Gemelli IRCCS (ID 3555/2021) and Ospedale Pediatrico Bambino Gesù IRCCS (1107_OPBG_2016). Written informed consent was obtained from all adult participants, and from parents of pediatric (2–17 years old) participants. The systematic review was not registered.

Author contributions

G.I. and N.S. conceived and designed the study. M.P. and N.K. performed the analysis. G.I., N.K. and S.P. performed the literature search. G.I. and G.C. supervised the sample collection and the clinical procedures. G.I., M.V.-C. and N.S. supervised the analysis. M.P., N.K., F. Armanini, F. Asnicar, F.B., A.B.-M., F.C., P.M. and F.P. contributed to data acquisition, data analysis or software development. G.I., S.P., L.M., G.Q., S.D.G., G.D.S., S.B., L.P., F.D.C., M.S., A.G. and G.C. contributed to the clinical procedures and sample collection. G.I., M.P., N.K., M.V.-C. and N.S. interpreted the data and wrote the manuscript. All authors provided critical revision of the manuscript and approved the final version for submission.

Competing interests

A.G. reports personal fees for consultancy for Eisai S.r.l., 3PSolutions, Real Time Meeting, Fondazione Istituto Danone, Sinergie S.r.l. Board MRGE and SanofiS.p.A; personal fees for acting as a speaker for Takeda S.p.A, AbbVie and Sandoz S.p.A; and personal fees for acting on advisory boards for VSL3 and Eisai. G.C. has received personal fees for acting as advisor for Ferring Therapeutics. G.I. has received personal fees for acting as speaker for Biocodex, Danone, Sofar, Malesci, Metagenics and Tillotts Pharma, and for acting as consultant/advisor for Ferring Therapeutics, Giuliani, Metagenics and Tillotts Pharma. N.S. reports consultancy and/or SAB contracts with Zoe, Roche, Ysopia, and Freya, Alia Therapeutics, speaker fees by Illumina, and is cofounder of PreBiomics. The other authors have no potential competing interest to disclose.

Additional information

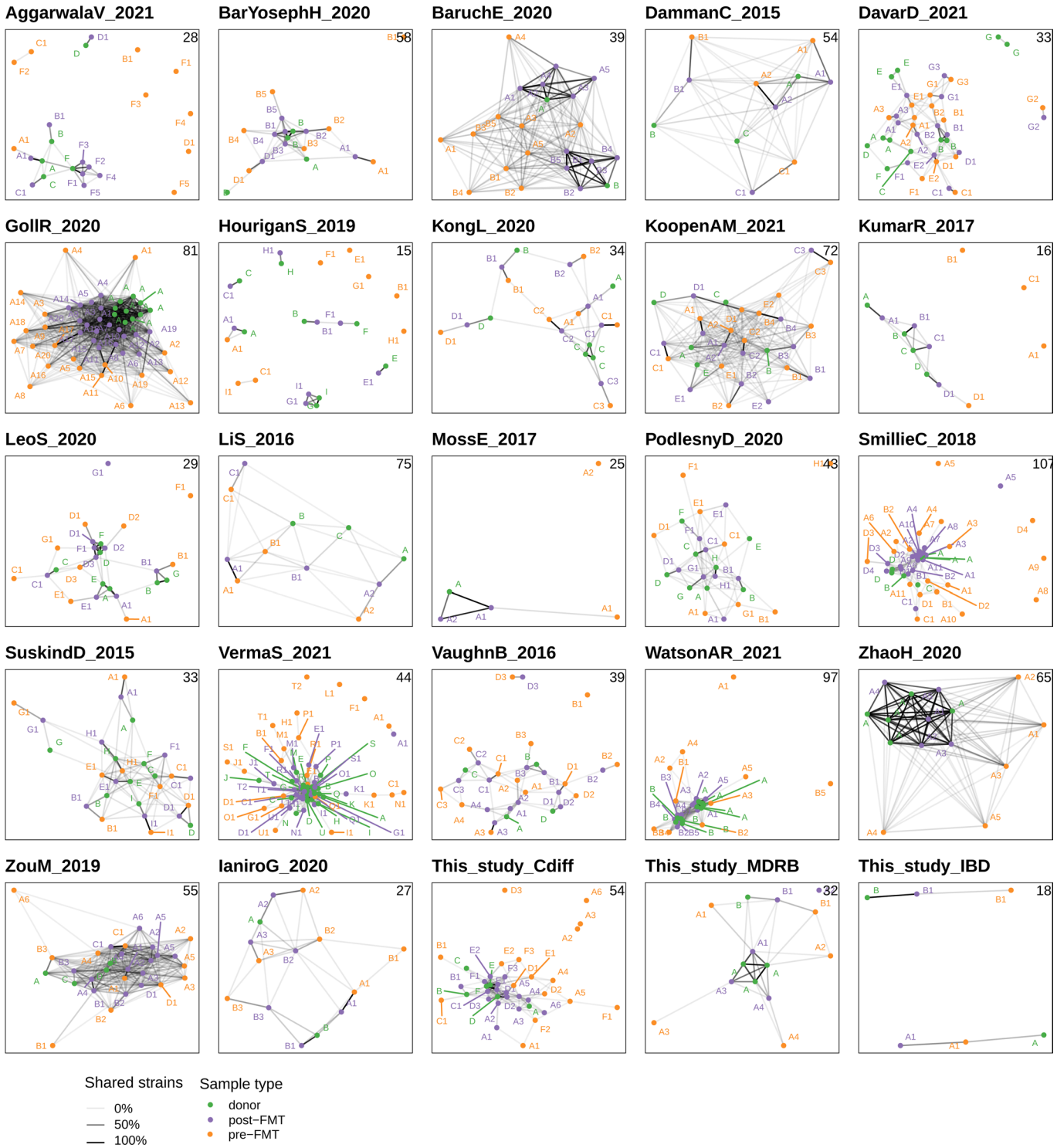
Extended data is available for this paper at https://doi.org/10.1038/s41591-022-01964-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01964-3.

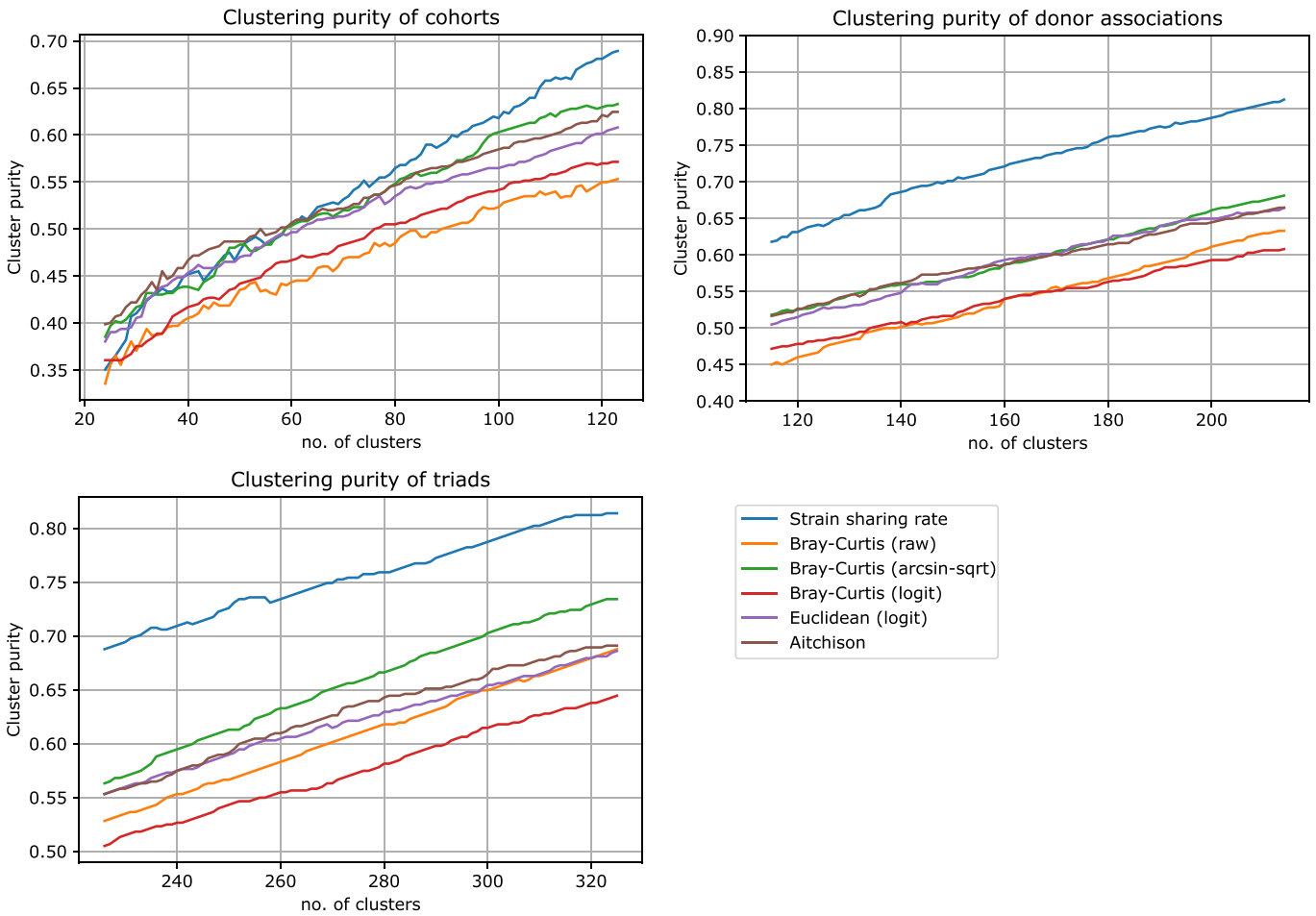
Correspondence and requests for materials should be addressed to Gianluca Ianiro or Nicola Segata.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Alison Farrell, in collaboration with the *Nature Medicine* team.

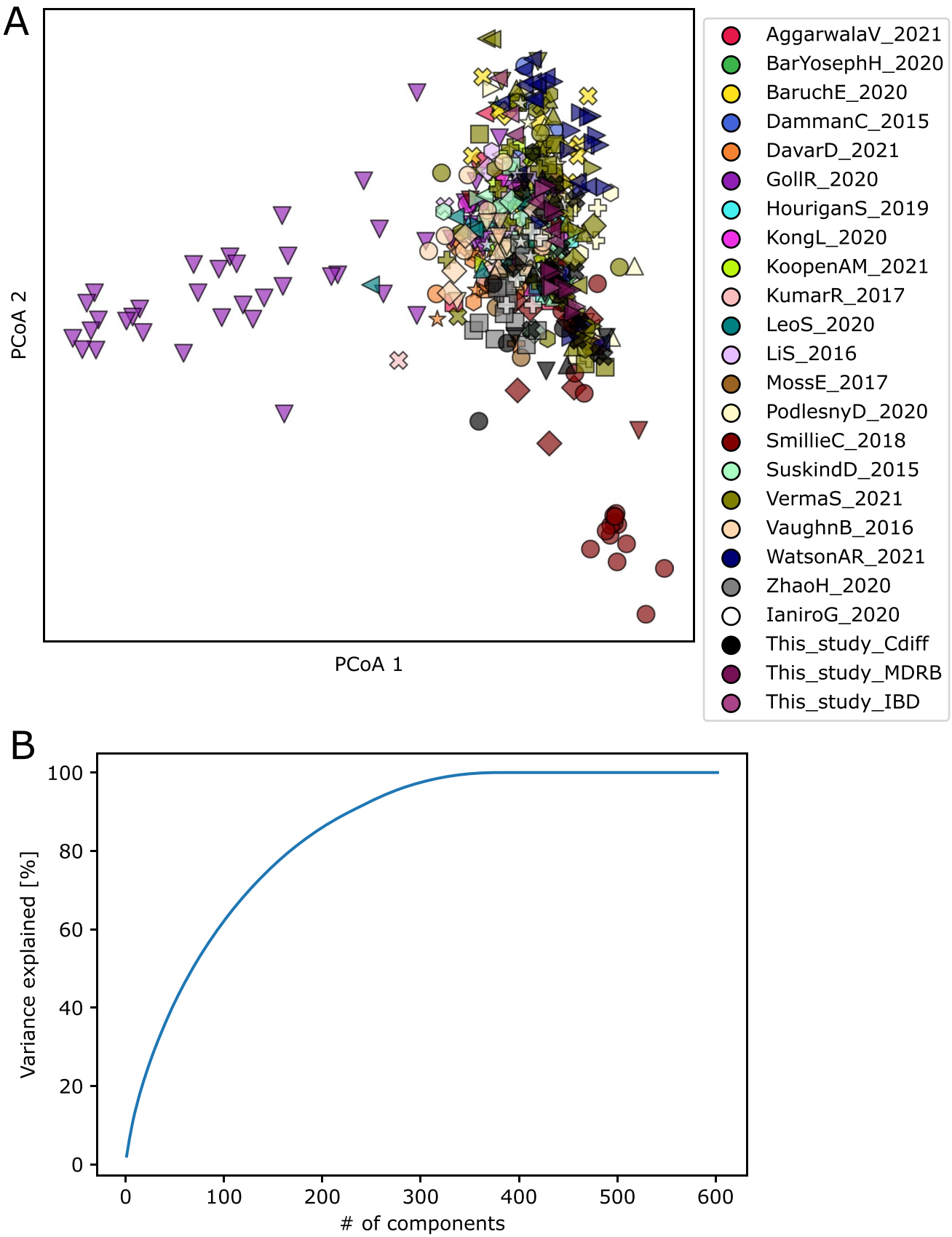
Reprints and permissions information is available at www.nature.com/reprints.



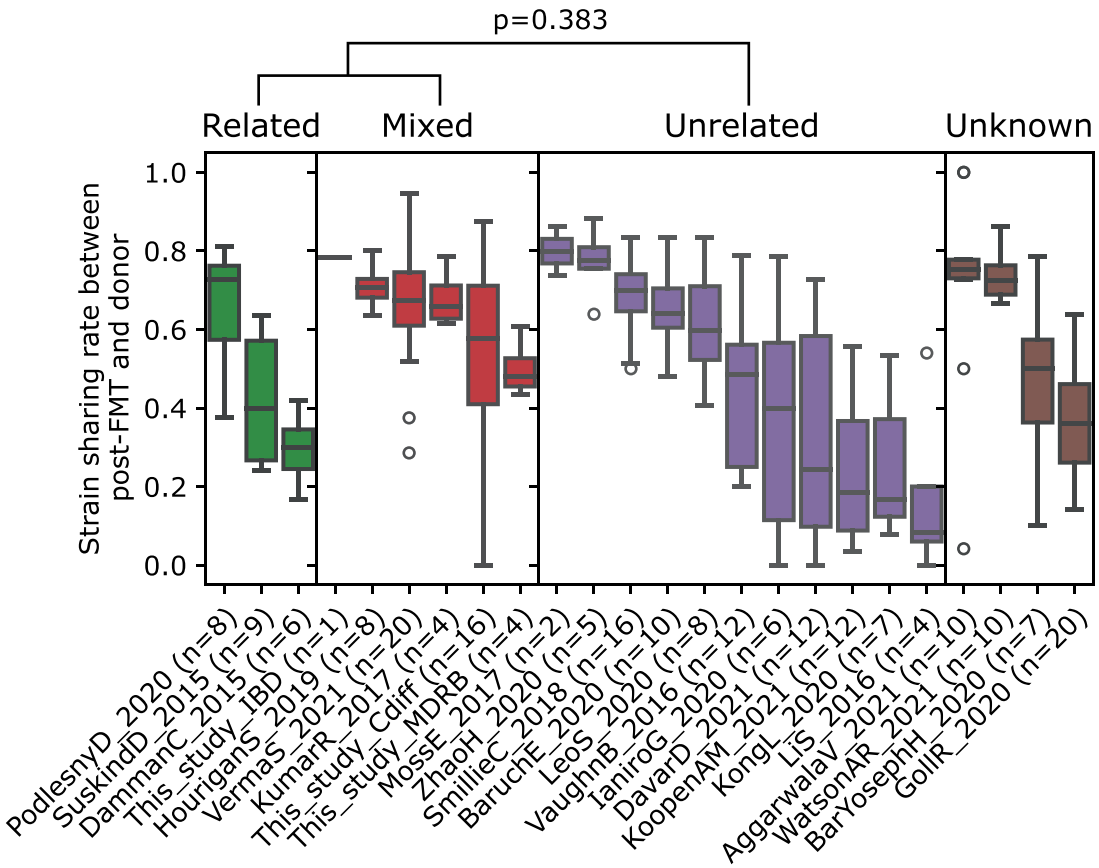
Extended Data Fig. 1 | Strain sharing networks for the datasets included in this study not shown in Fig. 1A. Each node corresponds to a sample and is colored by its role in FMT triads (recipient pre-FMT sample, recipient post-FMT sample, and donor's sample). Edge opacity is proportional to the number of shared strains between two samples (Methods) and only edges corresponding to at least 2 shared strains are shown. The structure of the networks illustrates how FMT triads tend to cluster together but with different clustering characteristics across cohorts.



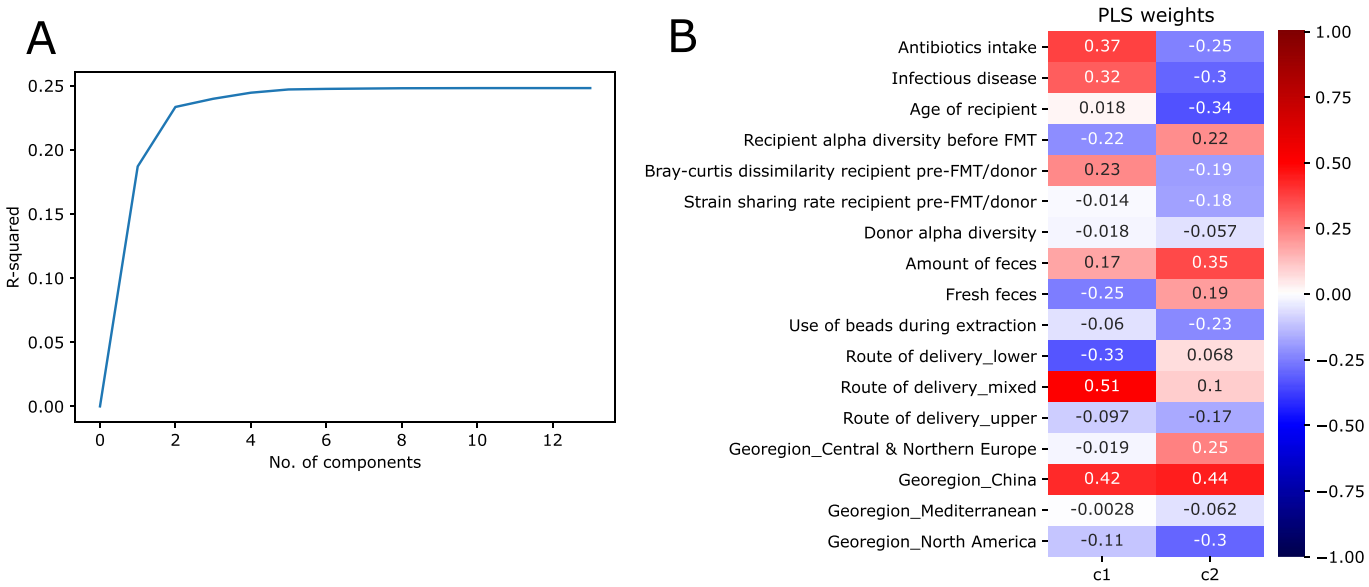
Extended Data Fig. 2 | The purity of K-medoids clustering with varying K shows that strain sharing rate outperforms β-diversity measures in clustering by donor associations and by FMT triads. In clustering by cohorts for the low number of clusters it gets outperformed by Aitchison distance, but catches up as the K increases.



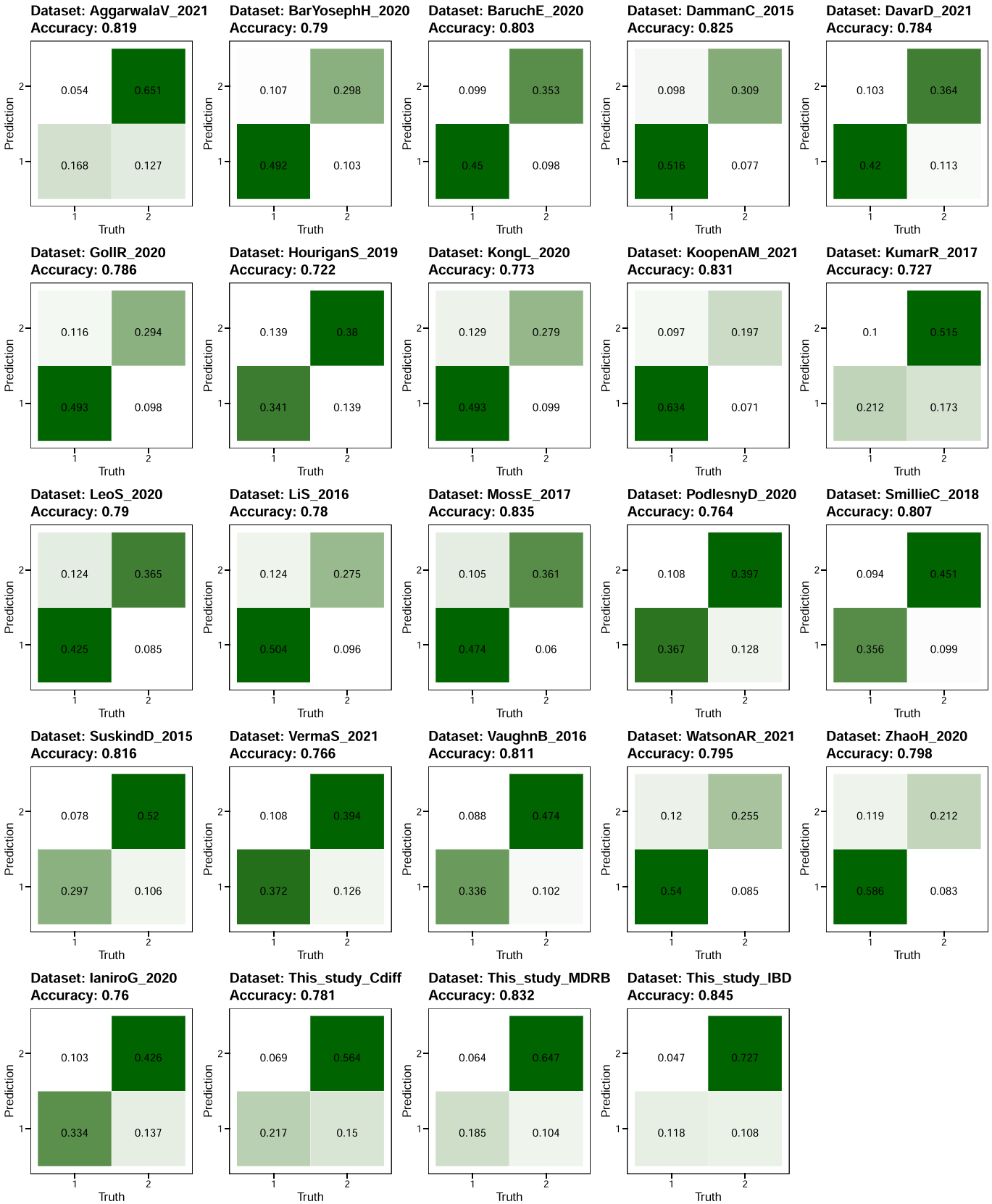
Extended Data Fig. 3 | PCoA ordination on strain sharing rate distances and variance explained by number of components, suggesting that two dimensions are not sufficient to linearly separate the clusters induced by dataset or donor batch effects. Unique combinations of color and shape correspond to samples associated with one donor subject.



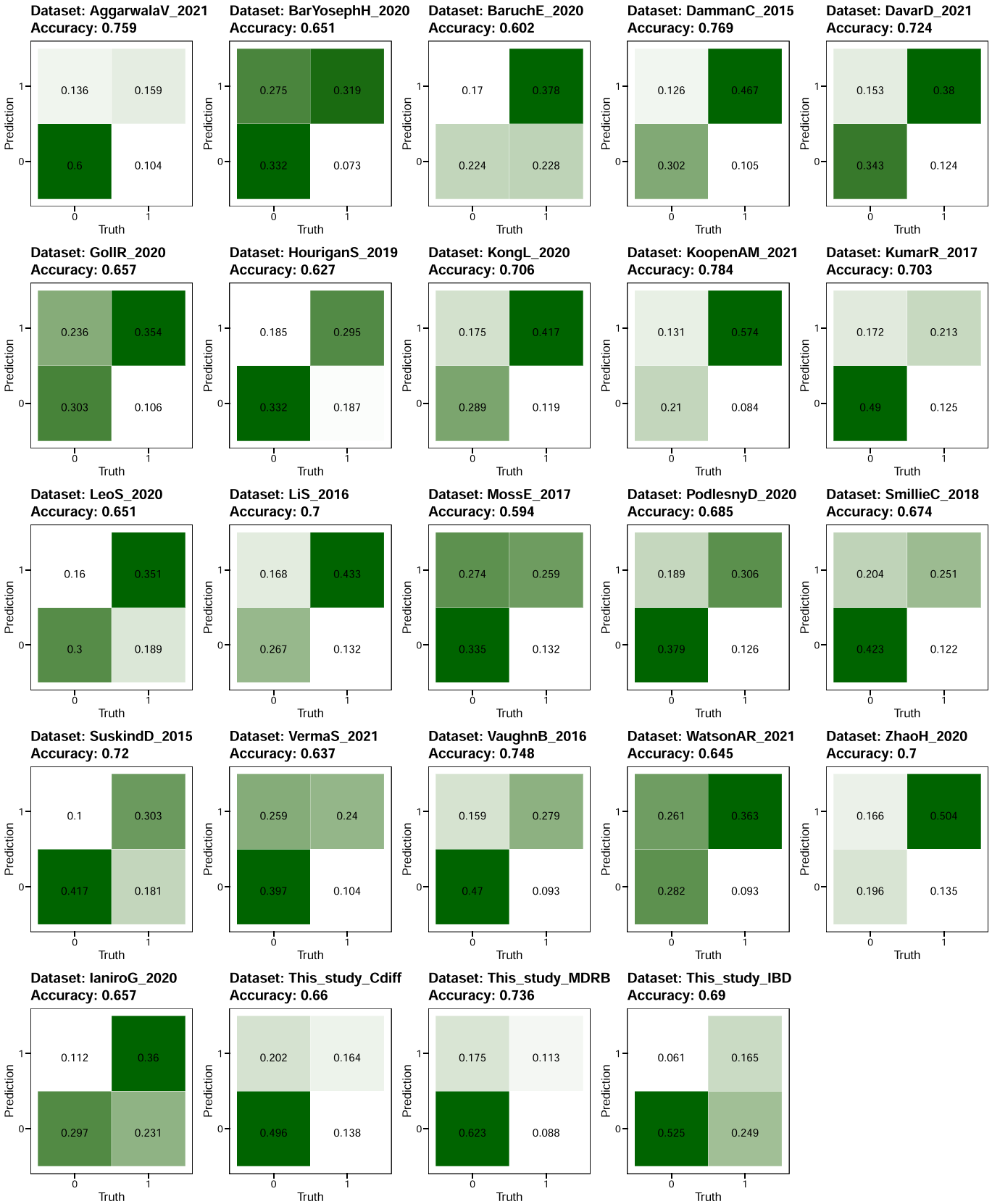
Extended Data Fig. 4 | Strain sharing rates between donor and post-FMT samples is non-significantly higher in datasets using related or a mixture of related and unrelated donors compared to those using only unrelated donors (related or mixed vs unrelated, permutation test, p=0.383). Box plots are defined as follows: the center line and upper and lower limit of the box correspond to the median, upper quartile and lower quartile respectively. The whiskers are defined by that data point that is at most 1.5 times higher than the upper quartile (upper whisker) or 1.5 times lower than the lower quartile (lower whisker).



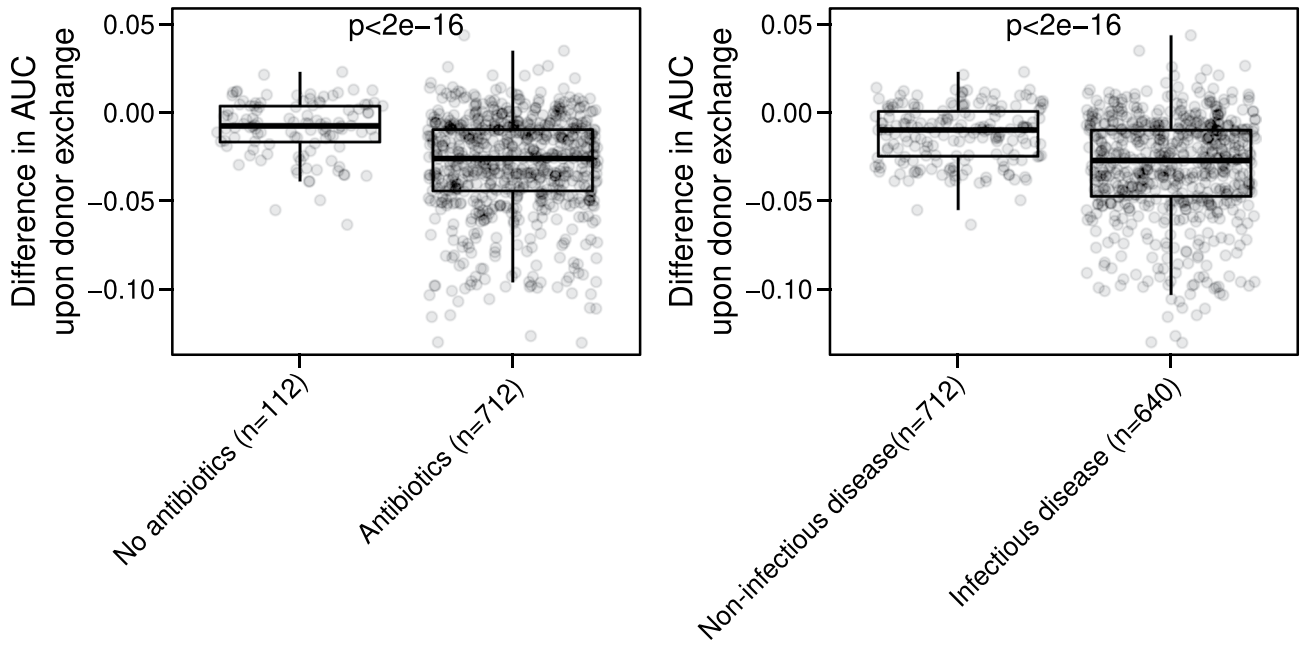
Extended Data Fig. 5 | Partial least squares regression of various variables of interest against strain engraftment rate. A) Most of the explained variance in strain engraftment rate is covered by the first two components. B) The weights of the variables in the first two components.



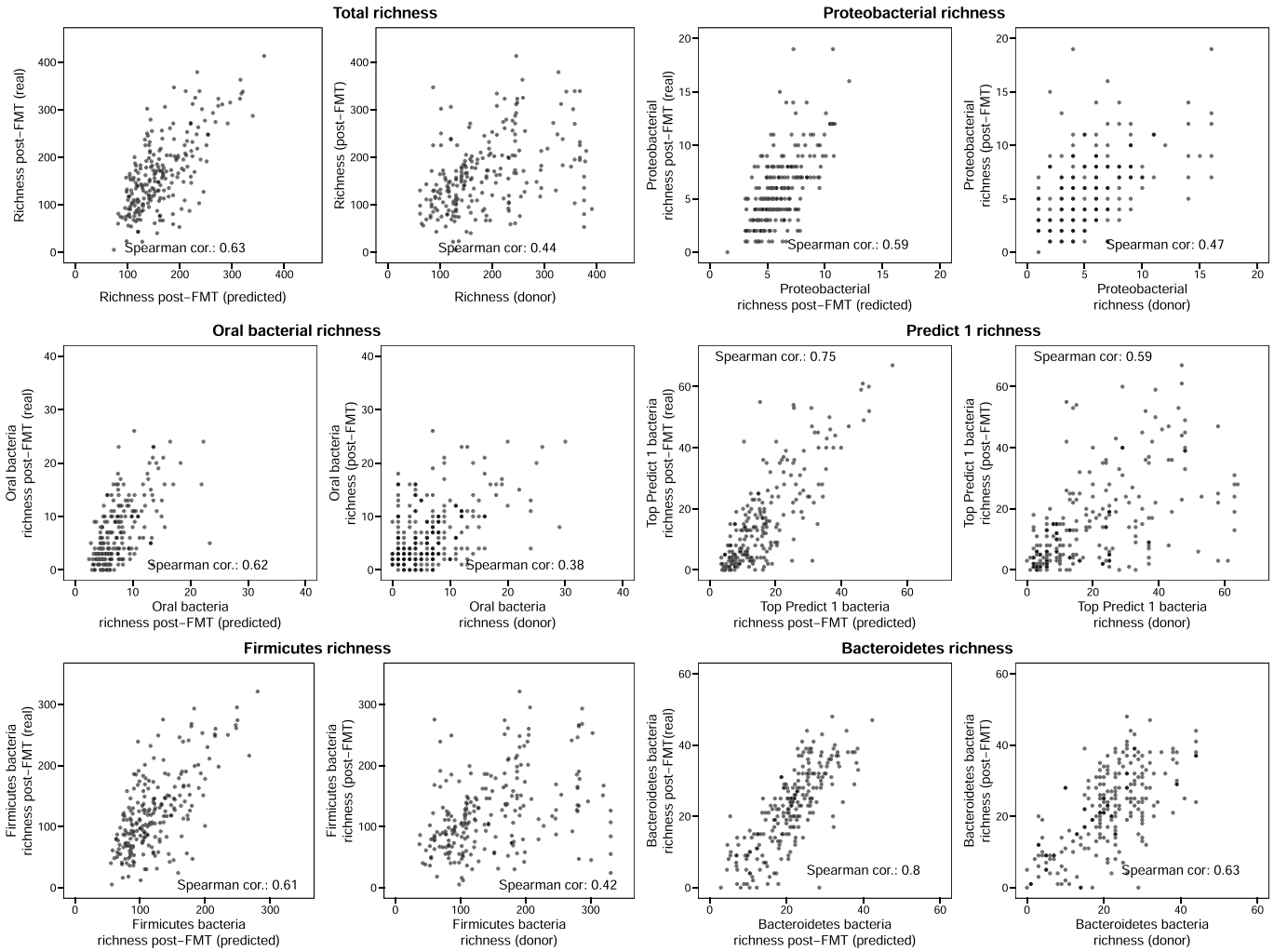
Extended Data Fig. 6 | Random forest classifier prediction accuracies of post-FMT species presence/absence (CV).



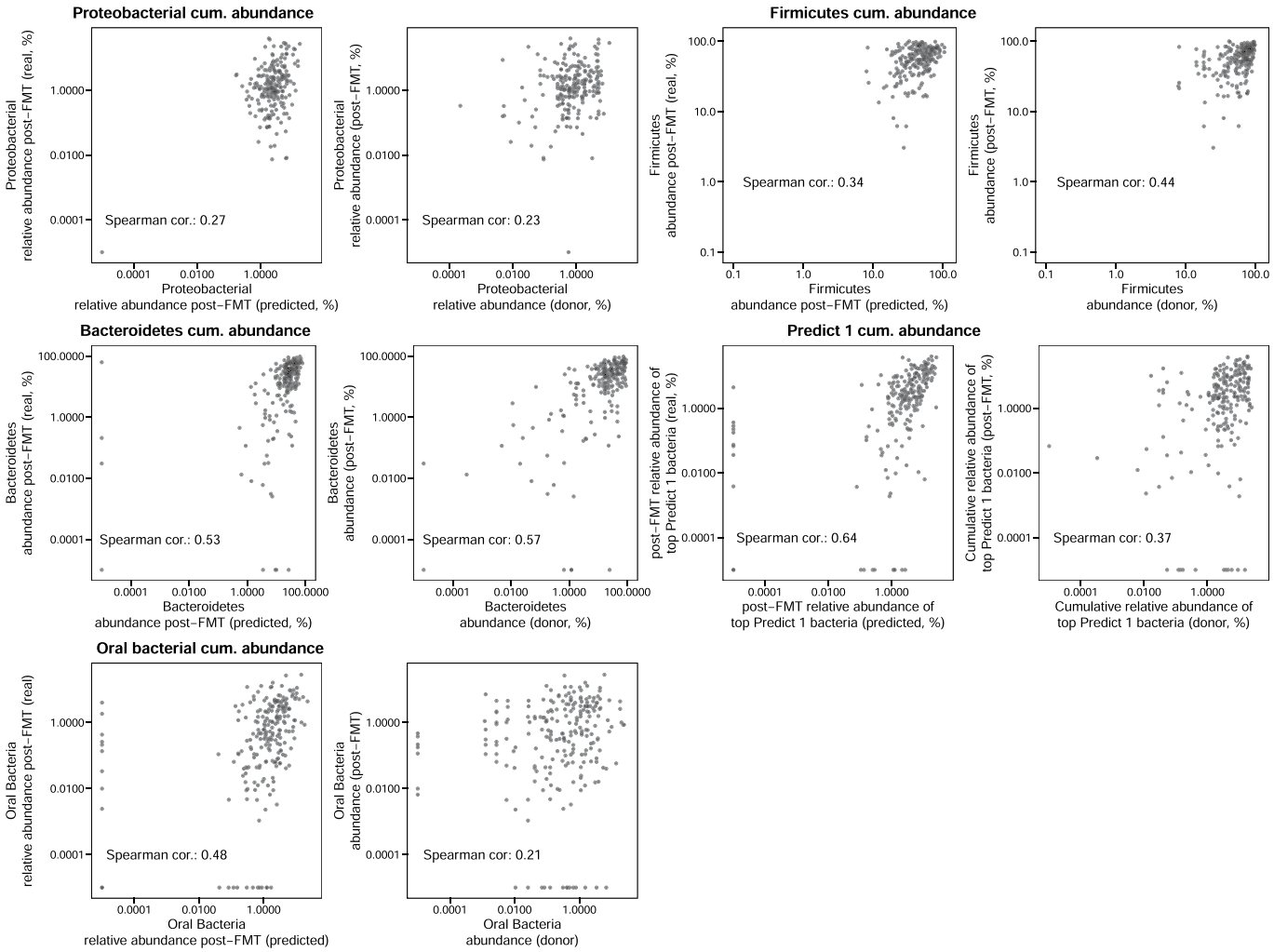
Extended Data Fig. 7 | Random forest classifier prediction accuracies of post-FMT species presence/absence (LODO).



Extended Data Fig. 8 | Boxplots of the difference in AUC upon simulated donor exchange. Mann-Whitney U-test two-tailed p<2e-16 for both infectious vs. non-infectious disease and antibiotics vs. no antibiotics comparisons. Box plots are defined as follows: the center line and upper and lower limit of the box correspond to the median, upper quartile and lower quartile respectively. The whiskers are defined by that data point that is at most 1.5 times higher than the upper quartile (upper whisker) or 1.5 times lower than the lower quartile (lower whisker).



Extended Data Fig. 9 | Comparisons of the predicted total species richness of bacterial groups in post-FMT samples. Predictions on the y-axis come from the RF classifier, predictions on the x-axis correspond to the cumulative richness in donor samples.



Extended Data Fig. 10 | Comparisons of the predicted cumulative abundance of bacterial groups in post-FMT samples. Predictions on the y-axis come from the RF regressor, predictions on the x-axis correspond to the cumulative abundance in donor samples.

nature portfolio

Corresponding author(s): Gianluca Ianiro, Nicola Segata

Last updated by author(s): Jul 14, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for microbiome data collection.
Data analysis	Newly-generated shotgun metagenomic sequences were pre-processed and quality controlled using the pipeline available at https://github.com/SegataLab/preprocessing and KneadData within bioBakery 3. Species-level profiling was performed using MetaPhlAn 3. Core gene fragments were aligned against the genomes of all SGBs using Bowtie2 (version 2.3.5.1; --sensitive option). Strain profiling was performed with a modified version of StrainPhlAn 3 using the custom SGB marker database. SGB phenotypes were predicted using Traitar (version 1.1.12). Statistical analyses and graphical representations were performed with R (packages curatedMetagenomicData v3.15, igraph 1.2.6, vegan 2.5-7, ggplot2 3.3.5, mlr3 0.10) and Python (packages scikit-learn version 0.24.2, scikit-learn-extra 0.2.0, SciPy 1.8.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Newly-generated shotgun metagenomics sequencing data are available at the European Nucleotide Archive under accession number PRJEB47909. Metadata are available in Supplementary Table 2 and in curatedMetagenomicData.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The sex of the study participants is reported in Supplementary Table 1 whenever available. Due to lack of availability in all published studies, this was not added as a covariate in the multivariate models.
Population characteristics	Population characteristics of the 24 studies included in the analysis are reported in Supplementary Tables 1 and 2. 17 cohorts including 147 FMT recipients, 98 donors and 882 samples were included in the analysis, together with three novel cohorts enrolled for this study including 23 recipients, 8 donors and 115 samples.
Recruitment	Three cohorts were newly enrolled, collected and sequenced in the context of this study (Fondazione Policlinico Gemelli IRCCS and Bambino Gesu Children's Hospital, Italy), as detailed in the Methods section. Participants that underwent FMT were recruited in the study, limitting the possibility of self-selection bias.
Ethics oversight	Study procedures of the newly-collected datasets were performed in compliance with the Declaration of Helsinki. Ethical approval was granted by Ethics Committees of Fondazione Policlinico Gemelli IRCCS (ID 3555/2021) and Ospedale Pediatrico Bambino Gesù IRCCS (1107_OPBG_2016). Written informed consent was obtained from all adult participants, and from parents of underaged participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In this meta-analysis we systematically searched PubMed, Scopus, and ISI Web of Knowledge for potentially eligible studies. In addition, three novel cohorts were newly-sequenced for this study, making it the largest set available. Due to lack of previous data, no power calculation could be performed as part of the study design.
Data exclusions	No data were excluded from the analyses.
Replication	The data included in this study was meta-analysed, no replication was performed.
Randomization	Samples were allocated to FMT triads based on previous published information (metadata of each study) or inferred based on strain sharing patterns when this information was not available. No new groupings were set as part of this analysis.
Blinding	All patients were administered FMT from donors (no placebo). The information on whether samples were from FMT donors or recipients were available to the researchers as this was needed to perform the statistical analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging