CORRIGENDUM

In the article by SUDHIR KUMAR and SUDHINDRA R. GADAGKAR (GENETICS **158**:1321-1327) entitled "Disparity index: A simple Statistic to measure and test the homogeneity of substitution patterns between molecular sequences" the analytical proof on pages 1321 - 1322) showing that $E(D_C) = E(N_d)$ is now rewritten (equations 1 - 10) as follows. This correction does not affect any of the results or conclusions of the article. In fact, the new proof now extends the equality in Equation (10) even when there is *among-site* (*i*) heterogeneity in substitution rates (see also Figure 1 in the original article), (*ii*) correlation in evolutionary rates/patterns, and (iii) variation in substitution patterns.

Let X and Y be two aligned DNA sequences of length L each. Let x_i be the count of the ith type of nucleotide (i = A, C, G, T) in sequence X, and y_i be the corresponding count in sequence Y. The composition distance (D_C) between X and Y is

$$D_C = \frac{1}{2}\sum_i (x_i - y_i)^2 \qquad (1)$$

The expected value of D_C is

$$E(D_C) = \frac{1}{2}\sum_i \left[E\left(x_i^2\right) + E\left(y_i^2\right) - 2E\left(x_i \times y_i\right)\right] \qquad (2)$$

Now,

$$E\left(x_i^2\right) = E\left[\left(\sum_{k=1}^{L} \alpha_i^k\right)^2\right], \qquad (3)$$

where $\alpha_i^k = 1$ if a site k in sequence X contains nucleotide i; and 0 otherwise.

Equation (3) can be expressed as

$$E\left(x_i^2\right) = E\left[\sum_{k=1}^{L}\left(\alpha_i^k\right)^2\right] + E\left[\sum_{k=1}^{L}\sum_{k'\neq k}^{L}\alpha_i^k\alpha_i^{k'}\right] = \sum_{k=1}^{L}P\left(\alpha_i^k\right) + \sum_{k}^{L}\sum_{k'\neq k}^{L}P\left(\alpha_i^k,\alpha_i^{k'}\right) \quad (4)$$

where $P\left(\alpha_i^k\right)$ is the probability of observing the *i*-th nucleotide at site *k* in sequence *X*, and $P\left(\alpha_i^k,\alpha_i^{k'}\right)$ is the joint probability of observing nucleotide *i* at sites *k* and k' in sequence *X*.

Similarly,

$$E\left(y_i^2\right) = \sum_{k=1}^{L}P\left(\beta_i^k\right) + \sum_{k}^{L}\sum_{k'\neq k}^{L}P\left(\beta_i^k,\beta_i^{k'}\right). \quad (5)$$

where $\beta_i^k = 1$ if a site *k* in sequence *Y* contains nucleotide *i*, and 0 otherwise. Also,

$P\left(\beta_i^k\right)$ is the probability of observing the *i*-th nucleotide at site *k* in sequence *Y*, and $P\left(\beta_i^k,\beta_i^{k'}\right)$ is the joint probability of observing nucleotide *i* at sites *k* and k' in sequence *Y*.

Furthermore,

$$E\left(x_i \times y_i\right) = \sum_{k=1}^{L}\sum_{k'=1}^{L}P\left(\alpha_i^k,\beta_i^{k'}\right) = \sum_{k=1}^{L}P\left(\alpha_i^k,\beta_i^k\right) + \sum_{k=1}^{L}\sum_{k'\neq k}^{L}P\left(\alpha_i^k,\beta_i^{k'}\right) \quad (6)$$

Now, when the evolutionary patterns are homogeneous and stationary in the two lineages,

$$\begin{aligned} P\left(\alpha_i^k\right) &= P\left(\beta_i^k\right), \text{ and} \\ P\left(\alpha_i^k, \alpha_i^{k'}\right) &= P\left(\beta_i^k, \beta_i^{k'}\right) = P\left(\alpha_i^k, \beta_i^{k'}\right). \end{aligned} \tag{7}$$

Using (7) and substituting (4)-(6) in (2), we get

$$\begin{aligned} E(D_{\mathrm{C}}) &= \sum_i \left[\sum_{k=1}^{L} P\left(\alpha_i^k\right) - \sum_{k=1}^{L} P\left(\alpha_i^k, \beta_i^k\right) \right] \\ &= \sum_{k=1}^{L} \left[\sum_i P\left(\alpha_i^k\right) - \sum_i P\left(\alpha_i^k, \beta_i^k\right) \right]. \end{aligned} \tag{8}$$

Because $\sum_i P\left(\alpha_i^k\right)$ is equal to 1 for each site, equation (8) is simplified as

$$E(D_{\mathrm{C}}) = L - \sum_{k=1}^{L} \sum_i P\left(\alpha_i^k, \beta_i^k\right) \tag{9}$$

Here, $\sum_{k=1}^{L} \sum_i P\left(\alpha_i^k, \beta_i^k\right)$ is the sum of probability of identity site-by-site, and thus gives the expected number of identical sites. Therefore, when subtracted from the sequence length L, the right hand side in (9) becomes the expected number of differences (N_{d}). That is,

$$E\left(D_{\mathrm{C}}\right) = E\left(N_{\mathrm{d}}\right). \tag{10}$$

This equality holds true for any number of states (e.g., 20 in amino acid sequences). Also note that the probability of identity is specified individually for each site in (9) and summed over all sites. Therefore, we do not need to assume that the substitution pattern is the same *among sites* or that the evolutionary rate is equal *among sites*. Furthermore, we do not need to assume site independence because the joint probabilities in equations

(4) and (5) are not required to be expressed as multiples of the individual probabilities to get equation (9). Therefore, equation (10) holds true under a variety of biologically realistic conditions and only requires that the evolutionary substitution pattern be homogeneous at individual sites between lineages.