Supplementary Figure 4. Impact of k-mer size and of the solid filter, as measured by the correlation with a taxonomic distance. The impact of two important parameters of Simka was measured on the 138 gut samples from the HMP project: the size of the k-mers k and the solid filter. Remember that the solid filter discards k-mers seen exactly once in a sample. The correlation is pretty robust to the choice of k as long as k is larger than 15, with an optimum correlation for $k = 21$. Remarkably, the solid filter has almost no impact on the correlation. As shown in Table 2 of the main manuscript, even if the solid filter discards 85% of the distinct k-mers, solid k-mers still represent more than 95% of the total base pairs in the HMP dataset, which explains the lack of difference.