Supplementary Material 1

Process for Shannon entropy calculation

For a column $\boldsymbol{C_n}$ in the database matrix $\boldsymbol{M}$:

1) For every row, check the nucleotide or gap that occurs in the row.
2) Increment corresponding sum of the relevant nucleotide or gap.
3) When all rows are iterated over, store the sum of each nucleotide and gap.

For every column in the aligned database:

1) The probability of each nucleotide or gap was calculated by dividing the number of occurrences of the specific nucleotide or gap over total number of occurrences of all nucleotides and gaps as shown below:

   $P(n_i)$ = Probability of nucleotide n_i = (number of n_i) / (total number of all nucleotides and gaps)

2) The probability $P(n_i)$ generated for each nucleotide was then multiplied with its natural log, $Ln\ (P(n_i))$.

3) Shannon entropy of a column was then calculated as the sum total of Shannon entropy of every nucleotide and gap by using the following formula.

$$-\sum_{1}^{i} p(n_{\;i}) \, ln \, p(n_i)$$

Process for generating reference sequence Shannon entropy

As every row $\boldsymbol{R_m}$ represented a reference sequence in the matrix $\boldsymbol{M}$, every column was iterated over to determine entropy value at the specific location:

1) A per reference Shannon entropy vector was generated by storing entropy value of a column $\boldsymbol{C_n}$ if a nucleotide was present at this location.
2) Gaps were ignored in this process, as they do not form part of the sequence.

Process for determining Shannon entropy of query read I_i

Custom vector, V_j

S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	…	S_n

Reference Sequence, R_j

Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	…	Z_n

Input Read, I_i

D_1	D_2	D_3	D_4	…	D_m

S_n: Entropy value at location n

N_n: Reference sequence nucleotide N at location n

D_m: Input read nucleotide D at location m

For an query read $\boldsymbol{I_i}$ having length $\boldsymbol{m}$ that aligned to a reference sequence $\boldsymbol{R_j}$:

1) The location of matches between reference sequence and query read were found.

2) For every match, the corresponding Shannon entropy value was taken from the database using the location of the match on the reference sequence.
    a. For example, if nucleotide $\boldsymbol{D_3}$ on query read matches nucleotide $\boldsymbol{Z_7}$ on reference sequence, the corresponding Shannon entropy value is $\boldsymbol{S_7}$.

Formulae for relative Shannon entropy metric:

$$Relative\ SE\ Score = {SE_{Read}}\Big/{SE_{Ref}}$$

$$Read\ Score = {SE_{Read}}\Big/{SE_{Ref}} * \frac{SE\ Coverage}{Alignment\ Length}$$

Sugarcane dataset sampling

Sugarcane leaf, stalk, root and rhizosphere soil samples were collected by Dr. Kelly Hamonts at Hawkesbury Institute for the Environment, Western Sydney University, Australia, in November 2014 from eight sugarcane fields growing three sugarcane varieties (KQ228, MQ239 and Q240) near Ingham, Queensland, Australia. In each field, 3 stools were randomly selected and samples were collected from 2 plants per stool. Samples were snap-frozen in liquid nitrogen on the field, transported to the laboratory on dry ice and stored at -80C. Frozen sugarcane tissue samples were ground using mortar and pestle and DNA was extracted from the resulting powder using the MoBio PowerPlant DNA extraction kit, following the manufacturer's instructions. The MoBIO PowerSoil DNA extraction kit was used to extract DNA from the soil samples. Bacterial 16S rRNA amplicon sequencing was performed by the NGS facility at Western Sydney University using Illumina Miseq (2x 301 bp PE) and the 341F/805R primer set.