Detecting Novel Associations in Large Data Sets

Article *in* Science · December 2011

DOI: 10.1126/science.1205438 · Source: PubMed

9 authors, including:

Yakir Reshef
Harvard University
28 PUBLICATIONS **3,649** CITATIONS
SEE PROFILE

Peter J Turnbaugh
University of California, San Francisco
164 PUBLICATIONS **86,604** CITATIONS
SEE PROFILE

Pardis C Sabeti
Harvard University
408 PUBLICATIONS **31,713** CITATIONS
SEE PROFILE

Science
AAAS

Detecting Novel Associations in Large Data Sets
David N. Reshef, *et al.*
Science **334**, 1518 (2011);
DOI: 10.1126/science.1205438

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3]*† Yakir A. Reshef,[2,4]*† Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10]‡ Pardis C. Sabeti[2,6]‡

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

Imagine a data set with hundreds of variables, which may contain important, undiscovered relationships. There are tens of thousands of variable pairs—far too many to examine manually. If you do not already know what kinds of relationships to search for, how do you efficiently identify the important ones? Data sets of this size are increasingly common in fields as varied as genomics, physics, political science, and economics, making this question an important and growing challenge (*1*, *2*).

One way to begin exploring a large data set is to search for pairs of variables that are closely associated. To do this, we could calculate some measure of dependence for each pair, rank the pairs by their scores, and examine the top-scoring pairs. For this strategy to work, the statistic we use to measure dependence should have two heuristic properties: generality and equitability.

By generality, we mean that with sufficient sample size the statistic should capture a wide range of interesting associations, not limited to specific function types (such as linear, exponential, or periodic), or even to all functional relationships (*3*). The latter condition is desirable because not only do relationships take many functional forms, but many important relationships—for example, a superposition of functions—are not well modeled by a function (*4*–*7*).

By equitability, we mean that the statistic should give similar scores to equally noisy relationships of different types. For example, we do not want noisy linear relationships to drive strong sinusoidal relationships from the top of the list. Equitability is difficult to formalize for associations in general but has a clear interpretation in the basic case of functional relationships: An equitable statistic should give similar scores to functional relationships with similar R^2 values (given sufficient sample size).

Here, we describe an exploratory data analysis tool, the maximal information coefficient (MIC), that satisfies these two heuristic properties. We establish MIC's generality through proofs, show its equitability on functional relationships through simulations, and observe that this translates into intuitively equitable behavior on more general associations. Furthermore, we illustrate that MIC gives rise to a larger family of statistics, which we refer to as MINE, or maximal information-based nonparametric exploration. MINE statistics can be used not only to identify interesting associations, but also to characterize them according to properties such as nonlinearity and monotonicity. We demonstrate the application of MIC and MINE to data sets in health, baseball, genomics, and the human microbiota.

The maximal information coefficient. Intuitively, MIC is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship. Thus, to calculate the MIC of a set of two-variable data, we explore all grids up to a maximal grid resolution, dependent on the sample size (Fig. 1A), computing for every pair of integers (x,y) the largest possible mutual information achievable by any x-by-y grid applied to the data. We then normalize these mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. We define the characteristic matrix $M = (m_{x,y})$, where $m_{x,y}$ is the highest normalized mutual information achieved by any x-by-y grid, and the statistic MIC to be the maximum value in M (Fig. 1, B and C).

More formally, for a grid G, let I_G denote the mutual information of the probability dis-

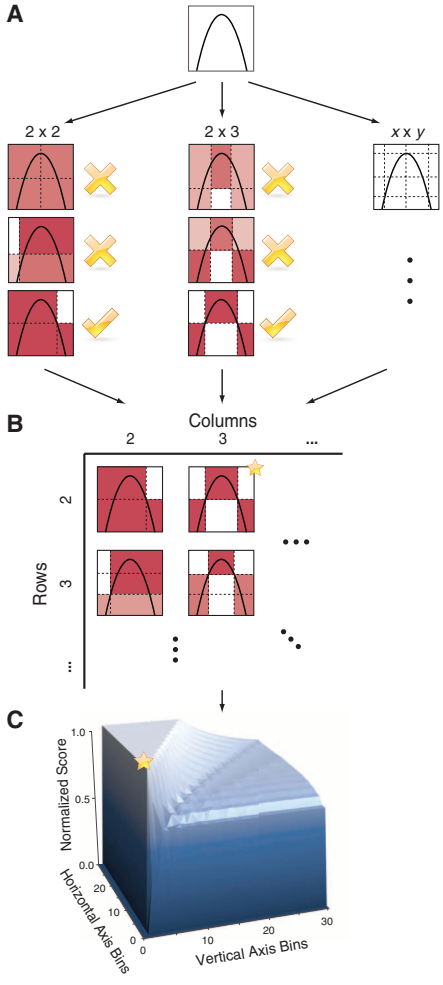


Fig. 1. Computing MIC (**A**) For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest induced mutual information. (**B**) The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score. (**C**) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid's corresponding location on the surface.

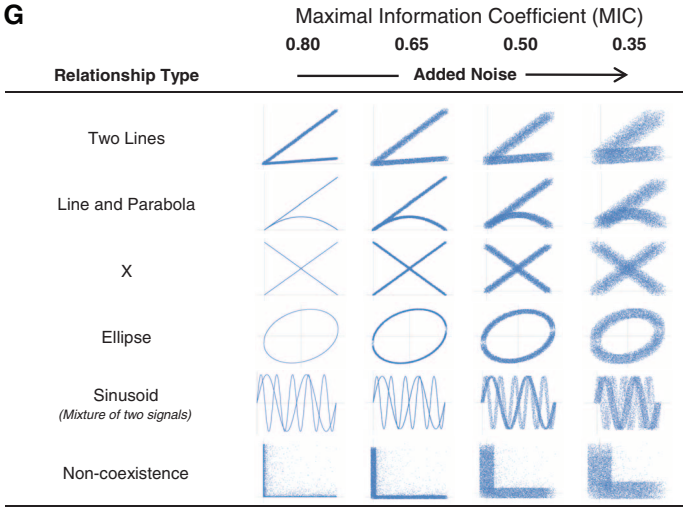
[1]Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [4]Department of Mathematics, Harvard College, Cambridge, MA 02138, USA. [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6]Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [7]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [8]Department of Biology, MIT, Cambridge, MA 02139, USA. [9]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [10]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: dnreshef@mit.edu (D.N.R.); yreshef@post.harvard.edu (Y.A.R.)
‡These authors contributed equally to this work.

tribution induced on the boxes of G, where the probability of a box is proportional to the number of data points falling inside the box. The (x,y)-th entry $m_{x,y}$ of the characteristic matrix equals $\max\{I_G\}/\log \min\{x,y\}$, where the maximum is taken over all x-by-y grids G. MIC is the maximum of $m_{x,y}$ over ordered pairs (x,y) such that $xy < B$, where B is a function of sample size; we usually set $B = n^{0.6}$ (see SOM Section 2.2.1).

A

Relationship Type	MIC	Pearson	Spearman	Mutual Information (KDE)	Mutual Information (Kraskov)	CorGC (Principal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Fourier frequency)	1.00	-0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.31	0.69	0.45	0.49	0.91
Parabolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Fourier frequency)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varying frequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76



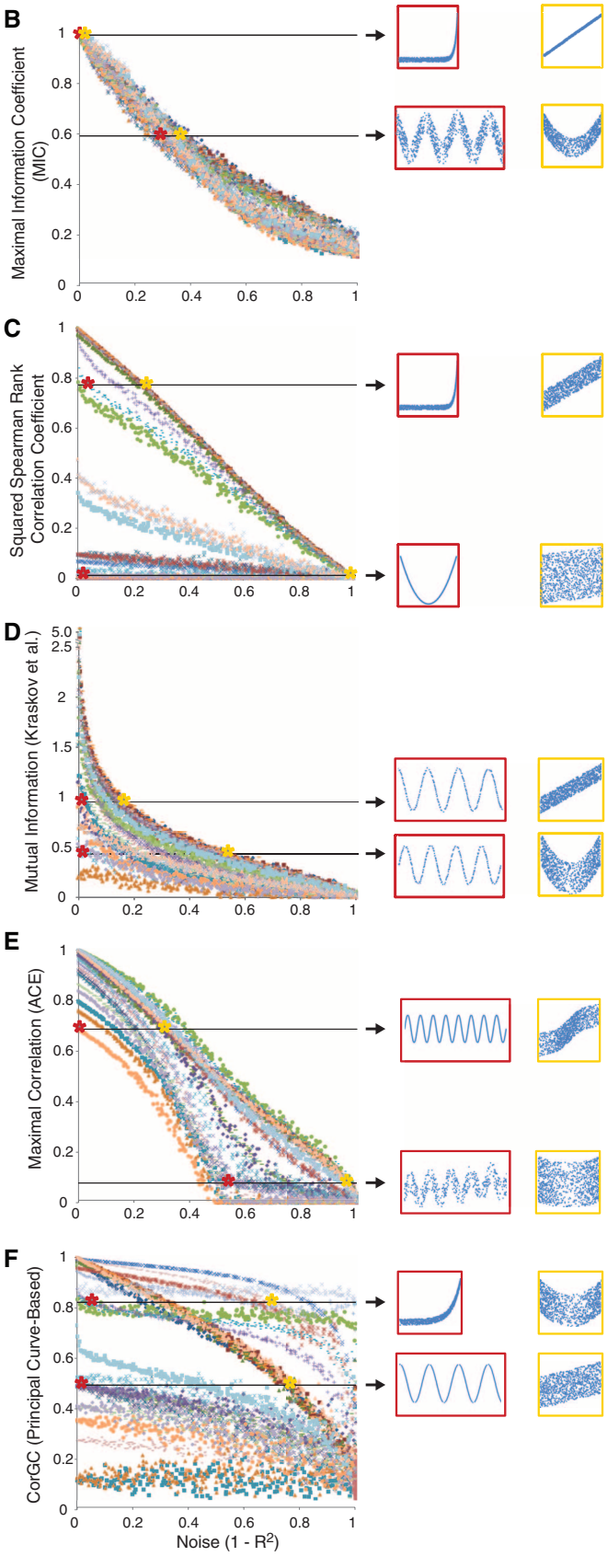


Fig. 2. Comparison of MIC to existing methods (**A**) Scores given to various noiseless functional relationships by several different statistics (*8*, *12*, *14*, *19*). Maximal scores in each column are accentuated. (**B** to **F**) The MIC, Spearman correlation coefficient, mutual information (Kraskov *et al*. estimator), maximal correlation (estimated using ACE), and the principal curve-based CorGC dependence measure, respectively, of 27 different functional relationships with independent uniform vertical noise added, as the R^2 value of the data relative to the noiseless function varies. Each shape and color corresponds to a different combination of function type and sample size. In each plot, pairs of thumbnails show relationships that received identical scores; for data exploration, we would like these pairs to have similar noise levels. For a list of the functions and sample sizes in these graphs as well as versions with other statistics, sample sizes, and noise models, see figs. S3 and S4. (**G**) Performance of MIC on associations not well modeled by a function, as noise level varies. For the performance of other statistics, see figs. S5 and S6.

Every entry of M falls between 0 and 1, and so MIC does as well. MIC is also symmetric [i.e., MIC(X, Y) = MIC(Y, X)] due to the symmetry of mutual information, and because I_G depends only on the rank order of the data, MIC is invariant under order-preserving transformations of the axes. Notably, although mutual information is used to quantify the performance of each grid, MIC is not an estimate of mutual information (SOM Section 2).

To calculate M, we would ideally optimize over all possible grids. For computational efficiency, we instead use a dynamic programming algorithm that optimizes over a subset of the possible grids and appears to approximate well the true value of MIC in practice (SOM Section 3).

Main properties of MIC. We have proven mathematically that MIC is general in the sense described above. Our proofs show that, with probability approaching 1 as sample size grows, (i) MIC assigns scores that tend to 1 to all never-constant noiseless functional relationships; (ii) MIC assigns scores that tend to 1 for a larger class of noiseless relationships (including superpositions of noiseless functional relationships); and (iii) MIC assigns scores that tend to 0 to statistically independent variables.

Specifically, we have proven that for a pair of random variables X and Y, (i) if Y is a function of X that is not constant on any open interval, then data drawn from (X,Y) will receive an MIC tending to 1 with probability one as sample size grows; (ii) if the support of (X,Y) is described by a finite union of differentiable curves of the form $c(t) = [x(t),y(t)]$ for t in [0,1], then data drawn from (X,Y) will receive an MIC tending to 1 with probability one as sample size grows, provided that dx/dt and dy/dt are each zero on finitely many points; (iii) the MIC of data drawn from (X,Y) converges to zero in probability as sample size grows if and only if X and Y are statistically independent. We have also proven that the MIC of a noisy functional relationship is bounded from below by a function of its R^2. (For proofs, see SOM.)

We tested MIC's equitability through simulations. These simulations confirm the mathematical result that noiseless functional relationships (i.e., $R^2 = 1.0$) receive MIC scores approaching 1.0 (Fig. 2A). They also show that, for a large collection of test functions with varied sample sizes, noise levels, and noise models, MIC roughly equals the coefficient of determination R^2 relative to each respective noiseless function. This makes it easy to interpret and compare scores across various function types (Fig. 2B and fig. S4). For instance, at reasonable sample sizes, a sinusoidal relationship with a noise level of $R^2 = 0.80$ and a linear relationship with the same R^2 value receive nearly the same MIC score. For a wide range of associations that are not well modeled by a function, we also show that MIC scores degrade in an intuitive manner as noise is added (Fig. 2G and figs. S5 and S6).

Comparisons to other methods. We compared MIC to a wide range of methods—including methods formulated around the axiomatic framework for measures of dependence developed by Rényi (*8*), other state-of-the-art measures of dependence, and several nonparametric curve estimation techniques that can be used to score pairs of variables based on how well they fit the estimated curve.

Methods such as splines (*1*) and regression estimators (*1*, *9*, *10*) tend to be equitable across functional relationships (*11*) but are not general: They fail to find many simple and important types of relationships that are not functional. (Figures S5 and S6 depict examples of relationships of this type from existing literature, and compare these methods to MIC on such relation-

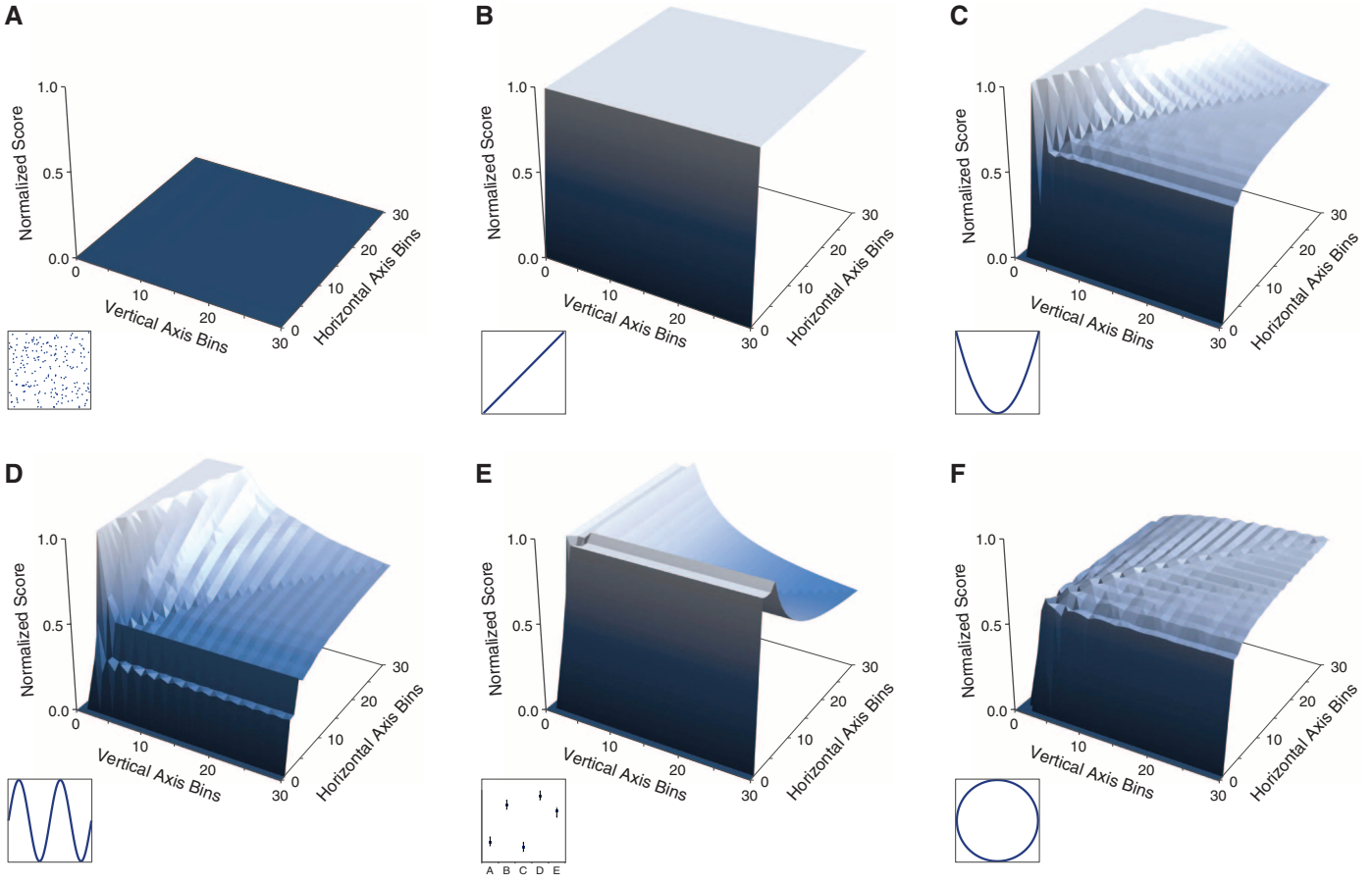


Fig. 3. Visualizations of the characteristic matrices of common relationships. (**A** to **F**) Surfaces representing the characteristic matrices of several common relationship types. For each surface, the x axis represents number of vertical axis bins (rows), the y axis represents number of horizontal axis bins (columns), and the z axis represents the normalized score of the best-performing grid with those dimensions. The inset plots show the relationships used to generate each surface. For surfaces of additional relationships, see fig. S7.

ships.) Although these methods are not intended to provide generality, the failure to assign high scores in such cases makes them unsuitable for identifying all potentially interesting relationships in a data set.

Other methods such as mutual information estimators (*12*–*14*), maximal correlation (*8*, *15*), principal curve–based methods (*16*–*19*, *20*), distance correlation (*21*), and the Spearman rank correlation coefficient all detect broader classes of relationships. However, they are not equitable even in the basic case of functional relationships: They show a strong preference for some types of functions, even at identical noise levels (Fig. 2, A and C to F). For example, at a sample size of 250, the Kraskov *et al.* mutual information estimator (*14*) assigns a score of 3.65 to a noiseless line but only 0.59 to a noiseless sinusoid, and it gives equivalent scores to a very noisy line ($R^2 = 0.35$) and to a much cleaner sinusoid ($R^2 = 0.80$) (Fig. 2D). Again, these

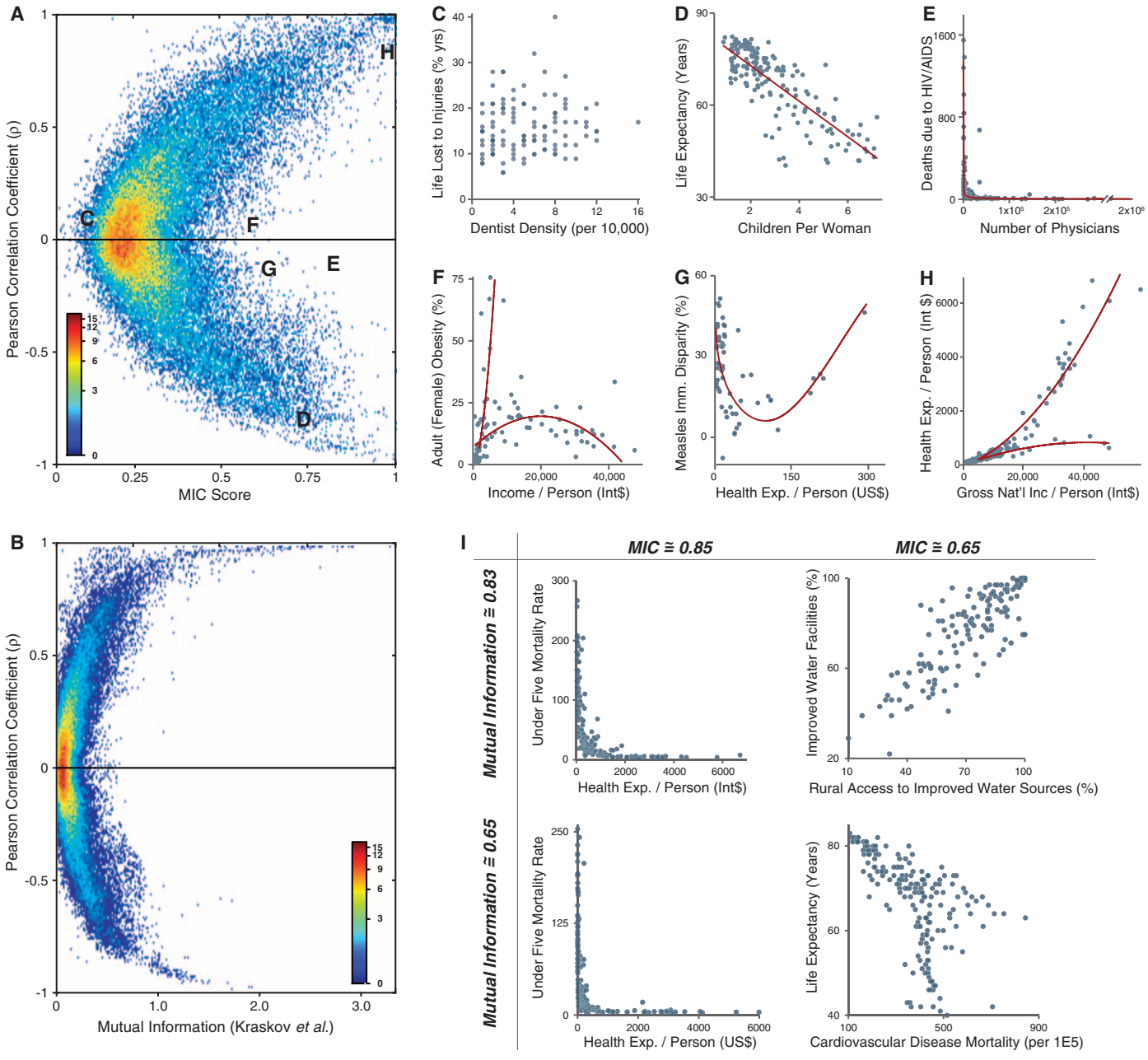


Fig. 4. Application of MINE to global indicators from the WHO. (**A**) MIC versus ρ for all pairwise relationships in the WHO data set. (**B**) Mutual information (Kraskov *et al.* estimator) versus ρ for the same relationships. High mutual information scores tend to be assigned only to relationships with high ρ, whereas MIC gives high scores also to relationships that are nonlinear. (**C** to **H**) Example relationships from (A). (C) Both ρ and MIC yield low scores for unassociated variables. (D) Ordinary linear relationships score high under both tests. (E to G) Relationships detected by MIC but not by ρ, because the relationships are nonlinear (E and G) or because more than one relationship is present (F). In (F), the linear trendline comprises a set of Pacific island nations in which obesity is culturally valued (*33*); most other countries follow a parabolic trend (table S10). (H) A superposition of two relationships that scores high under all three tests, presumably because the majority of points obey one relationship. The less steep minority trend consists of countries whose economies rely largely on oil (*37*) (table S11). The lines of best fit in (D) to (H) were generated using regression on each trend. (**I**) Of these four relationships, the left two appear less noisy than the right two. MIC accordingly assigns higher scores to the two relationships on the left. In contrast, mutual information assigns similar scores to the top two relationships and similar scores to the bottom two relationships.

results are not surprising—they correctly reflect the properties of mutual information. But this behavior makes these methods less practical for data exploration.

An expanded toolkit for exploration. The basic approach of MIC can be extended to define a broader class of MINE statistics based on both MIC and the characteristic matrix *M*. These statistics can be used to rapidly characterize relationships that may then be studied with more specialized or computationally intensive techniques.

Some statistics are derived, like MIC, from the spectrum of grid resolutions contained in *M*. Different relationship types give rise to different types of characteristic matrices (Fig. 3). For example, just as a characteristic matrix with a high maximum indicates a strong relationship, a symmetric characteristic matrix indicates a monotonic relationship. We can thus detect deviation from monotonicity with the maximum asymmetry score (MAS), defined as the maximum over *M* of $|m_{x,y} - m_{y,x}|$. MAS is useful, for example, for detecting periodic relationships with unknown frequencies that vary over time, a common occurrence in real data (*22*). MIC and MAS together detect such relationships more effectively than either Fisher's test (*23*) or a recent specialized test developed by Ahdesmäki *et al.* (figs. S8 and S9) (*24*).

Because MIC is general and roughly equal to R^2 on functional relationships, we can also define a natural measure of nonlinearity by MIC – ρ^2, where ρ denotes the Pearson product-moment correlation coefficient, a measure of linear dependence. The statistic MIC – ρ^2 is near 0 for linear relationships and large for nonlinear relationships with high values of MIC. As seen in the real-world examples below, it is useful for uncovering novel nonlinear relationships.

Similar MINE statistics can be defined to detect properties that we refer to as "complexity" and "closeness to being a function." We provide formal definitions and a performance summary of these two statistics (SOM section 2.3 and table S1). Finally, MINE statistics can also be used in cluster analysis to observe the higher-order structure of data sets (SOM section 4.9).

Application of MINE to real data sets. We used MINE to explore four high-dimensional data sets from diverse fields. Three data sets have previously been analyzed and contain many well-understood relationships. These data sets are (i) social, economic, health, and political indicators from the World Health Organization (WHO) and its partners (*7*, *25*); (ii) yeast gene expression profiles from a classic paper reporting genes whose transcript levels vary periodically with the cell cycle (*26*); and (iii) performance statistics from the 2008 Major League Baseball (MLB) season (*27*, *28*). For our fourth analysis, we applied MINE to a data set that has not yet been exhaustively analyzed: a set of bacterial abundance levels in the human gut microbiota (*29*). All relationships discussed in this section are significant at a false discovery rate of 5%; *p*-values and *q*-values are listed in the SOM.

We explored the WHO data set (357 variables, 63,546 variable pairs) with MIC, the commonly used Pearson correlation coefficient (ρ), and Kraskov's mutual information estimator (Fig. 4 and table S9). All three statistics detected many linear relationships. However, mutual information gave low ranks to many nonlinear relationships that were highly ranked by MIC (Fig. 4, A and B). Two-thirds of the top 150 relationships found by mutual information were strongly linear ($|\rho| \geq 0.97$), whereas most of the top 150 relationships found by MIC had $|\rho|$ below this threshold. Further, although equitability is difficult to assess for general associations, the results on some specific relationships suggest that MIC comes closer than mutual information to this goal (Fig. 4I). Using the nonlinearity measure MIC – ρ^2, we found several interesting relationships (Fig. 4, E to G), many of which are confirmed by existing literature (*30*–*32*). For example, we identified a superposition of two functional associations between female obesity and income per person—one from the Pacific Islands, where female obesity is a sign of status (*33*), and one from the rest of the world, where weight and status do not appear to be linked in this way (Fig. 4F).

We next explored a yeast gene expression data set (6223 genes) that was previously analyzed with a special-purpose statistic developed by Spellman *et al.* to identify genes whose transcript levels oscillate during the cell cycle (*26*). Of the genes identified by Spellman *et al.* and MIC, 70 and 69%, respectively, were also identified in a later study with more time points conducted by Tu *et al.* (*22*). However, MIC identified genes at a wider range of frequencies than did Spellman *et al.*, and MAS sorted those

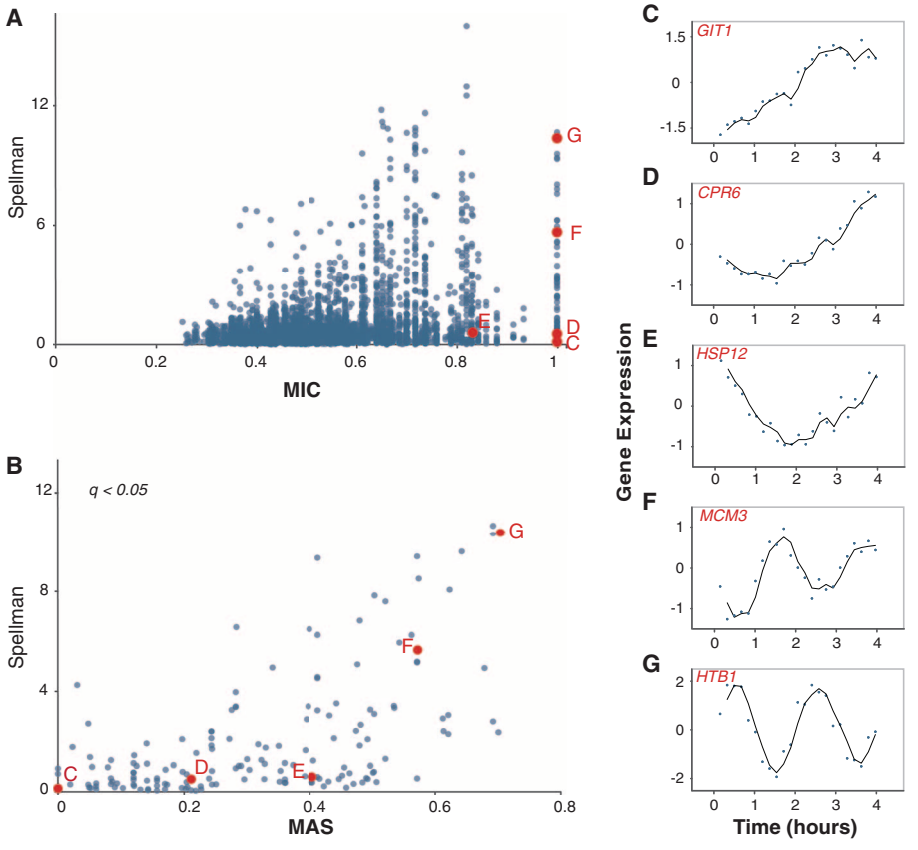


Fig. 5. Application of MINE to *Saccharomyces cerevisiae* gene expression data. (**A**) MIC versus scores obtained by Spellman *et al.* for all genes considered (*26*). Genes with high Spellman scores tend to receive high MIC scores, but some genes undetected by Spellman's analysis also received high MICs. (**B**) MAS versus Spellman's statistic for genes with significant MICs. Genes with a high Spellman score also tend to have a high MAS score. (**C** to **G**) Examples of genes with high MIC and varying MAS (trend lines are moving averages). MAS sorts the MIC-identified genes by frequency. A higher MAS signifies a shorter wavelength for periodic data, indicating that the genes found by Spellman *et al.* are those with shorter wavelengths. None of the examples except for (F) and (G) were detected by Spellman's analysis. However, subsequent studies have shown that (C) to (E) are periodic genes with longer wavelengths (*22*, *24*). More plots of genes detected with MIC and MAS are given in fig. S11.

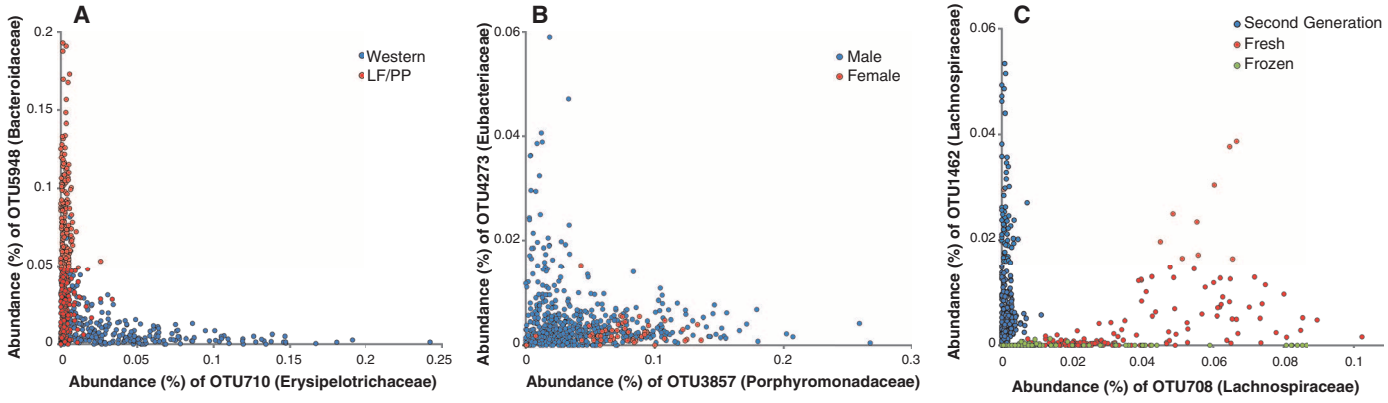
genes by frequency (Fig. 5). Of the genes identified by MINE as having high frequency (MAS > 75th percentile), 80% were identified by Spellman *et al*., while of the low-frequency genes (MAS < 25th percentile), Spellman *et al*. identified only 20% (Fig. 5B). For example, although both methods found the well-known cell-cycle regulator HTB1 (Fig. 5G) required for chromatin assembly, only MIC detected the heat-shock protein HSP12 (Fig. 5E), which Tu *et al*. confirmed to be in the top 4% of periodic genes in yeast. HSP12, along with 43% of the genes identified by MINE but not Spellman *et al*., was also in the top third of statistically significant periodic genes in yeast according to the more sophisticated specialty statistic of Ahdesmäki *et al*., which was specifically designed for finding periodic relationships without a prespecified frequency in biological systems (*24*). Because of MIC's generality and the small size of this data set (n = 24), relatively few of the genes analyzed (5%) had significant MIC scores after multiple testing correction at a false discovery rate of 5%. However, using a less conservative false discovery rate of 15% yielded a larger list of significant genes (16% of all genes analyzed), and this larger list still attained a 68% confirmation rate by Tu *et al*.

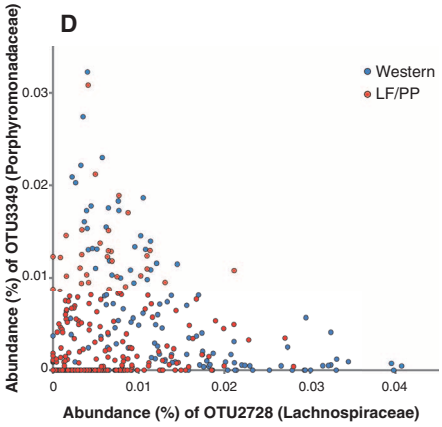
In the MLB data set (131 variables), MIC and ρ both identified many linear relationships, but interesting differences emerged. On the basis of ρ, the strongest three correlates with player salary are walks, intentional walks, and runs batted in. By contrast, the strongest three associations according to MIC are hits, total bases, and a popular aggregate offensive statistic called Replacement Level Marginal Lineup Value (*27*, *34*) (fig. S12 and table S12). We leave it to baseball enthusiasts to decide which of these statistics are (or should be!) more strongly tied to salary.

Our analysis of gut microbiota focused on the relationships between prevalence levels of the trillions of bacterial species that colonize the gut of humans and other mammals (*35*, *36*). The data set consisted of large-scale sequencing of 16*S* ribosomal RNA from the distal gut microbiota of mice colonized with a human fecal sample (*29*). After successful colonization, a subset of the mice was shifted from a low-fat, plant-polysaccharide–rich (LF/PP) diet to a high-fat, high-sugar "Western" diet. Our initial analysis identified 9472 significant relationships (out of 22,414,860) between "species"-level groups called operational taxonomic units (OTUs); significantly more of these relationships occurred between OTUs in the same bacterial family than expected by chance (30% versus 24 ± 0.6%).

Examining the 1001 top-scoring nonlinear relationships (MIC-ρ^2 > 0.2), we observed that a common association type was "noncoexistence": When one species is abundant the other is less abundant than expected by chance, and vice versa (Fig. 6, A, B, and D). Additionally, we found that 312 of the top 500 nonlinear relationships were affected by one or more factors for which data were available (host diet, host sex, identity of human donor, collection method, and location in the gastrointestinal tract; SOM section 4.8). Many are noncoexistence relationships that are explained by diet (Fig. 6A and table S13). These diet-explained noncoexistence relationships occur at a range of taxonomic depths—interphylum, interfamily, and intrafamily—and form a highly interconnected network of nonlinear relationships (Fig. 6E).

The remaining 188 of the 500 highly ranked nonlinear relationships were not affected by any of the factors in the data set and included many noncoexistence relationships (table S14 and Fig. 6D). These unexplained noncoexistence relationships may suggest interspecies competition and/or additional selective factors that shape gut microbial ecology and





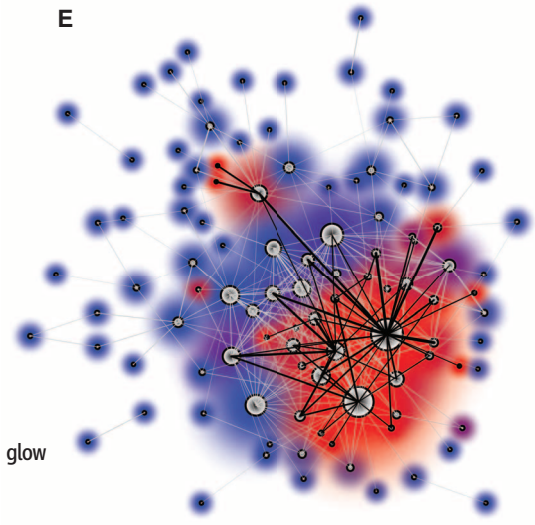


Fig. 6. Associations between bacterial species in the gut microbiota of "humanized" mice. (**A**) A noncoexistence relationship explained by diet: Under the LF/PP diet a *Bacteroidaceae* species-level OTU dominates, whereas under a Western diet an *Erysipelotrichaceae* species dominates. (**B**) A noncoexistence relationship occurring only in males. (**C**) A nonlinear relationship partially explained by donor. (**D**) A noncoexistence relationship not explained by diet. (**E**) A spring graph (see SOM Section 4.9) in which nodes correspond to OTUs and edges correspond to the top 300 nonlinear relationships. Node size is proportional to the number of these relationships involving the OTU, black edges represent relationships explained by diet, and node glow color is proportional to the fraction of adjacent edges that are black (100% is red, 0% is blue).

therefore represent promising directions for future study.

Conclusion. Given the ever-growing, technology-driven data stream in today's scientific world, there is an increasing need for tools to make sense of complex data sets in diverse fields. The ability to examine all potentially interesting relationships in a data set—independent of their form—allows tremendous versatility in the search for meaningful insights. On the basis of our tests, MINE is useful for identifying and characterizing structure in data.

References and Notes

1. T. Hastie, R. Tibshirani, J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Verlag, New York, 2009).
2. Science Staff, *Science* **331**, 692 (2011).
3. By "functional relationship" we mean a distribution (X,Y) in which Y is a function of X, potentially with independent noise added.
4. A. Caspi *et al*., *Science* **301**, 386 (2003).
5. R. N. Clayton, T. K. Mayeda, *Geochim. Cosmochim. Acta* **60**, 1999 (1996).
6. T. J. Algeo, T. W. Lyons, *Paleoceanography* **21**, PA1016 (2006).
7. World Health Organization Statistical Information Systems, *World Health Organization Statistical Information Systems (WHOSIS)* (2009); www.who.int/whosis/en/.
8. A. Rényi, *Acta Math. Hung.* **10**, 441 (1959).
9. C. J. Stone, *Ann. Stat.* **5**, 595 (1977).
10. W. S. Cleveland, S. J. Devlin, *J. Am. Stat. Assoc.* **83**, 596 (1988).
11. For both splines and regression estimators, we used R^2 with respect to the estimated spline/regression function to score relationships.
12. Y. I. Moon, B. Rajagopalan, U. Lall, *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **52**, 2318 (1995).
13. G. Darbellay, I. Vajda, *IEEE Trans. Inf. Theory* **45**, 1315 (1999).
14. A. Kraskov, H. Stögbauer, P. Grassberger, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 066138 (2004).
15. L. Breiman, J. H. Friedman, *J. Am. Stat. Assoc.* **80**, 580 (1985).
16. T. Hastie, W. Stuetzle, *J. Am. Stat. Assoc.* **84**, 502 (1989).
17. R. Tibshirani, *Stat. Comput.* **2**, 183 (1992).
18. B. Kégl, A. Krzyzak, T. Linder, K. Zeger, *Adv. Neural Inf. Process. Syst.* **11**, 501 (1999).
19. P. Delicado, M. Smrekar, *Stat. Comput.* **19**, 255 (2009).
20. "Principal curve-based methods" refers to mean-squared error relative to the principal curve, and CorGC, the principal curve-based measure of dependence of Delicado *et al*.
21. G. Székely, M. Rizzo, *Ann. Appl. Stat.* **3**, 1236 (2009).
22. B. P. Tu, A. Kudlicki, M. Rowicka, S. L. McKnight, *Science* **310**, 1152 (2005).
23. R. Fisher, Tests of significance in harmonic analysis. *Proc. R. Soc. Lond. A* **125**, 54 (1929).
24. M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, O. Yli-Harja, *BMC Bioinformatics* **6**, 117 (2005).
25. H. Rosling, Gapminder, *Indicators in Gapminder World* (2008); www.gapminder.org/data/
26. P. T. Spellman *et al*., *Mol. Biol. Cell* **9**, 3273 (1998).
27. Baseball Prospectus Statistics Reports (2009); www.baseballprospectus.com/sortable/
28. S. Lahman, The Baseball Archive, *The Baseball Archive* (2009); baseball1.com/statistics/
29. P. J. Turnbaugh *et al.*, The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).
30. L. Chen *et al*., *Lancet* **364**, 1984 (2004).
31. S. Desai, S. Alva, *Demography* **35**, 71 (1998).
32. S. Gupta, M. Verhoeven, *J. Policy Model.* **23**, 433 (2001).
33. T. Gill *et al*., Obesity in the Pacific: Too big to ignore. *Noumea, New Caledonia: World Health Organization Regional Office for the Western Pacific, Secretariat of the Pacific Community* (2002).
34. RPMLV estimates how many more runs per game a player contributes over a replacement-level player in an average lineup.
35. P. J. Turnbaugh *et al*., *Nature* **449**, 804 (2007).
36. R. E. Ley *et al*., *Science* **320**, 1647 (2008).
37. *The World Factbook 2009* (Central Intelligence Agency, Washington, DC, 2009).

Acknowledgments: We thank C. Blättler, B. Eidelson, M. D. Finucane, M. M. Finucane, M. Fujihara, T. Gingrich, E. Goldstein, R. Gupta, R. Hahne, T. Jaakkola, N. Laird, M. Lipsitch, S. Manber, G. Nicholls, A. Papageorge, N. Patterson, E. Phelan, J.Rinn, B. Ripley, I. Shylakhter, and R. Tibshirani for invaluable support and critical discussions throughout; and O. Derby, M. Fitzgerald, S. Hart, M. Huang, E. Karlsson, S. Schaffner, C. Edwards, and D. Yamins for assistance. P.C.S. and this work are supported by the Packard Foundation. For data set analysis, P.C.S. was also supported by NIH MIDAS award U54GM088558, D.N.R. by a Marshall Scholarship, M.M. by NSF grant 0915922, H.K.F. by ERC grant 239985, S.R.G. by the Medical Scientist Training Program, and P.J.T. by NIH P50 GM068763. Data and software are available online at http://exploredata.net.

Supporting Online Material

www.sciencemag.org/cgi/content/full/334/6062/1518/DC1
Materials and Methods
SOM Text
Figs. S1 to S13
Tables S1 to S14
References (*38*–*54*)

10 March 2011; accepted 5 October 2011
10.1126/science.1205438

The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution

Adam Ben-Shem,*† Nicolas Garreau de Loubresse,* Sergey Melnikov,* Lasse Jenner, Gulnara Yusupova, Marat Yusupov†

Ribosomes translate genetic information encoded by messenger RNA into proteins. Many aspects of translation and its regulation are specific to eukaryotes, whose ribosomes are much larger and intricate than their bacterial counterparts. We report the crystal structure of the 80*S* ribosome from the yeast *Saccharomyces cerevisiae*—including nearly all ribosomal RNA bases and protein side chains as well as an additional protein, Stm1—at a resolution of 3.0 angstroms. This atomic model reveals the architecture of eukaryote-specific elements and their interaction with the universally conserved core, and describes all eukaryote-specific bridges between the two ribosomal subunits. It forms the structural framework for the design and analysis of experiments that explore the eukaryotic translation apparatus and the evolutionary forces that shaped it.

Ribosomes are responsible for the synthesis of proteins across all kingdoms of life. The core, which is universally conserved and was described in detail by structures of prokaryotic ribosomes, catalyzes peptide bond formation and decodes mRNA (*1*). However, eukaryotes and prokaryotes differ markedly in other translation processes such as initiation, termination, and regulation (*2*, *3*), and eukaryotic ribosomes play a central role in many eukaryote-specific cellular processes. Accordingly, eukaryotic ribosomes are at least 40% larger than their bacterial counterparts as a result of additional ribosomal RNA (rRNA) elements called expansion segments (ESs) and extra protein moieties (*4*).

All ribosomes are composed of two subunits. The large 60*S* subunit of the eukaryotic ribosome (50*S* in bacteria) consists of three rRNA molecules (25*S*, 5.8*S*, and 5*S*) and 46 proteins, whereas the small 40*S* subunit (30*S* in bacteria) includes one rRNA chain (18*S*) and 33 proteins. Of the 79 proteins, 32 have no homologs in crystal structures of bacterial or archaeal ribosomes, and those that do have homologs can still harbor large eukaryote-specific extensions (*5*). Apart from variability in certain rRNA expansion segments, all eukaryotic ribosomes, from yeast to human, are very similar.

Three-dimensional cryoelectron microscopy (cryo-EM) reconstructions of eukaryotic ribosomes at 15 to 5.5 Å resolution provided insight into the interactions of the ribosome with several factors (*4*, *6*–*8*). A crystal structure of the *S. cerevisiae* ribosome at 4.15 Å resolution described the fold of all ordered rRNA expansion segments, but the relatively low resolution precluded localization of most eukaryote-specific proteins (*9*). Crystallographic data at a better resolution (3.9 Å) from the *Tetrahymena thermophila* 40*S* led to a definition of the locations and folds of all eukaryote-specific proteins in the

Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1 rue Laurent Fries, BP10142, Illkirch F-67400, France; INSERM, U964, Illkirch F-67400, France; CNRS, UMR7104, Illkirch F-67400, France; and Université de Strasbourg, Strasbourg F-67000, France.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: adam@igbmc.fr (A.B.-S.); marat@igbmc.fr (M.Y.)