Supplementary Material: Use of ChIP-Seq data for the design of a multiple promoter alignment method

Ionas Erb[1], Juan González-Vallinas[2], Giovanni Bussotti[1], Enrique Blanco[3], Eduardo Eyras[2,4] and Cédric Notredame[1]

[1]Centre for Genomic Regulation (CRG). Dr. Aiguader 88, E08003 Barcelona, Spain

[2]Universitat Pompeu Fabra. Dr. Aiguader 88, E08003, Barcelona, Spain

[3]Departament de Genètica / Institut de Biomedicina (IBUB), Universitat de Barcelona (UB). Av. Diagonal 643 08028 Barcelona (Spain).

[4]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, E08010, Barcelona, Spain

	CEBPA	HNF4a
Human	65 756 959	47 288 137
Mouse	78 073 807	44 956 816
Dog	61 395 140	47 339 609
Chicken	32 465 702	-

Table S1. Total reads mapped (sums over various individuals of one species).

	CEBPA	HNF4a
Human	75 791	39 300
Mouse	29 051	19 284
Dog	44 223	38 764
Chicken	18 799	-

Table S2. Total peaks called ($p<=10^{-6}$).

	CEBPA	HNF4a
Human	1399 (1.8%)	1494 (3.8%)
Mouse	523 (1.8%)	435 (2.3%)
Dog	560 (1.3%)	861 (2.2%)
Chicken	423 (2.3%)	-

Table S3. Peaks falling in orthologous promoter cliques.

	CEBPA	HNF4a
Chicken-Dog	46	-
Chicken-Human	101	-
Chicken-Mouse	33	-
Dog-Human	165	282
Dog-Mouse	83	135
Human-Mouse	184	159

Table S4. Promoter cliques that contain peaks from both species of a given species pair.

	CEBPA	HNF4a
Chicken-Dog	10	-
Chicken-Human	12	-
Chicken-Mouse	10	-
Dog-Human	64	88
Dog-Mouse	37	60
Human-Mouse	80	75

Table S5. Pairs of overlapping ChIP-Seq regions falling in orthologous promoter cliques (Pro-Coffee Alignments)

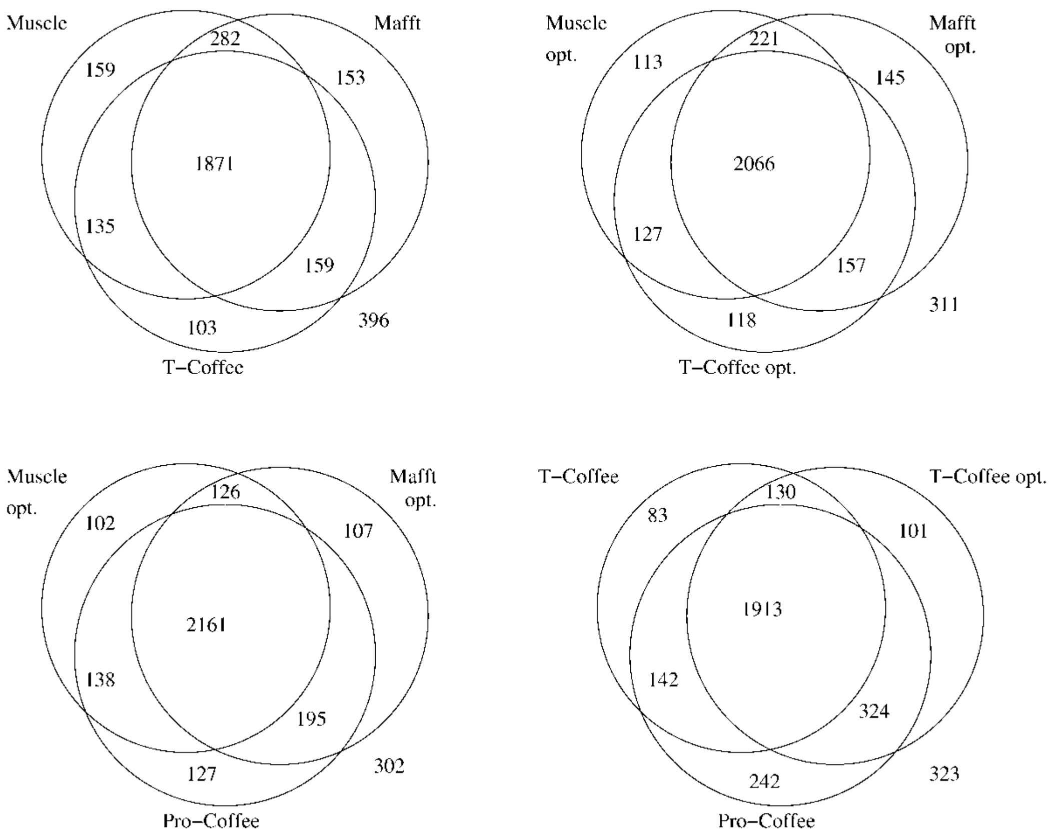

Figure S1. Venn diagrams for the homology benchmark. Numbers stand for correctly predicted orthologs from the full set of 3258 cliques. Top row: Optimization increases the intersection of methods, which goes up by 6% to 63.4% of all orthologs. Bottom left: Replacing the optimized T-Coffee by Pro-Coffee increases this figure by another 3%. In this case the percentage of orthologs no method is able to predict correctly is still 9.3%. Bottom right: Optimization and method improvement does not lead to a mere growth of the correct set but also causes a drift that leaves previously identified orthologs undiscovered. This is however compensated by larger sets of newly identified orthologs.

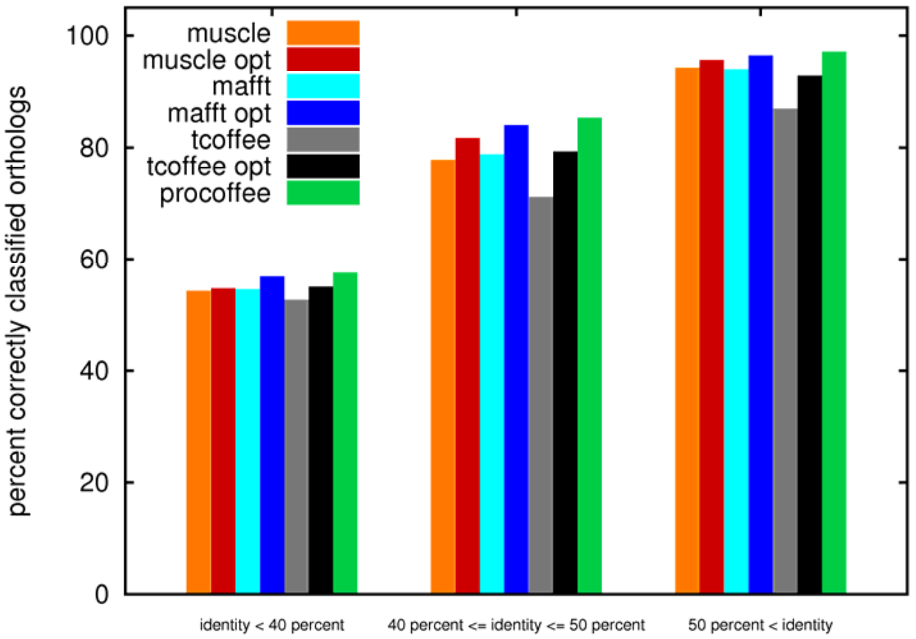

Figure S2. Homology benchmark on three subsets of orthologs whose alignments fall in a certain range of percent identity. Identity of orthologs was determined from averaging the quantity over alignments of the three best methods (optimized versions of mafft and muscle and procoffee). Then the ortholog classification was done on each subset separately, where the low identity set comprises 897 orthologs with average identity below 40%, the middle identity set 1659 orthologs with identities between 40% and 50%, and the high identity set 720 ortholgs with identities beyond 50%. Optimized methods perform consistently better on all sets and there is little effect of identity on performance ranking of the methods.

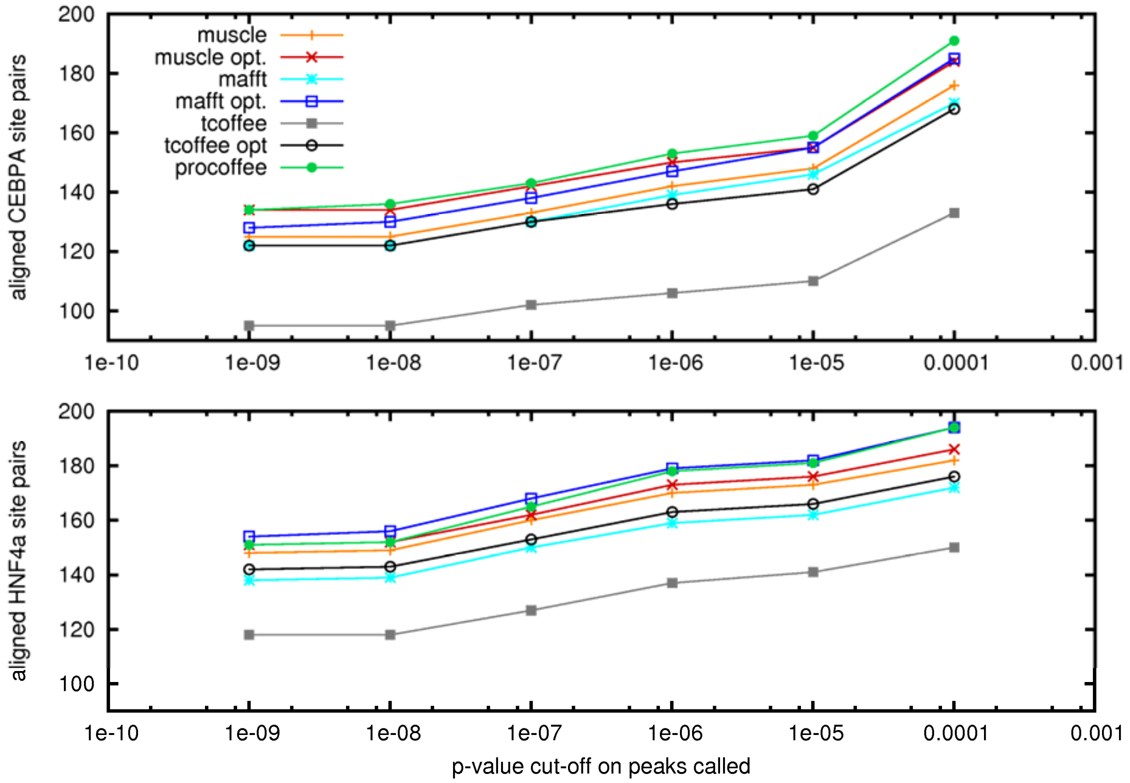

Figure S3. Total number of aligned binding sites depending on peak quality cut-off (use: p<=10^{-6}). Cut-off changes within this range of significance basically do not affect method rankings.

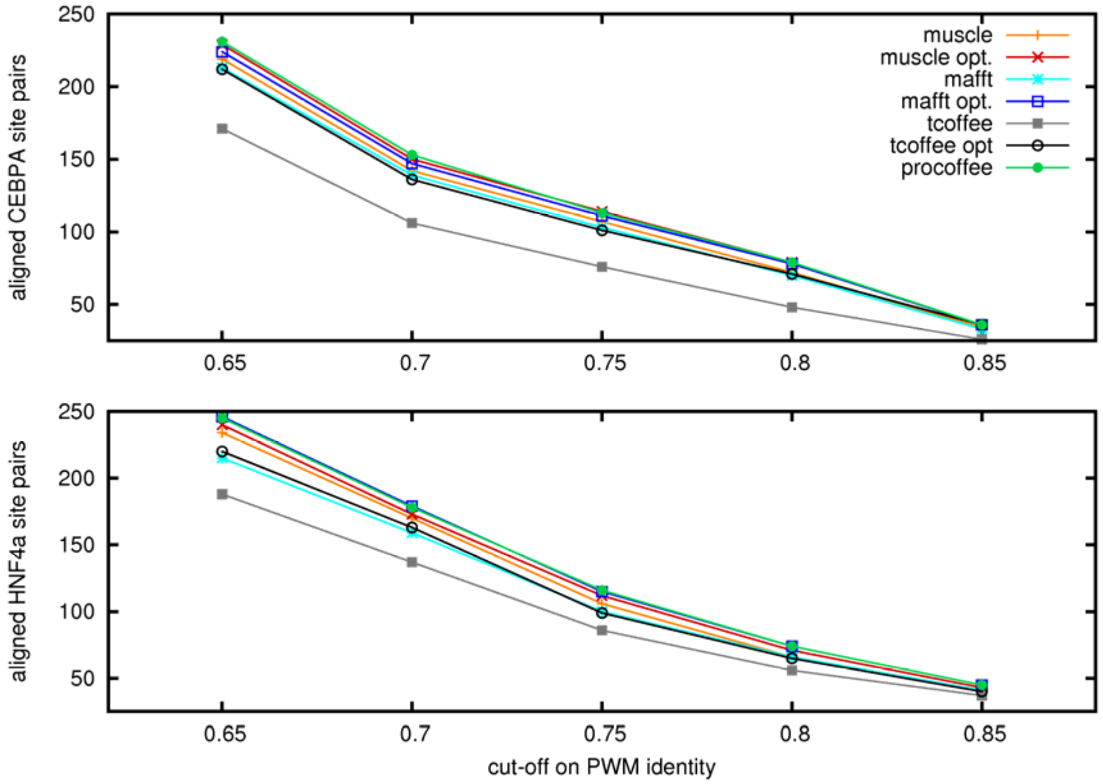

Figure S4. Total number of aligned binding sites depending on site quality (TFBS-PWM identity) cut-off (use: 70% identity). Note that we are only considering sites that fall in factor binding regions. Again, there is little effect on ranking of methods.

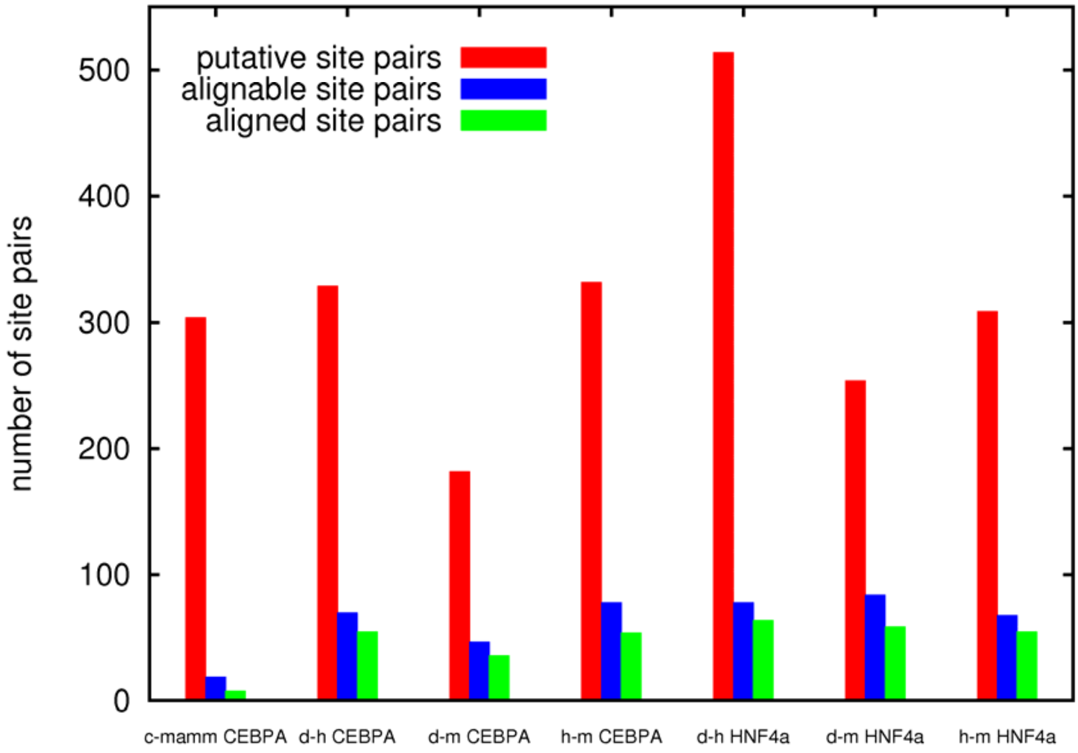

Figure S5. Aligned TFBS pairs compared to TFBS pairs contained in orthologous promoter cliques ("putative site pairs") and in overlapping ChIP-Seq regions ("alignable site pairs") using Pro-Coffee alignments. See Materials and methods for definitions.

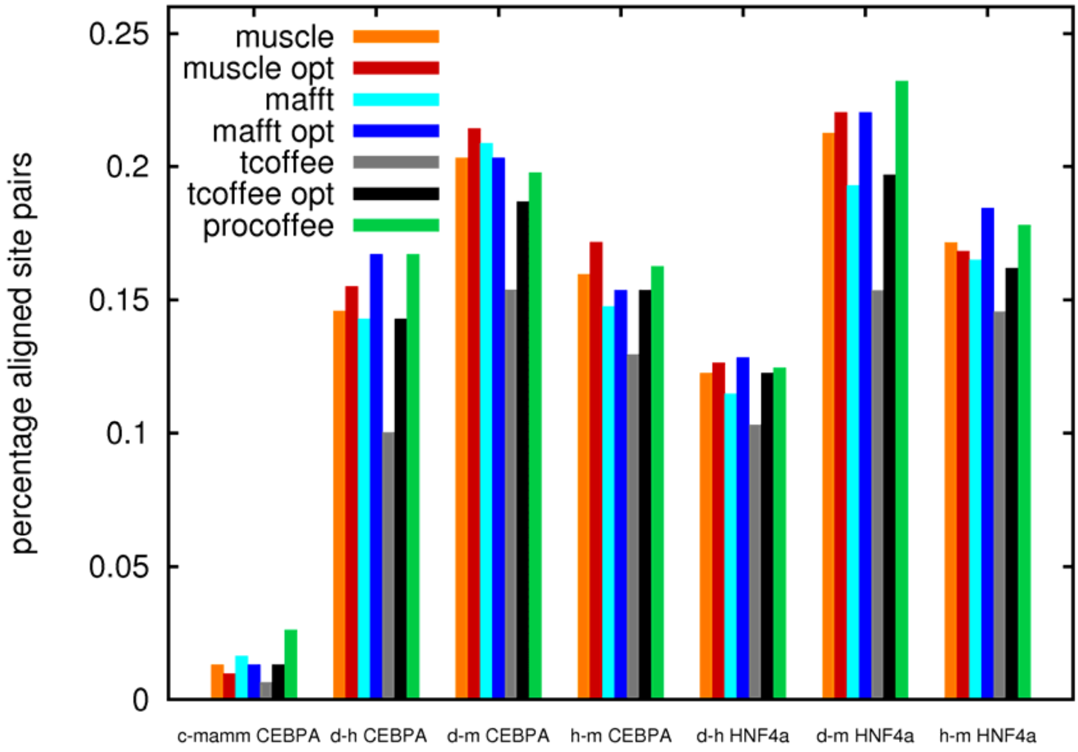

Figure S6. Aligned site pairs as a percentage of putative site pairs. For the example Pro-Coffee, these are the numbers represented by the green bars in Supplementary Figure S5 divided by the numbers represented by the corresponding red bars. Methods usually improve on these data sets when trained on the ortholog benchmark: Muscle and Mafft improve on 5, T-Coffee on all 7 data sets.

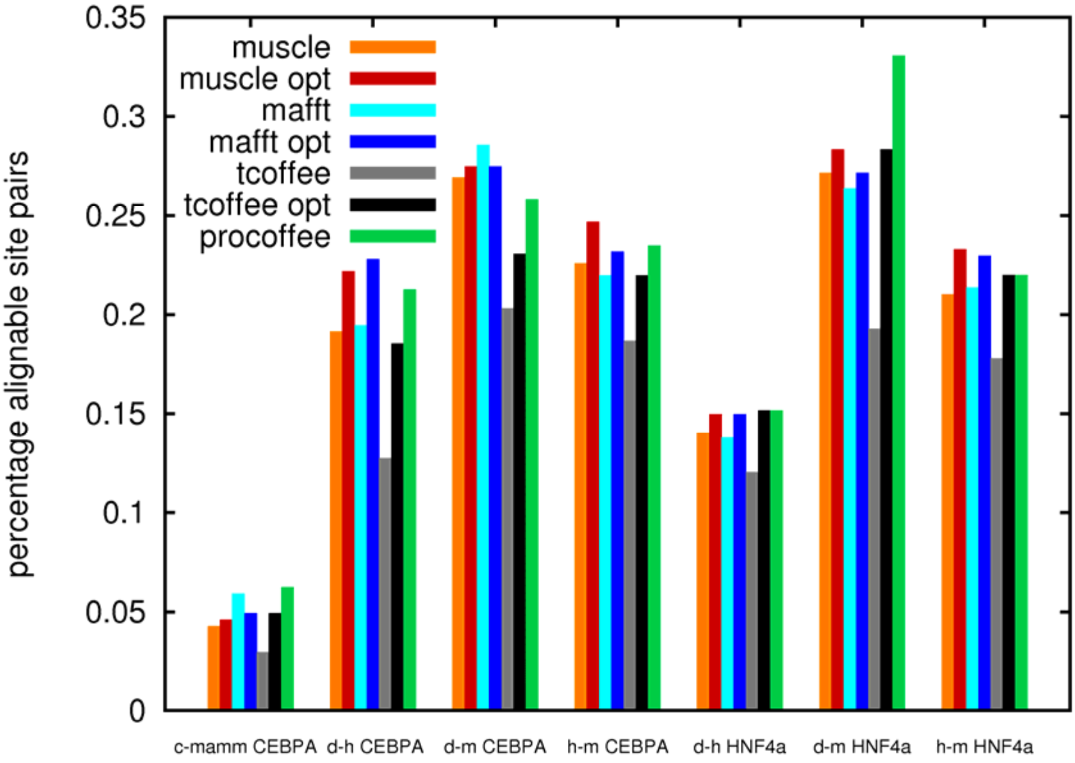

Figure S7. Percentage of alignable site pairs. For the example Pro-Coffee, these are the numbers represented by the blue bars in Supplementary Figure S5 divided by the numbers represented by the corresponding red bars. The increase in alignable site pairs with method tuning corresponds to an improvement of large-scale properties of alignments (more overlapping factor-bindig regions). It is this global property that mostly profits from tuning on the ortholog benchmark: Mafft improves on 5 data sets, Muscle and T-Coffee on all 7 data sets.

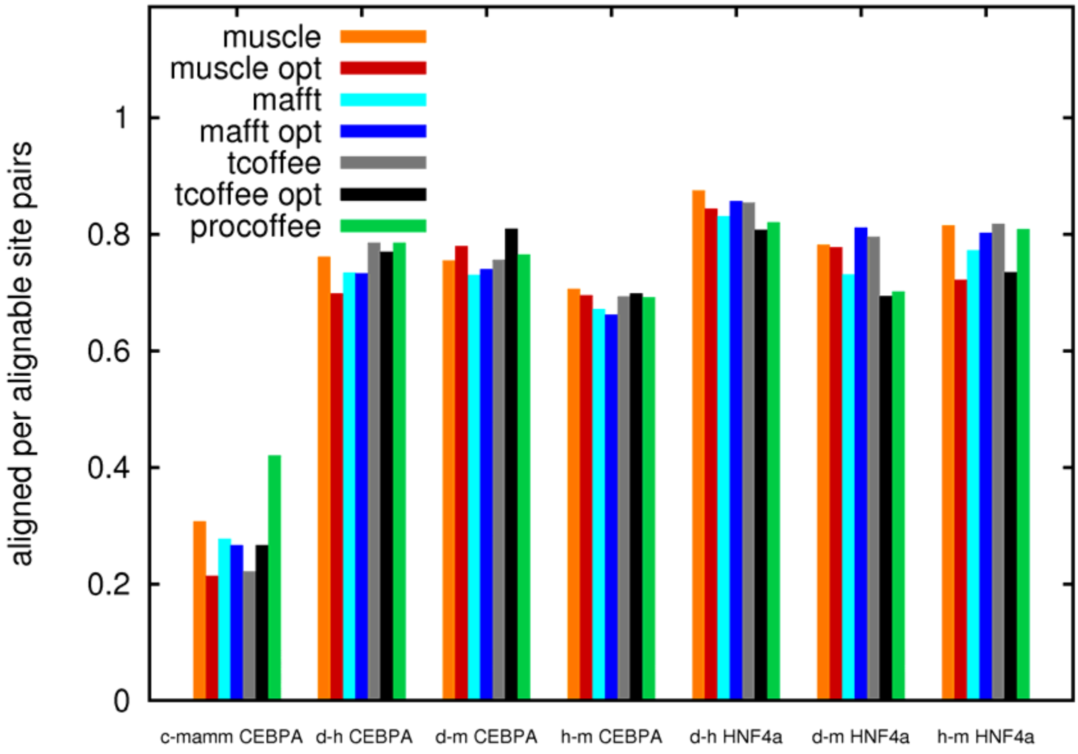

Figure S8. Ratio of aligned site pairs over alignable pairs. For the example Pro-Coffee, these are the numbers represented by the green bars in Supplementary Figure S5 divided by the numbers represented by the corresponding blue bars. An increase in aligned per alignable sites corresponds to an improvement of fine-grained alignment properties. Our method tuning on the ortholog benchmark usually does not improve this property, it can often even change to the worse: Mafft still improves on 4 data sets, Muscle on only 1, and T-Coffee on 3.