Title: Re-evaluating the evidence for fecal microbiota transplantation "super-donors" in inflammatory bowel disease

Short title: Re-evaluating FMT "super-donors"

Authors: Scott W. Olesen, PhD[1,2]; Ylaine Gerardin, PhD[3]

1 OpenBiome, Cambridge, MA
2 Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA
3 Finch Therapeutics, Somerville, MA

Emails: solesen@openbiome.org, ylaine@finchtherapeutics.com

ABSTRACT

Fecal microbiota transplantation (FMT) is a recommended treatment for recurrent *Clostridioides difficile* infection, and there is promise that FMT may be effective for conditions like inflammatory bowel disease (IBD). Previous FMT clinical trials have considered the possibility of a "donor effect", that is, that FMT material from different donors has different clinical efficacies. Here we lay out rigorous statistical methodology for detecting donor effects, finding that reliable detection of a donor effect requires trials with more than 200 FMT-treated patients. A re-evaluation of previous FMT clinical trials for IBD showed that while there is very little evidence for a non-zero donor effect, the existing data are also not inconsistent with substantial donor effects. Large-scale meta-analysis, combined with careful reporting from clinical trials, will be crucial in determining if donor effects are clinically relevant for IBD.

INTRODUCTION

The human microbiome is increasingly understood to play a key role in health and disease.[1] Fecal microbiota transplantation (FMT), the infusion of a healthy person's stool into a patient, is one method for manipulating the gut microbiome. FMT is recommended for treatment of recurrent *Clostridioides difficile* infection.[2,3] Although the specific mechanisms by which FMT cures *C. difficile* infection are not well understood, FMT is being investigated as a therapy for dozens of other microbiome-related indications.[1]

A key challenge in identifying FMT's specific mechanism, or mechanisms, is the complexity and diversity of human stool, a mixture of bacteria, viruses, fungi, microbe-derived molecules, and host-derived molecules which varies enormously from person to person.[4,5] It has therefore been hypothesized that different stools, stool donors, or matches of donors and recipients could have different abilities to treat disease. This concept has been referred to with terms like "donor effect", "super-donor", and "super-stool".[6–9]

If FMT's efficacy varied widely across donors, then rational, *a priori* selection of stool donors based on biomarkers that predict their donated FMT material's efficacy could improve the clinical practice of FMT.[10,11] Differences in FMT efficacy between stool donors or specific stool samples could be also an important starting point for scientific investigations into the "active ingredient" in FMT donations.[6] It is therefore important to

quantify the extent of donor variability in indications where FMT is a promising treatment.

There are no reports of a donor effect in the context of *C. difficile* infection, the condition for which the use of FMT is best studied.[12] However, multiple studies, mostly for ulcerative colitis, a kind of inflammatory bowel disease, have tested for donor effects.[8,9,13–20] In all of these studies, identifying a donor effect was a *post hoc* analysis, and in no case was multiple hypothesis correction, a critical methodology in *post hoc* analyses, employed. As we show herein, in at least two previous reports the specifics of the clinical design require adjustments to the statistical tests used, such that those reports inflated the apparent evidence for a donor effect.[13,14]

Furthermore, previous reports mostly analyzed donor effects with respect to the null hypotheses that all donors produce stool with equal clinical efficacy, or that particular features of stool, like its bacterial community diversity, are unrelated to clinical efficacy. We propose that, rather than merely asking if donor effects exist or not, we should ask about the size of the donor effect, whether it is clinically relevant, and whether it is large enough to be detected in the studies that are testing for it. It is inconceivable that all stool has precisely the same therapeutic effect. If 2 donors are each used to treat enough patients, the difference between their treatment efficacies would almost certainly become statistically significant. However, without quantifying the donor effect, it is difficult to determine if the effect is clinically relevant or to determine how many patients would be required to power a clinical trial to reliably detect a donor effect.

Here we re-evaluate the existing evidence for donor effects and evaluate the implications of that evidence for experimental design and clinical practice. First, we lay out easy-to-implement but statistically rigorous methods for evaluating the donor effect in *post hoc* analyses. Second, we use these methods to estimate the statistical power of FMT clinical trials to detect donor effects. Finally, we apply these statistical methods to the existing literature.

METHODS

Distinguishing types of stool superiority

Discussions of “super-donors” and “super-stool” have suggested that stool might be superior in at least 4 distinct but conceptually related senses. To avoid confusion in our re-evaluation of the evidence for donor effects, we distinguish between these definitions of stool superiority:

1. *Donor superiority*, in which particular donors are associated with better clinical outcomes for the recipient patients. For example, Moayyedi *et al.*[13] tested whether a particular donor (“donor B”) was associated with better patient outcomes, compared to the other donors.
2. *Donor characteristic superiority*, in which a donors with some particular characteristic are associated with better outcomes. For example, donor age, diet,

and host genetics have been suggested as potential factors in donor superiority.[8,16,21]

3. *Donation superiority*, in which donations with some particular characteristic are associated with better outcomes. For example, Vermeire *et al.*[22] tested whether stools with higher bacterial diversity are associated with better outcomes.
4. *Donor-recipient match superiority*, in which certain combinations of donor and recipient are associated with better outcomes,[8,9,20,23,24] analogous to how donors' and recipients' blood types are matched for blood transfusions. FMT studies have tested whether stool from "related" donors,[25] typically defined as first-degree relatives but also sometimes including spouses or partners, is associated with better patient outcomes.

These types of superiority are conceptually related and not mutually exclusive. For example, imagine it were the case that female donors were associated with better patient outcomes than male donors, making a case for "donor characteristic superiority". However, because donors have a clinical effect on the patient only via their donation, sex of donor cannot be the molecular mechanism by which some donations are more efficacious than others. It would have to be that sex determined or correlated with some component of the stool that made those donations more effective. In other words, there must be an underlying "donation characteristic superiority" that correlates with donor sex. Thus, even if donor-recipient match superiority might be the most accurate model, simpler types of superiority may be parsimonious.[7]

When pools of stool from multiple donors are used for FMT, it is typical to test for donor superiority (i.e., if the pools that include material from a particular donor are associated with better patient outcomes)[15] and pool characteristic superiority (e.g., if pools with higher bacterial α-diversity are associated with better patient outcomes).[17] One could also test if particular pools are associated with better patient outcomes.

Model of donor effects and quantification of donor effect size

To quantify the size of the donor effect, we modeled each donor as having a unique treatment efficacy, that is, the proportion of patients that would reach a positive clinical endpoint in a very large clinical trial. The donors' treatment efficacies are drawn from a distribution. Specifically, the log odds of the donors' treatment efficacies are normally distributed. We used this model of distribution in donor efficacies because it corresponds to the distribution of a random effect in a logistic regression, which can be fit to data using well-established statistical methods.[26]

Under this model, the size of the donor effect is related to the "width" of this distribution. If the donor effect is strong, the difference in treatment efficacies between donors tends to be large. We defined the effect size as the median difference in treatment efficacy between two randomly selected donors (see Supplemental Methods). For example, say the effect size is a median difference of 10 percentage points, and the mean treatment efficacy across donors is 50%. Then it would not be unusual if two donors selected at random had efficacies of 45% and 55%.

130

131 **Statistical power computations**

132 We performed statistical power computations to estimate what donor effect size can be
133 reliably detected in an FMT clinical trial for IBD. At a range of effect sizes, statistical
134 power was estimated as the proportion of 1,000 simulated clinical trials with $p < 0.05$.
135 For each simulated clinical trial:

136

137 1. Efficacies for donors (2, 3, 6, or 12 donors) were randomly drawn from a
138 distribution of log odds. The mean of the distribution of log odds was selected
139 based on typical FMT efficacy rates (30%).[9,13,15,27] The effect size determines the
140 standard deviation.
141 2. Patients treated with FMT (12, 24, 48, 96, or 192 patients) were distributed
142 evenly among the donors.
143 3. Patients' outcomes for patients were simulated based on their donors' efficacies.
144 4. A statistical test (Fisher-Freeman-Halton test, see below) was run, and the *p*-
145 value noted.

146

147 If the simulated power for any effect size reached 80%, we found the effect size
148 corresponding to 80% power by linear interpolation.

149

150 10 effect sizes were evaluated, ranging from approximately zero (10^{-6} standard
151 deviation in log odds) to the difference in log odds between the typical FMT efficacy rate
152 and typical placebo rate for a clinical indication. This design sets the upper limit on

153 donor efficacies such that, at most, those donors 1 standard deviation below the mean
154 (i.e., ~16% of all donors) have efficacies worse than typical placebo rates (5%).
155
156 **Studies re-analyzed**
157 We analyzed a set of studies from recent FMT clinical study literature that tested for the
158 hypothesis that one donor or donation characteristic was associated with improved
159 patient outcomes (Results). To simplify the analyses, we considered only a single
160 dichotomous outcome, whether a patient achieved the trial's primary clinical endpoint.
161 We also restricted microbiome analyses to consider only the bacterial α-diversity of
162 stool, as associations between donor material α-diversity and patient outcomes have
163 been investigated in multiple studies,[14,16,22] and the hypothesis that improved bacterial
164 diversity in stool is part of the motivation for using pools of stool from multiple donors in
165 FMT.[15,23,27]
166
167 **Statistical methodology**
168 Statistics were computed using the R programming language (version 3.6.0).[28] To test
169 the hypothesis that there is any difference in efficacy between donors (or pools of
170 donors), we used the Fisher-Freeman-Halton test (function *fisher.test*) on $2 \times D$
171 contingency tables, where D is the number of donors (or pools). In cases when the
172 Fisher-Freeman-Halton test is computationally infeasible, we use the χ^2 test (function
173 *chisq.test*). For 2×2 contingency tables (e.g., to test if one donor has a different
174 treatment efficacy than all the others), we used Fisher's exact test (function *exact2x2*)[29]
175 and the mid-p value to prevent the tests from being overly conservative.[30] We used

176 logistic regressions (function *glm*) to test for donation/pool characteristic superiority. We
177 used random effects logistic regressions (function *glmer* from the *lme4* package)[26] to
178 estimate the strength of the donor effect.
179
180 **Microbiome analysis**
181 We re-analyzed raw 16S rRNA sequencing data from three studies (Kump *et al.*, study
182 accession PRJEB11841; Jacob *et al.*, accession PRJNA388210; Goyal *et al.*, accession
183 PRJNA380944) using QIIME 2[31] (version 2019.7) and Deblur.[32] Paired reads were
184 joined (*vsearch* plugin, default parameters), quality filtered (*quality-filter* plugin, default
185 parameters), and denoised using Deblur (trim length 253 nucleotides, 1 minimum reads,
186 otherwise default parameters). To follow Kump *et al.*'s general analysis method, each
187 sample in the resulting table of amplicon sequence variants was downsampled to the
188 minimum number of ASV counts across samples. Bacterial α-diversity was computed as
189 richness, the number of unique ASVs present in each sample. For Goyal *et al.*, the
190 sequencing data was single-ended, so reads were not paired for that study. When a
191 donor had multiple samples, we associated a single α-diversity with each donor by
192 computing the average number of ASVs over that donor's samples.
193
194 **Code and data availability**
195 Computer code and underlying data to reproduce the results are online at
196 https://www.github.com/openbiome/donor-effects.
197
198

RESULTS

Statistical power of FMT clinical trials for IBD to detect a donor effect

Using simulations, we estimated the minimal strength of the donor effect required for an FMT clinical trial for IBD to reach 80% statistical power in detecting a donor effect with a Fisher-Freeman-Halton test (Figure 1). Trials' ability to reliably detect a donor effect increased with the strength of the donor effect, the number of patients, and the number of donors in the trial. More patients and donors allowed for reliable detection of weaker donor effects.

For trials with 24 patients and any number of donors, the strength of the donor effect required for reliable detection of that effect was implausibly large: it would require that a substantial fraction of donors caused their patients to perform worse than placebo (see Methods). Similarly, studies with 2 donors and 192 or fewer patients were also unable to detect plausible donor effects. Trials with 96 patients were able to reliably detect a donor effect when the median difference in treatment efficacy between randomly selected donors was 20 percentage points or more. In other words, the donor effect must be strong enough that, in a typical pair of donors, one might have a clinical efficacy of 20% and the other 40%. Even with 384 patients in a trial, the donor effect would need to be large enough that the median difference in clinical efficacy between randomly selected donors was more than 10 percentage points.

Re-evaluations of the evidence for a donor effect in IBD

222 The results of the statistical power computations above suggest that previous studies of
223 FMT for IBD are unlikely to have detected a donor effect. We therefore re-evaluated the
224 evidence for a donor effect in IBD from a selection of FMT clinical studies (Table 1).
225
226 *Rossen et al. 2015*
227 In this study,[9] 23 ulcerative colitis patients were treated with FMT. 17 patients received
228 stool from 1 of 15 donors; 6 patients received stool from 2 different donors. In the
229 original publication, the investigators report on the performance of 3 donors (Table 2).
230 There was no evidence of a donor effect (mid-$p > 0.05$, Fisher's exact test) for any of
231 the 3 donors. However, the study is small enough that the data are not inconsistent with
232 large variability between donors. For example, the upper 95% confidence interval on
233 donor A's odds ratio of successful patient outcome compared to all other donors is 29.
234 In other words, the data do not provide evidence of a donor effect, but they also cannot
235 definitively rule out a strong donor effect.
236
237 *Moayyedi et al. 2015*
238 In this study,[13] 38 ulcerative colitis patients were treated with FMT. Each patient was
239 treated with stool from 1 of 6 donors. Donors were used according to an adaptive
240 process, described in the original publication. Briefly, at the start of the trial, patients
241 were treated with stool from one of two donors (A or B). Material from donor B then
242 became unavailable, and patients were instead treated with stool from donor A or 1 of 4
243 other donors (C, D, E, and F). At this point, donor B became available again, and "[t]he
244 remaining participants allocated to active therapy all received FMT from donor B

245 exclusively, as [the investigators] had not experienced any success with donor A". In the
246 original study, donor B's performance (7 of 18 treated patients achieved the primary
247 endpoint) was compared against all other donors (2 of 20 patients; odds ratio 5.5 [95%
248 CI 1.0 to 44], mid-p = 0.048, Fisher's exact test).
249
250 In general, adaptive trial designs require special statistical methodologies.[33] However,
251 we were aware of no statistical methodologies designed to account for an adaptive
252 design like the one used in this trial. To investigate whether typical statistical tests like
253 Fisher's exact test or a χ^2 test accurately measure the statistical significance of the
254 collected data, we simulated a simplified adaptive donor selection process. Among the
255 first 24 patients, use all 6 donors 4 times each. Then exclusively use the best-
256 performing donor for the remaining 14 patients. Finally, compare the overall
257 performance of this donor against all the others. In these simulations, if all the donors
258 are the same, the probability of finding mid-$p < 0.05$ is approximately 9%, greater than
259 that 5% that would be expected if the test were accurate for that type of data
260 (Supplemental Results). The intuitive explanation is that, if all donors are the same, the
261 selection process is merely setting aside donors who had "bad luck" on their first
262 patients, and they are not allowed to recover their performance with a run of "good luck"
263 later on. This is not a flaw of Fisher's exact test. Instead, it is problem of applying a
264 statistical test to data that do not meet the assumptions of the test.
265
266 Although the adaptive procedure that led to extensive use of donor B may have
267 improved the probability of the trial's success,[7] the procedure was not pre-specified,

making it impossible to rigorously determine the degree to which the observed results are consistent with all donors being identical, that is, whether donor B simply had a "lucky" initial run. Thus, the value mid-$p \approx 0.05$ reported above is inaccurately optimistic, and a value closer to 0.09 is a better representation of the likelihood of a donor having the kind of success observed in this trial.

Paramsothy et al. 2017

In this study,[15] 41 ulcerative colitis patients were randomized to receive FMT, and another 37 were randomized to placebo and later received FMT. Each patient received FMT from 1 of 21 pools of donors. Each pool included 3 to 7 donors, out of 13 total donors.

The original publication compared the best performing individual donor against all the others (14 of 38 patients achieved the primary outcome, versus 7 of 40 assigned to other donors; odds ratio 2.7 [95% CI 0.96 to 8.2], mid-$p = 0.06$, Fisher's exact test). This result, although near mid-$p = 0.05$, was not subjected to multiple hypothesis correction, as there are 13 relevant hypotheses to be tested, one per donor. In terms of pools of donors, there was also no evidence that any particular pool was associated with better outcomes ($p = 0.57$, Fisher-Freeman-Halton test on 2×21 table of patient outcomes by pool).

To develop an estimate of the strength of the donor effect, we used a logistic regression, modeling the pool as a random effect. This model estimated that the median

difference between the efficacy of two randomly selected pools is 0 percentage points (95% CI 0 to 20 percentage points). In other words, the regression's best estimate of the variation in efficacy among pools is zero, even without applying any penalizations for model complexity. Thus, similar to Rossen *et al.*, the data do not provide evidence for a donor effect, but they are also not inconsistent with a strong effect, since the upper confidence limit on the strength of the donor effect is 20 percentage points.

Costello et al. 2019

In this study,[27] 38 ulcerative colitis patients received FMT from 1 of 11 pools. Each pool included stool from 3 or 4 donors, out of 19 total donors. Similar to Paramsothy *et al*., there was no evidence of heterogeneity in patient outcomes by donor pool ($p = 0.50$, Fisher-Freeman-Halton test on 2 × 11 table) nor evidence of better outcomes for any particular donor (mid-$p > 0.05$ for all donors, Fisher's exact test). Furthermore, a logistic regression estimated the median difference between two randomly selected donors as 0 percentage points (95% CI 0 to 26 percentage points).

Jacob et al. 2017

In this study,[17] 20 ulcerative colitis patients received FMT from 1 of 6 pools. Each pool included stool from 2 donors, out of 4 total donors. In this analysis, we analyzed only the 14 patients available for follow-up at 12 weeks, using remission at this timepoint as the outcome. Similar to Paramsothy *et al.* and Costello *et al.*, there was no evidence for differences in patient outcomes by pool (Fisher-Freeman-Halton test on 2 × 4 table, $p = 0.86$) or by donor (mid-$p > 0.05$ for all 4 donors, Fisher's exact test). A logistic

314 regression estimated the median difference in efficacy between two randomly selected
315 pools as zero (95% CI 0 to 31 percentage points).
316
317 The pools administered to 13 of the 14 patients were characterized using 16S rRNA
318 sequencing, allowing for an analysis of patient outcomes by bacterial characteristics of
319 the pools (Figure 2). However, there was no evidence for an association between
320 patient outcomes and the bacterial community α-diversity, measured as the number of
321 unique amplicon sequence variants (ASVs, i.e., denoised 100% OTUs), of the pool they
322 received (p = 0.18, Wilcoxon rank-sum test).
323
324 To develop confidence intervals around the strength of the donor effect, we fit a logistic
325 regression to the data, predicting patient outcomes from pool α-diversity. The model's
326 point estimates for pool efficacy was 20% (95% CI 2% to 76%) for the pool with the
327 lowest α-diversity (239 ASVs) but 84% (95% CI 25% to 99%) for the pool with the
328 highest diversity (487 ASVs). In other words, the data do not provide statistically
329 significant evidence for an association between pool α-diversity and patient outcomes,
330 but they cannot definitively rule out the possibility of a wide variation in efficacy between
331 pools.
332
333 *Meta-analysis of pool studies*
334 Paramsothy *et al*., Costello *et al.*, and Jacob *et al.* have sufficiently similar designs and
335 the available data to allow a meta-analysis. A logistic regression on the combined data
336 from all three studies, with the study as a fixed effect, estimated the median difference

337 in efficacy between two randomly selected pools as zero, with a smaller confidence
338 interval than for any individual study (95% CI 0 to 16 percentage points).
339
340 *Goyal et al. 2018*
341 In this study,[16] 21 patients with inflammatory bowel disease (any of Crohn's disease,
342 ulcerative colitis, or indeterminate colitis) received FMT and were available for follow-up.
343 We used remission at 30 days as the outcome. Each patient received stool from 1
344 donor, precluding any test of donor superiority. However, 16S rRNA sequencing was
345 performed, allowing for an analysis of donation characteristic superiority, similar to
346 Jacob *et al.* above.
347
348 Similar to that study, there was no evidence for an association between patient
349 outcomes and the α-diversity (p = 0.3, Wilcoxon rank-sum test), but a logistic regression
350 estimated a potentially wide but statistically insignificant variation in donation efficacy
351 across the range of α-diversities among donors, predicting an efficacy of 38% (95% CI
352 8% to 82%) for the donor with the lowest α-diversity (386 ASVs) and an efficacy of 75%
353 (95% CI 26% to 96%) for the donor with the highest diversity (1072 ASVs). Again, there
354 was no statistically significant association between patient outcomes and donor stool
355 characteristics, but the data are not inconsistent with a wide variation.
356
357 *Kump et al. 2017*
358 In this study,[14] 17 ulcerative colitis patients received FMT from 1 of 14 donors. 12
359 donors were used in 1 patient, 1 donor was used in 2 patients, and 1 donor was used in

3 patients. Like Jacob *et al.* and Goyal *et al*., this study tested whether increased α-diversity was associated with improved patient outcomes.

In the original analysis, the 4 patients who did not respond clinically to FMT were compared against the 4 patients who reached remission (Figure 2). 16S rRNA sequencing data was available for 28 donations associated with those 8 patients. Donors assigned to patients who received remission did not have statistically significantly higher α-diversity than those assigned to patients who did not respond (Wilcoxon rank-sum test, $p = 0.3$).

The original publication reported statistically significant differences between the α-diversity of stool samples associated with remission ($n = 16$) compared to samples associated with no response ($n = 12$; $p = 0.11$). Our analysis has a different conclusion, finding no support for a donor effect, principally because we analyzed the data as only 8 independent data points (i.e., one for each patient), while Kump *et al.* analyzed the data as if there were 28 independent data points (i.e., one for each donation that was administered to a patient that also has associated microbiome data). Repeated measurements of the material delivered to a patient do not constitute independent measures of the patient's outcome. In other words, when testing for an association between patient outcomes and a donation characteristic like α-diversity, the weight of the evidence is determined by the number of patient outcomes, not by the number of microbiome measurements.

As an extreme example, consider a study with two donors and two patients: donor A is used with a patient who achieves remission, while donor B is used with a patient who has no response to FMT. If 100 microbiome samples were taken from each of donors A and B, a difference between A's and B's samples could be found with great statistical confidence. However, this does not translate into confidence in the relationship between the microbiome and patient outcome.

Similar to the studies above, a logistic regression identified a potentially wide but not statistically significant variation in donation efficacy across the range of α-diversities, predicting an efficacy of 38% (95% CI 8% to 82%) for the donor with the lowest α-diversity (386 ASVs) and an efficacy of 75% (95% CI 26% to 96%) for the donor with the highest diversity (1072 ASVs).

Vermeire et al. 2016

In this study, 14 patients with inflammatory bowel disease (either of Crohn's disease or ulcerative colitis) received FMT, each from a different donor. The original study reported that the bacterial α-diversity of donations varied by patient outcomes, that patients with successful outcomes received more bacterially diverse donations (Wilcoxon rank-sum test, $p = 0.012$). However, the underlying bacterial sequencing data was not available for re-analysis.

Nishida et al. 2017

405 In this study, 41 patients with ulcerative colitis received FMT, each from a different
406 donor. The original study found no difference in the α-diversity of the stool from 7
407 donors whose patients met the primary endpoint, compared to 19 donors whose
408 patients did not (p = 0.69). The study also tested whether the abundance of 10 taxa
409 were differentially abundant among those two groups of donors, finding p-values less
410 than 0.05 for 2 of 9 taxa, neither of which was significant after Benjamini-Hochberg
411 multiple hypothesis correction.[34] The underlying bacterial sequencing data was not
412 available for re-analysis.
413
414
415 **DISCUSSION**
416
417 In simulation of FMT clinical trials for IBD, we found that trials would need to be very
418 large, with more than 200 patients receiving FMT, to achieve 80% statistical power to
419 detect a modest donor effect (median 10 percentage point difference in efficacy
420 between two randomly selected donors). In other words, if the donor effect were only
421 strong enough that randomly selected pairs of donors had a 10 percentage point
422 difference in efficacy —say, one having a 30% efficacy and the other a 40% efficacy—
423 then no existing trial of FMT for IBD had a design that could reliably detect such an
424 effect.
425
426 Concluding that existing FMT studies for IBD were simply not powered to detect donor
427 effects, we re-evaluated reports from those studies, finding no statistically rigorous

428 evidence for the superiority of particular donors and only very weak evidence for
429 variation in FMT's efficacy based on the α-diversity of the FMT material (specifically, a
430 single result from Vermeire *et al.*). In most cases, the statistical best estimate for the
431 strength of the donor effect is exactly zero. However, the small size of these studies
432 relative to the number of patients required to reliably detect a plausible donor effect also
433 means that the data cannot definitively rule out a large donor effect, as much as 20 or
434 30 percentage points between randomly selected donors. It therefore remains
435 undetermined if differences between donors are clinically relevant to FMT for IBD.
436
437 These sobering results are subject to multiple limitations. First, we measured the
438 associations between very simple measures. We considered only dichotomous patient
439 outcomes and the relatively simple microbiological characteristic of bacterial α-diversity
440 as determined by 16S rRNA taxonomic marker gene sequencing. Reliably detecting
441 donor effects may require more sophisticated methods. For example, Jacob *et al.* report
442 that donors' bacterial compositions meaningfully differed with respect to patient
443 outcomes (Figure 4C in that publication, PERMANOVA, $p = 0.044$). More proximal and
444 fine-grained measurements of patients' response to FMT may also provide more
445 sensitivity for detecting differences in donor efficacies. Dichotomous primary endpoints
446 and ordinal changes in patients' Mayo scores are subject to enormous sources of
447 variance in the patient's biology and disease history that may be hiding the more subtle
448 effects of differences between donors.
449

450 Second, we modeled the distribution of donors and pool efficacies as Gaussian in log
451 odds space. This approach allowed us to use mixed model logistic regression, a well-
452 established statistical procedure, to estimate the variance in donors' efficacies, but it is
453 not an intuitive measure of variation. Furthermore, this model of donor variability does
454 not require that all donors' efficacies be higher than the placebo rate. A model of donor
455 variability that does sensibly account for placebo rates was previously used in a study of
456 FMT clinical trial design,[7] but that model would require adaptation to make it suitable for
457 statistical inference.
458
459 In light of the number of patients required for a well-powered test for a donor effect, we
460 propose that an important path is meta-analysis across clinical trials, compiling
461 characterizations of donors' and patients' microbiomes and host factors, linked with
462 patient outcomes.[6] This approach will present difficult technical challenges, as data from
463 multiple studies be carefully collected, cleaned, and analyzed to make the analytical
464 results valid across those studies.[35] However, liberal sharing of clinical metadata and
465 microbiome data, including careful reporting of donor and donation characteristics, will
466 more easily enable inference across studies, and may provide the best direction forward
467 to characterizing the donor effect for FMT in IBD.
468
469
470 **Author contributions**

SWO conceived the study, obtained data, performed the analysis, and wrote the manuscript. YG obtained data; improved the analysis, interpretation, and presentation of the results; and revised the manuscript.

Acknowledgements

- Sudarshan Paramsothy, Nadeem Kaakoush, and Rotem Sadovsky for assistance in collecting the Paramsothy *et al.* data and for helpful conversations.
- Sam Costello for assistance in collecting the Costello *et al.* data and for helpful conversations.
- Eric J. Alm, Shrish Budree, Claire Duvallet, Justin O'Sullivan, Pratik Panchal, Mark B. Smith, and Duane Wesemann for helpful conversations.

483 **REFERENCES**

484

485 1. Marchesi, J. R. *et al.* The gut microbiota and host health: a new clinical frontier. *Gut*
486 **65**, 330–339 (2016).
487 2. McDonald, L. C. *et al.* Clinical Practice Guidelines for Clostridium difficile Infection in
488 Adults and Children: 2017 Update by the Infectious Diseases Society of America
489 (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clin. Infect. Dis.*
490 **66**, e1–e48 (2018).
491 3. Debast, S. B., Bauer, M. P. & Kuijper, E. J. European Society of Clinical Microbiology
492 and Infectious Diseases: Update of the Treatment Guidance Document for
493 Clostridium difficile Infection. *Clin. Microbiol. Infect.* **20**, 1–26 (2014).
494 4. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome.
495 *Nature* **505**, 559–563 (2014).
496 5. The Human Microbiome Project Consortium. Structure, function and diversity of the
497 healthy human microbiome. *Nature* **486**, 207–214 (2012).
498 6. Olesen, S. W., Leier, M. M., Alm, E. J. & Kahn, S. A. Searching for superstool:
499 maximizing the therapeutic potential of FMT. *Nat. Rev. Gastroenterol. Hepatol.* **15**,
500 387–388 (2018).
501 7. Olesen, S. W., Gurry, T. & Alm, E. J. Designing fecal microbiota transplant trials that
502 account for differences in donor stool efficacy. *Stat. Methods Med. Res.* **27**, 2906–
503 2917 (2018).

504 8. Wilson, B. C., Vatanen, T., Cutfield, W. S. & O'Sullivan, J. M. The Super-Donor
505 Phenomenon in Fecal Microbiota Transplantation. *Front. Cell. Infect. Microbiol.* **9**, 2
506 (2019).
507 9. Rossen, N. G. *et al.* Findings From a Randomized Controlled Trial of Fecal
508 Transplantation for Patients With Ulcerative Colitis. *Gastroenterology* **149**, 110-
509 118.e4 (2015).
510 10. Barnes, D. *et al.* Competitively Selected Donor Fecal Microbiota Transplantation:
511 Butyrate Concentration and Diversity as Measures of Donor Quality. *J. Pediatr.*
512 *Gastroenterol. Nutr.* **67**, 185–187 (2018).
513 11. Duvallet, C. *et al.* Framework for rational donor selection in fecal microbiota
514 transplant clinical trials. *medRxiv* (2019) doi:10.1101/19000307.
515 12. Osman, M., Abend, A., Panchal, P., Kassam, Z. & Budree, S. 88 - Does the
516 Donor Matter? Microbiome Sequencing to Evaluate Lower Donor Efficacy in Fecal
517 Microbiota Transplantation for Recurrent Clostridium Difficile Infection.
518 *Gastroenterology* **154**, S-25-S-26 (2018).
519 13. Moayyedi, P. *et al.* Fecal Microbiota Transplantation Induces Remission in
520 Patients With Active Ulcerative Colitis in a Randomized Controlled Trial.
521 *Gastroenterology* **149**, 102-109.e6 (2015).
522 14. Kump, P. *et al.* The taxonomic composition of the donor intestinal microbiota is a
523 major factor influencing the efficacy of faecal microbiota transplantation in therapy
524 refractory ulcerative colitis. *Aliment. Pharmacol. Ther.* **47**, 67–77 (2018).

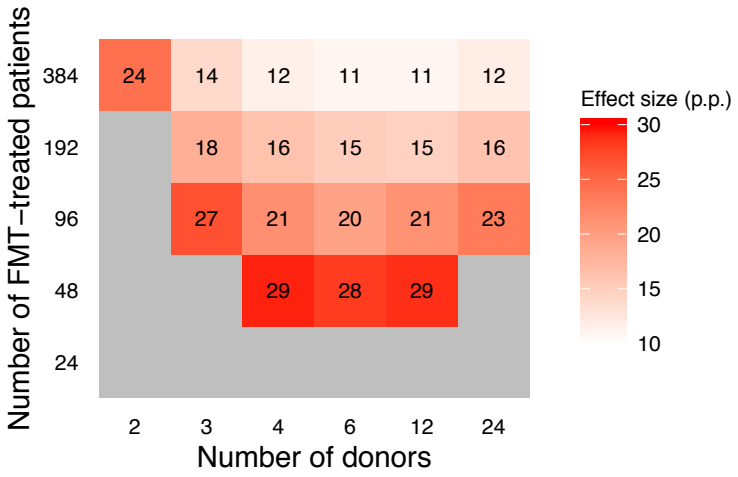
525 15. Paramsothy, S. *et al.* Multidonor intensive faecal microbiota transplantation for
526 active ulcerative colitis: a randomised placebo-controlled trial. *Lancet Lond. Engl.*
527 **389**, 1218–1228 (2017).
528 16. Goyal, A. *et al.* Safety, Clinical Response, and Microbiome Findings Following
529 Fecal Microbiota Transplant in Children With Inflammatory Bowel Disease. *Inflamm.*
530 *Bowel Dis.* **24**, 410–421 (2018).
531 17. Jacob, V. *et al.* Single Delivery of High-Diversity Fecal Microbiota Preparation by
532 Colonoscopy Is Safe and Effective in Increasing Microbial Diversity in Active
533 Ulcerative Colitis: *Inflamm. Bowel Dis.* **23**, 903–911 (2017).
534 18. Allegretti, J. R., Mullish, B. H., Kelly, C. & Fischer, M. The evolution of the use of
535 faecal microbiota transplantation and emerging therapeutic indications. *Lancet Lond.*
536 *Engl.* **394**, 420–431 (2019).
537 19. Nishida, A. *et al.* Efficacy and safety of single fecal microbiota transplantation for
538 Japanese patients with mild to moderately active ulcerative colitis. *J. Gastroenterol.*
539 **52**, 476–482 (2017).
540 20. Paramsothy, S. *et al.* Faecal Microbiota Transplantation for Inflammatory Bowel
541 Disease: A Systematic Review and Meta-analysis. *J. Crohns Colitis* **11**, 1180–1199
542 (2017).
543 21. Budree, S. *et al.* The Association of Stool Donor Diet on Microbial Profile and
544 Clinical Outcomes of Fecal Microbiota Transplantation in Clostridium Difficile
545 Infection. *Gastroenterology* **152**, S630–S631 (2017).

546 22. Vermeire, S. *et al.* Donor Species Richness Determines Faecal Microbiota
547 Transplantation Success in Inflammatory Bowel Disease. *J. Crohns Colitis* **10**, 387–
548 394 (2016).
549 23. Kazerouni, A. & Wein, L. M. Exploring the Efficacy of Pooled Stools in Fecal
550 Microbiota Transplantation for Microbiota-Associated Chronic Diseases. *PloS One*
551 **12**, e0163956 (2017).
552 24. Davido, B. *et al.* Impact of Fecal Microbiota Transplantation for Decolonization of
553 Multidrug-Resistant Organisms May Vary According to Donor Microbiota. *Clin. Infect.*
554 *Dis.* (2017) doi:10.1093/cid/cix963.
555 25. Khanna, S. *et al.* Changes in microbial ecology after fecal microbiota
556 transplantation for recurrent C. difficile infection affected by underlying inflammatory
557 bowel disease. *Microbiome* **5**, (2017).
558 26. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects
559 Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).
560 27. Costello, S. P. *et al.* Effect of Fecal Microbiota Transplantation on 8-Week
561 Remission in Patients With Ulcerative Colitis: A Randomized Clinical Trial. *JAMA*
562 **321**, 156–164 (2019).
563 28. R Core Team. *R: a language and environment for statistical computing*. (R
564 Foundation for Statistical Computing, 2019).
565 29. Fay, M. P. Confidence intervals that match Fisher's exact or Blaker's exact tests.
566 *Biostatistics* **11**, 373–374 (2010).
567 30. Mehta, C. R. & Hilton, J. F. Exact Power of Conditional and Unconditional Tests:
568 Going beyond the 2 × 2 Contingency Table. *Am. Stat.* **47**, 91–98 (1993).

569 31. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome
570 data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
571 32. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence
572 Patterns. *mSystems* **2**, (2017).
573 33. Bhatt, D. L. & Mehta, C. Adaptive Designs for Clinical Trials. *N. Engl. J. Med.*
574 **375**, 65–74 (2016).
575 34. Yekutieli, D. & Benjamini, Y. The control of the false discovery rate in multiple
576 testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
577 35. Gibbons, S. M., Duvallet, C. & Alm, E. J. Correcting for batch effects in case-
578 control microbiome studies. *PLoS Comput. Biol.* **14**, e1006102 (2018).
579

FIGURES AND TABLES

Figure 1: Statistical power calculations showing detectable effect sizes for varying trial designs. In simulated clinical trials, a number of patients (vertical axis) are evenly allotted to a number of donors (horizontal axis). Each tile's color and number show the minimum effect size (median difference in efficacy, measured in percentage points, between two randomly selected donors) required for the trial design to reach 80% statistical power. Gray tiles show that 80% statistical power cannot be achieved for plausible donor variations (i.e., for which most donors perform at least as well as placebo). p.p.: percentage points.



592 **Table 1**: Summary of re-analyses.

Data source	Indication	No. patients	No. donors	No. pools	Evidence for donor superiority?	Evidence for pool superiority?	Association between diversity and outcomes?
Rossen *et al.*	IBD	23	15	—	No	—	—
Moayyedi *et al.*	IBD	38	6	—	No*	—	—
Paramsothy *et al.*	IBD	78	13	21	No	No (0 p.p., 95% CI 0 to 20)†	—
Costello *et al.*	IBD	38	19	11	No	No (0 p.p., 95% CI 0 to 26)	—
Jacob *et al.*	IBD	21	4	6	No	No (0 p.p., 95% CI 0 to 31)	No
Pool meta-analysis‡	IBD	137	36	38	No	No (0 p.p., 95% CI 0 to 16)	—
Goyal *et al.*	IBD	21	21	—	—	—	No
Kump *et al.*	IBD	17	14	—	—	—	No
Vermeire *et al.*	IBD	14	14	—	—	—	Reported§
Nishida *et al.*	IBD	41‖	41‖	—	—	—	No

593 p.p.: Percentage points
594 *: See Results regarding the difficulty in matching statistical methods with the experimental design.
595 †: Median difference in clinical efficacy between two randomly selected pools.
596 ‡: Includes data from Paramsothy *et al.*, Costello *et al.*, and Jacob *et al.* Numbers of patients, donors, and pools are the sum of the values from
597 those studies.
598 ‖: Underlying data not available for re-analysis. Original publication reports that 14 patients were analyzed, but the figure (3a in Vermeire *et al.* only
599 reports data from 13.
600 §: Only 13 of 14 patients were analyzed for a donor effect.
601

602 **Table 2**: Tests of donor superiority, by donor, for Rossen *et al.* Donor labels are

603 arbitrary. “Success” means the patient reached the primary endpoint; “failure” means

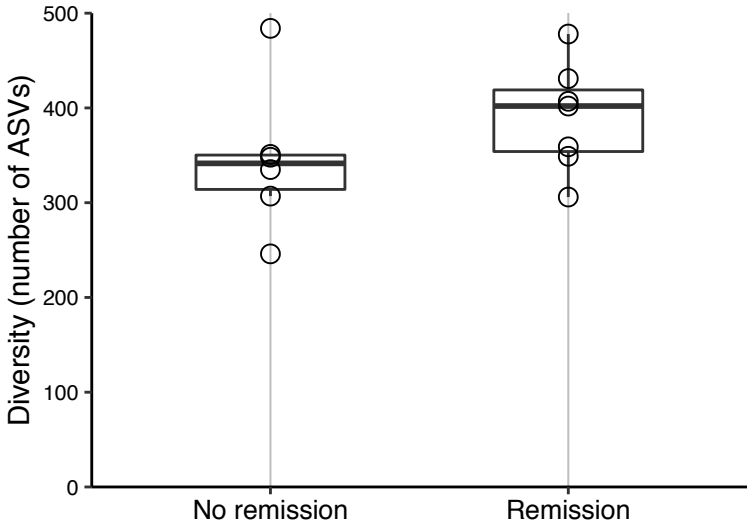
604 they did not.

	Patients treated using this donor		**Patients treated not using this donor**			
Donor	**Success**	**Fail**	**Success**	**Fail**	**Odds ratio (95% CI)**	**Mid-*p* (Fisher’s exact test)**
A	4	4	3	12	3.7 (0.43 to 39)	0.18
B	1	2	6	14	1.2 (0.02 to 27)	0.89
C	0	2	7	14	0.0 (0.0 to 13)	0.47

605

606

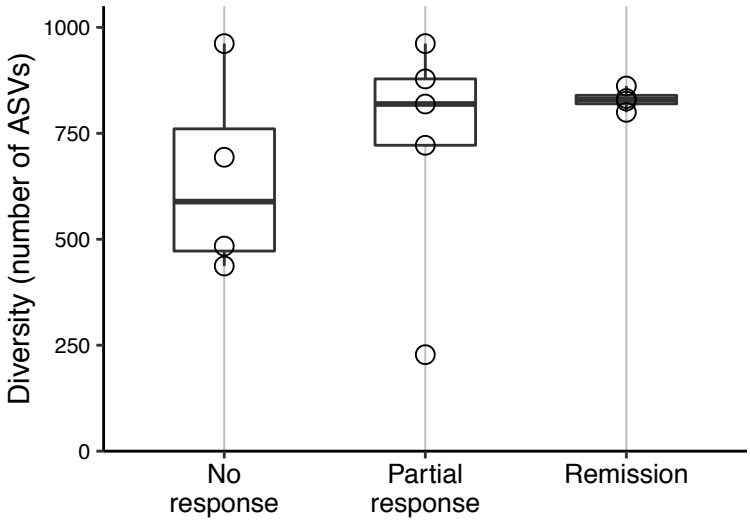
607 **Figure 2**: Bacterial α-diversity of FMT pools (vertical axis) do not significantly differ by
608 patient outcomes (columns), in Jacob *et al.* (p = 0.18, Wilcoxon rank-sum test). Each
609 point represents a single patient. The diversity shown for each patient (vertical axis) is
610 the average of the number of unique 16S rRNA amplicon sequence variants (ASVs) in
611 the sampled FMTs that were used in that patient.



612

613

614 **Figure 3**: Bacterial α-diversity do not differ by patient outcome in re-analysis of Kump *et*
615 *al*.'s data when the unit of analysis is the patient (Wilcoxon test, $p > 0.05$ for all 3
616 comparisons). Each point represents a single patient. The diversity shown for each
617 patient (vertical axis) is the average of the diversity of the sampled FMTs used in that
618 patient.



619

620

621 **SUPPLEMENTAL METHODS**

622 **Calculating the median difference in treatment efficacy between two randomly**
623 **selected donors**

624 To convert a mean and standard deviation, measured in terms of log odds ratios, into
625 the median difference in treatment efficacy between two randomly selected donors,
626 measured in terms of percentage points:

627

628 1. 2 million random variates are drawn from a normal distribution with that mean log
629 odds and standard deviation
630 2. These values are transformed into probabilities
631 3. The difference between subsequent values is calculated, and the median
632 determined

633

634

635 **SUPPLEMENTAL RESULTS**

636 **Moayyedi *et al.* adaptive simulation**

637 To determine this probability, we simulated 10,000 clinical trials. In each trial:

638

639 - 6 donors had 4 success/fail trials with probability of success 9 / 39 = 0.23.
640 - The best performing donor (or, in the case of a tie, a randomly selected donor
641 among the best performers) had 14 additional success/fail trials with the same
642 probability of success

- The performance of the donor with 18 patients was compared against the aggregate performance of the other donors, with 20 patients, using a Fisher's exact test.

The test returned mid-$p < 0.05$ in 8.6% of simulations ($n = 866$; 95% CI 8.1% to 9.2%).