Supplementary material to

Fast and Accurate Phylogeny Reconstruction using Filtered Spaced-Word Matches

Chris-André Leimeister, Salma Sohrabi-Jahromi, Burkhard Morgenstern

Submitted to OUP Bioinformatics, 2016

In our paper, we introduced a new way of estimating phylogenetic distances from genomic sequences. We first search for all *spaced-word matches* with respect to a given binary pattern P of *match* and *don't-care* positions. That is, we are looking for gapfree alignments of the same length as P where aligned nucleotides must *match* at the match positions while mismatches are possible at the *don't-care* positions. In short, we discard ('filter out') low-scoring and ambigous spaced-word matches and estimate the phylogenetic distance between the input sequences by looking at the nucleotides that are aligned to each other at the *don't-care* positions of the remaining spaced-word matches.

The number of match positions of the pattern P is called its *weight* w. For a given value of w, we calculate a pattern P with w *match* positions using our software tool *rasbhari* [2]. In our implementation, we are using a fixed number of 100 *don't-care* positions; the main parameter that can be adjusted by the user is the weight w; our default value is $w = 12$. The weight w influences the number of spaced-word matches that are considered and is therefore important for the program runtime. However, the value of w should not have any systematic influence on the estimated distances.

As one of the test cases in our paper, we applied our algorithm to a set of 13 bacterial genomes from the *Brucella* genus using values of $w = 10, 11, \ldots, 14$. With all these values, we obtained exactly the same tree topologie, which is in accordance with a tree published in the literature [1]; the trees are shown in the figures below.

References

[1] Jeffrey T. Foster, Stephen M. Beckstrom-Sternberg, Talima Pearson, James S. Beckstrom-Sternberg, Patrick S G Chain, Francisco F. Roberto, Jonathan Hnath, Tom Brettin, and Paul Keim. Whole-genome-based phylogeny and divergence of the genus brucella. *Journal of Bacteriology*, 191:2864–2870, 4 2009.

[2] Lars Hahn, Chris-André Leimeister, Rachid Ounit, Stefano Lonardi, and Burkhard Morgenstern. *rasbhari*: optimizing spaced seeds for database searching, read mapping and alignment-free sequence comparison. *PLOS Computational Biology*, 12(10):e1005107, 2016.

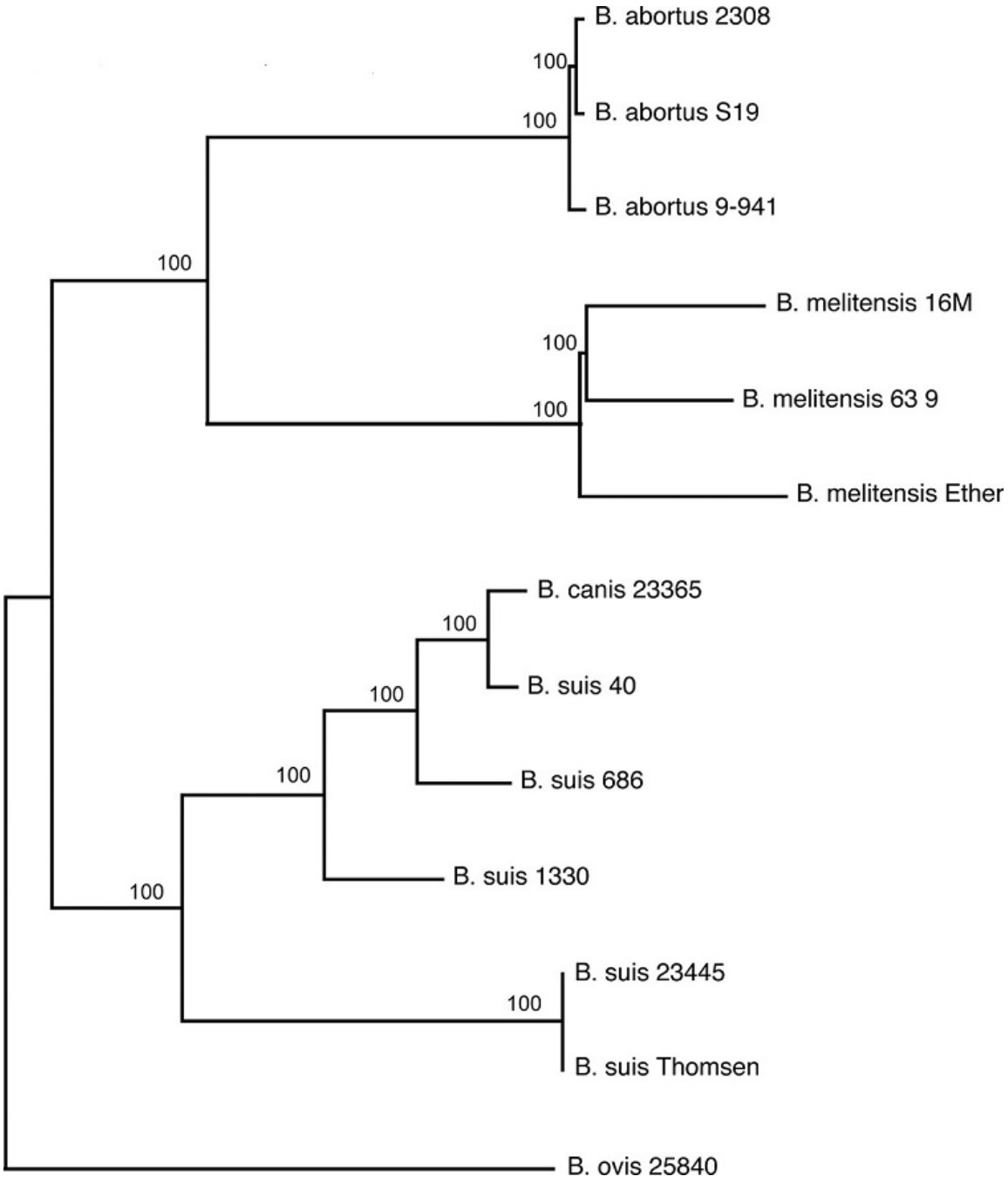

Figure 1: Reference tree from [1].

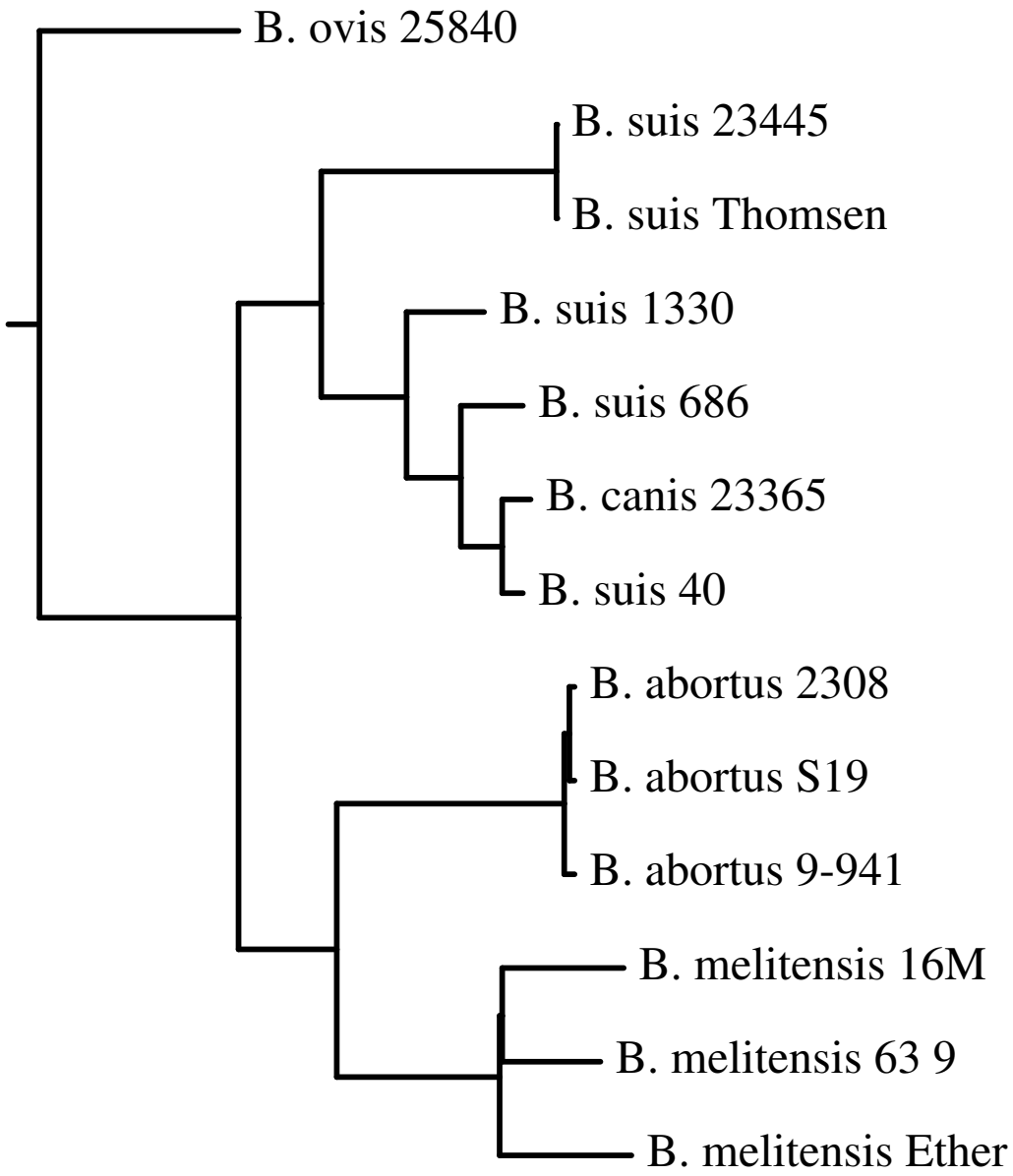

Figure 2: Tree calculated with distance matrix from $FSWM$ with $w = 10$

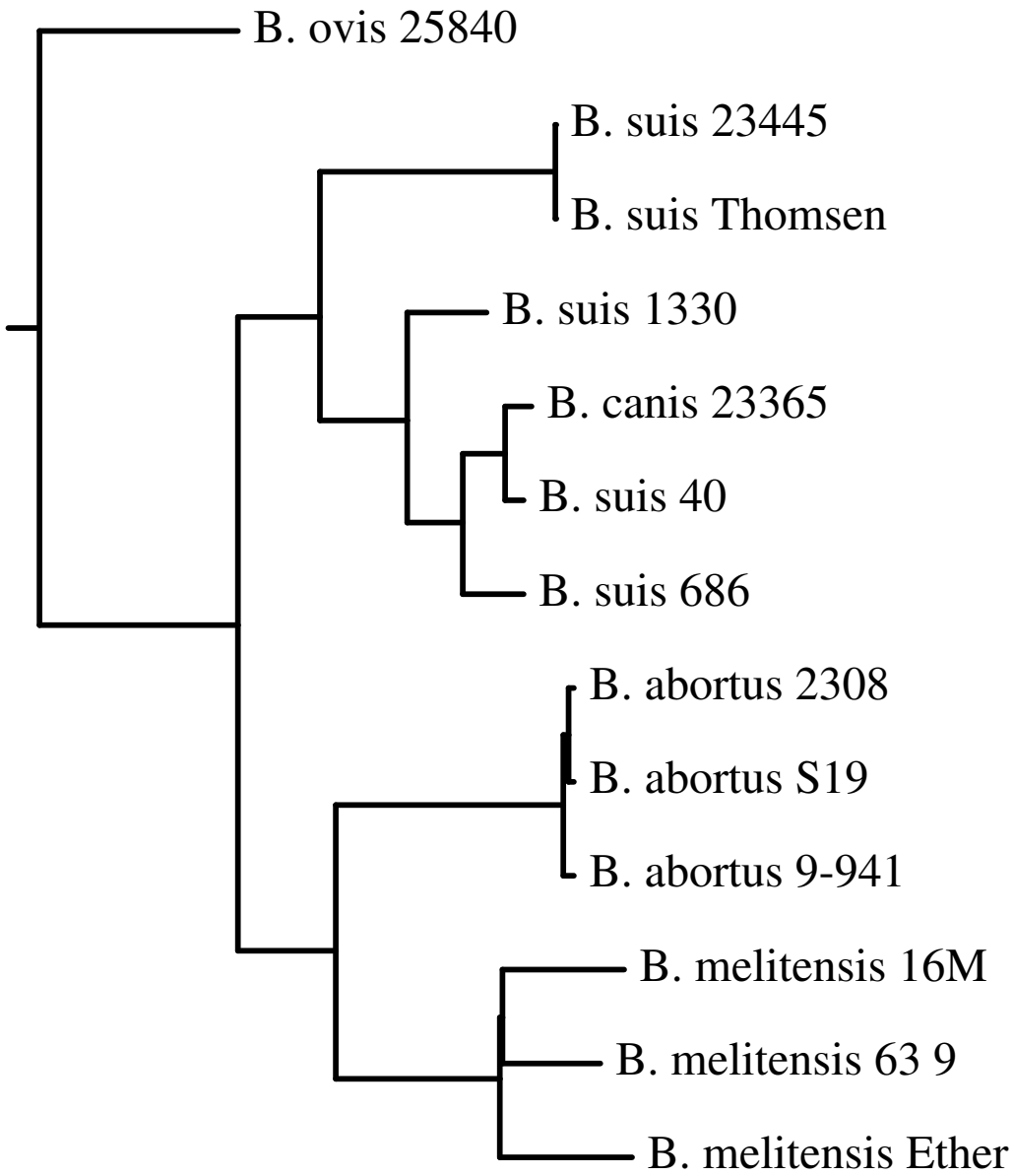

Figure 3: Tree calculated with distance matrix from $FSWM$ with $w = 11$

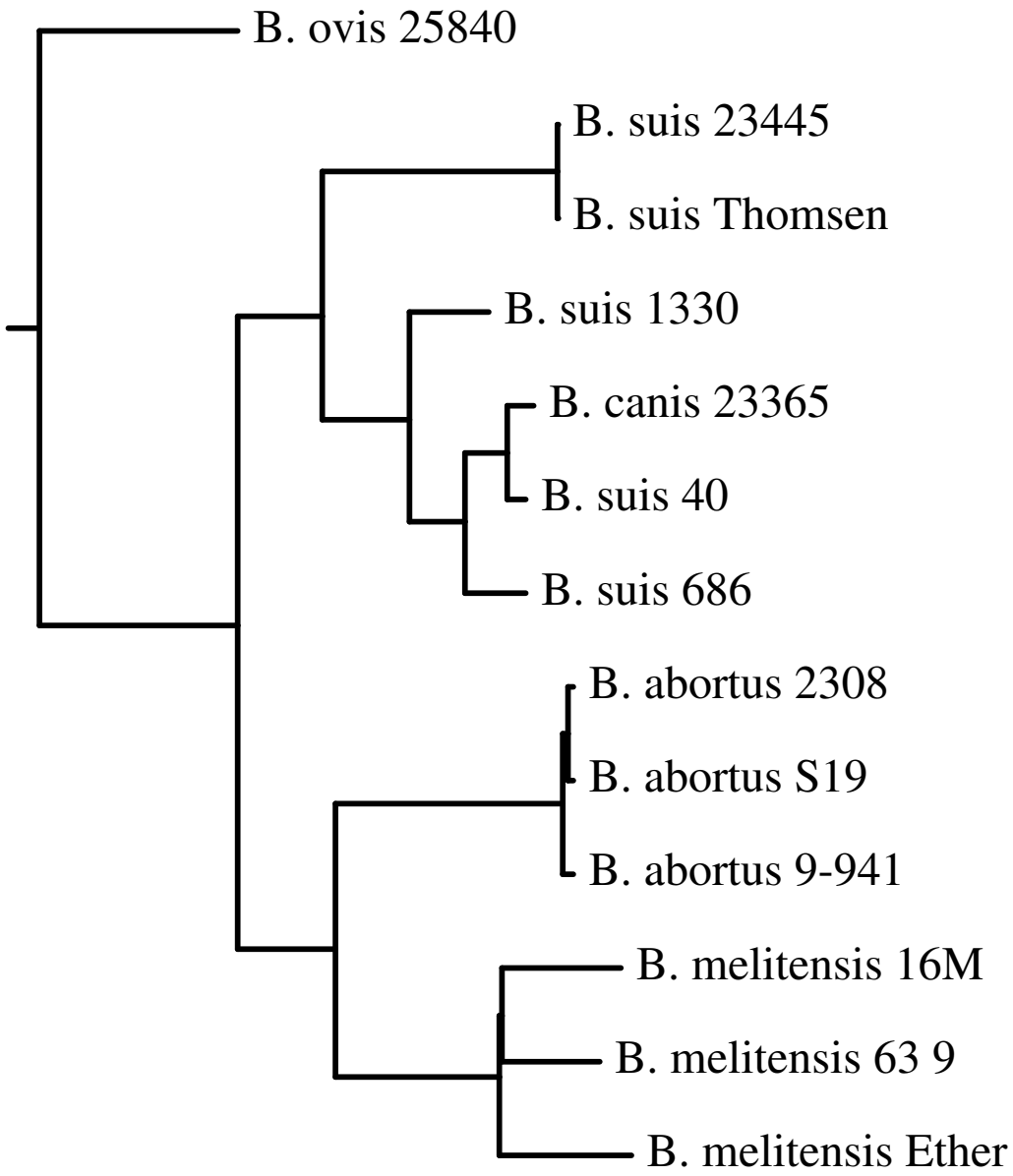

Figure 4: Tree calculated with distance matrix from *FSWM* with $w = 12$ (default value)

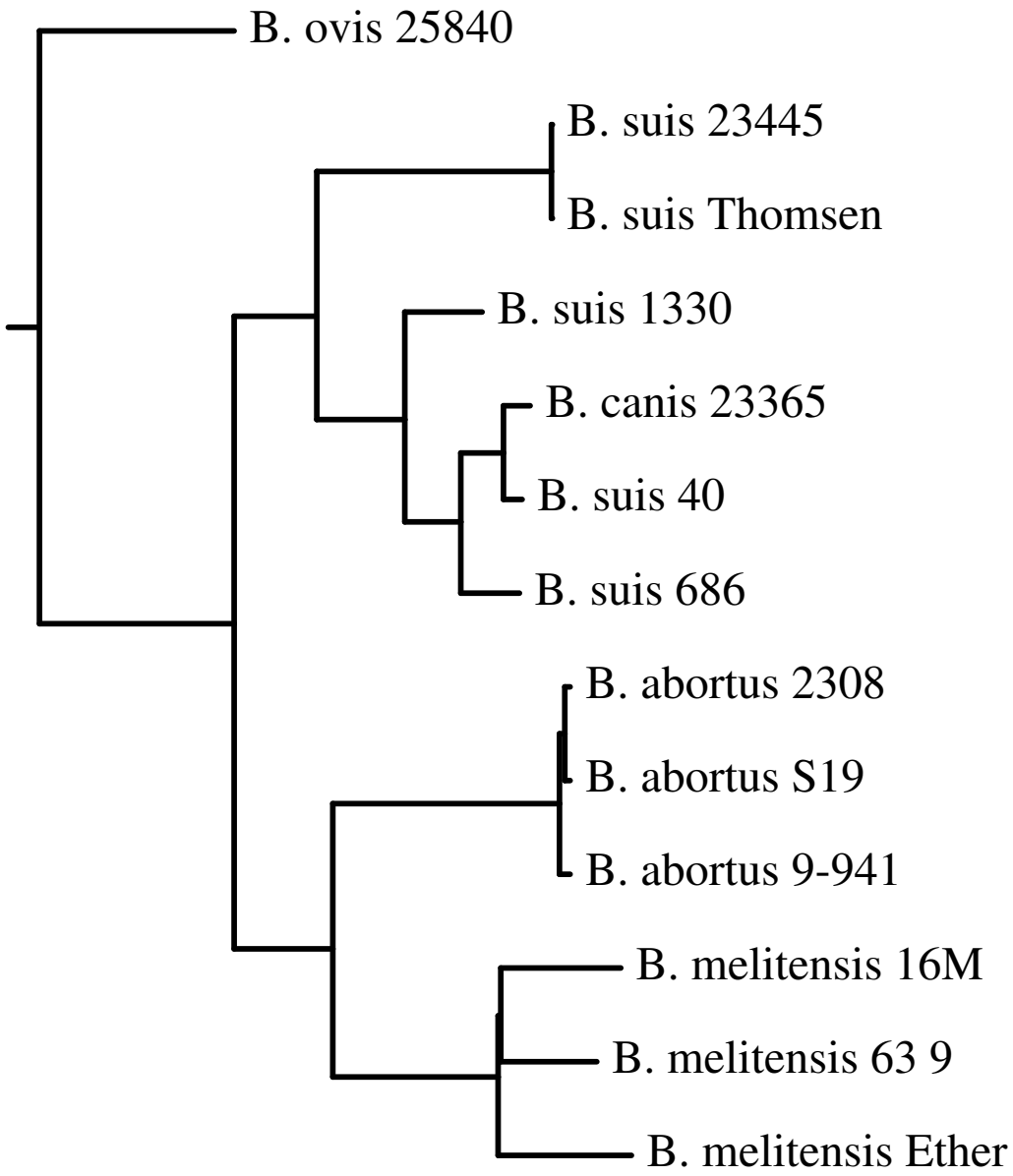

Figure 5: Tree calculated with distance matrix from $FSWM$ with $w = 13$

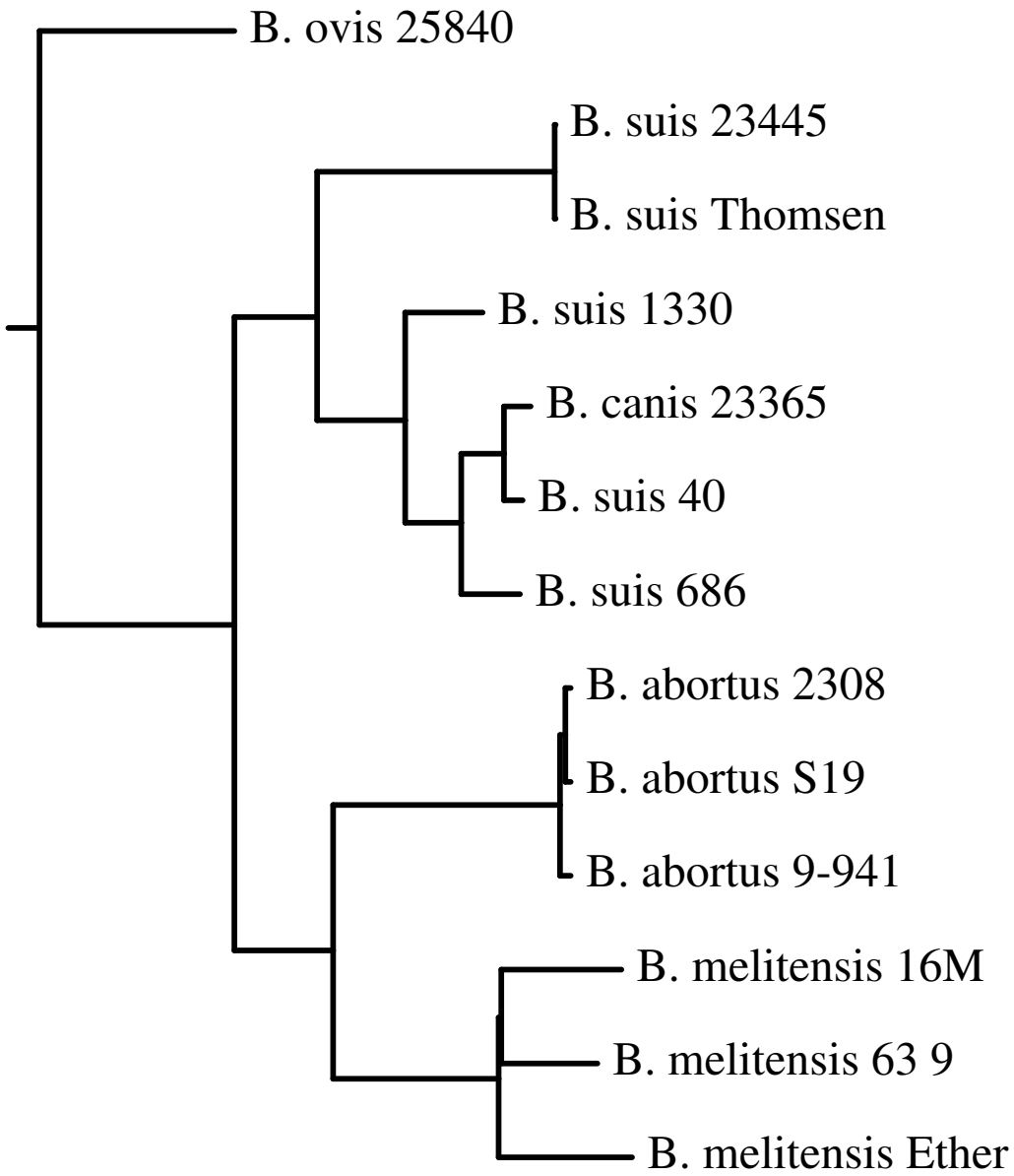

Figure 6: Tree calculated with distance matrix from $FSWM$ with $w = 14$