


RESEARCH ARTICLE

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

Han Hu^{1,2} | Yuxiang Tan¹ | Chenhao Li³ | Junyu Chen¹ | Yan Kou² |
Zhenjiang Zech Xu⁴ | Yang-Yu Liu⁵ | Yan Tan² | Lei Dai¹ 

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

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

³Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Richard B. Simches Research Center, Boston, Massachusetts, USA

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

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

Correspondence

Lei Dai, CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences 1068 Xueyuan Avenue, Shenzhen University Town, 518055 Shenzhen, China.
Email: lei.dai@siat.ac.cn

Abstract

Microbial strains of variable functional capacities coexist in microbiomes. Current bioinformatics methods of strain analysis cannot provide the direct linkage between strain composition and their gene contents from metagenomic data. Here we present *Strain-level Pangenome Decomposition Analysis* (StrainPanDA), a novel method that uses the pangenome coverage profile of multiple metagenomic samples to simultaneously reconstruct the composition and gene content variation of coexisting strains in microbial communities. We systematically validate the accuracy and robustness of StrainPanDA using synthetic data sets. 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 and pathogenicity). Finally, StrainPanDA has minimal requirements for computing resources and can be scaled to process multiple species in a community in parallel. In short, StrainPanDA can be applied to metagenomic data sets to detect the association between molecular functions and microbial/host phenotypes to formulate testable hypotheses and gain novel biological insights at the strain or subspecies level.

KEYWORDS

gene content profile, metagenomics, microbiome, pangenome, strain analysis

Highlights

- *Strain-level Pangenome Decomposition Analysis* (StrainPanDA) uses the pangenome coverage profile of multiple metagenomic samples to

Han Hu, Yuxiang Tan and Chenhao Li contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *iMeta* published by John Wiley & Sons Australia, Ltd on behalf of *iMeta* Science.

Yan Tan, Xbiome, Room 907, 9th Floor, Scientific Research Bldg, Tsinghua High-Tech Park, 518000 Shenzhen, China.
Email: yant@xbiome.com

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 31971513, 32061143023; National Key R&D Program of China, Grant/Award Number: 2019YFA0906700

simultaneously reconstruct the composition and gene content variation of coexisting strains in microbial communities.

- StrainPanDA allows accurate and robust inference of strain composition and gene content profiles on synthetic data sets.
- Linked reconstruction of strain composition and gene content profiles provided by StrainPanDA furthers our understanding of the relationship between microbial adaptation and strain-specific functions (e.g., nutrient utilization and pathogenicity).

INTRODUCTION

There is mounting evidence that multiple within-species variants coexist in microbiomes [1, 2]. Coexisting 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 the accessory genome can lead to substantial phenotypic differences (e.g., nutrient utilization, pathogenicity, and 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 studying 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 the population genetics of microbial strains [25], 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 profiles [4], 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 HGT [4]. Alternatively, the current pangenome-based method can infer the gene content of the dominant strain in a metagenomic sample [35] but fails to provide the abundance and gene contents of coexisting strains within the sample. Establishing the linkage of composition and gene contents of coexisting 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 known as *Strain-level Pangenome Decomposition Analysis* (StrainPanDA) to simultaneously reconstruct the composition and gene contents of coexisting strains using the pangenome coverage profile from metagenomic data (Figure 1). We validated the performance of StrainPanDA with a comprehensive collection of synthetic data sets 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 pairs [36] and patients treated with fecal microbiota transplantation (FMT) [29, 37]. We found that StrainPanDA was able to identify the 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.

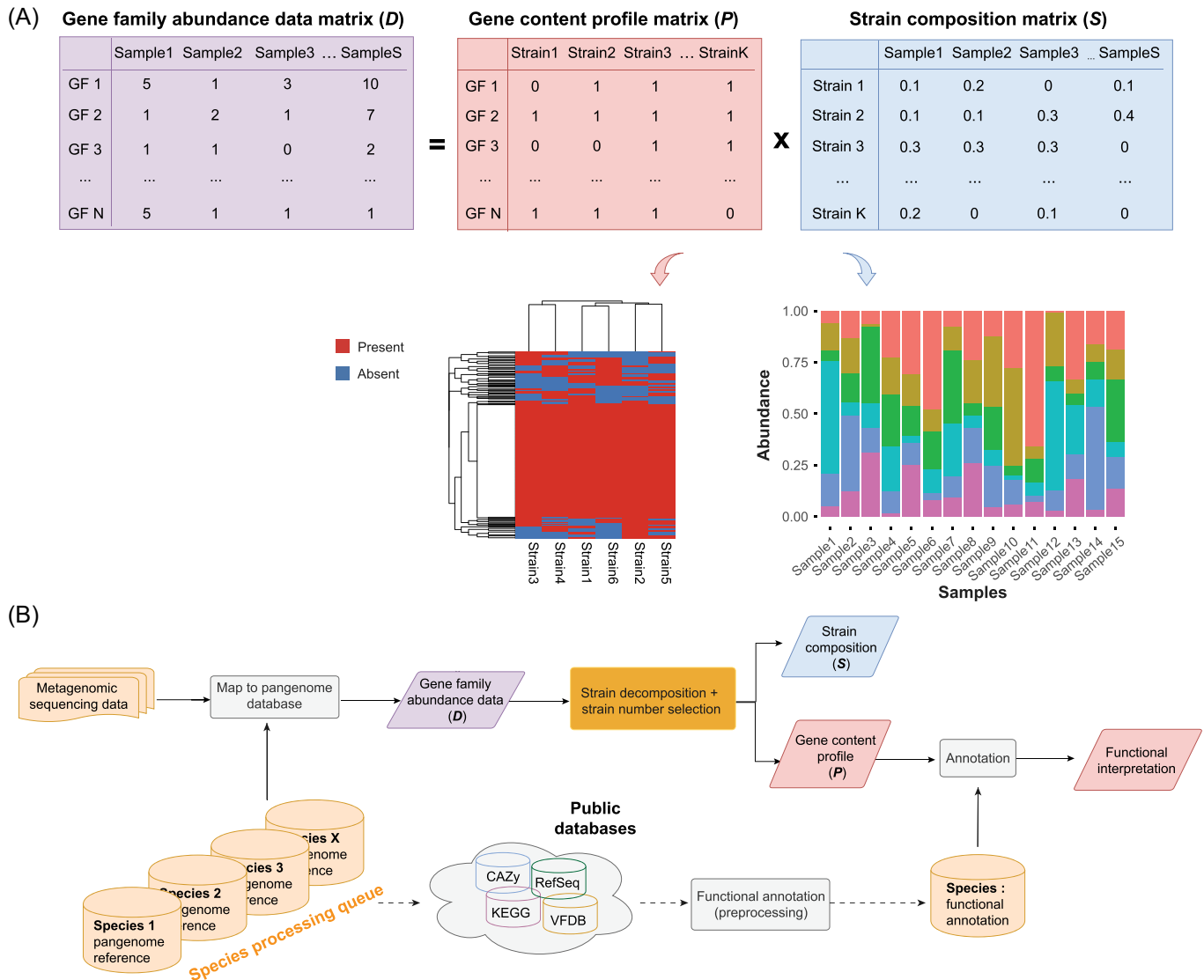


FIGURE 1 Illustration of the StrainPanDA workflow. (A) The gene family abundance data matrix D is decomposed into the product of two matrices P and S via nonnegative matrix factorization (Methods section). The gene content profile matrix P is a binary matrix that denotes the presence/absence of gene families in each strain. The strain composition matrix S represents the relative abundance of coexisting strains in each sample. In the illustrated example, the size of metagenomic samples $S = 15$ and the number of strains $K = 6$. (B) The workflow of StrainPanDA analysis is performed in a species-by-species manner, including mapping metagenomic reads to the pangenome database, strain decomposition, and functional annotation of gene family profiles. CAZy, Carbohydrate-Active enZymes; GF, gene family; KEGG, Kyoto Encyclopedia of Genes and Genomes; StrainPanDA, *Strain-level Pangenome Decomposition Analysis*; VFDB, Virulence Factor Database.

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 coexisting 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 profile [38]. Building on this intuition, the main algorithm of StrainPanDA aims to decompose the gene family abundance data matrix D into the product of two matrices, the gene content profile matrix P , and the strain composition matrix S (Figure 1A, see Methods section for details). Here the pangenome coverage profile from metagenomic data is represented by matrix D , where D_{ij} is the normalized count of gene family i in metagenomic sample j . The gene contents of coexisting

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 coexisting 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 nonnegative 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 coexisting strains.

The StrainPanDA software provides a fully automated workflow of strain analysis (Figure 1B), which imports raw sequencing data from multiple metagenomic samples, and performs reads mapping, strain decomposition, and downstream annotations (see Methods section for details). To assist the interpretation of gene content variation in strains, StrainPanDA incorporates functional annotation from several databases, including but not limited to Kyoto Encyclopedia of Genes and Genomes (KEGG) [42], Carbohydrate-Active enZymes (CAZy) [43], and Virulence Factor Database (VFDB) [44].

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 section). For synthetic mixtures of *Escherichia coli* strains (ranging from 2 to 8 strains, see Methods section), the strain composition predicted by StrainPanDA was overall close to the actual composition (Ground Truth; Figure 2A), and its performance was better at a lower number of coexisting strains (2 and 4 strains). 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]. At a lower number of coexisting strains (2 and 4 strains), the predicted strain composition by StrainPanDA was better than the state-of-the-art SNV-based methods, including StrainEst [30] and PStrain [33] (the latter was modified based on ConStrains [28]). While the results of StrainEst tended to include false positives at a lower number of coexisting strains, its performance was better at 6 and 8 strains. Furthermore, we generated synthetic mixtures of *E. coli* strains with varying levels of sequencing errors (Supporting Information Figure S2), sequencing depths (Supporting Information Figure S3), and different background noises (mixed with different

metagenomic data sets, Supporting Information Table S1 and Figure S4). 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*; Supporting Information Table S2, see Methods section). In comparison to other methods, StrainPanDA made the most accurate prediction of strain composition across all species when strain number was 4 (with JSD as 0.021 ± 0.006 ; Figure 2C).

While current SNV-based methods can reconstruct the composition of coexisting 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 (Figure 2D, Supporting Information Figure S5; precision = 0.91–0.96, recall = 0.87–0.96, for the four strains in a synthetic mixture). In particular, we note that StrainPanDA is able to infer the gene family profile of strains not included in the prebuilt 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 (Supporting Information Figure S6).

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 (Figure 2E). The predicted gene family profiles were robust to sequencing errors, sequencing depths, and the background of real metagenomic data (Supporting Information Table S4). Moreover, to demonstrate the ability of StrainPanDA to identify strain-specific genes, the pathogenic *E. coli* outbreak strain O104 [45] was introduced in a synthetic mixture with other *E. coli* strains (Supporting Information Table S5). All outbreak-related gene families were successfully recovered by StrainPanDA (Supporting Information Figure S7). Finally, the performance of StrainPanDA and Pangenome-based Phylogenomic Analysis (PanPhlAn)/PanPhlAn3 in inferring the gene content profiles were comparable (Supporting Information Figure S8). We note that PanPhlAn and PanPhlAn3 can only report the gene content profile of the “dominant strain” in a particular metagenomic sample (i.e., the strain with the highest relative abundance); in contrast,

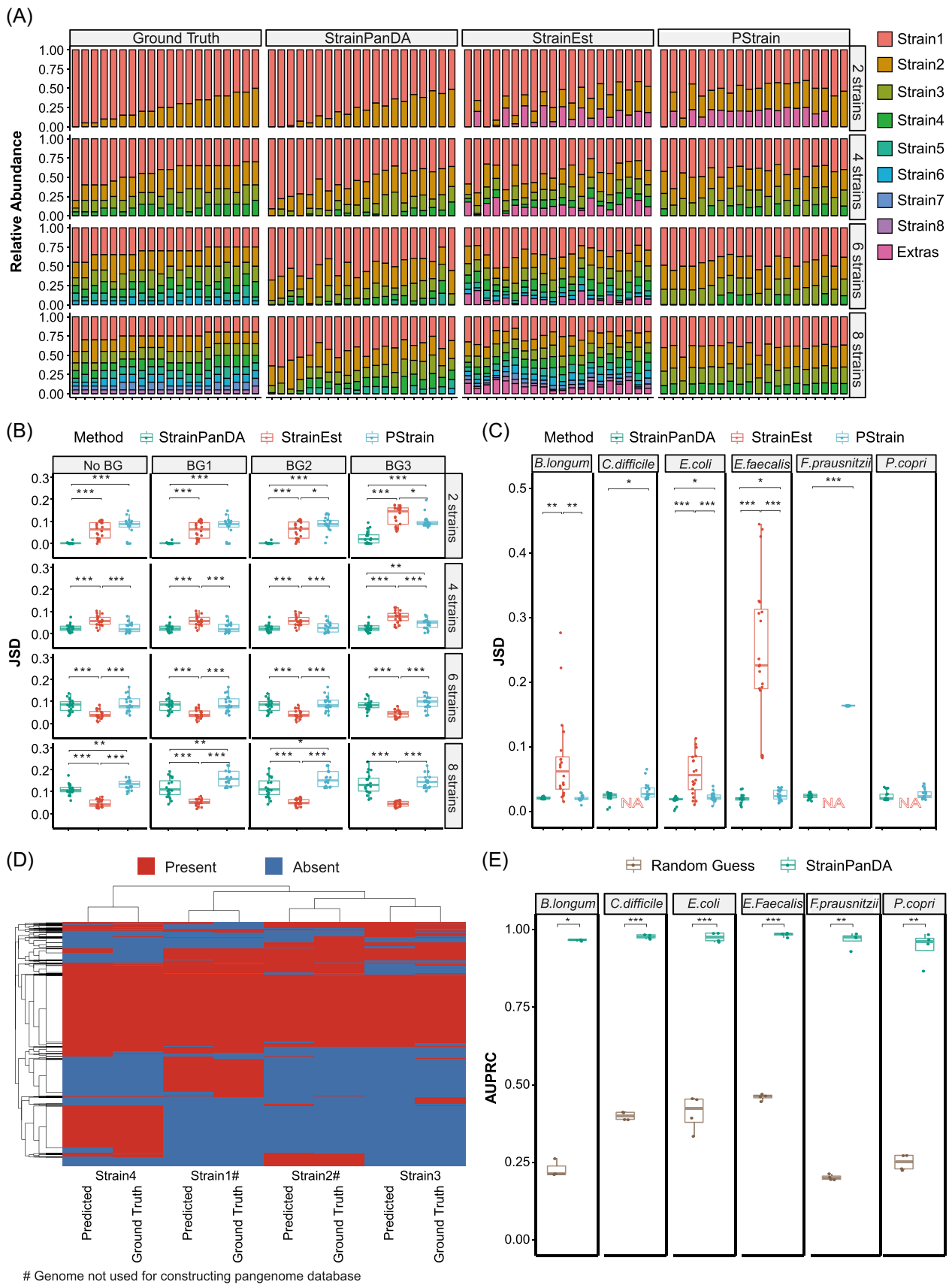


FIGURE 2 (See caption on next page)

StrainPanDA can identify the gene content profiles of all coexisting strains.

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

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

The direct inference of both the population structure and gene content variation at the strain level is crucial to understanding the ecology of microbial communities. Here we apply StrainPanDA to study the adaptation of coexisting bacterial subspecies in the infant gut microbiomes. We analyzed a previously published data set that includes gut metagenomic samples from ~100 mother–infant pairs (infants were sampled at three time points: newborn, 4 months, and 12 months) [36] (Supporting Information Table S6). At the species level, the authors found that the composition of the 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 microbiome [36].

We focused on the infraspecific analysis of *B. 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 (Supporting Information Figure S9). Interestingly, we discovered a clear pattern of succession in the subspecies composition of *B. longum* over time, that is, a shift in the dominant subspecies (Figure 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). On the basis of the diet history provided in the original study, we further grouped infant samples into two different categories: “discontinued breastfeeding” and “continued breastfeeding” between successive time points (Methods section). It was evident that the relative abundance of *B. longum* subspecies 1 was enriched in infants that continued breastfeeding (Figure 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 within-species competition in nutrient utilization functions (Figure 3C). On the basis of functional annotations of KEGG [42] and CAZy [43], 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, Figure 3C) that are key enzymes related to human milk oligosaccharide (HMO), including galactosidase, α -L-fucosidase, sialidase, and their corresponding CAZy groups (GH33, GH29, and GH95; Supporting Information Figures S12 and S13 and Table S7). 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 urea [51] from breast milk could confer a competitive advantage under breastfeeding, consistent with

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 *Escherichia coli* strains (pWGS data set, 1× sequencing depth, see Methods section). 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 data sets as background (WGSBG data set, 100-fold background, see Methods section). 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 reference and predicted gene family profiles of *E. coli* strains (the synthetic data used are 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 coexisting 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 section). p values from paired t test: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$. *Bifidobacterium longum*, *Clostridium difficile*, *Enterococcus faecalis*, *Faecalibacterium prausnitzii*, and *Prevotella copri*. pWGS, pure whole genome sequencing; StrainPanDA, Strain-level Pangenome Decomposition Analysis; WGSBG, whole genome sequencing background.

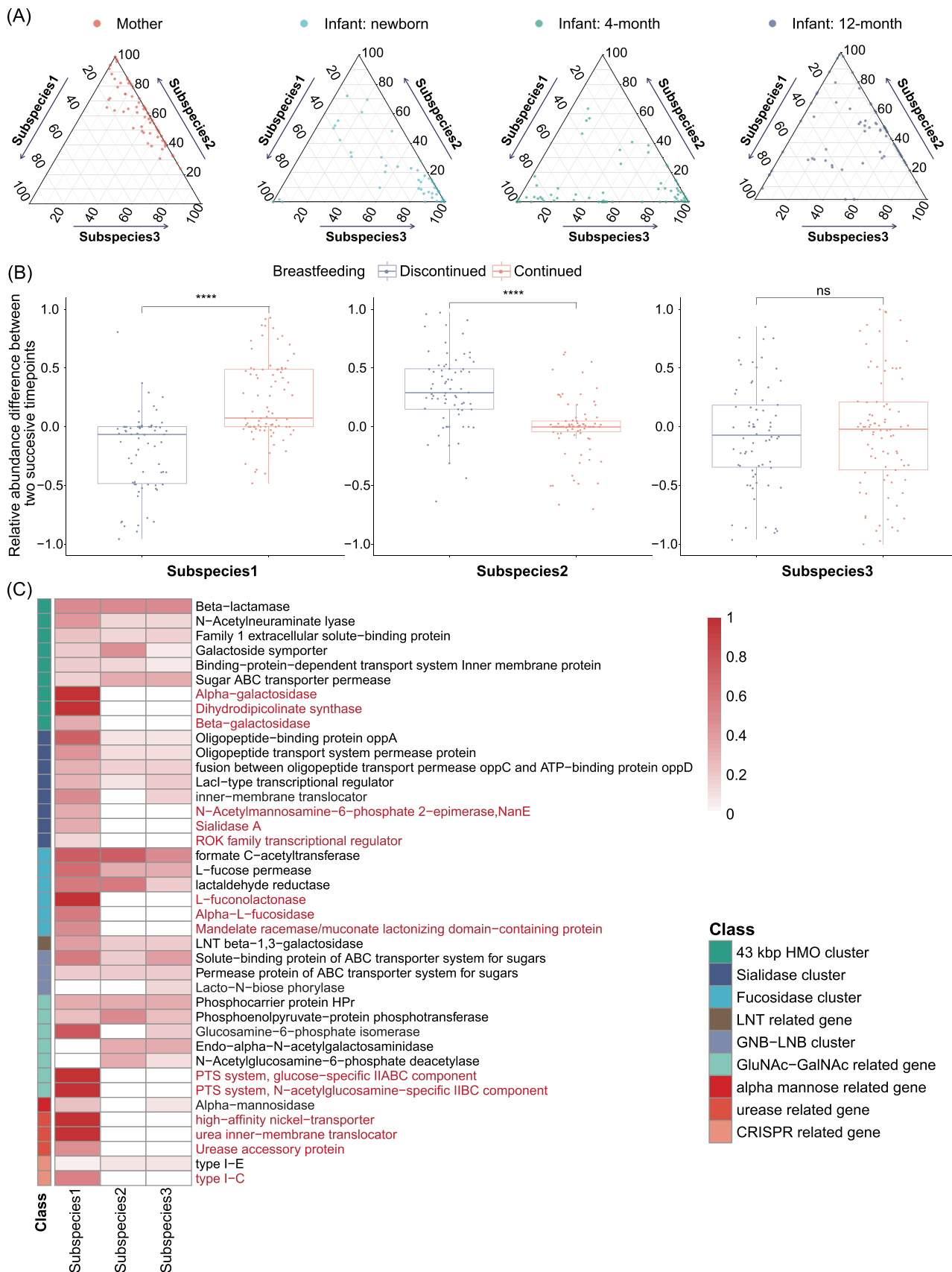


FIGURE 3 (See caption on next page)

our observations of its transient dominance at 4 months in infant gut metagenomes.

Our finding is consistent with previous reports [52] 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* (Supporting Information Figure S14), 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 ecology at the subspecies level.

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

FMT introduces bacterial strains from healthy donors into recipients and has a profound impact 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 disease [37], including 17 patients (eight in the FMT group and nine 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 (Supporting Information Table S8). Hierarchical clustering of the predicted subspecies compositional profiles revealed strong individual signatures, which remained stable throughout 24 weeks (Figure 4A). The pairwise distance in subspecies composition profiles among samples of the same individual (sampled at multiple time points, that is, the “intrasubject” group) was significantly

lower than the pairwise distance among samples of different individuals (i.e., the “intersubject” group; Figure 4B, $p < 10^{-15}$), 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 ($p < 0.001$), indicating the engraftment of donor strains and coexistence of donor and recipient strains [37]. 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; Figure 4B). Similarly, we applied StrainPanDA to analyze an independent FMT data set of metagenomic samples from patients with *C. difficile* infection [29]. We observed a clear pattern of subspecies engraftment in post-FMT gut metagenomes, consistent with SNV-based strain analysis in the original study [29] (Supporting Information Figure S15). 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 the 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 showed that the enrichment of Bacteroidetes species in patients relapsed after FMT [37]. We focused our analysis on *Bacteroides ovatus*, which was found to be enriched in relapsed individuals (false discovery rate [FDR]-adjusted $p = 0.2$; Figure 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 $p < 10^{-6}$, Supporting Information Figure S16). We found substantial gene content variation among different *B. ovatus* subspecies (Figure 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 competitive

FIGURE 3 Succession of *Bifidobacterium longum* subspecies in infant gut microbiome can be attributed to the selection of 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 months, and 12 months). Each dot represents one sample. (B) The shift in the relative abundance of *B. longum* subspecies between successive time points. According to the breastfeeding status at the subsequent time point, infants are divided into two groups (purple, discontinued breastfeeding, $N = 86$; red, continued breastfeeding, $N = 71$). **** $p < 0.00005$; ns, not significant; Student's t test. (C) Gene family profiles of predicted *B. longum* subspecies. Gene families related to the metabolism of host glycans, urease, and CRISPR (KEGG annotations) are selected for display. The color bar on the left indicates the class of gene clusters. Each row is a subclass of gene families and unique gene families of subspecies 1 are marked in red. The color scale in the heatmap indicates the normalized gene family coverage in the specific subclass (i.e., the fraction of detected gene families belonging to the subclass). HMO, human milk oligosaccharide; KEGG, Kyoto Encyclopedia of Genes and Genomes; ABC, ATP-binding cassette transporters; ATP, adenosine triphosphate; CRISPR, clustered regularly interspaced short palindromic repeat; Gal, galactose; Gal-NAc, *N*-acetylgalactosamine; Glu-NAc, *N*-acetylglucosamine; GNB, galacto-*N*-biose; LNB, lacto-*N*-biose; LNT, lacto-*N*-tetraose; PTS, phosphotransferase system; ROK, repressor, open reading frame, kinase.

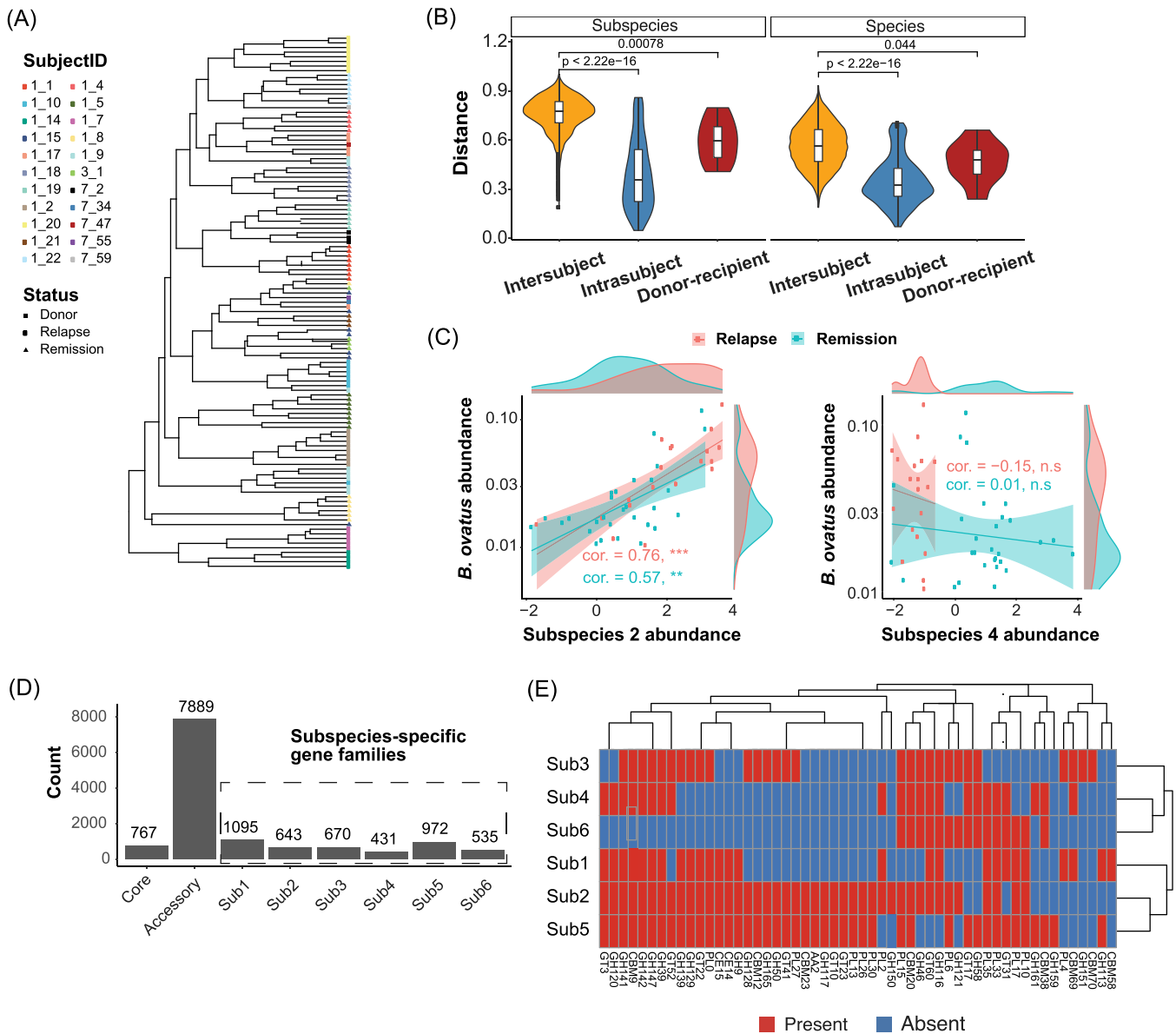


FIGURE 4 Analysis of post-FMT gut metagenomes reveals individualized subspecies profiles and the association between subspecies-specific functions and phenotypes. (A) Hierarchical clustering of predicted subspecies compositional profiles of common gut species (Supporting Information Table S8) reveals strong individual signatures. The subject IDs were collected from the original paper [37] 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 show the distribution of the corresponding variables. At the species level, *B. 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. FMT, fecal microbiota transplantation.

advantages (Figure 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 (Figure 4C and Supporting Information Figure S16). In addition, *B. ovatus* subspecies 2 carried several strain-specific

virulence factor genes (e.g., type IV secretion system and cholesterol-dependent cytolysin), which may contribute to the positive association between *B. ovatus* and post-FMT relapse (Supporting Information Figure S16). For example, cholesterol-dependent cytolysin is a pore-forming toxin that can disrupt the host plasma membrane [59], whose

integrity has been linked to inflammatory bowel disease [60]. In addition, we noted that *B. ovatus* subspecies 4 was more abundant in the remission group (Figure 4C, FDR-adjusted $p < 10^{-5}$); 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 (Supporting Information Figure S17).

In summary, we show that the linkage of strain composition and gene contents provided by StrainPanDA can greatly facilitate our understanding of microbial ecology 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, to simultaneously profile the composition of coexisting 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 prebuilt reference genome database. Furthermore, we applied StrainPanDA to metagenomic data sets 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 competitive advantage. We demonstrated that the linkage of strain abundance and gene contents could lead to direct functional interpretations and testable hypotheses.

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 the core genome have less than 70% of genes in common [3], 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 coexisting within-species variants. Our current method relies on the pangenome constructed from a collection of genomes for a given microbial species, thus it does not account for

gene content transfers between species. To account for interspecies gene transfer, one possible solution is to include a pool of “putative mobile elements” to expand the pangenome for each species. 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 on the comparison of StrainPanDA to other reference-based methods, it is worth noting that complementary approaches based on metagenome-assembled genomes (MAGs) can identify novel strains from metagenomic data. For example, DESMAN [61] can provide the predicted draft genome of each strain; other recently developed MAG-based methods include mixtureS [62], STRONG [63], and so forth. In contrast to reference-based methods, MAG-based methods can identify novel species and genes, yet the quality of MAG will greatly affect the results. In addition, 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 (Supporting Information Figures S3 and S4).

StrainPanDA is most suitable for the analysis of multiple metagenomic samples with shared within-species variants, such as longitudinal studies. While the analysis in this study focused on the human gut microbiome, StrainPanDA is broadly applicable to microbiomes in different environments, as long as the pangenomes of the target species are available. The performance of StrainPanDA, including the accuracy of predicted strain composition and gene content profiles, improves with sample size (Supporting Information Figure S18) and sequencing depth (Supporting Information Figure S3). To apply StrainPanDA on a typical metagenomic data set, 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 genetic mosaic or highly similar gene content profiles (i.e., lacking strain-unique features), thus StrainPanDA is most suitable for analysis at the level of subspecies [3]. Finally, in comparison to MAG-based methods, StrainPanDA has minimal requirements for computing resources (Supporting Information Figure S19) and can be scaled to process multiple species in parallel.

CONCLUSION

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 data sets, especially in the context of spatiotemporal characterization of microbiomes [64–67], will help elucidate novel associations between molecular functions and microbial/host phenotypes as well as microbial ecology at the infraspecies level.

METHODS

Generation of pangenome database and mapping of metagenomic data

The 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 National Center for Biotechnology Information (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 were mapped to the pangenome database by PanPhlAn [35] (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 reads per kilobase million [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 the 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 (Figure 1):

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

The gene family abundance data matrix \mathbf{D} is an $N \times S$ nonnegative matrix, where D_{ij} is the normalized count of gene family i in metagenomic sample j . The gene content

profile matrix \mathbf{P} is an $N \times K$ 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 the sample ($S_{ij} \geq 0$ 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, and 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 NMF, considering the nonnegative 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 the 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 Supporting Information Figure S1B). 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 1–12 (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 a 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 read simulator ART [73] (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 whole genome sequencing (pWGS): randomly draw reads from the WGS data of selected strains by seq-tk (<https://github.com/lh3/seqtk>, version 1.3; default parameter); and (4) pWGS data mixed with a real background metagenomic data set (whole genome sequencing background [WGSBG]): Three different metagenomic data sets (see Supporting Information Table S4; 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 a minimal abundance of *E. coli*. Strains of *E. coli* with pairwise genome-wide ANI between 95%–99% were selected to represent different subspecies (Supporting Information Table S3). In each synthetic data set of mixed strains, 20 combinations of strain composition were generated by Dirichlet distribution (Supporting Information Table S2). All synthetic data sets 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 data set was always 20×, and ~90 MB in size (1× sequencing depth as a unit and 20 units in total). To evaluate the effect of sample size, 400 synthetic mixtures of four 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 six species, including *B. longum*, *C. difficile*, *E. coli*, *E. faecalis*, *F. prausnitzii*, and *P. copri*, were generated separately (Sync6, pWGS at 5× sequencing depth; Supporting Information Table S2). For each species, the relative abundances of four strains (5%, 10%, 25%, and 60%) were permuted to generate 24 samples in total (Supporting Information Table S2). All six species were in the prebuilt databases of StrainPanDA and PStrain. Only *B. longum*, *E. coli*, and *E. faecalis* were in the prebuilt database of StrainEst, so the other species were excluded in the comparison to StrainEst (Figure 2C).

Evaluation of predicted strain composition

StrainEst (v1.2.4 through docker) and PStrain (downloaded from GitHub on May 23, 2021) were run with their prebuilt 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. *JSD* [74]: JSD between the predicted strain composition and actual strain composition is calculated by the distance function in phyloseq [75] (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 a lower dimension. The JSD is symmetric and is in the interval of [0, 1]. It reflects the dissimilarity in compositional profiles of strains, that is, JSD = 0 represents an exact prediction.
2. *MCC* [76]:

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

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.

Owing 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 was 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 prebuilt database of reference genomes, the ID of the annotated reference genome is directly compared with the actual strain to determine if the predicted strain is a true positive. If a strain in the synthetic mixture is not included in the prebuilt database of reference genomes, we used the phylogenetic tree to decide whether a predicted strain is a true positive (Supporting Information Table S4). Briefly, we generated a phylogenetic tree by parsnp [77] (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 a true positive.

Evaluation of predicted gene family profiles

For each microbial strain evaluated in benchmarking data sets, ErrFree reads at 5× sequencing depth were generated by ART simulator [73] (Sync-Single data set) based on its reference genome downloaded from NCBI. The Sync-Single data set contained three replicates for each strain and was used to generate the actual gene family profile of each strain by PanPhlAn and PanPhlAn3-v3.1 (default parameters for sensitive mode). The gene families found in two or more replicates were considered “present.” The actual gene family profile (reference) of each strain was compared with the predicted gene family profile (Figure 2D). The Jaccard distances between microbial strains’ predicted gene family profiles and their reference profiles (or the gene family profile of a randomly sampled reference genome) were computed (Supporting Information - Figure S5). The Precision-Recall curve of gene family profiles for each strain was generated by R package

PRROC [78] using the confidence scores to rank the gene families predicted (Supporting Information Figure S6). 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 data set of four strains (Sync O104 data set, pWGS, 5× sequencing depth). The outbreak-related genes curated from Scholz et al. [35] were used to evaluate StrainPanDA’s gene content prediction (Supporting Information Table S5). We also compared the predicted gene family profiles from StrainPanDA to the prediction from PanPhlAn and PanPhlAn3 (Supporting Information Figure S8).

Runtime evaluation

The runtime of StrainPanDA was measured with the time command in Linux. All these tests were run on a workstation of Intel(R) Xeon(R) Gold 6238 CPU @ 2.10 GHz and 16 GB memory. The runtime (seconds) as a function of sample size was estimated by running StrainPanDA with the downsampled synthetic mixture of four *E. coli* strains (See *Synthetic data of E. coli strains* in Methods section). The runtime (seconds) as a function of strain number was calculated by running StrainPanDA with the pWGS and 25-fold WGSBG data set.

Applications of StrainPanDA in metagenomic data

Case study: Mother–infant gut metagenomes

All available samples of ERP005989 were downloaded from EBI (eight samples failed, Supporting Information Table S6). On the basis of the diet history, infants without diet history were filtered. The rest of the 84 infants were split into three different groups (B_F_F, discontinued breastfeeding at 4 months; B_B_F, discontinued breastfeeding at 12 months; B_B_B, continued breastfeeding; the F_B_M sample was excluded; Supporting Information Table S6). 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 months, or between 4

and 12 months) were included. In the “discontinued breastfeeding” group, infants who stopped breastfeeding by 4 or 12 months were included. For functional interpretation of the subspecies, we grouped the gene families annotated to the same KEGG (downloaded September 1, 2021) ortholog or CAZy (downloaded July 31, 2019) 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] (Supporting Information Table S7). The gene family coverage was calculated as the fraction of detected genes belonging to the subclass (Figure 3C).

Case study: FMT donor–recipient metagenomes

Raw sequencing reads were downloaded from ENA (Accession: PRJNA625520 for the study on Crohn's disease [37], PRJEB23524 for the study on *C. difficile* infection study [29]). For Crohn's disease data set, species relative abundances were estimated using Kraken2 [18] (version 2.0.8-beta) with the miniKraken database (v2_8GB_201904_UPDATE). To identify species associated with remission or relapse, the Wilcoxon rank-sum test was conducted to select differentially abundant species using the mean relative abundances across different time points 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 usearch [69] (v11.0.667; “-ublast”). Alignments with identity >50% and query coverage >50% are kept. To annotate the gene families by CAZy, SeqKit [80] (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_dbcan [81] (2.0.11), which used DIAMOND [82] (2.0.8), HMMer [83] (3.3.2), and Hotpep [84] (2.0.8) with default parameters to predict the CAZy annotation. The CAZy annotations were selected only if it was predicted

by at least two programs. If a gene family was assigned by more than one CAZy annotation, only the first annotation was used. To annotate virulence factors, gene family centroids were mapped against the VFDB (April 9, 2021) by DIAMOND [82] (2.0.8; blastp, query coverage >50% and identity >50%).

AUTHOR CONTRIBUTIONS

Han Hu, Yan Tan, and Lei Dai conceived and supervised the study. Han Hu, Yuxiang Tan, and Chenhao Li developed the algorithm and performed the analysis on simulated and real metagenomic samples. Lei Dai, Han Hu, Yuxiang Tan, and Chenhao Li wrote the manuscript with inputs from Zhenjiang Zech Xu, Yang-Yu Liu, Yan Kou, and Yan Tan.

ACKNOWLEDGMENTS

We would like to thank Haokui Zhou, Ramnik Xavier, and members of the 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; and Reese Hitchings for proofreading the manuscript. This study was supported by the National Key R&D Program of China (No. 2019YFA0906700) and the National Natural Science Foundation of China (Nos. 31971513 and 32061143023).

CONFLICTS OF INTEREST

The results described in this manuscript support pending patent application CN202011146154.3A. Yan Tan is cofounder and shareholder with a personal financial interest in Xbiome. Han Hu and Yan Kou are employees of Xbiome. Lei Dai received research grant support from Xbiome and serves as an unpaid consultant to the company. The remaining author/authors declares/declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The source code of StrainPanDA is available on GitHub (<https://github.com/xbiome/StrainPanDA>) and Zenodo (<https://doi.org/10.5281/zenodo.6668661>). The experiment source code used in the manuscript is available from <https://github.com/xbiome/StrainPanDA-data/tree/main/example#readme>. Supplementary materials (figures, tables, scripts, graphical abstract, slides, videos, Chinese translated version, and update materials) may be found in the online DOI or iMeta Science <http://www.imeta.science/>.

ORCID

Lei Dai  <http://orcid.org/0000-0002-5598-5308>

REFERENCES

1. Poyet, Mathilde, Mathieu Groussin, Sean M. Gibbons, Julian Avila-Pacheco, Xiaofang Jiang, Sean M. Kearney, Allison R. Perrotta, et al. 2019. "A Library of Human Gut Bacterial Isolates Paired with Longitudinal Multiomics Data Enables Mechanistic Microbiome Research." *Nature Medicine* 25: 1442–52. <https://doi.org/10.1038/s41591-019-0559-3>
2. Costea, Paul I., Luis Pedro Coelho, Shinichi Sunagawa, Robin Munch, Jaime Huerta-Cepas, Kristoffer Forslund, Falk Hildebrand, et al. 2017. "Subspecies in the Global Human Gut Microbiome." *Molecular Systems Biology* 13: 960. <https://doi.org/10.15252/msb.20177589>
3. Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. "Diversity Within Species: Interpreting Strains in Microbiomes." *Nature Reviews Microbiology* 18: 491–506. <https://doi.org/10.1038/s41579-020-0368-1>
4. Vatanen, Tommi, Damian R. Plichta, Juhi Somani, Philipp C. Münch, Timothy D. Arthur, Andrew Brantley Hall, Sabine Rudolf, et al. 2019. "Genomic Variation and Strain-Specific Functional Adaptation in the Human Gut Microbiome During Early Life." *Nature Microbiology* 4: 470–9. <https://doi.org/10.1038/s41564-018-0321-5>
5. Ma, Bing, Michael T. France, Jonathan Crabtree, Johanna B. Holm, Michael S. Humphrys, Rebecca M. Brotman, and Jacques Ravel. 2020. "A Comprehensive Non-Redundant Gene Catalog Reveals Extensive Within-Community Intraspecies Diversity in the Human Vagina." *Nature Communications* 11: 940. <https://doi.org/10.1038/s41467-020-14677-3>
6. Scheuerl, Thomas, Meirion Hopkins, Reuben W. Nowell, Damian W. Rivett, Timothy G. Barraclough, and Thomas Bell. 2020. "Bacterial Adaptation Is Constrained in Complex Communities." *Nature Communications* 11: 754. <https://doi.org/10.1038/s41467-020-14570-z>
7. De Filippis, Francesca, Edoardo Pasolli, Adrian Tett, Sonia Tarallo, Alessio Naccarati, Maria De Angelis, Erasmo Neviani, et al. 2019. "Distinct Genetic and Functional Traits of Human Intestinal *Prevotella copri* Strains Are Associated with Different Habitual Diets." *Cell Host & Microbe* 25: 444–53. <https://doi.org/10.1016/j.chom.2019.01.004>
8. Carrow, Hannah C., Lakshmi E. Batachari, and Hiutung Chu. 2020. "Strain Diversity in the Microbiome: Lessons From *Bacteroides fragilis*." *PLOS Pathogens* 16: e1009056. <https://doi.org/10.1371/journal.ppat.1009056>
9. Bisanz, Jordan E., Paola Soto-Perez, Cecilia Noecker, Alexander A. Aksenov, Kathy N. Lam, Grace E. Kenney, Elizabeth N. Bess, et al. 2020. "A Genomic Toolkit for the Mechanistic Dissection of Intractable Human Gut Bacteria." *Cell Host & Microbe* 27: 1001–13. <https://doi.org/10.1016/j.chom.2020.04.006>
10. Ahern, Philip P., Jeremiah J. Faith, and Jeffrey I. Gordon. 2014. "Mining the Human Gut Microbiota for Effector Strains That Shape the Immune System." *Immunity* 40: 815–23. <https://doi.org/10.1016/j.immuni.2014.05.012>
11. Maini Rekdal, Vayu, Elizabeth N. Bess, Jordan E. Bisanz, Peter J. Turnbaugh, and Emily P. Balskus. 2019. "Discovery and Inhibition of an Interspecies Gut Bacterial Pathway for Levodopa Metabolism." *Science* 364: eaau6323. <https://doi.org/10.1126/science.aau6323>
12. Wilson, Matthew R., Yindi Jiang, Peter W. Villalta, Alessia Stornetta, Paul D. Boudreau, Andrea Carrá, Caitlin A. Brennan, et al. 2019. "The Human Gut Bacterial Genotoxin Colibactin Alkylates DNA." *Science* 363: eaar7785. <https://doi.org/10.1126/science.aar7785>
13. Zhang, Chenhong, and Liping Zhao. 2016. "Strain-Level Dissection of the Contribution of the Gut Microbiome to Human Metabolic Disease." *Genome Medicine* 8: 41. <https://doi.org/10.1186/s13073-016-0304-1>
14. Zhai, Rui, Xinhe Xue, Liying Zhang, Xin Yang, Liping Zhao, and Chenhong Zhang. 2019. "Strain-Specific Anti-Inflammatory Properties of Two *Akkermansia muciniphila* Strains on Chronic Colitis in Mice." *Frontiers in Cellular and Infection Microbiology* 9: 239. <https://www.frontiersin.org/article/10.3389/fcimb.2019.00239>
15. Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlan2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12: 902–3. <https://doi.org/10.1038/nmeth.3589>
16. Sun, Zheng, Shi Huang, Meng Zhang, Qiyun Zhu, Niina Haiminen, Anna Paola Carrieri, Yoshiki Vázquez-Baeza, et al. 2021. "Challenges in Benchmarking Metagenomic Profilers." *Nature Methods* 18: 618–26. <https://doi.org/10.1038/s41592-021-01141-3>
17. Milanese, Alessio, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. "Microbial Abundance, Activity and Population Genomic Profiling with mOTUs2." *Nature Communications* 10: 1014. <https://doi.org/10.1038/s41467-019-08844-4>
18. WoodDerrick, E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20: 257. <https://doi.org/10.1186/s13059-019-1891-0>
19. Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, et al. 2019. "1,520 Reference Genomes from Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses." *Nature Biotechnology* 37: 179–85. <https://doi.org/10.1038/s41587-018-0008-8>
20. Liu, Chang, Meng-Xuan Du, Rexiding Abuduaini, Hai-Ying Yu, Dan-Hua Li, Yu-Jing Wang, Nan Zhou, et al. 2021. "Enlightening the Taxonomy Darkness of Human Gut Microbiomes with a Cultured Biobank." *Microbiome* 9: 119. <https://doi.org/10.1186/s40168-021-01064-3>
21. Segata, Nicola. 2018. "On the Road to Strain-Resolved Comparative Metagenomics." *mSystems* 3: e00190-00117. <https://doi.org/10.1128/mSystems.00190-17>
22. Anyansi, Christine, Timothy J. Straub, Abigail L. Manson, Ashlee M. Earl, and Thomas Abeel. 2020. "Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data." *Frontiers in Microbiology* 11: 1925. <https://doi.org/10.3389/fmicb.2020.01925>
23. Yassour, Moran, Eeva Jason, Larson J. Hogstrom, Timothy D. Arthur, Surya Tripathi, Heli Siljander, Jenni Selvenius, et al. 2018. "Strain-Level Analysis of Mother-to-Child Bacterial Transmission During the First Few Months of Life." *Cell Host & Microbe* 24: 146–54. <https://doi.org/10.1016/j.chom.2018.06.007>

24. Asnicar, Francesco, Serena Manara, Moreno Zolfo, Tin Truong Duy, Matthias Scholz, Federica Armanini, Pamela Ferretti, et al. 2017. "Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling." *mSystems* 2: e00164-00116. <https://doi.org/10.1128/mSystems.00164-16>
25. Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure and Genetic Diversity From Metagenomes." *Genome Research* 27: 626–38. <https://doi.org/10.1101/gr.216242.116>
26. Korpela, Katri, Paul Costea, Luis Pedro Coelho, Stefanie Kandels-Lewis, Gonneke Willemssen, Dorret I. Boomsma, Nicola Segata, and Peer Bork. 2018. "Selective Maternal Seeding and Environment Shape the Human Gut Microbiome." *Genome Research* 28: 561–8. <https://doi.org/10.1101/gr.233940.117>
27. Shao, Yan, Samuel C. Forster, Evdokia Tsaliki, Kevin Vervier, Angela Strang, Nandi Simpson, Nitin Kumar, et al. 2019. "Stunted Microbiota and Opportunistic Pathogen Colonization in Caesarean-Section Birth." *Nature* 574: 117–21. <https://doi.org/10.1038/s41586-019-1560-1>
28. Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. 2015. "ConStrains Identifies Microbial Strains in Metagenomic Datasets." *Nature Biotechnology* 33: 1045–52. <https://doi.org/10.1038/nbt.3319>
29. Smillie, Christopher S., Jenny Sauk, Dirk Gevers, Jonathan Friedman, Jaeyun Sung, Ilan Youngster, Elizabeth L. Hohmann, et al. 2018. "Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation." *Cell Host & Microbe* 23: 229–40. <https://doi.org/10.1016/j.chom.2018.01.003>
30. Albanese, Davide, and Claudio Donati. 2017. "Strain Profiling and Epidemiology of Bacterial Species from Metagenomic Sequencing." *Nature Communications* 8: 2260. <https://doi.org/10.1038/s41467-017-02209-5>
31. Olm, Matthew R., Alexander Crits-Christoph, Keith Bouma-Gregson, Brian A. Firek, Michael J. Morowitz, and Jillian F. Banfield. 2021. "Instrain Profiles Population Microdiversity from Metagenomic Data and Sensitive Detects Shared Microbial Strains." *Nature Biotechnology* 39: 727–36. <https://doi.org/10.1038/s41587-020-00797-0>
32. Yassour, Moran, Tommi Vatanen, Heli Siljander, Anu-Maaria Hämäläinen, Taina Härkönen, Samppa J. Ryhänen, Eric A. Franzosa, et al. 2016. "Natural History of the Infant Gut Microbiome and Impact of Antibiotic Treatment on Bacterial Strain Diversity and Stability." *Science Translational Medicine* 8: 343ra381. <https://doi.org/10.1126/scitranslmed.aad0917>
33. Wang, Shuai, Yiqi Jiang, and Shuaicheng Li. 2020. "PStrain: An Iterative Microbial Strains Profiling Algorithm for Shotgun Metagenomic Sequencing Data." *Bioinformatics* 36: 5499–5506. <https://doi.org/10.1093/bioinformatics/btaa1056>
34. Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with Biobakery 3." *eLife* 10: e65088. <https://doi.org/10.7554/eLife.65088>
35. Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, et al. 2016. "Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics." *Nature Methods* 13: 435–8. <https://doi.org/10.1038/nmeth.3802>
36. Bäckhed, Fredrik, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, et al. 2015. "Dynamics and Stabilization of the Human Gut Microbiome During the First Year of Life." *Cell Host & Microbe* 17: 690–703. <https://doi.org/10.1016/j.chom.2015.04.004>
37. Kong, Lingjia, Jason Lloyd-Price, Tommi Vatanen, Philippe Seksik, Laurent Beaugerie, Tabassome Simon, Hera Vlamakis, Harry Sokol, and Ramnik J. Xavier. 2020. "Linking Strain Engraftment in Fecal Microbiota Transplantation with Maintenance of Remission in Crohn's Disease." *Gastroenterology* 159: 2193–2202. <https://doi.org/10.1053/j.gastro.2020.08.045>
38. Carr, Rogan, Shai S. Shen-Orr, and Elhanan Borenstein. 2013. "Reconstructing the Genomic Content of Microbiome Taxa Through Shotgun Metagenomic Deconvolution." *PLOS Computational Biology* 9: e1003292. <https://doi.org/10.1371/journal.pcbi.1003292>
39. Lee, Daniel D., and H. Sebastian Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401: 788–91. <https://doi.org/10.1038/44565>
40. Gaujoux, Renaud, and Cathal Seoighe. 2010. "A Flexible R Package for Nonnegative Matrix Factorization." *BMC Bioinformatics* 11: 367. <https://doi.org/10.1186/1471-2105-11-367>
41. Kim, Hyunsoo, and Haesun Park. 2007. "Sparse Non-Negative Matrix Factorizations Via Alternating Non-Negativity-Constrained Least Squares for Microarray Data Analysis." *Bioinformatics* 23: 1495–1502. <https://doi.org/10.1093/bioinformatics/btm134>
42. Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28: 27–30. <https://doi.org/10.1093/nar/28.1.27>
43. Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. "The Carbohydrate-Active Enzymes Database (CAZy) in 2013." *Nucleic Acids Research* 42: D490–D5. <https://doi.org/10.1093/nar/gkt1178>
44. Liu, Bo, Dandan Zheng, Qi Jin, Lihong Chen, and Jian Yang. 2019. "VFDB 2019: A Comparative Pathogenomic Platform with an Interactive Web Interface." *Nucleic Acids Research* 47: D687–D92. <https://doi.org/10.1093/nar/gky1080>
45. Mellmann, Alexander, Dag Harmsen, Craig A. Cummings, Emily B. Zentz, Shana R. Leopold, Alain Rico, Karola Prior, et al. 2011. "Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology." *PLOS ONE* 6: e22751. <https://doi.org/10.1371/journal.pone.0022751>
46. Sela, David A., Jarrod Chapman, Anthony Adeuya, Jae-Han Kim, Feng Chen, Terrence R. Whitehead, Alla Lapidus, et al. 2008. "The Genome Sequence of *Bifidobacterium longum* subsp. *infantis* Reveals Adaptations for Milk Utilization within the Infant Microbiome." *Proceedings of the National Academy of Sciences* 105: 18964–9. <https://doi.org/10.1073/pnas.0809584105>
47. Sela, David A. 2011. "Bifidobacterial Utilization of Human Milk Oligosaccharides." *International Journal of Food Microbiology* 149: 58–64. <https://doi.org/10.1016/j.ijfoodmicro.2011.01.025>

48. Katayama, Takane. 2016. "Host-Derived Glycans Serve as Selected Nutrients for the Gut Microbe: Human Milk Oligosaccharides and Bifidobacteria." *Bioscience, Biotechnology, and Biochemistry* 80: 621–32. <https://doi.org/10.1080/09168451.2015.1132153>
49. Duranti, Sabrina, Gabriele Andrea Lugli, Leonardo Mancabelli, Federica Armanini, Francesca Turroni, Kieran James, Pamela Ferretti, et al. 2017. "Maternal Inheritance of Bifidobacterial Communities and Bifidophages in Infants Through Vertical Transmission." *Microbiome* 5: 66. <https://doi.org/10.1186/s40168-017-0282-6>
50. Asakuma, Sadaki, Emi Hatakeyama, Tadasu Urashima, Erina Yoshida, Takane Katayama, Kenji Yamamoto, Hidehiko Kumagai, et al. 2011. "Physiology of Consumption of Human Milk Oligosaccharides by Infant Gut-Associated Bifidobacteria." *Journal of Biological Chemistry* 286: 34583–92. <https://doi.org/10.1074/jbc.M111.248138>
51. Schimmel, Patrick, Lennart Kleinjans, Roger S. Bongers, Jan Knol, and Clara Belzer. 2021. "Breast Milk Urea as a Nitrogen Source for Urease Positive *Bifidobacterium infantis*." *FEMS Microbiology Ecology* 97: fiab019. <https://doi.org/10.1093/femsec/fiab019>
52. Kujawska, Magdalena, Sabina Leanti La Rosa, Laure C. Roger, Phillip B. Pope, Lesley Hoyles, Anne L. McCartney, and Lindsay J. Hall. 2020. "Succession of *Bifidobacterium longum* Strains in Response to a Changing Early Life Nutritional Environment Reveals Dietary Substrate Adaptations." *iScience* 23: 101368. <https://doi.org/10.1016/j.isci.2020.101368>
53. Garrido, Daniel, Santiago Ruiz-Moyano, Nina Kirmiz, Jasmine C. Davis, Sarah M. Totten, Danielle G. Lemay, Juan A. Ugalde, et al. 2016. "A Novel Gene Cluster Allows Preferential Utilization of Fucosylated Milk Oligosaccharides in *Bifidobacterium longum* subsp. *longum* SC596." *Scientific Reports* 6: 35045. <https://doi.org/10.1038/srep35045>
54. Wang, Shaopu, Shuqin Zeng, Muireann Egan, Paul Cherry, Conall Strain, Emilene Morais, Patrick Boyaval, et al. 2021. "Metagenomic Analysis of Mother–Infant Gut Microbiome Reveals Global Distinct and Shared Microbial Signatures." *Gut Microbes* 13: 1911571. <https://doi.org/10.1080/19490976.2021.1911571>
55. Frese, Steven A., Andra A. Hutton, Lindsey N. Contreras, Claire A. Shaw, Michelle C. Palumbo, Giorgio Casaburi, Gege Xu, et al. 2017. "Persistence of Supplemented *Bifidobacterium longum* subsp. *infantis* EVC001 in Breastfed Infants." *mSphere* 2: e00501–e17. <https://doi.org/10.1128/mSphere.00501-17>
56. Sela, David A., and David A. Mills. 2010. "Nursing Our Microbiota: Molecular Linkages Between Bifidobacteria and Milk Oligosaccharides." *Trends in Microbiology* 18: 298–307. <https://doi.org/10.1016/j.tim.2010.03.008>
57. Smits, Loek P., Kristien E. C. Bouter, Willem M. de Vos, Thomas J. Borody, and Max Nieuwdorp. 2013. "Therapeutic Potential of Fecal Microbiota Transplantation." *Gastroenterology* 145: 946–53. <https://doi.org/10.1053/j.gastro.2013.08.058>
58. Ooijevaar, Rogier E., Elisabeth M. Terveer, Hein W. Verspaget, Ed J. Kuijper, and Josbert J. Keller. 2019. "Clinical Application and Potential of Fecal Microbiota Transplantation." *Annual Review of Medicine* 70: 335–51. <https://doi.org/10.1146/annurev-med-111717-122956>
59. Peraro, Matteo Dal, and F. Gisou van der Goot. 2016. "Pore-Forming Toxins: Ancient, but Never Really Out of Fashion." *Nature Reviews Microbiology* 14: 77–92. <https://doi.org/10.1038/nrmicro.2015.3>
60. Tan, Joel M. J., Nora Mellouk, Suzanne E. Osborne, Dustin A. Ammendolia, Diana N. Dyer, Ren Li, Diede Brunen, et al. 2018. "An ATG16L1-Dependent Pathway Promotes Plasma Membrane Repair and Limits *Listeria monocytogenes* Cell-to-Cell Spread." *Nature Microbiology* 3: 1472–85. <https://doi.org/10.1038/s41564-018-0293-5>
61. Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. "DESMAN: A New Tool for De Novo Extraction of Strains from Metagenomes." *Genome Biology* 18: 181. <https://doi.org/10.1186/s13059-017-1309-9>
62. Li, Xin, Haiyan Hu, and Xiaoman Li. 2021. "Mixtures: A Novel Tool for Bacterial Strain Genome Reconstruction from Reads." *Bioinformatics* 37: 575–7. <https://doi.org/10.1093/bioinformatics/btaa728>
63. Quince, Christopher, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S. Soyer, J. Kimberly Summers, Antoine Limasset, et al. 2021. "STRONG: Metagenomics Strain Resolution on Assembly Graphs." *Genome Biology* 22: 214. <https://doi.org/10.1186/s13059-021-02419-7>
64. Sheth, Ravi U., Mingqiang Li, Weiqian Jiang, Peter A. Sims, Kam W. Leong, and Harris H. Wang. 2019. "Spatial Metagenomic Characterization of Microbial Biogeography in the Gut." *Nature Biotechnology* 37: 877–83. <https://doi.org/10.1038/s41587-019-0183-2>
65. Chng, Kern Rei, Chenhao Li, Denis Bertrand, Amanda Hui Qi Ng, Junmei Samantha Kwah, Hwee Meng Low, Chengxuan Tong, et al. 2020. "Cartography of Opportunistic Pathogens and Antibiotic Resistance Genes in a Tertiary Hospital Environment." *Nature Medicine* 26: 941–51. <https://doi.org/10.1038/s41591-020-0894-4>
66. Palleja, Albert, Kristian H. Mikkelsen, Sofia K. Forslund, Alireza Kashani, Kristine H. Allin, Trine Nielsen, Tue H. Hansen, et al. 2018. "Recovery of Gut Microbiota of Healthy Adults Following Antibiotic Exposure." *Nature Microbiology* 3: 1255–65. <https://doi.org/10.1038/s41564-018-0257-9>
67. Chng, Kern Rei, Tarini Shankar Ghosh, Yi Han Tan, Tannistha Nandi, Ivor Russel Lee, Amanda Hui Qi Ng, Chenhao Li, et al. 2020. "Metagenome-Wide Association Analysis Identifies Microbial Determinants of Post-Antibiotic Ecological Recovery in the Gut." *Nature Ecology & Evolution* 4: 1256–67. <https://doi.org/10.1038/s41559-020-1236-0>
68. Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17: 132. <https://doi.org/10.1186/s13059-016-0997-x>
69. Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than BLAST." *Bioinformatics* 26: 2460–1. <https://doi.org/10.1093/bioinformatics/btq461>
70. Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9: 357–9. <https://doi.org/10.1038/nmeth.1923>

71. Danecek, Petr, K. Bonfield James, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10: giab008. <https://doi.org/10.1093/gigascience/giab008>
72. Pan, Weiwei., and Finale. Doshi-Velez. 2016. "AQ13A Characterization of the Non-Uniqueness of Nonnegative Matrix Factorizations." ArXiv abs/1604.0065300: n. pag. <https://doi.org/10.48550/arXiv.1604.00653>
73. Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. "ART: a Next-Generation Sequencing Read Simulator." *Bioinformatics* 28: 593–4. <https://doi.org/10.1093/bioinformatics/btr708>
74. Lin, Jianhua. 1991. "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory* 37: 145–51. <https://doi.org/10.1109/18.61115>
75. McMurdie, Paul J., and Susan Holmes. 2013. "phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PLOS ONE* 8: e61217. <https://doi.org/10.1371/journal.pone.0061217>
76. Matthews, Brian W. 1975. "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme." *Biochimica et Biophysica Acta (BBA)—Protein Structure* 405: 442–51. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
77. Treangen, Todd J., Brian D. Ondov, Sergey Koren, and Adam M. Phillippy. 2014. "The Harvest Suite for Rapid Core-Genome Alignment and Visualization of Thousands of Intraspecific Microbial Genomes." *Genome Biology* 15: 524. <https://doi.org/10.1186/s13059-014-0524-x>
78. Grau, Jan, Ivo Grosse, and Jens Keilwagen. 2015. "PRROC: Computing and Visualizing Precision-Recall and Receiver Operating Characteristic Curves in R." *Bioinformatics* 31: 2595–7. <https://doi.org/10.1093/bioinformatics/btv153>
79. Odamaki, Toshitaka, Ayako Horigome, Hirotsuke Sugahara, Nanami Hashikura, Junichi Minami, Jin-zhong Xiao, and Fumiaki Abe. 2015. "Comparative Genomics Revealed Genetic Diversity and Species/Strain-Level Differences in Carbohydrate Metabolism of Three Probiotic Bifidobacterial Species." *International Journal of Genomics* 2015: 567809. <https://doi.org/10.1155/2015/567809>
80. Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation." *PLOS ONE* 11: e0163962. <https://doi.org/10.1371/journal.pone.0163962>
81. Zhang, Han, Tanner Yohe, Le Huang, Sarah Entwistle, Peizhi Wu, Zhenglu Yang, Peter K. Busk, Ying Xu, and Yanbin Yin. 2018. "dbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation." *Nucleic Acids Research* 46: W95–W101. <https://doi.org/10.1093/nar/gky418>
82. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12: 59–60. <https://doi.org/10.1038/nmeth.3176>
83. Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLOS Computational Biology* 7: e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
84. Busk, Peter K., Bo Pilgaard, Mateusz J. Lezyk, Anne S. Meyer, and Lene Lange. 2017. "Homology to Peptide Pattern for Annotation of Carbohydrate-Active Enzymes and Prediction of Function." *BMC Bioinformatics* 18: 214. <https://doi.org/10.1186/s12859-017-1625-9>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hu, Han, Yuxiang Tan, Chenhao Li, Junyu Chen, Yan Kou, Zhenjiang Zech Xu, Yang-Yu Liu, Yan Tan, and Lei Dai. 2022. "StrainPanDA: Linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data." *iMeta* 1, e41. <https://doi.org/10.1002/imt2.41>