**Editor:**

**Issue 1:**

*First, the resubmission has format issues: (1) the PDF of the rebuttal does not display well (some mathematical fonts are missing?). Similar paragraphs correctly appear in the cover letter, but the reviewers only have access to the rebuttal, so they need a fully readable document. (2) A track-change version of the document is missing. It is customary to add it to the resubmission to facilitate the reviewers' work.*

**Response:**

We apologize for the corrupted fonts and equations in the rebuttal PDF. This has been fixed.

We had created a tracked changes version, but did not see a specific mechanism to upload it. We will upload the tracked changes versions as a supplementary document, if we cannot figure out another mechanism for uploading it.

**Issue 2:**

*Second and most importantly, the authors seem to have lightly accounted for one very important comment (although most of them have been addressed). Reviewer 2 made a very precise argument: the main methodological proposal according to the article (how to account for missing values while computing the Kt correlations) is equivalent to impute the missing values with a trivial MNAR mechanism, then to compute the Kt correlation, and finally to remove the imputed values (to keep the data unchanged and to avoid corrupting downstream processing). Reviewer 1 made a similar note, though in a softer way (Rev1-Iss1). The argument looks solid to me, but it does not hamper the publication potential as I acknowledge the vertue of application-oriented papers (to the benefit of exploring a specific biological dataset; or of providing nice packages to the community). However, the current discrepancies between the manuscript and the rebuttal pose a problem. Obviously, the reviewers base their evaluation on what the authors present, and the authors cannot ask for the reviewers to adopt a different perspective. Therefore, it is not possible to have an abstract that essentially promotes the theoretical novelty of the ICI-Kt, followed by its theory-oriented presentation, while rebutting things like "Some of our improvements are not rocket science [+ implementation-focused arguments]" (taken from response to Rev2-Iss7, but response to Rev2-Iss5, which stresses features other than ICI-Kt, is of the same taste); or like artificially stretching the difference between the proposed approach and the one based on imputation (see response to Rev1-Iss1, where the authors seem to consider that imputation necessarily alters the dataset, as if it were impossible to remove the imputed value after correlation computation). As the authors seem to acknowledge (based on their rebuttal) that a part of the theory presented is a bit overkill, I expect the paper to be restructured, as to feature results in accordance to what they answer to the reviewers.*

**Response:**

We understand the point being made by the reviewers. The base implementation is simple imputation of Kt. We discovered this implementation during the development of the software packages and realized that the ICI-Kt equations we had developed could be reformulated as simple imputation. However, this is only part of the ICI-Kt methodology, which is an information content interpretation of missing values due to left censorship. Thus, the ICI-Kt methodology includes the following parts:

1. Calculation of Kt using a left-censorship interpretation of missing values.
2. Validation of the presence of left-censorship with a statistical test.
3. The “local” and “global” scaling of ICI-Kt correlation values.
4. The tau\_max scaling of ICI-KT correlation values.
5. The total information content scaling of ICI-Kt correlation values by Completeness.

Given all of the feedback from the reviewers and editors, we believe it is best to frame the ICI-Kt methodology as an interpretation based on left censorship and information content. The software packages provide a set of tools that facilitate this interpretation. We have made extensive revisions to the Abstract, Introduction (now Background), and Discussion sections to frame ICI-Kt methodology from this perspective.

**Issue 3:**

Perhaps, a Software Article (https://bmcbioinformatics.biomedcentral.com/submission-guidelines/preparing-your-manuscript/software-article),

would be more suited to the features of the manuscript than a Research Article (https://bmcbioinformatics.biomedcentral.com/submission-guidelines/preparing-your-manuscript/research-article)?

To summarize, the manuscript will be sent back to the reviewers once (1) the manuscript is revised as to be in accordance with the reviewers' comments and responses; (2) the aforementionned format issues are dealt with. Beyond the imputation-equivalence issue, the general feeling of the reviewers is positive and as the majority of the comments have been scrupulously accounted for, I think finalizing the revision according to these expectations is worth it."

**Response:**

Based on the editor’s suggestion, we have reformatted the manuscript as a Software Article.

We thank the editor and reviewers for pushing us! We know the ICI-Kt methodology and software packages have high utility. We have been using them for years on many different datasets. We know they work. We have simply struggled to both present the methodology clearly and demonstrate that the software works. The new evaluations with respect to changes in dynamic range using simulated datasets was a huge improvement along these lines.

**Reviewer 1:**

*In the work “Information-Content-Informed Kendall-tau Correlation: Utilizing Missing Values”, Robert M Flight et al. propose a modified Kendall’s Tau-b statistic that can handle missing values under the assumption of missingness due to low abundance.*

*Correlation-based analyses are important to evaluate the strength of associations between different experimental conditions, as well as to construct interaction networks. However, there are very few correlation-based analyses that can handle missingness due to low abundance. Missingness due to low abundance is indeed an issue in many omics datasets. For instance, in mass-spectrometry-based proteomics, low-abundant and poorly ionizing peptides are more likely to go missing. Thus missing values are informative, and most existing methods that calculate correlations completely ignore missing values.*

*The method described by Flight et al. is – as far as I know – novel and addresses a real need as it provides a correlation estimator that can natively handle missingness due to low abundance. The authors found an elegant way to expand the Kendall’s Tau-b statistic by redefining concordant and discordant pairs to naturally incorporate missing values, under the assumption that missingness is due to left censoring. They provide both a local (which ignores points where both values are missing) and a global (which considers points where both values are missing as ties) version of their ICI-Kt correlation estimator.*

*Overall, the manuscript is well written and clearly structured. The fact that Flight et al. made their method available as both an R package and a Python module proves the efforts of the authors to ensure that their method can be adopted by a wide community of researchers in many different disciplines. Of note, even though the proposed method is an extension of the existing Kendall Tau-b, the implementation of their method in R is much faster than the base R cor “kendall” method. I also strongly appreciate the fact that the authors made all code and data openly available. For all these reasons, I would like to congratulate the authors for their hard work and recommend this work to be published in BMC Bioinformatics, provided that the authors would address my comments below.*

**Response:**

We thank the reviewer for their thorough review! Also, we appreciate the positive comments. We have put a lot of thought and effort into the development of ICI-Kt. We have similar perspectives to the reviewer in recognizing missingness in large datasets as a very important problem to solve.

**Issue 1:**

*Major comments*

*1. The authors write: “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.” I was wondering why they decided not to include this method in their comparison? I think it might be relevant for the readers to see how the new ICI-Kt estimator performs versus the Kendall Tau-b estimator where the missing values were imputed first with a value lower than the lowest value in the dataset.*

**Response:**

This comparison was presented in the manuscript, with the text:

“The normal Kendall tau correlation replacing missing with 0 has identical behavior to the ICI-Kt correlation.”

In Supplemental Figure S11, panels A and C, bottom panels, where one can observe a straight line of values between Kendall-0 (0 being less than all values in the data) and ICI-Kt. We have added further text to highlight this:

“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 the bottom panels of Figure S11A and Figure S11C.”

Although directly imputing with a minimum value is similar, we have been striving to make the point that this approach to handling missingness directly is distinct from imputation. Also, imputation often does not work well with other statistical methods and leads to problematic interpretations, which is why we implemented our R and Python packages to take datasets with missing values without adulterating the original data with imputed values. Moreover, imputation in general does not work well for analytical techniques where the detecting limit varies across analytical sample. We have added the following comment to the Introduction to emphasize this limitation with imputation methods.

“Imputation methods rely on very similar analytical detection limits between analytical samples. When this condition does not hold, imputation methods have reduced performance and lower interpretive value. For analytical techniques requiring complex sample handling and detection, the variability in the analytical detection level can be quite high.”

**Issue 2:**

*2. In addition to point (1), I think it would also be relevant to compare the ICI-Kt estimator to an analysis where the data is imputed first, and the correlation is calculated based on the imputed data. More specifically, I think they should include at least one imputation method where data is imputed under a missingness at random assumption, and at least one existing imputation method that assumes missingness due to low abundance.*

*Examples of the former include:*

*- k-nearest neighbors imputation*

*- Regularized Expectation-Maximization (RegEM) imputation assumes missingness at random. It uses ridge regression to regress the missing values on the observed values and fits these models with an iterative expectation-maximization (EM) algorithm until convergence. Missing values are imputed with the models’ estimates. RegEM imputation has been implemented in MatLab. The script can be downloaded from https://climate-dynamics.org/software/#regem.*

*Examples of the latter include:*

*- Imputation as implemented in the proteomics software Perseus: it first constructs a rescaled normal distribution with: (1) a downshifted mean equal to the average of all the observed data minus 𝑑 times the standard deviation of the observed data and (2) a standard deviation equal to 𝑤 times the standard deviation of the observed data. Missing values are then imputed with random draws from this rescaled distribution. The default values for 𝑤 and 𝑑 are 0.3 and 1.8, respectively.*

*- Imputation as implemented in the proteomics software MSstats: missing values are imputed with an Accelerated Failure Time (AFT) model.*

*- Imputation as implemented in the “Differential Enrichment analysis of Proteomics data” (DEP) R package, which imputes certain missing values under a missingness-by-low-abundance assumption and other missing values under an abundance-independent missingness assumption. More specifically, mixed imputation imputes features with missing values in all replicates of at least one experimental factor with “MinProb” imputation from the imputeLCMD R. MinProb constructs normal distributions with (1) downshifted means equal to the lower 1% quantile per sample and (2) standard deviations equal to the median of all peptide-wise standard deviations with less than 50% missing values. Missing values are imputed with random draws from this distribution. Missing values for features which have been identified at least once in all experimental factors are imputed with kNN imputation.*

*I realize implementing two additional methods (one abundance-independent and one abundance-independent) requires some extra work, but I am only asking to test them on a ground-truth dataset like the analysis in Fig. 3 or a modified version of this analysis (see my comment (4.) below).*

*https://github.com/statOmics/MSqRobHurdlePaper/tree/master/analyses/CPTAC is a repository where many of these methods have been tested, which could perhaps help for a quick implementation of two imputation methods.*

**Response:**

This is similar to the previous issue. We have added the basic imputation with a value ½ the lowest observed value in a dataset, where the dataset has low, medium, or high variability in the dynamic range across samples. See the new section “Differences in Dynamic Range and Correlation”, text, and new Figure 5.

“Another way that missing values appear is due to changes in dynamic range between samples, as some samples have features with higher values, and the fixed dynamic range of the instrumentation results in features with lower values to be “missing” in those samples. We created a set of 100 simulated samples with uniform noise on the log-scale, with relatively constant dynamic ranges, and introduced changes to the overall dynamic range using a random censor, at varying levels (see Methods). Possible different levels of censoring based on dynamic range were checked by first determining how many missing values would be introduced in each sample as the dynamic range was increased in increments of 0.1 (see Figure S12). Based on the number of values being censored, limits of 0.5, 1, and 1.5 were selected, representing low, medium, and high variability in dynamic range.

For each level of possible missingness introduced by changes to the dynamic range, correlation across all samples were calculated using all values (reference), as well as after missingness was added (trimmed), and using Pearson correlation with global imputation (Pearson Imputed) or ICI-Kt. Figure 5 demonstrates that it is only as the number of of missing values in one of the samples approaches 50% or more (500 of 1000 features) does the Pearson correlation with global imputation give correlation values closer to the known correlation with no missing values in any appreciable amount (points below the red lines in the top panels, and to the right of the red line in the histograms in the bottom panels). Points above the lines with slopes of -1 and 1 indicate that the difference of reference - trimmed is smaller in the ICI-Kt correlations, and below the lines indicate the difference is larger in the ICI-Kt correlations. This is further emphasized by the majority of the values are to the left of the red line at 0 in the difference histograms.”

**Issue 3:**

*3. The mathematical notations are not always clear to me. For example, I assume that |𝑝𝑎𝑖𝑟𝑠𝑐𝑜𝑛𝑐𝑜𝑟𝑑𝑎𝑛𝑡| refers to the number of concordant pairs. However, the use of “|” makes me think that this might also refer to an absolute value. If this is a indeed a number, I would also use “𝑛” with an appropriate index for clarity. Similarly, for clarity, it might also be a good idea to explicitly define the following symbols the first time they appear in the text: 𝑛0 (total number of pairs?), 𝑛𝑥𝑡𝑖𝑒 (number of ties in all (𝑥𝑖,𝑥𝑗) combinations?), 𝑛𝑦𝑡𝑖𝑒 (number of ties in all (𝑦𝑖,𝑦𝑗) combinations?), 𝑛𝑡𝑖𝑒 (total number of ties?).*

*I am also not sure about this formula: 𝑝𝑎𝑖𝑟𝑠𝑚𝑎𝑥𝑐𝑜𝑛𝑐𝑜𝑟𝑑𝑎𝑛𝑡=𝑛0−𝑛𝑥𝑡𝑖𝑒−𝑛𝑦𝑡𝑖𝑒−𝑛𝑡𝑖𝑒*

*If I take as an example two vectors 𝑋=[1,2,3] and 𝑌=[2,4,4], then I count two concordant pairs ((𝑋1,𝑌1),(𝑋2,𝑌2) and (𝑋1,𝑌1),(𝑋3,𝑌3)) and one tie due to equal y-values ((𝑋2,𝑌2),(𝑋3,𝑌3)). Using the formula, I get: 3-0-1-1 = 1, where I would expect 2.*

*Also, in the following formula: 𝑐𝑜𝑚𝑝=𝑛𝑓𝑒𝑎𝑡−(𝑚𝑖𝑠𝑠𝑖|𝑚𝑖𝑠𝑠𝑗)𝑛𝑓𝑒𝑎𝑡*

*Please clarify in the text that you are calculating the completeness index 𝑐𝑜𝑚𝑝 between samples 𝑖 and 𝑗, and that 𝑚𝑖𝑠𝑠𝑖 and 𝑚𝑖𝑠𝑠𝑗 are the numbers of missing features in these respective samples. Also please clarify that “|” refers to the “or” operator; in statistics, this could also refer to a conditional probability, which could cause confusion.*

*Is 𝑛𝑓𝑒𝑎𝑡 the total number of features identified in at least on of both samples or the total number of features identified in the dataset?*

**Response:**

The use of vertical lines on both sides of a set variable is standard notation for cardinality. See <https://en.wikipedia.org/wiki/Cardinality> . However, we agree that this will likely be confusing for other readers as well, and have modified the various formulas to be more clear, and have added descriptions of values in the accompanying text as well.

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

Where is the the number of concordant pairs and is the number of discordant pairs.”;

“However, the related tau-b statistic does handle the presence of tied values by adding the tied and values to the denominator, and in our special case of missing data, we can add the ties that result from both of and being missing into and as well.

Where is the total number of pairs, are the number of tied values in X, and are the number of paired values in Y.”

“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:

Where is the number of commonly tied values in both X and Y. Calculating a set of values between all experimental samples, we can take the maximum of the values, and use it to scale **all** of the obtained Kendall-tau values equally ().”

Also, thank you for catching the error in the description of the “pairsmax\_concordant” value, the last minus sign should have been a plus, and has been corrected in the manuscript (see formula above).

Thanks for pointing out the use of a single vertical line between two sets to represent the union of the two sets. This is nonstandard notation and has been replaced with the standard or symbol.

“ “

Where for any two samples *i* and *j*, is the total number of features or entries, and are the features missing in either sample *i* or *j*, with being the total number of missing entries in either sample *i* or *j*.”

**Issue 4:**

*4. Figure 3: what is the true correlation in the complete dataset? I would suggest to add this to each panel. Would it be possible to somehow indicate the number of missing values that has been introduced for each dot (e.g. using a color scale)? I would expect the stability of the correlation estimate in the light of removing more and more values to be a desirable property. However, here it seems that when more missingness in the lower range is introduced, the classic estimators of correlation are more stable than the ICI-Kt estimator, so to me Fig. 3 would suggest that one should prefer the classic estimators over the new ICI-Kt estimator as the classic estimators seem to be more stable.*

*However, the result from Fig. 3 seems somewhat odd to me. Consider the following R script:*

*library(ICIKendallTau)*

*### Generate some data and calculate correlations ###*

*correlations <- vector(mode = "list", length = 3)*

*names(correlations) <- c("complete", "missing", "imputed")*

*correlations <- lapply(correlations, function(x){*

*x <- vector(mode = "list", length = 4)*

*names(x) <- c("pearson", "spearman", "kendall", "ICI\_Kt")*

*return(x)*

*})*

*set.seed(1713314672)*

*for(m in 1:1000){*

*for(i in 1:3){*

*x <- 1:1000+rnorm(1000)\*5000*

*y <- 2\*x+5+rnorm(1000)\*5000*

*df <- data.frame(x = x,*

*y = y)*

*if(i == 1){*

*data <- df*

*} else if(i == 2){*

*data <- df*

*data$x[data$x < -5000] <- NA*

*data$y[data$y < -1000] <- NA*

*} else if(i == 3){*

*data <- df*

*data$x[data$x < -5000] <- -1e99*

*data$y[data$y < -1000] <- -1e99*

*}*

*for(k in 1:length(correlations[[i]])){*

*if(k != 4){*

*correlations[[i]][[k]] <- c(correlations[[i]][[k]], cor(na.omit(data)$x, na.omit(data)$y, method = names(correlations[[i]])[k]))*

*} else{*

*correlations[[i]][[k]] <- c(correlations[[i]][[k]], ici\_kendalltau(data, global\_na = NA)$cor['x', 'y'])*

*}*

*}*

*}*

*}*

*median(correlations$complete$pearson)*

*# 0.8945951*

*median(correlations$complete$spearman)*

*# 0.8847967*

*median(correlations$complete$kendall)*

*# 0.7047287*

*median(correlations$complete$ICI\_Kt)*

*# 0.7047287*

*median(correlations$missing$pearson)*

*# 0.7875461*

*median(correlations$missing$spearman)*

*# 0.7526054*

*median(correlations$missing$kendall)*

*# 0.5574929*

*median(correlations$missing$ICI\_Kt)*

*# 0.7476337*

*median(correlations$imputed$pearson)*

*# 0.4495594*

*median(correlations$imputed$spearman)*

*# 0.8451884*

*median(correlations$imputed$kendall)*

*# 0.6814186*

*median(correlations$imputed$ICI\_Kt)*

*# 0.748195*

*Based on this analysis, Spearman and Kendall correlation estimators that impute missing values with -1\*1099 actually seems to approach the original correlation much closer than the estimators that only include the complete cases, and hence looks more stable than the classic estimators. The same goes for the ICI-Kt estimator, whose estimate (0.74) also remains relatively close to the original correlation of 0.70, which is good. Would it be possible to either improve the explanation of the analysis in Fig. 3, and/or add an analysis in which, for example, missing values are being introduced progressively under the assumption of missingness by low abundance, while knowing the ground truth? Given the previous, I would expect the ICI-Kt estimator to perform very well in such an analysis.*

**Response:**

We believe the analysis you are asking for is actually handled in “Effect of Left Censoring VS Random Missing Data”, and Figure 4, where we do sequentially introduce missingness in the low end. Figure 3 is about introducing missingness randomly up to the lower 50% of the values. We have added the full sample correlation to the figure subtitle for Figure 3, and added to the figure caption, and added Supplemental Figure S10, where we plot the difference to the base sample sample correlation of 1 or -1. Both Figure 3 and Figure S10 now have points colored by how many values are missing.

We have also added the “true” correlation to the plots in Figure 4.

We believe the code provided above is incorrect, as the random sample generation should be done outside the loops over non-missing, missingness, or imputation, so that each introduction of missingness and calculation are done on the same sample. As mentioned above, we think the analysis of introducing a particular amount missingness between two samples with some variance is covered by the analysis presented in Figure 4.

“withr::local\_seed(1713314672)

### Generate some data and calculate correlations ###

correlations <- vector(mode = "list", length = 3)

names(correlations) <- c("complete", "missing", "imputed")

correlations <- lapply(correlations, function(x){

x <- list(pearson = vector("numeric", 1000),

spearman = vector("numeric", 1000),

kendall = vector("numeric", 1000),

ICI\_Kt = vector("numeric", 1000))

return(x)

})

#set.seed(1713314672)

for (m in 1:1000) {

#m = 1

x <- 1:1000 + rnorm(1000) \* 5000

y <- 2 \* x + 5 + rnorm(1000) \* 5000

df <- data.frame(x = x,

y = y)

for (i in 1:3) {

#i = 1

if (i == 1) {

data <- df

} else if (i == 2) {

data <- df

data$x[data$x < -5000] <- NA

data$y[data$y < -1000] <- NA

} else if (i == 3) {

data <- df

data$x[data$x < -5000] <- -1e99

data$y[data$y < -1000] <- -1e99

}

for (k in 1:length(correlations[[i]])) {

if (k %in% c(1, 2)) {

correlations[[i]][[k]][m] <- cor(na.omit(data)$x, na.omit(data)$y, method =

names(correlations[[i]])[k])

} else if (k == 3) {

# for the cases examined, this should be completely equivalent to method = "kendall", but much faster

correlations[[i]][[k]][m] <- ici\_kt(na.omit(data)$x, na.omit(data)$y)[1]

} else if (k == 4) {

correlations[[i]][[k]][m] <- ici\_kendalltau(t(data), global\_na = NA)$cor['x', 'y']

}

}

}

}

**Issue 5:**

*5. The authors write: “In contrast to ICI-Kt, the Pearson correlation calculated using only pairwise complete entries is constant over growing left-centered and random missingness.” However, in Fig. 4 panel A and B, I see a respective decrease and increase in correlation with increasing missingness. Can the authors clarify this?*

**Response:**

We meant to say this is “practically constant”, since the range is 0.004 or smaller. We just left out a clarifying adjective. We have revised as follows:

“In contrast to ICI-Kt, the Pearson correlation calculated using only pairwise complete entries is practically constant (i.e., range 0.004 or less) over growing left-centered and random missingness.”

**Issue 6:**

*Minor comments*

*- The explanation in the section “Number of Non-Missing Values and Median Abundance” is not very clear to me. As I understand it, the authors calculate the median value for each feature within each experimental factor. Then they record the minimum median value for each combination of experimental condition x number of features identified. It is not clear what is then being done with this value. The description that follows is also not clear to me. As I understand it, the authors first take the log10(median) for each experimental condition x feature combination. Then they say they take the median of these previous values. In my mind, this leave one single value per experimental condition. How is it possible to calculate a correlation based on only a single pair of values? I also do not understand the sentence “Additionally, the Kendall-tau correlation of the minimum median and number of present values was calculated.”. Can the authors clarify what they did here?*

**Response:**

We can see how this explanation is confusing. As an alternative that still demonstrates the issue, we have used the median rank across present samples and the number of missing values, and modified the methods and results:

“For each dataset, the samples were split by experimental factor or treatment of interest (see previous Methods for each dataset). For each feature, the rank of the feature was calculated for each sample where the feature was present, followed by the features median rank, as well as the number of samples the feature was missing from. Grouping the features by the number of missing values, we calculate the median of median ranks, as well as the minimum of median ranks for the visualization and correlation of the relationship of rank with missing values.”

**Issue 7:**

*- The authors write: “As we are interested in only those correlations at the low end of correlation (becoming the high end after a log-transform)”. However, a log-transform should keep the order of the values. Why do correlations at the low end become correlations at the high end after log-transformation?*

**Response:**

Thank you for catching this. The description is incorrect. We have updated the description of the outlier detection:

“For outlier detection, median sample-sample correlations within the sample treatment (genotype, condition, etc) is calculated, and 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 the subtraction and log-transform), we restrict to only those entries at the *high* end of the score distribution (using *visualizationQualityControl::determine\_outliers* [44]). This is equivalent to using the correlation component of the score described by Gierliński et al [14] and setting the other component weights to zero.”

**Issue 8:**

*- The authors write “we trimmed to features present in 25% or more of any of the sample classes”. I was wondering why the authors are imposing this trimming? I would expect that one of the advantages of the ICI-Kt estimator is that it can better deal with missing values caused by low abundance of the true unobserved, underlying value. I would assume that features with a lot of missing values are even more likely to be affected by left censoring. Hence, it would be interesting to also see the results of this analysis without trimming (I expect them to be even better). If they indeed are better, I think the authors might have a case to claim that trimming features with an arbitrary threshold for missing values might not be needed anymore thanks to their estimator. To me, it seems that this would increase the impact of their new estimator, but I leave it up to the authors if they would like to include this analysis or not.*

**Response:**

This is an issue of information content. Features with very low information content will just add noise. We do not believe that changing the inclusion cutoff of features in the feature – feature partitioning will change the results, the cutoff is not salient to the point we are trying to make, and the calculation of the feature – feature correlations at different cutoffs would require too much time to perform in time for revising the manuscript.

**Issue 9:**

*- One thing I wonder is: how big is the problem of missing values in each of the datasets presented? The authors should include the number and the percentages of missing values in each dataset in Table 1.*

**Response:**

This is a very good suggestion! Thank you. We have added Table 2, describing the number of values, number and percentage of missing values, as well as the results of a binomial test that tests if the missing values are likely to be the result of left-censorship.

**Issue 10:**

*- I noticed that the authors only included RNA-Seq, metabolomic and lipidomic datasets in their analysis, but they did not include proteomics. This is only a suggestion, but to increase the impact of this work, the authors could consider including a (label-free) mass-spectrometry-based proteomics dataset as well. The reason for this suggestion is that, as far as I know, missingness due to low abundance is much more of an issue in mass-spectrometry-based proteomics as compared to RNA-Seq. Furthermore, the reasons for missingness in label-free mass spectrometry-based proteomics have been discussed extensively (see e.g. PMIDs: 25855118, 26906401, 32227882), and various imputation methods that not always capture missingness not at random are still the gold standard for dealing with missing values in label-free MS-based data. Proving that their method also improves correlation estimates in MS-based proteomics could have a positive impact on the proteomics community and widen the impact of this work. However, such an analysis is not essential for the quality of the manuscript, so I only leave it as a suggestion to the authors.*

**Response:**

We have included 5 different datasets (3 datasets are variants where the treatment is indicated differently), which is more than most methods manuscripts of this type. Also, metabolomics and lipidomics datasets are both mass-spectrometry generated. But your point is valid about missingness in proteomics datasets. We are analyzing a proteomics dataset right now and used ICI-Kt for QA/QC, but we cannot include it until our collaborator publishes their results. On the time frame of this revision, we will keep to the current datasets.

**Issue 11:**

*- The authors write: “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).”: Is there a figure that shows this?*

**Response:**

Thank you for catching that we missed references to some of the supplemental figures. We have corrected this and added the references to the supplemental figures:

“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, as shown in Figures S21, S24, S27, S30 and S33 (see Methods).

**Issue 12:**

*- Fig. 4: It might be more informative to make to scales of the vertical axis equal in all figure panels. If this would make the changes invisible for some panels, the authors could for example include the current figure panels as an inset in these cases.*

**Response:**

We generated these figures on similar scales and it simply hides too many trends. We also have 15 graphs in this figure as it is, so adding insets would make it extremely busy. However, we have updated the figure legend and text to highlight the ranges.

“Pay attention to the different y-axis ranges across graphs, with A and B graphs having much smaller y-axis ranges.”

**Issue 13:**

*- Fig. 5 B: please clarify in the legend that this is for your ICI-Kt method.*

**Response:**

We changed Figure 5B figure legend as follows:

“Log of time taken to calculate all pairwise comparisons using ICI-Kt for different numbers of samples in multicore processing.”

**Issue 14:**

*- Table 2 and 3: why are some values shown in bold and others not?*

**Response:**

We have added the following to the table caption in Table 2 (now Table 3) to explain what the bolding means:

“Bolded entries indicate the sample was an outlier for that method.”

We have added the following to the table caption in Table 3 (now Table 4) to explain what the bolding means:

“Bolded entries indicate the highest partitioning ratio for the dataset.”

**Issue 15:**

*- Just a follow-up thought: if one could somehow get an estimate the fraction of values missing due to low abundance and the fraction of values missing at random, would it be possible to further improve your estimator? This could perhaps be interesting for a follow-up paper?*

**Response:**

That is an interesting idea. We have developed and added a statistical test for left censorship to the R and Python packages, after we had submitted this manuscript. We wanted a statistical test to determine when left censorship was an issue in a dataset or not. We applied the left censorship test to each dataset, and have added Table 2, with the total number of data entries in each dataset, the number and percent missing values, and the number of values that are “likely” missing due to left censorship. We have also included a section in the methods describing the statistical test for left censorship:

“**Binomial Test for Left Censorship**

For each dataset, the samples were first split by treatment. In each sample, the median abundance of features present in the sample are calculated. For any feature that is missing in any sample, the values in the present samples are compared to the median values in each sample. If the value is **less than** the median value in the sample, that is counted as a **success** in a binomial test, otherwise it is counted as a failuer. The number of successes and failures are aggregated across the treatment splits for calculation in a binomial test, with the null hypothesis as a ratio of 0.5.”

And describe them in the Results section:

“Additional support for the hypothesis that missing values are **primarily** due to being below the limit of detection is provided by the binomial test for left censorship (see Methods), with the results shown in Table 2. This test involves counting whether the non-missing values (for those features that are missing in at least one sample) are below the median of the samples where they are non-missing. For many of the datasets, the proportion of non-missing values below the median are in the range of 0.9 or higher, and only the two mass-spectrometry datasets (ratstamina and nsclc) show lower estimates. Even in those two specific cases, the proportions are 0.69 and 0.60, respectively.”

**Issue 16:**

*Spelling suggestions*

*- WT is the abbreviation of wild type. Contrary to “Snf2”, which is a gene name and hence should be in italics, I would not put “WT” in italics.*

**Response:**

Thanks for the catch! Fixed.

**Issue 17:**

*- “we only considered peaks matched that had a single assigned and a single un-assigned peak match together.” -> “we only considered peaks matched that had a single assigned or a single un-assigned peak match together.”?*

**Response:**

Thanks for the catch! Fixed as follows:

“For lipid category and class voting, we only considered peaks as matched when a single assigned or a single un-assigned peak match together, i.e. satisfying a bijection criterion.”

**Issue 18:**

*- “organism specific annotation packages” -> “organism-specific annotation packages”*

*- Fig. 6: “any of the method” -> “any of the methods”*

*- “rank based statistics” -> “rank-based statistics”*

**Response:**

Again, thanks for the catches! All fixed.

**Reviewer 2**

*Dealing with missing data is a very important topic in scientific research. It is always positive to see efforts to improve the approaches in the literature. Unfortunately I was not able to see a difference between the proposal here and the very standard idea that is commonly used for the type of missing data that is tackled. Perhaps I have missed something, but if so, then the explanations would require a great improvement, because the current version left me without doubt that this is already the very simplistic approach most people use. I would be glad to rethink it in the future.*

**Response:**

Confusing our methodology with imputation is common. However, there are clear distinctions. Imputation, especially simple imputation, is based on the assumption that missingness has similarities across analytical samples. While this holds for some datasets, this assumption starts to break down when the level of left censorship is variable across analytical samples. Most experimentalists would view this as variability in the dynamic range of detection across samples, which is common for complex analytical techniques often used in omics experiments. In these cases, ICI-Kt is “safer” for interpretation than the use of simple imputation. We have added this description to the Introduction.

“Imputation methods rely on very similar analytical detection limits between analytical samples. When this condition does not hold, imputation methods have reduced performance and lower interpretive value. For analytical techniques requiring complex sample handling and detection, the variability in the analytical detection level can be quite high.” We have also created a new simulated dataset with different levels of variability in dynamic range across 100 samples. With this dataset, we clearly show ICI-KT to be more robust than Pearson correlation with global imputation.

“Another way that missing values appear is due to changes in dynamic range between samples, as some samples have features with higher values, and the fixed dynamic range of the instrumentation results in features with lower values to be “missing” in those samples. We created a set of 100 simulated samples with uniform noise on the log-scale, with relatively constant dynamic ranges, and introduced changes to the overall dynamic range using a random censor at varying levels (see Methods). Possible different levels of censoring based on dynamic range were checked by first determining how many missing values would be introduced in each sample as the dynamic range was increased in increments of 0.1 (see Figure S12). Based on the number of values being censored, limits of 0.5, 1, and 1.5 were selected, representing low, medium and high variability of the dynamic range.

For each level of possible missingness introduced by changes to the dynamic range, correlation across all samples were calculated using all values (reference), as well as after missingness was added (trimmed), and using Kendall-tau or Pearson correlation with global imputation (Pearson Imputed), or ICI-Kt. Figure 5 demonstrates that it is only as the number of of missing values in one of the samples approaches 50% or more (500 of 1000 features) does the Pearson correlation with global imputation give correlation values closer to the known correlation with no missing values in any appreciable amount (points below the red lines in the top panels, and to the right of the red line in the histograms in the bottom panels). Points above the lines with slopes of -1 and 1 indicate that the difference of reference - trimmed is smaller in the ICI-Kt correlations, and below the lines indicate the difference is larger in the ICI-Kt correlations. This is further emphasized by the majority of the values are to the left of the line at 0 in the difference histograms.”

We have also added the following to the Discussion section:

“Global imputation methods rely on the assumption that samples have similar dynamic range of detection and thus an imputed value should be comparable between samples. However, the dynamic range of detection often has variability across samples. For complex analytical techniques often used in omics experiments, the variability in the dynamic range of detection can be quite high. Under these circumstances, ICI-Kt provides more robust results as compared to Pearson correlation with global imputation. This holds true for low, medium, and high variability in dynamic range across samples.”

**Issue 1:**

*The example of missing data of "left-censored due to analytes being below the effective detection limit of the instrument" is a particular type of missing not-at-random (MNAR) mechanism, albeit a very simplistic one, since one can often impute those missing values with a near null/zero value (effectively the reason for the missingness in this case is to be near zero). Therefore, this type of missingness is typically not the most problematic.*

**Response:**

Having analyzed quite a few omics datasets, left-censored missing values are a serious problem. Imputation can alleviate some issues for certain statistical methods while causing problems for other statistical methods. Safety in analysis is one of the reasons we developed ICI-Kt, where it can be used in QA/QC and other analyses without changing the dataset. Then, imputation is left until it is absolutely needed, rather than being an unthinking, automatic data transformation done at the beginning. One RNAseq dataset comes to mind, where imputation or dropping missing values hid outlier samples with about half of the detection gene expressions of the other samples, which affected downstream differential analyses.

**Issue 2:**

*When the authors state "when it comes to calculating correlation, there are very few methods that explicitly account for missing data that we know of.", they must be considering MNAR data even if they do not explicitly state, as the literature is filled up with ideas to handle MAR data. The same holds for other sentences, such as "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", as it is known that under a MAR mechanism, dropping missing values can be unbiased (with enough data, one finds the "correct" result).*

**Response:**

This is a valid point. We have made the left censorship condition clearer in the Abstract and Introduction.

“In omics data sets that are derived from analytical measurements, a major reason for missing values is that a specific measurable phenomenon falls below the detection limits of the analytical instrumentation (left-censorsed values).”

“Therefore, to better handle missingness due to left-censorship, we propose the information-content-informed Kendall-tau (ICI-Kt) correlation coefficient that allows left-censored missing values to carry explicit information in the determination of concordant and discordant pairs.”

“However, when it comes to calculating correlation, there are very few methods that explicitly account for left-censored missing data that we know of.”

**Issue 3:**

*On "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", it is unclear any the set of samples should be ordered. This does not seem to be a requirement for the calculations, but perhaps an artefact of the analyses in mind.*

**Response:**

Thank you for catching this. You are correct, there is no need for an ordering with respect to the samples.

**Issue 4:**

*The so-called classical definition of concordant and discordant pairs for Kendall's calculation that is presented in the beginning of the Methods section is valid only without ties, although this is not mentioned when it is presented.*

**Response:**

We clearly state this at the beginning of the next subsection “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 and values to the denominator, and in our special case of missing data, we can add the ties that result from and to and [18,19].”

**Issue 5:**

*The so-called "information-content-informed concordant pair" (as well as for the discordant) is exactly the same as taking the missing values as being smaller than any observed value, so that one cannot see the difference between that and imputation with a lower than minimum value. This is what one would call the standard procedure in the literature for the type of missing value that has been the focus of this manuscript.*

**Response:**

There are similarities with imputing a value below the lowest value across all samples (i.e., global minimum) and our implementation of ICI-Kt. However, we have built in several other steps and features for computing “local” and “global” ICI-Kt correlation definitions. There is also combining ICI-Kt with completeness. So the ICI-Kt methodology, R, and Python packages are more than just this part of the implementation. Plus, our approach is safer than initial imputation with respect to downstream analysis steps and when data transformations are utilized. We are presenting more than just one simple method here and demonstrating their utility for improving correlation interpretation. Moreover, we have efficient R and Python implementations that utilize multiprocessing, enabling practical application for calculating correlation matrices for large datasets. Many of the correlation matrix analyses presented in this manuscript are not practical even on HPC resources without our efficient implementations, especially not in R.

**Issue 6:**

*The definitions do not account for then X (respectively Y) is missing in both pairs, which would lead to some observations being neither concordant nor discordant. This is discussed later on about ties and the tau-b. The values n\_{xtie} and n\_{ytie} are not formally defined (same for n\_0, though one can infer from common knowledge. The handling of those would again match with imputing a particular value (lower than the minimum) and running the usual inferences, so there seems to be any difference from the "standard". This seems to be acknowledged in the Implementation Details section.*

**Response:**

Thank you for pointing out the lack of definitions for n\_{xtie} and n\_{ytie} and n\_0! We have fixed this as follows, substituting n\_{tot} for n\_0 and providing further definition in the text:

“

Where is the total number of pairs, are the number of tied values in X, and are the number of paired values in Y.”

**Issue 7:**

*The so-called local ICI-Kt matches with the usual procedure of disregarding pairs when X (or Y) are missing in both elements of the pair, while the global does not. The discussion on the maximum possible value might be interesting, though I cannot infer the exact meaning as the elements in the equation are not formally defined. If I would dare to guess the meaning of the so-called theoretical maxima, it seems that it is simply the calculation of the maximum value the well-known tau-b can get, which is a trivial computation. Afterwards, the so-called completeness seems to be a common simple value used to analyse the amount of missingness in the data.*

**Response:**

Some of our improvements are not rocket science; however, we provide highly optimized implementations in two very popular programming languages that straight-forwardedly provide all correlation results with efficient use of computational resources. The tau{max} may be trivial to derive for some; but this calculation is integrated into the packages, which means that others do not need to derive, implement, and test it.

We have also updated the definition of the tau\_{max}, and provided further definitions previously in the text (see above):

“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:

Where is the number of commonly tied values in both X and Y. Calculating a set of values between all experimental samples, we can take the maximum of the values, and use it to scale **all** of the obtained Kendall-tau values equally ().

**Issue 7:**

*The details about the implementation are useful. However, whether the Kendall tau is efficiently implemented in the packages used as baseline is a minor point, since it is known that it can be computed in O(n log n) time and space. This perhaps can be a very simple implementation contribution, though I assume that most practitioners using Kendall who have large data to process will make sure to use an O(n log n) procedure anyway (it is very straightforward to run in O(n log n)).*

**Response:**

We disagree! Software packages implemented to the high industrial standard demonstrated here are typically beyond the capabilities of most academic labs. The R implementation includes compiled C/C++ code and utilizes multiprocessing to fully use the CPU cores available in a very efficient manner. Likewise, the Python implementation uses Cythonization to convert and compile C++ code and also utilizes multiprocessing with shared memory for a highly efficient implementation. Furthermore, these packages enable the calculation of very large correlation matrices not possible with other implementations. These optimized packages required development and testing on high memory (1TB+ RAM) compute nodes. This level of package development and implementation is NOT minor and NOT straightforward.

**Issue 8:**

*Finally, a collection of experiments follow to illustrate the use of the methods. Given that the theoretical part shows that the procedure is the same as the standard approach for dealing with missing values (where the value is known to be smaller than the minimum in the data), I have not checked all those experimental details.*

**Response:**

There is no tangible issue to respond to.

**Issue 9:**

*Minor issues:*

*"since it’s introduction" -> its*

**Response:**

Thanks! Fixed.

**Reviewer 3**

*This article proposes a Kendall-tau coefficient adapted to the case of missing left-censored data (very common with omics data). In my opinion, the authors should rework the form of the article for clarity: highlighting the case in question (left-censored), revising certain notations, ... I recommend that this paper should be published in BMC Bioinformatics, after some minor revisions. I give my remarks in the following.*

**Response:**

We thank the reviewer for their positive comments!

**Issue 1:**

*Major comments*

*1. This is my major concern. The authors introduce the left-censored data problem in the introduction and talk about the fact that omics data are particularly concerned by this (Section Limit of Detection As a Cause for Missingness). The proposed Kendall-tau coefficient (ICI-Kt) only makes sense in some specific cases of missingness. The authors claim that in the Discussion and Conclusion part: “ICI-Kt explicitly treats leftcensored*

*missing values”. However, the authors should make it clear in the abstract, and also reiterate throughout the article, that the proposed Kendall coefficient does not handle all situations of missingness.*

**Response:**

We thank the reviewer for pointing out the need to highlight the left censorship use-case in the abstract and introduction! As authors, it can be hard to view the manuscript from the perspective of a new reader. We have made the following revisions to address this shortcoming.

Abstract: “In omics data sets that are derived from analytical measurements, a major reason for missing values is that a specific measurable phenomenon falls below the detection limits of the analytical instrumentation (left-censorsed values).”

“Therefore, to better handle missingness due to left-censorship, we propose the information-content-informed Kendall-tau (ICI-Kt) correlation coefficient that allows left-censored 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 left-censored missing data values as interpretable information, enabling both improved determination of outlier samples and improved feature-feature network construction, without explicitly using imputation.”

Introduction: “However, when it comes to calculating correlation, there are very few methods that explicitly account for left-censored missing data that we know of.”

**Issue 2:**

*2. In conclusion, it might be interesting if the authors could quickly discuss potential variants of their coefficient. If the data are missing when the values are large (right-censored data problem), the coefficient variant is fairly obvious. In other missing situations, where the lack of data depends on the value of the data itself (not missing at random) but is neither a leftcensored nor a right-censored case, what coefficient could be proposed? In addition, the numerical experiment in Figure 4 is interesting, and the fact that the proposed coefficient makes sense in the case of MAR data could be emphasized more.*

**Response:**

If software returned variables that are obviously right censored with something to indicate right censorship and the fact that the values are not truly known, then a right-censored variant could be easily implemented very similarly to the left-censored case. However, in our experience, detecting right-censored data is difficult, as truncated values are reported, with no indication of “censorship”. With respect to missing at random (MAR), this is tricky. Yes, ICI-Kt is useful for “penalizing” MAR. But sometimes you do not want to “penalize” MAR. This is why we added a left censorship statistical test to the package, but after we had submitted this manuscript. But to the reviewer’s comment, we have added a discussion about “penalizing” MAR to the discussion:

“ICI-Kt explicitly treats left-censored missing values as correlative information while preserving the full deleterious effects of random values on the correlation metric.”

**Issue 3:**

*3. In the paragraph “Theoretical Maxima”, the authors claim that the maximum number of concordance can be used to scale the obtained Kendalltau values. Have the authors tested this method in the “semi-realistic” dataset, for example?*

**Response:**

We have tested normalizing correlation values with the tau\_max on multiple datasets. Given the number of analyses and datasets presented in this manuscript, we had to pick and choose what to emphasize. Although it may be useful to scale the individual ICI-Kt values by their respective tau\_{max} values, we did not do this. Rather, we scale **all** the values in a dataset by the maximum tau\_{max} observed for a set of correlations. This means that all values are scaled by the same amount. We have clarified this in the manuscript:

“Calculating a set of values between all experimental samples, we can take the maximum of the values, and use it to scale **all** of the obtained Kendall-tau values equally ( ).

To show the effect of applying it, here is a plot of the median sample – sample correlations among the yeast samples un-scaled and scaled by max(tau\_{max}), showing that the scaling across samples is applied equally, and therefore does not change the determination of outliers for a group of samples.

A graph with a line

Description automatically generated

**Issue 4:**

*4. In the real datasets, could the authors give precise information on the percentage of missing data (in the form of a table, for example)? In addition, some of the results in the appendix are interesting. Overall, for the real datasets, I think it would be beneficial to add a summary table of the results to clarify the experiments.*

**Response:**

Thanks for pointing this out! This is a really good suggestion. We have added Table 2, with both the number and percentage of missing values, as well as results of the binomial test for left-censorship.

**Issue 5:**

*Minor comments*

*1. I find that the notations in the Methods section, although central, are introduced too little, making this part difficult to read, when simple examples could suffice to make it more enjoyable. (a) In the paragraph Additional Definitions of Concordant and Discordant Pairs to Include Missingness, I understand the authors’ choice to use the !x notation to indicate that a value is missing, since not all programming languages use NA, but I would still find it more readable if the authors used the NA notation. It’s a suggestion, but for me it would be clearer to note {xi = NA}&{xj} instead of !xi&xj . (b) In the formula, some notation are not explicitly given: n0 (page 5), ntie (page 6), nfeat (page 6), missi (page 6) (c) In the paragraph Considering Ties, it would be easier to read if the different cases were explicitly stated, as in the previous paragraph. In particular, it is not clear for me, when the authors say: “the ties that result from both of (xi, xj) and (yi, yj) being missing”. Does it mean that xi = NA, xj = NA and yi = NA, yj = NA ?*

**Response:**

We have to pick a single consistent notation. However, we have added additional description and examples to describe in terms of NA.

“ Here indicates and indicates , and & is synonymous with “and”.”

We have updated the tie definitions as follows:

“we can add the ties that result from and to and .”;

“

Where is the total number of pairs, are the number of tied values in X, and are the number of paired values in Y.”

We have updated the formula and definitions under **Theoretical Maxima:**

“

Where is the number of commonly tied values in both X and Y. Calculating a set of values between all experimental samples, we can take the maximum of the values, and use it to scale **all** of the obtained Kendall-tau values equally ().”

We also updated the formula and definitions for **Completeness:**

“

Where for any two samples *i* and *j*, is the total number of features or entries, and are the features missing in either sample *i* or *j*, with being the total number of missing entries in either sample *i* or *j*.”

**Issue 6:**

*2. It would be clearer to define IKC in the section Methods ( τ comp ).*

**Response:**

Thanks for the suggestion! We have defined IKC in the Methods section as follows:

“ICI-Kendall-tau with zeros replaced with NA (IK); and then scaled (multiplied) by the completeness metric (IKC).”

**Issue 7:**

*3. For clarity, the legend on figure 5a could be simpler.*

**Response:**

We wanted to make it explicit which function was used, so it cannot be made simpler without losing this explicitness.

**Issue 8:**

*4. ICI-Kt \* Completeness is IKC but IKC is scaled by completeness, thus noting “\*Completeness” is confusing for me.*

**Response:**

Thanks! We have reworded this as follows:

“(2) ICI-Kt \* completeness (i.e., ICI-Kt scaled by completeness; see IKC in Methods)”

**Issue 9:**

*5. The authors use “MAR” (page 16), without explaining what this typology actually stands for. I think the authors should either talk about the MCAR/MAR/MNAR typology, here the left-censored case is MNAR, or not use this acronym.*

**Response:**

Thanks! We decided to remove extraneous MCAR/MAR/MNAR terminology, since we want to focus on the term left-censorship.