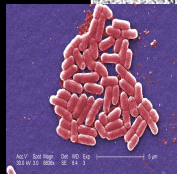
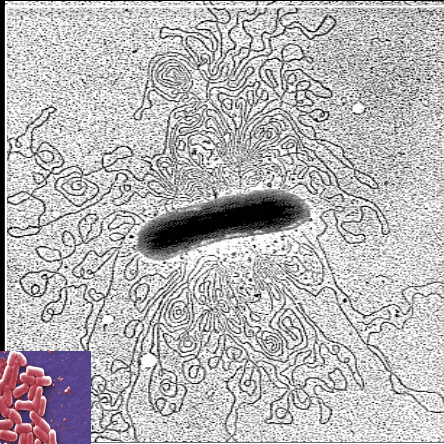


PSEUDOGENE DERIVED lncRNAs

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

E.coli



C. elegans

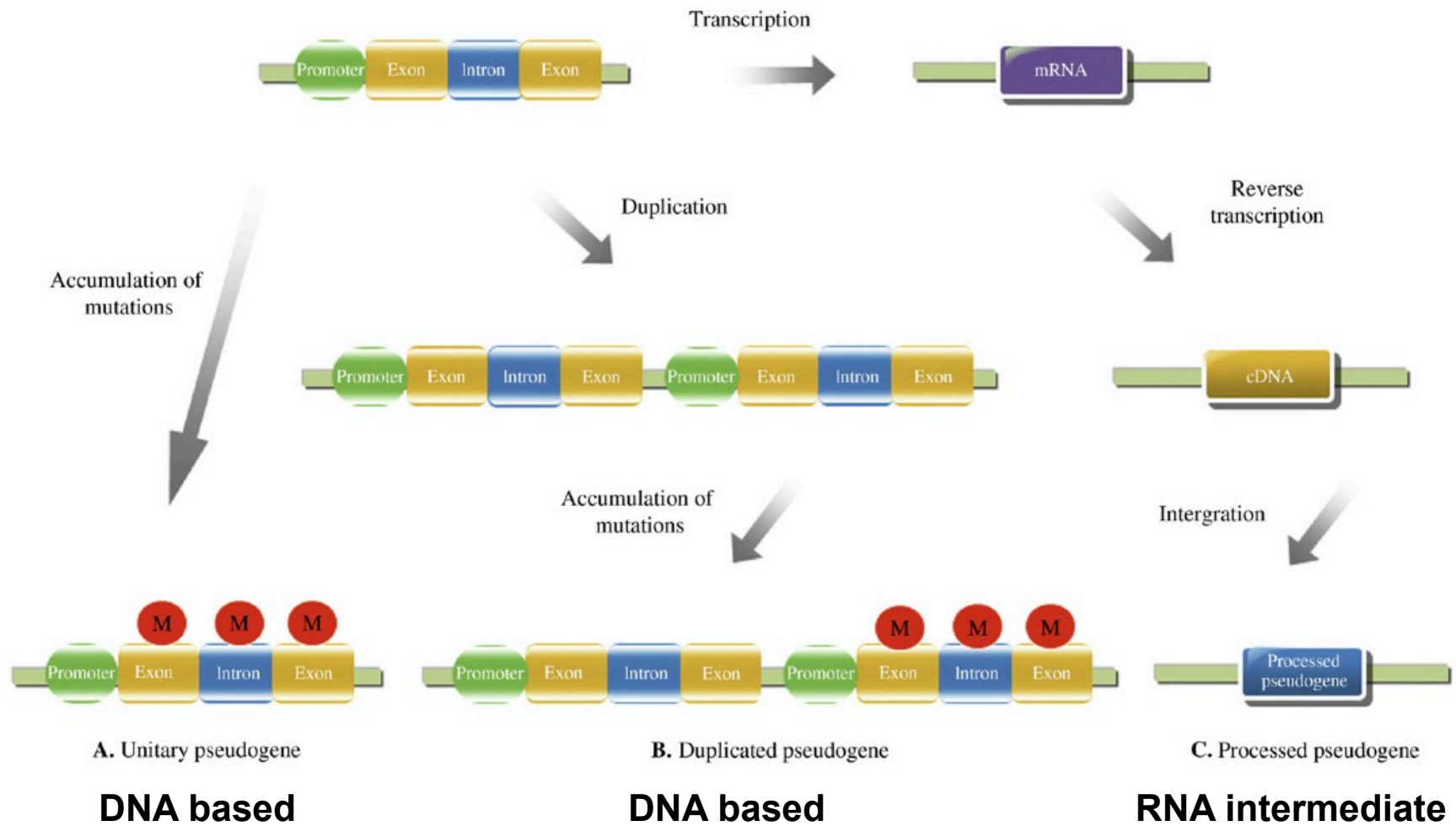


H. sapiens

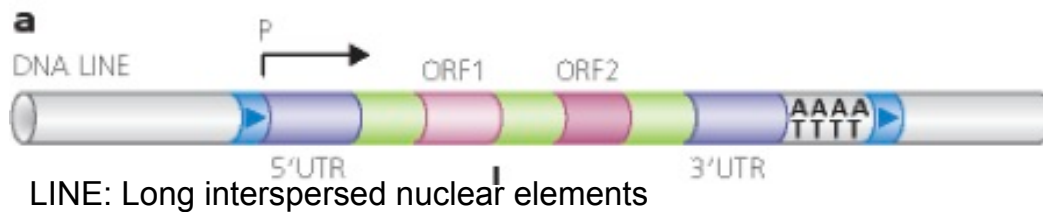


	Genome	5×10^6 bp	1×10^8 bp	3×10^9 bp
Chromosomes		1	6	23
Coding genes		6692	20541	21995
ncDNA		5%	60%	98%
non-coding RNA genes		15	23136	ca. 40000
miRNAs		0	224	4274
pseudogenes		21	1522	10616

Protein coding genes give rise to pseudogenes



Transposition of Retrotransposons

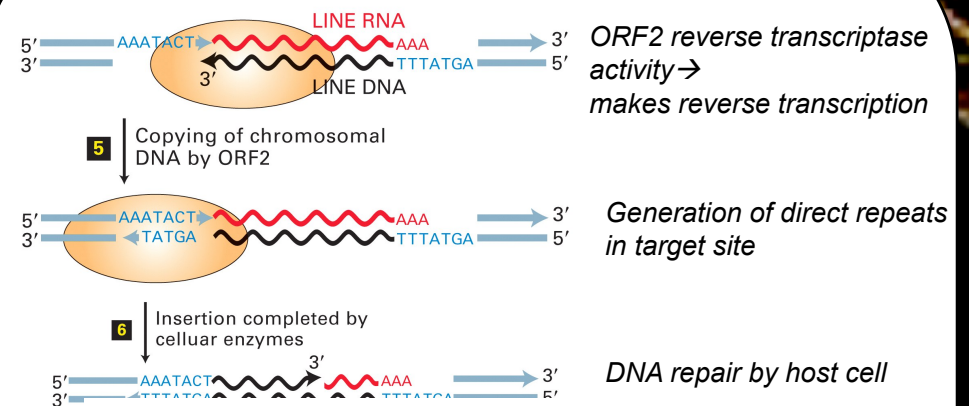
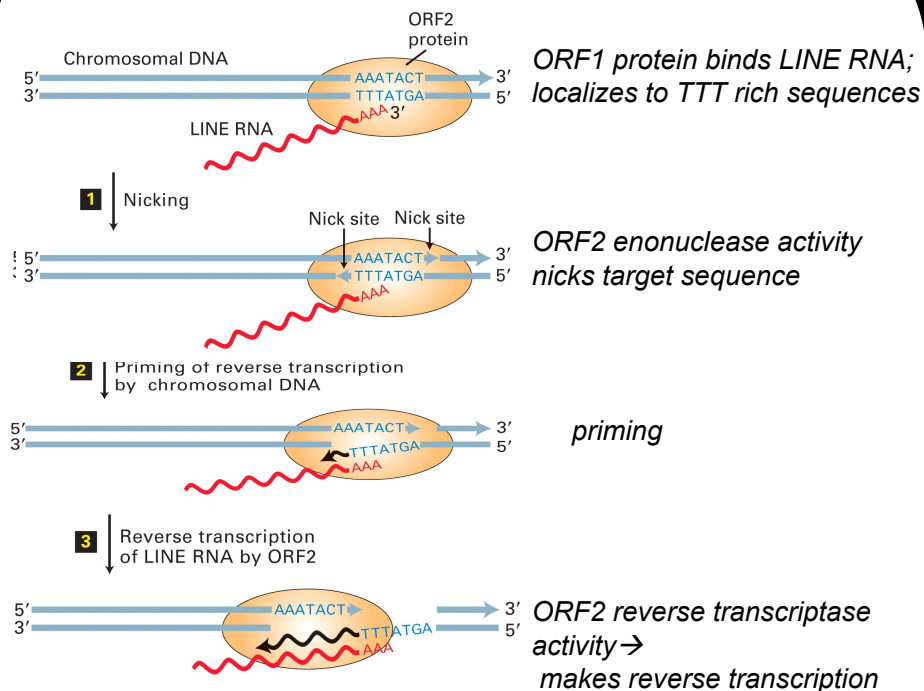


LINE elements (L1,L2,L3)

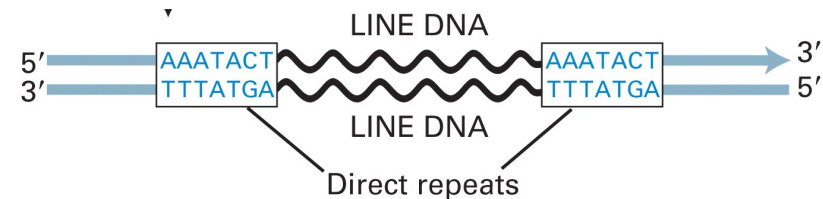
(21% of genome; 800.000 copies)

ORF1: RNA binding protein

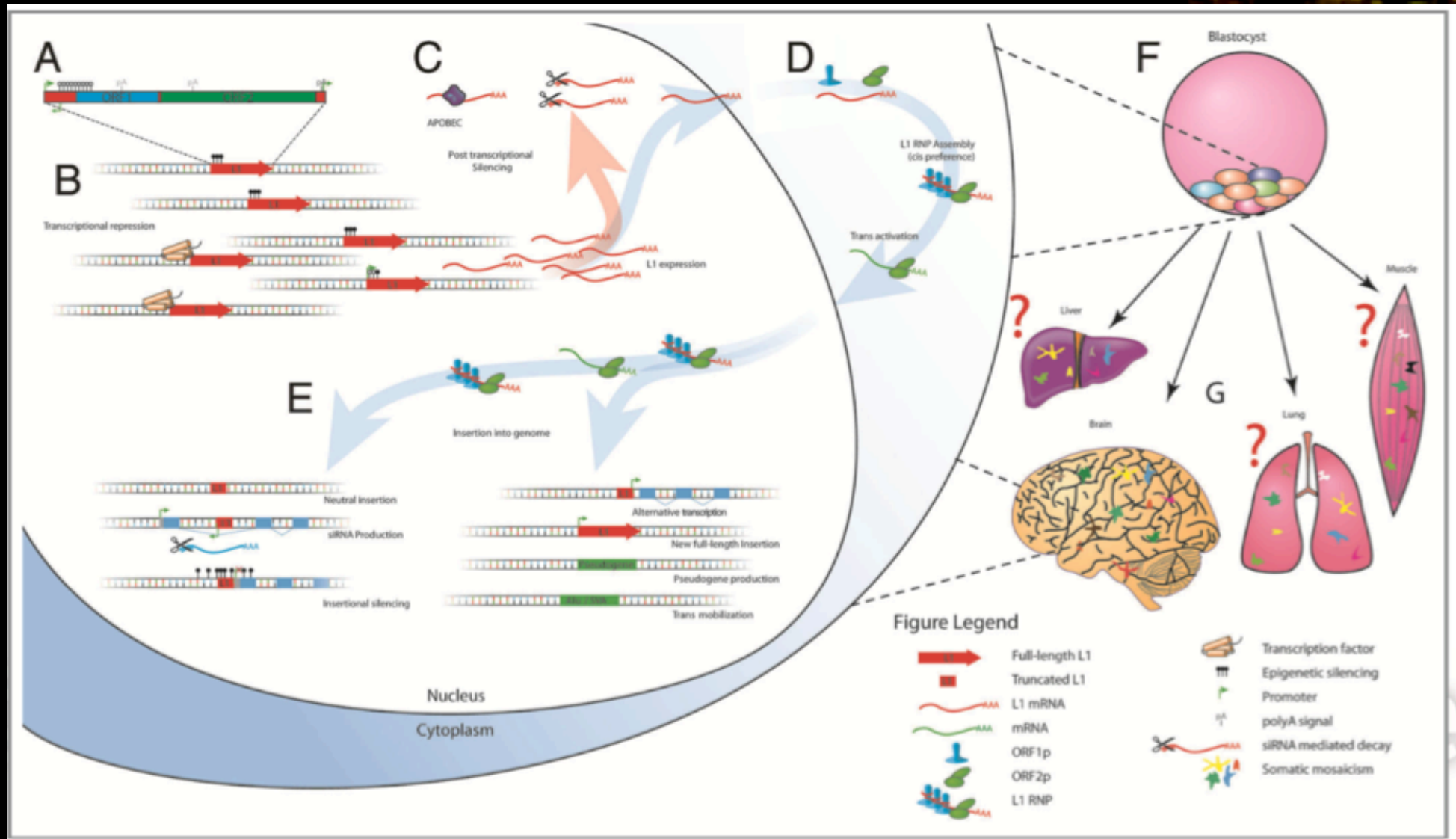
ORF2: Endonuclease, Reverse transcriptase



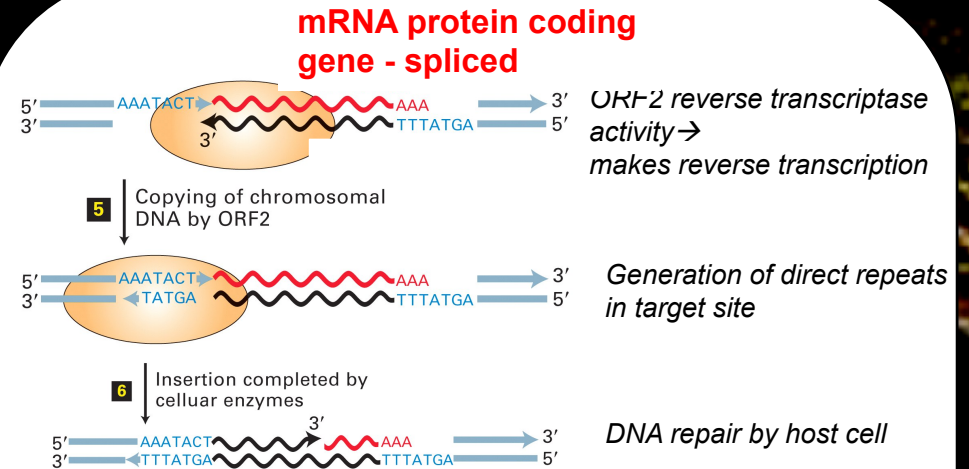
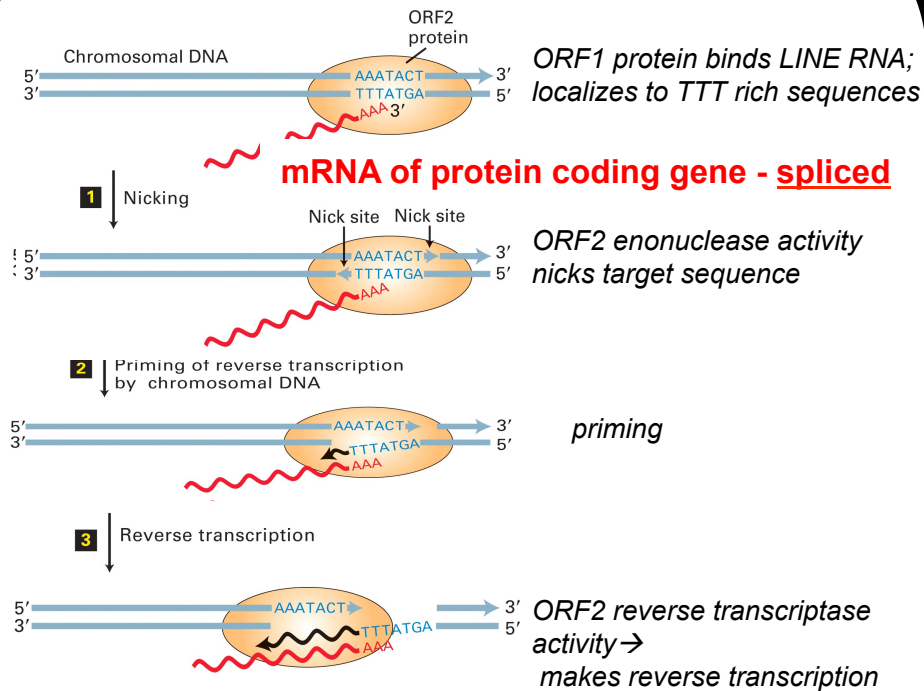
FINAL PRODUCT



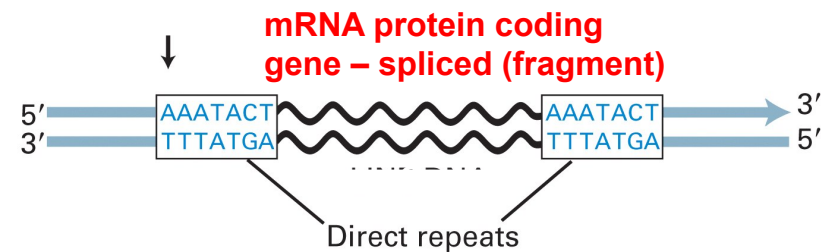
Retrotransposons can change genetic context



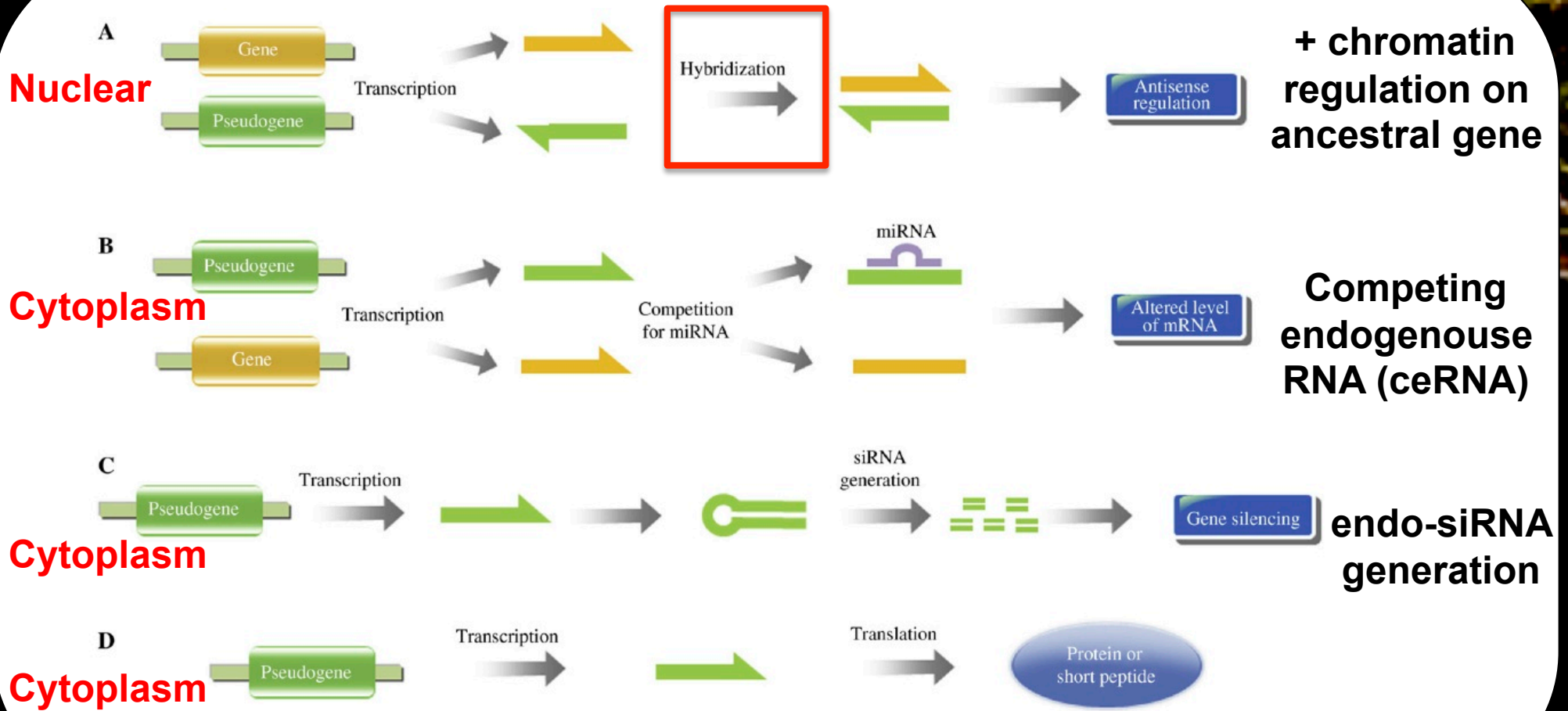
Retro-transposition machinery hijacks endogenous mRNAs



FINAL PRODUCT: PROCESSED PSEUDOGENE



Pseudogene derived RNAs can acquire new functions



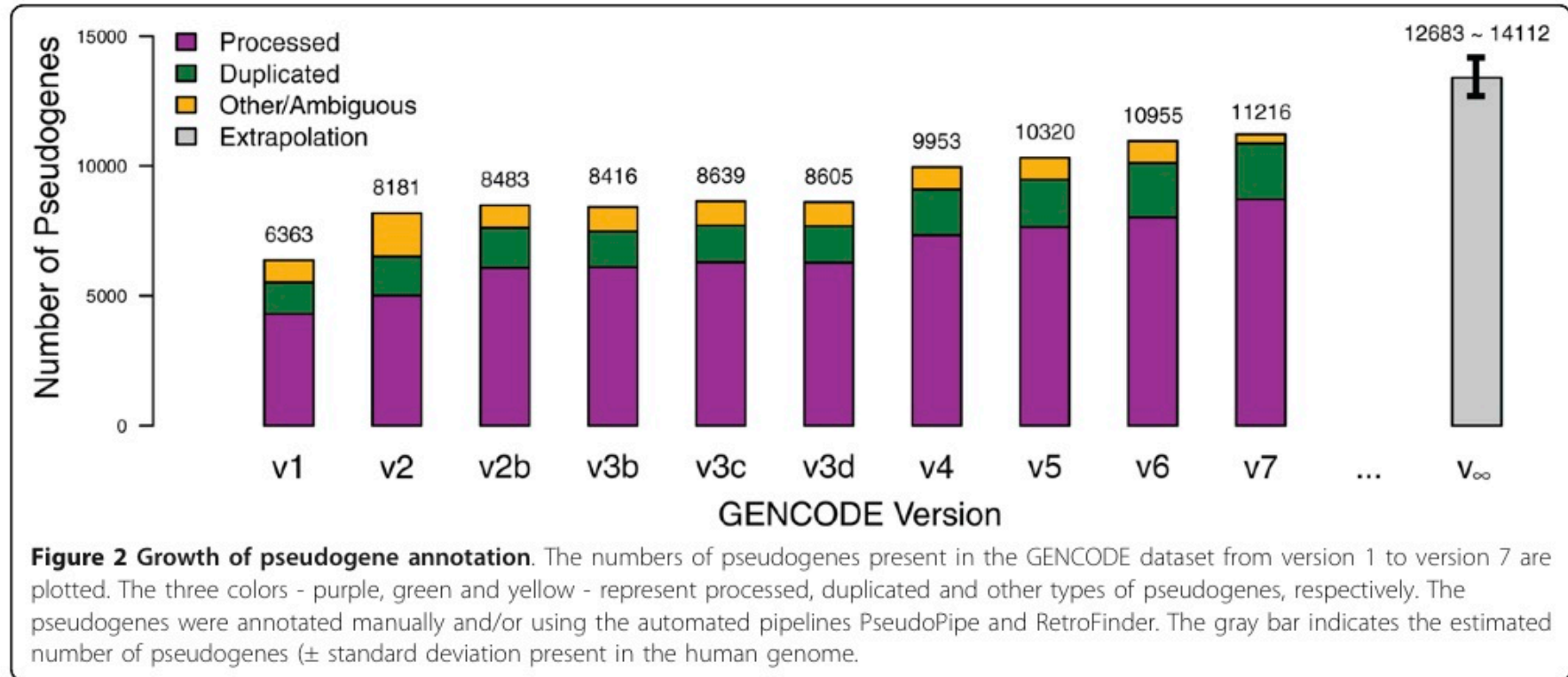
PSEUDOGENE BIOTYPES

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

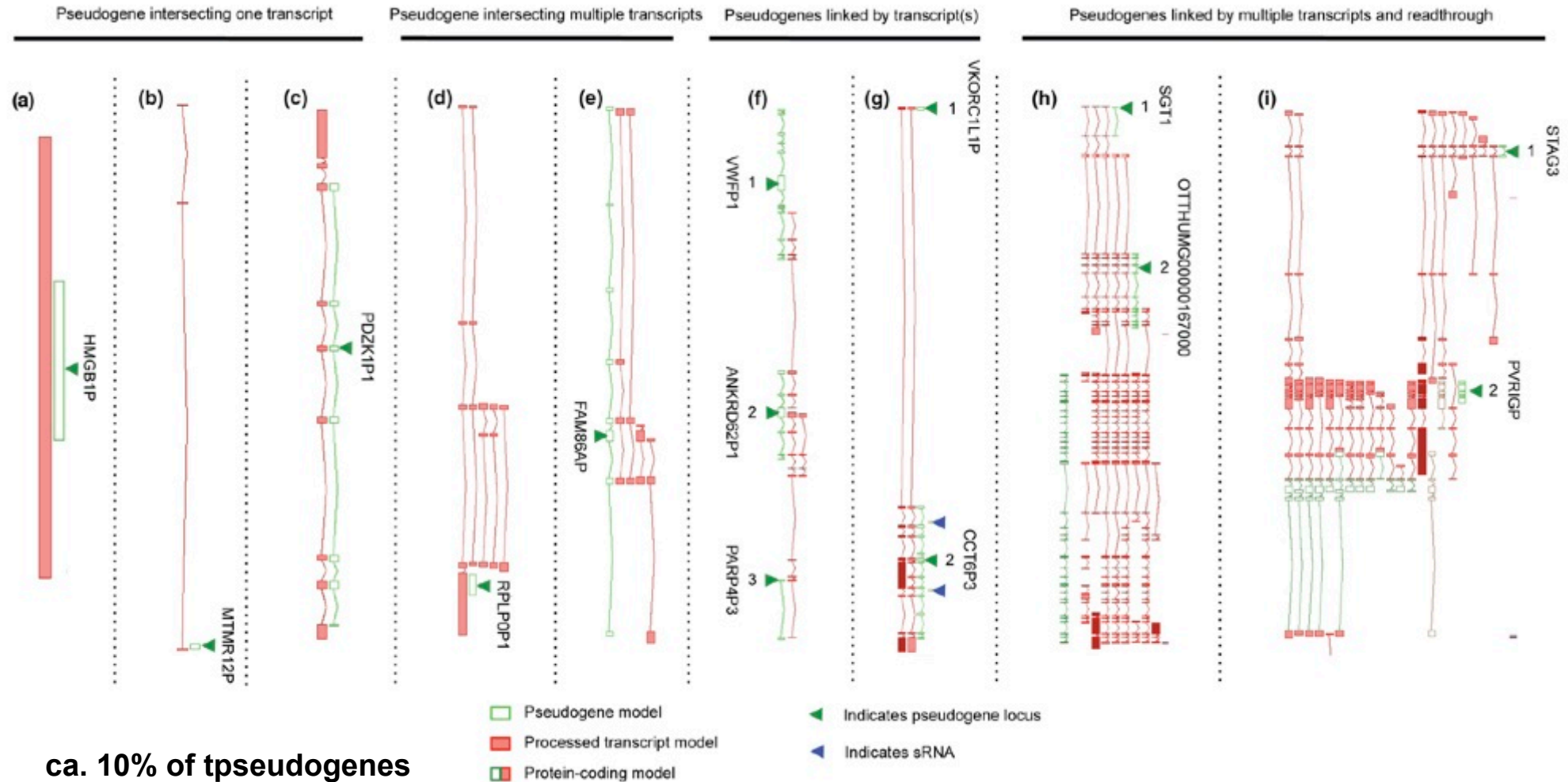
PSEUDOGENE BIOTYPES



*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

PSEUDOGENE LOCI (duplicated/processed) CAN BE USED BY OTHER FUNCTIONAL LOCI



ca. 10% of tpseudogenes

1. Pseudogene sequence creates a new alternatively spliced internal exon in the protein coding gene
2. Pseudogene sequence contributes to the 5' exon in the protein coding gene
3. Pseudogene sequence contributes to the 3' terminal exon of the protein-coding gene

→ Pseudogenes contribute to the evolution of protein coding genes

Figure 3 Complexity of transcribed pseudogenes. Screenshots of pseudogene annotation are taken from the Zmap annotation interface. The pseudogenes are represented as open green boxes and indicated by dark green arrowheads, exons of associated transcript models are represented as filled red boxes and connections are shown by red lines. The coding exons of protein-coding models are represented by dark green boxes and UTR exons as filled red boxes; protein-coding models are also indicated by red arrowheads. **(a-c)** Single pseudogene models intersecting with single transcript models. (a) The processed pseudogene High mobility group box 1 pseudogene (*HMGB1P*; HAVANA gene ID: OTTHUMG00000172132) and its associated unspliced (that is, single exon) transcript. (b) The processed pseudogene Myotubularin related protein 12 pseudogene (*MTMR12P*; HAVANA gene ID: OTTHUMG00000167532) and a spliced transcript model with three exons. (c) A duplicated pseudogene PDZ domain containing 1 pseudogene 1 (*PDZK1P1*; HAVANA gene ID: OTTHUMG00000013746) and a spliced transcript model with nine exons. **(d,e)** Single pseudogene models intersecting with multiple transcripts. (d) The processed pseudogene Ribosomal protein, large, P0 pseudogene 1 (*RPLPOP1*; HAVANA gene ID: OTTHUMG00000158396) and five spliced transcripts. (e) The duplicated pseudogene Family with sequence similarity 86, member A pseudogene (*FAM86AP*; HAVANA gene ID: OTTHUMG00000159782) and four spliced transcripts. **(f,g)** Groups of multiple pseudogenes that are connected by overlapping transcripts. (f) Three pseudogenes with single connecting transcripts: 1 is the duplicated pseudogene von Willebrand factor pseudogene 1 (*VWFP1*; HAVANA gene ID: OTTHUMG00000143725); 2 is a duplicated pseudogene ankyrin repeat domain 62 pseudogene 1 (*ANKRD62P1*; HAVANA gene ID: OTTHUMG00000149993); 3 is the duplicated pseudogene poly (ADP-ribose) polymerase family, member 4 pseudogene 3 (*PARP4P3*; HAVANA gene ID: OTTHUMG00000142831). Pseudogene 1 and 2 are connected by a seven exon transcript, pseudogenes 2 and 3 are connected by a nine exon transcript and there is a third transcript that shares two of its four exons with pseudogene 2. (g) Two pseudogenes with multiple connecting transcripts: 1 is the processed pseudogene vitamin K epoxide reductase complex, subunit 1-like 1 pseudogene (*VKORC1L1P*; HAVANA gene ID: OTTHUMG00000156633); 2 is the duplicated pseudogene chaperonin containing TCP1, subunit 6 (zeta) pseudogene 3 (*CCT6P3*; HAVANA gene ID: OTTHUMG00000156630). The two pseudogenes are connected by two transcripts that initiate at the upstream pseudogene and utilize a splice donor site within the single exon, which is also a splice donor site in the pseudogene's parent locus. Interestingly, the downstream locus hosts two small nucleolar RNAs (snoRNAs) that are present in the parent locus and another paralog. **(h)** A very complex case where multiple pseudogenes, connected by multiple transcripts, read through into an adjacent protein-coding locus: 1 is the duplicated pseudogene suppressor of G2 allele of SKP1 (*S. cerevisiae*) pseudogene (*SGT1P*; HAVANA gene ID: OTTHUMG00000020323); 2 is a novel duplicated pseudogene (OTTHUMG00000167000); and the protein-coding gene is *C9orf174*, chromosome 9 open reading frame 174 (OTTHUMG00000167001). **(i)** A similarly complex case where multiple pseudogenes, connected by multiple transcripts, read through into an adjacent protein-coding locus: 1 is a duplicated pseudogene stromal antigen 3 pseudogene (*STAGP3*; HAVANA gene ID: OTTHUMG00000156884); 2 is a duplicated pseudogene poliovirus receptor related immunoglobulin domain containing pseudogene (*PVRIGP*; HAVANA gene ID: OTTHUMG00000156886); and the protein-coding gene is *PILRB*, paired immunoglobulin-like type 2 receptor beta (OTTHUMG00000155363). sRNA, small RNA.

GENOMICS STRATEGIES TO IDENTIFY AND CLASSIFY PSEUDOGENES

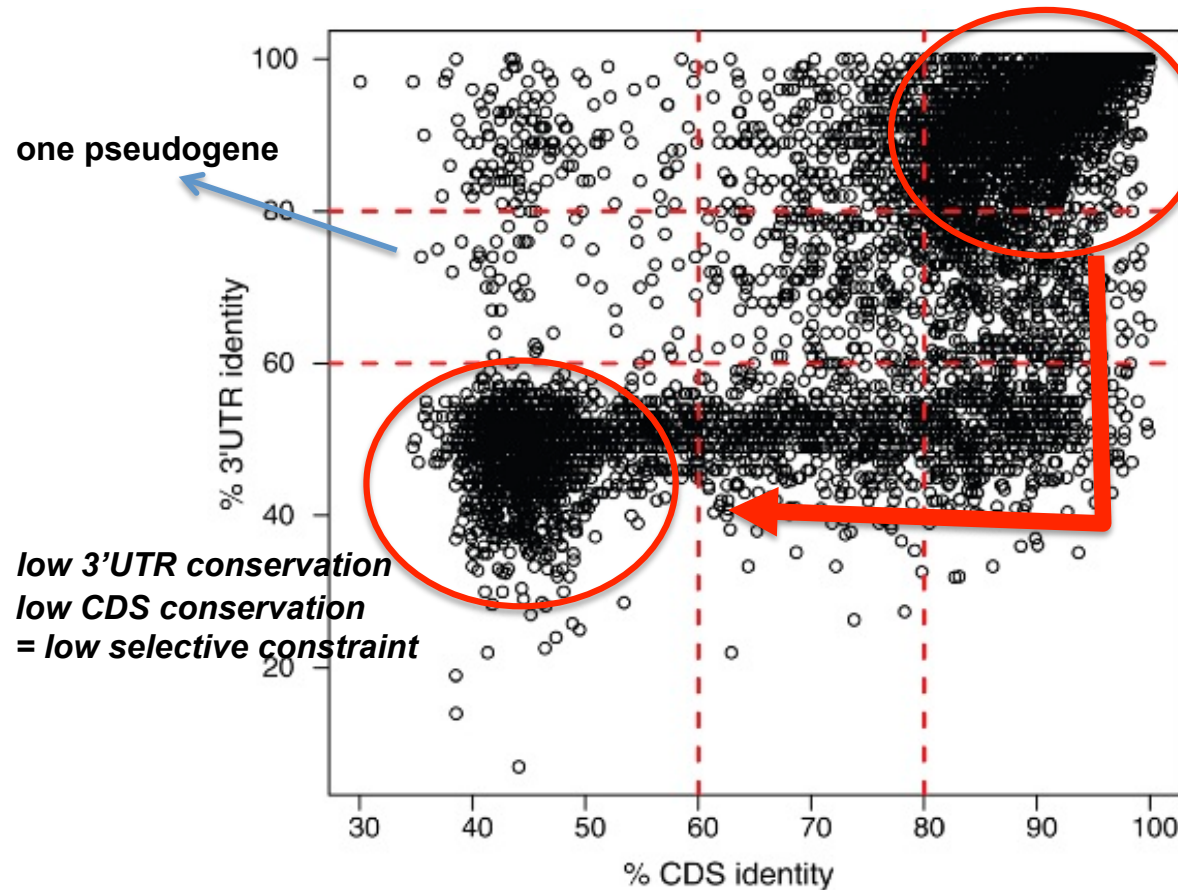
Table 3 Fields for pseudogene features in the psiDR annotation file **Pseudogene decoration resource**

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 ' <i>Parents of pseudogenes</i> '	
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 ' <i>Transcription of pseudogenes</i> '	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 ' <i>Chromatin signatures of pseudogenes</i> '	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 ' <i>Chromatin signatures of pseudogenes</i> '	1, active chromatin; 0, otherwise
Active Pol2* binding	Whether Pol2 binds to the upstream region of a pseudogene. Detailed in section ' <i>Upstream regulatory elements</i> '	1, active binding site; 0, otherwise
Active promoter region	Whether there are active promoter regions in the upstream of pseudogenes. Detailed in section ' <i>Upstream regulatory elements</i> '	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 ' <i>Evolutionary constraint on pseudogenes</i> '	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 pseudogenes**
- **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**

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



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

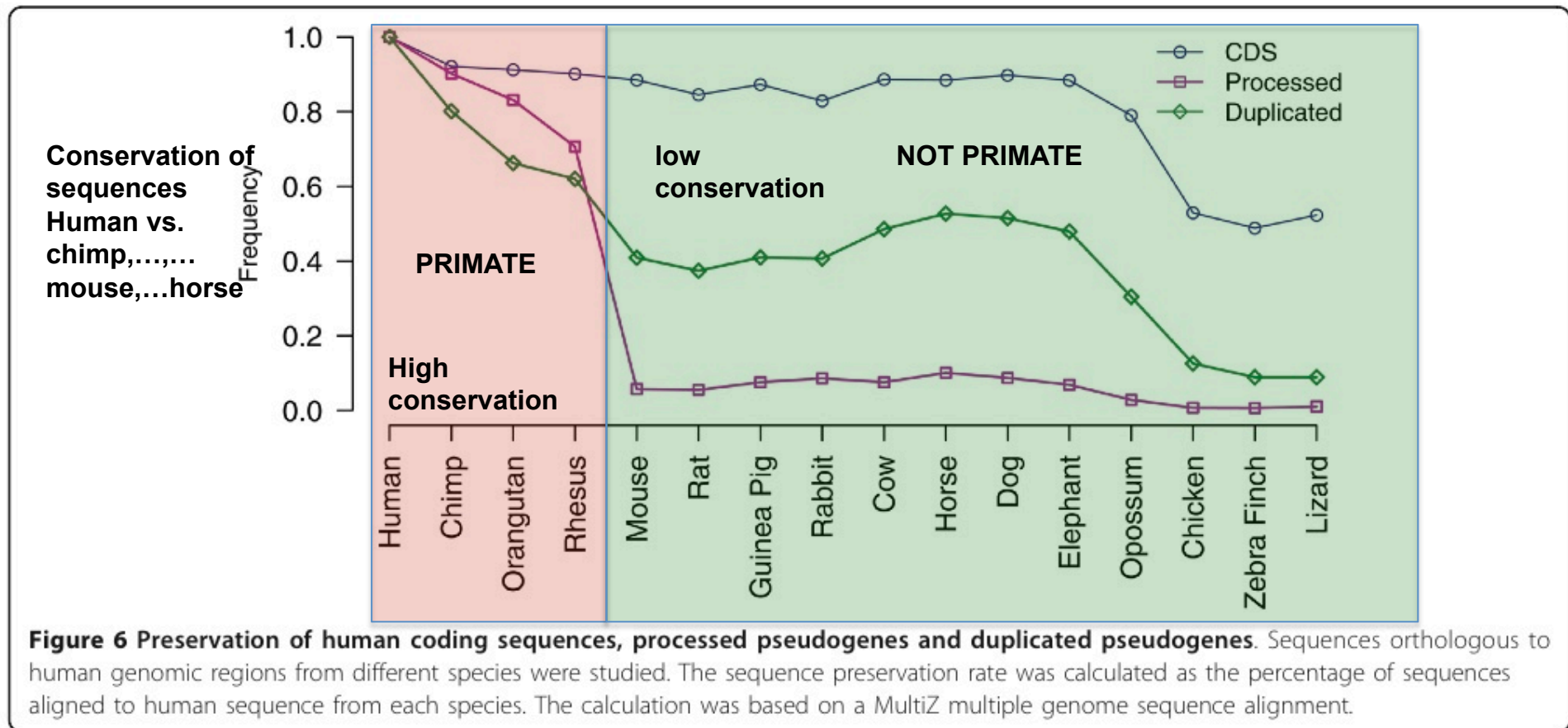
Most pseudogenes have similar sequence identity between CDS and UTR (high: >80%; low: <60%)

small group of pseudogenes
Have different selective constraint (CDS high/UTR low or CDS low/UTR high,...)
→ Mutation in CDS/UTR was rejected during evolution (= exists a selective constraint!!)
→ Higher selective pressure to stay with conserved UTR or CDS
→ mutations were rejected non-randomly

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

one pseudogene

Evolutionary constraint on pseudogenes



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].

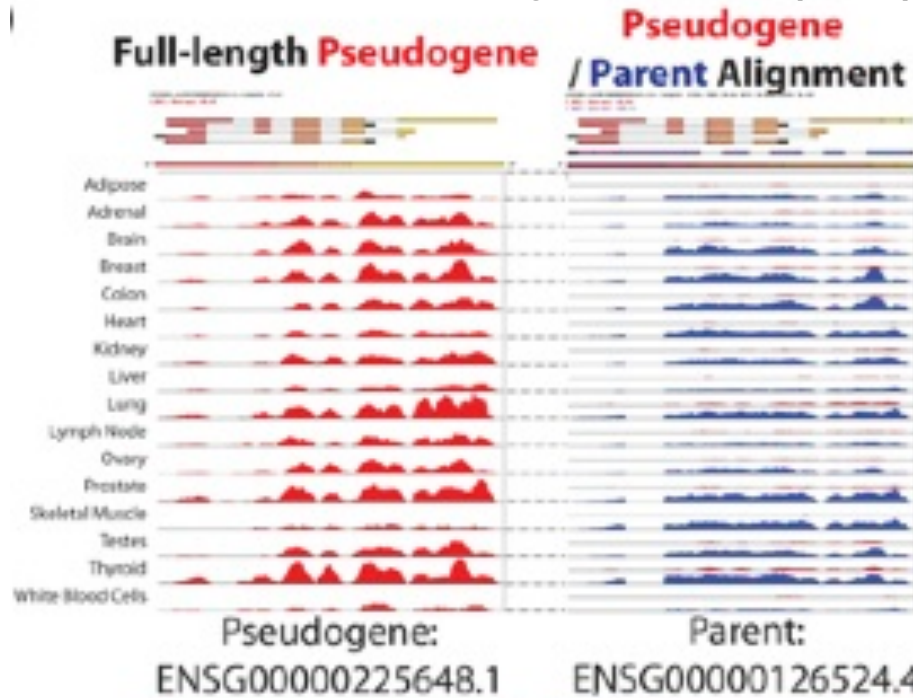
Features of transcribed pseudogenes

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



differential expression
parental/pseudogene

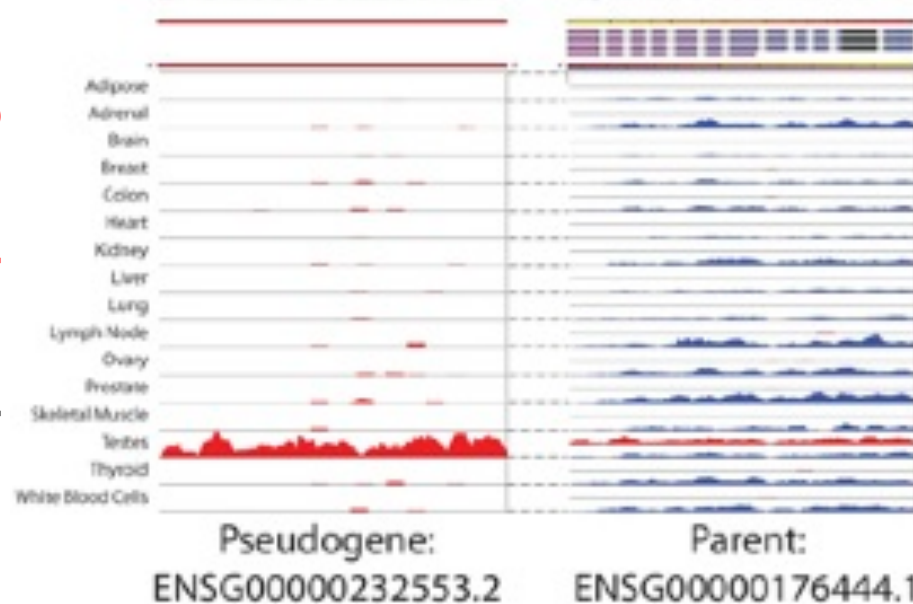
transcription of pseudogene
and parental gene

Pseudogenes with differential expression compared to parental genes are considered as “expressed”

Pseudogene expression levels are LOWER than coding gene expression

tissue
specific expression

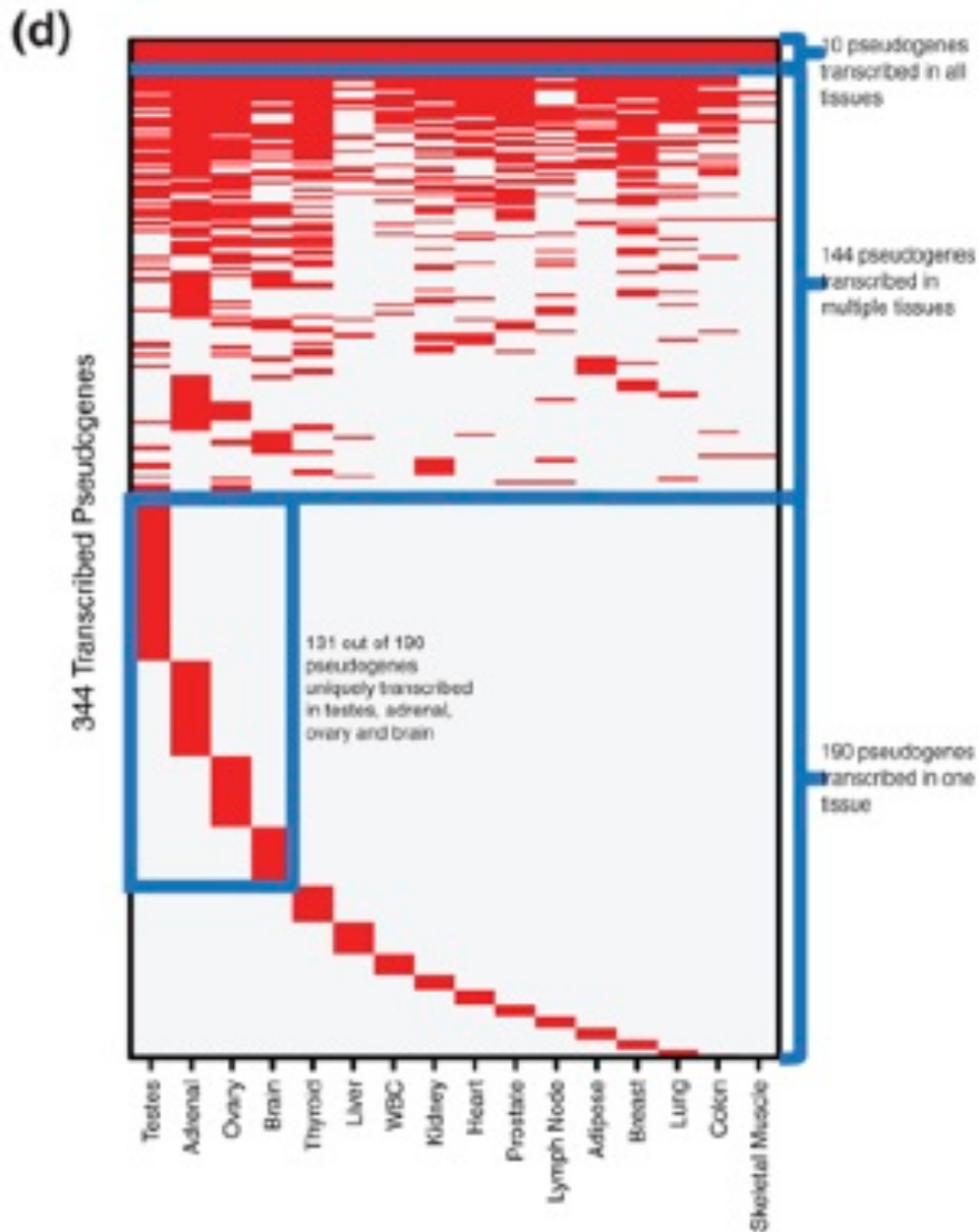
transcription of pseudogene



differential expression
parental/pseudogene

transcription of pseudogene
and parental gene

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 than 1 tissue
- 190/344 pseudogenes exclusively expressed in 1 tissue

**duplicated/processed
pseudogenes have
specific
regulatory elements!!**

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

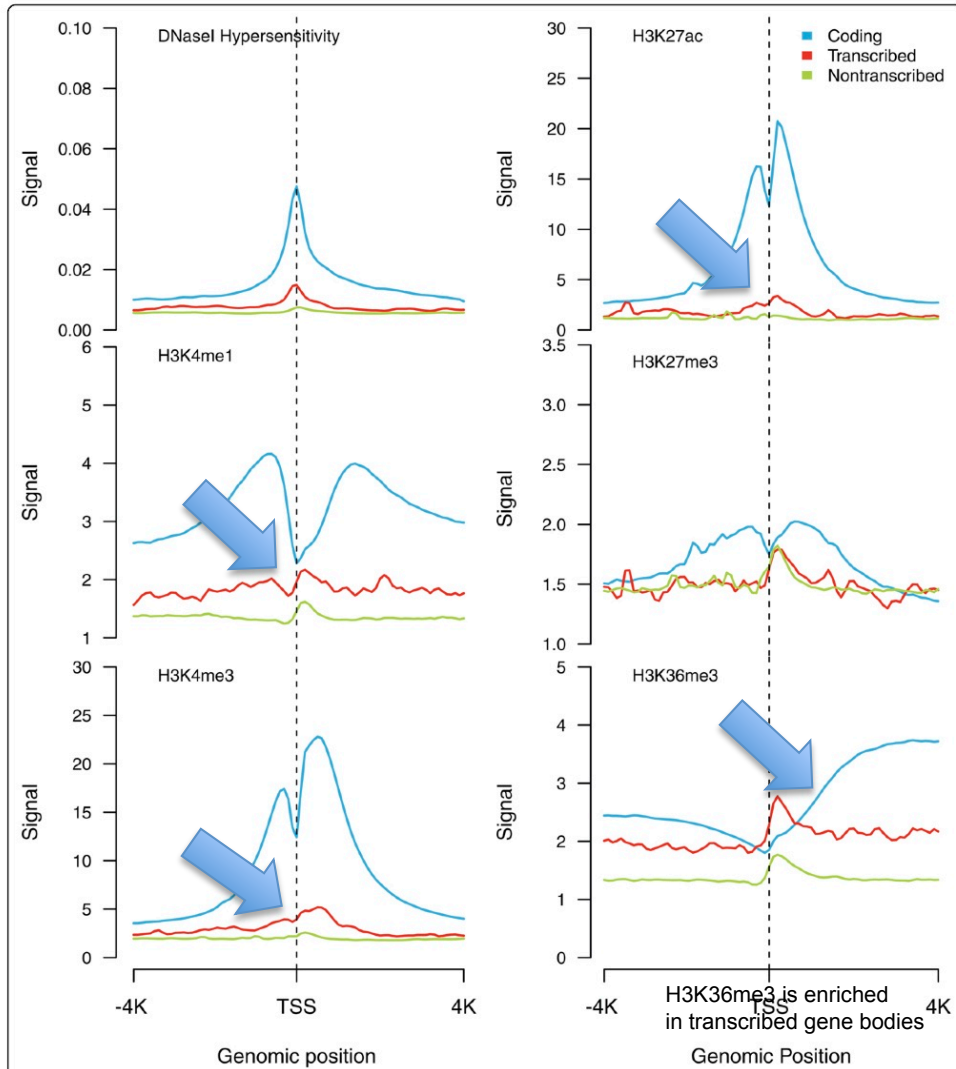
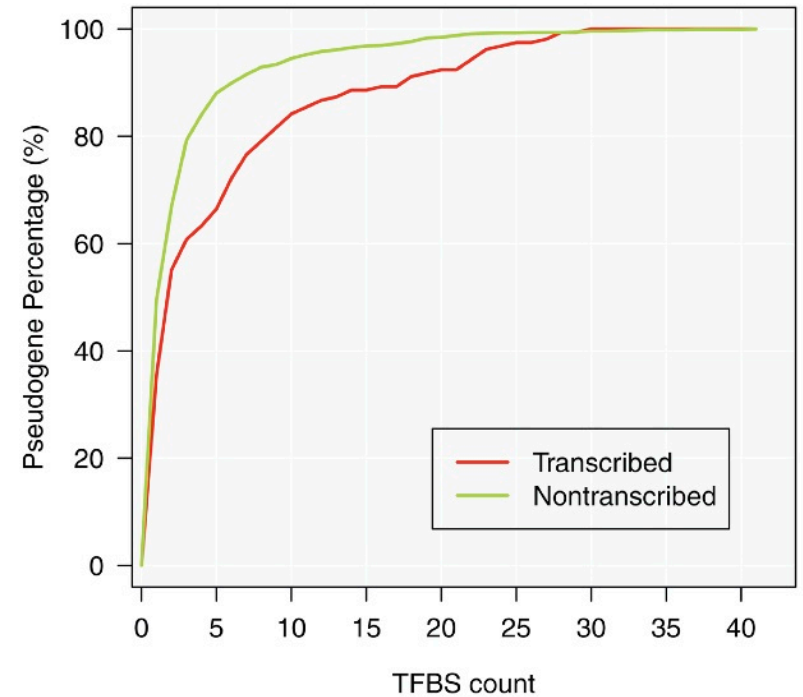


Figure 8 Chromatin signatures: DNaseI hypersensitivity and histone modification. Average chromatin accessibility profiles and various histone modifications surrounding the TSS for coding genes, transcribed pseudogenes, and non-transcribed pseudogenes. The coding gene histone modification profiles around 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.



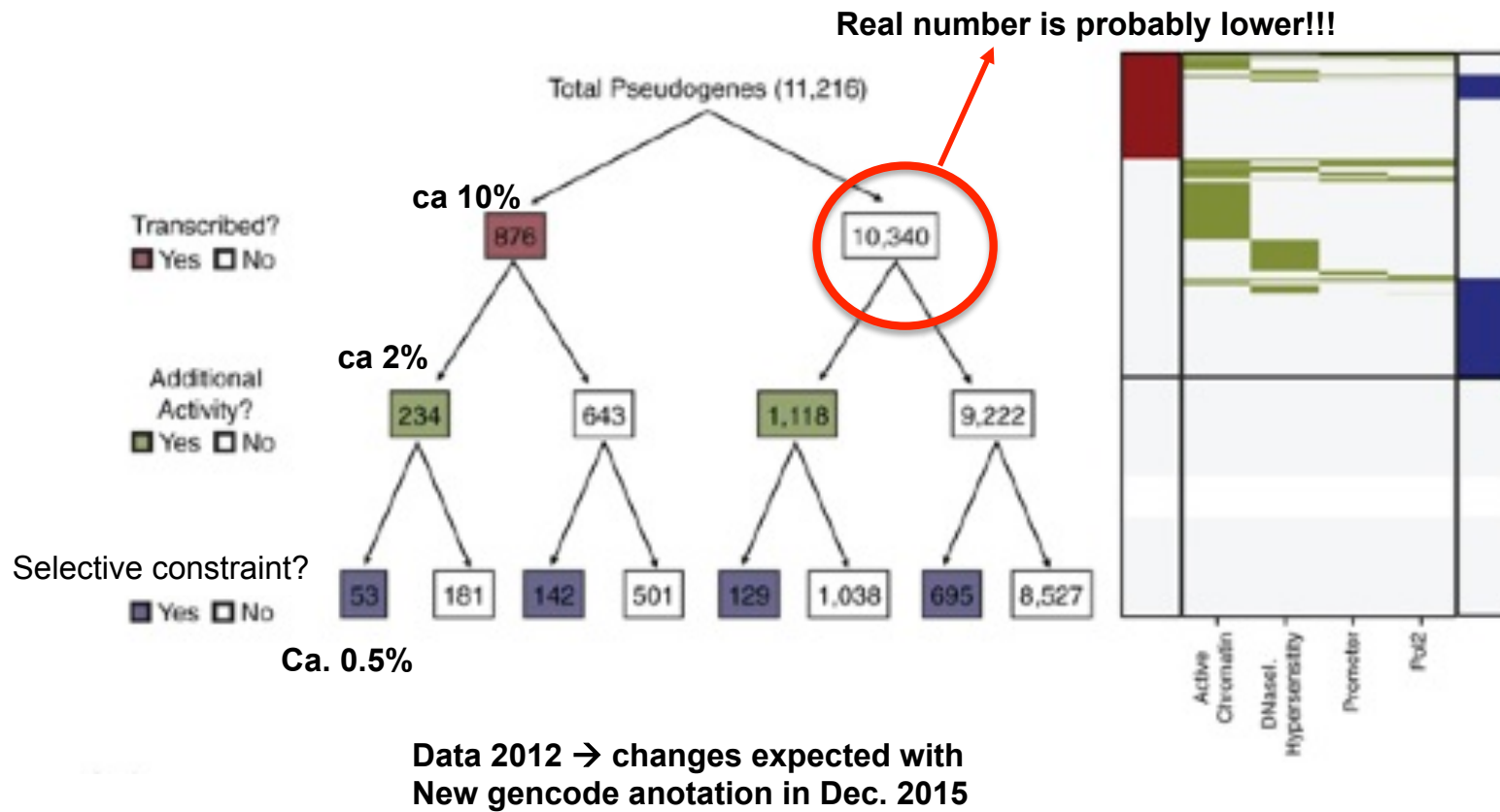
Frequency of transcription factor binding sites enriched in transcribed Pseudogenes vs non-transcribed pseudogenes

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



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

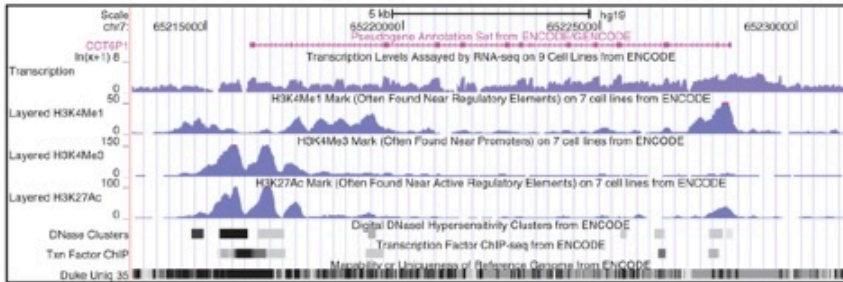
→ many pseudogenes show little or no activity

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

(b)

Transcribed With Additional Activity



Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor

Pseudogene
under selective
constraint
→ maintained

(c)

Transcribed Only

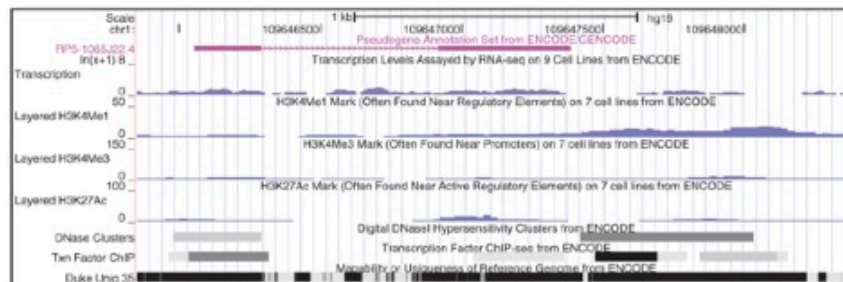


Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor

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:

(d)

Partially Active



Transcribed
DNase hypersensitive sites
Histonemarks
Transcription factor

A. dying pseudogene or
B. acquisition of critical
feature leading to the
resurrection to become a
functional pseudogene

Figure 12 Summary of pseudogene annotation and case studies. (a) A heatmap showing the annotation for transcribed pseudogenes including active chromatin segmentation, DNase 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 UGT7A2P

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 to be a ncRNA or gradually

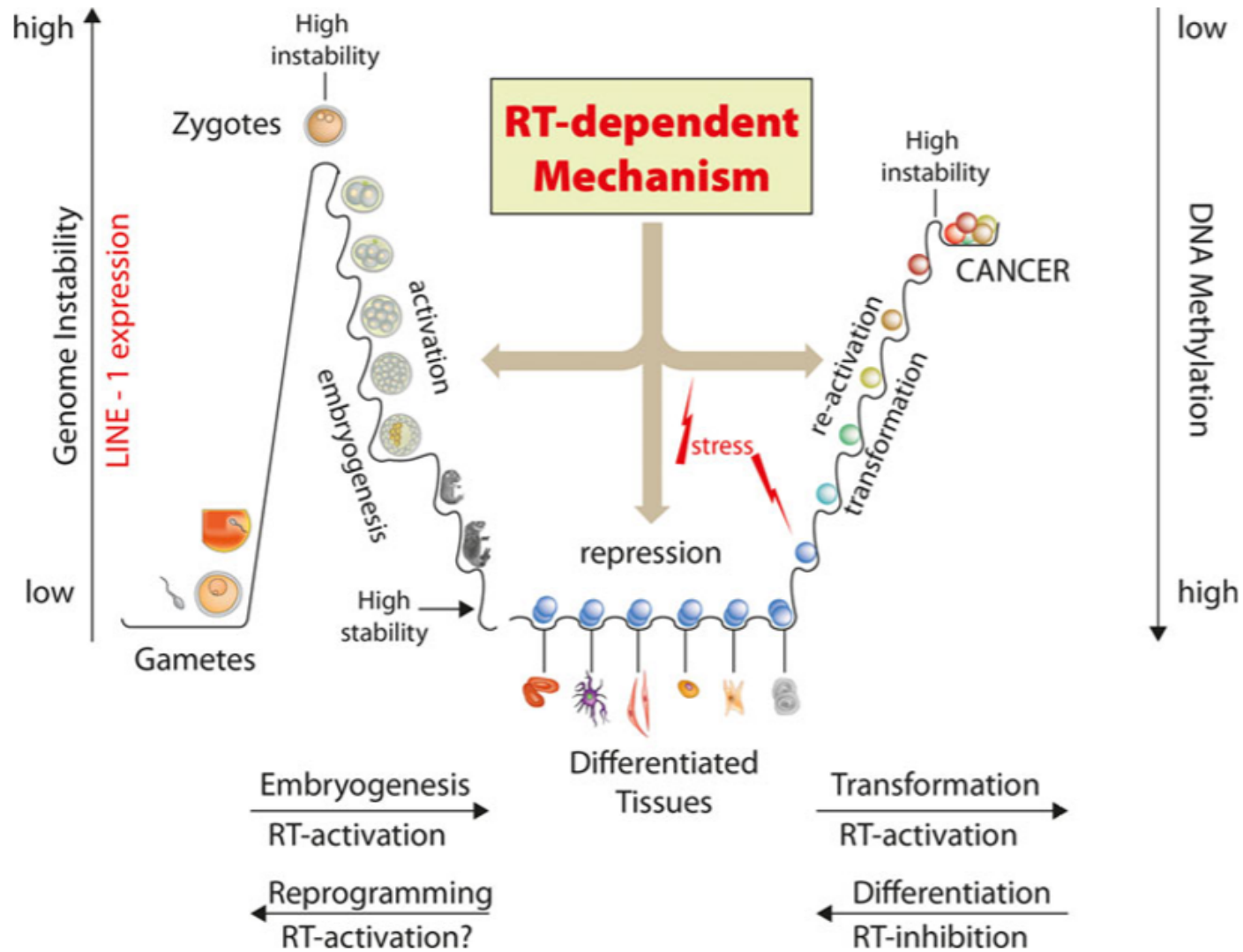
losing its functionality. Understanding why pseudogenes show partial activity may shed light on pseudogene evolution and function.

PSEUDOGENE lncRNAs EXAMPLE 1

***Pseudogene lncRNA that controls embryonic stem cell
Self-renewal***

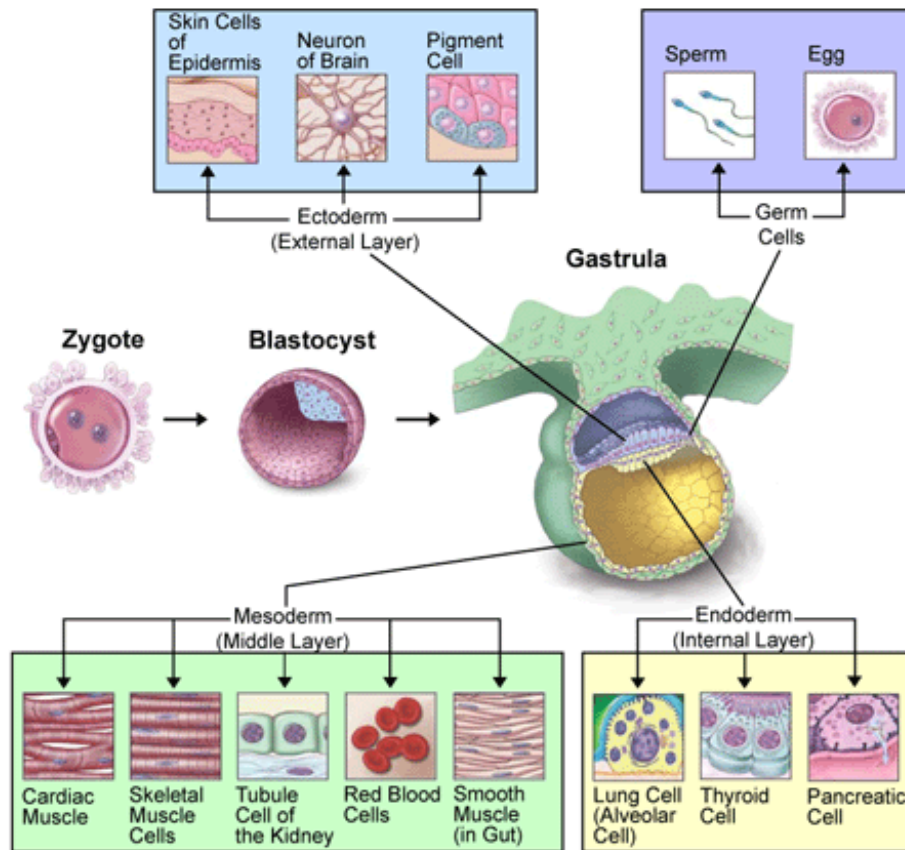
***A Oct4P4 pseudogene derived lncRNA silences the
ancestral Oct4 gene in trans***

Retrotransposon activity during development

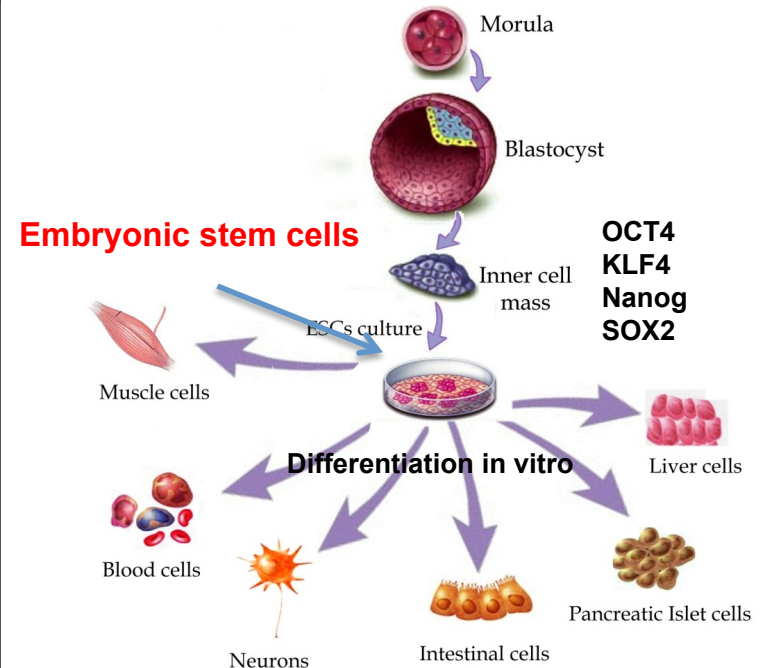


The inner cell mass of the blastocyst are the source of pluripotent embryonic stem cells

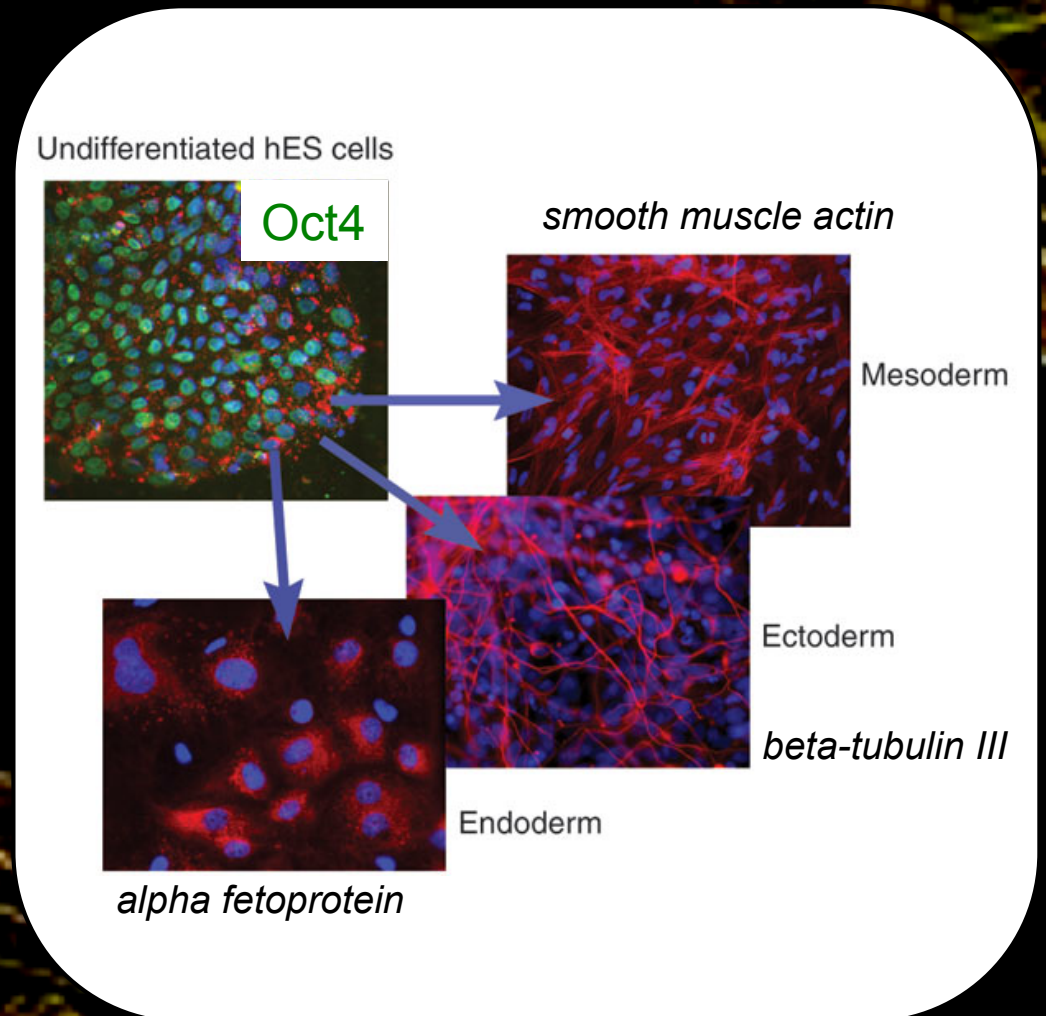
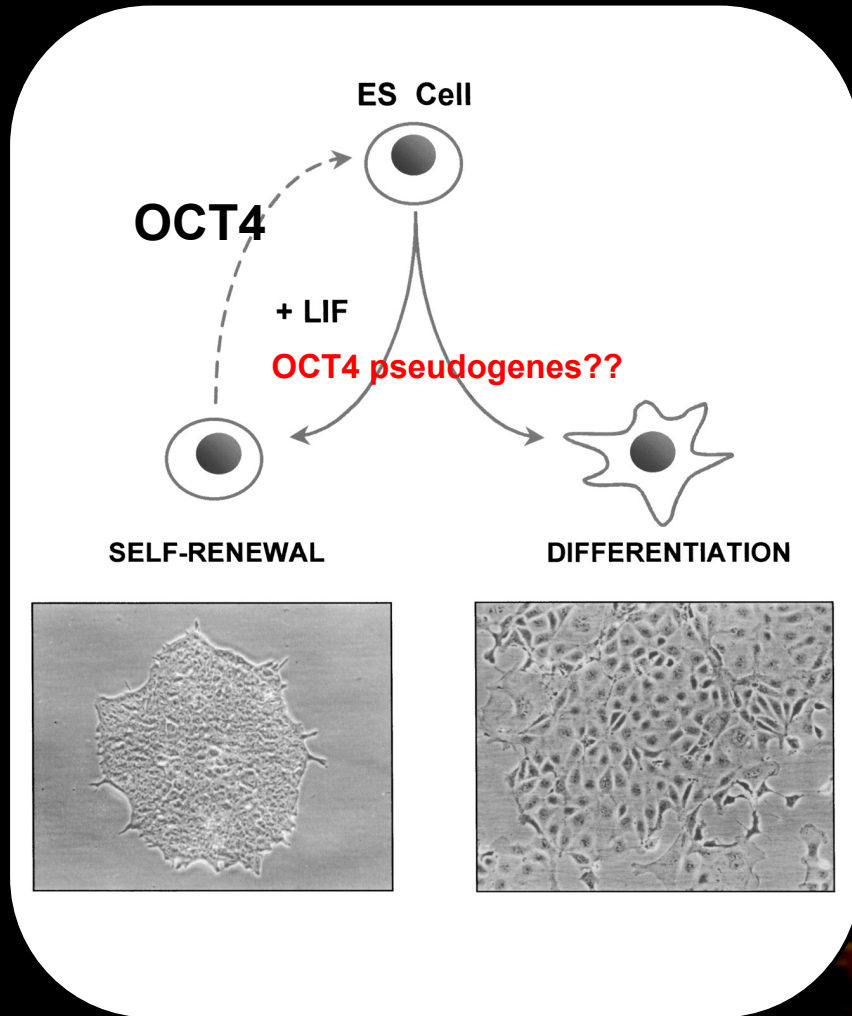
Blastocyst cells give rise to all organs and cell types



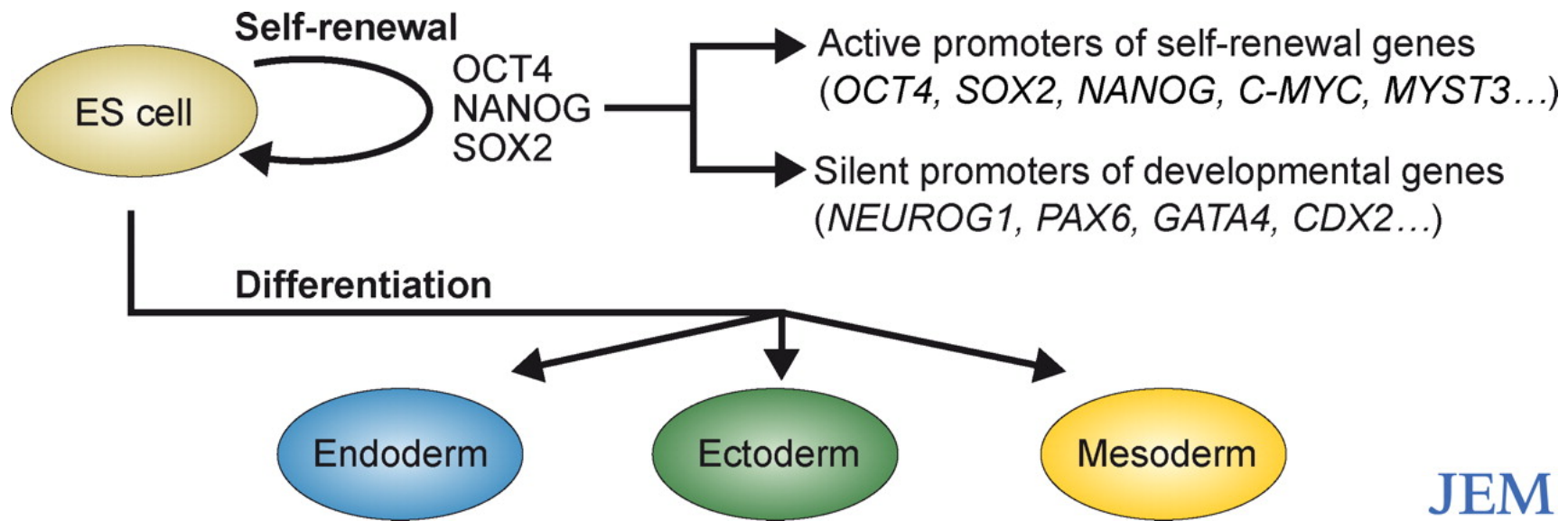
Blastocyst cells can be isolated and cultivated



OCT4 expressing ES cells have self-renewing and differentiation potential in vitro

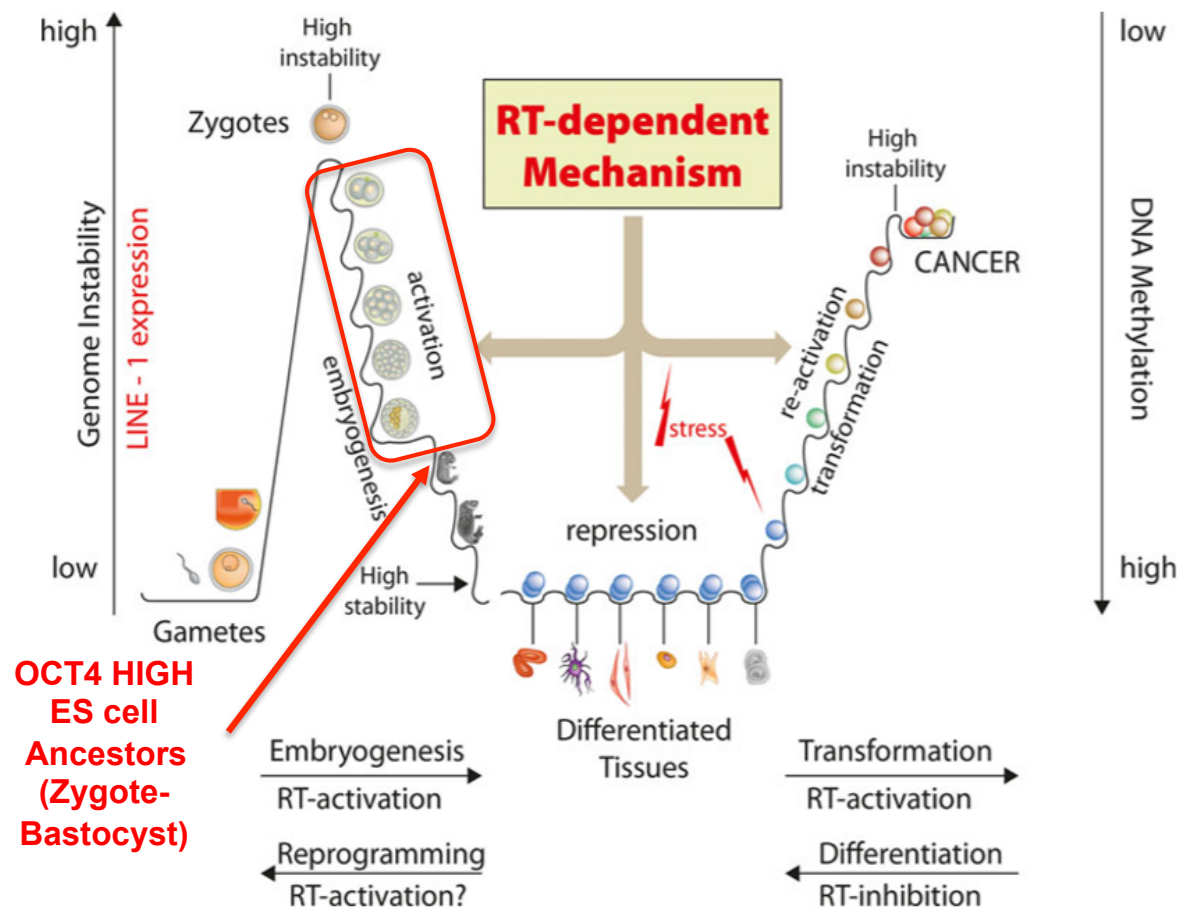


The self-renewal transcription factor Oct4 is essential for embryonic stem cell self-renewal



Nicolaj Strøyer Christophersen, and Kristian Helin J Exp Med
2010;207:2287-2295

Mouse and human contain several processed OCT4 pseudogenes

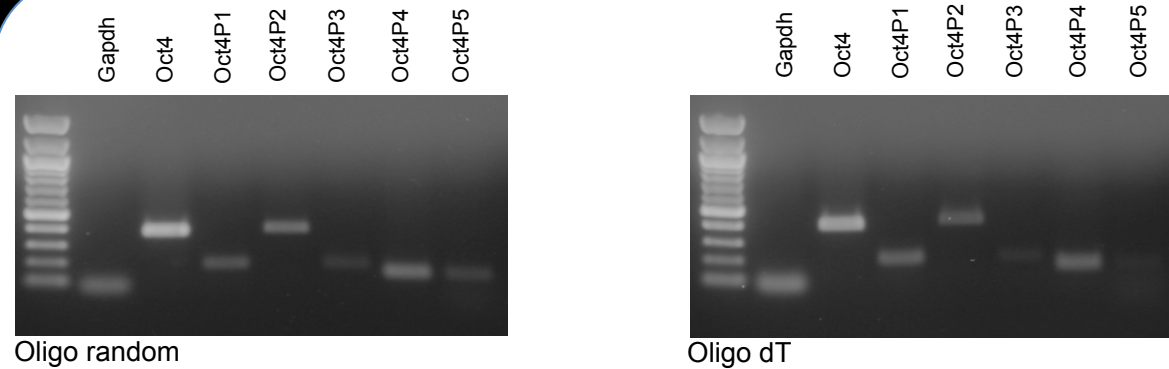
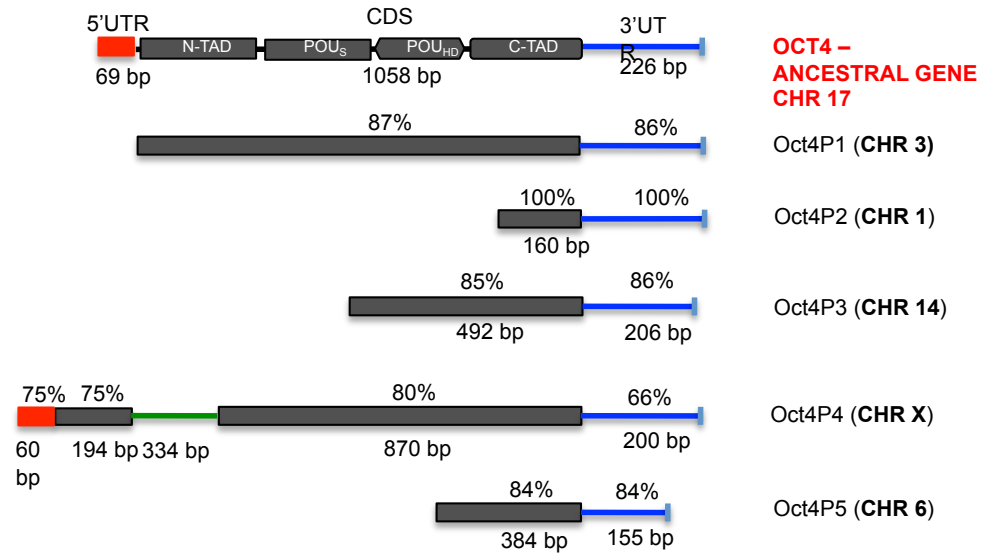


**OCT4 HIGH
ES cell
Ancestors
(Zygote-
Bastocyst)**

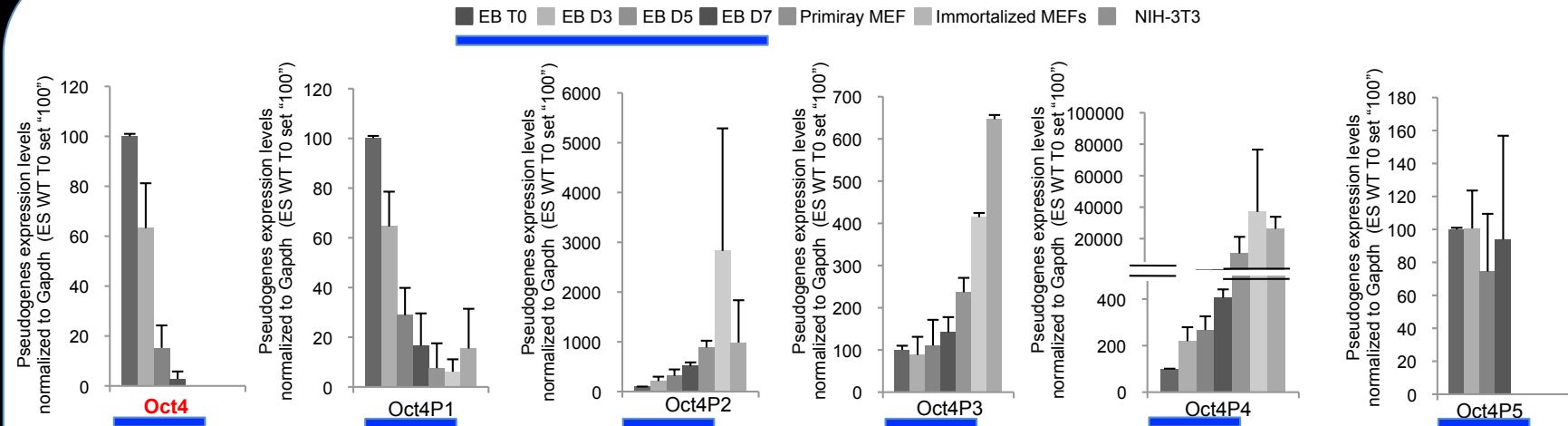
Human:
1 ancestral OCT4
7 processed OCT4
pseudogenes

Mouse:
1 nacestral Oct4
5 processed OCT4
pseudogenes

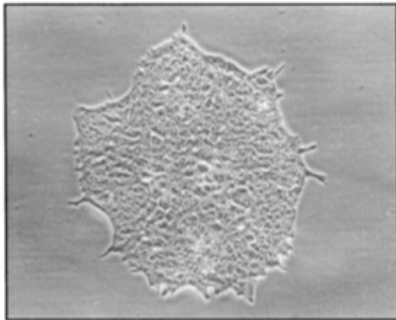
Ancestral OCT4 gave rise to 5 processed pseudogenes that are expressed in mESCs



Oct4 pseudogenes are tightly controlled during the differentiation of mESCs

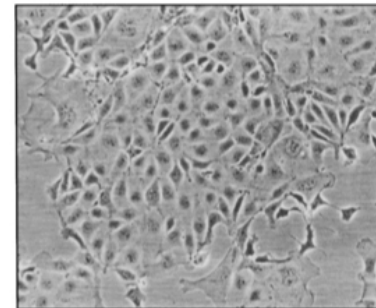


SELF-RENEWAL



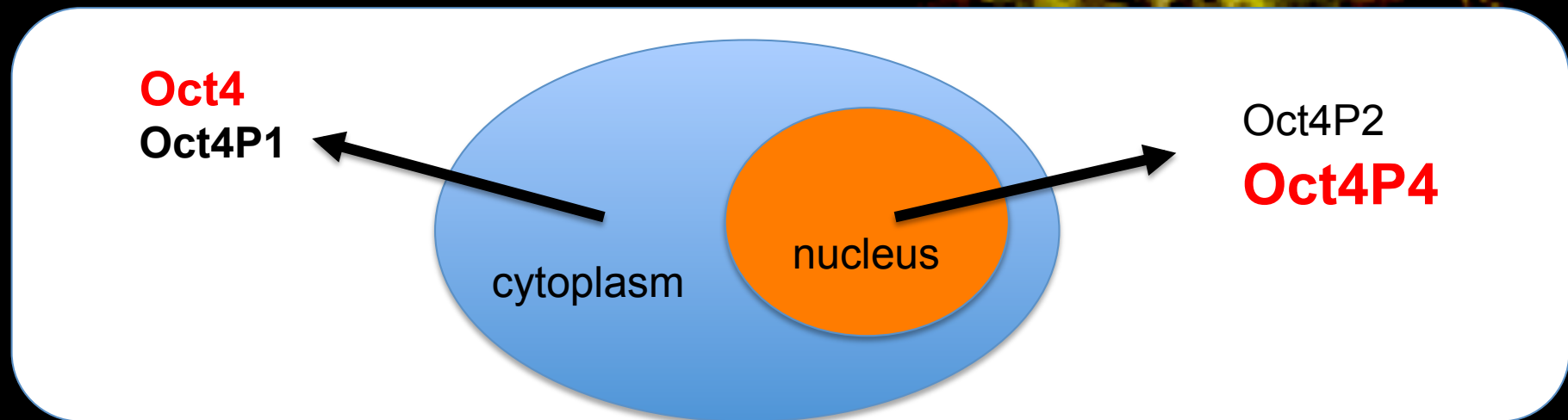
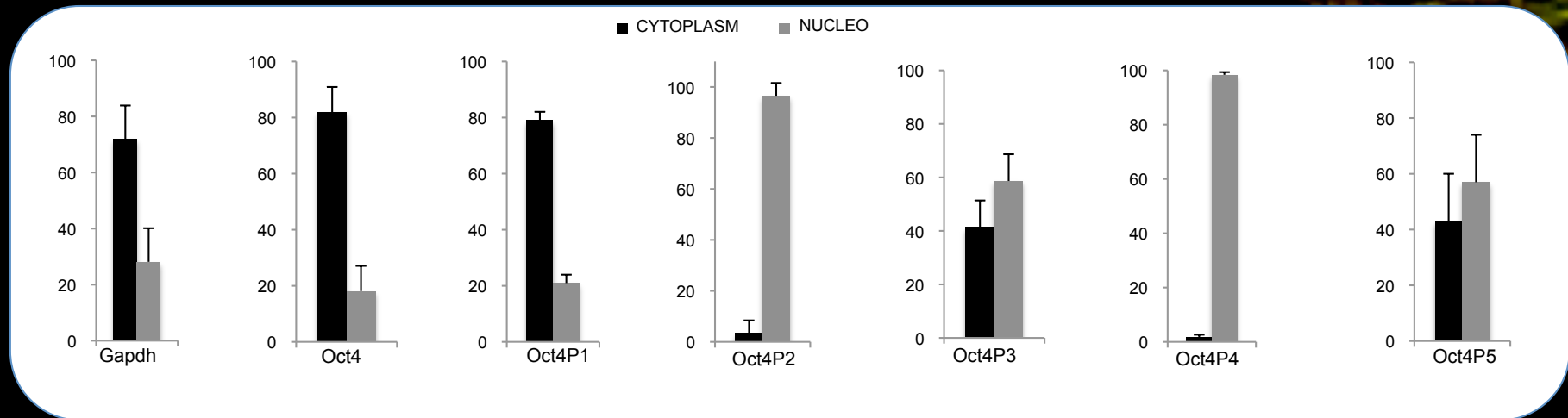
Oct4
Oct4P1 (-10X)

DIFFERENTIATION

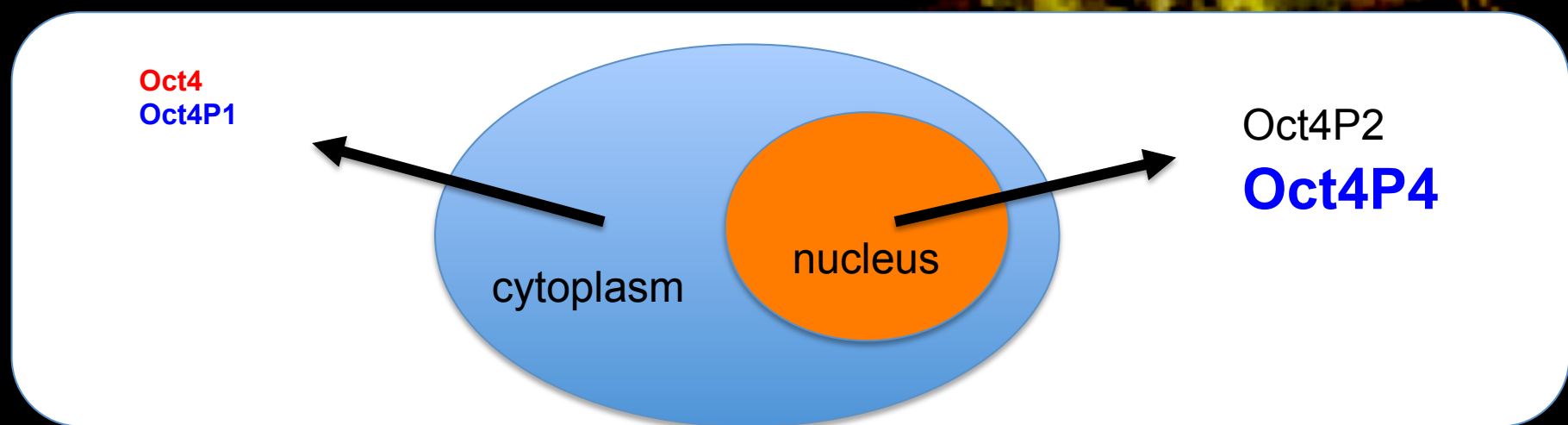
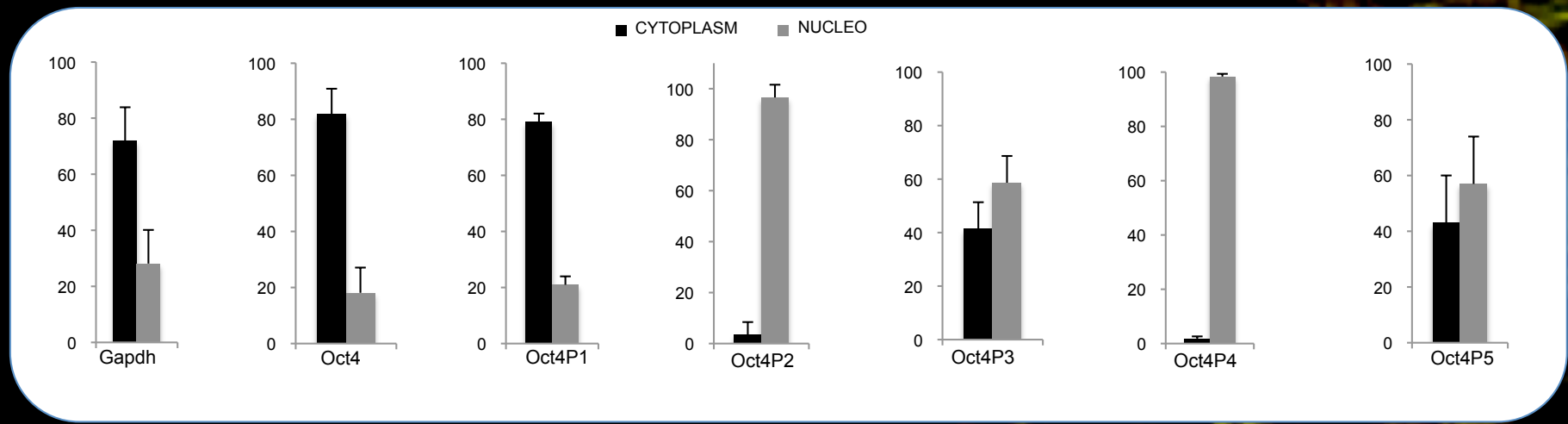


Oct4P2 (+9x)
Oct4P3 (+2x)
Oct4P4 (+4x;
Fiborbl. +200x)

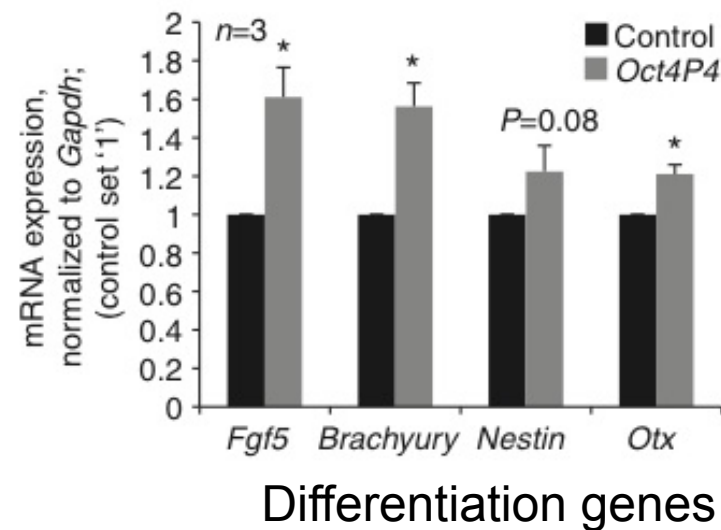
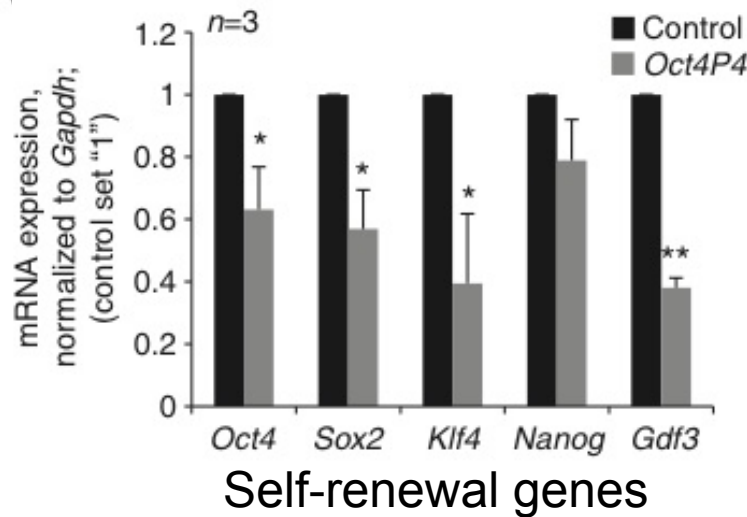
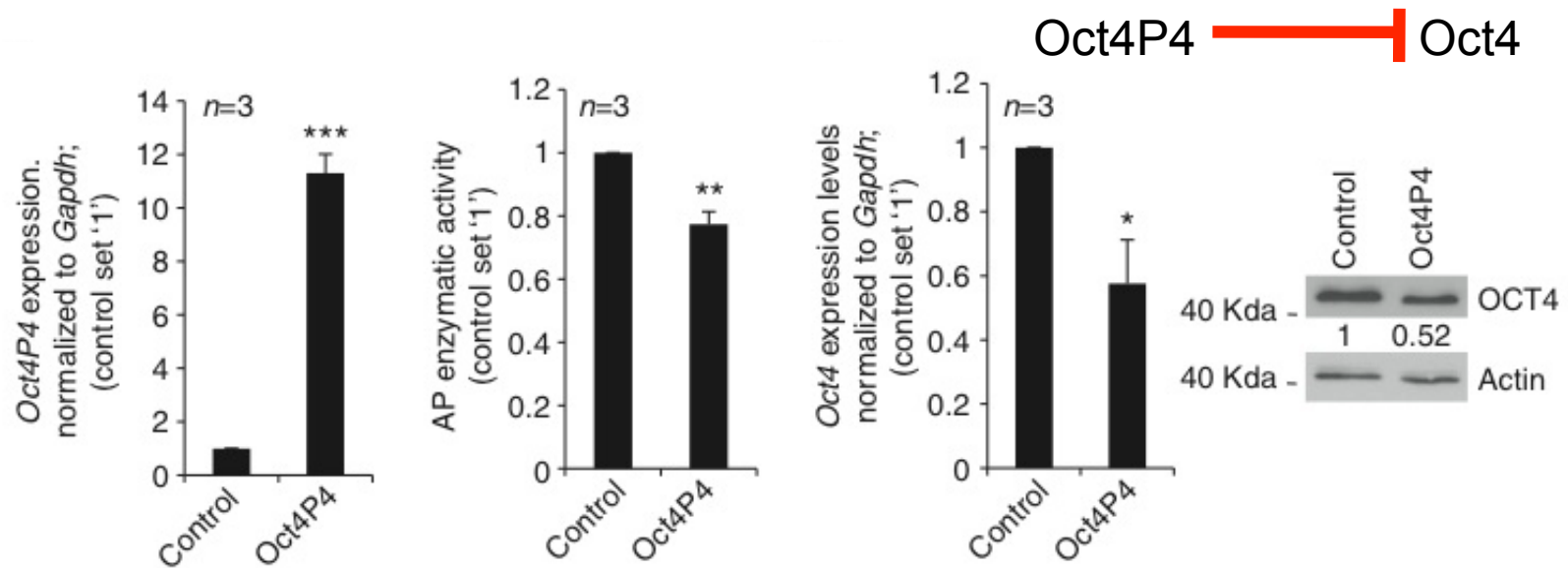
OCT4 pseudogenes are localized to nucleoplasm or cytoplasm



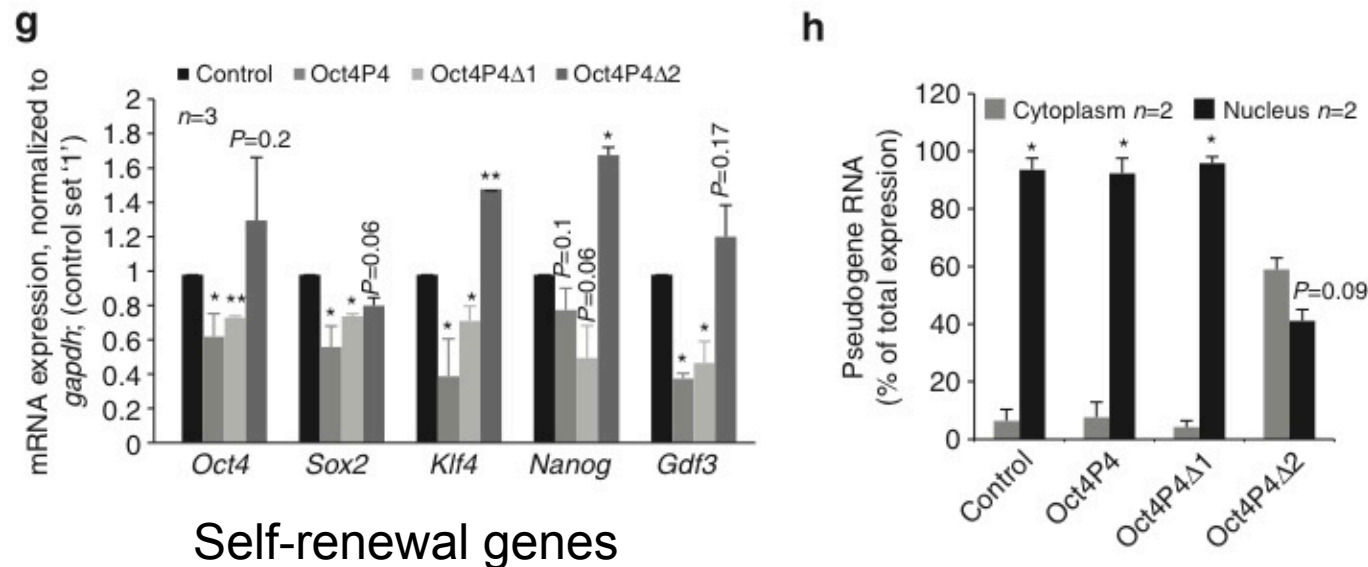
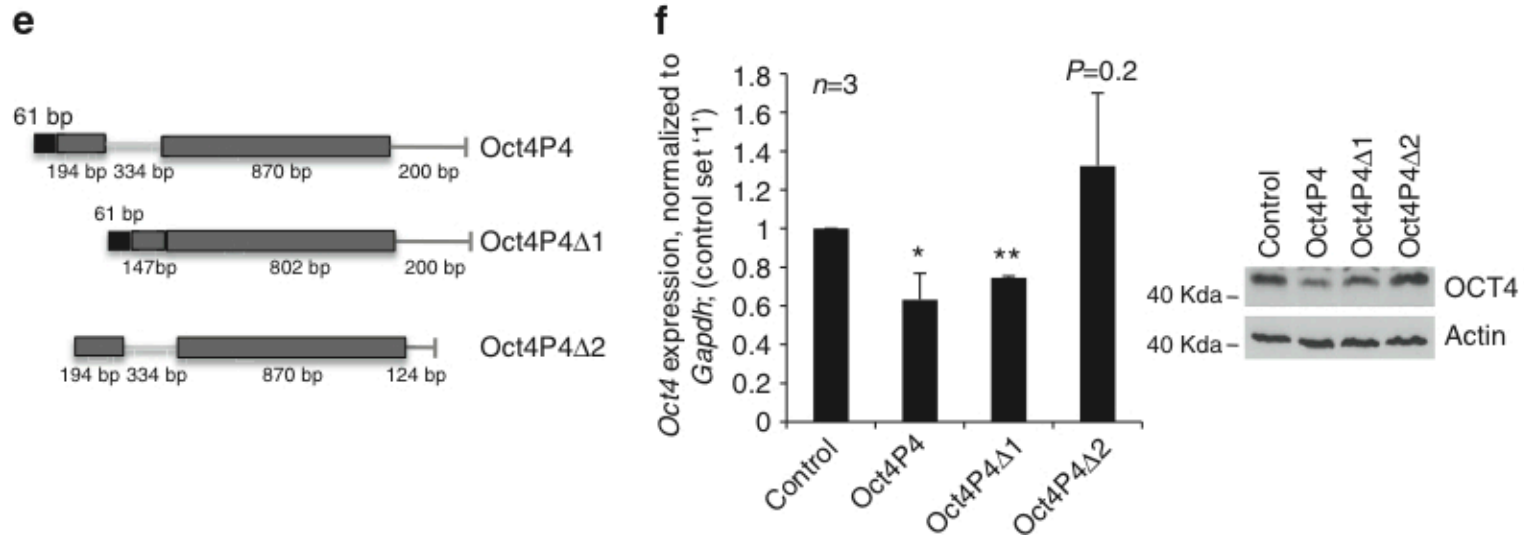
OCT4 pseudogenes are localized to nucleoplasm or cytoplasm



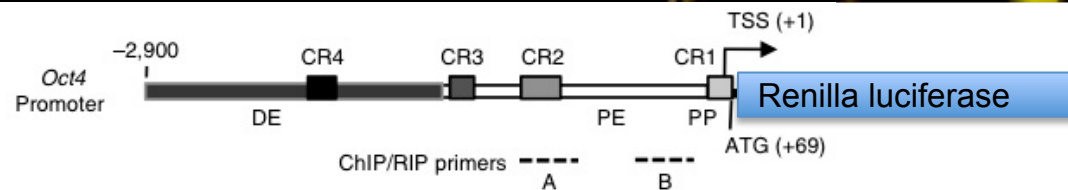
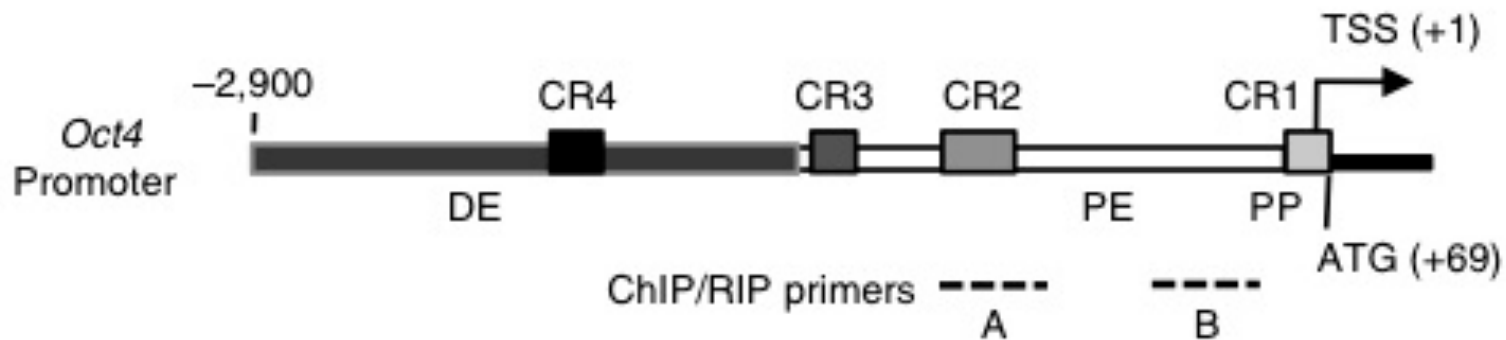
Nuclear OCT4P4 promotes mESC differentiation



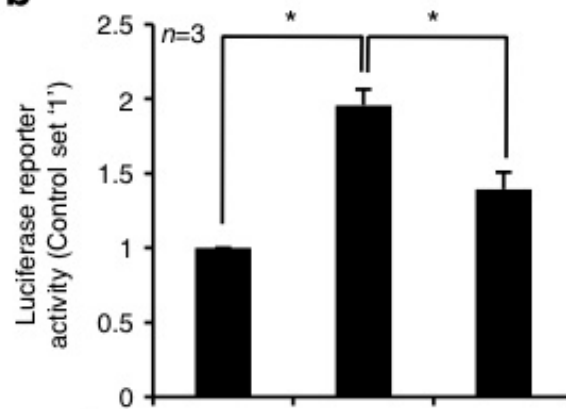
5' and 3' UTR homology domains are required to repress self-renewal marker genes



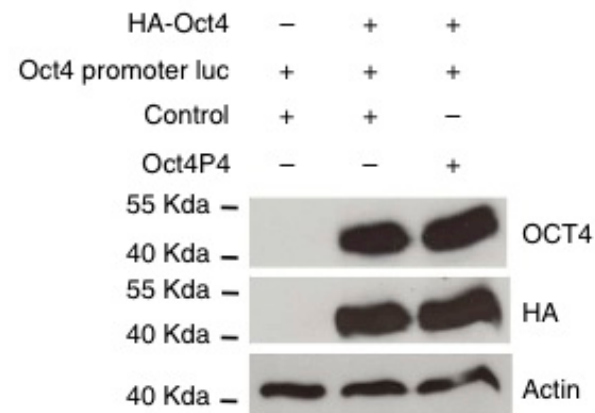
Oct4P4 interferes with the ancestral Oct4 promoter



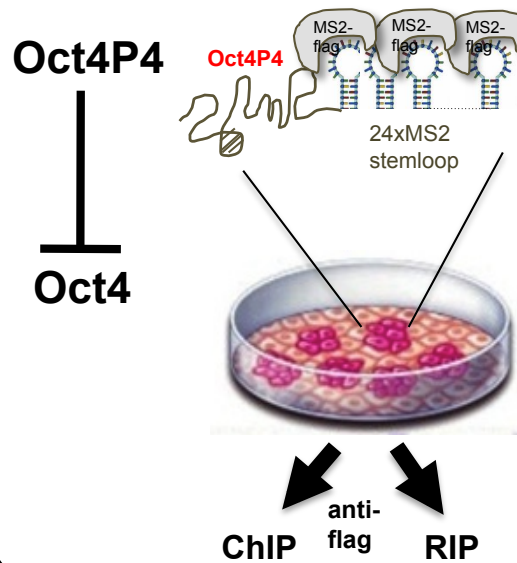
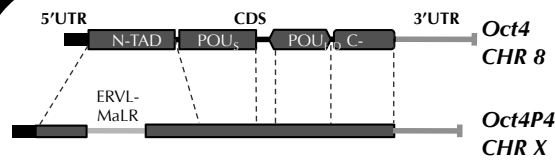
b



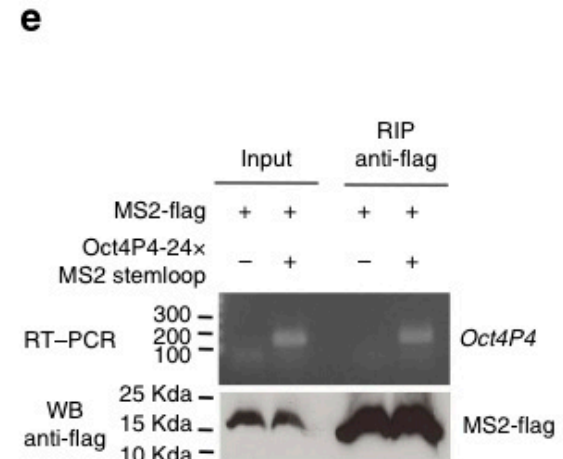
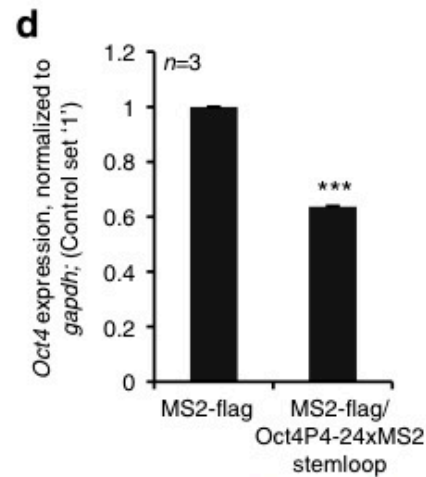
HA-Oct4	-	+	+
Oct4 promoter luc	+	+	+
Control	+	+	-
Oct4P4	-	-	+



A model system to study Oct4P4 lncRNA localization

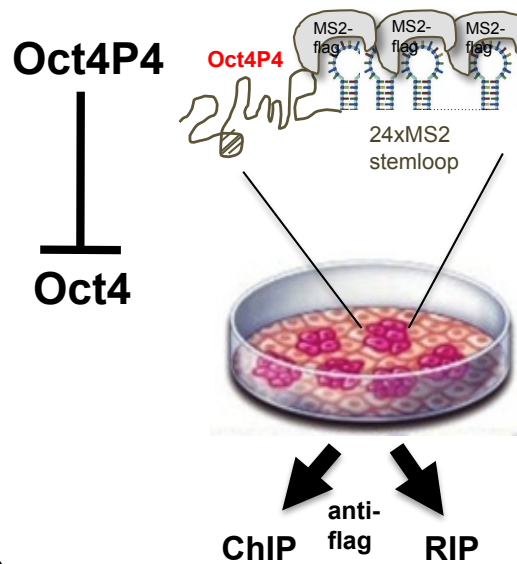
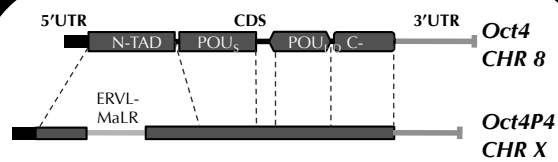


RNA immunoprecipitation anti-flag then RT-PCR for Oct4P4 lncRNA

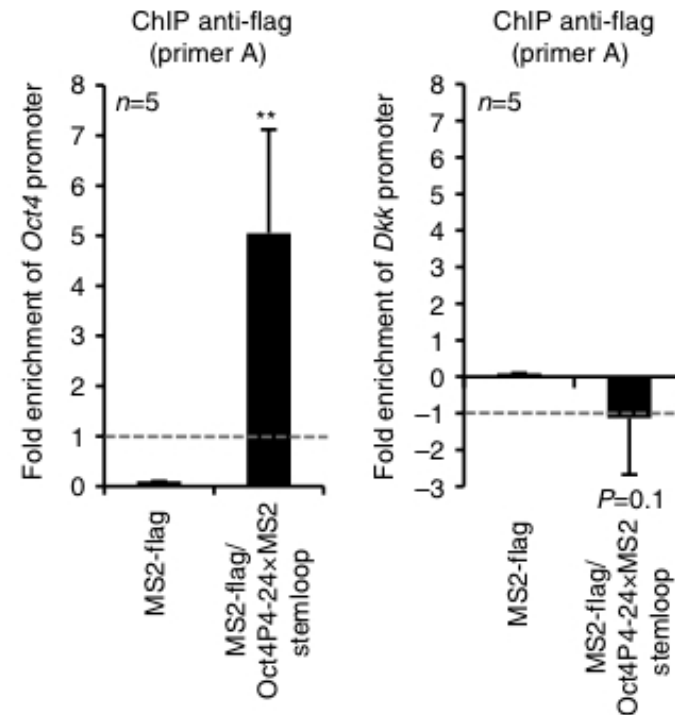


MS2 stem loop tagged Oct4P4 co-expressed with flag-MS2

A model system to study Oct4P4 lncRNA localization



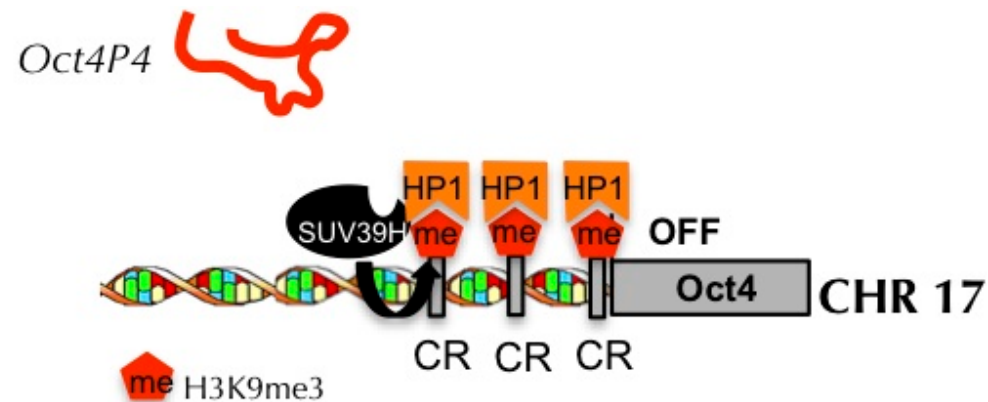
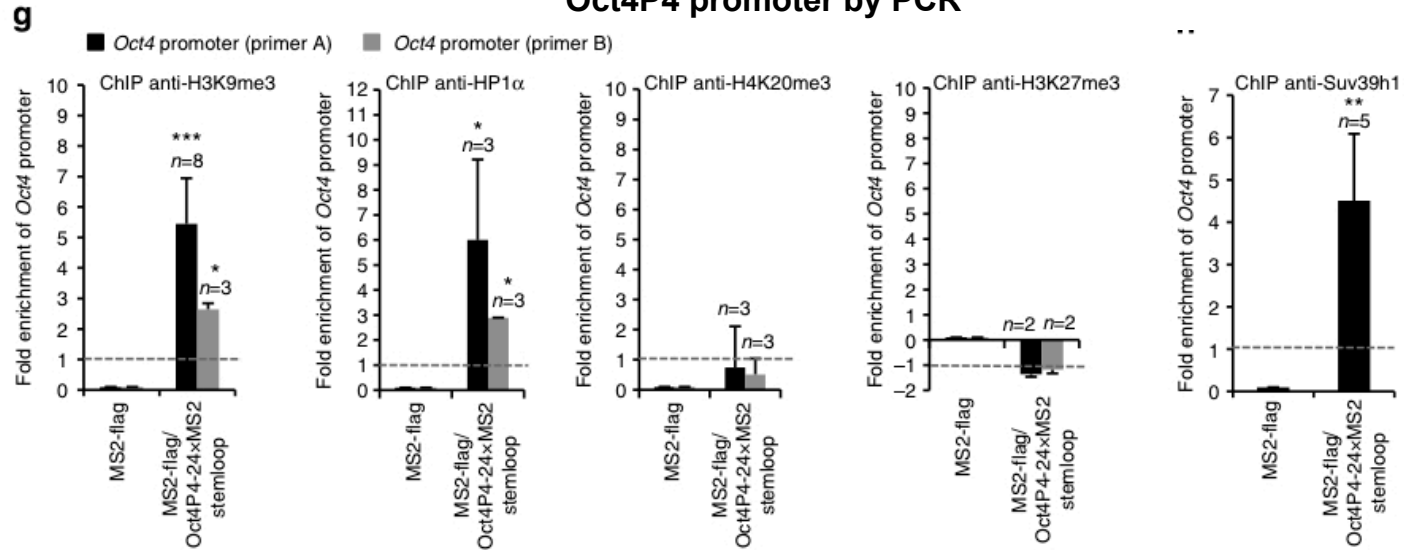
ChIP using anti-flag then use the immuno-precipitate to detect the *Oct4P4* promoter by PCR



Oct4P4-MS2 lncRNA localizes to *Oct4* promoter

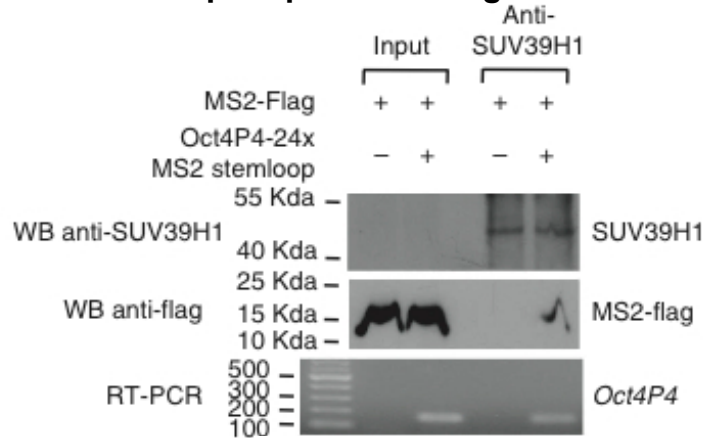
Oct4P4-MS2 directs Suv39h1 to Oct4 promoter

ChIP using specific antibodies then use the immuno-precipitate to detect the Oct4P4 promoter by PCR

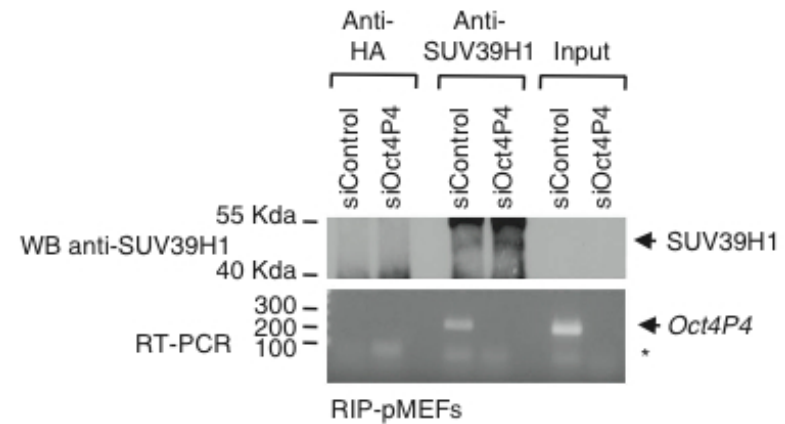


Oct4P4-MS2 directly interacts with Suv39h1

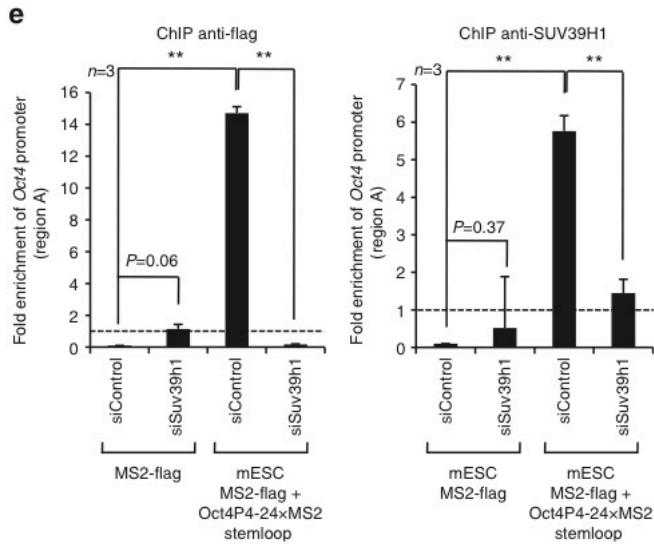
RNA immunoprecipitation using anti-SUV39h1;



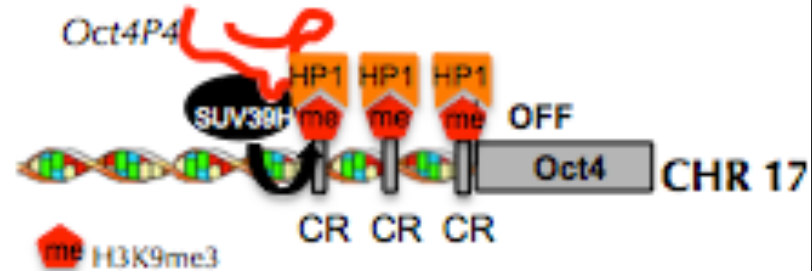
Western for MS2-flag



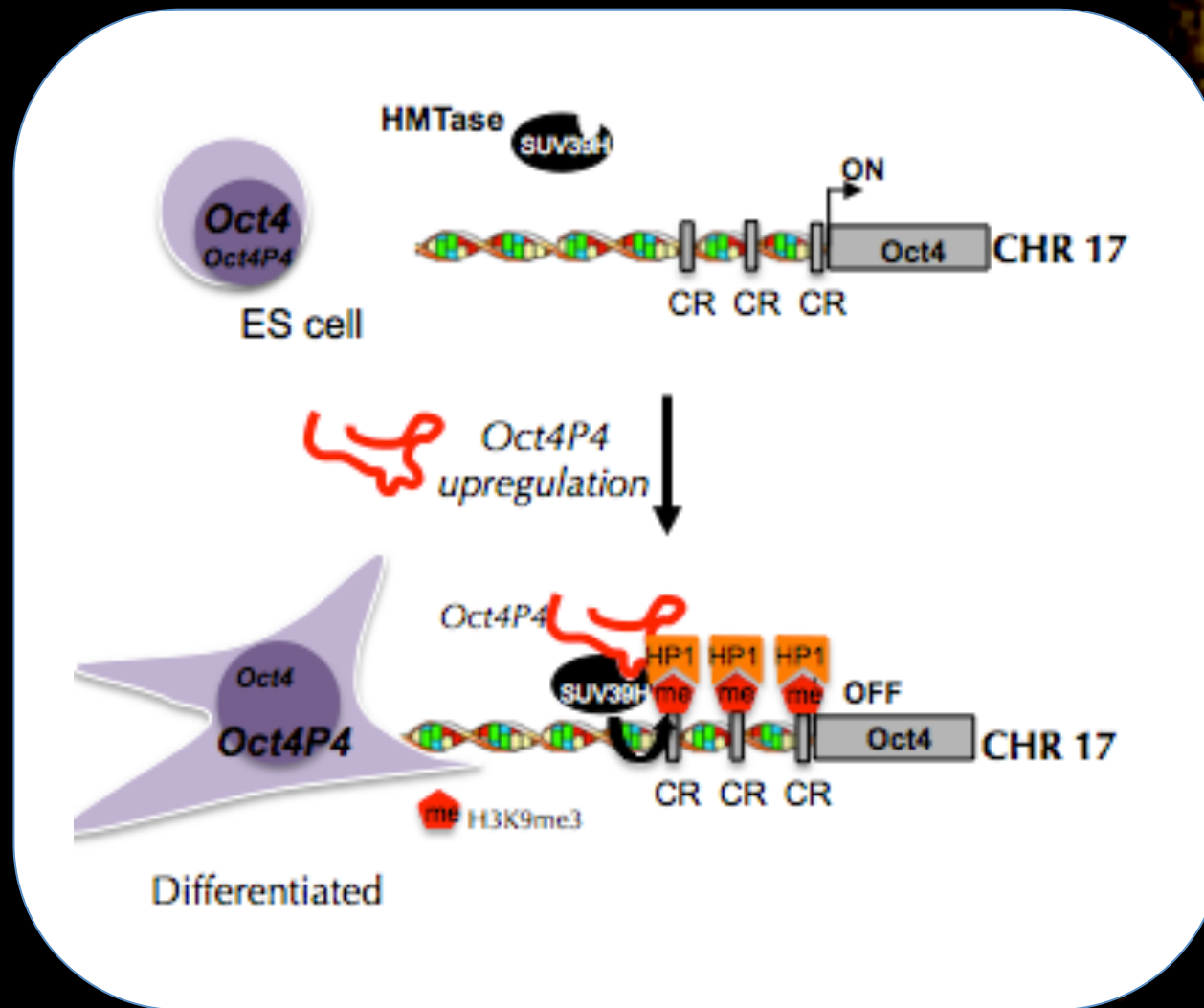
RT-PCR for Oct4P4 IncRNA



ChIP using specific antibodies then use the immuno-precipitate to detect the Oct4P4 promoter by PCR



Oct4P4-MS2 recruits Suv39h1 To direct silencing of the Oct4 promoter

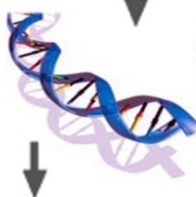


REVERSIBILITY???

INDUCING PLURIPOTENCY IN ADULT CELLS

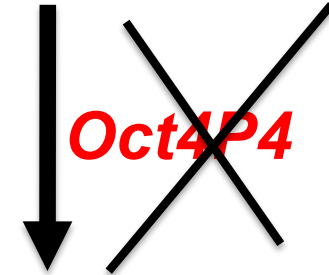


Adult cells



Genes inserted to induce reprogramming
Oct4, Klf4, Sox2, c-Myc

DIFFERENTIATED



Oct4P4

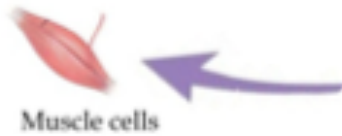
Reprogram into
ES like-cells



iPS cells

=induced pluripotent cells

**UNDIFFERENTIATED
(SELF-RENEWAL)**



Muscle cells



Liver cells



Blood cells



Neurons



Intestinal cells

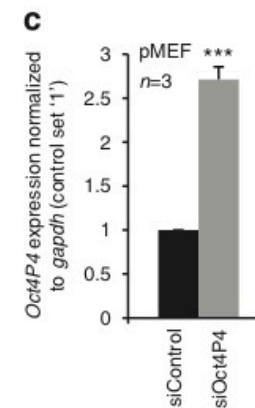
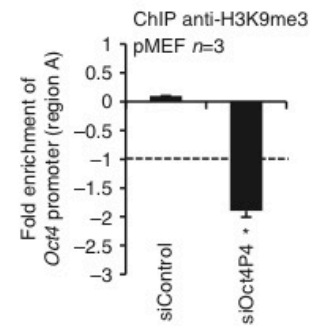
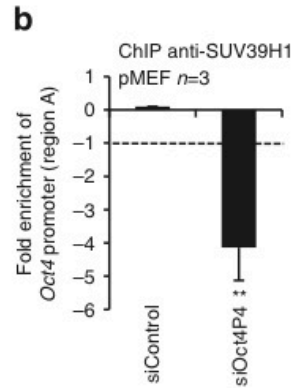
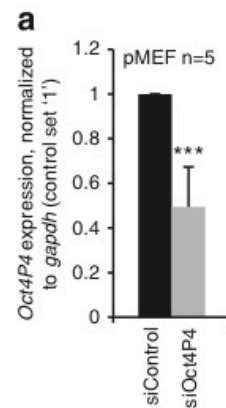
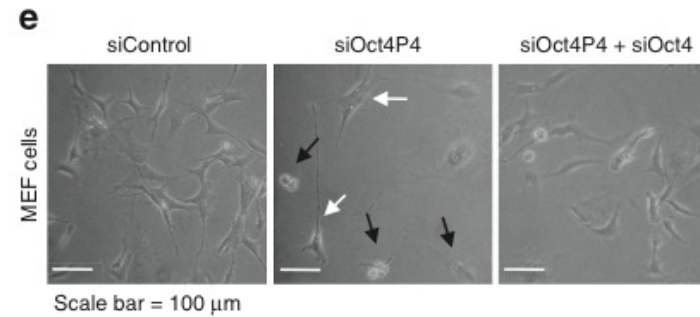
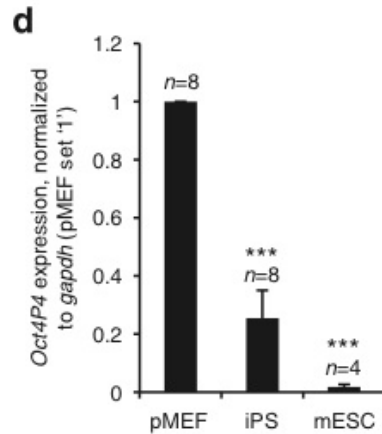


Pancreatic Islet cells

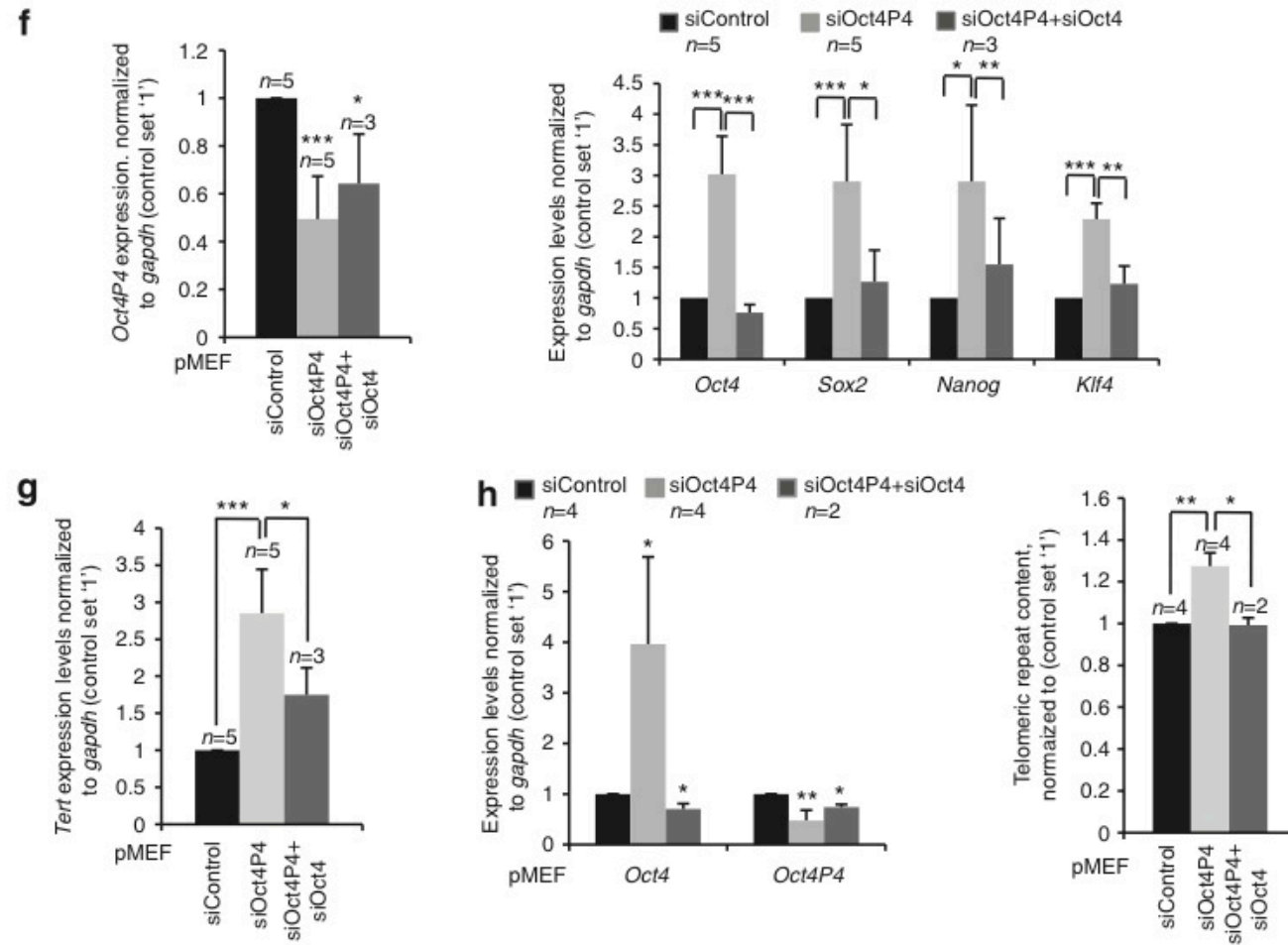
**INDUCE CELL
LINES SPECIFIC
DIFFERENTIATION**

Allotherapy/Cell therapy

Oct4P4 depletion in pMEFs causes the re-acquisition of self-renewal features

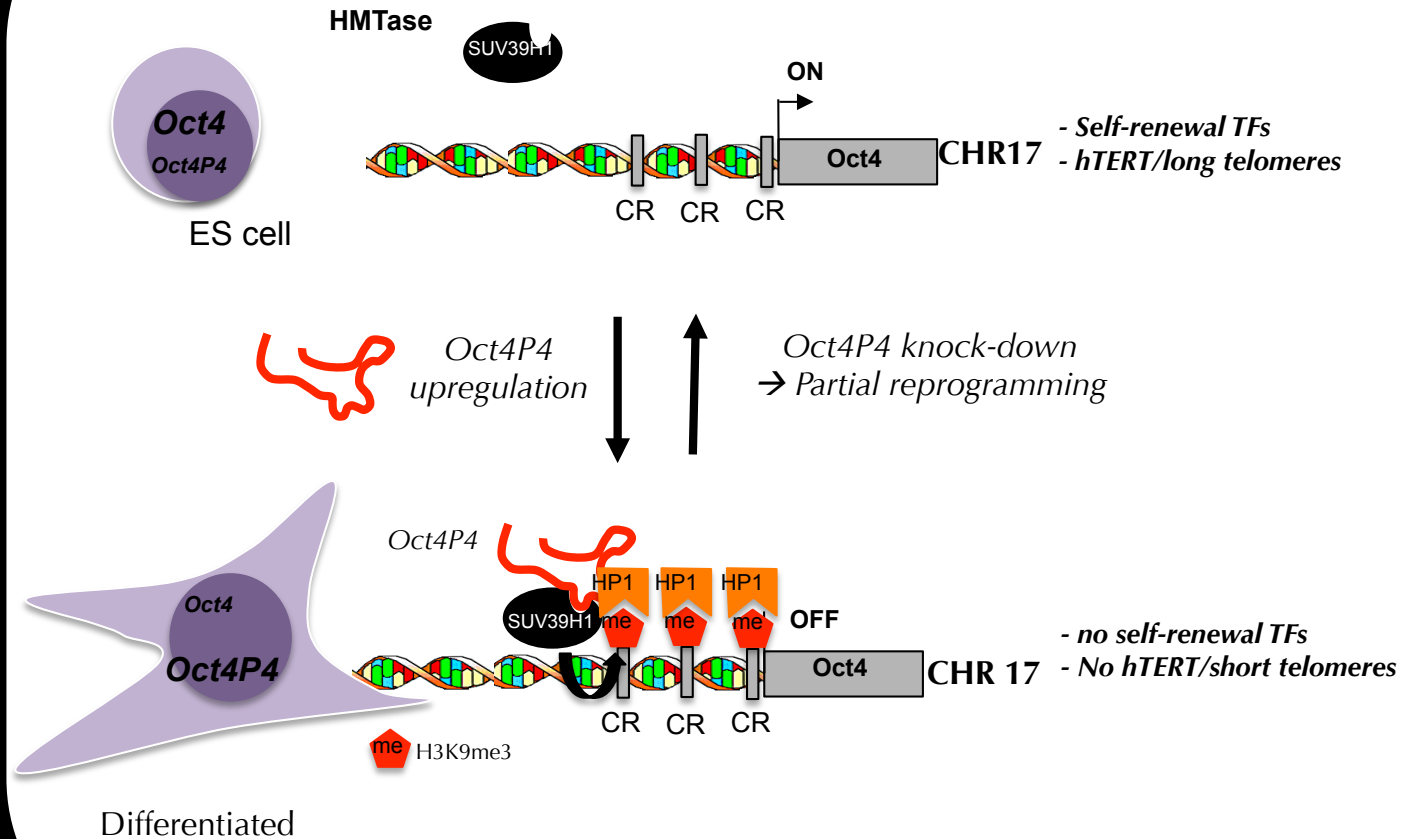


Oct4P4 depletion in pMEFs causes the re-acquisition of self-renewal features



Pseudogenes control the epigenetic status of ancestral genes

Oct4 pseudogene lncRNA silences ancestral gene



Scarola et al. Under review in Nat. Comm

Pseudogenes are powerful regulators of gene expression

