Genomics:

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

Mark Gerstein
Yale

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

See last slide for more info.

How might annotate a text?

Color is Function

Lines are Similarity

The Semicolon Wars

Brian Hayes

If you want to be a thoroughgoing world traveler, you need to learn 6,912 ways to say "Where is the toilet, please?" That's the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot programmer, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we've been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

Every programmer knows there is one true programming language. A new one every week

a good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they'll concede, but it's built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will cide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that's *not* what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in `x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C

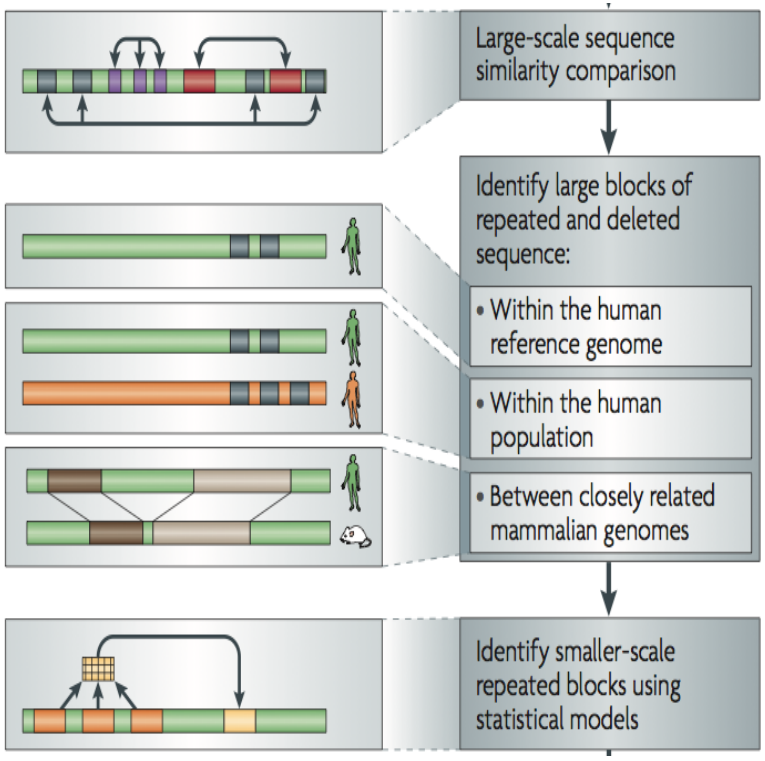
[B Hayes, Am. Sci. (Jul.- Aug. '06)]

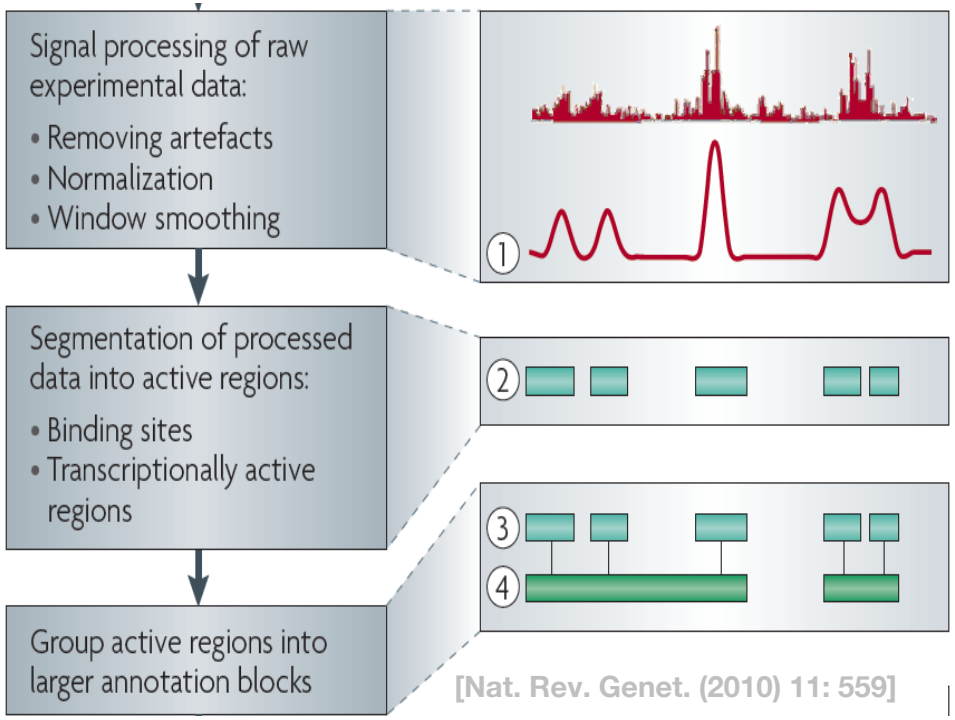
Non-coding Annotations: Overview

There are several collections of information "tracks" related to non-coding features

Sequence features, incl. **Conservation**

Functional Genomics
ChIP-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription



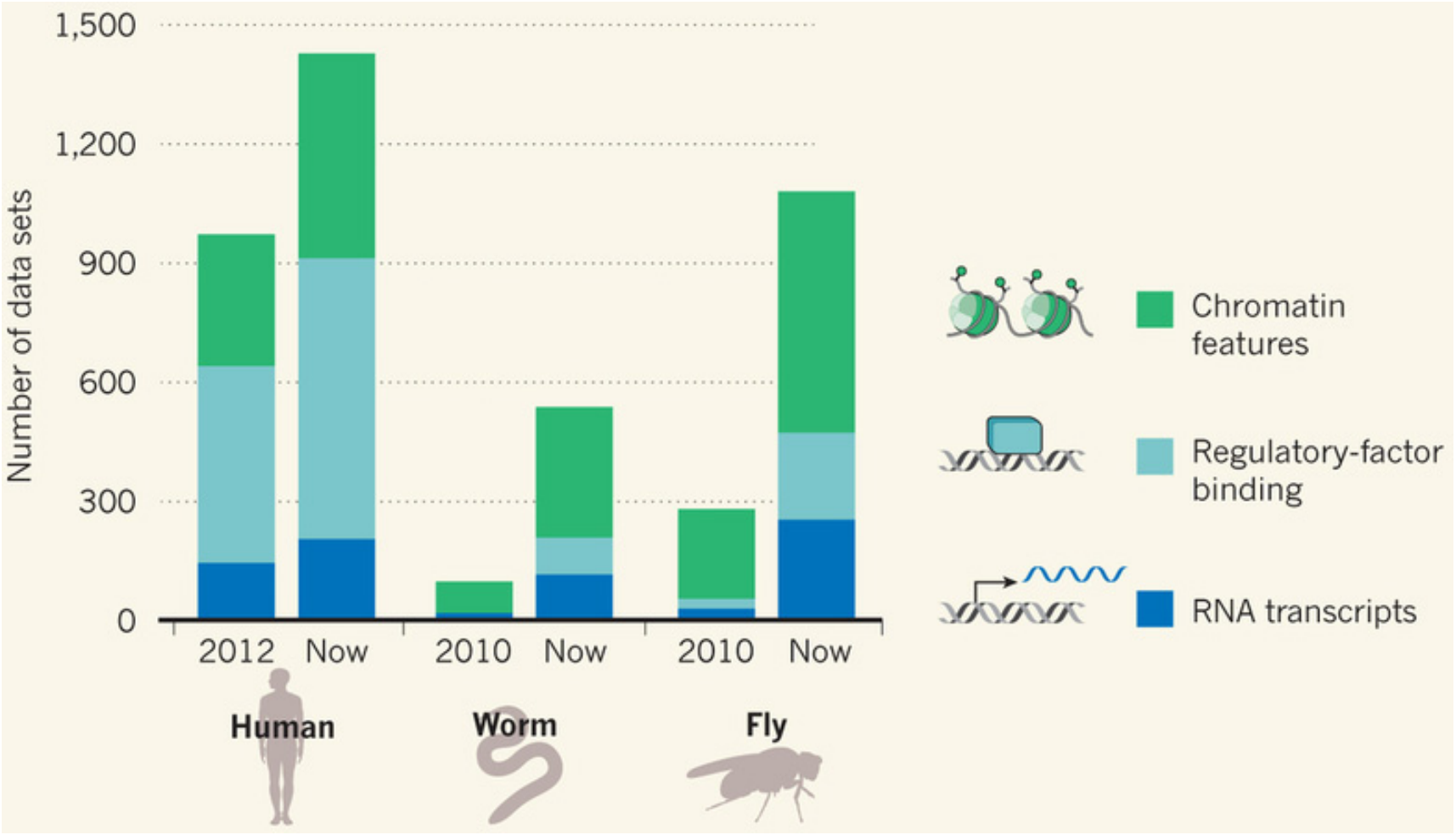


[Nat. Rev. Genet. (2010) 11: 559]

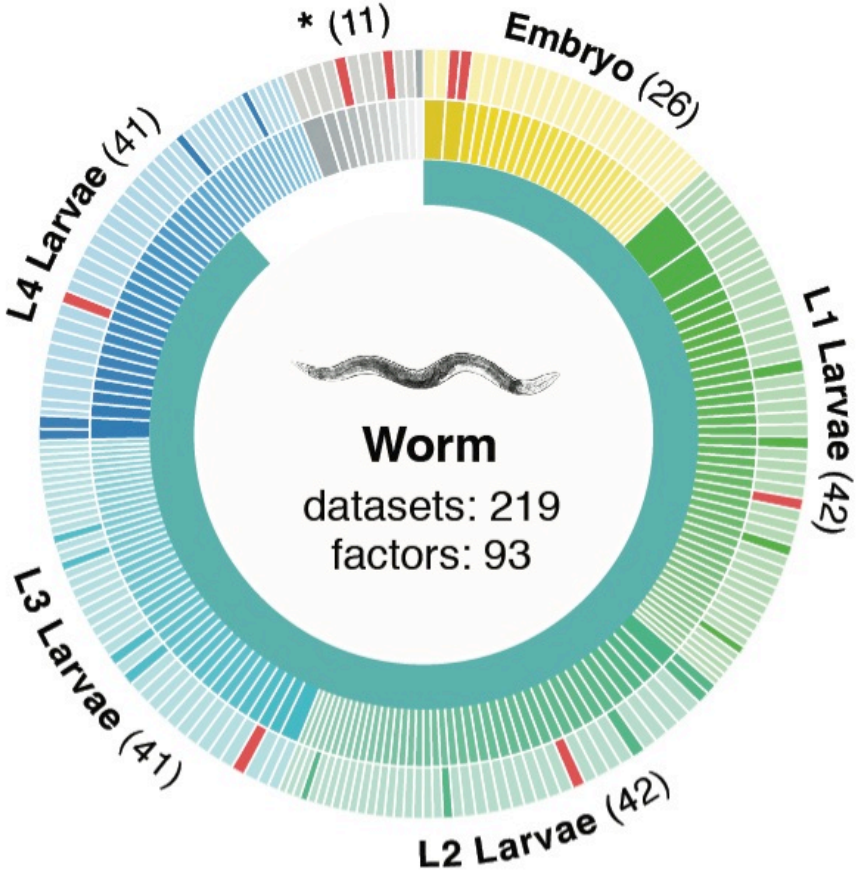
Comparative ENCODE Functional Genomics Resource

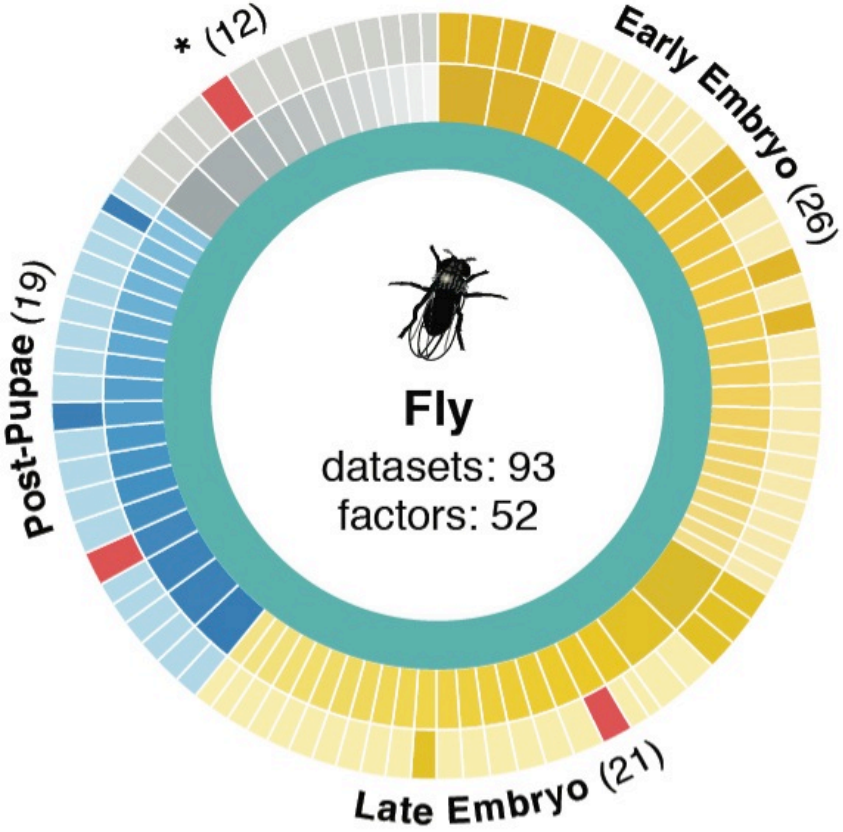
(EncodeProject.org/comparative)

- 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



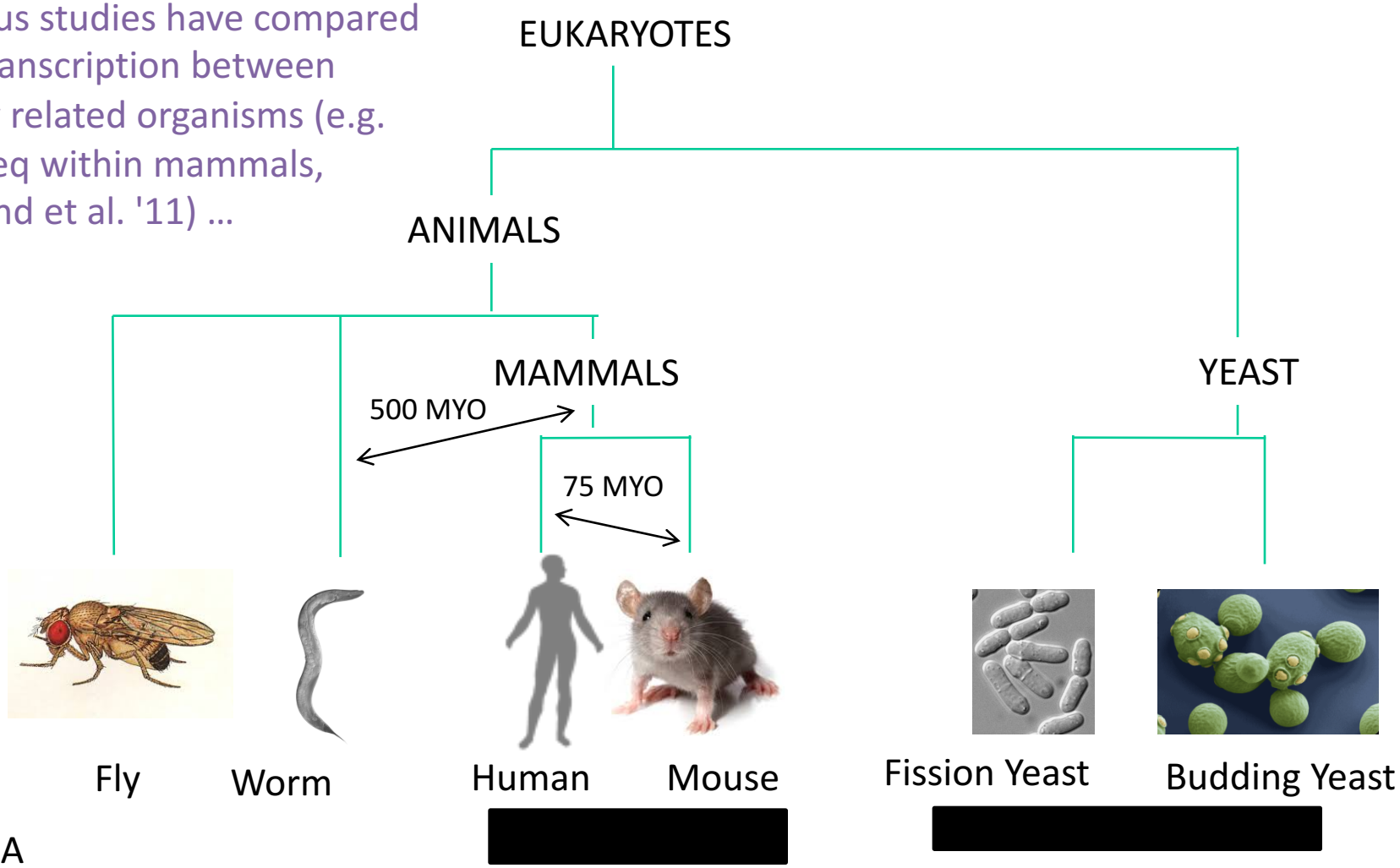


Organism	Major developmental stages
worm (*C. elegans*)	33 stages: 0, 0.5, 1, …, 12 hours, L1, L2, L3, L4, …, Young Adults, Adults
fly (*D. mel.*)	30 stages: 0, 2, 4, 6, 8,…, 20, 22 hours, L1-L4, Pupaes, Adults

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

Comparative ENCODE

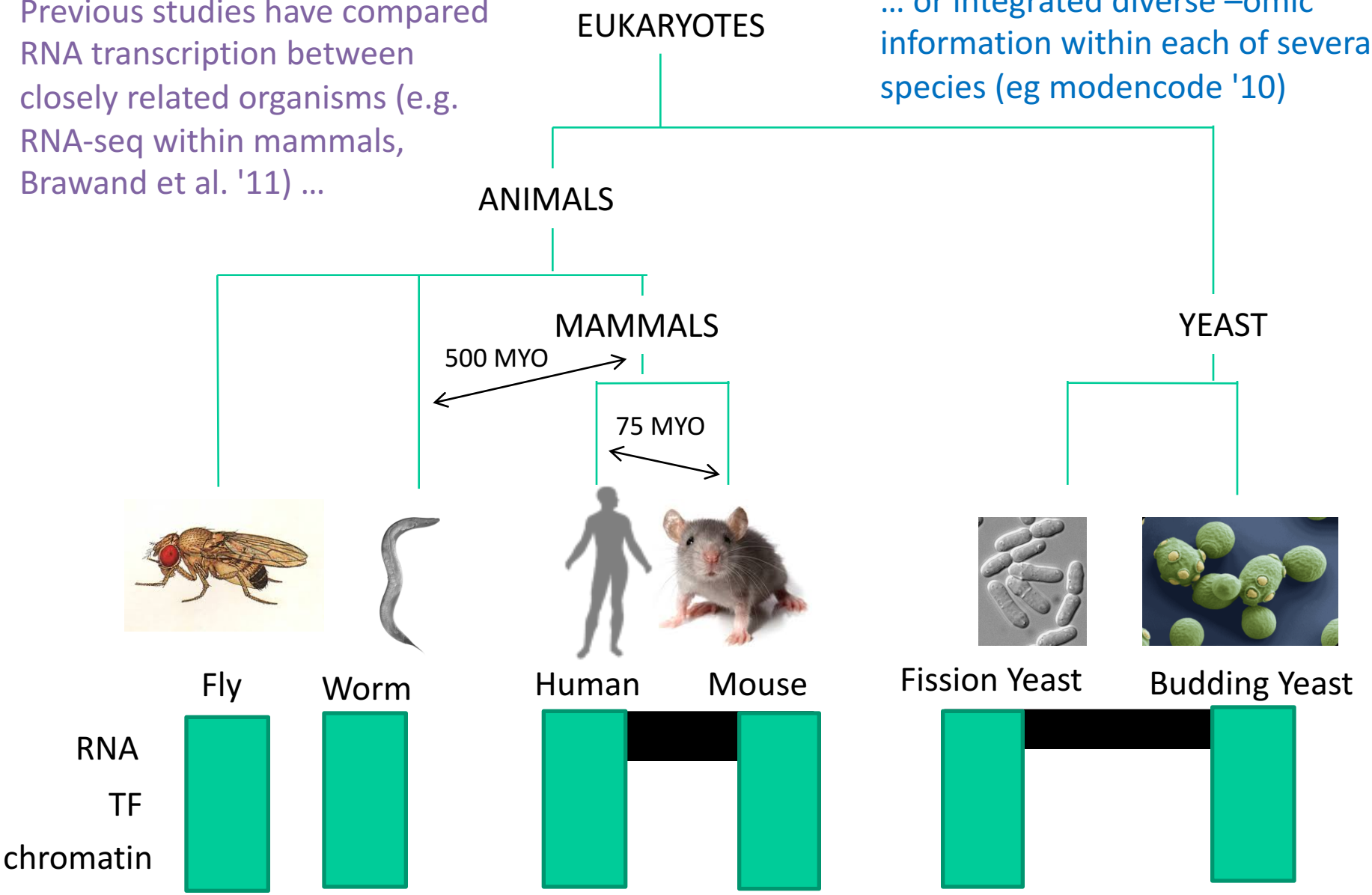
Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) …



Comparative ENCODE

Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) …

… or integrated diverse –omic information within each of several species (eg modencode '10)

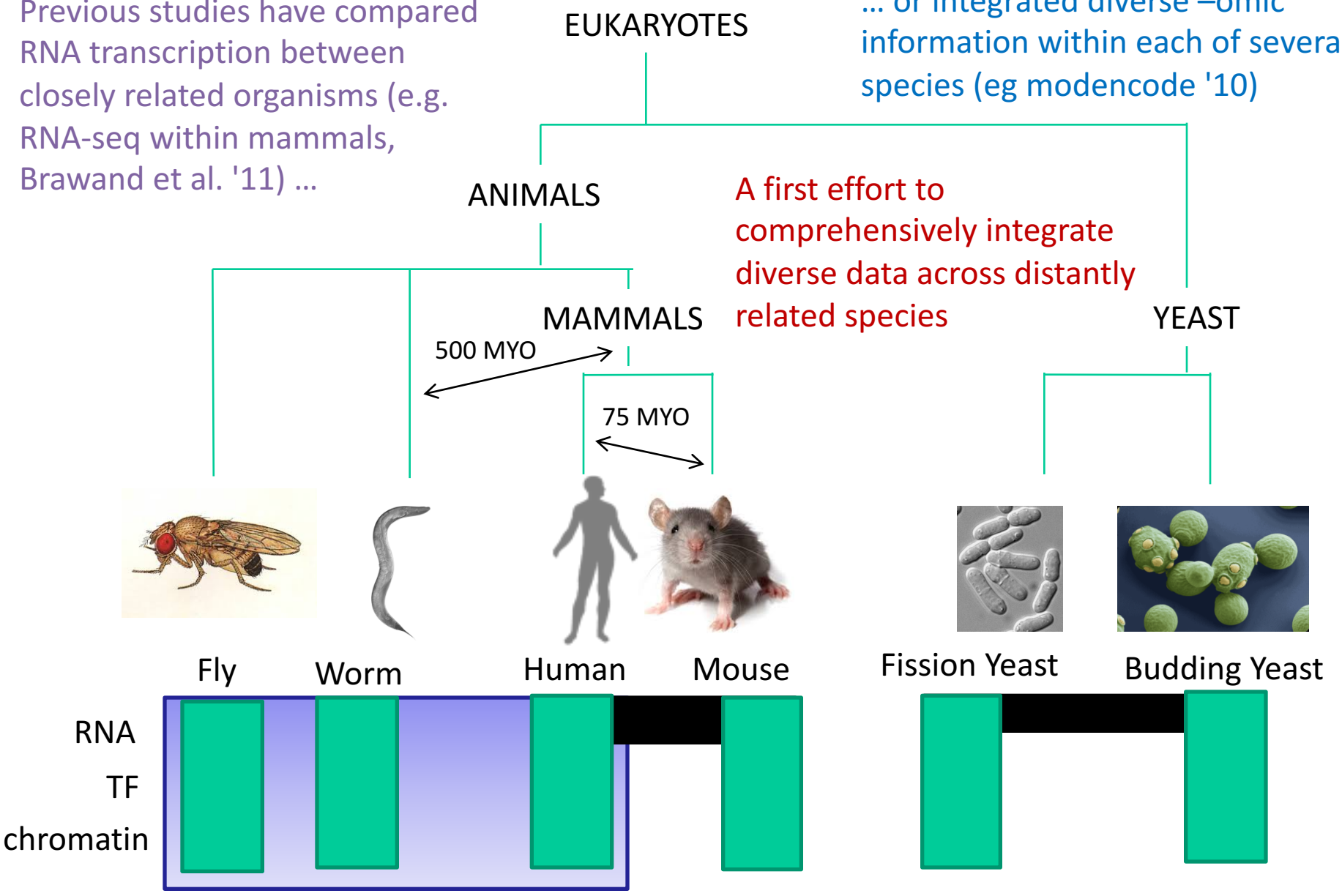


Comparative ENCODE

Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) …

… or integrated diverse –omic information within each of several species (eg modencode '10)

A first effort to comprehensively integrate diverse data across distantly related species



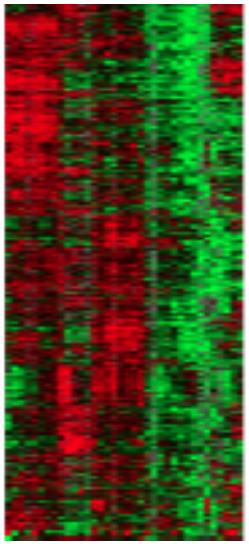
Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

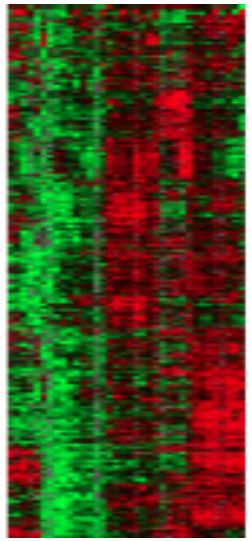
- Intro to **Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Expression clustering: revisiting an ancient problem



Species A

Species B

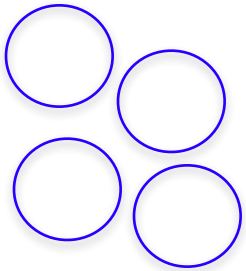


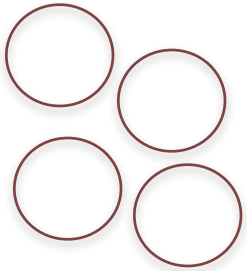
Clustering algorithm

Clustering algorithm

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

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

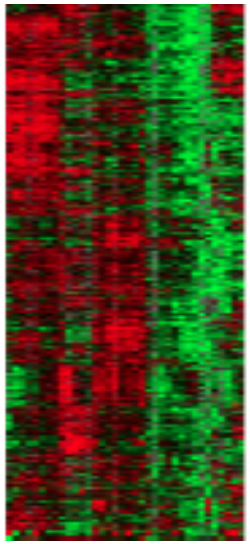




two independent sets of modules

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

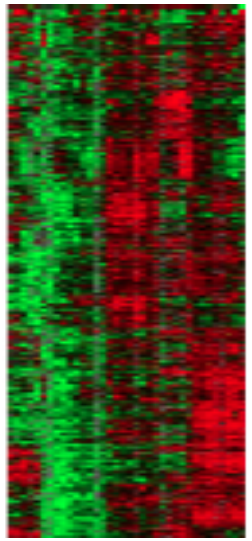
Expression clustering: revisiting an ancient problem



Species A

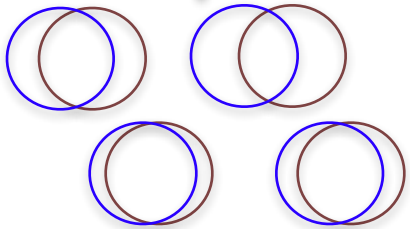
Orthologous pairs between species

Species B



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

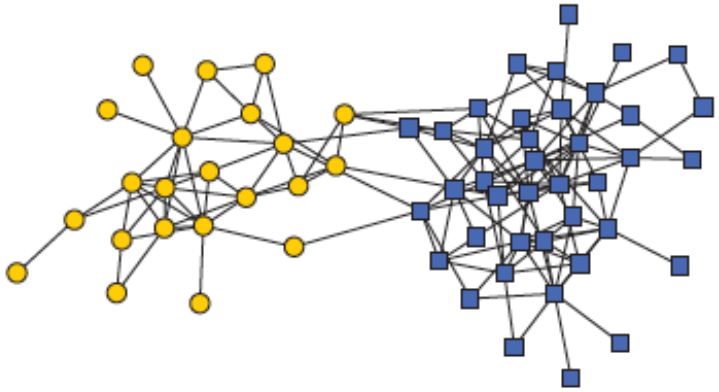
OrthoClust



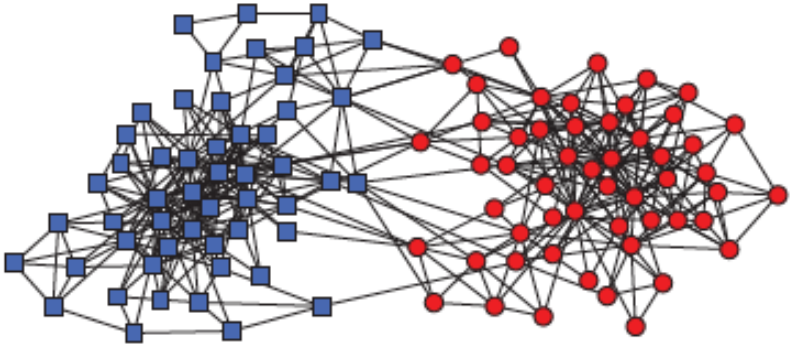
cross species modules

Yan et al. Genome Biol. 2014

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

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

adjacency matrix

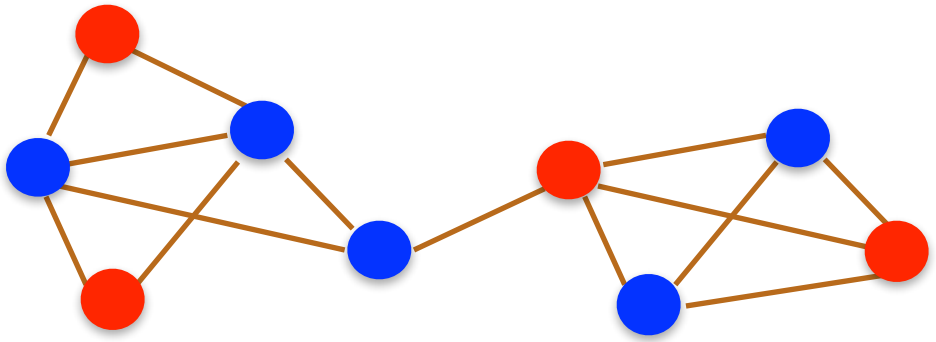
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q \approx 0$$

degree of node i

adjacency matrix

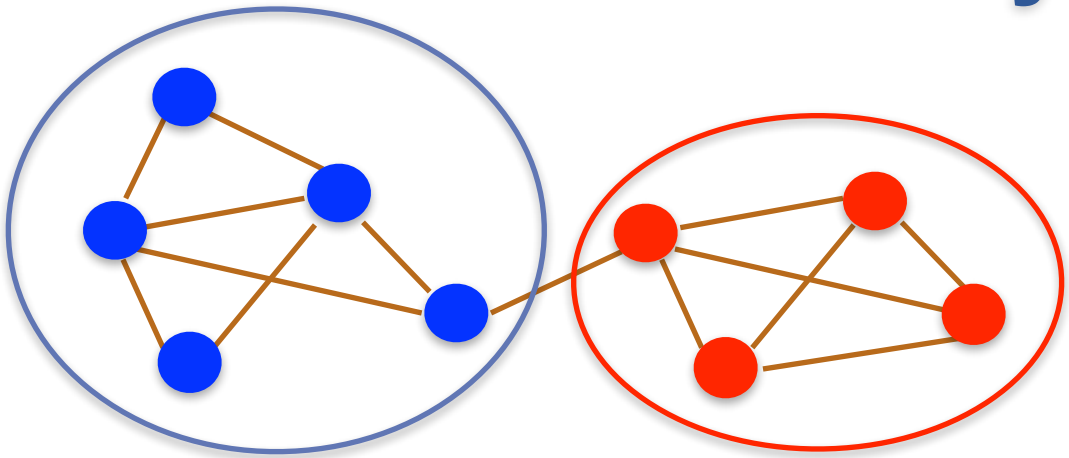
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

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

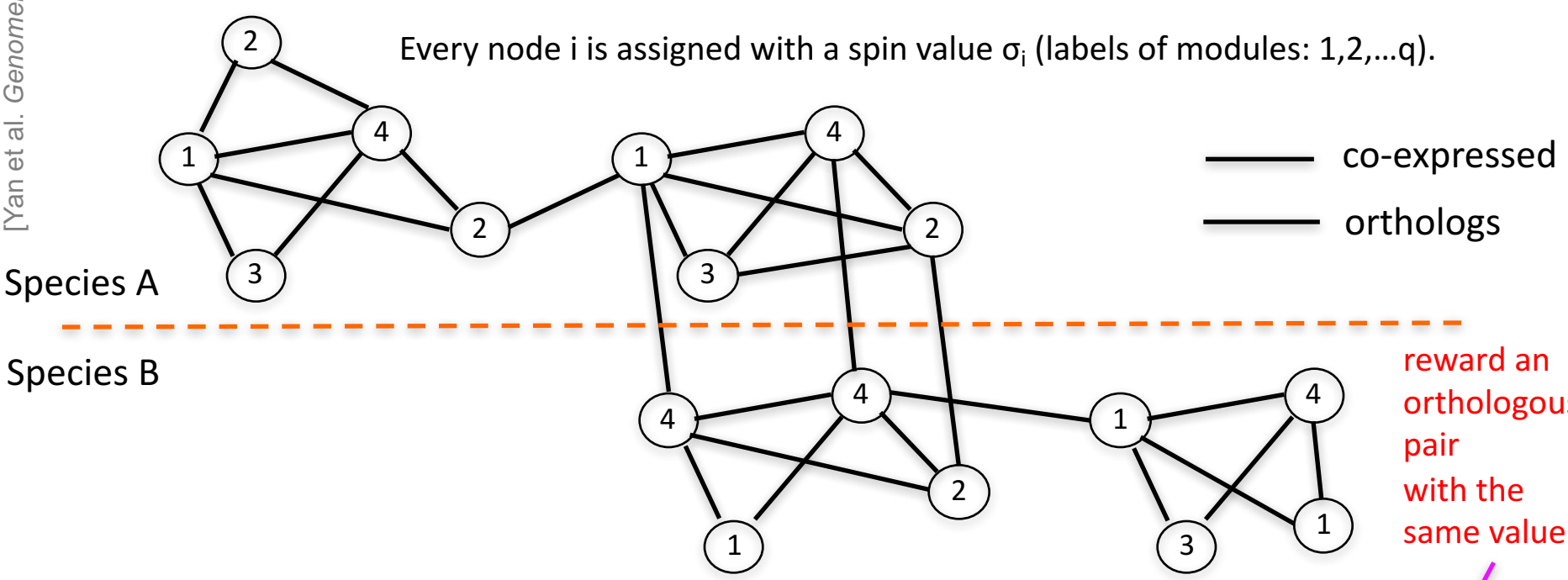
number of edges

expected number of edges between i and j

whether or not i, j are in the same module

A toy example [orthoclust]

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



$$H = \sum_{i,j}\left(-W_{ij}^{(A)} + p_{ij}^{(A)}\right)\delta_{\sigma_i\sigma_j} + \sum_{i',j'}\left(-W_{i'j'}^{(B)} + p_{i'j'}^{(B)}\right)\delta_{\sigma_{i'}\sigma_{j'}} - \kappa \sum_{(i,j')\in Ortho}\delta_{\sigma_i\sigma_{j'}}$$

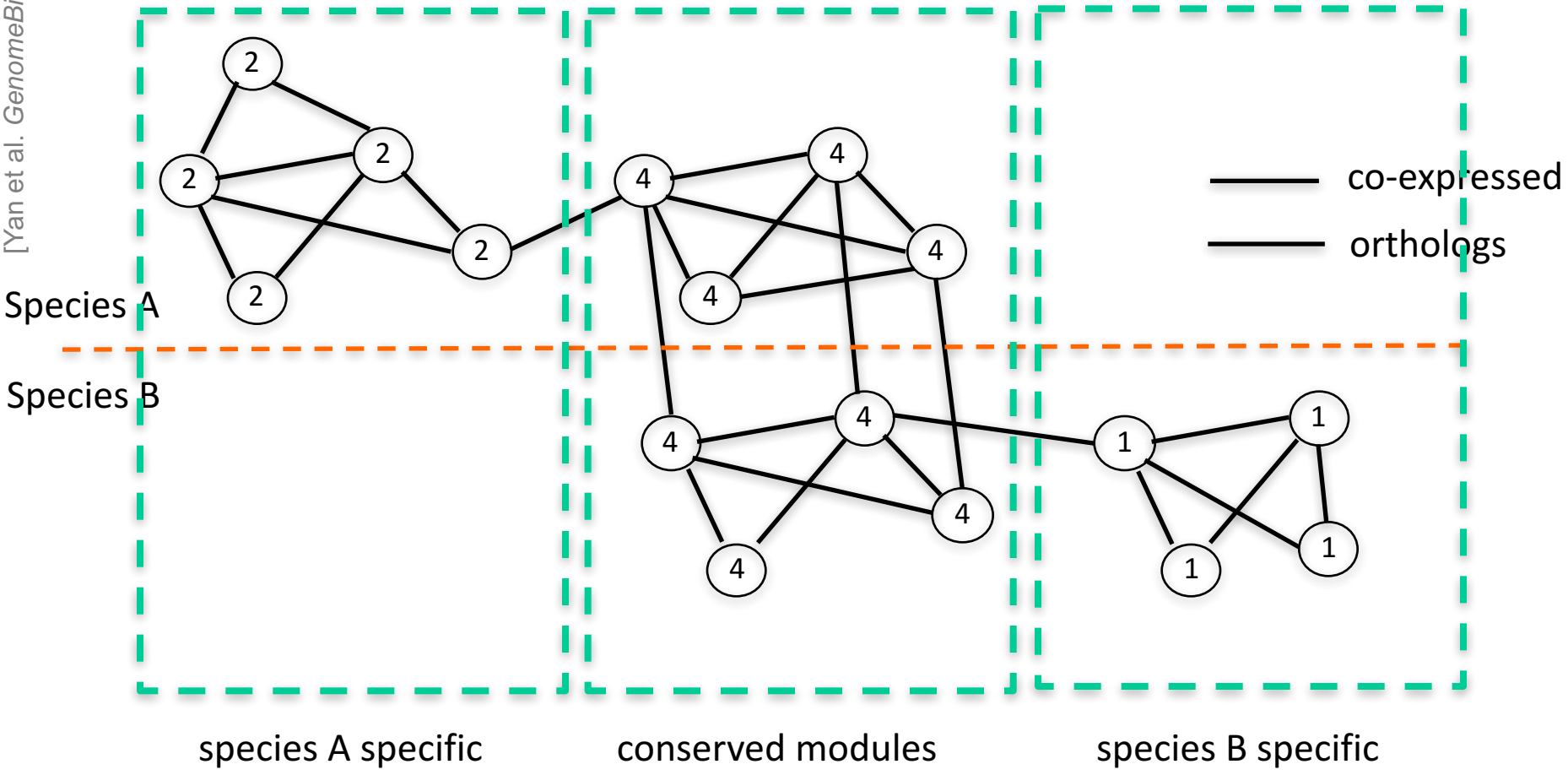
reward a co-expressed pair with the same value

punish a non co-expressed pair with the same value

Favorableness = "Modularity" in species A + "Modularity" in species B + consistency betw. A & B

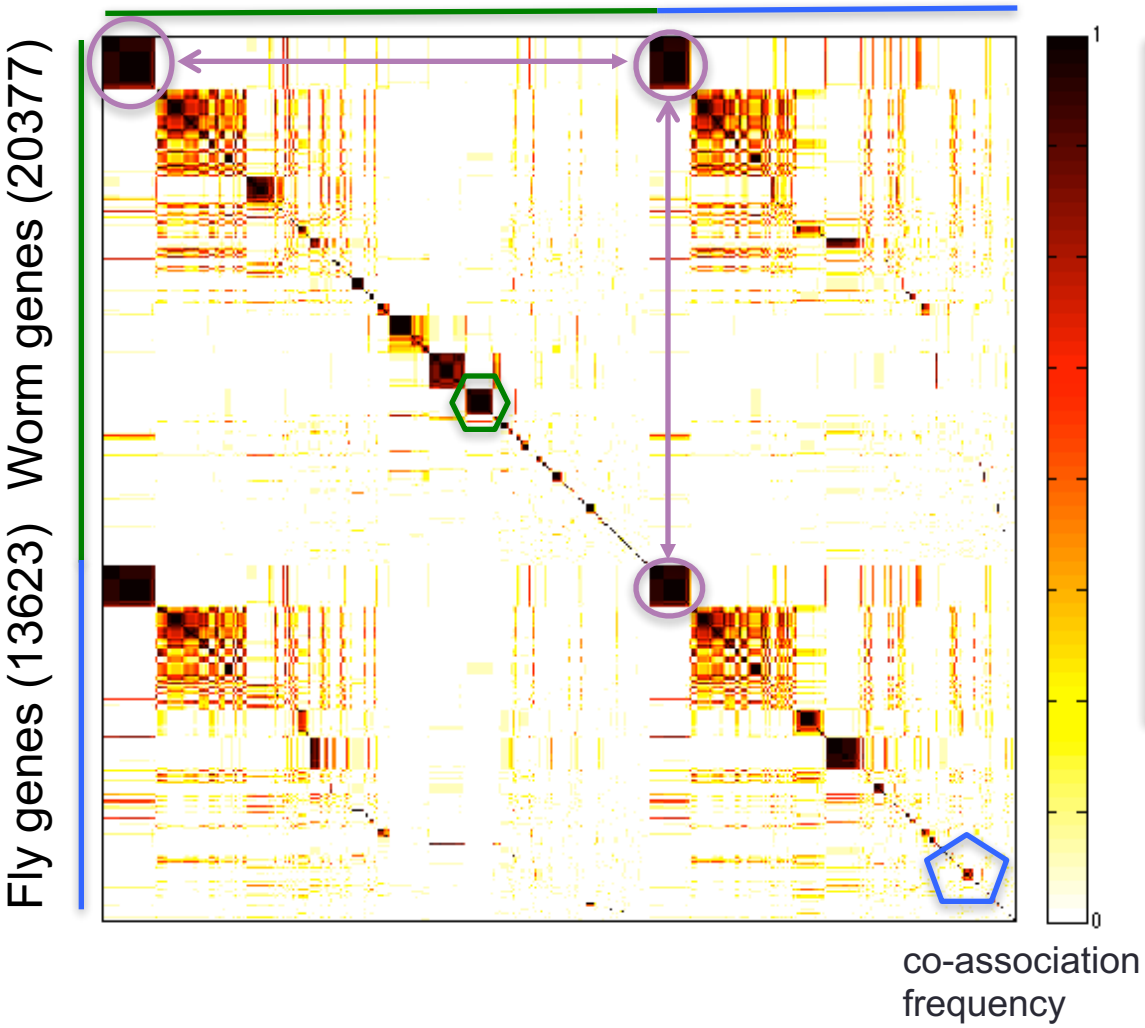
[Yan et al. *GenomeBiol* 15:R100 ('14)]

A toy example [orthoclust]



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.

Cross-species clusters for worm and fly



GO terms of conserved modules

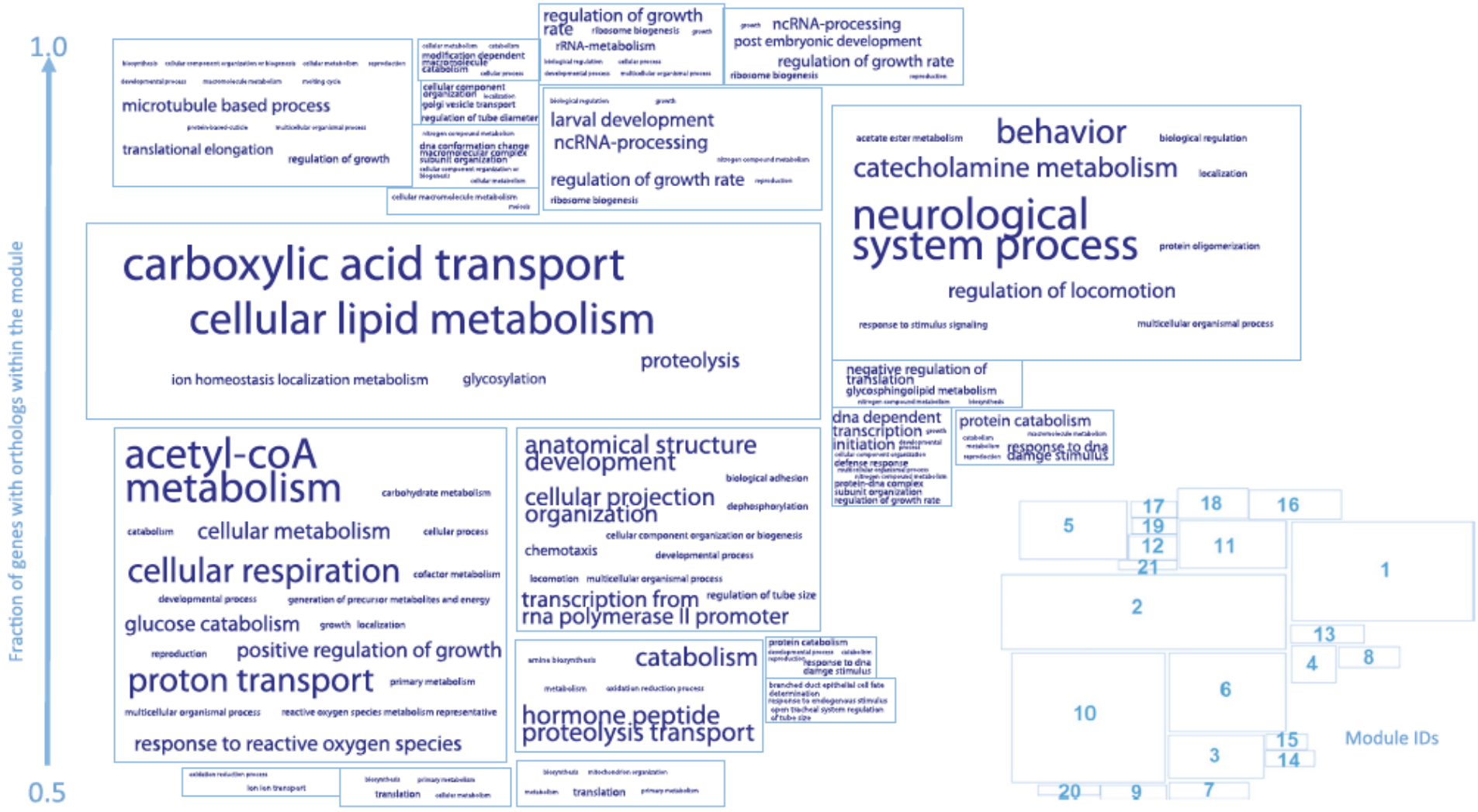
GO terms of specific modules

worm specific dauer entry

fly specific chitin activities

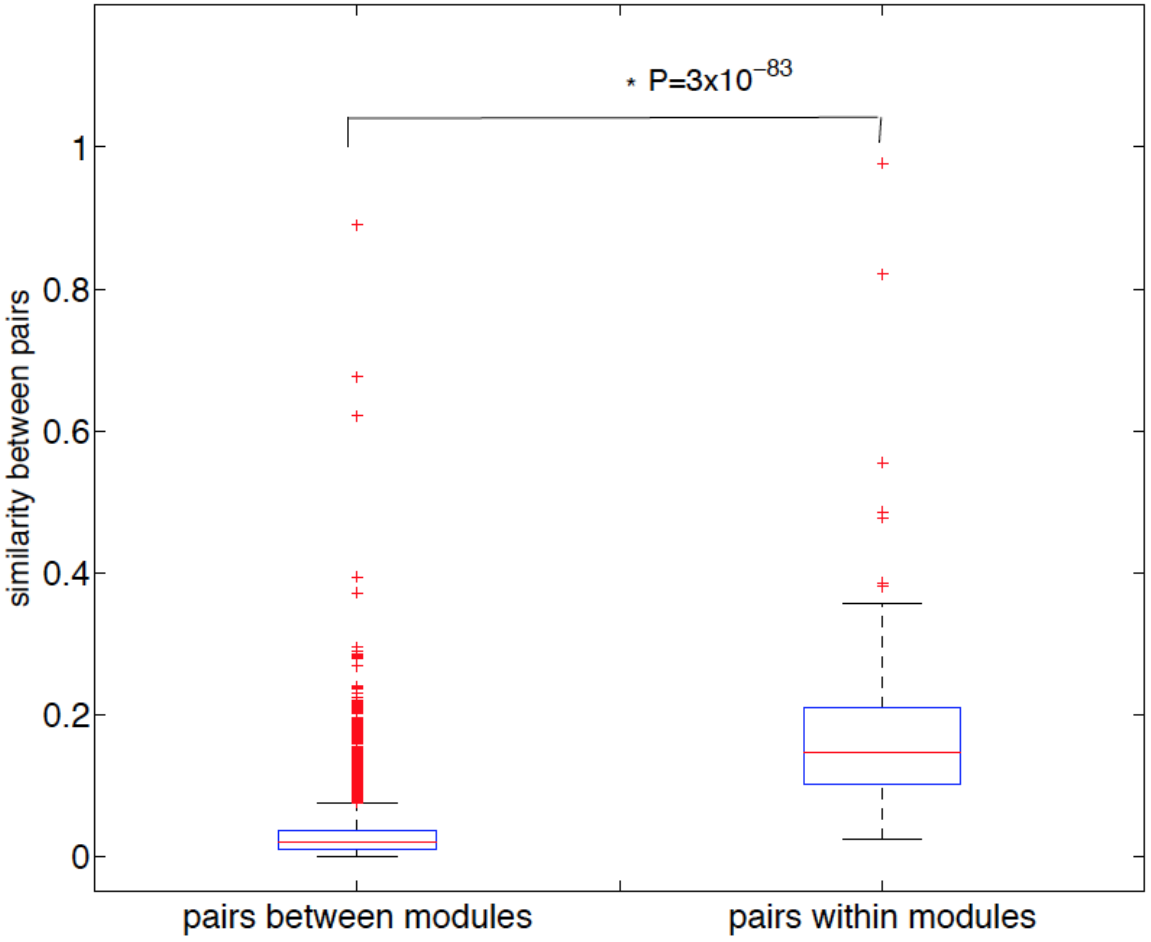
Yan KK et al. Genome Biology. 2014

More conserved modules



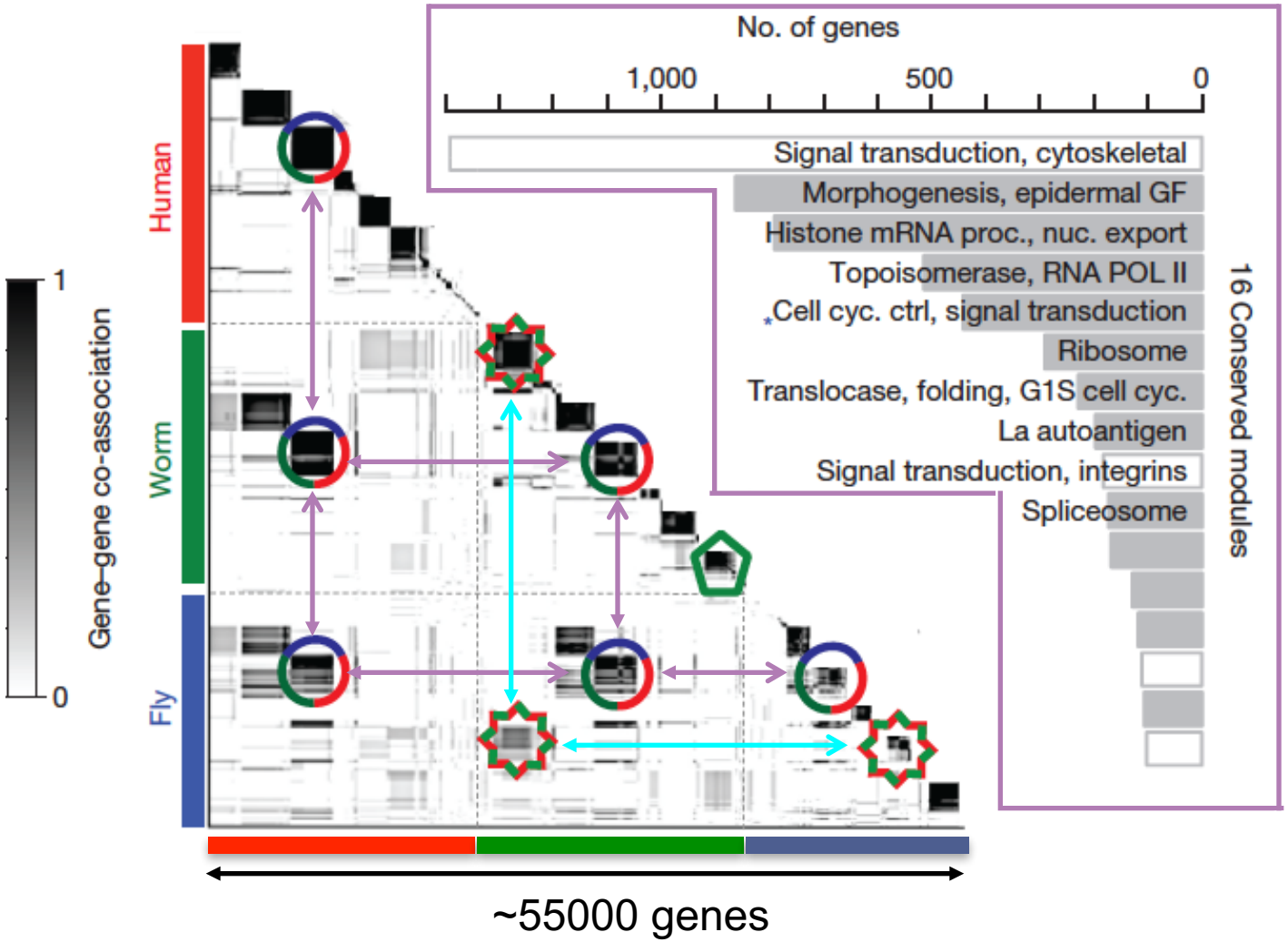
Yan KK et al. Genome Biology. 2014

Separation of modules in terms of GO



Yan KK et al. Genome Biology. 2014

Application for more than 2

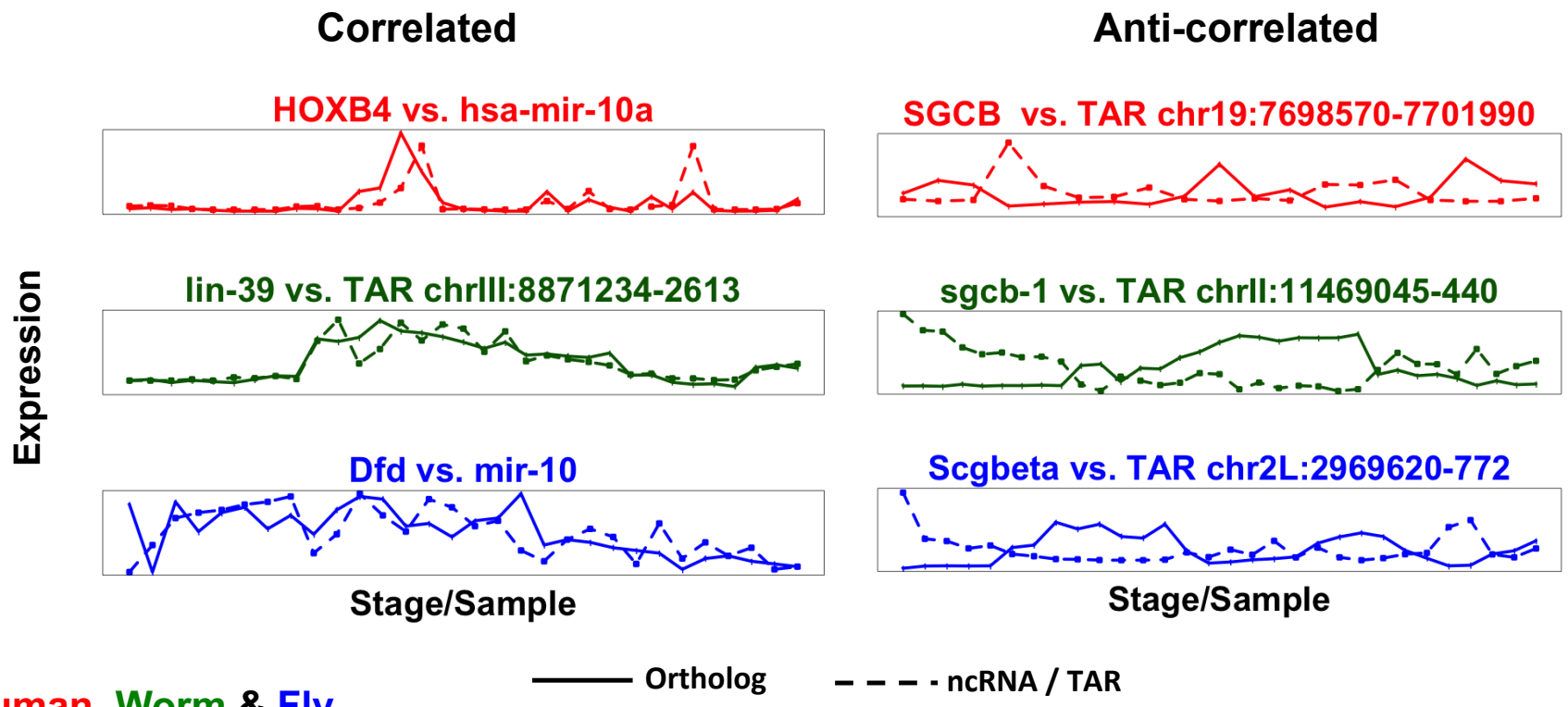


[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

ncRNAs associated with modules

Non-canonical transcription (TARs):

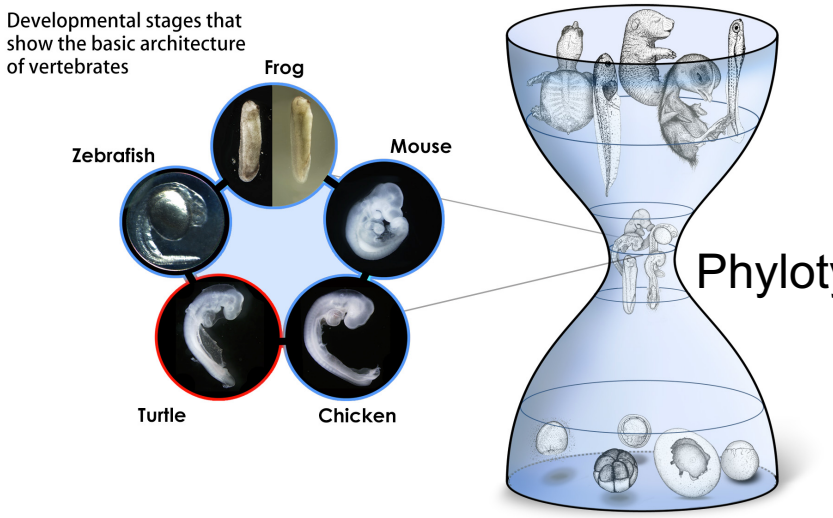
- Identify TARs that are significantly correlated and anti-correlated with genes in the 16 modules.



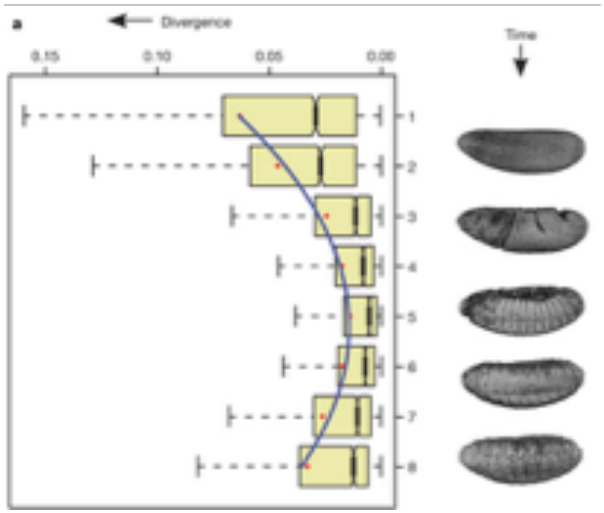
Human, Worm & Fly

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

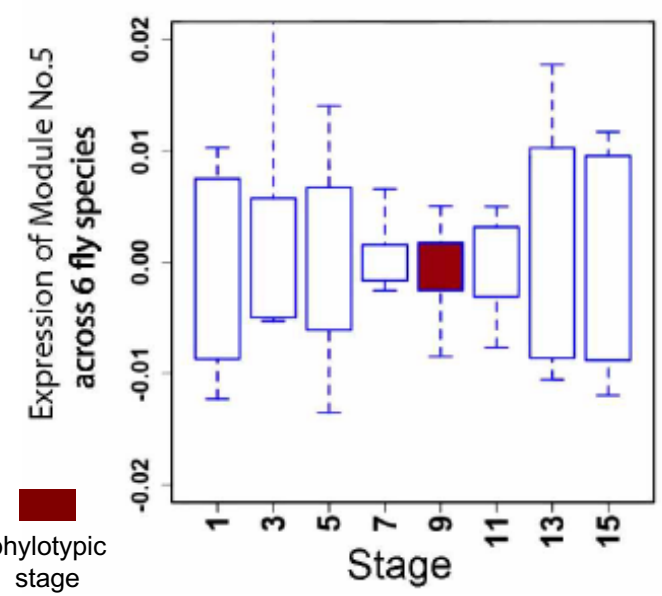
Conserved modules exhibit canonical hourglass behavior



Illustrations courtesy Naoki Irie



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



Canonical Inter-organism Behavior

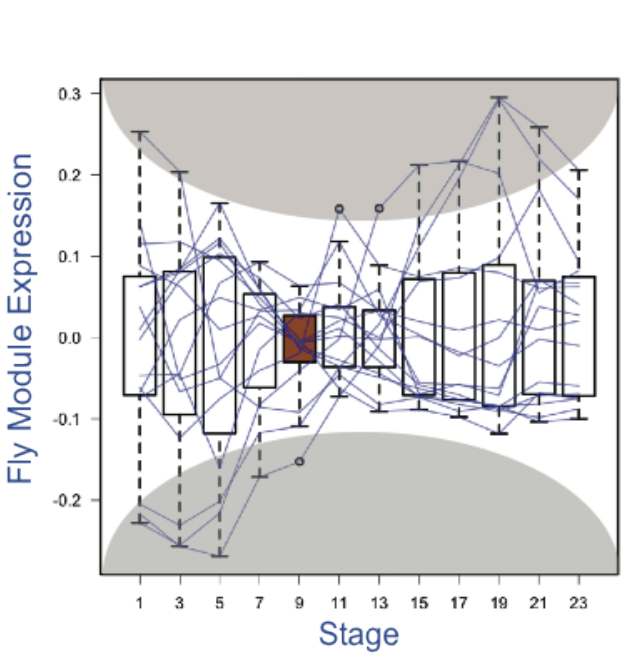
- "Hourglass hypothesis": all organisms go through a particular stage in embryonic development ("phylotypic" stage) where inter-organism expression differences of orthologous genes are smallest.
- **We identify modules (12 out of 16) which have this behavior at the phylotypic stage.**

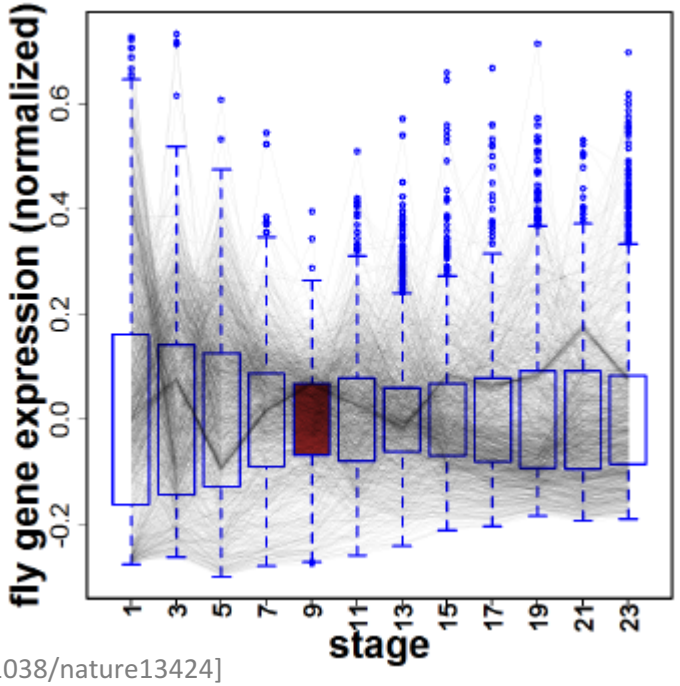
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

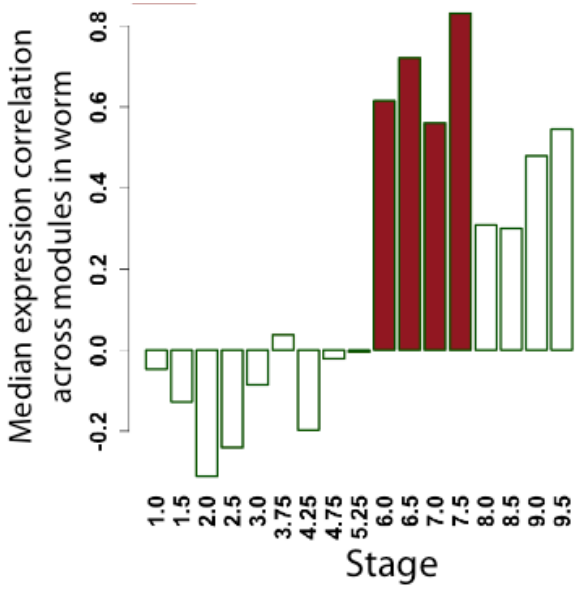
Hourglass Behavior

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).
- Strongly correlated correlation at phylotypic stage (worm).

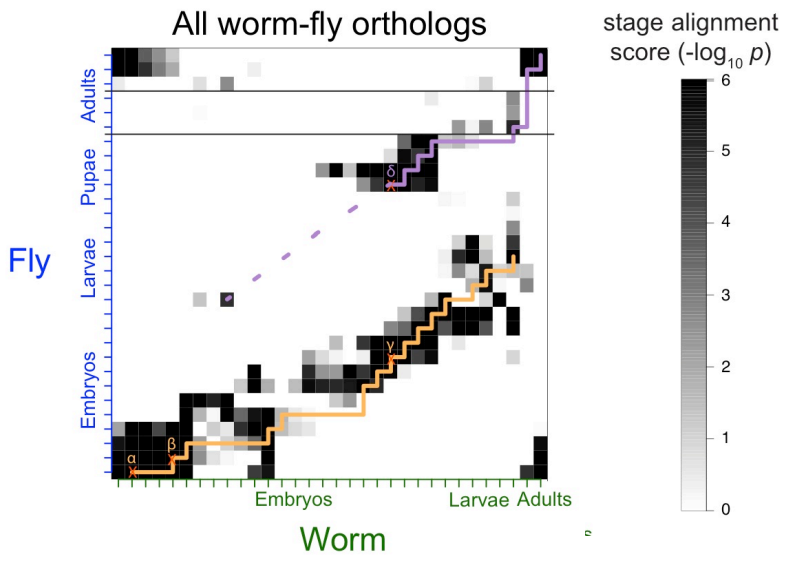






[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

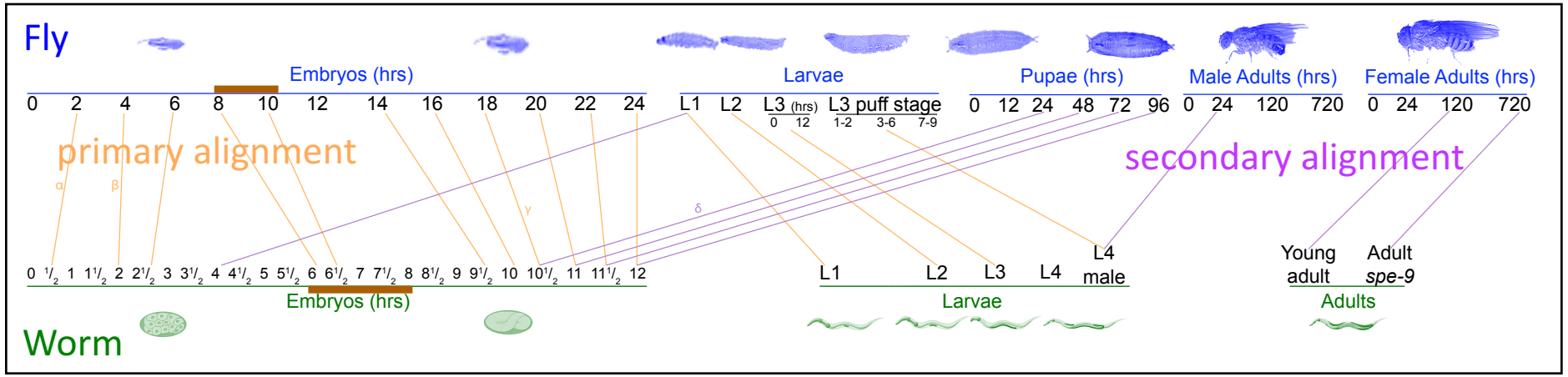
Alignment of Developmental Time-Course



For worm & fly find stage-specific genes

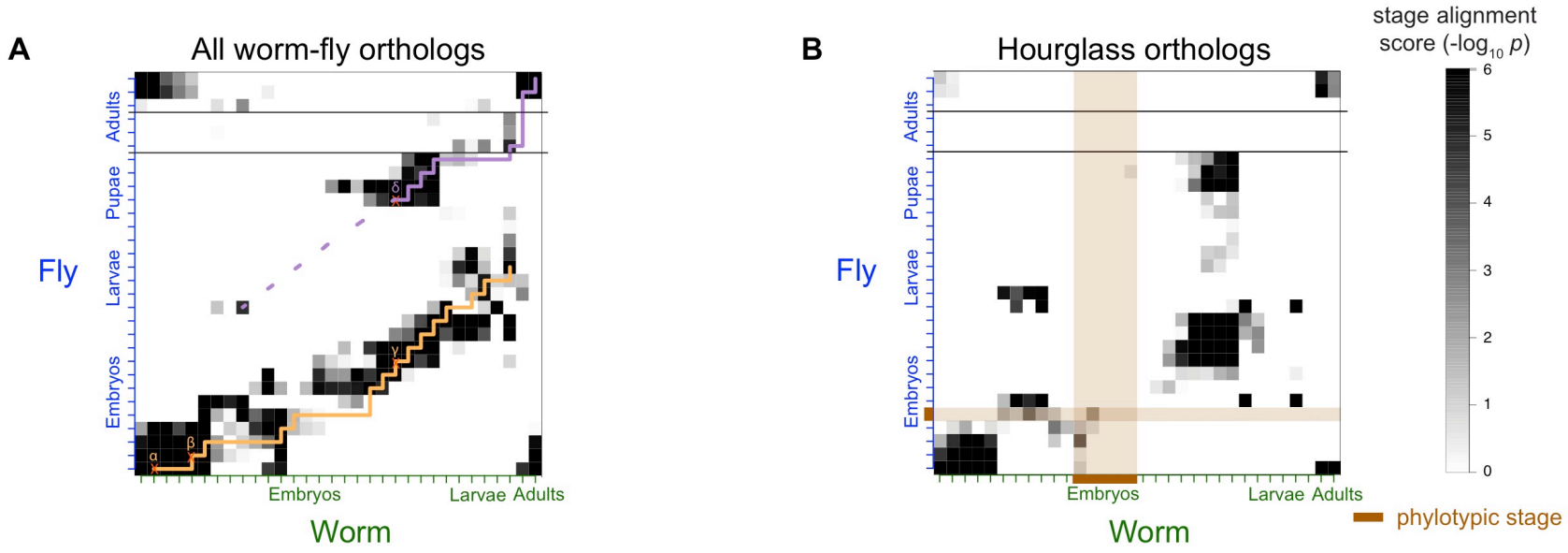
We can align developmental stages using fraction of shared orthologs between worm and fly amongst these

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

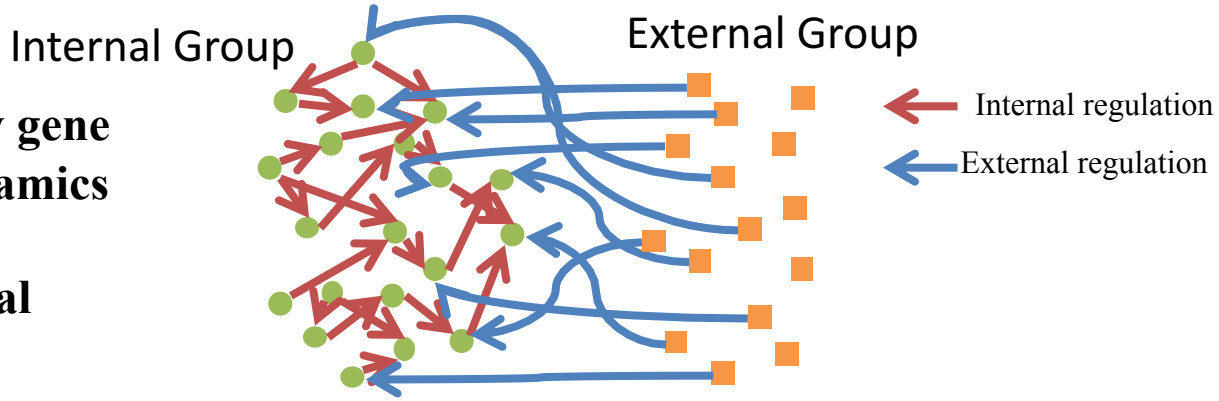
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

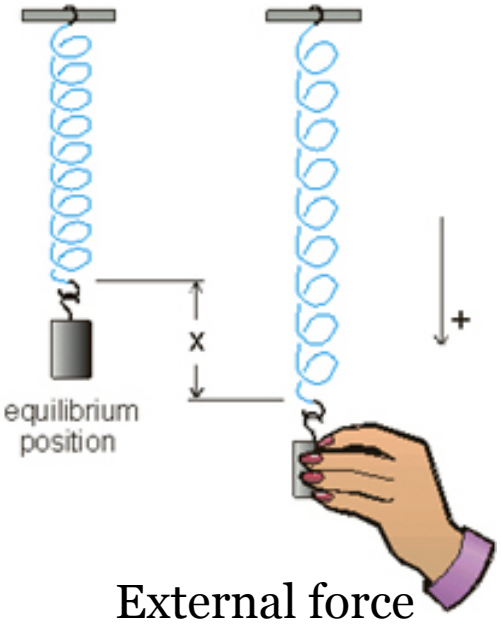
Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Internal & external gene regulatory networks

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



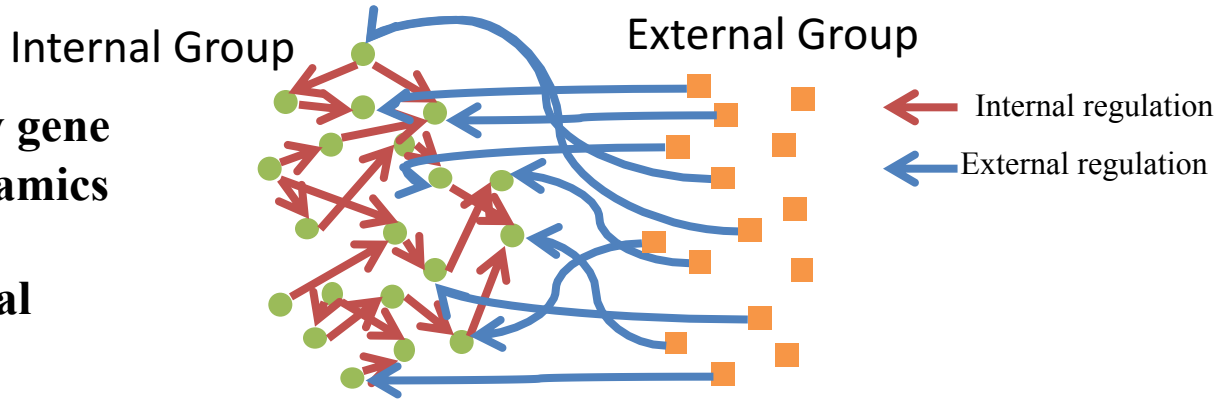


Interested system	Internal regulatory network	External regulatory network
Cross-species conserved genes	Conserved transcriptional factors (TFs)	Non-conserved TFs
Protein-coding genes	TFs	micro-RNAs
Individual's protein coding genes	Wild-type TFs	Somatic mutated TFs
Protein-coding genes in brain	Commonly expressed TFs	Brain-specific expressed TFs
Protein-coding genes in development	House-keeping TFs	Developmental TFs

[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*, '16]

Effective state space model for meta-genes

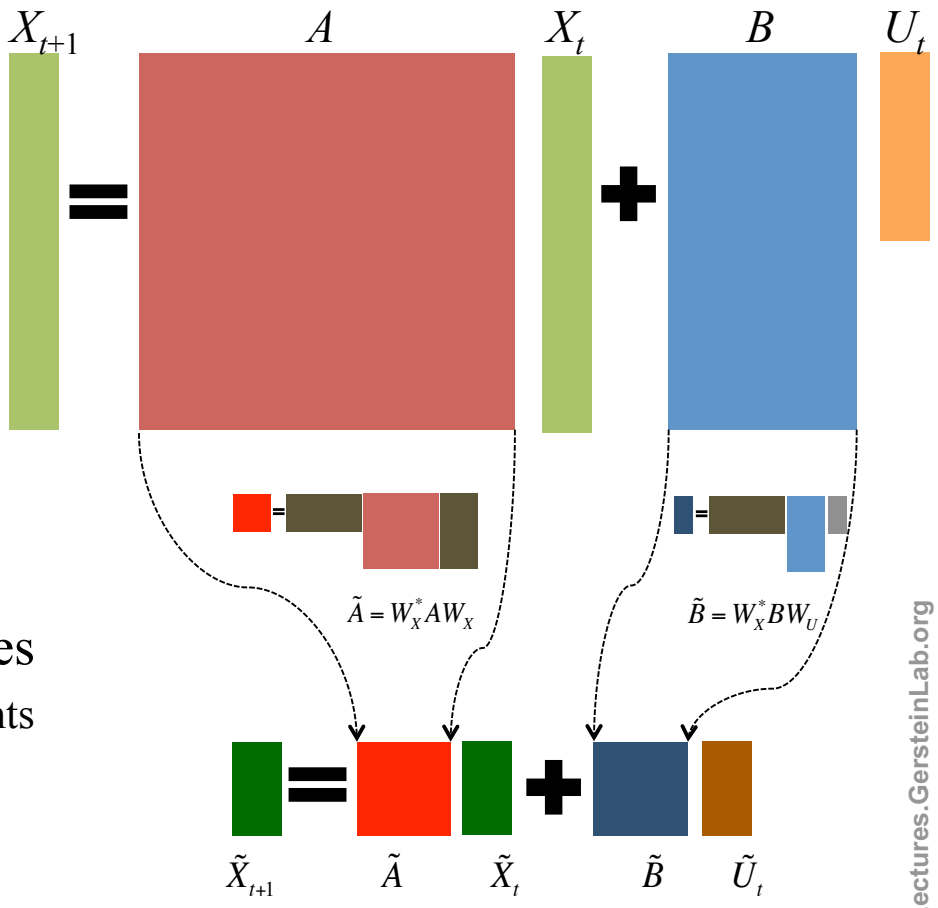
Not enough data to estimate state space model for genes
(e.g., 25 time points per gene to estimate 4 million elements of A or B for 2000 genes)

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

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 meta-genes)

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



[Wang et al. *PLOS CB*, '16]

Decomposition of internal and external-related dynamic components

$$
\begin{aligned}
X_t &= AX_{t-1} + BU_{t-1} \\
&= A(AX_{t-2} + BU_{t-2}) + BU_{t-1} \\
&= A^2X_{t-2} + ABU_{t-2} + BU_{t-1} \\
&= A^3X_{t-3} + A^2BU_{t-3} + ABU_{t-2} + BU_{t-1} \\
&= \cdots \\
&= A^{t-1}X_1 + A^{t-2}BU_1 + A^{t-3}BU_2 + \cdots + ABU_{t-2} + BU_{t-1} \\
&= \underbrace{A^{t-1}X_1}_{x_t^{INT}} + \underbrace{\sum_{k=1}^{t-2} A^k BU_{t-1-k}}_{x_t^{INTER}} + \underbrace{BU_{t-1}}_{x_t^{EXT}}
\end{aligned}
$$

X_t^{EXT}: externally driven dynamic component

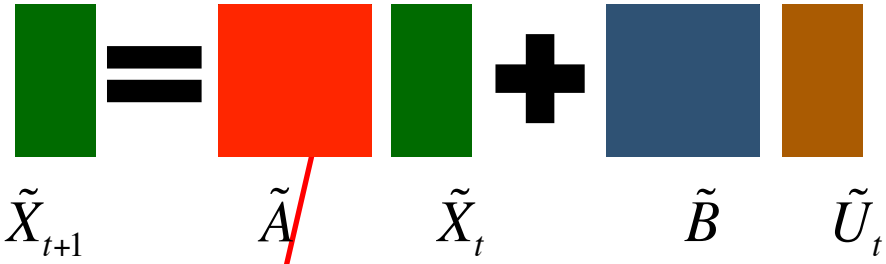
X_t^{INT}: Internally driven dynamic component

X_t^{INTER}: dynamic components driven by interactions between internal and external terms

* Subdivision of the rest of the terms $\sum_{k=1}^{t-2} A^k BU_{t-1-k} + BU_{t-1}$ is completely arbitrary

[Wang et al. *PLOS CB*, '16]

Canonical temporal expression trajectories from effective state space model

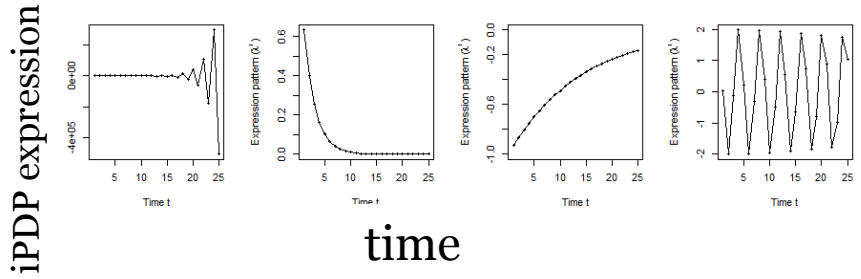


$\tilde{X}_{t+1}$ $\tilde{A}$ $\tilde{X}_t$ $\tilde{B}$ $\tilde{U}_t$

Internal driven dynamics

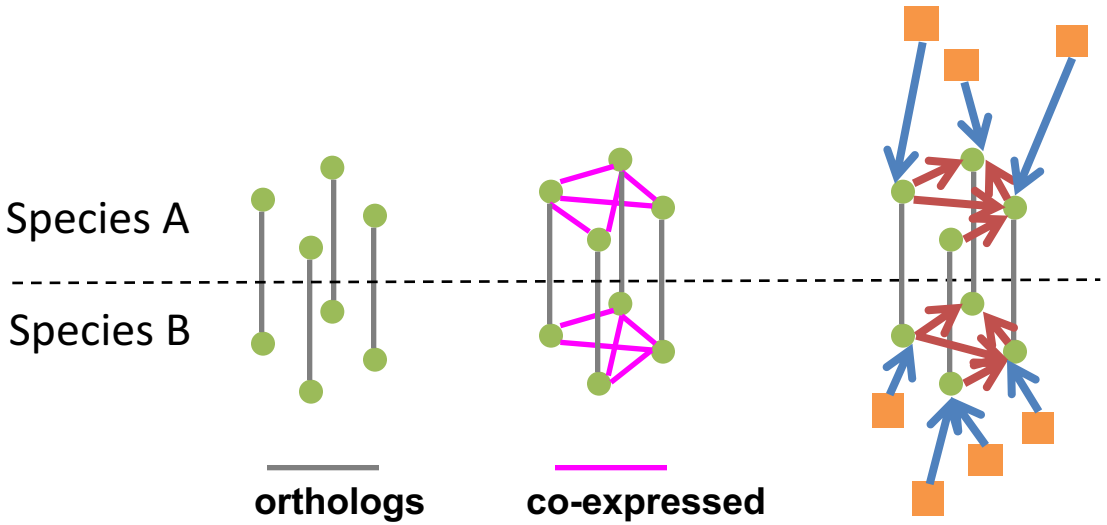
p^{th} internal principal dynamic pattern (iPDP): $[\lambda_p{}^1, \lambda_p{}^2, \ldots, \lambda_p{}^T]$, where λ_p is p^{th} eigenvalue of $\tilde{A}$.

Canonical temporal expression trajectories (e.g., degradation, growth, damped oscillation, etc.)



[Wang et al. *PLOS CB*, '16]

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

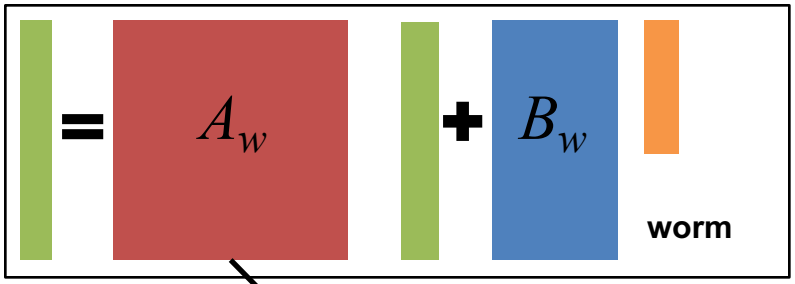
To what degree can't ortholog expression levels be predicted due to species-specific regulation

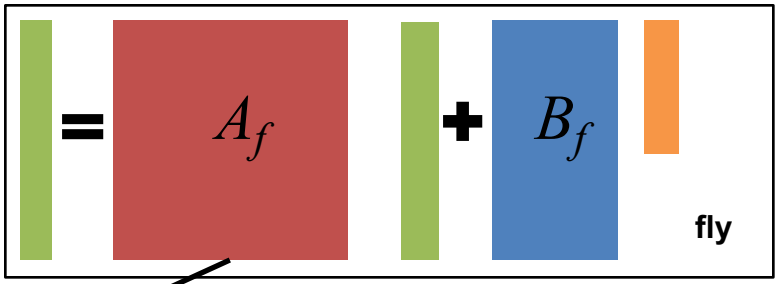
[Wang et al. *PLOS CB*, '16]

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

Not enough time samples!

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



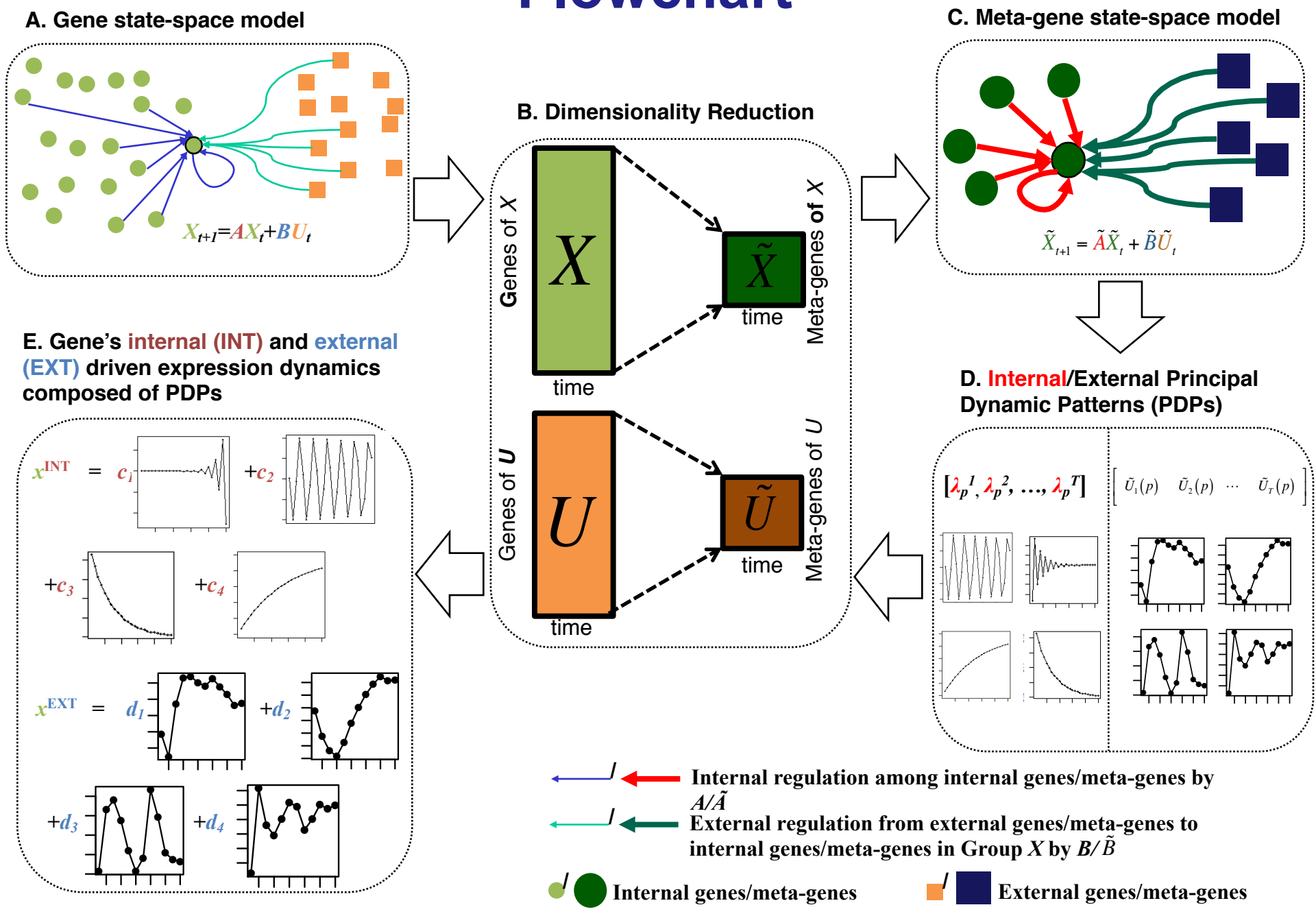


If A_w and A_f have similarities, cross-species conserved regulatory networks in embryonic development

Embryonic stem cells (ESCs)

[Wang et al. *PLOS CB*, '16]

Flowchart



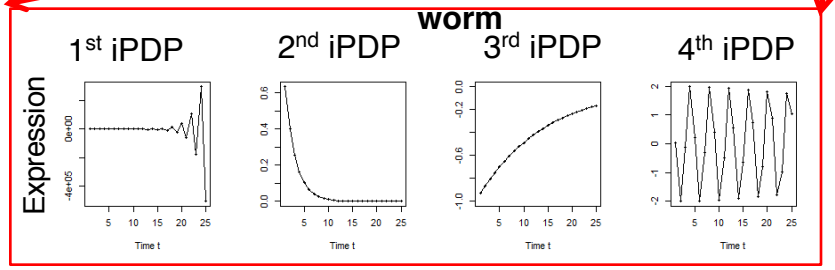
Orthologs have similar internal but different external dynamic patterns during embryonic development

Worm's effective state space model

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

A

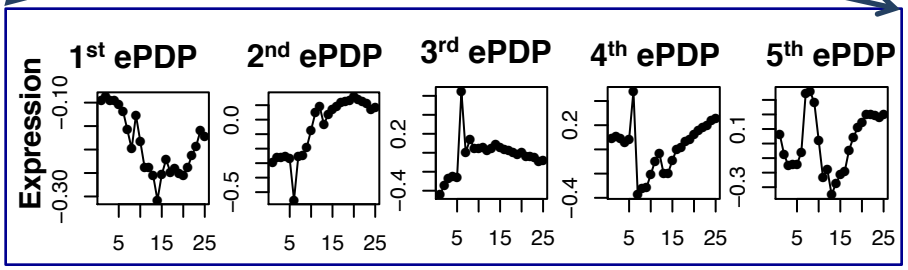
iPDPs: time exponentials of $\tilde{A}$ eigenvalues in worm



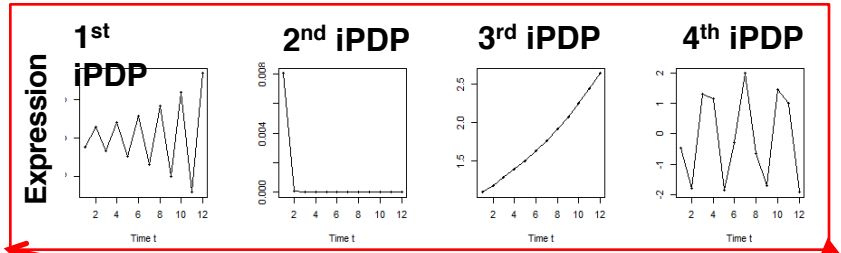
Similar iPDPs

B

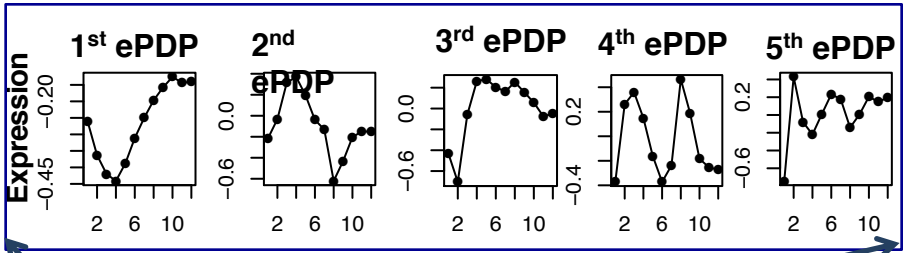
ePDPs: $\tilde{U}$ in worm



Different ePDPs



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



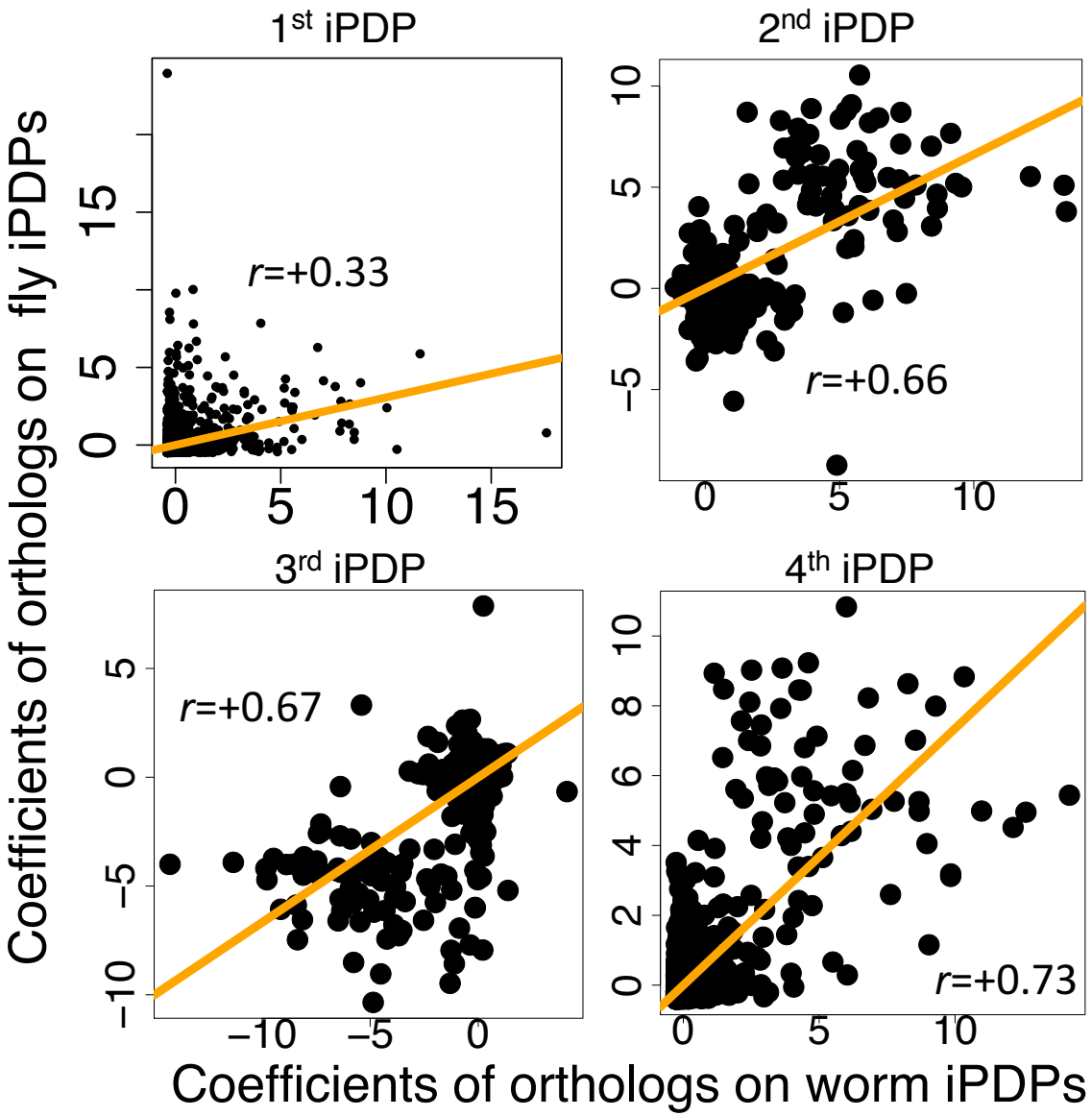
ePDPs: $\tilde{U}$ in fly

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

Fly's effective state space model

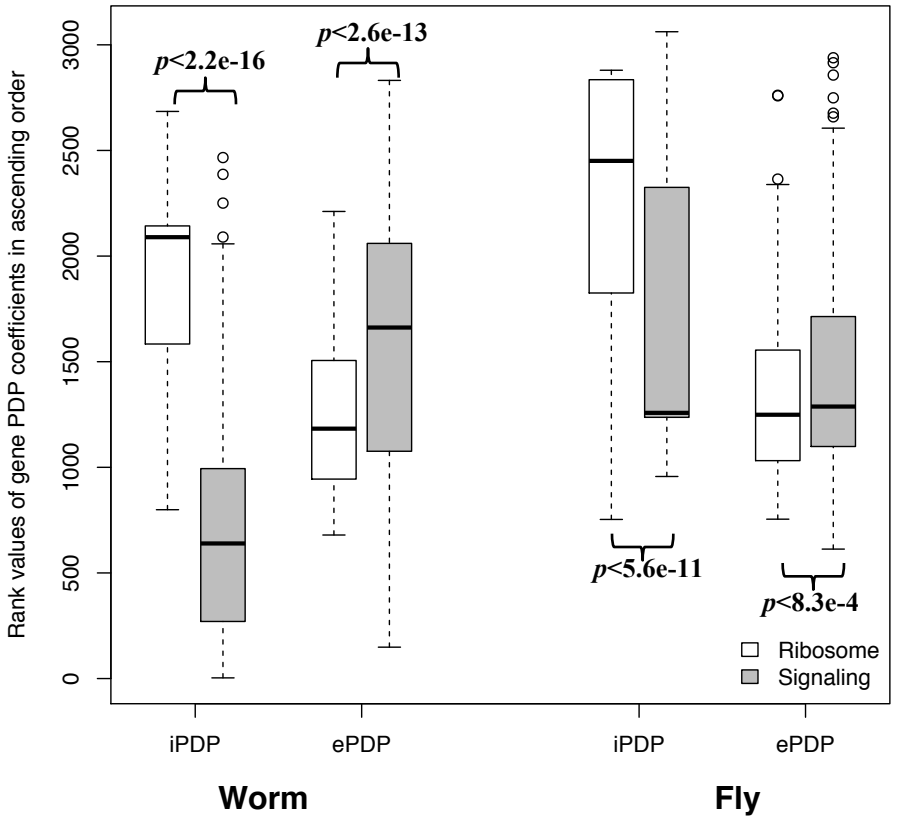
[Wang et al. *PLOS CB*, '16]

Orthologs have correlated iPDP coefficients



[Wang et al. *PLOS CB*, '16]

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

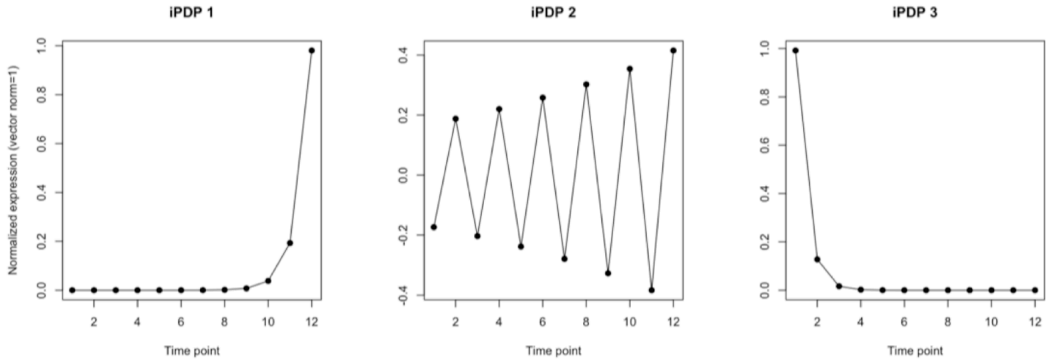


Ribosomal genes have significantly larger coefficients for the internal than external PDPs, but signaling genes exhibit the opposite trend

[Wang et al. *PLOS CB*, '16]

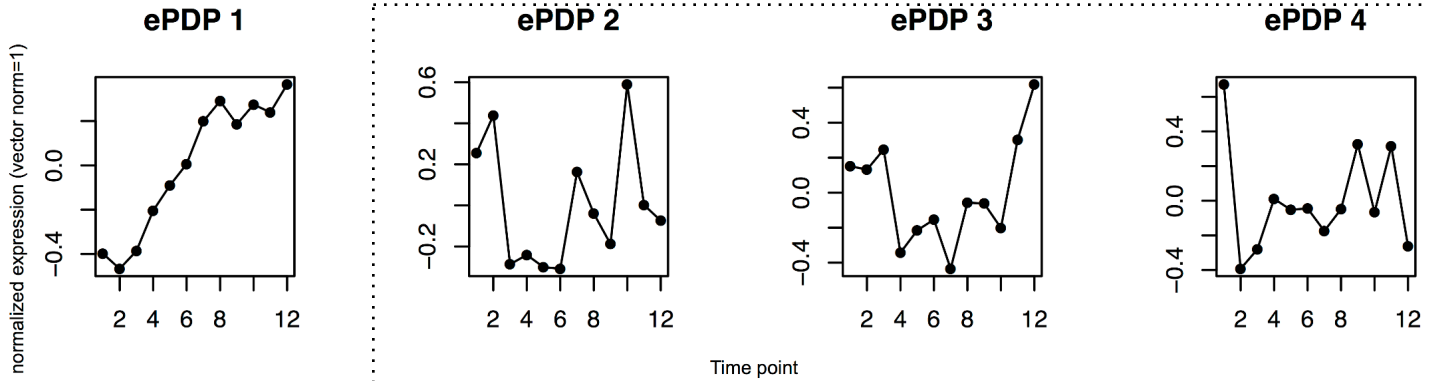
Breast cancer cell cycle under hormonal stimulation

Dataset	Group *X* (internal)	Group *U* (external)	Time samples of a full cell cycle
Human breast cancer cell cycle under hormonal stimulation	1132 metazoan conserved genes incl. 150 orthologous TFs	1870 non-conserved metazoan transcription factors	T=12 time points: 0, 4, 6, 8, 12, …, 28, 32 hours



Oscillated iPDP by conserved TFs

a full cell cycle



Oscillated ePDP by non-conserved TFs

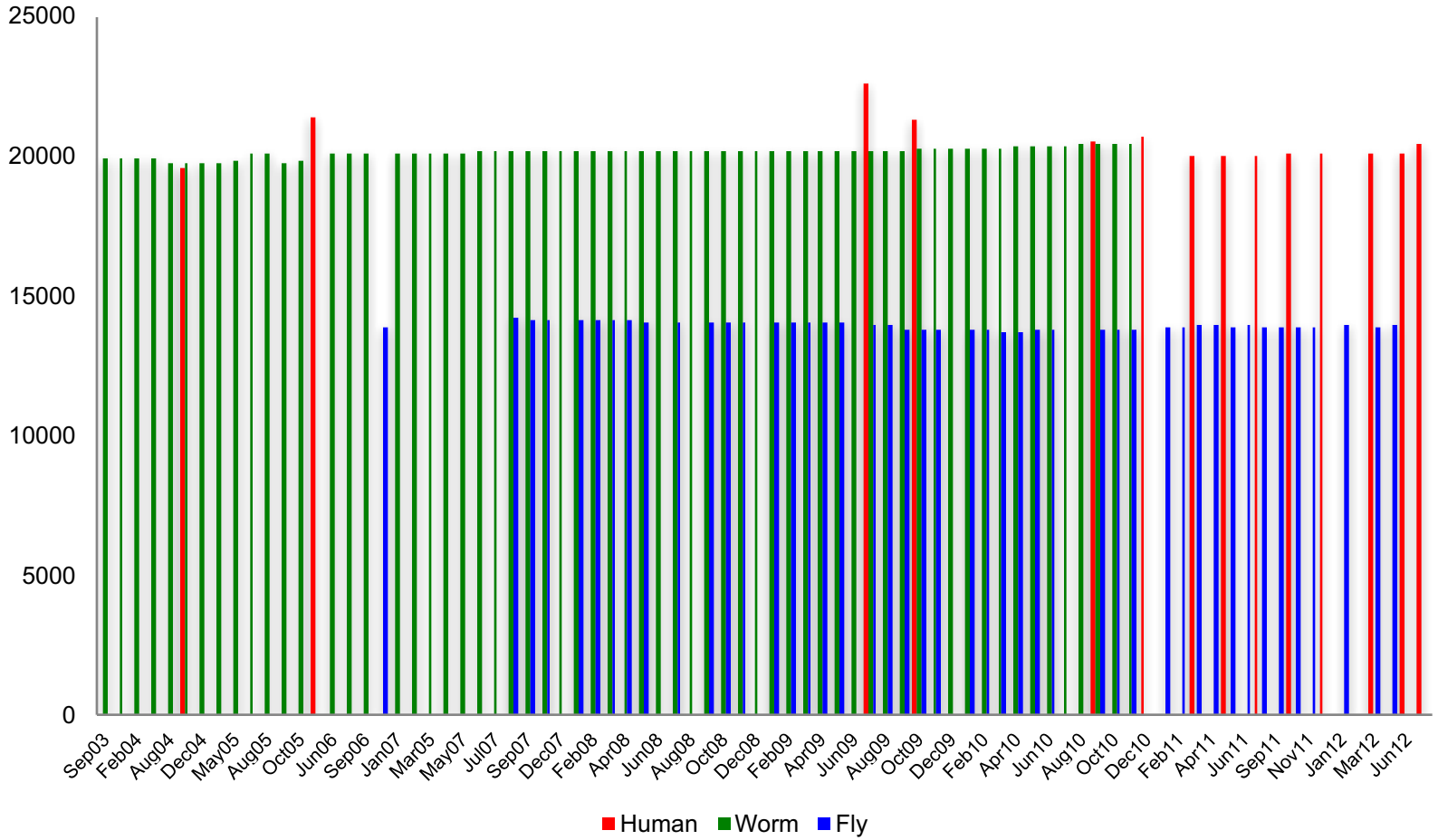
faster cycle due to hormone

[Wang et al. *PLOS CB*, '16]

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

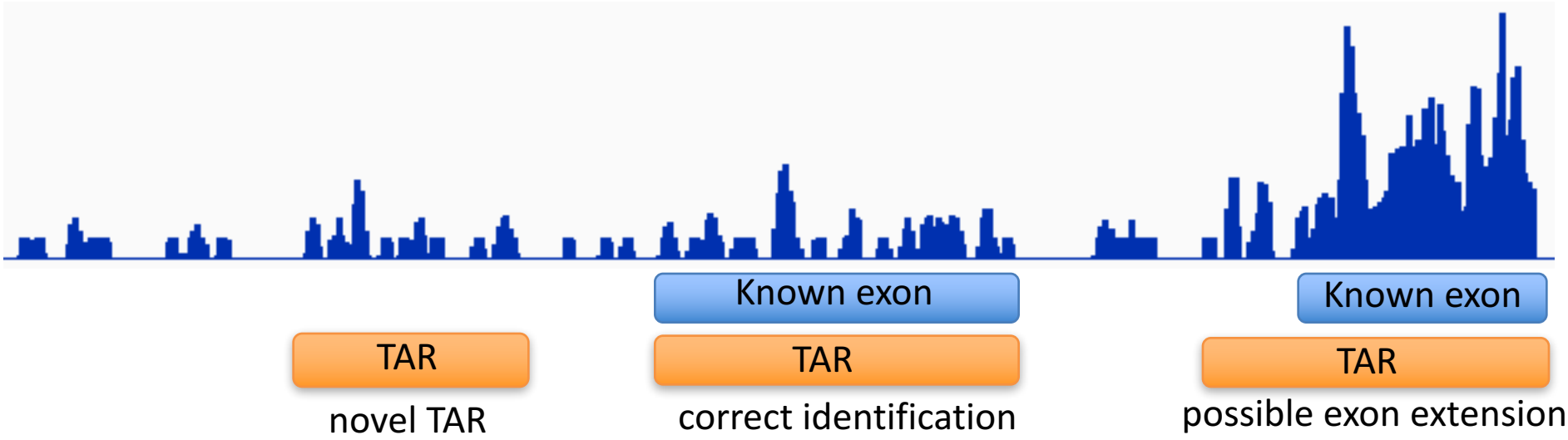
Protein-coding gene counts in worm, fly & human have stabilized & have remained fairly constant



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

Discovering Transcriptionally Active Regions (novel RNA contigs)

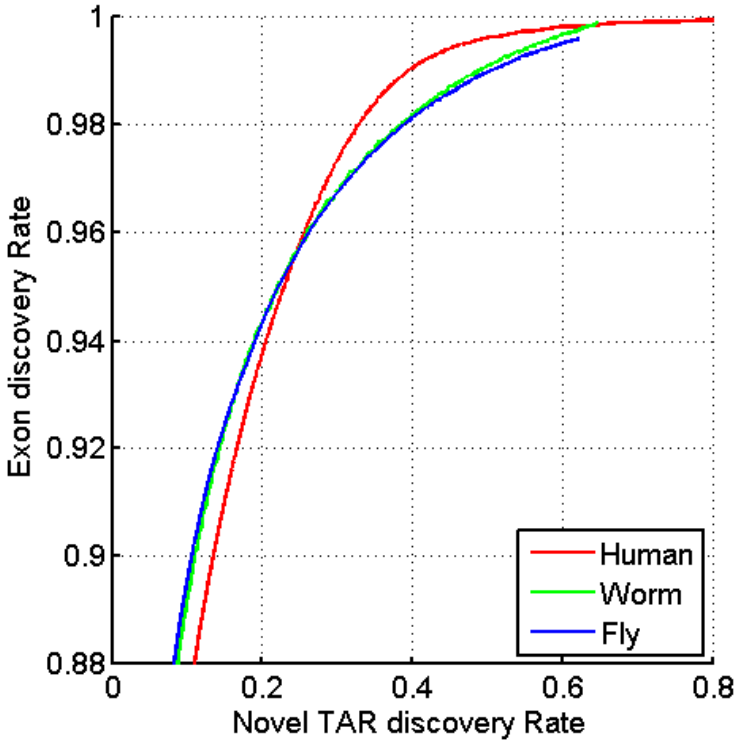
- Cluster reads setting minimum-run and maximum gap parameters for newly identified transcribed regions (TARs)
- Assess exon discovery rates for known genes and noncoding RNAs



ENCODE RNA-Seq Data

Uniform Annotation of non-coding Elements

- Uniformly processed the RNA-seq expression compendium and for identification of pervasively transcribed regions



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

Annotated ncRNAs

			Human			Worm			Fly		
			Elements	Genome Coverage		Elements	Genome Coverage		Elements	Genome Coverage	
				Kb	%		Kb	%		Kb	%
mRNAs (exons)			20,007	86,560	3.0	21,192	34,437	34.3	13,940	35,970	28.0
Pseudogenes			11,216	27,089	0.95	881	1,343	1.3	145	155	0.12
Annotated ncRNAs	Comparable ncRNAs	pri-miRNA	58	1,158	0.04	44	16	0.02	43	300	0.23
		pre-miRNAs	1,756	162	0.006	221	20	0.02	236	22	0.02
		tRNAs	624	47	0.002	609	45	0.04	314	22	0.02
		snoRNAs	1,521	168	0.006	141	16	0.02	287	34	0.03
		snRNAs	1,944	210	0.007	114	14	0.01	47	7	0.006
		lncRNAs	10,840	10,581	0.37	233	184	0.18	852	868	0.68
	Other ncRNAs		5,411	3,268	0.11	40,104	2,329	2.3	376	2,103	1.6
		nc-piRNA loci	88	1,272	0.04	35,329	449	0.45	27	1,473	1.1
Total			22,154	17,770	0.62	41,466	2,611	2.6	2,155	3,279	2.6

Identify non-canonical transcription in regions of the genome excluding mRNA exons, pseudogenes or annotated ncRNAs.

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

& Non-Canonical Transcription

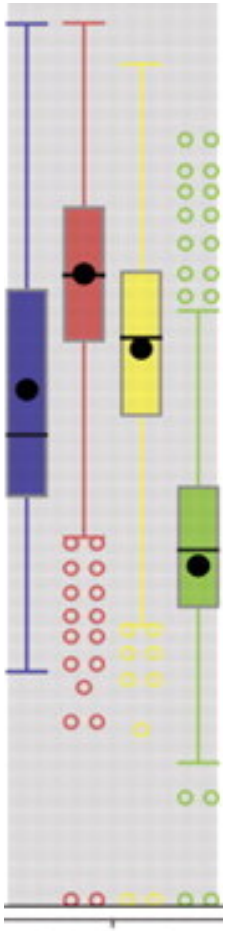
	Human			Worm			Fly		
	Elements	Genome Coverage		Elements	Genome Coverage		Elements	Genome Coverage	
		Kb	%		Kb	%		Kb	%
Total ncRNAs	22,154	17,770	0.62	41,466	2,611	2.6	2,155	3,279	2.6
Regions Excluding mRNAs, Pseudogenes or Annotated ncRNAs	283,816	2,731,811	95.5	143,372	63,520	63.3	60,108	89,445	69.6
Transcription Detected (TARs)	708,253	916,401	32.0	232,150	37,029	36.9	83,618	44,256	34.5
Supervised Predictions	104,016	13,835	0.48	2,525	392	0.39	599	164	0.13

- Similar fraction of non-canonical transcription of non-canonical transcription in human, worm and fly
  - 32-37% of each genome

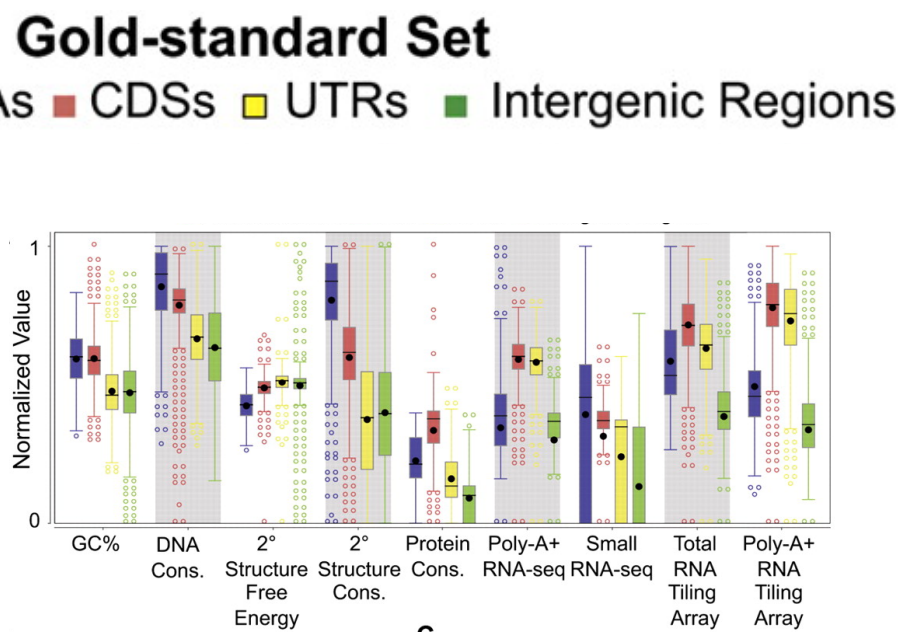
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

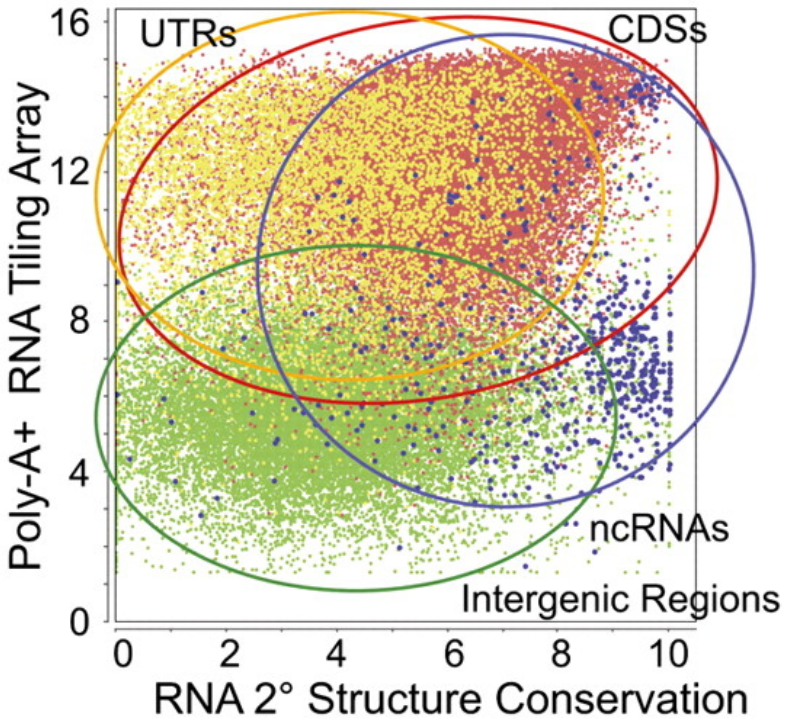
lncRNA: Machine-learning Identification of many candidate ncRNAs through evidence integration

- No single feature (e.g. expr. expts., conservation, or sec. struc.) finds all known ncRNAs => combine features in stat. model
- 90% PPV, 13 of 15 tested validate



Total RNA Tiling Array



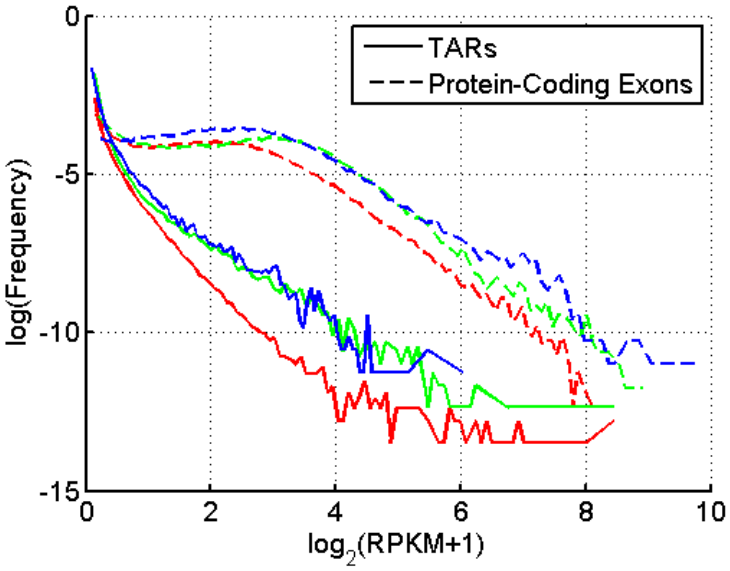


[Lu et al. Genome Res. 2011;21:276-285]

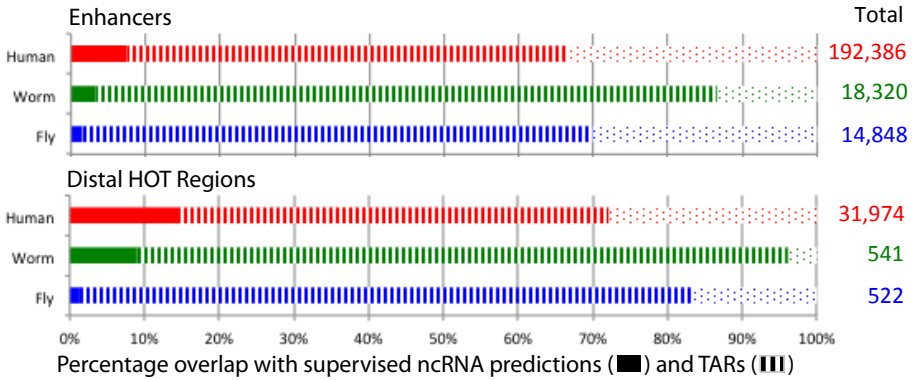
TAR Characterization

Non-canonical transcription (TARs):

- Mostly transcribed at lower levels than protein-coding genes.



- Enrichment for overlap of TARs with ENCODE enhancers and distal HOT regions -> potential enhancer RNAs (eRNAs).



Human, Worm & Fly

[ENCODE-modencode Transcriptome paper, Nature (in press), doi: 10.1038/nature13424]

HOT Regions = High TF Co-occupancy

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

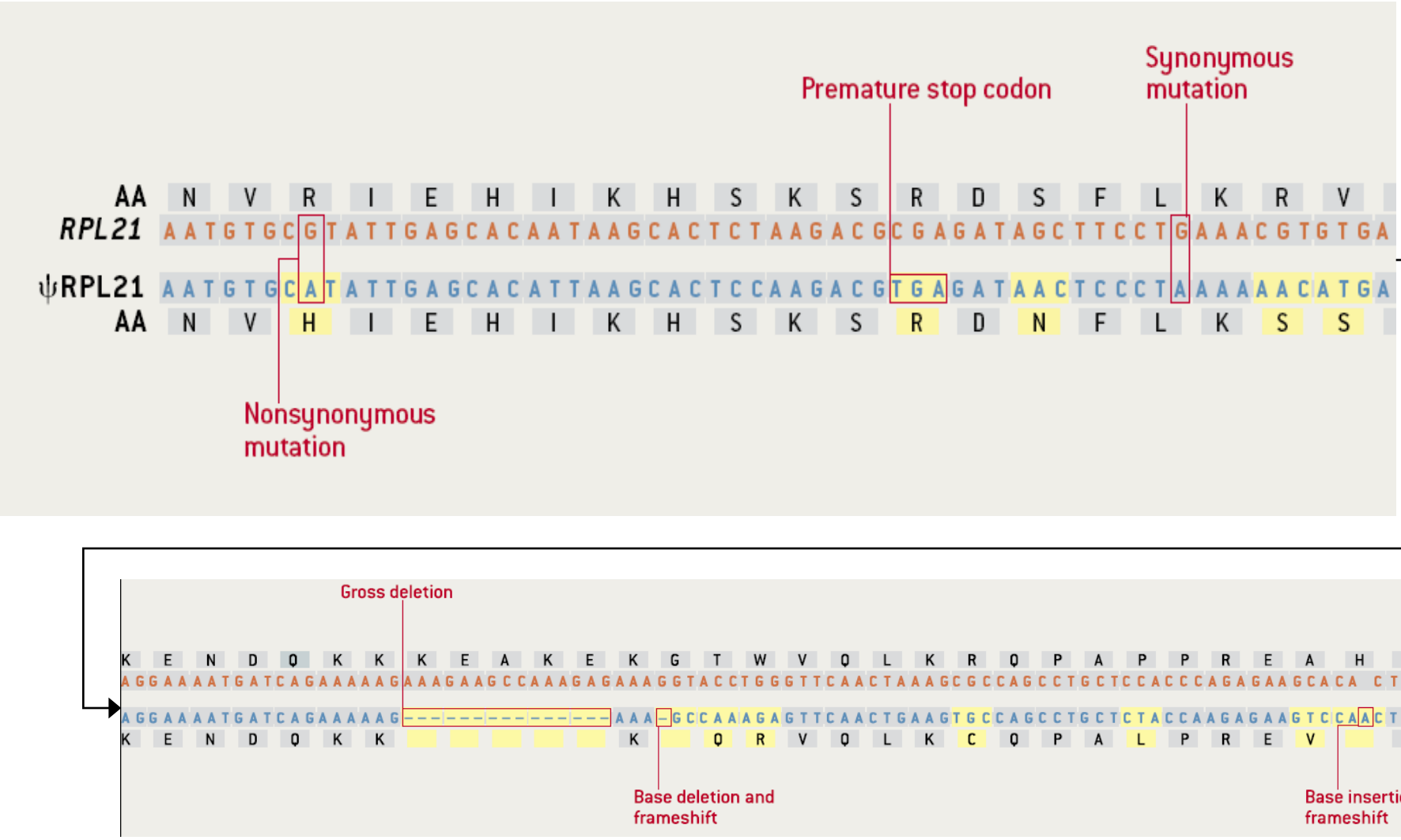
- Intro to **Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (ΨG)
  - Inheritable
  - Homologous to a functioning element – ergo a repeat!
  - Non-functional
    - No selection pressure so free to accumulate mutations
      - Frameshifts & stops
      - Small Indels
      - Inserted repeats (LINE/Alu)
    - **What does this mean?** no transcription, no translation?…

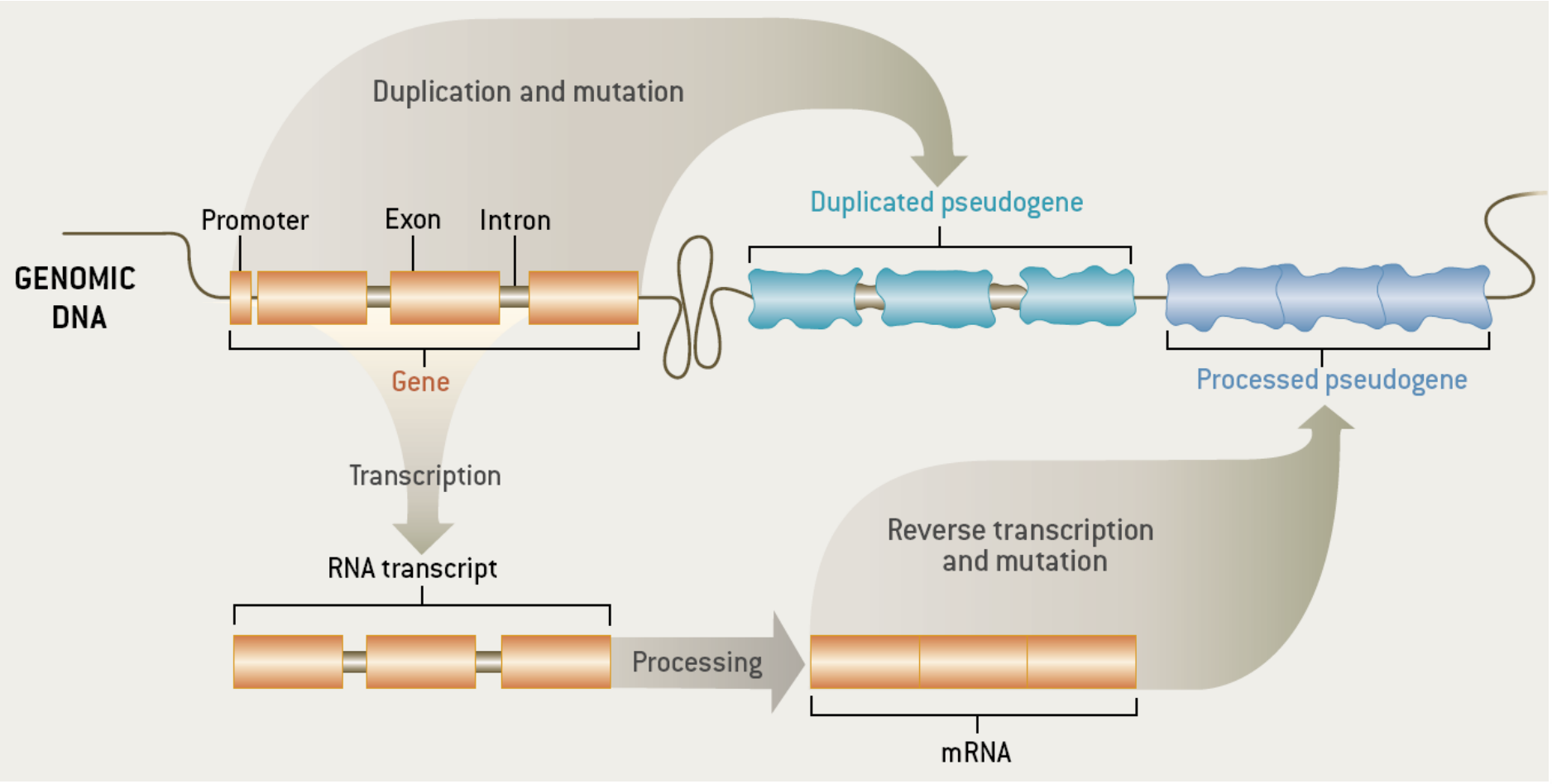
[Mighell et al. *FEBS Letts*, 2000]

Identifiable Features of a Pseudogene (ψRPL21)



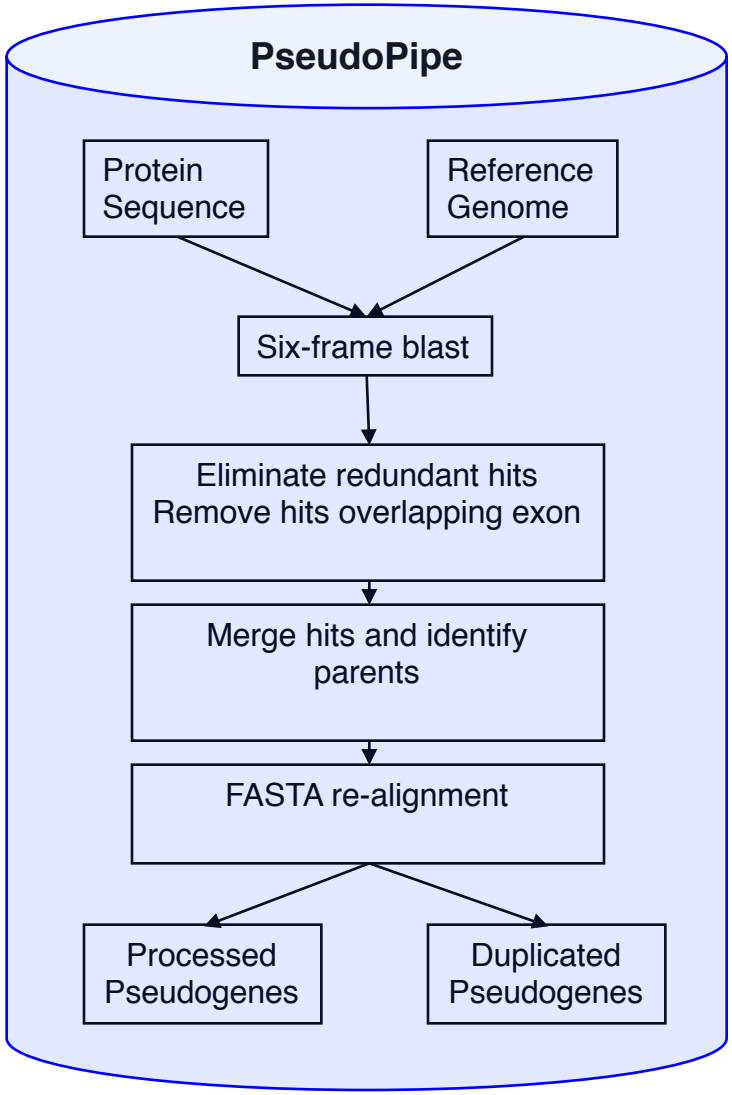
[Gerstein & Zheng. Sci Am 295: 48 (2006).]

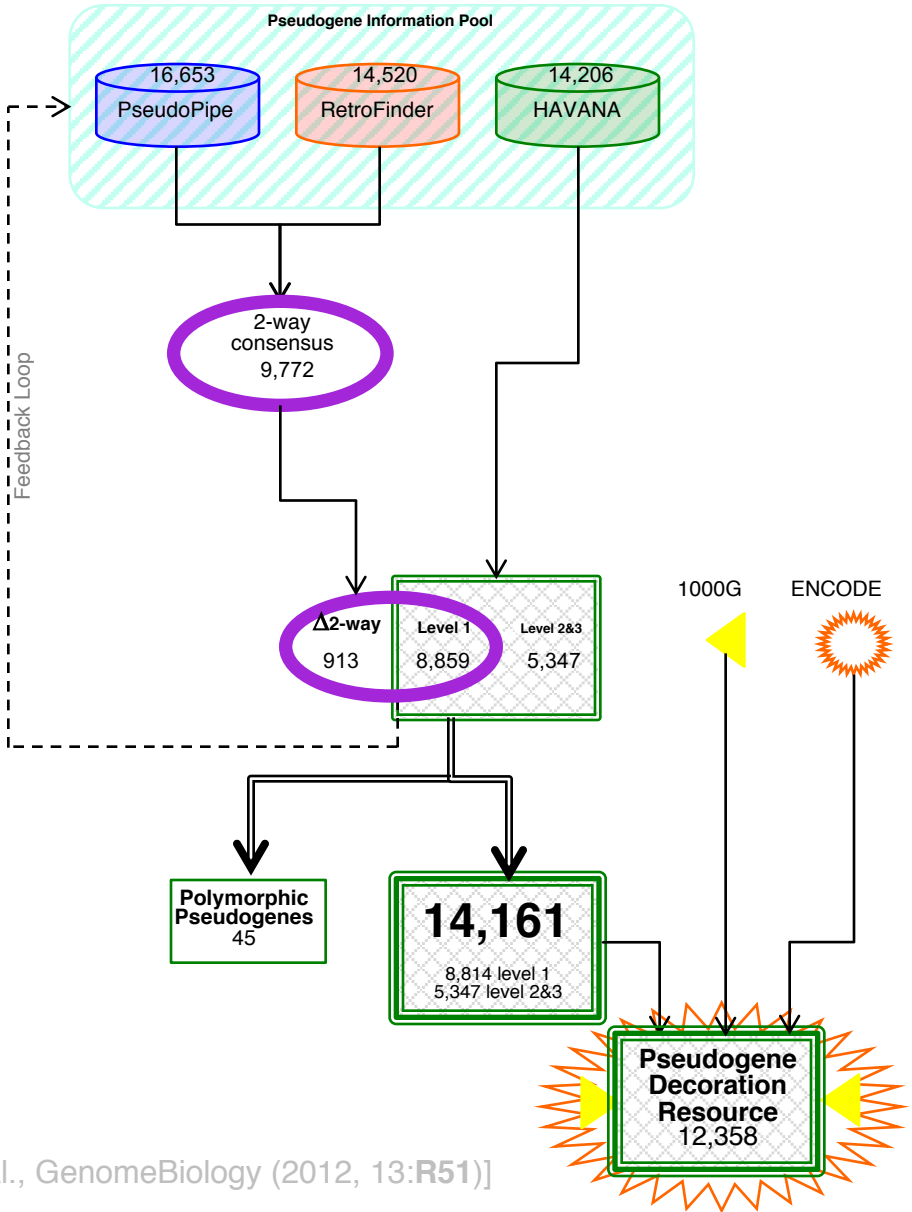
Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



[Gerstein & Zheng. Sci Am 295: 48 (2006).]

Genome-wide Annotation of Pseudogenes



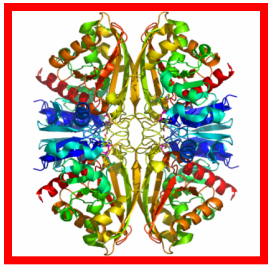


[Pei et al., GenomeBiology (2012, 13:**R51**)]

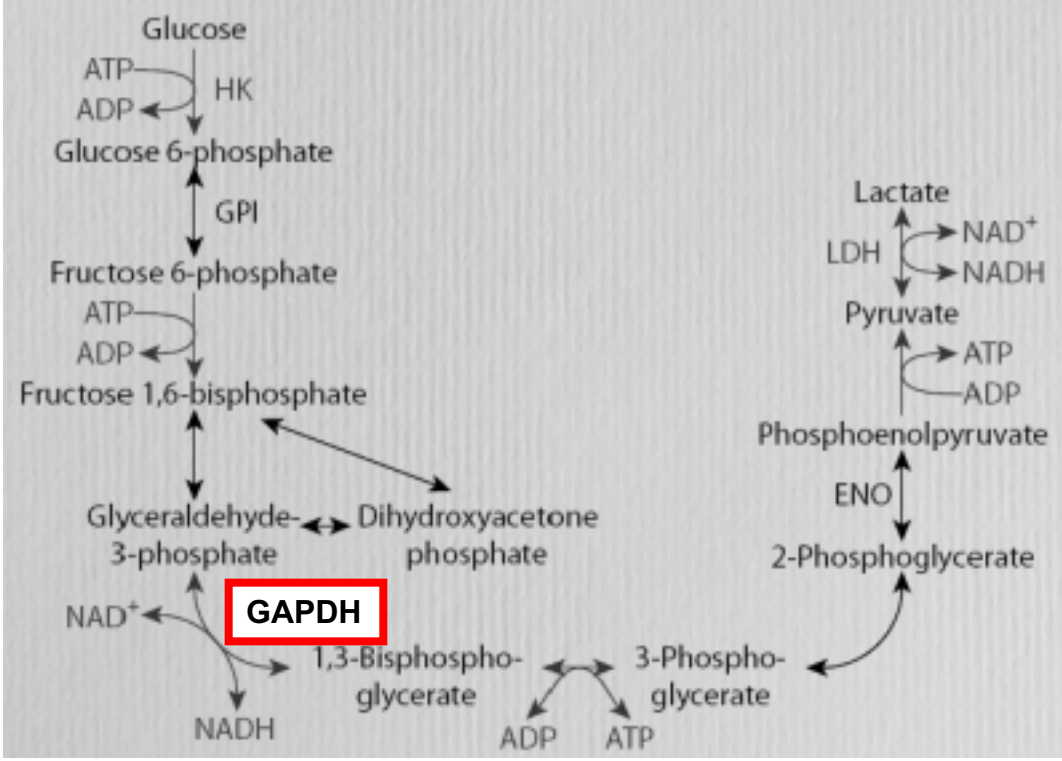
EX: Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



Processed/Duplicated

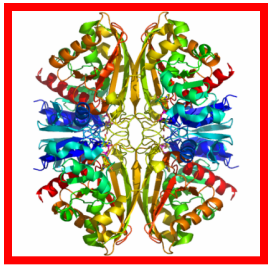


	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

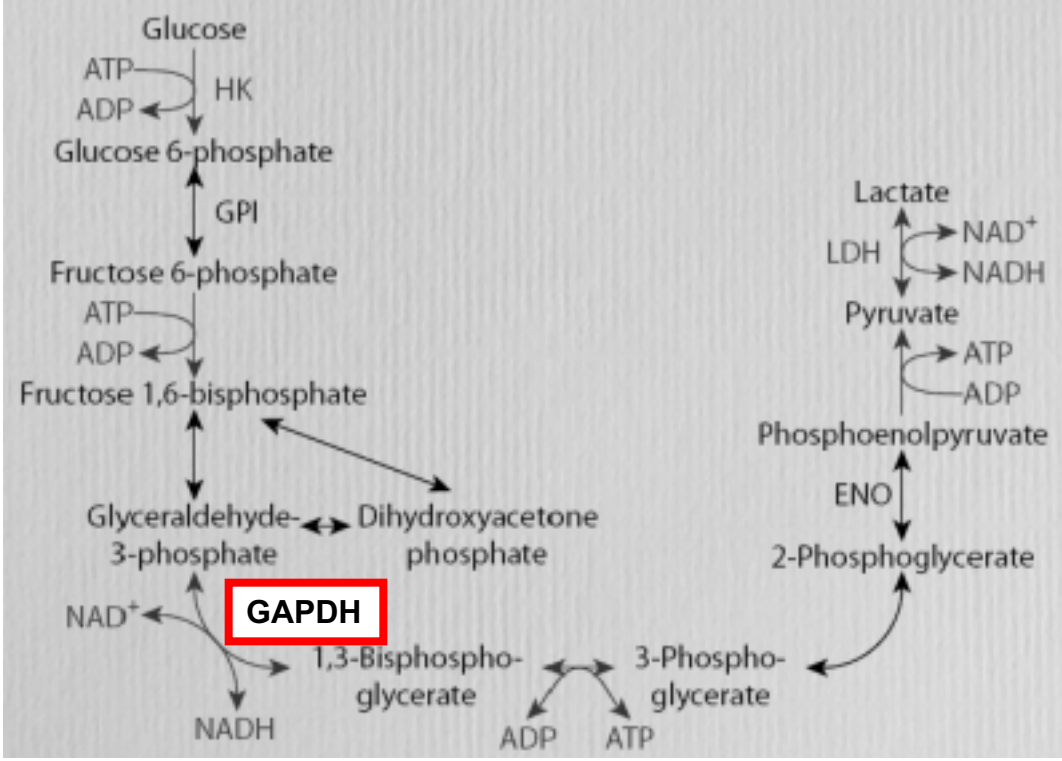
EX: Number of pseudogenes for each glycolytic enzyme

[Liu et al. BMC Genomics ('09)]

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



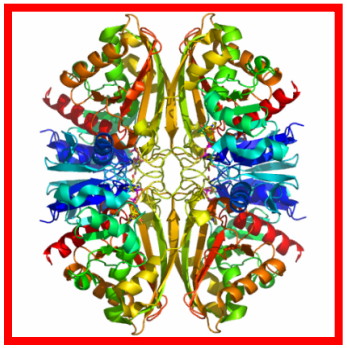
Processed/Duplicated



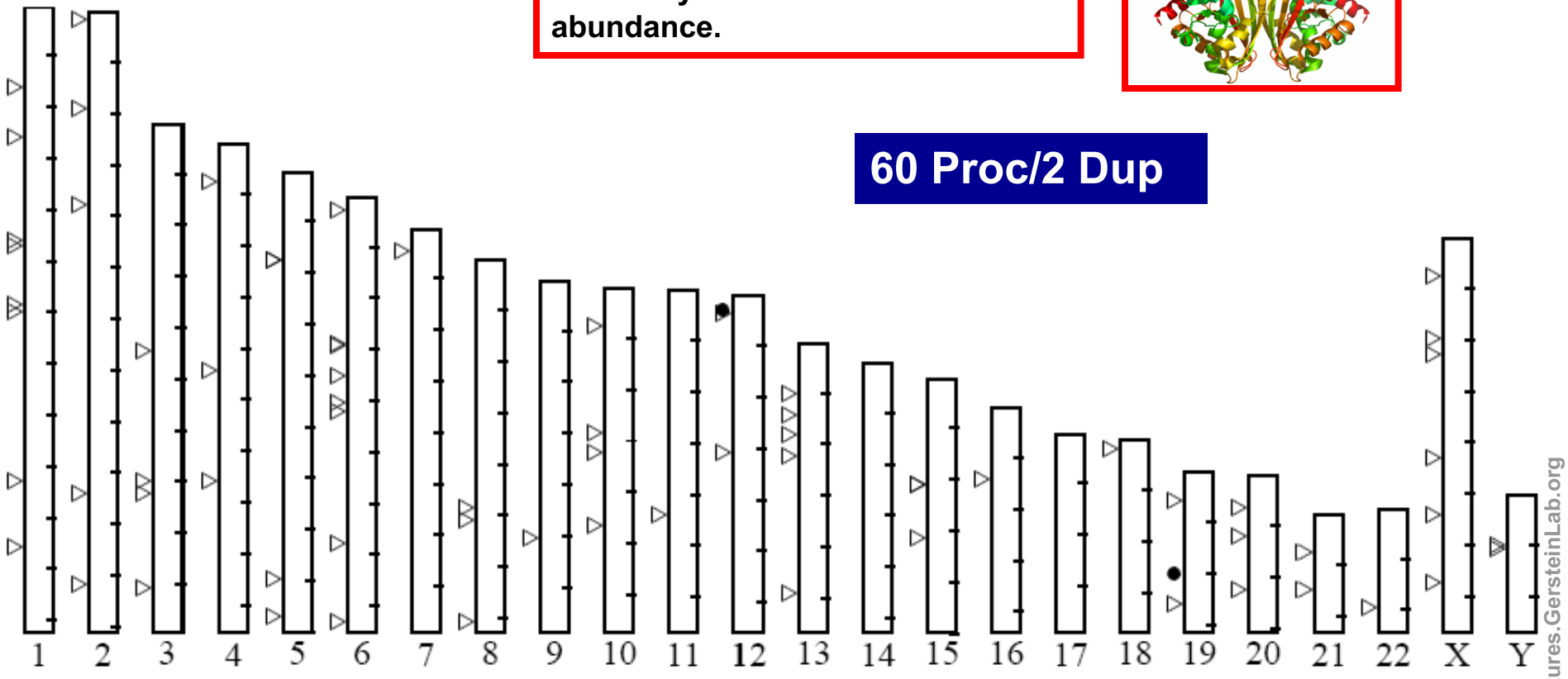
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
ENO	1/0	1/2	12/1	36/3	-	-	-	-	-
PK	2/0	3/0	10/3	4/1	-	-	-	-	-
LDH	10/2	9/1	27/7	25/4	-	-	-	-	-
Total	**97**	**91**	**422**	**463**	**4**	**1**	**0**	**0**	**0**

Distribution of human GAPDH pseudogenes

Large numbers of processed GAPDH pseudogenes in mammals comprise one of the biggest families but numbers not obviously correlated with mRNA abundance.



60 Proc/2 Dup



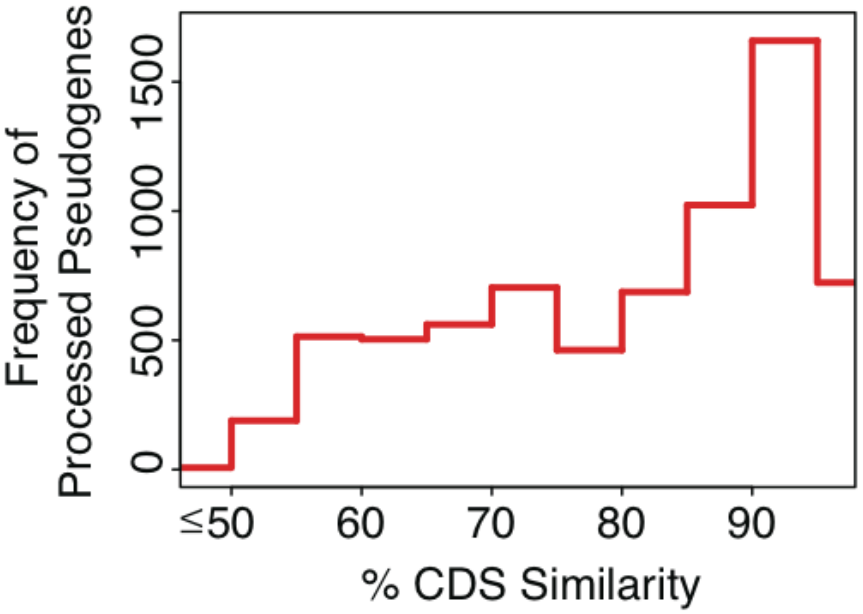
[Liu et al. BMC Genomics ('09, in press)]

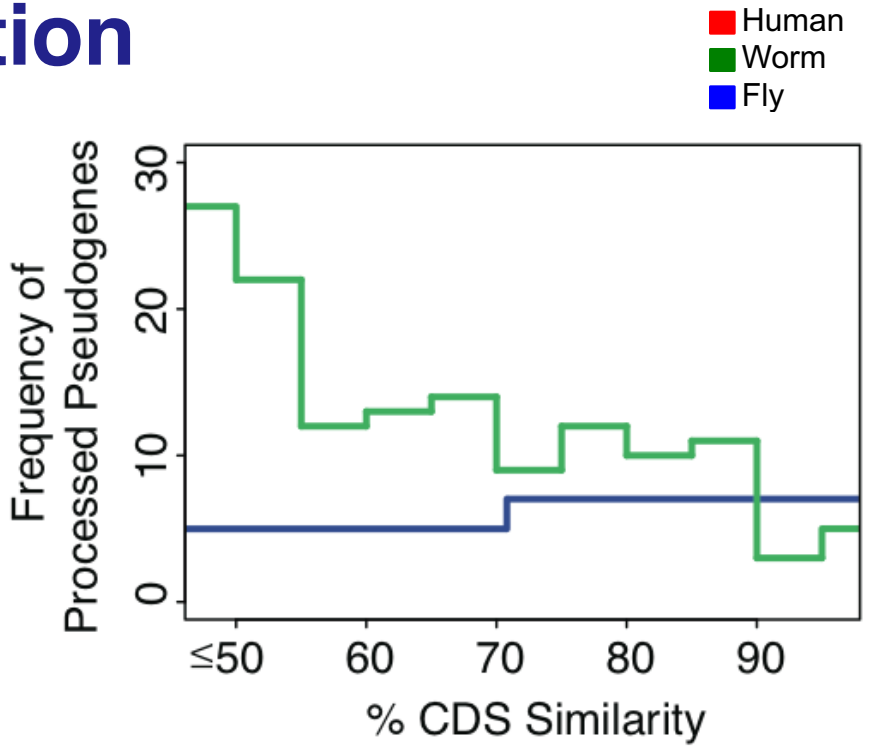
Annotation of Human Pseudogenes in Comparison to those in other Model Organisms

Organism	Total Pseudogenes	Biotype Distrbution (Processed / Duplicated)	ENCODE Functional Genomics Data	Completed Manual Annotation
Human	12,358	8908 / 2266	✔	✔
Worm	911	159 / 566	✔	✔
Fly	145	16 / 109	✔	✔
Zebrafish	229	21 / 177	✔	✔
Macaque	11,136	6570 / 1725	✗	✗
Mouse	13,169	7811 / 1827	✔	✗

[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Evolution

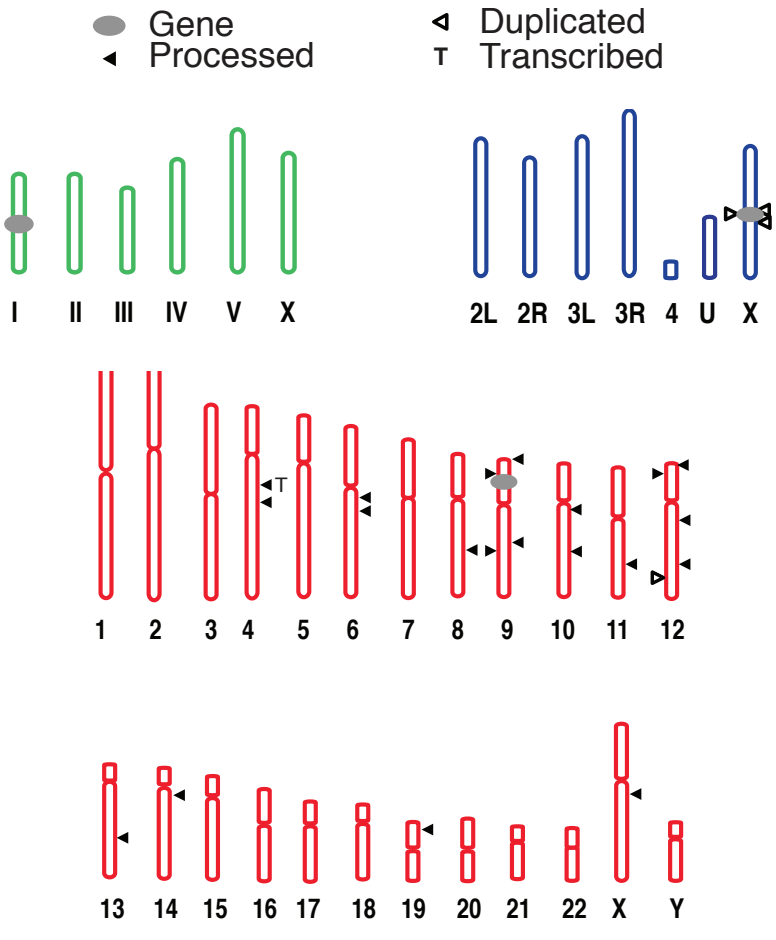


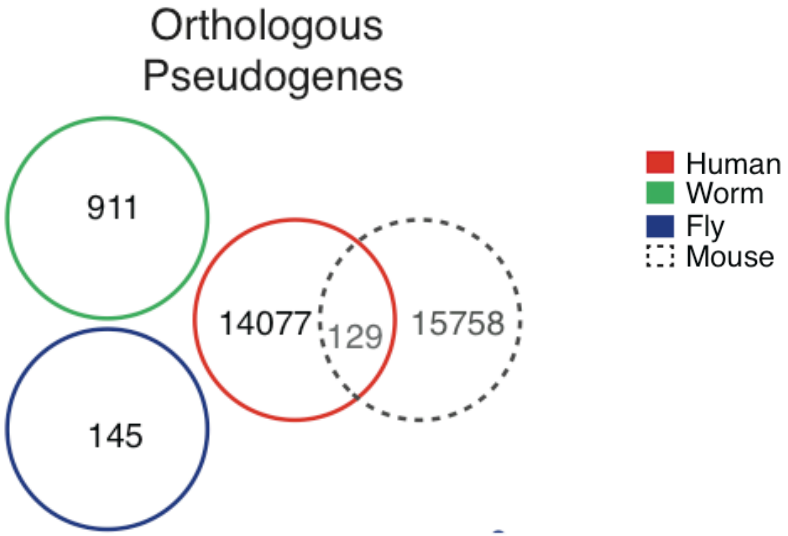


Organism	Defect / Pseudogene x MB		
	Insertion	Deletion	Stop
Human	4.4	4.9	2.4
Worm	**25.8**	7.45	2.5
Fly	7.9	**12.7**	1.1

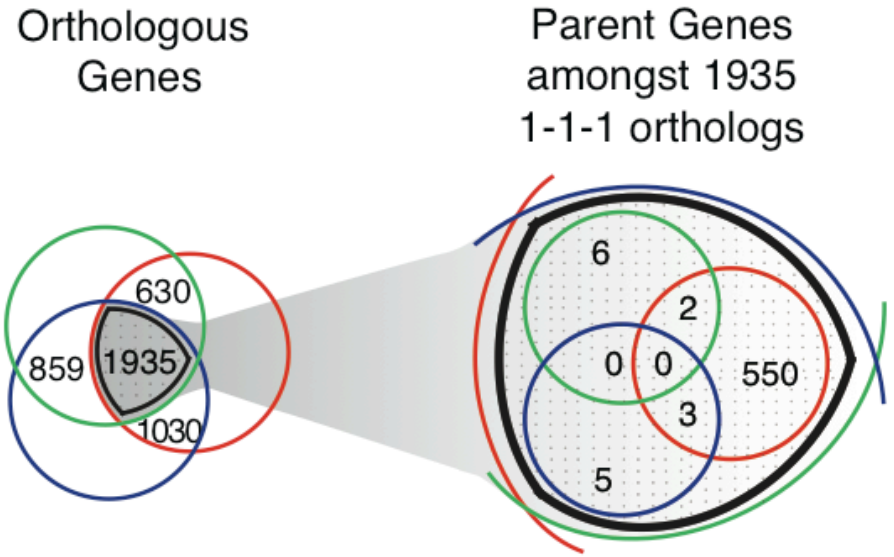
[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Case Study: Ribosomal Protein RpS6



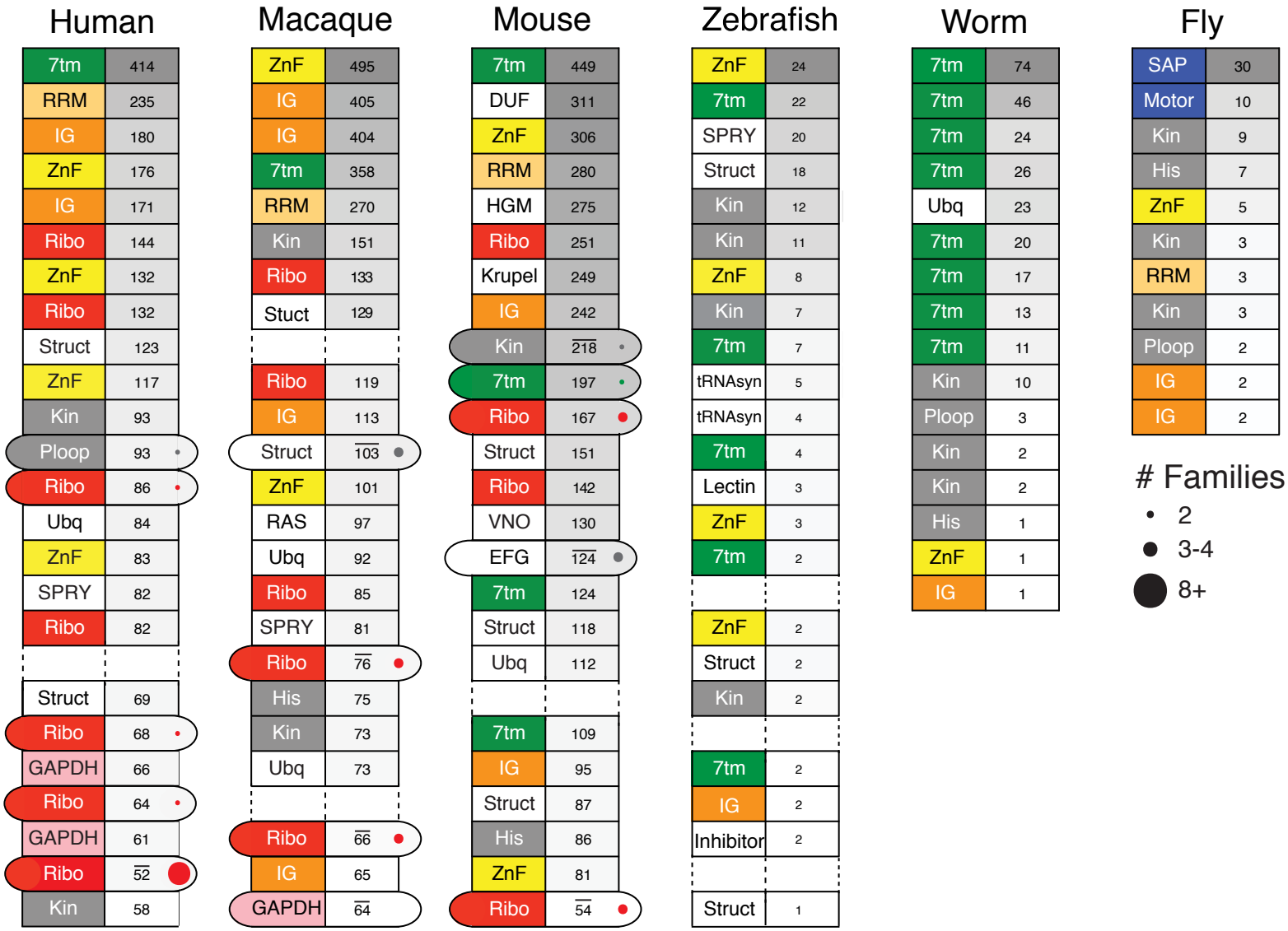


Great divergence in pseudogenes in terms of Orthologs & Paralogs



[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Divergence but
More interpretable Patterns in terms of Families

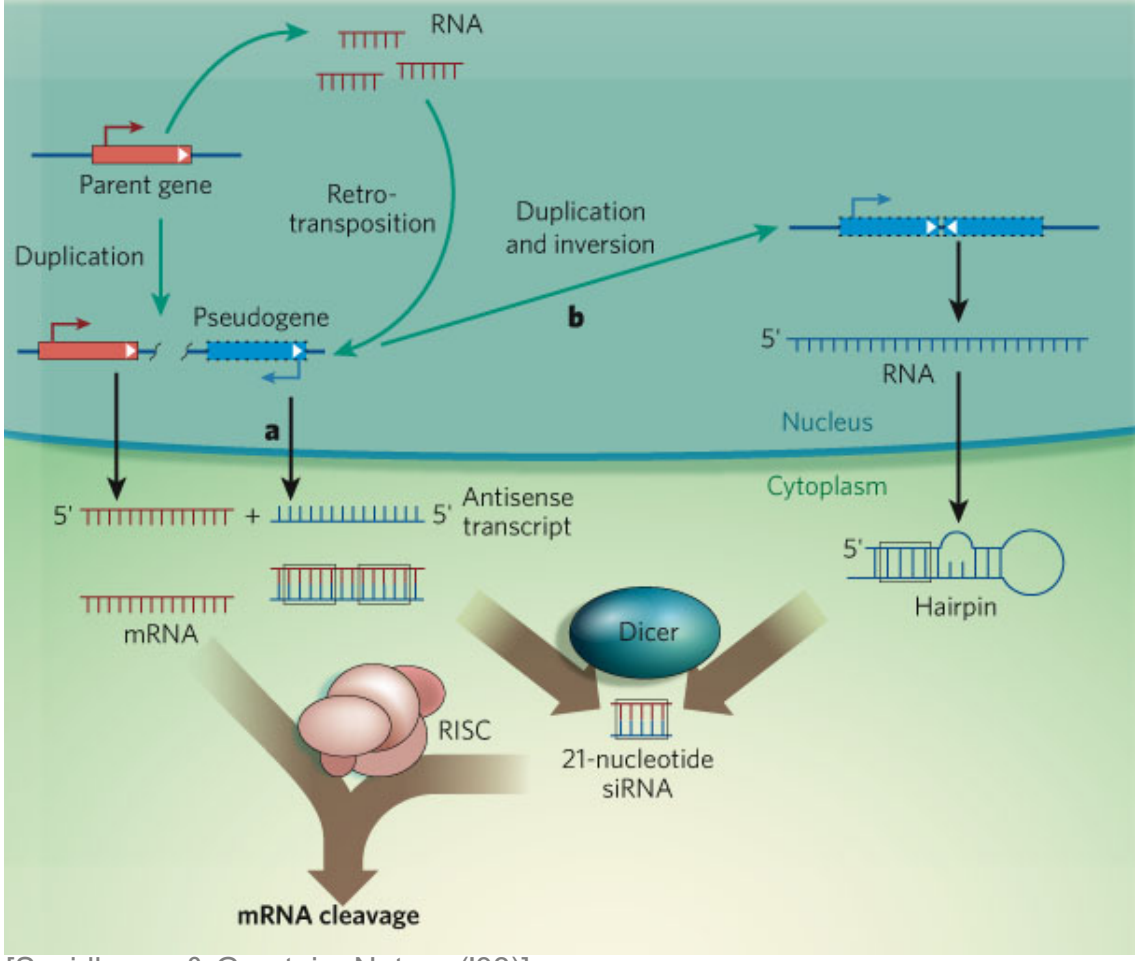


[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Examples & speculation on the function of pseudogene ncRNAs:

Regulating their parents

- via acting as **endo-siRNAs** [ex. in fly & mouse, '08 refs.]
- via acting as **miRNA decoys** [PTEN]
- via **inhibiting degradation** of parent's mRNA [makorin]



[Sasidharan & Gerstein, Nature ('08)]

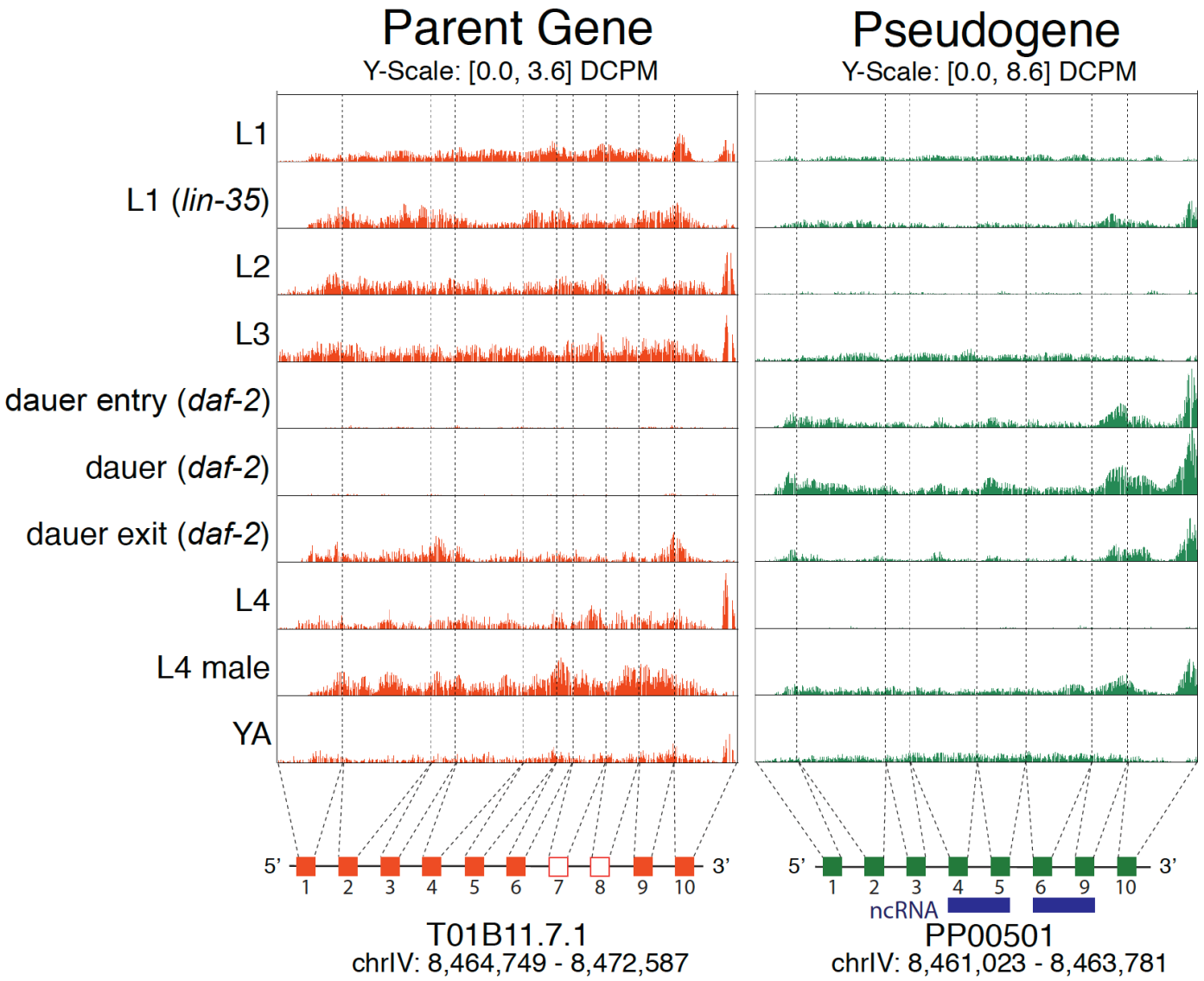
- Functional candidates:
  - SLIT-ROBO Rho GTPase activating protein 2B pseudogene
  - PRKY-004, Y-linked protein kinase pseudogene
  - Fer-1-like 4 (C. elegans), pseudogene

Alternatively, just last gasps of a dying gene

Czech *et al. Nature* 453: 798 ('08).
Ghildiyal *et al. Science* 320: 1077 ('08).
Kawamur *et al. Nature* 453: 793 ('08).
Okamura *et al. Nature* 453: 803 ('08).
Tam *et al. Nature* 453: 534 ('08).
Watanabe *et al. Nature* 453: 539 ('08).

Poliseno et al. Nature 465:1033 ('10).

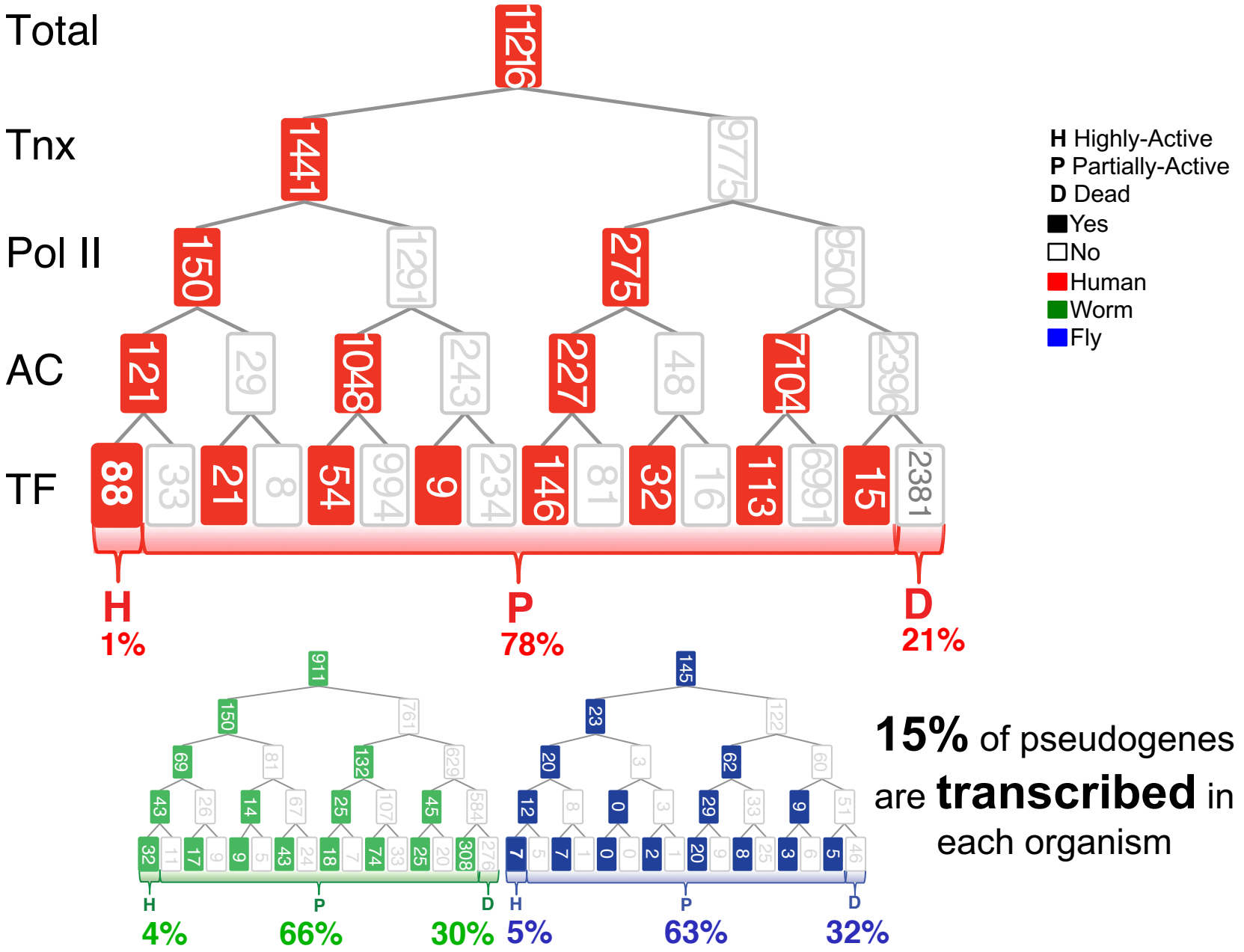
Pseudogene Transcription: interesting but tricky to ascertain



- Difficulty in ascertainment because of mis-mapping v parent
- One approach to this confound is look across mult. samples

[*Science* 330:6012]

Pseudogene Activity



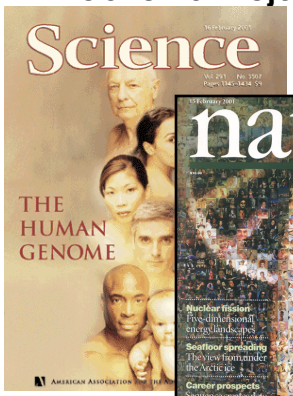
15% of pseudogenes are **transcribed** in each organism

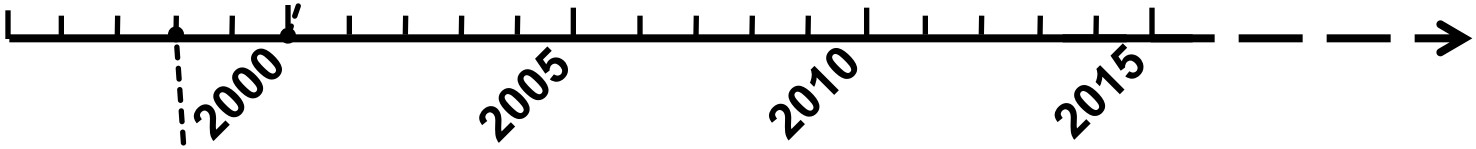
[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

The Human Genome Project



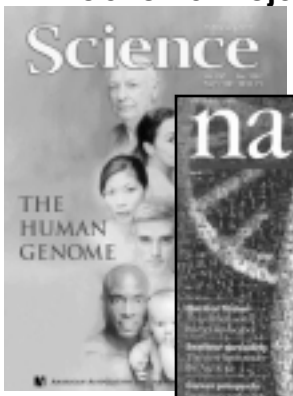


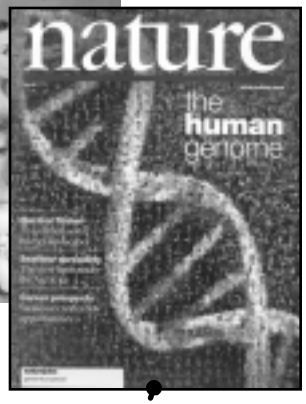
Worm Genome

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

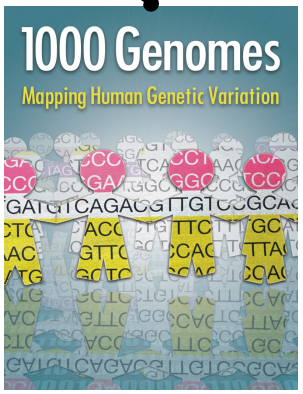
Comparative ENCODE

2000

2005

2010

2015



Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Production

The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE

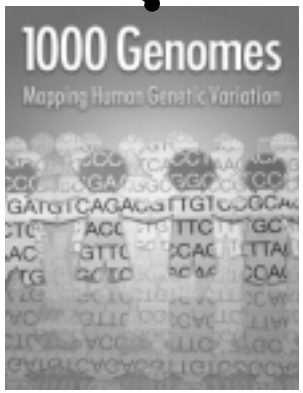
Epigenome Roadmap

2000

2005

2010

2015



Worm Genome

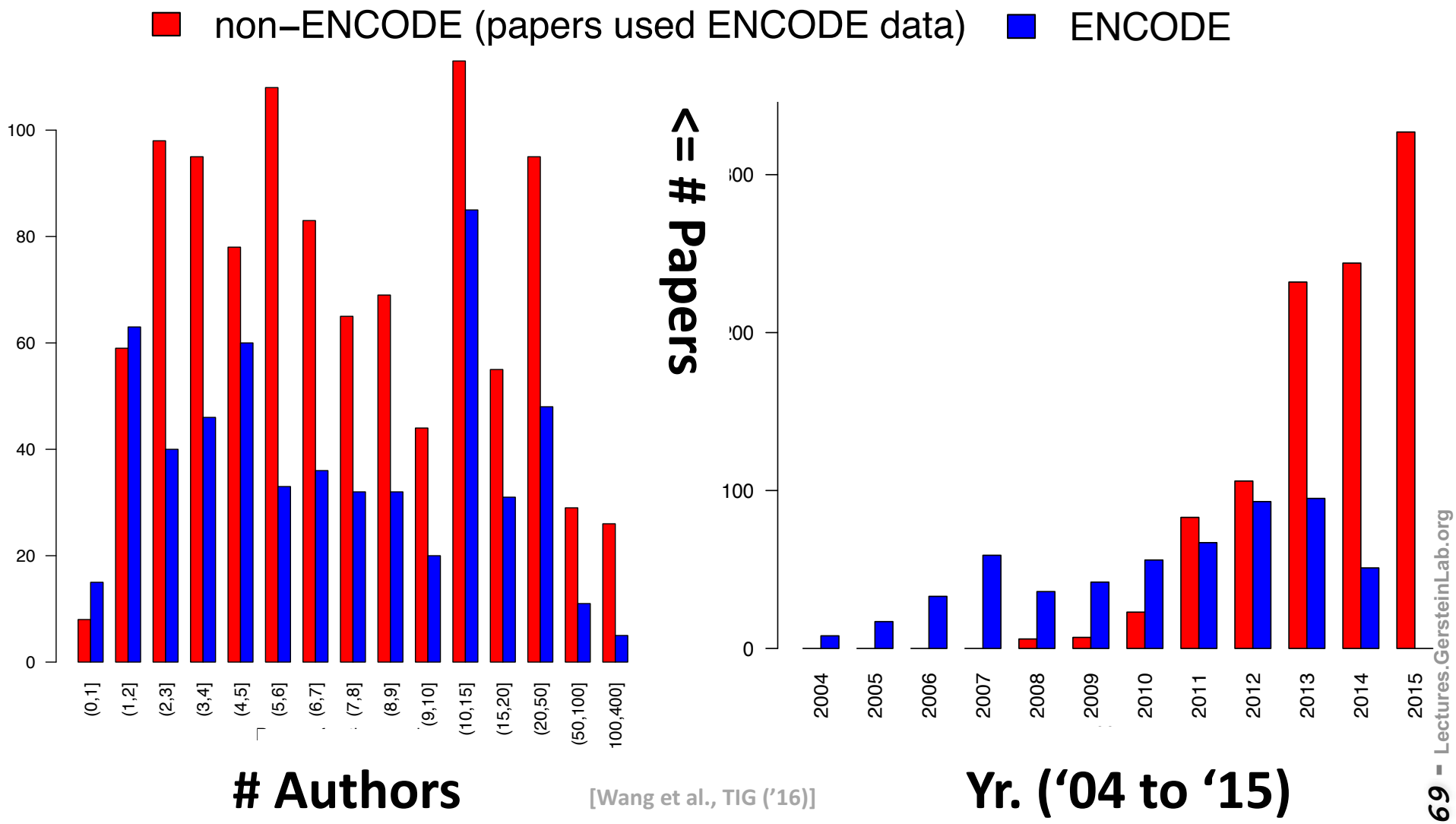
modENCODE

1000 Genomes Pilot

1000 Genomes Production

GTEx

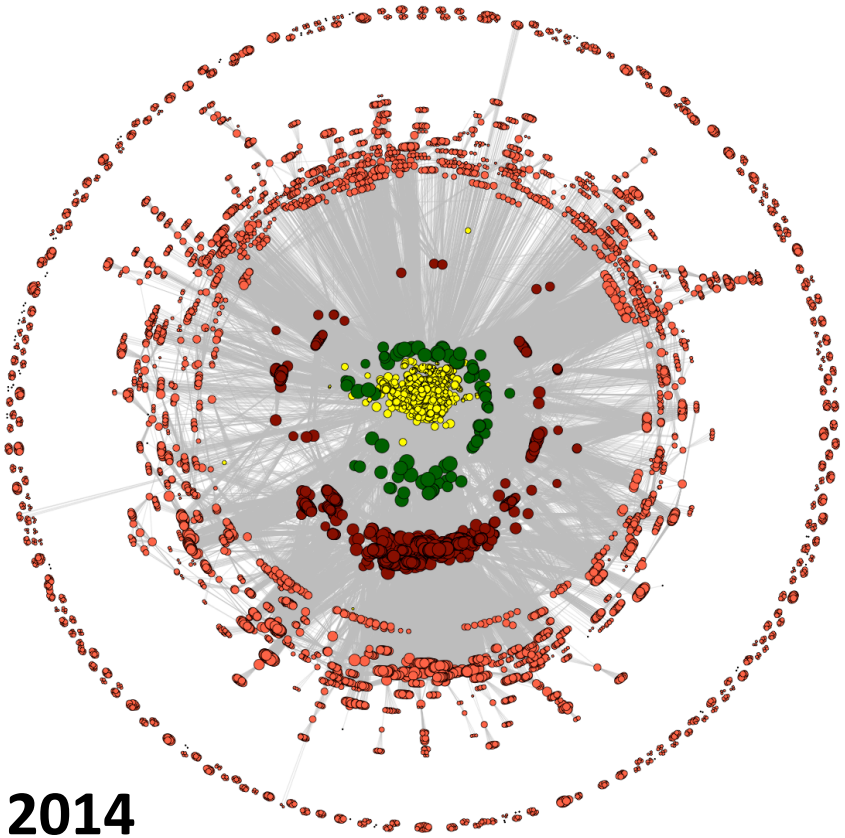
With help of M Pazin at NHGRI, identified: 702 community papers that used ENCODE data but were not supported by ENCODE funding &
558 consortium papers supported by ENCODE funding
(https://www.encodeproject.org/search/?type=Publication for up-to-date query)
Then identified 1,786 ENCODE members & 8,263 non-members .



Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



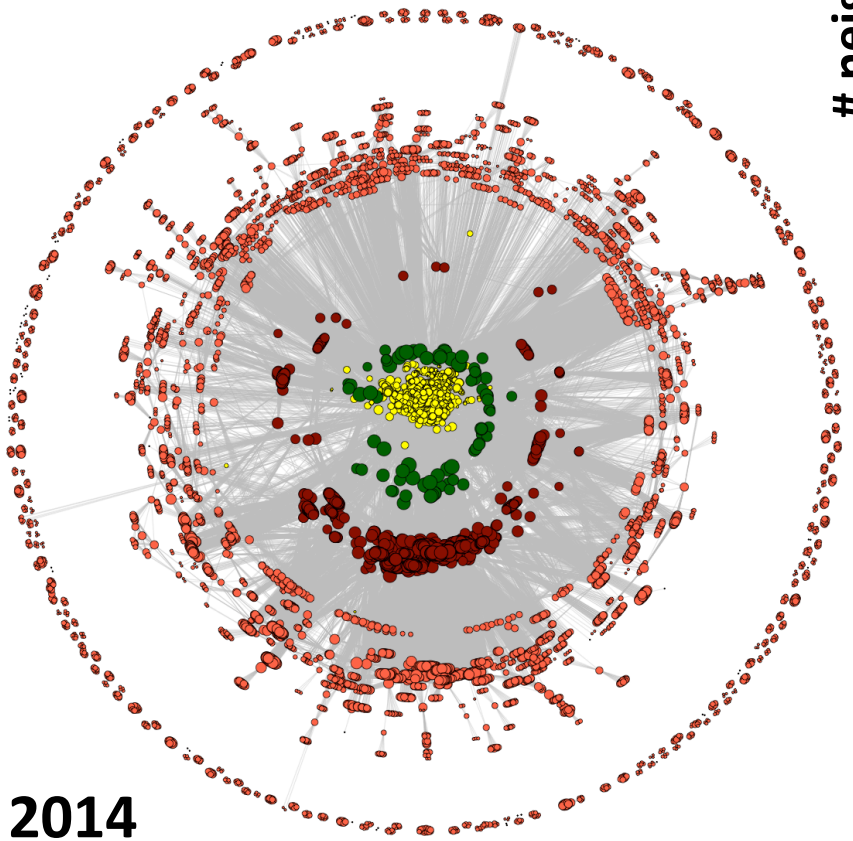
2014

[Wang et al., TIG ('16)]

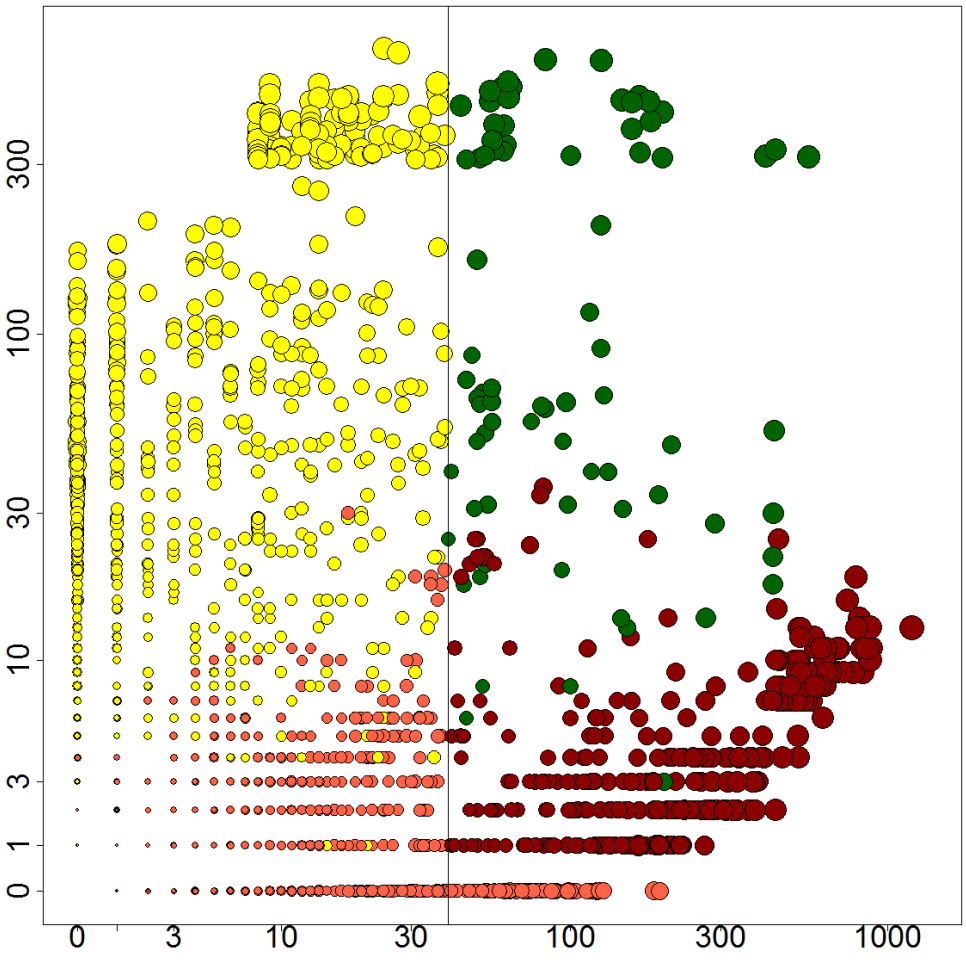
Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



2014

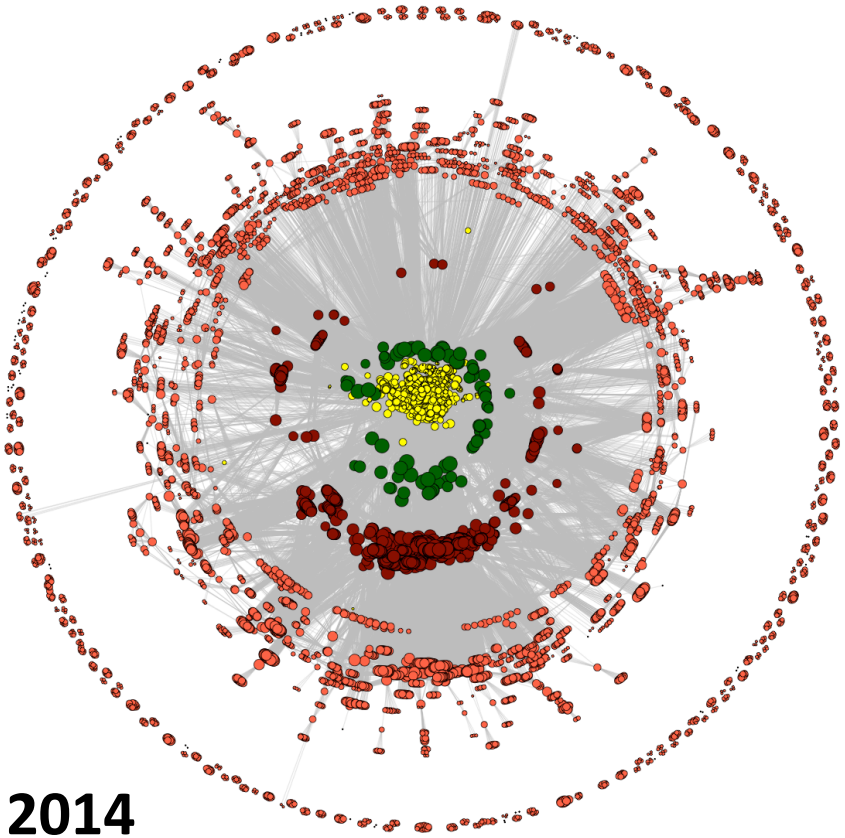


[Wang et al., TIG ('16)]

Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

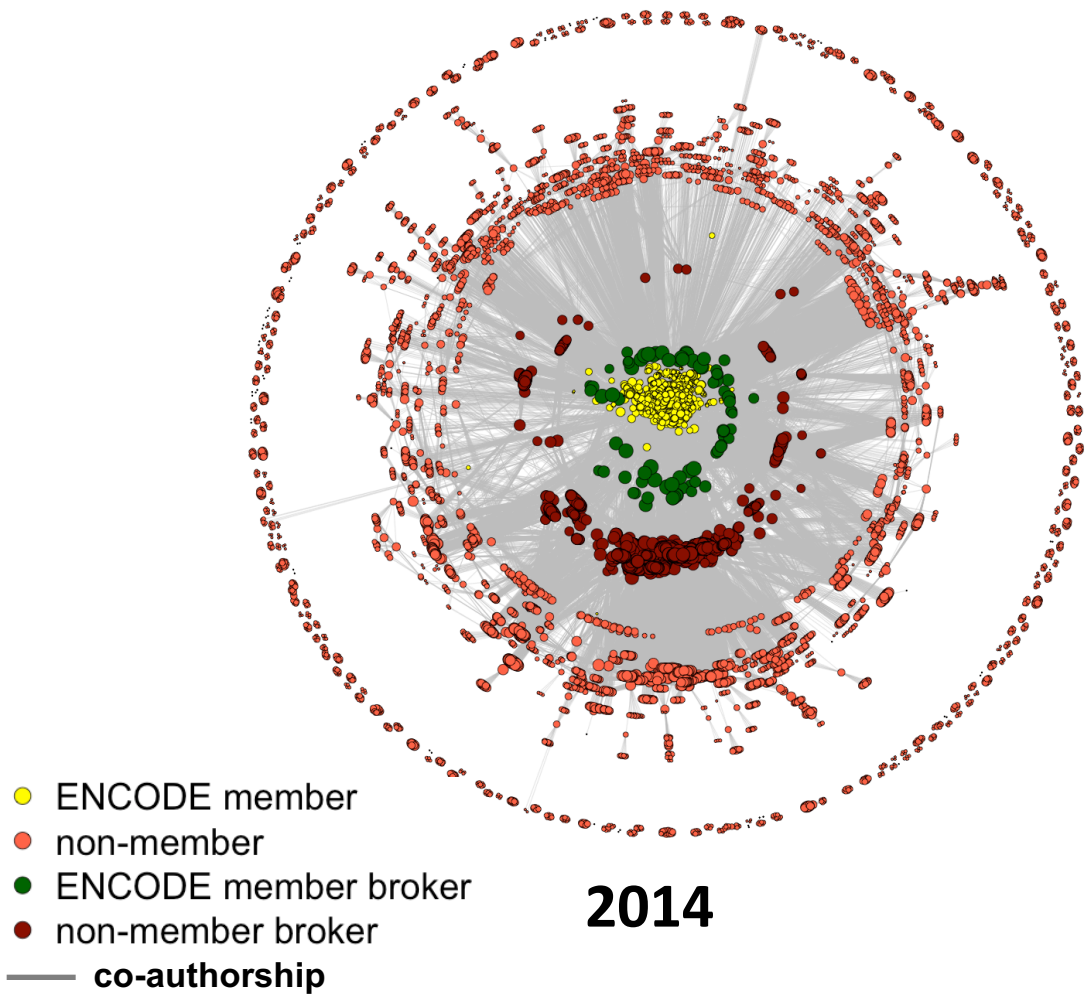
co-authorship



2014

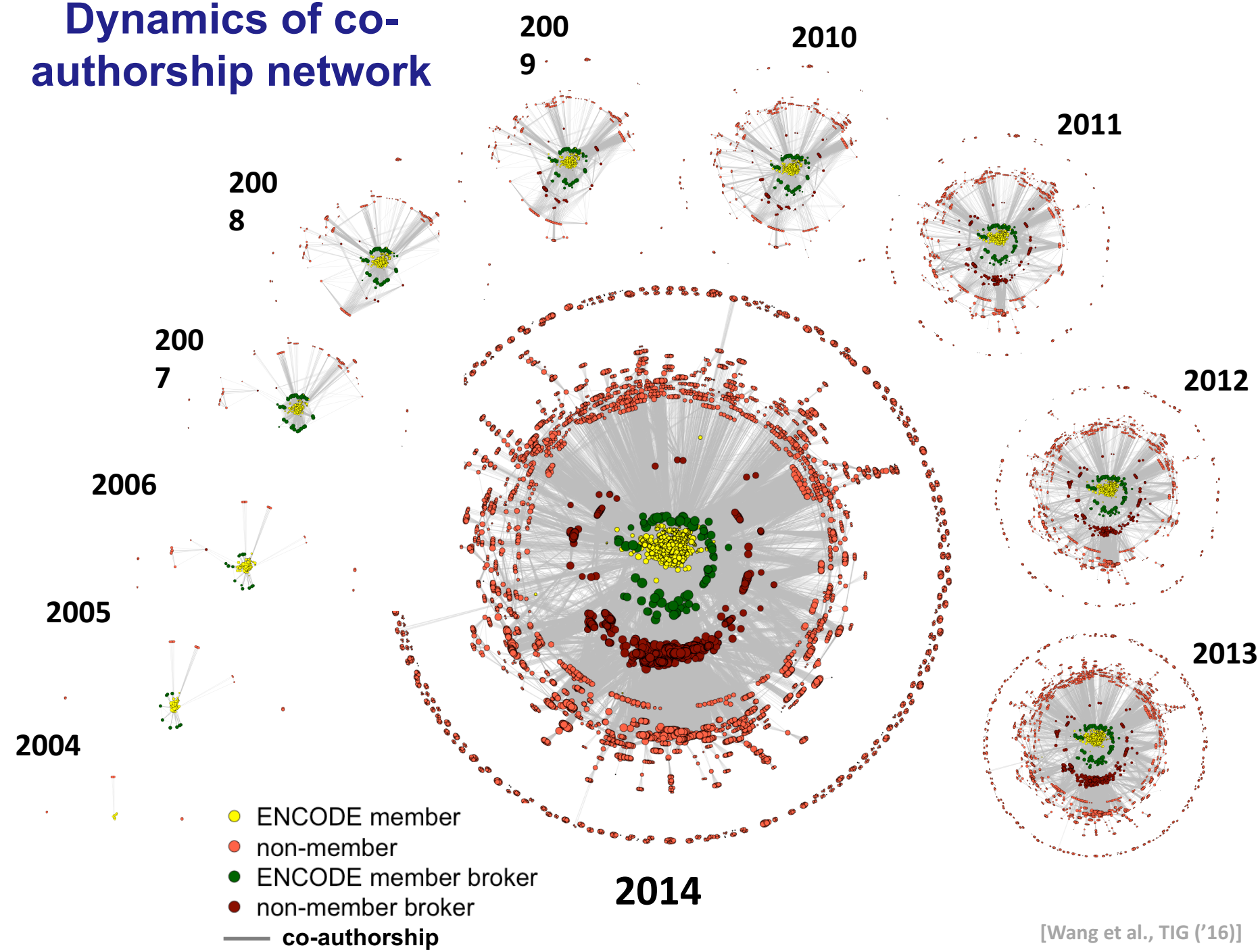
[Wang et al., TIG ('16)]

Dynamics of co-authorship network



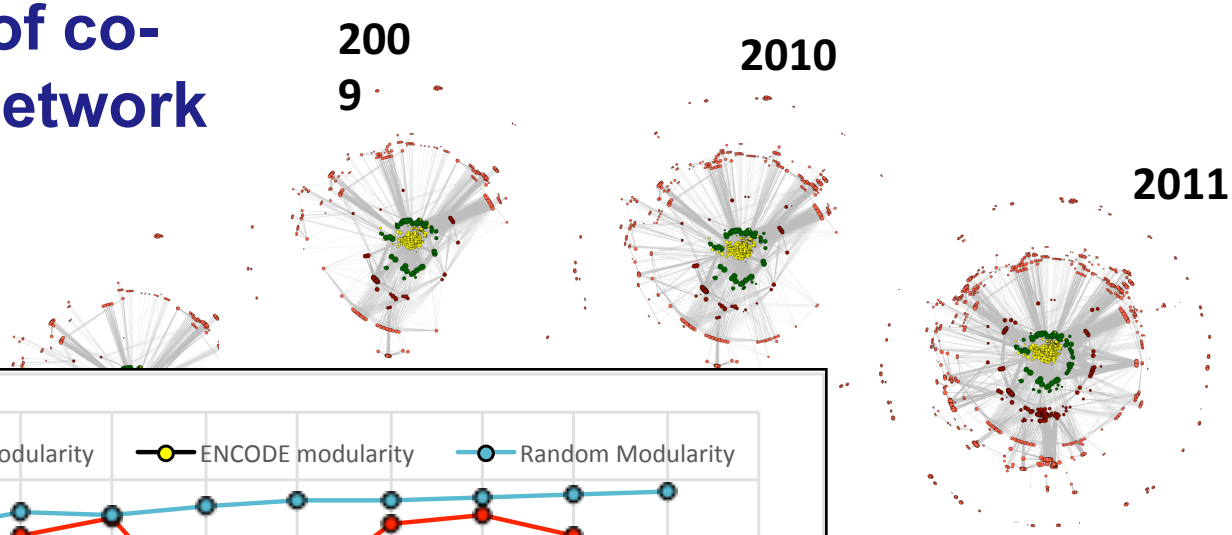
[Wang et al., TIG ('16)]

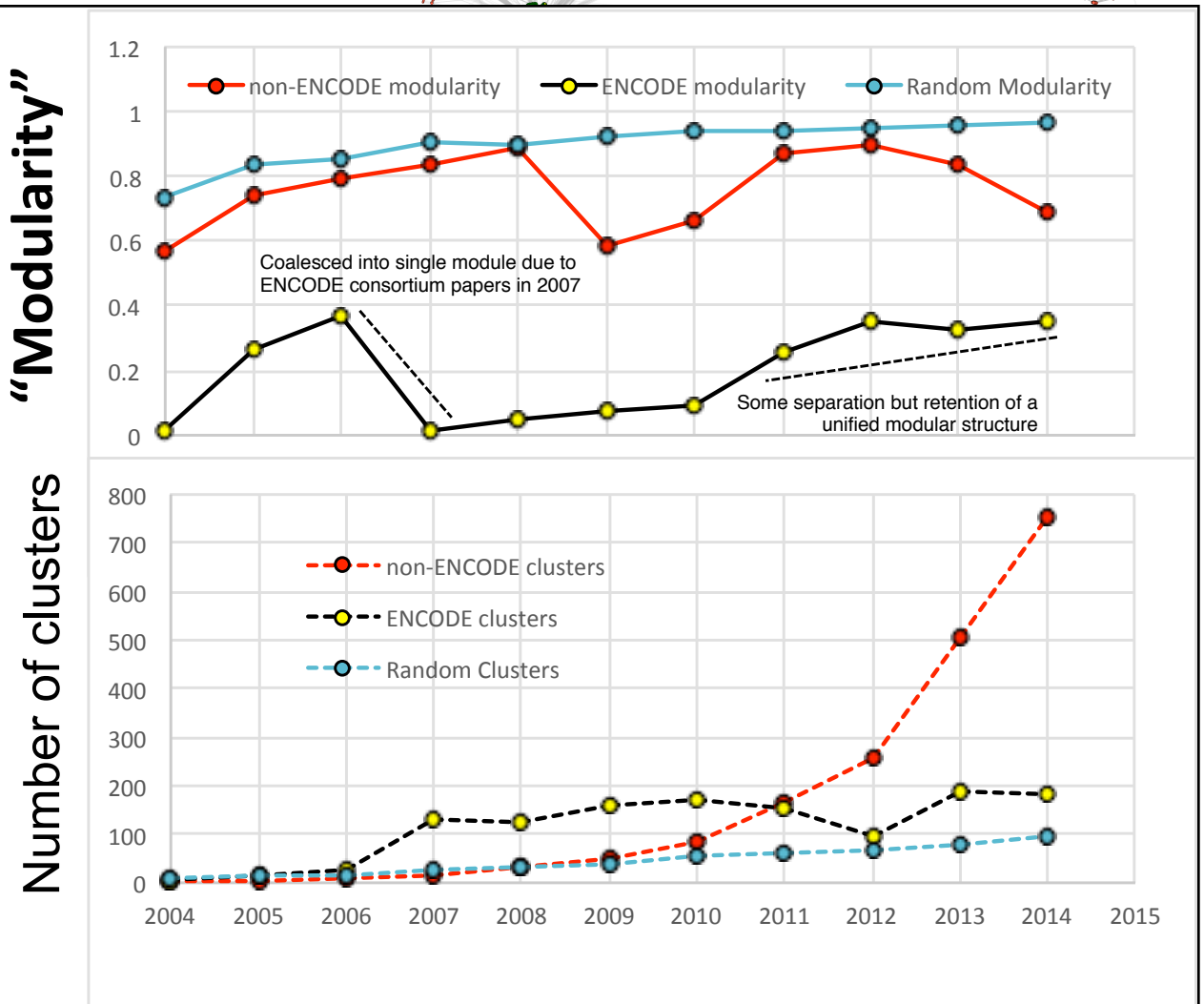
Dynamics of co-authorship network



[Wang et al., TIG ('16)]

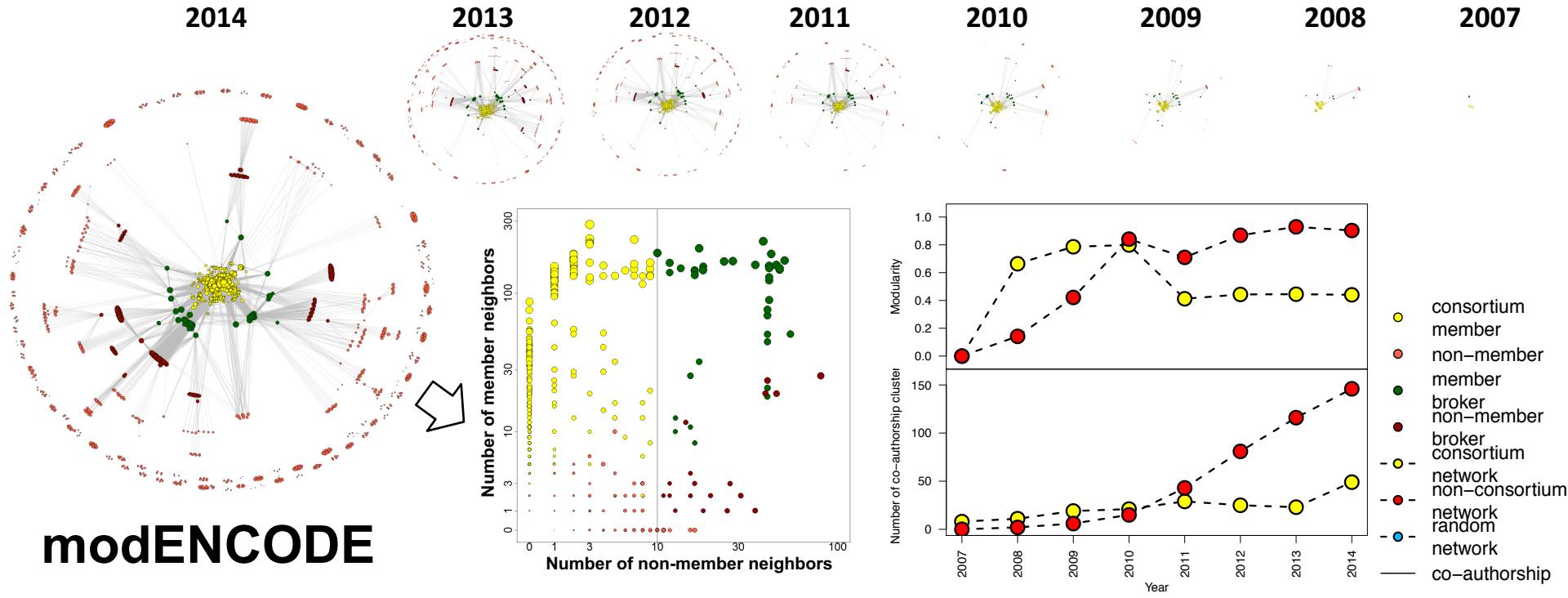
Dynamics of co-authorship network





[Wang et al., TIG ('16)]

Similar findings in terms of slow growth trends & broker scientists in the modENCODE consortium as for ENCODE



[Wang et al., TIG ('16)]

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Comparing Diverse Transcriptomes to Determine Deeply Conserved Aspects of Gene Expression

- 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)
  - Developmental 'hourglass' genes in 12 of these. They also exhibit intra-organism hourglass behavior.
  - Stage alignment of worm & fly development, strongest with hourglass genes
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones. Also, Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- Characterizing **ncRNAs & TARs**
  - Not much news in canonical gene models
  - Simple contig search (TARs) finds uniform density of non-canonical transcription
  - ML model shows few TARs similar to existing ones, but some enrichment for eRNAs
- **Pseudogenes**
  - Fundamentally repetitive elements
  - Collaborative assignment in results in ~14K
  - Impact of lineage-specific retro-transpositional burst – ie human v other metazoans is dominated (~80%) by retro-duplication ~40 MYA (Ribo. Proteins).
  - Many Pseudogenes with Low Levels of Biochemical Activity
    - ~15% transcribed & 80% w/ some activity
- Value of **publication patterns generated by the consortium**
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Acknowledgements

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 !

EncodeProject.org/comparative/transcriptome

DREISS.gersteinlab.org

D **Wang**, F He, S Maslov

Pseudogene.org/psicube

C **Sisu**, B Pei, Jing Leng, A Frankish, Yan Zhang, Suganthi Balasubramanian, Rachel Harte, Daifeng Wang, Michael Rutenberg-Schoenberg, Wyatt Clark, Mark Diekhans, Joel Rozowsky, Tim Hubbard, J Harrow

Publication patterns [“encode authors”]

D **Wang**

KK Yan, J Rozowsky, E Pan

github.com/gersteinlab/**OrthoClust**

KK **Yan**, D Wang, J Rozowsky, H Zheng, C Cheng

Hiring Postdocs. See gersteinlab.org/jobs !

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to 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