

Supplementary Material

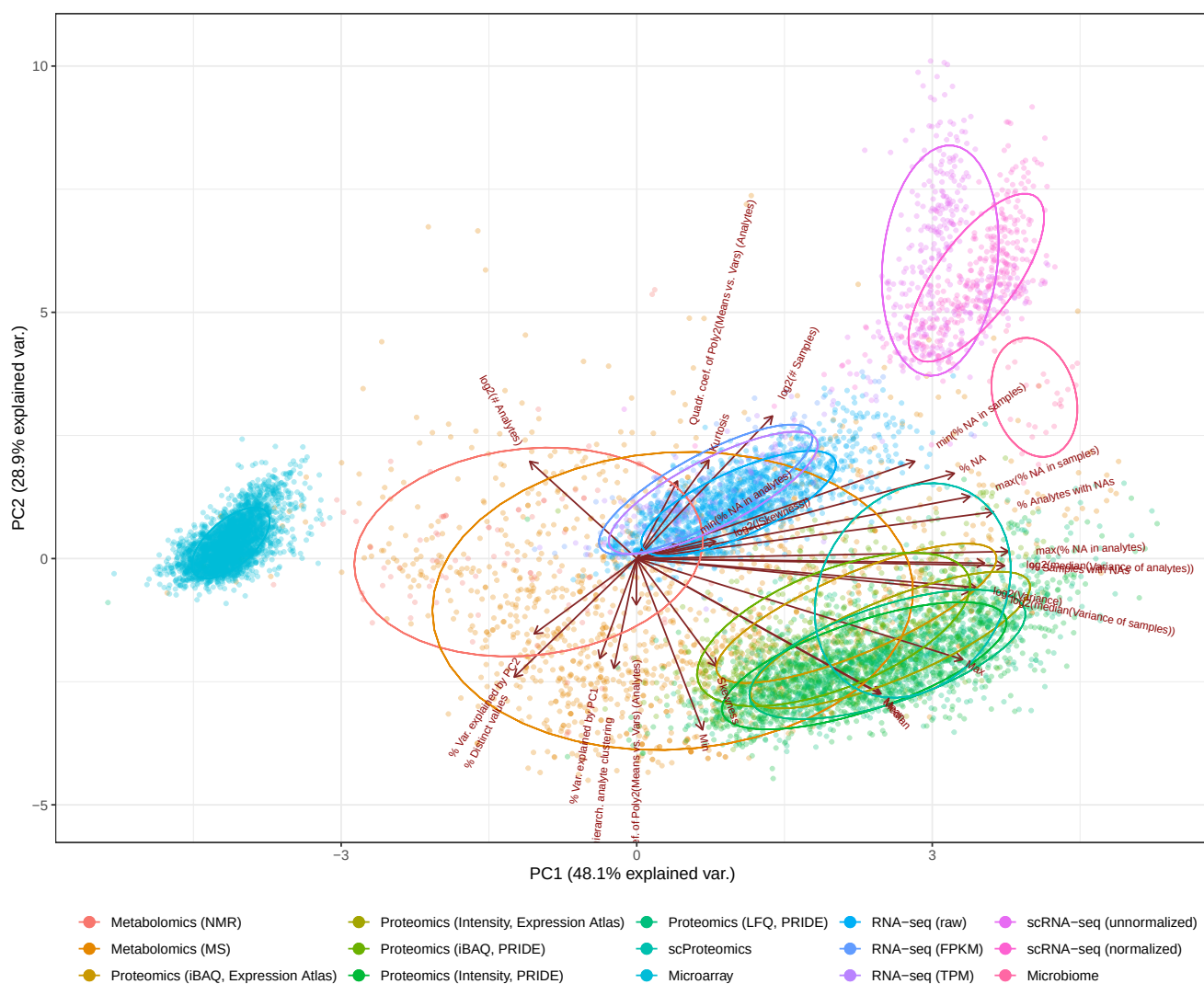


Figure S1. Clustering of the different data types using the singular value decomposition (SVD) implementation of the principal component analysis (PCA). As missing values are not allowed for SVD, due to many missing values the data characteristics 'Corr(Mean vs. % NA) (Samples)', 'Corr(Mean vs. % NA) (Analytes)', 'sd(Intensity w/ prob(NA) = 50% for sample)', and 'sd(Intensity w/ prob(NA) = 90% for sample)' were not included, and only those datasets with no missing values for the remaining data characteristics were included. The arrows indicate degree of contribution of the different data characteristics to the first two principal components (PCs), PC1 and PC2. The ellipses indicate the area which captures 68% of the datasets for a respective data type. One unnormalized scRNA-seq dataset and five metabolomics (MS) were lying outside of the depicted PC2 interval and are not included in this figure.

Table S1. Data type subgroups investigated in this study (with more than 5 entries). For further information regarding the higher-level data types see Table 1.

Data type subgroup	Alias	Higher-level data type	Number Of Datasets	Database Name
Metabolomics data generated by gas chromatography (GC)-mass spectrometry (MS)	Metabolomics (GC-MS)	Metabolomics (MS)	158	MetaboLights ¹³
Metabolomics data generated by liquid chromatography (LC)-MS	Metabolomics (LC-MS)	Metabolomics (MS)	991	MetaboLights ¹³
Metabolomics data generated by direct infusion (DI)-MS	Metabolomics (DI-MS)	Metabolomics (MS)	12	MetaboLights ¹³
Metabolomics data generated by flow injection analysis (FIA)-MS	Metabolomics (FIA-MS)	Metabolomics (MS)	22	MetaboLights ¹³
Label-free protein-level proteomics data generated in data-dependent acquisition (DDA) mode on a Thermo instrument and processed by MaxQuant ¹⁴	Proteomics (iBAQ, Intensity, LFQ], PRIDE, Thermo)	Proteomics (iBAQ, Intensity, LFQ], PRIDE)	iBAQ: 532, Intensity: 1419, LFQ: 1024	PRIDE ¹²
Label-free protein-level proteomics data generated in data-dependent acquisition (DDA) mode on a Bruker instrument and processed by MaxQuant ¹⁴	Proteomics (iBAQ, Intensity, LFQ], PRIDE, Bruker)	Proteomics (iBAQ, Intensity, LFQ], PRIDE)	iBAQ: 7, Intensity: 28, LFQ: 19	PRIDE ¹²
Label-free protein-level proteomics data generated in data-dependent acquisition (DDA) mode on a SCIEX instrument and processed by MaxQuant ¹⁴	Proteomics (iBAQ, Intensity, LFQ], PRIDE, SCIEX)	Proteomics (iBAQ, Intensity, LFQ], PRIDE)	iBAQ: 12, Intensity: 17, LFQ: 16	PRIDE ¹²
Normalized microarray data (normalized as is provided on Expression Atlas website https://www.ebi.ac.uk/gxa/home) generated using a Affymetrix microarray	Microarray (Affymetrix)	Microarray	2816	Expression Atlas ¹⁵
Normalized microarray data (normalized as is provided on Expression Atlas website https://www.ebi.ac.uk/gxa/home) generated using a Agilent microarray	Microarray (Agilent)	Microarray	127	Expression Atlas ¹⁵
Unnormalized single-cell RNA sequencing (scRNA-seq) data generated by SMART-like technologies	scRNA-seq (SMART-like, un-normalized)	scRNA-seq (unnormalized)	135	Single Cell Expression Atlas ^{15,18}
Unnormalized single-cell RNA sequencing (scRNA-seq) data generated by droplet-based technologies	scRNA-seq (Droplet-based, un-normalized)	scRNA-seq (unnormalized)	214	Single Cell Expression Atlas ^{15,18}
ScRNA-seq data generated by SMART-like technologies and normalised to counts per million	scRNA-seq (SMART-like, normalized)	scRNA-seq (normalized)	135	Single Cell Expression Atlas ^{15,18}
ScRNA-seq data generated by droplet-based technologies and normalised to counts per million	scRNA-seq (Droplet-based, normalized)	scRNA-seq (normalized)	214	Single Cell Expression Atlas ^{15,18}
Microbiome data containing taxon abundances from 16S amplicon studies	Microbiome (16S)	Microbiome	22	MicrobiomeDB ¹⁹
Microbiome data containing taxon abundances from whole genome shotgun sequencing (WGS) studies	Microbiome (WGS)	Microbiome	9	MicrobiomeDB ¹⁹

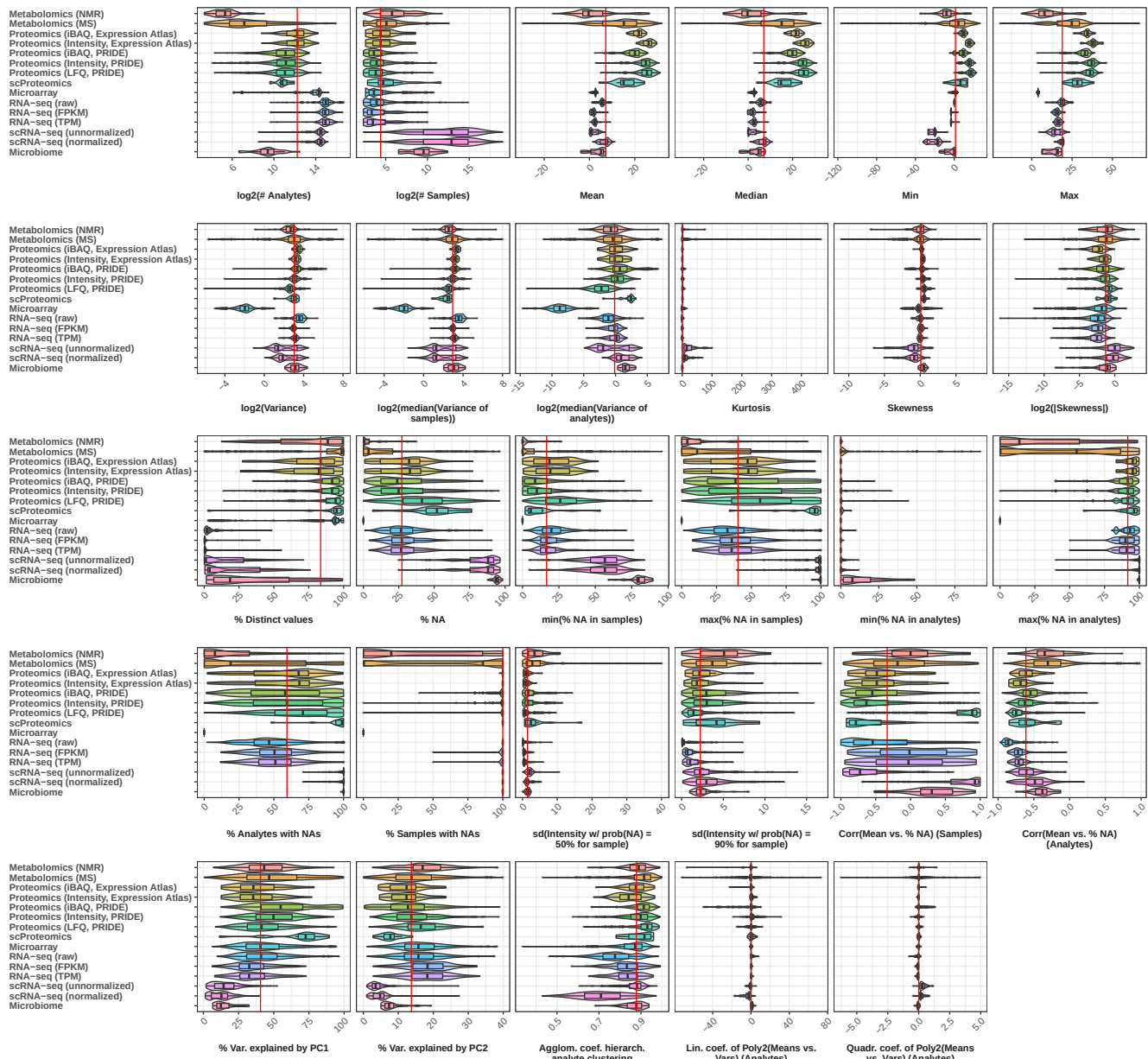


Figure S2. For most investigated data characteristics there are substantial differences between the data types. The vertical red lines represent the median of the median of each data type.

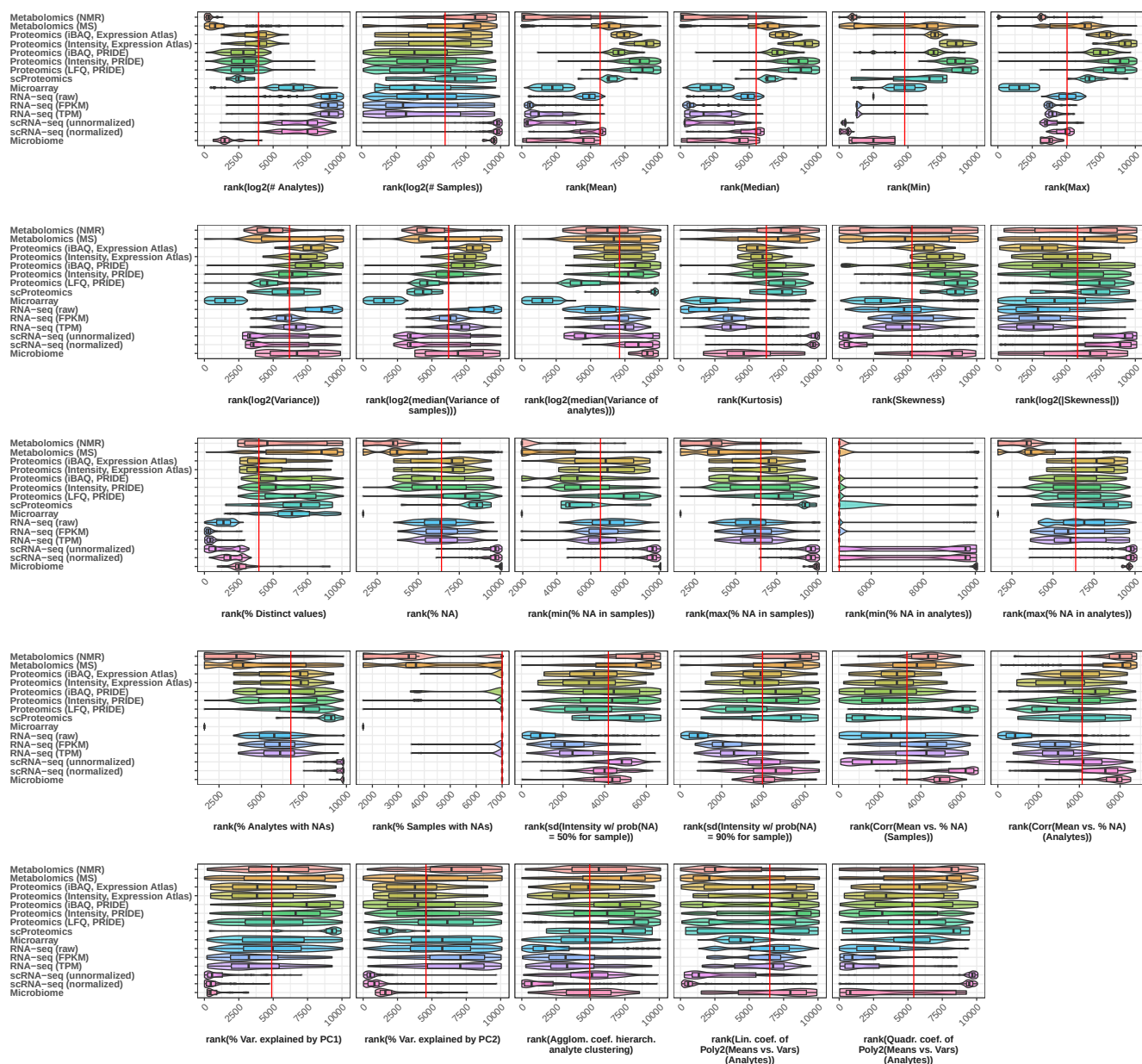


Figure S3. For most investigated data characteristics there are substantial differences between the data types, ranked for each data characteristic separately. The vertical red lines represent the median of the median of each data type.

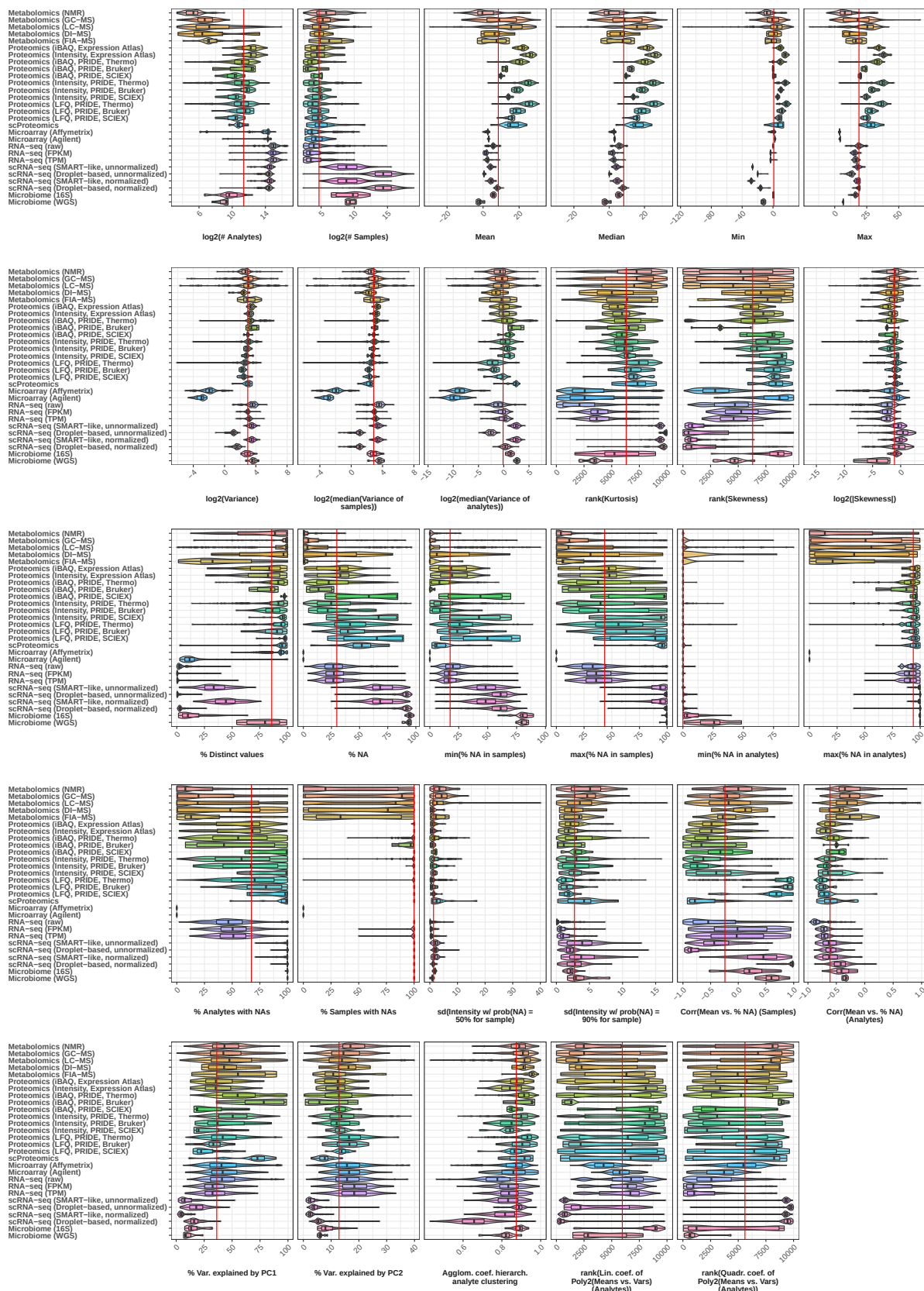


Figure S4. For most investigated data characteristics there are substantial differences between the data type subgroups. The vertical red lines represent the median of each data type. For better comparison, for Kurtosis, Skewness, Lin. coef. of Poly2(Mean vs. Vars)(Analytes), and Quadr. coef. of Poly2(Mean vs. Vars)(Analytes) the rank is displayed.

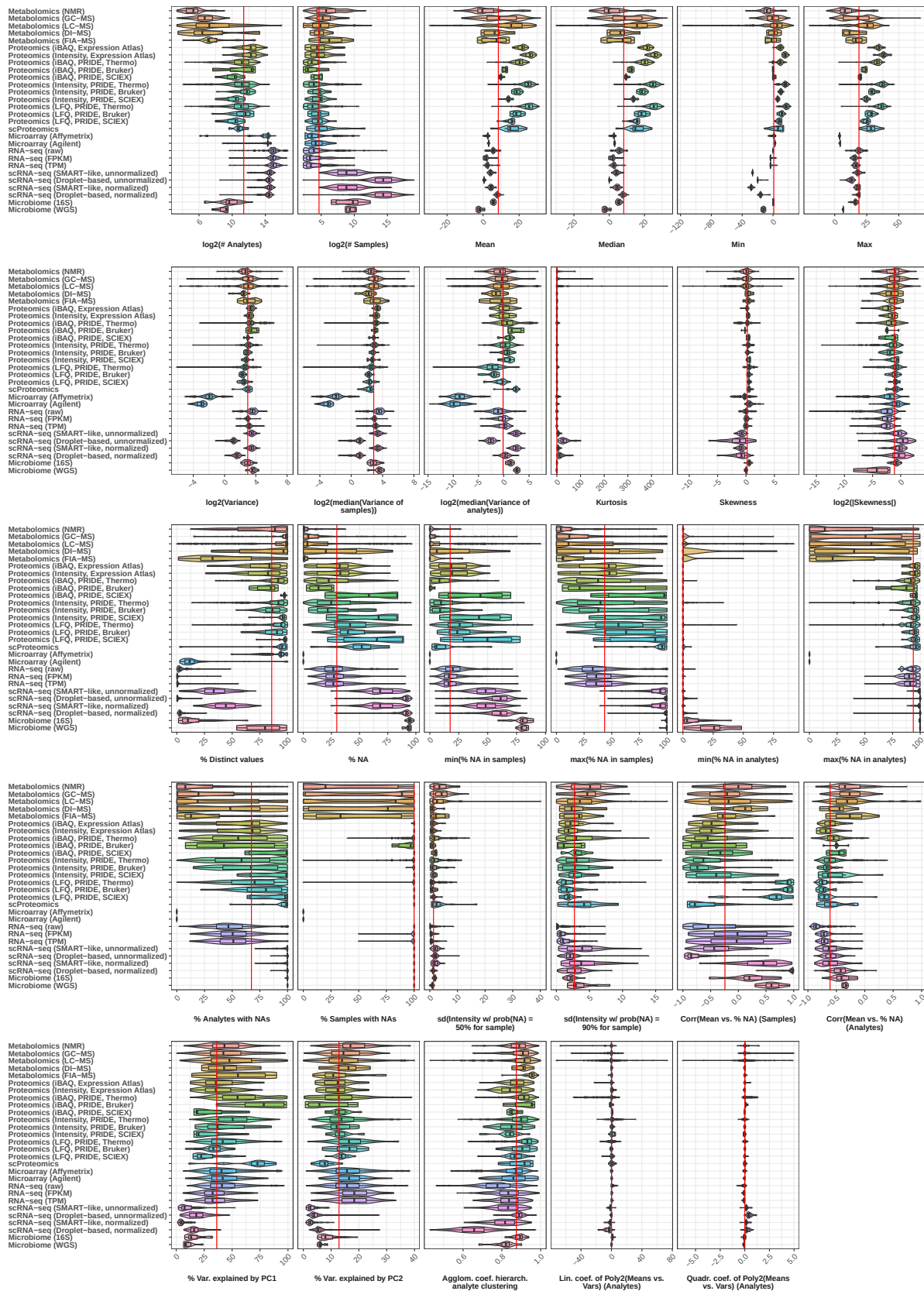


Figure S5. For most investigated data characteristics there are substantial differences between the data type subgroups. The vertical red lines represent the median of the median of each data type.

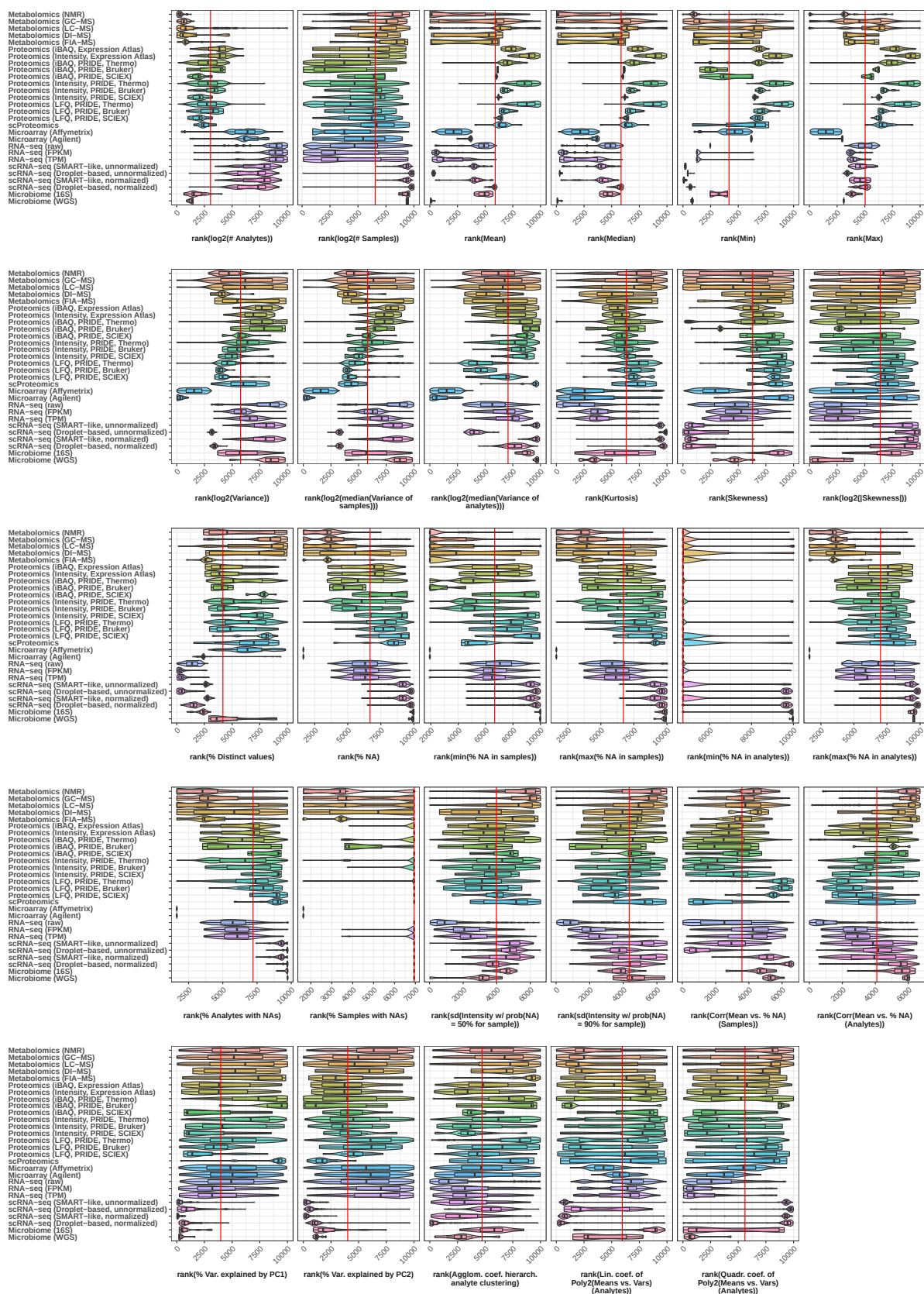


Figure S6. For most investigated data characteristics there are substantial differences between the data type subgroups, ranked for each data characteristic separately. The vertical red lines represent the median of the median of each data type.

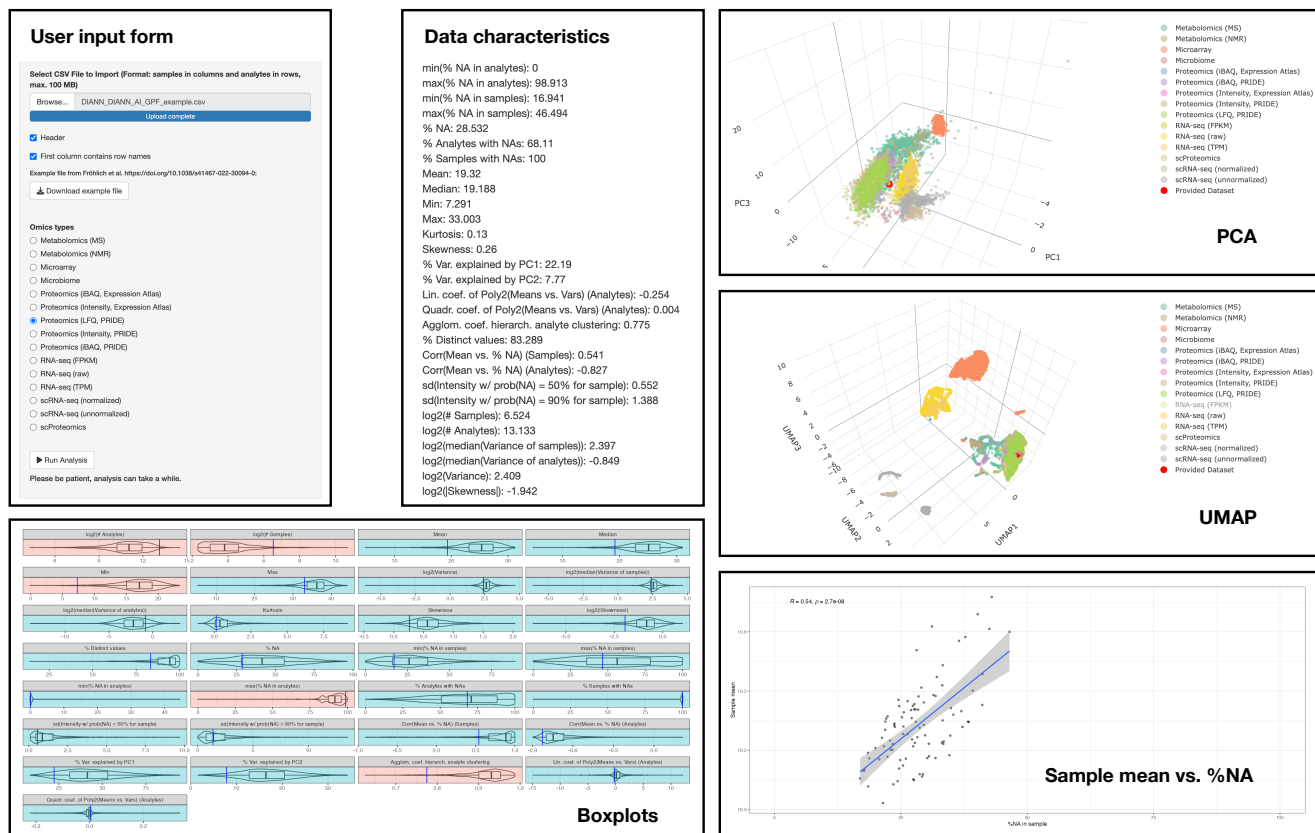
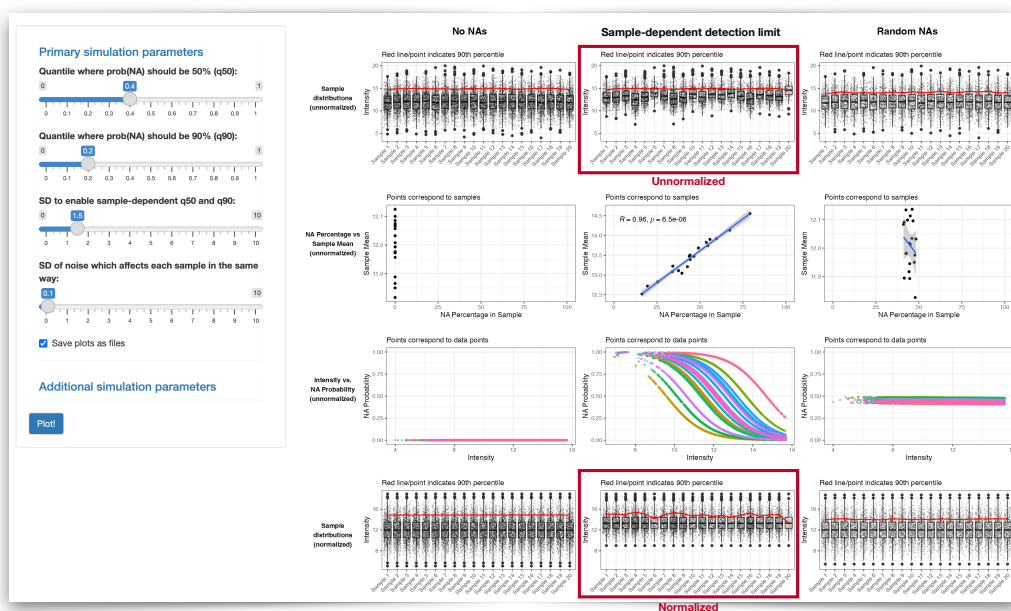


Figure S7. Components of the tool for assessing how representative a provided dataset is for a specific omics discipline based on its data characteristics, which is accessible at <https://omicscharacterization.imbi.uni-freiburg.de/>. Using the user input form, users can upload a dataset, which should be assessed, specifying the omics discipline to which it should be compared against using boxplots, showing a green background if the dataset lies between the 5th and 95th percentile box, and a red background otherwise. In addition, the data characteristics itself are displayed, complemented by principal component analysis(PCA) and uniform manifold approximation and projection (UMAP) plots which show where the provided dataset fits within the clusters of datasets of the different omics disciplines. Finally, the mean of a sample and the percentage of missing values in this sample are plotted against each other for the provided dataset, to assess possible correlations indicating if normalization should be performed.

(a)



(b)

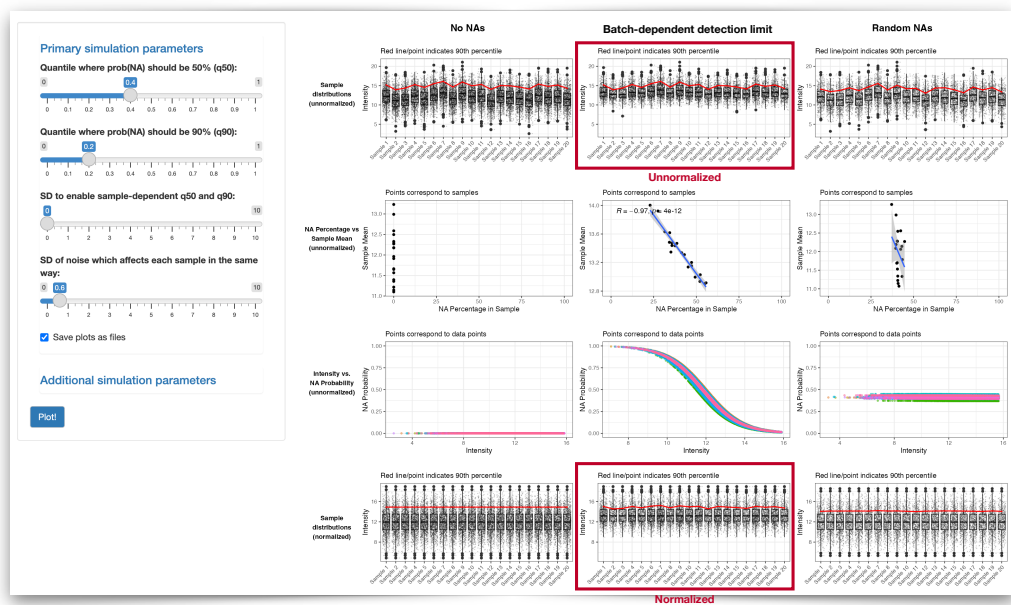


Figure S8. Simulation of datasets with (a) sample- and (b) batch-dependent detection limits accessible at <https://missingvaluesimulation.imbi.uni-freiburg.de>.

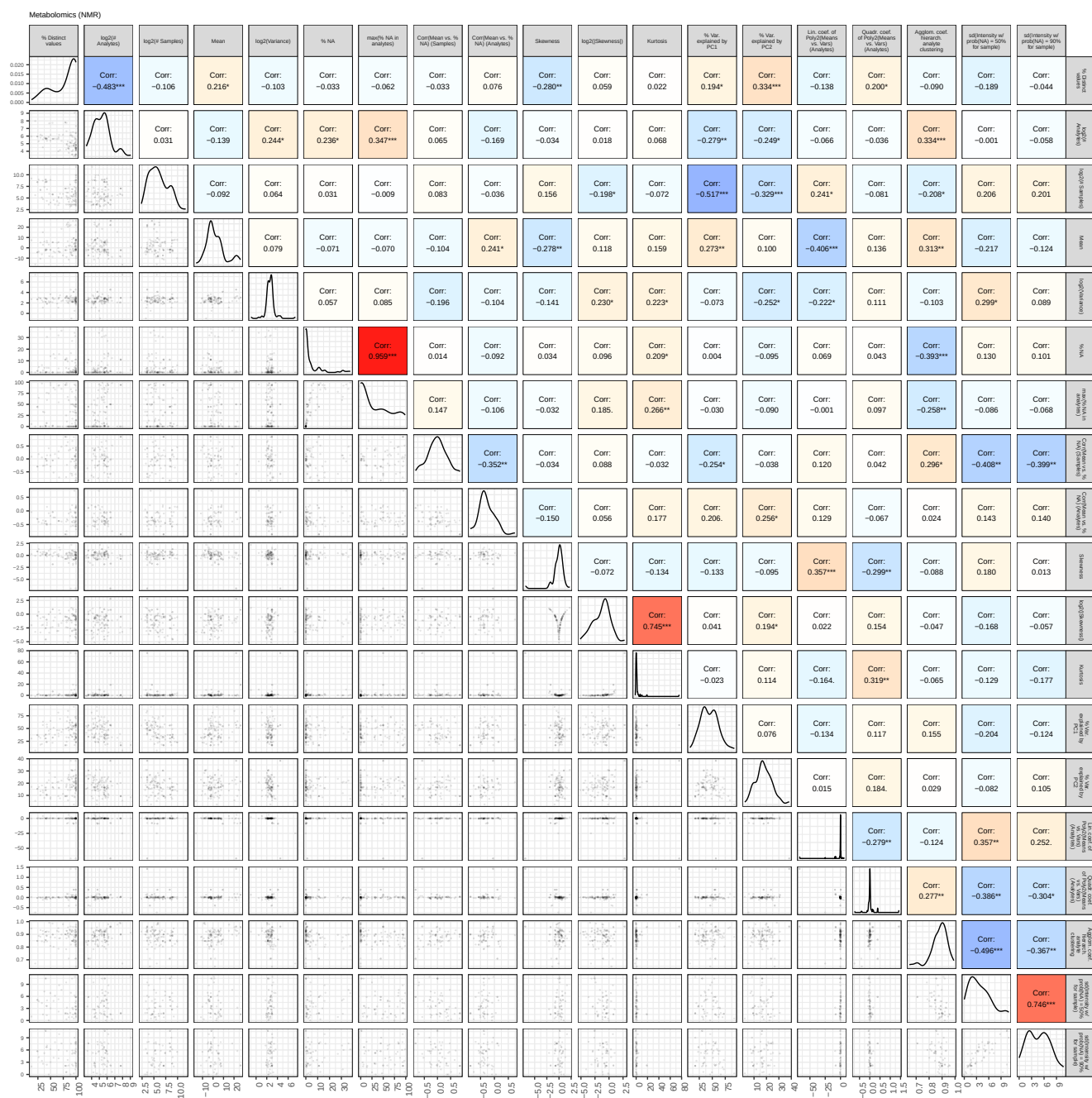


Figure S9. Spearman correlation plot of selected data characteristics over Metabolomics NMR datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

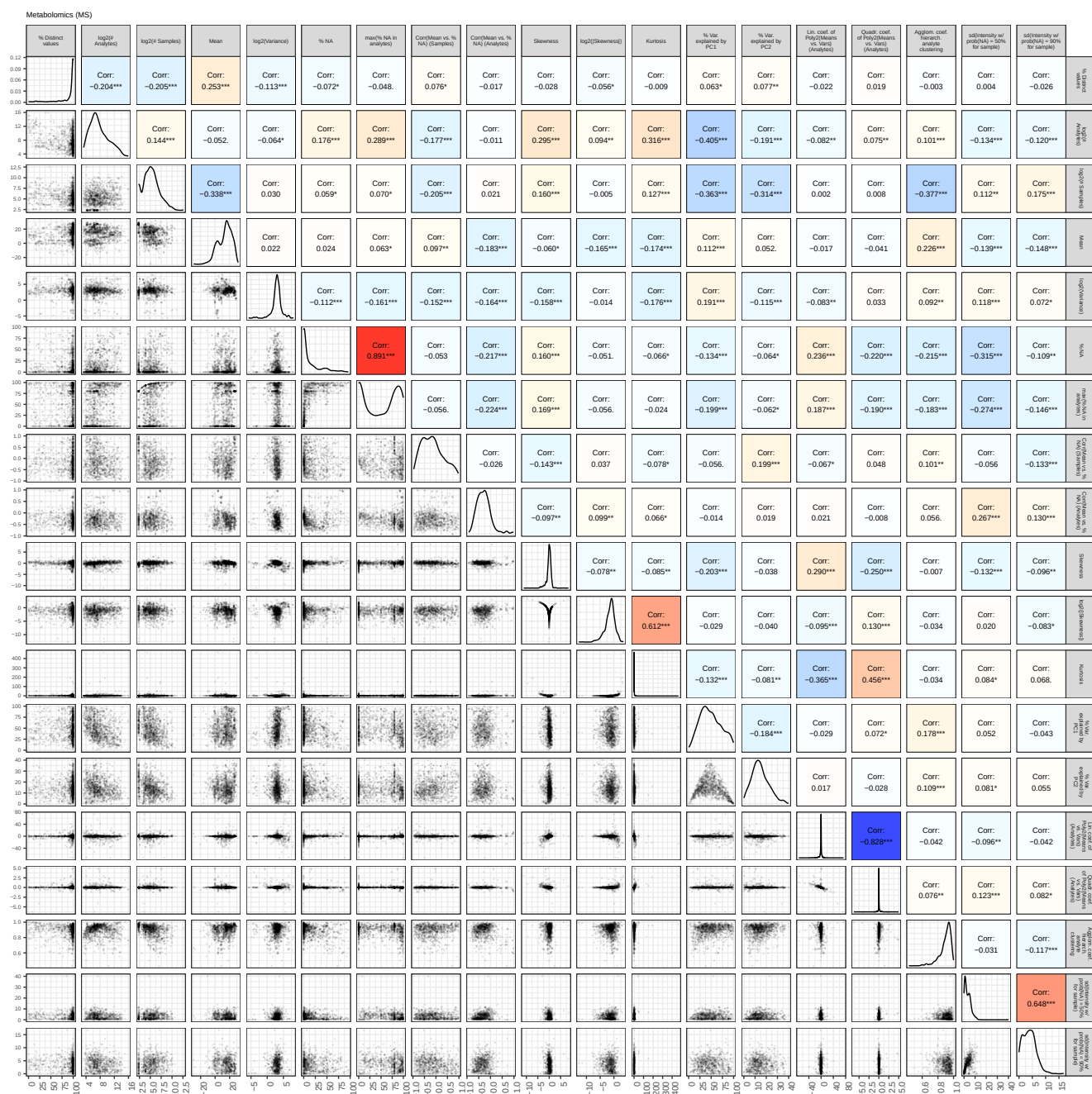


Figure S10. Spearman correlation plot of selected data characteristics over Metabolomics MS datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

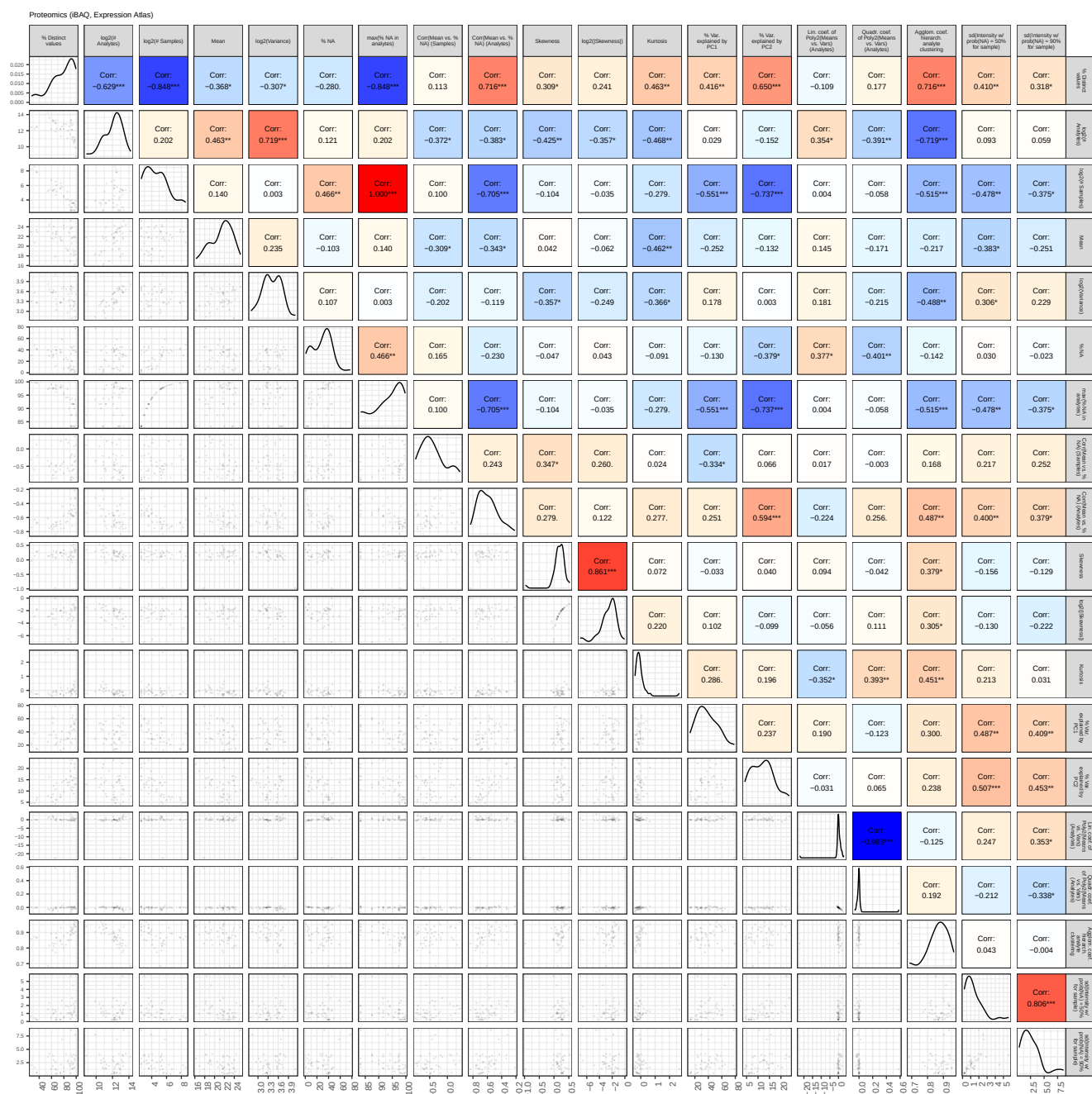


Figure S11. Spearman correlation plot of selected data characteristics over Proteomics (iBAQ, Expression Atlas) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

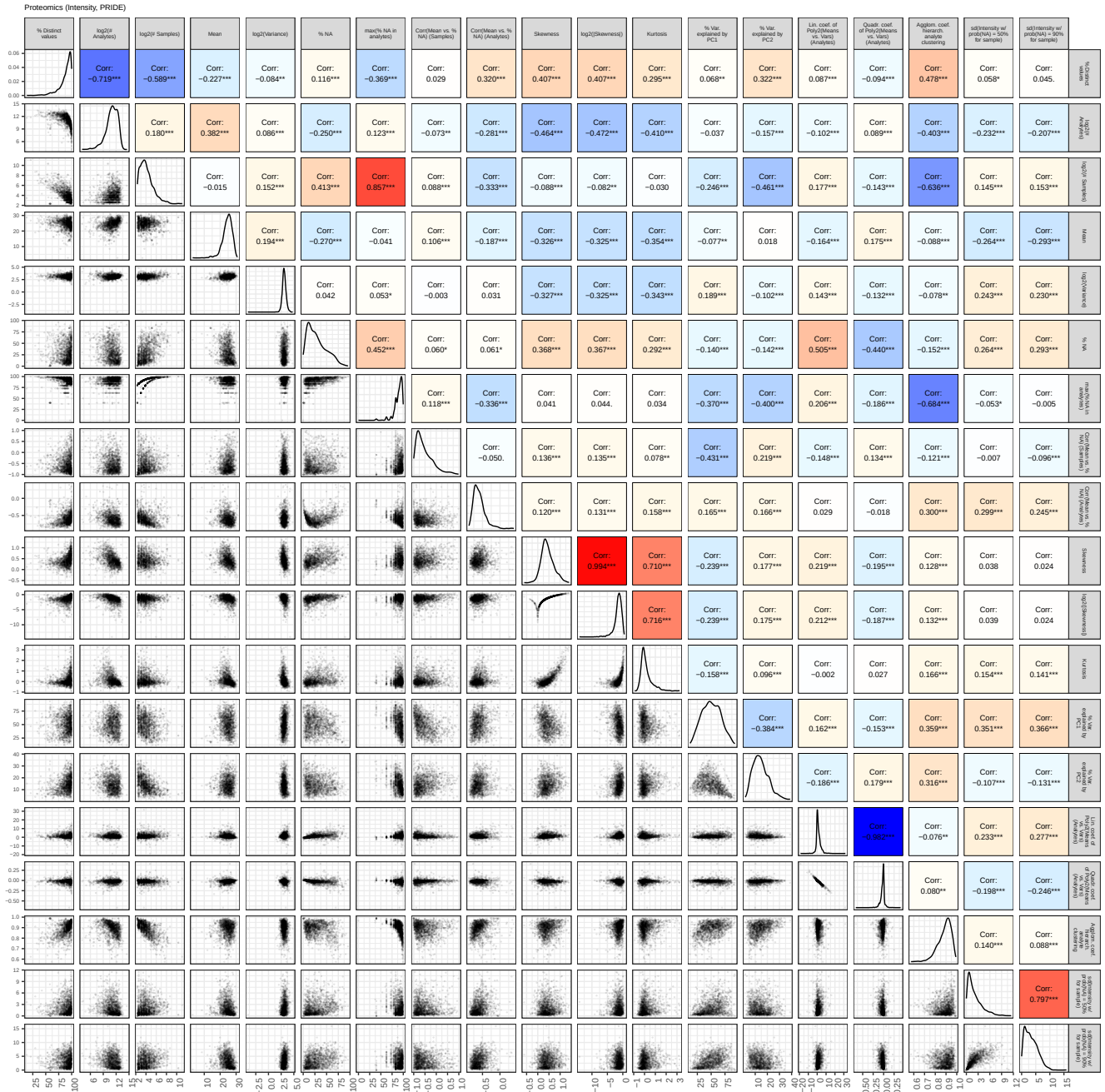


Figure S14. Spearman correlation plot of selected data characteristics over Proteomics (Intensity, PRIDE) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

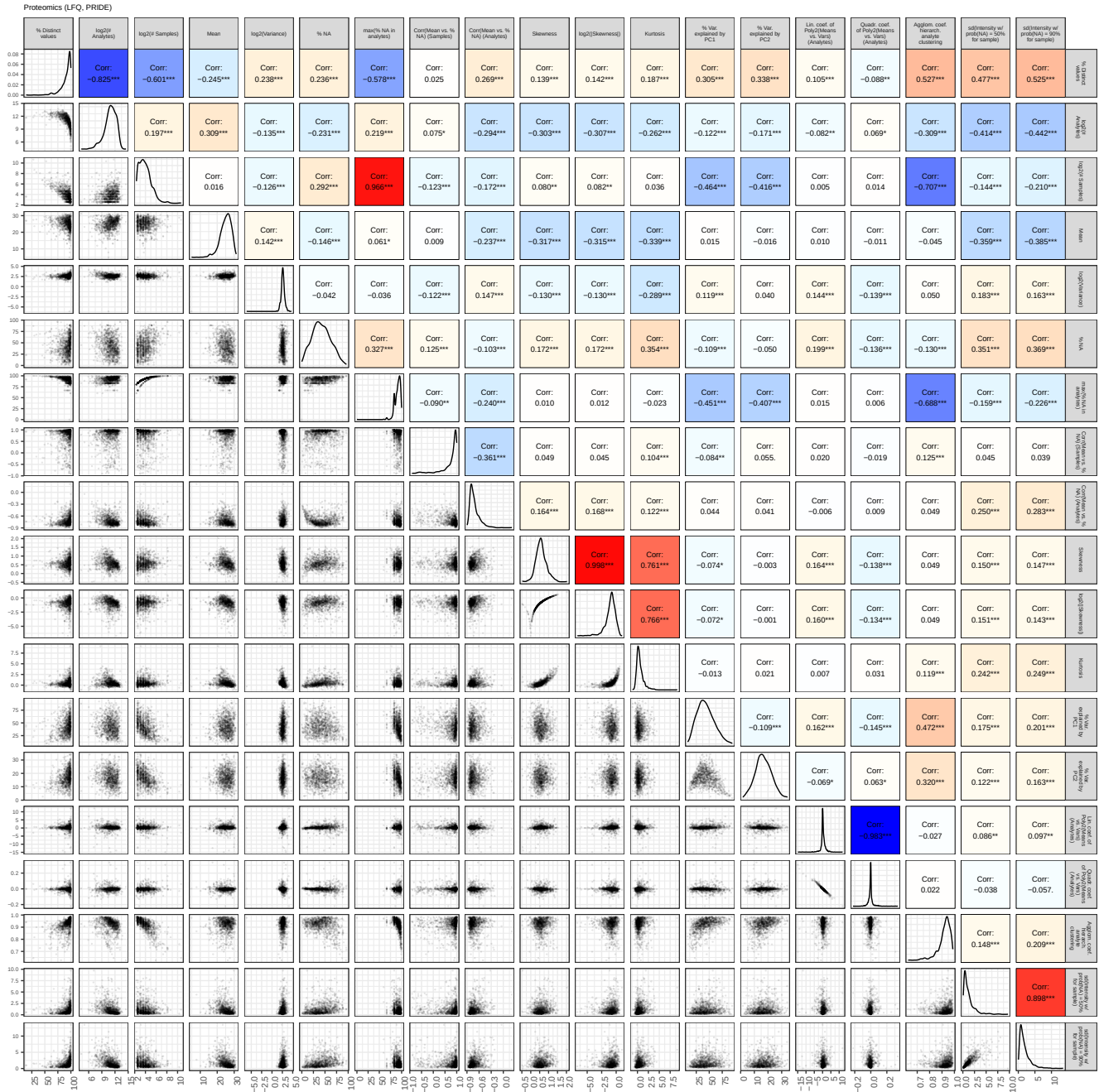


Figure S15. Spearman correlation plot of selected data characteristics over Proteomics (LFQ, PRIDE) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

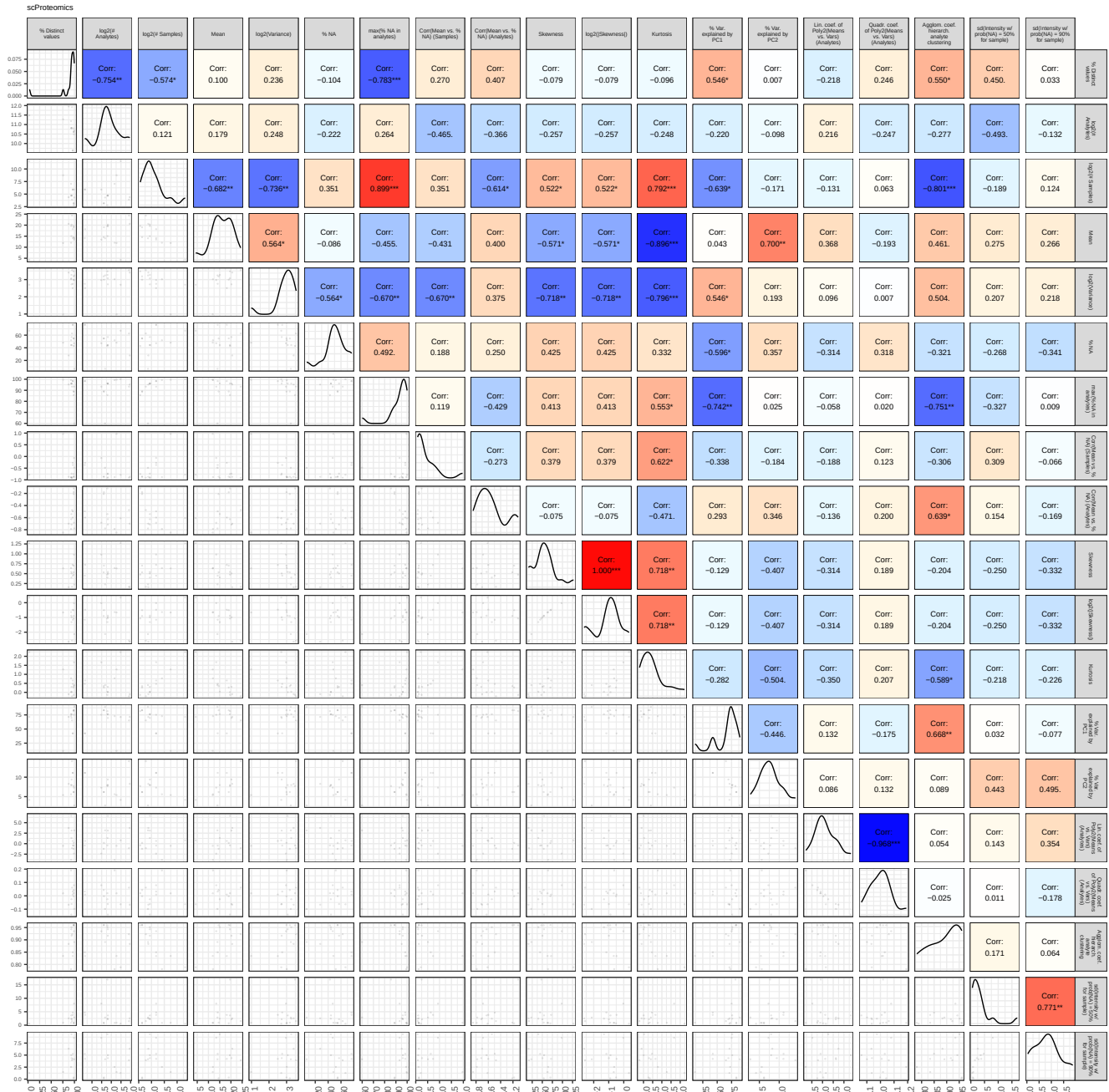


Figure S16. Spearman correlation plot of selected data characteristics over scProteomics datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

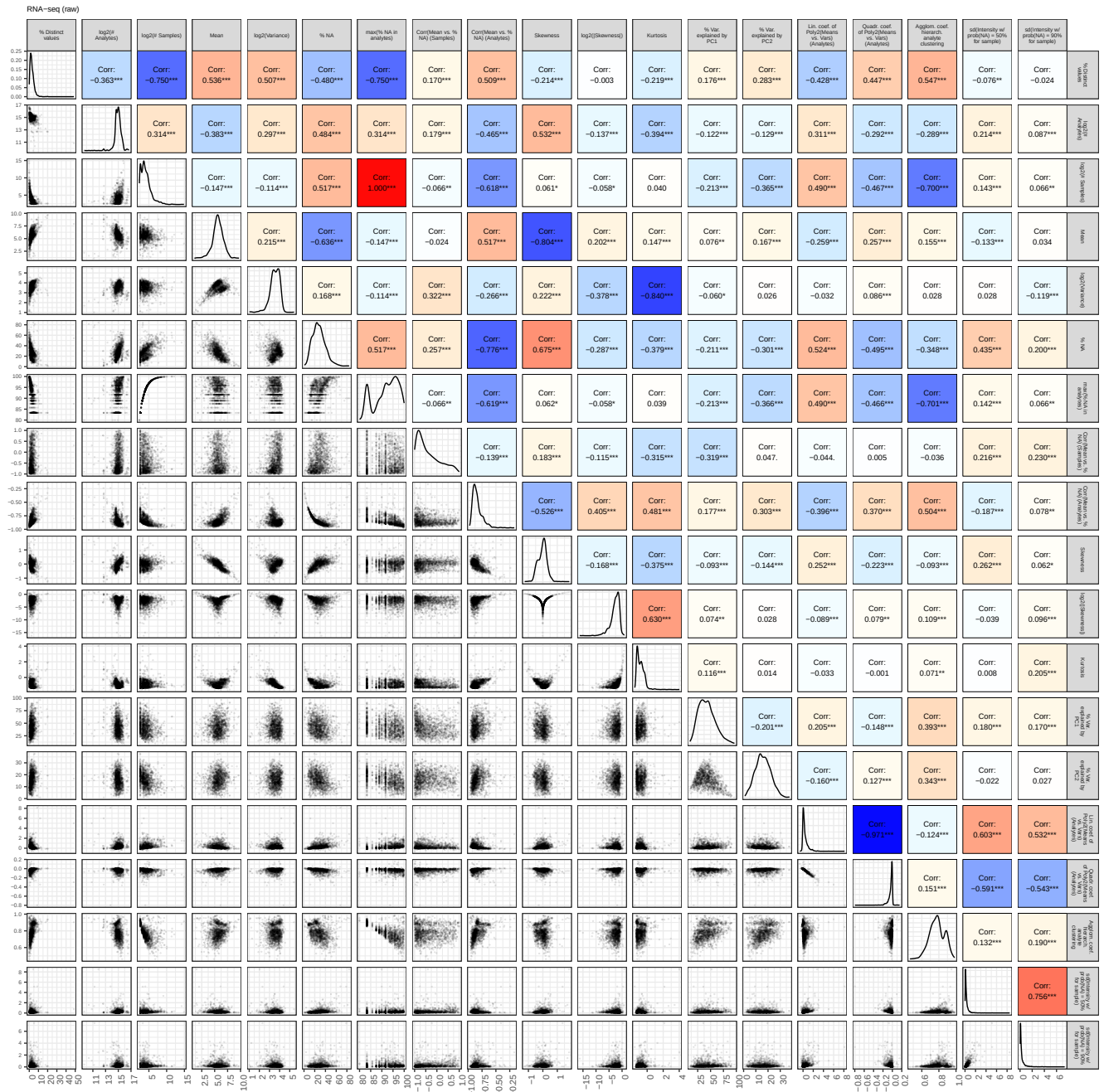


Figure S18. Spearman correlation plot of selected data characteristics over RNA-seq (raw) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

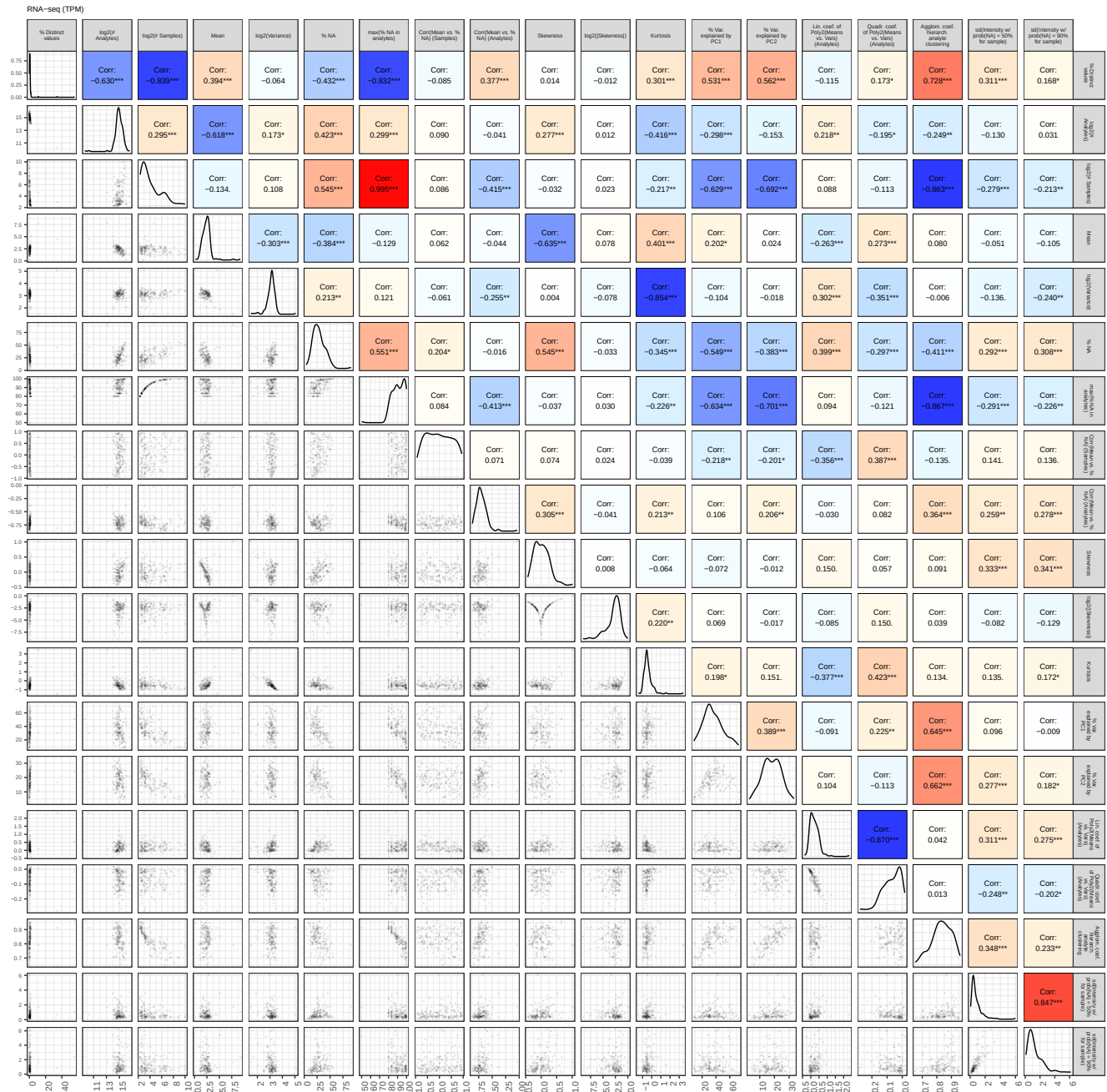


Figure S20. Spearman correlation plot of selected data characteristics over RNA-seq (TPM) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

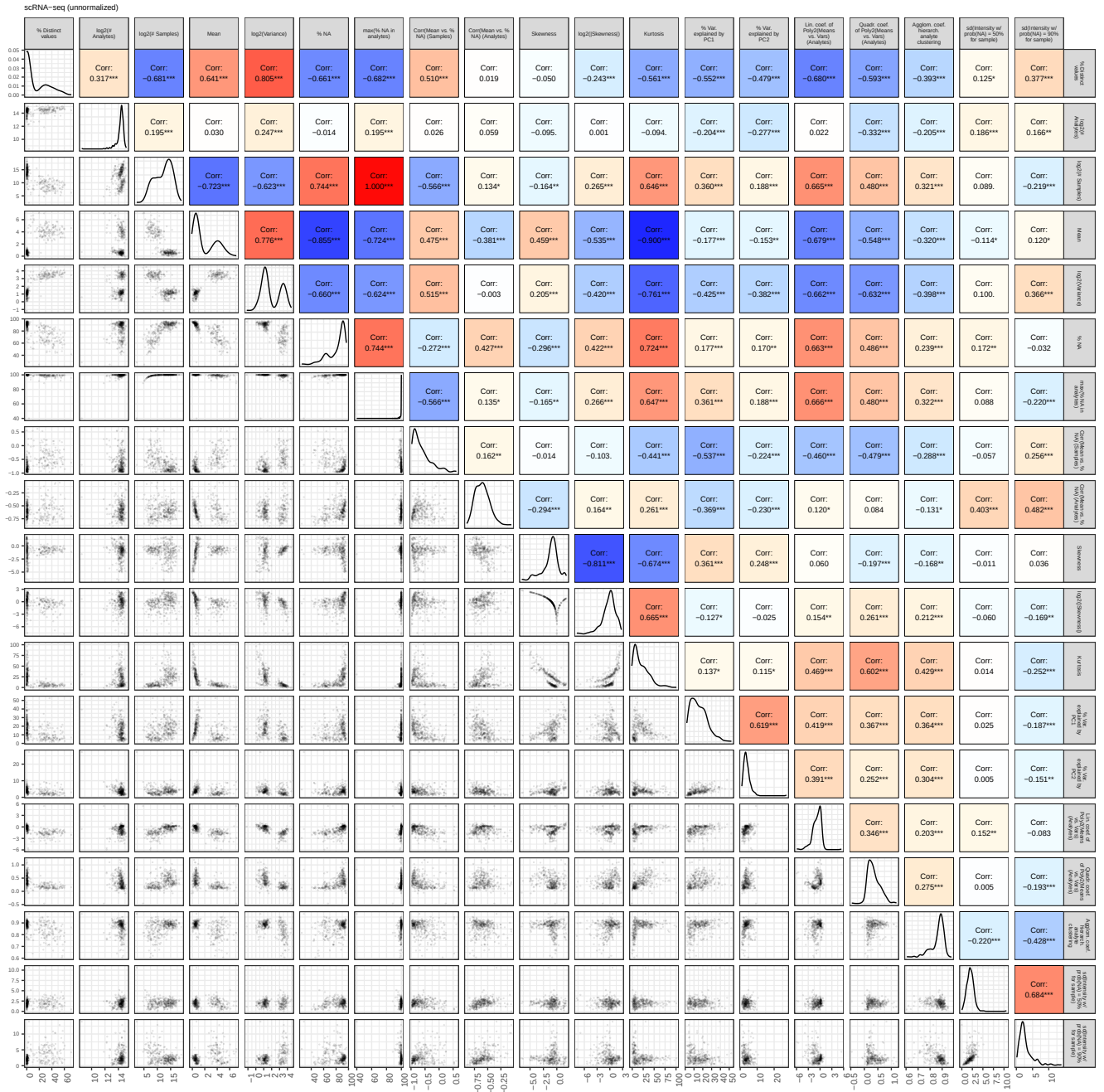


Figure S21. Spearman correlation plot of selected data characteristics over scRNA-seq (unnormalized) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

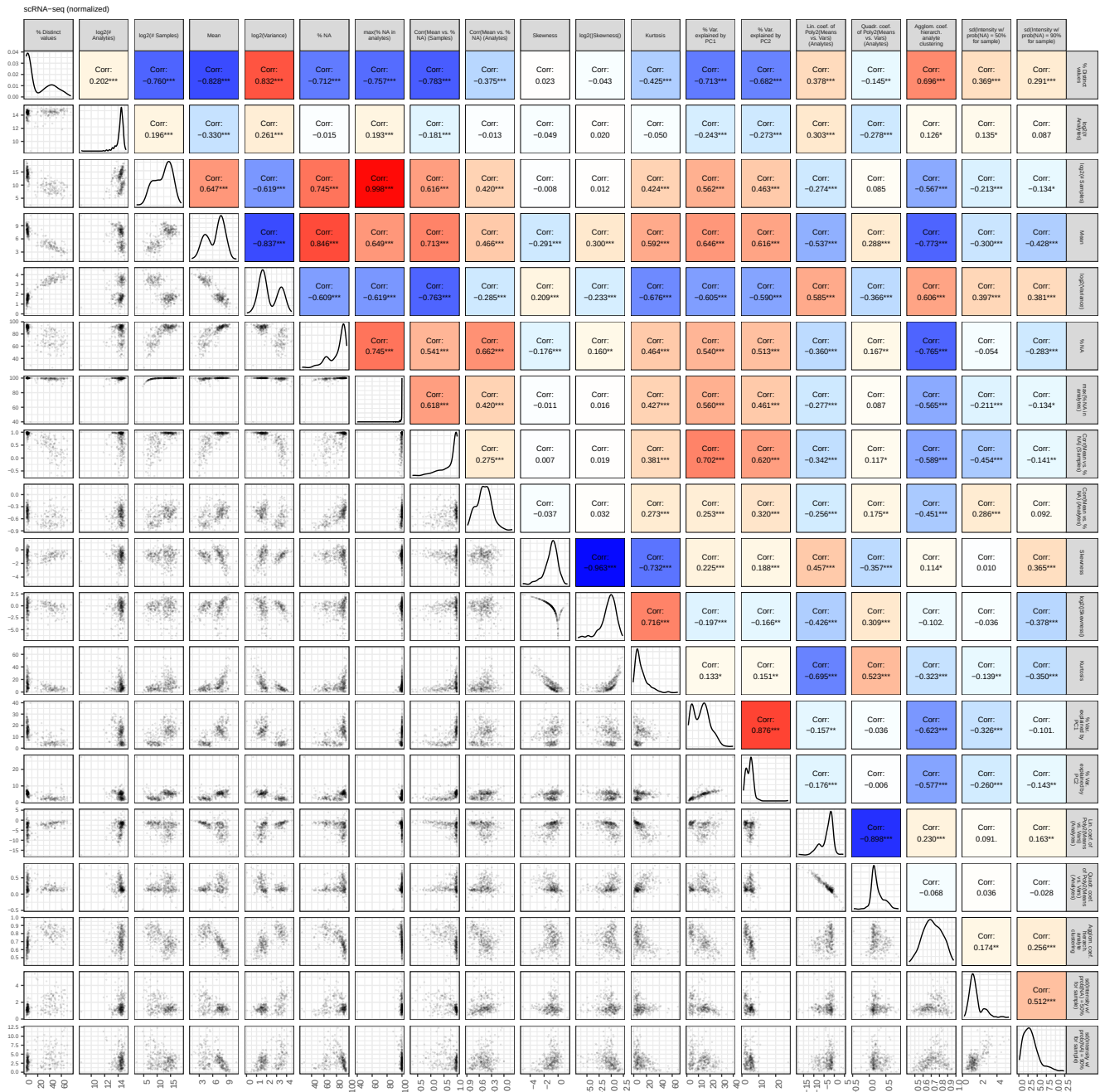
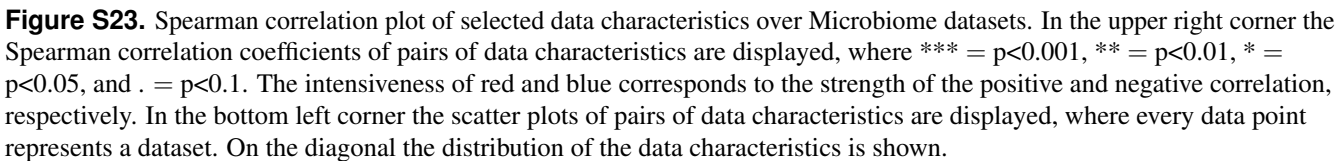


Figure S22. Spearman correlation plot of selected data characteristics over scRNA-seq (normalized) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.



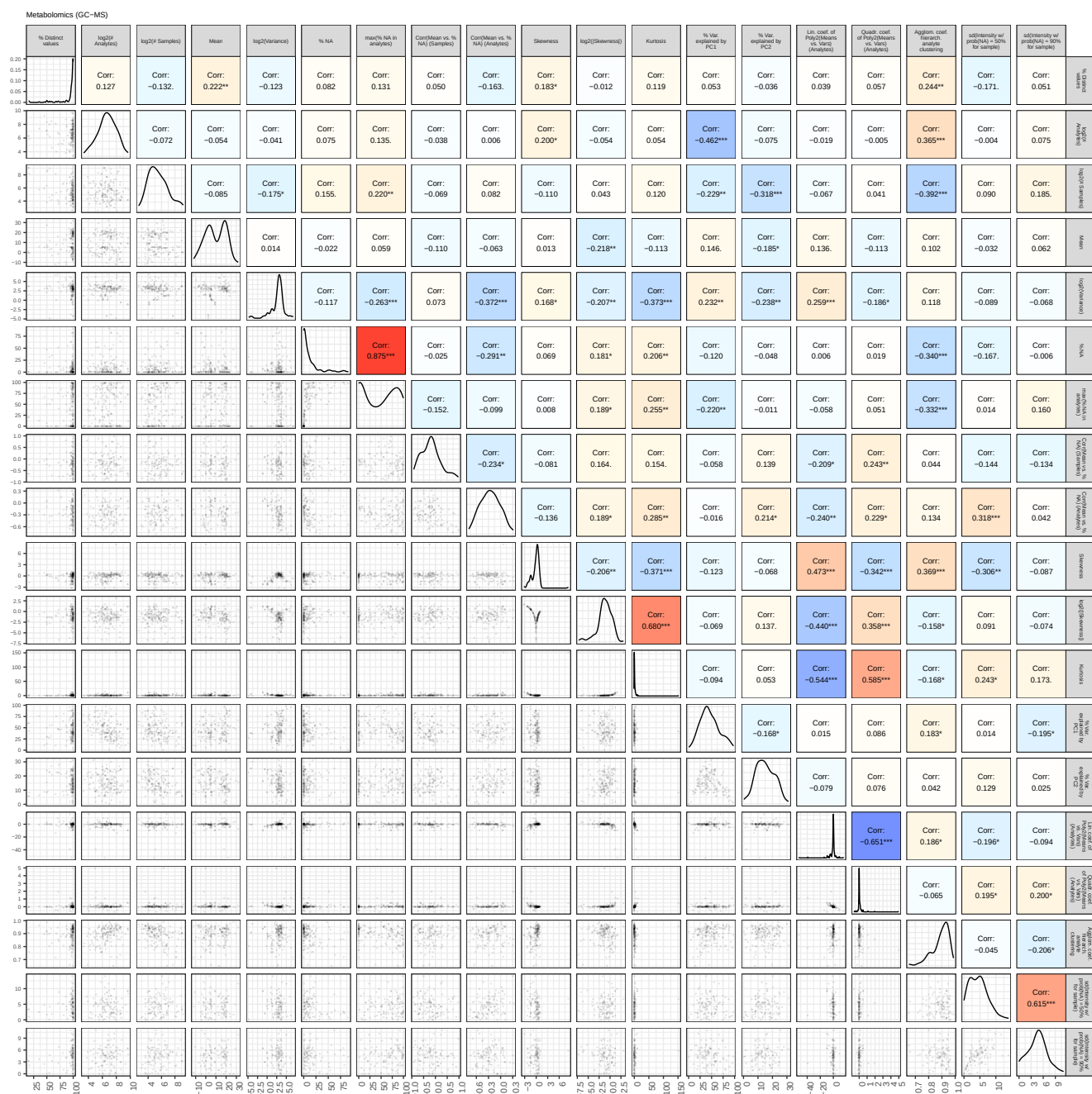


Figure S24. Spearman correlation plot of selected data characteristics over Metabolomics GC-MS datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

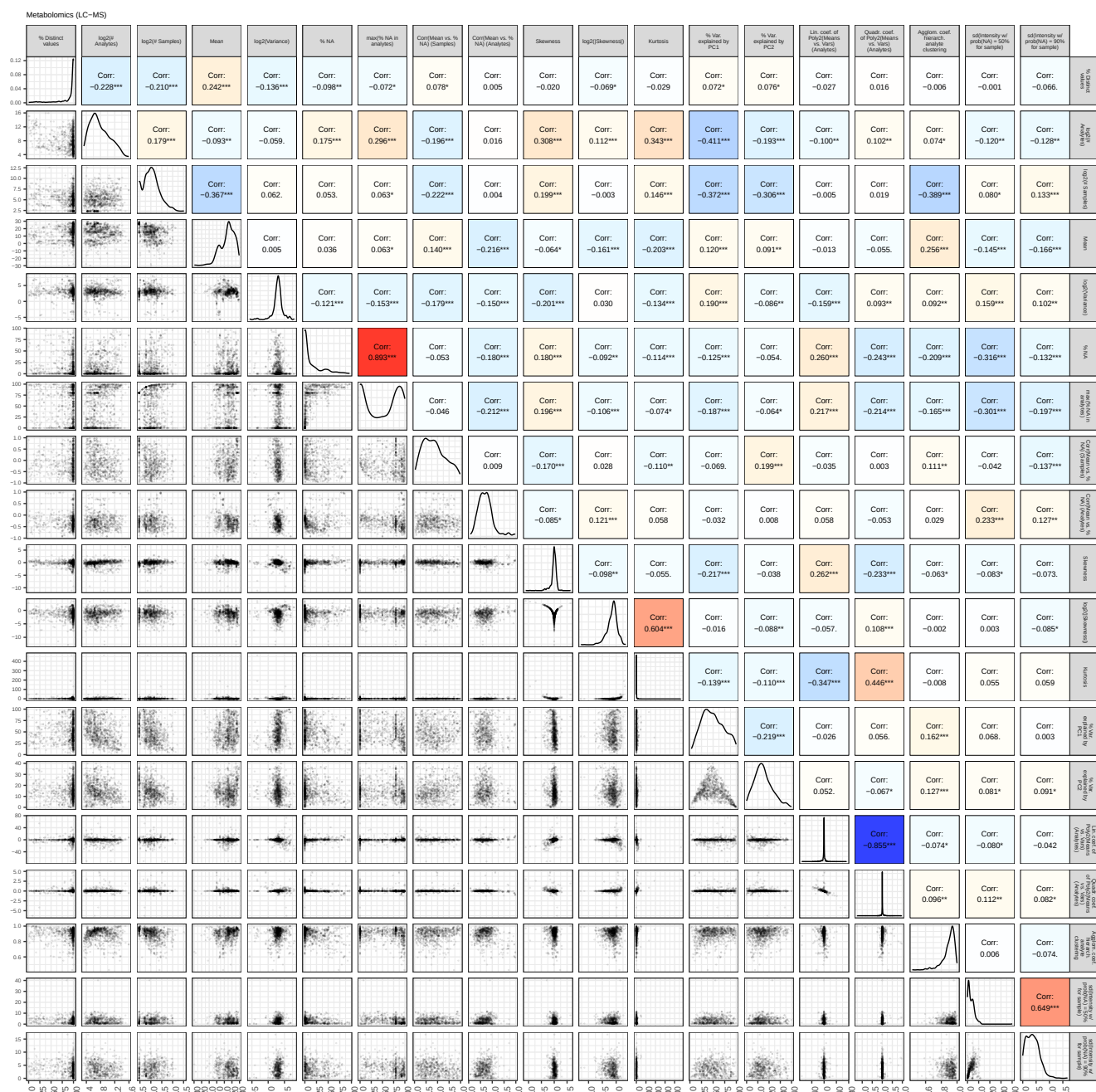


Figure S25. Spearman correlation plot of selected data characteristics over Metabolomics LC-MS datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

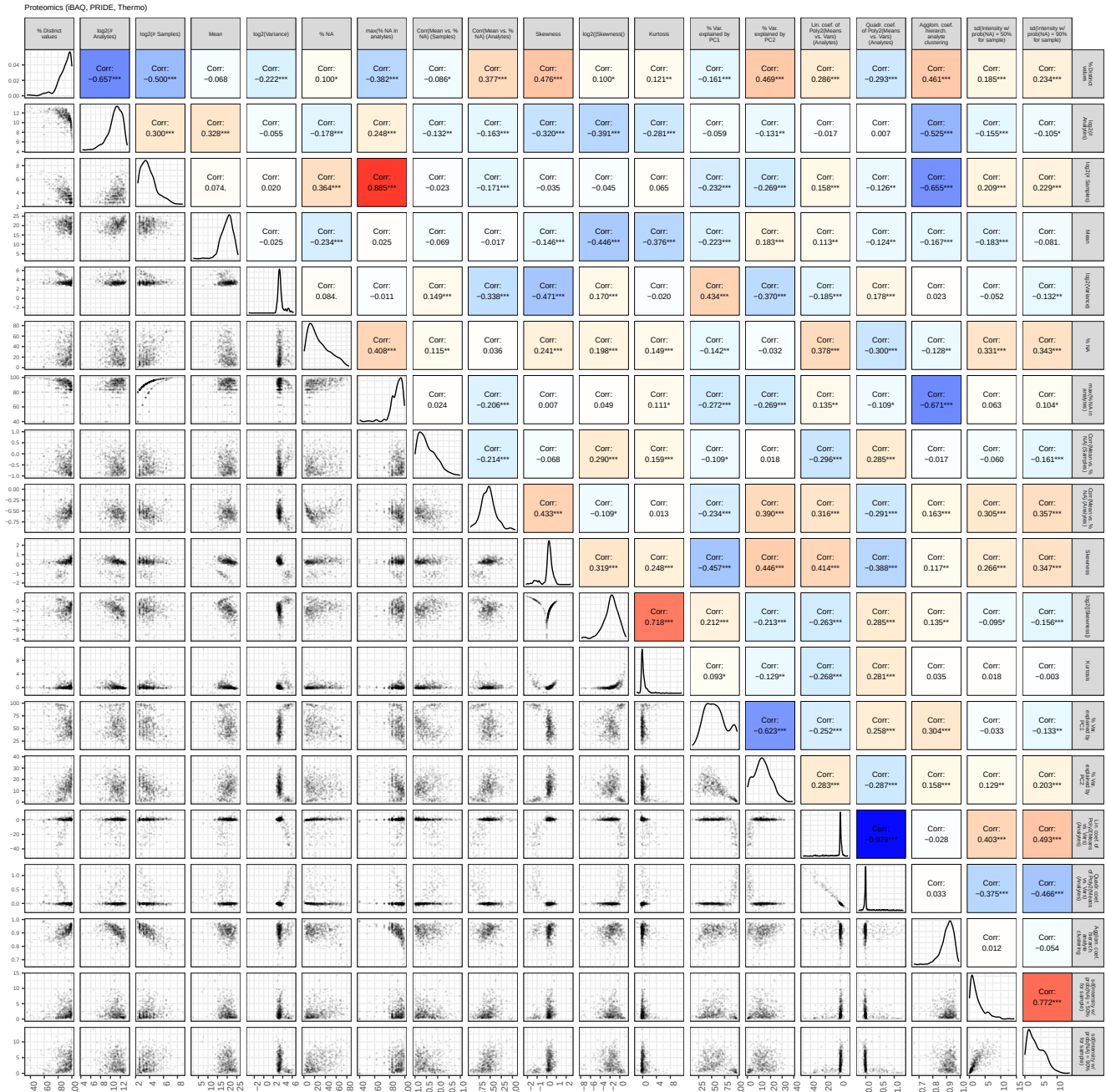


Figure S28. Spearman correlation plot of selected data characteristics over Proteomics (iBAQ, PRIDE, Thermo) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

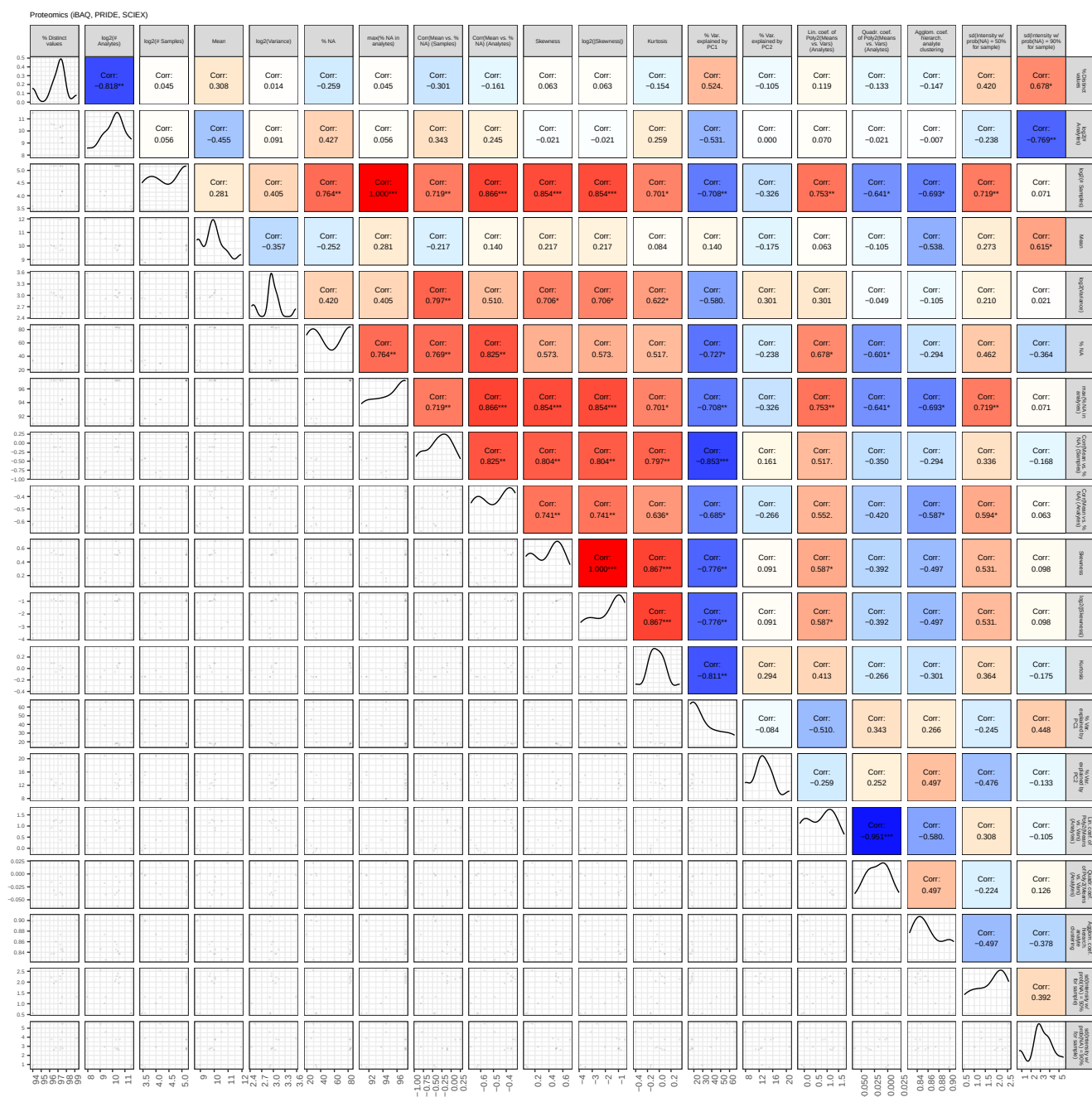


Figure S30. Spearman correlation plot of selected data characteristics over Proteomics (iBAQ, PRIDE, SCIEX) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

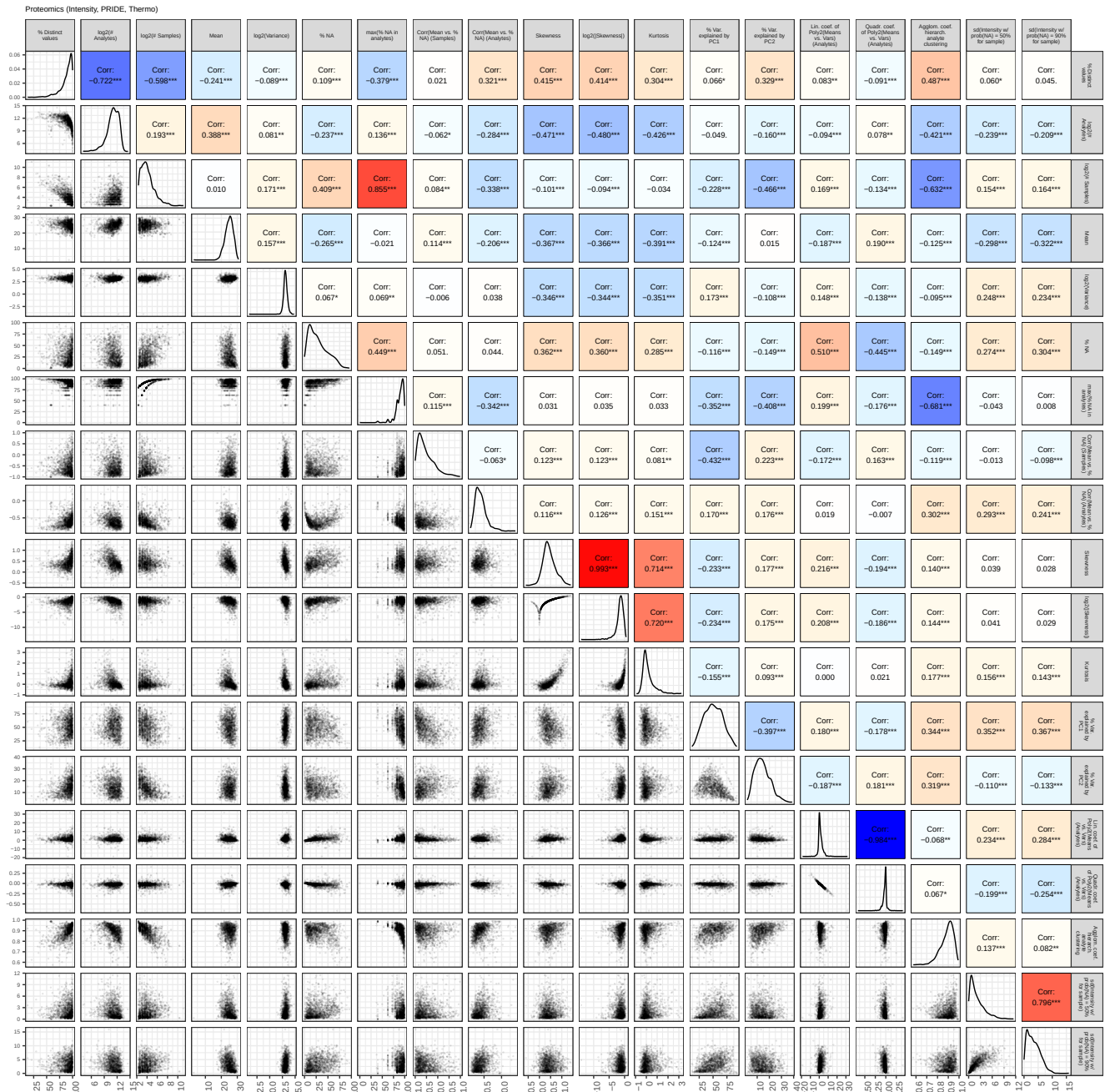


Figure S31. Spearman correlation plot of selected data characteristics over Proteomics (Intensity, PRIDE, Thermo) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

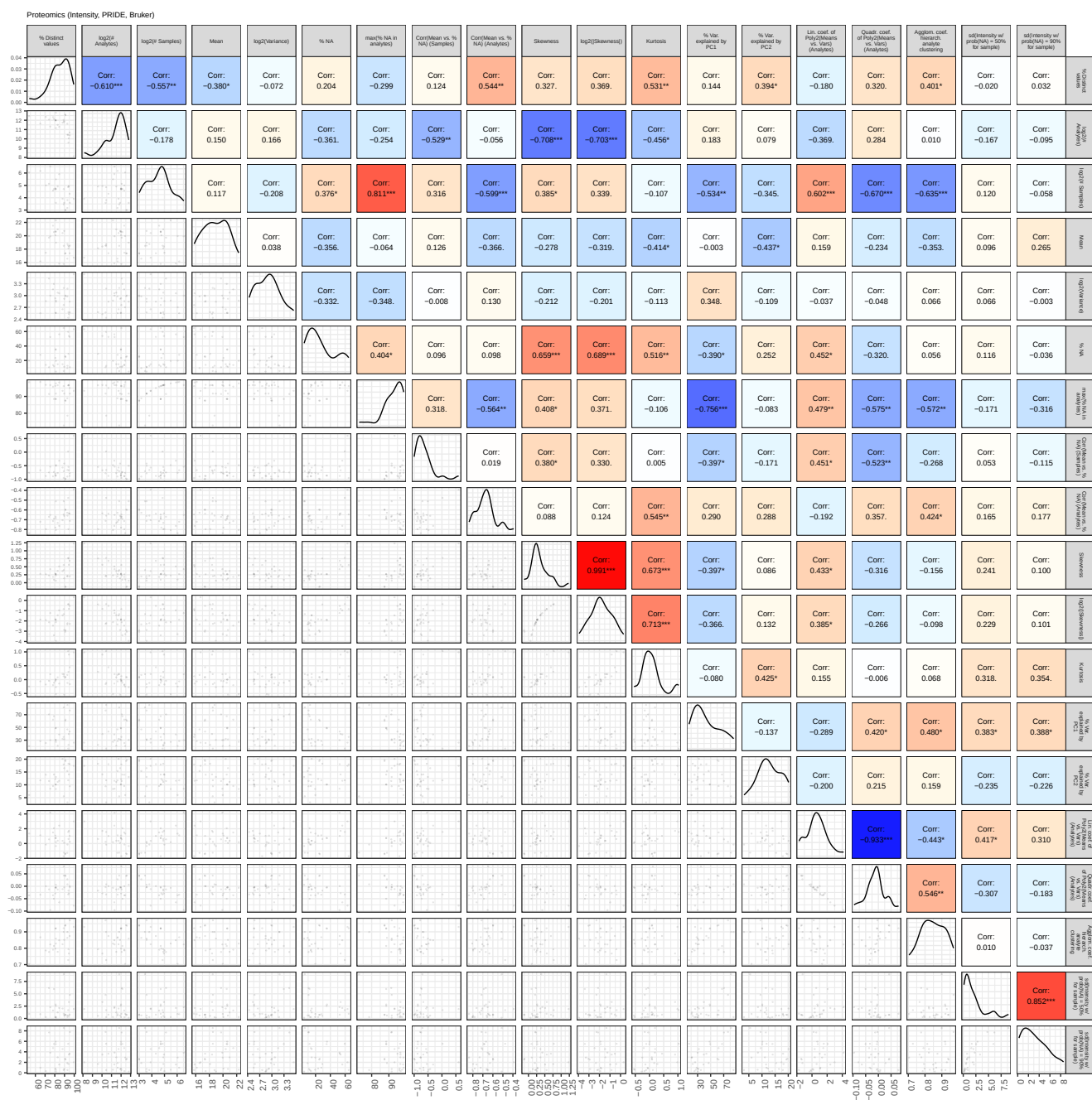


Figure S32. Spearman correlation plot of selected data characteristics over Proteomics (Intensity, PRIDE, Bruker) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

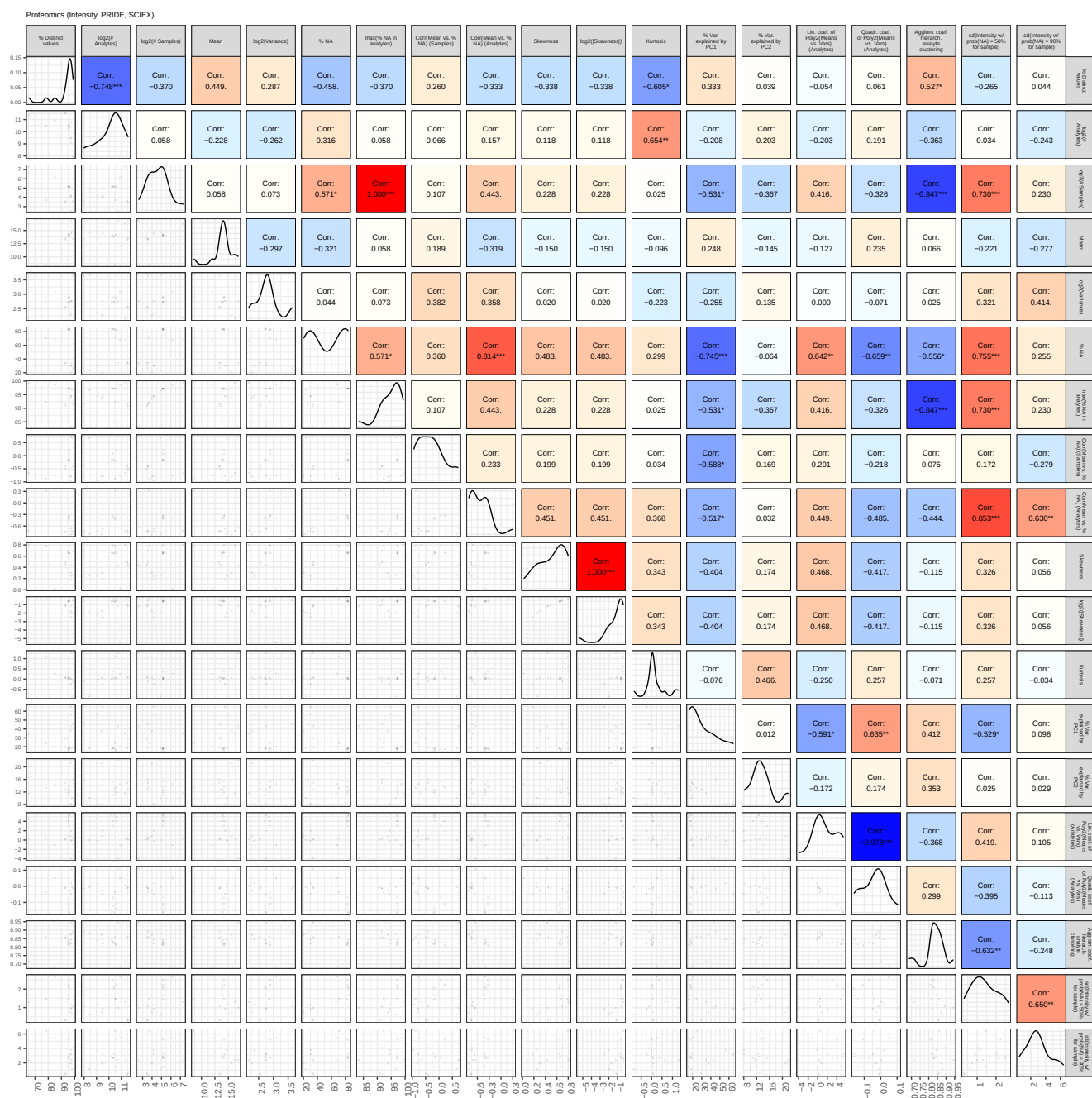


Figure S33. Spearman correlation plot of selected data characteristics over Proteomics (Intensity, PRIDE, SCIEX) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

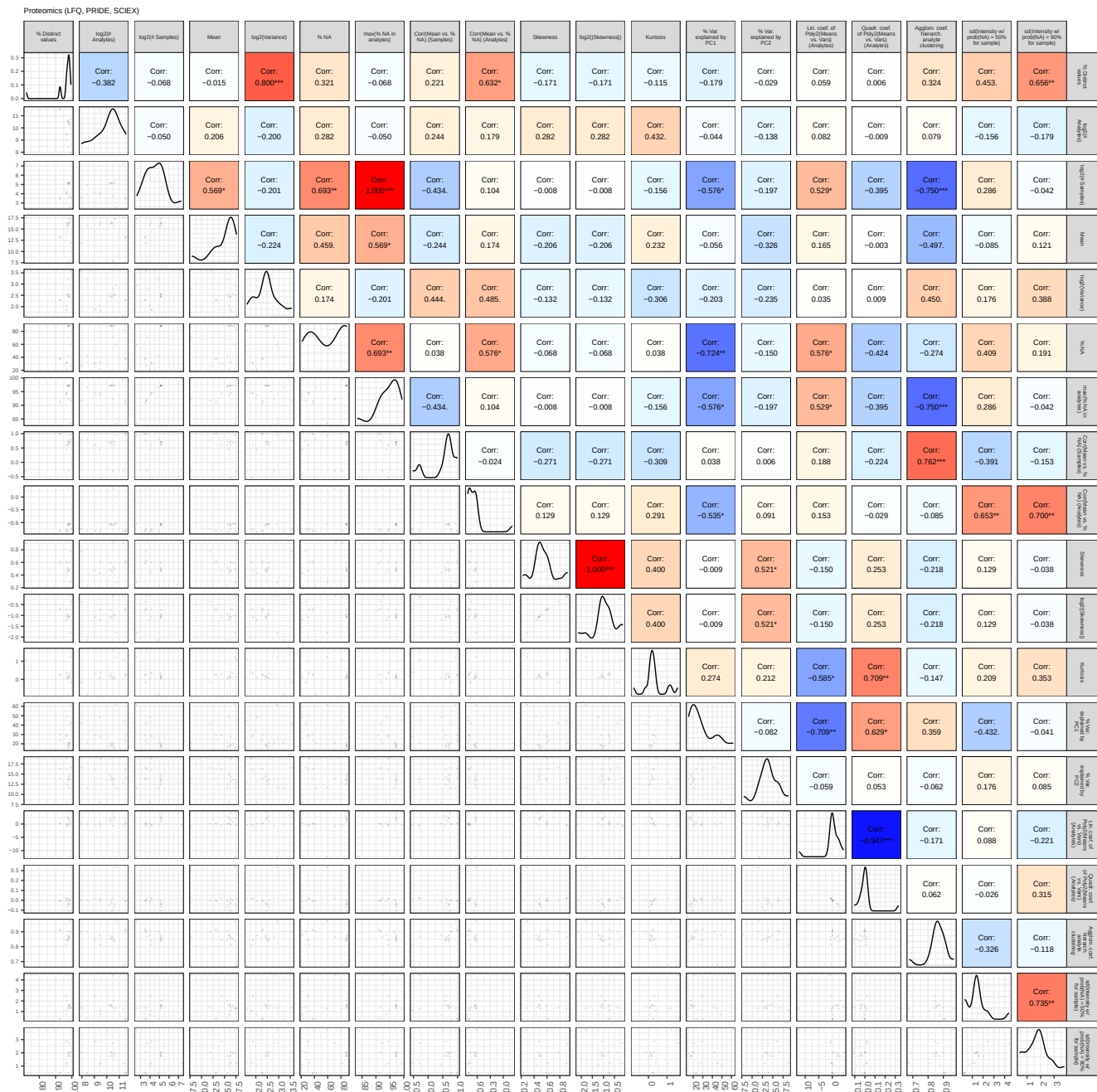
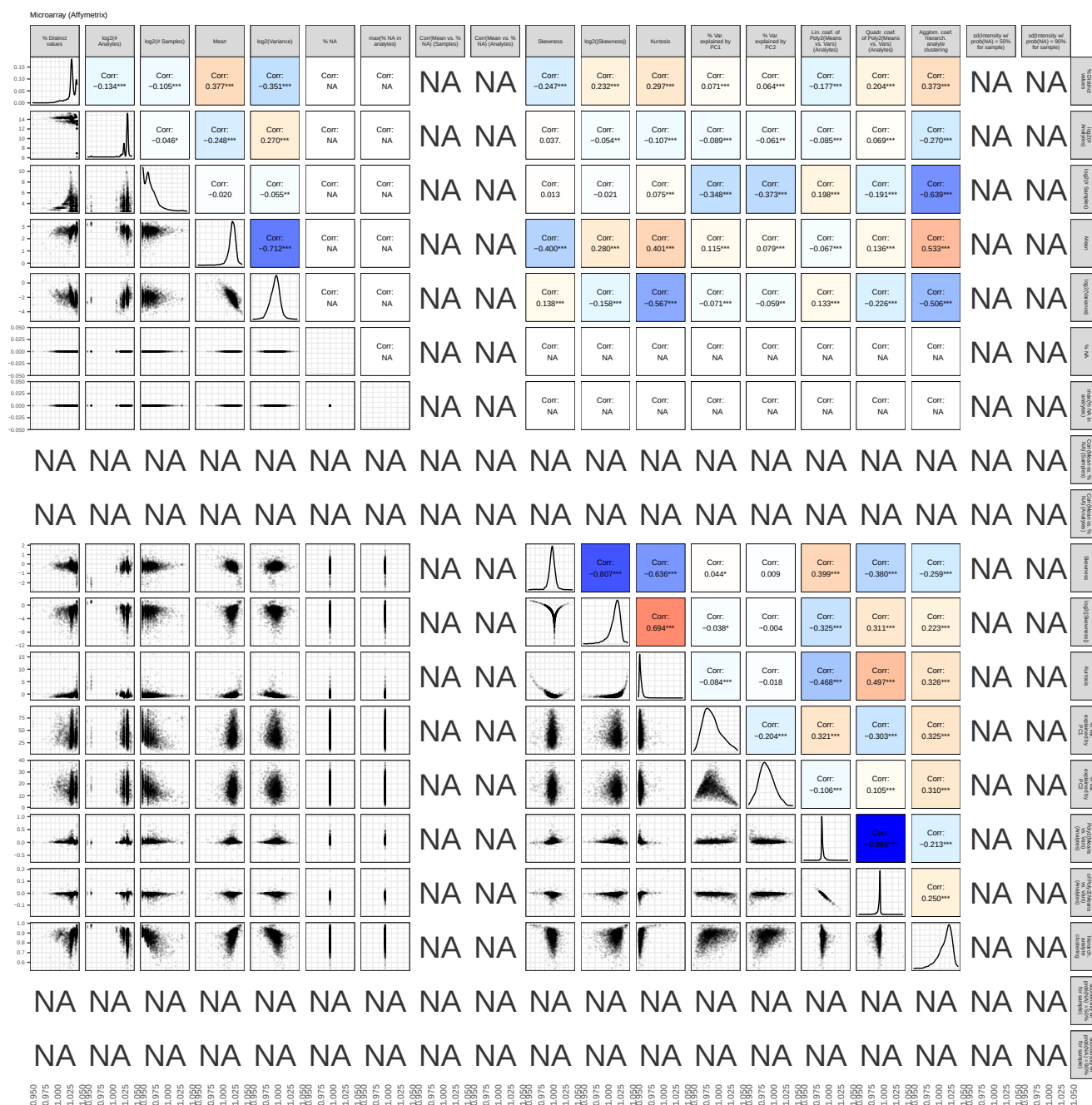
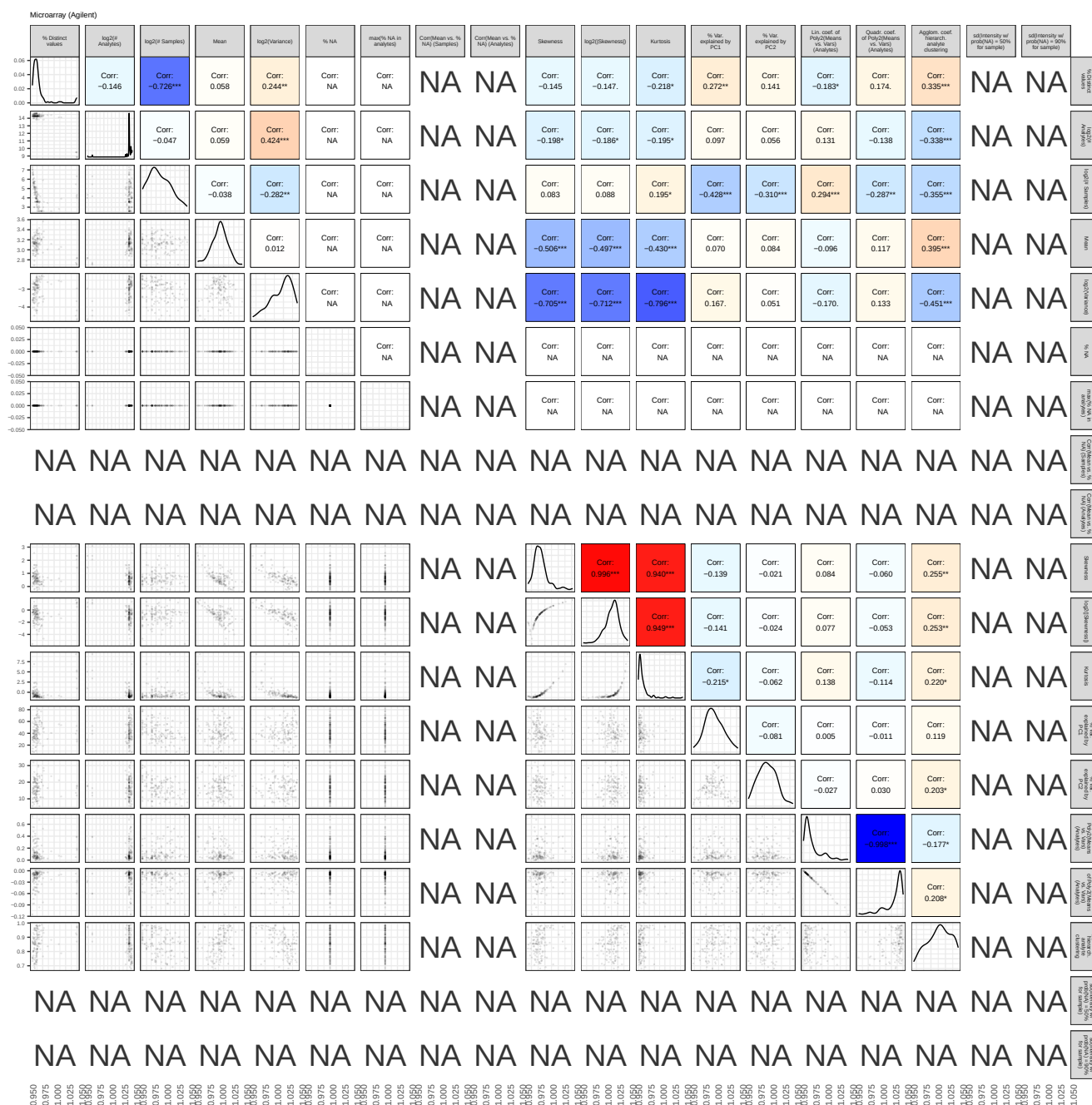


Figure S36. Spearman correlation plot of selected data characteristics over Proteomics (LFQ, PRIDE, SCIEX) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.





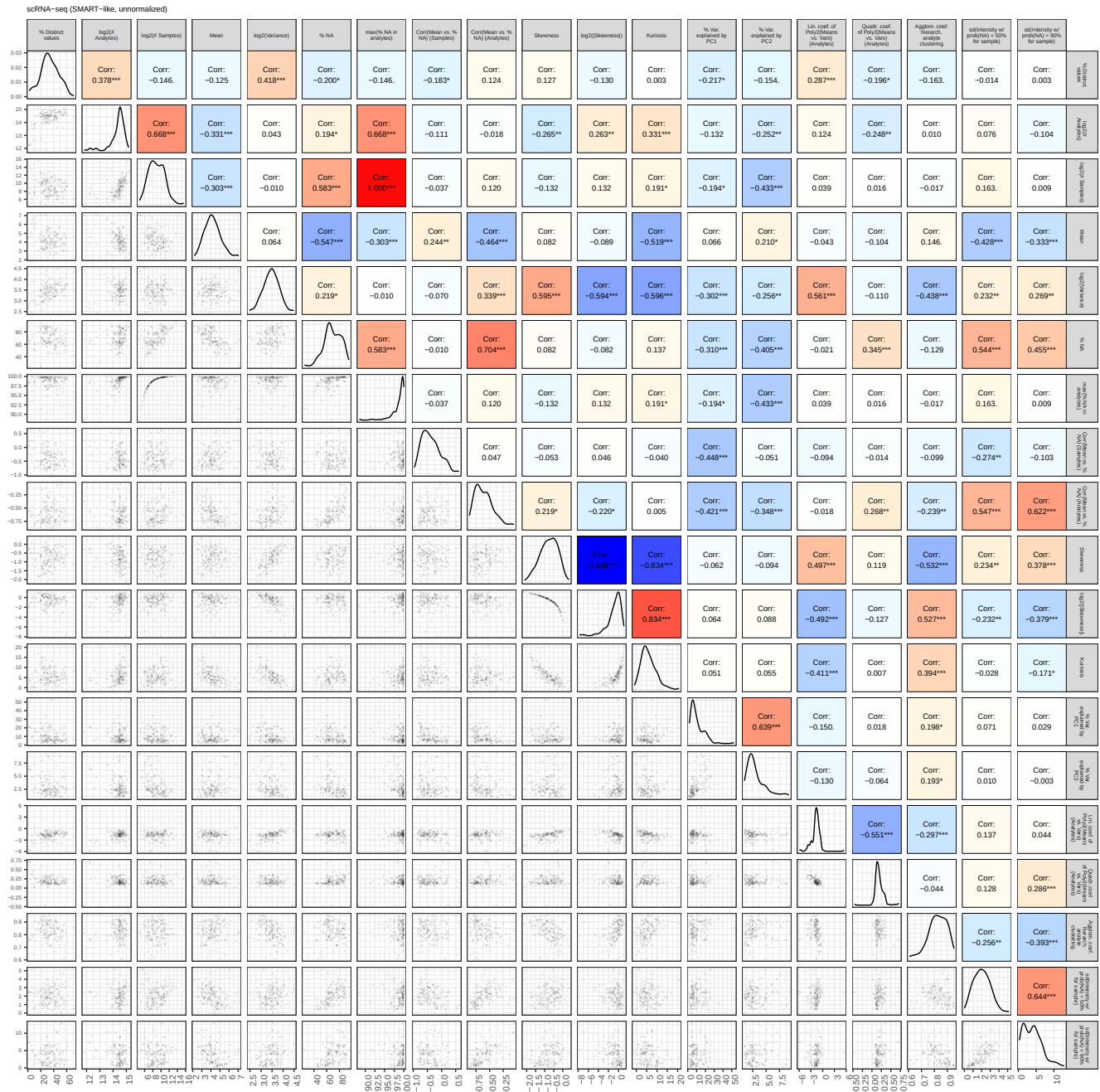
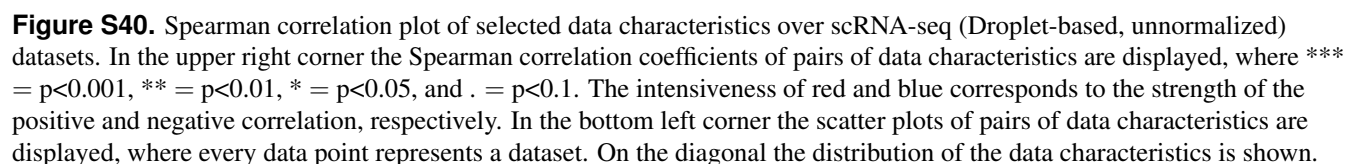


Figure S39. Spearman correlation plot of selected data characteristics over scRNA-seq (SMART-like, unnormalized) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.



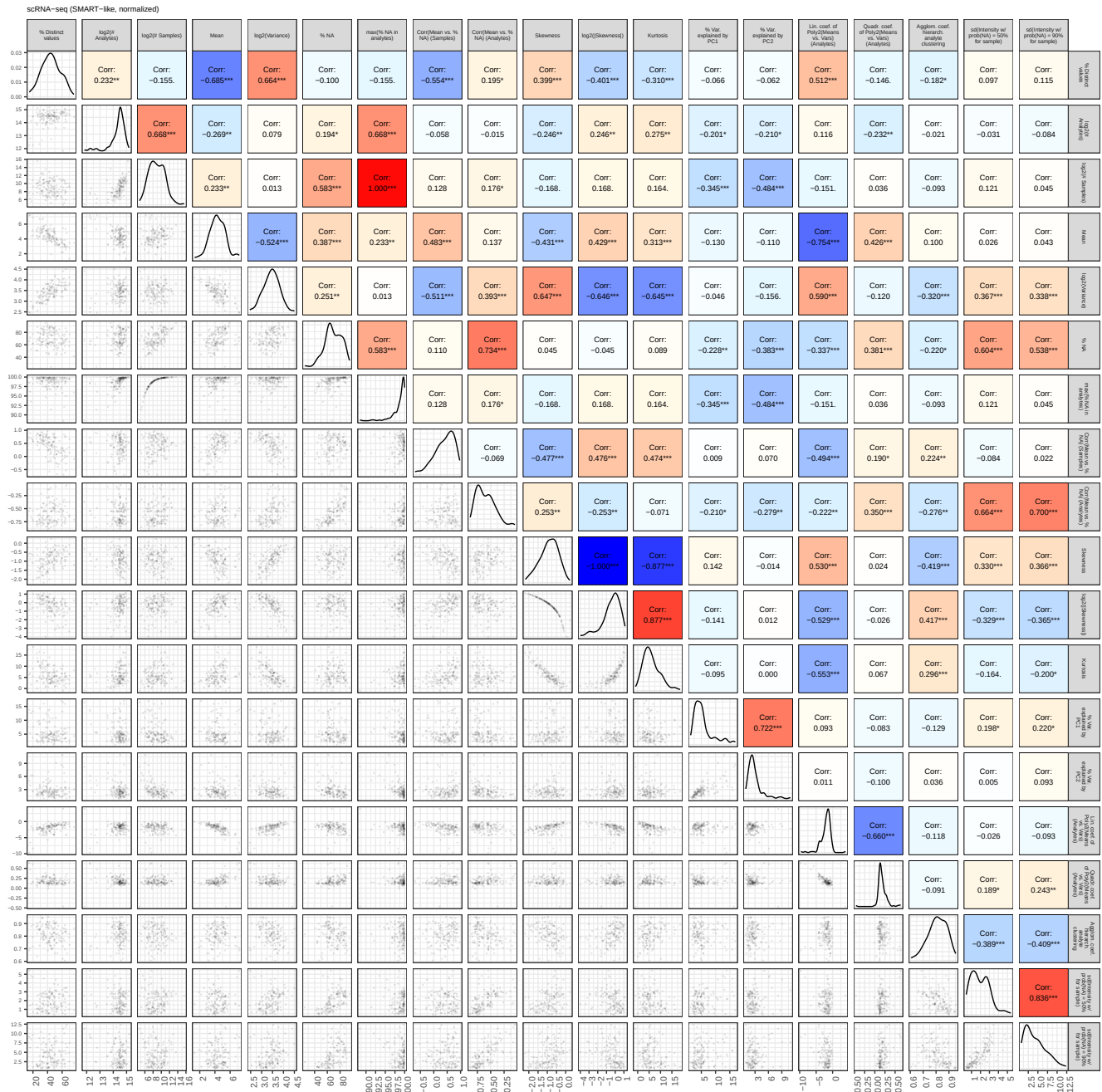


Figure S41. Spearman correlation plot of selected data characteristics over scRNA-seq (SMART-like, normalized) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

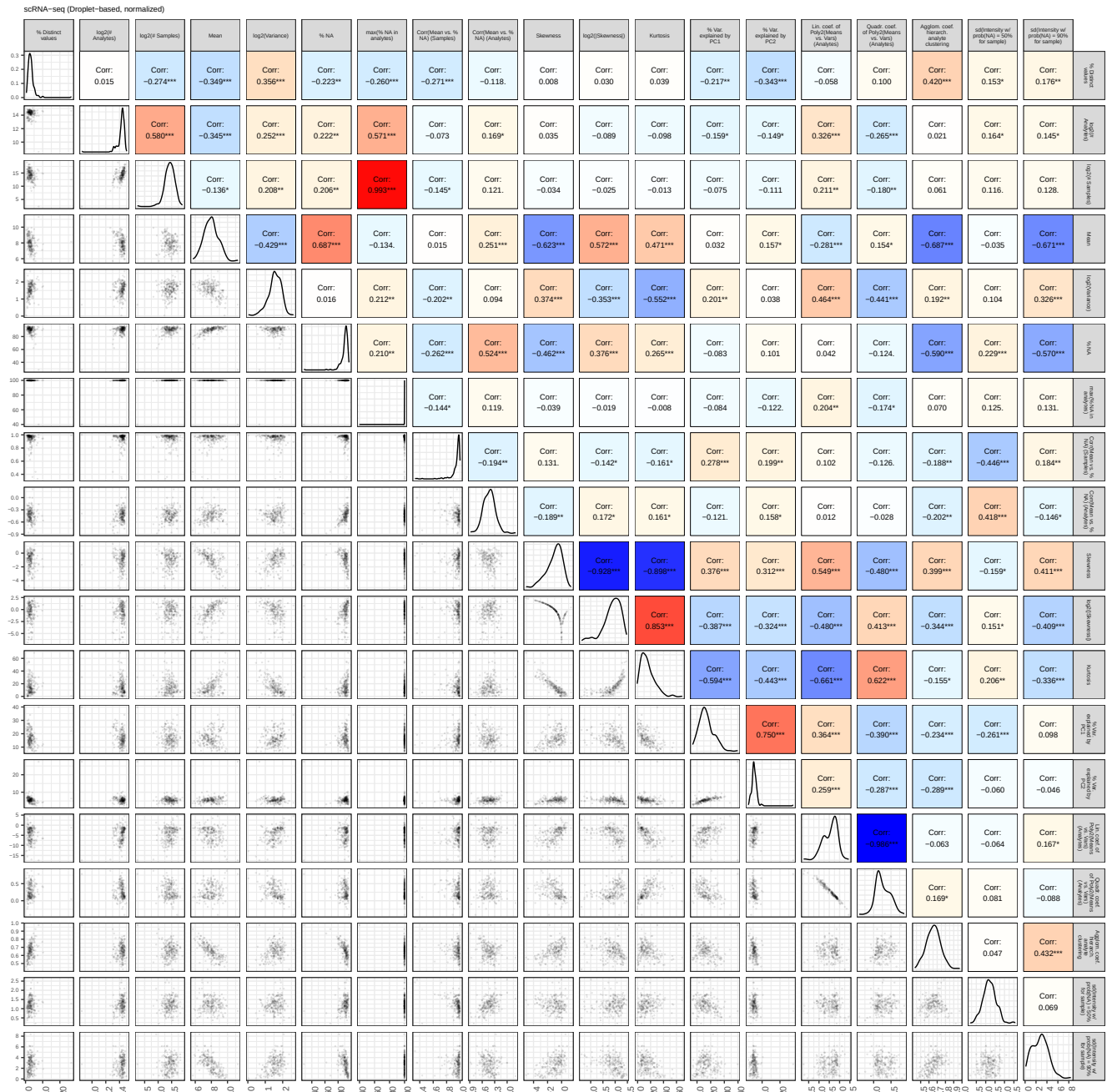


Figure S42. Spearman correlation plot of selected data characteristics over scRNA-seq (Droplet-based, normalized) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

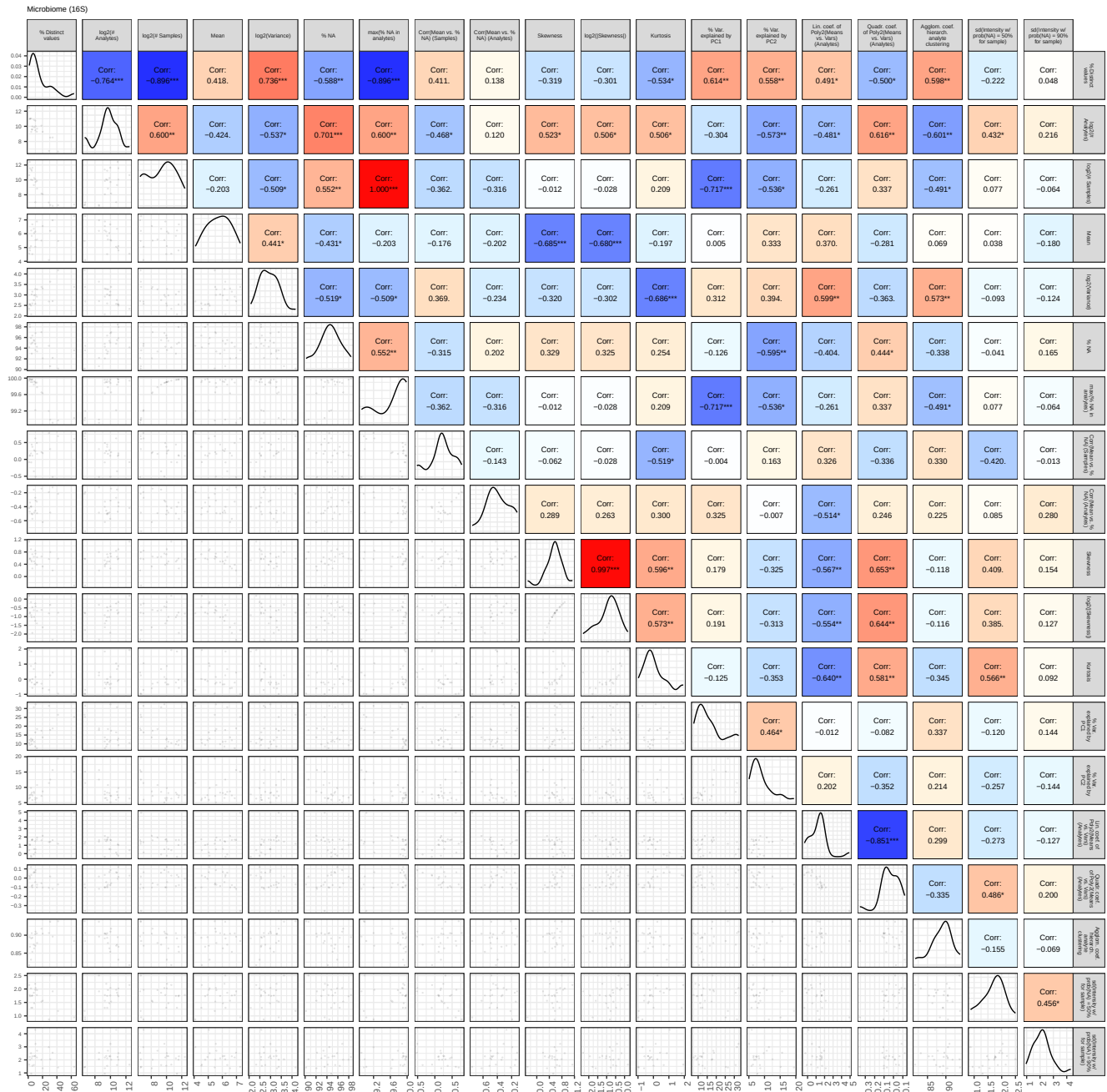


Figure S43. Spearman correlation plot of selected data characteristics over Microbiome (16S) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.

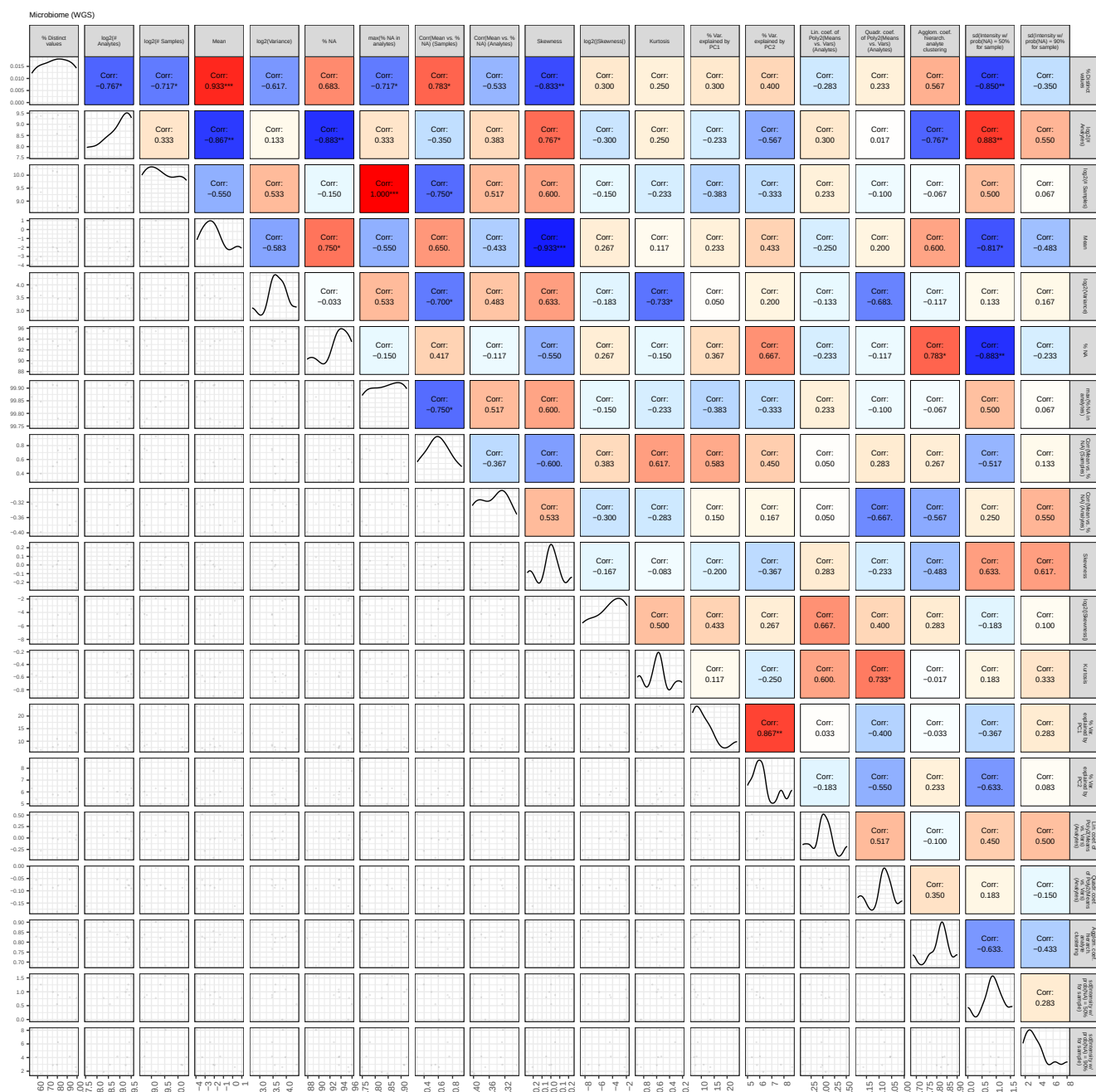


Figure S44. Spearman correlation plot of selected data characteristics over Microbiome (WGS) datasets. In the upper right corner the Spearman correlation coefficients of pairs of data characteristics are displayed, where *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, and . = $p < 0.1$. The intensiveness of red and blue corresponds to the strength of the positive and negative correlation, respectively. In the bottom left corner the scatter plots of pairs of data characteristics are displayed, where every data point represents a dataset. On the diagonal the distribution of the data characteristics is shown.