

C. difficile may be overdiagnosed in adults and is a prevalent commensal in infants

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted

September 12, 2023 (this version)

Sent for peer review

July 19, 2023

Posted to bioRxiv

June 27, 2023

Pamela Ferretti, Jakob Wirbel, Oleksandr M Maistrenko, Thea Van Rossum, Renato Alves, Anthony Fullam, Wasii Akanni, Christian Schudoma, Anna Schwarz, Roman Thielemann, Leonie Thomas, Stefanie Kandels, Rajna Hercog, Anja Telzerow, Ivica Letunic, Michael Kuhn, Georg Zeller, Thomas SB Schmidt, Peer Bork 

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany • Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Microbiology & Biogeochemistry, 1797 SZ, 't Horntje (Texel), Netherlands • Genomic Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany • biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg • Max Delbrück Centre for Molecular Medicine, Berlin, Germany • Yonsei Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea • Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

 https://en.wikipedia.org/wiki/Open_access

 <https://creativecommons.org/licenses/by/4.0/>

Abstract

Clostridioides difficile is an urgent threat in hospital-acquired infections world-wide, yet the microbial composition associated with *C. difficile*, in particular in *C. difficile* infection (CDI) cases, remains poorly characterised. To investigate the gut microbiome composition in CDI patients, we analysed 534 metagenomes from 10 publicly available CDI study populations. We then tracked *C. difficile* on a global scale, screening 42,900 metagenomes from 253 public studies. Among the CDI cohorts, we detected *C. difficile* in only 30% of the stool samples from CDI patients. However, we found that multiple other toxigenic species capable of inducing CDI-like symptomatology were prevalent. In addition, the majority of the investigated studies did not adhere to the recommended guidelines for a correct CDI diagnosis.

In the global survey, we found that *C. difficile* prevalence, abundance and biotic context were age-dependent. *C. difficile* is a rare taxon associated with reduced diversity in healthy adults, but common and associated with increased diversity in infants. We identified a group of species co-occurring with *C. difficile* exclusively in healthy infants, enriched in obligate anaerobes and in species typical of the healthy adult gut microbiome. *C. difficile* in healthy infants was therefore associated with multiple indicators of healthy gut microbiome maturation.

Our analysis raises concerns about potential CDI overdiagnosis and suggests that *C. difficile* is an important commensal in infants and that its asymptomatic carriage in adults depends on microbial context.

eLife assessment

This **important** study points out discrepancies between the clinical diagnosis of *Clostridioides difficile* infection and the lack of detectable *C. difficile* in gut microbiome samples, as well as different relationships between asymptomatic *C. difficile* carriage and adult or infant gut microbiota compositions. While the **solid** analysis of a comprehensive and diverse metagenomic dataset suggests an over-diagnosis of *C. difficile* infection and an under-diagnosis of other putative enteric pathogens, the work requires addressing the detection limitations of the approach to be more **convincing**. This work will interest microbiologists and clinicians concerned with understanding the role of *C. difficile* in gut microbiota health and dysbiosis.

Introduction

Clostridioides difficile (previously *Clostridium* or *Peptoclostridium difficile*) was first isolated in 1935 from the stools of healthy infants (Hall and O'toole 1935 [↗](#)), but later identified as a causative agent of pseudomembranous colitis (Bartlett et al. 1978 [↗](#)), a severe condition with potentially life-threatening outcomes. While not all *C. difficile* strains are toxigenic, *C. difficile* toxin A (TcdA) and toxin B (TcdB) can increase intestinal permeability and promote intense inflammation (Voth and Ballard 2005 [↗](#)), a condition referred to (Bartlett et al. 1978 [↗](#)) as *C. difficile* infection (CDI). CDI is typically characterised by diarrhoea and/or pseudomembranous colitis and often unresponsive to repeated antibiotic treatments. Known risk factors include hospitalisation, age (≥ 65 years of age), and broad-spectrum antibiotic therapy (Eze et al. 2017). CDI has an increasing clinical relevance (at an estimated annual economic burden of US\$6.3 billion in the US alone (Bartlett et al. 1978 [↗](#))) and is considered one of the most urgent threats in hospital-acquired infections (CDC, 2019 [↗](#)).

C. difficile has primarily been studied as a pathogen, with a strong focus on CDI aetiology (Smits et al. 2016 [↗](#)), mechanisms of *C. difficile* toxicity (Voth and Ballard 2005 [↗](#)) or efficacious therapies (Kociolek and Gerding 2016 [↗](#)). However, an increasing number of other enteropathogens has been linked to antibiotics-associated diarrhoea (AAD) with similar, and sometimes indistinguishable, symptomatology to CDI (Tang et al. 2016 [↗](#); Chia et al. 2018 [↗](#); Larcombe et al. 2018 [↗](#)), making the correct diagnosis, estimation and management of CDI particularly challenging.

Far less is known about asymptomatic *C. difficile* carriage among the healthy population, as sentinel studies are often limited in sample size and geography (Galdys et al. 2014 [↗](#); Kato et al. 2001 [↗](#); Cui et al. 2021; Mani et al. 2023) (Galdys et al. 2014 [↗](#); Kato et al. 2001 [↗](#)). Gut microbial composition associated with *C. difficile* presence in healthy subjects, in particular in infants where *C. difficile* is highly prevalent (Rousseau et al. 2012 [↗](#)), remains poorly understood.

Here we leverage public shotgun metagenomic data to characterise gut microbial signatures associated with CDI in adult and elderly patients and to quantify other AAD-associated species with similar signatures. We then contextualise our (Rousseau et al. 2012 [↗](#)) analysis by tracing *C. difficile* in a broad multi-habitat collection of 42,900 metagenomes. We investigate *C. difficile* prevalence, relative abundance and its associated microbial community structure and composition in the human gut over lifetime, with particular focus on healthy infants, where several lines of evidence point to *C. difficile* being a commensal and a hallmark species of healthy infant gut microbiome maturation.

***C. difficile* is metagenomically detectable in only 30% of patients with a CDI diagnosis**

We conducted a metagenomic meta-analysis of the gut microbiome associated with CDI, analysing 534 fecal samples across 10 geographically diverse study populations (Vincent et al. 2016 [\[1\]](#); Smillie et al. 2018 [\[2\]](#); Podlesny and Florian Fricke; Kim et al. 2020 [\[3\]](#); Langdon et al. 2021 [\[4\]](#); Monaghan et al. 2020 [\[5\]](#); Stewart et al. 2019 [\[6\]](#); Kumar et al. 2017 [\[7\]](#); Watson et al.), including patients diagnosed with CDI, diseased control subjects without CDI (D-Ctr, see Methods), and healthy subjects (H-Ctr) (Figure 1A [\[8\]](#) and Supplementary Table 1).

Using mOTUs-based species-level taxonomic profiling, *C. difficile* was detectable in just 30% of CDI samples at established metagenomic detection limits (Milanese et al. 2019 [\[9\]](#)) (Figure 1A [\[8\]](#)), with considerable variance across study populations (9.2%-92.3%), and with much lower prevalence among diseased (2.6%) and healthy (1.1%) controls (Supplementary Figure 1A). *C. difficile* toxin genes *tcdA* and *tcdB* were detectable in about half of *C. difficile*-positive CDI metagenomes, but not in any *C. difficile*-negative CDI samples nor controls, indicating a lower sensitivity of toxin-based detection (Figure 1A [\[8\]](#), see Methods). We confirmed that marker gene-based *C. difficile* detection was not limited by sequencing depth (linear mixed-effect model $p=0.35$, ANOVA, adjusted $p=0.54$; $R^2=0.01$), and that detection was not skewed towards metagenomic detection limits (Parks et al. 2021 [\[10\]](#)). Neither species richness (Figure 1B [\[11\]](#)) nor composition (Supplementary Figure 2) differed between *C. difficile*-positive and -negative CDI samples, suggesting that *C. difficile* presence was not associated with characteristic community-level shifts in these generally disbalanced microbiomes.

The microbiome of CDI patients is characterised by an enrichment of enteropathogens beyond *C. difficile*

Several other species were enriched in prevalence or abundance in CDI patients relative to controls (Figure 1C [\[12\]](#), Supplementary Figure 3). Among these, we identified several opportunistic enteropathogens known to induce antibiotic-associated diarrhoea (AAD), resulting in a CDI-like diarrhoeal symptomatology (Chia et al. 2018 [\[13\]](#); Chia et al. 2017 [\[14\]](#); Kiu and Hall 2018 [\[15\]](#); Zollner-Schwetz et al. 2008 [\[16\]](#); Larcombe et al. 2016 [\[17\]](#); Larcombe et al. 2018 [\[18\]](#); Högenauer et al. 1998 [\[19\]](#); Fakhkhari et al. 2022), such as *Staphylococcus aureus*, *Klebsiella oxytoca*, *Enterococcus faecalis*, *Enterobacter cloacae*, *Citrobacter amalonaticus*, *Clostridium innocuum*, *Clostridium perfringens*, and *Pseudomonas aeruginosa*. In particular, *C. innocuum* was highly prevalent (63.2%) in CDI samples, and significantly enriched relative to both healthy and diseased controls (Figure 1C [\[12\]](#), Wilcoxon test $p=8.2 \times 10^{-16}$ and $p=4.3 \times 10^{-3}$, respectively). 94% of all CDI samples contained at least one of these other AAD species (*C. difficile* excluded) with variable prevalence across study populations (Fisher's test, $p=2.2 \times 10^{-16}$, OR=29.62; Supplementary Figure 1A and 1B [\[11\]](#)). Presence and potential toxigenicity of several enteropathogens was further confirmed by the detection of characteristic toxin genes that were likewise enriched in CDI samples (Figure 1D [\[20\]](#)), in particular enteropathogenic *E. coli* and *S. flexneri* toxin genes (see Methods for full list). Overall, CDI samples, independently of *C. difficile* presence, were significantly enriched ($p=2.2 \times 10^{-16}$) in species carrying toxin genes in terms of cumulative relative abundance, indicating that CDI symptomatology could be entirely or partially driven by other enteropathogens than *C. difficile*.

To integrate these univariate associations of individual species into characteristic and predictive multi-species signatures of CDI, we next trained a series of LASSO-regularised logistic regression models in a leave-one-study-out validation approach (see Methods). As expected, *C. difficile* was the most predictive species for CDI samples, although the enrichment of other putative AAD enteropathogens (such as *S. aureus* or *C. perfringens*) was also highly characteristic, in particular

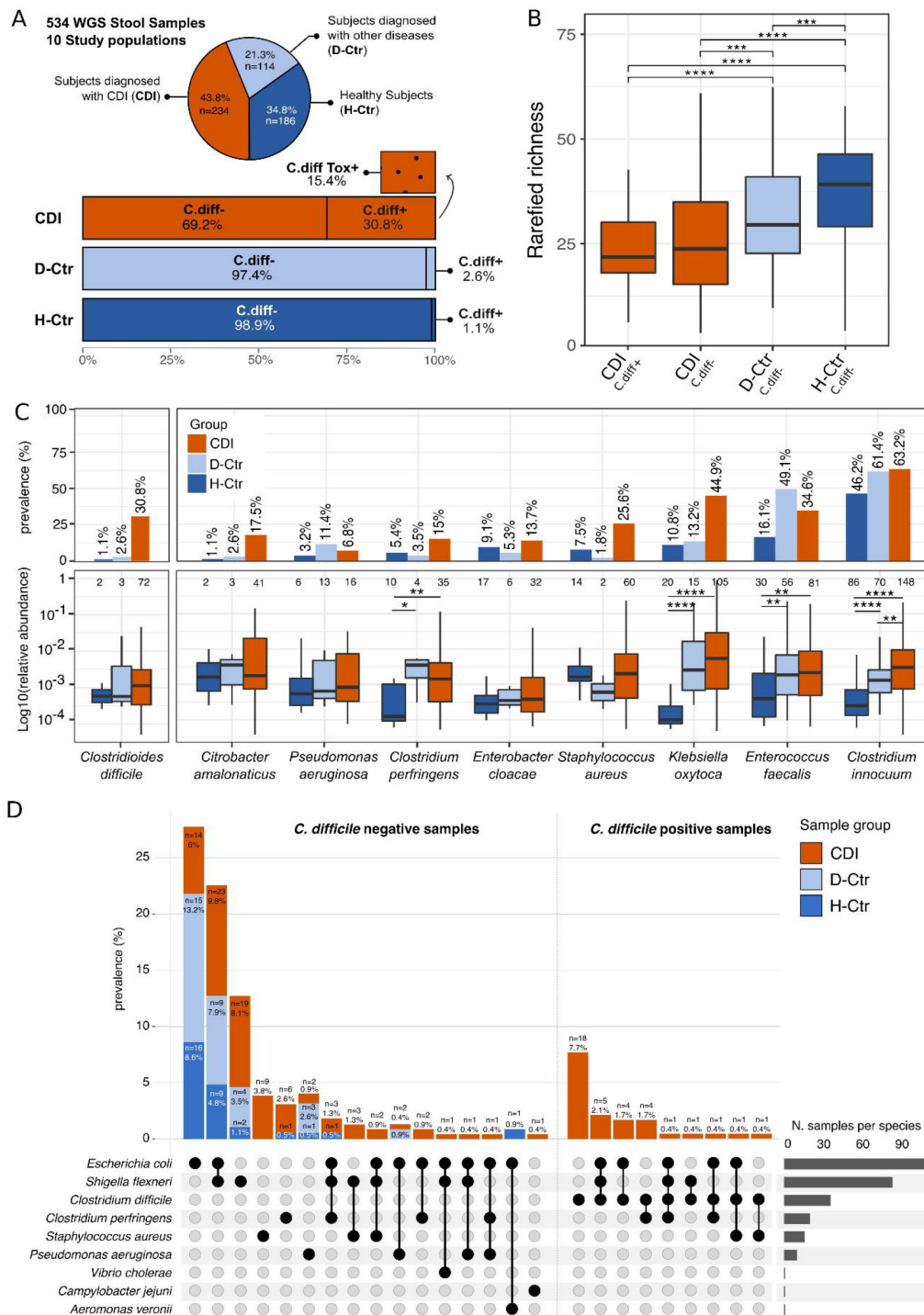


Figure 1.

Metagenomic meta-analysis of the gut microbiome associated with CDI

(A) On top, fecal metagenomes used in the meta-analysis from 10 public CDI or diarrhoeal study populations, divided by groups: subjects diagnosed with CDI (CDI, n=234), subjects diagnosed with other diseases than CDI (D-Ctr, n=114), and healthy subjects (H-Ctr, n=186) samples. See Methods for further group description. At the bottom, fraction of *C. difficile* positive samples per group. *C. difficile* toxin genes were found only among CDI *C. difficile* positive samples. (B) Rarefied species richness across the three groups. (C) Prevalence, relative abundance of species that induce antibiotic-associated diarrhoea (AAD) among CDI and controls. (D) Prevalence of single and multiple toxigenic species found among *C. difficile* positive and *C. difficile* negative samples, divided by sample group. Mean comparison p-values calculated using Wilcoxon-test, * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, **** for $P \leq 0.0001$, empty when non significant.

in CDI samples where *C. difficile* was not detectable (**Figure 2A**). Moreover, several oral species (such as *Veillonella parvula*, *Veillonella atypica* or *Rothia dentocariosa*) were enriched in CDI gut metagenomes, implying increased oral-intestinal microbial transmission (Schmidt et al., 2019). Enrichment of common probiotics (*Lactobacillus casei*, *L. plantarum* or *L. fermentum*) in CDI is likely due to the widespread use of probiotic therapy among these patients (Golić et al. 2017; Na and Kelly 2011).

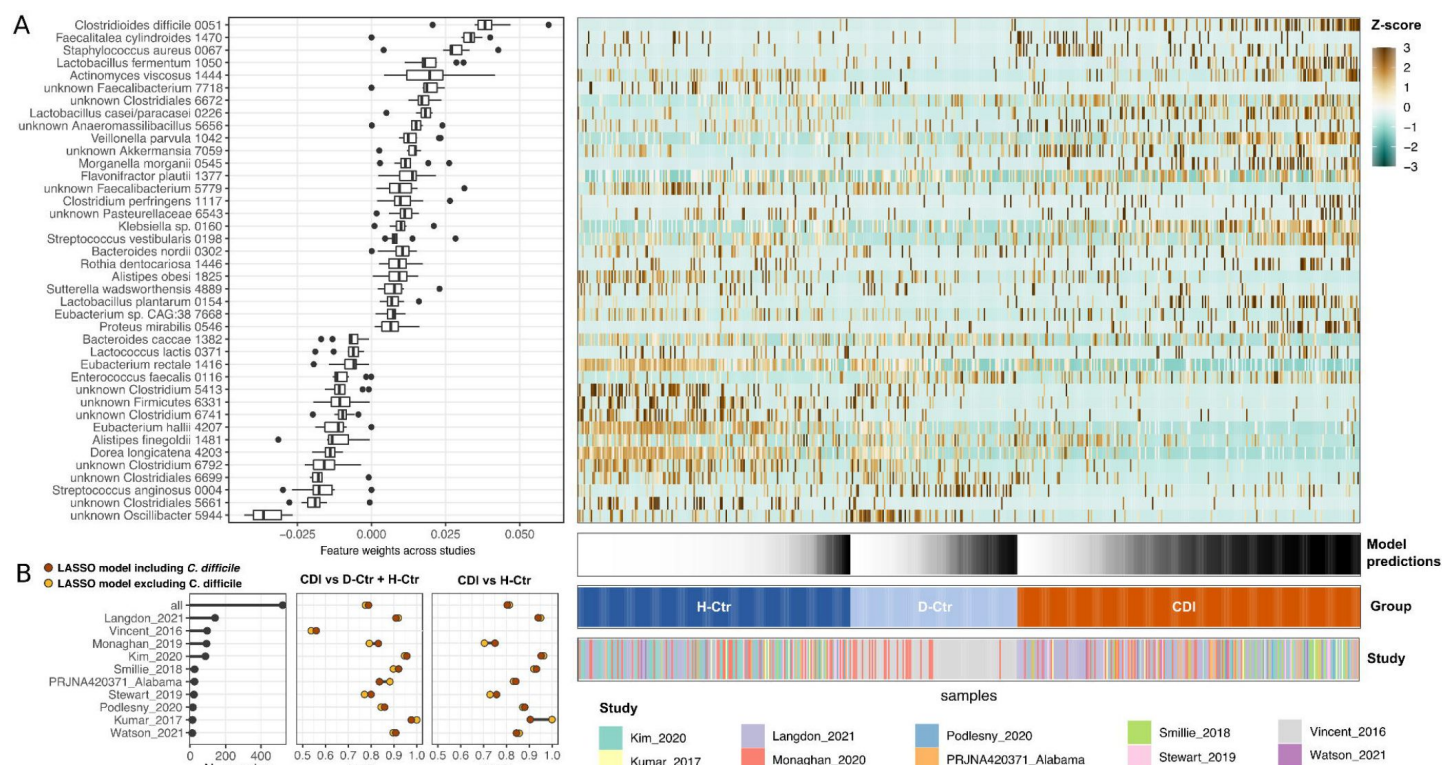
Using these models, CDI samples were distinguishable from diseased and healthy controls with high accuracy, at an area under the receiver operating characteristics curve (AUROC) of 0.78, ranging between 0.56 and 0.98 depending on the study population (**Figure 2B**), in line with a previous 16S-based investigation (Schubert et al. 2014). Model performance moderately improved when considering only healthy controls (overall AUROC = 0.81) (**Figure 2B**). To assess the importance of *C. difficile* to predicting CDI state, we next re-trained models under explicit exclusion of *C. difficile*. Interestingly, this did not lead to a noticeable reduction in accuracy (**Figure 2B**), indicating that the presence or enrichment of *C. difficile* was indeed not an essential defining feature of CDI samples. Rather, *C. difficile* appeared to be just one of several AAD species characterising the CDI-associated microbiome in varying constellations, forming a loose clique instead of fixed co-occurring species groups.

Extensive asymptomatic carriage of *C. difficile* across age and geography

Given that *C. difficile* appears to have a much lower prevalence in CDI patients than anticipated, we next studied its prevalence in an extended collection of 253 publicly available studies, including a novel cohort of 447 samples from the citizen science project MyMicrobes, for a total of 42,900 shotgun metagenomic samples (**Supplementary Table 2**). This collection included healthy and diseased subjects of all ages, with samples from multiple body sites, hosts, environments and geographic locations (**Figure 3A** and **3B**). Statements from this point onwards refer to this extended collection of samples.

C. difficile was detected almost exclusively in human and animal feces (8.7% and 1.7% prevalence, respectively) with sporadic detection in hospital surfaces (0.2%), the respiratory tract of diseased patients (0.4%) and freshwater (0.9%) (**Figure 3C** and **Supplementary Table 3**). *C. difficile* was also detectable in the colon, ileum and cecum lumen, and biopsies from the colon and rectal mucosa of healthy subjects. We again confirmed in this larger dataset that there was no relationship between *C. difficile* detection and sequencing depth (linear mixed-effect model $p=0.36$; logistic regression, ANOVA adjusted $p=0.52$, **Supplementary Figure 4A and 4B**). Nevertheless, we adjusted all subsequent analyses for total sequencing depth.

In healthy human populations, both *C. difficile* prevalence and relative abundance in fecal samples varied considerably across age ranges. Infants aged one year or younger showed by far the highest *C. difficile* carriage, up to 76.5% prevalence at the age of 8 to 10 months (**Figure 4A**), at an average relative abundance of 1.1% (and up to 50% in a two-weeks-old healthy preterm infant). *C. difficile* prevalence remained elevated (43.4%) between 1 and 4 years of age, in contrast to previous reports (Jangi and Thomas Lamont 2010; Lees et al. 2016). With increasing age, *C. difficile* was increasingly rare, found in less than 1% of the healthy population (6.5% in adolescents, 0.8% in adults) (**Figure 4A** and **Supplementary Table 3**) at average abundances of 0.2%, suggesting a considerably lower asymptomatic carriage among adults than previously estimated (up to 15% (Crobach et al. 2018)). *C. difficile* carriage varied across geography (**Supplementary Figure 5A**) and was lower in non-westernised populations, at 4.8% and 0.4% in infants and adults, respectively.



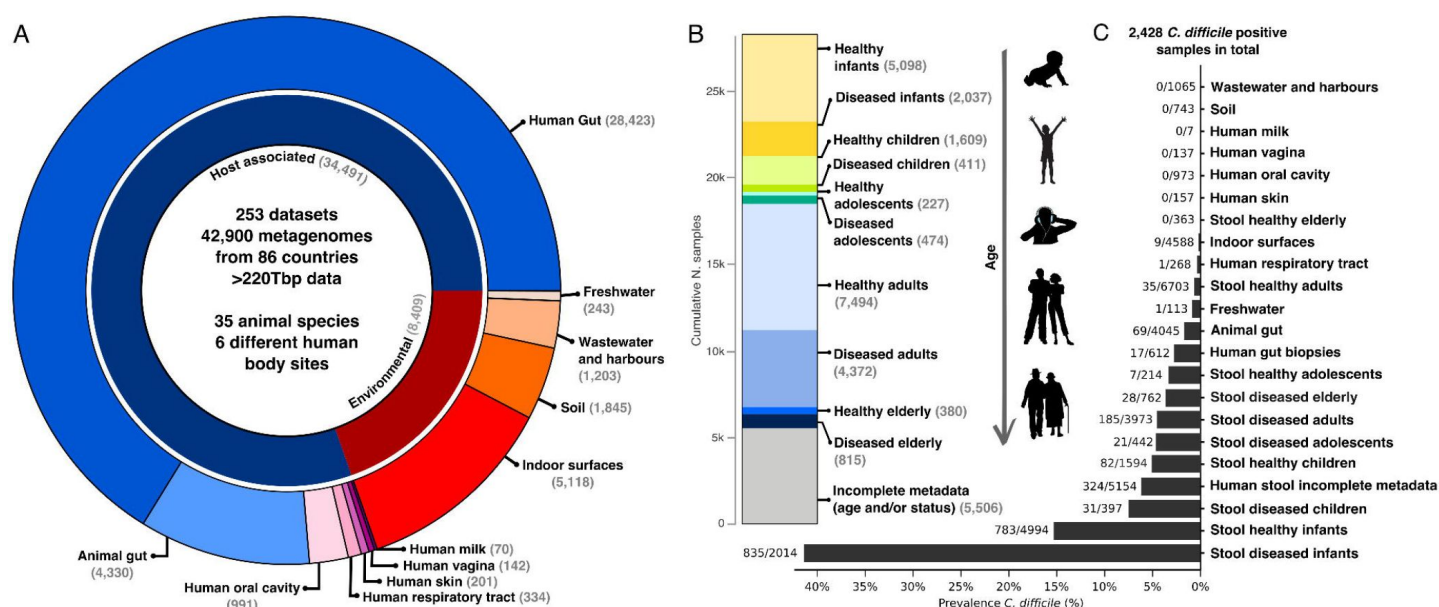


Figure 3.

C. difficile tracking in a global dataset collection

(A) Overview of the global dataset collection composed of 42,900 samples, from 253 publicly available studies, including host-associated and environmental samples (internal ring), and their subcategories (external ring). The human gut was the most represented environment (66.3%, $n=28,423$) in our collection. (B) Stratification of the human gut samples, divided by health status and age group (see Methods for detailed age group description, numbers refer to the number of samples prior to read filtering). (C) Prevalence of *C. difficile* positive samples per category. Total values refer to the number of samples after the initial read filtering and before time series dereplication (see Methods). From this point onwards, results refer to this extended collection of 42,900 samples.

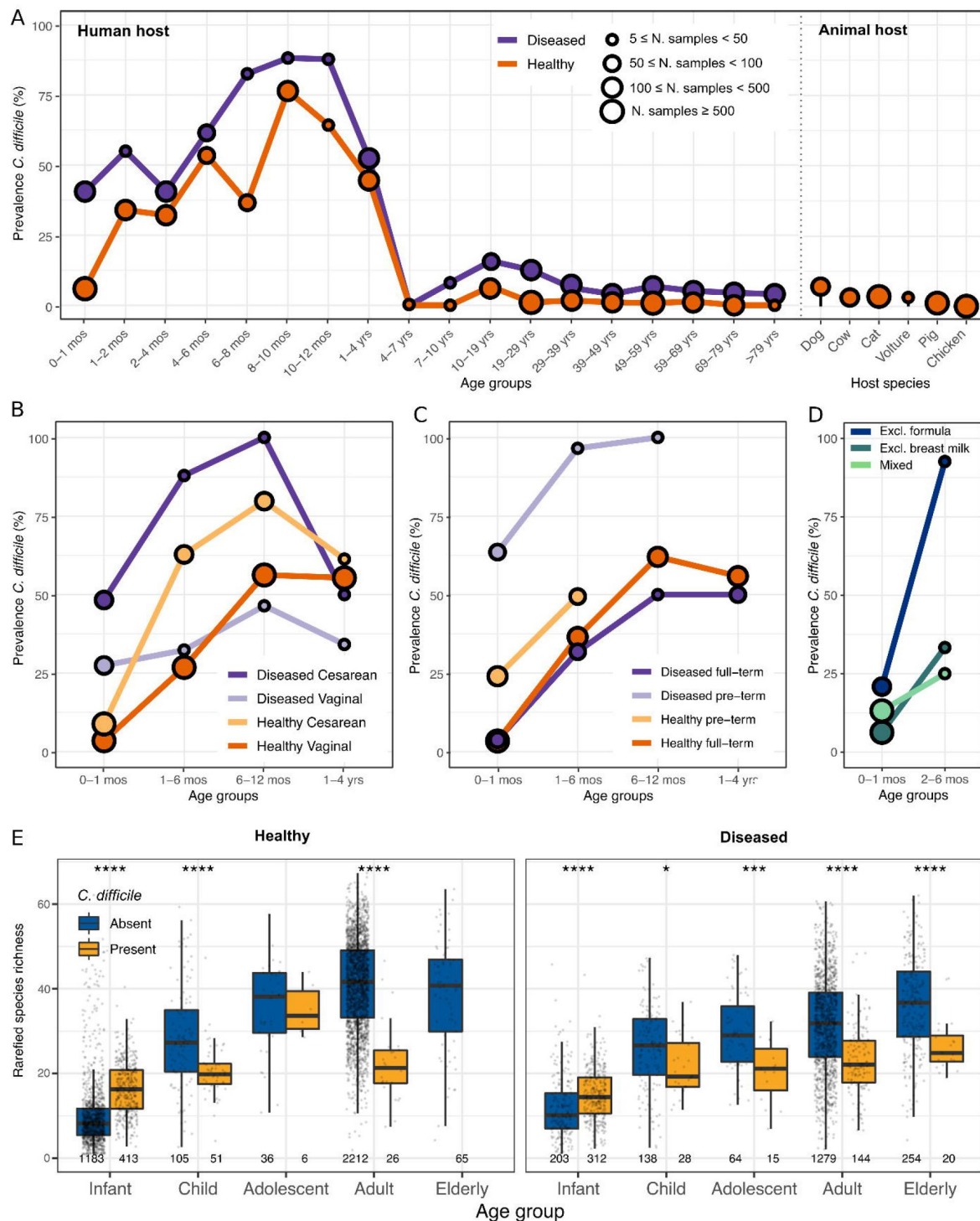


Figure 4.

C. difficile prevalence and associated microbial diversity

(A) *C. difficile* prevalence in stool samples in healthy and diseased human (left) and animal (right) hosts over lifetime. *C. difficile* prevalence divided by (B) health status and delivery mode, (C) health status and prematurity and (D) feeding mode within the first semester of life. For animal samples, only species with at least ten total samples are shown. (E) Species richness of human stool samples, with and without *C. difficile*, across age groups in healthy (left) and diseased subjects (right). See Methods for details on disease and age group definitions. For samples belonging to time series, only the representative samples are included in the estimations (see Methods). Mean comparison p-values calculated using t-test, * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, **** for $P \leq 0.0001$, empty when non significant.

***C. difficile* is enriched upon antibiotic treatment**

Diseased or antibiotics-treated subjects exhibited consistently increased *C. difficile* carriage rates across all age groups (Figure 4A). Prevalence was highest among infants taking antibiotics (81.1%, relative abundance 2.2×10^{-2}) and infants suffering from cystic fibrosis (78.7%, 4.0×10^{-3}) or neonatal sepsis/necrotizing enterocolitis (53.7%, 3.2×10^{-2} ; Supplementary Figure 5B). An elevated *C. difficile* asymptomatic carriage rate in cystic fibrosis patients is well documented (Deane et al. 2021) and may be the result of long-term antibiotics usage and frequent exposure to the hospital environment (Bauer et al. 2014). Among adolescents, adults and elderly, *C. difficile* prevalence was highest in patients suffering from CDI (31%, see above) or undefined diarrhoea (25.9%), followed by subjects taking antibiotics (on average 25% across age groups, relative abundance 4.6×10^{-3}). Inflammatory bowel diseases, liver diseases, diabetes, and various forms of cancer were likewise associated with an increased *C. difficile* carriage relative to the healthy background population (Supplementary Figure 5B).

***C. difficile* is a frequent commensal in the healthy infant gut**

C. difficile was among the most prevalent species in infants during the first year of life. Carriage was increased in infants born via C-section (28.7%; logistic regression, adjusted ANOVA $p = 3.0 \times 10^{-3}$, $\log OR = 2.0 \times 10^{-5}$) compared to age- and health status-matched vaginally born infants (19.6%) (Figure 4B and Supplementary Figure 6). Likewise, prevalence was moderately increased among pre-term (39.14%; $p = 1.7 \times 10^{-1}$, $\log OR = 1.1 \times 10^{-5}$) compared to full-term healthy infants (23.8%) (Figure 4C and Supplementary Figure 6). Exclusively formula-fed infants were more frequently colonised by *C. difficile* (43.2%; $p = 5.6 \times 10^{-6}$, $\log OR = 3.1 \times 10^{-5}$), compared to partially or exclusively breastfed infants during the first 6 months of life (14.2% and 8.6% respectively) (Figure 4D), in line with previous reports (McDonald et al. 2018; Drall et al. 2019).

C. difficile was not maternally acquired at birth: across 829 mother-infant pairs included in the dataset, we found no single case of vertical transmission of *C. difficile* from any of the maternal body sites investigated (gut, vagina and oral cavity). Rather, *C. difficile* was likely sourced from the social environment (i.e. other infants, family members or pets) during early life, in line with previous observations on *Clostridia* (Korpela et al. 2018). Leveraging longitudinal data, we observed that the first appearance of *C. difficile* was not arbitrarily distributed over the first year of life, but instead concentrated in two defined time windows, between the 2nd-4th and 8th-10th months of age (Supplementary Figure 7). The first interval coincided with the increased strain influx from environmental sources to the infant gut reported in a previous study (Ferretti et al. 2018), probably due to the start of mouthing. The second interval, dominated by samples from the UK, overlaps with the paid maternity leave length in that country (Thevenon et al. 2016), suggesting that it could be due to the start of day care. C-section and premature birth were associated with increased rate of *C. difficile* appearance, compared to vaginally born and full-term infants (Supplementary Figure 7). Together these results suggest that *C. difficile* is acquired from the environment, and that increased exposure to novel environmental sources increases the chances of *C. difficile* acquisition. In addition, during the first year of life, the gut microbiome of healthy infants colonised with *C. difficile* was significantly more similar, in terms of composition, to that of their mothers (Supplementary Figure 8).

These vast differences in *C. difficile* carriage across age did not translate into differential toxin burdens. *C. difficile* toxin genes were detectable in metagenomes of healthy *C. difficile*-positive infants (22.4%) and adults (18.6%) at similar rates (Supplementary Figure 9).

***C. difficile* occurs in distinct biotic and physiological contexts in infants and adults**

The *C. difficile*-associated microbiome varied considerably across age ranges. *C. difficile* carriage was associated with a significantly decreased overall species richness in subjects above 1 year of age, independently of the health status (Figure 4E, see Methods for the definition of diseased). However, in healthy infants we observed the opposite trend: species richness was higher in subjects carrying *C. difficile* (Figure 4E and Supplementary Figure 10A), irrespective of delivery mode, gestational age or geography (Supplementary Figure 10B and 10C). Species richness was also elevated in *C. difficile*-positive diseased infants, albeit to a lesser degree. In line with these trends, *C. difficile* carriage was associated with higher community evenness (lower dominance of individual species) in infants, but a more uneven community structure in adults (Supplementary Figure 10D).

To further characterise how the biotic context of *C. difficile* in the gut changes over lifetime, we studied its co-occurrence with other intestinal species, stratified by age group and health status. We broadly distinguished two groups of species based on their co-occurrence relationships with *C. difficile* (Figure 5), independently of the presence of its toxin genes. The first set of species (“Group 1”) consistently co-occurred with *C. difficile* across all age groups. Part of this group are *Clostridium paraputrificum* and *Clostridium neonatale* (Supplementary Figure 11A), in line with previous reports (Daquigan et al., 2017), an association that was conserved also in animal hosts (Supplementary Figure 11B). Group 1 was enriched in facultative anaerobic or aerobic species and comprised several known AAD enteropathogens, such as *K. oxytoca*, *C. perfringens*, *C. amalonaticus* or *C. innocuum*. A second set of species (“Group 2”) co-occurred with *C. difficile* exclusively in healthy infants, but showed strong avoidance patterns in later life stages and, to a lesser extent, in diseased infants. Group 2 was characterised by obligate anaerobic species typical of the healthy adult gut microbiome and included several common butyrate producers (such as *Roseburia* sp., *Faecalibacterium prausnitzii* or *Anaerostipes hadrus*) (Figure 5 and Supplementary Figure 12).

To investigate whether healthy infants carried specific *C. difficile* strains, we performed an analysis of metagenomic single nucleotide variants (SNVs). Despite the distinct biotic context associated with *C. difficile* presence in healthy infants at the species level, *C. difficile* strain populations found in this group did not cluster separately from those identified in other age groups and health status, with consistent patterns of toxin-gene carriage (Supplementary Figure 13). Therefore, the general asymptomatic nature of *C. difficile* carriage in healthy infants is likely determined by its surrounding microbial community and host age, rather than specific strain groups.

Discussion

C. difficile is predominantly studied in context of its pathogenicity, yet its role in the human gut is arguably much more nuanced. Our data suggests that *C. difficile* may be both overestimated as a pathogen among the elderly and underestimated as a commensal among infants, and that its ambivalent role in the gut is malleable throughout the host’s lifespan depending on external triggers (like antibiotics treatment) and biotic context.

We detected fecal *C. difficile* in just 30% of cases diagnosed with CDI, consistent with previous smaller-scale reports based on 16S rRNA (Daquigan et al. 2017; Seekatz et al. 2016) and WGS metagenomics (Vincent et al. 2016; Zhou et al. 2016), even when combined with laboratory-based tests (Seekatz et al. 2016). This surprisingly low prevalence was neither due to metagenomic detection limits, as sequencing depth was not associated with *C. difficile* prevalence, nor could it be explained by (temporary) reduction of *C. difficile* abundance below detection limits

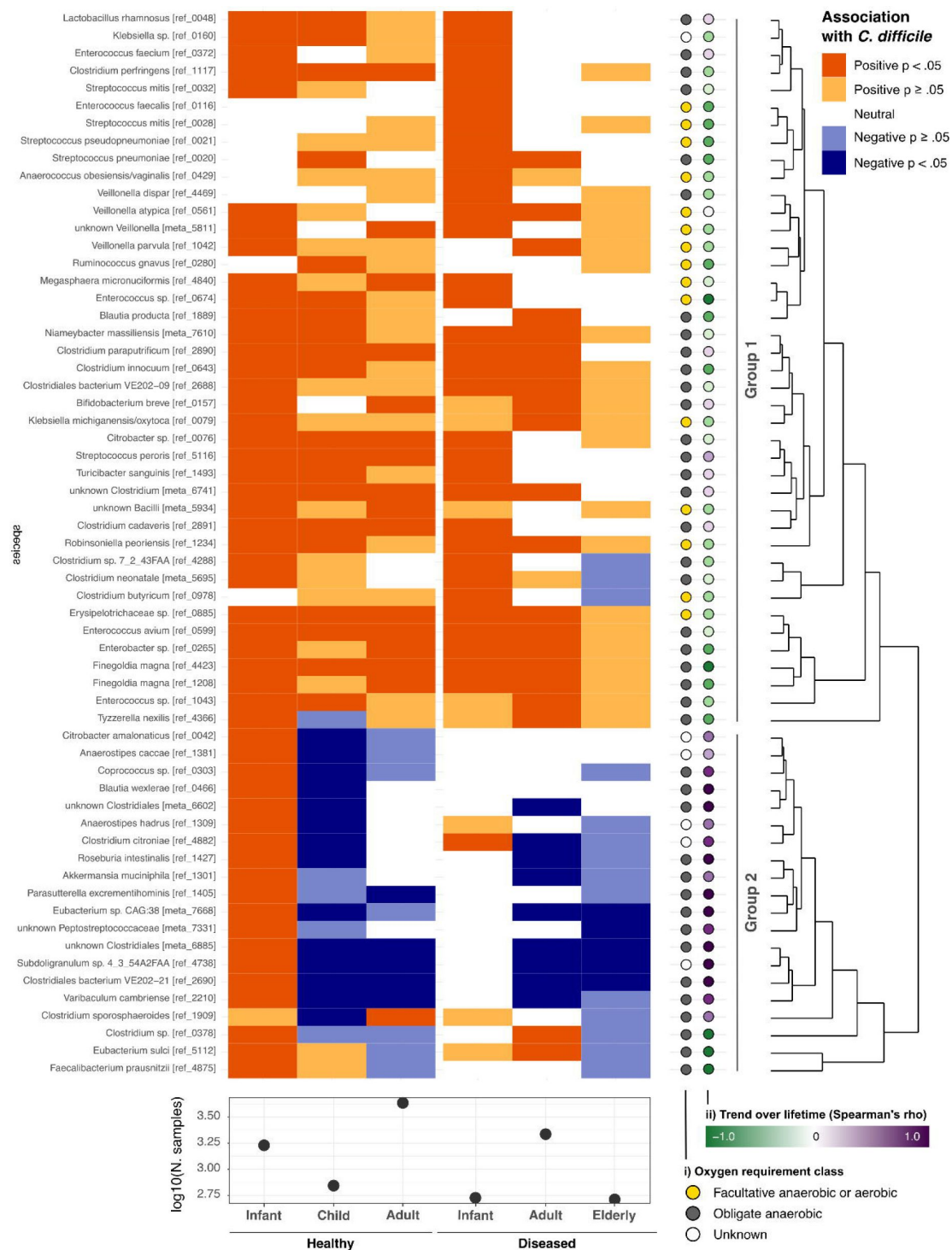


Figure 5.

Biotic context associated with *C. difficile*

Species associated with *C. difficile*, divided by age group and health status: significant positive associations are shown in dark orange, significant negative ones in dark blue (p -values adjusted using BH). Lighter shades indicate non significant associations. Only species significantly associated with *C. difficile* in at least one age group are shown. Number of samples per each category shown in the lower part. See Methods for details on age group definitions. Species group separation was not affected by presence of *C. difficile* toxin genes (data not shown). On the right, per species annotation of their i) oxygen requirement (see also Supplementary Figure 12 for the differential enrichment across the two groups) and ii) trend over lifetime. Positive Spearman's rho values indicate species more commonly found in the gut microbiome of healthy adults, negative values the opposite trend.

by antibiotics, as stricter inclusion criteria for samples from active and symptomatic CDI cases did not change the observation. A more likely explanation for the under-detection of *C. difficile* among CDI patients is a lack of diagnostic specificity. Challenges in CDI diagnosis are further complicated by the vast array of protocols and diagnostic algorithms available in clinical practice (Jo et al. 2022). Diarrhoeal symptoms, often the initial diagnostic prompt, are not specific to CDI (Bartlett and Gerding 2008; Polage et al. 2012; Jackson et al. 2016; Reich et al. 2019) and pseudomembranous colitis, long thought to be characteristic (and often eponymously referred to as “*C. difficile* colitis”), may indeed be caused by other enteropathogens, such as *C. innocuum* or *E. coli* (Gateau et al. 2018; Humphries et al. 2013; Burnham and Carroll 2013; Lee et al. 2021; McDonald et al. 2018; Crobach et al. 2016; Tenover et al. 2011; Rodriguez et al. 2016; Martínez-Meléndez et al. 2017; Litvin et al. 2009). Additional challenges include the large variability in sensitivity and specificity in diagnostic laboratory tests (Gateau et al. 2018; Humphries et al. 2013; Burnham and Carroll 2013; Lee et al. 2021; McDonald et al. 2018; Crobach et al. 2016; Tenover et al. 2011; Rodriguez et al. 2016; Martínez-Meléndez et al. 2017; Litvin et al. 2009; BBL™ Clostridium difficile Selective ...), contradicting results between cultivation and toxin detection assays (Polage et al. 2015) and an excessive reliance on single molecular tests (Polage et al. 2015). Recommended guidelines discourage stand-alone tests for CDI diagnosis and recommend a two-step procedure (Crobach et al. 2016; Gateau et al. 2018), but due to costs, test availability, capacity and turnaround times, diagnostic algorithms vary between hospitals and often deviate from recommended procedures (Tenover et al. 2011). Despite the recent improvements in the CDI diagnosis guidelines awareness, low compliance is still reported (Viprey et al. 2023). In our meta-analysis, almost two thirds (62.5%) of studies deviated from recommended best practices for CDI diagnostic protocols.

Moreover, *C. difficile*'s role in causing antibiotics-associated diarrhoea (AAD) or pseudomembranous colitis (PMC) may be overstated, as simultaneous colonisation of multiple enteropathogens is very common among CDI patients (Hensgens et al. 2014; Spina et al. 2015). Enteropathogens like *C. perfringens*, *C. innocuum*, *S. aureus*, *E. coli* or *S. flexneri*, among others, are known to induce AAD or PMC (Chia et al. 2018; Kiu and Hall 2018; Zollner-Schwetz et al. 2008; Larcombe et al. 2018; Högenauer et al. 1998; Larcombe et al. 2016) and were prevalent and enriched among CDI samples, as were their characteristic toxin genes – in particular also in samples where *C. difficile* was not detectable. Indeed, while *C. difficile* was the most individually characteristic species of CDI, its presence was not an essential input to microbiome-based classification models accurately predicting CDI-diagnosed samples. Taken together, this suggests that *C. difficile* is often not the most parsimonious explanation for clinical manifestations of CDI that may instead be driven by alternative enteropathogens, implying mis- or over-diagnosis in practice and an overestimation of the real burden of *C. difficile* relative to other AAD species. Differential diagnosis against multiple enteropathogens may therefore stratify patients with CDI-like symptoms, towards adapted therapeutic interventions.

Widespread asymptomatic carriage in healthy individuals is a further indicator that *C. difficile* plays a complex role in the gut, beyond pathogenicity. We tracked *C. difficile* across an extensive set of 42,900 metagenomes sampled from multiple habitats and hosts, including $n=28,423$ (66.3%) human gut samples which covered a wide range of subject ages, disease states and geographic origins. *C. difficile* was particularly common among infants, peaking at >75% prevalence at 8-10 months of age. On average *C. difficile* was present in 25% of healthy infants aged 1 year or younger, in line with previous studies (Lees et al. 2016; Jangi and Thomas Lamont 2010).

In healthy infants, the occurrence of *C. difficile* coincided with an increased resemblance to the maternal gut microbial community, general increase in microbiome richness and with several obligate anaerobic commensals, butyrate producers and a generally more adult-like microbiota, indicating that *C. difficile* may be a transient hallmark of healthy gut microbiome maturation (Bäckhed et al. 2015; Yassour et al. 2016; Ferretti et al. 2018; Chu et al. 2017; Rodríguez et al. 2015). In contrast, *C. difficile* presence in adulthood was associated with decreased

microbiome richness and an enrichment in facultative and obligate aerobes, generally associated with dysbiosis (Rigottier-Gois 2013 [↗](#); Rivera-Chávez et al. 2016 [↗](#)). We found multiple lines of evidence suggesting that the transition between these apparent dichotomic states (commensal in infancy, pathobiont in adulthood), begins at around 10 (± 2) months of age. We hypothesise that *C. difficile* could be an indicator species of a larger-scale and fundamental change taking place in the microbial ecology of the gut.

C. difficile toxin genes were likewise prevalent among infants, in line with previous reports (Rousseau et al. 2012 [↗](#); Kubota et al. 2016 [↗](#); Mani et al. 2023), indicating asymptomatic carriage of toxigenic strains. Indeed, *C. difficile* testing is generally discouraged by paediatric guidelines even in presence of diarrhoeal symptoms (McDonald et al. 2018 [↗](#); Schutze et al. 2013) and toxin concentration may not be a reliable proxy for inferring disease and its severity (Davis et al. 2016 [↗](#); Kubota et al. 2016 [↗](#)). The apparent protection from CDI symptoms has been ascribed to a lack of toxin receptors in the immature infant gut, although this early hypothesis (Chang et al. 1986 [↗](#)) has since been called into question (Keel and Songer 2007 [↗](#); Eglow et al. 1992 [↗](#)). The biotic context of *C. difficile* suggests additional or alternative explanations. We observed a significant enrichment in butyrate producers co-occurring with *C. difficile* in healthy infants, but not later in life nor in diseased infants. High levels of butyrate have been linked to inflammation inhibition, regulation of cell-to-cell tight junctions and increased mucin production (and therefore mucosal layer thickness and integrity) (Cornick et al. 2015 [↗](#); Willemsen et al. 2003 [↗](#)), all considered protective against *C. difficile* toxins (Pruitt and Lacy 2012 [↗](#)).

Our analyses are descriptive and arguably inherently limited by the underlying metagenomic data, such as DNA extraction bias against endospores (Felczykowska et al. 2015 [↗](#)) or general detection limits due to finite sequencing depth. Moreover, although some longitudinal data was available, time series were too sparse to infer causal relationships, such as e.g. the possible role of *C. difficile* in a partially deterministic community succession in the infant gut. To our knowledge, ours is the largest single-species metagenomic survey to date, demonstrating the utility of metagenomic meta-studies as a nuanced approach to gut microbial ecology, beyond *C. difficile*. Overall, our results suggest that the ability of *C. difficile* to induce disease may be context-dependent and multifactorial (i.e. influenced by the rest of the microbiome composition, toxin presence and host age). Further study of *C. difficile*'s association with health outcomes should therefore adopt a more holistic view of the ambivalent role of this species over the human lifespan.

Methods

Human study

We compiled a metagenomic dataset of both public (see below) and 447 newly generated samples. The novel cohort is part of the citizen science-based my.microbes project (<http://my.microbes.eu/> [↗](#)) (Voigt et al. 2015 [↗](#)), conducted in accordance with the WMA Declaration of Helsinki and with the approval of the EMBL Bioethics Internal Advisory Board. Sample collection, processing and sequencing were performed as previously described in (Voigt et al. 2015 [↗](#); Costea et al. 2017 [↗](#); Hildebrand et al. 2021 [↗](#)). This cohort includes human fecal (n=437) and oral (n=10) samples, from a total of 142 subjects (aged 0-97 years). Self-reported subject metadata are included in **Supplementary Table 4**.

Public data overview

CDI meta-analysis

In the CDI meta-analysis (see **Supplementary Table 1**), out of 534 samples in total (from as many subjects, with an average age of 60 ± 19.9 years), 234 were identified as CDI, 114 as diseased controls and 186 as healthy controls. Samples from subjects diagnosed with CDI in the original study population were classified as CDI in our meta-analysis. Samples were identified as controls in any of the following cases: (i) H-Ctr: healthy subject, reported as “control” in the metadata of the original study population; or (ii) D-ctr: diseased subject, that either had diarrhoeal symptoms but negative CDI diagnostic outcome, or was diarrhoea-free but diagnosed for a disease other than CDI. The majority of the samples in the latter group (D-Ctr) belong to hospitalised subjects. Study-specific CDI diagnostic criteria are indicated, when published in the original studies, in Supplementary Figure 1.

Global meta-analysis

We analysed a total of 42,900 publicly available metagenomic samples (collectively >220Tbp) from 253 different studies, including the novel cohort of 447 samples from the citizen science project MyMicrobes described above (see **Supplementary Table 2**). The collection includes samples from 86 countries, 35 animal species and 6 different human body sites: gastrointestinal tract (stools, rectal swabs and biopsies), vagina, skin, oral cavity, respiratory tract and human milk. The selected environmental samples come from potentially faecally contaminated habitats, such as wastewater, freshwater, indoor surfaces, soil and polluted harbor marine waters.

Public data download

Metagenomes publicly available on the 22nd of April 2021 were downloaded using fetch-data (Coelho et al. 2021 [DOI](#)). The collection process aimed at leveraging the largest possible set of samples and to maximise the number of covered environments, with the only inclusion criteria being samples sequenced with whole genome sequencing on Illumina platforms. No minimum number of samples per study or minimum sequencing depth threshold was applied at this stage.

Metadata curation

Manually curated metadata for each dataset include: health status, age, geography (country and continent name), westernised lifestyle or not, diagnosis for *C. difficile* infection (CDI), use of antibiotics, delivery mode, premature birth (pre-term or full-term), and sex. Subjects were categorised by age group, defined as follows: $0 \leq \text{infant} \leq 12$ months, $1 \text{ year} < \text{child} \leq 10$ years, $10 \text{ years} < \text{adolescent} \leq 18$ years, $18 \text{ years} < \text{adult} \leq 65$ years, and elderly > 65 years of age. If a specific age was not available, a range of age was provided in alternative. Given the broad range of perturbations in the gut microbiome associated with antibiotics intake, we ad-hoc defined “diseased” samples as any sample taken from a subject with any medically diagnosed disease or syndrome and/or intake of one or more antibiotics at the time of the sampling.

Filtering on read count

Two consecutive read count filtering steps were performed on all samples: (i) samples with zero reads mapping to mOTU marker genes (Milanese et al. 2019 [DOI](#)) were discarded ($n=1,091$) and (ii) samples with less than 59 reads mapping to mOTU marker genes, corresponding to 95%ile calculated on the remaining samples, were discarded ($n=2,075$). An additional third read count filtering was performed only on human gut stool samples, corresponding to 99%ile, removing samples with less than 100 reads ($n=132$ samples). The resulting filtered dataset included 39,338 samples, of which 26,610 were human fecal metagenomes.

Identification of timeseries-representative samples

Out of 26,610 samples, 24,690 had subject-level metadata and were associated with 12,352 unique subjects. For 3,573 subjects, multiple timepoints were available. The mean number of timepoints per subject was 2 (3.2 for infants, 1.7 for children, 3.5 for adolescents, 1.7 for adults and 1.3 for elderly), with a maximum of 205 time points per subject.

In order to avoid under- or over-estimating *C. difficile* prevalence, only one sample per time series was used in cross-sectional analyses. We distinguished three cases for each time series:

1. *C. difficile* observed in all samples. In this case, the sample with the highest *C. difficile* read count is selected as representative and the corresponding subject was considered *C. difficile* positive.
2. *C. difficile* observed in none of the samples. In this case, the sample of the first time point was selected as representative and the corresponding subject was considered *C. difficile* negative.
3. *C. difficile* observed in only some samples. In this case, the sample with the highest *C. difficile* read count was selected as representative and the corresponding subject was considered *C. difficile* positive.

1,920 samples had no subject identifier. In this case, we assumed one sample per subject. One representative sample for each time series was used for all downstream analyses, if not specified otherwise. The timeseries dereplication procedure described above has also been applied to the CDI study populations before downstream analysis.

Metagenomic data processing

Taxonomic profiling

Metagenomes were taxonomically profiled at the species level with mOTUs v2.0 (Milanese et al. 2019 [DOI](#)), requiring the confident detection of at least two taxonomic marker genes. All data analyses were conducted in the R Statistical Computing framework v3.5 or higher. Only *C. difficile* positive samples were considered for *C. difficile* relative abundance estimation.

Microbiome diversity

Local community diversities ('alpha' diversities) of human gut samples were estimated by iteratively rarefying taxonomic count tables to 100 marker gene-mapping reads and computing average Hill diversities at $q=0$ (species richness), $q=1$ (exponential Shannon entropy) and $q=2$ (inverse Simpson index), as well as evenness measures as ratios thereof. Unless otherwise stated, results in the main text refer to taxa richness. Differences in alpha diversity were tested using ANOVA followed by post hoc tests and Benjamini-Hochberg correction, as specified in the main text.

Prevalence and abundance estimations of *C. difficile*

Prevalence estimations of *C. difficile* over life time were based on human gut samples (stools, rectal swabs and biopsies) where the precise subject age in months was available (samples with missing or too broad or age ranges were discarded). Both *C. difficile* mOTUs2.0 ("ref_mOTU_0051" and "ref_mOTU_0052") were considered in downstream analysis. Abundance estimations included only *C. difficile* human stool samples with precise age metadata.

Species co-occurrence analysis

Human gut stool samples, with at least 100 reads per sample and with known age group and health status were considered for this analysis. One representative sample per time series was considered. Fisher's exact test, followed by Benjamini-Hochberg correction, were applied to identify co-occurring species. Meaningful positive association was identified for species with adjusted p-value <0.05 and logarithm of the odds ratio >1, while meaningful negative association was identified for species with adjusted p-value <0.05 and logarithm of the odds ratio < -1.

Machine learning modelling

L1-regularised LASSO logistic regression models to predict CDI status were built using the SIAMCAT R package (Wirbel et al. 2021 [\[1\]](#)) with 10-fold cross-validation. For this analysis, we focused on the subset of 10 CDI or diarrhoea-associated datasets (**Supplementary Table 1**) and then trained two different sets of models: one set of models to distinguish CDI samples and samples from healthy controls (excluding controls from diseased subjects) and another set of models to distinguish CDI samples and any type of control samples. In order to minimise overfitting and to counter batch effects (Wirbel et al. 2021 [\[1\]](#)), we pooled datasets across studies in a leave-one-study-out approach. In short, all except one study were jointly processed to train a LASSO model that was then used to predict the left-out study. Additionally, to check if the microbial signature for CDI was independent of *C. difficile*, we trained another set of models with the same cross-validation splits but excluded *C. difficile* from the feature table. Feature weights were extracted from the models, normalised by the absolute sum of feature weights, and averaged across cross-validation folds. For the heatmap in **Figure 2** [\[2\]](#), all microbial species that were assigned non-zero weights in at least 80% of cross-validation folds were included.

Linear mixed-effect model

To test for differential abundance of microbial species between CDI and non-CDI samples while taking into account possible confounding factors, we employed linear mixed-effect models as implemented in the lmerTest package (Kuznetsova et al. 2017 [\[3\]](#)). After filtering for prevalence (prevalence of at least 5% in three or more studies), we tested the log-transformed abundance of each microbial species using a linear mixed-effect model with “CDI status” as fixed and “Study” and “Age group” as random effects. Effect size and p-values were extracted from the model and p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. In addition, we used a linear mixed-effect model to test for a relationship between sequencing depth and the ability to detect *C. difficile* in CDI study populations as well as on the wider global set of samples (considering only dereplicated samples from time series). In this analysis, study was considered as a random effect.

Pathogenic toxin genes detection in metagenomes

Identification of toxin-coding genes and their assignment was performed using the VFDB database (Liu et al. 2019 [\[4\]](#)), downloaded in March 2021. In our analysis we did not investigate the presence of the *C. difficile* binary toxin gene, since this toxin alone has not been associated with disease severity (Goldenberg and French 2011 [\[5\]](#)). Mapping against VFDB was performed via BWA (Li and Durbin 2009 [\[6\]](#)) and filtering via NGLess (Coelho et al. 2019 [\[7\]](#)), using 99% as minimal alignment percentage identity threshold and 75bp as minimal read length match. Toxin gene profiles were computed with gffquant (https://github.com/grp-bork/gff_quantifier [\[8\]](#), version 1.2.3) using the 1overN count model. Toxin genes included in VFDB refer to the following strains: *C. difficile* 630, *E. coli* CFT073, O44:H18 042 and O157:H7 str. EDL933, *S. flexneri* 2a str.301, *P. aeruginosa* PAO1, *C. perfringens* str.13 and SM101, *S. aureus* RN4220, subsp. aureus MW2 and N315. The comprehensive list of virulence-related genes for each of these strains is available on the VFDB website (<http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi> [\[9\]](#)).

Subspecies analysis

C. difficile subspecies detection was performed on all *C. difficile* positive samples with MetaSNV2 (Van Rossum et al. 2021 [\[1\]](#)). Reads from 2,428 samples were mapped against proGenomes v1 (Mende et al. 2017 [\[2\]](#)) species representatives genomes for 3 species in the *Clostridium* genus: i) specI_v2_0051 (NCBI taxonomy ID 272563, PRJNA78) *Clostridioides difficile*, ii) specI_v2_0052 (NCBI taxonomy ID 1151292, PRJNA85757) *Clostridioides difficile* and iii) specI_v2_1125 (NCBI taxonomy ID 1408823, PRJNA223331) *Clostridioides mangenotii*. By mapping to multiple genomes and only using the uniquely mapped reads, we focused on the species-specific core genomic regions. Mappings that had at least 97% identity and a match length of at least 45bp were kept. Mapping and filtering was performed using BWA (Li and Durbin 2009 [\[3\]](#)) and NGLess (Coelho et al. 2019 [\[4\]](#)). Reads that mapped uniquely across the 3 reference genomes were used to call single nucleotide variants (SNVs) using metaSNV2 with default parameters. Of the initial 2383 samples, 197 passed the requirements for robust SNV calling (8%) and 179 could be used to detect subspecies presence. Substructure within the population was assessed in the resultant SNV profiles according to a previously reported approach (Li and Durbin 2009 [\[3\]](#); Costea et al. 2017 [\[5\]](#)). Briefly, dissimilarities between samples were calculated based on SNV abundance profiles and the resultant distance matrix was tested for clusters using the Prediction Strength algorithm (Yuan et al. 2020 [\[6\]](#)). Distance matrix is plotted using R and the pheatmap package with average clustering. Six samples with extreme dissimilarity to all other samples were removed from the distance matrix for illustrative purposes (SAMN08918181, SAMN09980608, SAMN10722477, SAMN13091313, SAMN13091317, SAMN13091322).

Building SNP-tree

SNPs obtained via metaSNV2 were arranged into an alignment PHY input format file suitable for IQ-TREE2 tree-builder (Minh et al. 2020 [\[7\]](#)). Sites which had multiple variants/alleles (in this case only SNPs) were replaced with alternative non-reference allele if they were supported by >50% of reads mapped against specI_v2_0051 (NCBI taxonomy ID 272563, PRJNA78), see ‘Subspecies analysis’ section for mapping details. This results in forced fixing of alleles across polymorphic sites within each sample. To include specI_v2_0052 (NCBI taxonomy ID 1151292, PRJNA85757) *Clostridioides difficile* and specI_v2_1125 (NCBI taxonomy ID 1408823, PRJNA223331) *Clostridioides mangenotii* into SNP input file for IQTREE2 we used Mauve (Darling et al., 2004 [\[8\]](#)) to 1) order contigs of both genomes against specI_v2_0051 genome, 2) align ordered contigs, 3) export SNP profile from Mauve and further merged alignment SNP profile on metaSNV2 output file using matching positions. Missing coverage was kept as gaps in the alignment and tree construction. Following command was used to build phylogenetic tree: `iqtree -s >INPUT snp file< -m GTR+ASC -B 1000`.

Availability of data and materials

This study did not generate new unique reagents. Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contacts, Sebastian Schmidt (sebastian.schmidt@embl.de) and Peer Bork (bork@embl.de). Novel raw metagenomic sequencing data have been uploaded to the European Nucleotide Archive under the accession number PRJEB50977. The accession numbers relative to the other public studies included in this meta-analysis are available in Supplementary Table 2. The filtered taxonomic profiles and associated metadata used for the analyses are available in Supplementary Tables 4-6. The analysis code and commands used in the project are deposited in the GitHub repository at https://github.com/grp-bork/cdifficile_Ferretti_2022 [\[9\]](#).

Acknowledgements

We are thankful to the members of the groups of Peer Bork, Georg Zeller, Thomas Dandekar (University of Würzburg) and Dr. Athanasios Typas (EMBL) for fruitful discussions. We are also grateful to Lia Oken for her continuous support and her availability for brainstorming sessions. We acknowledge funding from EMBL and the European Research Council (MicrobioS grant no. ERC-AdG-669830 to P.B.). This work comprises results from Pamela Ferretti's doctoral thesis.

Author contributions

Conceptualization, P.F., T.S.B.S. and P.B.; data collection, P.F., T.S.B.S., R.A., A. F., W.A. and M.K.; metadata curation, P.F., T.S.B.S., R.A., A.S., R.T., L.T, I.L. and M.K.; novel samples collection and processing, S. K., R.H. and A.T.; data analysis, P.F., J.W., O.M.M., T.VR., R.A., A. F., W.A., C.S. and T.S.B.S.; writing – original draft, P.F. and T.S.B.S.; writing – review & editing, P.F., J.W., O.M.M., T.VR, C.S., M.K, G.Z. T.S.B.S. and P.B.; supervision, T.S.B.S., G.Z. and P.B.; funding acquisition, P.B.

Declaration of interests

The authors have no conflicts of interest to declare.

Supplementary Figure Legends

Supplementary figure 1. (A) Prevalence of *C. difficile* and other antibiotic-associated diarrhoea (AAD) species, divided by study. For *C. difficile* only, the portion of samples with toxigenic *C. difficile* is shown (dotted segments). CDI diagnosis procedure is shown on top of each study. Each row represents the diagnostic algorithm (combination of tests) used to diagnose CDI. Tests performed are indicated with black dots (white if not). For example, in “Langdon et al. 2021 [DOI](#)”, CDI was diagnosed if symptoms were present and toxigenic culture for *C. difficile* was positive, or if symptoms were present and the enzymatic immunoassays for *C. difficile* was positive, or if pseudomembranous colitis was identified. Diagnostic protocols abbreviations: “EIA”: Enzymatic ImmunoAssays, “GDH”: glutamate dehydrogenase, “NAAT” nucleic acid amplification test (including PCR), “PMC”: pseudomembranous colitis. (B) Prevalence of the number of AAD species (*C. difficile* not included) identified in each CDI sample, divided by *C. difficile* positivity.

Supplementary Figure 2. Species-level composition for CDI, diseased control and healthy control samples as seen by mOTUs v2.0. Species with minimum relative abundance ≥ 0.01 and prevalence ≥ 0.1 are shown.

Supplementary figure 3. Species significantly enriched (in yellow) or depleted (in blue) in terms of relative abundance in CDI compared to diseased and healthy controls, as identified by linear mixed-effect model analysis. Species known to cause antibiotic-associated diarrhoea (AAD) are highlighted in red. *C. difficile* is not included in the analysis. Study and age group were considered as nuisance variables (and modelled as random effects). To be noted that *Enterobacteriaceae* sp. includes *Shigella flexneri* and *Escherichia coli*.

Supplementary figure 4. (A) To check the influence of sequencing depth on the detection of *C. difficile*, the marker gene coverage values across all samples were binned into ten quantiles. The proportion of *C. difficile*-positive samples is shown in the barplot. A possible association between coverage quantile and *C. difficile* detection was tested using a linear mixed-effect model with Study

as random effect (P-value=0.63). (B) Adjusted p-values for ANOVA analysis aiming to explain variance in *C. difficile* presence by the available metadata, either as single factors (two top rows) or by a combination of multiple factors (sequential, bottom rows, in the exact order shown along the X axis), in data sets that were broken down by age group (infants versus adults).

Supplementary figure 5. (A) Geographical distribution across continents and countries for *C. difficile* prevalence in the stools of healthy infants (left) and adults (right). Only countries with *C. difficile* prevalence above 1% are shown. Prevalence was calculated based on presence/absence of *C. difficile* from mOTUs v2.0³⁰ taxonomic profiles. (B) Prevalence of *C. difficile* across subjects diagnosed with specific diseases (left) and subjects taking antibiotics (right). Average *C. difficile* prevalence is shown with grey bars, age-group specific prevalence is highlighted with colored dots. Number of datasets used for prevalence estimations is reported in the lower part. Abbreviations: NEC: “Necrotizing enterocolitis”; ASCVD: “atherosclerotic cardiovascular disease”; T2D: “Type 2 diabetes”; CRC: “Colorectal cancer”; ADA: “advanced adenoma”; NAA: “non-advanced adenoma”. See Methods for details on age group definitions.

Supplementary figure 6. *C. difficile* prevalence in infants and children, divided by possible combinations of health status, prematurity and delivery mode.

Supplementary figure 7. Time of the first appearance of *C. difficile* in infants and children timeseries. The grey line indicates the total number of samples per age interval, considering all samples independently of health status, delivery mode, gestational age and diet.

Supplementary figure 8. Community similarity (Bray-Curtis index) of healthy infant-mother pairs in presence or absence of *C. difficile* in stools across the first four years of life. Only full-term infant samples were included. Mean comparison p-values calculated using t-test, * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, **** for $P \leq 0.0001$, empty when non significant.

Supplementary figure 9. Prevalence of *C. difficile* toxin genes in the stools of healthy and diseased subjects over lifetime, based on the detection of either one or both of *C. difficile* Toxin A (*tcdA*) or Toxin B genes (*tcdB*) (see Methods for details). Prevalence is calculated on the number of *C. difficile* positive samples. *C. difficile* toxin genes were exclusively found among human stool samples. See Methods for details on age group definitions.

Supplementary figure 10. Alpha diversity calculations in *C. difficile* positive and negative samples. (A) Species richness in healthy infants with detailed age resolution. (B-C) species richness in infant samples divided by delivery mode, gestational age and health status. (D) Species evenness (as the inverse Simpson index divided by Richness) across age groups and health status. Species richness for continents with at least five *C. difficile* positive samples are shown. The number of samples per group is shown under each boxplot. Mean comparison p-values calculated using t-test, * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, **** for $P \leq 0.0001$, empty when non significant. See Methods for details on age group definitions.

Supplementary figure 11. Species with co-occurrence ($\log OR > 0$) or co-exclusion ($\log OR < 0$) pattern with *C. difficile* across gut stool samples from humans (A) and animals (B). Labelled species are highlighted in aquamarine and violet, for humans and animals respectively. *Clostridium paraputrificum* and *Clostridium neonatale* were among the very few species significantly associated with *C. difficile* in both human and animal stools.

Supplementary Figure 12. Spearman’s Rho values for the species listed in [Figure 5](#), across the two groups. Mean comparison p-values calculated using t-test, * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, **** for $P \leq 0.0001$, empty when non significant.

Supplementary Figure 13. SNV similarity across *C. difficile* positive samples from our global metagenomic survey. Tree rooted with *Clostridioides mangenotii* as outgroup. Missing metadata are shown in white. All metagenomes have ANI similarity >95%. See Methods for detailed analysis description.

Supplemental Table Legends

Supplementary Table 1. Overview of the public metagenomic CDI study populations included in the CDI meta-analysis, divided by sample group.

Supplementary Table 2. Overview of the public metagenomic studies included in the global meta-analysis.

Supplementary Table 3. Mean *C. difficile* prevalence and relative abundance estimation across age groups and health status.

Supplementary Tables 4. Samples metadata before read filtering and timeseries dereplication steps.

Supplementary Table 5. Samples taxonomic profiles, as provided by mOTUs v2.0.

Relevance of this study

To our knowledge, this is the first large-scale metagenomic meta-analysis on *Clostridioides difficile*. We investigated *C. difficile* prevalence in 42,900 metagenomic samples from 86 countries and used machine learning to identify a global CDI-specific microbial signature. Our findings suggest that *C. difficile* may be overdiagnosed in the clinical context, with other enteropathogens able to induce CDI-like symptomatology, erroneously contributing to the *C. difficile* burden estimation. Disentangling the real burden of *C. difficile* from other pathogens is a fundamental step towards the development of adapted therapeutic interventions. Furthermore, this study provides new insights into the gut microbiome composition in presence of *C. difficile*, not only in CDI patients, but also across other diseases and in healthy subjects. In particular, we found that *C. difficile* colonisation was associated with multiple indicators of a healthy gut microbiome development in infants, an age group in which *C. difficile* has been so far understudied. This study also represents a blueprint for future large-scale species-specific metagenomic meta-analyses.

References

- Bäckhed F., Roswall J., Peng Y., Feng Q., Jia H., Kovatcheva-Datchary P., Li Y., Xia Y., Xie H., Zhong H., Khan M. T., Zhang J., Li J., Xiao L., Al-Aama J., Zhang D., Lee Y. S., Kotowska D., Colding C., Wang J. (2015) **Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life** *Cell Host & Microbe* **17**
- Bartlett J. G., Gerding D. N. (2008) **Clinical recognition and diagnosis of Clostridium difficile infection** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **46**:S12–S18
- Bartlett J. G., Moon N., Chang T. W., Taylor N., Onderdonk A. B. (1978) **Role of Clostridium difficile in antibiotic-associated pseudomembranous colitis** *Gastroenterology* **75**:778–782
- Bauer M. P., Farid A., Bakker M., Hoek R. A. S., Kuijper E. J., van Dissel J. T. (2014) **Patients with cystic fibrosis have a high carriage rate of non-toxigenic Clostridium difficile** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **20**:O446–O449
- (1970) **BBLTM Clostridium difficile Selective Agar. (n.d.). Retrieved October 21, 2021, from <https://www.bd.com/resource.aspx?IDX=20564>**
- Burnham C.-A. D., Carroll K. C. (2013) **Diagnosis of Clostridium difficile infection: an ongoing conundrum for clinicians and for clinical laboratories** *Clinical Microbiology Reviews* **26**:604–630
- CDC (2019) **CDC. (2019). Antibiotic resistance threats in the United States, 2019. 10.15620/cdc:82532 Antibiotic resistance threats in the United States 2019 <https://doi.org/10.15620/cdc:82532>**
- Chang T. W., Sullivan N. M., Wilkins T. D. (1986) **Insusceptibility of fetal intestinal mucosa and fetal cells to Clostridium difficile toxins** *Zhongguo Yao Li Xue Bao = Acta Pharmacologica Sinica* **7**:448–453
- Chia J.-H., Feng Y., Su L.-H., Wu T.-L., Chen C.-L., Liang Y.-H., Chiu C.-H. (2017) **Clostridium innocuum is a significant vancomycin-resistant pathogen for extraintestinal clostridial infection** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **23**:560–566
- Chia J.-H., Wu T.-S., Wu T.-L., Chen C.-L., Chuang C.-H., Su L.-H., Chang H.-J., Lu C.-C., Kuo A.-J., Lai H.-C., Chiu C.-H. (2018) **Clostridium innocuum is a vancomycin-resistant pathogen that may cause antibiotic-associated diarrhoea** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **24**:1195–1199
- Chu D. M., Ma J., Prince A. L., Antony K. M., Seferovic M. D., Aagaard K. M. (2017) **Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery** *Nature Medicine* **23**:314–326

- Coelho L. P., Alves R., Del Río Á. R., Myers P. N., Cantalapiedra C. P., Giner-Lamia J., Schmidt T. S., Mende D. R., Orakov A., Letunic I., Hildebrand F., Van Rossum T., Forslund S. K., Khedkar S., Maistrenko O. M., Pan S., Jia L., Ferretti P., Sunagawa S., Bork P. (2021) **Towards the biogeography of prokaryotic genes** *Nature* <https://doi.org/10.1038/s41586-021-04233-4>
- Coelho L. P., Alves R., Monteiro P., Huerta-Cepas J., Freitas A. T., Bork P. (2019) **NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language** *Microbiome* **7**
- Cornick S., Tawiah A., Chadee K. (2015) **Roles and regulation of the mucus barrier in the gut** *Tissue Barriers* **3**
- Costea P. I., Coelho L. P., Sunagawa S., Munch R., Huerta-Cepas J., Forslund K., Hildebrand F., Kushugulova A., Zeller G., Bork P. (2017) **Subspecies in the global human gut microbiome** *Molecular Systems Biology* **13**
- Crobach M. J. T., Planche T., Eckert C., Barbut F., Terveer E. M., Dekkers O. M., Wilcox M. H., Kuijper E. J. (2016) **European Society of Clinical Microbiology and Infectious Diseases: update of the diagnostic guidance document for Clostridium difficile infection** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **22**:S63–S81
- Crobach M. J. T., Vernon J. J., Loo V. G., Kong L. Y., Péchiné S., Wilcox M. H., Kuijper E. J. (2018) **Understanding Clostridium difficile Colonization** *Clinical Microbiology Reviews* **31** <https://doi.org/10.1128/CMR.00021-17>
- Daquigan N., Seekatz A. M., Greathouse K. L., Young V. B., White J. R. (2017) **High-resolution profiling of the gut microbiome reveals the extent of Clostridium difficile burden** *NPJ Biofilms and Microbiomes* **3**
- Darling A. C. E., Mau B., Blattner F. R., Perna N. T. (2004) **Mauve: multiple alignment of conserved genomic sequence with rearrangements** *Genome Research* **14**:1394–1403
- Davis M. Y., Zhang H., Brannan L. E., Carman R. J., Boone J. H. (2016) **Rapid change of fecal microbiome and disappearance of Clostridium difficile in a colonized infant after transition from breast milk to cow milk** *Microbiome* **4**
- Deane J., Fouhy F., Ronan N. J., Daly M., Fleming C., Eustace J. A., Shanahan F., Flanagan E. T., Dupont L., Harrison M. J., Haworth C. S., Floto A., Rea M. C., Ross R. P., Stanton C., Plant B. J. (2021) **A multicentre analysis of Clostridium difficile in persons with Cystic Fibrosis demonstrates that carriage may be transient and highly variable with respect to strain and level** *The Journal of Infection* **82**:363–370
- Drall K. M., Tun H. M., Morales-Lizcano N. P., Konya T. B., Guttman D. S., Field C. J., Mandal R., Wishart D. S., Becker A. B., Azad M. B., Lefebvre D. L., Mandhane P. J., Moraes T. J., Sears M. R., Turvey S. E., Subbarao P., Scott J. A., Kozyrskyj A. L. (2019) **Clostridioides difficile Colonization Is Differentially Associated With Gut Microbiome Profiles by Infant Feeding Modality at 3–4 Months of Age** *Frontiers in Immunology* **10** <https://doi.org/10.3389/fimmu.2019.02866>
- Eglow R., Pothoulakis C., Itzkowitz S., Israel E. J., O’Keane C. J., Gong D., Gao N., Xu Y. L., Walker W. A., LaMont J. T. (1992) **Diminished Clostridium difficile toxin A sensitivity in newborn rabbit ileum is associated with decreased toxin A receptor** *The Journal of Clinical Investigation* **90**:822–829

- Felczykowska A., Krajewska A., Zielińska S., Łoś J. M. (2015) **Sampling, metadata and DNA extraction - important steps in metagenomic studies** *Acta Biochimica Polonica* **62**:151–160
- Ferretti P., Pasolli E., Tett A., Asnicar F., Gorfer V., Fedi S., Armanini F., Truong D. T., Manara S., Zolfo M., Beghini F., Bertorelli R., De Sanctis V., Bariletti I., Canto R., Clementi R., Cologna M., Crifò T., Cusumano G., Segata N. (2018) **Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome** *Cell Host & Microbe* **24**:133–145
- Galdys A. L., Nelson J. S., Shutt K. A., Schlackman J. L., Pakstis D. L., Pasculle A. W., Marsh J. W., Harrison L. H., Curry S. R. (2014) **Prevalence and duration of asymptomatic Clostridium difficile carriage among healthy subjects in Pittsburgh, Pennsylvania** *Journal of Clinical Microbiology* **52**:2406–2409
- Gateau C., Couturier J., Coia J., Barbut F. (2018) **How to: diagnose infection caused by Clostridium difficile** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **24**:463–468
- Goldenberg S. D., French G. L. (2011) **Lack of association of tcdC type and binary toxin status with disease severity and outcome in toxigenic Clostridium difficile** *The Journal of Infection* **62**:355–362
- Golić N., Veljović K., Popović N., Djokić J., Strahinić I., Mrvaljević I., Terzić-Vidojević A. (2017) **In vitro and in vivo antagonistic activity of new probiotic culture against Clostridium difficile and Clostridium perfringens** *BMC Microbiology* **17**
- Hall I. C., O'toole E. (1935) **INTESTINAL FLORA IN NEW-BORN INFANTS: WITH A DESCRIPTION OF A NEW PATHOGENIC ANAEROBE BACILLUS DIFFICILIS.** *American Journal of Diseases of Children* **49**:390–402
- Hensgens M. P. M., Dekkers O. M., Demeulemeester A., Buiting A. G. M., Bloembergen P., van Benthem B. H. B., Le Cessie S., Kuijper E. J. (2014) **Diarrhoea in general practice: when should a Clostridium difficile infection be considered? Results of a nested case-control study.** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **20**:O1067–O1074
- Hildebrand F., Gossmann T. I., Frioux C., Özkurt E., Myers P. N., Ferretti P., Kuhn M., Bahram M., Nielsen H. B., Bork P. (2021) **Dispersal strategies shape persistence and evolution of human gut bacteria** *Cell Host & Microbe* **29**:1167–1176
- Högenauer C., Hammer H. F., Krejs G. J., Reisinger E. C. (1998) **Mechanisms and management of antibiotic-associated diarrhea** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **27**:702–710
- Humphries R. M., Uslan D. Z., Rubin Z. (2013) **Performance of Clostridium difficile toxin enzyme immunoassay and nucleic acid amplification tests stratified by patient disease severity** *Journal of Clinical Microbiology* **51**:869–873
- Jackson M., Olefson S., Machan J. T., Kelly C. R. (2016) **A high rate of alternative diagnoses in patients referred for presumed Clostridium difficile infection** *Journal of Clinical Gastroenterology* **50**:742–746

- Jangi S., Thomas Lamont J. (2010) **Asymptomatic Colonization by Clostridium difficile in Infants: Implications for Disease in Later Life** In *Journal of Pediatric Gastroenterology & Nutrition* **51**:2–7 <https://doi.org/10.1097/mpg.0b013e3181d29767>
- Jo H. B., Ham S. Y., Jung J., Moon S. M., Kim N.-H., Song K.-H., Park J. S., Park K. U., Kim E. S., Kim H. B. (2022) **Prevalence of and factors associated with inappropriate Clostridioides difficile testing in a teaching hospital in Korea** *Antimicrobial Resistance and Infection Control* **11**
- Kato H., Kita H., Karasawa T., Maegawa T., Koino Y., Takakuwa H., Saikai T., Kobayashi K., Yamagishi T., Nakamura S. (2001) **Colonisation and transmission of Clostridium difficile in healthy individuals examined by PCR ribotyping and pulsed-field gel electrophoresis** *Journal of Medical Microbiology* **50**:720–727
- Keel M. K., Songer J. G. (2007) **The distribution and density of Clostridium difficile toxin receptors on the intestinal mucosa of neonatal pigs** *Veterinary Pathology* **44**:814–822
- Kim J., Cho Y., Seo M.-R., Bae M. H., Kim B., Rho M., Pai H. (2020) **Quantitative characterization of Clostridioides difficile population in the gut microbiome of patients with C. difficile infection and their association with clinical factors** *Scientific Reports* **10** <https://doi.org/10.1038/s41598-020-74090-0>
- Kiu R., Hall L. J. (2018) **An update on the human and animal enteric pathogen Clostridium perfringens** *Emerging Microbes & Infections* **7**
- Kociolek L. K., Gerding D. N. (2016) **Breakthroughs in the treatment and prevention of Clostridium difficile infection** *Nature Reviews. Gastroenterology & Hepatology* **13**:150–160
- Korpela K., Costea P., Coelho L. P., Kandels-Lewis S., Willemsen G., Boomsma D. I., Segata N., Bork P. (2018) **Selective maternal seeding and environment shape the human gut microbiome** *Genome Research* **28**:561–568
- Kubota H., Makino H., Gawad A., Kushiro A., Ishikawa E., Sakai T., Akiyama T., Matsuda K., Martin R., Knol J., Oishi K. (2016) **Longitudinal Investigation of Carriage Rates, Counts, and Genotypes of Toxigenic Clostridium difficile in Early Infancy** *Applied and Environmental Microbiology* **82**:5806–5814
- Kumar R., Yi N., Zhi D., Eipers P., Goldsmith K. T., Dixon P., Crossman D. K., Crowley M. R., Lefkowitz E. J., Martin Rodriguez J., Morrow C. D. (2017) **Identification of donor microbe species that colonize and persist long term in the recipient after fecal transplant for recurrent Clostridium difficile** In *npj Biofilms and Microbiomes* **3** <https://doi.org/10.1038/s41522-017-0020-7>
- Kuznetsova A., Brockhoff P. B., Christensen R. H. B. (2017) **lmerTest Package: Tests in Linear Mixed Effects Models** In *Journal of Statistical Software* **82** <https://doi.org/10.18637/jss.v082.i13>
- Langdon A., Schwartz D. J., Bulow C., Sun X., Hink T., Reske K. A., Jones C., Burnham C.-A. D., Dubberke E. R., Dantas G. (2021) **Microbiota restoration reduces antibiotic-resistant bacteria gut colonization in patients with recurrent Clostridioides difficile infection from the open-label PUNCH CD study** *Genome Medicine* **13**
- Larcombe S., Hutton M. L., Lyras D. (2016) **Involvement of Bacteria Other Than Clostridium difficile in Antibiotic-Associated Diarrhoea** *Trends in Microbiology* **24**:463–476

- Larcombe S., Hutton M. L., Riley T. V., Abud H. E., Lyras D. (2018) **Diverse bacterial species contribute to antibiotic-associated diarrhoea and gastrointestinal damage** *The Journal of Infection* **77**:417–426
- Lee H. S., Plechot K., Gohil S., Le J. (2021) **Clostridium difficile: Diagnosis and the Consequence of Over Diagnosis** *Infectious Diseases and Therapy* **10**:687–697
- Lees E. A., Miyajima F., Pirmohamed M., Carrol E. D. (2016) **The role of Clostridium difficile in the paediatric and neonatal gut - a narrative review** *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* **35**:1047–1057
- Li H., Durbin R. (2009) **Fast and accurate short read alignment with Burrows-Wheeler transform** *Bioinformatics* **25**:1754–1760
- Litvin M., Reske K. A., Mayfield J., McMullen K. M., Georgantopoulos P., Copper S., Hoppe-Bauer J. E., Fraser V. J., Warren D. K., Dubberke E. R. (2009) **Identification of a pseudo-outbreak of Clostridium difficile infection (CDI) and the effect of repeated testing, sensitivity, and specificity on perceived prevalence of CDI** *Infection Control and Hospital Epidemiology: The Official Journal of the Society of Hospital Epidemiologists of America* **30**:1166–1171
- Liu B., Zheng D., Jin Q., Chen L., Yang J. (2019) **VFDB 2019: a comparative pathogenomic platform with an interactive web interface** *Nucleic Acids Research* **47**:D687–D692
- Martínez-Meléndez A., Camacho-Ortiz A., Morfin-Otero R., Maldonado-Garza H. J., Villarreal-Treviño L., Garza-González E. (2017) **Current knowledge on the laboratory diagnosis of Clostridium difficile infection** *World Journal of Gastroenterology: WJG* **23**:1552–1567
- McDonald L. C., Gerding D. N., Johnson S., Bakken J. S., Carroll K. C., Coffin S. E., Dubberke E. R., Garey K. W., Gould C. V., Kelly C., Loo V., Shaklee Sammons J., Sandora T. J., Wilcox M. H. (2018) **Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA)** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **66**:e1–e48
- Mende D. R., Letunic I., Huerta-Cepas J., Li S. S., Forslund K., Sunagawa S., Bork P. (2017) **proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes** *Nucleic Acids Research* **45**:D529–D534
- Milanese A., Mende D. R., Paoli L., Salazar G., Ruscheweyh H.-J., Cuenca M., Hingamp P., Alves R., Costea P. I., Coelho L. P., Schmidt T. S. B., Almeida A., Mitchell A. L., Finn R. D., Huerta-Cepas J., Bork P., Zeller G., Sunagawa S. (2019) **Microbial abundance, activity and population genomic profiling with mOTUs2** *Nature Communications* **10**
- Minh B. Q., Schmidt H. A., Chernomor O., Schrempf D., Woodhams M. D., von Haeseler A., Lanfear R. (2020) **IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era** *Molecular Biology and Evolution* **37**:1530–1534
- Monaghan T. M., Sloan T. J., Stockdale S. R., Blanchard A. M., Emes R. D., Wilcox M., Biswas R., Nashine R., Manke S., Gandhi J., Jain P., Bhotmange S., Ambalkar S., Satav A., Draper L. A., Hill C., Kashyap R. S. (2020) **Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome** *Gut Microbes* **12**

Na X., Kelly C. (2011) **Probiotics in clostridium difficile Infection** *Journal of Clinical Gastroenterology* **45**:S154–S158

Parks D. H., Rigato F., Vera-Wolf P., Krause L., Hugenholtz P., Tyson G. W., Wood D. L. A. (2021) **Evaluation of the Microba Community Profiler for Taxonomic Profiling of Metagenomic Datasets From the Human Gut Microbiome** *Frontiers in Microbiology* **0** <https://doi.org/10.3389/fmicb.2021.643682>

Parks T., Chapman T., Culver E., Scarborough M. (2011) **Persistent diarrhoea after presumed Clostridium difficile infection at Oxford Radcliffe Hospitals NHS Trust: an unrecognised entity** *Gut* **60**:A72–A72

Podlesny D., Florian Fricke W. (1970) **Podlesny, D., & Florian Fricke, W. (n.d.). Microbial Strain Engraftment, Persistence and Replacement after Fecal Microbiota Transplantation.** **10.1101/2020.09.29.20203638** <https://doi.org/10.1101/2020.09.29.20203638>

Polage C. R., Gyorko C. E., Kennedy M. A., Leslie J. L., Chin D. L., Wang S., Nguyen H. H., Huang B., Tang Y.-W., Lee L. W., Kim K., Taylor S., Romano P. S., Panacek E. A., Goodell P. B., Solnick J. V., Cohen S. H. (2015) **Overdiagnosis of Clostridium difficile Infection in the Molecular Test Era** *JAMA Internal Medicine* **175**:1792–1801

Polage C. R., Solnick J. V., Cohen S. H. (2012) **Nosocomial diarrhea: evaluation and treatment of causes other than Clostridium difficile** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **55**:982–989

Pruitt R. N., Lacy D. B. (2012) **Toward a structural understanding of Clostridium difficile toxins A and B** *Frontiers in Cellular and Infection Microbiology* **2**

Reich N., Payne M., Sharma A., Champagne S., Romney M. G., Hinch M., Leung V., Lowe C. F. (2019) **Prospective Review of Clostridioides difficile Testing Indications to Inform Local Laboratory Stewardship Initiatives** *Infection Prevention in Practice* **1**

Rigottier-Gois L. (2013) **Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis** *The ISME Journal* **7**:1256–1261

Rivera-Chávez F., Zhang L. F., Faber F., Lopez C. A., Byndloss M. X., Olsan E. E., Xu G., Velazquez E. M., Lebrilla C. B., Winter S. E., Bäuml A. J. (2016) **Depletion of Butyrate-Producing Clostridia from the Gut Microbiota Drives an Aerobic Luminal Expansion of Salmonella** *Cell Host & Microbe* **19**:443–454

Rodriguez C., Warszawski N., Korsak N., Taminiau B., Van Broeck J., Delmée M., Daube G. (2016) **Laboratory identification of anaerobic bacteria isolated on Clostridium difficile selective medium** *Acta Microbiologica et Immunologica Hungarica* **63**:171–184

Rodríguez J. M., Murphy K., Stanton C., Ross R. P., Kober O. I., Juge N., Avershina E., Rudi K., Narbad A., Jenmalm M. C., Marchesi J. R., Collado M. C. (2015) **The composition of the gut microbiota throughout life, with an emphasis on early life** *Microbial Ecology in Health and Disease* **26**

Rousseau C., Poilane I., De Pontual L., Maherault A.-C., Le Monnier A., Collignon A. (2012) **Clostridium difficile carriage in healthy infants in the community: a potential reservoir for pathogenic strains** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **55**:1209–1215

Schmidt T. S., Hayward M. R., Coelho L. P., Li S. S., Costea P. I., Voigt A. Y., Wirbel J., Maistrenko O. M., Alves R. J., Bergsten E., de Beaufort C., Sobhani I., Heintz-Buschart A., Sunagawa S., Zeller G., Wilmes P., Bork P. (2019) **Extensive transmission of microbes along the gastrointestinal tract** *eLife* **8** <https://doi.org/10.7554/eLife.42693>

Schubert A. M., Rogers M. A. M., Ring C., Mogle J., Petrosino J. P., Young V. B., Aronoff D. M., Schloss P. D. (2014) **Microbiome data distinguish patients with *Clostridium difficile* infection and non-*C. difficile*-associated diarrhea from healthy controls** *mBio* **5**:e01021–14

Schutze G. E., Willoughby R. E. (2013) **Committee on Infectious Diseases, & American Academy of Pediatrics** *Clostridium difficile* infection in infants and children. *Pediatrics* **131**:196–200

Seekatz A. M., Rao K., Santhosh K., Young V. B. (2016) **Dynamics of the fecal microbiome in patients with recurrent and nonrecurrent *Clostridium difficile* infection** *Genome Medicine* **8**

Smillie C. S., Sauk J., Gevers D., Friedman J., Sung J., Youngster I., Hohmann E. L., Staley C., Khoruts A., Sadowsky M. J., Allegretti J. R., Smith M. B., Xavier R. J., Alm E. J. (2018) **Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation** *Cell Host & Microbe* **23**:229–240

Smits W. K., Lyras D., Lacy D. B., Wilcox M. H., Kuijper E. J. (2016) ***Clostridium difficile* infection** *Nature Reviews. Disease Primers* **2**

Spina A., Kerr K. G., Cormican M., Barbut F., Eigentler A., Zerva L., Tassios P., Popescu G. A., Rafila A., Eerola E., Batista J., Maass M., Aschbacher R., Olsen K. E. P., Allerberger F. (2015) **Spectrum of enteropathogens detected by the FilmArray GI Panel in a multicentre study of community-acquired gastroenteritis** *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **21**:719–728

Stewart D. B., Wright J. R., Fowler M., McLimans C. J., Tokarev V., Amaniera I., Baker O., Wong H.-T., Brabec J., Drucker R., Lamendella R. (2019) **Integrated Meta-omics Reveals a Fungus-Associated Bacteriome and Distinct Functional Pathways in *Clostridioides difficile* Infection** *In mSphere* **4** <https://doi.org/10.1128/msphere.00454-19>

Tang D. M., Urrunaga N. H., De Groot H., von Rosenvinge E. C., Xie G., Ghazi L. J. (2014) **Pseudomembranous Colitis: Not Always Caused by *Clostridium difficile*** *Case Reports in Medicine* **2014**

Tang D. M., Urrunaga N. H., von Rosenvinge E. C. (2016) **Pseudomembranous colitis: Not always *Clostridium difficile*** *Cleveland Clinic Journal of Medicine* **83**:361–366

Tenover F. C., Baron E. J., Peterson L. R., Persing D. H. (2011) **Laboratory diagnosis of *Clostridium difficile* infection can molecular amplification methods move us out of uncertainty?** *The Journal of Molecular Diagnostics: JMD* **13**:573–582

Thevenon O., Adema W., Clarke C. (2016) **Background brief on fathers' leave and its use** *OECD* <https://doi.org/10.13140/RG.2.2.27717.24808>

Van Rossum T., Costea P. I., Paoli L., Alves R., Thielemann R., Sunagawa S., Bork P. (2021) **metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes** *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btab789>

- Vincent C., Miller M. A., Edens T. J., Mehrotra S., Dewar K., Manges A. R. (2016) **Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and *Clostridium difficile* colonization or infection** *Microbiome* **4**
- Voigt A. Y., Costea P. I., Kultima J. R., Li S. S., Zeller G., Sunagawa S., Bork P. (2015) **Temporal and technical variability of human gut metagenomes** *Genome Biology* **16**
- Voth D. E., Ballard J. D. (2005) ***Clostridium difficile* toxins: mechanism of action and role in disease** *Clinical Microbiology Reviews* **18**:247–263
- Watson A. R., Fuessel J., Veseli I., DeLongchamp J. Z., Silva M., Trigodet F., Lolans K., Shaiber A., Fogarty E., Runde J. M., Quince C., Yu M. K., Soylev A., Morrison H. G., Lee S. T. M., Kao D., Rubin D. T., Jabri B., Louie T., Murat Eren A (1970) **Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut** <https://doi.org/10.1101/2021.03.02.433653>
- Willemsen L. E. M., Koetsier M. A., van Deventer S. J. H., van Tol E. A. F. (2003) **Short chain fatty acids stimulate epithelial mucin 2 expression through differential effects on prostaglandin E(1) and E(2) production by intestinal myofibroblasts** *Gut* **52**:1442–1447
- Wirbel J., Zych K., Essex M., Karcher N., Kartal E., Salazar G., Bork P., Sunagawa S., Zeller G. (2021) **Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox** *Genome Biology* **22**
- Yassour M., Vatanen T., Siljander H., Hämäläinen A.-M., Härkönen T., Ryhänen S. J., Franzosa E. A., Vlamakis H., Huttenhower C., Gevers D., Lander E. S., Knip M., Xavier R. J. (2016) **Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability** *Science Translational Medicine* **8**
- Yuan Q., Guo X., Ren Y., Wen X., Gao L. (2020) **Cluster correlation based method for lncRNA-disease association prediction** *BMC Bioinformatics* **21**
- Zhang S., Palazuelos-Munoz S., Balsells E. M., Nair H., Chit A., Kyaw M. H. (2016) **Cost of hospital management of *Clostridium difficile* infection in United States-a meta-analysis and modelling study** *BMC Infectious Diseases* **16**
- Zhou Y., Wylie K. M., El Feghaly R. E., Mihindukulasuriya K. A., Elward A., Haslam D. B., Storch G. A., Weinstock G. M. (2016) **Metagenomic Approach for Identification of the Pathogens Associated with Diarrhea in Stool Specimens** *Journal of Clinical Microbiology* **54**:368–375
- Zollner-Schwetz I., Högenauer C., Joainig M., Weberhofer P., Gorkiewicz G., Valentin T., Hinterleitner T. A., Krause R. (2008) **Role of *Klebsiella oxytoca* in antibiotic-associated diarrhea** *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **47**:e74–e78

Author information

Pamela Ferretti

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-1707-9013](https://orcid.org/0000-0002-1707-9013)

Jakob Wirbel

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-4073-3562](https://orcid.org/0000-0002-4073-3562)

Oleksandr M Maistrenko

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany, Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Microbiology & Biogeochemistry, 1797 SZ, 't Horntje (Texel), Netherlands
ORCID iD: [0000-0003-1961-7548](https://orcid.org/0000-0003-1961-7548)

Thea Van Rossum

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-3598-5001](https://orcid.org/0000-0002-3598-5001)

Renato Alves

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-7212-0234](https://orcid.org/0000-0002-7212-0234)

Anthony Fullam

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-0884-8124](https://orcid.org/0000-0002-0884-8124)

Wasiu Akanni

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0002-2075-2387](https://orcid.org/0000-0002-2075-2387)

Christian Schudoma

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0003-1157-1354](https://orcid.org/0000-0003-1157-1354)

Anna Schwarz

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany

Roman Thielemann

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany
ORCID iD: [0000-0003-2582-6178](https://orcid.org/0000-0003-2582-6178)

Leonie Thomas

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany

Stefanie Kandels

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117
Heidelberg, Germany
ORCID iD: [0000-0002-4194-4927](https://orcid.org/0000-0002-4194-4927)

Rajna Hercog

Genomic Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany
ORCID iD: [0000-0003-2587-9739](https://orcid.org/0000-0003-2587-9739)

Anja Telzerow

Genomic Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany
ORCID iD: [0000-0001-9855-0809](https://orcid.org/0000-0001-9855-0809)

Ivica Letunic

biobyte solutions GmbH, Bothestr 142, 69126 Heidelberg
ORCID iD: [0000-0003-3560-4288](https://orcid.org/0000-0003-3560-4288)

Michael Kuhn

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117
Heidelberg, Germany
ORCID iD: [0000-0002-2841-872X](https://orcid.org/0000-0002-2841-872X)

Georg Zeller

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117
Heidelberg, Germany
ORCID iD: [0000-0003-1429-7485](https://orcid.org/0000-0003-1429-7485)

Thomas SB Schmidt

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117
Heidelberg, Germany
ORCID iD: [0000-0001-8587-4177](https://orcid.org/0000-0001-8587-4177)

Peer Bork

European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117
Heidelberg, Germany, Max Delbrück Centre for Molecular Medicine, Berlin, Germany, Yonsei
Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea, Department of Bioinformatics,
Biocenter, University of Würzburg, Würzburg, Germany
For correspondence: bork@embl.de
ORCID iD: [0000-0002-2627-833X](https://orcid.org/0000-0002-2627-833X)

Editors

Reviewing Editor

María Zambrano

CorpoGen, Colombia

Senior Editor

Wendy Garrett

Harvard T.H. Chan School of Public Health, United States of America

Reviewer #1 (Public Review):

Summary:

The authors present a comprehensive meta-analysis of *Clostridioides difficile* (CD) occurrence across 42,900 metagenomes from 253 public studies, largely representing stool samples from human adults, infants, and with a smaller fraction of samples from non-gut body sites and from environmental samples (e.g., non-human animals, wastewater, soil, etc.). In particular, the authors looked at adults who were healthy, diseased (but not with *C. diff*), and with diagnosed *C. diff* infection (CDI) and found that CD occurrence was fairly low: ~30% in adult CDI samples, ~2% in adult diseased samples, and ~1% in healthy samples. CD was much more prevalent in infants (15 and 40% in healthy and diseased infants, respectively). These findings, if they hold true, would be significant because they would suggest an over-diagnosis of CDI and an under-diagnosis of other putative enteric pathogens (also enriched in CDI samples) across the population. Furthermore, these results suggest that the asymptomatic carriage of CD in adults (~1-5%, depending on demographics) may be much lower than some prior estimates (some as high as ~30-40%).

Strengths:

The authors have done an admirable job pulling down an enormous data set for this CD-focused meta-analysis, which is a valuable service to the field. The results push against some common wisdom in the field, in terms of the prevalence of CD in CDI patients, which will be impactful if they hold up to further scrutiny. Furthermore, the identification of commensal bacteria that are positively or negatively associated with CD presence in both healthy and diseased people at different periods of the lifespan (infant, child, and adult), is a valuable synthesis with potential translational value. The manuscript is clearly written and the figures are presented well. The methodology is robust, although I have a few suggestions for improvement.

Weaknesses:

My main critique relates to detection limitations, both in terms of sequencing depth and read-mapping. Given that CD detection is the root of the main conclusions reported here, this deserves some additional care. The authors have already done some work to address this by including sequencing depth in their linear mixed effects model, which is great. Furthermore, they were conservative with how they labeled CD positive/negative individuals with multiple time points (i.e., if you had CD detected at any point, this sample was selected for the cross-sectional analysis, and that individual was labeled as CD positive). I have a few additional suggestions to explore this issue, which I outline in the recommendations for authors.

Reviewer #2 (Public Review):

Summary:

C. difficile infection (CDI) is clinically important as a hospital-acquired infection and a frequent cause of antibiotic-associated diarrhea, which is associated with high morbidity and mortality and increases in prevalence. It is also the prime example of a disease that is associated with gut microbiome dysbiosis and successfully treated with fecal microbiota transfer, highlighting the important but unclear functional or structural role of this bacterial pathogen and the condition of CDI for the gut microbiome, which is the focus of this study.

Ferretti et al. assembled an impressive gut metagenome dataset from previous and ongoing microbiome studies, which involves a large number of samples from patients with CDI or other diarrheal and non-diarrheal diseases and from healthy individuals, as well as from infants, adolescents, and adults. The authors analyze the prevalence and relative abundance

of *C. difficile* in this dataset in relation to CDI diagnosis, host age and disease background, and the composition of the remaining microbiota. They detect *C. difficile* only in a minority of samples labelled as originating from CDI patients but frequently identify other pathogens and their toxin genes in the same samples. In infants, they detect *C. difficile* at high frequency and relative abundance in samples without clinical symptoms. They associate *C. difficile* presence in infant samples with 'multiple indicators of healthy gut microbiome maturation' and suggest 'distinct biotic and physiological contexts in infants and adults' for *C. difficile*.

Strengths:

The manuscript provides an important overview of the complex relationship of *C. difficile* with the gut microbiome of healthy and diseased infants and adults, mostly due to the large studied dataset and convincing applied analysis that underlies the presented findings. This includes a number of interesting findings including, for example, that CDI can be reliably predicted based on taxonomic microbiota compositions, without including *C. difficile* itself or that *C. difficile* in infants appears not to originate from maternal sources.

Weaknesses:

Inconsistent associations of *C. difficile* with what is clinically labeled CDI, as well as the frequent detection of *C. difficile* in healthy infants, have been reported before and the manuscript does not reveal to what extent this bacterium reflects or even directly influences the gut microbiome of infants and adults. Whether the increased microbiota diversity, richness, and compositional similarity of *C. difficile*-positive infants to their mothers is sufficient to associate this bacterium with "healthy gut microbiome maturation" seems questionable, since *C. difficile* was also found to be more prevalent in preterm infants, formula-fed or antibiotic-treated infants, and infants born by C-section, all of which are typically considered detrimental influences on microbiota development. The conclusion that "C. difficile may be a transient hallmark of healthy gut microbiome maturation" therefore appears too strong.

In addition, the statement that "its asymptomatic carriage in adults depends on microbial context" is not sufficiently supported by the presented data. Apparently, the authors are unable to define or measure "asymptomatic carriage", as they convincingly show that many patients diagnosed with "CDI" appear not to carry *C. difficile*, suggesting that neither asymptomatic nor symptomatic "CDI" conditions are necessarily linked to *C. difficile*.

The manuscript includes a large number of samples from poorly defined, but diverse patient backgrounds. It might be helpful to better define these samples (e.g. fecal samples vs. other gut samples) and to specify subcategories for samples from "diseased control subjects without CDI". Maybe this information could help validate the interesting suggestion from the manuscript that *C. difficile* may be (one of several) dysbiosis marker rather than the cause of (CDI) dysbiosis.

The phylogenetic analysis of *C. difficile* from metagenomic sequence data seems to suggest that there is a large mostly toxin gene-free cluster that is only identified in infants (Supplementary Figure 13). Could this indicate that there are, in fact, less pathogenic *C. difficile* lineages that are more prevalent in infants?

The authors argue in the Discussion that "Differential diagnosis against multiple enteropathogens may therefore stratify patients with CDI-like symptoms, towards adapted therapeutic interventions." It might be helpful to expand this discussion of different clinical options that could be adapted to highlight the clinical applicability of the presented findings.