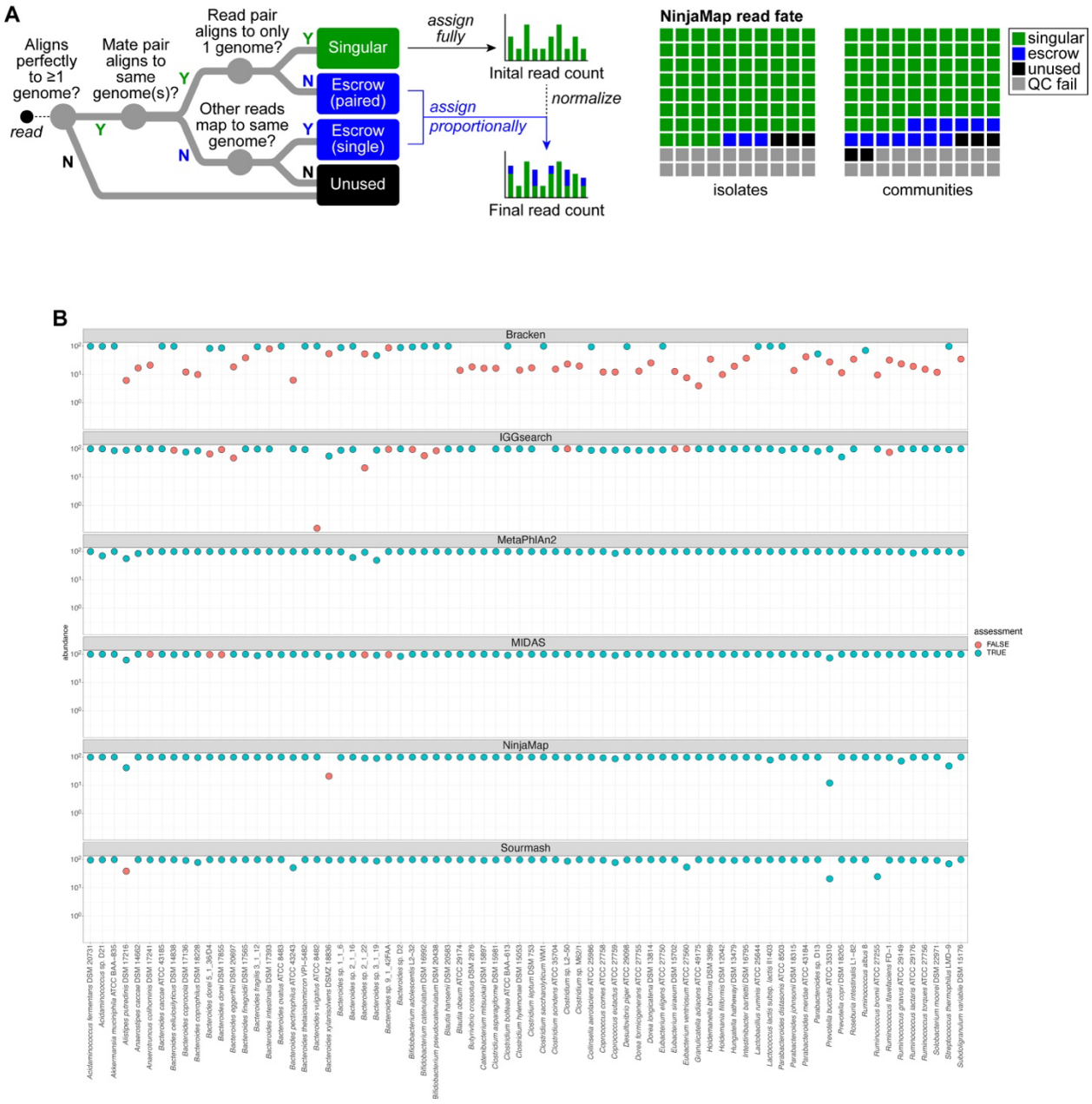
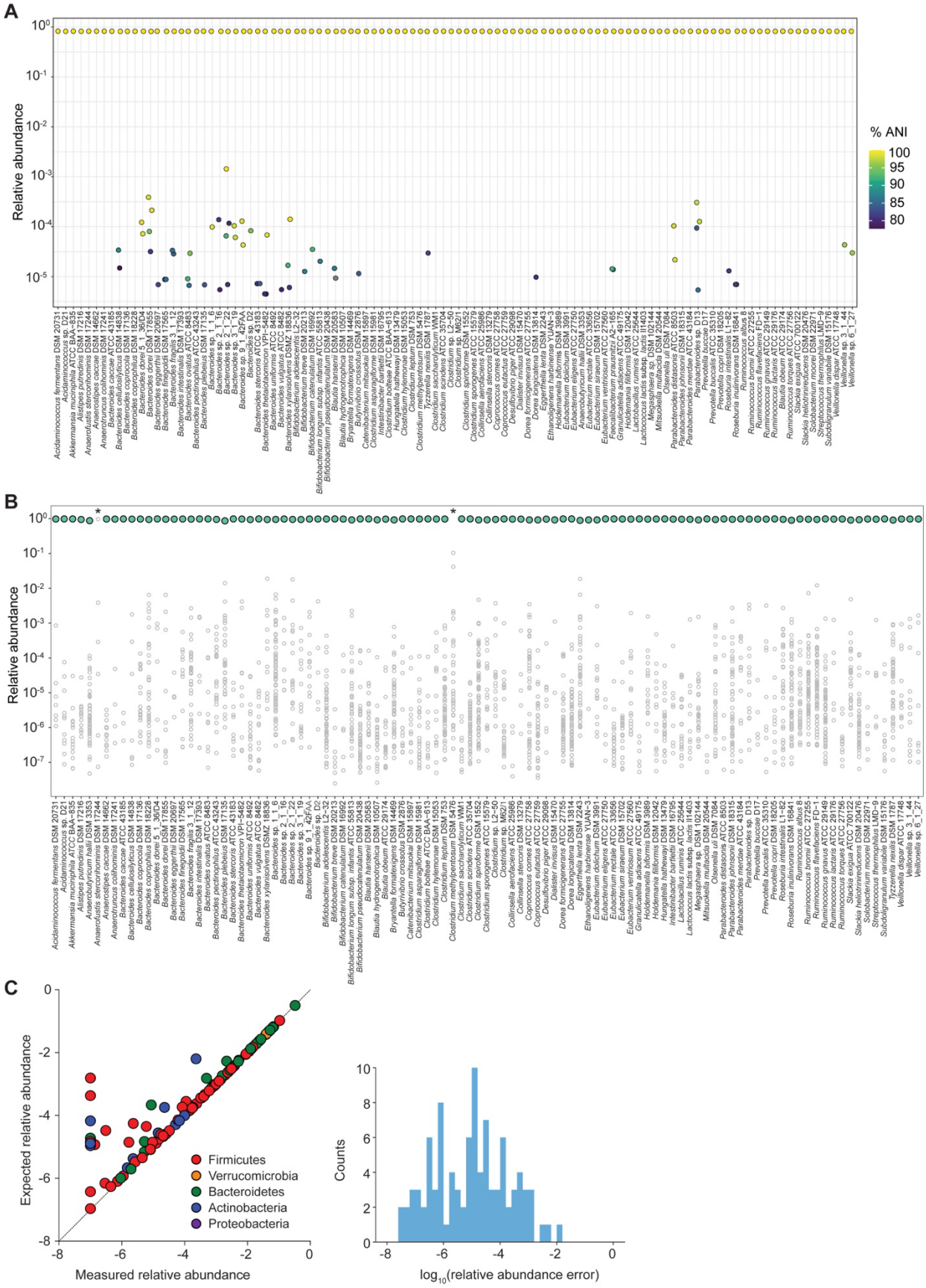
Data S2: Standardization and benchmarking of NinjaMap, related to STAR Methods.

Table of Contents

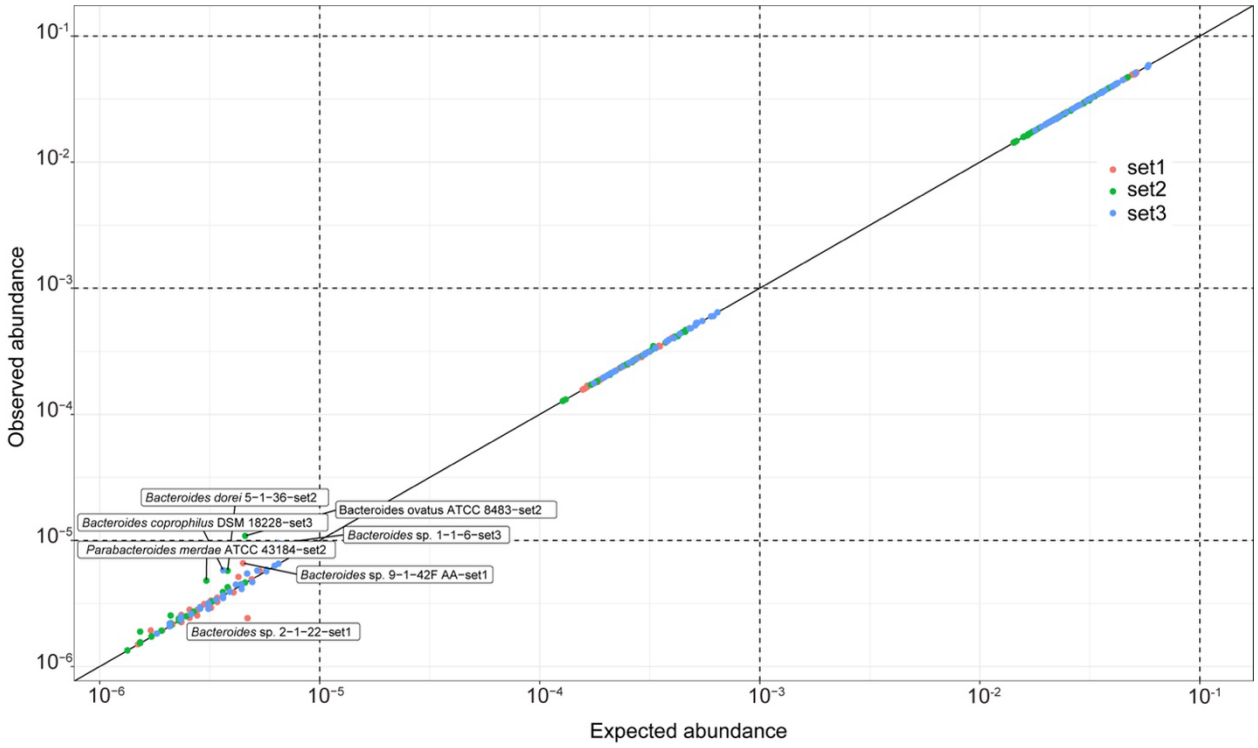


A sensitive and specific read mapping pipeline. (**A**) A schematic of NinjaMap, an algorithm that quantifies strain abundances in defined communities with high accuracy. Reads that match a single genome unambiguously are assigned to that genome, and reads that match multiple genomes are placed in escrow. An initial estimate of the relative abundance of each strain is computed from the unambiguous alignments and used to assign escrow reads proportionally. The final read counts are then normalized to obtain relative abundances. (**B**) To benchmark the performance of NinjaMap, we compared it with five publicly available tools: Kraken/Bracken, IGGSearch, MetaPhlAn2, MIDAS, and Sourmash. We used isolate sequencing data from 74

strains and assessed the top hits for accuracy. MetaPhlAn2 achieved the most accurate results by identifying all 74 strains correctly, followed by Sourmash and NinjaMap (73 each), and then by Midas (69), IGGSearch (56), and Bracken (30). Since MetaPhlAn2 and Sourmash do not classify the complete set of reads, but instead rely on the presence and identification of particular marker genes or k-mer signatures to infer community composition, we decided to forgo their use. Instead, for all analyses with defined communities we used NinjaMap, which can assign taxonomies to >97% of reads post-QC with near-MetaPhlAn2 accuracy.


A
Relative abundance
% ANI
100
95
90
85
80
B
Relative abundance
C
Expected relative abundance
Measured relative abundance
Firmicutes
Verrucomicrobia
Bacteroidetes
Actinobacteria
Proteobacteria
Counts
log10(relative abundance error)

Performance of NinjaMap on simulated and actual single-strain sequencing data. (**A**) NinjaMap analysis of simulated (*in silico*-generated) sequencing data for each strain in hCom1. The yellow dot represents the relative abundance of the strain that was sequenced; the remaining dots represent the relative abundance of other strains that were called in the sample. Points are colored based on percent average nucleotide identity (ANI) values compared to the strain being sampled; a small amount of jitter has been added to aid in viewing overlapping circles. Mapping to the correct genome was achieved for the vast majority of strains, even when other highly similar genomes are present in the database. (**B**) NinjaMap analysis of data derived from sequencing each strain in the community individually. The green dot represents the relative abundance of the strain that was sequenced; gray dots represent the relative abundance of other strains in the sample. The two samples with an asterisk instead of a green circle were contaminated in this experiment, so the strains in question (*Anaerofustis stercorihominis* DSM 17244 and *Clostridium methylpentosum* DSM 5476) were not present. We subsequently replaced those stocks with pure, uncontaminated cultures. (**C**) Expected mismapping in a community. Left: Comparison of the measured relative abundances of fecal samples of hCom1-colonized mice at week 4 (**Figure 3D**) versus the relative abundances expected from isolate mis-mapping, which was calculated for strain i by multiplying the rate of mismapping into strain i from strain j by the measured relative abundance of strain j and summing across all strains j. Right: The difference between the measured and expected relative abundances across all strains.



NinjaMap accuracy and sensitivity for *in silico* metagenomic sequencing data. NinjaMap was applied to Grinder-simulated metagenomic sequencing reads generated from pooled hCom1 community strains in three separate trials. The expected abundances were almost perfectly correlated with the observed abundances computed using NinjaMap. Thus, NinjaMap is able to accurately measure strain abundances as low as 10^{-6} in the context of these simulated mixed communities of known composition.