
GRECO TFBS Benchmarking Initiative

Towards a representative set of TFs

Dr. Oriol Fornes

Deputy Group Leader @ Wasserman Lab
Centre for Molecular Medicine and Therapeutics
BC Children's Hospital Research Institute
University of British Columbia

oriol@cmmt.ubc.ca

 @OFornes

February 18, 2020

GitHub repository

<https://github.com/oriolfornes/GRECO>

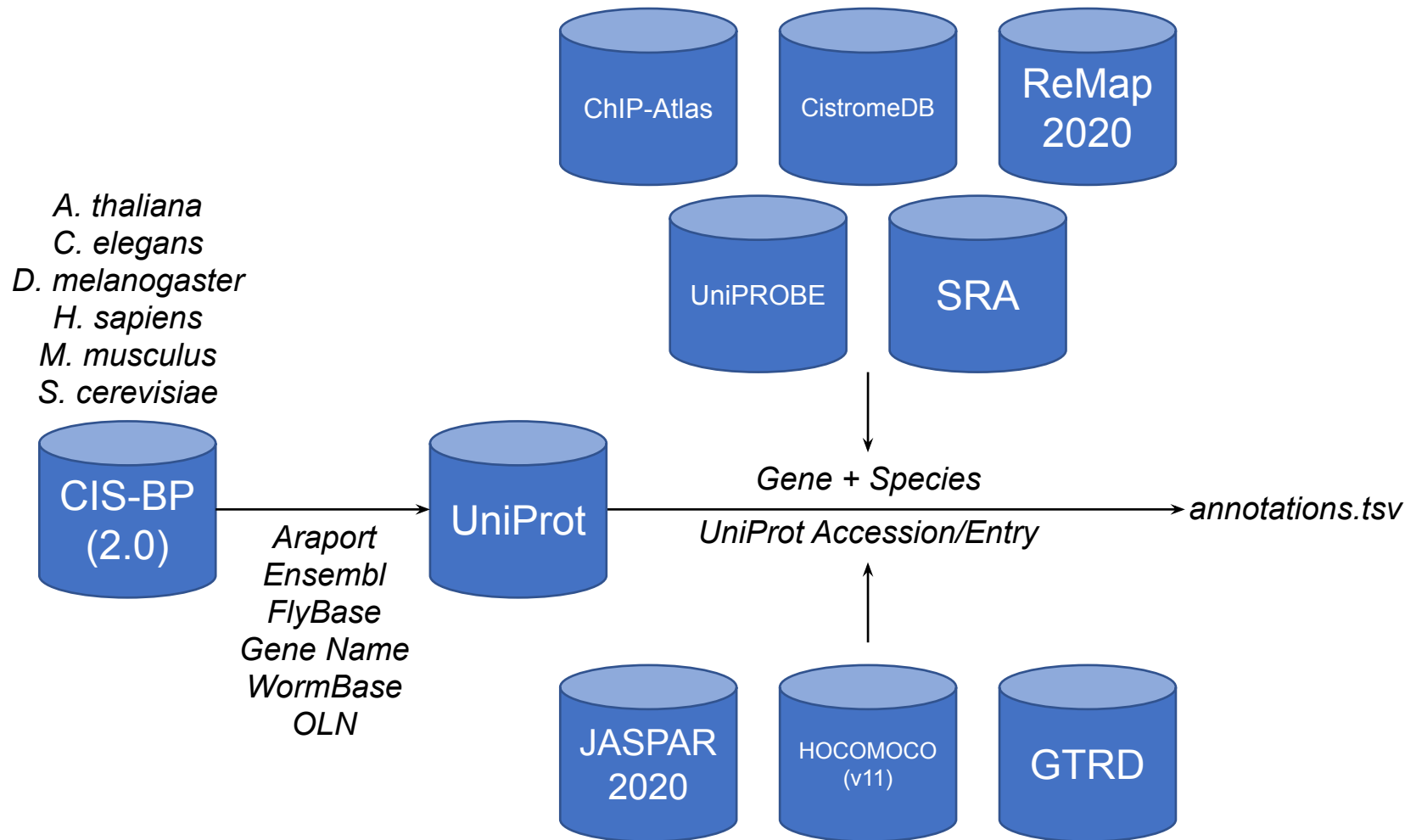
Outline

- **Aim**
- **Annotation**
- **Clustering**
- **Results**
- **Next steps**

Aim

- Obtain a representative set of TFs:
 1. Supported by high-quality experimental data;
 2. From different model organisms; and
 3. From different structural families.

Annotation



Annotation

<https://raw.githubusercontent.com/oriolfornes/GRECO/master/annotations.tsv>

Clustering

- For each TF (*i.e.* query)... (sort by amount of experimental evidence):
 1. Identify the query's Pfam DBD(s) using hmmscan with the "--domtblout" option and E-value thresholds for models and domains of 10^{-5} and 10^{-2} , respectively;
 2. Search for TFs homologous to the query using BLAST+;
 3. Select homologs:
 - with the same DBD composition than the query; and
 - whose BLAST+ alignment with the query is above the [Rost's sequence identity curve](#);
 4. For each selected homolog, if the amino acid sequence identity of the query and homolog DBDs is greater than the DBD-specific motif inference thresholds from CIS-BP, cluster the TFs together.

Clustering

<https://github.com/oriolforbes/GRECO/blob/master/Data/Clusters/TFs.json>

Results

- Triads: *i.e.* TFs with support by *in vivo* and **at least two** *in vitro* methods
- Species:
 1. *Drosophila melanogaster* **1**
 2. *Homo sapiens* **57**
 3. *Mus musculus* **28**
- Families:
 1. C2H2 ZF **15**
 2. C2H2 ZF,MADF **1**
 3. CUT,Homeodomain **1**
 4. DM **1**
 5. E2F **1**
 6. Ets **4**
 7. Forkhead **6**
 8. GATA **2**
 9. Homeodomain **13**
 10. Homeodomain,POU **1**
 11. Homeodomain,Paired box **1**
 12. Nuclear receptor **14**
 13. RFX **2**
 14. Rel **2**
 15. SAND **1**
 16. Sox **4**
 17. bHLH **9**
 18. bZIP **8**

Results

- [Duos](#): *i.e.* TFs with support by *in vivo* and *in vitro* methods
- Species:
 1. Arabidopsis thaliana **98**
 2. Caenorhabditis elegans **30**
 3. Drosophila melanogaster **39**
 4. Homo sapiens **279**
 5. Mus musculus **130**
 6. Saccharomyces cerevisiae **32**
- Families: **62**

Results

- [Multiple evidence](#): *i.e.* TFs with support by **at least two** methods
- Species:
 1. Arabidopsis thaliana **98**
 2. Caenorhabditis elegans **30**
 3. Drosophila melanogaster **40**
 4. Homo sapiens **290**
 5. Mus musculus **141**
 6. Saccharomyces cerevisiae **32**
- Families: **62**

Next steps

- Ensure that the representative set of TFs contains only sequence-specific DNA-binding TFs
- For each representative TF, ensure that the mapped experimental data correspond to that TF