GigaScience, 7, 2018, 1–8

doi: 10.1093/gigascience/giy032
Advance Access Publication Date: 2 April 2018
Technical Note

TECHNICAL NOTE

A practical tool for maximal information coefficient analysis

Davide Albanese, Samantha Riccadonna, Claudio Donati and Pietro Franceschi*

Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy

*Correspondence address. Computational Biology Unit, Research and Innovation Centre, Fondazione Edmund Mach, via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy E-mail: pietro.franceschi@fmach.it

Abstract

Background: The ability of finding complex associations in large omics datasets, assessing their significance, and prioritizing them according to their strength can be of great help in the data exploration phase. Mutual information-based measures of association are particularly promising, in particular after the recent introduction of the TIC_e and MIC_e estimators, which combine computational efficiency with superior bias/variance properties. An open-source software implementation of these two measures providing a complete procedure to test their significance would be extremely useful. **Findings:** Here, we present MICtools, a comprehensive and effective pipeline that combines TIC_e and MIC_e into a multistep procedure that allows the identification of relationships of various degrees of complexity. MICtools calculates their strength assessing statistical significance using a permutation-based strategy. The performances of the proposed approach are assessed by an extensive investigation in synthetic datasets and an example of a potential application on a metagenomic dataset is also illustrated. **Conclusions:** We show that MICtools, combining TIC_e and MIC_e, is able to highlight associations that would not be captured by conventional strategies.

Keywords: maximal information coefficient; MIC; TIC; equitability; multiple testing; permutation test; power of statistical significance; false discovery rate; FDR

Introduction

With the growing popularity of high-throughput quantitative technologies, it is now common to characterize living systems by measuring thousands of variables over a wide range of conditions. In these large datasets, the number of potential associations between variables is enormous. Computational and statistical methods should be able to highlight the significant ones (striking a balance between flexibility and statistical robustness) and to prioritize the more relevant for downstream analysis. Traditionally, the presence of a potential relationship between two variables X and Y is assessed on the basis of a certain measure of association that is often able to reveal specific types of relationships but it is blind to others. Then, once the measure is computed, its significance is tested against the null hypothesis of no association. For linear associations, the Pearson correlation coefficient is the natural choice, while the Spearman rank coefficient represents a more flexible alternative for general monotonic relationships. In the exploratory analysis of datasets produced by modern -omics technologies, this conventional approach shows its limits, because a huge number of potential associations needs to be screened without any *a priori* information on their form. In these cases, it would be desirable to use a measure of dependence that ranks the relationships according to their strength, regardless of the type of association. A measure with this property has been defined *equitable* [1], and a consistent mathematical framework for the definition of equitability has been proposed [2–6]. The second challenge faced in the

Received: 6 November 2017; **Revised:** 8 March 2018; **Accepted:** 22 March 2018

unsupervised screening of large datasets is that the number of associations to be tested is usually huge and the statistical assessment of significance has to face well-known multiplicity issues [7,8].

Recently, a family of measures based on the concept of mutual information has been proposed, and one of the most popular (and debated) members of this family, the maximal information coefficient (MIC), has been shown to have good equitability [1]. Unfortunately, MIC does not have state-of-the-art power [9,10], and its heuristic estimator, APPROX-MIC, is computationally demanding [5]. These two drawbacks have severely hampered the application of MIC to large datasets. In order to overcome these limitations, two new MIC-based measures, the MIC_e—a consistent estimator of the MIC population value (MIC^*) and the related TIC_e (total information coefficient) statistics—have been proposed [5]. Both quantities can be calculated more efficiently than APPROX-MIC and have better bias/variance properties [5]. In particular, TIC_e is characterized by high power, which has been obtained at the cost of equitability, while MIC_e performs better on this side, showing reduced performances in terms of power. These two MIC-based measures compensate each other, and their combination is extremely promising as a data exploration tool. In particular, a two-step procedure can be applied where TIC_e is used to perform efficiently for a high-throughput screening of all the possible pairwise relationships and assess their significance, while MIC_e is used to rank the subset of significant associations in terms of strength [5]. Despite the potential of this approach, an efficient software implementation of these two measures and of a statistical procedure to test the significance of each association controlling multiplicity issues is still lacking.

Here we present MICtools, an open-source and easy-to-use software that provides:

- an efficient implementation of TIC_e and MIC_e estimators [11];
- a permutation-based strategy for estimating TIC_e empirical p values;
- several methods for multiple testing correction, including the Storey q value to control the false discovery rate (FDR); and
- the MIC_e estimates for each association called significant.

Methods

MICtools implements a multistep procedure to identify relevant associations among a large number of variables, assess their statistical significance, and rank them according to the strength of the relationship. Starting from M variable pairs x_i and y_i measured in n samples, the procedure can be broken into 4 steps (Fig. 1):

1. Estimating the empirical TIC_e null distribution by permutations.
2. Computing TIC_e statistics and the empirical p value for each variable pairs.
3. Applying a multiple testing correction strategy in order to control the family-wise error rate (FWER) or the FDR [12].
4. Using MIC_e to estimate the strength of the relationships called significant.

The pipeline can be run as a sequence of subcommands implemented into the main command `mictools` (Fig. 1).

The empirical TIC_e null distribution

Since TIC_e depends only on the rank-order of the vectors x_i and y_i [1], the empirical null distribution can be estimated for a given sample size and set of parameters by performing R permutations of the elements of the vectors y_i and by calculating the set of null TIC_e statistics $t_1^0, \ldots, t_R^0$. Two parameters control the estimation of the null distribution of TIC_e: the parameter B controlling the maximal-allowed grid resolution and the number of permutations R. In the current implementation, B was set to the default value 9, which guarantees good performance in terms of statistical power against independence in most situations [10]. However, different values of B can be chosen; for example, $B = 4$ for less complex alternative hypothesis, $B = 12$ for more complex associations [10]. With regard to the number of permutations, instead, the results obtained on the synthetic datasets (see Additional File 2, Figs. A2 and A3 and Additional File 1, Table A2) empirically indicate that 200,000 permutations represent a reasonable choice for the dataset SD1 (see the Synthetic datasets section).

Computing the TIC_e and its associated empirical p values for each variable pair

The TIC is computed for each (nonpermuted) variable pair, obtaining a set of TIC_e values t_i (with $i = \{1, \ldots, M\}$). For each t_i, the p value p_i is estimated as the fraction of values of the empirical null distribution that exceeds t_i [13]:

$$p_i = \frac{1 + \#\{r : t_r^0 \geq t_i, \quad r = 1, \ldots, R\}}{1 + R}$$

Multiple testing correction

Considering the large number of tests of independence performed, it is necessary to correct the p values for multiplicity. In general, this can be done either by controlling the FWER or the FDR. The first approach aims at controlling the probability of making at least one type I error in the set of tests, and this is done by decreasing the significance threshold of each individual test (as in Bonferroni correction). In the case of FDR, the presence of false positives (FPs) is accepted and what is controlled is their fraction among the associations called significant. This is done by estimating the distribution of the p values under the hypothesis of independence and comparing it with the observed one. MICtools implements several state-of-the-art strategies to accomplish this task. For all the examples presented here, we have used the Storey method for estimating the q values to control the FDR [7]. Assuming a uniform distribution for the null p values, the fraction of associations for which the null is true (π_0) is estimated directly from the shape of distribution at high p values. π_0 is then used to calculate the q value for the i^{th} association as the minimum FDR that can be obtained varying the significance threshold (h):

$$q(p_i) = \min_{h \geq p_i} \mathrm{FDR}(h) = \min_{h \geq p_i} \frac{\pi_0 M h}{\#\{p_i \leq h\}}.$$

Briefly, setting a q value cutoff to 0.05, we accept an FDR of 5% at most. To check the method assumptions, MICtools provides the empirical distribution of p values as a diagnostic plot.

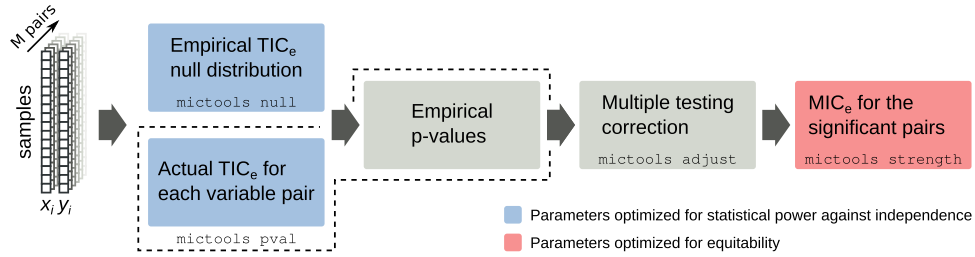

Figure 1: The MICtools pipeline. Each step is implemented as a subcommand of the `mictools` main command. `mictools null` estimates the empirical TIC$_e$ null distribution of the M variable pairs (x_i, y_i). `mictools pval` computes the TIC$_e$ values and estimates their p values (boxes within the dashed line). The multiple testing correction is performed by `mictools adjust`. Finally, `mictools strength` estimates the MIC$_e$ value for the subset of significant relationships. The color of the boxes highlights the criterion used for parameter optimization.

Table 1: Default values of the α parameter vary according to the number of samples

Number of samples	α parameter
$n < 25$	0.85
$25 \leq n < 50$	0.80
$50 \leq n < 250$	0.75
$250 \leq n < 500$	0.70
$500 \leq n < 1{,}000$	0.65
$1{,}000 \leq n < 2{,}500$	0.60
$2{,}500 \leq n < 5{,}000$	0.55
$5{,}000 \leq n < 10{,}000$	0.50
$10{,}000 \leq n < 40{,}000$	0.45
$n > 40{,}000$	0.40

Computing the MICe on the significant relationships

Finally, the strength of the associations that pass the significance threshold is estimated using the MIC$_e$ estimator. In this case, we define the B parameter as a function of the number of samples n, $B(n) = n^{\alpha}$ [1]. The default values are optimized for equitability [6] and summarized in Table 1.

Findings

Two synthetic datasets (SD1 and SD2) were created in order to assess the statistical power (or recall, i.e., the fraction of nonindependent relationships that were recovered at a given significance level) and the ability to control the FDR. The analyses were performed varying the number of samples (SD1) and the effect chance [14], i.e., the percentage of nonindependent variable pairs (SD2). Both datasets contain a set of independent variables and a fixed number of variable pairs X and Y related by associations in the form $Y = f(X) + \eta$, where $f(X)$ is a function and η is a noise term. To characterize the performance of MICtools in the presence of associations that could not be described by a function, a series of Madelon datasets [15,16] was also analyzed. The main characteristics of the three synthetic datasets are summarized in Table 2. Finally, the proposed pipeline was applied to the analysis of an environmental/metagenomic dataset that has been recently made available within the Tara project, a global-scale characterization of plankton using high-throughput metagenomic sequencing [17].

Synthetic datasets

SD1 contains 60,000 associations between variable pairs X and Y. The effect chance was set to 1%. The relationships between the 600 nonindependent variable pairs were randomly chosen among six types of functional associations, namely, cubic, exponential (2^x), line, parabola, sigmoid, and spike (see Table S3 in [1]). The noise term η is a random variable with uniform distribution in the range of $f(X)$ multiplied by an intensity factor k_η. Different values of k_η were chosen randomly among 18,000 values obtained joining the following three sequences: the first ranging from 0.05 to 1 (with steps of 0.0001), the second ranging from 1 to 2 (with steps of 0.0002), and the third ranging from 2 to 9 (with steps of 0.002). Using these values, the coefficient of determination (R^2) between Y and the noiseless function $f(X)$ ranges from approximately 0 to 1. The remaining 99% (59,400) ofassociations were defined with X and Y randomly generated from a uniform distribution between 0 and 1. To characterize the effect of the sample size, we created 20 replicates of SD1 for an increasing number of samples ($n \in \{25, 50, 100, 250, 1{,}000\}$), for a total of 100 datasets. Considering that the fraction of true positive associations was known, this design of experiment allowed us to quantify the statistical power and the performances in terms of FDR of the proposed pipeline. The results for 2×10^5 permutations are summarized in Fig. 2 and in Additional File 1, Table A1. The dependence of the power and of the number of FPs from the number of samples is shown in Fig. 2A and Fig. 2B. The power increases with the number of samples, reaching 75% for a sample size of 100. As expected, considering that we used the Storey q value as a strategy to control the FDR, the number of FPs also grows for increasing sample size (Fig. 2B) to keep the false discovery rate constant (0.05 in this case). Fig. 2C shows the observed FDR, which is almost equal to the expected value of 0.05 for all sample sizes. In Fig. 2D we show the values of MIC$_e$ as a function of the coefficient of determination (R^2) between Y and the noiseless function $f(X)$ for the associations that pass the significance filter (i.e., associations with q values <0.05). As expected, MIC$_e$ and R^2 were always linearly correlated, especially for the larger sample sizes [5] (Fig. 2D, upper panel). Moreover, we found that for small sample sizes, only relationships with relatively high values of R^2 passed the significance filter. This effect decreases with increasing number of samples, showing that the pipeline is able to identify relationships with more noise, provided that a sufficient number of experimental points is available. This effect is clearly visible in Fig. 2E, where we show the statistical power as a function of the strength of the relationships for different sample sizes. While on less noisy associations (having R^2 close to 1) the pipeline shows high power also for smaller sample sizes, a high number of samples is needed to attain high power for very noisy relationships (having R^2 close to 0). Upon closer inspection, panel D in Fig. 2 also shows that the power depends on the form of the association. For instance, red points (corresponding to cubic functional forms) are hardly

Table 2: Characteristics of the three synthetic datasets analyzed in this work

Characteristic	SD1	SD2	Madelon
N. associations	60,000	60,000	19,900
Effect chance (%)	1	1, 2, 5, 10, 20, 50	1
N. samples	25, 50, 100, 250, 1,000	100	50, 250, 1,000, 2,500, 5,000
N. replicates	20	20	1
Total n. replicates	100	120	7

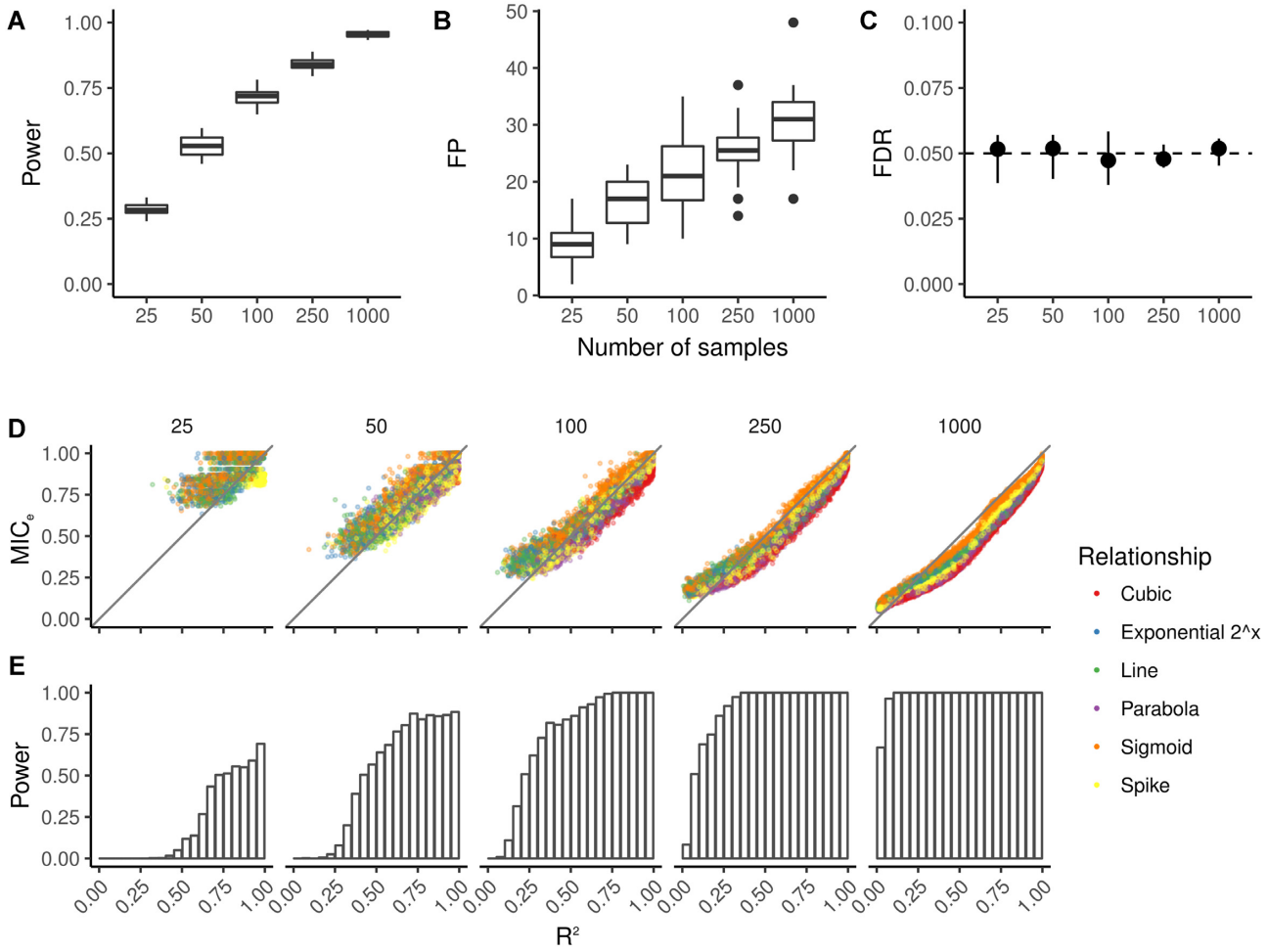

Figure 2: Analysis on SD1 dataset at the 0.05 significance level. A) Statistical power, B) number of FPs, and C) FDR for varying number of samples *n*. Each range represents the results of the 20 replicates. D) MIC_e values and E) statistical power at different levels of R^2, for increasing number of samples (from 25 to 1,000, plots from left to right). Only significant relationships, i.e., relationships with $q < 0.05$, are shown.

visible for sample sizes smaller than 100, while sigmoidal, linear, and exponential relationships can be identified for all sample sizes, albeit with a power that depends on the amount of noise. This finding can be easily interpreted considering that more complex relationships (e.g., polynomials of higher order) are defined by a higher number of parameters that makes them more difficult to distinguish from random associations if the number of points is limited. A clearer representation of this phenomenon is included in Additional File 2 (Fig. A1). Moreover, the downward bias in terms of equitability, especially for the more complex relationships (Fig. 2D and A1), is a result of the core approximation algorithm EQUICHARCLUMP, which speeds up the computation of MIC_e [5,18]. The EQUICHARCLUMP parameter *c* controls the coarseness of the discretization in the grid search phase; by default, it is set to 5, providing good performance in most settings [10].

As anticipated, SD1 was also used to investigate the dependence of the performances of MICtools on the number of independent permutations used to estimate the empirical null distribution. Figures A2 and A3 (Additional File 2) show the FDR and the power as a function of the number of samples and of the number of permutations. The plots indicate that for all the combinations of the two parameters, the measured FDR was consistent with the expected value 0.05 (Additional File 2, Fig. A2 and

Additional File 1, Table A2) and that the true value is always included in the shaded interquartile area. As expected, the variability is stronger for the smaller dataset (25 samples); however, with such a small number of samples above 2×10^5 permutations, the median measured FDR stabilizes around 0.05. Figure A3a (Additional File 2) shows the expected increase in power with the number of samples, from 0.25 to almost 1. The median value does not show a strong dependence on the number of permutations. Figure A3b indicates that below 100 samples, at least 2×10^5 permutations are needed to obtain stable values of power and that its variability is anyway larger for small sample sets. In MICtools, the default value of the number of permutations is set to 2×10^5, and the parameter can be optimized by the user on the basis of the characteristics of the dataset under analysis.

SD2 was generated to characterize how the effect of chance, i.e., the fraction of nonrandom associations, affected the performance of MICtools. Similar to SD1, SD2 contains a subset of variable pairs X and Y related by associations of the form $Y = f(X) + \eta$, where η is defined as in SD1. The number of samples was fixed at n= 100, and the total number of associations was 60,000. For each effect chance value (1%, 2%, 5%, 10%, 20%, and 50%),we generated 20 independent datasets, for a total of 120. The power, number of FPs, and FDR as a function of the effect chance are shown in Fig. 3, panels A, B, and C, respectively (see also Additional File 1, Table A3). In Fig. 3A, we can observe that the statistical power grows with the effect chance, while the actual FDR remains constant. In fact, an increase of effect chance corresponds to a decrease of the fraction of relationships for which the null is true, π_0 (effect chance $= 1 - \pi_0$). Consequently, an increase in the p-value threshold and therefore a growth of power is expected in order to maintain the FDR cutoff constant [7,14].

The Madelon classification dataset

The analysis of SD1 and SD2 demonstrates that MICtools is able to identify the relationships described by analytic functions with additive noise. However, more general forms of nonrandom associations are possible. Consider, for instance, the presence of clusters that might indicate the presence of subpopulations. To test the ability of MICtools to identify this type of association, we created seven datasets with an increasing number of samples $n \in \{50, 250, 500, 1{,}000, 2{,}500, 5{,}000\}$ with a structure similar to the Madelon binary classification dataset [15,16] (http://archive.ics.uci.edu/ml/machine-learning-databases/madelon/Dataset.pdf) using the `datasets.make_classification()` function available in the scikit-learn library [19]. Each dataset contains four clusters (two for each class), placed on the vertices of a five-dimensional foursided hypercube. Each cluster was composed by normally distributed points ($\sigma = 1$). The five dimensions defining the hypercube constitutes the five "informative" features. Fifteen other "redundant" features were generated as random linear combinations of the informative features and added to the dataset. Finally, 180 random variables without predictive power were added, for a total of 200. In this type of setting, the number of associations to be tested was 19,900= $(200 \times 199)/2$. Among them, 190 are "real" (the relationships between the variables belonging to the "informative" and "redundant"). Figure 4 a summarizes the results of the analysis. Panel A shows the association called significant (q-value cutoff set to 0.05) on a Hive plot [20] as a function of the number of samples. Each branch of the Hive represents a type of variable (informative: 5 variables; redundant: 15; random: 180). The blue lines identify true positives (associations between nonindependent variables correctly identified), while FPs (incorrectly identified associations between independent variables) are marked in red. This representation clearly shows that, as expected, the number of true positives increases with the number of samples. A more quantitative representation of the effect of the number of samples on the number of false negatives (FNs) (nonindependent associations incorrectly rejected) is shown in panel B. Again, an increase in the number of samples is beneficial because it reduces the number of FNs. Panel C shows the effect of n on the FDR, which is always approximately constant and very close to the theoretical value of 0.05.

On the basis of these results, we conclude that with a relatively low number of samples, MICtools is able to identify nonfunctional associations typical of cluster structures efficiently. It is interesting to note that the associations among the informative variables started to be recovered when at least 250 samples were considered, while the associations between informative/redundant and redundant/redundant variables were significant for a smaller number of samples (50). This apparently odd behavior is due to the different nature of the association among the variables. Binary associations among informative variables are indeed characterized by the presence of clusters, while redundant associations are constructed by linear combinations. In accordance to the results discussed for SD1, the statistical power of the procedure depends on the type of association; with a smaller number of samples, the results are biased toward less complex association patterns.

Identifying ecological niches: the Tara dataset

The Tara Oceans project is a large multinational effort for the study of plankton on a global scale [17]. Within the project, a large study of the microbiota in water samples from the oceans, characterized using metagenomic techniques, has been recently made available. To illustrate the added value of using MICtools to analyze such large datasets, we downloaded the annotated 16S $_{mi}$tags [21] OTU count table of 139 water samples from http://ocean-microbiome.embl.de/companion.html, together with the accompanying metadata on temperature and chemical composition [22]. MICtools was used to identify the existence of significant relationships between the environmental variables and the taxonomic composition of the microbiota. The genus relative abundances, the environment variables, and the sample metadata are available in Additional File 1, Tables A4, A5, and A6, respectively. By using a q-value cutoff of 0.01, we found significant associations between the relative abundances of 279 taxa with water temperature and of 287 taxa with oxygen (Fig. 5, panels B and C, respectively). To highlight the novel information provided by MICtools, Spearman rank correlation coefficients and their associated p values were also calculated as in [23] (the default for the `cor.test()` function in the R environment). By using the Spearman coefficient alone, we could identify a subset of the relations identified by MICtools, namely, 194 taxa were associated with temperature and 191 taxa were associated with oxygen concentration. Conversely, almost all relationships identified with Spearman correlation were also identified by MICtools. While the Spearman coefficient-based approach identified associations well described by monotonic functions (Fig. 5E and 5F), MICtools was able to highlight the presence of more complex relationships between the taxa and the environmental parameters. As an example, we found a sharp increase in the *Alcaligenaceae* genus at an oxygen concentration of 200 μmol kg^{-1} (Fig. 5D) and a slow increase in the *Sphingomonadaceae* genus as a function of temperature. In both cases, by highlighting the sam-

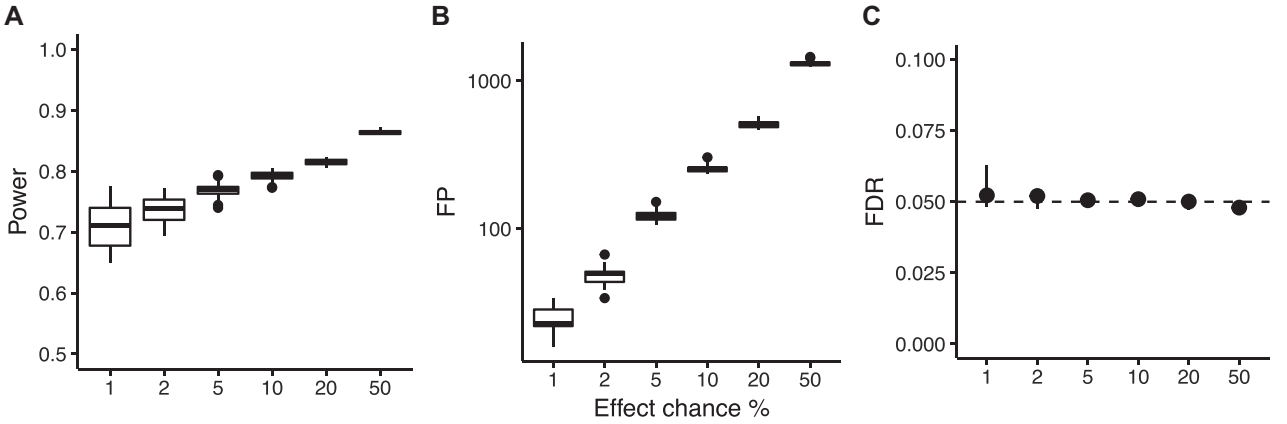

Figure 3: Analysis on SD2 dataset at the 0.05 significance level. A) Statistical power, B) number of FPs, and C) FDR for increasing effect chance. Each range represents 20 replicated datasets.

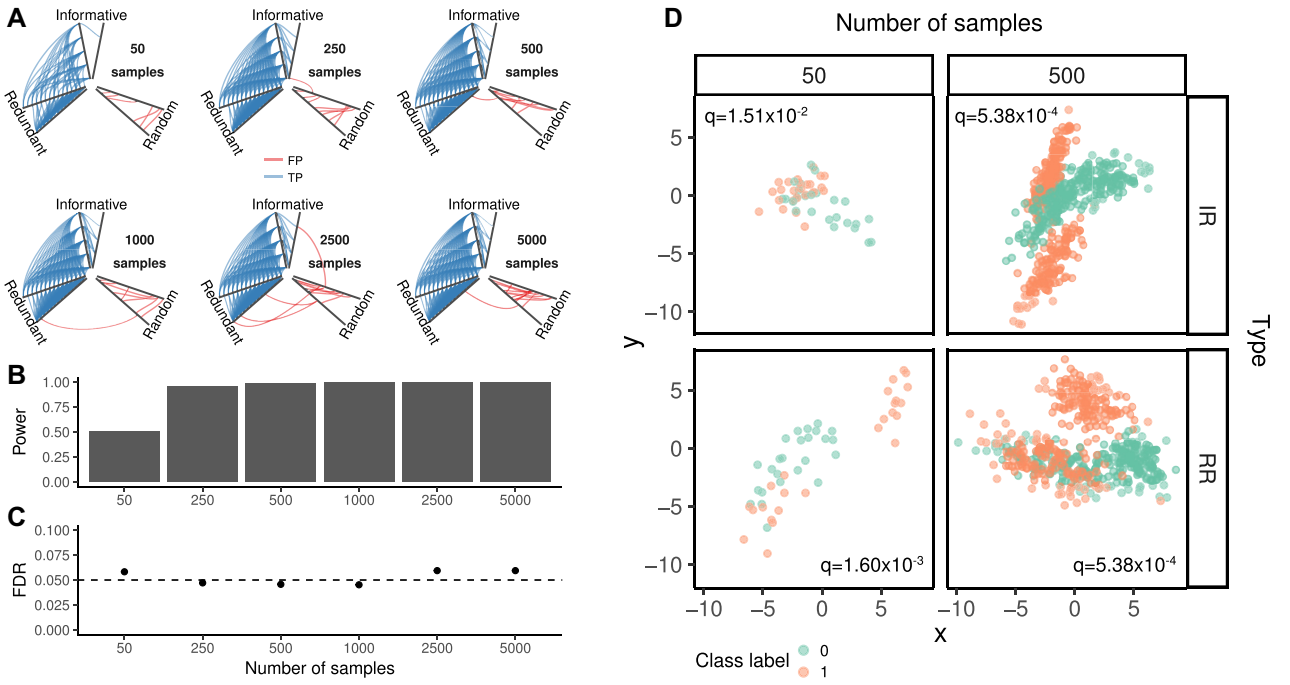

Figure 4: Madelon dataset. A) Hive plots of the detected association for increasing number of samples. The variables are grouped as "informative" (5), "redundant" (15), and "random" (180). True positives (associations between nonindependent variables passing the significance test) are in blue; false positives (associations between independent variable passing the significance test) are in red. B) Power and C) false discovery rate as a function of the number of samples. D) Example of significant relationships between informative and redundant (IR) and redundant (RR) variables within the Madelon datasets with 50 and 500 samples.

ples on the basis of their specific aquatic layer of reference, it is possible to see that the complex aggregation patterns identified by MICtools are associated with specific ecological niches. These results show the advantage of the use of the proposed approach as an automatic screening tool in the data exploration phase. The lists of the relationships identified by MICtools and by the Spearman coefficient-based procedure are available in Additional File 1, Tables A7 and A8, respectively.

Implementation details

MICtools is a Python-based open source software (licensed under GPLv3). MICtools requires the minepy [11] (https://minepy.readthedocs.io), Statsmodels [24], NumPy, SciPy, pandas, and Matplotlib scientific libraries. MICtools can handle different types of experiments:

- given a single dataset X with M variables and n samples, MICtools evaluates the $\frac{M\times(M-1)}{2}$ possible associations;
- given two datasets, X (of size $M \times n$) and Y (of size $K \times n$), MICtools evaluates all the pairwise relationships between the variables of the two datasets (for a total of $M \times K$ associations);
- given two datasets, X (of size $M \times n$) and Y (of size $K \times n$), MICtools evaluates all row-wise relationships, i.e., only the variable pairs x_i and y_i (for $i = 1, \ldots,$ min (M, K)) will be tested;

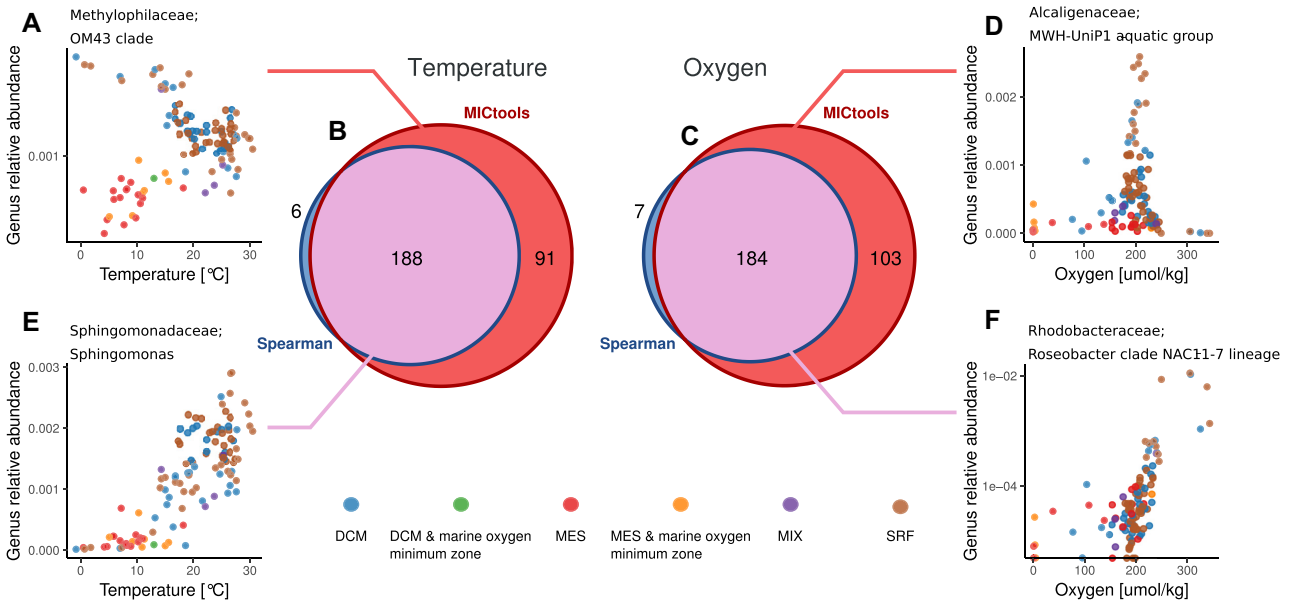

Figure 5: Tara dataset: Venn diagrams of the significant relationships between the genus-level relative abundances and two environmental variables, temperature (B) and oxygen (C), identified by MICtools and the Spearman coefficient-based procedure ($q < 0.01$). A, D) The relationships between the OM43 clade and temperature and between the MWH-UniP1 aquatic group are detected only by MICtools. E, F) Two monotonic relationships identified by both methods. Abbreviations: DCM, deep chlorophyll maximum layer; MES, mesopelagic zone; MIX, subsurface epipelagic mixed layer; SRF, surface water layer.

- for each experiment listed above, if the sample classes are provided, the analysis will be performed within each class, independently.

For multiple testing correction, MICtools makes available the strategies implemented in Statsmodels and a Python implementation of the Storey q-value method [7]. The indicative number of relationships tested per second during the empirical null estimation (using the TIC_e) and the strength estimation (MIC_e) for an increasing number of samples are listed in Additional File 2, Fig. A3.

MICtools source and the documentation are available at https://github.com/minepy/mictools. The Docker (https://www.docker.com/) image containing MICtools and the minepy library is available at https://hub.docker.com/r/minepy/mictools/ and installable with the command `docker pull minepy/mictools`.

Availability of source code and requirements

- Project name: MICtools
- Project home page: https://github.com/minepy/mictools
- Research Resource Identification Initiative ID (RRID), SciCrunch.org: SCR_016121
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: minepy, Statsmodels, NumPy, SciPy, pandas, Matplotlib
- License: GNU GPLv3

Availability of supporting data

The Tara dataset is available at http://ocean-microbiome.embl.de/companion.html. Code snapshots and additional tables are available in the GigaScience GigaDB repository [25].

Additional file

Additional File 1.xlsx, Additional File 2.pdf

Abbreviations

FDR, false discovery rate; FN, false negative; FP, false positive; FWER, family-wise error rate; MIC, maximal information coefficient; SD, synthetic dataset; TIC, total information coefficient.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Autonomous Province of Trento (Accordo di Programma).

Author Contributions

D.A., S.R., C.D., and P.F. conceived the manuscript. D.A. and P.F. developed the methodology. D.A. wrote the software. D.A., S.R., and C.D. analyzed the data. D.A., S.R., C.D., and P.F. wrote the manuscript.

References

1. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. Science 2011;**334**(6062):1518–24.
2. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. Proc Natl Acad Sci 2014;**111**(9):3354–9.
3. Murrell B, Murrell D, Murrell H. R2-equitability is satisfiable. Proc Natl Acad Sci 2014;**111**(21):E2160–E2160.

4. Reshef DN, Reshef YA, Mitzenmacher M, et al. Cleaning up the record on the maximal information coefficient and equitability. Proc Natl Acad Sci 2014;**111**(33):E3362–E3363.
5. Reshef YA, Reshef DN, Finucane HK, et al. Measuring dependence powerfully and equitably. J Mach Learn Res 2016;**17**(212):1–63.
6. Reshef YA, Reshef DN, Sabeti PC, et al. Equitability, interval estimation, and statistical power. arXiv preprint 2015 May.
7. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003 Aug;**100**(16):9440–5.
8. Franceschi P, Giordan M, Wehrens R. Multiple comparisons in mass-spectrometry-based -omics technologies. Trends Analyt Chem 2013;**50**:11–21.
9. Simon N, Tibshirani R. Comment on "detecting novel associations in large data sets" by Reshef Et Al, Science Dec 16, 2011. arXiv preprint arXiv:1401.7645 2014 Jan.
10. Reshef DN, Reshef YA, Sabeti PC, et al. An Empirical Study of Leading Measures of Dependence. arXiv preprint 2015 May.
11. Albanese D, Filosi M, Visintainer R, et al. Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics 2012;**29**(3):407–8.
12. Storey JD. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol 2002;**64**(3):479–98.
13. North BV, Curtis D, Sham PC. A note on the calculation of empirical P values from Monte Carlo procedures. Am J Hum Genet 2002;**71**(2):439–41.
14. Krzywinski M, Altman N. Points of significance: comparing samples—part II. Nat Methods 2014;**11**(4):355–6.
15. Guyon I, Elisseeff A. An introduction to feature extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh LA , editors. Feature Extraction. Studies in Fuzziness and Soft Computing, vol. 207, Berlin, Heidelberg: Springer; 2006.
16. Guyon I, Gunn S, Nikravesh M, et al. Feature Extraction: Foundations and Applications, Berlin, Heidelberg. Springer; 2008.
17. Bork P, Bowler C, de Vargas C, et al. Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. Science 2015;**348**(6237):873.
18. Reshef D, Reshef Y, Mitzenmacher M, et al. Equitability Analysis of the Maximal Information Coefficient, with Comparisons. arXiv preprint arXiv:13016314 2013.
19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learning Res 2011;**12**:2825–30.
20. Krzywinski M, Birol I, Jones SJM, et al. Hive plots–rational approach to visualizing networks. Brief Bioinform 2012;**13**(5):627–44.
21. Logares R, Sunagawa S, Salazar G, et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol 2014;**16**(9):2659–71.
22. Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science 2015;**348**(6237):1261359.
23. Best DJ, Roberts DE. Algorithm AS 89: the upper tail probabilities of Spearman's rho. Appl Stat 1975;**24**(3):377.
24. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python; 2010.
25. Albanese, D; Riccadonna, S; Donati, C; Franceschi, P et al. Supporting data for "A practical tool for maximal information coefficient analysis" GigaScience Database. http://dx.doi.org/10.5524/100427; 2018.