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,✉]

2024-01-24 14:25:16.256154

[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 directly 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 virtue of their "missingness". Therefore, we propose the 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, metabolomics, and lipidomics experiments, we demonstrate that the ICI-Kt allows for the inclusion of missing data values as interpretable information, enabling both improved determination of outlier samples and improved feature-feature network construction, without explicitly using imputation. Moreover, our implementation of ICI-Kt uses a mergesort-like algorithm that provides O(nlog(n)) computational performance, a significant improvement over the Kendall-tau correlation available in base R.

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, DNA methylation, etc.) 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]. 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].

Often, the first step in omics level analyses is to examine the sample-sample (dis)similarities in various ways using exploratory data analysis or EDA. This can include the examination of decomposition by principal components analysis (PCA), sample-sample pairwise distances, or sample-sample pairwise correlations to highlight biological and batch groups [10–12], double check the appropriateness of planned analyses [13], and check if any samples should be removed prior to statistical analysis (outlier detection and removal) [14]. Outlier detection, in particular, is often required for successful omics data analysis, as 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 [10–12,14,15].

All analytical methods, and in particular the analytical methods used in omics where many analytes are being measured simultaneously, suffer from missing measurements. Some analytes will be missing at random because of spurious 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. Also, 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. However, 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 are either ignored or imputed to zero (or another value) and then included in the correlation calculation. The two most common approaches for ignoring (i.e. dropping) values is to only use those measurements that are common across all samples (complete) or that are common between two samples being compared (pairwise complete). Both dropping or imputing missing values are likely to cause the calculated sample-sample correlation values to deviate from the real sample-

sample correlation values, especially with respect to specific data interpretation perspectives.

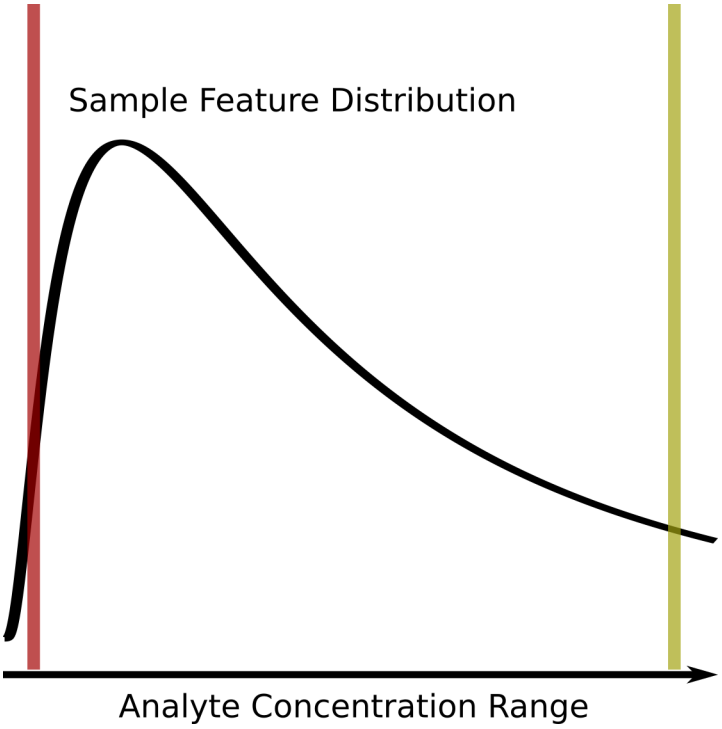

Figure 1. Graphical description of the left-censored data problem. An example density plot of the analyte concentrations for a single experimental sample is shown as a solid black curve. The true analyte concentration range covers the full range of the density distribution, with the minimum on the left (red vertical line), and the maximum on the right (yellow vertical line). Below certain concentrations, shown by the red line, the instrument returns either missing values (NA), zeros, or some other floored values, resulting in a left-censored distribution. Above certain concentrations, highlighted by the yellow line, typically the values returned will be identical (or flagged depending on the instrument). Which analytes will have concentrations below the red detection limit line 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.

To create a correlation measure that is capable of working with missing values, we would not be interested in creating a completely new correlation metric from scratch, but modifying an existing one. Of the three commonly used correlation measures, Pearson, Spearman, and Kendall-tau, Spearman and Kendall-tau seem most appropriate for modification as they solely use ranks in the calculation of their coefficients. Modifying Pearson would either involve imputing new values, or finding a way to calculate the covariances **with** missingness included. While Spearman uses ranks, many of the

modifications for handling identical ranks and ties do not seem amenable to working with missing values. In contrast, Kendall-tau's use of concordant and discordant pair counts seems most amenable to the creation of new definitions that incorporate missingness while still working within the original definition of the correlation coefficient, as shown below.

In this work, 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, information-content-informed Kendall-tau (ICI-Kt). We examine the effect of missing values on various simulated and real data-sets, comparing ICI-Kt with other simpler methods of handling the missing values, namely removing them or imputing them to zero. Given the detrimental effects of including outlier samples in differential analyses, we also evaluate the ability of ICI-Kt to capture sample-sample pairwise similarities and the determination of outlier samples prior to differential analyses. Finally, we examine how feature-feature networks derived from various feature-feature correlations change within annotations depending on the correlation method used.

All of the code used for this manuscript is available on zenodo [16].

Methods

Additional Definitions of Concordant and Discordant Pairs to Include Missingness

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 [17]:

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

We can expand the concordant and discordant pair definitions to include missing values (e.g. `NA` in R). Here $!\,x$ indicates a missing value and & is synonymous with "and". The information-content-informed concordant pair definitions are then:

- $x_i > x_j$ 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 then:

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

These additional definitions make it possible to calculate a Kendall-tau correlation that **incorporates missing values**, information-content-informed Kendall-tau (ICI-Kt).

Considering Ties

Tied values do not contribute to either of the concordant or discordant pair counts, and the original Kendall-tau formula for 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 [18,19].

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

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 correlation over multiple comparisons with the same set of features will consider the same number of pair comparisons. This is useful when analyzing and interpreting correlations from a large number of experimental samples, not just two samples.

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.

Completeness

As an addition to the correlation value, 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.

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

Implementation Details

We produced an initial reference implementation in *R* [20] 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). We note here that we **do not** consider this imputation of missing values.

After this initial implementation, we used the *SciPy* [21] *kendalltau* code as a model to create a mergesort based implementation in *Rcpp* that has a complexity of O(nlog(n)), in comparison to the pairwise testing that has a complexity of O(n^2) [22]. 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* [23–25]) 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 [26].

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 trivial [27]. 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 [28]. 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 EGFR Lung Tumor RNA-Seq Data Set

This RNA-Seq dataset is from null and knock-in EGFR genotype mice mutants. Genotypes include Null (no mutant-EGFR inducible expression), Heterozygous (only one copy of mutant-EGFR), and Homozygous (two copies of mutant EGFR). For each mutant, cells 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 Chen et al [29] for the full experimental details. Counts were normalized using *DESeq2* [30] with either the genotype, the growth environment, or the combination of them as the grouping factor. Samples were grouped by their genotype, growth environment, or their combination for the calculation of median abundance across samples.

Yeast RNA-Seq Data Set

These data were generated and reported as part of two publications evaluating replicate data and differential gene expression using wild-type (*WT*) and *Snf2* deletion mutant [14,31]. Summarized gene level counts were obtained from a GitHub project maintained by the Barton group [32]. It should be noted that the data are also available from two figshare repositories [33,34]. Counts were normalized using *DESeq2* [30] with the sample genotype of *WT* or *Snf2* deletion as the experimental factor. Samples were grouped by their genotype for the calculation of median abundance across samples.

The original outliers reported by Gierliński et al. [14] 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 (*pairwise-complete-observations*).

Adenocarcinoma Recount RNA-Seq Data Set

We downloaded the recount2 TCGA lung cancer data [35], 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. Counts were normalized using *DESeq2* [30] with the tissue status as normal or tumor as the experimental factor. Samples were grouped by the tissue status of normal or tumor for the calculation of median abundance across samples.

Non-Small Cell Lung Cancer Lipidomics Data Set

We previously analyzed a lipidomics data set from non-small-cell lung cancer (NSCLC) [5,36]. Because the assigned peaks represent such a small amount of the overall data, we reprocessed the full dataset to match peaks across samples. We first calculated the assigned peaks standard deviations in parts per million (PPM) to define an appropriate cutoff for matching peaks across samples. A value of 0.5 ppm looked to be wide enough to capture the variances in location of the assigned peaks. For each peak, an interval of 0.5 ppm to either side of the peak was defined. Starting with a random peak, all peaks with an overlapping interval to the starting peak were deemed to be the *same* peak across samples and aggregated. All of the overlapping peaks are removed from the sample peak list, and

then the next available peak chosen, and the process repeated until all peaks were accounted for. This process was repeated with four different randomizations of the full peak list to verify that the number of final aggregated peaks was not dependent on the starting peak list order. All iterations showed that the number of final aggregated peaks were within 100 of each other, where we result in close to 34,000 aggregated peaks. Peak intensities in each sample were normalized by the median abundance of all quantified peaks in a sample. For details on the samples, their processing, and generation of assignments, see Mitchell et al [5]. Samples were grouped by the instrument (two different instruments were used to collect the data) and whether they were taken from normal (nearby non-tumor) or tumor tissue for the calculation of median abundance across samples. The previously assigned peaks were matched to the un-assigned peaks by finding peaks within a 0.5 ppm cutoff. For lipid category and class voting, we only considered peaks matched that had a single assigned and a single un-assigned peak match together.

Rat Stamina and Feeding Metabolomics Data Set (Rat)

This metabolomics data was collected as part of a study on muscle selection of fuel source and oxidative capacity in outbred rats [37]. This data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench, https://www.metabolomicsworkbench.org where it has been assigned Project ID (PR000016). The data can be accessed directly via it's Project DOI: https://dx.doi.org/10.21228/M86P45 [38]. The mass-spectrometry peak intensities, and sample metadata were taken directly from the MWTAB json file provided by Metabolomics Workbench. Peak intensities were normalized by the non-zero peak abundances in each sample. Samples were grouped by the combination of feeding status (ad lib or calorie restricted) and oxidative capacity (low or high) for the calculation of median abundance across samples.

Feature Annotations

Pathway annotations for gene transcripts in the RNA-Seq datasets were obtained from Reactome pathways in the *reactome.db* Bioconductor package as well as organism specific annotation packages [39,40]. Metabolite annotations for the rat stamina metabolites are based on Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that map from the provided KEGG compound IDs to various KEGG pathways [41]. Compound-pathway annotations were fetched using the *kegg-pull* Python package [42]. Lipid features from the NSCLC lipidomics with assignments are classified into one or more lipid categories using our lipid classifier tool [43]. For each assigned peak, we used voting of the lipid categories and classes from multiple assignments to ascribe a single category and lipid class (where possible).

Number of Non-Missing Values and Median Abundance

For each dataset, the samples were split by experimental factor of interest (see previous Methods for each dataset). Within each experimental factor, for each feature, the median for present values (or non-zero) was calculated, as well as the number of samples the feature was observed in. Subsequently, for all the medians with the same number of

present values, the minimum median value was recorded. For correlations, the median of the $log10(median)$ across all present values was used to include data for Kendall-tau correlation. Additionally, the Kendall-tau correlation of the minimum median and number of present values was calculated.

Correlation Methods

For each dataset, we calculated correlations using a variety of methods. In each dataset, there were either zero values or NA values to represent missingness. To start, we replaced all missing values with 0, and then either set them to NA or left them as zero as appropriate. ICI-Kendall-tau with zeros replaced with NA (IK); and then scaled by the completeness metric (IKC). Kendall-tau, with zero replaced with NA, and then using pairwise-complete-observations (K). Pearson, with zeros, using pairwise-complete-observations (P). Pearson, with zeros replaced with NA, using pairwise-complete-observations (PN0). Pearson, with a $log(x + 1)$ transform applied, using pairwise-complete-observations (PL1). Pearson, with a $log(x)$ transform, and then setting infinite values to NA values, using pairwise-complete-observations (PL).

Outlier Detection

For outlier detection, median sample-sample correlations within the sample class (genotype, condition, etc) is calculated, a $log(cor_{median})$ calculated to transform it into a score, and then outliers determined using *grDevices::boxplot.stats*, which by default are at 1.5X the whiskers in a box-and-whisker plot. As we are interested in only those correlations at the *low* end of correlation (becoming the high end after a log-transform), we restrict to only those entries at the *high* end of the score distribution (using *visualizationQualityControl::determine_outliers* [44]).

Sample-sample correlations are calculated with different sets of features. For each feature, we calculate the number and fraction of samples in a sample class that feature had non-zero or non-missing values in, making it possible to filter to a subset of features that are present in a minimum required number of samples.

For each cutoff of non-missing values, we also calculate the median and median-absolute-deviation (MAD) of the sample median correlations, and then the differences of median and MAD between two classes of samples: SNF2 and WT for yeast; normal and tumor for adenocarcinoma; wt and null for "egfrgenotype"; tumor and normal for one instrument for NSCLC; Ad lib high and low for "ratstamina".

Feature-Feature Networks and Partitioning

For each of the datasets, we trimmed to features present in 25% or more of **any** of the sample classes (see above for the various classes of samples in each dataset). Various correlation measures are calculated between all remaining features (Correlation Methods). For any given correlation, we generate the feature-feature network for that dataset-correlation combination. The dataset correlations are transformed to partial correlation. From the distribution of partial correlation values, we consider the fraction of values that make up the 2.5 % of the tail values (for a total of 5%) as the **significant** partial

correlations that can be used as actual edges in the network. The network is then trimmed to only the edges that have a positive weight. For each feature annotation (see Feature Annotations), we calculate three sums of the edge weights.

1. The total sum of edge weights for all edges with features that are annotated to one or more of the annotations (*annotated*).
2. The *within* annotation edge weight sum, where both start and end nodes are annotated to the same annotation.
3. The *outer* annotation edge weight sum, where the start node is part of the annotated set, and the end node is annotated to one of the other annotations.

The partitioning ratio (or q-ratio) is calculated as follows:

$$Q = \sum_{i=1}^{annot} \frac{within_i}{annotated} - \left(\frac{outer_i}{annotated}\right)^2$$

The partitioning ratio was originally designed as a method to determine the optimal clustering of networks, where each member of the network has only a single label [45–48]. In those cases, the partitioning ratio should range between 0 and 1 for non-partitioned and fully partitioned networks, respectively. None of the annotation sources we use have single labels for any features, and therefore the partitioning ratios have a much wider range. However, we expect that **better** partitioning of the network will be reflected by **more positive** partitioning ratios.

Computing Environment

Most calculations were run on a virtual machine running Ubuntu 18.04.6 LTS, 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 Intel Xeon CPUs (E7-4820 v4 @ 2.00GHz) with hyperthreading, 3TB of RAM, an 8TB solid-state-drive array, and a 100Gbps Mellanox ConnectX-4 adapter, provided by the Kentucky Research Informatics Cloud (KyRIC). KyRIC manages the virtual machines via an OpenStack instance. We used the *targets* package to manage calculation dependencies [49]. For the comparisons of time taken using different numbers of samples to evaluate the algorithmic complexity, calculations were run on a single laptop Intel i5-10210U core clocked at 1.6 GHz.

Results

Datasets

In Table 1, we provide a summary of the number of features (measurements), samples, treatments or conditions, and the number of biological replicates per condition for each of the experimental datasets.

Table 1. Number of features, samples, treatments, and biological replicates in each condition for each dataset.

Dataset	Measurement	Features	Samples	Conditions	Replicates
yeast	RNA-Seq	6,887	96	2	48, 48
adenocarcinoma	RNA-Seq	57,655	264	2	28, 236
egfrgenotype	RNA-Seq	30,875	56	3	18, 20, 18
egfrgenotypetumorculture	RNA-Seq	30,875	56	4	20, 14, 11, 11
typeandtumorculture	RNA-Seq	30,875	56	12	6, 4, 5, 3, 7, 5, 3, 5, 7, 5, 3, 3
ratstamina	Metabolomics	532	42	4	9, 12, 12, 9
nsclc	Lipidomics	33,986	146	4	38, 39, 35, 34

Limit of Detection As a Cause for Missingness

We are aware of only one previous investigation of the causes of missingness in metabolomics datasets [45]. In Do et al. [45], the authors showed that there was a limit of detection (LOD) effect, with a dependence on the day the samples were run. However, one thing that seemed to be missing in the analysis was normalizing each sample for sample-to-sample variation. We feel that this is critical, given that even with measurements acquired by Metabolon (the company that performed the measurements in Do et al.), the standards do not fully correct for sample-to-sample variation. Unfortunately, the KORA4 metabolomics dataset from Do et al. is not publicly available, so we could not attempt to redo their analysis of missing values with the same dataset.

For each experimental group of samples in each of the datasets we investigated (three RNA-seq, one metabolomics, one lipidomics), we calculated the median abundance and number of measurement values present across samples (i.e. N-Present) for each feature (gene or metabolite, see Methods). As shown in Figure 2A for the yeast RNA-seq dataset (also see Figures S1-S6 in Supplemental Materials for the other datasets), there is a monotonic relationship between the median abundance and the number of measurement values present. Moreover, as N-Present increases, there is clearly a minimum median value below which the values do not cross as illustrated in Figure 2B (see also Figures S1-S6 in Supplemental Materials for the other datasets). Given this relationship of the minimum median value observed and N-Present, we believe that the **majority** of missing values in many -omics datasets are due to being below the LOD.

This makes the ICI-Kt appropriate for use in many -omics datasets by incorporating missing values due to being below the LOD as useful information in the correlation calculation.

The LOD figures for alternative groupings of the EGFR genotype samples also demonstrate another interesting property (Figures S2-S4), in that incorrect groupings of RNA-Seq samples at the normalization stage result in negatively correlating relationships of the

number of present values and the lowest observed median value as illustrated in Figure S3B.

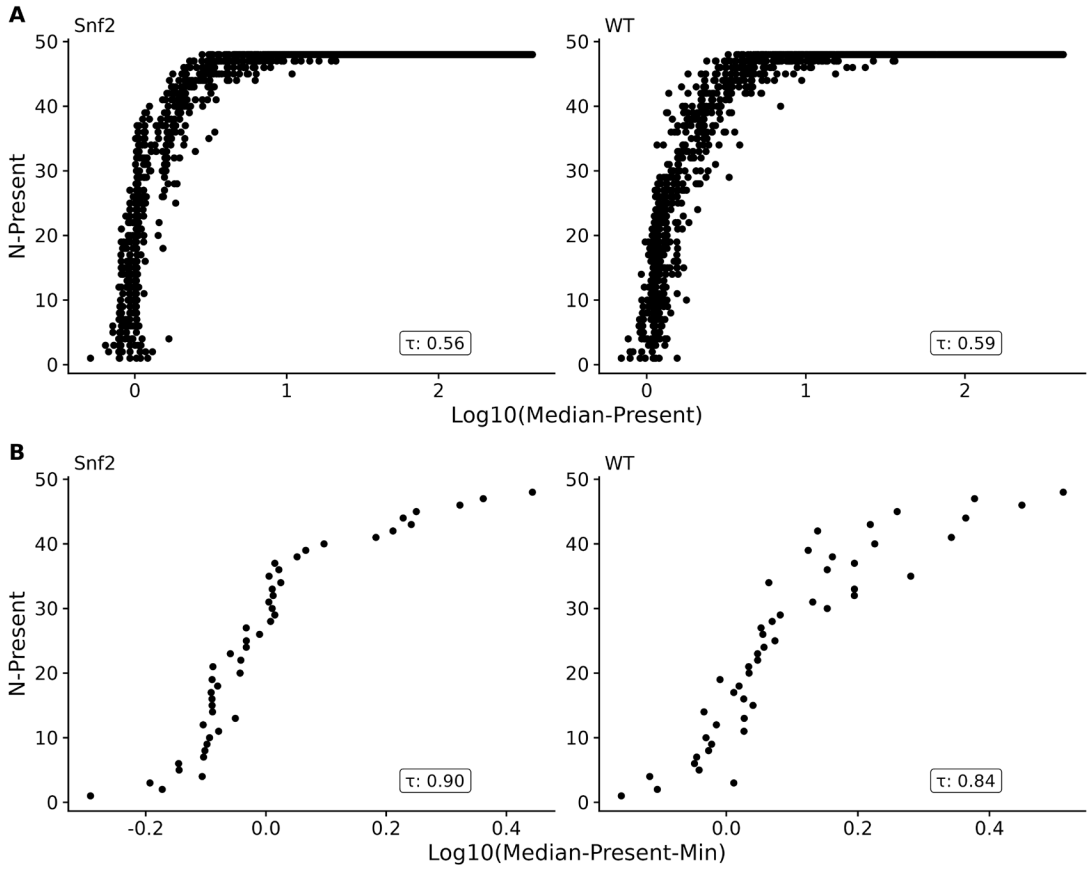

Figure 2. Yeast dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

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 S7-S9).

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).

In Figure 3, we can see that as missing values are added, only the ICI-Kt correlation values change in any significant way as illustrated by the wider range of ICI-Kt values on the y-axes versus the much narrower range of Pearson and Kendall tau correlation values on the x-axes. 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 x-axis scales compared to the y-axis), on the order of only 0.005 for both cases. 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).

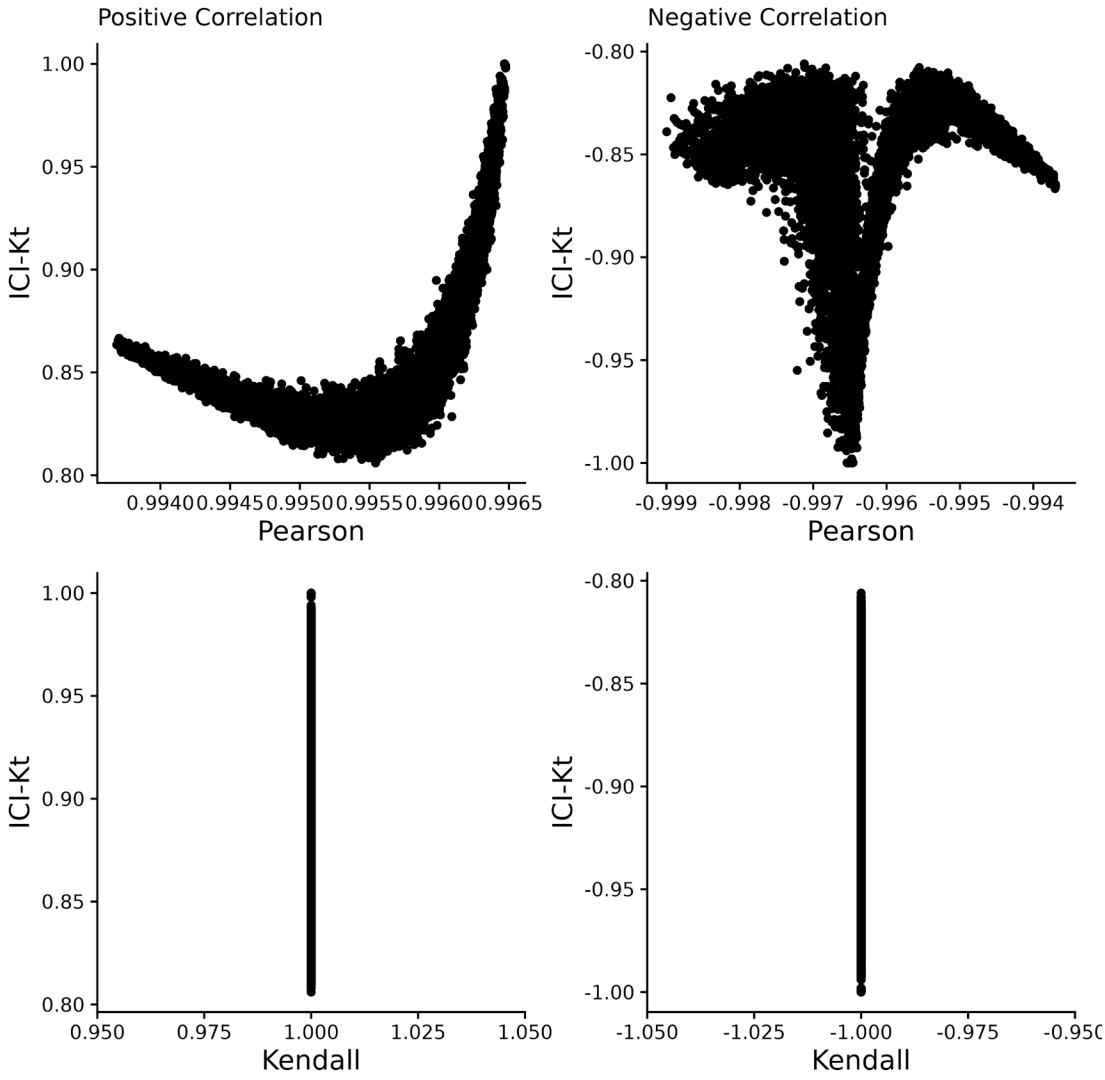

Figure 3. Comparing the correlation values obtained by Pearson, Kendall, and ICI-Kt correlation as an increasing number of missing values in the bottom half of either sample for both positively and negatively correlated samples. A subset of 10,000 points was used for visualization.

Effect of Left Censoring VS Random Missing Data

Figure 4 demonstrates the effect of introducing left-censored versus random missingness in five different measures of correlation, including the ICI-Kt, the normal Kendall-tau with

pairwise-complete-observations, the normal Kendall-tau replacing missing with 0, Pearson with *pairwise-complete-observations*, 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. In contrast to ICI-Kt, the Pearson correlation calculated using only pairwise complete entries is **constant** over growing left-centered and random missingness. When replacing missing values with zero, Pearson correlation demonstrates a small decrease in the correlation value with growing left-centered missingness due to the zero values causing some deviation from linearity. Pearson correlation 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 this special case where zero is lower than all of the values in the dataset, ICI-Kt and Kendall-tau replaced with zero result in identical correlation values, as shown in Figure S10A and Figure S10C. 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.

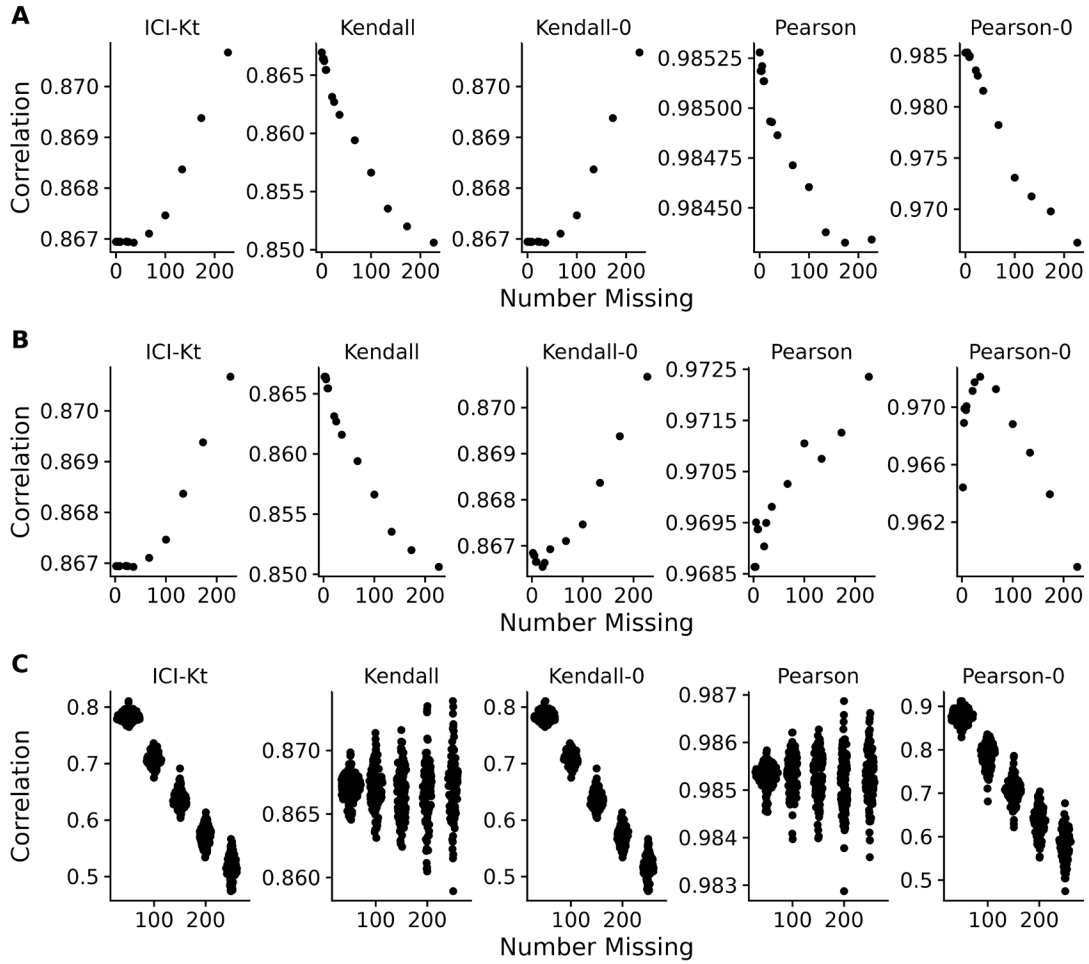

Figure 4. Effect of introducing missing values from a cutoff (**A** & **B**) or randomly (**C**) on different measures of correlation, including ICI-Kt, Kendall with pairwise complete, Kendall replacing missing with 0, Pearson with pairwise complete, and Pearson replacing missing with 0. **A**: Missing 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.

A common way missing data is handled in correlation calculations is to ignore them completely and use the pairwise complete cases to calculate the Pearson correlation coefficient. As shown in Figure 4C, this results in a complete misestimation of the changed correlative structure introduced by missing at random (MAR) entries. ICI-Kt, in contrast, incorporates the missingness in a sensical way, and the resulting correlation values fall as MAR entries are introduced.

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 5A shows that base R `cor` "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 S11).

As we implemented a relatively easy way to make use of asynchronous compute resources via the *future* package, we also verified that increases in run time using multiple cores was reasonable as well. Figure 5B shows that as samples are added, the run time quickly increases logarithmically, similar to increasing samples on a single core (see Figure 5A).

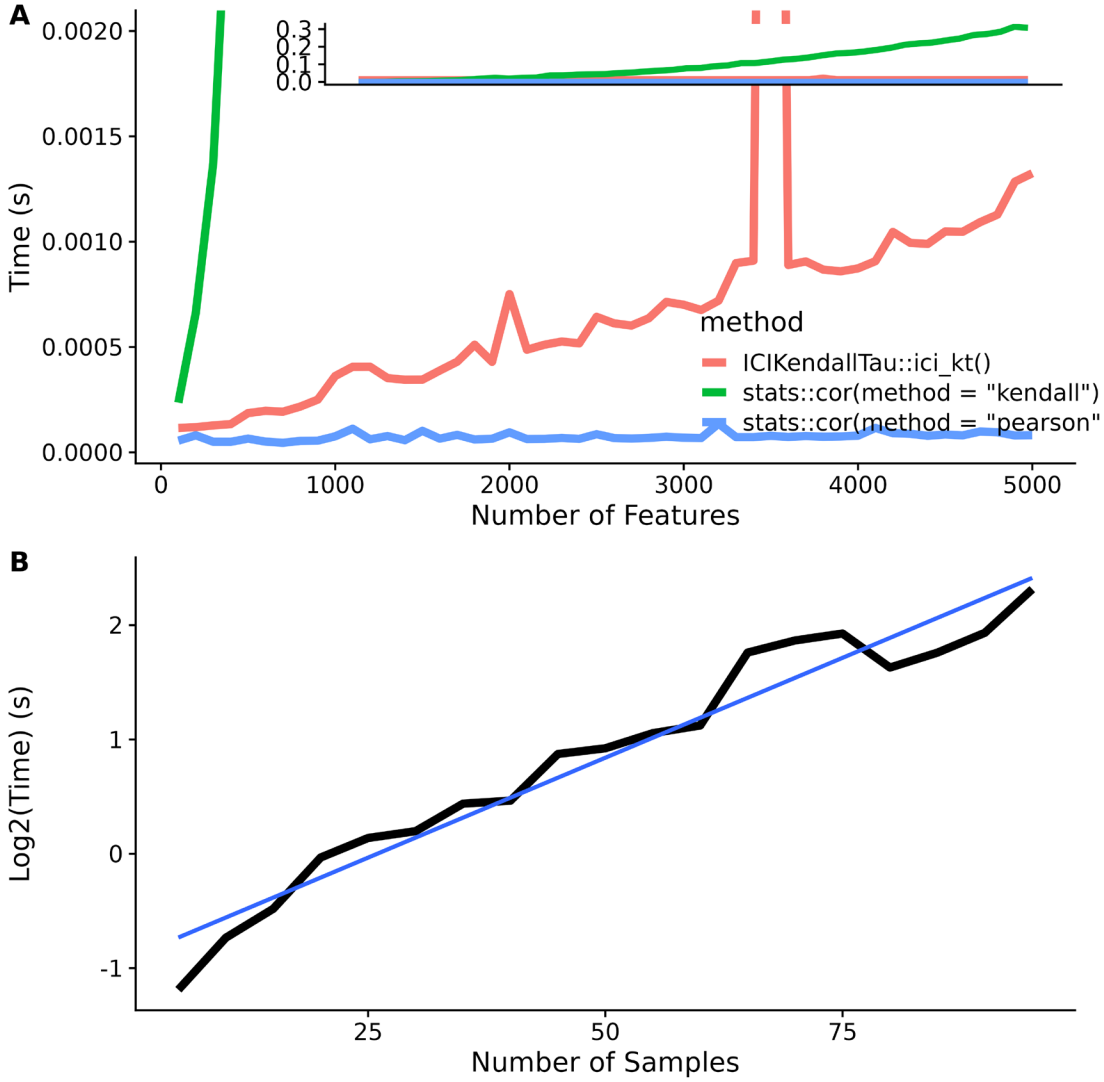

Figure 5. (A) Comparison of the *Rcpp mergesort ici_kt*, and base R's *cor* function using the "pearson" and "kendall" methods as the number of features are increased from 100 to 5000. Inset is the expanded view showing the full range of computation time in seconds. (B) Log of time taken to calculate all pairwise comparisons for different numbers of samples in multicore processing.

Utility for Large Omics' Data Sets

Detecting Outlier Samples

We analyzed datasets from RNA-Seq (3), lipidomics (1), and small molecule metabolomics (1) for possible outliers using a variety of different correlation and correlation-related metrics (see Correlation Methods): (1) ICI-Kt; (2) ICI-Kt scaled by *completeness* (see Methods); (3) Kendall-tau; (4) Pearson correlation; (5) Pearson correlation with no zeros; (6) Pearson on $log(x+1)$; and (7) Pearson on $log(x)$. We primarily concentrate on the ICI-Kt, ICI-Kt * completeness, and Pearson $log(x+1)$ here for one dataset, but all outliers detected by each method are shown in the Supplemental materials (Figures S12-S16, Tables S1-S6). 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, condition, etc).

Figure 6 visualizes the sample-specific median of each correlation-related metric calculated from the Yeast RNA-Seq data set. Table 2 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 S12 and Table S1).

Gierliński et al [14] 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 greater than 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 (also see PN0). This version of Pearson correlation recreates the median sample-sample correlations observed in the original report. We denote the actual samples recorded as outliers by Gierlinski and coworkers with "Manuscript". In Figure S12 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 correlation) 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 greater than 1.5X 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: WT.22, WT.25, 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 greater than -2.8 (evaluated by RMF eyeball on a zoomed in graph and using a ruler), a value that is never stated in the manuscript, and which seems arbitrary from the data. In addition, given the proportional variance inherent in RNA-Seq count data, Pearson correlation on non-transformed counts should not be used, as the low abundance counts will tend to force the correlation values to be arbitrarily high in the absence of large biological heterogeneity, as observed for these samples compared to the other datasets that involve individuals as biological replicates.

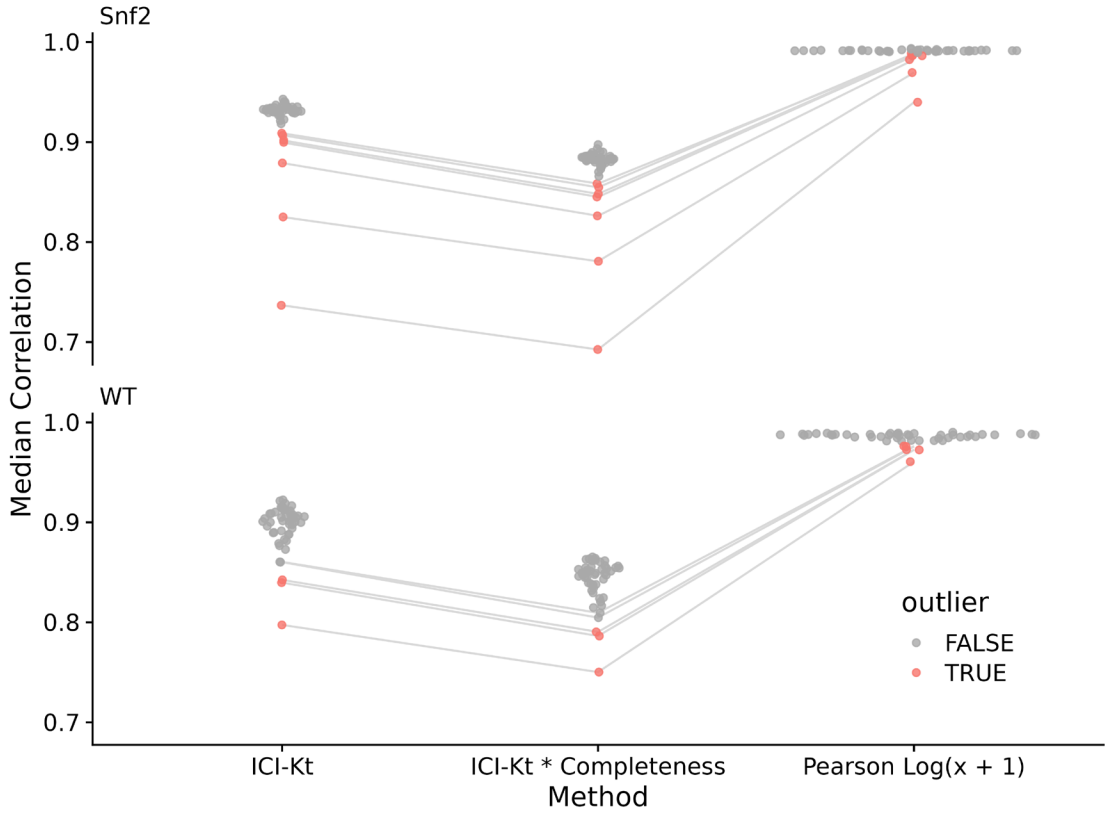

Figure 6. Median correlations for each of the yeast RNA-Seq sample to all other samples in the same group, using either ICI-Kt, ICI-Kt * completeness or Pearson Log(x + 1) correlation. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

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

Sample	ICI-Kt	ICI-Kt * Completeness	Pearson Log(x + 1)
Snf2.10	**0.907**	**0.855**	**0.989**
Snf2.31	**0.902**	**0.848**	**0.987**
Snf2.35	**0.909**	**0.858**	**0.986**
Snf2.15	**0.900**	**0.845**	**0.986**
Snf2.25	**0.879**	**0.826**	**0.982**
Snf2.13	**0.825**	**0.781**	**0.970**
Snf2.06	**0.737**	**0.693**	**0.940**
WT.36	0.860	0.810	**0.976**
WT.28	0.860	0.805	**0.976**
WT.25	**0.843**	**0.790**	**0.973**
WT.34	**0.840**	**0.786**	**0.973**
WT.21	**0.797**	**0.750**	**0.961**

In the case of the adenocarcinoma dataset (Figure S14 and Table S3), although there are not many outliers in these relatively large groups, Kendall-tau (K) values results in the most outliers, mostly in the *tumor* samples. Also of note is that ICI-Kt * Completeness (IKC) is the only measure where the variance measure for both groups is similar (Table S4) (1.8X). For both ICI-Kt and the various Pearson correlation, the tumor samples variance measures are 3.1X and 3.6X greater than normal, respectively.

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 (see Figures S17-S31). Although, in general the correlations remain the same, or show some downward trend with increasing fractions of samples required to be non-missing, IKC shows a different trend. Overall, this is due to there being a higher likelihood of pairwise samples being *complete* with the requirement for a feature to be in a greater fraction of a class of samples.

With the exception of the yeast dataset, the base Pearson (PB) and Pearson without zeros (PN0) were the most likely methods to show higher changes in median and median-absolute-differences (MAD) between groups of samples (see Methods).

Partitioning of Feature-Feature Networks

Feature-feature correlations (for those features present in 25% or more of the samples of a class) for each dataset were calculated using each of the correlation methods (see Methods). From those feature-feature correlations, partial correlations are calculated, and significant edges retained. Feature annotations from either Reactome pathways (RNA-Seq),

KEGG pathways (metabolomics), or lipid class (lipidomics) were used to determine if the edges within annotations were more present than outside of annotations. The biochemical rationale for this analysis is that stronger positive partial correlations should be reflected within known pathways versus between pathways.

Table 3 lists the summed partitioning values for each dataset and method. For each of the datasets examined, ICI-Kt or ICI-Kt completeness gave the most positive summed partitioning values.

Table 3. Partitioning ratios for feature-feature correlation networks generated using each method for each dataset.

Method	Yeast RNA-Seq	EGFR Genotype RNA-Seq	Rat Stamina Metabolomics	NSCLC Lipidomics
ICI-Kt	**-0.918**	**-0.720**	0.045	-0.139
ICI-Kt * Completeness	-0.935	-1.011	**0.395**	**-0.112**
Kendall-tau	-0.943	-0.785	-0.756	-0.272
Pearson Base	-1.001	-0.844	0.009	-0.146
Pearson Log(x + 1)	-1.008	-0.858	-0.055	-0.180
Pearson No Zeros	-3.317	-0.885	-0.703	-0.322
Pearson Log(x)		-0.952	-0.373	-0.291

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 [50] 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. In the case of using sample-sample correlation to detect outliers,

imputation does not solve any of the issues related to discovering outliers, as it should be applied **after** outlier samples are removed, otherwise the imputed values may not be useful. When used to create feature-feature networks based on partial correlations derived from the feature-feature correlations, ICI-Kt based methods also showed the best partitioning of features based on their known pathway annotations. 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, 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.

As demonstrated with the 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). We are heavily indebted to the University

of Kentucky Center for Computational Sciences (CCS) provided Kentucky Research Informatics Cloud (KyRIC), an NSF supported computational resource (NSF ACI1626364) for access to large virtual machines that allowed for methods development and the running of many of the larger datasets in this manuscript.

Supplemental materials

In addition to the over 30 supplemental figures and tables, the following Zenodo item includes all instructions for accessing public domain datasets, methods used to generate results, and full results generated: https://zenodo.org/doi/10.5281/zenodo.6309187

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. Love MI, Anders S, Kim V, Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Research; 2016 Nov. Report No.: 4:1070. doi:10.12688/f1000research.7035.2

11. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research; 2018 Dec. Report No.: 5:1408. doi:10.12688/f1000research.9005.3

12. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research; 2016 Aug. Report No.: 5:1438. doi:10.12688/f1000research.8987.2

13. 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

14. 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

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

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

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

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

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

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

21. 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

22. 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

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

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

25. 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

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

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

28. Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS, Smith K. Cython: The Best of Both Worlds. Computing in Science & Engineering. 2011;13: 31–39. doi:10.1109/MCSE.2010.118

29. 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

30. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15: 550. doi:10.1186/s13059-014-0550-8

31. 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

32. 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

33. 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

34. 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

35. 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

36. Flight RM, Moseley HNB, Mitchell JM. Untargeted lipidomics of non-small cell lung carcinoma data and code. figshare; 2021. doi:10.6084/m9.figshare.14199521.v3

37. Overmyer KA, Evans CR, Qi NR, Minogue CE, Carson JJ, Chermside-Scabbo CJ, et al. Maximal Oxidative Capacity during Exercise Is Associated with Skeletal Muscle Fuel Selection and Dynamic Changes in Mitochondrial Protein Acetylation. Cell Metabolism. 2015;21: 468–478. doi:10.1016/j.cmet.2015.02.007

38. Metabolomics Workbench. PR000016. Metabolomics Workbench; 2016. doi:10.21228/M86P45

39. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33: D428–432. doi:10.1093/nar/gki072

40. Ligtenberg W. reactome.db: A set of annotation maps for reactome. 2023.

41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30. doi:10.1093/nar/28.1.27

42. Huckvale E, Moseley HNB. kegg_pull: a software package for the RESTful access and pulling from the Kyoto Encyclopedia of Gene and Genomes. BMC Bioinformatics. 2023;24: 78. doi:10.1186/s12859-023-05208-0

43. Mitchell JM, Flight RM, Moseley HNB. Deriving Lipid Classification Based on Molecular Formulas. Metabolites. 2020;10: E122. doi:10.3390/metabo10030122

44. Flight RM, Moseley HN. visualizationQualityControl: Development of visualization methods for quality control. 2021. Available:
https://moseleybioinformaticslab.github.io/visualizationQualityControl
https://github.com/moseleybioinformaticslab/visualizationQualityControl

45. Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics. 2018;14: 128. doi:10.1007/s11306-018-1420-2

46. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology. 2011;5: 21. doi:10.1186/1752-0509-5-21

47. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69: 026113. doi:10.1103/PhysRevE.69.026113

48. White S, Smyth P. A Spectral Clustering Approach To Finding Communities in Graphs. Proceedings of the 2005 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2005. pp. 274–285. doi:10.1137/1.9781611972757.25

49. Landau WM. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. Journal of Open Source Software. 2021;6: 2959. doi:10.21105/joss.02959

50. 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

2024-01-24 14:23:14.260331

Limits of Detection Issues in Datasets

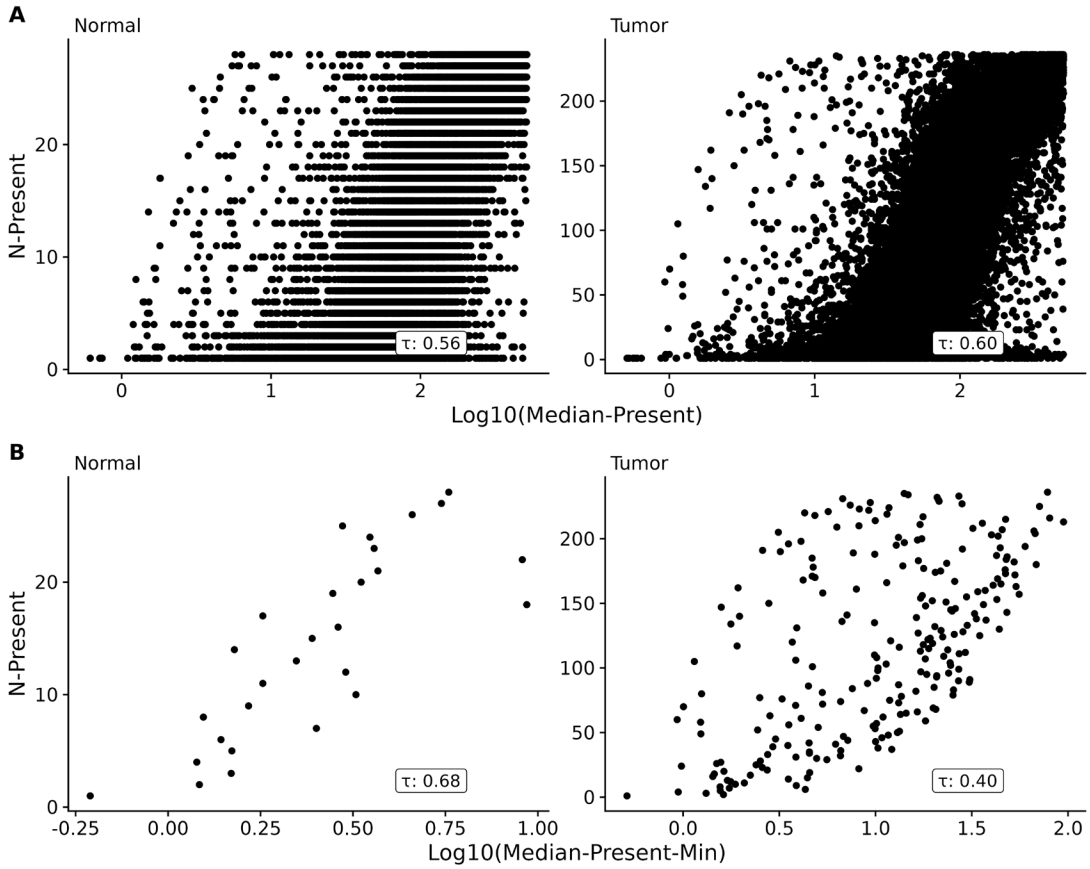

Figure S1. Adenocarcinoma dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

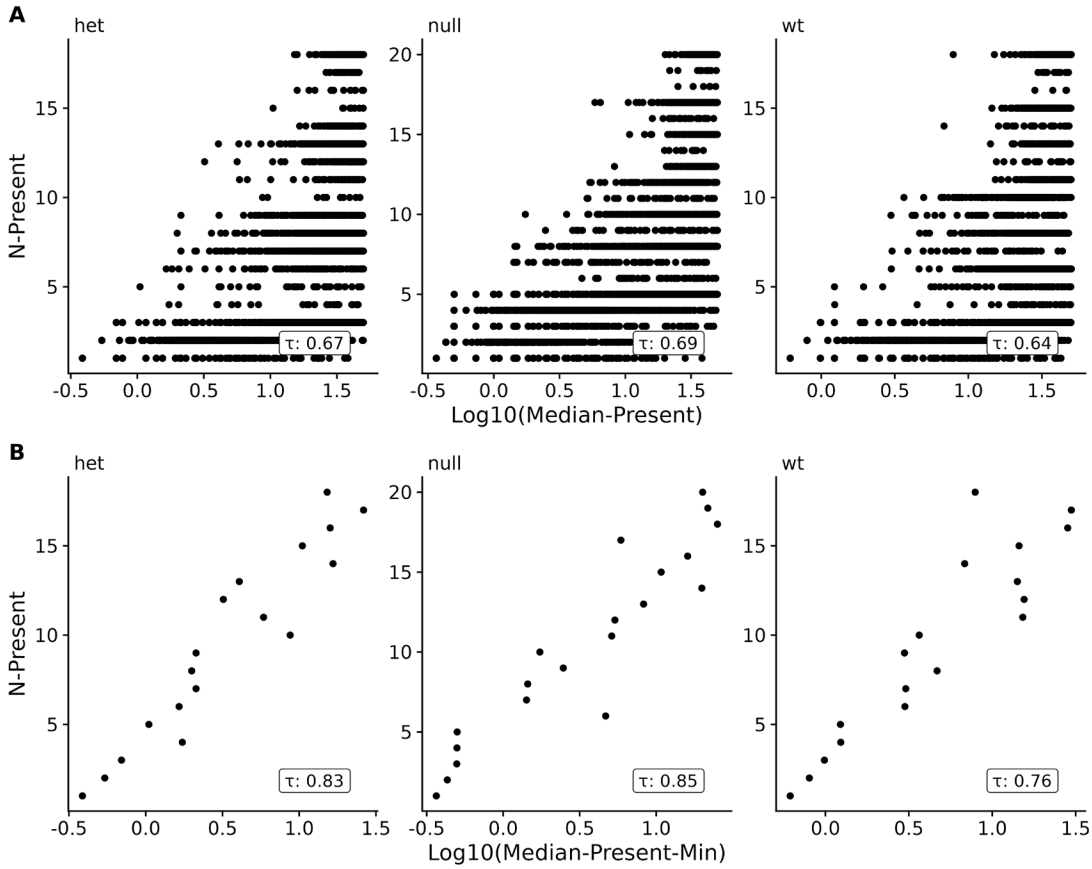

Figure S2. EGFR genotype dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

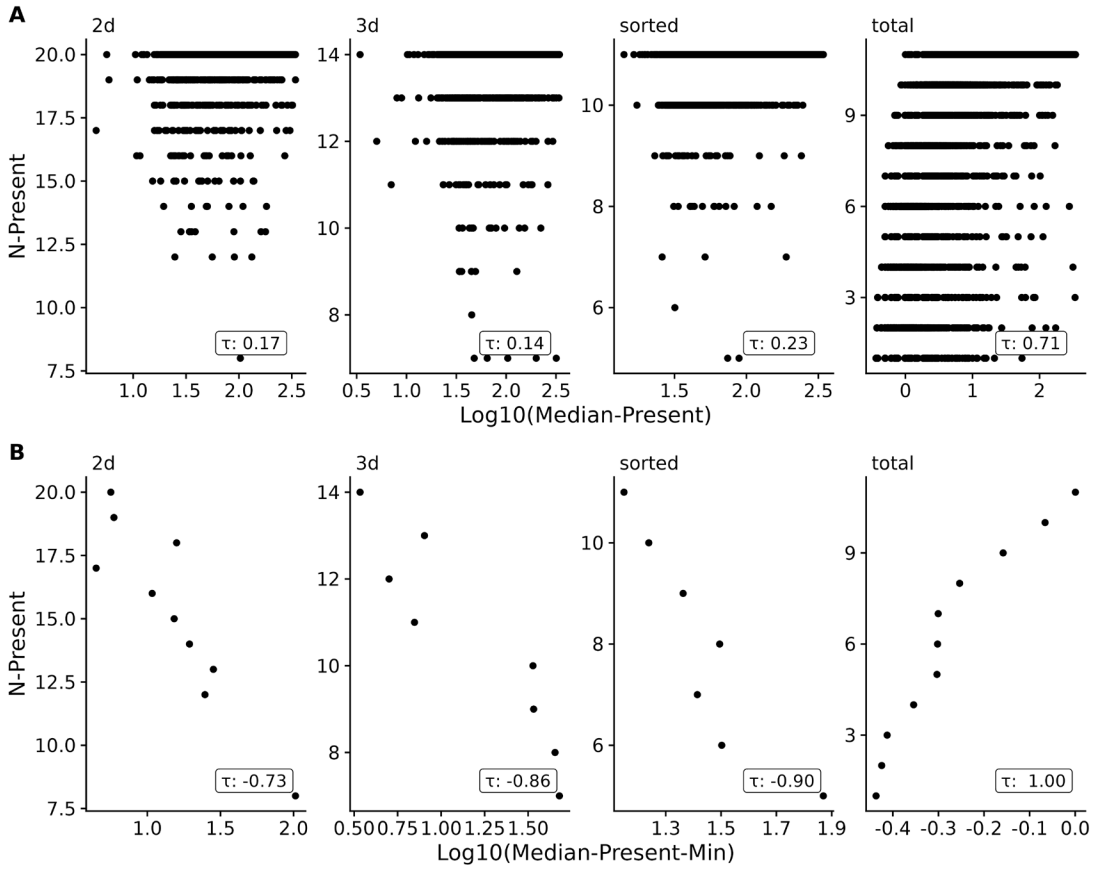

Figure S3. EGFR genotype and tumor culture dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

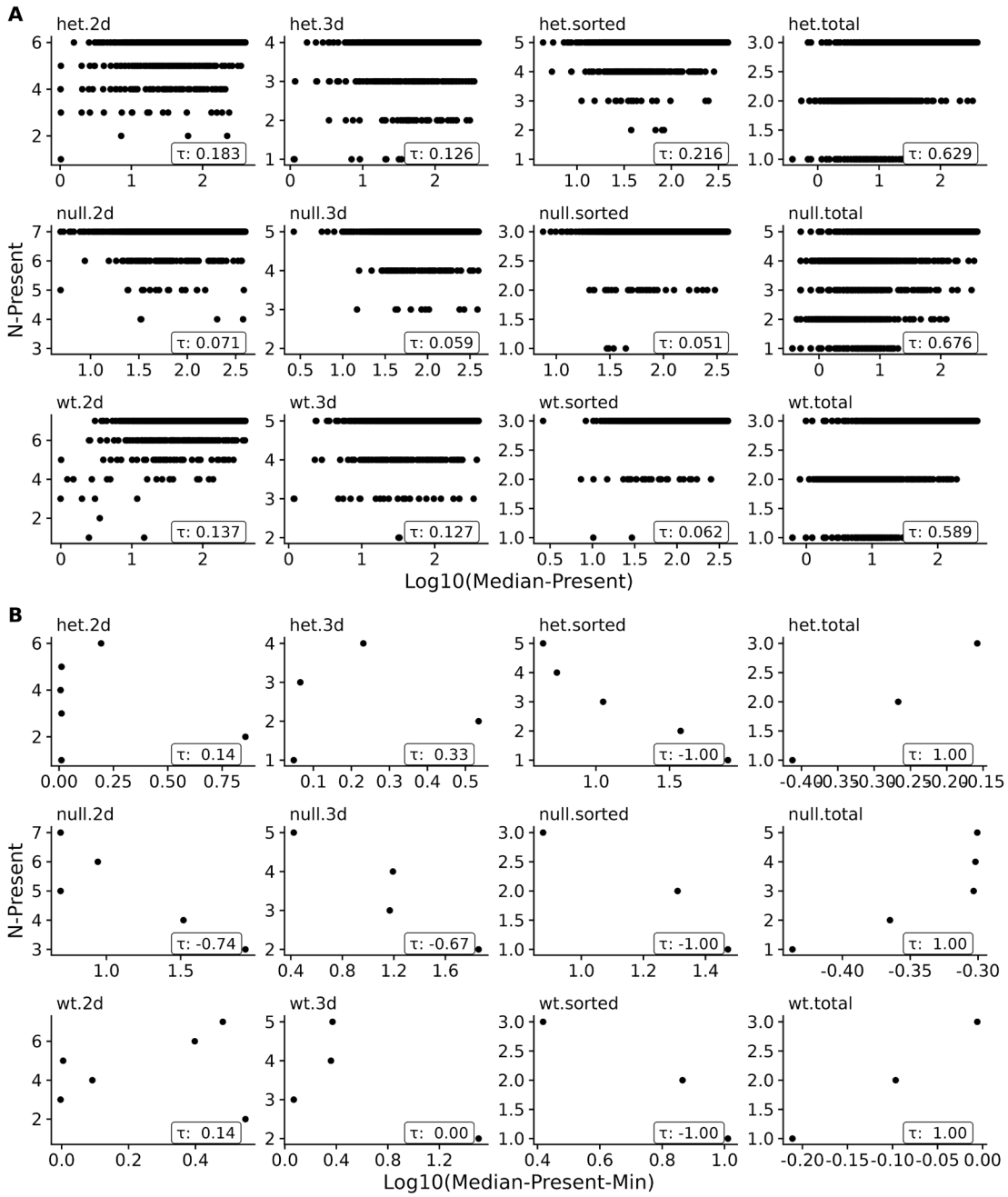

Figure S4. EGFR genotype and tumor culture dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

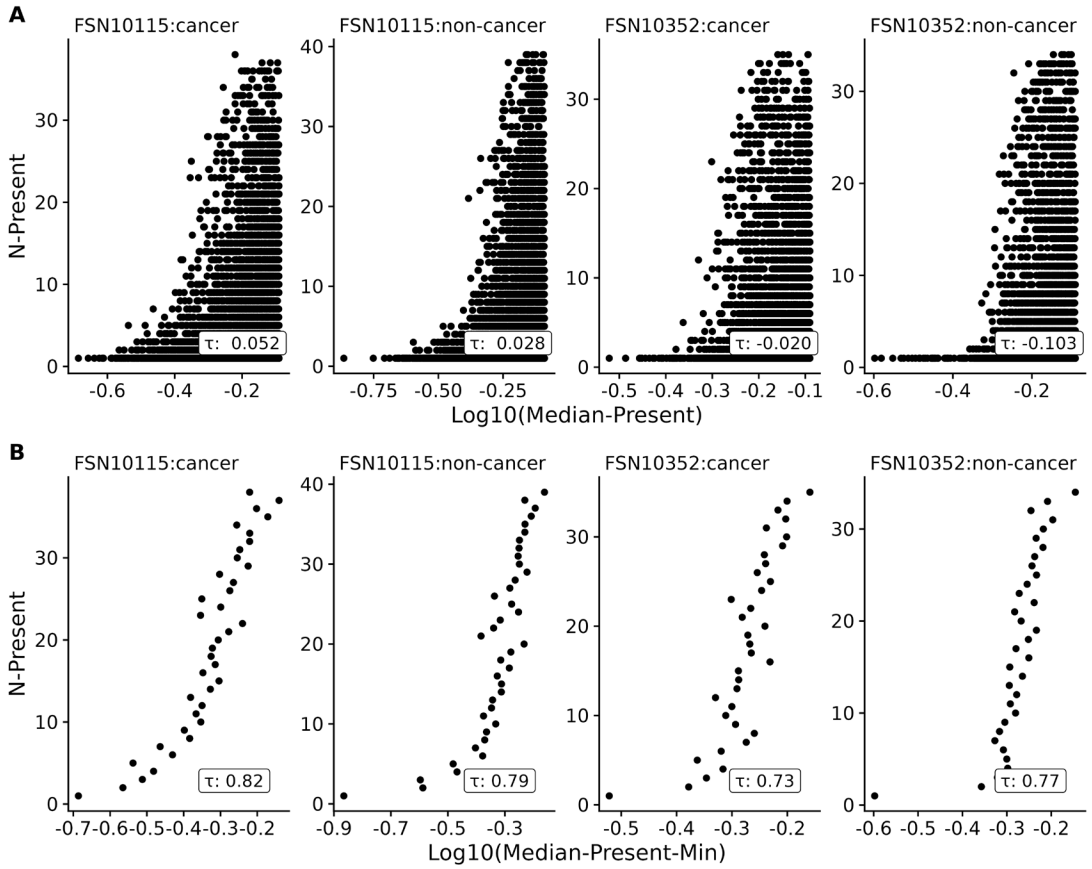

Figure S5. Non-small-cell-lung-cancer (NSCLC) lipidomics dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

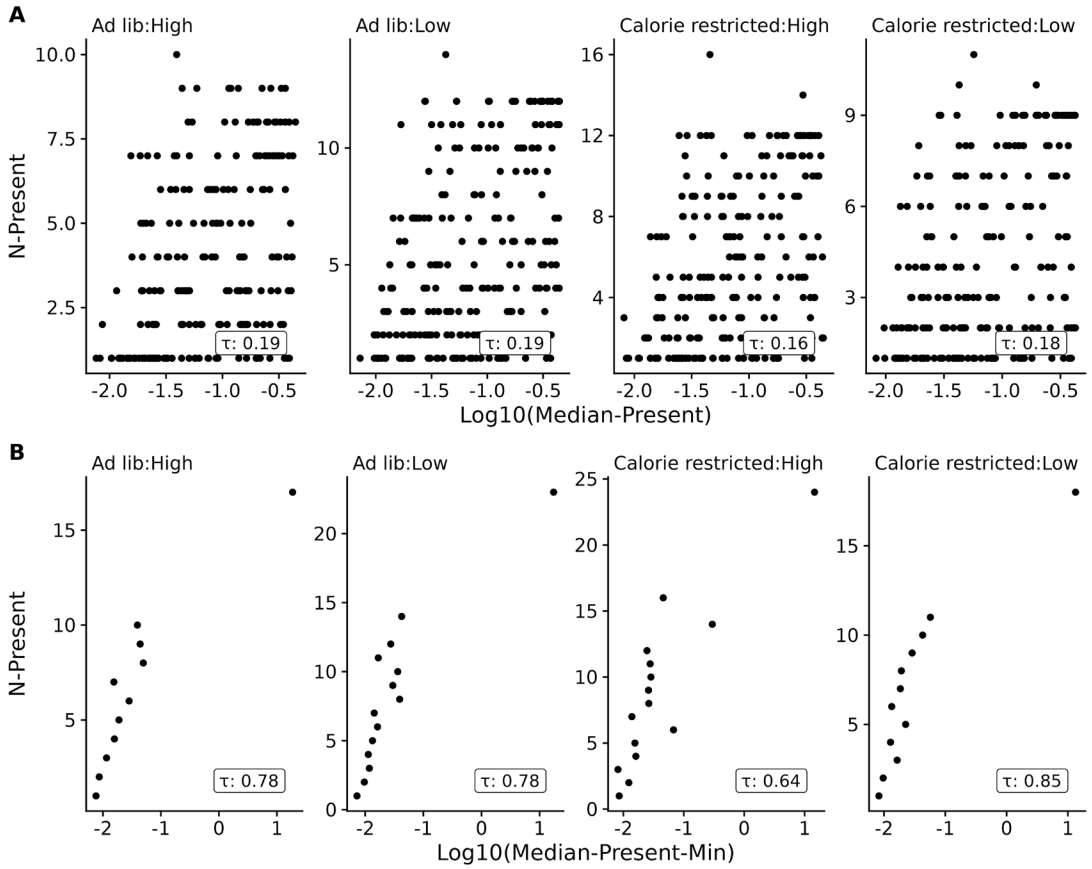

Figure S6. Rat stamina dataset count medians by number of samples feature was present in. **A**: Log10 of medians for present values compared to the number of samples the feature was present in. **B**: Log10 of minimum median for each value of number of samples the feature was present in.

Simulated Data

Simple Data Set

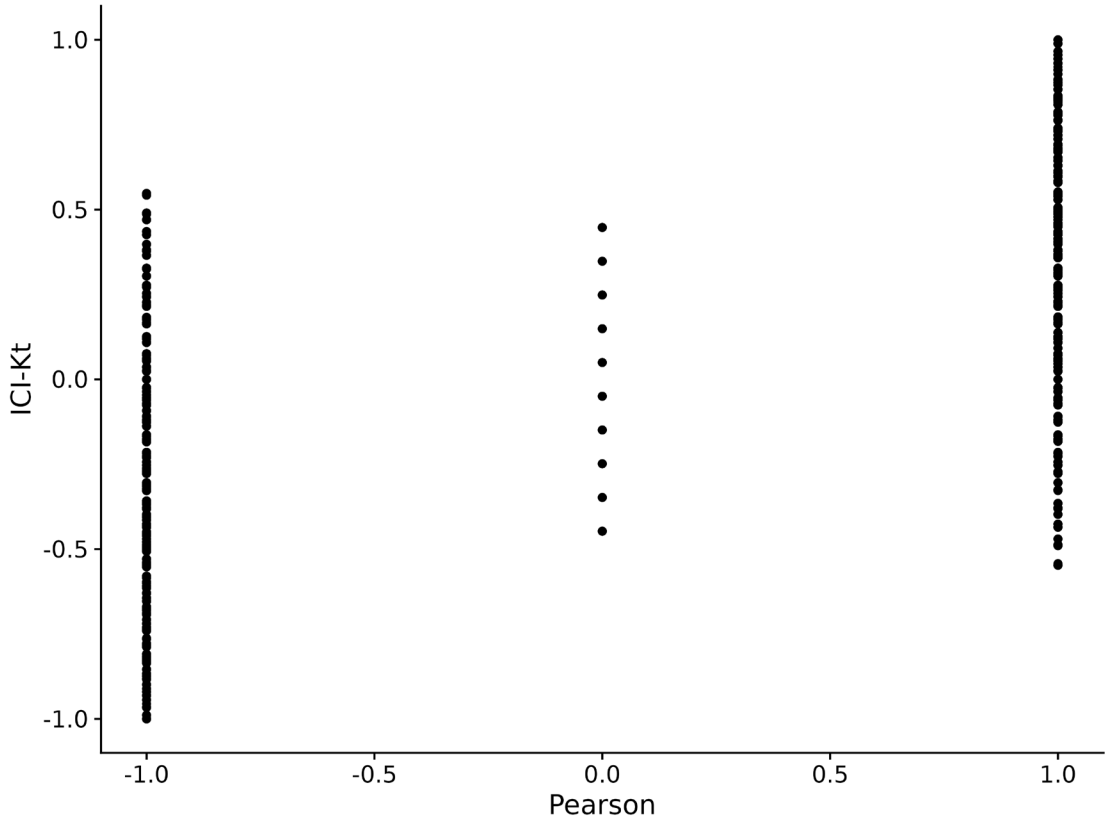

Figure S7. Comparison of ICI-Kt and Pearson correlations for perfectly positive and negatively correlated samples, systematically replacing values with NA. NA values from Pearson were replaced with zero for this comparison.

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 S9. 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.

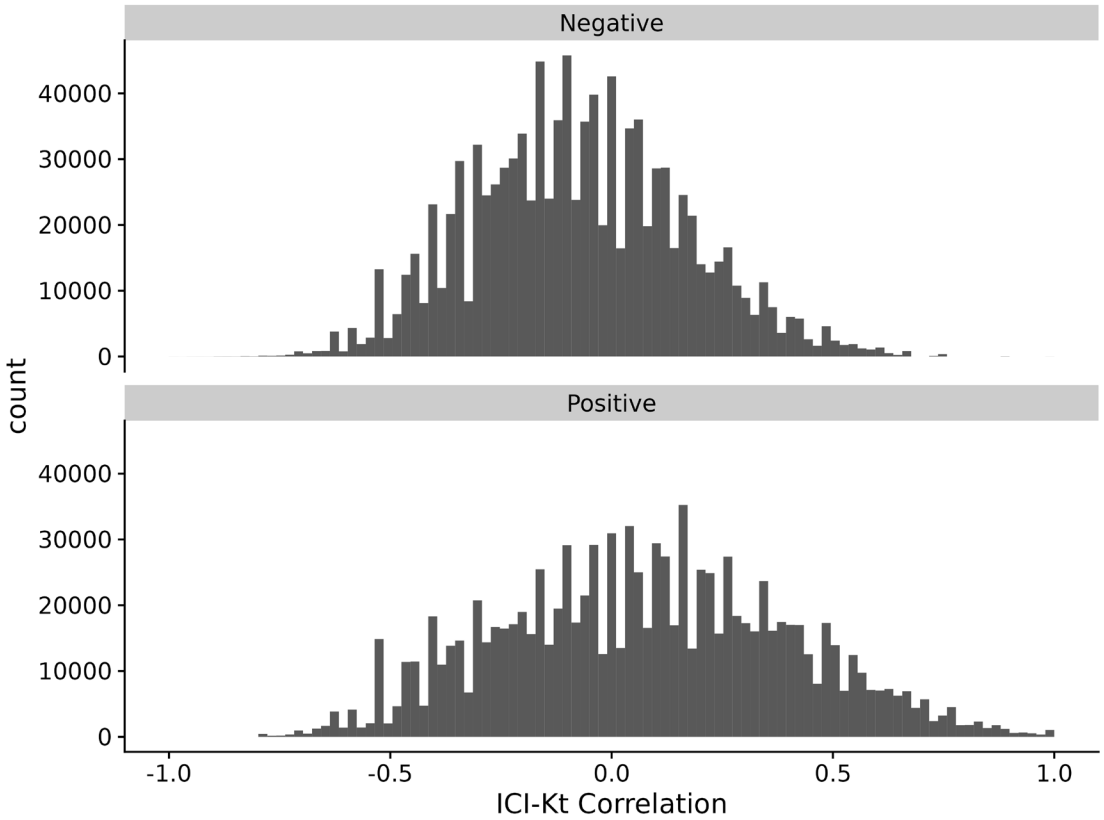

Figure S8. ICI-Kendall-tau correlation as missing values are varied between two samples.

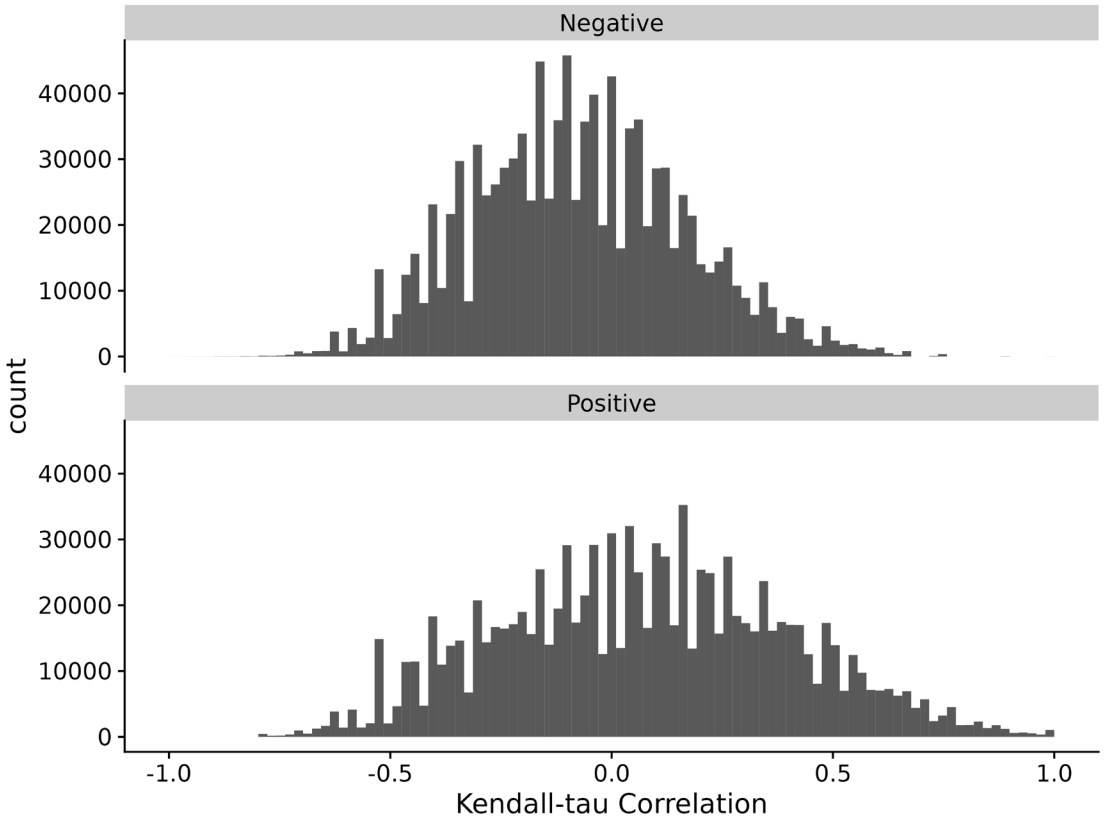

Figure S9. Kendall-tau correlation as missing values are varied between two samples and replaced with 0 before calculating Kendall-tau.

Semi-Realistic Data Set

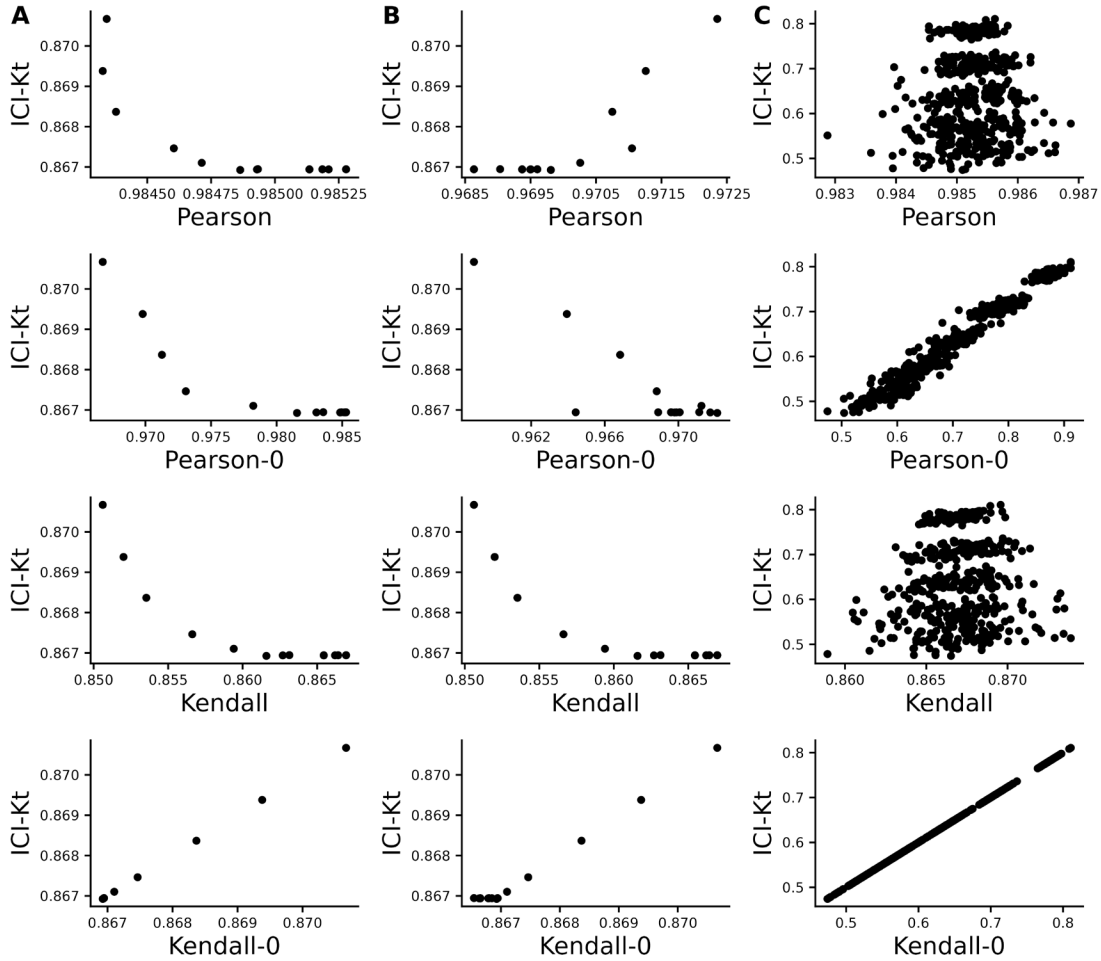

Figure S10. Effect of introducing missing values from a cutoff (A & B) or randomly (C) on different measures of correlation, including ICI-Kt, Kendall with pairwise complete, Kendall replacing missing with 0, Pearson with pairwise complete, and Pearson replacing missing with 0. A) Missing values introduced 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.

Performance

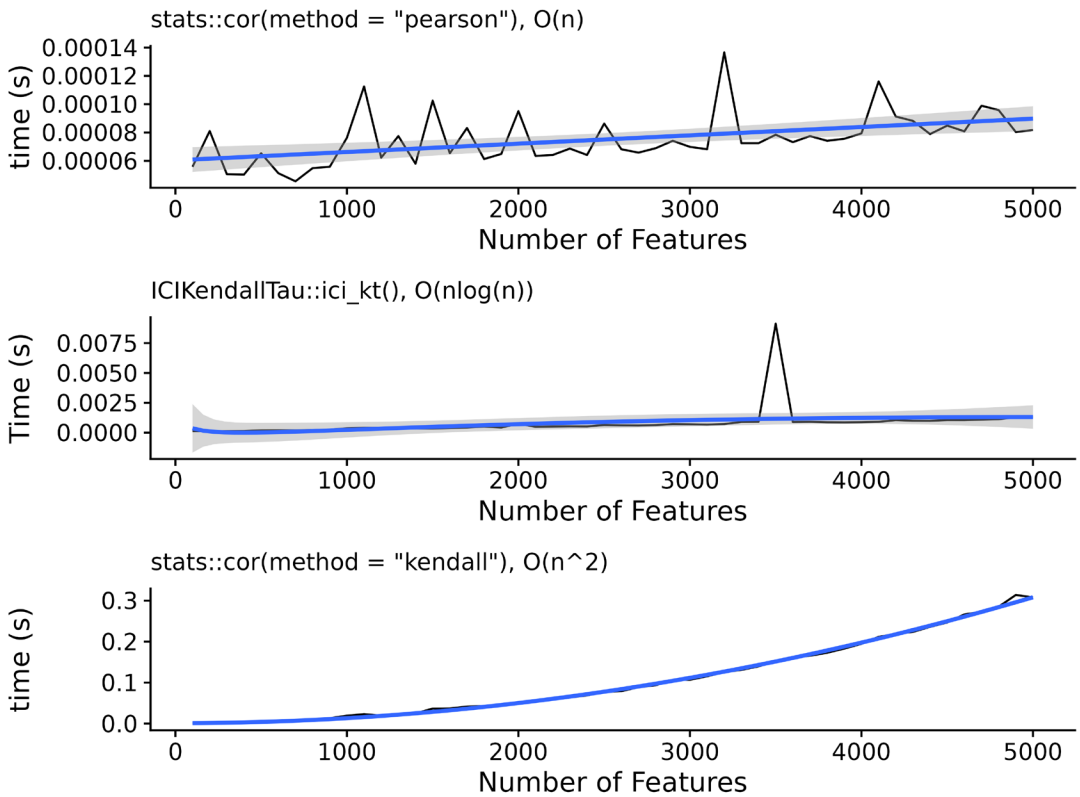

Figure S11. Time in seconds needed as a function of the number of features, with a fitted line for the assumed complexity for each of the methods tested, including R's Pearson correlation, the ICI-Kt mergesort, and R's Kendall-tau correlation algorithm.

Outlier Samples

Gierlinski Yeast

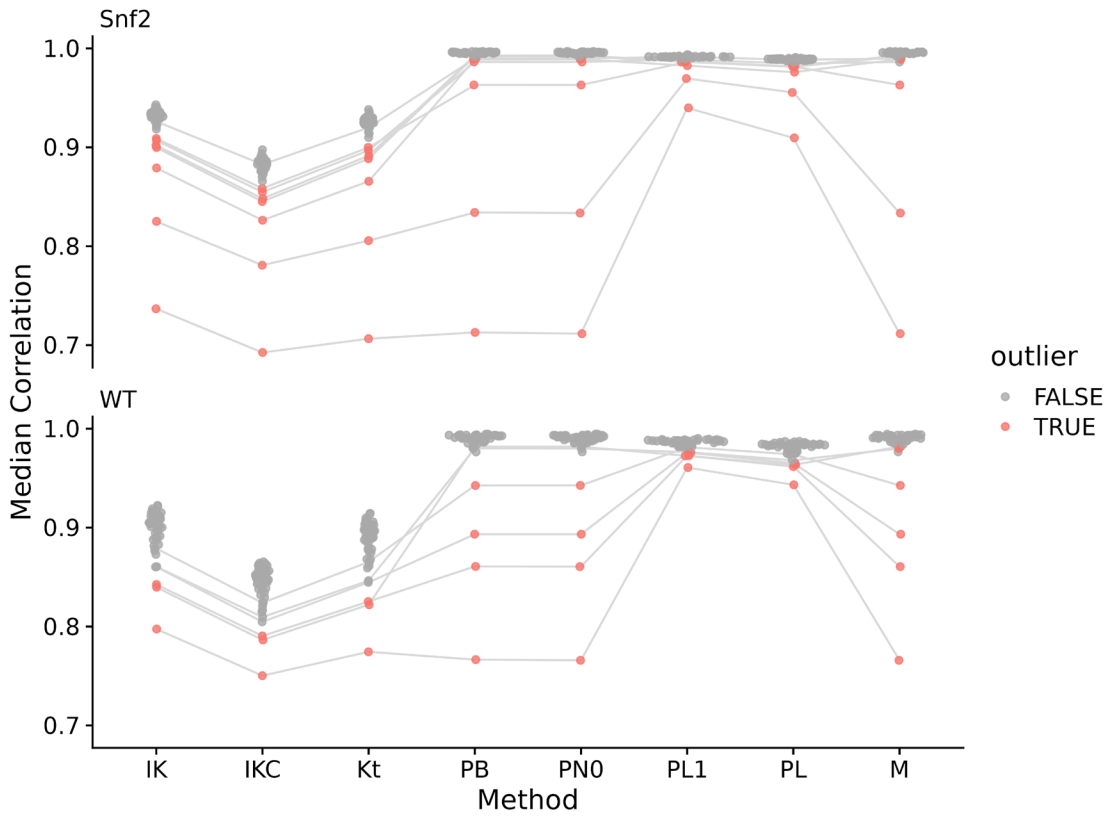

Figure S12. Median correlations for each of the yeast RNA-Seq samples to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; M: Manuscript; PL1: Pearson Log(x + 1); PN0: Pearson No Zeros. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

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.

Sample	IK	IKC	Kt	PB	PN0	PL1	PL	M
Snf2.31	**0.902**	**0.848**	**0.891**	0.993	0.993	**0.987**	**0.982**	0.993
Snf2.15	**0.900**	**0.845**	**0.888**	0.992	0.992	**0.986**	**0.981**	0.992
Snf2.24	0.926	0.883	0.920	**0.990**	**0.990**	0.991	0.988	0.990

Sample	IK	IKC	Kt	PB	PN0	PL1	PL	M
Snf2.25	**0.879**	**0.826**	**0.866**	**0.989**	**0.989**	**0.982**	**0.976**	**0.989**
Snf2.10	**0.907**	**0.855**	**0.897**	**0.986**	**0.986**	**0.989**	0.985	0.986
Snf2.35	**0.909**	**0.858**	**0.900**	**0.963**	**0.963**	**0.986**	**0.982**	**0.963**
Snf2.13	**0.825**	**0.781**	**0.806**	**0.834**	**0.834**	**0.970**	**0.955**	**0.834**
Snf2.06	**0.737**	**0.693**	**0.707**	**0.713**	**0.712**	**0.940**	**0.909**	**0.712**
WT.34	**0.840**	**0.786**	**0.822**	0.982	0.982	**0.973**	**0.963**	**0.982**
WT.36	0.860	0.810	0.846	0.980	0.980	**0.976**	0.968	**0.980**
WT.22	0.879	0.824	0.865	**0.943**	**0.943**	0.981	0.974	**0.943**
WT.28	0.860	0.805	0.845	**0.893**	**0.893**	**0.976**	**0.965**	**0.893**
WT.25	**0.843**	**0.790**	**0.825**	**0.861**	**0.861**	**0.973**	**0.962**	**0.861**
WT.21	**0.797**	**0.750**	**0.774**	**0.767**	**0.766**	**0.961**	**0.943**	**0.766**

EGFR RNASeq Genotype

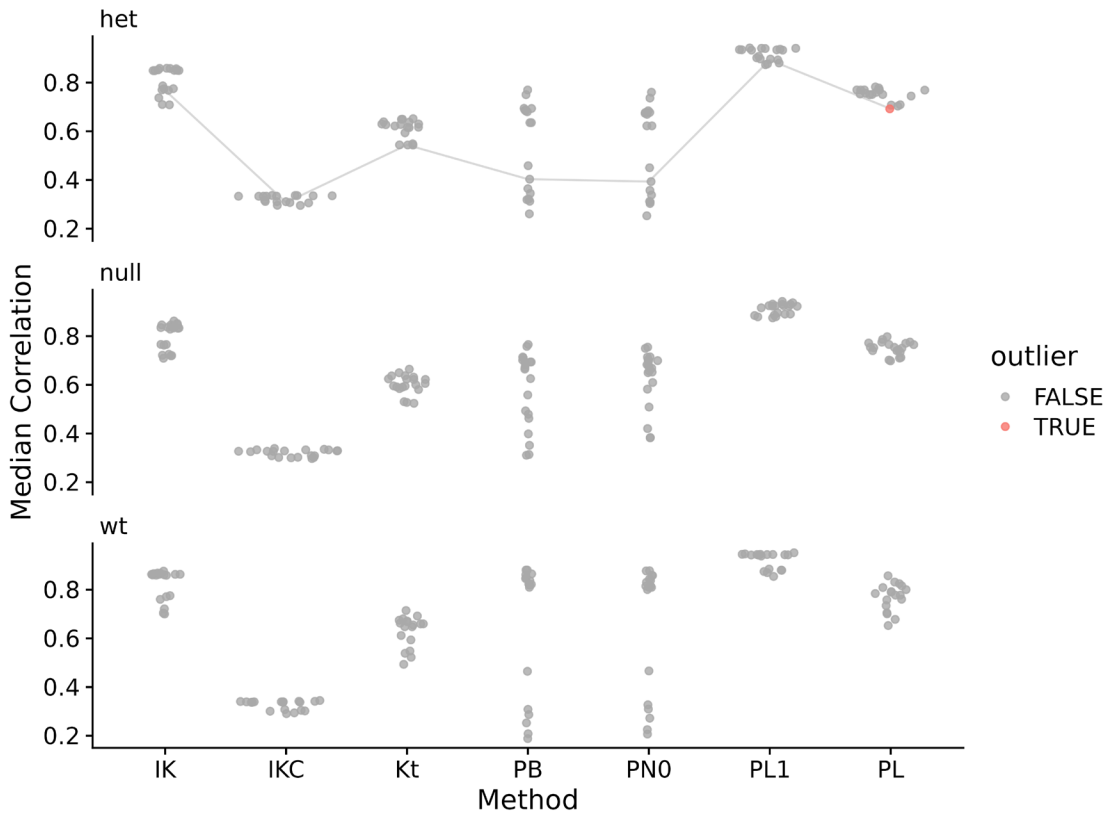

Figure S13. Median correlations for each of the Brainson RNA-Seq samples to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

Table S2. EGFR RNA-Seq dataset using genotype as the grouping variable 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.

Sample	IK	IKC	Kt	PB	PN0	PL1	PL
783-het.sorted	0.768	0.307	0.544	0.403	0.394	0.893	**0.692**

TCGA Adenocarcinoma

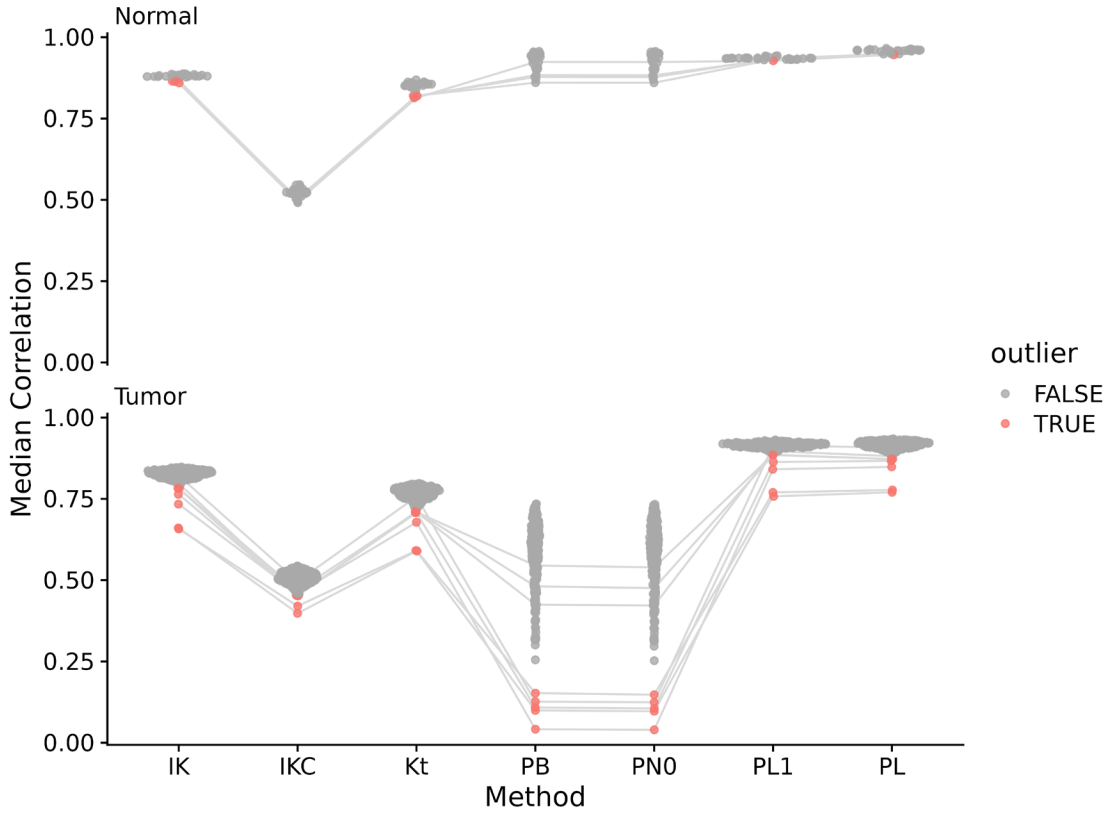

Figure S14. Median correlations for each of the TGCA adenocarcinoma RNA-Seq samples to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness;

PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

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.

Sample	IK	IKC	Kt	PB	PN0	PL1	PL
N.24	**0.864**	0.504	**0.819**	0.878	0.877	0.934	0.949
N.13	**0.865**	0.504	**0.818**	0.883	0.883	0.933	0.949
N.18	**0.867**	0.499	**0.820**	0.860	0.860	0.932	0.948
N.22	**0.859**	0.495	**0.814**	0.924	0.924	**0.928**	**0.946**
T.39	0.821	0.496	0.757	**0.126**	**0.124**	0.914	0.908
T.122	0.798	**0.453**	**0.710**	0.425	0.422	0.897	0.881
T.180	**0.783**	0.467	**0.709**	0.545	0.539	**0.886**	**0.872**
T.220	**0.782**	**0.452**	**0.712**	0.481	0.475	**0.885**	**0.872**
T.38	**0.764**	0.454	**0.707**	**0.109**	**0.105**	**0.863**	**0.867**
T.261	**0.734**	0.459	**0.678**	**0.042**	**0.040**	**0.841**	**0.848**
T.40	**0.657**	**0.420**	**0.591**	**0.153**	**0.148**	**0.770**	**0.777**
T.260	**0.661**	**0.398**	**0.590**	**0.099**	**0.097**	**0.758**	**0.770**

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

method	Normal	Tumor
IK	0.004	0.012
IKC	0.011	0.019
Kt	0.008	0.019
PB	0.022	0.092
PN0	0.021	0.091
PL1	0.002	0.008

method	Normal	Tumor
PL	0.003	0.012

NSCLC Lipidomics

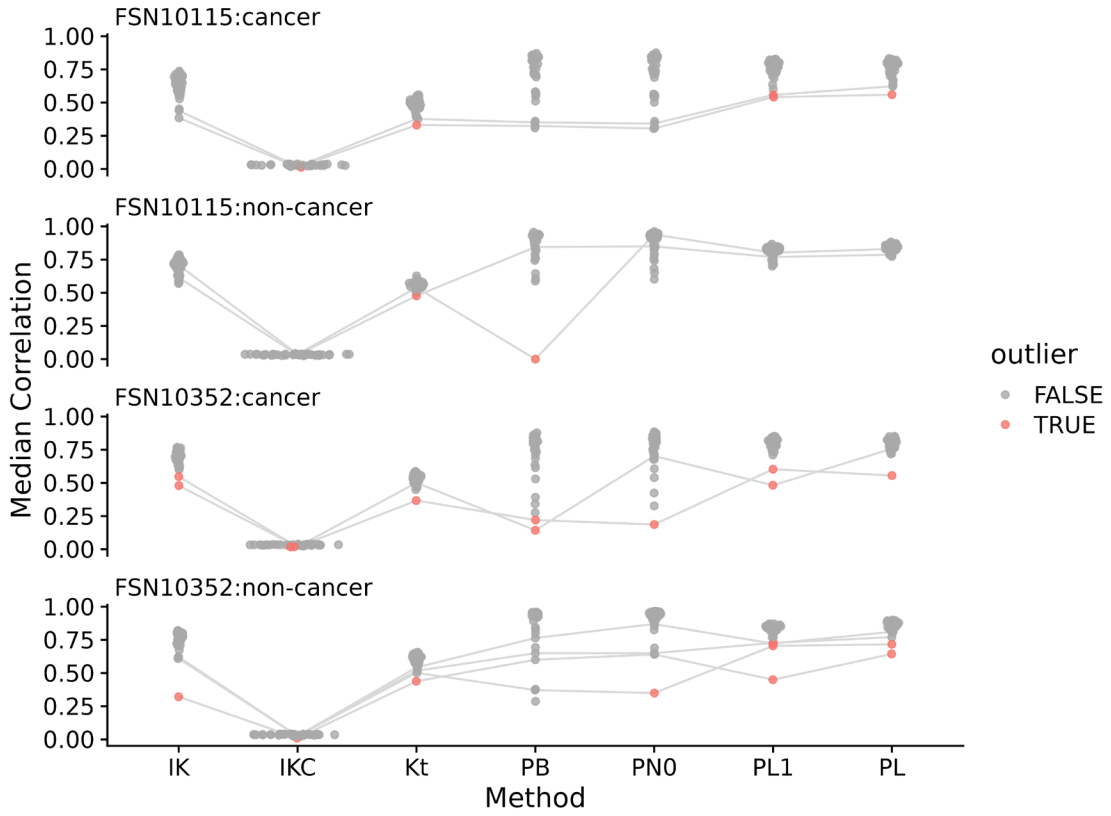

Figure S15. Median correlations for each of the NSCLC lipidomics samples to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt: Kendall-tau. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

Table S5. NSCLC lipidomics 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.

Sample	IK	IKC	Kt	PB	PN0	PL1	PL
172Cpos	0.440	0.019	0.377	0.351	0.342	**0.558**	0.623

Sample	IK	IKC	Kt	PB	PN0	PL1	PL
73Cpos	0.385	**0.013**	**0.331**	0.324	0.305	**0.541**	**0.560**
73Npos	0.708	0.037	0.538	**-0.000**	0.939	0.802	0.830
78Npos	0.615	0.032	**0.477**	0.845	0.849	0.769	0.787
uk003cposb	**0.547**	**0.018**	**0.367**	**0.220**	**0.187**	**0.603**	**0.556**
47Cpos	**0.480**	**0.019**	0.501	**0.143**	0.703	**0.482**	0.758
155Nposr	0.609	0.025	0.517	0.650	0.650	**0.727**	0.771
201Npos	0.621	0.026	0.542	0.764	0.870	**0.722**	0.811
169Npos	0.621	0.026	0.502	0.372	**0.350**	**0.705**	**0.717**
148Npos	**0.322**	**0.010**	**0.438**	0.600	0.640	**0.450**	**0.644**

Rat Stamina Metabolomics

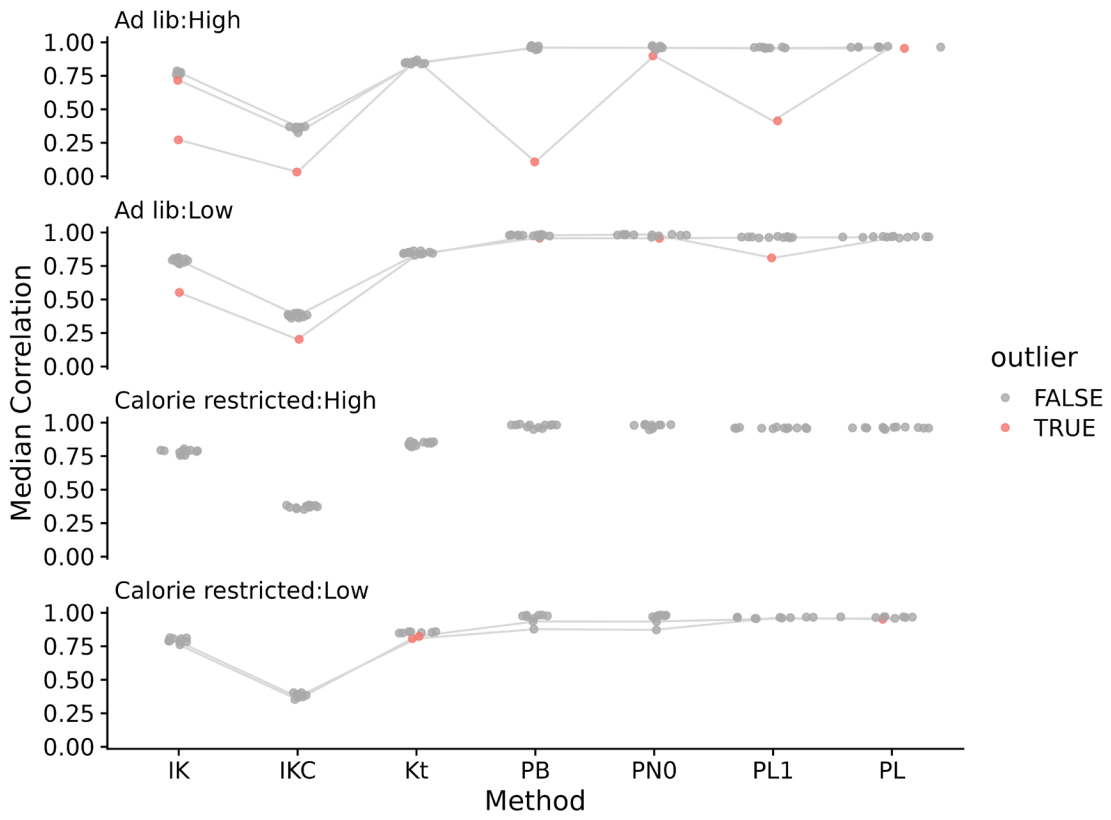

Figure S16. Median correlations for each of the rat stamina metabolomics samples to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; PB: Pearson Base, PN: Pearson No Zero; PL1: Pearson Log(x + 1); PL: Pearson Log(x); Kt:

Kendall-tau. Points are colored red if they were an outlier using that correlation method. For a sample that is considered an outlier in any of the method, lines are drawn connecting them between methods.

Table S6. Rat stamina metabolomics 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.

Sample	IK	IKC	Kt	PB	PN0	PL1	PL
s00009502	**0.717**	0.325	0.848	0.959	0.959	0.957	0.963
s00009501	0.776	0.371	0.844	0.961	0.960	0.955	**0.955**
s00009481	**0.273**	**0.032**	0.869	**0.108**	**0.899**	**0.414**	0.970
s00009471	0.794	0.380	0.838	**0.957**	**0.957**	0.962	0.965
s00009470	**0.553**	**0.203**	0.829	0.979	0.985	**0.810**	0.959
s00009484	0.783	0.372	**0.808**	0.878	0.872	0.961	**0.953**
s00009508	0.762	0.353	**0.823**	0.935	0.934	0.956	0.959

Effect of Increasing Presence in Samples

Yeast Samples

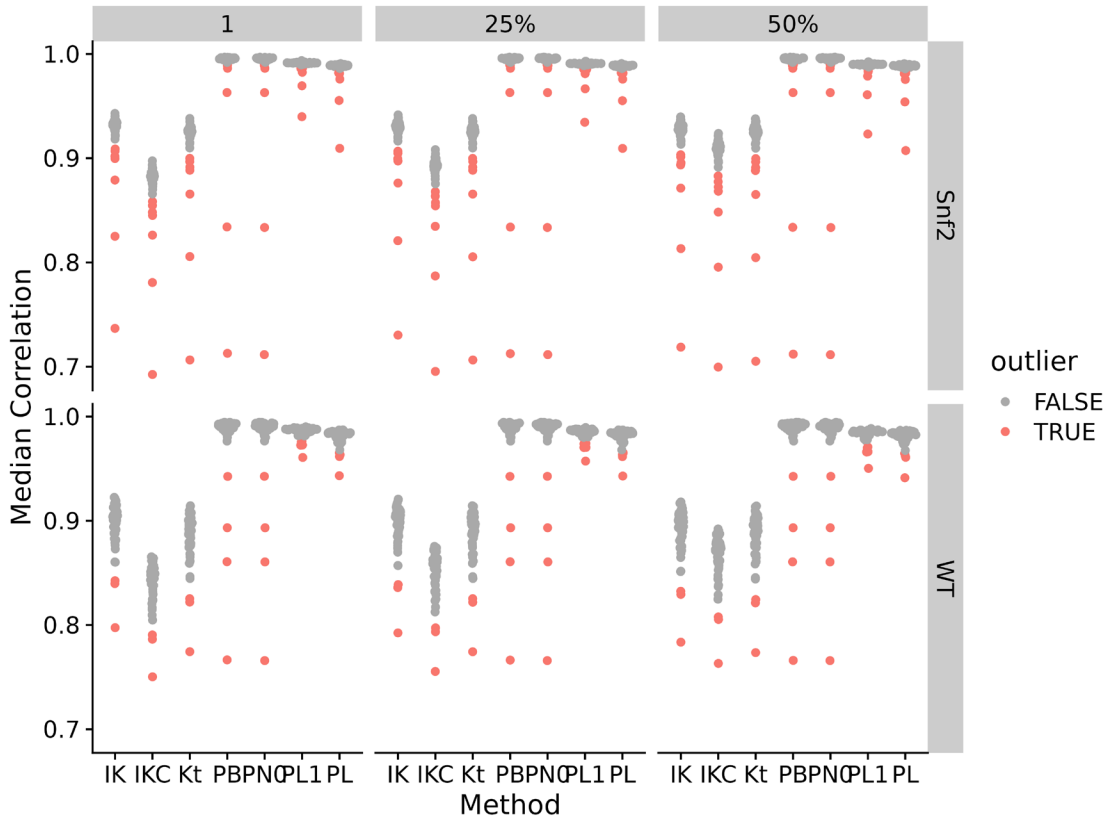

Figure S17. Median correlations by correlation method and applying different fractional cutoffs. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

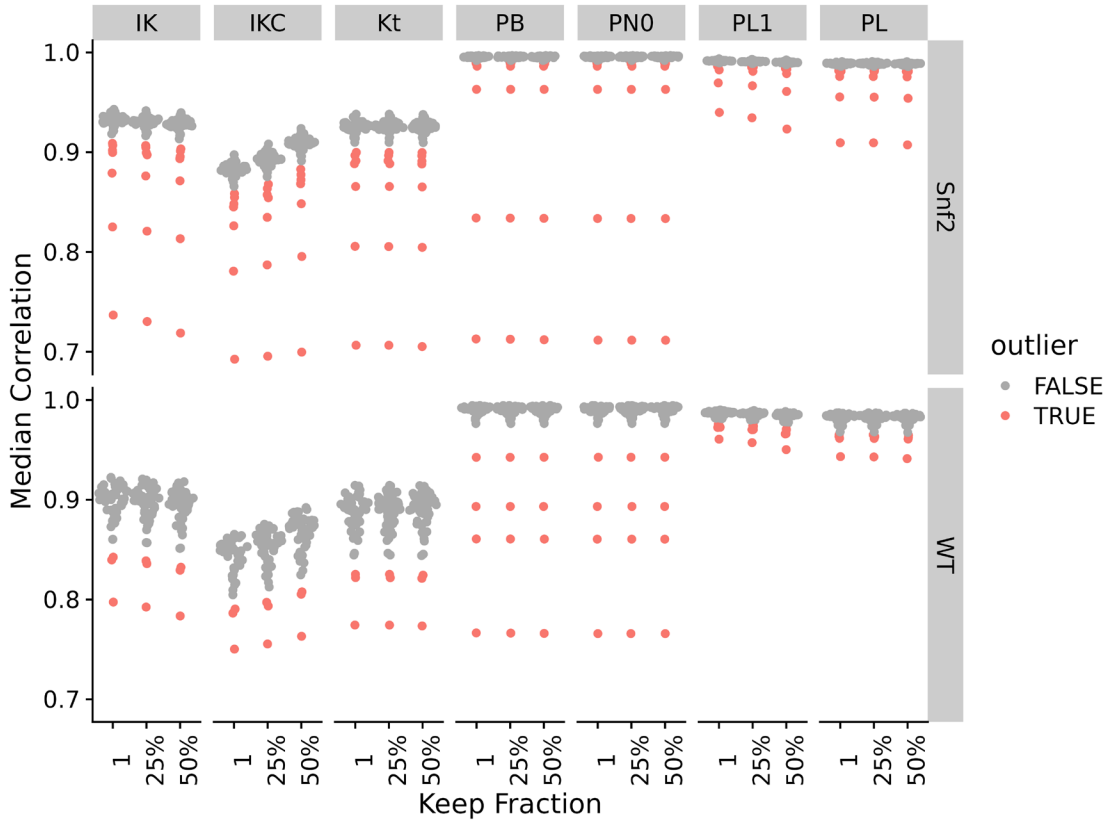

Figure S18. Median correlations by applying different fractional cutoffs and different correlation methods. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

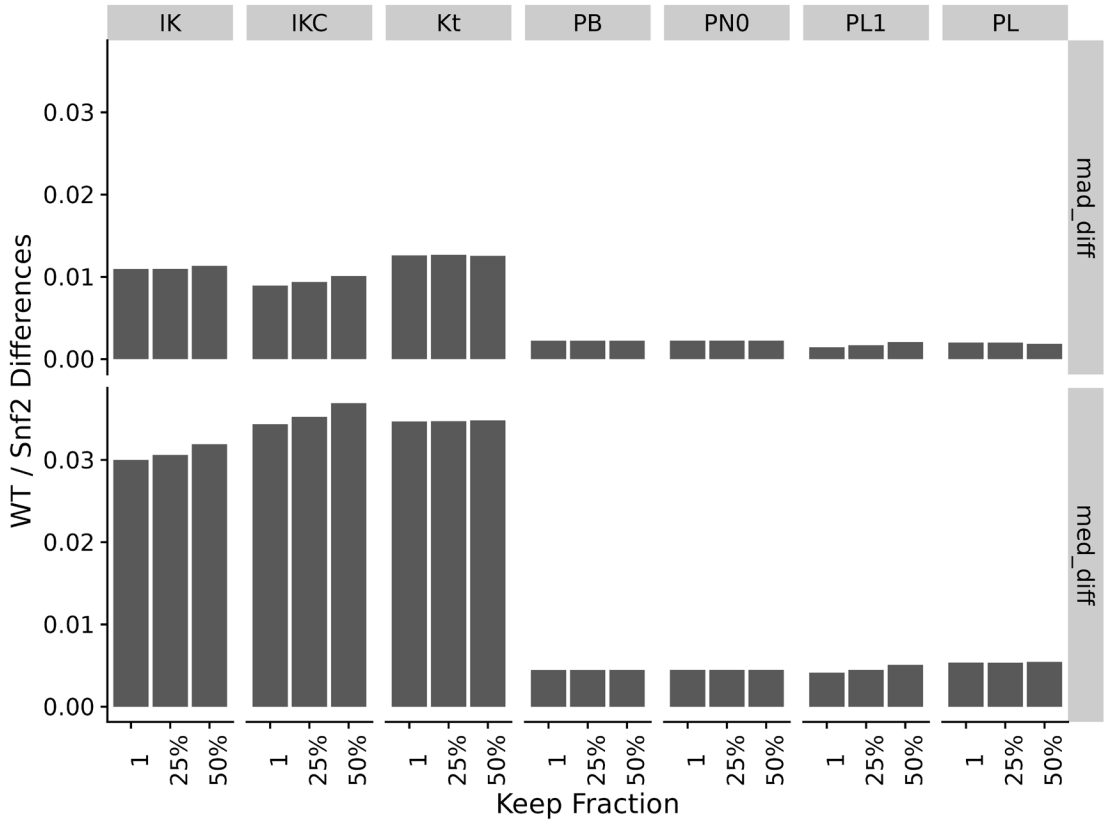

Figure S19. Difference between WT and Snf2 group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

EGFR Genotype Samples

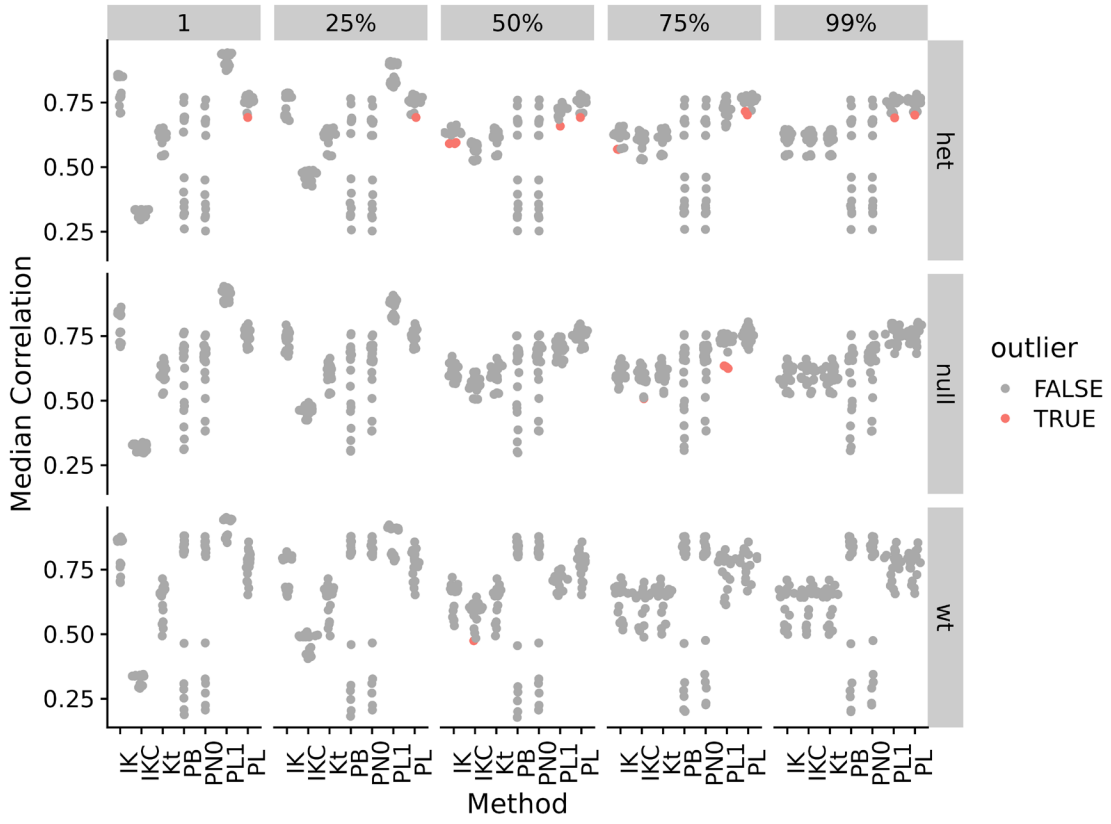

Figure S20. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

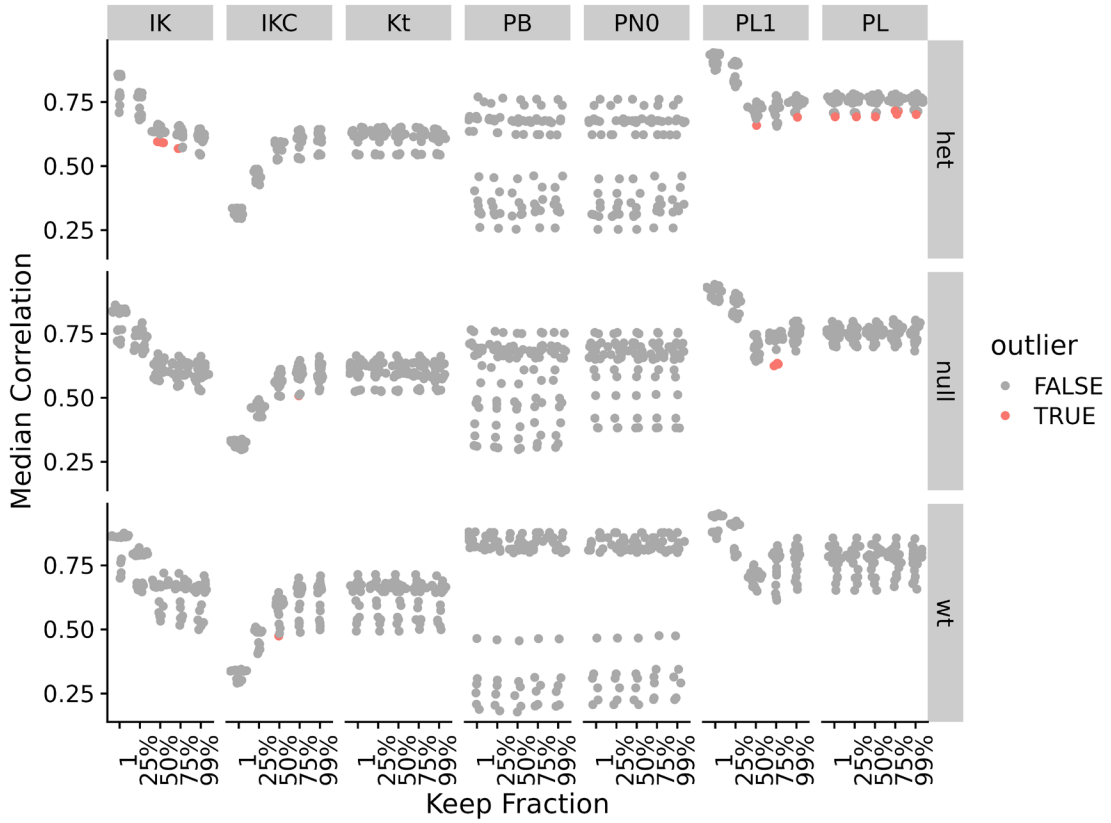

Figure S21. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

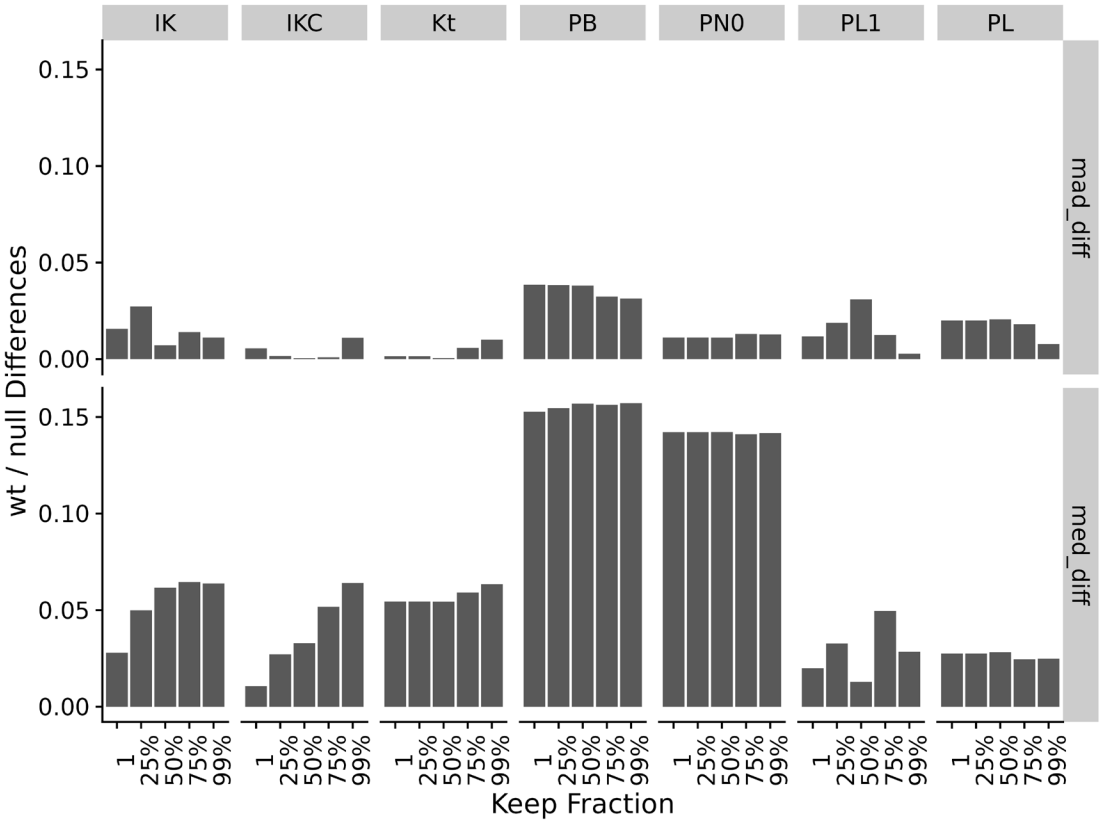

Figure S22. Difference between *wt* and *null* sample group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

Adenocarcinoma

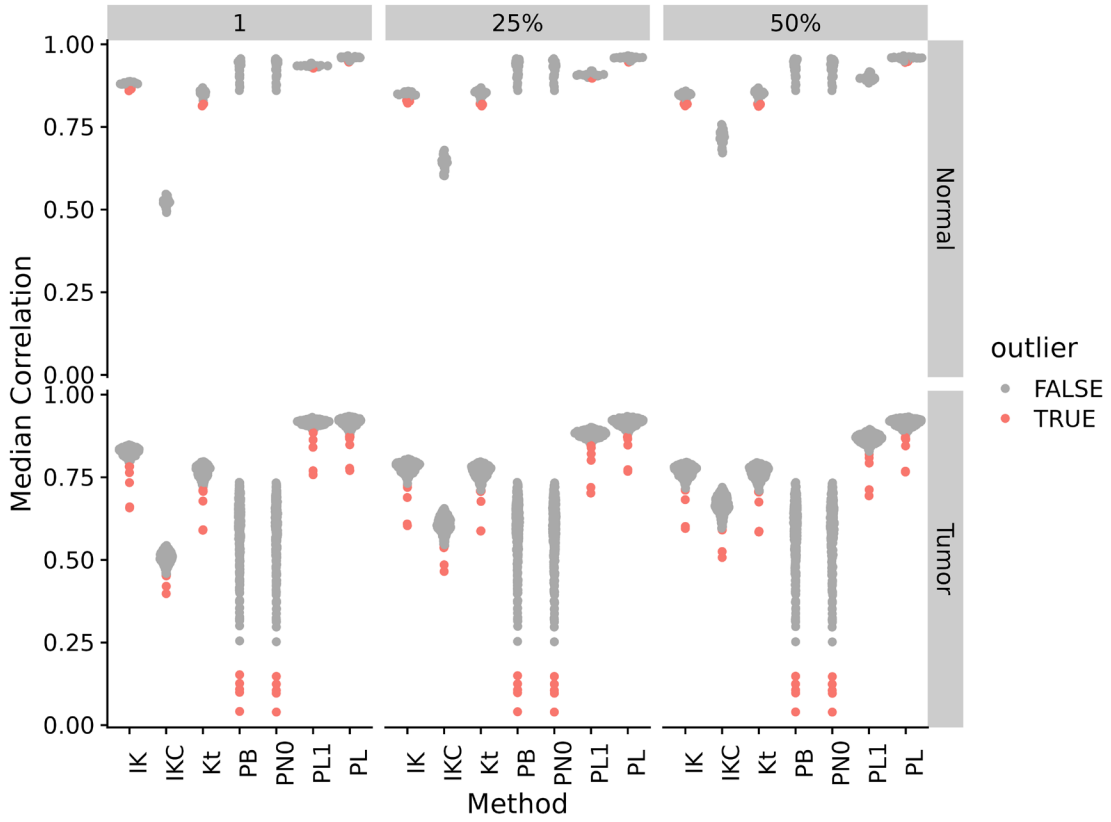

Figure S23. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

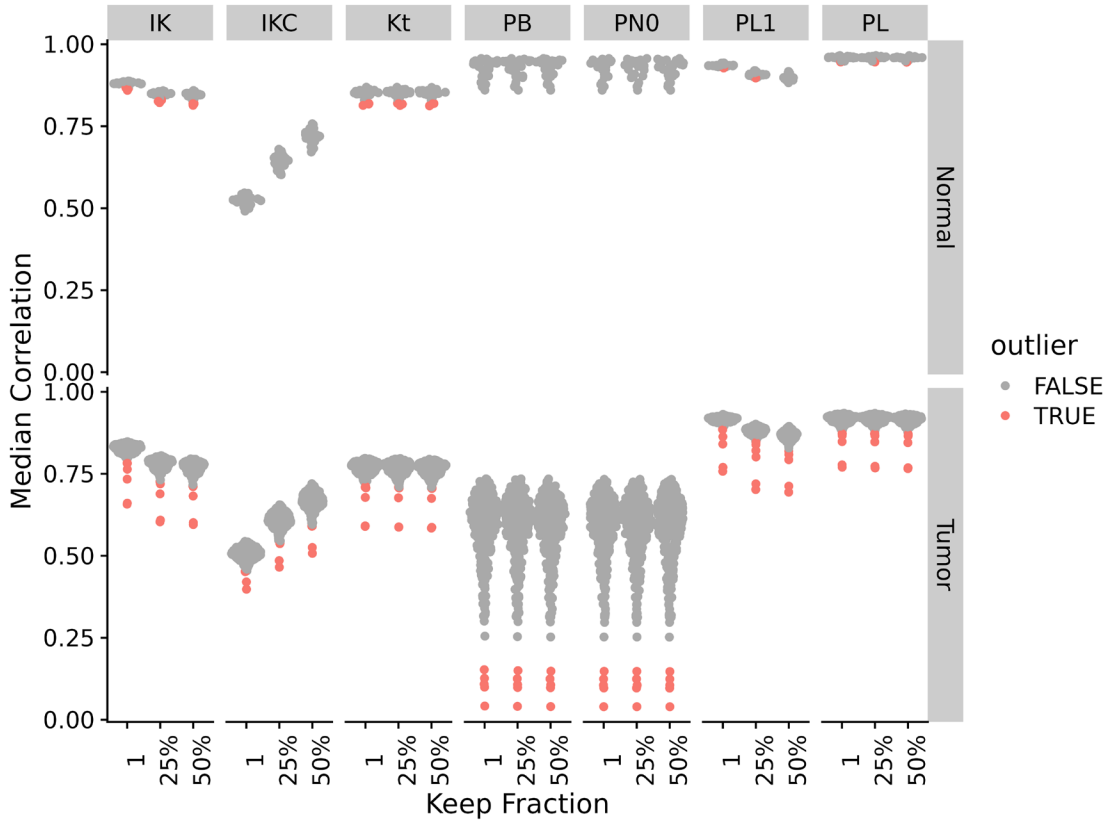

Figure S24. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

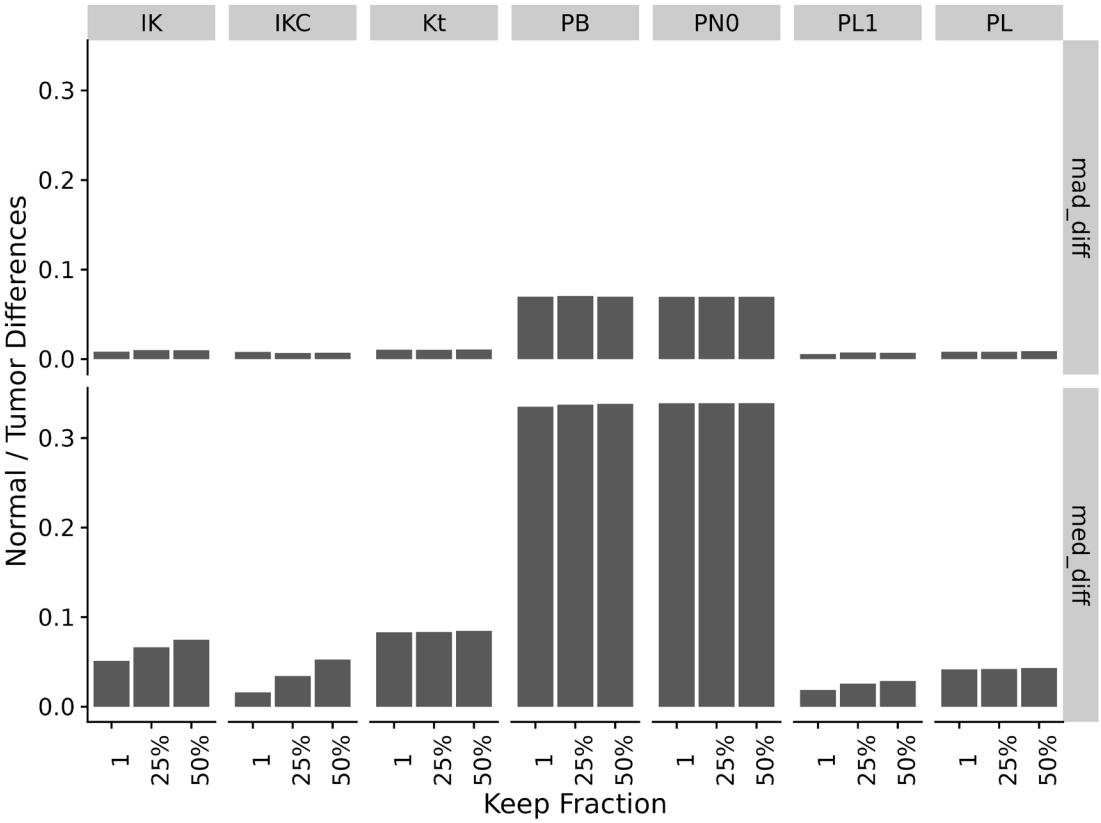

Figure S25. Difference of normal and tumor medians and MADs by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

NSCLC

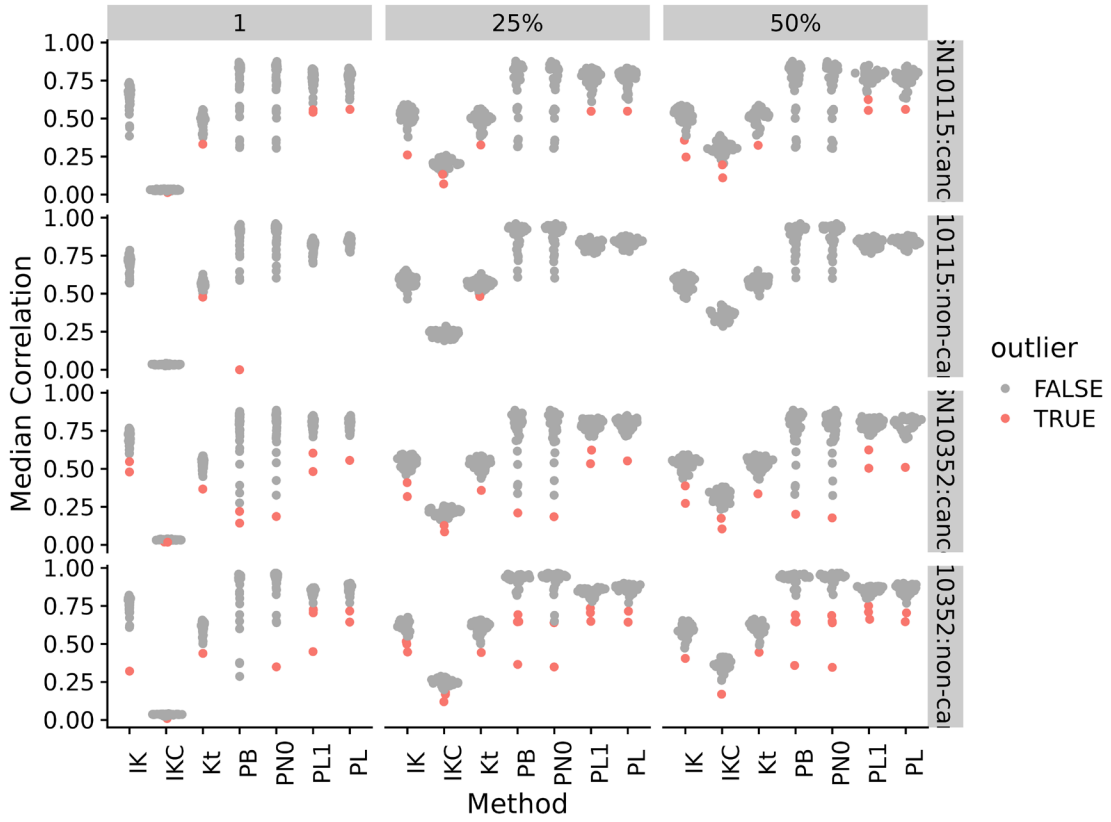

Figure S26. Median correlations by correlation method and applying different fractional cutoffs for NSCLC lipidomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

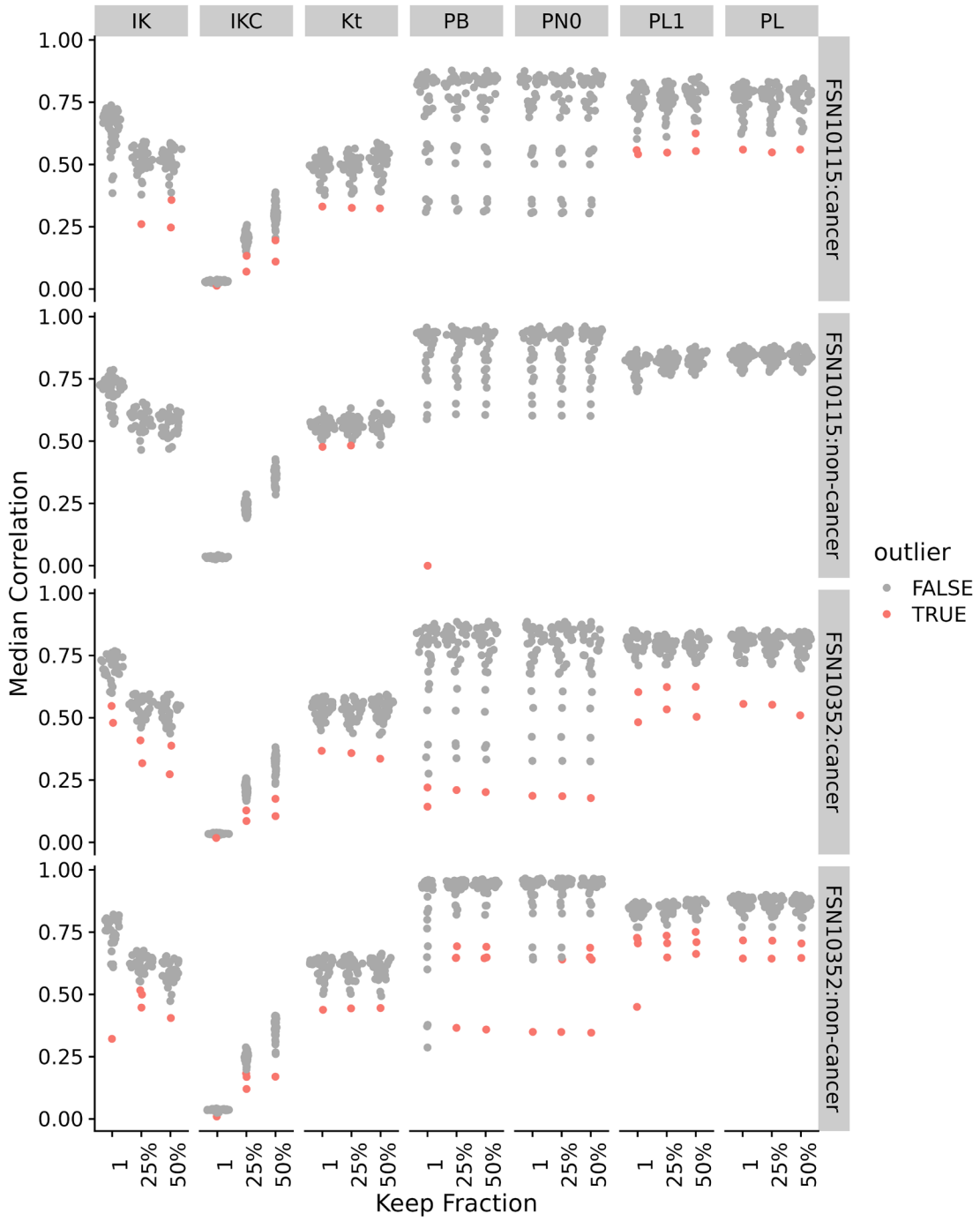

Figure S27. Median correlations by correlation method and applying different fractional cutoffs for NSCLC lipidomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

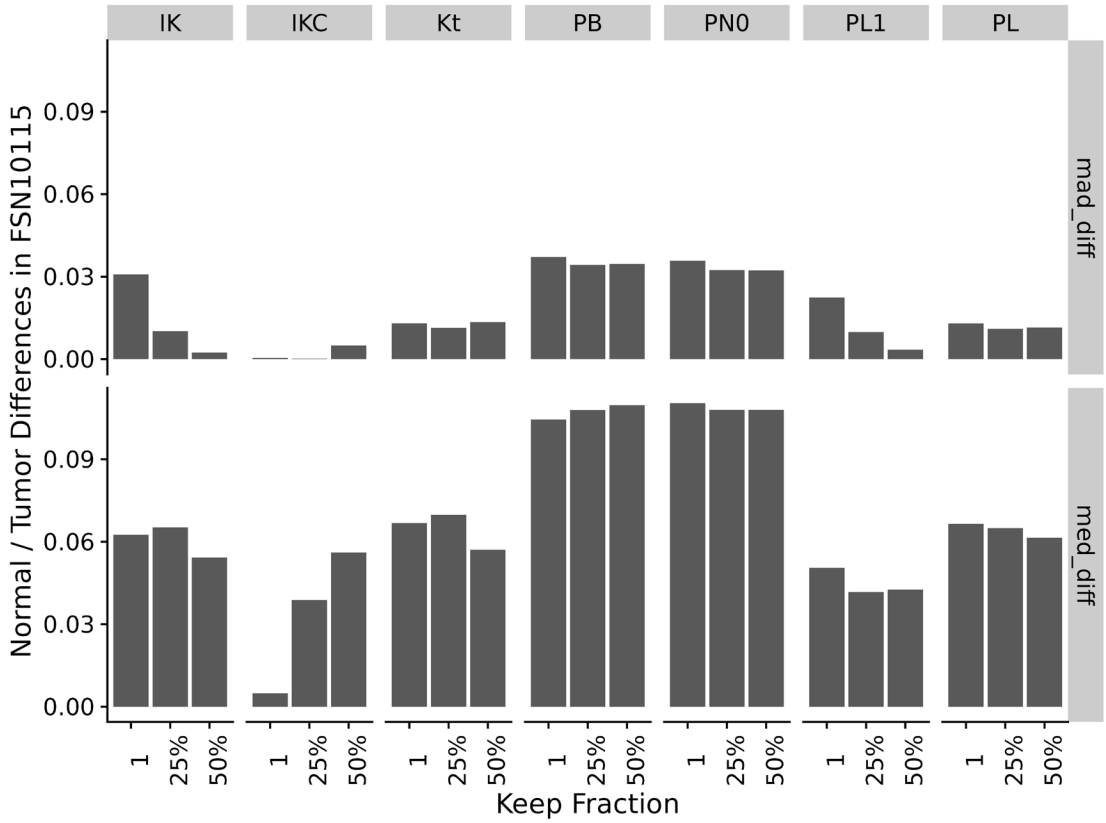

Figure S28. Difference of normal and tumor medians and MADs by correlation method and applying different fractional cutoffs for NSCLC lipidomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

Rat Stamina

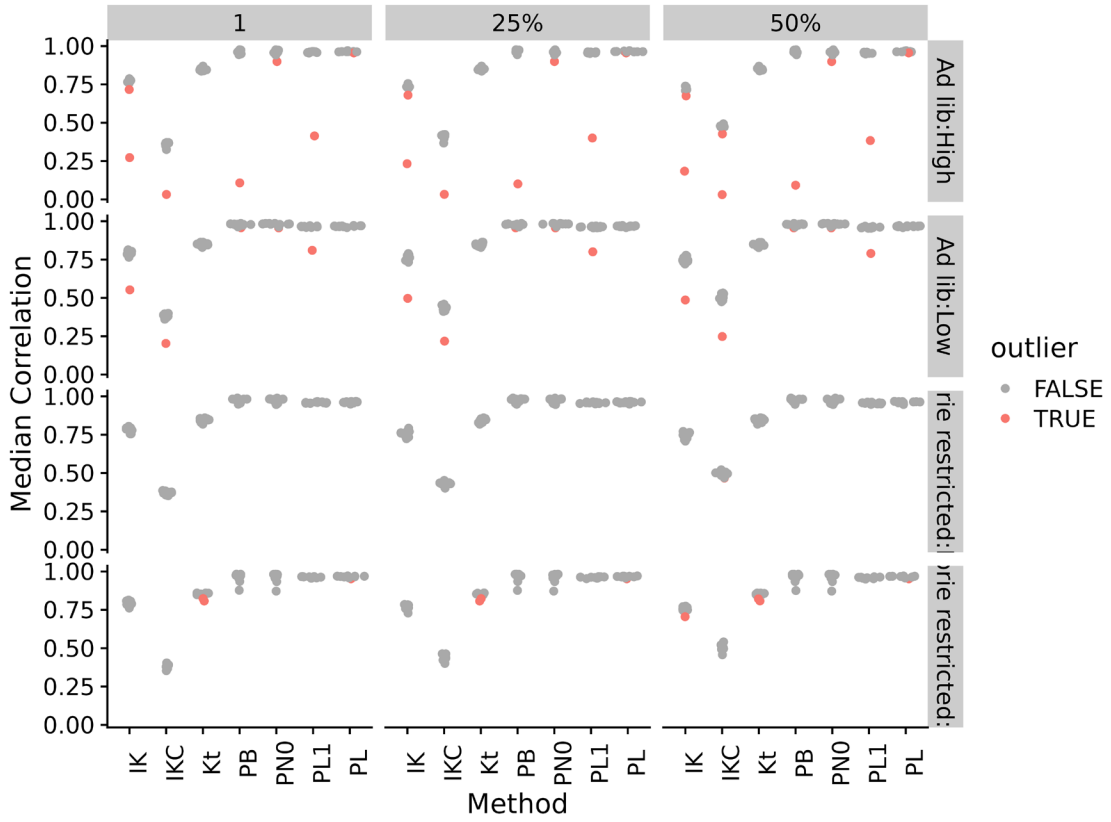

Figure S29. Median correlations by correlation method and applying different fractional cutoffs for rat stamina metabolomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

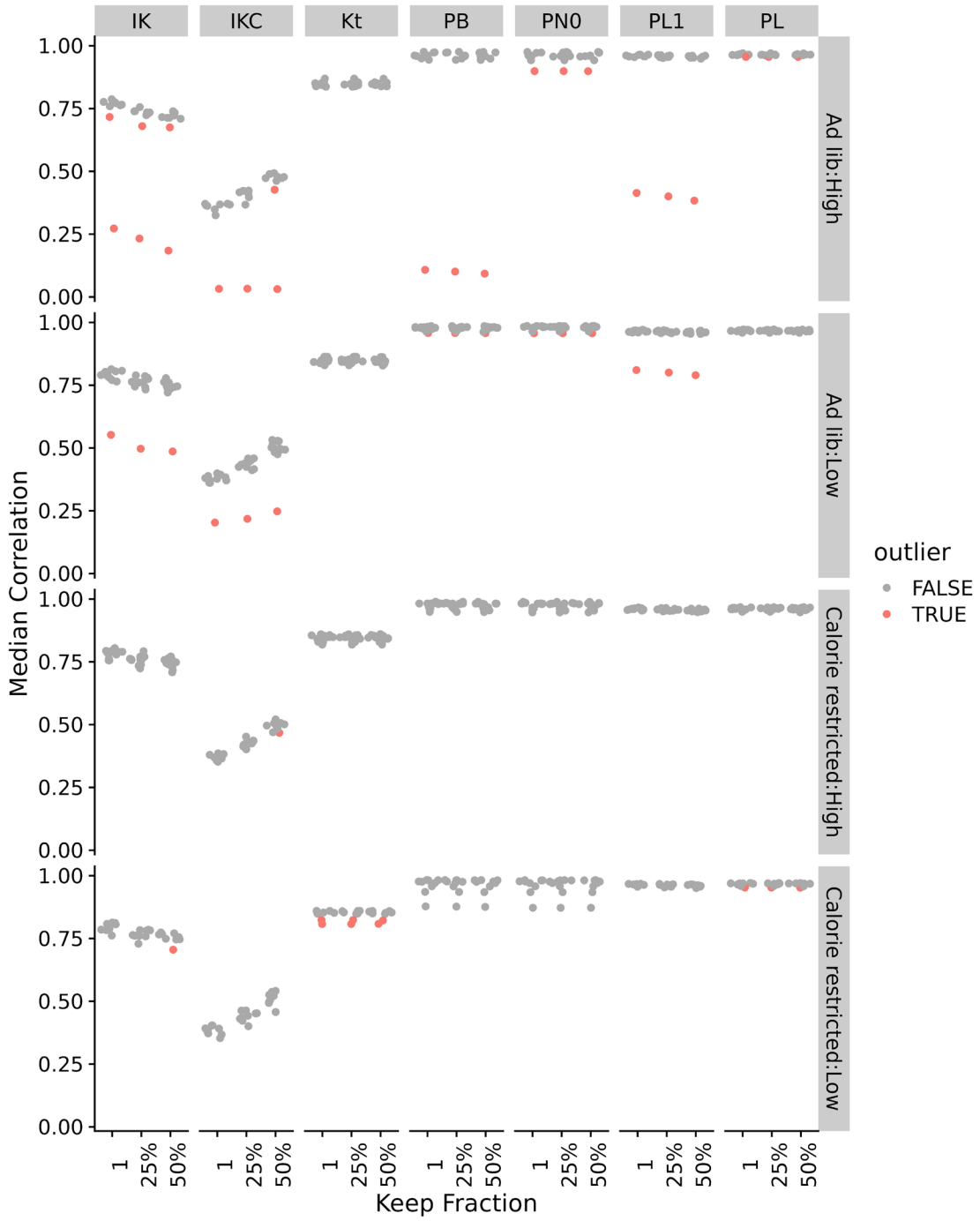

Figure S30. Median correlations by correlation method and applying different fractional cutoffs for rat stamina metabolomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

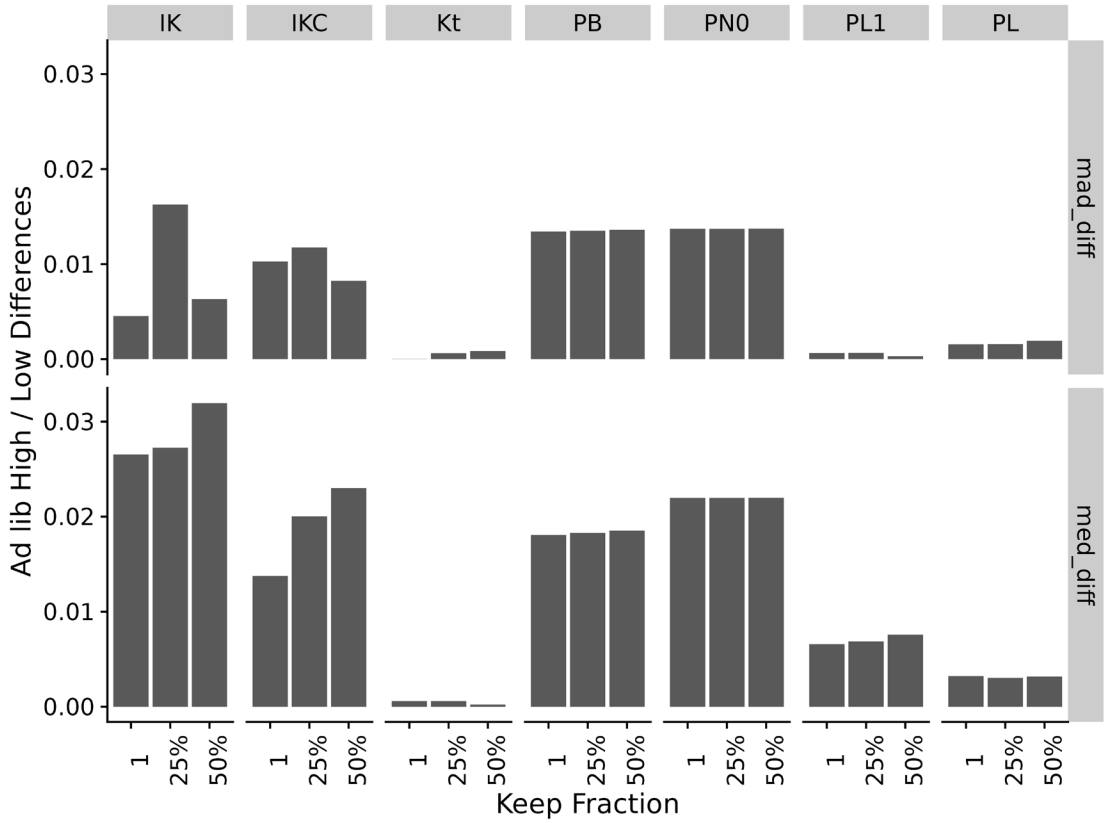

Figure S31. Difference of Ad lib feeding in high and low stamina medians and MADs by correlation method and applying different fractional cutoffs for rat stamina metabolomics data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt * Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

References