# Supplementary Online Materials for Lawrence *et al.*, "tSFM 1.0: tRNA Structure-Function Mapper"

*Authors*
T.J. Lawrence, F. Hadi-Nezhad, I. Grosse, and D.H. Ardell

February 27, 2021

# Contents

## 0.1 Availability of Code and Data

Complete code and data to generate our results are provided separately through the Github repository at https://github.com/tlawrence3/tSFM.

## 0.2 Mathematical Presentation of Statistics Calculated in `tSFM v1.0`

`tSFM v1.0` has been written to function in a general setting for the analysis of any sequence family encompassing multiple (sub-)functions, whether a noncoding RNA family, protein superfamily, or family of related genes or genetic elements. It infers structural alphabets and functional sets from its inputs. RNA base modifications could in principle be symbolized and processed directly by `tSFM v1.0`. Nevertheless, in the sequel, we pre-specify simple RNA, protein and DNA alphabets for clarity.

### 0.2.1 Definitions and Preliminaries

Let $\mathcal{T}$ be a set of taxa. Let $\mathcal{B}_\mathcal{R} \equiv \{A, C, G, U\}$ be the set of RNA nucleobases, $\mathcal{B}_\mathcal{D} \equiv \{A, C, G, T\}$ be the set of DNA nucleobases, $\mathcal{A}$ be the union of all amino acids encoded by any taxon $t \in \mathcal{T}$, and $\mathcal{C}$ be a set of structural aligned sites across a DNA element or gene, noncoding RNA, or protein family, such as the set of all Sprinzl coordinates [1] in the consensus structure of tRNAs. Then we define the set of single-site features $\mathcal{D}$ as $\mathcal{D} \equiv \mathcal{B}_\mathcal{R} \times \mathcal{C}$ for RNAs, $\mathcal{D} \equiv \mathcal{A} \times \mathcal{C}$ for proteins, or $\mathcal{D} \equiv \mathcal{B}_\mathcal{D} \times \mathcal{C}$ for DNA elements.

Let $\mathcal{P} \equiv \mathcal{B}_\mathcal{R} \times \mathcal{B}_\mathcal{R}$ be the set of all 16 base-pairs or base mis-pairs in the Cartesian square of $\mathcal{B}_\mathcal{R}$, and let $\mathcal{Q} \subset \mathcal{C} \times \mathcal{C}$ be the set of all pairs of structurally aligned coordinates involved in potential base-pairs in the consensus secondary structure of an RNA family. Then we define the set of paired-site features $\mathcal{R}$ in that RNA family as $\mathcal{R} \equiv \mathcal{P} \times \mathcal{Q}$, and the set of single- and paired-site features in that RNA family as $\mathcal{S} \equiv \mathcal{D} \cup \mathcal{R}$. Otherwise, for a protein or DNA element/gene family, let $\mathcal{S} \equiv \mathcal{D}$.

Let $\mathcal{F}$ be a set of all functional classes of the protein, RNA, or DNA element/gene family; for example, in eukaryotic tRNA genes, $\mathcal{F}$ contains 20 classes of elongator tRNAs and one initiator tRNA class.

Given taxa $t, u \in \mathcal{T}$, let $F$ and $G$ be random variables specifying the probabilities that a random sequence belonging to the family under study isolated from taxon $t$ (respectively $u$) is of functional class $f \in \mathcal{F}$. We assume $F$ and $G$ have respective probability mass functions $p_F(f)$ and $p_G(f)$ estimated by the fractions of annotated genes or elements in genomes of class $f$ in taxa $t$ and $u$ respectively. Similarly, given feature $s \in \mathcal{S}$, let $M$ and $N$ be random variables specifying the conditional probabilities that a random gene/element, RNA or protein isolated from taxon $t$ (respectively $u$) containing feature $s \in \mathcal{S}$ is of functional class $f \in \mathcal{F}$. We assume that $M$ and $N$ have probability mass functions $p_M(f|s)$ and $p_N(f|s)$ respectively, estimated by the fraction of annotated genes, RNAs or proteins of class $f$ from taxon $t$ or $u$ respectively that contain feature $s$.

### 0.2.2 Statistics for Class-Informative Features (CIFs)

`tSFM v1.0` estimates the structure-conditioned functional information of feature $s$ in taxon $t$ as $\hat{I} = \hat{H}(F) - \hat{H}(M)$, where $\hat{H}(\cdot)$ is one of three different estimators of Shannon entropy [2–4]. The proportion of functional information $\hat{h}(f|s,t)$ attributed to functional class $f$ given feature $s$ and taxon $t$ is estimated using Gorodkin letter-heights [5] as described previously [6].

### 0.2.3 Statistics for Divergences of CIFs between Taxa

`tSFM v1.0` re-implements ID logos and KLD logos [7] to contrast CIFs between taxa. To estimate divergence in the structure-conditioned functional information of feature $s \in \mathcal{S}$ between two taxa $t, u \in \mathcal{T}$, we separately estimate the structure-conditioned functional information of the feature in taxa $t$ and $u$, obtaining respective values $\hat{I}$ and $\hat{J}$, and compute two Information Difference (ID) values: the ID $\hat{\Delta}_{t|u} \equiv \max(\hat{I} - \hat{J}, 0)$ of $s$ in $t$ with $u$ as background; and the ID $\hat{\Delta}_{u|t} \equiv \max(\hat{J} - \hat{I}, 0)$ of $s$ in $u$ with $t$ as background. To contrast functional associations of a feature $s$ in taxa $t, u \in \mathcal{T}$, we compute two Kullback-Leibler Divergences (KLD): the KLD $\hat{D}_{t|u} \equiv \hat{D}\,(M\|N)$ of $s$ in $t$ with $u$ as background if at least one sequence contains the feature $s$ in taxon $u$, otherwise $\hat{D}_{t|u} \equiv 0$; and the KLD $\hat{D}_{u|t} \equiv \hat{D}\,(N\|M)$ of $s$ in $u$ with $t$ as background if at least one sequence contains the feature $s$ in taxon $t$, otherwise $\hat{D}_{u|t} \equiv 0$. If the probability mass of any functional class is zero in $M$ or $N$, we add pseudo-counts. The proportion of gain or change attributed to a specific functional class $f$ of a feature in taxon $t$ compared to $u$ is computed as $(p_M(f|s)/p_F(f))/(p_N(f|s)/p_G(f))$. Here, if feature $s$ is not observed in functional type $f$ in taxon $u$, we add pseudo-counts before computing $p_N(f|s)$, assuming $p_F(f) > 0$ and $p_G(f) > 0$ for all $f \in \mathcal{F}$ .

## 0.3 Significance Calculations for CIFs

`tSFM v1.0` estimates the statistical significance of $\hat{I}$ and $\hat{h}(f|s,t)$ by repeatedly randomly permuting functional class labels over the input gene sequence data and computing the fraction of random permutation replicates with $\hat{I}^* \geq \hat{I}$ or $\hat{h}^*(f|s,t) \geq \hat{h}(f|s,t)$ respectively (starred estimators are computed from permutation pseudo-replicates).

## 0.4 Significance Calculations for CIF Divergences

`tSFM v1.0` estimates the statistical significance of an ID value $\hat{\Delta}$ or KLD value $\hat{D}$ for a given feature $s \in \mathcal{S}$ by repeatedly randomly permuting taxon labels over input gene sequence data containing feature $s$ and estimating a $p$-value based on these random permutations using algorithm Approximate. To compute the $p$-value $\Pr(x \geq x_0)$ of a KLD or ID statistic $x_0 > 0$ for a given $s$, up to $R$ (by default 10,000) permutation replicates $\{y_i^*\}_{i=1}^R$ of $x_0$ are generated by permuting taxon labels over sequences containing $s$ and recomputing the statistic. If $S$ (by default 10) permutation replicates in $\{y_i^*\}_{i=1}^R$ exceed $x_0$, the algorithm terminates and returns $P_{\text{ECDF}}$. The default value of the "target" permutation number $T$ to start to attempt fitting of the Generalized Pareto Distribution (GPD) function $F_{\text{GPD}}$ to a

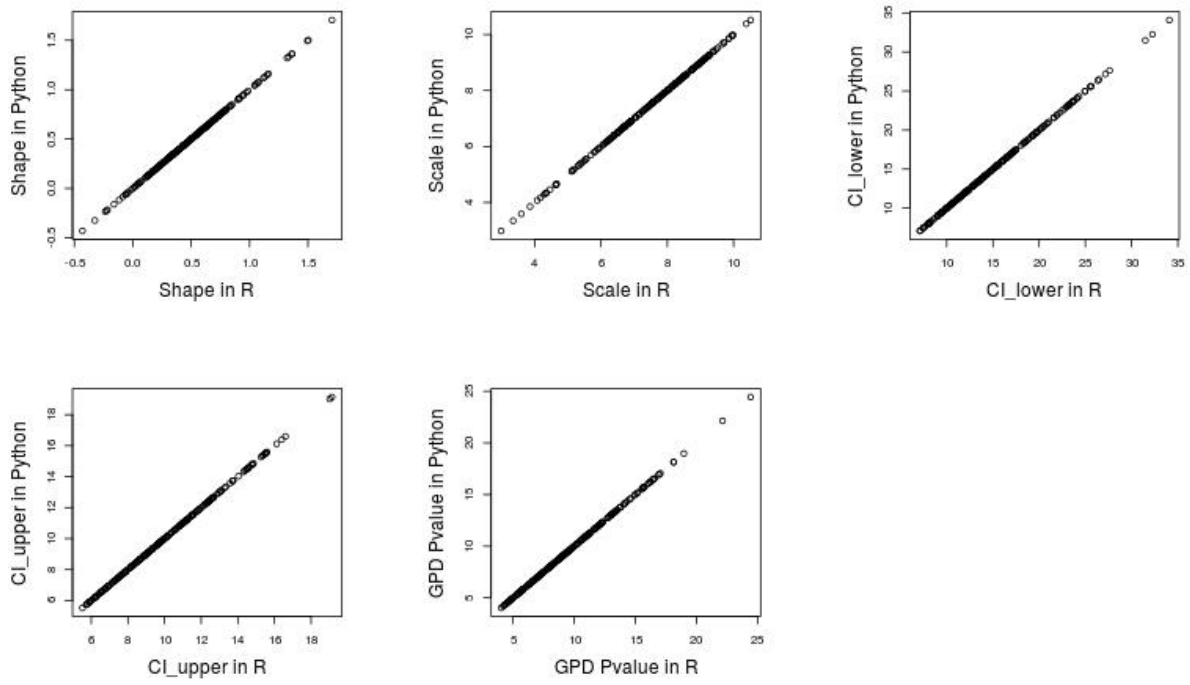


Figure S1: Validation of Python-based GPD-estimation procedure, GPD-based $p$-value calculation, and confidence intervals for the GPD-based $p$-value calculation against the Multivariate Extreme Values Distribution package in R (`mev v1.13.1`) for 237 randomly chosen features from the ENRIETTII clade against humans.

set of peaks-over-threshold $\{z_i^*\}_{i=1}^V$ is 500, where $V < T$ is the number of observed peaks-over-threshold. We test goodness-of-fit of the peaks $\{z_i^*\}_{i=1}^V$ to the fitted GPD function with a decision criterion of $p_G \geq 0.05$, where $p_G$ is the $p$-value of the Anderson-Darling statistic. The default preferred number of peaks-over-threshold $U$ for fitting to the GPD is 250.

To compensate for multiple testing, `tSFM` computes FWER or FDR values using the `statsmodels` library [8]. Options include the Bonferroni, Sidak, Holm, Holm-Sidak, Simes-Hochberg, or Hommel methods for FWER control, or the Benjamini-Hochberg, Benjamini-Yekutieli or Gavrilov-Benjamini-Sarkar methods for FDR control. The default option is BH for the Benjamini-Hochberg FDR. Users may optionally test only single-site features, only paired-site features or both, as well as only information, letter-heights, or both. We did not compute $p$-values for Information Differences of 0.

## 0.5 Calculation of Confidence Intervals for CIF Divergence $p$-Values

For ECDF-based $p$-value estimates $P_{\text{ECDF}}$ that attained $S = 10$ exceedances, we computed $100(1-\alpha)\%$ confidence intervals by a standard normal approximation for the standard error

**Algorithm 1:** APPROXIMATE $p$-value $\Pr(x \geq x_0)$ for KLD or ID CIF Divergence $x_0$

**Input:** $x_0, \{y_i^*\}_{i=1}^R \in \mathbb{R}^+; R, S, T, U \in \mathbb{Z}^+$, U $<$ T $<$ R, S $<$ R

**Output:** Approximation to $\Pr(x \geq x_0)$

$N \leftarrow 0; M \leftarrow 0;$
**while** $N < R$ **do**
  $N \leftarrow N + 1;$
  $w_N^* \leftarrow (y_N^*)^5;$
  **if** $y_N^* \geq x_0$ **then**
    $M \leftarrow M + 1;$
  **if** $M = S$ **then**
    **return** $P_{ECDF} \leftarrow M/N$
  **if** $N = T$ **then**
    $V \leftarrow \min U, N/3;$
    **repeat**
      $t \leftarrow (w^*_{(N-V-1)} + w^*_{(N-V)})/2;$
      **for** $i \leftarrow 1$ ***to*** $V$ **do**
        $z_i^* \leftarrow (w^*_{(N-V-1+i)} - t)$
      Estimate parameters $\hat{\xi}, \hat{\sigma}$ for $F_{\mathrm{GPD}}(z \mid \hat{\xi}, \hat{\sigma})$ from $\{z_i^*\}_{i=1}^V$ by MLE;
      Calculate $p_G$ for goodness-of-fit of $F_{\mathrm{GPD}}(z \mid \hat{\xi}, \hat{\sigma})$ to $\{z_i^*\}_{i=1}^V$ by Anderson-Darling Test;
      **if** $p_G < 0.05$ **then**
        $V \leftarrow V - 10;$
    **until** $V < 10 \vee p_G \geq 0.05;$
    **if** $p_G \geq 0.05$ **then**
      $P_{\mathrm{GPD}} \leftarrow (V/N)(1 - F_{\mathrm{GPD}}(((x_0)^5 - t) \mid \hat{\xi}, \hat{\sigma}));$
      **if** $P_{GPD} > 0$ **then**
        **return** $P_{GPD}$
    $T \leftarrow \min 2T, R$
**return** $P'_{ECDF} \leftarrow (M + 1)/(N + 1);$

of a binomial proportion $\sqrt{pq/n}$ with $q = 1 - p$. For ECDF-based $p$-value estimates $P'_{\text{ECDF}}$ that did not attain 10 exceedances, we did not compute confidence intervals. For GPD-based $p$-value estimates $P_{\text{GPD}}$, we computed $100(1 - \alpha)\%$ confidence intervals by the "boundary method" outlined in [9] and described in algorithm BOUNDARY. Given GPD shape and scale parameter estimates $\hat{\xi}, \hat{\sigma}$ and a set of peaks-over-threshold $\{z_i^*\}_{i=1}^{V}$ associated with CIF divergence $x_0$, we computed the observed Fisher Information Matrix in pure Python using expressions for partial derivatives of log-likelihood sums given in [10], and other matrix calculations in NumPy as detailed in algorithm BOUNDARY. The estimates of the GPD shape parameters $\hat{\xi}$ that we observed for our data were nearly always greater than negative one-half, which is technically required for the MLE asymptotics underlying both MLE estimation of the GPD parameters and their confidence intervals by the "boundary method" approach [9, 11]. However, one ID statistic for feature A1 from MAJOR clade against humans with moderate sample sizes (18 sequences in humans and 13 in MAJOR) resulted in a GPD fit with shape parameter $\xi = -0.58$ and a $p$-value for fit to GPD of about 9%. tSFM includes the values of the GPD parameter estimates in its output.

As shown in Fig. S1 we validated our Python-based GPD-estimation procedure, GPD-based $p$-value calculation, and confidence intervals for the GPD-based $p$-value calculation against the Multivariate Extreme Values Distribution package in R (`mev v1.13.1`) for 237 randomly chosen features from the ENRIETTII clade against humans and found good agreement.

## 0.6 Significance of CIF Divergences from [12] with Confidence Intervals

Supplementary Figs. S2 and S3 show $p$-value calculations with 95% confidence intervals for KLD and (respectively) ID CIF divergences between humans and two clades of trypanosomes derived from the data of [12].

**Algorithm 2:** BOUNDARY Confidence Interval for $P_{\text{GPD}}$

**Input:** $x_0, \{z_i^*\}_{i=1}^{V} \in \mathbb{R}^+, V, N \in \mathbb{Z}^+$ with $V < N$, $\hat{\xi}, \hat{\sigma} \in \mathbb{R}$, $\alpha \in [0, 0.5]$

**Output:** Upper and Lower $(1-\alpha)\%$ Confidence Limits $\{\hat{L}, \hat{U}\}$ of $P_{\text{GPD}}$

Calculate observed Fisher Information Matrix $F$ from $\{z_i^*\}_{i=1}^{V}$ and density $f_{\text{GPD}}(z \mid \hat{\xi}, \hat{\sigma})$;

Calculate Moore-Penrose Generalized Inverse $F^{-1}$ from $F$;

Compute Singular Value Decomposition $WDW^T$ of $F^{-1}$, with $D = \text{diag}\,(d_1, d_2)$;

Compute $C_{\sqrt{\alpha}}$, the $100(1-\sqrt{\alpha}/2)\%$ quantile of the standard normal distribution;

$\xi_{0,1} \leftarrow \hat{\xi} - C_{\sqrt{\alpha}}\sqrt{d_1}$; $\sigma_{0,1} \leftarrow \hat{\sigma} - C_{\sqrt{\alpha}}\sqrt{d_2}$;

$\xi_{0,2} \leftarrow \hat{\xi} + C_{\sqrt{\alpha}}\sqrt{d_1}$; $\sigma_{0,2} \leftarrow \hat{\sigma} + C_{\sqrt{\alpha}}\sqrt{d_2}$;

$\begin{pmatrix} \xi_1 \\ \sigma_1 \end{pmatrix} \leftarrow W \begin{pmatrix} \xi_{0,1} - \hat{\xi} \\ \sigma_{0,1} - \hat{\sigma} \end{pmatrix} + \begin{pmatrix} \hat{\xi} \\ \hat{\sigma} \end{pmatrix}$;

$\begin{pmatrix} \xi_2 \\ \sigma_2 \end{pmatrix} \leftarrow W \begin{pmatrix} \xi_{0,2} - \hat{\xi} \\ \sigma_{0,2} - \hat{\sigma} \end{pmatrix} + \begin{pmatrix} \hat{\xi} \\ \hat{\sigma} \end{pmatrix}$;

$P_{11} \leftarrow (1 - F_{\text{GPD}}(((x_0)^5 - t) \mid \xi_1, \sigma_1))$;

$P_{12} \leftarrow (1 - F_{\text{GPD}}(((x_0)^5 - t) \mid \xi_1, \sigma_2))$;

$P_{21} \leftarrow (1 - F_{\text{GPD}}(((x_0)^5 - t) \mid \xi_2, \sigma_1))$;

$P_{22} \leftarrow (1 - F_{\text{GPD}}(((x_0)^5 - t) \mid \xi_2, \sigma_2))$;

$\hat{L_r} \leftarrow \min P_{11}, P_{12}, P_{21}, P_{22}$;

$\hat{U_r} \leftarrow \max P_{11}, P_{12}, P_{21}, P_{22}$;

Compute $C_\alpha$, the $100(1-\alpha/2)\%$ quantile of the standard normal distribution;

$\hat{L_{nr}} \leftarrow (V/N) - C_\alpha\sqrt{(V/N)(1-(V/N))/N}$;

$\hat{U_{nr}} \leftarrow (V/N) + C_\alpha\sqrt{(V/N)(1-(V/N))/N}$;

$\hat{L} \leftarrow \hat{L_r}\hat{L_{nr}}$;

$\hat{U} \leftarrow \hat{U_r}\hat{U_{nr}}$;

**return** $\{\hat{L}, \hat{U}\}$;

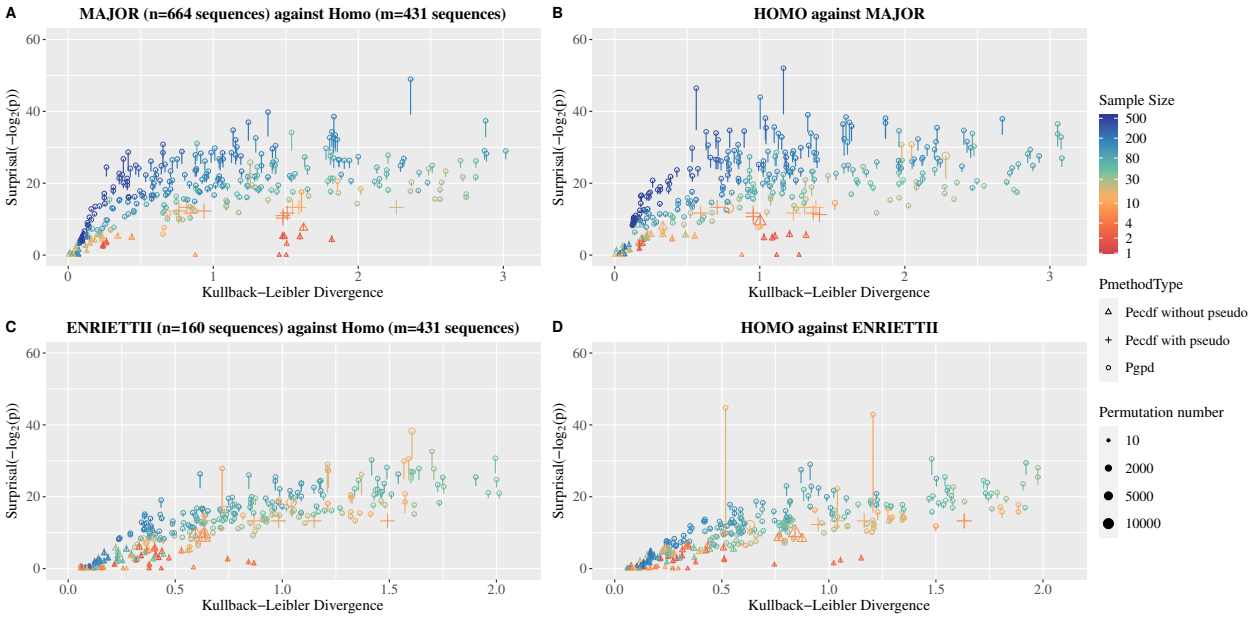


Figure S2: Permutation-based surprisals of KLD values of shared tRNA features for two different clades of *Leishmania* against humans as a function of magnitude of KLD signal. The MAJOR clade pools data from nine genomes of *Leishmania* and thus contains more data than the ENRIETTII clade, which pools data from only two genomes. $P$-values are calculated by algorithm APPROXIMATE. Confidence intervals for $P_{\text{GPD}}$ are calculated by algorithm BOUNDARY. Confidence intervals for $P_{\text{ECDF}}$ are based on the standard error for a binomial proportion. The $x$-axes shows KLD signals of features measured in bits and $y$-axes show $-\log_2$ of the permutation $p$-value of that signal. Colors represent the harmonic mean of conditional sample sizes of sequences carrying a feature in the two clades. A) KLD for MAJOR clade against humans. B) KLD for humans against MAJOR clade. C) KLD for ENRIETTII clade against humans. D) KLD for humans against ENRIETTII clade.

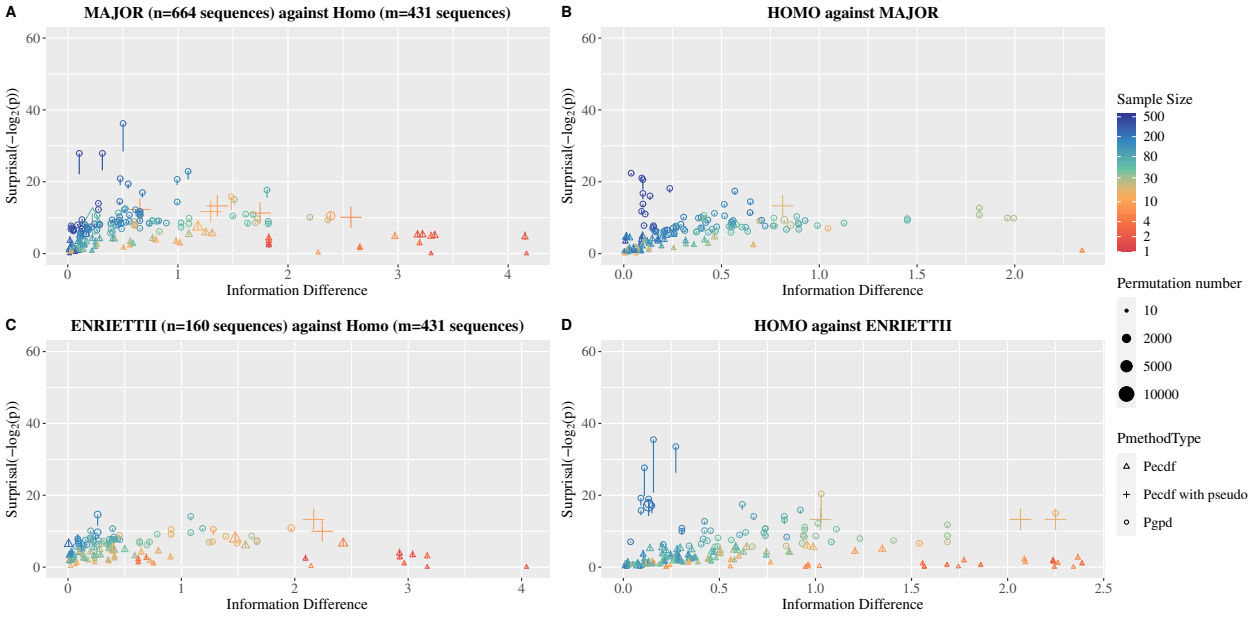


Figure S3: Permutation-based surprisals of ID values of shared tRNA features for two different clades of *Leishmania* against humans as a function of magnitude of KLD signal. $P$-values are calculated by algorithm APPROXIMATE. Confidence intervals for $P_{\text{GPD}}$ are calculated by algorithm BOUNDARY. Confidence intervals for $P_{\text{ECDF}}$ are based on the standard error for a binomial proportion. The $x$-axes shows KLD signals of features measured in bits and $y$-axes show $-\log_2$ of the permutation $p$-value of that signal. Colors represent the harmonic mean of conditional sample sizes of sequences carrying a feature in the two clades. A) ID for MAJOR clade against humans. B) ID for humans against MAJOR clade. C) ID for ENRIETTII clade against humans. D) ID for humans against ENRIETTII clade.

# Bibliography


[1] M Sprinzl, C Horn, M Brown, A Ioudovitch, and S Steinberg. "Compilation of tRNA sequences and sequences of tRNA genes." In: *Nucleic Acids Res.* 26.1 (Jan. 1998), pp. 148–53. ISSN: 0305-1048 (page 1).

[2] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. "Information content of binding sites on nucleotide sequences". In: *J. Mol. Biol.* 188.3 (Apr. 1986), pp. 415–431. ISSN: 0022-2836. DOI: 10.1016/0022-2836(86)90165-8 (page 2).

[3] G A Miller. "Note on the bias of information estimates." In: *Information Theory in Psychology II-B*. Glencoe, IL: Free Press, 1955, pp. 95–100 (page 2).

[4] Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. "Entropy and information in neural spike trains: Progress on the sampling problem". In: *Physical Review E* 69.5 (May 2004), p. 056111. ISSN: 1539-3755. DOI: 10.1103/PhysRevE.69.056111. URL: http://www.ncbi.nlm.nih.gov/pubmed/15244887%20https://link.aps.org/doi/10.1103/PhysRevE.69.056111 (page 2).

[5] J. Gorodkin, L.J. Heyer, S. Brunak, and G.D. Storomo. "Displaying the information contents of structural RNA alignments: the structure logos". In: *Bioinformatics* 13.6 (Dec. 1997), pp. 583–586. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/13.6.583 (page 2).

[6] Eva Freyhult, Vincent Moulton, and David H Ardell. "Visualizing bacterial tRNA identity determinants and antideterminants using function logos and inverse function logos." In: *Nucleic Acids Res.* 34.3 (Jan. 2006), pp. 905–16. ISSN: 1362-4962. DOI: 10.1093/nar/gkj478 (page 2).

[7] Eva Freyhult, Yuanyuan Cui, Olle Nilsson, and David Ardell. "New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria." In: *Biochimie* 89.3 (Oct. 2007), pp. 1276–1288. ISSN: 0300-9084. DOI: 10.1016/j.biochi.2007.07.013. URL: http://www.sciencedirect.com/science/article/pii/S0300908407001897 (page 2).

[8] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010 (page 3).

[9] Dylan Glotzer, Vladas Pipiras, Vadim Belenky, Bradley Campbell, and Timothy Smith. "Confidence intervals for exceedance probabilities with application to extreme ship motions". In: *REVSTAT Statistical Journal* 15.4 (2017), pp. 537–563 (page 5).

[10] Bradley Campbell, Vadim Belenky, and Vladas Pipiras. "Application of the envelope peaks over threshold (EPOT) method for probabilistic assessment of dynamic stability". In: *Ocean Engineering* 120 (2016), pp. 298–304. ISSN: 0029-8018. DOI: https://doi.

org/10.1016/j.oceaneng.2016.03.006. URL: http://www.sciencedirect.com/science/article/pii/S002980181600113X (page 5).

[11] Theo A. Knijnenburg, Lodewyk F. A. Wessels, Marcel J. T. Reinders, and Ilya Shmulevich. "Fewer permutations, more accurate P-values". In: *Bioinformatics* 25.12 (May 2009), pp. i161–i168. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp211. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/12/i161/16887761/btp211.pdf. URL: https://doi.org/10.1093/bioinformatics/btp211 (page 5).

[12] Paul Kelly, Fatemeh Hadi-Nezhad, Dennis Y. Liu, Travis J. Lawrence, Roger G. Linington, Michael Ibba, and David H. Ardell. "Targeting tRNA-synthetase interactions towards novel therapeutic discovery against eukaryotic pathogens". In: *PLOS Neglect. Trop. D.* 14.2 (Feb. 2020), pp. 1–30. DOI: 10.1371/journal.pntd.0007983. URL: https://doi.org/10.1371/journal.pntd.0007983 (page 5).