

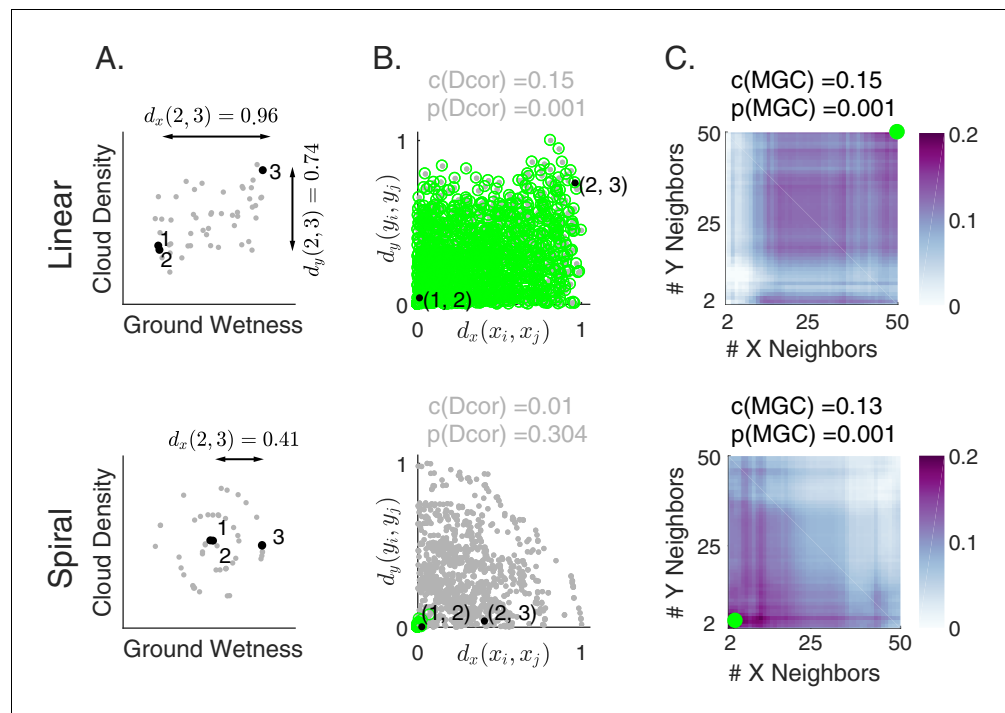


---

## Figures and figure supplements

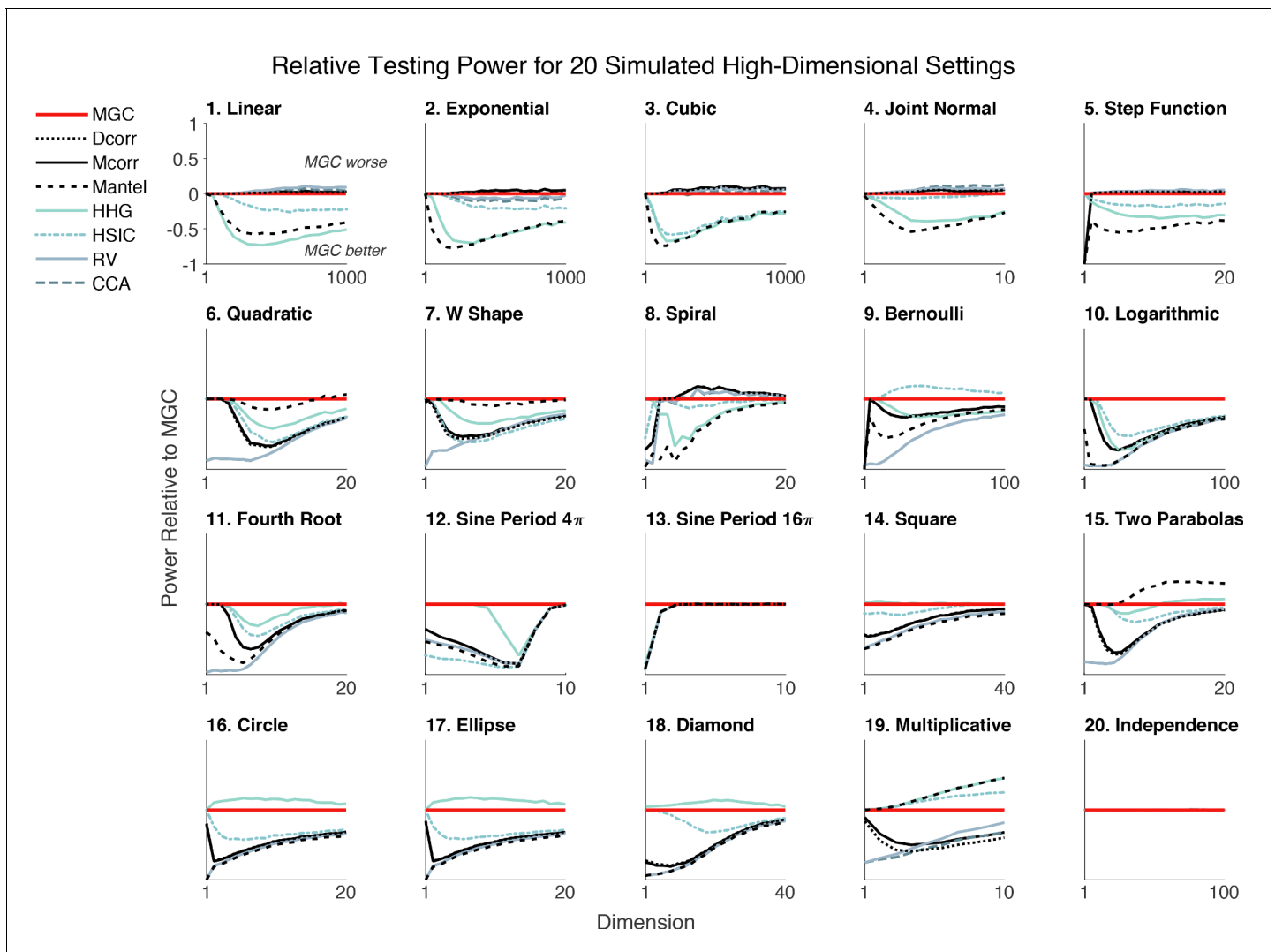
Discovering and deciphering relationships across disparate data modalities

**Joshua T Vogelstein *et al***



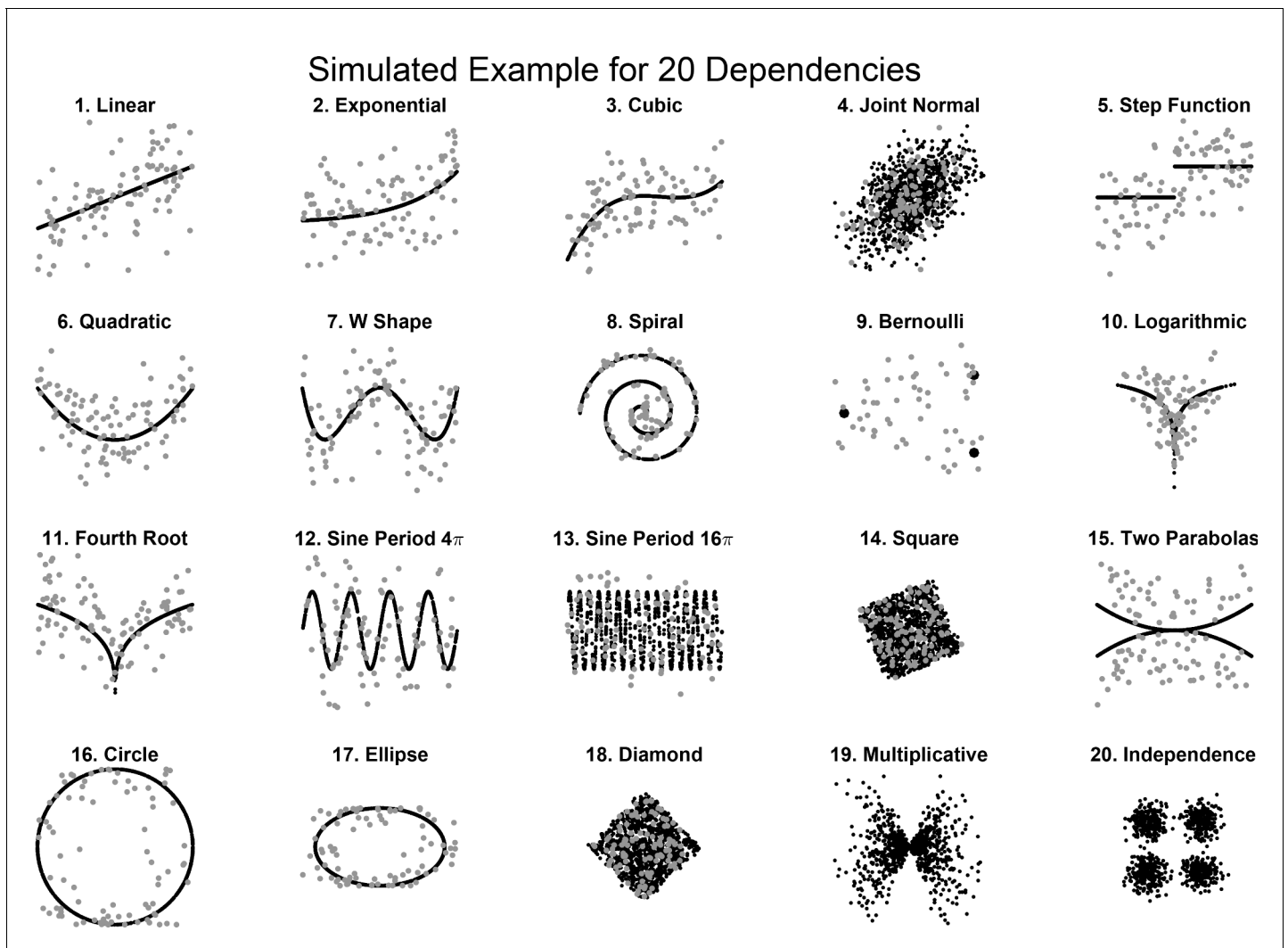
**Figure 1.** Illustration of Multiscale Graph Correlation (MGC) on simulated cloud density ( $x_i$ ) and grass wetness ( $y_i$ ). We present two different relationships: linear (top) and nonlinear spiral (bottom; see Materials and methods for simulation details). (A) Scatterplots of the raw data using 50 pairs of samples for each scenario. Samples 1, 2, and 3 (black) are highlighted; arrows show  $x$  distances between these pairs of points while their  $y$  distances are almost 0. (B) Scatterplots of all pairs of distances comparing  $x$  and  $y$  distances. Distances are linearly correlated in the linear relationship, whereas they are not in the spiral relationship. Dcorr uses all distances (gray dots) to compute its test statistic and p-value, whereas MGC chooses the local scale and then uses only the local distances (green dots). (C) Heatmaps characterizing the strength of the generalized correlation at all possible scales (ranging from 2 to  $n$  for both  $x$  and  $y$ ). For the linear relationship, the global scale is optimal, which is the scale that MGC selects and results in a p-value identical to Dcorr. For the nonlinear relationship, the optimal scale is local in both  $x$  and  $y$ , so MGC achieves a far larger test statistic, and a correspondingly smaller and significant p-value. Thus, MGC uniquely detects dependence and characterizes the geometry in both relationships.

DOI: <https://doi.org/10.7554/eLife.41690.003>



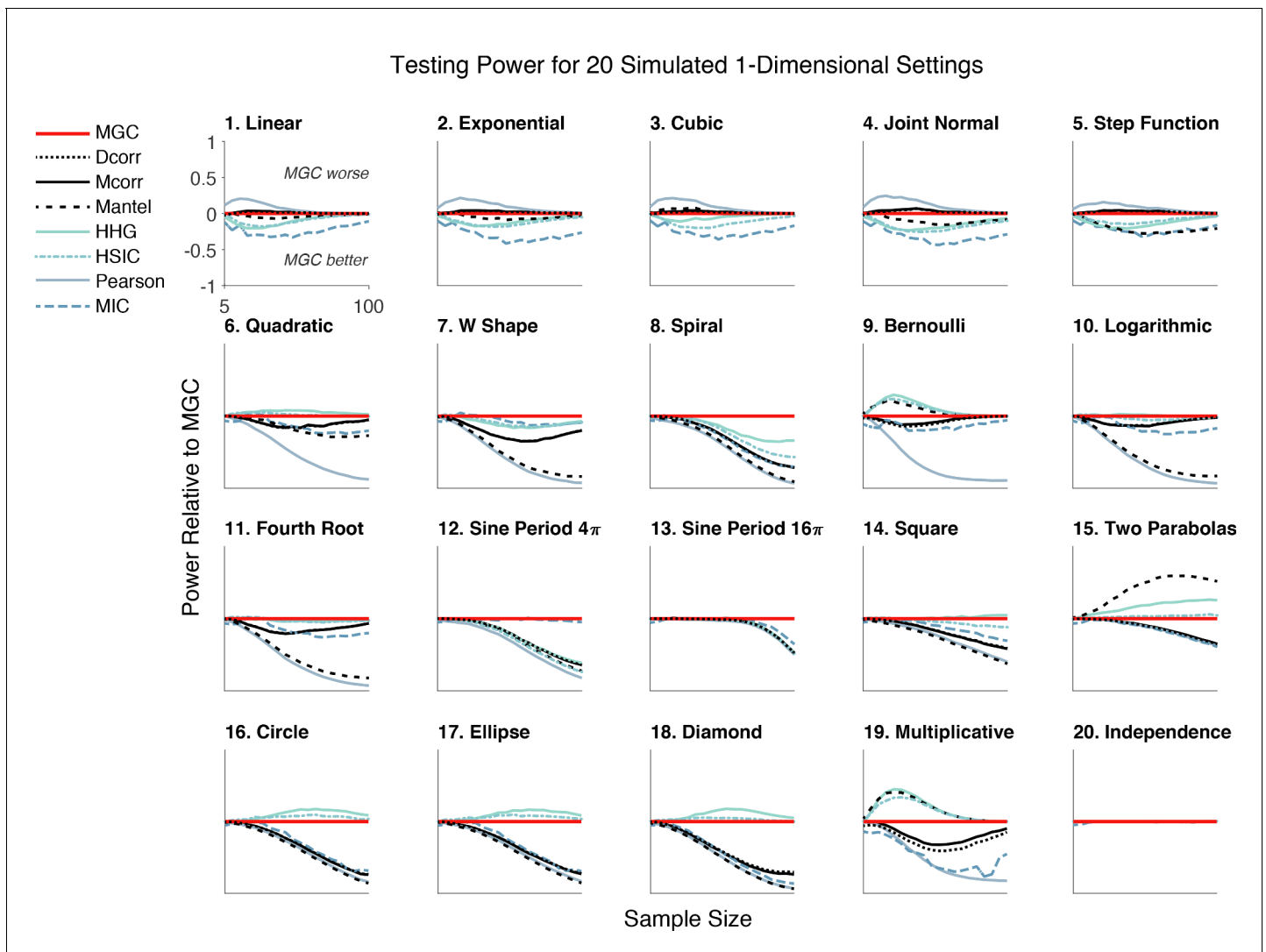
**Figure 2.** An extensive benchmark suite of 20 different relationships spanning polynomial, trigonometric, geometric, and other relationships demonstrates that MGC empirically nearly dominates eight other methods across dependencies and dimensionalities ranging from 1 to 1000 (see Materials and methods and **Figure 2—figure supplement 1** for details). Each panel shows the testing power of other methods relative to the power of MGC (e.g. power of MGC minus the power of MGC) at significance level  $\alpha = 0.05$  versus dimensionality for  $n = 100$ . Any line below zero at any point indicates that that method's power is less than MGC's power for the specified setting and dimensionality. MGC achieves empirically better (or similar) power than all other methods in almost all relationships and all dimensions. For the independent relationship (#20), all methods yield power 0.05 as they should. Note that MGC is always plotted 'on top' of the other methods, therefore, some lines are obscured.

DOI: <https://doi.org/10.7554/eLife.41690.004>



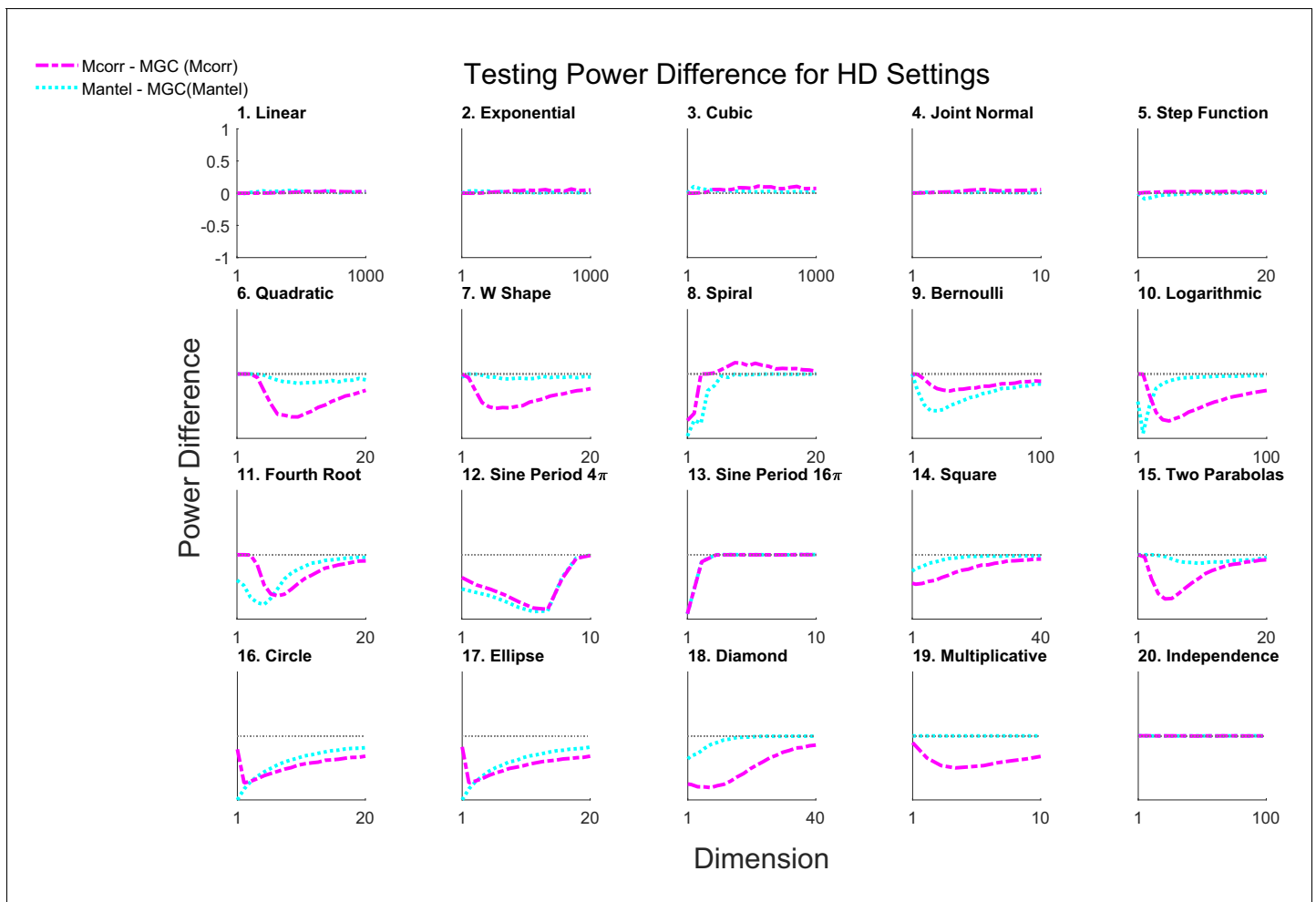
**Figure 2—figure supplement 1.** Visualization of the 20 dependencies at  $p = q = 1$ . For each,  $n = 100$  points are sampled with noise ( $\kappa = 1$ ) to show the actual sample data used for one-dimensional relationships (gray dots). For comparison purposes,  $n = 1000$  points are sampled without noise ( $\kappa = 0$ ) to highlight each underlying dependency (black dots). Note that only black points are plotted for type 19 and 20, as they do not have the noise parameter  $\kappa$ .

DOI: <https://doi.org/10.7554/eLife.41690.005>



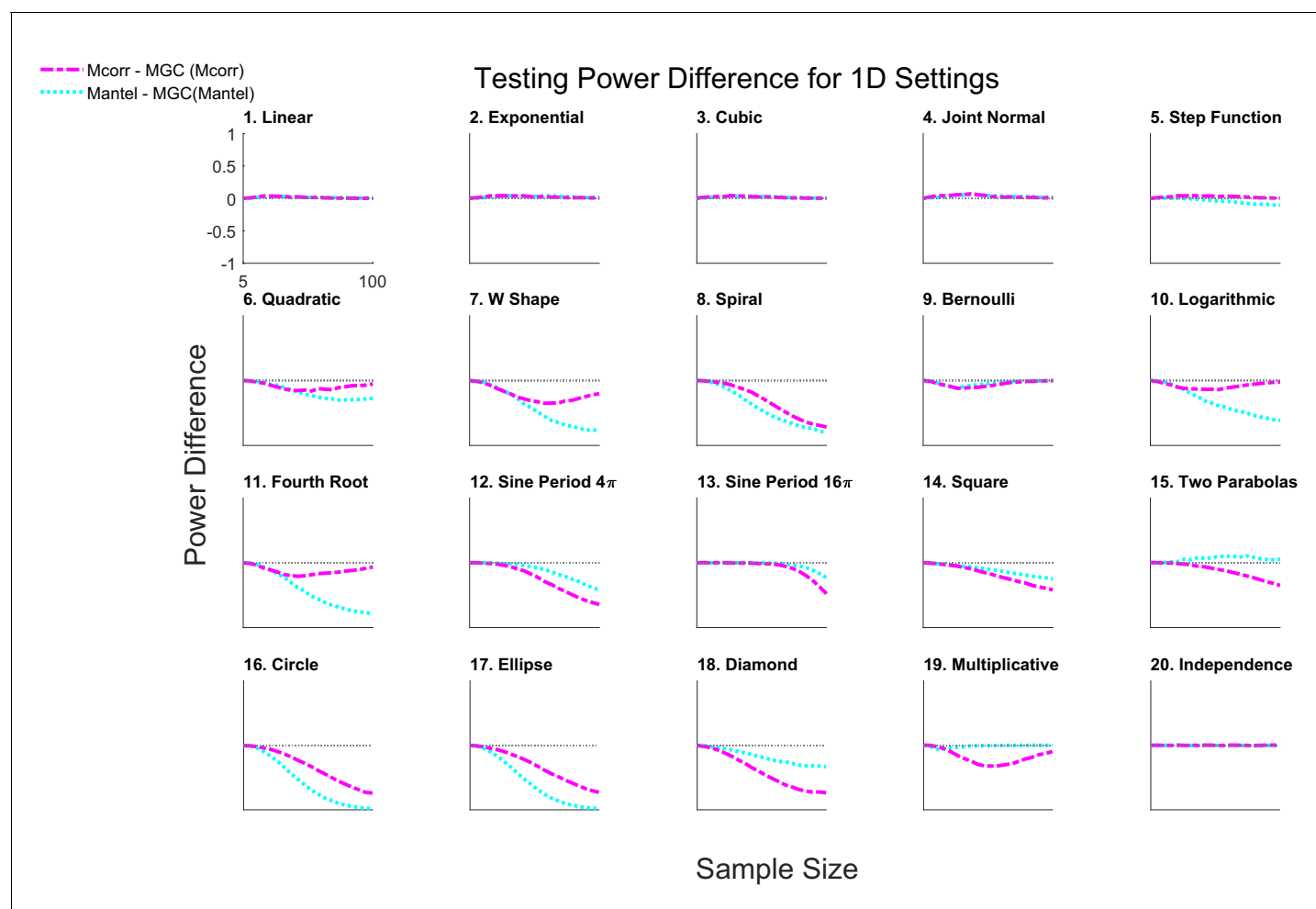
**Figure 2—figure supplement 2.** The same power plots as in **Figure 2**, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100. MGC empirically achieves similar or better power than the previous state-of-the-art approaches on most problems. Note that MIC is included in 1D case; RV and CCA both equal PEARSON in 1D; KENDALL and SPEARMAN are too similar to PEARSON in power and thus omitted in plotting.

DOI: <https://doi.org/10.7554/eLife.41690.006>



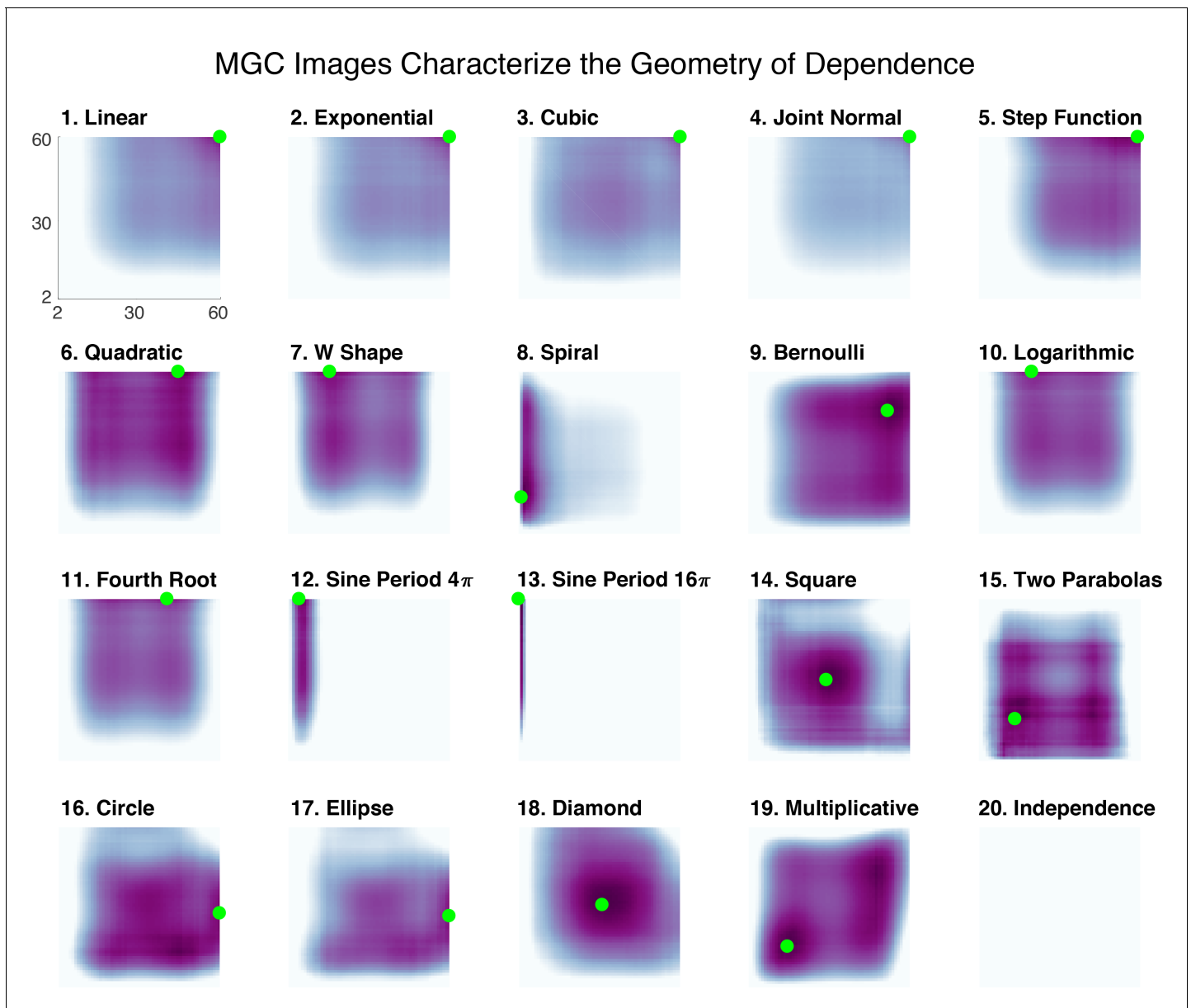
**Figure 2—figure supplement 3.** The same set-ups as in **Figure 2**, comparing different MGC implementations versus its global counterparts. The default MGC builds upon M<sub>CORR</sub> throughout the paper, and we further consider MGC on MANTEL to illustrate the generalization. The magenta line shows the power difference between M<sub>CORR</sub> and MGC, and the cyan line shows the power difference between MANTEL and the MGC version of MANTEL. Indeed, MGC is able to improve the global counterpart in testing power under nonlinear dependencies, and maintains similar power under linear and independent dependencies.

DOI: <https://doi.org/10.7554/eLife.41690.007>



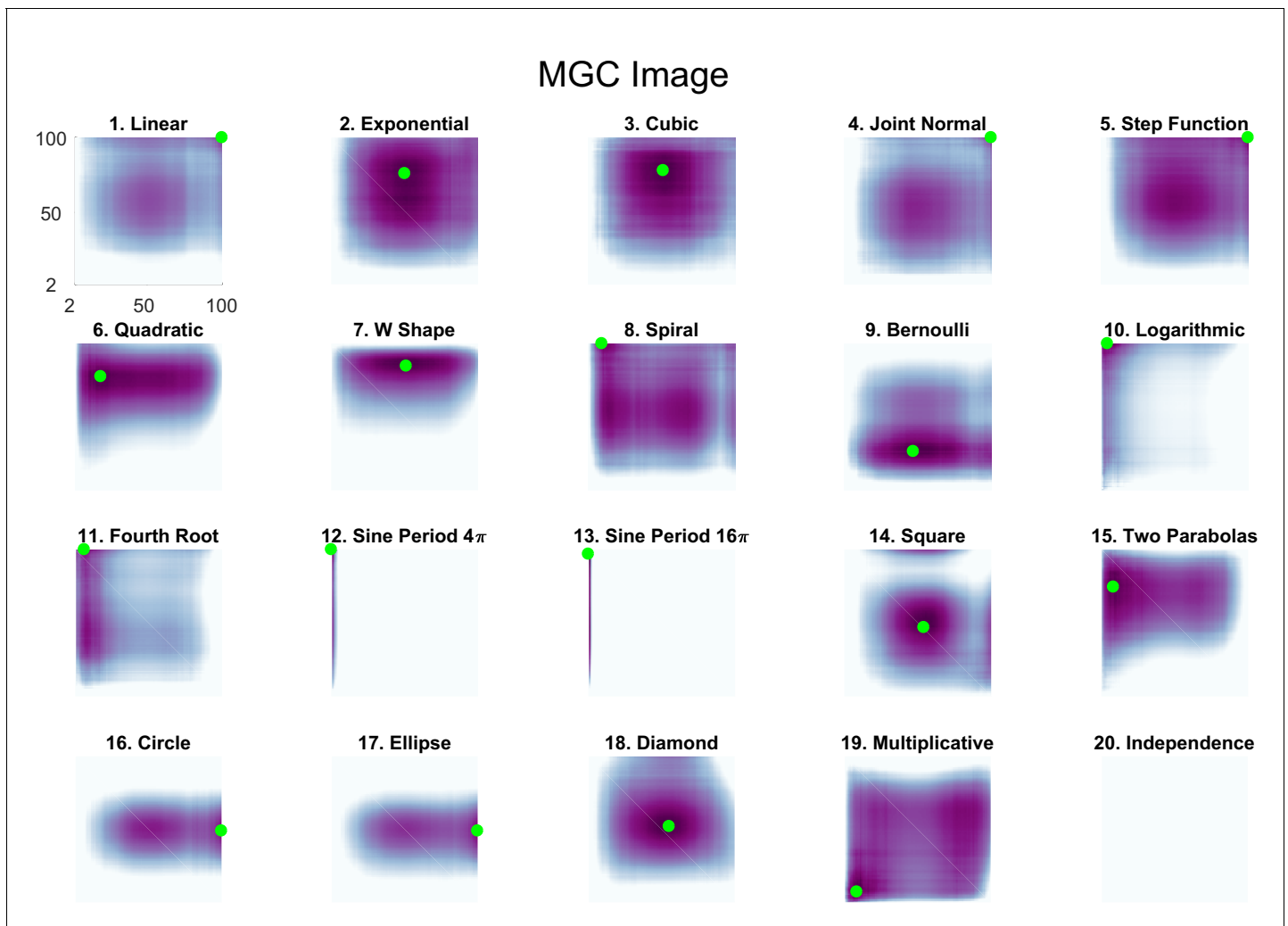
**Figure 2—figure supplement 4.** The same power plots as in **Figure 3**, except the 20 dependencies are one-dimensional with noise, and the x-axis shows sample size increasing from 5 to 100.

DOI: <https://doi.org/10.7554/eLife.41690.008>



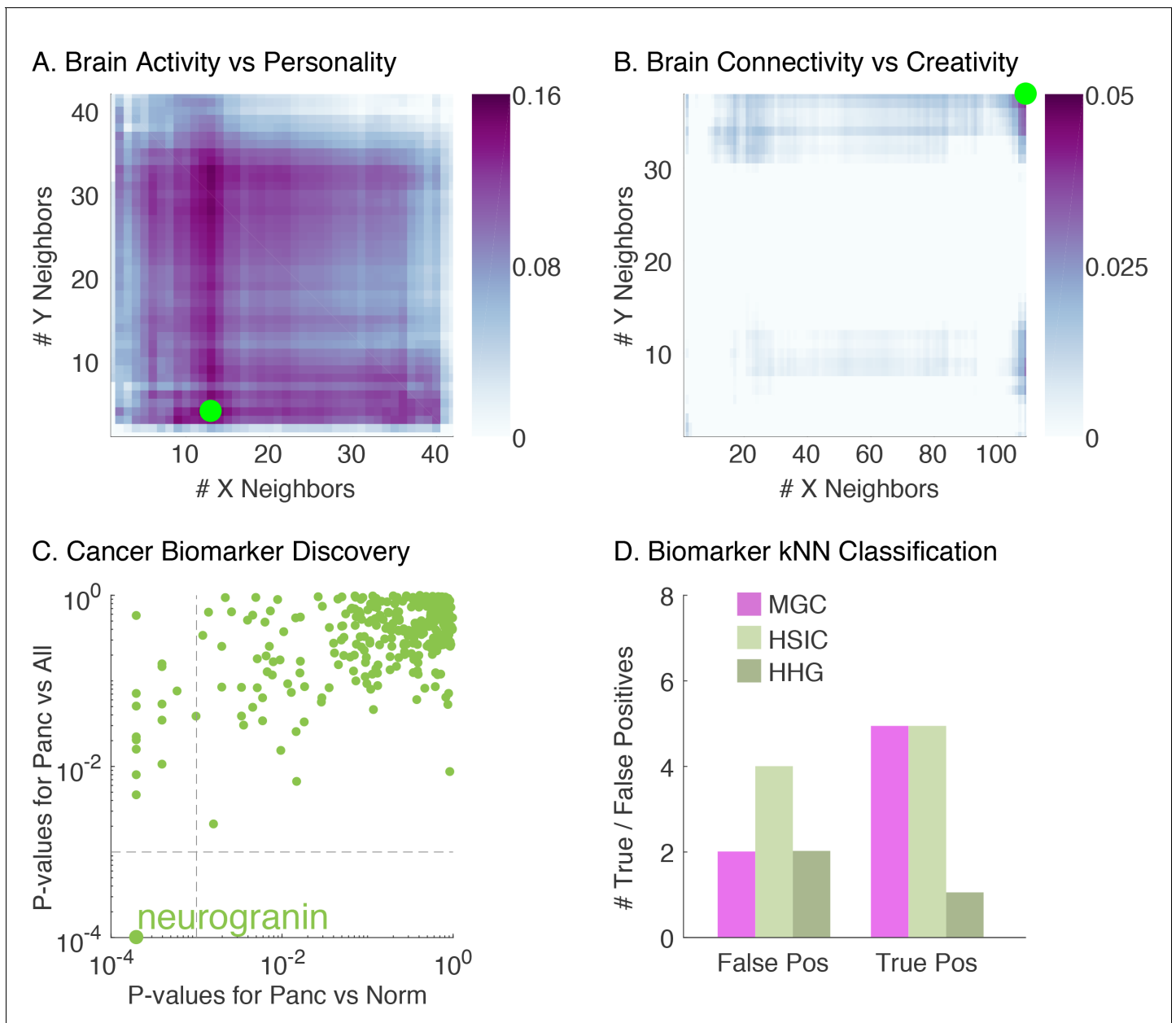
**Figure 3.** The Mgc-Map characterizes the geometry of the dependence function. For each of the 20 panels, the abscissa and ordinate denote the number of neighbors for  $X$  and  $Y$ , respectively, and the color denotes the magnitude of each local correlation. For each simulation, the sample size is 60, and both  $X$  and  $Y$  are one-dimensional. Each dependency has a different Mgc-Map characterizing the geometry of dependence, and the optimal scale is shown in green. In linear or close-to-linear relationships (first row), the optimal scale is global, that is the green dot is in the top right corner. Otherwise the optimal scale is non-global, which holds for the remaining dependencies. Moreover, similar dependencies often share similar Mgc-Maps and similar optimal scales, such as (10) logarithmic and (11) fourth root, the trigonometric functions in (12) and (13), (16) circle and (17) ellipse, and (14) square and (18) diamond. The Mgc-Maps for high-dimensional simulations are provided in **Figure 3—figure supplement 1**.

DOI: <https://doi.org/10.7554/eLife.41690.012>



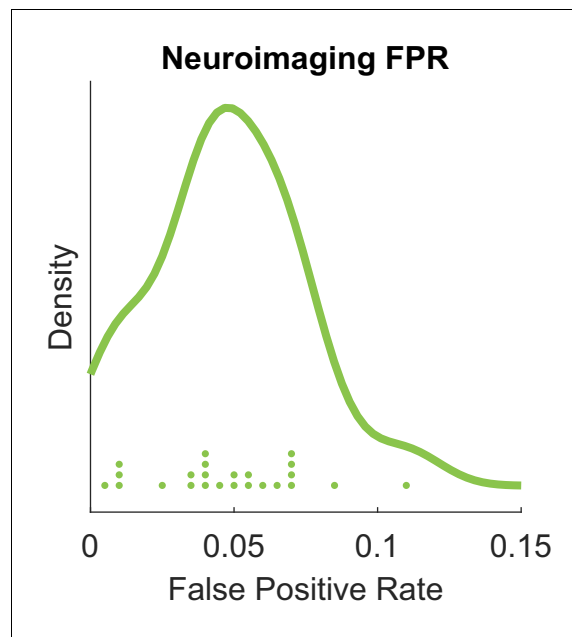
**Figure 3—figure supplement 1.** The Mgc-Map for the 20 panels for high-dimensional dependencies. For each simulation, the sample size is 100, and the dimension is selected as the dimension such that Mgc has a testing power above 0.5. It has similar behavior and interpretation as the one-dimensional power maps in **Figure 3**, that is the linear relationships optimal scales are global, and similar dependencies share similar Mgc-Maps.

DOI: <https://doi.org/10.7554/eLife.41690.013>



**Figure 4.** Demonstration that Mgc successfully detects dependency, distinguishes linearity from nonlinearity, and identifies the most informative feature in a variety of real data experiments. (A) The Mgc-Map for brain activity versus personality. Mgc has a large test statistic and a significant p-value at the optimal scale (13, 4), while the global counterpart is non-significant. That the optimal scale is non-global implies a strongly nonlinear relationship. (B) The Mgc-Map for brain connectivity versus creativity. The image is similar to that of a linear relationship, and the optimal scale equals the global scale, thus both Mgc and M<sub>CORR</sub> are significant in this case. (C) For each peptide, the x-axis shows the p-value for testing dependence between pancreatic and healthy subjects by Mgc, and the y-axis shows the p-value for testing dependence between pancreatic and all other subjects by Mgc. At critical level 0.05, Mgc identifies a unique protein after multiple testing adjustment. (D) The true and false positive counts using a k-nearest neighbor (choosing the best  $k \in [1, 10]$ ) leave-one-out classification using only the significant peptides identified by each testing method. The peptide identified by Mgc achieves the best true and false positive rates, as compared to the peptides identified by Hsic or HHG.

DOI: <https://doi.org/10.7554/eLife.41690.014>



**Appendix 1—figure 1.** We demonstrate that MGC is a valid test that does not inflate the false positives in screening and variable selection. This figure shows the density estimate for the false positive rates of applying MGC to select the 'falsely significant' brain regions versus independent noise experiments; dots indicate the false positive rate of each experiment. The mean  $\pm$  standard deviation is  $0.0538 \pm 0.0394$ .

DOI: <https://doi.org/10.7554/eLife.41690.025>