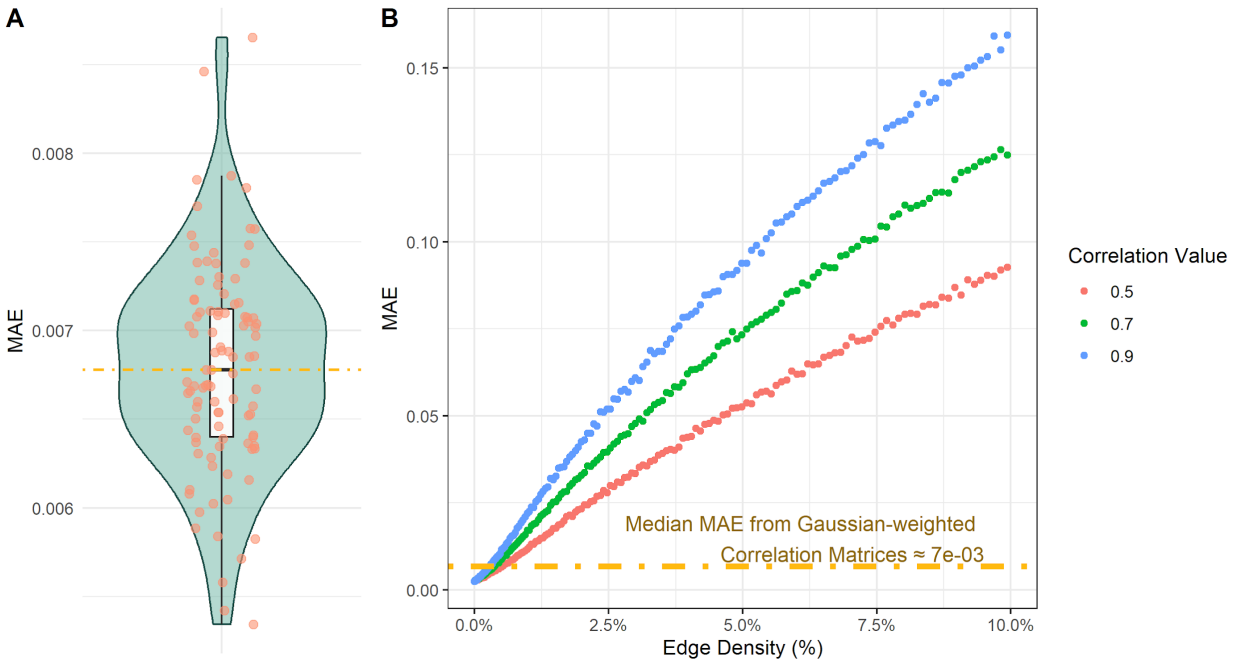
Correlation measures in metagenomic data: the blessing of dimensionality. Supplementary Information

Alessandro Fuschi[1], Thi Dong Binh Tran[2], Hoan Nguyen[2], George M. Weinstock[2,3], Daniel Remondini[1,*] , Alessandra Merlotti[1]

*Corresponding author

1. Department of Physics and Astronomy, University of Bologna, Bologna 40127, IT.
2. The Jackson Laboratory for Genomic Medicine, Farmington, CT 06030 USA
3. Dept. Genetics and Genome Science, University of Connecticut Health Center, Farmington, CT 06032 USA

Supplementary Section S1: effect of dense correlation



Impact of Correlation Density on Bias in Correlation Estimates Using CLR Normalization Across Varying Densities of Artificially Imposed Correlations. A) MAE distribution from 100 unique randomly generated datasets, where correlation weights in each dataset of 500 dimensions and 10,000 samples are Gaussian-distributed via `mvtnorm`, mimicking realistic data scenarios. The distribution showcases that the median MAE is on the order of $\cong 10^{-3}$, indicating negligible biases when CLR normalization is applied to realistic correlation structures. B) Effects of correlation under an extreme scenario, in which all nonzero correlation matrix values are identical and equal to 0.5, 0.7 or 0.9 respectively. The plot shows MAE as a function of an increasing number of completely correlated species in the matrix, ranging from 0% to 10% of all possible matrix elements (corresponding to the link density of a network constructed by such correlation matrix imposing a threshold over non-zero values). Even in this unrealistic scenario, when link density reaches up to 1-2% for fully connected blocks of taxa (close to the bottom left of the plot), the observed MAE in correlation estimates remains below 0.05.

Supplementary Section S2: pseudocode for ZINB data generation

```
// Fit ZINB Model parameters using OTUs from HMP2
params_ZINB_HMP2=fitZINBParameters(OTUs_HMP2);

// Perform 100 iterations of simulation
for (iteration in 1:100) {

   // generate ZINB random parameters using the ecdf of the real
   // distributions of the ZINB parameters
   random_params_ZINB=randomZINBParameters(params_ZINB_HMP2);

   // Generate synthetic dataset with D=200
   syntheticData = generateSyntheticDataset(D=200,
par=random_params_ZINB);

   // Loop over varying levels of sparsity (phi) and correlation (r)
   for (phi in seq(0, 0.95, by=0.025)) {
      for (r in seq(-0.9, 0.9, by=0.05)) {

         // Modify variables I and J in the dataset
         modifyVariables(syntheticData, I, J, phi, r);

         // Record error for current sparsity and correlation
         err_phi_r = recordError(syntheticData, I, J);
      }
   }
}
```

Pseudo-code illustrating the methodology for quantifying correlation errors across different sparsity levels (Φ) and pairwise correlations (r) between two taxa in synthetic datasets. The model parameters are fitted using OTUs from HMP2 and randomly generated for each iteration to ensure distribution independence. The process involves 100 iterations of synthetic data generation, modification, and error recording, followed by the calculation of the mean absolute error (MAE) for each Φ and r combination.