Accurate estimation of intraspecific microbial gene content variation in metagenomic data with MIDAS v3 and StrainPGC

Supplemental Materials

Extended hCom2 benchmarking results

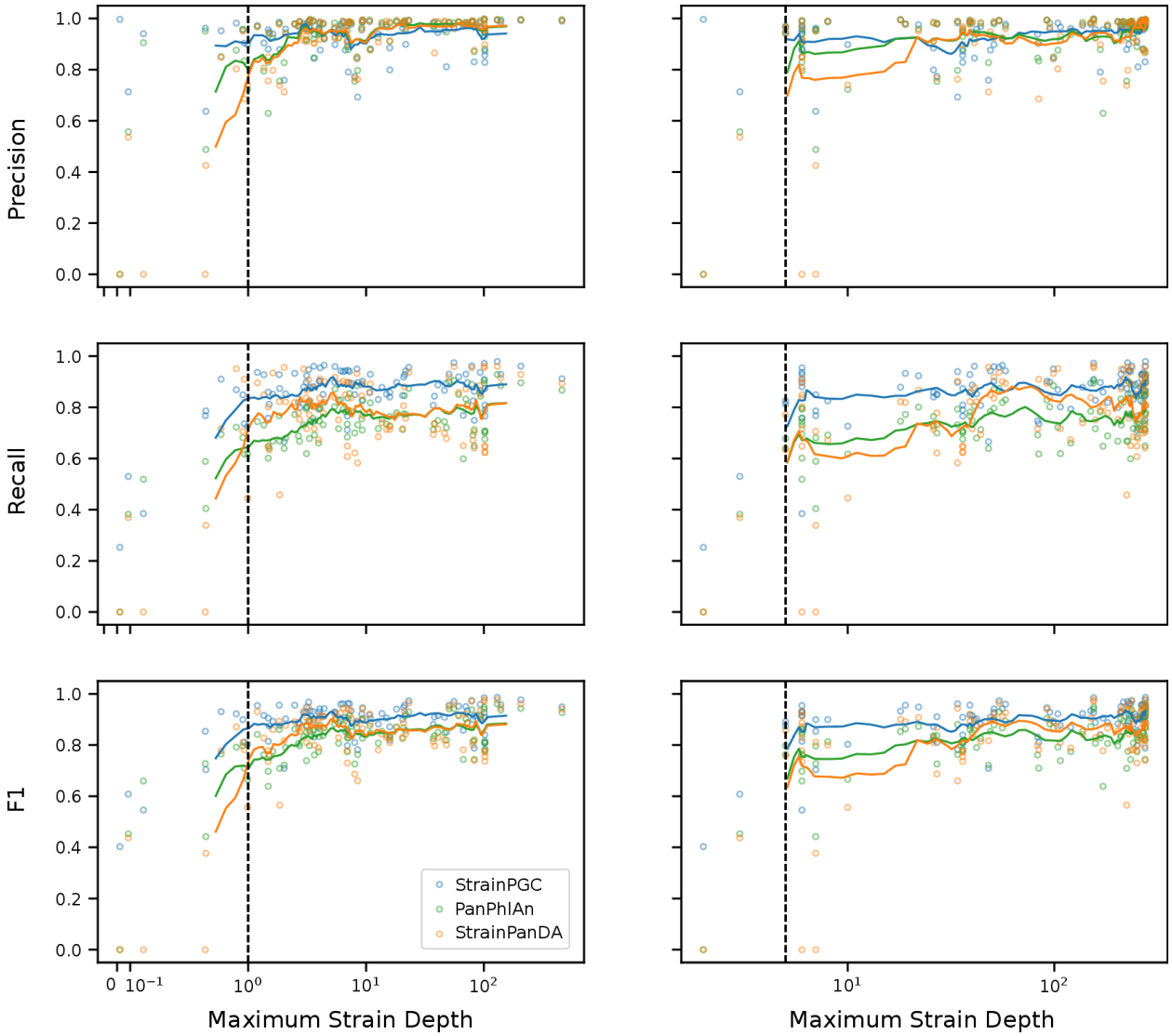

Figure S1: Relationship between sequencing depth or number of samples and the accuracy of gene content estimation. *Points represent the performance of each tool (colors) on each of the 97 benchmark strains. For the left column, the x-axis is the maximum estimated depth of the genotype-matched strain across strain-pure samples, and for the right column it is the total number of strain-pure samples identified for that strain. Trend lines are a rolling average over the 10 nearest points. The dotted vertical line indicates the 1× depth and 5 strain-pure samples, after which the mean performance stabilizes for StrainPGC.*

Extended pangenome results

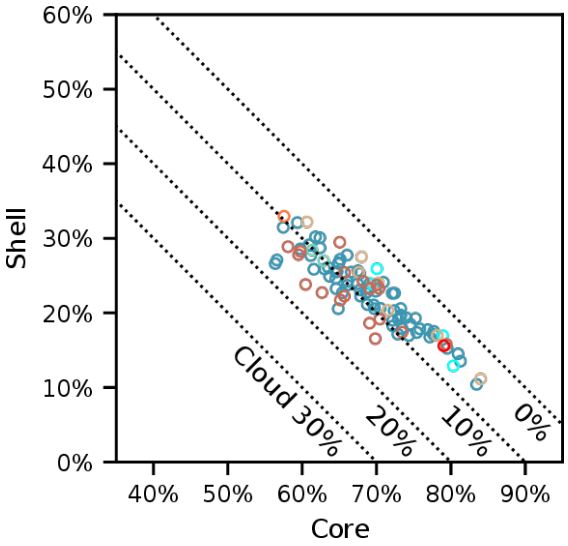

Figure S2: Per-genome core, shell, and cloud gene fractions in reference genomes. *Equivalent results to Fig. 4B, here calculated using reference genomes for comparison to StrainPGC-based gene content estimates.*

Simulated *E. coli* spike-in validation

We performed an additional benchmarking study to validate our approach in datasets with substantially more strain-diversity, for strains with more divergence from the reference set, and with a limited number of strain-pure samples. To keep the simulated metagenomeic data as realistic as possible, we opted to construct samples with novel strains by “spiking” simulated reads from recently sequenced isolates into real metagenomes from the HMP2 study. Due to an abundance of studies with wild *E. coli* isolates, and our particular focus on this species throughout, we identified five novel *E. coli* genomes from a recently published project (Davidova-Gerzova et al. 2023). These isolates varied greatly in their relatedness to the UHGG

reference genomes, including very distantly related strains with a genotype dissimilarity of 0.077. These strains are as novel relative to the reference database as would be expected for *E. coli* found in the human gut; only 0.8% of UHGG genomes had a closest match genotype-dissimilarity of more than 0.077.

We selected five HMP2 samples, all from one subject (C3022), where *E. coli* was not detected. Into these, we spiked-in simulated reads at 1×, 2×, 4×, 8×, and 16× depths with a separate set of reads for each strain. We combined all 25 of these additional, synthetic samples with the full HMP2 dataset, and then re-ran our integrated workflow. We matched the inferred strains to each of the ground-truth genomes based on genotype similarity and evaluated the StrainPGC gene content estimates as in the hCom2 benchmark.

We found that the performance of StrainPGC in these simulations with non-reference *E. coli* genomes is consistent with the overall performance on the hCom2 (synthetic community) benchmark. This is despite the fact that the metagenomes were much more complex and some strains were more dissimilar to the closest reference genome. Specifically, we found a median F1 score across all strains of 0.92, equivalent to the median F1 of 0.91 from the hCom2 benchmark. We do not find a negative relationship between the divergence of the benchmark genome and performance. StrainPGC performance was nearly equivalent for the least diverged (F1 of 0.89) and most diverged genomes (F1 of 0.92). We conclude that it is reasonable to expect similar performance for other strains and datasets, even when the number of strains for a species is large and when strains are more diverged from the reference database.

Supplementary Table S2*: Performance on five E. coli genomes in an* in silico *spike-in experiment.*

GenBank Accession	**Closest UHGG Reference**	**Closest Genotype Dissimilarity**	**Precision**	**Recall**	**F1**
GCF_030198905.1	GUT_GENOME144970	0.0039	0.97	0.87	0.92
GCF_030202075.1	GUT_GENOME140957	0.0078	0.96	0.87	0.92
GCF_030204715.1	GUT_GENOME144767	0.0011	0.97	0.82	0.89
GCF_030205145.1	GUT_GENOME144552	0.030	0.96	0.87	0.91
GCF_030205875.1	GUT_GENOME144360	0.077	0.97	0.87	0.92

Sensitivity of StrainPGC performance to depth ratio and correlation score thresholds

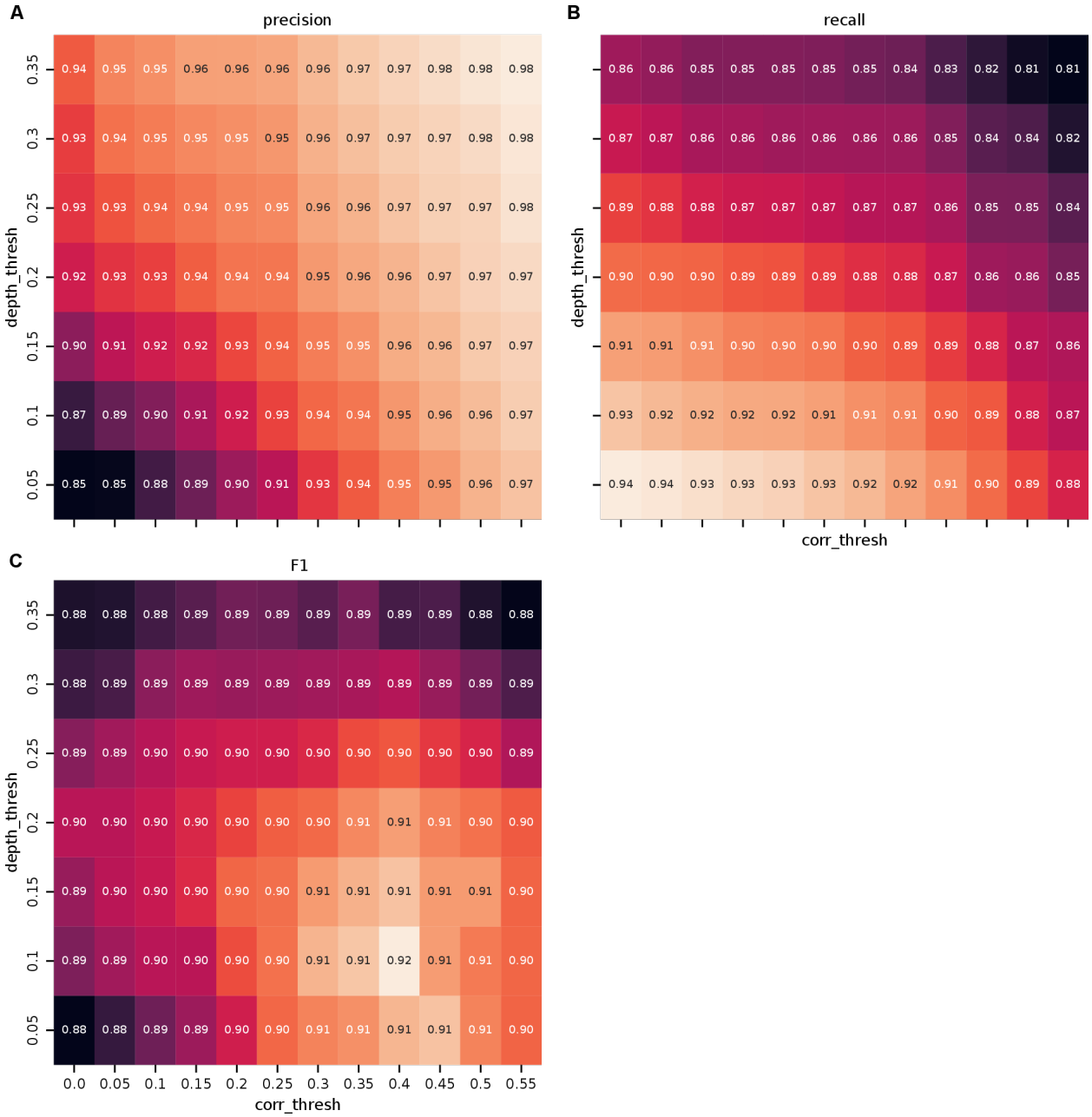

Figure S3: Threshold depth ratio and correlation score parameter search. *Median performance across 97 hCom2 benchmark genomes at every combination of 11 correlation score thresholds (x-axis) and 7 depth ratio thresholds (y-axis). Panels represent median precision (A), recall (B), and F1 score (C). The best performance (F1 score) was achieved at a depth ratio threshold of 0.1 and correlation threshold of 0.40. We used a slightly more conservative depth threshold of 0.2 throughout the rest of this work, which decreased the median F1 score negligably from 0.916 to 0.908.*