**C. difficile** is overdiagnosed in adults and a commensal in infants

Pamela Ferretti1, Jakob Wirbel1, Oleksandr M Maistrenko1,2, Thea Van Rossum1, Renato Alves1, Anthony Fullam1, Wasiu Akanni1, Christian Schudoma1, Anna Schwarz1, Roman Thielemann1, Leonie Thomas1, Michael Kuhn1, Georg Zeller1, Thomas SB Schmidt1 and Peer Bork1,3,4,5

**Affiliations:**

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

2 Present address: Royal Netherlands Institute for Sea Research (NIOZ), Department of Marine Microbiology & Biogeochemistry, 1797 SZ, ’t Horntje (Texel), Netherlands

3 Max Delbrück Centre for Molecular Medicine, Berlin, Germany

4 Yonsei Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea

5 Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

x These authors contributed equally to this work.

*Correspondence: sebastian.schmidt@embl.de and Bork@embl.de

**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. Here, we analysed 534 metagenomes from 10 publicly available CDI study populations. While we detected *C. difficile* in only 30% of CDI samples, multiple other toxigenic species capable of inducing CDI-like symptomatology were prevalent, raising concerns about CDI overdiagnosis. We further tracked *C. difficile* in 42,814 metagenomic samples from 253 public studies. We found that *C. difficile* prevalence, abundance and association with other bacterial species is age-dependent. In healthy adults, *C. difficile* is a rare taxon associated with an overall species richness reduction, while in healthy infants *C. difficile* is a common member of the gut microbiome and its presence is associated with a significant increase in species richness. More specifically, we identified a group of species co-occurring with *C. difficile* exclusively in healthy infants, enriched in obligate anaerobes and in species typically found in the gut microbiome of healthy adults. Overall, gut microbiome composition in presence of *C. difficile* in healthy infants is associated with multiple
parameters linked to a healthy gut microbiome maturation towards an adult-like state. Our results suggest that *C. difficile* is a commensal in infants, and that its asymptomatic carriage is dependent on the surrounding microbial context.

**Keywords:** *Clostridioides difficile*, metagenomics, CDI, antibiotic-associated diarrhea, infant microbiome, overdiagnosis

**Introduction**

*Clostridioides difficile* (previously *Clostridium* or *Peptoclostridium difficile*) was first isolated in 1935 from the stools of healthy infants, but later identified as a causative agent of pseudomembranous colitis, 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 damage the host by increasing intestinal permeability and promoting intense inflammation, a condition referred to as *C. difficile* infection (CDI). CDI is typically characterised by diarrhea and/or pseudomembranous colitis and often unresponsive to repeated antibiotic treatments. CDI has an increasing clinical relevance (at an estimated annual economic burden of US$6.3 billion in the US alone) and is considered one of the most urgent threats in hospital-acquired infections. *C. difficile* has primarily been studied as a pathogen, with a strong focus on CDI aetiology, mechanisms of *C. difficile* toxicity or efficacious therapies. However, an increasing number of other enteropathogens has been linked to antibiotics-associated diarrhea (AAD) with similar, and sometimes indistinguishable, symptomatology to CDI, 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. Gut microbial composition associated with *C. difficile* presence in healthy subjects, in particular in infants where *C. difficile* is highly prevalent, remains poorly understood.

Here we leverage public shotgun metagenomic data to characterise the gut microbial signature associated with CDI in adult and elderly patients and to quantify other AAD-associated species with similar signatures. We then contextualise our analysis by tracing *C. difficile* in a broad multi-habitat collection of 42,814 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.
Results

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

We first conducted a metagenomic meta-analysis of the gut microbiome associated with CDI, analysing 534 fecal samples across 10 geographically diverse study populations14–22, including patients diagnosed with CDI, diseased subjects without CDI (D-Ctr, see Methods), and healthy subjects (H-Ctr) (Figure 1A and Supplementary Table 1).

Using mOTUs-based species-level taxonomic profiling23, *C. difficile* was detectable in just 30% of CDI samples at established metagenomic detection limits23 (Figure 1A), 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, 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; R²=0.01), and that detection was not skewed towards the theoretical metagenomic detection limit of 10⁻⁵ relative abundance24. Neither species richness (Figure 1B) 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 diagnosed 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, Supplementary Figure 3). Among these, we identified several opportunistic enteropathogens known to induce antibiotic-associated diarrhea (AAD), resulting in a CDI-like diarrheal symptomatology8,10,25–29, such as *Klebsiella oxytoca, Citrobacter amalonaticus, Clostridium innocuum, Clostridium perfringens, Staphylococcus aureus, Enterococcus faecalis, Enterobacter cloacae*, 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, Wilcoxon test p=8.2x10⁻¹⁶ and p=4.3x10⁻⁵, 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=2.2x10⁻¹⁶,
OR=29.62; Supplementary Figure 1A and 1B). Presence and potential toxigenicity of several enteropathogens was further confirmed by the detection of characteristic toxin genes (see Methods) that were likewise enriched in CDI samples (Figure 1D). In particular, enteropathogenic *E. coli* and *S. flexneri* were, either alone or in combination with other species, the most common species with toxin genes. Overall, CDI samples, independently of *C. difficile* presence, were significantly enriched (*p*=2.22×10⁻¹⁶) 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 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 transmission⁵⁰ among these patients. 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 patients³¹,³² or to the increased antibiotic resistance of these species³³. 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 investigation⁴⁴. 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, for a total of 42,814 shotgun metagenomic samples (Supplementary Table 2), including healthy and diseased subjects of all ages, with samples from multiple body sites, hosts, environments and geographic locations (Figure 3A and 3B). Results from this point onwards refer to this extended collection of samples.

*C. difficile* was detected almost exclusively in human and animal feces (5.9% and 1.7% prevalence, respectively) with sporadic detection in hospital surfaces (0.19%), the respiratory tract of diseased patients (0.37%) and freshwater (0.88%) (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 is no relationship between *C. difficile* detection and sequencing depth (linear mixed effect model p=0.36; logistic regression, ANOVA adjusted p=0.44, Supplementary Figure 4A). 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 (44.5%) between 1 and 4 years of age, in contrast to previous reports. Among older individuals, *C. difficile* was increasingly rare, found in less than 1.06% of the healthy population (7.6% in adolescents, 0.92% 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%). *C. difficile* carriage varied across geography (Supplementary Figure 4B) and was lower in non-westernised populations, at 4.76% and 0.35% in infants and adults, respectively.

*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 2.22) and infants suffering from cystic fibrosis (78.7%,
4x10^{-3}) or neonatal sepsis/necrotizing enterocolitis (55.3%, 3.2x10^{-2}; Supplementary Figure 5). An elevated \textit{C. difficile} asymptomatic carriage rate in cystic fibrosis patients is well documented\textsuperscript{38} and may be the result of long-term antibiotics usage and frequent exposure to the hospital environment\textsuperscript{39}. Among adolescents, adults and elderly, \textit{C. difficile} prevalence was highest in subjects taking antibiotics (on average 43.6% across age groups, relative abundance 4.4^{-3}), followed by patients suffering from CDI (30%, see above) or undefined diarrhea (25.9%). Inflammatory bowel diseases, liver diseases, diabetes, and various forms of cancer were likewise associated with an increased \textit{C. difficile} carriage relative to the healthy background population (Supplementary Figure 5).

\textbf{C. difficile is a frequent commensal in the healthy infant gut}

\textit{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 (30.16%; logistic regression, adjusted ANOVA \textit{p}=2.1x10^{-5}, logOR=3.1x10^{-5}) compared to age- and health status-matched vaginally born infants (24.94%) (Figure 4B and Supplementary Figure 6). Likewise, prevalence was moderately increased among pre-term (36.17%; \textit{p}=3.3x10^{-4}, logOR=2.5x10^{-5}) compared to full-term healthy infants (25.62%) (Figure 4C and Supplementary Figure 6). Exclusively formula-fed infants were more frequently colonised by \textit{C. difficile} (43.18%; \textit{p}=4.8x10^{-9}, logOR=5x10^{-5}), compared to partially or exclusively breastfed infants during the first 6 months of life (14.23% and 8.6% respectively) (Figure 4D and Supplementary Figure 6), in line with previous reports\textsuperscript{40,41}.

\textit{C. difficile} was not maternally acquired during birth: across 820 mother-infant pairs included in the dataset, we found no single case of vertical transmission of \textit{C. difficile} from any of the maternal body sites investigated (gut, vagina and oral cavity). Rather, \textit{C. difficile} was likely sourced from the environment during early life, in line with previous observations on \textit{Clostridia}\textsuperscript{42}. Leveraging longitudinal data, we observed that the first appearance of \textit{C. difficile} was not arbitrarily distributed over the first year of life, but instead concentrated in two defined time windows, between the 2\textsuperscript{nd}-4\textsuperscript{th} and 8\textsuperscript{th}-10\textsuperscript{th} 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\textsuperscript{43}, 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\textsuperscript{44}, suggesting that it could be due to the start of day care. C-section and premature birth were associated with increased rate of \textit{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 (22.6%) and adults (26.8%) 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). 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 reports45, 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 \textit{Roseburia} sp., \textit{Faecalibacterium prausnitzii} or \textit{Anaerostipes hadrus}) (Figure 5 and Supplementary Figure 12).

To investigate whether healthy infants were associated with specific \textit{C. difficile} strains, we performed an analysis of metagenomic single nucleotide variants (SNVs). Despite the distinct biotic context associated with \textit{C. difficile} presence in healthy infants at the species level, \textit{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 \textit{C. difficile} carriage in healthy infants is likely determined by its surrounding microbial community and host age, rather than specific strain groups.

**Discussion**

\textit{C. difficile} has been predominantly studied in the context of its pathogenicity, but our analysis suggests that considering \textit{C. difficile} exclusively as a pathogen is falling short. \textit{C. difficile}’s role in causing antibiotics-associated diarrhea (AAD) or pseudomembranous colitis (PMC) may have been overestimated as our results strongly suggest that symptoms can also be explained by other AAD or PMC associated species, which would indicate CDI over-diagnosis in clinical practice. We detected fecal \textit{C. difficile} in just 30% of cases diagnosed with CDI. This is consistent with the absence of \textit{C. difficile} among CDI samples reported in previous 16S rRNA\textsuperscript{45,46} and WGS metagenomic studies\textsuperscript{14,47}, even when combined with laboratory-based tests\textsuperscript{46}. Although the sensitivity of our analysis is dependent on metagenomic detection limits in principle, we found that variance in sequencing depth did not explain the observed variability in \textit{C. difficile} detection. While it is possible that \textit{C. difficile} might have been successfully eradicated by the antibiotic treatments following the CDI diagnosis but anteceding the sampling, the CDI samples included in this analysis were labelled as CDI in the original studies. In this context, we hypothesise that the under-detection of \textit{C. difficile} among CDI patients is likely due to a lack of diagnostic specificity. Proper CDI diagnosis is surprisingly challenging in current health systems, with only marginal consensus on the use of the vast array of protocols and diagnostic algorithms available in the clinical practice. Diarrheal symptoms, often the initial diagnostic prompt, are not
specific to CDI 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. Additional challenges include the large variability in sensitivity and specificity in diagnostic laboratory tests, contradicting results between cultivation and toxin detection assays, and an excessive reliance on single molecular tests. Recommended guidelines discourage stand-alone tests for CDI diagnosis and recommend a two-step procedure, but due to costs, test availability, capacity and turnaround times, diagnostic algorithms vary between hospitals and often deviate from recommended procedures. In our CDI meta-analysis, almost two thirds (62.5%) of the studies with published CDI diagnostic protocols did deviate from recommended best practices. The use of multiple alternative protocols within studies, combined with the lack of published per-sample diagnostic information, prevented further investigation of variability in C. difficile detection rates between different diagnostic protocols.

Simultaneous colonisation of multiple enteropathogens is very common among CDI patients. We found several species known to induce AAD or PMC enriched in CDI samples, along with their characteristic toxin genes, independently of C. difficile presence. In particular, at least one of these species is present in 90% of C. difficile negative CDI samples (toxins were detected in 43% of samples). While C. difficile was the most individually characteristic species in CDI, it was not an essential input to microbiome-based classification models accurately predicting the disease. Enteropathogens like C. perfringens, C. innocuum, S. aureus, E. coli or S. flexneri, among others, may contribute to or indeed drive clinical manifestations of CDI. Our data strongly suggests that differential diagnosis against multiple enteropathogens may stratify patients with CDI-like symptoms, towards adapted therapeutic interventions. Disentangling the real burden of C. difficile from other AAD species could also aid its classification, which remains inconsistent, with most studies classifying C. difficile as a pathogen, some as an opportunistic pathogen, and few others as a pathobiont. This is also attributed to the limited study of C. difficile in healthy individuals and hence potential commensal features. To address this, we tracked C. difficile across an extensive set of 42,814 metagenomes sampled from multiple habitats and hosts. Samples from the human gut (n=28,347, 66.2%) covered a wide range of subject ages, disease states and geographic origins. C. difficile was very common among healthy and diseased 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.
In healthy infants, we could associate the occurrence of *C. difficile* 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. Indeed, an increase in species richness and number of obligate anaerobes are among the community-wide changes that the infant gut microbiome undergoes during its healthy development, as shown in several previous studies. In contrast, *C. difficile* presence in adulthood was associated with decreased microbiome richness and an enrichment in facultative and obligate aerobes. Increased oxygen tolerance in the adult gut has been associated with dysbiotic states found in IBD patients and cases of *Salmonella* proliferation. 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, indicating asymptomatic carriage of toxigenic strains. Indeed, *C. difficile* testing is generally discouraged by pediatric guidelines even in presence of diarrheal symptoms and toxin concentration may not be a reliable proxy for inferring disease and its severity. The apparent protection from CDI symptoms has been ascribed to a lack of toxin receptors in the immature infant gut, although this early hypothesis has since been called into question. 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), all considered protective against *C. difficile* toxins.

Our analyses are descriptive and arguably inherently limited by the underlying metagenomic data, such as DNA extraction bias against endospores, incomplete correlation of gene dosage and activity 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 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.

**Online Methods**

**Data overview**

**CDI meta-analysis**

In the CDI meta-analysis, out of 534 samples in total (from as many subjects, with an average age of 60±19.88 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 diarrheal symptoms but negative CDI diagnostic outcome, or was diarrhea-free but diagnosed for a disease other than CDI. The majority of the samples in the latter group (D-Ctr) belong to hospitalised subjects.

**Global meta-analysis**

We analysed a total of 42,814 publicly available metagenomic samples (collectively >220Tbp) from 253 different studies (see Supplementary Table 2). The collection includes samples from 84 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.
Data download

Metagenomes publicly available on the 22\textsuperscript{nd} of April 2021 were downloaded using fetch-data\textsuperscript{83}. Only shotgun metagenomic samples sequenced using Illumina platforms have been included. No minimum number of samples per study or minimum sequencing depth threshold was applied during data download.

Metadata curation

Manually curated metadata for each dataset include: health status, age, geography (country and continent name), westernised lifestyle or not, diagnosis for \textit{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, $1$ year $< \text{child} \leq 10$ years, $10$ years $< \text{adolescent} \leq 18$ years, $18$ 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 were discarded ($n=1,091$) and (ii) samples with less than 59 reads, corresponding to 95\textsuperscript{th} percentile calculated on the remaining samples, were discarded ($n=2,083$). An additional third read count filtering was performed only on human gut stool samples, corresponding to 99\textsuperscript{th} percentile, removing samples with less than 100 reads ($n=138$ samples). The resulting filtered dataset included 39,502 samples, of which 26,784 were human fecal metagenomes.

Identification of timeseries-representative samples

Out of 26,784 samples, 24,864 had subject-level metadata and were associated with 12,402 unique subjects. For 3,576 subjects, multiple timepoints were available. The mean number of timepoints per subject was 2 (3.18 for infants, 1.68 for children, 3.56 for adolescents, 1.68 for adults and 1.34 for elderly), with a maximum of 205 time points per subject.

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

(ii) *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.

(iii) *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.

No subject metadata were available for 1,920 samples. 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\(^{23}\), 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 \textit{C. difficile} mOTUs2.0 ("ref\_mOTU\_0051" and "ref\_mOTU\_0052") were considered in downstream analysis. Abundance estimations included only \textit{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\textsuperscript{84} with 10-fold cross-validation. For this analysis, we focused on the subset of 10 CDI or diarrhea-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\textsuperscript{84}, 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 \textit{C. difficile}, we trained another set of models with the same cross-validation splits but excluded \textit{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, 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\textsuperscript{85}. 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 \textit{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-codifying genes and their assignment was performed using the VFDB database\textsuperscript{86}, as downloaded in March 2021. In our analysis we did not investigate the presence of the binary toxin gene, since this toxin alone has not been associated with disease severity\textsuperscript{87}. Mapping against VFDB was performed via BWA\textsuperscript{88} and filtering via NGLess\textsuperscript{89}, using 99\% as minimal alignment percentage identity threshold and 75bp as minimal read length match. Toxin gene profiles were computed with gffquant (\url{https://github.com/grp-bork/gff_quantifier}, version 1.2.3) using the 1overN count model. Toxin genes included in VFDB refer to the following strains: \textit{C. difficile} 630, \textit{E. coli} CFT073, O44:H18 042 and O157:H7 str. EDL933, \textit{S. flexneri} 2a str.301, \textit{P. aeruginosa} PAO1, \textit{C. perfringens} str.13 and SM101, \textit{S. aureus} RN4220, subsp. aureus MW2 and N315.

**Subspecies analysis**

\textit{C. difficile} subspecies detection was performed on all \textit{C. difficile} positive samples with MetaSNV2\textsuperscript{90}. Reads from 2441 samples were mapped against the proGenomes v1\textsuperscript{91} species representatives genomes for 3 species in the Clostridium genus: i) spec1_v2_0051 (NCBI taxonomy ID 272563, PRJNA78) \textit{Clostridioides difficile}, ii) spec1_v2_0052 (NCBI taxonomy ID 1151292, PRJNA85757) \textit{Clostridioides difficile} and iii) spec1_v2_1125 (NCBI taxonomy ID 1408823, PRJNA223331) \textit{Clostridioides mangenotii}. By mapping to multiple genomes and only using the uniquely mapped reads, we essentially focus 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\textsuperscript{88} and NGLess\textsuperscript{89}. 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 2441 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 previously reported approach. 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. 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 IQTREE2 tree-builder. 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) *Clostridiodides difficile* and specI_v2_1125 (NCBI taxonomy ID 1408823, PRJNA223331) *Clostridiodides mangenotii* into SNP input file for IQTREE2 we used Mauve 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.

**Data availability**

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/pamela314/cdifficile.

Acknowledgements

We are thankful to the members of the groups of Peer Bork, Georg Zeller and Ed Kuiper, Prof. 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.

Figure & Table Legends

Figure 1. (A) On top, fecal metagenomes used in the meta-analysis from 10 public CDI or diarrheal 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 \(C.\ difficle\) positive samples per group. \(C.\ difficle\) toxin genes were found only among CDI \(C.\ difficle\) positive samples. (B) Rarefied species richness across the three groups. (C) Prevalence, relative abundance of species that induce antibiotic-associated diarrhea (AAD) among CDI and controls. (D) Prevalence of single and multiple toxigenic species found among \(C.\ difficle\) positive and \(C.\ difficle\) negative samples, divided by sample group. Mean comparison p-values calculated using Wilcoxon-test.

Figure 2. (A) Microbial species signature associated with CDI, healthy controls and diseased controls, as seen by LASSO model and (B) its associated AUC values for all samples as well as for single study populations, when comparing CDI to healthy and diseased controls combined (left) and CDI with healthy controls only (right). In the latter, the AUC value of Vincent_2016 is left intentionally blank as only diseased controls (D-Ctr) are available for this study population (see Supplementary Table 1).

Figure 3. (A) Overview of the global dataset collection composed of 42,814 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 (66.2%, n=28,347) 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 42,814 samples.

**Figure 4.** (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 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.

**Figure 5.** 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.

**Supplementary figure 1.** (A) Prevalence of C. difficile and other antibiotic-associated diarrhea (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”, 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 diarrhea (AAD) are highlighted in red. C. difficile is not included in the analysis. Study and age group were considered as nuisance variables (and modeled as random effects). To be noted that Enterobacter sp. (ref_mOTU_0036) includes Shigella flexneri and Escherichia coli.

Supplementary figure 4. (A) 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). (B) 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 v2.0 taxonomic profiles.

Supplementary figure 5. 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 \textit{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 \textit{C. difficile} in stools across the first four years of life. Only full-term infant samples were included.

Supplementary figure 9. Prevalence of \textit{C. difficile} toxin genes in the stools of healthy and diseased subjects over lifetime, based on the detection of either one or both of \textit{C. difficile} Toxin A (TcdA) or Toxin B genes (TcdB) (see Methods for details). Prevalence is calculated on the number of \textit{C. difficile} positive samples. \textit{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 \textit{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 Continents with at least five \textit{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. See Methods for details on age group definitions.

Supplementary figure 11. Species with co-occurrence (logOR>0) or co-exclusion (logOR<0) pattern with \textit{C. difficile} across gut stool samples from humans (n=14,095) (A) and animals (n=3,967) (B). Labelled species are highlighted in aquamarine and violet, for humans and animals respectively. \textit{Clostridium paraputrificum} and \textit{Clostridium neonatale} were among the very few species significantly associated with \textit{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.

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

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 Table 4-5. Samples metadata before and after timeseries dereplication.

Supplementary Table 6. Samples filtered taxonomic profiles, as provided by mOTUs v2.0.
References


34. Schubert, A. M. et al. Microbiome data distinguish patients with Clostridium difficile infection


38. Deane, J. et al. 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. J. Infect. 82, 363–370 (2021).


44. Thevenon, O., Adema, W. & Clarke, C. Background brief on fathers’ leave and its use.


55. Burnham, C.-A. D. & Carroll, K. C. Diagnosis of Clostridium difficile infection: an ongoing


75. Davis, M. Y., Zhang, H., Brannan, L. E., Carman, R. J. & Boone, J. H. Rapid change of fecal microbiome and disappearance of Clostridium difficile in a colonized infant after


