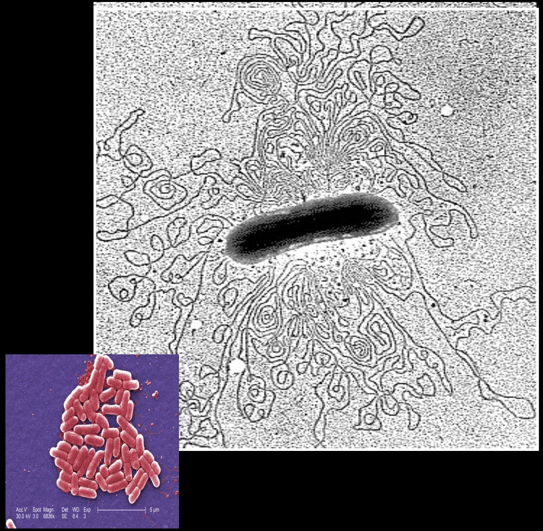
PSEUDOGENE DERIVED lncRNAs

Reason 1:
The non-coding genome (r)evolution

E.coli



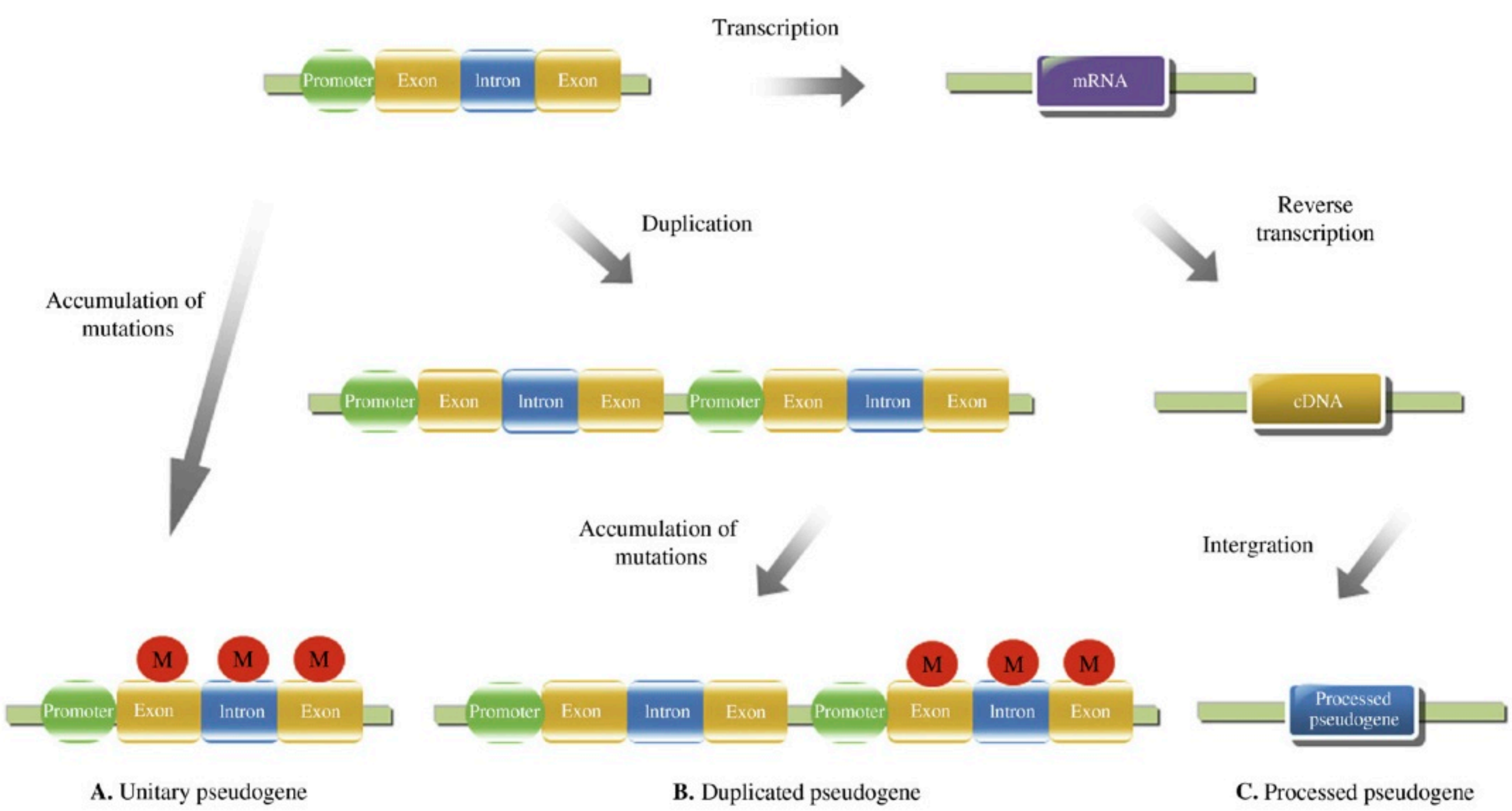
C. elegans

H. sapiens

Genome	5x10^{6} bp	1x10^{8} bp	3x10^{9} bp
Chromosomes	1	6	23
Coding genes	6692	20541	21995
ncDNA	5%	60%	**98%**
non-coding RNA genes	15	23136	ca. 40000
miRNAs	0	224	4274
pseudogenes	21	1522	10616

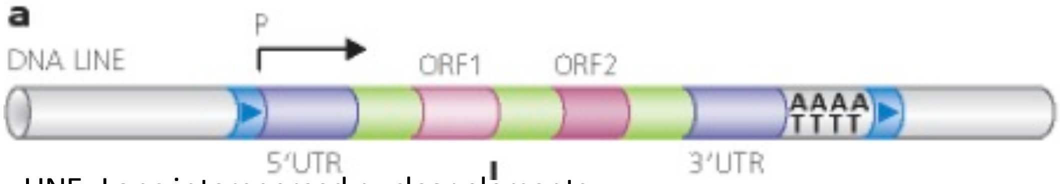
ENSEMBL 11/2014

Protein coding genes give rise to pseudogenes



DNA based **DNA based** **RNA intermediate**

Transposition of Retrotransposons



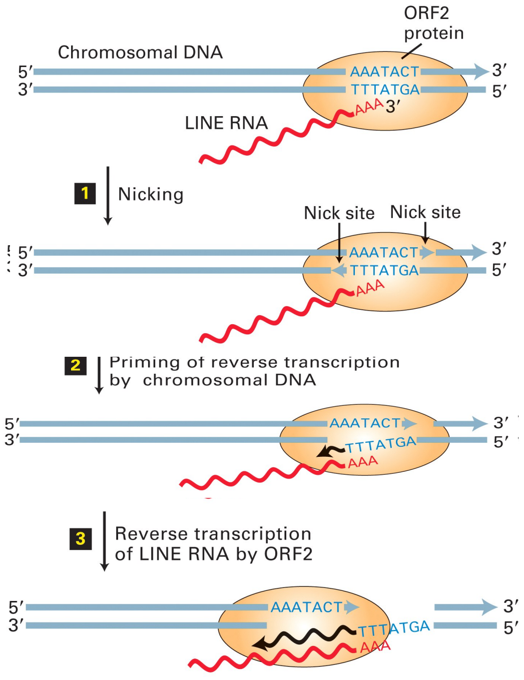
LINE: Long interspersed nuclear elements

LINE elements (L1,L2,L3)

(21% of genome; 800.000 copies)

ORF1: RNA binding protein

ORF2: Endonucelase, Reverse transcriptase

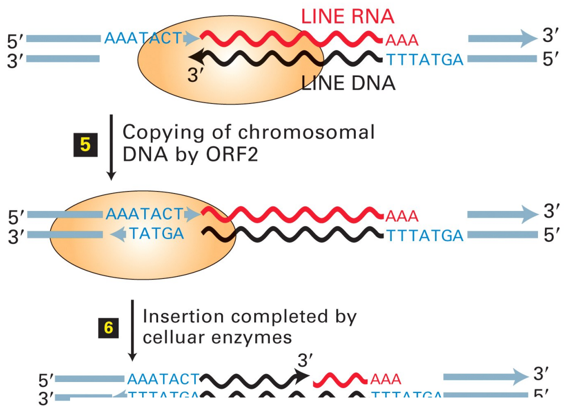

ORF2 protein
Chromosomal DNA
5′
3′
AAATACT
TTTATGA
3′
5′
LINE RNA
AAA
1
Nicking
Nick site
Nick site
2
Priming of reverse transcription by chromosomal DNA
3
Reverse transcription of LINE RNA by ORF2

ORF1 protein binds LINE RNA; localizes to TTT rich sequences

ORF2 enonuclease activity nicks target sequence

priming

ORF2 reverse transcriptase activity→ makes reverse transcription

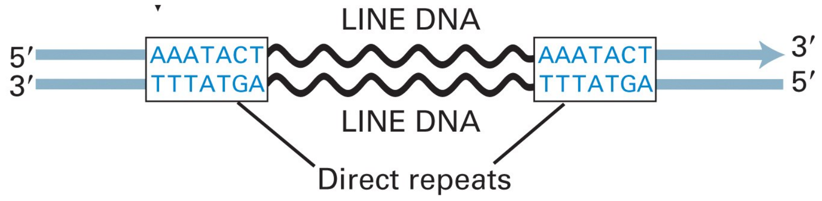

LINE RNA
5′
3′
AAATACT
AAA
TTTATGA
3′
5′
LINE DNA
5
Copying of chromosomal DNA by ORF2
TATGA
6
Insertion completed by celluar enzymes

ORF2 reverse transcriptase activity→ makes reverse transcription

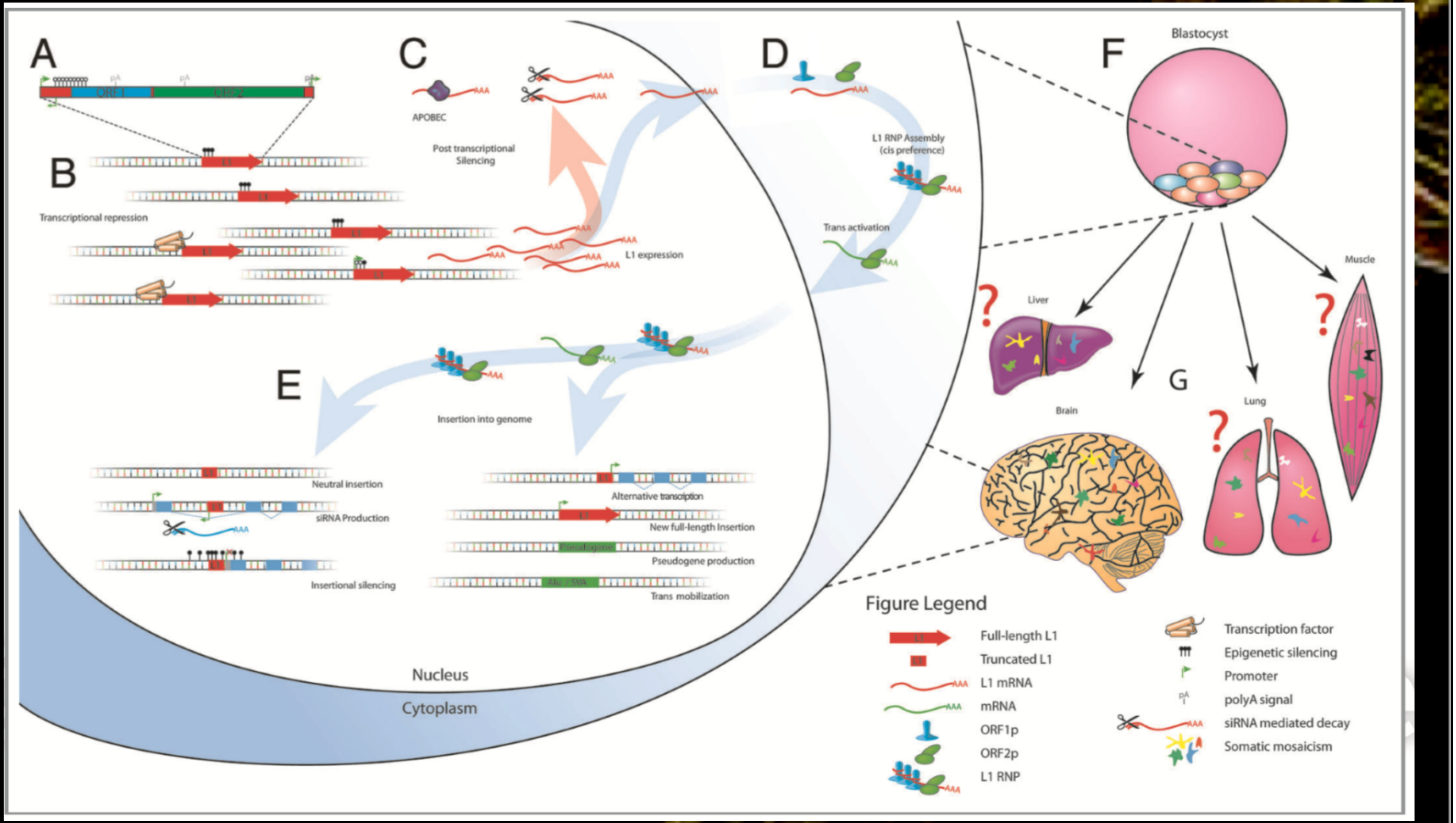
Generation of direct repeats in target site

DNA repair by host cell

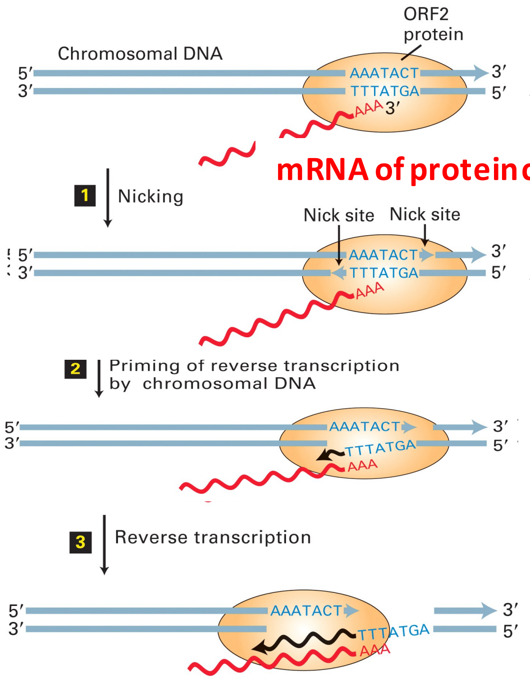
FINAL PRODUCT



Retrotransposons can change genetic context



Retro-transposition machinery hijacks endogenous mRNAs



ORF1 protein binds LINE RNA; localizes to TTT rich sequences

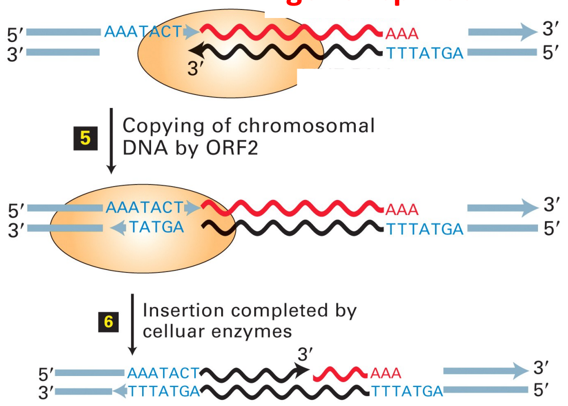
mRNA of protein coding gene - spliced

ORF2 enonuclease activity nicks target sequence

priming

ORF2 reverse transcriptase activity→ makes reverse transcription

mRNA protein coding gene - spliced



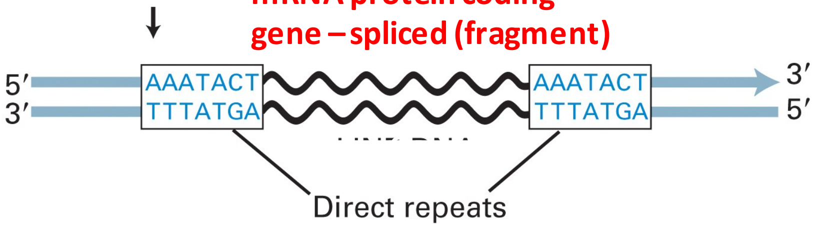
ORF2 reverse transcriptase activity→ makes reverse transcription

Generation of direct repeats in target site

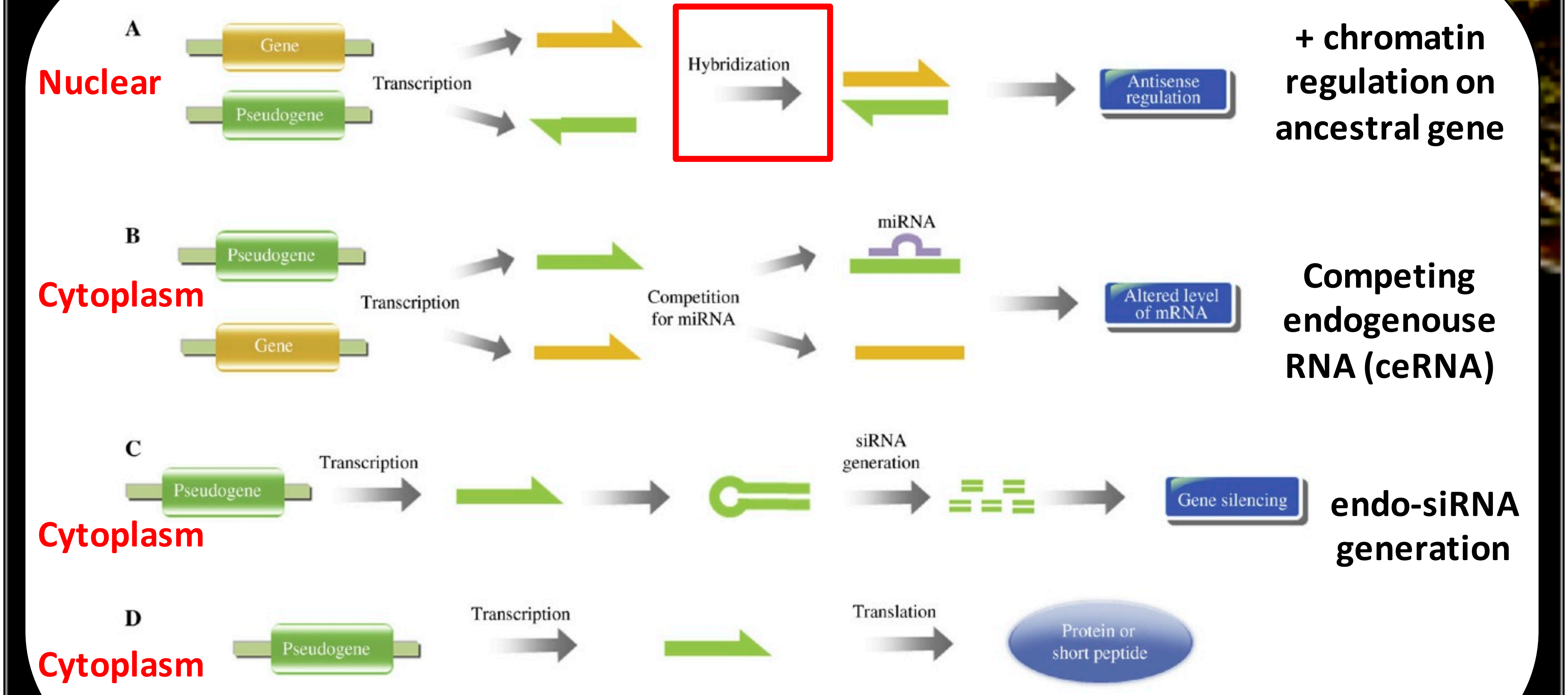
DNA repair by host cell

FINAL PRODUCT: PROCESSED PSEUDOGENE

mRNA protein coding gene – spliced (fragment)



Pseudogene derived RNAs can acquire new functions



Nuclear

+ chromatin regulation on ancestral gene

Cytoplasm

Competing endogenouse RNA (ceRNA)

Cytoplasm

endo-siRNA generation

Cytoplasm

PSEUDOGENE BIOTYPES

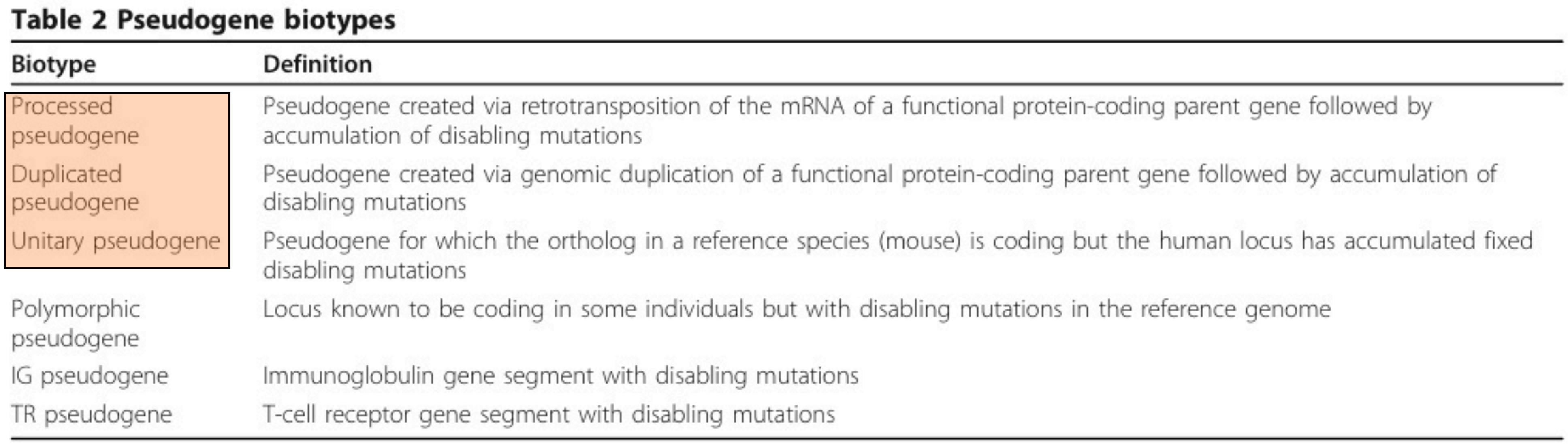


Table 2 Pseudogene biotypes

Biotype	Definition
Processed pseudogene	Pseudogene created via retrotransposition of the mRNA of a functional protein-coding parent gene followed by accumulation of disabling mutations
Duplicated pseudogene	Pseudogene created via genomic duplication of a functional protein-coding parent gene followed by accumulation of disabling mutations
Unitary pseudogene	Pseudogene for which the ortholog in a reference species (mouse) is coding but the human locus has accumulated fixed disabling mutations
Polymorphic pseudogene	Locus known to be coding in some individuals but with disabling mutations in the reference genome
IG pseudogene	Immunoglobulin gene segment with disabling mutations
TR pseudogene	T-cell receptor gene segment with disabling mutations

Duplicated/Unitary pseudogenes: can bring regulatory sequences, often spliced

Processed pseudogenes: hitch hike on regulatory elements dispersed throughout throughout the genome

PSEUDOGENE BIOTYPES

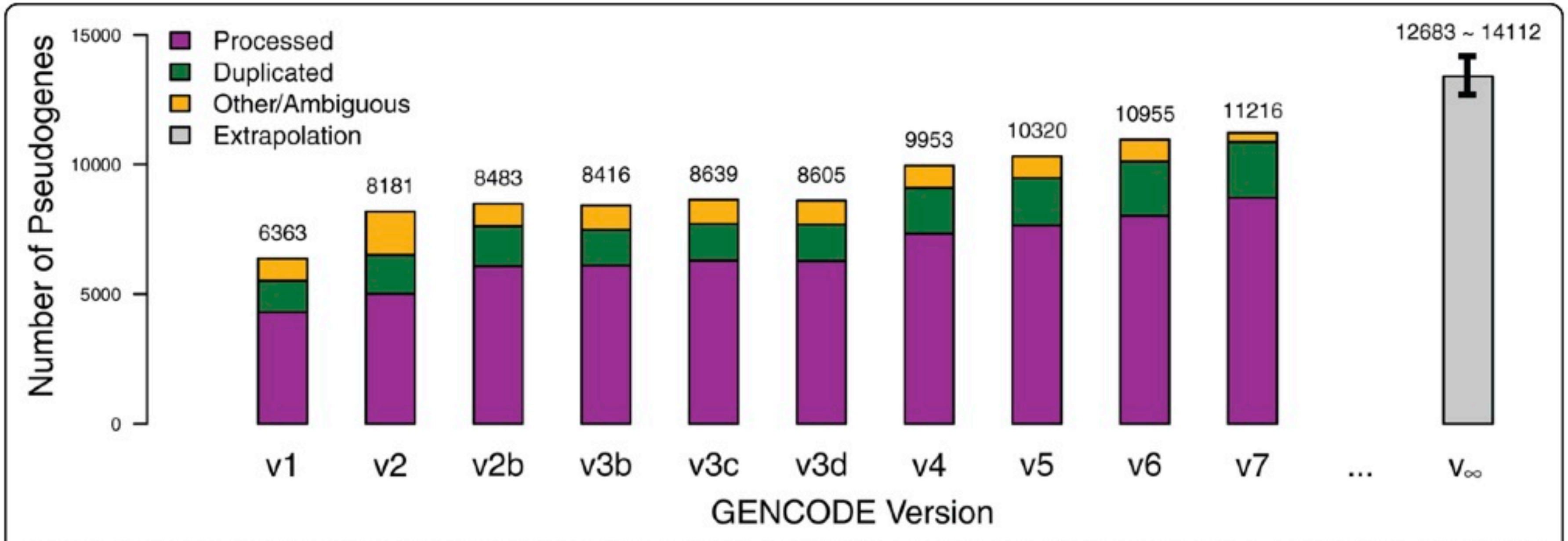
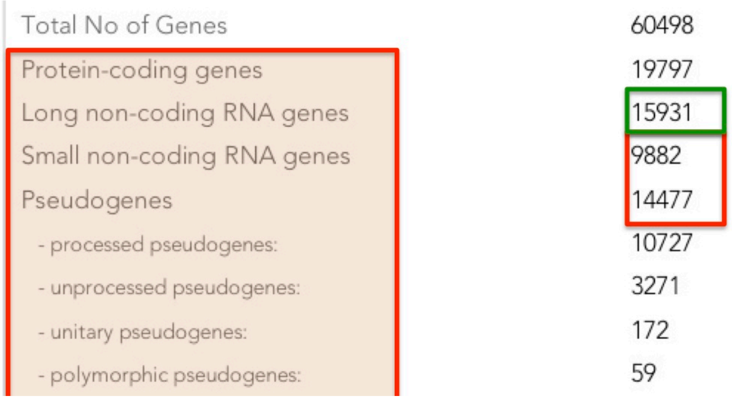


Figure 2 Growth of pseudogene annotation. The numbers of pseudogenes present in the GENCODE dataset from version 1 to version 7 are plotted. The three colors - purple, green and yellow - represent processed, duplicated and other types of pseudogenes, respectively. The pseudogenes were annotated manually and/or using the automated pipelines PseudoPipe and RetroFinder. The gray bar indicates the estimated number of pseudogenes (± standard deviation present in the human genome.

The majority of pseudogenes are processed pseudogenes:
Burst of retro-transposition events in recent phase of evolution



Total No of Genes	60498
Protein-coding genes	19797
Long non-coding RNA genes	15931
Small non-coding RNA genes	9882
Pseudogenes	14477
- processed pseudogenes:	10727
- unprocessed pseudogenes:	3271
- unitary pseudogenes:	172
- polymorphic pseudogenes:	59

GENOMICS STRATEGIES TO IDENTIFY AND CLASSIFY PSEUDOGENES

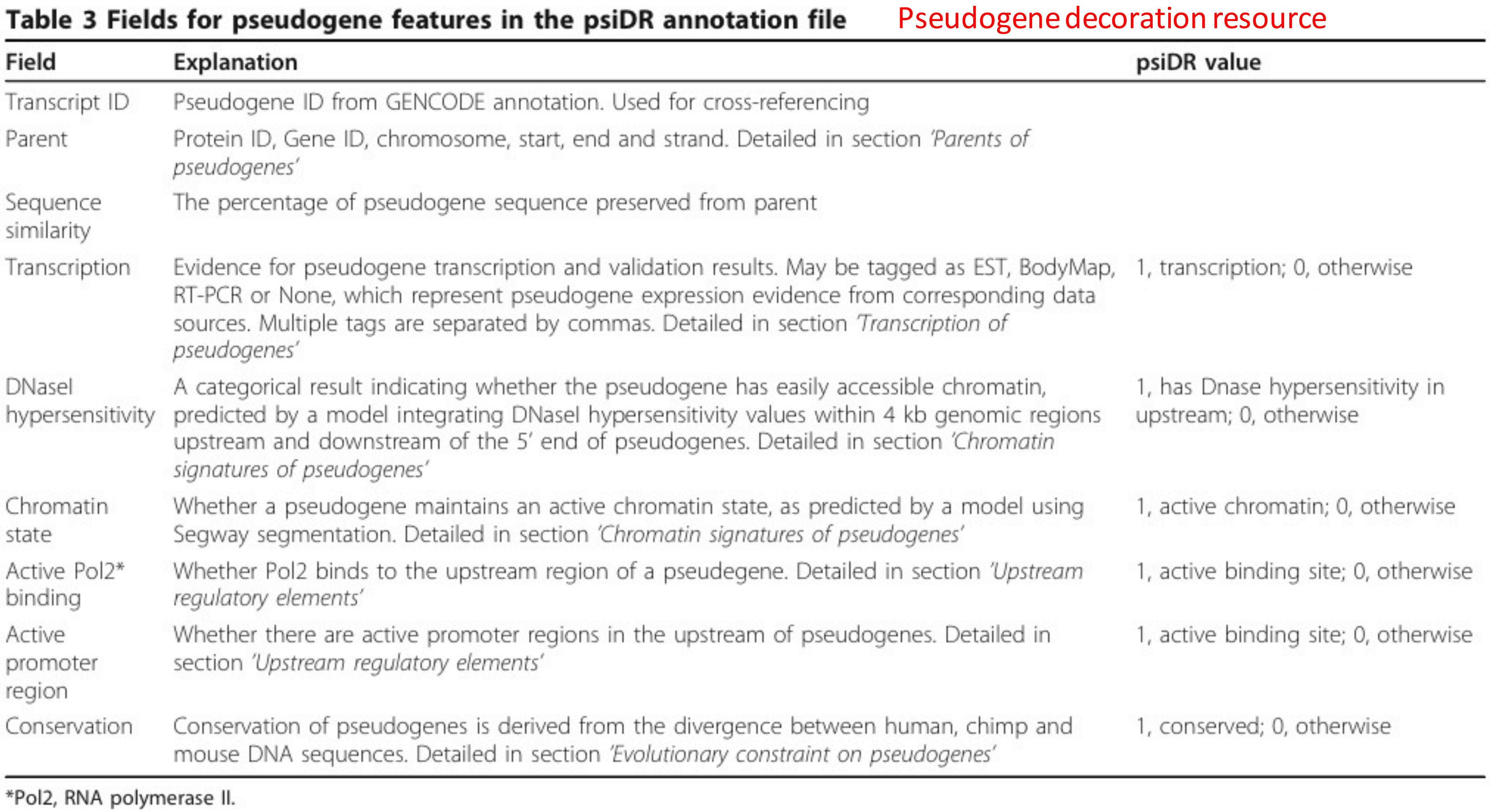


Table 3 Fields for pseudogene features in the psiDR annotation file

Pseudogene decoration resource

Field	Explanation	psiDR value
Transcript ID	Pseudogene ID from GENCODE annotation. Used for cross-referencing	
Parent	Protein ID, Gene ID, chromosome, start, end and strand. Detailed in section *'Parents of pseudogenes'*	
Sequence similarity	The percentage of pseudogene sequence preserved from parent	
Transcription	Evidence for pseudogene transcription and validation results. May be tagged as EST, BodyMap, RT-PCR or None, which represent pseudogene expression evidence from corresponding data sources. Multiple tags are separated by commas. Detailed in section *'Transcription of pseudogenes'*	1, transcription; 0, otherwise
DNaseI hypersensitivity	A categorical result indicating whether the pseudogene has easily accessible chromatin, predicted by a model integrating DNaseI hypersensitivity values within 4 kb genomic regions upstream and downstream of the 5′ end of pseudogenes. Detailed in section *'Chromatin signatures of pseudogenes'*	1, has Dnase hypersensitivity in upstream; 0, otherwise
Chromatin state	Whether a pseudogene maintains an active chromatin state, as predicted by a model using Segway segmentation. Detailed in section *'Chromatin signatures of pseudogenes'*	1, active chromatin; 0, otherwise
Active Pol2* binding	Whether Pol2 binds to the upstream region of a pseudegene. Detailed in section *'Upstream regulatory elements'*	1, active binding site; 0, otherwise
Active promoter region	Whether there are active promoter regions in the upstream of pseudogenes. Detailed in section *'Upstream regulatory elements'*	1, active binding site; 0, otherwise
Conservation	Conservation of pseudogenes is derived from the divergence between human, chimp and mouse DNA sequences. Detailed in section *'Evolutionary constraint on pseudogenes'*	1, conserved; 0, otherwise

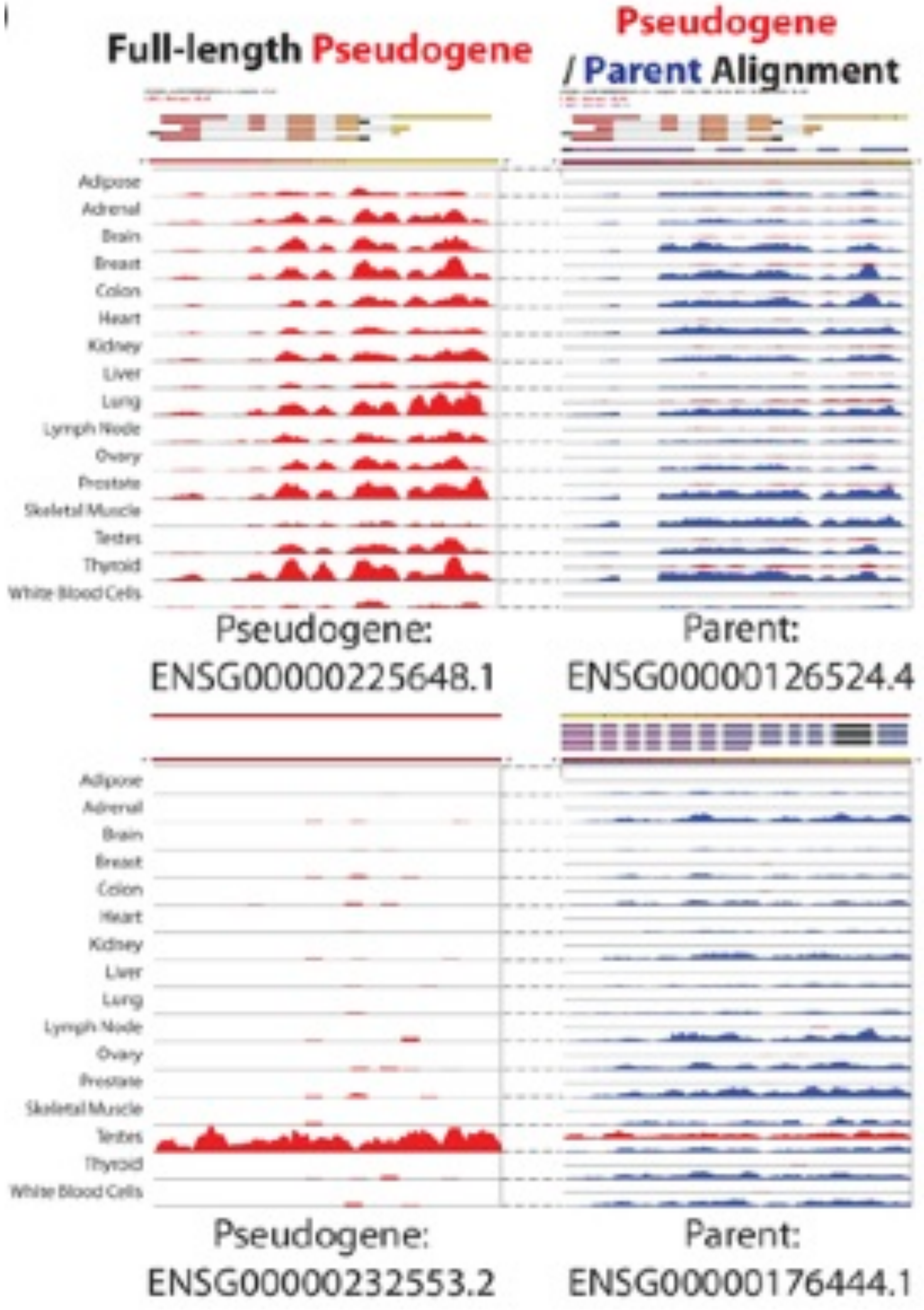
*Pol2, RNA polymerase II.

- Parent gene/ancestral gene = functional gene with greatest sequence similarity
- Ancestral gene can be identified for ca. 90% of pseduogenes
- 10% of pseudogenes are highly degraded and is derived from a parent gene with highly similar paralogs
Or parent gene contains a commonly found functional domain
-NOTE: most parental genes have only 1 pseudogene
-NOTE: some parental genes – mainly housekeeping genes - have MANY pseudogenes:
-Robosomal protein L21: 143 pseudogenes
-Gapdh: 68 pseudogenes

Features of transcribed pseudogenes

Problem: precise analysis of RNA-seq/array data: high sequence similarity pseudogene – parental gene

2012: ca 9000 pseudogenes: 873 are transcribed according to STRINGENT psiDR parameters (real number is higher)



tissue specific expression

transcription of pseudogene

transcription of pseudogene and parental gene

differential expression parental/pseudogene

Pseudogene expression levels are LOWER than coding gene expression

Pseudgenes are expressed in a different manner compared to parental mRNAs (different tissues)

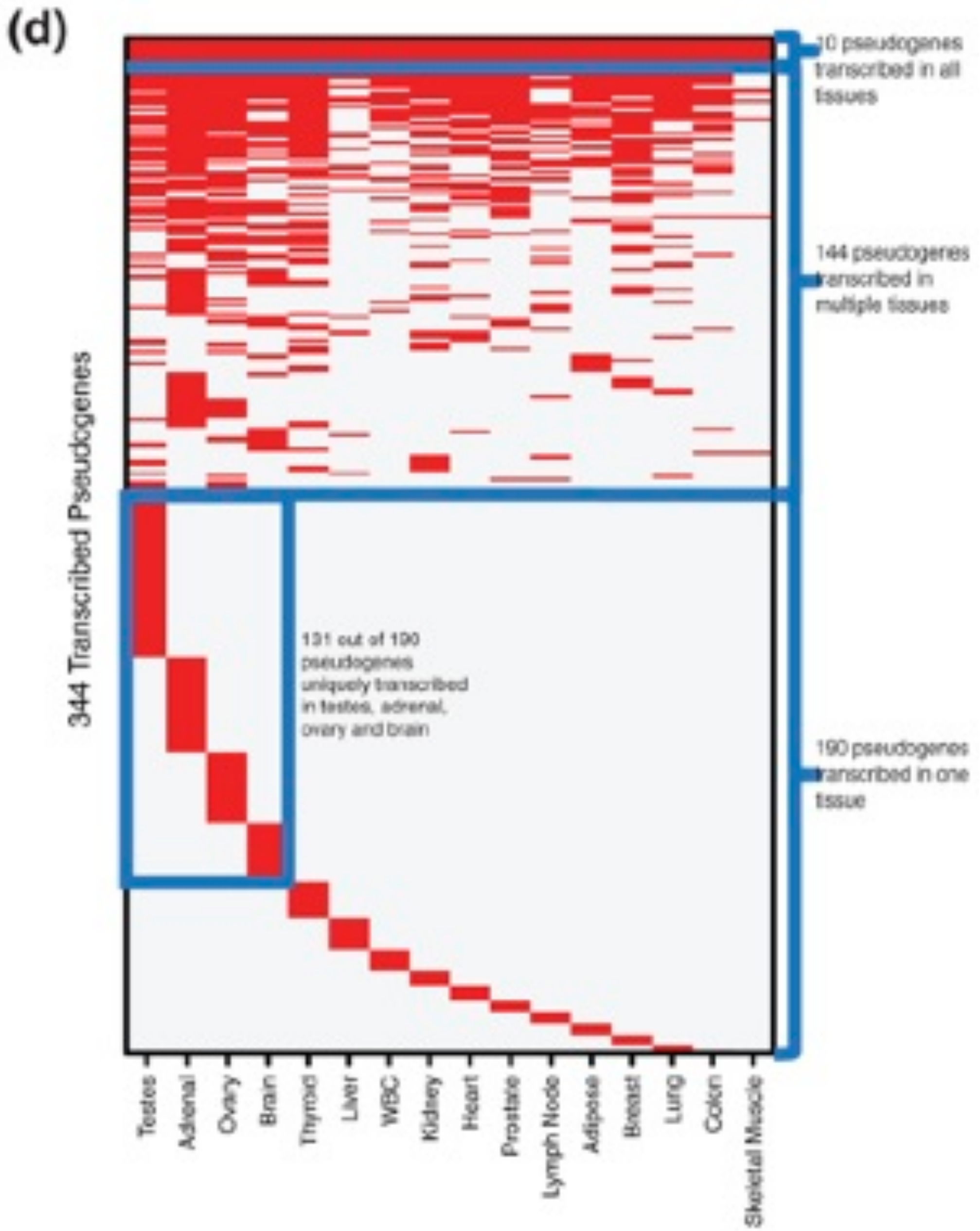
tissue specific expression

transcription of pseudogene

transcription of pseudogene and parental gene

differential expression parental/pseudogene

The majority of pseudogenes show tissue specific expression



Categories:

-Expressed in all tissues
(10 out of 344 tested pseudogenes)

-144/344 pseudogenes expressed in more then 1 tissue

-190/344 pseudogenes exclusively expressed in 1 tissue

duplicated/processed pseudogenes have specific regulatory elements!!

Evolutionary constraint on pseudogenes

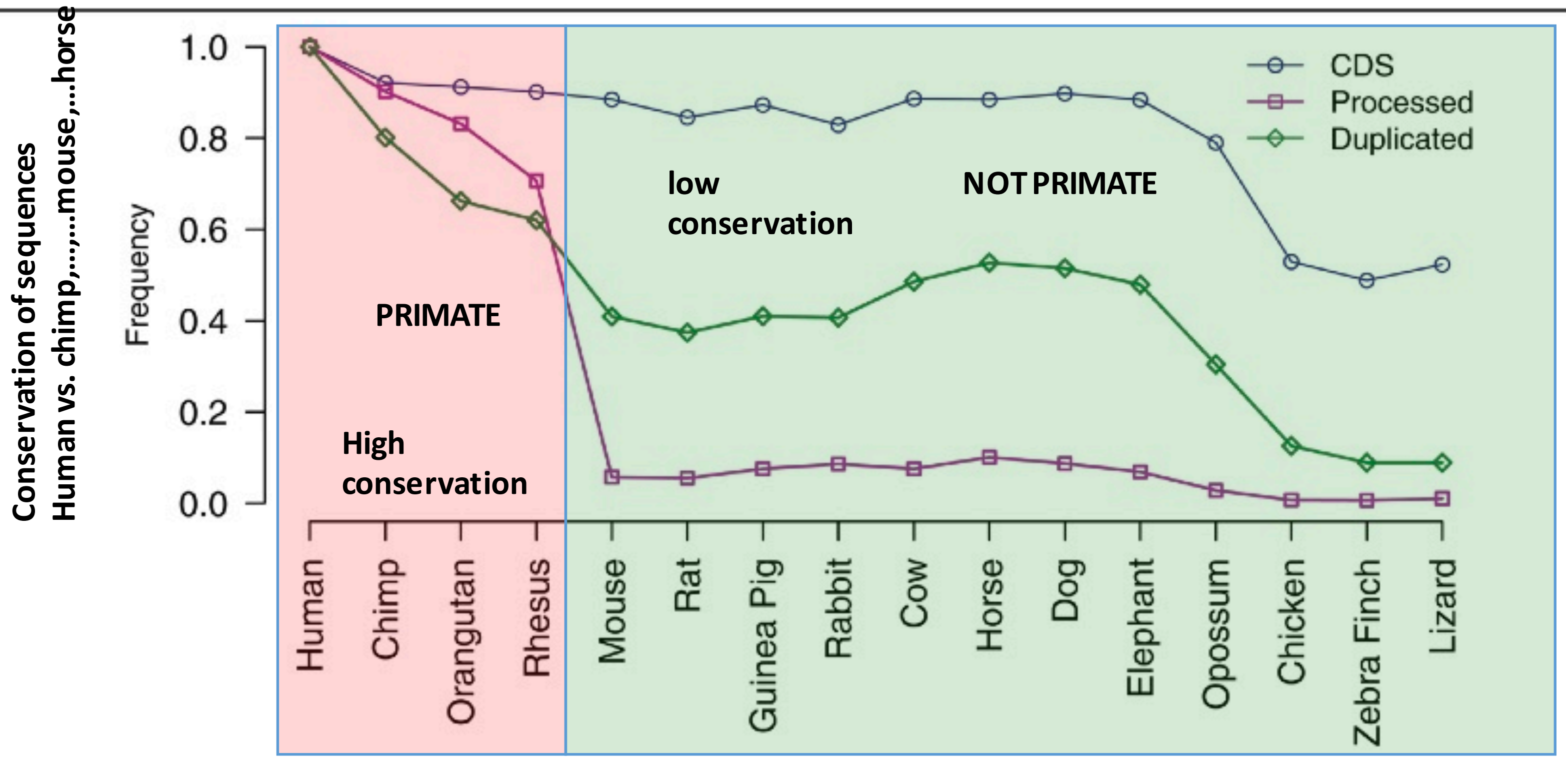
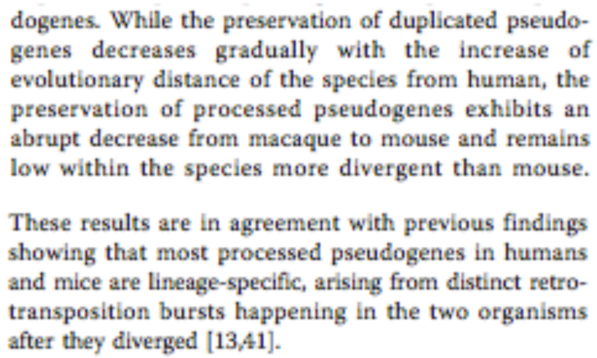


Figure 6 Preservation of human coding sequences, processed pseudogenes and duplicated pseudogenes. Sequences orthologous to human genomic regions from different species were studied. The sequence preservation rate was calculated as the percentage of sequences aligned to human sequence from each species. The calculation was based on a MultiZ multiple genome sequence alignment.



dogenes. While the preservation of duplicated pseudogenes decreases gradually with the increase of evolutionary distance of the species from human, the preservation of processed pseudogenes exhibits an abrupt decrease from macaque to mouse and remains low within the species more divergent than mouse.

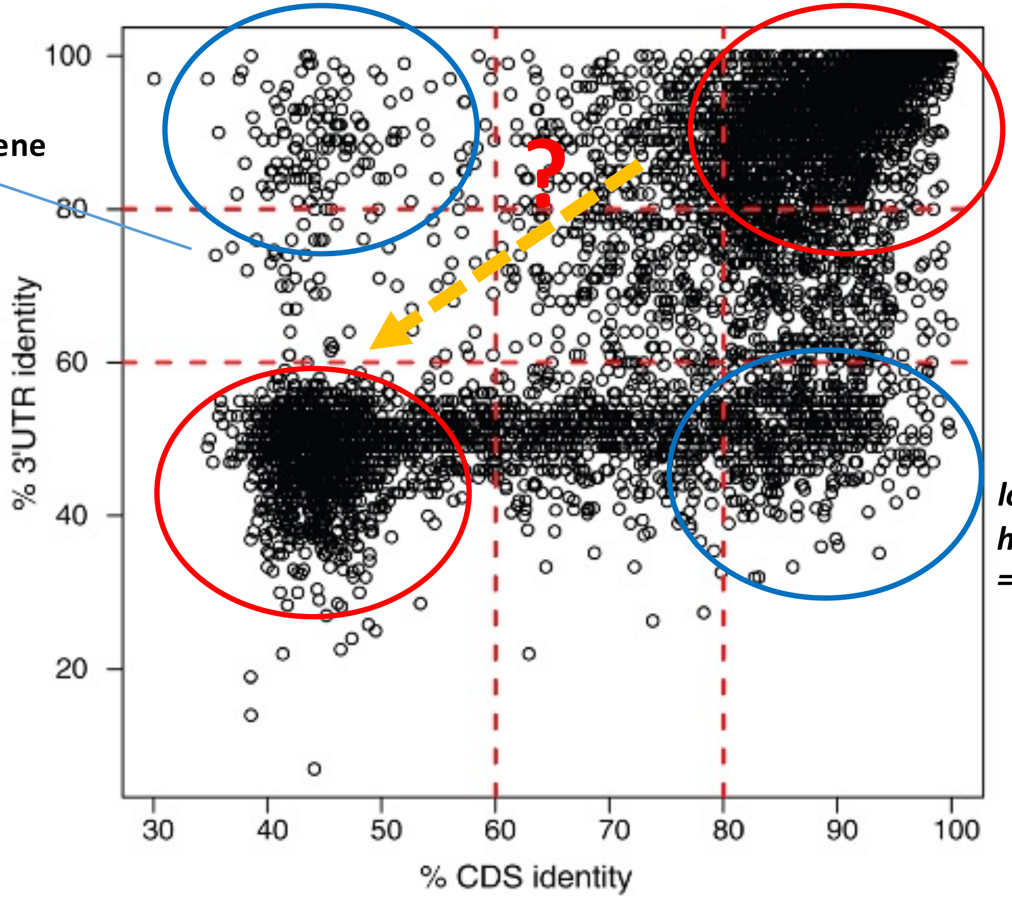
These results are in agreement with previous findings showing that most processed pseudogenes in humans and mice are lineage-specific, arising from distinct retrotransposition bursts happening in the two organisms after they diverged [13,41].

Selective constraint in pseudogen lncRNAs

high 3'UTR conservation
low CDS conservation
= high selective constraint

Sequence identity between parental and pseudogenes with focus on coding sequence (CDS) and 3'UTR

one pseudogene



High CDS conservation

High 3'UTR conservation
=high selective constraint

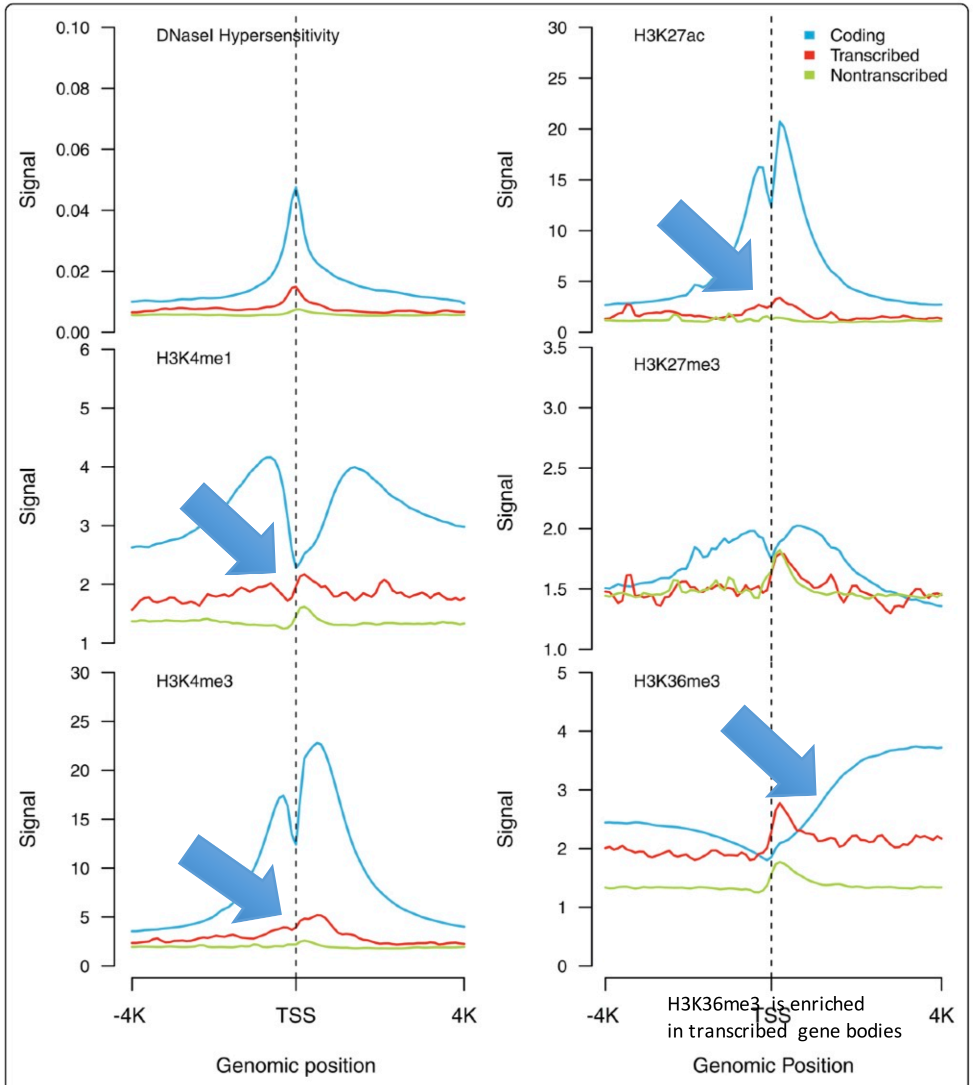
Most pseudogenes have Similar sequence identity between
CDS and UTR (high: >80%; low: <60%)

low 3'UTR conservation
high CDS conservation
= high selective constraint

Mutations were rejected by natural selection
non-randomly. Certain regions in the sequence may be
under higher evolutionary constraint than the others.

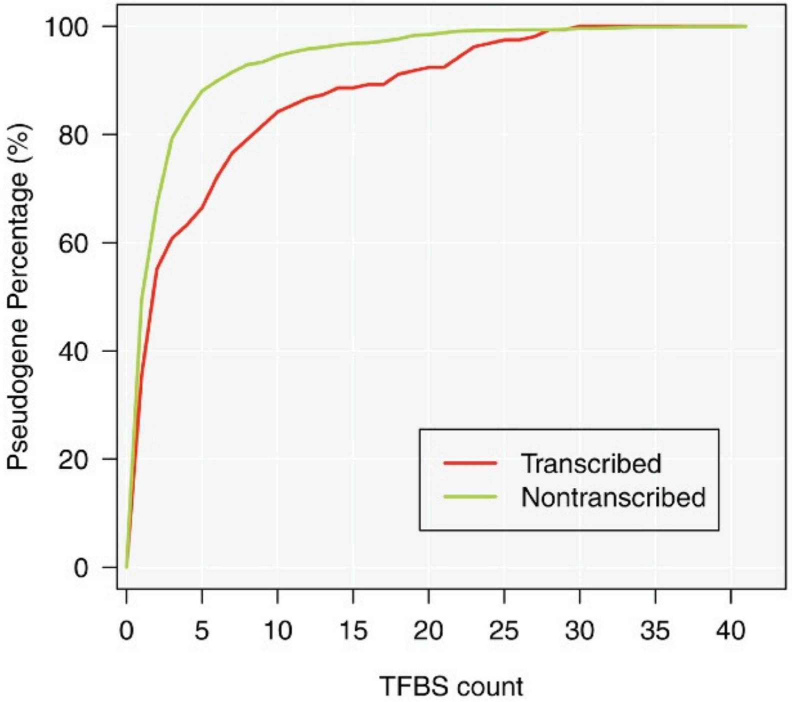
Inconsistency implies that mutations were rejected by natural selection non-randomly. Certain regions in the sequence may be under higher evolutionary constraint than the others. We identified 998 pseudogenes showing a high (>80%) sequence identity to parent CDS and simultaneously poor (<60%) sequence identity to the 3' UTR, and 36 pseudogenes with high (>80%) sequence identity to the parent 3' UTR and small (<60%) sequence identity to CDS.

Chromatin at transcriptional start sited of transcribed pseudogenes is similar to coding genes



H3K36me3 is enriched in transcribed gene bodies

Figure 8 Chromatin signatures: DNaseI hypersensitivity and histone modification. Average chromatin accessibility profiles and various histone modifications surrounding the TSS for coding genes, transcribed pseudogenes, and non-transcribed pseudogenes. The coding gene histone modification profiles around the TSS follow known patterns - for example, enrichment of H3K4me1 around 1 kb upstream of the TSS and the H3K4me3 peaks close to the TSS [63]. Transcribed pseudogenes also show stronger H3K4 signals than non-transcribed pseudogenes. H3K27me3, a marker commonly associated with gene repression [64], showed depletion around the TSS for the coding gene and a distinctive peak in the same region for the pseudogenes. H3K36me3 also shows a similar pattern as H3K27me3 at TSSs, which may relate to nucleosome depletion.

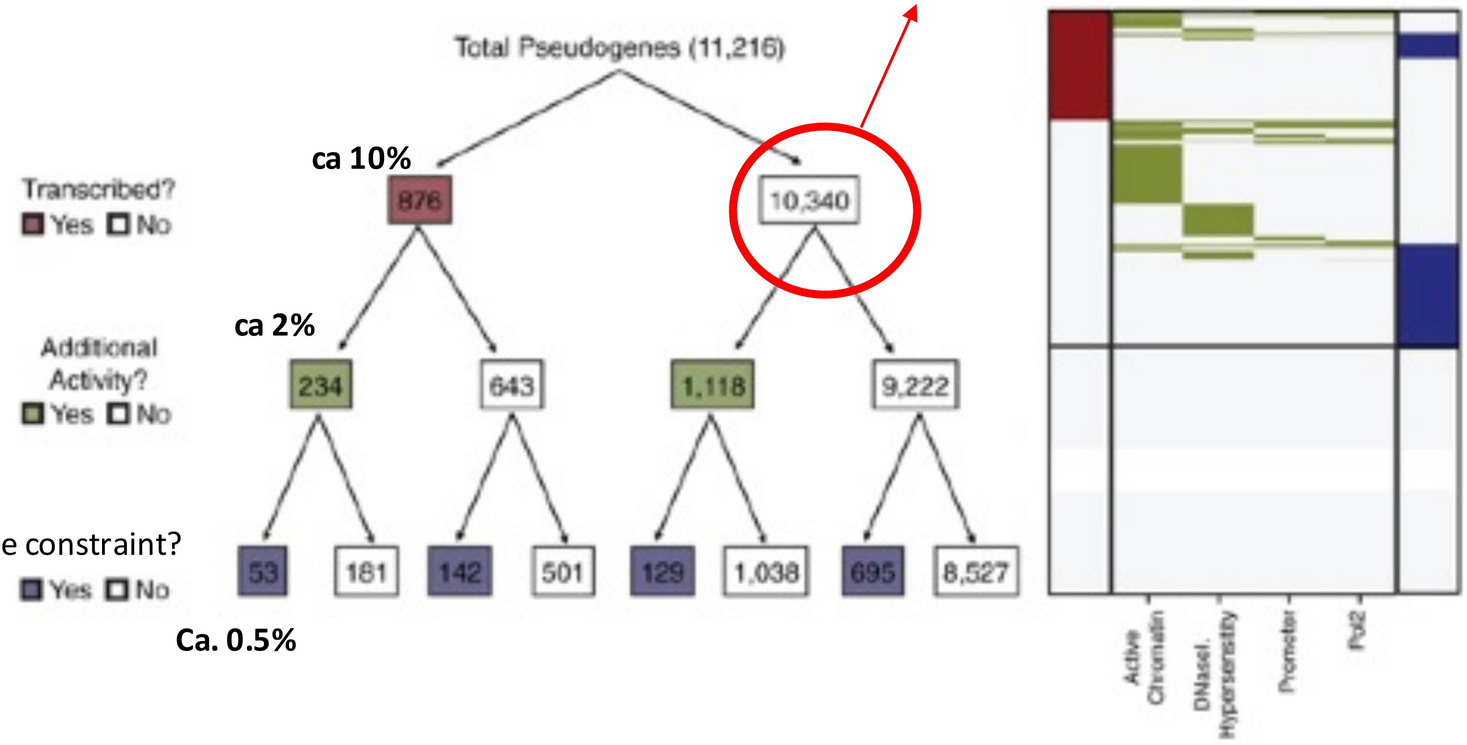


Frequency of transcription factor binding sites enriched in transcribed Pseudogenes vs non-transcribed pseudogenes

Transcribed pseudogenes
resemble coding genes; however:
Peaks are not as clear defined =
average chromatin marks are less concentrated:
Reason:
→lower expression
→ expressed pseudogenes do not show marks in an uniform manner

Pseudogenes are a diversified group of genetic elements

K562 cancer cell line



Real number is probably lower!!!

ca 10%

ca 2%

Selective constraint?

Ca. 0.5%

→ **few pseudogenes show consistently active signals across all biological features that describe gene activity**

→ **many pseudogenes show little or no activity**

Figure 12 Summary of pseudogene annotation and case studies. **(a)** A heatmap showing the annotation for transcribed pseudogenes including active chromatin segmentation, DNaseI hypersensitivity, active promoter, active Pol2, and conserved sequences. Raw data were from the K562 cell line. **(b)** A transcribed duplicated pseudogene (Ensembl gene ID: ENST00000434500.1; genomic location, chr7: 65216129-65228323)

Pseudogenes are a diversified group of genetic elements

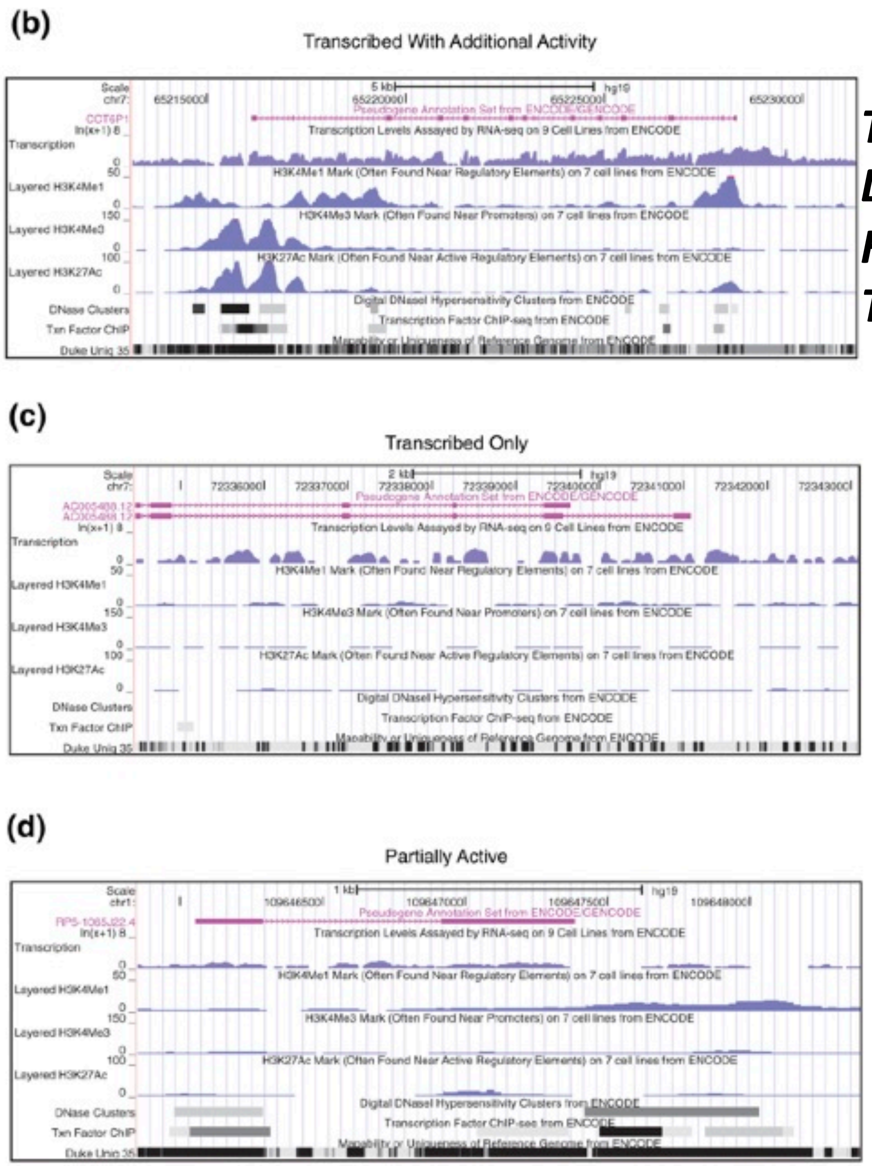


Figure 12 Summary of pseudogene annotation and case studies. **(a)** A heatmap showing the annotation for transcribed pseudogenes including active chromatin segmentation, DNaseI hypersensitivity, active promoter, active Pol2, and conserved sequences. Raw data were from the K562 cell line. **(b)** A transcribed duplicated pseudogene (Ensembl gene ID: ENST00000434500.1; genomic location, chr7: 65216129-65228323) showing consistent active chromatin accessibility, histone marks, and TFBSs in its upstream sequences. **(c)** A transcribed processed pseudogene (Ensembl gene ID: ENST00000355920.3; genomic location, chr7: 72333321-72339656) with no active chromatin features or conserved sequences. **(d)** A non-transcribed duplicated pseudogene showing partial activity patterns (Ensembl gene ID: ENST00000429752.2; genomic location, chr1: 109646053-109647388). **(e)** Examples of partially active pseudogenes. E1 and E2 are examples of duplicated pseudogenes. E1 shows *UGT1A2P*

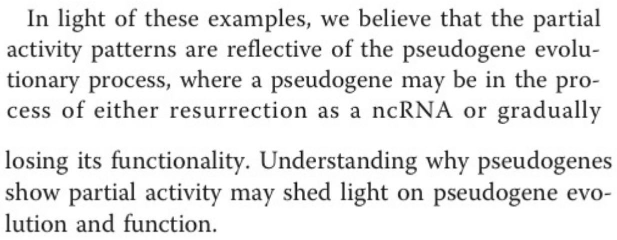
Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor

Pseudogene
under selective constraint
→ maintained

Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor

Pseudogenes
under low selective constraints
→This stage also involves
acquisition of new splice sites –
resembles a stage of testing new
mutations for evolutionary
advantage. Result:
A. dying pseudogene or
B. acquisition of critical feature
leading to the resurrection to
become a functional pseudogene

Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor



In light of these examples, we believe that the partial activity patterns are reflective of the pseudogene evolutionary process, where a pseudogene may be in the process of either resurrection as a ncRNA or gradually losing its functionality. Understanding why pseudogenes show partial activity may shed light on pseudogene evolution and function.