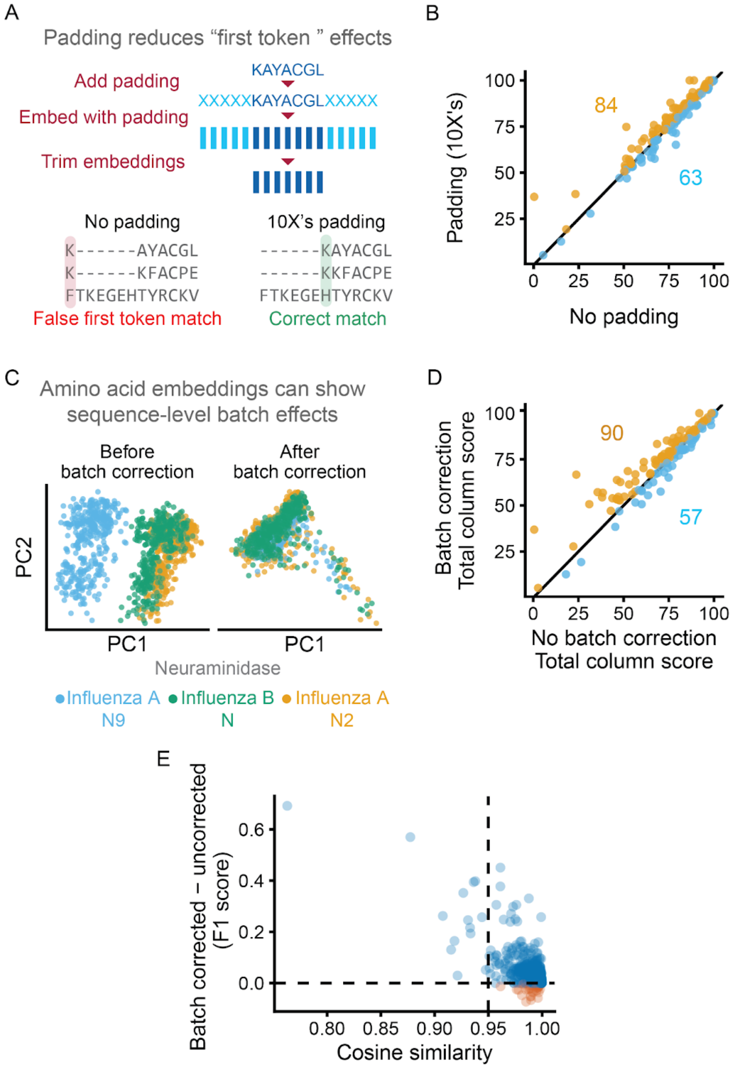
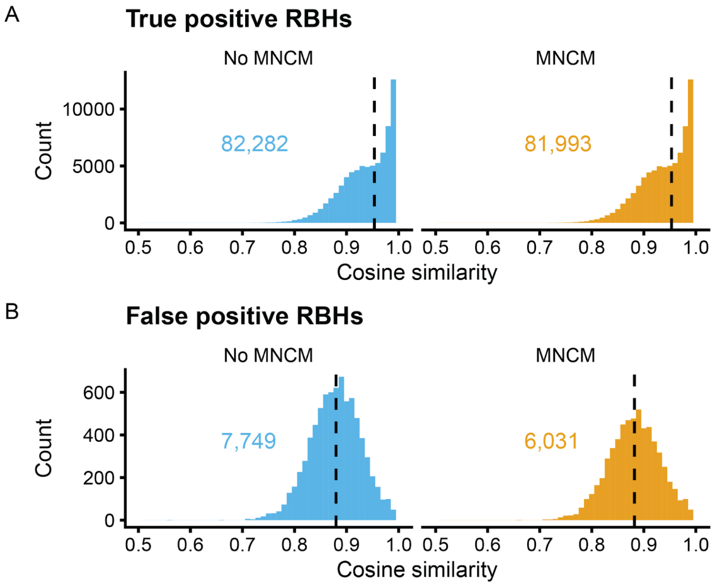
SUPPLEMENTAL MATERIAL

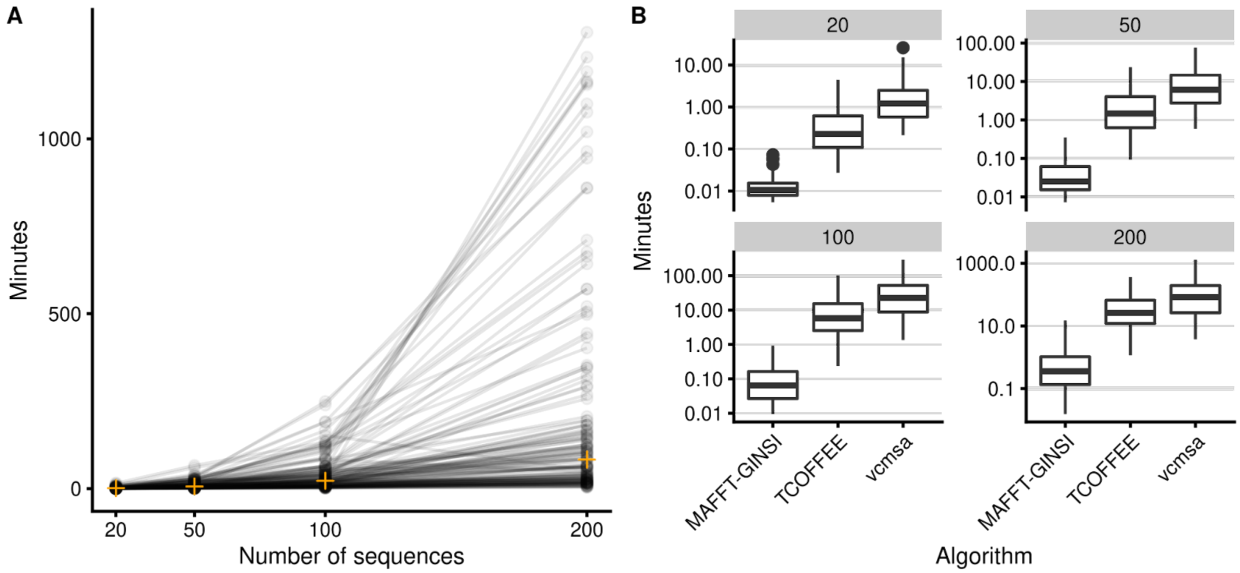
Supplemental Figures



Supplemental Figure 1: Illustration of sequence padding and sequence-specific effect correction. (A) For sequence padding, ten X's are added to the start and end of each sequence prior to embedding. Embedding positions corresponding to these twenty X's are trimmed prior to use. (B) Adding padding generally increases the Total Column scores for the benchmarking set of 147 alignments of 20 proteins. (C) Amino acid embeddings can show sequence-specific effects. (D) When run on our benchmark set, sequence-specific effect correction has a beneficial effect on alignment accuracy. (E) For pairs of sequences in the HOMSTRAD dataset (each pair is shown as a dot), batch correction generally increases $F1\, score = (2 * precision * recall)/(precision + recall)$ of reciprocal best hits as compared to gold standard alignments, particularly for pairs of sequences with cosine similarity of less than 0.95. For each pair of sequences, precision is computed as the fraction of MNCM-filtered RBHs that aligned together in the gold standard, and recall is computed as the fraction of aligned amino acids in the gold standard alignment that are in the MNCM-filtered RBHs.



Supplemental Figure 2. Maximum non-crossing matching (MNCM) without sequence-specific effect correction removes false positive reciprocal best hits (RBHs). (A) (Left) Histogram of cosine similarity scores of RBHs that are in the gold standard alignments. (Right) Histogram of cosine similarity scores of RBHs that are in the gold standard alignments, after MNCM filtering removes RBH pairs. MNCM keeps 99.6% of the initial true positive RBH pairs. (B) (Left) Histogram of cosine similarity scores of RBHs that are not in the gold standard alignments. (Right) Histogram of cosine similarity scores of RBHs that are not in the gold standard alignments, after MNCM removes RBH pairs. MNCM filtering removes 22.2% of the initial false positive RBHs. In each histogram, vertical dotted lines show the median cosine similarity value of the depicted RBHs.



Supplemental Figure 3: (A) Relationship between number of input sequences and time to align them for 100 different sequence-specific effect corrected protein alignments. Each line represents one protein with 20, 50, 100, and 200 sequences, and orange crosses show median values for each alignment size. Occasionally, additional sequences can simplify the clustering problem, reducing the number of iterations and time required to perform the alignment. vcMSA is run on a NVIDIA A100 GPU and a 2.6 GHz AMD EPYC Rome CPU with 64 GB of memory. (B) Relative time to completion for 20, 50, 100, and 200 sequences by MAFFT-GINSI, T-Coffee and vcMSA, shown as boxplots. MAFFT-GINSI and T-Coffee are run on a 2.6 GHz AMD EPYC Rome CPU with 64 GB of memory.