Supplemental Materials

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

2022-01-10 21:08:32

## Simulated Data

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

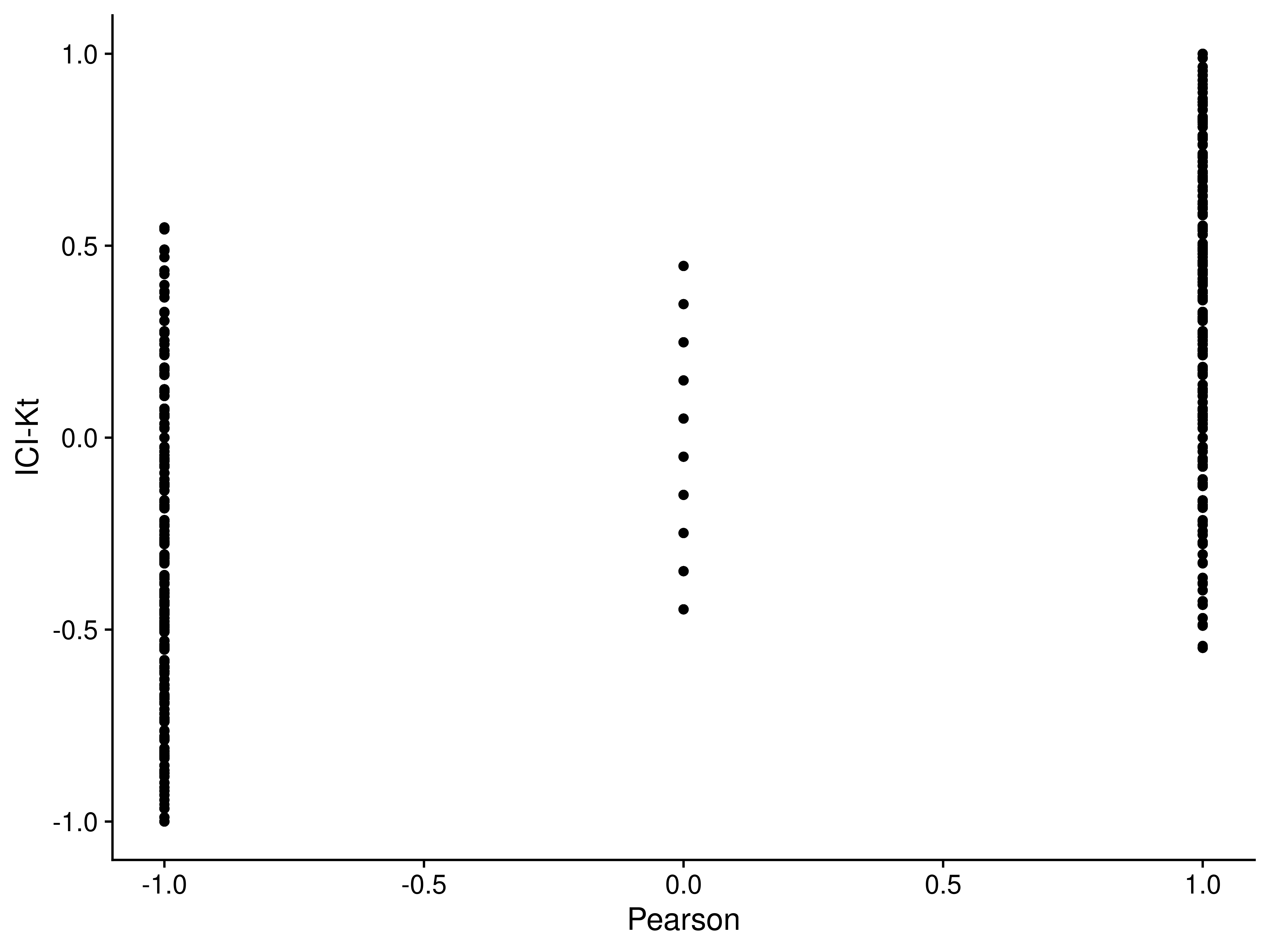


Figure S1. Comparison of ICI-Kt and Pearson correlations for perfectly positive and negatively correlated samples, systematically replacing values with NA. NA values from Pearson were replaced with zero for this comparison.

We can also examine the full set of positive and negative correlations generated as we vary the number of missing entries between two positively correlated samples and two negatively correlated samples. These distributions are shown in Figure S3. We can see that the distributions from both ICI-Kt and Kendall-tau are the same, which is expected given we replaced missing values (NA) with zero *within* the ICI-Kt code, and replaced missing values (NA) with zero prior to calculating Kendall-tau correlations.

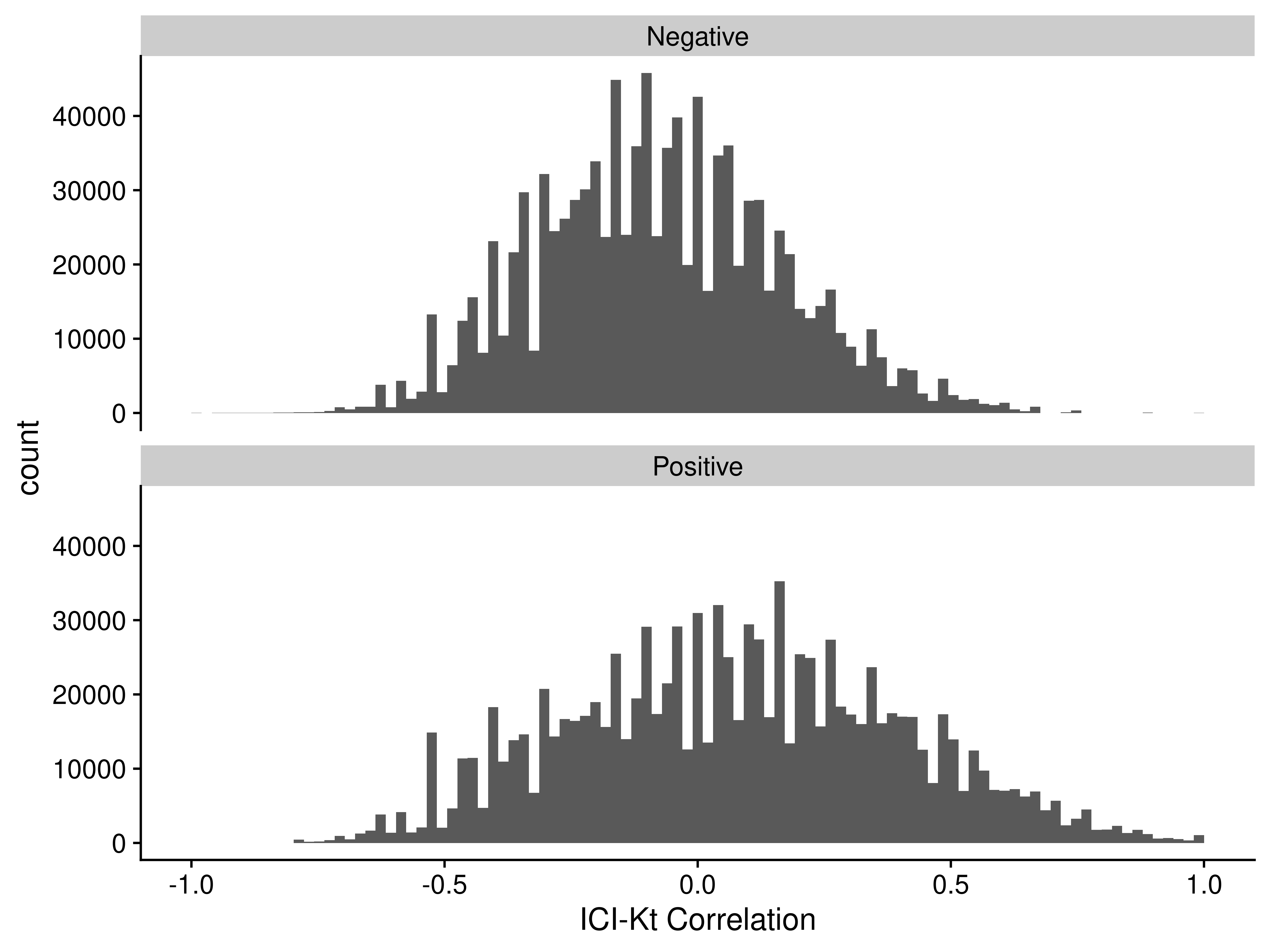


Figure S2. ICI-Kendall-tau correlation as missing values are varied between two samples.

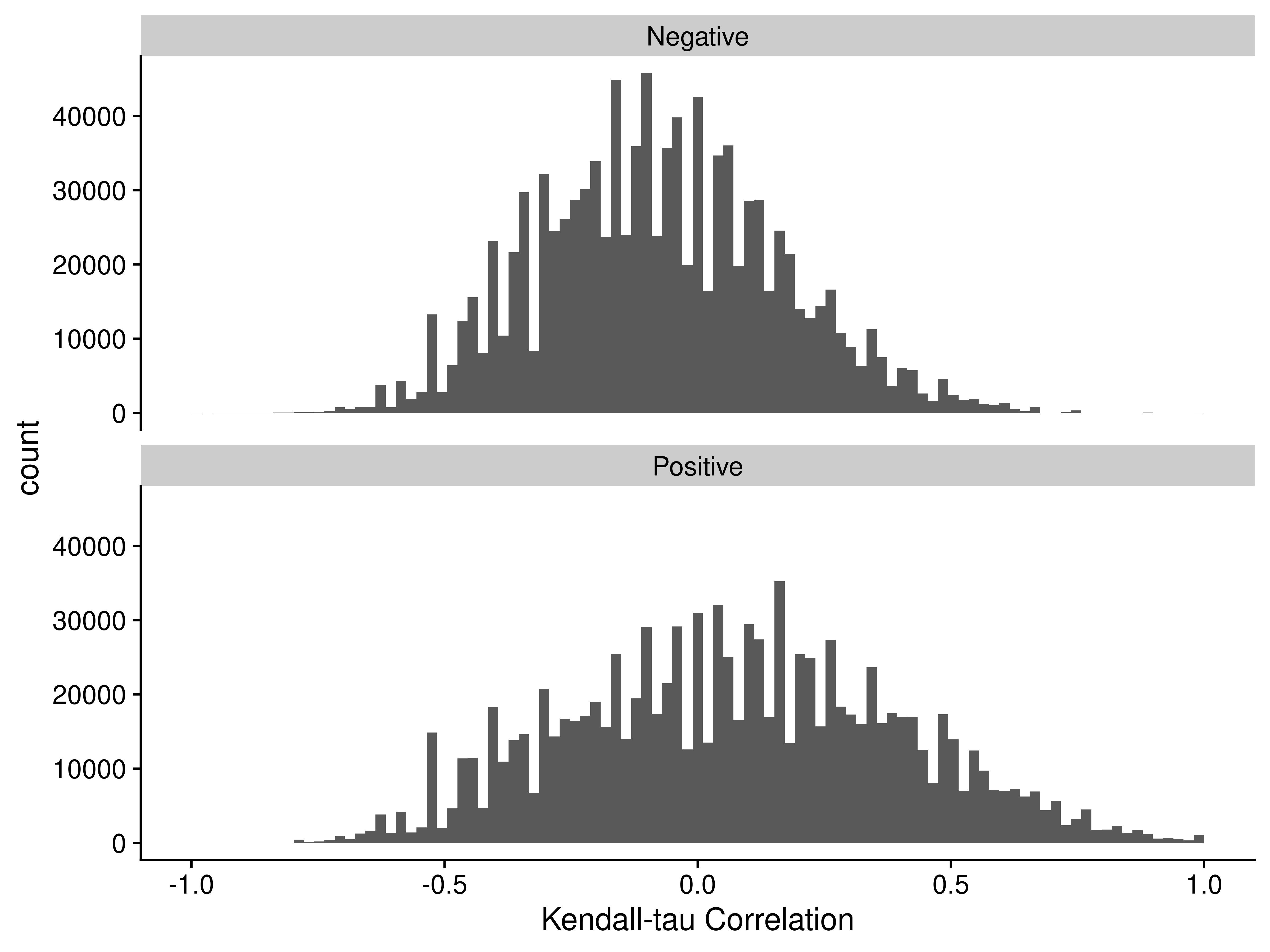


Figure S3. Kendall-tau correlation as missing values are varied between two samples and replaced with 0 before calculating Kendall-tau.

## Algorithmic Complexity

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



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

## Outlier Samples

### Gierlinski Yeast

Gierlinski et al (Gierliński et al. 2015) proposed a combination of median sample-sample correlation, gene count outlier fraction, and a chi-square test of read coverage per gene to score each biological replicate from a group of samples. Outside of defining this combined score, they do not describe the actual criteria of saying a replicate is “bad.” In this work, we used **only** the sample-sample median correlation to identify “bad” or “outlier” samples, using the boxplot.stats function to determine outliers, where an outlier is defined as samples that are > 1.5x away from the limits of the box defining the distribution.

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

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

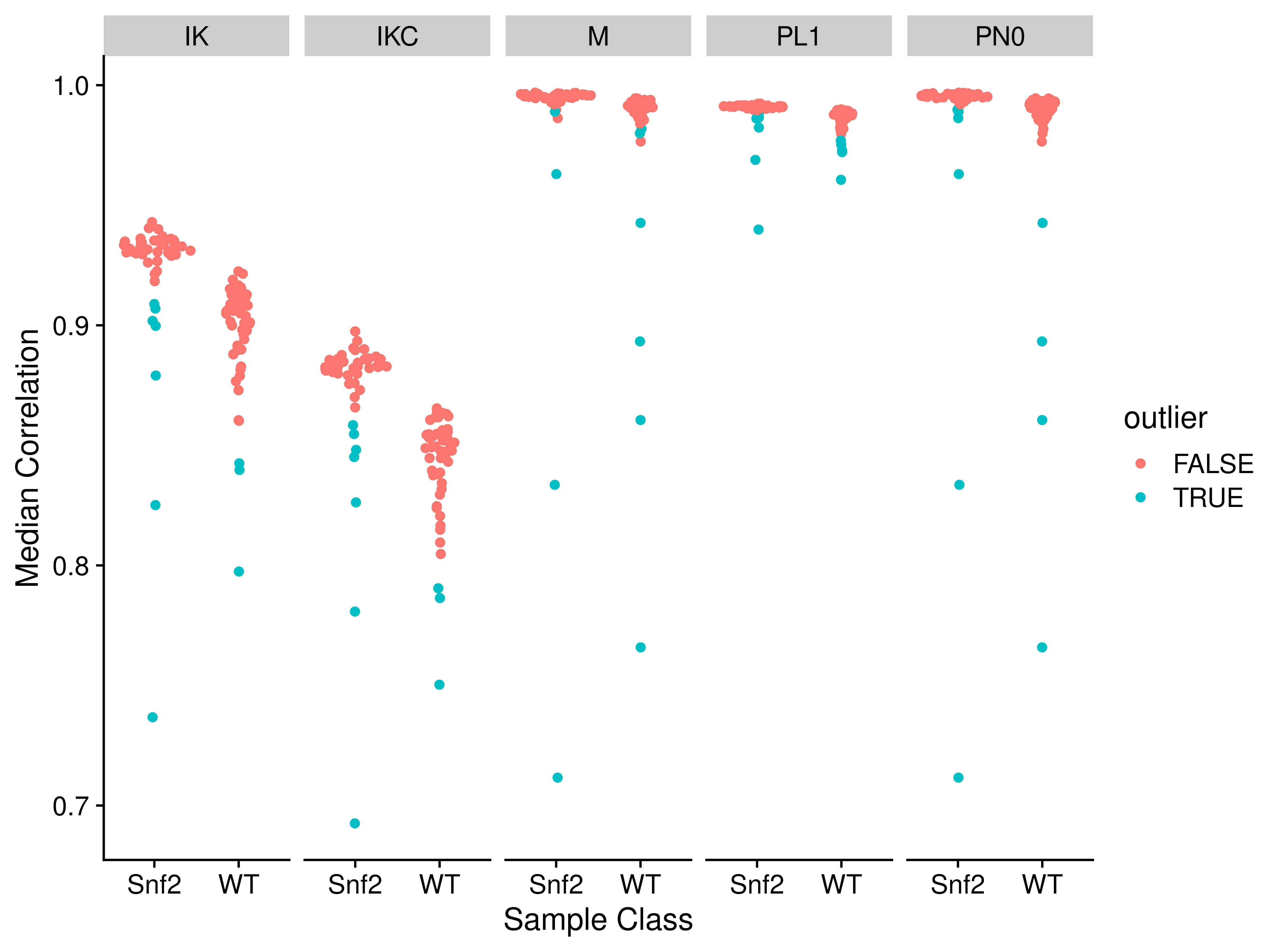


Figure S5. Median correlations for each of the yeast RNA-Seq sample to all other samples in the same group, using different correlation measures on different data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; M: Manuscript; PL1: Pearson Log(x + 1); PN0: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

Table S1. Yeast 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; M: Manuscript; PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

|  | IK | | IKC | | PL1 | | PN0 | | M | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | Outlier | Correlation | Outlier | Correlation | Outlier | Correlation | Outlier | Correlation | Outlier | Correlation |
| Snf2.31 | X | 0.90 | X | 0.85 | X | 0.99 |  | 0.99 |  | 0.99 |
| Snf2.15 | X | 0.90 | X | 0.85 | X | 0.99 |  | 0.99 |  | 0.99 |
| Snf2.24 |  | 0.93 |  | 0.88 |  | 0.99 | X | 0.99 |  | 0.99 |
| Snf2.25 | X | 0.88 | X | 0.83 | X | 0.98 | X | 0.99 | X | 0.99 |
| Snf2.10 | X | 0.91 | X | 0.85 | X | 0.99 | X | 0.99 |  | 0.99 |
| Snf2.35 | X | 0.91 | X | 0.86 | X | 0.99 | X | 0.96 | X | 0.96 |
| Snf2.13 | X | 0.83 | X | 0.78 | X | 0.97 | X | 0.83 | X | 0.83 |
| Snf2.06 | X | 0.74 | X | 0.69 | X | 0.94 | X | 0.71 | X | 0.71 |
| WT.34 | X | 0.84 | X | 0.79 | X | 0.97 |  | 0.98 | X | 0.98 |
| WT.36 |  | 0.86 |  | 0.81 | X | 0.98 |  | 0.98 | X | 0.98 |
| WT.22 |  | 0.88 |  | 0.82 |  | 0.98 | X | 0.94 | X | 0.94 |
| WT.28 |  | 0.86 |  | 0.80 | X | 0.98 | X | 0.89 | X | 0.89 |
| WT.25 | X | 0.84 | X | 0.79 | X | 0.97 | X | 0.86 | X | 0.86 |
| WT.21 | X | 0.80 | X | 0.75 | X | 0.96 | X | 0.77 | X | 0.77 |

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

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

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

## Brainson RNASeq Samples

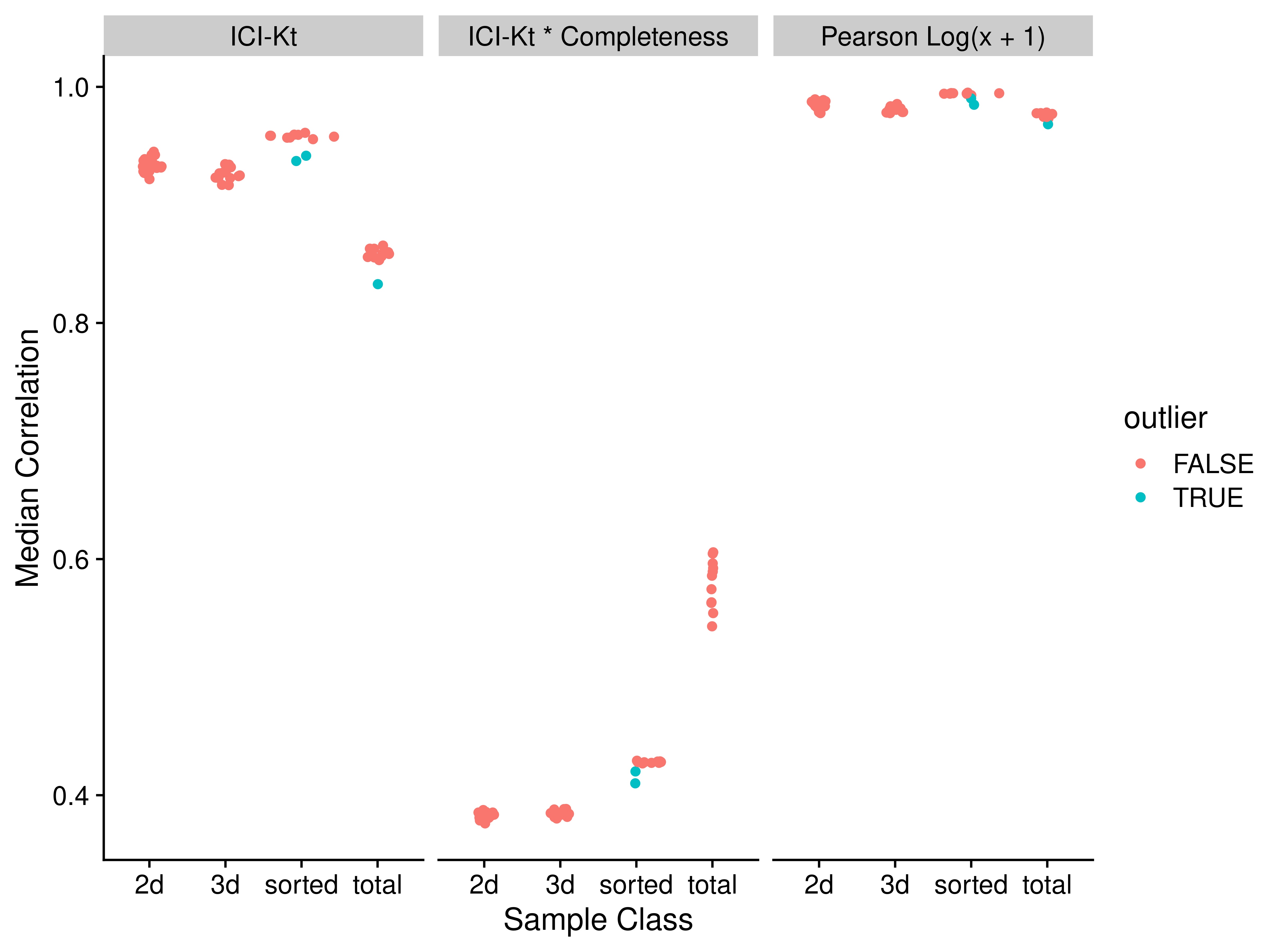


Figure S6. Median correlations for each of the Brainson RNA-Seq sample to all other samples in the same group, using different correlation measures on different data.

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

Table S2. Brainson dataset median correlation values and outlier determination for each outlier from each of the correlation methods.

|  | ICI-Kt | | ICI-Kt \* Completeness | | Pearson Log(x + 1) | |
| --- | --- | --- | --- | --- | --- | --- |
| Sample | Outlier | Correlation | Outlier | Correlation | Outlier | Correlation |
| 1004-het.sorted | X | 0.94 | X | 0.42 | X | 0.99 |
| 783-het.sorted | X | 0.94 | X | 0.41 | X | 0.98 |
| 1263wt\_S1.total | X | 0.83 |  | 0.56 | X | 0.97 |

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

## TCGA Adenocarcinoma

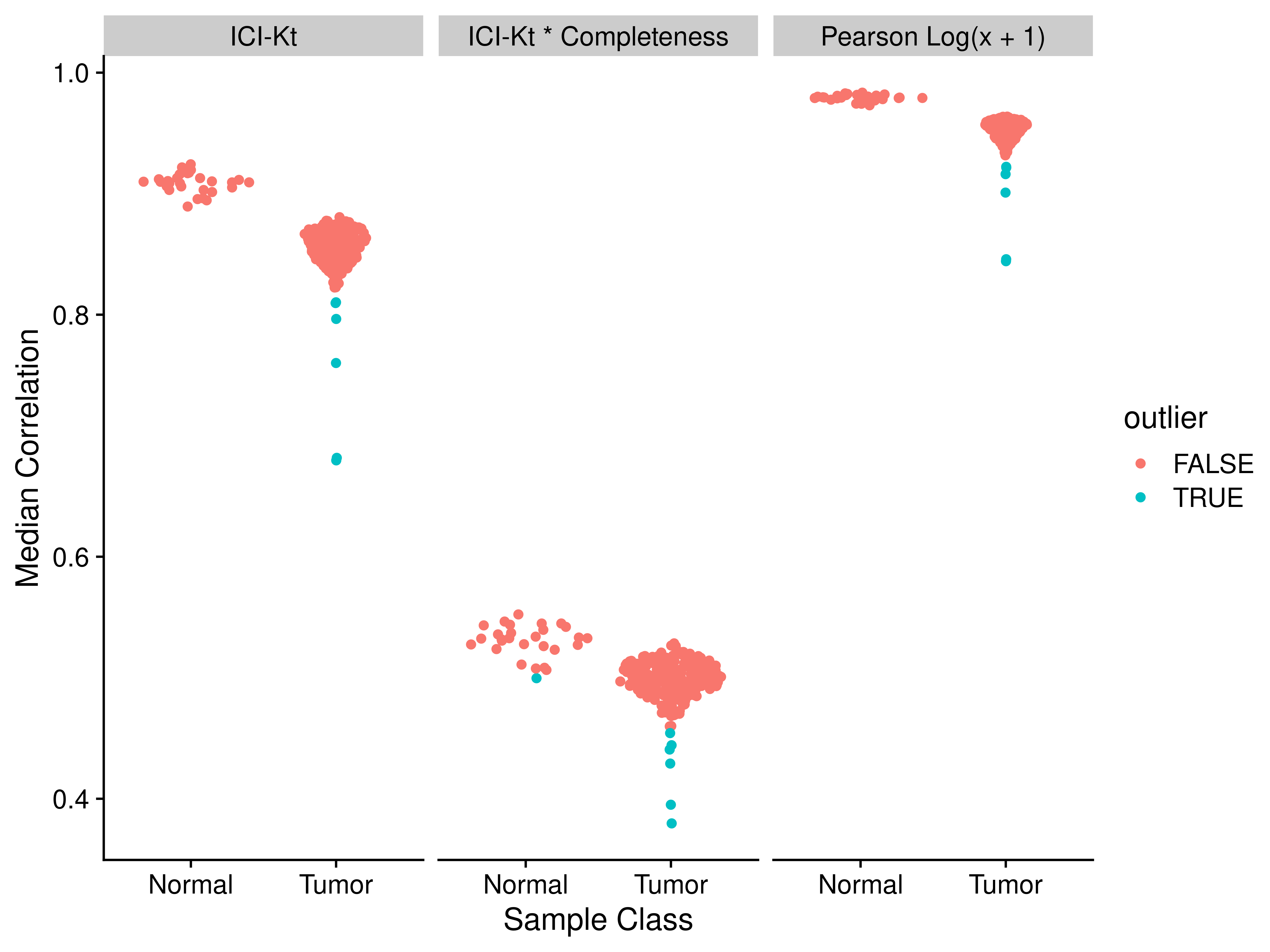


Figure S7. TCGA Adenocarcinoma outliers using various measures of correlation.

Table S2. Brainson dataset median correlation values and outlier determination for each outlier from each of the correlation methods.

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

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

| method | Normal | Tumor |
| --- | --- | --- |
| ICI-Kt | 0.006 | 0.014 |
| ICI-Kt \* Completeness | 0.012 | 0.015 |
| Pearson Log(x + 1) | 0.002 | 0.007 |

As Figure S7 and Table S3, although there are not many outliers in these relatively large groups, Pearson correlation using log-transformed values results in the most outliers. Also of note is that ICI-Kt \* Completeness is the only measure where the variance measure for both groups is similar. For both ICI-Kt and Pearson correlation, the tumor samples variance measures are 2x and 3x greater than normal, respectively.

## Effect of Increasing Presence in Samples

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

### Yeast Samples

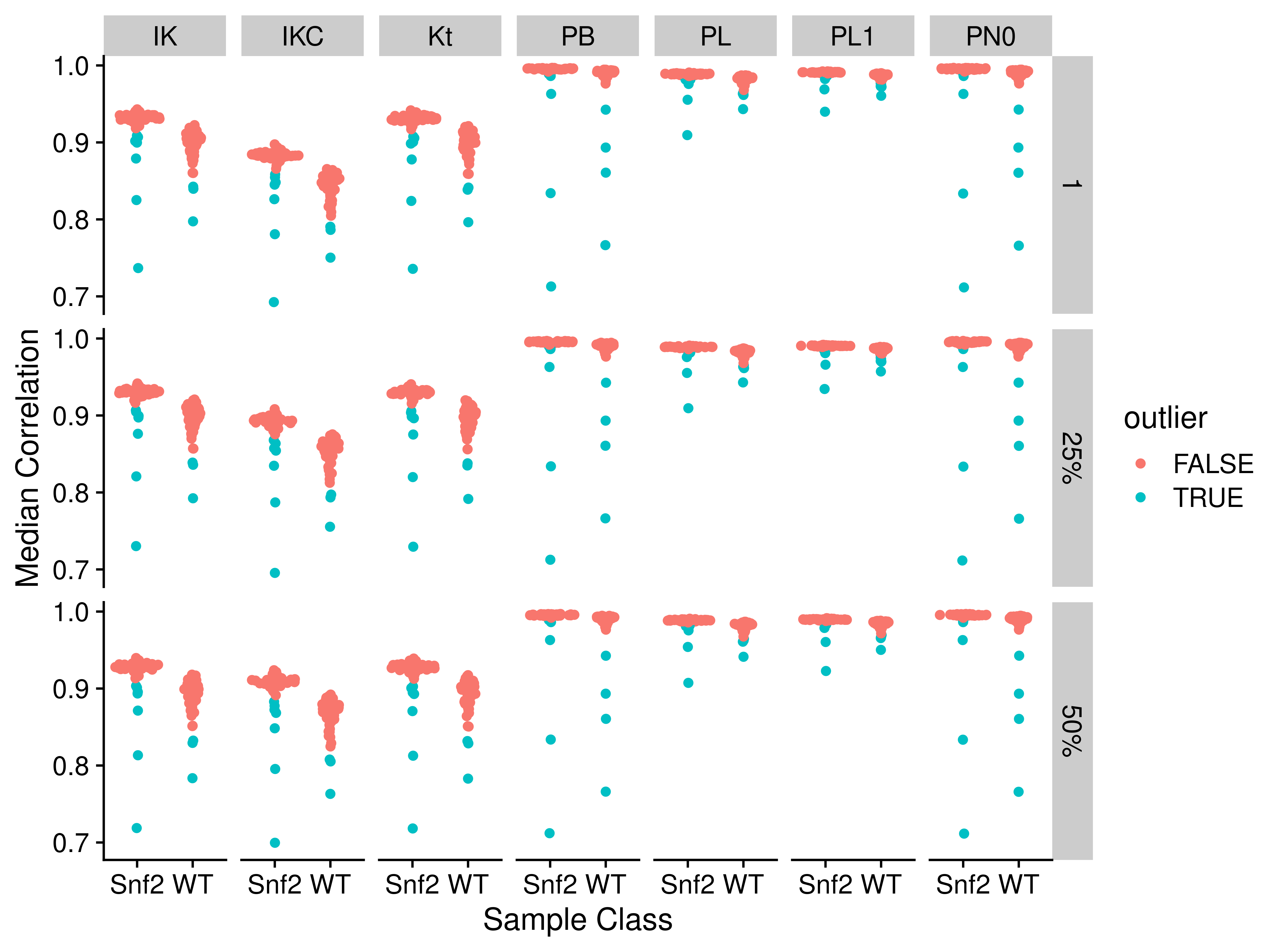


Figure S8. Median correlations by correlation method and applying different fractional cutoffs. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson No Zeros.

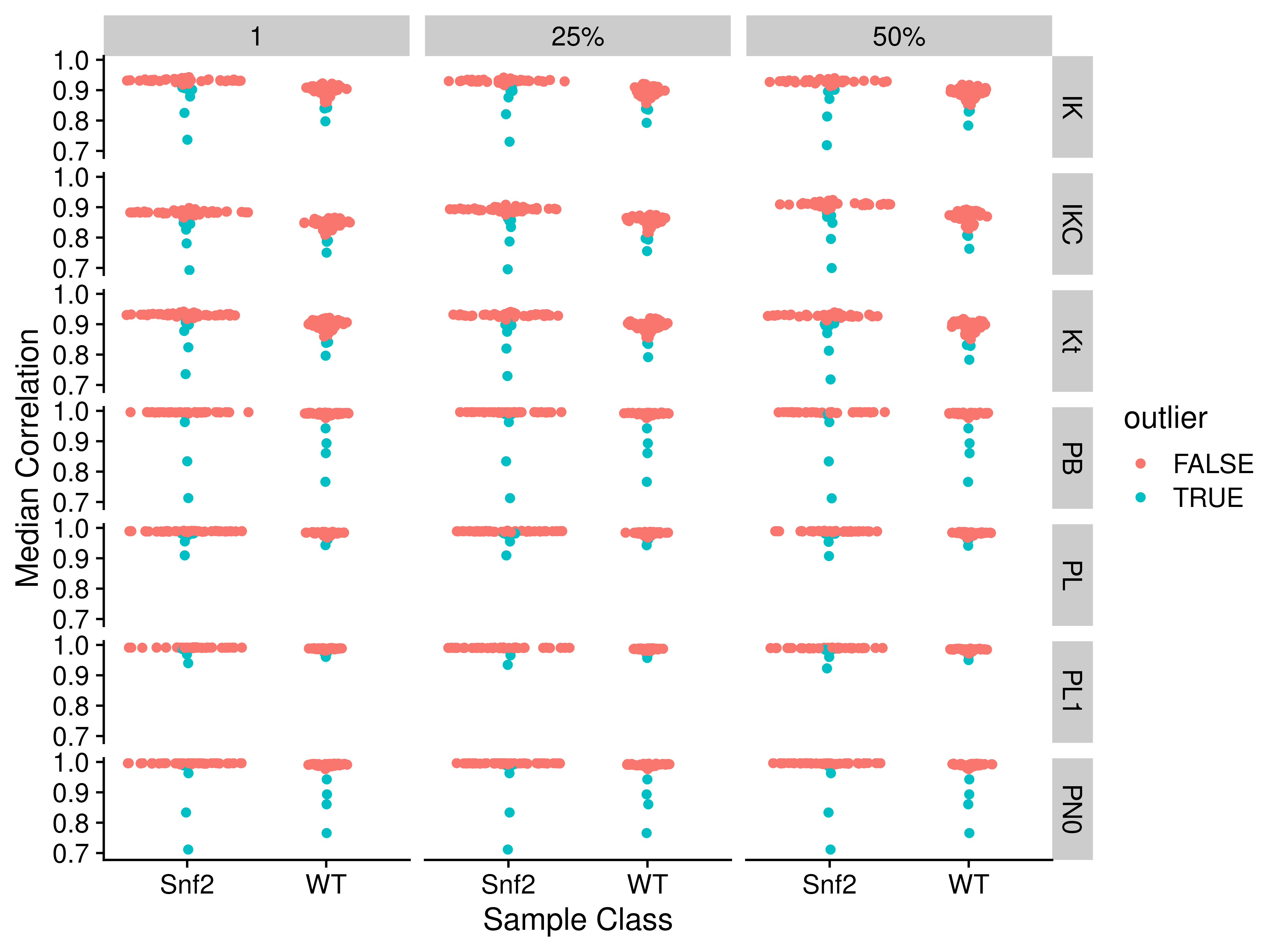


Figure S9. Median correlations by applying different fractional cutoffs and different correlation methods. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x + 1); PN0: Pearson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

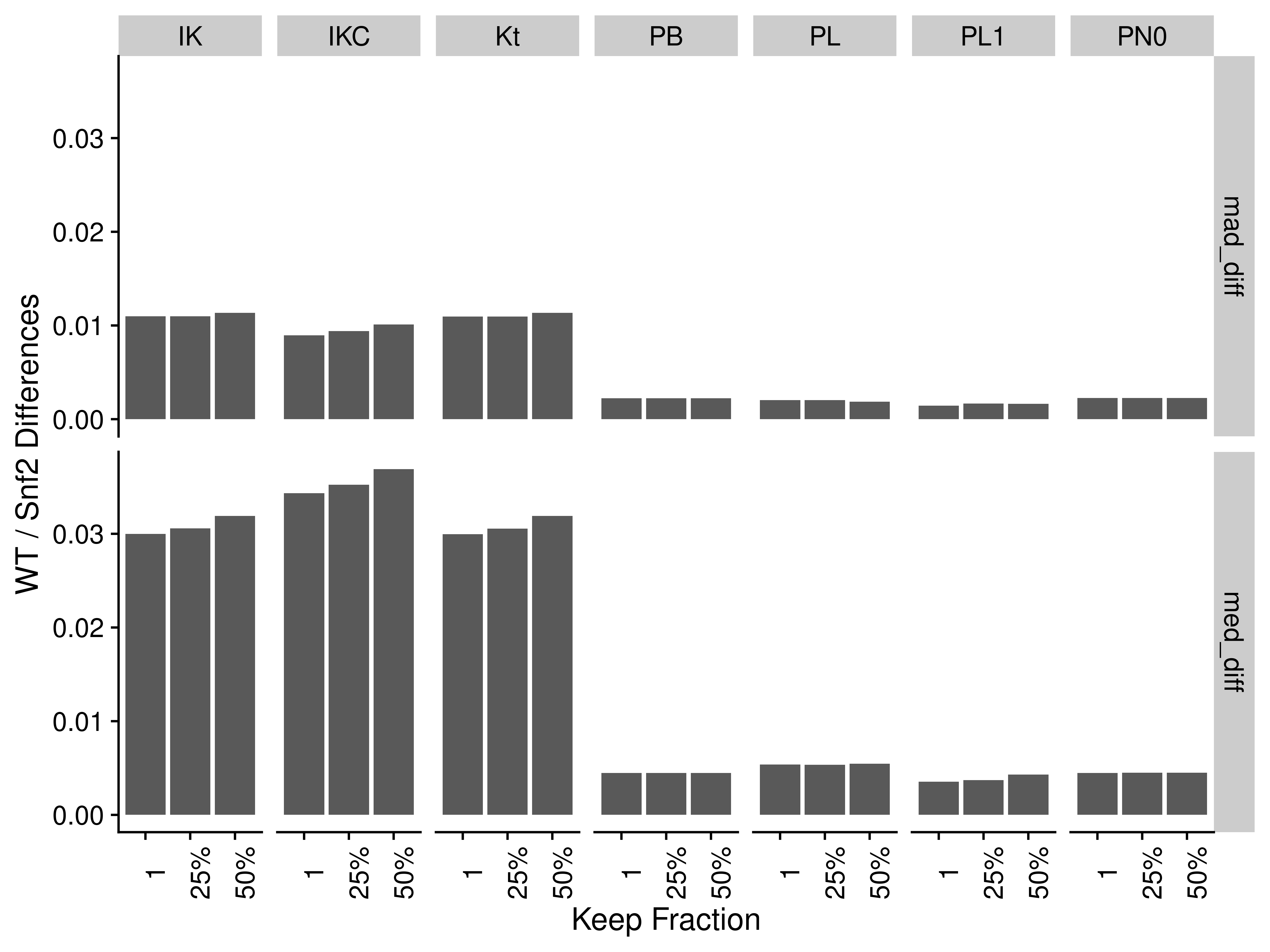


Figure S10. Difference between WT and Snf2 group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson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

### Brainson Samples

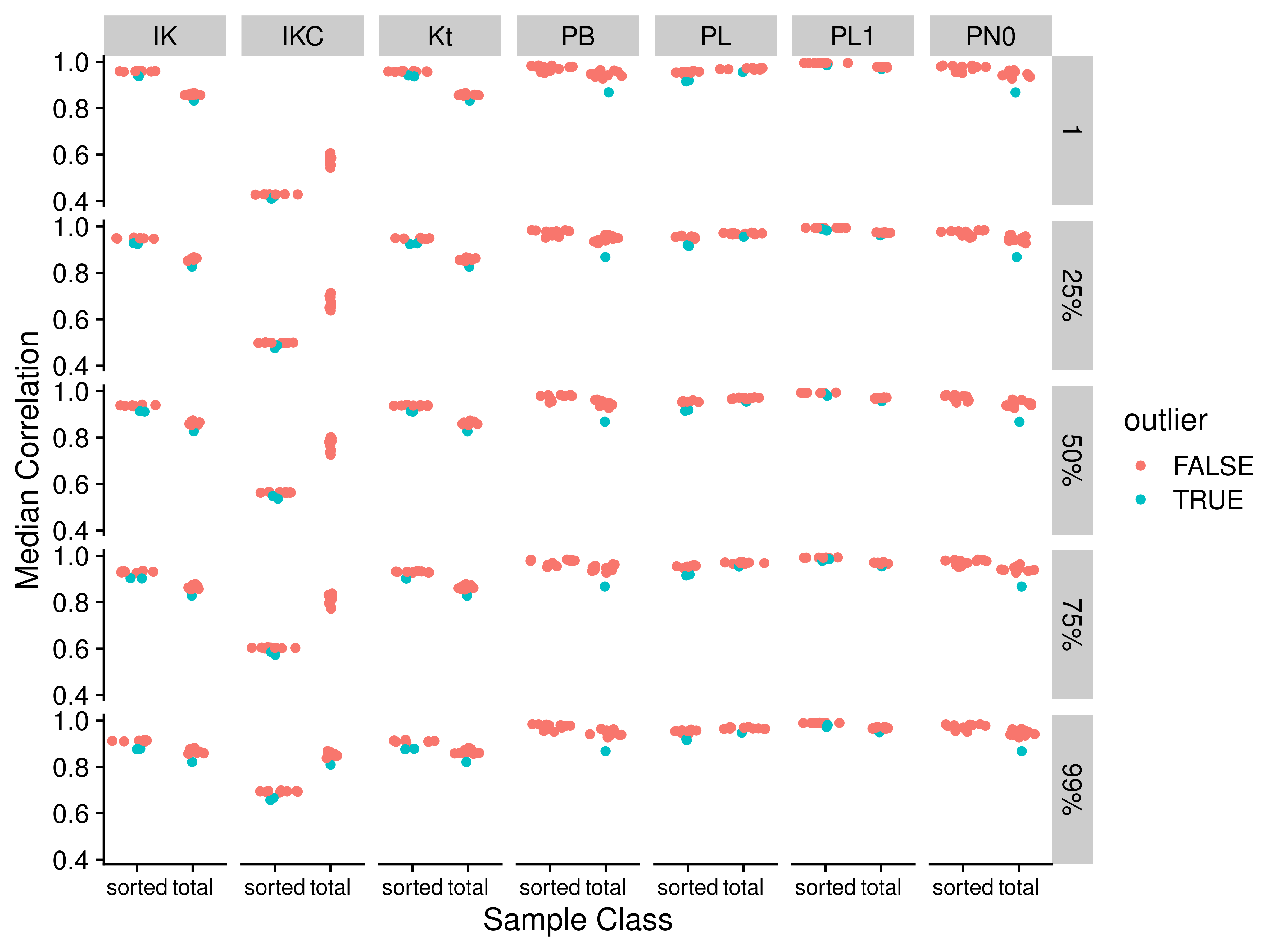


Figure S11. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x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

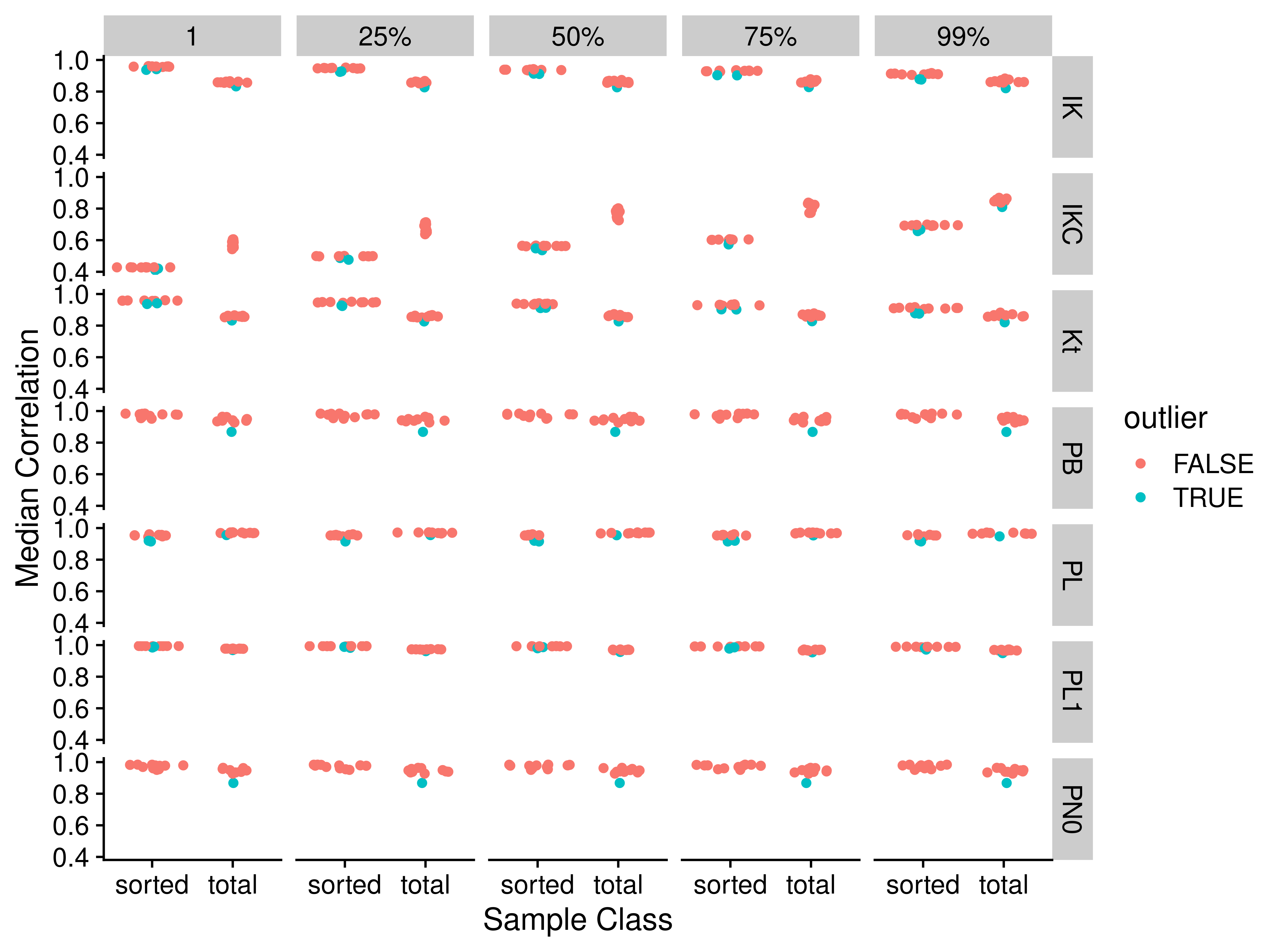


Figure S12. Median correlations by correlation method and applying different fractional cutoffs for Brainson RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson Log(x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

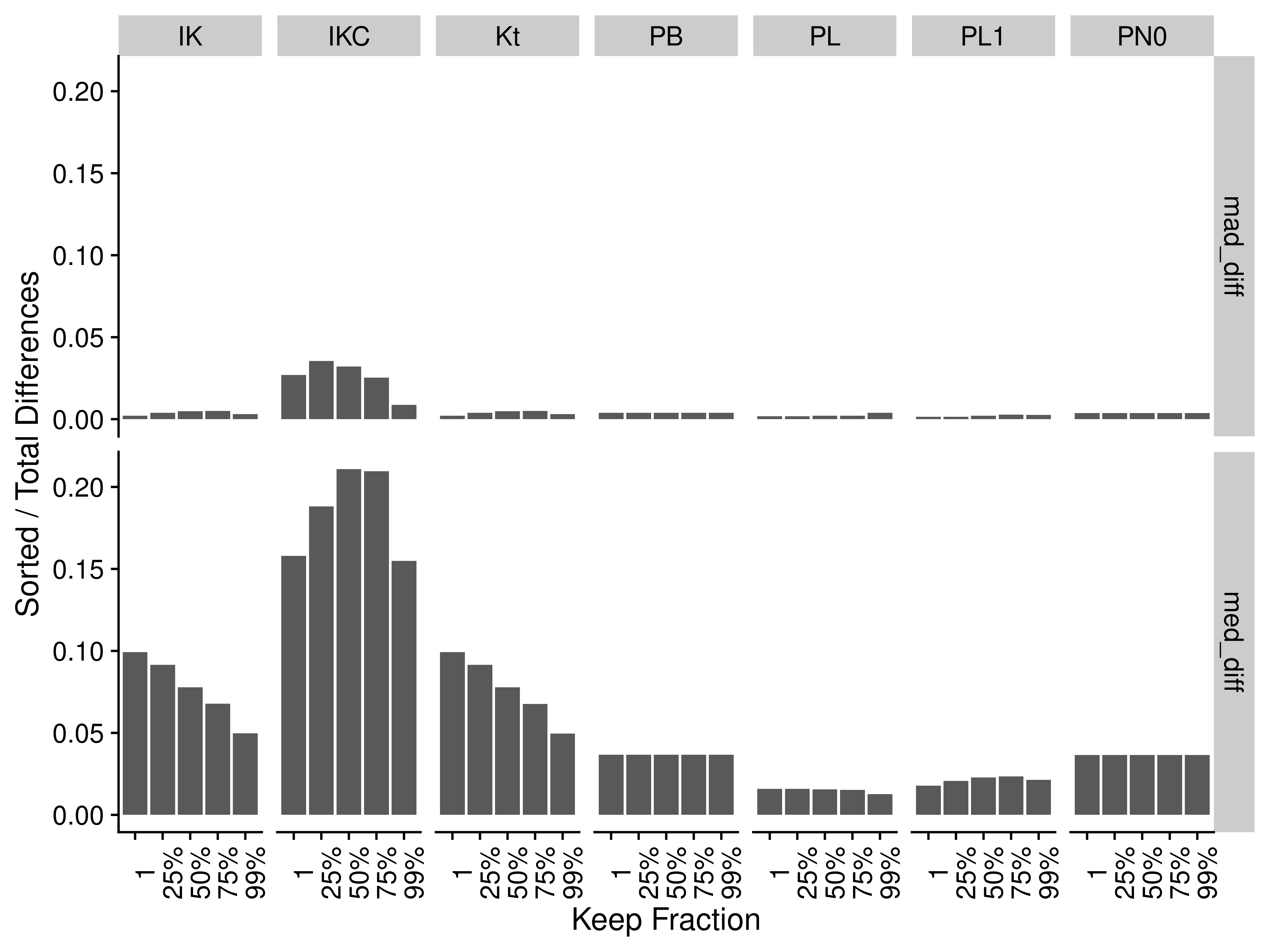


Figure S13. Difference between sorted and total sample group medians and MAD for the different correlation methods and gene presence fractions. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x);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

### Adenocarcinoma

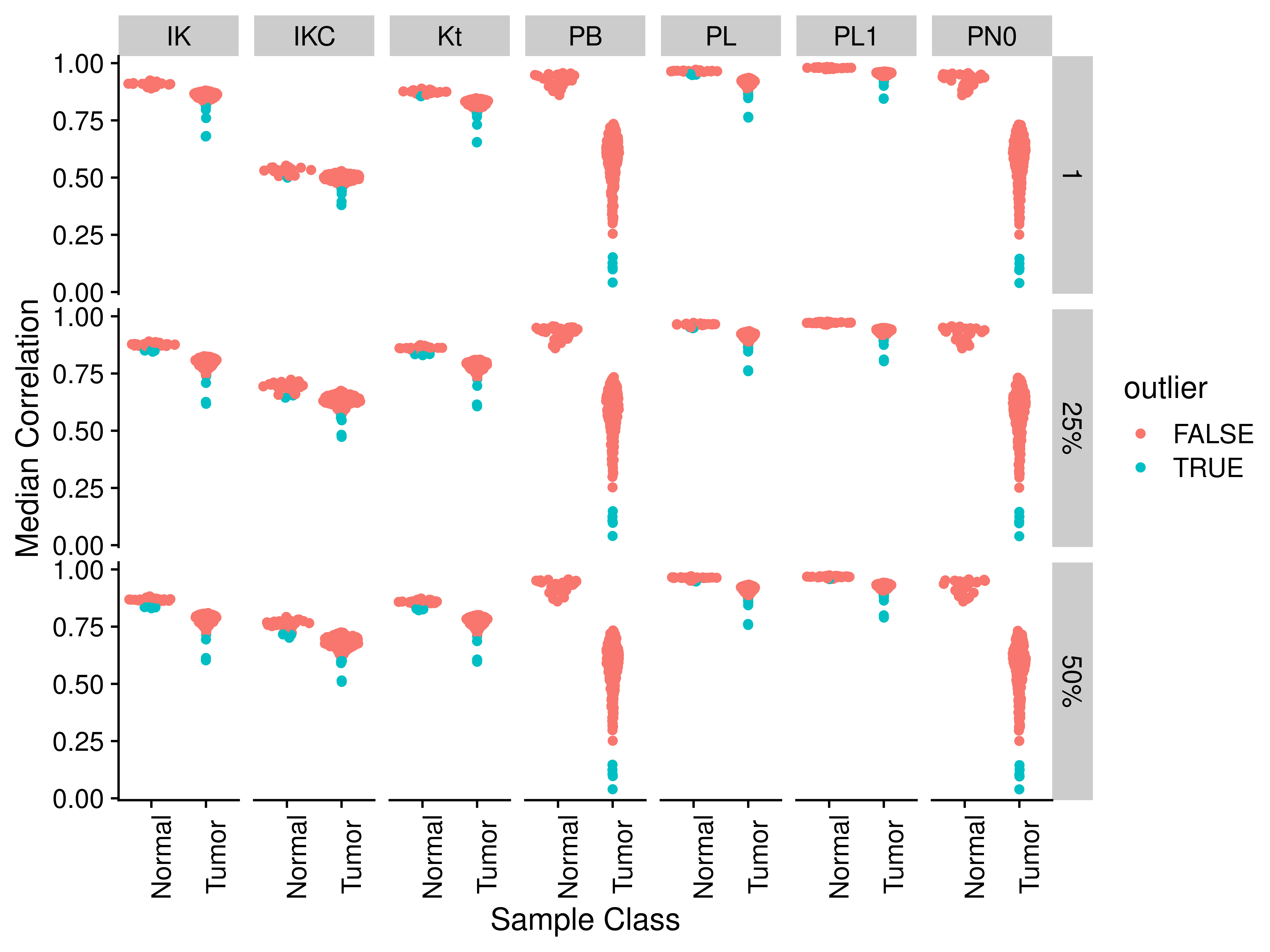


Figure S14. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson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

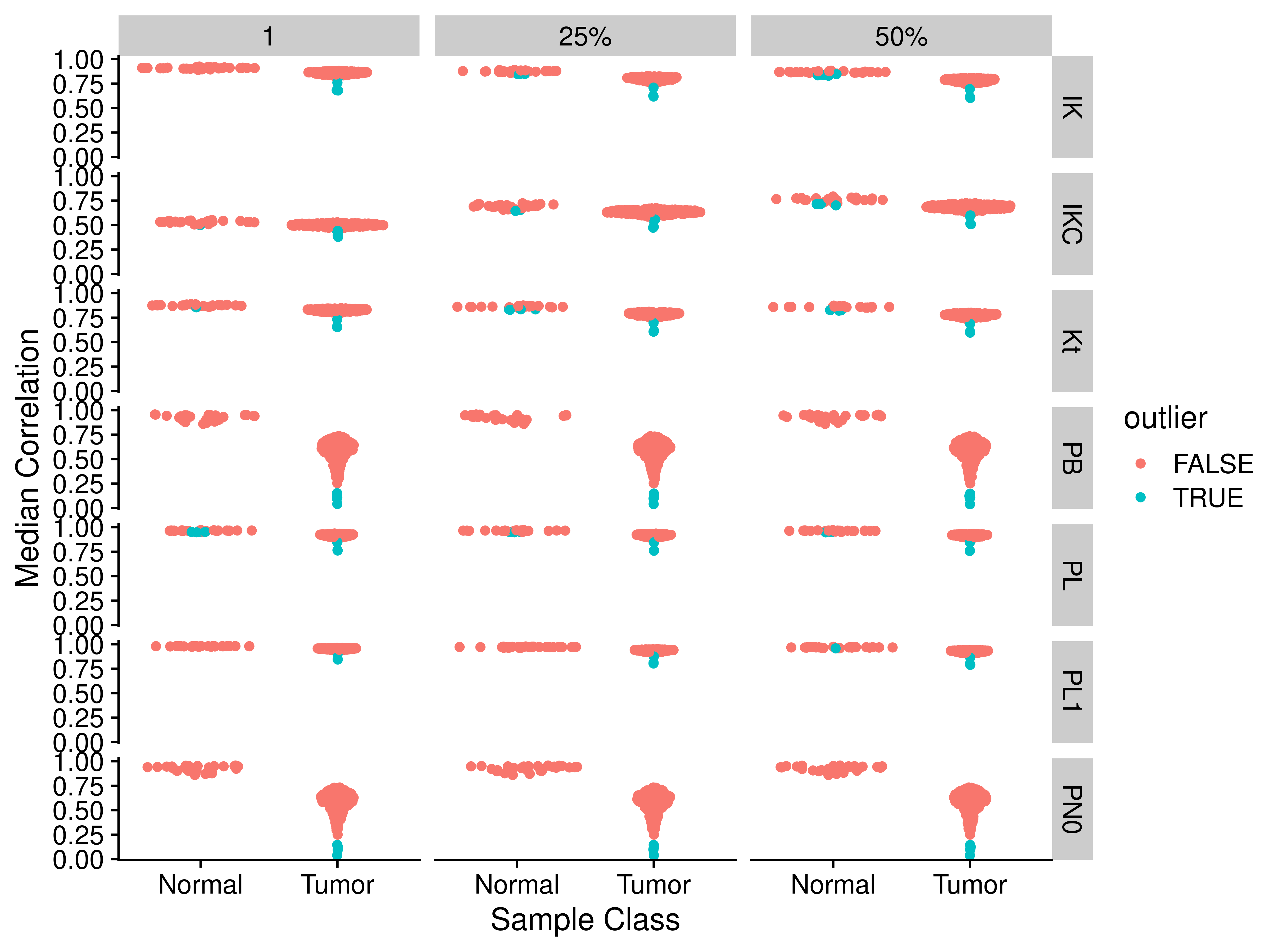


Figure S15. Median correlations by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values); PL: Pearson Log(x); PL1: Pearson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

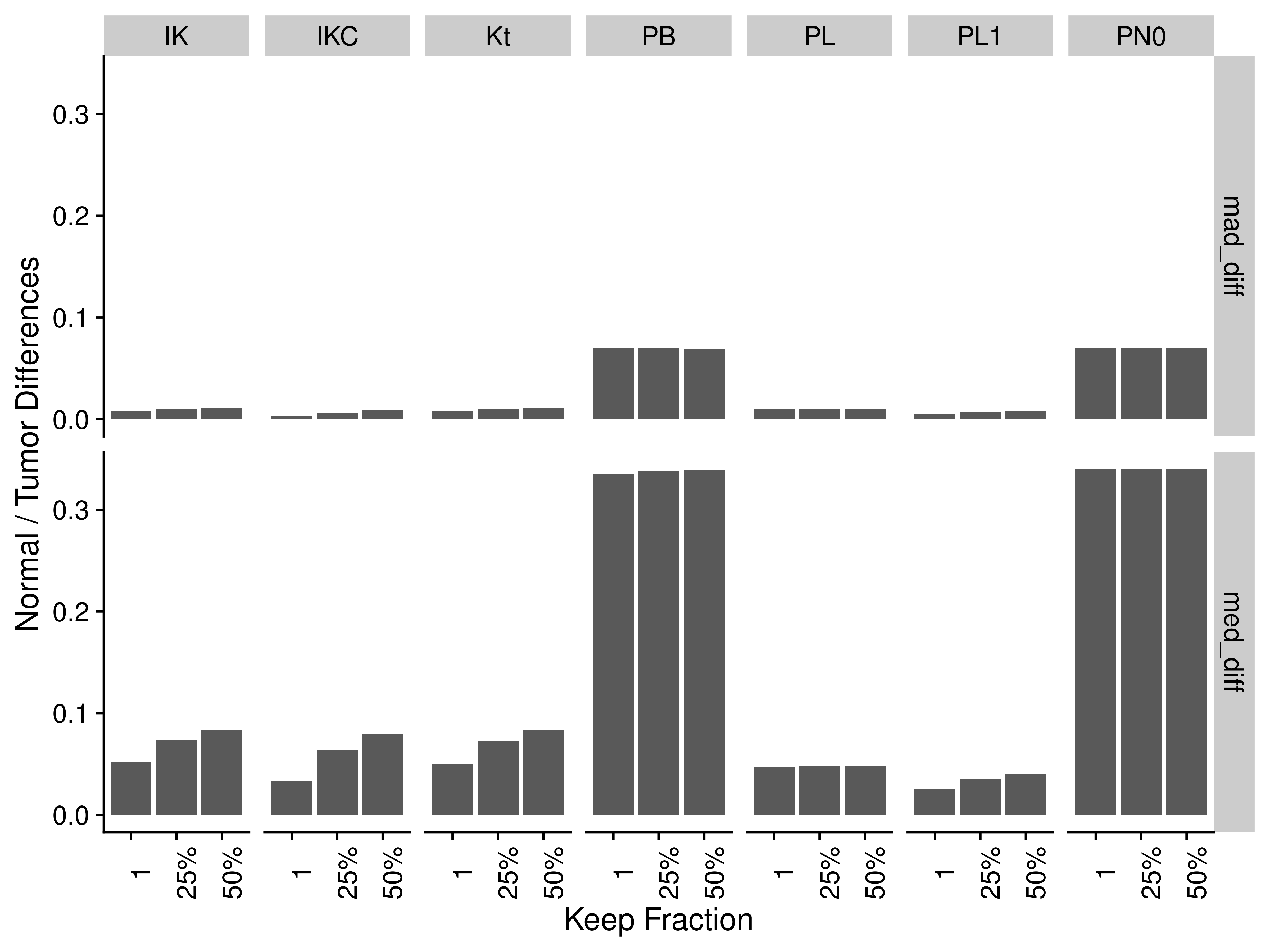


Figure S16. Difference of normal and tumor medians and MADs by correlation method and applying different fractional cutoffs for TCGA adenocarcinoma RNA-seq data. Abbreviations for different measures and data are: IK: ICI-Kt; IKC: ICI-Kt \* Completeness; Kt: Kendall-tau; PB: Pearson Base (raw values);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

## References

Gierliński, Marek, Christian Cole, Pietà Schofield, Nicholas J. Schurch, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, et al. 2015. “Statistical Models for RNA-Seq Data Derived from a Two-Condition 48-Replicate Experiment.” *Bioinformatics* 31 (22): 3625–30. <https://doi.org/10.1093/bioinformatics/btv425>.