# 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 the same score as the true alignment, best alignments have lowest $d_{seq}$, and imperfect alignments have a different score than the true alignment when at least one method found a perfect alignment.

| | tri-mg | MAFFT | PRANK* | MACSE | ClustalOmega |
|---|---|---|---|---|---|
| $d_{seq}$ | 0.00214 | 0.01392 | 0.02001 | 0.01351 | 0.02691 |
| Perfect alignments | 5722.00000 | 5408.00000 | 4706.00000 | 2860.00000 | 2937.00000 |
| Best alignments | 5152.00000 | 4833.00000 | 4748.00000 | 3754.00000 | 2595.00000 |
| Imperfect alignments | 1066.00000 | 1380.00000 | 2082.00000 | 3928.00000 | 3851.00000 |
| 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 the same score as the true alignment, best alignments have lowest $d_{seq}$, and imperfect alignments have a different score than the true alignment when at least one method found a perfect alignment.

| | tri-ecm | MAFFT | PRANK* | MACSE | ClustalOmega |
|---|---|---|---|---|---|
| $d_{seq}$ | 0.00244 | 0.01462 | 0.02051 | 0.01366 | 0.02838 |
| Perfect alignments | 5524.00000 | 5350.00000 | 4664.00000 | 2864.00000 | 2970.00000 |
| Best alignments | 4957.00000 | 4731.00000 | 4764.00000 | 3731.00000 | 2652.00000 |
| Imperfect alignments | 1189.00000 | 1363.00000 | 2049.00000 | 3849.00000 | 3743.00000 |
| 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 7666 simulated sequence pairs. Perfect alignments have the same score as the true alignment, best alignments have lowest $d_{seq}$, and imperfect alignments have a different score than the true alignment when at least one method found a perfect alignment.

| | mar-mg | MAFFT | PRANK* | MACSE | ClustalOmega |
|---|---:|---:|---:|---:|---:|
| $d_{seq}$ | 0.00216 | 0.01564 | 0.01927 | 0.01474 | 0.02968 |
| Perfect alignments | 5678.00000 | 5208.00000 | 4733.00000 | 2799.00000 | 2891.00000 |
| Best alignments | 5348.00000 | 4684.00000 | 4860.00000 | 3794.00000 | 2576.00000 |
| Imperfect alignments | 1055.00000 | 1525.00000 | 2000.00000 | 3934.00000 | 3842.00000 |
| F1-score pos selection | 0.98451 | 0.83737 | 0.88698 | 0.79310 | 0.68648 |
| F1-score neg selection | 0.99842 | 0.98308 | 0.98858 | 0.98122 | 0.96817 |

* PRANK produced 47 empty alignments, calculations are based on 7619 alignments.

Table 4: Accuracy of COATi codon-marginal-ecm, PRANK, MAFFT, ClustalOmega, and MACSE on 7717 simulated sequence pairs. Perfect alignments have the same score as the true alignment, best alignments have lowest $d_{seq}$, and imperfect alignments have a different score than the true alignment when at least one method found a perfect alignment.

| | mar-ecm | MAFFT | PRANK* | MACSE | ClustalOmega |
|---|---:|---:|---:|---:|---:|
| $d_{seq}$ | 0.00234 | 0.01514 | 0.01889 | 0.01428 | 0.02818 |
| Perfect alignments | 5685.00000 | 5339.00000 | 4779.00000 | 2846.00000 | 2979.00000 |
| Best alignments | 5221.00000 | 4844.00000 | 4881.00000 | 3828.00000 | 2677.00000 |
| Imperfect alignments | 1090.00000 | 1436.00000 | 1996.00000 | 3929.00000 | 3796.00000 |
| F1-score pos selection | 0.98053 | 0.84153 | 0.89902 | 0.80370 | 0.70905 |
| F1-score neg selection | 0.99800 | 0.98339 | 0.98966 | 0.98196 | 0.96990 |

* PRANK produced 40 empty alignments, calculations are based on 7677 alignments.

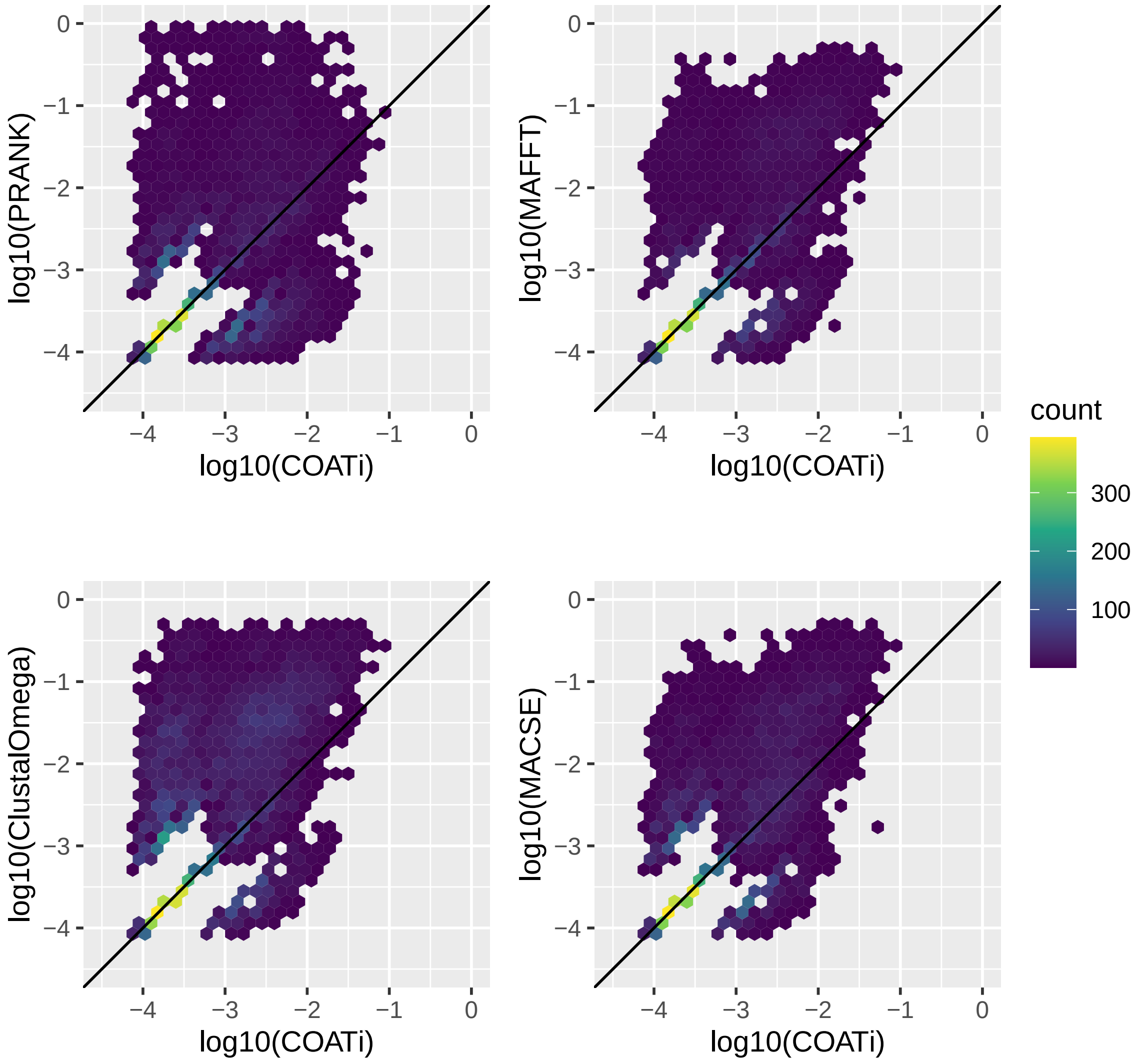


Figure 1: Comparison of log10-transformed $d_{seq}$ data with pseudocounts between COATi codon-triplet-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 2.06e-79$.

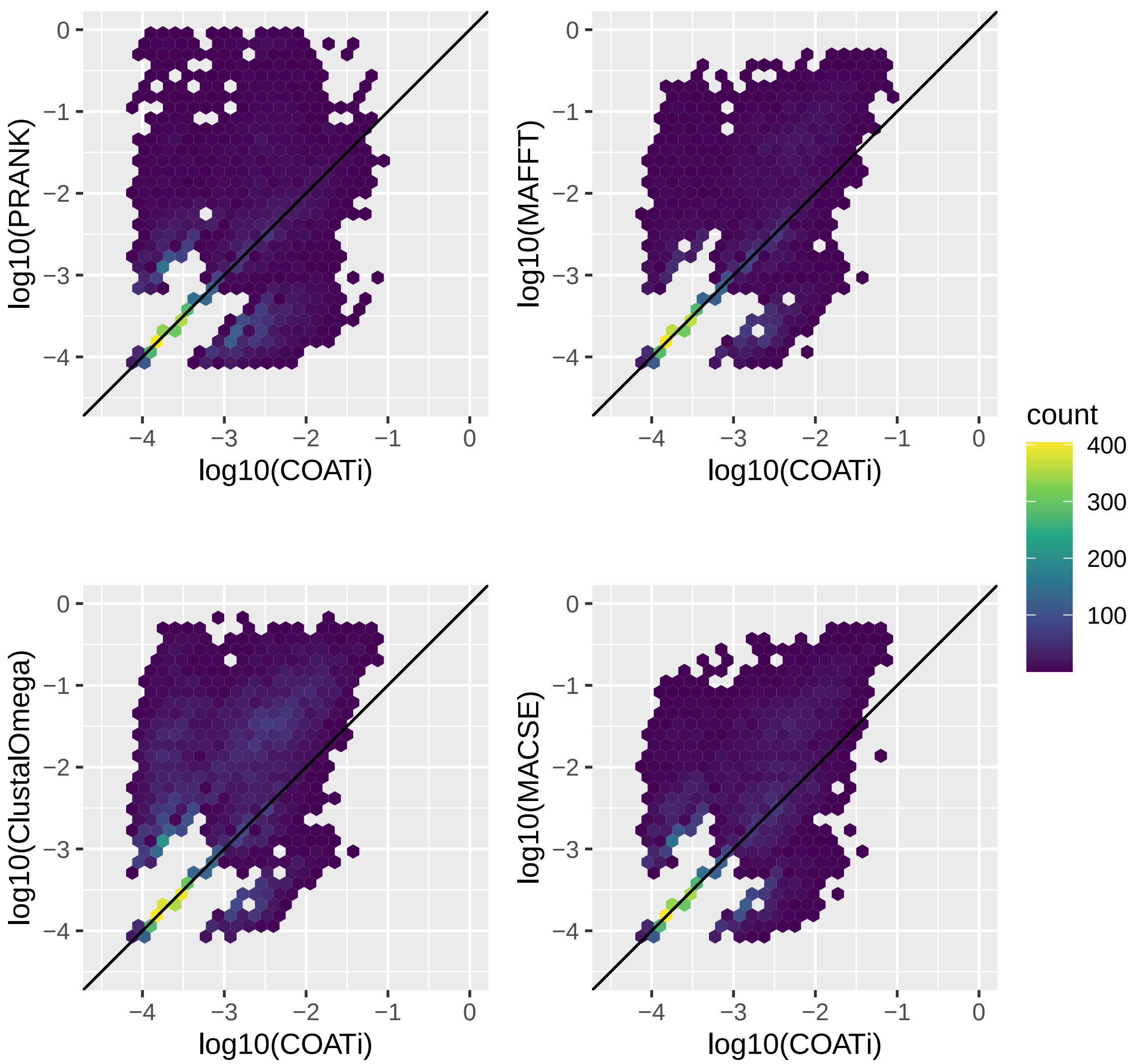


Figure 2: Comparison of log10-transformed $d_{seq}$ data with pseudocounts between COATi codon-triplet-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 8.15e-53$.

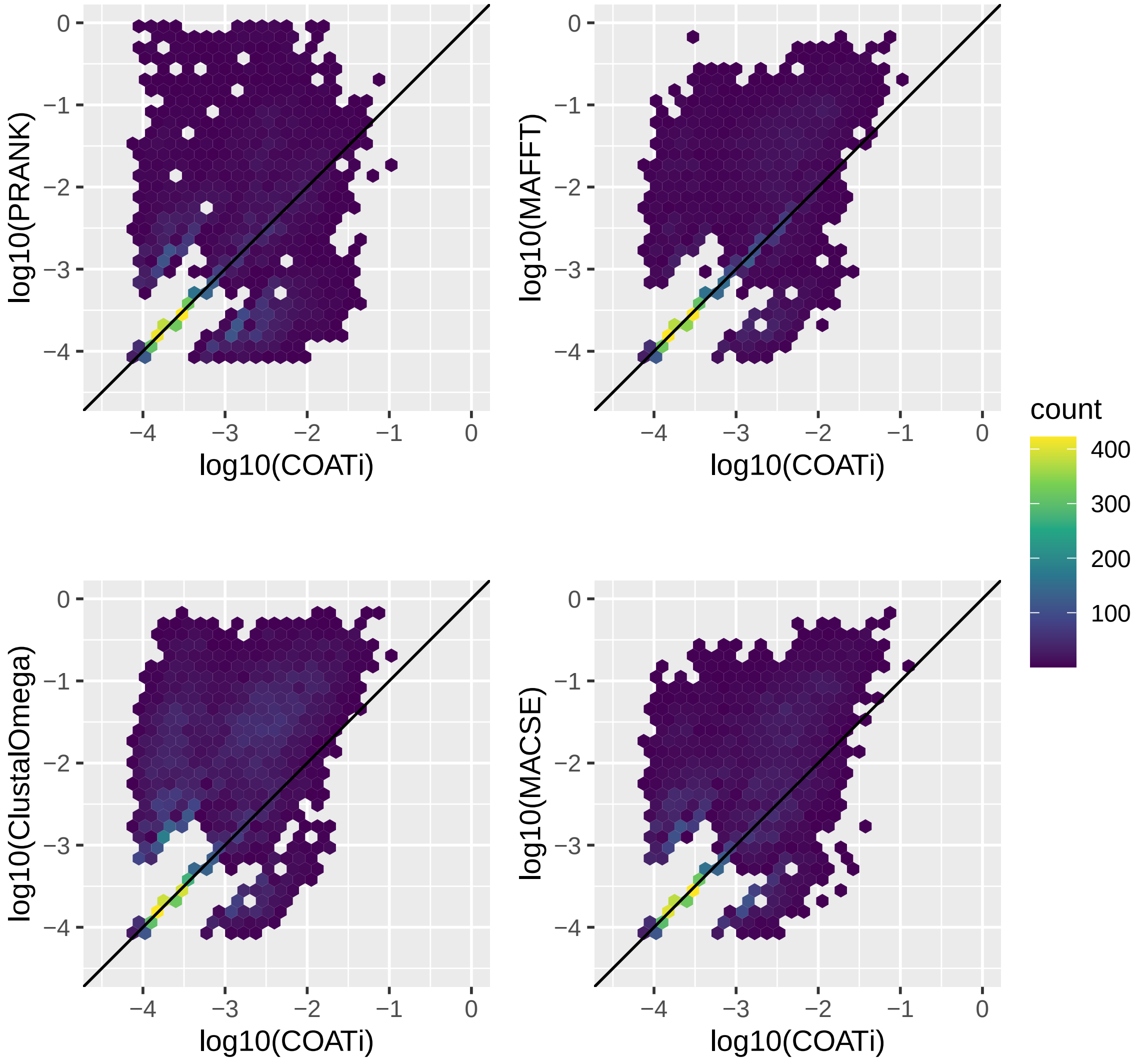


Figure 3: Comparison of log10-transformed $d_{seq}$ data with pseudocounts between COATi codon-marginal-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 2.65e - 80$.

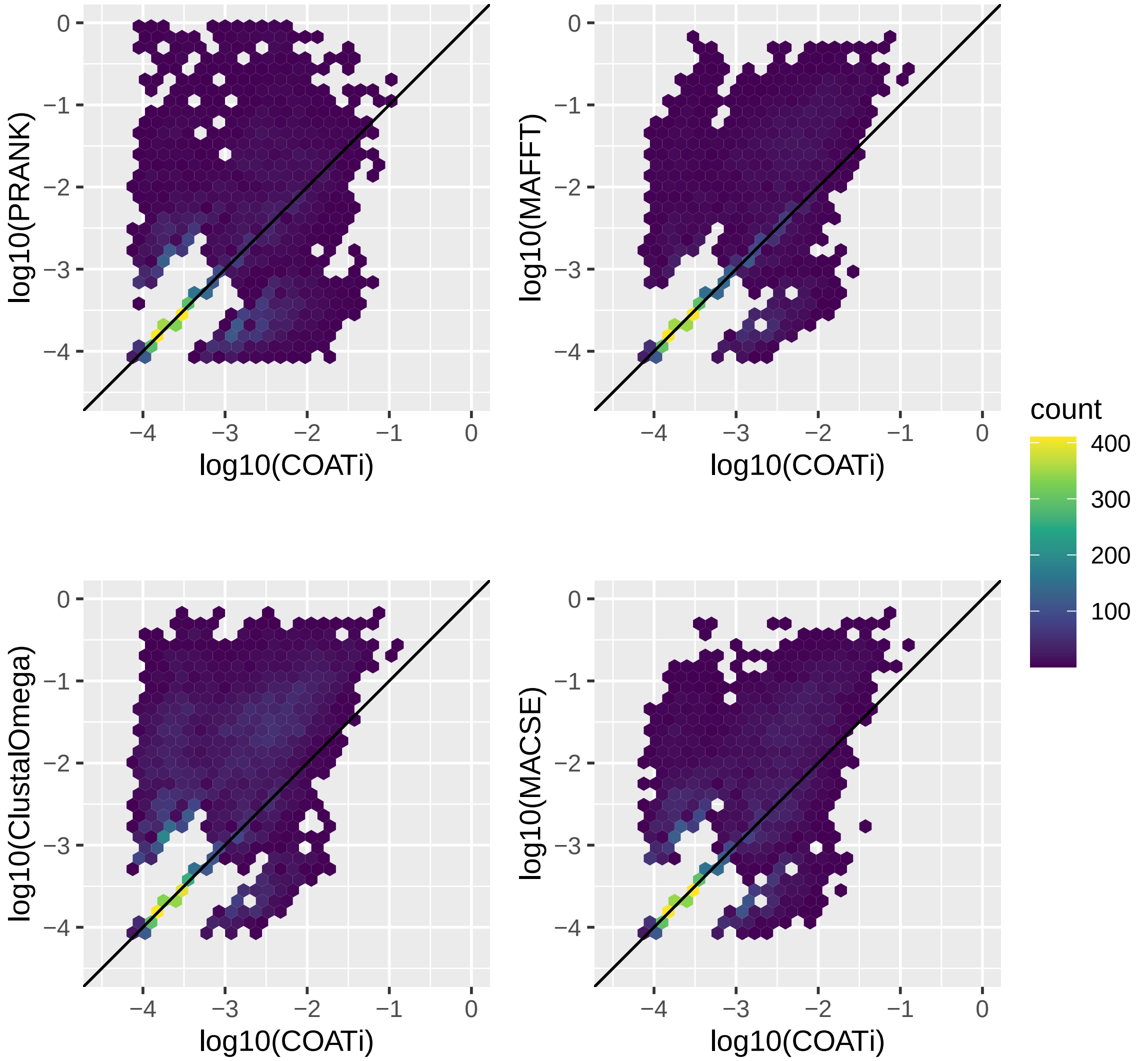
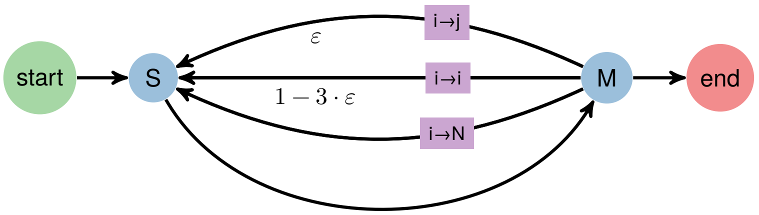


Figure 4: Comparison of log10-transformed $d_{seq}$ data with pseudocounts between COATi codon-marginal-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 3.37e - 59$.



| **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** | |
|---|---|
| $\varepsilon$: | 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.