Comprehensive function annotation of metagenomes and microbial genomes using a deep learning-based method

Mary Maranga[1], Pawel Szczerbiak[1], Valentyn Bezshapkin[1], Vladimir Gligorijevic[2,5], Richard Bonneau[2,5], Tomasz Kosciolek[1*], Tommi Vatanen[3,4,6*]

[1] Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland
[2] Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, USA
[3] Broad Institute, Cambridge, MA, USA
[4] Liggins Institute, University of Auckland, New Zealand
[5] Prescient Design, a Genentech Accelerator, New York, NY, USA
[6] University of Helsinki, Finland

*corresponding authors
tomasz.kosciolek@uj.edu.pl
tommi.vatanen@helsinki.fi

Abstract

Comprehensive protein function annotation is key to understanding microbiome-related disease mechanisms in the host organisms. We have developed a new metagenome analysis workflow integrating *de novo* genome reconstruction, taxonomic profiling and deep learning-based functional annotations from DeepFRI. We validate DeepFRI functional annotations by comparing them to orthology-based annotations from eggNOG. Further, we demonstrate the usage of the workflow using 1,070 infant metagenome samples from the DIABIMMUNE cohort. We have generated a sequence catalogue comprising a collection of 7,174 metagenome-assembled genomes (MAGs), of which 2,255 are high quality near-complete bacterial genomes. These genomes encode 1.9 million non-redundant genes. We found high concordance (70%) between GO annotations predicted by DeepFRI and eggNOG. Importantly, we show that DeepFRI improved the annotation coverage, with nearly all the genes in the gene catalogue obtaining GO molecular function annotations, although the annotations are less specific compared to eggNOG. The pan-genome analysis of 42 bacterial species revealed a striking difference between

DeepFRI and eggNOG annotations. eggNOG was shown to annotate more genes coming from well-studied organisms such as *Escherichia coli* while DeepFRI on the other hand is not sensitive to taxa. This work contributes to the understanding of the functional signature of the human gut microbiome.

Introduction

The advent of sequencing technologies has resulted in a substantial increase in metagenomic studies sequencing DNA of microbial consortia (microbiomes) inhabiting various hosts and environments. This has led to a significant boost in known microbial genomes contributing substantially to our understanding of the genetic diversity encoded in microbiomes. Although metagenomics provides a great potential to characterize difficult-to-cultivate and uncultivated microbes, our ability to link specific genes to disease phenotypes is hampered by poor understanding of the functions and roles of the majority of the newly discovered microbial genes (Li et al., 2014; Pasolli et al., 2019).

Comprehensive functional annotation is crucial in identifying disease-causing functional changes in proteins, detecting antibiotic resistance genes (Baric et al., 2016; Tuan et al., 2019), and designing new therapeutic strategies. Labor-intensive laboratory experiments still provide the most reliable way of functionally annotating genes and the proteins they encode (Tiwari & Srivastava, 2014). However, due to the rapid increase in the number of uncharacterized proteins, such experimental methods and manual curation fail to scale up to accommodate such a large amount of protein sequence data. This has created a huge sequence-to-function gap which is still widening, because sequencing is high-throughput, while functional characterization continues to be relatively slow. The overall proportion of human gut microbiome protein-coding genes of which functions cannot be predicted range between 40-60% depending on the annotation method (Li et al 2014; Vatanen et al., 2018). A recent study reported that 27.3% of genes present in the Unified Human Gastrointestinal Genome (UHGG) catalogue could not be mapped to functional databases, while 14.2% of genes matching to COG database were labelled as “function unknown” (Almeida et al., 2021). Automated and scalable methods to predict functions for microbial genes and proteins are needed to address this gap.

Most functional annotation methods rely on inferring homology across databases such as Uniprot (Bateman et al., 2021) and the NCBI's reference sequence (RefSeq) database (O'Leary et al., 2016). Homology-based annotation methods are fast, however, they suffer low functional annotation coverage. Deep learning methods have been considered as an effective complement (Gligorijevic et al., 2019; Kulmanov et al., 2018; Rifaioglu et al., 2019), since they are able to predict protein functions at large-scale irrespective of sequence homology. Metagenome functional annotation can be performed either by assembling reads into contigs, followed by reconstruction of complete and accurate metagenome-assembled genomes (MAGs) and mapping predicted genes to annotated sequences (Quince et al., 2017), or by directly mapping individual reads to annotated gene sequences (Bose et al., 2015; Franzosa et al., 2018; Kim et al., 2016). Various read mapping-based annotation workflows have been developed including HUMAnN2 (Franzosa et al., 2018), HUMAnN3 (Beghini et al., 2021) and MG-RAST (Keegan et al., 2016). These pipelines align short reads directly to a reference catalogue of microbial sequence to estimate functional profiles (Eng et al., 2020). Such approach is, however, limited by the used reference database and fails to annotate genes that lack a homologous counterpart in the reference collection.

Reconstruction of MAGs has proved to be a successful strategy for novel genome discovery and characterization of the functional potential of complex microbial communities (Almeida et al., 2019). The metagenomic reads are assembled into contigs and grouped into different bins based on depth of coverage and tetranucleotide frequency across samples (Alneberg et al., 2014; Strous et al., 2012). Several studies have employed this technique providing new insights into the genetic diversity of the human gut microbiome and paving ways for exploring microbial dark matter (Parks et al., 2017; Nayfach et al., 2019).

Widely used schemes for classifying protein functions include the Gene Ontology (GO) (Carbon et al., 2019), Enzyme Commission (EC) numbers (Martínez Cuesta et al., 2015), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017). Gene ontology (GO) terms are attractive as they offer an accurate description of the protein functions levels and relationships between those annotations. The system represents function in a Directed Acyclic Graph (DAG) type

relationship where protein attributes are divided into three main categories: molecular function (MF), biological process (BP) and cellular component (CC). Importantly, GO terms facilitate more comprehensive inter-method comparisons through these protein annotation relationships even when the specificity of annotations differ between methods.

Here we describe a new metagenome assembly workflow integrating *de novo* genome reconstruction, taxonomic profiling and deep learning-based gene annotations from DeepFRI (Gligorijevic et al., 2021). We validate the functional annotations by comparing them to orthology based annotations from eggNOG (Huerta-Cepas et al., 2019). We further show that such an approach almost exhaustively annotates millions of metagenome-assembled genes, although the annotations are, on average, less specific compared to the homology-based annotations from eggNOG. Finally, we demonstrate integration of gene annotations and metagenome-assembled pangenomes using >1,000 infant metagenomes from the DIABIMMUNE cohort. In summary, combining gene annotations from DeepFRI with the commonly used orthology-based annotations helps understanding the roles of novel microbial genes observed in metagenomic surveys and could eventually close the sequence-to-function gap hindering most microbiome studies.

Methods

Workflow architecture

The workflow is a fully automated pipeline for metagenomic assembly, binning of metagenome-assembled genomes, construction of gene catalogue and functional annotation. It integrates state-of-the-art bioinformatic tools via Docker containers. The workflow is implemented using Workflow Description Language (WDL) and allows flexibility of bioinformatic tools versioning and scalability of memory in high-performance computing environments. This allows large-scale functional annotations of metagenomics data by leveraging high-quality protein information to annotate functions with higher coverage. It takes raw paired-end Illumina reads (short reads) as input and performs data analysis in five phases (1) Assembly of sequencing reads into contigs, gene prediction and clustering to generate a gene catalogue, (2)

functional annotation of the gene catalogue, (3) Binning of assembled contigs into metagenome-assembled genomes (MAGs), (4) Taxonomic annotation of MAGs and (5) Mapping between MAGs and functionally annotated gene catalogue (Figure 1). We have developed a custom Python script for mapping between metagenomic species and functionally annotated gene catalogue. On completion, the workflow provides various tabular output files such as the metagenome species (MAGs) constructed, non-redundant gene catalogue, and functional annotations as tab-separated output tables.

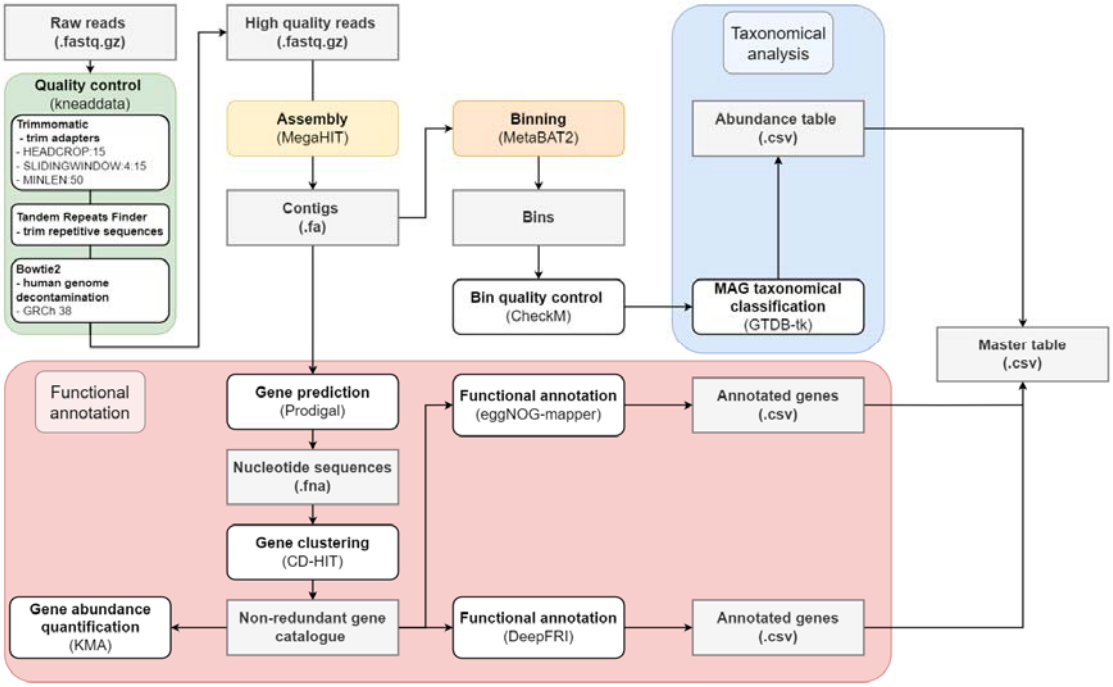

Figure 1: Schematic workflow overview.

Quality control and assembly

Metagenomics data usually contains sequencing artifacts such as low-quality reads, host contamination and adapter sequences, which compromise downstream analysis. The quality control phase utilizes kneadData https://huttenhower.sph.harvard.edu/ kneaddata/. KneadData combines Trimmomatic for adapter sequence and low-quality base trimming (Bolger et al., 2014), Bowtie2 (Langmead & Salzberg, 2012) for removal of host-derived reads and Tandem Repeat Finder for removing tandem repeats in DNA (Benson, 1999).

After the quality control process, the workflow uses MegaHIT (v2.4.2) (Li et al., 2015) to assemble each sample individually into contigs. MegaHIT has been shown to reconstruct known genomes accurately by employing succinate data structures known as the *de Bruijn* graph which decomposes reads into k-mers, thus using minimal memory requirements (Quince et al., 2017; Sczyrba et al., 2017). Additionally, it captures more genes allowing higher resolution of functional profiling and phenotype to genotype analysis of specific microbial communities. The assembled contigs less than 500 bp long are then filtered out to yield high-quality contigs.

Construction of the gene catalogue and gene abundance estimation

For gene annotation, open reading frames from each contig (length ≥ 500 bp) are predicted using Prodigal (Hyatt et al., 2010). The gene products are then clustered at 95% sequence identity using CD-HIT (v4.8.1) (Fu et al., 2012) to generate the non-redundant gene catalogue. Thereafter, gene abundance quantification is computed by mapping paired-end reads from each sample against the gene catalogue using KMA tool (Clausen et al., 2018). Hits to a sequence are weighted based on the target sequence length and further normalized to produce counts in copies per million (CPM) units.

Binning of contigs to metagenome-assembled genomes and taxonomic classification

Prediction of metagenome assembled genomes provides novel biological insights into genetic diversity within the microbial communities. Our pipeline implements an adaptive binning approach using Metabat2 (v.2.3.0) (Kang et al., 2019) by clustering contigs into bins (metagenome-assembled genomes, MAGs) based on the use of tetra-nucleotide frequencies, differential abundance and coverage across samples. The workflow supports binning for each sample individually thus recovering a great number of high-quality bins. Moreover, for each MAG, completeness and contamination are determined using single-copy genes with CheckM (v1.0.12) (Parks et al., 2015). Users can then select MAGs of desired quality for downstream

analyses. We select near-complete, high-quality (HQ) genomes as those having ≥90% genome completeness and <5% contamination for downstream analysis.

Taxonomic classification is conducted using GTDB-Tk (v2.1.0) (Chaumeil et al., 2020). GTDB-Tk allows robust taxonomic classification by identifying single-copy marker genes from the Genome Taxonomy Database reference genomes using HMMER (Eddy, 2011). The domain-specific marker gene alignments are then concatenated into multiple sequence alignment and used to construct a reference tree by a likelihood-based phylogenetic inference algorithm. The tool improves the resolution of microbial taxa by classifying MAGs based on their position in the GTDB reference tree, evolutionary divergence, and average nucleotide identity to the reference genomes.

Functional annotation

Representative cluster centroids of the gene catalogue are annotated with predicted gene functions using two complementary approaches; deep neural networks in DeepFRI (Gligorijevic et al.,2019) and orthology based annotation in eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (Huerta-Cepas et al., 2019). eggNOG provides accurate function by inferring orthology relationships and evolutionary history of proteins and is widely used for gene functional annotations. The annotated genes are mapped to various ontologies thus accurately assigning each gene to its function. These ontologies include Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) for metabolic pathway analysis, COGs and SMART/Pfam domains for each group.

On the other hand, DeepFRI (Gligorijevic et al., 2021) is a novel deep learning-based function annotation method which applies Graph Convolutional Networks (GCNs) to learn and extract features from protein sequences and structures. The method achieves accurate prediction of GO terms in two stages, (1) the first step utilizes long short term memory (LSTM)-based language model, pre-trained on sequences from Pfam database, to extract features from PDB protein sequences, then (2) GCN architecture learns structure to function relationships using three graph convolutional layers. DeepFRI has the ability to predict protein functions accurately irrespective of sequence homology and provide interpretability of its predictions via saliency maps.

Integration of gene functions, quantification and metagenome assembled genomes

We have developed a custom Python script for generating mapping between functionally annotated non-redundant genes and metagenome assembled genomes (MAGs). The script takes in a variety of input files: gene catalogue including cluster members, assembled contigs and metagenome assembled genomes (MAGs) (including taxonomic annotations from GTDB and CheckM quality control information).

Considering that the gene catalogue was created by clustering genes into a representative sequence (centroid) using CD-HIT (Fu et al., 2012) at a sequence identity threshold of 95%, we then propagate functions within each gene cluster and further annotate genes within MAGs. This establishes a link between microbial genes and species allowing reconstruction of the functional potential of individual taxa. Moreover, integrating functional and taxonomical results allows us to determine organisms contributing to various functions. Thus, providing a deeper and more comprehensive insight into identification of specific species or genes associated with outcome of interest.

Quantification of individual gene abundances across a given metagenome is critical for understanding how variation in the functional composition can impact human health as well as in understanding how microbes adapt to various environments. We perform quantifications of each gene and then apply the abundance values to quantify gene functions from DeepFRI and eggNOG. To estimate per-sample gene abundances, the quality-controlled reads are aligned directly against the non-redundant gene catalogue KMA tool (Clausen et al., 2018) with default parameters. The number of reads mapping to each gene is then used as a proxy for its abundance in the sample. For example, the total number of reads mapped to each gene is normalized by the length of the target gene sequence to produce counts in copies per million (CPM) units. The relative abundance of each gene is calculated as follows:

$$Relative\ abundance = \frac{Number\ of\ mapped\ reads\ to\ a\ target\ gene\ *\ 1{,}000{,}000}{Total\ read\ counts}$$

Metagenome assembled genomes (MAGs) include core genes that have specific and specialized function as well as accessory genes that are variably present in the genomes. We use MAGs to construct pan-genomes in reference free manner. To construct a pan-genome of a given species, we recruit near-complete MAGs (≥90% genome completeness and <5%) and species with at least 10 genomes. We then define genes present in ≥90% of MAGs of each species as “core” genes, while genes present in <90% of MAGS are considered as “accessory” genes. This provides additional dimension to analysis of the functional repertoire of core and accessory genomes and gives a glimpse into the functional diversity within species.

Results

Undescribed diversity in infant gut metagenomes

To demonstrate feasibility and robustness and to highlight several use cases of the workflow data products, we analyzed metagenomic data from the DIABIMMUNE cohort (Vatanen et al., 2019) (https://diabimmune.broadinstitute.org/diabimmune/). This constituted a longitudinal sample series from individual infants and young children from Finland, Estonia and Russia. Overall, we aggregated metagenome shotgun sequencing data from 1,070 samples including 202 samples from Estonia, 586 samples from Finland and 282 samples from Russia. Average number of reads per sample was 1.49×10^7 paired end reads (minimum 9.28×10^3; maximum 8.09×10^7).

The first step for functional analysis was the construction of a non-redundant gene catalogue. We performed *de novo* assembly of the metagenomes resulting in a total of 17 million long contigs of length ≥500 bp mean 16,510 contigs per sample (minimum 94; maximum 60,979 contigs), collectively harboring approximately 21.9 million ORFs, as predicted by Prodigal. Clustering of these genes into gene families with >95% sequence similarity resulted in a catalogue of 1.9 million non-redundant genes (Table 1).

Metabat binning of contigs produced a collection of 7,174 MAGs, of which 2,255 were near-complete bacterial genomes (>90% completeness, <5% contamination) (Table 1 and Supplementary figure 1) which corresponded to 2 near-complete MAGs

per sample, on average. Taxonomic annotation of each genome was carried out using GTDB-tk. The high-quality genomes spanned 202 bacterial species, with most MAGs representing phyla *Firmicutes* (884 genomes) *Bacteroidota* (823 genomes), *Actinobacteriota* (296) and genera *Bacteroides* (614 genomes), *Bifidobacterium* (268 genomes), and *Faecalibacterium* (124 genomes) (Supplementary figure 2b-c).

To visualize the distribution of high quality near-complete genomes across phylogeny, we build a maximum likelihood phylogenetic tree based on high quality near-complete genomes. The tree was constructed by PhyloPhLAN (Asnicar et al., 2020) and visualized using Interactive Tree of Life (v5.6.2) (Letunic & Bork, 2019). The genomes spanned 10 bacterial class, including *Coriobacteriia*, *Actinomycetia*, *Bacilli*, *Negativicutes, Clostridia*, *Campylobacteria*, *Alphaproteobacteria*, *Gammaproteobacteria*, *Verrucomicrobiae*, and *Bacteroidia* (Figure 5a). We observed high phylogenetic dispersion in *Faecalibacterium* genera. *Faecalibacterium* has been shown to be highly diverse and consists multiple different phylotypes (Pasolli et al., 2019).

Table 1: Assembly and gene prediction statistics for the Diabimmune dataset.

Assembly	**Statistics**
Contigs	17M
Total MAGs reconstructed	7,174
Near-complete MAGs	2,255
Non-redundant gene catalogue	1.9M

To integrate information on gene homology from the gene catalogue and taxonomic information obtained through MAGs, we tried to provide further insights into the following question: are 95% gene families taxon specific or shared across genomes of different species, genera or even families?. Interestingly we found that a notable portion of the gene families had multiple taxonomies assigned, 55,006 (3%) at family level, 208,885 (11%) at genus level, and 276,042 (14%) at species level (Supplementary figure 2a). We hypothesized that these shared genes could be

attributed to the horizontal gene transfer between species although additional studies are required to conclusively confirm this.

Functional representation of the DIABIMMUNE infant metagenomes

To elucidate the functional composition of the gut microbiota we annotated genes using DeepFRI and eggNOG. We focussed on the molecular function branch of Gene Ontology and only considered DeepFRI annotations with a probability threshold of 0.2. We first validated the predicted annotations by comparing the level of concordance between both methods based on the specificity of the GO terms, expressed as Shannon information content. The information content (IC) is used to quantify the specificity of a GO term in the context of the entire set of annotations, such that GO terms annotating many genes are considered to be general and hence contain low information content while rare occurring GO terms are considered more specific and contain higher information (IC values) (Gaudet et al., 2009). We propagated the Gene Ontology annotations upward through the child-parent relationships using a list of information content values obtained from (Gligorijevic et al., 2021), where IC was calculated as follows:

$$IC(GOi) = -log_2 Prob(GOi),$$

where, Prob(GO_i) is the probability of GO term *i* occurring in the UniProt-GOA database (n_i/n, where n_i equals number of annotations with a term *i* and *n* is total number of annotated proteins in UniProt-GOA) (Gligorijevic et al., 2021).

By comparing consensus in the gene ontology annotations predicted by eggNOG and DeepFRI, we observed a decreasing trend in DeepFRI where genes with high information content annotation were less in comparison to eggNOG (Figure 2a). DeepFRI gave a high number of general annotations while eggNOG was shown to predict more specific MF-GO terms (IC > 12). Next, we wanted to assess whether genes annotated by both methods gained function at the same information content level. We further stratified annotations into different categories; (1) concordance annotations, where eggNOG and DeepFRI annotations agree (genes obtained functions at the same information content level); (2) discordant, where eggNOG and DeepFRI annotations disagree but annotations from both methods exist; (3)

DeepFRI and eggNOG unique, genes obtained function from only one method. Our results demonstrated high concordance annotations between eggNOG and DeepFRI (Figure 2b-c). Of all the genes annotated by both methods, the proportion of concordance annotations was 70%, on average. Additionally, DeepFRI had a high number of unique annotations thus expanding the functional landscape of metagenome-assembled genes

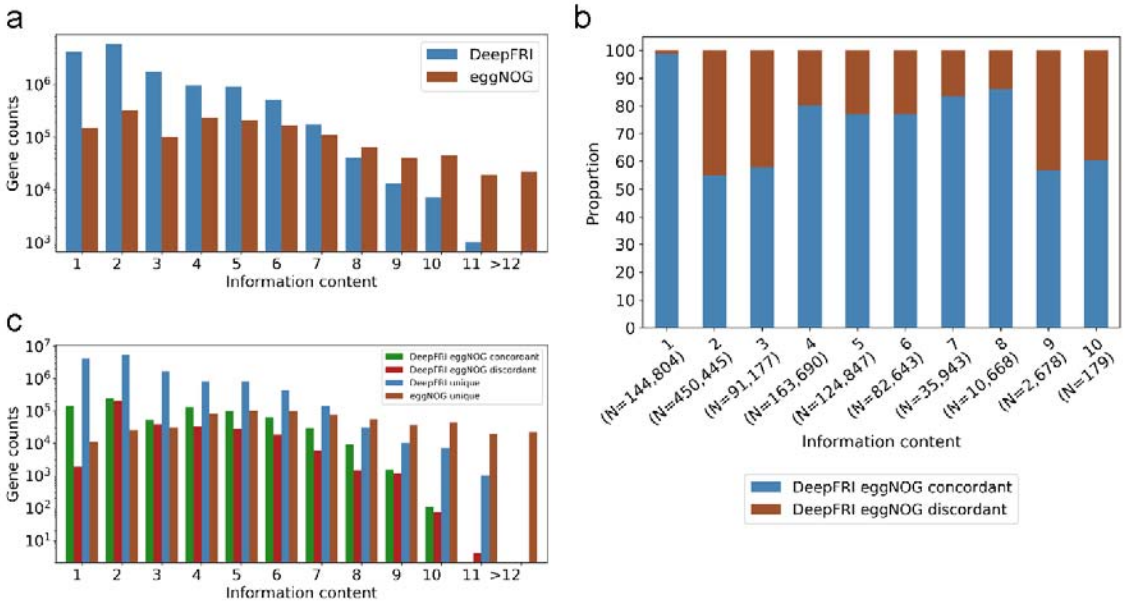

Figure 2: Concordance between DeepFRI and eggNOG annotations. (a) Difference in the annotation information content between eggNOG and DeepFRI. (**b)** Proportion of concordant and discordant annotations between DeepFRI and eggNOG per each information content level. (**c)** Consensus between DeepFRI and eggNOG annotations. See Table S2 for the list of information content values.

We then looked into functions encoded in the microbial genes. In total nearly all the genes in the gene catalogue (1,895,540; 99%) obtained GO molecular function annotations by DeepFRI at 0.2 probability threshold whereas only (219,634; 12%) of genes obtained similar annotations by eggNOG (Table 2). Further, we complemented the annotations by considering functional annotations from eggNOG free text description which consolidates information obtained from several source databases such as SMART/Pfam (Huerta-Cepas et al., 2019). A total of 1,372,653 (72%) genes obtained functional information from eggNOG free text description. We then compared differences in gene sets annotated by both methods. The venn diagram (Figure 3b) shows the intersection among genes annotated by DeepFRI and

eggNOG, with 219,003 genes in common across the two gene sets. A total of 1.6M genes were unique to DeepFRI and 631 genes unique to eggNOG gene ontology. Additionally, 1.3M gene sets were common across DeepFRI and eggNOG free text description annotations (Figure 3c).

The gene ontology protein function description is represented as a directed acyclic graph (DAG), such that a gene/protein can have multiple GO terms annotated, and a term can have multiple relationships to broader parent terms and more specific child term (child-parent relationship). For example, the molecular function term 'ATP binding' has parent terms 'binding'. For downstream analysis, we filtered out the general gene ontology terms using a subset of 614 informative GO term list obtained from previous work by Vatanen et al., 2017, following the method proposed in (Huang et al., 2007; Zhou et al., 2002). The informative GO categories are obtained by traversing the GO tree from the root and selecting terms that satisfy the following parameters; (i) terms that are associated with more than k=2,000 proteins and (ii) each of their descendant term contains less than 2,000 proteins (k=2,000 equates to approximately 1 of every 5,000 UniRef50 protein families). This gene ontology informative set provides more resolution to extensively studied processes. Filtering of gene ontology terms resulted in 365,472 genes annotated by DeepFRI and 156,704 genes annotated by eggNOG. The Venn diagram (Figure 3d) shows the intersection of 84,089 genes common between DeepFRI and eggNOG gene ontology. A total of 281,383 genes were unique to DeepFRI and 72,615 genes unique to eggNOG.

Finally, we evaluated the relative abundance of annotated genes between DeepFRI and eggNOG considering only genes annotated with informative gene ontology sets. We computed the proportion of metagenomic gene abundance with functional annotation as follows: (1) paired-end reads are mapped to the gene catalogue using the KMA tool, (2) per-gene alignment statistics are then weighted based on the target sequence length to produce abundance values for each gene family. We observed an increase in the annotation coverage at metagenome level by DeepFRI with an average coverage of 25.6% in comparison to eggNOG which had an average coverage of 14.5% (Figure 3a). The main molecular functions annotated by DeepFRI and eggNOG are ATP binding (GO:0005524), structural constituent of ribosome

(GO:0003735) and zinc ion binding (GO:0008270) (Supplementary figure 4).

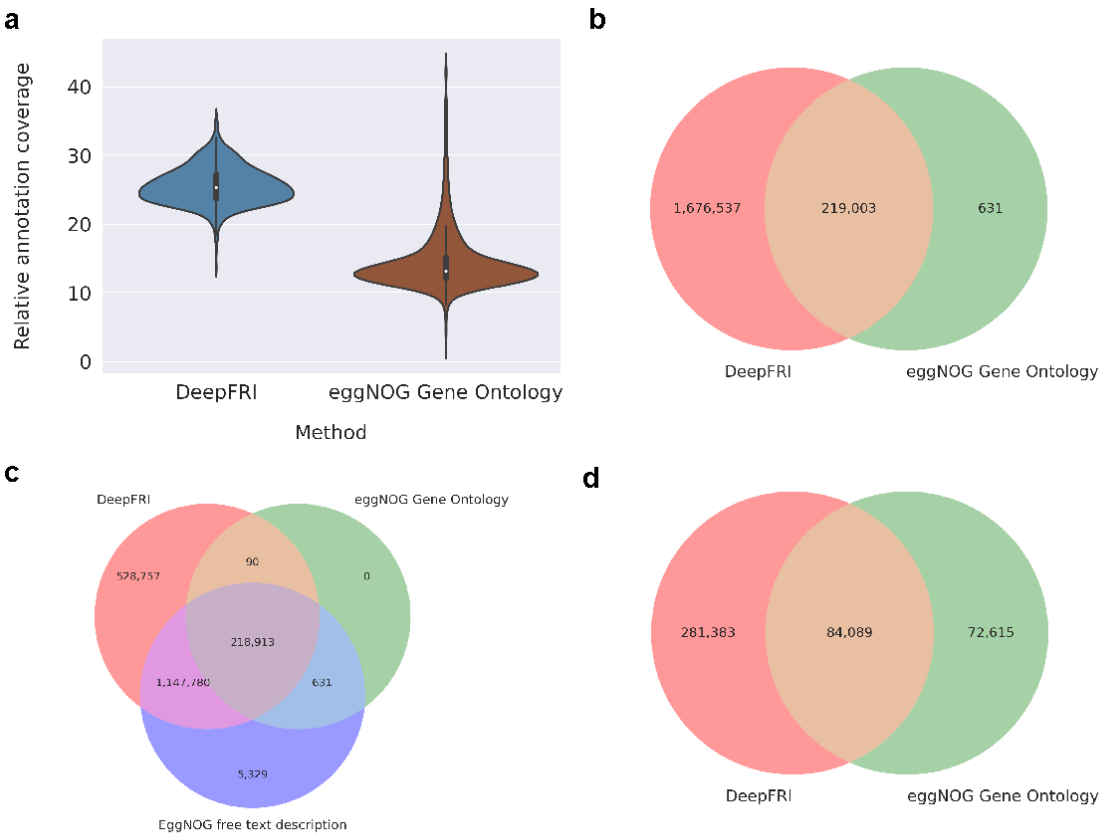

Figure 3: Functional capacity of the gut microbiota. (a) Proportion of metagenomics gene abundance with functional annotation by DeepFRI and eggNOG (informative gene ontology terms). The annotation is weighted by relative abundances normalized in copies per million. (**b)** Venn diagram of the number of gene sets annotated by DeepFRI and eggNOG gene ontology (all GO terms). (**c)** 3-way comparisons of gene sets annotated by DeepFRI, eggNOG gene ontology (all GO terms) and eggNOG free text description. (**d)** Venn diagram comparisons of gene sets annotated by DeepFRI and eggNOG using only informative gene ontology terms.

Table 2: Genes annotated by DeepFRI and eggNOG

Method	**DeepFRI molecular function**	**eggNOG Molecular function**	**eggNOG free text description**
Annotated genes	1,895,540 (99%)	219,634 (12%)	1,372,653 (72%)

Pan-genome diversity patterns within the infant metagenomes

The pan-genome can be defined as the entire gene repertoire of all strains in a species (Tettelin et al., 2005). Genes in a pan-genome are classified into two categories: core and accessory genes. The core genes are shared by genomes within a species while accessory genes are present in a subset of the genomes within a species. The pan-genomes were constructed in the following manner: we recruited near-complete metagenome assembled genomes (≥90% completeness and <5% contamination) and included species with at least ten independent near-complete MAGs. We then considered the genes present in ≥90% of MAGs of each species as “core” genes (accounting for the incompleteness of the MAGs), while genes present in <90% of MAGS were considered as “accessory” genes. Combined core and accessory genes constituted the pan-genome of a species.

The pan-genomes covered a total of 42 bacterial species (Supplementary figure 3a) and consisted of a total of 70,997 core genes (on average, 1,690 core genes per pan-genome, ranging from 573 to 3,612 genes) and 355,761 accessory genes (on average, 8,470 accessory genes per pan-genome, ranging from 1,222 to 27,879 genes) (Supplementary figure 3b). Each species had distinctly different core genome and pan-genome sizes. Our pan-genome analysis showed that *Veillonella parvula* and *Bifidobacterium pseudocatenulatum* contained the smallest core genomes with 573 and 753 genes respectively. In contrast, *Escherichia coli* contained the largest core genome (3,612 genes). To investigate differences in intra-species gene richness, we analysed the pan-genome size in relation to the number of genomes (MAGs). We found some associations between the number of MAGs and the size of the pan-genome (Figure 4c). *Bacteroides ovatus and Blautia wexlerae* had a strong negative correlation between the number of genomes and the size of the associated pangenome (Pearson correlation r -0.1). Additionally, *Bifidobacterium longum, Bifidobacterium bifidum and Akkermansia muciniphilla* showed moderate correlation (Pearson correlation r 0.4). However, this association was not observed in all the species, *Prevotella copri* and *Agathobacter faecis* had the same number of MAGs (a total of 31 MAGs) but different pan-genome sizes of 9,348 and 10,307 genes.

To explore the diversity within the pangenomes, we compared the accessory gene repertoire of each species. We observed a wide variation in the gene content between various species. *Bacteroides* species had the highest number of shared genes (Figure 4b). This observation is consistent with a recent study that showed that *Bacteroides* species exchange genes within the genus more frequently through horizontal gene transfer compared to other species such as *Bifidobacterium* (Groussin et al., 2021). Interestingly, species such as *Veillonella parvula*, *Akkermansia muciniphila*, and *Parasutterella excrementhominis, Sutterella wadsworthensis* accessory genomes harbored very low numbers of shared genes 76, 56, 6 and 1 respectively and the rest of the accessory genome was unique (Figure 4b). *Bacteroides dorei* is one of the most diverse and prominent members in the infant gut playing an essential role in immune activation (Vatanen et al., 2017, Vatanen et al., 2019). We wanted to look into its genomic diversity by comparing its accessory genome to other *Bacteroides* species. The analysis showed that *Bacteroides dorei* harboured 7,514 unique accessory genes representing 38% of its accessory genome (Supplementary figure 3c). One plausible explanation for this diversity could be caused by gene gain through horizontal gene transfer.

To obtain a better understanding of the functions encoded in the core and accessory genomes, we annotated the pan-genomes using DeepFRI and eggNOG. Here, we used only informative GO terms. There was a striking difference between DeepFRI and eggNOG annotations. The results show that eggNOG was able to ascribe more functions coming from well-studied species *such as Escherichia coli.* The core genome of *Escherichia coli* had the highest annotation level, (1,449; 40%) of core genes obtaining GO term annotations from eggNOG and (880; 24%) from DeepFRI (Figure 4a). This could be attributed to the fact that *E. coli* is an extensively studied model organism (Ghatak et al., 2019). Moreover, this observation was supported by our phylogeny analysis that showed that eggNOG annotated more genes belonging to *Gammaproteobacteria* (Figure 5b). On the other hand, DeepFRI is not sensitive to taxa, we observed diversity of functional enrichment between core and accessory genomes. Functional enrichment of DeepFRI annotations was high in *Bacteroides vulgatus*, with a total of 4,383 (7%) annotated genes compared to eggNOG which annotated 928 (4.5%) genes.

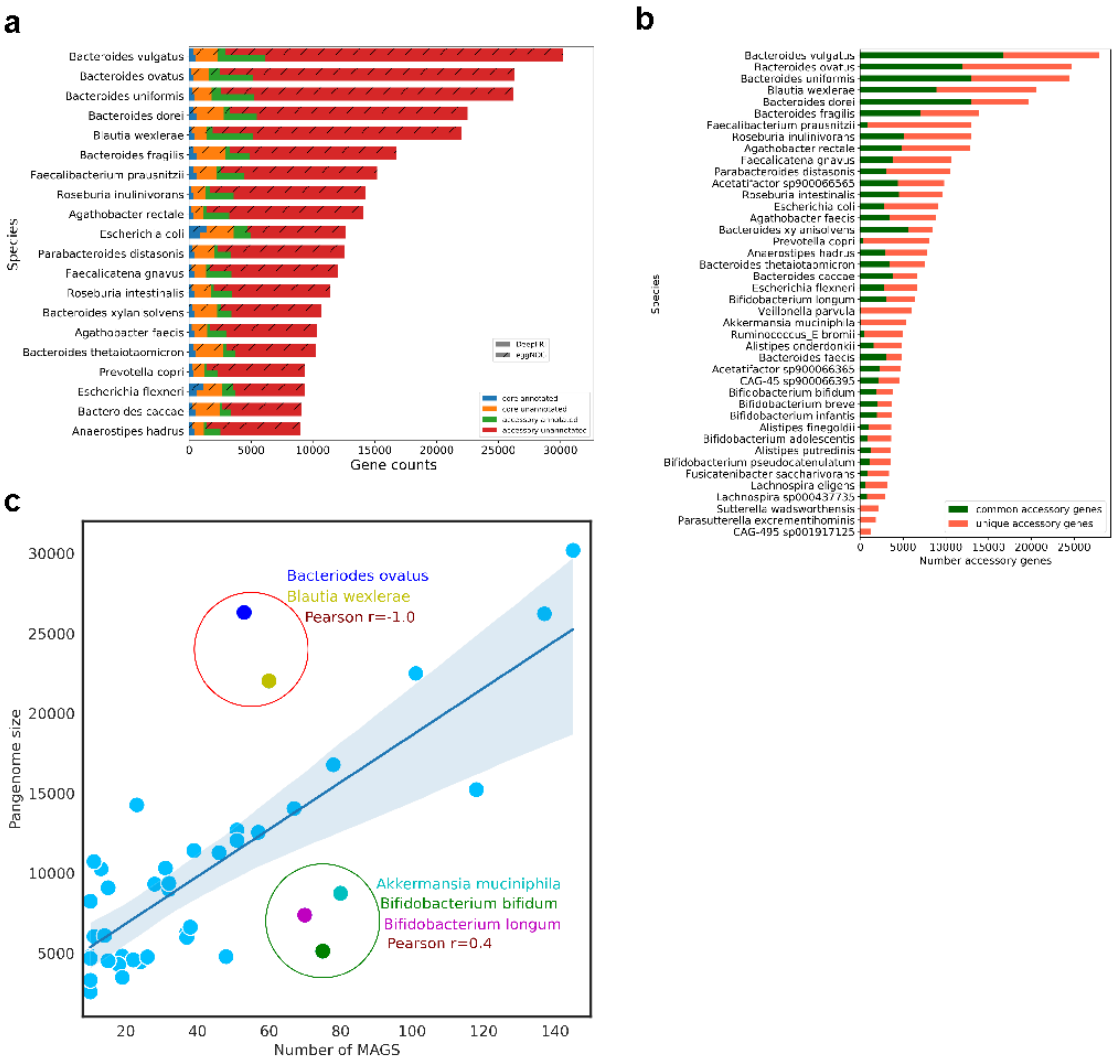

Figure 4: Pan-genome patterns within the infant metagenomes. (**a)** Size of the core and accessory genomes per species stratified by the functional annotation of genes using eggNOG and DeepFRI (known versus unknown function). Entries are ordered according to the size of the pangenome (20 out of 42 species used to construct the pangenomes). (**b)** Number of unique (not shared with other species) and shared accessory genes per pangenome. See Table S1 for a full pan-genome size list. (**c)** Pan-genome size in relation to the number of genomes (MAGs)

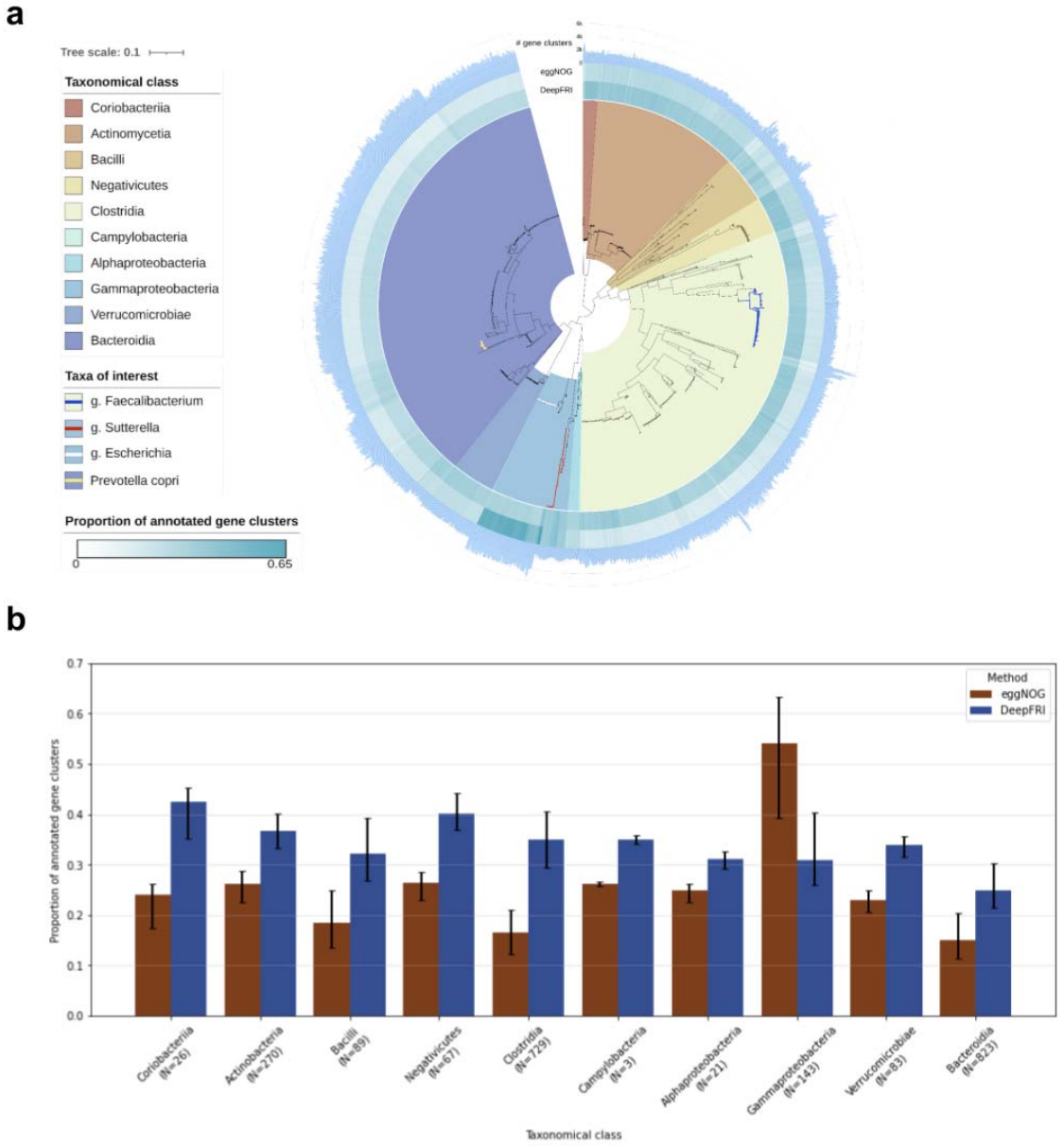

Figure 5: Phylogenetic analysis and taxonomic annotation of high-quality metagenome-assembled genomes. (a) Maximum-likelihood phylogenetic tree of 2,255 high-quality near-complete genomes. The taxonomy of the MAGs was assigned by GTDB-Tk. The innermost layer corresponds to 10 bacteria class. The second and third ring represent the proportion of genes annotated by DeepFRI and eggNOG. Bars in the outermost layer indicate the number of gene families per MAG. (**b)** Distribution of annotated genes per class

Discussion

Functional assignment of human gut metagenomic sequences is crucial for unlocking the microbiome clinical potential. Here we provide a standalone and fully automated workflow implementing metagenomic assembly, construction of MAGs and non-redundant gene catalogue and comprehensive gene function predictions. The workflow performs different steps of a metagenomics analysis in a highly reproducible and scalable workflow using Workflow Description Language (WDL). Its features provide functional inference of each gene, including taxonomic classification through better linkage of marker genes in the GTDB reference tree, as well as visualization of near-complete genomes. The workflow provides full-length complete genes required for a robust understanding of the functional repertoire encoded in the microbial genomes.

To demonstrate the usage of our pipeline, we tested it on infant gut metagenomes from the DIABIMMUNE cohort (Vatanen et al., 2019). Metagenomic single-sample assembly strategy employed here (Li et al., 2015) represents a scalable methodology for large-scale metagenomes analysis and profiling of functional diversity within the microbial communities. Additionally, the pipeline supports binning of contigs to high quality genomes using Metabat2 (Kang et al., 2019). We recovered MAGs with high completeness (90%) and low contamination (5%). Construction of complete or near-complete MAGs have enabled identification of valuable new genomes from rare species and quantification of intrapopulation diversity (Albertsen et al., 2013; Bowers et al., 2017; Wibowo et al., 2021). It has also been shown to identify novel genes with distinct functional properties associated with human diseases (Nayfach et al., 2019).

For functional annotation we utilized the advantage of deep learning method ability to process massive amounts of data. Graph convolutional networks (GCN)-based DeepFRI ensures efficient and accurate mapping of protein sequences to function annotation at a large scale. It enables discovery of novel protein functions by combining two aspects of information, features learned from protein sequences and contact graphs derived from 3-dimensional structures (Gligorijevic et al., 2021).

Combining DeepFRI with commonly used orthology-based method eggNOG not only improves our annotation capabilities but also helps us understand function better. Moreover, we provide an additional layer of function information obtained from eggNOG such as Smart/Pfam annotations (Mistry et al., 2021).

Metagenomics provides a great potential to characterize the functional capacity of the microbial populations, however, the functional diversity within the human gut microbiome remains poorly characterized as most genes still lack functional annotation (Almeida et al., 2021). By leveraging high quality protein information, we demonstrate that DeepFRI significantly improved the functional annotation coverage of the assembled genes (99%). Further validation of the GO terms predicted by DeepFRI showed a high level of concordance with eggNOG annotations, 70% on average. This indicates that DeepFRI produces reproducible and relevant predictions for the biological interpretation. We show that despite DeepFRI giving less specific (generic) GO term annotations, it provides a more comprehensive functional landscape of almost all metagenome-assembled genes. This increase in annotation coverage represents a key step towards novel understanding of the human gut microbiome, thus alleviating this sequence-to-function knowledge gap. Moreover, we provide an important functional resource for future interventions leveraging the gut microbiome to improve health.

Another critical aspect of our workflow is the full support of pan-genome analysis. Pan-genome analysis provides an undiscovered wealth of information for studying the diversity within species. Here we provide a way to construct pangenomes in a reference-free manner from near-complete metagenome assembled species (≥90% completeness and < 5% contamination). MAGs serve as an important resource in pan-genome analyses, especially where unculturable species or species without reference genomes are being studied. The constructed pan-genomes were of high quality and thus provided a better understanding of the functional repertoire encoded in core and accessory genes. The functional predictions generated from the species pan-genomes revealed a striking difference between DeepFRI and eggNOG annotations. Notably, DeepFRI annotated more functions coming from members of *Bacteroides* species. The pangenome analysis done here could be leveraged for

further studies to understand species-specific gene functions or identification of newly acquired functions of biomedical relevance.

There exist certain limitations in the workflow. First, we only tested our workflow on the infant gut microbiome. More evaluation of the workflow is required to demonstrate the performance on other complex metagenomes such as human skin metagenomes, environmental microbiome (seawater/soil metagenomes) and other diabetes specific benchmarks. Second, DeepFRI only provides GO annotations, therefore, it is difficult to profile microbial antibiotic resistomes (antibiotic-resistant genes), as well as genes encoding carbohydrate-active enzymes (CAZymes), which are important in health and disease.

Given the potential and limitations highlighted above. Future integrations of high-quality structure information coming from DeepFRI GCN predictions will allow large-scale functional annotations of metagenomics data with higher accuracy. Additionally, we also plan to incorporate KEGG orthology (Kanehisa et al., 2017) and evaluate genes and pathways associated with disease.

Conclusions

We provide a robust and reproducible pipeline for large-scale metagenomics functional annotation, where quality controlled short read sequences can be fully processed into annotated files for downstream analysis and visualization. The pipeline is available as a Docker image and makes use of standard tools for metagenome analysis and functional annotation. Its implementation in the WDL workflow gives it the advantage of extensive parallelization using high performance computing and cloud computing environments. Additionally, its modular architecture enables tool versioning while allowing users to tweak the different parameters to their specific needs. The Docker images used in the workflow are publicly available at https://github.com/microbiome-immunity-project/metagenome_assembly

Code availability. The Docker images and scripts used in the workflow are available at https://github.com/microbiome-immunity-project/metagenome_assembly

Acknowledgements

MM and TK are supported by grant NAWA PPN/PPO/2018/1/00014. TV acknowledges the use of Broad Institute computing facility.

Author contributions

MM – performed analyses, contributed analytical tools, wrote the paper
PS – contributed analytical tools, performed analyses
VB – performed analyses
VG – contributed analytical tools
RB – supervised the project
TK – conceived the study, supervised the project, wrote the paper
TV – conceived the study, contributed analytical tools, supervised the project, wrote the paper

Competing interests

All authors declare that there is no competing interest

References

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–538. https://doi.org/10.1038/nbt.2579

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., & Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, *568*(7753), 499–504. https://doi.org/10.1038/s41586-019-0965-1

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146.

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., Sanders, J. G., 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. *Nature Communications*, *11*(1), 1–10. https://doi.org/10. 1038/ s41467-020-16366-7

Baric, R. S., Crosson, S., Damania, B., Miller, S. I., & Rubin, E. J. (2016). Next-generation high-throughput functional annotation of microbial genomes. *MBio*, *7*(5), 7–10. https://doi.org/10.1128/mBio.01245-16

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. Da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa 1100

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *ELife*, *10*, 1–42. https://doi.org/10.7554 /eLife.65088

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10. 1093/ bioinformatics/btu170

Bose, T., Haque, M. M., Cvsk, R., & Mande, S. S. (2015). COGNIZER: A framework for functional annotation of metagenomic datasets. *PLoS ONE*, *10*(11), 1–16. https://doi.org/10.1371/journal.pone.0142102

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731. https://doi.org/10.1038/nbt.3893

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still going strong. *Nucleic Acids Research*, *47*(D1), D330–D338. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Clausen, P. T. L. C., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics*, *19*(1), 1–8. https://doi.org/10.1186/s12859-018-2336-6

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Eng, A., Verster, A.J., Borenstein, E. (2020). MetaLAFFA: a flexible, end-to-end, distributed computing-compatible metagenomic functional annotation pipeline. *BMC Bioinformatics*, *21*(1), 1-9. https://doi.org/10.1186/s12859-020-03815-9

Franzosa, E. A., Mciver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*, *15*(11), 962–968. https://doi.org/10.1038/s4

1592-018-0176-y

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gaudet, P., Chisholm, R., Berardini, T., Dimmer, E., Engel, S. R., Fey, P., Hill, D. P., Howe, D., Hu, J. C., Huntley, R., Khodiyar, V. K., Kishore, R., Li, D., Lovering, R. C., McCarthy, F., Ni, L., Petri, V., Siegele, D. A., Tweedie, S., & Lewis, S. E. (2009). The gene ontology's reference genome project: A unified framework for functional annotation across species. *PLoS Computational Biology*, *5*(7), 1–8.

Ghatak, S., King, Z. A., Sastry, A., & Palsson, B. O. (2019). The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Research*, *47*(5), 2446–2454. https://doi.org/10.1093/nar/gkz030

Gligorijevic, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B., Fisk, I., Vlamakis, H., Xavier, R., Knight, R., Cho, K., & Bonneau, R. (2021). Structure-Based Protein Function Prediction using Graph Convolutional Networks. *Nature Communications*, *12*(3168), 1–14.

Groussin, M., Poyet, M., Sistiaga, A., Kearney, S. M., Moniz, K., Noel, M., Hooker, J., Gibbons, S. M., Segurel, L., Froment, A., Mohamed, R. S., Fezeu, A., Juimo, V. A., Lafosse, S., Tabe, F. E., Girard, C., Iqaluk, D., Nguyen, L. T. T., Shapiro, B. J., &Alm, E. J. (2021). Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*, *184*(8), 2053-2067.e18.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.

Nucleic Acids Research, *47*(D1), D309–D314.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 1–11. https://doi.org/10.1186/1471-2105-11-119

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *2019*(7), 1–13.

Keegan, K. P., Glass, E. M., & Meyer, and F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods in Molecular Biology*, *1399*(6000), 167–182. https://doi.org/10.1007/978-1-4939 -3369-3

Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, *34*(4), 660–668. https://doi.org/10.1093/bioinformatics /btx624

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Letunic, I., & Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), 256–259. https://doi.org/ 10.1093/nar/gkz239

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*(10), 1674–1676. https://doi.org/ 10.1093/bioinformatics/btv033

Li, J., Wang, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., & Bork, P. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, *32*(8), 834–841. https://doi.org/10.1038/nbt.2942

Martínez Cuesta, S., Rahman, S. A., Furnham, N., & Thornton, J. M. (2015). The Classification and Evolution of Enzyme Function. *Biophysical Journal*, *109*(6), 1082–1086. https://doi.org/10.1016/j.bpj.2015.04.020

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, *568*(7753), 505–510. https://doi.org/10.1038/s41586-019-1058-x

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., & Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/ 10.1093/nar/ gkv1189

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., 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(3), 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

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, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., & McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071.

Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. E. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Frontiers in Microbiology*, *3*(DEC), 1–11. https://doi.org/10.3389/fmicb.2012.00410

Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., & Atalay, V. (2019). DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Scientific Reports*, *9*(1), 1–16. https://doi.org/10.1038/s41598-019-43708-3

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit Y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., & Fraser, C. M. (2005). Erratum: Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome” (Proceedings of the National Academy of Sciences of the United States of America (September 27, 2005) 102, 39 (13950-13955)). *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16530

Vatanen, T., Kostic, A. D., D’Hennezel, E., Siljander, H., Franzosa, E. A., Yassour, M., Kolde, R., Vlamakis, H., Arthur, T. D., Hämäläinen, A. M., Peet, A., Tillmann, V., Uibo, R., Mokurov, S., Dorshakova, N., Ilonen, J., Virtanen, S. M., Szabo, S. J., Porter, J. A., & Xavier, R. J. (2017). Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*, *165*(4), 842–853.

https://doi.org/10. 1038/s41586-021-03532-0

Vatanen, T., Franzosa, E. A., Schwager, R., Tripathi, S., Arthur, T. D., Vehik, K., Lernmark, Å., Hagopian, W. A., Rewers, M. J., She, J. X., Toppari, J., Ziegler, A. G., Akolkar, B., Krischer, J. P., Stewart, C. J., Ajami, N. J., Petrosino, J. F., Gevers, D., Lähdesmäki, H., & Xavier, R. J. (2018). The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature*, *562*(7728), 589–594. https://doi.org/10.1038/s41586-018-0620-2

Vatanen, T., Plichta, D. R., Somani, J., Münch, P. C., Timothy, D., Hall, A. B., Rudolf, S., Oakeley, E. J., Ke, X., Young, A., Haiser, H. J., Kolde, R., Yassour, M., & Luopajärvi, K. (2019). Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol.*, *4*(3), 470–479.

Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., Martínez-Hernández, A., Luber, J. M., Kirstahler, P., Blohm, T., Smiley, F. E., Arnold, R., Ballal, S. A., Pamp, S. J., Russ, J., & Kostic, A. D. (2021). Reconstruction of ancient microbial genomes from the human gut. *Nature*, *594*(7862), 234–239. https://doi.org/10.1038/s41586-021-03532-0