A Bayesian Model of the DNA Barcode Gap

Supplementary Information

Jarrett D. Phillips

1 Mathematical Details

For a given random variable X, its cumulative distribution function (CDF) is defined by

$$F_X(t) = \mathbb{P}(X \leq t) = 1 - \mathbb{P}(X > t). \tag{1}$$

Rearranging Equation (1) gives

$$\mathbb{P}(X > t) = 1 - F_X(t). \tag{2}$$

Equations (1)-(4) within the main text can thus be expressed in terms of empirical cumulative distribution functions (ECDFs) as follows, since the true underlying CDFs, $F(\cdot)$, are unknown *a priori*, and therefore must be estimated using available data:

$$\begin{aligned} p_x &= \mathbb{P}(d_{ij} \geq a) \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) \end{aligned} \tag{3}$$

$$\begin{aligned} q_x &= \mathbb{P}(d_{XY} \leq b) \\ &= \hat{F}_{d_{XY}}(b) \end{aligned} \tag{4}$$

$$\begin{aligned} p'_x &= \mathbb{P}(d_{ij} \geq a') \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{XY} = a') \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{XY} = a') \end{aligned} \tag{5}$$

$$\begin{aligned} q'_x &= \mathbb{P}(d'_{XY} \leq b) \\ &= \hat{F}_{d'_{XY}}(b) \end{aligned} \tag{6}$$

noting that $\mathbb{P}(d_{ij} \geq b) = \hat{F}_{d_{ij}}(b) + \mathbb{P}(d_{ij} = b) = 0$, $\mathbb{P}(d_{XY} \leq a) = \hat{F}_{d_{XY}}(a) = 0$, and $\mathbb{P}(d'_{XY} \leq a') = \hat{F}_{d'_{XY}}(a') = 0$ (see **Figure 1** in main text). Given n increasing-ordered data points, the (discrete) ECDF, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[x_i \leq t]}$, comprises a step function having jump discontinuities of size $\frac{1}{n}$ at each sample observation (x_i), excluding ties (or steps of weight $\frac{i}{n}$ with tied observations), where $\mathbb{1}(x)$ is the indicator function. Equations (3)-(6) herein clearly demonstrate the asymmetric directionality of the proposed metrics. Furthermore, calculation of the DNA barcode gap estimators given in the main text is straightforward as they implicitly account for both total distribution area and overlap. Nevertheless, the total areas bounded by intraspecific and interspecific distributions on [a, b], and combined distributions on [a', b] (see **Figure 1** in the main text) are given by

$$\begin{aligned} A &= p_x + q_x \\ &= \mathbb{P}(d_{ij} \geq a) + \mathbb{P}(d_{XY} \leq b) \\ &= 1 - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \mathbb{P}(d_{ij} = a) + \hat{F}_{d_{XY}}(b) \end{aligned} \tag{7}$$

$$\begin{aligned} A' &= p'_x + q'_x \\ &= \mathbb{P}(d_{ij} \geq a') + \mathbb{P}(d'_{XY} \leq b) \\ &= 1 - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \\ &= \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \mathbb{P}(d_{ij} = a') + \hat{F}_{d'_{XY}}(b) \end{aligned} \tag{8}$$

whose values lie in [0, 2]. Similarly, the degree of distribution overlap is

$$\begin{aligned} O &= \hat{F}_{\min(d_{ij}, d_{XY})}(b) - \hat{F}_{\min(d_{ij}, d_{XY})}(a) \\ O' &= \hat{F}_{\min(d_{ij}, d'_{XY})}(b) - \hat{F}_{\min(d_{ij}, d'_{XY})}(a') \end{aligned}$$

whose support is on [0, 1].