Searching for orthologs in un-annotated genome assemblies with fDOG – Assembly

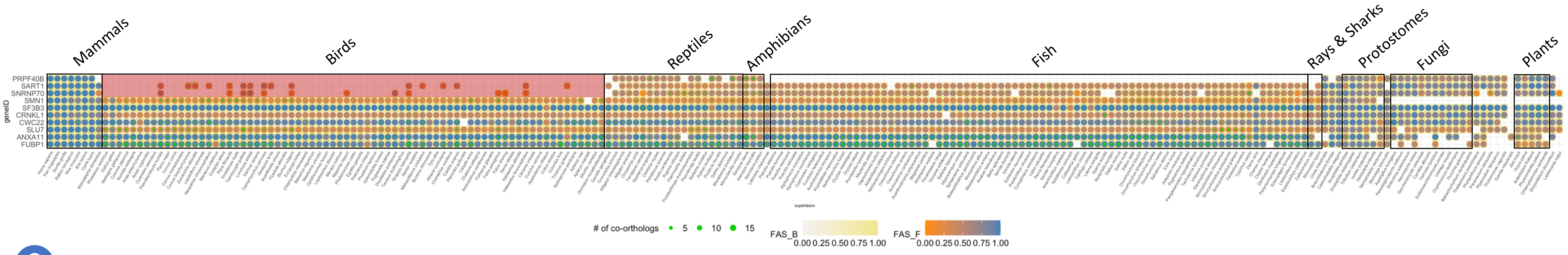
Hannah Mülbaier, Vinh Tran and Ingo Ebersberger

Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience,
Goethe University Frankfurt, Germany

SCAN ME

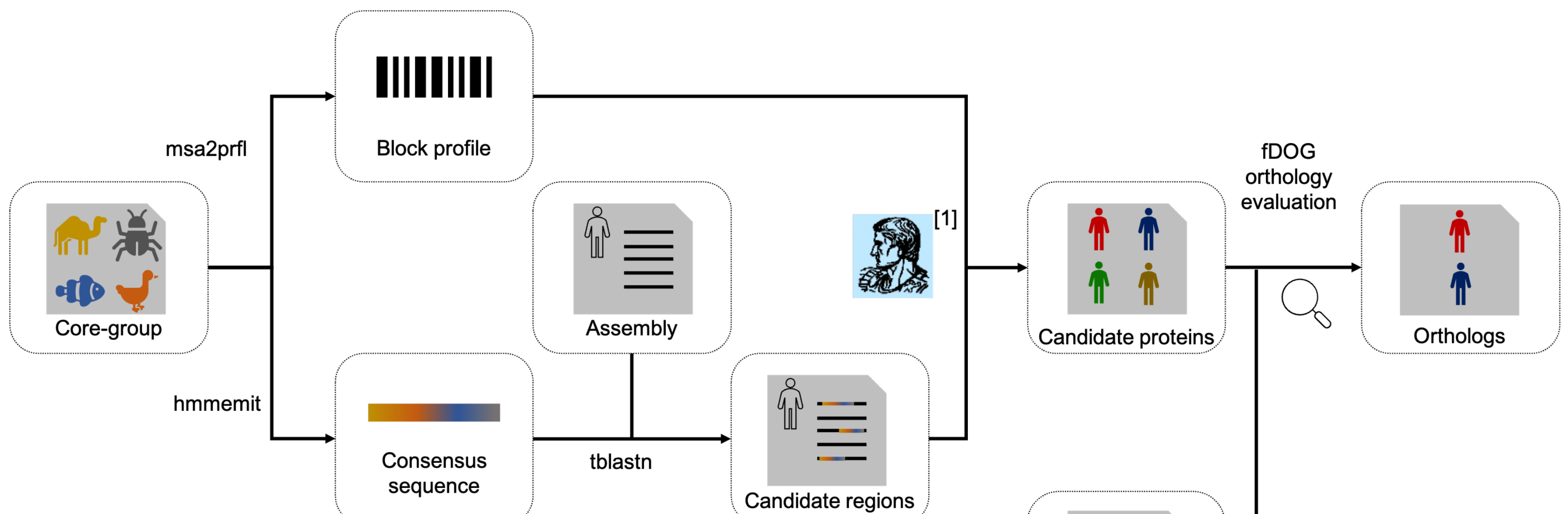
Motivation & Background

The identification of orthologs in the genomes of newly sequenced species is a relevant step for their integration into a broad range of evolutionary and functional studies. Numerous approaches varying in computational complexity, sensitivity and specificity have been developed for this purpose. However, one dependency is common to all tools: they require comprehensively annotated gene sets as input where any overlooked gene will result in a missed ortholog. Here, we present fDOG – Assembly, a targeted profile-based ortholog search tool that can identify orthologs in un-annotated genome assemblies.



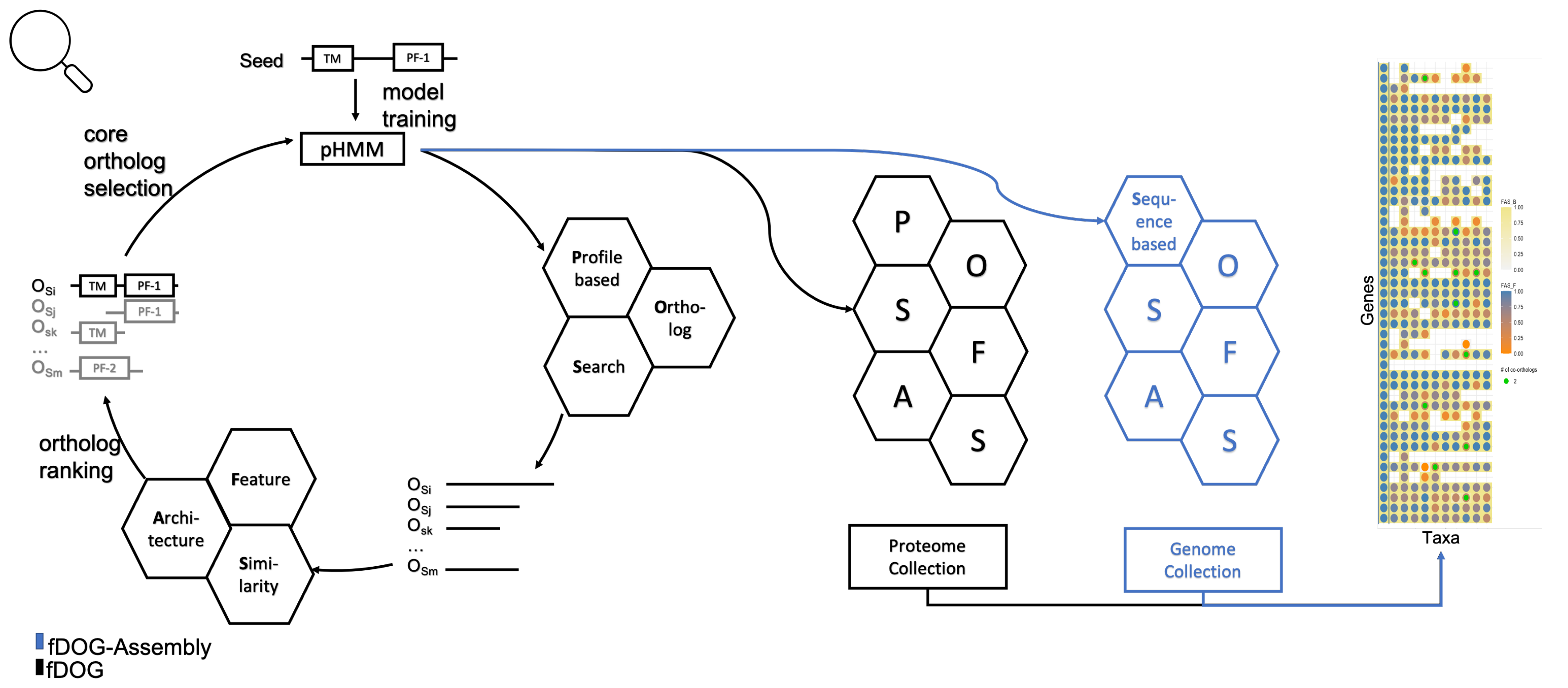
Gene loss or artefact?

Ortholog search pipeline



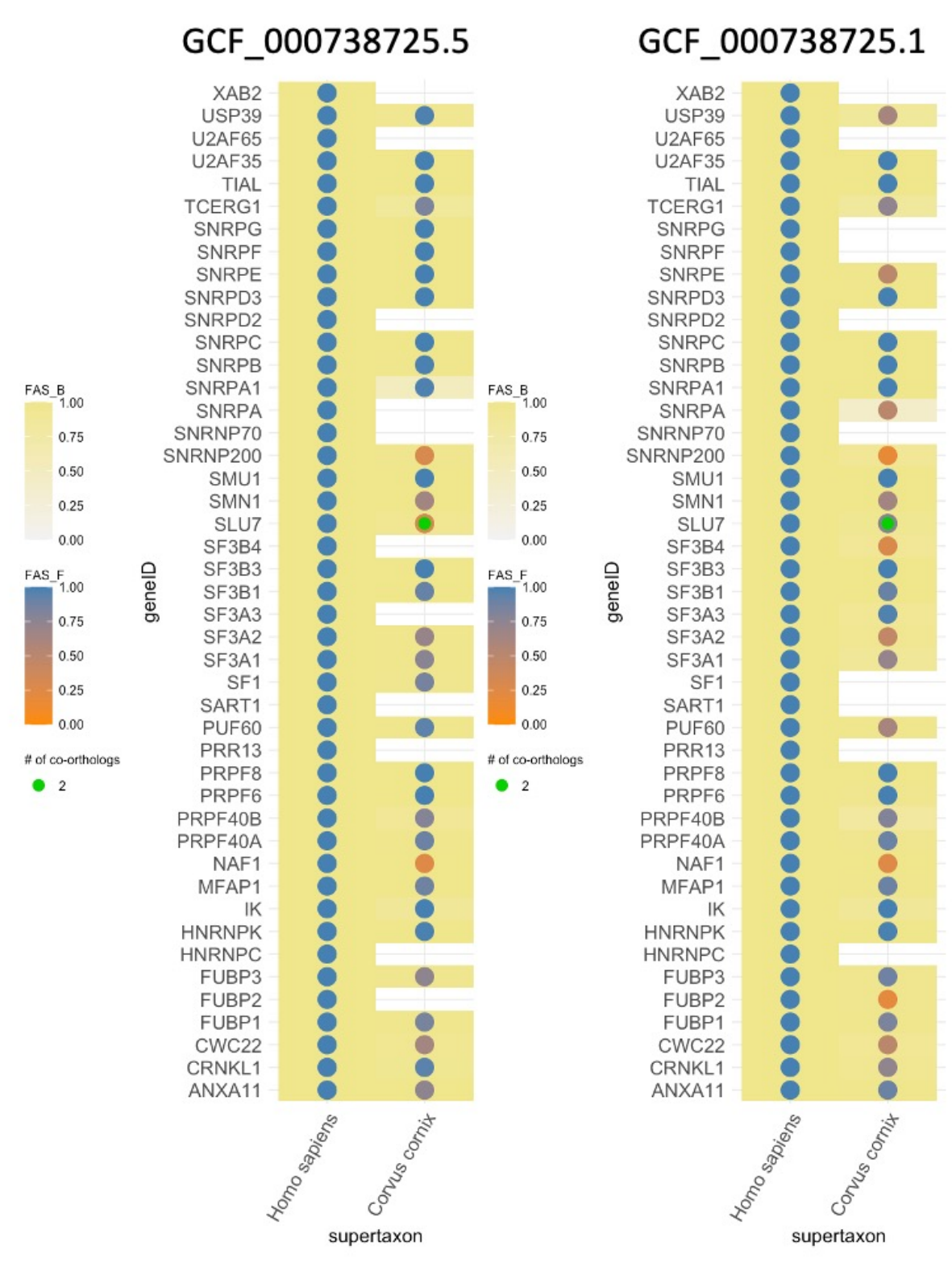
Core-group: Ortholog group representing gene of interest

Reference species: A species chosen by the user which is used as reference to test if reciprocity criterion is fulfilled

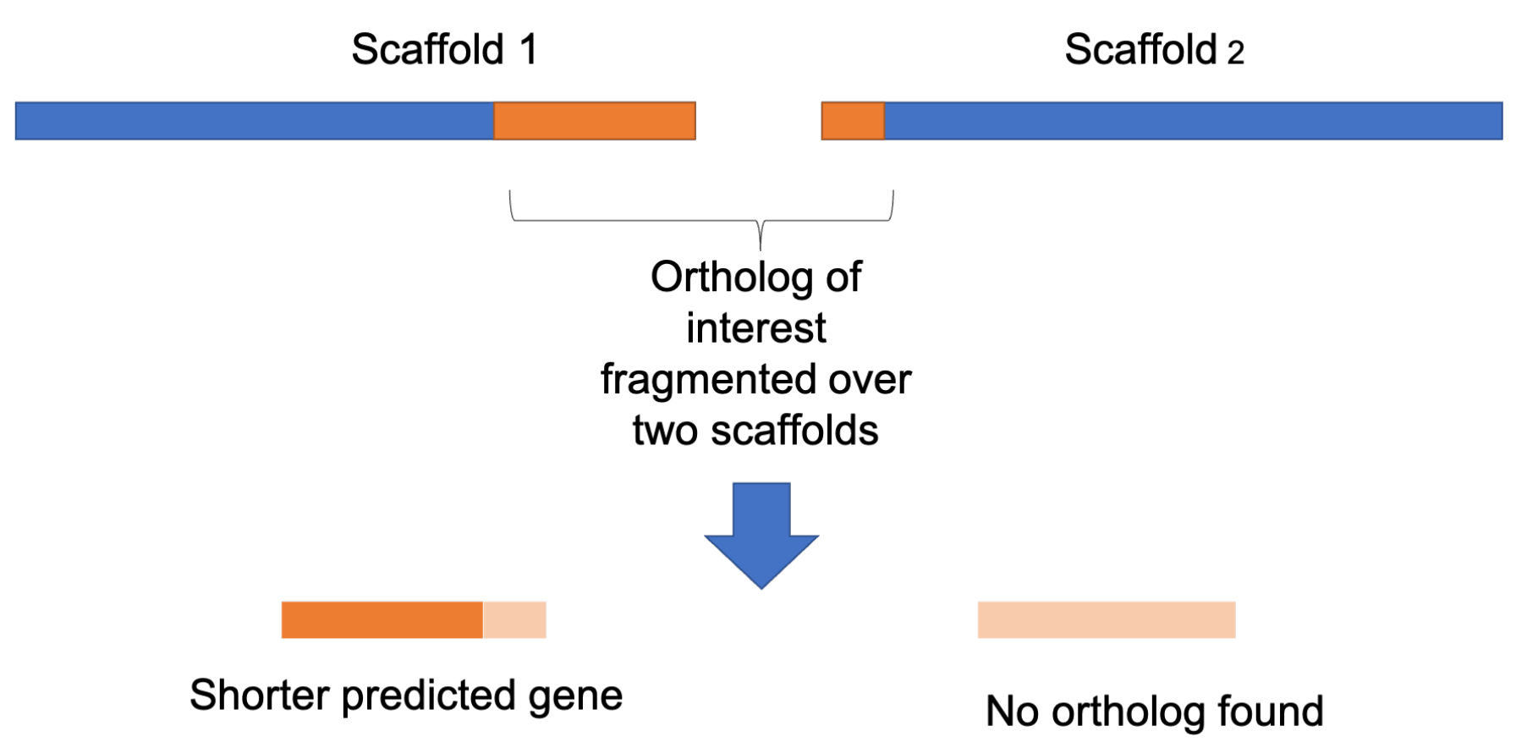


Assembly quality: a limiting factor

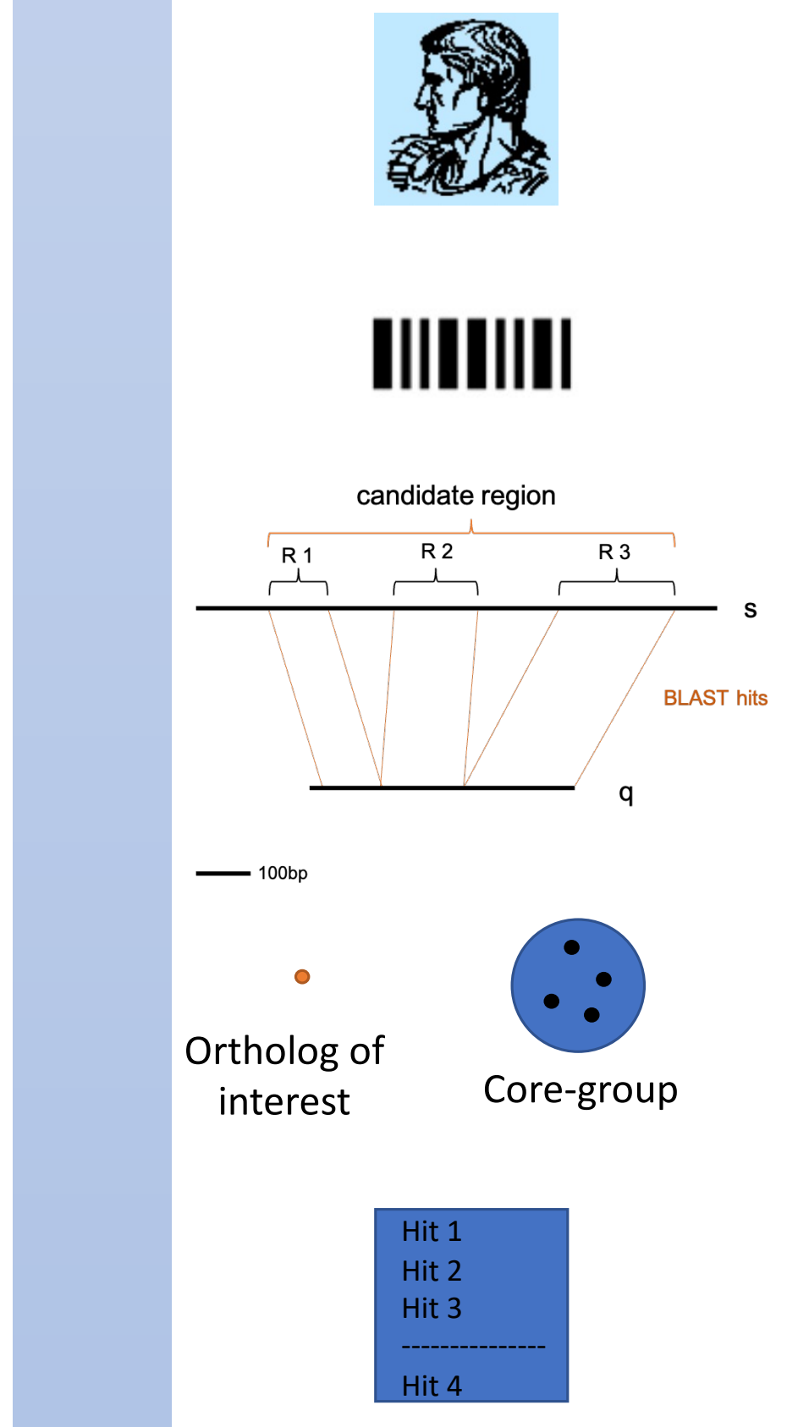
1 Different assembly versions can lead to different presence/absence patterns



2 Draft assemblies can result in fragmented or missed orthologs



Reasons for incomplete gene predictions



- Gene prediction with Augustus[3] is only guided by block profiles, additional hints are not available
- The Block profile is not significant enough and will therefore not be used during gene prediction
- The candidate regions forwarded to Augustus are too small because the tblastn[4] hit locations do not match the expected intron length
- Ortholog of interest differs significantly from the core-group and will therefore not be found during tblastn search
- E-value cut-off was chosen too low

Sensitivity is up to you

fDOG - Assembly offers different parameters to adapt the sensitivity and precision:

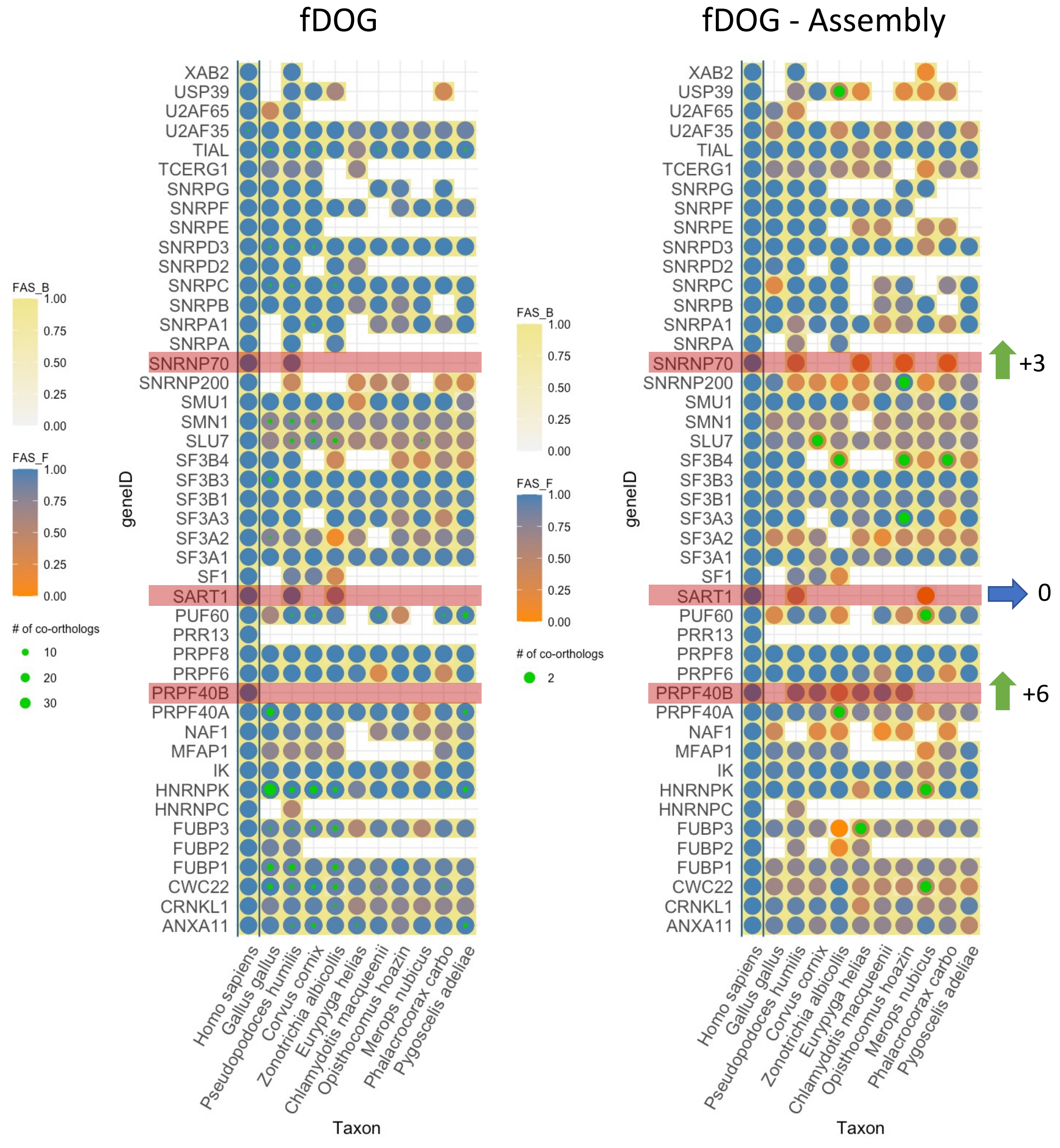
- A more sensitive search with the parameter –checkCoorthologsRef
- E-value cut-off can be changed by the user
- The user can increase the parameters –avItron and –lengthExtension which were used to compute the size of the candidate regions

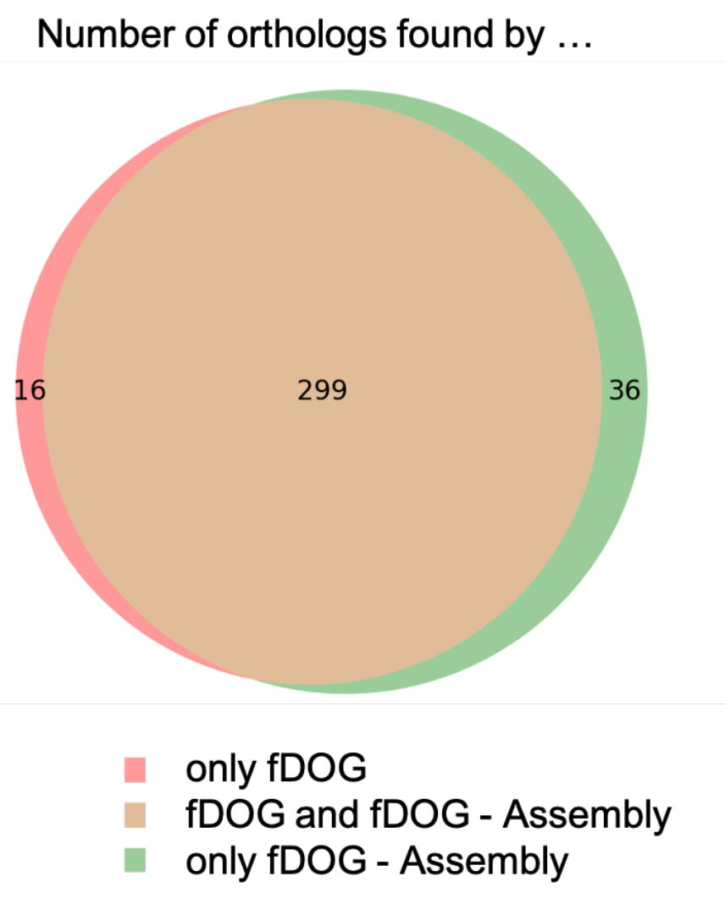
Costs computational time

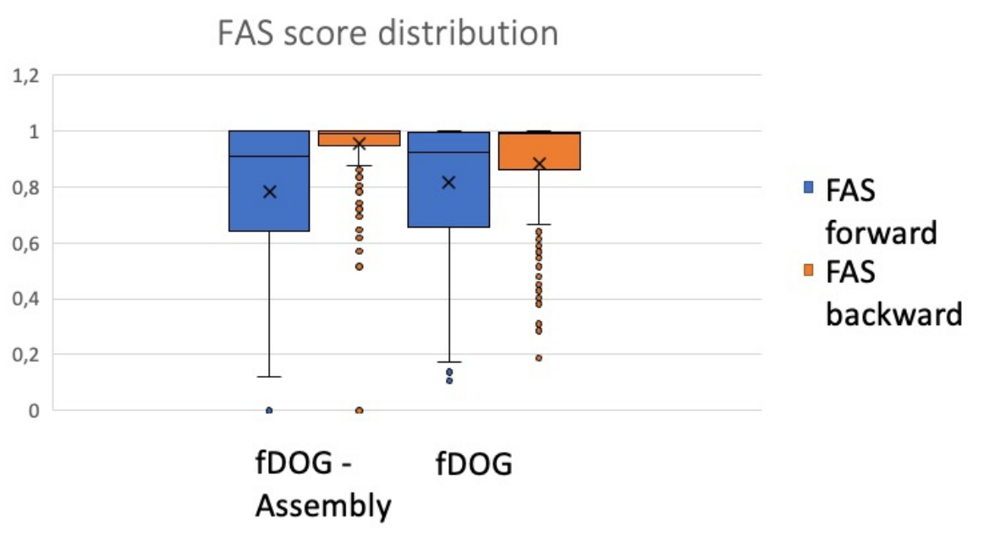
Initial benchmark

Benchmark setting:

- 45 proteins involved in alternative splicing
- ortholog search in 10 annotated NCBI RefSeq gene sets with fDOG[2]
- ortholog search in the corresponding genome assemblies with fDOG - Assembly







Overlapping genomic locations of orthologous found by both fDOG and fDOG – Assembly: 98%

Take home

- fDOG – Assembly can search in un-annotated genome assemblies which allows to by-pass time and resource-demanding gene annotations
- Initial benchmark revealed a performance that is comparable to the ortholog search in fully annotated gene sets
- fDOG – Assembly already includes different parameters which can improve the sensitivity or adapt the ortholog search to the species set of interest

Contact

Hannah Mülbaier
Applied Bioinformatics Group,
Goethe University Frankfurt, Germany
hannah.muelbaier@gmail.com

Ingo Ebersberger
Applied Bioinformatics Group,
Goethe University Frankfurt, Germany
ebersberger@bio.uni-frankfurt.de

References

[1] http://bioinf.uni-greifswald.de/augustus/
[2] Zhen Jiang, Claudia Carlantoni, Srinivas Allanki, Ingo Ebersberger, Didier Y. R. Stainier; Tek (Tie2) is not required for cardiovascular development in zebrafish. Development 1 (2020)
[3] Oliver Keller, Martin Kollmar, Mario Stanke, Stephan Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments, Bioinformatics, Volume 27, Issue 6 (2011)
[4] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. (2009)