

Supplemental Document 3: The Effect of the Data on the Performance

The outlier detection depends on a score distribution, hence, the number of sequences that are present in a subset is envisaged to affect performance. Then, certain sequences present in both a complete proteome dataset and SwissProt, which is somewhat smaller than most of the complete proteome-derived datasets, sometimes had different fates. Hence, we analysed the BAHd dataset (total 2003 sequences) to see if the size of the offered dataset affects performance.

We created ten randomized sequence sets for four different subset size conditions. Sets tagged as Small contain 40 subsets of ~50 sequences, Medium sets have 20 subsets of ~99 sequences whereas each Large set has 10 subsets of ~199 sequences. Each of the Small, Medium and Large subsets thereby randomly represent ~2.5, 5 and 10% of all sequences, respectively. The Variable datasets emulate the sizes of the different initial crude BAHdomes, ranging from 4 to 368 sequences. Additionally, we combined all the sequences in a single dataset. All of these datasets were subjected to the default pipeline (12345) and a pipeline lacking the outlier removers (134). Due to the large size of the single combined dataset, cleaning this dataset was performed using FAMSA rather than MAFFT G-INS-i as the alignment method. The originally prepared datasets were treated using both pipelines each using FAMSA or MAFFT G-INS-i. Figure SD3-1A shows the design.

We first quantified the number of sequences removed by each module for each dataset (Figure SD3-1B). SSR removes all sequences below a certain length determined by the reference sequence and shows no differences. The PR module removes more sequences from larger datasets but the effective number is very small (<20 as compared to ~200 for CGSR) and with high coefficients of variation (CV) in all the randomized datasets. The clear trend seen for the PR module is not seen for the NHHR module which is explained by the higher complexity of the MSAs made of sets with many more NFHs. All other modules show differences but no trend that indicates the subset size affects the modules' performance. Hence, size appears to affect the PR module and the remaining differences are likely better explained by random variation. An interesting detail is the fact that CGSR, which removes the highest number of sequences following SSR (between 194 to 215), shows low CVs, which suggests it is reliable.

A second question is whether the same sequences are removed and to what extent. For this, sequences were tagged for each module and final result (accepted) and we determined the cumulative tag scores for both the default and the 134 pipelines. Figure SD3-1C shows the incidence of tag scores for each module as part of the default pipeline, whereas Figure SD3-1D summarizes the total numbers for both pipelines. By far the largest part of the sequences (>90%) was either always (tag score Accepted = 43) or never (tag score Accepted = 0) accepted. The GIR and CGSR modules do never identify 94% and 87% of all sequences, respectively, and always detect 1% and 6% of all sequences, leaving 5 to 7% of sequences that are detected in some sets only. Both the NHHR and the PR module data show a skew towards never removed.

Together, these results indicate that even though dataset size may have an impact on modules' performance, Seqrutinator shows a high consistency on identifying NFHs.

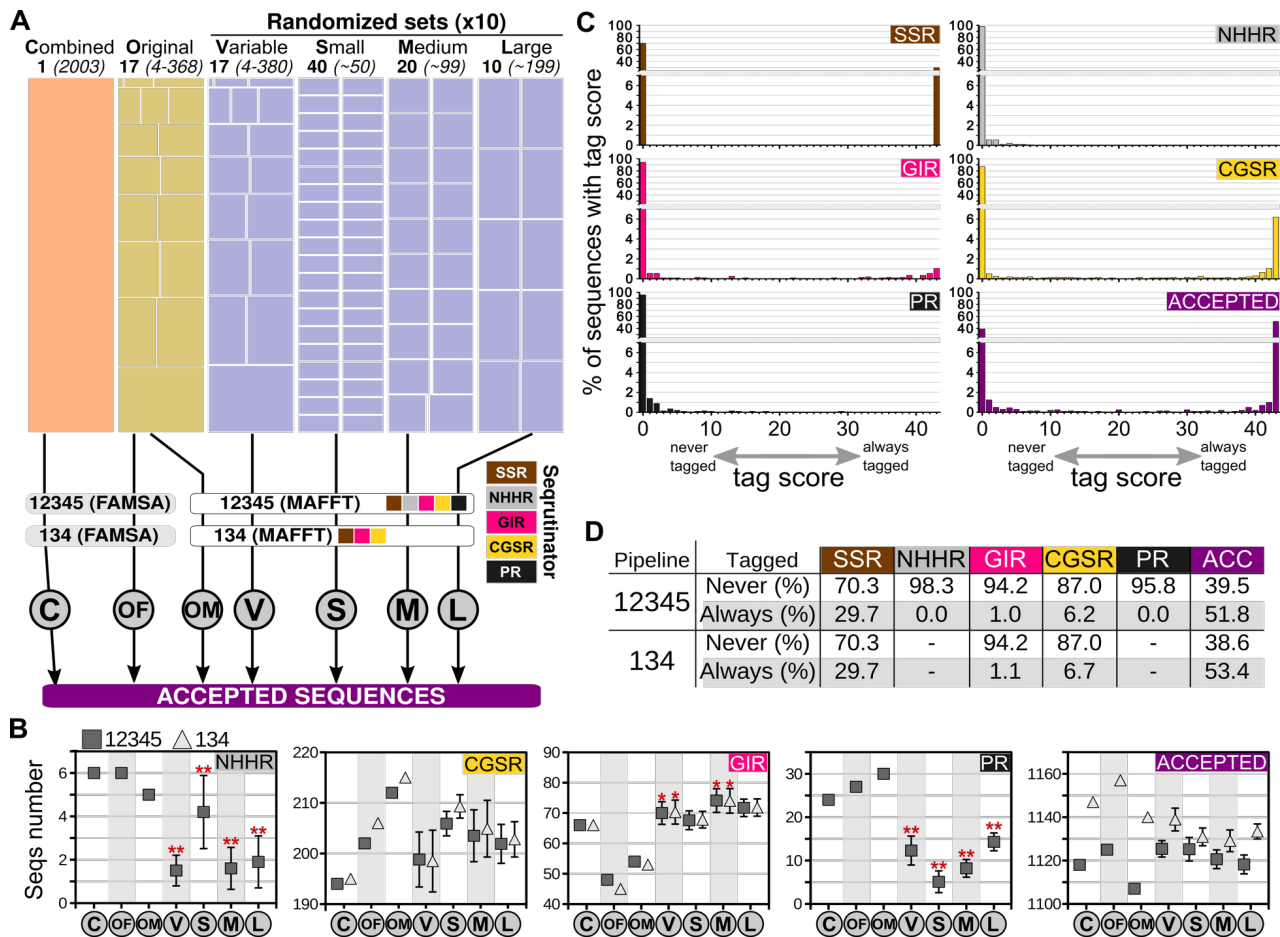


Figure SD3-1: The Effect of Set Size on Seqrutinator Performance. A: Design of Analysis. Four sets of randomized datasets representing a different condition (V Variable; S: Small; M: Medium; L: Large (L) sizes) were compared with the combined and the original dataset. Dataset subset number (in bold), and the number of sequences per subset (in parentheses) are indicated. Ten randomized sets for each condition were analysed using the standard pipeline (12345) as well as a pipeline that lacks outlier removers NHHR and PR (134). FAMSA was used in order to allow the analysis of the combined (C) dataset and used also for the original dataset (OM and OF; Original MAFFT and Original FAMSA) in order to allow comparison. **B: Net Effect of Dataset on BAHD Sequence Removal.** The number of sequences removed by each module in either the default pipeline or the pipeline without outlier removers was determined. C: Combined dataset (All sequences); OM: MAFFT-G-INS-i aligned original sequence sets; OF: FAMSA aligned original sequence sets; V Variable; S: Small; M: Medium; L: Large (L) sizes. For V, S, M and L ten randomized datasets were tested, average and s.d. are shown. Coefficient of variation, CV is indicated when >10% (**) or >5% (*). **C Default Sequence Fate.** The tag score is the number of times a sequence was either removed by each module or finally accepted, ranging from 0 to 43. The histograms show the incidence of the tag scores. **D: Final Sequence Fate.** Shown are the numbers of always rejected and always accepted for the 12345 and 134 pipelines.