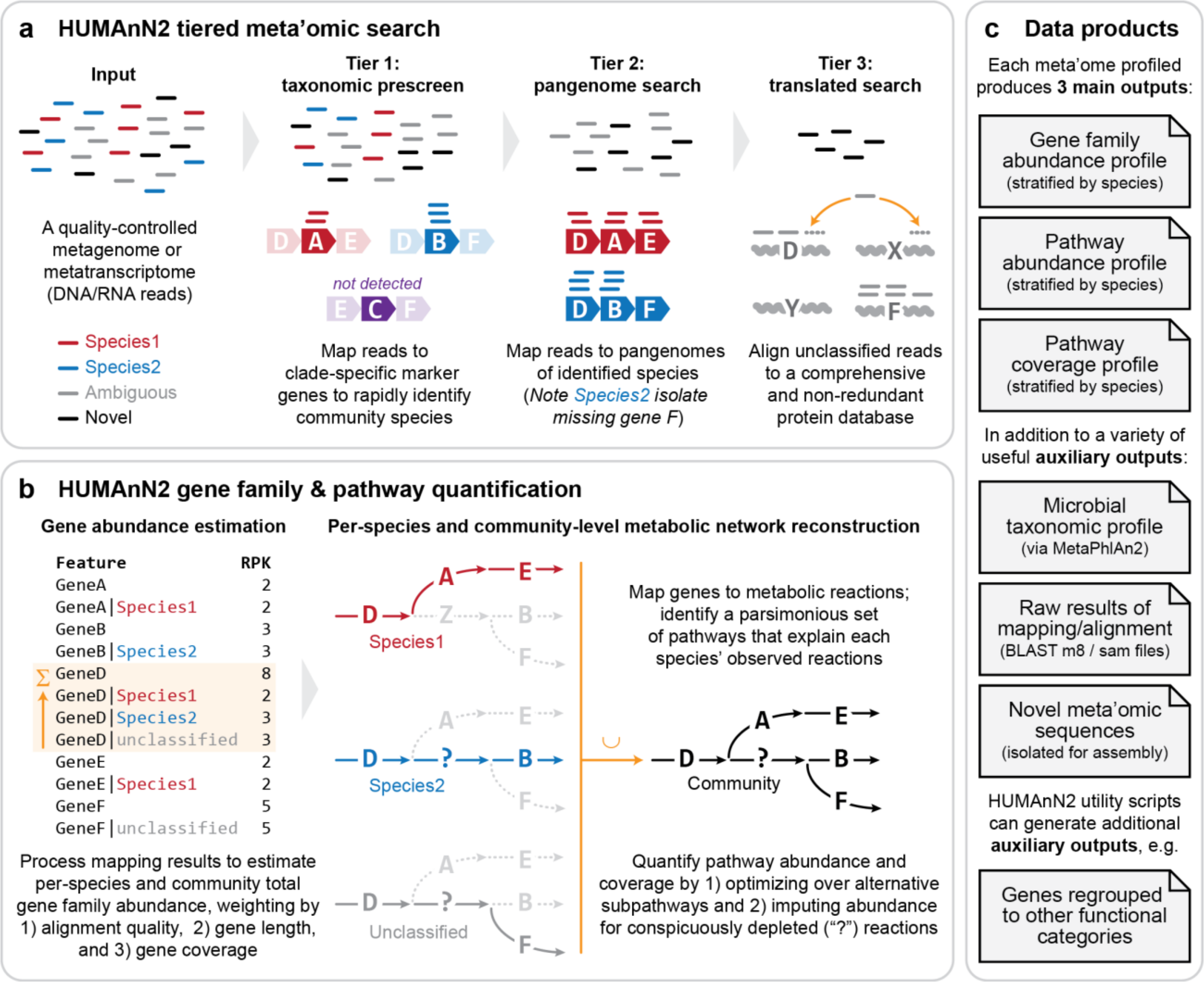
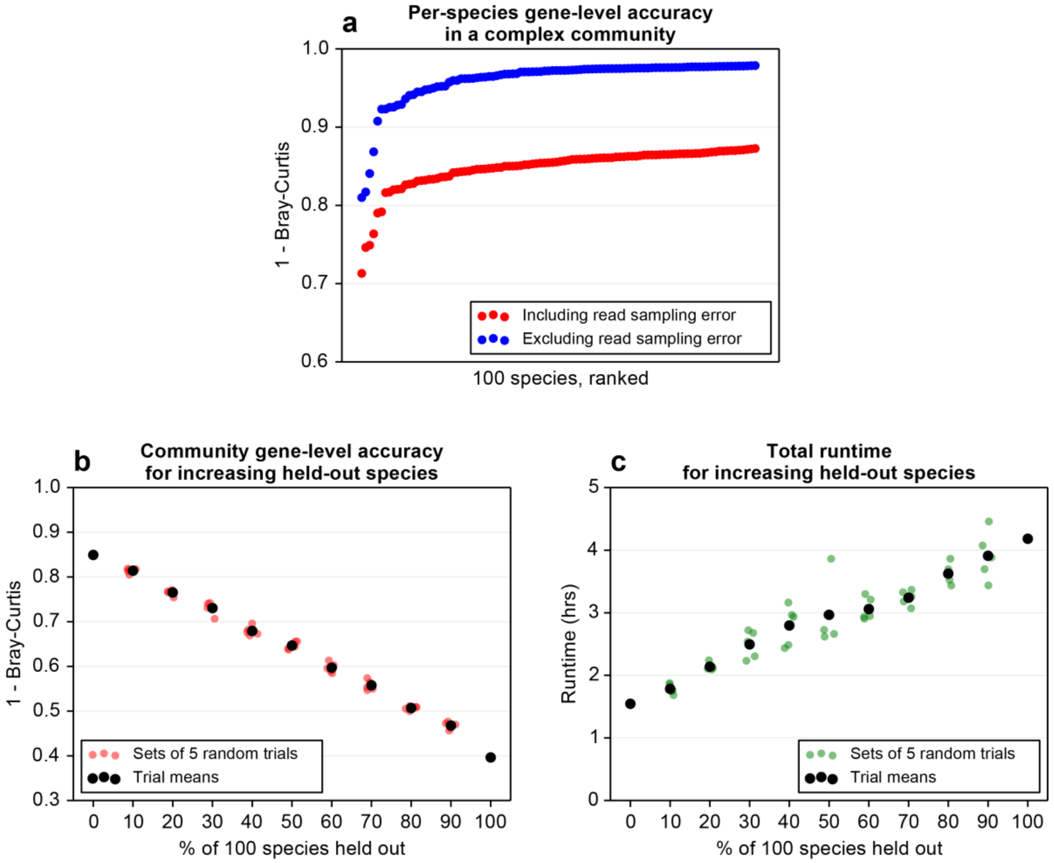
SUPPLEMENTARY FIGURES

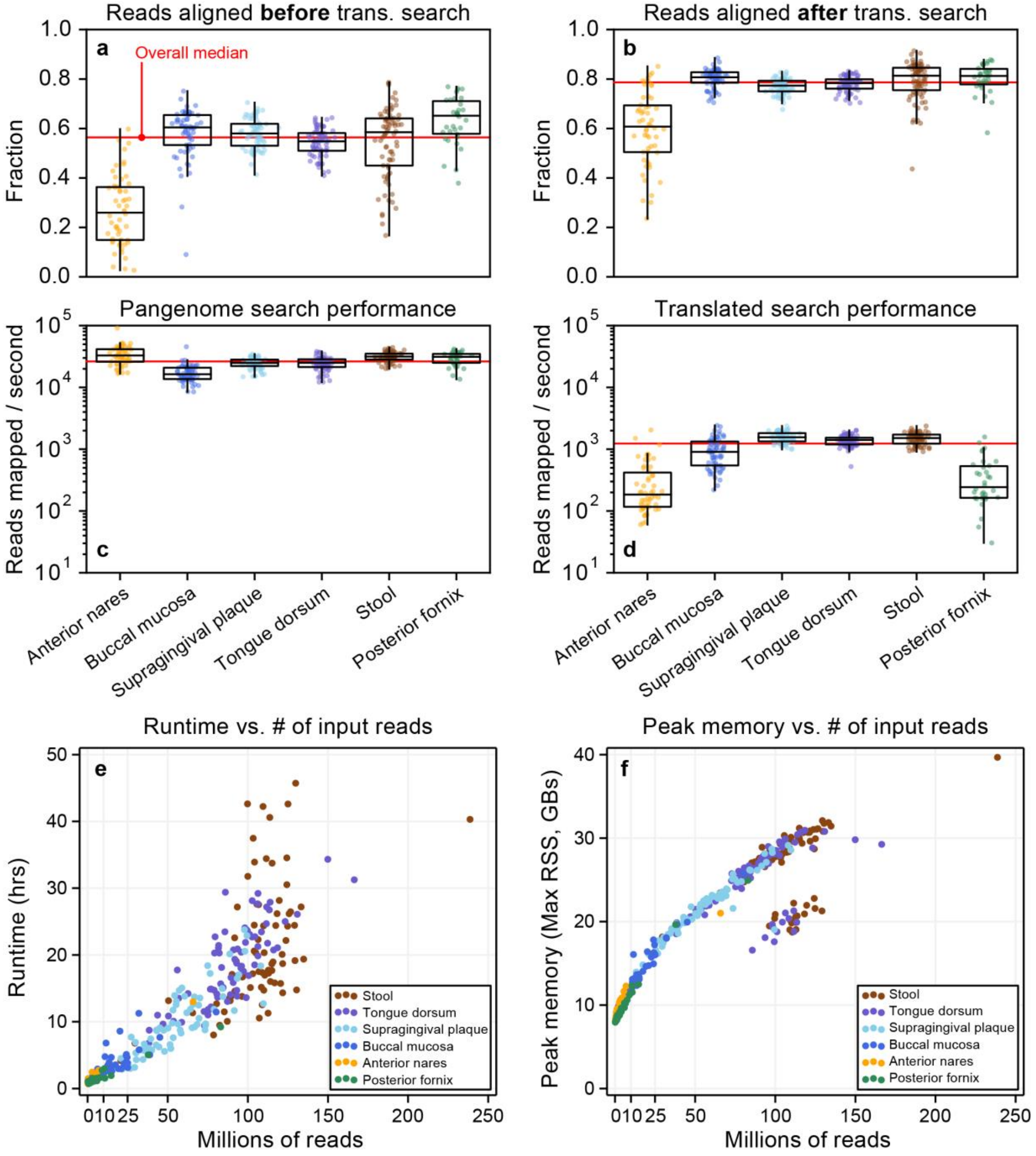
Supplementary Figure 1. Expanded overview of the HUMAnN2 method. (**a**) HUMAnN2 implements a tiered meta'omic search that aims to explain the origin of microbial community DNA or RNA reads based on the pangenomes of detected microbes before falling back to more computationally expensive translated search. (**b**) The tiered search produces alignments of reads to coding sequences of known or ambiguous taxonomy. These alignments are processed in a species-specific manner to calculate gene family abundance and reconstruct community metabolic pathways. (**c**) HUMAnN2 thus provides, for each community meta'ome: per-gene abundances, pathway presence/absence calls and abundances, and downstream visualization and statistical tests.


a HUMAnN2 tiered meta'omic search
Input
Tier 1: taxonomic prescreen
Tier 2: pangenome search
Tier 3: translated search
A quality-controlled metagenome or metatranscriptome (DNA/RNA reads)
D A E
D B F
not detected
E C F
D A E
D B F
D X
Y F
Species1
Species2
Ambiguous
Novel
Map reads to clade-specific marker genes to rapidly identify community species
Map reads to pangenomes of identified species (Note Species2 isolate missing gene F)
Align unclassified reads to a comprehensive and non-redundant protein database
b HUMAnN2 gene family & pathway quantification
Gene abundance estimation
Per-species and community-level metabolic network reconstruction
Feature RPK
GeneA 2
GeneA|Species1 2
GeneB 3
GeneB|Species2 3
∑ GeneD 8
GeneD|Species1 2
GeneD|Species2 3
GeneD|unclassified 3
GeneE 2
GeneE|Species1 2
GeneF 5
GeneF|unclassified 5
Process mapping results to estimate per-species and community total gene family abundance, weighting by 1) alignment quality, 2) gene length, and 3) gene coverage
D A E Z B F
Species1
D ? A E B F
Species2
D ? A E B F
Unclassified
∪
D A E ? B F
Community
Map genes to metabolic reactions; identify a parsimonious set of pathways that explain each species' observed reactions
Quantify pathway abundance and coverage by 1) optimizing over alternative subpathways and 2) imputing abundance for conspicuously depleted ("?") reactions
c Data products
Each meta'ome profiled produces 3 main outputs:
Gene family abundance profile (stratified by species)
Pathway abundance profile (stratified by species)
Pathway coverage profile (stratified by species)
In addition to a variety of useful auxiliary outputs:
Microbial taxonomic profile (via MetaPhlAn2)
Raw results of mapping/alignment (BLAST m8 / sam files)
Novel meta'omic sequences (isolated for assembly)
HUMAnN2 utility scripts can generate additional auxiliary outputs, e.g.
Genes regrouped to other functional categories

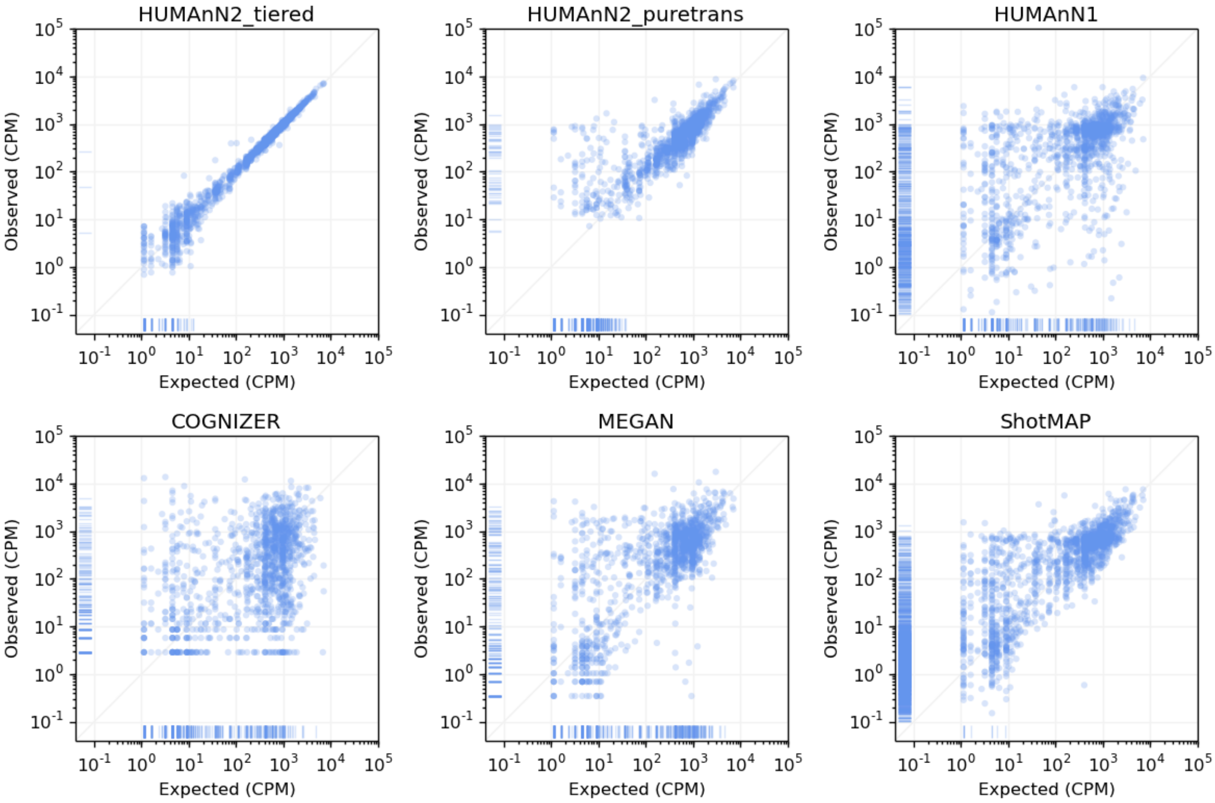
Supplementary Figure 2. Reference hold-out analysis of a complex synthetic metagenome. We constructed and analyzed with HUMAnN2 a 100-member mock-even synthetic metagenome containing only non-human associated species (~2x coverage per species). (**a**) Variation in the number of reads sampled per gene (compared with a genome's average fold-coverage) makes a non-trivial contribution to the error in per-species gene abundance estimation in HUMAnN2 (roughly 0.1 Bray-Curtis dissimilarity units). (**b**) Accuracy of community-level gene family abundance estimation decreases linearly with the number of community species missed by HUMAnN2's taxonomic prescreen (simulated here by excluding sets of species from the underlying pangenome reference collection). (**c**) HUMAnN2's overall runtime increases linearly as more species are excluded from the taxonomic prescreen (which results in more work being done during translated search). Runtimes reflect execution using 8 CPU cores.



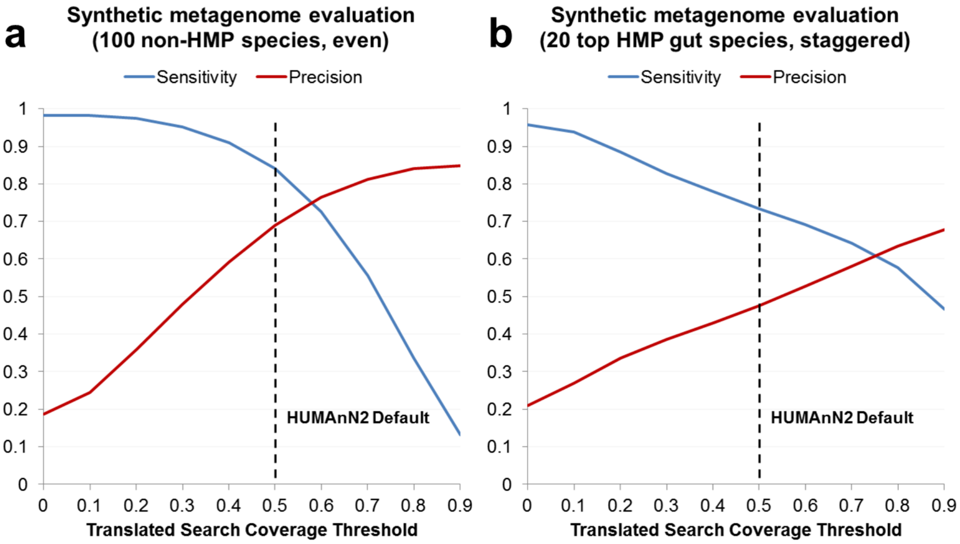
Supplementary Figure 3. HUMAnN2 tiered search performance on human metagenomes. We applied HUMAnN2's tiered search to profile 397 first-visit HMP metagenomes on Harvard University's Odyssey Research Computing Cluster (8 CPU cores per job). Sample counts per body site were as follows: 54 for anterior nares, 65 for buccal mucosa, 68 for supragingival plaque, 73 for tongue dorsum, 76 for stool, and 34 for posterior fornix. (**a**) At most body sites, ~60% of reads were explained by detected pangenomes, with (**b**) an additional ~20% explained by downstream translated search (~80% total). Pangenome search performance (**c**) consistently exceeded translated search performance (**d**) by 1-2 orders of magnitude. From smallest to largest, box plot elements in panels a-d represent the lower inner fence, first quartile, median, third quartile, and upper inner fence. Horizontal red lines indicate the median value over all samples. (**e**) Total runtime is largely dictated by the number of reads passed to translated search, and (for HMP samples with <100M reads) was approximately linear in the number of input reads (~1 hr/5M input reads). (**f**) Peak memory use was sublinear in the number of input reads and very predictable. The cluster of outliers in (f) results from large samples that were requeued during their runs: these samples resumed later in the HUMAnN2 workflow and hence display smaller peak memory use.



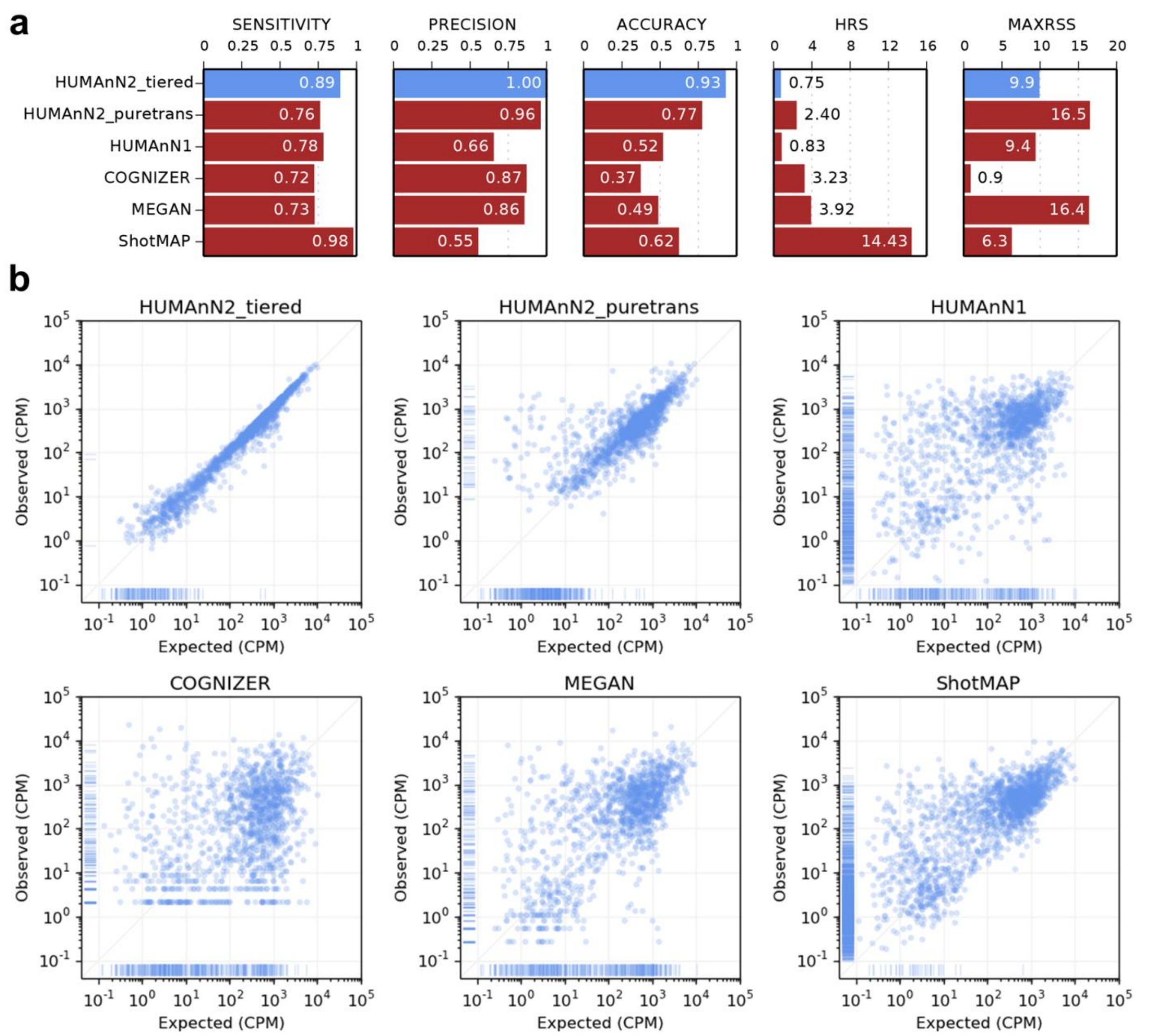
Supplementary Figure 4: HUMAnN2 compared with other methods (details). We profiled a 10M read synthetic gut metagenome using HUMAnN2 (tiered and pure translated search modes), HUMAnN1[1], COGNIZER[2], MEGAN[3], and ShotMAP[4] to produce profiles of COG[5] abundance. Here, expected (gold standard) and observed COG abundances are compared in units of copies per million (CPMs; i.e. raw abundance normalized by gene length and number of mapped reads). HUMAnN2's tiered search was considerably more accurate than the other methods based on pure translated search. HUMAnN2's pure translated search showed better agreement than other translated search methods, with its largest source of error being underreporting of low-abundance COGs (false negatives). This behavior is expected from the translated search coverage filters used in HUMAnN2, which we use to limit false positive detection events (i.e. COGs with zero expected abundance and non-zero observed abundance). Ticks in the x- and y-axis margins represent zero values; x-axis ticks are false negatives and y-axis ticks are false positives.



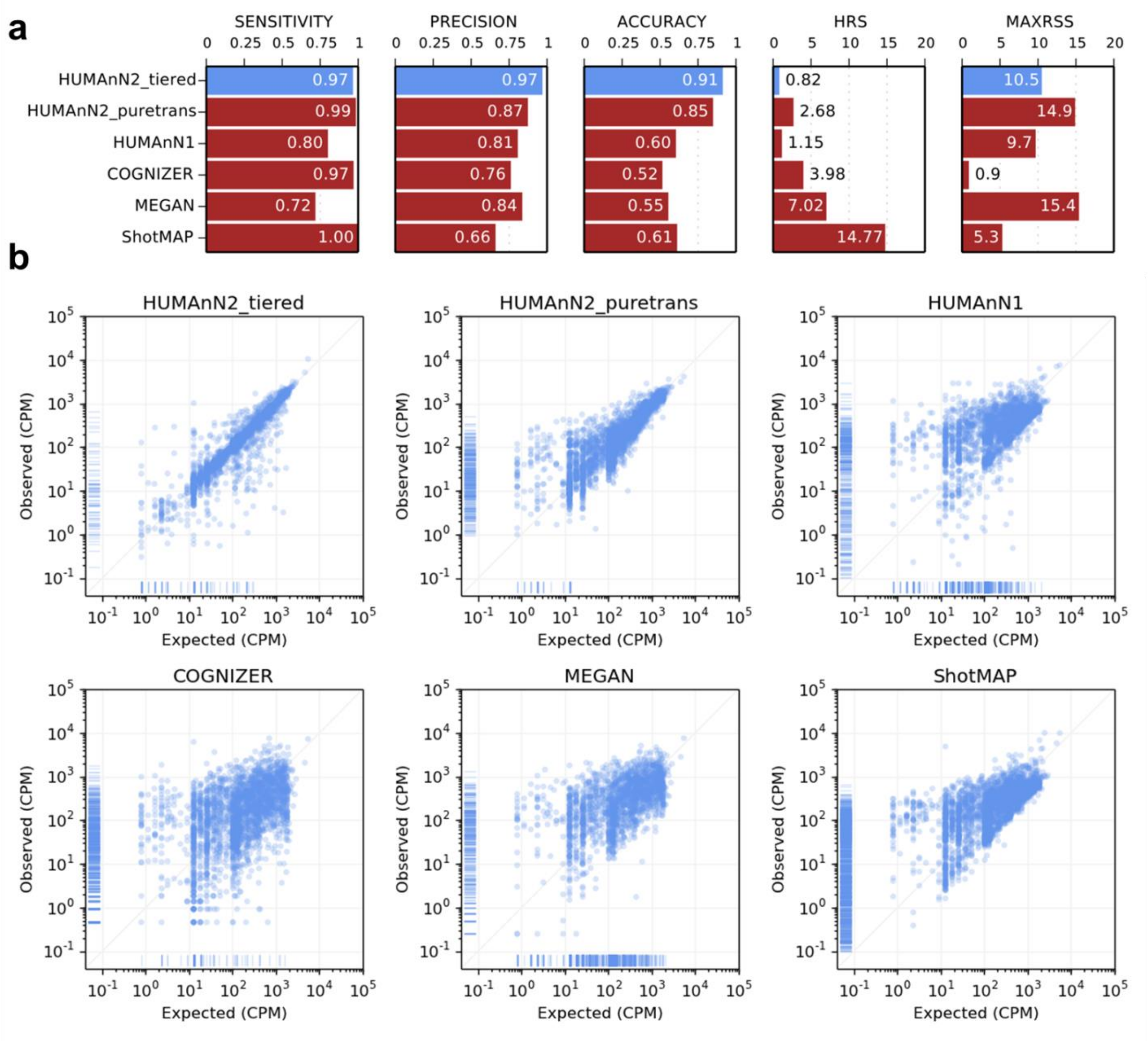
Supplementary Figure 5. Protein coverage thresholds in translated search. If two largely unrelated proteins share local sequence homology, reads drawn from the homologous region will map to both proteins, potentially resulting in false positive detection events. To limit such events, we require a threshold fraction of sites in a protein to recruit reads during translated search before considering the protein “detected.” We evaluated potential thresholds by analyzing the results of pure translated search of synthetic metagenomes versus the UniRef90 database. Trade-offs between sensitivity and precision are shown for the 100-member, even, non-human-associated metagenome in (**a**), and the 20-member, staggered, human-gut-associated metagenome in (**b**). When all community genomes are well-covered, a 50% coverage threshold (HUMAnN2’s default) yields a marked increase in precision with only minor loss of sensitivity (a). Loss of sensitivity is higher at this threshold when rare (low-coverage) genomes are included, as genes in low-coverage genomes often fail to meet the coverage threshold due to insufficient read sampling (b). These evaluations do not reflect any additional post-processing of translated search results (e.g. weighting by alignment quality), which provide additional accuracy improvements.



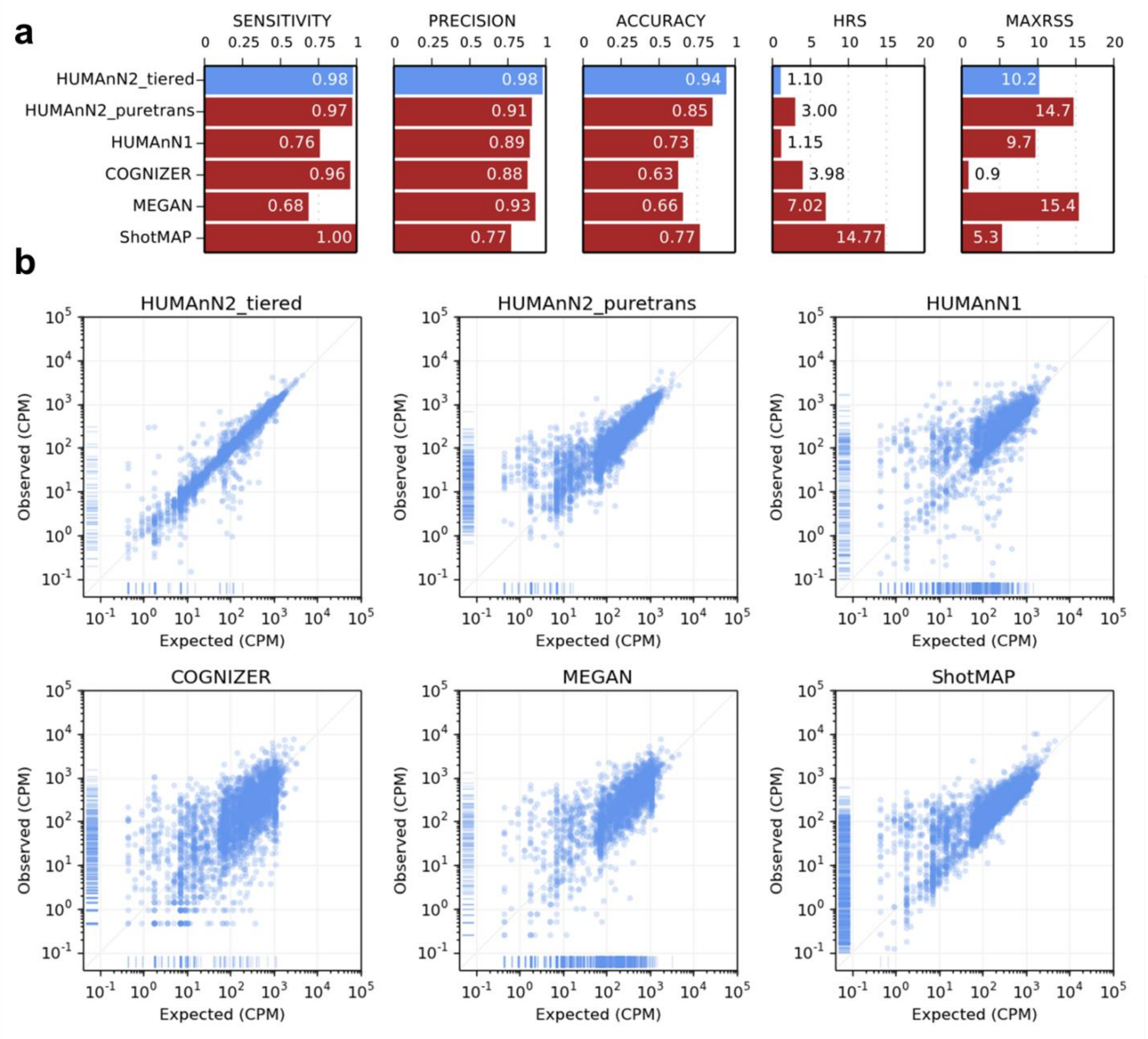
Supplementary Figure 6: HUMAnN2 compared with other methods: synthetic metatranscriptome evaluation. We profiled a 10M-read synthetic gut metatranscriptome using HUMAnN2 (tiered and pure translated search modes), HUMAnN1, COGNIZER, MEGAN, and ShotMAP to produce profiles of community-level COG transcript abundance. Twenty species' genomic abundance values were geometrically staggered (as in the gut metagenome evaluation), while genes (transcripts) were sampled within-species following a log-normal distribution [ln *N*(0, 1)]. (**a**) Measures of methods' accuracy and performance in this evaluation. All methods were allowed to use 8 CPU cores and up to 30 GB of memory. This panel is analogous to Fig. 1e (which focuses on metagenomic COG abundance in the same synthetic community). (**b**) Observed versus expected COG transcript abundance across the six methods. This panel is analogous to Supplementary Fig. 4. CPM refers to "copies per million." Ticks in the x- and y-axis margins represent zero values; x-axis ticks are false negatives and y-axis ticks are false positives.



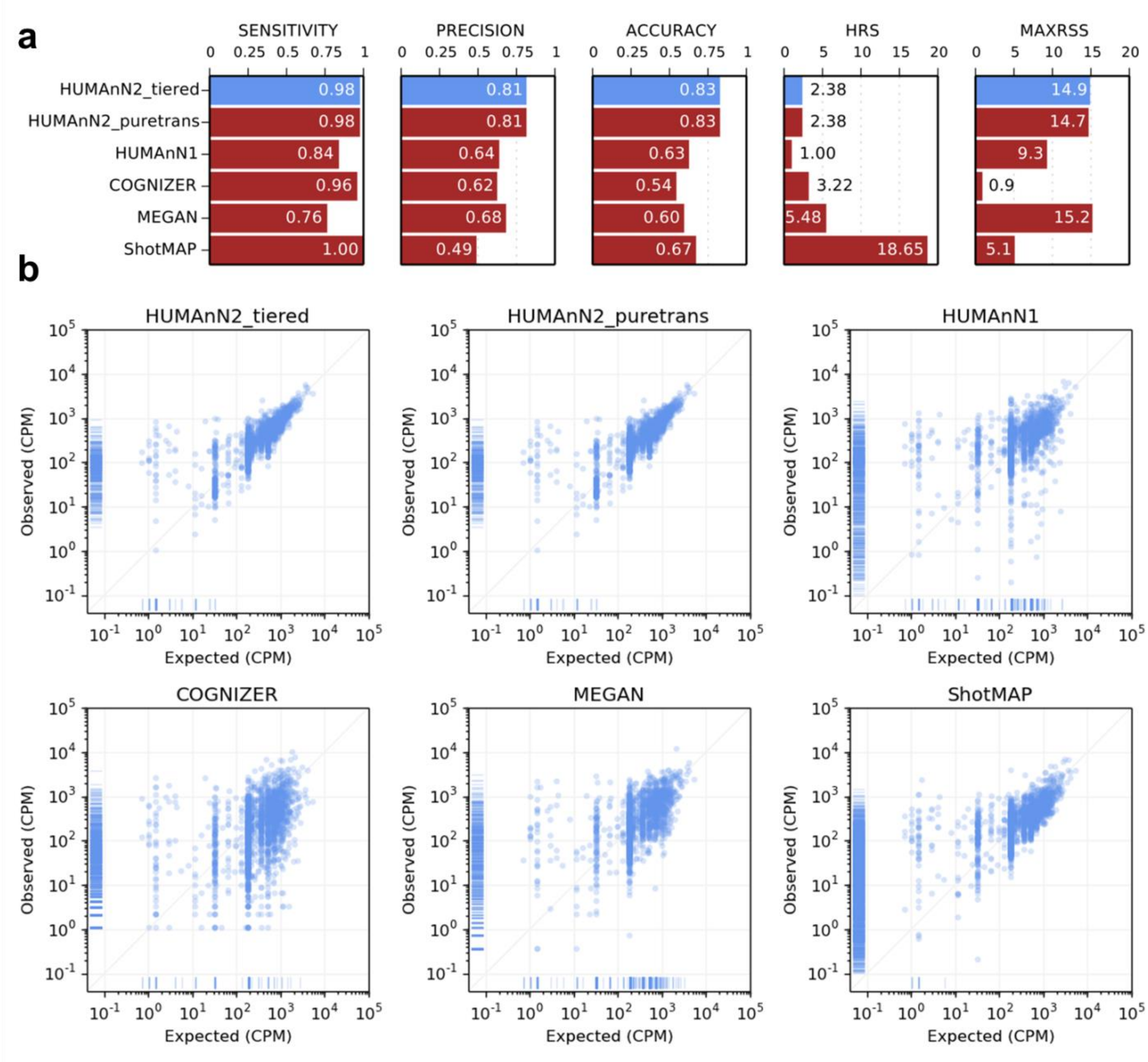
Supplementary Figure 7: HUMAnN2 compared with other methods: novel isolates of known species, UniRef90-based COG gold standard. We profiled a 10M-read synthetic metagenome using HUMAnN2 (tiered and pure translated search modes), HUMAnN1, COGNIZER, MEGAN, and ShotMAP to produce profiles of community-level COG abundance. Twenty recent, new isolates of known species (i.e. species present in HUMAnN2's pangenome database) were sampled at staggered relative abundance. (**a**) Measures of methods' accuracy and performance in this evaluation. All methods were allowed to use 8 CPU cores and up to 30 GB of memory. This panel and analysis are analogous to those in Fig. 1e. (**b**) Observed versus expected COG transcript abundance across the six methods. This panel is analogous to Supplementary Fig. 4. CPM refers to "copies per million." Ticks in the x- and y-axis margins represent zero values; x-axis ticks are false negatives and y-axis ticks are false positives.



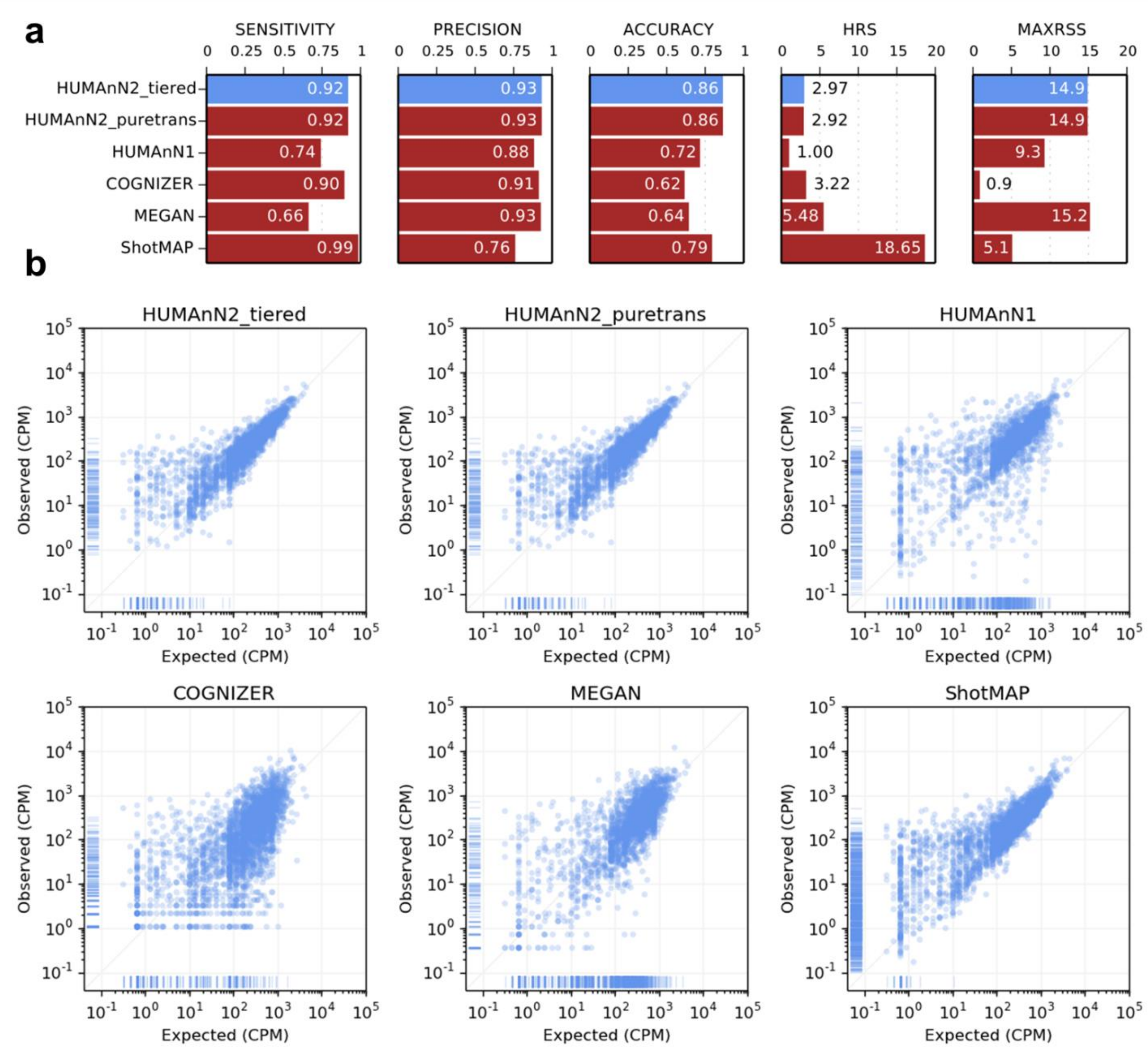
Supplementary Figure 8: HUMAnN2 compared with other methods: novel isolates of known species, UniRef50-based COG gold standard. This figure mirrors Supplementary Fig. 6 above, except that COG annotations are defined based on co-clustering with UniRef50 families (rather than UniRef90). Similarly, HUMAnN2 was run in UniRef50 mode. These changes will tend to favor sensitivity over specificity during both isolate genome annotation and profiling. (**a**) Accuracy and performance of the six functional profiling methods. (**b**) Observed versus expected COG abundance.



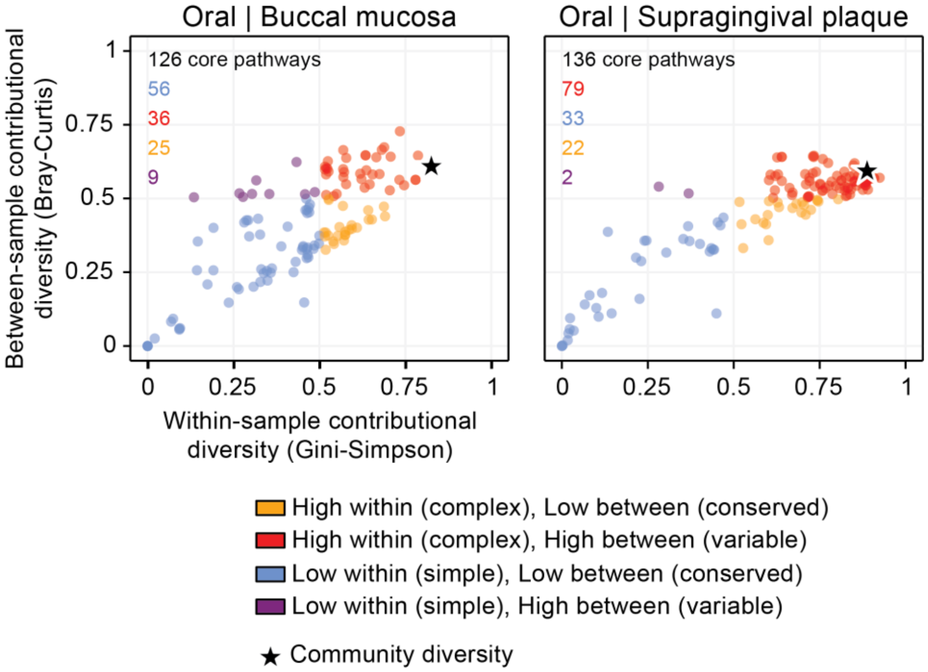
Supplementary Figure 9: HUMAnN2 compared with other methods: isolates of novel species, UniRef90-based COG gold standard. We profiled a 10M-read synthetic metagenome using HUMAnN2 (tiered and pure translated search modes), HUMAnN1, COGNIZER, MEGAN, and ShotMAP to produce profiles of community-level COG abundance. Twenty recent, new isolates of novel species (i.e. species not present in HUMAnN2's pangenome database) were sampled at staggered relative abundance. Note that, in this context, HUMAnN2's tiered search relies entirely on the translated search phase to explain sample reads. (**a**) Measures of methods' accuracy and performance in this evaluation. All methods were allowed to use 8 CPU cores and up to 30 GB of memory. This panel and analysis are analogous to those in Fig. 1e. (**b**) Observed versus expected COG transcript abundance across the six methods. This panel is analogous to Supplementary Fig. 4. CPM refers to "copies per million." Ticks in the x- and y-axis margins represent zero values; x-axis ticks are false negatives and y-axis ticks are false positives. Vertical striping of "expected COG abundance" results from single-copy COGs that were only assigned to one genome (and hence all have the same expected coverage).



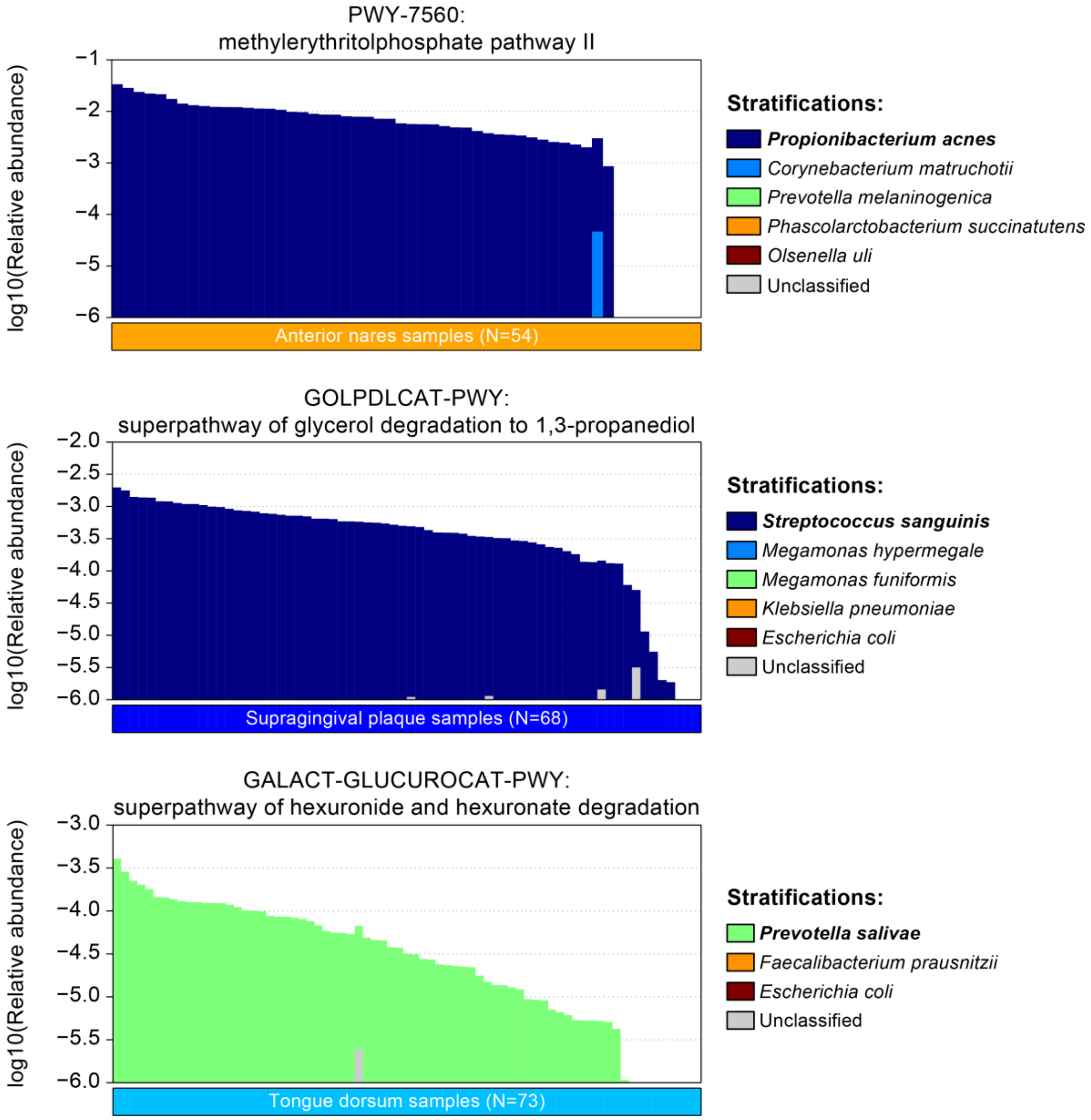
Supplementary Figure 10: HUMAnN2 compared with other methods: isolates of novel species, UniRef50-based COG gold standard. This figure mirrors Supplementary Fig. 8 above, except that COG annotations are defined based on co-clustering with UniRef50 families (rather than UniRef90). Similarly, HUMAnN2 was run in UniRef50 mode. These changes will tend to favor sensitivity over specificity during both isolate genome annotation and profiling. (**a**) Accuracy and performance of the six functional profiling methods. (**b**) Observed versus expected COG abundance.



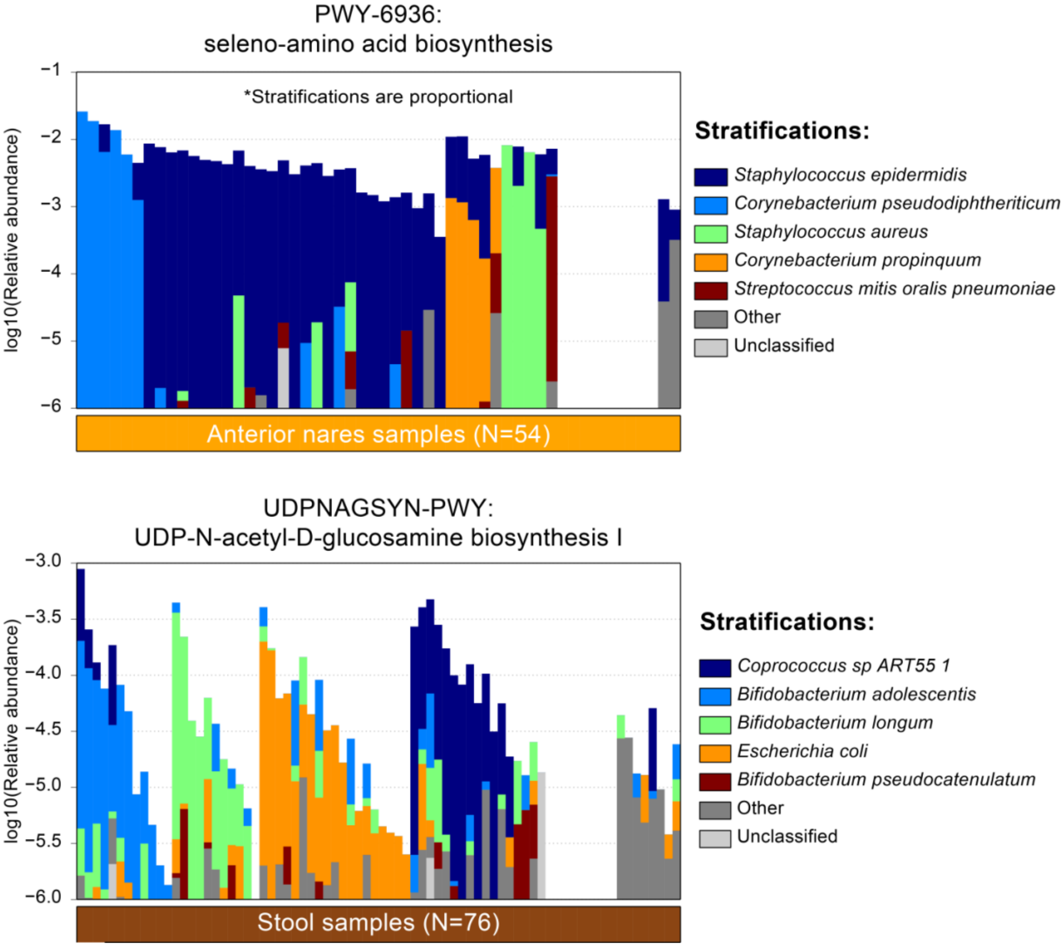
Supplementary Figure 11. Contributional diversity at additional oral sites. This figure follows the format of Fig. 2 from the main text and includes data for two additional oral body sites: buccal mucosa and supragingival plaque. Stars indicate background species-level community diversity.



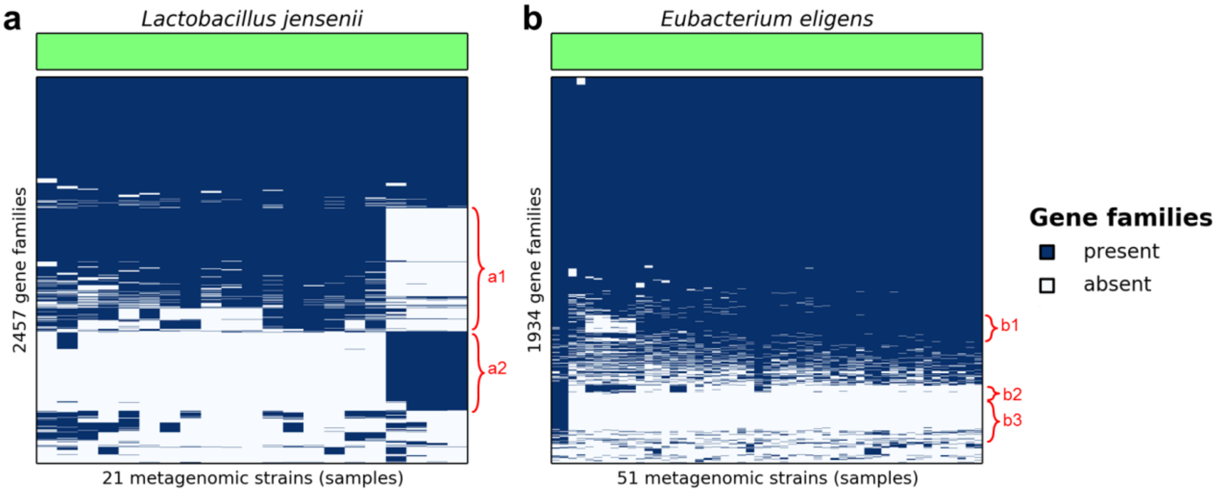
Supplementary Figure 12. Additional examples of core human microbiome pathways with low within-subject and low between-subject contributional diversity. Bar heights represent the total relative abundance of the pathway and are log-scaled. Contributions of individual species/other/unclassified are linearly scaled within the total bar height.



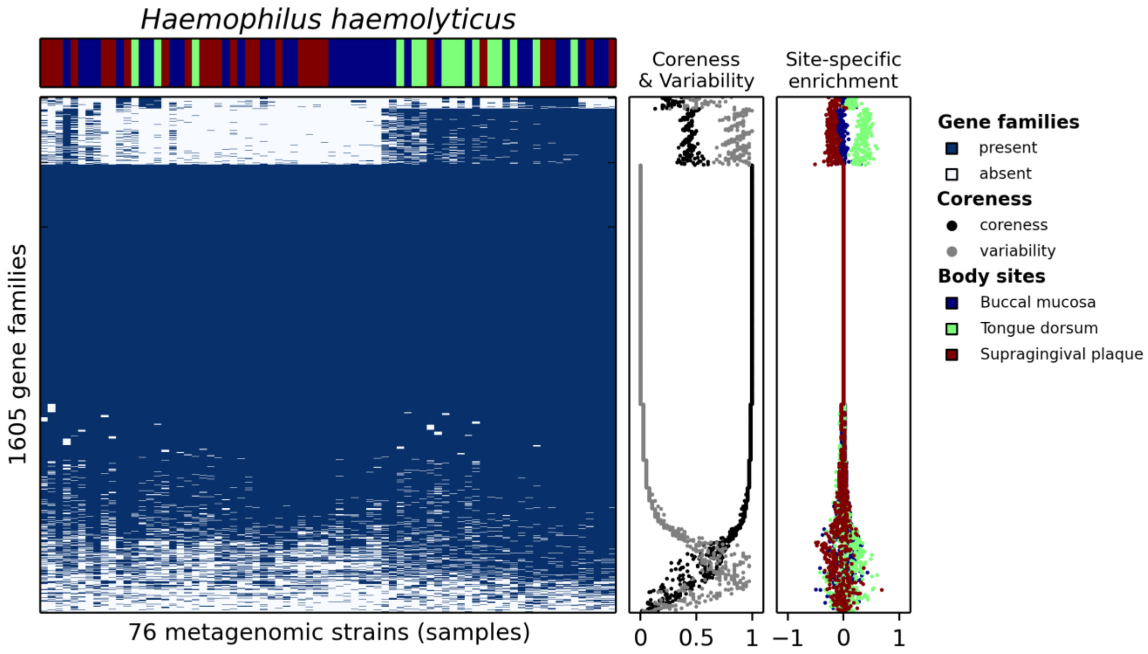
Supplementary Figure 13. Non-vaginal examples of human microbiome pathways with simple but varied contributional diversity. Bar heights represent the total relative abundance of the pathway and are log-scaled. Contributions of individual species/other/unclassified are linearly scaled within the total bar height.



Supplementary Figure 14. Examples of subspecies-level functional variation (gene level). (**a**) Strains of *Lactobacillus jensenii* were well-represented in 21 HMP posterior fornix samples. At least two (2) subspecies-level clades appear to be present, defined by the presence of gene block a1 or a2 (highlighted). (**b**) Strains of *Eubacterium eligens* were well-represented in 51 HMP stool samples. At least three (3) subspecies-level clades appear to be present, defined by the presence/absence of gene blocks b1, b2, and b3 (highlighted).



Supplementary Figure 15. Example of potential niche-adapted subspecies of *Haemophilus haemolyticus*. Metagenomic "strains" (UniRef90 gene family presence/absence profiles) of this species differ across the three oral sites where it was detected. Right-side plots illustrate the coreness, variability, and site-specific enrichment of individual genes. Variability peaks at 1.0 for genes detected in exactly 50% of samples. Site-specific enrichment peaks at 1.0 when the gene is 100% prevalent in a focal site and 0% prevalent in all other sites (with -1 corresponding to the exact opposite scenario).



SUPPLEMENTARY NOTES

Supplementary Note 1: Reference hold-out analysis of a complex synthetic metagenome

Following the procedures described in the main text and **Methods**, we produced a 10M-read synthetic metagenome containing even contributions from 100 non-human-associated species (which equated to just over 2x genomic fold-coverage per species). Species were defined as "human-associated" if they appeared (with any abundance) in the MetaPhlAn2[6] profiles of Human Microbiome Project[7] metagenomes. The 100 species were selected at random, but constrained to species that were i) quantifiable by MetaPhlAn2 and ii) present in ChocoPhlAn and its underlying isolate genome catalogue. As in the synthetic human gut metagenome application described in the main text, isolate genomes were randomly mutated at 3% of positions to simulate novel isolates of known species.

Analysis of the complex synthetic metagenome in HUMAnN2's tiered mode produced estimates of gene family (UniRef90[8]) abundance. We compared those profiles to gold standards based on i) species' genomic fold-coverage values and ii) per-species, per-gene read sampling (with the latter accounting for the fact that some genes are sampled more or less at random). As expected, random sampling induces a non-trivial error that cannot be explained by HUMAnN2: even if HUMAnN2 "places" each read into the correct gene family, it is not aware that some of these families were over- or under-sampled during synthetic metagenome creation. Indeed, while most species' gene-level accuracy (1 - Bray-Curtis dissimilarity) ranged from ~0.8 to 0.9 using the fold-coverage-based gold standard, this range improved to ~0.9 to 1.0 when using the sampled-reads-based gold standard (**Supplementary Fig. 2**). Hence, even in this complex mock community, HUMAnN2's per-gene abundance estimation is performing close to optimally for the majority of species.

We further applied this complex synthetic metagenome to perform a reference hold-out analysis. This procedure evaluates how HUMAnN2's accuracy (and runtime) scale depending on the proportion of the community that is identified in the first tier of the tiered search (taxonomic prescreen). To accomplish this, we randomly selected subsets of the 100 species in the community as "unquantifiable" (deleting them from the MetaPhlAn2 taxonomic prescreen output). Reads from these species then pass directly to translated search and are not mapped to their corresponding pangenomes. We considered hold-out sets of size 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 species, with five replicates per hold-out set size (with the exception of 0 and 100). Larger hold-out sets correspond to more poorly-characterized communities (in which more alignment will be done during translated search). Conversely, smaller hold-out sets correspond to better-characterized communities (in which more alignment can in principle be done during the pangenome search).

Consistent with the synthetic human gut metagenome analysis from the main text, community-level accuracy was strongest when no species were held out of the pangenome search (85%) and least-accurate when all species were held out of the pangenome search, i.e. a pure translated search analysis (40%; **Supplementary Fig. 2**). We observed that accuracy decreased linearly for increasing numbers of held-out species, with relatively little variation across the sets of five replicates. This suggests that the accuracy of tiered search is more a function of the number of species identified during the taxonomic prescreen and not the specific nature of those species. Put another way, the loss of accuracy incurred by searching a species' reads against a comprehensive protein database (versus the species' own pangenome) appears to be fairly well conserved across these 100 species.

Also as expected, HUMAnN2's runtime increased with the number of held-out species, which equates to more reads being passed to the (slower) translated search tier of the tiered search (**Supplementary Fig. 2**). Runtime was shortest when none of the 100 species were held out (1.5 hours, 8 CPU cores) and longest when all were held out (pure translated search, 4.2 hours). As in the synthetic gut metagenome example from the main text, most time during tiered search was spent during the translated search phase, even for small numbers of held-out species. This underscores the incredible speed of nucleotide-level search against a sample-specific pangenome relative to comprehensive translated search.

Supplementary Note 2: Unclassified abundance in synthetic evaluations

Reads that failed to align to a pangenome in the second tier of HUMAnN2's tiered search pass to the final search tier: comprehensive translated search. Although the synthetic metagenomes used in our evaluations only included reads from mock isolate genomes of characterized species, it is not surprising for a subset of those reads to fail to map to the corresponding species' pangenomes and be passed to translated search (resulting in an unexpected "unclassified" stratum in the output): this will be the case for i) coding reads that were highly diverged from the corresponding sequence in the pangenome database, ii) coding reads that spanned multiple genes, and iii) reads drawn from (apparent) non-coding regions. Reads of the first type can still be assigned to correct gene families during translated search, which does not impact community-level error, but does increase stratified error. Reads of all three types were vulnerable to spurious mapping during translated search, resulting in false positive detection events (though HUMAnN2 is optimized to reduce such errors; see **Online Methods** and **Supplementary Fig. 5**). Passage of reads into translated search made a minority contribution (~11%) to the total absolute error of the stratified evaluation.

Supplementary Note 3: Additional HUMAnN2 performance considerations

HUMAnN2's tiered search was 3.0x faster than pure translated search in the synthetic gut metagenome evaluation and 2.4x faster in the synthetic complex metagenome evaluation (introduced in Supporting Note 1). Bypassing translated search entirely (which is reasonable as a first-pass in well-characterized communities) offered a further ~2x speedup in both scenarios. Notably, this is consistent with a 5-6x speedup for tiered search relative pure translated search in the limit where no reads are passed to translated search. A less-extreme approach for accelerating HUMAnN2 translated search is to limit its scope (database) to proteins annotated to metabolic reactions. This approach limits the coverage of "unclassified" gene families but not the coverage of "unclassified" pathway abundance. Using a reaction-filtered protein database offers close to a 2x speedup during tiered search of well-characterized communities and a >3x speedup of pure translated search.

HUMAnN2's translated search performance is similar when using UniRef90 versus UniRef50 databases (~1.5K reads processed / second), despite the former being ~2x larger. This is due to the fact that finding valid alignments to UniRef50 (>50% amino acid identity) requires a more sensitive search than finding valid alignments to UniRef90 (>90% identity). This in turn requires relaxing the seed parameters for alignments, which results in more spurious seed events, and thus considering a greater proportion of the UniRef50 database (relative to UniRef90). Using HUMAnN2 in "UniRef50 mode" will automatically adjust the translated search step to operate in this more sensitive manner (individual search parameters can also be independently tuned by the user). The taxonomic prescreen and pangenome search steps proceed identically in UniRef90 mode versus UniRef50 mode, with the appropriate species-stratified gene families (i.e. UniRef90 gene families versus UniRef50 gene families) included in the main output. "UniRef50" mode is potentially preferred when profiling poorly characterized communities. In such cases, more sequencing reads are expected to pass to translated search, where they may benefit from the relaxed UniRef50 homology thresholds.

We evaluated HUMAnN2 in UniRef50 mode using the synthetic gut metagenome introduced in the main text. In tiered mode, community-level sensitivity, precision, and overall accuracy (1 - Bray-Curtis distance) were similar to those calculated in the UniRef90 evaluation: 86%, 75%, and 86%. Precision and accuracy were slightly reduced due to the greater influence of spurious mapping during the translated search phase (a natural consequence of relaxing the stringency of the translated search in UniRef50 mode). This difference was more apparent when profiling with pure translated search, where overall accuracy was reduced to 40% from 67% in the UniRef90 evaluation. However, this difference is somewhat misleading, as it results in part from read mass being split between target UniRef50 gene families (i.e. those spiked into the synthetic metagenome) and close homologs: a further consequence of less-stringent translated search. However, when evaluating HUMAnN2's pure translated search on more coarse-grained functions (where homologous UniRef50s tend to sum to the same higher-level group), accuracy was higher and more in-line with UniRef90-based results. For example, when regrouping UniRef50 and UniRef90 abundance to KEGG Orthogroups (KOs)[9], overall accuracy was 77% in the UniRef50 evaluation and 81% in the UniRef90 evaluation.

Supplementary Note 4: Extended methods for HUMAnN2 applications

Analysis of HMP metagenomes

We profiled 397 quality-controlled shotgun metagenomes from the first wave of the Human Microbiome Project[7] (downloaded from http://hmpdacc.org). We selected samples to highlight the first sampling visit from HMP subjects at six major body sites (nares, buccal mucosa, supragingival plaque, tongue, gut, and posterior fornix). Longitudinal samples and technical replicates were excluded. These samples were profiled with HUMAnN2's tiered search under default settings using UniRef90 as a protein database. Mapping rates and performance of the tiered search are reported in **Supplementary Fig. 3**.

Relative to other human body sites, skin (nares) metagenomes contained smaller fractions of reads that were explainable during pangenome search (median ~25%) with a considerable boost in total mapping occurring during translated search (median ~60% of reads mapped after translated search). There are a number of possible explanations for this observation, including an abundance of DNA reads derived from i) uncharacterized cellular microbes, ii) DNA viruses (not included in this pangenome analysis), iii) residual human reads that survived host decontamination, and iv) other sources of non-microbial contamination (all of which would map during translated search but not during pangenome search). To distinguish among these possibilities, we examined the taxonomic distribution of gene families that recruited reads during translated search of nares samples (as inferred from LCA annotations of UniRef90 sequence families). Averaging over samples, 84.4% of unclassified mapping was assigned to bacteria-specific protein families, 2.5% to virus-specific protein families, and 7.0% to eukaryote-specific protein families [including 1.3% to Basidiomycota (a fungal phylum), 1.3% to Streptophyta (a plant phylum), and 4.0% to Chordata (the phylum to which humans belong)]; the remaining ~6% of unclassified abundance could not be assigned at the kingdom level. This suggests that the majority of reads mapped during translated search derived from previously uncharacterized bacteria, with a minority derived from viruses/contamination.

Analysis of Red Sea metagenomes

We applied HUMAnN2 to profile 45 metagenomes collected during the 2011 KAUST Red Sea Expedition[10]. First, MetaPhlAn2 was run on all samples as a “joint prescreen” for all samples. In other words, if a species was detected in any of the 45 metagenomes, then its pangenome was included in the analysis of all samples. This “dataset-specific” (versus sample-specific) pangenome was used to add robustness to missing MetaPhlAn2 marker genes, which we expect to be more common in less-well-characterized communities. 388 pangenomes were selected in this manner, including many examples from known marine clades, such as *Nitrospina*, *Nitrospira*, *Prochlorococcus*, *Synechococcus*, *Synechocystis*, *Candidatus* Pelagibacter, alphaproteobacterium (SAR11 clade), *Polymorphum*, and SAR116 cluster. Each sample was then profiled against the custom pangenome database and UniRef50 protein database via tiered search. UniRef50 abundances were regrouped to KO abundances using the “humann2_regroup_table” script and UniRef50-to-KO mapping to enable comparison with HUMAnN1-based profiles. KO abundance was normalized over KO-mapped reads during HUMAnN1-HUMAnN2 comparison and over total reads in all other applications.

Compared with earlier HUMAnN1 profiles of the same samples, HUMAnN2 explained similar fractions of reads (16.2–42.5% with mean 26.2% versus 16.6–35.5% with mean 25.5%). This likely reflects a balance between the larger databases underlying HUMAnN2 combined with its stricter mapping criteria. Of the 6,919 KOs identified by either method, 4,609 were identified by both (shared KOs), 1,166 by HUMAnN1 only, and 1,144 by HUMAnN2 only (including the examples discussed above). HUMAnN1-only KOs contained examples of “retired” KOs that are no longer annotated to UniRef sequences; the vast majority (>95%) of mapped reads aligned to shared KOs. Relative abundance estimates among shared KOs were extremely well correlated between HUMAnN 1 and 2, despite being derived from independent sequence collections (KEGG vs. UniRef50). Notably, there was a decrease in correlation strength at greater depths, particularly in samples from the Southern Red Sea. This depth- and latitude-dependent effect may arise from the poorer representation of mesopelagic (deep) taxa in the reference genome databases.

Analysis of strain-level functional variation

It is only meaningful to discuss variation in the contribution of a function by a species when that species is otherwise well-covered. Without this assurance, variable contribution could be driven by undersampling of the species (or simply species presence/absence). Taking inspiration from PanPhlAn[11], we examined stratified HUMAnN2 gene family profiles of HMP metagenomes for species that recruited reads to i) a large number (≥500) of UniRef90 gene families with ioi) reasonably *strong* coverage (median non-zero abundance ≥10 RPK), and iii) reasonably *consistent* coverage (3rd quartile / 1st quartile of non-zero abundances <2.5). We refer to such species as "well-represented" in a sample. The first two criteria ensure that a metagenomic strain's coverage of its corresponding pangenome is sufficiently high for zero-abundance gene families to be believed as "non-detect" events, while the third and final criterion ensures that a single, dominant strain of the species is present. We focused on strains of species that were well-represented in at least 10 metagenomes from a given body site. To add robustness to false positive gene calls, we binarized strains' continuous gene abundance values to presence/absence (1/0) data using the square root of the strain's median non-zero abundance as a threshold for "confident detection." These binary strain representations (and their constituent gene families) could then be compared/clustered across samples using Jaccard distance as a similarity measure.

To identify functional enrichments among variably encoded genes from a given species, we assigned each gene a score corresponding to its prevalence (after binarization) in samples where the species was well-represented. We then applied a simple functional enrichment analysis (term-positive versus term-negative Wilcoxon rank-sum test) to these scores, focusing on a reduced ("informative"[12]) set of GO Biological Process terms bundled with HUMAnN2 (these terms are all associated with ≥1,000 UniRef90 identifiers, but no term has a descendant term with ≥1,000 associations). Enrichments were carried out per-species and subjected to false discovery rate (FDR) correction[13].

Analysis of IBDMDB multi'omic data

We profiled paired metatranscriptomes and metagenomes from subjects enrolled in the iHMP IBD pilot (http://ibdmdb.org) using HUMAnN2 (version 0.7.0). Sequencing reads were processed with KneadData (version 0.5.1) including quality trimming (MAXINFO:80:0.5 MINLEN:50) and length filtering followed by decontamination. Specifically, all reads were filtered against the human genome (hg19 build), and metatranscriptomic reads were additionally filtered against the NCBI human genome RNA database and SILVA rRNA database[14]. Only meta'omes with at least 1 million reads were including for further analyses, which resulted in 78 paired metatranscriptomes and metagenomes. Metagenomes were analyzed by HUMAnN2 under default settings. Metatranscriptomes were similarly analyzed, but conditioned upon the species detected in their corresponding metagenomes (i.e. each DNA/RNA sample pair was mapped against the pangenomes of species detected in the DNA sample). We isolated pathways for diversity analysis following the same procedures described for HMP samples in the main text (i.e. pathways present in the majority of samples and with a low, average "unclassified" component to facilitate contributional diversity calculation).

Supplementary Note 5: Synthetic gut metatranscriptome evaluation

We simulated metatranscriptomes following a modified version of our protocol for simulating metagenomes (**Online Methods**). Instead of sampling synthetic fragments from entire isolate genomes, synthetic metatranscriptome fragments were sampled from isolate genomes' protein-coding regions only. Sampling from a given coding region ("transcript") was proportional to the product of i) the target metagenomic abundance of the transcript's source species, ii) the length of the transcript, and iii) a random draw from ln N(0, 1) (which adds log-normal variation in gene expression within-species). These proportions were further applied to produce gold-standard abundances for gene families within each synthetic metatranscriptome, which could then be summed to other transcript-level abundances (e.g. COGs).

We evaluated the six functional profiling methods considered in Fig. 1e on a synthetic gut metatranscriptome (**Supplementary Fig. 6**). This metatranscriptome incorporated the same human gut species and staggered species abundances as the synthetic gut metagenome from the main text (see **Fig. 1b**). Trends in COG quantification accuracy and method performance on the synthetic gut metatranscriptome were similar to those observed in the metagenomic evaluation. In particular, HUMAnN2's tiered search achieved sensitivity, precision, and accuracy of 89%, 100%, and 93%, respectively. Most methods (including HUMAnN2) exhibited slightly reduced sensitivity in the metatranscriptomic evaluation, which we can attribute to undersampling of low-abundance transcripts in low-abundance species.

Supplementary Note 6: Evaluations on novel isolate genomes

In order to evaluate functional profiling methods' robustness to novel genomes, we constructed two additional synthetic metagenomes from isolate genomes deposited in IMG[15] in 2017. These genomes are therefore newer than the methods and/or underlying databases being evaluated, including HUMAnN2's databases. We restricted our selection to genomes that were tagged as "finished" and "high quality" in IMG. The first of these synthetic metagenomes (the "novel strains" metagenome) was drawn from 20 novel isolates of species that were already represented in HUMAnN2's pangenome database (i.e. having one or more existing isolate genomes) and which could be correctly identified by MetaPhlAn2. The second of these synthetic metagenomes (the "novel species" metagenome) was drawn from 20 novel isolates of novel species, with "novel species" defined here as "not present in HUMAnN2's pangenome database" and "not classified to the species level by MetaPhlAn2." In both cases, we disallowed adding more than one new isolate from the same species. Subject to these constraints, the two sets of 20 isolate genomes were chosen at random. The novel strains metagenome provides an additional benchmark for functional profiling of a real-world community composed of known species, while the novel species metagenome provides a benchmark for a real-world uncharacterized community.

Reads were sampled from the novel isolate genomes following the procedures used to make the synthetic gut metagenome in the main text (**Online Methods**). Briefly, the species' relative abundances were geometrically staggered, and 5M genomic fragments were sampled from their genomes in proportion to relative abundance and genome size. Fragments were then converted to pairs of 100-nt sequencing reads (10M reads total, 1 Gnt) using ART[16]. Unlike the main-text evaluations, we did not spike additional synthetic biological sequence variation into these metagenomes, as they are already expected to contain natural biological sequence variation (relative to reference genomes in the tools' databases).

In order to make the "previously unseen" nature of these genomes the one new variable in these evaluations, we constructed their gold standards in the same way as before: annotating the new isolate genomes against UniRef90 (and Uniref50) by sequence alignment and then inferring COG annotations from UniProt[17] (following the procedure for pangenome annotation introduced in **Online Methods**). This annotation procedure worked well for the novel isolates of known species, with 83% of IMG-defined ORFs annotated to a UniRef90 family, and 91% annotated to a UniRef50 family. However, for the novel isolates of novel species, only 20% of ORFs could be assigned to UniRef90, along with a more reasonable 78% assigned to UniRef50s. Hence, while we used a UniRef90-based gold standard for all other evaluations, we concluded that this would not be optimal for the "novel species" metagenome. Instead, we constructed both UniRef90- and UniRef50-based COG gold standards for each new metagenome, and performed a separate evaluation on each. Compared to the UniRef90-based gold standard, the UniRef50-based gold standard offers better coverage but potentially worse specificity. A further coverage/specificity trade-off was offered by IMG's own COG annotations, which are based on best BLAST hits. In our view, this trade-off was not reasonable, as >50% of IMG's COG annotations were based on alignments with <40% amino acid identity. Treating such remote-homology annotations as a "gold standard" would reward best-BLAST-hit approaches to functional profiling, which we have previously shown to be suboptimal[1], while penalizing approaches that favor specificity.

In the context of the "novel isolates" metagenome (with a UniRef90-based gold standard), trends in accuracy and performance across the six functional profiling methods were similar to those observed for the synthetic gut metagenome comparison (see **Fig. 1d**). Specifically, HUMAnN2's tiered search was more accurate and 3x faster than its own pure translated search implementation, which was in turn more accurate than the four alternative translated search implementations (**Supplementary Fig. 7**). These results are due in part to HUMAnN2 i) identifying that the isolate genomes in this community belonged to known species, ii) mapping many of their reads to homologous sequences in the existing pangenomes, and iii) forwarding unmapped reads (e.g. from novel pangenome content) to translated search. Trends were similar in the UniRef50-based evaluation of this metagenome (**Supplementary Fig. 8**). The accuracy of all methods improved in this context, though improvements were more notable for the non-HUMAnN2 methods, which tend to favor sensitivity over specificity (and hence benefit from higher-coverage, relaxed-homology annotations).

In the evaluation on the "novel species" metagenome, HUMAnN2's tiered search did not identify any known species in the community (as expected), and thus forwarded all community reads to translated search. Hence, for this metagenome, HUMAnN2 in "tiered search" and "pure translated search" produce identical results in roughly the same amount of time. (Note that, while the taxonomic prescreen does not identify any known species in the tiered mode, it still contributes to total runtime. However, its contribution was negligible compared with the duration and variability of translated search execution.) For the sake of consistency, we performed an evaluation on this community using UniRef90-defined COGs, as in all preceding evaluations (**Supplementary Fig. 9**). While HUMAnN2 performed well in that context, precision-based comparisons must be considered critically due to the low coverage of UniRef90 annotations in this community (as discussed above). The high-coverage, UniRef50-based evaluation is easier to interpret (**Supplementary Fig. 10**). While HUMAnN2 remained among the most efficient methods in that evaluation, several methods (including HUMAnN2) produced reasonably accurate profiles. This finding is consistent with the fact that the UniRef50 gold standard was constructed by translated search with relaxed homology thresholds: mirroring the translated search methods used by the various profiling methods in this evaluation.

Supplementary Note 7: Comparison of HUMAnN2 with metagenomic assembly

Unlike reference-based approaches, metagenomic assembly is ideal for identifying novel gene content in a microbial community, as well as novel genomes and gene arrangements. However, relative to reference-based functional profiling, we expect assembly to struggle to reconstruct known genes from a metagenome that are not well-covered by sequencing reads. (This assertion applies even at very high read depth, where a fraction of community genes will recruit reads at <1x coverage, and hence be detectable by mapping but not assembly.) To evaluate this hypothesis, we applied a modern metagenomic assembler (metaSPAdes[18]) to the synthetic gut and complex metagenomes introduced above and in the main text (with modest but realistic read depths of 10M reads / 1 Gnt). We then i) called ORFs in the resulting contigs, ii) annotated those ORFs against UniRef90 following the procedures used to annotate HUMAnN2's pangenome database (with adjustments to avoid penalizing partially assembled ORFs), and iii) compared assembled versus expected gene families.

More specifically, we ran metaSPAdes using 8 CPU cores, consistent with the resources allotted to HUMAnN2 and the other reference-based profilers in their evaluations. Synthetic reads were provided to metaSPAdes as paired FASTQ files without prior filtering (again, equivalent to the evaluation of the reference-based profilers). Resulting contigs were passed to Prodigal[19] for open reading frame (ORF) calling in "-p meta" mode. We annotated called ORFs against UniRef90 using a modified version of the procedure we applied to annotate HUMAnN2's pangenome database (**Online Methods**). Specifically, translated ORFs were aligned against UniRef90 using DIAMOND[20]. The best hit (if any) with >90% amino acid alignment identity, >80% coverage of the ORF, and >50% coverage of the UniRef90 representative sequence was assigned as an annotation for the ORF. Note that the required coverage of the UniRef90 representative sequence (50%, versus 80% during pangenome annotation) was relaxed to avoid penalization of partially assembled ORFs. We explored a modified version of this protocol in which ORFs were annotated to *all* UniRef90 representatives that matched the above conditions (instead of just the best match): this ultimately tended to reduce precision with only minor gains in sensitivity (~1%), and so the best-match approach was favored in the evaluation. Species-specific genes were defined on a per-metagenome basis as ORFs that mapped uniquely to a particular UniRef90 family.

Assembling and annotating the synthetic gut metagenome took 3.4 hours using 8 CPU cores (~4x longer than HUMAnN2’s tiered search) but with surprisingly low peak memory use (Max RSS, 7.4 GB). Precision was high at 85%, indicating that relatively few ORFs were assigned to incorrect gene families. However, sensitivity was low at 31% (compared with HUMAnN2’s 91% on the same dataset), indicating that many expected UniRef90 families were not successfully reconstructed. On a per-genome level, 5% of species-specific genes were identified from a genome at ~1x fold-coverage, increasing to 69% for a genome at 35x fold-coverage (compared with 93-100% for HUMAnN2). Thus, while sensitivity of assembly was particularly poor at low coverage, recovering full genomes from the synthetic metagenome was challenging even at high fold-coverage.

In the 100-species complex synthetic metagenome, precision of assembly was similarly high at 92% and sensitivity low at 23%. However, performance was considerably worse (10.0 hours runtime; 30.3 GB peak memory use), despite the input dataset being of the same size (10M reads, 1 Gnt). This is in stark contrast to HUMAnN2's performance, which scaled more predictably with input size: real-world runtime was roughly linear in the input size, and memory use scaled sub-linearly, with small metagenomes (up to 10M reads, 1 Gnt) requiring <16 GB of memory and larger metagenomes (up to 100M reads, 10 Gnt), requiring <32 GB (**Supplementary Fig. 3**). In contrast, assembling large metagenomes can require 100s of GBs of memory: necessitating specialized hardware and limiting the number of samples that can be processed in parallel in cluster environments. Thus, for profiling known genomes and gene families, HUMAnN2 can be considerably more accurate and resource-efficient than an assembly-based approach.

REFERENCES (SUPPLEMENTARY)

1. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology* **8**, e1002358 (2012).

2. Bose, T., Haque, M.M., Reddy, C. & Mande, S.S. COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. *PloS one* **10**, e0142102 (2015).

3. Huson, D.H. et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS computational biology* **12**, e1004957 (2016).

4. Nayfach, S. et al. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. *PLoS computational biology* **11**, e1004573 (2015).

5. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 41 (2003).

6. Truong, D.T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods* **12**, 902-903 (2015).

7. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).

8. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. & Wu, C.H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)* **31**, 926-932 (2015).

9. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457-462 (2016).

10. Thompson, L.R. et al. Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *The ISME journal* (2016).

11. Scholz, M. et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature methods* **13**, 435-438 (2016).

12. Zhou, X., Kao, M.C. & Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12783-12788 (2002).

13. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 289-300 (1995).

14. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41**, D590-596 (2013).

15. Markowitz, V.M. et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic acids research* **40**, D115-122 (2012).

16. Huang, W., Li, L., Myers, J.R. & Marth, G.T. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* **28**, 593-594 (2012).

17. UniProt Consortium. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-212 (2015).

18. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new versatile metagenomic assembler. *Genome research* (2017).

19. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 119 (2010).

20. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60 (2015).