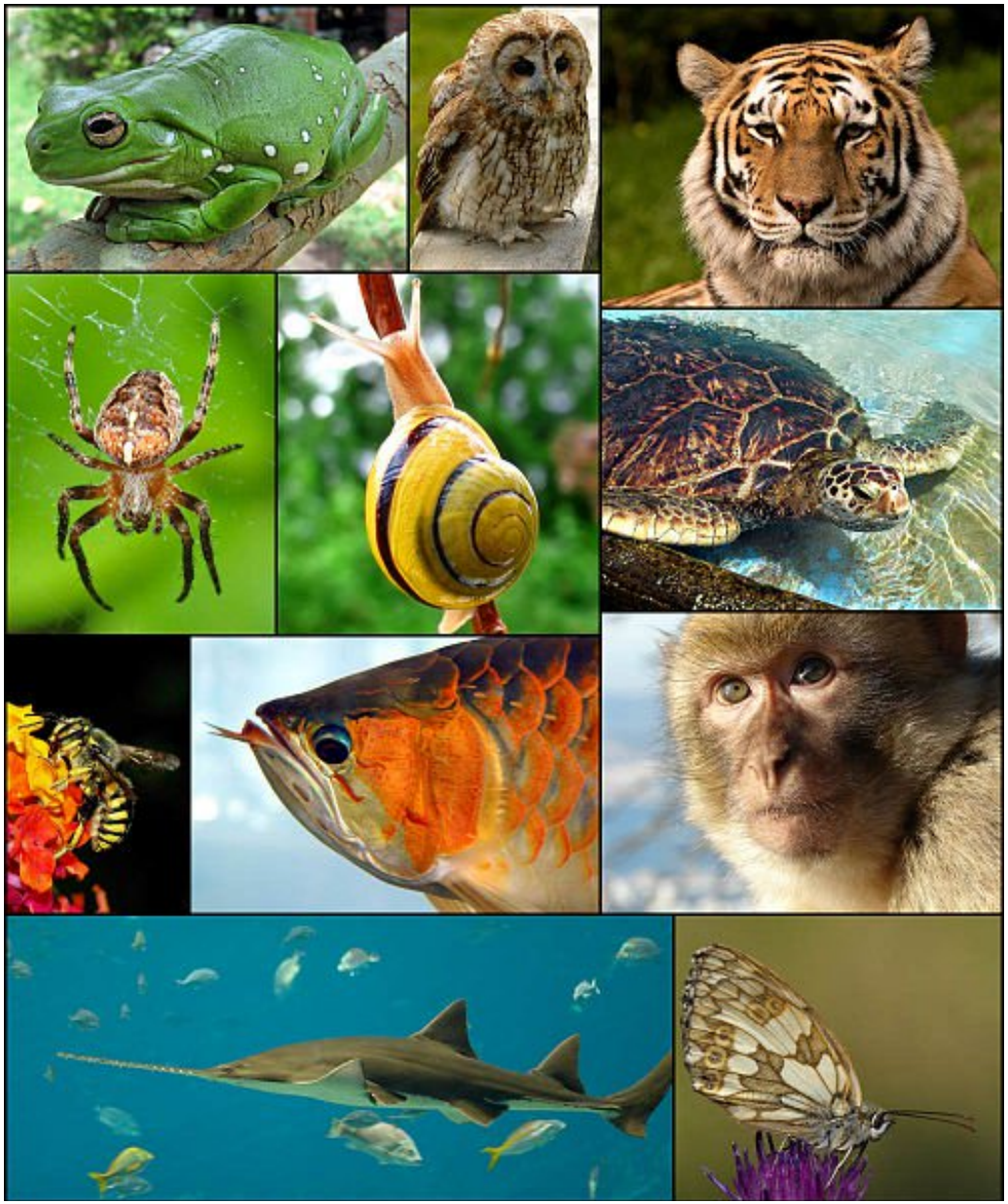
Building from scratch: *de novo* gene birth

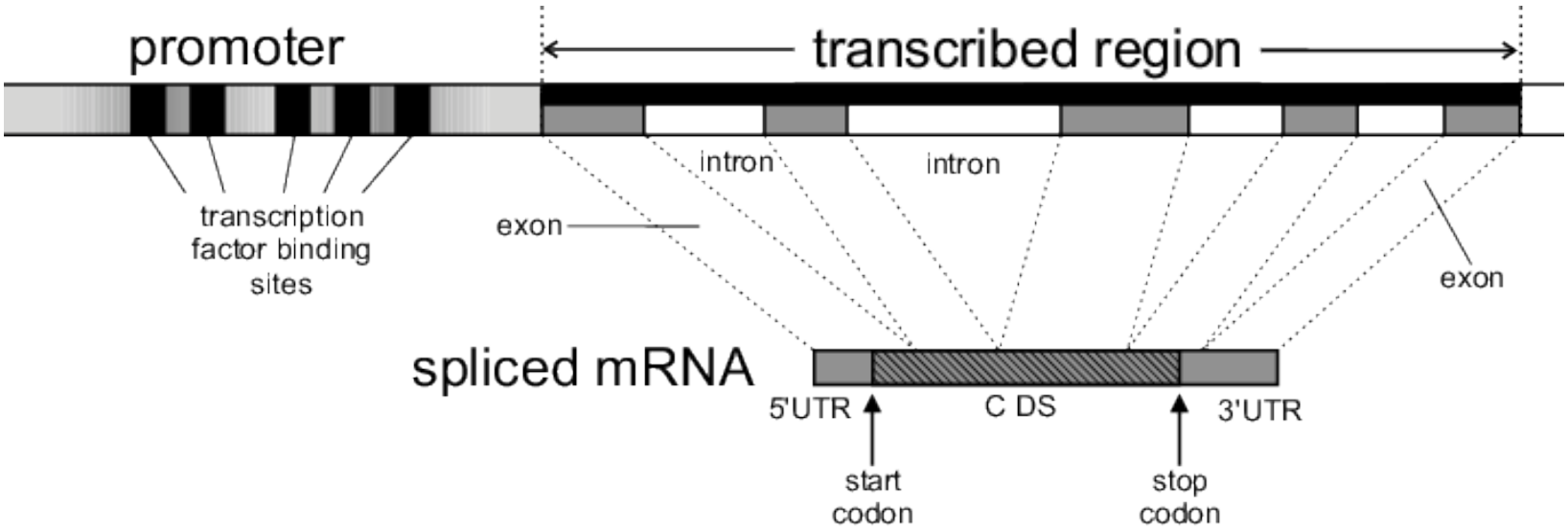
M.Mar Albà

ICREA - Fundacio IMIM - Universitat Pompeu Fabra (UPF), PRBB, Barcelona
evolutionarygenomics.imim.es

BSC, March 26 201



Genes


promoter
transcribed region
intron
intron
transcription factor binding sites
exon
exon
spliced mRNA
5'UTR
C DS
3'UTR
start codon
stop codon

How are new genes gained?

- Use of existing coding sequences to make a similar protein: **gene duplication, horizontal gene transfer**

How are new genes gained?

- Use of existing coding sequences to make a similar protein: **gene duplication, horizontal gene transfer**

- Use of existing coding sequences to make a different protein: **overprinting**

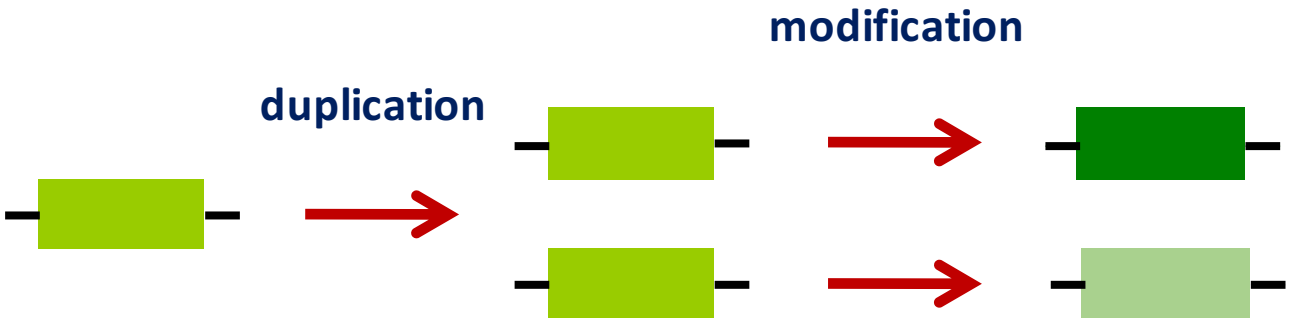
How are new genes gained?

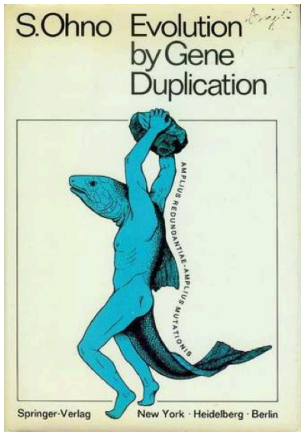
- Use of existing coding sequences to make a similar protein: **gene duplication, horizontal gene transfer**

- Use of existing coding sequences to make a different protein: **overprinting**

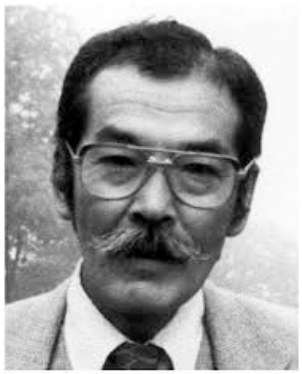
- Exaptation of non-coding sequences into a coding function: ***de novo* gene birth**

Gene duplication



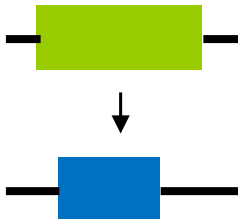


1970



Overprinting

translation of an alternative frame



original protein

new protein

Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 2421–2425, April 1984
Evolution

Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence

(oligomeric repeats/nylon oligomer degrading enzymes)

SUSUMU OHNO

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 9489–9493, October 1992
Evolution

Origins of genes: "Big bang" or continuous creation?

(overlapping genes/new genes/thyroid hormone receptor α2/plant viruses/human immunodeficiency virus)

PAUL K. KEESE* AND ADRIAN GIBBS†

De novo gene birth

non-coding

no protein

coding

new protein

DOI: 10.1534/genetics.106.069245

Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade

David J. Begun,*[,1] Heather A. Lindfors,* Andrew D. Kern* and Corbin D. Jones[†]

*Section of Evolution and Ecology, University of California, Davis, California 95616 and [†]Department of Biology

Plant J. 2009 May;58(3):485-98. doi: 10.1111/j.1365-313X.2009.03793.x. Epub 2008 Jan 18.

Identification of the novel protein QQS as a component of the starch Arabidopsis leaves.

Li L[1], Foster CM, Gan Q, Nettleton D, James MG, Myers AM, Wurtele ES.

Letter

Recent de novo origin of human protein-coding

David G. Knowles and Aoife McLysaght[1]

Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

DOI: 10.1534/genetics.107.084491

Origin of Primate Orphan Genes: A Comparative Genomics Approach

Macarena Toll-Riera, Nina Bosch, Nicolás Bellora, Robert Castelo, Lluis Armengol, Xavier Estivill, M. Mar Albà

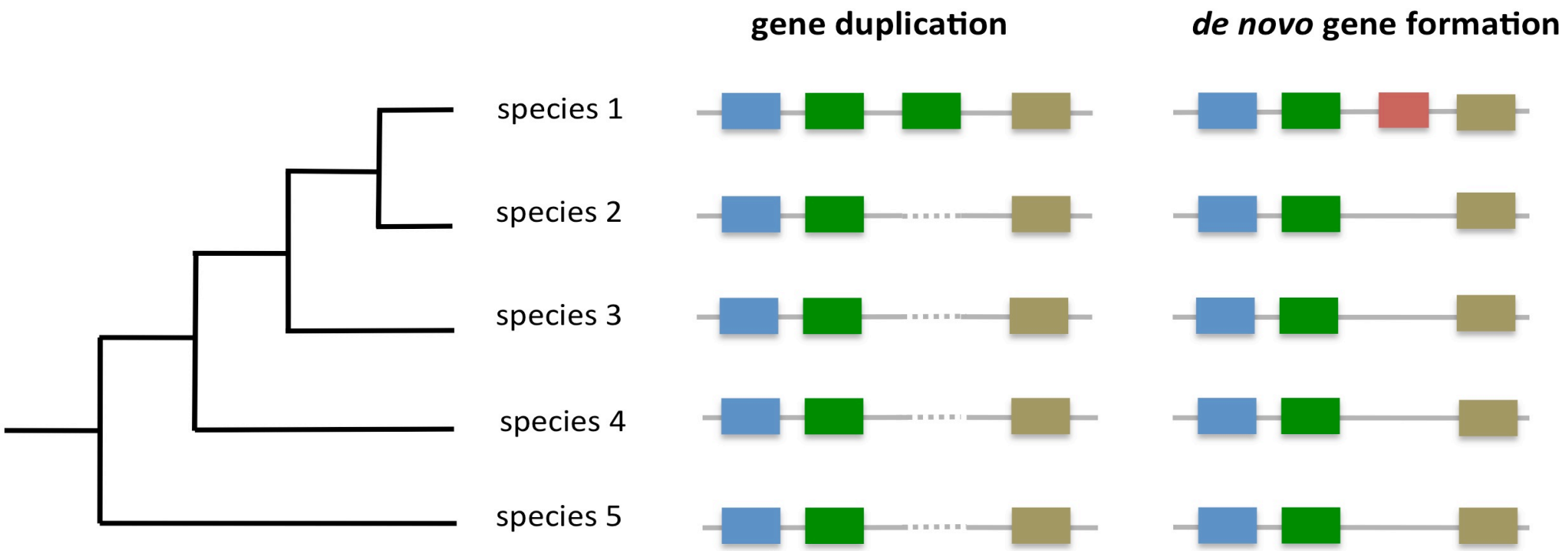
Molecular Biology and Evolution, Volume 26, Issue 3, 1 March 2009, Pages 603–612, https://doi.org/10.1093/molbev/msn281

De Novo Origination of a New Protein-Coding Gene in *Saccharomyces cerevisiae*

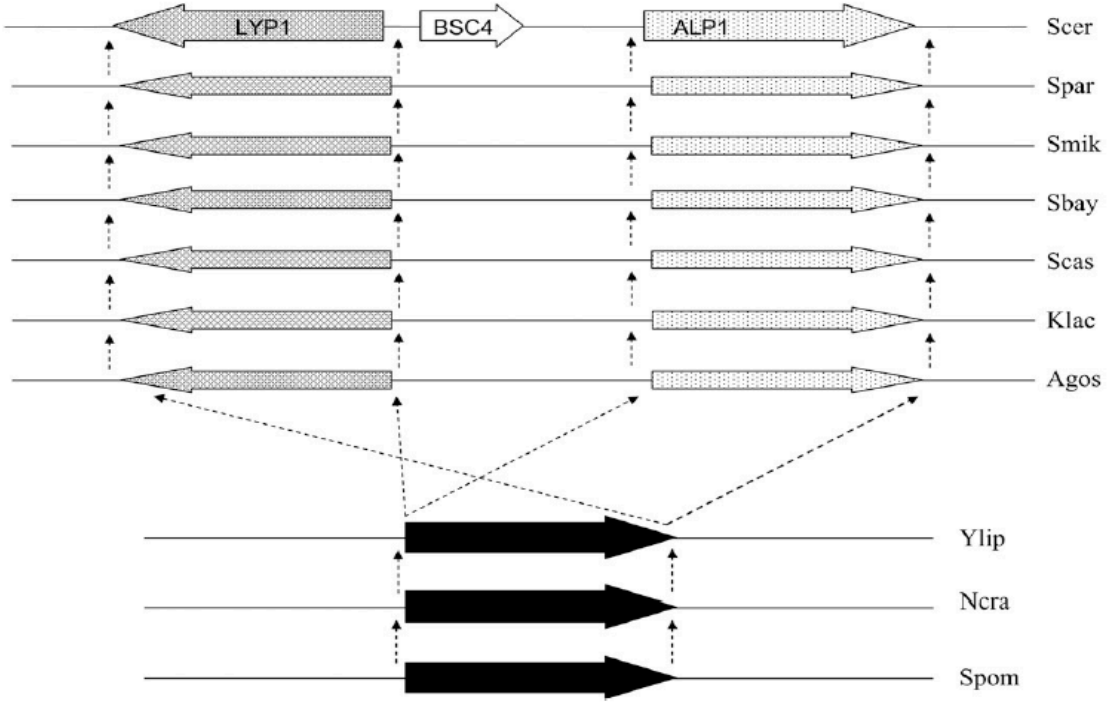
Jing Cai,*[,†,1] Ruoping Zhao,*[,1] Huifeng Jiang*[,†] and Wen Wang*[,2]

*CAS–Max Planck Junior Research Group on Evolutionary Genomics, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (CAS), Kunming, Yunnan 650223, China and [†]Graduate School of Chinese Academy of Sciences, Beijing 100049, China

Comparing genomes



One example: *BSC4* in *S.cerevisiae*



Cai et al., 2008

BSC4: 131 amino acid long protein involved in DNA repair

Human *de novo* genes involved in cancer

Gene	Product	Description
De novo *genes*		
CLLU1	CLLU1	This gene was first annotated from a screen for upregulated genes in CLL[123] and was one of the first identified human-specific *de novo* genes[4]
PART1	PART1	This primate-specific *de novo* gene has been implicated as a tumour suppressor gene[124]
MYCNOS (also known as *NCYM*)	Ncym	*De novo* human gene that stabilizes its antisense gene, the oncogene *MYCN*, in neuroblastomas[111]
PBOV1	PBOV1	*De novo* human gene associated with positive clinical outcomes in cancer[112]
GR6 (also known as *LINC01565*)	GR6	*De novo* human- and chimpanzee-specific gene that is normally expressed early in fetal development. Ectopic expression is associated with leukaemia[13,125]

McLysaght & Hurst, 2016

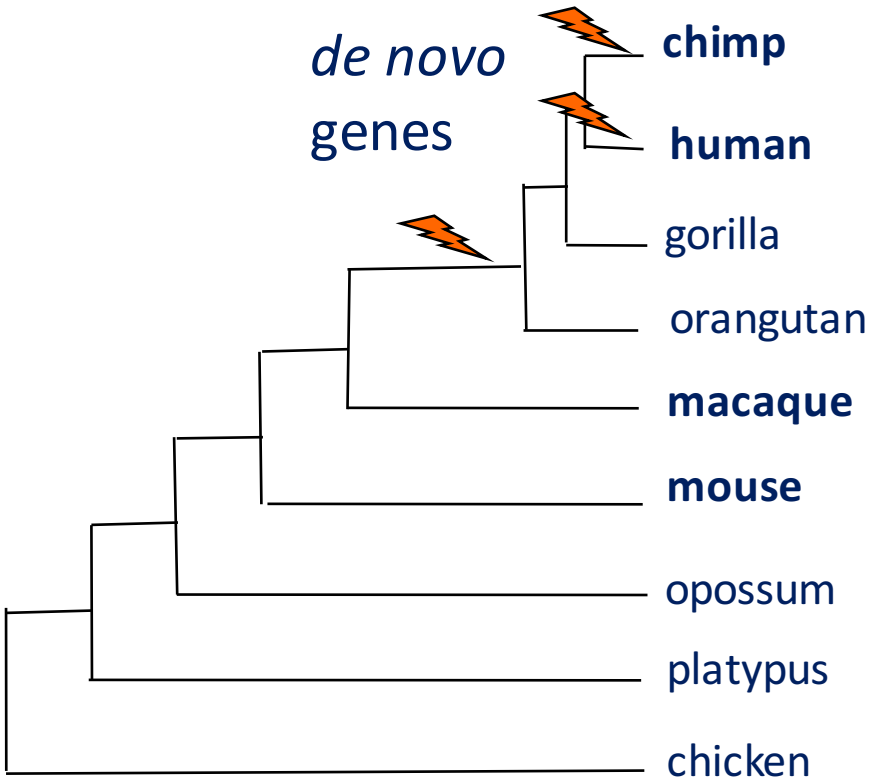
Steps in the formation of *de novo* genes

1. **Gain of transcription**
2. **Gain of translation**
3. **Gain of protein function**

Steps in the formation of *de novo* genes

1. **Gain of transcription**
2. **Gain of translation**
3. **Gain of protein function**

Comparison of transcriptomes



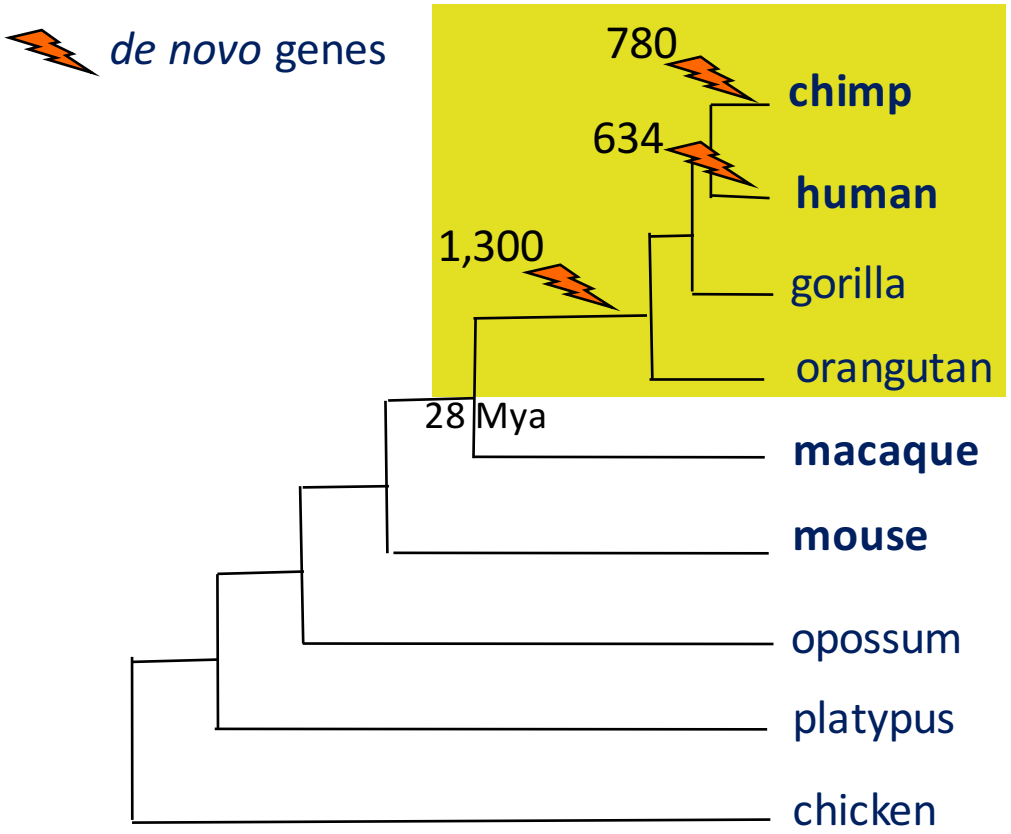
Deep transcriptome sequencing

RNA-Seq
polyA+ selection
strand-specific
100 nt x 2

≥2 ind., 4 tissues (testis, brain, heart, liver)
≈ 140 Mreads/sample

Ruiz-Orera et al., 2015 (Plos Genetics)

Comparison of transcriptomes and genomes



Deep transcriptome sequencing

RNA-Seq
polyA+ selection
strand-specific
100 nt x 2

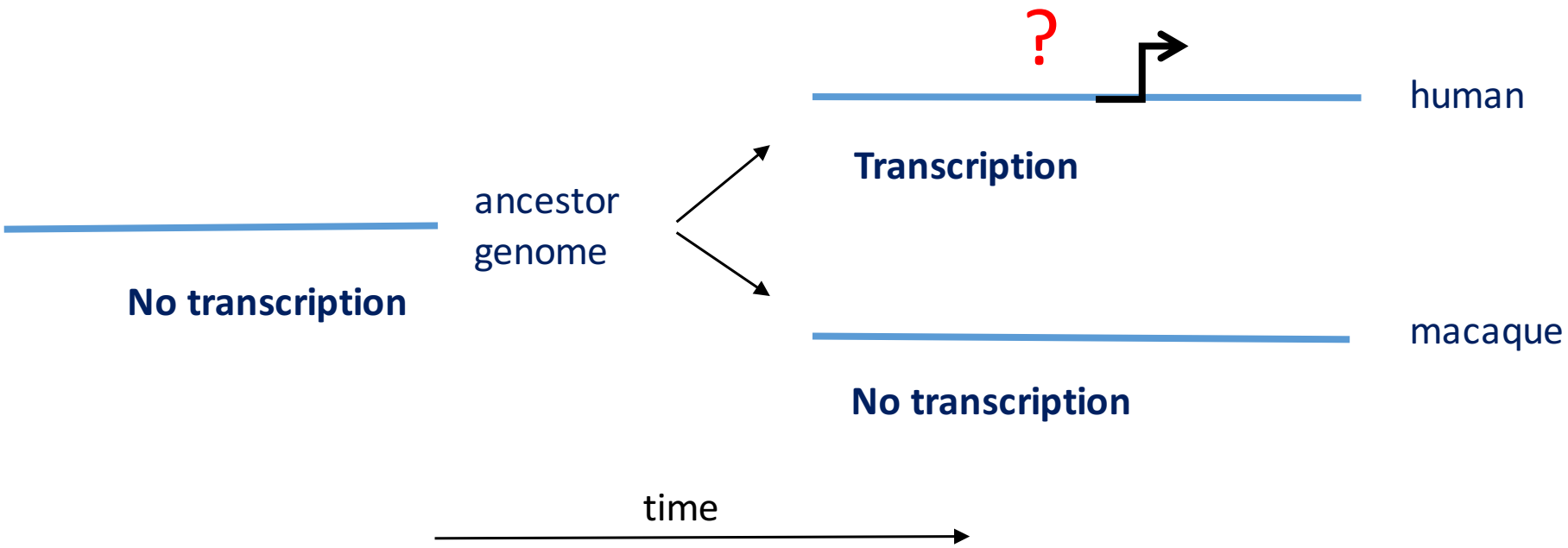
≥2 ind., 4 tissues (testis, brain, heart, liver)
≈ 140 Mreads/sample

Ruiz-Orera et al., 2015 (Plos Genetics)

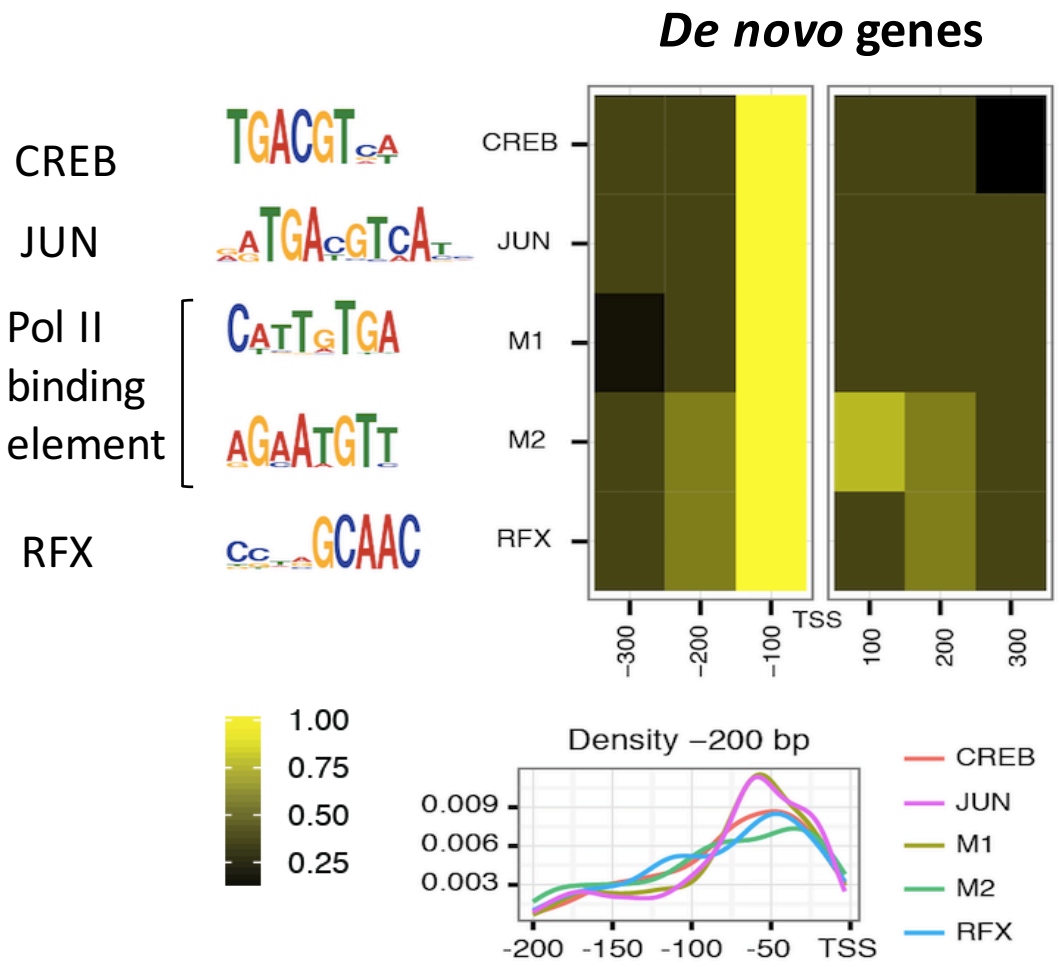
Some facts about human *de novo* genes

- **We found 1,934 putative *de novo* genes (about 5% of the human genes)**
- **The majority of them were not annotated in the databases (novel/lncRNAs)**
- **They were shorter than conserved genes**
- **They were expressed at lower levels than conserved genes**
- **Many were preferentially expressed in testis**

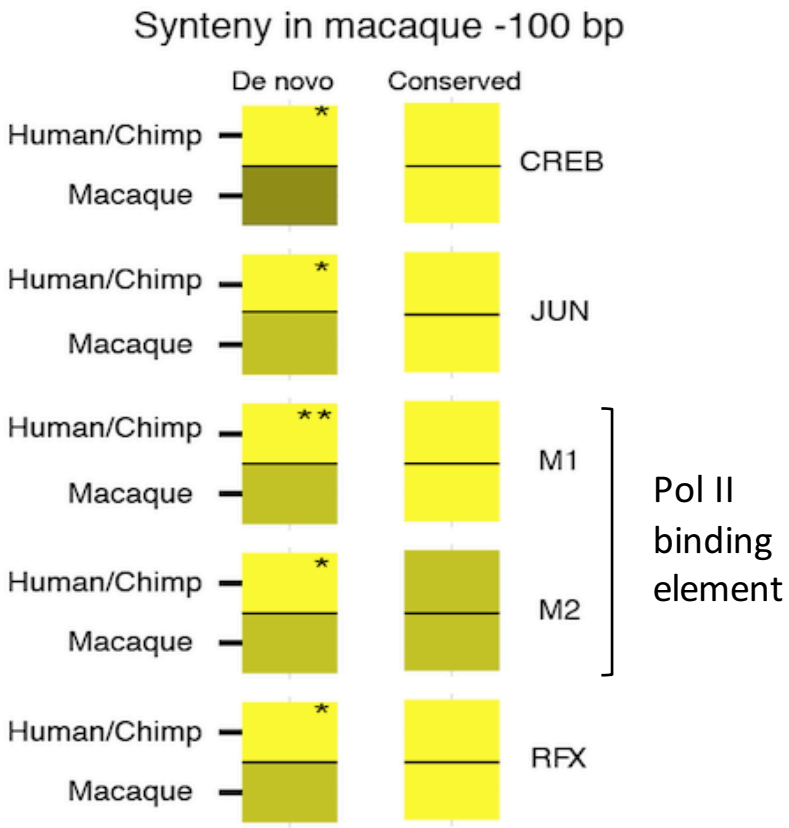
How is transcription initiated?



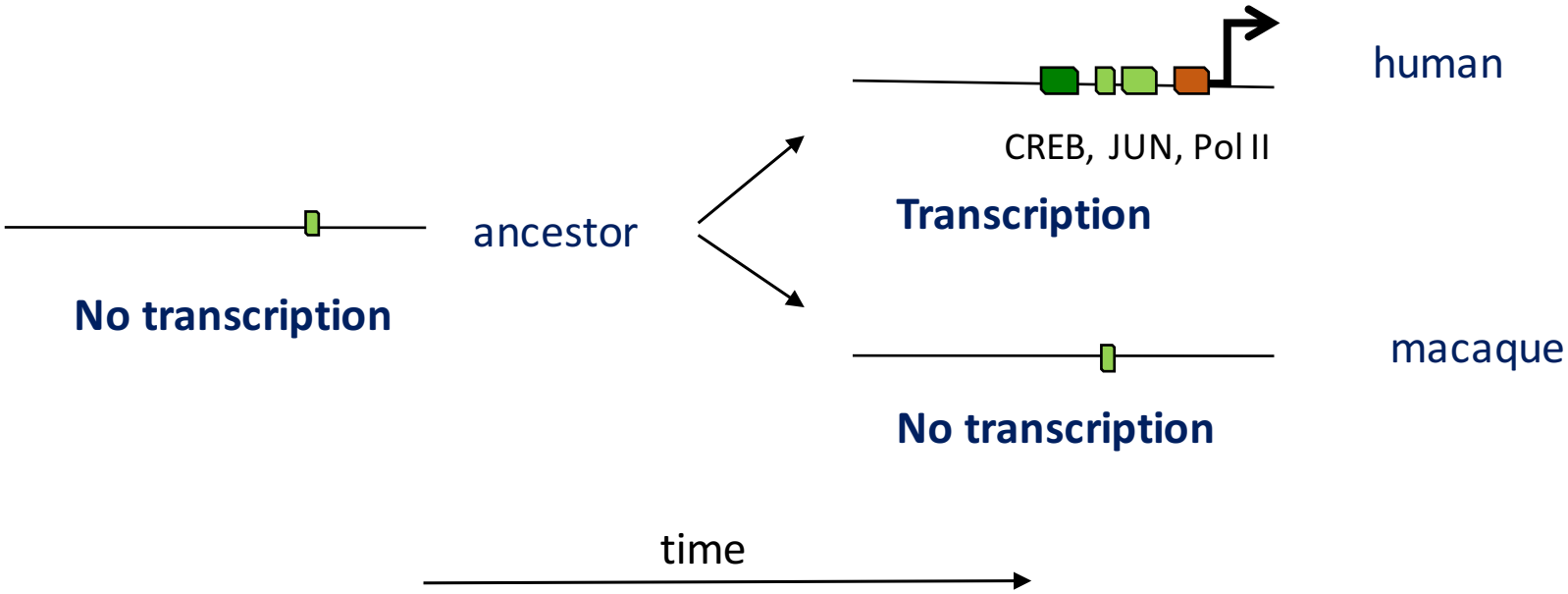
Five motifs are enriched in the promoters of *de novo* genes



The motifs are less abundant in macaque



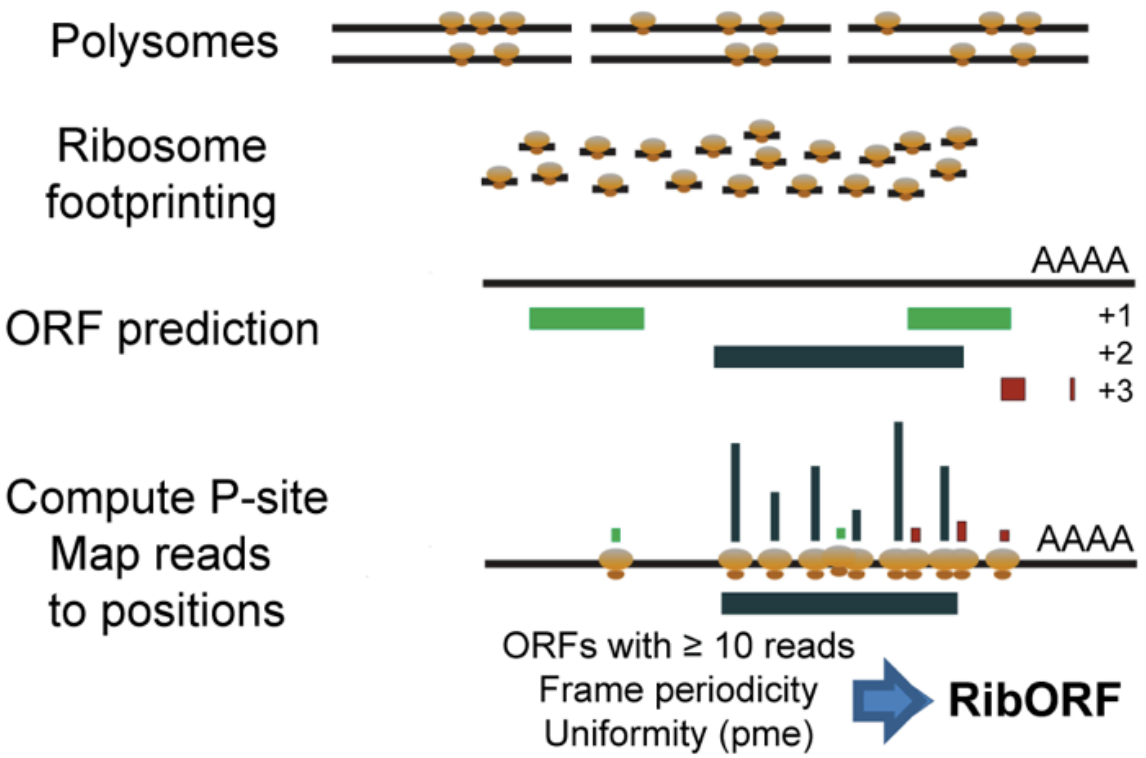
Transcription of *de novo* genes is associated with the formation of novel promoters



Steps in the formation of *de novo* genes

1. **Gain of transcription**
2. **Gain of translation**
3. **Gain of protein function**

Ribosome profiling (Ribo-Seq)



Many lncRNAs are associated with ribosomes

Table 2. Fraction of transcripts associated with ribosomes

	codRNA			lncRNA		
	Expressed	Associated with ribosomes (RP)		Expressed	Associated with ribosomes (RP)	
		Total	Stringent		Total	Stringent
Mouse	14,245	14,196 (99.7%)	13,918 (97.7%)	476	390 (81.9%)	367 (77.1%)
Human	17,011	16,630 (97.8%)	16,617 (97.7%)	934	403 (43.1%)	343 (36.7%)
Zebrafish	12,595	11,643 (92.4%)	11,637 (92.4%)	2392	726 (30.4%)	684 (28.6%)
Fruit fly	8041	8031 (99.9%)	7623 (94.8%)	28	22 (78.6%)	10 (35.7%)
Arabidopsis	19,162	18,879 (98.5%)	10,329 (53.9%)	139	93 (66.9%)	68 (48.9%)
Yeast	4740	4547 (95.9%)	4335 (91.5%)	21	6 (28.6%)	6 (28.6%)

Stringent: number of transcripts significant at $p < 0.05$ using 3′UTRs as a null model (see 'Materials and methods' for more details).

DOI: 10.7554/eLife.03523.004

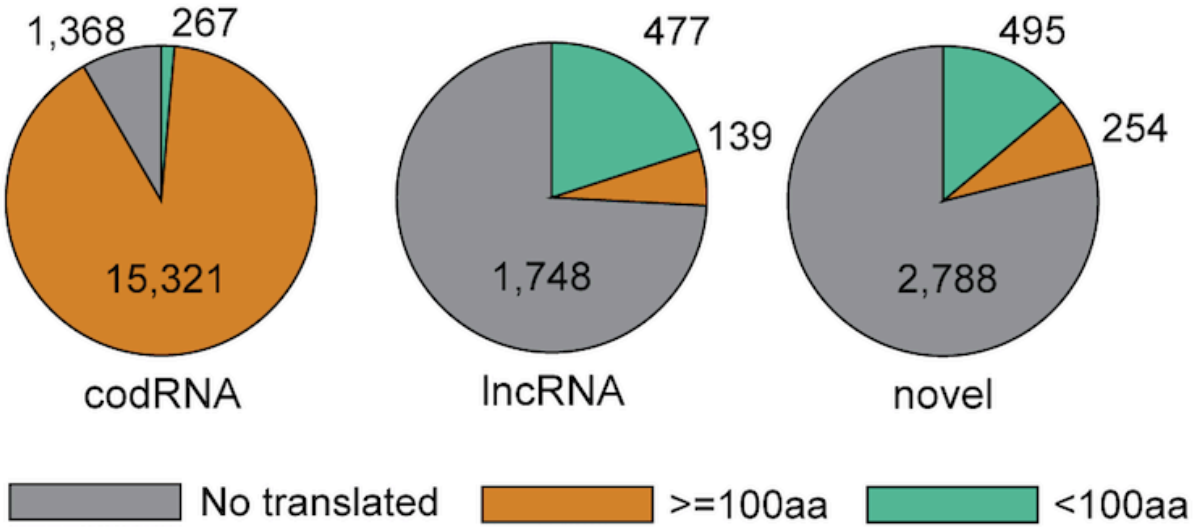
Ruiz-Orera et al., 2014 (eLife)

Analysis of the mouse transcriptome

Mouse Ribo-Seq data

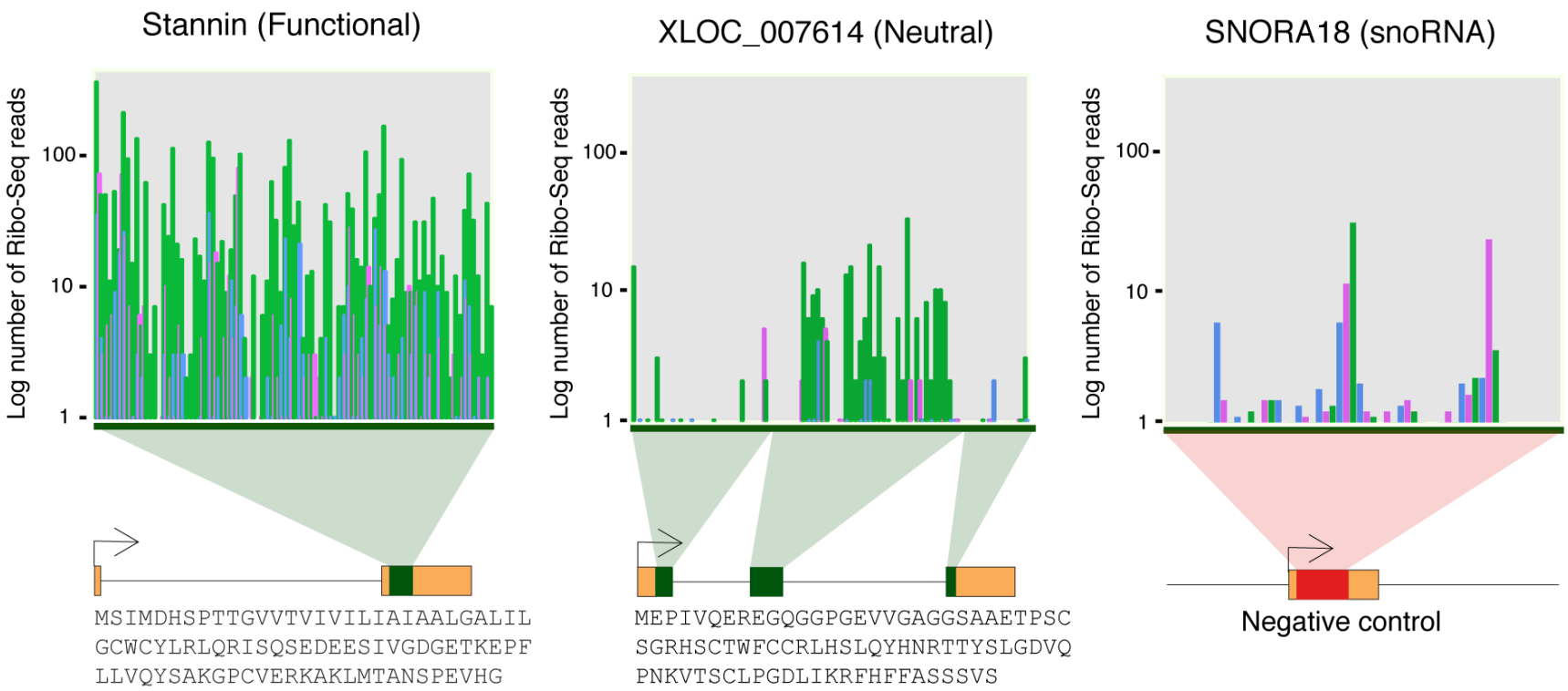
Brain
Testis
Neutrophils
Heart
Skeletal muscle
Splenic B cells
Neural ES cells
Hipoccampus

Transcripts with translated ORFs



Ruiz-Orera et al., 2018 (Nature Ecol & Evol)

Three nucleotide periodicity



Conservation of the translated ORFs

Mouse translated ORFs

↓

Comparison with human and rat transcriptomes,
and annotated proteomes from 101 species

↙ ↘

Mouse conserved ORFs (conserved, C)

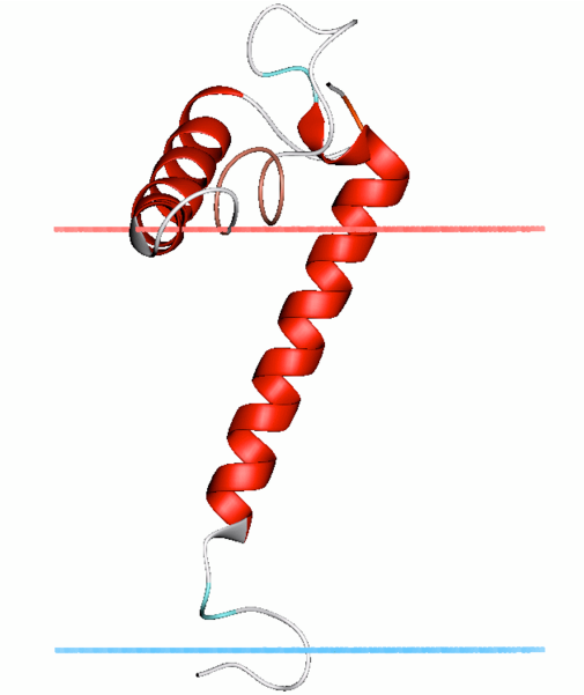
90%

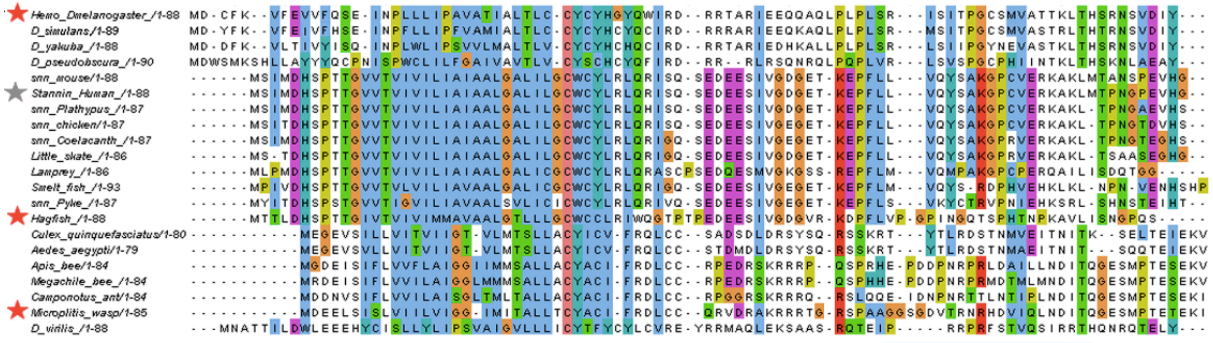
Mouse-specific ORFs (non-conserved, NC)

10%

Mouse conserved ORFs

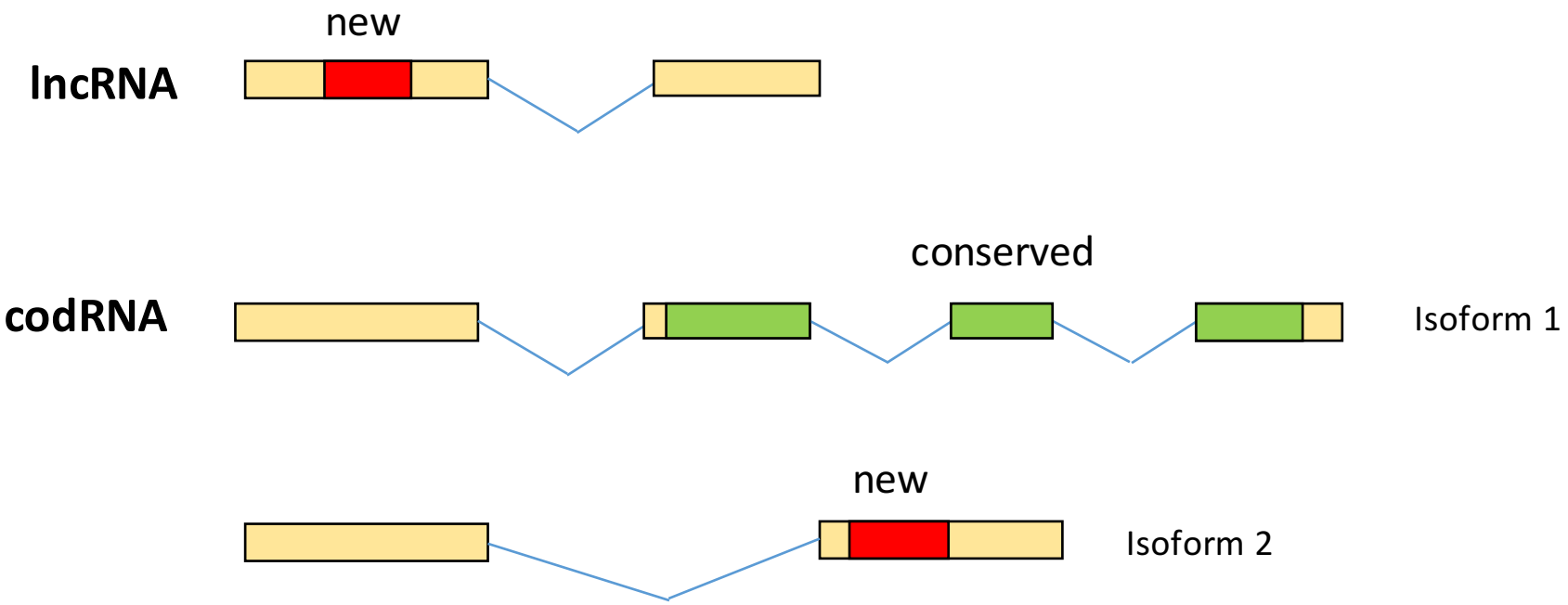
Example: stannin (88 amino acids)





Pueyo et al., 2016.

Mouse-specific ORFs



Identification of selection signatures

PN/PS: number of non-synonymous to synonymous polymorphisms

Calculated for sets of ORFs with evidence of translation by Ribo-Seq

PN/PS (Obs/Exp) < 1 purifying selection

PN/PS (Obs/Exp) = 1 no purifying selection

Coding score

Per hexamer:

$$CS_{hexamer(i)} = \log\left(freq_{coding}(hexamer(i)) \Big/ freq_{non-coding}(hexamer(i))\right)$$

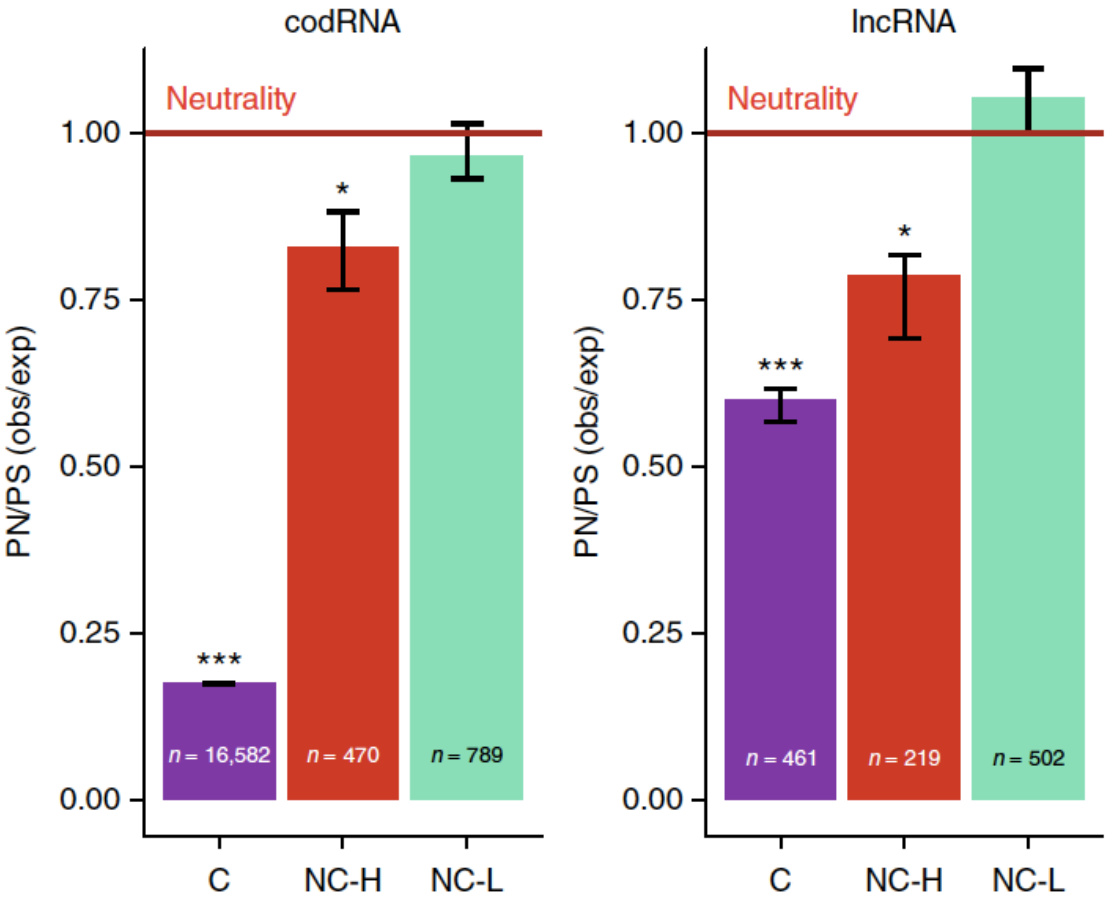
In a sequence:

$$CS_{ORF} = \frac{\sum_{i=1}^{i=n} CS_{hexamer(i)}}{n}$$

In coding sequences coding score is usually larger than 0 and in non-coding sequences it is lower than 0

De novo genes generally have low coding scores compared to other genes

Pervasive translation of ORFs



C: conserved transcripts

NC: non-conserved transcrips

NC-H

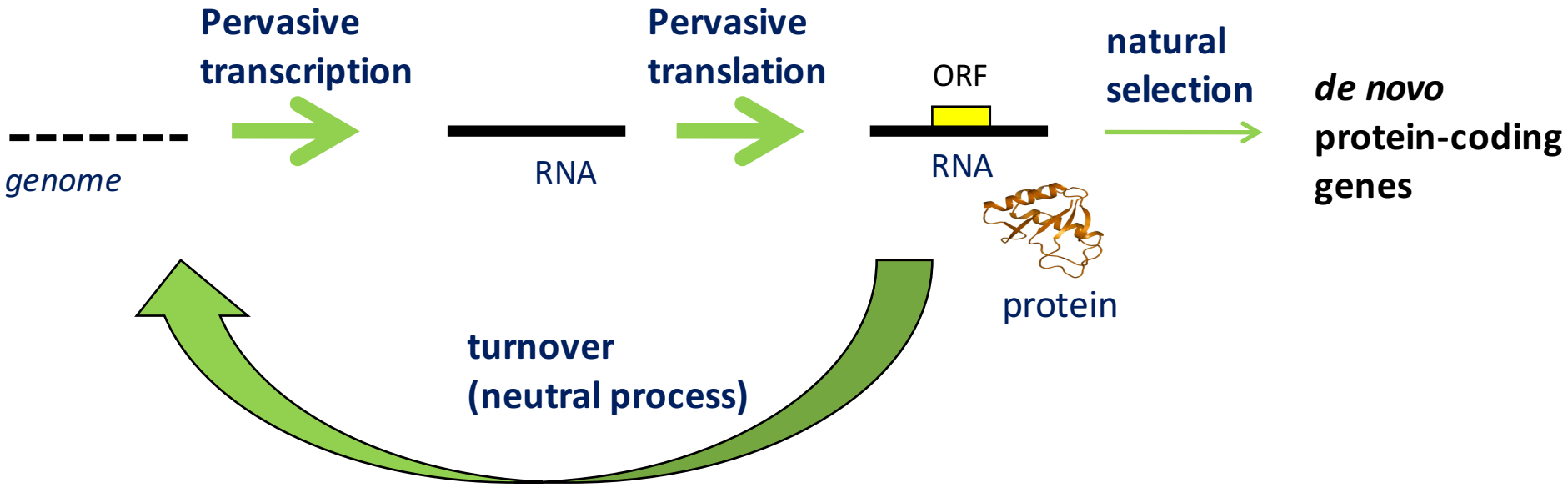
high coding score (> 0.1)

NC-L

low coding score

Ruiz-Orera et al., 2018 (Nature Ecol & Evol)

The life cycle of genes



Steps in the formation of *de novo* genes

1. **Gain of transcription**
2. **Gain of translation**
3. **Gain of protein function**

Can “random” proteins be functional?

François Jacob (1977, Evolution and Tinkering)
“The probability that a functional protein would appear *de novo* by random association of amino acids is practically zero.”

Can “random” proteins be functional?

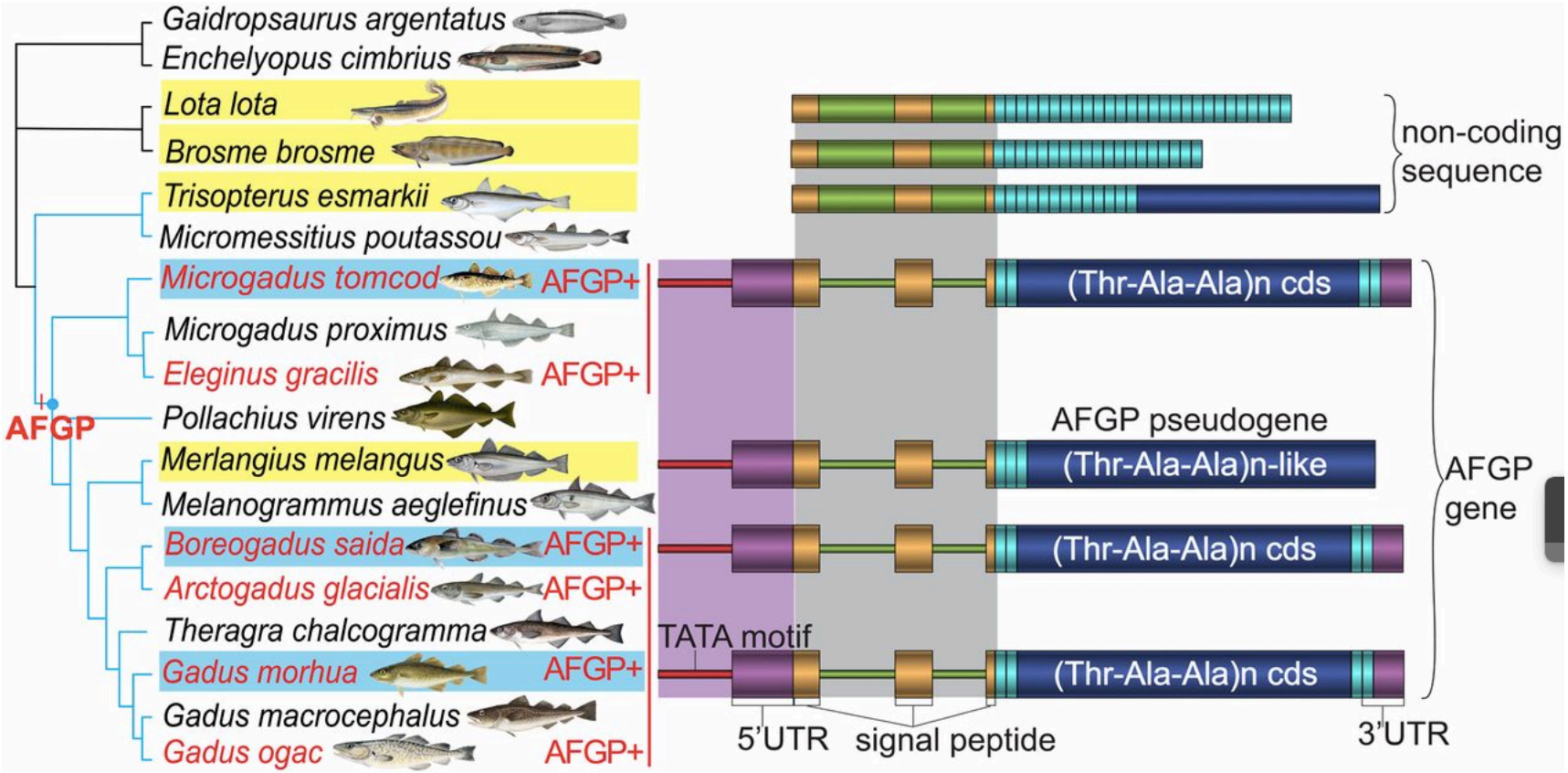
François Jacob (1977, Evolution and Tinkering)
“The probability that a functional protein would appear *de novo* by random association of amino acids is practically zero.”

But..

Keefe & Szostak (2001 Nature). ATP-binding proteins obtained from a random sequence library.

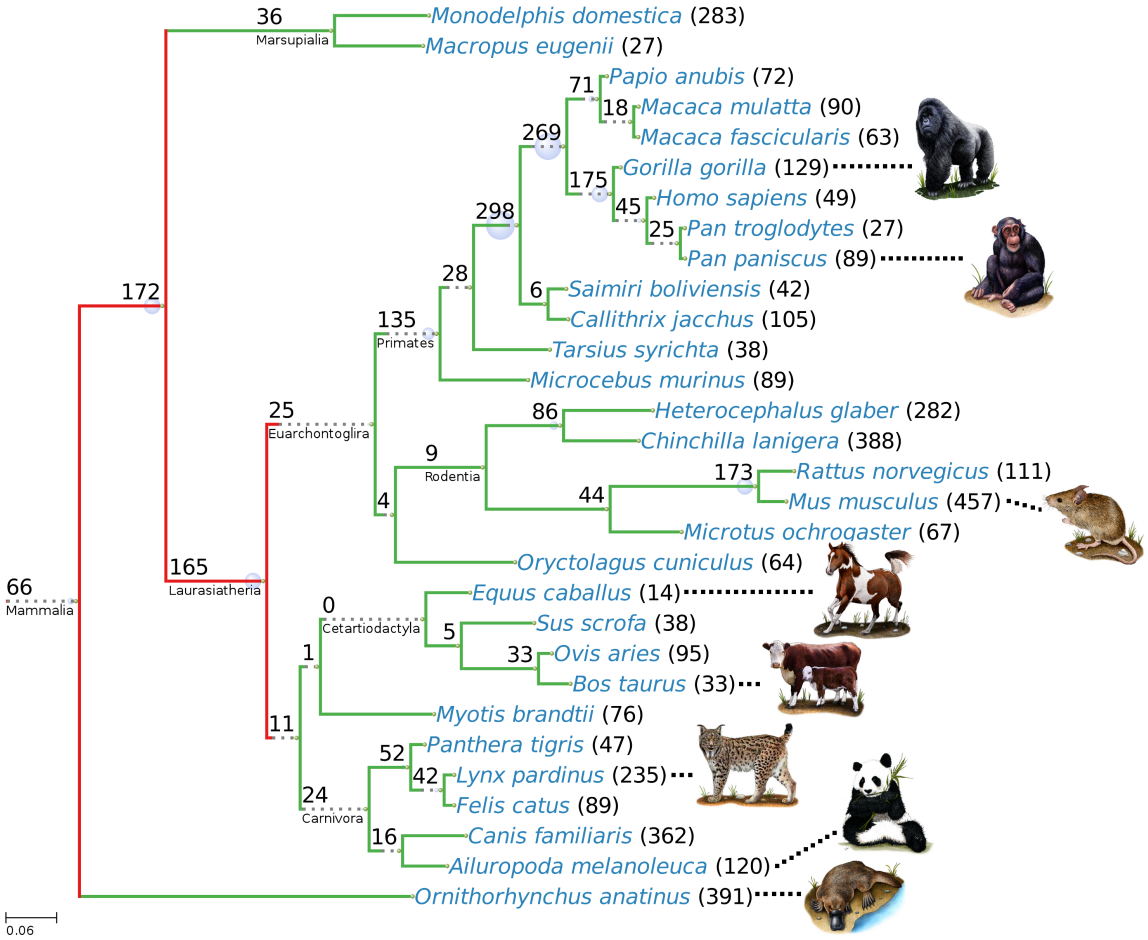
Neme et al. (2017, Nature Ecol. Evol.). Random sequences are an abundant source of bioactive RNAs or peptides

Antifreeze glycoproteins in codfishes



Zhuang et al., 2007

Identification of mammalian-specific genes



Villanueva-Cañas et al., 2017 (Genome Biol & Evol)

Young proteins have specific properties

Functions of mammalian-specific genes

Main Functions of Mammalian-Specific Genes

Enriched Function	Representative Terms	*N* Genes	Corrected *P* Value
1. Immune response	1.1 immune response (GO)	14	2.2E-3
	1.2 cytokine activity (GO)	13	1.6E-10
	1.3 Jak-STAT signaling pathway (KEGG)	6	5.5E-4
2. Reproduction	2.1 reproductive process in a multicellular organism (GO)	12	1.3E-3
	2.2 spermatogenesis (GO)	10	9.2E-4
3. Secreted protein	3.1 extracellular region (GO)	64	1.8E-15
	3.2 secreted (Uniprot)	59	2.8E-14
	3.3 signal peptide (Uniprot)	60	7.0E-10

Conclusions

- Pervasive transcription and translation can generate abundant raw material for *de novo* gene birth

- *De novo* generated proteins contribute to evolutionary innovation, most of them remain uncharacterized.

Acknowledgements

Evolutionary Genomics Group
Macarena Toll-Riera
José Luis Villanueva Cañas
Jorge Ruiz Orera
Will Blevins

Collaborators
Xavier Messeguer (UPC)
Pol Verdaguer-Grau (UPC)
Juan Antonio Subirana (RACAB)
Jéssica Hernández (UPF)
Tomàs Marqués-Bonet (UPF)
Cristina Chiva (CRG-UPF)
Eduard Sabidó (CRG-UPF)