UPF Bioinformatics Course Projects

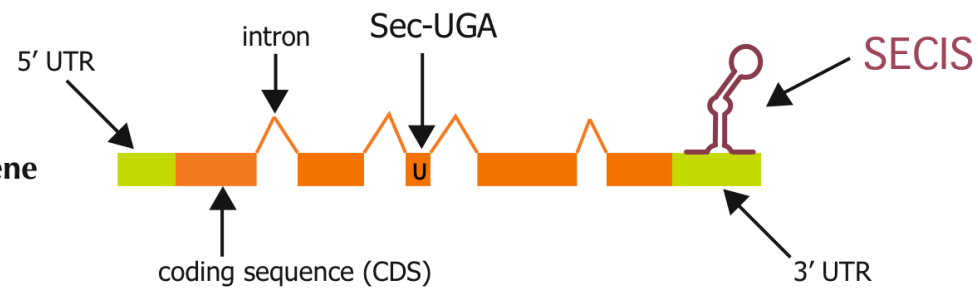
- Students guide 2019/2020 -

Aida Ripoll (PhD Student)
Didac Santesmasses (PhD)

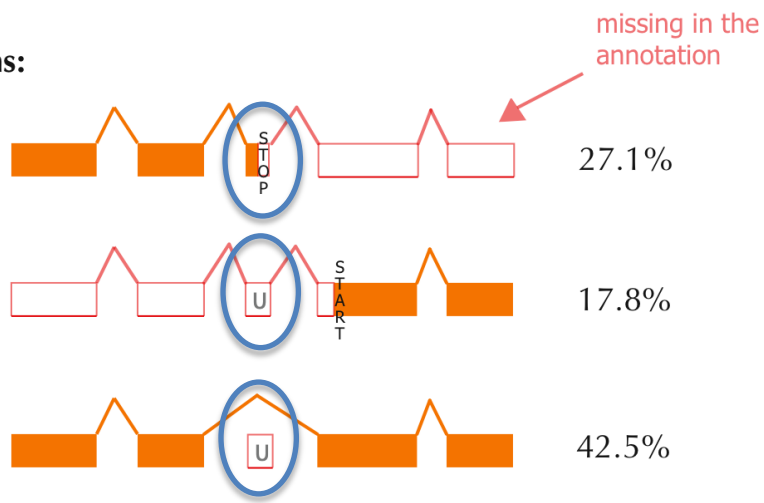
Bioinformatics and genomics programme
Roderic Guigó's group
Centre for Genomic Regulation, Barcelona

Selenoproteins are **generally misannotated**

Example of selenoprotein gene



Typical coding sequence misannotations:



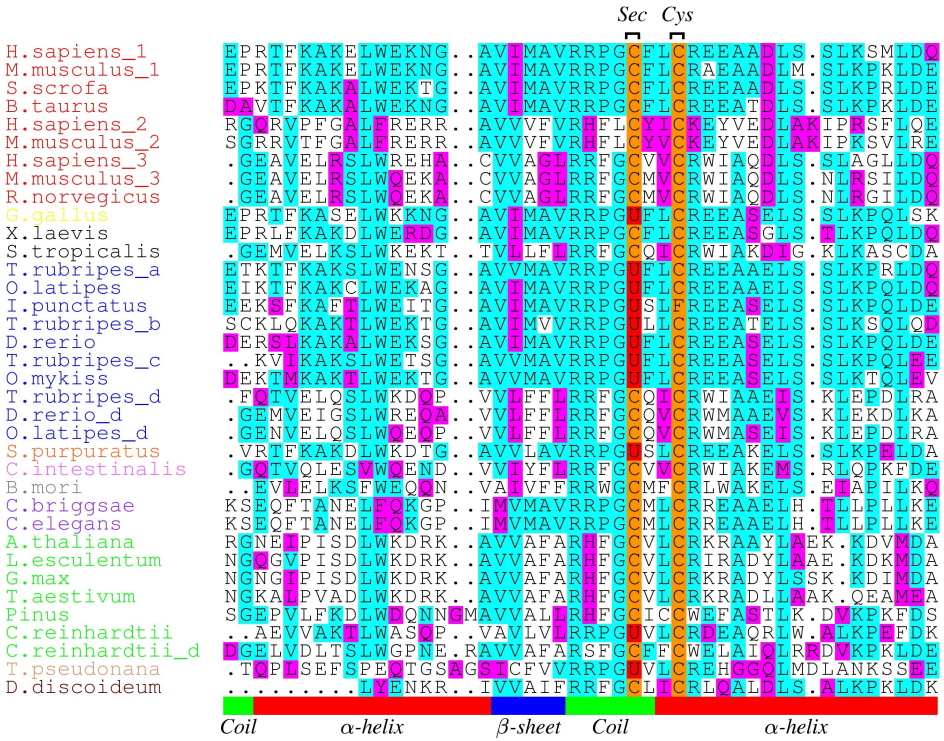
Percentages are computed by comparison Selenoprofiles-Ensembl annotations – see Mariotti and Guigó, 2010 - Bioinformatics.

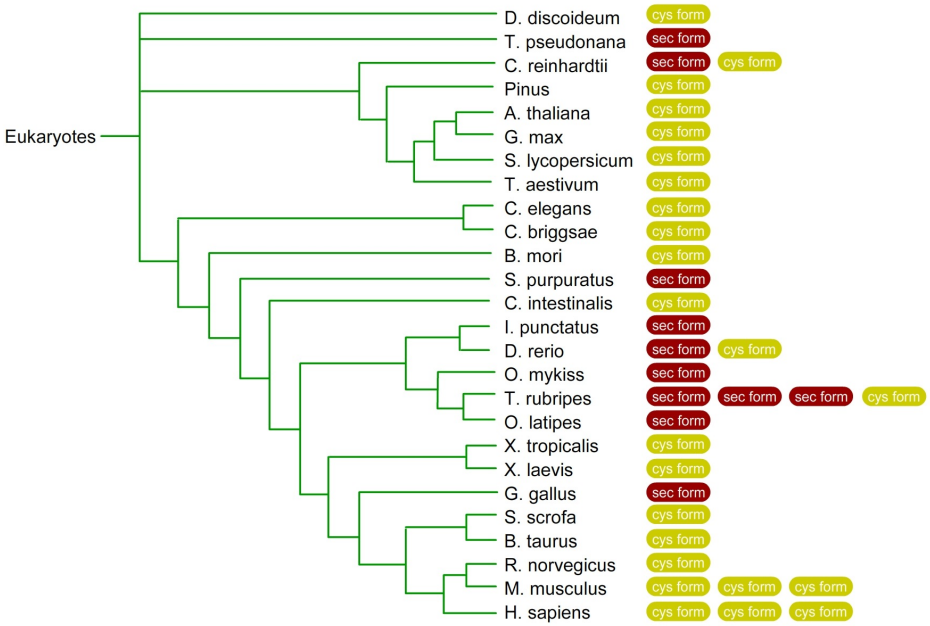
Selenoprotein families include:

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

- **Cysteine homologues** (Cys-containing proteins)

- orthologues
- paralogues

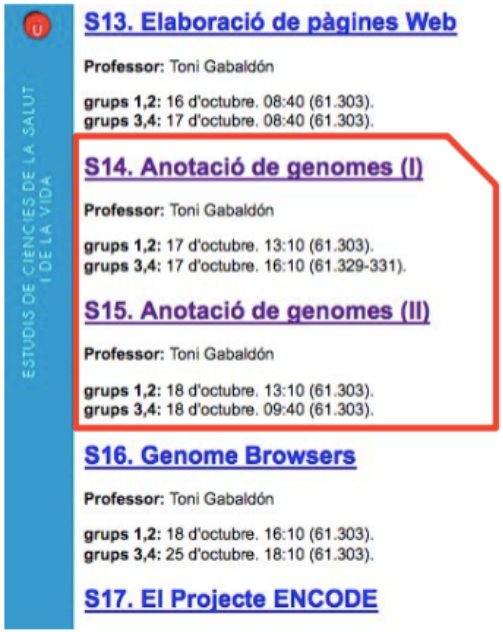




Protocol overview

Tools:

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



http://bioinformatica.upf.edu/

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

  http://seblastian.crg.es/

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/

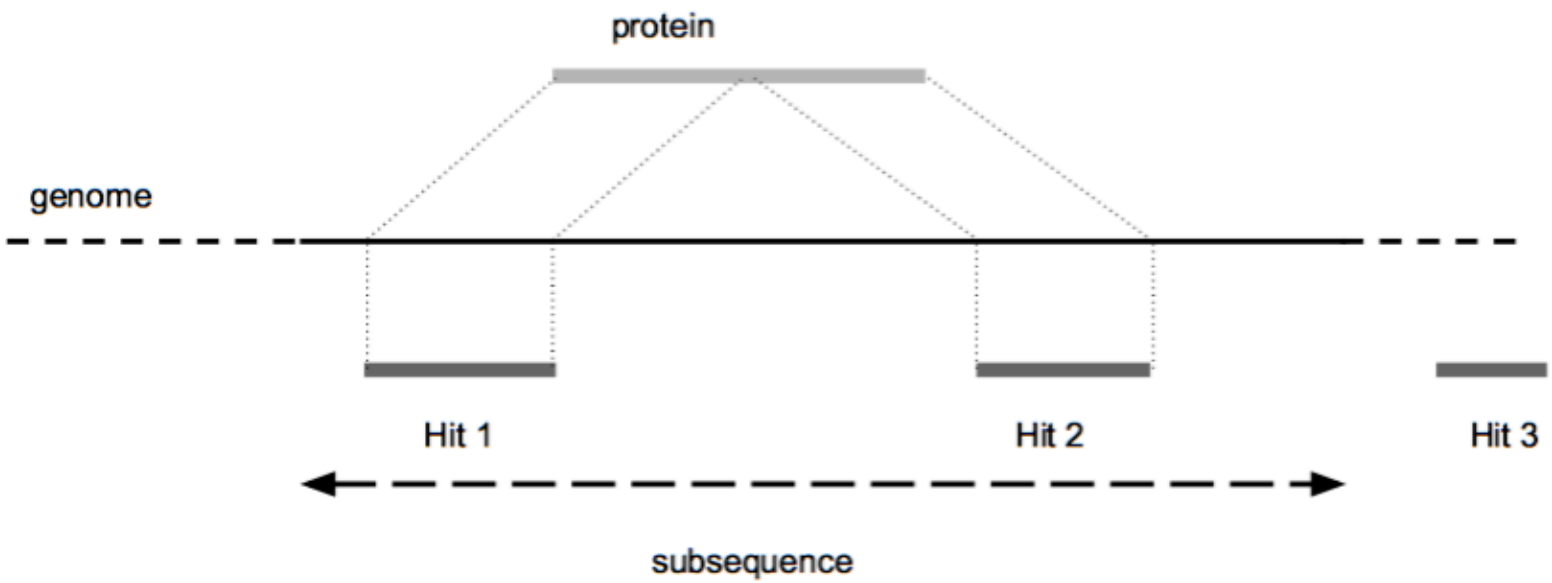
  http://www.uniprot.org

- **Past year projects:**

  http://bioinformatica.upf.edu/

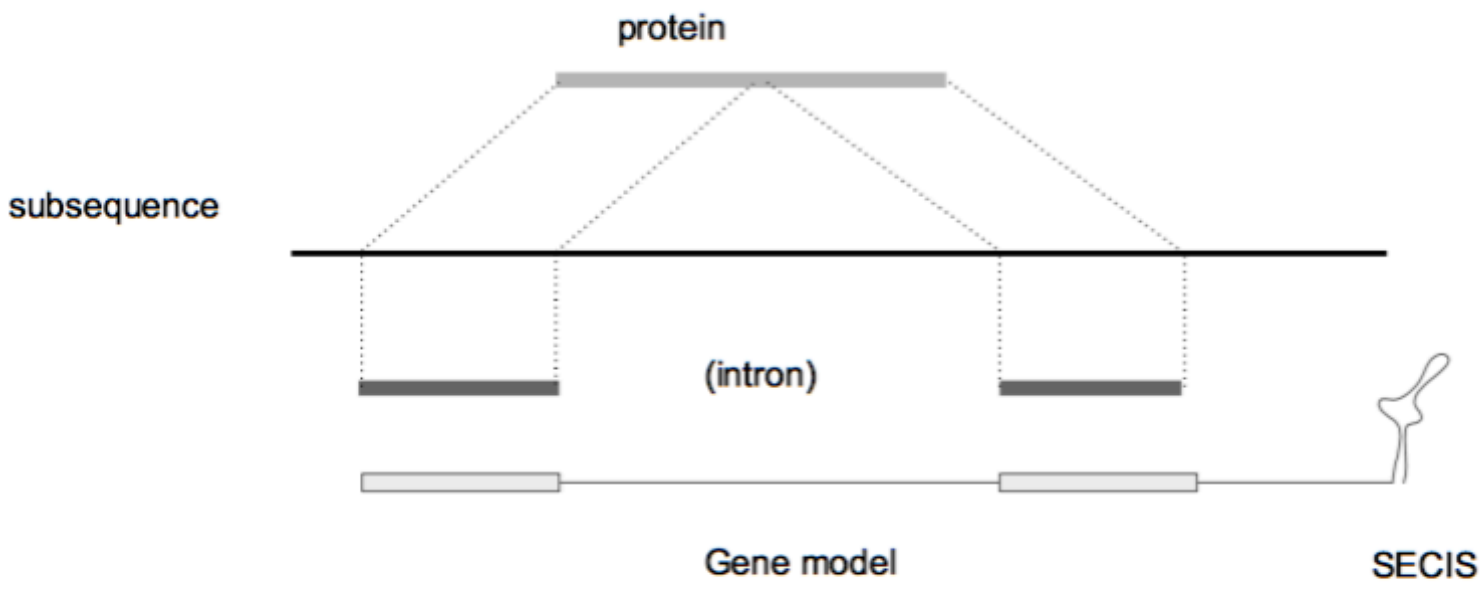
Protocol steps

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

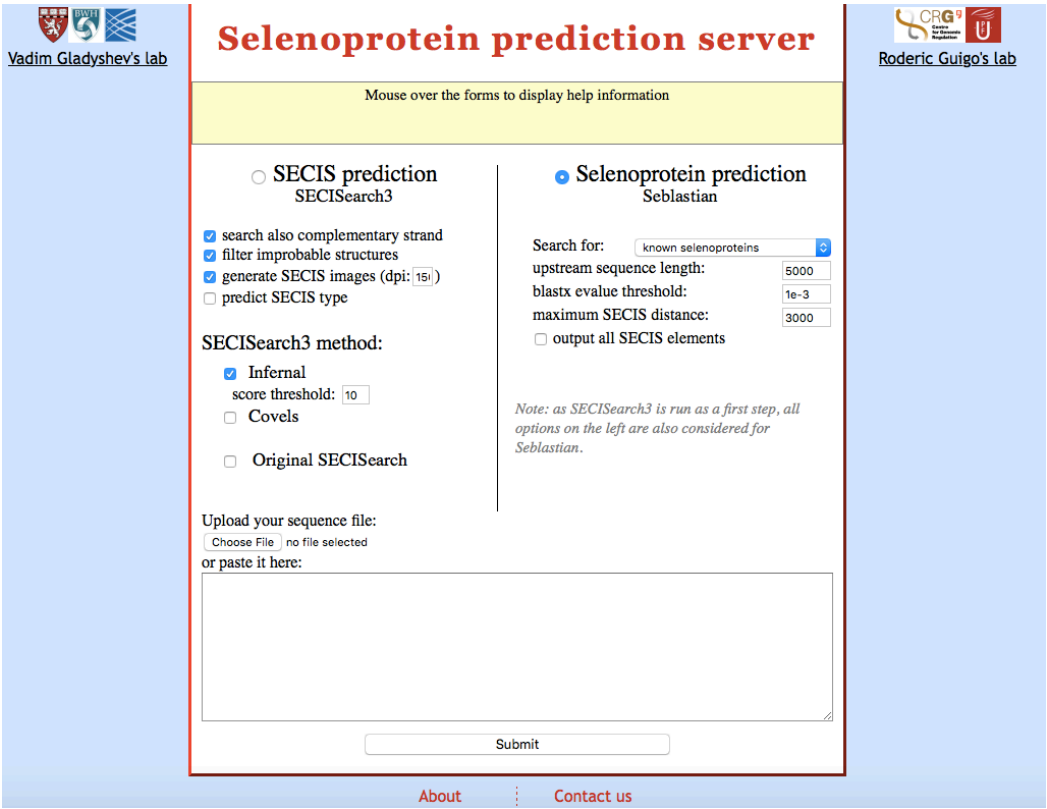
- **Fastatranslate:** translate coding sequences careful with the selenocysteine codon character! It is a good idea to substitute the "*" with "X" or "U" as multiple sequence alignment programs just ignore "*"

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*).

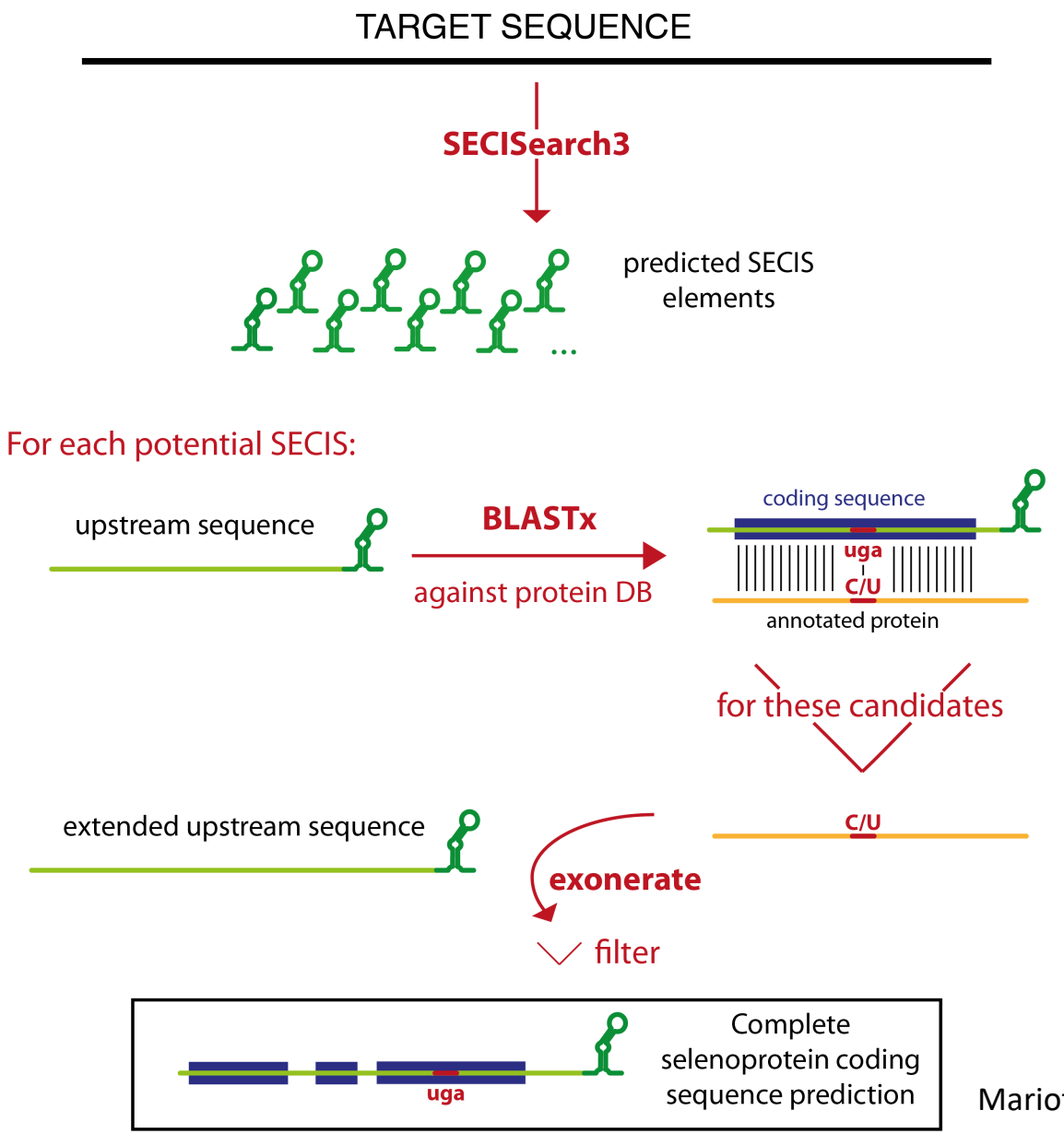
Protocol steps

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



http://seblastian.crg.es/

Seblastian



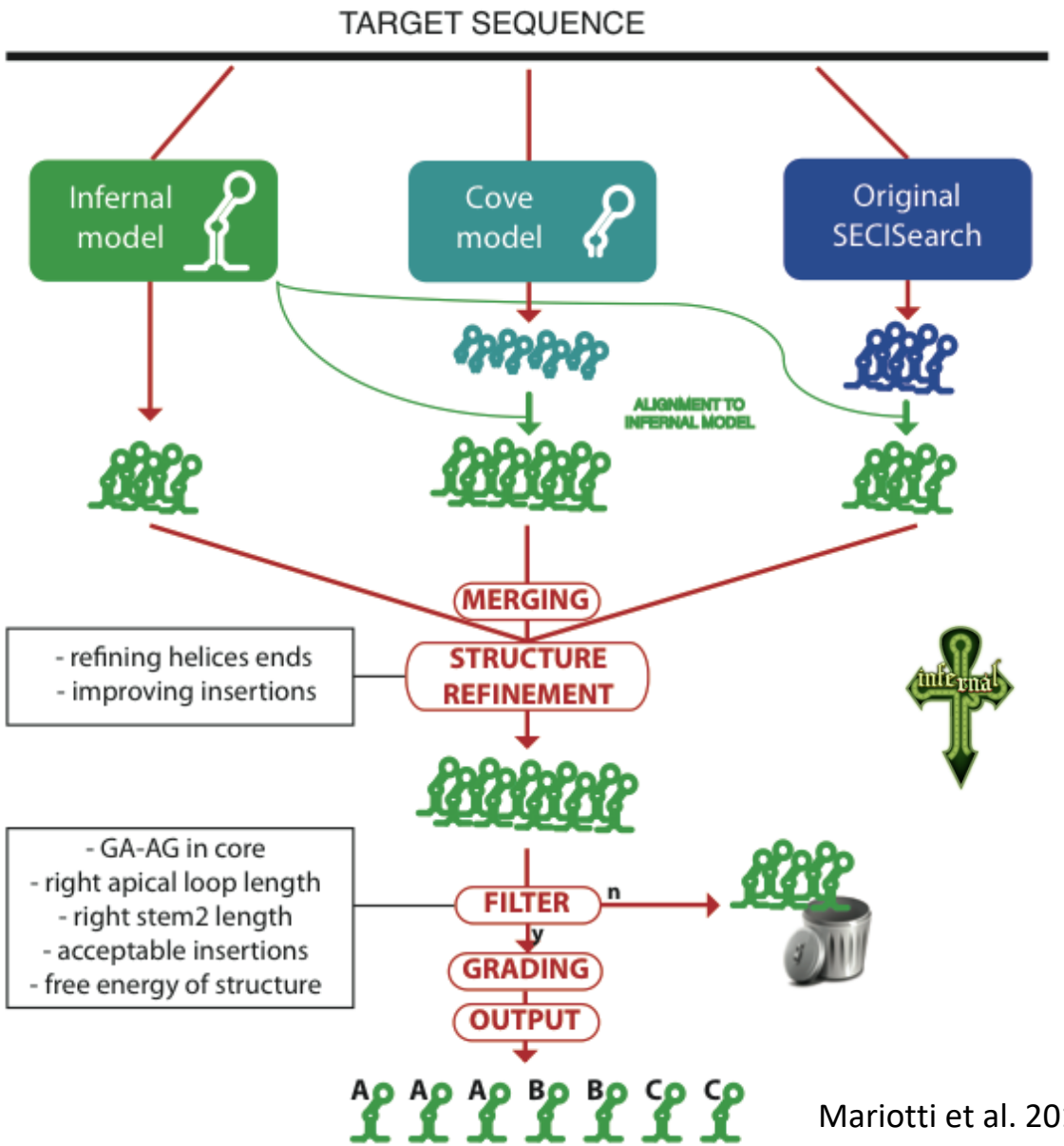
Assumptions:

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

UPF Human Biology.
Bioinformatics Courses 2007-2020

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
- **Oral presentation** (05/12/2019)

Notes for the project

- Results must be presented in a **web page with the structure of a scientific paper**

✓ **Protein** sequence (+SECIS elements)

✓ **Genes** in gff format -absolute coordinates-

- 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 as query.

- Report also the **genes** of **selenoprotein machinery**: SecS, eEFsec, pstk, secp43, SBP2, SPS1, (SPS2).

- In some cases, the predicted protein can be **located** in **more than one contigs/scaffolds**. You will notice this if you try to predict the protein in both of them, and you pay attention at both MSA performed by T-coffee.

Notes for the project

- **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)

- **HTML language** (Web page)

✓ https://www.w3schools.com/html/default.asp

✓ https://getbootstrap.com/

Notes for the project

- We **already provide** you, together with the genome:
  - ✓ BLAST data base for the genome
  - ✓ Indexed genome

- **Scaffolds/Contigs lengths** can be found in *genomes.lengths* file [/NFS_UPF/soft/genomes/2019/***Genus_specie***/*genome.fa*]

- **Fastatranslate** (option -F 1) to consider only the 1st ORF.

- Before performing the **Multiple Sequence Alignment (MSA) with T-Coffee**, substitute the "*" with "X" or "U" as multiple sequence alignment programs just ignore "*"

- **Seblastian** and **SECISearch3 web servers:**

  Input: Nucleotide sequence (*fastasubseq* file)

  * DO NOT take into account other nucleotide bases different than A, C, G, T, a, c, g, t, or N. Then, in case you have one of the other symbols from ambiguity code, one solution could be substituting them by an N.

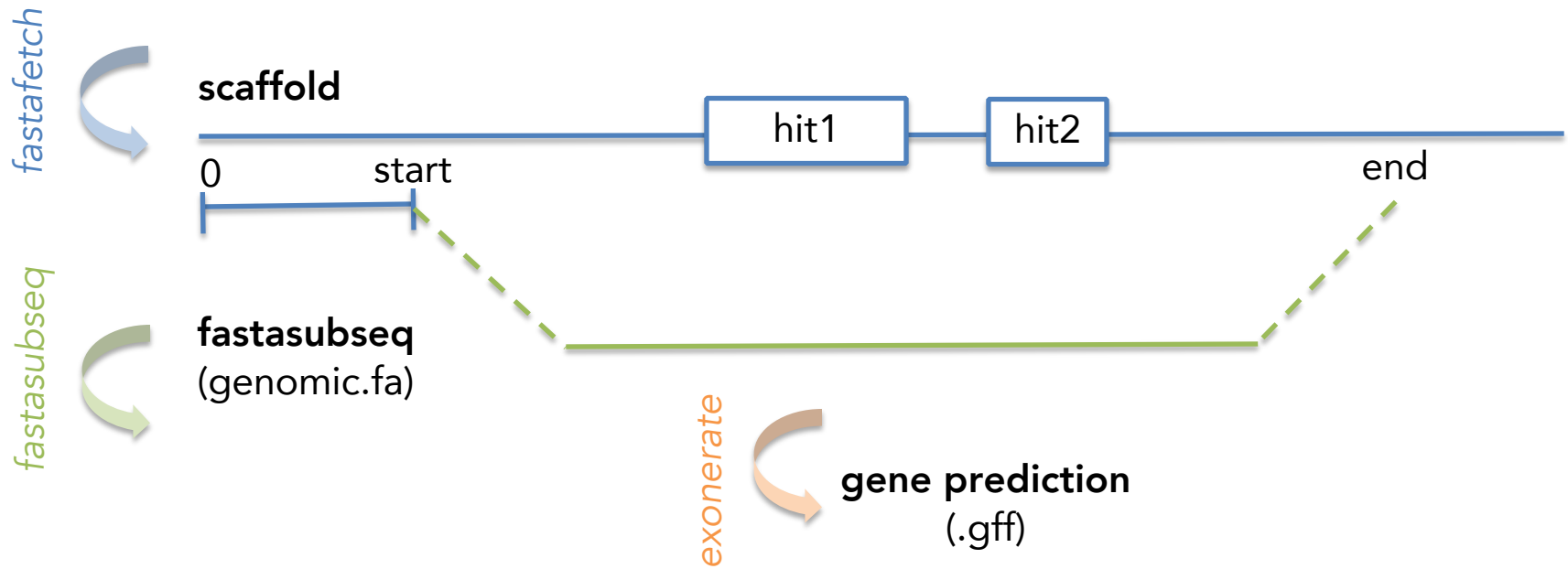
Johnson A.D. An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics.* 2010; 26(10): 1386-1389.

Notes for the project

- **Genes prediction (GFF format):** Conversion of **relative** to **absolute coordinates**

Apart from obtaining the protein sequence predictions, you should obtain the gene predictions in .gff format considering the absolute coordinates.

Remember that, as you made your prediction using the *fastasubseq* file, you will be predicting the genes (.gff file from exonerate) with the relatives coordinates instead of the absolute coordinates. Then, to generate the .gff files with absolute coordinates, you will have to convert the your .gff files with relatives coordinates (.gff file from exonerate) considering the **start** you decided to give to the **fastasubseq program.** [In this case, **start**: start_hit1 nt - 50.000 nt]



About PERL…

- Dealing with **directories** (open, listing the content, etc..) and **files** (*read* –input-, *write* –output-)
  - ✓ https://perldoc.perl.org/functions/open.html
  - ✓ http://perlmeme.org/faqs/file_io/directory_listing.html

- **RegExpr** (Regular Expression): match, substitution of patterns
  - ✓ https://www.tutorialspoint.com/perl/perl_regular_expressions.htm
  - ✓ http://jkorpela.fi/perl/regexp.html

- **Special variables** (e.g., $_)
  - ✓ https://perlmaven.com/the-default-variable-of-perl
  - ✓ https://www.tutorialspoint.com/perl/perl_special_variables.htm

- **Running external program** (e.g., call blast, fastafetch, etc.. from Perl)
  - ✓ https://perlmaven.com/running-external-programs-from-perl

- **Split, Join** and options to print variables
  - ✓ https://perlmaven.com/perl-split
  - ✓ https://alvinalexander.com/perl/edu/qanda/plqa00007.shtml

Common pitfalls

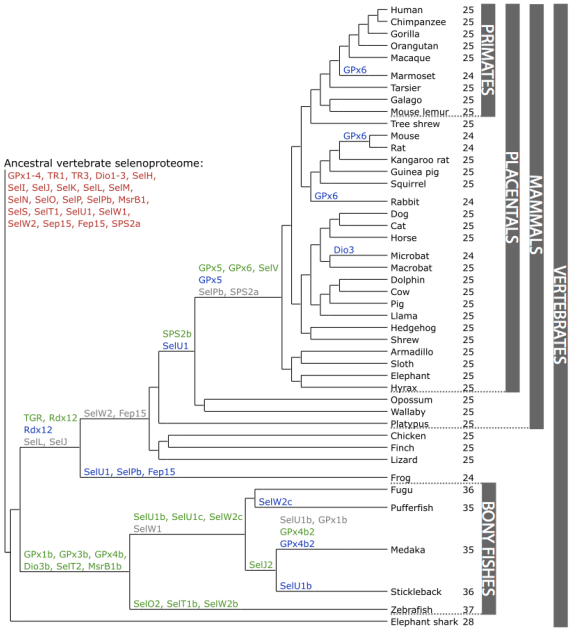
- Know what to **expect**

- Zero, one or many genes?

  Careful with **superfamilies** and **gene duplications**

  (consider the phylogenetic context)

- **Genomic** context

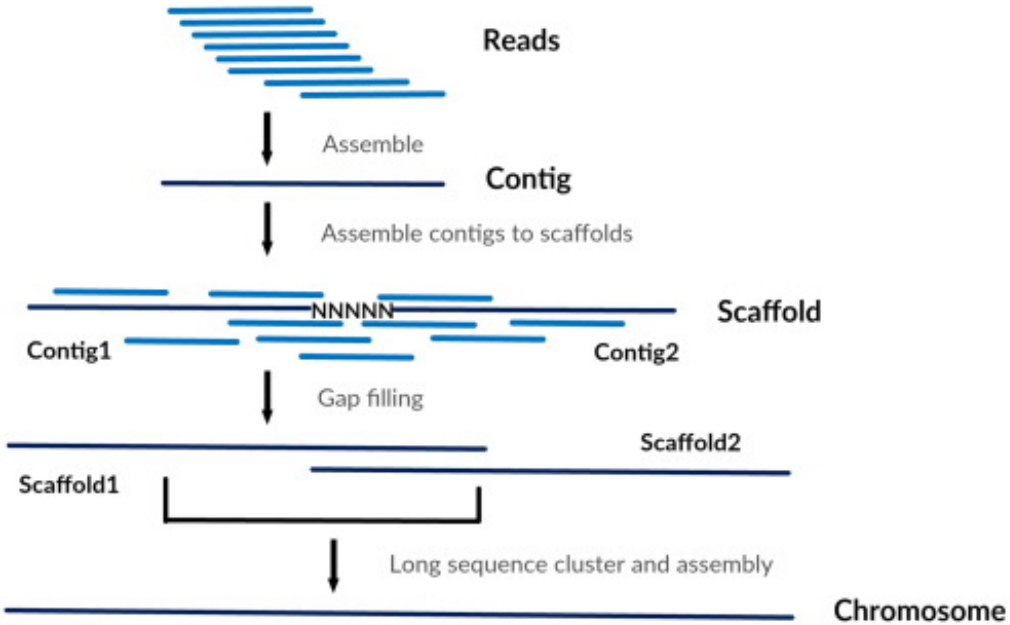


Common pitfalls

- **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



Technical issues

- Genomes (vertebrates) -

[blast formatted and indexed]

in /mnt/NFS_UPF/soft/genomes/2019/***Genus_specie/***

- Access to the FileSystem of the classrooms -

[NO access to the cluster]

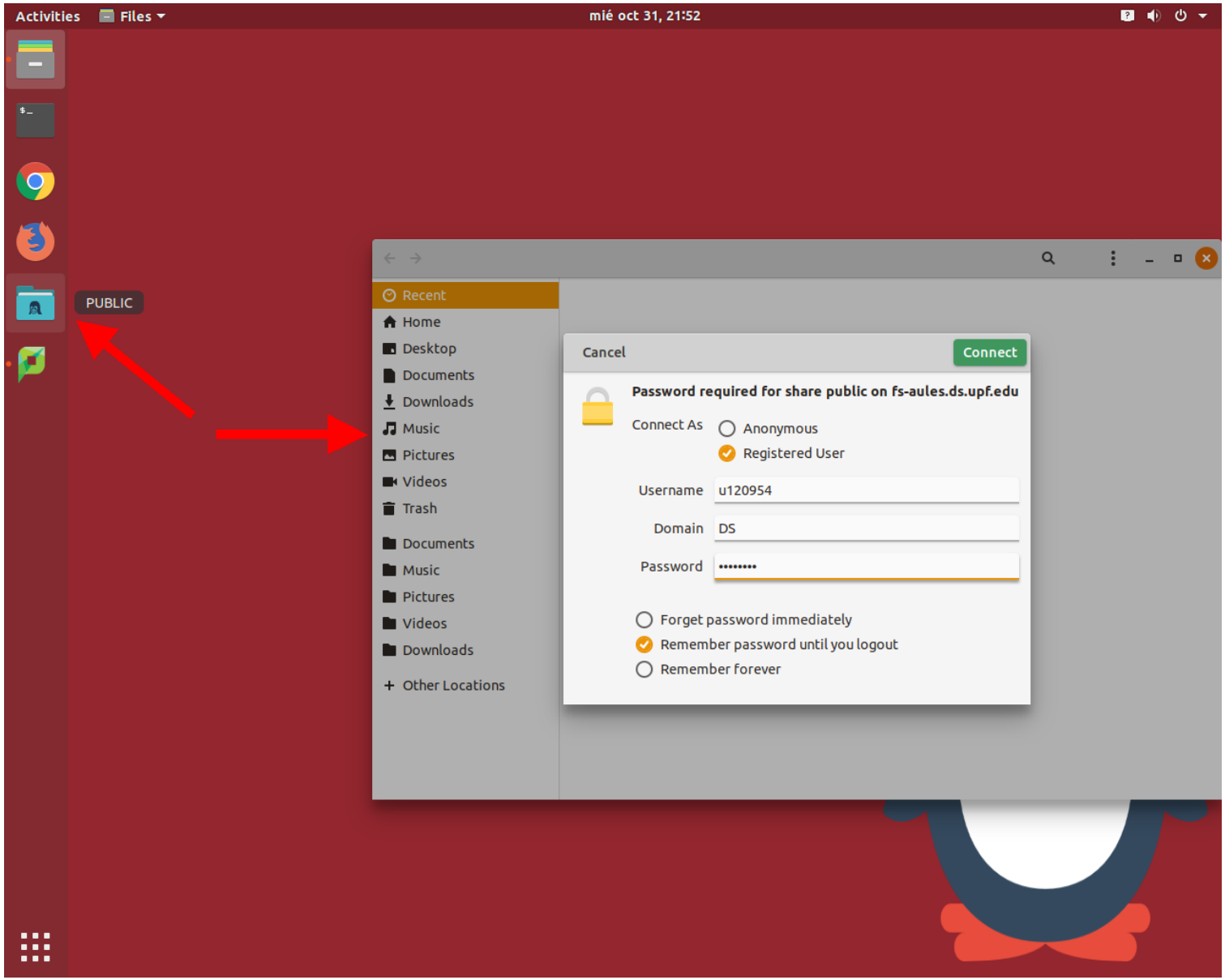
Username: uXXXXXX (UPF identification code UNIS)
Password: DDMMYYYY

Ubuntu: Classrooms

- **File explorer** (Nautilus) → Other locations → Connect to server: smb:**//fs-aules.ds.upf.edu/PUBLIC/**

*** Mount point:** /mnt/NFS_UPF/

Ubuntu → File explorer (Nautilus)



Technical issues

- Softwares -

Installed in the Mount Point (/mnt/NFS_UPF/soft/)

- **NCBI Blast**: /mnt/NFS_UPF/soft/ncbi-blast-2.7.1+/bin/
- **Exonerate:** /mnt/NFS_UPF/soft/exonerate/exonerate-2.2.0-x86_64/bin/exonerate
- **T-Coffee:** /mnt/NFS_UPF/soft/tcoffee/bin/t_coffee
- **fastaseqfromGFF.pl:** /mnt/NFS_UPF/soft/fasta/fastaseqfromGFF.pl

These programs have been linked to the **/bin directory of each student.** So, they do not need to do use the complete path (mentioned above) every time you need to use them.

(!) Genewise: It needs to be installed every time you open your session in the computers of the room:
sudo apt install wise

Groups and Species

Group	Subgroup	Supervisor	E-mail	Specie
Group 101	1	Aida Ripoll	aidaripollcladellas@gmail.com	*Callopanchax toddi*
	2	Edgar Garriga	edgano@gmail.com	*Anarrhichthys ocellatus*
Group 102	3	Diego Garrido	diego.garrido@crg.eu	*Colinus virginianus*
	4	Beatrice Borsari	beatrice.borsari@crg.eu	*Craseonycteris thonglongyai*
Group 103	5	Laura Jimenez	laurajimenez2095@gmail.com	*Mungos mungo*
	6	Toni de Dios	tonidedios94@gmail.com	*Datnioides undecimradiatus*
	7	Toni de Dios	tonidedios94@gmail.com	*Cricetomys gambianus*
	8	Aitor Serres	aitor.serres@upf.edu	*Kobus ellipsiprymnus*
Group 104	9	Miquel Angel Schikora/td>	miki.s.t@hotmail.com	*Laticauda laticaudata*
	10	Veronica De Pinho	veronica.mixao@crg.eu	*Carettochelys insculpta*
	11	Hrant Hovhannisyan	grant.hovhannisyan@gmail.com	*Varanus komodoensis*
Support supervisor	-	Not determined	-	-