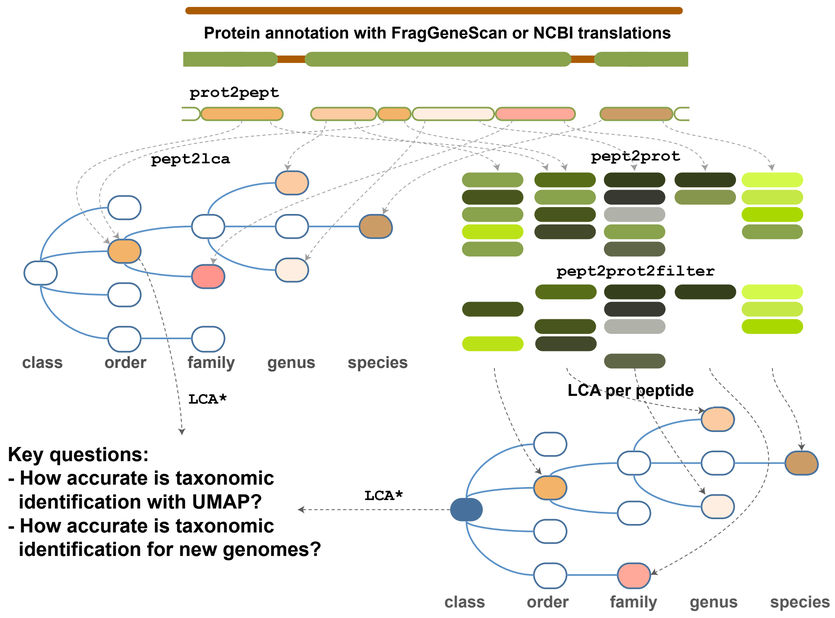
Benchmarking the Unipept Metagenomics Analysis Pipeline

T. Naessens, B.T. Habtemariam, R. Deklerck, S. Houbracken, M. Niklaus, I. Melckenbeeck
Promoter: Prof. Dr. Peter Dawyndt

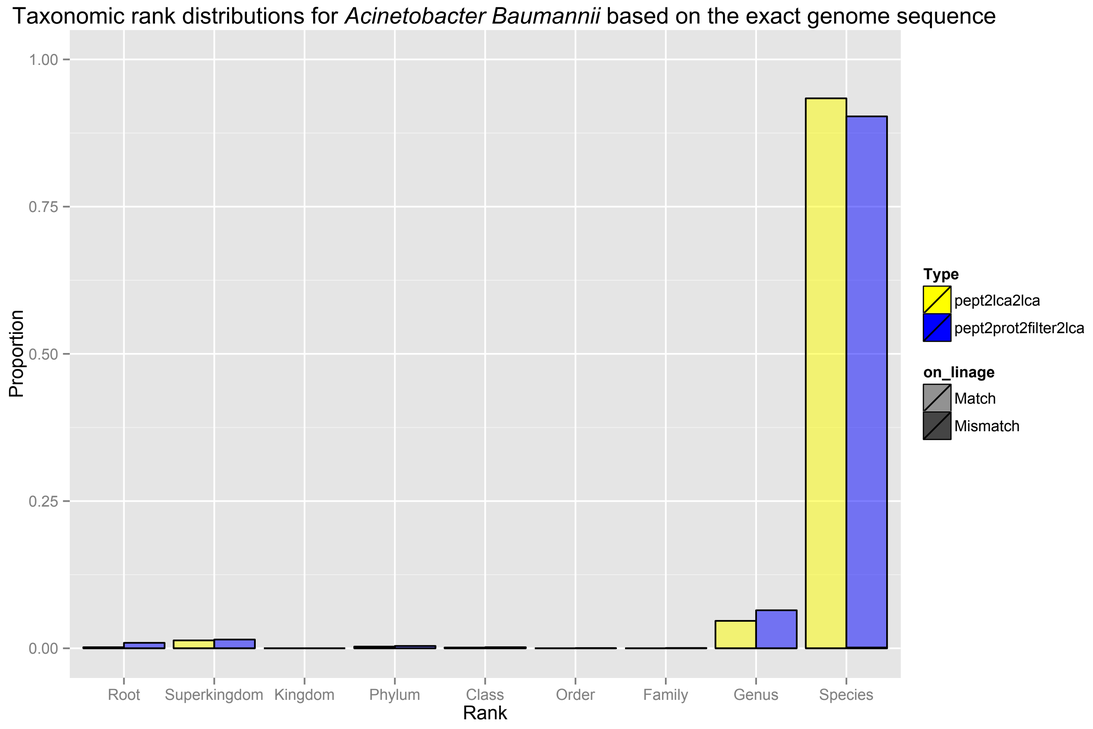
Abstract



Context The Unipept Metagenomics Analysis Pipeline (UMAP) is an approach to solve the problem of taxonomic identification with metaproteomics. This is achieved by predicting all proteins on each DNA strand of a metagenomics sample, running the Unipept metaproteomics pipeline on these proteins, as indicated in the left hand side of the picture on the left, and aggregating them back to one resulting taxon. This last step is done using a novel LCA* algorithm, which exploits the fact that the proteins all originate from one DNA strand.
Approach To benchmark, we run the UMAP on both completely sequenced genomes and simulated reads from those genomes and compare the results with a separate analysis on those genomes, but where proteins that were found in Uniprot to be originating from that genome, are filtered out. This allows us to simulate what would happen if the UMAP is being run on unknown genomes while still producing comparable results.
Results To summarise obtained results, we have found that the UMAP performs very well for known genomes, where on average 97% is mapped on the species level. For simulated reads where no error was introduced, this number is reduced to 74% and lowers the more error is introduced. On the simulated unknown genomes, 38% is mapped to the species level.

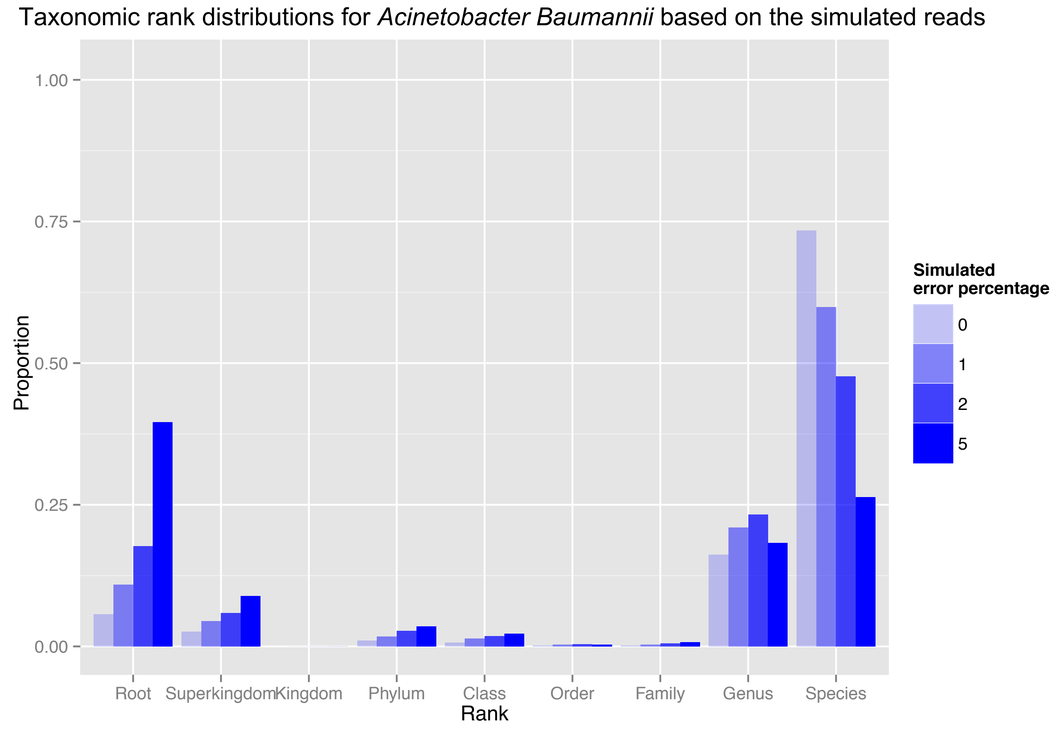
Benchmarking results for *Acinetobacter baumannii*

Taxonomic rank distributions for *Acinetobacter Baumannii* based on the exact genome sequence



The barplot above shows the taxonomic rank distribution of the results of both toolchains for the taxonomic identification of the peptides from the *Acinetobacter Baumannii* organism. The yellow bars show the result with the default UMAP toolchain, where the blue bars correspond with the different approach where the proteins from the original sequence have been filtered out in the identification process.
A specific level of identification is obtained for both toolchains. We also see a shift to the less specific ranks when we filter out the proteins already occurring in the originating genome. This shift is expected as this filter step causes a loss of specific information.

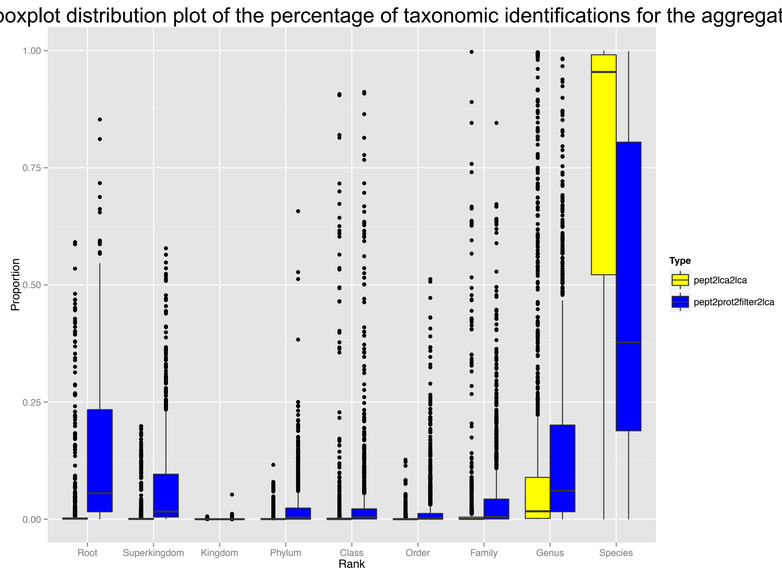
Taxonomic rank distributions for *Acinetobacter Baumannii* based on the simulated reads



The above barplot shows the distribution of identifications found at the different levels in simulated reads on one genome with read lengths of 500 with 0%, 1%, 2% and 5% error rates.
As can be seen in the plot, the identification of the proteins is about 20% less accurate when using reads with 0% than when using the exact genome sequence. When using reads without errors, almost 75% of the proteins are mapped to the species rank and 20% to genus. Introducing errors in the reads predictably hampers the identification of the peptides, resulting in a worse specific identification when the error rate increases.

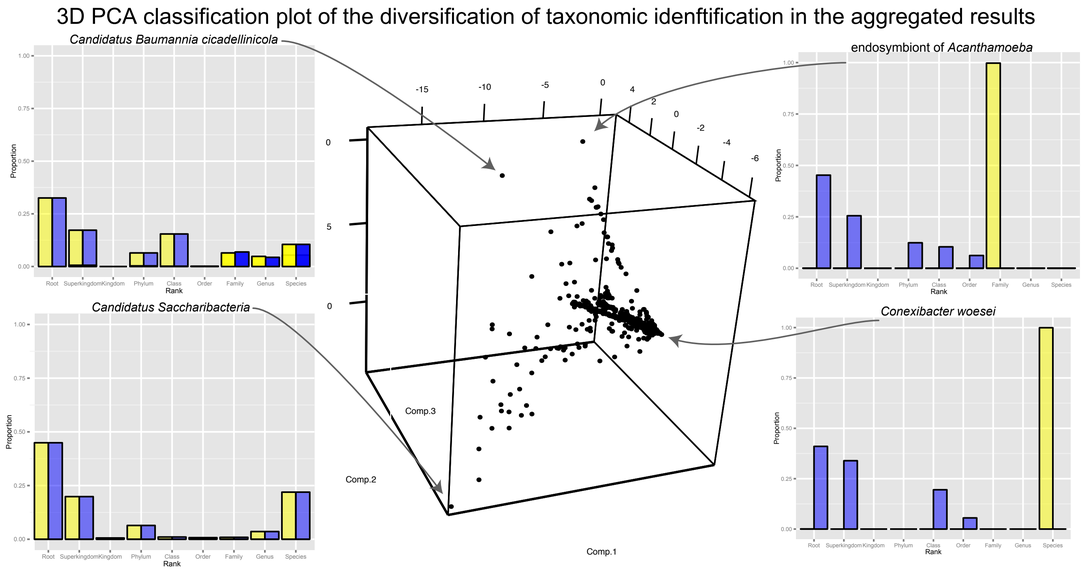
Total aggregated results for 1145 genomes

Per rank boxplot distribution plot of the percentage of taxonomic identifications for the aggregated results



Identification of proteins from 1145 genomes with (blue) and without (yellow) filtering out proteins that were found to be originating from the initial genomes. Without filtering, the identifications are on average for 97% mapped to species rank, with the remaining being mapped at genus rank. Enabling the filter step results in a general shift to the less specific ranks, which is expected as specific information is being filtered out.

3D PCA classification plot of the diversification of taxonomic idenftification in the aggregated results



Above is the PCA plot of the identification results for 1145 completely sequenced genomes. This plot clusters the genomes by the way their proteins are classified. Most are clustered along one line, but some outliers can be seen. The most extreme outliers have been accompanied by their correspondig bar plots. The reason for these outliers can vary from wrongly identified or classified genomes in the source database, genomes with very less or very much specific peptides or proteins, etc.