6 Supplement Materials

6.1 Bootstrap procedure

Here is the outline of our bootstrap procedure we have used in the main manuscript.

```
Algorithm 1. Input: Alignments $D_1$ and $D_2$
    Output: An estimated p-value, under the null $H_0 : ||v(T_1) - v(T_2)|| = 0$

  1. Compute tree estimates $v(\hat{T_1}), v(\hat{T_2})$ from the data $D_1,\ D_2$.
  2. Compute $d = ||v(\hat{T_1}) - v(\hat{T_2})||$. //Using the formula in (3).
  3. Let $p = 0$ and for $i = 1$ to $N$ do
      • Take a bootstrap sample $D_1^*$ from the columns of alignment $D_1$ and a bootstrap sample ${D_2}^*$ from the columns of $D'$.
      • Compute tree estimates $v({T_1}^*), v({T_2}^*)$ from the data $D_1^*, {D_2}^*$.
      • Test the condition: if $d \leq ||v(\hat{T_1}) - v({T_1}^*)|| + ||v(\hat{T_2}) - v({T_2}^*)||$ then set $p = p + 1$.
        // the bound in (2) for $p$.
  4. Set $p = \frac{p}{N}$ and return $p$ as the p-value of the hypothesis test.
```

During the "Test the condition" step, one could alternatively use the less conservative condition $d \leq \max(||v(\hat{T_1}) - v({T_1}^*)||, ||v(\hat{T_2}) - v({T_2}^*)||)$

6.2 Proof of Proposition 1

Proof. For any $a \in \mathbb{R}^m$, let a^T be the transpose of a. Since $x_1, x_2, y_1, y_2 \in \mathbb{R}^m$ are mutually independent, the bilinearity of dot-product gives $\mathbb{E}(x_1^T x_2) = \mathbb{E}(x_1)^T \mathbb{E}(x_2)$ and similarly $\mathbb{E}(y_1^T y_2) = \mathbb{E}(y_1)^T \mathbb{E}(y_2)$ and $\mathbb{E}(x_1^T y_1) = \mathbb{E}(x_1)^T \mathbb{E}(y_1)$. Also for any vectors $a, b \in \mathbb{R}^m$, we have the identity $a^T b = b^T a$.

Then we have

$$\begin{aligned} & \mathbb{E}(||x_1 - y_1||^2) \\ = \ & \mathbb{E}((x_1 - y_1)^T (x_1 - y_1)) \\ = \ & \mathbb{E}(x_1^T x_1 + y_1^T y_1 - x_1^T y_1 - y_1^T x_1) \\ = \ & \mathbb{E}(x_1^T x_1) + \mathbb{E}(y_1^T y_1) - \mathbb{E}(x_1^T y_1) - \mathbb{E}(y_1^T x_1) \\ = \ & \mathbb{E}(x_1^T x_1) + \mathbb{E}(y_1^T y_1) - \mathbb{E}(x_1)^T \mathbb{E}(y_1) - \mathbb{E}(y_1)^T \mathbb{E}(x_1) \\ = \ & \mathbb{E}(x_1^T x_1) + \mathbb{E}(y_1^T y_1) - 2\mu_x^T \mu_y. \end{aligned} \tag{5}$$

But

$$\begin{aligned}
& \mathbb{E}(x_1^T x_1) \\
= \; & \mathbb{E}\left[(x_1 - x_2 + x_2)^T (x_1 - x_2 + x_2)\right] \\
= \; & \mathbb{E}[(x_1 - x_2)^T (x_1 - x_2) + x_2^T (x_1 - x_2) \\
& +(x_1 - x_2)^T x_2 + x_2^T x_2] \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mathbb{E}\left[x_2^T (x_1 - x_2)\right] \\
& +\mathbb{E}\left[(x_1 - x_2)^T x_2\right] + \mathbb{E}(x_2^T x_2) \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mathbb{E}(x_2^T x_1) - \mathbb{E}(x_2^T x_2) \\
& +\mathbb{E}(x_1^T x_2) - \mathbb{E}(x_2^T x_2) + \mathbb{E}(x_2^T x_2) \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mathbb{E}(x_2^T x_1) - \mathbb{E}(x_2^T x_2) \\
& +\mathbb{E}(x_1^T x_2) \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mu_x^T \mu_x - \mathbb{E}(x_2^T x_2) \\
& +\mu_x^T \mu_x \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + 2\mu_x^T \mu_x - \mathbb{E}(x_2^T x_2) \\
= \; & \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + 2\mu_x^T \mu_x - \mathbb{E}(x_1^T x_1)
\end{aligned}$$

Thus we have

$$2\mathbb{E}(x_1^T x_1) = \mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + 2\mu_x^T \mu_x. \tag{6}$$

By dividing the both sides of the equation in (6) by 2 we have:

$$\mathbb{E}(x_1^T x_1) = \frac{1}{2}\mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mu_x^T \mu_x. \tag{7}$$

Similarly we have

$$\mathbb{E}(y_1^T y_1) = \frac{1}{2}\mathbb{E}\left[(y_1 - y_2)^T (y_1 - y_2)\right] + \mu_y^T \mu_y. \tag{8}$$

Then we substitute $\mathbb{E}(x_1^T x_1)$ and $\mathbb{E}(y_1^T y_1)$ in equations (7) and (8) into the equation in (5), we have

$$\begin{aligned}
& \mathbb{E}(||x_1 - y_1||^2) \\
= \; & \left[\tfrac{1}{2}\mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \mu_x^T \mu_x\right] \\
& + \left[\tfrac{1}{2}\mathbb{E}\left[(y_1 - y_2)^T (y_1 - y_2)\right] + \mu_y^T \mu_y\right] \\
& -2\mu_x^T \mu_y \\
= \; & \tfrac{1}{2}\mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \tfrac{1}{2}\mathbb{E}\left[(y_1 - y_2)^T (y_1 - y_2)\right] \\
& + \left[\mu_x^T \mu_x + \mu_y^T \mu - 2\mu_x^T \mu_y\right] \\
= \; & \tfrac{1}{2}\mathbb{E}\left[(x_1 - x_2)^T (x_1 - x_2)\right] + \tfrac{1}{2}\mathbb{E}\left[(y_1 - y_2)^T (y_1 - y_2)\right] \\
& +||\mu_x + \mu_y||^2.
\end{aligned} \tag{9}$$

$\square$

6.3 Supplement for simulation studies

Throughout the remaining subsections, Condition (i) refers to the condition tested in Algorithm 1: $d \leq ||v(\hat{T}_1) - v({T_1}^*)|| + ||v(\hat{T}_2) - v({T_2}^*)||$. Condition (ii) refers to the alternative condition $d \leq \max(||v(\hat{T}_1) - v({T_1}^*)||, ||v(\hat{T}_2) - v({T_2}^*)||)$.

6.4 Results with biological data sets

	1000K SD				600K SD				100K SD			
Data set	i	ii	H	SH	i	ii	H	SH	i	ii	H	SH
Sp1_Gene0_1 / Sp1_Gene0_2	NR	NR	NR	NR	NR	NR	R	NR	NR	NR	NR	NR
Sp1_Gene1_1 / Sp1_Gene1_2	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sp1_Gene2_1 / Sp1_Gene2_2	NR	R	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sp1_Gene3_1 / Sp1_Gene3_2	NR	NR	NR	NR	NR	R	NR	NR	NR	R	NR	NR
Sp1_Gene4_1 / Sp1_Gene4_2	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sp1_Gene5_1 / Sp1_Gene5_2	NR	NR	NR	NR	NR	R	NR	NR	NR	R	NR	NR
Sp1_Gene6_1 / Sp1_Gene6_2	NR	R	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sp1_Gene7_1 / Sp1_Gene7_2	NR	NR	NR	NR	NR	NR	NR	NR	NR	R	R	NR
Sp1_Gene8_1 / Sp1_Gene8_2	NR	NR	NR	NR	NR	NR	R	NR	NR	NR	NR	NR
Sp1_Gene9_1 / Sp1_Gene9_2	NR	NR	NR	NR	NR	R	R	NR	NR	NR	NR	NR

Table 4: Testing Type I error by comparing the identical gene trees generated by `Mesquite`. We tested ten gene trees for each species depth (i.e. 30 different gene trees in total) generated under the same species tree. We used two sets of sequences generated under the HKY model with the same tree for each test. We give the statements of reject or non-reject the null hypothesis for the three species depths of 1000,000, 600,000 and 100,000, with fixed population size of 100,000. We used the species tree on right with each species depth in Figure 2. H represents the results with the method in [11] and SH represents the Shimodaira–Hasegawa test [23]. R means the method rejected the null hypothesis and NR means it did not reject the null hypothesis.

Data set	1000K SD			600K SD			100K SD		
	i	ii	dist	i	ii	dist	i	ii	dist
Sp1_Gene0_1 / Sp1_Gene0_2	1.00	0.79	0.046	0.98	0.14	0.071	0.87	0.10	0.142
Sp1_Gene1_1 / Sp1_Gene1_2	0.99	0.43	0.066	0.99	0.80	0.051	0.87	0.18	0.116
Sp1_Gene2_1 / Sp1_Gene2_2	0.33	0.00	0.105	1.00	0.44	0.061	1.00	0.70	0.081
Sp1_Gene3_1 / Sp1_Gene3_2	0.98	0.34	0.063	0.76	0.00	0.092	0.70	0.04	0.136
Sp1_Gene4_1 / Sp1_Gene4_2	0.94	0.11	0.072	0.99	0.37	0.063	1.00	0.41	0.103
Sp1_Gene5_1 / Sp1_Gene5_2	0.55	0.05	0.09	0.75	0.04	0.087	0.81	0.03	0.134
Sp1_Gene6_1 / Sp1_Gene6_2	0.71	0.02	0.091	0.99	0.73	0.055	1.00	0.80	0.090
Sp1_Gene7_1 / Sp1_Gene7_2	0.94	0.15	0.077	1.00	0.60	0.057	0.57	0.01	0.159
Sp1_Gene8_1 / Sp1_Gene8_2	0.98	0.34	0.061	0.97	0.11	0.067	0.95	0.57	0.083
Sp1_Gene9_1 / Sp1_Gene9_2	1.00	0.58	0.058	0.78	0.04	0.09	0.92	0.16	0.112

Table 5: We list the p-values estimated by our method. We tested ten gene trees for each species depth (i.e. 30 different gene trees in total) generated under the same species tree. We used two sets of sequences generated under the HKY model with the same tree for each test. The numbers in the table are p-value for our hypothesis testing. We give p-values for the three species depths of 1000,000, 600,000 and 100,000, with fixed population size of 100,000. We used the species tree on right with each species depth in Figure 2. mean dist represents the average of the normalized difference of means with posterior distributions given the original sequence data sets.

Data set	1000K SD				600K SD				100K SD			
	i	ii	H	SH	i	ii	H	SH	i'	ii'	H	SH
Sp1_g0 / Sp1_g1	NR	NR	R	NR	NR	NR	NR	NR	R	R	R	R
Sp1_g0 / Sp1_g2	NR	R	NR	NR	NR	NR	NR	NR	NR	R	R	NR
Sp1_g0 / Sp1_g3	NR	R	R	NR	NR	R	R	NR	NR	NR	NR	NR
Sp1_g0 / Sp1_g4	NR	R	R	NR	NR	NR	NR	NR	NR	R	R	R
Sp1_g0 / Sp1_g5	NR	R	NR	NR	NR	R	R	NR	NR	R	R	NR
Sp1_g0 / Sp1_g6	NR	R	NR	NR	NR	NR	NR	NR	NR	NR	NR	NR
Sp1_g0 / Sp1_g7	R	R	R	NR	NR	NR	R	NR	NR	R	R	NR
Sp1_g0 / Sp1_g8	NR	R	NR	NR	NR	R	R	NR	NR	R	R	R
Sp1_g0 / Sp1_g9	NR	R	NR	NR	NR	R	R	NR	NR	R	R	NR
Sp1_g0 / Sp2_g1	R	R	R	NR	R	R	R	R	NR	R	R	NR
Sp1_g0 / Sp2_g2	R	R	R	R	R	R	R	R	NR	R	NR	NR
Sp1_g0 / Sp2_g3	R	R	R	R	R	R	R	R	NR	R	R	NR
Sp1_g0 / Sp2_g4	R	R	R	R	R	R	R	R	NR	R	R	R
Sp1_g0 / Sp2_g5	R	R	R	R	R	R	R	R	NR	R	NR	NR
Sp1_g0 / Sp2_g6	R	R	R	R	R	R	R	R	NR	R	R	R
Sp1_g0 / Sp2_g7	R	R	R	R	R	R	R	R	R	R	R	R
Sp1_g0 / Sp2_g8	R	R	R	R	R	R	R	R	NR	R	R	NR
Sp1_g0 / Sp2_g9	R	R	R	R	R	R	R	R	NR	R	NR	R

Table 6: The first nine lines compare different genes derived from the same species tree (the species tree on left in each species depth in Figure 2), the second nine lines compare two genes from different species trees. We give the statements of reject or non-reject the null hypothesis for the three species depths of 1,000,000, 600,000 and 100,000, with fixed population size of 100,000. H represents the results with the method in [11] and SH represents the Shimodaira–Hasegawa test [23]. R means the method rejected the null hypothesis and NR means it did not reject the null hypothesis.

Data set	1000K SD				600K SD				100K SD			
	i	ii	mean dist	true dist	i	ii	mean dist	true dist	i,	ii,	mean dist	true dist
Sp1_g0 / Sp1_g1	0.91	0.07	0.079	0.044	1.00	0.88	0.047	0.044	0.00	0.00	0.284	0.290
Sp1_g0 / Sp1_g2	0.64	0.01	0.093	0.06	0.97	0.32	0.067	0.039	0.12	0.00	0.205	0.182
Sp1_g0 / Sp1_g3	0.23	0.00	0.114	0.074	0.84	0.04	0.088	0.056	0.95	0.21	0.118	0.172
Sp1_g0 / Sp1_g4	0.31	0.01	0.105	0.083	1.00	0.71	0.055	0.051	0.29	0.00	0.204	0.164
Sp1_g0 / Sp1_g5	0.32	0.00	0.111	0.082	0.38	0.00	0.107	0.088	0.31	0.00	0.192	0.229
Sp1_g0 / Sp1_g6	0.37	0.00	0.101	0.055	0.94	0.12	0.078	0.039	0.98	0.49	0.107	0.096
Sp1_g0 / Sp1_g7	0.00	0.00	0.208	0.159	0.98	0.20	0.076	0.049	0.19	0.00	0.207	0.136
Sp1_g0 / Sp1_g8	0.64	0.02	0.087	0.078	0.17	0.00	0.116	0.128	0.05	0.00	0.24	0.269
Sp1_g0 / Sp1_g9	0.64	0.02	0.093	0.079	0.62	0.00	0.101	0.054	0.07	0.00	0.236	0.270
Sp1_g0 / Sp2_g1	0.00	0.00	0.192	0.202	0.00	0.00	0.194	0.176	0.16	0.00	0.218	0.248
Sp1_g0 / Sp2_g2	0.00	0.00	0.207	0.245	0.00	0.00	0.253	0.252	0.48	0.00	0.181	0.156
Sp1_g0 / Sp2_g3	0.00	0.00	0.18	0.209	0.00	0.00	0.193	0.188	0.12	0.00	0.234	0.256
Sp1_g0 / Sp2_g4	0.00	0.00	0.237	0.269	0.00	0.00	0.199	0.232	0.05	0.00	0.244	0.237
Sp1_g0 / Sp2_g5	0.00	0.00	0.213	0.233	0.00	0.00	0.281	0.283	0.29	0.00	0.202	0.205
Sp1_g0 / Sp2_g6	0.00	0.00	0.192	0.245	0.00	0.00	0.259	0.236	0.08	0.00	0.234	0.276
Sp1_g0 / Sp2_g7	0.00	0.00	0.218	0.253	0.00	0.00	0.225	0.231	0.00	0.00	0.308	0.283
Sp1_g0 / Sp2_g8	0.00	0.00	0.177	0.213	0.00	0.00	0.225	0.232	0.15	0.00	0.222	0.218
Sp1_g0 / Sp2_g9	0.00	0.00	0.209	0.235	0.00	0.00	0.225	0.213	0.46	0.00	0.188	0.259

Table 7: Shown are the p-values calculated using our method on eighteen pairs of gene trees. The first nine lines compare different genes derived from the same species tree (the species tree on left in each species depth in Figure 2), the second nine lines compare two genes from different species trees. We give p-values for the three species depths of 1,000,000, 600,000 and 100,000, with fixed population size of 100,000. mean dist represents the average of the normalized difference of means with posterior distributions given the original sequence data sets and true dist represents the normalized distance between true trees.

Data set	i	ii	dist
Gopher-louse (dataset 1)	0.64	0.02	0.120
Gopher-louse (dataset 2)	0.40	0.02	0.128
Gopher-louse (dataset 3)	0.84	0.07	0.117
Gopher-louse (dataset 4)	0.59	0.03	0.122
Grass-endophyte (*tefA* gene)	0.04	0.00	0.073
Grass-endophyte (*tubB* gene)	0.08	0.01	0.088
Grass-endophyte (*tefA_tubB* genes)	0.00	0.00	0.073

Table 8: The numbers in this table are p-values for our statistical hypothesis testing estimated by our method with the grass-endophyte data sets from [22]. dist represents the average of the normalized difference of means with posterior distributions given the original sequence data sets.

Data set	i	ii	Permuted	dist
(- E2368)				
endophyte 16-taxa (lolC vs tubB)	0.34	0.00	0.00	0.100
endophyte 16-taxa (lolC vs tefA)	0.23	0.00	0.00	0.105
endophyte 16-taxa (tefA vs tubB)	0.87	0.05	0.00	0.08
(- E1125)				
endophyte 16-taxa (lolC vs tubB)	0.67	0.01		0.093
endophyte 16-taxa (lolC vs tefA)	0.87	0.04		0.073
endophyte 16-taxa (tefA vs tubB)	0.95	0.16		0.069

Table 9: The numbers in this table are p-values for our statistical hypothesis testing estimated by our method with the endophyte data sets from *tefA*, *tubB*, and *lolC* genes. dist represents the average of the normalized difference of means with posterior distributions given the original sequence data sets. (- E2368) means that they are with the data sets after removing E2368 and (- E1125) means that these are the results with the data sets after removing E1125.

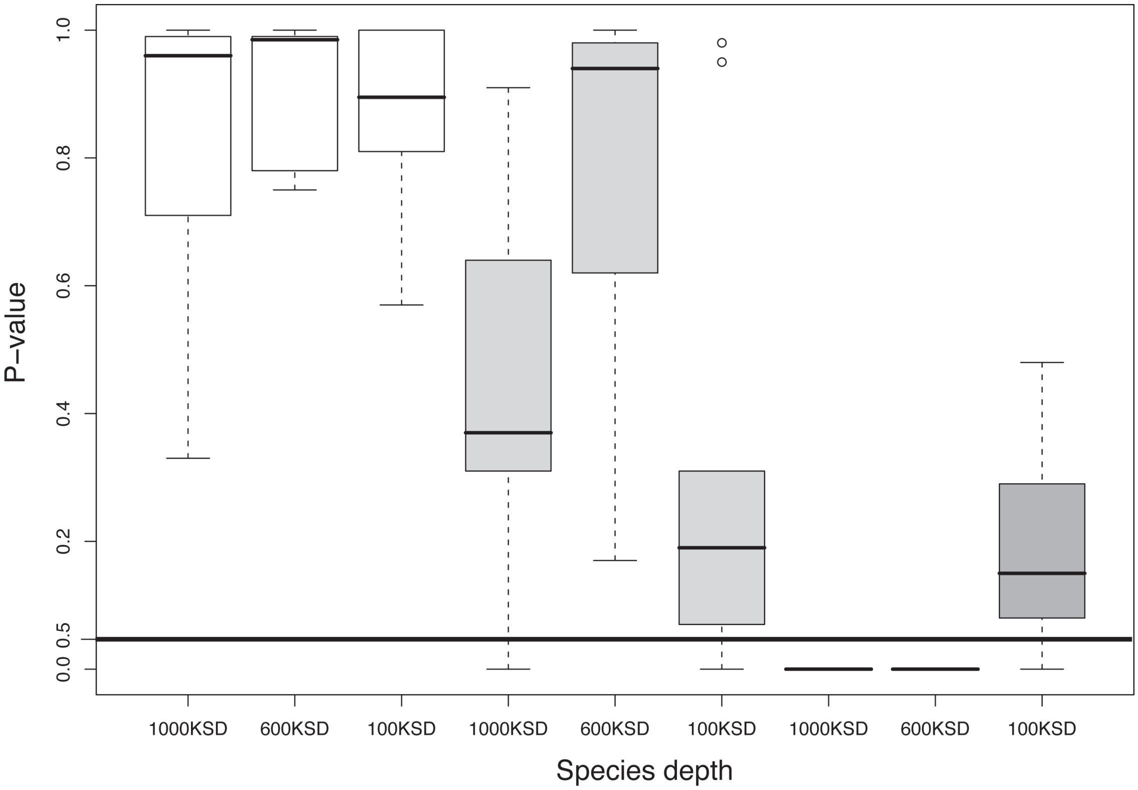

Figure 5: Box plots of p-values for simulated data. We generated ten gene trees for each species depth (i.e. 30 different gene trees in total) for each species tree. Sequences for each gene tree were generated using the HKY model. Species depths of $1,000,000$, $600,000$ and $100,000$ were used, with fixed population size of $100,000$. White box plots are for testing Type I error by comparing identical gene trees generated by `Mesquite` (we used the species tree on the right in Figure 2), light grey box plots represent the p-values with two different gene trees generated from the same species tree under the coalescence model (we used the species tree on right in Figure 2), and dark grey box plots summarize the p-values with gene trees generated from two different species trees. Some boxes of p-values are identically zero. Box plots were computed in `R` [28] using `boxplot`

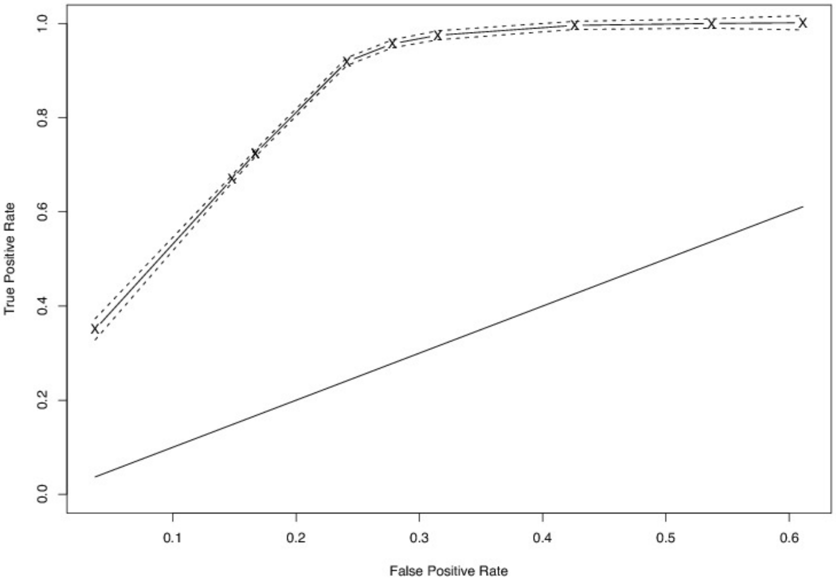

Figure 6: ROC curve for our method with the condition (i). The line labeled 'X' is the ROC curve for our method, and the solid line corresponds to True Positive Rate = False Positive Rate. We used the software `R` [28] to fit the curve using the function `loess`. The dotted lines are for 95% confidence intervals of the fit.