The GENCODE Pseudogene Resource

Baikang Pei
Group Meeting, Gerstein lab
06/19/2012

Pseudogene Annotation

- Automated prediction by PseudoPipe and RetroFinder pipelines

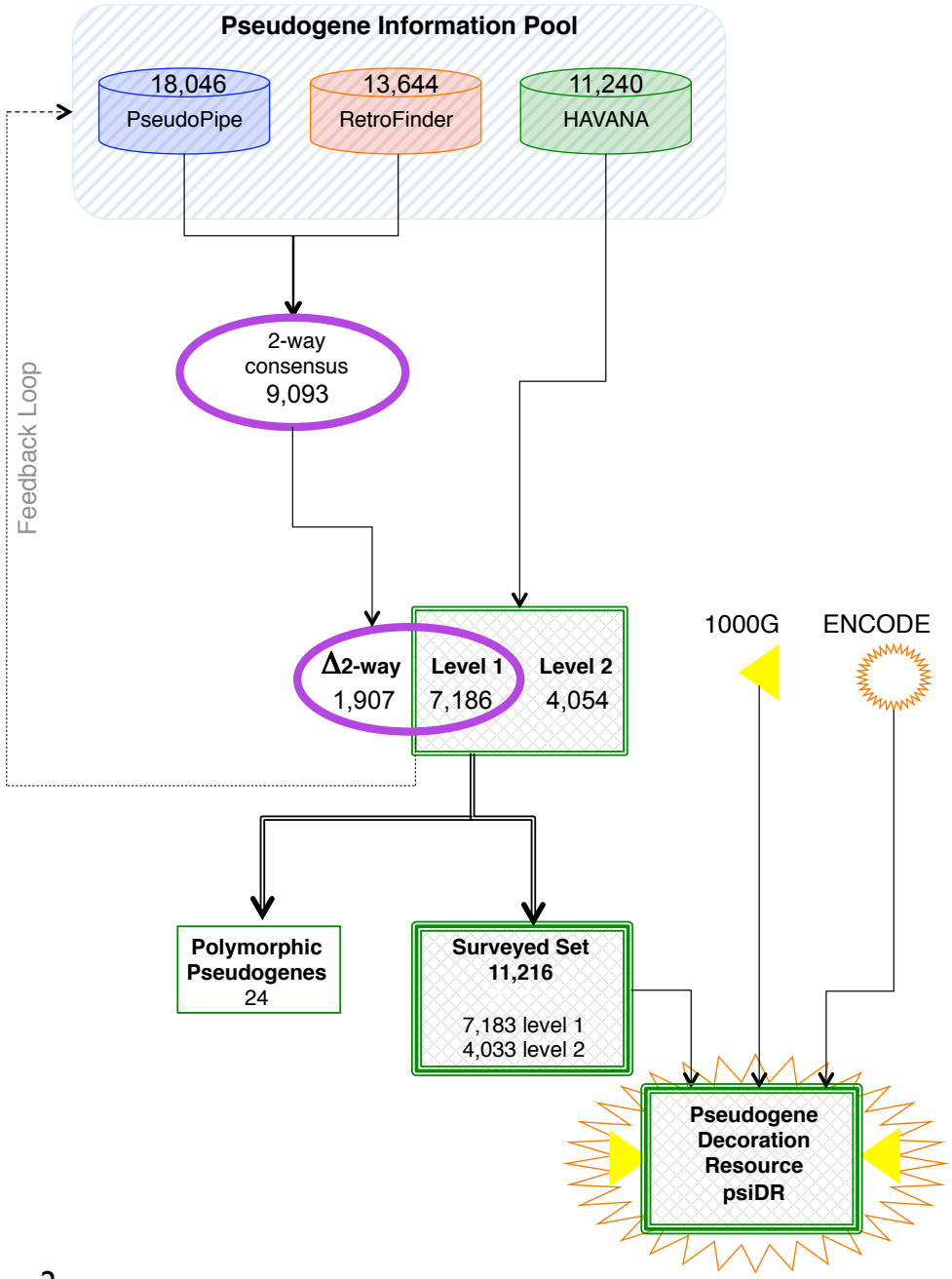
- Manual annotation is carried out by HAVANA

- Automated predictions are re-investigated by manual annotation

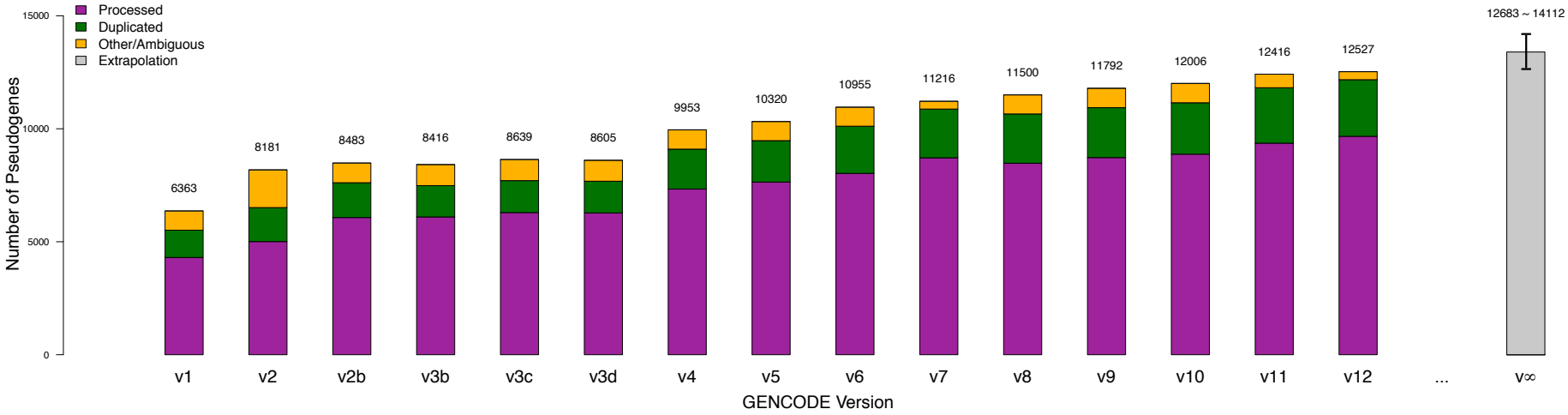
Pseudogene Extraction

- Manually annotated pseudogene transcripts are extracted from Gencode 7 GTF file

- Totally 11,216 pseudogene transcripts are collected.



Growth of GENCODE Pseudogene Annotation



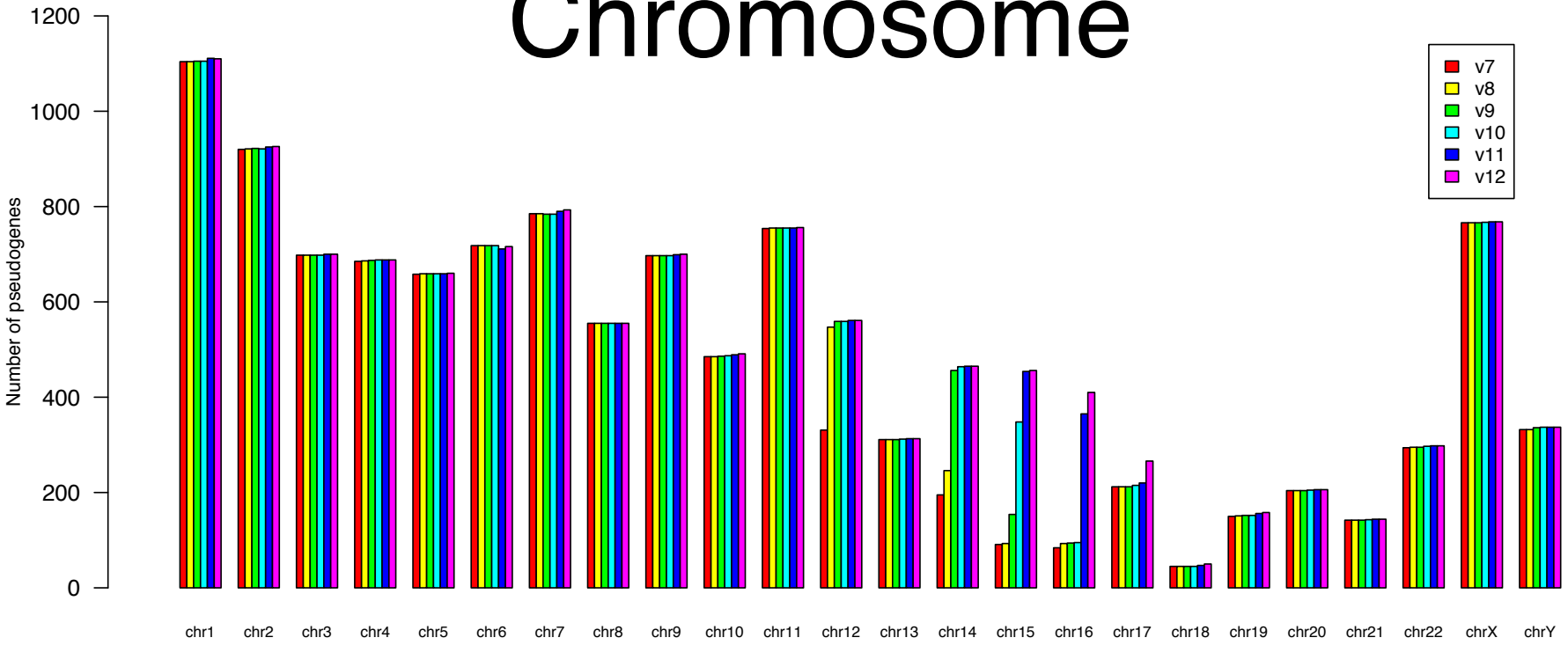
Pseudogene annotation status of Gencode v7:

Chromosome 1-11, 20, 21, 22, X and Y have been fully annotated, others are partially done.

Extrapolation methods:

1. Linear extrapolate from fully annotated chromosomes to partially annotated regions -> 12,683
2. Compare manual annotation and pseudopipe results on fully annotated regions, and then extrapolate to whole genome -> 14,112

Pseudogenes in Each Chromosome



Compare to previous Gencode release

	Add		Remove		Overall change
	Level 1	Level 2	Level 1	Level 2	
v8	184	115	11	4	284
v9	198	119	10	15	292
v10	144	104	21	13	214
v11	266	255	68	43	410
v12	130	89	74	34	111

Compare to Gencode v7

	Add		Remove		Overall change
	Level 1	Level 2	Level 1	Level 2	
v8	184	115	11	4	284
v9	389	224	27	10	576
v10	533	325	48	20	790
v11	799	566	113	52	1200
v12	927	643	184	75	1311

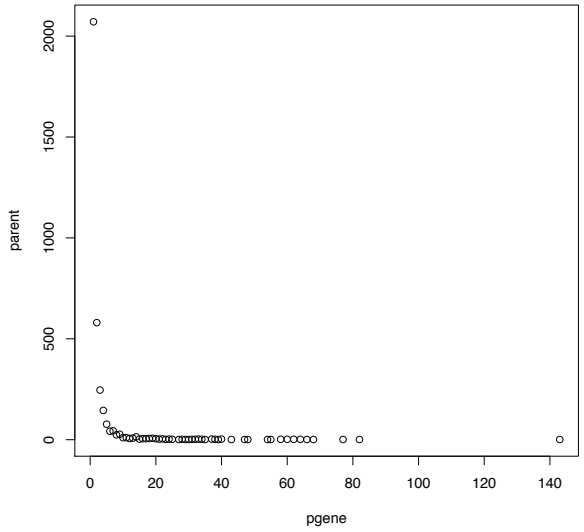
Identify Pseudogene Parents

Three sources for pseudogene parents:

- Manual annotation of parents (for Gencode v6)
- Unique high sequence similarity (>90%) identified by BLAT a pseudogene against the human genome
- Pseudopipe predictions

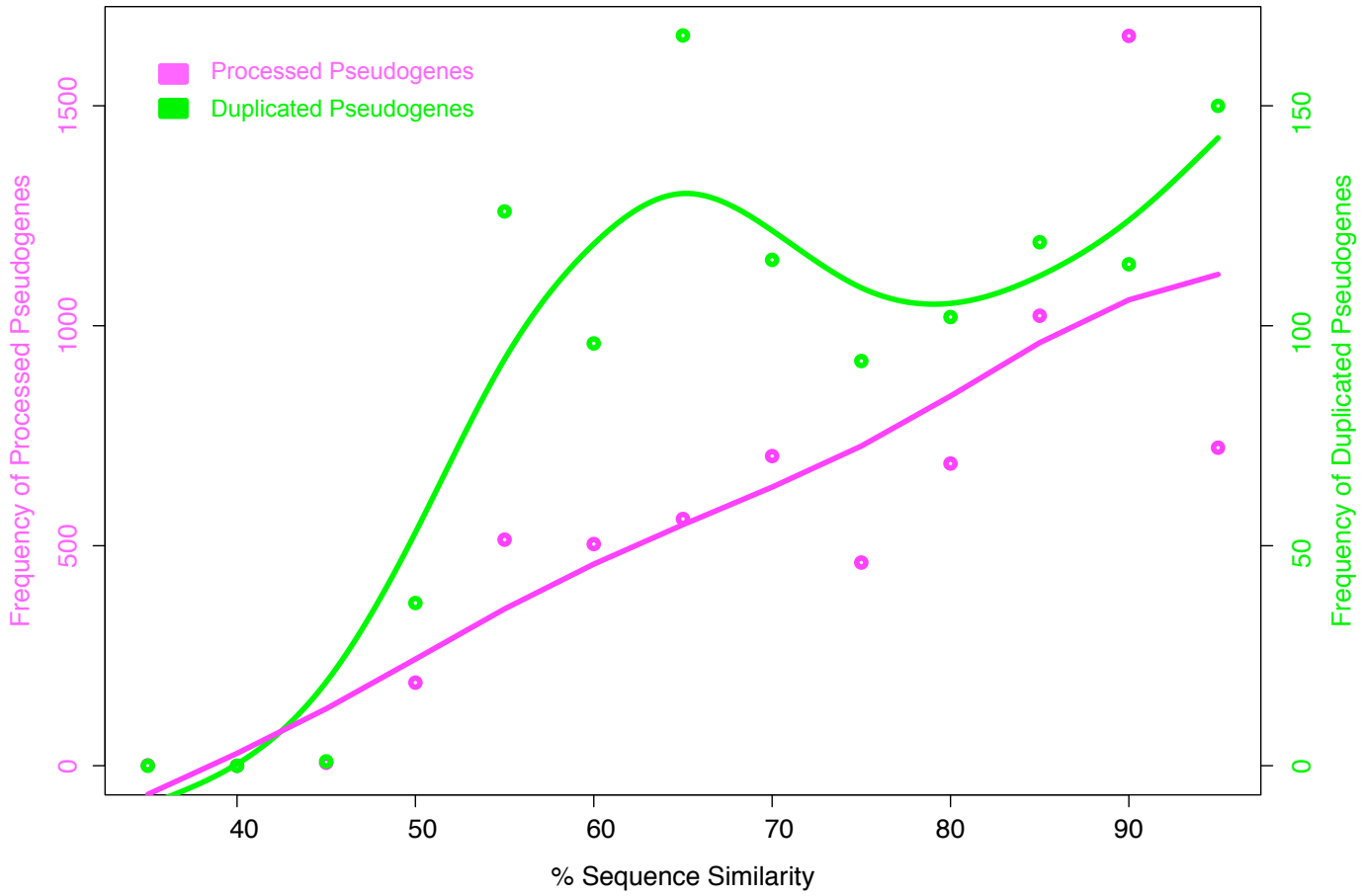
Results:

3,391 parent genes for 9,368 pseudogenes are identified, where the parents of the other pseudogenes are still ambiguous.

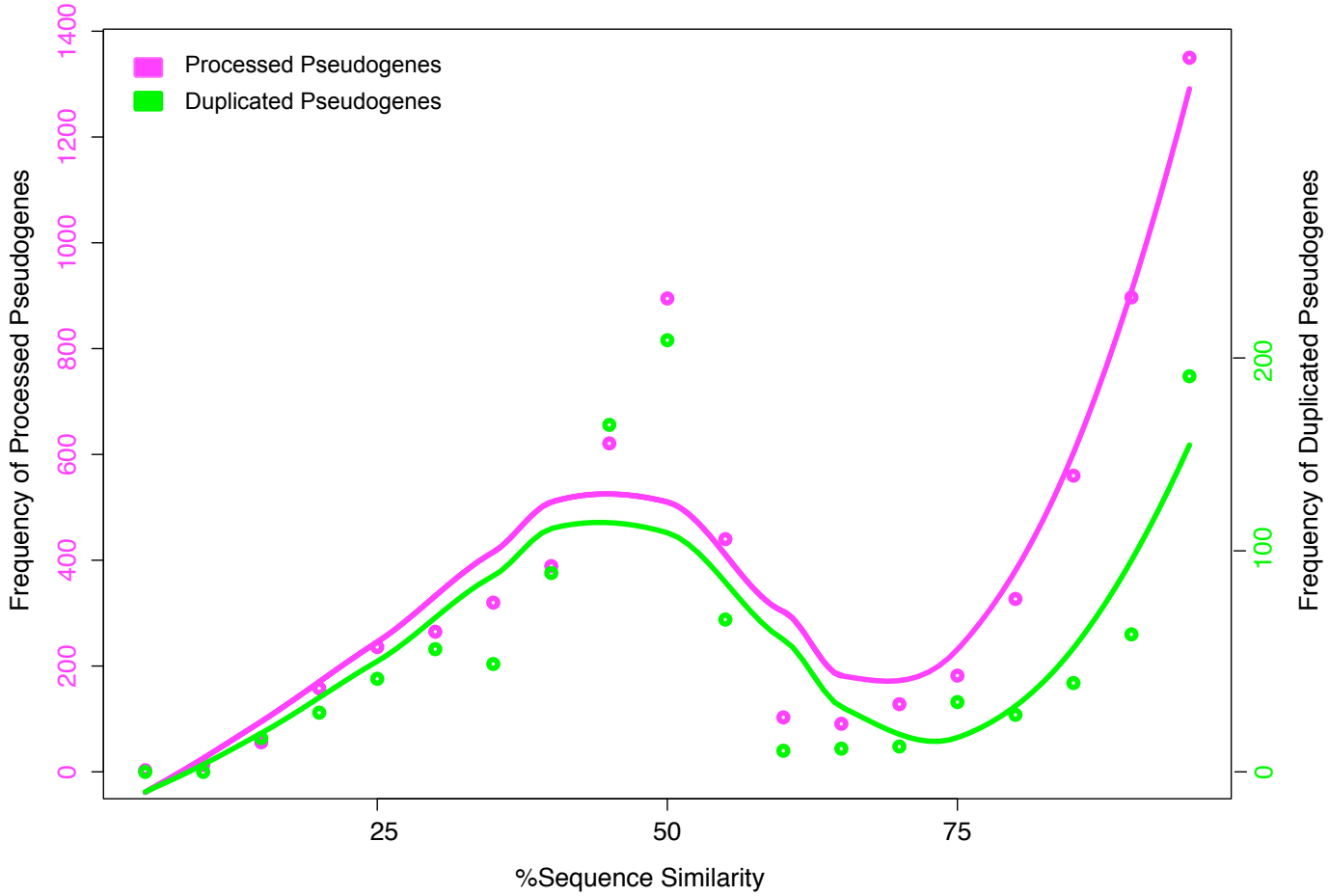


Parent	Number of pseudogenes
RPL21	143
RPL23A	82
RPL7A	77
GAPDH	68
RPL31	66
PPIA	64
RPL7	64
OR7E24	62
RPS2	62
RPS3A	60
HNRNPA1	60

Pseudogene Sequence Identity to Parent CDS



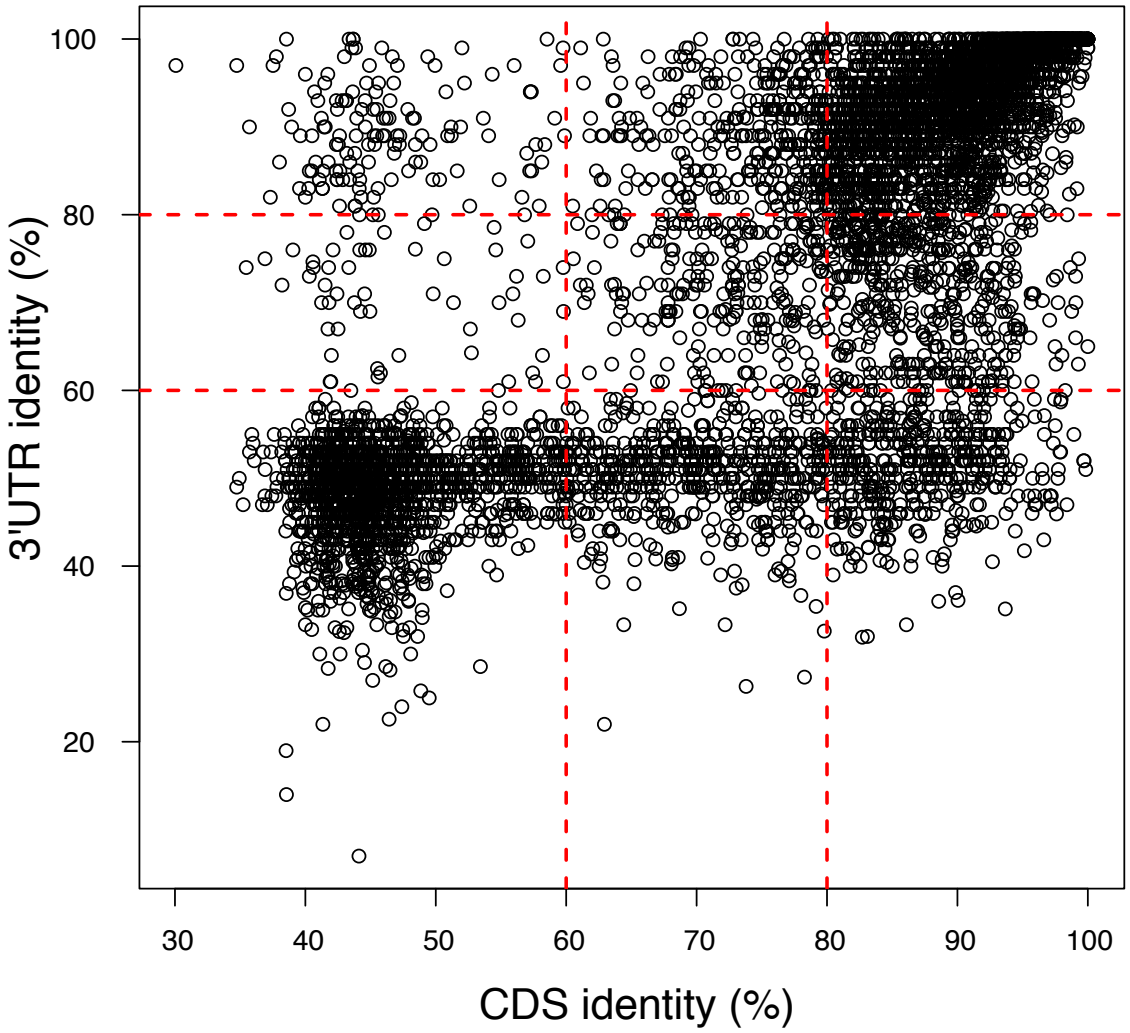
Pseudogene Sequence Identity to Parent 3' UTR



3' UTR alignment:

Pseudogenes are extended for 2 kb at 3' ends and aligned to the 3' UTR of parent genes.
Sequence identities are from a 100 bp sliding window.

Pseudogene Sequence Identity to Parent



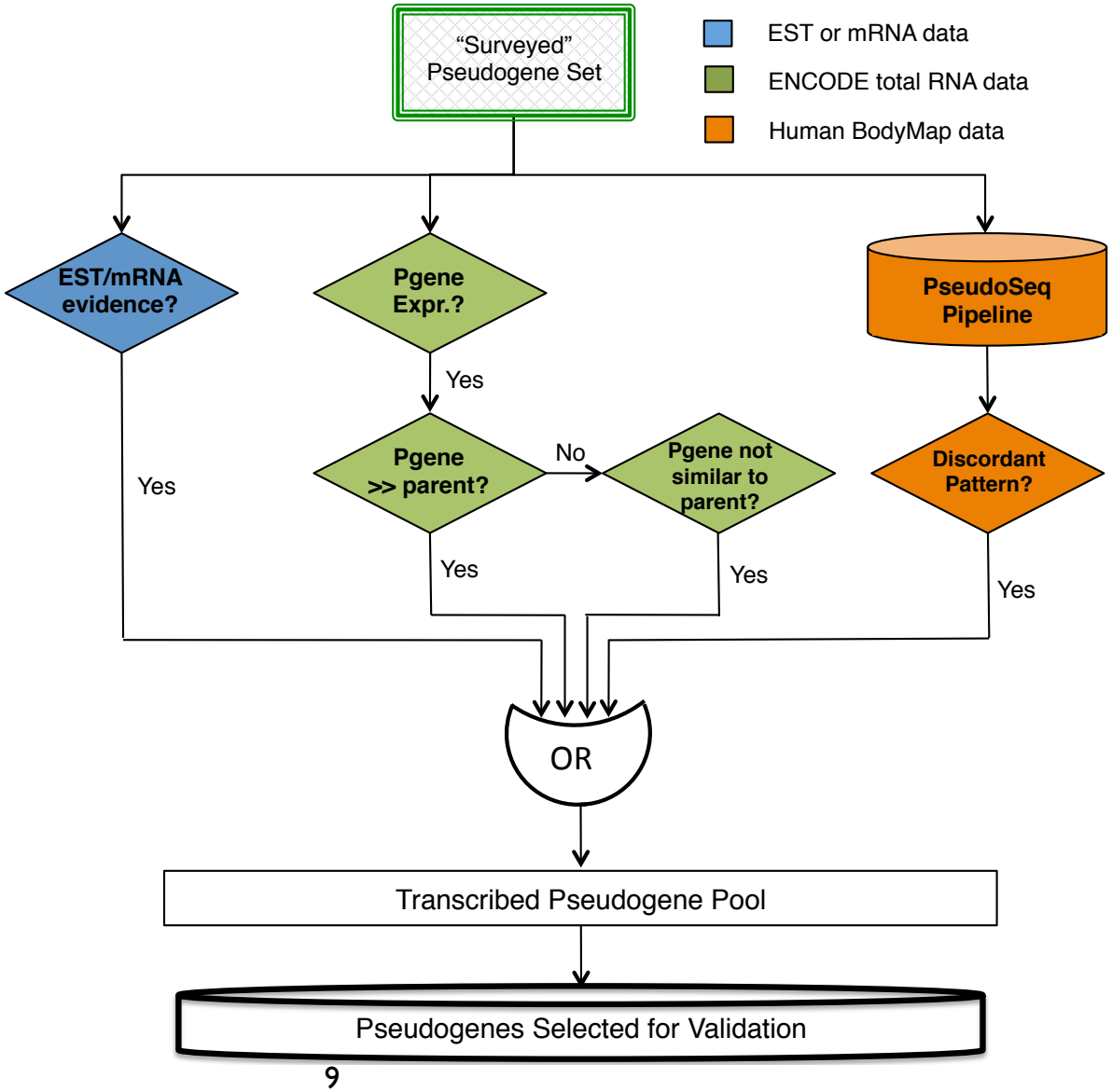
Pipeline to Identify Transcribed Pseudogenes

876 transcribed pseudogenes:

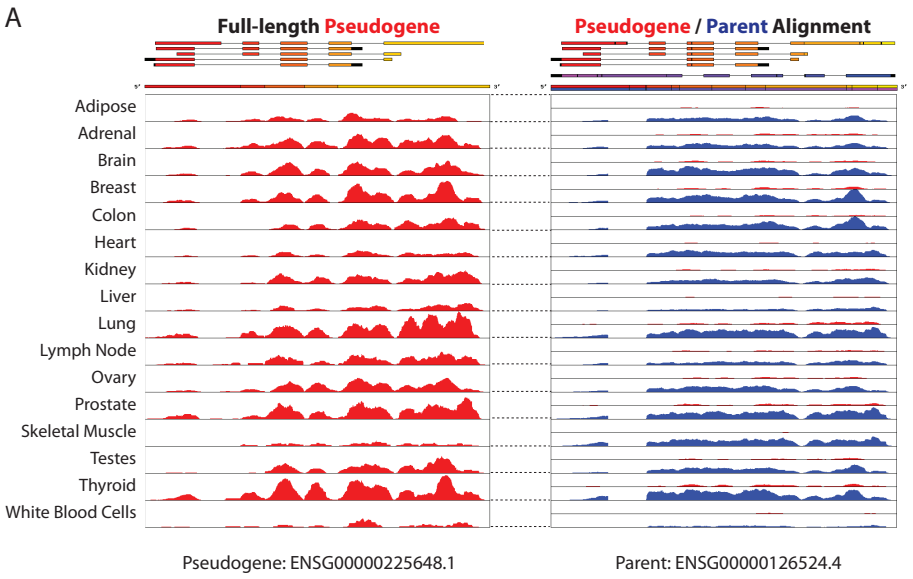
- 422 from EST evidence;

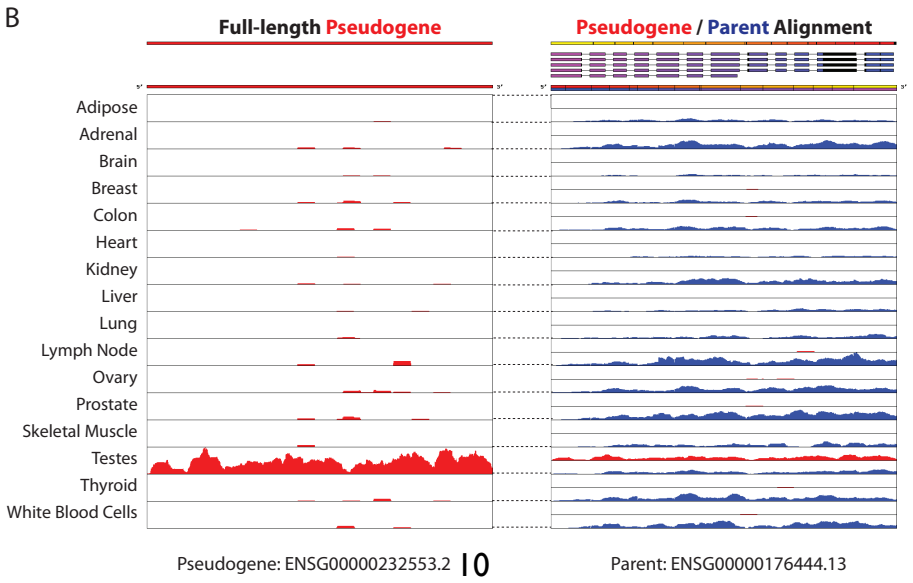
- 344 from pseudoSeq pipeline on BodyMap data;

- 110 from total RNA data of GM12878 and K562.

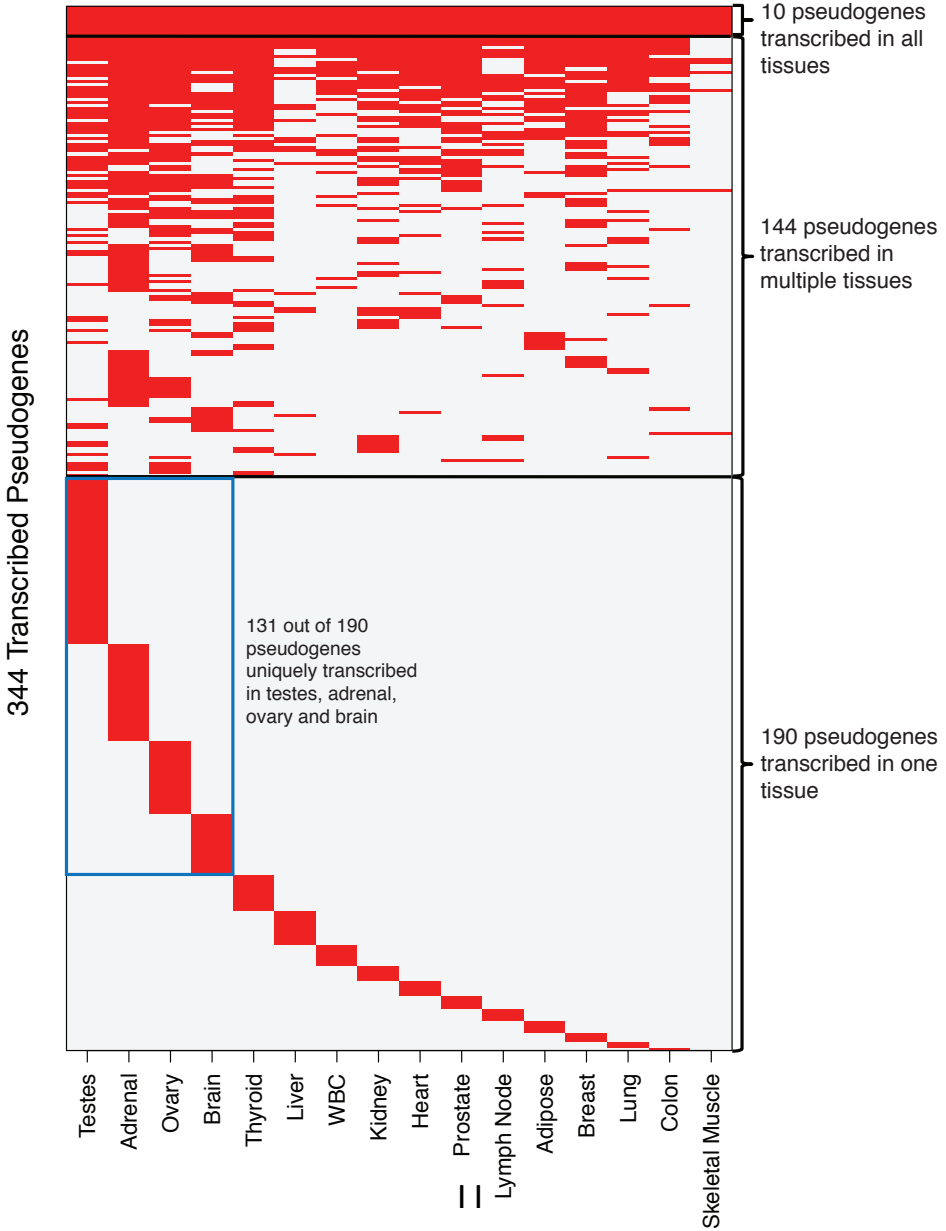


Transcribed Pseudogenes by PseudoSeq





Transcribed Pseudogenes by PseudoSeq

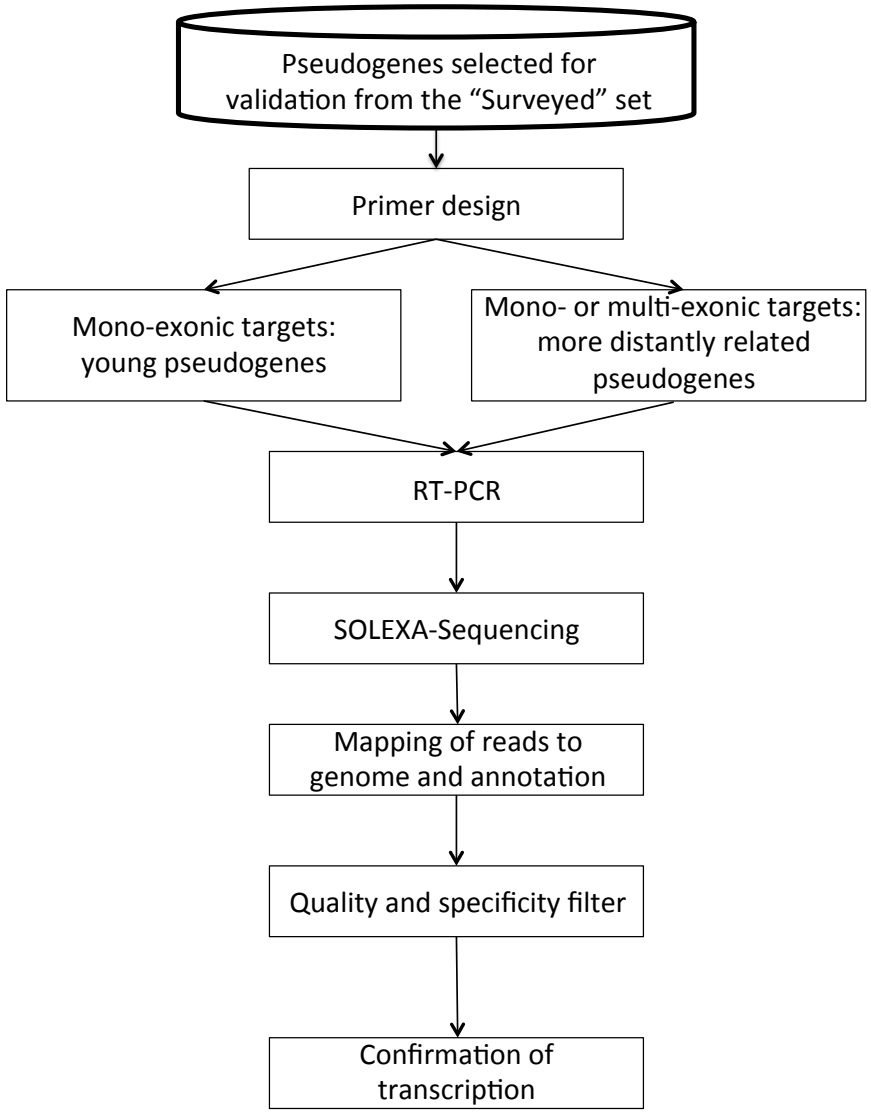


Validation of Transcribed Pseudogene

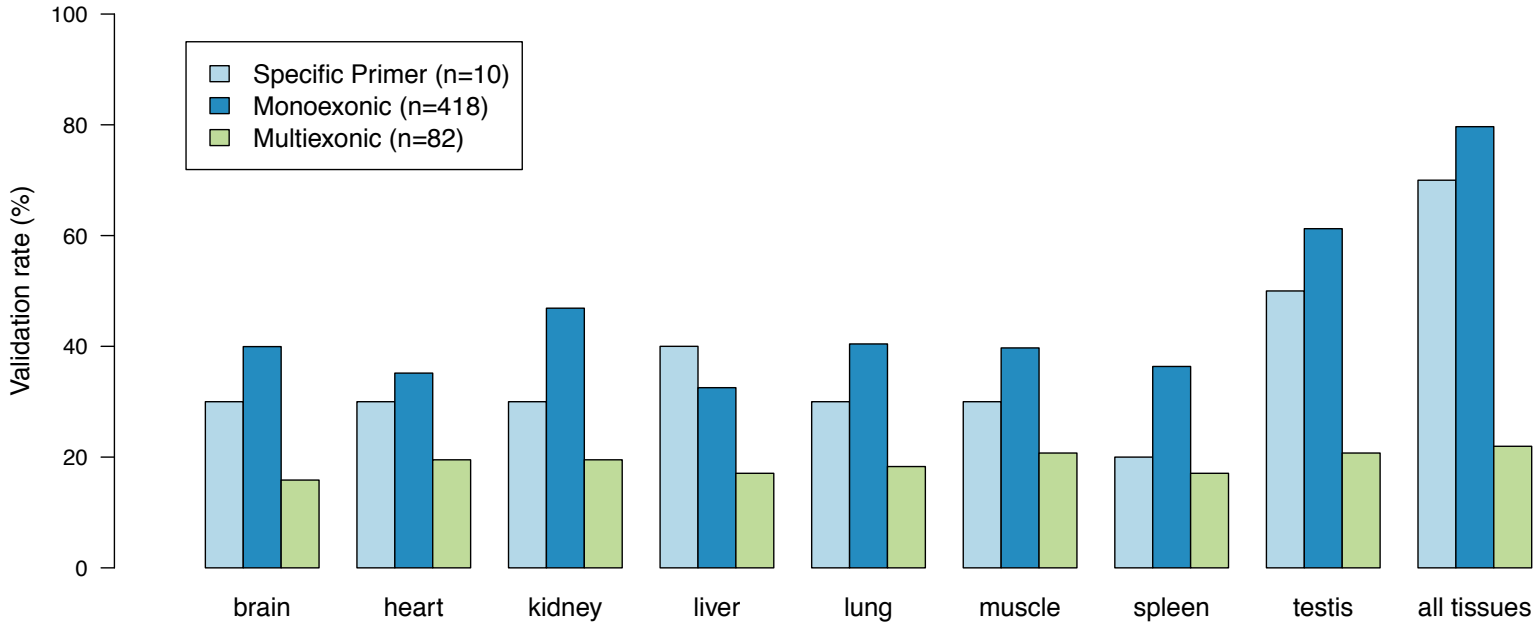
Mono-exonic RT-PCR:
Target to pseudogene exons. One target for each pseudogene;

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



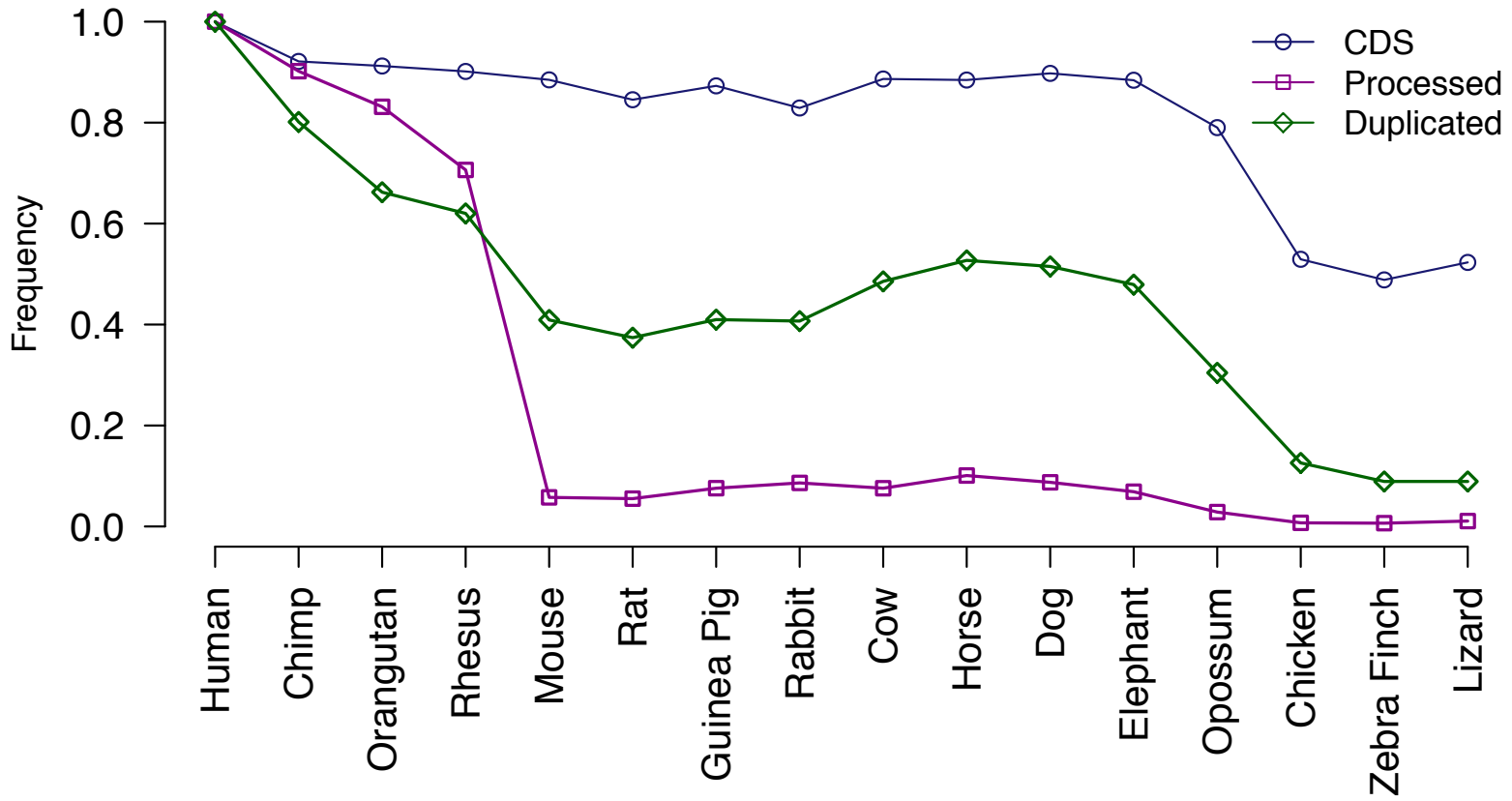
Total number of transcribed pseudogenes being validated: 469

- 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

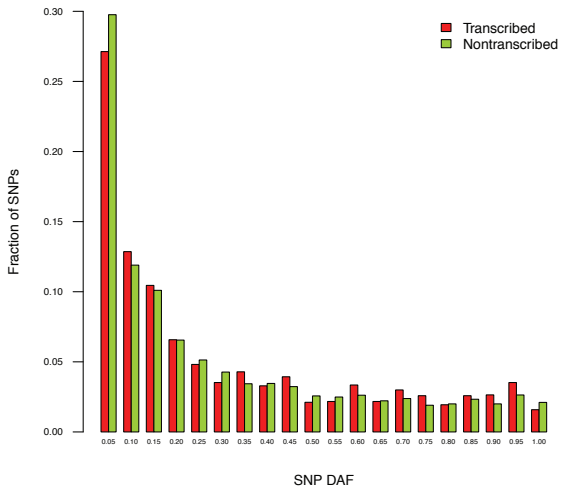
Overall validation rate: 75.5% (354 out of 469)

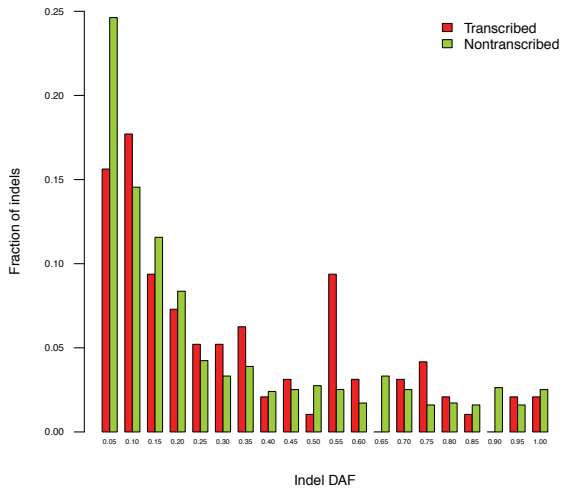
- Specific primer: 70% (7 out of 10)
- Monoexonic: 79.7% (333 out of 418)
- Multiexonic: 22.0% (18 ouf of 82)

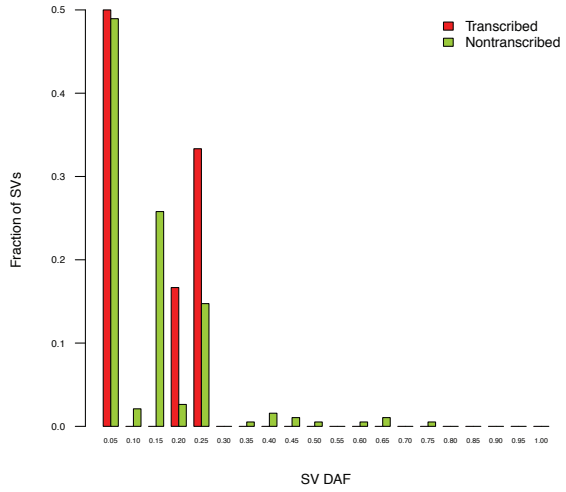
Sequence Preservation

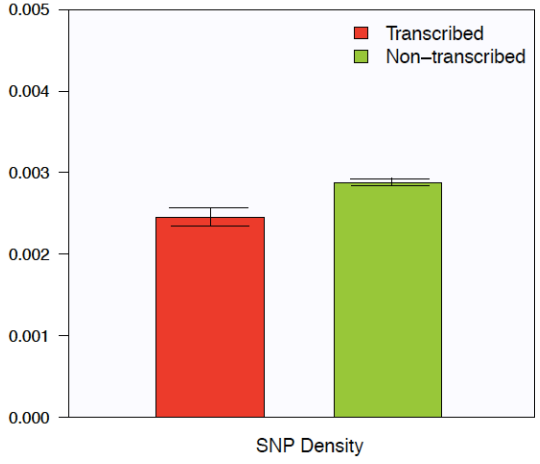


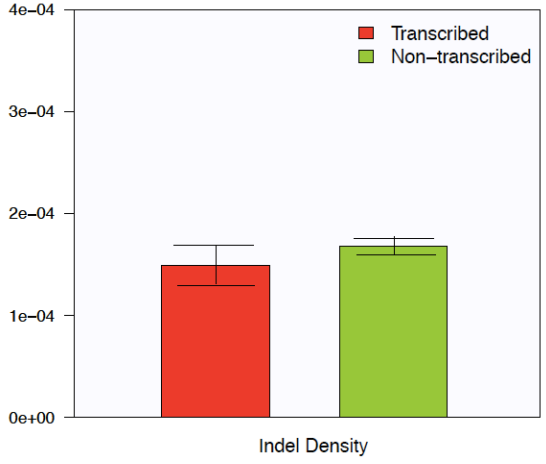
Selection Pressure on Transcribed Pseudogenes

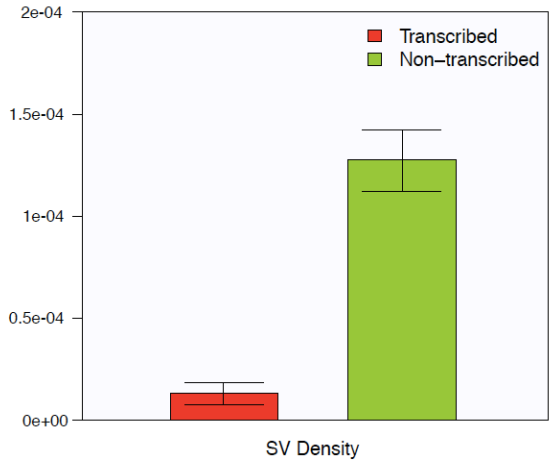










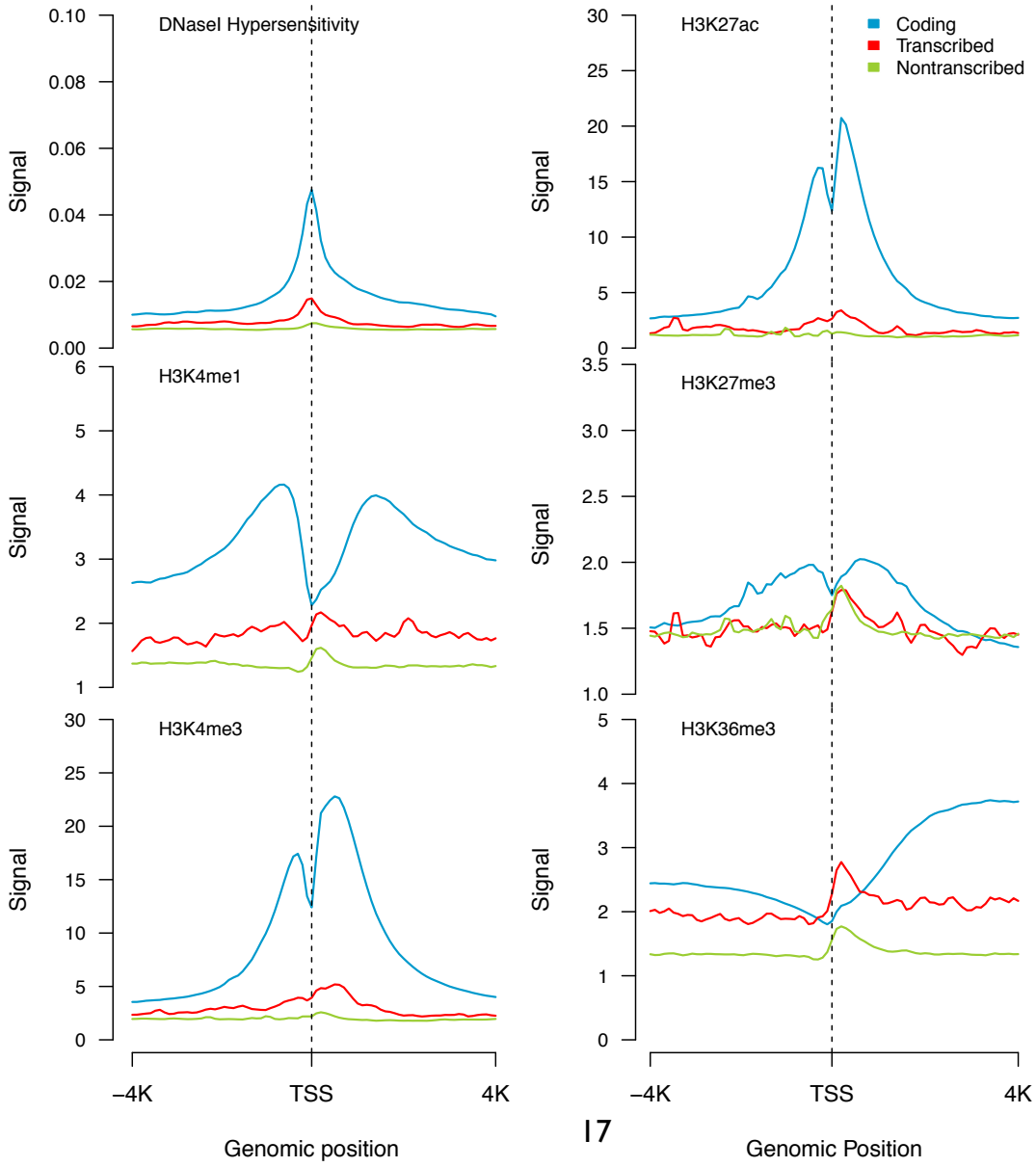


No differences in evolutionary selection are detected between transcribed and non-transcribed pseudogenes

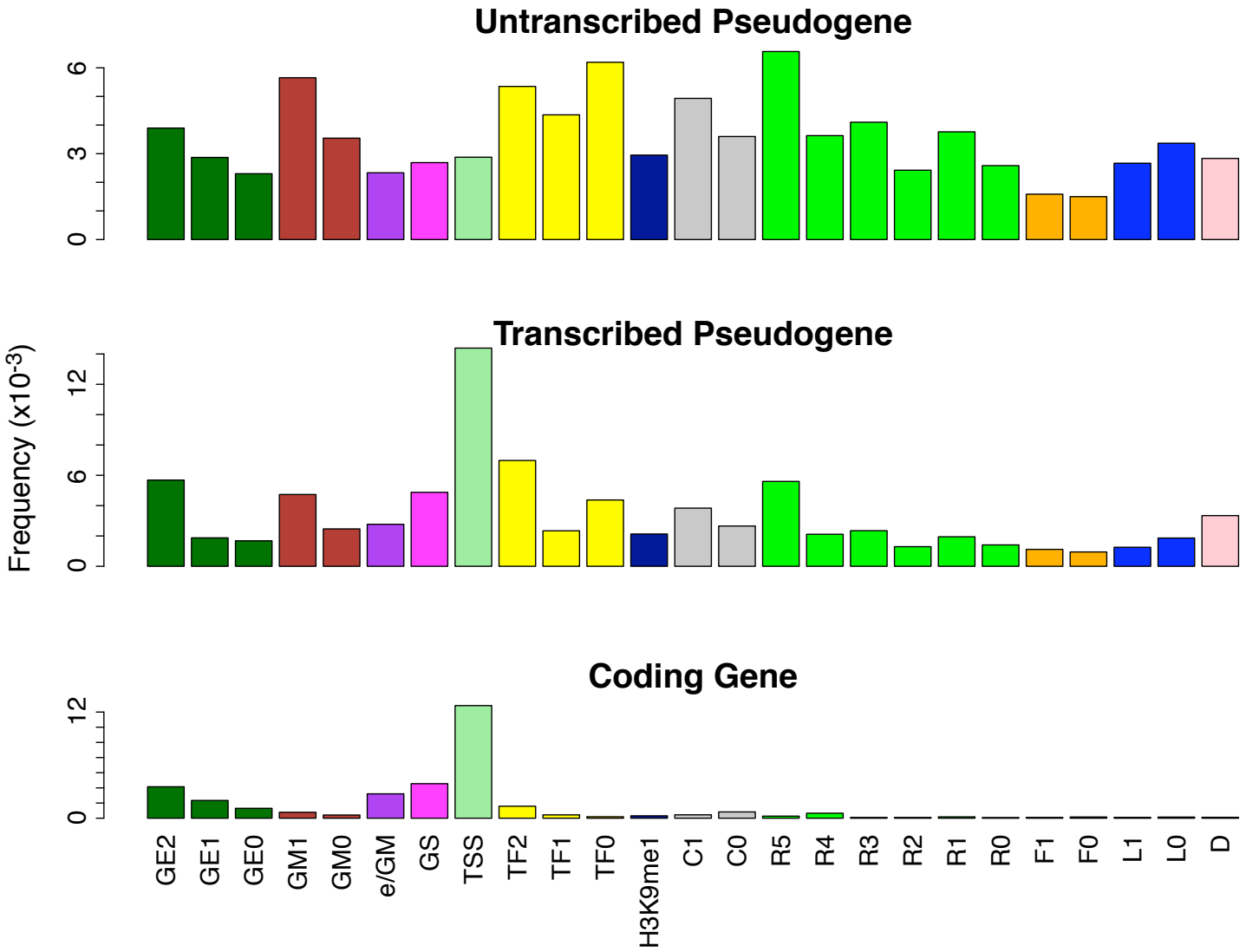
Evolutionary Constraints of Pseudogenes

- Human pseudogene sequences are aligned to chimp and mouse genome;
- Substitution rate of each pseudogene is calculated;
- Assume the substitution follows Poisson distribution, the pseudogenes with significantly less substitution are considered as conserved. The background substitution rate is set at 0.015 for chimp and 0.5 for mouse;
- Error rate in multiple hypothesis testing is controlled by setting FDR to 0.05
- 1,019 pseudogenes are calculated as conserved.

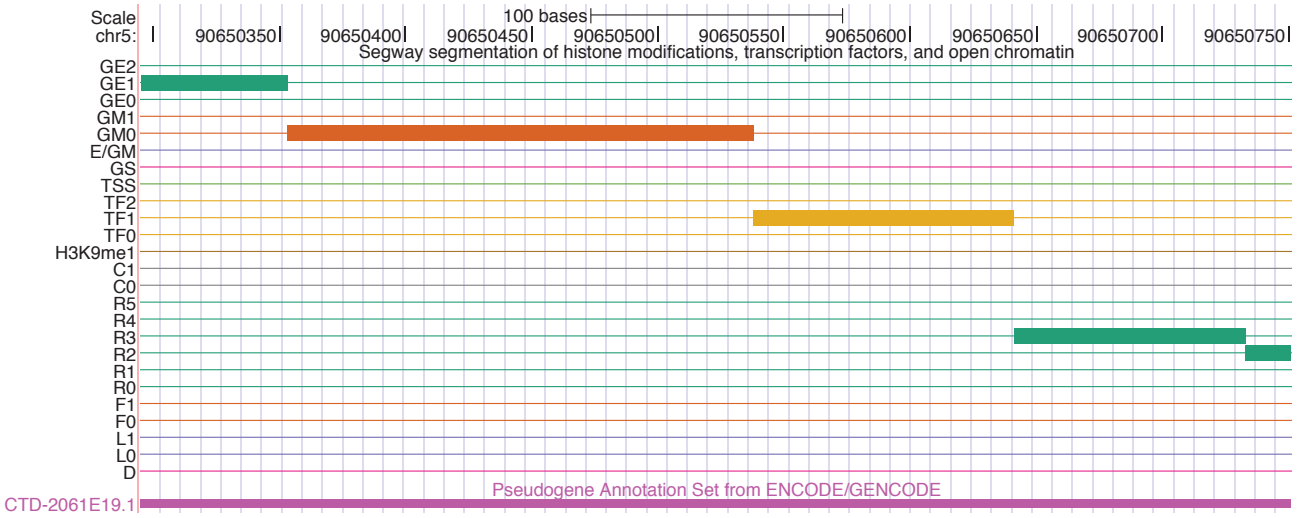
Chromatin Signatures of Pseudogenes

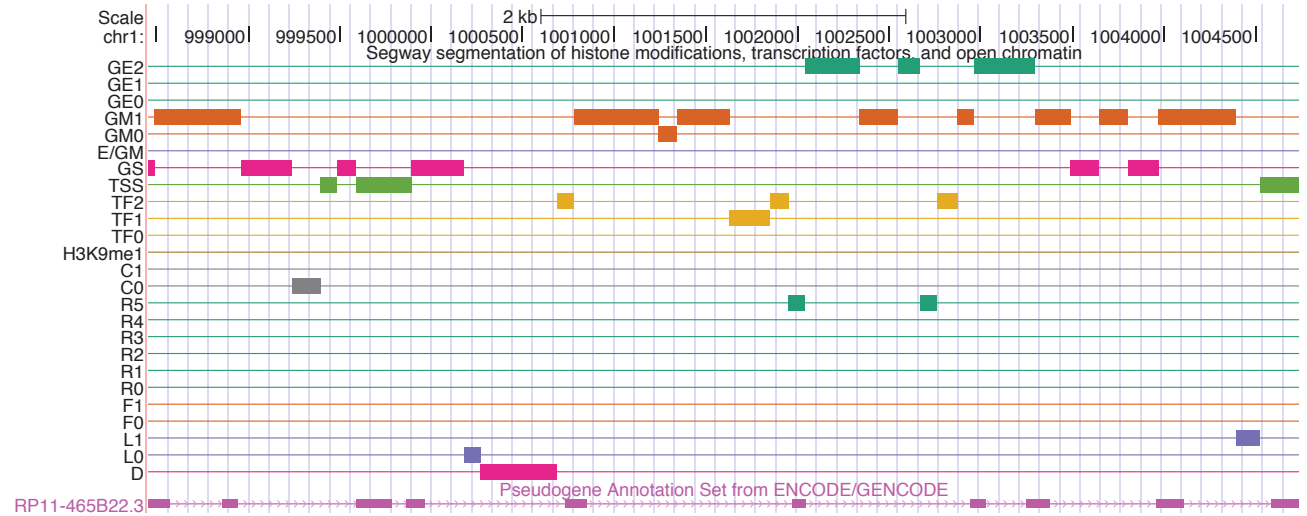


Chromatin State Segmentation



Examples of Transcribed Pseudogenes





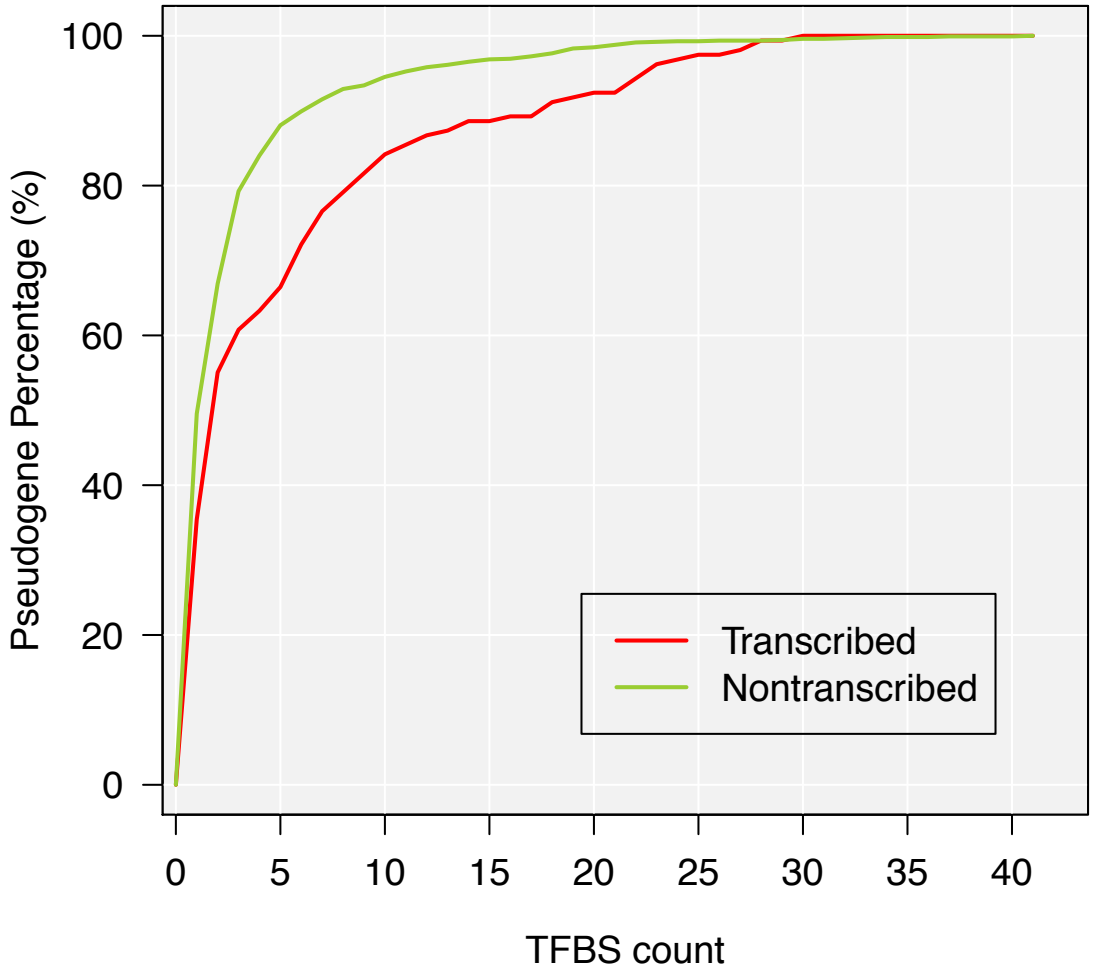
Transcription Factor Binding Sites of Pseudogenes

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

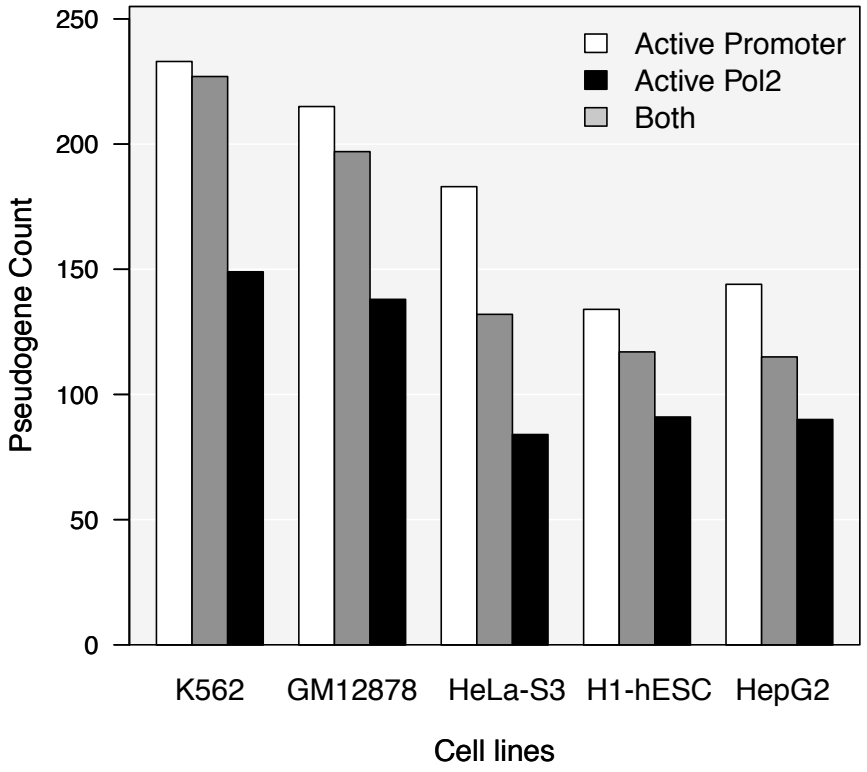


Pseudogenes with Active Upstream Sequences

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

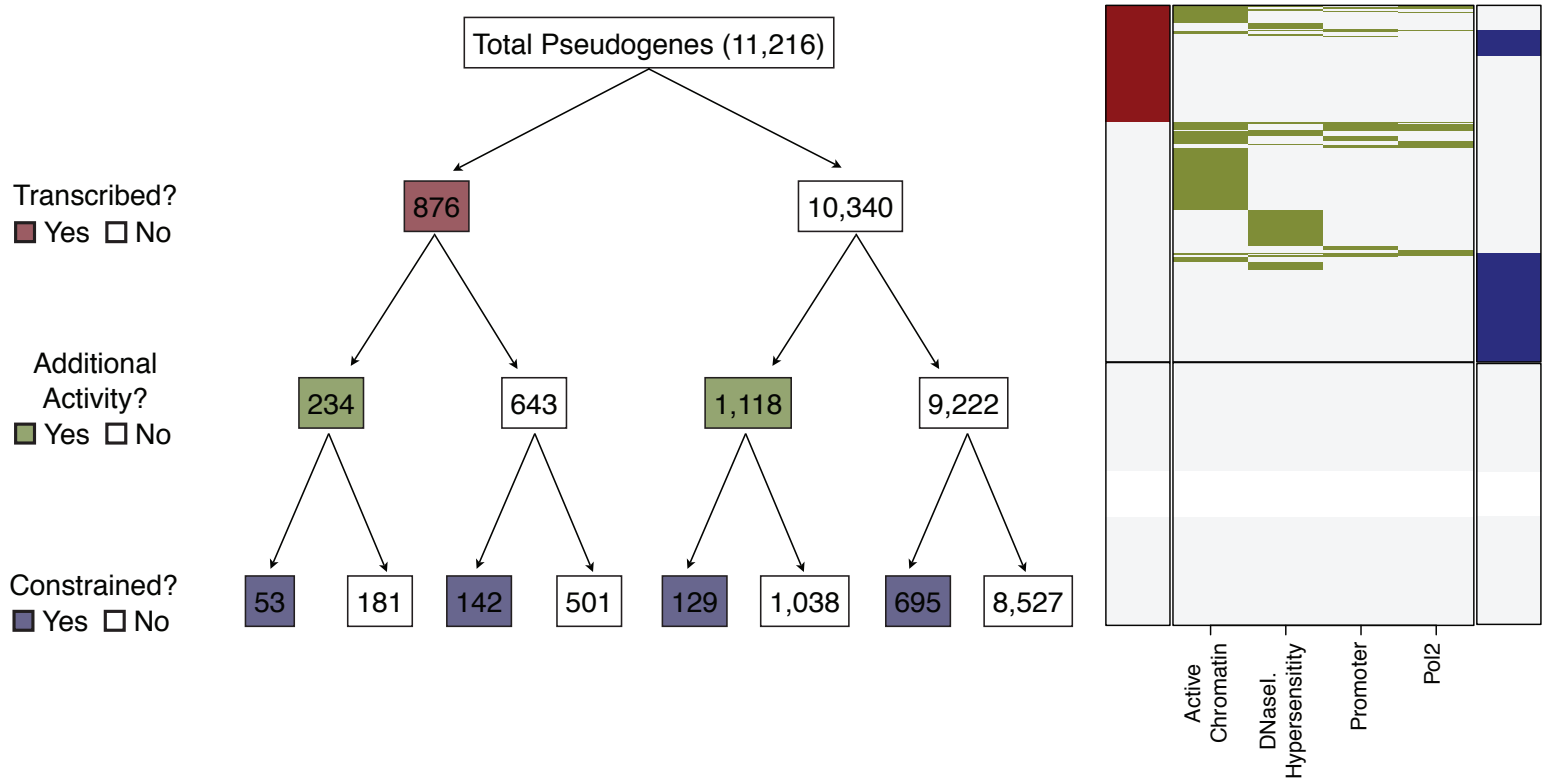
Active Pol2 bindings are from upper 5% of Pol2 binding peaks, in terms of peak widths and heights, plus binding of Pol2 co-factors;

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



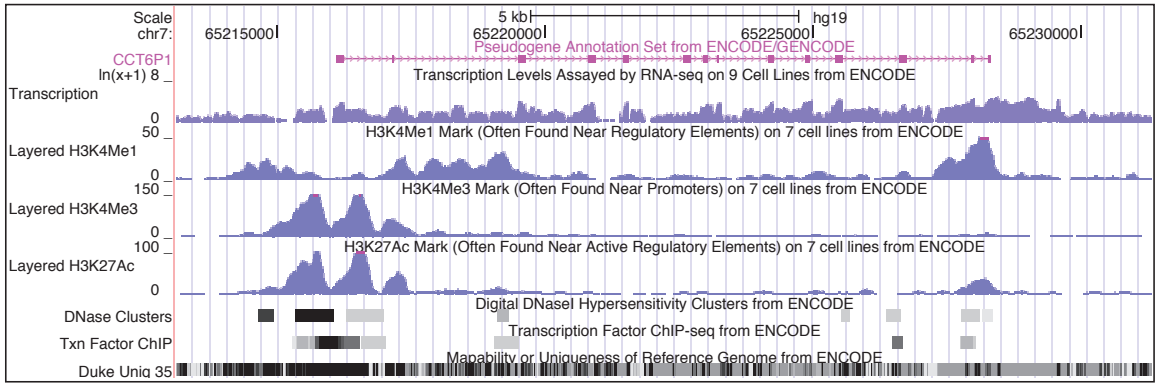
	K562	Gm12878	Helas3	H1hesc	Hepg2
K562	-	0.30	0.29	0.22	0.27
Gm12878	0.33	-	0.33	0.27	0.32
Helas3	0.31	0.31	-	0.30	0.39
H1hesc	0.24	0.27	0.29	-	0.27
Hepg2	0.26	0.32	0.33	0.33	-

Partial Activity of Pseudogenes

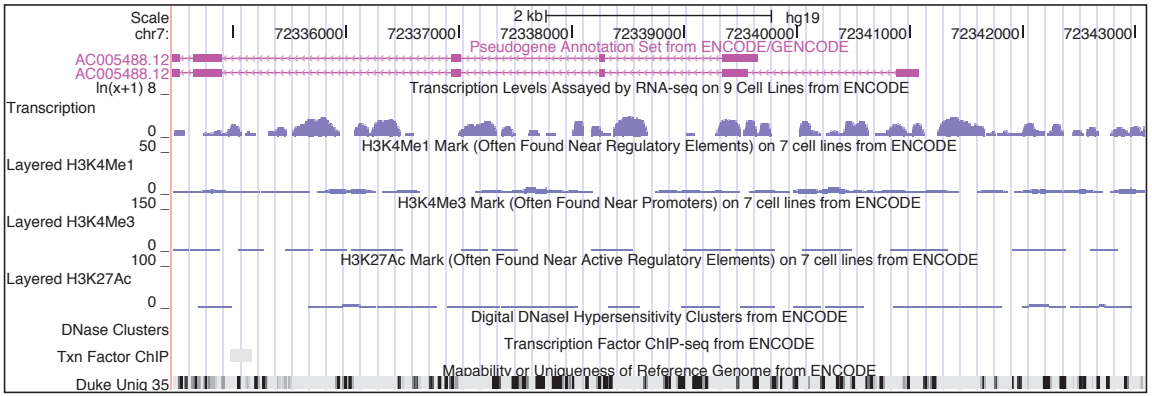


Case Studies of Pseudogenes Activity

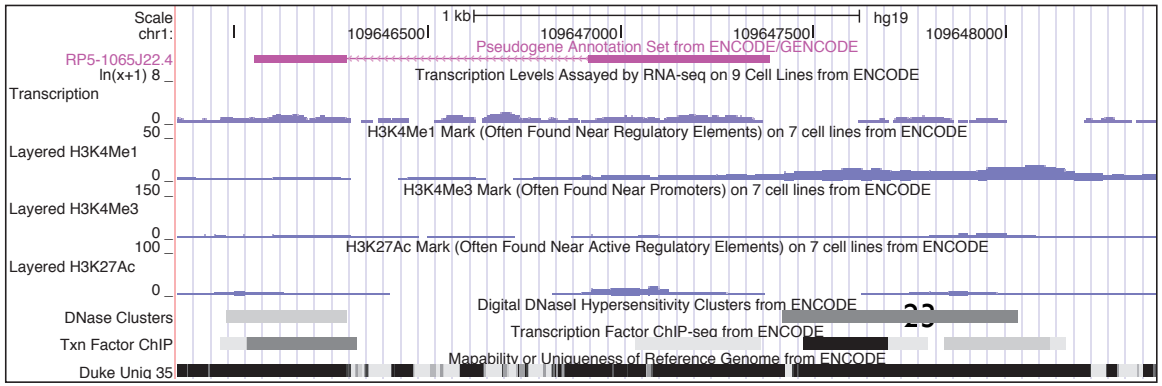
Transcribed With Additional Activity



Transcribed Only



Partially Active



Summary

- Genome-wide pseudogene annotation by combining automated pipelines and manual annotation;
- Identify parents for pseudogenes and examine their sequence similarities;
- Pseudogene transcription, experimental validation and tissue specificity;
- Evolutionary constraints on pseudogenes;
- Chromatin signatures of pseudogenes;
- Upstream regulatory elements of pseudogenes;
- Summarize the partial activity of pseudogenes into a resource file psiDR

Future Studies

• Expand current study to the more up-to-date Gencode annotation;

• Integrate data for pseudogenes activities, sequences and evolutionary constraints to predict their potential regulatory roles;

• Combine other genomic data, such as DNA methylation, ChIA-PET and HITS-CLIP, to study the pseudogenes activities and their regulation;

• Comparative study with mouse, worm and fly pseudogenes;

• Study the upstream regions of pseudogenes for their co-evolution with pseudogene exons, and its possible relationships with pseudogene activities

Acknowledgement

Sanger
- Adam Frankish
- Jeniffer Harrow
- Tim Hubbard

University of Lausanne
- Cedric Howald
- Alexandre Reymond

CRG
- Andrea Tanzer

UCSC
- Rachel Harte
- Mark Diekhans

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