Supplemental Material

SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes

Elmar Pruesse [1,2],
Jörg Peplies [3]
Frank Oliver Glöckner [1,2]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany,
[2]Jacobs University Bremen, Bremen, Germany
[3]Ribocon GmbH, 28359 Bremen, Germany

1 Algorithm

1.1 Positional Variability by Parsimony (PVP)

The “positional variability by parsimony” (PVP) function of ARB computes a per-column conservation profile from a MSA and a phylogenetic tree. For each column, the number of transitions and transversions required to explain the tree given the aligned sequence data is computed. The sum of transitions and transversions divided by the number of observed bases, capped at 0.55 and corrected for not observed mutations using the Jukes Cantor formula (Jukes and Cantor, 1969). From this rate, we compute the scoring weight as 0.5-log(rate). This weight is capped at 20. Columns containing gap characters in more than 80% of the sequences are assigned a weight of 1.

The PVP statistic applied may be chosen dynamically for each candidate sequence based on classification meta-data in the reference MSA database. This allows using for example domain specific statistics. The name of each PVP statistic stored in the reference database is compared with the configured classification attribute of each reference sequence. If a name is a prefix of the attribute for a majority of the sequences, the corresponding PVP statistic is used instead of a globally configured PVP statistic.

2 Results

All figures displaying mean accuracy (all except Fig. S2) are shown twice, once using a linear (labeled A) and once using a logarithmic scale on the y-axis (labeled B).

References

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *New York: Academic Press*, pages 21–132.

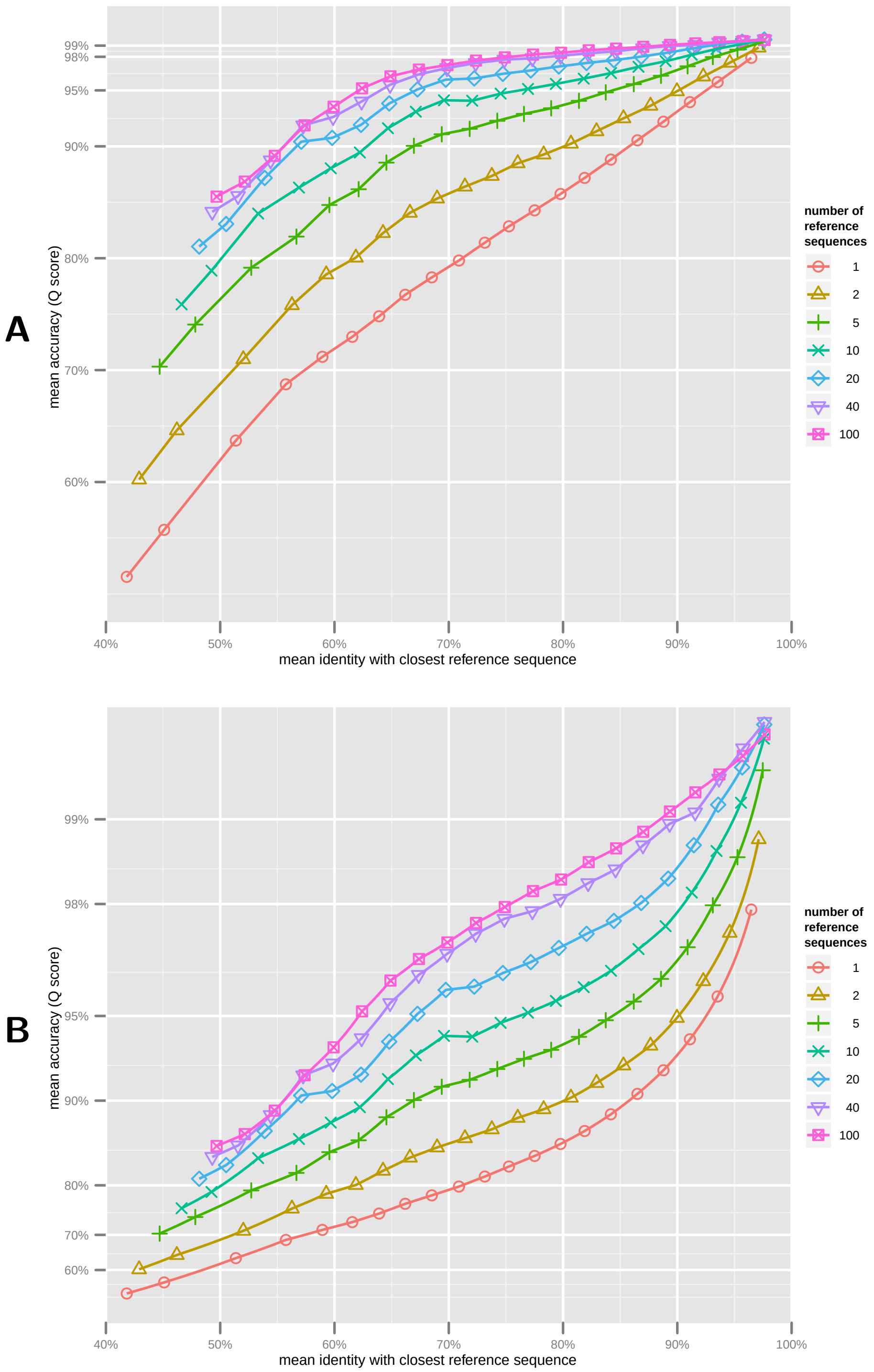

Figure S1: Effect of increasing the number of reference sequences on alignment accuracy at different levels of identity with the reference alignment.

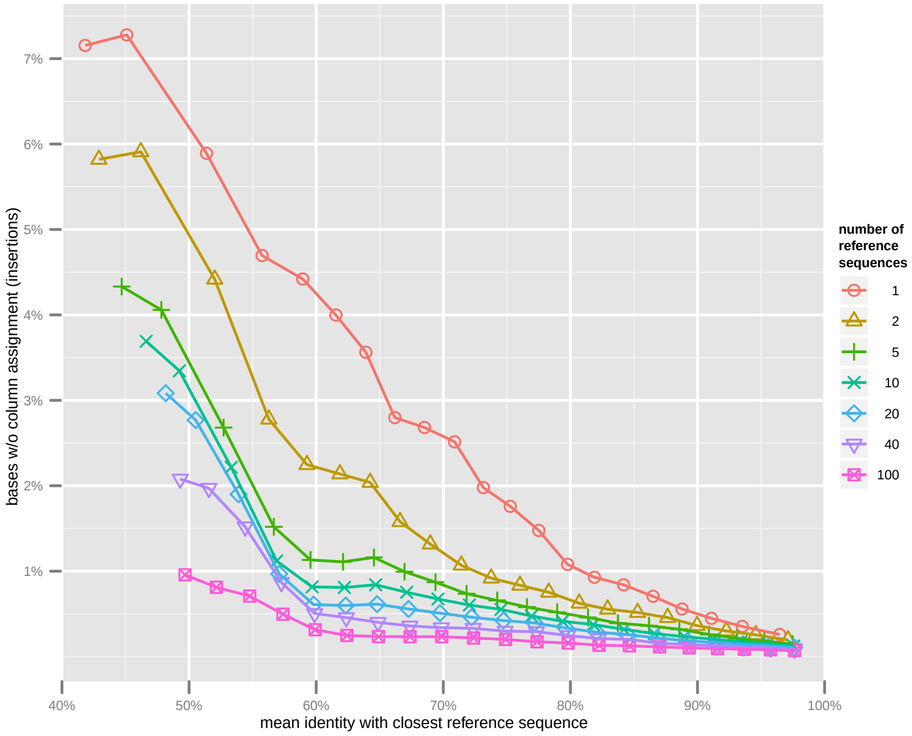

Figure S2: Effect of increasing the number of reference sequences on the fraction of "insertions" with respect to the alignment template.

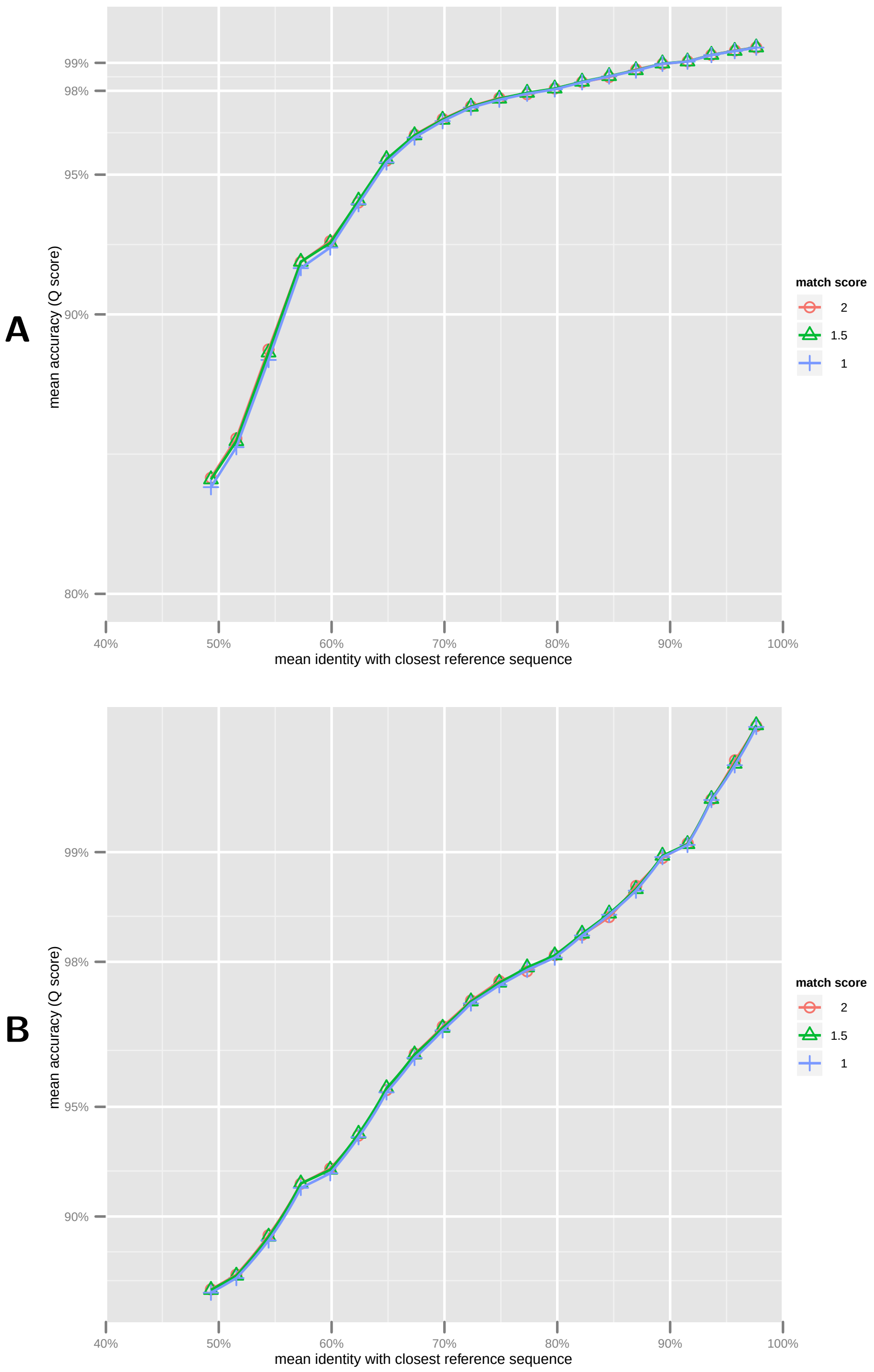

Figure S3: Using a mismatch score of -1 a match score of 2 is minimally better than a match score of 1 or 1.5. The difference, however, is almost beyond the resolution of this figure.

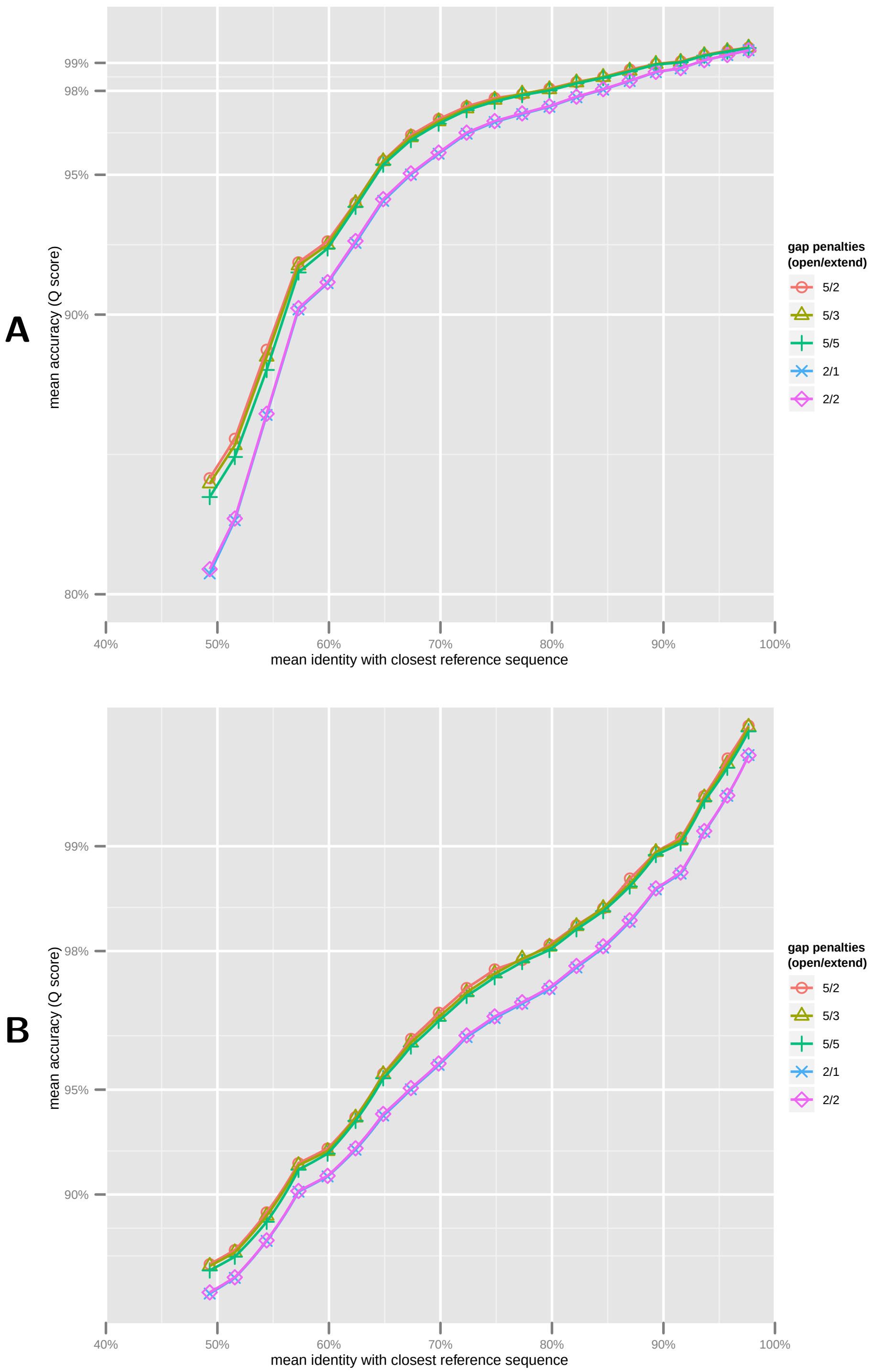

Figure S4: A gap open penalty of 5 works better than a gap open penalty of 2. Among the tested gap extend penalties using a gap open penalty of 5, a gap extension penalty of 2 is best by a small margin.

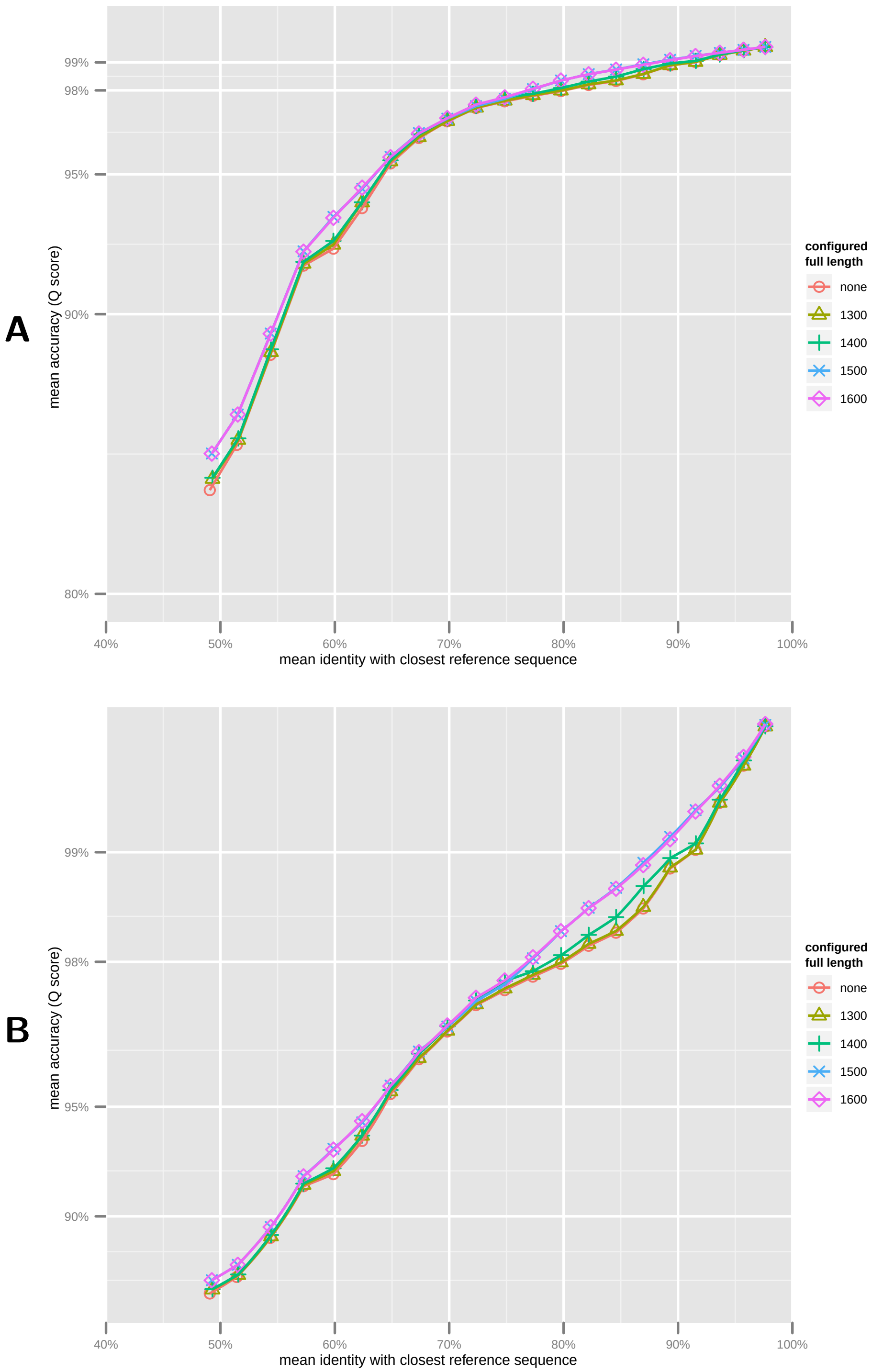

Figure S5: Requiring that at least one sequence of 1500 or 1600 bp length be included in the reference set improves averaged results.

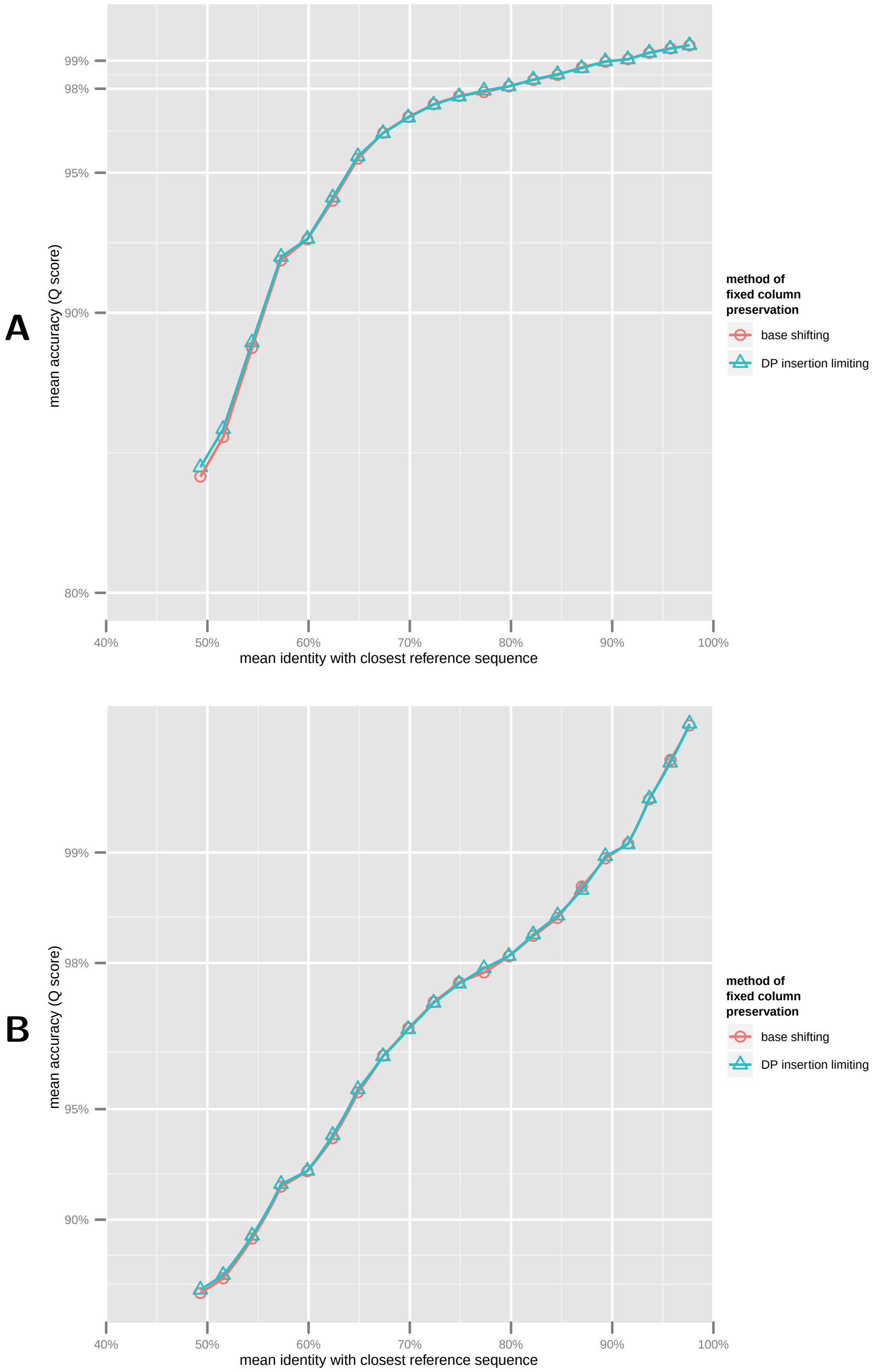

Figure S6: The method used for column preservation makes little difference to alignment accuracy using the SSU benchmark. Considering that the SSU alignment is more than 30 times wider than typical SSU sequences and that special care was taken to have sufficient alignment space between bases in the construction of the alignment, this is not unexpected.

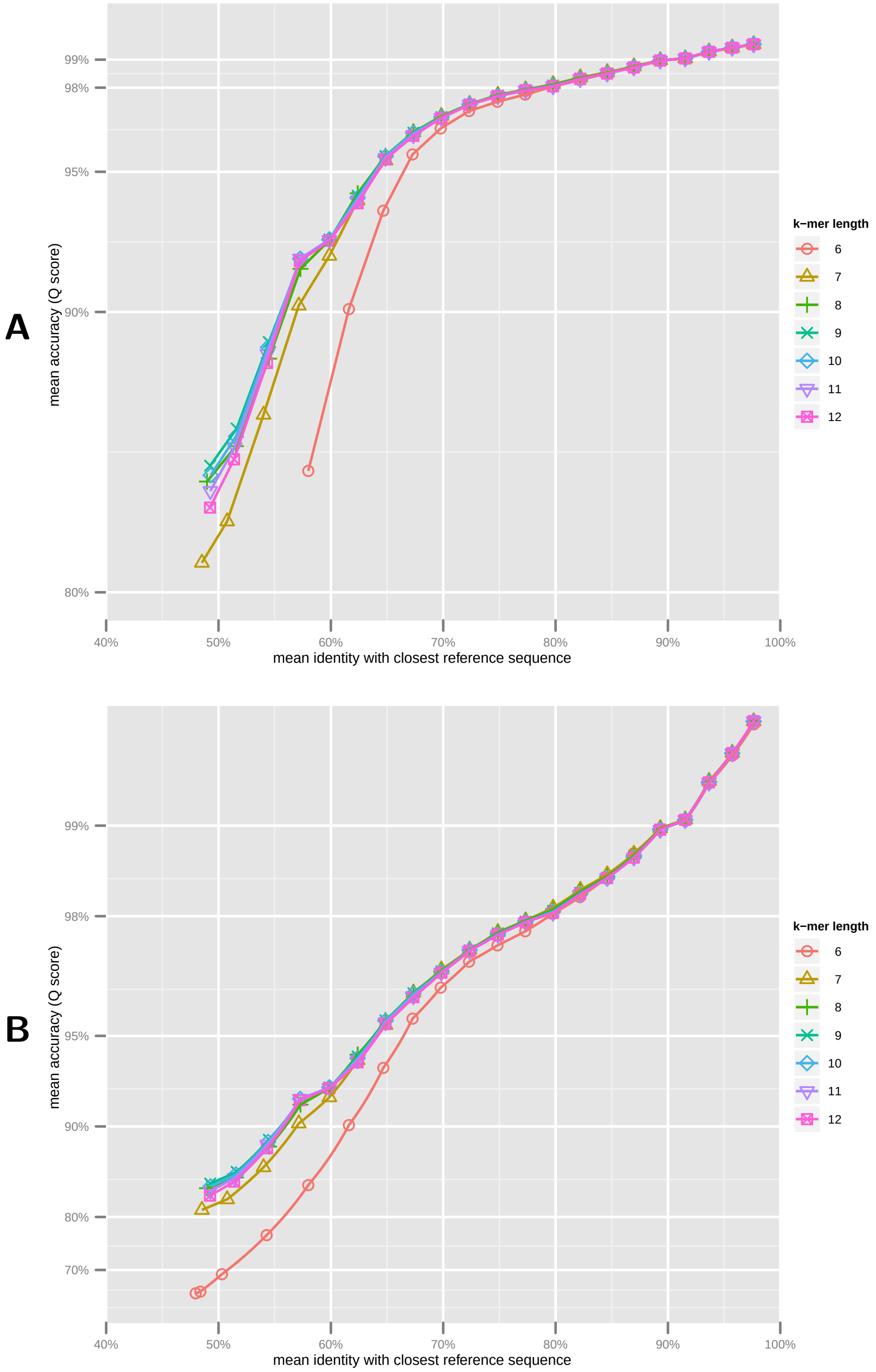

Figure S7: Varying the length of the *k*meres used to find reference sequences has little impact, values between 8 and 12 work well for SSU sequences.

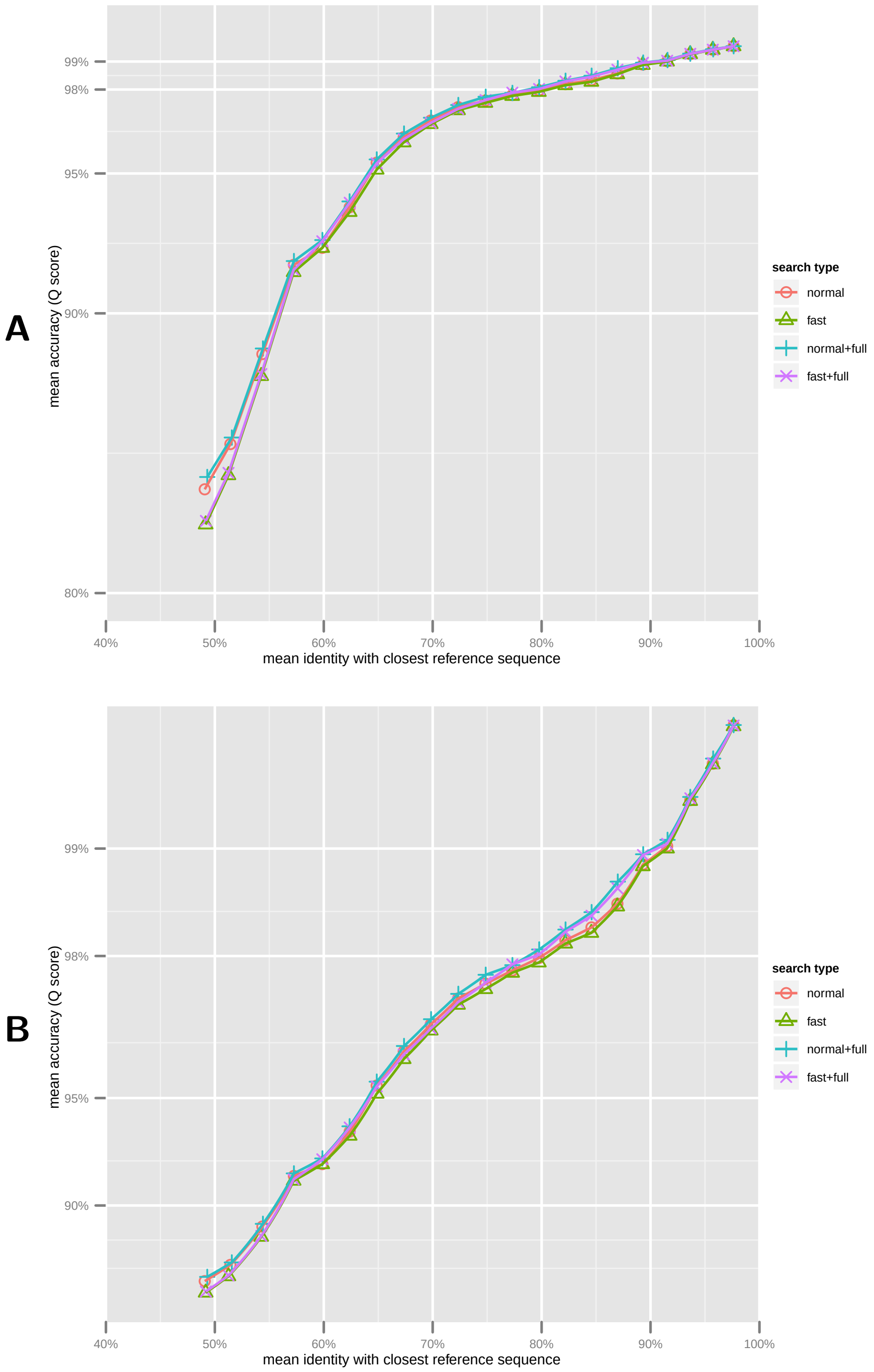

Figure S8: Using the "fast mode" of the $k$mer search provided by the ARB PT server ignores $k$meres that do not begin with 'A'. The impact to alignment accuracy is visible. The graphs labled "+full" require that at least one sequence of at least 1400 bp be included in the reference.

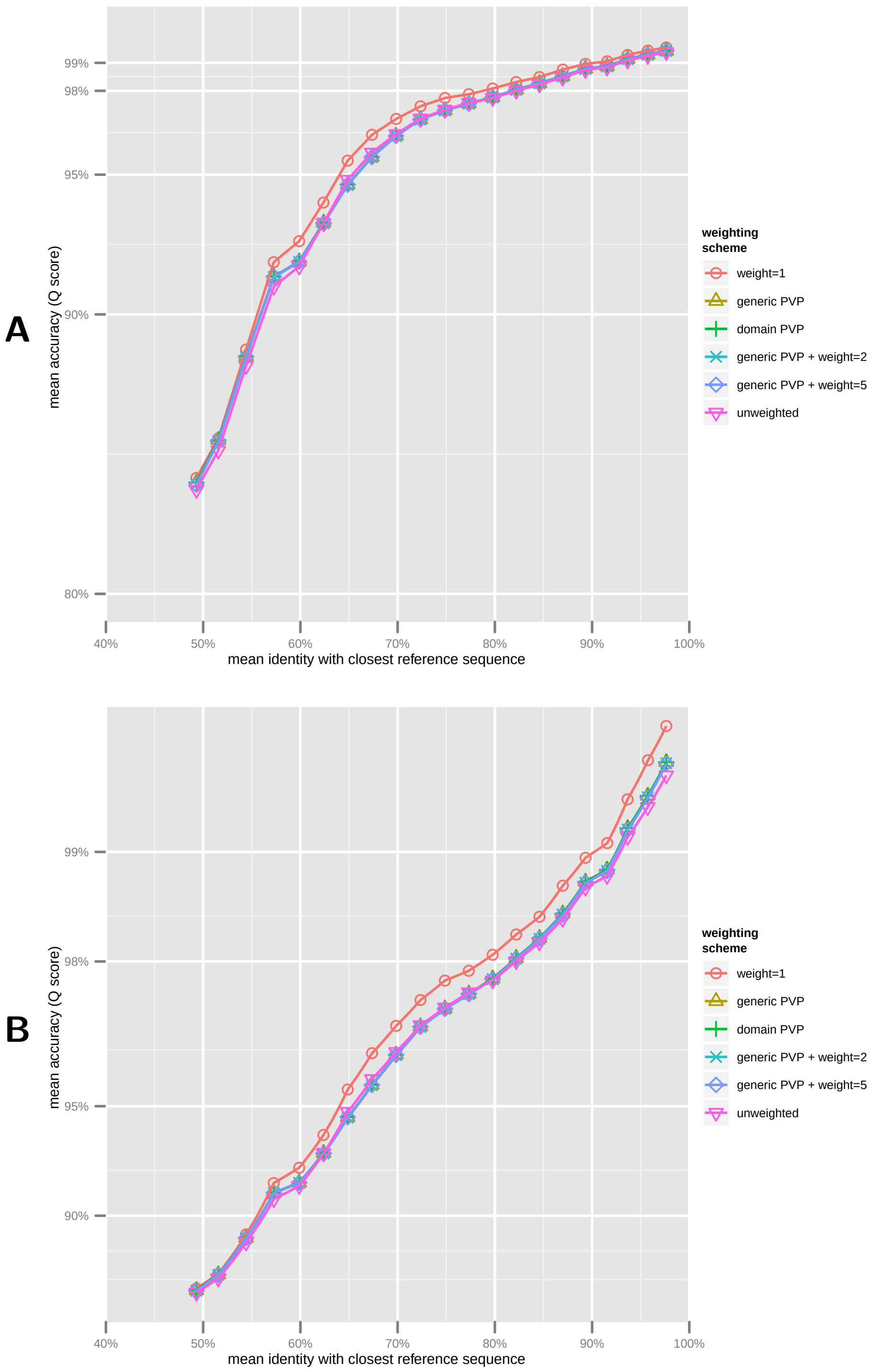

Figure S9: All weighting schemes and combinations thereof provide some improvement to unweighted alignment. Using only the base frequency among the selected reference sequences performs significantly better than all other tested methods.