Supplementary file 2. Villanueva-Cañas, J.L., Laurie, S., Albà, M.M. (2013). Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biology and Evolution*.

FIGURE 1

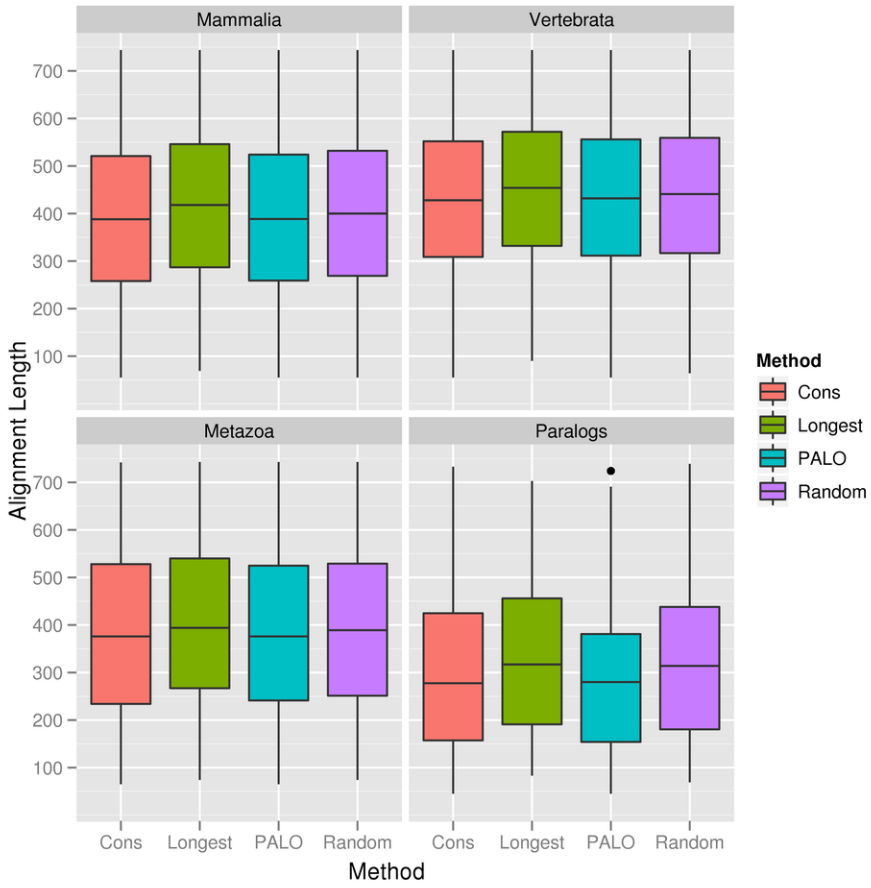

Figure 1. Alignment length using different methods. The area within the box contains 50% of the data; the horizontal line is the average. Except for the Metazoa dataset, differences between Longest and PALO were significant at p<10-4 using a Mann-Whitney test. Alignments were computed using MAFFT.

FIGURE 2

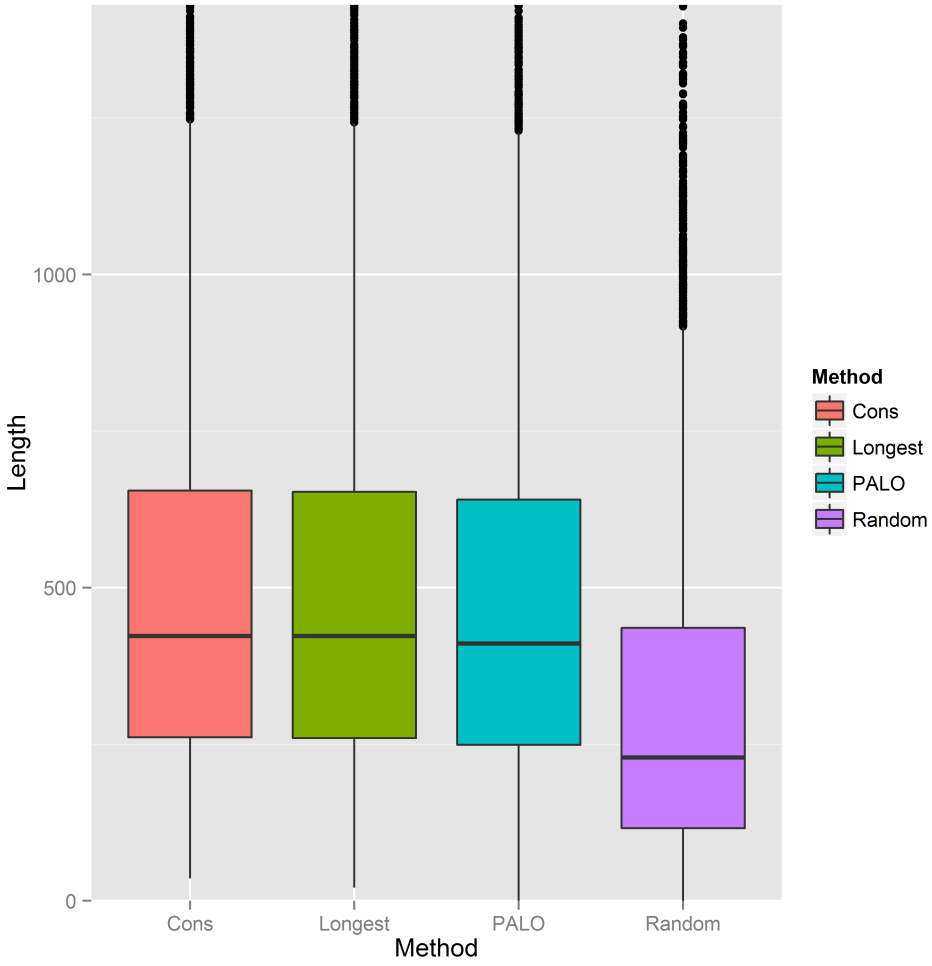

Figure 2. Alignment length in the Mammalia dataset excluding columns with indels. The area within the box contains 50% of the data; the horizontal line is the average. Differences with Cons are only significant for Random ($p<0.05$, Mann-Whitney test).

TABLE 1

Combinations	Dataset	Method	% Hit Cons	Conservation score Mean	Median	SD
2 to 5	Mammalia	PALO	80.00	65.46	68	21.08
		Longest**	16.89	59.15	60	19.10
		Random**	37.70	56.71	58	23.36
		Cons	-	66.95	68	19.47
	Vertebrata	PALO	74.05	48.84	47	18.39
		Longest**	20.76	45.21	43	16.99
		Random**	40.83	44.07	43	19.47
		Cons	-	50.12	48	17.49
	Metazoa	PALO	70.37	22.56	20	10.87
		Longest	24.07	20.56	19	10.43
		Random	38.89	20.17	18	11.26
		Cons	-	23.59	21	10.83
	Paralogs	PALO	59.09	47.04	42	27.83
		Longest**	20.78	41.23	36	22.43
		Random**	14.94	31.55	23	27.29
		Cons	-	51.89	48	24.81
6 to 12	Mammalia	PALO	73.58	65.39	69	21.24
		Longest**	18.32	60.29	63	19.19
		Random**	22.08	46.69	48	26.43
		Cons	-	67.35	70	19.04
	Vertebrata	PALO	64.96	48.40	48	18.81
		Longest**	19.67	45.67	44	17.03
		Random**	18.85	36.04	34	20.66
		Cons	-	50.20	49	17.38
	Metazoa	PALO	63.21	25.27	22	13.49
		Longest*	28.30	23.21	22	11.55
		Random**	26.42	18.45	16	12.83
		Cons	-	26.67	24	12.76
	Paralogs	PALO	59.09	47.04	42	27.83
		Longest**	20.78	41.23	36	22.43
		Random**	14.94	31.55	23	27.29
		Cons	-	51.89	48	24.81

(see next page)

Combinations	Dataset	Method	% Hit Cons	Conservation score Mean	Median	SD
13 to 30		PALO	73.88	69.60	73	19.67
	Mammalia	Longest**	14.07	65.09	69	18.24
		Random**	11.94	41.06	39	28.31
		Cons	-	71.49	75	17.65
		PALO	58.19	50.01	49	18.27
	Vertebrata	Longest**	18.07	47.44	47	16.71
		Random**	13.24	33.86	32	21.60
		Cons	-	51.81	51	17.14
		PALO	59.22	26.70	25	14.54
	Metazoa	Longest	24.27	25.39	24	12.03
		Random**	17.48	19.18	16	14.03
		Cons	-	28.36	26	13.89
		PALO	59.09	47.04	42	27.83
	Paralogs	Longest**	20.78	41.23	36	22.43
		Random**	14.94	31.55	23	27.29
		Cons	-	51.89	48	24.81
31 to 100		PALO	68.48	70.48	75	20.90
	Mammalia	Longest**	15.06	65.82	69	18.99
		Random**	6.26	35.81	29	28.59
		Cons	-	72.53	77	18.62
		PALO	50.35	49.70	49	17.92
	Vertebrata	Longest**	16.16	47.27	46	16.11
		Random**	6.09	28.40	27	21.68
		Cons	-	51.91	50	16.65
		PALO	63.08	28.45	24	16.62
	Metazoa	Longest	20.00	26.28	24	15.15
		Random**	18.46	17.09	14	16.84
		Cons	-	29.28	25	16.28
		PALO	59.09	47.04	42	27.83
	Paralogs	Longest**	20.78	41.23	36	22.43
		Random**	14.94	31.55	23	27.29
		Cons	-	51.89	48	24.81

(see next page)

				Conservation score		
Combinations	**Dataset**	**Method**	**% Hits Cons**	**Mean**	**Median**	**SD**
> 100		PALO	55.41	72.05	76	19.98
	Mammalia	Longest**	18.47	67.91	72	18.30
		Random**	2.37	33.00	25	29.36
		Cons	-	75.32	79	16.56
		PALO	39.10	51.78	53	16.45
	Vertebrata	Longest*	8.97	49.40	50	15.06
		Random**	1.28	23.95	16	21.46
		Cons	-	54.91	55	15.35
		PALO	65.38	32.88	32	16.41
	Metazoa	Longest	19.23	29.69	26	15.44
		Random**	3.85	14.23	12	11.69
		Cons	-	33.73	32	15.98
		PALO	59.09	47.04	42	27.83
	Paralogs	Longest**	20.78	41.23	36	22.43
		Random**	14.94	31.55	23	27.29
		Cons	-	51.89	48	24.81

Table 1. Sequence conservation statistics for the different methods by number of protein isoform combinations. Data is for gene families in which PALO selected a different combination than Longest. % Hit Cons: percentage of cases in which the different methods results in the same protein isoform combination as Cons. For MAFFT alignment Conservation scores the Mean, Median and standard deviation (SD) were calculated. Differences in the Alignment Score distribution between each method and Cons were evaluated by a Mann-Whitney test, * p-value < 0.05, ** p-value < 10^{-3}.

TABLE 2

DATASET	Method	%Hit Cons	Mean	Median	SD
Mammalia	PALO*	88.5	69.58	73	20.17
	Longest**	72.35	68.15	72	19.85
	Random**	33.28	45.15	46	29.14
	Cons	-	69.58	73	20.17
Vertebrata	PALO*	74.4	46.7	47	19.78
	Longest**	60.6	45.86	46	19.1
	Random**	30.01	32.14	30	22.28
	Cons	-	47.64	48	19.32
Metazoa	PALO	82.82	22.19	18	15.03
	Longest**	74.19	21.79	18	14.56
	Random**	36.35	15.24	12	14.03
	Cons	-	22.65	19	14.94
Paralogs	PALO	72.59	30.73	20	29.43
	Longest**	63.53	29.68	20	27.94
	Random**	33.88	21.5	10	25.94
	Cons	-	32.12	22	29.06

Table 2. Sequence conservation statistics using different methods to select the protein isoforms to align for the complete datasets. Alignments were constructed with MAFFT. % Hit Cons: percentage of cases in which the methods results in the same protein isoform combination as Cons. Differences in the Conservation Score distribution between each method and Cons were evaluated by a Mann-Whitney test,* p-value < 0.05,** p-value < 10^{-3}. In all datasets PALO matches the Cons alignment in a higher proportion of cases than Longest and also shows a higher conservation score than Longest.

TABLE 3

Dataset	Method	% Alignments with gaps	%gaps in alignments	p-value[a]	Mean n gaps	p-value[b]	Mean length gaps
Mammalia	Cons	85.99	12.32	-	4.84	-	2.54
3,827	PALO	85.78	14.33	0.006737	5.27	0.0004089	2.71
	Longest	99.84	19.64	2.20E-016	6.10	2.20E-016	3.21
	Random	97.41	44.29	2.20E-016	5.21	2.20E-016	8.48
Vertebrata	Cons	97.6	16.16	-	10.97	-	1.47
1,836	PALO	97.38	18.54	0.005226	11.31	0.08928	1.63
	Longest	99.94	22.56	2.20E-016	12.26	5.45E-008	1.84
	Random	99.29	43.45	2.20E-016	8.89	2.95E-006	4.88
Metazoa	Cons	99.54	17.29	-	11.14		1.55
221	PALO	99.54	19.31	4.16E-001	11.19	7.52E-001	1.72
	Longest	100	24.4	1.36E-007	12.09	8.57E-002	2.01
	Random	100	42.71	2.20E-016	9.77	1.52E-001	4.36
Paralogs	Cons	83.76	18.91		5.21		3.62
154	PALO	83.76	23.73	5.01E-001	5.62	4.39E-001	4.22
	Longest	98.7	31.86	7.56E-008	6.51	4.77E-003	4.89
	Random	94.8	49.77	6.80E-013	5.34	4.57E-001	9.32

Table 3. Statistics for indels. Data is for families in which PALO selected a different combination than Longest. Gene families for which Longest and PALO selected different protein isoforms. p-value is for a Mann-Whitney test against Cons, for [a]number of positions containing gaps in the alignments and for [b]number of gaps in the alignments