Information-Content-Informed Kendall-tau Correlation: Utilizing Missing Values

Robert M Flight[1,2,3], Praneeth S Bhatt[4], and Hunter NB Moseley[1,2,3,5,6,✉]

[1] Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA
[2] Department of Molecular & Cellular Biochemistry, University of Kentucky, Lexington, KY 40536, USA
[3] Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky, Lexington, KY 40536, USA
[4] Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40506
[5] Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA
[6] Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, KY 40536, USA

✉ Correspondence: Hunter NB Moseley <hunter.moseley@uky.edu>

Abstract

Almost all correlation measures currently available are unable to handle missing values. Typically, missing values are either ignored completely by removing them or are imputed and used in the calculation of the correlation coefficient. In both cases, the correlation value will be impacted based on a perspective that missing data represents no useful information. However, missing values occur in real data sets for a variety of reasons. In omics data sets that are derived from analytical measurements, the primary reason for missing values is that a specific measurable phenomenon falls below the detection limits of the analytical instrumentation. These missing data are not missing at random, but represent some information by their "missingness." Therefore, we propose an information-content-informed Kendall-tau (ICI-Kt) correlation coefficient that allows missing values to carry explicit information in the determination of concordant and discordant pairs. With both simulated and real data sets from RNA-seq experiments, we demonstrate that the ICI-Kt allows for the inclusion of missing data values as interpretable information. Moreover, our implementation of ICI-Kt uses a mergesort-like algorithm that provides O(nlog(n)) computational performance. Finally, we show that approximate ICI-Kt correlations can be calculated using smaller feature subsets of large data sets with significant time savings, which has practical computational value when feature sizes are very large.

The ICI-Kt correlation calculation is available in an R package and Python module on GitHub at https://github.com/moseleyBioinformaticsLab/ICIKendallTau and https://github.com/moseleyBioinformaticsLab/icikt, respectively.

Introduction

Correlation as a measure of the relatedness or similarity of two or more sets of data has a long history, with the mathematical technique being used (and abused) in various scientific fields since it's introduction [1,2]. More recently, correlation calculations have become a cornerstone statistical method in the analysis and integration of varied omics' datasets, especially the big five omics: genomics, transcriptomics, proteomics, metabolomics, and epigenomics [3]. Correlation between biomolecular features (nucleotide variants, RNA transcripts, proteins, metabolites) may be used to evaluate the relationship strength between pairs of the features as well as to detect and derive correlative structures between groups of features [4]. Moreover, feature-feature correlations can be used to evaluate a dataset based on expected biochemical correlations, for example higher feature-feature correlations within lipid categories versus between lipid categories [5]. Moreover, correlation is a foundational method for generating biomolecular feature-feature interaction networks, like those provided by STRING [6], Genemania [7], and WCGNA [8]. Feature-feature correlation may also be used to inform which features are used for imputation of missing values [9]. Correlation between biological samples as measured from omics data sets may be used for evaluating whether all the samples from a particular condition are truly related to each other, as well as to inform the choice of analysis, and to check for outliers prior to other statistical analyses [10,11]. Outlier detection, in particular, is often required for successful omics data analysis, since any misstep during the experimentation, sample collection, sample preparation, or analytical measurement of individual samples can inject high error and/or variance into the resulting data set [11,12].

All analytical methods, and in particular the analytical methods used in omics where many analytes are being measured simultaneously; suffer from missing measurements. Of course some analytes will be missing at random because of random issues with either the instrument or the particular sample, but a larger number of missing measurements are left-censored due to analytes being below the effective detection limit of the instrument, as shown in Figure 1. Some analytical instruments are purposely designed to floor measurements when they occur below a certain signal to noise ratio threshold. Imputation of missing measurements in omics samples is an active area of research, which we will not cover here beyond to say that it is worthwhile and very necessary in many instances. But when it comes to calculating correlation, there are very few methods that explicitly account for missing data that we know of. In many cases, missing values may be imputed to zero (or another value) and then included in the correlation calculation. Imputation or replacement with zero is likely to lead to high inflation of Pearson and related correlation measures. Other approaches keep only those features with no missing values across experimental samples (complete cases), or keep non-missing between any pair of experimental samples (pairwise complete cases). Each of these are likely to deviate from the real sample-sample correlation values, especially with respect to specific data interpretation perspectives.

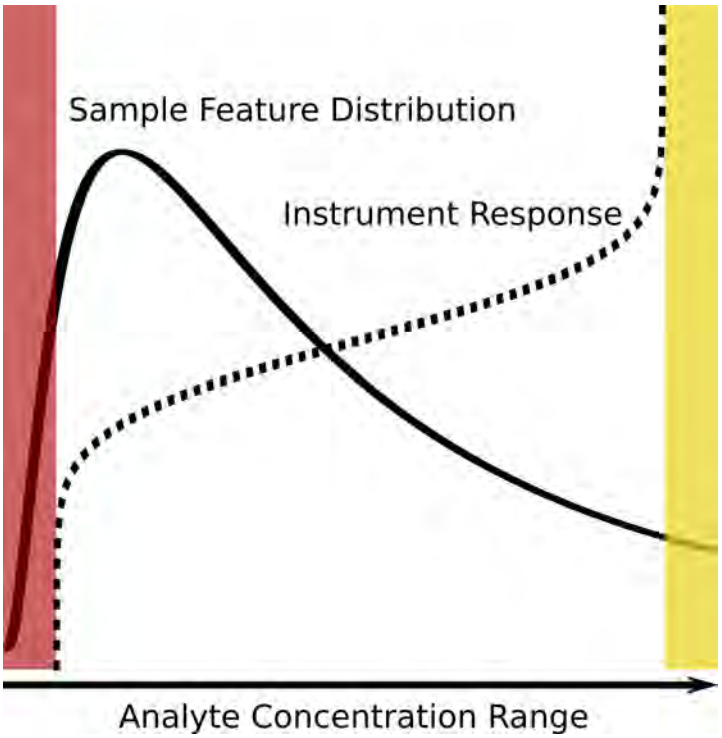

Figure 1. Graphical description of the left-censored data problem. The true analyte concentration range covers the full range of the density distribution, with the minimum on the left, and the maximum on the right. An example density plot of the analyte concentrations for a single experimental sample is shown as a solid line, and the instrument response to concentration as a dashed line. Below certain analyte concentrations, highlighted as the red region, the instrument returns missing values, 0, or some other floored value. Above certain concentrations, highlighted as the yellow region, all values returned will also be identical (or flagged depending on the instrument). Which analytes will have concentrations in the red region may vary from sample to sample due to the overall sample composition, as well as natural variances (non-systematic error) within each experimental sample.

Assuming that a majority of missing values are not missing at random, but rather result from left-censored distributions due to the analyte being below the effective detection limit (see Figure 1), we propose that these missing values do in fact encode useful information that can be incorporated into correlation calculations, especially the Kendall-tau paired rank correlation coefficient. In this manuscript, we propose new definitions of concordant and discordant rank pairs that include missing values, as well as methods for incorporating missing values into the number of tied values for the equivalent of the modified Kendall-tau coefficient. We examine the effect of missing values on simulated data, comparing our information-content-informed Kendall-tau (ICI-Kt) method with other simpler methods of handling the missing values, namely removing them or imputing them to zero. We also evaluate the ability of ICI-Kt to capture outlier samples using three different RNA-seq datasets and comparing to a standard method using Pearson correlation. Finally, we examine the ability to recapitulate ICI-Kt correlation values calculated on a large feature

omics data set using feature subsets, which is useful when the execution time of very large feature data sets becomes computationally prohibitive.

All of the code used for this manuscript is available on zenodo [13].

Methods

Definition of Concordant and Discordant Pairs

In the simplest form, the Kendall-tau correlation can be defined as:

$$\tau = \frac{|pairs_{concordant}| - |pairs_{discordant}|}{|pairs_{concordant}| + |pairs_{discordant}|}$$

In this case a pair are any two x-y points, x_i, y_i and x_j, y_j, with $i \neq j$, composed from two joint random variables X and Y, where x_i represents the *ith value* in X and y_i represents the *ith value* in Y. In an omics context, X and Y can represent feature vectors for two experimental samples or two specific features across an ordered set of samples. A concordant pair has the following classical definition:

- $x_i > x_j$ and $y_i > y_j$
- $x_i < x_j$ and $y_i < y_j$

A discordant pair has the following classical definition [14]:

- $x_i > x_j$ and $y_i < y_j$
- $x_i < x_j$ and $y_i > y_j$

Assuming that the majority of missing values are not missing at random, but due to being below the effective detection limit, we can expand the concordant and discordant pair definitions to include missing values. Here $!\,x$ indicates a missing value and & is synonymous with "and." The information-content-informed concordant pair definitions are below:

- $x_i > x_j$ and $y_i > y_j$
- $x_i < x_j$ and $y_i < y_j$
- $x_i > x_j$ and $y_i \& !\,y_j$
- $x_i < x_j$ and $!\,y_i \& y_j$
- $x_i \& !\,x_j$ and $y_i > y_j$
- $!\,x_i \& x_j$ and $y_i < y_j$
- $x_i \& !\,x_j$ and $y_i \& !\,y_j$
- $x_i \& x_j$ and $!\,y_i \& y_j$

The information content informed discordant pair definitions are below:

- $x_i > x_j$ and $y_i < y_j$
- $x_i < x_j$ and $y_i > y_j$

- $x_i > x_j$ and $!\,y_i \& y_j$
- $x_i < x_j$ and $y_i \&!\,y_j$
- $x_i \&!\,x_j$ and $y_i < y_j$
- $!\,x_i \& x_j$ and $y_i > y_j$
- $x_i \&!\,x_j$ and $!\,y_i \& y_j$
- $!\,x_i \& x_j$ and $y_i \&!\,y_j$

Considering Ties

Tied values do not contribute to either of the concordant or discordant pair counts, and the original Kendall-tau formula which calculates the tau-a statistic does not consider the presence of tied values. However, the related tau-b statistic does handle the presence of tied values by adding the tied x and y values to the denominator, and in our special case of missing data, we can add the ties that result from both of (x_i, x_j) and (y_i, y_j) being missing into n_{xtie} and n_{ytie} as well [15,16].

$$\tau = \frac{|pairs_{concordant}| - |pairs_{discordant}|}{\sqrt{(n_0 - n_{xtie})(n_0 - n_{ytie})}}$$

We can also consider commonly missing values in X and Y specially as well. In the first instance, we remove those x-y points where both values are missing. We refer to this case as the "local" ICI-Kt correlation. It is most appropriate for the comparison of only two experimental samples, where we are concerned with what values are present in the two experimental samples, with the odd case of missingness.

The other case, where we leave ties resulting from points with missing X and Y, we refer to as the "global" ICI-Kt correlation. In this case, every single sample-sample correlation will consider the same number of pair comparisons. This is more useful when analyzing and interpreting correlations across multiple experimental samples, not just two feature vectors.

Theoretical Maxima

The "global" case also provides an interesting property, whereby we can calculate the theoretical maximum correlation that would be possible to observe given the lowest number of shared missing values. This value can be useful to scale the rest of the observed correlation values across many sample - sample correlations, providing a value that scales an entire dataset appropriately. For any pairwise comparison of two vectors (from experimental samples for example), we can calculate the maximum possible Kendall-tau for that comparison by defining the maximum number of concordant pairs as:

$$tau_{max} = \frac{|pairsmax_{concordant}|}{\sqrt{(n_0 - n_{xtie})(n_0 - n_{ytie})}}$$

Where:

$$pairsmax_{concordant} = n_0 - n_{xtie} - n_{ytie} - n_{tie}$$

Calculating a set of tau_{max} values between all experimental samples, we can take the maximum of those values, and use it to scale all of the obtained Kendall-tau values.

Implementation Details

We produced an initial reference implementation in R [17] where the various concordant and discordant pair definitions were written as simple logical tests to allow further exploration and validation of faster implementations. In practice, it is possible to replace the missing values with a value that is smaller than all of the values in the two sets of values under consideration, and then do tests of signs to define concordant and discordant pairs (see the definitions of concordant and discordant pairs above).

After this initial implementation, we used the *SciPy* [18] *kendalltau* code as a model to create a mergesort based implementation that has a complexity of O(nlog(n)), in comparison to the pairwise testing that has a complexity of O(n^2) [19]. This provides a 460X speedup of the runtime to compare two feature vectors with lengths of 40,000. In comparison to directly counting the concordant and discordant pairs, some of the values in the mergesort implementation can reach the 32 bit default limit in C++. Therefore, we used 64 bit integers and floats where necessary in the *Rcpp* code.

C++ (via *Rcpp* [20–22]) and R code implementations are in the src/kendallc.cpp and R/kendalltau.R files of the ICIKendallTau R package, hosted at https://github.com/MoseleyBioinformaticsLab/ICIKendallTau. The version of the package used in this manuscript is available on zenodo [23].

When a large number of samples need to be compared, it may be useful to split the comparisons across compute instances (across hyperthreaded cores, physical cores, or physical compute nodes). The *furrr* R package makes the definition of compute clusters easy [24]. For a large correlation matrix computation, we first define the sample-sample comparisons to be performed, then check how many instances of compute are available (defined by a previous call to *furrr*), and finally split the comparisons into a list that can be easily distributed using the *furrr::future_map* function.

We also implemented a package in Python 3 with the same functionality and similar interface to the R package, hosted on GitHub (https://github.com/MoseleyBioinformaticsLab/icikt) and Python Package Index (https://pypi.org/project/icikt). Appropriate portions of the package were *cythonized* to achieve similar execution performance to the R package. We used the *multiprocessing* Python package to distribute the sample-sample comparisons calculations across cores with very little overhead. This package includes both an application programming interface (API) with equivalent parameter options to the R package implementation and a command line interface (CLI) for generating straight-forward correlation matrix analyses without the need to write Python scripts.

Simulated Data Sets

Simulated feature vectors (analytical samples) are generated by drawing 1000 random values from a log-normal distribution with a mean of 1 and standard deviation (sd) of 0.5

and sorting them in ascending order. The negative analytical sample has values sorted in descending order. Missing value indices are generated by randomly sampling up to 499 of the lowest values in each sample. For the negative sample, the indices are also subtracted from 1000 to cause them to be at the lower end of the feature distribution. Finally, missing indices were only inserted into one of the two samples being compared before calculating the correlation. The missing indices are replaced with NA, and then correlations between the analytical samples are calculated.

Another, more realistic, simulated data set is generated by drawing 1000 random values from a log-normal distribution, and adding noise from a normal distribution with a mean of zero and sd of 0.2 to create two statistical samples. Missing values are created in these statistical samples via two methods: 1) by creating intensity cutoffs from 0 to 1.5 in 0.1 increments, values below the cutoff set to missing or zero depending on the calculation; 2) randomly sampling locations in the two-sample matrix ranging from zero to 300 in increments of 50 and setting the indices to missing or zero.

Brainson RNA-Seq Data Set

This RNA-Seq dataset is from null and knock-in EGFR mice mutants. Genotypes include Null (no mutant-EGFR inducible expression), Heterozygous (only one copy of mutant-EGFR), Homozygous (two copies of mutant EGFR). For each mutant, they were also grown in different ways and sequenced: (1) 2D plates; (2) 3D organoids; (3) cells sorted and selected using FACS; (4) total tumor without sorting. See [25] for the full experimental details.

Yeast RNA-Seq Data Set

Summarized gene level counts were obtained from a GitHub project maintained by the Barton group [26]. It should be noted that the data are also available from two figshare repositories [27,28]. These data were generated and reported as part of two publications evaluating replicate data and differential gene expression [11,29].

The original outliers reported by Gierliński et al are based on a combination of median correlations, feature outliers, and RNA-seq coverage. Given that we want to compare outliers based on correlation, we re-determined outliers using only correlation calculated by replacing zero counts with missing values and calculating sample - sample pairwise Pearson correlations on the raw feature counts using the genes present in both samples.

Adenocarcinoma Recount RNA-Seq Data Set

We downloaded the recount2 TCGA lung cancer data [30], extracted the scaled counts, and trimmed to the Stage I adenocarcinoma samples, and those genes that have a non-zero count in at least one of the samples.

Pearson and Kendall-tau correlations are calculated with zero, as well as replacing zero values with NA.

Completeness

As an addition to the correlation, we also calculate the *completeness* between any two samples. We first measure the number of entries missing in either of the samples being compared, and subtract that from the total number of features in the samples. This defines how many features are potentially *complete* between the two samples. This number, over the total number of features defines the *completeness* fraction.

$$comp = \frac{n_{feat} - (miss_i | miss_j)}{n_{feat}}$$

Subsampling of Features

In addition to using the full set of features to calculate sample-sample correlations, it is possible to select a subsample of features and only use this feature subset to calculate correlations. Three methods were implemented: 1) random selections; 2) top set of features by variance; and 3) top set of features by variance of principal components.

For (1), random selection, feature subsamples of 1 to 10% by 1% increments, and then in 5% increments to 90% were taken using the *sample* R function. For (2), selection by variance, feature variances across samples were calculated, and then ranked by the variances. For (3), selection by variance of the principal components, principal component analysis (PCA) decomposition was performed on the full data set, the variance contribution of each PC calculated, and then the percent variance for each principal component (PC) is used to determine the number of features with high loadings to select from the given PC. In this feature subsampling approach, weaker PCs contribute proportionately fewer features to the feature subset.

Pearson correlation values were calculated using both raw and log-transformed values, where the starting data had zero values replaced by NA (missing). In both the raw and log-transformed cases, final missing values were handled by either: (1) missing (NA) values left as missing (Pearson); (2) missing (NA) values replaced with 0 (Pearson 0). In the Pearson case, only features that are non-missing in both samples being compared are used for the correlation (*pairwise.complete.obs*).

Principal Component Evaluation of Subsample Coverage

The principal component decomposition of the full adenocarcinoma dataset was used to define the loadings for each principal component. For a given principal component (PC), the sum of the absolute values of the loadings defines the total loadings for that PC. The absolute values of the loadings for the subsampled features are also summed, and then divided by the total loadings previously calculated using all features. The loading fraction is calculated for each subsample of features, across all PCs. The responsible variance for each PC was also calculated as the variance of the scores in each PC. Further, the Kendall-tau correlation between the loading fraction and the PC variances was calculated. The median loading fraction across all PCs was also calculated, as well as fractional difference of the median loading fraction to the intended fraction (the number of subsampled features over the total number of features).

Computing Environment

Most calculations were run on a virtual machine with 80 virtual cores, and 1 TB of RAM. The virtual machine is running on top of a 50 node cluster, each with 4 10-core processors, 3TB of RAM and an 8TB solid-state-drive array, provided by the Kentucky Research Informatics Cloud (KyRIC). KyRIC manages the virtual machines via an OpenStack instance. We used the *drake* package to manage calculation dependencies [31]. For the comparisons of time taken using different numbers of samples to evaluate the algorithmic complexity, calculations were run on a single laptop Intel i7 core clocked at 2.2 GHz.

Results

Comparison To Other Correlation Measures

We compared the ICI-Kt correlation to both Pearson and regular Kendall-tau-b correlations as calculated by the built-in R functions using simulated data sets with missing values (Figures S1-S3).

We created two samples with 1000 observations each drawn from a log-normal distribution, and sorted in each case. The *true* correlation for each of the Kendall and Pearson correlations were calculated, and then missingness was introduced in the lower range of values (see Methods).

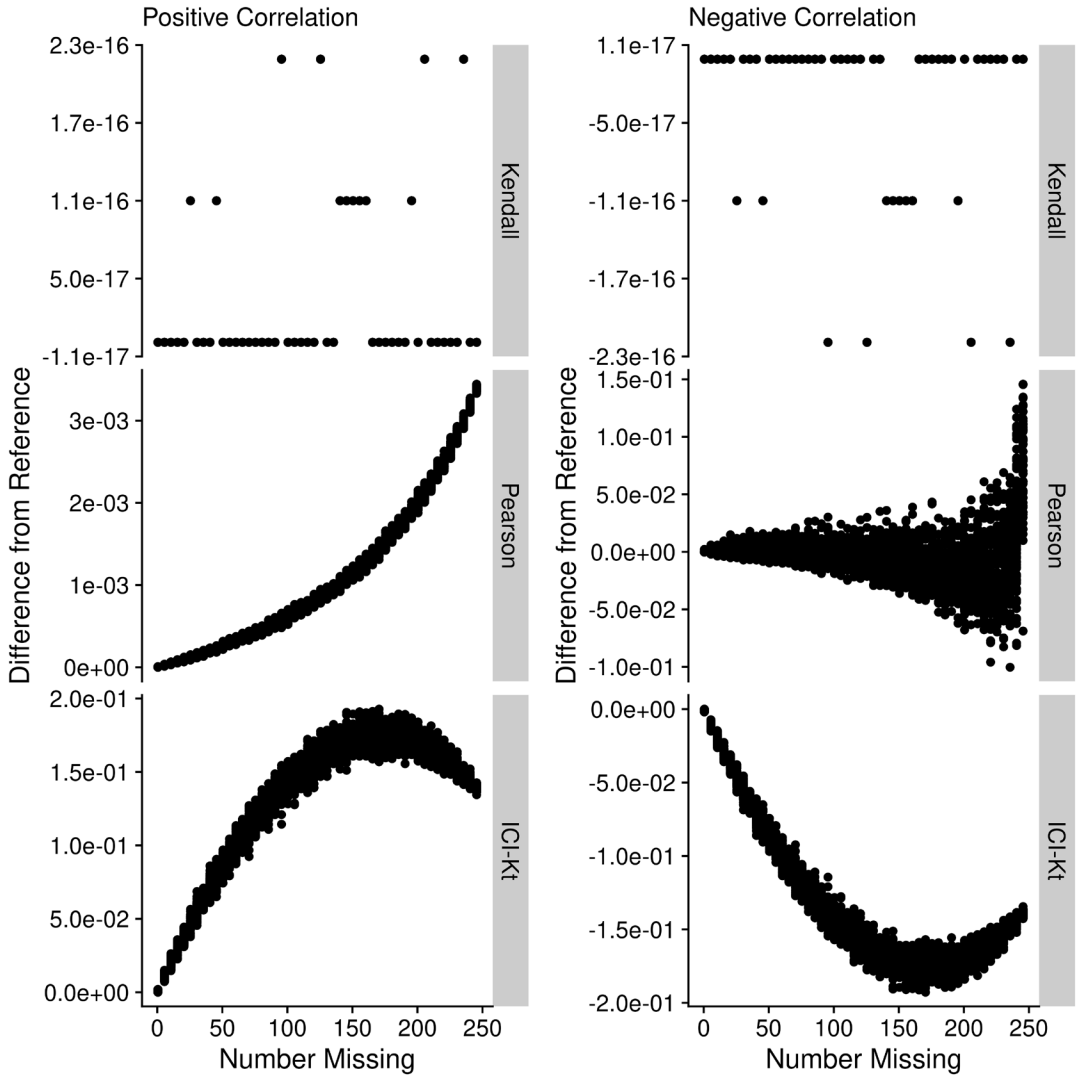

Figure 2. The effect on Kendall, Pearson and ICI-Kt correlations as increasing number of missing values in the bottom half of either sample for both positively and negatively correlated samples. A subset of 10,000 points was used for visualization. Pay attention to the orders of magnitude differences between the y-axis scales.

In Figure 2, we can see that as missing values are added, only the ICI-Kt correlation values change in any significant way. As the number of missing values increase, the ICI-Kt values drop or increase by up to 0.2. Similarly, Pearson correlation is also affected, but the degree of change in the correlation values are much less (notice the orders of magnitude differences in the y-axis scales), on the order of only 0.003 for the positive case and 0.1 for the negative case. Also, the lack of symmetry in the Pearson negative comparisons is due to the lack of linear correlation between the samples. These results demonstrate that the ICI-Kt correlation has quantitative sensitivity to missing values over the normal Kendall tau correlation and linear Pearson correlation where points with missing values are ignored (pairwise complete).

Effect of Left Censoring VS Random Missing Data

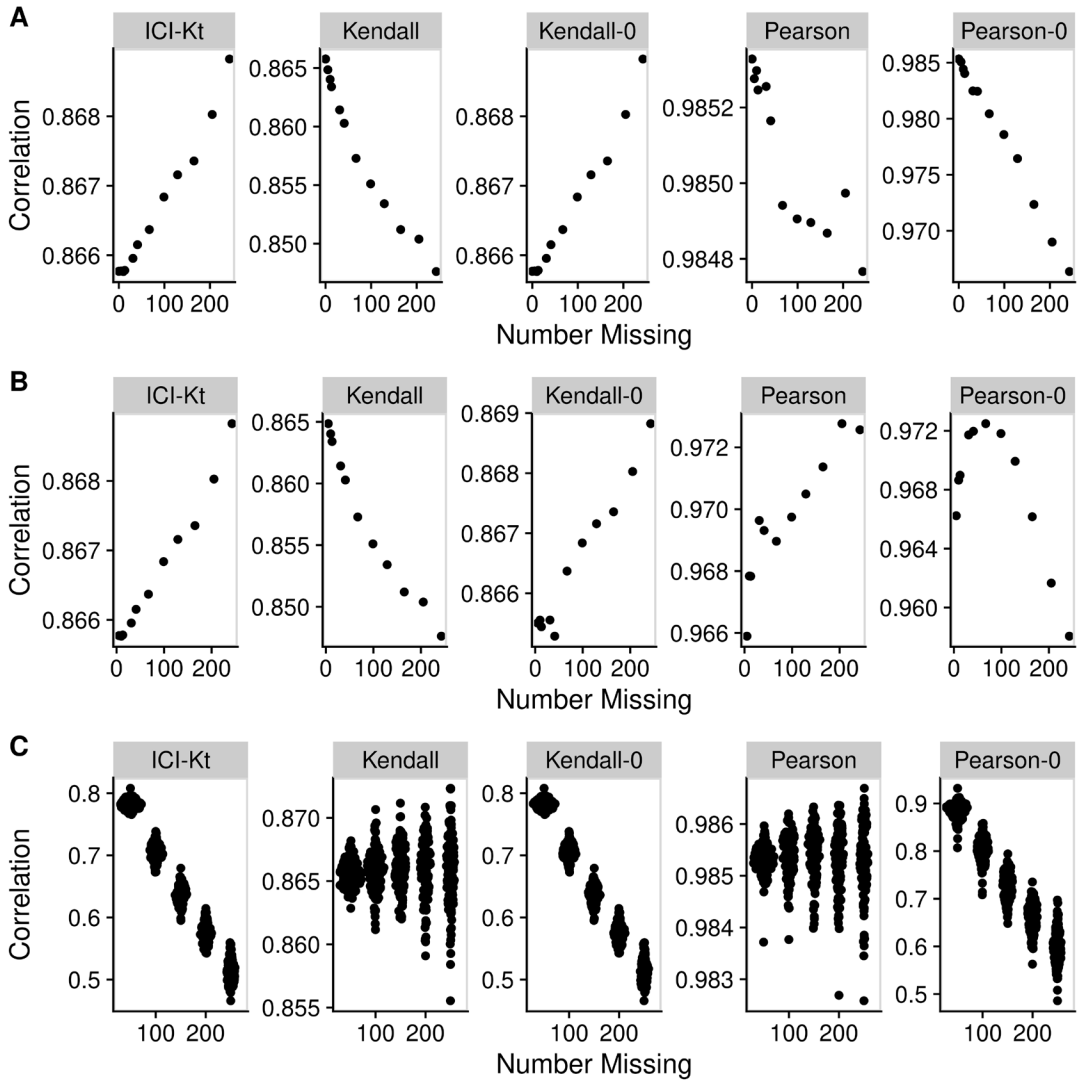

Figure 3. Effect of introducing missing values from a cutoff (A & B) or randomly (C) on different measures of correlation, including ICI-Kt, Kendall with pairwise complete, Kendall replacing missing with 0, Pearson with pairwise complete, and Pearson replacing missing with 0. A) Missing values introduced by setting an increasing cutoff. B) Missing values introduced by setting an increasing cutoff, and then log-transforming the values before calculating correlation. C) Missing values introduced at random. For the random case, each sample of random positions was repeated 100 times.

Figure 3 demonstrates the effect of introducing left-center versus random missingness in five different measures of correlation, including the ICI-Kt, the normal Kendall tau with pairwise complete, the normal Kendall tau replacing missing with 0, Pearson with pairwise complete, and Pearson replacing missing with 0. The ICI-Kt correlation demonstrates a slight increase from the starting 0.86 correlation value with growing left-centered missingness caused by a slight reinforcement of the correlation, while with growing

random missingness, the ICI-Kt correlation drops precipitously due to the large increase in discordant pairs caused by the random missing values. The normal Kendall tau correlation with pairwise complete has a small decrease in the correlation value with growing left-centered missingness caused by a loss of supporting pairs, while this correlation has a near constant average value with growing random missingness. The normal Kendall tau correlation replacing missing with 0 has identical behavior to the ICI-Kt correlation. The Pearson correlation with pairwise complete is constant over growing left-centered and random missingness. The Pearson correlation replacing missing with 0 demonstrates a small decrease in the correlation value with growing left-centered missingness due to the zero values causing some deviation from linearity while this metric drops precipitously with growing random missingness with a magnitude similar to the ICI-Kt and normal Kendall tau replacing missing with 0. Overall, the ICI-Kt and the normal Kendall-tau replacing missing with zero have the desirable characteristics of maintaining the correlation with growing left-centered missing while sharply dropping the correlation with growing random missingness. In a naive treatment of the left-centered missing data, if the values below the cutoff are set to missing followed by log-transforming the values and subsequently setting missing values to 0, then the Kendall tau correlation replacing missing with 0 will show some very odd patterns at low intensity cutoffs due to the introduction of discordant pairs. Likewise, Pearson correlation replacing missing with 0 shows a parabolic effect with increasing missing values.

Algorithmic Performance

We compared the performance of our *Rcpp mergesort* implementation to the base R *cor* function using both "pearson" and "kendall" methods on a single core with increasing numbers of features.

Figure 4 shows that the "kendall" method is the slowest by far. Zoomed out (inset), the ICI-Kt appears as fast as using "pearson," but upon zooming in, it's possible to see that ICI-Kt is using increasing amounts of time much faster than "pearson." In fact, regressing the time to number of samples using a formula with the expected algorithmic complexity, "pearson" shows O(n), ICI-Kt is O(nlog(n)), and "kendall" is O(n^2) (see Figure S4).

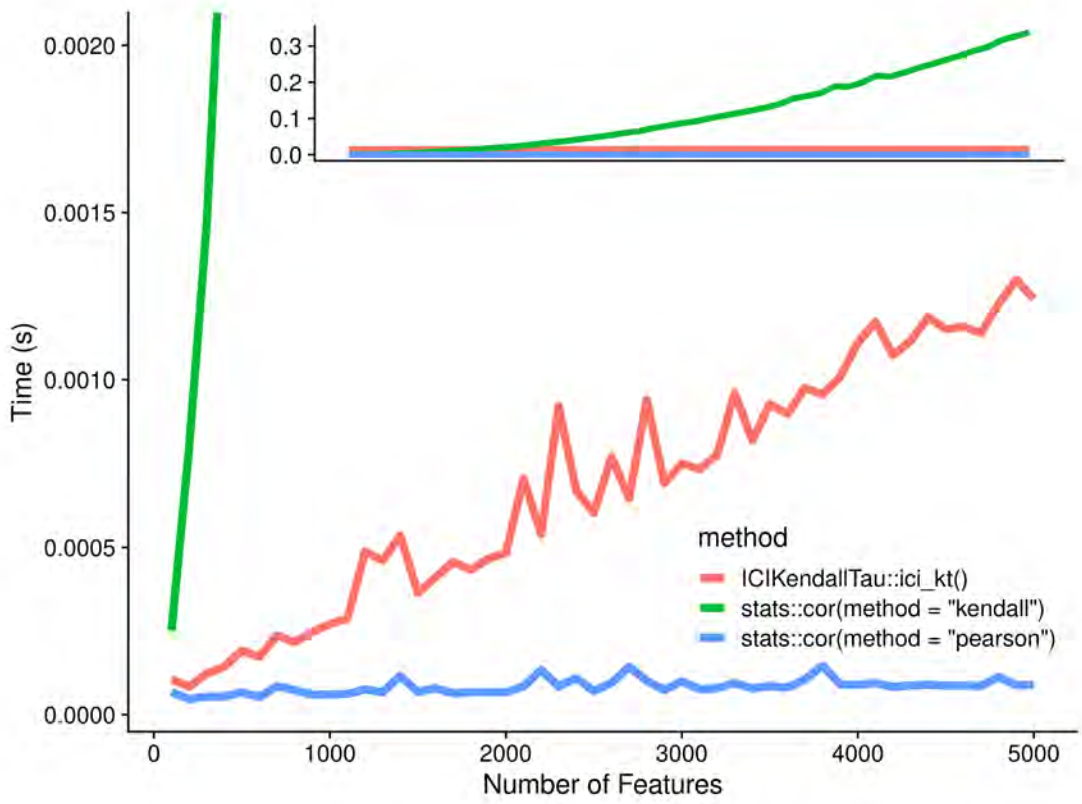

Figure 4. Comparison of the *Rcpp mergesort ici_kt*, and base R's *cor* function using the "pearson" and "kendall" methods as the number of features are increased from 100 to 5000. Inset is the expanded view showing the full range of computation time in seconds.

Multi-Processing Performance

In addition to an evaluation of the mergesort performance, we also wanted to evaluate the multi-processing performance. For this, we ran with small numbers of samples **without** parallelization enabled, and then with larger numbers of samples **with** parallelization enabled.

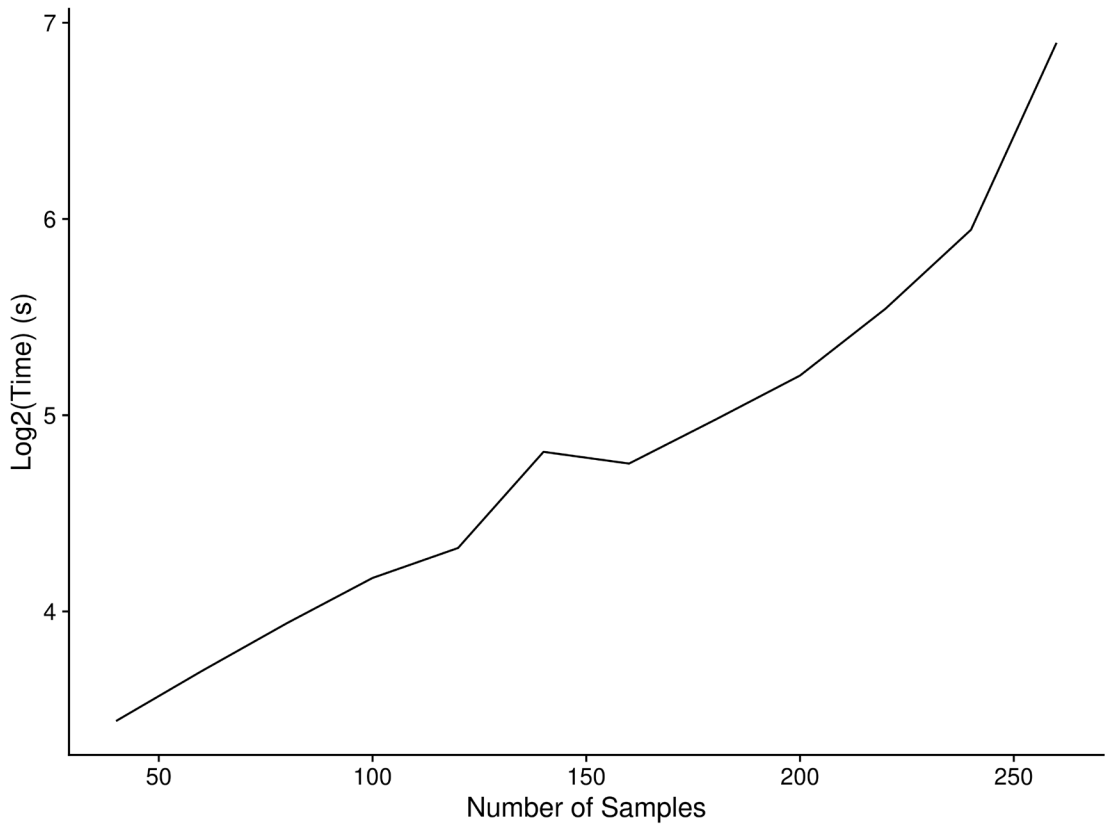

Figure 5. Time taken by ICI-Kt as a function of the number of samples with multi-processing at the R level, using 80 cores.

As samples are added, the run time quickly increases logarithmically, similar to increasing samples on a single core (see Figure 5).

Utility in Large Omics' Data Sets

Detecting Outlier Samples

We analyzed three RNA-Seq datasets for possible outliers using a variety of different correlation and correlation-related metrics: (1) ICI-Kt; (2) ICI-Kt scaled by *completeness* (see Methods); (3) Pearson correlation; (4) Pearson using pairwise complete features; (5) Pearson on log(value + 1); (6) Pearson on log(value); and (7) Kendall-tau. We primarily concentrate on the ICI-Kt, ICI-Kt * completeness, and Pearson log(x + 1) here, but all outliers by method are shown in the Supplemental materials (Figures S5-S7, Tables S1, S2). Additionally, we compared how the outliers change if we require that a feature is present in at least a certain fraction of samples of a class with each different correlation measure (Figures S8-S16). For each dataset and correlation-related metric, the median correlation for a sample was calculated amongst the samples for each sub-group of samples corresponding to an experimental condition of interest (disease, mutant, how cells were grown).

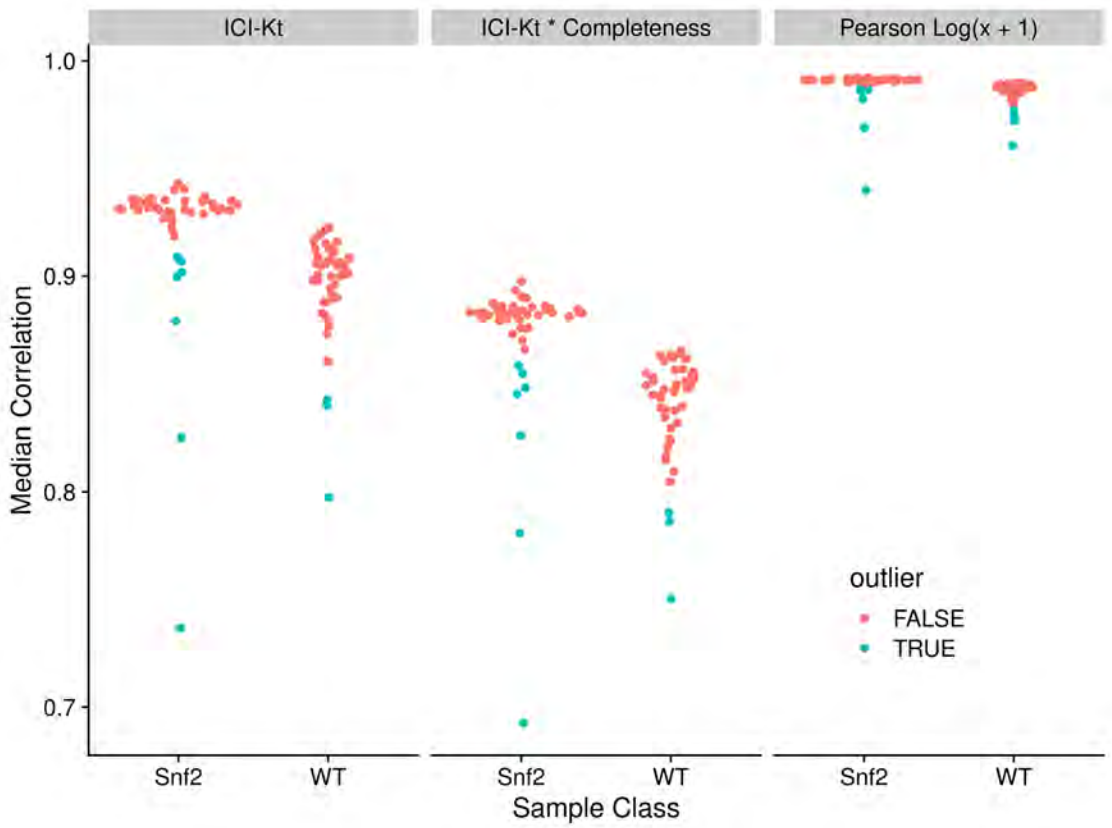

Figure 6. Median correlations for each of the yeast RNA-Seq sample to all other samples in the same group, using either ICI-Kt, ICI-Kt * completeness or Pearson Log(x + 1) correlation.

Table 1. Yeast dataset median correlation values and outlier determination for each outlier from each of the correlation methods.

	ICI-Kt		ICI-Kt * Completeness		Pearson Log(x + 1)	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
Snf2.10	X	0.91	X	0.85	X	0.99
Snf2.15	X	0.90	X	0.85	X	0.99
Snf2.31	X	0.90	X	0.85	X	0.99
Snf2.35	X	0.91	X	0.86	X	0.99
Snf2.25	X	0.88	X	0.83	X	0.98
Snf2.13	X	0.83	X	0.78	X	0.97
Snf2.06	X	0.74	X	0.69	X	0.94

	ICI-Kt		ICI-Kt * Completeness		Pearson Log(x + 1)	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
WT.36		0.86		0.81	X	0.98
WT.28		0.86		0.80	X	0.98
WT.34	X	0.84	X	0.79	X	0.97
WT.25	X	0.84	X	0.79	X	0.97
WT.21	X	0.80	X	0.75	X	0.96

Figure 6 visualizes the sample-specific median of each correlation-related metric calculated from the Yeast RNA-Seq data set. Table 1 lists the outliers detected based on each sample-specific median metric value.
In this example, ICI-Kt shows superior sensitivity for outlier detection as compared to the other metrics and is likely related to the increased spread of median metric values as compared to Pearson. We also note that these outliers seem to be a superset of the outliers noted by Gierliński et al. (see Figure S5 and Table S1).

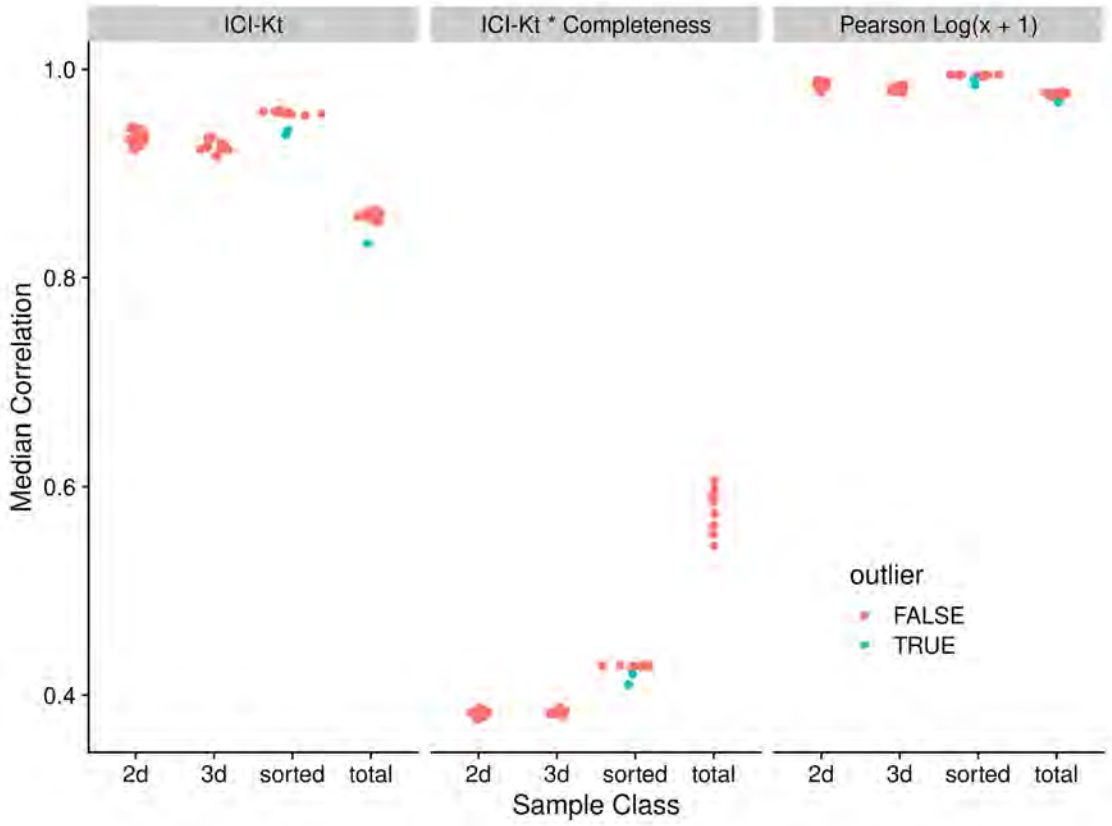

Figure 7. Median correlations for each of the Brainson RNA-Seq samples to all other samples in the same group, using either ICI-Kt, ICI-Kt * completeness or Pearson Log(x + 1) correlation.

Table 2. Brainson dataset median correlation values and outlier determination for each outlier from each of the correlation methods.

	ICI-Kt		ICI-Kt * Completeness		Pearson Log(x + 1)	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
1004-het.sorted	X	0.94	X	0.42	X	0.99
783-het.sorted	X	0.94	X	0.41	X	0.98
1263wt_S1.total	X	0.83		0.56	X	0.97

Figure 7 visualizes the sample-specific median of each metric calculated from the Brainson RNA-Seq data set. Likewise, Table 2 lists the outliers detected based on each metric. In this example, Pearson shows superior sensitivity to outlier detection, likely due to the larger spread in median Pearson correlation values.

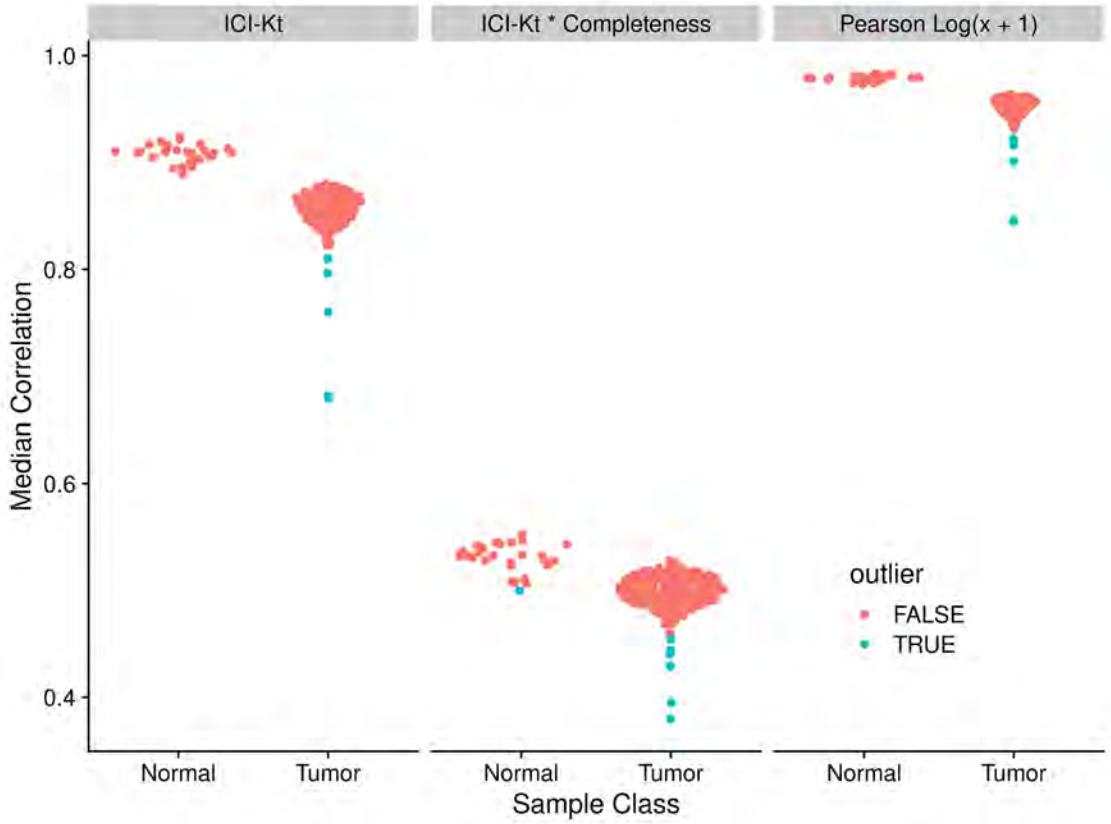

Figure 8. Median correlations for each of the TCGA adenocarcinoma RNA-Seq samples to all other samples in the same group, using either ICI-Kt, ICI-Kt * completeness or Pearson Log(x + 1) correlation.

Table 3. TCGA adenocarcinoma dataset median correlation values and outlier determination for each outlier from each of the correlation methods.

	ICI-Kt		ICI-Kt * Completeness		Pearson Log(x + 1)	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
N.186		0.89	X	0.50		0.97
T.204		0.83	X	0.45		0.94
T.171	X	0.81		0.46	X	0.92
T.215	X	0.81	X	0.44	X	0.92
T.12	X	0.80	X	0.43	X	0.92
T.261	X	0.76	X	0.44	X	0.90
T.259	X	0.68	X	0.38	X	0.85
T.14	X	0.68	X	0.40	X	0.84

Figure 8 visualizes the sample-specific median metric value calculated from the TCGA RNA-Seq data set, with Table 3 listing the outliers detected from each metric. In this example, the ICI-Kt * completeness correlation demonstrated superior sensitivity for outlier detection, likely due to its better sensitivity to measurement completeness. The wider spread of measurement completeness is expected, since TCGA is an aggregate of samples collected from various human cancer patents as compared to the first two examples which were collected from specific experiments in cell culture (Table S4). The wider spread of measurement completeness is expected, since TCGA is an aggregate of samples collected from various human cancer patients as compared to the first two examples which were collected from specific experiments in cell culture. However, the outliers detected are not a complete superset of the outliers detected by ICI-Kt and Pearson, probably due to the integration of completeness and correlation.

Feature Subsets for Extremely Large Datasets

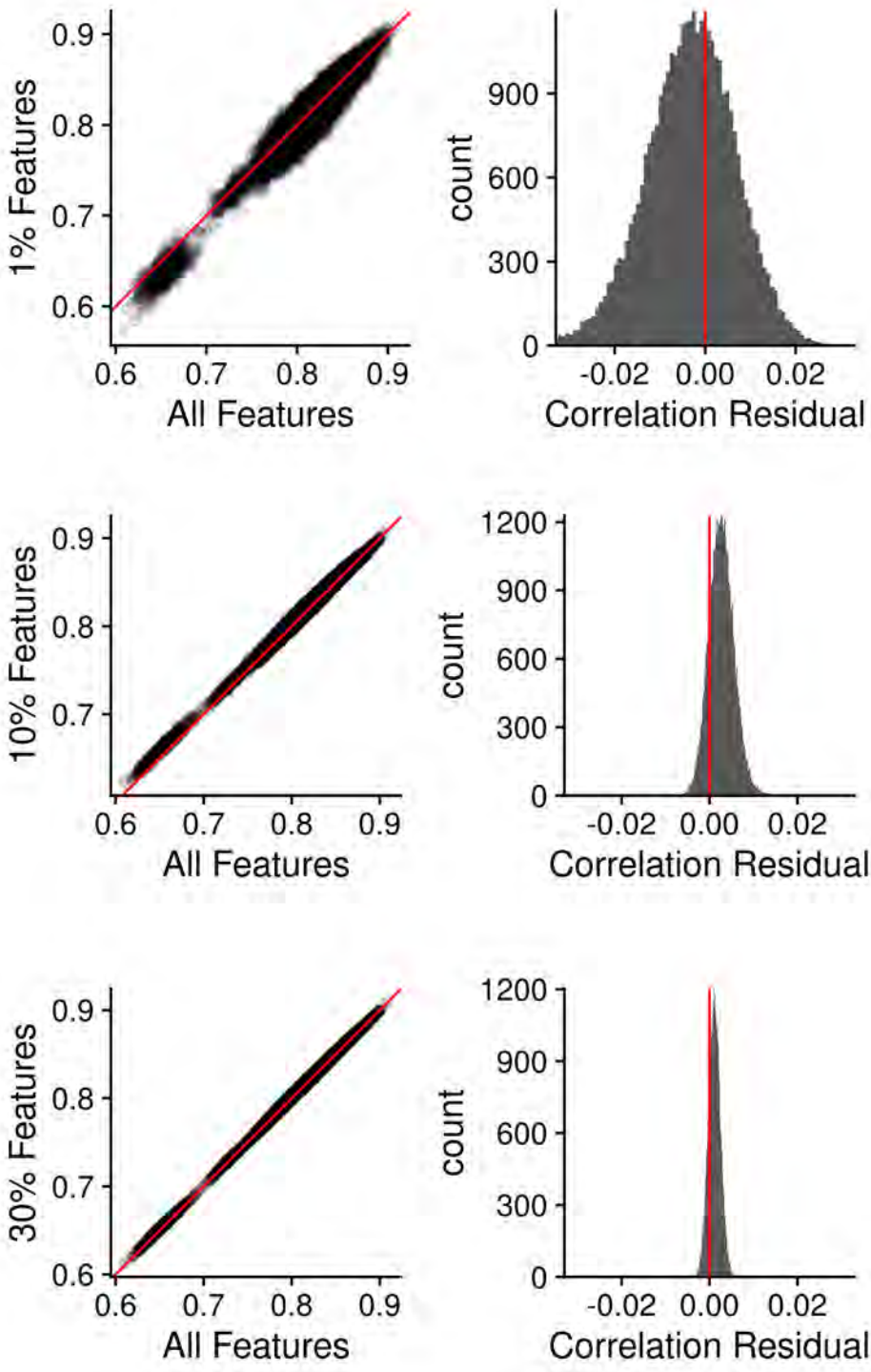

Figure 9. Sample-sample ICI-Kt correlations using all features compared with 1%, 10% and 30% random subsets of features. Left: The sample-sample correlations using all features

are plotted against the sample-sample correlations using a random subset. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (random subset - all features).

Given the length of time it takes to run the calculations for larger datasets (in particular RNA-Seq data), we wondered if "good enough" results could be obtained using subsamples of the features across the samples. Our starting data is the same set of stage I (I, Ia, and Ib) adenocarcinoma tumor and normal tissue RNA-Seq samples from the *recount2* project used for outlier detection, with 55128 features that have a non-zero value in at least one sample, and 264 samples.

In our original *Rcpp* implementation using pairwise comparisons, the adenocarcinoma pairwise comparisons took 6.7 hours, on an 80 core virtual machine. With the mergesort implementation, that time was greatly reduced to just over a minute (69 seconds). However, given that there may still be cases where the number of comparisons are so large that run time is still a concern, we investigated the effect of estimating sample-sample correlations using a subset of the features.

Figure 9 compares the sample-sample correlations from a random percent subsample of features to those obtained using all of the features. Both the pairwise plot (left) and histogram of residual differences (right) show that the correlations become much closer to those obtained when using all of the features as the percentage of subsampled features is increased. However, there is a consistent positive bias in the estimate, regardless of the number of subsamples used (see also Figure S17). This bias is small, because we are comparing the **raw** correlation values, and not the rescaled values. When the rescaled values (using the maximum correlation) are compared, the positive bias becomes larger, due to the differences in the maximum correlation (results not shown).

For the Pearson and Kendall correlation methods, the subsampled correlation results are almost identical to the ICI-Kt, with slightly higher residual differences for both Pearson and Kendall over ICI-Kt, and then lower residuals if the missing values (NA) from log-transformation of zeros are replaced with zero (see Figures S18-S26).

We also tested Pearson correlation subsets without log-transforming the count data first. As shown in Figure 10, as well as Figures S22 and S23, when the transcriptomics data is not log-transformed, Pearson correlation using subsets of features deviates wildly from those obtained using the full data set, much more than ICI-Kt or any of the other methods used.

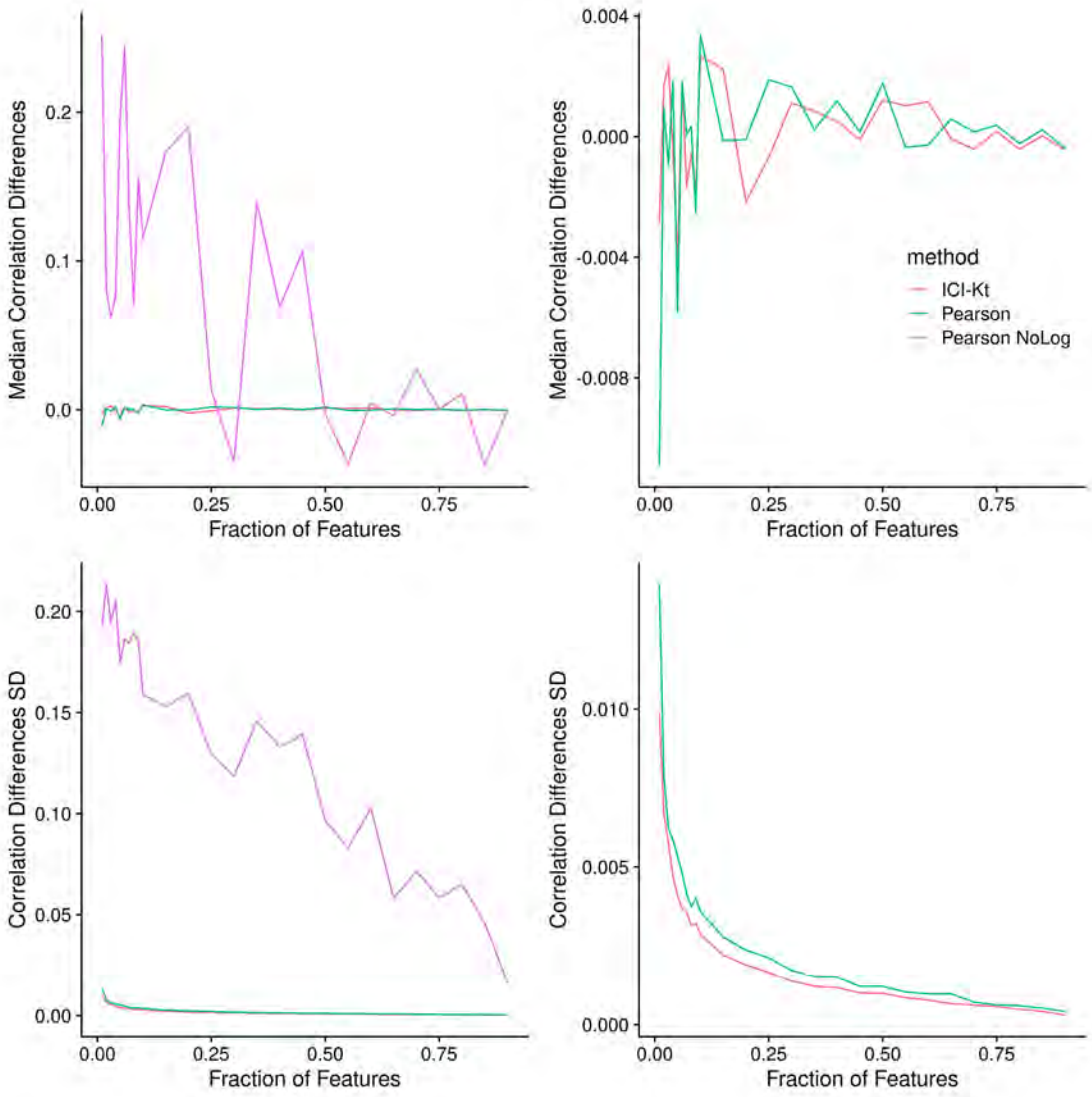

Figure 10. Median (top) and standard deviation (SD, bottom) of correlation differences using subsets as the fraction of features used increases, for different methods, including ICI-Kt, Pearson on log-transformed data (Pearson), and Pearson without log-transforming the data (Pearson NoLog). Left shows all three methods compared to each other, while the right omits Pearson NoLog to show the much smaller ranges of values for the other two methods in comparison to the Pearson NoLog.

In addition to the random feature subsamples, we also implemented feature selection methods based on loading contributions to each PC, as well as feature variances in the full data set (see Methods). As shown in Figure 11, both of these feature selection methods perform extremely poorly, even at the relatively large fraction of 50% of the features.

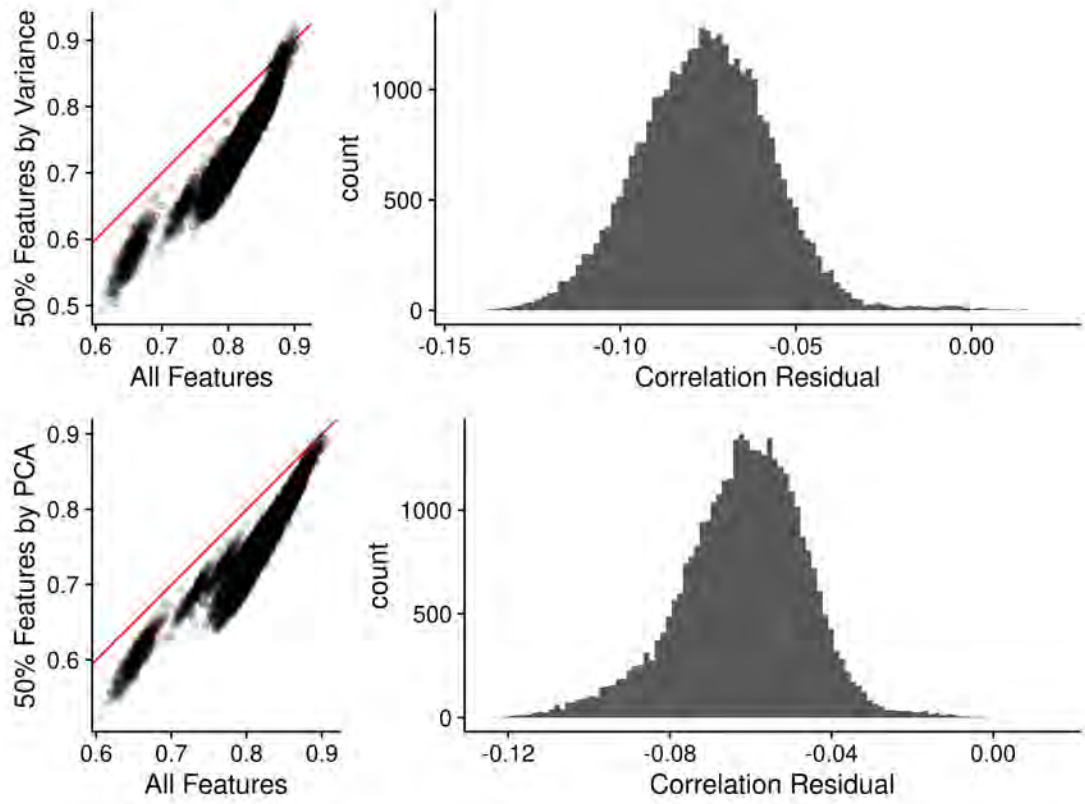

Figure 11. Sample-sample ICI-Kt correlations using all features compared with 50% non-random subsets of features. Left: The sample-sample correlations using all features are plotted against the sample-sample correlations using 50% of features chosen by variance contribution or PCA. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (subset - all features).

Outside of the differences between the subset and full set of correlations, we also evaluated each of the feature subsets using the correlation between loading fraction and principal component contributed variance (see Figure 12 A & B), and the difference between the median loading fraction and the intended fraction (C & D).

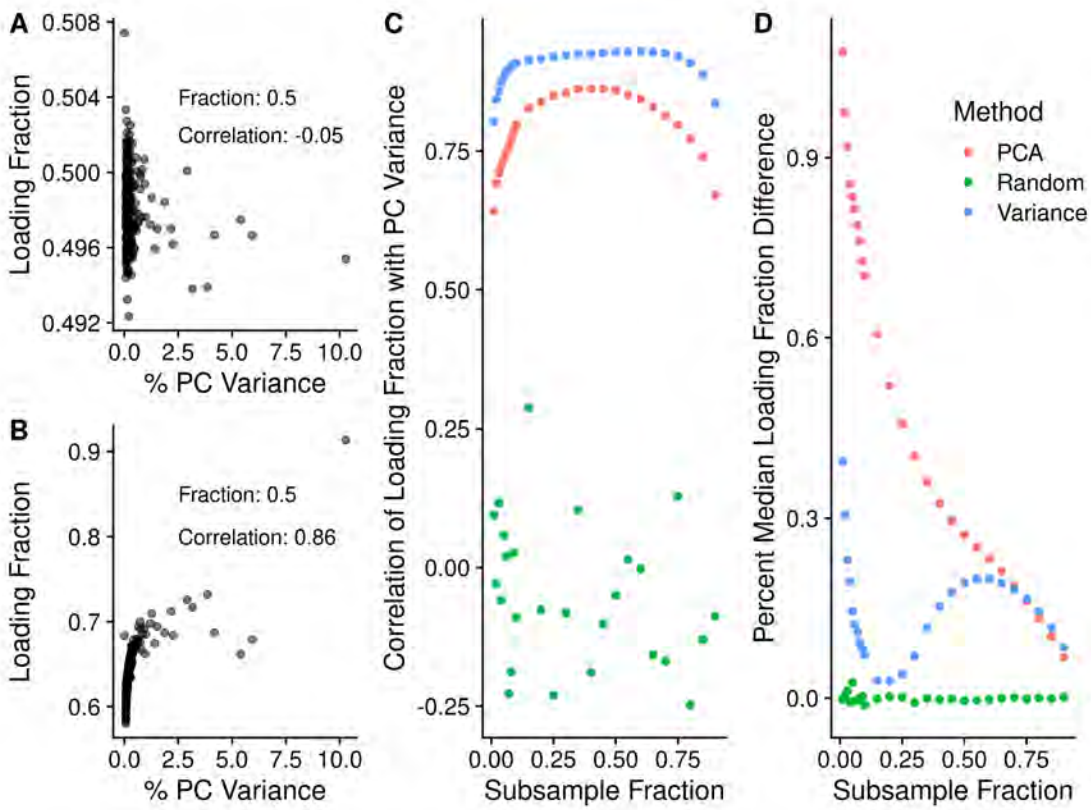

Figure 12. A) Random feature subsample of 0.5, with % principal component variance against the loading fraction. B) Variance informed feature subsample of 0.5, % principal component variance against the loading fraction. C) Subsample feature fraction plotted against the correlation of the loading fraction with PC variance for each type of subsampling. D) Subsample feature fraction plotted against the percent difference of the median loading fraction to the intended loading fraction, which is also the subsample fraction.

Figure 12 shows that for both the PCA and variance informed subsampling methods, there is a relationship between loading fraction and the percent PC variance. In contrast, the random subsampling shows no relationship of fraction to percent PC variance. The actual loading fraction covered by the subsample also varies widely from the intended fraction for the PCA and variance informed subsampling, while the random subsample fraction remains close to the intended fraction, with some minor deviation at the very low fractions.

Discussion and Conclusion

Left-censored distributions in analytical measurements of biological samples are common in biological and biomedical research, because of detection limits of the analytical instrumentation, which produces missing measurement values for all the analytes below these detection limits. As far as we are aware, previous efforts in this area are concerned with either 1: attempting to come up with better imputation methods prior to calculating correlation values; or 2: finding ways to still estimate the correlation in the face of missing

values, generally by finding maximum likelihood estimates. In the case of (1), there are many imputation methods, and new methods are still being developed, although they tend to be new combinations of old methods to handle the various types of missing values. For (2), the maximum likelihood methods generally apply to Pearson and similar types of correlation, as they benefit from the use of regression in the presence of missing data. Alvo and Cabilio's work from 1995 [32] is one of the only works we are aware of that attempts to create a general framework for rank based statistics in the presence of missing data. But, in our understanding their framework applies to data that is missing at random versus non-random missing-values, as is the case for analytes that are below the detection limit. Additionally, there does not appear to be a software implementation of Alvo and Cabilio's method available. As far as we know, information-content-informed Kendall-tau (ICI-Kt) is the first correlation method that explicitly attempts to utilize non-random missing values that occur due to being below the detection limit. ICI-Kt explicitly treats left-censored missing values as correlative information while preserving the full deleterious effects of non-random missing values on the correlation metric. Moreover, the ICI-Kt can be combined with measurement completeness to create a composite metric that is quite sensitive to overall data quality on a sample-specific level. Also, this ICI-Kt * completeness metric may have applications in cluster detection of single-cell omics data sets.

The implementations of the ICI-Kt in the presented R and Python packages provide a rich set of options for customizing the correlation calculation for a variety of use-cases and interpretations. These packages handle missing values in log-transformed data in a safe manner and have O(nlogn) performance, making them computationally practical for real-world omics data sets. Also, these packages provide multiprocessing implementations that take full advantage of modern multi-core central processing units. Furthermore, given the high amount of correlated variance in most real-world high-feature data sets, we demonstrate that random feature subsetting can be utilized to effectively estimate correlation for very large data sets or in situations where sub-second computational performance is needed for non-large data sets. However, care must be taken both in transforming data if using a method with an expectation of linearity, and selecting subsets of features. When using Pearson correlation, the log-transformation of RNA-seq data is absolutely necessary when using feature subsets. From the results presented here, random feature subsetting was superior to variance-selective feature subsetting.

As demonstrated with the three RNA-Seq datasets analyzed here, the "best" correlation-related metric will likely depend on the specific dataset and the specific data analysis step. Many factors affect this, especially correlation linearity and the modality of measurement value distributions. We would humbly suggest that for most omics datasets, the application of several correlation-related metrics simultaneously would be the best approach for outlier detection in quality control and quality assessment steps. Where one metric lacks outlier detection sensitivity, another metric will prove sensitive. Therefore, ICI-Kt and composite metrics derived from it should be considered as useful additions to the omics data analysis toolkit.

Author Contributions

RMF wrote the code in the ICIKendallTau R package, tested the ICI-Kt correlation code, and wrote all of the analysis code for this manuscript. PSB wrote the icikt Python package. HNBM conceived of the ICI-Kt correlation metric, provided input into code structures, and supervised the analyses and interpretation of results. All authors contributed to the writing of the manuscript.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported in part by grants NSF 2020026 (PI Moseley), NSF ACI1626364 (Griffioen, Moseley), P30 CA177558 (PI Evers) via the Markey Cancer Center Biostatistics and Bioinformatics Shared Resource Facility (MCC BB-SRF), P20 GM121327 (PD St. Clair), and P42 ES007380 (PI Pennell) via the Data Management and Analysis Core (DMAC).

References

1. Pearson K. Notes on the History of Correlation. Biometrika. 1920;13: 25–45. doi:10.1093/biomet/13.1.25

2. Rodgers JL, Nicewander WA. Thirteen Ways to Look at the Correlation Coefficient. The American Statistician. 1988;42: 59–66. doi:10.1080/00031305.1988.10475524

3. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32: 2847–2849. doi:10.1093/bioinformatics/btw313

4. Fukushima A, Kusano M, Redestig H, Arita M, Saito K. Integrated omics approaches in plant systems biology. Curr Opin Chem Biol. 2009;13: 532–538. doi:10.1016/j.cbpa.2009.09.022

5. Mitchell JM, Flight RM, Moseley HNB. Untargeted Lipidomics of Non-Small Cell Lung Carcinoma Demonstrates Differentially Abundant Lipid Classes in Cancer vs. Non-Cancer Tissue. Metabolites. 2021;11: 740. doi:10.3390/metabo11110740

6. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45: D362–D368. doi:10.1093/nar/gkw937

7. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, et al. GeneMANIA update 2018. Nucleic Acids Res. 2018;46: W60–W64. doi:10.1093/nar/gky311

8. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

9. Faquih T, van Smeden M, Luo J, le Cessie S, Kastenmüller G, Krumsiek J, et al. A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. Metabolites. 2020;10. doi:10.3390/metabo10120486

10. Flight RM, Wentzell PD. Preliminary exploration of time course DNA microarray data with correlation maps. OMICS. 2010;14: 99–107. doi:10.1089/omi.2009.0096

11. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31: 3625–3630. doi:10.1093/bioinformatics/btv425

12. Moseley HNB. Error Analysis and Propagation in Metabolomics Data Analysis. Computational and Structural Biotechnology Journal. 2013;4: e201301006. doi:10.5936/csbj.201301006

13. Flight RM, Moseley HNB. manuscript.ICIKendallTau. 2022. Available: https://dx.doi.org/10.5281/zenodo.6309187

14. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938;30: 81–93. doi:10.2307/2332226

15. Kendall MG. The treatment of ties in ranking problems. Biometrika. 1945;33: 239–251. doi:10.1093/biomet/33.3.239

16. Kendall MG. Rank correlation methods. First Edition. C. Griffin; 1948.

17. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

18. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

19. Knight WR. A Computer Method for Calculating Kendall's Tau with Ungrouped Data. Journal of the American Statistical Association. 1966;61: 436–439. doi:10.2307/2282833

20. Eddelbuettel D, Francois R. Rcpp: Seamless R and C++ Integration. Journal of Statistical Software. 2011;40: 1–18. doi:10.18637/jss.v040.i08

21. Eddelbuettel D. Seamless R and C++ Integration with Rcpp. New York: Springer-Verlag; 2013. doi:10.1007/978-1-4614-6868-4

22. Eddelbuettel D, Balamuta JJ. Extending R with C++: A Brief Introduction to Rcpp. The American Statistician. 2018;72: 28–36. doi:10.1080/00031305.2017.1375990

23. Flight RM, Moseley HNB. ICIKendallTau. 2022. Available: https://doi.org/10.5281/zenodo.6311002

24. Vaughan D, Dancho M. furrr: Apply mapping functions in parallel using futures. 2021. Available: https://CRAN.R-project.org/package=furrr

25. Chen F, Liu J, Flight RM, Naughton KJ, Lukyanchuk A, Edgin AR, et al. Cellular Origins of EGFR-Driven Lung Cancer Cells Determine Sensitivity to Therapy. Advanced Science. 2021;8: 2101999. doi:10.1002/advs.202101999

26. Cole C, Gierlinski M, Schofield P, Schurch N. Profiling Differential Gene Expression with a 48 replicate experiment. Geoff Barton's Computational Biology Group; 2021. Available: https://github.com/bartongroup/profDGE48

27. Barton G, Blaxter M, Cole C, Gharbi K, Gierliński M, Schofield P, et al. SNF2 knock-out yeast gene read counts from a 48 replicate experiment. figshare; 2015. pp. 0 Bytes. doi:10.6084/M9.FIGSHARE.1425502.V1

28. Barton G, Blaxter M, Cole C, Gharbi K, Gierliński M, Schofield P, et al. Wild-type yeast gene read counts from 48 replicate experiment. figshare; 2015. pp. 0 Bytes. doi:10.6084/M9.FIGSHARE.1425503.V1

29. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22: 839–851. doi:10.1261/rna.053959.115

30. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017;35: 319–321. doi:10.1038/nbt.3838

31. Landau WM. The drake R package: a pipeline toolkit for reproducibility and high-performance computing. Journal of Open Source Software. 2018;3: 550. doi:10.21105/joss.00550

32. Alvo M, Cabilio P. Rank correlation methods for missing data. Can J Statistics. 1995;23: 345–358. doi:10.2307/3315379

Supplemental Materials

Robert M Flight, Praneeth S Bhatt, and Hunter NB Moseley

2022-03-29 10:29:51

Simulated Data

The simplest simulated data set includes three samples, where two are perfectly correlated, and the third perfectly anti-correlated. We then created missing values (replaced value with NA) in each sample systematically, and computed the ICI-Kt, Pearson, and Kendall-tau correlations.

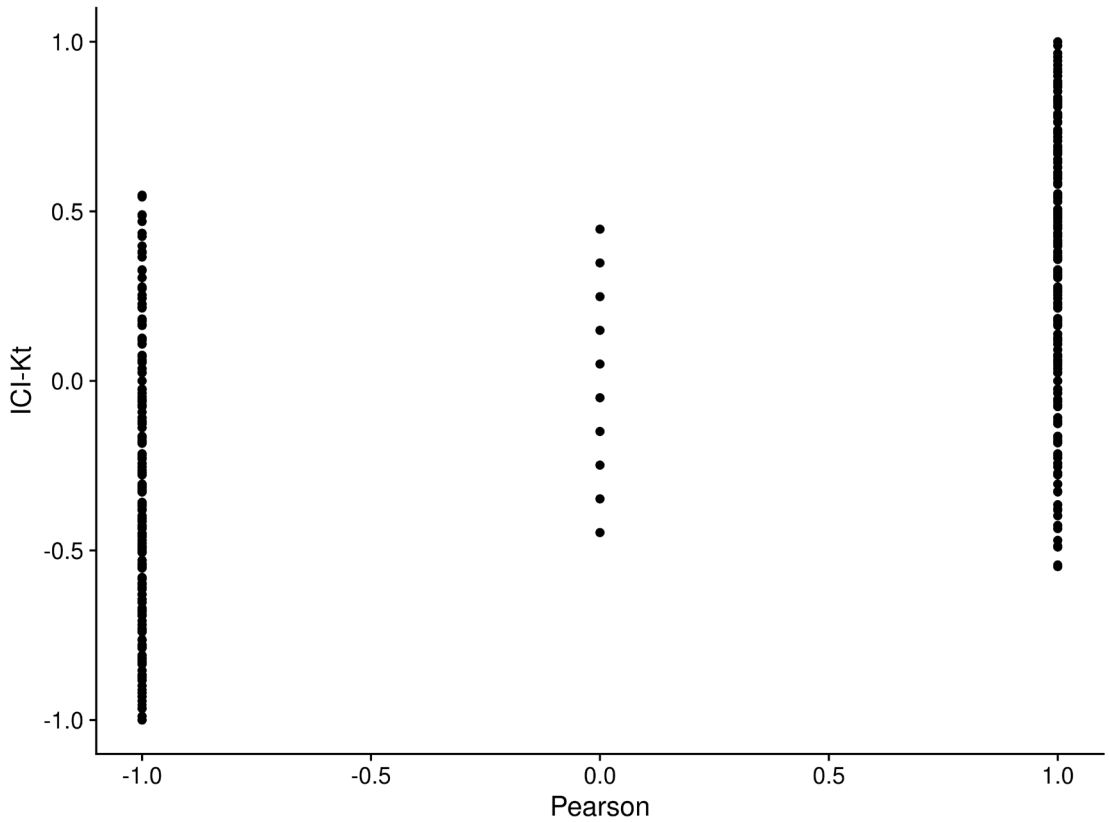

Figure S1. Comparison of ICI-Kt and Pearson correlations for perfectly positive and negatively correlated samples, systematically replacing values with NA. NA values from Pearson were replaced with zero for this comparison.

We can also examine the full set of positive and negative correlations generated as we vary the number of missing entries between two positively correlated samples and two negatively correlated samples. These distributions are shown in Figure S3. We can see that the distributions from both ICI-Kt and Kendall-tau are the same, which is expected given

we replaced missing values (NA) with zero *within* the ICI-Kt code, and replaced missing values (NA) with zero prior to calculating Kendall-tau correlations.

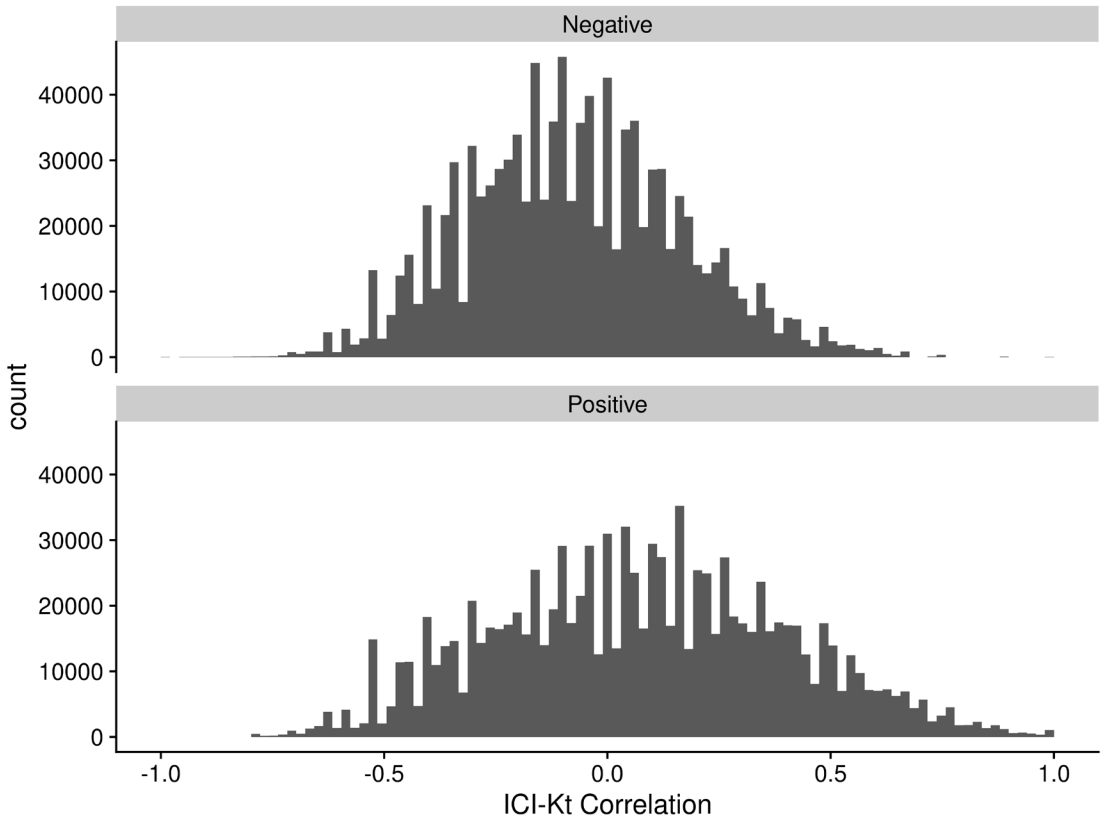

Figure S2. ICI-Kendall-tau correlation as missing values are varied between two samples.

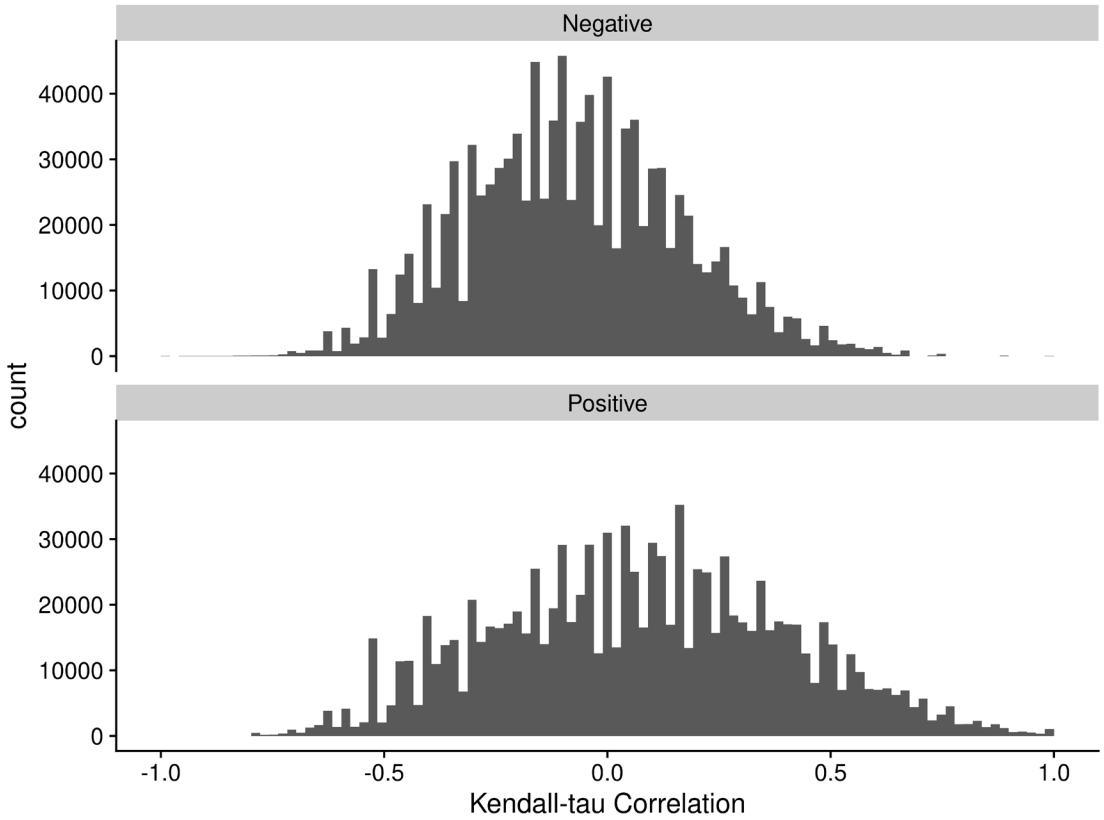

Figure S3. Kendall-tau correlation as missing values are varied between two samples and replaced with 0 before calculating Kendall-tau.

Algorithmic Complexity

In addition to comparing their performance, we can check that the algorithmic complexity fits the theoretically expected complexity by fitting a regression line of the run time to the number of items. The run times and fitted lines for each method (R's Pearson, the ICI-Kt, and R's Kendall-tau) are shown in Figure S4.

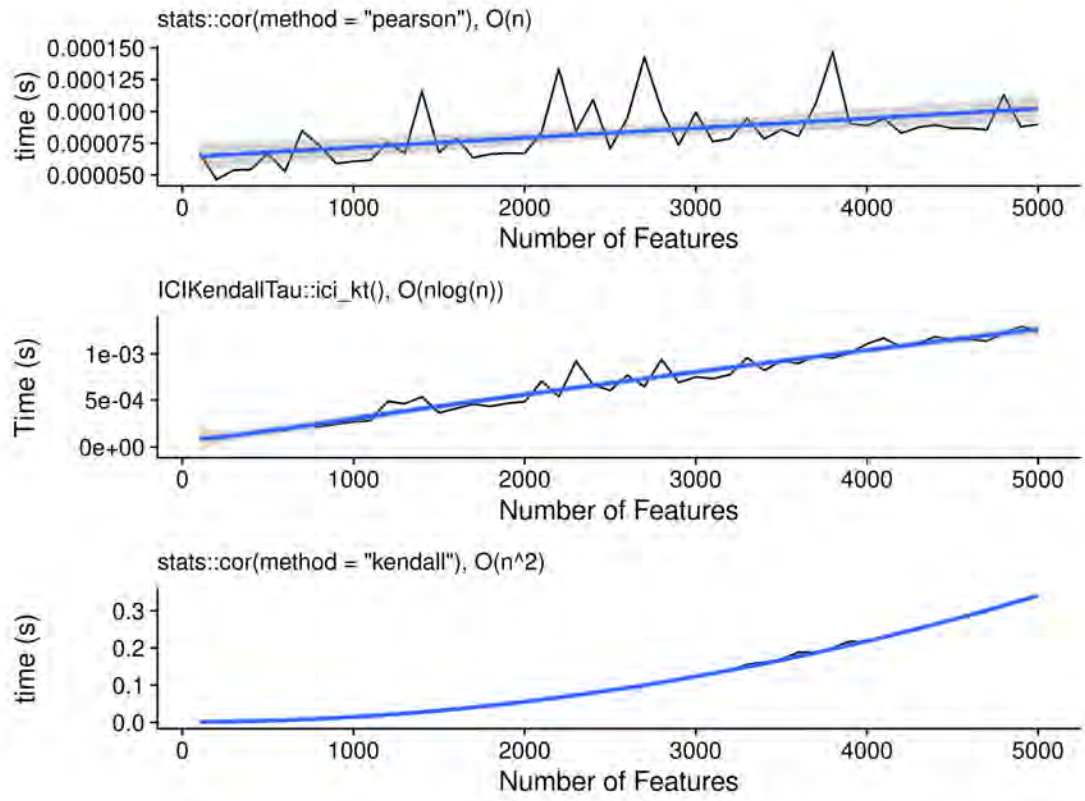

Figure S4. Time in seconds needed as a function of the number of features, with a fitted line for the assumed complexity for each of the methods tested, including R's Pearson correlation, the ICI-Kt mergesort, and R's Kendall-tau correlation algorithm.

Outlier Samples

Gierlinski Yeast

Gierliński et al [1] proposed a combination of median sample-sample correlation, gene count outlier fraction, and a chi-square test of read coverage per gene to score each biological replicate from a group of samples. Outside of defining this combined score, they do not describe the actual criteria of saying a replicate is "bad." In this work, we used **only** the sample-sample median correlation to identify "bad" or "outlier" samples, using the boxplot.stats function to determine outliers, where an outlier is defined as samples that are > 1.5x away from the limits of the box defining the distribution.

Here, we calculated Pearson correlation using the raw counts, and with only those genes that had non-zero reads in both samples. This version of Pearson correlation recreates the median sample-sample correlations observed in the original report.

Finally, we note the actual samples recorded as outliers by Gierlinski and coworkers, noted as "Manuscript."

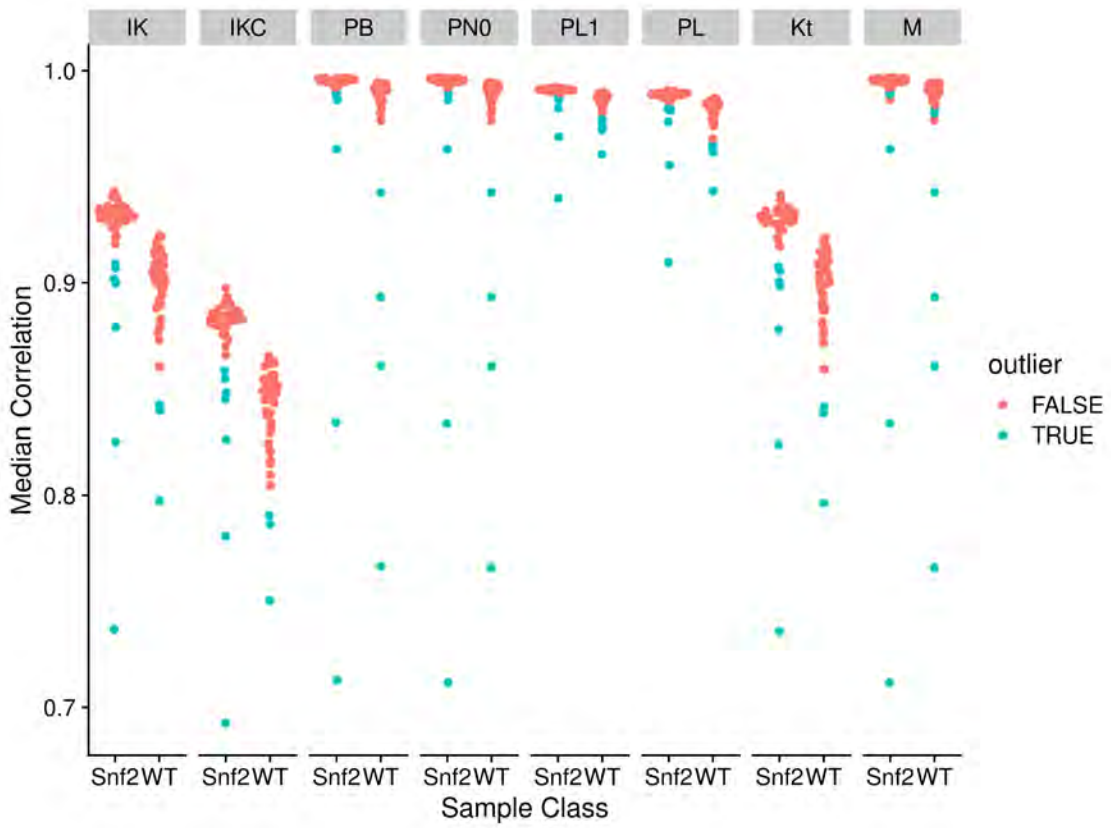

Figure S5. Median correlations for each of the yeast RNA-Seq sample to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; M: Manuscript; PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

In Figure S5 and Table S1 we can see how the determination of outliers was not made solely on the basis of correlation alone, but on a combination of factors that lead to some of the higher correlating samples (using raw counts and Pearson correlaion) being considered outliers where lower correlating samples were not listed as being outliers.

Regardless, using ICI-Kt or ICI-Kt * Completeness in this instance, the outliers using the simple distribution summary statistics and an outlier having to be > 1.5 median error, the outliers are mostly a superset of the outliers determined by Gierlinski et al, with the exception of three samples specific to their data: Snf2.24, WT.22, and WT.28.

It should be noted that it seems that Gierlinski et al used an eyeball cutoff for the outliers based on the combined score of correlation, outlier fraction, and read coverage chi-square fitness, with samples having a combined log-score > -2.8 (evaluated by RMF eyeball on a zoomed in graph and a ruler), a value that is never stated in the manuscript, and which seems arbitrary from the data.

Brainson RNASeq Samples

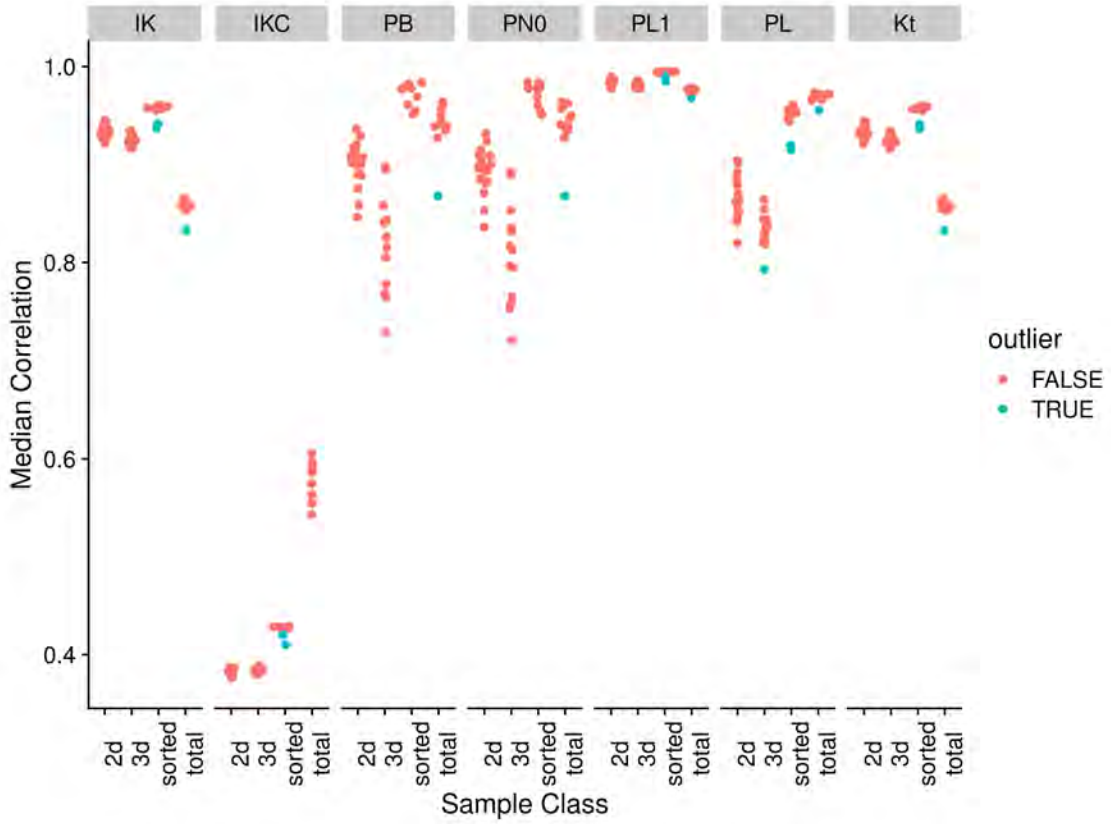

Figure S6. Median correlations for each of the Brainson RNA-Seq sample to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau.

Figure S6 shows the effect of including the pairwise completeness measure, where the "total" tumor samples have a higher correlation than the "sorted" samples where a specific subset of cells were collected from each sample and sequenced. This is due to having more genes with non-zero reads in general over the sorted samples.

Table S2 indicates the outliers determined with each correlation method. Discussions with the Brainson group have determined that some of the samples in the sorted group were run on different days with different technicians, which may explain the two sorted outliers.

TCGA Adenocarcinoma

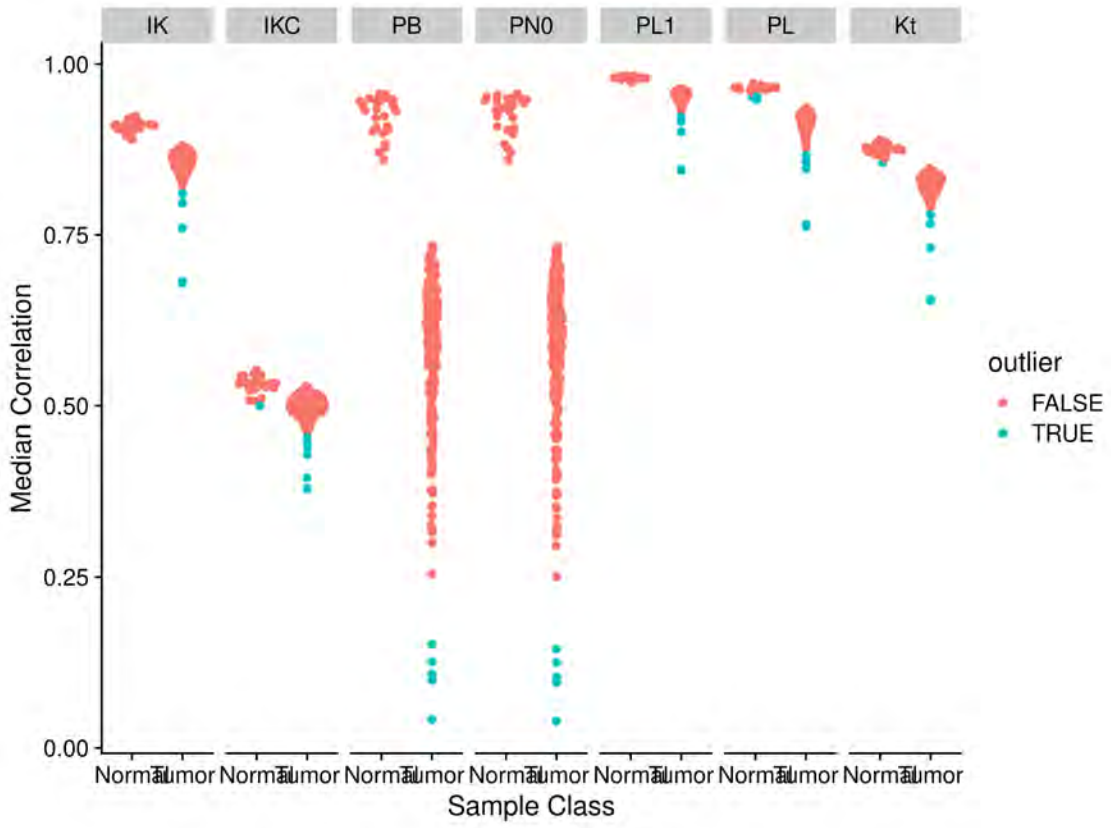

Figure S7. Median correlations for each of the TGCA adenocarcinoma RNA-Seq sample to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau.

Table S4. Adenocarcinoma median-absolute-deviations (MAD) for each group of sample median correlations in each type of tissue.

method	Normal	Tumor
IK	0.006	0.014
IKC	0.012	0.015
PB	0.022	0.092
PN0	0.021	0.091
PL1	0.002	0.007
PL	0.003	0.013

method	Normal	Tumor
Kt	0.006	0.013

As Figure S7 and Table S3, although there are not many outliers in these relatively large groups, Pearson correlation using log-transformed values results in the most outliers. Also of note is that ICI-Kt * Completeness is the only measure where the variance measure for both groups is similar. For both ICI-Kt and Pearson correlation, the tumor samples variance measures are x and x greater than normal, respectively.

Effect of Increasing Presence in Samples

One sure way to limit the effect of missing values on correlation values is to impose the condition that a gene has a non-zero count in a minimum percentage of the samples in one of the classes. In all of the example plots above, and the tables in the main manuscript text, a gene had to be present in at least **one sample**. Here, we explore what happens if we increase the number of samples required for a gene to be present in at 25% and 50% of a class for the yeast samples and adenocarcinoma samples, and 25%, 50%, 75% and all samples of a class for the Brainson samples. We also examine the effect using all possible correlation measures.

Yeast Samples

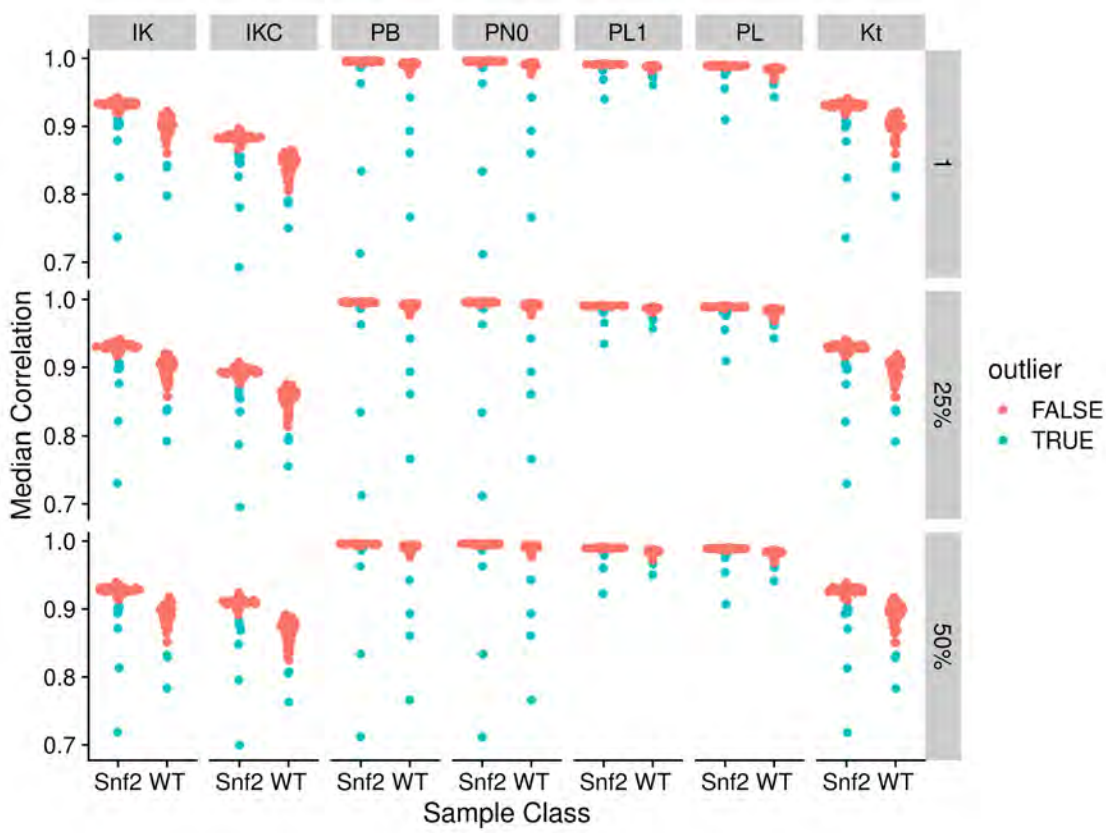

Figure S8. Median correlations by correlation method and applying different fractional cutoffs. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

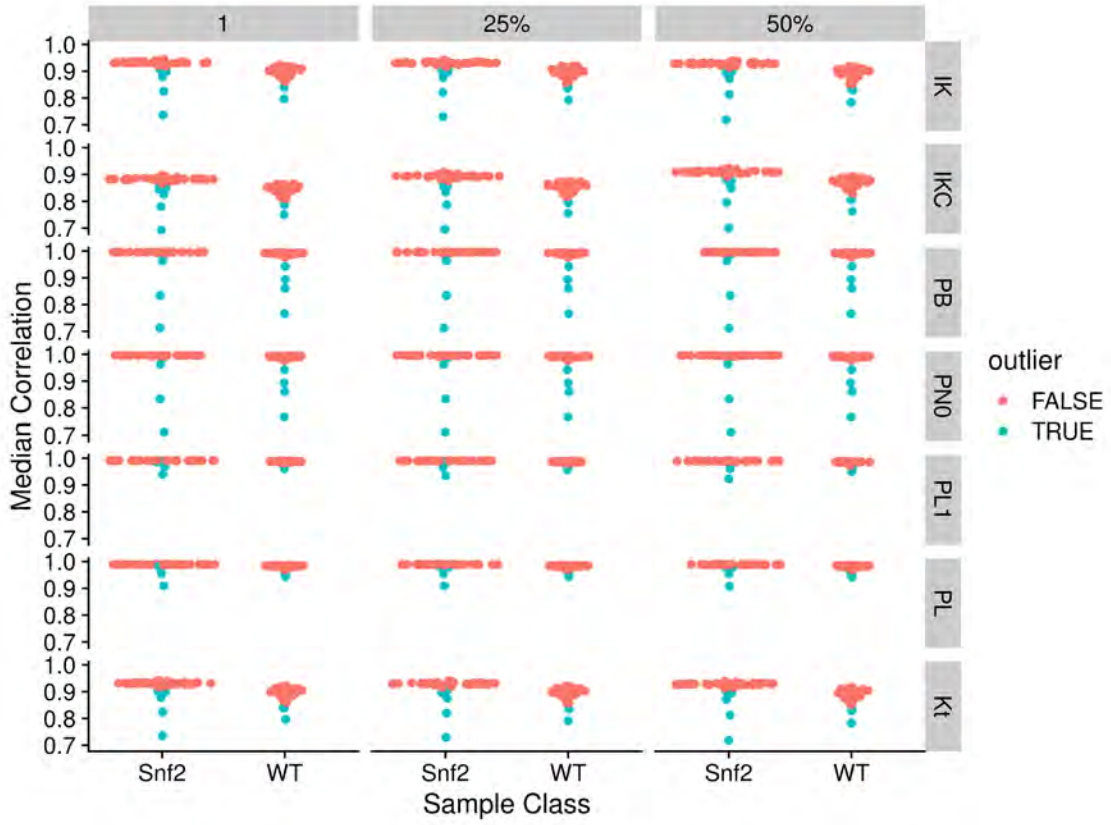

Figure S9. Median correlations by applying different fractional cutoffs and different correlation methods. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

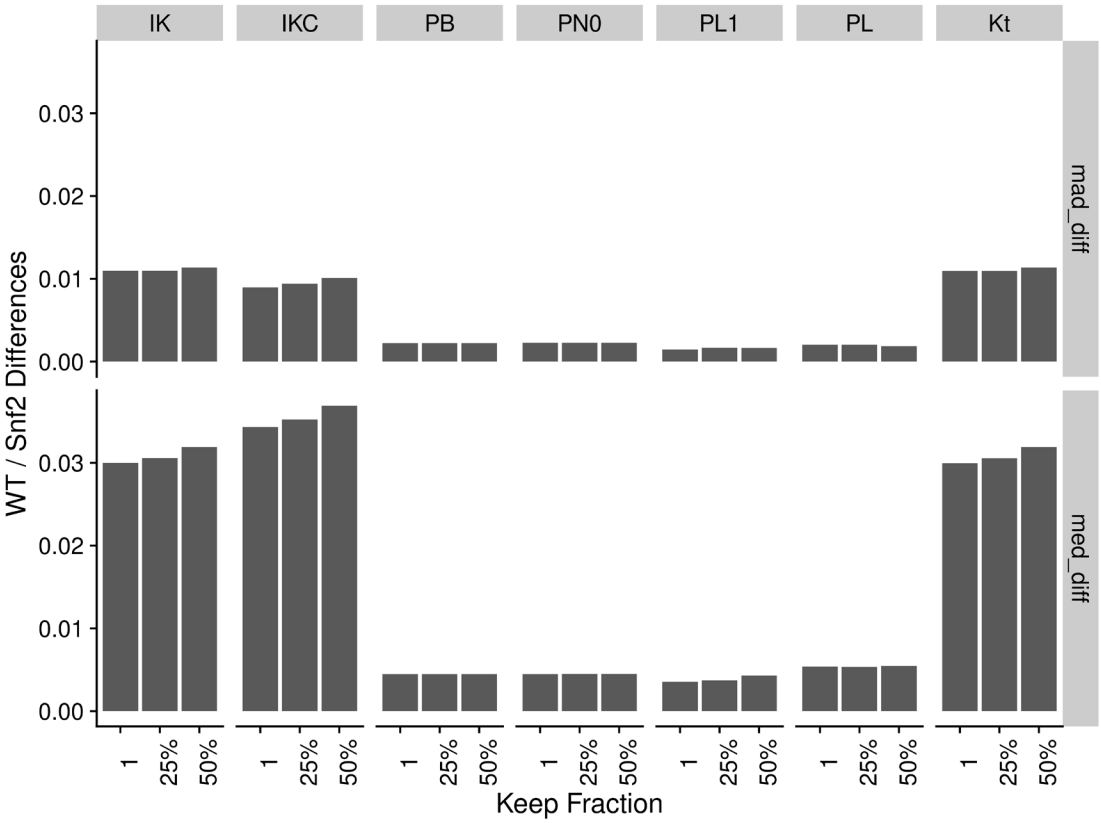

Figure S10. Difference between WT and Snf2 group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

Brainson Samples

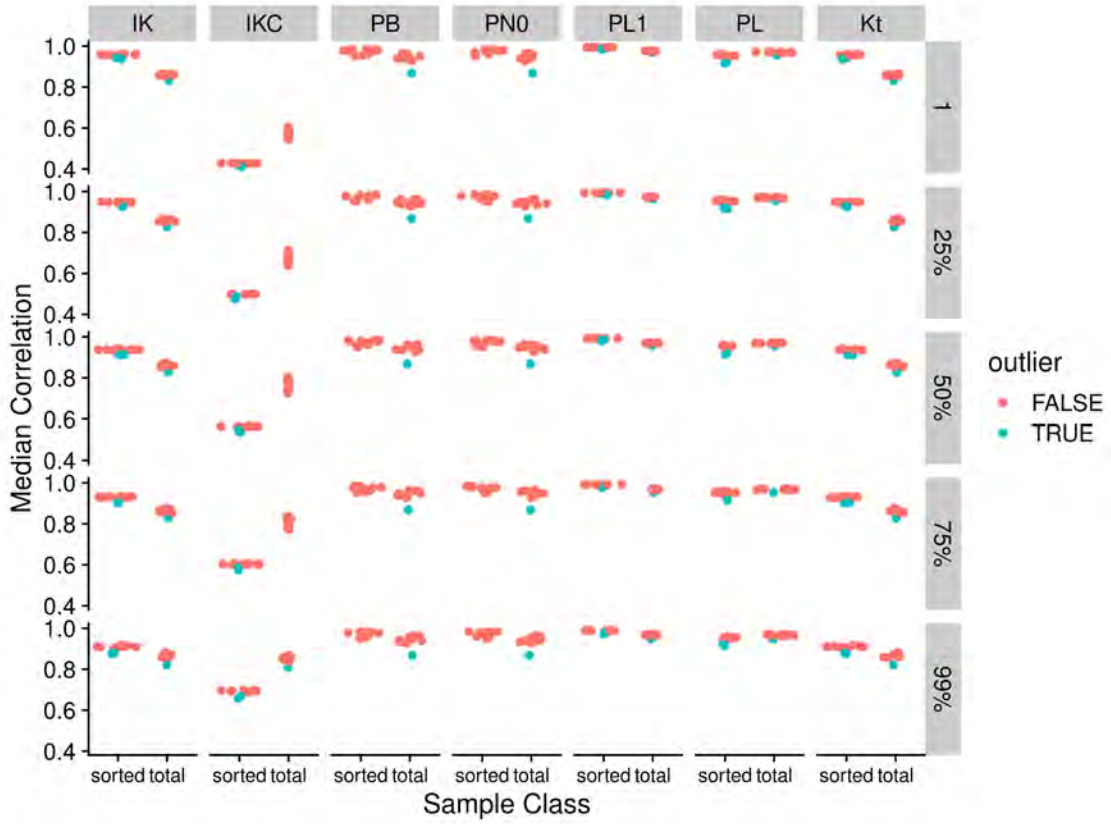

Figure S11. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

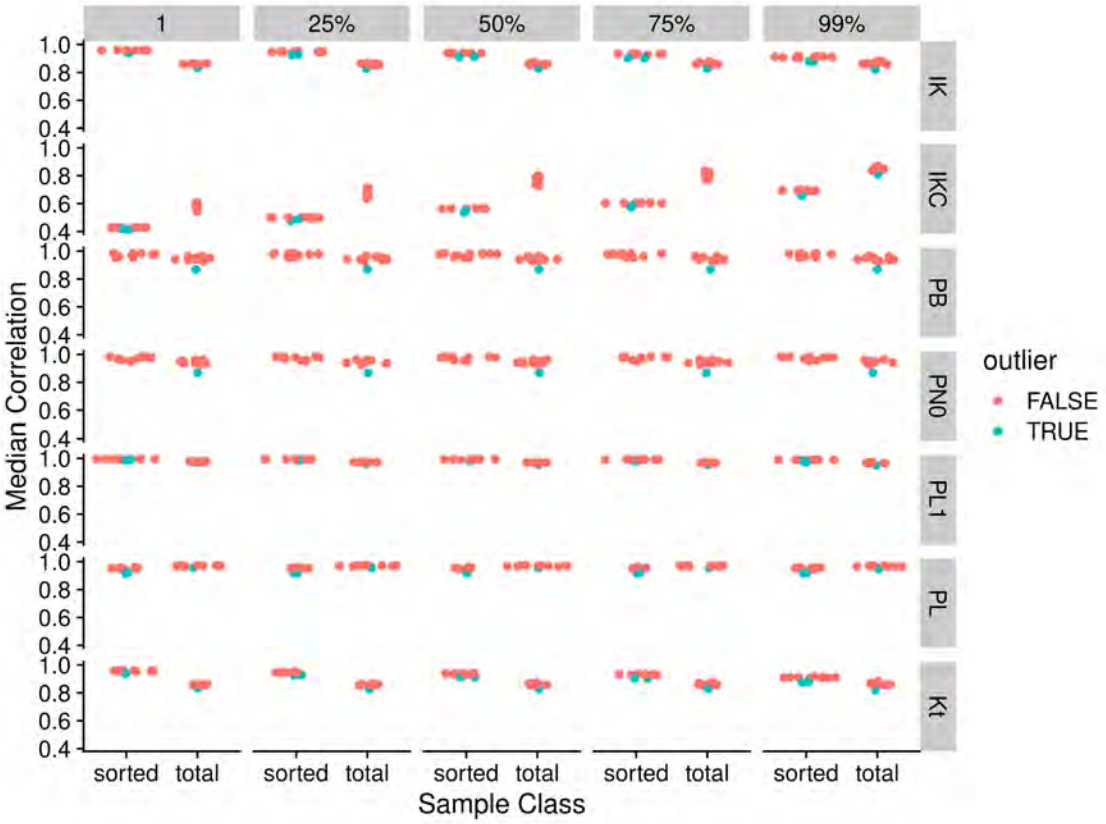

Figure S12. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

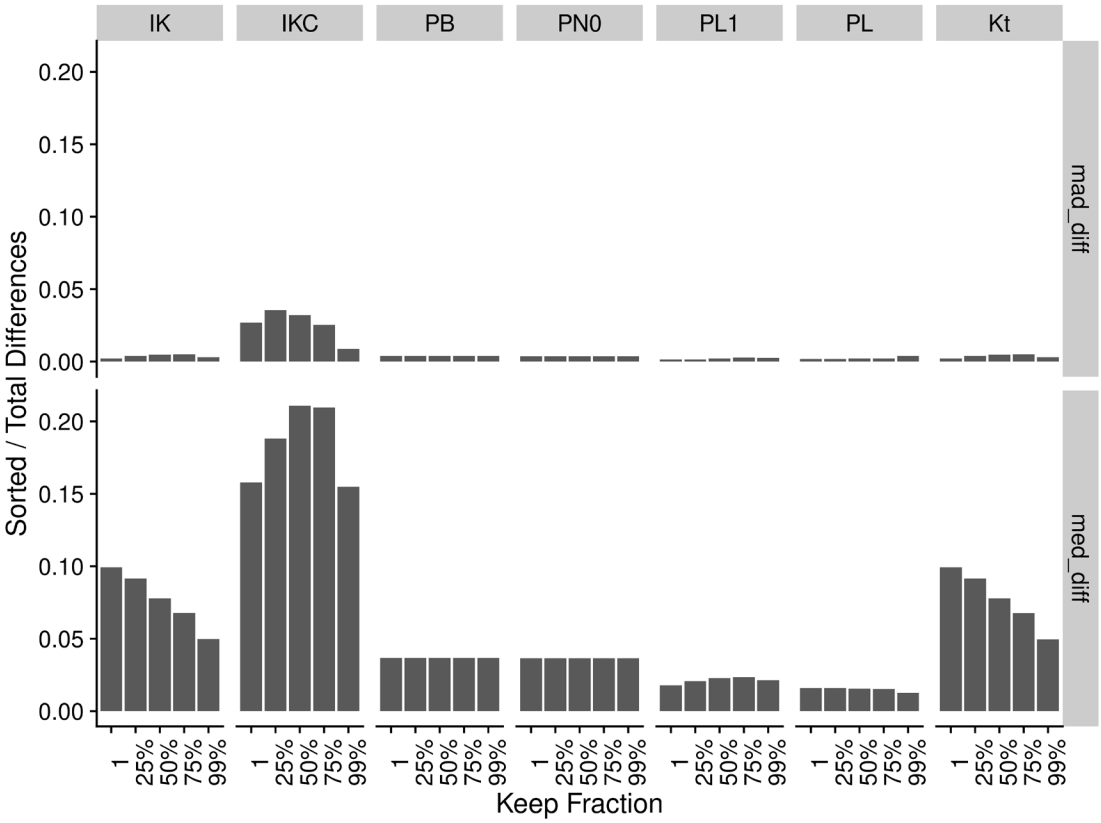

Figure S13. Difference between sorted and total sample group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

Adenocarcinoma

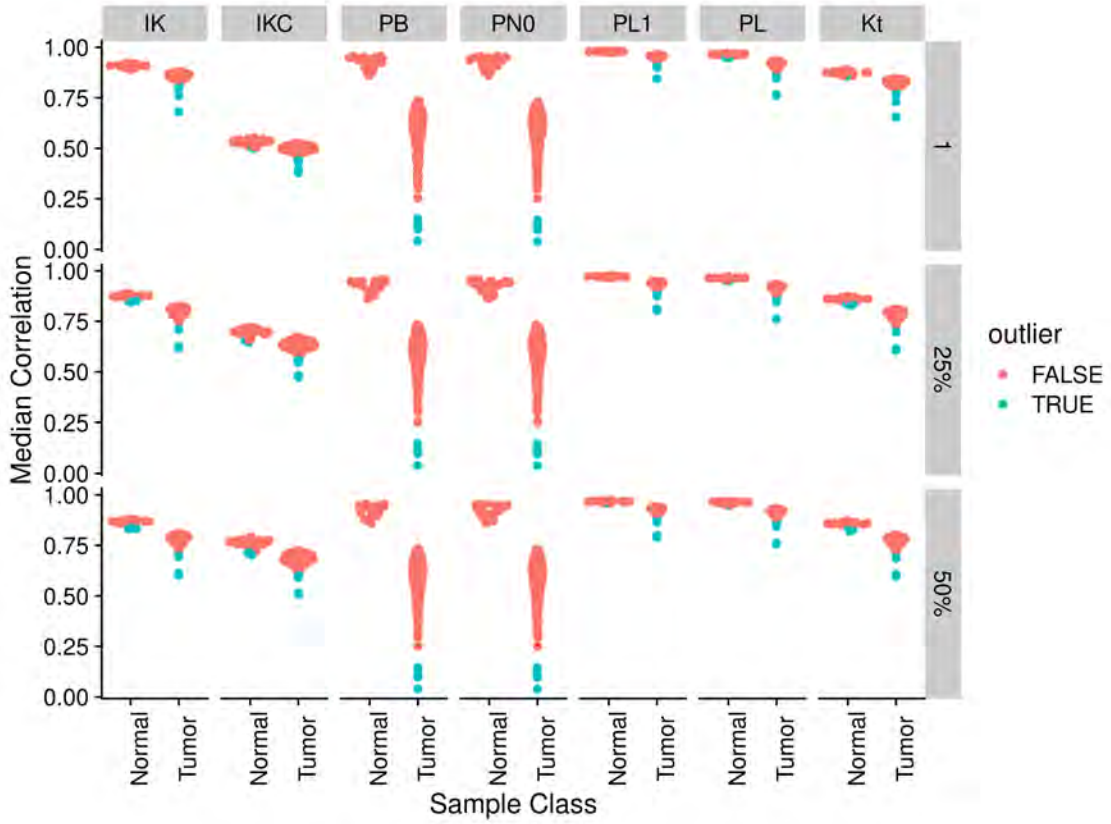

Figure S14. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

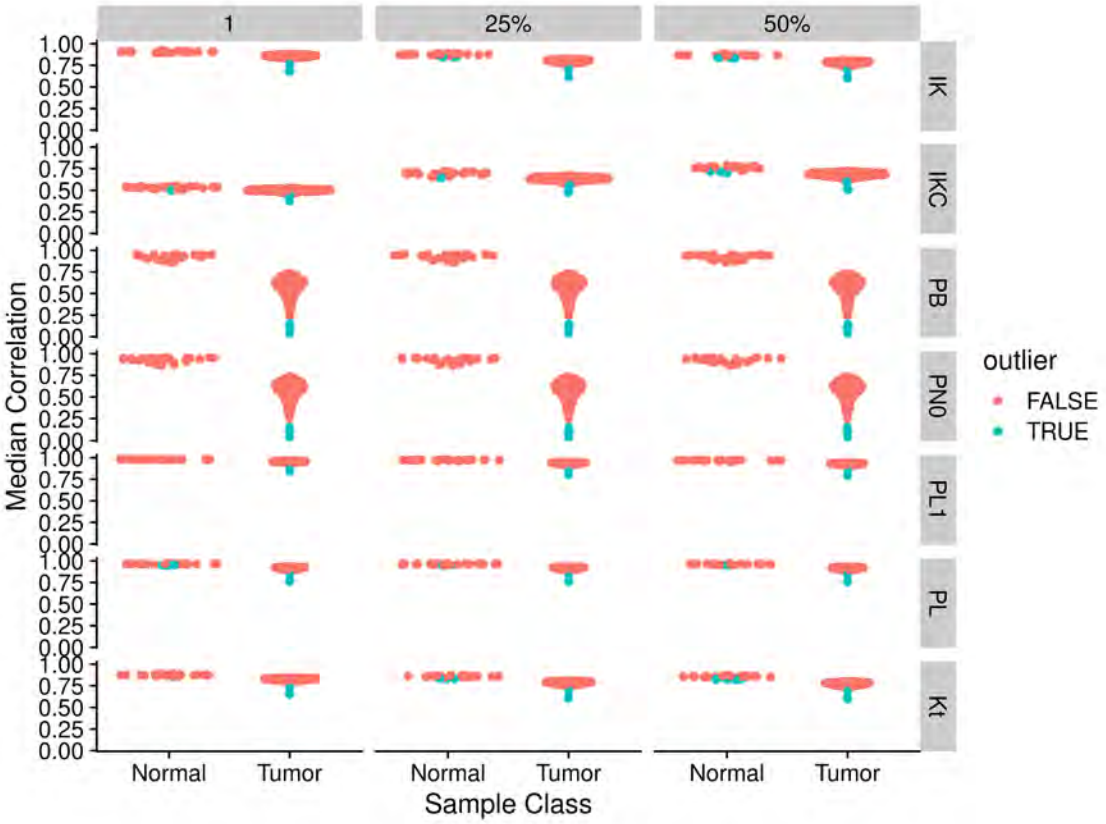

Figure S15. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

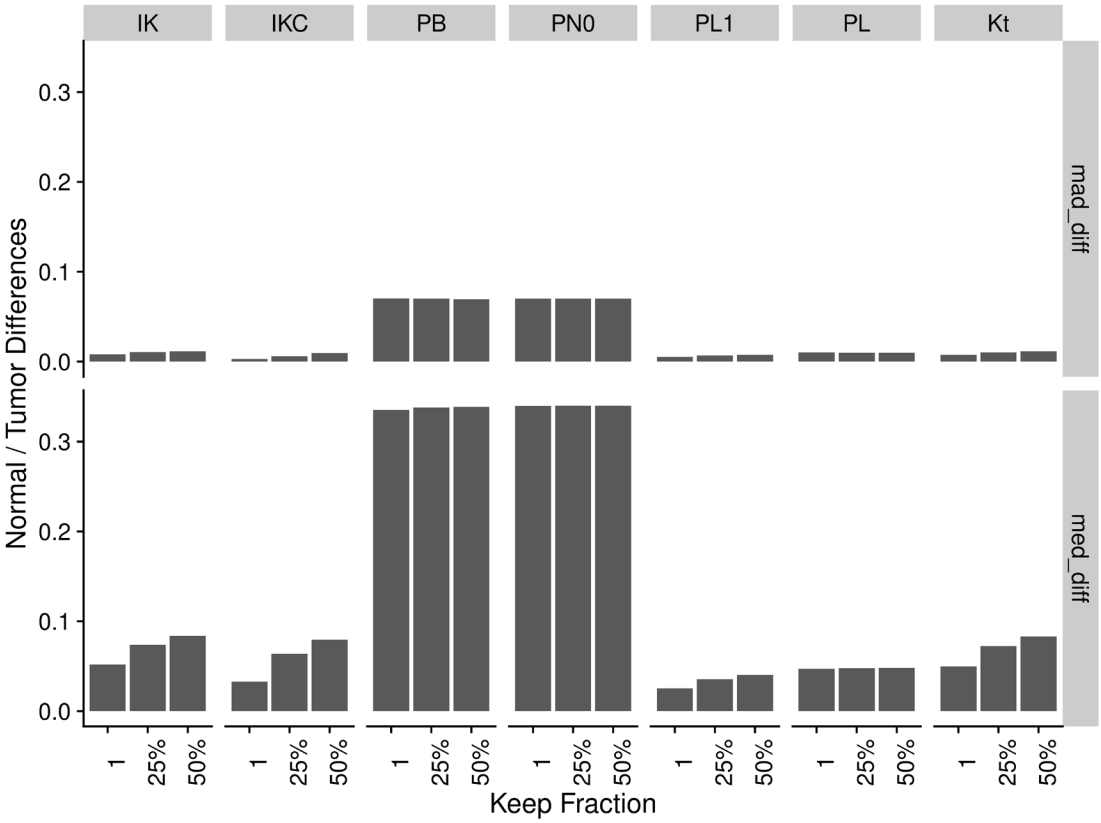

Figure S16. Difference of normal and tumor medians and MADs by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

Subsampling Correlation Differences

ICI-Kt

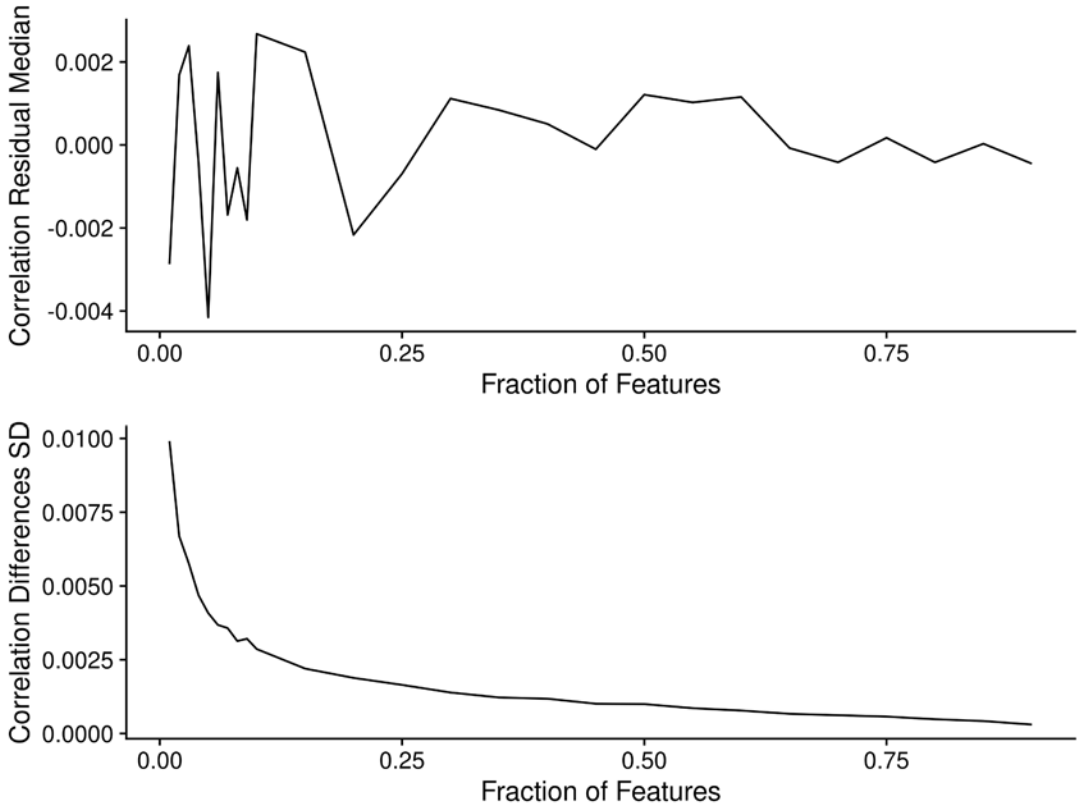

Figure S17. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample ICI-Kt correlations.

Pearson

Straight Pearson

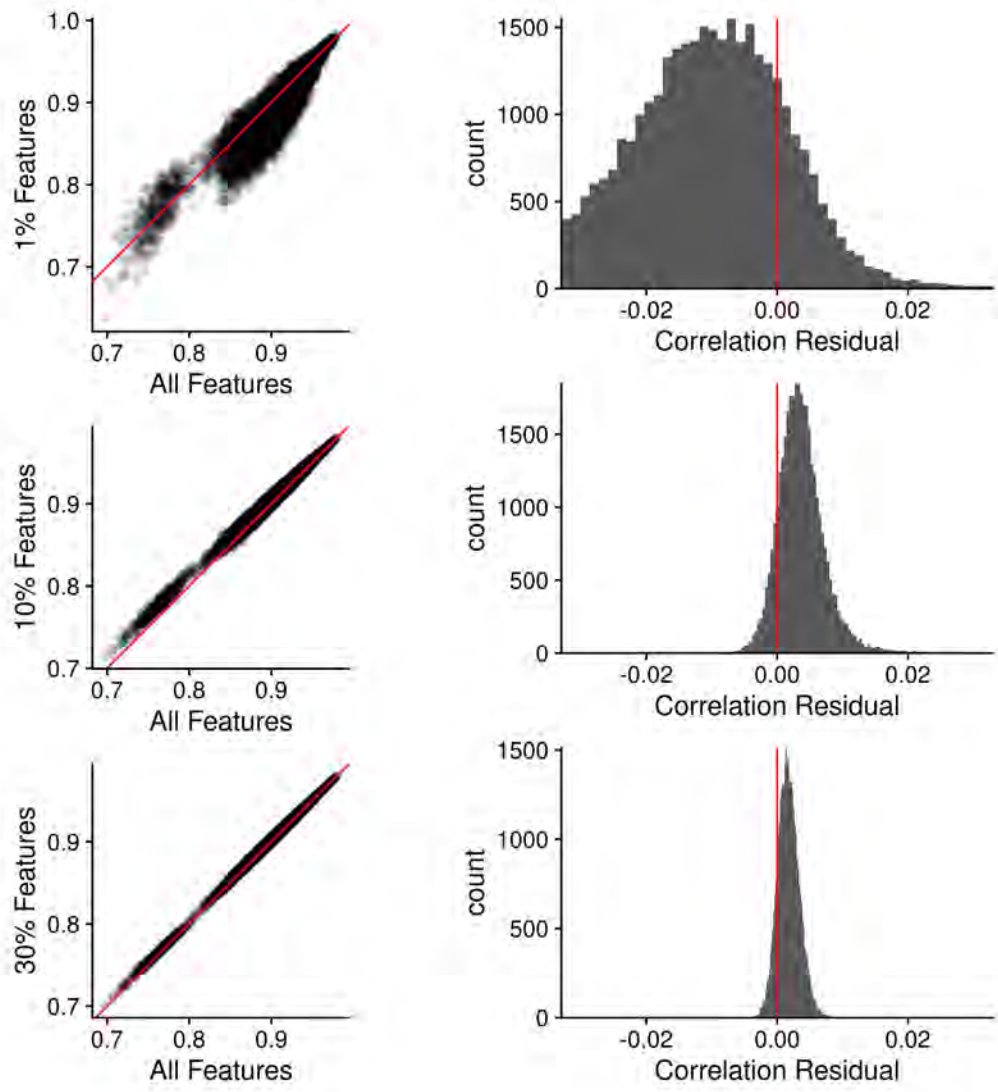

Figure S18. Sample-sample Pearson correlations using all features compared with 1%, 10% and 30% random subsets of features. Left: The sample-sample correlations using all features are plotted against the sample-sample correlations using a random subset. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (random subset - all features).

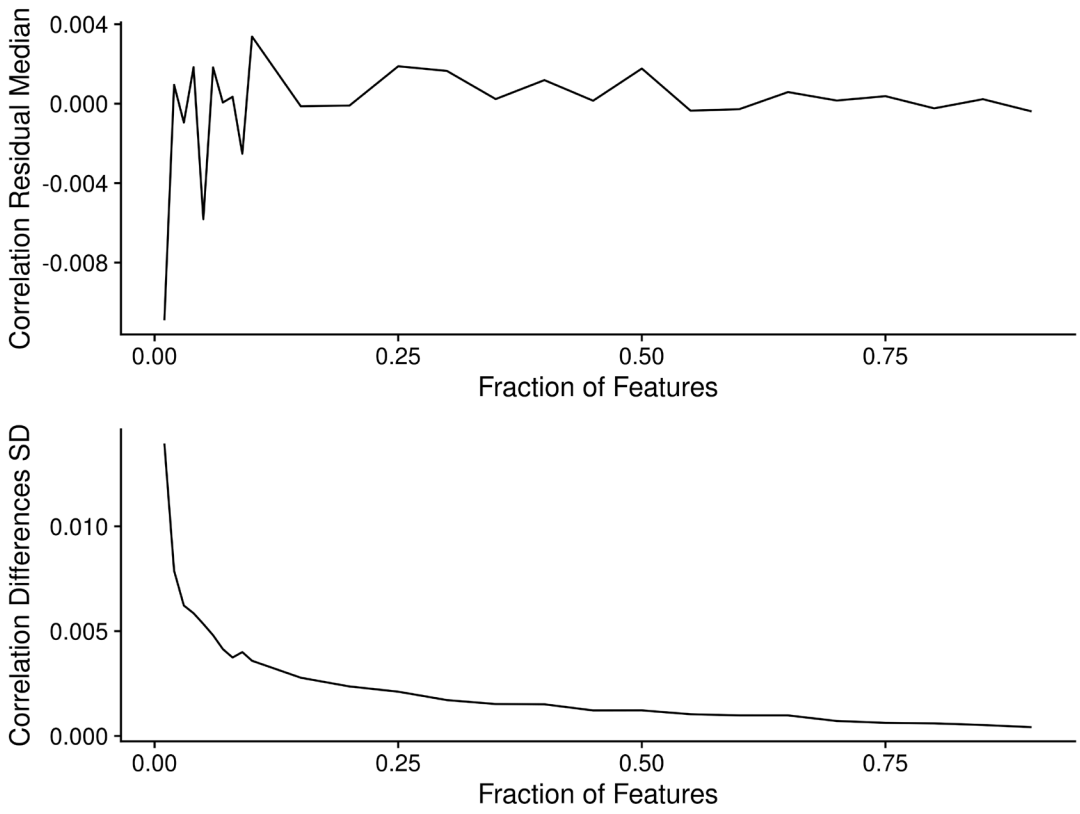

Figure S19. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample Pearson correlations.

Replacing Missing with Zero

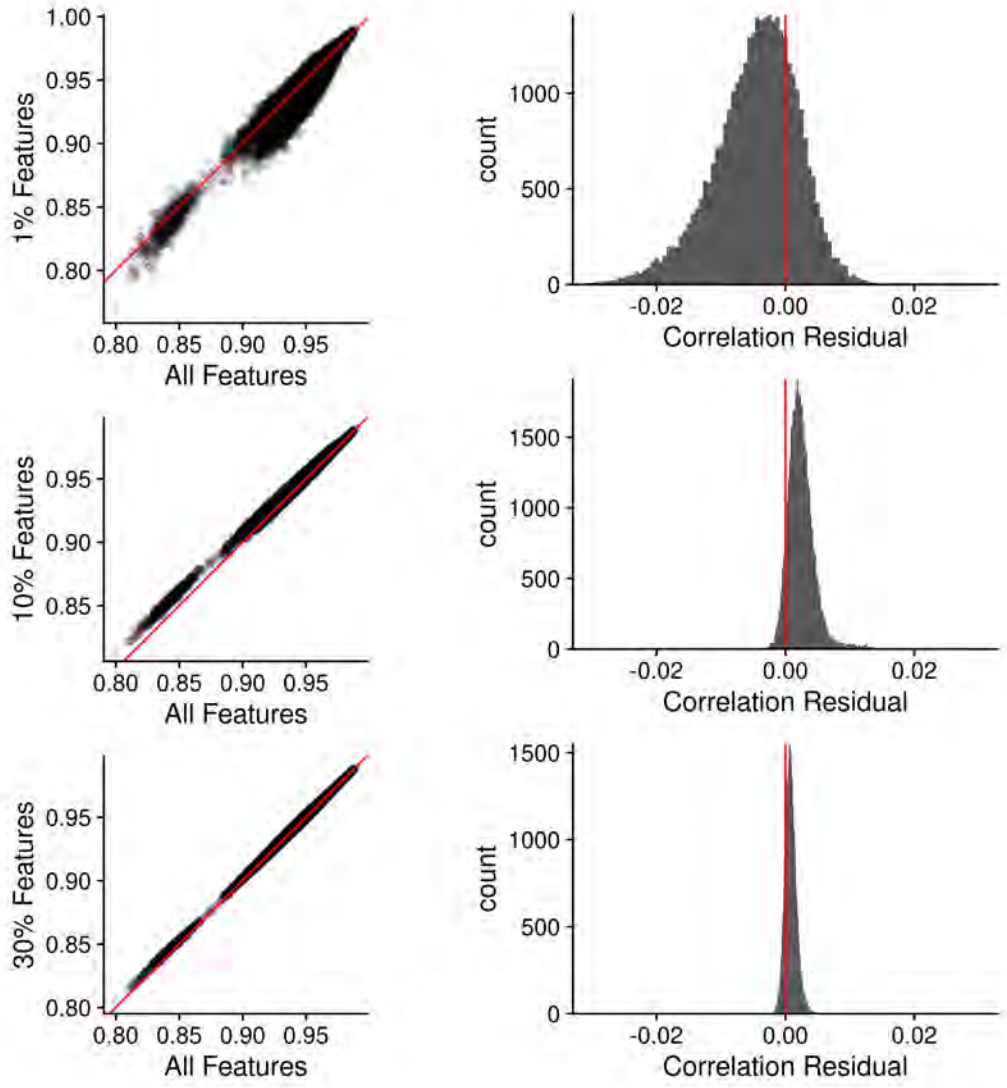

Figure S20. Sample-sample Pearson correlations using all features compared with 1%, 10% and 30% random subsets of features. Left: The sample-sample correlations using all features are plotted against the sample-sample correlations using a random subset. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (random subset - all features).

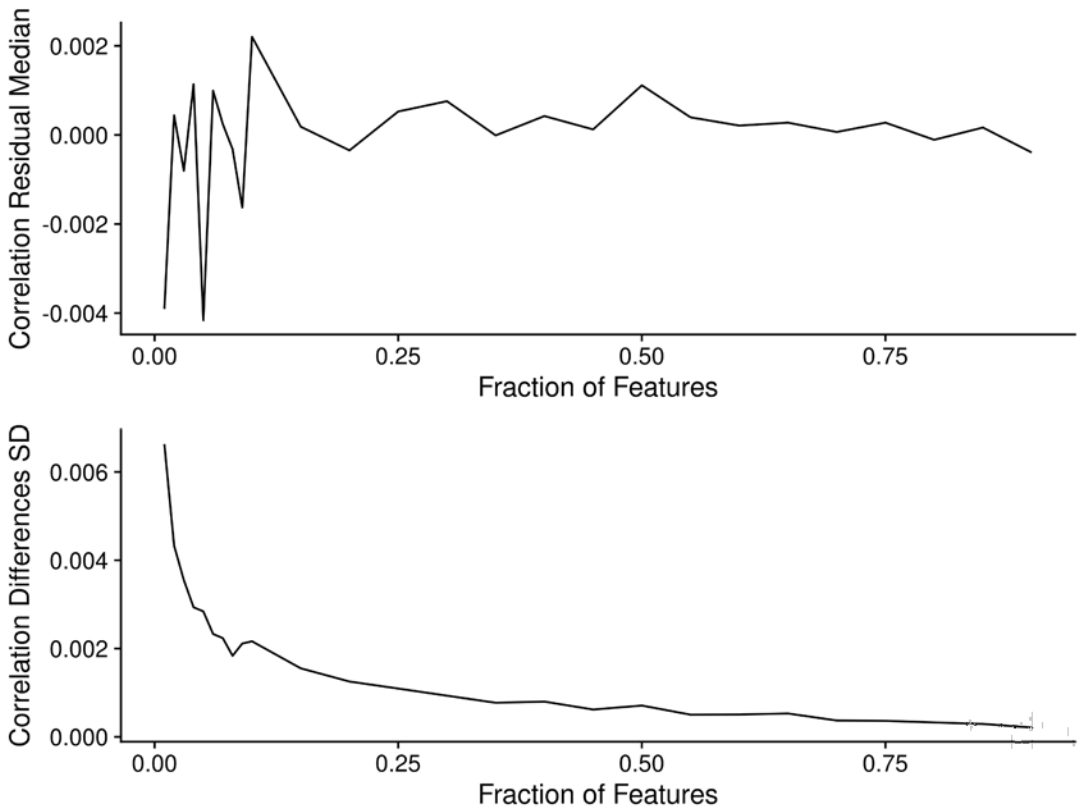

Figure S21. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample Pearson correlations.

Without Log-Transformation

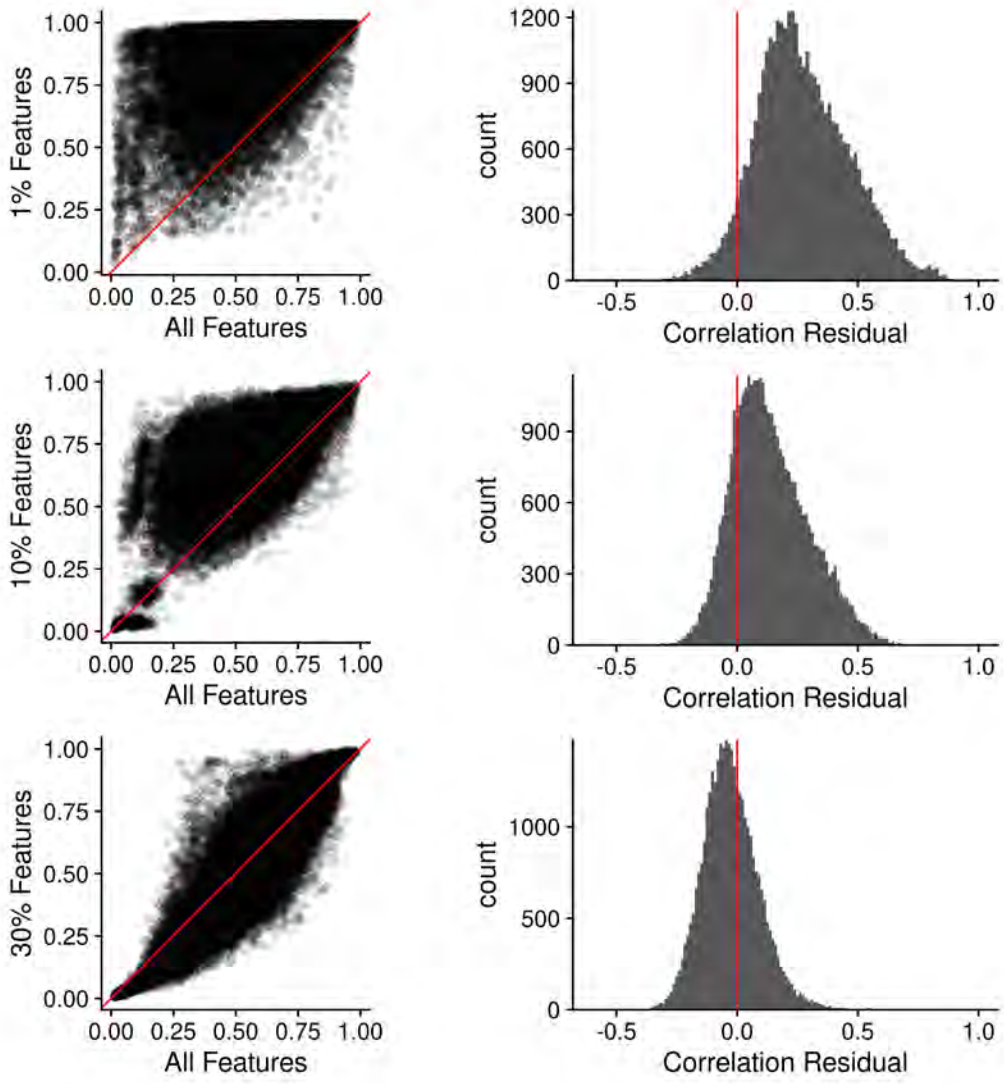

Figure S22. Sample-sample Pearson correlations using all features compared with 1%, 10% and 30% random subsets of features. Left: The sample-sample correlations using all features are plotted against the sample-sample correlations using a random subset. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (random subset - all features).

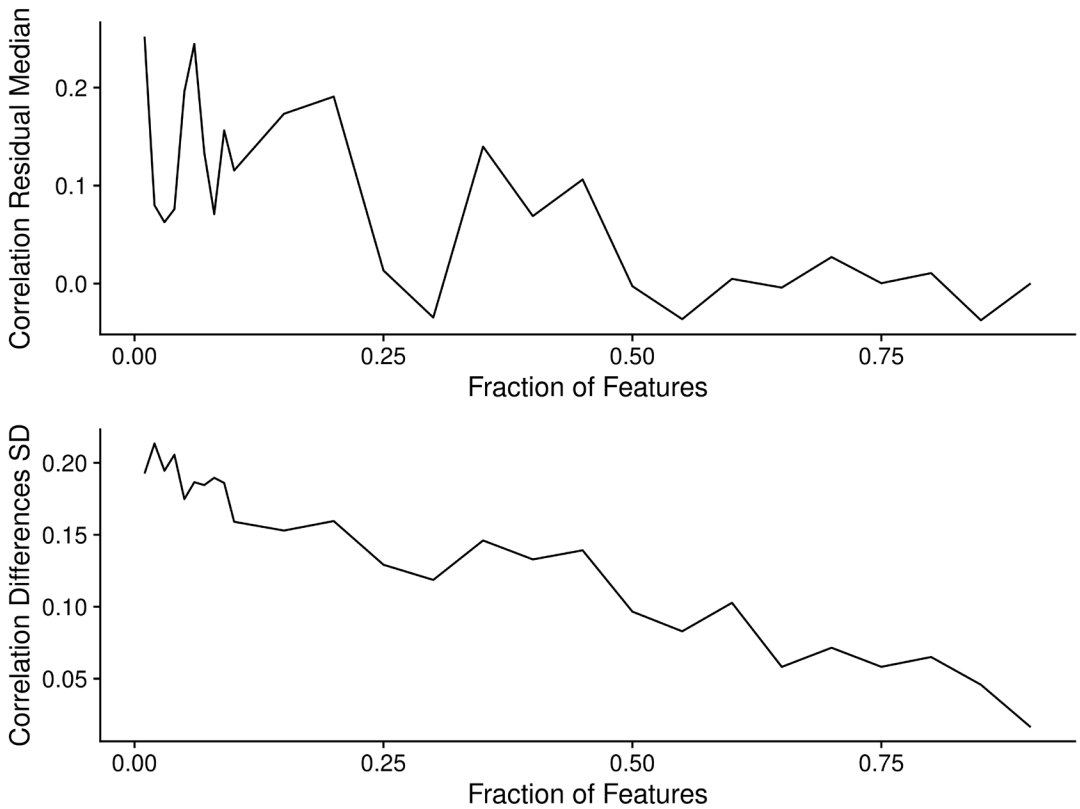

Figure S23. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample Pearson correlations.

Kendall

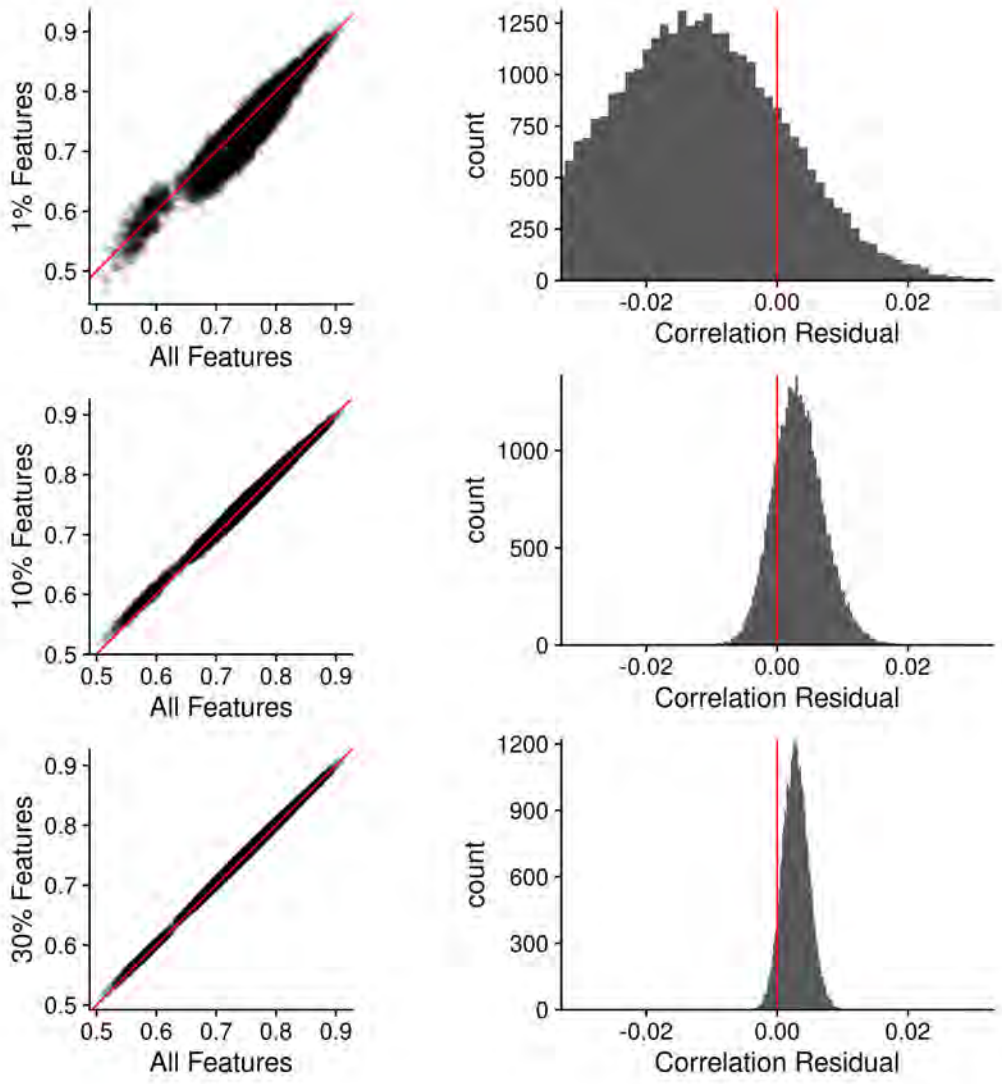

Figure S24. Sample-sample Kendall-tau correlations using all features compared with 1%, 10% and 30% random subsets of features. Left: The sample-sample correlations using all features are plotted against the sample-sample correlations using a random subset. Red line indicates perfect agreement. Right: Histogram of the residual differences of the sample-sample correlations, where the difference is the (random subset - all features).

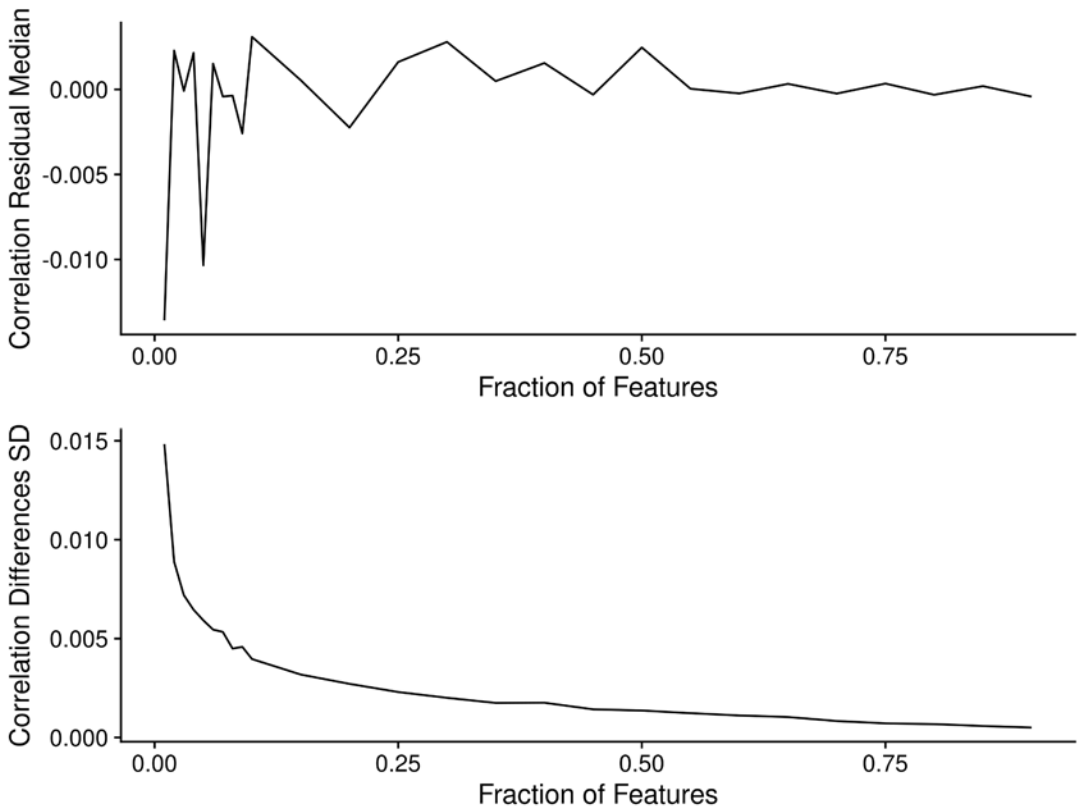

Figure S25. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample Kendall-tau correlations.

Comparisons

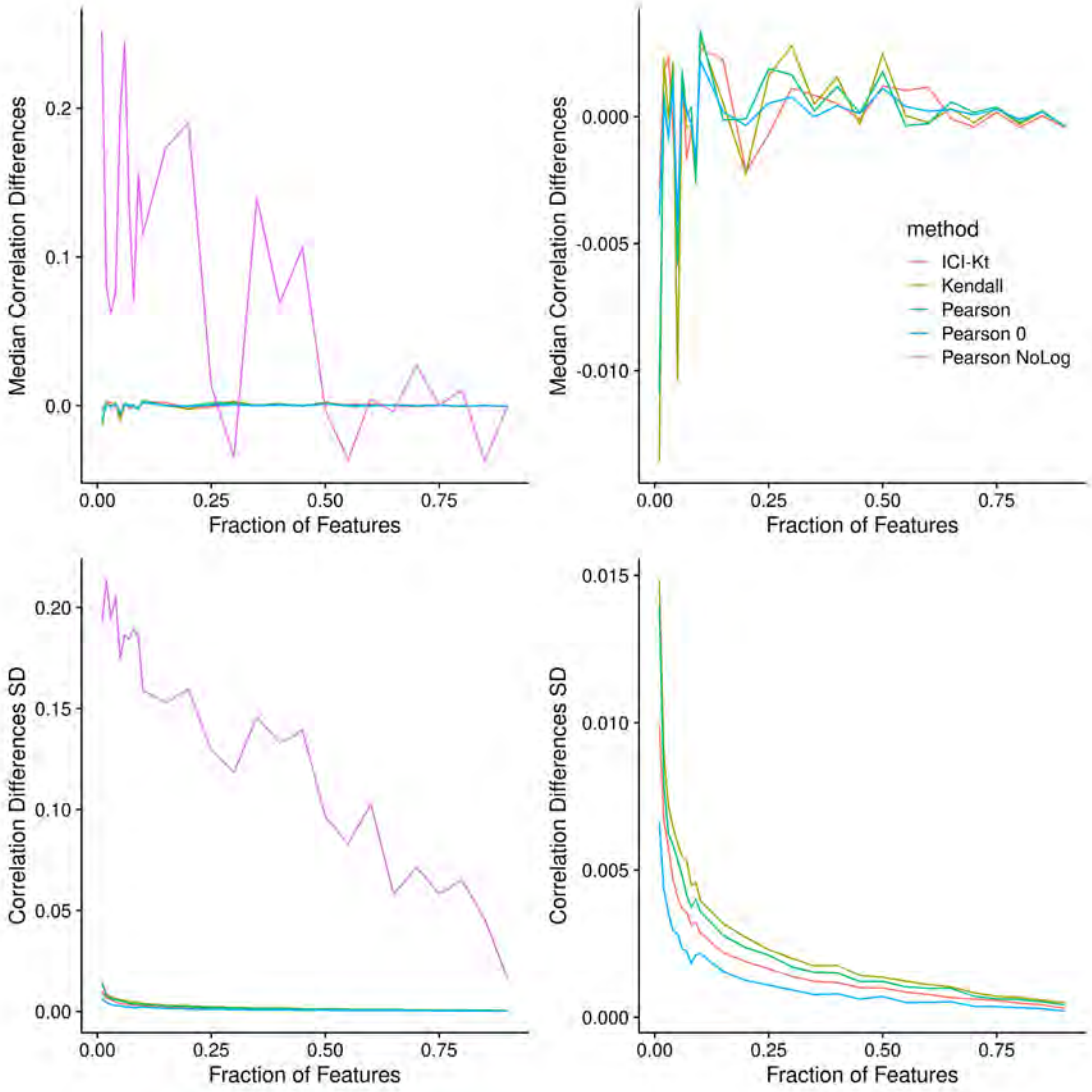

Figure S26. Median differences in correlation (top) and standard deviation of differences (bottom) as increasing fractions of features are used to calculate the sample-sample correlations.

References

1. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31: 3625–3630. doi:10.1093/bioinformatics/btv425

Supplemental Tables

Table S1. Yeast dataset median correlation values and outlier determination for each outlier from each of the correlation methods. Abbreviations for different measures are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau.

	IK		IKC		PB		PN0		PL1		PL		Kt		M	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
Snf2.31	X	0.90	X	0.85		0.99		0.99	X	0.99	X	0.98	X	0.90		0.99
Snf2.15	X	0.90	X	0.85		0.99		0.99	X	0.99	X	0.98	X	0.90		0.99
Snf2.24		0.93		0.88	X	0.99	X	0.99		0.99		0.99		0.92		0.99
Snf2.25	X	0.88	X	0.83	X	0.99	X	0.99	X	0.98	X	0.98	X	0.88	X	0.99
Snf2.10	X	0.91	X	0.85	X	0.99	X	0.99	X	0.99		0.98	X	0.91		0.99
Snf2.35	X	0.91	X	0.86	X	0.96	X	0.96	X	0.99	X	0.98	X	0.91	X	0.96
Snf2.13	X	0.83	X	0.78	X	0.83	X	0.83	X	0.97	X	0.96	X	0.82	X	0.83
Snf2.06	X	0.74	X	0.69	X	0.71	X	0.71	X	0.94	X	0.91	X	0.74	X	0.71
WT.34	X	0.84	X	0.79		0.98		0.98	X	0.97	X	0.96	X	0.84	X	0.98
WT.36		0.86		0.81		0.98		0.98	X	0.98		0.97		0.86	X	0.98
WT.22		0.88		0.82	X	0.94	X	0.94		0.98		0.97		0.88	X	0.94

	IK		IKC		PB		PN0		PL1		PL		Kt		M	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
WT.28		0.86		0.80	X	0.89	X	0.89	X	0.98	X	0.97		0.86	X	0.89
WT.25	X	0.84	X	0.79	X	0.86	X	0.86	X	0.97	X	0.96	X	0.84	X	0.86
WT.21	X	0.80	X	0.75	X	0.77	X	0.77	X	0.96	X	0.94	X	0.80	X	0.77

Table S2. Brainson dataset median correlation values and outlier determination for each outlier from each of the correlation methods. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau.

	IK		IKC		PB		PN0		PL1		PL		Kt	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
330-WT.3d		0.92		0.38		0.84		0.83		0.98	X	0.79		0.92
1004-het.sorted	X	0.94	X	0.42		0.95		0.95	X	0.99	X	0.92	X	0.94
783-het.sorted	X	0.94	X	0.41		0.98		0.98	X	0.98	X	0.92	X	0.94
1173Null_S12.total		0.85		0.59	X	0.87	X	0.87		0.97		0.97		0.85
1263wt_S1.total	X	0.83		0.56		0.93		0.93	X	0.97	X	0.96	X	0.83

Table S3. TGCA adenocarcinoma dataset median correlation values and outlier determination for each outlier from each of the correlation methods. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau.

	IK		IKC		PB		PN0		PL1		PL		Kt	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation

	IK		IKC		PB		PN0		PL1		PL		Kt	
Sample	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation	Outlier	Correlation
N.227		0.90		0.51		0.88		0.88		0.98	X	0.95		0.86
N.120		0.90		0.51		0.88		0.88		0.97	X	0.95		0.86
N.154		0.89		0.51		0.86		0.86		0.97	X	0.95		0.86
N.186		0.89	X	0.50		0.92		0.92		0.97	X	0.95	X	0.86
T.13		0.85		0.49	X	0.13	X	0.12		0.95		0.90		0.82
T.204		0.83	X	0.45		0.67		0.66		0.94		0.89		0.80
T.171	X	0.81		0.46		0.54		0.54	X	0.92	X	0.87	X	0.78
T.215	X	0.81	X	0.44		0.48		0.47	X	0.92	X	0.87	X	0.78
T.12	X	0.80	X	0.43	X	0.11	X	0.10	X	0.92	X	0.86	X	0.77
T.261	X	0.76	X	0.44	X	0.04	X	0.04	X	0.90	X	0.85	X	0.73
T.259	X	0.68	X	0.38	X	0.10	X	0.10	X	0.85	X	0.77	X	0.65
T.14	X	0.68	X	0.40	X	0.15	X	0.15	X	0.84	X	0.76	X	0.66