StrainPanDA: linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data

Han Hu^{1,2*}, Yuxiang Tan^{1,*}, Chenhao Li^{3,*}, Junyu Chen¹, Yan Kou², Zhenjiang Zech Xu⁴, Yang-Yu Liu⁵, Yan Tan^{2,#}, Lei Dai^{1,#}

¹ CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

² Xbiome, Scientific Research Building, Tsinghua High-Tech Park, Shenzhen, China

³ Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

⁴ State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang, China.

⁵ Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

*: These authors contributed equally to this work

#: Correspondence: lei.dai@siat.ac.cn, yant@xbiome.com

Abstract

Background: Microbial strains of variable functional capacities co-exist in microbiomes. Current bioinformatics methods of strain analysis cannot provide the direct linkage between strain composition and their gene contents from metagenomic data.

Methods: Here we present StrainPanDA (Strain-level Pangenome Decomposition Analysis), a novel method that uses the pangenome coverage profile of multiple metagenomic samples to simultaneously reconstruct the composition and gene content variation of co-existing strains in microbial communities.

Results: We systematically validate the accuracy and robustness of StrainPanDA using synthetic datasets. To demonstrate the power of gene-centric strain profiling, we then apply StrainPanDA to analyze the gut microbiome samples of infants, as well as patients treated with fecal microbiota transplantation. We show that the linked reconstruction of strain composition and gene content profiles is critical for understanding the relationship between microbial adaptation and strain-specific functions (*e.g.*, nutrient utilization, pathogenicity).

Conclusions: StrainPanDA can be applied to metagenomic datasets to detect association between molecular functions and microbial/host phenotypes to formulate testable hypotheses and gain novel biological insights at the strain or subspecies level.

Introduction

There is mounting evidence that multiple within-species variants co-exist in microbiomes^{1,2}. Co-existing microbial cells of the same species can have substantial variations in their gene contents (*i.e.*, accessory genome), which is largely generated by horizontal gene transfer (HGT)^{3–5}. The intraspecies variation in accessory genome can lead to substantial phenotypic differences (*e.g.*, nutrient utilization, pathogenicity, antibiotic resistance) and plays an important role in microbial adaptation across environments^{6–9}. Moreover, many health outcomes linked to host-associated microbiomes have been found to be consequences of the function of individual strains^{8,10–14}.

Metagenomic sequencing has revolutionized microbiome studies by providing a culture-independent approach to study the composition and function of complex microbial communities. Commonly used tools for metagenomic analysis, known as metagenomics profilers, typically provide species-level taxonomic composition^{15–18}. In parallel with the rapid increase of sequenced microbial isolates from culturomics studies^{1,9,19,20}, high-resolution analyses of metagenomic data have revealed notable within-species variations^{21,22}. Methods that enable strain-level analysis of metagenomes have been used for tracking strain transmission or dispersal^{23,24}, studying population genetics of microbial strains²⁵, and typing strains of specific interest^{26–32}.

The gene content profile of a microbial strain determines its biological function. To date, the majority of strain-level analysis methods use single nucleotide variants (SNVs) to identify strain composition^{25,28,29,33,34}. By assuming an association between SNV haplotypes and gene content profiles⁴, SNV-based methods can indirectly profile the within-species gene content variation. However, for many species, it has been shown that SNV haplotypes cannot capture microbial genetic diversification resulting from HGT⁴. Alternatively, current pangenome-based method can infer the gene content of the dominant strain in a metagenomic sample³⁵, but fails to provide the abundance and gene contents of co-existing strains within the sample. Establishing the linkage of composition and gene contents of co-existing within-species variants can provide crucial insights into microbial adaptation and microbiome-host interactions, but is inaccessible from currently available reference-based bioinformatics tools.

To meet the increasing needs of strain-level functional inference from metagenomics data, we developed a novel method StrainPanDA (Strain-level Pangenome Decomposition Analysis) to simultaneously reconstruct the composition and gene contents of co-existing strains using the pangenome coverage profile from metagenomic data (**Fig. 1**). We validated the performance of StrainPanDA with a comprehensive collection of synthetic datasets and showed that StrainPanDA was able to accurately infer strain composition and gene content profiles from metagenomic data. To demonstrate the practical use of StrainPanDA in human microbiome studies, we analyzed longitudinal gut microbiome samples of mother-infant pairs³⁶ and patients treated with fecal microbiota transplantation (FMT)^{29,37}. We found that StrainPanDA was able to identify association between strain-specific functions and microbial adaptation (or host phenotypes), leading to novel biological insights at the infraspecific level and testable hypotheses of molecular mechanisms.

Results

Decomposition of the pangenome coverage profile to infer strain composition and gene content

The pangenome coverage profile of a microbial species from metagenomic data is composed of the gene contents of all co-existing strains. If there are multiple metagenomic samples with varying strain compositions, in principle it is possible to infer the composition of strains within the sample as well as the gene contents of each strain from the pangenome coverage profile³⁸. Building on this intuition, the main algorithm of StrainPanDA aims to decompose the gene family abundance data matrix $\mathbf{D}$ into the product of two matrices, the gene content profile matrix $\mathbf{P}$ and the strain composition matrix $\mathbf{S}$ (**Fig. 1A**, see **Methods** for details). Here the pangenome coverage profile from metagenomic data is represented by matrix $\mathbf{D}$, where D_{ij} is the normalized count of gene family i in metagenomic sample j . The gene contents of co-existing strains are represented by binary matrix $\mathbf{P}$, where the element P_{ij} indicates the presence/absence of gene family i in strain j . The composition of co-existing strains across samples is represented by $\mathbf{S}$, where S_{ij} is the relative abundance of strain i in sample j ($S_{ij} \geq 0$ and $\sum_i S_{ij} = 1$). In the implementation of StrainPanDA, the gene family abundance matrix $\mathbf{D}$ is decomposed by non-negative matrix factorization (NMF)^{39–41} to solve for matrices $\mathbf{P}$ and $\mathbf{S}$. This processing allows

StrainPanDA to simultaneously delineate the composition and gene contents variation of co-existing strains.

The StrainPanDA software provides a fully automated workflow of strain analysis (**Fig. 1B**), which imports raw sequencing data from multiple metagenomic samples, performs reads mapping, strain decomposition, and downstream annotations (see **Methods** for details). To assist the interpretation of gene content variation in strains, StrainPanDA incorporates functional annotation from several databases, including but not limited to KEGG⁴², CAZy⁴³, and VFDB⁴⁴.

StrainPanDA provides accurate predictions of strain composition and gene family profiles in synthetic data

We validated the performance of StrainPanDA using synthetic metagenomic data (**Methods**). For synthetic mixtures of *E. coli* strains (ranging from 2 to 8 strains, see **Methods**), the strain composition predicted by StrainPanDA was close to the actual composition (Ground Truth) (**Fig. 2A**). For quantitative comparison, we calculated the Jensen-Shannon divergence (JSD) and Matthews Correlation Coefficient (MCC) between the predicted and actual strain composition of simulated samples. JSD and MCC have been widely used in the evaluation of strain analysis tools^{28,30,33}. The predicted strain composition by StrainPanDA was better than or similar to the state-of-the-art SNV-based methods, including StrainEst³⁰ and PStrain³³ (the latter was modified based on ConStrains²⁸) (**Fig. 2B, Supplementary Fig. 4**). Furthermore, we generated synthetic mixtures of *E. coli* strains with varying levels of sequencing errors (**Supplementary Fig. 2**), sequencing depths (**Supplementary Fig. 3**) and different background noise (mixed with different metagenomic datasets, **Supplementary Table 1, Supplementary Fig. 4**). In comparison to SNV-based methods, the performance of StrainPanDA in predicting strain composition was robust.

To evaluate the performance of StrainPanDA in different bacterial species, we generated synthetic data for common human gut bacterial species (*Bifidobacterium longum*, *Clostridium difficile*, *Enterococcus faecalis*, *Faecalibacterium prausnitzii* and *Prevotella copri*) (**Supplementary Table 2**, see **Methods**). In comparison to other methods, StrainPanDA made the most accurate prediction of strain composition across all species (with JSD as 0.021 ± 0.006) (**Fig. 2C**).

While current SNV-based methods can reconstruct the composition of co-existing strains from metagenomic samples, they could not directly provide the gene contents of the predicted strains. Here we show that StrainPanDA allows simultaneous reconstruction of strain composition in each metagenomic sample and the gene content variations among strains. In synthetic mixtures of *E. coli* strains, the predicted gene family profiles by StrainPanDA were close to the actual profiles (**Fig. 2D**, **Supplementary Fig. 5**; precision=0.91-0.96, recall=0.87-0.96, for the 4 strains in synthetic mixture). In particular, we note that StrainPanDA is able to infer the gene family profile of strains not included in the pre-built reference genome database. The area under the Precision-Recall Curve (AUPRC) was over 0.95 for all *E. coli* strains, indicating that StrainPanDA was able to reconstruct the gene contents of microbial strains with high sensitivity and precision (**Supplementary Fig. 6**).

We further evaluated the predicted gene family profiles of the human gut bacterial species included in the synthetic data. The AUPRC was on average above 0.9 and significantly better than random guesses (**Fig. 2E**). The predicted gene family profiles were robust to sequencing errors, sequencing depths and the background of real metagenomic data (**Supplementary Table 4**). Moreover, to demonstrate the ability of StrainPanDA to identify strain-specific genes, the pathogenic *E. coli* outbreak strain O104⁴⁵ was introduced in a synthetic mixture with other *E. coli* strains (**Supplementary Table 5**). All outbreak-related gene families were successfully recovered by StrainPanDA (**Supplementary Fig. 7**). Finally, the predicted gene family profiles for the dominant strain from StrainPanDA were overall more similar to the ground truth compared to the predictions from PanPhlAn³⁵, which is a pangenome based method that can only profile the gene content of the dominant strain in each sample (**Supplementary Fig. 8**).

Taken together, our benchmarking results demonstrate that StrainPanDA provides accurate predictions of compositional profiles and gene contents of co-existing strains from metagenomic samples. In the following sections, we will demonstrate the application of StrainPanDA in two longitudinal metagenomic datasets to elucidate the diversity of gut microbiome at the subspecies level.

Succession of *B. longum* subspecies in infant gut microbiome is associated with breastfeeding patterns and the selection on nutrient utilization

The direct inference of both the population structure and gene content variation at the strain-level is crucial to understanding the ecology and evolution of microbial communities. Here we apply StrainPanDA to study the adaptation of co-existing bacterial subspecies in the infant gut microbiomes. We analyzed a previously published dataset that includes gut metagenomic samples from ~100 mother-infant pairs (infants were sampled at three time points: newborn, 4-month and 12-month)³⁶ (**Supplementary Table 6**). At the species level, the authors found that the composition of infant gut microbiome had distinctive features at each sampled time point, and the cessation of breastfeeding was clearly associated with the maturation of an infant gut microbiome into an adult-like microbiome³⁶.

We focused on the infraspecific analysis of *Bifidobacterium longum*, which is known to play an important role in the development of the infant gut microbiome^{36,46–50} and was found to be enriched in 4-month infant samples in this study (**Supplementary Fig. 9**). Interestingly, we discovered a clear pattern of succession in the subspecies composition of *B. longum* over time, *i.e.*, a shift in the dominant subspecies (**Fig. 3A**). Among the three *B. longum* subspecies predicted by StrainPanDA, *B. longum* subspecies 2 was dominant in the gut microbiomes of mothers. In the gut microbiomes of infants, *B. longum* subspecies 3 was most prevalent for newborns, while subspecies 1 transiently increased at the intermediate time point (at 4-months) and then was outcompeted by subspecies 2 (at 12-months). Based on the diet history provided in the original study, we further grouped infant samples into two different categories, “discontinued breastfeeding” and “continued breastfeeding” between successive timepoints (**Methods**). We found clear evidence that the relative abundance of *B. longum* subspecies 1 was enriched in infants that continued breastfeeding (**Fig. 3B**). By contrast, once breastfeeding was discontinued, *B. longum* subspecies 1 was taken over by subspecies 2.

The association between *B. longum* subspecies composition and breastfeeding patterns suggests natural selection on nutrient utilization functions (**Fig. 3C**). Based on functional annotations of KEGG⁴² and CAZy⁴³, we found clear variations in nutrient utilization genes among the predicted *B. longum* subspecies. *B. longum* subspecies 1 had unique gene families (marked in red, **Fig. 3C**) that are key enzymes related to human milk oligosaccharide (HMO), including galactosidase, alpha-L-fucosidase,

sialidase and their corresponding CAZy groups (GH33, GH29 and GH95) (**Supplementary Fig. 12-13** and **Supplementary Table 7**). In addition, the urease-related gene families were only found in the gene family profile of subspecies 1. Therefore, the unique functional potential of *B. longum* subspecies 1 in utilizing HMO and urea⁵¹ from breast milk could confer a fitness advantage under breastfeeding, consistent with our observations of its transient dominance at 4 months in infant gut metagenomes.

Our finding is consistent with previous reports⁵² on HMO utilization genes in some *B. longum* strains as well as observed changes in the frequency of *B. longum* subspecies *infantis* after weaning^{4,46,47,53,54}. The functional profile of subspecies 1, in comparison to *B. longum* reference genomes, suggests that it may correspond to the *B. longum* subsp. *infantis* (**Supplementary Fig. 14**), which has been isolated from infants and known to be associated with breastfeeding^{46,55,56}.

Overall, we show that StrainPanDA is able to identify associations between strain-specific functions (via reconstruction of gene contents) and adaptation (via reconstruction of strain composition), leading to novel biological insights and testable hypotheses about microbial evolution at the subspecies level.

Analysis of post-FMT gut metagenomes reveals individualized subspecies profiles and subspecies-specific functions

Fecal microbiota transplantation (FMT) introduces bacterial strains from healthy donors into recipients and has profound impacts on the structure and function of the recipient's gut microbiota^{57,58}. Here we apply StrainPanDA to analyze the metagenomic samples from FMT recipients in a clinical trial to treat Crohn's disease³⁷, including 17 patients (8 in the FMT group and 9 in the sham group) and multiple samples (4-8 time points) for each patient. We analyzed the subspecies composition of commonly observed bacterial species in the human gut metagenome (**Supplementary Table 8**). Hierarchical clustering of the predicted subspecies compositional profiles revealed strong individual signatures, which remained stable throughout 24 weeks (**Fig. 4A**). The pairwise distance in subspecies composition profiles among samples of the same individual (sampled at multiple time points, *i.e.*, the "intrasubject" group) was significantly lower than the pairwise distance among samples of different individuals (*i.e.*, the "intersubject" group) (**Fig. 4B**, $p\text{-value} < 10^{-15}$), similar to the pattern at the species level.

Furthermore, we found that the dissimilarity in subspecies composition between paired FMT donors and recipients (*i.e.*, the “donor-recipient” group) was significantly lower than the “intersubject” group (p-value < 0.001), indicating the engraftment of donor strains and co-existence of donor and recipient strains³⁷. We noted that the engraftment of donor gut bacteria was more obvious at the subspecies level (effect size=1.4) than at the species level (effect size=0.78) (**Fig. 4B**). Similarly, we applied StrainPanDA to analyze an independent FMT dataset of metagenomic samples from patients with *Clostridium difficile* infection²⁹. We observed a clear pattern of subspecies engraftment in post-FMT gut metagenomes, consistent with SNV-based strain analysis in the original study²⁹ (**Supplementary Fig. 15**). Overall, we show that StrainPanDA is able to delineate the difference in subspecies composition among individuals and track the transmission of strains.

To elucidate the potential role of gut microbiome in the maintenance of remission in Crohn’s disease patients, we further investigated the strain-level genetic signatures associated with post-FMT clinical outcomes. The original study has noted the enrichment of Bacteroidetes species in patients relapsed after FMT³⁷. We focused our analysis on *Bacteroides ovatus*, which was found to be enriched in relapsed individuals (FDR-adjusted p-value = 0.2) (**Fig. 4C-E**). Among the predicted subspecies of *B. ovatus*, the relative abundance of subspecies 2 was positively correlated with the abundance of species-level *B. ovatus* in gut metagenomes (Spearman correlation = 0.64, and FDR-adjusted p-value < 10⁻⁶, **Supplementary Fig. 16**). We found substantial gene content variation among different *B. ovatus* subspecies (**Fig. 4D**). Interestingly, we found that *B. ovatus* subspecies 2 had more CAZy family genes than others, indicating its functional potential to utilize diverse carbon sources and potential fitness advantages (**Fig. 4E**). Thus, the strain-specific metabolic functions of *B. ovatus* subspecies 2 may explain its dominance within the species (~20%) as well as its positive correlation with species abundance (**Fig. 4C** and **Supplementary Fig. 16**). In addition, *B. ovatus* subspecies 2 carried several strain-specific virulence factor genes (*e.g.*, type IV secretion system, cholesterol-dependent cytolysin), which may contribute to the positive association between *B. ovatus* and post-FMT relapse (**Supplementary Fig. 16**). For example, cholesterol-dependent cytolysin is a pore forming toxin that can disrupt the host plasma membrane⁵⁹, whose integrity has been linked to inflammatory bowel disease⁶⁰. In addition, we noted that *B. ovatus* subspecies 4 was more abundant in the remission group

(**Fig. 4C**, FDR-adjusted p -value $< 10^{-5}$); whether this *B. ovatus* subspecies contributes to post-FMT remission remains to be validated in future studies. Similarly, we performed StrainPanDA analysis for *Bacteroides vulgatus*, which was also enriched in relapsed individuals, and found clear functional variation among its subspecies (**Supplementary Fig. 17**).

In summary, we show that the linkage of strain composition and gene contents provided by StrainPanDA can greatly facilitate our understanding of microbial ecology and evolution beyond the species level. For microbes closely related to host health, this linkage helps formulate testable hypotheses on the association between molecular functions (*e.g.*, pathogenicity) and clinical outcomes, which can be directly tested in experiments of isolated microbial strains.

Discussion

Here we report a novel method, StrainPanDA (Strain-level Pangenome Decomposition Aalysis), to simultaneously profile the composition of co-existing strains and their corresponding gene content from metagenomics data. Our benchmarking results showed that StrainPanDA provided accurate and robust predictions from synthetic data. The predicted strain composition was better than or comparable to state-of-the-art methods; meanwhile, the predicted gene content profile was close to the actual profile, even for strains not included in the pre-built reference genome database. Furthermore, we applied StrainPanDA to metagenomic datasets to resolve within-species variation of bacterial taxa of interest. For example, we found that the composition of *B. longum* subspecies in infant gut microbiomes was associated with dietary shifts and the unique functional potential of certain *B. longum* subspecies in utilizing nutrients from breast milk might confer a fitness advantage. We demonstrated that the linkage of strain abundance and gene contents could lead to direct functional interpretations and testable hypothesis.

To study within-species gene content variation, current SNV-based methods implicitly assume the association between SNV haplotypes and gene content. However, many microbial genomes with high similarity in core genome have less than 70% of genes in common³, indicating that the indirect inference of gene content by SNV-based methods may be insufficient. In contrast, StrainPanDA adopts the pangenome-based approach to directly infer the gene content of multiple co-existing within-species

variants. With increasing availability of sequenced microbial isolates from environmental and host-associated microbiomes, we expect the pangenome database to expand accordingly. The prediction of StrainPanDA relies on the pangenome database, but it is not limited to the profiles of available reference genomes; thus, StrainPanDA can also be used to identify novel strains, as long as the relevant gene families are included in the pangenome. Although we focused the comparison to reference-based methods, it is worth noting that complementary approaches based on metagenome-assembled genomes (MAGs) such as DESMAN^{61,62,63} can also identify novel strains from metagenomic data. However, the MAG-based methods require much higher sequencing depth than reference-based methods, which prohibits their application to species with low abundance. In comparison, sequencing depth is not a limiting factor for StrainPanDA (**Supplementary Fig. 3-4**).

StrainPanDA is designed to fit the analysis of multiple metagenomic samples with shared within-species variants, such as longitudinal studies. The performance of StrainPanDA, including the accuracy of predicted strain composition and gene content profiles, improves with sample size (**Supplementary Fig. 18**) and sequencing depth (**Supplementary Fig. 3**). To apply StrainPanDA on a typical metagenomic dataset, it would be desirable to have at least 10 samples and the relative abundance of the species of interest to be above 1%. Due to the nature of StrainPanDA's algorithm, it may be difficult to disentangle within-species variants with highly similar gene content profiles, thus StrainPanDA is most suitable for analysis at the level of subspecies³. Finally, in comparison to MAG-based methods, StrainPanDA has minimal requirements for computing resources (about 400 minutes on a 20-sample dataset for one species) and can be scaled to process multiple species in parallel.

Conclusions

In summary, we show that StrainPanDA is able to provide accurate profiling of strain composition and gene content from metagenomic data. We envision that the application of StrainPanDA to the rapidly increasing metagenomic datasets, especially in the context of spatio-temporal characterization of microbiomes⁶⁴⁻⁶⁷, will help elucidate novel associations between molecular functions and microbial/host phenotypes as well as microbial ecology and evolution at the infraspecies level.

Figure 2. Validation of StrainPanDA using synthetic metagenomic data. (A) Comparison between the actual strain composition (Ground Truth) and the strain composition predicted by StrainPanDA and existing tools (StrainEst and PStrain) in synthetic mixtures of *E. coli* strains (pWGS dataset, 1× sequencing depth, see **Methods**). The number of actual *E. coli* strains in the mixture ($n = 2, 4, 6$ and 8) is shown in rows. Each stacked bar is one simulated sample. Strains are displayed by the order of sorted relative abundance. If the number of predicted strains exceeds the number of actual strains, the extra strains are grouped into “Extras”. (B) Jensen-Shannon Divergence (JSD) between the actual and predicted strain composition. No BG: No Background; BG1/BG2/BG3: synthetic data of *E. coli* strains mixed with three different metagenomic datasets as background (WGSBG dataset, 100-fold background, see **Methods**). Each dot represents one simulated sample ($n = 20$). (C) JSD between the actual and predicted strain composition is evaluated for different microbial species. Each dot represents one simulated sample ($n = 24$). Outputs not available are marked as “NA”. (D) The actual and predicted gene family profiles of *E. coli* strains (the synthetic data used is the same as panel A). Each row is one gene family, and each column is one strain. Hierarchical clustering is based on Euclidean distance. (E) The area under the Precision-Recall Curve (AUPRC) for the gene family profiles of co-existing strains is evaluated for different microbial species. Each dot represents the AUPRC of one strain ($n = 4$ strains). Brown dots represent random guesses of gene family profiles (see **Methods**). P values from t-test: * $P < 0.05$, *** $P < 0.001$, **** $P < 0.0001$.

Figure 3. Succession of *B. longum* subspecies in infant gut microbiome can be attributed to the selection on nutrient utilization. (A) Ternary plots of the predicted composition of three subspecies of *B. longum* from mothers and infants of multiple time points (newborns, 4-month, 12-month). Each dot represents one sample. (B) The shift in the relative abundance of *B. longum* subspecies between

318 successive time points. According to the breastfeeding status at the subsequent time point, infants are
 319 divided into two groups (purple: discontinued breastfeeding, N=86; red: continued breastfeeding,
 320 N=71). ****: p-value <0.00005, ns: not significant, Student's t-test. (C) Gene family profiles of
 321 predicted *B. longum* subspecies. Gene families related to the metabolism of host glycans, urease and
 322 CRISPR (KEGG annotations) are selected for display. The color bar on the left indicates the class of
 323 gene clusters. Each row is a subclass of gene families and unique gene families of subspecies 1 are
 324 marked in red. The color scale in heatmap indicates the normalized gene family coverage in the specific
 325 subclass (*i.e.*, the fraction of detected gene families belonging to the subclass).

Figure 4. Analysis of post-FMT gut metagenomes reveals individualized subspecies profiles and the association between subspecies-specific functions and phenotypes. (A) Hierarchical clustering of predicted subspecies compositional profiles of common gut species (**Supplementary Table 8**) reveals strong individual signatures. The subject IDs were collected from the original paper (Kong *et al.* 2020) and marked by different colors. **(B)** The dissimilarity (Bray-Curtis dissimilarity) in subspecies composition and species composition between samples. The pairs are classified into three groups for comparison: intersubject (samples from different individuals), intrasubject (samples from the same individuals) and donor-recipient (FMT donor vs. the post-FMT sample of the recipient). **(C)** The relationship between the relative abundance of *Bacteroides ovatus* and its subspecies (normalized by centered log-ratio transformation). Lines represent fitted linear regression (shaded areas: 95% confidence interval). The density plots on the side shows the distribution of the corresponding variables. At the species level, *Bacteroides ovatus* is enriched in the relapse group. **(D)** The summary of pangenome information of predicted *B. ovatus* subspecies. **(E)** Gene family profiles of Carbohydrate-Active enZymes (CAZy) in predicted *B. ovatus* subspecies. CAZy genes shared by all the subspecies are not shown.

Materials and methods

Generation of pangenome database and mapping of metagenomic data

Pangenome database of bacterial species analyzed in this study was created following the steps recommended by PanPhlAn (version 1.2.8)³⁵. For each bacterial species, genomes were downloaded from NCBI. Average Nucleotide Identity (ANI) between genomes was calculated by mash (version 1.1)⁶⁸. Representative strains (pairwise ANI $\leq 99\%$) were selected and used as reference genomes for pangenome construction. The annotated genes were extracted from the reference genomes and clustered into gene families at 95% identity by usearch (v7)⁶⁹ to create the pangenome database. Shotgun metagenomic data was mapped to the pangenome database by PanPhlAn³⁵ (version 1.2.8), which used Bowtie2 (version 2.4.1)⁷⁰ and SAMtools (version 0.1.19)⁷¹ to map and count the reads, respectively. A gene family profile was generated by summing up the read counts of genes (normalized by RPKM) belonging to the same gene family. The gene family profiles of all metagenomic samples were grouped into a single gene family profile matrix. To account for potential noise in reads mapping, the gene family abundance was trimmed to 0 if RPKM value was below the cutoff (10, by default). After trimming, gene families absent in all samples were removed from further analysis. In addition, samples were filtered out if the number of gene families detected was below $0.9 \times g_{min}$ (g_{min} is the minimum number of gene families found in all reference genomes).

StrainPanDA algorithm

The core algorithm of StrainPanDA decomposes the gene family abundance data matrix ($\mathbf{D}$) of the microbial species of interest into the product of two matrices (**Fig. 1**):

$$\mathbf{D} = \mathbf{P} \cdot \mathbf{S}$$

The gene family abundance data matrix $\mathbf{D}$ is a $N \times S$ non-negative matrix, where D_{ij} is the normalized count of gene family i in metagenomic sample j . The gene content profile matrix $\mathbf{P}$ is a $N \times K$ binary matrix, where P_{ij} is 1 if the gene family i is present in strain j and 0 otherwise. The strain composition matrix $\mathbf{S}$ is a $K \times S$ matrix, where S_{ij} is the relative abundance of strain i in sample j ($S_{ij} \geq 0$ and

$\sum_i S_{ij} = 1$). N is the number of gene families in the pangenome of the microbial species of interest, S is the number of metagenomic samples, K is the number of strains (*i.e.*, factorization rank).

To estimate $\mathbf{P}$ and $\mathbf{S}$, we approximate the solution $\mathbf{P}'$ and $\mathbf{S}'$ using non-negative matrix factorization (NMF), considering the non-negative constraints on both matrices (optimized using the “snmf/r” algorithm implemented in the R package ‘NMF’^{40,41}, version 0.21.0). The addition of sparsity constraints (*i.e.*, regularization terms in the objective function) ensures uniqueness of factorization^{41,72}. The $\mathbf{S}'$ matrix is then scaled into relative abundances. We binarize the approximated $\mathbf{P}'$ matrix, following the assumption that the matrix elements corresponding to “present” gene families should have higher values than “absent” gene families, and the matrix elements should have a tight distribution due to the expectation that $\mathbf{P}$ is a binary matrix (see **Supplementary Fig. 1B**). Briefly, we find the peak of the probabilistic density curve (p_{max}) for each strain j , where the number of matrix elements on the right of the peak ($P_{ij} > p_{max}$) is equal to the expected number of gene families of the species of interest (*i.e.*, averaged over all reference genomes in the pangenome database). We then cut the density curve at θ between the selected peak and 0 ($\theta = 0.5 \times p_{max}$, by default), where the gene families with a weight greater than θ are considered as present. The confidence score C_{ij} for gene family i in sample j was assigned to every gene family:

$$C_{ij} = \begin{cases} 1, & p'_{ij} \geq \theta \\ \frac{\theta - p'_{ij}}{\theta}, & p'_{ij} < \theta \end{cases}.$$

The confidence scores were used to rank gene presence predictions for generating the Precision-Recall curves in the benchmarking experiments.

To select the proper number of strains (*i.e.*, the rank of NMF), we parsimoniously select the least number of strains from a range of 1 to 12 (by default) satisfying the following criteria: 1) The mean relative abundance across all the samples of any strain should be greater than τ_2 ($\tau_2 = 0.1$ by default), 2) the number of gene families of all strains should be greater than $\tau_3 \times g_{min}$ ($\tau_3 = 0.5$ by default), g_{min} is the minimum number of gene families found in all reference genomes, and 3) the gene family profiles between a pair of strains should have Jaccard distance larger than τ_1 ($\tau_1 = 0.1$ by default). The

program also provides an option to accept user-specified number of strains set. In this study, we did not set the number of strains a priori in benchmarking and applications of StrainPanDA.

Benchmarking StrainPanDA with synthetic data

Synthetic data of *E. coli* strains. We generated four types of simulated sequencing reads: 1) Error-Free (ErrFree): pick random fragments from the reference genome of *E. coli* by ART simulator⁷³ (version 2016.06.05; parameter: -ef -ss HS25 -l 150 -m 270 -s 27). 2) ART with sequencing errors (ARTerr): use ART to add sequencing errors on top of ErrFree reads (parameter: -ss HS25 -l 150 -m 270 -s 27). 3) pure WGS (pWGS): randomly draw reads from the whole genome sequencing data of selected strains by seq-tk (<https://github.com/lh3/seqtk>, version 1.3; default parameter). 4) pWGS data mixed with a real background metagenomic dataset (WGSBG): Three different metagenomic datasets (see **Supplementary Table 4**. BG1: IBD, BG2: FMT, BG3: MI, as shown in **Figure 2**) were used to mix with the pWGS data of *E. coli* at different ratios (1, 5, 10, 25 and 100-fold). Metagenomic samples were analyzed by Kraken2 (version 2.1.1; database: minikraken2_v2_8GB_201904) to ensure minimal abundance of *E. coli*. Strains of *E. coli* with pairwise genome wide ANI between 95%-99% were selected to represent different subspecies (**Supplementary Table 3**). In each synthetic data set of mixed strains, 20 combinations of strain composition were generated by Dirichlet distribution (**Supplementary Table 2**). All synthetic datasets were generated by the SimStr pipeline (<https://github.com/xbiome/StrainPanDA/tree/main/SimStr>). For each strain, its genome size was considered 1× sequencing depth and used to calculate the number of reads to generate. The minimum relative abundance (*i.e.*, frequency) of a strain was set as 5% and as one unit. For example, for *E. coli* synthetic data of 1× sequencing depth that we refer to in this study, the data size of the strain with 5% frequency was ~4.5 megabases (MB), while the total depth of each sample in this dataset was always 20x, and ~90MB in size (1x sequencing depth as a unit and 20 units in total). To evaluate the effect of sample size, 400 synthetic mixtures of 4 strains were generated by Dirichlet distribution. The synthetic data were separated into 10 runs (40 samples each) and further downsampled to 20, 15, 10 and 5 samples in each run.

Synthetic data of gut bacterial species. Synthetic data of 6 species, including *B. longum*, *C. difficile*, *E. coli*, *E. faecalis*, *F. prausnitzii* and *P. copri*, were generated separately (Sync6, pWGS at 5× sequencing depth) (**Supplementary Table 2**). For each species, the relative abundances of 4 strains (5%, 10%, 25% and 60%) were permuted to generate 24 samples in total (**Supplementary Table 2**). All 6 species were in the pre-built databases of StrainPanDA and PStrain. Only *B. longum*, *E. coli* and *E. faecalis* were in the pre-built database of StrainEst, so the other species were excluded in the comparison to StrainEst (**Fig. 2C**).

Evaluation of predicted strain composition. StrainEst and PStrain were run with their pre-built database and default parameters (StrainEst: <ftp://ftp.fmach.it/metagenomics/strainest/ref/>; PStrain: <https://github.com/wshuai294/PStrain>). The strain compositional profiles predicted by different methods were evaluated and compared by SimStr. For the predicted strain composition shown in Figure 2A (stacked bar plots), strains with relative abundance below 0.01 were filtered and the remaining strains were sorted by their relative abundance (rescaled to 1) in decreasing order. After sorting, the predicted strains in the lower tail exceeding the number of simulated strains were grouped into “Extras”.

Two commonly used metrics were used to evaluate the performance of predicted strain composition of different methods:

1) Jensen-Shannon divergence⁷⁴ (JSD): JSD between the predicted strain composition and actual strain composition is calculated by the distance function in phyloseq⁷⁵ (R package) on the sorted relative abundance (in decreasing order). If the number of predicted strains is different from the actual number of strains, zeros were appended to the vector with lower dimension. The JSD is symmetric and is in the interval of [0,1]. It reflects the dissimilarity in compositional profiles of strains, *i.e.*, JSD = 0 represents an exact prediction.

2) Matthews correlation coefficient⁷⁶ (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. The MCC ranges from -1 to 1, where 1 represents an exact prediction, 0 represents a random prediction and -1 represents total disagreement.

Due to the lack of strain annotations from PStrain, we only computed MCC for strain composition predicted by StrainPanDA and StrainEst. For StrainEst, the predicted strains were directly annotated by reference genomes. For StrainPanDA, based on the predicted gene family profile, the Jaccard distance ($JD(A, B) = 1 - |A \cap B| \div |A \cup B|$) between the predicted strain and all reference genomes were calculated. The reference genome with the smallest Jaccard distance to the predicted strain was used for annotation. If a strain in the synthetic mixture is included in the pre-built database of reference genomes, ID of the annotated reference genome is directly compared to the actual strain to determine if the prediction strain is true positive. If a strain in the synthetic mixture is not included in the pre-built database of reference genomes, we used the phylogenetic tree to decide whether a predicted strain is true positive (**Supplementary Table 4**). Briefly, we generated a phylogenetic tree by parsnp⁷⁷ (version 1.5.1, default parameter) including genomes of the strains used in synthetic mixtures and all the reference genomes. If the annotated reference genome of a predicted strain is within the cutoff of phylogenetic distance (cutoff = 0.05 for *E. coli*, corresponding to ANI ~99%) from an actual strain, it is considered as true positive.

Evaluation of predicted gene family profiles. For each microbial strain evaluated in benchmarking datasets, ErrFree reads at 5× sequencing depth were generated by ART simulator⁷³ (Sync-Single dataset) based on its reference genome downloaded from NCBI. The Sync-Single dataset had three replicates for each strain and was used to generate the actual gene family profile of each strain by PanPhlAn (default parameters). The gene families found in two or more replicates were considered as “present”. The actual gene family profile (Ground Truth) of each strain was compared to the predicted gene family profile (**Fig. 2D**). The Jaccard distances between the predicted gene family profiles of microbial strains and their ground truth profiles (or the gene family profile of a randomly sampled reference genome) were computed (**Supplementary Fig. 5**). The Precision-Recall curve of gene family profiles for each strain was generated by R package PRROC⁷⁸ using the confidence scores to rank the gene families

predicted (**Supplementary Fig. 6**). For random guesses, 1000 random gene family profiles were generated by sampling N gene families from the pangenome as “present”, where N is the average number of gene families present in reference genomes. To demonstrate the ability of StrainPanDA to identify strain-specific genes, the pathogenic *E. coli* strain O104 (GCF_002983645) was introduced in a synthetic dataset of 4 strains (Sync O104 dataset, pWGS, 5× sequencing depth). The outbreak related genes curated from Segata *et al.*³⁵ were used to evaluate StrainPanDA’s gene content prediction (**Supplementary Table 5**). We also compared the predicted gene family profiles from StrainPanDA to the prediction from PanPhlAn (**Supplementary Fig. 8**). To simplify the output, only samples with the highest abundance of one strain were used for PanPhlAn prediction.

Applications of StrainPanDA in metagenomic data

Case study: mother-infant gut metagenomes. All available samples of ERP005989 were downloaded from EBI (8 samples failed, **Supplementary Table 6**). Based on diet history, infants without diet history were filtered. The rest of 84 infants were split into 3 different groups (B_F_F: discontinued breastfeeding at 4-month; B_B_F: discontinued breastfeeding at 12-month; B_B_B: continued breastfeeding; the F_B_M sample was excluded) (**Supplementary Table 6**). Samples without enough coverage on *B. longum* gene families were filtered by StrainPanDA at the preprocessing step and excluded from the downstream analysis. In the “continued breastfeeding” group, infants that kept breastfeeding between successive time points (*i.e.*, between newborn and 4-month, or between 4-month and 12-month) were included. In the “discontinued breastfeeding” group, infants which stopped breastfeeding by 4-month or 12-month were included. For functional interpretation of the subspecies, we grouped the gene families annotated to the same KEGG ortholog or CAZy family. To further analyze the key functions related to breastfeeding, we curated a set of KEGG orthologs from related references^{51,79}. All the KEGG orthologs were further grouped into subclasses and classes based on the literature^{51,79} (**Supplementary Table 7**). The gene family coverage was calculated as the fraction of detected genes belonging to the subclass (**Fig. 3C**).

Case study: FMT donor-recipient metagenomes. Raw sequencing reads were downloaded from ENA (Accession: PRJNA625520 for the study on Crohn's disease³⁷, PRJEB23524 for the study on *C. difficile* infection study²⁹). For the Crohn's disease dataset, species relative abundances were estimated using Kraken2¹⁸ (version 2.0.8-beta) with the miniKraken database (v2_8GB_201904_UPDATE). To identify species associated with remission or relapse, Wilcoxon rank sum test was conducted to select differentially abundant species using the mean relative abundances across different timepoints for each subject (samples collected before FMT or after relapse were discarded). The relative abundances of subspecies were predicted by StrainPanDA and normalized by the centered-log-ratio transformation for calculating Spearman correlation with the species abundances. For functional annotation of the subspecies, virulence factors and CAZy annotations were taken from the species-specific databases constructed as described above.

Functional annotation of gene families. To annotate the gene families by KEGG, gene family representative sequences were mapped against KEGG orthologs (release 2020-07-20) using usearch⁶⁹ (v11.0.667; '-ublast'). Alignments with identity > 50% and query coverage > 50% are kept. To annotate the gene families by CAZy, seqkit⁸⁰ (0.15.0) was used to translate the gene family centroids into six open reading frames. The translated amino acid was used as the input of run_dbcan⁸¹ (2.0.11), which used DIAMOND⁸² (2.0.8), HMMer⁸³ (3.3.2) and Hotpep⁸⁴ (2.0.8) with default parameters to predict the CAZy annotation. The CAZy annotations was selected only if it was predicted by at least two programs. If a gene family was assigned by more than one CAZy annotations, only the first annotation was used. To annotate virulence factors, gene family centroids were mapped against the Virulence Factor Database (VFDB, Apr 9 2021) by DIAMOND⁸² (2.0.8) (blastp, query coverage >50% and identity > 50%).

Declarations

Acknowledgements

We would like to thank Haokui Zhou, Ramnik Xavier and members of Lei Dai lab for constructive comments on the manuscript. We would like to thank Jinhui Tang and Yong Liang for their assistance in bioinformatics analysis, Xi Wang and Dongdong Xu for their support in maintaining the computing environment, Reese Hitchings for proofreading the manuscript.

Software availability

The source code of StrainPanDA is available on Github (<https://github.com/xbiome/StrainPanDA>).

Authors' contributions

H.H., Y.T. and L.D. conceived and supervised the study. H.H., Y-X.T. and C.L. developed the algorithm and performed the analysis on simulated and real metagenomic samples. L.D., H.H., Y-X.T. and C.L. wrote the manuscript with inputs from all co-authors.

Funding

This research was supported by National Key R&D Program of China (No.2019YFA0906700) and National Natural Science Foundation of China (No.31971513, No. 32061143023) to L.D.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The results described in this manuscript support pending patent application CN202011146154.3A. Y.T. is co-founder and shareholder with a personal financial interest in Xbiome. H.H. and Y.K. are employees of Xbiome. L.D. received research grant support from Xbiome and serves as an unpaid consultant to the company. The other authors do not have competing interests.

References

1. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nature Medicine* **25**, 1442–1452 (2019).
2. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, 960 (2017).
3. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology* 1–16 (2020) doi:10.1038/s41579-020-0368-1.
4. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life. *Nat Microbiol* **4**, 470–479 (2019).
5. Ma, B. *et al.* A comprehensive non-redundant gene catalog reveals extensive within-community intraspecies diversity in the human vagina. *Nature Communications* **11**, 940 (2020).
6. Scheuerl, T. *et al.* Bacterial adaptation is constrained in complex communities. *Nature Communications* **11**, 754 (2020).
7. De Filippis, F. *et al.* Distinct Genetic and Functional Traits of Human Intestinal *Prevotella* copri Strains Are Associated with Different Habitual Diets. *Cell Host Microbe* **25**, 444–453.e3 (2019).
8. Carrow, H. C., Batachari, L. E. & Chu, H. Strain diversity in the microbiome: Lessons from *Bacteroides fragilis*. *PLOS Pathogens* **16**, e1009056 (2020).
9. Bisanz, J. E. *et al.* A Genomic Toolkit for the Mechanistic Dissection of Intractable Human Gut Bacteria. *Cell Host Microbe* **27**, 1001–1013.e9 (2020).
10. Ahern, P. P., Faith, J. J. & Gordon, J. I. Mining the human gut microbiota for effector strains that shape the immune system. *Immunity* **40**, 815–823 (2014).
11. Rekdal, V. M., Bess, E. N., Bisanz, J. E., Turnbaugh, P. J. & Balskus, E. P. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science* **364**, (2019).
12. Wilson, M. R. *et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, (2019).
13. Zhang, C. & Zhao, L. Strain-level dissection of the contribution of the gut microbiome to human metabolic disease. *Genome Med* **8**, 41 (2016).
14. Zhai, R. *et al.* Strain-Specific Anti-inflammatory Properties of Two *Akkermansia muciniphila* Strains on Chronic Colitis in Mice. *Front. Cell. Infect. Microbiol.* **0**, (2019).
15. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902–903 (2015).
16. Sun, Z. *et al.* Challenges in benchmarking metagenomic profilers. *Nat Methods* **18**, 618–626 (2021).
17. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications* **10**, 1014 (2019).

18. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).
19. Zou, Y. *et al.* 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* **37**, 179–185 (2019).
20. Liu, C. *et al.* Enlightening the taxonomy darkness of human gut microbiomes with a cultured biobank. *Microbiome* **9**, 119 (2021).
21. Segata, N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **3**, (2018).
22. Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M. & Abeel, T. Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. *Front. Microbiol.* **11**, (2020).
23. Yassour, M. *et al.* Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. *Cell Host & Microbe* **24**, 146-154.e4 (2018).
24. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, e00164-16 (2017).
25. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
26. Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **28**, 561–568 (2018).
27. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
28. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology* **33**, 1045–1052 (2015).
29. Smillie, C. S. *et al.* Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. *Cell Host & Microbe* **23**, 229-240.e5 (2018).
30. Albanese, D. & Donati, C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nature Communications* **8**, 2260 (2017).
31. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).
32. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine* **8**, 343ra81-343ra81 (2016).
33. Wang, S., Jiang, Y. & Li, S. PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data. *Bioinformatics* **36**, 5499–5506 (2020).
34. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).
35. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods* **13**, 435–438 (2016).

36. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
37. Kong, L. *et al.* Linking Strain Engraftment in Fecal Microbiota Transplantation With Maintenance of Remission in Crohn’s Disease. *Gastroenterology* **159**, 2193–2202.e5 (2020).
38. Carr, R., Shen-Orr, S. S. & Borenstein, E. Reconstructing the Genomic Content of Microbiome Taxa through Shotgun Metagenomic Deconvolution. *PLOS Computational Biology* **9**, e1003292 (2013).
39. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
40. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
41. Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* **23**, 1495–1502 (2007).
42. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
43. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490–D495 (2014).
44. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* **47**, D687–D692 (2019).
45. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* **6**, e22751 (2011).
46. Sela, D. A. *et al.* The genome sequence of *Bifidobacterium longum* subsp. *infantis* reveals adaptations for milk utilization within the infant microbiome. *PNAS* **105**, 18964–18969 (2008).
47. Sela, D. A. Bifidobacterial utilization of human milk oligosaccharides. *Int J Food Microbiol* **149**, 58–64 (2011).
48. Katayama, T. Host-derived glycans serve as selected nutrients for the gut microbe: human milk oligosaccharides and bifidobacteria. *Biosci Biotechnol Biochem* **80**, 621–632 (2016).
49. Duranti, S. *et al.* Maternal inheritance of bifidobacterial communities and bifidophages in infants through vertical transmission. *Microbiome* **5**, 66 (2017).
50. Asakuma, S. *et al.* Physiology of consumption of human milk oligosaccharides by infant gut-associated bifidobacteria. *J Biol Chem* **286**, 34583–34592 (2011).
51. Schimmel, P., Kleinjans, L., Bongers, R. S., Knol, J. & Belzer, C. Breast milk urea as a nitrogen source for urease positive *Bifidobacterium infantis*. *FEMS Microbiology Ecology* **97**, (2021).
52. Kujawska, M. *et al.* Succession of *Bifidobacterium longum* Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations. *iScience* **23**, 101368 (2020).
53. Garrido, D. *et al.* A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in *Bifidobacterium longum* subsp. *longum* SC596. *Scientific Reports* **6**, 35045 (2016).

54. Wang, S. *et al.* Metagenomic analysis of mother-infant gut microbiome reveals global distinct and shared microbial signatures. *Gut Microbes* **13**, 1911571 (2021).
55. Frese, S. A. *et al.* Persistence of Supplemented *Bifidobacterium longum* subsp. *infantis* EVC001 in Breastfed Infants. *mSphere* **2**, (2017).
56. Sela, D. A. & Mills, D. A. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends Microbiol* **18**, 298–307 (2010).
57. Smits, L. P., Bouter, K. E. C., de Vos, W. M., Borody, T. J. & Nieuwdorp, M. Therapeutic Potential of Fecal Microbiota Transplantation. *Gastroenterology* **145**, 946–953 (2013).
58. Ooijselaar, R. E., Terveer, E. M., Verspaget, H. W., Kuijper, E. J. & Keller, J. J. Clinical Application and Potential of Fecal Microbiota Transplantation. *Annual Review of Medicine* **70**, 335–351 (2019).
59. Peraro, M. D. & van der Goot, F. G. Pore-forming toxins: ancient, but never really out of fashion. *Nature Reviews Microbiology* **14**, 77–92 (2016).
60. Tan, J. M. J. *et al.* An ATG16L1-dependent pathway promotes plasma membrane repair and limits *Listeria monocytogenes* cell-to-cell spread. *Nature Microbiology* **3**, 1472–1485 (2018).
61. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).
62. Quince, C. *et al.* STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology* **22**, 214 (2021).
63. Vicedomini, R., Quince, C., Darling, A. E. & Chikhi, R. Strawberry: automated strain separation in low-complexity metagenomes using long reads. *Nat Commun* **12**, 4485 (2021).
64. Sheth, R. U. *et al.* Spatial metagenomic characterization of microbial biogeography in the gut. *Nat Biotechnol* **37**, 877–883 (2019).
65. Chng, K. R. *et al.* Cartography of opportunistic pathogens and antibiotic resistance genes in a tertiary hospital environment. *Nat Med* **26**, 941–951 (2020).
66. Palleja, A. *et al.* Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat Microbiol* **3**, 1255–1265 (2018).
67. Chng, K. R. *et al.* Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nat Ecol Evol* **4**, 1256–1267 (2020).
68. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).
69. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
71. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

72. Pan, W. & Doshi-Velez, F. A Characterization of the Non-Uniqueness of Nonnegative Matrix Factorizations. *arXiv:1604.00653 [cs, stat]* (2016).
73. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
74. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* **37**, 145–151 (1991).
75. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217 (2013).
76. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451 (1975).
77. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology* **15**, 524 (2014).
78. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
79. Odamaki, T. *et al.* Comparative Genomics Revealed Genetic Diversity and Species/Strain-Level Differences in Carbohydrate Metabolism of Three Probiotic Bifidobacterial Species. *International Journal of Genomics* **2015**, e567809 (2015).
80. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**, e0163962 (2016).
81. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95–W101 (2018).
82. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).
83. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011).
84. Busk, P. K., Pilgaard, B., Lezyk, M. J., Meyer, A. S. & Lange, L. Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **18**, 214 (2017).