Comparative Genome Analysis:

Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

M Gerstein, Yale

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

See last slide for
references & more info.

How might we annotate a human text?

Color is Function

Lines are Similarity

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

The Semicolon Wars

Brian Hayes

IF YOU WANT TO BE a thorough-going 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.

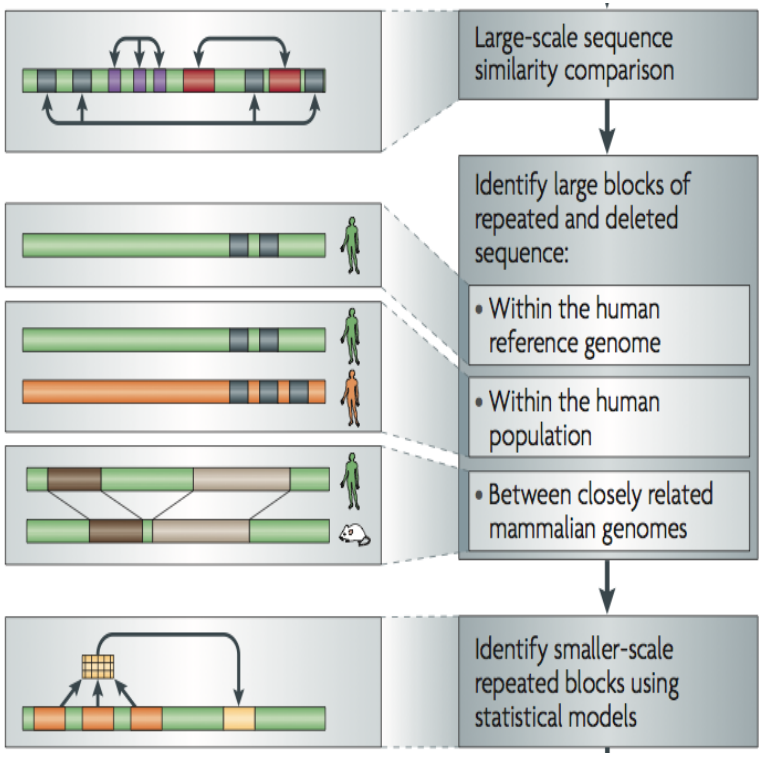
Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in `x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C

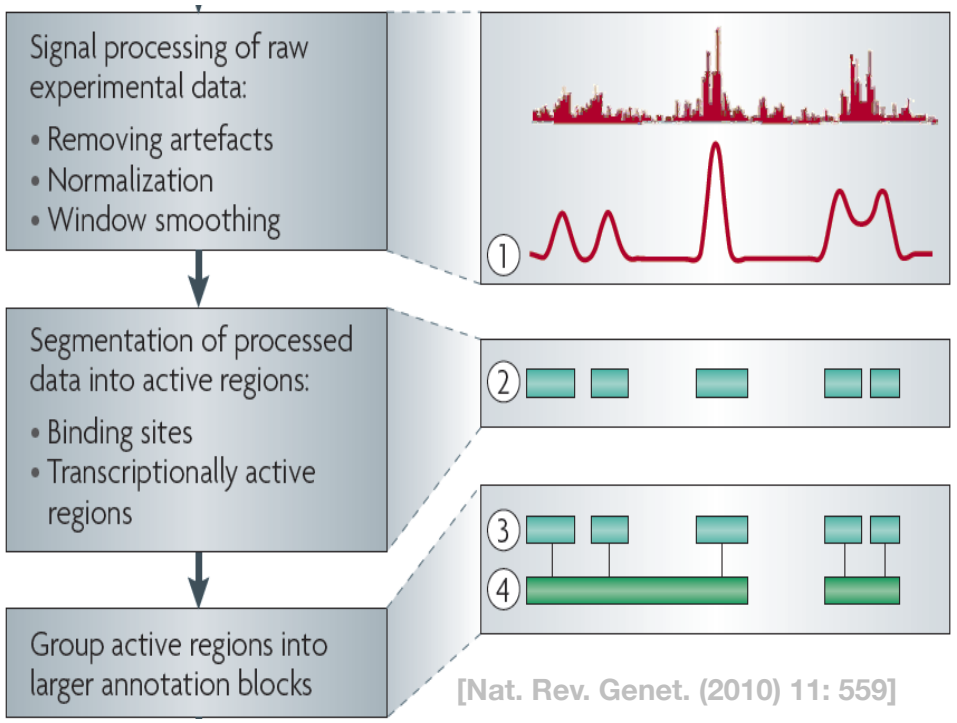
Non-coding Annotations: Overview

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

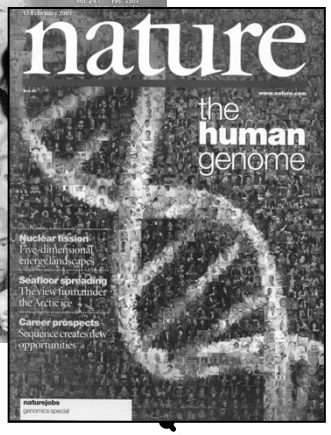
Sequence features, incl. **Conservation**

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

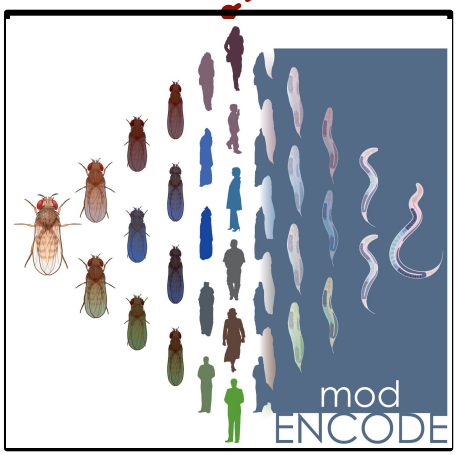




[Nat. Rev. Genet. (2010) 11: 559]



ENCODE Consortium & various annotation rollouts

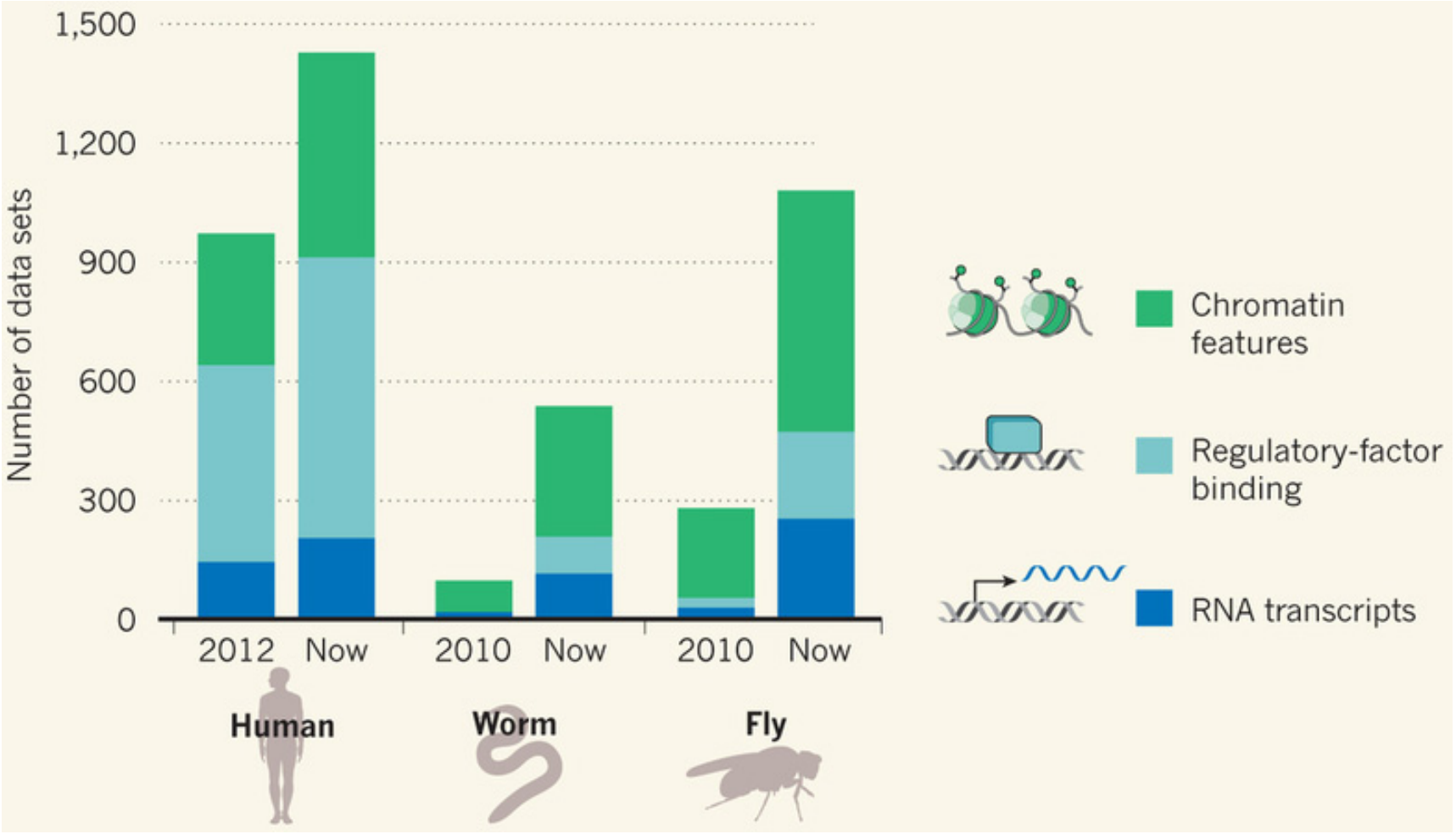


1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

Comparative ENCODE Functional Genomics Resource

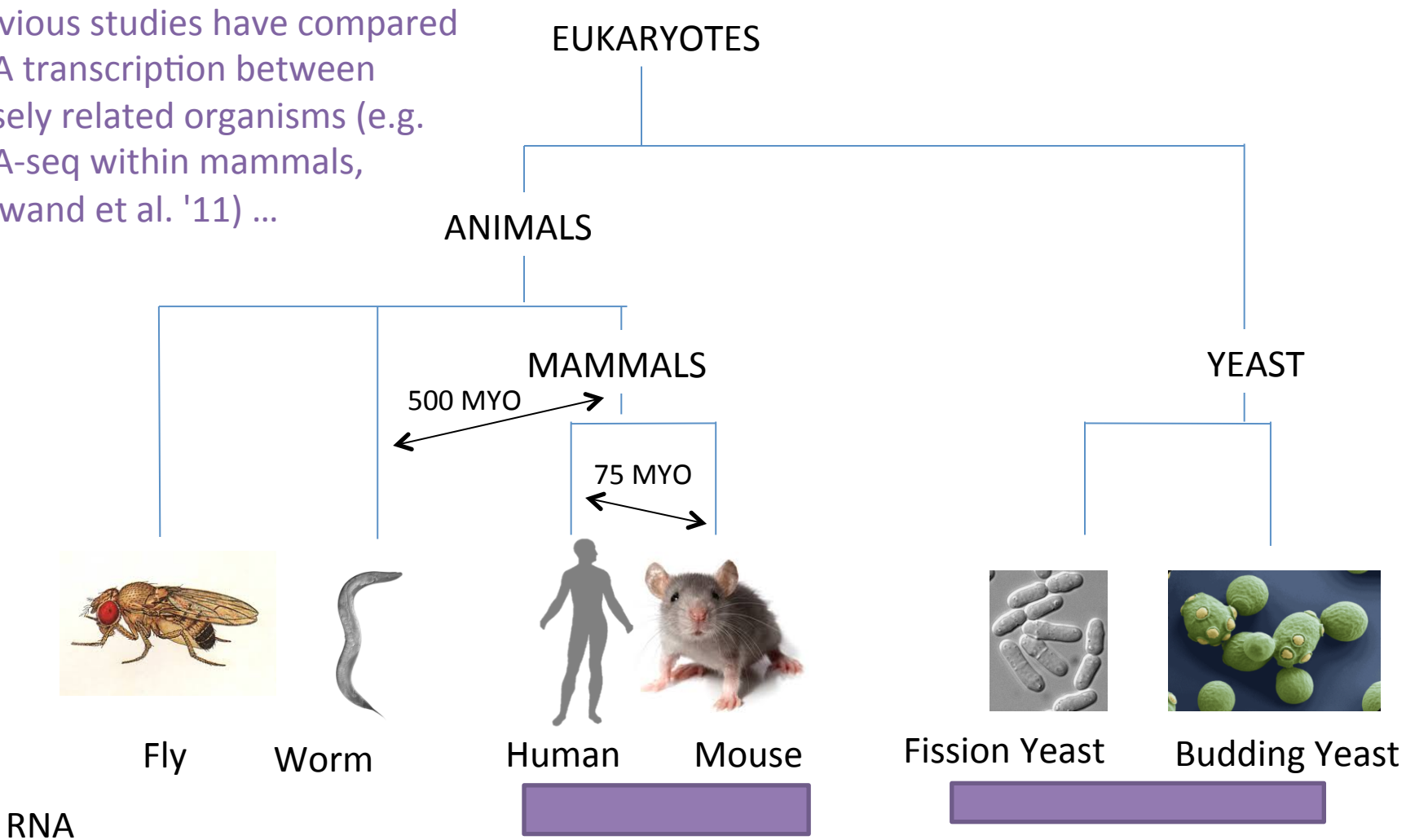
(EncodeProject.org/modENCODE.org)

- 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)



Comparative ENCODE

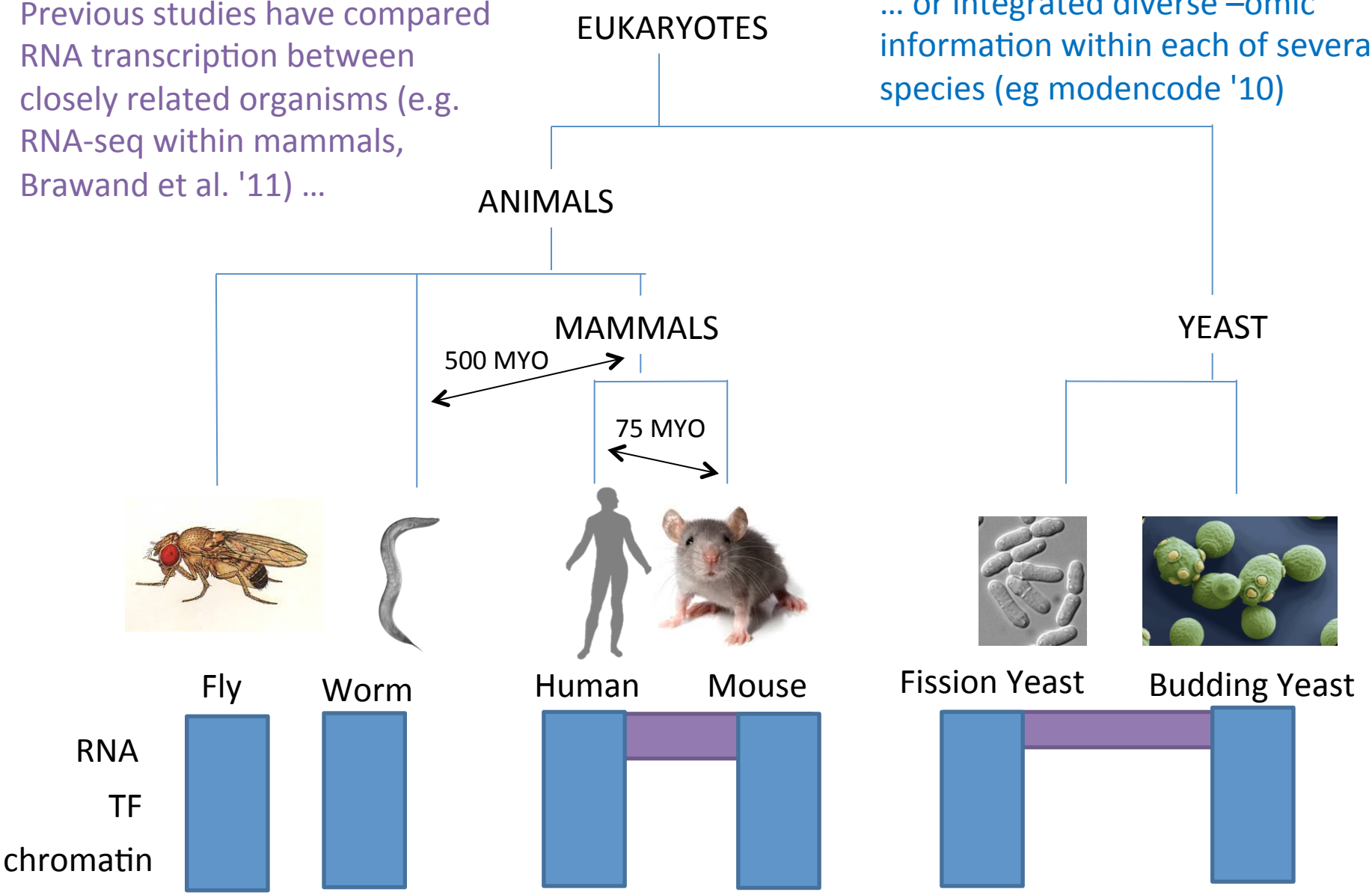
Previous studies have compared RNA transcription between closely related organisms (e.g. RNA-seq within mammals, Brawand et al. '11) …



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)

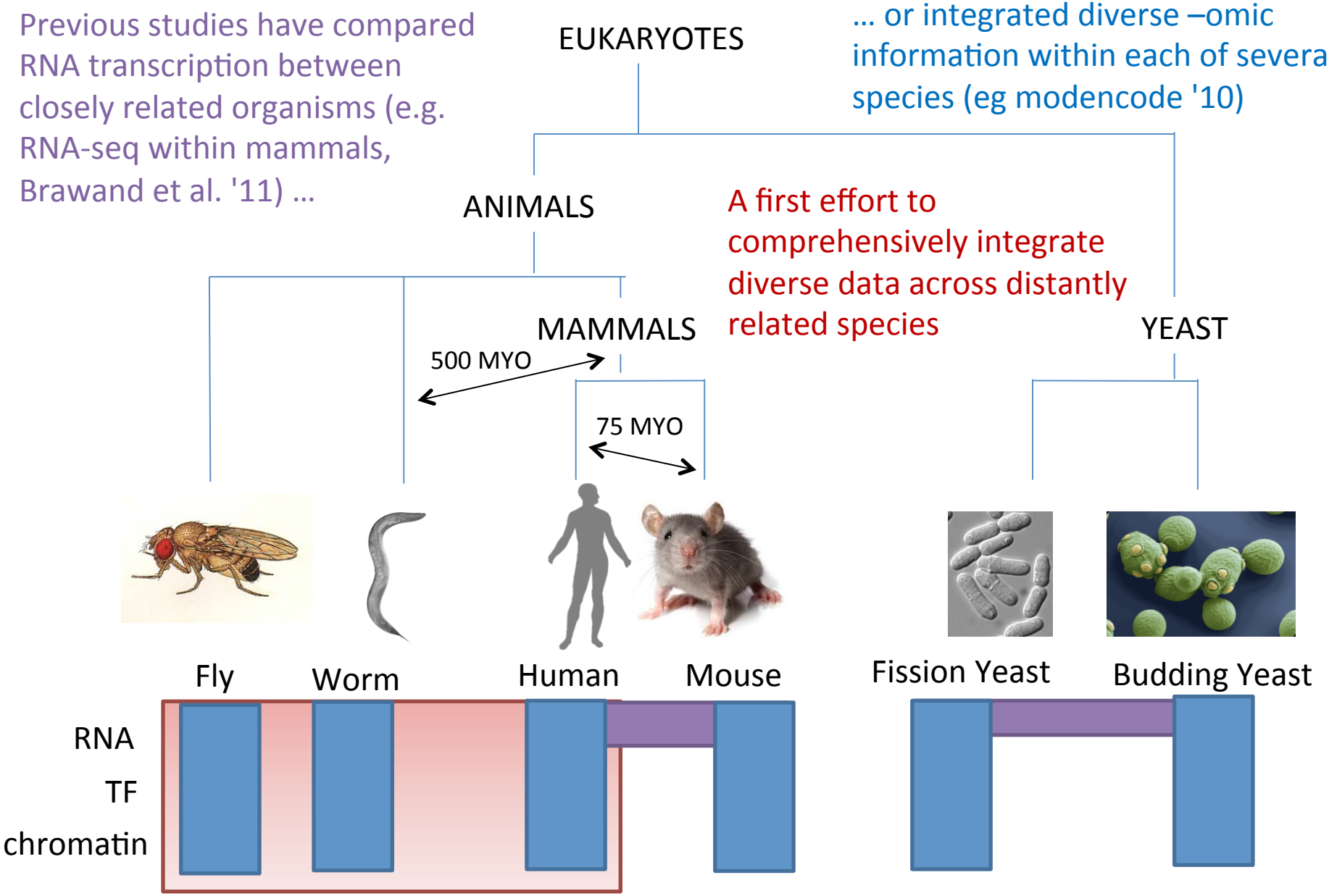


Comparative ENCODE

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)

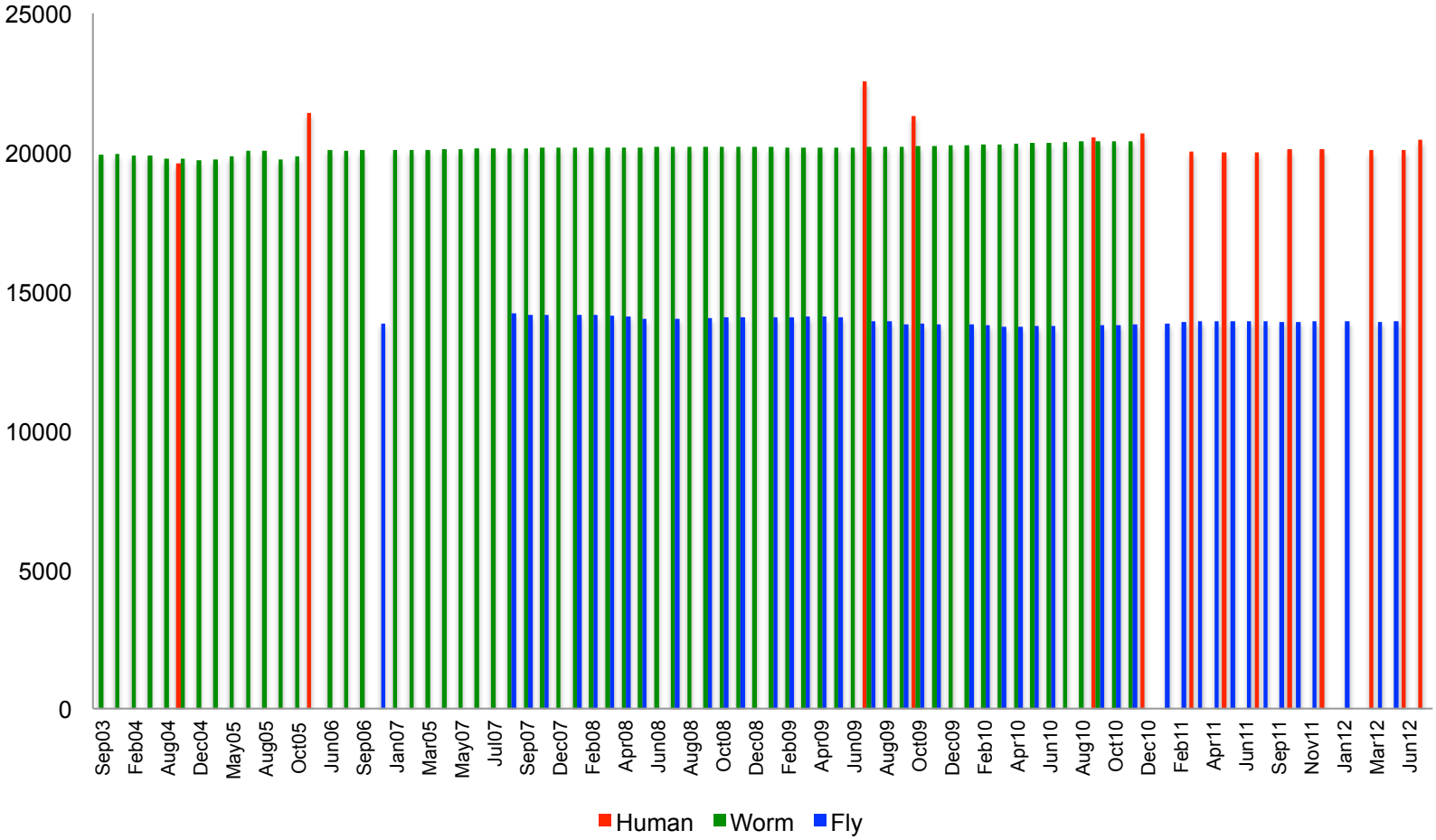
A first effort to comprehensively integrate diverse data across distantly related species



Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

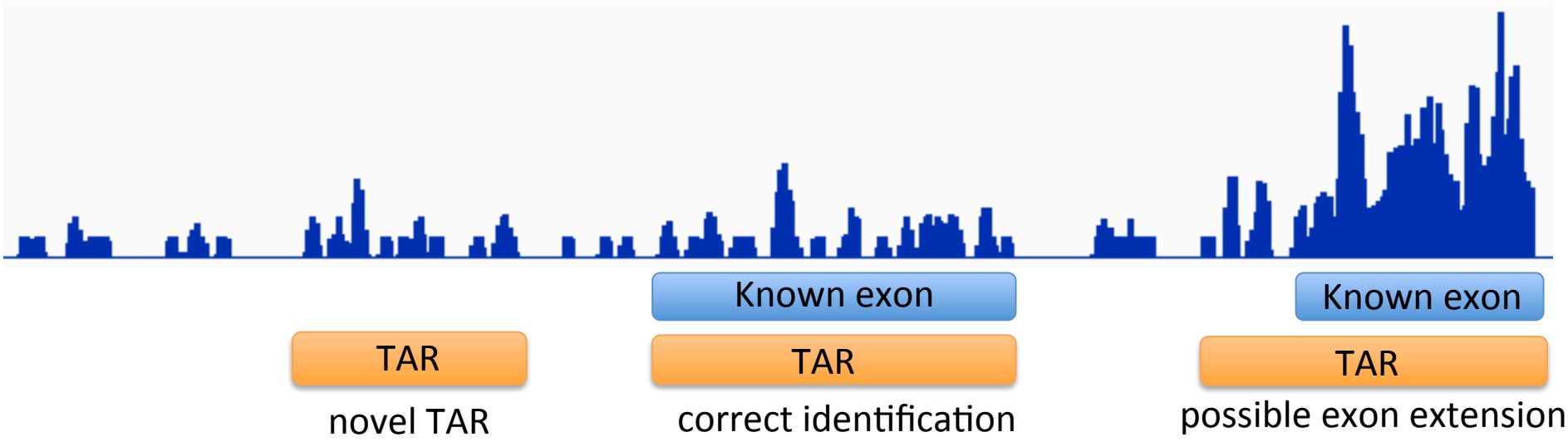
Protein-coding gene counts in worm, fly & human have stabilized & have remained fairly constant



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

Discovering Transcriptionally Active Regions (novel RNA contigs)

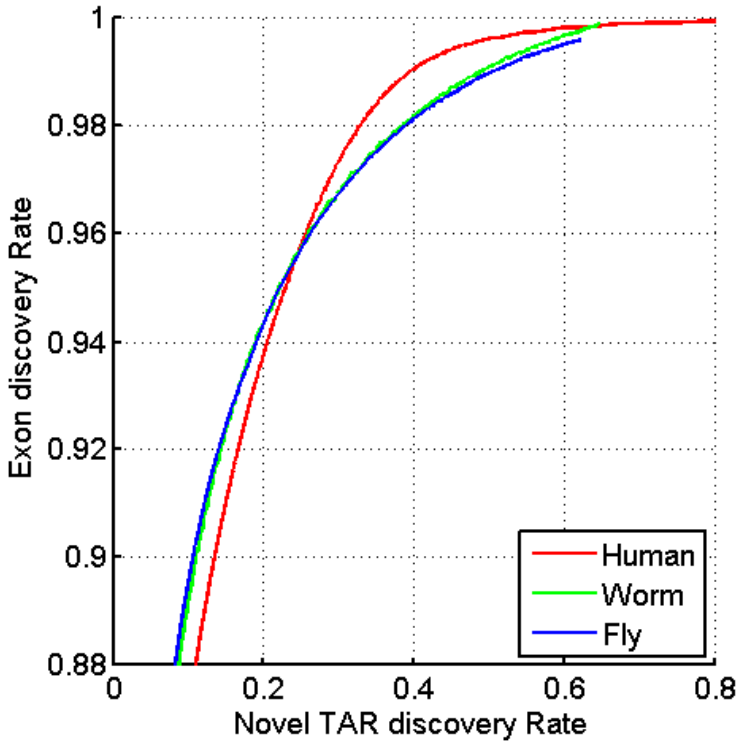
- Cluster reads setting minimum-run and maximum gap parameters for newly identified transcribed regions (TARs)
- Assess exon discovery rates for known genes and noncoding RNAs



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

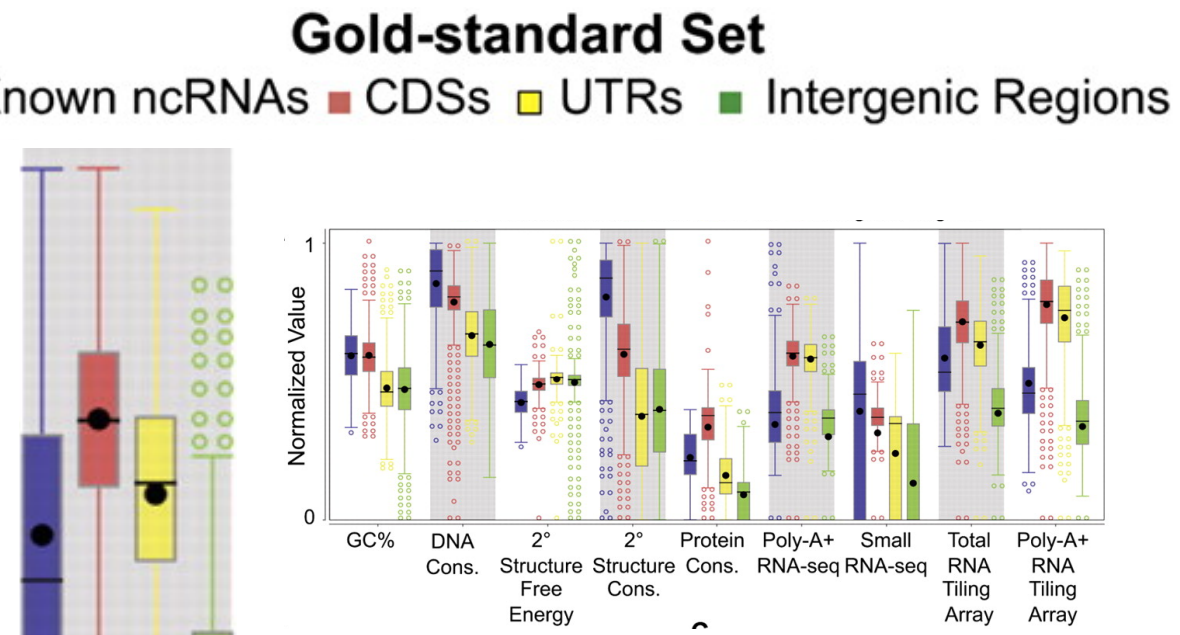
	Human			Worm			Fly		
	Elements	Genome Coverage		Elements	Genome Coverage		Elements	Genome Coverage	
		Kb	%		Kb	%		Kb	%
Total ncRNAs	22,154	17,770	0.62	41,466	2,611	2.6	2,155	3,279	2.6
Regions Excluding mRNAs, Pseudogenes or Annotated ncRNAs	283,816	2,731,811	95.5	143,372	63,520	63.3	60,108	89,445	69.6
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

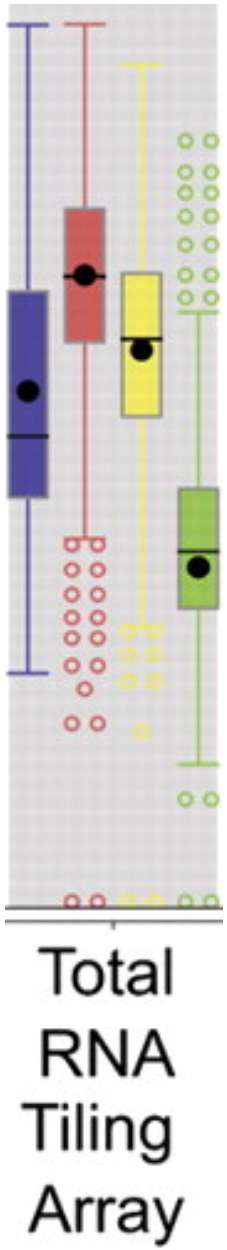
- Similar fraction of non-canonical transcription of non-canonical transcription in human, worm and fly
  - 32-37% of each genome

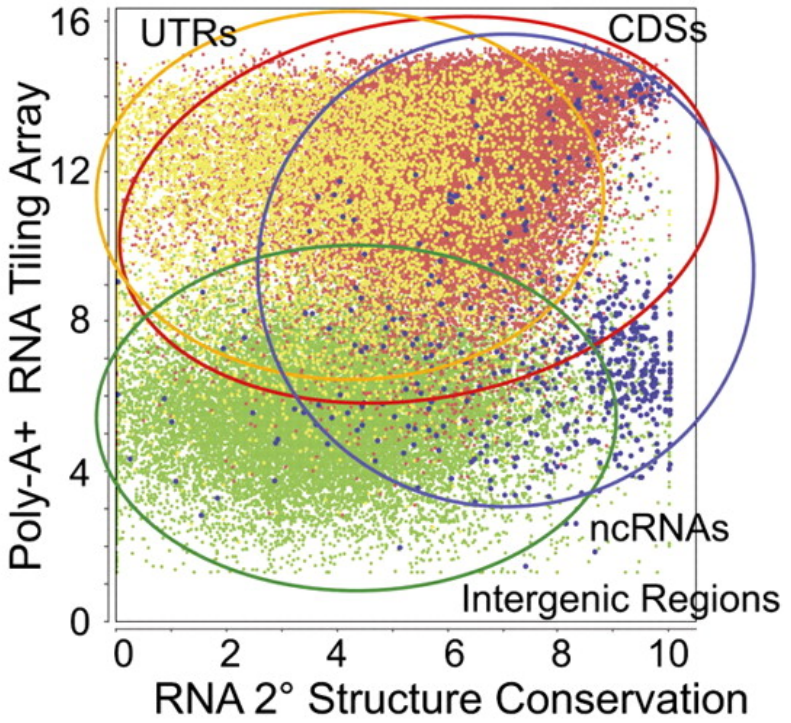
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

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





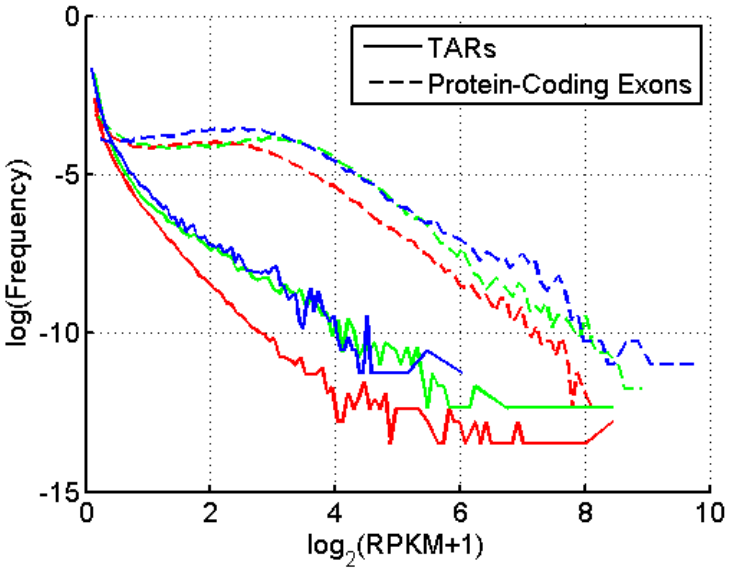


[Lu et al. Genome Res. 2011;21:276-285]

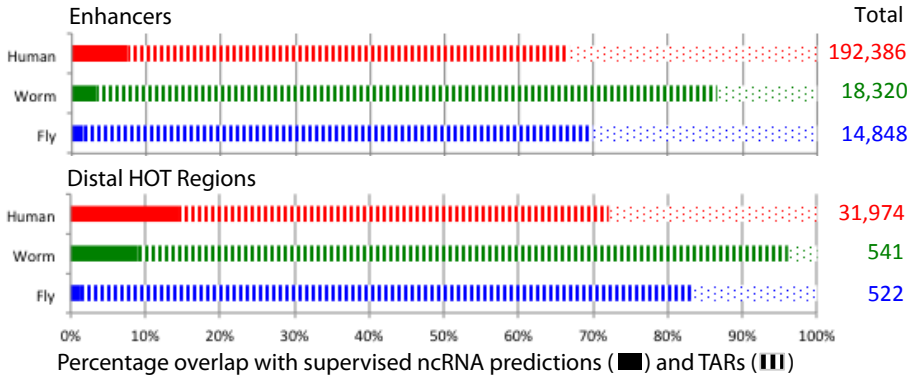
TAR Characterization

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

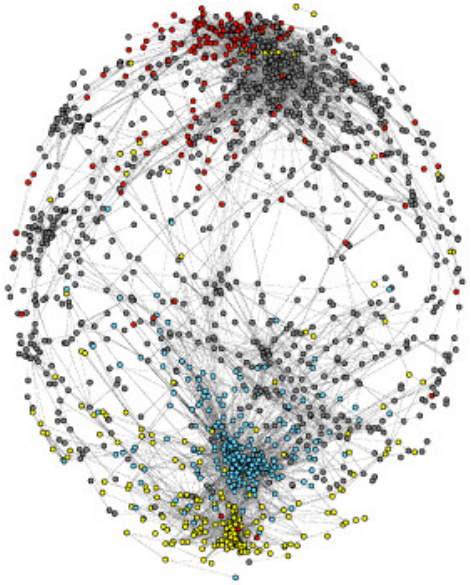
Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

Gene coexpression clustering

Gene expression can vary over a range of features such as time, tissue, etc.

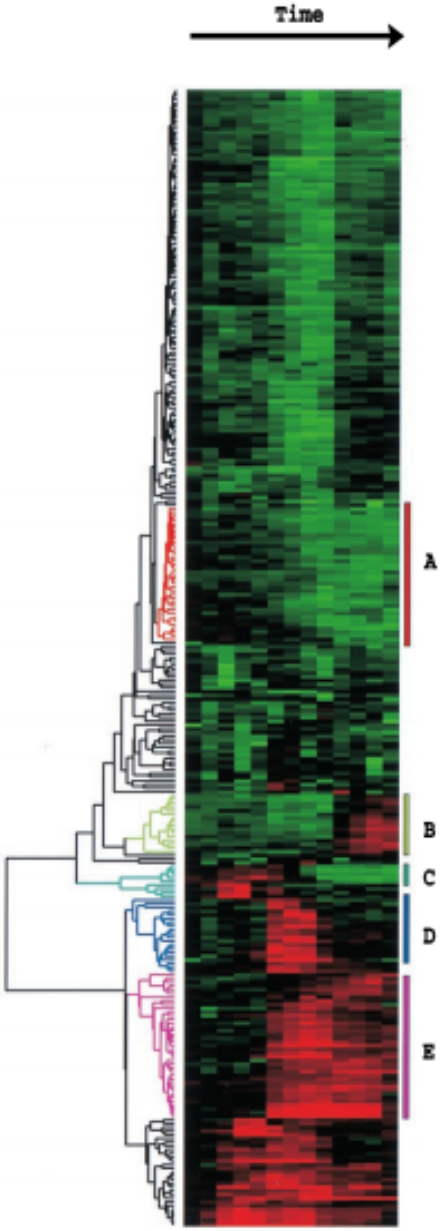
Clustering genes based on their expression profiles can help reveal relationships.



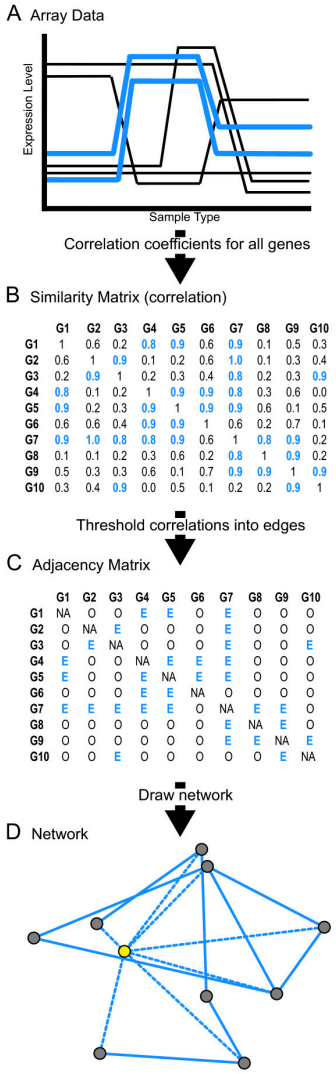
rRNA Proc
Prot Synth
Ubiquitin

Gene co-expression networks can reveal **functional groupings**

M. R. J. Carlson et al., BMC genomics 7, 40 (2006).



M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, PNAS 95, 14863-14868 **(1998).**



Gene co-expression data can also be viewed from a **network perspective**

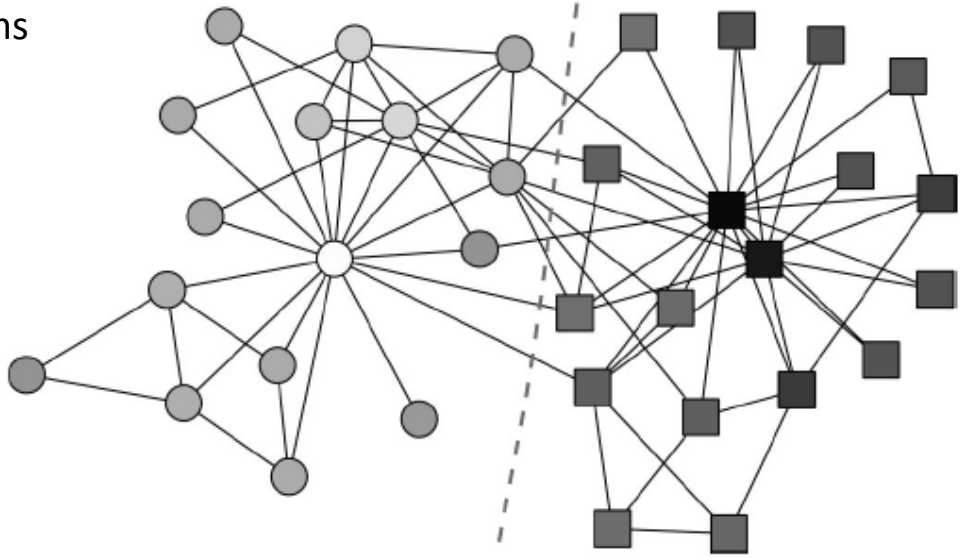
M. R. J. Carlson et al., BMC genomics 7, 40 (2006).

Modular organization in networks

Intuition:
To divide vertices into groups such that connections within groups are relatively dense while those between groups are sparse.

Method:
Many. A common one is to maximize an objective function **(modularity function)** over all possible divisions of a network.

Zachary karate club network



Newman PNAS 2006

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

sum over nodes within a group (module)

m: total number of edges

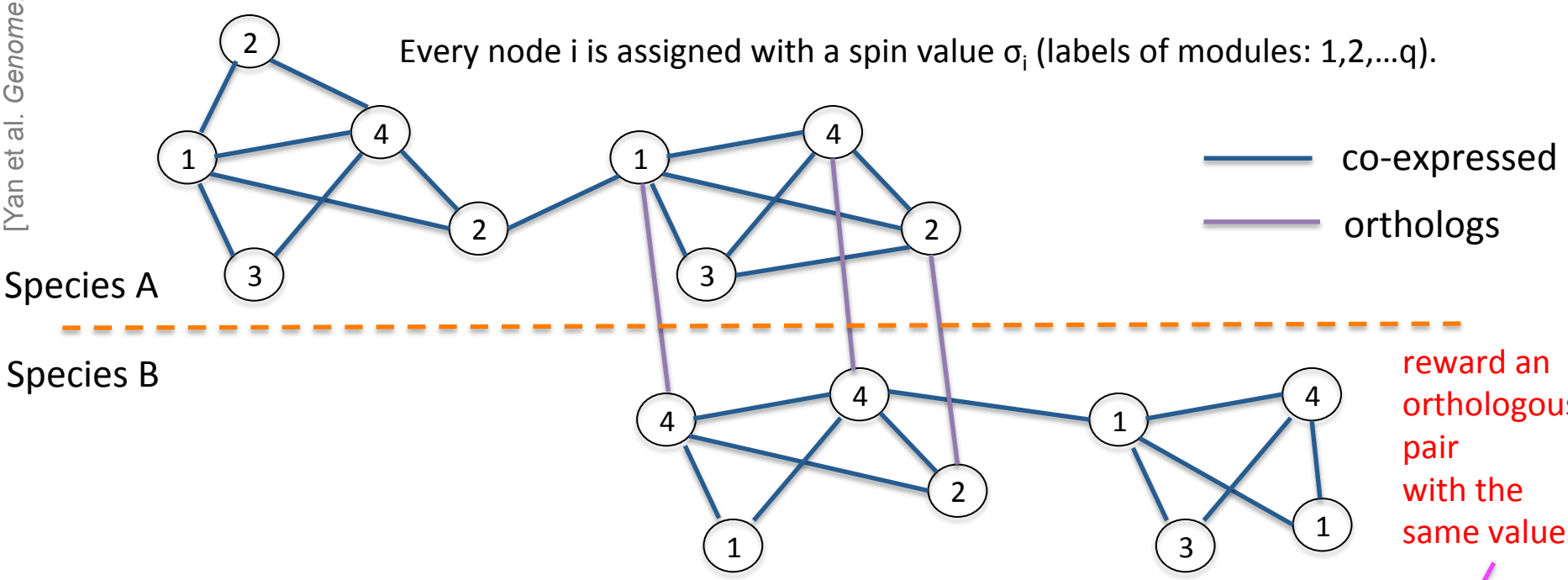
negative contribution of non-links
(if two nodes are not connected, they should not be in the same group)

$\frac{k_i k_j}{2m} = p_{ij}$ =expected number of edges between i and j in a null model

positive contribution of links
(if two nodes are connected, they should be in the same group)

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

A toy example [orthoclust]



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

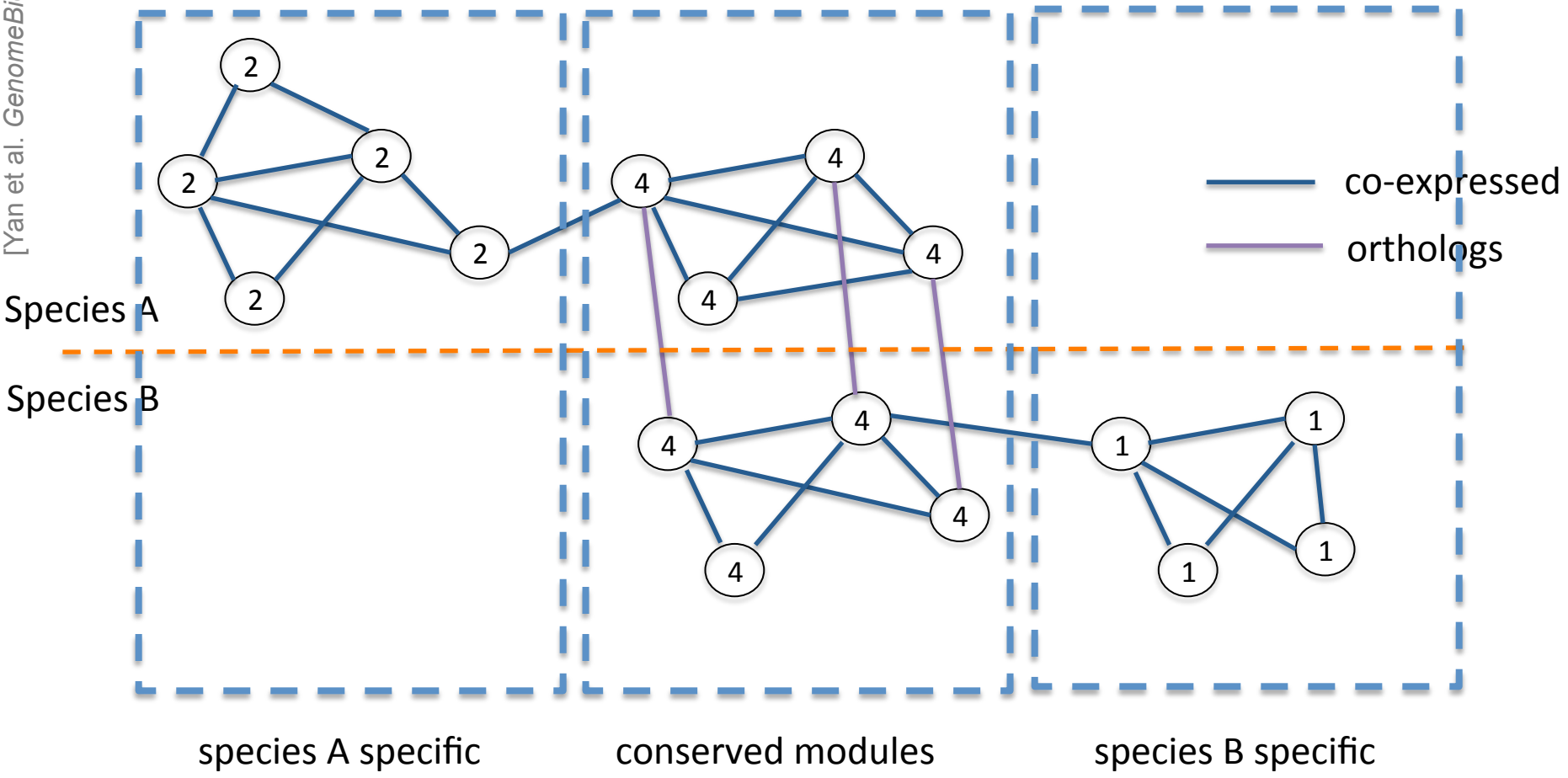
reward a co-expressed pair with the same value

punish a non co-expressed pair with the same value

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

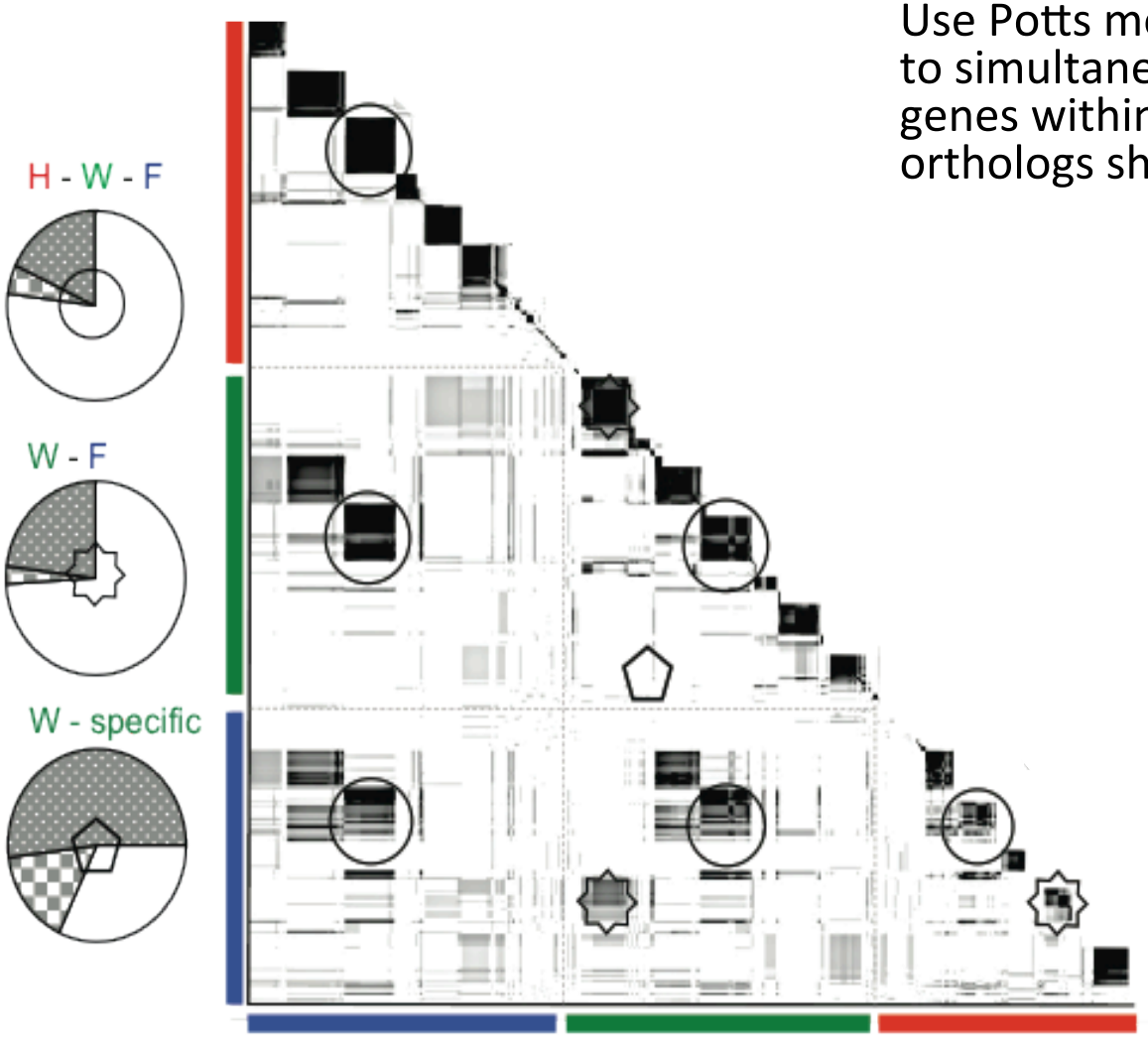
[Yan et al. *GenomeBiol* 15:R100 ('14)]

A toy example [orthoclust]



The ground state configuration correspond to three modules: 1, 2, 4.

Cross-Species Co-expression Clustering



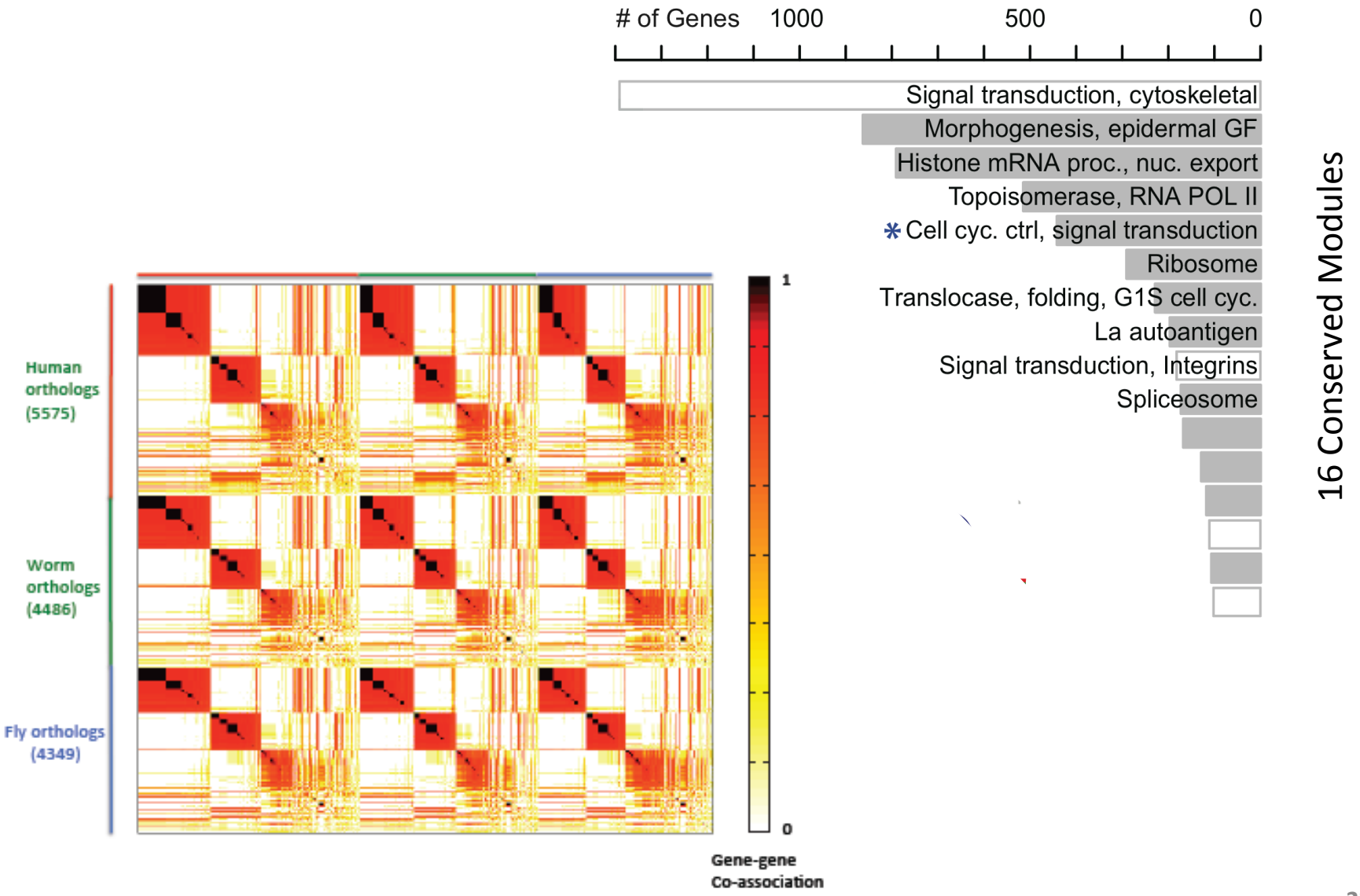
Use Potts model (generalized Ising model) to simultaneously cluster co-expressed genes within an organism as well as orthologs shared between organisms.

Conservation of modules across # of species

□ >30 ▣ 2-30 ▩ 1

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

16 Conserved Modules

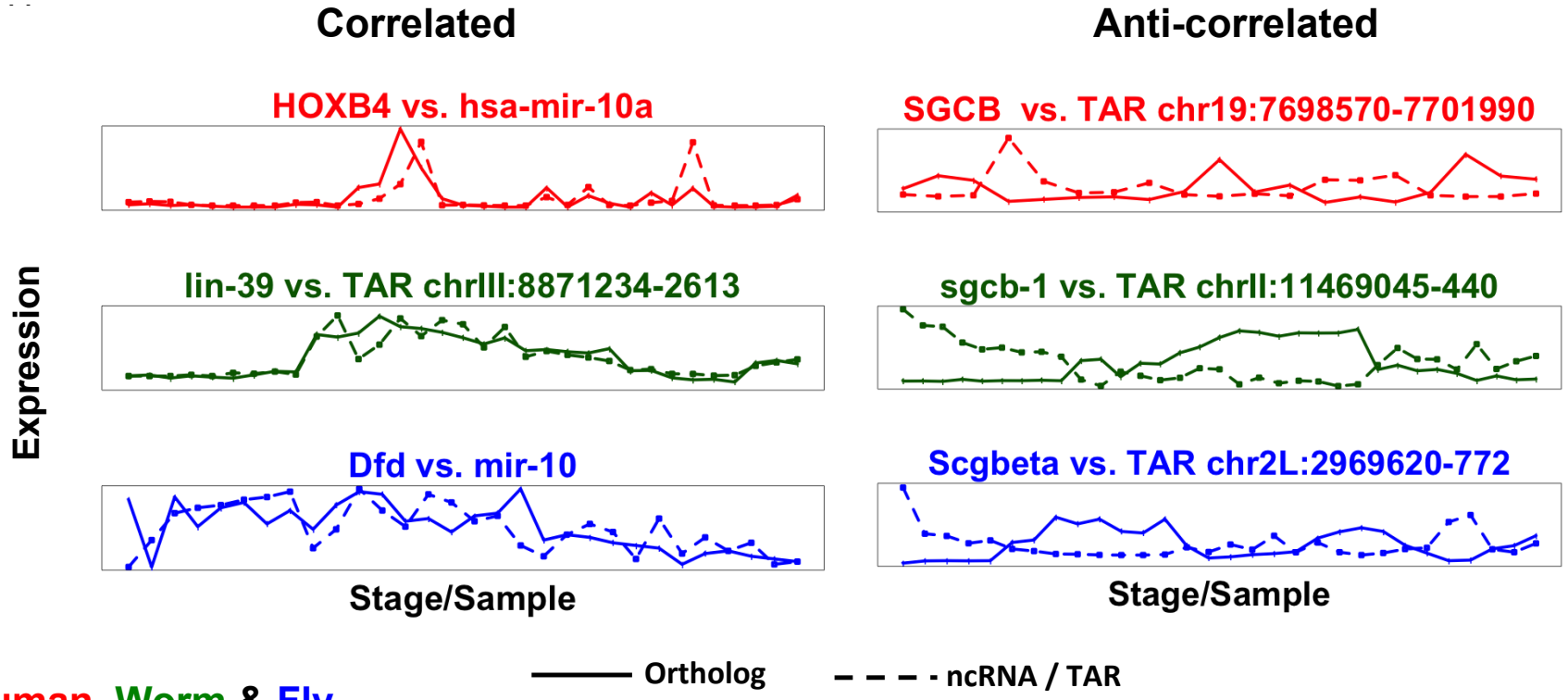


[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

ncRNAs associated with modules

Non-canonical transcription (TARs):

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

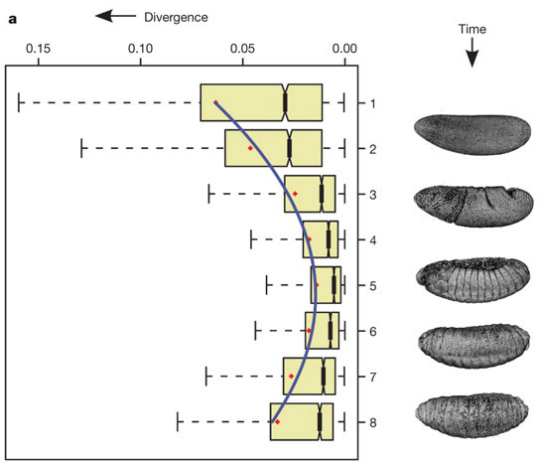


Human, **Worm** & **Fly**

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

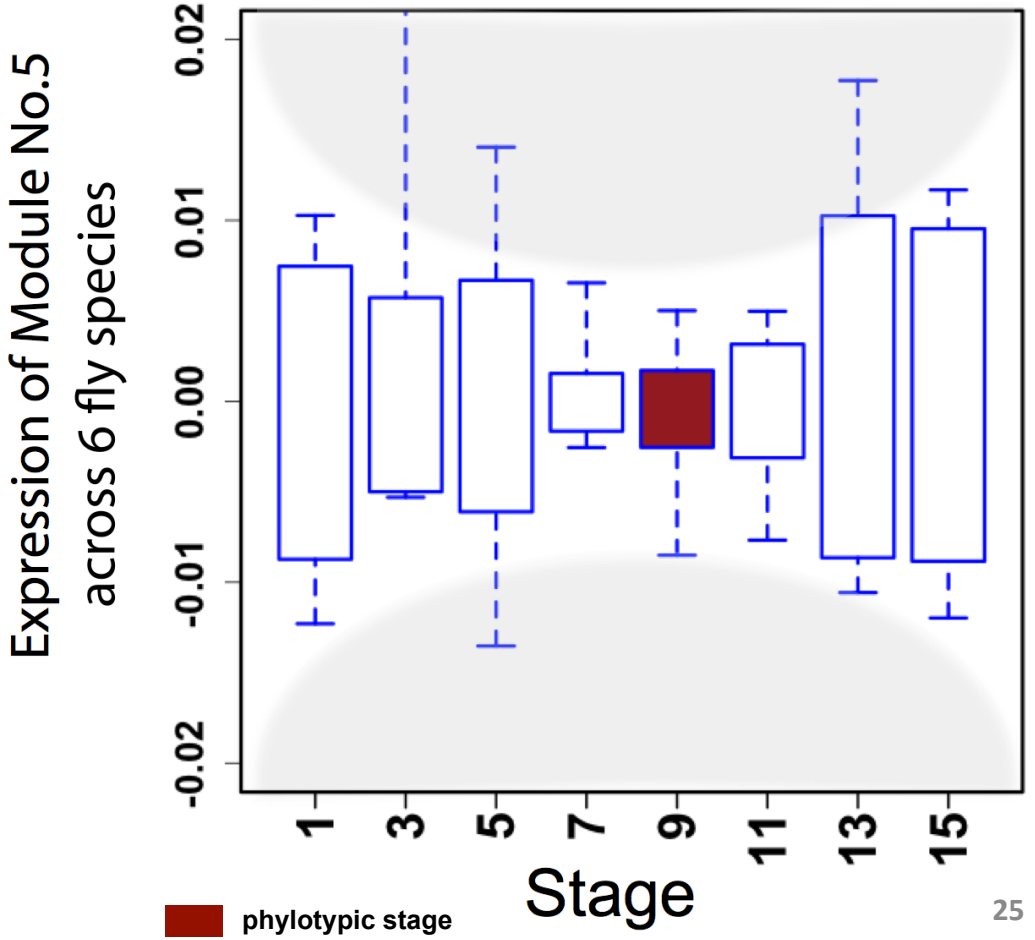
Hourglass Behavior

Temporal expression divergence is minimized during the phylotypic period.
[AT Kalinka et al.
Nature 468, 811-814 (2010)
doi:10.1038/nature09634]



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.



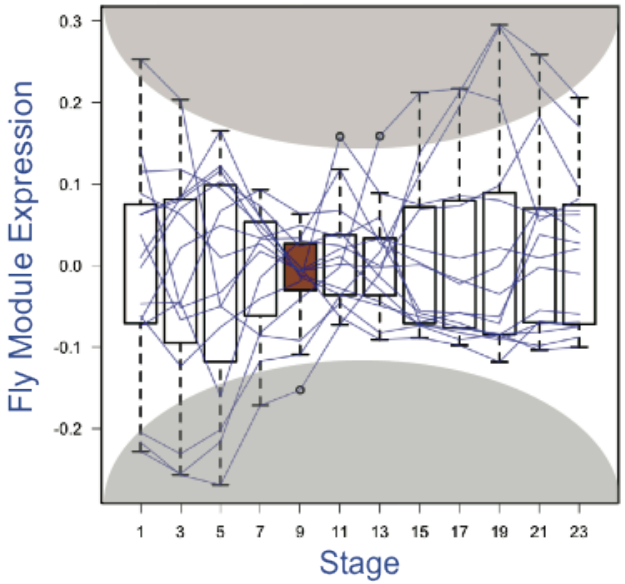
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

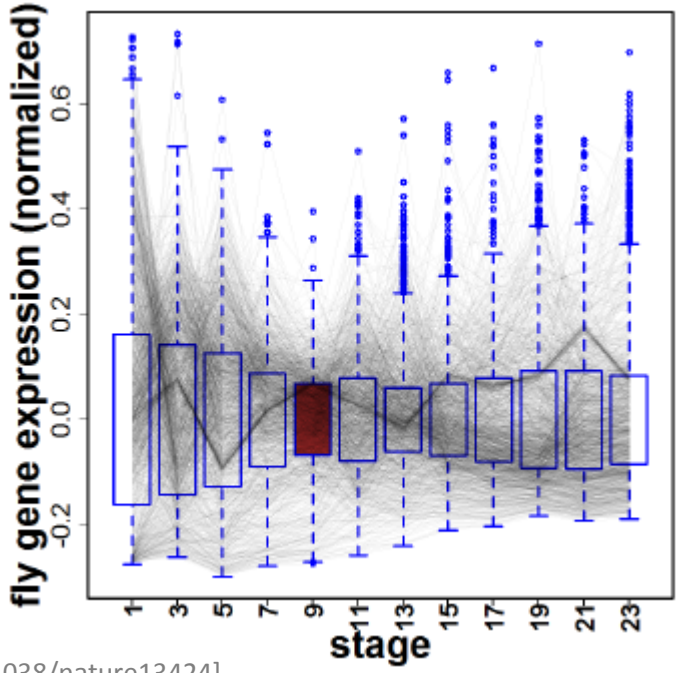
Hourglass Behavior

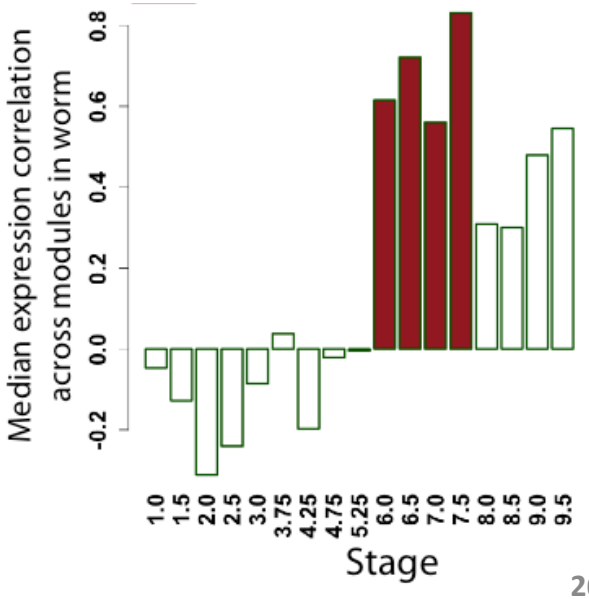
phylotypic stage

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).

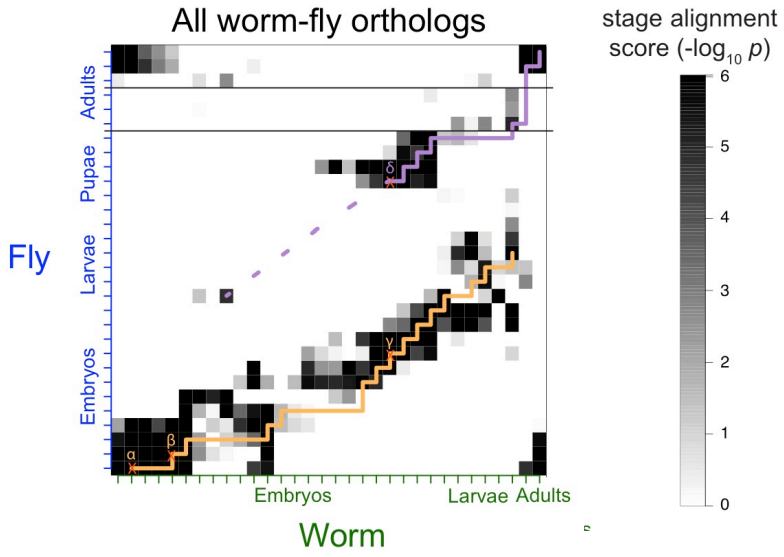






[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

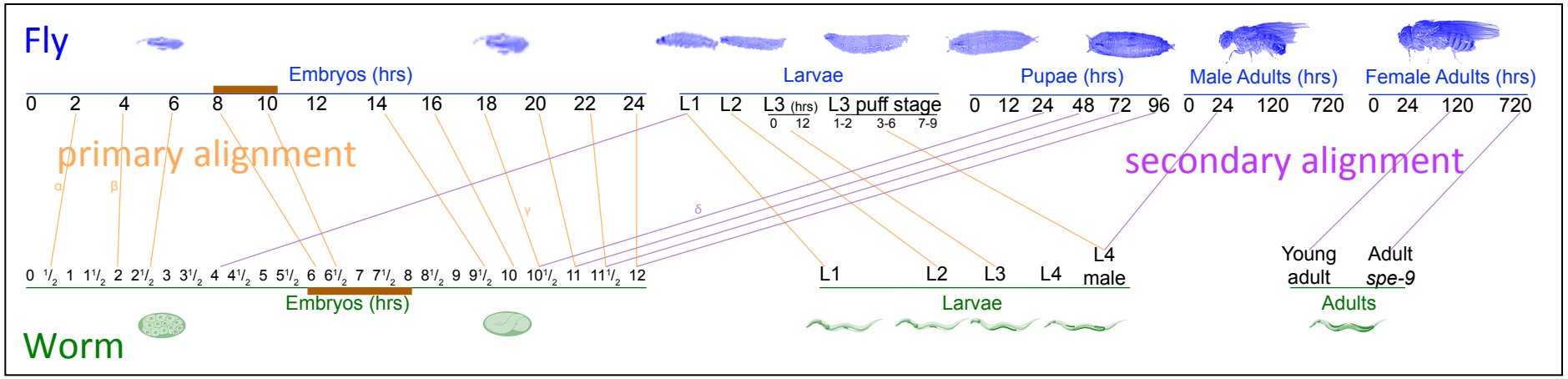
Alignment of Developmental Time-Course



For worm & fly find stage-specific genes

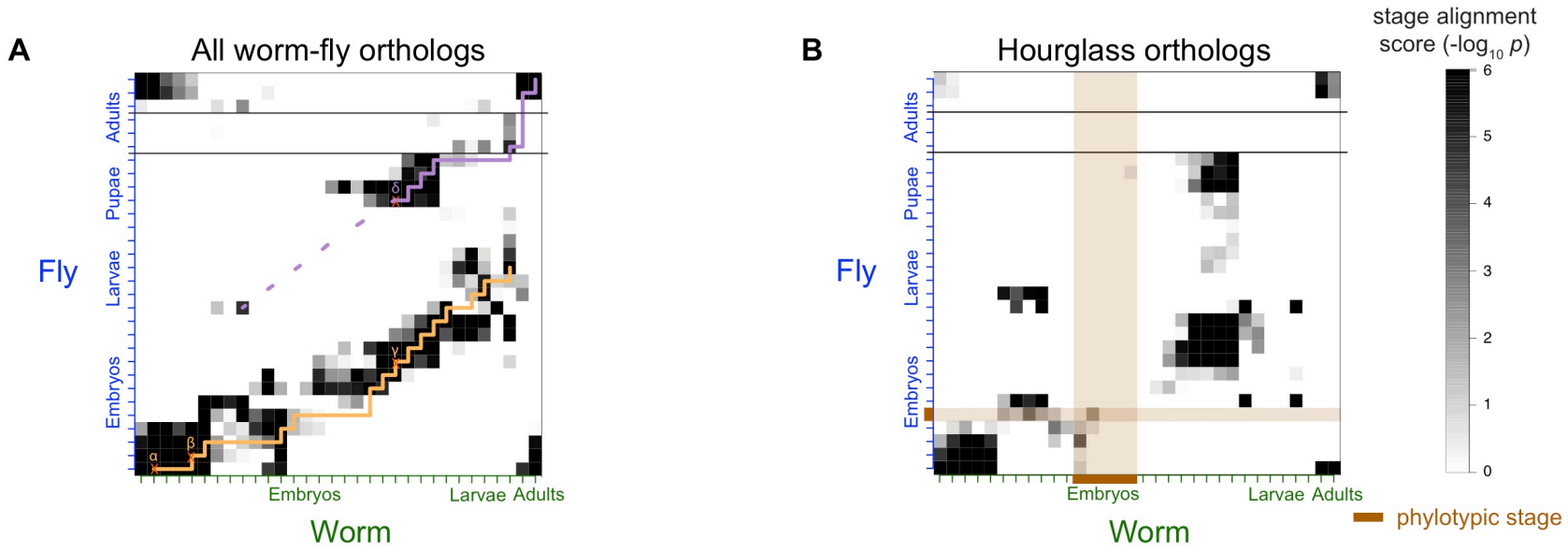
We can align developmental stages using fraction of shared orthologs between worm and fly amongst these

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

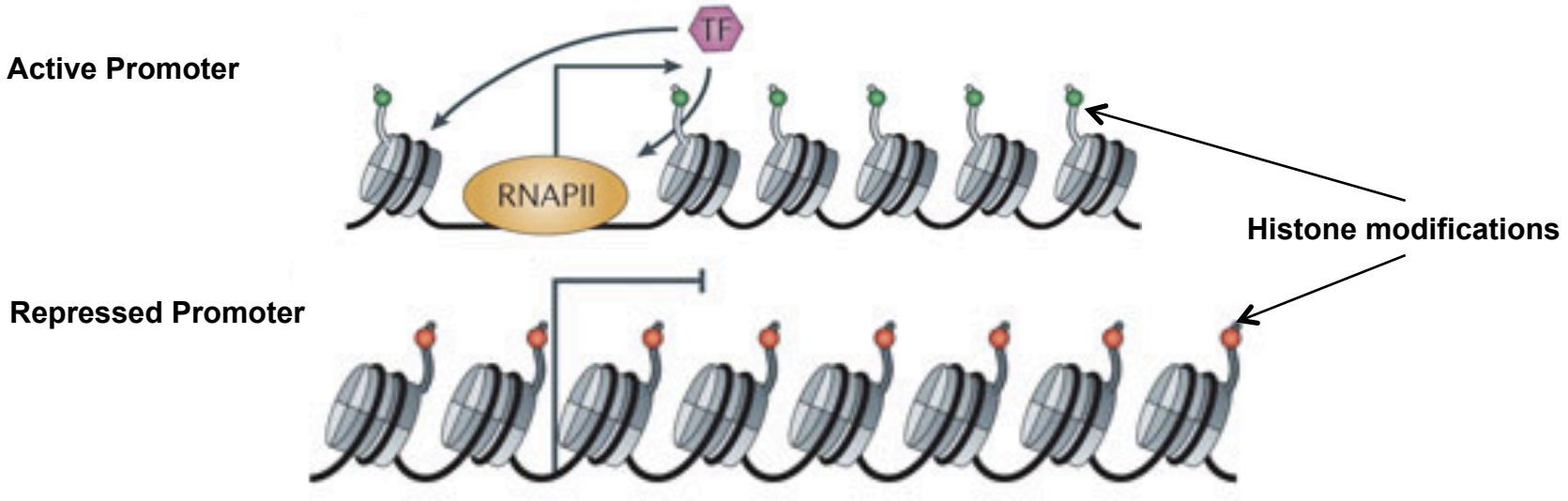
- By definition genes in hourglass modules are not phylotypic stage specific, hence the gap

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

Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

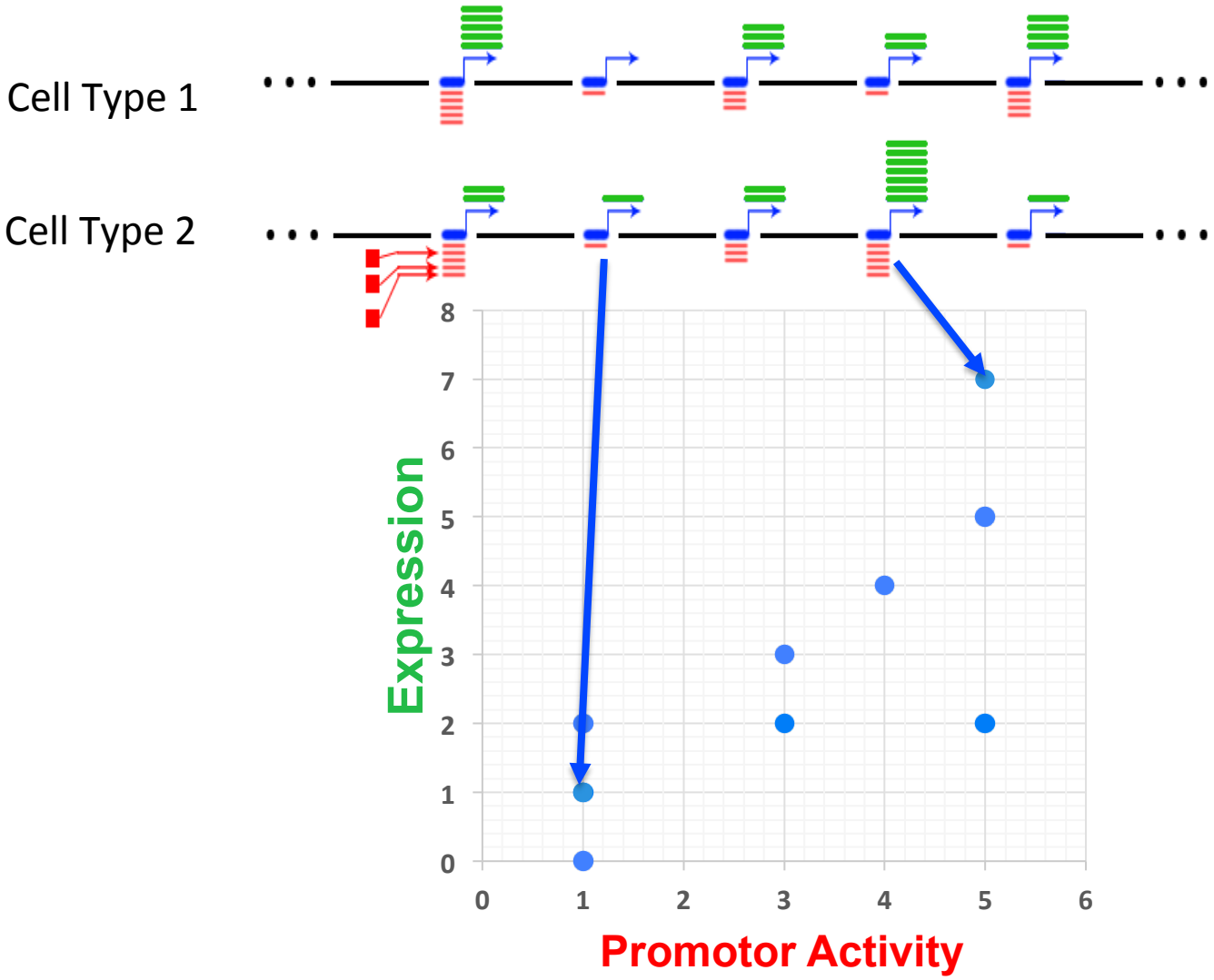
Focus on Promoters

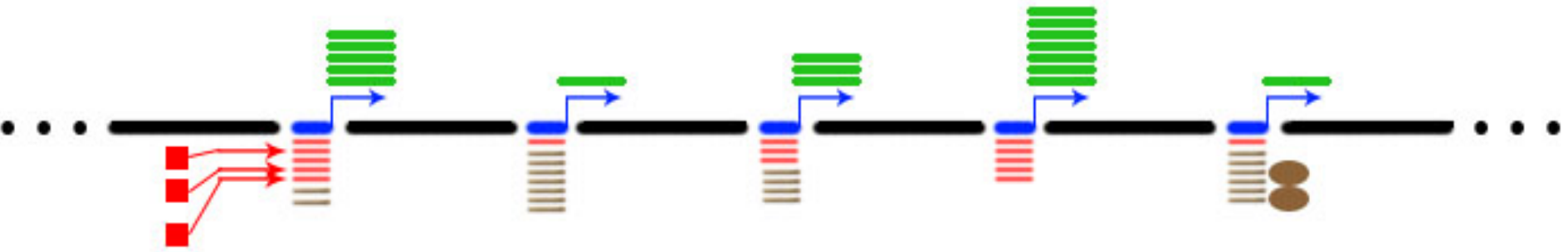


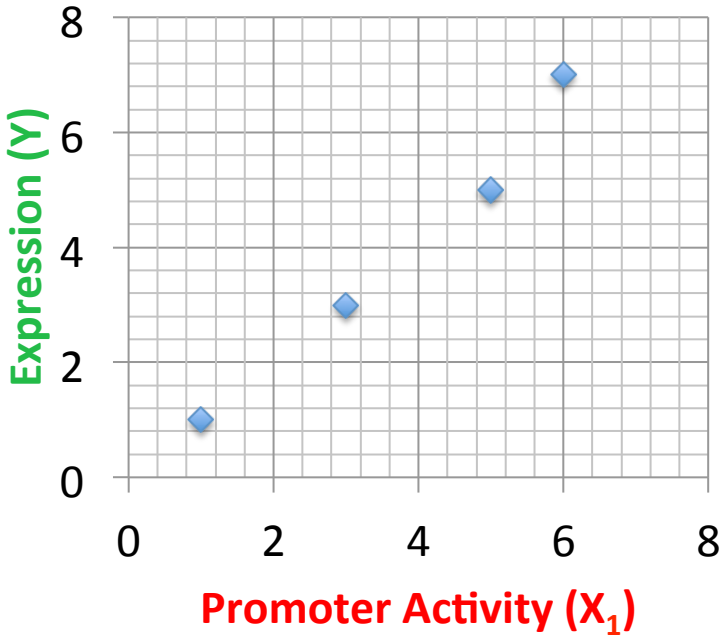
- Key Questions
  - How do we define the active regions of promoter?
  - For an active promoter, how do we relate it bound TFs, its epigenetic marks & its chromatin state to the level of transcription?
  - Are these definitions & relationships conserved between very different species?

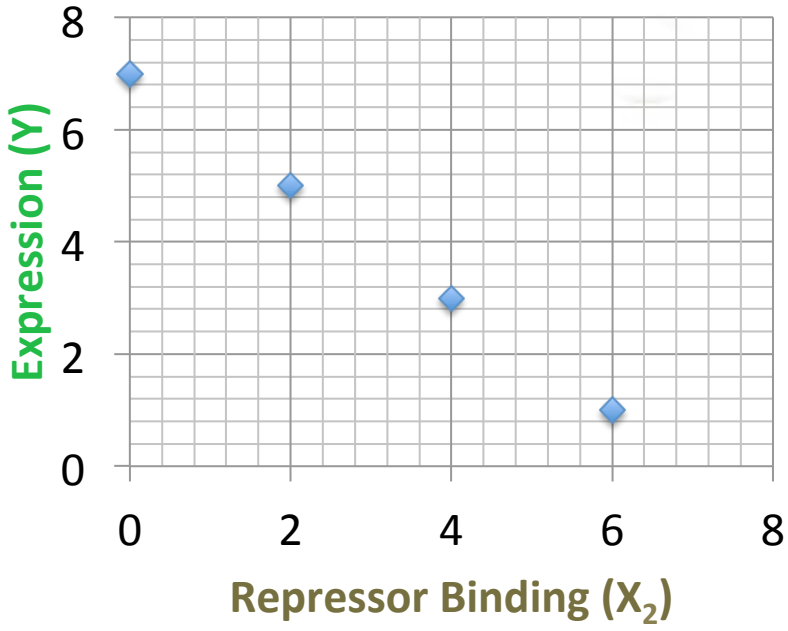
Zhou *et al. Nat. Rev. Genetics* 2011

Relating Genomic Inputs to Outputs



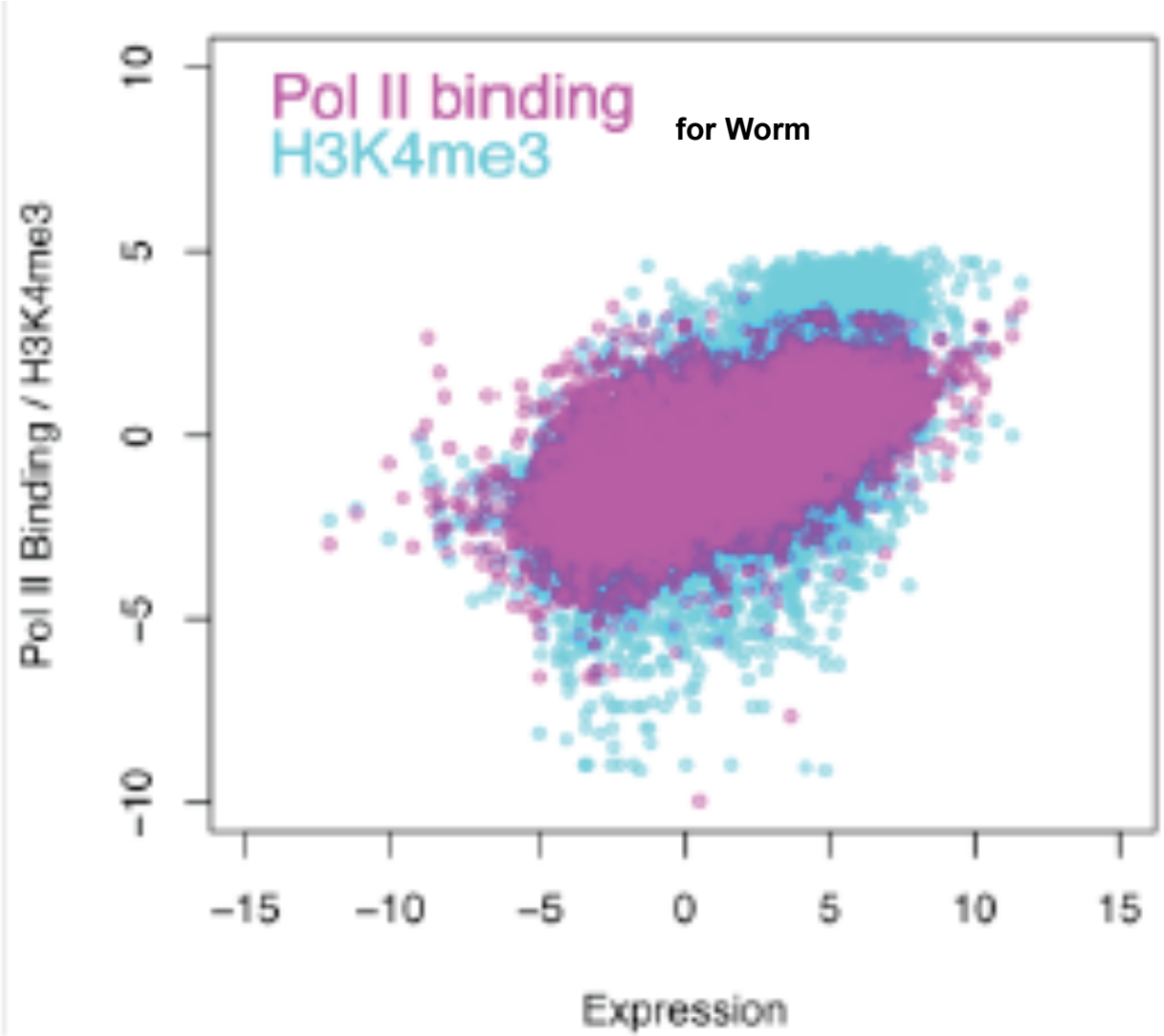



Expression (Y)
0
2
4
6
8
Promoter Activity (X_1)


Expression (Y)
0
2
4
6
8
Repressor Binding (X_2)

$$Y = aX_1 + bX_2 + c$$

Inputs v Outputs:
Upstream Binding/Modification v Expression

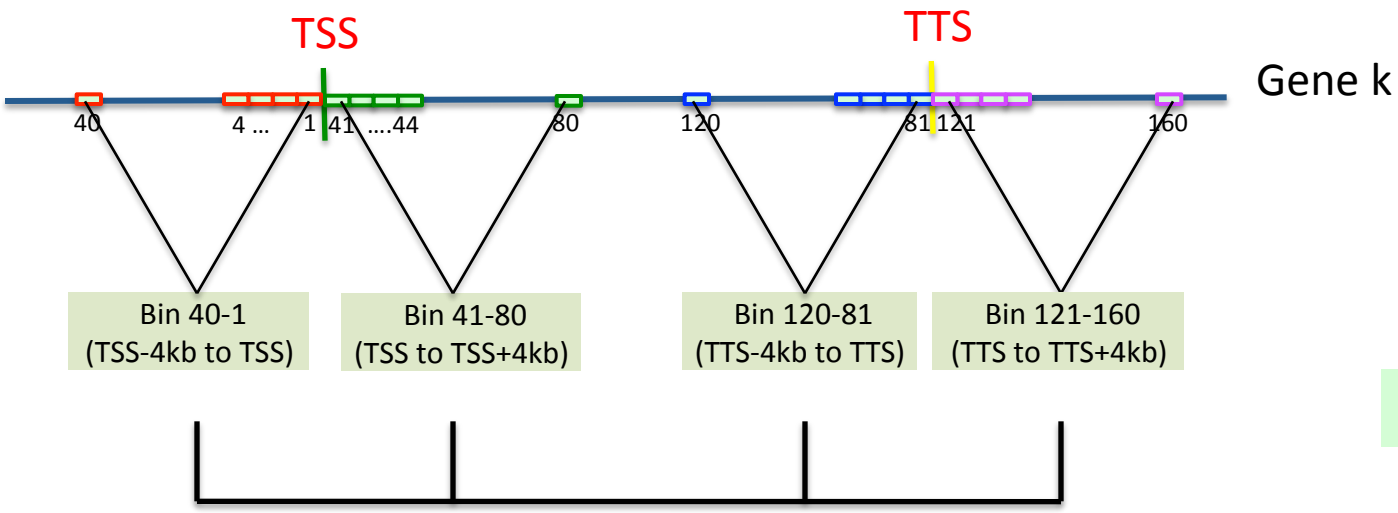


PCC: Pol II,
0.33;
H3K4me3,
0.28

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

Histone Modification (HM) model

[Cheng et al. ('11) *Genome Biol.* 12: R15]

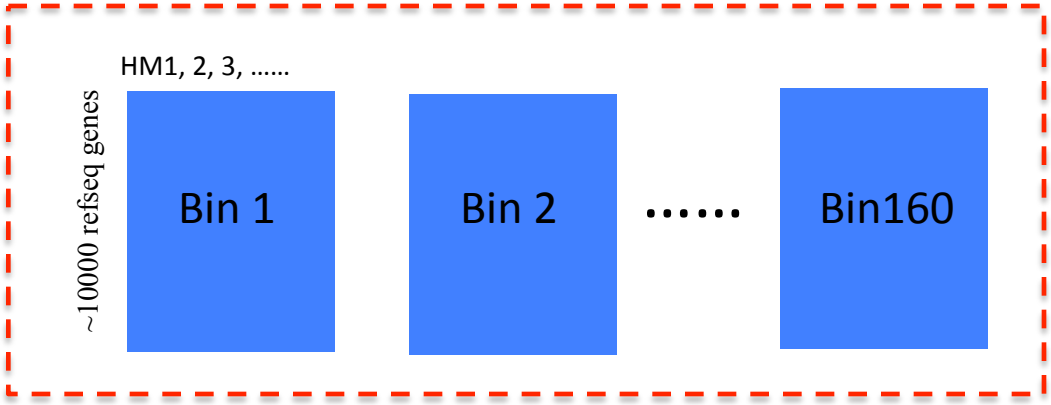


Chromatin features:
Histone modifications

RNA-Seq data

Prediction target:
Gene expression level

Predictors



His. mods around TSS & TTS are clearly related to level of gene expression, in a position-dependent fashion

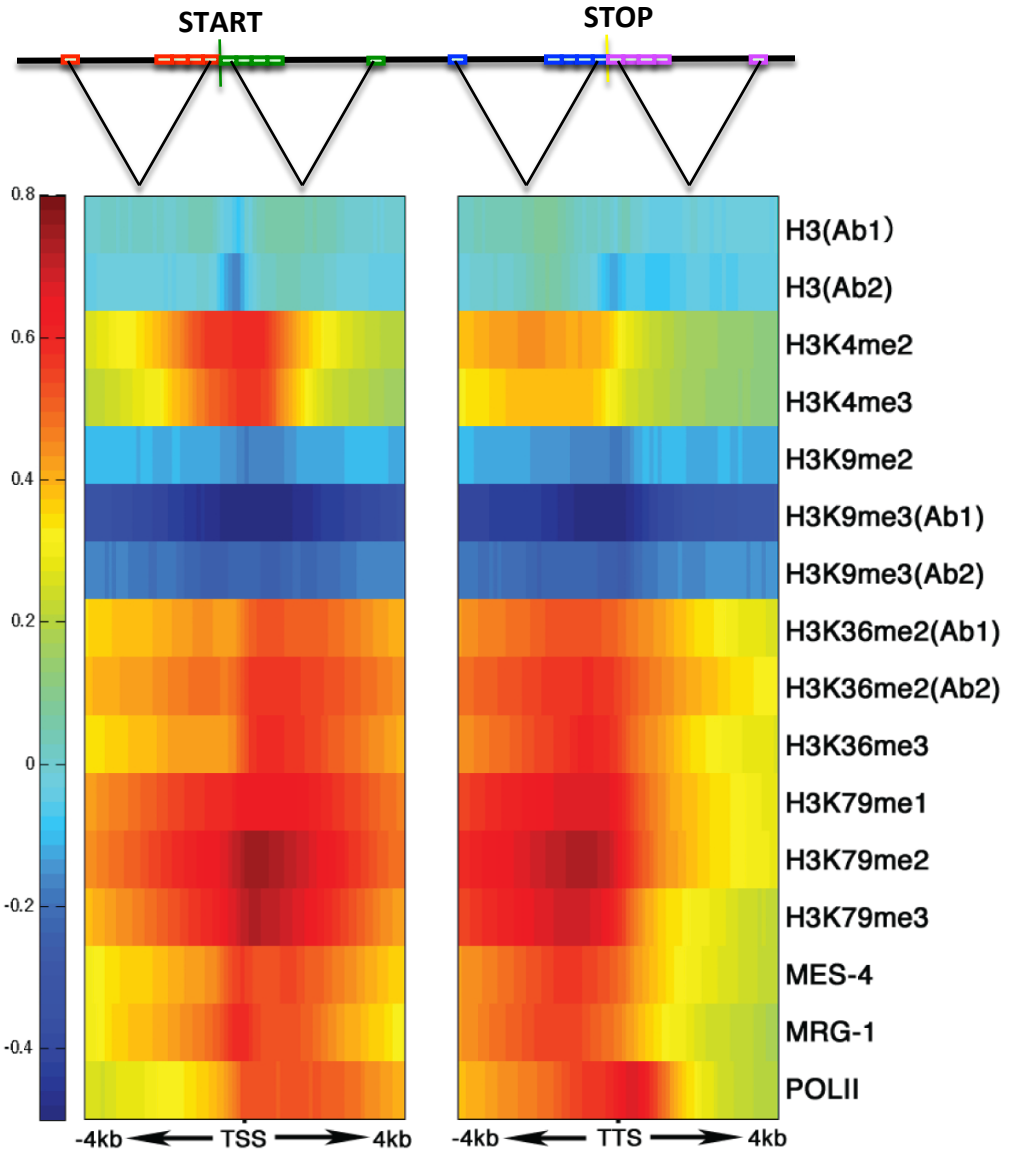
Early work in '09/'10

Science 330:6012 [here]

Also:

Ouyang, Zhou, Wong (’09) *PNAS*;

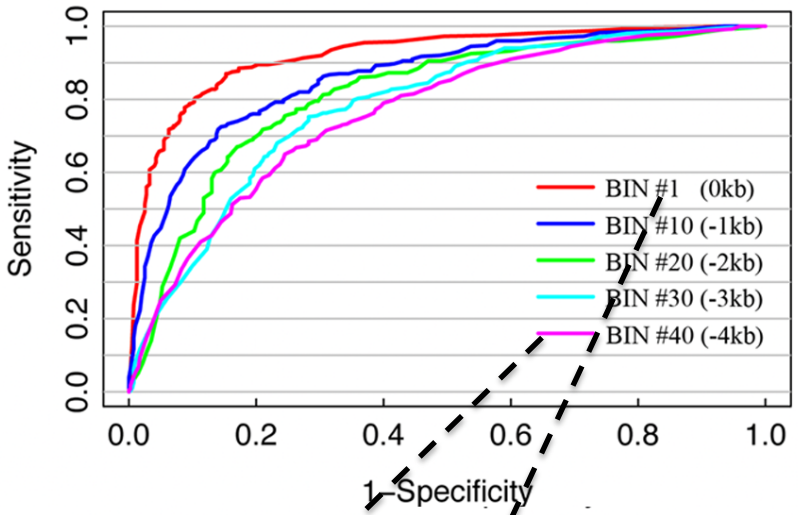
Karlic et al. & Vingron (‘10) *PNAS*



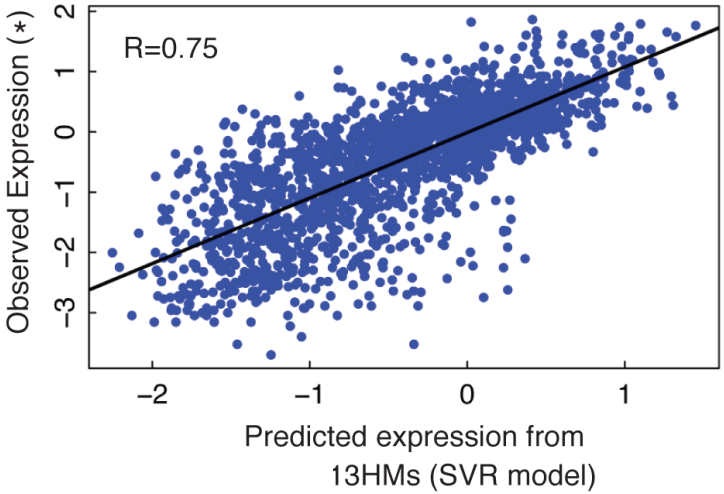
Integrate all histone modifications to predict gene expression levels

[Cheng et al. ('11) *Genome Biol.* 12: R15]

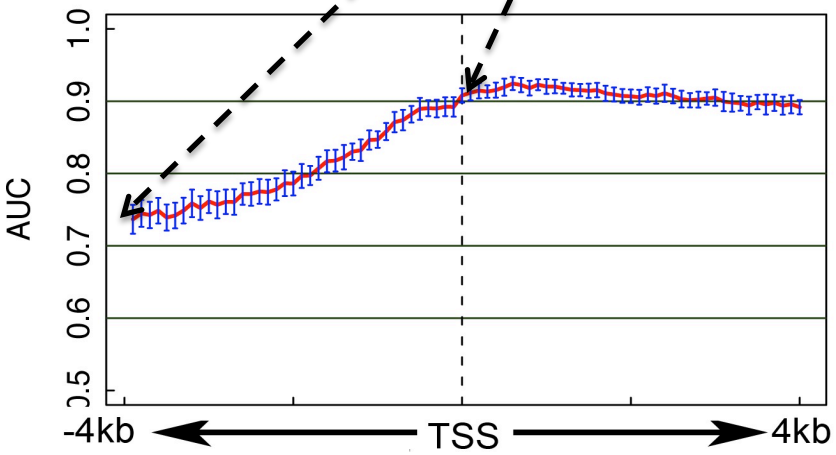
Classify H/L genes (SVM)

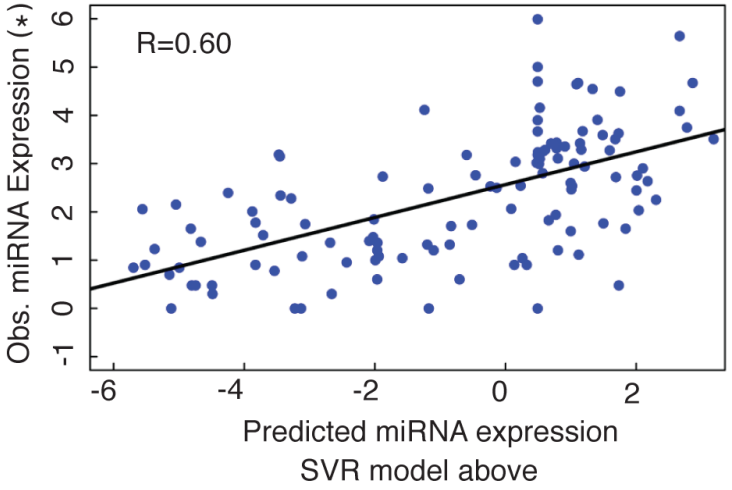


Predict expression values



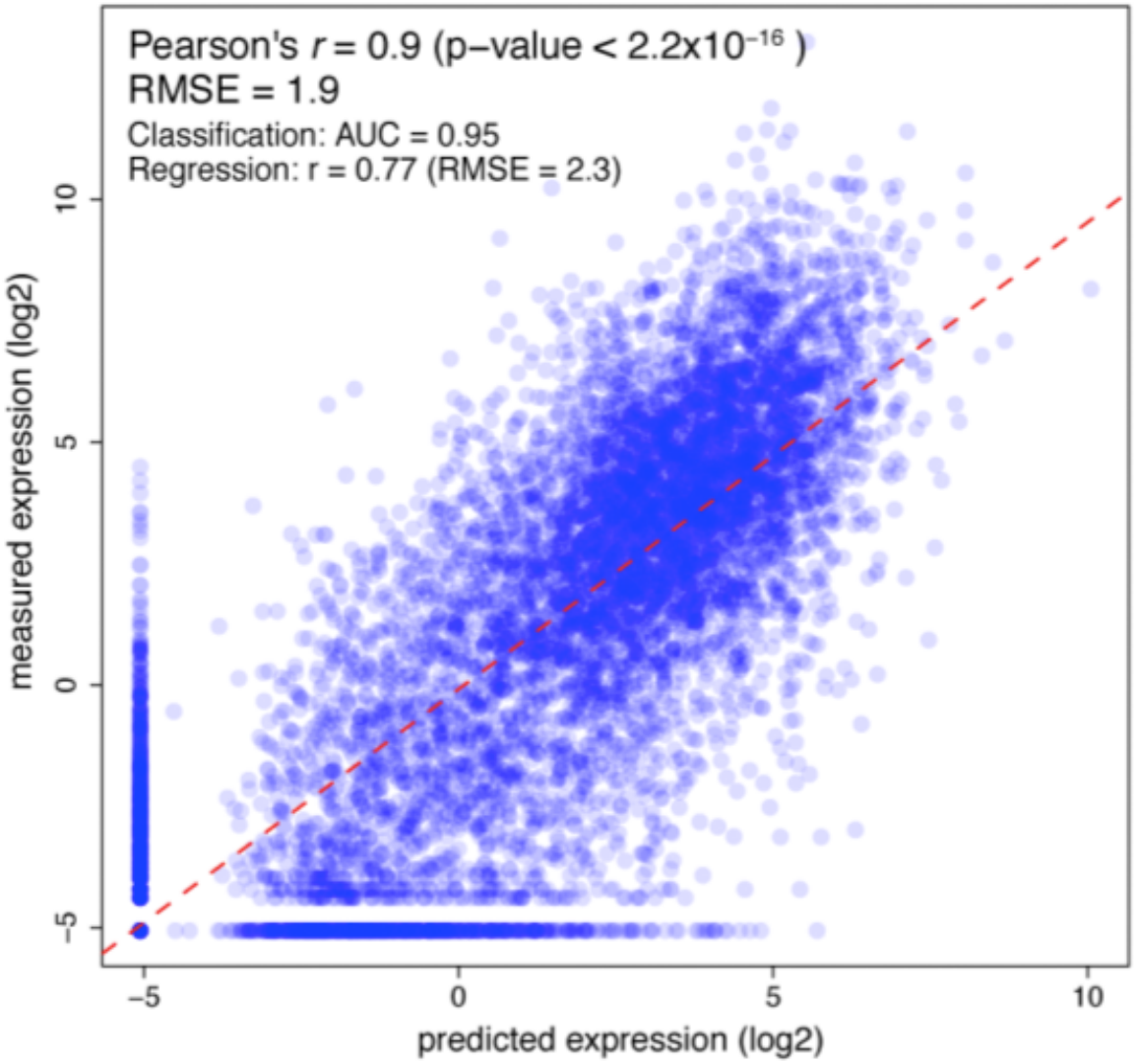
Magnitude of Prediction from a "bin" around the TSS

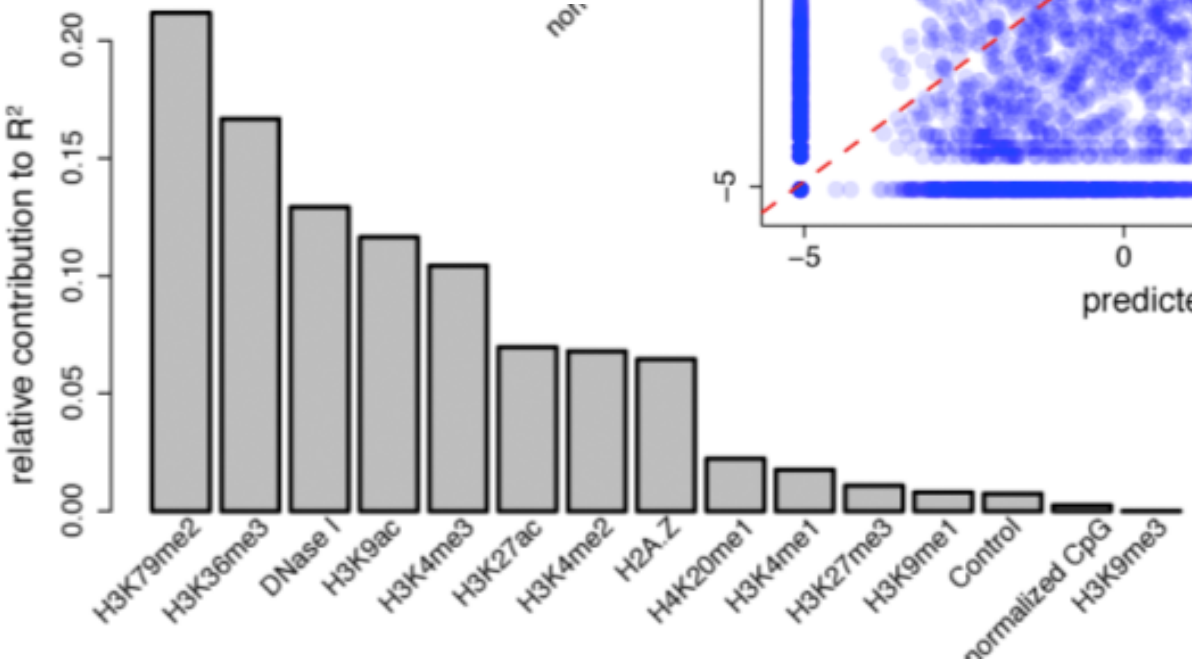




* = LOG_{10}RPKM

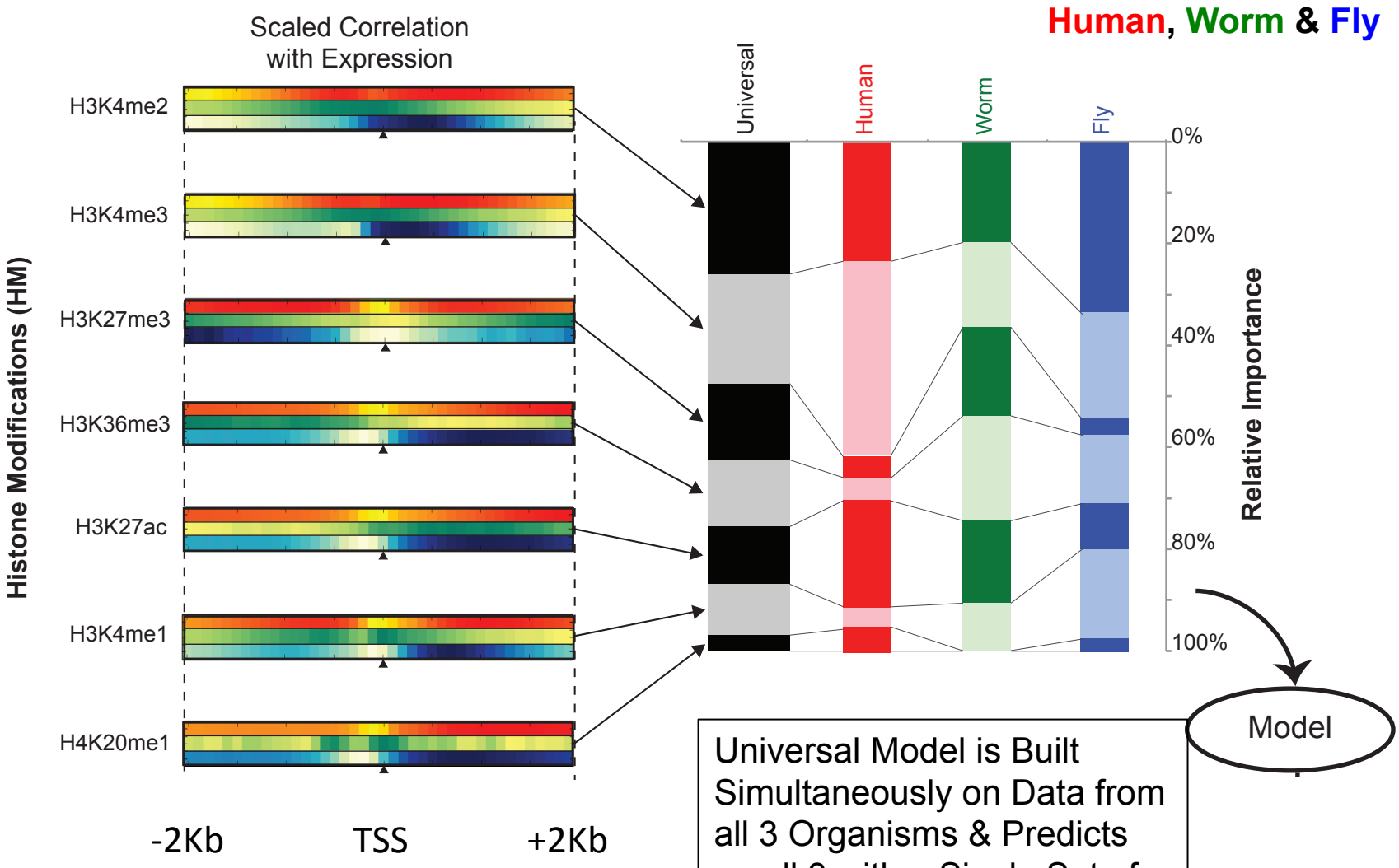
Human ENCODE Results





[Encode Consortium, *Nature* ('12)]

Comparison of Models for Gene Expression, Building a Universal Model

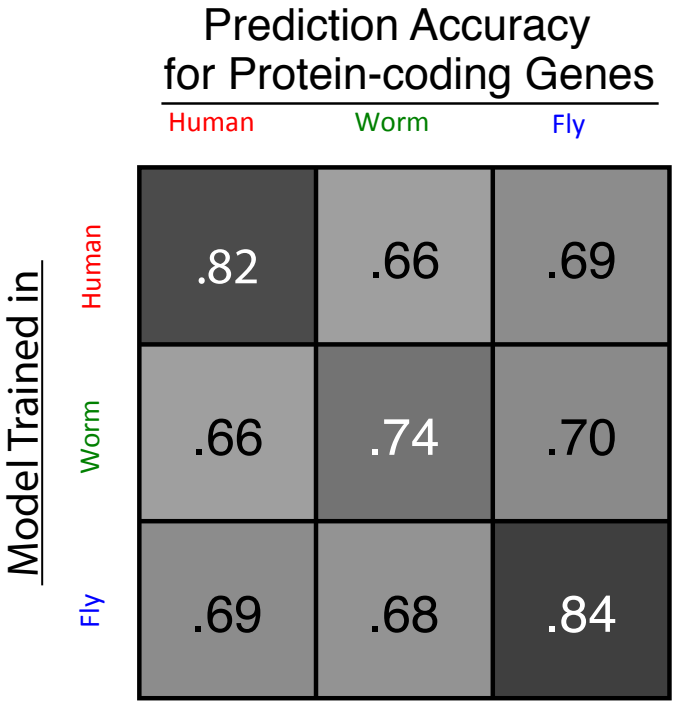


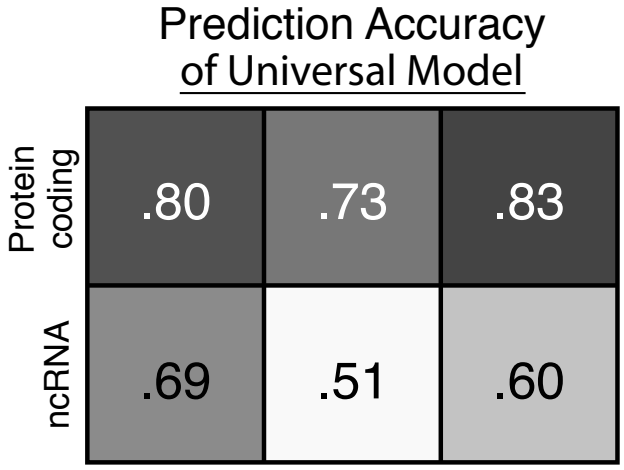
Universal Model is Built Simultaneously on Data from all 3 Organisms & Predicts on all 3 with a Single Set of Parameters

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

Performance of Universal, cross-organism Model

- works almost as well as species specific models
- works for both mRNAs and ncRNAs



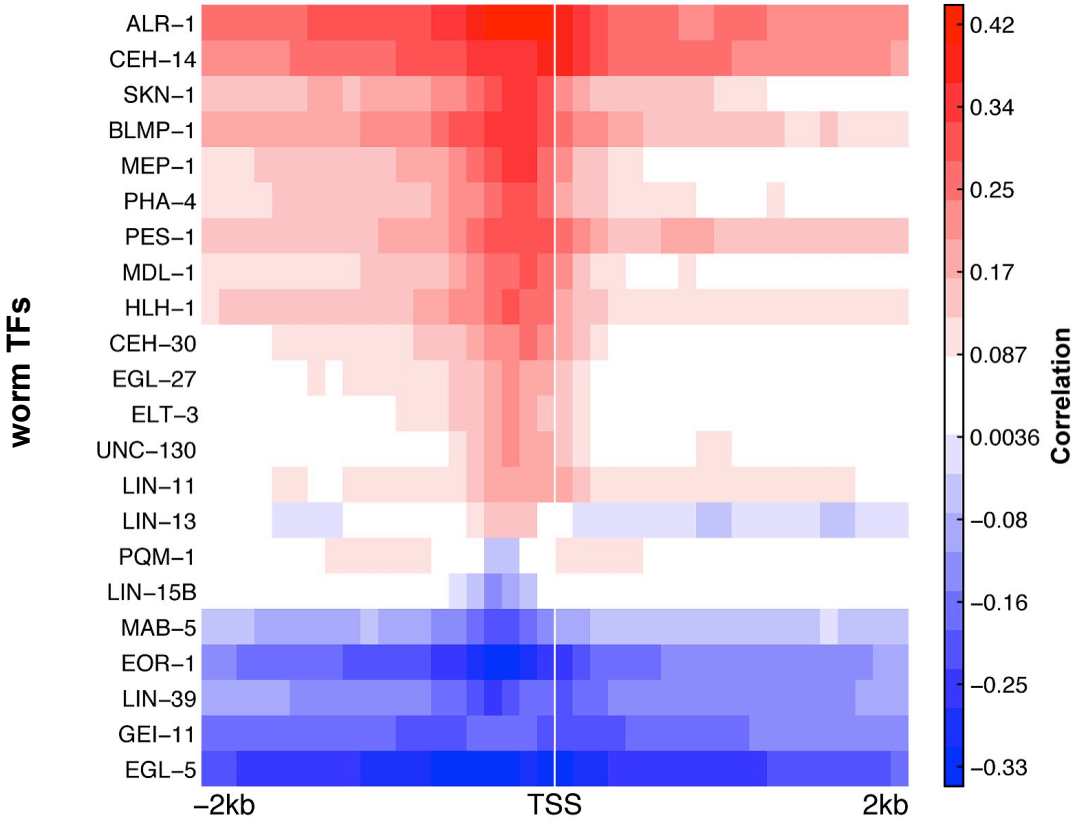


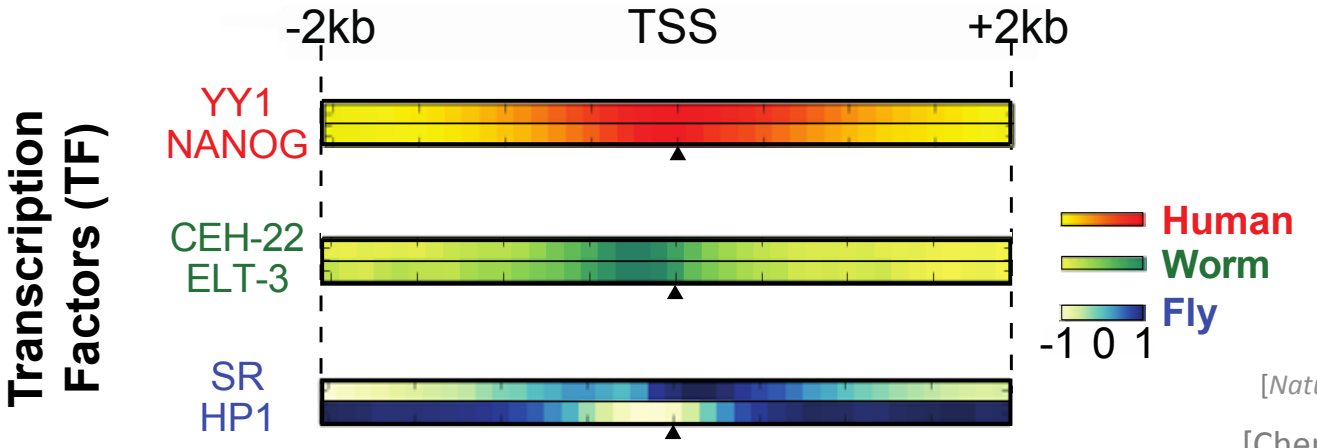
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

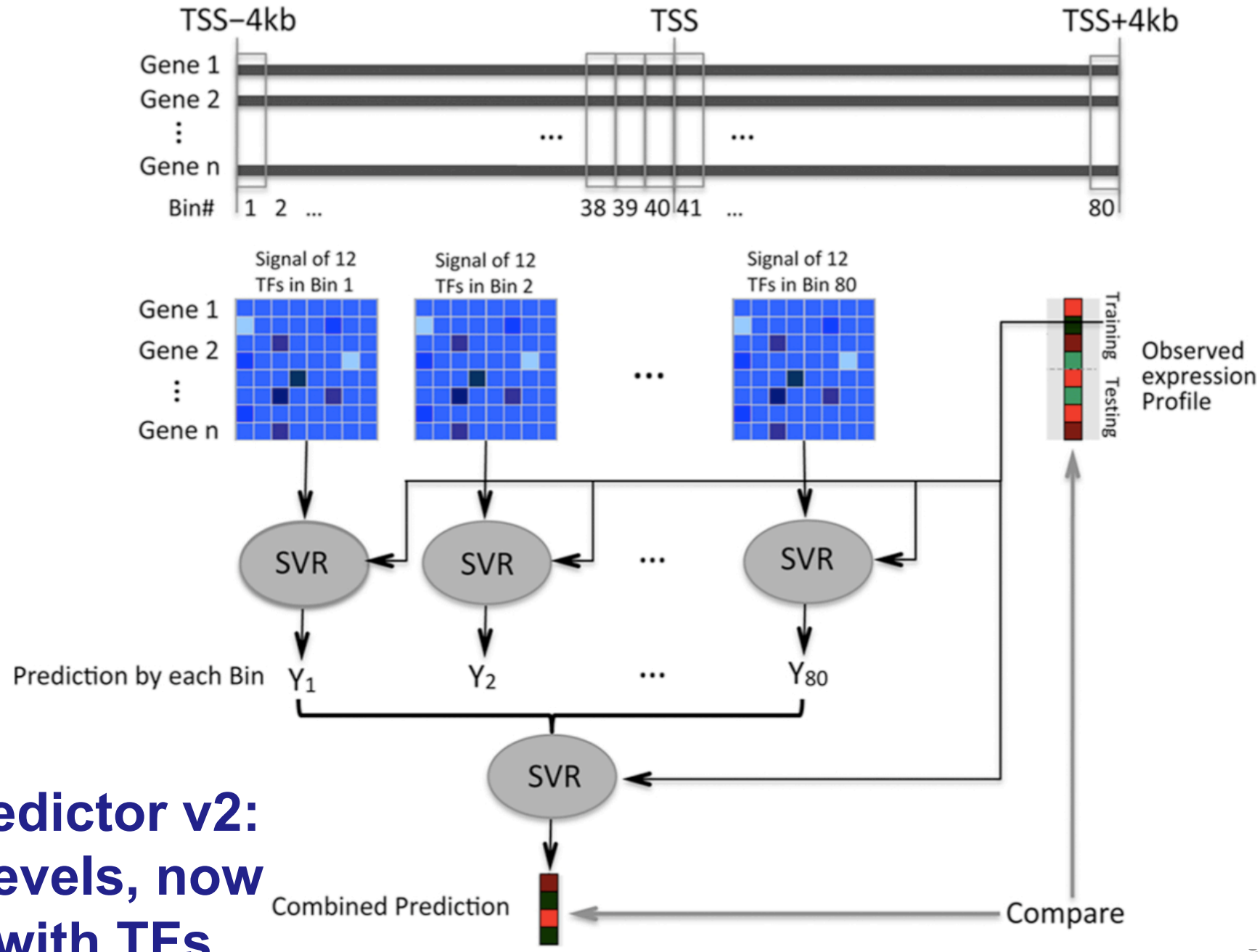
Doing a Model with TFs: Positive and negative regulators from correlating TF signal at TSS with gene expression





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

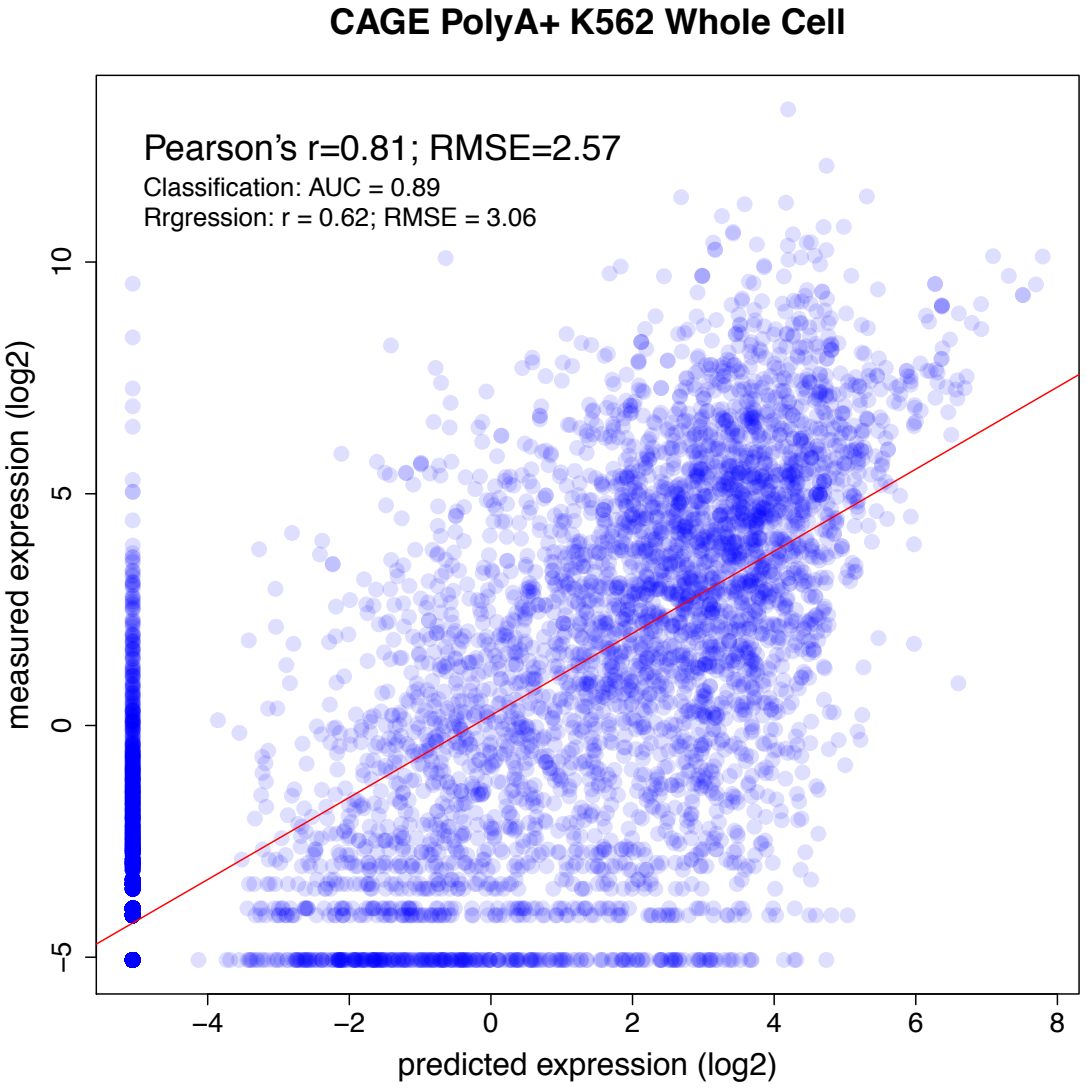
[Cheng et al. ('11) *PLOS CB*]

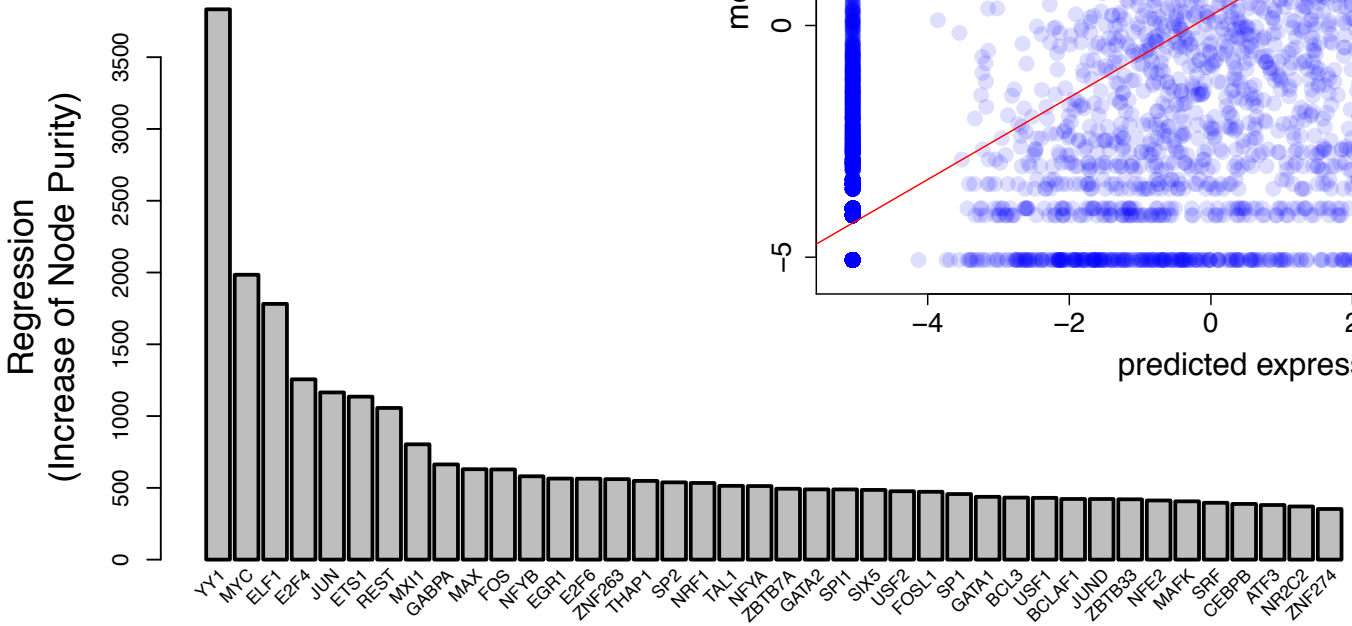


Predictor v2: 2-levels, now with TFs

[Cheng et al. *NAR* ('11)]

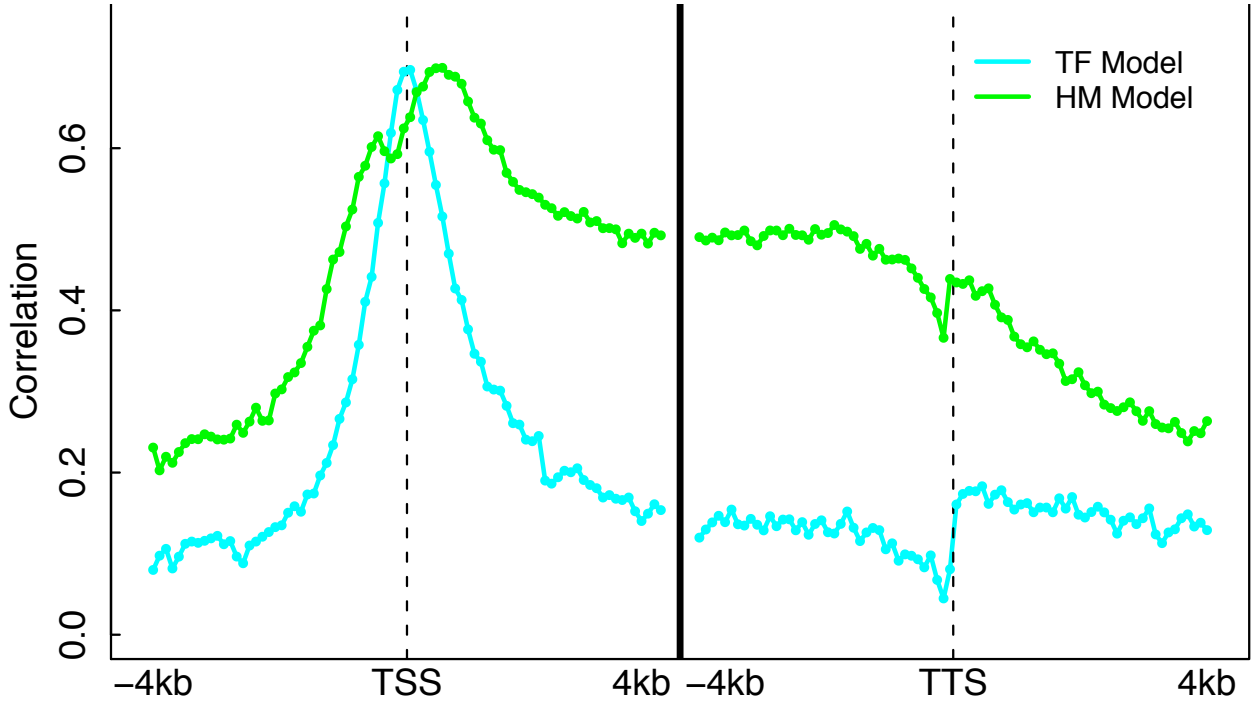
Human Results





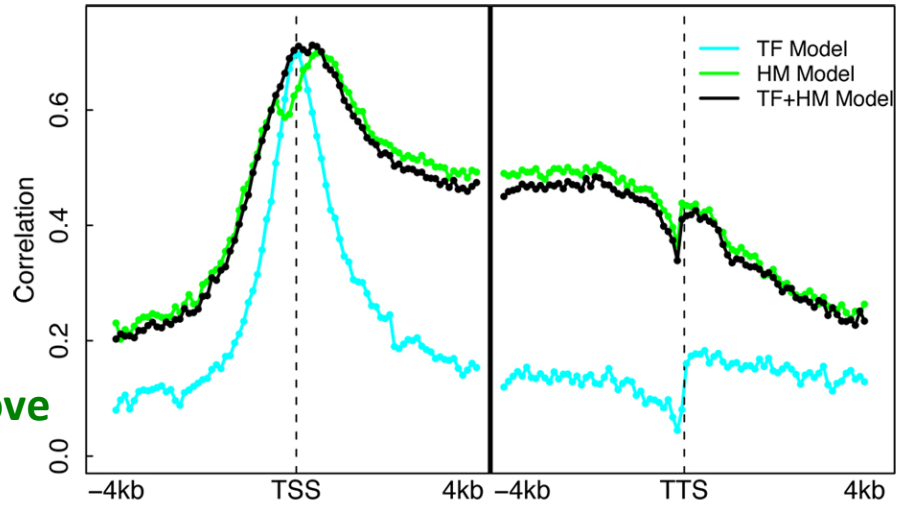
[Encode Consortium, *Nature* ('12)]

Models Illuminates Different Regions of Influence for TFs vs HMs



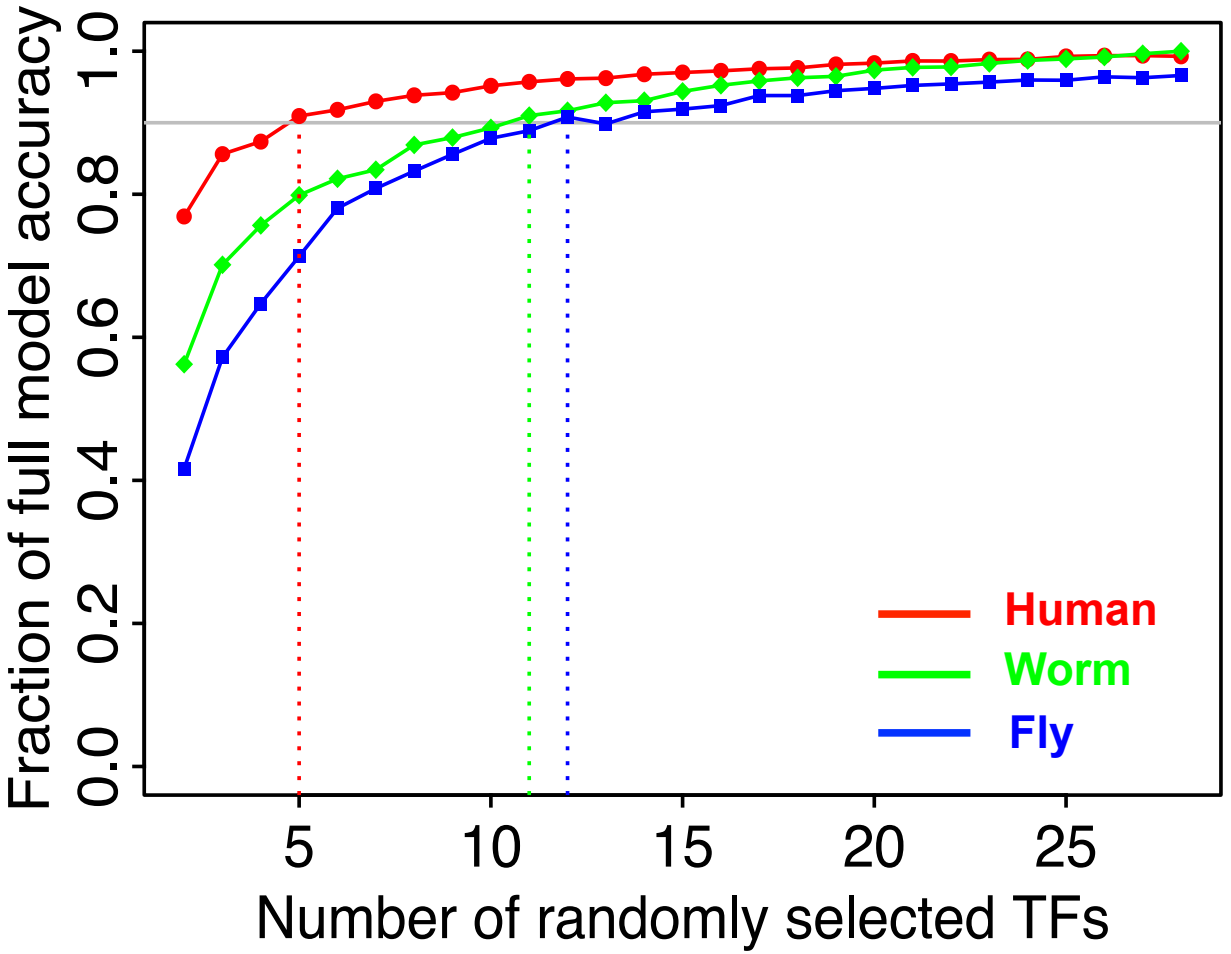
- Datasets
  - ChIP-Seq for 12 TFs (Chen et al. 2008)
  - ChIP-Seq for 7 HMs (Meissner et al.'08; Mikkelsen et al. '07)
  - RNA-Seq (Cloonan et al. 2008)

A TF+HM model that combine TF and HM features does NOT improve accuracy!



Cheng et al. 2011, Nucleic Acids Res.

TF model accuracy only needs a small number of TFs for high accuracy (>90%)



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

Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for 'prediction'
  - Surprisingly, a few TFs are quite predictive

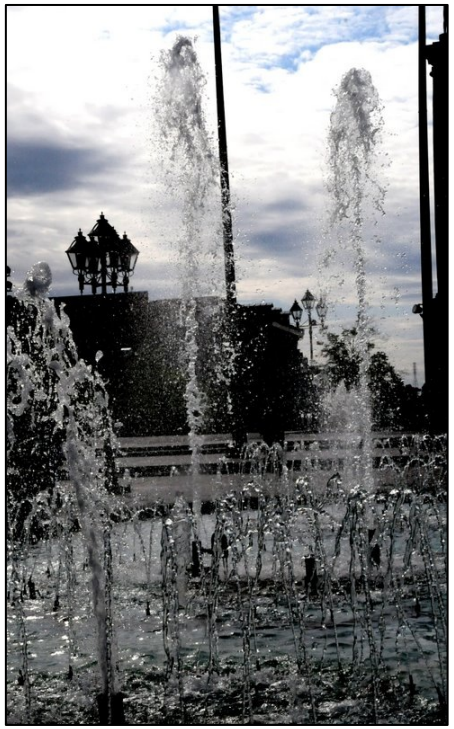
Applications of Machine Learning for Comparing Transcriptomes of Distant Organisms

- **Intro to Comparative ENCODE**
  - Lots of Matched Data for Comparative Analysis
- **App. #1: 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
- **App. #2: 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
- **App. #3: HM Models Relating Gene Expression to Promoter Activity**
  - Works for ncRNAs as well as genes
  - Universal cross-species model uses same set of parameters across diverse phyla
- **App. #4: Similarly constructed TF Models**
  - Variable importance of regions around genes for HMs & TFs
  - TF & HM signals are redundant for ‘prediction'
  - Surprisingly, a few TFs are quite predictive

Acknowledgements

modENCODE/ENCODE Transcriptome subgroup

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 !

TF-v-expr:

Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M

worm-HM:

Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C

ENCODE:

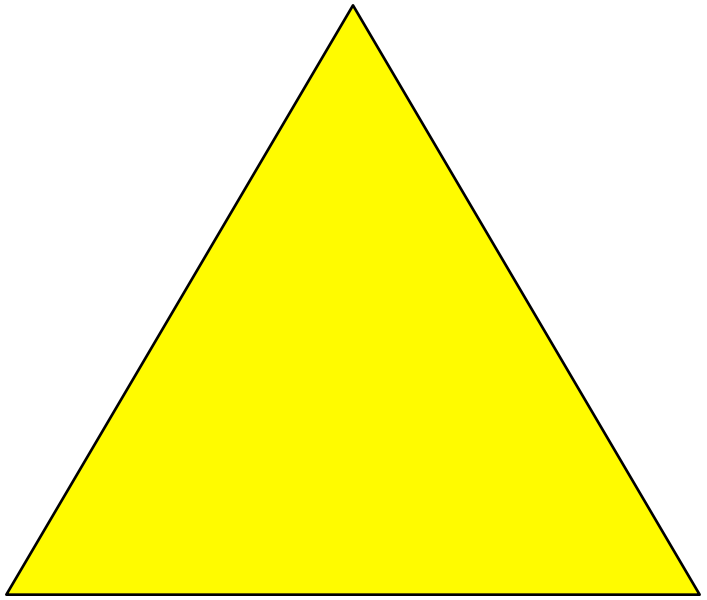
Chao Cheng, Roger Alexander, **Renqiang Min**, Kevin Y. Yip, Jing Leng, Joel Rozowsky, Koon-kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie A Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, Roderic Guigó Serra, Ewan Birney, Zhiping Weng, Michael Snyder

Models Acknowledgements

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