A Bayesian Model of the DNA Barcode Gap

Supplementary Information

Jarrett D. Phillips

Equations (1)-(4) within the main text can be expressed in terms of empirical cumulative distribution functions (ECDFs) as follows, since the true underlying CDFs, $F(\cdot)$, are unknown

$$p_x = \mathbb{P}(d_{ij} \geq a) \quad = 1 - \hat{F}_{d_{ij}}(a) = \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) \tag{1}$$

$$q_x = \mathbb{P}(d_{XY} \leq b) = \hat{F}_{d_{XY}}(b) \tag{2}$$

$$p'_x = \mathbb{P}(d_{ij} \geq a') \quad = 1 - \hat{F}_{d_{ij}}(a') = \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') \tag{3}$$

$$q'_x = \mathbb{P}(d'_{XY} \leq b) = \hat{F}_{d'_{XY}}(b) \tag{4}$$

noting that $\hat{F}_{d_{ij}}(b) = 1$, $\hat{F}_{d_{XY}}(a) = 0$, and $\hat{F}_{d'_{XY}}(a') = 0$ (see **Figure 1** in main text). Given n increasing-ordered data points, the 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 $\frac{i}{n}$ with tied observations), where $\mathbb{1}(x)$ is the indicator function. Equations (1)-(4) herein clearly demonstrate the asymmetric directionality of the metrics. Furthermore, calculation of the DNA barcode gap estimators given herein is straightforward as they implicitly account for both total distribution area and overlap. Nevertheless, the total area bounded by intraspecific and interspecific distributions is given by the joint ECDF. Thus,

$$A = p_x + q_x = 1 - \hat{F}_{d_{ij}}(a) + \hat{F}_{d_{XY}}(b) = \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a) + \hat{F}_{d_{XY}}(b) \tag{5}$$

$$A' = p'_x + q'_x = 1 - \hat{F}_{d_{ij}}(a') + \hat{F}_{d_{XY}}(b) = \hat{F}_{d_{ij}}(b) - \hat{F}_{d_{ij}}(a') + \hat{F}_{d_{XY}}(b) \tag{6}$$

whose values lie in [0, 2].