Recerca de selenoproteïnes en el genoma d'organismes eucariotes

- Bioinformàtica 2019/20 -

Aida Ripoll (PhD Student)
Didac Santesmasses (PhD)

Bioinformatics and genomics programme
Roderic Guigó's group
Centre for Genomic Regulation, Barcelona

What are **selenoproteins**?

What are **selenoproteins**?

❖ **Role of selenium (Se)**

- One of the **nine** essential trace elements
- **Vital** functions (homeostasis)

Se deficiency → pathophysiological status (Keshan disease)
Se excess → toxic

What are **selenoproteins**?

❖ **Role of selenium (Se)**

- One of the **nine** essential trace elements
- **Vital** functions (homeostasis)
  - *Se deficiency* → pathophysiological status (Keshan disease)
  - *Se excess* → toxic

❖ **Selenoproteins** (Se-containing proteins)

- All **3 domains of life**: eukarya, archea, and eubacteria
- Human **selenoproteome** is encoded in 25 selenoprotein genes (but the number varies for different taxa)
- Sometimes the **orthologue** of a selenoprotein has **Cys** instead of Sec
- Mostly **redox enzyms** → antioxidant protection capacity

What are **selenoproteins**?

❖ Role of selenium (Se)

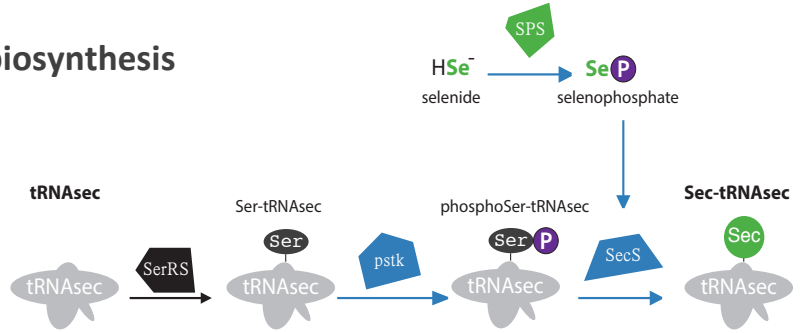
- One of the **nine** essential trace elements
- **Vital** functions (homeostasis)
  - *Se deficiency* → pathophysiological status (Keshan disease)
  - *Se excess* → toxic

❖ **Selenoproteins** (Se-containing proteins)

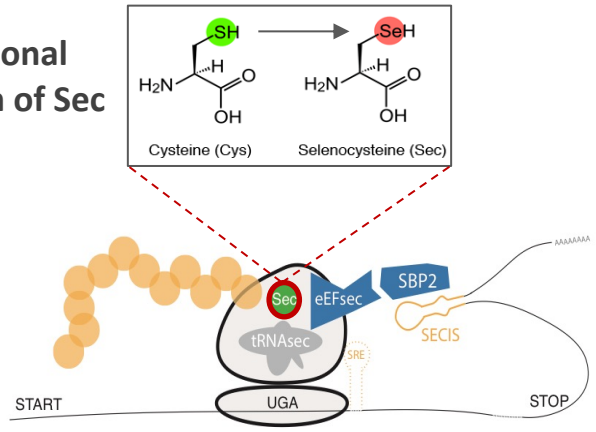
- All **3 domains of life**: eukarya, archea, and eubacteria
- Human **selenoproteome** is encoded in 25 selenoprotein genes (but the number varies for different taxa)
- Sometimes the **orthologue** of a selenoprotein has **Cys** instead of Sec
- Mostly **redox enzyms** → antioxidant protection capacity

❖ **Sec biosynthesis** & **insertion** mechanism

I. Sec biosynthesis



II. Cotranslational incorporation of Sec

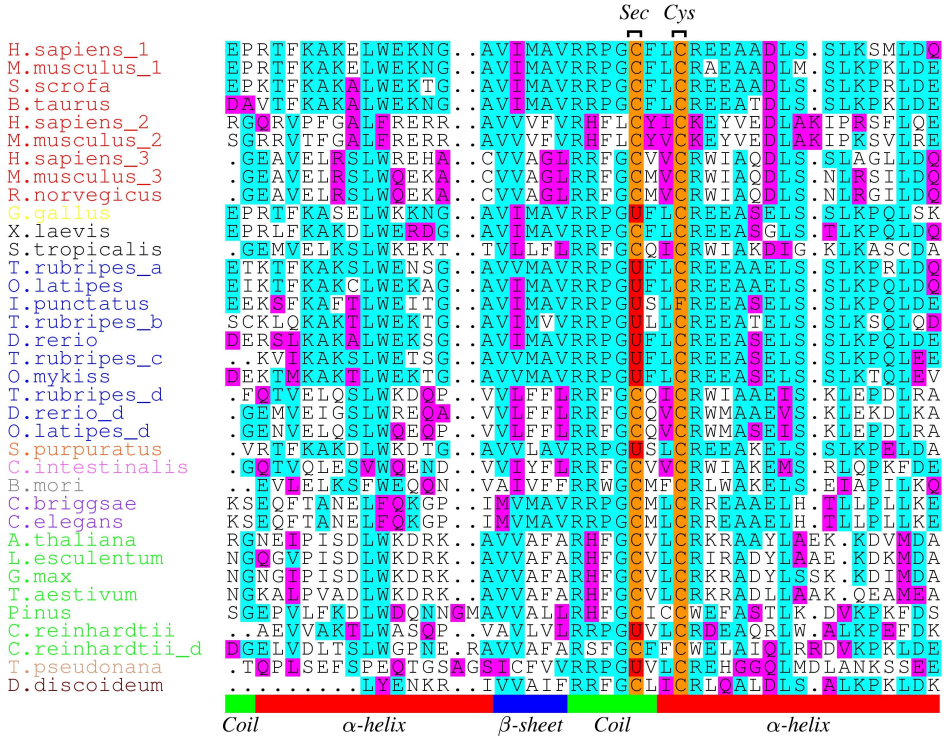


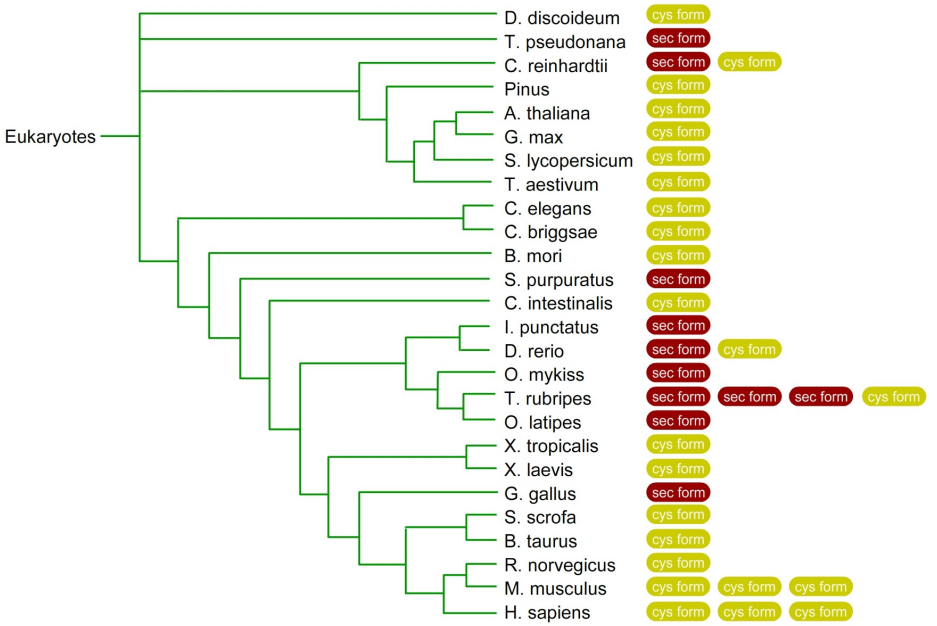
Extracted from:
Mariotti M et al. Mol Biol Evol. 2016;33(9):2441-2453

Selenoprotein families include:

- **Selenoproteins** (Sec-containing proteins)

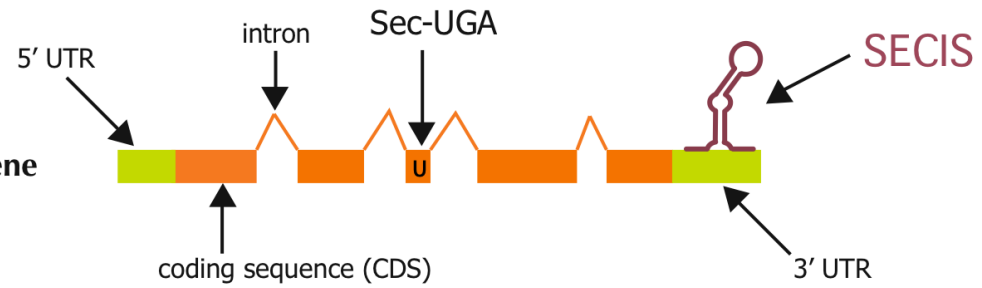
- **Cysteine homologues** (Cys-containing proteins)
  - Orthologues (*speciation* event)
  - Paralogues (*duplication* event)



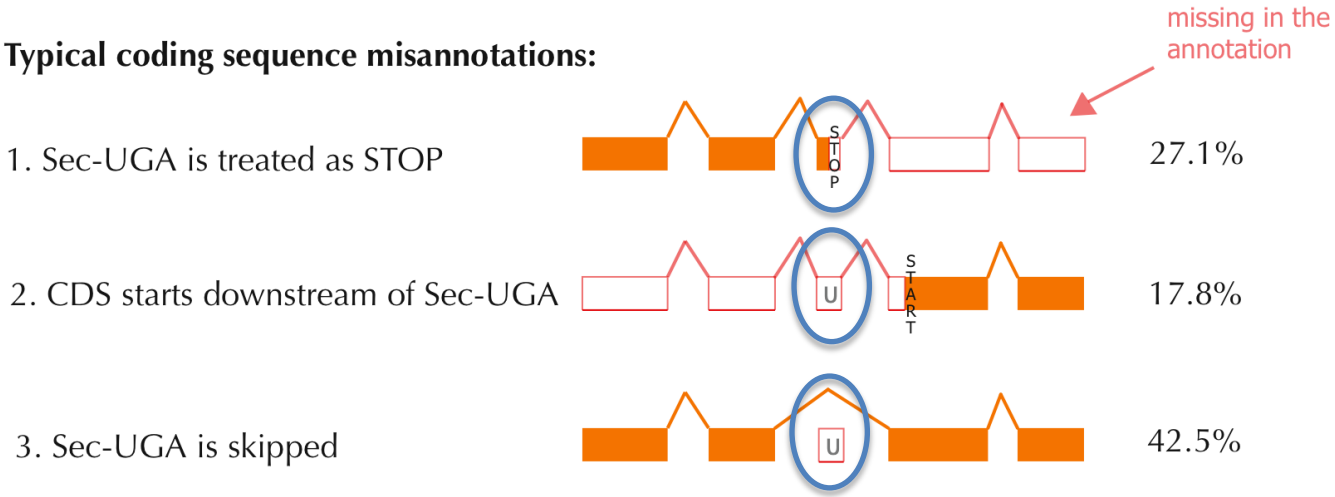


Selenoproteins are **generally misannotated**

Example of selenoprotein gene



Typical coding sequence misannotations:



Percentages are computed by comparison Selenoprofiles-Ensembl annotations – see Mariotti and Guigó, 2010 - Bioinformatics.

Bioinformatics methods for **selenoproteins**

- **De *novo***: Selenogeneid (Castellano et al. 2001)

- **Homology based approaches:**
  - UGA / Sec or UGA / Cys alignments (e.g. Kryukov et al. 2003)
  - Selenoprofiles (Mariotti and Guigó 2010)
  - Seblastian (Mariotti et al. 2013)

- **SECIS prediction:**
  - SECISearch (Kryukov et al. 2003)
  - SECISearch3 (Mariotti et al. 2013)

- **tRNA-Sec prediction:**
  - Secmarker

OPEN ACCESS Freely available online

PLoS one

Composition and Evolution of the Vertebrate and Mammalian Selenoproteomes

Marco Mariotti[1,2☯], Perry G. Ridge[3☯], Yan Zhang[1,4☯], Alexei V. Lobanov[1], Thomas H. Pringle[5], Roderic Guigo[2], Dolph L. Hatfield[6], Vadim N. Gladyshev[1*]

1 Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **2** Center for Genomic Regulation and Universitat Pompeu Fabra, Barcelona, Spain, **3** Department of Biochemistry and Redox Biology Center, University of Nebraska, Lincoln, Nebraska, United States of America, **4** Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **5** Sperling Foundation, Eugene, Oregon, United States of America, **6** Laboratory of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America

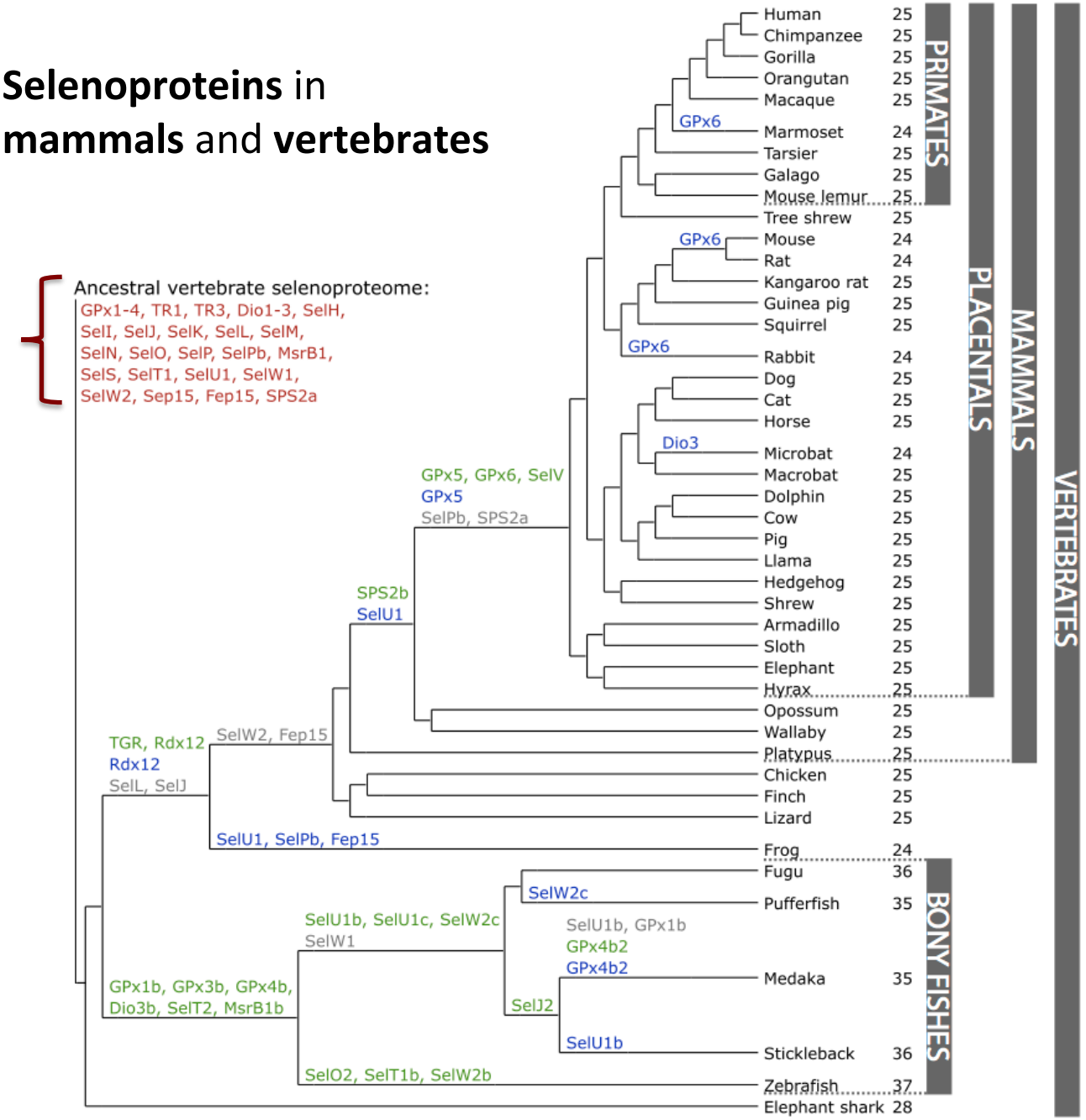
Abstract

Background: Selenium is an essential trace element in mammals due to its presence in proteins in the form of selenocysteine (Sec). Human genome codes for 25 Sec-containing protein genes, and mouse and rat genomes for 24.

Methodology/Principal Findings: We characterized the selenoproteomes of 44 sequenced vertebrates by applying gene prediction and phylogenetic reconstruction methods, supplemented with the analyses of gene structures, alternative splicing isoforms, untranslated regions, SECIS elements, and pseudogenes. In total, we detected 45 selenoprotein subfamilies. 28 of them were found in mammals, and 41 in bony fishes. We define the ancestral vertebrate (28 proteins) and mammalian (25 proteins) selenoproteomes, and describe how they evolved along lineages through gene duplication (20 events), gene loss (10 events) and replacement of Sec with cysteine (12 events). We show that an intronless selenophosphate synthetase 2 gene evolved in early mammals and replaced functionally the original multiexon gene in placental mammals, whereas both genes remain in marsupials. Mammalian thioredoxin reductase 1 and thioredoxin-glutathione reductase evolved from an ancestral glutaredoxin-domain containing enzyme, still present in fish. Selenoprotein V and GPx6 evolved specifically in placental mammals from duplications of SelW and GPx3, respectively, and GPx6 lost Sec several times independently. Bony fishes were characterized by duplications of several selenoprotein families (GPx1, GPx3, GPx4, Dio3, MsrB1, SelJ, SelO, SelT, SelU1, and SelW2). Finally, we report identification of new isoforms for several selenoproteins and describe unusually conserved selenoprotein pseudogenes.

Conclusions/Significance: This analysis represents the first comprehensive survey of the vertebrate and mammal selenoproteomes, and depicts their evolution along lineages. It also provides a wealth of information on these selenoproteins and their forms.

Selenoproteins in **mammals** and **vertebrates**



20 duplications

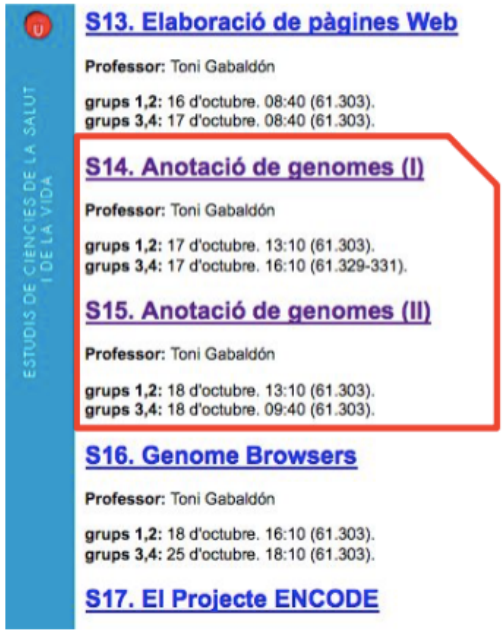
9 gene losses

13 Sec → Cys

Protocol overview

Tools:

- **BLAST** - typically **tblastn**
- **Exonerate** - protein2genome mode
- **Genewise**
- **T-coffee**



http://bioinformatica.upf.edu/
http://bioinformaticaupf.crg.eu

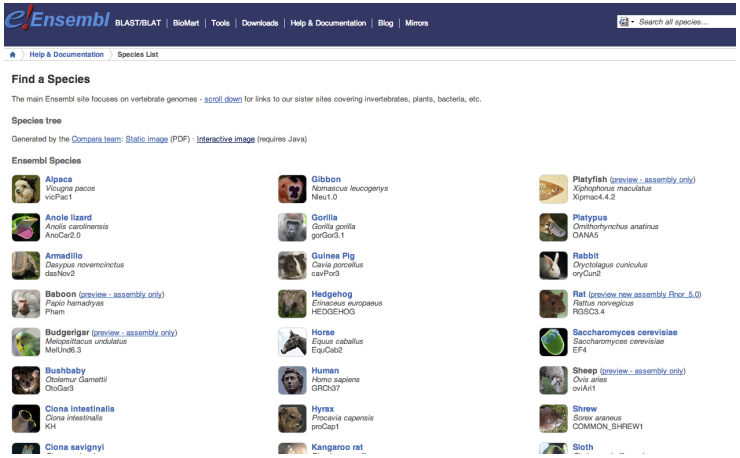
- Webserver with **SECISearch3** and **Seblastian**:

  http://seblastian.crg.es/

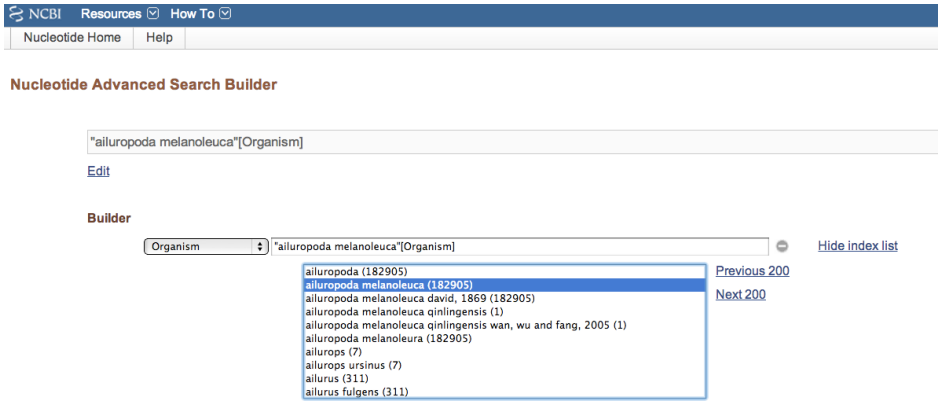
Useful resources

Your **assigned genomes** will be available in the **UPF computers** when you will start the project

- **Ensembl:** Collection of genomes (and annotations)



- **NCBI nucleotide:** Collection of all sequences (genomes, ESTs, etc)



Protocol steps

1st step: Get selenoprotein sequences

- **SelenoDB 2.0 (and 1.0)** SelenoDB

  http://www.selenodb.org (2.0; automatic annotation)

  http://www1.selenodb.org (1.0; manually curated, less species)

- **Protein databases**

  https://www.ncbi.nlm.nih.gov/protein/

  http://www.uniprot.org

- **Past year projects:**

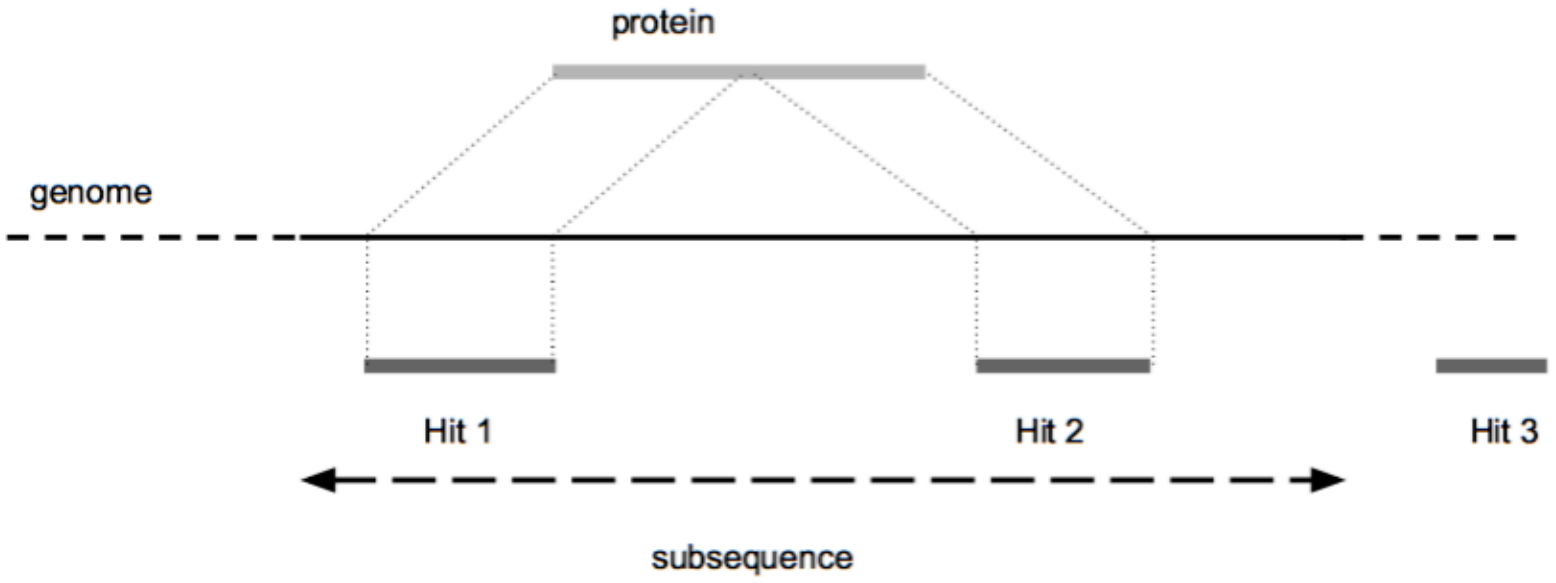
  http://bioinformatica.upf.edu/

NCBI BLAST programs… reminder

Search Name	Query Type	Database Type	Translation
blastn	Nucleotide	Nucleotide	None
tblastn	Peptide	Nucleotide	Database
blastx	Nucleotide	Peptide	Query
blastp	Peptide	Peptide	None
tblastx	Nucleotide	Nucleotide	Query and Database

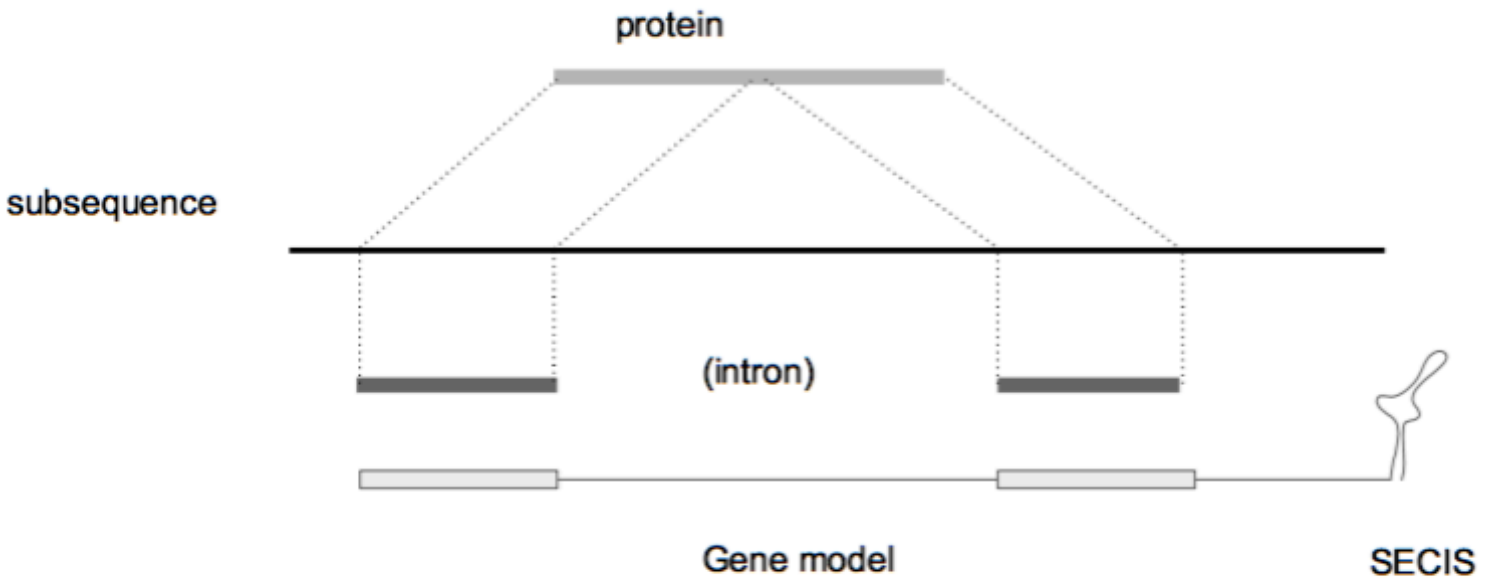
Protocol steps

- **Tblastn:** locate gene exons (independent blast hits)



Protocol steps

- **Exonerate** or **genewise:** multi-exonic gene model
- **Seblastian:** SECIS + selenoprotein prediction



Protocol steps

Gene finding tools: fastasuite **(exonerate)**

- **Fastafetch:** extracting a single sequence from a multifasta (requires previous run of fastaindex)

- **Fastasubseq:** getting a subsequence of a single sequence, careful with indexes, 0-based! Transform gene positions to absolute coordinates.

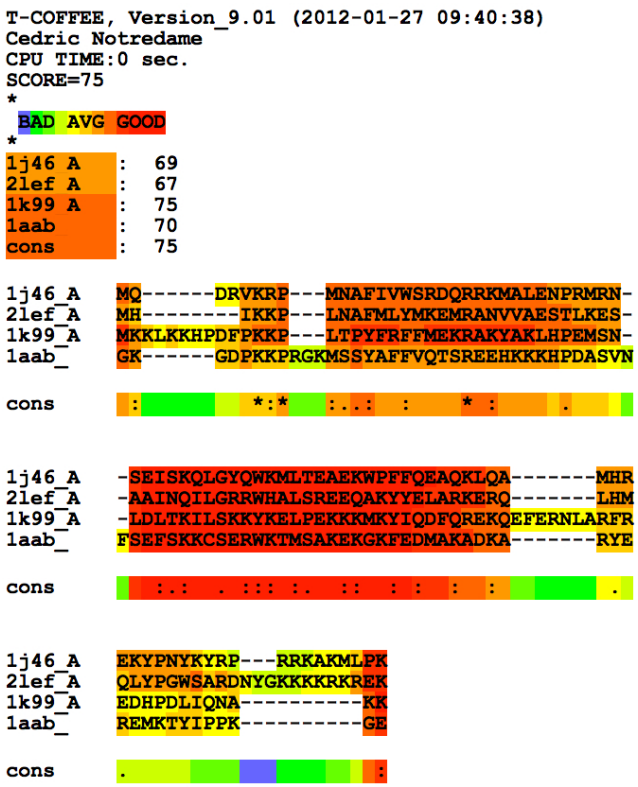
- **Exonerate/Genewise:** predict the gene and align it with the sequence of the selenoprotein that encodes, and also recognizes the exons.

- **fastaseqFromGFF.pl:** obtain the cDNA sequence that encodes the final protein. We get it from the subsequence and the file that contains the exons.

- **Fastatranslate:** translate coding sequences careful with the selenocysteine codon character! It is a good idea to substitute the "*" with "X" or "U" as multiple sequence alignment programs just ignore "*"

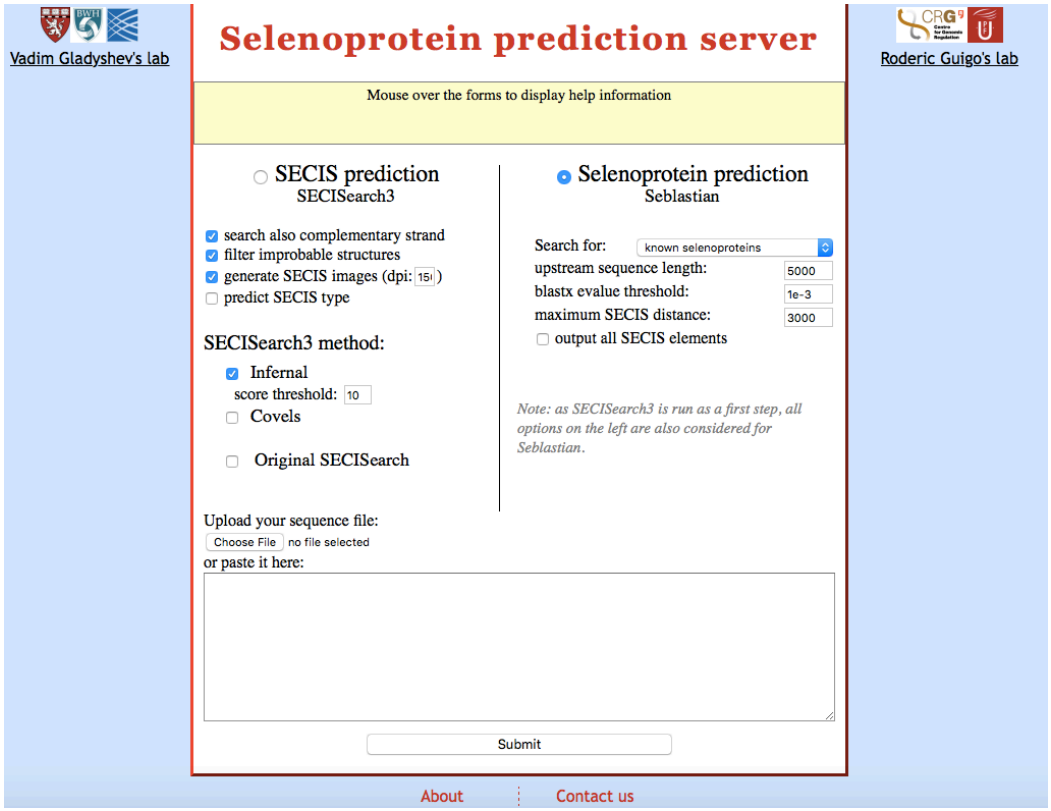
Protocol steps

- **Tcoffe:** compare two sequences, in this case we compare the known sequence (*query protein*) with the homologous sequence of the the genome (*predicted protein*).



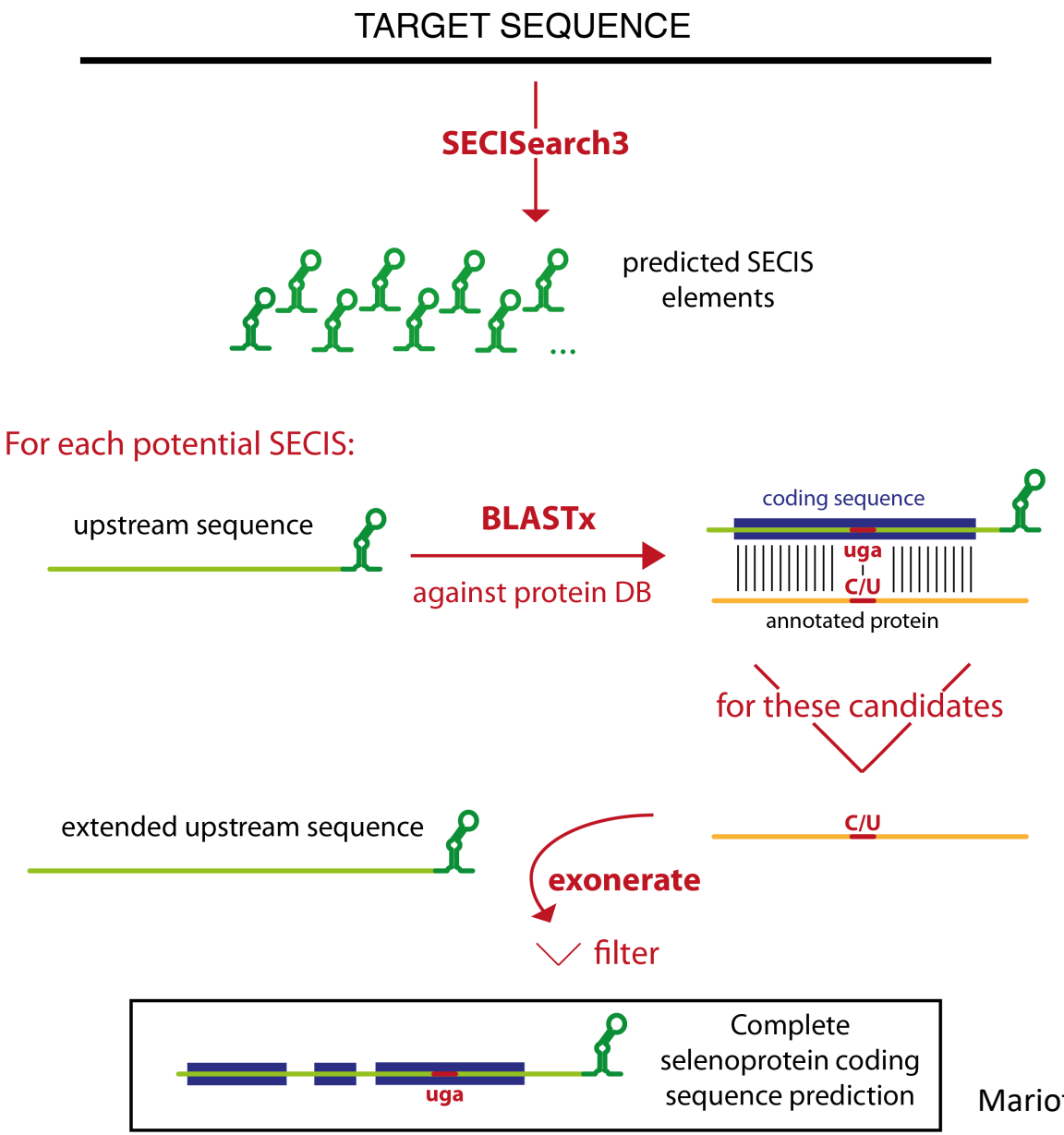
Protocol steps

Seblastian: Predict SECIS in the 3'UTR (using SECISearch3), and then searches upstream for selenoprotein coding sequences.



http://seblastian.crg.es/

Seblastian



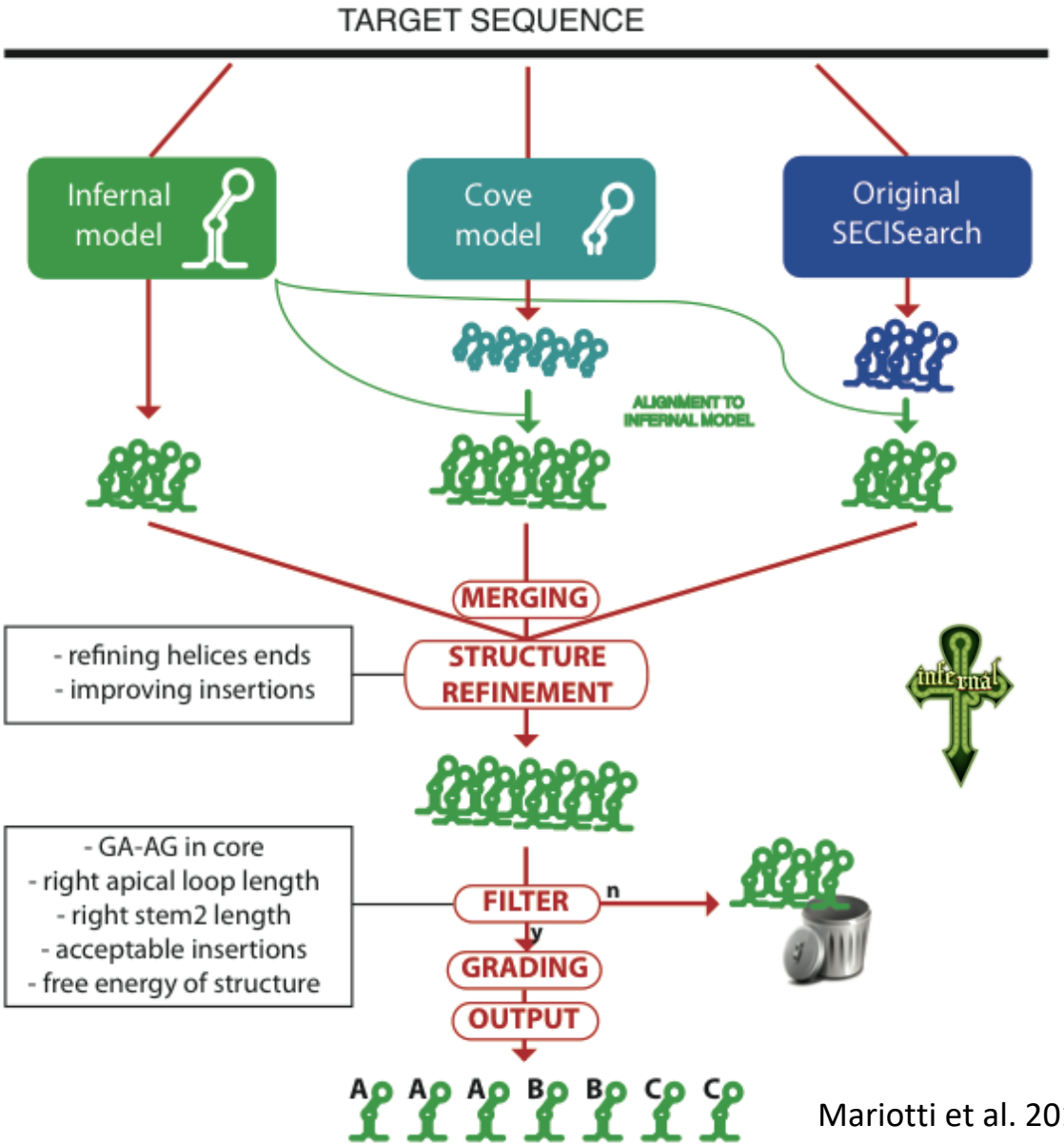
Assumptions:

the presence of a detectable SECIS within acceptable genomic distance from the Sec-UGA

annotated homologue(s) (Sec/Cys) in the reference protein database

Mariotti et al. 2013

SECISearch 3



Based on a manually curated 2ndary structure alignment

Combines up to 3 methods to ensure maximum sensitivity

Filter and grading procedure based on manual inspection of hundreds of SECIS elements

Infernal: inference of RNA alignments

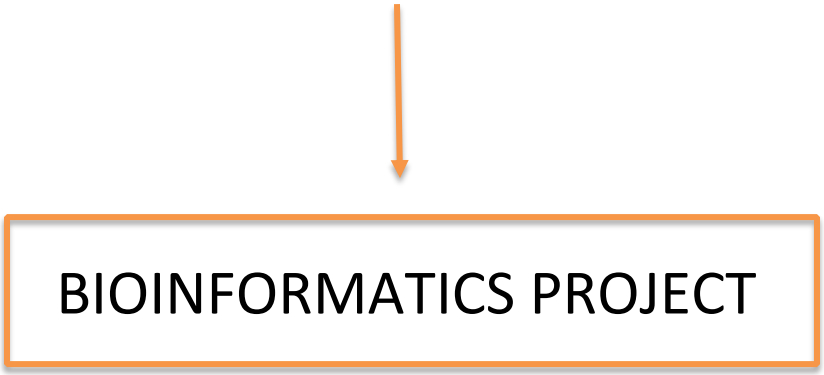
infernal home | rfam database | eddy lab | janelia farm

Mariotti et al. 2013

Selenoproteins as **test case**

- Selenoproteins have the peculiar characteristic of possessing a **UGA codon**, recoded because of the presence of the **SECIS element**.

- If you learn how to predict selenoproteins, you are able to do the same with any "*standard*" protein family.



Find all **selenoprotein-related genes**

in a **vertebrate genome**

UPF Human Biology.
Bioinformatics Courses 2007-2019

2007/08 – 2008/09: find all selenoproteins in a given protist genome
2009/10 – 2011/12: find a given selenoprotein family in all protist genomes
2012/13 – **2019/20**: find all selenoproteins in a given **vertebrate** genome

http://bioinformatica.upf.edu/

Projectes de l'assignatura de Bioinformàtica

Facultat de Ciències de la Salut i de la Vida

Universitat Pompeu Fabra

Curs 2012/2013

1A: Ailuropoda melanoleuca *AM. Barrios, A. Bellot,* *S. Castany, M. De Manuel*	**1B: Cricetulus griseus** *J. Fernandez, J. Gomez,* *FD. Jurquiza, A. Lopez*	**1C: Mustela putorius furo** *M. Perez, L. Taberner,* *G. Vilajosana, I. Villate*
2A: Nomascus leucogenys *M. Alemany, H. Costa,* *A. Escrig, I. Gafarot*	**2B: Saimiri boliviensis** *P. Garcia, J. Latorre,* *R. Martinez, H. Palma*	**2C: Sarcophilus harrisii** *G. Rodriguez, E. Ros,* *AM. Saludes, H. Xicoy*
3A: Chrysemys picta bellii *C. Bitlloch, G. Clua,* *J. Domingo, P. Gelabert*	**3B: Meleagris gallopavo** *J. Jancyte, L. Mateo,* *A. Olle, M. Perera, C. Perez*	
4A: Pelodiscus sinensis *SU. Abad, A. Almeyda,* *A. Azagra, R. Bartomeus*	**4B: Gadus morhua** *O. Bover, N. Cortell,* *B. Grau, E. March*	**4C: Latimeria chalumnae** *A. Martinez, A. Perlas,* *T. Robert, S. Walsh*

Projects 2019-2020
selenoproteins in vertebrates

http://bioinformatica.upf.edu/

- **Web page**: Structure of a scientific paper
- **Wikipedia**: Species description

  https://ca.wikipedia.org/wiki/Viquiprojecte:Curs_Bioinform%C3%A0tica_UPF_2018
- **SelenoDB**: Insert your selenoprotein genes predictions into a real world database. Available to the scientific community.

Selenoproteins in *Miichthys miiuy*

Selenoproteins are a group of proteins characterized by the presence of, at least, one Selenocysteine (Sec) residue in its chain. Since this residue is codified by UGA, which is normally considered as a stop codon, some of this proteins are dismissed in genome databases.

Moreover, the inclusion of Selenocysteine residue depends on the presence of an element called Selenocystein Insertion Sequence (SECIS), which is a secondary mRNA structure that allows the insertion of a selenocysteine instead of a stop codon.

The aim of our study is to predict the selenoproteins of *Miichthys miiuy*, a Japanese benthic fish, performing an homology-based in silico search. In order to assess the characteristics of the *Miichthys miiuy*'s selenoproteome, we have compared the genome of this species with *Danio rerio*'s and *Homo sapiens*'s selenoproteins annotations obtained from SelenoDB. For the prediction, different bioinformatic tools such as BLAST, Exonerate, Genewise, T_coffee, Seblastian and SECISearch3 were needed. Additionally, we have designed an automatic program to speed up the process.

Our results show a high conservation between Zebrafish' and *Miichthys miiuy*' selenoproteome. We have found 33 selenoproteins, 8 Cys-containing homologous proteins, 5 machinery proteins and 11 proteins related to selenium metabolism.

This study contributes with the identification of selenoproteins in new-sequenciated organisms.

Portada
Article a l'atzar
Articles de qualitat

Comunitat
Portal viquipedista
Canvis recents
La taverna
Contacte
Xat
Donatius
Ajuda

Eines
Què hi enllaça
Canvis relacionats
Pàgines especials
Enllaç permanent
Informació de la pàgina
Element a Wikidata
Citau aquest article

Imprimeix/exporta
Crear un llibre
Baixa com a PDF

Miichthys miiuy

Miichthys miiuy és una espècie de peix de la família dels escènids i de l'ordre dels perciformes.

Contingut [amaga]

Morfologia [modifica | modifica el codi]

Els masclesc poden assolir 70 cm de longitud total.[5][6] Com la resta de peixos de la família Sciaenidae, *M. miiuy* és conegut per tenir uns otòlits excepcionalment grans que els doten d'un sistema auditiu molt desenvolupat.[7] Aquests peixos s'anomenen sovint peixos tambors o corballs a causa dels sons que produeixen amb les seves bufetes natatòries.

Hàbitat [modifica | modifica el codi]

És un peix de clima temperat i demersal que viu entre 15-100 m de fondària.[5][8] Eviten les aigües clares, prefereixen viure en estuaris, badies i riberes de rius fangosos. Són organismes carnívors bentònics.[7]

Miichthys miiuy	
Taxonomia	
Super-regne	Eukaryota
Regne	Animalia
Fílum	Chordata
Classe	Actinopterygii
Ordre	Perciformes
Família	Sciaenidae
Gènere	*Miichthys*
Espècie	***Miichthys miiuy*** (Basilewsky, 1855)[1][2][3]
Nomenclatura	
Sinònims	• *Argyrosomus miiuy* (Basilewsky, 1855) • *Miichthys imbricatus* (Matsubara, 1937) • *Nibea imbricata* (Matsubara, 1937) • *Otolithus fauvelii* (Peters, 1881) • *Sciaena miiuy* (Basilewsky, 1855)[4]


SelenoDB
Release 2.0
Home
Statistics
Download
Documentation
Help
Author
email address
Specie
-- select a specie --
+ Add New Specie
Gene name
-- select a gene family --
Forward
Reverse
Promoter
Promoter start
Promoter end
Exons
Exon start
Exon end
+ Add
Protein
Sec / U
Protein start
Protein start
Secis
Secis start
Secis end
+ Add
Residue
Sec / U
Residue start
Residue end
+ Add
Sequence
CHR
Sequence start on the c
Sequence end on the cl
ENS00075
Ensembl
http://www.ensembl.org
http://www.ensembl.org
Choose File
No file chosen
gff format
Fetch Sequence
Submit
Home
Cite
License
Disclaimer
Faq
Contact us

Notes for the project

- Results must be presented in a **web page** with the **structure of a scientific paper**
  - ✓ Aminoacid sequences; SECIS sequences
  - ✓ Genes in GFF format

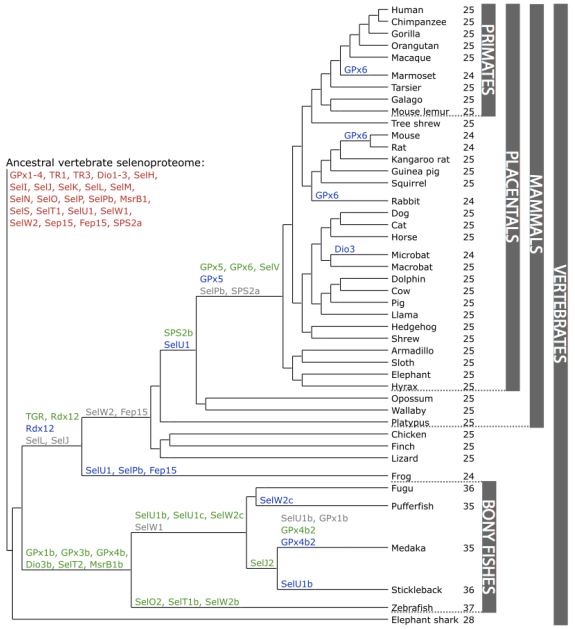
- All **genes** should be **as complete as possible**: starting with a AUG, ending with a STOP codon, and with an identified SECIS element downstream.
- **Ignore alternative isoforms** (if any), just choose one
- Report also the **genes** of:
  - ✓ **selenoprotein machinery**: SEPSECS, EEFSEC, PSTK, SBP2, SECP43, SEPHS1, SEPHS2.
  - ✓ **Cys-containing homologs**

- **Other helpful resources** to biologically interpret and visualize the results (**phylogenetic trees**):

- phyloT: https://phylot.biobyte.de/ (from NCBI taxonomy → .nw)
- iTOL: https://itol.embl.de/ (.nw)
- Etetoolkit: http://etetoolkit.org/treeview/ (.nw or .msa)
- Phylogeny.fr: http://www.phylogeny.fr/simple_phylogeny.cgi (.mfa)

Common pitfalls

- Know what to **expect**
- **Zero, one or many genes?**
  (!) careful with superfamilies and gene duplications
- **Genomic** context

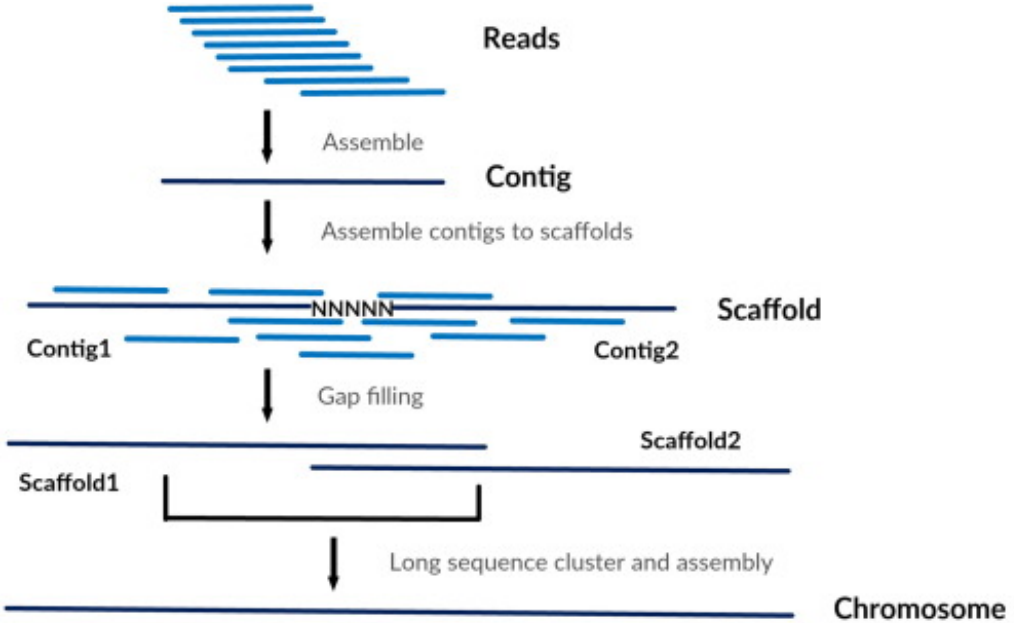


Common pitfalls

- **Contigs** and **Scaffolds**

✓ **Contig**: a contiguous stretch of nucleotides resulting from the assembly of several reads

✓ **Scaffold**: several contigs stitched together wit NNNs in between



Evaluation

The projects will be **evaluated** based on:

- ✓ **Methods:** scripting is encouraged (different levels of automation)
- ✓ **Results:** you are expected to find all selenoprotein-related genes in your assembly
- ✓ **Discussion:** interpret your results logically
- ✓ **Presentation:** the web page should present the work as clearly as possible (including Wikipedia entry)