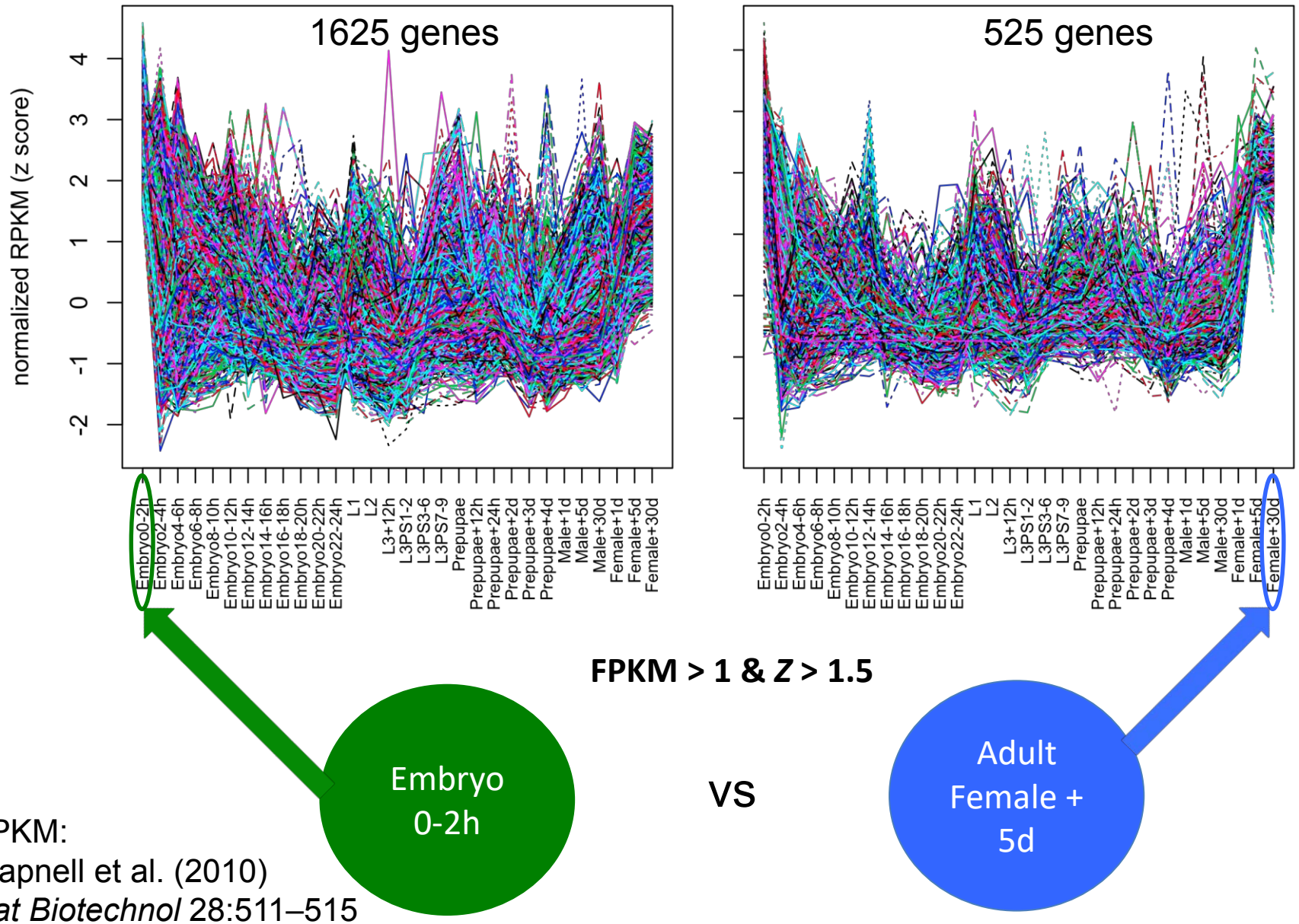
vs

Adult Female + 5d

FPKM:
Trapnell et al. (2010)
Nat Biotechnol 28:511–515

NOT specific

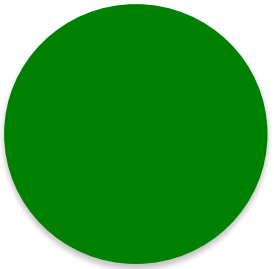
Stage **associated** genes



FPKM:
Trapnell et al. (2010)
Nat Biotechnol 28:511–515

“Associated” vs. “Specific”

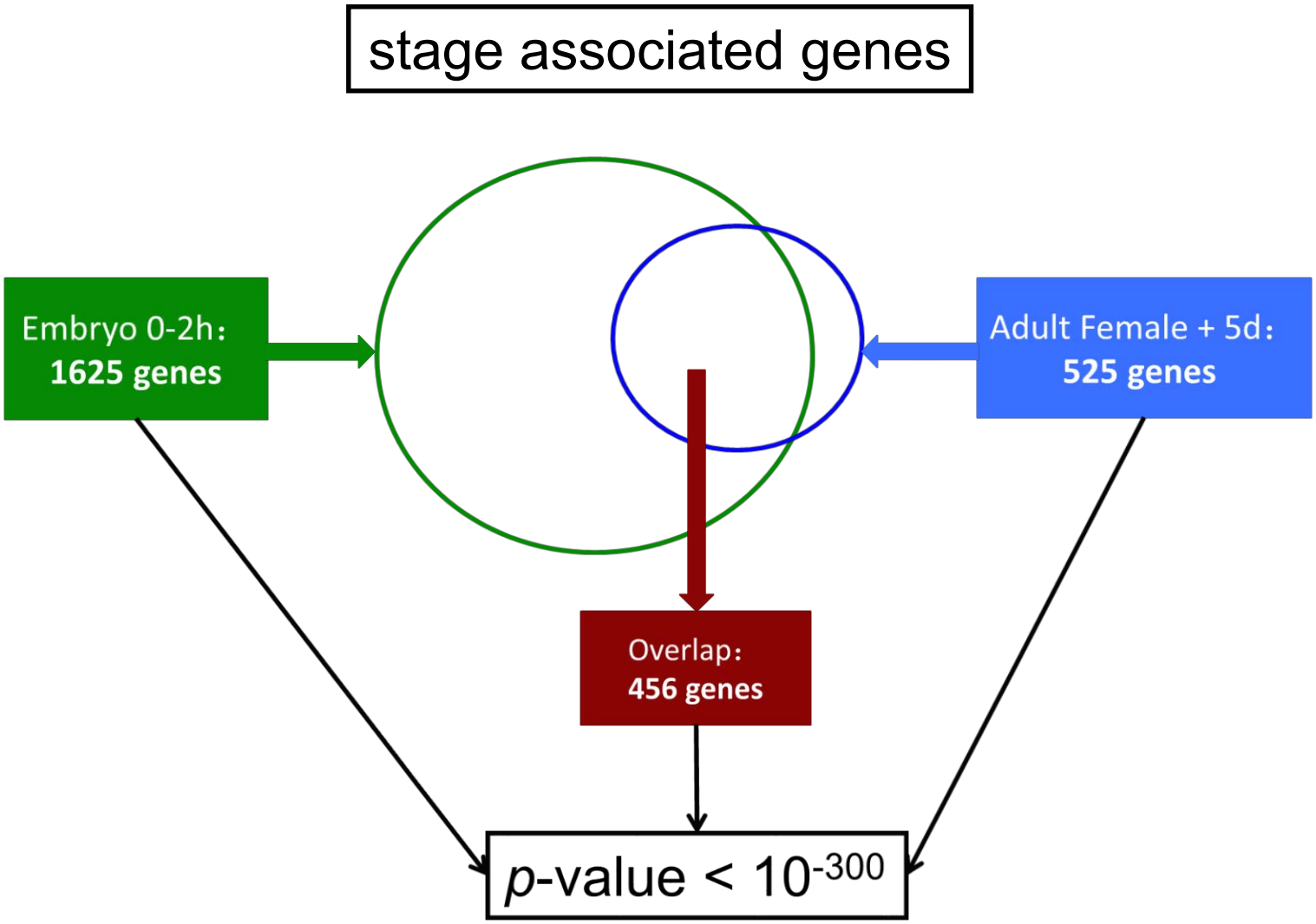




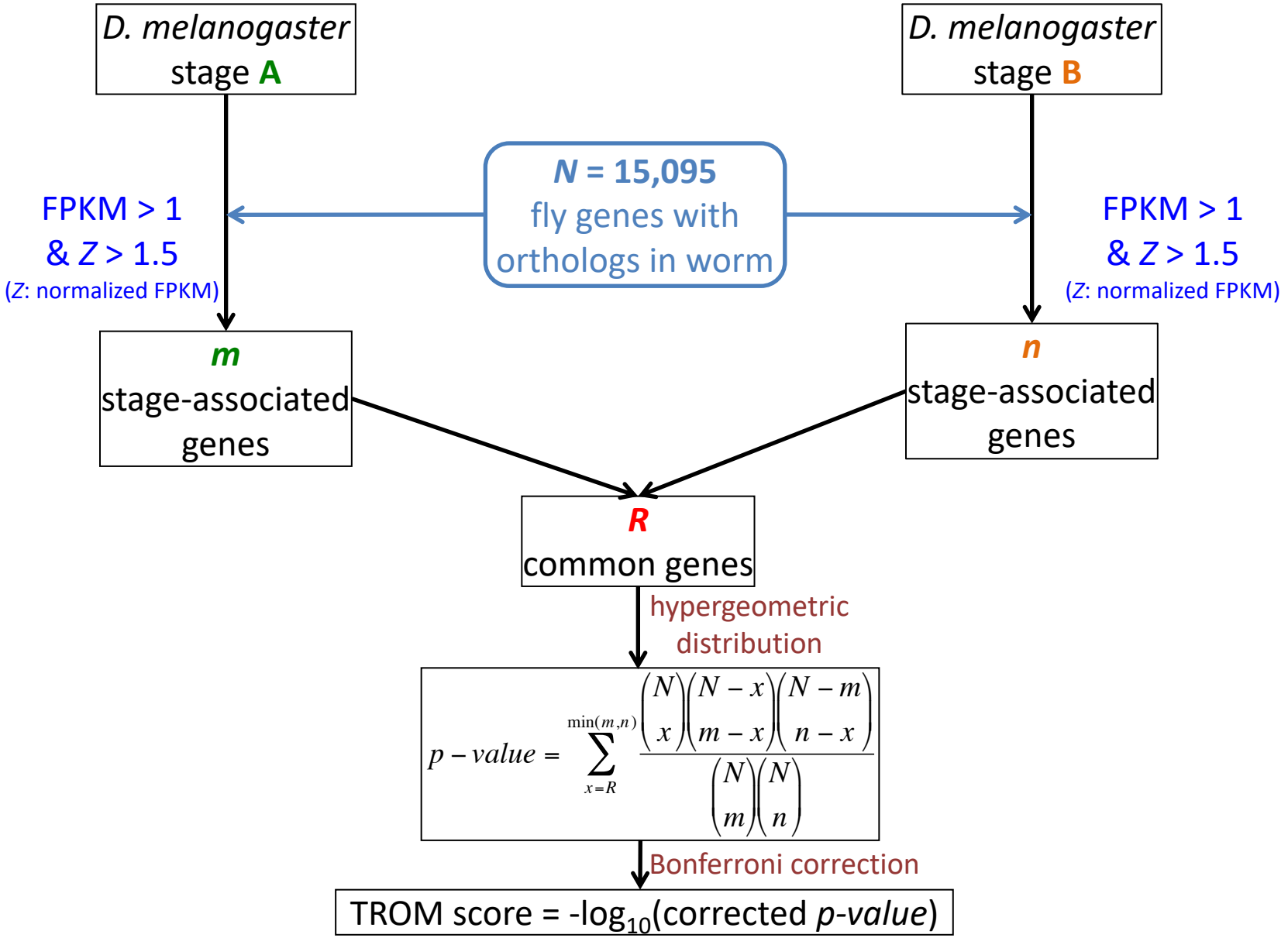
Specific:	Red	Blue	Green Circle
Associated:	Red Square	Blue Square	Green Circle

- Associated genes can be shared by a subset of samples
- # associated genes ≥ # specific genes

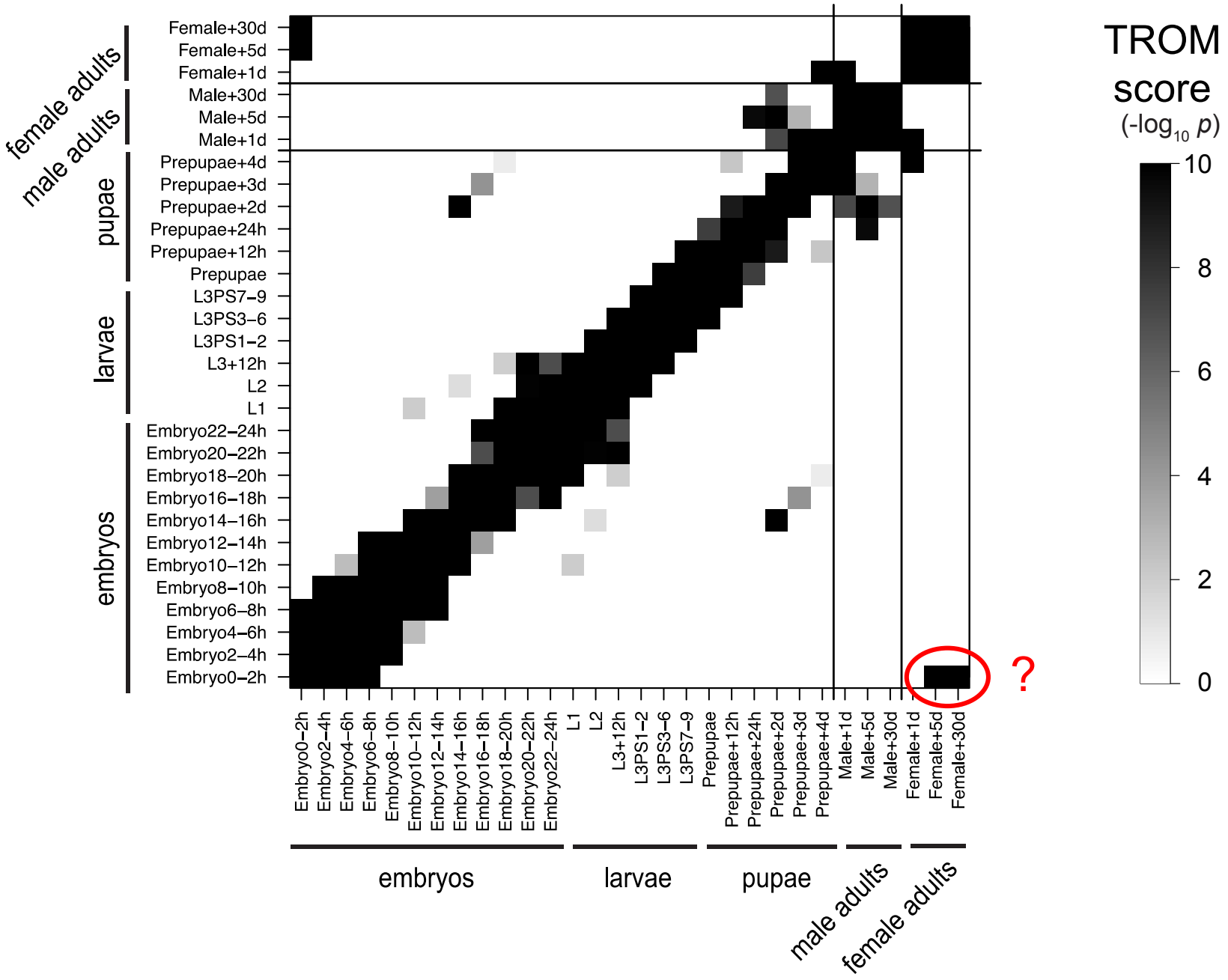
Within-species stage mapping



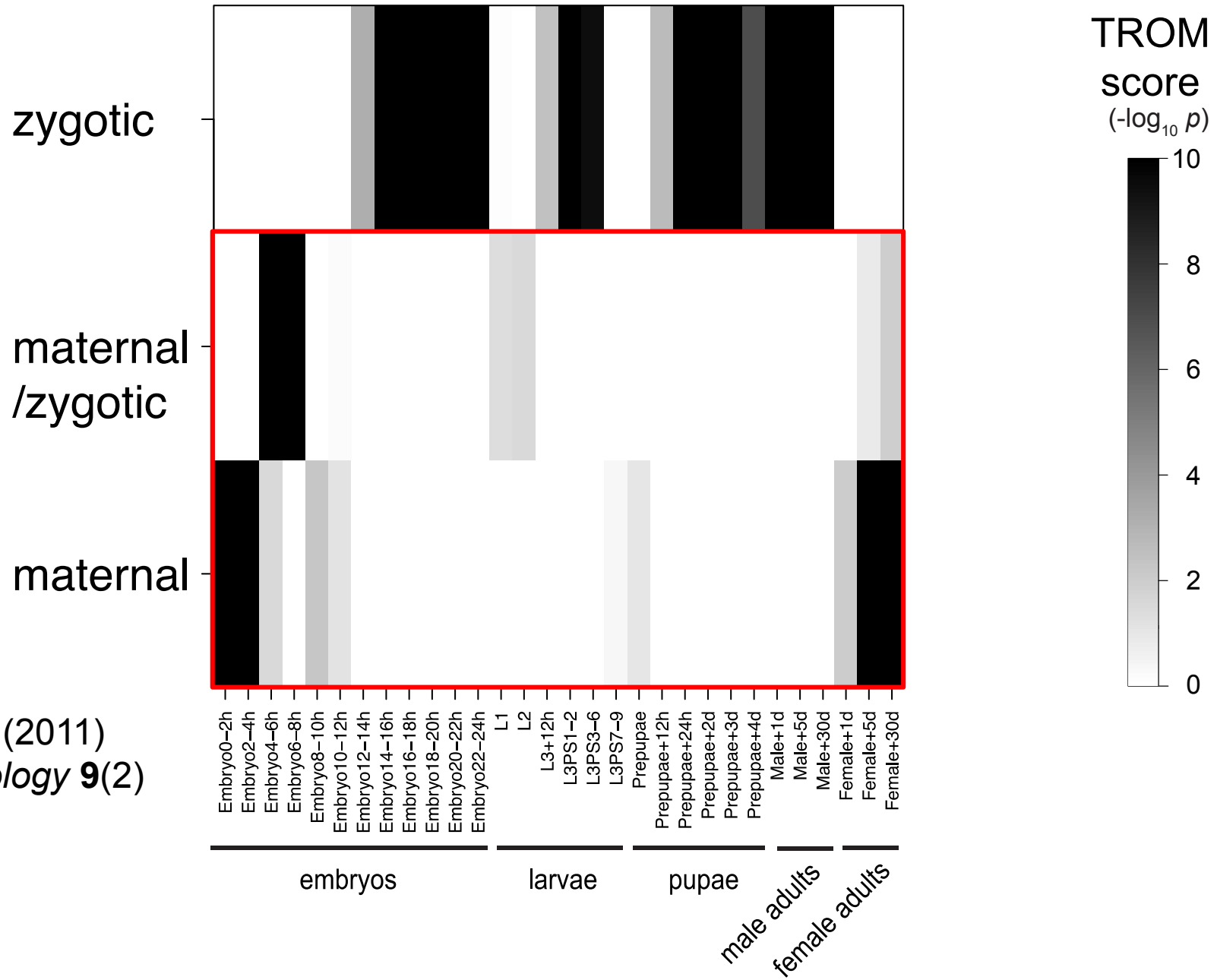
Stage/tissue/cell mapping within species (e.g. fly) (explanation of *p*-values)



D. melanogaster stage mapping results

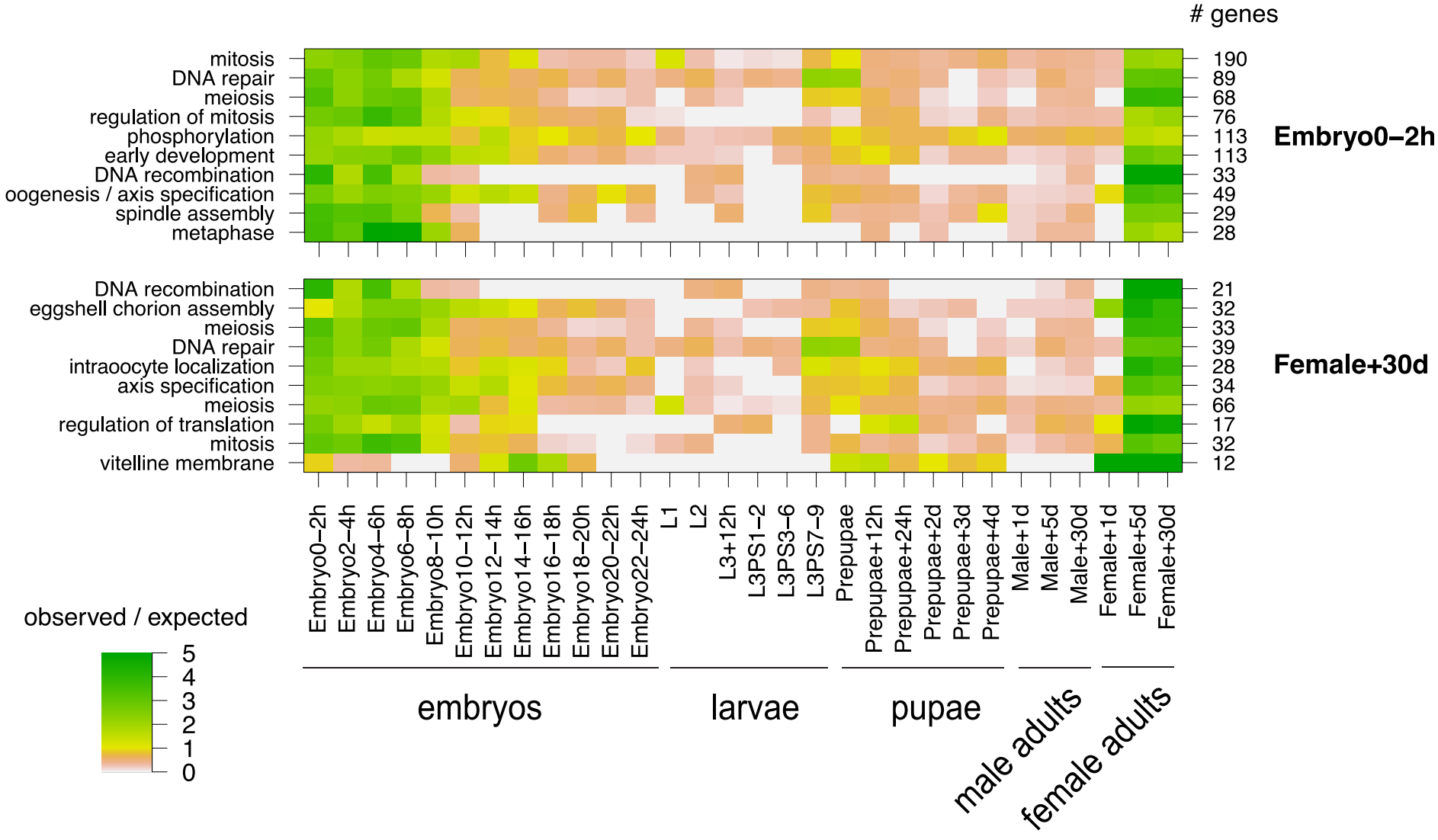


D. melanogaster stages vs. gene classes



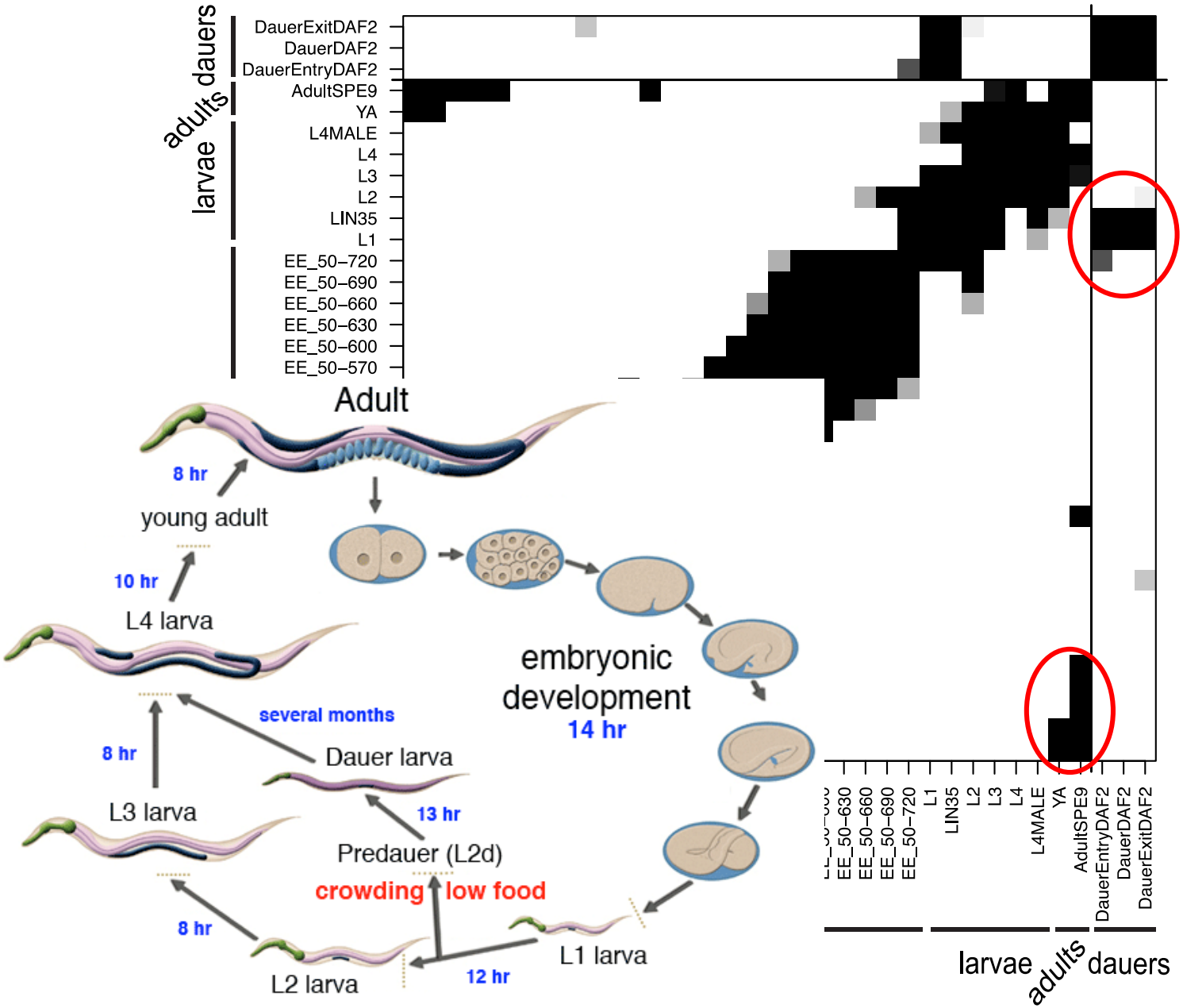
Lott el al (2011)
PLoS biology **9**(2)

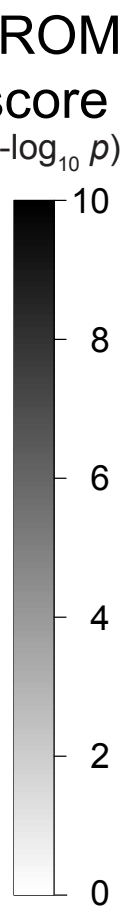
Top enriched GO terms



D. melanogaster stages

C. elegans stage mapping results

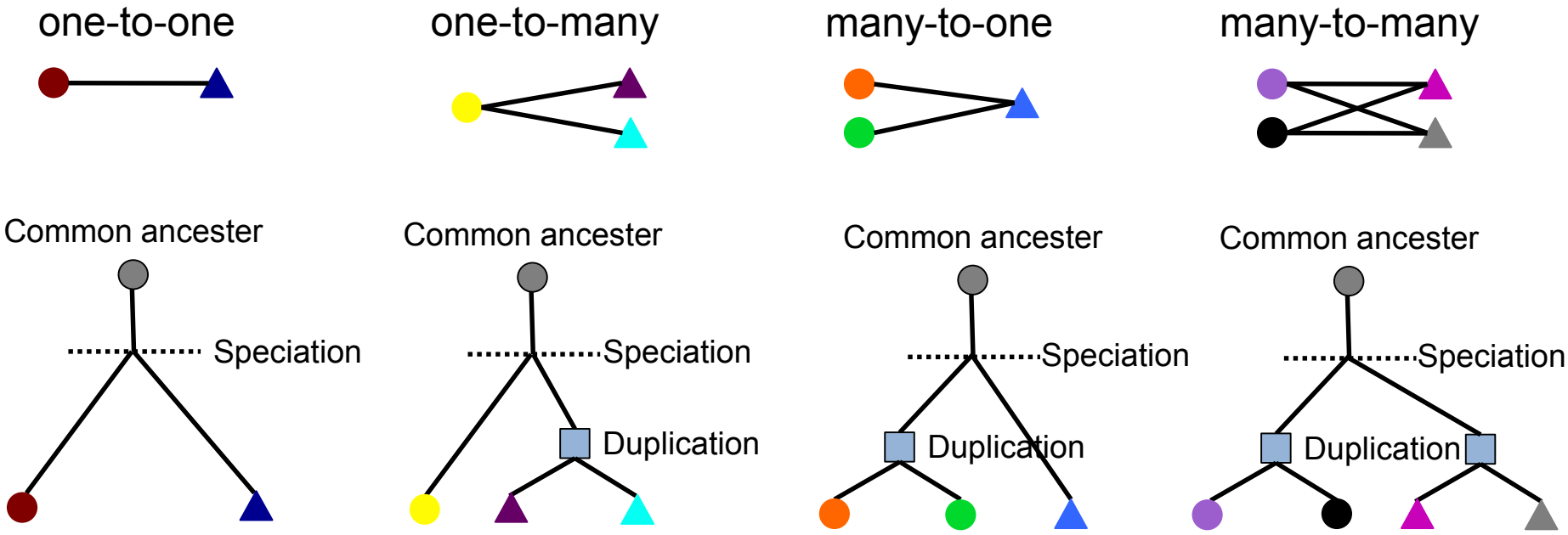




Question 2

- Between fly and worm, is there a clear mapping between their developmental stages in terms of orthologous gene expression?

Orthologs

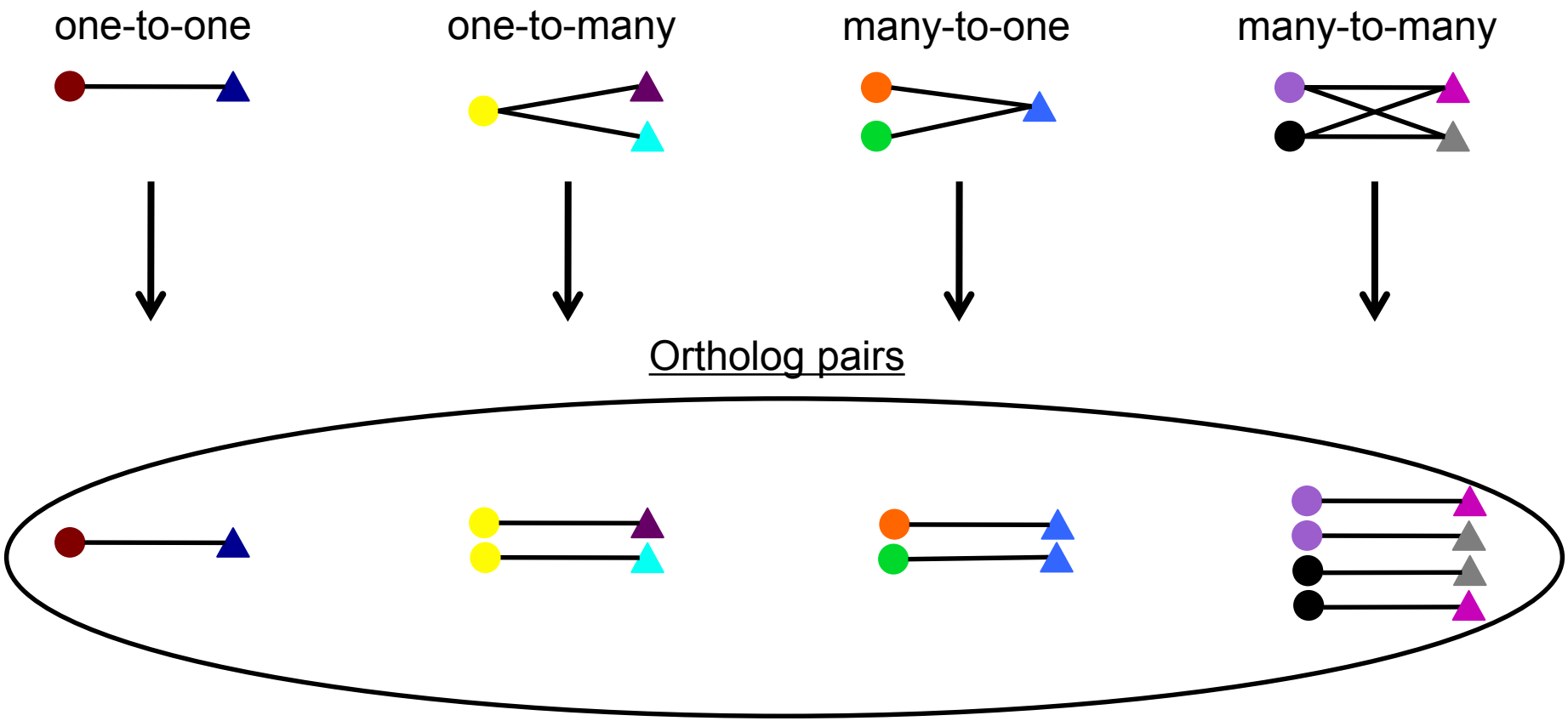


modENCODE orthologs:
Wu el al (2014) *bioRxiv* doi: http://dx.doi.org/10.1101/005736.

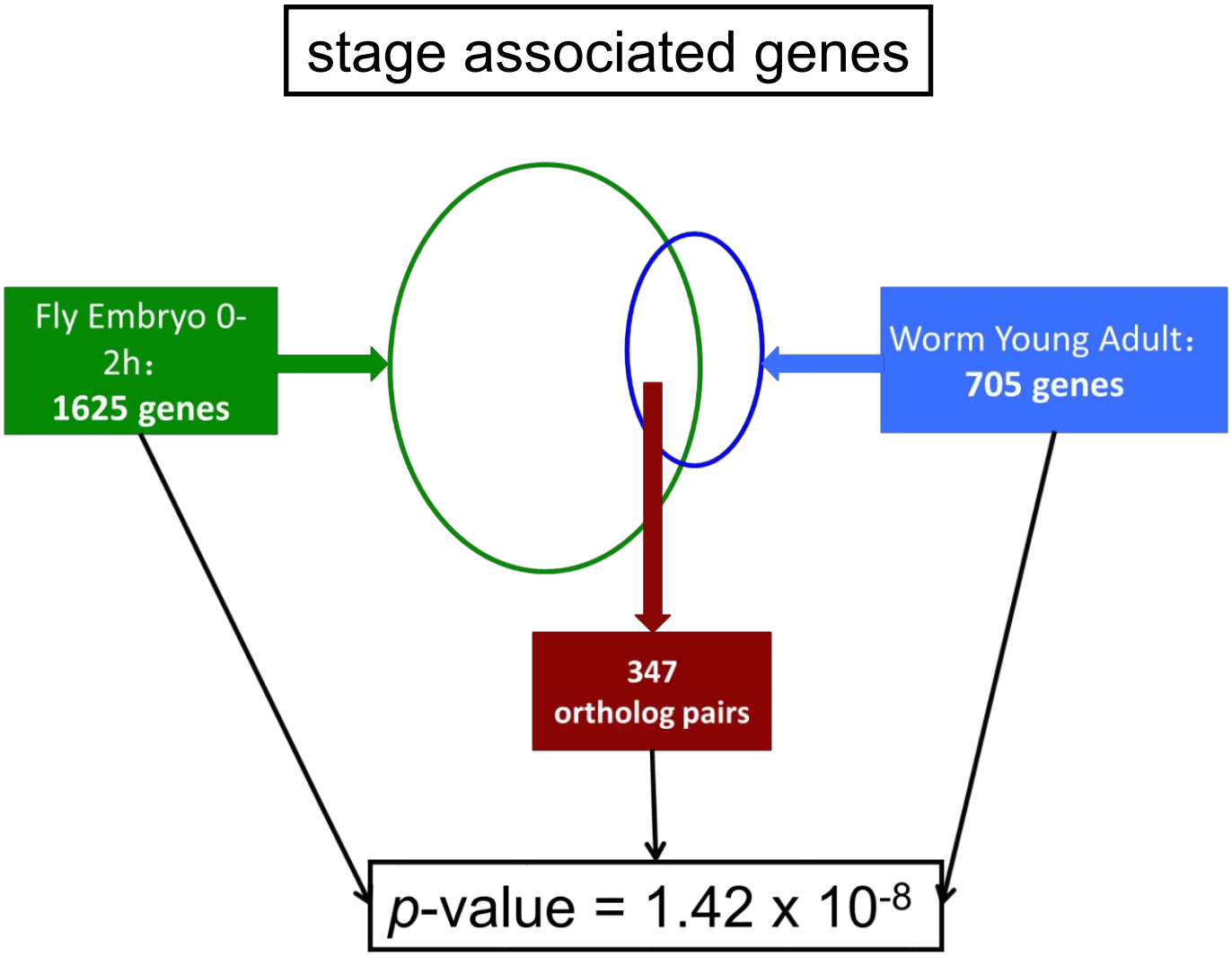
Question 2

- Between fly and worm, is there a clear mapping between their developmental stages in terms of orthologous gene expression?

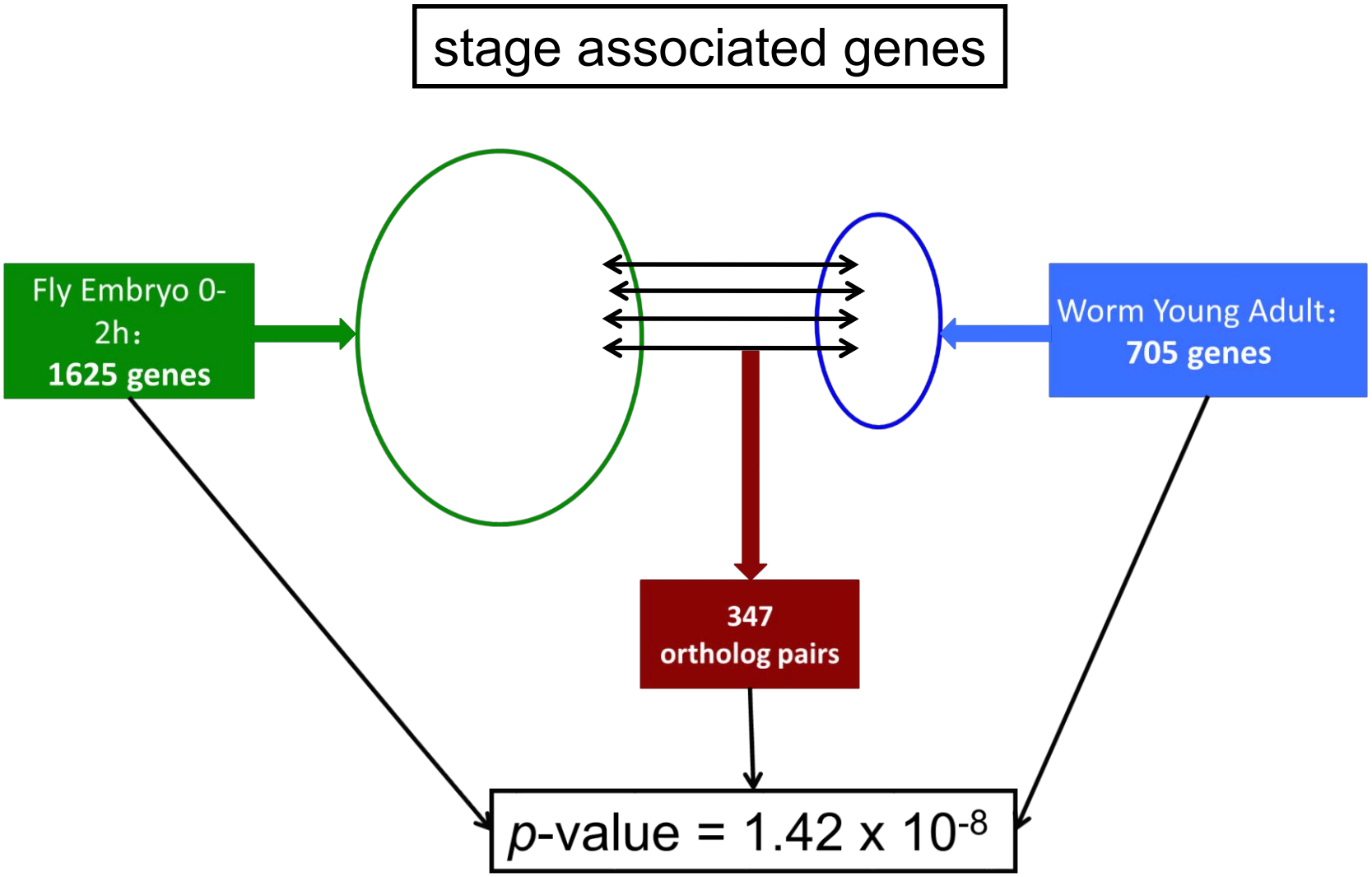
Orthologs



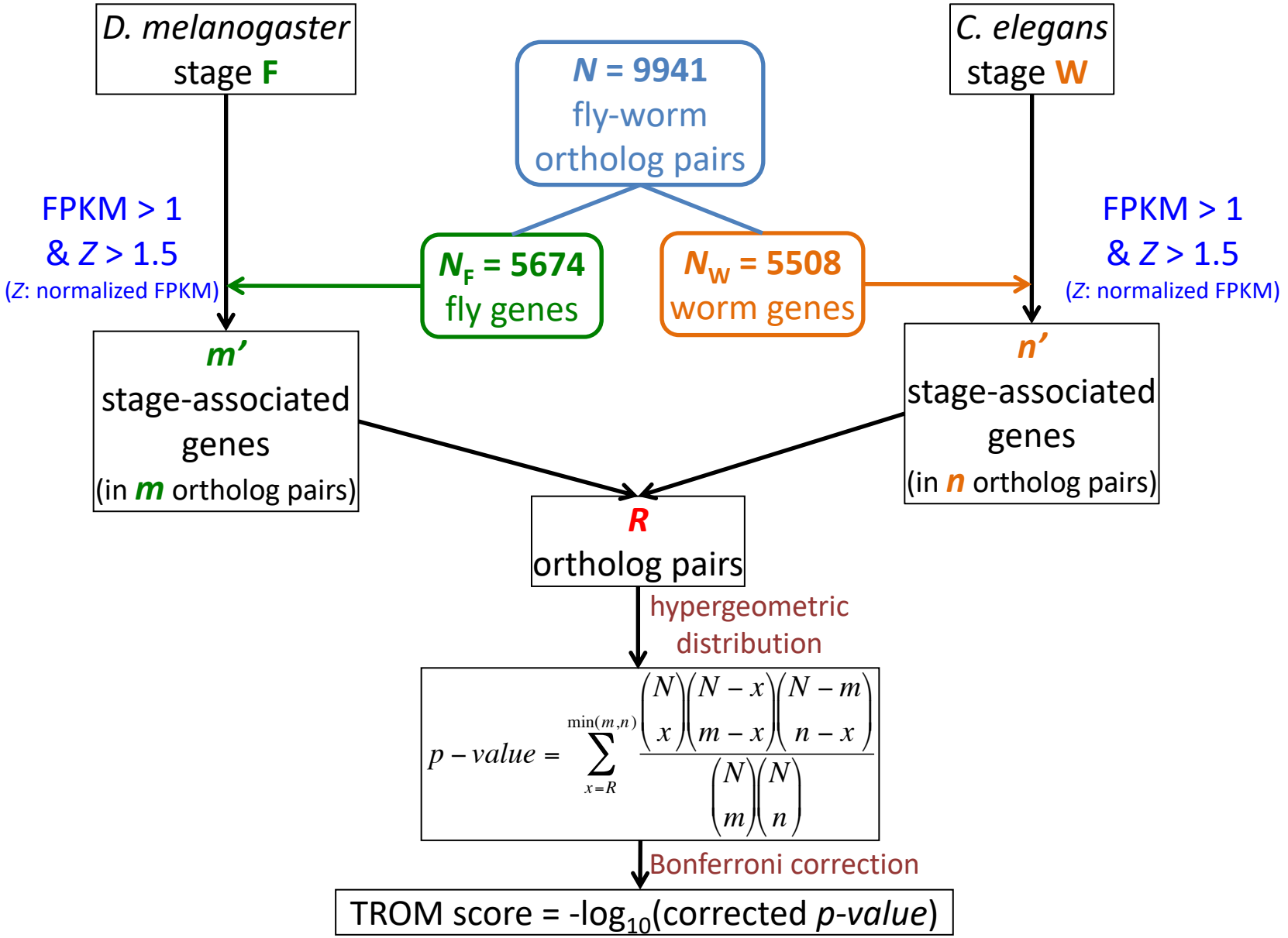
Between-species stage mapping



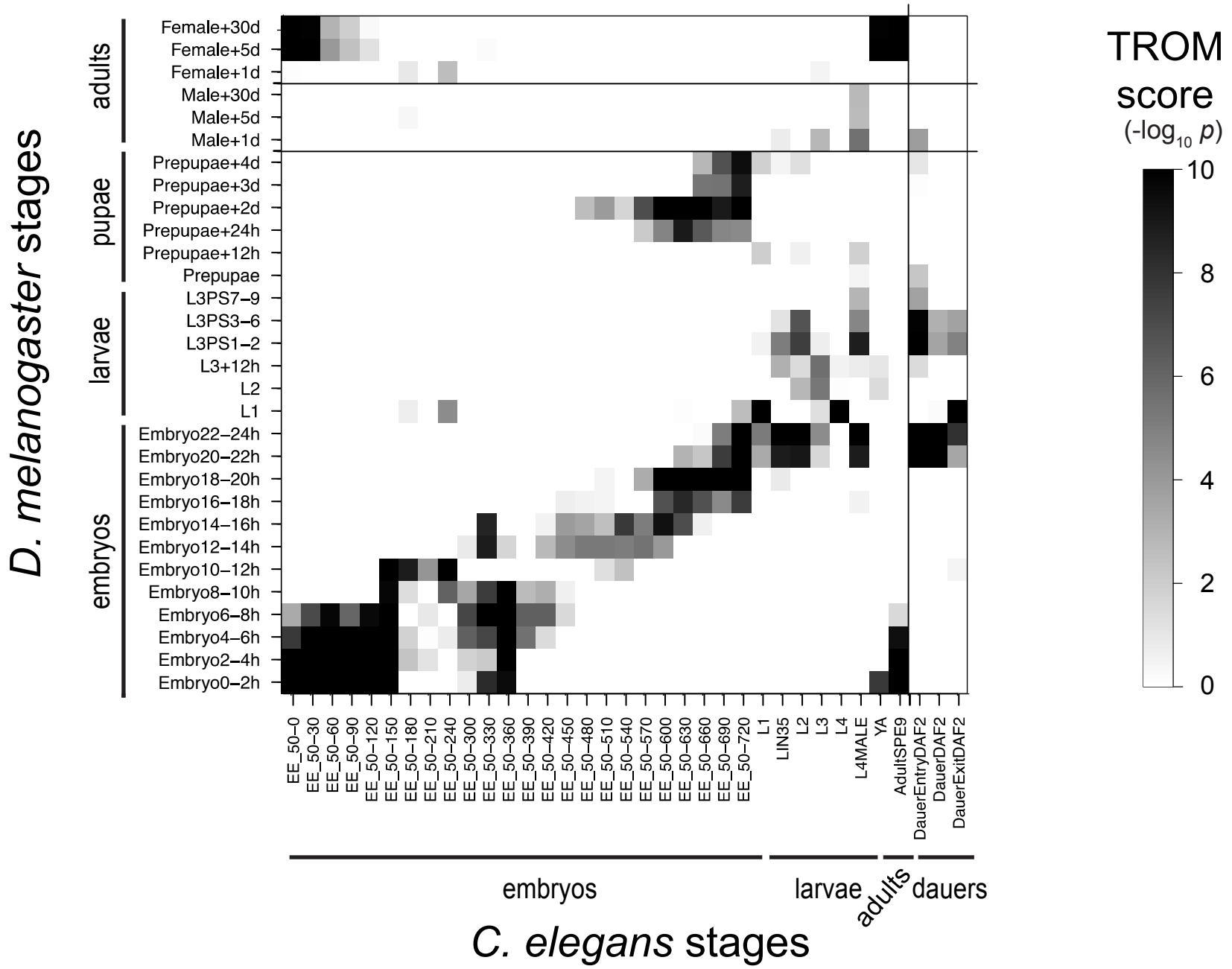
Between-species stage mapping



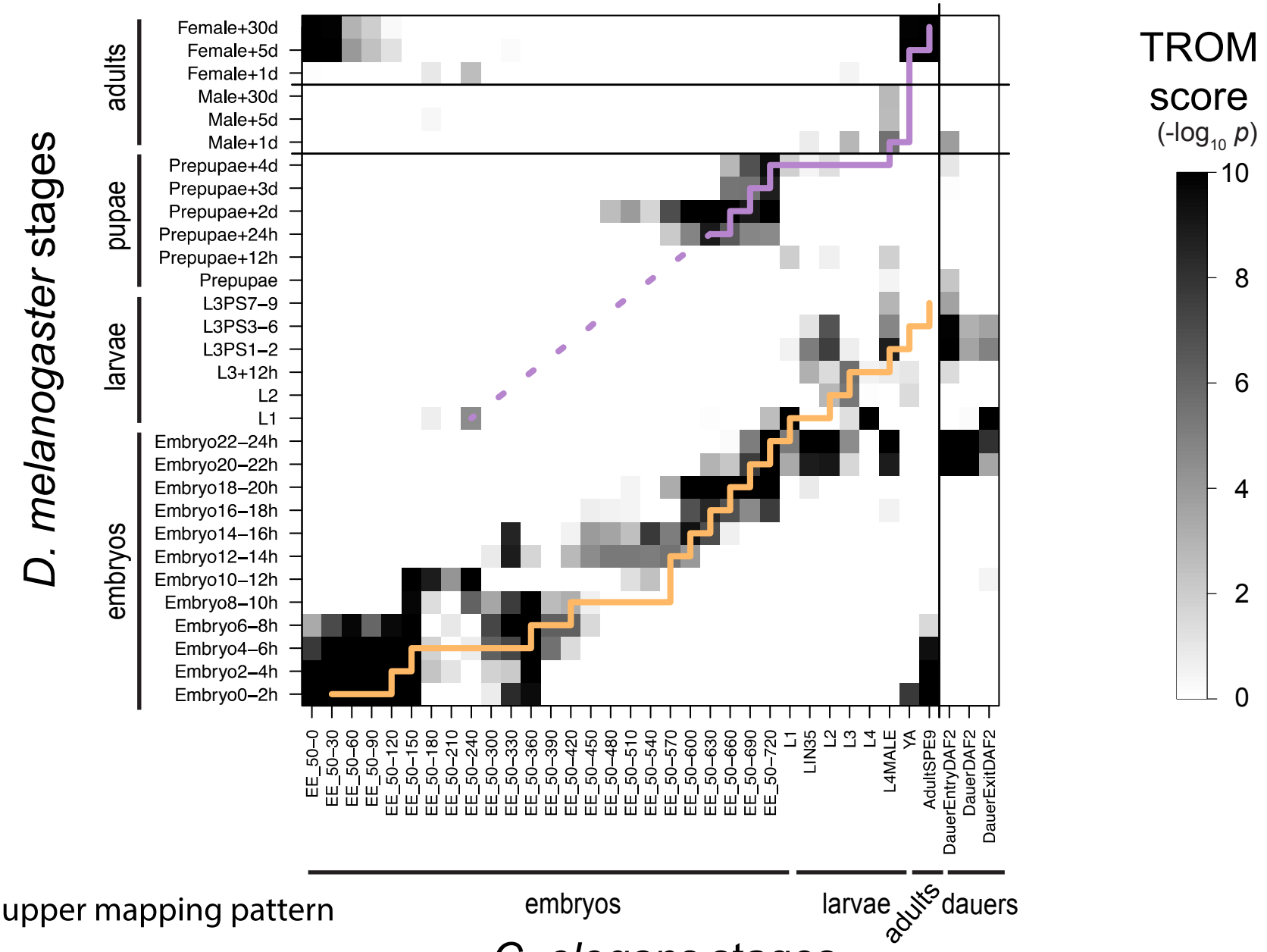
Stage/tissue/cell mapping between fly and worm (explanation of *p*-values)



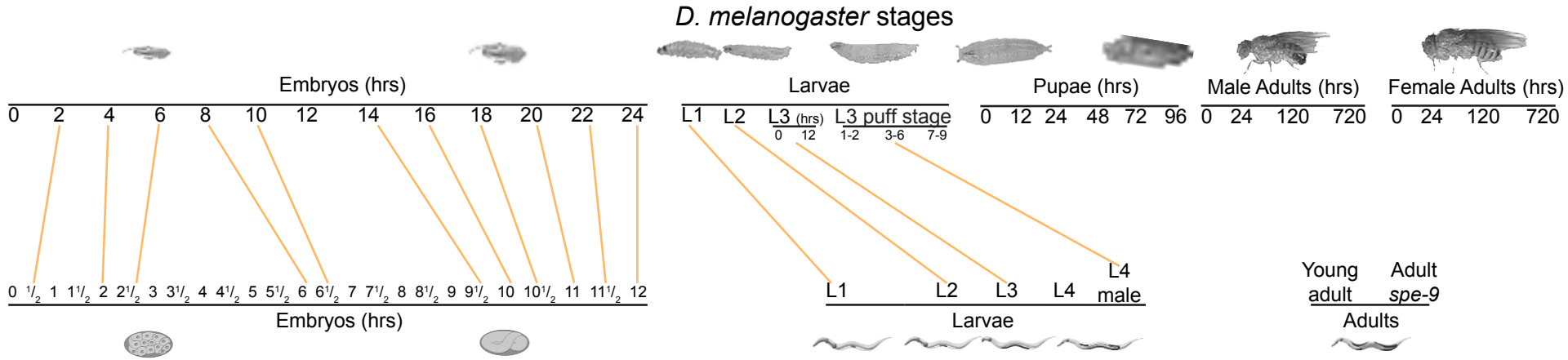
C. elegans vs. *D.melanogaster* stage mapping results



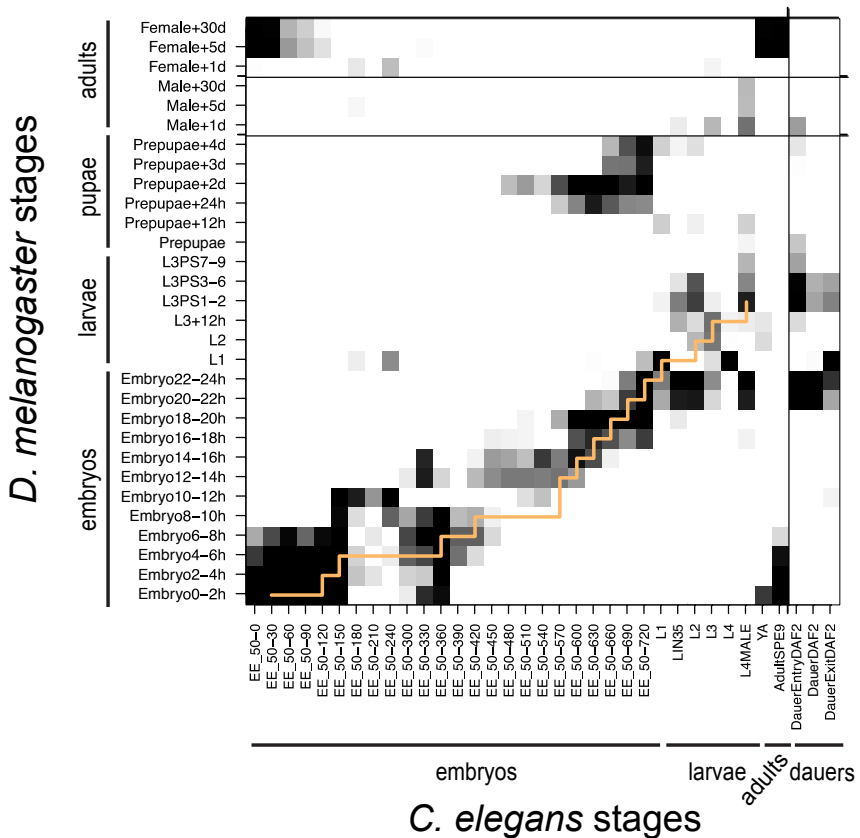
C. elegans vs. *D.melanogaster* stage mapping results



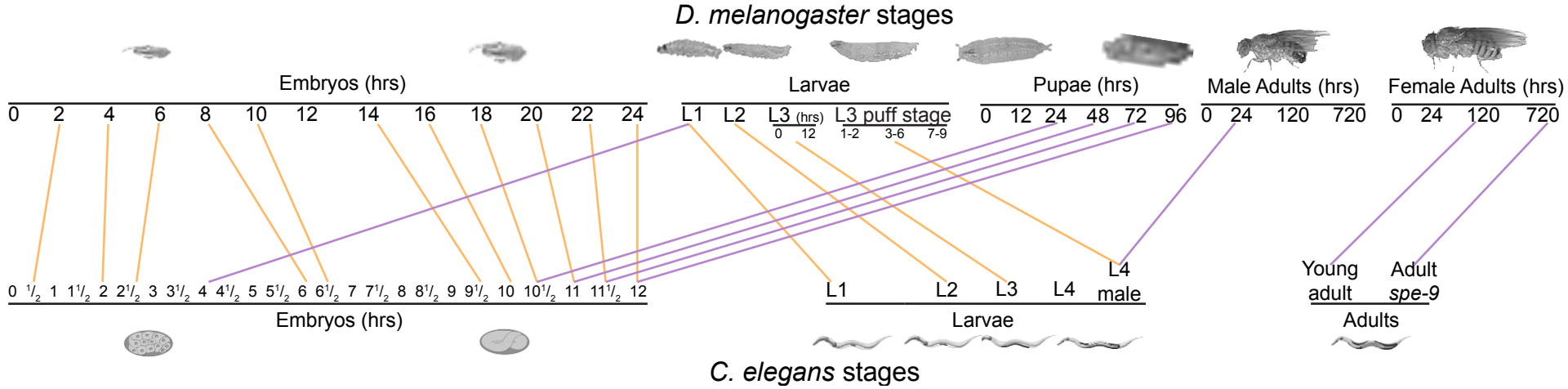
C. elegans vs. *D.melanogaster* stage mapping results

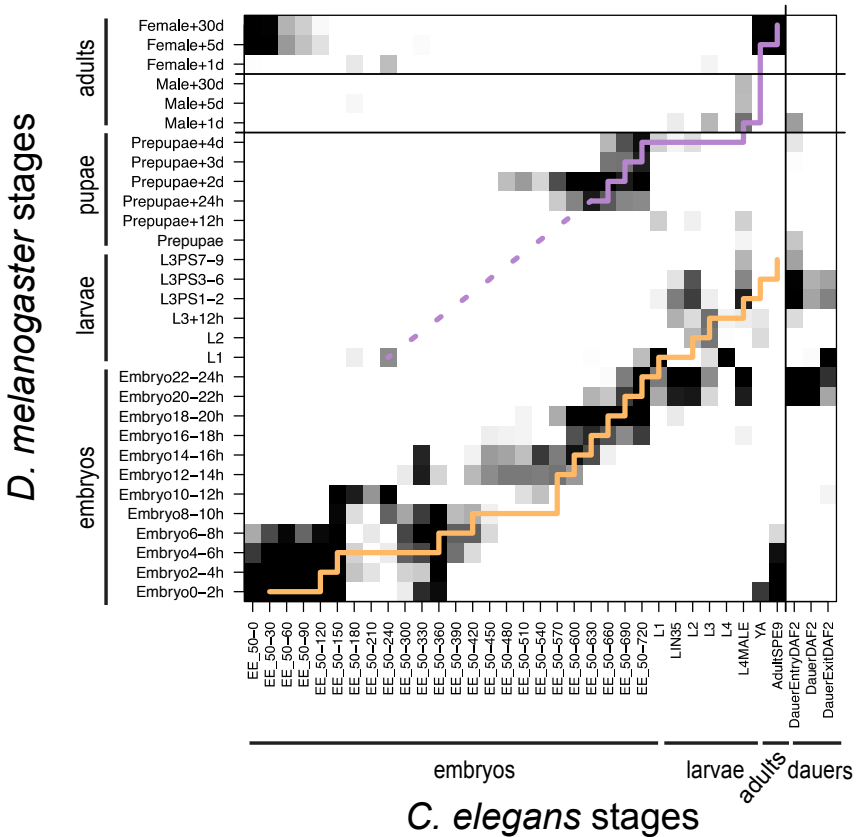


C. elegans stages



C. elegans vs. *D.melanogaster* stage mapping results





Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data

- Li, J.J., Huang, H., Bickel, P.B., and Brenner, S.E. (2014). ***Genome Research*** 24(7):1086-1101.

LETTER

OPEN

doi:10.1038/nature13424

Comparative analysis of the transcriptome across distant species

- Gerstein, M.B. et al. (2014). ***Nature*** 512(7515):445-448.

Acknowledgements

- Computational analysis
  - Haiyan Huang, Peter Bickel, and Steven Brenner (UC Berkeley)


Haiyan Huang


Peter Bickel


Steven Brenner

  - Dat Duong (previous MS student at UC Berkeley, currently at UCLA)
  - Mark Biggin (LBNL)
- Fly RNA-Seq data
  - Sue Celniker, Roger Hoskins (LBNL)
- Worm RNA-Seq data
  - Robert Waterston (U Washington)
  - LaDeana Hillier (WUSTL)

Technical aspects of TROM

- Selection of the *Z*-score threshold
- Performance as classification scores

Selection of the *Z*-score threshold

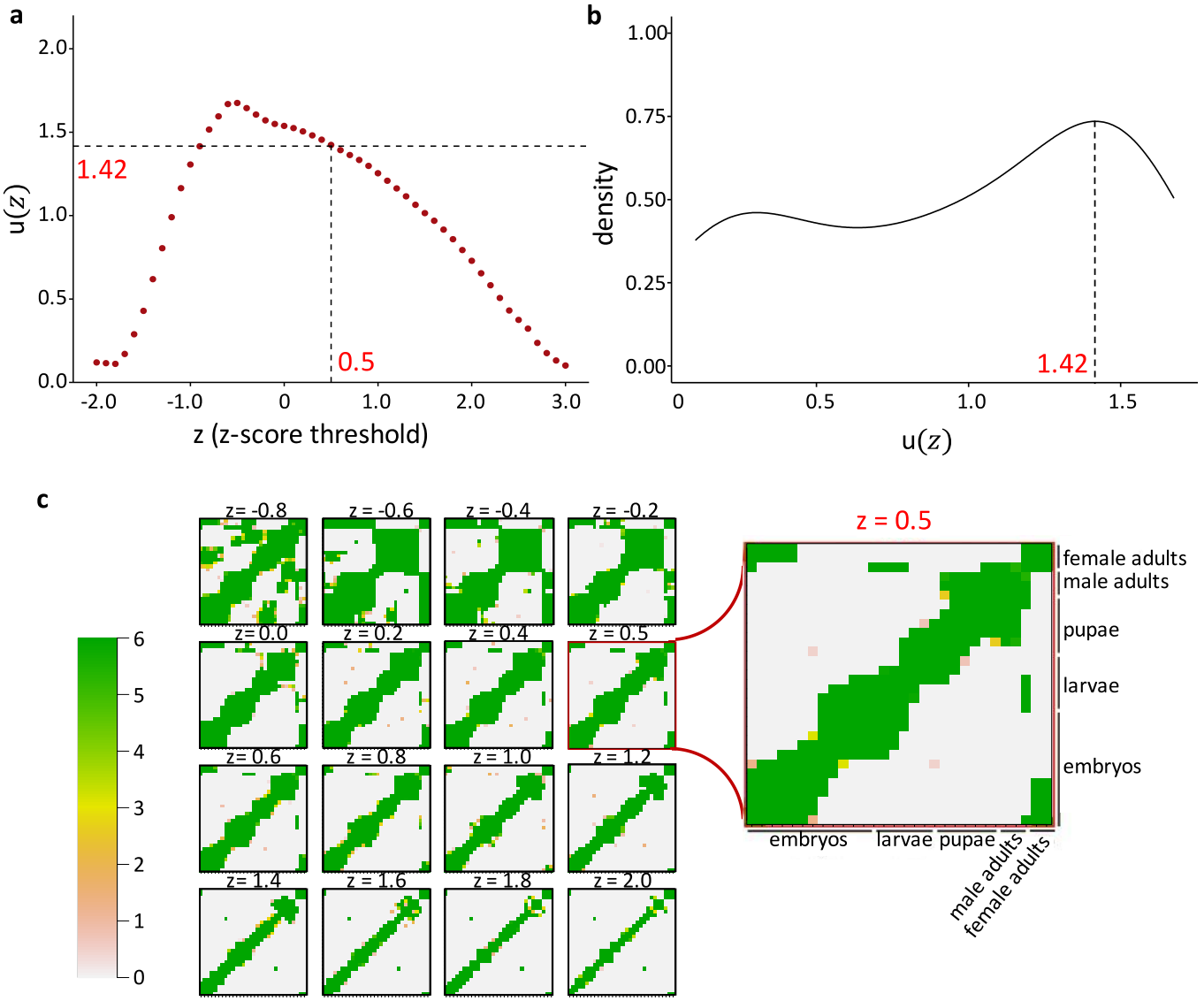
- Objective function

$$u(z) = \log_{10}\left(\frac{\sum_{i=1}^{p}\sum_{j=1,i\neq j}^{p} a_{ij}(z)}{p^2 - p} + 1\right)$$

  - p the number of biological samples,
  - $A(z) = \left(a_{ij}(z)\right)_{p\times p}$ is the TROM matrix based on threshold z.

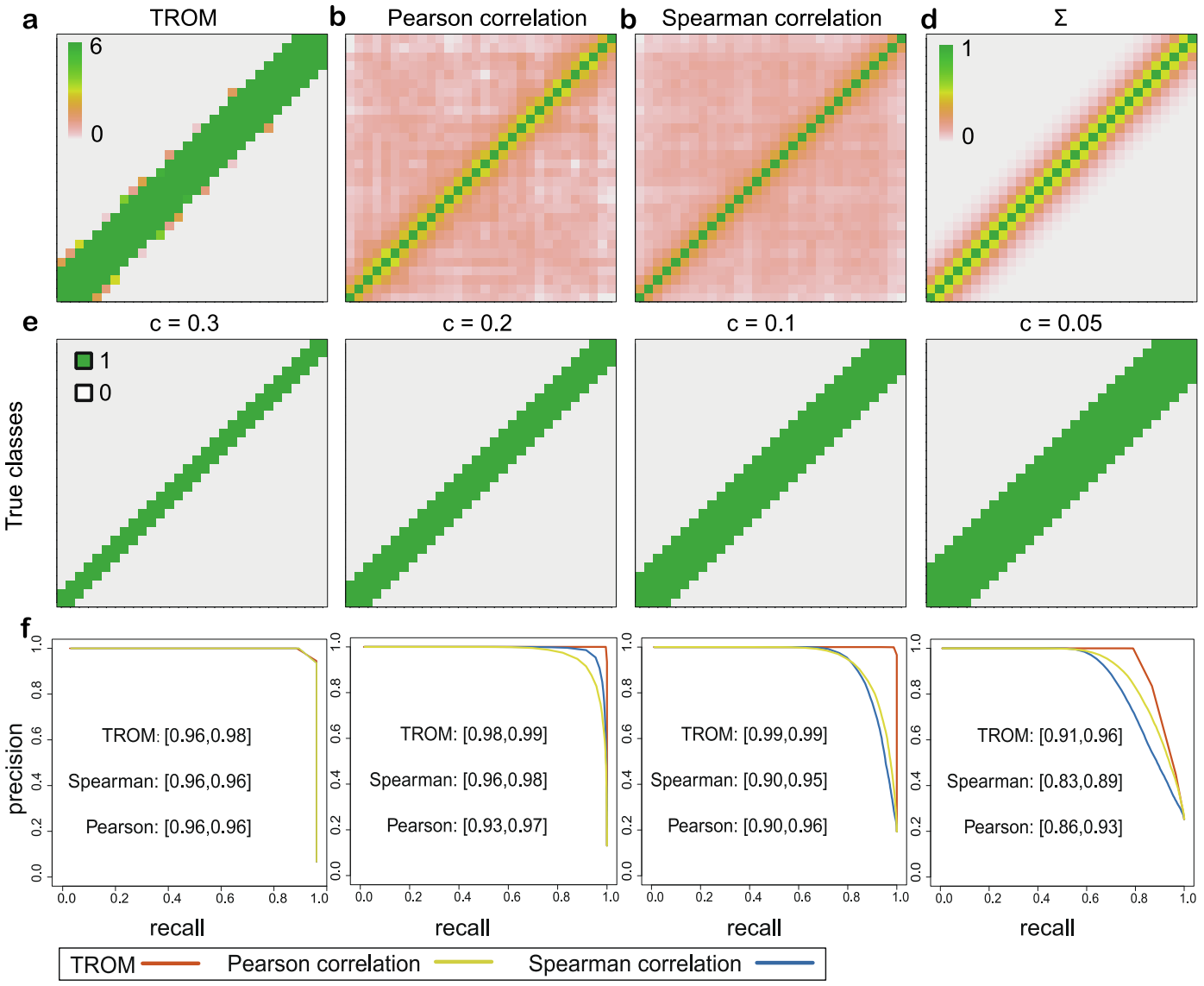
- $\lim_{z\to-\infty} u(z) = 0$ and $\lim_{z\to+\infty} u(z) = 0$.

Selection of the *Z*-score threshold



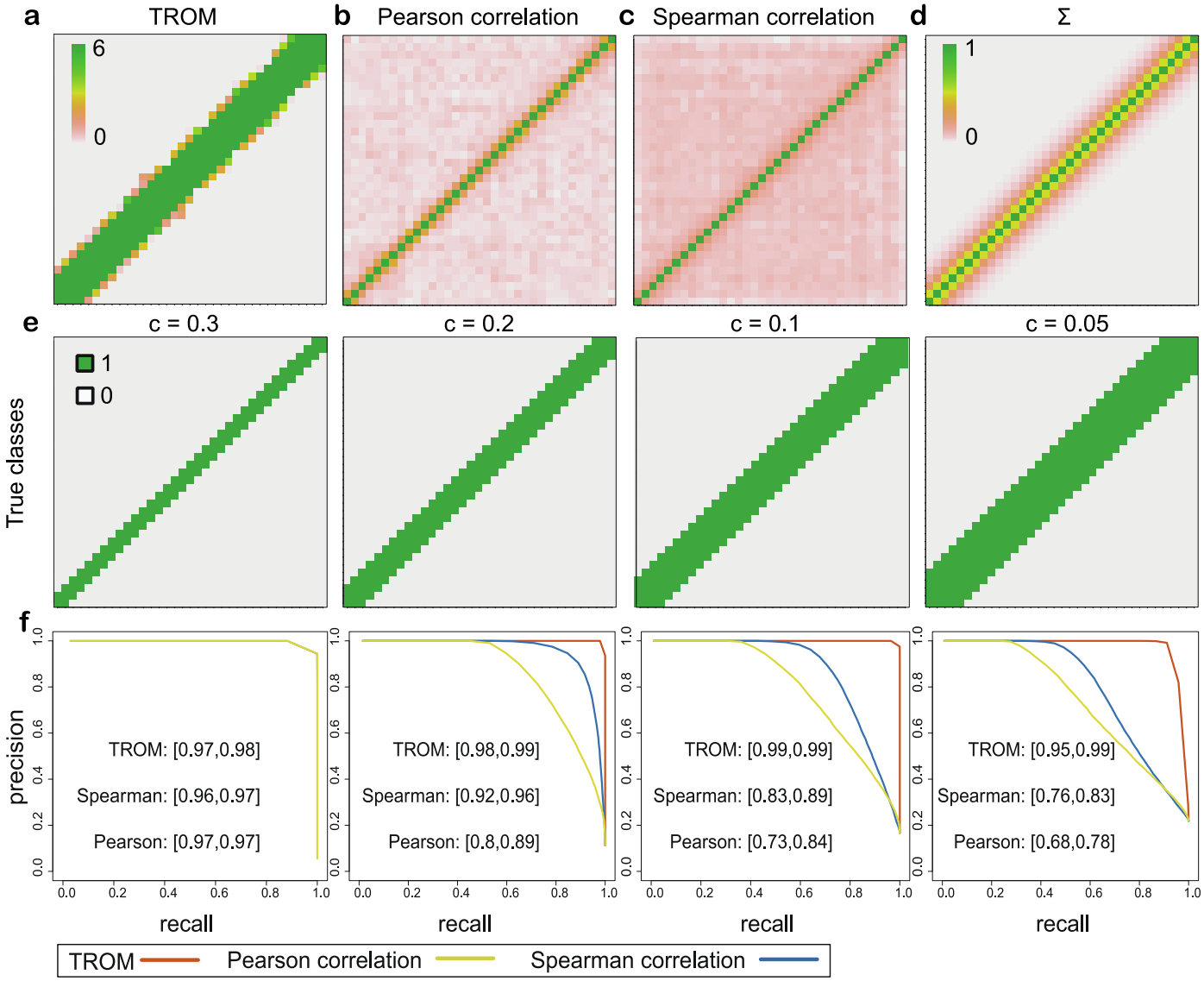
Performance as classification scores

Simulation based on *D. melanogaster* RNA-seq data



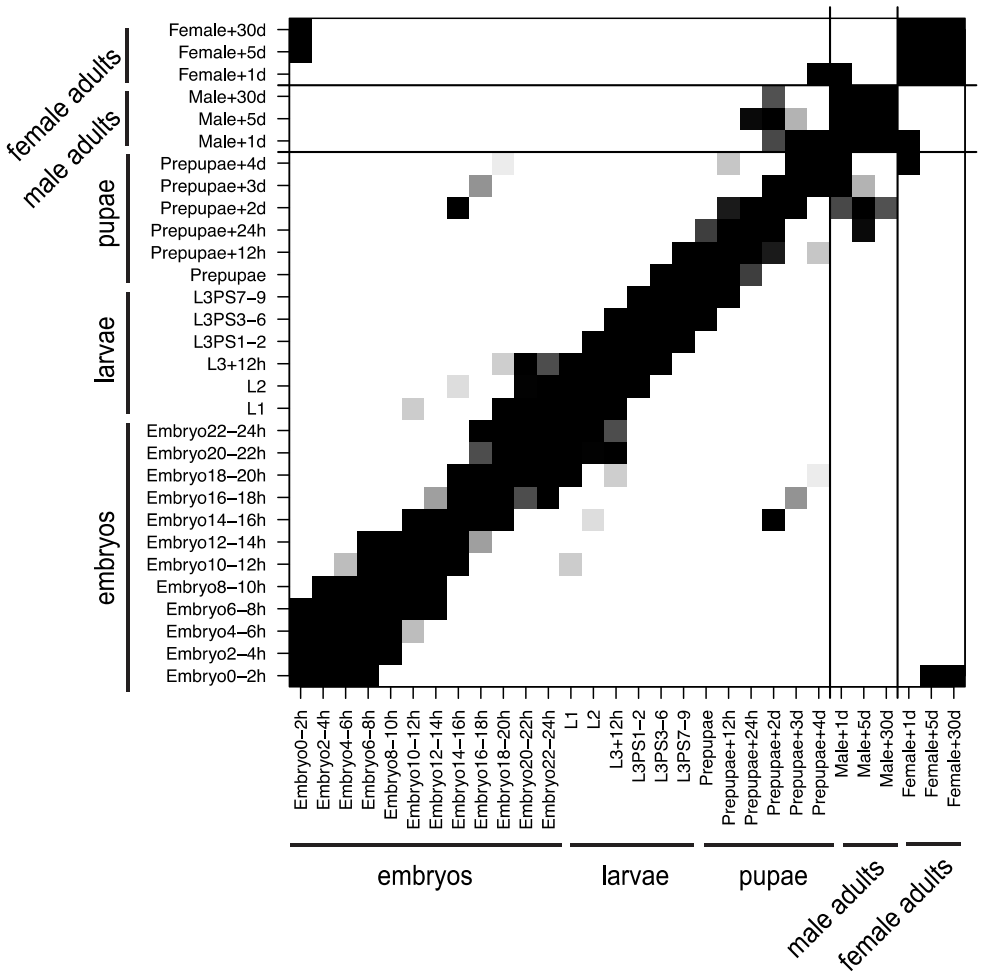
Performance as classification scores

Simulation based on *C. elegans* RNA-seq data

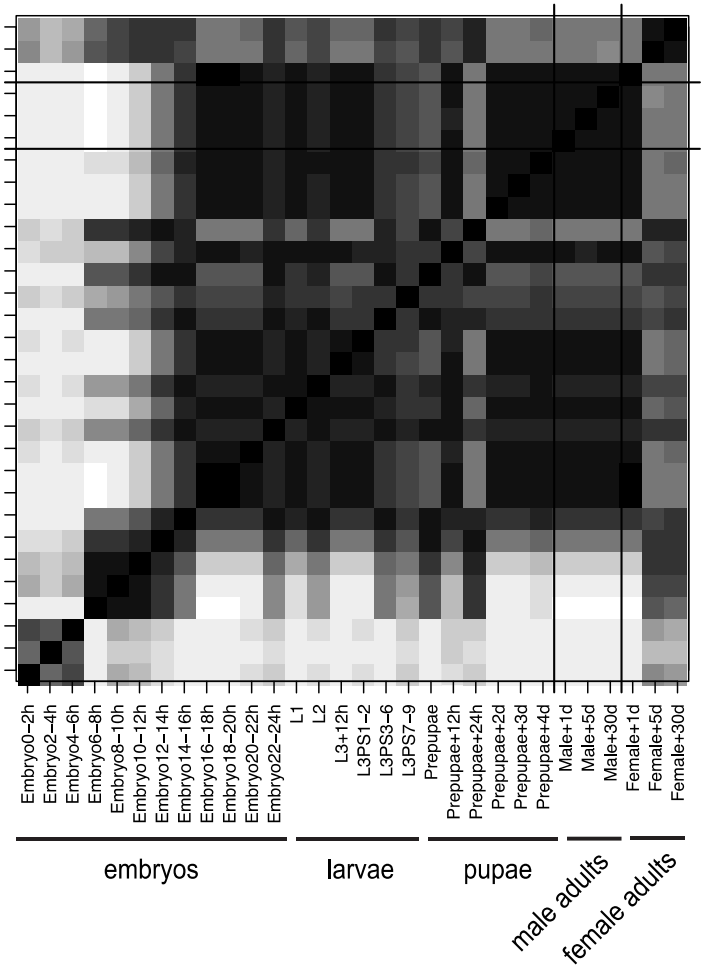


TROM vs. Pearson correlation

D. melanogaster

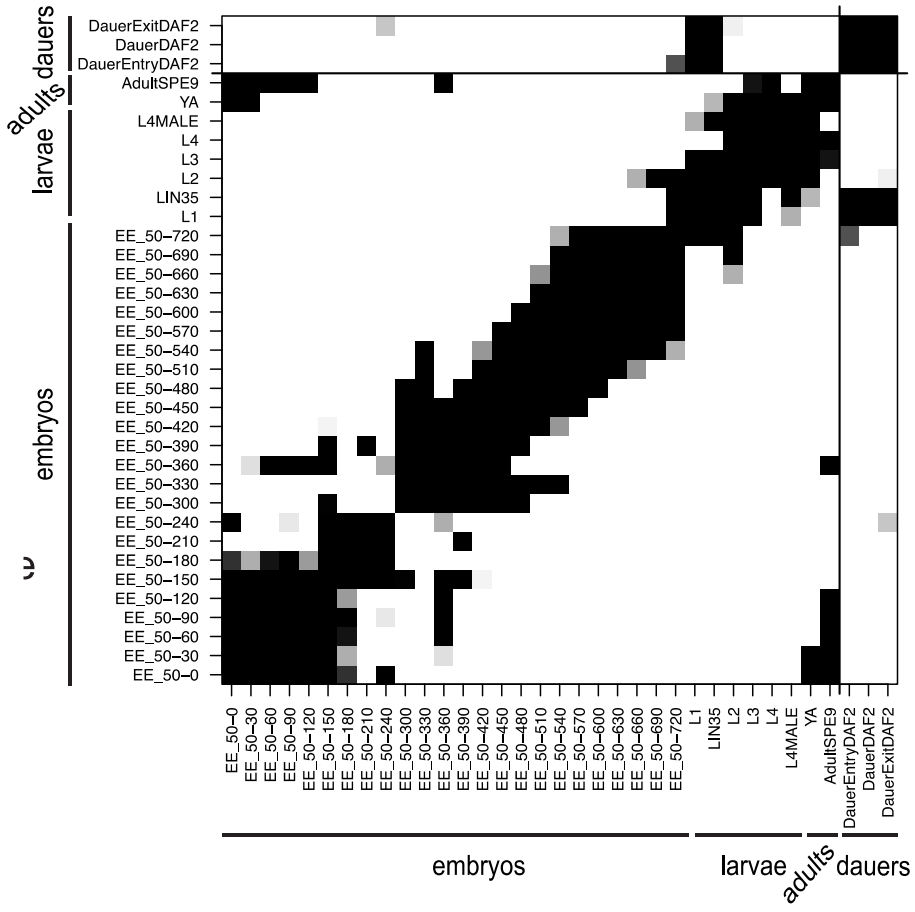


vs.

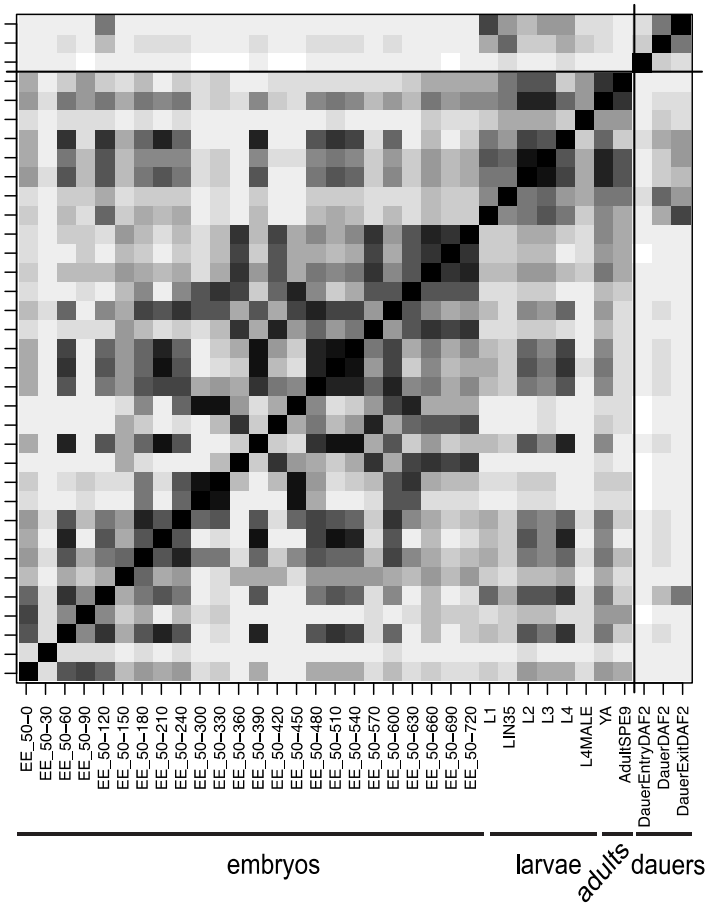


TROM vs. Pearson correlation

C. elegans

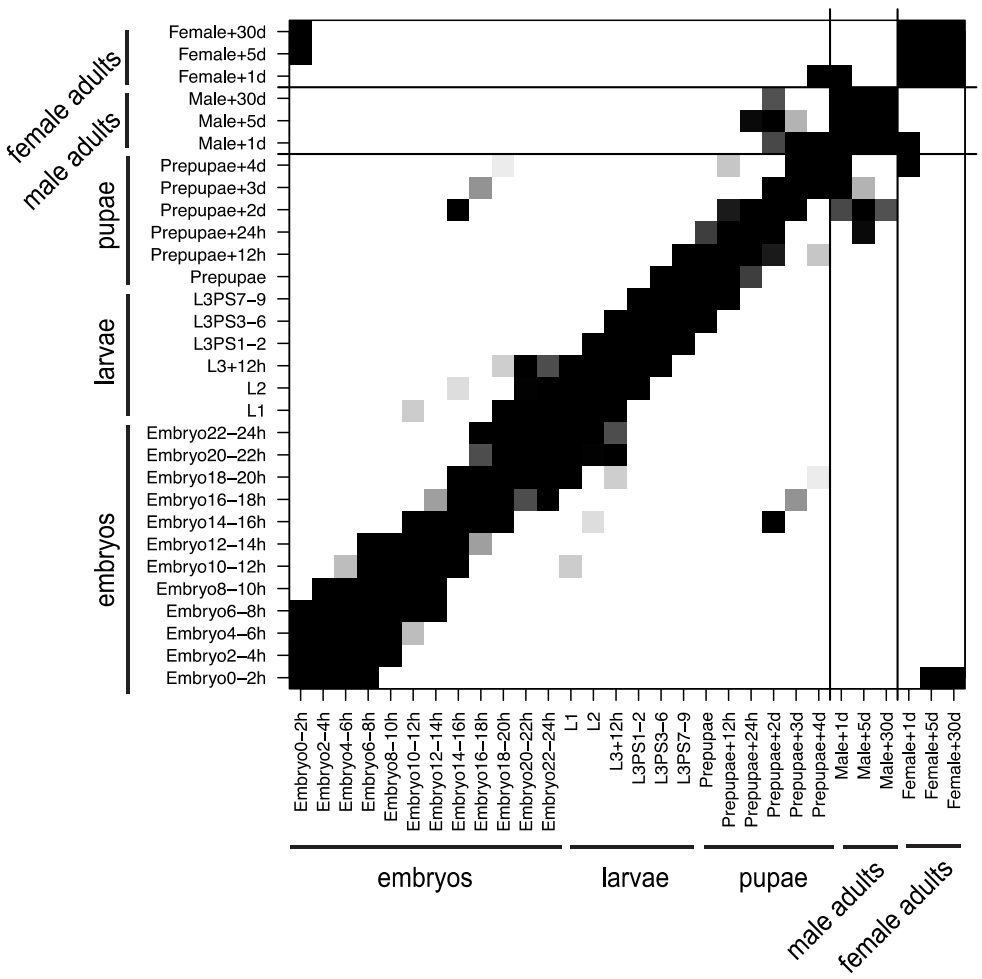


vs.

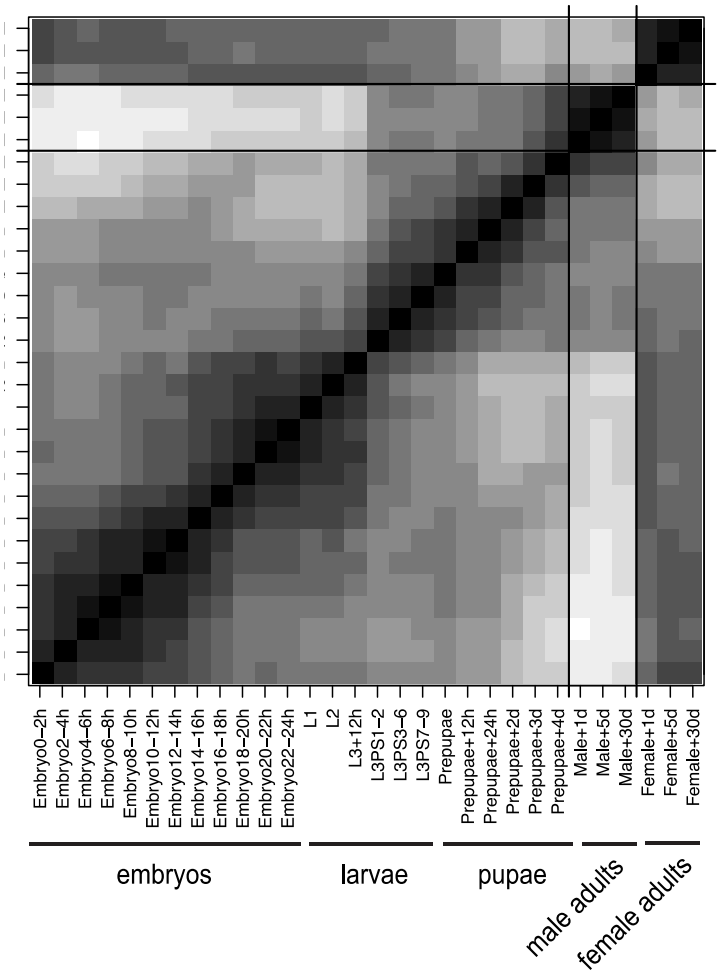


TROM vs. Spearman's rank correlation

D. melanogaster

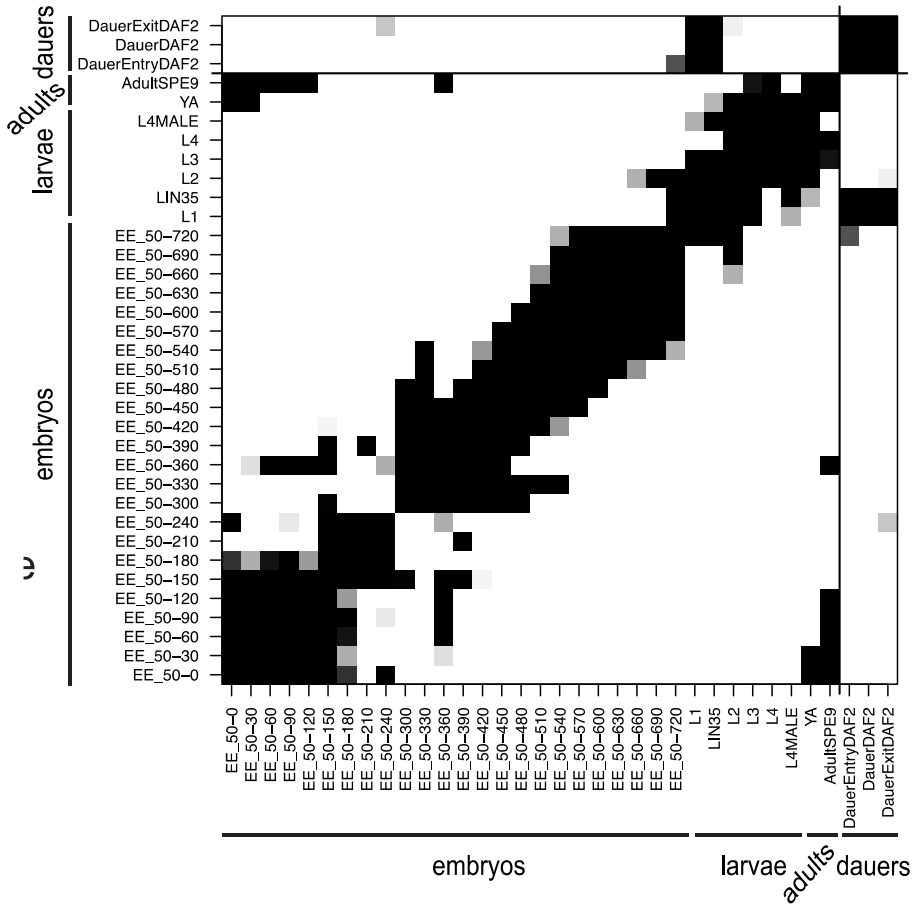


vs.

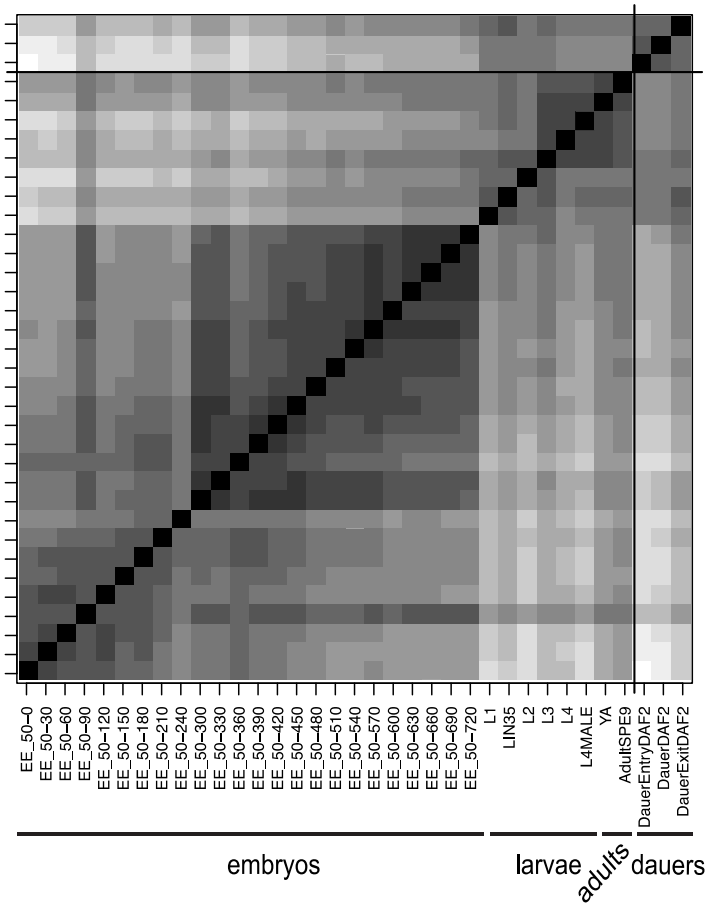


TROM vs. Spearman's rank correlation

C. elegans



VS.



TROM: a new measure for comparing

Statistics in Biosciences
June 2017, Volume 9, Issue 1, pp 105–136 | Cite as

TROM: A Testing-Based Method for Finding Transcriptomic Similarity of Biological Samples

Authors | Authors and affiliations

Wei Vivian Li, Yiling Chen, Jingyi Jessica Li

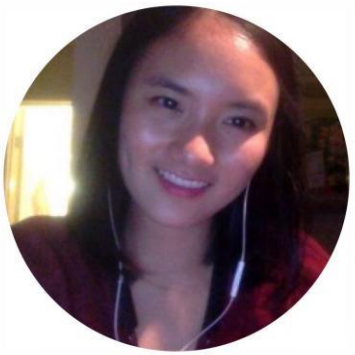
Article
First Online: 29 August 2016

- R package available on CRAN
- Questions can be sent to jli@stat.ucla.edu

Acknowledgements

- Wei Vivian Li
  - Ph.D. student at UCLA
- Yiling Chen
  - Undergraduate student, now Ph.D. student at UCLA

Wei Vivian Li

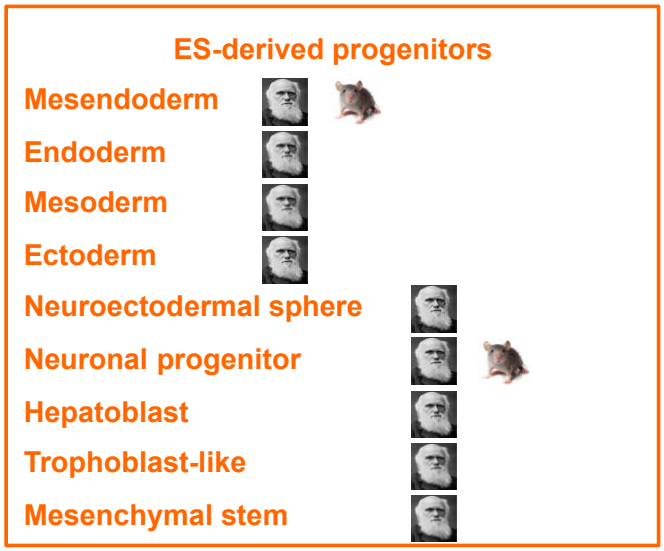


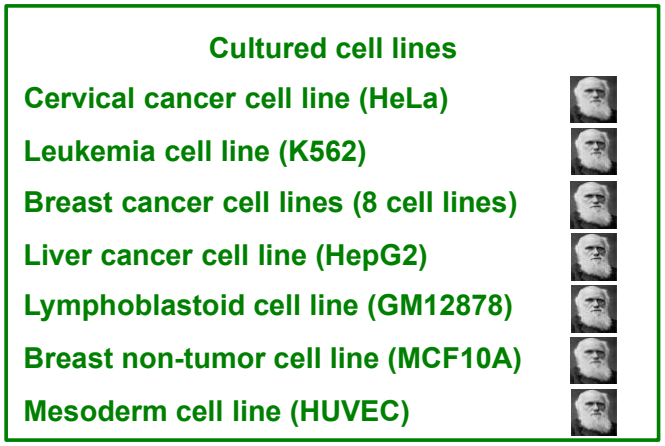
Yiling Chen

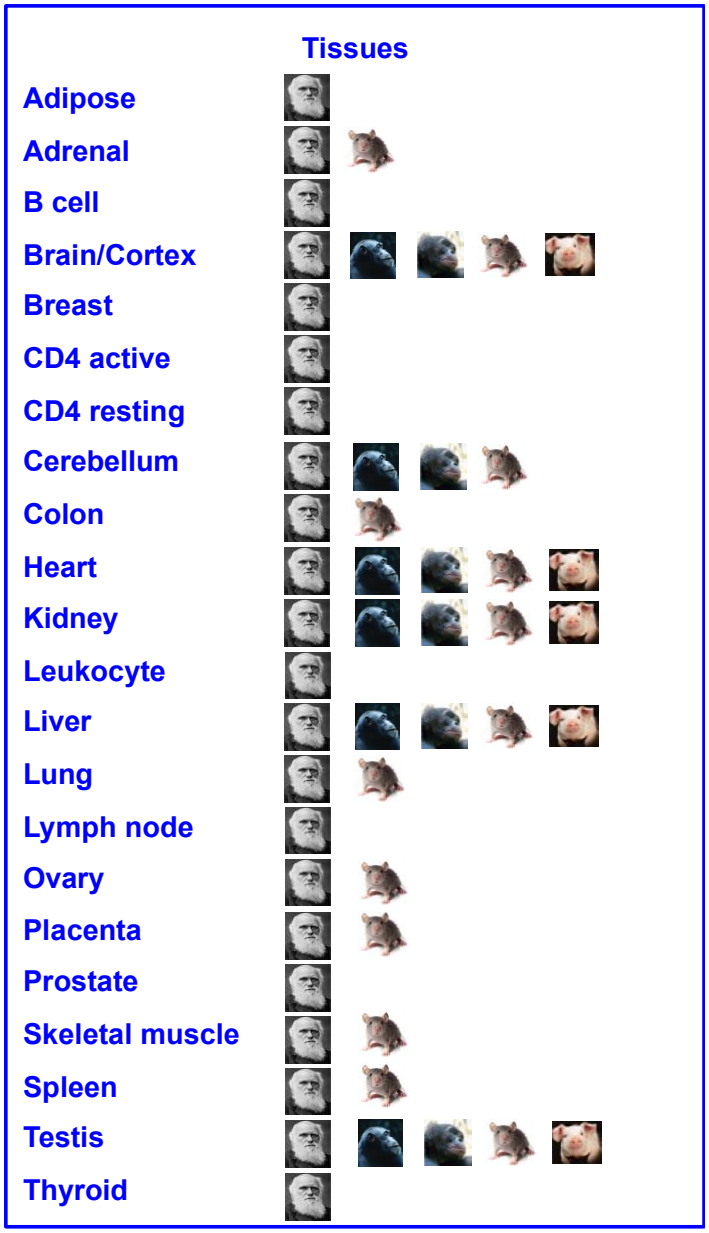
Motivating example 2

RNA-seq data

courtesy to Yu-Cheng T. Yang (Tsinghua University)

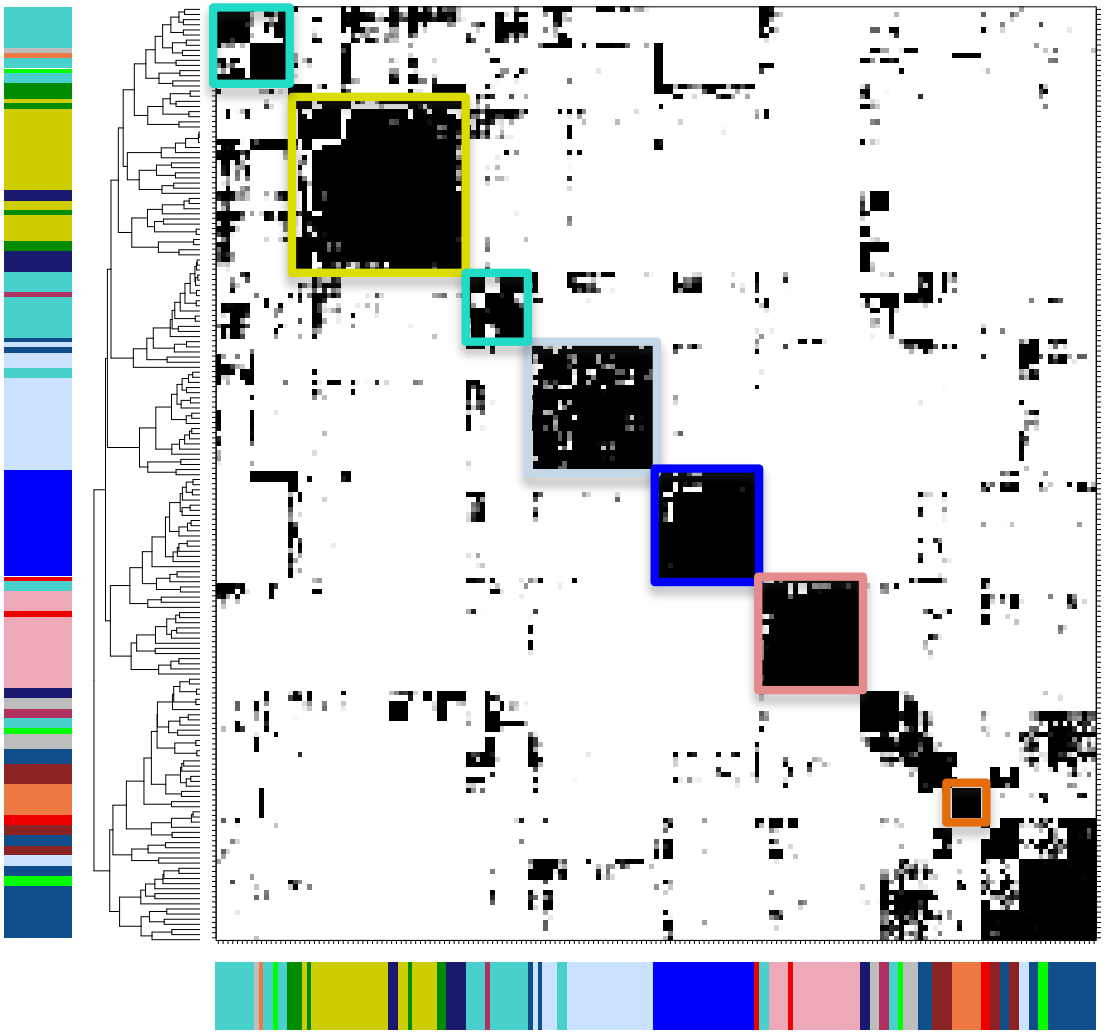






Transcriptome mapping within human

Within-human mapping using protein-coding genes



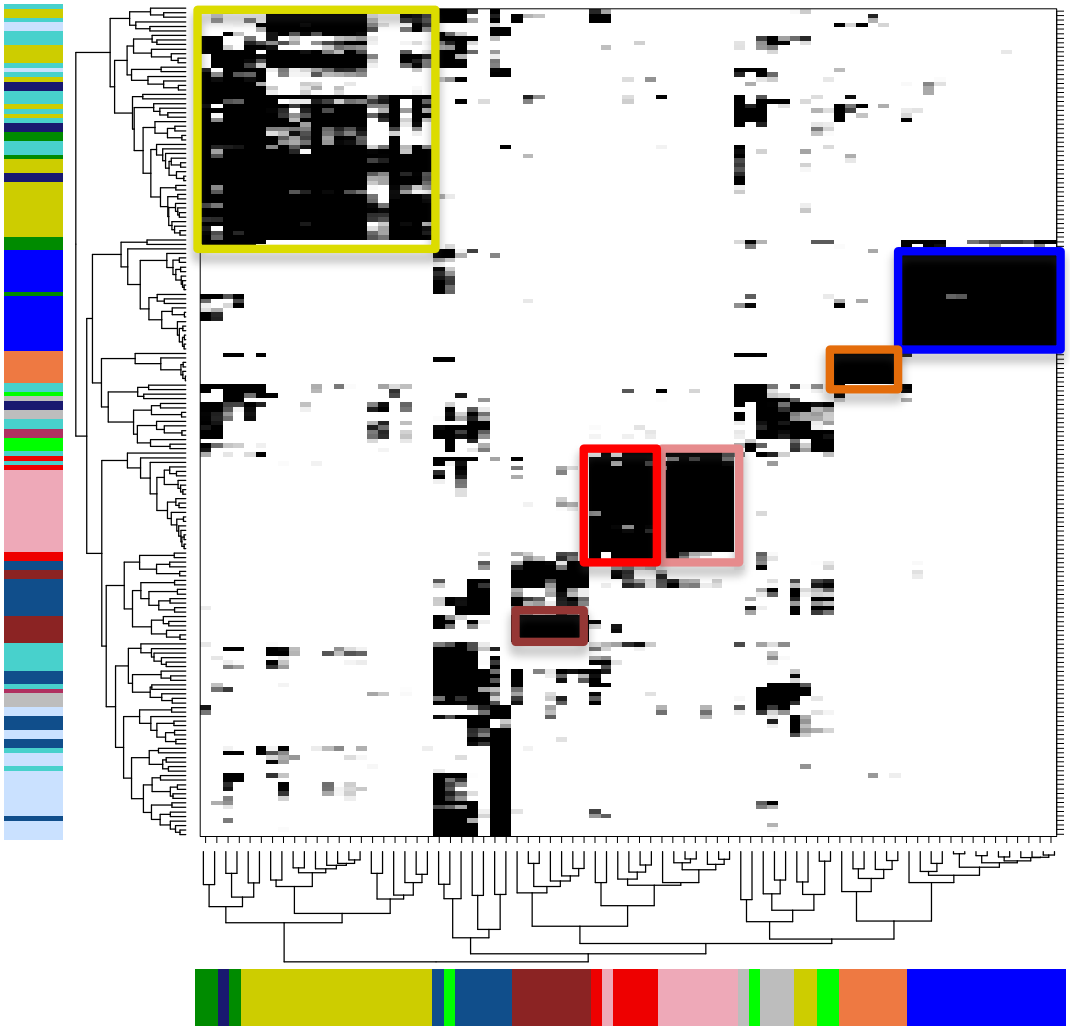
Mapping score
(-log$_{10}$ Bonferroni p)

0
1
2
3
4
5
>6

ESC and iPSC
Germ layers
Neural progenitors
Other progenitors
Neural tissues
Immune cells & tissues
Heart & skeletal muscle
kidney
Liver
Placenta & trophoblast
Ovary
testis
Other tissues
Cultured cell lines

Transcriptome mapping between species

Human-mouse mapping using protein-coding genes



Mapping score
(-log$_{10}$ Bonferroni p)

0
1
2
3
4
5
>6

ESC and iPSC
Germ layers
Neural progenitors
Other progenitors
Neural tissues
Immune cells & tissues
Heart & skeletal muscle
kidney
Liver
Placenta & trophoblast
Ovary
testis
Other tissues
Cultured cell lines

Paper

Nucleic Acids Research

Volume 45, Issue 4
28 February 2017

Large-scale mapping of mammalian transcriptomes identifies conserved genes associated with different cell states

Yang Yang, Yu-Cheng T. Yang, Jiapei Yuan, Zhi John Lu ✉, Jingyi Jessica Li ✉

Nucleic Acids Research, Volume 45, Issue 4, 28 February 2017, Pages 1657–1672,
https://doi.org/10.1093/nar/gkw1256
Published: 14 December 2016 **Article history** ▼

Acknowledgements

- Lu Lab (Tsinghua University, China)
  - Yu-Cheng T. Yang, Yang Yang, and Dr. Zhi (John) Lu

Yu-Cheng T. Yang

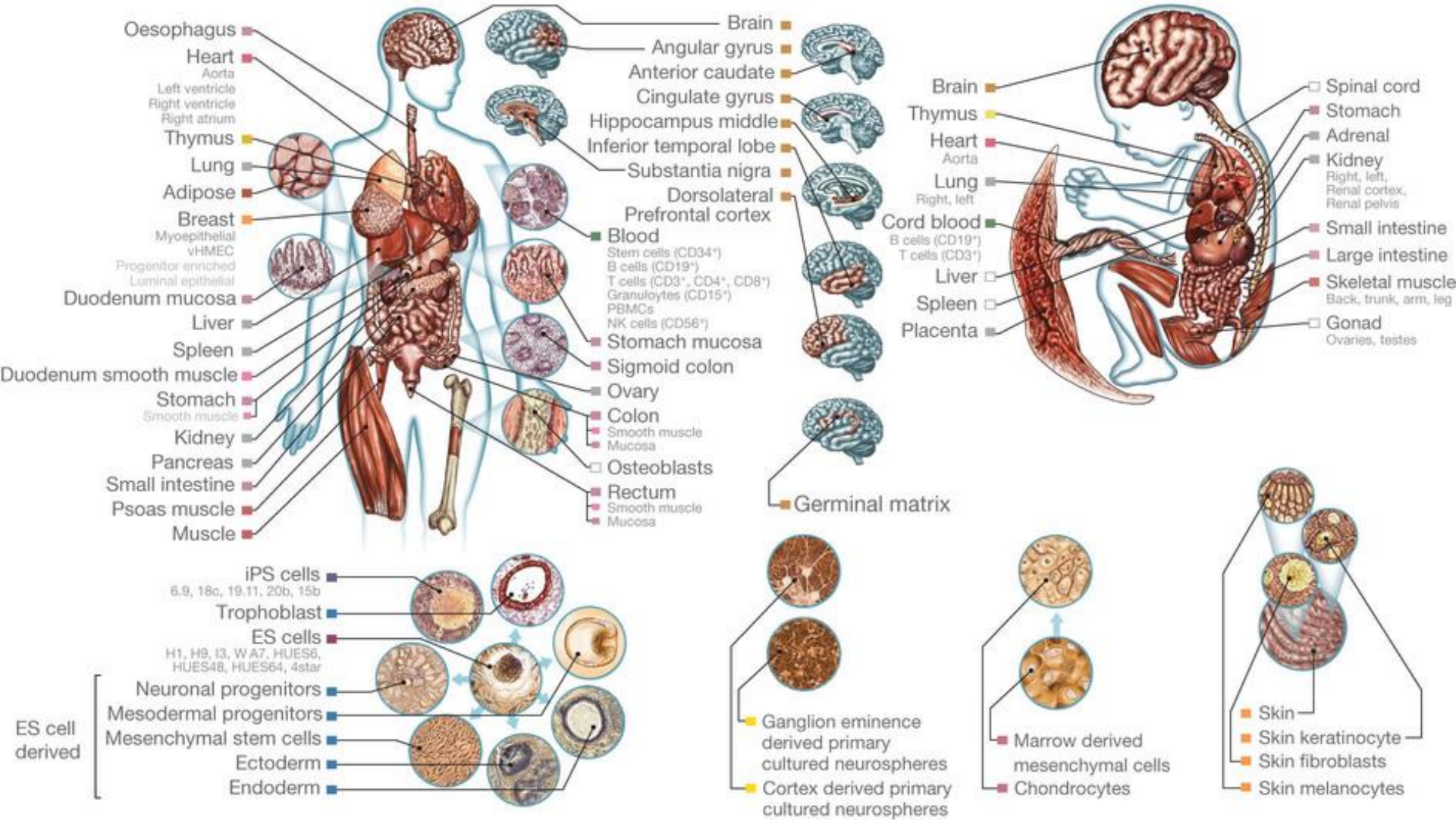
Yang Yang

Zhi (John) Lu

Part II
Epigenome Overlap Measure (EPOM)

ChIP-seq data

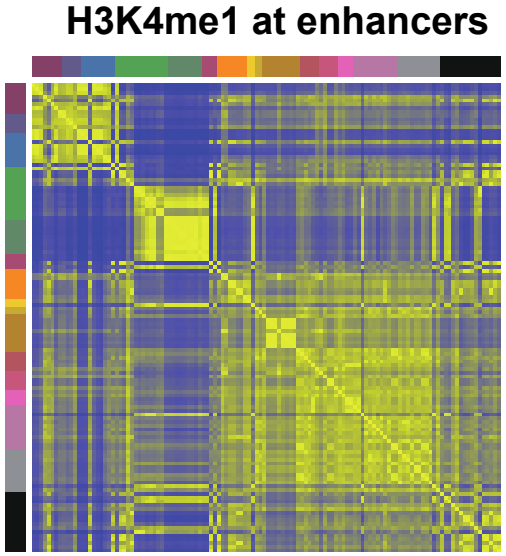
Motivating example 3

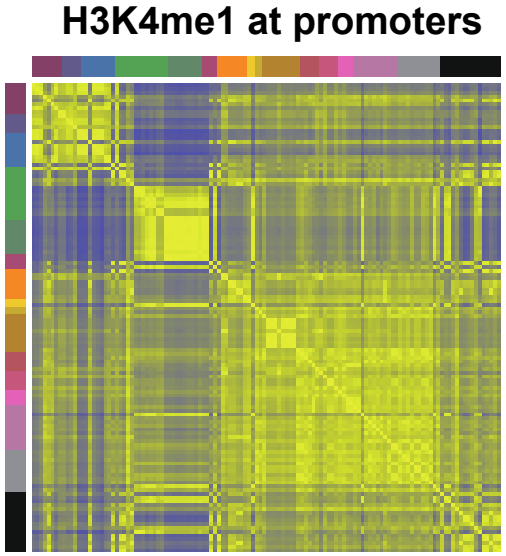


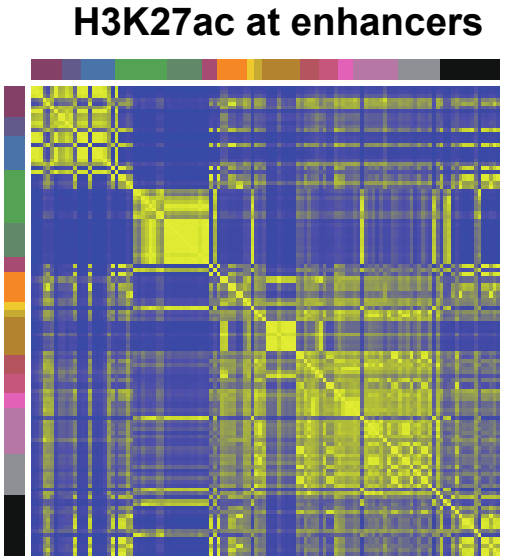
Kundaje et al. “Integrative analysis of 111 reference human epigenomes”. *Nature* (2015).

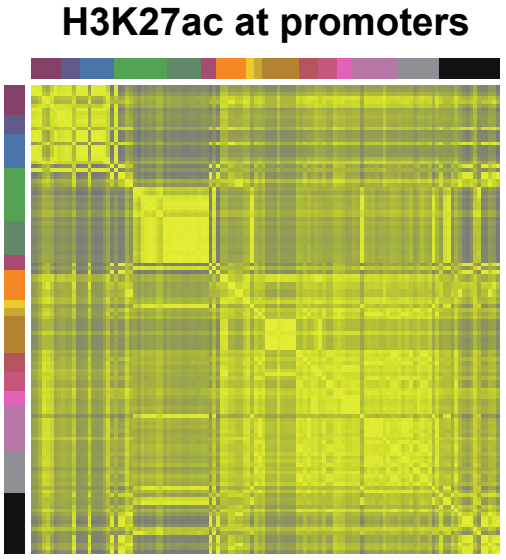
Why not correlation analysis?

- ESC (embryonic stem cell)
- iPSC (induced pluripotent stem cells)
- ES-deriv (ESC-derived cells)
- Blood & T cell
- HSC (hematopoietic stem cell) & B-cell
- Mesench (mesenchymal stem cells)
- Epithelial
- Neurosph (neurosphere)
- Thymus
- Brain
- Muscle
- Heart
- Sm. Muscle (smooth muscle)
- Digestive
- Other
- ENCODE2012

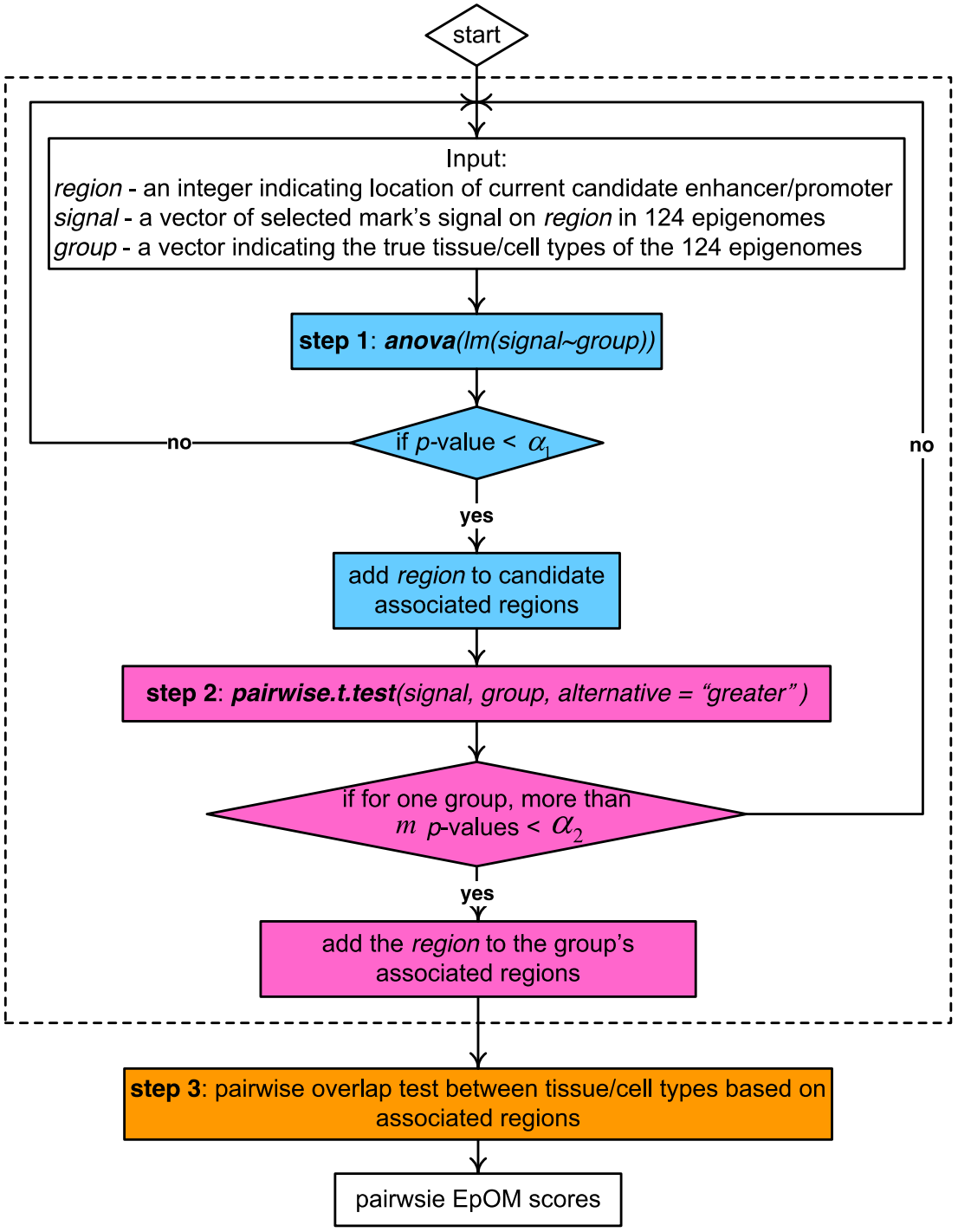




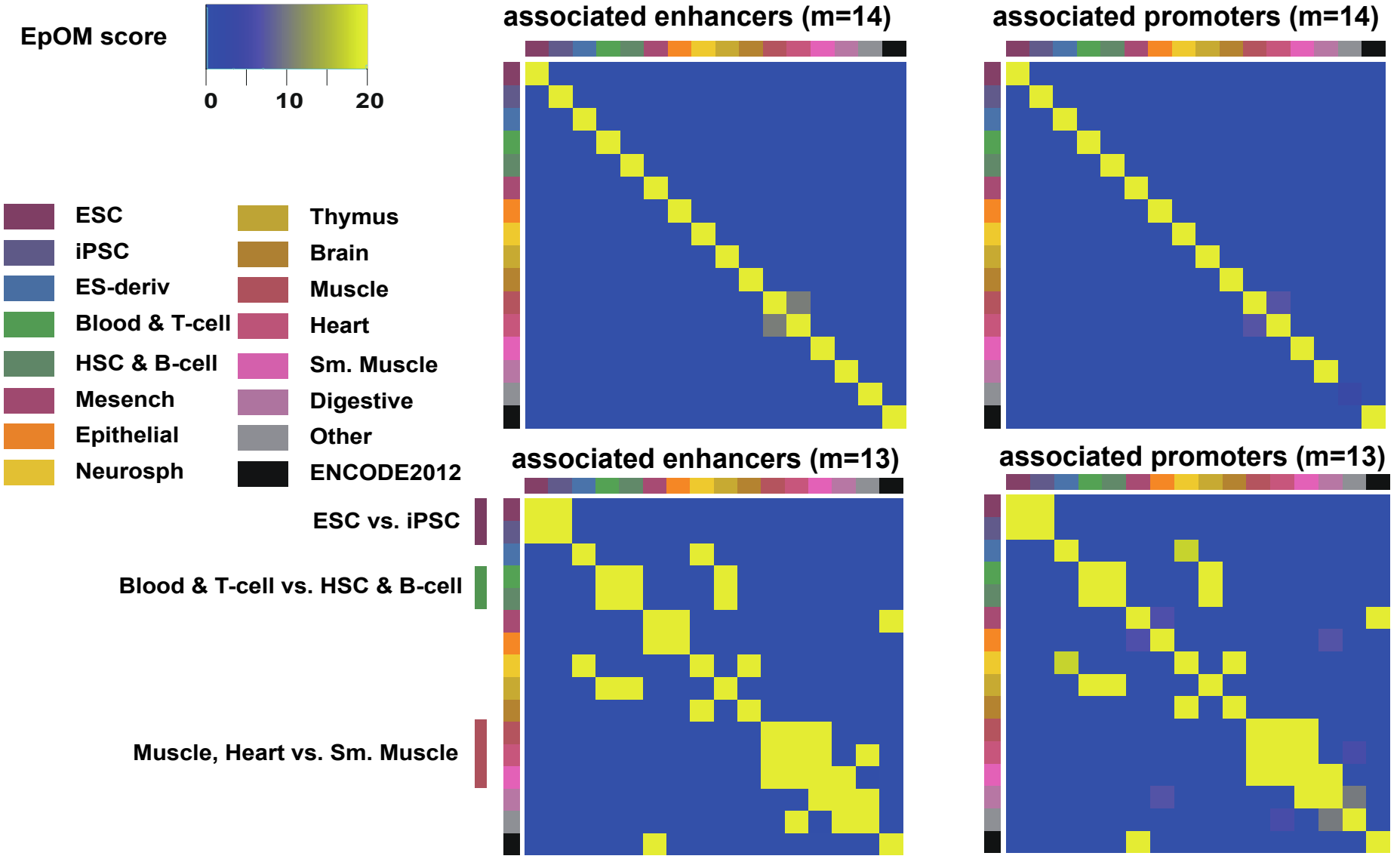




EPOM outline



EPOM results



GO enrichment of associated enhancers and promoters in ESC vs. iPSC

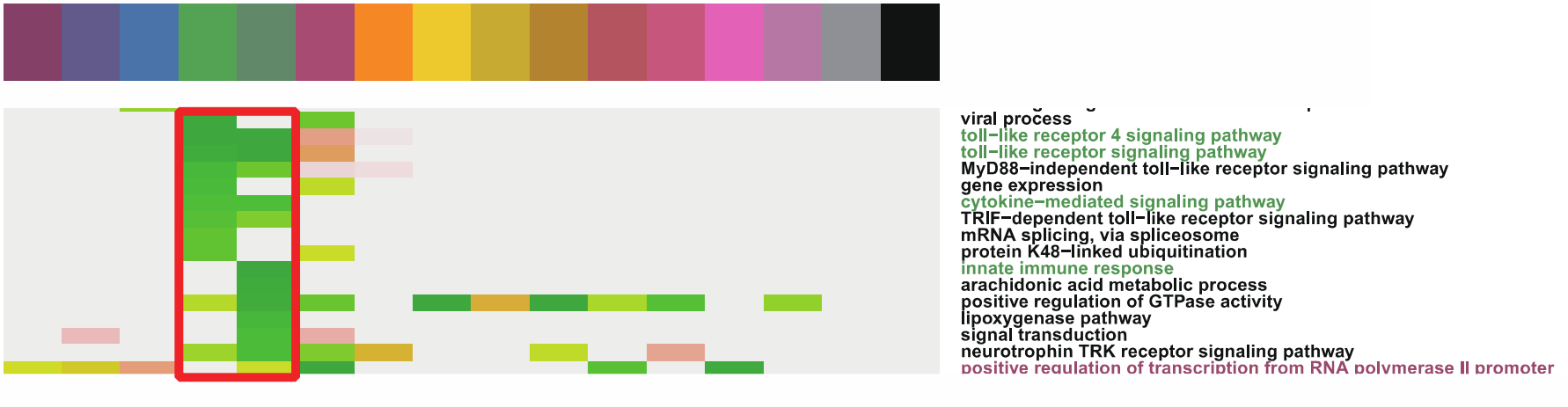
associated enhancers' potential target genes

fibroblast growth factor receptor signaling pathway
synaptic transmission
negative regulation of transcription from RNA polymerase II promoter
neuron migration
chondrocyte development
positive regulation of transcription, DNA−templated
stem cell maintenance
BMP signaling pathway
positive regulation of synapse assembly
axon guidance
canonical Wnt signaling pathway
negative regulation of canonical Wnt signaling pathway
adenylate cyclase−activating dopamine receptor signaling pathway
ionotropic glutamate receptor signaling pathway

associated promoters' potential target genes

fibroblast growth factor receptor signaling pathway
negative regulation of canonical Wnt signaling pathway
negative regulation of Wnt signaling pathway
synaptic transmission
I−kappaB phosphorylation
phosphatidylinositol−3−phosphate biosynthetic process
negative regulation of transcription from RNA polymerase II promoter
neuron migration
protein O−linked glycosylation via threonine
snoRNA catabolic process
ammonium transport
endocrine pancreas development
positive regulation of transcription from RNA polymerase II promoter
vascular endothelial growth factor signaling pathway
positive regulation of transcription, DNA−templated
regulation of alpha−amino−3−hydroxy−5−methyl−4−isoxazole propionate
negative regulation of transcription, DNA−templated
negative regulation of steroid biosynthetic process

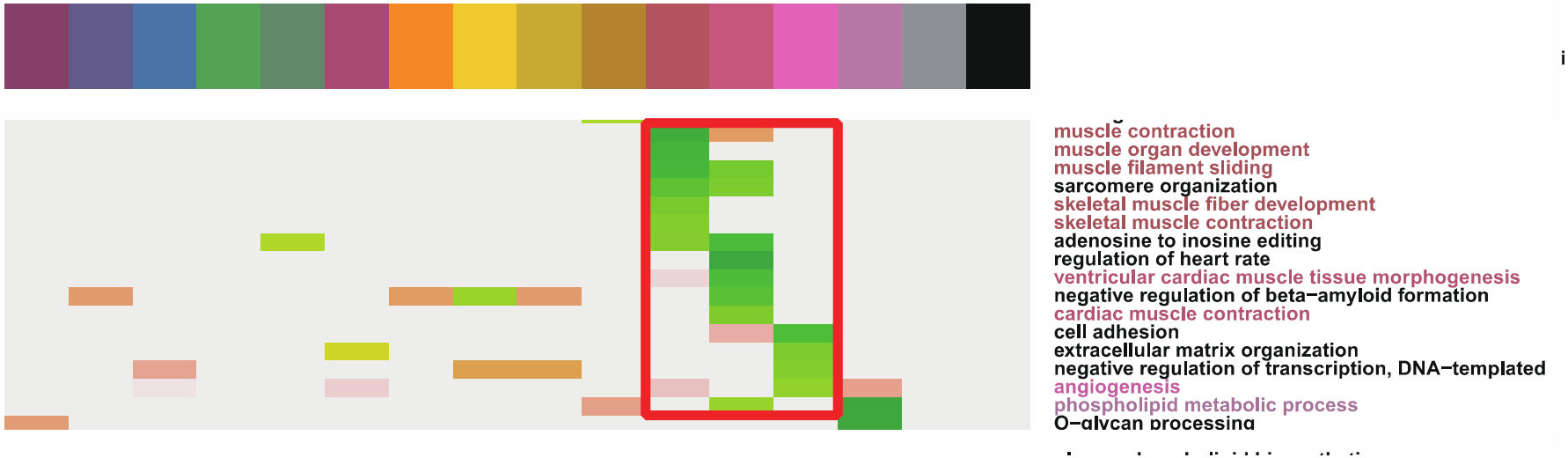
GO enrichment of associated enhancers and promoters in Blood & T-cell vs. HSC & B-cell



Common top enriched GO terms:

- toll-like receptor 4 signaling pathway
- toll-like receptor signaling pathway
- Notch signaling pathway
- MyD88-independent toll-like receptor signaling pathway
- cytokine-mediated signaling pathway
- neurotrophin TRK receptor signaling pathway

GO enrichment of associated enhancers and promoters in Muscle, Heart and Sm. Muscle



Common top enriched GO terms:

- muscle filament sliding
- sarcomere organization
- fibroblast growth factor receptor signaling pathway
- adenosine to inosine editing
- positive regulation of GTPase activity
- HAC1-type intron splice site recognition and cleavage

Disease ontology enrichment analysis of associated enhancers (promoters)

tissue/cell type	top enriched DO terms
Blood & T-cell, HSC & B-cell	hypersensitivity reaction disease (celiac disease), hematopoietic system disease (lymphopenia) and immune system cancer (lymphoma and leukemia)
Epithelial	gastric adenocarcinoma
Brain	disease of mental health (attention deficit hyperactivity disorder, alchohol dependence and schizophrenia), major depressive disorder, neurodegenerative disease (Alzheimer's and Parkinson's)
Muscle	cardiovascular system disease, cardiomyopathy
Heart	cardiovascular system disease, diabetes mellitus, kidney disease
Sm. Muscle	coronary artery disease
Digestive	hepatocellular carcinoma, pancreatic cancer, gastrointestinal system disease (ulcerative colitis and esophageal cancer)

Paper

BMC Genomics

Home About Articles Submission Guidelines

PROCEEDINGS | Open Access

Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states

Wei Vivian Li, Zahra S. Razaee and Jingyi Jessica Li

BMC Genomics 2016 **17(Suppl 1)**:S10
https://doi.org/10.1186/s12864-015-2303-9
Published: 11 January 2016

Software

https://github.com/ruochenj/EPOM_R

Acknowledgements

- Wei Vivian Li
  - Ph.D. student at UCLA
- Zahra Razaee
  - Former Ph.D. student
- Ruochen Jiang Razaee
  - M.S. student at UCLA

Wei Vivian Li

Zahra Razaee

Ruochen Jiang