Report on Transcriptome Analysis

modENCODE Joint AWG call

13 July 2012

Datasets

- agreed-upon “expression compendium”
  - total RNA
  - ENCODE Tier 1
- developmental time courses (worm, fly)
- matched embryonic datasets

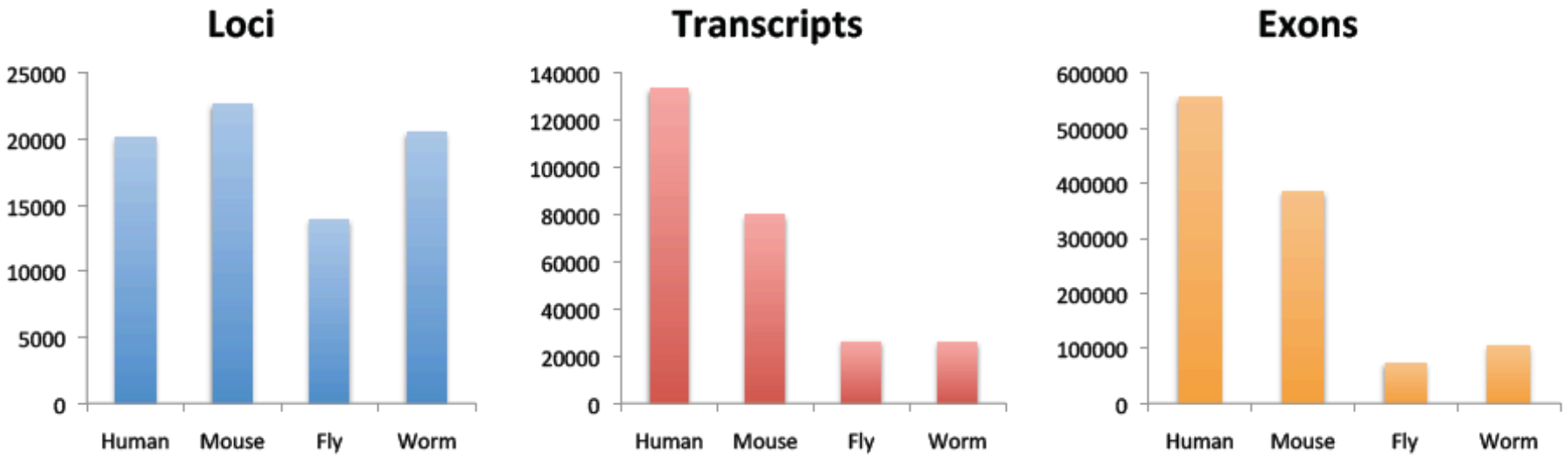
Transcription Paper Outline

- Comparison of protein-coding genes
  - Comparison with existing annotations (Hillier, Davis, Brown)
  - Splicing complexity (Graveley)
  - Comparison of select orthologs (Mortazavi, Harrow, Celniker)
- Comparison on non-coding RNAs (Brown, Lai, Gerstein, Guigo, Samsonova)
- Comparison of pseudogenes (Gerstein)
- Analysis of relationship of upstream regions to transcript level (Gerstein, Weng)
- Expression clustering (Brenner, Gerstein)

Comparison with existing annotations

- Because of the difficulty of assembling full transcripts with short reads and comparing their expression across species, we will focus on comparing transcript elements:
  - Transcript Start Sites (TSSs)
  - Transcript End Sites (TESs)
  - Splice Junctions (SJ)
  - de novo exons
  - de novo genes
  - de novo transcripts
  - Expression values for each above element
  - Expression values for the annotations

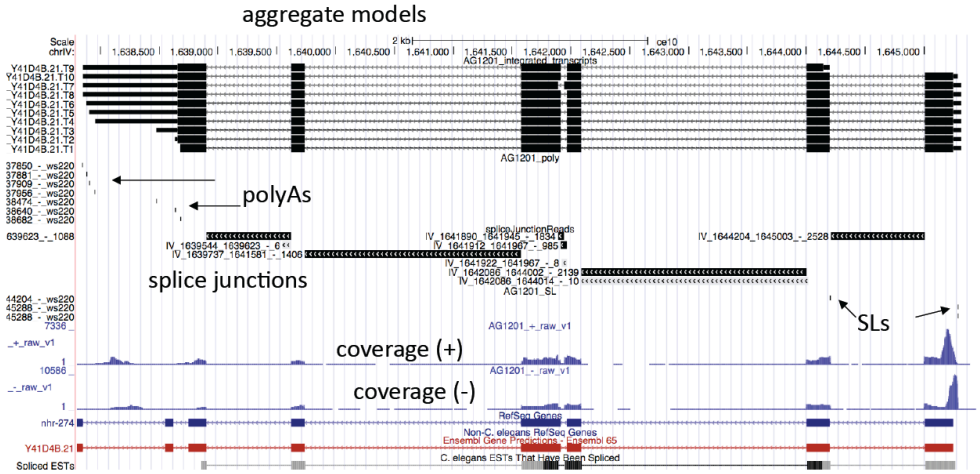
Number of protein-coding genes



Adam Frankish

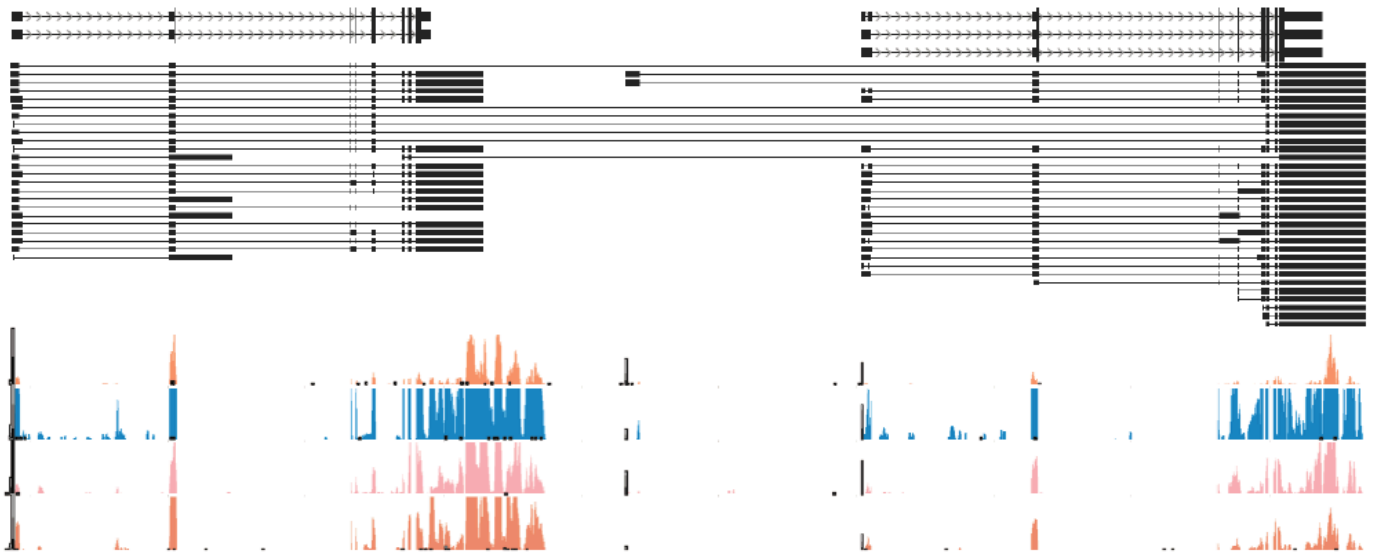
Finding all isoforms of a gene can be difficult

Simple Case



C. elegans refseq models and spliced ESTs

Hard Case

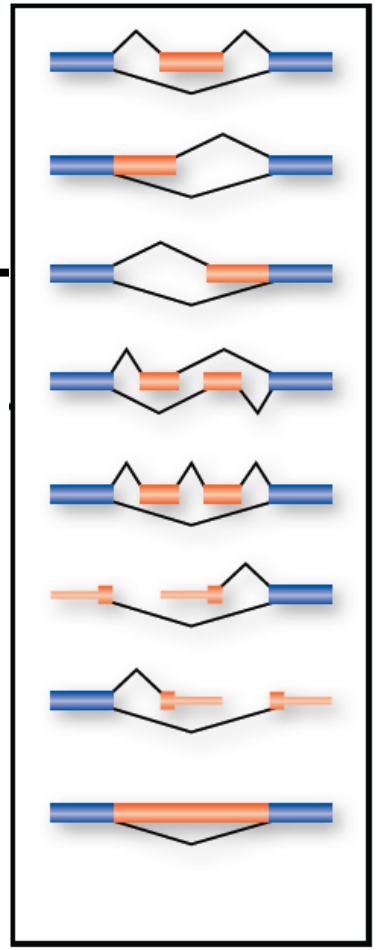


Analysis of Splicing Complexity

Transcript Annotations

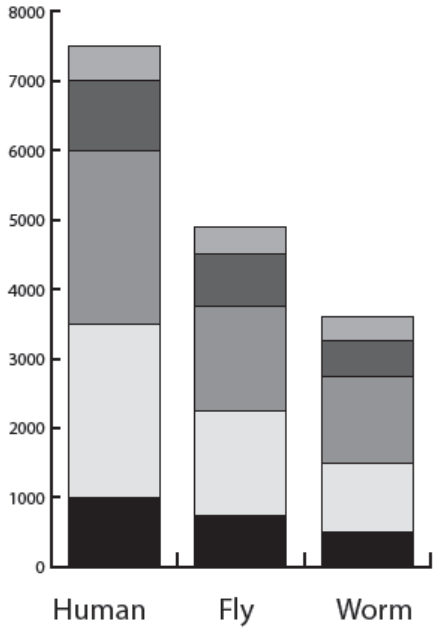
↓

Event Classifications



↓

Comparison of Event Types

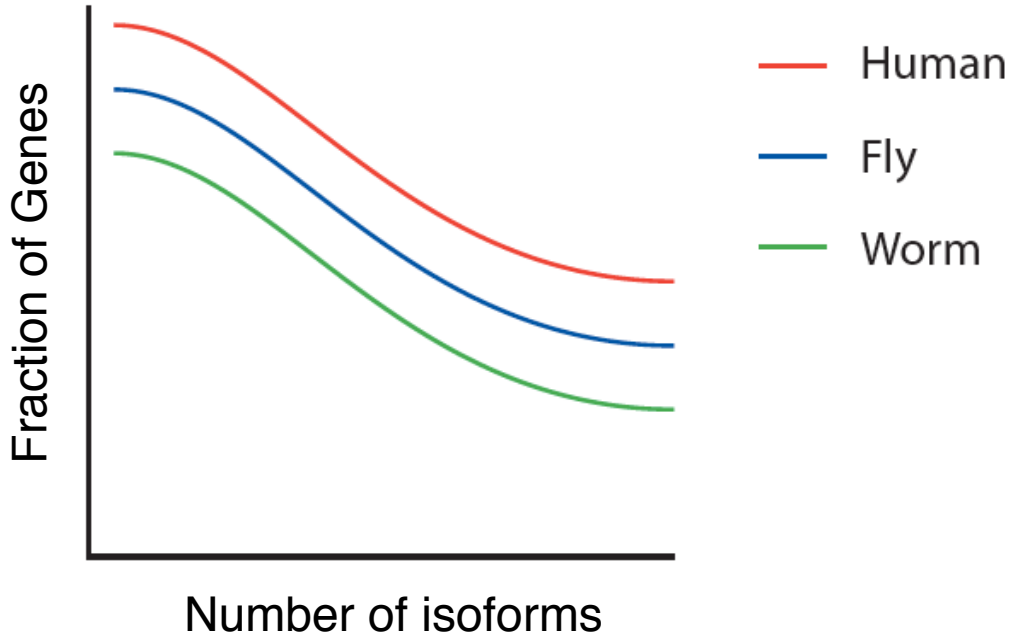


Analysis of Splicing Complexity

- For all three species, compare motifs and conservation at splice sites for constitutive vs. alternative exons, and highly switching vs. low switching.

- Analyze number of isoforms per gene.

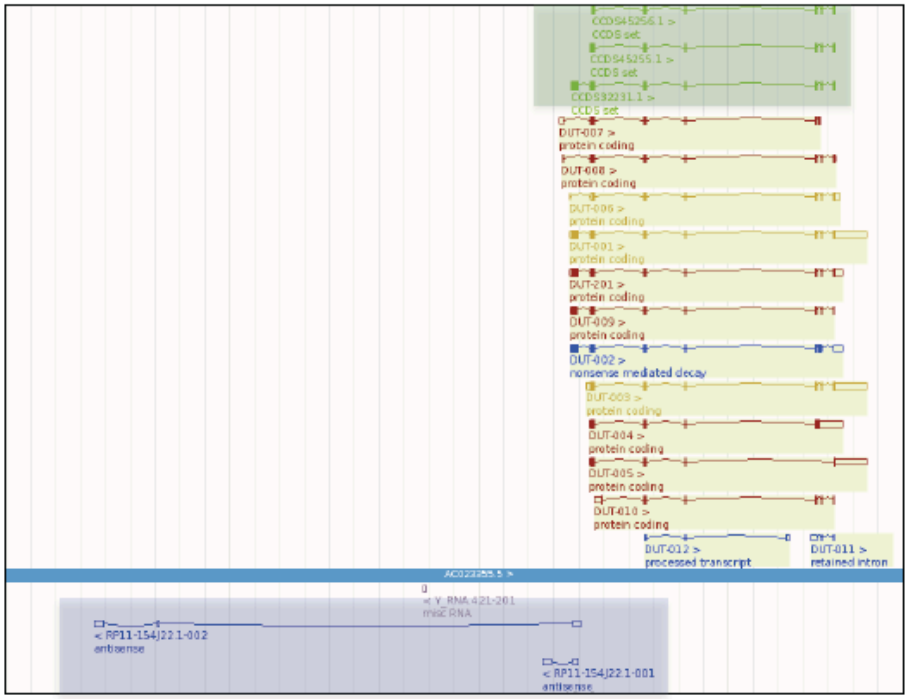
  Highlight outliers (Dscam, etc.)



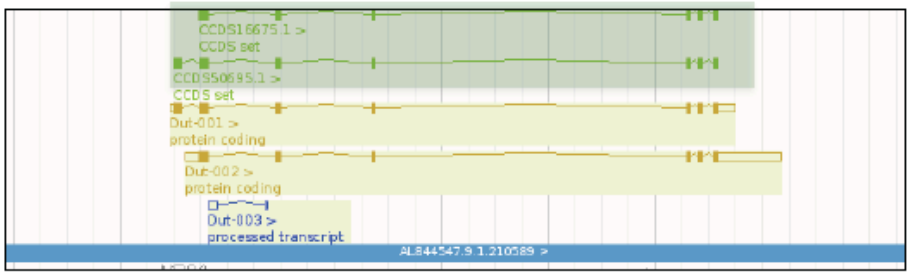
Comparison of select orthologs

Case Study: DUT / Dut / dUTPase / dut-1

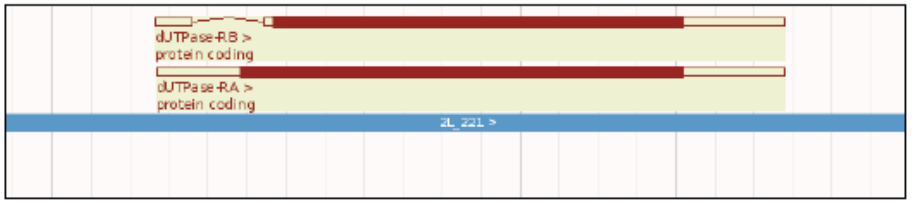
Human DUT



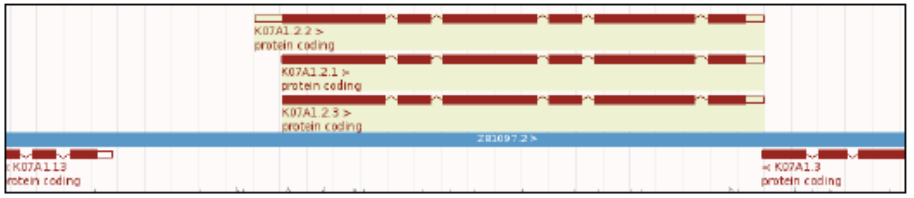
Mouse Dut



Fly dUTPase



Worm dut-1

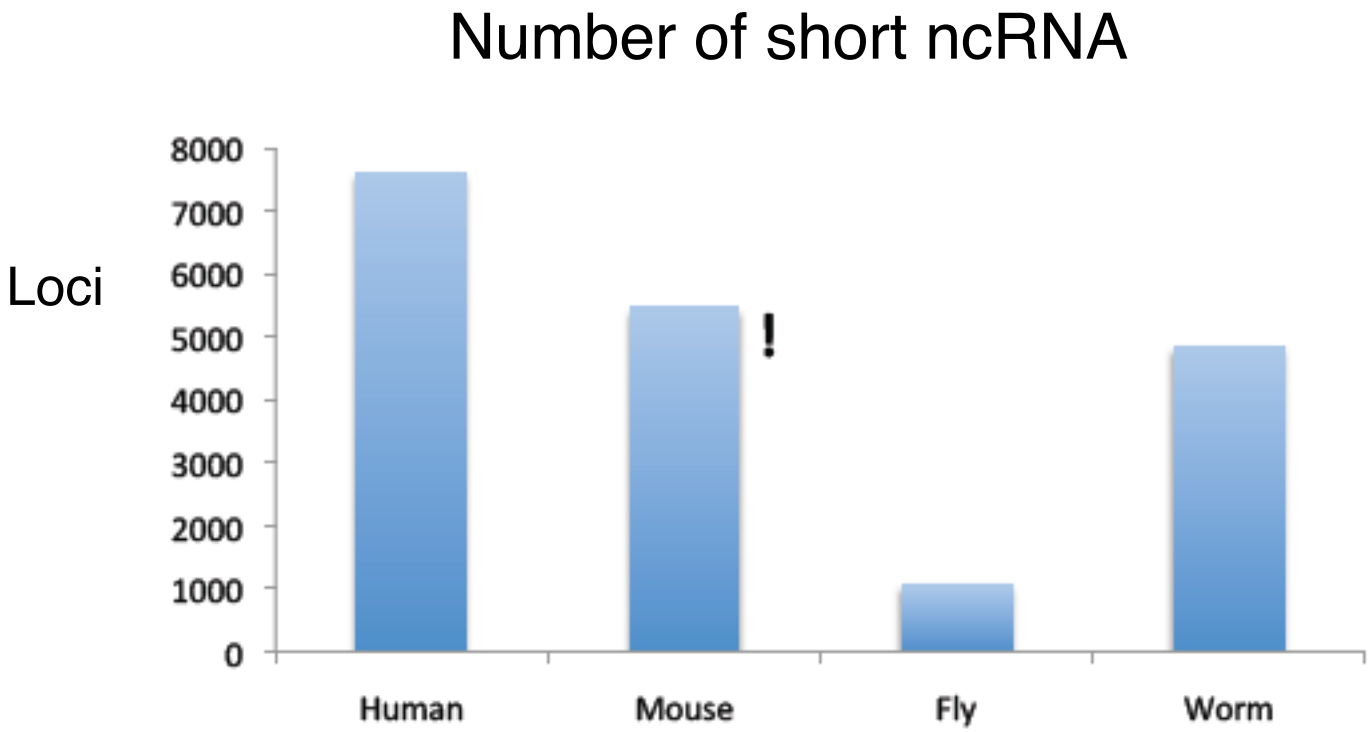


Adam Frankish

Comparison of non-coding RNAs

- How much of the nc genome is transcribed?
  - per megabase
  - across entire agreed-upon “expression compendium”
  - in ~matched embryonic stages
  - Ubiquitous vs Stage- / Cell-line specific transcription
- You cannot directly compare annotations (Gencode vs Flybase vs Wormbase)
- so, use a tiered approach; build a table or pie chart
  - first compare the existing annotations
  - incRNA algorithm
    - breakdown by RNA class
  - *de novo* mapping / TAR calling
    - issues: repeats, multi-mapped vs unique reads

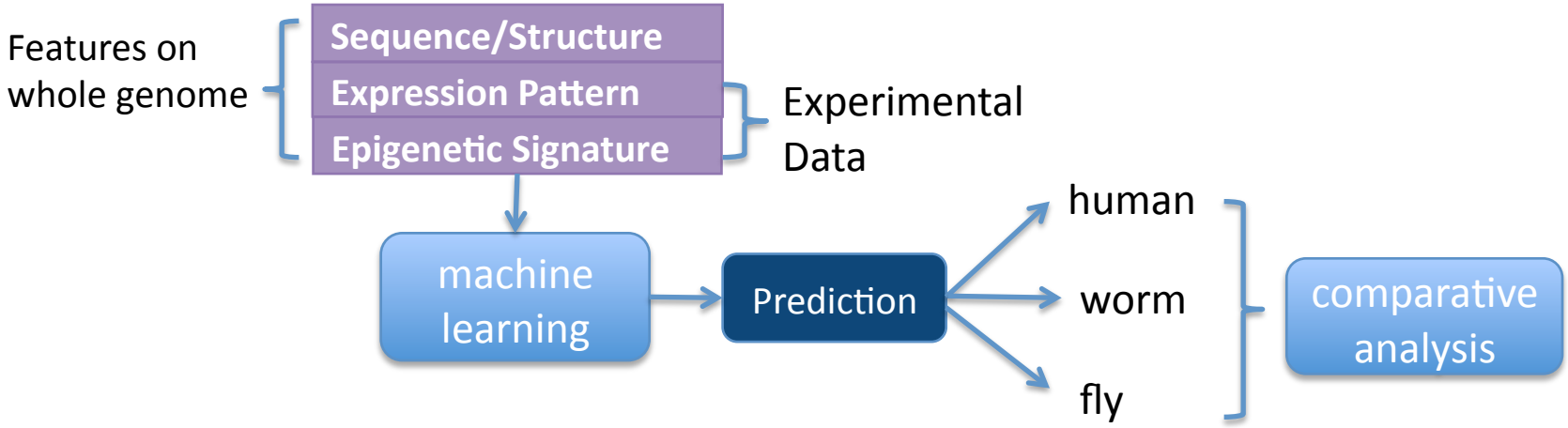
Comparison of existing annotations



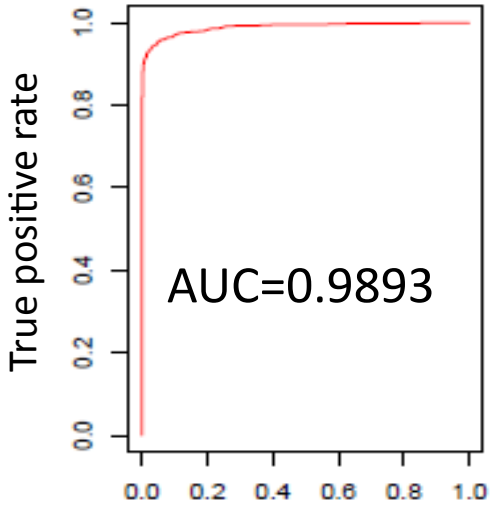
rRNA, tRNA, miRNA, snRNA, snoRNA (! mouse excludes tRNA)

Adam Frankish

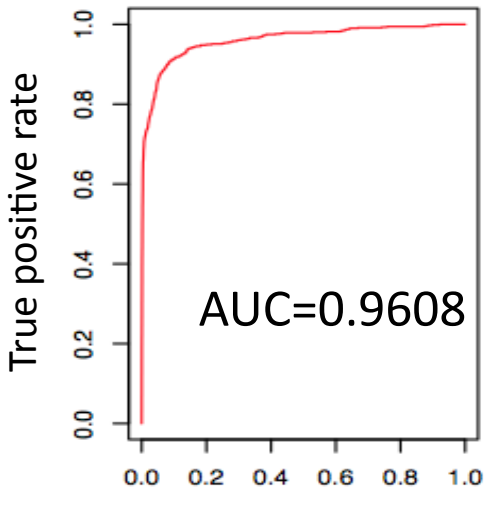
incRNA algorithm



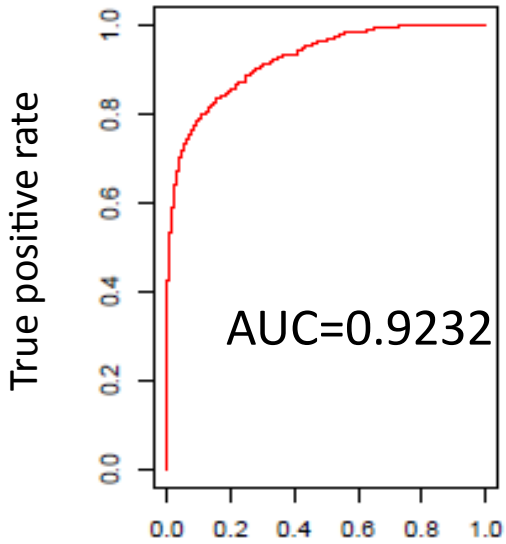
Results for known types of ncRNAs:



fly



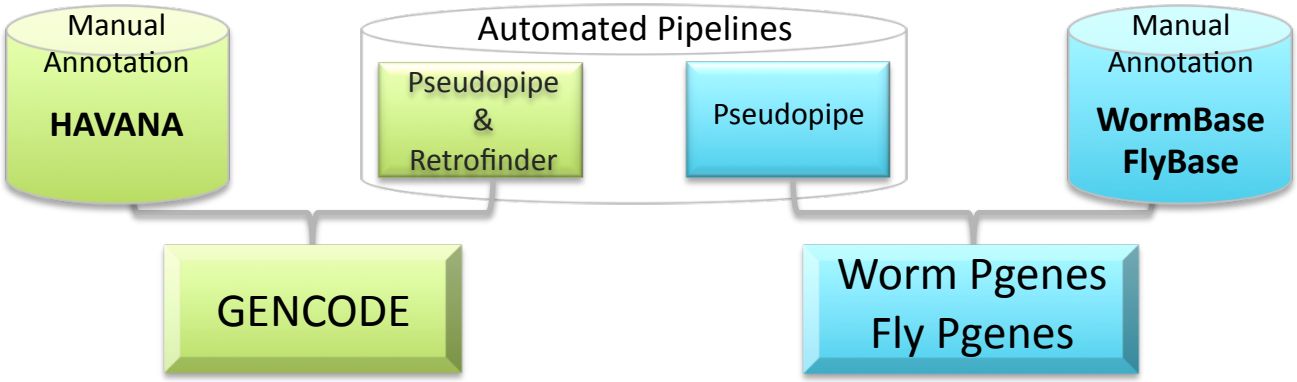
worm



human

Comparison of pseudogenes

- Pseudogenes annotated using automated pipelines intersected with manual curation



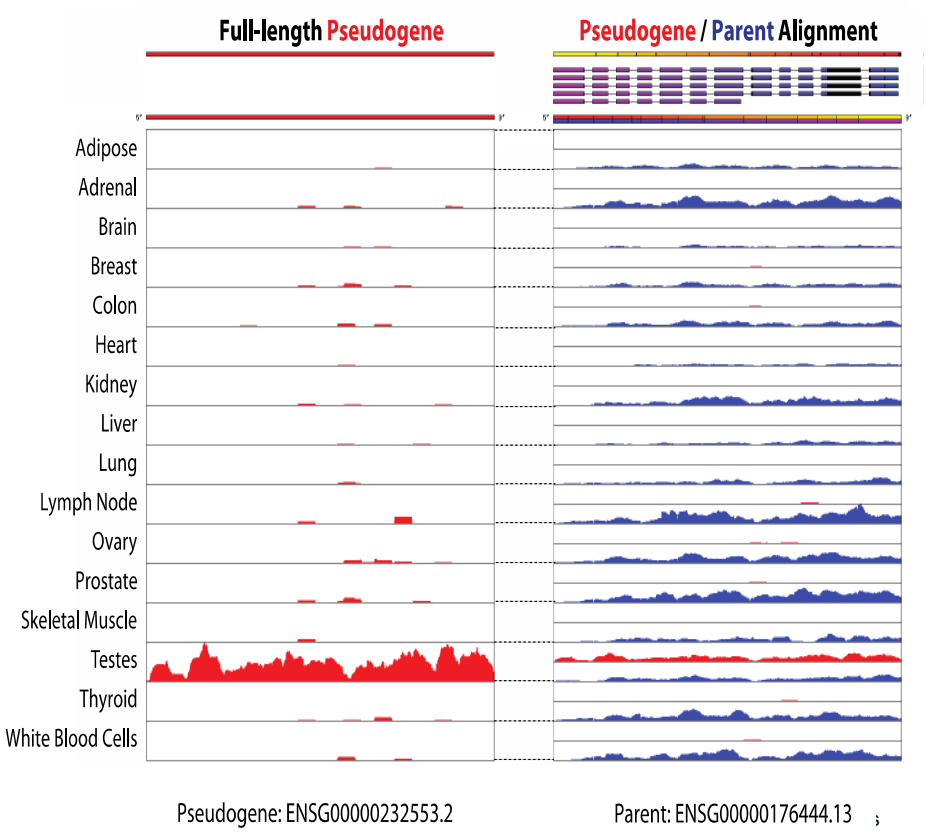
	Human – GENCODE	Worm	Fly
Total	11240 (14112*)	1198	529
Duplicated	2158	538	119
Processed	8715	255	95
Ambiguous	23	405	315
Others**	344		

* Estimated total number of pseudogenes in human genome.

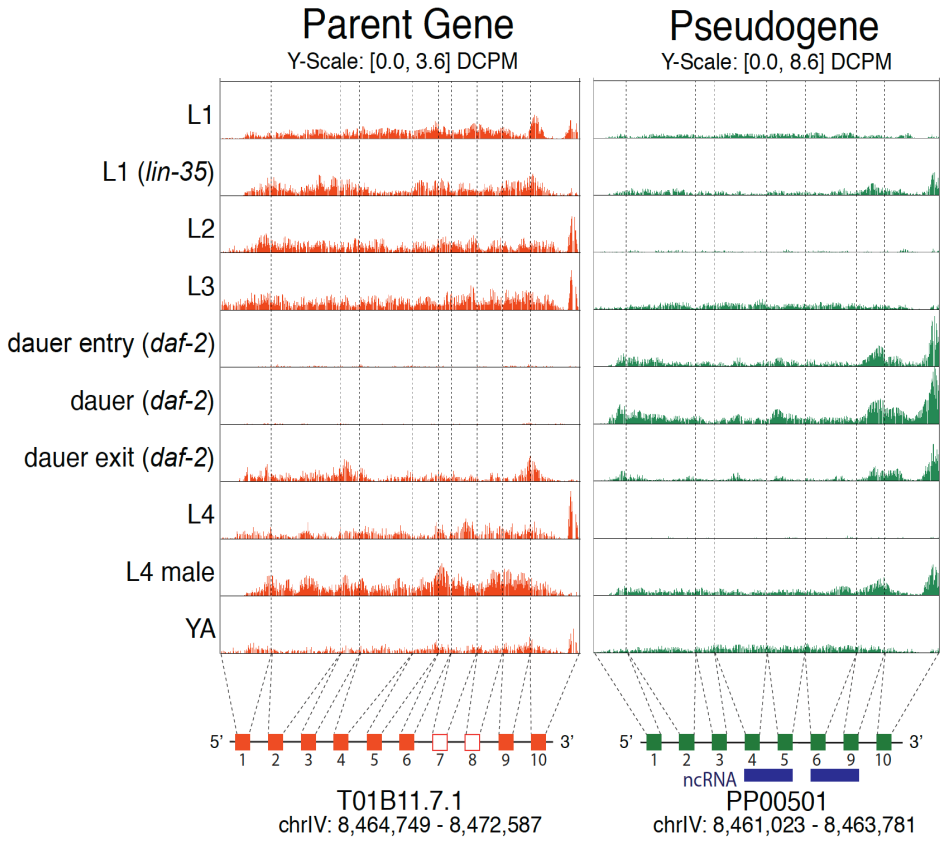
** Including Unitary (138), IG (161) TR V (21) and polymorphic (24) pseudogenes

*Transcribed Pseudogenes

Human

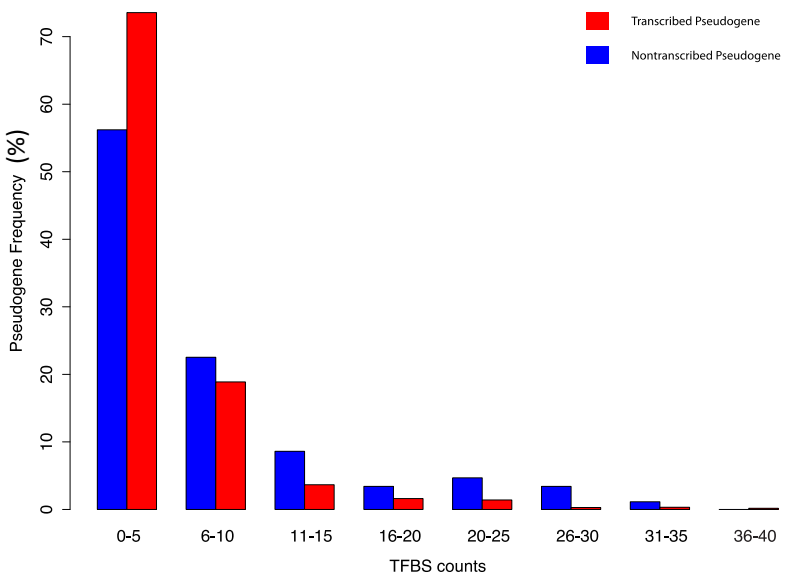


Worm

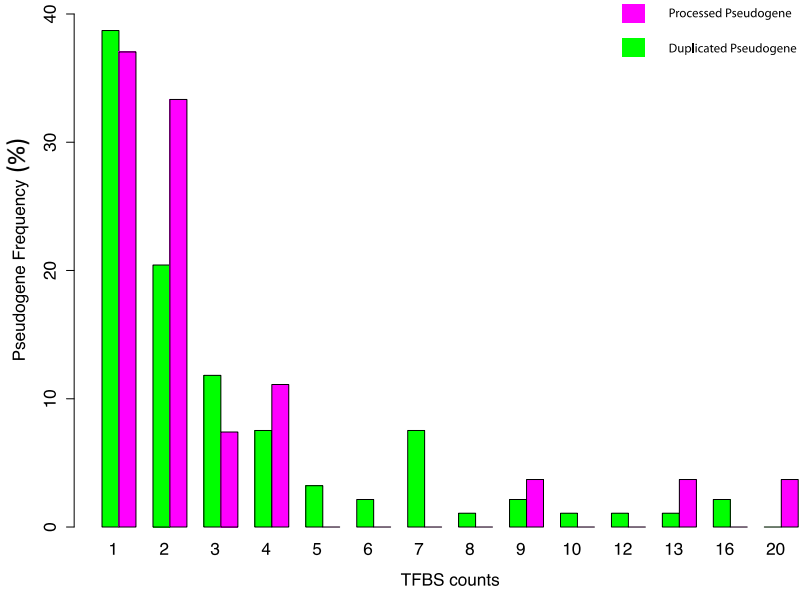


*Transcription Factor Binding Sites

Human



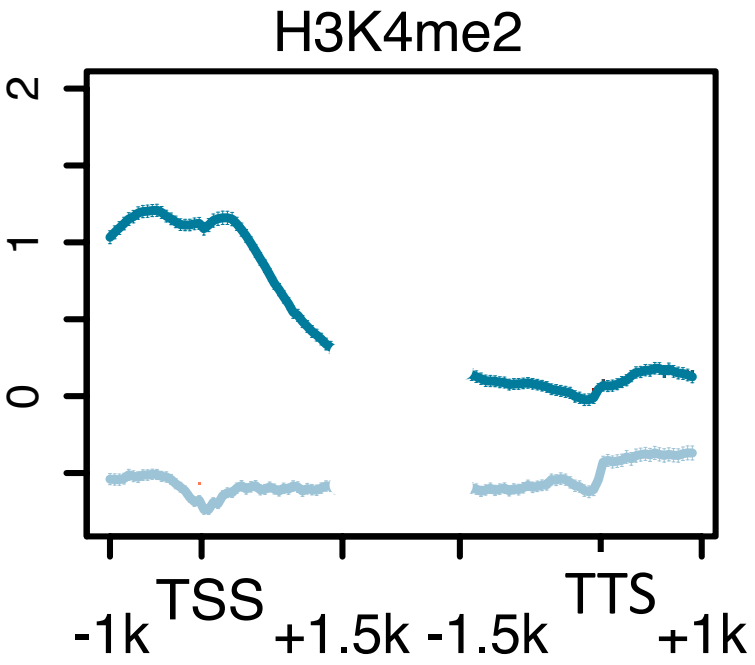
Worm

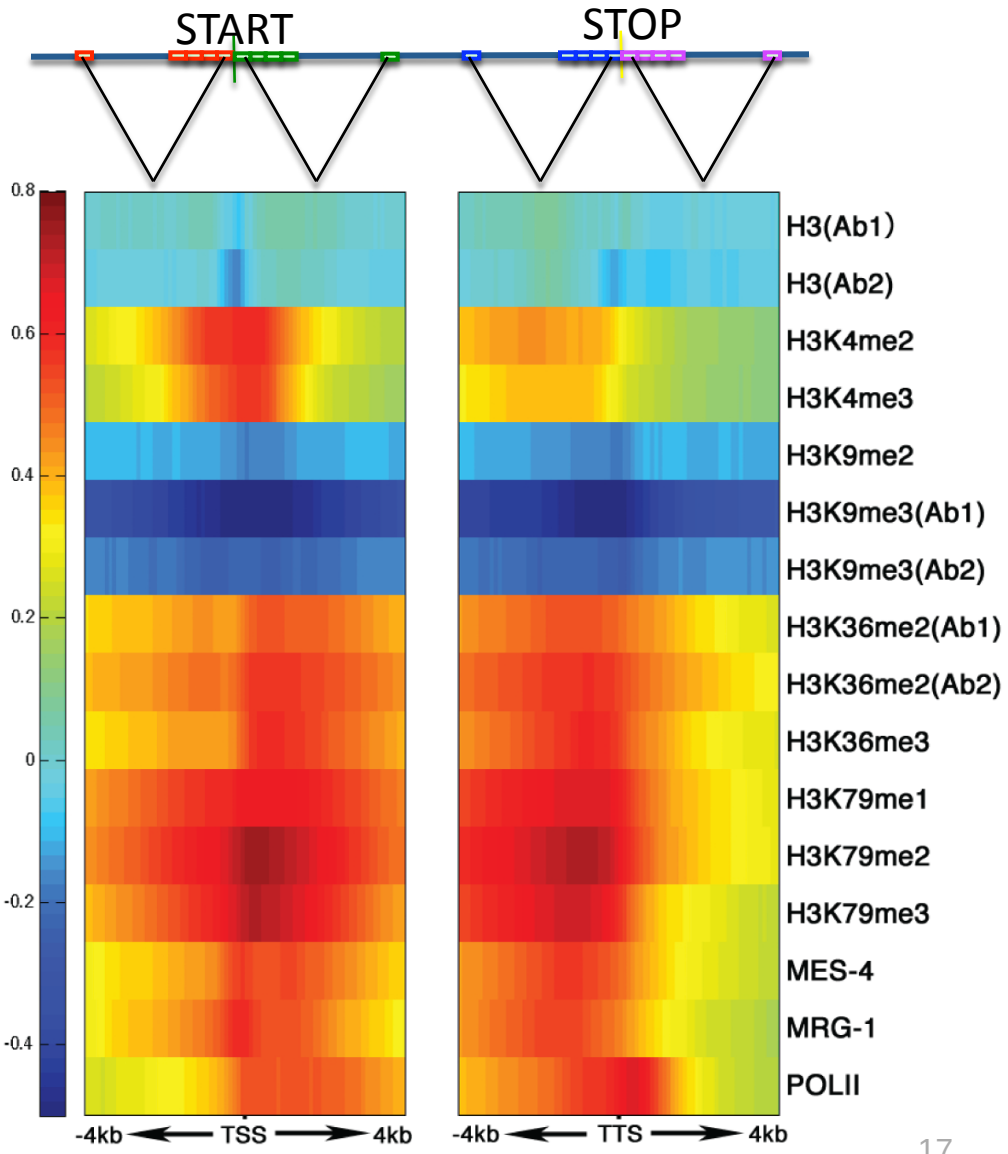


- TFBS were selected within 2kb upstream of the pseudogene start site
- 95 (58) duplicated and 29 (20) processed pseudogenes had TFBS in the upstream region

Analysis of relationship of upstream regions to transcript level

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



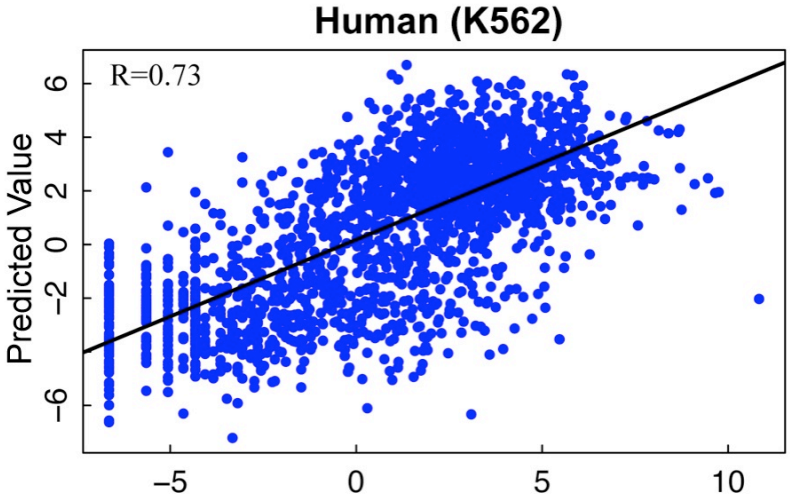


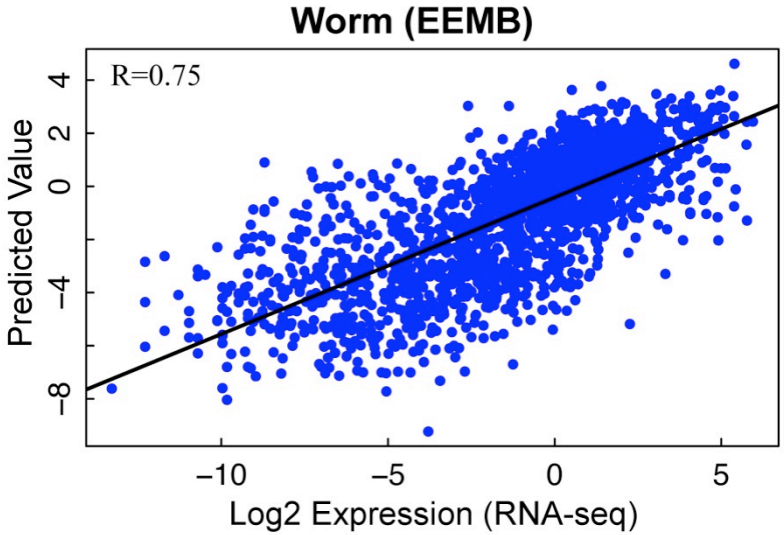
Correlation between Signal and expression

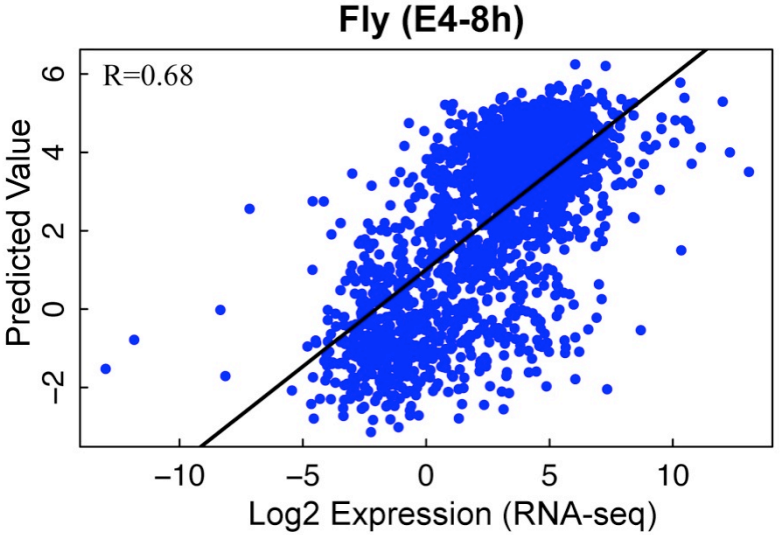
Chao Cheng

Application of HM model in 3 species: Consistent Performance

>50% of variation of expression levels can be explained by HMs







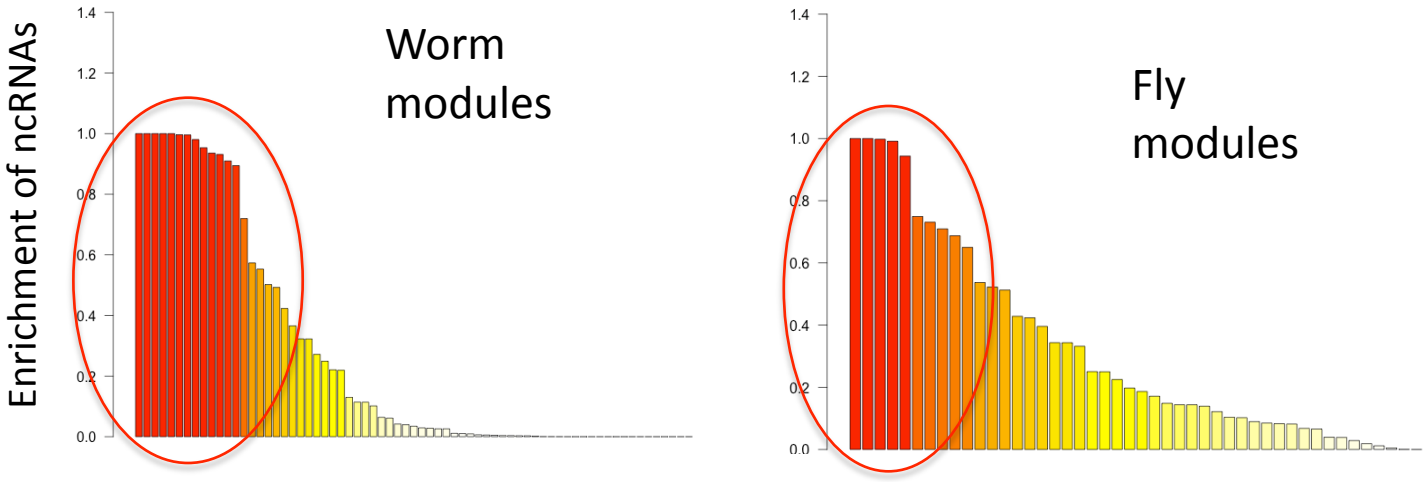
Chao Cheng

Expression clustering of protein-coding and ncRNA genes in embryo development

Species	Developmental stages	Protein-coding genes*	Non-coding RNAs*	Co-expression modules**
Worm (C. elegans)	111	9114	855	69
Fly (D. mel.)	50	8340	357	46

* >80% valid samples, coeff. of variance > 1 in the modENCODE finalized datasets in June 2012
** clustering via weighted gene co-expression network analysis (WGCNA)

Many co-expression modules are enriched with ncRNAs (red circles).

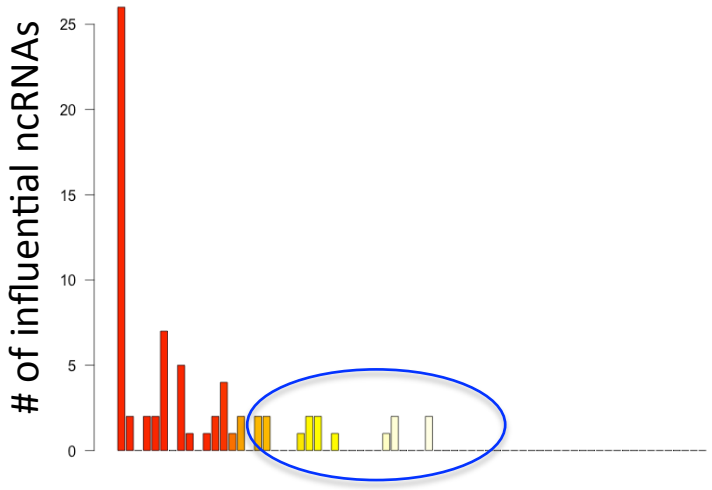


Daifeng Wang

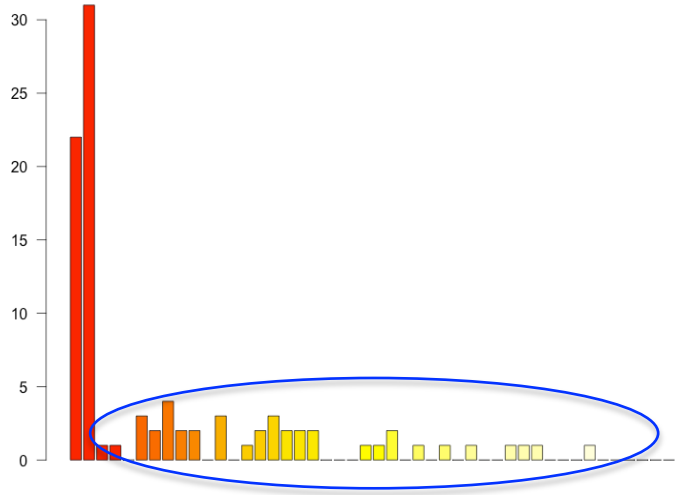
Influence of ncRNA hubs on protein-coding co-expression modules

Influential ncRNAs (high network centrality) exist in modules NOT enriched with ncRNAs (blue circles).

Worm modules



Fly modules

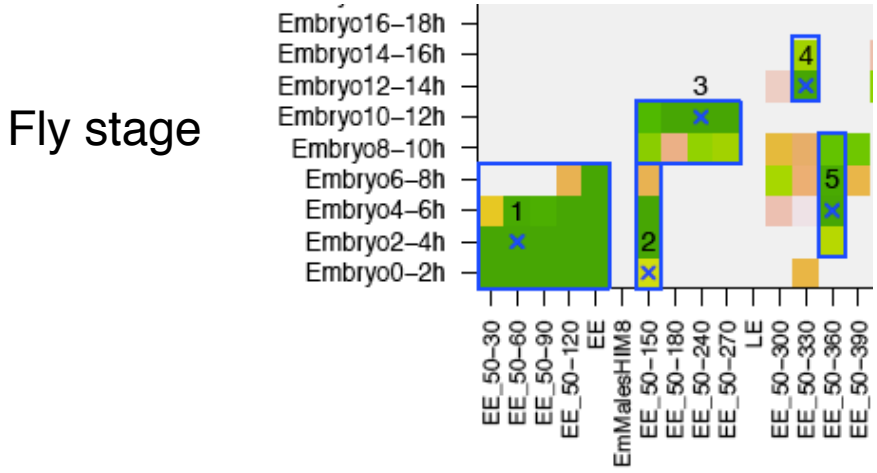


Daifeng Wang

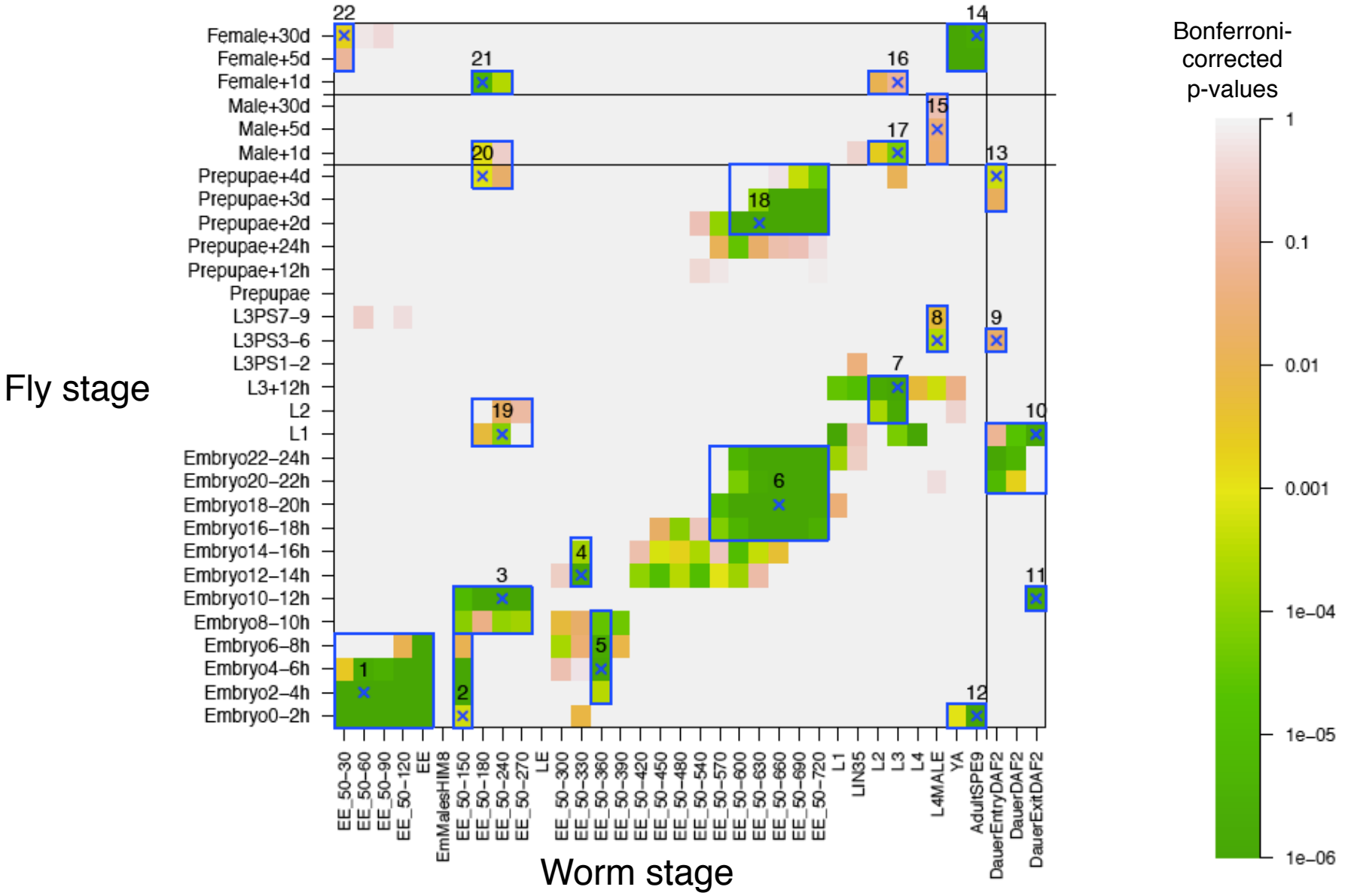
Developmental stage mapping between worm and fly based on co-expression clustering of orthologs

- Gene expression threshold: FPKM >=1 and z >= 1.5
- Significance calculated from fraction of orthologs co-expressed between pairs of stages compared to hypergeometric expectation

- Cluster numbering facilitates follow-on analysis:



Developmental stage mapping between worm and fly based on co-expression clustering of orthologs



Jingyi Jessica Li, Peter Bickel, Haiyan Huang, Steven Brenner

END

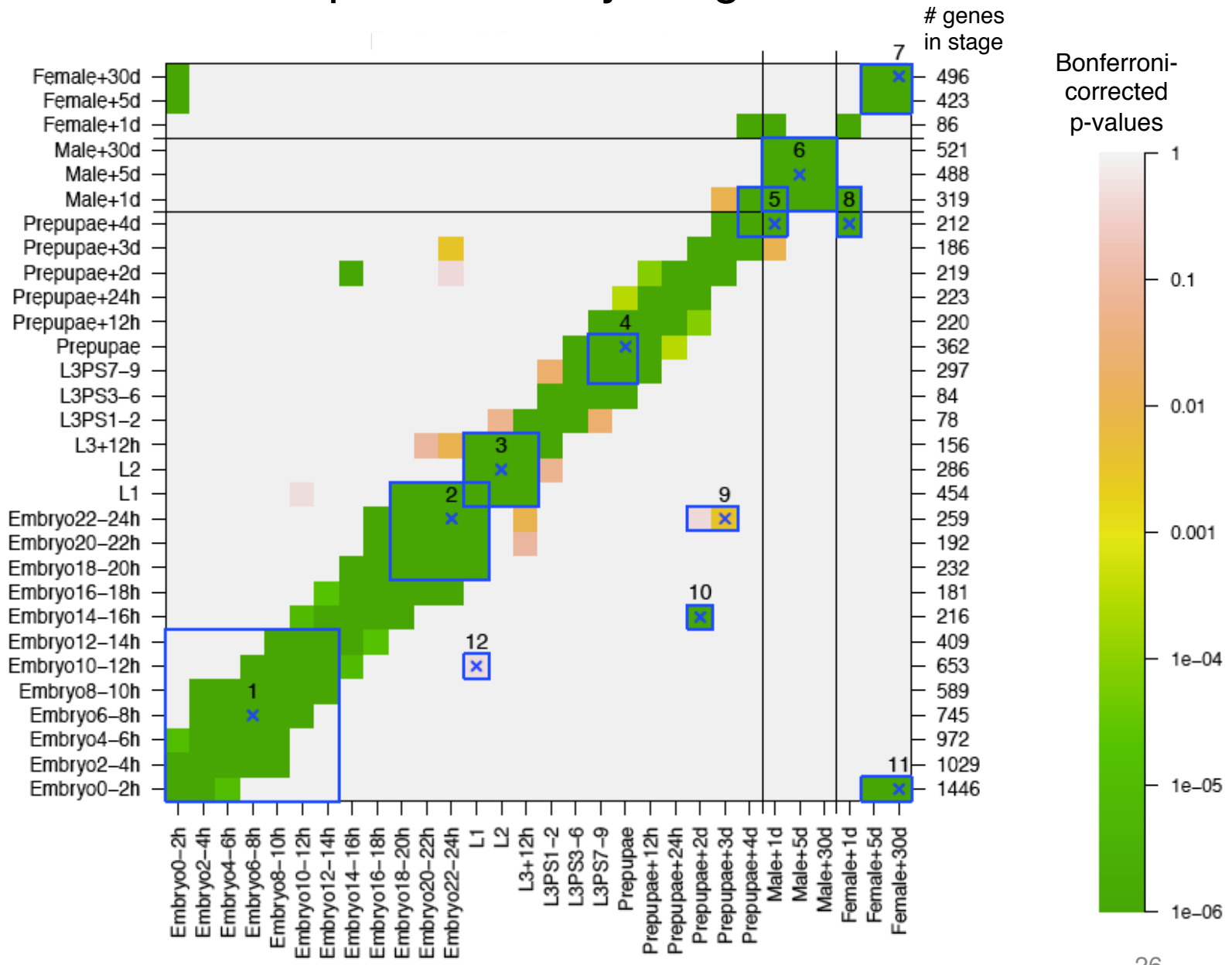
Production Stats - Worm

	Samples	**Total** Reads	**Total Unique** Reads
Embryonic Time Course	106	1,633,419,670	1,031,557,649
Life Stages	70	2,401,311,389	1,420,342,487
Other Species	54	1,779,775,463	946,431,824
Pathogens	11	702,645,329	489,536,643
Tissues	183	3,560,398,393	1,322,552,917
Totals		10,077,550,244	5,210,421,520

Production Stats - Fly

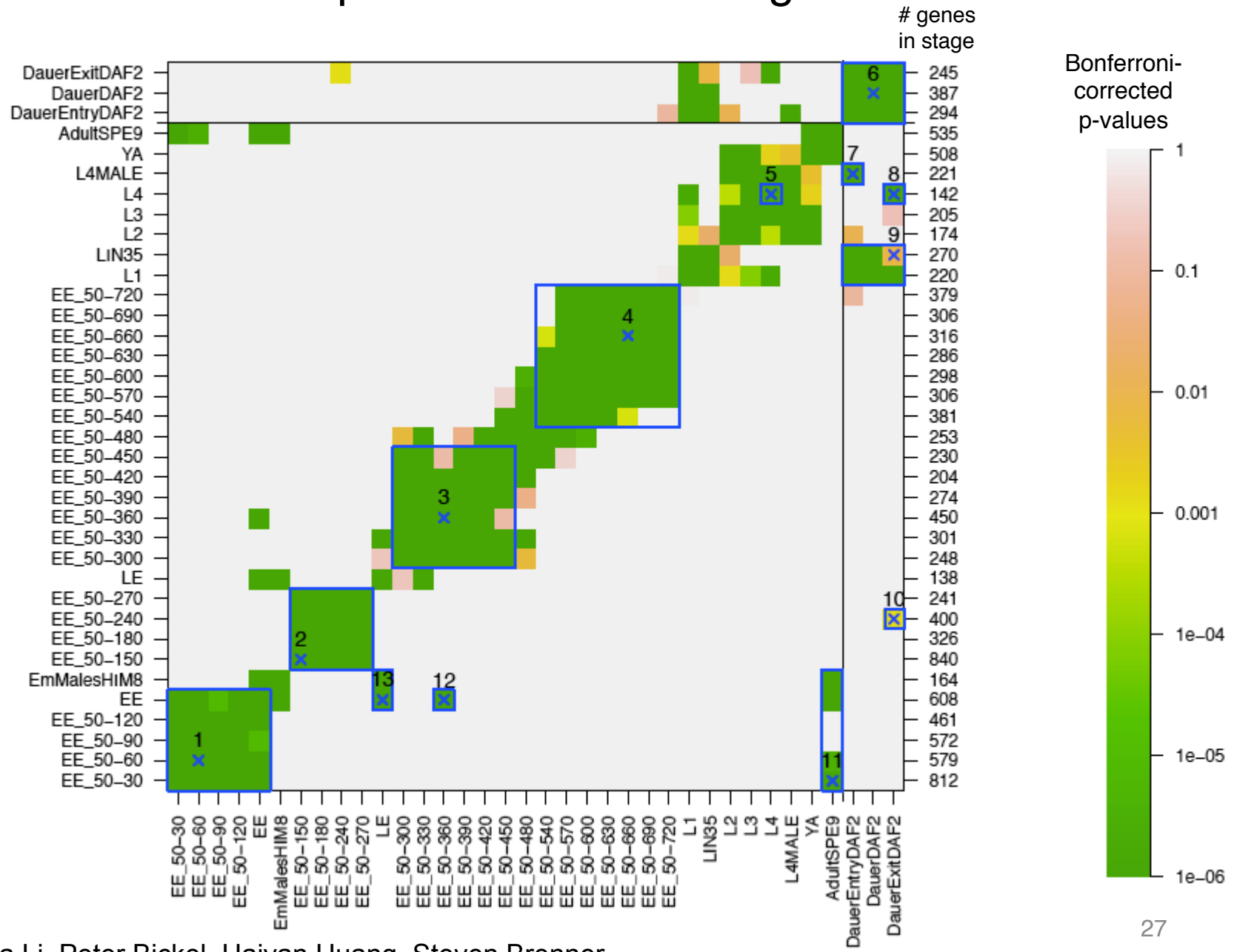
Experiment	Samples	Total Reads	Total Unique Reads	Total Unique bp
Cell Lines	25	1,677,980,920	1,272,452,612	96,706,398,512
Tissues	29	4,265,585,752	3,667,365,400	278,719,770,400
Treatment	21	6,495,812,560	4,949,215,447	376,140,373,972
Poly(A) Tail Enrichment	29	845,610,153	638,882,610	48,555,078,360
Developmental Time Course*	30	3,538,880,404	2,282,408,273	171,180,620,475
Genome Resequencing	25	943,927,826	N/A	71,738,514,776
Total	247	17,767,797,615	12,810,324,342	1,043,040,756,495

Comparison of Fly Stages



Jingyi Jessica Li, Peter Bickel, Haiyan Huang, Steven Brenner

Comparison of Worm Stages



Jingyi Jessica Li, Peter Bickel, Haiyan Huang, Steven Brenner