**Practical 6: Orthology Prediction**

**Group number: 2**

**Group members: Maximilian Senftleben, Zhong Hao Daryl Boey**

**Summary:** The practical involved searching across the various available databases for suitable genes to perform the orthology test, followed by searching for suitable databases for the genes selected. The same genes were then searched-for across the various databases, with the hits compared and quality assessed. We also reviewed the concepts behind the various databases, and their relevance to various purposes. In addition the results were used as a benchmark to compare OMA, InParanoid, and DB Phylome, and their ability to accurately predict orthologs.

KEY QUESTIONS

1. Summarize shortly this practical.

The practical involved searching across the various available databases for suitable genes to perform the orthology test, followed by searching for suitable databases for the genes selected. The same genes were then searched-for across the various databases, with the hits compared and quality assessed. We also reviewed the concepts behind the various databases, and their relevance to various purposes.

2. Pick at least three databases that store orthologs for three of your selected genes (links

provided under Material & Tools); describe the used algorithms of the databases you are

comparing and motivate your choice of databases.

Gene 1: Q9ZZW7, present in OMA, InParanoid, and Tree Explorer (DB Phylome).

Gene 2: P59932, present in OMA, InParanoid, and Tree Explorer.

Gene 3: G4FFG1, present in OMA, InParanoid, and Tree Explorer.

Therefore we used these 3 databases, as they contained the genes we selected. In addition, OMA uses a comprehensive search algorithm which first involves a Smith-Waterrnan alignment, using evolutionary distance to identify close homologs, followed by clustering. This results in a broad but specific search.

InParanoid uses a different method to find orthologs, by using pair-wise similarity scores from NCBI BLAST, followed by the creation of a seed pair of orthologs, adding sequences from the 2 reference proteomes that are similar to the seed pair, and members of this ortholog group are called inparalogs. This clustering method is faster than phylogenomic methods, and can also provide confidence values for hits, which is useful in assessing quality of the predictions.

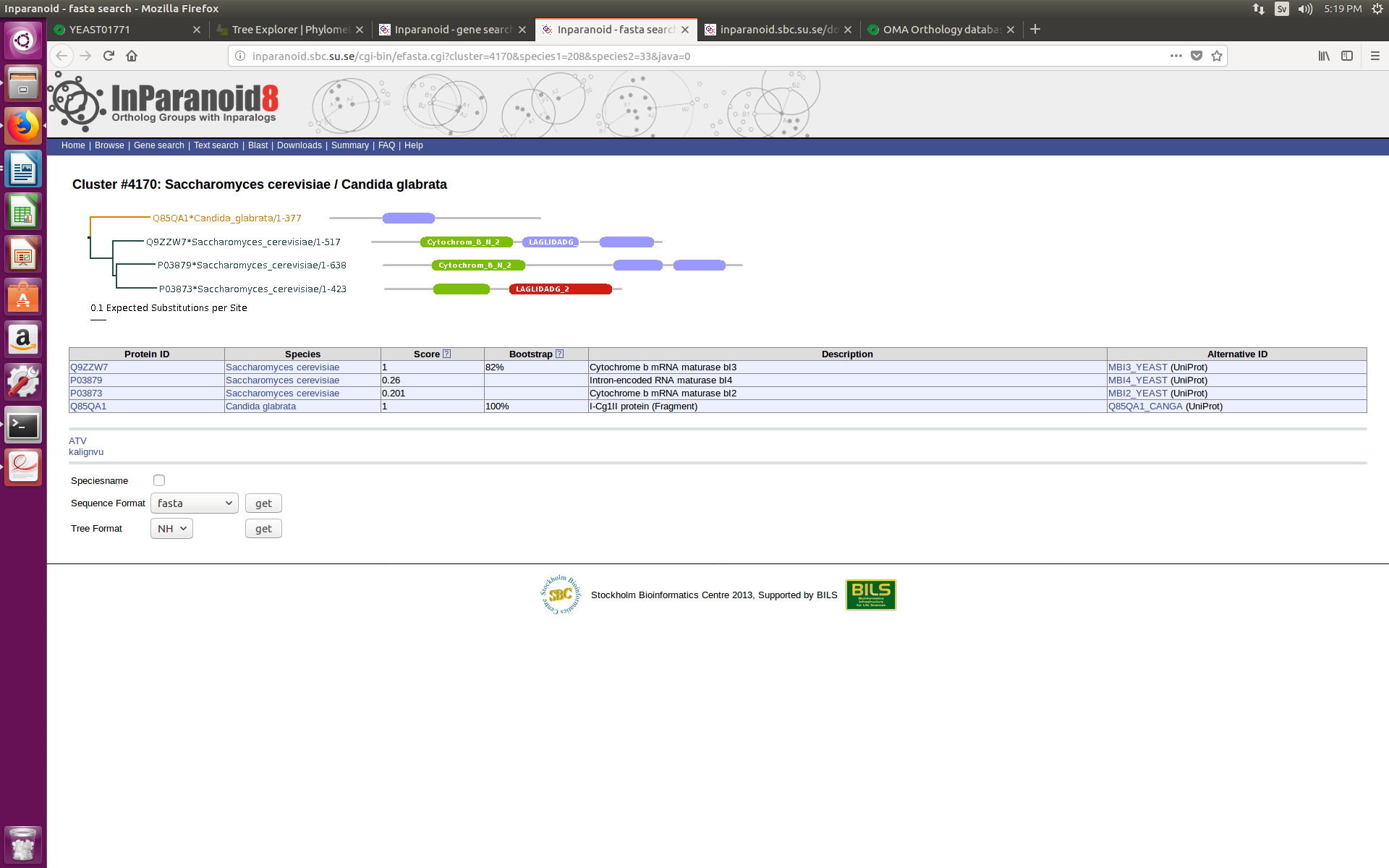
DB Phylome employs a phylogenomic method to search for orthologs, using evolutionary tree analysis.

3. Discuss the achieved results with the different algorithms, especially the differences

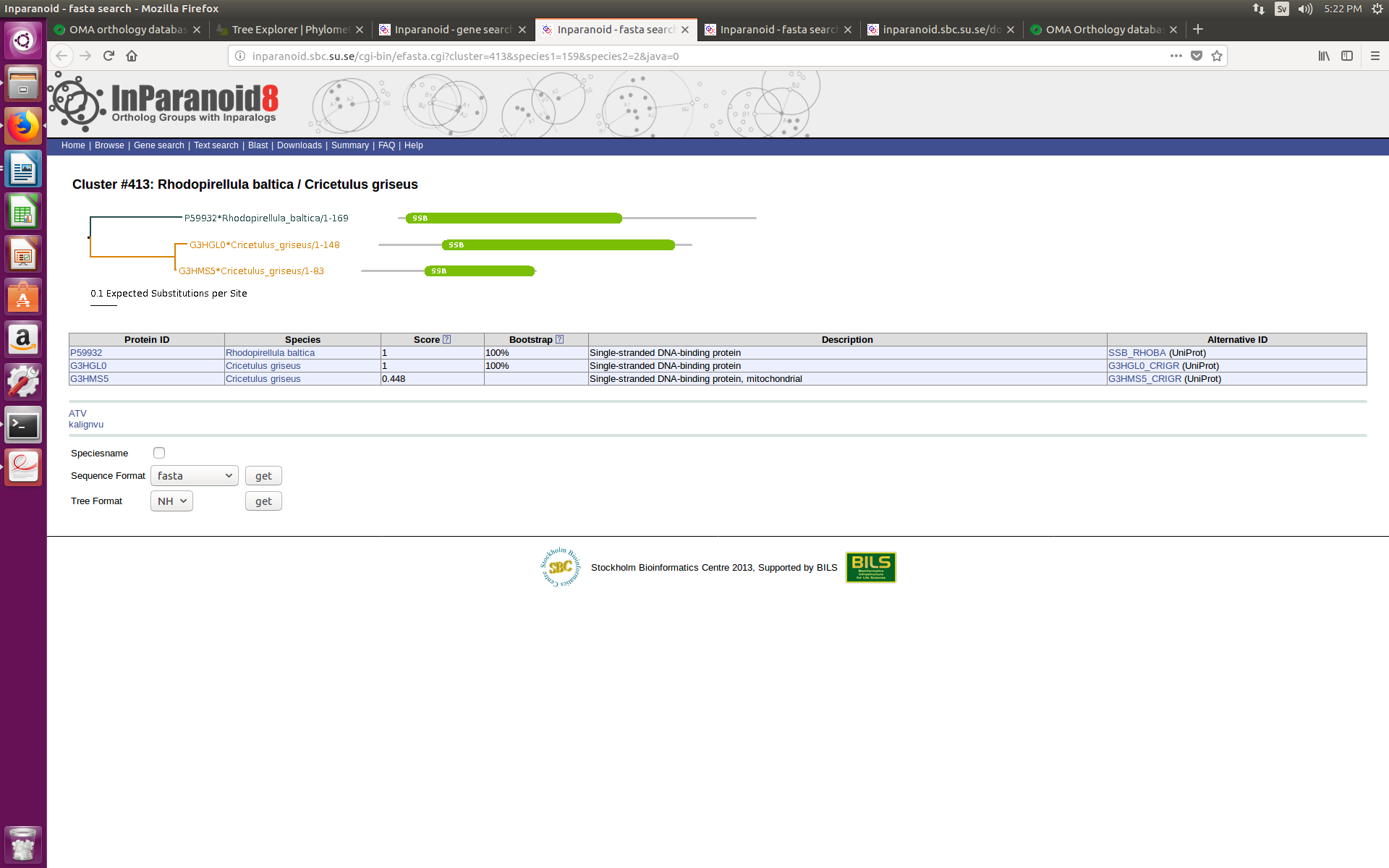
between their predictions (pairs, ortholog groups):

a. How do the predicted orthologs differ? Which are missing or are the same?

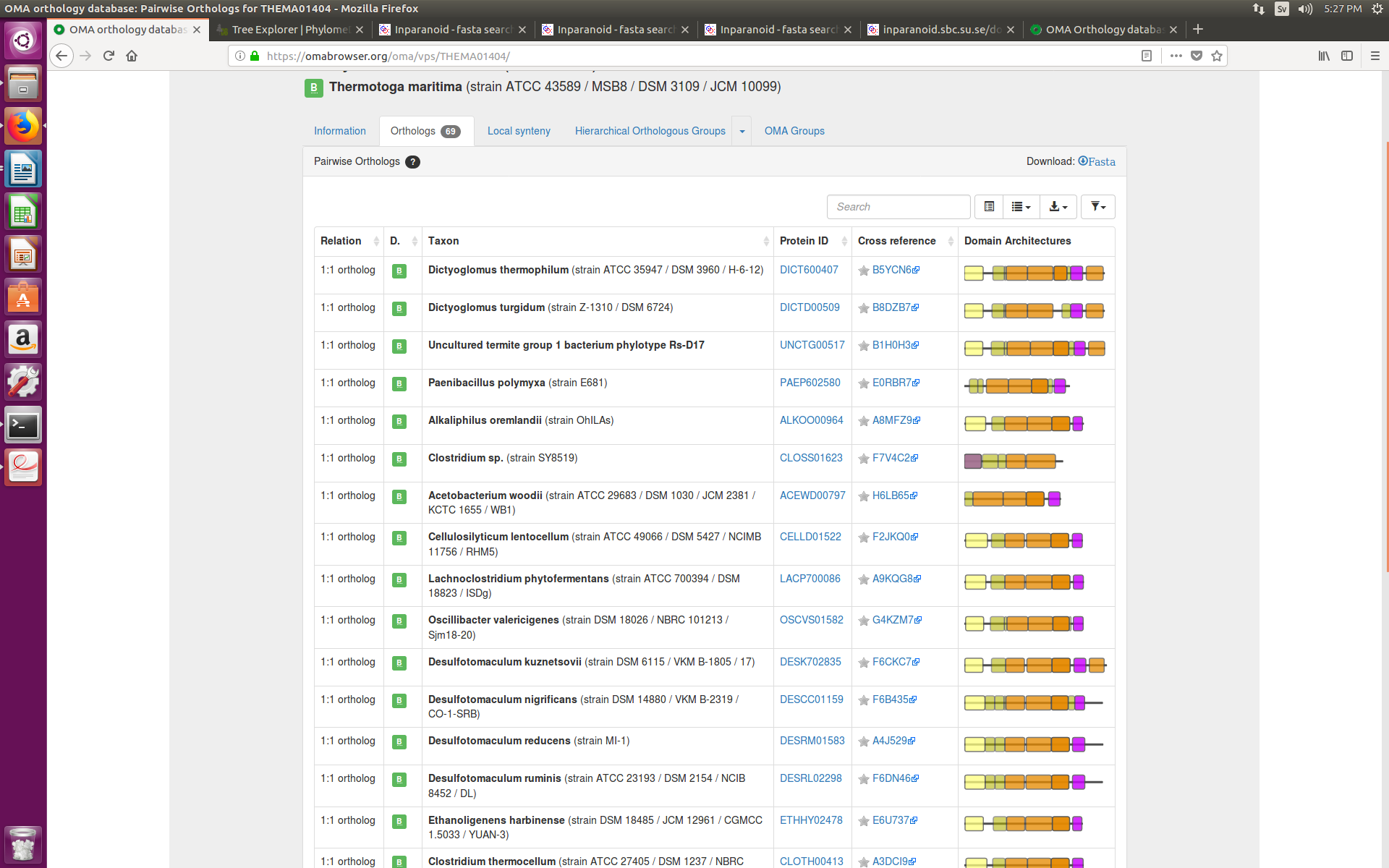
Gene 1, did not show any ortholog hits in OMA, returned a comprehensive tree for DB phylome (26 leaves), and multiple clusters for InParanoid.

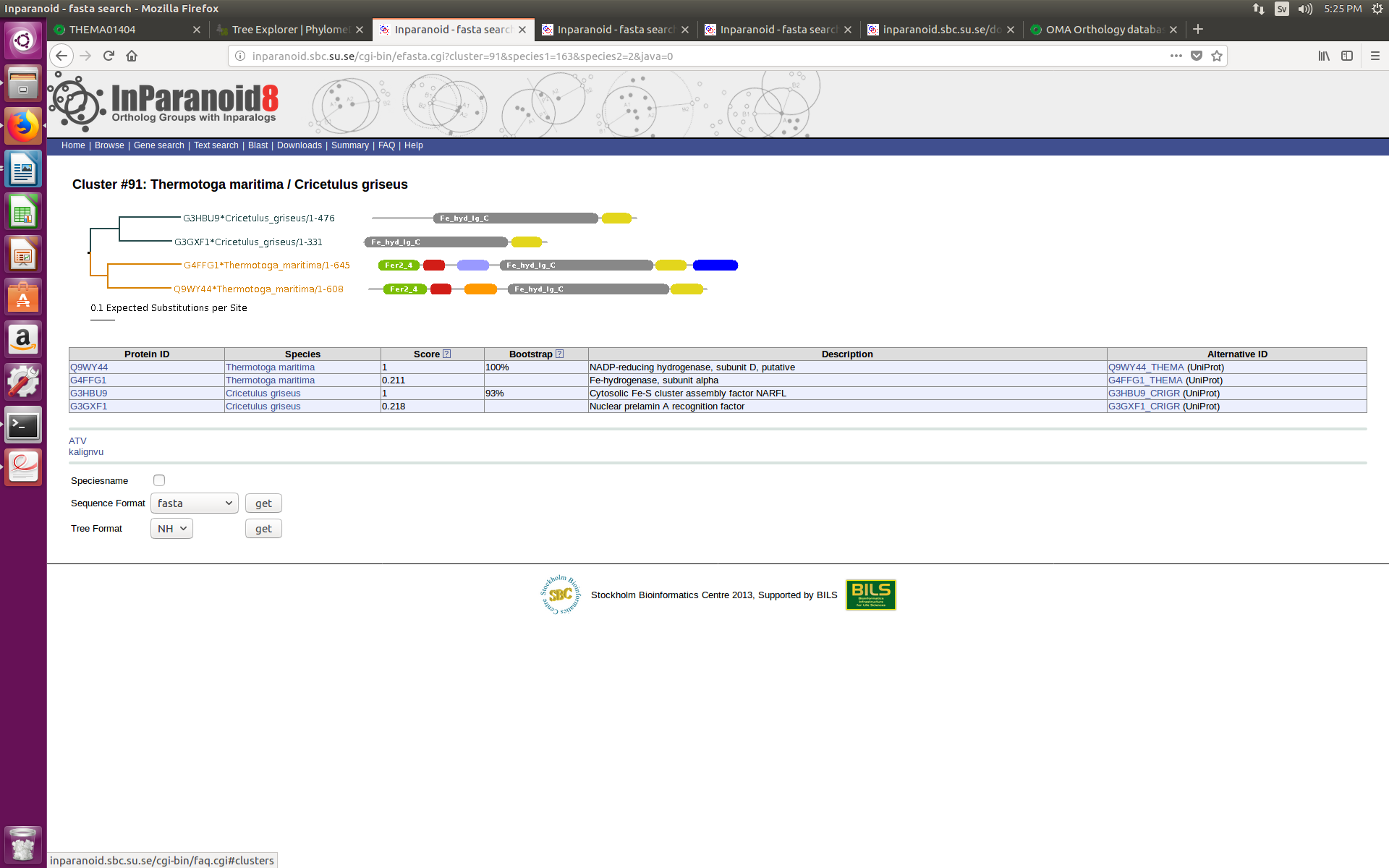


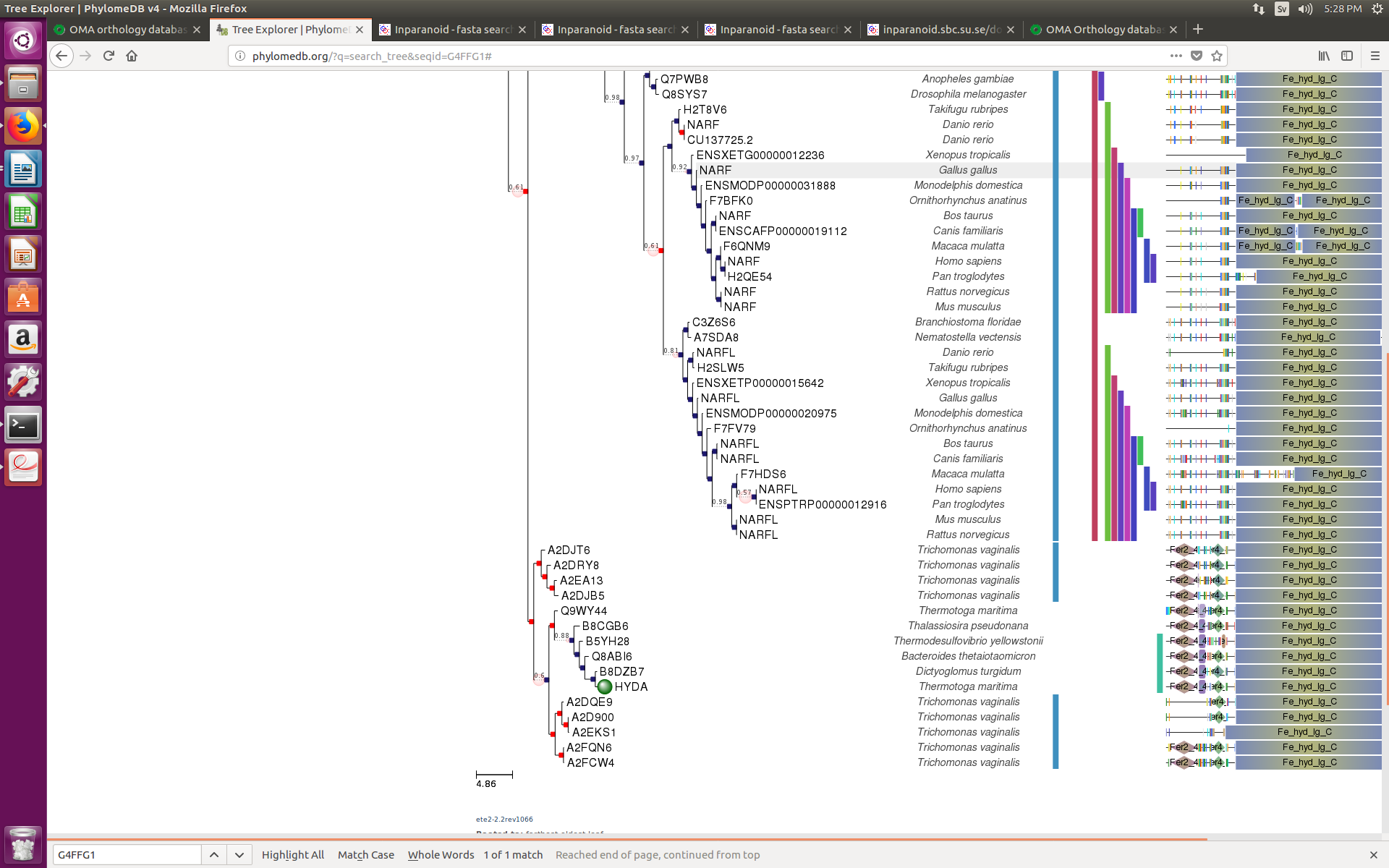
Gene 2 returned 1632 orthologs for OMA, 44 leaves for DB phylome, and multiple clusters for InParanoid.



Gene 3 returned 69 orthologs in OMA, 72 leaves for DB phylome, and multiple clusters for InParanoid.







In the instance of gene 3, all the databases returned a close ortholog for B8DZB7, a gene from *D. turgidum*.

It is immediately apparent that the 3 databases produce results that are reported in very different fashions, due to varying conceptions.

b. Can you find orthologs in one database that are either missing or appear as out-paralogs in another database? Why do you think this happens?

While conducting a search for homologs in gene 3, the next evolutionarily closest ortholog reported by Phylome DB was Q8ABI6 from the organism *B. thetaiotaomicron*. However scanning the other 2 databases for this ortholog did not return a hit for this search. This is highly likely due to the varying construction methods used to produce the databases, as well as different reference genomes and databases used in construction. Therefore some databases may contain references or hits that other databases may have excluded.

c. How big are the ortholog groups for your selected genes in the databases you compare?

Gene 1, did not show any ortholog hits in OMA, returned a comprehensive tree for DB phylome (26 leaves), and multiple clusters for InParanoid. Gene 2 returned 1632 orthologs for OMA, 44 leaves for DB phylome, and multiple clusters for InParanoid. Gene 3 returned 69 orthologs in OMA, 72 leaves for DB phylome, and multiple clusters for InParanoid.

d. What can you say about the quality of orthology predictions with the databases you compare?

It is hard to compare the quality of the predictions across the databases as they are completely differently developed tools, but InParanoid very handily clusters the ortholog hits into groups, with bootstrap scores and scores reported, allowing easy understanding of the quality of the hits, while OMA and Phylome DB are devoid of such a function.