Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition: supplementary material

Supplementary Methods 1

The procedure for computing OFDEG [1] is given as follows: given an input sequence – S – with total length L, the reference n-mer profile $T_n(S)$ is defined as the total count of all n-mers in S, where n is a user-defined basis. The n-mer profiles are computed for randomly selected subsamples $T_n(s_i)$ of length l_i, $\{l_i \in \mathbb{Z}\colon n \le l_i \le L\}$. From this it is clear that when $l_i = L$ the n-mer profiles are equivalent, $T_n(S) \equiv T_n(s_i)$, and $\|T_n(S) - T_n(s_i)\| = 0$. These subsample profiles are then compared with the reference profile using a Euclidean distance metric, and the geometric mean of all samples is calculated by

$$E_i = \frac{1}{N_s}\sum_{i=1}^{N_s} \|T_n(S) - T_n(s_i)\| \tag{1}$$

where N_s is the total number of samples. Regressing the E_i terms over the set of l_i, $\forall i \in 1 \ldots G$, where G is the number of discrete subsample lengths, gives the linear trend of the *compositional error* over subsample length:

$$\text{OFDEG} = \frac{\sum_i^G l_i E_i - \frac{1}{G}\sum_i^G l_i \sum_i^G E_i}{\sum_i^G l_i^2 - \frac{1}{G}\left(\sum_i^G E_i\right)^2} \tag{2}$$

The quantity G can be defined in terms of step-size between subsequent subsample lengths and N_s can be defined in terms of the fold-coverage of the original sequence. In this study $G = 0.1 \times L$ and $N_s = 3 \times L$. Further work on OFDEG has shown its validity and given an insight into its theoretical underpinnings (Maheswararajah S, Saeed I and Halgamuge SK, *Internal Report*, 2011)

Supplementary Methods 2

Dinculeotide odds ratio

The dinucleotide odds ratio [2] is given by:

$$\rho_i^2 = \frac{f(n_1 n_2)}{f(n_1)f(n_2)}, \tag{3}$$

where $f(n_1n_2)$ is the frequency of the dinucleotide n_1n_2, and $f(n_1)$ and $f(n_2)$ are the mononucleotide frequencies of nucleotides n_1 and n_2.

Tetranucleotide frequency

Tetranucleotide frequency is the most widely adopted feature for clustering metagenomic sequences. The maximal-order Markov normalisation [3] for a given sequence is given by:

$$\rho_i^4 = \frac{f(n_1 n_2 n_3 n_4) f(n_2 n_3)}{f(n_1 n_2 n_3) f(n_2 n_3 n_4)}. \tag{4}$$

Z-score normalisation

Given the maximal order Markov model from the dinucleotide and trinucleotide frequency components, the z-score transform is used to assess the statistical significance of each tetramer [4]. The mean and variance used in the normalisation are calculated as follows. The expected value for a given tetramer is calculated as:

$$E\left(n_1 n_2 n_3 n_4\right) = \frac{N\left(n_1 n_2 n_3\right) N\left(n_2 n_3 n_4\right)}{N\left(n_1 n_2\right)}, \tag{5}$$

and the variance is calculated using:

$$\sigma^2\left(n_1 n_2 n_3 n_4\right) = E\left(n_1 n_2 n_3 n_4\right) \times \frac{\left[N\left(n_2 n_3\right) - N\left(n_1 n_2 n_3\right)\right]\left[N\left(n_2 n_3\right) - N\left(n_2 n_3 n_4\right)\right]}{N\left(n_2 n_3\right)^2}, \tag{6}$$

which gives:

$$Z\left(n_1 n_2 n_3 n_4\right) = \frac{N\left(n_1 n_2 n_3 n_4\right) - E\left(n_1 n_2 n_3 n_4\right)}{\sqrt{\sigma^2\left(n_1 n_2 n_3 n_4\right)}} \tag{7}$$

Supplementary Figures

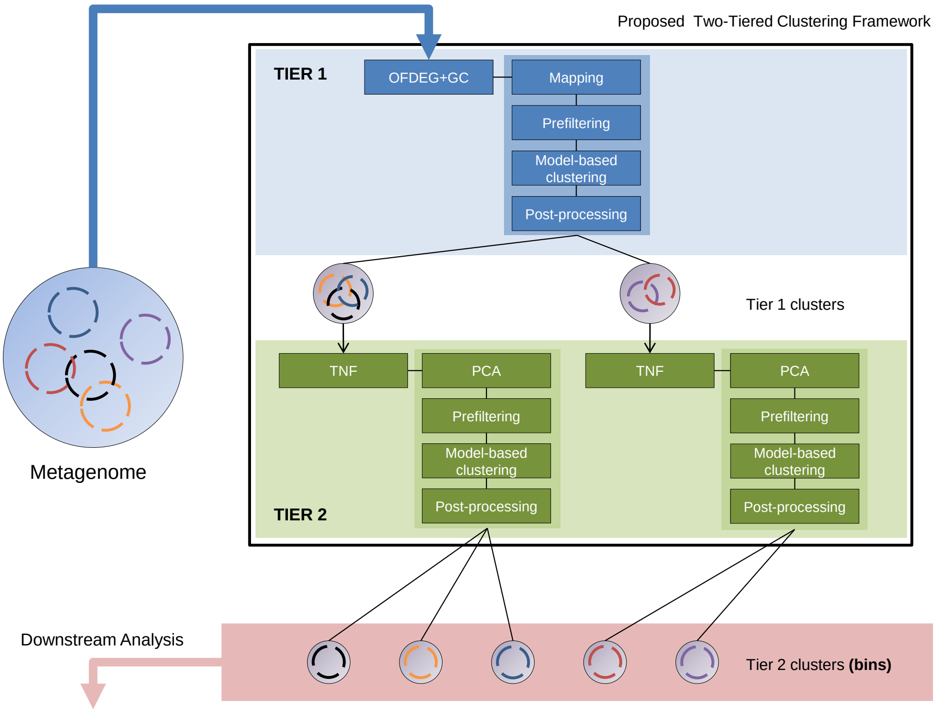

Figure 1: Block diagram of the proposed framework

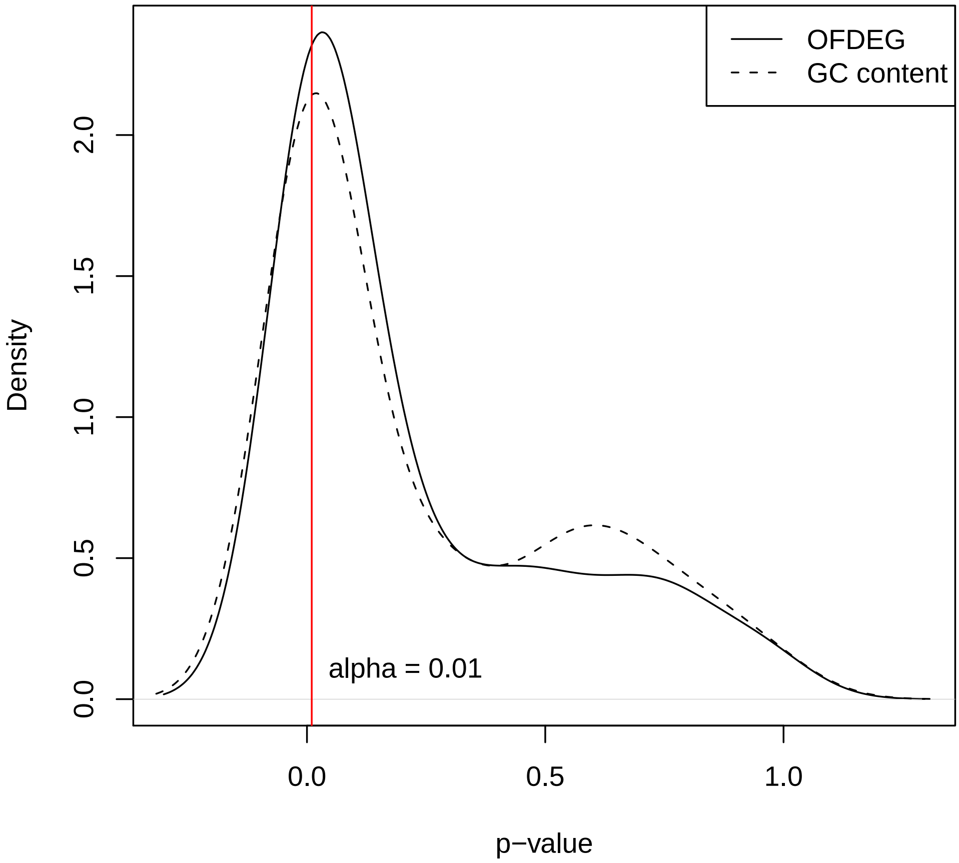

Figure 2: The resulting p-values of the Shapiro-Wilk test for normality. The p-values were calculated for each of the 124 genomes used for evaluation, where randomly selected 5 kbp fragments were used to model a microbial population. Each genome was tested separately, and it was found that the assumption of normality holds at a significance level of 0.01.

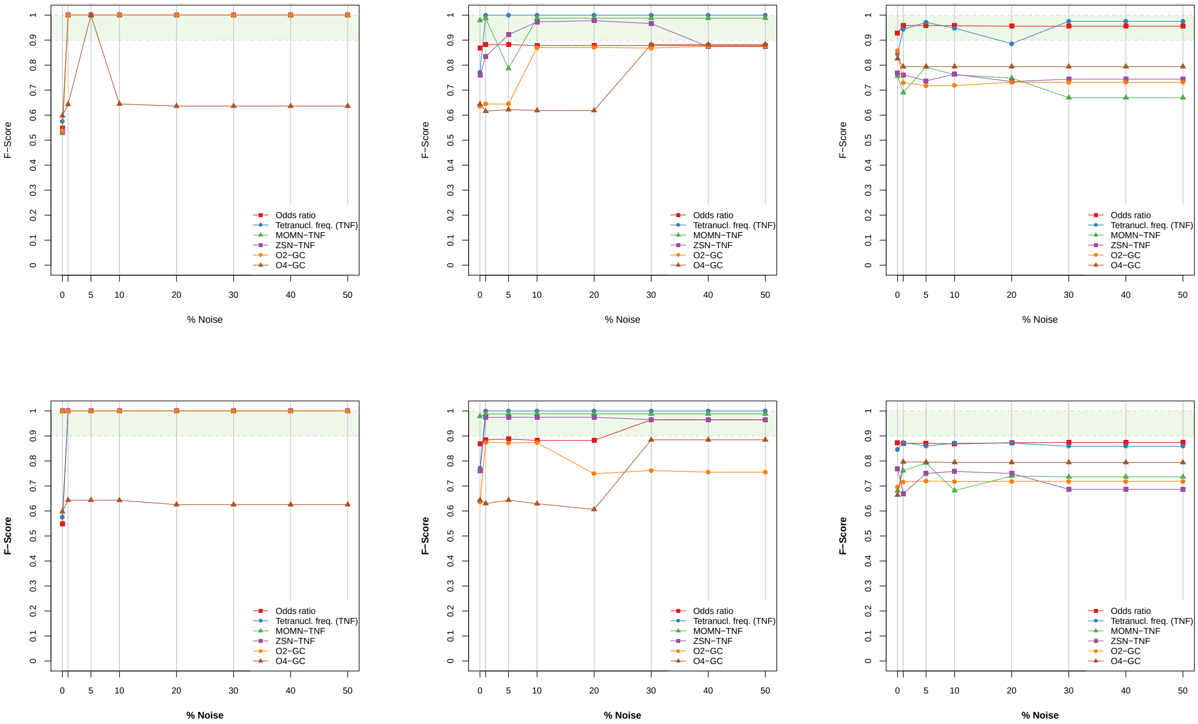

Figure 3: Parameter selection for the NNVE method. The columns (left-right) of images correspond to the simLC, simMC and sim-BG data sets, while the rows (top-bottom) correspond to a neighbourhood size of 5, 10 and 15. It was found that a neighbourhood size of 5 and an initial noise estimate of 0.01, unless otherwise specified, produces adequate results for the metagenomes used in this study.

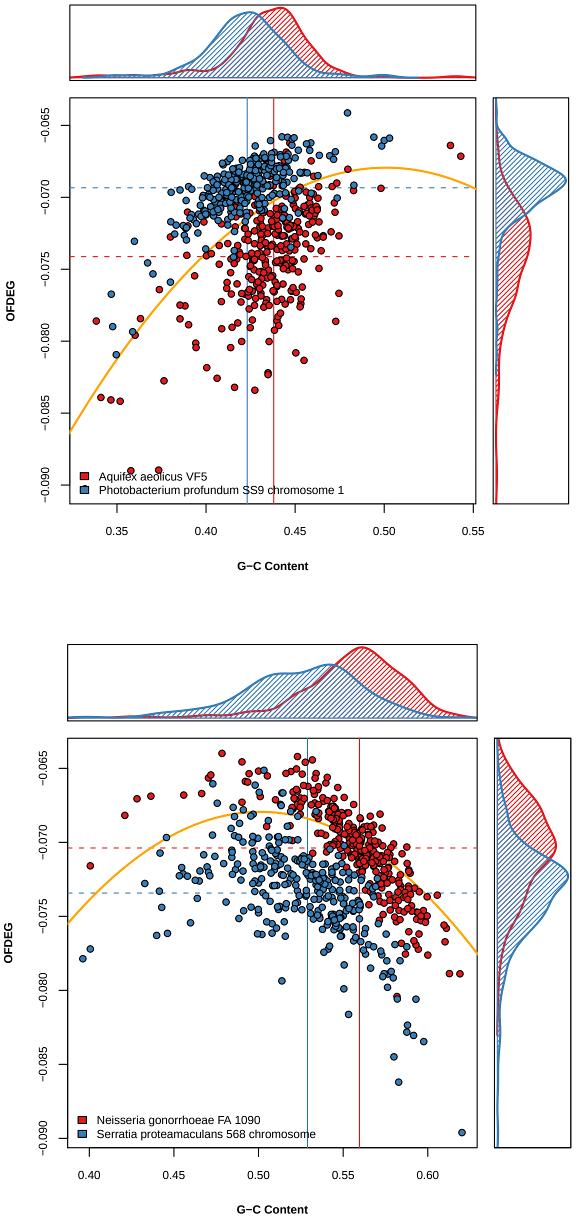

Figure 4: The joint distribution of OFDEG and GC content clusters sequences of distinct organisms that otherwise cannot be distinguished by each feature alone. Shown here are the joint distributions for randomly sampled genomic fragments of: (TOP) *Aquifex aeolicus* VF5, and the *Photobacterium profundum* SS9 chromosome 1; and (BOTTOM) *Neisseria gonorrhoeae* FA 1090, and *Serratia proteamaculans* 568. Randomly sampled fragments of the *Aquifex aeolicus* VF5 genome (GC= 43%) and the *Photobacterium profundum* SS9 (GC= 41%) cannot be accurately separated based on marginal GC content alone, but can be distinguished when used in conjunction with OFDEG. With reference to the same figure, fragments of the *Neisseria gonorrhoeae* FA 1090 genome (GC= 52%) and the *Serratia proteamaculans* 568 chromosome (GC= 55%) are also shown to be separable using both features despite significant overlap in the marginal distributions

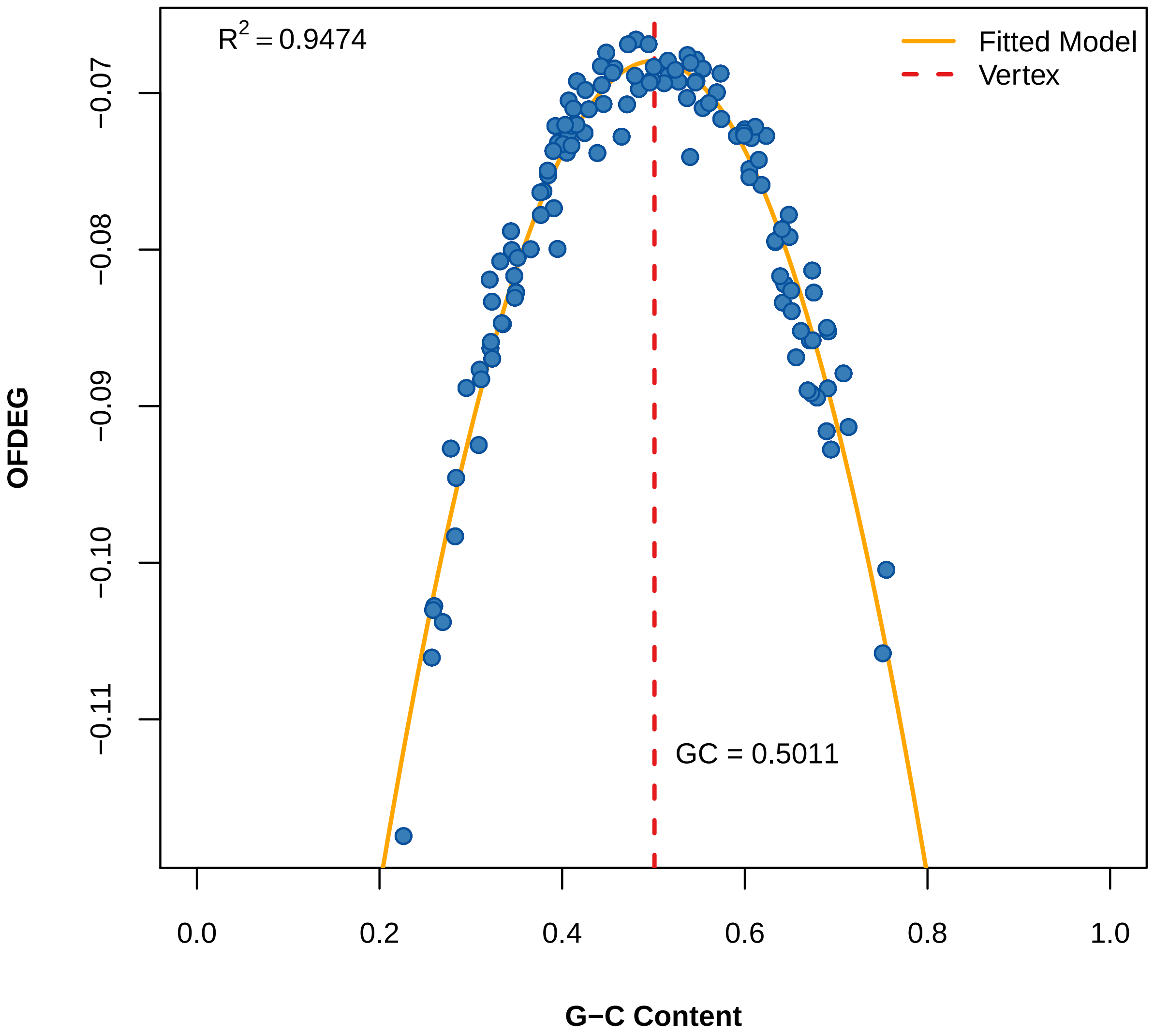

Figure 5: The distribution of OFDEG and GC content can be represented by a one-dimensional manifold. The distribution shown here was constructed using samples of 124 full genomes representative of a diverse range of fully sequenced prokaryotic genomes.

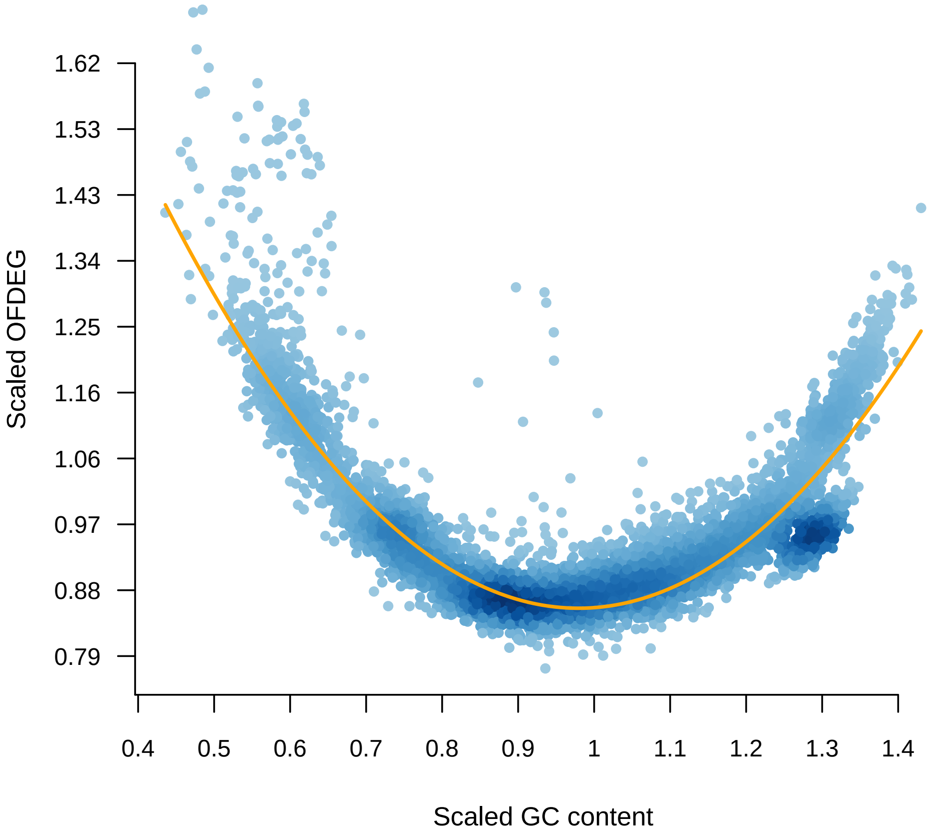

Figure 6: The raw distribution of the sim-BG sequences in the OFDEG and GC content space. The orange line represents the precomputed principal curve; note that the absolute value of OFDEG is shown here.

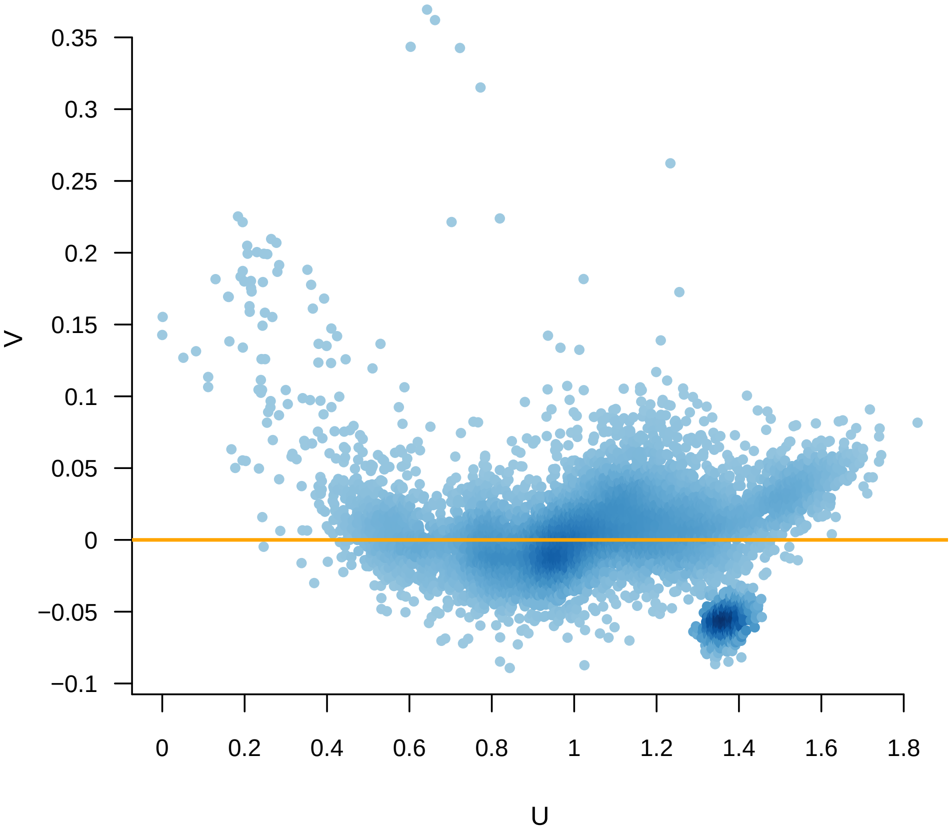

Figure 7: The scaled and mapped sequences of the sim-BG data set from the raw OFDEG and GC content space to the projection onto the precomputed prinicpal curve.

Supplementary Data 1

The variation in the V1 and V2 regions of the 16S rRNA sequences were analysed using the Ribosmal Database Project classifier [6]. The remaining unassigned sequences were subsequently classified using GF (Tseng et al., BMC Bioinf., *manuscript submitted*). The results of the 16S rRNA diversity analysis are shown in Figure 8.

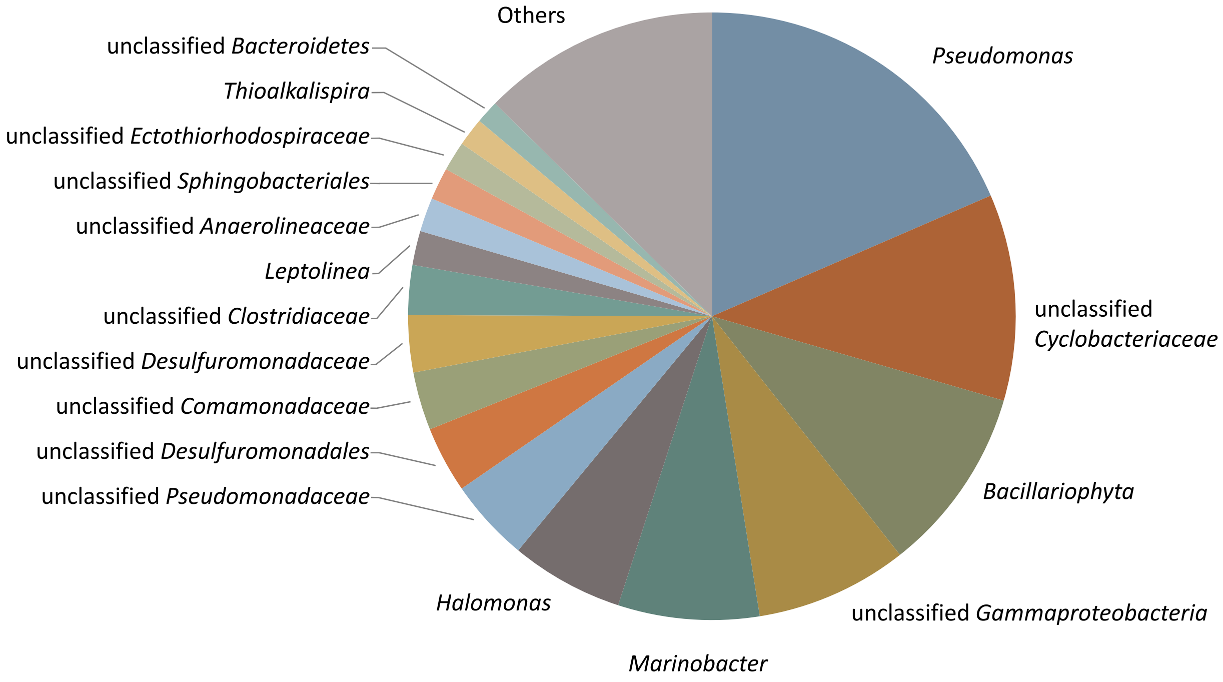

Figure 8: The 16S rRNA diversity analysis of the novel mud volcano metagenome.

Supplementary Data 2

The KEGG analysis of each bin reveals highly accurate predictions of the expected metabolic potential of the mud volcano sample (Fig. 9).

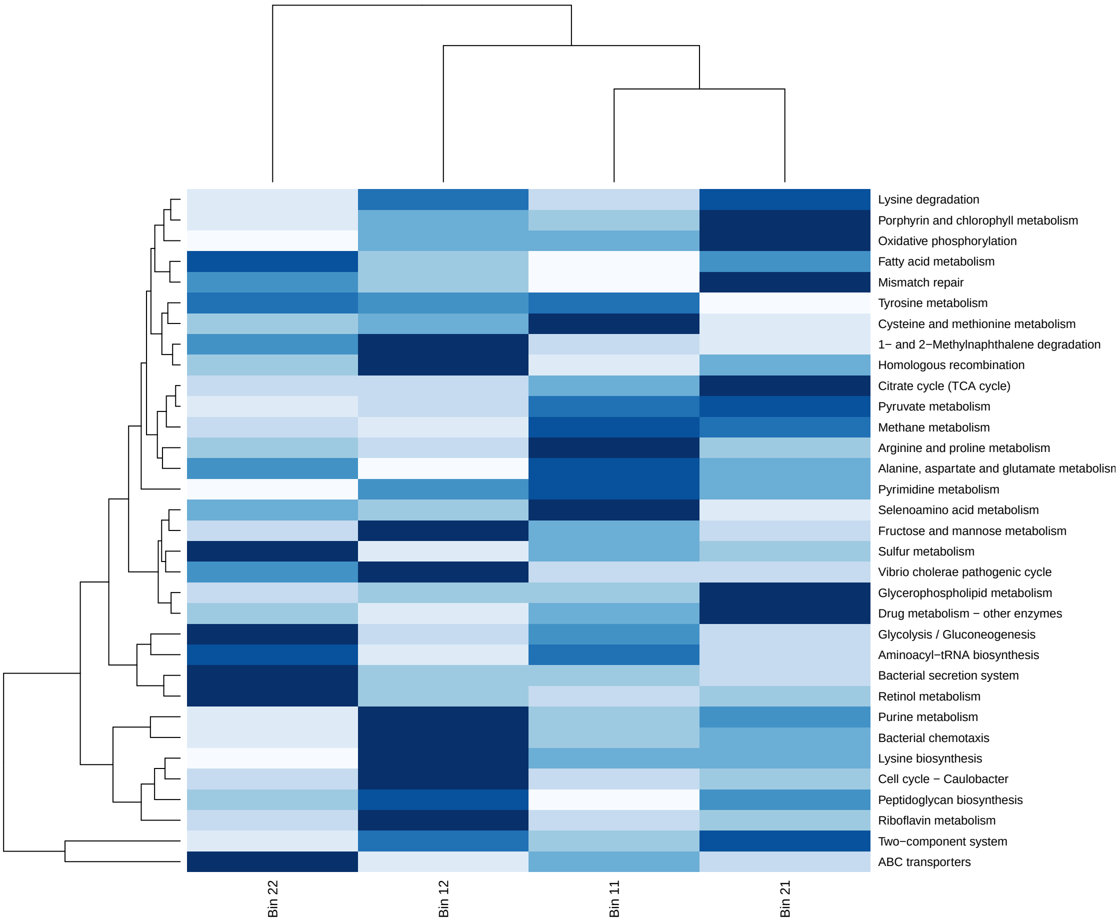

Figure 9: Results of the KEGG analysis. The intensity represents the number of genes assigned to a KEGG category (using MEGAN), where the total number of assignments for each profile is normalised over each bin.

Bibliography

[1] I. Saeed and S.K. Halgamuge. The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics*, 10(S3):S10, 2009.

[2] S. Karlin, J. Mrazek, and A.M. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, 179(12):3899–3913, 1997.

[3] J. Mrazek. Phylogenetic signals in dna composition: Limitations and prospects. *Mol. Biol. Evol.*, 26(5):1163–1169, 2009.

[4] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glockner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, 6(9):938–47, 2004.

[5] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinf.*, 5(163), 2004.

[6] J.R. Cole, Q. Wang, E. Cardenas, J. Fish, B. Chai, J. Farris, A.S. Kulam-Syed-Mohideen, D.M. McGarrell, T. Marsh, G.M. Garrity, and J.M. Tiedje. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, 37(Database issue):141–5, 2008.