

Identification of Transcribed Pseudogenes in Human Genome

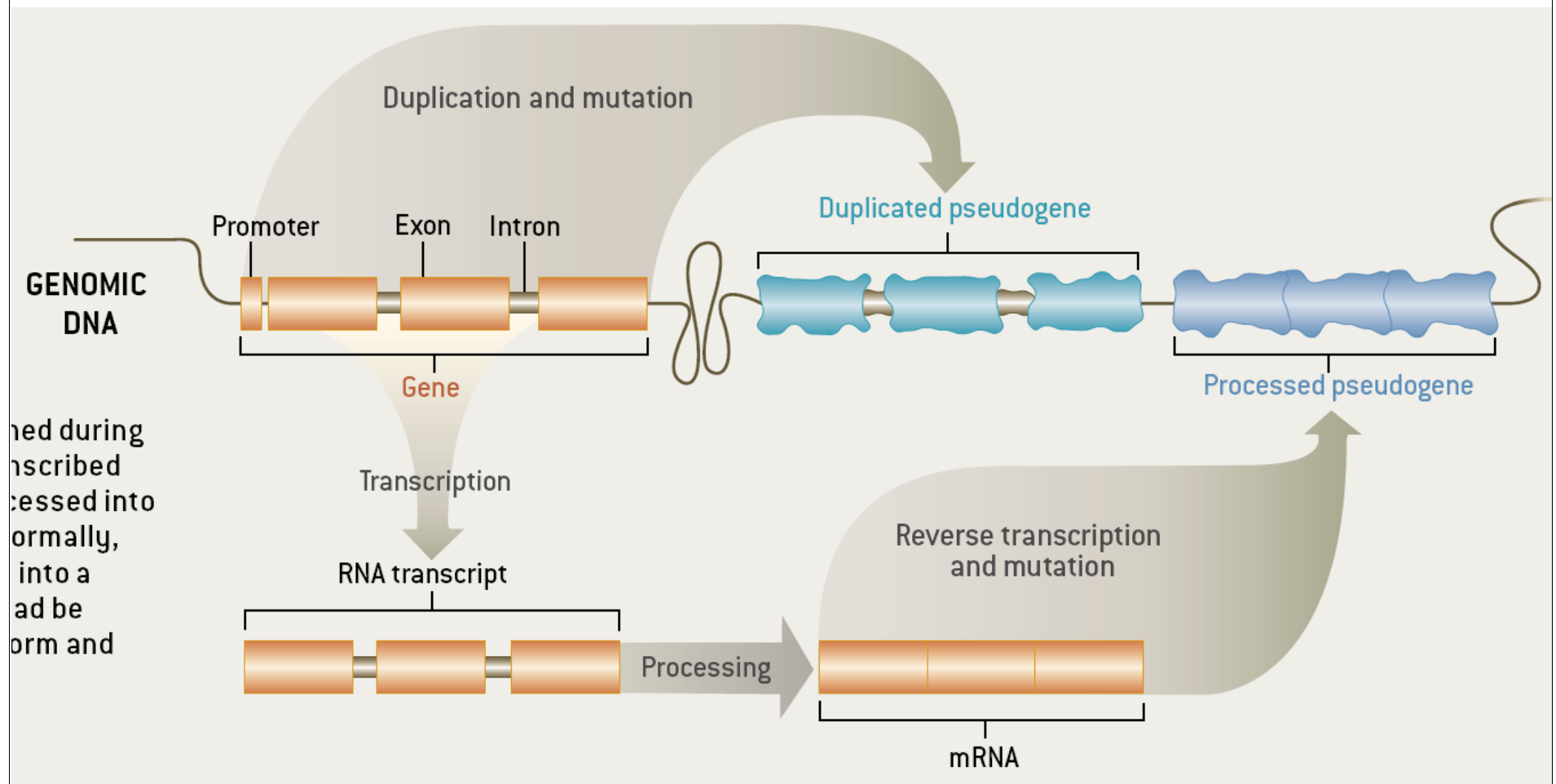
Baikang Pei, Yale

NLM Informatics Training Conference 2013

Pseudogenes are among the most interesting intergenic elements

- Formal Properties of Pseudogenes (Ψ G)
 - Inheritable
 - Homologous to a functioning element
 - 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?...

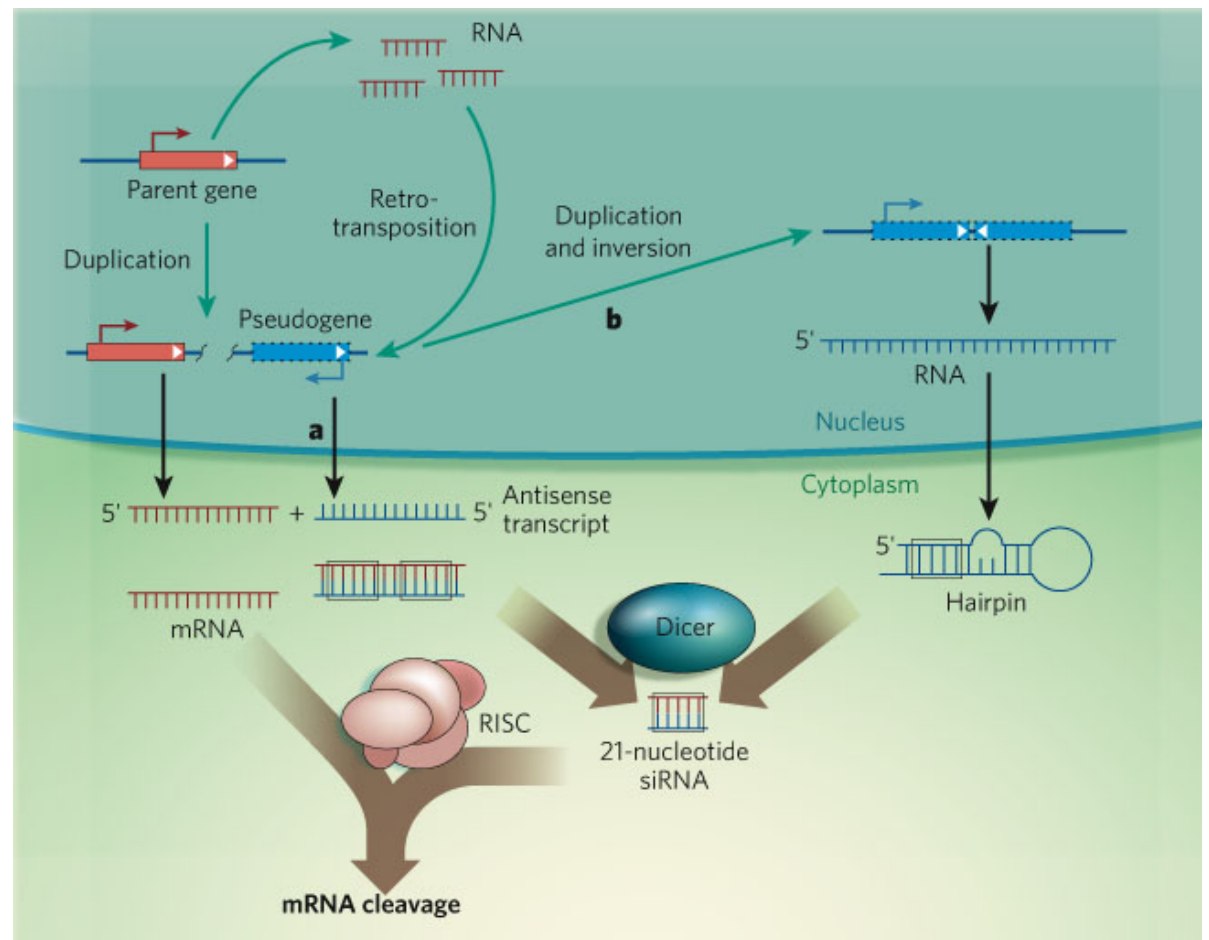
Two Major Genomic Remodeling Processes Give Rise to Distinct Types of Pseudogenes



Examples & speculation on the function of pseudogene ncRNAs:

Regulating their parents

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



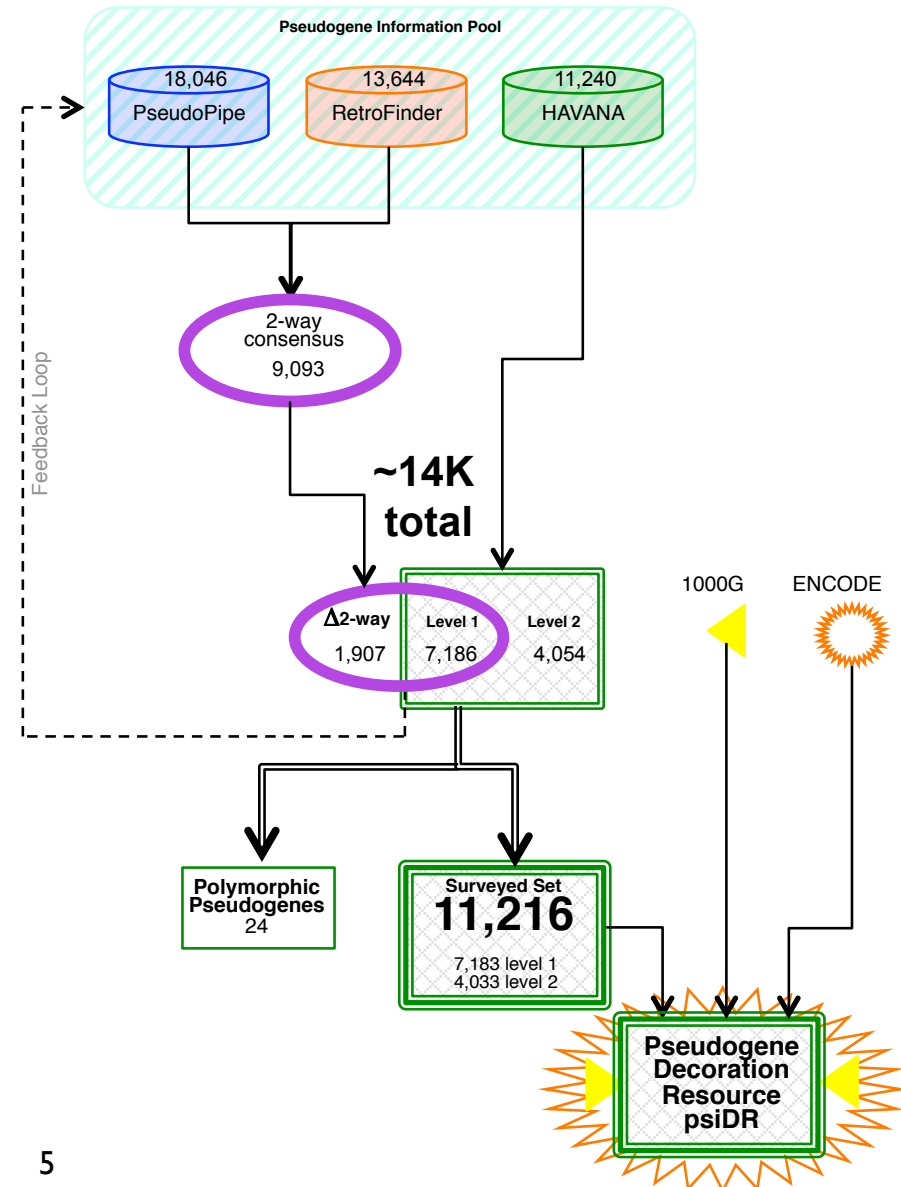
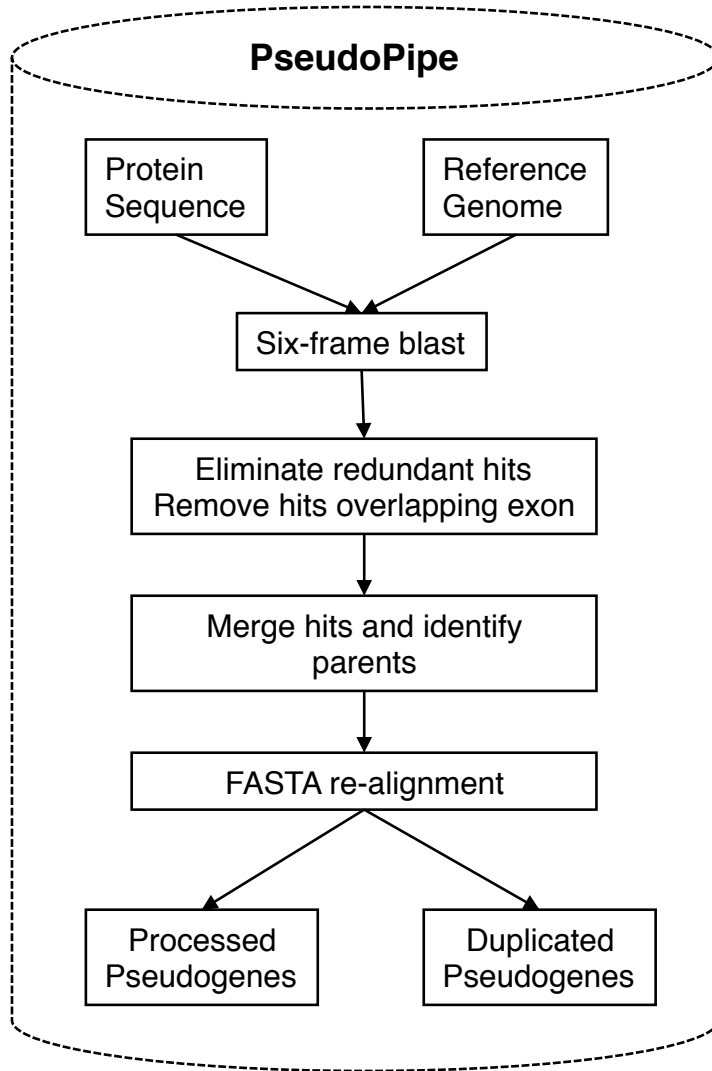
[Sasidharan & Gerstein, Nature ('08)]

Alternatively,
just last gasps
of a dying
gene

Czech *et al.* Nature 453: 798 ('08).
 Ghildiyal *et al.* Science 320: 1077 ('08).
 Kawamura *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)

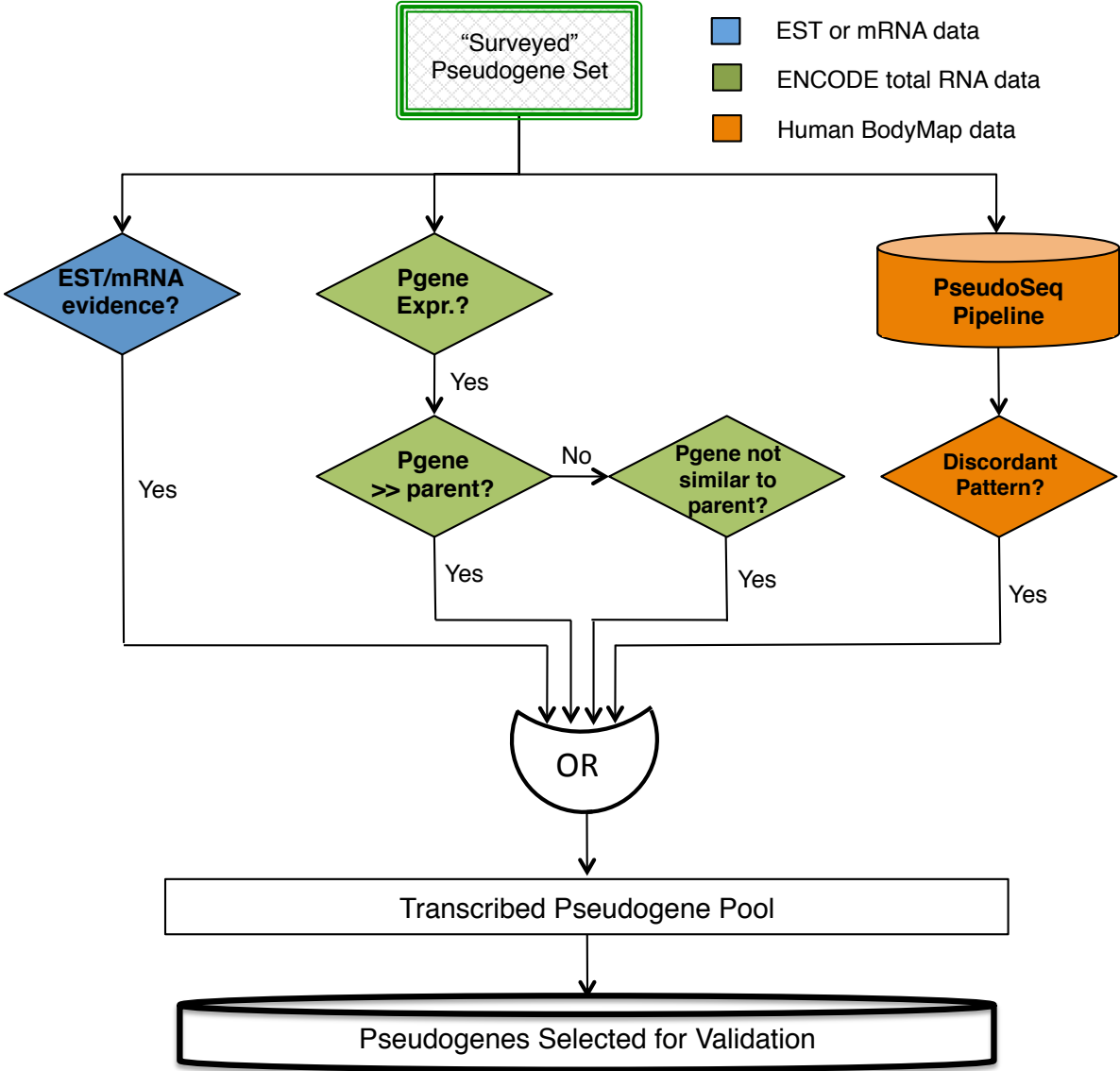
Genome-wide Annotation of Pseudogenes



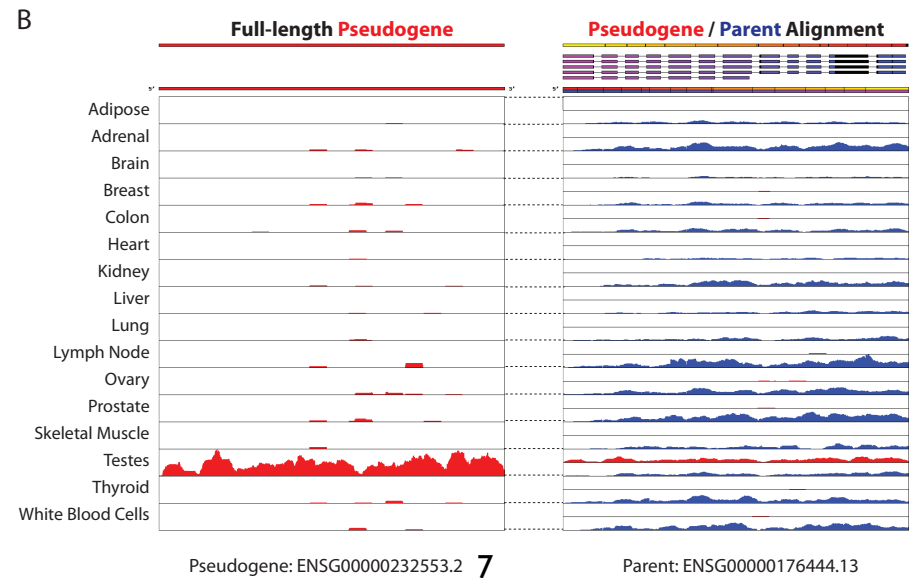
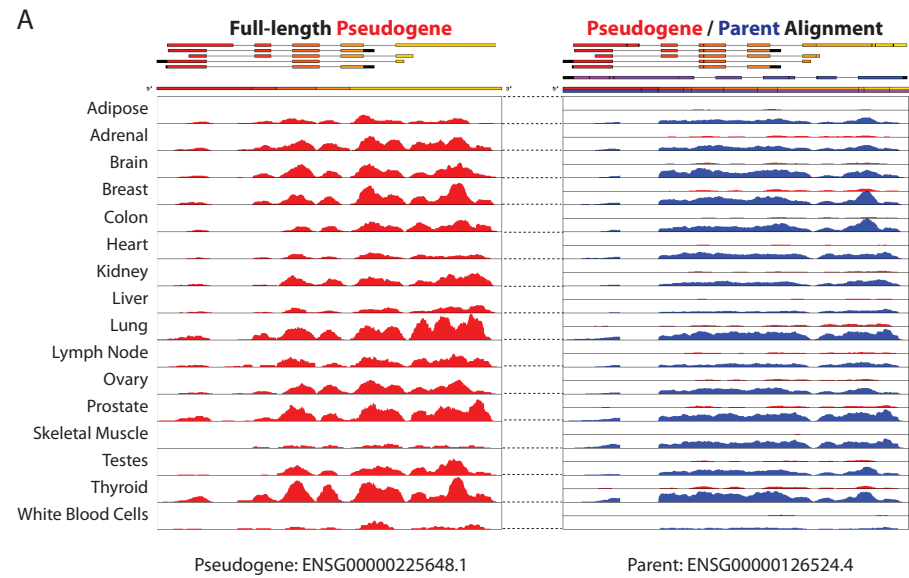
Pipeline to Identify Transcribed Pseudogenes

876 transcribed pseudogenes:

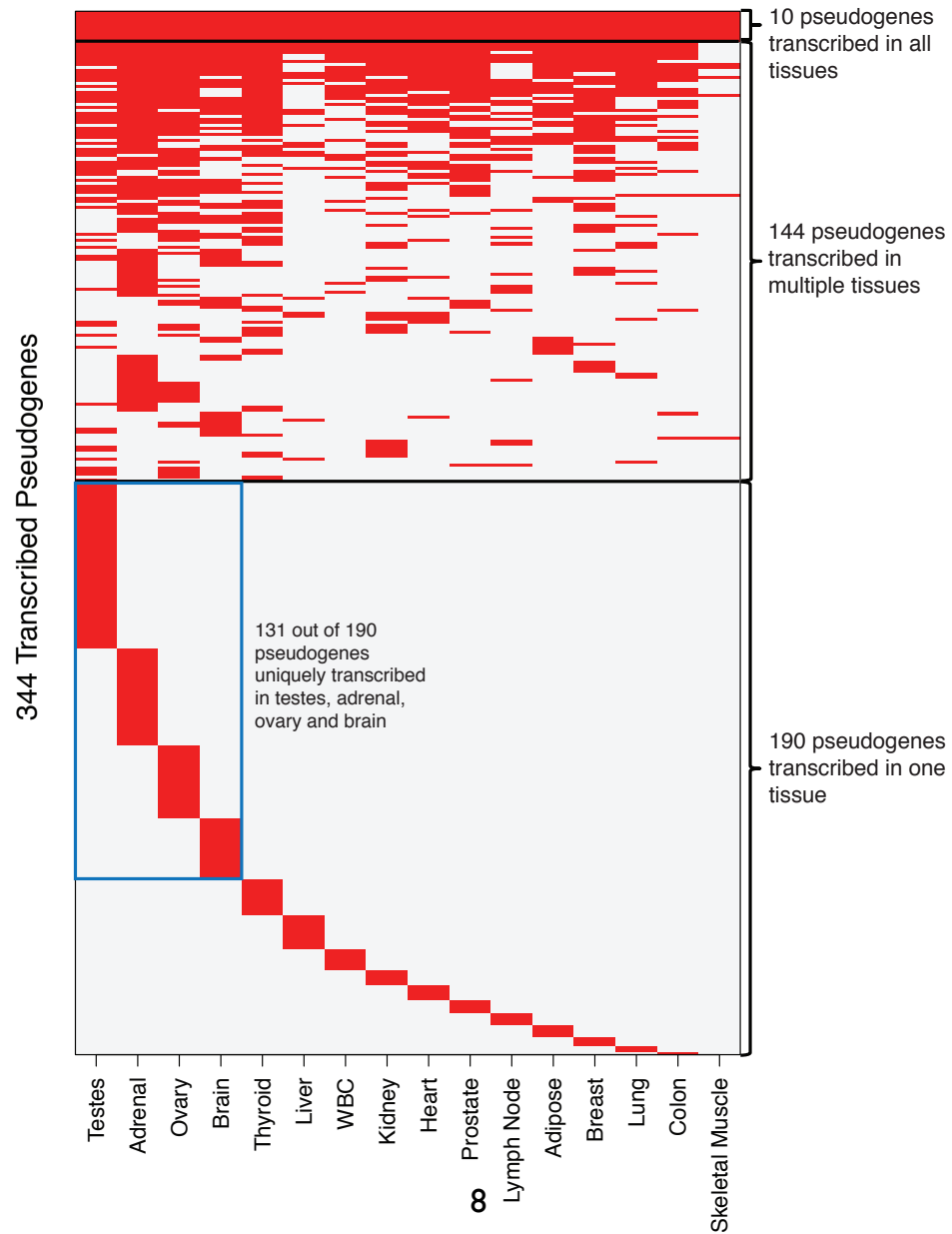
- 422 from EST evidence;
- 344 from pseudoSeq pipeline on BodyMap data;
- 110 from total RNA data of GM12878 and K562.



Transcribed Pseudogenes by PseudoSeq



Transcribed Pseudogenes by PseudoSeq

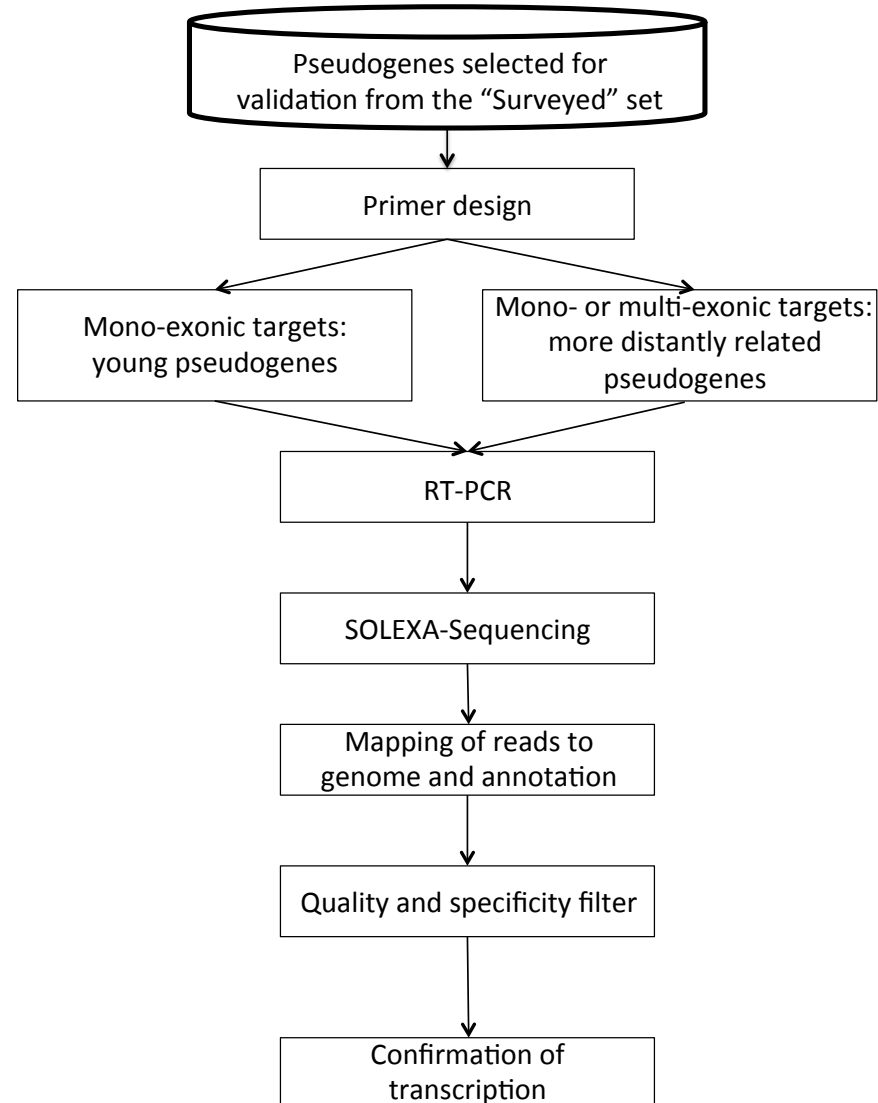


Validation of Transcribed Pseudogene

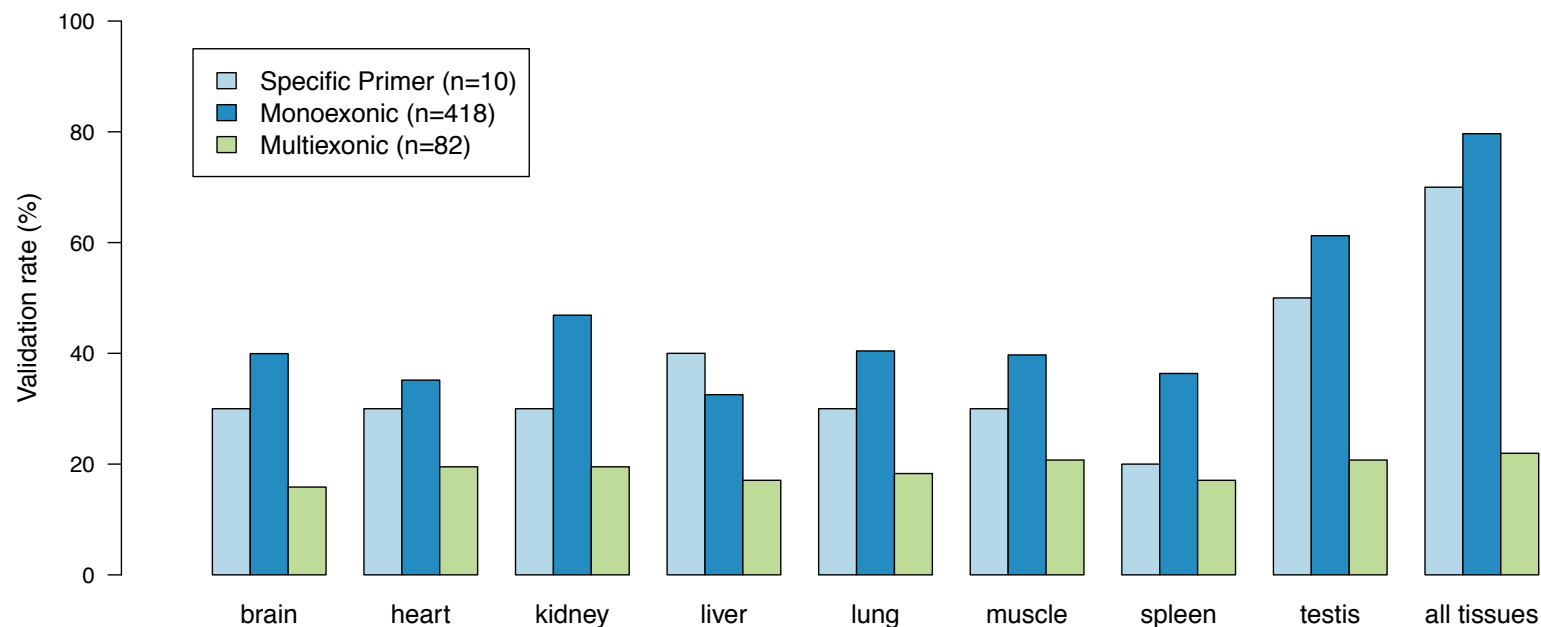
Mono-exonic RT-PCR:
Target to pseudogene exons. One target for each pseudogene;

Multi-exonic RT-PCR:
Target to exon-exon junctions. Multiple targets for each pseudogene;

Statistical model to make sure reads mapped to pseudogene annotation are indeed from pseudogene transcription, but not from parents.



Validation of Transcribed Pseudogene



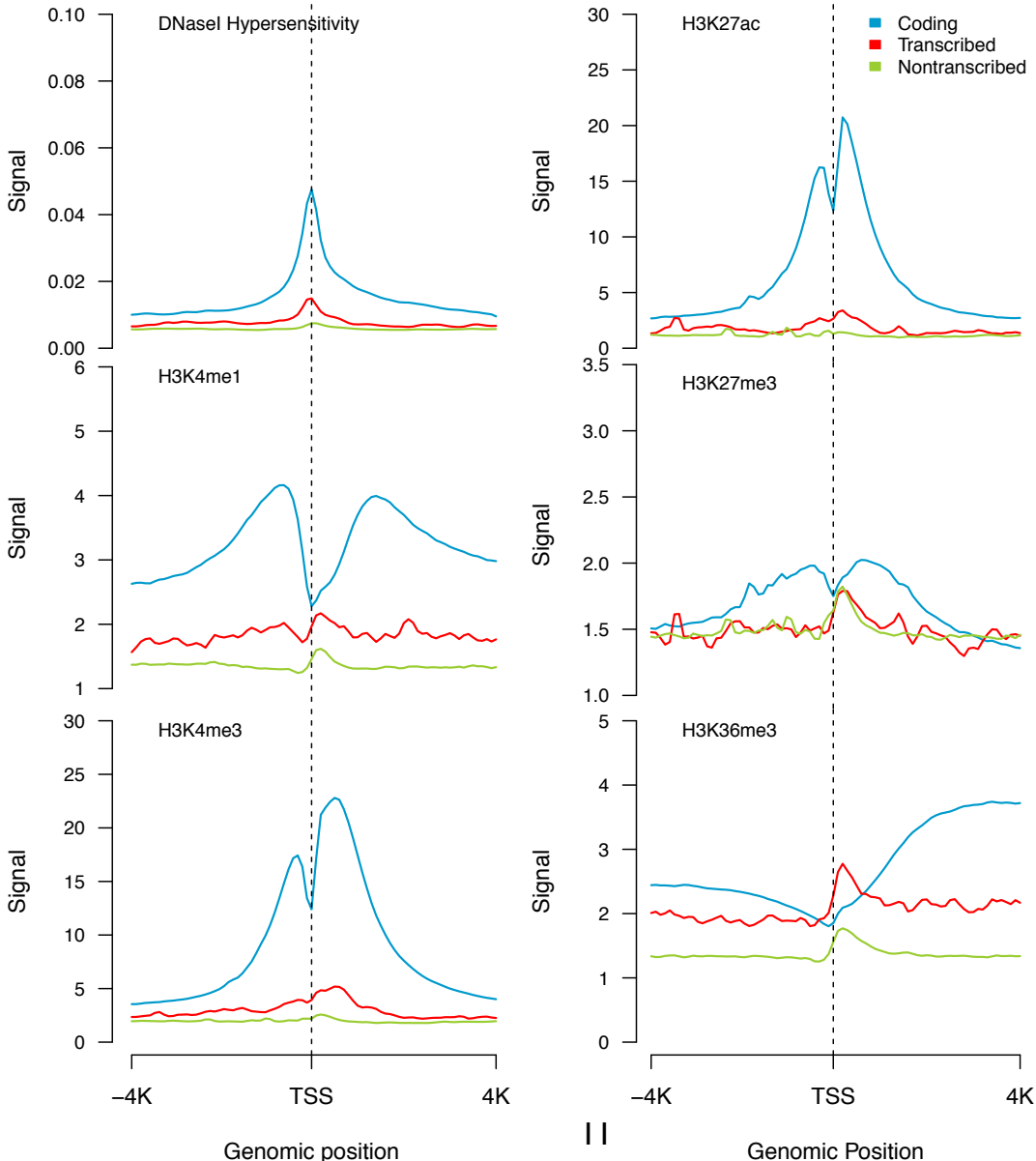
Total number of transcribed pseudogenes being validated: 469

- 94 from EST pipeline;
- 97 from totalRNA pipeline;
- 271 from BodyMap data pipeline;
- 7 are manually chosen due to their discordant expression patterns of pseudogenes and parents

Overall validation rate: 75.5% (354 out of 469)

- Specific primer: 70% (7 out of 10)
- Monoexonic: 79.7% (333 out of 418)
- Multiexonic: 22.0% (18 out of 82)

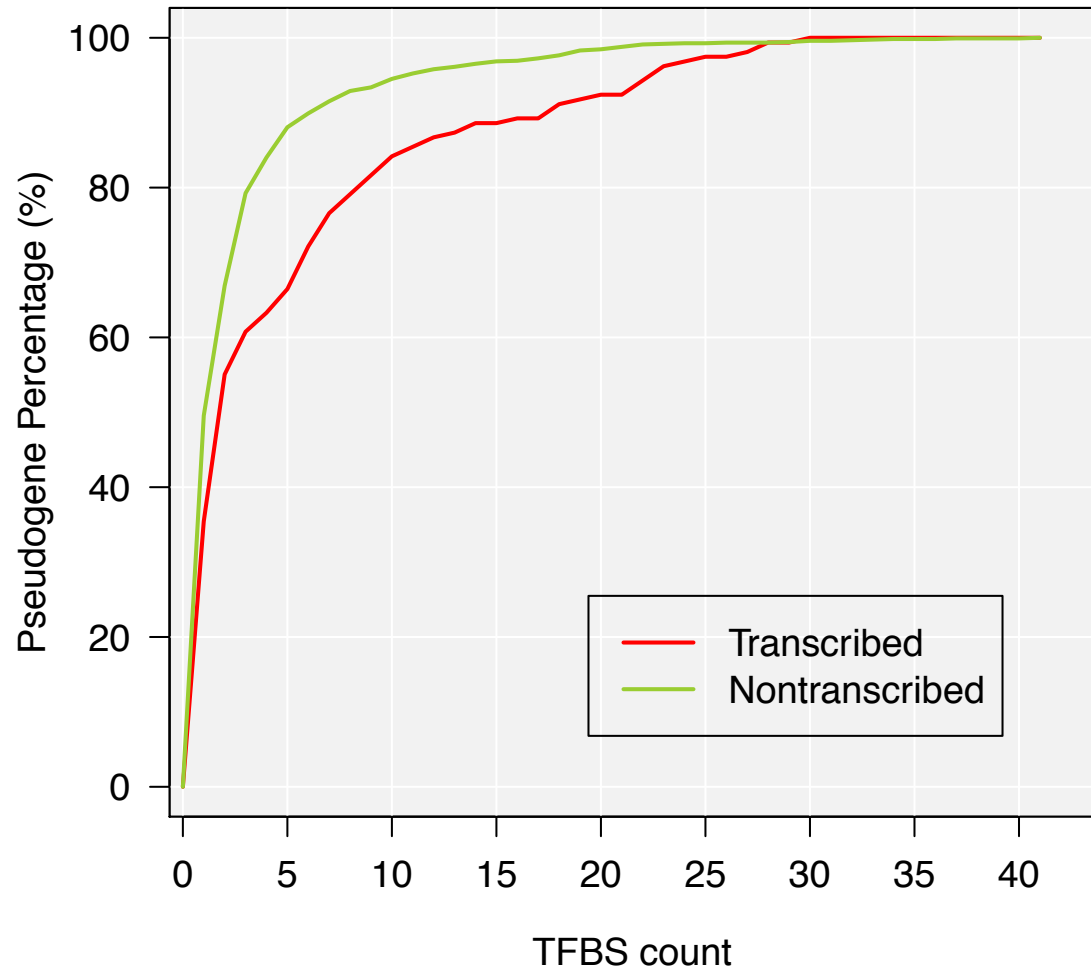
Chromatin Signatures of Pseudogenes



Transcription Factor Binding Sites of Pseudogenes

TF binding sites in upstream regions of pseudogenes in K562:

- Most pseudogenes have 0 or very few TFBS in their upstream regions
- Transcribed pseudogenes have more TFBS than non-transcribed pseudogenes (p-value = $3.8e-3$)
- Similar results in GM12878, HeLa-S3, h1-Hesc and HepG2 cell lines

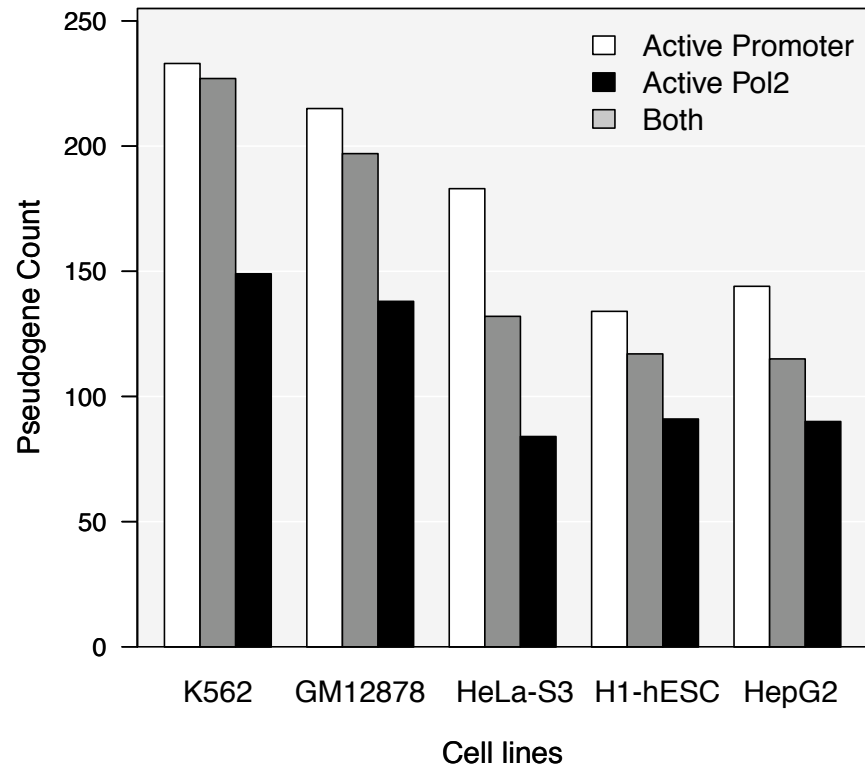


Pseudogenes with Active Upstream Sequences

Active promoters predicted by Kevin Yep's random forest model, using open chromatin, histone modification and TFBS data;

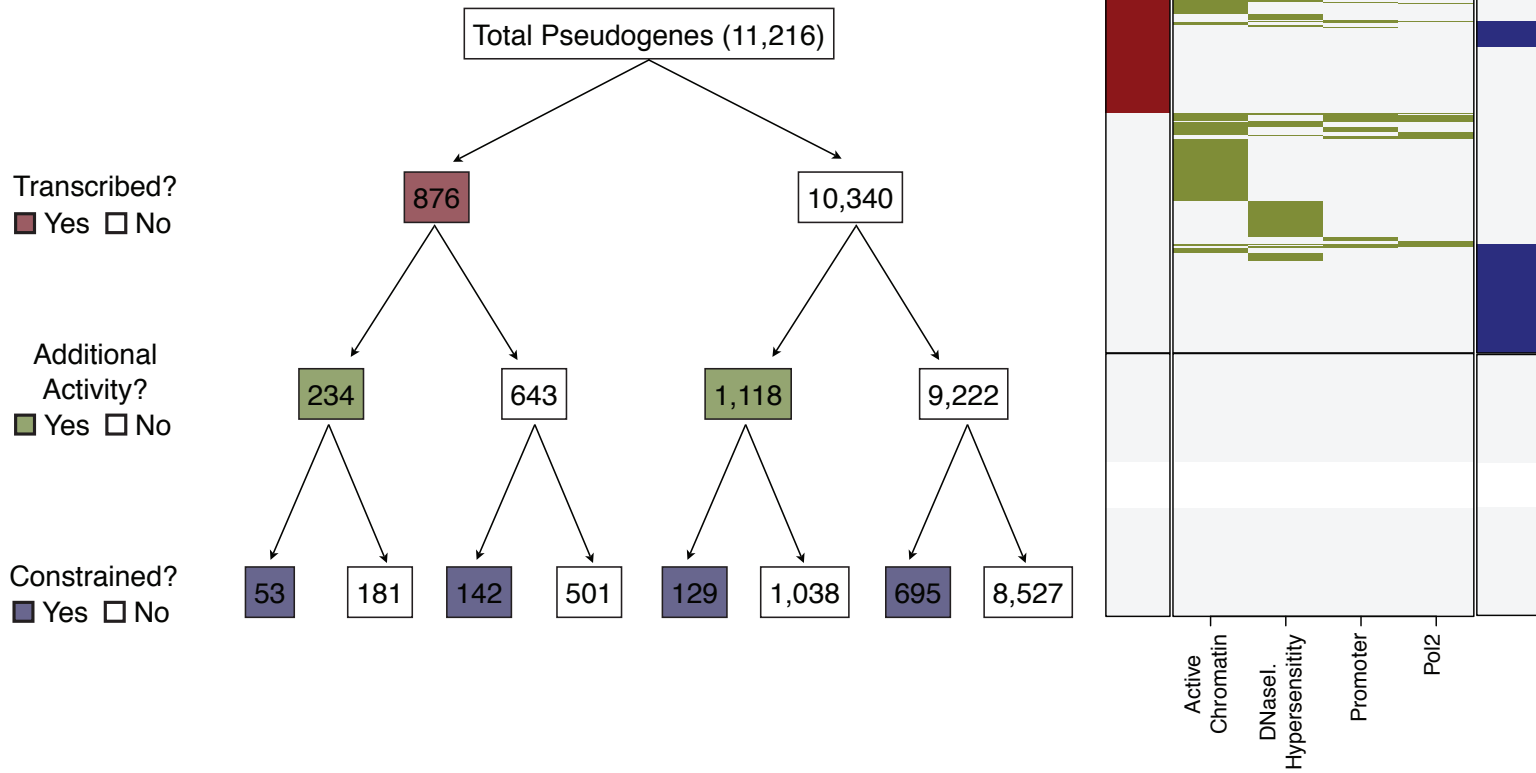
Active Pol2 bindings are from upper 5% of Pol2 binding peaks, in terms of peak widths and heights, plus binding of Pol2 co-factors;

Both active promoters and Pol2 binding sites are more abundant in upstream of transcribed pseudogenes than that of non-transcribed pseudogenes



	K562	Gm12878	Helas3	H1hesc	Hepg2
K562	-	0.30	0.29	0.22	0.27
Gm12878	0.33	-	0.33	0.27	0.32
Helas3	0.31	0.31	-	0.30	0.39
H1hesc	0.24	0.27	0.29	-	0.27
Hepg2	0.26	0.32	0.33	0.33	-

Partial Activity of Pseudogenes



Summary

- About 10% of human pseudogenes are transcribed
- Transcribed pseudogenes show tissue specificity
- Very few transcribed pseudogenes show universal activity, instead most exhibit partial activity

Acknowledgement

Sanger

Adam Frankish
Jeniffer Harrow
Tim Hubbard

UCSC

Rachel Harte
Mark Diekhans

University of Lausanne

Cedric Howald
Alexandre Reymond

Gerstein lab

Cristina Sisu
Lukas Habegger
Jasmine Mu
Suganthi Balasubramanian
Annotation subgroup
Mark Gerstein

CRG

Andrea Tanzer