Transcriptome Analysis:

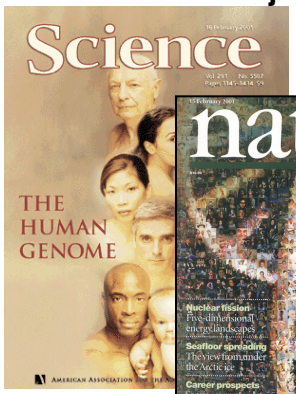
Expression Clustering across Distant Organisms

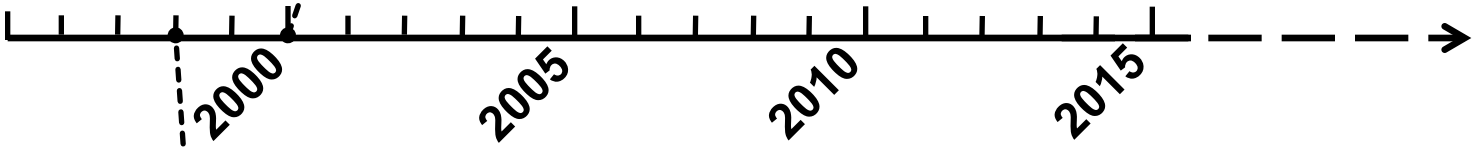
M Gerstein, Yale

See last slide for references & more info. (Background image from http://www.genomenewsnetwork.org/articles/04_02/leukemia.shtml)

Slides freely downloadable from **Lectures.GersteinLab.org** & “tweetable” (via @markgerstein)

The Human Genome Project



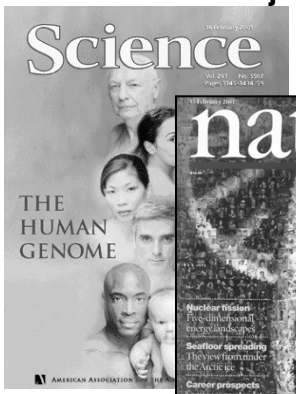


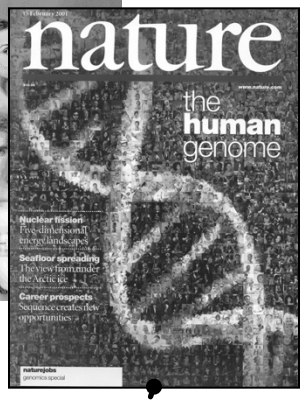
Worm Genome

The Human Genome Project

ENCODE Pilot

ENCODE Production





2000

2005

2010

Worm Genome

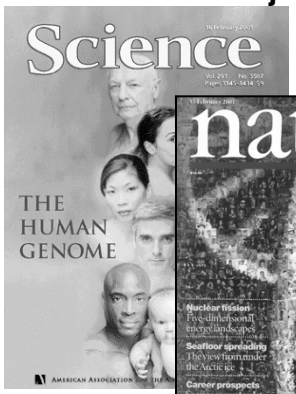
modENCODE

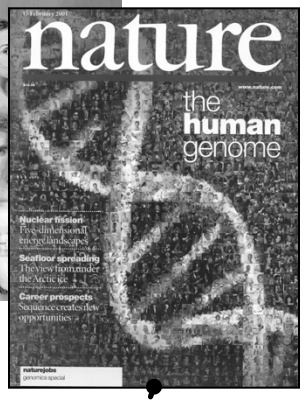
The Human Genome Project

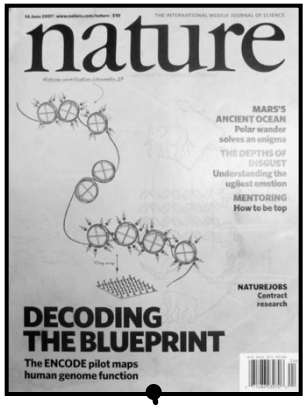
ENCODE Pilot

ENCODE Production

Comparative ENCODE





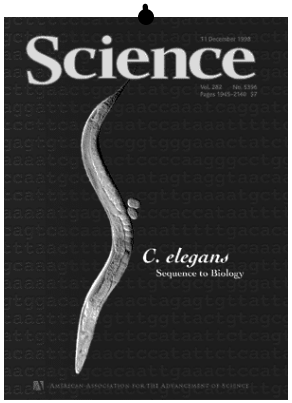


2000

2005

2010

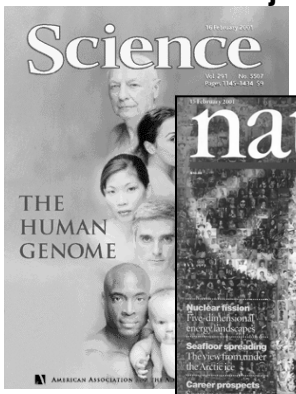
2015

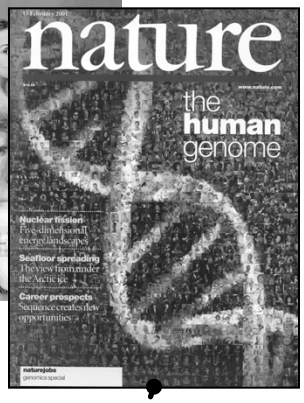


Worm Genome

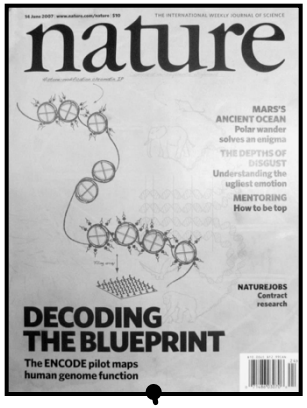
modENCODE

The Human Genome Project





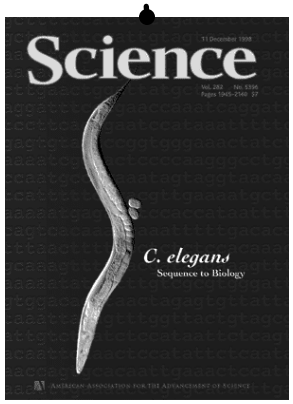
ENCODE Pilot



ENCODE Production

Comparative ENCODE

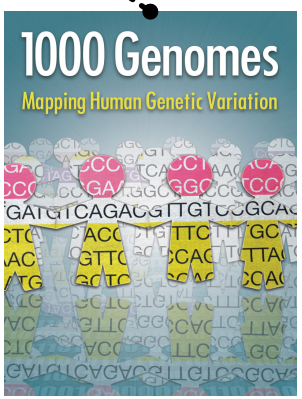
2000 2005 2010 2015



Worm Genome

modENCODE

1000 Genomes Pilot



1000 Genomes Production

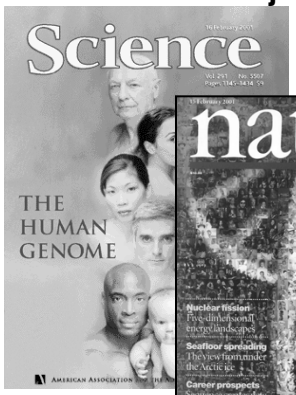
The Human Genome Project

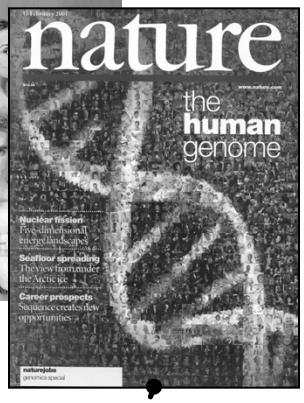
ENCODE Pilot

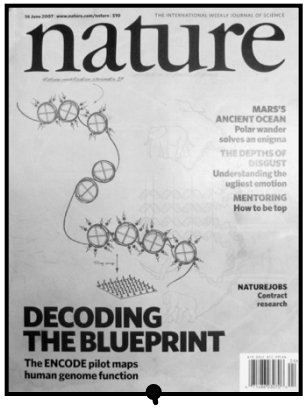
ENCODE Production

Comparative ENCODE

Epigenome Roadmap





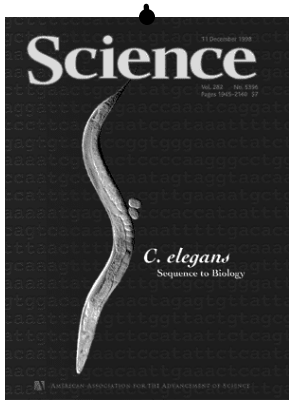


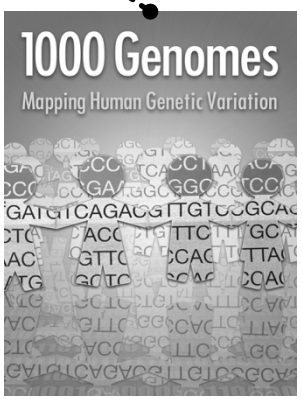
2000

2005

2010

2015





Worm Genome

modENCODE

1000 Genomes Pilot

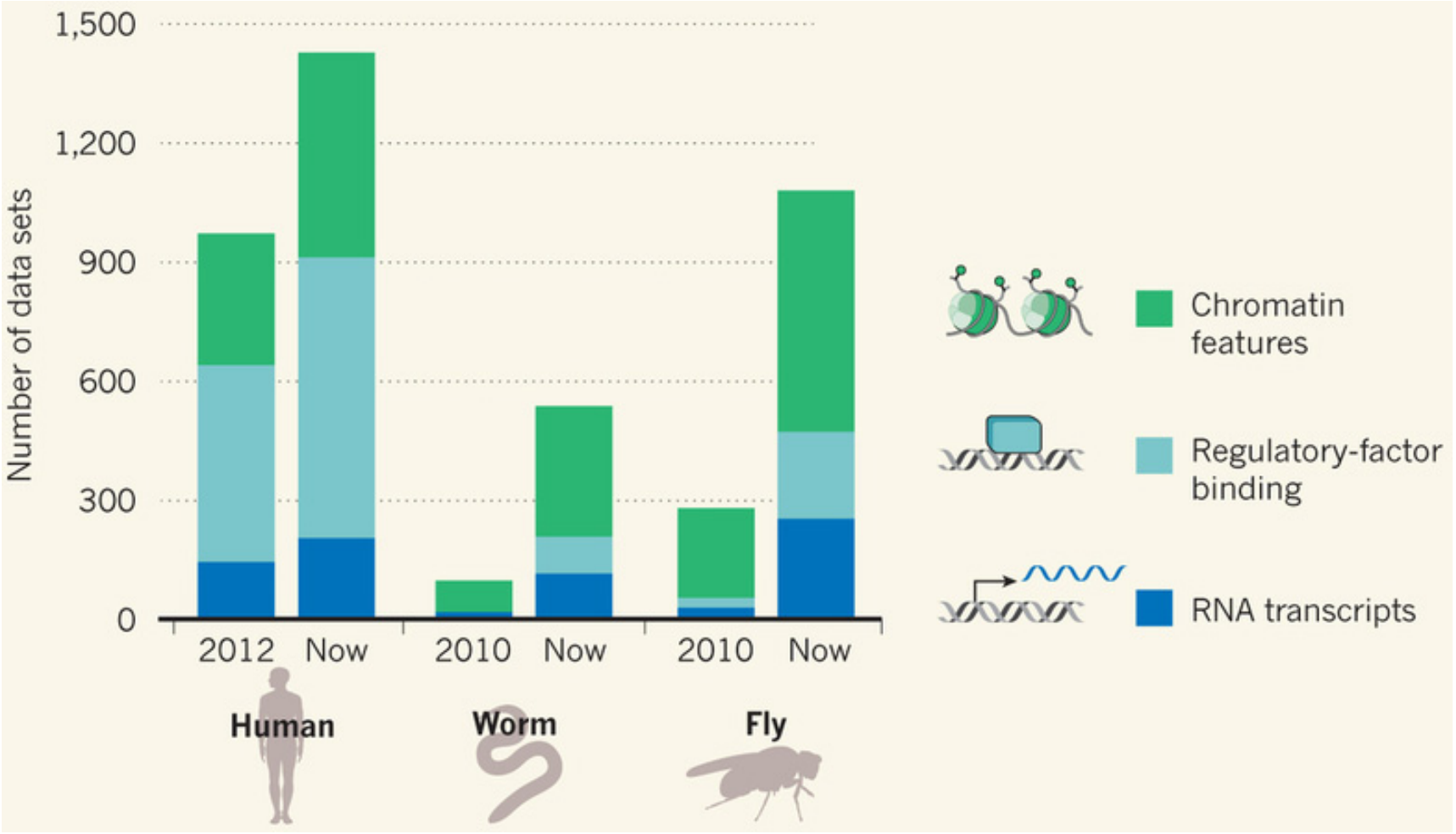
1000 Genomes Production

GTEx

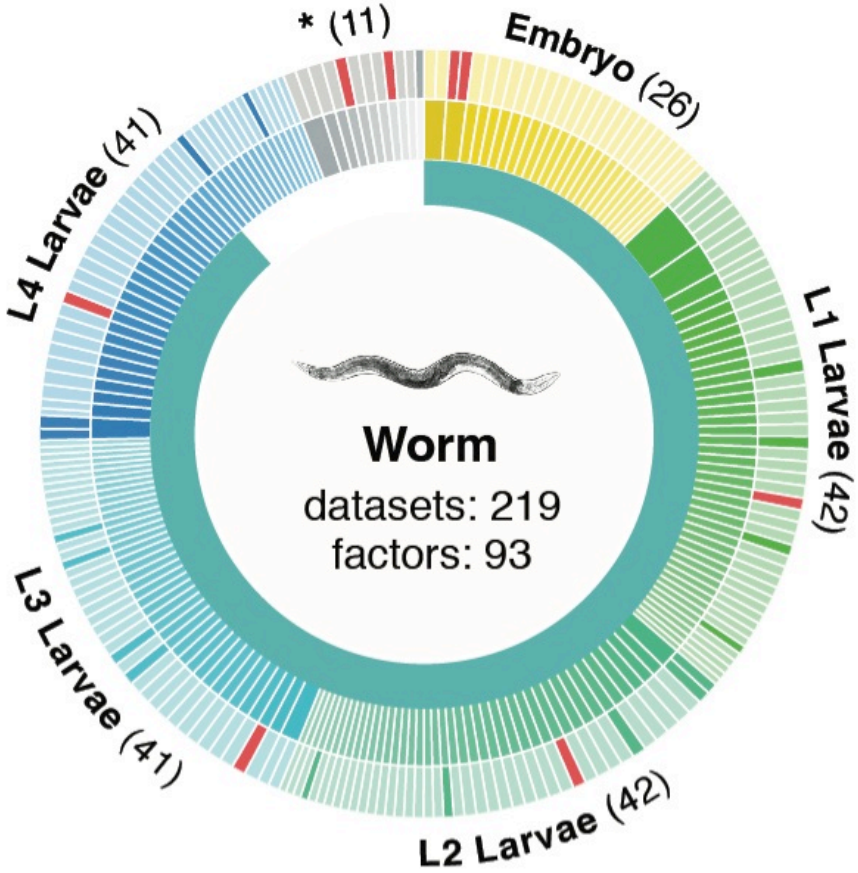
Comparative ENCODE Functional Genomics Resource

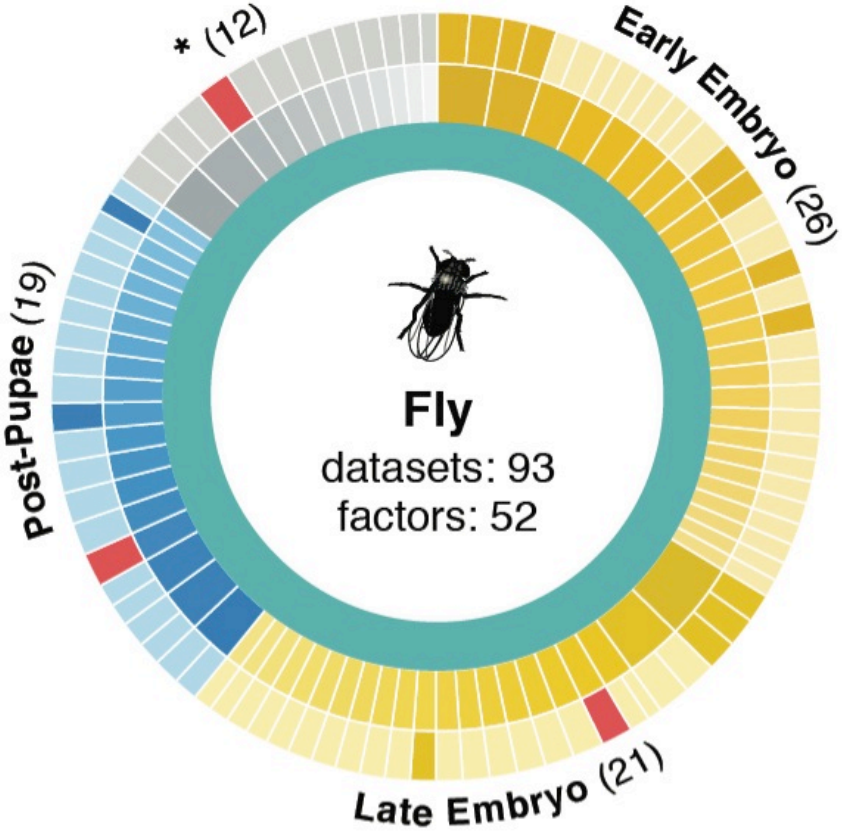
(EncodeProject.org/comparative)

- Broad sampling of conditions across transcriptomes & regulomes for human, worm & fly
  - embryo & ES cells
  - developmental time course (worm-fly)
- In total: ~3000 datasets (~130B reads)



Time-course gene expression data of worm & fly development





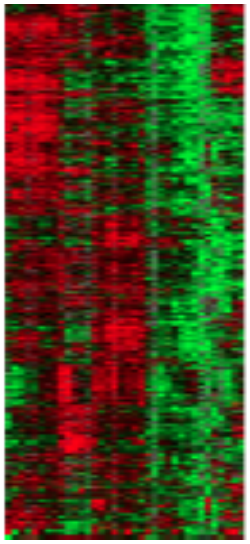
Organism	Major developmental stages
worm (*C. elegans*)	33 stages: 0, 0.5, 1, …, 12 hours, L1, L2, L3, L4, …, Young Adults, Adults
fly (*D. mel.*)	30 stages: 0, 2, 4, 6, 8,…, 20, 22 hours, L1-L4, Pupaes, Adults

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

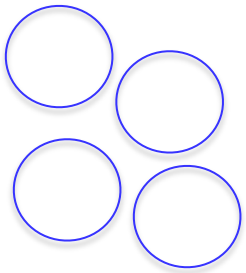
Expression clustering: revisiting an ancient problem



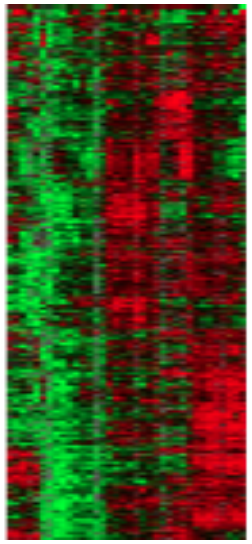
Species A

Clustering algorithm

co-expressed genes responsible for the same function in a species

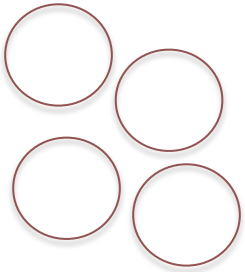


Species B



Clustering algorithm

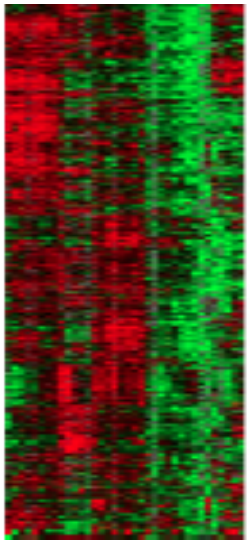
co-expressed genes responsible for the same function in a species



two independent sets of modules

Eisen MB et al. PNAS 1998
Langfelder P et al. BMC Bioinfo. 2008
Tamayo P et al. PNAS 1999
Kluger Y et al. Genome Res. 2003

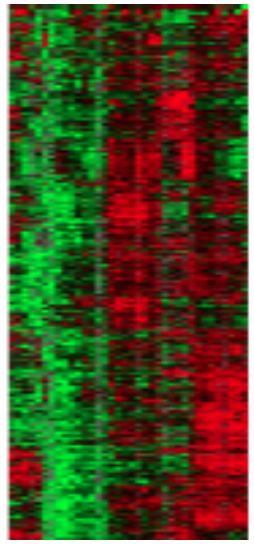
Expression clustering: revisiting an ancient problem



Species A

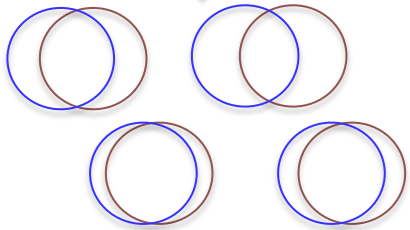
Orthologous pairs between species

Species B



A novel unified framework to integrate co-expression data across species

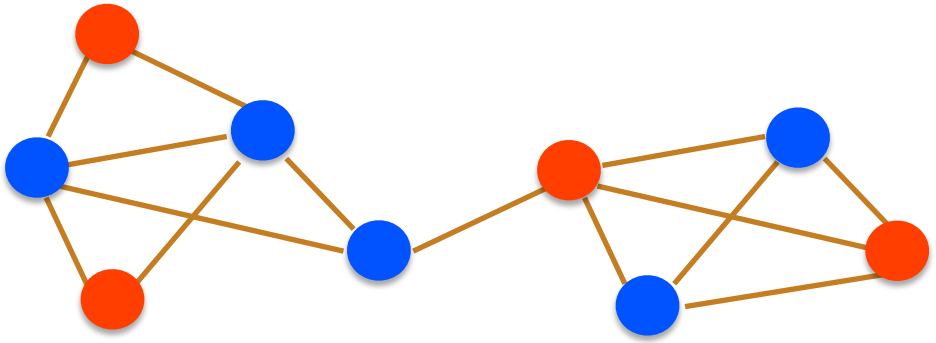
OrthoClust



cross species modules

Yan et al. Genome Biol. 2014

Network modularity



$$Q \approx 0$$

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

adjacency matrix

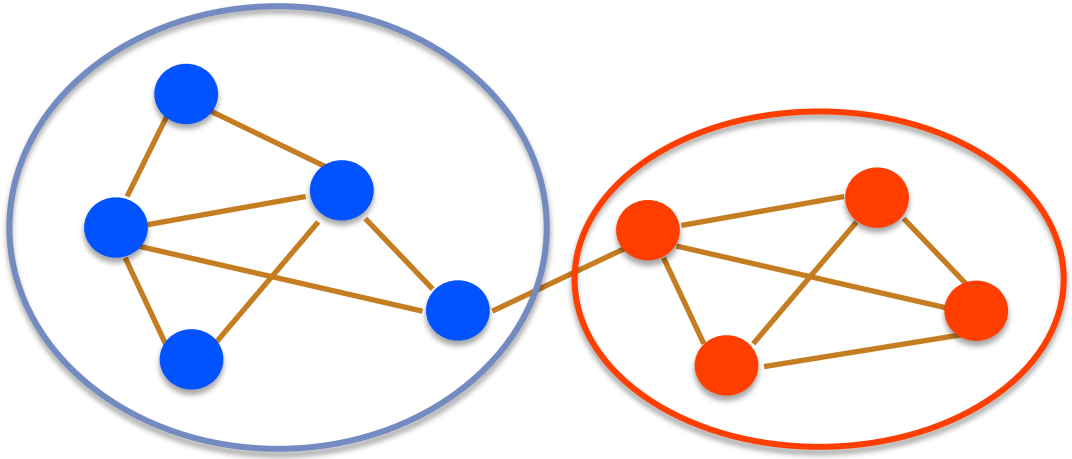
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q = Q_{max}$$

Optimization problem

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

adjacency matrix

degree of node i

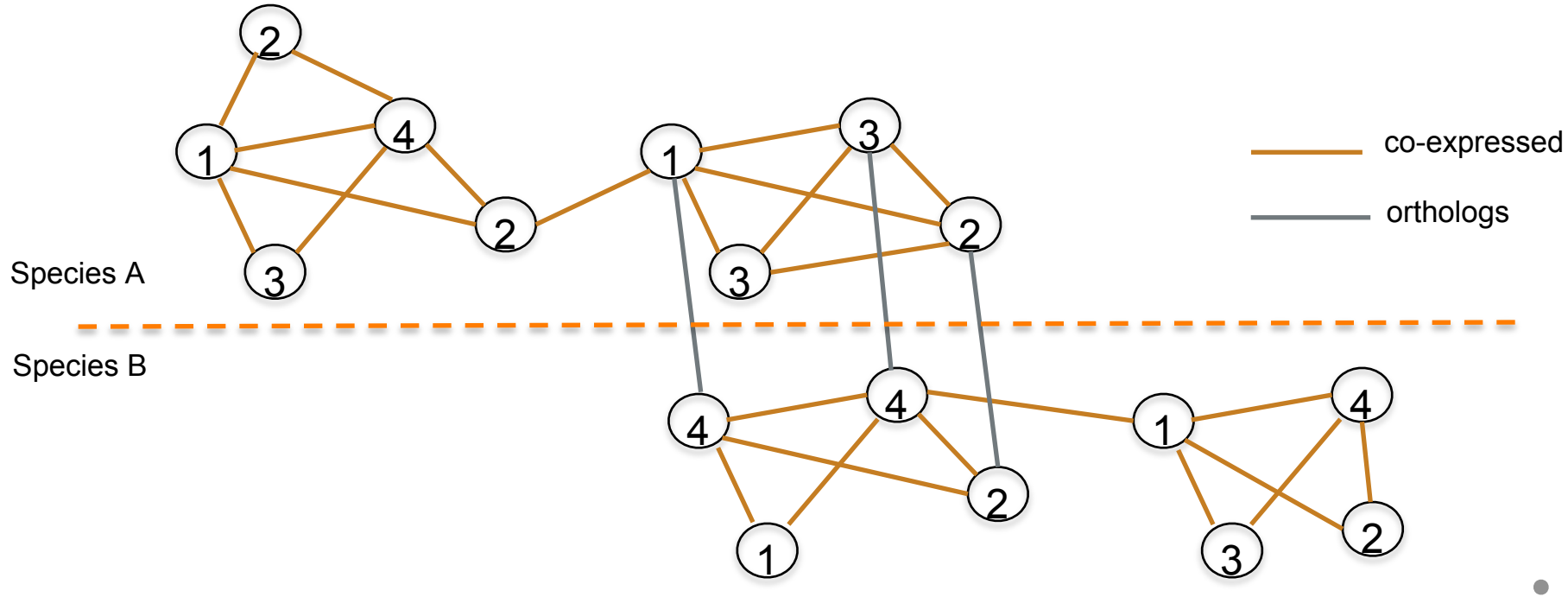
whether or not i, j are in the same module

number of edges

expected number of edges between i and j

OrthoClust: toy example

Every node i is assigned with a label σ_i (labels of modules: 1,2,…q).

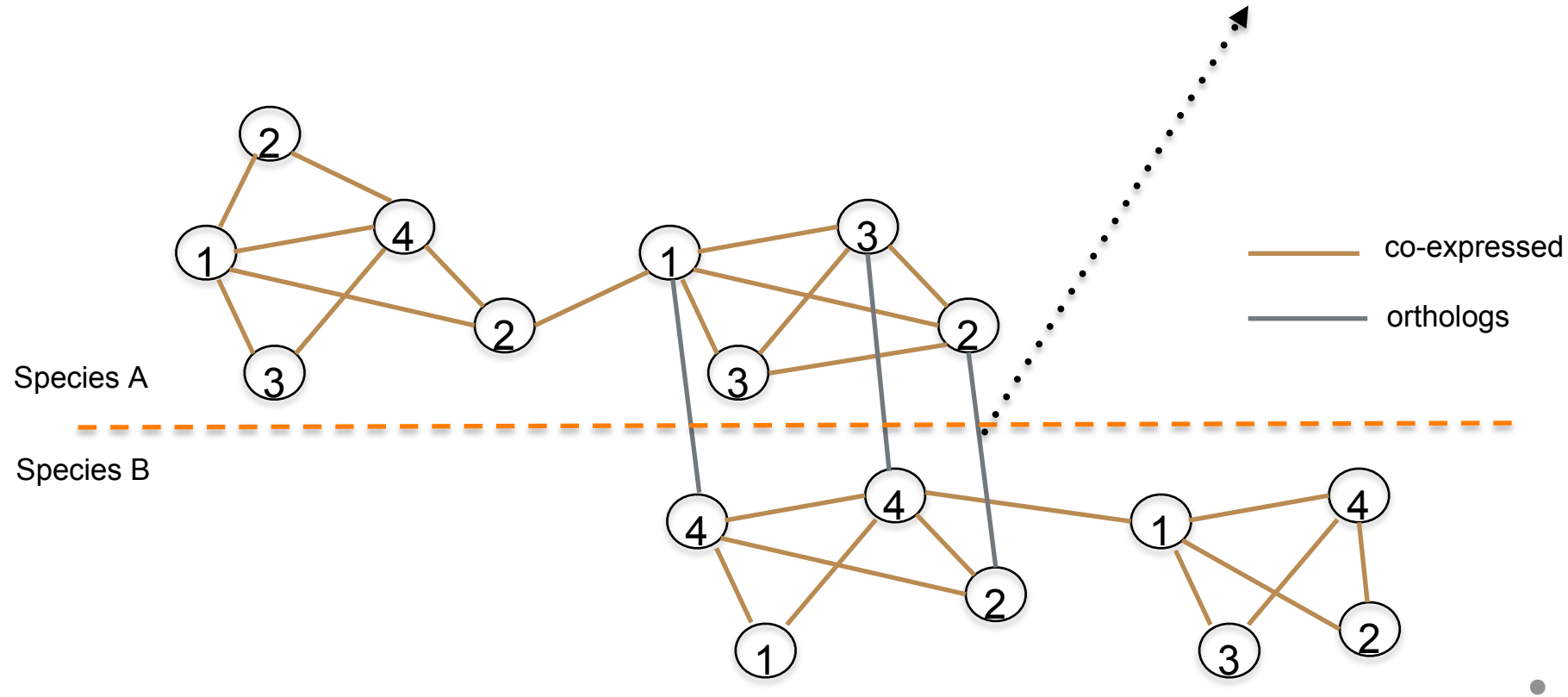


Yan KK et al. Genome Biology. 2014

OrthoClust: toy example

Every node i is assigned with a label σ_i (labels of modules: 1,2,…q).

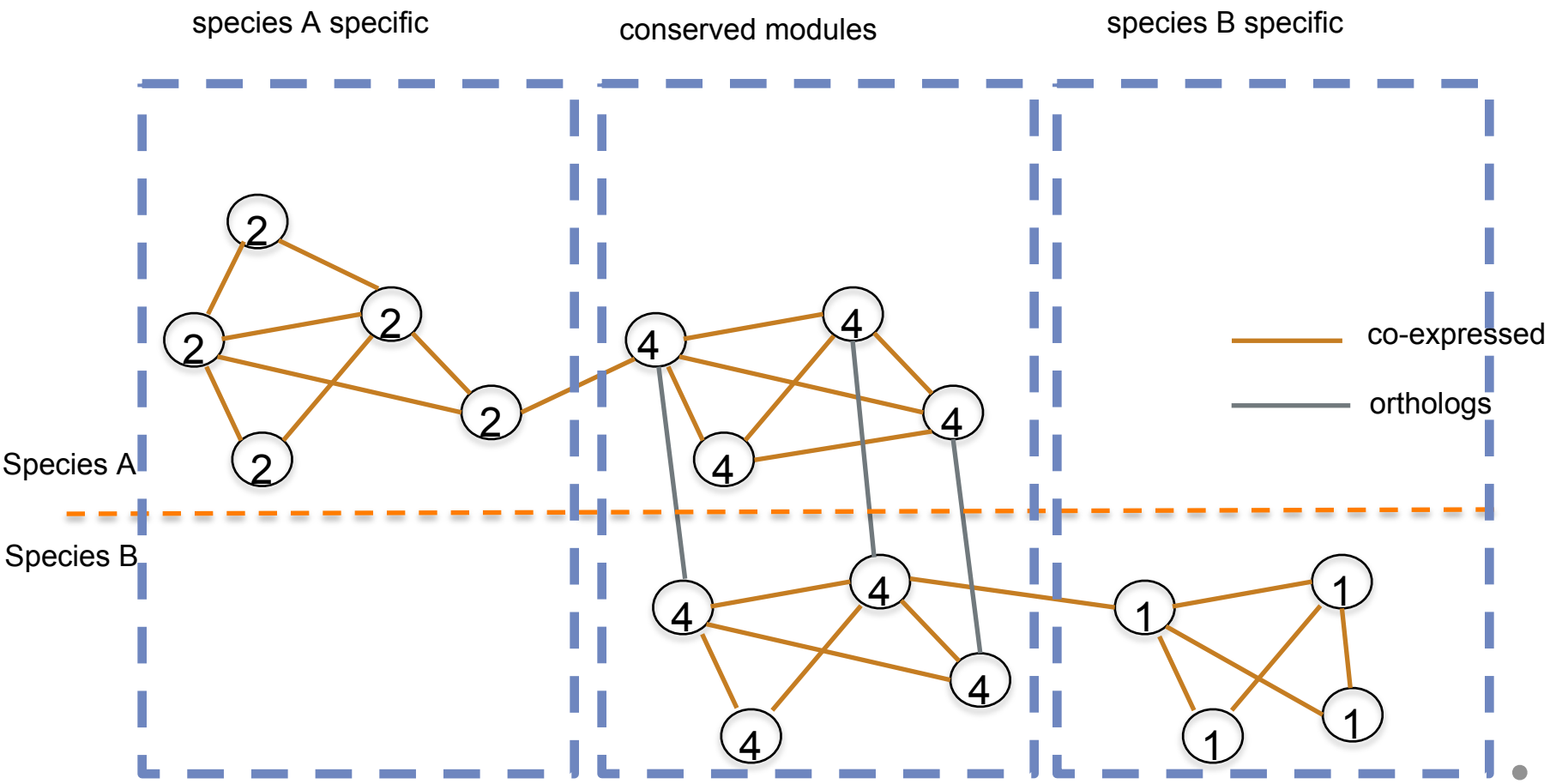
$$H = Q^{(A)} + Q^{(B)} - \kappa' \sum_{(i,j')\in(A,B)} \delta_{\sigma_i \sigma'_j}$$



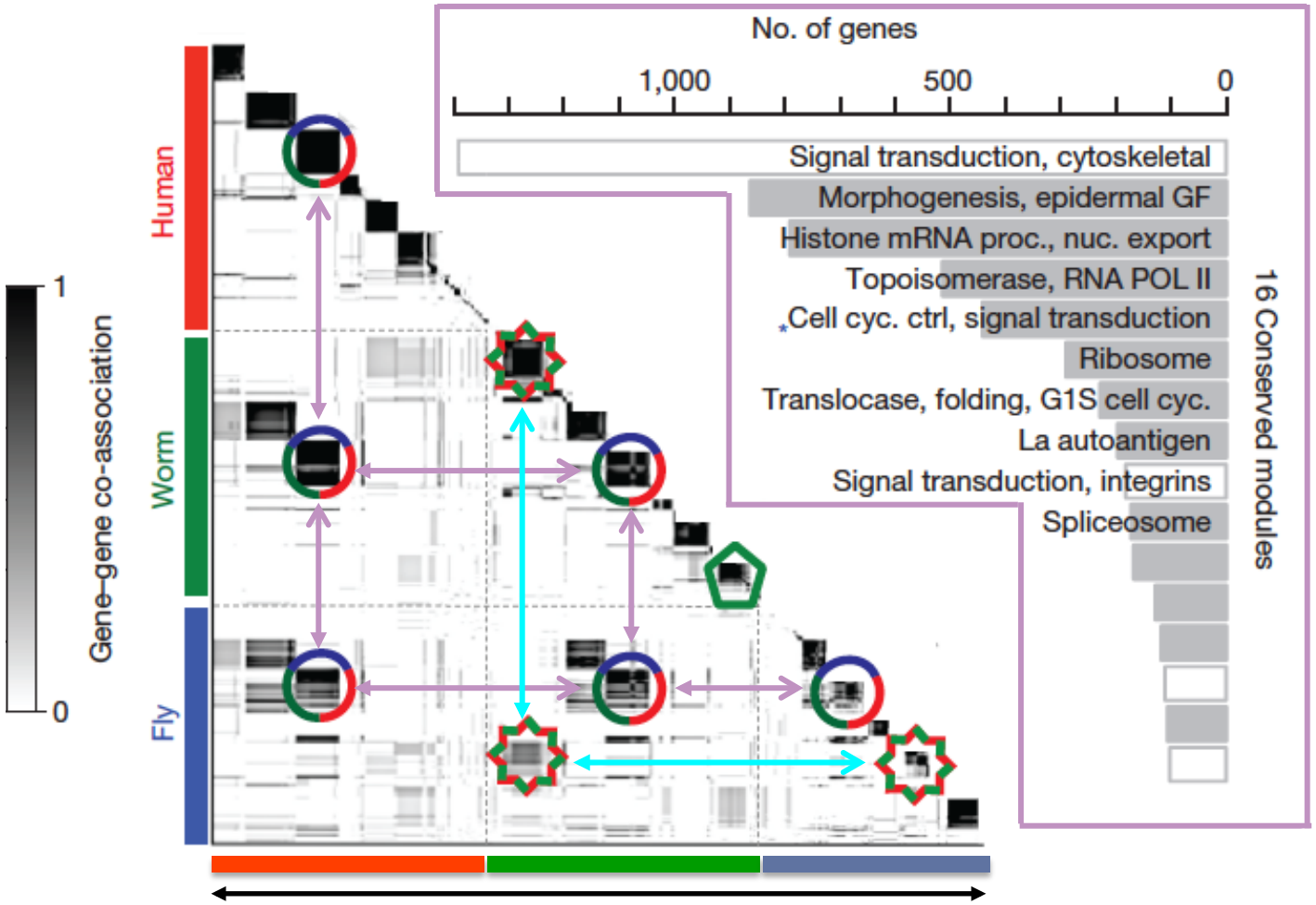
Yan KK et al. Genome Biology. 2014

OrthoClust: toy example

Use Potts model (generalized Ising model) to simultaneously cluster co-expressed genes within an organism as well as orthologs shared between organisms. Here, the ground state configuration correspond to three modules: 1, 2, 4.



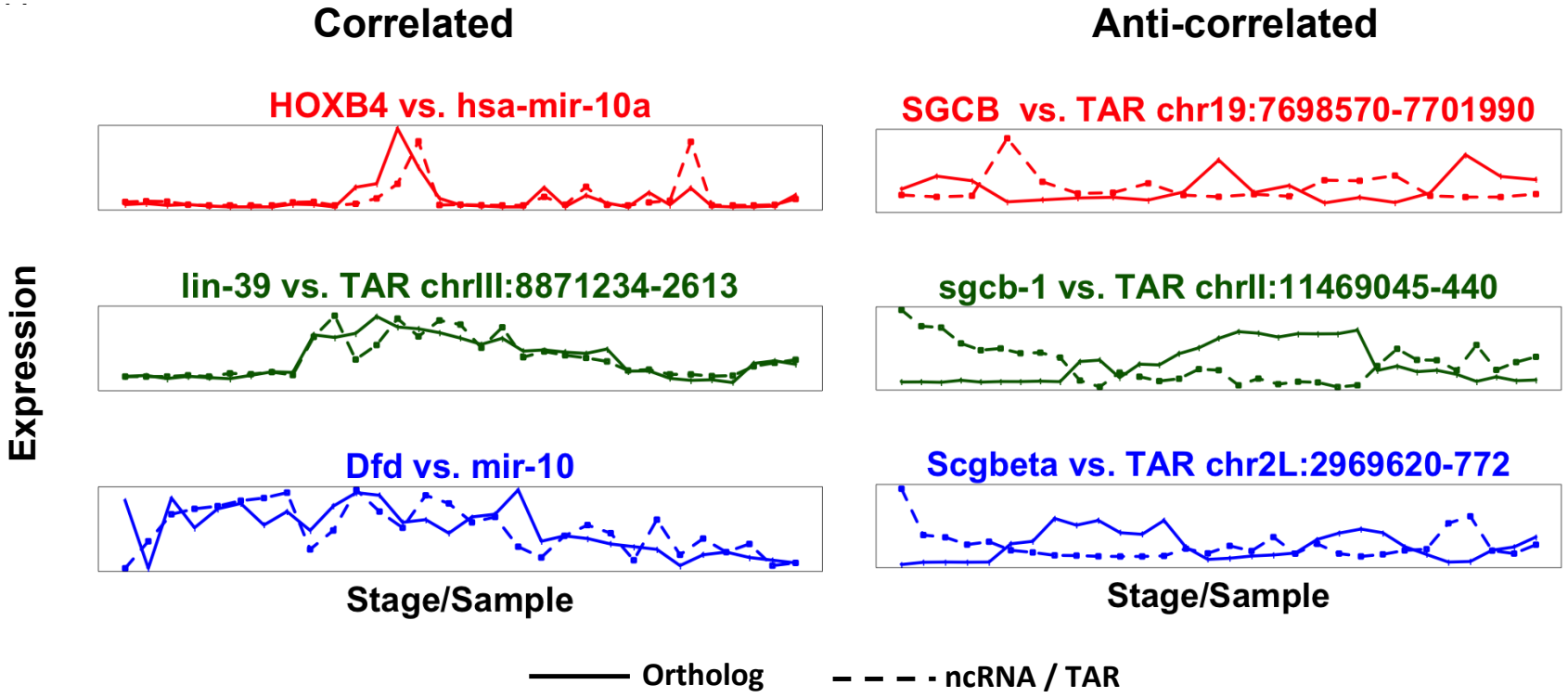
Application for 3 species



[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

ncRNAs associated with modules

- Identify ncRNAs & TARs that are significantly correlated and anti-correlated with genes in the 16 modules.



Human, **Worm** & **Fly**

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

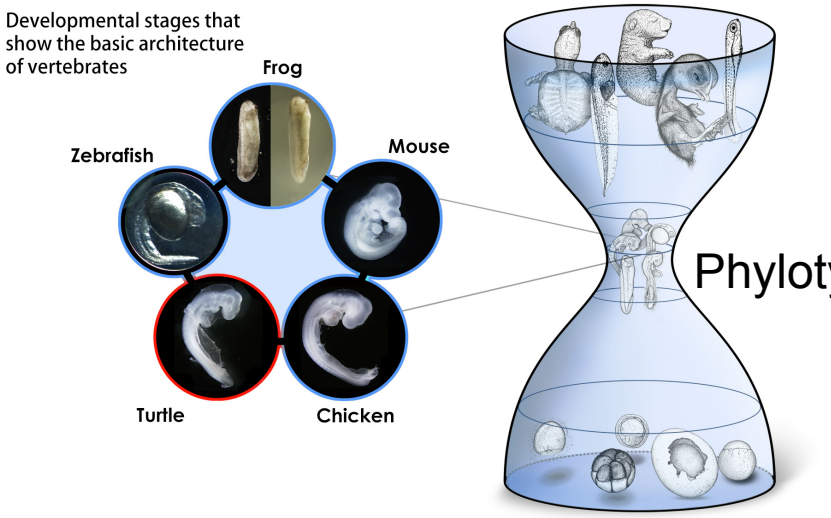
Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

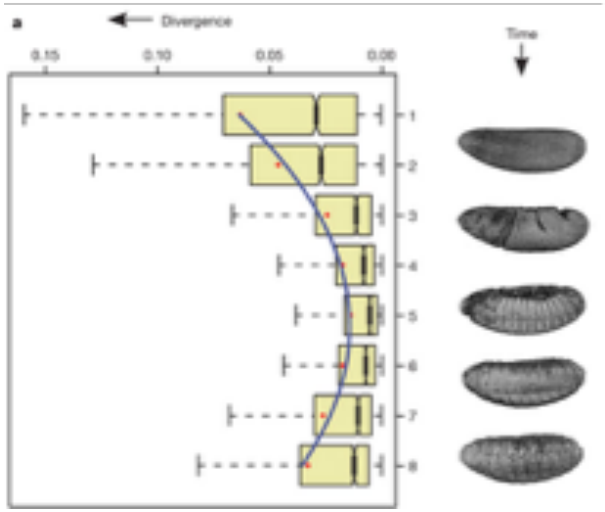
Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

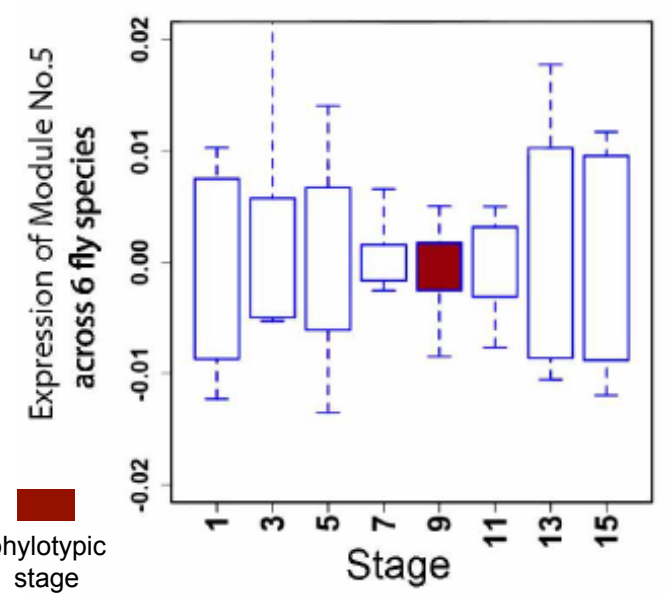
Conserved modules exhibit canonical hourglass behavior



Illustrations courtesy Naoki Irie



Expression divergence across species is minimized during phylotypic stage (Kalinka et al. Nature 2010)



Canonical Inter-organism Behavior

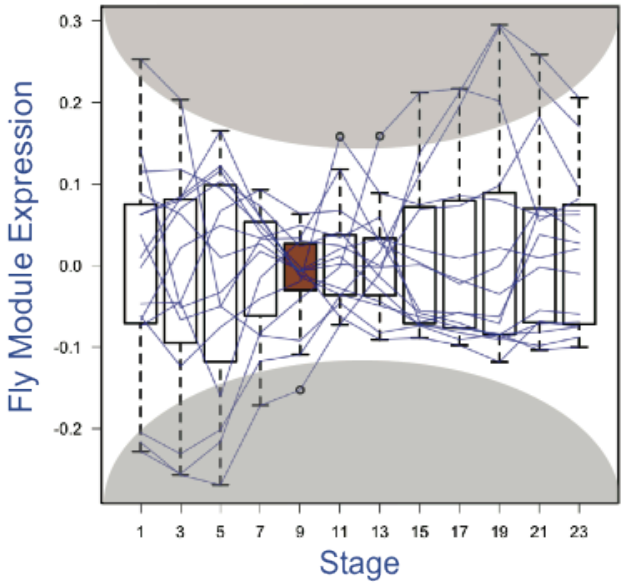
- "Hourglass hypothesis": all organisms go through a particular stage in embryonic development ("phylotypic" stage) where inter-organism expression differences of orthologous genes are smallest.
- **We identify modules (12 out of 16) which have this behavior at the phylotypic stage.**

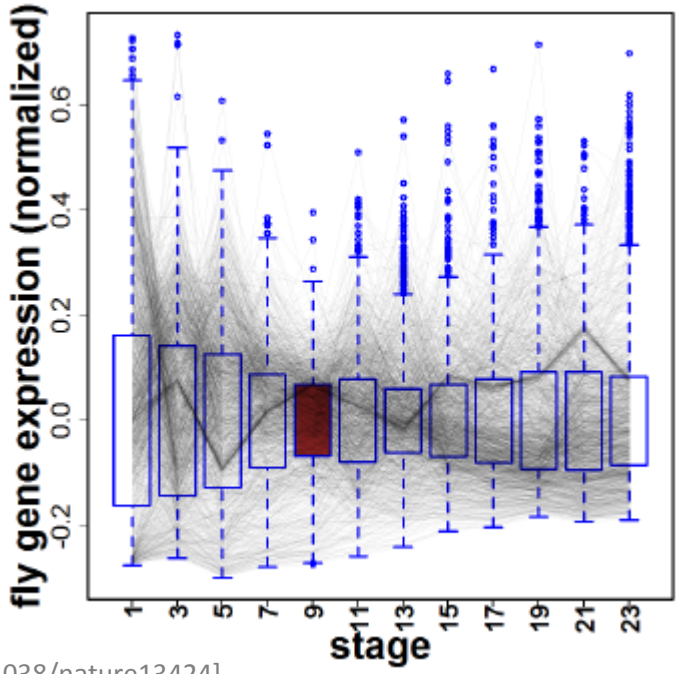
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Hourglass Behavior

Intra-organism Behavior also Present

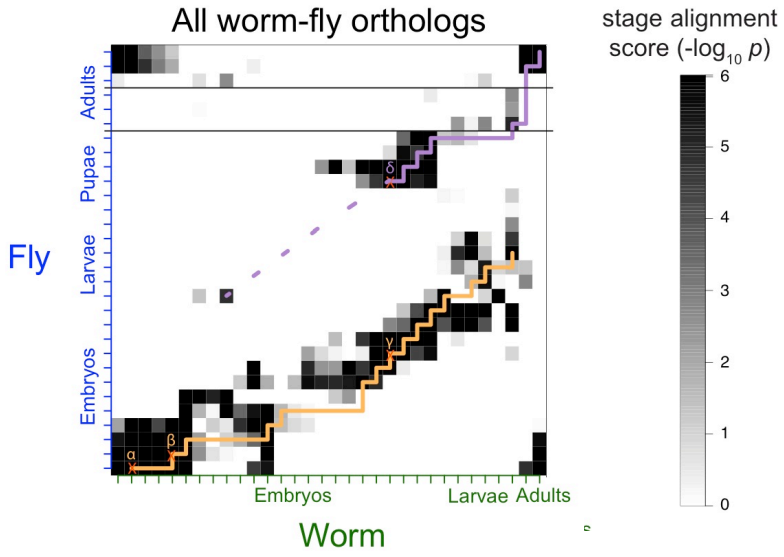
- We observe that the expression of genes across 12 modules are the most tightly coordinated at the phylotypic stage (fly).





[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

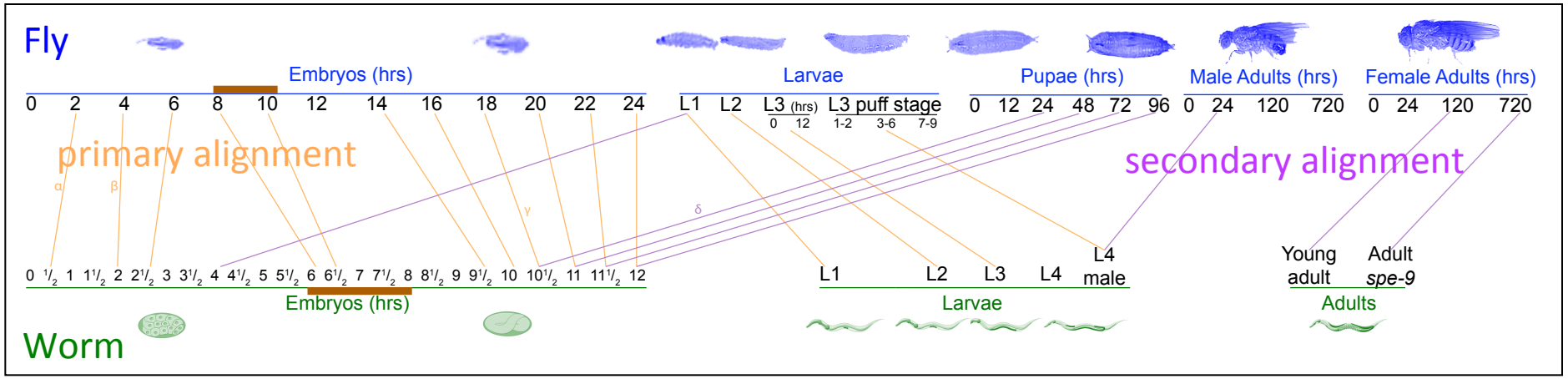
Alignment of Developmental Time-Course



For worm & fly find stage-specific genes

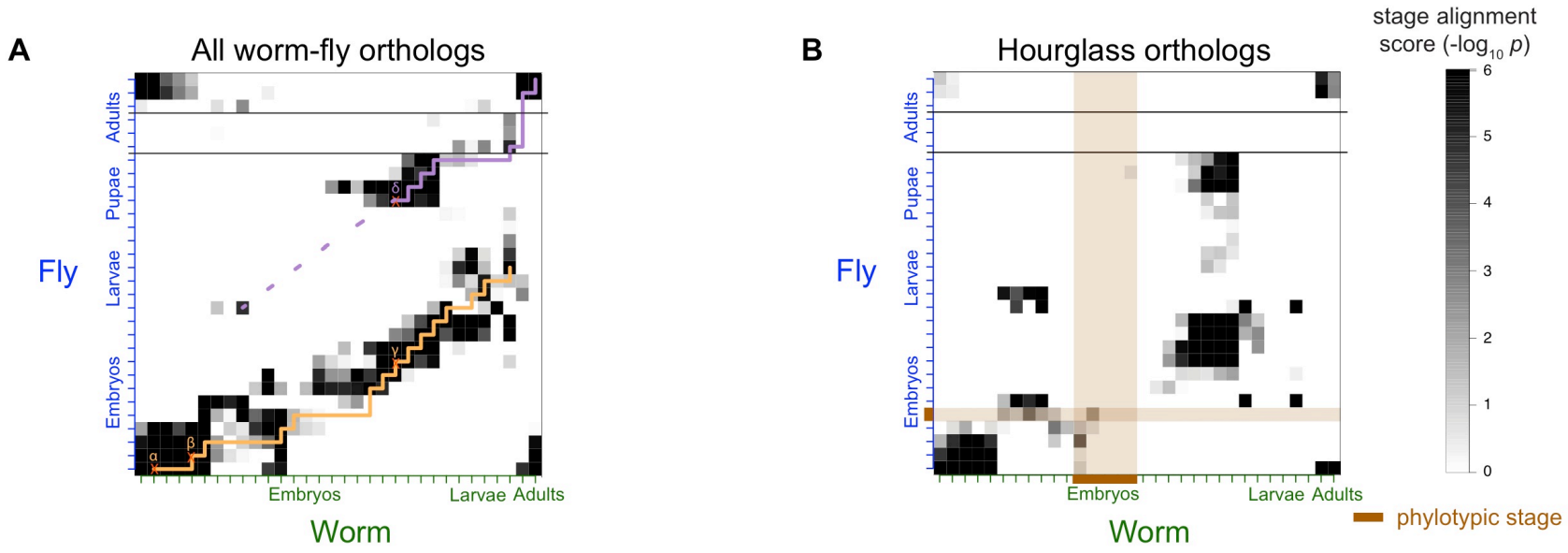
We can align developmental stages using fraction of shared orthologs between worm and fly amongst these

Reuse of genes from LE in worm in fly pupa



[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Alignment of Developmental Time-Course



Using only orthologs in 12 "hourglass" modules show stronger alignment except for absence of genes at the phylotypic stage

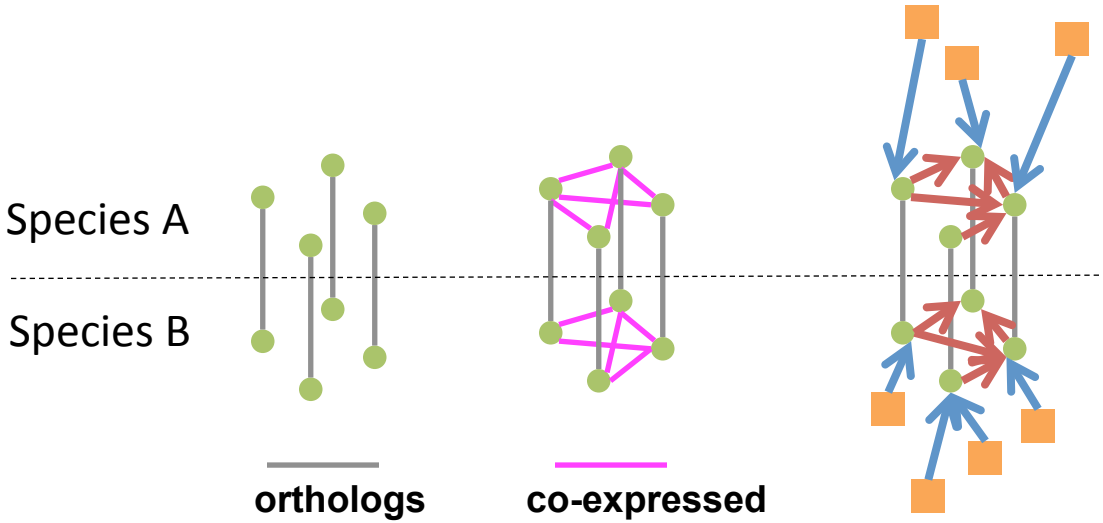
- By definition genes in hourglass modules are not phylotypic stage specific, hence the gap

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

Are gene regulations among orthologs conserved across species?



Are gene regulatory networks among orthologs conserved across species?

Regulation among orthologs (**internal**)

Regulation from species-specific factors (**external**)

Orthologous genes (orthologs)

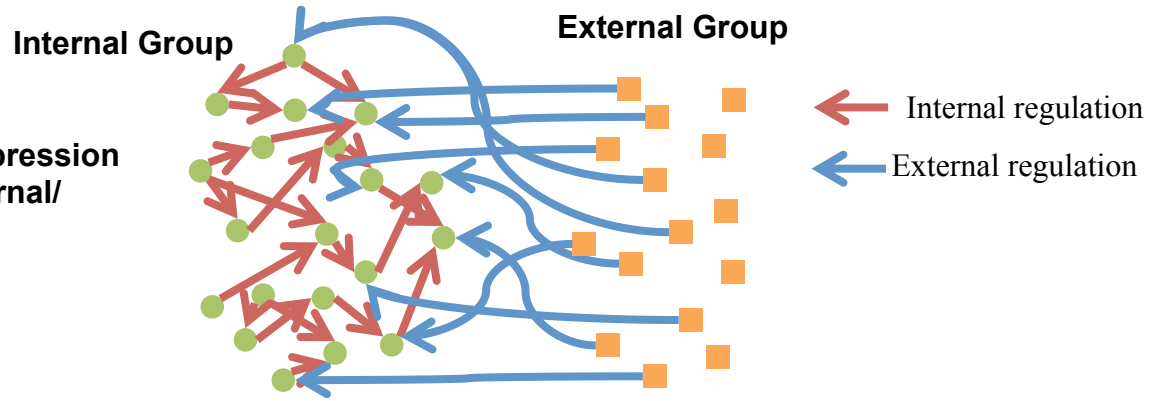
Species-specific transcription factors

To what degree can't ortholog expression levels be predicted due to species-specific regulation

[Wang et al. *PLOS CB* (in revision, '15)]

State-space model for internal and external gene regulatory networks

How to identify gene expression dynamics driven by internal/ external regulation?



State space model

$$X_{t+1} = A X_t + B U_t$$

State: Gene expression vector of Group *X* at time *t*+1

A_{ij} captures temporal casual influence from Gene *i* to Gene *j* in internal group

State: Gene expression vector of internal group at time *t*

Control: Gene expression vector of external factors at time *t*

B_{kl} captures temporal casual influence from external factor *k* to Gene *l* in internal group

[Wang et al. *PLOS CB* (in revision, '15)]

Effective state space model for meta-genes

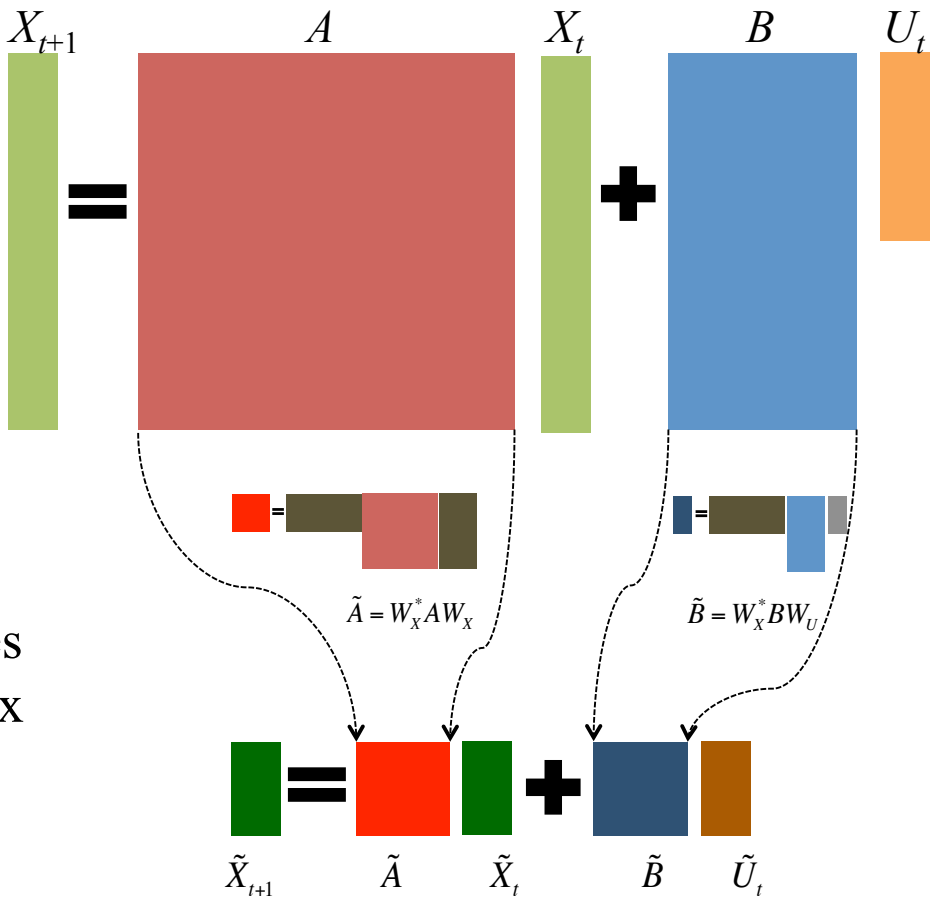
Not enough data to estimate state space model for genes (e.g., 91K time points to estimate 11.5M elements of A & B in worm)

$$X_{t+1} = AX_t + BU_t$$

Dimensionality reduction from genes to meta-genes (e.g., SVD)

Effective state space model for meta-genes (e.g., 250 time points to estimate 50 matrix elements if 5 worm meta-genes)

$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$$



[Wang et al. *PLOS CB* (in revision, '15)]

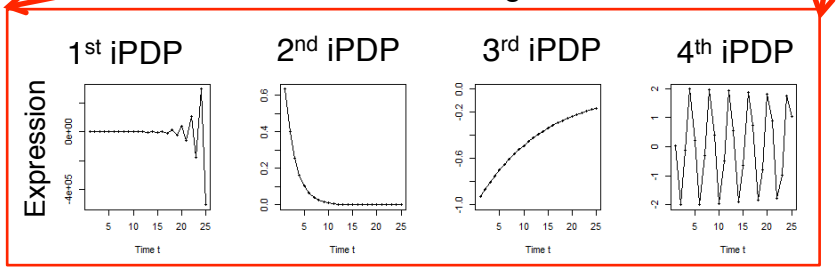
Orthologs have similar internal but different external dynamic patterns during embryonic development

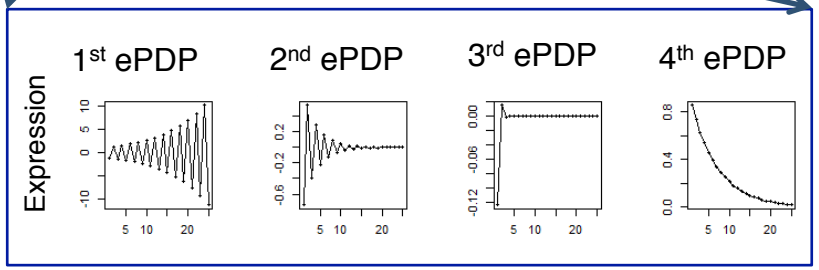
Worm's effective state space model

$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$$

iPDPs: time exponentials of $\tilde{A}$ eigenvalues in worm

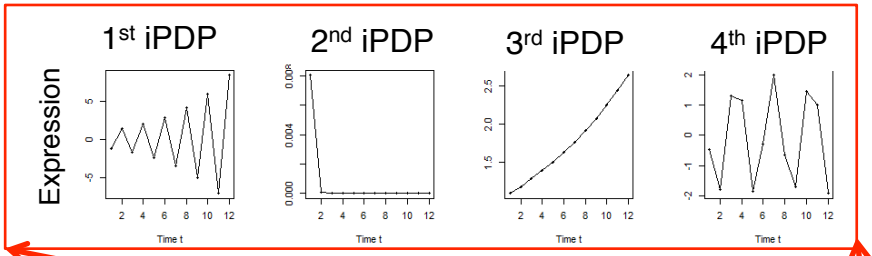
ePDPs: time exponentials of $\tilde{B}$ eigenvalues in worm

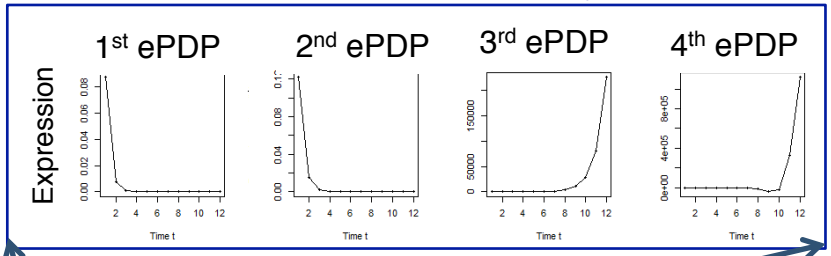




Similar iPDP canonical trajectories

Different ePDP canonical trajectories





iPDPs: time exponentials of $\tilde{A}$ eigenvalues in fly

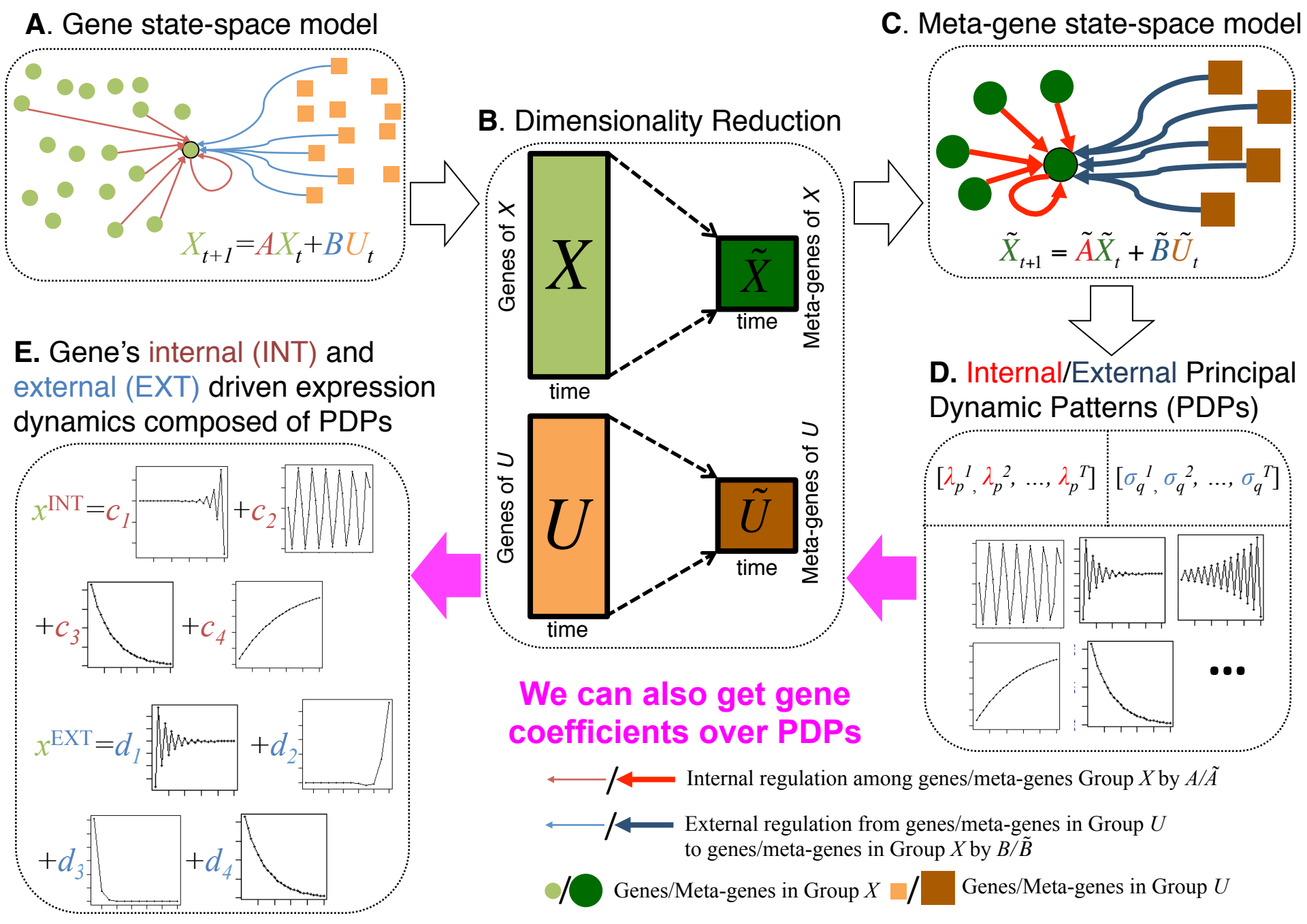
ePDPs: time exponentials of $\tilde{B}$ eigenvalues in fly

Fly's effective state space model

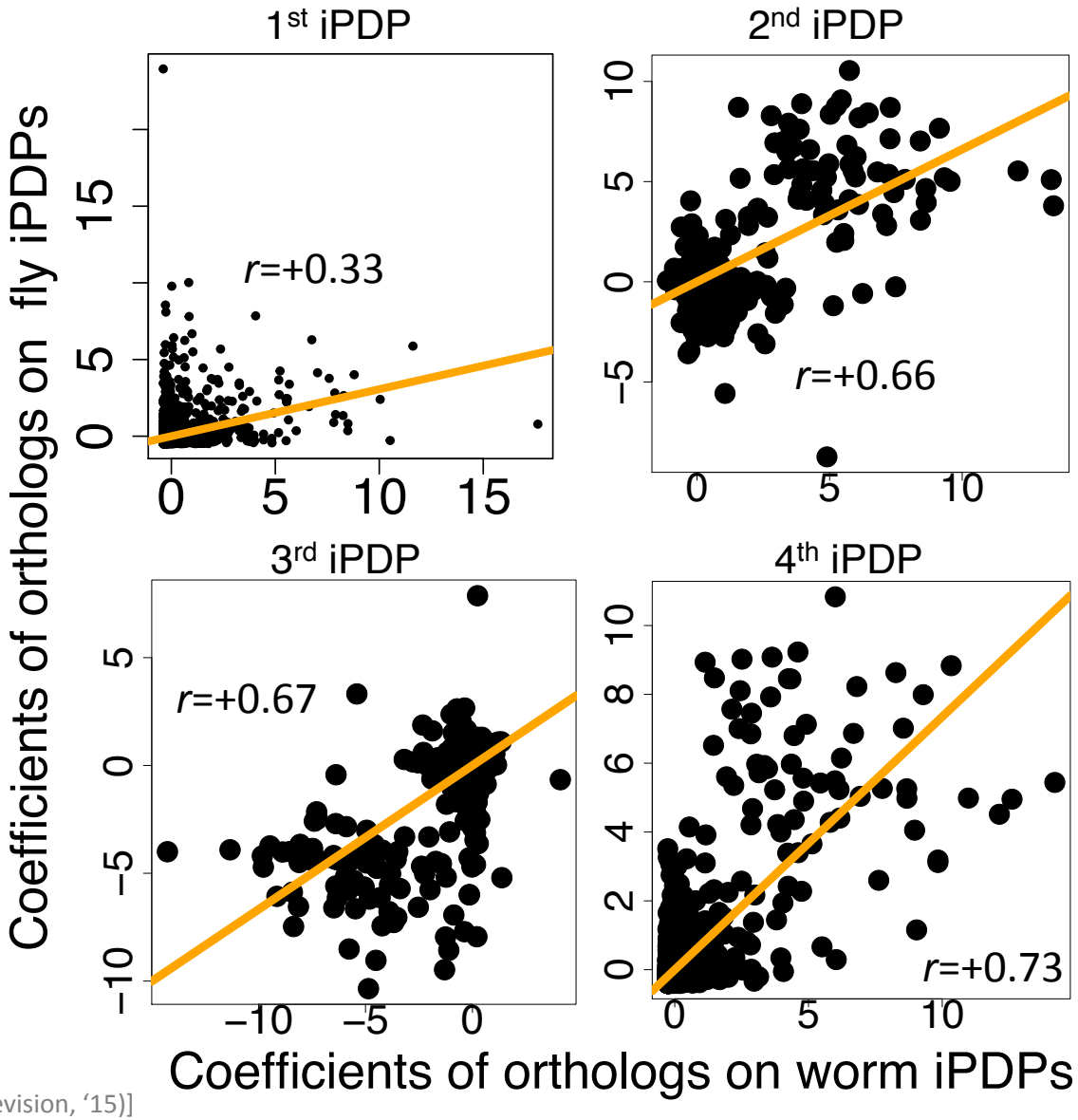
$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$$

[Wang et al. *PLOS CB* (in revision, '15)]

[Wang et al. *PLOS CB* (in revision, '15)]



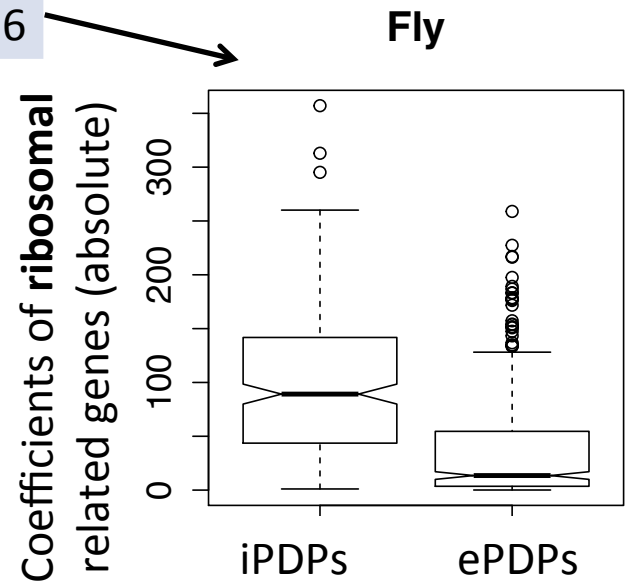
Orthologs have correlated iPDP coefficients



[Wang et al. *PLOS CB* (in revision, '15)]

Evolutionarily conserved and younger genes exhibit the opposite internal and external PDP coefficients

iPDP coeffs > ePDP coeffs	Worm	Fly
Ribosomal genes	$p<0.001$	$p<2.2e-16$



Ribosomal genes have significantly larger coefficients for the internal than external PDPs, but signaling genes exhibit the opposite trend

iPDP coeffs < ePDP coeffs	Worm	Fly
Signaling genes	$p<7e-4$	$p<6e-4$

* *p*-values from KS-test

[Wang et al. *PLOS CB* (in revision, '15)]

Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

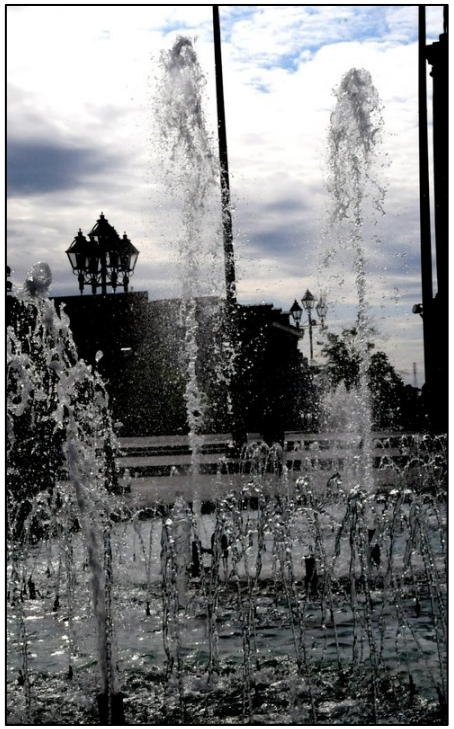
Transcriptome Analysis: Expression Clustering across Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering, Cross-species**
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **Relating Clusters to Hourglass Genes**
  - Developmental 'hourglass' genes in 12 of the clusters. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones**
  - Using dimensionality reduction to help determine internal & external drivers
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

Acknowledgements

modENCODE/ENCODE Transcriptome group [EncodeProject.org/comparative]

Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B. Brown, Carrie A. Davis, LaDeana Hillier, Cristina Sisu, **Jingyi Jessica Li,** Baikang Pei, Arif O. Harmanci, Michael O. Duff, Sarah Djebali, Roger P. Alexander, Burak H. Alver, Raymond K. Auerbach, Kimberly Bell, Peter J. Bickel, Max E. Boeck, Nathan P. Boley, Benjamin W. Booth, Lucy Cherbas, Peter Cherbas, Chao Di, Alex Dobin, Jorg Drenkow, Brent Ewing, Gang Fang, Megan Fastuca, Elise A. Feingold, Adam Frankish, Guanjun Gao, Peter J. Good, Phil Green, Roderic Guigó, Ann Hammonds, Jen Harrow, Roger A. Hoskins, Cédric Howald, Long Hu, Haiyan Huang, Tim J. P. Hubbard, Chau Huynh, Sonali Jha, Dionna Kasper, Masaomi Kato, Thomas C. Kaufman, Rob Kitchen, Erik Ladewig, Julien Lagarde, Eric Lai, Jing Leng, **Zhi Lu,** Michael MacCoss, Gemma May, Rebecca McWhirter, Gennifer Merrihew, David M. Miller, Ali Mortazavi, Rabi Murad, Brian Oliver, Sara Olson, Peter Park, Michael J. Pazin, Norbert Perrimon, Dmitri Pervouchine, **Valerie Reinke**, Alexandre Reymond, Garrett Robinson, Anastasia Samsonova, Gary I. Saunders, Felix Schlesinger, Anurag Sethi, Frank J. Slack, William C. Spencer, Marcus H. Stoiber, Pnina Strasbourger, Andrea Tanzer, Owen A. Thompson, Kenneth H. Wan, Guilin Wang, Huaien Wang, Kathie L. Watkins, Jiayu Wen, Kejia Wen, Chenghai Xue, Li Yang, Kevin Yip, Chris Zaleski, Yan Zhang, Henry Zheng, **Steven E. Brenner, Brenton R. Graveley, Susan E. Celniker, Thomas R Gingeras, Robert Waterston**



Hiring Postdocs. See gersteinlab.org/jobs !

Models
Acknowledgements

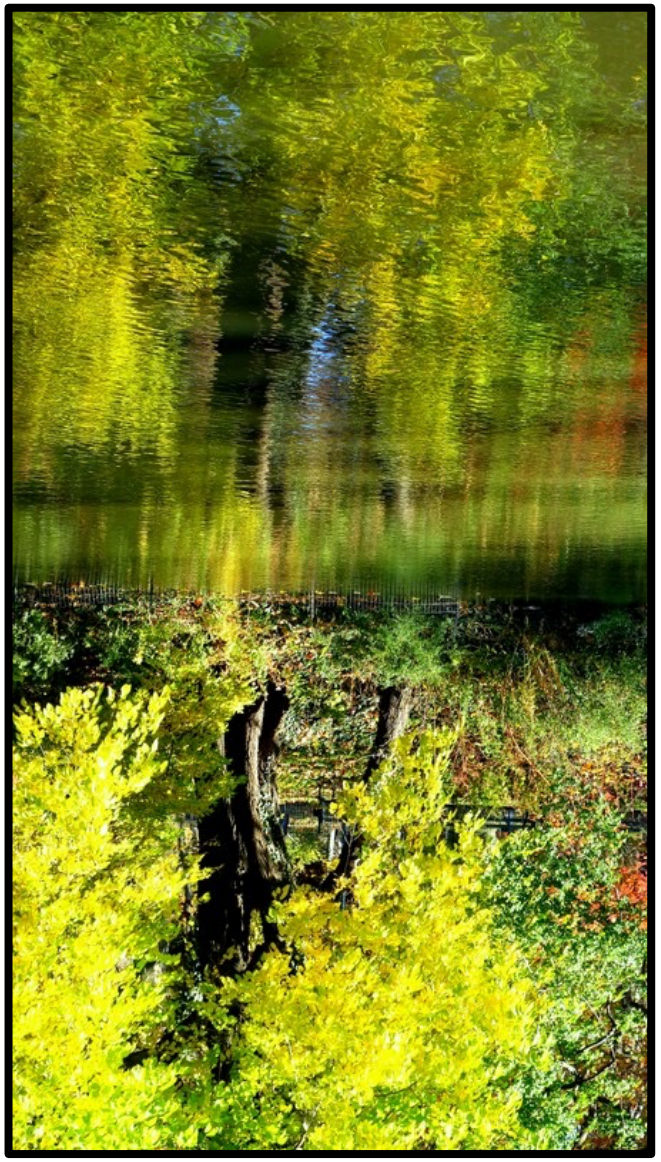
ORTHOCLUST.gersteinlab.org :

KK Yan, D Wang,

J Rozowsky, H Zheng, C Cheng

DREISS.gersteinlab.org

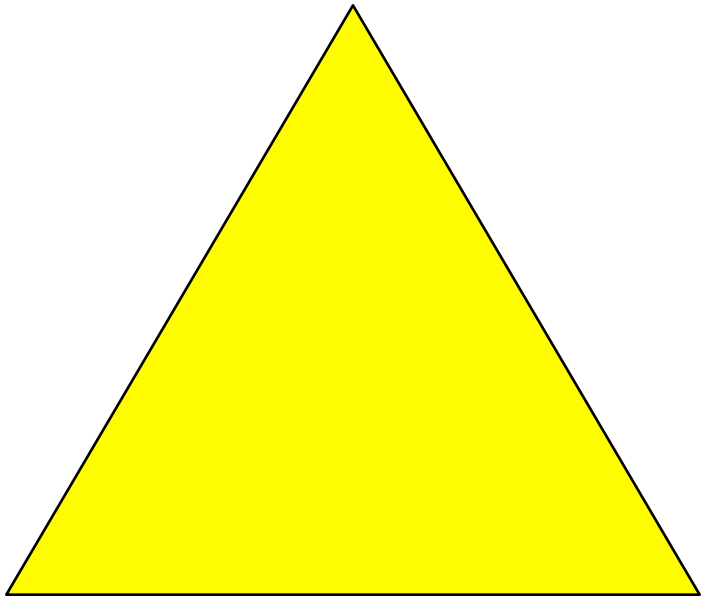
D Wang, F He, S Maslov



Hiring Postdocs. See gersteinlab.org/jobs !

Default Theme

- Default Outline Level 1
  - Level 2



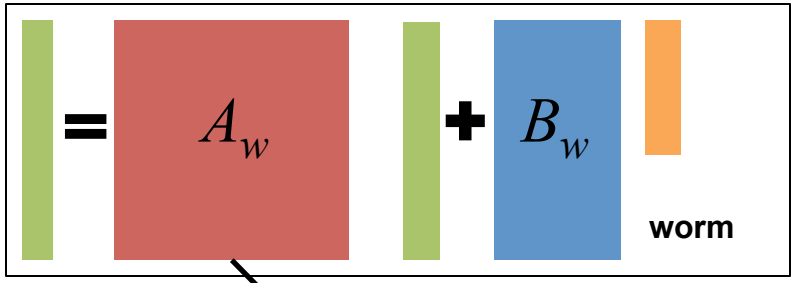
Info about content in this slide pack

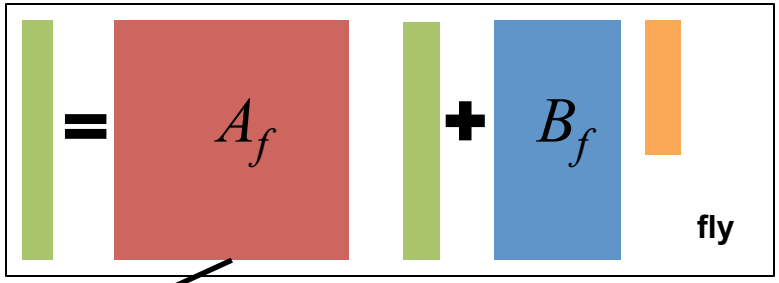
- PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2012 (and beyond). Please read statement at http://www.gersteinlab.org/misc/permissions.html . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to appropriate place on gersteinlab.org).
- Paper references in the talk were mostly from Papers.GersteinLab.org.
- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info . In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt

Are there any conserved regulatory networks between worm and fly during embryonic development?

Dataset	Internal Group	External Group	Developmental stages	# of unknown parameters in A and B	# of available time samples
worm (*C. elegans*)	N_1=3147 worm-fly orthologs (incl. ortholog TFs)	N_2=509 worm-specific transcription factors	T=25 time points: 0, 0.5, 1, …, 12 hours	3147*3147+3147*509=11.5M	3147*25+509*25=91400
fly (*D. mel.*)		N_2=442 fly-specific transcription factors	T=12 time points: 0, 2, 4, 6, 8,…, 20, 22 hours	3147*3147+3147*442=11.3M	3147*25+442*25=89725

No enough time samples!





If A_w and A_f have similarities, cross-species conserved regulatory networks in embryonic development

Embryonic stem cells (ESCs)