Supplementary Materials for " COATi: statistical pairwise alignment of protein coding sequences"

Juan J. Garcia Mesa, Ziqi Zhu, Reed A. Cartwright

Table 1: Accuracy of COATi codon-triplet-mg, PRANK, MAFFT, ClustalOmega, and MACSE on 7719 simulated sequence pairs. Perfect alignments have $d_{seq} = 0$, best alignments have lowest d_{seq}, and imperfect alignments have $d_{seq} > 0$ when at least one aligner found a perfect alignment.

	tri-mg	MAFFT	PRANK*	MACSE	ClustalOmega
d_{seq}	0.00214	0.01392	0.02001	0.01351	0.02691
Perfect alignments	5419.00000	5217.00000	4659.00000	3743.00000	2913.00000
Best alignments	6151.00000	5513.00000	4996.00000	3880.00000	2967.00000
Imperfect alignments	1147.00000	1350.00000	1883.00000	2824.00000	3654.00000
F1-score pos selection	0.98238	0.86069	0.88445	0.81206	0.70909
F1-score neg selection	0.99822	0.98556	0.98838	0.98282	0.97030

* PRANK produced 50 empty alignments, calculations are based on 7669 alignments.

Table 2: Accuracy of COATi codon-triplet-ecm, PRANK, MAFFT, ClustalOmega, and MACSE on 7678 simulated sequence pairs. Perfect alignments have $d_{seq} = 0$, best alignments have lowest d_{seq}, and imperfect alignments have $d_{seq} > 0$ when at least one aligner found a perfect alignment.

	tri-ecm	MAFFT	PRANK*	MACSE	ClustalOmega
d_{seq}	0.00244	0.01462	0.02051	0.01366	0.02838
Perfect alignments	5195.00000	5143.00000	4618.00000	3733.00000	2923.00000
Best alignments	5901.00000	5433.00000	5030.00000	3874.00000	2985.00000
Imperfect alignments	1257.00000	1310.00000	1814.00000	2720.00000	3530.00000
F1-score pos selection	0.97196	0.84955	0.87737	0.80256	0.71317
F1-score neg selection	0.99721	0.98456	0.98792	0.98249	0.97111

* PRANK produced 41 empty alignments, calculations are based on 7637 alignments.

Table 3: Accuracy of COATi codon-marginal-mg, PRANK, MAFFT, ClustalOmega, and MACSE on 7661 simulated sequence pairs. Perfect alignments have $d_{seq} = 0$, best alignments have lowest d_{seq}, and imperfect alignments have $d_{seq} > 0$ when at least one aligner found a perfect alignment.

	mar-mg	MAFFT	PRANK*	MACSE	ClustalOmega
d_{seq}	0.00201	0.01374	0.01813	0.01310	0.02631
Perfect alignments	4979.00000	5157.00000	4711.00000	3704.00000	2869.00000
Best alignments	5686.00000	5432.00000	5037.00000	3842.00000	2927.00000
Imperfect alignments	1551.00000	1374.00000	1792.00000	2827.00000	3662.00000
F1-score pos selection	0.96442	0.85775	0.89164	0.81797	0.72281
F1-score neg selection	0.99649	0.98547	0.98924	0.98356	0.97158

* PRANK produced 50 empty alignments, calculations are based on 7611 alignments.

Table 4: Accuracy of COATi codon-marginal-ecm, PRANK, MAFFT, ClustalOmega, and MACSE on 7710 simulated sequence pairs. Perfect alignments have $d_{seq} = 0$, best alignments have lowest d_{seq}, and imperfect alignments have $d_{seq} > 0$ when at least one aligner found a perfect alignment.

	mar-ecm	MAFFT	PRANK*	MACSE	ClustalOmega
d_{seq}	0.00209	0.01436	0.02099	0.01382	0.02827
Perfect alignments	5037.00000	5125.00000	4662.00000	3702.00000	2808.00000
Best alignments	5735.00000	5447.00000	5002.00000	3837.00000	2869.00000
Imperfect alignments	1496.00000	1408.00000	1852.00000	2831.00000	3725.00000
F1-score pos selection	0.97003	0.85421	0.88352	0.79471	0.71379
F1-score neg selection	0.99692	0.98474	0.98824	0.98132	0.97045

* PRANK produced 35 empty alignments, calculations are based on 7675 alignments.

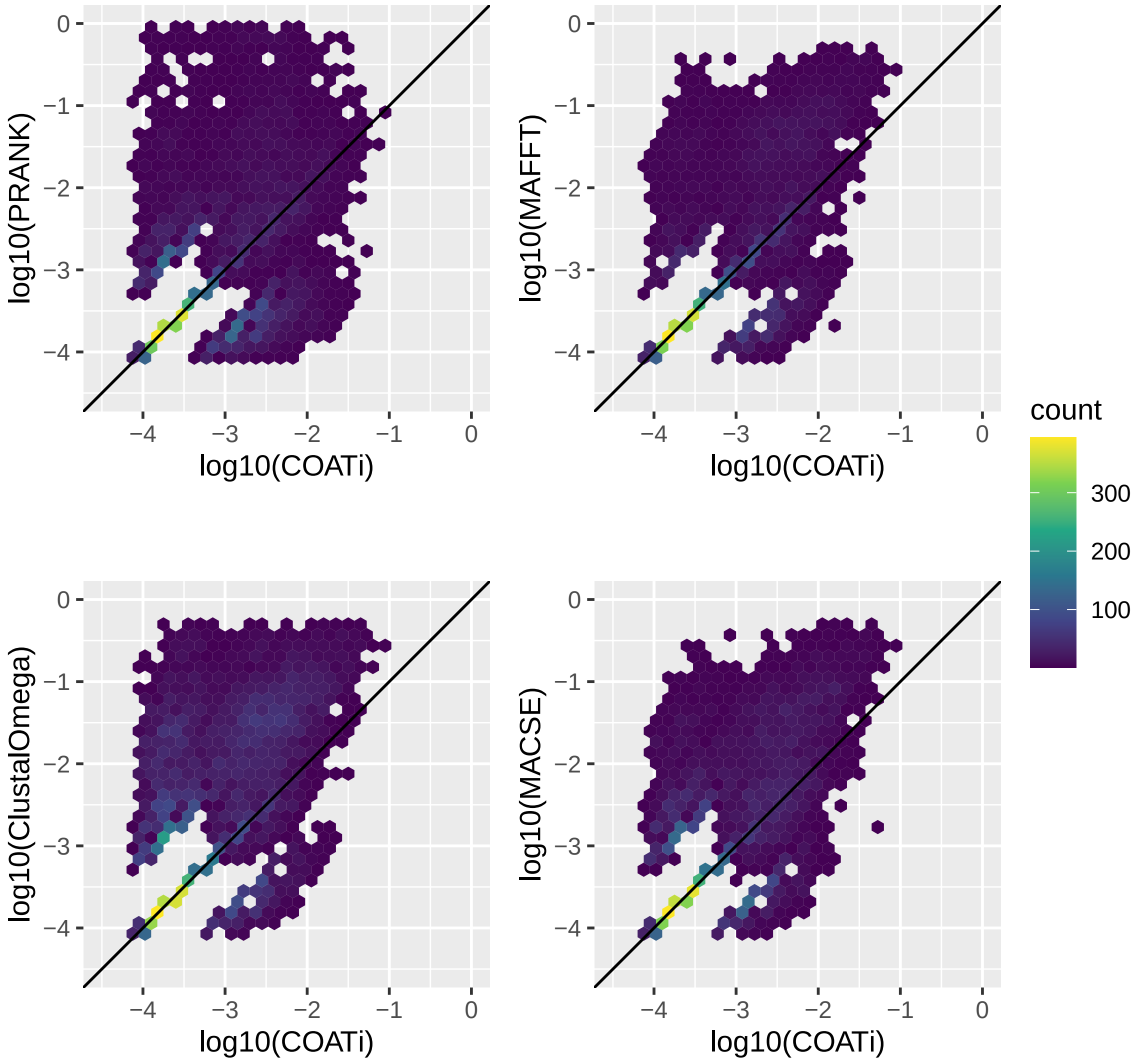


Figure 1: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-triplet-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 2.06e - 79$.

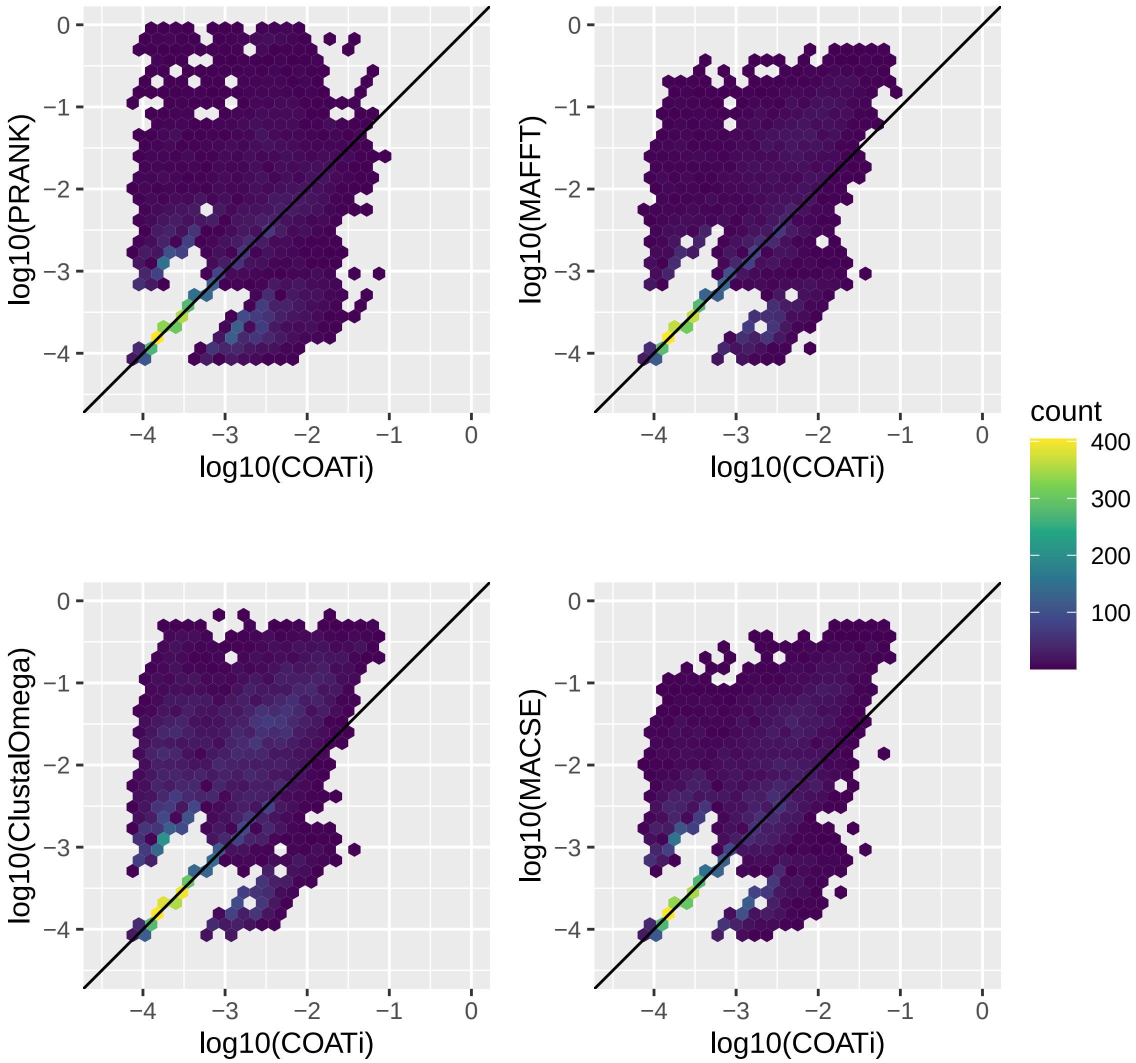


Figure 2: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-triplet-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 8.15e-53$.

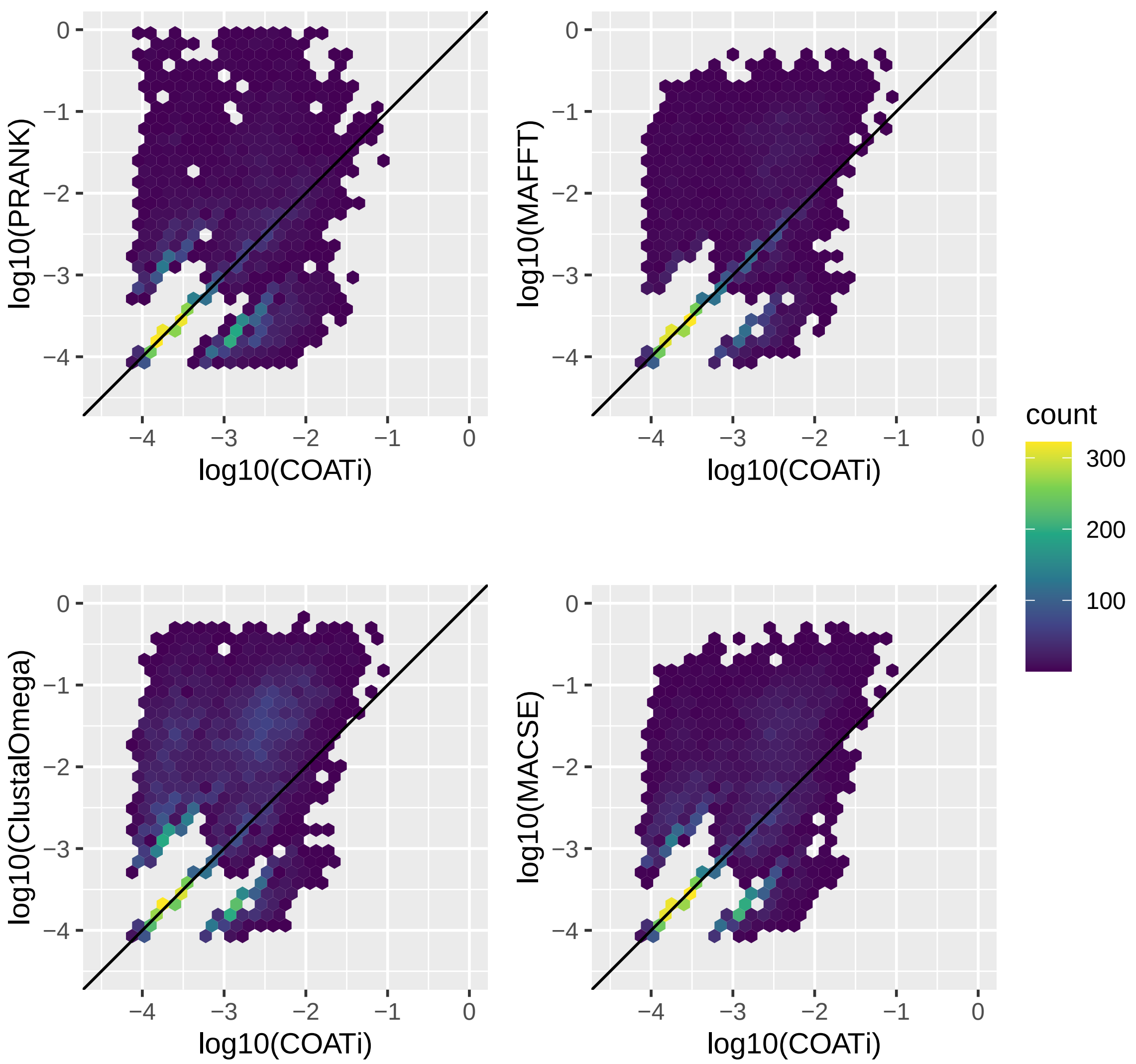


Figure 3: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-marginal-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 6.07e-40$.

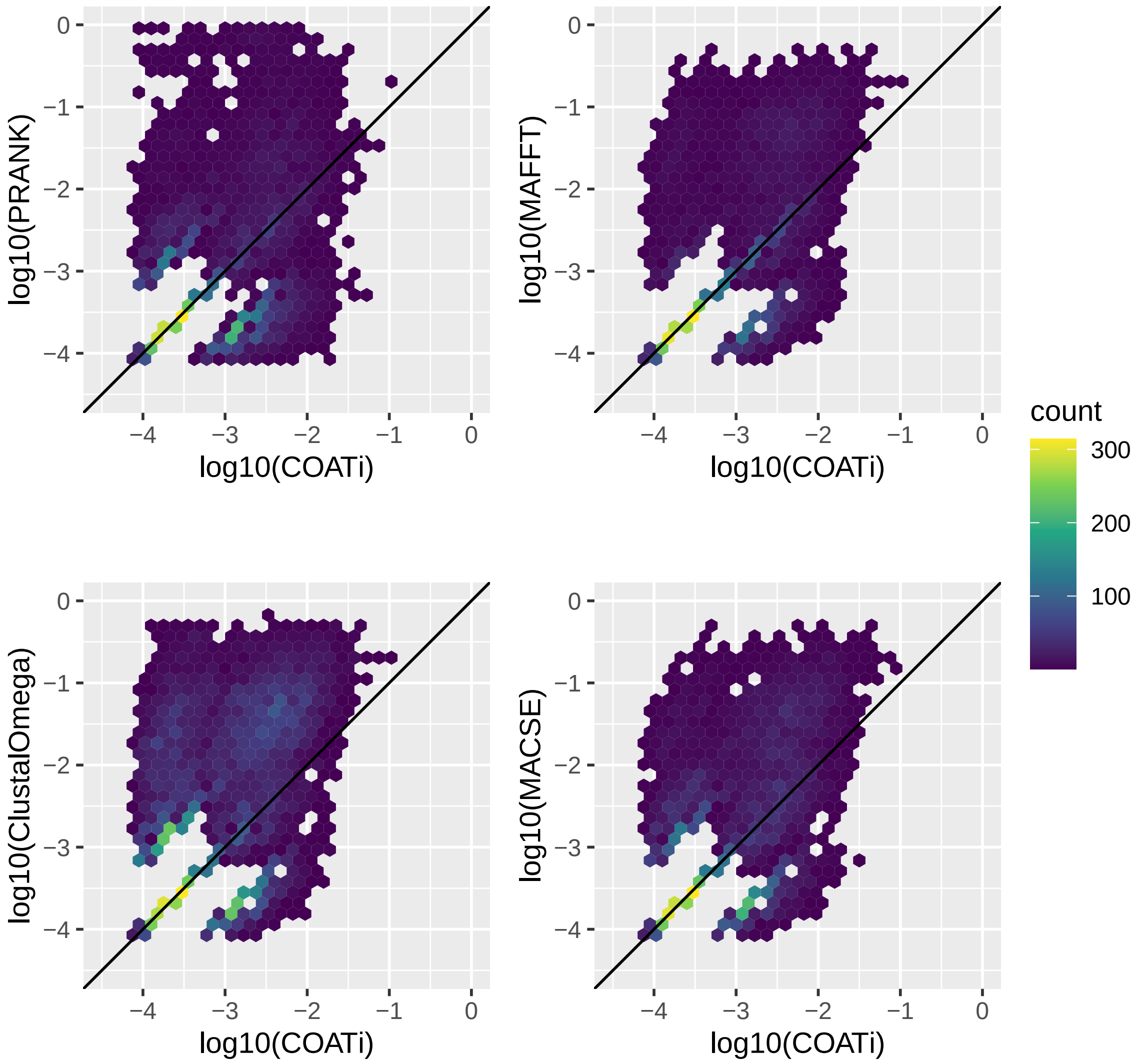
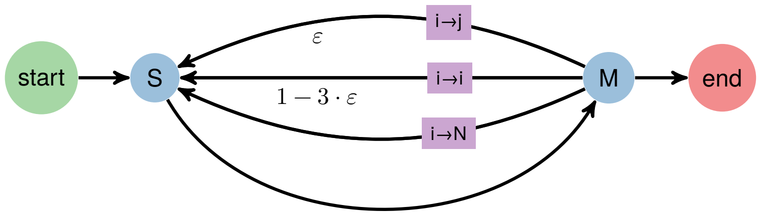


Figure 4: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-marginal-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 2.12e - 42$.



Sequences	
$i \to i$:	matching input to intermediate base
$i \to j$:	mismatching input to intermediate base
$i \to N$:	input to intermediate ambiguous base N

Parameters	
ε:	base calling error weight

Figure 5: Base calling error FST. Arcs from M to S generate matches; however, here they can introduce single-nucleotide errors, which can generate stop codon artifacts.

Supplementary Methods

Ks and Ka represent the number of substitutions per synonymous and non-synonymous sites. The ratio of nonsynonymous (Ka) to synonymous (Ks) nucleotide substitution rates indicates the selective pressures acting on genes. If the ratio is significantly greater than 1, it suggests positive selective pressure, meaning that nonsynonymous substitutions occur more frequently than synonymous substitutions. A ratio around 1 can indicate either neutral evolution at the protein level or a mixture of positive and negative selective pressures. If the ratio is less than 1, it indicates a pressure to maintain protein sequence, known as purifying selection. Ks and Ka are calculated using the R package seqinr v.4.2-30 (Charif and Lobry 2007).

References

Charif, D., and J. R. Lobry. 2007. "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis." In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, edited by U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, 207–32. Biological and Medical Physics, Biomedical Engineering. New York: Springer Verlag.