Prediction and evolution of selenoproteins in eukaryotic genomes

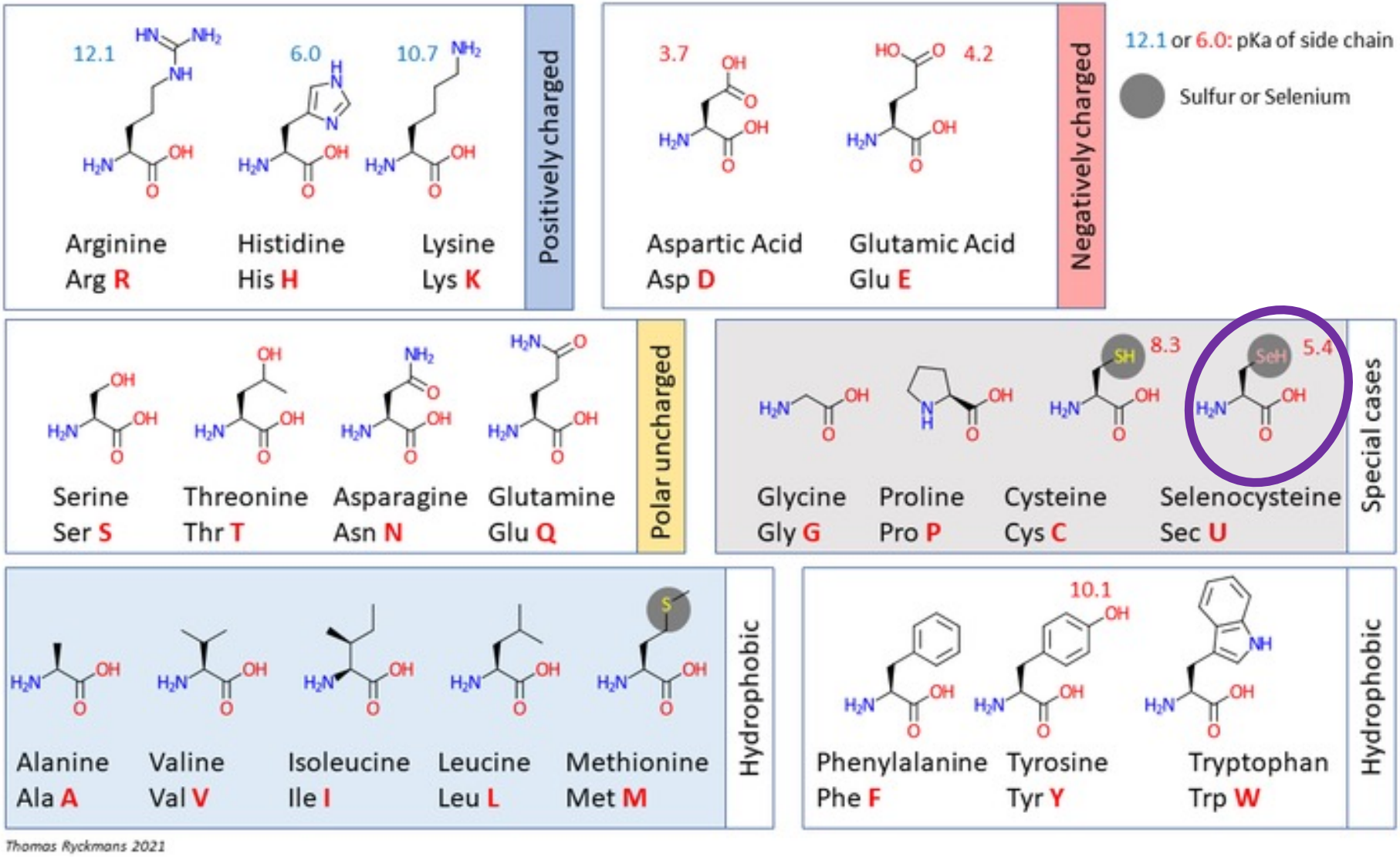
Bioinformatics,
Universitat Pompeu Fabra

Marco Mariotti
Universitat de Barcelona
www.mariottigenomicslab.com

What are selenoproteins?

Selenoproteins are proteins that incorporate selenocysteine, the 21st aminoacid

Selenocysteine (Sec or U)



Thomas Ryckmans 2021

Role of selenium

- Selenium is an essential nutrient for vertebrates, many other animals, and microorganisms

- Selenium deficiency leads to disease

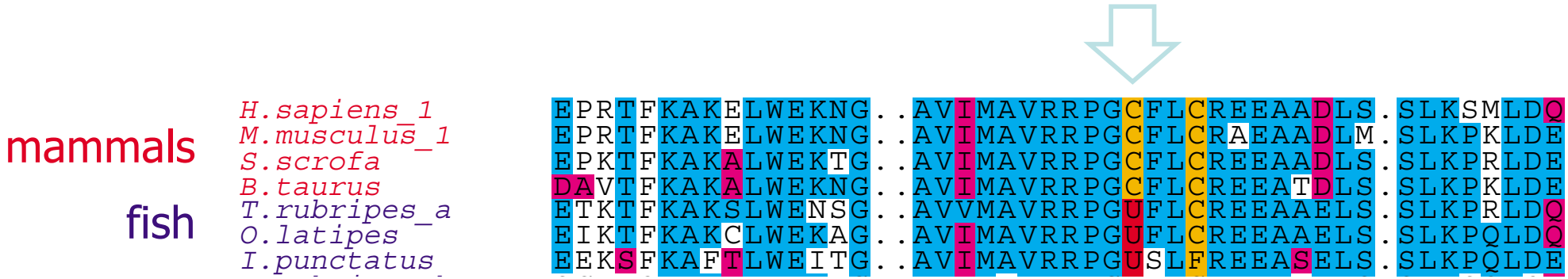
  Keshan disease (myocardial necrosis): named after the Keshan province in China, whose lands have low levels of selenium

- Excess selenium can be toxic

Selenium is found in cells mostly in selenoproteins

- About 25 selenoproteins in mammals
- Their number varies for different taxa
  - 3 selenoproteins in *Drosophila melanogaster*
  - 1 selenoprotein in *Caenorhabditis elegans*
- Typically oxidoreductase enzymes, with functions in redox homeostasis (e.g. antioxidant defense)
- Sometimes the orthologue of a selenoprotein has Cys instead of Sec.

SelenoU is a selenoprotein in fishes, but it is not in humans



What are selenoproteins?

Selenoproteins are proteins that incorporate selenocysteine, the 21st aminoacid

Why do we usually talk about 20 amino acids instead of 21?

The selenocysteine codon?

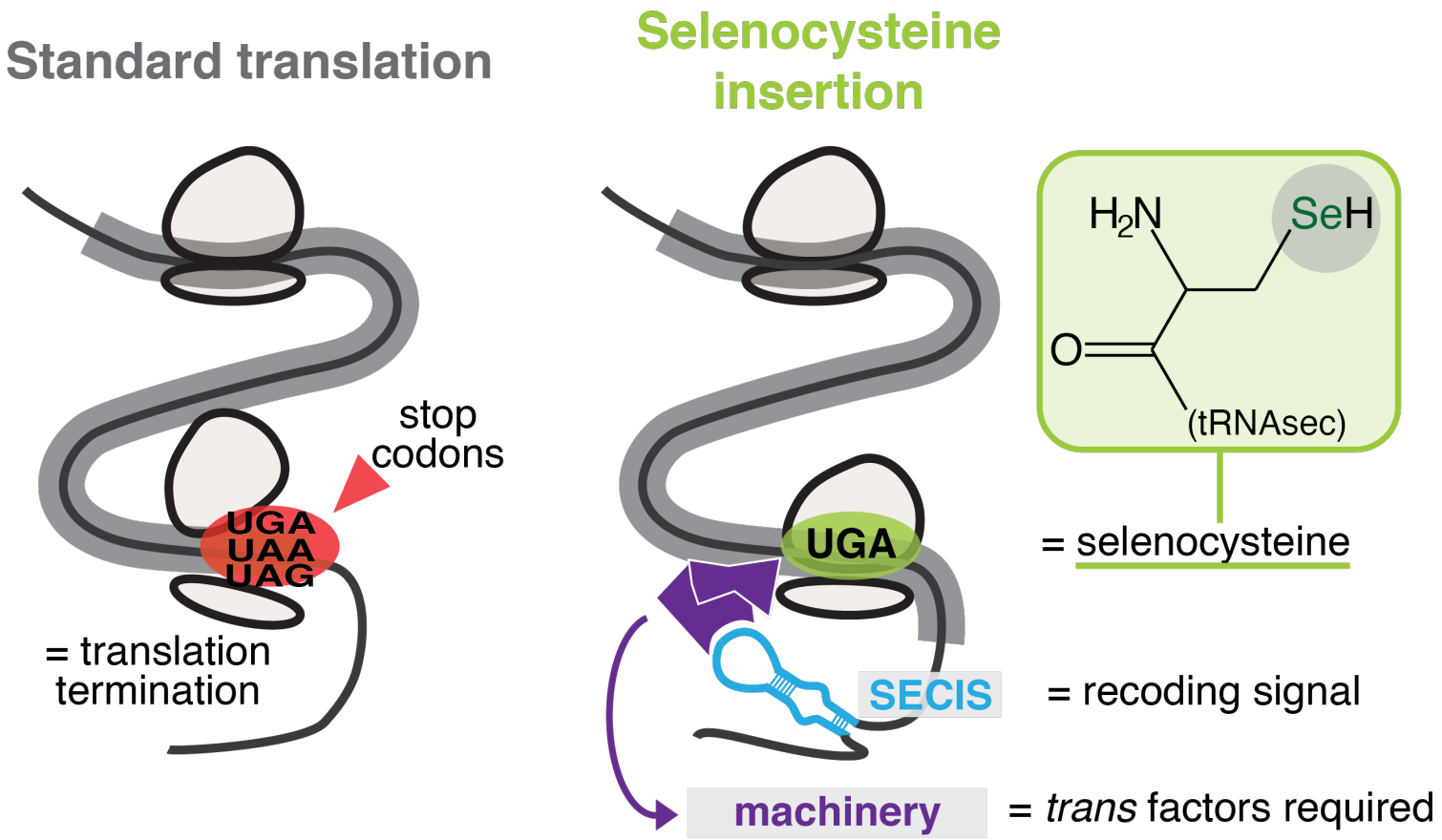
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

The selenocysteine codon = UGA

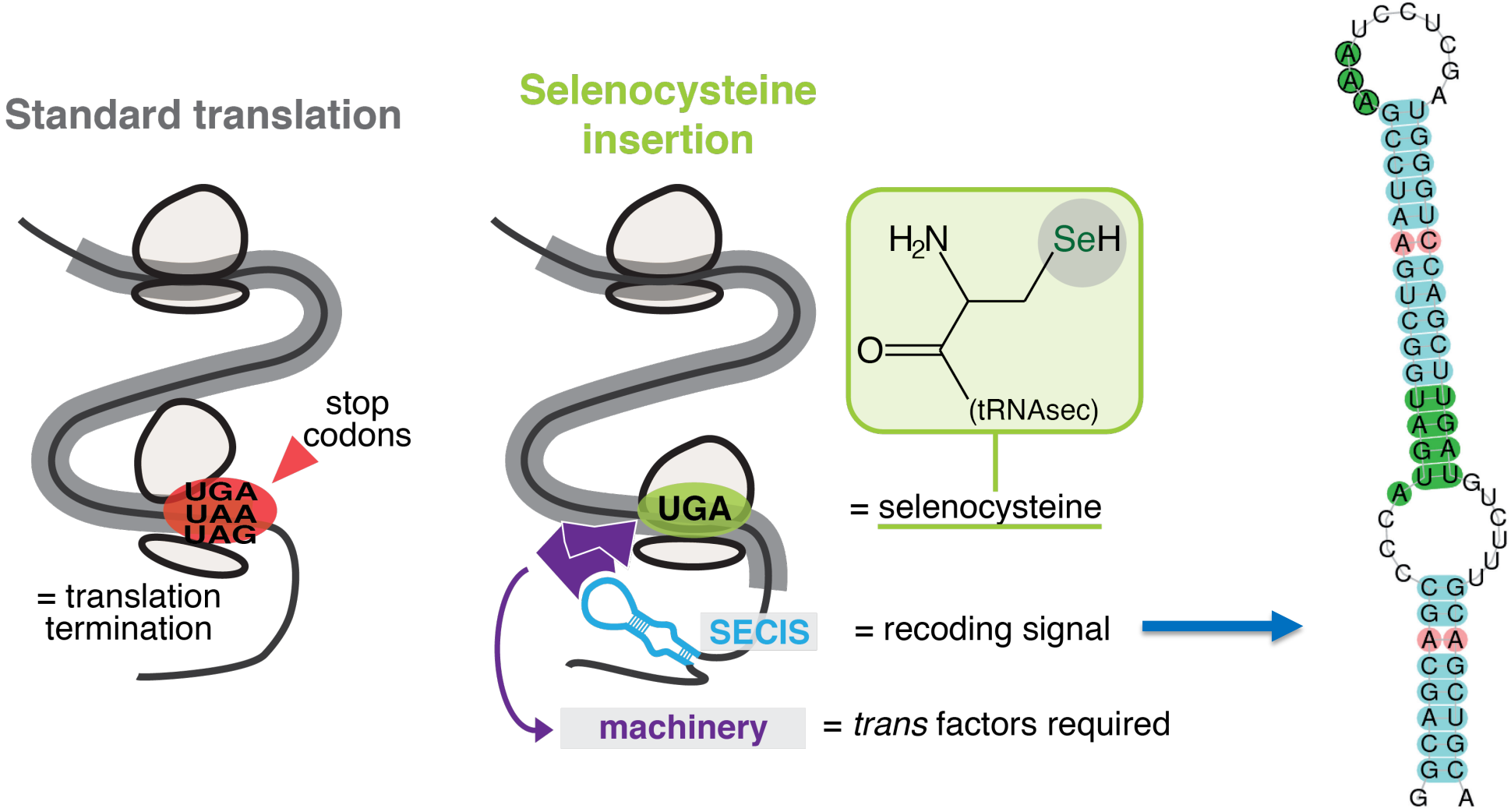
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop/Sec
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

But UGA is a stop codon in 99.9% of cases!

Recoding of UGA for selenocysteine insertion

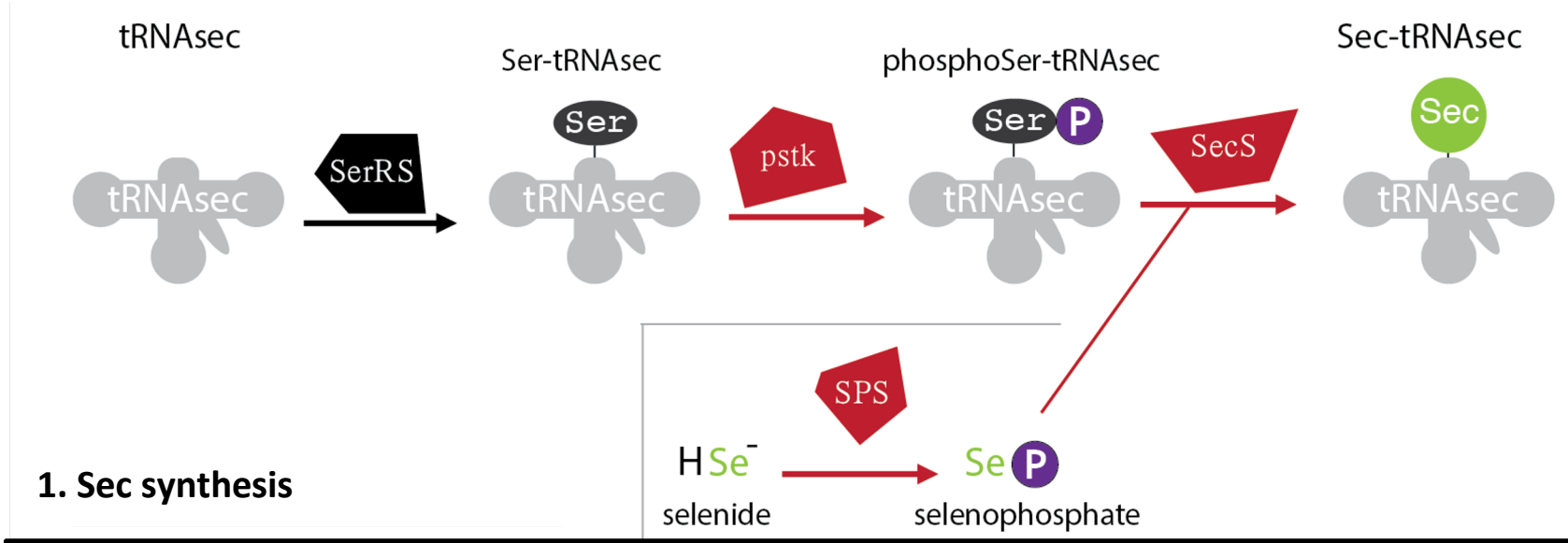


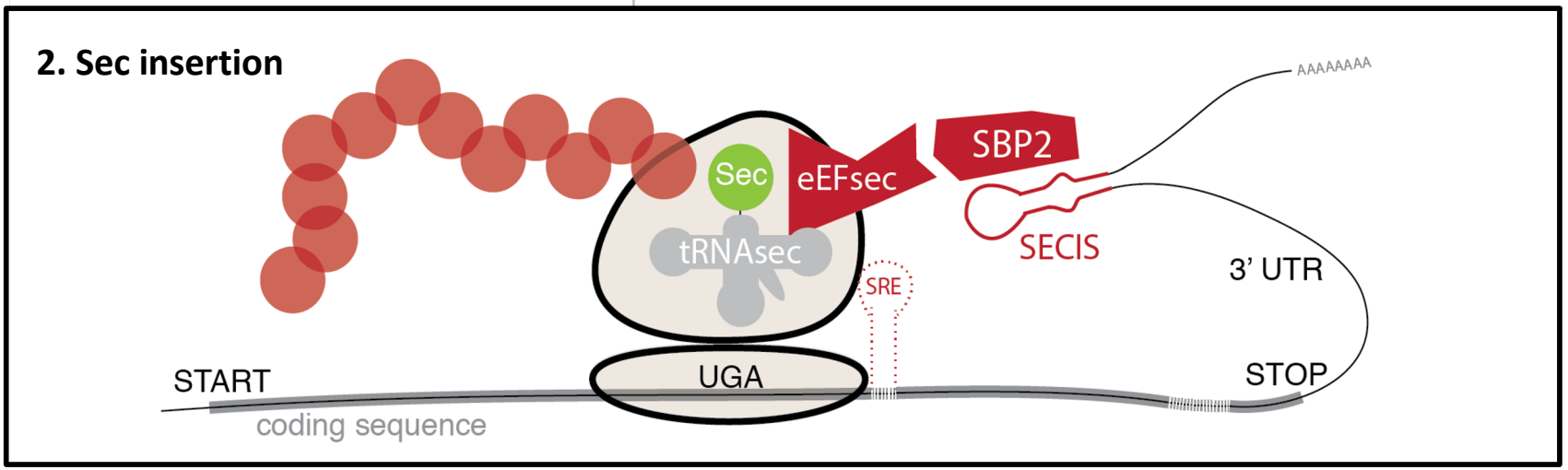
Recoding of UGA for selenocysteine insertion



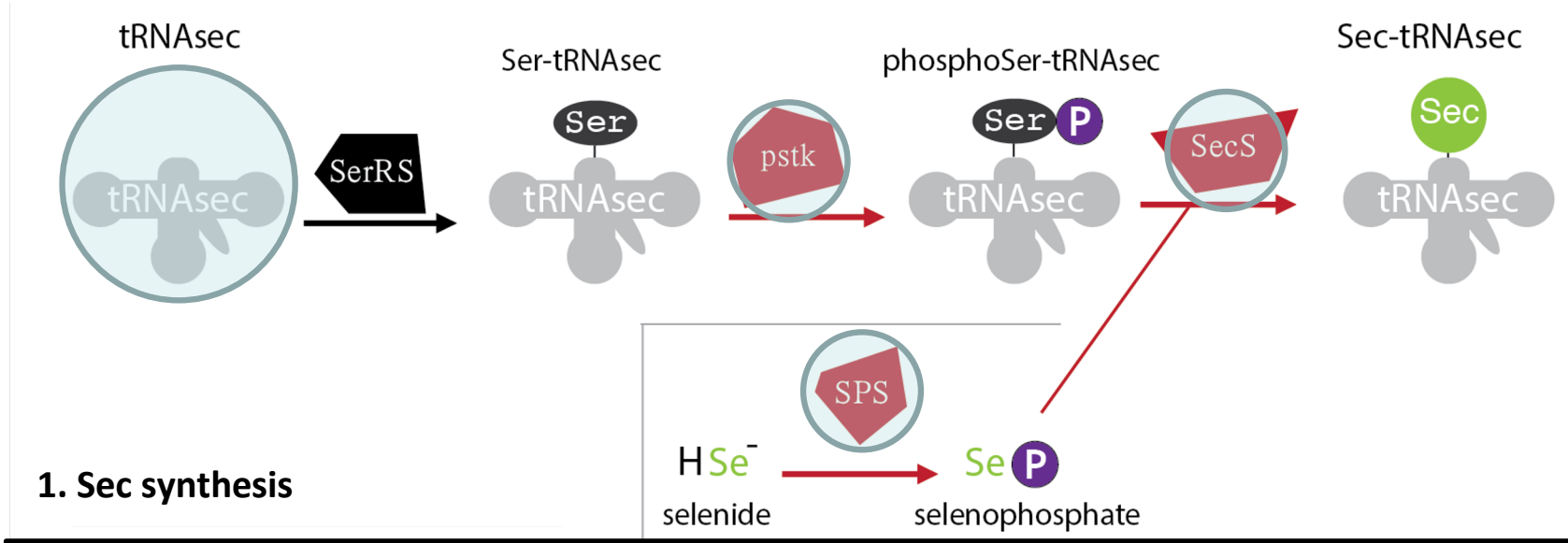
recoding signal in 3'UTR
(downstream of coding sequence)= SECIS element

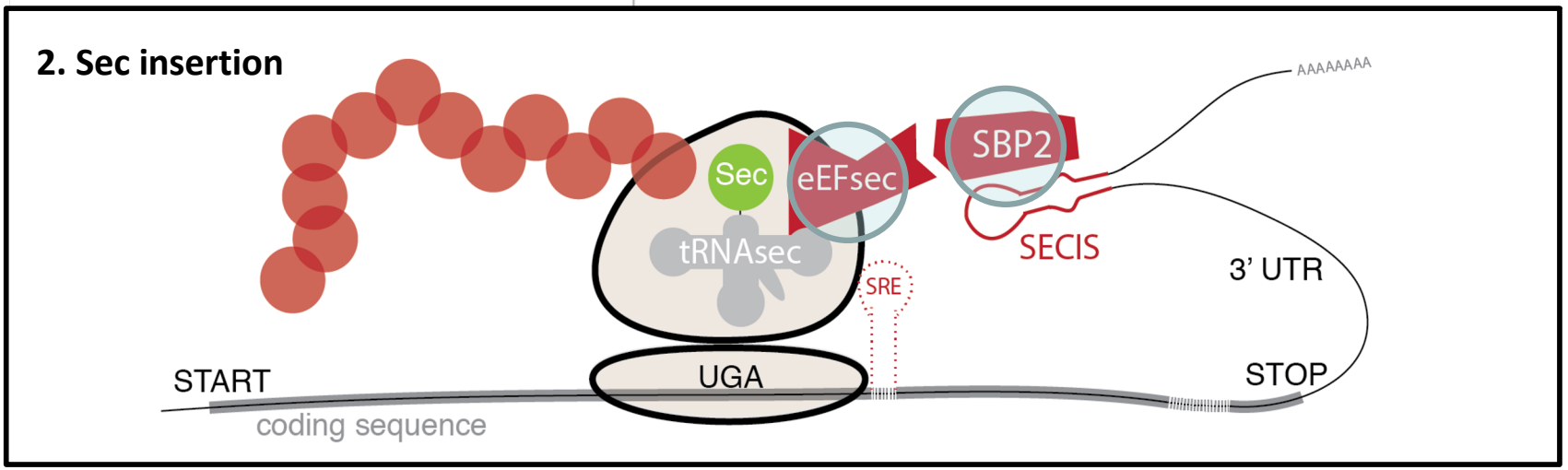
Pathway for selenocysteine biosynthesis and insertion





Pathway for selenocysteine biosynthesis and insertion





Selenocysteine machinery

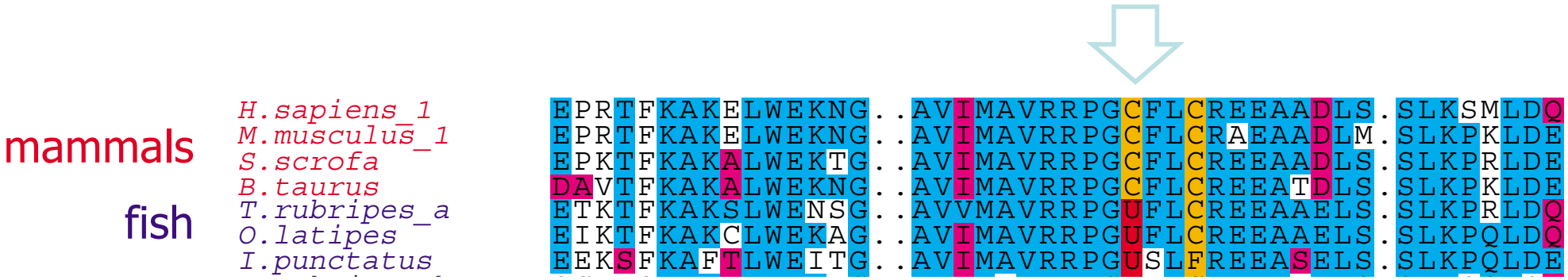
= factors required for selenoprotein synthesis

- Synthesis of selenocysteine
  - tRNAsec
  - SPS2 / Sephs2
  - SecS / SepSecS
  - Pstk
- Incorporation of Sec into selenoproteins
  - SBP2 / SECISBP2
  - EFsec / eEFsec

Why selenocysteine?

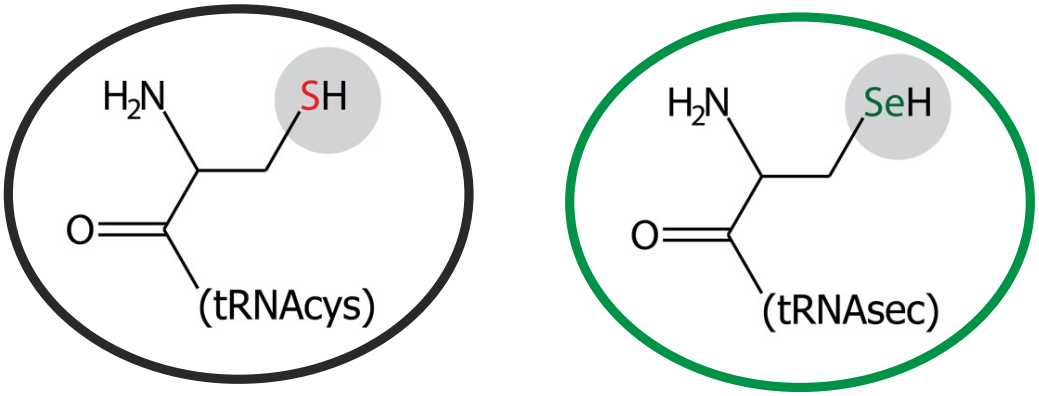
- It is metabolically expensive and complex to make!

- … also, it seems the same function can be achieved with cysteine instead of selenocysteine:

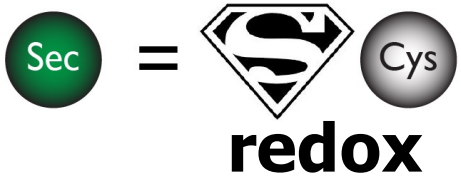


Why selenocysteine?

Cysteine vs **Selenocysteine**



provides enhanced biochemical properties for **redox reactions**



Selenocysteine is

- more reactive and
- more resistant to permanent oxidation

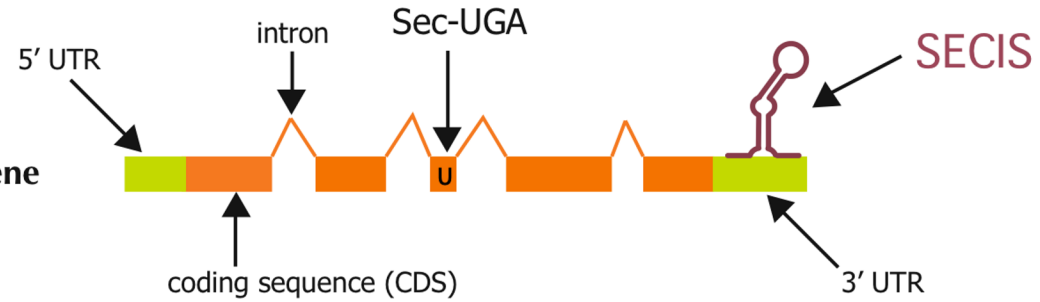
than **cysteine**

Computational identification of selenoproteins

The identification of selenocysteine has been compared to **finding a needle in a haystack**

Selenoproteins are usually incorrectly annotated

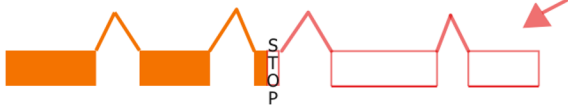
Example of selenoprotein gene



Typical coding sequence misannotations:

missing in the annotation

1. Sec-UGA is treated as STOP



27.1%

2. CDS starts downstream of Sec-UGA

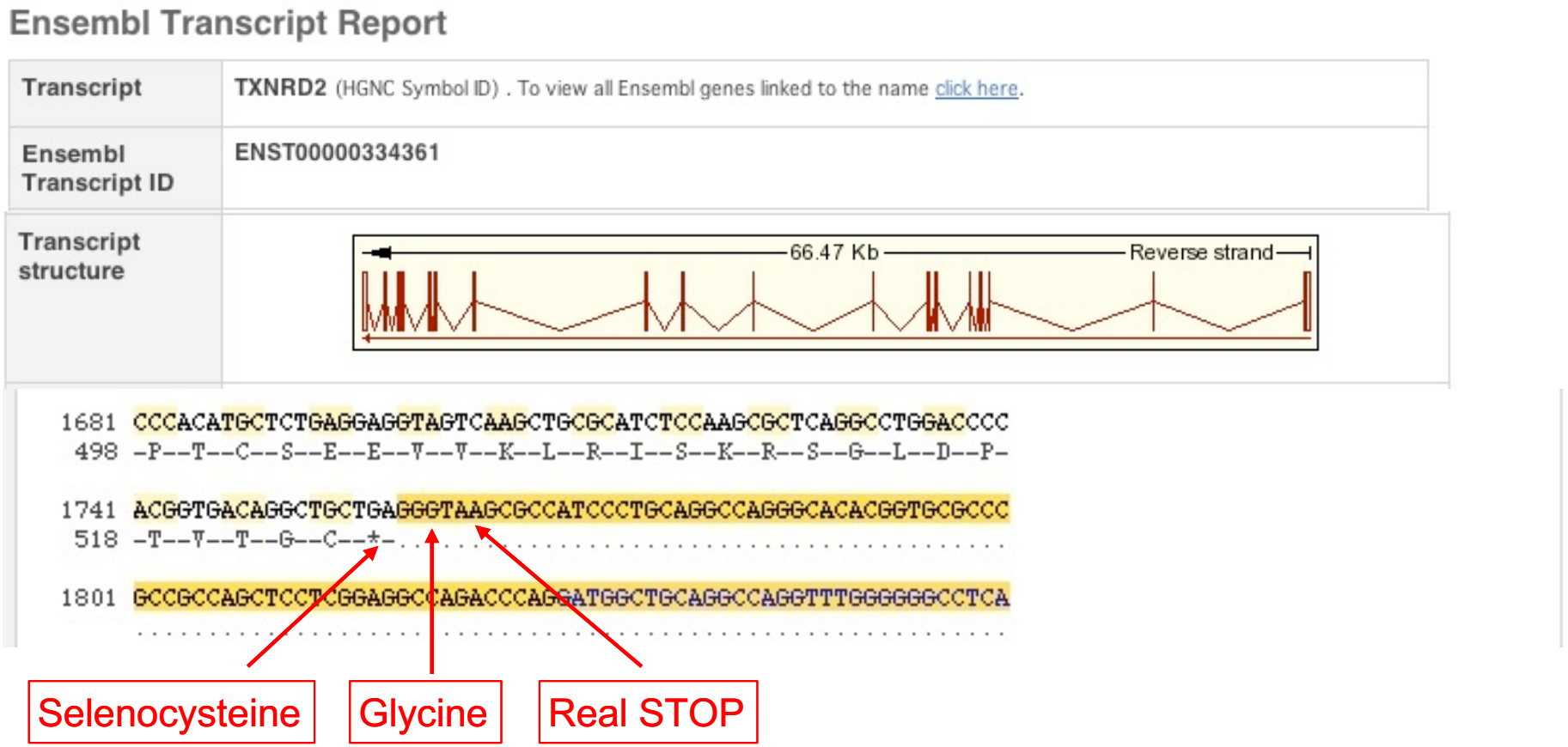
17.8%

3. Sec-UGA is skipped

42.5%

Selenoproteins are usually incorrectly annotated

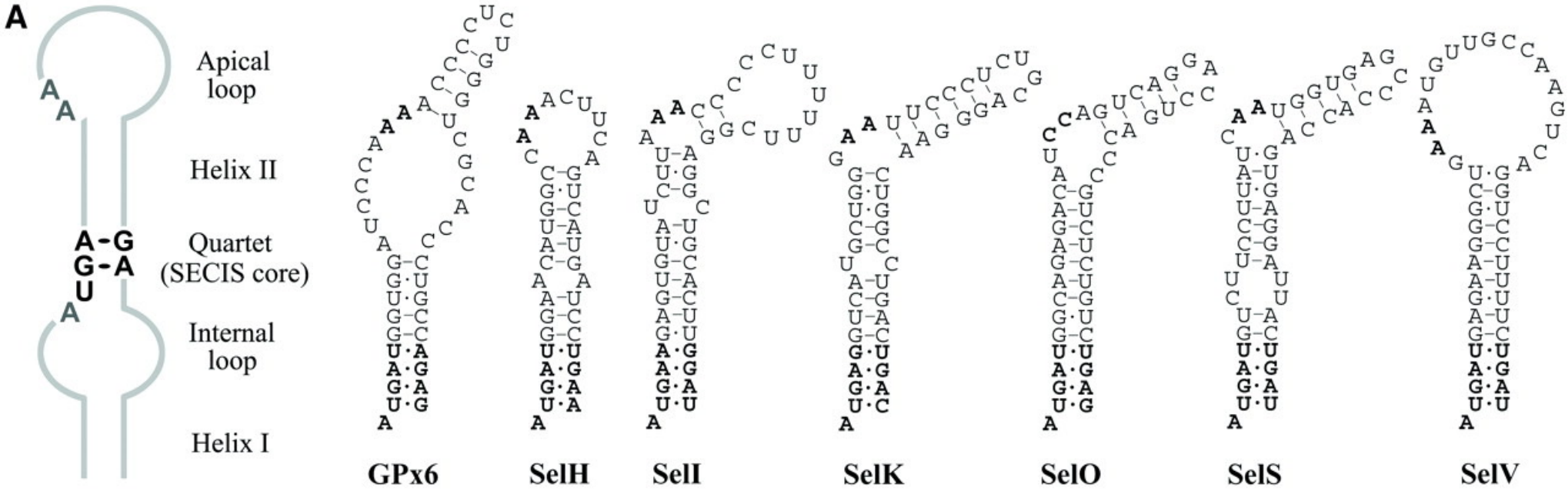
Selenocysteine is the penultimate residue in several selenoproteins



Selenoprotein identification

- SECIS prediction

Finding SECIS



Strategies:

- pattern-based: SECISearch v1
- Covariance model: SECISearch v3

https://seblastian.crg.es/ (Mariotti et al., 2013)

Selenoprotein finding in Drosophila

(Castellano et al. EMBO Reports 2:697-702, 2001)

Drosophila genome published in 2000

SECIS predicted	35876
SECIS thermo assessment	1220

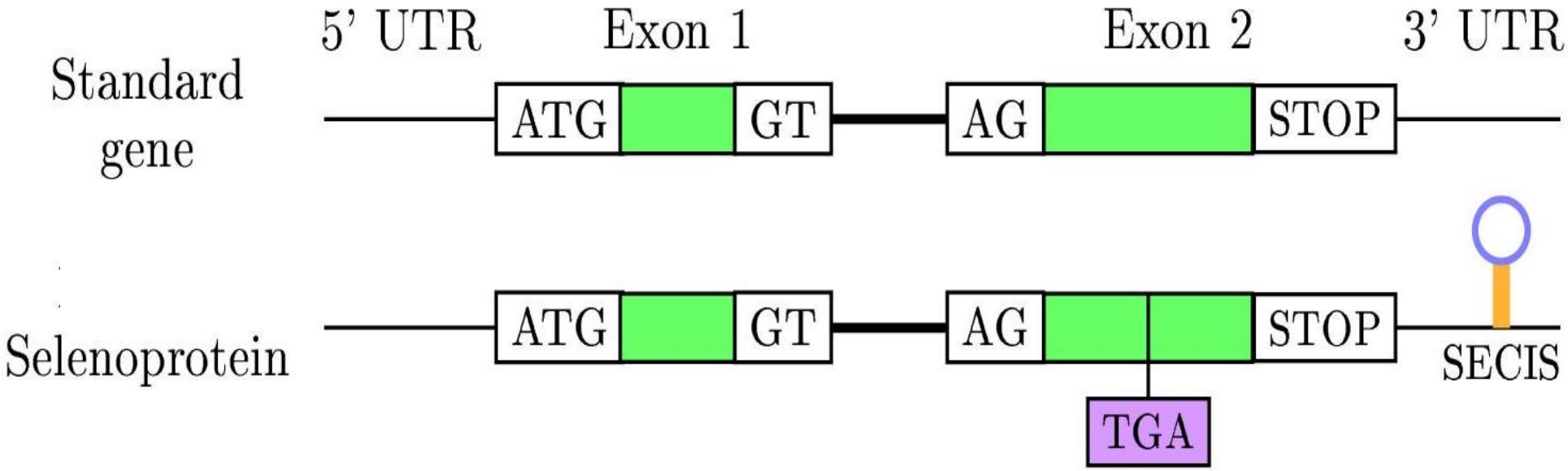
Too many False Positives!
Need some other criteria

Selenoprotein identification

- *De novo* prediction of coding sequences
- SECIS prediction

Selenoprotein search: codon bias across TGA

- Sequence "signals" to discriminate coding from non-coding
  - Build gene structures (made of coding exons)



Selenoprotein finding in Drosophila

(Castellano et al. EMBO Reports 2:697-702, 2001)

Drosophila genome published in 2000

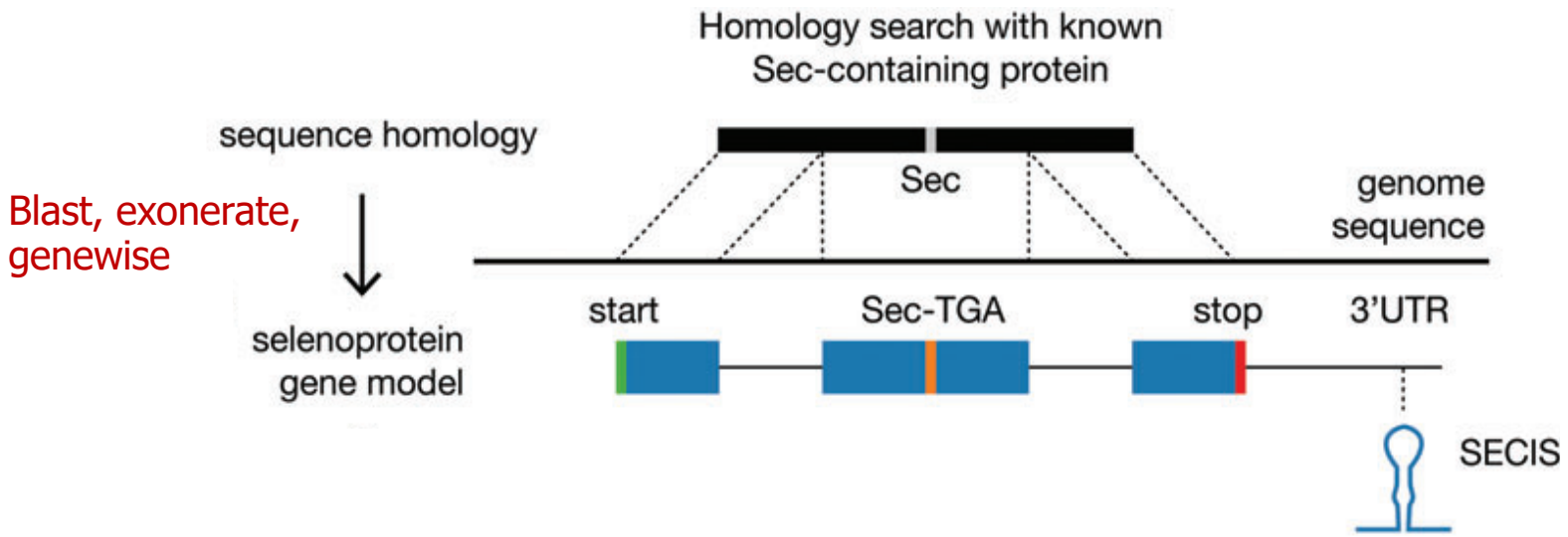
SECIS predicted	35876
SECIS thermo assessment	1220
Genes predicted	12194
Predicted Selenoproteins	(4)
Real Selenoproteins	3

Selenoprotein identification

- *De novo* prediction of coding sequences
- SECIS prediction
- Homology-based prediction of coding sequences

Homology-based selenoprotein finding

Exploit conservation of selenoproteins between species



- Simplest case: "map" selenoproteins from one species to another
- More complex strategies can be used to discover novel selenoproteins

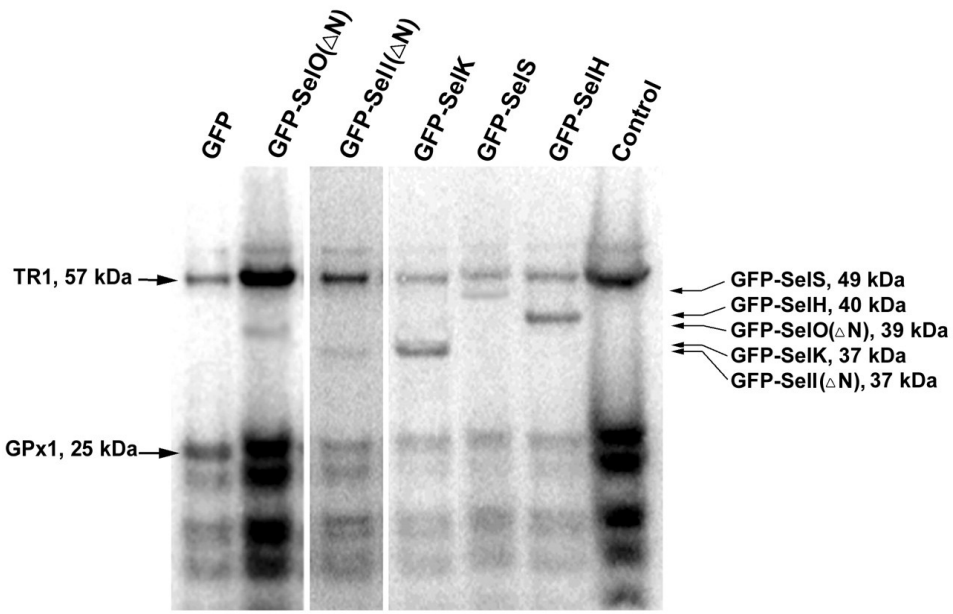
Human genome draft published in 2001

+

Mouse genome draft published in 2002

Characterization of mammalian selenoproteins

(Kryukov et al., Science 300:1439-1443, 2003)



B

Selenoprotein	Chromosomal location (number of exons)	Sec location in protein (length of protein)	Selenoprotein structure
15kDa	1p22.3 (5)	93 (162)	
DI1	1p32.3 (4)	126 (249)	
DI2	14q31.1 (2)	133 (265)	
DI3	14q32	144 (278)	
GPx1	3p21.31 (2)	47 (201)	
GPx2	14q23.3 (2)	40 (190)	
GPx3	5q33.1 (5)	73 (226)	
GPx4	19p13.3 (7)	73 (197)	
GPx6	**6p22.1 (5)**	**73 (221)**	
H	**11q12.1 (4)**	**44 (122)**	
I	**2p23.3 (10)**	**387 (397)**	
K	**3p21.31 (5)**	**92 (94)**	
M	22q12.2 (5)	48 (145)	
N	1p36.11 (12)	428 (556)	
O	**22q13.33 (9)**	**667 (669)**	
P	5p12 (4)	59, 300,318,330, 345, 352, 367, 369, 376, 378 (381)	
R	16p13.3 (4)	95 (116)	
S	**15q26.3 (6)**	**188 (189)**	
SPS2	-	60 (448)	
T	3q24 (6)	36 (182)	
TR1	12q23.3 (15)	498 (499)	
TR2	3q21.2 (16)	655 (656)	
TR3	22q11.21 (18)	522 (523)	
V	**19q13.13 (6)**	**273 (346)**	
W	19q13.32 (6)	13 (87)	

Characterization of Mammalian Selenoproteomes

Gregory V. Kryukov,[1] Sergi Castellano,[2] Sergey V. Novoselov,[1] Alexey V. Lobanov,[1] Omid Zehtab,[1] Roderic Guigó,[2] Vadim N. Gladyshev[1]*

In the genetic code, UGA serves as a stop signal and a selenocysteine codon, but no computational methods for identifying its coding function are available. Consequently, most selenoprotein genes are misannotated. We identified selenoprotein genes in sequenced mammalian genomes by methods that rely on identification of selenocysteine insertion RNA structures, the coding potential of UGA codons, and the presence of cysteine-containing homologs. The human selenoproteome consists of 25 selenoproteins.

In the universal genetic code, 61 codons encode 20 amino acids, and 3 codons are terminators. However, the UGA codon has a dual function in that it signals both the termination of protein synthesis and incorporation of the amino acid selenocysteine (Sec) (*1–3*). Available computational tools lack the ability to correctly assign UGA function. Consequently, there are numerous examples of misinterpretations of UGA codons as both Sec codons (*4*) and terminators (*5, 6*), including annotations of the human genome (*7, 8*), where no selenoproteins have been correctly predicted. With 18 human selenoprotein genes previously discovered (*3*), the estimates of the actual number of such genes vary greatly (*9*). All previously characterized selenoproteins except selenoprotein P (*10*) contain single Sec residues that are located in enzyme-active sites and are essential for their activity. Thus, misidentification of UGA codons leads to a loss of crucial biological and functional information. Sec is cotranslationally incorporated into nascent polypeptides in response to UGA codons when a specific stem-loop structure, designated the Sec insertion sequence (SECIS) element, is present in the 3′ untranslated regions (UTRs) in eukaryotes and in archaea, or immediately downstream of UGA in bacteria (*1, 11–13*). Trans-acting factors, including Sec tRNA, Sec-specific elongation factor, selenophosphate synthetase (SPS), Sec synthase, and a SECIS-binding protein, are also required for Sec biosynthesis and insertion (*1, 3, 13–15*). Most known selenoprotein genes have homologs, in which Sec is replaced with cysteine (Cys). However, these proteins are poor catalysts as compared with selenoproteins (*3*).

We hypothesized that the UGA dual-function problem could be solved by identifying selenoprotein genes in sequenced genomes and assigning terminator functions to the remaining in-frame UGAs. The requirement of SECIS elements for Sec insertion and the presence of Cys-containing homologs of selenoproteins suggested two independent bioinformatics methods for selenoprotein identification. In addition, we used an ob-

[1]Department of Biochemistry, University of Nebraska, Lincoln, NE 68588–0664, USA. [2]Grup de Recerca en Informàtica Biomèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, Centre de Regulació Genòmica, Doctor Aiguader 80, 08003 Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: vgladyshev1@unl.edu

Sergi Castellano

- **Grau,** Universitat de Barcelona
- **Doctorat** l'any 2007
- **PostDoc**
  - Marla Berry, Universitat de Hawaii
  - Andrew G. Clark, Cornell University
  - Sean Eddy, Janelia Farm
- **Group Leader**
  - Max Plank Institute for Evolutionary Antropology, Leipzig
  - University College London

The first eukaryotic selenoproteomes

- Manual analysis of selenoproteins in the first genomes published

SelenoDB

Release 1.0

Welcome

SelenoDB is a joint effort between experimental and computational labs to create, mantain and update a database of eukaryotic selenoprotein genes, proteins, SECIS elements and related molecules. Selenoproteins are routinely mispredicted by automatic annotation systems and, therefore, misannotated in most genomic databases. We aim to provide correct annotations for the growing number of known selenoprotein genes.

Current efforts are directed towards the construction of an initial set of genomic annotations in selected sequenced organisms using *ad hoc* computational tools and **manually curated** predictions. Computational approaches include *ab initio* and comparative gene prediction together with RNA secondary structure predictions.

Feel free to use these data in your research provided the source is acknowledged. Functional annotation will be open to the selenium community in the near future.

Please, select the Virginia lab below to access the database.

Virginia Nebraska Barcelona Hawaii

- A public source of selenium -

PhD students on selenoproteins at the Guigó lab

2007: Sergi Castellano

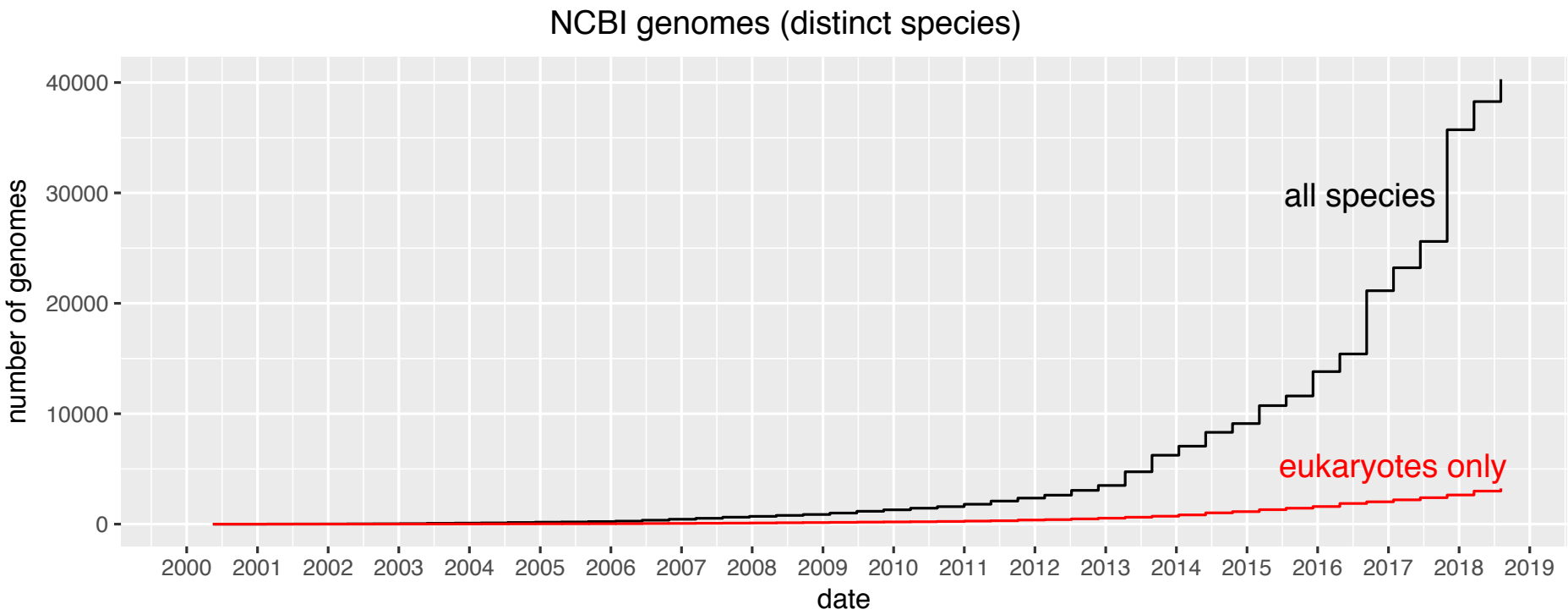
2009: Charles Chapple

2013: Marco Mariotti

2016: Didac Santesmasses

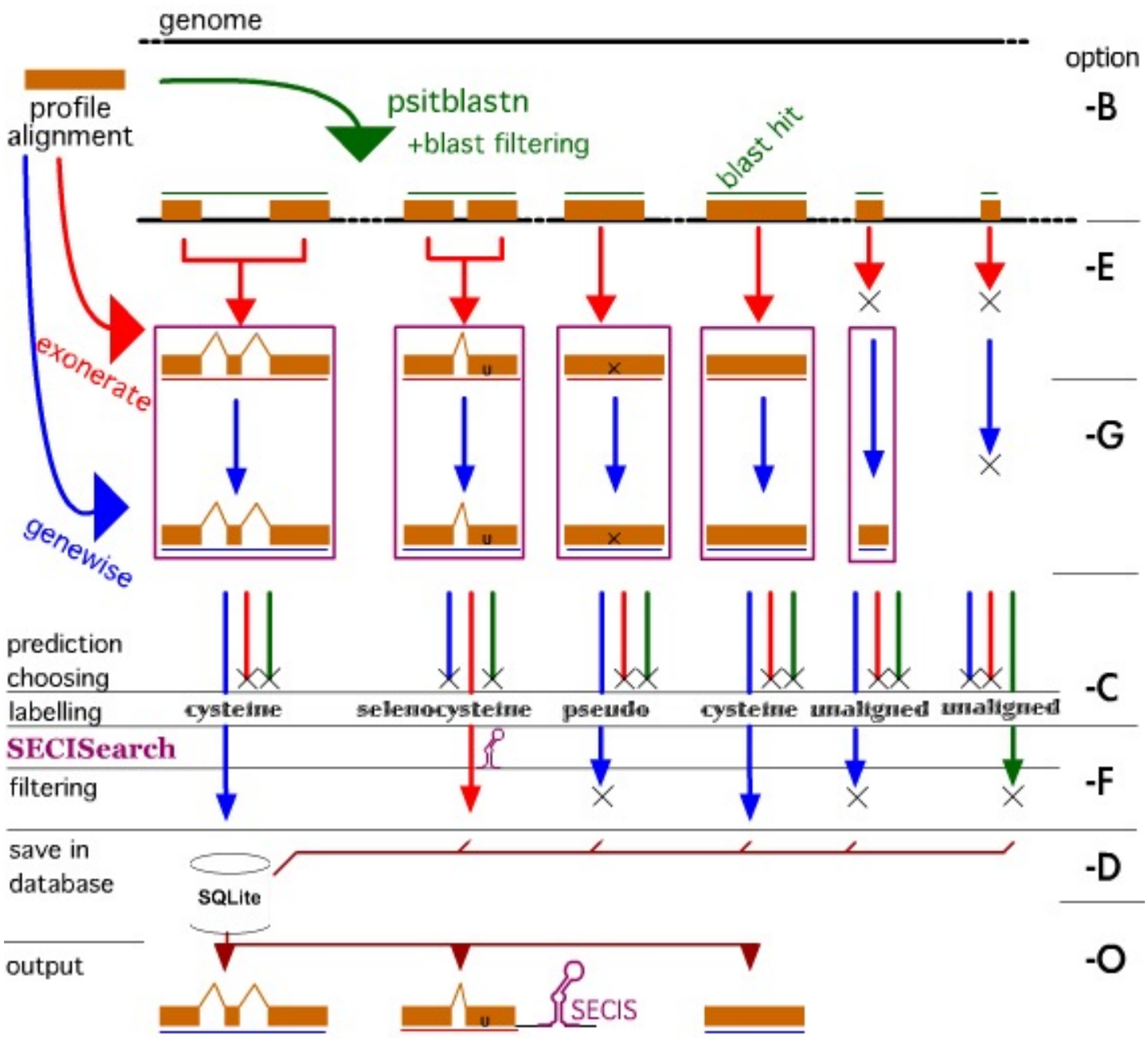
New automated methods

The explosion of sequencing



From manual curation to automated gene finding

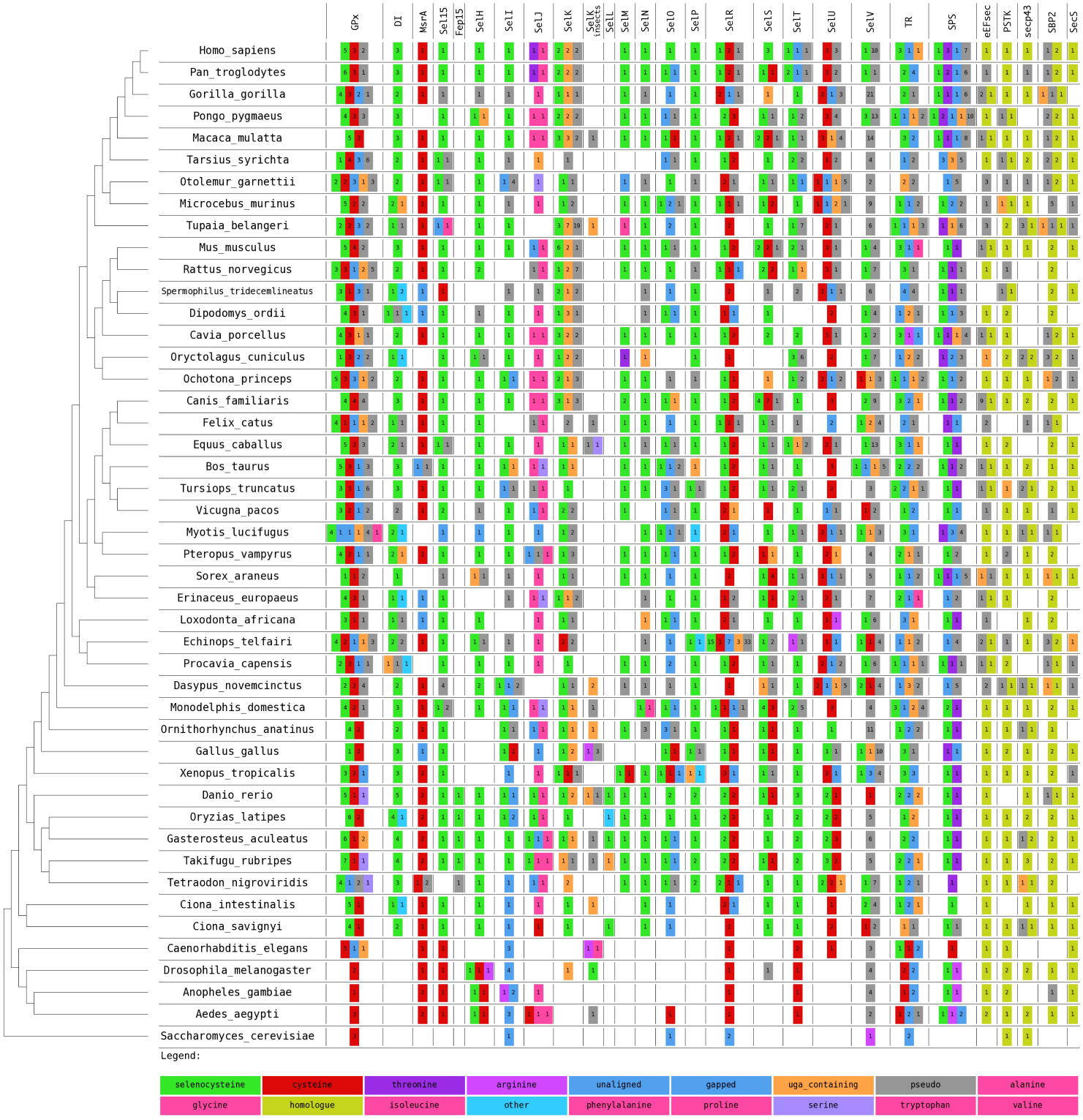
Selenoprofiles



Mariotti et al. 2010

Large scale automatic prediction of known selenoprotein families

Ensembl genomes (old release)



Mariotti et al. 2010

Marco Mariotti

- **Grau,** Università di Bologna
- **Doctorat** l'any 2013
- **PostDoc**
  - Toni Gabaldon
    CRG

  - John Atkins
    University College Cork, Ireland

  - Vadim Gladyshev
    Harvard Medical School, Boston

- **Ramón y Cajal / Group leader**
  Universitat de Barcelona

Didac Santesmasses

- **Grau,** Universitat Pompeu Fabra
- **Doctorat** l'any 2017
- **PostDoc**
  - John Atkins
    University College Cork, Ireland

  - Vadim Gladyshev
    Harvard Medical School, Boston

PLOS | COMPUTATIONAL BIOLOGY

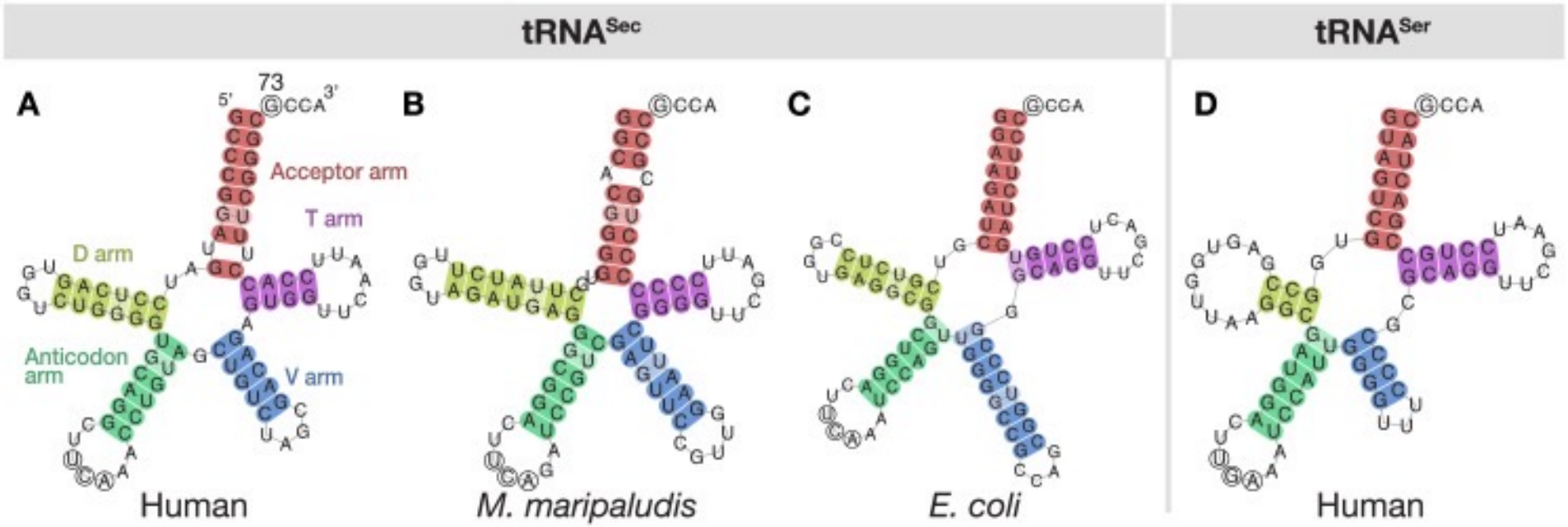
RESEARCH ARTICLE

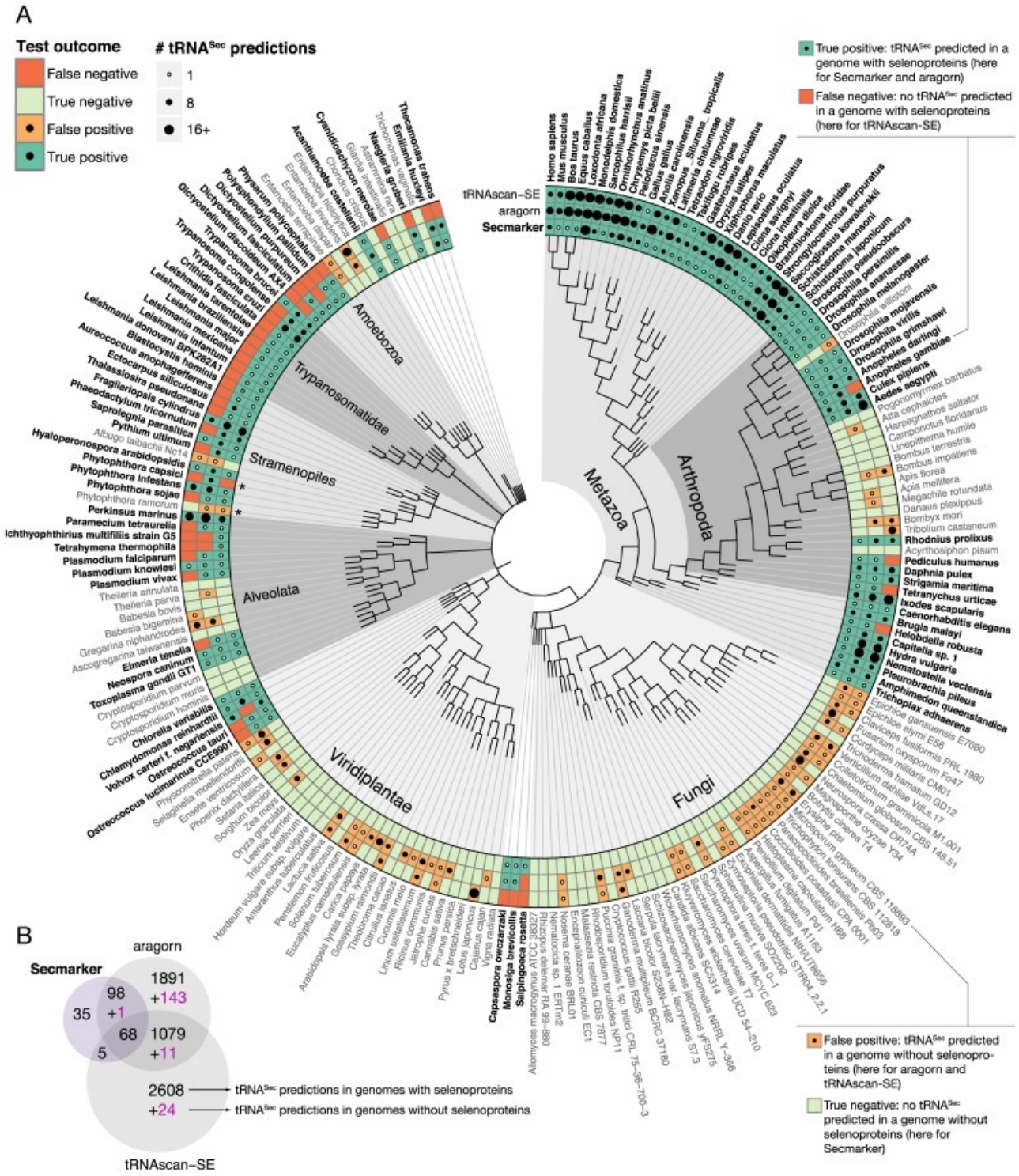
Computational identification of the selenocysteine tRNA (tRNA^{Sec}) in genomes

Didac Santesmasses[1,2,3]*, **Marco Mariotti**[1,2,3,4]*, **Roderic Guigó**[1,2,3]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain, **2** Universitat Pompeu Fabra (UPF), Barcelona, Spain, **3** Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Barcelona, Spain, **4** Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

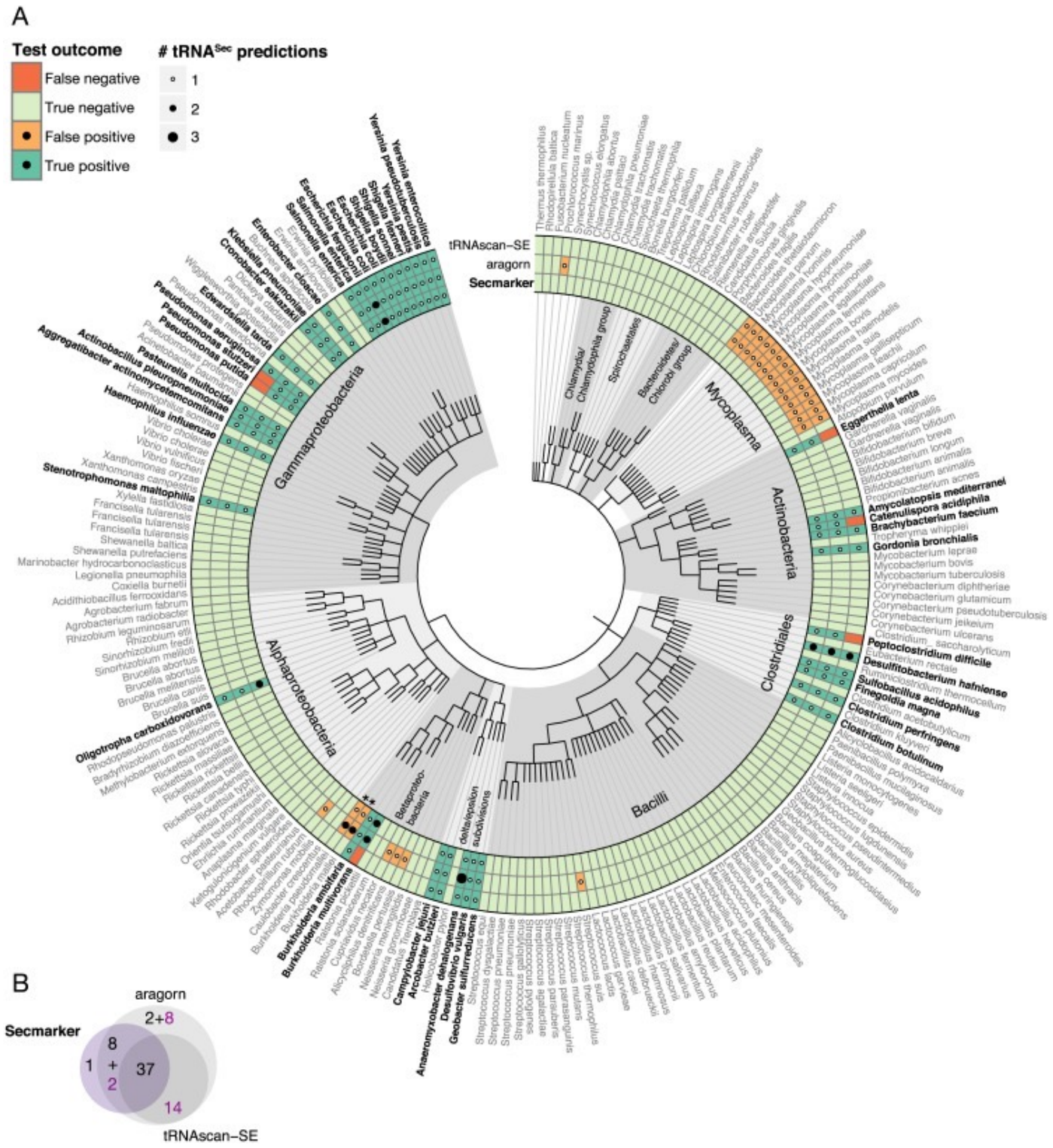
* didac.santesmasses@crg.cat (DS); marco.mariotti@crg.cat (MM)





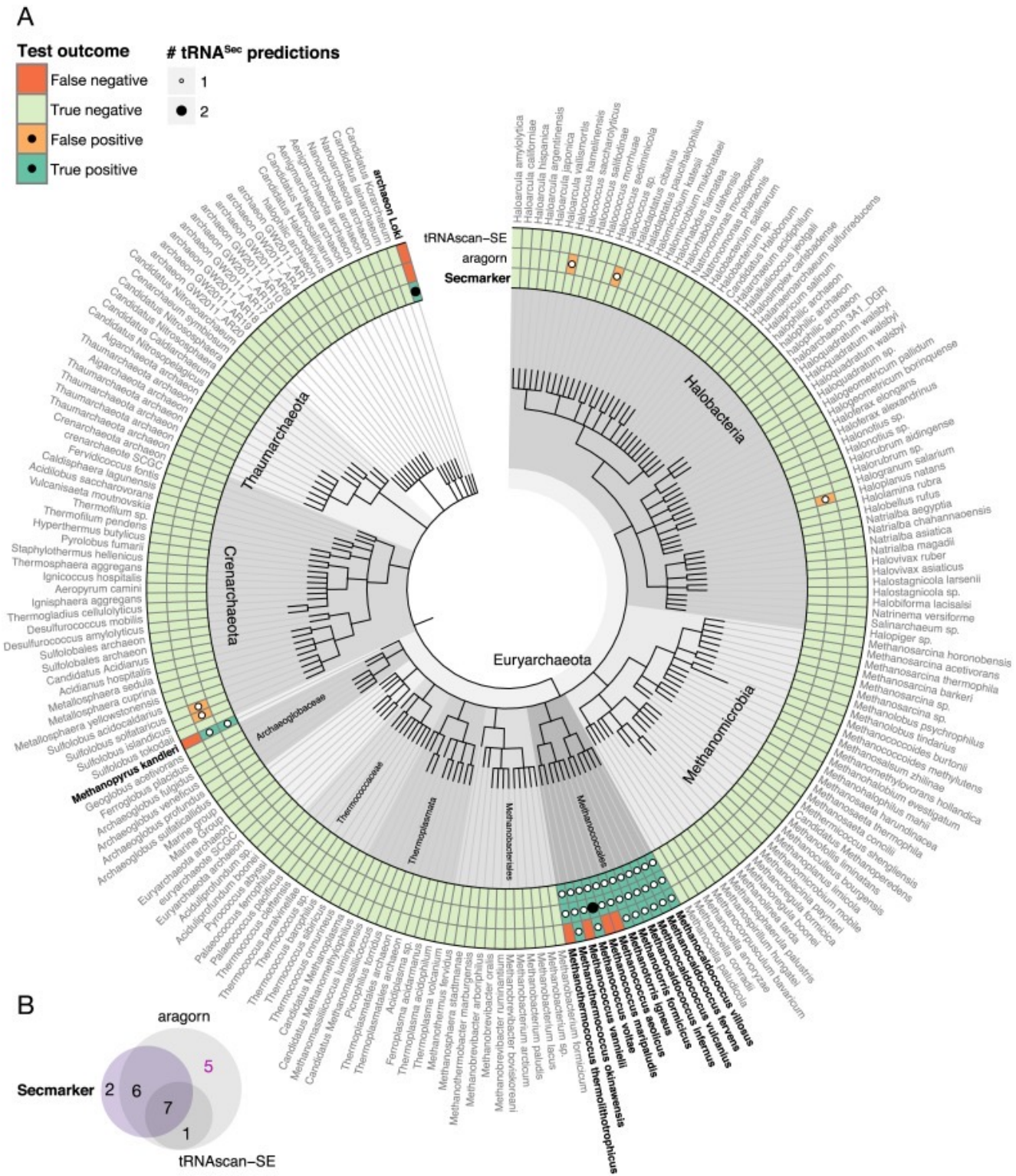
Eukaryotes

Tracing which species encode selenocysteine



Bacteria

Tracing which species encode selenocysteine



Archaea

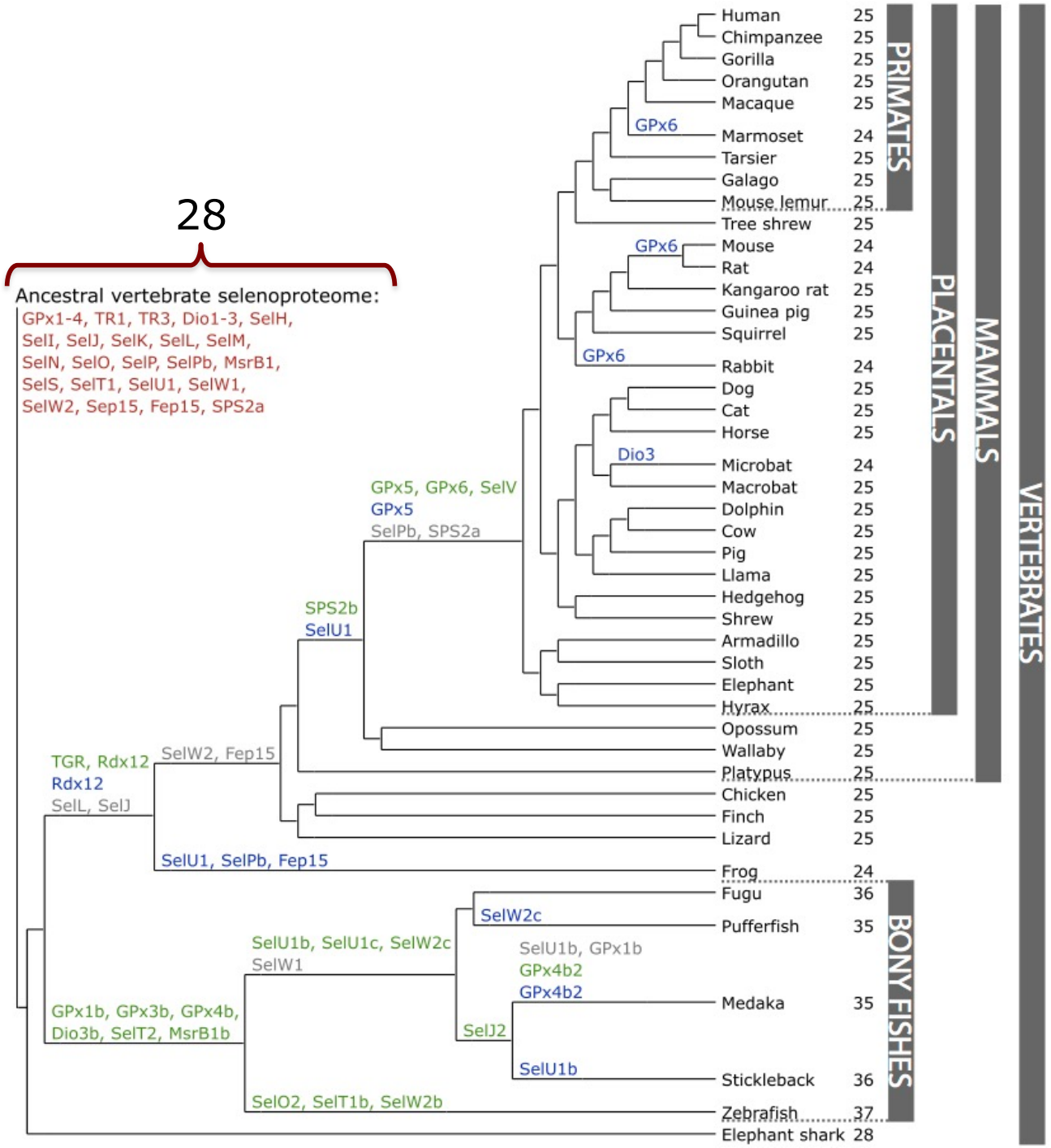
Tracing which species encode selenocysteine

Evolution of selenoproteins

- 25 selenoproteins in human, 24 in mouse
- Present in the three domains of life
- Their number varies for different taxa
  - 3 selenoproteins in *Drosophila melanogaster*
  - 1 selenoprotein in *Caenorhabditis elegans*

The increasing number of genomes available allows to trace the evolution of selenoproteins

Selenoproteins in mammals / vertebrates



20 duplications

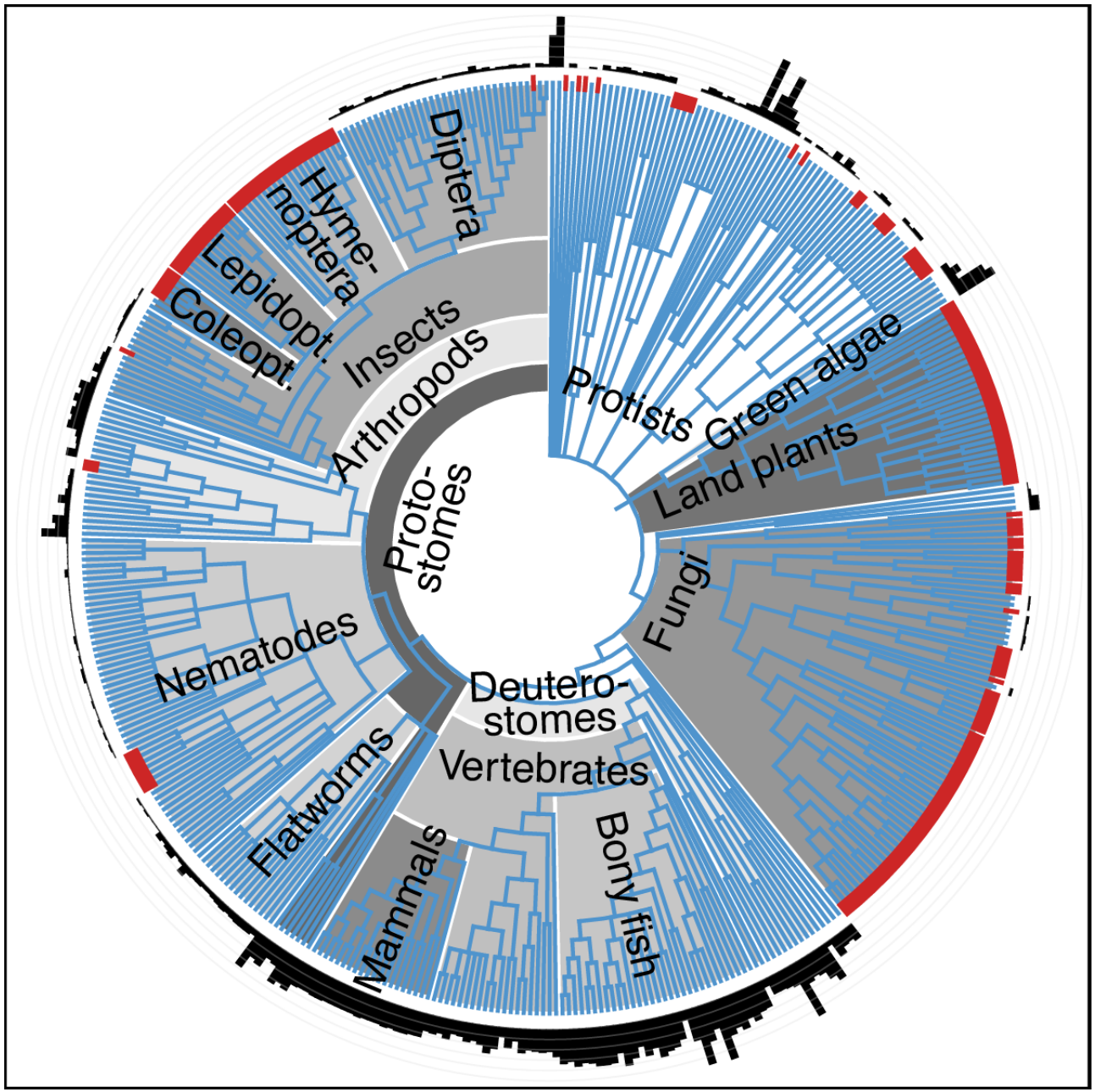
9 gene losses

13 Sec → Cys

Vertebrates have a quite conserved selenoproteome

Mariotti et al. 2012

Selenoproteomes of eukaryotes



- Black bar: proportional to number of selenoproteins
- Red: Sec was lost
- Vertebrates have "many" selenoproteins
- Sec lost in many insects

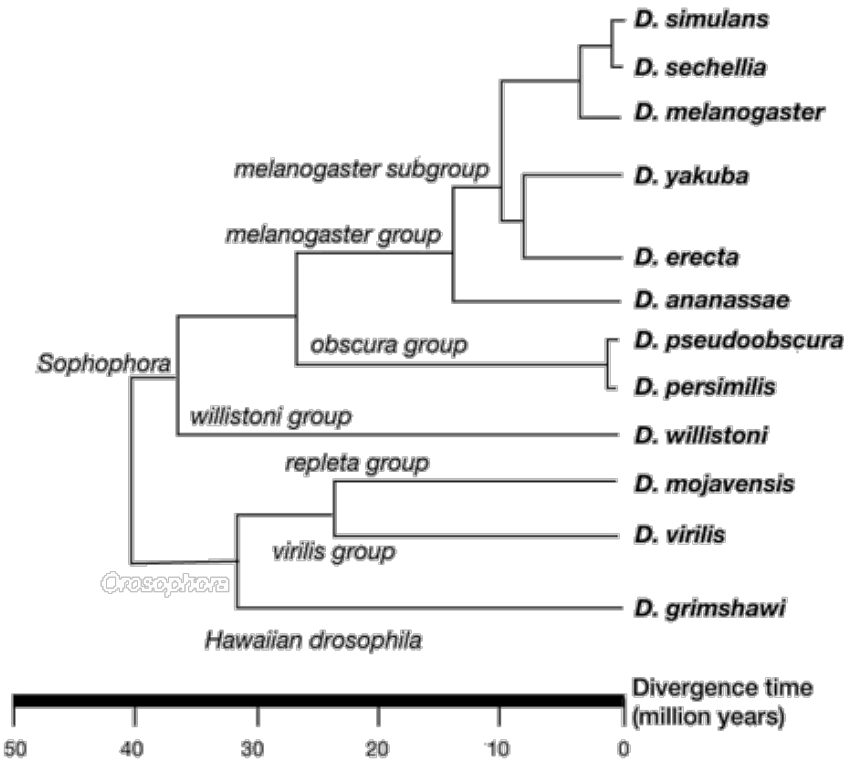
Charles Chapple

- **Grau,** University of York, UK

- **Doctorat** l'any 2009

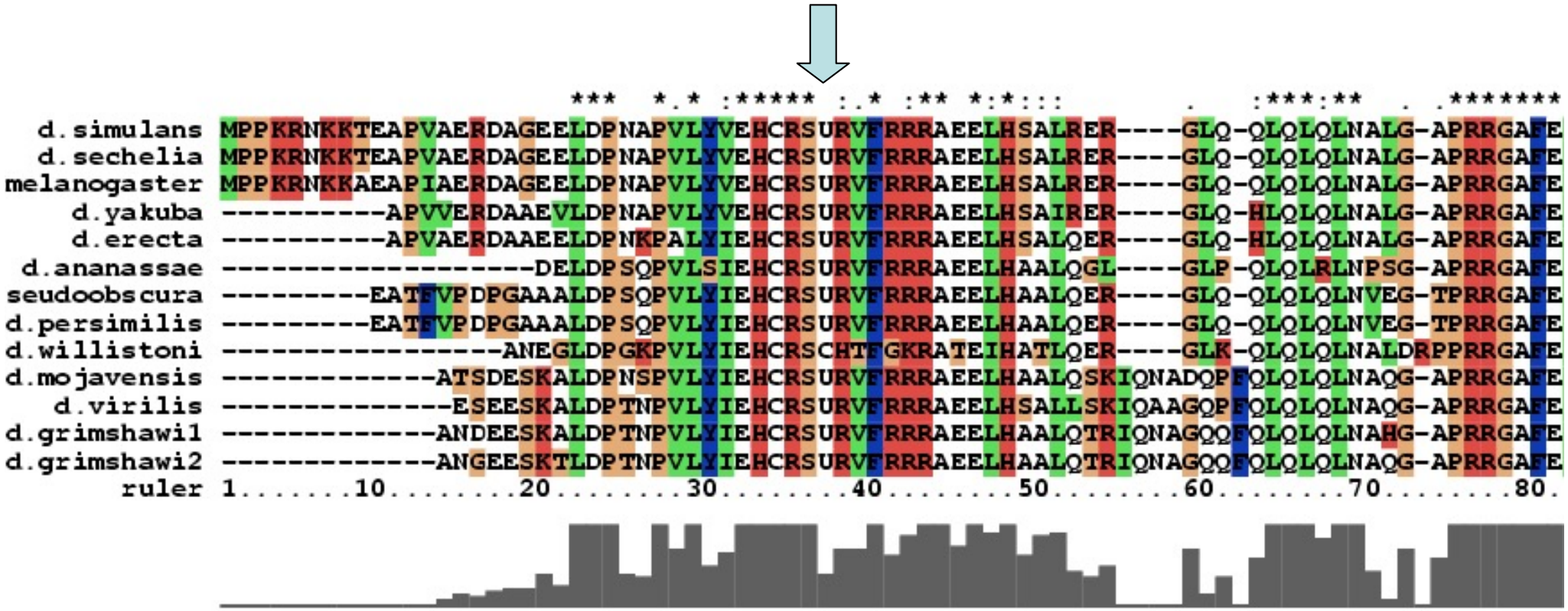
- **PostDoc**
  -Christine Brune
  INSERM, Marseille

- **Head of bioinformatics**,
  Saphetor



12 Drosophila genomes published in 2007

SelenoH alignment across fly genomes



D.willistoni lacks some of the genes involved in selenoprotein metabolism

- tRNA-Sec
- SPS2
- EFSec
- SecS

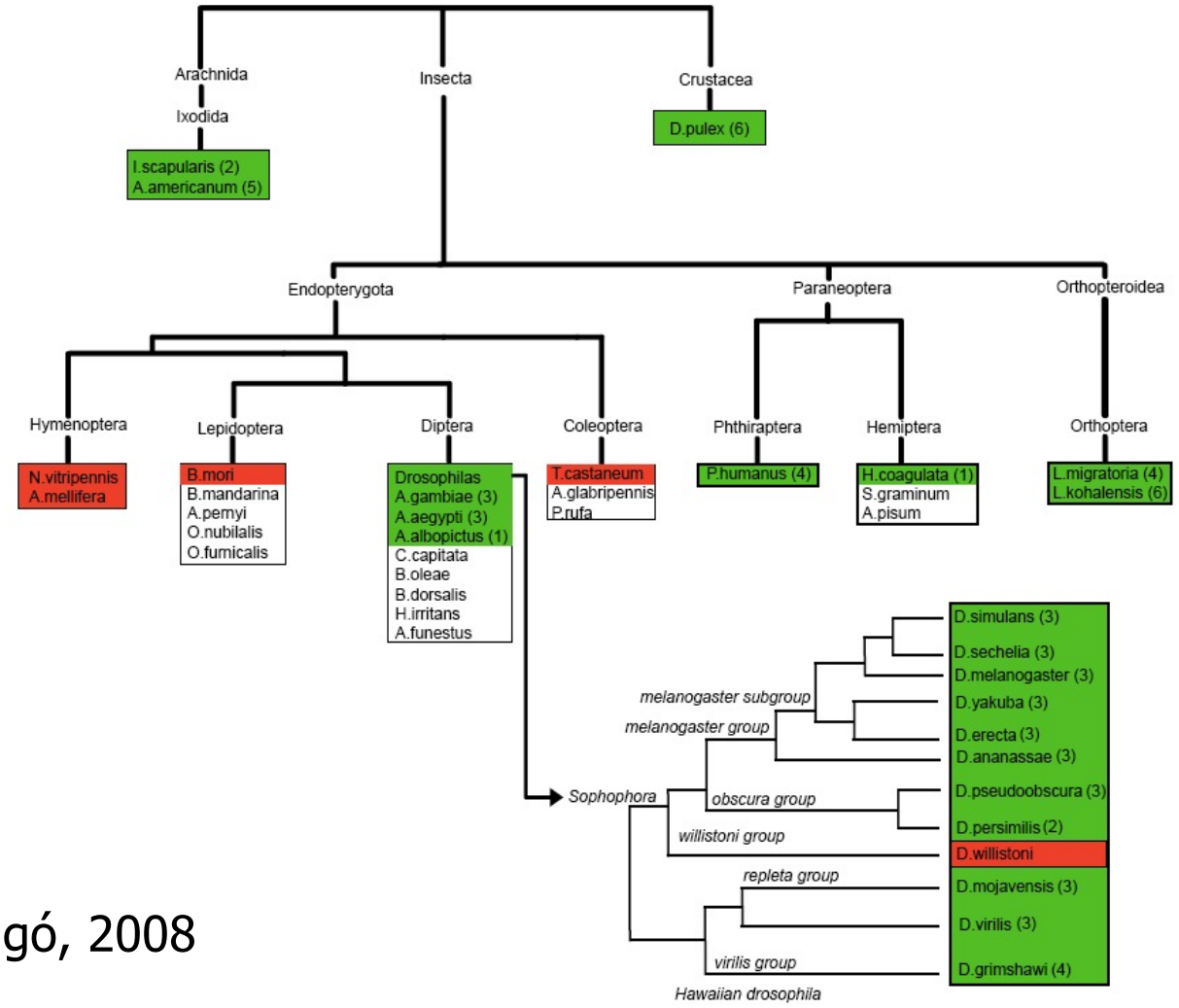
The first animal known to lack selenoproteins

OPEN ACCESS Freely available online

PLoS one

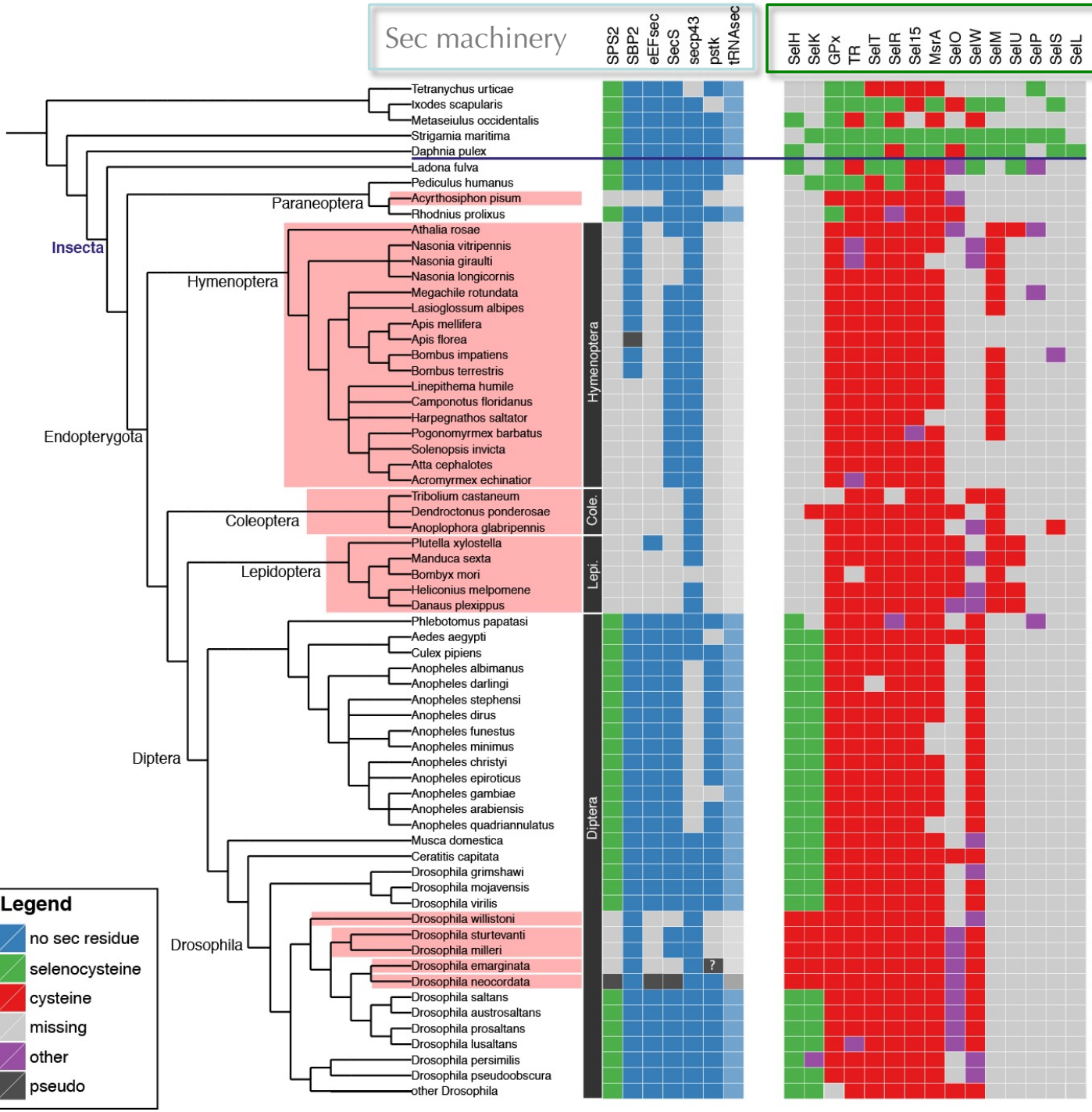
Relaxation of Selective Constraints Causes Independent Selenoprotein Extinction in Insect Genomes

Charles E. Chapple[1], Roderic Guigó[2]*



Chapple and Guigó, 2008

Selenoprotein extinction in insects

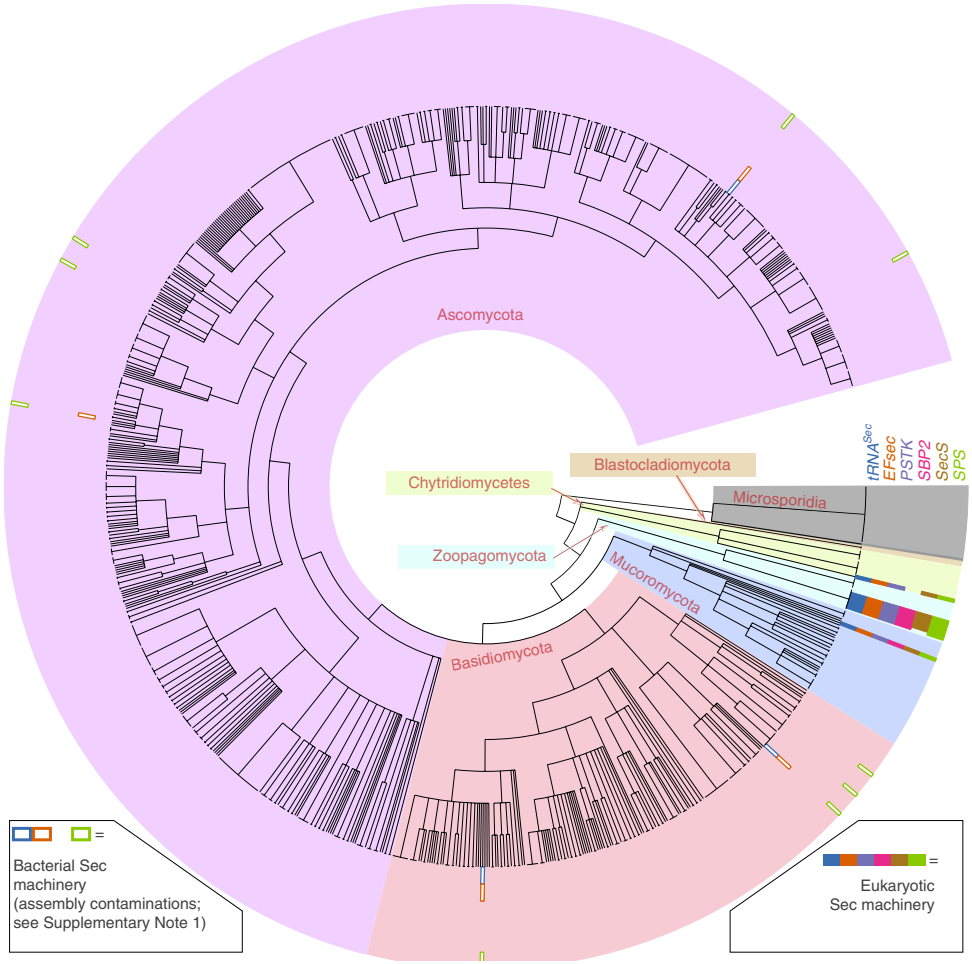


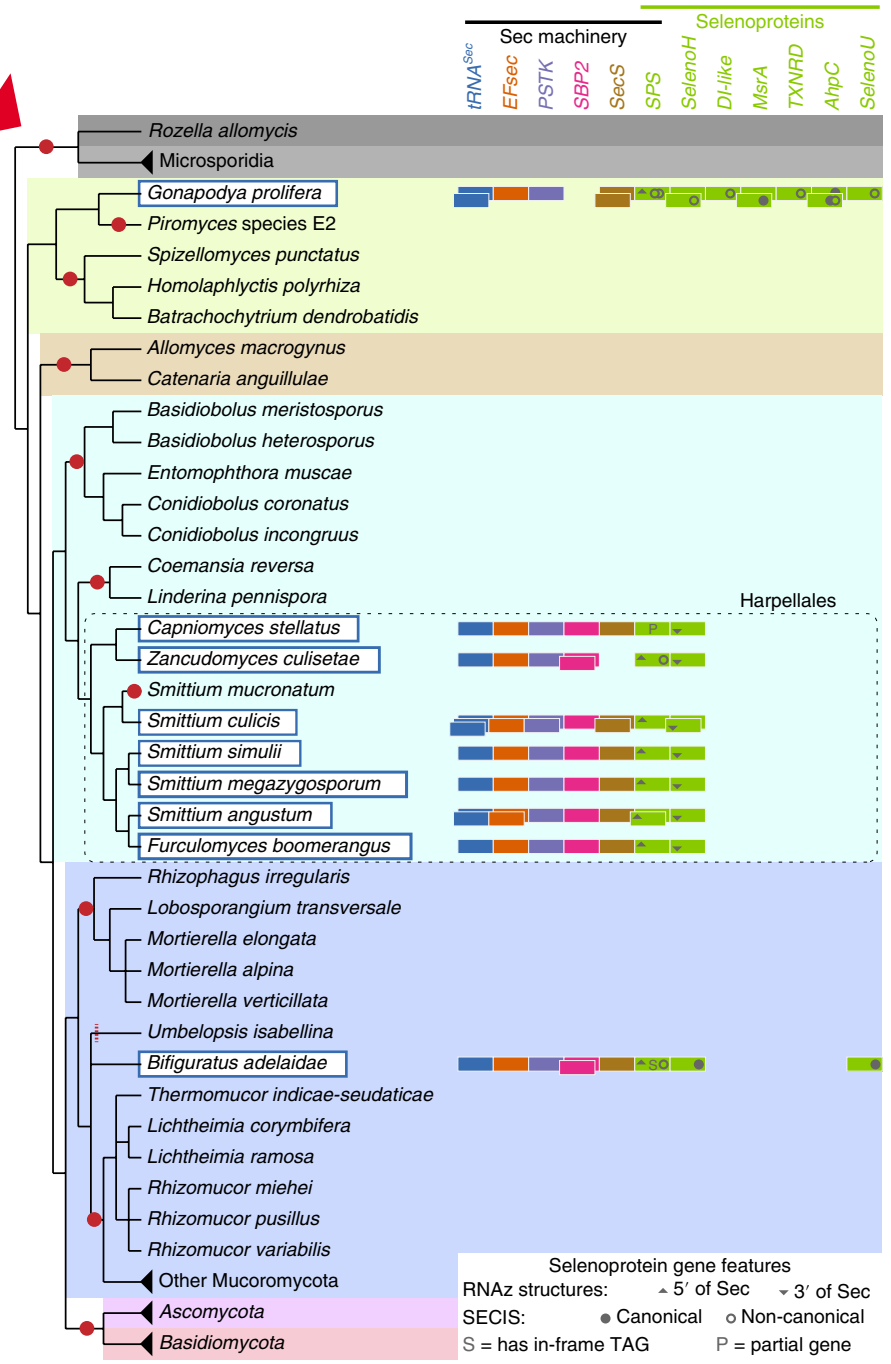
Many insects lost all selenoproteins independently

No selenoproteins, no Sec machinery

Selenoprotein extinctions in fungi

Loss of all selenoproteins





Mariotti et al. 2019

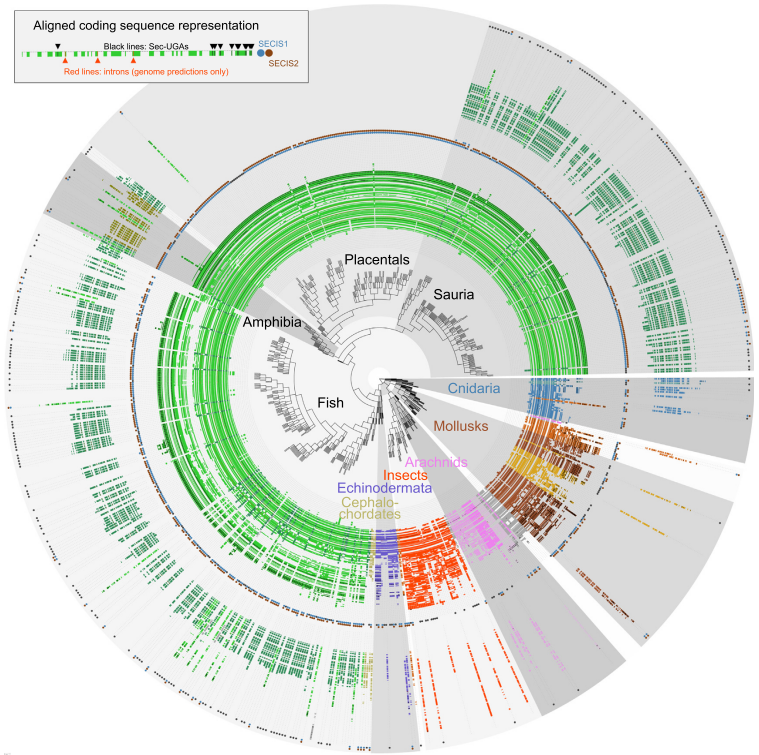
Selenocysteine evolution

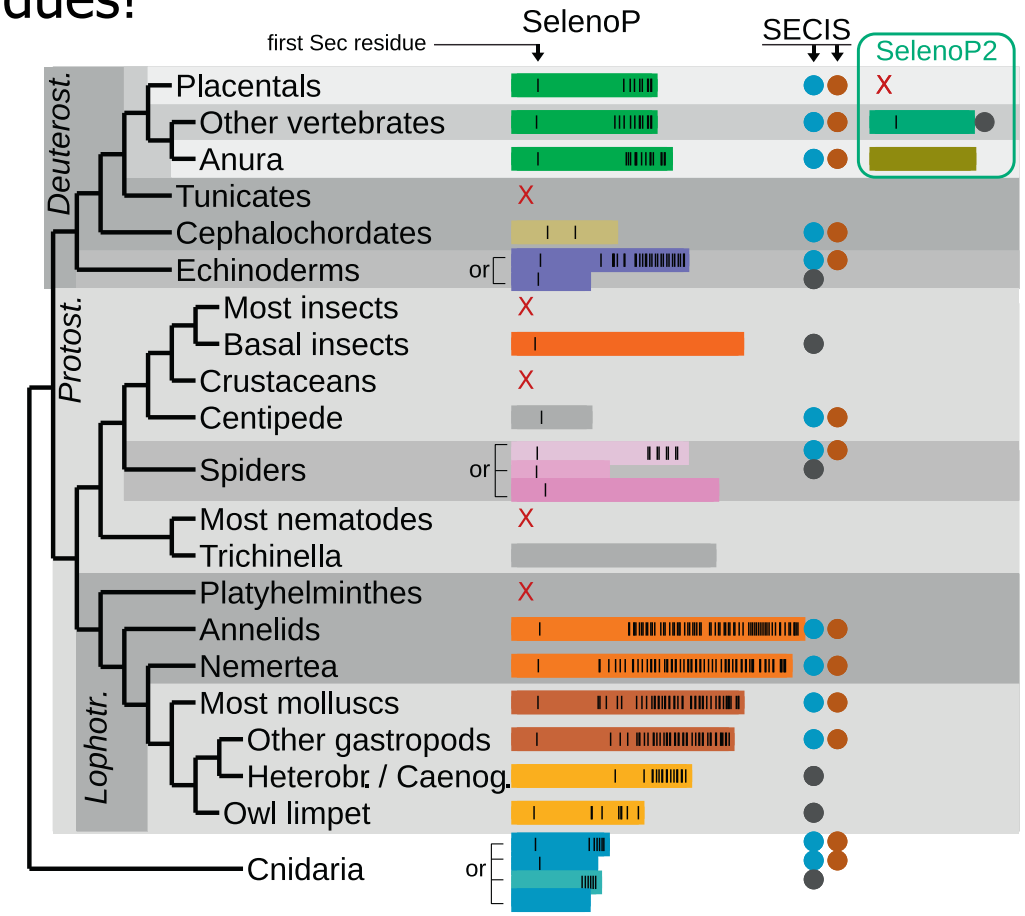
It is dynamic: selenoproteomes respond to various evolutionary forces, such as:

- Selenium availability
- Sec “usefulness” as redox catalyzer
- Competition with translation termination

SelenoP: a multi-Sec selenoprotein

- SelenoP transport Sec across the plasma
- It contains 10 Sec residues in human and mouse, with 2 SECISes
- Variable number of Sec across species: proxy for Sec usage
- Some mollusks have >100 Sec residues!





Baclaocos et al. 2019

Acknowledgments

Roderic Guigó

Vadim N. Gladyshev
Didac Santesmasses

John F. Atkins
Pavel V. Baranov
Janinah Baclaocos

Sergi Castellano

Charles Chapple

Montserrat Corominas,
Florenci Serras

Toni Gabaldón

Gustavo Salinas