JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Examining repeats with databases

Miguel Andrade
Faculty of Biology,
Johannes Gutenberg University
Mainz, Germany
andrade@uni-mainz.de

RepeatsDB

RepeatsDB | Browse | Search | About | Help | Stats | Search | Search

News

RepeatsDB v2.0
2016.10.11 More Info

Start

For a fast search use the top-right search box. Alternatively, visit the browse and search pages

Contact Info

For questions and/or comments, please use the contact form

Citing RepeatsDB

RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures.
Paladin L, Hirsh L, Piovesan D, Andrade-Navarro MA, Kajava AV, Tosatto SC
Nucleic Acids Research 2016
Go to PubMed | Go to NAR

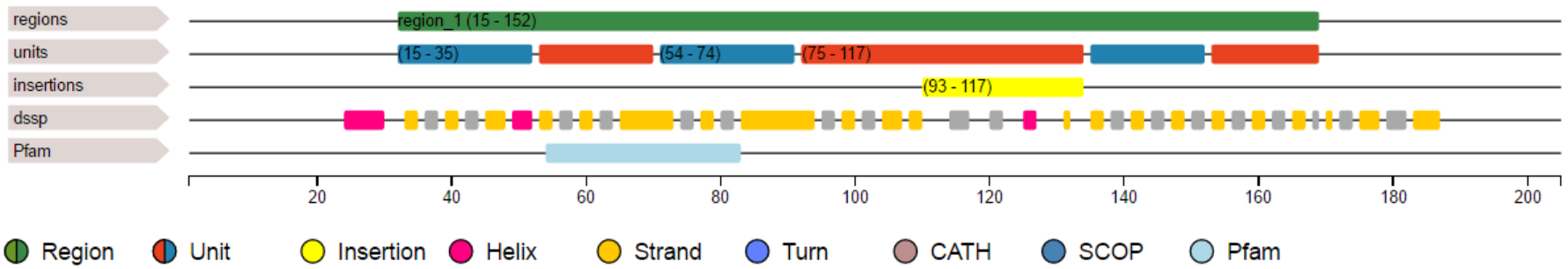
RepeatsDB 2.0

RepeatsDB is a database of annotated tandem repeat protein structures. The database provides unit position, classification and reference to other databases. To start using RepeatsDB, please try the search box (top-right corner), the advanced search or the browse page.

	II	Fibrous structures stabilized by interchain interactions	12
	III	Elongated structures whose repeat units require one another to maintain structure	2388
	IV	Closed structures whose repeat units need one another to maintain structure	2883
	V	'Beads on a string' structures whose repeat units are large enough to fold independently	215
TOTAL			5498

RepeatsDB

http://repeatsdb.bio.unipd.it/protein/3vbpA

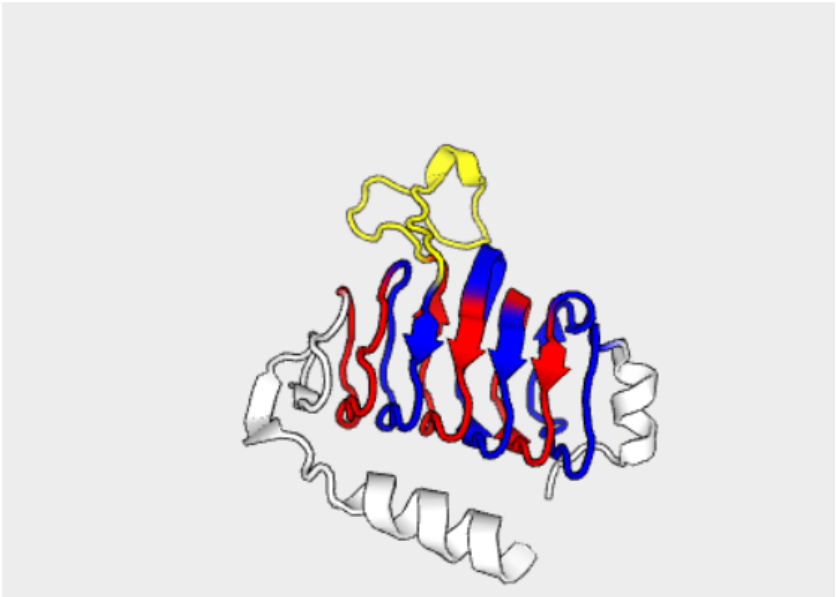


205 Sequence viewer

Search in sequence.. (Reg

```
  1 MGSHHHHHHE NLYFQGHMNS FYSQEELKKI GFLSVGKNVL
 41 ISKKASIYNP GVISIGNNVR IDDFCILSGK VTIGSYSHIA
 81 AYTALYGGEV GIEMYDFANI SSRTIVYAAI NDFSGNALMG
121 PTIPNQYKNV KTGKVILKKH VIIGAHSIIF PNVVIGEGVA
161 VGAMSMVKES LDDWYIYVGV PVRKIKARKR KIVELENEFL
201 KSMNS
```

Structure viewer



Exercise 1. Examining repeat structures in repeatsDB

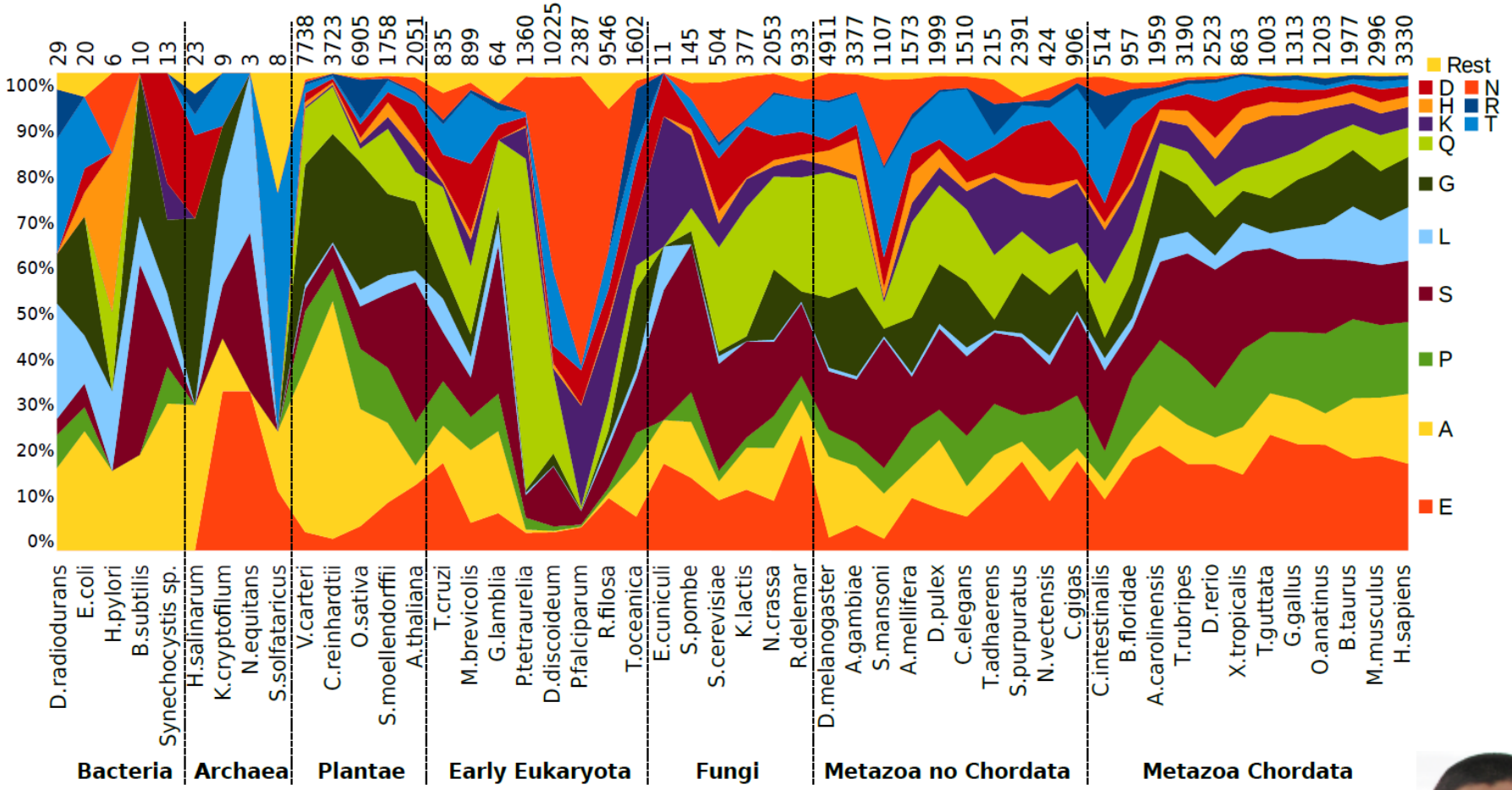
- Go to RepeatsDB: http://repeatsdb.bio.unipd.it/

- Go to Search, then select database RepeatsDB, and Search field, No.units

- Choose one example (you might look for an unreviewed one) and write the name in your card.

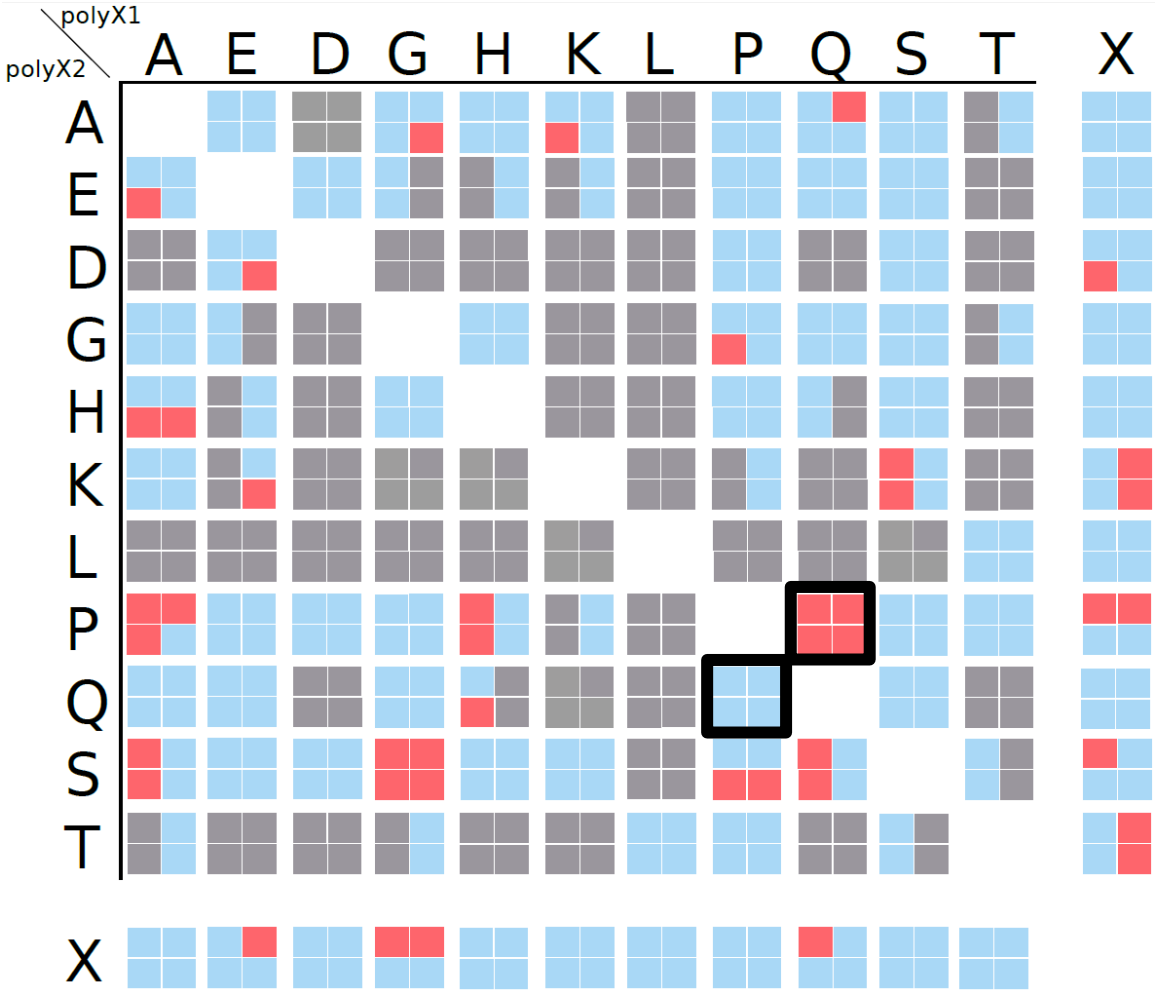
- Check the assignment of the units. Is it correct? What about insertions? Are there mistakes? Write an evaluation in the card: Looks good / one repeat wrong / many mistakes / total mess

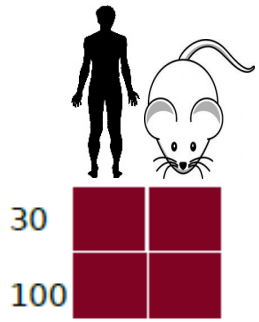
Evolution of homorepeats in 50 species



Pablo Mier

Co-occurrence of homorepeats





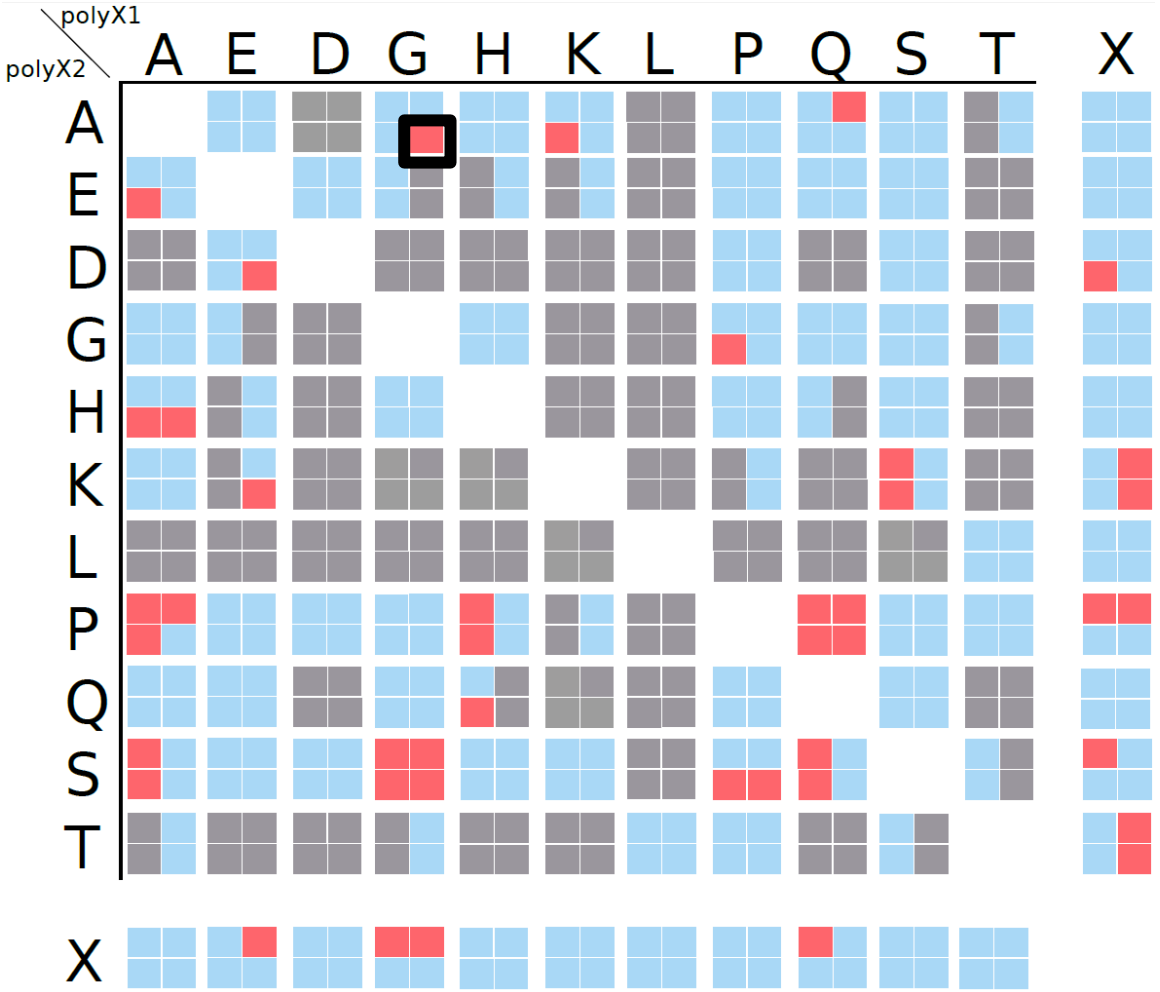
The ocurrence of the polyX pair is not significant

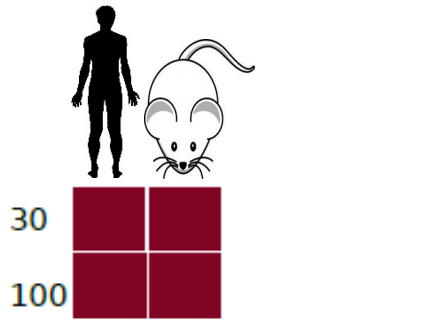
The order is not significant for this distance

The order is significant for this distance

polyQ followed by polyP dependency

Co-occurrence of homorepeats





The ocurrence of the polyX pair is not significant

The order is not significant for this distance

The order is significant for this distance

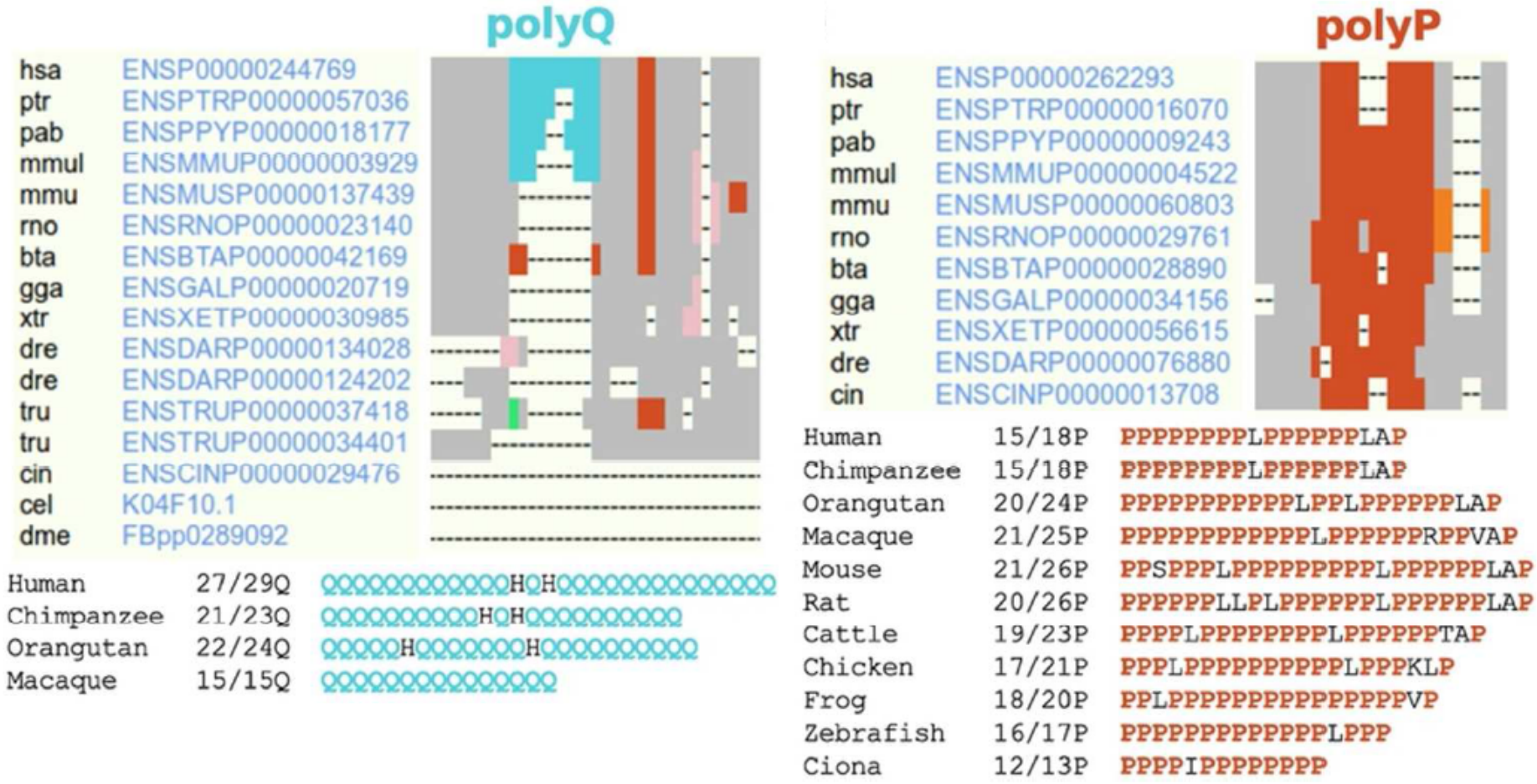
Species specific differences

Context and evolution of homorepeats

dAtabase of **P**olyX **E**volution

Mier *et al.* (2016) Bioinformatics

Context and evolution of homorepeats



dAtabase of **P**olyX **E**volution

Mier *et al.* (2016) Bioinformatics

Exercise 2. Viewing homorepeats in an alignment with dAPE

- Find a human protein starting with the letter in your card. Use UniProt advanced search: Organism “Human”, Entry Name [ID]: “a*” (for a). Write the Entry Name in your card.

- Copy the Entry ID (e.g. P10275).

- Go to the dAPE web page: http://cbdm-01.zdv.uni-mainz.de/~munoz/polyx/

- Use the Entry ID in option A

- Hit the “Report the evolution of its polyX" button

- Which polyX has the human protein? Which other polyX not in the human protein were found in the orthologs? Write these in your card.

Exercise 2. Viewing homorepeats in an alignment with dAPE

*{ **Mandatory** → Choose between **Input option A** and **Input option B** to start the execution of dAPE. }*

dAPE helps assessing the evolution of homorepeats and their protein context. It uses by default a weak cutoff (4 out of 6 identical amino acids) to help in the identification of emerging and disappearing homorepeats.

[Input option A]

Input one **EnsemblProtein ID**, UniProt AC or UniProt ID to get its polyX and their evolution using orthology data.

Name here (Organisms in our database)

**Execution time depends solely on the query length, from one second (~500 amino acids) to around 40 seconds (~4000 amino acids).*

Example 1 → one Ensembl ID (ENSP00000244769, human ataxin 1) as query. Precomputed result.
Example 2 → one UniProt AC (P42858, human huntingtin) as query. Precomputed result.

[Input option B]

Upload a file with one or more **protein sequence/s**, in ***fasta format***.

Datei auswählen Keine ausgewählt

or paste the sequence/s here:

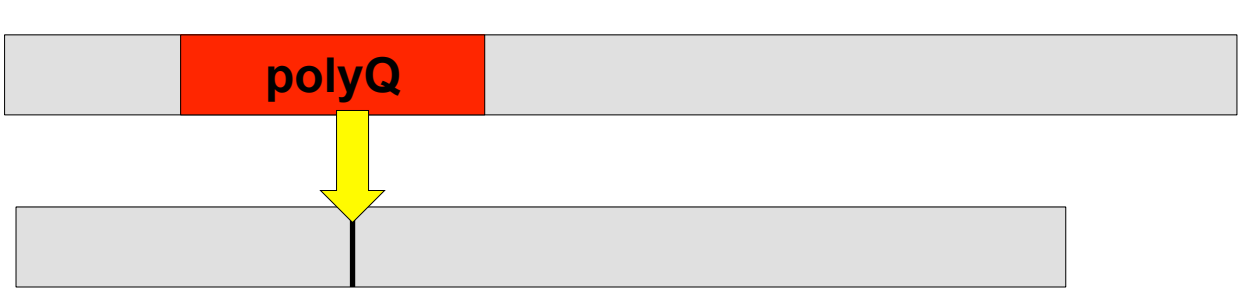
**Only the first 20 sequences will be computed and examined for polyX regions.*

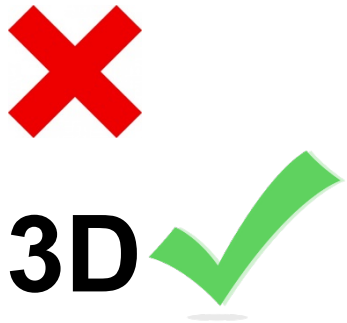
Example 3 → a set of orthologous sequences as input; clustered with **FastaHerder2**, in one cluster.
Example 4 → a set of sequences from different orthologous groups as query. Clustered with **FastaHerder2**, in more than one cluster.

Report the evolution of its polyX!

3D context of polyQ

Franziska Totzeck
Pablo Mier

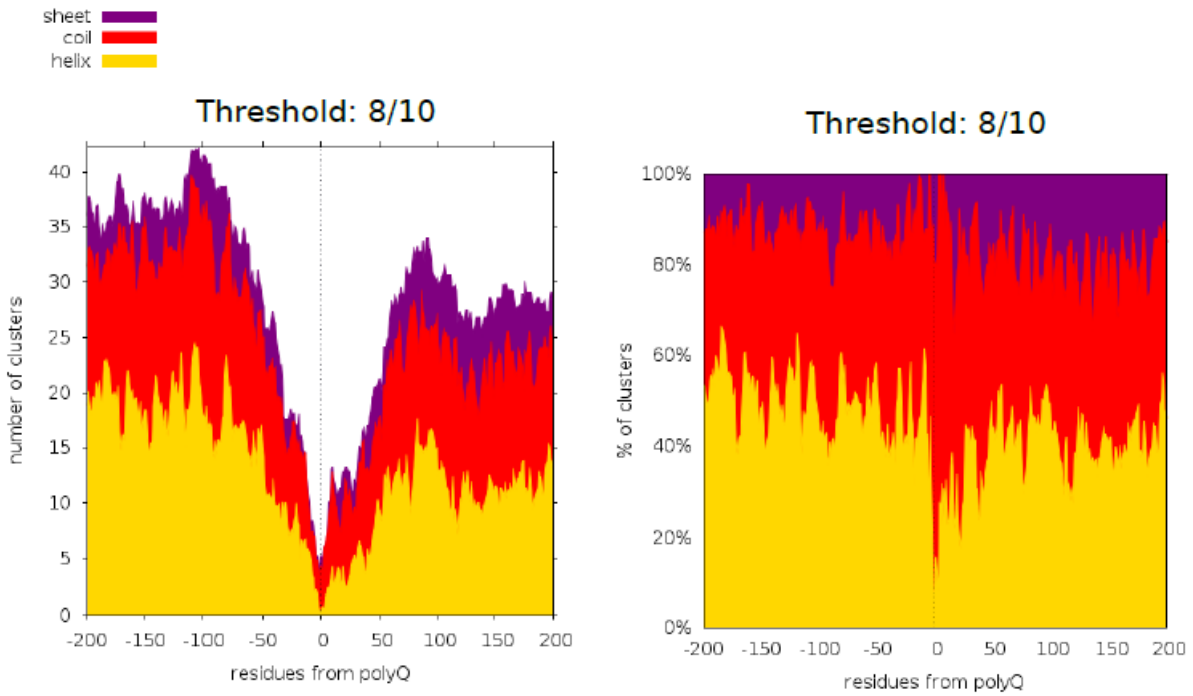




Totzek et al. *PLoS ONE (2017)*

3D context of polyQ

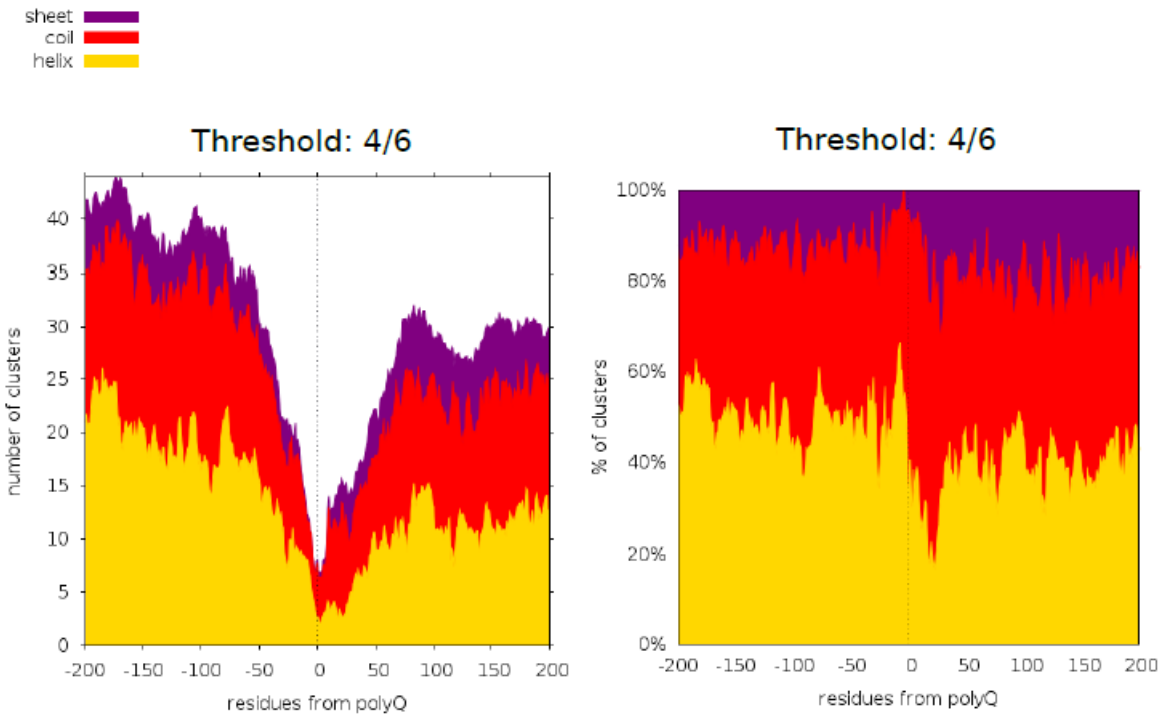
Franziska Totzeck
Pablo Mier



Totzek et al. *PLoS ONE (2017)*

3D context of polyQ

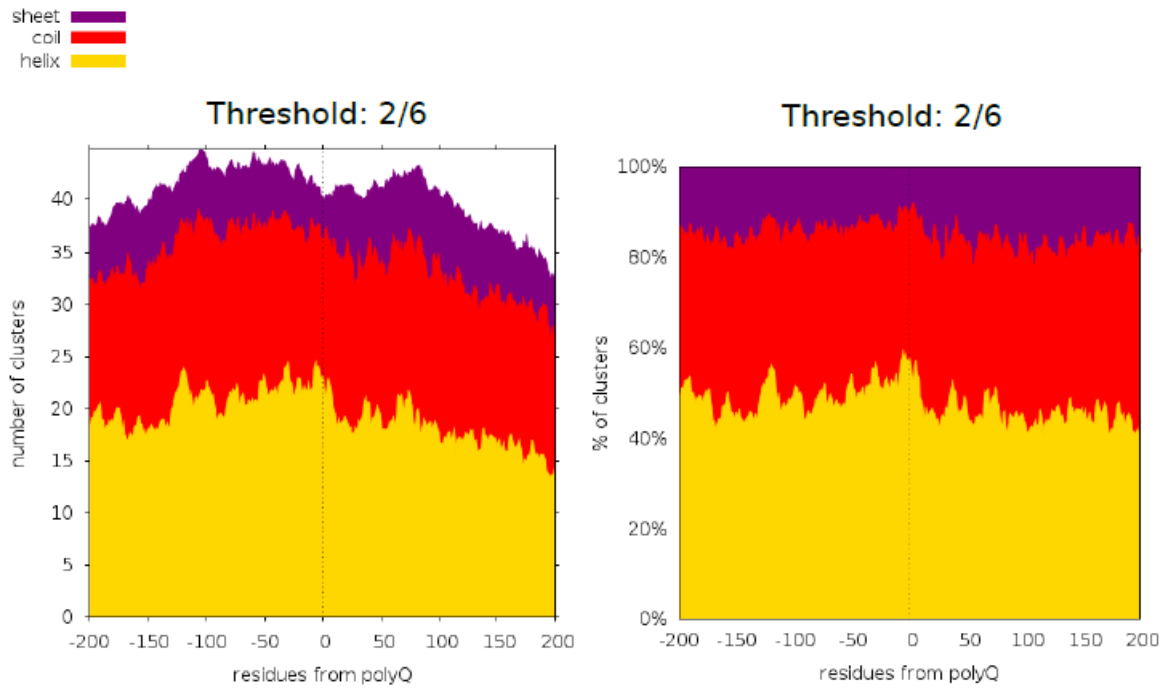
Franziska Totzeck
Pablo Mier



Totzek et al. *PLoS ONE (2017)*

3D context of polyQ

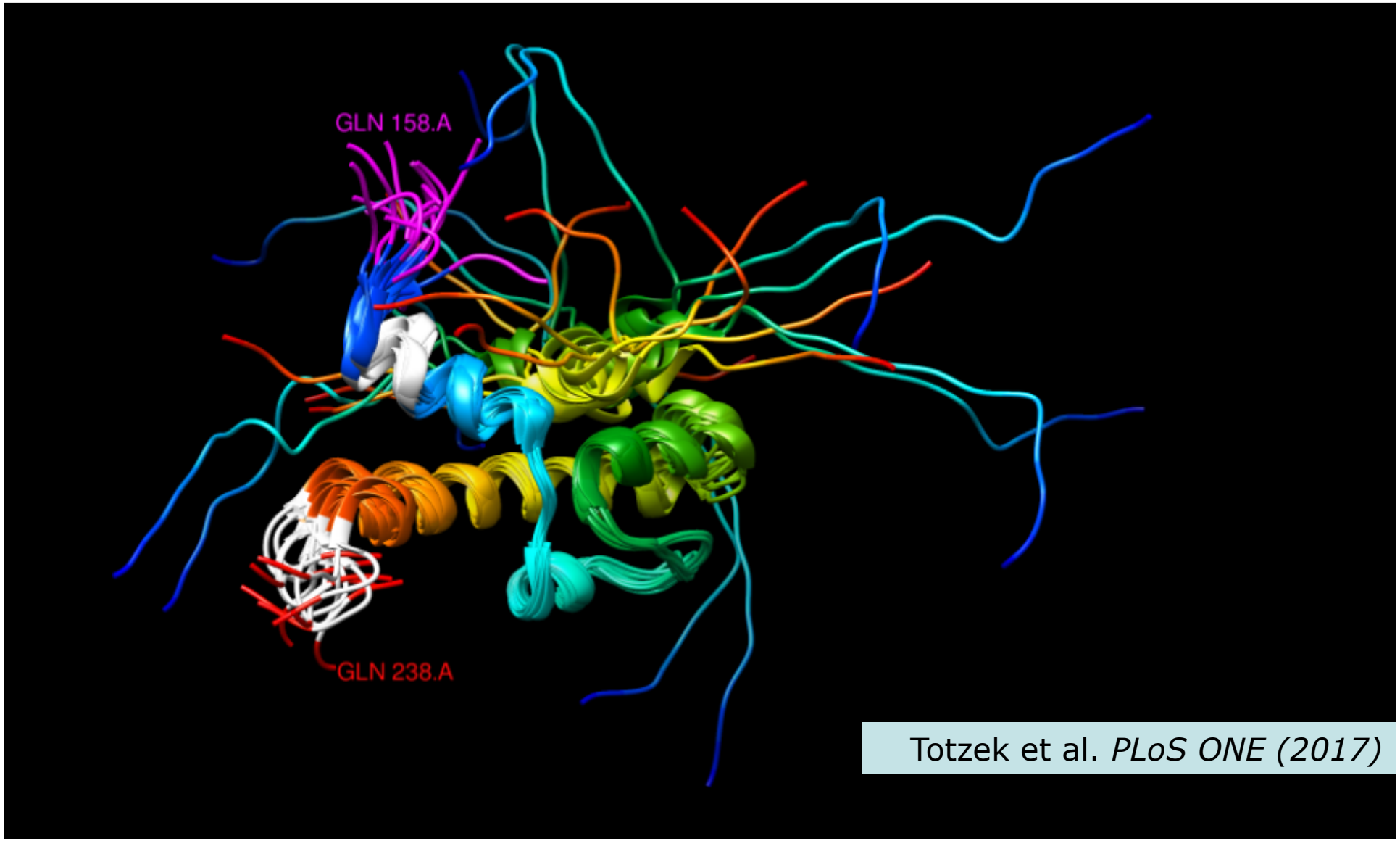
Franziska Totzeck
Pablo Mier



Totzek et al. *PLoS ONE (2017)*

3D context of polyQ

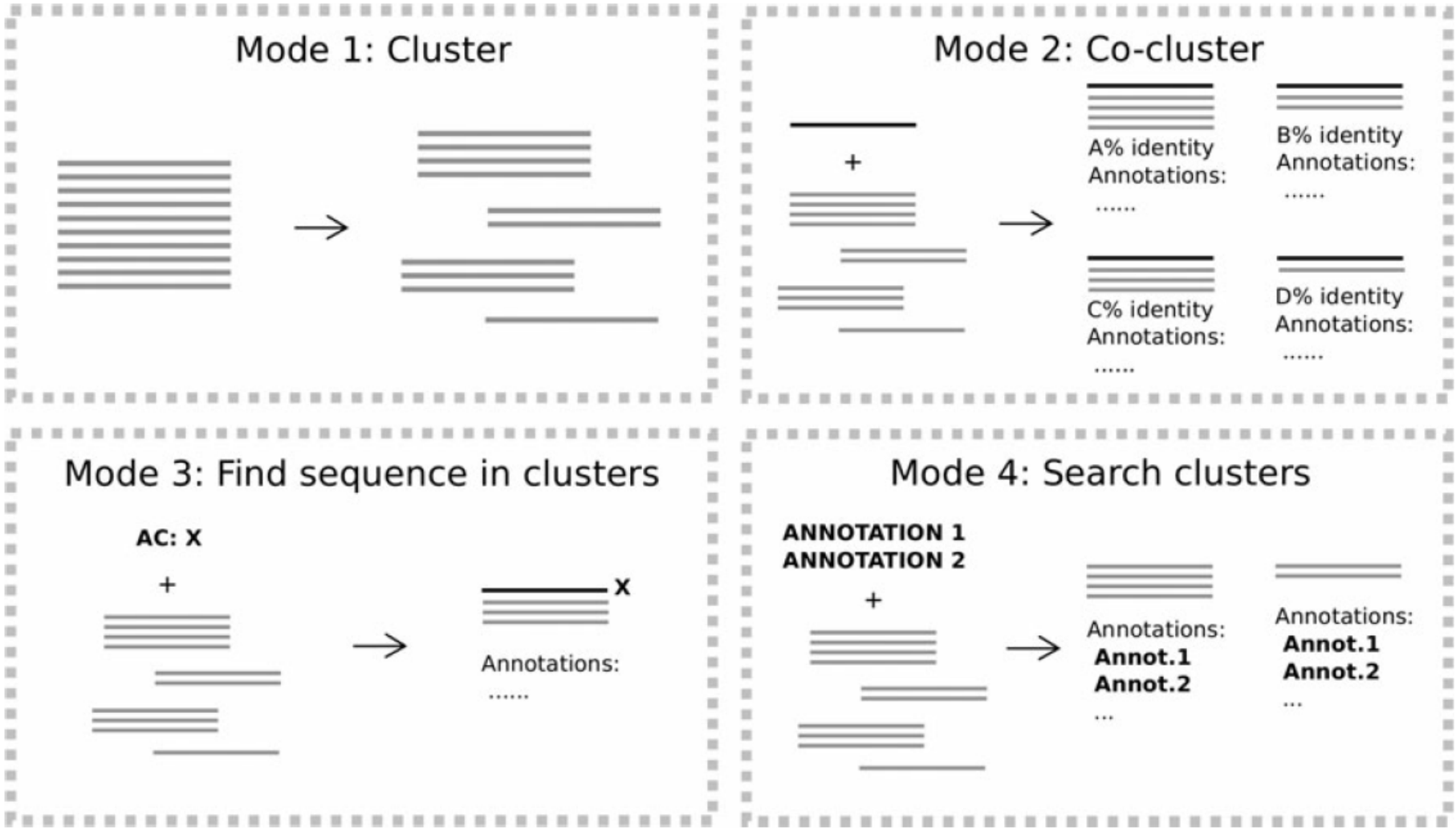
Franziska Totzeck
Pablo Mier



Totzek et al. *PLoS ONE (2017)*

Fasta Herder 2

Pablo Mier



Mier and Andrade (2016) *J. Comp. Biol.*

Fasta herder 2

MODE 4: SEARCH CLUSTERS

SEARCH CLUSTERS using a combination of selected annotations.

Cluster **length subrange** from [ ] to [ ]

Cluster **number of proteins** from [ ] to [ ]

Must the cluster have at least one sequence with...

...PDB annotation? DM

...PMID information? DM

...Pfam information? DM

...polyP regions? DM	**...polyE regions?** DM	**...polyK regions?** DM	**...polyS regions?** DM	**...polyQ regions?** DM
...polyH regions? DM	**...polyD regions?** DM	**...polyR regions?** DM	**...polyG regions?** DM	**...polyA regions?** DM
...polyC regions? DM	**...polyF regions?** DM	**...polyI regions?** DM	**...polyL regions?** DM	**...polyM regions?** DM
...polyN regions? DM	**...polyT regions?** DM	**...polyV regions?** DM	**...polyW regions?** DM	**...polyY regions?** DM

In the cluster, the following **Pfam domain/s...**

MUST be present in at least one protein: [ ] (Pfam domain names separated by "+")

MUST NOT be present in any protein: [ ] (Pfam domain names separated by "+")

In the cluster, ...

there **MUST** be at least one sequence from the following organism/s: [ ] (taxonomic id from an organism, e.g. **9606** for *H.sapiens*, or taxon name, e.g. ***Homo***)*

there **MUST NOT** be any sequence from the following organism/s: [ ] (taxonomic id from an organism, e.g. **9606** for *H.sapiens*, or taxon name, e.g. ***Homo***)*

if more than one, separate them by "+", e.g.* ***9606+Homo*

SUBMIT GO MODE 4! | What's this? | *e.g.* example 1, example 2

Exercise 3. Find a 3D of a polyQ ortholog

- Go to FASTAHERDER2: http://cbdm-01.zdv.uni-mainz.de/~munoz/fh2/

- Find a cluster containing polyQ and a PDB using mode 4

- Find the structure surrounding the place of polyQ insertion

- Any problems?

Exercise 3. Find a 3D of a polyQ ortholog

•Go to FASTAHERDER2:
http://cbdm-01.zdv.uni-mainz.de/~munoz/fh2/

•Find a cluster containing polyQ and a PDB using mode 4

•Find the structure surrounding the place of polyQ insertion

•Any problems?

•If yes, then use this example:
Species: “Escherichia coli”, PDB “yes”, and polyQ “yes”

Get the E. coli sequence and the one with polyQ and align them. Can you see the polyQ insertions?

Compare to PDB:4JNF (from DNAK_ECOLI P0A6Y8)

Exercise 3. Find a 3D of a polyQ ortholog

•FH2 mode 3 with C4YKT4 / *Candida albicans* 288 aa with two polyQ

Align and compare to P01123 *S. cerevisiae* 206 alpha-helix and polyQ inserts after. See PDB 2BCG chain Y = YPT1