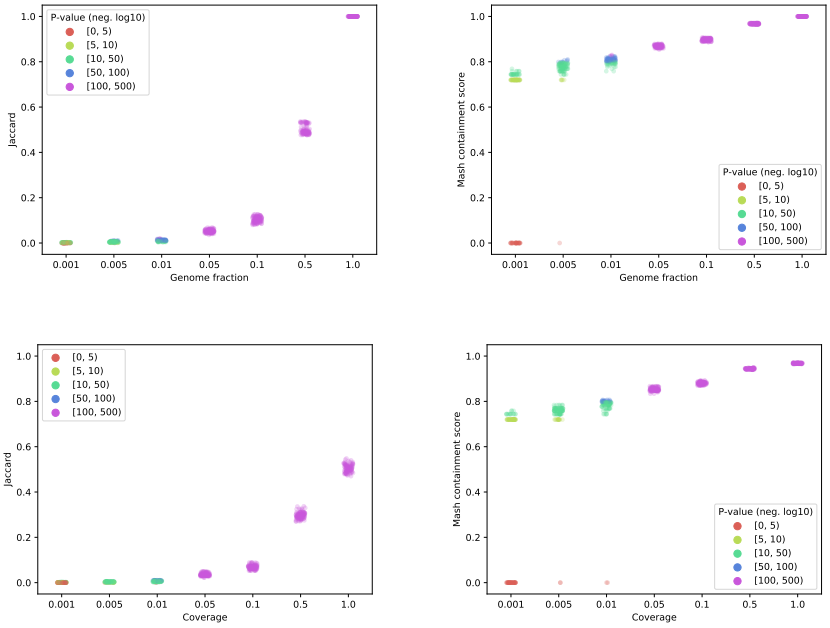

Figure S1 Relation of Mash containment scores to genome coverage. Experiments were performed by screening samples of an *E. coli* genome and screen against the complete reference genome. In the top charts ("Genome fraction" on x-axis), contiguous substrings of the genome were used, with lengths corresponding to the indicated fractions of the complete genome length. In the bottom charts ("Coverage" on x-axis), the indicated value was provided as the target coverage for simulating reads, which were then given to Mash Screen. The raw Jaccard scores (left plots) correspond closely to the actual fraction taken for contiguous subsequences, or to roughly half the read coverage (due to sequencing errors). The Mash containment score (right plots), however, is closer to the true identity than the true fraction contained because it models k-mer survival through mutational divergence rather than low coverage or structural change. Reads were simulated with ART (version "Mount Rainier"), using `-l 100` (100-base read length) and `-ss HS20` (Illumina HiSeq 2000 platform profile).