nature research

Corresponding author(s): Nicola Segata

Last updated by author(s): Feb 11, 2019

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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 commercial software has been used for data collection, all the software used is reported in the paper
Data analysis	All data analysis has been performed with open source software as comprehensively described in the methods of the paper

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Nucleotide sequences for the two new Italian cohorts are available in the Sequence Read Archive (SRA) under the accession number SRP136711. MetaPhlAn2 and HUMANn2 profiles for the new cohorts were also added to the curatedMetagenomicData R package along with their corresponding metadata. Validation Cohort1 is available in the European Nucleotide Archive (ENA) under the study identifier PRJEB27928, Validation Cohort2 is available in the DDBJ databases under the accession number DRA006684.

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	This is a meta-analysis of 9 shotgun metagenomic datasets for a total of 969 samples, details on the sample size of each dataset is reported in the paper
Data exclusions	Samples not passing quality-control have been excluded. Details of the exclusion criteria are reported in the paper
Replication	This is a meta-analysis so the main focus is indeed on the reproducibility. We thus used cross validaton, cross-dataset prediction, leave-one-dataset-out validation, and independent validation on additional cohorts. qPCR measurements were also done in triplicates
Randomization	All datasets are from a case/control design as described in the original publications and in our method section
Blinding	Samples were collected from treatment- and diagnosis- naive subjects

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
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	We report this information in Table 1 and Suppl. Table 1
Recruitment	All subjects meeting the inclusion criteria were enrolled until the required sample size was reached
Ethics oversight	The two new clinical studies performed here were approved by the relevant ethics committees (Cohort1: Ethics committee of Azienda Ospedaliera "SS. Antonio e Biagio e C. Arrigo" of Alessandria, Italy, protocol N. Colorectal_miRNA_CEC2014 and Cohort2: Ethics committee of European Institute of Oncology of Milan, Italy, protocol N. R107/14-IEO 118) and informed consent was obtained from all participants

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