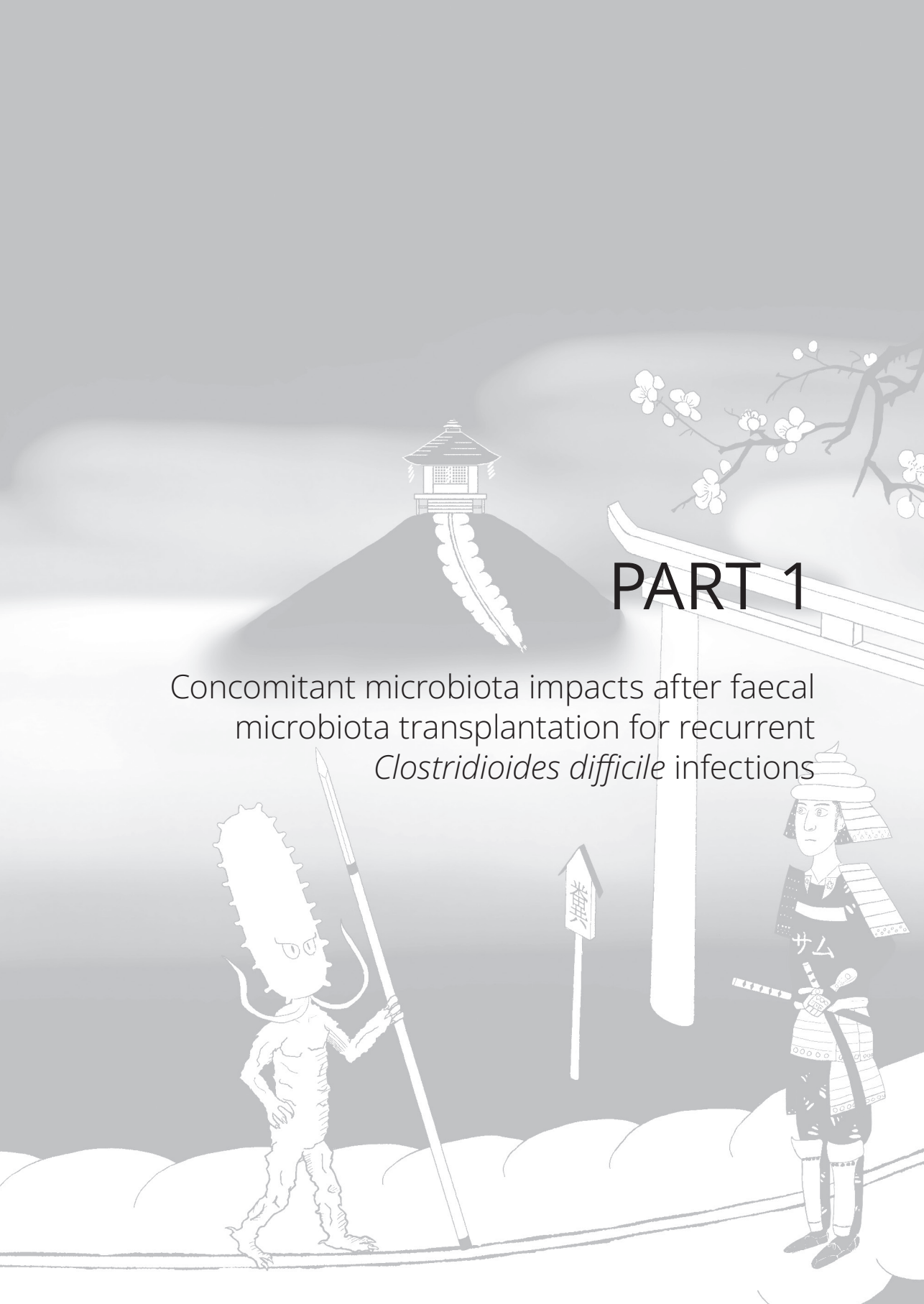
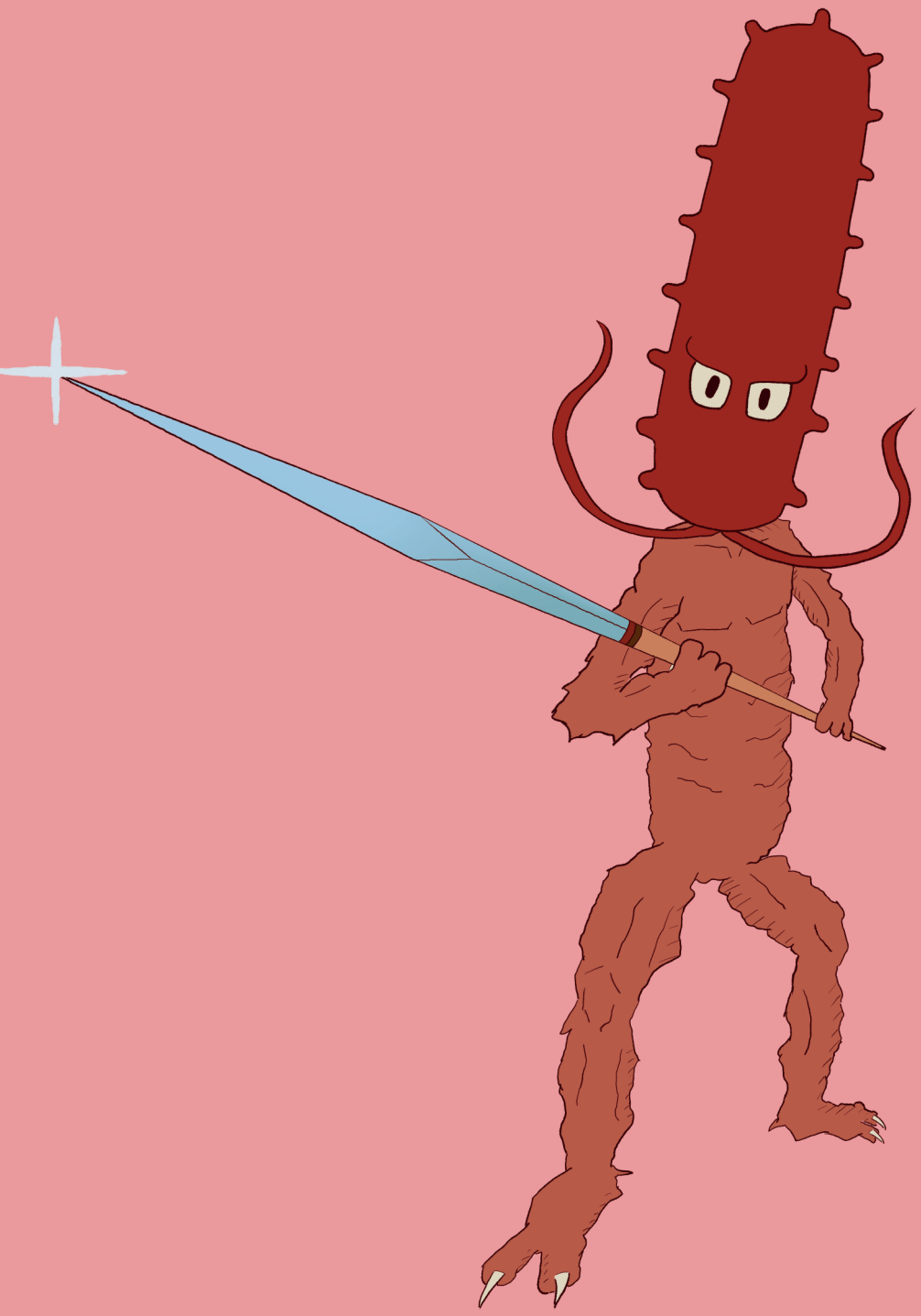
Guardians of the gut: harnessing bioinformatics to study the gut microbiome and faecal microbiota transplantation in intestinal disorders

Nooij, S.



PART 1

Concomitant microbiota impacts after faecal microbiota transplantation for recurrent *Clostridioides difficile* infections



Chapter 2.1

Faecal microbiota transplantation influences procarcinogenic *Escherichia coli* in recipient recurrent *Clostridioides difficile* patients

Sam Nooij[1,2,3], Quinten R. Ducarmon[1,3], Jeroen F.J. Laros[3,4,5,6], Romy D. Zwittink[1,3], Jason M. Norman[7], Wiep Klaas Smits[1,3], Hein W. Verspaget[8,9], Josbert J. Keller[8,10], Elisabeth M. Terveer[1,2,3], Ed J. Kuijper[1,2,3,6] on behalf of the working group of the Netherlands Donor Feces Bank

[1]Experimental Bacteriology, Department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands

[2]Netherlands Donor Feces Bank, Leiden, Netherlands

[3]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, Leiden, the Netherlands

[4]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

[5]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands

[6]National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

[7]Vedanta Biosciences Inc., Cambridge, Massachusetts, United States of America

[8]Department of Gastroenterology, Leiden University Medical Center, Leiden, the Netherlands

[9]Department of Biobanking, Leiden University Medical Center, Leiden, the Netherlands

[10]Department of Gastroenterology, Haaglanden Medical Center, The Hague, the Netherlands

Gastroenterology, Volume 161, Issue 4, 1218 - 1228.e5, October 2021

Abstract

Background and aims

Patients suffering from multiple recurrent *Clostridioides difficile* infections (rCDI) have a disturbed gut microbiota, which can be restored by faecal microbiota transplantation (FMT). Despite extensive screening, healthy faeces donors may carry bacteria in their intestinal tract that could have long-term health effects, such as potentially procarcinogenic *pks*$^{+}$ *Escherichia coli*. Here, we aim to determine whether the *pks* abundance and persistence of *pks*$^{+}$ *E. coli* is influenced by *pks*-status of the donor faeces.

Methods

In a cohort of 49 rCDI patients treated with FMT and matching donor samples – the largest cohort of its kind, to our knowledge – we retrospectively screened faecal metagenomes for *pks*$^{+}$ *E. coli* and compared the presence of *pks* in patients before and after treatment, and to their respective donors.

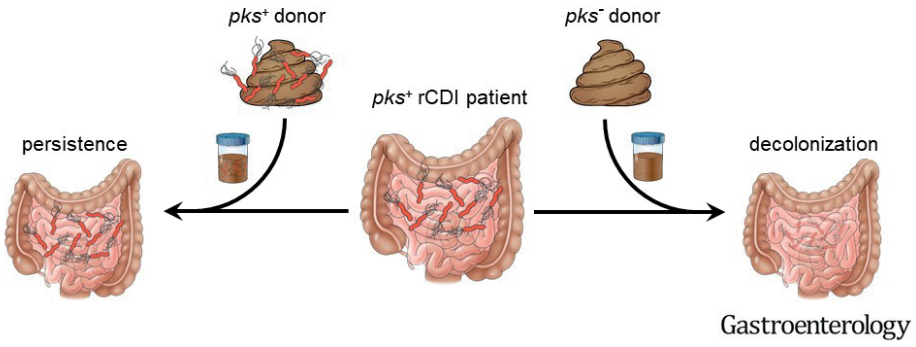
Results

The *pks* island was more prevalent ($P = .026$) and abundant ($P < .001$) in rCDI patients (pre-FMT; 27/49 = 55%, median: 0.46 reads per kilobase per million (RPKM) *pks*) than in healthy donors (3/8 donors (37.5%), 11/38 samples (29%), median: 0.01 RPKM *pks*). The *pks*-status of patients post-FMT depended on the *pks*-status of the donor suspension by which the patient was treated ($P = .046$). Particularly, persistence (8/9 cases) or clearance (13/18) of *pks*$^{+}$ *E. coli* in *pks*$^{+}$ patients was correlated to *pks* in the donor ($P = .004$).

Conclusion

We conclude that FMT contributes to *pks*$^{+}$ *E. coli* persistence or eradication in rCDI patients, but that donor-to-patient transmission of *pks*$^{+}$ *E. coli* is unlikely.

Graphical abstract



Introduction

Multiple recurrent *Clostridioides difficile* infection (rCDI) may occur as a complication after *C. difficile* infection (CDI) and is associated with a disturbance of the colonic microbiota.[1–3] Patients with rCDI can be treated with faecal microbiota transplantation (FMT), which has a cure rate of up to 89%.[1,4] The mechanism of action appears to be restoration of a healthy microbiome.[5] The donor faecal microbiota derives from carefully screened, healthy donors.[4,6,7] Donor screening focuses on the prevention of transfer of pathogens and the potential transmission of microbiota-associated disorders. In recent years there has been increasing attention for the causative role of gut bacteria in the development of colorectal cancer, which may pose a long-term risk.[8]

In 2019, Drewes and colleagues studied transmission and clearance of putative procarcinogenic bacteria, including colibactin-encoding *Escherichia coli*, by FMT in eleven paediatric rCDI patients.[9] Colibactin-producing *E. coli*, also known as *pks*$^+$ *E. coli*, has gained much attention lately, because it can be present in healthy and diseased people, has been used as a probiotic (strain Nissle 1917) and is now suspected to contribute to colorectal carcinogenesis.[10–12] The *pks*$^+$ *E. coli* carries a polyketide synthase (*pks*) gene island of about 54 kb long with 19 genes that encode the machinery to produce the nonribosomal peptide-polyketide hybrid genotoxin colibactin.[13–15] Colibactin can induce double-strand DNA breaks, which cause specific mutational signatures found in colorectal carcinomas.[10,13–15]

Here, we determine the effect of FMT from healthy donors on *pks*$^+$ *E. coli* in rCDI patients by retrospectively screening 49 matching triplets of deep-sequenced faecal metagenomes. Our results show that patients carry higher levels of *pks*$^+$ *E. coli* than donors and that the potential procarcinogen is more likely to persist when *pks*$^+$ donor faeces is used for FMT. When donor material is used with non-detectable levels of *pks*, FMT can contribute to clearance of *pks*$^+$ *E. coli*. Thus, FMT is unlikely to lead to transmission of *pks*$^+$ *E. coli*, but can reduce the levels of this procarcinogenic bacterium in rCDI patients.

Materials and methods

Faecal Microbiota Transplantation treatment

After careful screening of a multidisciplinary FMT expert panel, 49 patients suffering from multiple recurrent CDI were treated with FMT with a faecal suspension provided by the Netherlands Donor Faeces Bank. The complete procedure of donor recruitment, selection and screening, faecal suspension processing, evaluation of FMT requests, FMT treatment and patient follow-up has been described before.[4,16] In short, these patients received a vancomycin treatment for a minimum of four days (median: 12 days) until 24

hours before FMT. Eleven patients had a shorter antibiotics course or an unrecorded duration. Eight patients were treated with fidaxomicin or an unrecorded antibiotic. FMT was administered as a single treatment by a nasoduodenal tube, except in two patients that received a single infusion by colonoscopy. Differences in antibiotic pre-treatment regimen and mode of FMT were not found to influence the study results. The 198 mL faecal suspensions contained 60 gram of donor faeces in saline solution with 10% v/v glycerol and were stored until use at -80°C. All donors are healthy individuals between the ages of 18 and 60 and are extensively screened for disorders associated with microbiota and potential transmissible diseases.

Sample preparation and sequencing

Stool samples of patients were collected for sequencing approximately one day prior to FMT, after antibiotic pre-treatment, and three weeks after FMT. Stool samples of donors used for FMT treatment were also used for sequencing. Stool samples were stored at -80 °C. DNA extraction and sequencing was provided by DNA Genotek Inc. (Ottawa, Ontario, Canada). In short, DNA was extracted using the automated KingFisher™ Flex platform (Thermo Fisher Scientific Inc., Waltham, MA, USA) and MoBio PowerMag® +ClearMag® microbiome RNA/DNA isolation kit (QIAGEN Inc., Hilden, Germany). Sequencing libraries were prepared using Illumina's NexteraXT protocol and libraries were sequenced on the Illumina NextSeq platform to a target depth of 10 million reads per sample (2x150bp paired-end; Illumina, San Diego, CA, USA).

For patients P2 and P5 no matching donor metagenome was available, so instead they were matched to donor metagenomic data from samples taken 19 and 7 days later than those used for FMT, respectively.

Sequence analysis

Raw metagenomic reads were randomly subsampled to 10 million read pairs to facilitate *de novo* assembly using BBtools' reformat.sh (v. 38.79; parameters: 'samplereadstarget=10000000 sampleseed=20'; http://sourceforge.net/projects/bbmap/). The subsampled datasets were then analysed with the assembly-based workflow Jovian version 0.9.6,[17] using the NCBI BLAST nt database from 13 March 2020 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). In short, Jovian trims raw reads (trimmomatic v. 0.38,[18] parameters 'PE SLIDINGWINDOW:5:30 MINLEN:50'), removes human reads by mapping to the human genome (bowtie2 v. 2.3.4.3,[19] parameters '--local', samtools v. 1.9 and bedtools v. 2.27.1;[20,21] reference genome 'GRCh38.p7' (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.33/)), assembles filtered reads into scaffolds (SPAdes v. 3.11.0,[22] parameters '--meta --only-assembler -k 21,33,55,77') and filters scaffolds to a minimum length of 500 nucleotides, classifies scaffolds taxonomically using a Lowest Common Ancestor approach (BLAST v. 2.9.0 parameters '-evalue 0.05 -max_target_seqs 250, -max_hsps 1',[23] MGKit v. 0.3.4 using quantile threshold .97 and bitscore threshold 100)[24] and quantifies scaffolds by mapping filtered reads back to the

scaffolds (BWA-mem v. 0.7.17,[25] default parameters). From Jovian, we used 1) the quality-trimmed reads from which human reads are removed, 2) assembled scaffolds and 3) taxonomic classifications and quantifications of scaffolds. These reads and scaffolds were screened for *Escherichia coli* and the *pks* island (19 *clb* genes from GenBank accession ID AM229678) using the custom workflow 'Jovian screener' (v. 0.1).[26] This screening extension to Jovian works in three ways. 1) It extracts the species of interest from the taxonomic profile results table (based on the 'tax_name' column). 2) It looks for the (gene) sequences of interest in all scaffolds using BLAST (v. 2.9.0, parameters '-task blastn -perc_identity 75 -qcov_hsp_perc 0 -evalue 1E-20 -culling_limit 1 -max_hsps 1'), which are then filtered to include only matches that cover at least half (50%) of the sequences of interest. And 3) it maps trimmed and filtered reads to the sequences of interest (BWA v. 0.7.17, samtools v. 1.10), after which mapped reads are deduplicated (using Picard MarkDuplicates v. 2.23.3 (https://broadinstitute.github.io/picard/)) and counted (samtools v. 1.10). The resulting tables were analysed and visualised using R version 4.0.2 (Rmarkdown scripts online at https://git.lumc.nl/snooij/pks-in-rcdi-metagenomes-analysis).[27]

We renamed the 'hypothetical protein (clb)' from AM229678 '*clbS*', because it is identical to accession number KX683217 (Escherichia coli c*lbS* gene, complete CDS). Reads per kilobase per million (RPKM) values were calculated by dividing the number of mapped reads by the length of the respective gene (coding sequence), multiplied by 1,000, divided by the number of trimmed and filtered reads, multiplied by one million.

Pilot screening for other procarcinogenic bacteria

Next to *pks* from *E. coli*, we also screened metagenomes for the presence of *Fusobacterium nucleatum* adhesin A (*fadA*; accession ID AVQ22939), *Bacteroides fragilis* toxin (*bft* 1-3; accession IDs AB026624, AB026625 and AB026626), and *Campylobacter jejuni* cytolethal distending toxin (*cdtA, cdtB, cdtC*; acccession ID AB274791). The *fadA* and *cdt* genes were not detected, and *bft* was detected in two patients (P7 pre- and post-FMT, and P42 post-FMT). As these genes are almost completely absent from the current dataset and absent from donors, we did not analyse these further.

Improving the identification of *pks* by combining data from all 19 *clb* genes

To improve the sensitivity of screening for the *pks* island in metagenomics data compared to using a single marker gene only (e.g. *clbB*), we assessed whether the genes always co-occur and may be considered as one unit. In assembled scaffolds, the *clb* genes are mostly assembled on one or a few scaffolds when depth is sufficiently high, suggesting consistent close genomic proximity (Supplementary File 1). For the mapping approach, we compared the RPKM values of *clb* genes in each metagenomics sample with one another and calculated Spearman correlations (Supplementary Figures S1-3). Correlations between *clb* genes are high (r = .54-.99, P < .01), and also higher than

correlations with flanking regions and more distant *E. coli* marker genes. Therefore, we conclude that the *pks* island is present in genomes as complete island and for its detection and quantification we combine all hits to *clb* genes to calculate average *pks* values.

Selecting representative *E. coli* marker genes

To calculate the abundance of *E. coli* in a similar way as the *pks* genes to facilitate the estimation of *pks* island copies per genome, we also mapped the reads to a number of putative single-copy marker genes. The putative marker genes are *dxs*, *rodA* and *uidA* (accession IDs: AF035440, M22857, and S69414, respectively),[28,29] MultiLocus Sequence Typing (MLST) genes *adk, fumC, gyrB icd, mdh, purA* and *recA* from Enterobase,[30] and *adk, gcl, zwf, mdh, metA* and *ppk* by Adiri and colleagues.[31] Based on the depth of coverage, we found that the MLST genes sometimes showed uneven coverage – especially in metagenomes with more non-*E. coli* Enterobacteriaceae. This indicates that these genes may have regions that are homologous to genes from other species, so the MLST genes were excluded. Next, we compared RPKM values for *dxs, rodA* and *uidA* to the relative abundance of *E. coli* and Enterobacteriaceae in all analysed metagenomes. The *uidA* gene has the best correlation with *E. coli* abundance (Supplementary Figure S4A; Spearman correlation $r = .97$, $P < 2.2e\text{-}16$). The *dxs* and *rodA* genes also appear in metagenomes with no *E. coli* (Supplementary Figure S4 B-C) and show better correlations with Enterobacteriaceae (Supplementary Figure S4 E-F, Spearman correlations $r = 0.96$ and $r = 0.97$, respectively). Even though *uidA* is the most specific marker gene, it is not sufficiently present in all metagenomes. Therefore, we calculated the average RPKM of *dxs*, *rodA* and *uidA*, whichever are present, to represent the quantity of *E. coli* genomes in the metagenomes, which was then used to calculate the ratio of the average *pks* island copies per *E. coli* genome.

Comparing *E. coli* assembled scaffolds between subjects

We assessed whether our data allows us to distinguish within-species variants of *E. coli* that could be used to conclusively demonstrate transfer or persistence. By using average nucleotide identity (pyANI version 0.2.10, parameter '-m ANIb')[32] we compared *E. coli* scaffolds from patients with *pks*$^+$ *E. coli* before and after FMT and their donors, and *E. coli* from patients that had no *pks* before FMT but were *pks*$^+$ after FMT and the corresponding donors. Average nucleotide identities (ANI) between *E. coli* scaffolds from different patients were as high as within patients pre- and post-FMT, so they could not be used to support persistence. Furthermore, donor scaffolds were too short to make a reasonable argument for transmission based on ANI. Where possible, *E. coli* assemblies were typed *in silico* using the multi-locus sequence typing (MLST) software 'mlst' (https://www.github.com/tseemann/mlst) and the PubMLST database (downloaded on 14 April 2021)[33]. MLST was successful only for 20 patients and for two patients provided result for both pre-FMT and post-FMT (Supplementary Table S1). These two patients, P21 and P36, had the same sequence type before and after

FMT, suggesting persistence of the same dominant variant. Both these patients lost *pks* after FMT, so these sequence types do not carry *pks*.

Data availability

All raw metagenomics data, from which human-derived reads have been excluded by mapping to a human reference genome (bowtie2 v.2.4.2, using parameter '--very-sensitive-local' to the GRCh38.p7 assembly), have been deposited in the European Nucleotide Archive under accession number PRJEB44737.

Results

Study characteristics

We screened the faecal metagenomes of 49 rCDI patients, pre- and post-FMT, and 38 corresponding metagenomes from donor faeces, of eight donors of the Netherlands Donor Feces Bank, for *pks*$^+$ *E. coli* via read mapping against a *pks* island reference (GenBank accession ID AM229678) and *E. coli* single-copy marker genes (*dxs*, AF035440; *rodA*, M22857; *uidA*, S69414). Eleven donor suspensions were used to treat two patients each, thus yielding 49 sample triplets. The patients were between 27 and 92 years old (median: 73.5), including 31 females and 18 males. Only patients were included that had been treated with a single FMT and for which pre- and post-FMT samples were available. The donors were between 24 and 46 years old (median: 33); four females and four males. Samples were collected between April 2016 and March 2018. For patients P2 and P5 no donor metagenomic data were available from the date of FMT, but from 19 and 7 days later, respectively. For other patients, corresponding donor data were available from the same sample as used for FMT. All patients were cured from rCDI, except for three patients that had a recurrence within three weeks after FMT (P12, P25 and P26).

The *pks* island is absent in most donors or detected in very low quantities

To quantify *pks* levels in metagenomes, we averaged normalised RPKM values of 19 *clb* genes (colibactin-encoding, AM229678). Subjects with zero reads mapped to *clb* genes in their metagenome are considered *pks*-negative (*pks*$^-$) and subjects with one or more reads mapped to one or more *clb* genes are *pks*$^+$. Five out of eight healthy donor metagenomes were always negative for the *pks* island (Figure 1). One donor had multiple samples with *pks*$^+$ metagenomes (D5, see below). Another donor (D2) provided only one sample, which was *pks*$^+$, and a third donor (D6) had one *pks*-positive sample out of nine used for transplantation. Levels of *pks* among positive donors range from 0.0013 to 0.0498 RPKM (median: 0.010). In total, 11 out of 38 donor samples were *pks*$^+$ (29%).

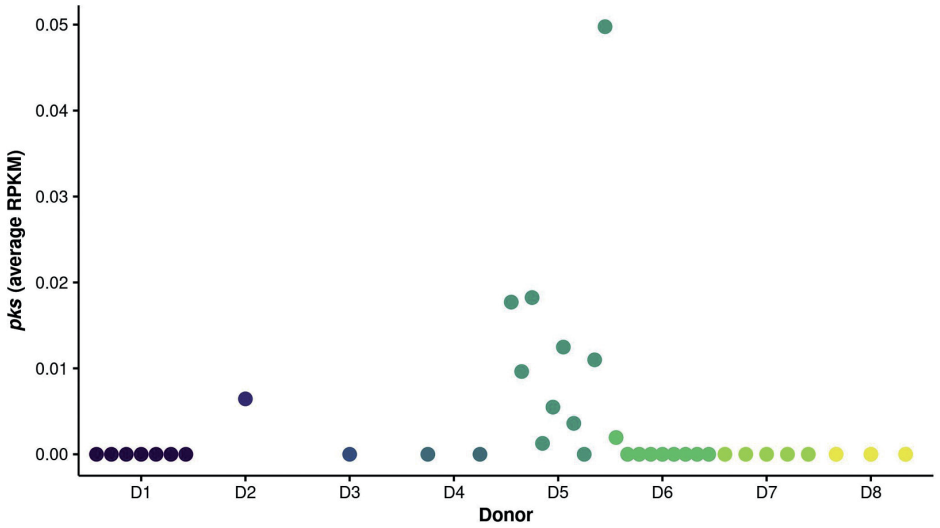


Figure 1. Levels of *pks* in donor metagenomes. Reads from donor metagenomes were mapped to the *pks* island (AM227896) to quantify the *pks* island per sample. Values are normalised by gene length and library size (reads per kilobase per million; RPKM) and the average RPKM value of the 19 *clb* genes is calculated as average *pks*. Out of eight donors, three had *pks*⁺ samples of which only one had multiple *pks*⁺ samples. Out of 38 donor samples included, 11 were *pks*⁺.

Healthy donors may persistently carry *pks*

To study the temporal dynamics of *pks* colonisation of donors, we evaluated longitudinal samples from one donor that was consistently *pks*$^+$ (D5). Out of 15 metagenomes total from donor D5, of which 10 were used for transplantation, 13 had traces of the *pks* island (Figure 2A). Both the first and last samples were positive, suggesting that this donor was persistently colonised with a *pks*$^+$ bacterium during the donation period of six months. Levels of *pks* range from 0 to .05 RPKM and correlate loosely with the relative abundance of *E. coli* (Spearman r = .43, P = .11) or Enterobacteriaceae (Spearman r = .1, P = .72). *E. coli* was not always detected in the metagenomes of donor D5 (Figure 2B-C). It seems most likely that due to the low abundance of *E. coli*, the sequencing depth in the metagenome was low, yielding only short genomic regions that are shared between *E. coli* and other Enterobacteriaceae resulting in these being classified as Enterobacteriaceae.

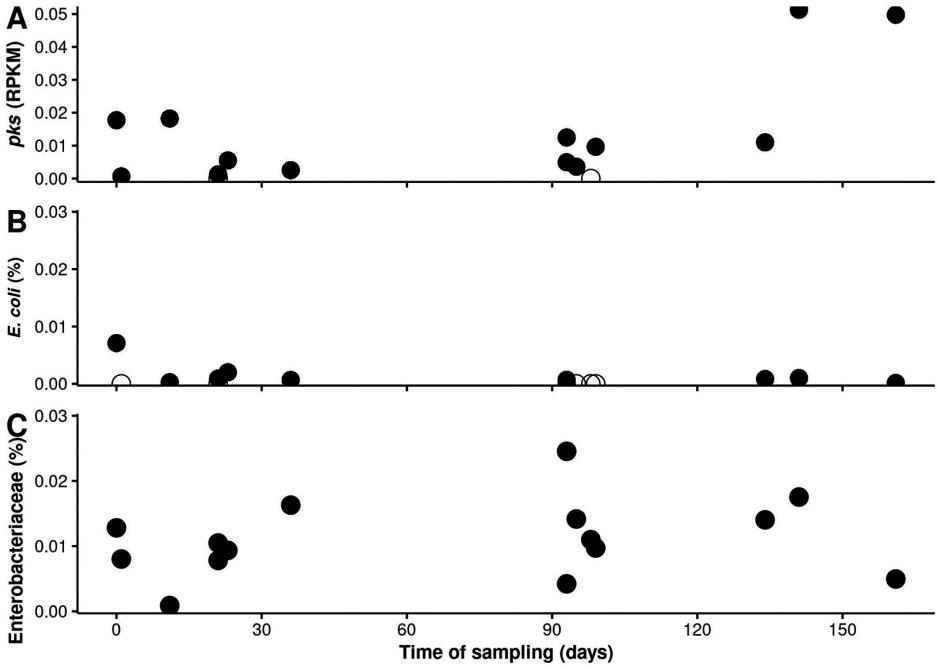


Figure 2. Levels of *pks* levels and relative abundances of *E. coli* and Enterobacteriaceae over time in one healthy donor. To gain more insight in the duration of colonisation by *pks*$^+$ *E. coli* in healthy donors, we quantified average *pks* levels of all available metagenomes of D5 by read mapping and we determined the relative abundances of *E. coli* and Enterobacteriaceae for each metagenome using Jovian. A) Average normalised values of *pks* island (*clb* genes) in donor D5 over approximately six months' time. Samples from days 21 and 93 are duplicates; both are shown. *Pks* levels are generally very low, near the limit of detection. The donor seems to have been persistently colonised by a *pks*$^+$ species for at least half a year, even though two samples in the first and fourth months were *pks*$^-$. B-C) Relative abundances of *E.* coli and Enterobacteriaceae (including *E. coli*) in each metagenomic sample of donor D5. For the duplicate samples only the higher values are shown. There is no clear correlation between either abundance and the level of *pks*. Empty circles denote that no *pks* or *E. coli* was detected at that time point.

The *pks* island is frequent and abundant in pre-FMT rCDI patient metagenomes

Genes from the *pks* island were detected in 27/49 patients before FMT treatment (55%), and 19/49 patients after FMT (39%). Overall, *pks* was more prevalent in pre-FMT rCDI patient metagenomes than in donor metagenomes (χ^2, $P = .026$, odds ratio = 2.246, 95% confidence interval = 1.010–4.589). Patient metagenomes reached *pks* levels up to 59.7 RPKM, corresponding to roughly 0.3% of all analysed reads. The median *pks* levels of *pks*$^+$ patient metagenomes were 0.461 RPKM and 1.348 RPKM, respectively for pre- and post-FMT. The *pks* level in *pks*$^+$ patient metagenomes was significantly higher than in donors (pairwise Wilcoxon rank sum exact test, $P < .001$ for both pre- and post-FMT), but was not different between pre- and post-FMT ($P = .66$; Figure 3A). When including

the *pks*-negative metagenomes in the comparison, the differences are less apparent and *pks* levels of donors only differ from patients before FMT (Supplementary Figure S5).

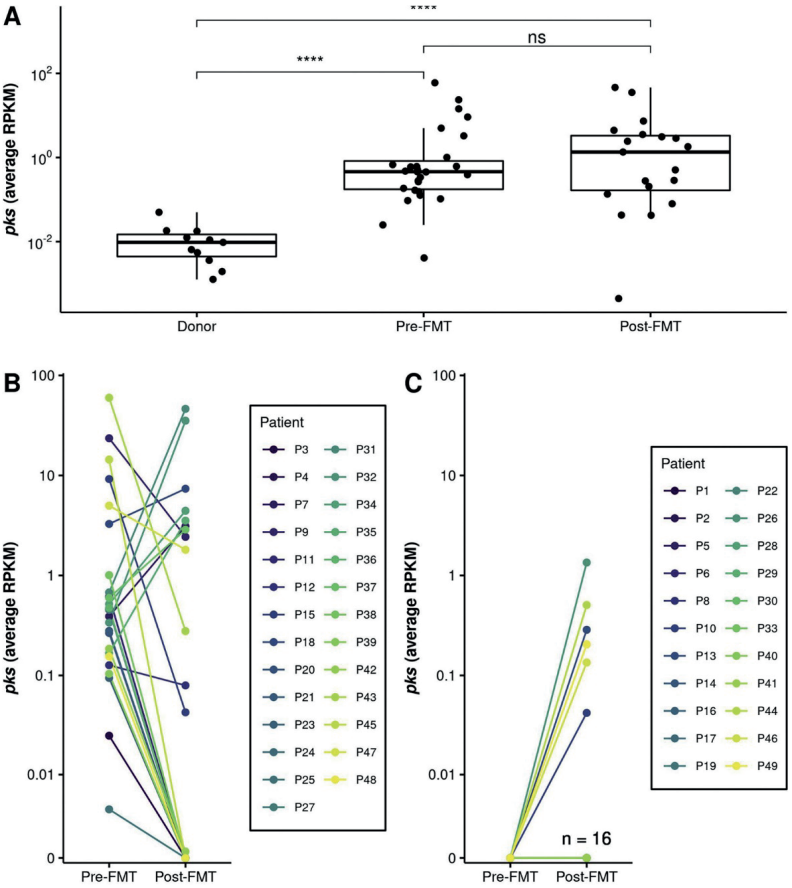


Figure 3. Levels of *pks* levels in *pks*+ faecal donor and rCDI patient metagenomes before and after FMT treatment and changes in *pks* in patients after FMT. To assess the effect of FMT on the *pks* levels in patients, we mapped reads to a *pks* island reference and quantified *pks* by normalising read counts by gene length and library size (RPKM), and averaged values over all 19 *clb* genes. Y-axes use a log10-scale. A) Levels of *pks* in *pks*$^+$ donor samples were lower than those in patients pre- and post-FMT (Wilcoxon rank sum test, $P < .001$). Levels of *pks* were not different between those found in patients pre-FMT and those post-FMT. B) Changes in *pks* levels in patients that were *pks*$^+$ before FMT. Twenty-seven patients were *pks*$^+$ prior to FMT. Twenty of them have lower *pks* levels after FMT, of which 14 lost *pks*. Seven patients had increased levels of *pks* after FMT. C) Changes in FMT levels in patients that were *pks*$^-$ before FMT. Twenty-two patients had no detectable *pks* prior to FMT. Six patients acquired *pks* post-FMT and sixteen remained *pks*$^-$. The number of patients with zero *pks* after FMT is written in the figure to clarify the overlapping dots.

Fecal microbiota transplantation changes *pks* levels in *pks*$^+$ patients with recurrent *Clostridioides difficile* infection

As *pks* levels – when detected – are much higher in rCDI patients than in healthy donors (Figure 3A), we determined whether treatment by FMT leads to a reduction in *pks* levels in the subset of *pks*$^+$ rCDI patients. Overall, there is no significant difference between *pks* levels pre- and post-FMT in patients that had *pks* prior to FMT (paired Wilcoxon signed rank exact test, $P = .2901$). However, when stratified by *pks*-status of the donor we find that the change of *pks* level in the rCDI patient after FMT depends on the *pks*-status of the donor (Fisher's exact test, $P = .0006$).

Twenty-seven patients' metagenomes were *pks*$^+$ prior to FMT treatment, and *pks* levels varied more than thousandfold. Twenty (74%) had decreased levels of *pks* after FMT, of which 14 (52%) were reduced to levels below our detection limit (Figure 3B). Eighteen out of these 20 (90%) were treated with a *pks*$^-$ donor suspension (Table 1, Supplementary Figure S6). In seven cases the level of *pks* increased, of which six were treated with a *pks*$^+$ donor suspension (86%).

Table 1. Summary of *pks*-status for each step in the FMT treatments of the 49 included patients. Forty-nine patients, of which 22 were *pks*$^-$ and 27 *pks*$^+$ prior to FMT, were treated with 35 *pks*$^-$ and 14 *pks*$^+$ donor suspensions. After FMT, 30 patients are *pks*$^-$ and 19 are *pks*$^+$.

Donor suspensions	Patients pre-FMT	Patients post-FMT
35 *pks* -	17 *pks* -	12 *pks* -
		5 *pks* +
	18 *pks* +	13 *pks* -
		5 *pks* +
14 *pks* +	5 *pks* -	4 *pks* -
		1 *pks* +
	9 *pks* +	1 *pks* -
		8 *pks* +

Patients who were *pks*-negative before FMT mostly remained negative

We assessed the effect of FMT on patients that had no detectable levels of *pks* before FMT to evaluate the possibility of transmission from a *pks*$^+$ donor. Overall, as expected with a conversion of *pks*-status in a *pks*-negative population, the *pks* level increased after FMT (paired Wilcoxon signed rank test, $P = .03603$). Twenty-two patients had no detectable *pks* prior to FMT. Sixteen (73%) of them remained negative after FMT and six patients acquired *pks* after FMT (Table 1). Interestingly, from the six patients who became *pks*$^+$ after FMT, five were treated with a *pks*$^-$ donor suspension derived from a total of three different donors (83%; Figure 3C, Supplementary Figure S6). Out of

the 16 patients that remained negative, 12 were treated with *pks*$^-$ donor suspensions (75%). The average *pks* level of the six patient metagenomes that acquired *pks* after FMT is 0.420 RPKM (median: 0.245 RPKM). The *pks* levels in patients who acquired *pks* vary more than tenfold and are generally low compared to *pks* levels in patients with persistent *pks* (median: 2.874 RPKM), although the difference is not significant (Wilcoxon rank sum test, $P = .08742$). Importantly, however, whether a patient in this group of *pks*$^-$ rCDI patients demonstrated increased *pks*-levels after FMT is not significantly correlated with the *pks*-status of the donor (Fisher's exact test, $P = 1$), suggesting that an increase in *pks* is as likely to derive from the patient itself as from the donor.

Persistence or clearance of *pks* in patients after FMT depends on presence of *pks* in donors

In addition to testing the effect of FMT on the quantity of *pks* in patients, we assessed the likelihood of *pks* persistence, clearance and transmission based on the presence of *pks* in patients and their donors. Thirty-five rCDI patients were treated with a *pks*$^-$ donor suspension of which twenty-five were *pks*$^-$ after FMT (71%; Table 1) and four had reduced *pks* (11%; Supplementary Figure S6). Fourteen patients were treated with a *pks*$^+$ donor suspension and nine patients were *pks*$^+$ after FMT (64%). Based on these numbers, we find a direct dependence between the post-FMT *pks*-status and *pks* in the donor suspension (χ^2 with Yates' continuity correction, $P = .046$). The effect of FMT on post-FMT *pks*-status was strongest in patients that were already colonised with a *pks*$^+$ *E. coli* before FMT (Fisher's exact test, $P = .004$). Out of 27 patients that were *pks*$^+$ pre-FMT, 9 were treated with a *pks*$^+$ donor suspension of whom 8 remained *pks*$^+$ post-FMT. Of the 18 patients who were treated with a *pks*-free donor suspension, 5 contained *pks* post-FMT. There was no such effect in patients in which we detected no *pks* pre-FMT (Fisher's exact test, $P = 1$). Although our metagenomic analyses could not demonstrate persistence or transmission using within-species variants (see Methods), taken together, these results indicate that FMT contributes to persistence or clearance rather than donor-to-patient transmission of *pks* in rCDI patients.

Variable within-species *pks* island prevalence suggests persistence of *pks*$^+$ *E. coli* in patients

Using single-copy marker genes *dxs*, *rodA* and *uidA* (encoding D-1-deoxyxylulose 5-phosphate synthase, rod-shape determining protein, and beta-glucuronidase, respectively) to quantify *E. coli* genomes, we calculated the average number of *pks* island copies per *E. coli* genome, or prevalence of *pks* among *E. coli* variant subpopulations. *E. coli* is detected in less than half of the donor metagenomes (14/38; full circles in Figure 4A) and is detected in five out of eleven *pks*$^+$ metagenomes (from donors D2 and D5). *E. coli* markers are present in nearly all patients pre-FMT (48/49; Figure 4B) and *pks* per *E. coli* ratios that are often near 0 indicate that *pks*$^+$ *E. coli* variants are often a minority variant. Post-FMT, still many patients harbour *E. coli* species (46/49;

Figure 4C) and when *pks* is present the *pks* per *E. coli* ratios are often higher. Taken together, the results suggest that rCDI patients treated with vancomycin or fidaxomicin before FMT may have high abundances of different *E. coli* subpopulations or variants. When a patient is colonised with a *pks*$^+$ *E. coli*, the *pks*$^+$ *E. coli* is often a minority of *E. coli* variants. After treatment by FMT using a donor suspension that contains little if any *E. coli*, the number of different *E. coli* variants decreases. And when a *pks*$^+$ variant is present post-FMT, this is often a larger fraction or the only *E. coli* variant, possibly indicating that this variant is more persistent than other *E. coli* variants.

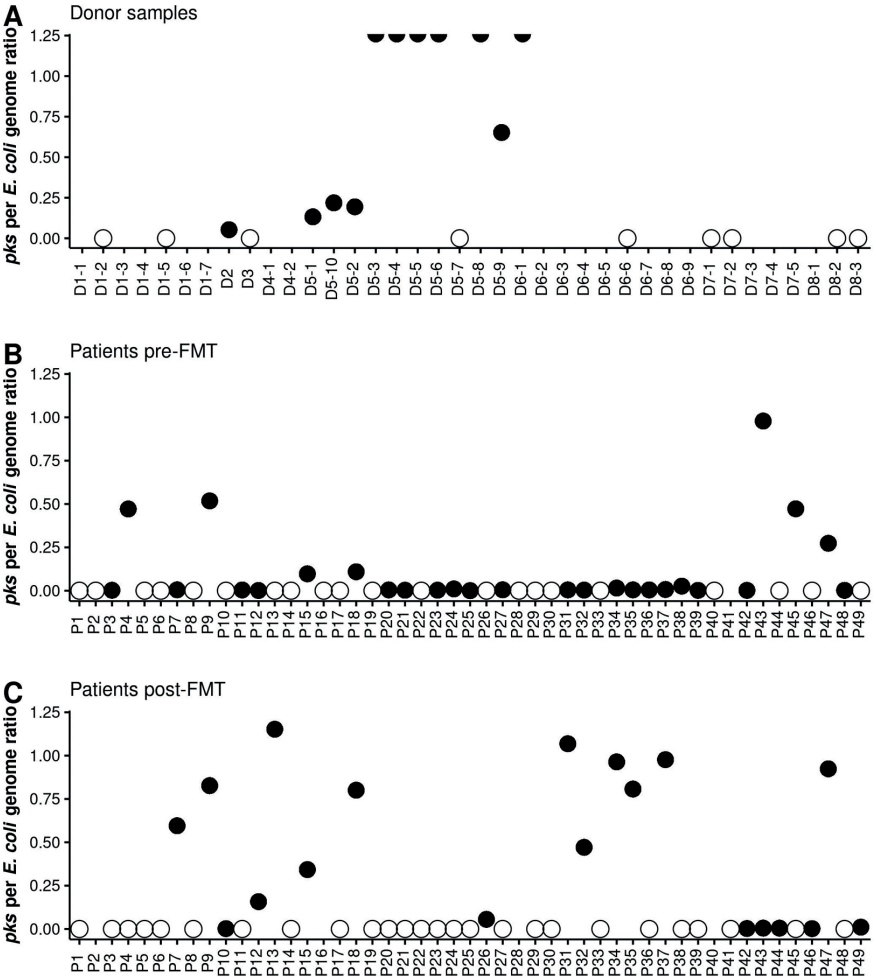


Figure 4. Average number of *pks* island copies per *E. coli* genome, based on three marker genes. Average *pks* island abundances per sample are calculated as the average RPKM value of all 19 *clb* genes. *E. coli* genome abundances are estimated based on the average RPKM value of the single-copy marker genes *dxs*, *rodA* and *uidA*. Empty circles indicate presence of *E. coli* and absence of *pks*, black semicircles at the top of a figure indicate presence of *pks* and absence of *E. coli* and missing circles indicate neither *E. coli* nor *pks* was detected in the metagenome. A) *E. coli* is detected in fourteen donor samples of which five samples from two different donors are *pks*⁺. In these samples, there are between 0.05 and 0.65 *pks* island copies per *E. coli* genome. Six donor metagenomes have *pks* without *E. coli* markers. B) Nearly all patients have *E. coli* in their faecal metagenome pre-FMT (all but P41). *pks* per *E. coli* ratios vary from zero (*E. coli* without the *pks* island) to one (all *E. coli* genomes in the sample have the *pks* island). In most *pks*⁺ cases, it is a minority of *E. coli* genomes that have the *pks* island. C) After FMT, *E. coli* is still detected in most patients and the average number of *pks* copies per *E. coli* genome varies from zero to approximately one. Fewer patients are *pks*⁺ after FMT and *pks* per *E. coli* ratios are often higher: in nine cases the majority of *E. coli* genomes has the *pks* island.

Discussion

In this cohort study of 49 FMT-treated adult rCDI patients and their corresponding healthy faeces donors, we screened faecal metagenomes for the presence of *pks*$^+$ *E. coli* to assess its prevalence and the effect of FMT. To our knowledge, this is the first study to employ metagenomic screening for potentially procarcinogenic bacteria in the largest FMT cohort to date. We find that *pks* was present in a minority of healthy donor samples and, when present, *pks* levels were low. Of eight donors, one donor was likely persistently colonised with a *pks*$^+$ bacterium for six months. In contrast, rCDI patients had high levels of *pks* and high abundances of *E. coli* in their metagenomes. FMT resulted in changes in both status and levels of *pks* detected in the deep-sequenced metagenomes. In particular, we find that treatment of *pks*$^+$ rCDI patients with a *pks*$^-$ donor suspension reduces *pks* levels, in some cases to below the detection limit, whereas treatment with a *pks*$^+$ donor suspension leads to persistence of *pks*$^+$ *E. coli*.

A previous study elucidating the effects of FMT on the presence of intestinal procarcinogenic bacteria in eleven paediatric rCDI patients reported similar findings.[9] Drewes and colleagues found that FMT supported decolonisation depending on the microbiota of both patient and donor. We find that FMT supports decolonisation of *pks*$^+$ bacteria in rCDI patients treated with *pks*$^-$ donor suspensions. Drewes and colleagues reported transmission of putative procarcinogenic bacteria during FMT. Though we cannot exclude the possibility of transmission on the basis of our data, we consider it more likely that *pks*$^+$ *E. coli* variants persist in the patients than that they are transferred from the donor, as *E. coli* is much lower in abundance in donor material than in patients, and engraftment of low-abundant bacteria in the donor is less likely.[34] A study specifically designed to address the question of transmission of procarcinogenic bacteria during FMT, using culturing and whole-genome sequencing, is needed to investigate this aspect further. Other potentially procarcinogenic species were studied previously in the context of FMT, namely enterotoxigenic *Bacteroides fragilis* and *Fusobacterium nucleatum*.[9] We also screened the current set of faecal metagenomes for the presence of these and other putative procarcinogenic bacteria (the *bft* toxin gene from enterotoxigenic *B. fragilis*, the *fadA* adhesin gene from *Fusobacterium nucleatum*, and the *cdtA-C* genes from *Campylobacter jejuni*),[35] but we only detected *bft* in two metagenomes from two patients. As copy numbers of the *bft*, *clbB* and *fadA* genes was reported to be similar in colonised patients,[9] and we readily detect *clb* genes (from the *pks* island), we do not expect this to be a technical limitation of our method, but possibly reflecting differences in the study population. As the other procarcinogenic bacteria appeared to be absent and could not be analysed further, we focused our analyses on *pks*$^+$ *E. coli*.

In a few faecal metagenomes, especially from donors, we detected *pks* in absence of *E. coli*. Possibly, *E. coli* was present below the limit of detection or the *pks* gene island

belonged to a different, pathogenic, species of Enterobacteriaceae that may also carry the *pks* island.[36] Although we detected *Klebsiella*, *Enterobacter*, and *Citrobacter* species using the assembly-based approach in six out of eleven donor samples with *pks*, but that were negative for *E. coli* based on the three single-copy marker genes, pathogenic isolates of these species are unlikely to be present in healthy donors due to thorough screening. Therefore, we consider it more likely that *E. coli* was present below the detection limit of our assay and is in fact the host of the *pks* island.

Earlier studies on procarcinogenic intestinal bacteria applied different technical approaches for detection and quantification. Often, detection of the *pks* island was based on the longest gene (*clbB*; 9.6 kb); either by PCR or by metagenomic screening.[9,11] Metagenomics-based assays require high sequencing depths to reliably detect single genes. We combine data of all 19 *clb* genes, which we show are highly intercorrelated, to improve the sensitivity of our assay and allowing us to detect very low levels of *pks*. Thereby we increased the number of *pks*$^+$ samples by eight compared to detection based on only *clbB*. The *pks* levels (as RPKM values) in our screened faeces donors are slightly lower than previously reported for healthy controls (donors: 0 RPKM median, 0.05 maximum; healthy controls: 0.06 median), while *pks* levels in *pks*$^+$ rCDI patients (pre-FMT: 0.46 median, post-FMT: 1.35 median) are higher than those found in inflammatory bowel disease patients (0.18 RPKM median).[11] This suggests that the screened and approved donors of the National Donor Feces Bank may have lower *pks* levels than average healthy controls. In addition, these rCDI patients seem to have been exposed to *pks* levels higher than those in inflammatory bowel disease patients. By comparing prevalence of *pks* in different groups we come to similar conclusions. The prevalence of *pks* among healthy donors (10 out of 38 samples; 26.3%) is comparable to other studies, which reported prevalence of *pks* among healthy controls between 20 and 30%.[10,11,36] The prevalence of *pks* among rCDI patients is higher (27/49 pre-FMT and 19/49 post-FMT; 55.1% and 38.8%) than in healthy donors. The numbers are between those reported for inflammatory bowel disease (~33%),[11] and those reported among colorectal cancer patients (~60%).[10] However, the reported prevalence is often based on fewer than 50 patients and may therefore not accurately reflect the prevalence in the general population. Regardless, the *pks* levels and prevalence we find are approximately in the same range as described before and suggest that the relatively old rCDI patients might be at an increased risk of developing colorectal cancer.

The age difference of approximately 40 years between our donors and patients is of note, as this may have been a confounding factor in the current study, assuming that age is correlated with *pks*$^+$ *E. coli* colonisation. To our knowledge, however, there has been no evidence indicating that a higher age is associated with a higher incidence of *pks*$^+$ *E. coli* colonisation. Furthermore, it has been suggested that rCDI is correlated to colorectal carcinogenesis.[37] These hypotheses may be tested in a future study by screening a cohort of healthy people of similar age as the rCDI patients and follow

them over time to compare the number of occurrences of colorectal cancer between these two groups.

In conclusion, we find that *pks* is prevalent and can be highly abundant in rCDI patients. FMT with *pks*$^-$ donor suspensions generally decreases *pks* levels of patients and *pks* persists in *pks*$^+$ patients when a *pks*$^+$ donor is used. We find no clear evidence of donor-to-patient transmission of *pks*$^+$ *E. coli* and if it were to occur, the impact may be negligible given the age and comorbidity of most patients in this particular cohort. Therefore, we think that current screening protocols for FMT donors are safe and routinely screening for *pks* is not required, at least for the treatment of rCDI. It is currently unknown how long *pks*$^+$ *E. coli* persists in cured rCDI patients and on what time-frame this may contribute to colorectal carcinogenesis. Further research addressing the long-term health effects is needed to evaluate the cancer risk due to *pks*$^+$ *E. coli* in rCDI and other patients, since this might offer opportunities for early intervention in the development of colorectal cancer.

Acknowledgments

The authors wish to thank dr. Szymon M. Kiełbasa and the other members of the interdepartmental next-generation sequencing discussion group at the Leiden University Medical Center for their advice on the analysis and statistics, and the Experimental Bacteriology group for helpful comments during work discussions. Also, the authors thank dr. Bastian V.H. Hornung for assistance with handling the dataset.

Ethics approval

Patients provided informed consent for collection of stool samples and outcome data of FMT for research purposes, which was approved by the Medical Ethics Committee at the Leiden University Medical Center (P15.145).

Contributors

SN, QRD, EMT and EJK conceived and designed the study. EJK, EMT and RDZ supervised the study. HWV, JJK, EMT and EJK were involved in the treatment of patients and collection of samples. SN, QRD and JFJL designed the analysis approach. JMN contributed materials and oversaw logistics. SN performed bioinformatics analyses and wrote the manuscript. SN, QRD, JFJL, RDZ, WKS, HWV, JJK, EMT and EJK discussed results and implications. All authors contributed to and approved the final manuscript.

Competing interests

The Netherlands Donor Feces Bank received an unrestricted grant from Vedanta Biosciences, Inc. (Cambridge, MA, USA) from 2017 to 2020. Metagenomic sequencing was funded by Vedanta Biosciences. The authors received no financial support for the research, authorship and/or publication. EJK has performed research for Cubist, Novartis and Qiagen and has participated in advisory forums of Astellas, Optimer, Actelion, Pfizer, Sanofi Pasteur and Seres Therapeutics. WKS has performed research for Cubist and has received speaker fees from Promega. JMN is an employee of Vedanta Biosciences Inc. The companies had no role in this study or writing of the manuscript. The other authors declare no competing interests.

Abbreviations used in this paper

rCDI (multiple recurrent *C. difficile* infection), FMT (faecal microbiota transplantation), *pks* (polyketide-synthase), RPKM (reads per kilobase per million), ST (sequence type)

Supplementary Material

Supplementary figures and document are available online on the publisher's website:

https://www.gastrojournal.org/article/S0016-5085(21)03120-6/fulltext#app-1

References

1. Smits WK, Lyras D, Lacy DB, et al. Clostridium difficile infection. Nature Reviews Disease Primers 2016;2:1–20.
2. Chang JY, Antonopoulos DA, Kalra A, et al. Decreased Diversity of the Fecal Microbiome in Recurrent Clostridium difficile—Associated Diarrhea. J Infect Dis 2008;197:435–438.
3. Nood E van, Vrieze A, Nieuwdorp M, et al. Duodenal infusion of donor feces for recurrent Clostridium difficile. N Engl J Med 2013;368:407–415.
4. Terveer EM, Vendrik KE, Ooijevaar RE, et al. Faecal microbiota transplantation for Clostridioides difficile infection: Four years' experience of the Netherlands Donor Feces Bank. United European Gastroenterology Journal 2020;8:1236–1247.
5. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nature Reviews Gastroenterology & Hepatology 2016;13:508–516.
6. Cammarota G, Ianiro G, Kelly CR, et al. International consensus conference on stool banking for faecal microbiota transplantation in clinical practice. Gut 2019;68:2111–2121.
7. Keller JJ, Ooijevaar RE, Hvas CL, et al. A standardised model for stool banking for faecal microbiota transplantation: a consensus report from a multidisciplinary UEG working group. United European Gastroenterology Journal 2020:2050640620967898.
8. Scott AJ, Alexander JL, Merrifield CA, et al. International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. Gut 2019;68:1624–1632.
9. Drewes JL, Corona A, Sanchez U, et al. Transmission and clearance of potential procarcinogenic bacteria during fecal microbiota transplantation for recurrent *Clostridioides difficile*. JCI Insight 2019;4.
10. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature 2020;580:269–273.
11. Dubinsky V, Dotan I, Gophna U. Carriage of Colibactin-producing Bacteria and Colorectal Cancer Risk. Trends in Microbiology 2020;0.
12. Olier M, Marcq I, Salvador-Cartier C, et al. Genotoxicity of Escherichia coli Nissle 1917 strain cannot be dissociated from its probiotic activity. Gut Microbes 2012;3:501–509.
13. Nougayrède J-P, Homburg S, Taieb F, et al. Escherichia coli Induces DNA Double-Strand Breaks in Eukaryotic Cells. Science 2006;313:848–851.
14. Li Z-R, Li J, Cai W, et al. Macrocyclic colibactin induces DNA double-strand breaks via copper-mediated oxidative cleavage. Nat Chem 2019;11:880–889.
15. Dziubańska-Kusibab PJ, Berger H, Battistini F, et al. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. Nature Medicine 2020:1–7.
16. Terveer EM, Beurden YH van, Goorhuis A, et al. How to: Establish and run a stool bank. Clin Microbiol Infect 2017;23:924–930.
17. Schmitz D, Zwagemaker F, Nooij S, et al. *Jovian, user-friendly Public Health toolkit*. Zenodo; 2020. Available at: https://doi.org/10.5281/zenodo.3929618.

18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–2120.

19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012;9:357–359.

20. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–2079.

21. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–842.

22. Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27:824–834.

23. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.

24. Rubino F, Creevey CJ. MGkit: Metagenomic Framework For The Study Of Microbial Communities. 2014. Available at: https://figshare.com/articles/poster/MGkit_Metagenomic_Framework_For_The_Study_Of_Microbial_Communities/1269288/1 [Accessed November 10, 2020].

25. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] 2013. Available at: http://arxiv.org/abs/1303.3997 [Accessed November 10, 2020].

26. Nooij S. *jovian-screener*. Zenodo; 2020. Available at: https://doi.org/10.5281/zenodo.4288679.

27. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available at: https://www.R-project.org/.

28. Lee C, Kim J, Shin SG, et al. Absolute and relative QPCR quantification of plasmid copy number in Escherichia coli. Journal of Biotechnology 2006;123:273–280.

29. Chern EC, Siefring S, Paar J, et al. Comparison of quantitative PCR assays for Escherichia coli targeting ribosomal RNA and single copy genes. Letters in Applied Microbiology 2011;52:298–306.

30. Wirth T, Falush D, Lan R, et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Molecular Microbiology 2006;60:1136–1151.

31. Adiri RS, Gophna U, Ron EZ. Multilocus sequence typing (MLST) of Escherichia coli O78 strains. FEMS Microbiol Lett 2003;222:199–203.

32. Pritchard L, Glover RH, Humphris S, et al. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods 2015;8:12–24.

33. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 2010;11:595.

34. Smillie CS, Sauk J, Gevers D, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host & Microbe 2018;23:229-240.e5.

35. He Z, Gharaibeh RZ, Newsome RC, et al. Campylobacter jejuni promotes colorectal tumorigenesis through the action of cytolethal distending toxin. Gut 2019;68:289–300.

36. Putze J, Hennequin C, Nougayrède J-P, et al. Genetic Structure and Distribution of the Colibactin Genomic Island among Members of the Family Enterobacteriaceae. Infection and Immunity 2009;77:4696–4703.

37. Fukugaiti MH, Ignacio A, Fernandes MR, et al. High occurrence of Fusobacterium nucleatum and Clostridium difficile in the intestinal microbiota of colorectal carcinoma patients. Braz J Microbiol 2015;46:1135–1140.