COATi: statistical pairwise alignment of protein coding sequences

Supplementary Materials

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

Contents

1 Aligner Commands

We evaluated five different aligners. Below are the commands that we used to run them. We have abbreviated the commands for clarity, stripping out unimportant arguments. Complete workflows can be found in our coati-testing repository on Github.

- COATi: `coati alignpair -m tri-mg ...`
- ClustalΩ v1.2.4: `clustalo ...`
- PRANK v.150803: `prank -codon ...`
- MACSE v2.06: `java -jar macse.jar -fs_lr 10 -stop_lr 10 -prog alignSequences ...`
- MAFFT v7.505: `mafft --preservecase --globalpair --maxiterate 1000 ...`

COATi can use different alignment models for pairwise alignment. Below are the commands that we used to run different models.

- tri-mg: `coati alignpair -m tri-mg ...`
- tri-ecm: `coati alignpair -m tri-ecm ...`
- mar-mg: `coati alignpair -m mar-mg ...`
- mar-ecm: `coati alignpair -m mar-ecm ...`

2 FST Alignment Example

When using the triplet models (tri-mg and tri-ecm), COATi uses the OpenFST library to generate best alignments by composing the input and output sequences with the COATi FST model. While the COATi FST model is too large to display, we can show the result of a composition.

Fig. S1 shows a graph depicting the FST that results from composing the COATi FST with the input sequence "CTC" and the output sequence "CTG". Every path through this FST represents one possible way to align "CTC" and "CTG", and the sum of all weights along a path is the total weight of the respective alignment. Here, the weights of each arc are in negative-log space. Note that this graph has been optimized, and weight has been pushed towards the initial state. The weight of any specific arc may not be directly mapable to a weight described in the model.

Fig. S2 is the best alignment between "CTC" and "CTG", as determined by the shortest path algorithm. Figs. S1 and S2 were produced by the OpenFST library. A bold circle represents a starting node, and a double circle represents a termination node.

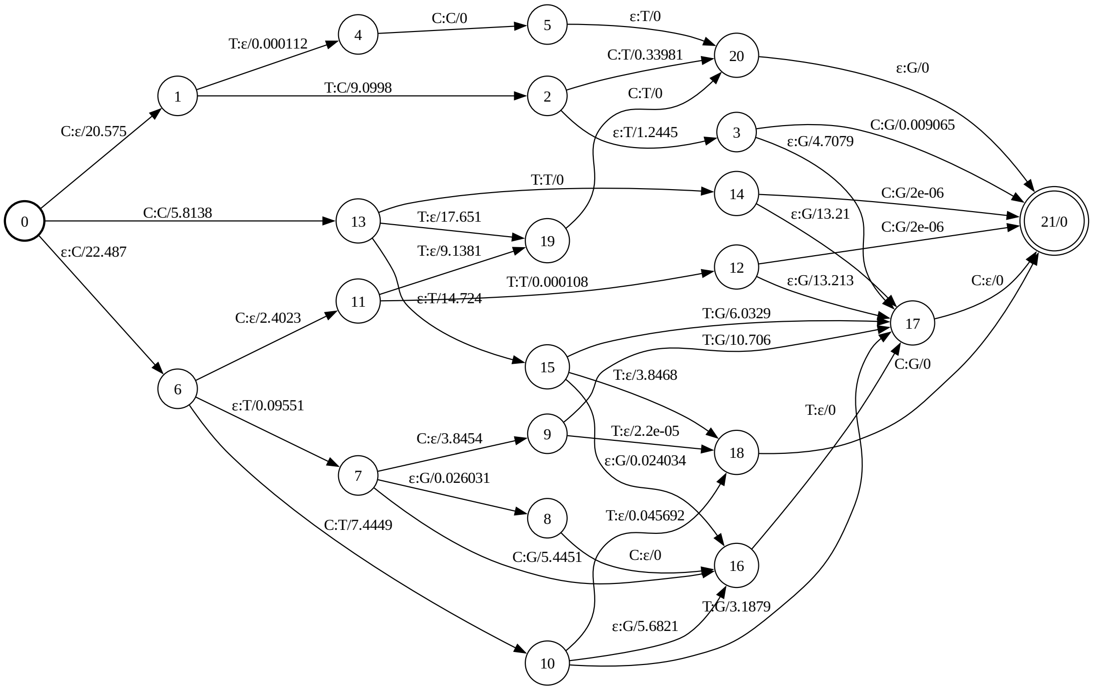

Figure S1: The FST of all possible alignments between ‘CTC’ and ‘CTG’.

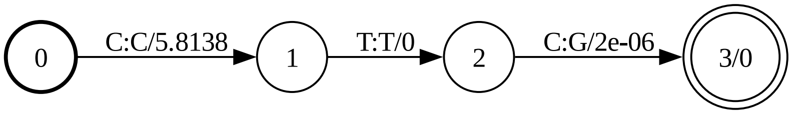

Figure S2: The best alignment of ‘CTC’ and ‘CTG’.

Table 1: Accuracy of COATi codon-triplet-mg, PRANK, MAFFT, ClustalOmega, and MACSE on 7761 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Triplet-MG94	0.00221	5793	5139	1048	0.98073	0.99809
MAFFT	0.01471	5292	4692	1549	0.84314	0.98411
PRANK*	0.01828	4725	4774	2116	0.86749	0.98698
MACSE	0.01399	2861	3737	3980	0.79456	0.98199
ClustalOmega	0.02929	2893	2615	3948	0.68691	0.96938

* PRANK produced 42 empty alignments, calculations are based on 7719 alignments.

Table 2: Accuracy of COATi codon-triplet-ecm, PRANK, MAFFT, ClustalOmega, and MACSE on 7761 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Triplet-ECM	0.00238	5689	5045	1118	0.97803	0.99779
MAFFT	0.01451	5338	4677	1469	0.86048	0.98549
PRANK*	0.01903	4803	4851	2004	0.89250	0.98912
MACSE	0.01352	2903	3787	3904	0.82181	0.98359
ClustalOmega	0.02801	2979	2624	3828	0.72337	0.97244

* PRANK produced 69 empty alignments, calculations are based on 7692 alignments.

Table 3: Accuracy of COATi codon-marginal-mg, PRANK, MAFFT, ClustalOmega, and MACSE on 7755 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Marginal-MG94	0.00222	5808	5220	1075	0.97671	0.99766
MAFFT	0.01505	5301	4782	1582	0.85147	0.98455
PRANK*	0.01974	4856	5015	2027	0.89928	0.99000
MACSE	0.01429	2855	3893	4028	0.81569	0.98349
ClustalOmega	0.02870	2901	2610	3982	0.72399	0.97171

* PRANK produced 60 empty alignments, calculations are based on 7695 alignments.

Table 4: Accuracy of COATi codon-marginal-ecm, PRANK, MAFFT, ClustalOmega, and MACSE on 7767 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Marginal-ECM	0.00229	5781	5135	1081	0.97052	0.99710
MAFFT	0.01473	5379	4813	1483	0.85011	0.98491
PRANK*	0.01953	4830	4918	2032	0.87752	0.98790
MACSE	0.01400	2953	3893	3909	0.78977	0.98159
ClustalOmega	0.02918	2892	2611	3970	0.67847	0.96785

* PRANK produced 49 empty alignments, calculations are based on 7718 alignments.

Table 5: Accuracy of COATi codon-triplet-mg, PRANK, MAFFT, ClustalOmega, and MACSE on 7798 simulated sequence pairs with gorilla as the reference. 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Triplet-MG94	0.00217	5870	5162	1030	0.98450	0.99853
MAFFT	0.01445	5450	4803	1450	0.84704	0.98508
PRANK*	0.02126	4942	5026	1958	0.89805	0.99042
MACSE	0.01340	2966	3989	3934	0.79608	0.98260
ClustalOmega	0.02860	3034	2711	3866	0.69147	0.97026

* PRANK produced 35 empty alignments, calculations are based on 7763 alignments.

Table 6: Accuracy of COATi codon-triplet-mg, PRANK, MAFFT, ClustalOmega, MACSE, and codon-triplet-mg with gorila as the reference on 4003 of the 7761 simulated sequence pairs where the gorilla sequence was simulated without early stop codons, incomplete codons, or ambiguous nucleotides. 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.

	dseq	Perfect alns	Best alns	Imperfect alns	F1 pos selection	F1 neg selection
Triplet-MG	0.00113	3501	2890	309	0.99278	0.99932
MAFFT	0.00586	3162	2704	648	0.91064	0.99137
PRANK	0.00358	2829	2673	981	0.90332	0.99084
MACSE	0.00448	2552	2434	1258	0.87234	0.98857
ClustalOmega	0.02099	1772	1554	2038	0.75960	0.97686
Triplet-MG-gor-ref	0.00118	3463	2816	347	0.98993	0.99904

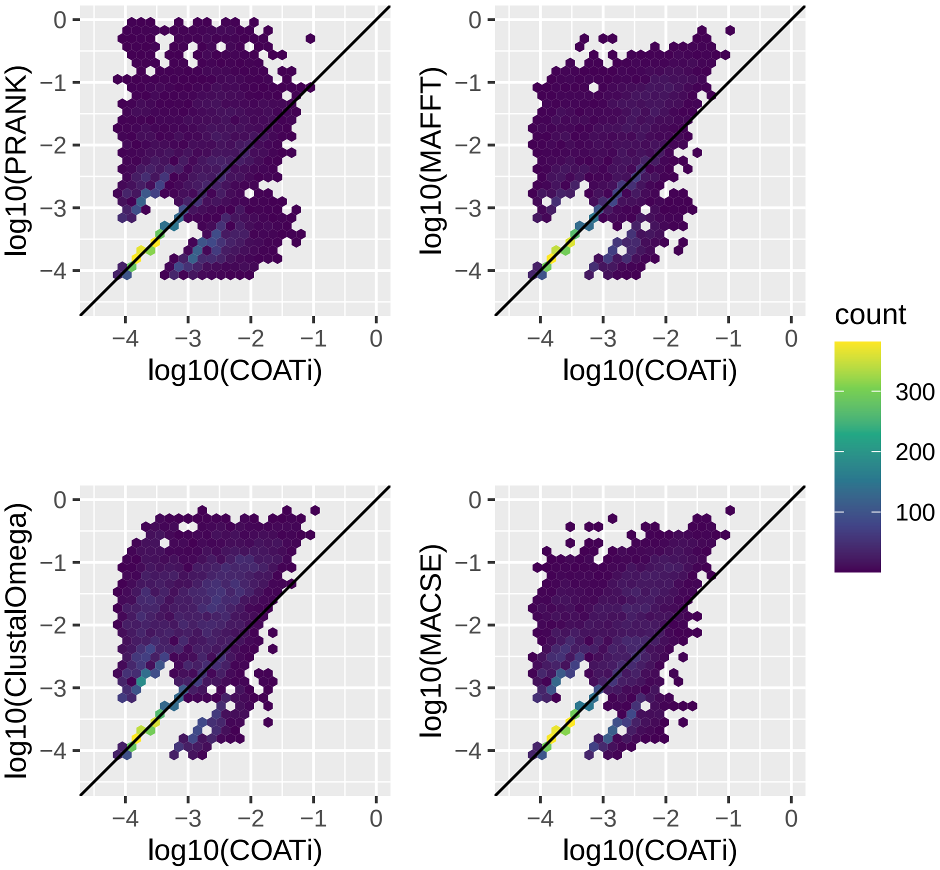

Figure S3: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-triplet-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 1.25e-76$.

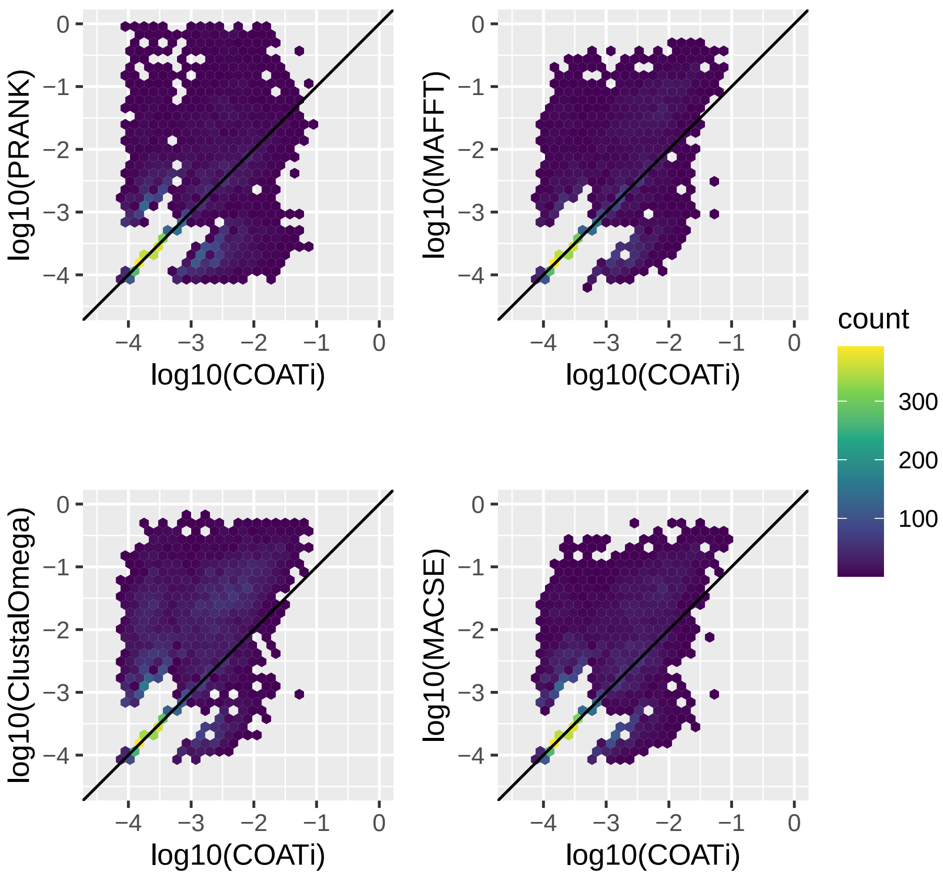

Figure S4: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-triplet-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 3.23e - 48$.

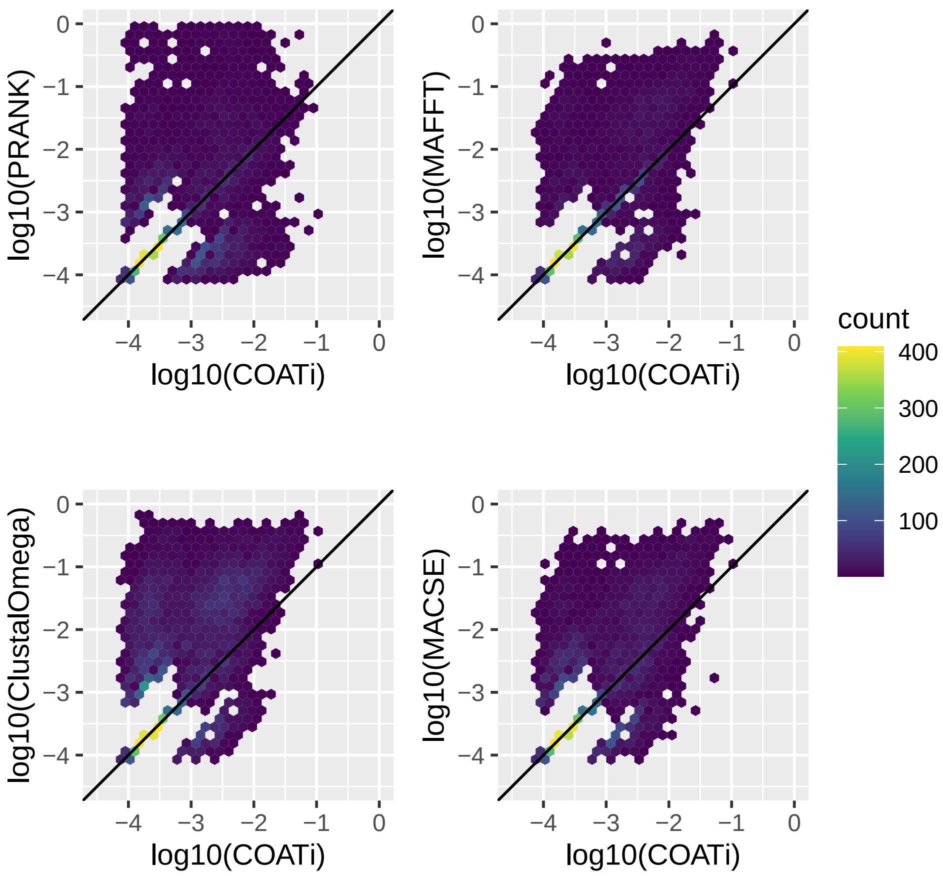

Figure S5: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-marginal-mg and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 1.99e-53$.

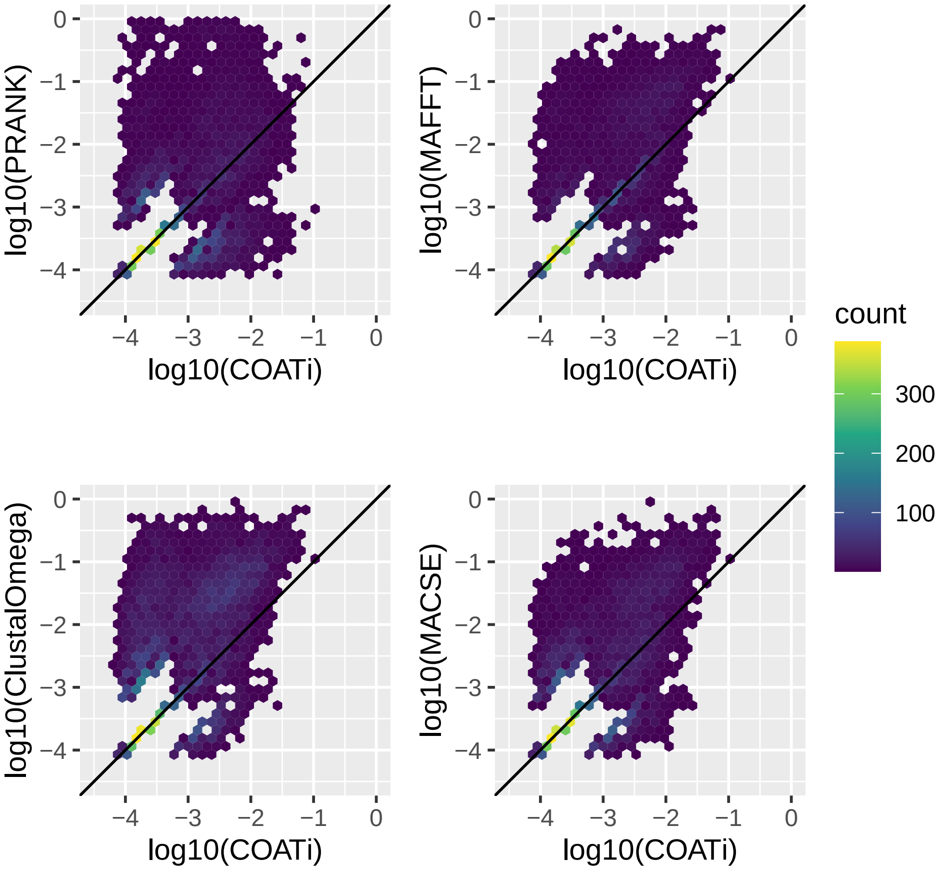

Figure S6: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-marginal-ecm and PRANK, MAFFT, ClustalOmega, and MACSE. COATi was significantly more accurate than other aligners; all p-values were $\leq 1.44e - 52$.

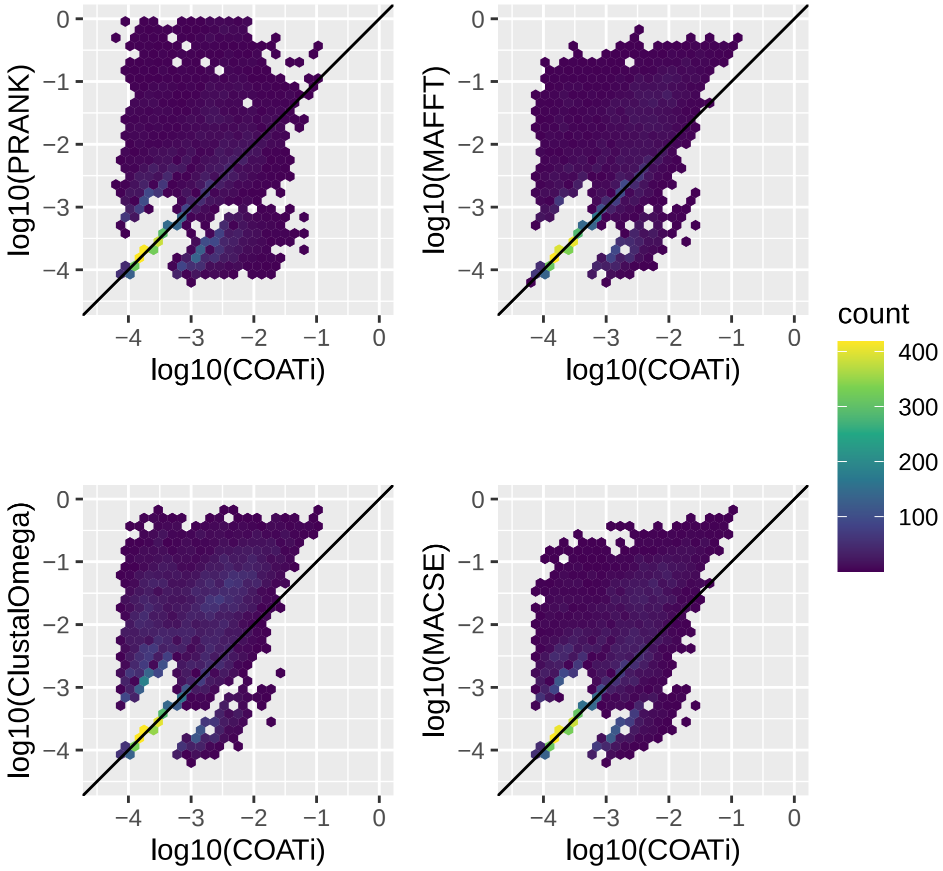

Figure S7: Comparison of log10-transformed d_{seq} data with pseudocounts between COATi codon-triplet-mg and PRANK, MAFFT, ClustalOmega, and MACSE with gorilla as the reference. COATi was significantly more accurate than other aligners; all p-values were $\leq 1.75e - 64$.

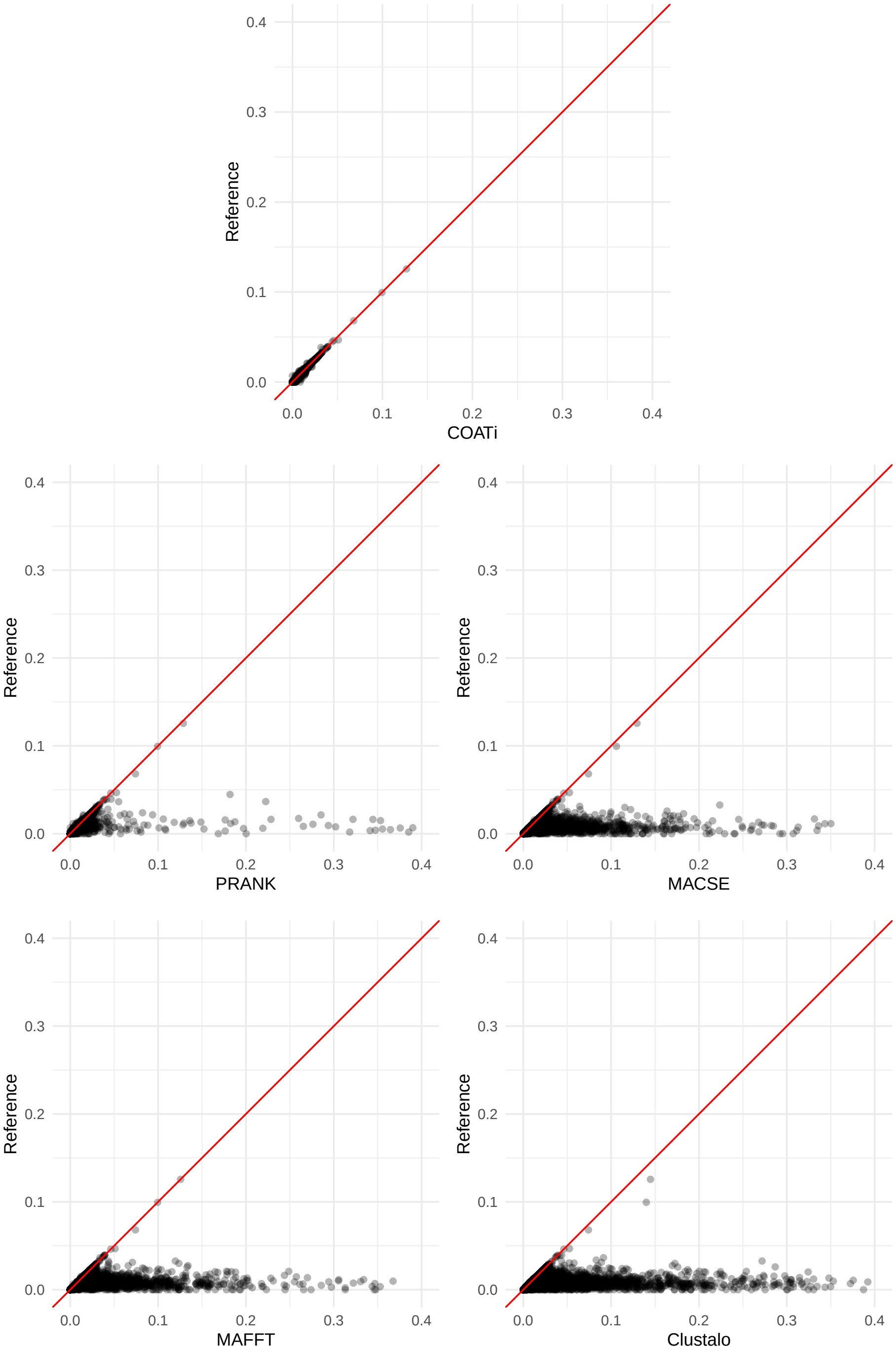

Figure S8: Alignments generated by COATi produce accurate evolutionary distances. Benchmark alignments were realigned by COATi and four other methods as described in the main text. Kimura two-parameter (K2P) distances were estimated from the benchmark alignments and the estimated alignments. Each panel is a scatter plot comparing the benchmark disances (Reference) with distances estimated via one of the aligners. COATi clearly performs better than other tools, which tend to drastically overestimate some distances.