SUPPLEMENTARY TEXT 1

Derivation of eq. [1] in main text

We will derive an equation that estimates the distance between two nucleic acid sequences before introduction of sequencing errors. The derivation will use Bayes's theorem.

Let X and Y be the letters of two sequences (Supplementary Figure 7A). $X_{[k]}$ and $Y_{[k]}$ refer to a letter at a given nucleotide position k within the sequences, and sequences have a total of n positions. Before sequencing (introduction of errors), the letters were originally X_{orig} and Y_{orig}. After sequencing (introduction of errors), some letters change, and the observed sets of letters become X_{obs} and Y_{obs}. Let D_{orig} be letters in X_{orig} and Y_{orig} that are different (when compared at a given position k), and D_{obs} be letters different between X_{obs} and Y_{obs}. The letters that are identical are I_{orig} and I_{obs}. We partition D_{obs} as 1) D_{obs1}, which originate from D_{orig}, and 2) D_{obs2}, which originate from I_{orig}.

Our goal is to calculate the original distance, $P(D_{orig})$, or distance before introduction of errors. It is defined as $P(D_{orig}) = n_{Dorig}/n$, where n_{Dorig} is the number of positions in D_{orig}. We will calculate it from 1) the observed distance, $P(D_{obs})$, or distance after introduction of errors, and 2) the error rates p_x and p_y (defined below).

Distance at a given position

To estimate the distance between X and Y in total, we will first estimate the distance at a given position in X and Y. $P(D_{obs[k]})$ is the observed distance at position k and is either 1 (letters different) or 0 (letters identical). Following Supplementary Figure 7B, we can partition it as

$$P(D_{obs[k]}) = P(D_{obs1[k]}) + P(D_{obs2[k]}) \tag{1}$$

By Bayes's theorem

$$P(D_{obs1[k]}) = \frac{P(D_{obs1[k]}|D_{orig[k]})P(D_{orig[k]})}{P(D_{orig[k]}|D_{obs1[k]})} \tag{2}$$

and

$$P(D_{obs2[k]}) = \frac{P(D_{obs2[k]}|I_{orig[k]})P(I_{orig[k]})}{P(I_{orig[k]}|D_{obs2[k]})} \tag{3}$$

We note $P\left(D_{orig[k]}\right) = 1 - P(I_{orig[k]})$, substitute eq. [2] and [3] into [1], and solve for $P\left(D_{orig[k]}\right)$ to give

$$P(D_{orig[k]}) = \frac{\left[P(D_{obs[k]})P\left(I_{orig[k]} \mid D_{obs2[k]}\right) - P\left(D_{obs2[k]} \mid I_{orig[k]}\right)\right] P\left(D_{orig[k]} \mid D_{obs1[k]}\right)}{P\left(D_{obs1[k]} \mid D_{orig[k]}\right) P\left(I_{orig[k]} \mid D_{obs2[k]}\right) - P\left(D_{obs2[k]} \mid I_{orig[k]}\right) P\left(D_{orig[k]} \mid D_{obs1[k]}\right)} \tag{4}$$

Next, we find expressions for $P\left(D_{orig[k]} \mid D_{obs1[k]}\right)$, $P\left(I_{orig[k]} \mid D_{obs2[k]}\right)$, $P\left(D_{obs1[k]} \mid D_{orig[k]}\right)$, and $P\left(D_{obs2[k]} \mid I_{orig[k]}\right)$. Because all $D_{obs1[k]}$ originate from $D_{orig[k]}$ and all $D_{obs2[k]}$ originate from $I_{orig[k]}$ (Supplementary Figure 7A),

$$P\left(D_{orig[k]} \mid D_{obs1[k]}\right) = 1 \tag{5}$$

and

$$P\left(I_{orig[k]} \mid D_{obs2[k]}\right) = 1 \tag{6}$$

We partition $P\left(D_{obs1[k]} \mid D_{orig[k]}\right)$ and $P\left(D_{obs2[k]} \mid I_{orig[k]}\right)$ as

$$P\left(D_{obs1[k]} \mid D_{orig[k]}\right) = P\left(\alpha\right)_{[k]} + P\left(\beta\right)_{[k]} + P\left(\gamma\right)_{[k]} + P\left(\delta\right)_{[k]} \tag{7}$$

and

$$P\left(D_{obs2[k]} \mid I_{orig[k]}\right) = P\left(A\right)_{[k]} + P\left(B\right)_{[k]} + P\left(\Gamma\right)_{[k]} + P\left(\Delta\right)_{[k]} \tag{8}$$

with terms defined in Supplementary Table 4. For example, $P\left(\alpha\right)_{[k]} = (1 - p_{x[k]})(1 - p_{y[k]})$ is the probability that neither $X_{[k]}$ nor $Y_{[k]}$ change (no errors were introduced) after sequencing, given the letters were different before sequencing (i.e., $X_{[k]}$ and $Y_{[k]}$ belong to D_{orig}). The terms $p_{x[k]}$ and $p_{y[k]}$ are probabilities for change (error rates) for $X_{[k]}$ and $Y_{[k]}$, respectively. We assume all errors are substitutions (not insertions or deletions), giving one of three equally probable outcomes per position per sequence. Quality scores (Q) can be used to calculate the error rates [e.g., $p_{x[k]} = 10^{(-Q_{X[k]}/10)}$].

Substituting expressions for Supplementary Table 4 into eq. [7] and [8] gives

$$P\left(D_{obs1[k]} \mid D_{orig[k]}\right) = -\frac{1}{3}p_{x[k]} - \frac{1}{3}p_{y[k]} + \frac{4}{9}p_{x[k]} \times p_{y[k]} + 1 \tag{9}$$

and

$$P\left(D_{obs2[k]} \mid I_{orig[k]}\right) = p_{x[k]} + p_{y[k]} + \frac{4}{3}p_{x[k]} \times p_{y[k]} \tag{10}$$

By substituting eq. [5], [6], [9], and [10] into eq. [4], we yield

$$P\left(D_{orig[k]}\right) = \frac{9P(D_{obs[k]}) - 9p_{x[k]} - 9p_{y[k]} + 12p_{x[k]} \times p_{y[k]}}{-12p_{x[k]} - 12p_{y[k]} + 16p_{x[k]} \times p_{y[k]} + 9} \tag{11}$$

We can derive eq. [11], with the same result, if we follow Supplementary Figure 7B and partition as $P\left(I_{obs[k]}\right)$ as $P\left(I_{obs[k]}\right) = P(I_{obs1[k]}) + P(I_{obs2[k]})$ (not shown).

Distance across all n

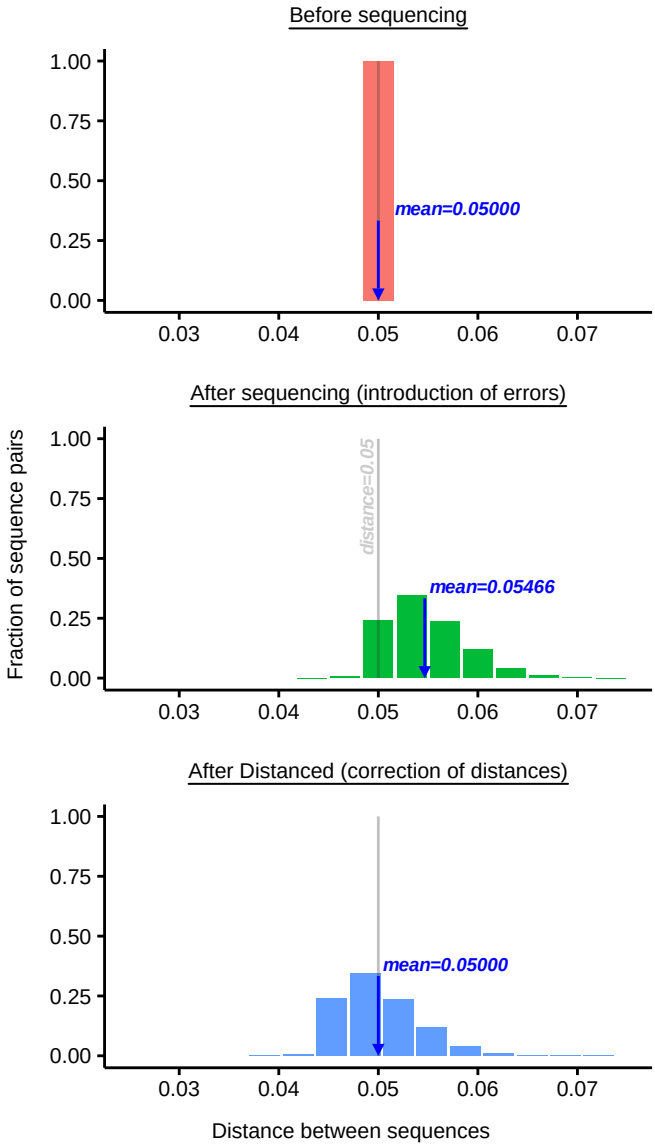
To estimate distance of X and Y in total, we average $P\left(D_{orig[k]}\right)$ across all n positions

$$P\left(D_{orig}\right) = \sum_{k=1}^{n} [P\left(D_{orig[k]}\right)]\frac{1}{n} \tag{12}$$

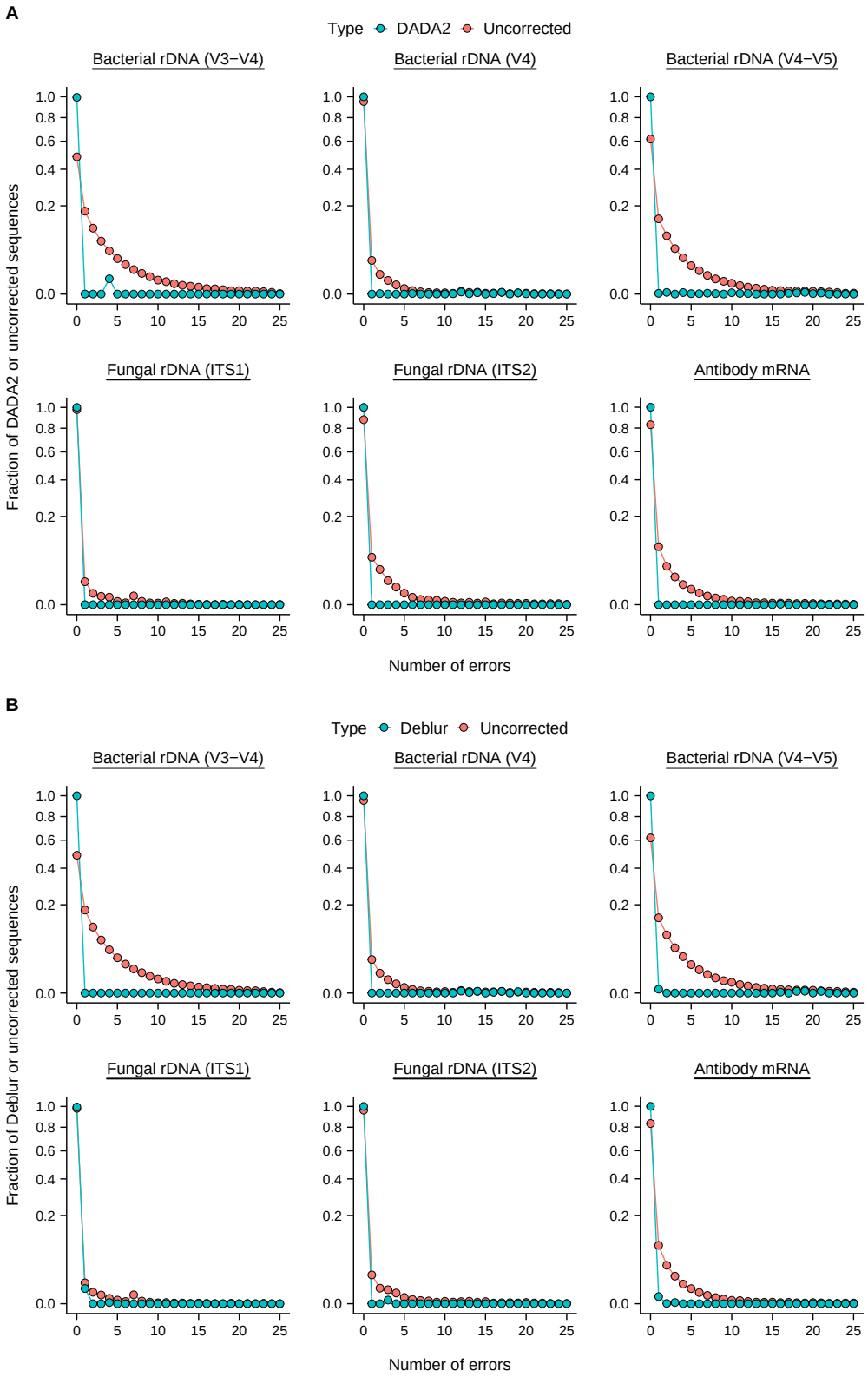
This approach assumes all changes (errors) occur independently (i.e., an error occurring at $k = 1$ does not change the probability of an error at $k = 2$). Eq. [12] can be expanding by substituting in eq. [11], giving

$$P\left(D_{orig}\right) = \sum_{k=1}^{n} [\frac{9P(D_{obs[k]}) - 9p_{x[k]} - 9p_{y[k]} + 12p_{x[k]} \times p_{y[k]}}{-12p_{x[k]} - 12p_{y[k]} + 16p_{x[k]} \times p_{y[k]} + 9}]\frac{1}{n} \tag{13}$$

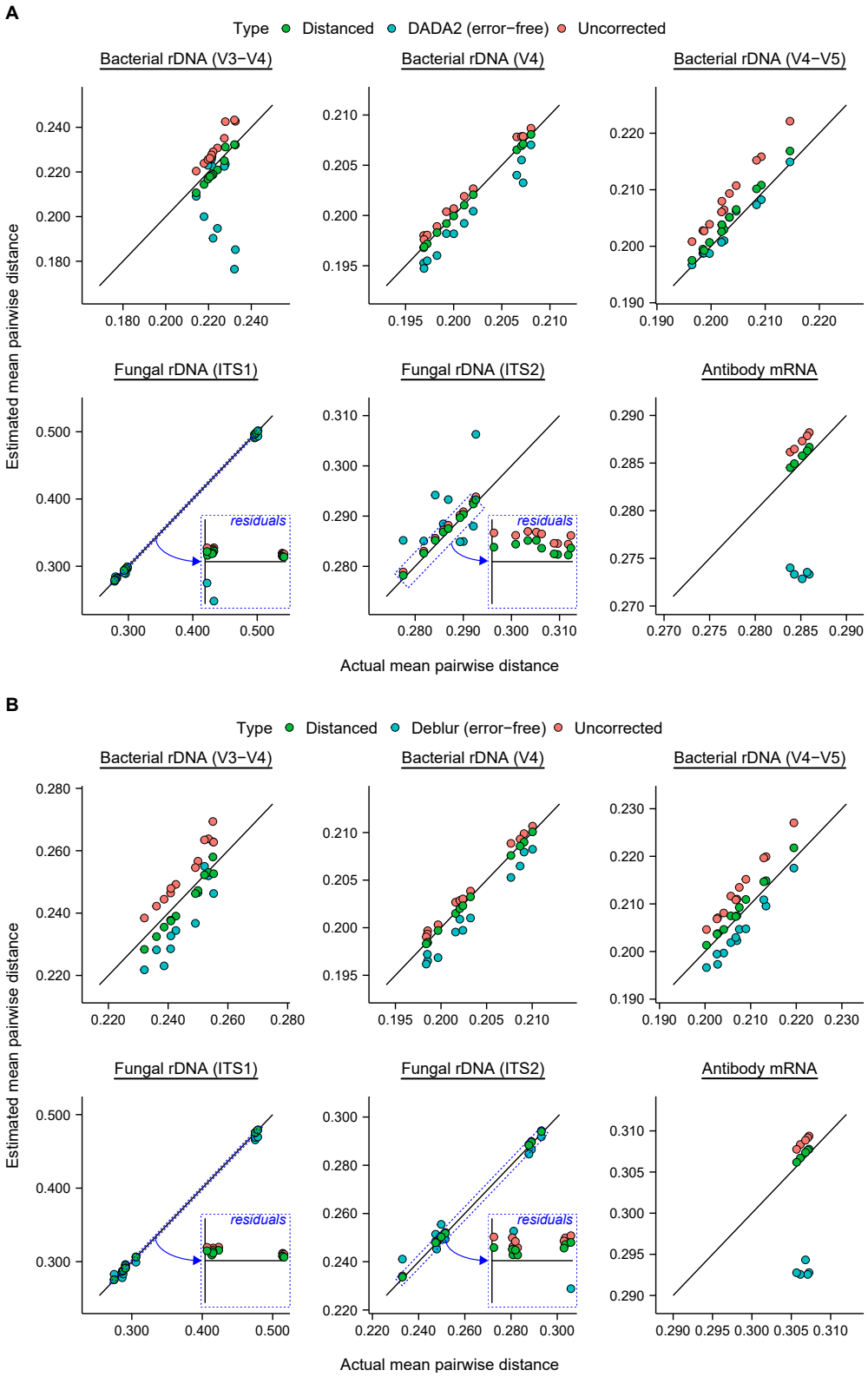
Eq. [13] is eq. [1] in the main text.



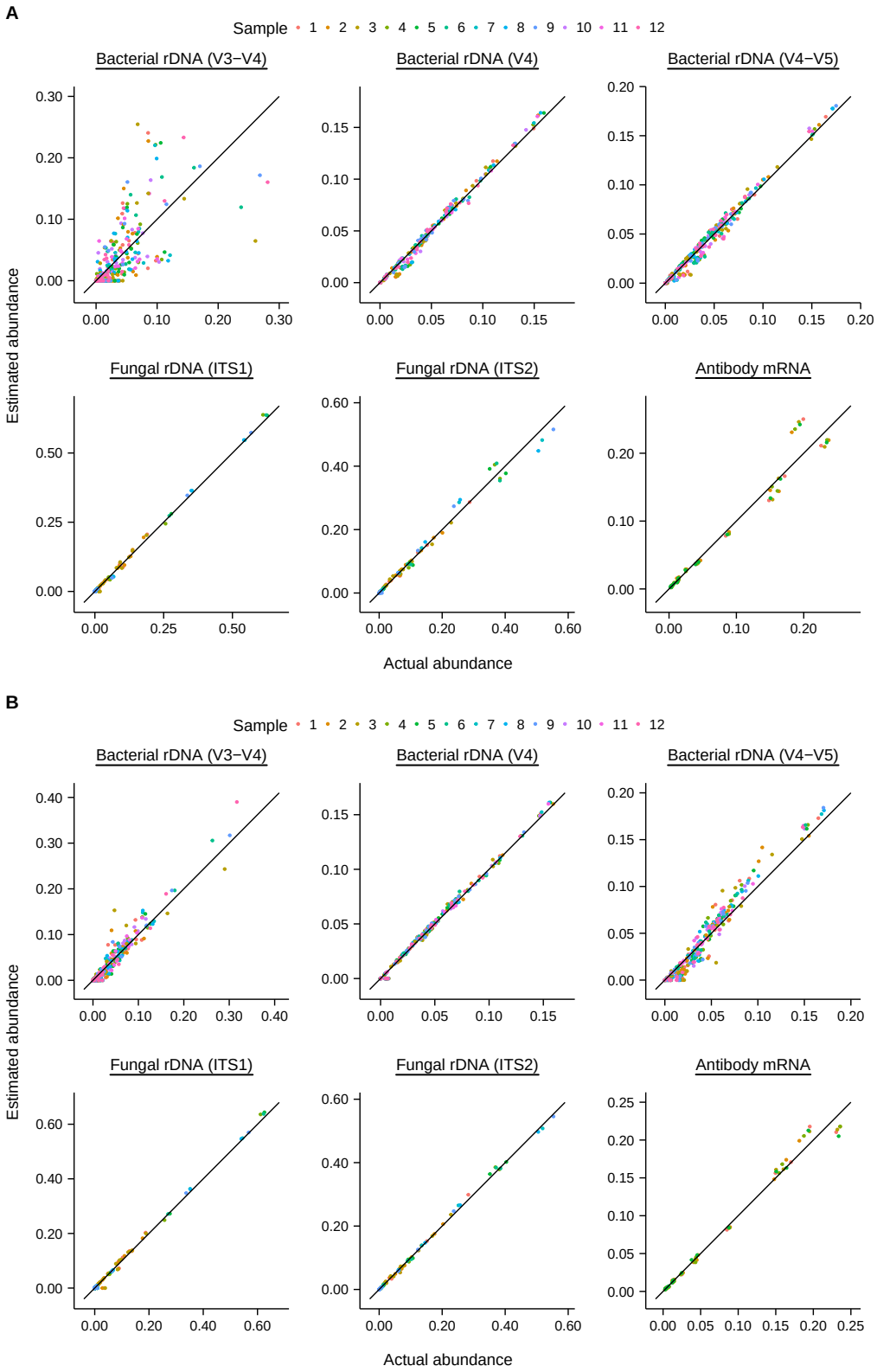
Supplementary Figure 1. Distanced accurately corrects distances of simulated sequence reads. The mean distance after applying Distanced equaled that before sequencing, indicating the correction was accurate. The original distance (0.05) was arbitrary, and the correction was accurate for other values (not shown).



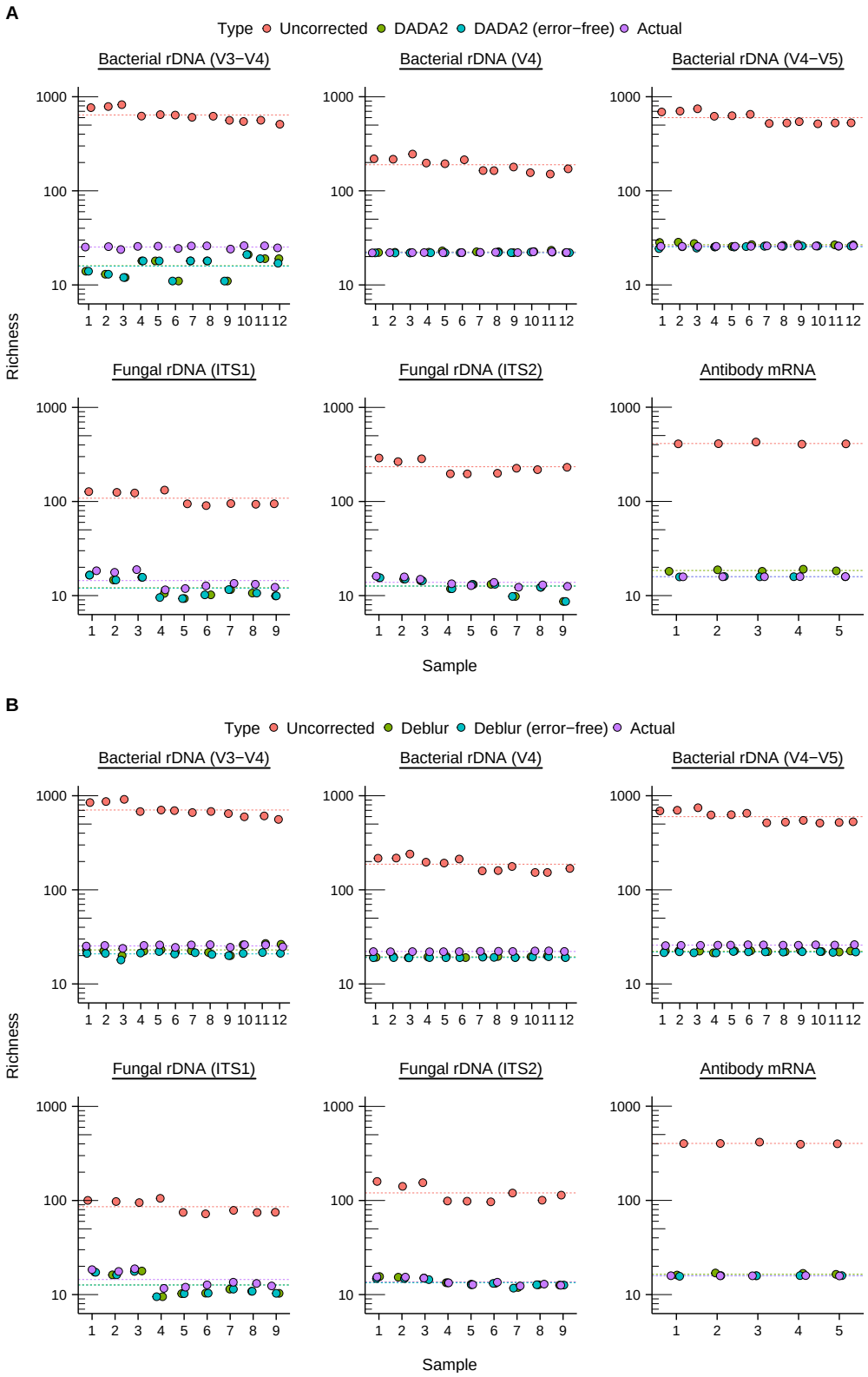
Supplementary Figure 2. Sequences outputted by DADA2 and Deblur have few errors. (A) Frequency of errors for DADA2. (B) Frequency of errors for Deblur. Values for when using no correction for sequencing errors are shown for comparison.



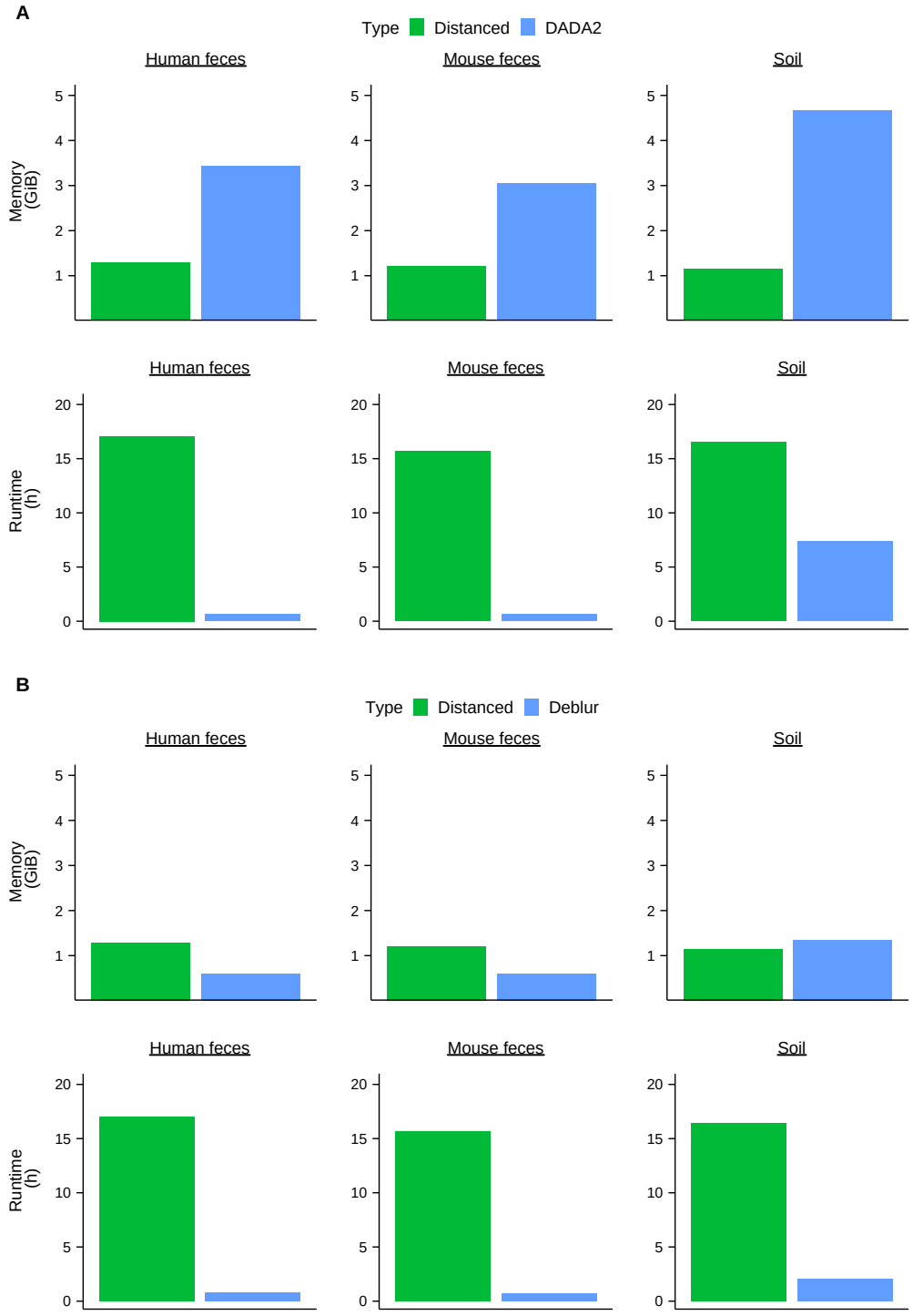
Supplementary Figure 3. After manually correcting errors, DADA2 and Deblur still poorly estimate alpha diversity (mean pairwise distance). Estimates from Distanced are shown alongside those from (A) DADA2 and (B) Deblur. Values are as Figure 3, except DADA2 and Deblur sequences have been manually corrected to remove all remaining errors. Errors were corrected by finding a matching reference sequence for each DADA2 or Deblur sequence. After correcting errors, DADA2 and Deblur sequences differed from the actual sequences only in their abundance. Thus, poor estimates of mean pairwise distance are due to poor estimates of abundance.



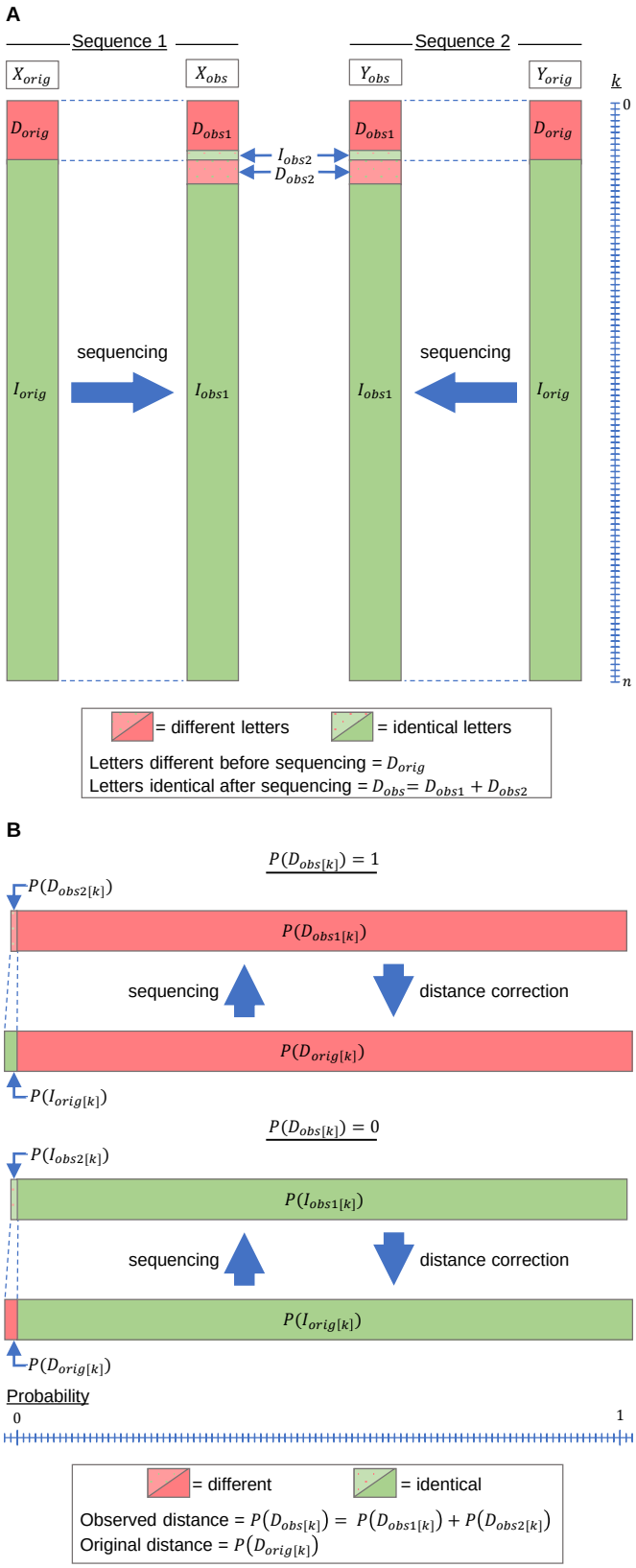
Supplementary Figure 4. DADA2 and Deblur distort the abundance of sequences. Abundance of sequences outputted by (A) DADA2 and (B) Deblur.



Supplementary Figure 5. DADA2 and Deblur can underestimate richness, but still produce estimates close to actual values. (A) DADA2. (B) Deblur. Errors in sequence letters were corrected by DADA2 or Deblur. For comparison, we show estimates of richness obtained when either 1) using no correction or 2) manually correcting all remaining errors in DADA2 sequences [DADA2 (error-free)] or Deblur sequences [Deblur (error-free)].



Supplementary Figure 6. Distanced requires modest memory but has a long runtime. Computational resources required by Distanced are shown alongside those for (A) DADA2 and (B) Deblur. Sequences analyzed were from the V4 region of ribosomal DNA for bacterial communities from human feces, mouse feces, and soil. Distances between sequences were estimated by Distanced, and errors in sequence letters were corrected by DADA2 or Deblur. Mean pairwise distance was calculated using Distanced or a custom R script. This calculation was iterated 100 times per sample. There are twelve samples per community, and resources are for the combined analysis of all twelve samples. The resources reported are only for steps of the analysis done with Distanced, DADA2, or Deblur. Steps done with VSEARCH or R scripts are not included (see Figure 2).



Supplementary Figure 7. *Caption on next page*

Supplementary Figure 7. Two DNA sequences before and after sequencing, illustrating terms in our equations for estimating distances between sequences. (A) The full sequences. (B) A given sequence position k. Letters in the two sequences are X and Y. Letters between X and Y that are different are D, and identical letters are I. The subscripts $orig$ and obs (e.g., in X_{orig} and X_{obs}) refer to conditions before and after sequencing. Each letter has a position k, and there are a total of n positions. In (A), letters in I_{orig} are grouped separately from letters in D_{orig}, though they would be interspersed in a real sequence. Letters in D_{obs1}, D_{obs2}, I_{obs1}, and I_{obs2}, are grouped in the same way. In (B), we show the condition where letters $X_{obs[k]}$ and $Y_{obs[k]}$ are different $[P\left(D_{obs[k]}\right) = 1]$. We also show the condition where letters $X_{obs[k]}$ and $Y_{obs[k]}$ are identical $[P\left(D_{obs[k]}\right) = 0]$.