Supplementary Material 3 - Influence of varying e-value and % query identity thresholds on raw hit counts

(a) The GOS dataset (longer, Sanger sequence reads) resulted in more robust hits, with a less severe drop in hit counts as e-value and percent identity were changed in the less strict ranges. The GOS cut-off was set at e-values less than 10^{-4} and sequence identity greater than 20%. (b) The viromes dataset (short, average <100 bp, pyrosequencing reads) resulted in poor similarity results, with the number of hits at higher, less stringent e-values dropping off drastically. Thresholds were chosen to minimize these likely false-positive hits at e-values less than 10^{-4} and sequence identity greater than 10%.