

## Supplementary results

We evaluated the influence of input files generated by different BLASTn searches against NCBI nt and Diamond blastx searches against UniProt Reference Proteomes on taxonomic assignment by BlobTools. Values for F-scores are listed in *Table S1*, values for precision and recall in *Table S2*, and the number of bases annotated as TP, FP, TN, FN for the different similarity searches for each taxon at the taxonomic rank of order in *Table S3*.

In the context of sequence similarity searches for taxonomic annotations, FNs arise mainly through phylogenetic distance of the target sequence to the ones in the database, resulting in the lack of results (*i. e.* 'no-hit'). However, both FNs and FPs can also arise through taxonomically mis-annotated sequences in the database, *i. e.* when a genome sequence is contaminated with sequences from another genome. For instance, if the target organism is *E. coli*, a FN could originate from a *C. elegans* genome contaminated with bacterial DNA, as the bacterial sequence in the assembly would be tagged with the TaxID of *C. elegans*. Simultaneously, this result would count as a FP if the target organism is *C. elegans*, since a non-nematode sequence is tagged with the TaxID of *C. elegans*.

Sequence similarity searches carried out against databases without taxonomic masking yielded results mostly congruent with taxonomic annotation through read mapping. FP taxonomic annotations in BLASTn searches were only found for a small amount of *E. coli* sequences. One such sequence is 'contig\_8499' in the combined assembly of both simulated libraries (blobtools.assembly.A\_B.fasta), which was assembled from read pairs originating from the X chromosome of the *C. elegans* reference (position 5,909,163 to 5,911,151). However, BLASTn searches identified this sequence as a Tn10 transposon with the taxonomic annotation of Enterobacterales. A likely explanation for this case is the transposition of this sequence into a fosmid clone during the *C. elegans* genome project, as has been noted by the WormBase database (see [http://www.wormbase.org/species/c\\_elegans/feature/WBsf977957#0123--10](http://www.wormbase.org/species/c_elegans/feature/WBsf977957#0123--10)).

No FPs were observed for Diamond blastx searches, as these are restricted to protein coding sequences. However, variation in the number of FNs was greater and caused mainly by the absence of hits. Diamond blastx searches exhibited the highest number of FNs, especially for *H. sapiens* and *E. coli* sequences. This was more pronounced when using the parameters `--max_target_seqs 1`. If more hits are supplied, the number of FNs decreases. If BLASTn searches were provided in addition to Diamond searches (using taxrule 'bestsumorder'), recall ranges between 0.9993 and 1.0.

Taxonomic masking of sequence databases to simulate phylogenetic distance between query sequences and those in the database, revealed more complex patterns of interaction between search parameters, which varied between the taxa. This variation is unsurprising since taxonomic masking was carried out at different, non-analogous taxonomic ranks. Similar to unmasked searches, FPs are not a major concern. Lowest values of precision (ranging from 0.8167 to 0.8642) were observed for *E. coli* sequences when using Diamond blastx alone. Number of FNs bases are also most extreme in taxonomic annotations based on Diamond blastx searches alone (for *H. sapiens* and *P. aeruginosa* sequences) or when using the BLASTn parameter `-culling_limit 10` (for *P. aeruginosa* sequences). Hence, we discourage the use `-culling_limit` for similarity searches, which are used for taxonomic annotation through BlobTools.

The best results for this dataset were obtained using the BLASTn parameter `-max_target_seqs 10` in combination with Diamond blastx parameter `--max_target_seqs 1` (lowest precision and recall, 0.9996 and 0.9374, respectively). However, a much faster search with acceptable outcome can be achieved by changing the BLASTn parameters to `-max-target-seqs 1 --max_hsp 1` (lowest precision and recall, 0.9997 and 0.9042, respectively).

Although we simulated phylogenetic distance between the sequences in the assembly and the sequences in the databases through taxonomic masking, it should be noted that the results discussed here are by no means universal. If an organism has not been sequenced previously, the number of FNs will be high and if it was, but the data is contaminated with other taxa or the organism is a contaminant itself, the number of FPs and FNs will be greater.