Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

Mark Gerstein, Yale

Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein). See last slide for more info.

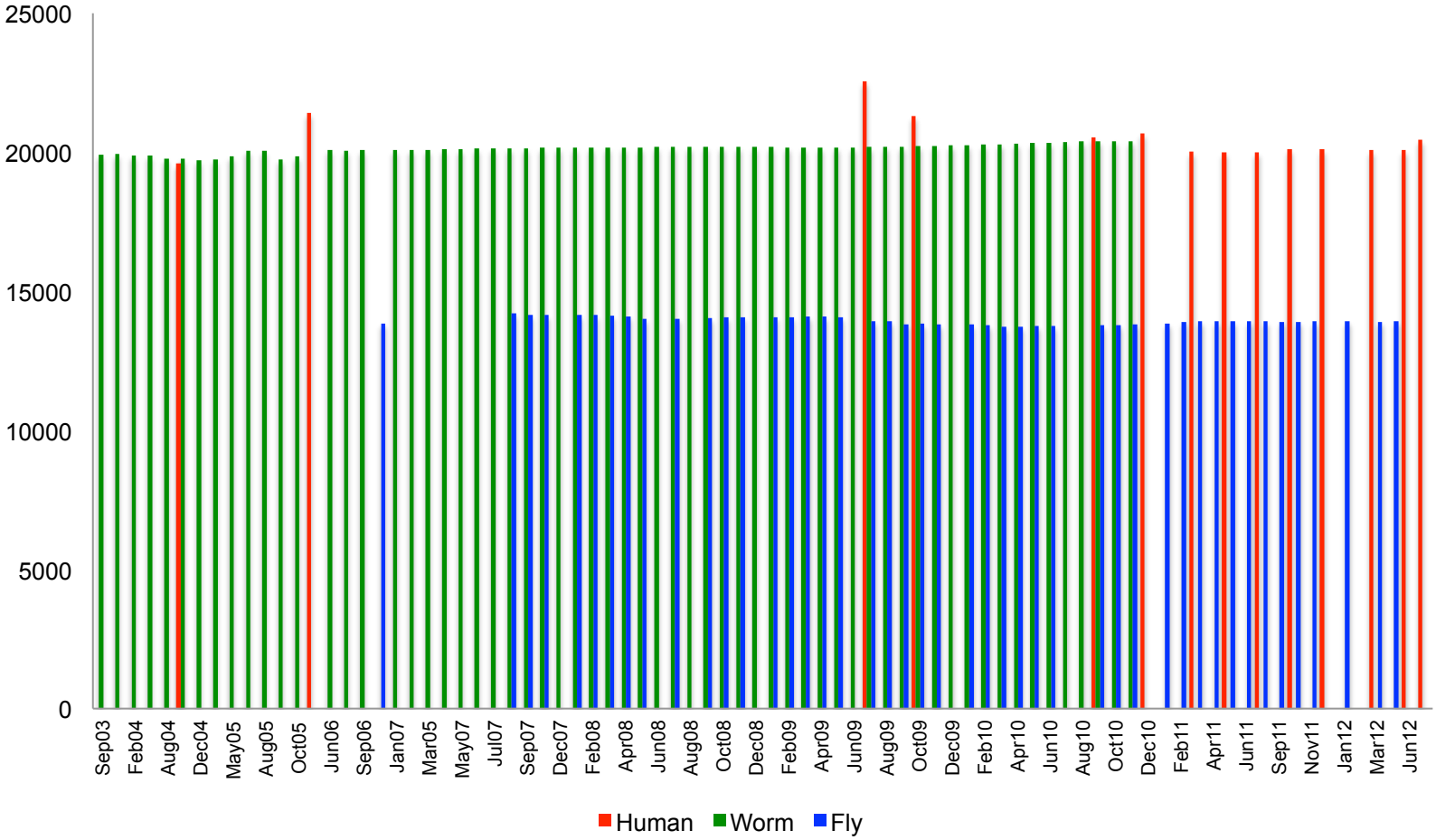
Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - The advent of Nextgen seq.
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - **The advent of Nextgen seq.**
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

During the genome annotation era, protein-coding gene counts in worm, fly & human have remained fairly constant



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

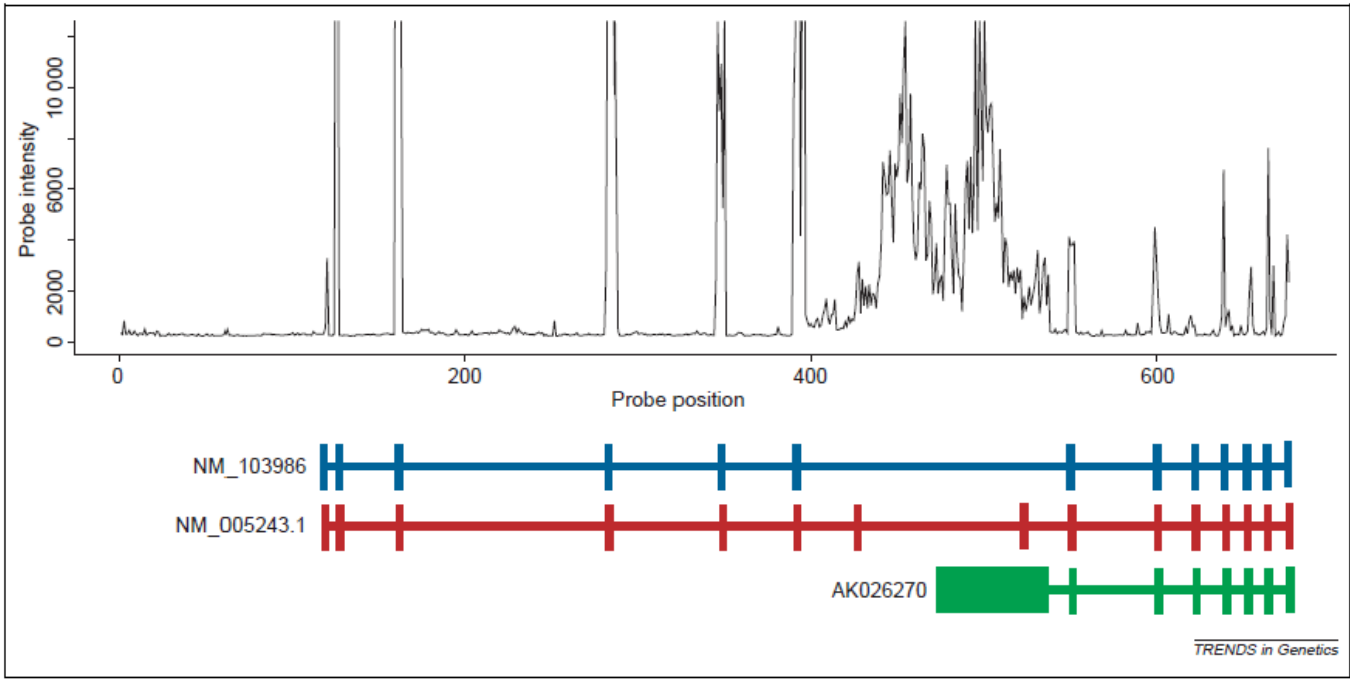
Discovery of Pervasive Transcription: Dark Matter of the Genome

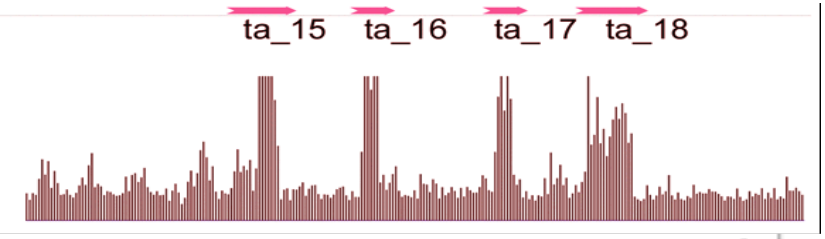
- With the advent of custom tiling arrays it was shown that a significant portion of the human genome (outside known protein coding genes) is transcribed
  - Chr 21/22:
    Kapranov et al….Gingeras (‘02) Science
    - “When compared with the sequence annotations available for these chromosomes, it is noted that as much as an order of magnitude more of the genomic sequence is transcribed than accounted for by the predicted and characterized exons.”
  - Also: Rinn et al… Snyder (‘03) Genes & Dev.
  - Whole Genome:
    Bertone et al. (‘04) Science & Cheng et al. (‘05) Science

[Gravitational lensing by dark matter in Abell 1689 – HST (NASA, ESA)]

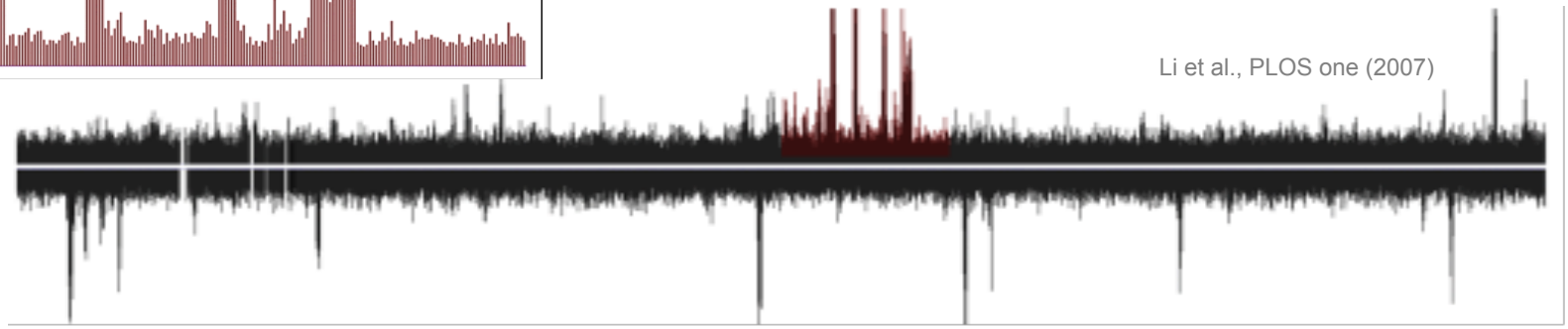
Noisy Raw Signal from Tiling Arrays (Transcription)

Johnson et al. (2005) TIG, 21, 93-102.



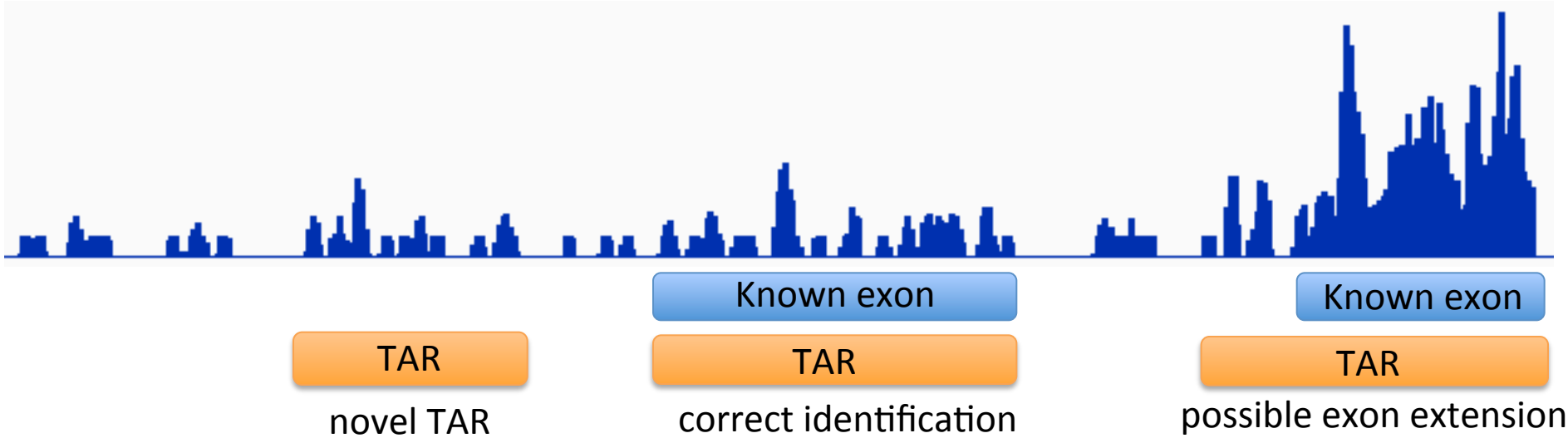


Li et al., PLOS one (2007)



TARs (novel RNA contigs) from Segmenting Transcriptional Signal

- Cluster reads setting minimum-run and maximum gap parameters for newly identified transcribed regions (TARs) [called TransFrags by Gingeras et al.]

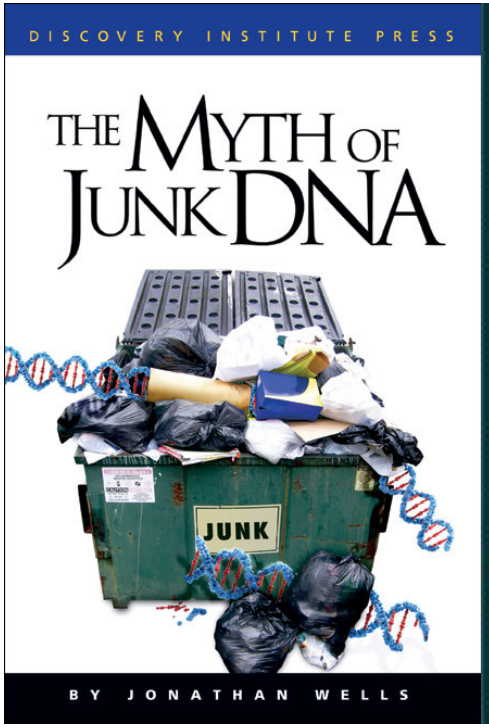


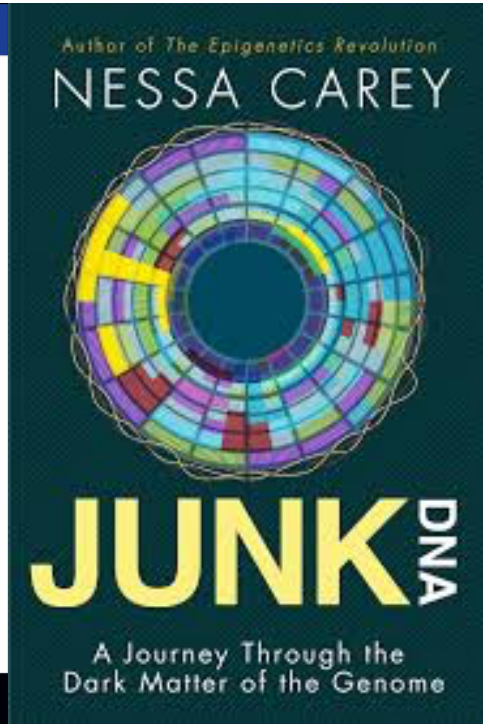
ENCODE RNA-Seq Data

Controversy of Pervasive Transcription

Over the last decade this result has been somewhat controversial (Clark et al ('11) PLoS Bio)

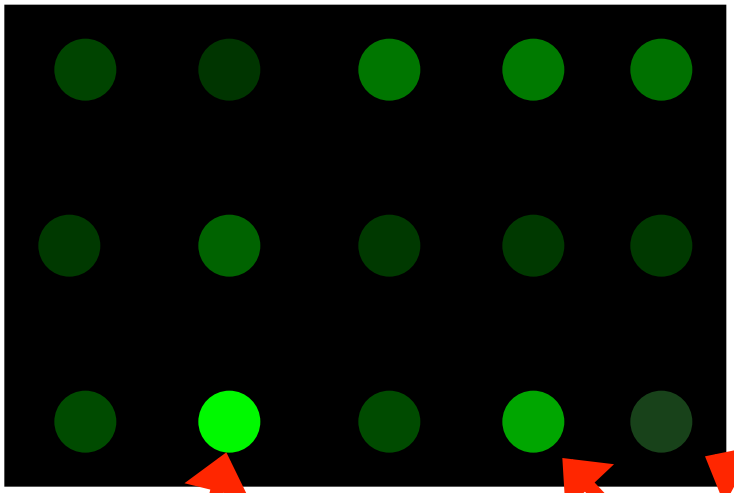
"Current estimates indicate that only about 1.2% of the mammalian genome codes for amino acids in proteins. However, mounting evidence over the past decade has suggested that the vast majority of the genome is transcribed, well beyond the boundaries of known genes, a phenomenon known as pervasive transcription. Challenging this view, an article published in PLoS Biology by van Bakel et al. **concluded that 'the genome is not as pervasively transcribed as previously reported' and that the majority of the detected low-level transcription is due to technical artefacts** and/or background biological noise."



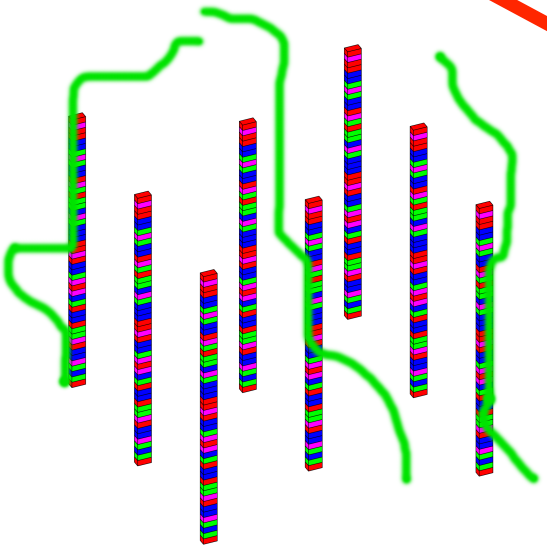


Cross-Hyb. – Specific & Non-specific

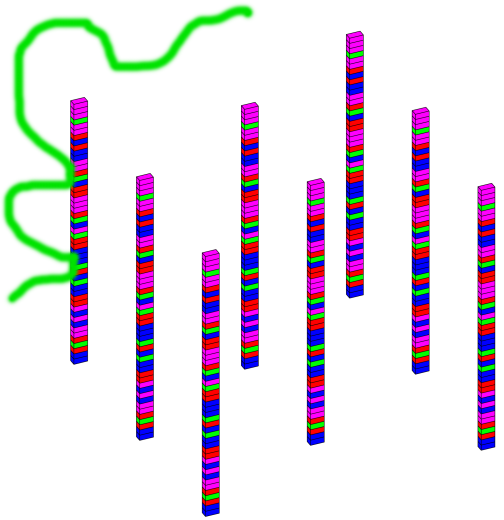
- Perfect match (PM): probe binding intended target
- Specific cross-hyb.: probes binding non-PM targets with a small number of mismatches
- Non-specific cross-hyb.: probes binding targets with many mismatches, due to general stickiness of oligos



Perfect Match

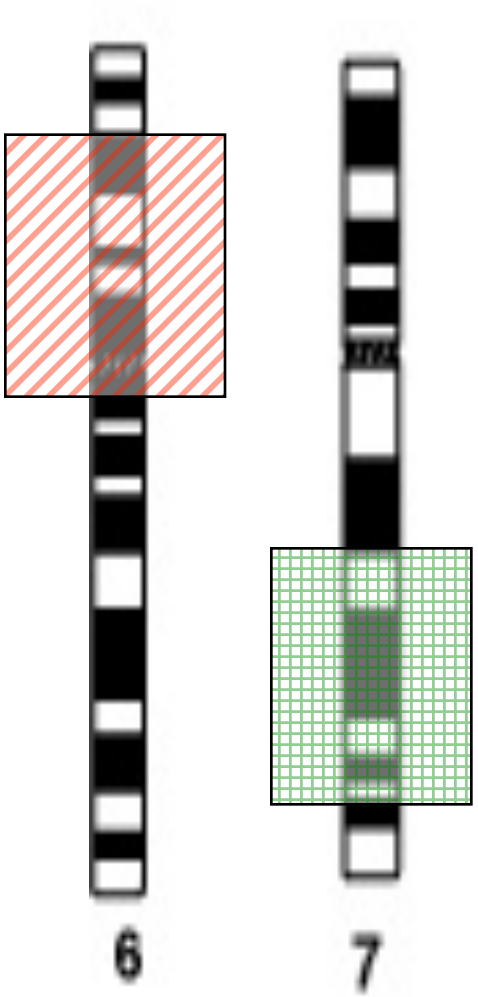


Specific Cross-hyb.



Non-specific Cross-hyb.

Proof of principle test to "exploit" this



- Using Cheng et al. (2005), predict gene expression levels (and profiles across tissues) for genes on part of chr. #6
- ...Based on closest cross-hyb tiles on part of chr. #7
- Then compare to measured levels and profile on #6

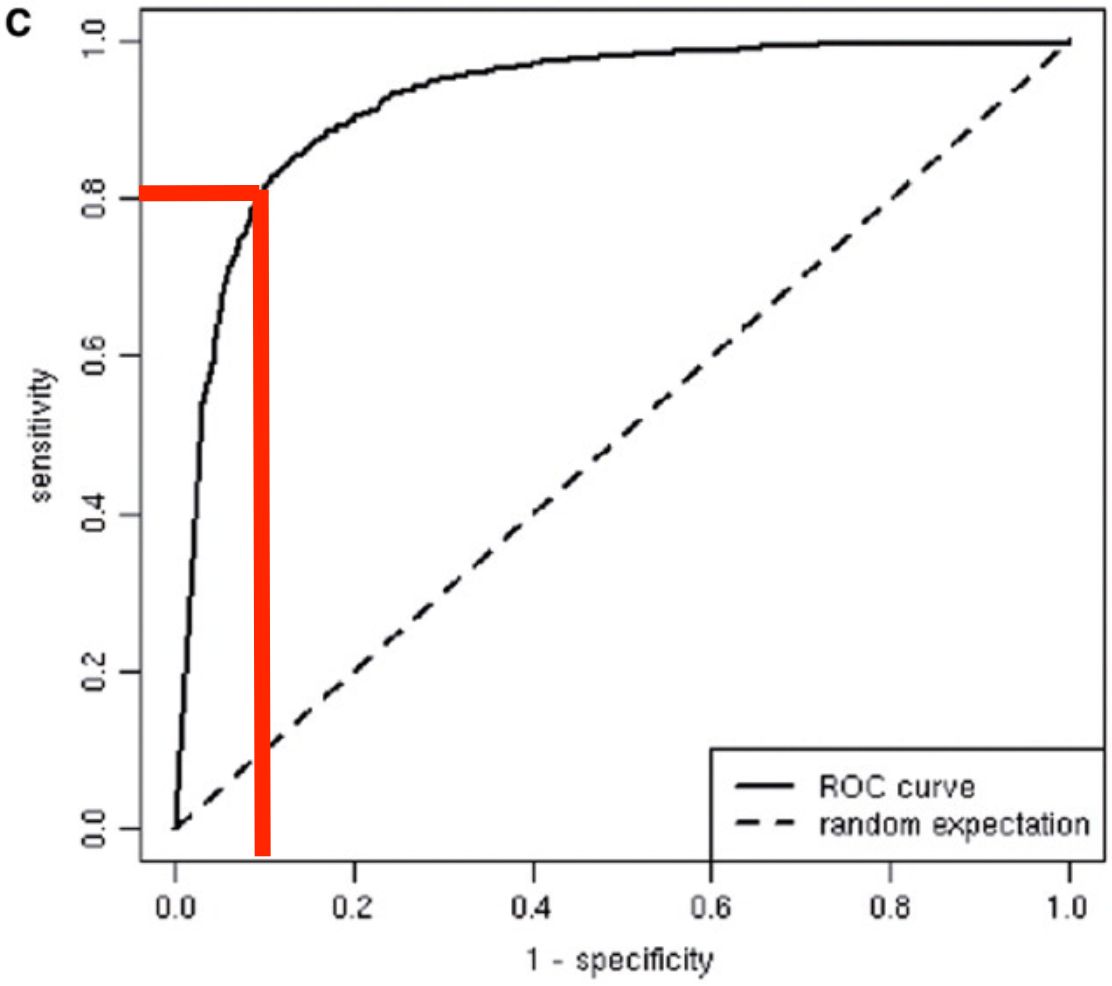
Figure from http://www.members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm

Source: Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

Very Strong ROC Curve: Most genes are accurately detected using nearest-neighbor features' signals

- **Illustrates great magnitude of cross-hyb. on hi-density arrays**

- Gold std. set of known expressed genes. How well do we find.
- A set of known positives was defined as the Refseq genes with at least 75% transfrag coverage. A set of known negatives was constructed by permuting the sequences in the set of known positives. For various thresholds, sensitivity and specificity were computed and then plotted.

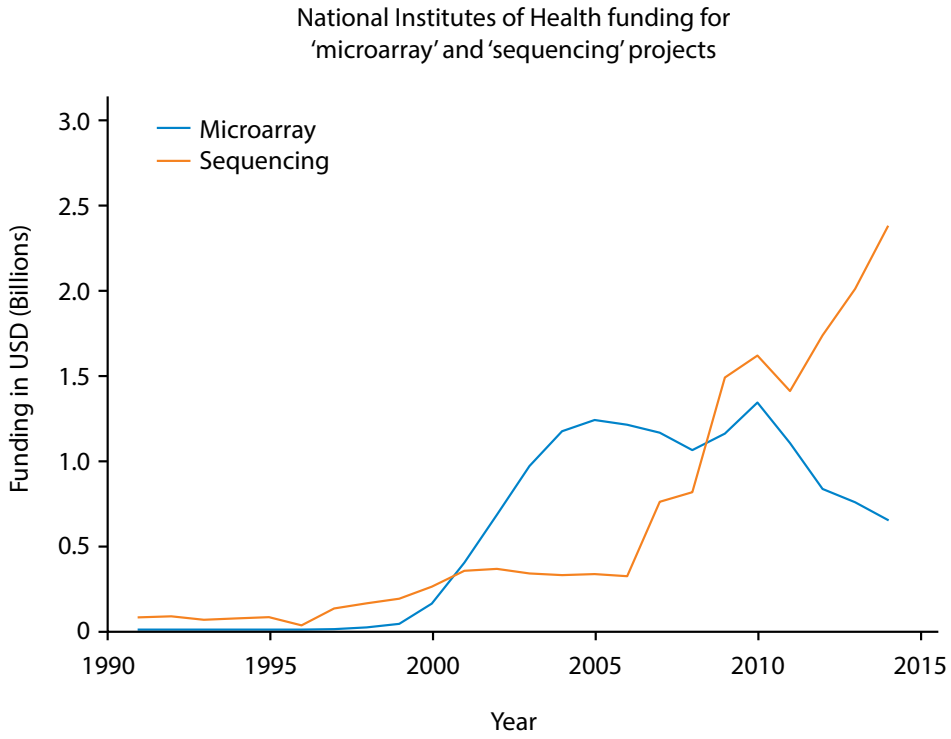


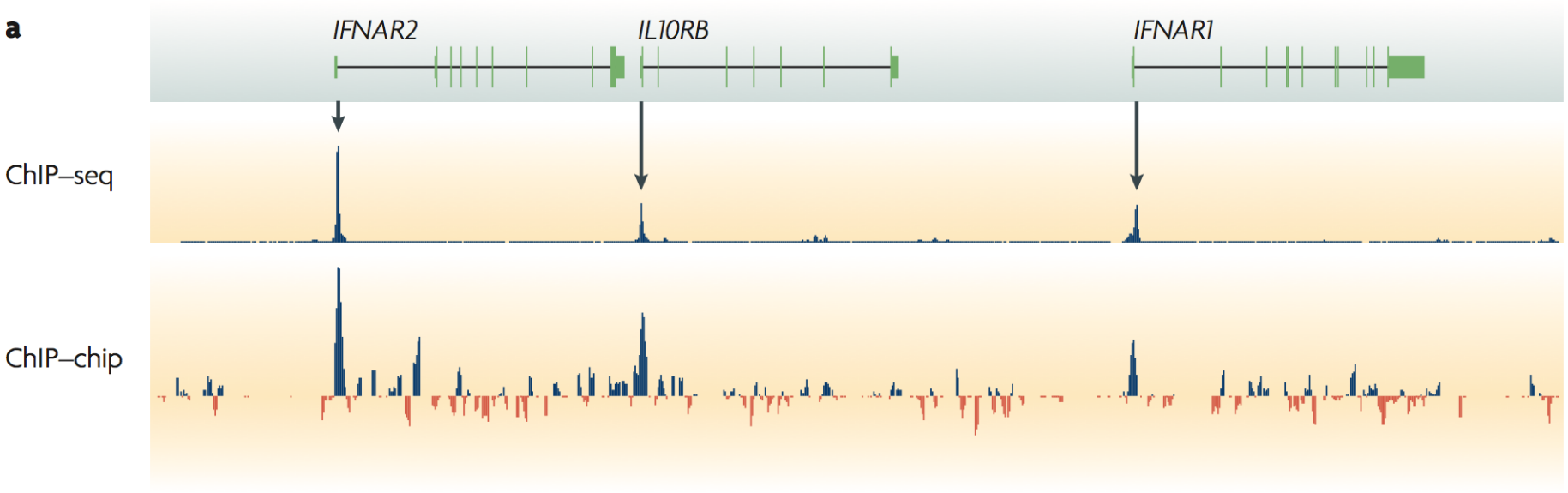
Royce, T. E. et al. Nucl. Acids Res. 2007 35:e99

Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - **The advent of Nextgen seq.**
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

Advent of Next-gen Sequencing: Much Cleaner Signals than Tiling Arrays, Supplanting this Technology

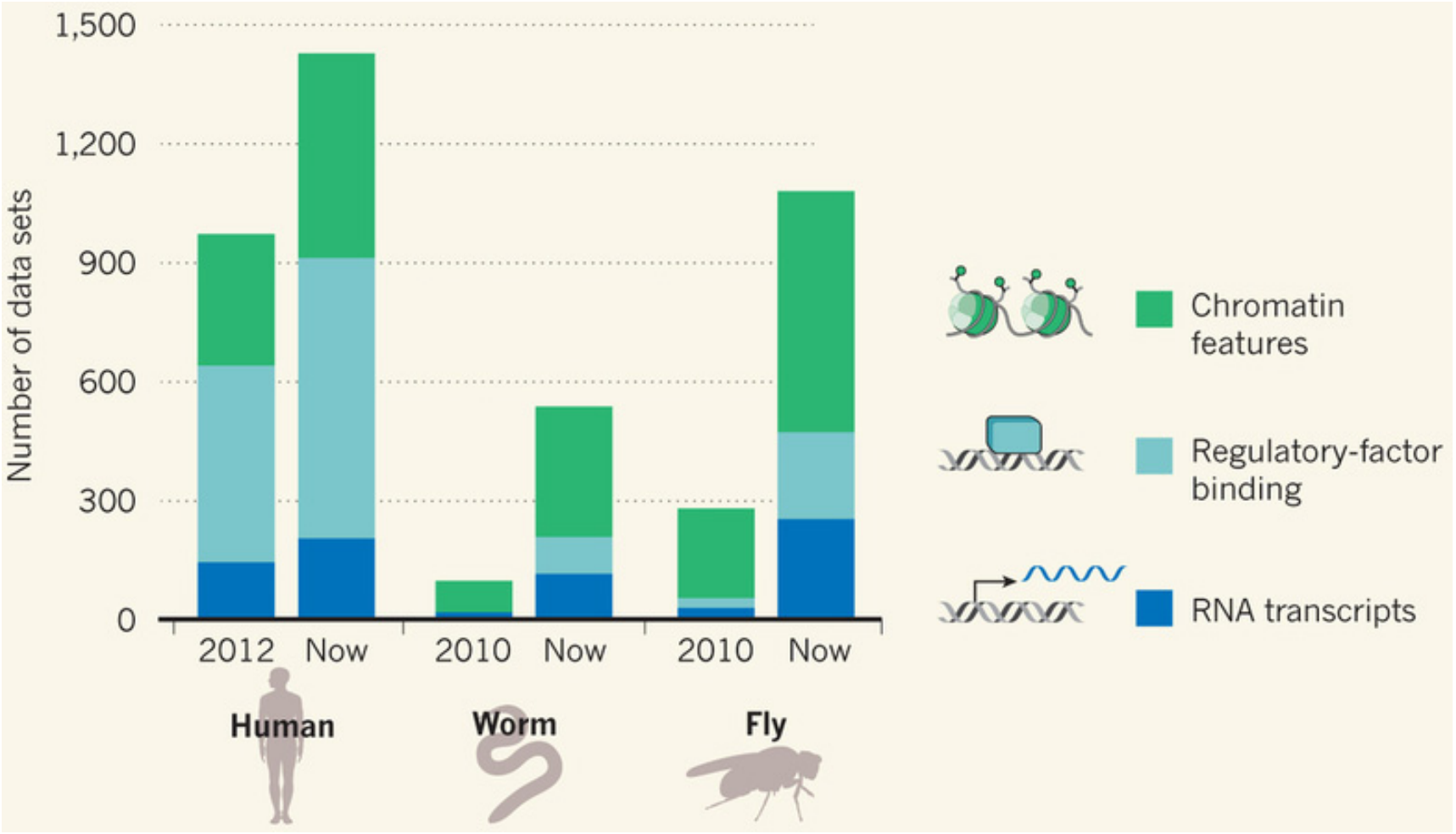




Comparative ENCODE Functional Genomics Resource

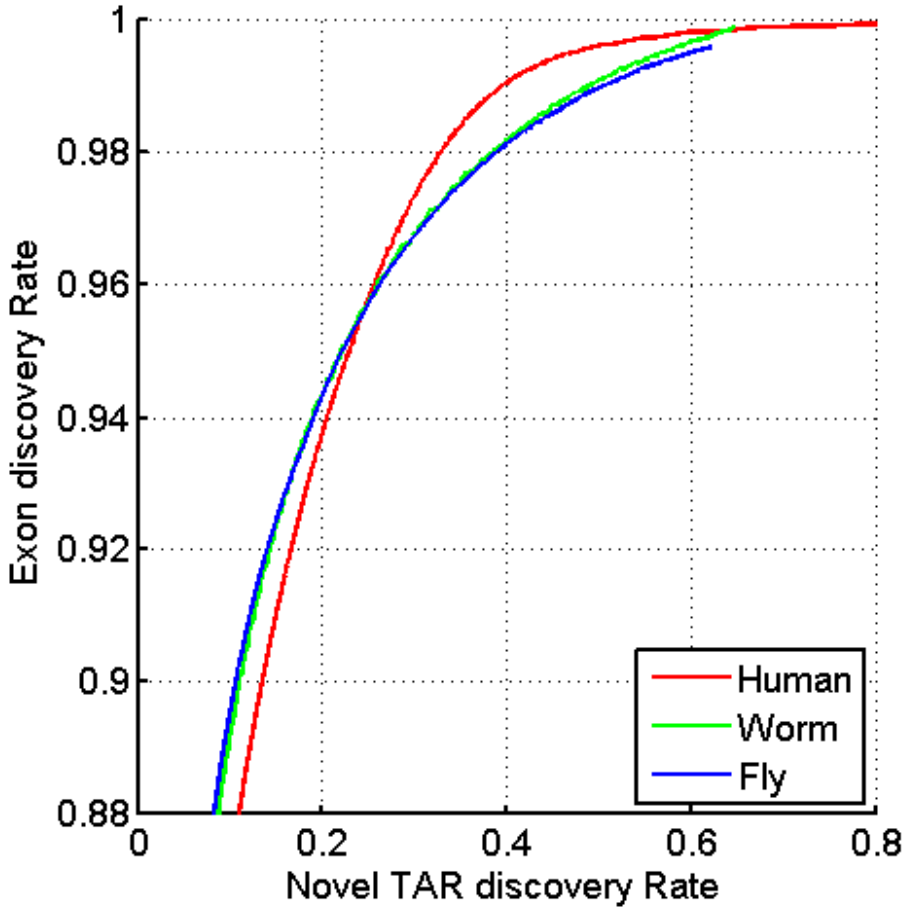
(EncodeProject.org/modENCODE.org)

- 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)



Uniform Annotation of non-coding Elements

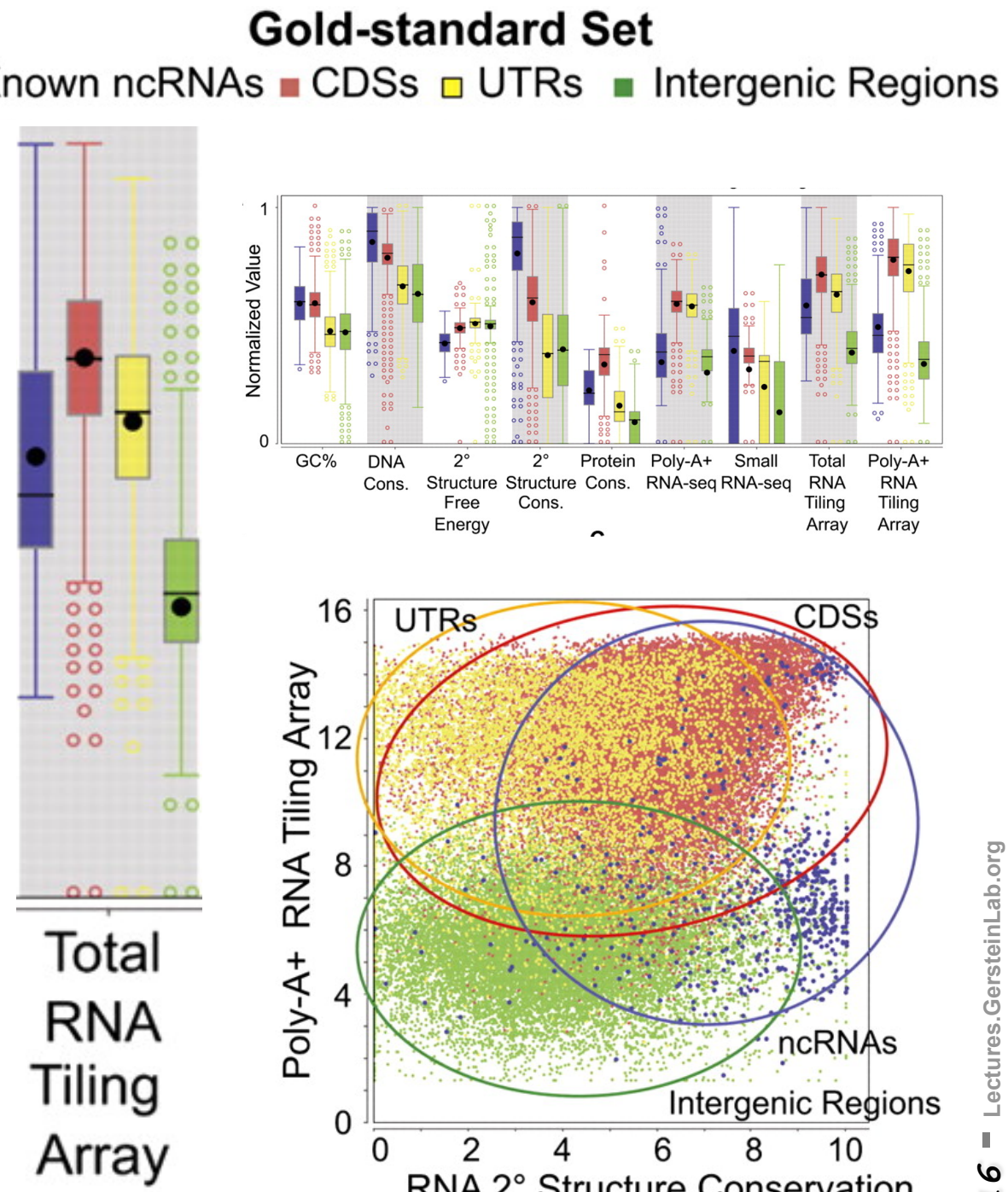
- Uniformly processed the RNA-seq expression compendium



[*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]

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

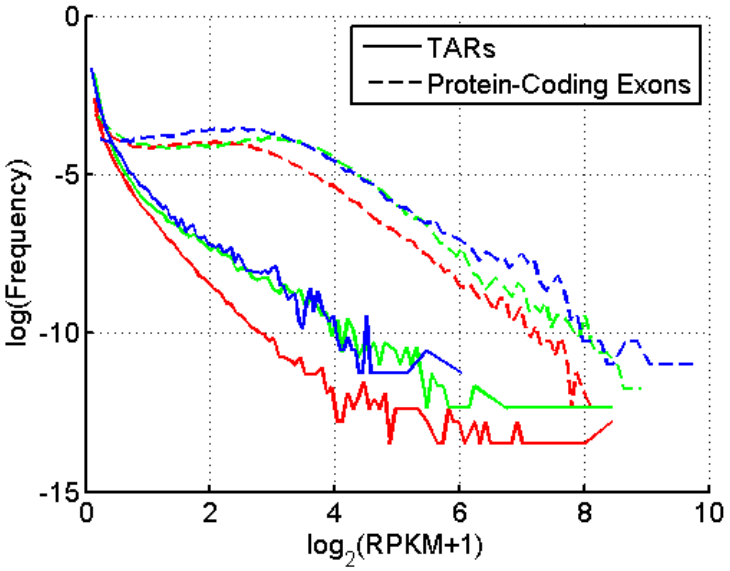
- Similar fraction of non-canonical transcription of non-canonical transcription in human, worm and fly
  - 32-37% of each genome

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

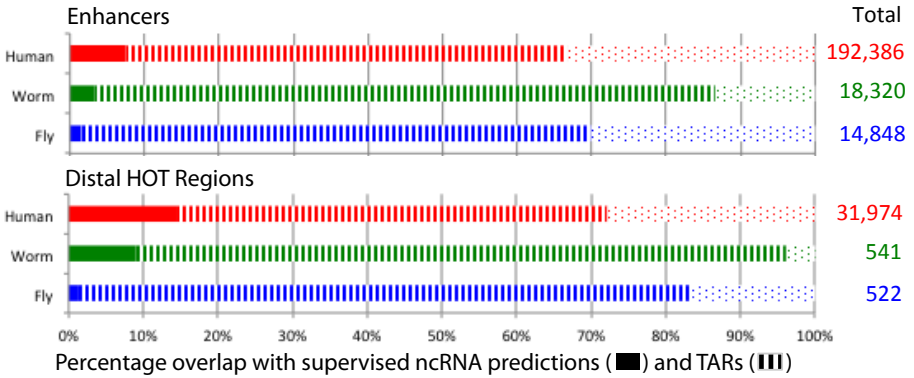
TAR Characterization

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).

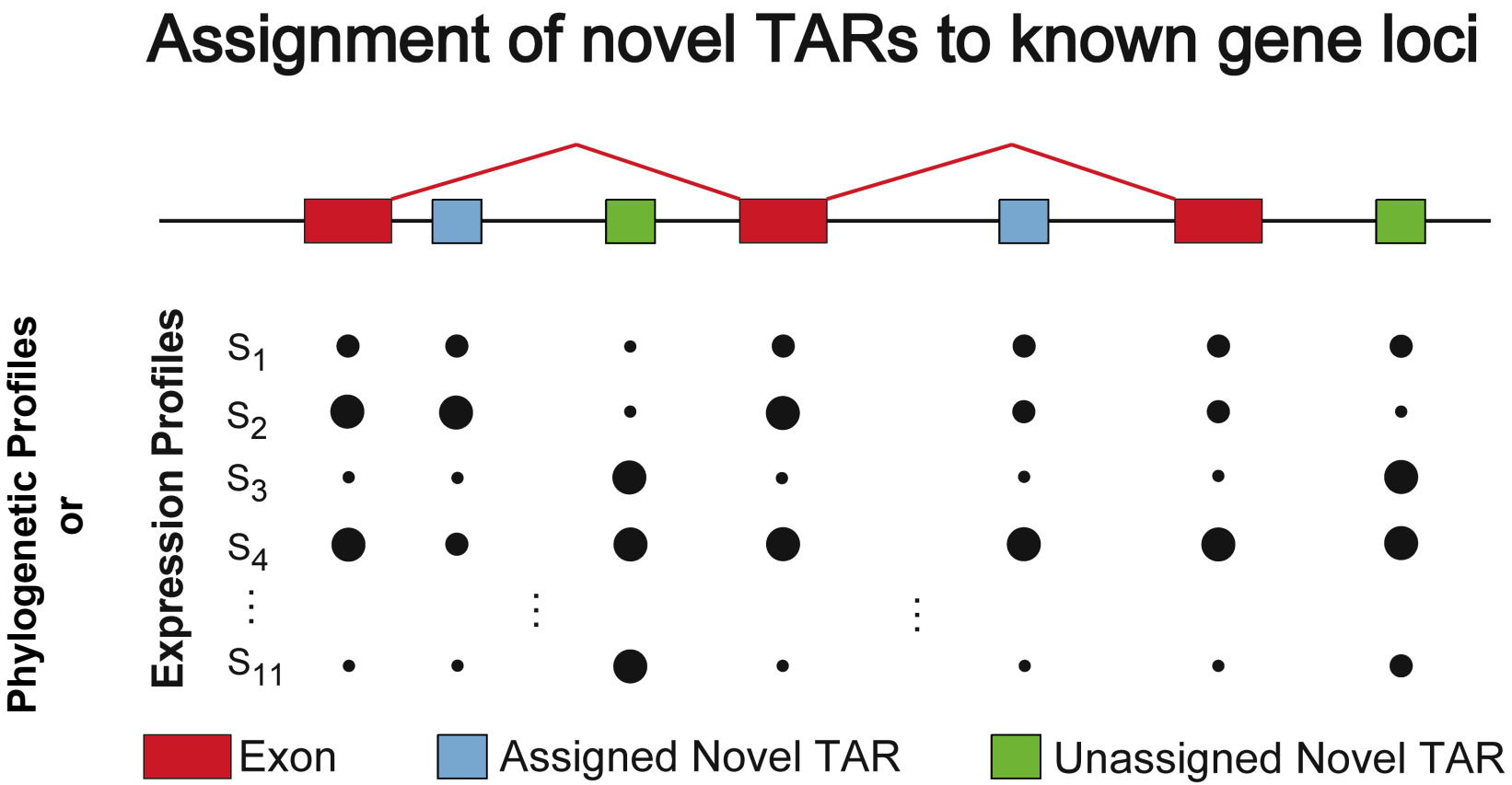


Human, Worm & Fly

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

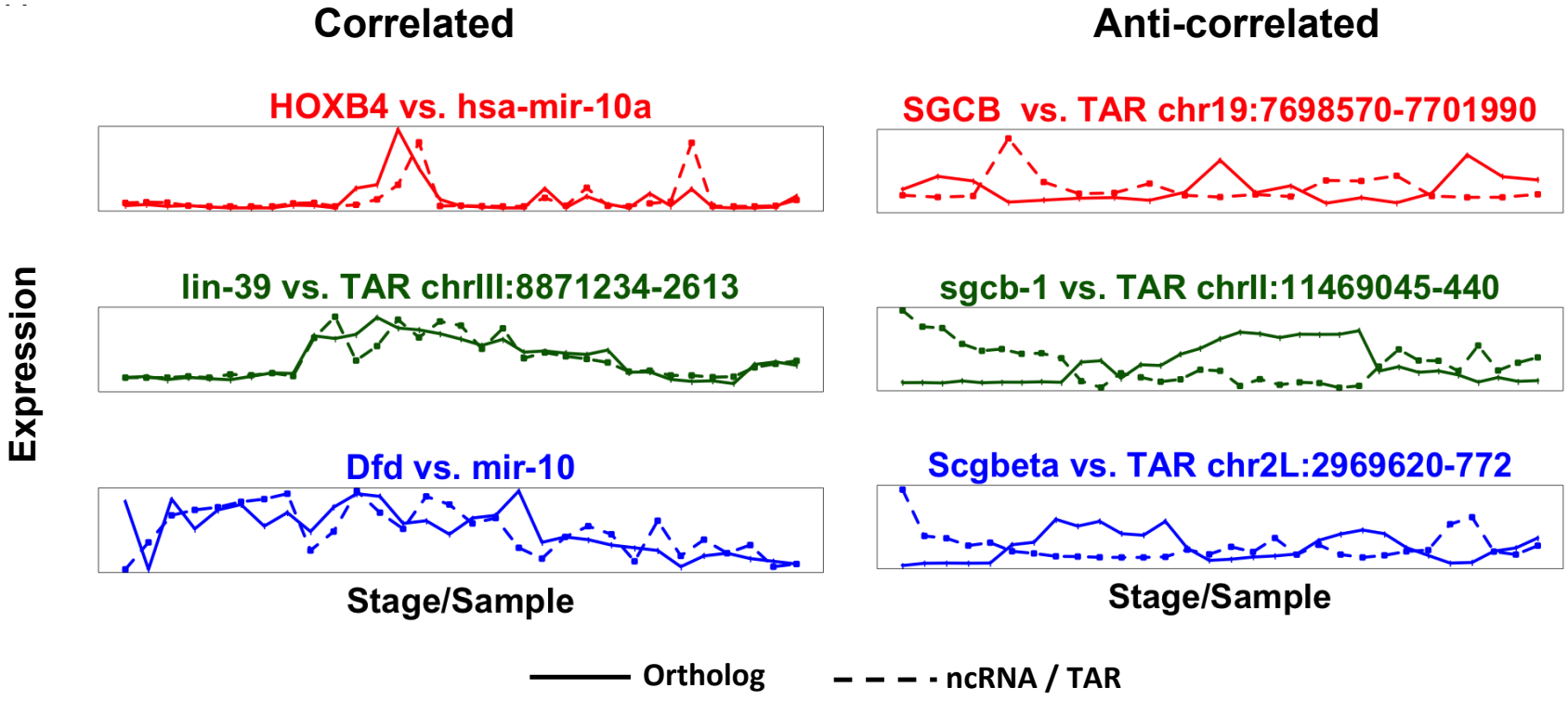
HOT Regions = High TF Co-occupancy

Clustering & Classifying Blocks of Un-annotated Transcription into larger units



Rozowsky et al. Gen. Res. (2007)

ncRNAs/TARS can be clustered with known genes



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

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

Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

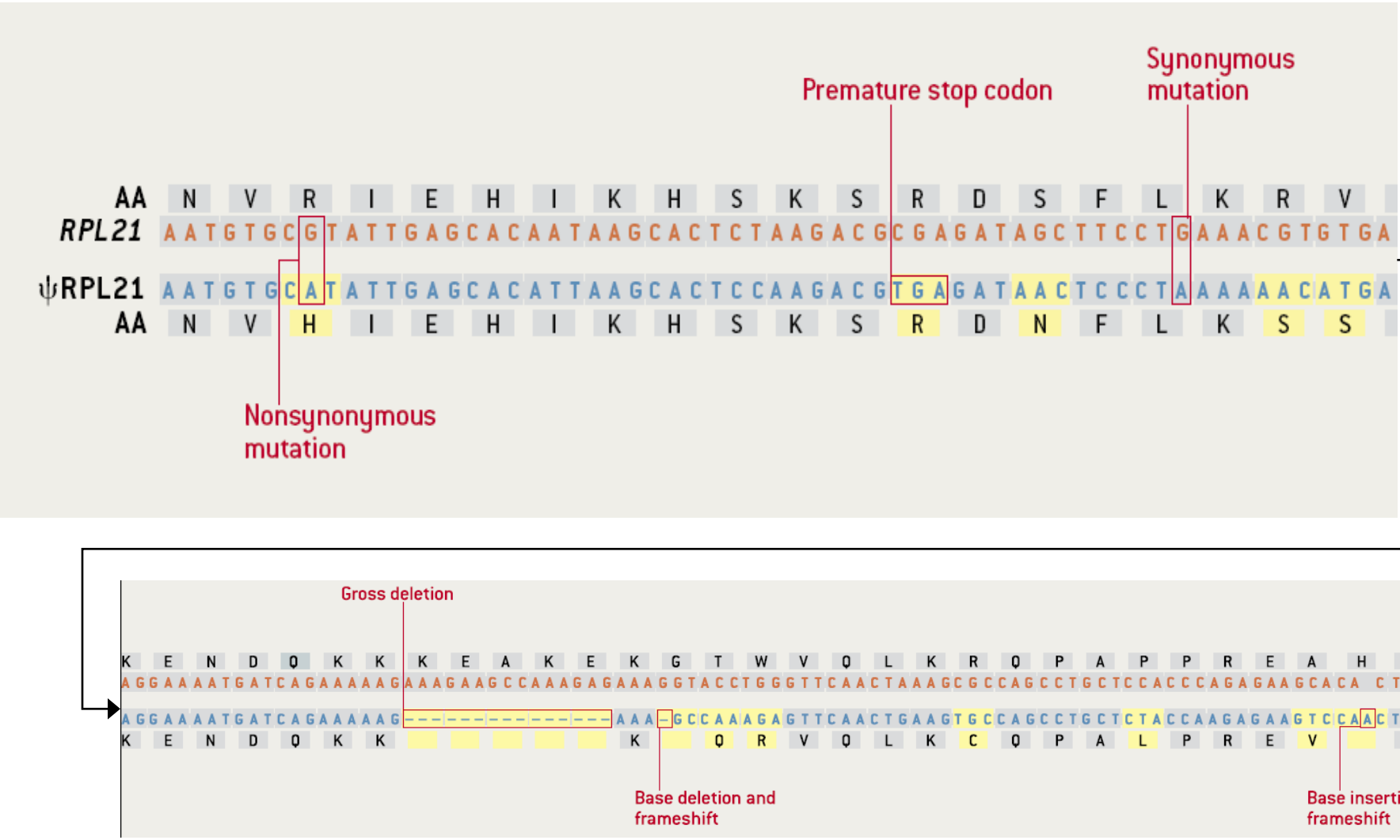
- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - **The advent of Nextgen seq.**
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

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?…

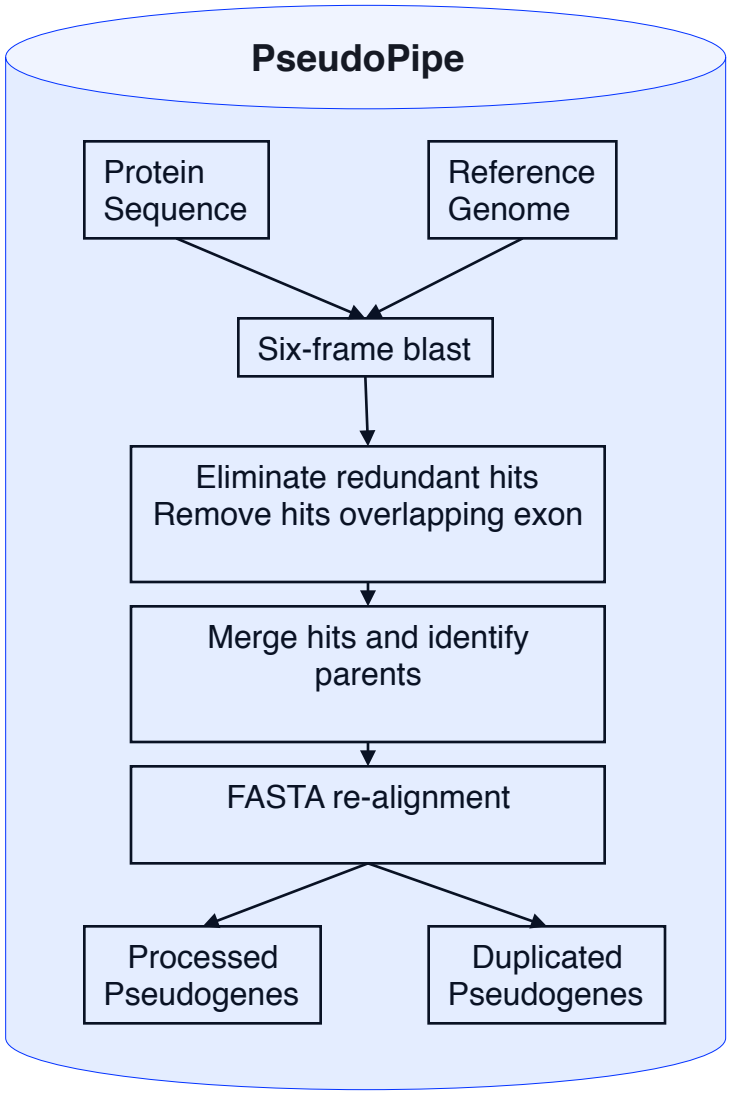
[Mighell et al. *FEBS Letts*, 2000]

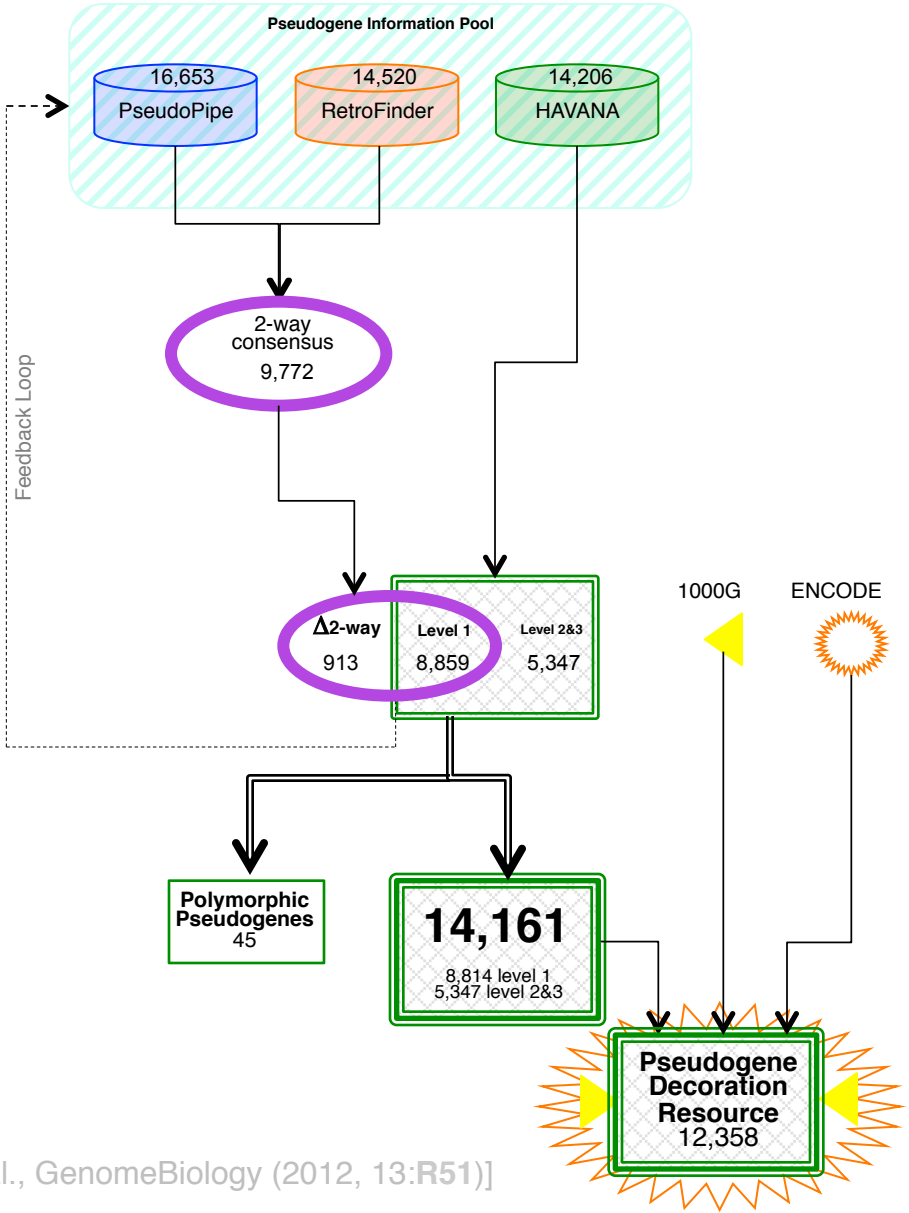
Identifiable Features of a Pseudogene (ψRPL21)



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

Genome-wide Annotation of Pseudogenes



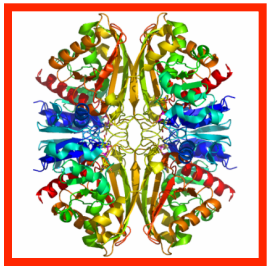


[Pei et al., GenomeBiology (2012, 13:**R51**)]

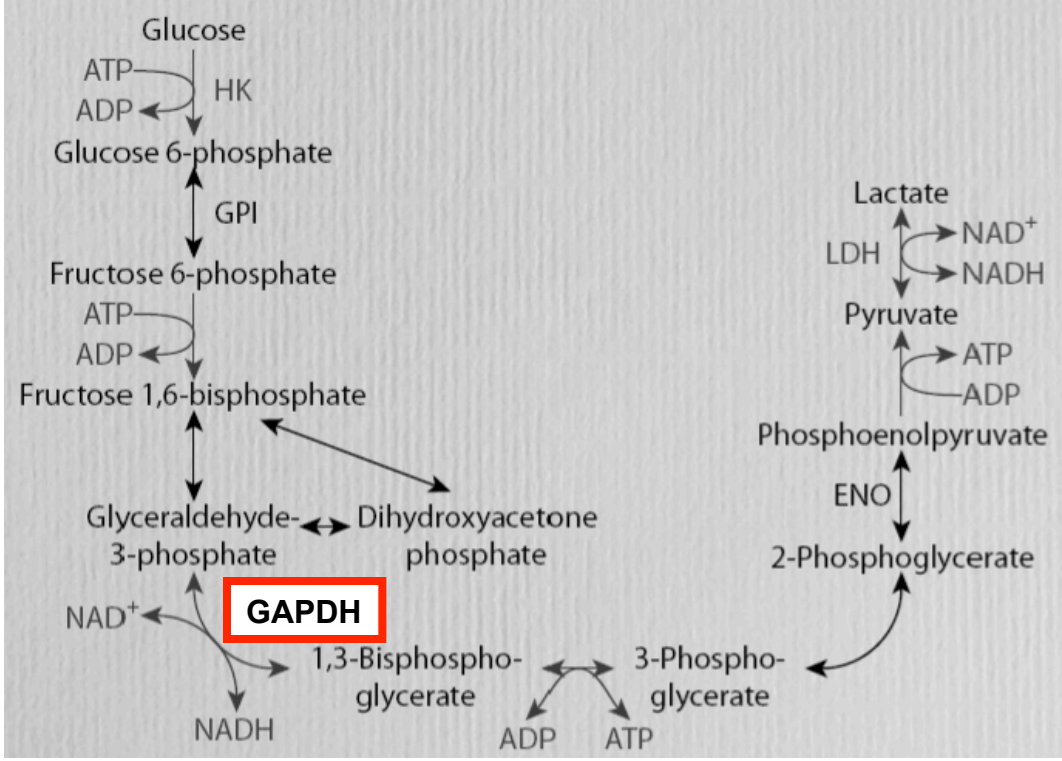
EX: Number of pseudogenes for each glycolytic enzyme

[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

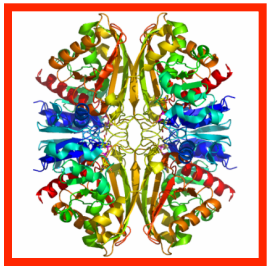


	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60/2	47/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	9/0	3/0	-	-	-	-	-
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**

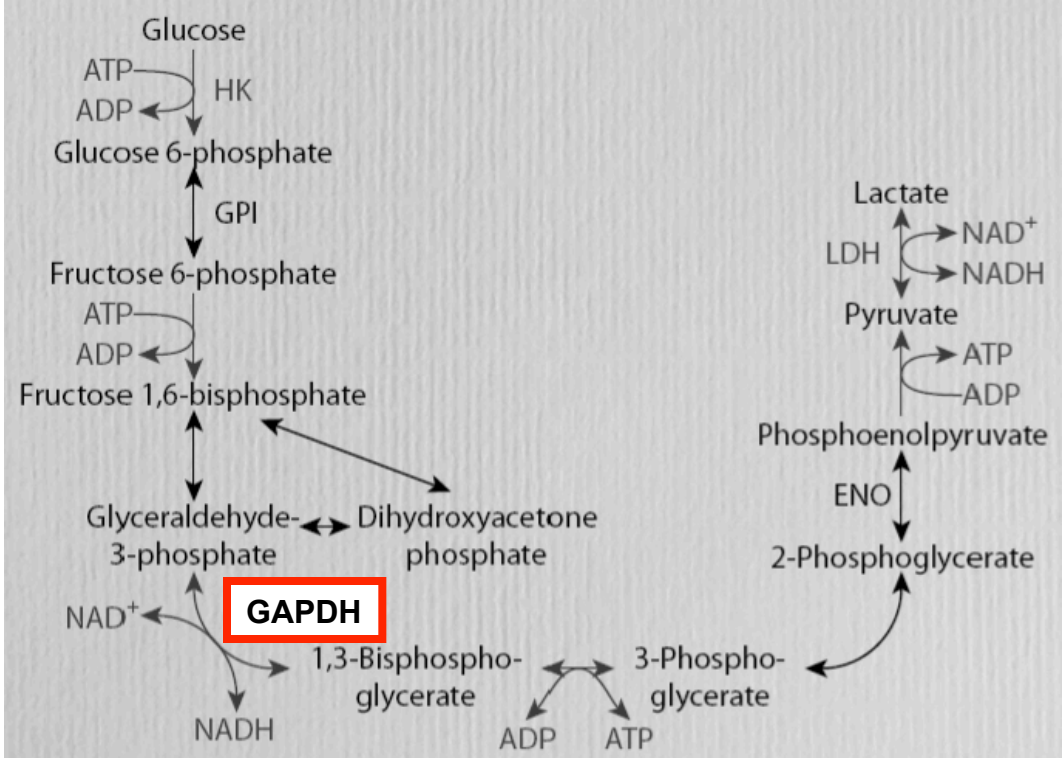
EX: Number of pseudogenes for each glycolytic enzyme

[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.



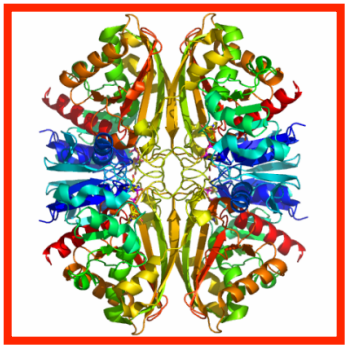
Processed/Duplicated



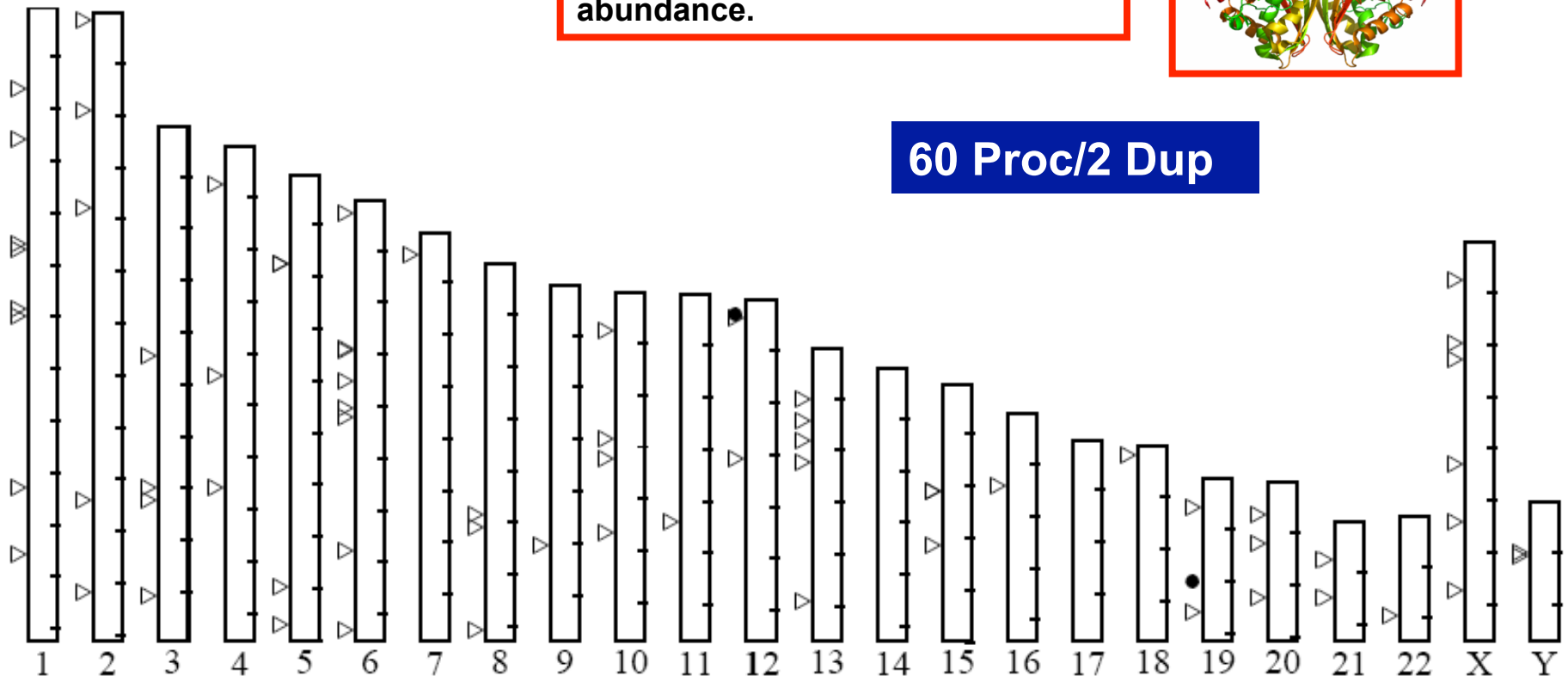
	Human	Chimp	Mouse	Rat	Chicken	Zebrafish	Pufferfish	Fruitfly	Worm
HK	1/0	1/2	0/1	-	0/2	-	-	-	-
GPI	-	-	1/0	-	-	-	-	-	-
PFK	-	-	-	-	-	0/1	-	-	-
ALDO	1/1	1/1	11/0	7/0	0/1	-	-	-	-
TPI	3/0	2/1	6/1	3/1	-	-	-	-	-
GAPDH	60 Proc/2 Dup	7/3	285/46	329/35	0/1	-	-	-	-
PGK	1/1	1/2	2/0	12/0	-	-	-	-	-
PGM	12/0	13/1	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.



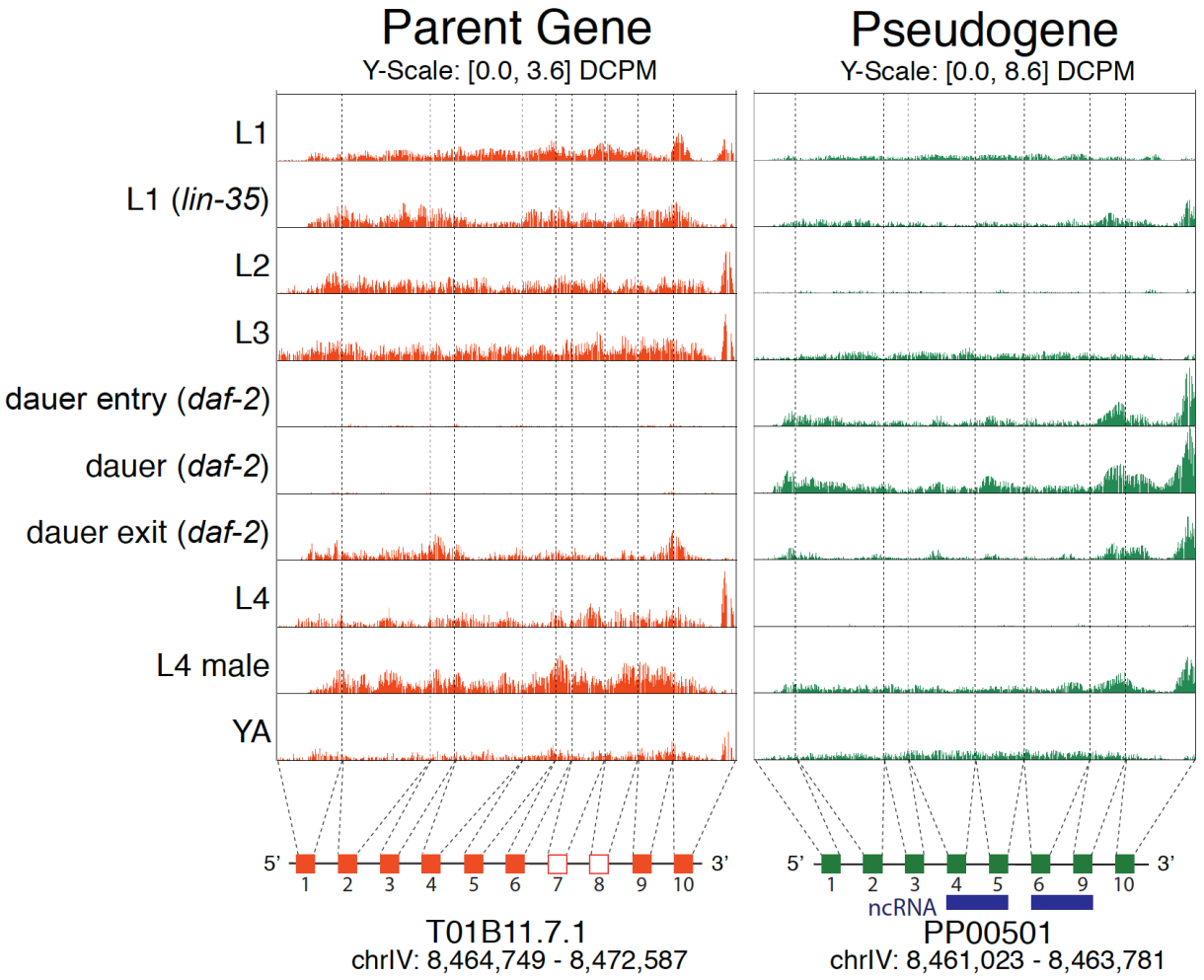
60 Proc/2 Dup



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

Calling transcribed pseudogenes (while guarding against mis-mapping)

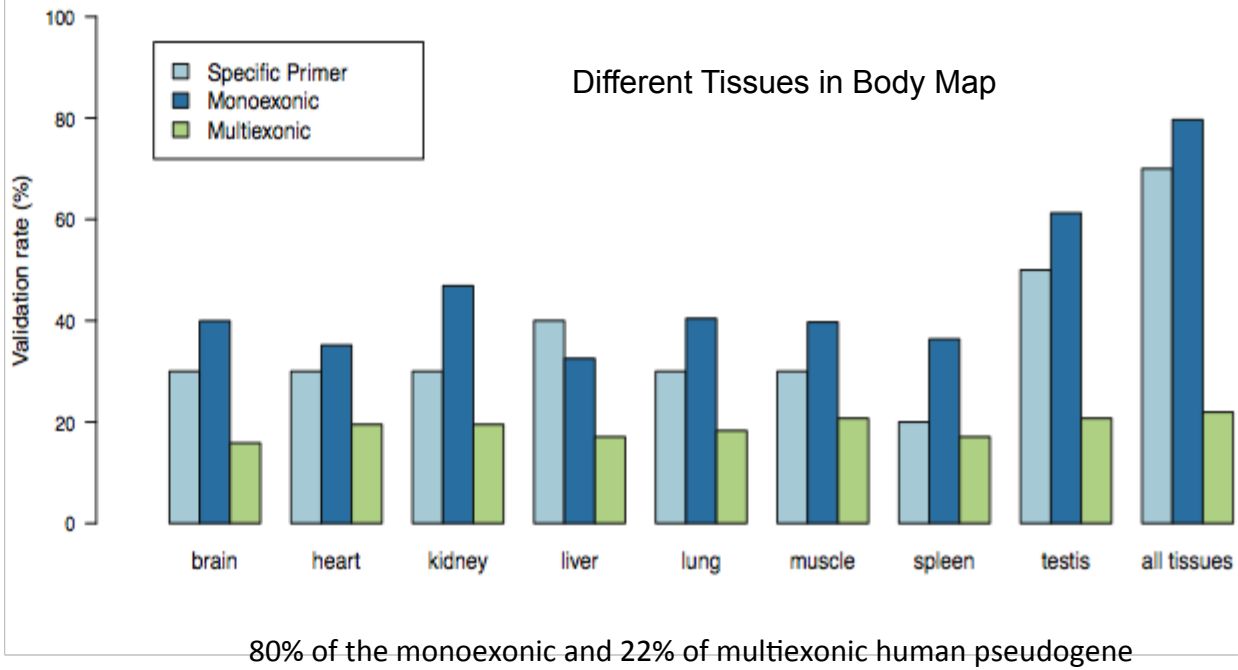
- Counting **uniquely mapped reads in RNA-seq**:
  - RPKM > 2
  - **1441** human transcribed pseudogenes
- Using **ESTs**



Looking for a discordant expression (misses pseudogenes co-expressed w. parent)

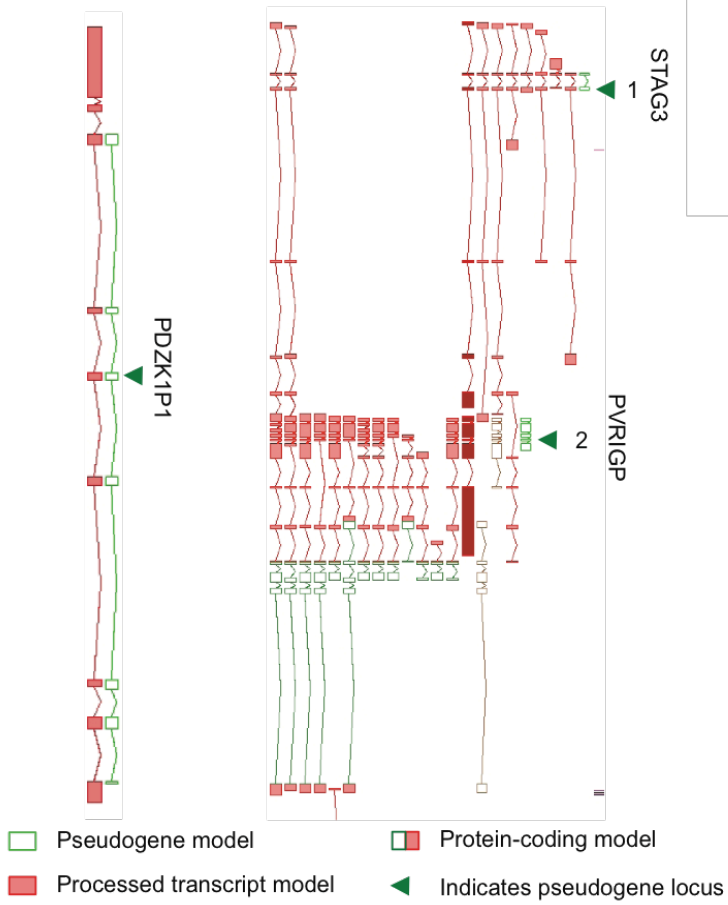
[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Validation of Pseudogene Transcription



Different Tissues in Body Map

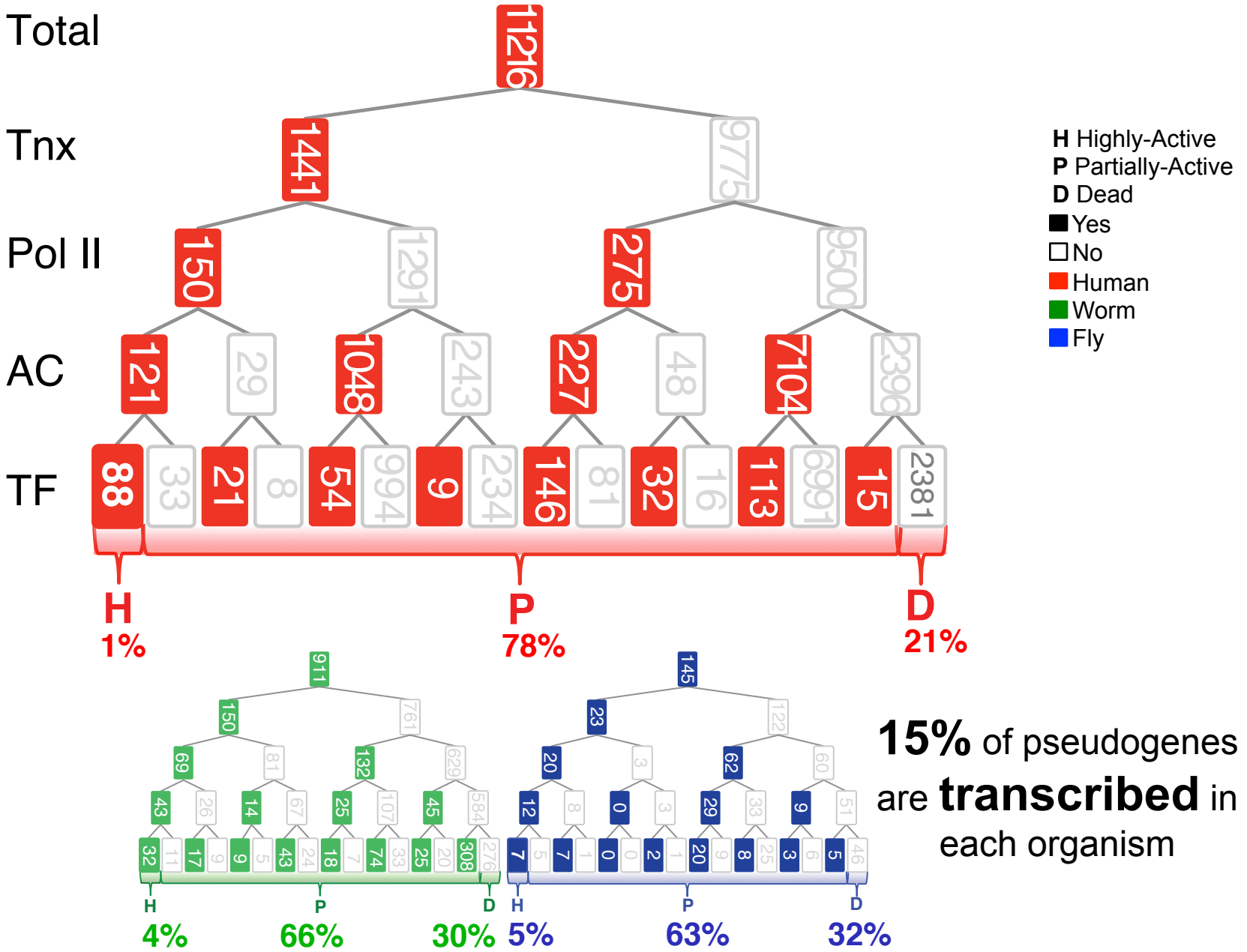
80% of the monoexonic and 22% of multiexonic human pseudogene models validated using RT-PCR-Seq; 57 of 76 in toal



Simple & Complex Ex of Pseudo-gene Trans-cription

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

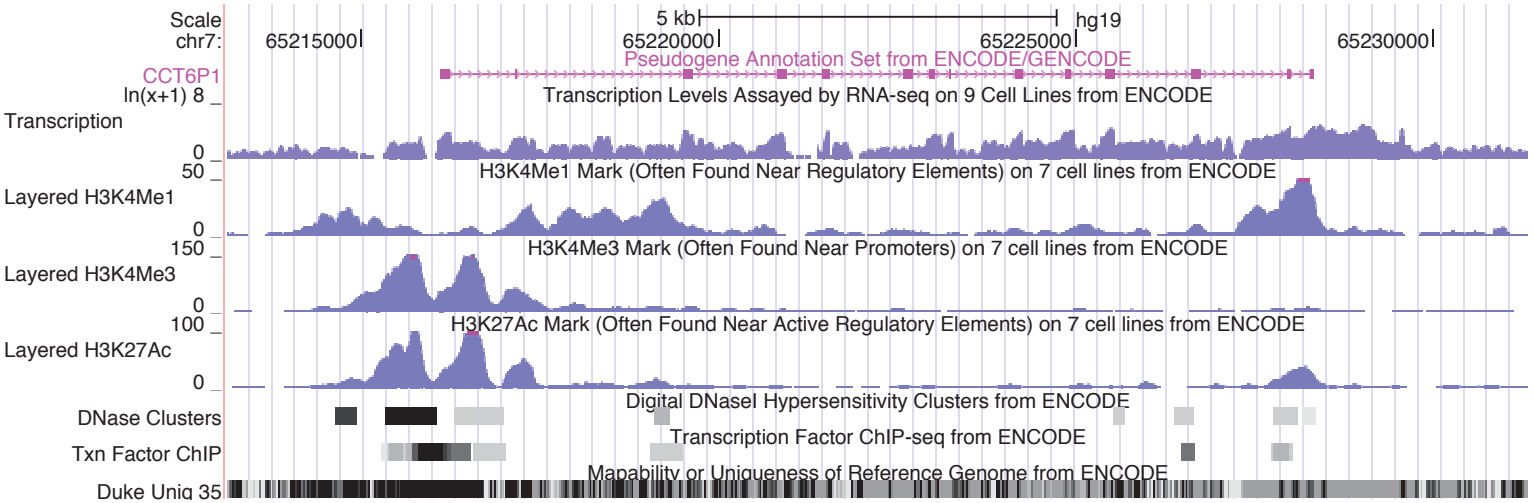
Pseudogene Activity



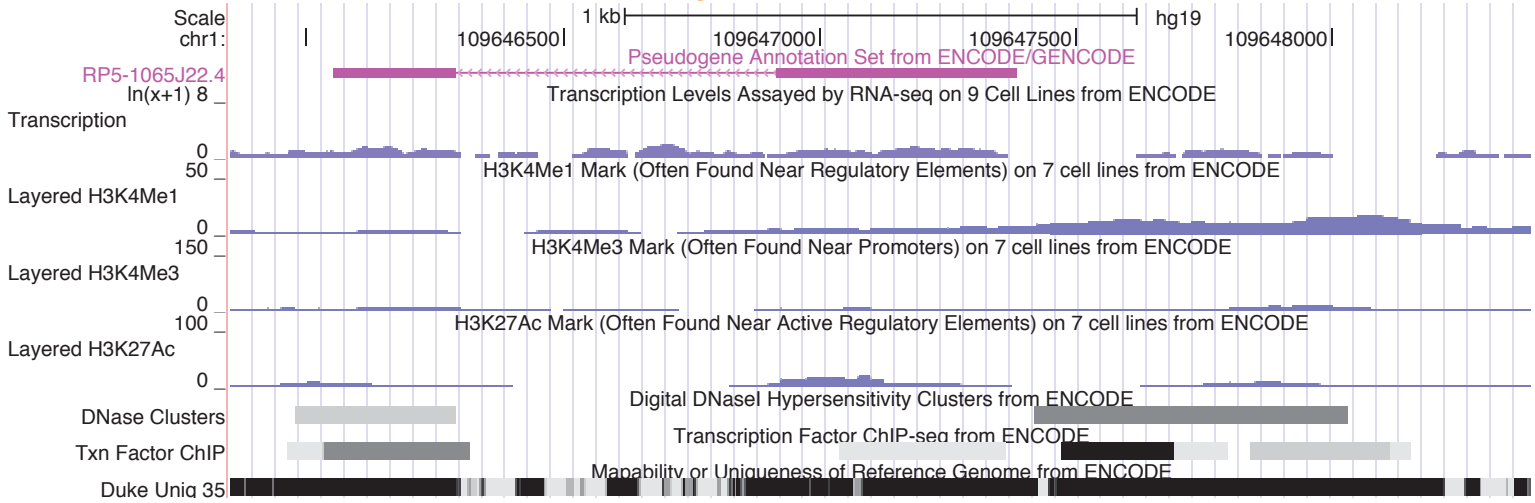
[Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Partial Pseudogene Activity

Transcribed with Additional Activity



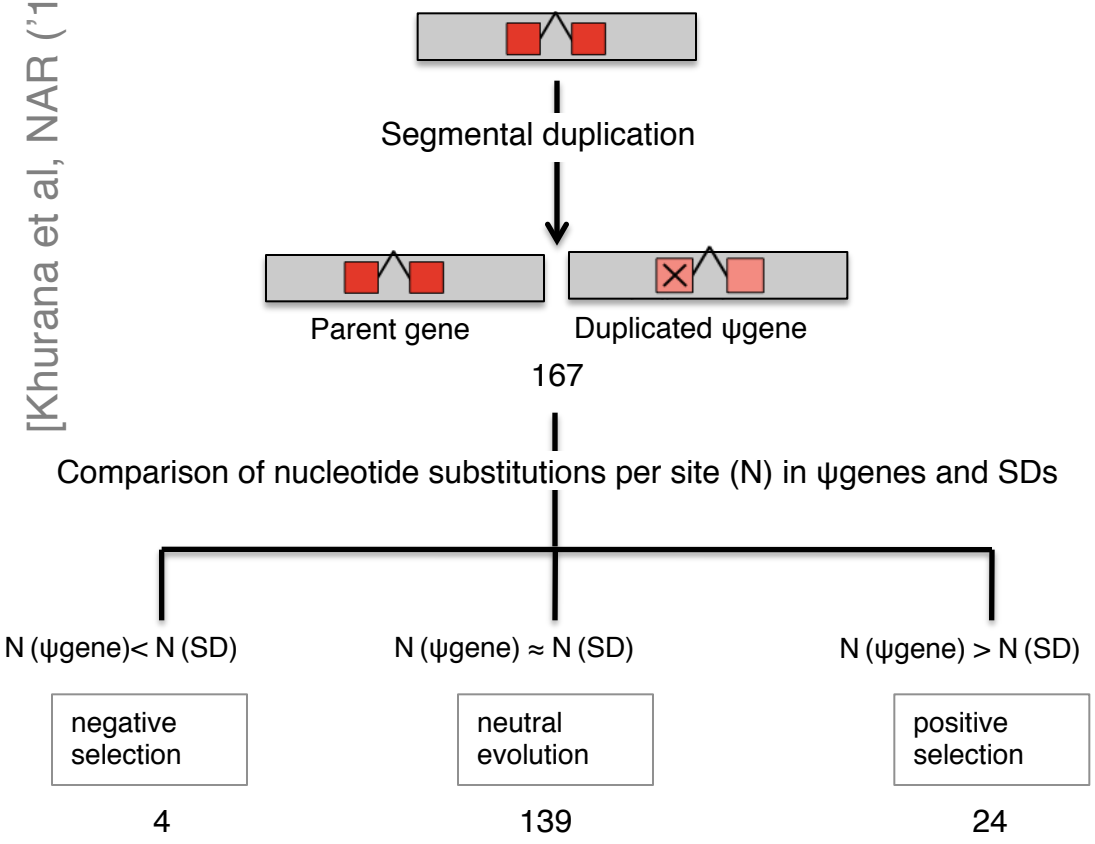
Partially Active

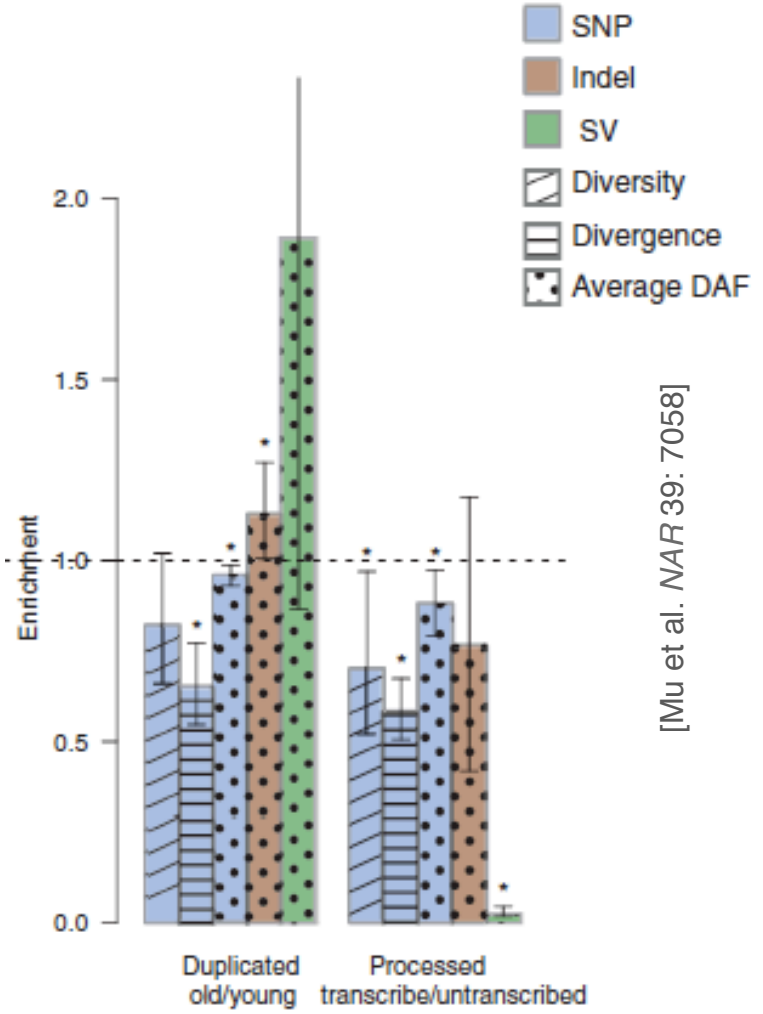


Signatures of Selection on Some Pseudogenes

- **Ka/Ks** conventional measure of selection for genes, shows no signal for pgenes
- Signature for selection on some SD pgenes (16%), derived from intersecting with UW SD DB & looking for differential conservation of neighborhood vs. center of pgene
- Weak signature for greater selection on transcribed pgenes using 1000G polymorphisms

[Khurana et al, NAR ('10)]



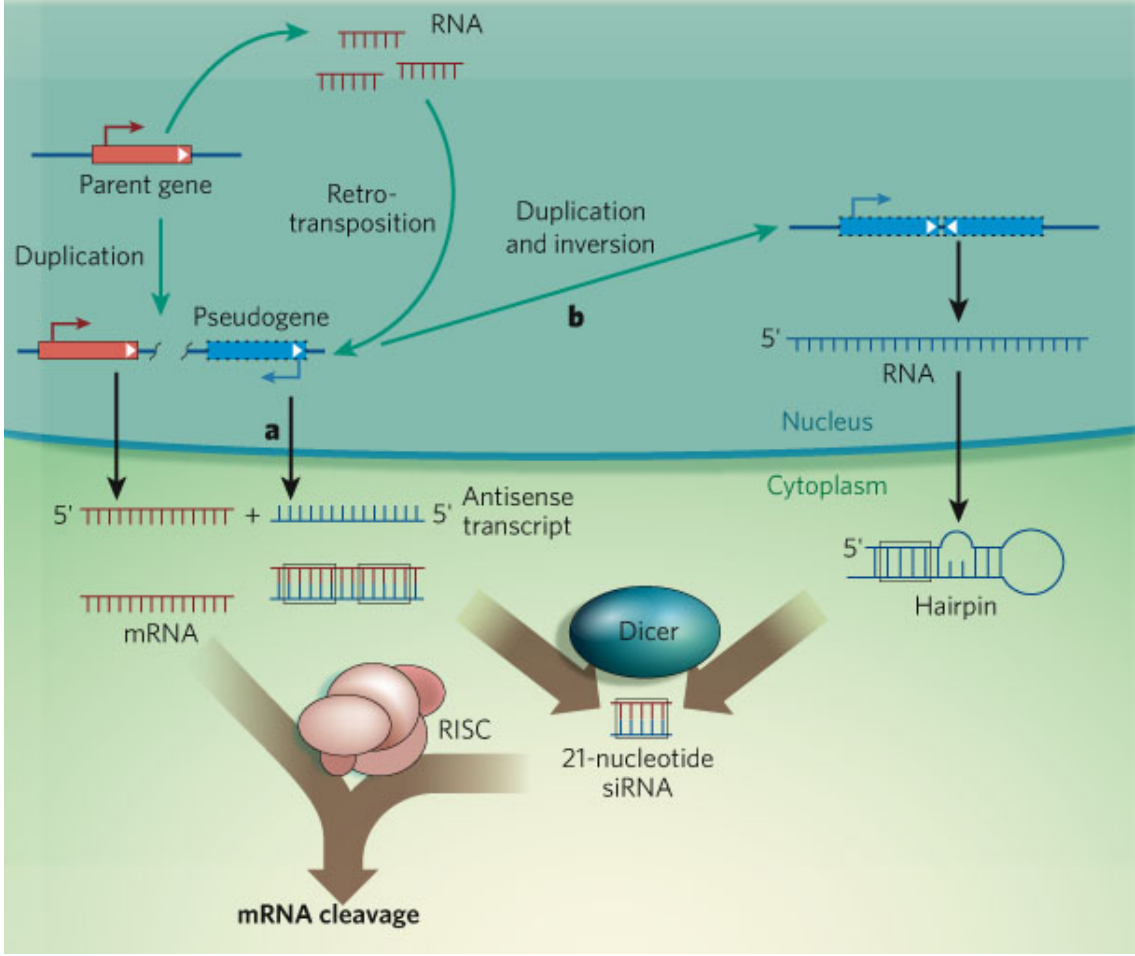


[Mu et al. *NAR* 39: 7058]

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, BRAF]
- via **inhibiting degradation** of parent's mRNA
  [makorin]



[Sasidharan & Gerstein, Nature ('08)]

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)
Karreth et al. Cell ('15).

Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - **The advent of Nextgen seq.**
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

Prevalence of noncoding transcription in the human genome, in relation to lncRNAs: From noisy TARs to regulatory pseudogenes

- **The Discovery of Pervasive Transcription**
  - Segmenting a noisy signal from Tiling Arrays into TARs
  - The problem: were the results accurate?
  - Cross-hyb.
- **Pervasive Transcription, Take 2**
  - The advent of Nextgen seq.
  - Evidence integration: incRNA
  - Now consistent cross-organism results: ~1/3 of the un-annotated genome is transcribed
- **Drilling into one type of pervasive transcription: Transcribed Pseudogenes**
  - Tricky definition & great prevalence (~14K)
  - Many transcribed: ~15%
    - Validation (RT-pcr)
    - Lots of other supporting evidence (other func. genomics expt. + selection)
  - Ideas on a regulatory biological role (endo-siRNAs, sponges, &c)

Acknowledgements

Early Pervasive Transcription
P Bertone, V Stolc, TE Royce, JS Rozowsky, AE Urban, X Zhu, JL Rinn, W Tongprasit, M Samanta, S Weissman, M Snyder

TAR Definition
J Rozowsky, D Newburger, F Sayward, J Wu, G Jordan, J Korbel, U Nagalakshmi, J Yang, D Zheng, R Guigo, T Gingeras, S Weissman, P Miller, M Snyder

Cross-Hyb.
T Royce, J Rozowsky

archive.gersteinlab.org/proj/**incrna**
Z Lu, **KY Yip**, G Wang, C Shou, LW Hillier, E Khurana, A Agarwal, R Auerbach, J Rozowsky, C Cheng, M Kato, D Miller, F Slack, M Snyder, RH Waterston, V Reinke

GAPDH Pseudogenes
Y-J Liu, D Zheng, S Balasubramanian, N Carriero, E Khurana, R Robilotto

Pseudogene.org/psicube
C Sisu, B Pei, J Leng, **A Frankish,** Y Zhang, S Balasubramanian, R Harte, D Wang, M Rutenberg-Schoenberg, W Clark, M Diekhans, J Rozowsky, T Hubbard, **J Harrow**

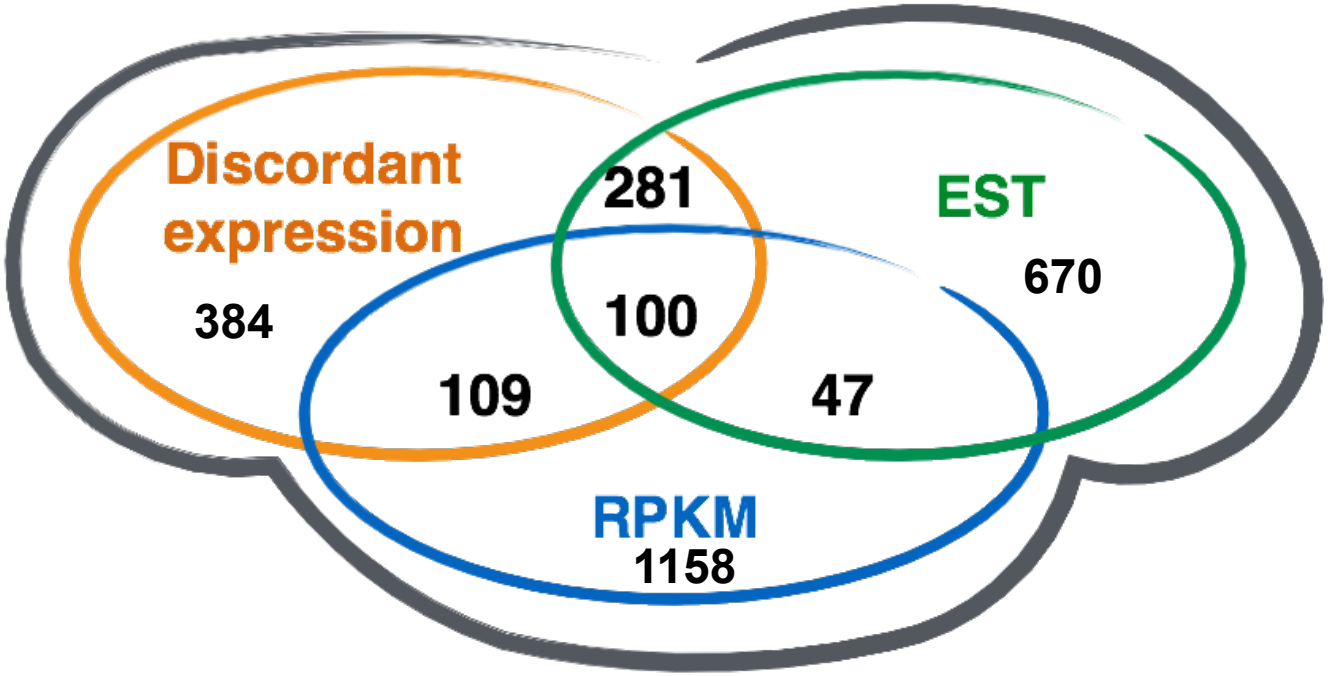
EncodeProject.org/comparative

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**

Extra

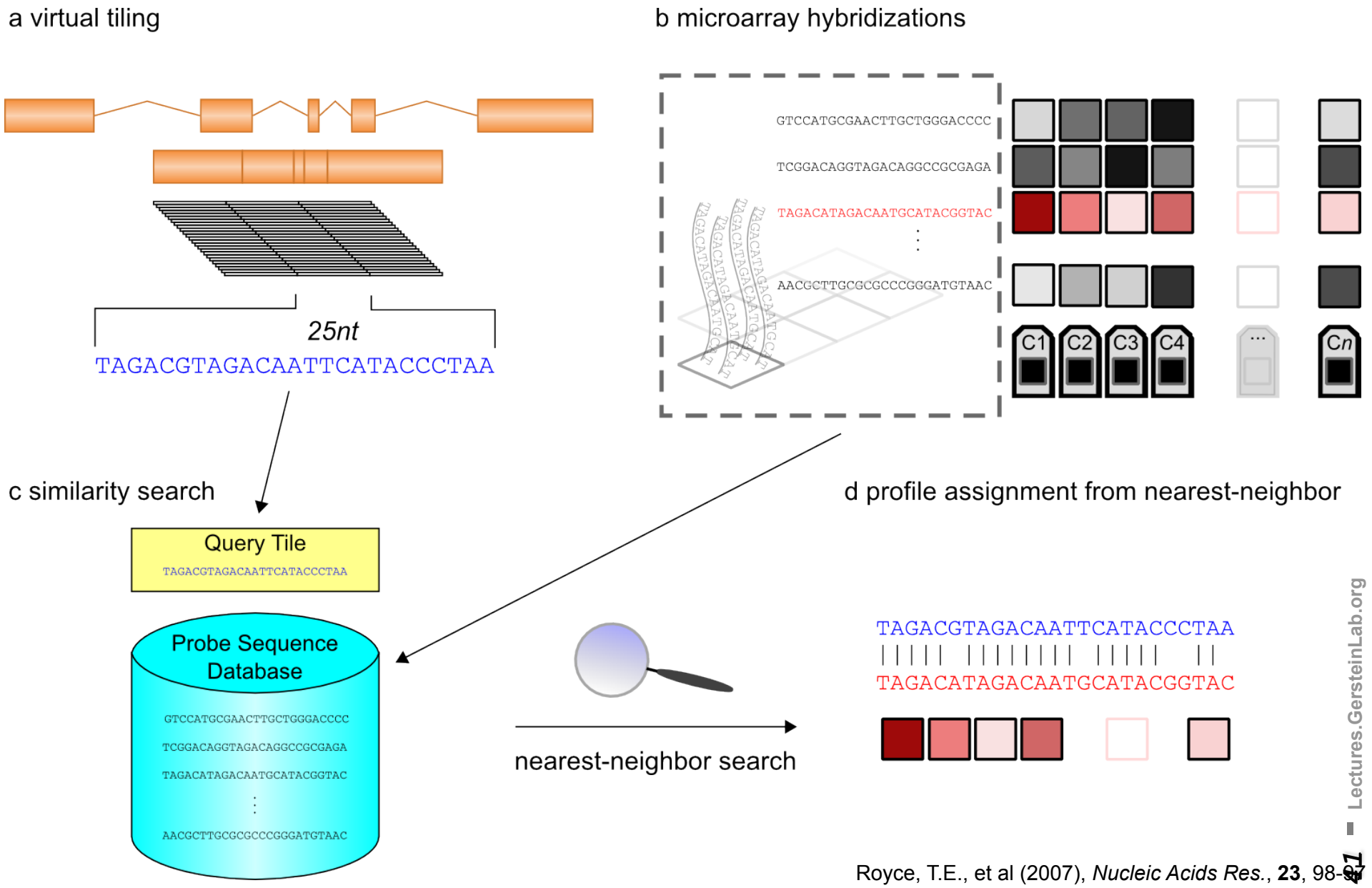
Transcribed pseudogene consensus



- **100** tri-way transcribed pseudogenes
- **2241** transcribed pseudognes showing at least one transcription evidence

[Pei et al., GenomeBiology (2012, 13:**R51**)] [Sisu et al. PNAS ('14); doi: 10.1073/pnas.1407293111]

Nearest Nbr Search on Virtual Tiling



Royce, T.E., et al (2007), *Nucleic Acids Res.*, **23**, 98-97

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - 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