modEncode Pseudogenes: Fly, Worm, and Human

Cristina Sisu

Group Meeting

8th August 2012

Overview

- Sequence analysis
- Transcription factors
- Pseudogene transcription
- Chromatin features
- Data summary
- Conclusions & future work

Data & Pipelines

Human:

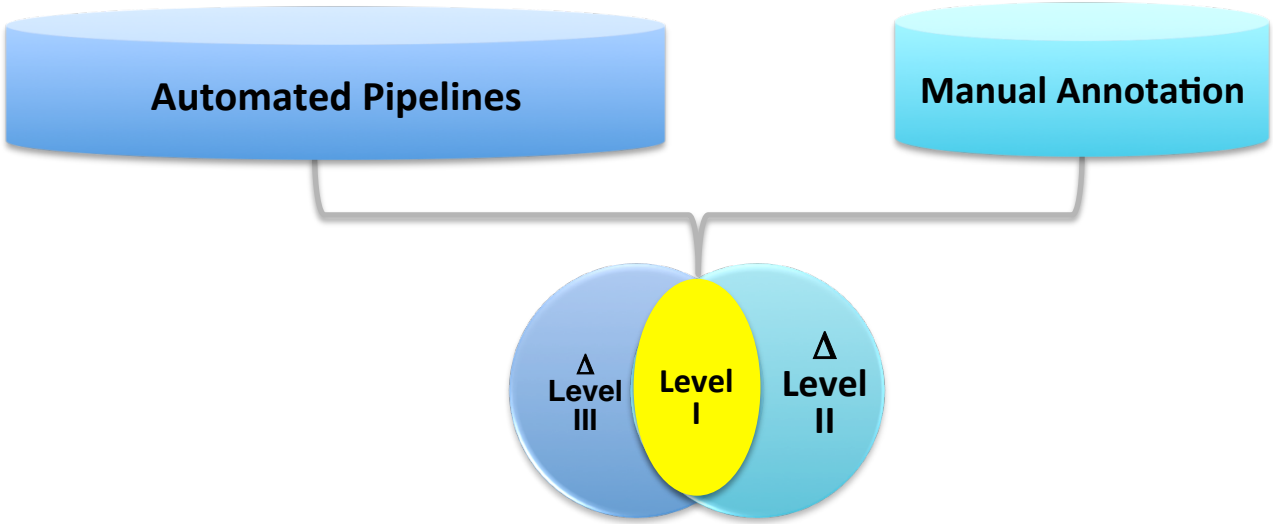
- Havana
- PseudoPipe & RetroFinder

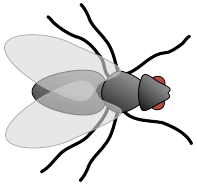
Worm:

- WormBase – release WS220
- PseudoPipe

Fly:

- FlyBase – build 5 release 5.45
- PseudoPipe





Annotation

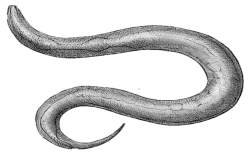
	Total	Duplicated	Processed	Ambiguous
PseudoPipe on FlyBase 5.25	529	119	95	315
PseudoPipe on FlyBase 5.39	312	46	75	191
FlyBase 5.39 (Ensembl)	170			
Intersection 80% (PPipe5.39-Flybase5.39)	114	33	32	49
FlyBase 5.45 (flybase.org)	**184**			
Intersection 80% (PPipe5.39-Flybase5.45)	**113**	**34**	**31**	**48**
Δ – PseudoPipe 5.39	*199*	*12*	*44*	*143*
Δ – FlyBase 5.45 with Parents	*~10*			

TO USE:
113 + *Δ FlyBase 5.45 with Parents*

Δ PseudoPipe 5.39

duplicated: 34
processed: 31
ambiguous 48

Δ FlyBase 5.45 with Parents



Annotation

	Total	Processed	Duplicated	Ambiguous
PseudoPipe - WS201	1198	253	538	508(153)
PseudoPipe - WS220	2267	313	719	1235
WormBase WS220 (Ensembl)	**1441**			
Intersection 80% (Ensembl – PPipeWS220)	1014	164	445	405
Intersection 80% (PPipeWS201-PPipeWS220)	1138	235	521	382
Intersection 80% (Ensembl – PPipeWS201-PPipeWS220)	**948**	**154**	**438**	**356**
Δ PseudoPipe WS220	*440*	*159*	*281*	-
Δ WS220 – with parents	*TBD*			

TO USE:
948 + *Δ WS220 with parents*

Δ – PseudoPipe WS220

duplicated: 154
processed: 438
Ambiguous 356

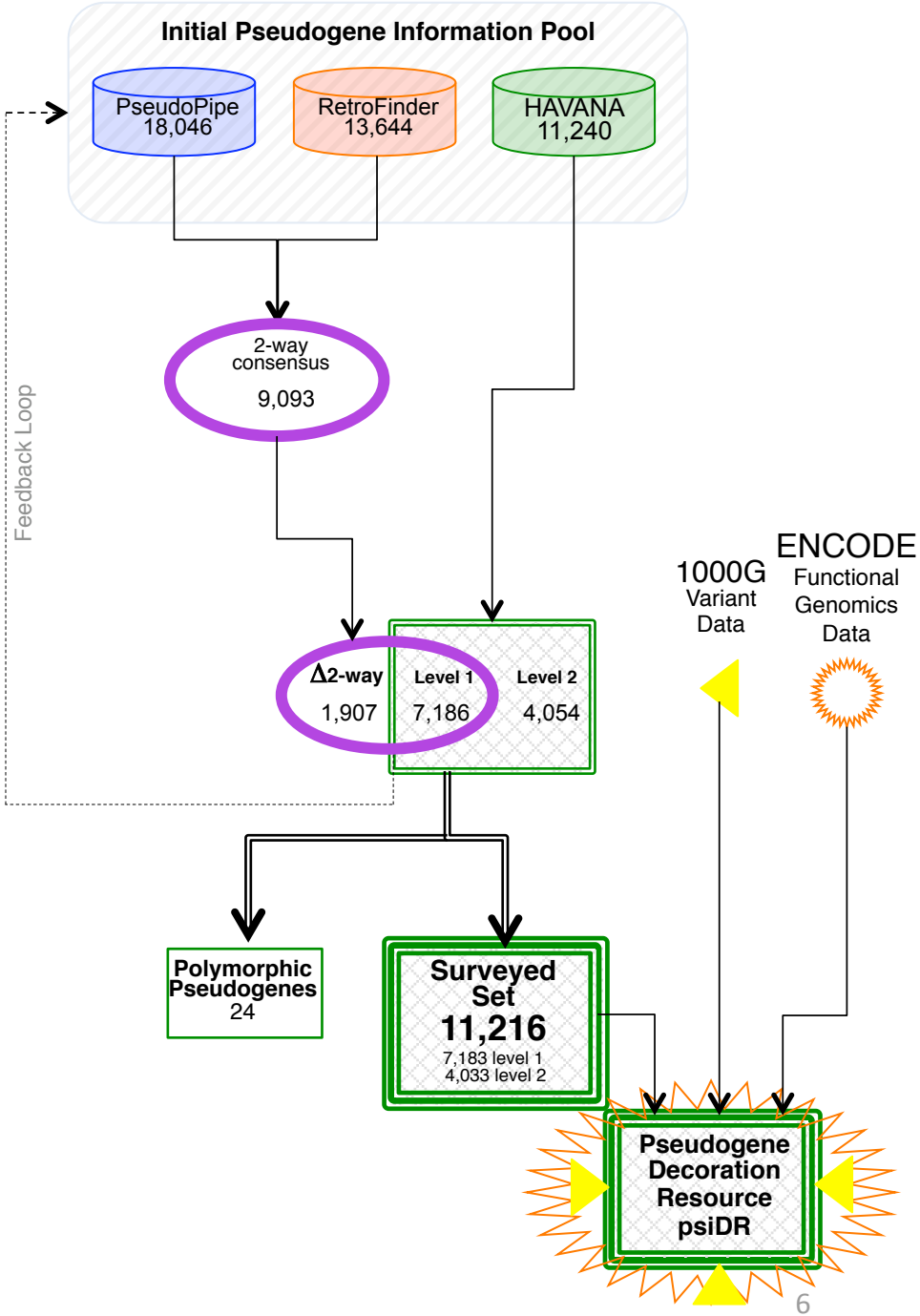
Δ – WS220 with Parents

Annotation

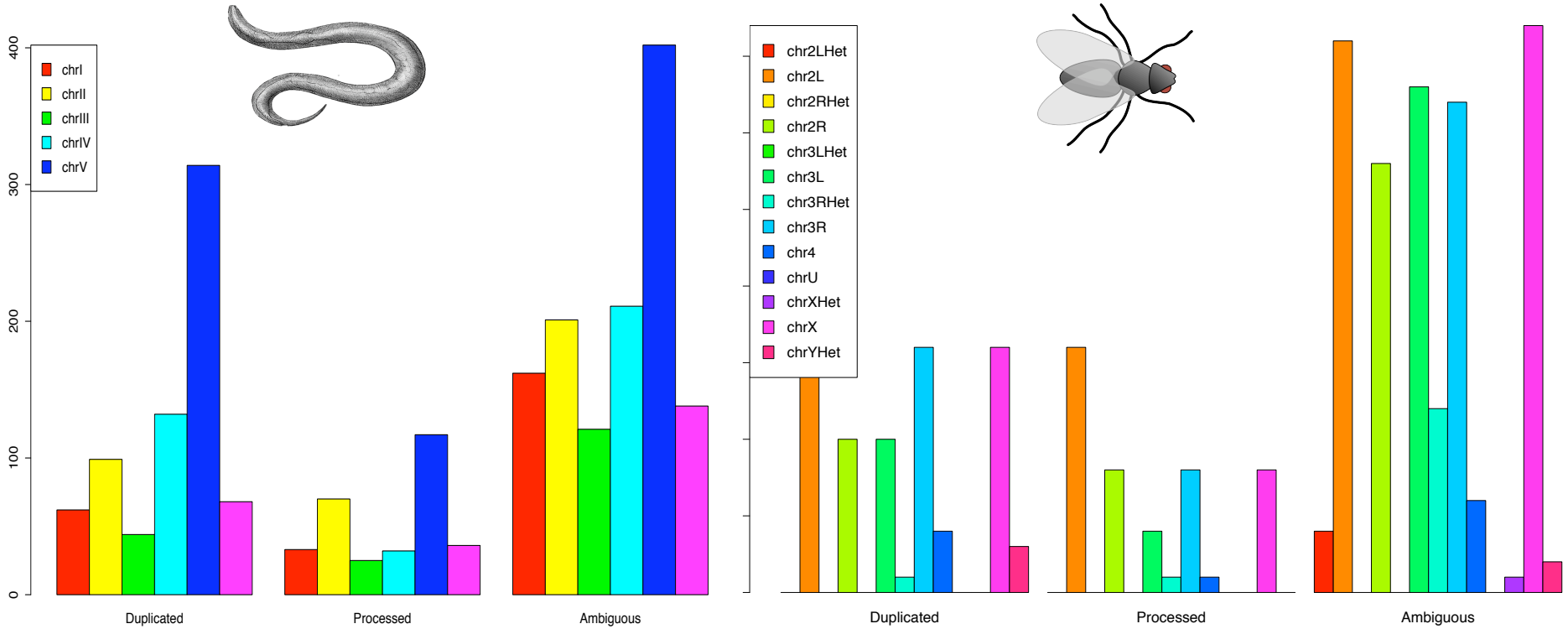
Pseudogene	Human GENCODE v7
Total	11240
Duplicated	2158
Processed	8715
Ambiguous	23
Others*	344
Total Estimated	14112**

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

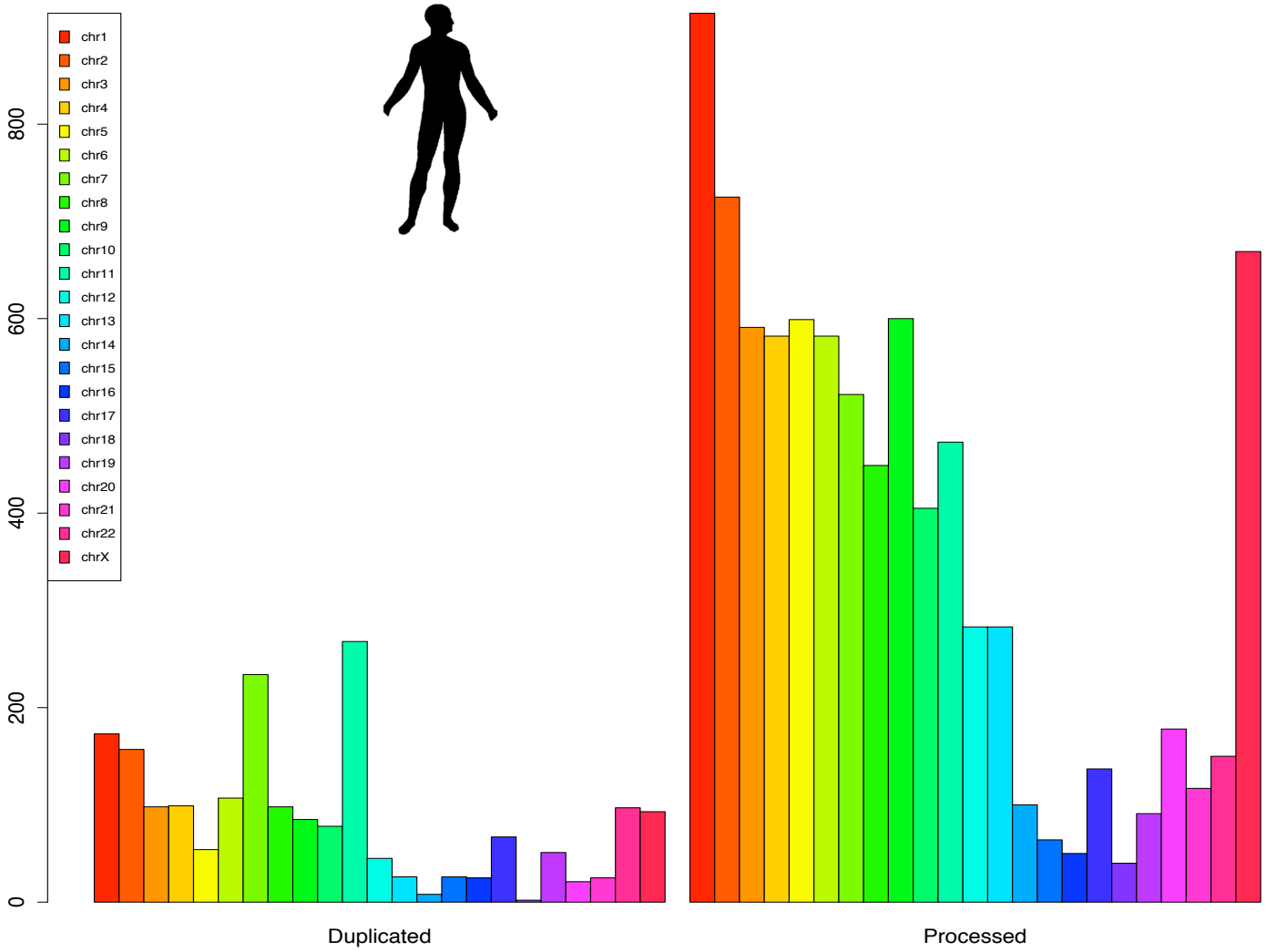
** Automatic pipelines accuracy is estimated using the manually annotated pseudogenes as a benchmark and the results are extrapolated to the whole genome



per Chromosome Distribution



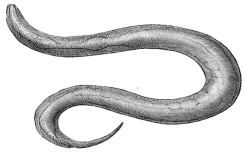
per Chromosome Distribution (2)

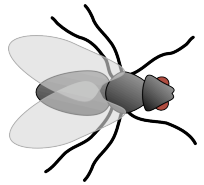


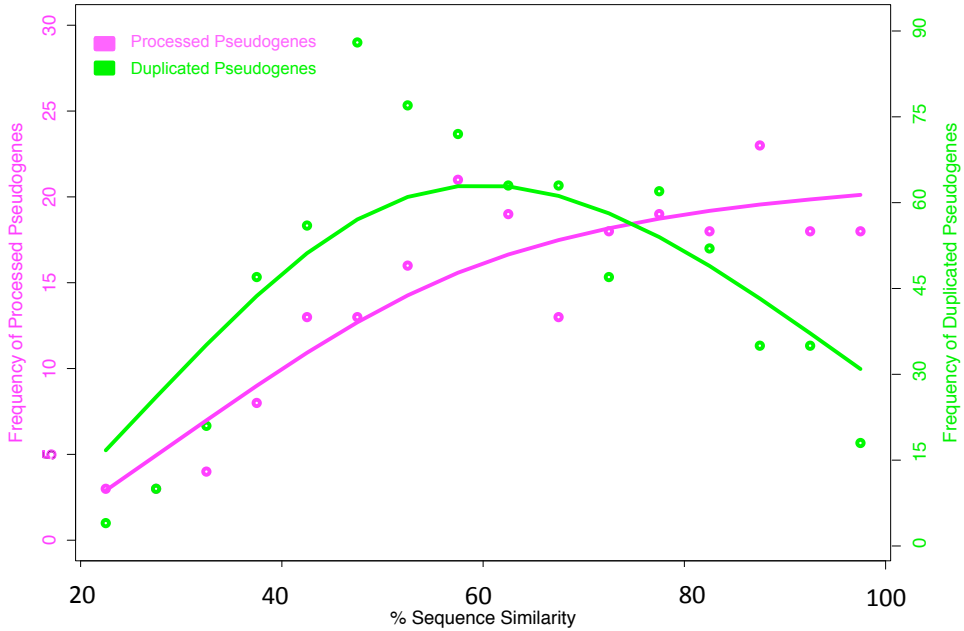
SEQUENCE ANALYSIS

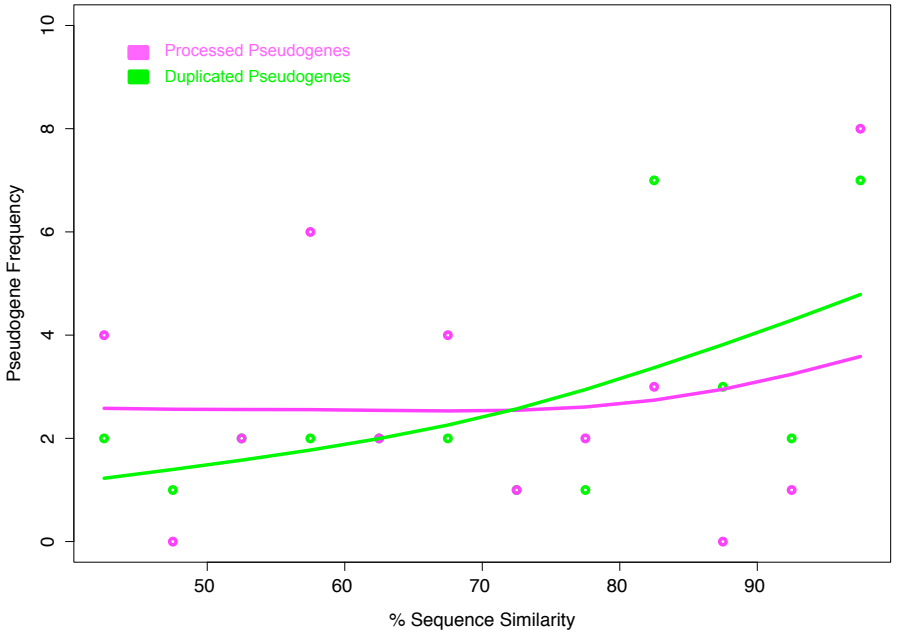
- Gives an indication of the time line of pseudogene formation
- Comparison of CDS & 3’ UTR

CDS Comparison

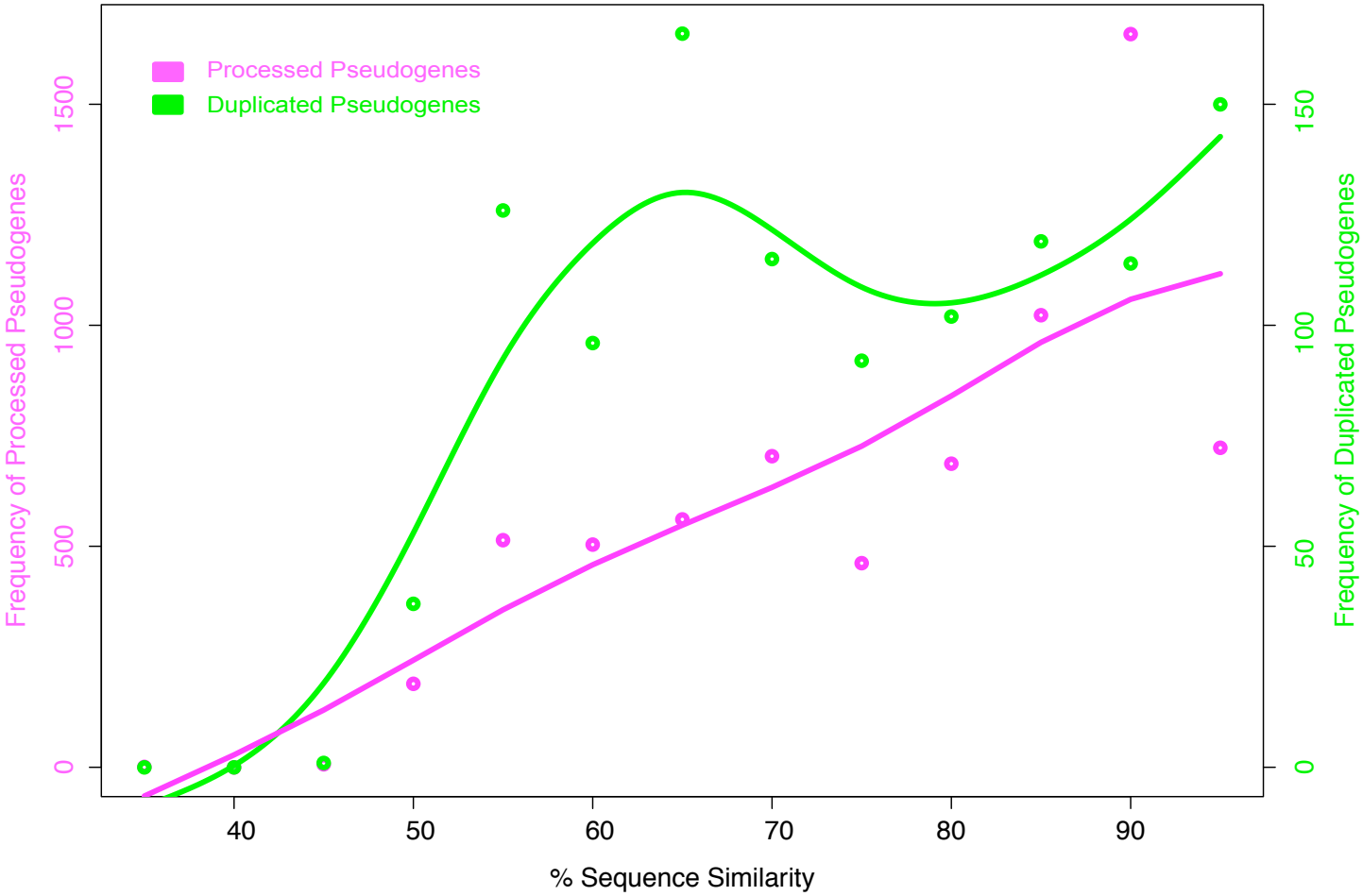




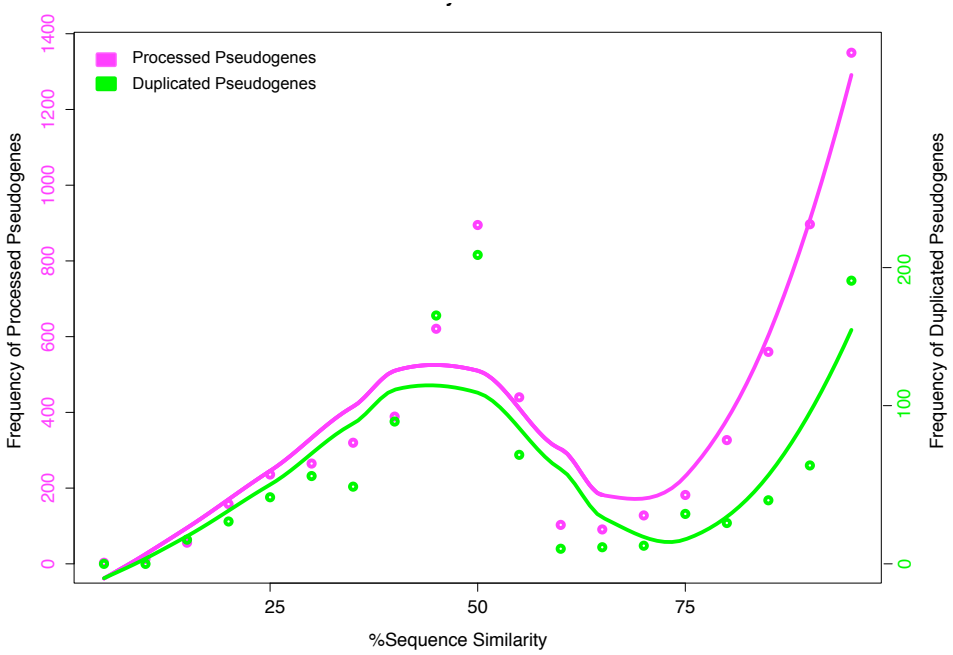


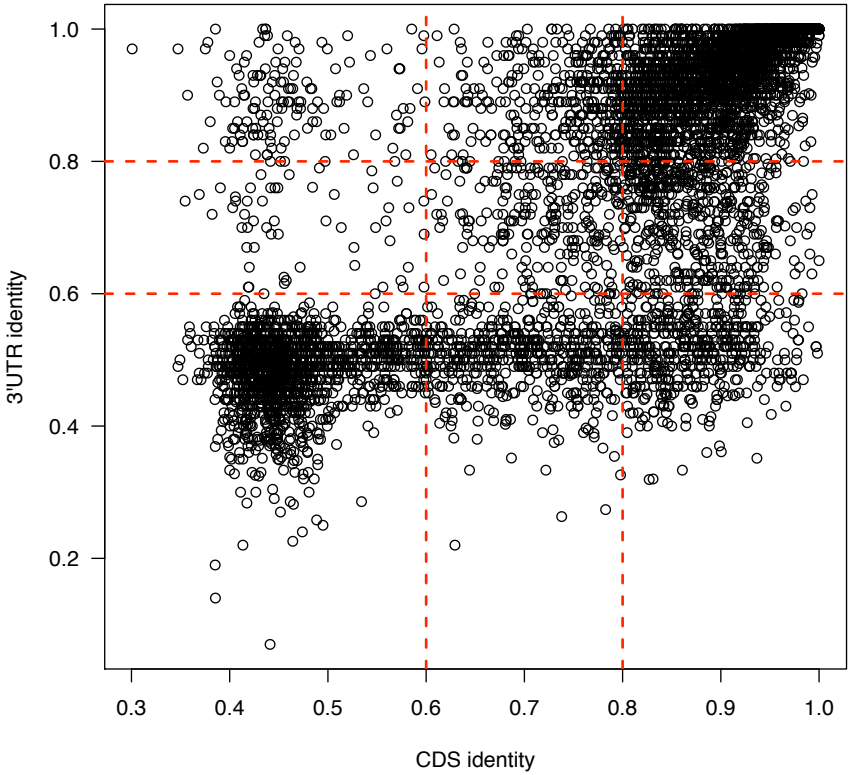


CDS Comparison (2)



CDS vs UTR


Processed Pseudogenes
Duplicated Pseudogenes
Frequency of Processed Pseudogenes
Frequency of Duplicated Pseudogenes
%Sequence Similarity
0
200
400
600
800
1000
1200
1400
100
25
50
75

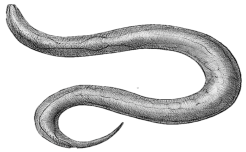

3'UTR identity
CDS identity
0.2
0.4
0.6
0.8
1.0
0.3
0.5
0.7
0.9

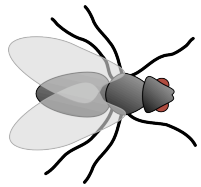
TRANSCRIPTION FACTORS

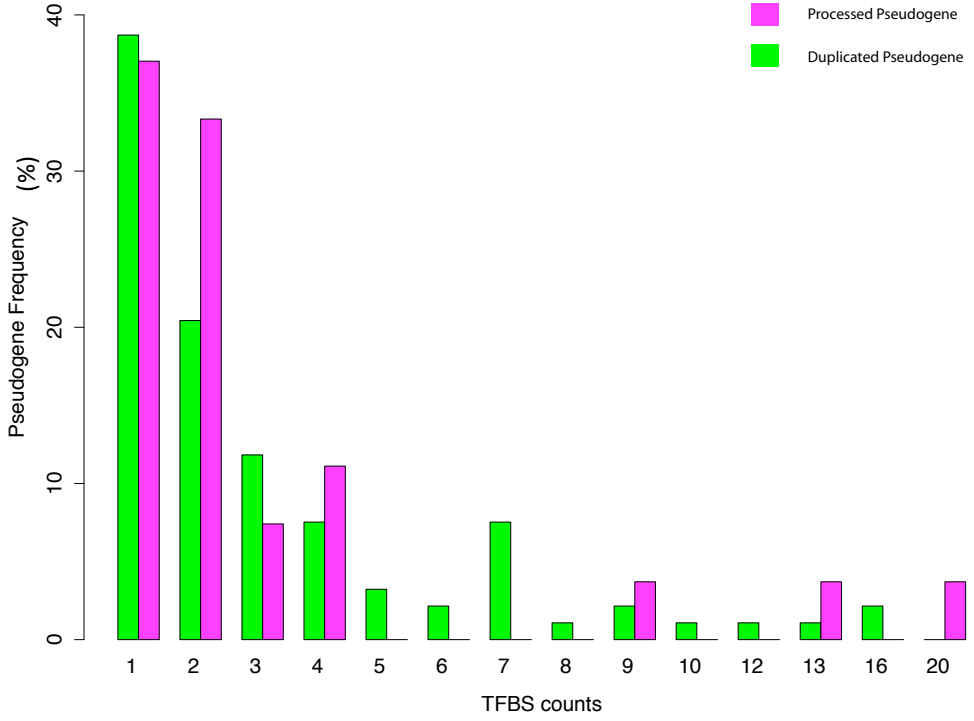
Indicator of potential activity:

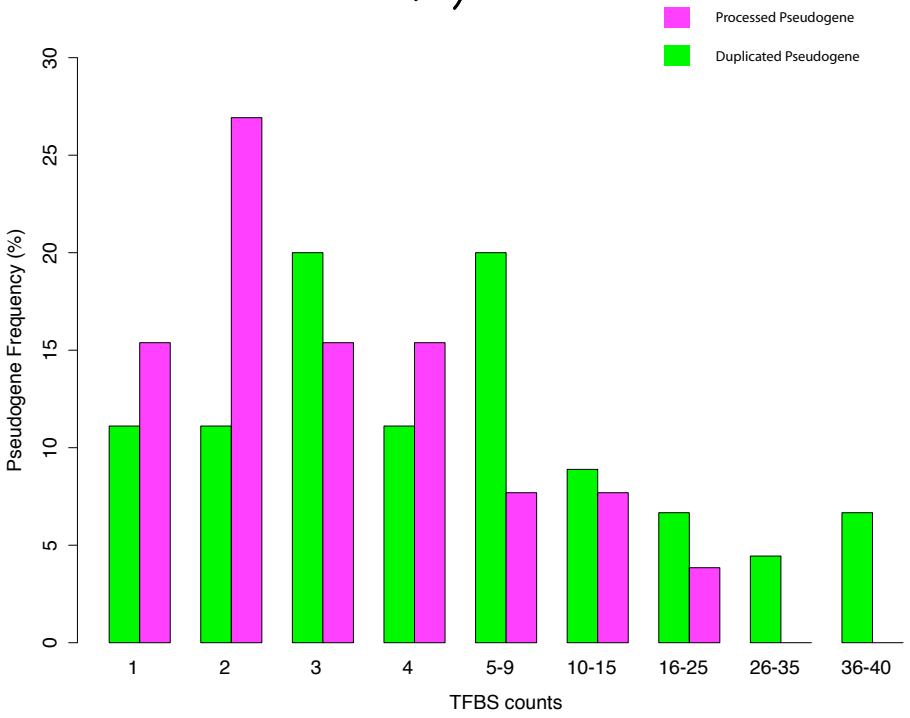
- Worm: 47 TFs within 2kb upstream of the pseudogenes start site
- Fly: 36 TFs within 2kb upstream of the pseudogenes start site

TFBS Distribution

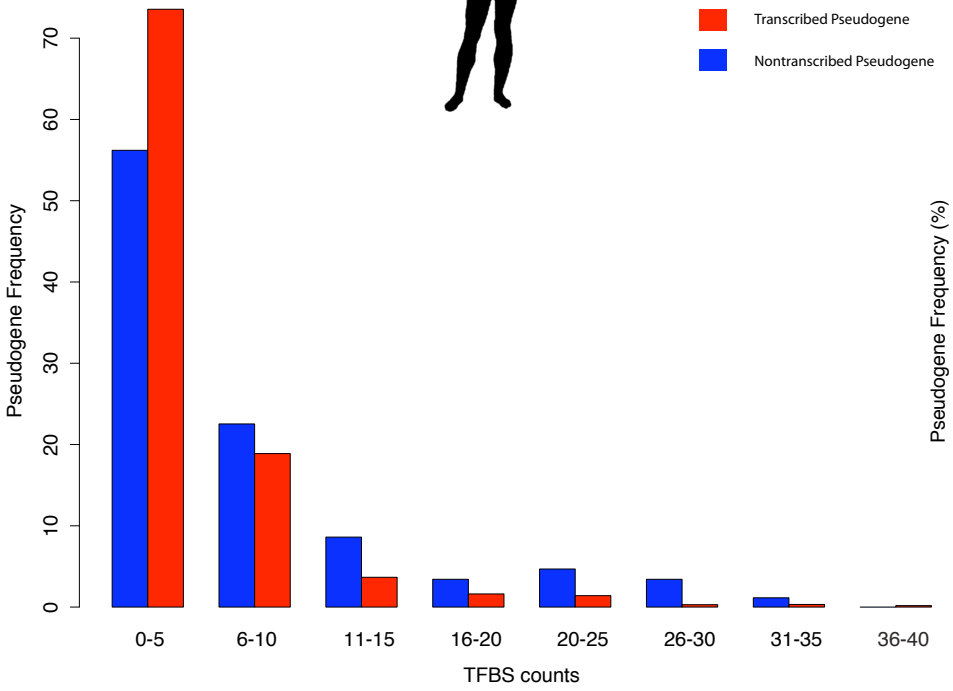


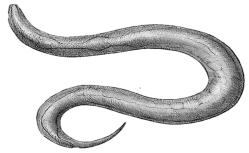


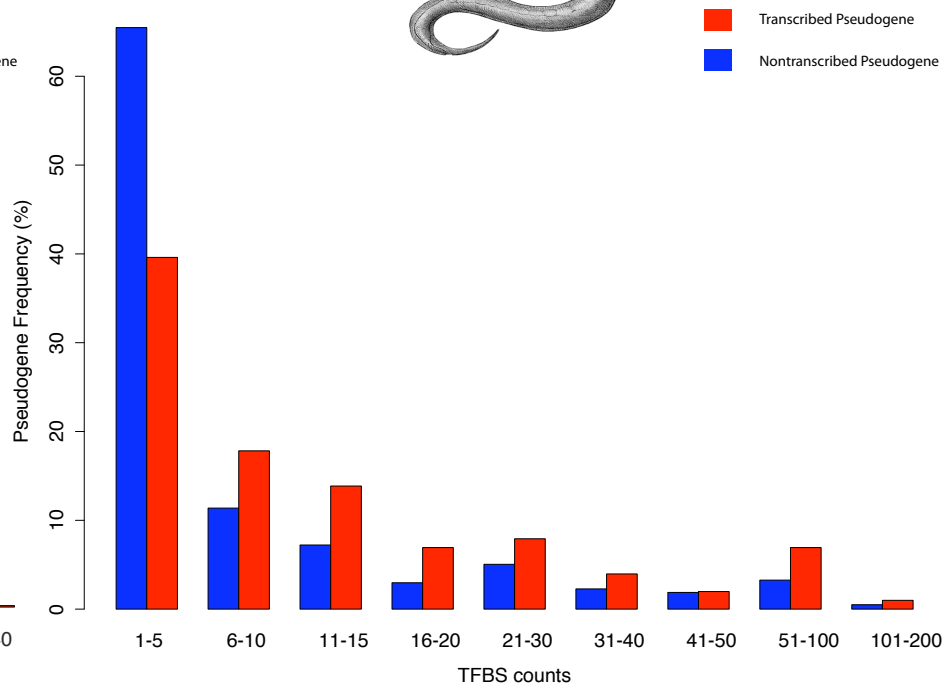




TFBS Distribution (2)







CHROMATIN FEATURES

- Open Chromatin
- Histone Modifications
- Segmentation

Histone Modifications

- Fly data:
  - Chip-chip: 30 assay factors
  - Chip-seq: 6 assay factors
- Worm data:
  - Chip-seq: 4 assay factors
  - Chip-chip: 21 assay factors

- Lesson from humans:
  - Histone modifications alone do not provide a direct understanding of pseudogene activity

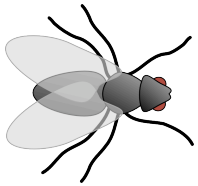
=> use SEGMENTATION

Unsupervised pattern discovery in human chromatin structure through genomic segmentation

Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes & William Stafford Noble

Affiliations | Contributions | Corresponding author

Nature Methods **9**, 473–476 (2012) | doi:10.1038/nmeth.1937
Received 01 July 2011 | Accepted 14 February 2012 | Published online 18 March 2012



Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*

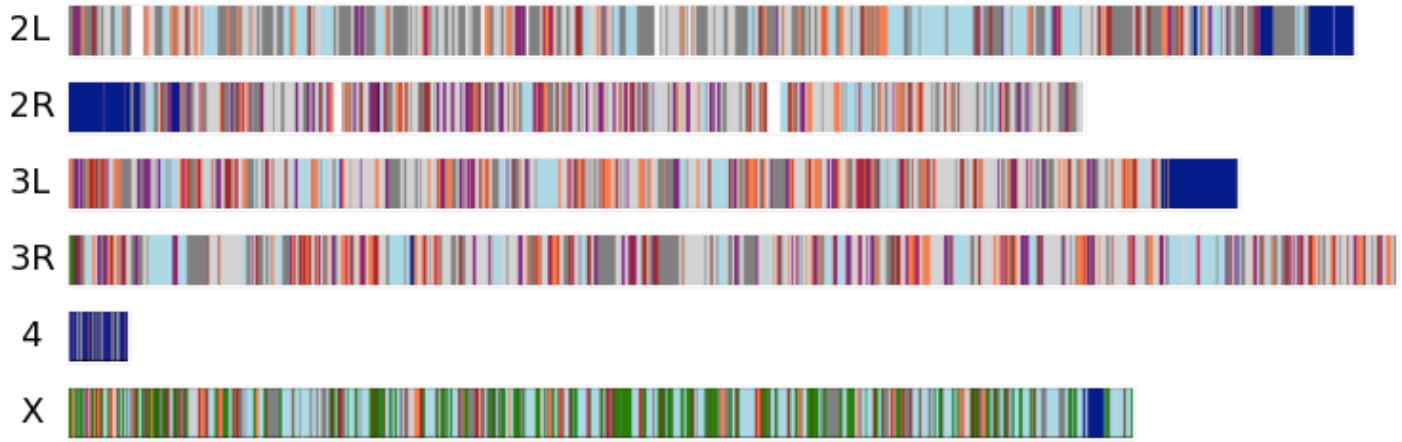
Peter V. Kharchenko, Artyom A. Alekseyenko, Yuri B. Schwartz, Aki Minoda, Nicole C. Riddle, Jason Ernst, Peter J. Sabo, Erica Larschan, Andrey A. Gorchakov, Tingting Gu, Daniela Linder-Basso, Annette Plachetka, Gregory Shanower, Michael Y. Tolstorukov, Lovelace J. Luquette, Ruibin Xi, Youngsook L. Jung, Richard W. Park, Eric P. Bishop, Theresa K. Canfield, Richard Sandstrom, Robert E. Thurman, David M. MacAlpine, John A. Stamatoyannopoulos, Manolis Kellis + *et al.*

Affiliations | Contributions | Corresponding authors

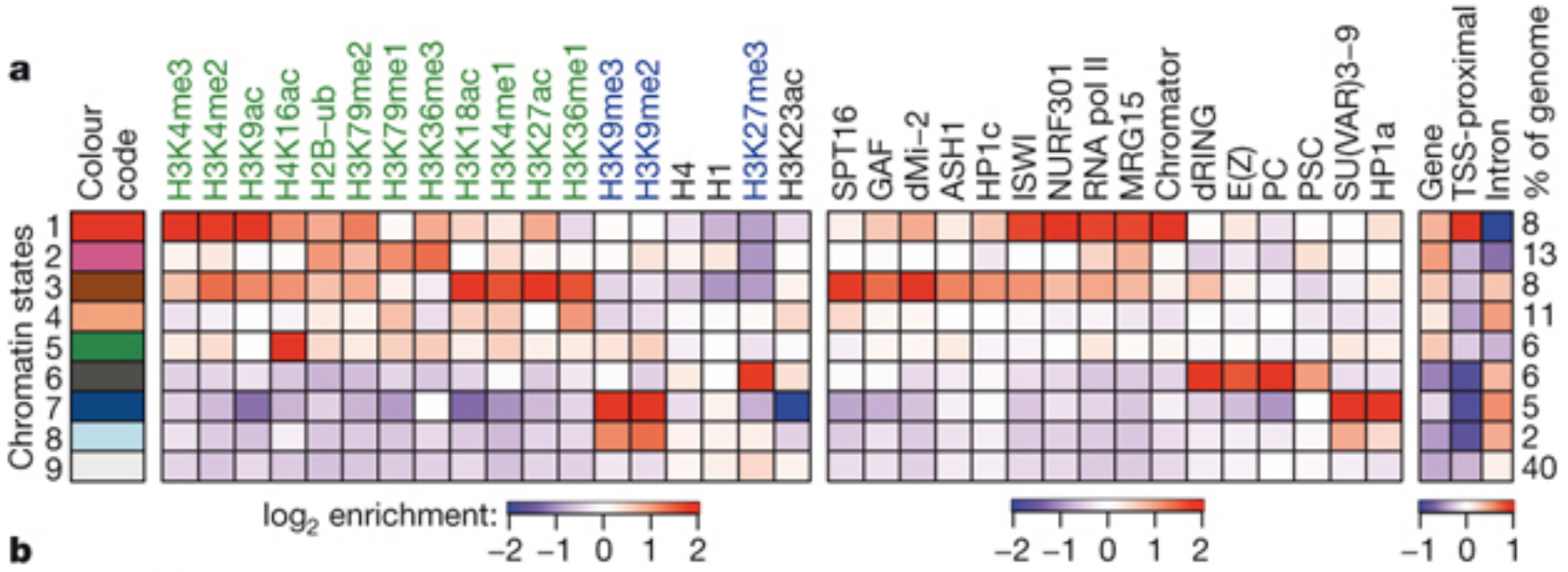
Nature **471**, 480–485 (24 March 2011) | doi:10.1038/nature09725
Received 02 September 2010 | Accepted 06 December 2010 | Published online 22 December 2010
| Corrected online **23 March 2011**

Fly Chromatin Segmentation

- 9 chromatin states:
  - Promoter and TSS
  - Transcription elongation
  - Regulatory regions (enhancers)
  - Active introns
  - Active genes on the Male X
  - Polycomb-mediated repression
  - Pericentromeric heterochromatin
  - Heterochromatin-like embedded in euchromatin
  - Transcriptionally silent, intergenic



Fly Chromatin Segmentation



- A 9-state model of prevalent chromatin states found in S2 and BG3 cells.
- Each chromatin state (row) is defined by a combinatorial pattern of enrichment (red) or depletion (blue) for specific chromatin marks (first panel, columns; active marks in green, repressive in blue).
- The second panel shows average enrichment of chromosomal proteins.
- The third panel shows fold over/under-representation of genic and TSS-proximal (±1 kb) regions relative to the entire tiled genome. The enrichment of intronic regions is relative to genic regions associated with each state.

PV Kharchenko *et al. Nature* **000**, 1-6 (2010) doi:10.1038/nature09725

TRANSCRIPTIONAL ACTIVITY

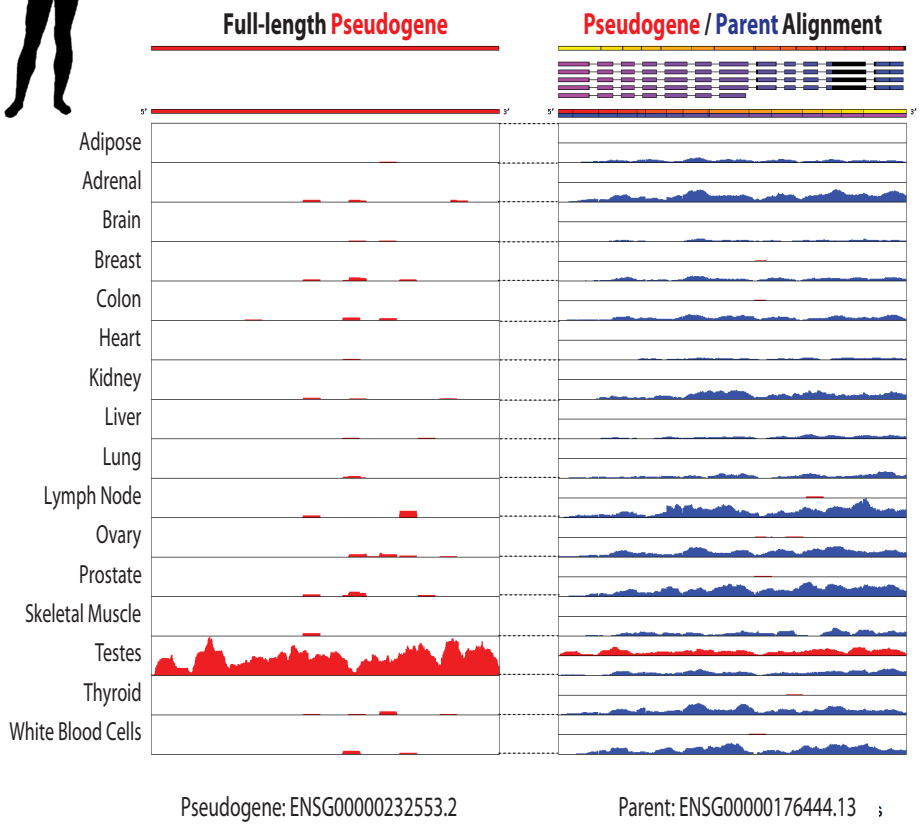
Human: - EST evidence
- Human BodyMap
- Encode cell lines

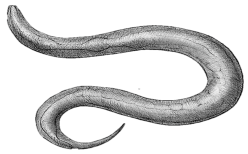
Worm: - RNAseq data

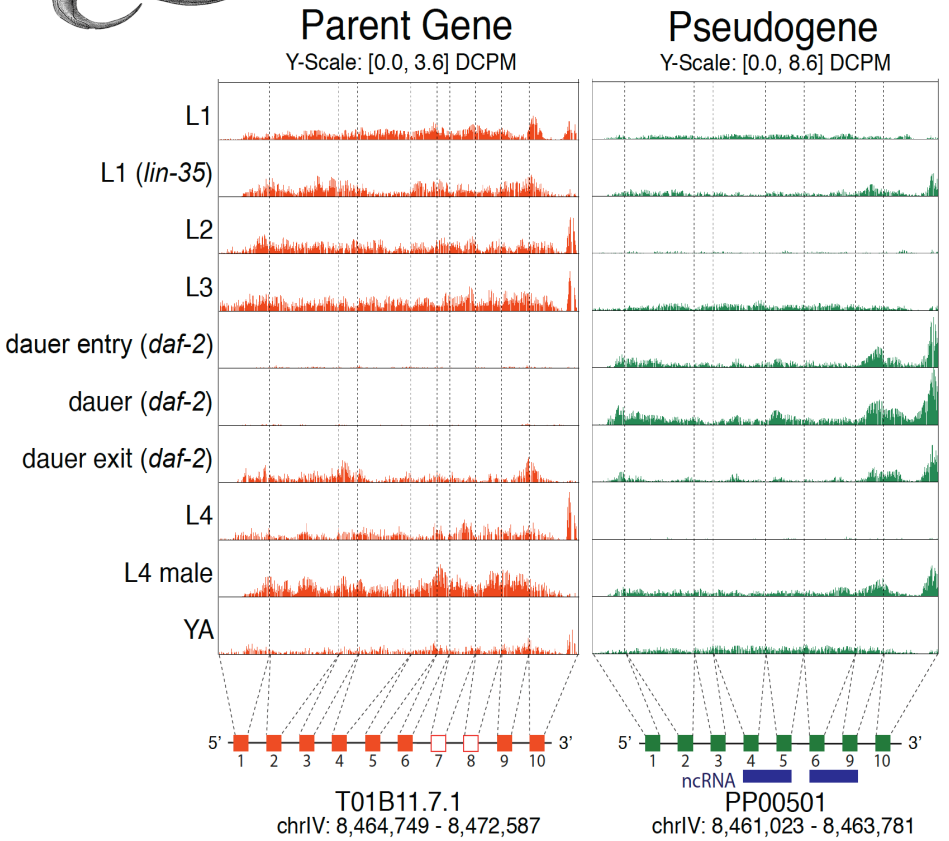
Transcribed Pseudogenes

- **Human** – 876 transcribed pseudogenes:
  - 422 from EST evidence
  - 344 from PseudoSeq on BodyMap data
  - 110 from total RNA data from ENCODE cell lines
- **Worm WS201:**
  - 323 transcribed pseudogenes – using PseudoSeq
    - 191 have strong evidence
    - 132 have concordant expression patterns with the parents
- **Worm WS220**:
  - **148 transcribed pseudogenes with strong evidence**
- **Fly** – to be determined

Transcribed Pseudogenes





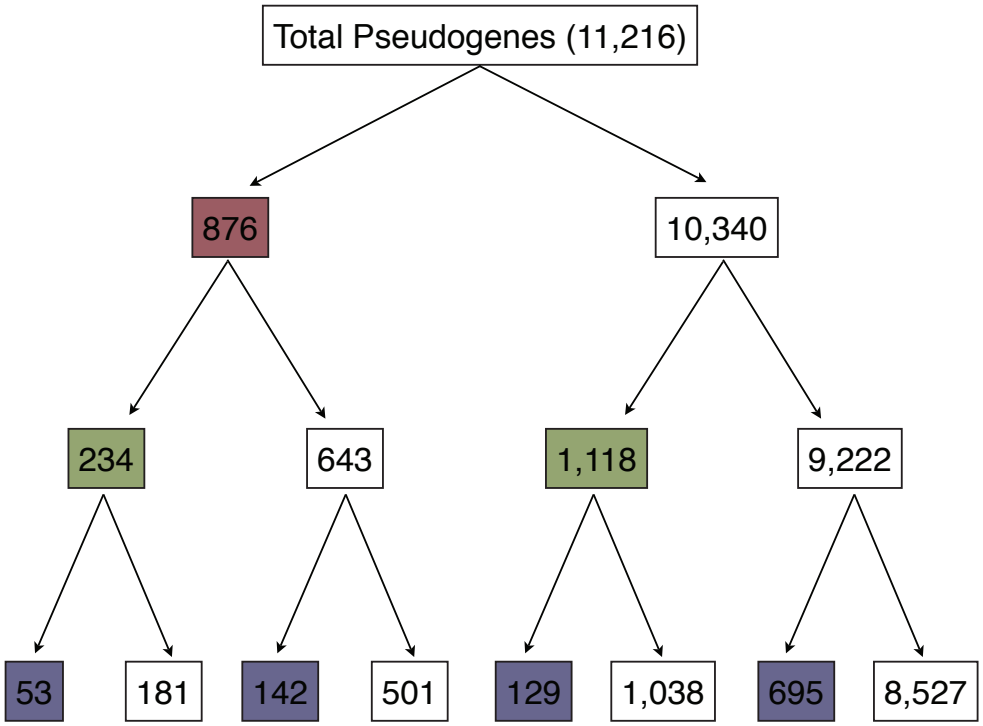


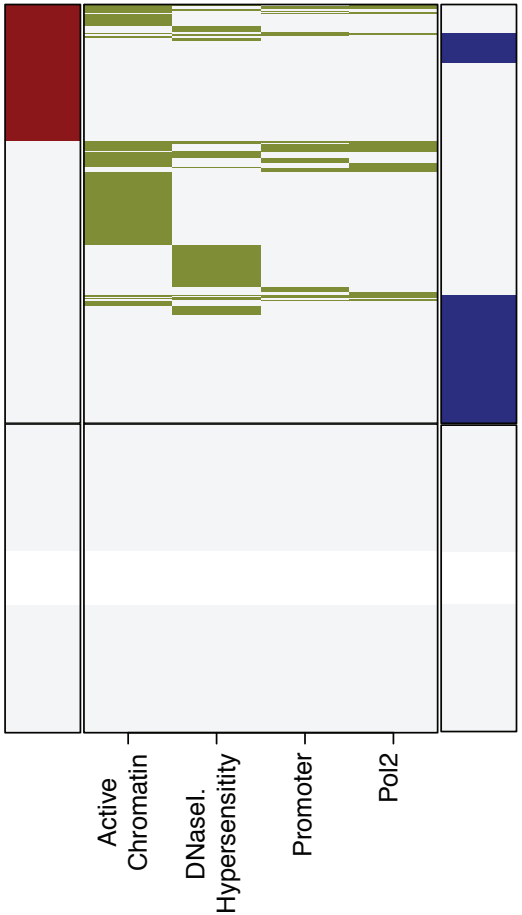
Pseudogene Activity

Transcribed?
Yes No

Additional Activity?
Yes No

Constrained?
Yes No





Data Summary

Data	Worm	Fly	Human
Annotation	Yes	Yes	Yes
* Pseudogenes	Yes	Yes	Yes
* Parent Genes	Yes	Yes	Yes
* UTR	No	Yes	Yes
Transcription Evidence	Yes	Not-yet	Yes
TFs	Yes	Yes	Yes
Pol2	Yes	Yes	Yes
Promoter	No	No	Yes
DnaseI Hypersensitivity	No	No	Yes
Segmentation	No	Yes	Yes
Histone Modifications	Yes	Yes	Yes
Developmental Stages	Yes	Yes	No

Conclusions & Future Work

- Human Pseudogenes available in GENCODE v7
  - To be updated to GENCODE v10
- Creation of a comprehensive & reliable data set for Worm & Fly pseudogenes:
  - Level 1: validated using both manual and automated annotation
  - Level 2: validated manually only
  - Level 3: validated using computational pipelines only
- Sequence similarity:
  - Worm processed pseudogens have similar pattern distribution to Human analogues,
  - Worm duplicated pseudogenes peak around 50% similarity to parents
  - Pseudogenes with high sequence similarity to parents are less frequent in worm than in human
- TFBS
  - Similar distribution for worm and human for transcribed vs non transcribed pseudogenes
  - Higher number of worm TFBS in the up
- To Do:
  - Paralogues comparison for Human/Wrom/Fly
  - Upstream Elements: analyse Pol2 Binding & Promoters
  - Catalog transcribed pseudogenes in worm & fly

Acknowledgement

Becky

Baikang

Mark

All of you for listening!