Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3

Francesco Beghini [*,1], Lauren J. McIver [*,2], Aitor Blanco-Míguez [1], Leonard Dubois [1], Francesco Asnicar [1], Sagun Maharjan [2,3], Ana Mailyan [2,3], Andrew Maltez Thomas [1], Paolo Manghi [1], Mireia Valles-Colomer [1], George Weingart [2,3], Yancong Zhang [2,3], Moreno Zolfo [1], Curtis Huttenhower [^,2,3], Eric A. Franzosa [^,2,3], Nicola Segata [^,1,4]

1. Department CIBIO, University of Trento, Italy
2. Harvard T.H. Chan School of Public Health, Boston, MA, USA
3. The Broad Institute of MIT and Harvard, Cambridge, MA, USA
4. IEO, European Institute of Oncology IRCCS, Milan, Italy
* Joint first authors
^ Joint senior authors
Correspondence to: chuttenh@hsph.harvard.edu, franzosa@hsph.harvard.edu, nicola.segata@unitn.it

Abstract

Culture-independent analyses of microbial communities have advanced dramatically in the last decade, particularly due to advances in methods for biological profiling via shotgun metagenomics. Opportunities for improvement continue to accelerate, with greater access to multi-omics, microbial reference genomes, and strain-level diversity. To leverage these, we present bioBakery 3, a set of integrated, improved methods for taxonomic, strain-level, functional, and phylogenetic profiling of metagenomes newly developed to build on the largest set of reference sequences now available. Compared to current alternatives, MetaPhlAn 3 increases the accuracy of taxonomic profiling, and HUMAnN 3 improves that of functional potential and activity. These methods detected novel disease-microbiome links in applications to CRC (1,262 metagenomes) and IBD (1,635 metagenomes and 817 metatranscriptomes). Strain-level profiling of an additional 4,077 metagenomes with StrainPhlAn 3 and PanPhlAn 3 unraveled the phylogenetic and functional structure of the common gut microbe *Ruminococcus bromii*, previously described by only 15 isolate genomes. With open-source implementations and cloud-deployable reproducible workflows, the bioBakery 3 platform can help researchers deepen the resolution, scale, and accuracy of multi-omic profiling for microbial community studies.

Introduction

Studies of microbial community biology continue to be enriched by the growth of culture-independent sequencing and high-throughput isolate genomics (Almeida et al., 2020, 2019; Forster et al., 2019; Parks et al., 2017; Pasolli et al., 2019; Poyet et al., 2019; Zou et al., 2019). Shotgun metagenomic and metatranscriptomic (i.e. “meta-omic”) measurements can be used to address an increasing range of questions as diverse as the transmission and evolution of strains in situ (Asnicar et al., 2017; Ferretti et al., 2018; Truong et al., 2017; Yassour et al., 2018), the mechanisms of multi-organism biochemical responses in the environment (Alivisatos et al., 2015; Blaser et al., 2016), or the epidemiology of the human microbiome for biomarkers and therapy (Gopalakrishnan et al., 2018; Le Chatelier et al., 2013; Thomas et al., 2019; Zeller et al., 2014). Using such analyses for accurate discovery, however, requires efficient ways to integrate hundreds of thousands of (potentially fragmentary) isolate genomes with community profiles to detect novel species and strains, non-bacterial community members, microbial phylogeny and evolution, and biochemical and molecular signaling mechanisms. Correspondingly, this computational challenge has necessitated the continued development of platforms for the detailed functional interpretation of microbial communities.

The past decade of metagenomics has seen remarkable growth both in the biology accessible via high-throughput sequencing and in the methods for doing so. Beginning with the now-classic questions of “who’s there?” and “what are they doing?” in microbial ecology (Human Microbiome Project Consortium, 2012), shotgun metagenomics provide a complementary means of taxonomic profiling to amplicon-based (e.g. 16S rRNA gene) sequencing, as well as functional profiling of genes or biochemical pathways (Morgan et al., 2013; Quince et al., 2017; Segata et al., 2013). More recently, metagenomic functional profiles have been joined by metatranscriptomics to also capture community regulation of gene expression (Lloyd-Price et al., 2019). Methods have been developed to focus on all variants of particular taxa of interest within a set of communities (Pasolli et al., 2019), to discover new variants of gene families or biochemical activities (Franzosa et al., 2018; Kaminski et al., 2015), or to link the presence and evolution of closely related strains within or between communities over time, space, and around the globe (Beghini et al., 2017; Karcher et al., 2020; Tett et al., 2019). Critically, all of these analyses (and the use of the word “microbiome” throughout this manuscript) are equally applicable to both bacterial and non-bacterial community members (e.g. viruses and eukaryotes) (Beghini et al., 2017; Olm et al., 2019; Yutin et al., 2018). Finally, although not addressed in depth by this study, shotgun meta-omics have increasingly also been combined with other community profiling techniques such as metabolomics (Heinken et al., 2019; Lloyd-Price et al., 2017; Sun et al., 2018) and proteomics (Xiong et al., 2015) to provide richer pictures of microbial community membership, function, and ecology.

Methods enabling such analyses of meta-omic sequencing have developed in roughly two complementary types, either relying on metagenomic assembly or using largely assembly-independent, reference-based approaches (Quince et al., 2017). The latter is especially supported by the corresponding growth of fragmentary, draft, and finished microbial isolate genomes, and their consistent annotation and clustering into genome groups and pan-genomes (Almeida et al., 2020, 2019; Pasolli et al., 2019). Most such methods focus on addressing a single profiling task within (most often) metagenomes, such as taxonomic profiling (Lu et al., 2017; Milanese et al., 2019; Truong et al., 2015; Wood et al., 2019), strain identification (Luo et al., 2015; Nayfach et al., 2016; Scholz et al., 2016; Truong et al., 2017), or functional profiling (Franzosa et al., 2018; Kaminski et al., 2015; Nayfach et al., 2015; Nazeen et al., 2020). In a few cases,

platforms such as the bioBakery (McIver et al., 2018), QIIME 2 (Bolyen et al., 2019), or MEGAN (Mitra et al., 2011) integrate several such methods within an overarching environment. While not a primary focus of this study, metagenomic assembly methods enabling the former types of analyses (e.g. novel organism discovery or gene cataloging (Lesker et al., 2020; Stewart et al., 2019)) have also advanced tremendously (Li et al., 2015; Nurk et al., 2017) and are now reaching a point of integrating microbial community and isolate genomics, particularly for phylogeny (Asnicar et al., 2020; Zhu et al., 2019). These efforts have also led to increased consistency in microbial systematics and phylogeny, facilitating the types of automated, high-throughput analyses necessary when manual curation cannot keep up with such rapid growth (Asnicar et al., 2020; Chaumeil et al., 2019).

Here, to further increase the scope of feasible microbial community studies, we introduce a suite of updated and expanded computational methods in a new version of the bioBakery platform. The bioBakery 3 includes updated sequence-level quality control and contaminant depletion guidelines (KneadData), MetaPhlAn 3 for taxonomic profiling, HUMAnN 3 for functional profiling, StrainPhlAn 3 and PanPhlAn 3 for nucleotide- and gene-variant-based strain profiling, and PhyloPhlAn 3 for phylogenetic placement and putative taxonomic assignment of new assemblies (metagenomic or isolate). Most of these tools leverage an updated ChocoPhlAn 3 database of systematically organized and annotated microbial genomes and gene family clusters, newly derived from UniProt/UniRef (Suzek et al., 2007) and NCBI (NCBI Resource Coordinators, 2014). Our quantitative evaluations show each individual tool to be more accurate and, typically, more efficient than its previous version and other comparable methods, increasing sensitivity and specificity by sometimes more than 2-fold (e.g. in non-human-associated microbial communities). Biomarker identifications in 1,262 colorectal cancer (CRC) metagenomes, 1,635 inflammatory bowel disease (IBD) metagenomes, and 817 metatranscriptomes show both the platform's efficiency and its ability to detect hundreds of species and thousands of gene families not previously profiled. Finally, in 4,077 human gut metagenomes containing *Ruminococcus bromii*, the bioBakery 3 platform permits an initial integration of assembly- and reference-based metagenomics, discovering a novel biogeographical and functional structure within the clade's evolution and global distribution. All components are available as open-source implementations with documentation, source code, and workflows enabling provenance, reproducibility, and local or cloud deployment at http://segatalab.cibio.unitn.it/tools/biobakery and http://huttenhower.sph.harvard.edu/biobakery.

Results

The bioBakery provides a complete meta-omic tool suite and analysis environment, including methods for individual meta-omic (and other microbial community) processing steps, downstream statistics, integrated reproducible workflows, standardized packaging and documentation via open-source repositories (GitHub, Conda, PyPI, and R/Bioconductor), grid- and cloud-deployable images (AWS, GCP, and Docker), online training material and demonstration data, and a public community support forum. For any sample set, quality control, taxonomic profiling, functional profiling, strain profiling, and resulting data products and reports can all be generated with a single workflow, while maintaining version control and provenance logging. All of the methods themselves, the associated training material, quality control using KneadData, and packaging for distribution and use have been updated in this version. For example, Docker images have been scaled down in size to optimize use in cloud environments, and workflows have been ported to AWS (Amazon Web Services) Batch and Terra/Cromwell (Google Compute Engine) to reduce costs through the use of spot and pre-emptive instances, respectively. All base images and dependencies have been updated as well, including the most recent Python (v3.7+) and R (v4.0+, see **Methods**). New and updated documentation of all tools, including detailed instructions on installation in different environments and package managers, is available at http://huttenhower.sph.harvard.edu/biobakery.

High-quality reference sequences for improved meta-omic profiling

The majority of methods within the bioBakery 3 suite leverage a newly-updated reference genome and gene cataloging procedure, the results of which are packaged as ChocoPhlAn 3 (**Fig. 1A**) (McIver et al., 2018). ChocoPhlAn uses publicly available genomes and standardized gene calls and gene families to generate markers for taxonomic and strain-level profiling of metagenomes with MetaPhlAn 3, StrainPhlAn 3, and PanPhlAn 3, phylogenetic profiling of genomes and MAGs with PhyloPhlAn 3, and functional profiling of metagenomes with HUMAnN 3.

ChocoPhlAn 3 is based on a genomic repository of 99.2k high-quality, fully annotated reference microbial genomes from 16.8k species available in the UniProt Proteomes portal as of January 2019 (UniProt Consortium, 2019) and the corresponding functionally-annotated 87.3M UniRef90 gene families (Suzek et al., 2015). From this resource, ChocoPhlAn initially generates annotated species-level pangenomes associating each microbial species with its sequenced genomes and repertoire of UniRef-based gene (nucleotide) and protein (amino acid sequence) families. These pangenomes provide a uniform shared resource for subsequent profiling across bioBakery 3. HUMAnN 3 and PanPhlAn 3 are directly based on complete pangenomes for overall functional and strain profiling, whereas other tools use additional information annotated onto the catalog. PhyloPhlAn 3 focuses on the subset of conserved core gene families (i.e. present in almost all strains of a species) for inferring accurate phylogenies, and both MetaPhlAn 3 and StrainPhlAn 3 further refine core gene families into species-specific unique gene families to generate unambiguous markers for metagenomic species identification and strain-level genetic characterization.

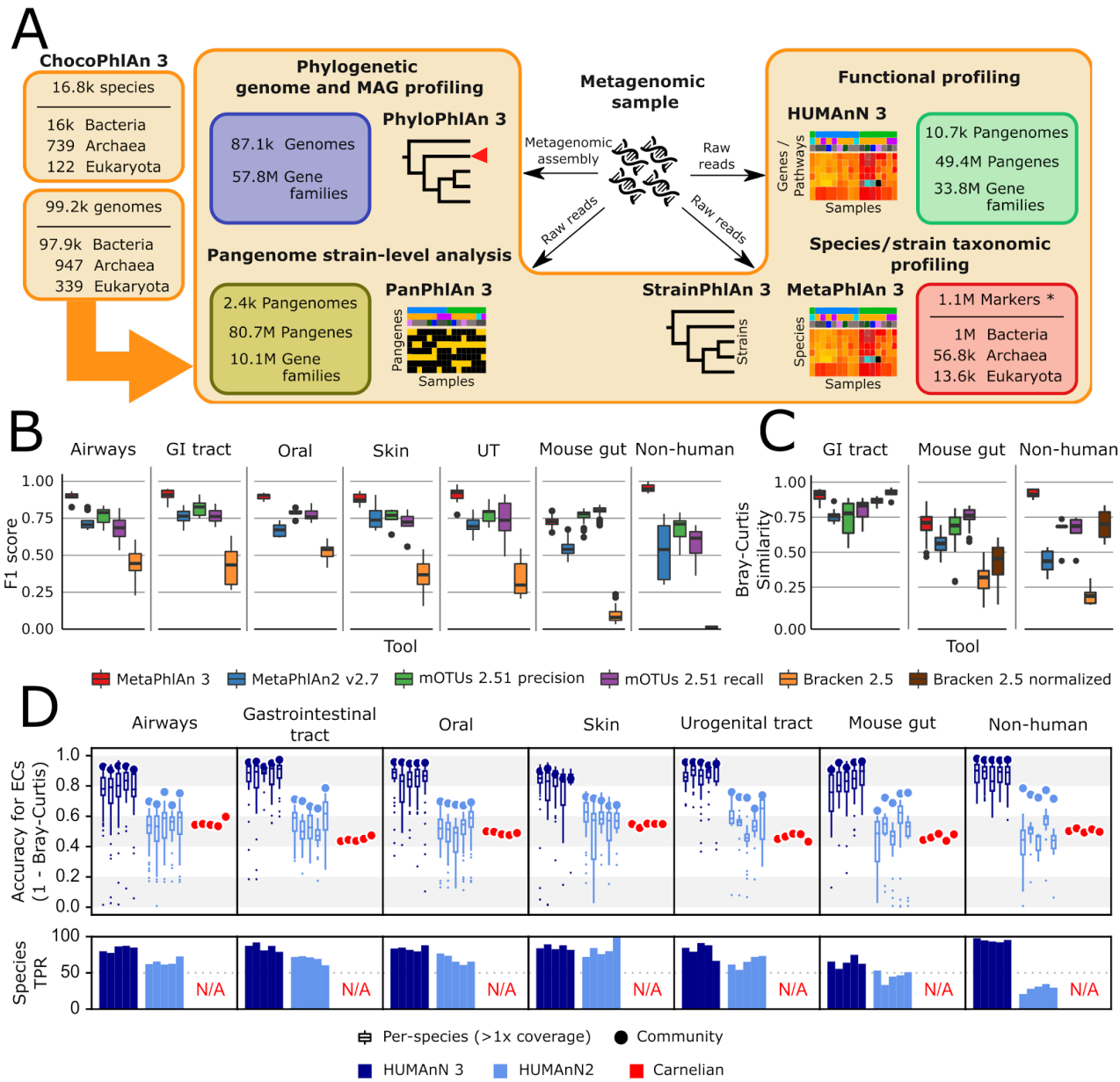

Figure 1: bioBakery 3 includes new microbial community profiling approaches that outperform previous versions and current methods. (A) The newly developed ChocoPhlAn 3 consolidates, quality controls, and annotates isolate-derived reference sequences to enable metagenomic profiling in subsequent bioBakery methods. (*The 1.1M MetaPhlAn 3 markers also comprise for 61.8k viral markers from MetaPhlAn 2 (Truong et al., 2015); other version descriptions in (Asnicar et al., 2020; Scholz et al., 2016; Truong et al., 2017)) **(B)** MetaPhlAn 3 was applied to a set of 113 total evaluation datasets provided by CAMI (Fritz et al., 2019) representing diverse human-associated microbiomes and 5 datasets of non-human-associated microbiomes (**Table S1**). MetaPhlAn 3 showed increased performance compared with the previous version MetaPhlAn 2 (Truong et al., 2015), mOTUs2 (Milanese et al., 2019), and Bracken 2.5 (Lu et al., 2017). We report here the F1 scores (harmonic mean of the species-level precision and recall, see **Fig. S1** for other evaluation scores). **(C)** MetaPhlAn 3 better recapitulates relative abundance profiles both from human and murine gastrointestinal metagenomes as well from non-human-associated communities compared to the other currently available tools (full results in **Fig. S1**). Bracken is reported both using its original estimates based on the fraction of reads assigned to each taxon and after re-normalizing them using the genome lengths of the taxa in the gold standard to match the taxa abundance estimate of the other tools. **(D)** Compared with HUMAnN 2 (Franzosa et al., 2018) and Carnelian (Nazeen et al., 2020), HUMAnN 3 produces more accurate estimates of EC abundances and displays a higher species true positive rate compared to HUMAnN 2.

MetaPhlAn 3 increases the accuracy of quantitative taxonomic profiling

MetaPhlAn estimates the relative abundance of microbial taxa in a metagenome using the coverage of clade-specific marker genes (Segata et al., 2012; Truong et al., 2015). Such marker genes are chosen so that essentially all of the strains in a clade (species or otherwise) possess such genes, and at the same time no other clade contains orthologs close enough to incorrectly map metagenomic reads. MetaPhlAn 3 incorporates 13.5k species (more than twice than MetaPhlAn 2) with a completely new set of 1.1M marker genes (84±47 mean±SD markers per species) selected by ChocoPhlAn 3 from the set of 16.8k species pangenomes. The adoption of UniRef90 gene families permitted the efficient expansion of the core-gene identification procedure, which is followed by a mapping of potential core genes against all available whole microbial genomes to ensure unique marker identification (see **Methods**). This restructuring of the marker selection process has been combined with several improvements and extensions of the algorithm, including optimized quality control during marker alignments and an estimation of the metagenome fraction composed of unknown microbes (**Table S2**).

We evaluated the taxonomic profiling performance of MetaPhlAn 3 using 118 synthetic metagenomes spanning 113 synthetic samples from the 2nd CAMI Challenge (Fritz et al., 2019; Sczyrba et al., 2017) through the OPAL benchmarking framework (Meyer et al., 2019) These represent typical microbiomes from five human-associated body sites and the murine gut, and we complemented them with 5 additional newly-generated synthetic non-human-associated metagenomes (see **Methods**). In addition to MetaPhlAn 3, the comparative evaluation considered MetaPhlAn 2.7 (Truong et al., 2015), mOTUs 2.51 (Milanese et al., 2019) (latest database available as of July 2020), and Bracken 2.5 (using a database built after the April 2019 RefSeq release) (Lu et al., 2017; Wood et al., 2019). These three profiling tools have consistently been shown to outperform other methods across multiple evaluations (McIntyre et al., 2017; Meyer et al., 2019; Milanese et al., 2019; Sczyrba et al., 2017; Truong et al., 2015; Ye et al., 2019).

MetaPhlAn 3 outperformed all the other profilers across all considered types of communities when assessing the F1 score (**Fig. 1B**), which is a measure combining the fraction of species actually present in the metagenomes that are correctly detected (recall, **Fig. S1**) and the fraction of species predicted to be present that were actually included in the synthetic metagenome (precision, **Fig. S1**). With a very low number of false positive species detected, MetaPhlAn 3 (avg 8.51 s.d. 5.12) also maximized precision (**Fig. S1**) with respect to the other tools (avg 9 s.d. 4.78 for mOTUs in high precision mode, the closest competitor on precision). On recall, Bracken and mOTUs in high-recall mode were in several cases superior to MetaPhlAn 3, but at the cost of a very high number of false positives (on average 729 species for Bracken and 39 for mOTUs high-recall, for a total of 86,077 and 4,655 false positive species across the synthetic metagenomes). MetaPhlAn can further minimize false positives by requiring a higher fraction of positive markers for positive species ("--stat_q" parameter, **Fig. S2**), but overall the F1 measure with default settings remains higher than the other evaluated tools across the panel of synthetic metagenomes in our evaluation.

In addition to more accurate species detection, MetaPhlAn 3 also quantified taxonomic abundance profiles more accurately compared to MetaPhlAn 2, mOTUs2, and Bracken based on Bray-Curtis dissimilarities in most datasets (**Table S3**, **Fig. 1C**). While it was slightly outperformed by mOTUs (only in high-recall mode) on the synthetic mouse gut dataset, even in this case, correlation-based measures (Pearson Correlation Coefficient between estimated and expected relative abundances) found MetaPhlAn 3 to be more accurate (r=0.73) than the other considered profilers (MetaPhlAn 2 r=0.63, mOTUs2 precision r=0.60, mOTUs2 recall r=0.71, Bracken r=0.43). Additionally, because Bracken estimates the fraction of reads belonging to each taxon rather than its relative abundance,

we also re-normalized its estimates based on genome length of the target species. This improved Bracken's performance on taxonomic abundances (but not false positives or false negatives, see **Methods**), but even so they were comparable with MetaPhlAn 3 in only some of the simulated environments (**Fig. S1**). Overall, this confirms that MetaPhlAn 3 is superior to its previous version and is more accurate than other currently available tools in the large majority of simulated environment-specific datasets.

In addition to improvements in accuracy, MetaPhlAn 3's computational efficiency also compares favorably with alternatives and with its previous version. It is >3x faster than MetaPhlAn 2 (10.0k vs. 2.9k reads/second on a Xeon Gold 6140) and almost matches the speed of Bracken (11k reads per second). MetaPhlAn 3 memory usage is slightly higher (2.6Gb for a complete taxonomic profiling run) than MetaPhlAn 2 (2.1Gb), but outperforms the other methods (4.3 Gb for mOTUs and 32.5 Gb for Bracken, **Fig. S2**, **Table S4**).

HUMAnN 3 accurately quantifies species' contributions to community function

HUMAnN 3 functionally profiles genes, pathways, and modules from metagenomes, now using native UniRef90 annotations from ChocoPhlAn species pangenomes. We compared its performance against HUMAnN 2 (Franzosa et al., 2018), and the recently published Carnelian (Nazeen et al., 2020) when profiling the 30 CAMI and 5 additional synthetic metagenomes introduced above (see **Methods** and **Fig. 1**). Carnelian was selected because it was published subsequent to HUMAnN 2 and, more importantly, follows the HUMAnN strategy of estimating the relative abundance of molecular functions directly from shotgun meta-omic sequencing reads rather than assembled contigs (albeit by a different approach). While HUMAnN 2 and 3 can both natively estimate the relative abundances of a wide variety of functional features from a metagenome (by first quantifying and then manipulating UniRef90 or UniRef50 abundances), we selected level-4 enzyme commission (EC) categories as a basis for comparison with Carnelian, as the method's authors provided a precomputed index for EC quantification (Nazeen et al., 2020).

HUMAnN 3 produced highly accurate estimates of community-level EC abundances across the 30 CAMI metagenomes (mean ± SD of Bray-Curtis similarity=0.93 ± 0.03, **Fig. 1**). HUMAnN 2 followed with an accuracy of 0.70 ± 0.04 and Carnelian at 0.49 ± 0.04. While HUMAnN 3 benefits in part from access to a more up-to-date sequence database, we note that HUMAnN 2's database (c. 2014) predates the Carnelian method by several years, and so recency cannot be the only explanation for this trend. For example, Carnelian uses a mean sequence length per EC during abundance estimation, a choice which may contribute additional error relative to HUMAnN's sum over per-sequence estimates. We observed similar trends in accuracy among the three methods using F1 score to prioritize presence/absence calls over abundance (**Fig. S3**). HUMAnN 2 and Carnelian were notably similar with respect to sensitivity (0.72 ± 0.05 vs. 0.74 ± 0.04, respectively) but not precision (0.95 ± 0.02 vs. 0.60 ± 0.08). This difference is attributable in part to HUMAnN's use of database sequence coverage filters (see **Methods**) to reduce false positives, an approach introduced for translated search in HUMAnN 2 and expanded to nucleotide search in HUMAnN 3 (**Fig. S4**).

One of the main advantages of HUMAnN 3 (and 2) compared with other functional profiling systems (including Carnelian) is their ability to stratify community functional profiles according to contributing species. This feature is additionally more accurate and useful in HUMAnN 3 as a function of its broader pangenome catalog. Across the CAMI metagenomes, EC accuracy for species with at least 1x mean coverage depth was 0.81 ± 0.16 for HUMAnN 3 and 0.51 ± 0.15 for HUMAnN 2 (mean ± SD within-species Bray-Curtis similarity; **Fig. 1**). HUMAnN 3 (via MetaPhlAn

3) additionally tended to detect more expected species in this coverage range compared with HUMAnN 2, a major driver of its improved community-level accuracy. As previously noted (Franzosa et al., 2018), HUMAnN's within-species function sensitivity is naturally lower for species below 1x coverage in a sample, as many of their genes will not have been sampled at all during the sequencing process. Per-species precision, however, remained high with HUMAnN independent of coverage and, following refinements in alignment post-processing, was slightly improved in v3 compared with v2 (0.95 ± 0.08 vs. 0.91 ± 0.07).

Carnelian was the most computationally efficient of the three methods, analyzing the CAMI metagenomes in 26.4 ± 2.7 CPU-hours (mean ± SD) compared with 38.1 ± 12.8 CPU-hours for HUMAnN 2 and 52.5 ± 19.2 CPU-hours for HUMAnN 3 (**Fig. S3**). Trends in peak memory use (MaxRSS) were similar, with Carnelian requiring 11.9 ± 0.0 GB versus HUMAnN 2's 17.0 ± 0.3 GB and HUMAnN 3's 21.5 ± 1.9 GB. We attribute these differences in large part to the sizes of the sequence spaces over which the methods search: while Carnelian focuses only on a subset of sequences annotatable to EC terms, HUMAnN aims to first quantify 10s of millions of unique UniRef90s, of which only 12.5% are ultimately annotated by ECs. The increased runtime of HUMAnN 3 compared to HUMAnN 2 is likewise attributable to the former's larger translated search database (87.3M vs. 23.9M UniRef90 sequences), as the translated search tier is the rate-limiting step of the HUMAnN algorithm even when most sample reads are explained in the preceding nucleotide-level search tiers (**Fig. S5**). This phenomenon also explains the greater runtime variability of HUMAnN, as runtimes vary inversely with the (a priori unknown) fraction of sample reads explained before the translated search tier (Franzosa et al., 2018). Notably, by bypassing the translated search step, HUMAnN 3 could explain the majority of CAMI metagenomic reads (70.9 ± 9.6% per sample) in only 5.8 ± 0.8 CPU-hours (a 9x speed-up; **Fig. S5**), although this is generally only appropriate for communities known to be well-covered by related reference sequences.

Evaluations on a set of synthetic metagenomes enriched for non-human-associated species resulted in similar relative accuracy and efficiency trends among the three methods (**Fig. 1** and **Fig. S3**). Hence, HUMAnN 3's strong performance is not restricted to microbial communities assembled from host-associated species. Moreover, MetaPhlAn 3's improved sensitivity for non-host-associated species increased both the accuracy and performance of HUMAnN 3 relative to HUMAnN 2 (by enabling a larger fraction of reads to be explained during the faster and more accurate pangenome search step). Finally, we evaluated HUMAnN 3's accuracy at the level of individual UniRef90 protein families (**Fig. S5**). As previously noted (Franzosa et al., 2018), the challenge of differentiating globally homologous UniRef90 protein sequences using short sequencing reads results in a reduction of community and per-species accuracy relative to broader gene families. However, because these homologs tend to share similar functional annotations, this error is smoothed out when individual UniRef90 abundances are combined in HUMAnN's downstream steps (as seen in the EC-level evaluation; **Fig. 1**).

MetaPhlAn 3 and HUMAnN 3 expand the link between the microbiome and colorectal cancer with a meta-analysis of 1,262 metagenomes

To illustrate the potential of bioBakery 3's updated profiling tools and to extend our understanding of the microbial signatures in colorectal cancer (CRC), we expanded our previous work to meta-analyze both existing and newly available CRC metagenomic cohorts for a total of 1,262 samples (600 control and 662 CRC samples) from 9 different datasets spanning 8 different countries (Feng et al., 2015; Gupta et al., 2019; Thomas et al., 2019; Vogtmann et al., 2016; Wirbel et al., 2019; Yachida et al., 2019; Yu et al., 2017; Zeller et al., 2014). The resulting integrated

profiles are available for download (**Table S5**) and included in the new release of curatedMetagenomicData (Pasolli et al., 2017).

MetaPhlAn 3 identified a total of 1,083 species detected at least once (172 considered "prevalent" when defined as present in >5% of samples at >0.1% relative abundance), of which 505 species (52 prevalent) were previously not reported by MetaPhlAn 2 due to the expansion of the genome database (or in some cases because of changes in the NCBI taxonomy). In addition, 82 species present in the MetaPhlAn 2 database were not detected by MetaPhlAn 2 but are now identified in the samples by MetaPhlAn 3, due to the expanded sequence catalog, improved marker discovery procedure, and increased sensitivity to low-abundance species (Thomas et al., 2019).

We found 121 species significantly associated with CRC (FDR q<0.05 and Q-test for heterogeneity >0.05; **Table S6**) by a meta-analysis of standardized mean differences using a random-effects model on arcsine-square-root-transformed relative abundances (see **Methods**). Association coefficients were also concordant with previous MetaPhlAn 2-based results using a fraction of the samples (Thomas et al., 2019), including the three species with the highest effect sizes: *Fusobacterium nucleatum*, *Parvimonas micra*, and *Gemella morbillorum*. We also identified three additional species not present in the previous MetaPhlAn 2 database that were among those most strongly associated with CRC (effect size >0.35): *Dialister pneumosintes*, *Ruthenibacterium lactatiformans*, and *Eisenbergiella tayi* (**Fig. 2B**, **Fig. S6**, **Fig. S7A**). Among these species, *Dialister pneumosintes* is typically oral, further reinforcing the role of oral taxa in CRC, and *R. lactatiformans* was reported as part of a consortium of bacteria able to increase colonic IFNγ+ T-cells (Tanoue et al., 2019). The expanded number of species detectable by MetaPhlAn 3 also strengthened the previously-observed pattern of increased richness in CRC-associated microbiomes - in contrast to the stereotype of decreased diversity during dysbiosis - in large part due to low-level addition of typically oral microbes to the baseline gut microbiome (**Fig. 2C**).

Functional profiling of this expanded CRC meta-analysis with HUMAnN 3 identified 4.3M UniRef90 gene families, corresponding to 549 MetaCyc pathways and 2,895 ECs. 120 MetaCyc pathways were significantly associated with CRC (Wilcoxon rank-sum test FDR q<0.05 and Q-test for heterogeneity >0.05) (**Fig. S7B**), of which 59 (49.1%) overlapped previous results, including e.g. the increased abundance of starch degradation V (**Table S6**) in healthy individuals. This pathway encodes functions for extracellular breakdown of starch by an amylopullulanase enzyme, which has both pullulanase and α-amylase activity (Flint et al., 2012). *Bifidobacterium breve* and other *Bifidobacterium* spp have been shown to encode amylopullulanases and attach to starch particles, and they have also been reported for their potential protective role against carcinogenesis here and previously (Sivan et al., 2015). Among the 20 disease-associated pathways with the highest significance, only 3 were present in the previous meta-analysis, with the majority exhibiting significant heterogeneity in the random effects model (possibly due to the inclusion of additional geographically distinct cohorts here). Large and diverse cohorts combined with improved taxonomic and functional profiling available via bioBakery 3 thus have the possibility to extend and refine microbiome biomarkers in CRC and other conditions.

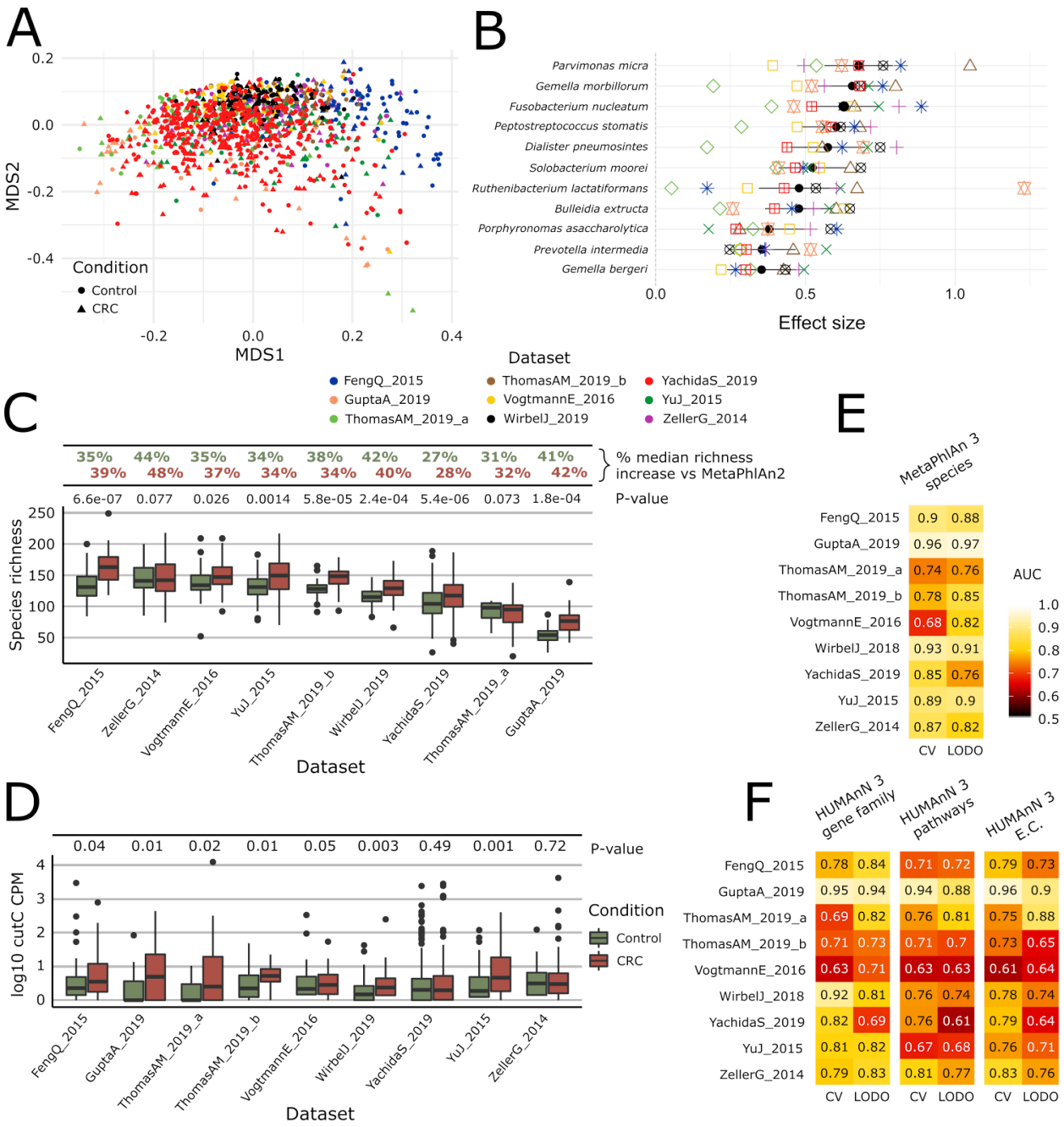

Figure 2: Meta-analysis with MetaPhlAn 3 and HUMAnN 3 expands taxonomic and functional associations with the CRC microbiome. (A) We considered a total of nine independent datasets (1,262 total samples) that highly but not completely overlap in composition based on ordination (multidimensional scaling) of weighted UniFrac distances (Lozupone and Knight, 2005) computed from the MetaPhlAn 3 species relative abundances. **(B)** Meta-analysis based on standardized mean differences and a random effects model yielded 11 MetaPhlAn 3 species significantly (Wilcoxon rank-sum test FDR P<0.05) associated with colorectal cancer at effect size>0.35 (see **Methods**). **(C)** Species richness is significantly higher in CRC samples compared to control (Wilcoxon rank-sum test P<0.05 in 7/9 datasets), and the expanded MetaPhlAn 3 species catalog detects more species compared to MetaPhlAn 2 (CRC mean median increase 37.1%, controls mean median increase 36.3%). **(D)** Distribution of *cutC* gene relative abundance (log10 count-per-million normalized) from HUMAnN 3 gene family profiles supporting the potential link between choline metabolism and CRC (Thomas et al., 2019). **(E)** Random forest (RF) classification using MetaPhlAn 3 features and HUMAnN 3 features **(F)** confirms that CRC patients can be predicted at (treatment-naive) baseline from the composition of their gut microbiome with performances reaching ~0.85 cross-validated or leave-one-dataset-out (LODO) ROC AUC (see **Methods**).

Improvements in HUMAnN 3 also allowed us to directly test functional hypotheses in the context of the CRC microbiome. Specifically, we previously showed that the abundance of the microbial gene encoding for the choline trimethylamine-lyase (*cutC*) is significantly higher in CRC patients (Thomas et al., 2019), using a customized ShortBRED database (Kaminski et al., 2015) due to incompleteness of reference sequences previously available to HUMAnN 2. HUMAnN 3 was instead able to directly profile relative abundances of 113 UniRef90 gene families annotated as *cutC* orthologs and identified 909 metagenomes in this data collection carrying at least one UniRef90 gene family annotated as *cutC*. These confirmed an increase of *cutC* relative abundance in CRC samples compared to controls (Wilcoxon rank-sum test $P<0.05$ in 6 of the 9 datasets, meta-analysis $P<0.0001$) and thus a potential role of TMA-producing dietary choline metabolism in the gut for this malignancy. Interestingly, a meta-analysis performed on the relative abundances of the L-carnitine dioxygenase gene (*yeaW*), a gene also involved in the trimethylamine synthesis, revealed only weak associations with disease status (Wilcoxon rank-sum test $P<0.05$ in 3 of the 9 datasets, meta-analysis $P=0.095$, **Fig. S8**, **Fig. S9**), possibly reflecting a stronger effect of dietary choline on CRC risk compared to carnitine.

MetaPhlAn 3 and HUMAnN 3 also proved accurate when combining CRC microbiomes using more purely discriminative models such as random forests (RFs), reaching 0.85 average AUC for CRC (vs. control) sample classification in leave-one-dataset-out evaluations using taxonomic features (LODO, minimum 0.76 for the YachidaS_2019 and ThomasAM_2019_a datasets, maximum 0.97 for the GuptaA_2019 dataset; **Fig. 4F**, **Fig. S10**). As in previous studies (Pasolli et al., 2016; Thomas et al., 2019), RFs using functional features performed similarly (0.69 Cross Validation and 0.71 LODO ROC AUC on pathways relative abundance), indicating a tight link between strain-specific taxonomy and gene carriage in this setting. When the classification model was used for assessing features' importance, several new taxa were identified compared to MetaPhlAn 2 and metabolic pathways or EC-numbers relative to HUMAnN 2 (**Fig. S10**), further confirming the relevance of the new reference sequences and annotations available to be profiled in bioBakery 3.

Longitudinal taxonomic and functional meta-omics of IBD

To further demonstrate the utility of MetaPhlAn 3 and HUMAnN 3 on combined meta-omic sequencing datasets, including identification of expression-level biomarkers, we applied the updated methods to 1,635 shotgun metagenomes (MGX) and 817 shotgun metatranscriptomes (MTX) derived from the stool samples of the HMP2 Inflammatory Bowel Disease Multi-omics Database (IBDMDB) cohort (http://ibdmdb.org; see **Methods**). Compared with previously published profiles of the samples generated with MetaPhlAn 2 and HUMAnN 2 (Lloyd-Price et al., 2019) (**Fig. 3A**), the v3 methods' profiles 1) identified more species pangenomes (MGX medians 40 vs. 48, MTX medians 40 vs. 47); 2) explained larger fractions of sample reads by mapping to pangenomes (MGX medians 54 vs. 63%, MTX medians 12 vs. 22%); and 3) explained larger total fractions of sample reads after falling back to translated search (MGX medians 69 vs. 75%, MTX medians 20 vs. 31%). Note that reduced MTX mapping rates (relative to MGX rates) result from enrichment for high-quality but non-coding RNA reads, which are unmapped by design in both HUMAnN 2 and 3. The v3 profiles thus promise increased understanding even of an already well-characterized dataset.

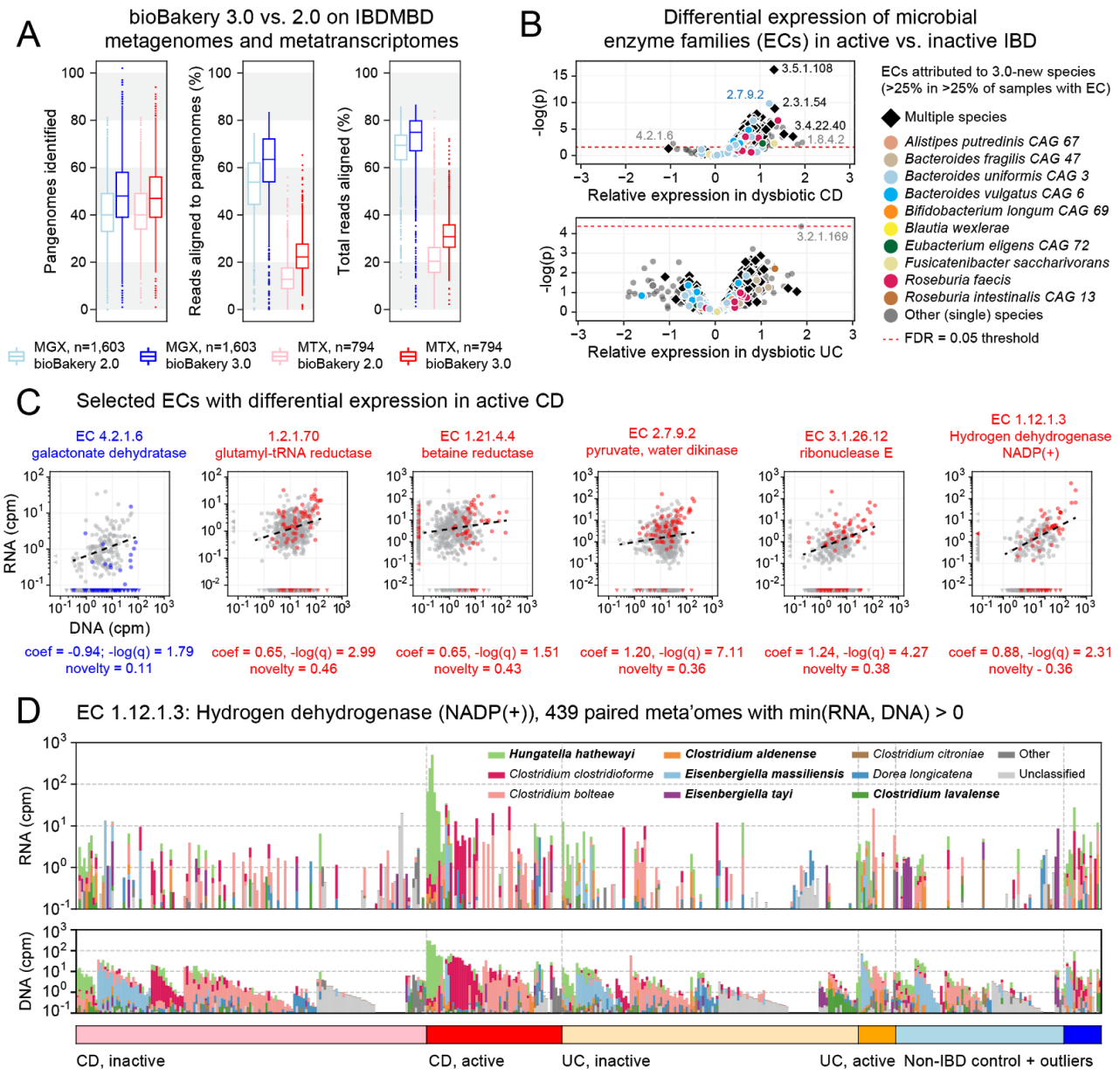

Figure 3: Longitudinal taxonomic and functional meta-omics of IBD. (A) Comparison of MetaPhlAn and HUMAnN profiles of IBDMDB metagenomes and metatranscriptomes using v2 and v3 software (sequencing data and v2 profiles downloaded from http://ibdmdb.org). **(B)** >500 Enzyme Commission (EC) families were significantly [linear mixed-effects (LME) models, FDR q<0.05] differentially expressed in active CD relative to inactive CD; only a single EC met this threshold for active UC. ECs (points) are colored to highlight large contributions from one or more species that were new or newly classified in MetaPhlAn 3 (independent of the strength of their association with active IBD). **(C)** Selected examples of EC families that were differentially expressed in active CD. Colored points correspond to active CD samples; all other samples are gray. The first example (blue) is the only EC to be down-regulated in active CD (as indicated by CD active samples falling below the best-fit RNA vs. DNA line). To match the associated LME models (see **Methods**), best-fit lines exclude samples where an EC's RNA or DNA abundance was zero (such samples are shown as triangles in the x:y margins). **(D)** Species contributions to RNA (top) and DNA (bottom) abundance of EC 1.12.1.3. The 7 strongest contributing species are colored individually; bold names indicate new species in MetaPhlAn 3. Samples are sorted according to the most abundant contributor and then grouped by diagnosis. The tops of the stacked bars (representing community total abundance) follow the logarithmic scale of the y-axis; species' contributions are linearly scaled within that height.

To that end, we applied a mixed-effects model to identify microbial biomarkers of disease activity within the Crohn's disease (CD) and ulcerative colitis (UC) subpopulations of the HMP2 cohort (see **Methods**). More specifically, we examined abundance profiles of EC families from 817 paired HMP2 metagenomes and metatranscriptomes in search of differences in functional activity between active (dysbiotic) and inactive (non-dysbiotic) time points from longitudinally sampled CD and UC patients. We identified 558 ECs whose residual expression was significantly different (FDR q<0.05) in active CD compared with inactive CD and a single EC that was differentially expressed in active UC (protein O-GlcNAcase, EC 3.2.1.169; **Fig. 3B**). The relative absence of biomarkers for active UC may result both from its generally more benign phenotype (Lloyd-Price et al., 2019) and from the smaller number of active UC samples (n=23) compared with active CD samples (n=76); as a result, we focused our subsequent analyses on expression differences within the CD subcohort.

Of the >500 significantly differentially expressed ECs in active CD, all but one were "over-expressed" (i.e. their residual expression after controlling for DNA copy number was higher than expected in active CD; see **Fig. 3B**). Hence, while many species (and their encoded functions) are known to be lost entirely during active IBD (Lloyd-Price et al., 2019), it seems to be rare for functions to be maintained by the community but not utilized. The one notable example of an "under-expressed" function was galactonate dehydratase (EC 4.2.1.6; **Fig. 3C**). This enzyme was encoded and highly expressed by *Faecalibacterium prausnitzii* in both control and inactive CD samples. While galactonate dehydratase was still metagenomically abundant in active CD (where it was contributed primarily by *Escherichia coli*), it was not highly expressed under those conditions. Related observations were made previously using a mouse model of colitis monocolonized with commensal *E. coli* (Patwa et al., 2011). There, microarray-based measurements found a number of enzymes in the galactonate utilization pathway, including galactonate dehydratase, to be among the most strongly down-regulated in comparison with wild-type mice. These results suggest that galactonate metabolism is either infeasible (e.g. due to low bioavailability) or otherwise suboptimal (e.g. due to the presence of preferred energy sources) in the inflamed gut, thus leading to its down-regulation by "generalist" pathobionts like *E. coli*.

From the many over-expressed functions in active CD, we focused for illustrative purposes on examples that were encoded non-trivially by species either new or newly classified in MetaPhlAn 3 ("3.0-new species"; **Fig. 3C**). To aid in this process, we defined an *h*-index-inspired "novelty" score (s) for each EC equal to the largest percentile p of samples with the EC in which p percent of its copies were contributed by 3.0-new species. For example, an EC with s=0.25 indicates that at least 25% of the EC's copies were from 3.0-new species in at least 25% of samples with the EC. The previously mentioned galactonate dehydratase thus had a low novelty score (s=0.11) resulting from dominant contributions of *F. prausnitzii* and *E. coli* (which are not new to MetaPhlAn 3).

Conversely, the highest novelty score was observed for glutamyl-tRNA reductase (EC 1.2.1.70, s=0.46), a highly-transcribed housekeeping gene that received large contributions from the 3.0-new species *Roseburia faecis*, *Phascolarctobacterium faecium*, and *Ruminococcus bicirculans*. Betaine reductase (EC 1.21.4.4, s=0.43), conversely, is much more specific and was contributed in part by 3.0-new species *Hungatella hathewayi*; this is notable as a rare example of a function that was often detectable from community RNA but not DNA (indicating high expression from a small pool of gene copies). Pyruvate, water dikinase (EC: 2.7.9.2) and Ribonuclease E (EC 3.1.26.12) were among the strongest signals of over-expression in active CD by both effect size and statistical significance; these functions were also characterized by large contributions of 3.0-new species (s=0.36 and 0.38, respectively). Ribonuclease E and a final example, hydrogen dehydrogenase

NADP(+) (EC 1.12.1.3), are also representative of the degree to which metagenomic copy number (DNA abundance) tends to be a strong driver of transcription (RNA abundance) in the gut microbiome, and thus the need to account for the former when estimating functional activity. The 3.0-new *H. hathewayi* expresses this enzyme highly in a subset of active CD samples, thus contributing to the enzyme's overall association with active CD.

Population-scale subspecies genetics (StrainPhlAn) and pangenomics (PanPhlAn) of *Ruminococcus bromii*

Strain-level characterization of taxa directly from metagenomes is an effective cultivation-free means to profile the population structure of a microbial species across geography or other conditions (Scholz et al., 2016; Truong et al., 2017) and to track strain transmission (Ferretti et al., 2018). These functionalities are incorporated into (i) StrainPhlAn 3, which infers strain-level genotypes by reconstructing sample-specific consensus sequences from MetaPhlAn 3 markers (Zolfo et al., 2019) (ii) PanPhlAn 3, which identifies strain-specific combination of genes from species' pangenomes; and (iii) PhyloPhlAn 3, which performs precise phylogenetic placement of isolate and metagenome-assembled genomes (MAGs) using global and species-specific core genes (Asnicar et al., 2020) (see **Methods**). ChocoPhlAn 3 automatically quantifies and annotates the distinct types of conservation metrics necessary to identify these markers, all updated in bioBakery 3 (**Table S2**).

Ruminococcus bromii is a common gut microbe that is surprisingly understudied due to its fastidious anaerobicity and general non-pathogenicity (Ze et al., 2012) but it is prevalent in over half of typical gut microbiomes. This made its population genetics, geographic association, and genomic variability of particular interest to assess via StrainPhlAn and PanPhlAn. From the meta-analysis of 7,783 gut metagenomes integrated for a previous study (Pasolli et al., 2019), we considered the 4,077 metagenomes in which *R. bromii* was found present with a relative abundance above 0.05% according to MetaPhlAn 3. StrainPhlAn SNV-based analysis of the 124 *R. bromii*-specific marker genes across the 3,375 samples with sufficient markers' coverage (see **Methods**) revealed a complex population structure not previously recapitulated by the only fifteen genomes available from isolate sequencing (**Fig. 4A**). Sub-clade prediction (see **Methods**) highlighted two sub-species clades that are particularly divergent within the phylogeny (**Fig. S11C-D**); interestingly, the first one (Cluster 1) is mainly composed of strains retrieved from Chinese subjects and from cohorts with a non-Westernized lifestyle (**Fig. 4A**; Cluster 1). StrainPhlAn 3 can thus rapidly reconstruct complex strain-level phylogenies from metagenomes (5,700 seconds using 20 CPUs), and with the integration of PhyloPhlAn 3's improvements specifically for strain-level manipulation of alignments and phylogenies (Asnicar et al., 2020), surpasses the previous version of the software in accuracy and sensitivity (67.4% more strain profiled, **Fig. S11A-B**).

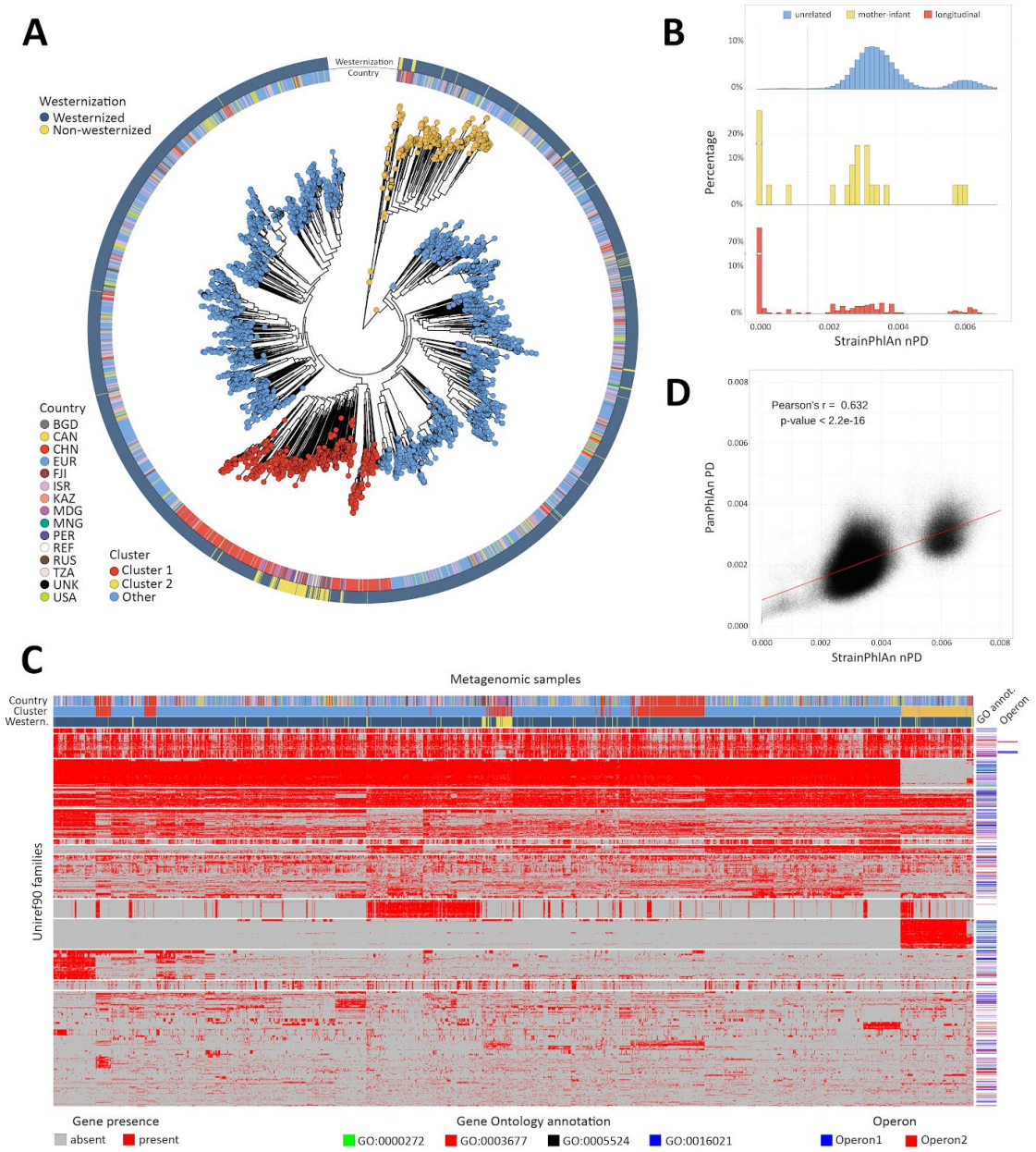

Figure 4: Population-scale strain-level phylogenetic and pangenomic analyses of *Ruminococcus bromii* from over 4,000 human gut metagenomes. (A) StrainPhlAn 3 profiling revealed stratification of *Ruminococcus bromii* clades with genetic content and variants frequently structured with respect to geographic origin and lifestyle. Genetically divergent subclades were identified, labeled as “Cluster 1” (mainly composed of strains retrieved from Chinese subjects) and a subspecies-like Cluster 2. **(B)** Strain tracking of *R. bromii*. While unrelated individuals from diverse populations very rarely share highly genetically similar strains, pairs of related strains are readily detected by StrainPhlAn from longitudinal samples from the same individuals (quantifying short- and medium-term strain retention at about 75%) and in mother-infant pairs (confirming this species is at least partially vertically transmitted). Normalized phylogenetic distances (nPD) were computed on the StrainPhlAn tree. **(C)** PanPhlAn 3 gene profiles of *R. bromii* strains from metagenomes highlights the variability and the structure of the accessory genes across datasets (core genes were removed for clarity). A total of 6,151 UniRef90 gene families from the *R. bromii* pangenome were detected across the 2,679 of the 4,077 samples in which a strain of this species was present at a sufficient abundance to be profiled by PanPhlAn. The 13 highest-rooted gene clusters are shown, highlighting co-occurrence of blocks likely to be functionally related. The most common GO annotations are also reported together with two operons containing genes verified to be on the same locus by analysis of the reference genomes in the PanPhlAn 3 database. **(D)** Genetic (SNV on marker genes from StrainPhlAn 3) and genomic (gene presence/absence from PanPhlAn 3) distances between *R. bromii* strains are correlated (Pearson's r=0.632, p-value<2.2e-16) pointing at generally consistent functional divergence in this species.

StrainPhlAn 3 also extends the ability of reference-based approaches to infer the genetic identity of strains across samples as previously explored (Ferretti et al., 2018; Truong et al., 2017). Specifically for *R. bromii*, different individuals tend to carry different strains diverged with a roughly normal distribution of genetic identities (mean 3.54e-3 normalized phylogenetic distance, **Fig. 4B**). However, the genetic differences between Cluster 1 and Cluster 2 were generally greater, with a lower peak and higher distances (mean 6.1e-3, **Fig. 4B**). For carriers of either clade, within-subject strain retention tended to be high as expected (i.e. low divergence); at distinct time points (average 261.35 s.d. 239.86 days, first quartile 72 days, third quartile 386 days, 3,537 comparisons in total), most of the strain distances (76.4%) approached zero (compared to 1% of comparisons for inter-individual differences, **Fig. 4B**). In addition to detecting these two genetically distinct clades and quantifying within-individual strain retention, a final distribution of higher intra-individual distances clearly captured (rare) strain replacement by *R. bromii* strains (i) in the same or (ii) in a different main cluster in the species' phylogeny. Mother-infant pairs showed a similar dynamic (**Fig. 4B**), with sporadic vertical transmission (~33.3%) (Ferretti et al., 2018; Korpela et al., 2018; Yassour et al., 2018) mixed with strain loss, replacement, and acquisition from other environmental or human sources (Korpela et al., 2018). This analysis highlighted the high precision of StrainPhlAn 3 in detecting strain identity across samples and thus the potential of using it for tracking the transmission network of specific individual strains within and between subjects.

PanPhlAn 3 provides a complementary form of strain analysis by constructing pangenome presence-absence (rather than individual nucleotide variant) genotypes (see **Methods**). Using eight *R. bromii* reference genomes, PanPhlAn 3 revealed the presence of 6,151 UniRef90 pangenes across 2,679 samples with sufficient depth to permit confident strain-specific gene repertoire reconstruction (**Fig. 4C**). This mirrored the genetic divergence of *R. bromii* Clusters 1 and 2, while also highlighting a range of functional differences annotatable to genes unique to the two clusters: Cluster 1 and Cluster 2 showed a total of 797 and 601 UniRef90 families specific to them (Fisher's exact test, FDR $q<0.05$). Although most of these gene families do not have precise functional annotations, these sets of genes should be prioritized in experimental characterization efforts to unravel the sub-species diversity of *R. bromii*, and UnireRef90-to-GO ID mapping also highlighted an enrichment of membrane proteins in Cluster 2. Interestingly, other clusters of co-occurring genes were independent of phylogenetic structure and also verified to be on the same locus on at least two reference genomes in the PanPhlAn 3 database (**Fig. 4C**) providing a new approach at identifying and annotating potential laterally-mobile elements.

StrainPhlan 3 and PanPhlAn 3 can thus be combined with PhyloPhlAn 3 (Asnicar et al., 2020) and HUMAnN 3 to provide multiple, complementary, culture-independent means to investigate the strain-level diversity of taxa in the microbiome, from new data or by re-using thousands of publicly available metagenomes. It is notable that these approaches tend to be consistent with each other (e.g. for *R. bromii,* Pearson's $r=0.632$, $P<2.2e-16$, **Fig. 4D**), while providing different benefits and drawbacks: PanPhlAn used with HUMAnN input is computationally efficient, used from whole pangenomes has higher sensitivity, and StrainPhlAn tends to have higher specificity. Together, the bioBakery 3 components provide an integrated platform for applying strain-level comparative genomics, taxonomic, and functional profiling to meta-omic microbial community studies.

Discussion

Here, we introduce and validate the set of expanded microbial community profiling methods making up the bioBakery 3 platform, including quality control (KneadData), taxonomic profiling (MetaPhlAn), strain profiling (StrainPhlAn and PanPhlAn), functional profiling (HUMAnN), and phylogenetics (PhyloPhlAn), largely relying on the underlying data resource of ChocoPhlAn 3 genomes and pangenomes. These modules are each more accurate and, often, more efficient than their previous versions and current alternatives, particularly for challenging (e.g. non-human-associated) metagenomes and for multi-omics (e.g. metatranscriptomes). In the process of these evaluations, we detected three species newly associated with CRC (*Dialister pneumosintes*, *Ruthenibacterium lactatiformans*, and *Eisenbergiella tayi*), over 500 enzyme families metatranscriptomically upregulated by diverse microbes in IBD, and two new phylogenetically, genomically, and biogeographically distinct subclades of *Ruminococcus bromii*.

These results highlight the degree to which meta-omic approaches can now realize the potential of culture-independent sequencing for characterizing microbial community dynamics, interactions, and evolution that are only active *in situ* and not *in vitro*. Since early studies of environmental and host-associated microbial communities (Gill et al., 2006; Tyson et al., 2004; Venter et al., 2004), it has been clear that many aspects of intercellular and inter-species signaling, short- and long-term evolution, and regulatory programs are exercised by microbes in their natural settings and extremely difficult to recapitulate in a controlled setting. This is supported by the extent to which “dark matter” not previously characterized in the laboratory pervades host-associated and (especially) environmental metagenomes (Parks et al., 2017), with most communities containing a plurality, majority, or sometimes supermajority of novel and/or uncharacterized sequences (Almeida et al., 2019; Pasolli et al., 2019). The bioBakery 3 begins to overcome this challenge by combining a greatly expanded set of reference sequences with ways of “falling back” gracefully when encountering new sequences, while also paving the way for further integration of assembly-based discovery in the future (discussed below). Critically, this now permits large collections of meta-omes to be used in ways only previously possible with large isolate genome or transcriptome collections, e.g. strain-level integrative comparative genomics, near-real-time epidemiology and evolution, and detailed gene content prediction and metabolic modeling. Results such as the heterogeneity of maternal-infant strain transmission and retention, or the globally stratified distribution of subspecies clades, would be extremely challenging to discover by other means.

Methodologically, it is notable that these new meta-omic analysis types have been enabled by several years of improved experimental fidelity, denoising, and quality control approaches. These effectively retain only the “best” subset of reads from large, noisy meta-omes for each analysis of interest, e.g. only the most unique sequences for taxonomic identification, or only the most evolutionarily informative loci for phylogeny. Meta-omes are uniquely positioned for broad reuse and discovery since different “best” subsets of each dataset can be used to answer different questions. The development of meta-omic analysis methods thus parallels that of genome-wide association studies or transcriptomics, inasmuch as early methods were later refined to provide much greater accuracy and scalability through removal of low-quality measurements, within- and between-study normalization approaches, statistical methods to reliably separate signal from noise, and biological annotation of previously uncharacterized loci. Similarly, methods for amplicon-based community profiling have progressed from noise- and chimera-prone stitching and clustering to near-exact sequence variant tracking (Callahan et al., 2016). Fortunately, continued decreases in

sequencing prices and increases in protocol efficiency have now made shotgun meta-omics nearly as affordable as amplicon sequencing in many settings. The challenge, of course, is that each metagenome combines many different noise sources: there is no single, whole genome to finish; host, microbial, and contaminant sequences are not always easily differentiated; there is no one set of “true” underlying variants (since each organism might be represented by multiple strains); and millions of microbial gene products remain functionally uncharacterized (Thomas and Segata, 2019).

Notably, the bioBakery provides one of very few environments currently capable of integrating both metagenomes and metatranscriptomes to begin overcoming these uncertainties (Franzosa et al., 2018). As introduced above, microbial community transcriptomes can be highly unintuitive to interpret, as transcript abundance is always influenced both by expression level and by underlying DNA copy number, i.e. abundance of the expressing taxon. Since both sequence-based DNA and RNA profiles are typically compositional (relative, not absolute, abundances), there is not always a simple way to account for these effects. HUMAnN 3 provides initial within- and between-species normalization options that can be combined with the statistical models of differential expression described here, making e.g. the >500 transcripts overexpressed in Crohn’s disease particularly noteworthy. *Hungatella hathewayi* was uniquely responsible for many of these, an organism not previously associated with IBD in humans (Schaubeck et al., 2016). While many of its overexpressed transcripts are core or housekeeping processes, indicative of general bioactivity in the inflamed gut (comparable to that of e.g. *Escherichia coli* (Lloyd-Price et al., 2019)), others such as betaine reductase are much more specific. This enzyme contributes directly to trimethylamine (TMA) formation (Rath et al., 2019), one of the more noteworthy microbial metabolites implicated in human disease via its transformation to proatherogenic trimethylamine-oxide (TMAO) (Tang et al., 2013). Conversely, the only transcript differentially regulated in ulcerative colitis, underexpressed *F. prausnitzii* galactonate dehydratase, contrasts its utility in polysaccharide degradation under non-inflamed conditions with the upregulation of alternative, more host-antagonistic energy sources in *E. coli* during inflammation (Lloyd-Price et al., 2019). Both of these examples are only analyzable due to the highly specific assignment of meta-omic reads to individual community members’ gene families, in combination with appropriate downstream statistical methods for multi-omics.

Finally, it is striking that metagenomically-derived comparative genomics has only recently been able to reach the scale and scope previously possible with microbial isolates. The genomic epidemiology of pathogens has driven the latter - recently in viral outbreaks such as COVID-19 (Lu et al., 2020) and Ebola (Gire et al., 2014), and in many bacterial conditions such as cholera (Weill et al., 2017) or pneumonia (Croucher et al., 2011). Since metagenomes can simultaneously access all community members with relatively little bias, such studies are now possible with organisms previously overlooked due to the absence of obviously associated phenotypes or convenient culture techniques (Manara et al., 2019; Pasolli et al., 2019). *Ruminococcus bromii* is one such example; despite being over 50% prevalent among typical human gut communities, only 15 isolates were previously sequenced, precluding epidemiology or phylogenetics. In addition to making a novel sub-species phylogenetic and biogeographic structure apparent, the combination of MetaPhlAn, HUMAnN, PanPhlAn, StrainPhlAn, and PhyloPhlAn together confirmed that most *R. bromii* strains are “personal” (i.e. specific to and retained within individuals, like most microbiome members), rarely transmissible across hosts, and that genomic differences characterize each subspecies (suggesting a degree of functional adaptation and specialization). Such results are in principle possible with any combination of metagenomic and isolate taxa and genes of interest, richly integrating culture-independent data with hundreds or thousands of isolate genomes.

Of course, many challenges remain both for improvement of the bioBakery platform and for the field as a whole. Both experimental and computational accessibility of non-bacterial microbial community members remains limited. While accurate, bioBakery 3's capacity for non-bacterial profiling is only slightly improved from the previous version by the expansion of available eukaryotic microbial reference sequences. These components of metagenomes - and, for RNA viruses, metatranscriptomes - are often measured with surprising heterogeneity during the initial generation of sequencing data themselves (Zolfo et al., 2019), suggesting necessary improvements in analytical quality control and normalization as well. The visibility of species with particularly high genetic diversity within individual communities also remains limited; in most cases, only the most dominant strain of each taxon per community is currently analyzable, again for both experimental (e.g. sequencing depth) and analytical reasons (Quince et al., 2017). This is true both for reference-based and for assembly-based approaches, the latter of which are often also stymied by highly diverse taxa (Pasolli et al., 2019). A final area of improvement for the bioBakery, relatedly, is the increased integration between reference-based and assembly-based approaches - begun here via PhyloPhlAn 3 - in order to better leverage MAGs (Almeida et al., 2020), SGBs (Pasolli et al., 2019), and novel gene families.

We thus anticipate improved integration of reference- and assembly-based meta-omic analyses to be one of the main areas of future development for the bioBakery, along with expanded methods for other types of multi-omics in addition to transcription. There will also be a continued focus on quality control and precision, enabling new types of functional analysis within microbial communities (e.g. bioactivity and gene function prediction) without sacrificing sensitivity to rare or novel community members. Finally, we are also committed to the platform's availability with well-documented, open-source implementations, training material, and pre-built locally-executable and cloud-deployable packaging. Feedback on any aspect of the methods or their applications in diverse host-associated or environmental microbiome settings can be submitted at https://forum.biobakery.org, and we hope the bioBakery will continue to provide a flexible, convenient, reproducible, and accurate discovery platform for microbial community biology.

Methods

The bioBakery 3 is a set of computational methods for the analysis of microbial communities from meta-omic data that produce taxonomic, functional, phylogenetic, and strain-level profiles to be interpreted directly or included in downstream statistical analyses (**Fig. 1A**). After read-level quality control by KneadData, MetaPhlAn 3 estimates the set of microbial species (and corresponding higher taxonomic clades) present in a sample and their relative abundances. StrainPhlAn 3 deepens genetic characterization by refining strain-level genotypes of species identified by MetaPhlAn 3. HUMAnN 3 focuses instead on the identification and quantification of the molecular functions encoded in the metagenome or expressed in the metatranscriptome, which can be resolved by PanPhlAn 3 into gene presence-absence strain-level genotypes. PhyloPhlAn 3, as previously reported (Asnicar et al., 2020), provides a comprehensive means to interpret the draft genomes produced by assembly-based metagenomic tools. These bioBakery 3 modules are generally based on an underlying dataset of functionally-annotated isolate microbial genes and genomes produced by ChocoPhlAn 3 to quality-control and annotated UniProt derivatives. This currently includes 99,227 genomes and 87.3M gene families, almost 100-fold greater than the data types included in the first bioBakery release (Segata and Huttenhower, 2011).

The AnADAMA scientific workflow manager

Most bioBakery 3 tools are integrated into reproducible workflows (the “bioBakery workflows”, http://huttenhower.sph.harvard.edu/biobakery_workflows) using the AnADAMA (Another Automated Data Analysis Management Application) task manager, currently v2 (http://huttenhower.sph.harvard.edu/anadama2). Briefly, this wraps doit (http://pydoit.org), a Python-based dependency manager, to provide a simple but scalable language for analysis task definition, version and provenance tracking, change management, documentation, grid and cloud deployment of large compute tasks, and automated reporting. AnADAMA operates in a make-like manner using targets and dependencies of each task to allow for parallelization. In cases where a workflow is modified or input files change, only those tasks impacted by the changes will be rerun. Essential information from all tasks is recorded, using the default logger and command line reporters, to ensure reproducibility. The information logged includes command line options provided to the workflow, the function or command executed for each data modification task, versions of tracked executables, and any output and data products from each task. It can optionally be used to chain together subsequent bioBakery 3 tasks and/or to parallelize them efficiently across multiple files or datasets.

KneadData read-level quality control

The bioBakery 3 includes a simple quality control module for raw sequences, KneadData (http://huttenhower.sph.harvard.edu/kneaddata), which automates a set of typical best practices for raw metagenome and metatranscriptome read cleaning and validation. These include:

- Trimming of 1) low-quality bases (default: 4-mer windows with mean Phred quality <20), 2) truncated reads (default: <50% of pre-trimmed length), and 3) adapter and barcode contaminants using Trimmomatic (Bolger et al., 2014).
- Removal of overrepresented sequences (default: > 0.1% frequency) using FastQC (Andrews and Others, 2010) and low-complexity sequences using TRF (Benson, 1999).

- Depletion of host-derived sequences by mapping with bowtie2 (Langmead and Salzberg, 2012) against an expanded human reference genome (including known “decoy” and contaminant sequences (Breitwieser et al., 2019)) and optionally other hosts (e.g. mouse) reference genomes and/or transcriptomes.

- Depletion of microbial ribosomal and structural RNAs by mapping against SILVA (Yilmaz et al., 2014) in metatranscriptomes.

It is recommended that KneadData be applied to raw sequences prior to further analyses, and the bioBakery workflows do this for all sequence types by default.

The ChocoPhlAn 3 pipeline

We developed the ChocoPhlAn pipeline to organize microbial reference genomes according to their taxonomy and to compute the relevant sequence and annotation data for subsequent bioBakery modules. At a high level, after retrieval of UniProt genomes and gene annotations, species-specific pangenomes (i.e. the set of gene families of a species present in at least one of its genomes) are generated using all the microbial reference genomes passing initial quality control. Core genomes (i.e. gene families present in all the genomes of a species) are then identified from the whole set of pangenomes and used as markers in PhyloPhlAn 3. Core genomes are also processed for the extraction of unique marker genes (i.e. core gene families uniquely associated with one species) that constitute the marker database for MetaPhlAn 3 and StrainPhlAn 3. Finally, functionally annotated pangenomes are processed to serve as references for PanPhlAn 3 and HUMAnN 3.

Data retrieval

ChocoPhlAn relies on the UniProt core data resources (UniProt Consortium, 2019) (release January 2019) and on the NCBI taxonomy and genomes repositories (NCBI Resource Coordinators, 2014) (release January 2019). The two basic sequence data types considered in ChocoPhlAn are the raw genomes of all available microbes and all the microbial proteins/genes identified on these genomes. The main supporting structure for a genome is the underlying microbial taxonomy, whereas the microbial proteins are organized in protein families clustered at multiple stringency parameters.

We adopted the NCBI taxonomy database (NCBI Resource Coordinators, 2014) for use by ChocoPhlAn as it is the one on which our genomic repository, UniProt, is also based. The full taxonomy was downloaded from the NCBI FTP server (ftp.ncbi.nlm.nih.gov/pub/taxonomy/) on January 24 2019. We identified and tagged species with “unidentified”, “sp.”, “Candidatus”, “ bacterium “, and several other keywords as low-quality species. Specifically, the regular expressions used to filter low-quality taxonomic annotations are:

“ (C|c)andidat(e|us) | _sp(_.*|$) | (.*_|^)(b|B)acterium(_.*|) | .*(eury|)archaeo(n_|te|n$).* | .*(endo|)symbiont.* | .*genomosp_.* | .*unidentified.* | .*_bacteria_.* | .*_taxon_.* | .*_et_al_.* | .*_and_.* | .*(cyano|proteo|actino)bacterium_.*) ”

All reference genomes available through UniProt Proteomes and linked to the public DDBJ, ENA, and GenBank repositories were then considered. Genomes are included by UniProt into UniProt Proteomes only if they are fully annotated and have a number of predicted CDSs falling within a statistically defined range of published proteomes from neighbouring species (“What are proteomes?,” 2020). We considered all UniProt Proteomes genomes assigned to the archaeal and

bacterial domain. For micro-eukaryotes, we considered all genomes assigned to the following manually selected genera: *Blastocystis, Candida, Saccharomyces, Cryptosporidium, Entamoeba, Aspergillus, Cryptococcus, Cyclospora, Cystoisospora, Giardia, Leishmania, Malassezia, Neosartorya, Pneumocystis, Toxoplasma, Trachipleistophora, Trichinella, Trichomonas,* and *Trypanosoma*.

Reference genomes ('fasta' format, suffix '.fna') and the associated genomic annotation ('.gff') of each proteome were downloaded from the NCBI GenBank FTP server (ftp.ncbi.nlm.nih.gov/genomes/all/GCA) by retrieving URLs from the assembly_summary_genbank.txt file (ftp.ncbi.nlm.nih.gov/genomes/genbank/assembly_summary_genbank.txt) using the GCA accession included in the UniProt Proteomes resource (01/24/2019). Starting from a total of 111,825 UniProt Proteomes entries, we discarded 12,598 proteomes missing the GenBank accession, ending up with 99,227 genomes (997 Archaea, 97,941 Bacteria, 339 Eukaryota).

The microbial proteins (and genes) associated to at least one UniProt Proteome and considered by ChocoPhlAn are retrieved from the UniProt Knowledgebase (UniProtKB) and the UniProt Archive (UniParc) databases. Proteins included in UniProtKB have been derived from the translation of the CDSs of all available reference genomes included in UniProt Proteomes. ChocoPhlAn 3 also retrieves and includes relevant data present in the UniProtKB entries (retrieved from ftp.uniprot.org/pub/databases/uniprot/ as XML files uniprot_sprot.xml.gz, uniprot_trembl.xml.gz, uniparc_all.xml.gz) such as functional, phylogenomic, and protein domain annotations (KEGG, KO, EggNOG, GO, EC, Pfam) (El-Gebali et al., 2019; Huerta-Cepas et al., 2016; Kanehisa and Goto, 2000; The Gene Ontology Consortium, 2019), accessions for cross-referencing entries with external databases (GenBank, ENA, and BioCyc) (Clark et al., 2016; Karp et al., 2019; Leinonen et al., 2011), name of the gene that encodes for the protein, and proteome accession.

We processed a total of 203.9M proteins included in both UniProtKB and UniParc, and 126.9M of them were associated with a UniProt Proteome entry. The Bacteria domain tallied the highest number of proteins (194.8M), whereas Archaea and Eukaryotes accounted for 5.0M and 4.0M proteins respectively.

In order to reduce the redundancy of the database, we use the UniRef90 clustering of UniProtKB proteins provided by UniProt. In brief, UniProtKB are clustered at different thresholds of sequence identity (100, 90, 50) and made available through the UniProt Reference Clusters (UniRef) resource (Suzek et al., 2015). UniRef90 clusters are generated by clustering unique sequences (UniRef100, which combines identical UniProtKB proteins in a single cluster) via CD-HIT (Li and Godzik, 2006) until August 2019, and via MMseqs2 (Steinegger and Söding, 2018) afterward. Sequences in UniRef90 clusters have at least 90% sequence identity (Suzek et al., 2015). UniRef50 clusters are generated by clustering the UniRef90 cluster seed sequences, and each cluster contains proteins with at least 50% identity. Both UniRef90 and UniRef50 require each protein to overlap at least 80% with the cluster's longest sequence. UniRef entries considered in ChocoPhlAn 3 contain the sequence of a representative protein, the accession IDs of all the entries included in the cluster, the accessions to the UniProtKB and UniParc records, and the accessions of the other associated UniRef cluster are included in the UniProt entries.

A total of 292.1M UniRef clusters were processed (172.3M, 87.3M, and 32.5M for UniRef100, UniRef90, and UniRef50, respectively) and associated with each protein and each genome in ChocoPhlAn 3.

Pan-proteome generation

We then generate pan-proteomes for each species represented at least by one UniProt Proteome. We define a species' pan-proteome as the non-redundant representation of the species' protein-coding potential. These are obtained for each species by considering the unique UniRef90 and UniRef50 protein families present in the genomes assigned at the species level and below.

For each pan-protein, we compute several scores. We define a 'coreness' score for a UniRef90 family as the number of genomes included in the species' pan-proteome having a protein belonging to the UniRef family, and the 'uniqueness' score as the number of pan-proteomes of other species possessing the same pan-protein. We then also considered a 'uniqueness_sp' score, a variant of the 'uniqueness' score obtained excluding those species that were previously tagged as low-quality species. Alongside the 'uniqueness' score, we compute the 'external_genomes' as the number of genomes (rather than species or species' pan-proteomes) of other species' pan-proteomes possessing the same pan-protein. These scores were computed for both UniRef50 and UniRef90 protein families.

In ChocoPhlAn 3 we consider a total of 22,096 species' pan-proteomes and a total of 87.3M UniRef90 core proteins (i.e. with coreness > 0.7, avg. 3,952 s.d. 6,311 per species).

Generation of MetaPhlAn 3 markers

MetaPhlAn relies on a set of unique and species-specific nucleotide markers that were updated in MetaPhlAn 3 starting from the ChocoPhlAn 3 pan-proteomes. We initially filtered out species having taxonomies previously tagged as low quality using the species-level genome bin (SGB) system (Pasolli et al., 2019). "Low-quality" species that were assigned to the same SGB were merged and only the representative SGB was taken into account.

This merging procedure occurred for a total of 1,328 species (6%) that were merged as they were unlikely to be distinguishable in metagenomic samples and would potentially lead to false-positive taxonomic assignments (see **Table S7** for the merged species). For the cases in which multiple species included by the NCBI taxonomy into a "species-group" showed a high number of markers with a high 'uniqueness' score (>30), we proceeded to identify unique markers for the whole species groups. This occurred for the following species groups: *Streptococcus anginosus* group, *Lactobacillus casei* group, *Bacillus subtilis* group, *Enterobacter cloacae* complex, *Pseudomonas syringae* group, *Pseudomonas stutzeri* group, *Pseudomonas putida* group, *Pseudomonas fluorescens* group, *Pseudomonas aeruginosa* group, *Streptococcus dysgalactiae* group, and *Bacillus cereus* group. In all these cases, the pangenomes were built by merging all the species-level pangenomes and treating them as a single species.

In the first step of the marker discovery procedure, we use the pan-proteome built using the UniRef90 clusters considering all proteins with a length between 150 and 1,500 amino acids. Starting from the coreness and uniqueness scores, we applied an iterative approach in order to find up to 150 unique markers whenever possible and retaining only those species with a minimum of 10 unique markers. We classify candidate markers into unique and quasi-markers according to the 'uniqueness' value: markers having zero 'uniqueness' are reported as 'unique markers'. When no unique markers can be identified, the less-stringent thresholds used in the marker discovery procedure allows the identification of the so-called 'quasi-markers', markers having non-null values of 'uniqueness'.

The iterative approach started with the definition of four tiers of unique markers according to a combination of the values of 'coreness', 'uniqueness', and 'external_genomes'. Tier 'A' includes pan-proteins with a coreness score higher than 80%, not shared with more than 2 other pan-proteomes considering both UniRef90 and UniRef50 clustering score ('Uniqueness_NR90' and 'Uniqueness_NR50'), and not present in more than 10 single genomes when considering the UniRef90 and 5 single genomes when considering UniRef50 ('External_genomes_NR90' and 'External_genomes_NR50'), respectively. Tier 'B' includes markers with 'coreness' values between 70% and 80%, 'Uniqueness_NR90', and 'Uniqueness_NR50' values of 5, and values of 'External_genomes_NR90' and 'External_genomes_NR50' lower than 15 and 10 genomes, respectively. Markers that did not meet the previous criteria were included in the 'C' tier, which includes markers with 'coreness' values between 50% and 70%, 'Uniqueness_NR90' less than 10, 'Uniqueness_NR50' less than 15, 'External_genomes_NR90' less than 25, and 'External_genomes_NR50' less than 20. Markers for the species having only one genome included in the pan-proteome, for which the definition of coreness is trivial, were classified as tier 'U', provided that they have zero 'Uniqueness'.

The definition of specific tiers allows the retrieval of the maximum number of unique markers. Marker discovery procedure was performed iteratively for each tier. Candidate markers that meet the tier-defined thresholds were ranked using a score function defined as follows:

$$Score = S_{coreness} * S_{uniqueness50} * S_{uniqueness90}$$

Where

$$S_{coreness} = \sqrt{coreness_{\%}}$$

$$S_{uniqueness90} = -log(1 - \frac{10^4 - min(10^4, uniqueness_{90})}{10^4 - 10^{-4}}) * \frac{1}{5}$$

$$S_{uniqueness50} = -log(1 - \frac{10^4 - min(10^4, uniqueness_{50})}{10^4 - 10^{-4}}) * \frac{1}{5}$$

The score function as defined will prioritize the selection of candidate markers highly conserved in the clade (high 'coreness' value) but shared with the smallest possible number of other species (low values of 'uniqueness'). Tier type is assigned to each candidate marker, and if more than 50 candidate markers were identified, we selected up to 150 markers from the ranked list. If not enough markers were identified (less than 50), the procedure was repeated using the subsequent tier's thresholds. If no markers were identified using tier C thresholds, the species was discarded.

Nucleotide sequences for each marker selected with this procedure are then considered as entries for the MetaPhlAn database. To refine the number of species estimated by the 'uniqueness' parameter, marker sequences were split into non-overlapping chunks of 150bp and mapped versus an index built using all the reference genomes used for the marker identification process using bowtie2 (version 2.3.4.3, parameters '-a --very-sensitive --no-unal --no-hq --no-sq'). We accounted for a newly identified species based on the 'uniqueness' parameter if at least 150 consecutive nucleotides of the marker sequence were found in the identified target reference genome.

We performed an additional step of curation for markers for species with genomes obtained with Co-Abundance gene Groups (CAGs) (Nielsen et al., 2014). To reduce the number of false positives, we removed the CAG species if more than 50% of its markers were shared with the species that gave the taxonomy to the CAG genome.

Each marker has associated an entry in the MetaPhlAn database which includes the species for which the sequence is a marker, the list of species sharing the marker, the sequence length, and the taxonomy of the species. Viral markers were taken from the v20_m200 MetaPhlAn2 database.

Altogether, this identified a total of 1.1M markers for 13,475 species (**Table S8**).

MetaPhlAn 3 taxonomic profiling

The raw reads in a metagenomic sample are mapped by MetaPhlAn 3 to a database of 1.1M markers using bowtie2 (Langmead and Salzberg, 2012). The default bowtie2 mapping parameters are those of the 'very-sensitive' preset but are customizable via the MetaPhlAn 3 settings. In MetaPhlAn 3 the input can be provided as a single FASTQ file (optionally compressed), multiple FASTQs in a single archive, or as a pre-performed mapping. Internally, MetaPhlAn 3 estimates the coverage of each marker and computes the clade's coverage as the robust average of the coverage across the markers of the same clade. The clade's coverages are then normalized across all detected clades to obtain the relative abundance of each taxon as previously described (Segata et al., 2012; Truong et al., 2015).

In version 3, we further optimized the parameter of the robust average which excludes the top and bottom quantiles of the marker abundances ("stat_q" parameter). This is now set by default to 0.2 (i.e. excludes the 20% of markers with the highest abundance as well as the 20% of markers with the lowest abundance). To further improve the quality of the read mapping, we adopted quality controls before and after mapping by discarding low-quality sequences and alignments (reads shorter than 70bp and alignment with a MAPQ value less than 5).

We also introduced a new feature for estimating the "unknown" portion of the taxonomic profile that would correspond with taxa not present in current databases; this is computed by subtracting from the total number of reads the average read depth of each taxon normalized by its taxon-specific average genome length. Additionally, the new output format for MetaPhlAn 3 by default includes the NCBI taxonomy ID of each profiled clade, allowing for better comparisons between tools and tracking of the species name in case of taxonomic reassignment.

Finally, alongside the default MetaPhlAn output format, profiles can be now reported using the CAMI output format defined by (Belmann et al., 2015; "BioBoxes RFC," 2020) that can be used for performing benchmarks with the OPAL framework (Meyer et al., 2019). To support post-profiling analyses, a convenience R script for computing weighted and unweighted UniFrac distances (Lozupone and Knight, 2005) from MetaPhlAn profiles is now available in the software repository (metaphlan/utils/calculate_unifrac.R), alongside the phylogeny (in Newick format) comprising all MetaPhlAn 3 taxa. The improvements and addition in MetaPhlAn 3 compared to the previous MetaPhlAn 2 version are summarized in **Supplementary Table 2**.

StrainPhlAn 3 strain profiling

StrainPhlAn performs genotyping at the strain level by reconstructing sample-specific consensus sequences of MetaPhlAn markers and using them for multiple-sequence alignment and phylogenetic modeling (Truong et al., 2017). StrainPhlAn 3 improves the original implementation in several aspects: (i) the integration of an improved and validated pipeline for consensus sequence generation (Zolfo et al., 2019), (ii) the integration of PhyloPhlAn 3 (Asnicar et al., 2020) which improves the quality of the phylogenetic modeling and the flexibility of the analysis, and (iii) a

refined algorithm for filtering samples not supported by enough species' markers and markers not enough conserved across strains and samples.

StrainPhlAn 3 takes as input the alignment results from the MetaPhlAn 3 profiling (i.e. the mapping of the metagenomic samples against the MetaPhlAn species-specific markers) as well as the MetaPhlAn 3 markers' database. For each sample, StrainPhlAn 3 reconstructs high-quality consensus sequences of the species-specific markers by considering, at each position of the marker, the nucleotide with the highest frequency among the reads mapping against the marker and covering that position. By default, consensus markers reconstructed with less than 8 reads or with a breadth of coverage (i.e. fraction of the marker covered by reads) lower than 80% are discarded ("--breadth_threshold" parameter). Ambiguous bases are defined as positions in the alignment with quality lower than 30 or high polymorphisms (major allele dominance lower than 80%) and are considered for the threshold on the breadth of coverage as unmapped positions.

After marker reconstruction, the filtering algorithm discards samples with less than 20 markers, as well as markers present in less than 80% of the samples ("--sample_with_n_markers" and "--marker_in_n_samples" parameters, respectively). Then, markers are trimmed by removing the leading and trailing 50 bases ("--trim_sequences" parameter), since these are usually supported by lower coverage due to the boundary effect during mapping, and a polymorphic rates report is generated for optional inspection by the user. Finally, filtered samples and markers are processed by PhyloPhlAn 3 for phylogenetic reconstruction. By default, reconstructed sequences are mapped against the markers database using BLASTn (Altschul et al., 1990), multiple sequence alignment is performed by MAFFT (Katoh and Standley, 2013) and phylogenetic trees are produced by RAxML (Stamatakis, 2014) Due to the reconstruction of a strain-level phylogeny, PhyloPhlAn was set to run with "--diversity low" parameter.

Phylogenetic trees produced by StrainPhlan 3 can also be used to identify identical strains across samples, which can be exploited, for example, in strain transmission analyses (Ferretti et al., 2018; Shao et al., 2019). This is now supported by the newly-added "strain_transmission.py" script. This script processes the phylogenetic tree produced by StrainPhAn together with metadata describing relations between the samples (e.g. longitudinal samples or samples with a relation of interest such as mother/infant pairings) to infer strain transmission events. First, using the phylogenetic tree, a pairwise distance matrix is generated and normalized by the total branch length of the tree. Using the distance matrix and the associated metadata, a threshold defining identical strains is inferred selecting the first percentile of the distribution of the non-related-samples distances (i.e. setting an upper bound on the theoretical false-discovery rate at 1%). If longitudinal samples are provided, only one is considered per subject, and samples not included in the metadata are considered as non-related. Finally, related sample pairs with a distance smaller than the inferred threshold are reported as potential transmission events.

HUMAnN 3 data and algorithm updates

Functional potential profiling of microbial communities is performed by HUMAnN using pangenomes annotated with UniRef90 on all species detectable per sample with MetaPhlAn. ChocoPhlAn pangenomes used by HUMAnN for functional profiling are directly available as the species pan-proteomes annotated with the UniRef90 clusters. To obtain a nucleotide representation of each pan-proteome, we identified a representative of the cluster for each pan-protein by selecting a UniProtKB or UniParc entry taxonomically assigned to the desired species. Each cluster representative was used for extracting the nucleotide sequence from the source reference genome and the several functional annotations from different systems (GO terms

(Ashburner et al., 2000), KEGG modules (Kanehisa et al., 2014), KO identifiers, Pfam accessions (Finn et al., 2014), EC numbers (Bairoch, 2000), and eggNOG accessions (Powell et al., 2014)) associated with the UniProtKB entry. Alongside the functional annotations, we associated each UniRef90 cluster with its corresponding UniRef50 cluster in order to provide multiple levels of functional resolution.

HUMAnN 3 implements a number of new options for fine-tuning the steps in its tiered search (e.g. passing custom search parameters to bowtie2 (Langmead and Salzberg, 2012) and DIAMOND (Buchfink et al., 2015) in the pangenome and translated search steps, respectively). We performed a round of additional accuracy and performance tuning on these new parameters prior to the main evaluations of the paper. To minimize overfitting potential, we conducted initial tuning of HUMAnN 3 on the above-described human-like synthetic metagenome, which featured a structure and species composition that were distinct from those of the CAMI and nonhuman synthetic metagenomes used in downstream inter-method comparisons (**Fig. 1**).

We first considered two new options when assigning reads to species pangenomes: 1) requiring pangene sequences to be covered above a threshold fraction of sites before any alignments to those sequences were accepted ("database sequence coverage filtering") and 2) allowing a read to align to multiple pangenes instead of the single target favored by bowtie2's default settings (as used in HUMAnN 2). Coverage filtering (new option 1) was already implemented in HUMAnN 2 for post-processing translated search results, where it was shown to increase UniRef90-level specificity considerably at a small cost to sensitivity (Franzosa et al., 2018). We observed similar results here in the context of pangenome search; as a result, HUMAnN 3 now imposes (separately tunable) database-sequence coverage filters during its pangenome and translated search steps (both default to 50%; **Fig. S4**). Conversely, allowing a read to hit up to 5 pangenes (new option 2, as implemented via bowtie2's "-k 5" setting) had very little impact on accuracy and is not enabled by default in HUMAnN 3.

We additionally considered new options to tune the stringency and memory usage of DIAMOND 0.9 during translated search. The most impactful of these was reducing the identity threshold for per-read alignment to UniRef90 from 90% (the HUMAnN 2 default) to 80% (the new default for HUMAnN 3; **Fig. S4**). While the former value was chosen to respect the average identity among UniRef90 family members, the 80% threshold is more forgiving of variation within read-length windows of a protein-level UniRef90 alignment. Coupled with HUMAnN's database sequence coverage filter, the 80% threshold correctly aligns considerably more reads during translated search without compromising specificity.

While HUMAnN 2 accepted DIAMOND's (default) top-20 database targets per query read, we newly evaluated the top 1 and top 5 targets, as well as any targets within 1, 2, or 10% of the best hit's score. We selected the "within 1% score of the best hit" filter (DIAMOND's "--top 1" option) as a new default for HUMAnN 3 on the basis of a marked increase in UniRef90 specificity with minimal loss of sensitivity. Finally, we explored tuning DIAMOND's memory via the "--block-size (-b)" and "--index-chunks (-c)" flags. We found the achievable increases in speed to be small relative to their corresponding memory requirements, and so HUMAnN 3 continues to favor DIAMOND's default, lower-memory configuration.

PanPhlAn 3 with expanded databases and functional annotations

PanPhlAn performs strain-level metagenomic profiling by identifying the species-specific gene repertoire composition inside individual metagenomic samples (Scholz et al., 2016). It maps

metagenomes against the pangenome of a species of interest using bowtie2 (Langmead and Salzberg, 2012). After coverage normalization (by summing the gene coverage of all genes in a gene family and dividing it by the average gene length of that family), PanPhlAn builds a coverage curve of genes' families across each sample and assesses which of these gene families are present or absent. This leads to the creation of a binary matrix of gene family presence/absence across all samples.

Compared to the previous versions, in PanPhlAn 3 we adopt a new ChocoPhlAn 3 pre-computed pangenome database of 2,298 species built from species included in MetaPhlAn 3 for which at least 2 reference genomes are available. For species having more than 200 reference genomes available, the pangenome is made using a representative subset of 200 genomes maximizing the Mash distances between them (Ondov et al., 2016). PanPhlAn pangenomes from the database are composed of a FASTA file of all contigs, pre-computed bowtie2 indexes and a tab-separated values file containing the UniRef90 ID of the gene family as well as gene name, position in genomes, on contigs, and functional and structural annotations

Moreover, new functionalities include a script for quick visualization of the presence/absence matrix with functionalities for clustering of gene family's profiles across samples. An empirical p-value can be computed for each cluster based on the ratio between the sum of the genes' lengths of one group and its total span along the contig. Thus a significantly "close" genes group can be identified and computation of empirical p-values assessing whether or not the genetic proximity of these families along the contigs could be considered significant. This eases the detection and identification of mobile elements in metagenomic samples.

PhyloPhlAn 3

PhyloPhlAn 3 is an easy-to-use method to perform taxonomic contextualization and phylogenetic analysis of microbial genomes and of metagenome-assembled genomes (MAGs). PhyloPhlAn among its databases exploits both the set of core genes and of reference genomes identified by ChocoPhlAn 3 and extracted from the 111,825 UniProt Proteomes for each taxonomic species. The methods, performance, and examples of PhyloPhlAn are described elsewhere (Asnicar et al., 2020) and refers to the same version incorporated into bioBakery 3. In brief, the core genes included in the PhyloPhlAn 3 database are used to identify sequence homologs in the input genomes and MAGs that are then aligned, concatenated, and used for phylogeny reconstruction. A set of MAGs previously analyzed (Pasolli et al., 2019) can also be included to provide phylogenetic contextualization of newly assembled MAGs. PhyloPhlAn 3 thus provides the methodology to integrate assembly-based methods and phylogenetic analysis into the bioBakery 3 analysis framework.

Synthetic metagenomes and gold standards for bioBakery 3 evaluations

We tuned and evaluated MetaPhlAn 3 and HUMAnN 3 using multiple different synthetic metagenomes of known species and gene content. The first set included synthetic metagenomes and gold-standard taxonomic profiles from the CAMI challenge representing five human body site-specific microbiomes and the murine gut microbiome (Fritz et al., 2019; Sczyrba et al., 2017). All such CAMI metagenomes were used for the evaluation of taxonomic profiling methods (including MetaPhlAn 3) while the first five lexically ordered metagenomes from each environment (human body sites and mouse gut) were used for the evaluation of functional profiling methods (including HUMAnN 3).

Second, because gold standard functional profiles were not provided for the CAMI metagenomes, we generated them ourselves by 1) functionally annotating the genomes sampled to build the CAMI metagenomes (and then 2) weighting their functional contributions according to mean coverage depth per "sample". Notably, this approach to gold-standard construction does not account for gene-to-gene variation in read sampling along the length of community genomes. As a result, comparing the gold standards with functional profiles derived directly from the metagenome underestimates the profiles' accuracy (by ~0.1 units of Bray-Curtis distance at the UniRef90 level).

We applied procedures for community genome annotation developed during HUMAnN2 benchmarking to aid in gold-standard construction (Franzosa et al., 2018). Briefly, we first identified and translated open reading frames (ORFs) within the CAMI genomes using Prodigal (Hyatt et al., 2010), and then aligned the translated ORFs against the v3 UniRef90 and UniRef50 sequence databases using DIAMOND (Buchfink et al., 2015). Each ORF was assigned to the best-scoring UniRef90 family to which it aligned with at least 90% identity and 80% mutual coverage (if any); similarly, ORFs were assigned to the best-scoring UniRef50 family to which they aligned with at least 50% identity. Functional annotations were then transferred from UniRef90 and UniRef50 representatives to the corresponding ORFs, with UniRef90-derived, enzyme commission (EC) annotations forming the basis for the main functional profiling evaluation (**Fig. 1** and **Fig. S3**).

We constructed additional synthetic metagenomes by sampling sequencing reads from curated microbial genome sets using ART (Huang et al., 2012) with an Illumina HiSeq 2500 error model. One such group of metagenomes (abbreviated synphlan-nonhuman) was designed to mirror the sequencing depth and community structure of the CAMI metagenomes: i.e. inclusive of 30-million, 150-nt paired-end sequencing reads sampled from species genomes with a log-normal abundance distribution. However, the synphlan-nonhuman metagenomes are distinct from the CAMI metagenomes in that they exclude genomes of human-associated microbial species (defined as species detected in MetaPhlAn 3 profiles of metagenomes from the Expanded Human Microbiome Project, HMP1-II (Lloyd-Price et al., 2017)). In addition, 50% of species sampled for the synphlan-nonhuman metagenomes were associated with at least two sequenced isolate genomes and 50% of species pairs were congeneric sisters. We constructed an additional synthetic metagenome (synphlan-humanoid) based on the top-50 most abundant species detected from HMP1-II metagenomes to use for initial tuning of HUMAnN 3 (**Fig. S4**). This metagenome contained 10-million, 100-nt paired-end reads sampled evenly from underlying species genomes. We constructed gold standard taxonomic profiles for these metagenomes based on the sampled genomes' taxonomic annotations and target sampling coverage; we constructed gold standard functional profiles based on UniProt-derived annotations of the species' protein-coding genes.

Evaluation of MetaPhlAn 3 and HUMAnN 3 on synthetic data

To assess the performance of MetaPhlAn 3, we compared it with its previous version, MetaPhlAn 2 (Truong et al., 2015), alongside mOTUs2 (Milanese et al., 2019) and Bracken (Lu et al., 2017; Wood et al., 2019). We profiled a total of 118 synthetic metagenomes spanning different ecosystems: (i) 49 synthetic metagenomes (10 Airways, 10 Gastrointestinal Tract, 10 Oral, 10 Skin, 9 Urogenital tract) provided by the 2nd CAMI challenge (Sczyrba et al., 2017) resemble the composition of the Human Microbiome as described by the Human Microbiome Project (Turnbaugh et al., 2007); (ii) 64 synthetic metagenomes generated by CAMISIM and modeled after the murine gut microbiome (Fritz et al., 2019); (iii) 5 synthetic metagenomes including non-human associated species (see above).

Each software was run using default parameters as described in their respective user manuals. Additionally, mOTUs2 was run with parameters "-C recall" and "-C precision" in order to increase precision and recall, respectively. When not directly available from the tool (MetaPhlAn 2 and Bracken), output profiles were converted into the CAMI output format as described by the BioBoxes RFC (Belmann et al., 2015; "BioBoxes RFC," 2020) in order to benchmark with the OPAL framework (Meyer et al., 2019) (version 1.0.5).

From the panel of measures computed by OPAL, we selected a subset (precision, recall, F1 score) for comparisons (**Table S9**). Additionally to these measures, we computed the Pearson Correlation Coefficient between the predicted and expected relative abundance and the Bray-Curtis similarity index using arcsin square-root normalized relative abundances (**Table S3**).

MetaPhlAn 3 includes markers describing species groups, a case is not taken into account by OPAL. To perform the evaluation, we expanded the species group to represent all contained species and considered a true positive if the expected species matches one species taxonomically placed under the species group. In case of no matches, we consider as false positive only one species.

We also assessed the performance in terms of run-time and memory usage. We profiled five HMP samples (SRS014235, SRS011271, SRS064645, SRS023346, SRS048870) with all the aforementioned software (using only one thread) and tracked every second of the execution till the end of process the resident set size (RSS) memory usage using ps.

We evaluated HUMAnN 3, HUMAnN 2 (Franzosa et al., 2018), and Carnelian (Nazeen et al., 2020) on 30 CAMI metagenomes and the 5 synphlan-nonhuman metagenomes. Evaluations of HUMAnN 3 were carried out using version 3.0.0-alpha of the software, MetaPhlAn 3, bowtie2 version 2.3.5.1, and DIAMOND version 0.9.24. Evaluations on HUMAnN 2 were carried out using version 0.11.1 of the software, MetaPhlAn version 2.7.5, bowtie2 version 2.3.5.1, and DIAMOND version 0.8.36 (HUMAnN 2 is not compatible with DIAMOND version 0.9). HUMAnN 3 and 2 were run with their default settings and full-size databases alongside the "--threads 6" option. UniRef90 abundance profiles were converted to EC abundance profiles (to facilitate comparisons with Carnelian) using the "uniref90_level4ec" option of the humann_regroup_table script.

We evaluated Carnelian version 1.0.0 following installation and usage instructions given at http://cb.csail.mit.edu/cb/carnelian/ and https://github.com/snz20/carnelian. Specifically, we first converted synthetic metagenome reads to FASTA format (this step was not counted toward the total runtime of the Carnelian method). Reads were then scanned for peptide fragments using "carnelian.py translate" wrapping FragGeneScan (Rho et al., 2010) version 1.31 with the "-n 3" option. Peptides were then assigned to EC categories using "carnelian.py predict" wrapping Vowpal Wabbit 8.1.1 and the EC-2010-DB model supplied at the above URLs. Finally, adjusted EC abundances were estimated using "carnelian.py abundance" and the average EC family gene lengths supplied with the software and a fragment size of 150 (to match the reads of the CAMI and synphlan-nonhuman metagenomes).

All method calls were made with the humann_benchmark utility script to track total runtime and memory usage (maximum resident set size, MaxRSS). Runtimes were converted to equivalent CPU-hours. For multi-step computations, CPU-hours were summed while the overall maximum MaxRSS was retained. Predicted EC abundances were sum-normalized to 1.0 at the community and per-species levels prior to Bray-Curtis dissimilarity computations.

Colorectal cancer microbiome meta-analysis

We applied the new MetaPhlAn 3 and HUMAnN 3 on a set of human gut metagenomes profiling colorectal cancer patients and controls, updating our previous meta-analyses performed with MetaPhlAn 2 and HUMAnN 2 (Thomas et al., 2019; Wirbel et al., 2019). To the previous meta-analysis, we added two more datasets that became available afterward (Gupta et al., 2019; Yachida et al., 2019). In total, we analyzed 1,262 metagenomes from 10 datasets (for a total of CRC metagenomes and 600 controls, **Table S10**). The dataset was stratified by country of origin with the exception of the two Italian cohorts published in (Thomas et al., 2019) which were kept separate due to differences in the DNA extraction protocols. Results were thus computed on nine distinct sub-cohorts.

MetaPhlAn 3 and HUMAnN 3 were used for the taxonomic and functional profiling of all sub-cohorts. Meta-analysis on the species-level, pathways, UniRef90 gene families, and enzyme commission (EC) categories relative abundances were performed on the sub-cohorts as previously described (Thomas et al., 2019). In brief, relative abundances were arcsine-square-root transformed, Cohen's D was computed by the escalc function (metafor R package (Viechtbauer, 2010) to model random effects, and I^2 estimates and Cochran's Q-test were used for quantifying study-heterogeneity and assessing their statistical significance. Multidimensional scaling analysis was performed on the Weighted UniFrac distance (vegan "cmdscale" and rbiom "unifrac" function (Oksanen et al., 2008) computed on the relative abundance data adjusted for study batch effect with MMUPHin (Ma, 2019) and normalized using arcsin-square root. Alpha-diversity analysis was performed on the data after being rarefied to the 10th percentile of the read depth in each cohort.

We used MetAML (Pasolli et al., 2016) to feed species-level and pathway-level relative abundances to a Random Forest model (Breiman, 2001). Age was also added to the feature-set, as this covariate has been shown to improve microbiome predictions in CRC (Ghosh et al., 2020). MetAML executed the Random-Forest implementation by Scikit-Learn v.0.22.2 with the following parameters: 10,000 estimator trees, square-root as the proportion of feature sampled in entrance to each estimator, no-maximum depth for the trees, 1 sample as the minimum amount for each leaf of each tree, "gini" as impurity criterion. Considering each cohort, we tested the taxonomical and the functional potential profiles in the CRC prediction problem in a standard cohort-specific cross-validation as well as on the more reproducible leave-one-dataset-out (LODO) setting (Thomas et al., 2019; Wirbel et al., 2019).

UniRef90 *cutC* gene family IDs were selected from the UniRef90 database included in HUMAnN 3. Species richness was calculated by tallying species with non zero relative abundance. Differential species richness and *cutC* abundance tests were performed using the Wilcoxon rank-sum test, wilcox.test, as implemented in the 'stats' R package.

HMP2 IBD metagenome and metatranscriptome profiling

We applied MetaPhlAn 3 and HUMAnN 3 to 1,635 metagenomes and 817 metatranscriptomes from the HMP2 Inflammatory Bowel Disease (IBD) Multi-omics Database (IBDMDB) (Lloyd-Price et al., 2019). We took advantage of previously quality-controlled sequencing data from this cohort as downloaded from http://ibdmdb.org (June 2020). Following the standard bioBakery workflow (McIver et al., 2018) for combined meta-omic sequencing data, we processed the HMP2 metagenomes using HUMAnN 3.0.0.alpha.1 (including taxonomic prescreening performed by MetaPhlAn 3). We then processed the paired HMP2 metatranscriptomes using their corresponding metagenomic taxonomic profiles as guides for pangenome selection. To quantify improved

performance in bioBakery 3, we compared the HUMAnN logs produced during the runs described above with logs downloaded from http://ibdmdb.org describing analyses of the same samples using MetaPhlAn 2.6.0 and HUMAnN 2.11.0.

To identify expression-level microbial metabolic biomarkers of IBD activity from the HMP2 dataset, we sum-normalized UniRef90 gene family abundance profiles to "copies per million" (CPM) units and then summed UniRef90 CPMs according to enzyme commission (EC) annotations using HUMAnN utility scripts. We then compared community-level EC expression with other sample properties using a mixed effects model implemented in R's lmerTest package (Kuznetsova et al., 2017) (using subject as a random effect to account for repeated longitudinal sampling):

$$log(RNA) \sim log(DNA) + diagnosis + diagnosis : active + age + antibiotics + (1|subject)$$

For a given EC, we evaluated the above model over paired meta-omes in which the EC's metatranscriptomic abundance (RNA) and metagenomic abundance (DNA) were both non-zero; ECs were excluded if they failed to satisfy this condition in at least 10% of paired meta-omes. This approach avoids interpreting RNA non-detection as strong evidence of "down-regulation" (relative to DNA abundance, identifying zero RNA reads for a feature is more common due to the wide dynamic range of gene expression values and the large fraction of sequencing depth absorbed by non-coding RNAs).

The inclusion of DNA abundance as a covariate in the above model accounts for the strong dependence between a function's gene (metagenomic) copy number and its metatranscriptomic abundance. Thus, associations between EC RNA and other covariates can be interpreted as associations with "residual expression" (potentially reflecting up- or down-regulation of community genes independent of changes in metagenome structure). Subject age at study enrollment and per-sample antibiotics exposure were included as additional clinical covariates. The statistical significance of model covariates was assessed after performing Benjamini-Hochberg FDR correction on model p-values batched by covariate and level.

We focused on associations between residual EC expression and subject diagnosis and disease activity with diagnosis. Here, subject diagnosis was divided broadly into Crohn's disease (CD; n=49), ulcerative colitis (UC; n=30), and non-IBD controls (n=27). Due to the longitudinal nature of the HMP2 dataset, subjects diagnosed with CD and UC experienced variation in disease severity over the course of the study. The effects of disease activity on the microbiome were previously quantified as a "dysbiosis score" (Lloyd-Price et al., 2019) measuring ecological deviation from the control microbiome population. Samples from CD and UC patients that deviated most strongly by this measure were classified as "active." Of 788 paired meta-omes considered here, 363 were from CD patients (76 with "active" CD), 227 were from UC patients (23 with "active" UC"), and 198 were from non-IBD controls. Consistent with earlier analyses of the HMP2 dataset (Lloyd-Price et al., 2019), we did not detect significant differences in EC expression as a function of diagnosis alone (i.e. independent of disease activity), as non-active IBD meta-omes tend to be similar to those from control patients.

Strain-level analysis of *Ruminococcus bromii*

For *Ruminococcus bromii* population genetic analysis, from the 9,316 metagenomes spanning 46 datasets considered by Pasolli et al. (Pasolli et al., 2019), we selected 4,077 samples in which *R. bromii* was found present with a relative abundance above 0.05%. Strain-level profiling with StrainPhlAn 3 was performed using default parameters. 702 samples were discarded due to the

low number and/or poor quality of the reconstructed markers (samples having less than 20 markers and markers present in less than the 80% of the samples are excluded). 124 *R. bromii* MetaPhlAn 3 markers were used to generate a multiple sequence alignment. A phylogenetic distance matrix was produced by the dismat function from the EMBOSS package (Rice et al., 2000) (Kimura 2-parameter distance correction) using the multiple sequence alignment file produced by StrainPhlAn. Prediction strength analysis performed on the phylogenetic distance matrix using the prediction.strength function included in the "fpc" R package (Hennig, 2010) version 2.2 revealed the presence of 4 optimal clusters (strength threshold 0.8). PAM clustering was subsequently applied on the phylogenetic distance matrix using the "cluster" R package (Kaufman and Rousseeuw, 2009) version 2.1. The phylogenetic tree generated by PhyloPhlAn was plotted with GraPhlAn (Asnicar et al., 2015). For visualization purposes, European countries were grouped into the EUR group. Tree cluster colors were assigned by considering the most common cluster assigned to leaves, and clusters 3 and 4 were joined into the "Others" group for the sake of discussion. In order to detect possible events of vertical transmission of R. bromii, we executed the "strain_transmission.py" script using as input the phylogenetic tree produced by StrainPhlAn.

Pangenome-based strain-level analysis was performed on the same selected set of samples using PanPhlAn 3 with the *R. bromii* pangenome composed of 8 reference genomes available on NCBI (GCA_002834165, GCA_002834225, GCA_002834235, GCA_003466165, GCA_003466205, GCA_003466225, GCA_900101355, and GCA_900291485). After mapping the metagenomic samples to the pangenome, a binary matrix of presence/absence was built using the PanPhlAn profiling script with default options for strain detection and filtering (--min_coverage 2 --left_max 1.25 --right_min 0.75). The resulting matrix describes the presence/absence of 6,151 UniRef90 families across 2,679 metagenomics samples and 8 reference genomes.

In order to simplify the visualization of these results, we first discarded the genes families present in less than 2 samples or absent in 5 or less samples. Then, the Jaccard distance based on presence/absence fingerprint was computed for both genes families and samples. Hierarchical clustering was built using the Ward criterion ("ward.D2" in R "hclust" function). A second more stringent filtering removed all genes families present in more than 95% or less than 5% of the remaining samples.

For assessing the correlation between the strain-level genomics and pangenomics results, we compared the phylogenetic distance distributions retrieved from the StrainPhlAn and PanPhlAn analyses. We used RAxML version 8.2.4 (Stamatakis, 2014) to generate phylogenetic distances between samples from PanPhlAN results. PanPhlAn information was coded as the presence-absence fingerprint of each sample and distances were computed using the substitution model based on these two states (argument -m MULTICAT of RAxML). One outlier sample was discarded due to mislabelled genomes. The StrainPhlAn phylogenetic distances were produced during the execution of the "strain_transmission.py" script. Correlation between PanPhlAn and StrainPhlAn pairwise distances was calculated using the Pearson correlation Coefficient.

Data Availability

Human and murine synthetic metagenomes and gold standards provided by the CAMI Challenge are available at https://data.cami-challenge.org/participate.

Non-human synthetic metagenomes and gold standards are available at http://segatalab.cibio.unitn.it/tools/biobakery/. CRC metagenomic datasets analyzed in the

meta-analysis are available in the Sequence Read Archive under accession numbers PRJEB7774, PRJNA531273, PRJNA447983, PRJDB4176, PRJEB12449, PRJEB27928, PRJDB4176, PRJEB10878, and PRJEB6070. Sequences and data for the Integrative Human Microbiome Project are available at the IBDMDB website (https://ibdmdb.org/) and deposited in SRA under accession number PRJNA398089.

Taxonomic profiles, functional profiles, and sample metadata of the CRC datasets are available as **Table S5** and **Table S10**. Taxonomic profiles and functional profiles of the HMP IBDMDB dataset are newly available at https://ibdmdb.org/.

Profiles are also available through the curatedMetagenomicData R package (Pasolli et al., 2017). The full list of metagenomic datasets and samples used for the strain-level analysis of *Ruminococcus bromii* is reported in **Table S1** from (Pasolli et al., 2019). *Ruminococcus bromii* reference genomes are deposited in GenBank under accession GCA_002834165, GCA_002834225, GCA_002834235, GCA_003466165, GCA_003466205, GCA_003466225, GCA_900101355 and GCA_900291485.

Funding

The work was supported by the European Research Council (ERC-STG project MetaPG) to NS; by MIUR 'Futuro in Ricerca' (grant No. RBFR13EWWI_001) to NS; by the European H2020 program (ONCOBIOME-825410 project and MASTER-818368 project) to NS; by the National Cancer Institute of the National Institutes of Health (1U01CA230551) to NS; by the Premio Internazionale Lombardia e Ricerca 2019 to NS; by the Harvard Chan Microbiome Analysis Core (CH); by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (R24DK110499 and U54DE023798) to CH; by Cancer Research UK Grand Challenge award C10674/A27140 to Wendy Garrett (CH); by the Juvenile Diabetes Research Foundation (3-SRA-2016-141-Q-R) to CH; and by the National Human Genome Research Institute of the National Institutes of Health (R01HG005220) to Raphael Irizarry (CH).

References

Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein PC, Gilbert JA, Green JL, Jansson JK, Knight R, Maxon ME, McFall-Ngai MJ, Miller JF, Pollard KS, Ruby EG, Taha SA, Unified Microbiome Initiative Consortium. 2015. MICROBIOME. A unified initiative to harness Earth's microbiomes. *Science* **350**:507–508.

Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 2019. A new genomic blueprint of the human gut microbiota. *Nature* **568**:499–504.

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*. doi:10.1038/s41587-020-0603-3

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**:403–410.

Andrews S, Others. 2010. FastQC: a quality control tool for high throughput sequence data.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**:25–29.

Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N. 2017. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**. doi:10.1128/mSystems.00164-16

Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, Sanders JG, Zolfo M, Kopylova E, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N. 2020. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**:2500.

Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**:e1029.

Bairoch A. 2000. The ENZYME database in 2000. *Nucleic Acids Res* **28**:304–305.

Beghini F, Pasolli E, Truong TD, Putignani L, Cacciò SM, Segata N. 2017. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *ISME J*. doi:10.1038/ismej.2017.139

Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. 2015. Bioboxes: standardised containers for interchangeable bioinformatics software. *Gigascience* **4**:47.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**:573–580.

BioBoxes RFC. 2020. https://github.com/bioboxes/rfc

Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL, Knight R, Maxon ME, Northen TR, Pollard KS, Brodie EL. 2016. Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *MBio* **7**. doi:10.1128/mBio.00714-16

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W,

Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**:852–857.

Breiman L. 2001. Random Forests. *Mach Learn* **45**:5–32.

Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res* **29**:954–960.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**:59–60.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**:581–583.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. doi:10.1093/bioinformatics/btz848

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. *Nucleic Acids Res* **44**:D67–72.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**:430–434.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**:D427–D432.

Feng Q, Liang S, Jia H, Stadlmayr A, Tang L, Lan Z, Zhang D, Xia H, Xu X, Jie Z, Su L, Li X, Li X, Li J, Xiao L, Huber-Schönauer U, Niederseer D, Xu X, Al-Aama JY, Yang H, Wang J, Kristiansen K, Arumugam M, Tilg H, Datz C, Wang J. 2015. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* **6**:6528.

Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, Armanini F, Truong DT, Manara S, Zolfo M, Beghini F, Bertorelli R, De Sanctis V, Bariletti I, Canto R, Clementi R, Cologna M, Crifò T, Cusumano G, Gottardi S, Innamorati C, Masè C, Postai D, Savoi D, Duranti S, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Milani C, Mangifesta M, Anzalone R, Viappiani A, Yassour M, Vlamakis H, Xavier R, Collado CM, Koren O, Tateo S, Soffiati M, Pedrotti A, Ventura M, Huttenhower C, Bork P, Segata N. 2018. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe* **24**:133–145.e5.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**:D222–30.

Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. 2012. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**:289–306.

Forster SC, Kumar N, Anonye BO, Almeida A, Viciani E, Stares MD, Dunn M, Mkandawire TT, Zhu A, Shao Y, Pike LJ, Louie T, Browne HP, Mitchell AL, Neville BA, Finn RD, Lawley TD. 2019. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat Biotechnol* **37**:186–192.

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**:962–968.

Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, Lesker TR, Belmann P, DeMaere MZ, Darling AE, Sczyrba A, Bremges A, McHardy AC. 2019. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**:17.

Ghosh TS, Das M, Jeffery IB, O'Toole PW. 2020. Adjusting for age improves identification of gut microbiome alterations in multiple diseases. *Elife* **9**. doi:10.7554/eLife.50240

Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* **312**:1355–1359.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang P-P, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K,

Foday M, Yillah M, Kanneh F, Robert W, Massally JLB, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**:1369–1372.

Gopalakrishnan V, Spencer CN, Nezi L, Reuben A, Andrews MC, Karpinets TV, Prieto PA, Vicente D, Hoffman K, Wei SC, Cogdill AP, Zhao L, Hudgens CW, Hutchinson DS, Manzo T, Petaccia de Macedo M, Cotechini T, Kumar T, Chen WS, Reddy SM, Szczepaniak Sloane R, Galloway-Pena J, Jiang H, Chen PL, Shpall EJ, Rezvani K, Alousi AM, Chemaly RF, Shelburne S, Vence LM, Okhuysen PC, Jensen VB, Swennes AG, McAllister F, Marcelo Riquelme Sanchez E, Zhang Y, Le Chatelier E, Zitvogel L, Pons N, Austin-Breneman JL, Haydu LE, Burton EM, Gardner JM, Sirmans E, Hu J, Lazar AJ, Tsujikawa T, Diab A, Tawbi H, Glitza IC, Hwu WJ, Patel SP, Woodman SE, Amaria RN, Davies MA, Gershenwald JE, Hwu P, Lee JE, Zhang J, Coussens LM, Cooper ZA, Futreal PA, Daniel CR, Ajami NJ, Petrosino JF, Tetzlaff MT, Sharma P, Allison JP, Jenq RR, Wargo JA. 2018. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**:97–103.

Gupta A, Dhakan DB, Maji A, Saxena R, P K VP, Mahajan S, Pulikkan J, Kurian J, Gomez AM, Scaria J, Amato KR, Sharma AK, Sharma VK. 2019. Association of Flavonifractor plautii, a Flavonoid-Degrading Bacterium, with the Gut Microbiome of Colorectal Cancer Patients in India. *mSystems* **4**. doi:10.1128/mSystems.00438-19

Heinken A, Ravcheev DA, Baldini F, Heirendt L, Fleming RMT, Thiele I. 2019. Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome* **7**:75.

Hennig C. 2010. fpc: Flexible procedures for clustering. *R package version* **2**:0–3.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**:593–594.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**:D286–93.

Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486**:207–214.

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119.

Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. 2015. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol* **11**:e1004557.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**:27–30.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**:D199–205.

Karcher N, Pasolli E, Asnicar F, Huang KD, Tett A, Manara S, Armanini F, Bain D, Duncan SH, Louis P, Zolfo M, Manghi P, Valles-Colomer M, Raffaetà R, Rota-Stabelli O, Collado MC, Zeller G, Falush D, Maixner F, Walker AW, Huttenhower C, Segata N. 2020. Analysis of 1321 Eubacterium rectale genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations. *Genome Biol* **21**:138.

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P. 2019. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* **20**:1085–1093.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–780.

Kaufman L, Rousseeuw PJ. 2009. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, Segata N, Bork P. 2018. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* **28**:561–568.

Kuznetsova A, Brockhoff PB, Christensen RHB, Others. 2017. lmerTest package: tests in linear mixed effects models. *J Stat Softw* **82**:1–26.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**:357–359.

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jorgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clement K, Dore J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, Meta, H. I. T. consortium, Bork P, Wang J, Ehrlich SD, Pedersen O. 2013. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**:541–546.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G. 2011. The European Nucleotide Archive. *Nucleic Acids Res* **39**:D28–31.

Lesker TR, Durairaj AC, Gálvez EJC, Lagkouvardos I, Baines JF, Clavel T, Sczyrba A, McHardy AC, Strowig T. 2020. An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. *Cell Rep* **30**:2909–2922.e6.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**:1674–1676.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659.

Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D, Vázquez-Baeza Y, White RA 3rd, IBDMDB Investigators, Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H, Xavier RJ, Huttenhower C. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**:655–662.

Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**:61–66.

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**:8228–8235.

Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci* **3**:e104.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* **33**:1045–1052.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**:565–574.

Manara S, Asnicar F, Beghini F, Bazzani D, Cumbo F, Zolfo M, Nigro E, Karcher N, Manghi P, Metzger MI, Pasolli E, Segata N. 2019. Microbial genomes from non-human primate gut metagenomes expand the primate-associated bacterial tree of life with over 1000 novel species. *Genome Biol* **20**:299.

Ma S. 2019. MMUPHin. Bioconductor. doi:10.18129/B9.BIOC.MMUPHIN

McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S, Hasan NA, Subramanian P, Moffat K, Levy S, Lonardi S, Greenfield N, Colwell RR, Rosen GL, Mason CE. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* **18**:182.

McIver LJ, Abu-Ali G, Franzosa EA, Schwager R, Morgan XC, Waldron L, Segata N, Huttenhower C. 2018. bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**:1235–1237.

Meyer F, Bremges A, Belmann P, Janssen S, McHardy AC, Koslicki D. 2019. Assessing taxonomic metagenome profilers with OPAL. *Genome Biol* **20**:51.

Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**:1014.

Mitra S, Stärk M, Huson DH. 2011. Analysis of 16S rRNA environmental sequences using MEGAN. *BMC*

Genomics **12 Suppl 3**:S17.

Morgan XC, Segata N, Huttenhower C. 2013. Biodiversity and functional genomics in the human microbiome. *Trends Genet* **29**:51–58.

Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, Pollard KS, Sharpton TJ. 2015. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. *PLoS Comput Biol* **11**:e1004573.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. 2016. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* **26**:1612–1625.

Nazeen S, Yu YW, Berger B. 2020. Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biol* **21**:47.

NCBI Resource Coordinators. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **42**:D7–17.

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD, MetaHIT Consortium. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**:822–828.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**:824–834.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens HH, Wagner H, Oksanen MJ, Suggests M. 2008. The vegan package. *Community ecology package* **10**.

Olm MR, West PT, Brooks B, Firek BA, Baker R, Morowitz MJ, Banfield JF. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7**:26.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**:1–14.

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**:1533–1542.

Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**:649–662.e20.

Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, Beghini F, Malik F, Ramos M, Dowd JB, Huttenhower C, Morgan M, Segata N, Waldron L. 2017. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* **14**:1023–1024.

Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput Biol* **12**:e1004977.

Patwa LG, Fan T-J, Tchaptchet S, Liu Y, Lussier YA, Sartor RB, Hansen JJ. 2011. Chronic intestinal inflammation induces stress-response genes in commensal Escherichia coli. *Gastroenterology* **141**:1842–51.e1–10.

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, Jensen LJ, von Mering C, Bork P. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42**:D231–9.

Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, Kearney SM, Perrotta AR, Berdy B, Zhao S, Lieberman TD, Swanson PK, Smith M, Roesemann S, Alexander JE, Rich SA, Livny J, Vlamakis H, Clish C, Bullock K, Deik A, Scott J, Pierce KA, Xavier RJ, Alm EJ. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat Med* **25**:1442–1452.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to

analysis. *Nat Biotechnol* **35**:833–844.

Rath S, Rud T, Pieper DH, Vital M. 2019. Potential TMA-Producing Bacteria Are Ubiquitously Found in Mammalia. *Front Microbiol* **10**:2966.

Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**:e191.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**:276–277.

Schaubeck M, Clavel T, Calasan J, Lagkouvardos I, Haange SB, Jehmlich N, Basic M, Dupont A, Hornef M, von Bergen M, Bleich A, Haller D. 2016. Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. *Gut* **65**:225–237.

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* **13**:435–438.

Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin H-H, Liao Y-C, Silva GGZ, Cuevas DA, Edwards RA, Saha S, Piro VC, Renard BY, Pop M, Klenk H-P, Göker M, Kyrpides NC, Woyke T, Vorholt JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC. 2017. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat Methods* **14**:1063–1071.

Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. 2013. Computational meta'omics for microbial community studies. *Mol Syst Biol* **9**:666.

Segata N, Huttenhower C. 2011. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* **6**:e24704.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**:811–814.

Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, Kumar N, Stares MD, Rodger A, Brocklehurst P, Field N, Lawley TD. 2019. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**:117–121.

Sivan A, Corrales L, Hubert N, Williams JB, Aquino-Michaels K, Earley ZM, Benyamin FW, Lei YM, Jabri B, Alegre M-L, Chang EB, Gajewski TF. 2015. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science* **350**:1084–1089.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. *Nat Commun* **9**:2542.

Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. 2019. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* **37**:953–961.

Sun L, Xie C, Wang G, Wu Y, Wu Q, Wang X, Liu J, Deng Y, Xia J, Chen B, Zhang S, Yun C, Lian G, Zhang X, Zhang H, Bisson WH, Shi J, Gao X, Ge P, Liu C, Krausz KW, Nichols RG, Cai J, Rimal B, Patterson AD, Wang X, Gonzalez FJ, Jiang C. 2018. Gut microbiota and intestinal FXR mediate the clinical benefits of metformin. *Nat Med* **24**:1919–1929.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**:1282–1288.

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**:926–932.

Tang WHW, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, Wu Y, Hazen SL. 2013. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *N Engl J Med* **368**:1575–1584.

Tanoue T, Morita S, Plichta DR, Skelly AN, Suda W, Sugiura Y, Narushima S, Vlamakis H, Motoo I, Sugita K, Shiota A, Takeshita K, Yasuma-Mitobe K, Riethmacher D, Kaisho T, Norman JM, Mucida D, Suematsu M, Yaguchi T, Bucci V, Inoue T, Kawakami Y, Olle B, Roberts B, Hattori M, Xavier RJ, Atarashi K, Honda K. 2019. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature*

565:600–605.

Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, De Filippis F, Magnabosco C, Bonneau R, Lusingu J, Amuasi J, Reinhard K, Rattei T, Boulund F, Engstrand L, Zink A, Collado MC, Littman DR, Eibach D, Ercolini D, Rota-Stabelli O, Huttenhower C, Maixner F, Segata N. 2019. The Prevotella copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations. *Cell Host Microbe* **0**. doi:10.1016/j.chom.2019.08.018

The Gene Ontology Consortium. 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**:D330–D338.

Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* **25**:667–678.

Thomas AM, Segata N. 2019. Multiple levels of the unknown in microbiome research. *BMC Biol* **17**:48.

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**:902–903.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**:626–638.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449**:804–810.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**:37–43.

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**:D506–D515.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66–74.

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. *J Stat Softw* **36**:1–48.

Vogtmann E, Hua X, Zeller G, Sunagawa S, Voigt AY, Hercog R, Goedert JJ, Shi J, Bork P, Sinha R. 2016. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS One* **11**:e0155362.

Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, Fawal N, Keddy KH, Salje H, Moore S, Mukhopadhyay AK, Bercion R, Luquero FJ, Ngandjio A, Dosso M, Monakhova E, Garin B, Bouchier C, Pazzani C, Mutreja A, Grunow R, Sidikou F, Bonte L, Breurec S, Damian M, Njanpop-Lafourcade B-M, Sapriel G, Page A-L, Hamze M, Henkens M, Chowdhury G, Mengel M, Koeck J-L, Fournier J-M, Dougan G, Grimont PAD, Parkhill J, Holt KE, Piarroux R, Ramamurthy T, Quilici M-L, Thomson NR. 2017. Genomic history of the seventh pandemic of cholera in Africa. *Science* **358**:785–789.

What are proteomes? 2020. . *UniProt*. https://www.uniprot.org/help/proteome

Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat Med* **25**:679–689.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**:257.

Xiong W, Giannone RJ, Morowitz MJ, Banfield JF, Hettich RL. 2015. Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut. *J Proteome Res* **14**:133–141.

Yachida S, Mizutani S, Shiroma H, Shiba S, Nakajima T, Sakamoto T, Watanabe H, Masuda K, Nishimoto Y, Kubo M, Hosoda F, Rokutan H, Matsumoto M, Takamaru H, Yamada M, Matsuda T, Iwasaki M, Yamaji T, Yachida T, Soga T, Kurokawa K, Toyoda A, Ogura Y, Hayashi T, Hatakeyama M, Nakagama H, Saito Y, Fukuda S, Shibata T, Yamada T. 2019. Metagenomic and metabolomic analyses reveal distinct

stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat Med* **25**:968–976.

Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, Siljander H, Selvenius J, Oikarinen S, Hyöty H, Virtanen SM, Ilonen J, Ferretti P, Pasolli E, Tett A, Asnicar F, Segata N, Vlamakis H, Lander ES, Huttenhower C, Knip M, Xavier RJ. 2018. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* **24**:146–154.e4.

Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**:779–794.

Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* **42**:D643–8.

Yu J, Feng Q, Wong SH, Zhang D, Liang QY, Qin Y, Tang L, Zhao H, Stenvang J, Li Y, Wang X, Xu X, Chen N, Wu WKK, Al-Aama J, Nielsen HJ, Kiilerich P, Jensen BAH, Yau TO, Lan Z, Jia H, Li J, Xiao L, Lam TYT, Ng SC, Cheng AS-L, Wong VW-S, Chan FKL, Xu X, Yang H, Madsen L, Datz C, Tilg H, Wang J, Brünner N, Kristiansen K, Arumugam M, Sung JJ-Y, Wang J. 2017. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**:70–78.

Yutin N, Makarova KS, Gussow AB, Krupovic M, Segall A, Edwards RA, Koonin EV. 2018. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**:38–46.

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**:766.

Ze X, Duncan SH, Louis P, Flint HJ. 2012. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME J* **6**:1535–1543.

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao J-Y, Wu Z, Xu ZZ, Cantrell K, Yang Y, Sayyari E, Rabiee M, Morton JT, Podell S, Knights D, Li W-J, Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun* **10**:5477.

Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, Segata N. 2019. Detecting contamination in viromes using ViromeQC. *Nat Biotechnol* **37**:1408–1412.

Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, Wan D, Jiang R, Su L, Feng Q, Jie Z, Guo T, Xia Z, Liu C, Yu J, Lin Y, Tang S, Huo G, Xu X, Hou Y, Liu X, Wang J, Yang H, Kristiansen K, Li J, Jia H, Xiao L. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* **37**:179–185.