BACKGROUND

Review of the mathematics of function logos from [re?]:
A Structural feature is defined as a state $x \in X$ of a sequence at an alignment position l, $1 \leqslant l \leqslant L$, where L is equal to the length of the sequence, and $X = \{A, C, G, U, -\}$. The possible functional classes of tRNAs are denoted by $Y = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y \}$. Functional information $I_l(Y|x)$ that a state x confers about the frequencies of different classes Y at position l is calculated for each feature x_l as $I_l(Y|x) = H(Y) - e(n_l(x)) - H_l(Y|x)$.
Function logo is a symbol-stacked-bar graph with positions on x axis and informations on y axis where each symbol within a bar is one of the functional classes of tRNAs in set Y and symbols are sorted based on their height. Height of each symbol y ($y \in Y$) for feature x_l is proportional to the frequency of sequences of class y with that feature and is calculated as $(\sum_{w \in Y} p_l(w|x)/p(w))I_l(Y|x)$ [ref?].

TSFM METHOD FOR CALCULATING INFORMATION DIFFERENCE AND KULLBACKELEIBLER DIVERGENCE DIFFERENCE

We used aligned tRNA sequences of Human and each of the TriTryp classes of genomes as input. we divided the dataset into two alignments as foreground dataset G^F and background dataset G^B. each of Homo and TriTryp are considered once as foreground and once as background for further analysis. We assumed that we have at least one sequence of each 21 tRNA classes annotated for each of the background and foreground dataset.

Information Difference (ID) logos

ID logos [?] visualize the evolutionary gain or loss of functional information between two genomes referred as foreground and background. Information difference of foreground vs background for a feature X_l is calculated as $\Delta I(Y|X_l = x) = I^F(Y|X_l = x) - I^B(Y|X_l = x)$. In our information difference function logos the negative information difference is not displayed and the height of each symbol within a bar is proportional to $\frac{p^F(y|x_l)/p^F(y)}{p^B(y|x_l)/p^B(y)}$. $p^F(y|x_l)$ is the posterior probability of class membership given feature X_l; in other words, it is probability of a tRNA gene with feature X_l being of class y. In this work, posterior probability of both background and foreground is calculated using pseudo counts when one of the 21 tRNA classes is absent for a feature X_l in either background or foreground or both. pseudo counts are applied by adding one to the frequency of each of 21 classes. (note that based on our assumption prior will never becomes zero.)

KullbackeLeibler divergence difference (KLD) logos

KLD logos [?] complement the Information Difference (ID) measure to visualize the changes in the functional associations of features of foreground genome against background genome. We calculate the height of KLD logos in each stack as $D_{KL}(Y|X_l = x) = D_{KL}(P^F(y|x_l)\|p^B(y|x_l)) = \sum_{y \in Y} p^F(y|x_l) \log_2(\frac{p^F(y|x_l)}{p^B(y|x_l)})$ and height of each symbol within a bar proportional to $\frac{p^F(y|x_l)/p^F(y)}{P^B(y|x_l)/p^B(y)}$. The posterior probabilities are calculated with the same criteria described for ID logos. Further, we added two other restrictions in calculating the KL-divergence to overcome non-smoothness of this measure. First, We know that when $\lim_{p \to \infty} p \log p = 0$. So, to avoid calculating logarithm of zero, whenever the foreground posterior probability for a feature X_l becomes zero, we set the KLD to zero. Second, we know that when $p \neq 0$ and $\lim_{q \to 0}$, $kld(p\|q)$ is defined as ∞. which means that when one event e is possible in foreground distribution ($P^F(y|x_l) > 0$) and not possible in background distribution ($P^B(y|x_l) = 0$) then two distributions are absolutely different. However distributions are derived from observations and we must take into account the possibility of unseen events. In order to overcome this issue whenever the sample size of background having a feature X_l is less than a small number (here we used 6), which will make the KLD a larg number, we set the KLD to zero; in other words, we cannot say that two distributions are very different based on a small sample size so we won't display the KLD for this feature.