SUPPORTING INFORMATION

Peak Alignment: Peaks detected in the mean spectrum are used as landmarks to conduct peak alignment across the samples. For every detected peak in the mean spectrum, e.g. a peak with ppm=X_i, we search for peaks located in the range (X_i-delta, X_i+delta) with a predefined value delta in each spectrum. For example, if a peak is found in that neighborhood for sample j, the height of that peak is then assigned to be associated with X_i and sample j. If more than two peaks are identified in the same neighborhood for one spectrum, the highest one will be obtained. In the case where no peaks are found in that neighborhood, a missing value will be imposed. After the peak alignment step, all the peak locations in the mean spectrum are screened one by one. For example, for a specific ppm X_i, if there are more than 85% non-missing values observed across all the samples, then that peak will be saved and all the missing values associated with it will be imputed by the intensity at ppm=X_i from the corresponding samples. Otherwise, peak ppm=X_i is removed together with all the associated peaks. All the detected peaks are subject to visual examination.

Table S1. Latin square design of spike-in dataset. The spike-in dataset consists of 6 types of mixtures spiked with 6 known compounds at various concentrations. The 6 mixtures undergo further dilutions to generate18 samples.

Metabolite name	Mixture1 (uM)	Mixture2 (uM)	Mixture3 (uM)	Mixture4 (uM)	Mixture5 (uM)	Mixture6 (uM)
succinate	200	6400	3200	1600	800	400
fumarate	400	200	6400	3200	1600	800
2-ketoglutarate	800	400	200	6400	3200	1600
4-hydroxybenzoate	1600	800	400	200	6400	3200
4-hydroxyphenylacetate	3200	1600	800	400	200	6400
nicotinate	6400	3200	1600	800	400	200

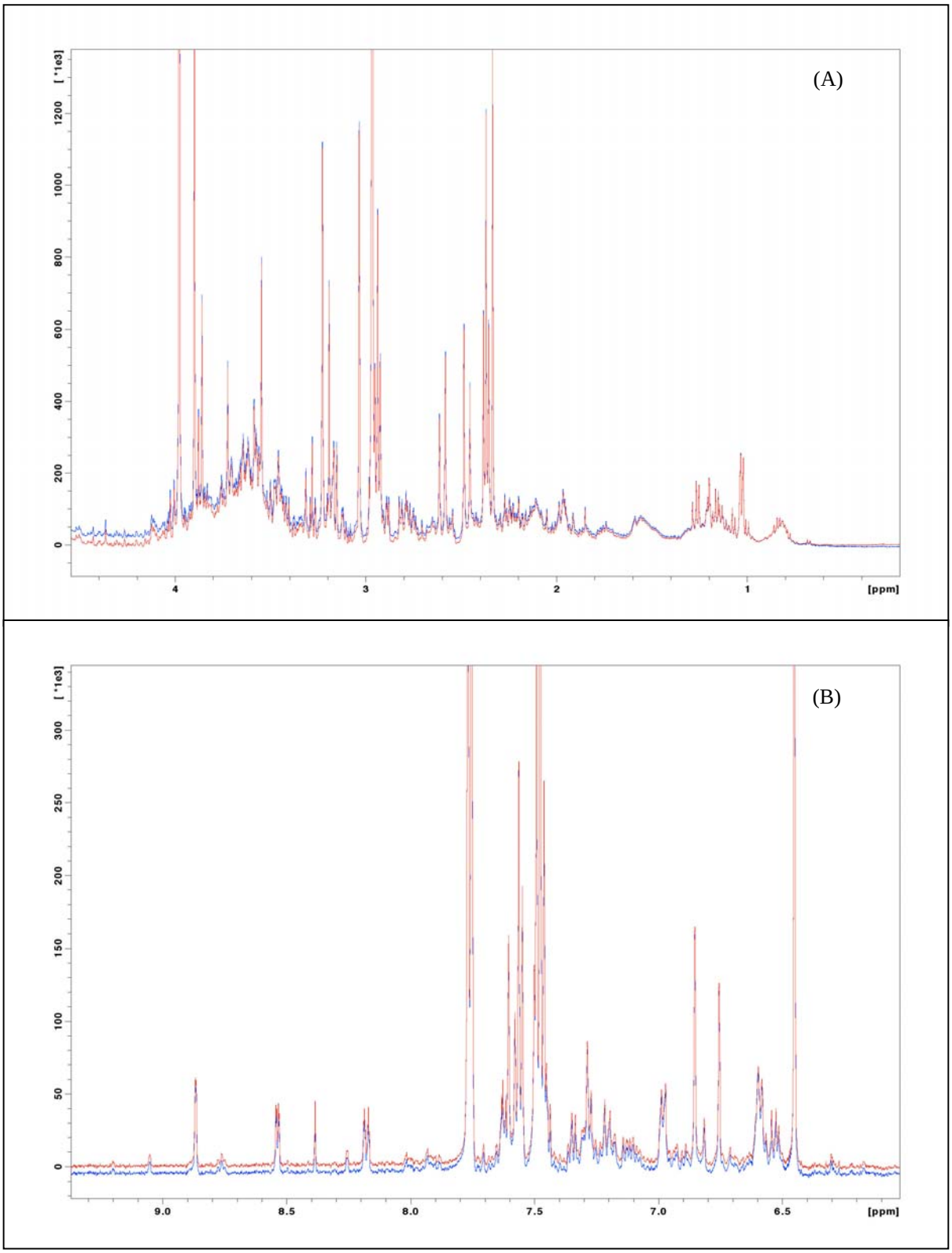

Figure S1. Typical ^{1}H NMR spectra (**A**: aliphatic region, **B**: aromatic region) of a spiked human urine. Red spectra are after baseline correction. Blue spectra are prior to baseline correction.

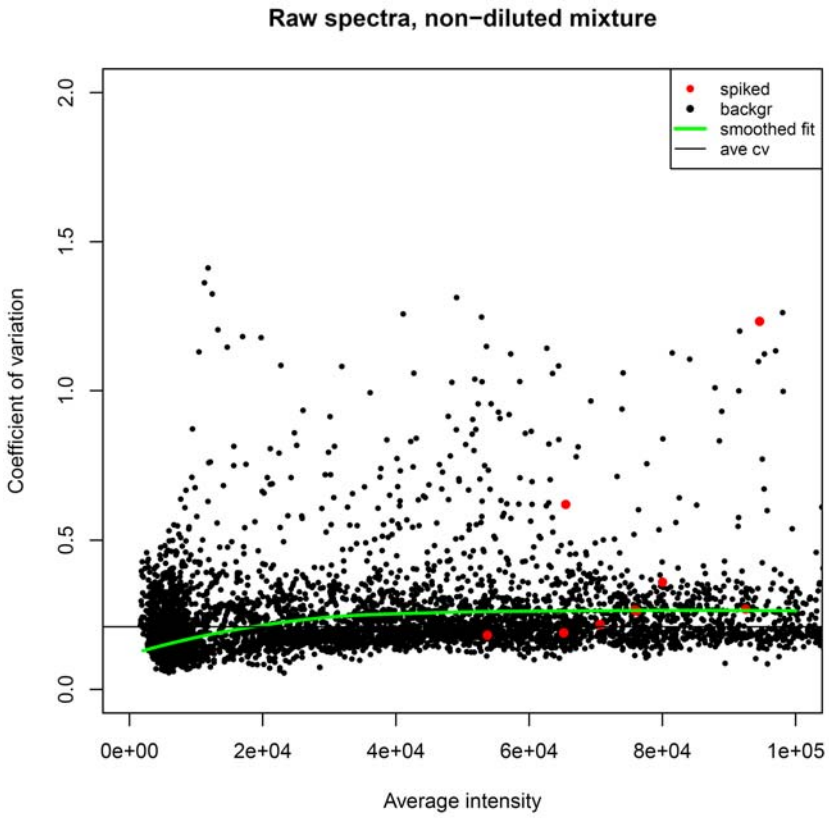

Figure S2. Relationship between average peak intensities (x axis) and peak-specific coefficients of variation (y axis) in the raw baseline-corrected spectra. To avoid dilution artifacts, only the first (non-diluted) set of mixtures was used to produce the plot. The coefficient of variation is roughly constant for all signal intensities, indicating that the variation increases with the signal mean, and therefore a multiplicative model of the measurement error is appropriate. The results are similar for other dilutions.

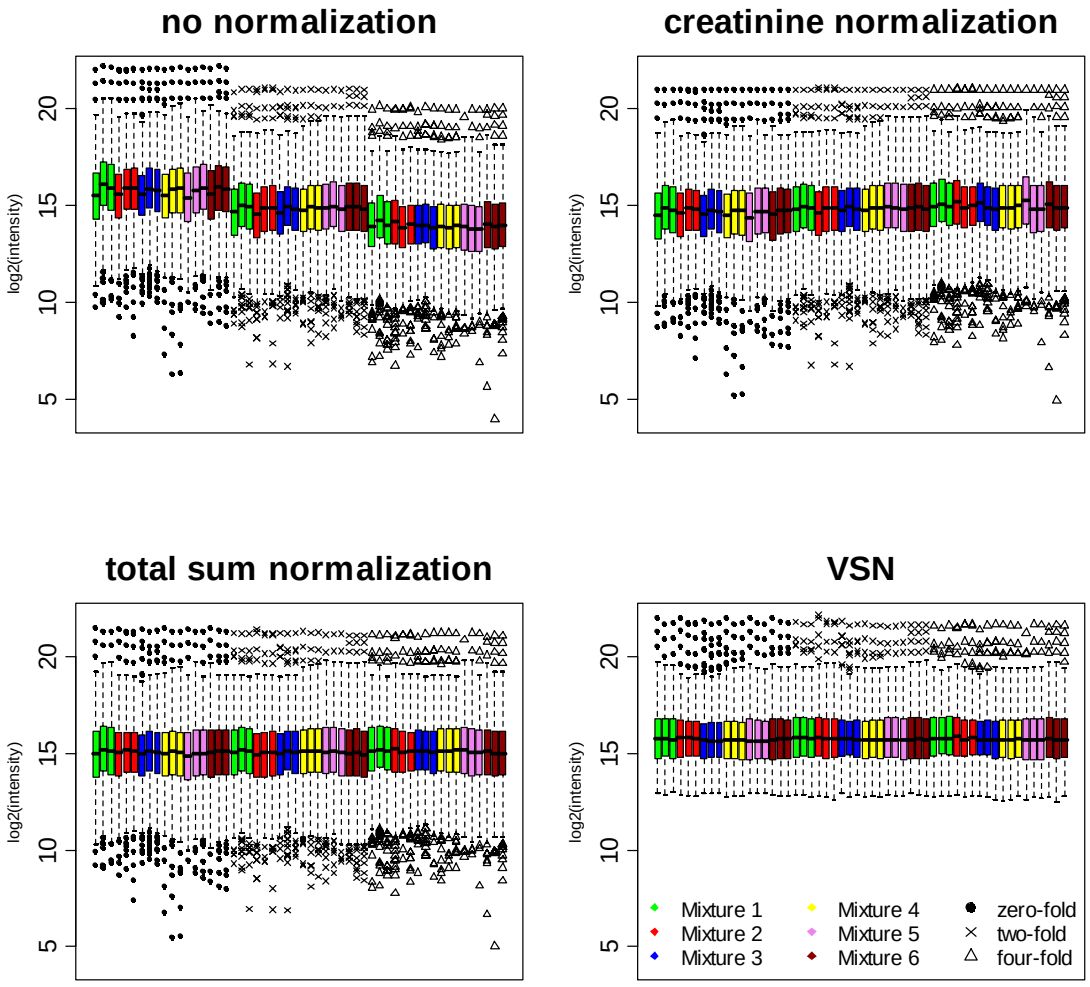

Figure S3. Boxplots of the normalized and log-transformed intensities of peaks detected in the 54 spectra for the spike-in dataset. Horizontal lines are quantiles of the distributions, and each box contains 50% of the peak intensities of the peaks. Individual points outside of the boxes indicate outlying peaks. The spike-in mixtures are color-coded, and dilution levels are differentiated according to the shapes of the outlying peaks. Features from spiked compounds comprise approximately 8% of the total peak area.

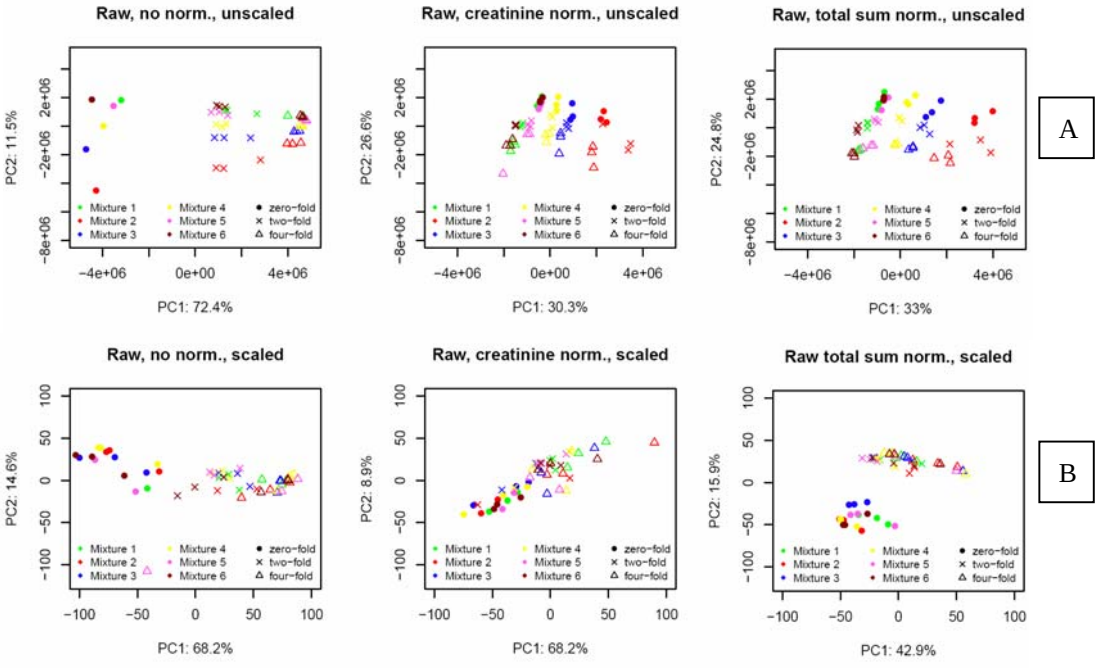

Figure S4. PCA score plots for the 54 spectra in the spike-in dataset, which utilize raw baseline-corrected spectra and (A) no scaling, and (B) auto scaling. X and Y axes are the first and the second principal components, respectively. Colors indicate the six mixture types, and shapes indicate the three dilution levels. PCA applied to the unscaled raw spectra produces an overall inferior performance as compared to peak picked and log transformed data in Figure 2. The procedure fails distinguishing between different mixture types when scaling is used. Since VSN incorporates the log as part of its normalization procedure, there is no alternative to VSN on the raw scale.

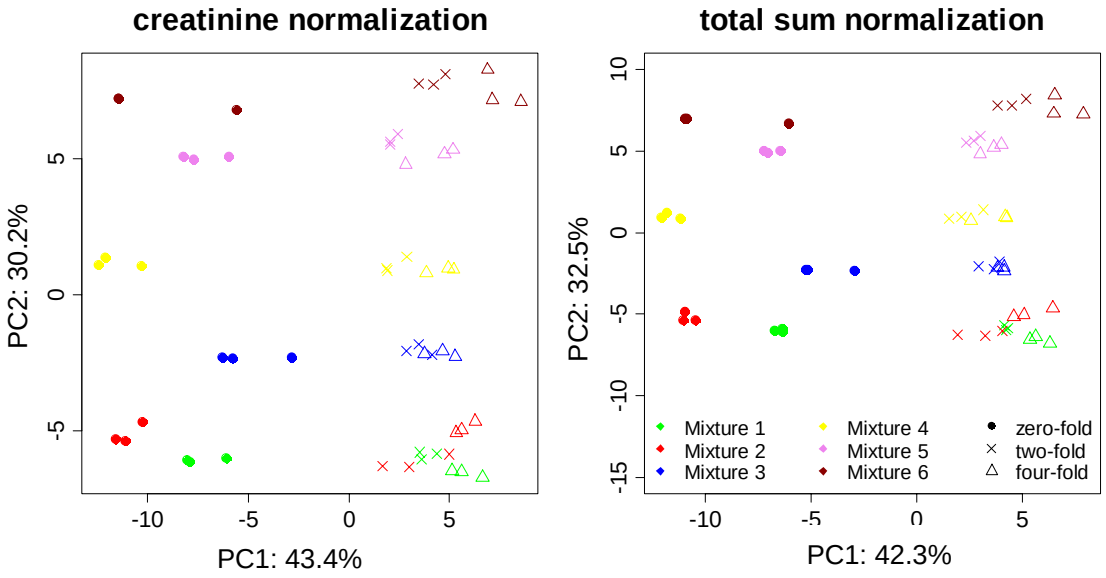

Figure S5. PCA score plots of the 54 spectra in the spike-in dataset using peak picked data on the original scale. X and Y axes are the first and the second principle components, respectively. Colors indicate the six mixture types, and shapes indicate the three dilution levels. Since VSN incorporates the log as part of its normalization procedure, there is no equivalent to VSN on the original scale.

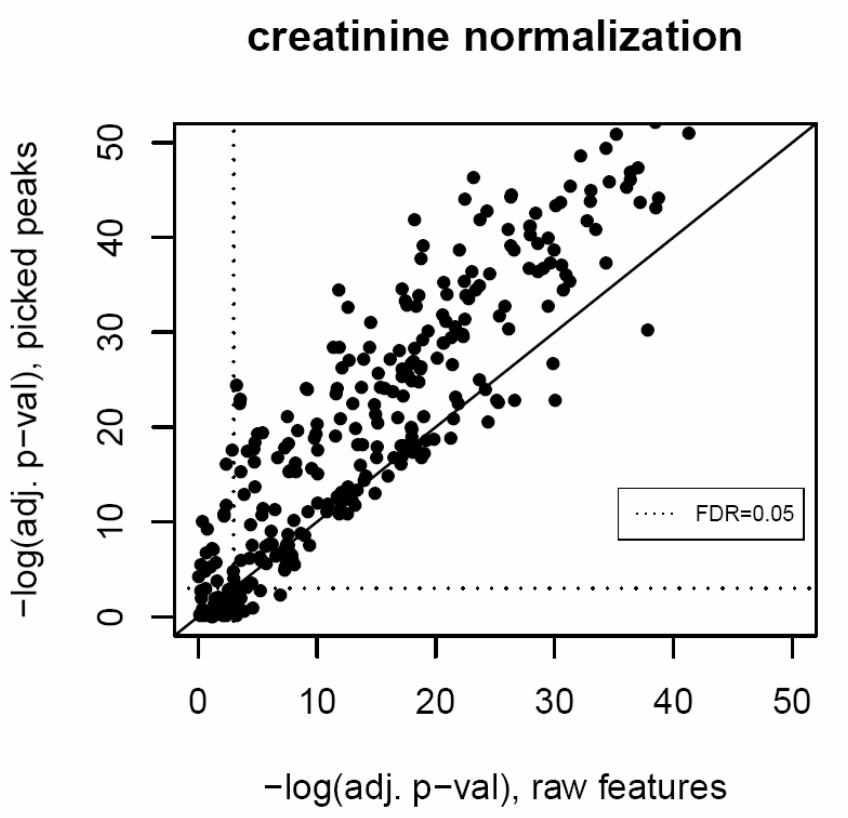

Figure S6. P-values of all pairwise comparisons of 6 mixtures, separately for each spiked metabolite, after log transform and creatinine normalization. X axis: p-values of tests in the raw spectra; y axis: p-values of tests in the peak picked spectra. The p-values of each pairwise comparison are adjusted for multiple testing using the Benjamini-Hochberg procedure, and the dotted lines correspond to the FDR cutoff of 0.05 for each of the spectra. The p-values are shown on the –log scale, i.e. larger values correspond to stronger evidence of differential abundance. The solid line represents a perfect agreement.

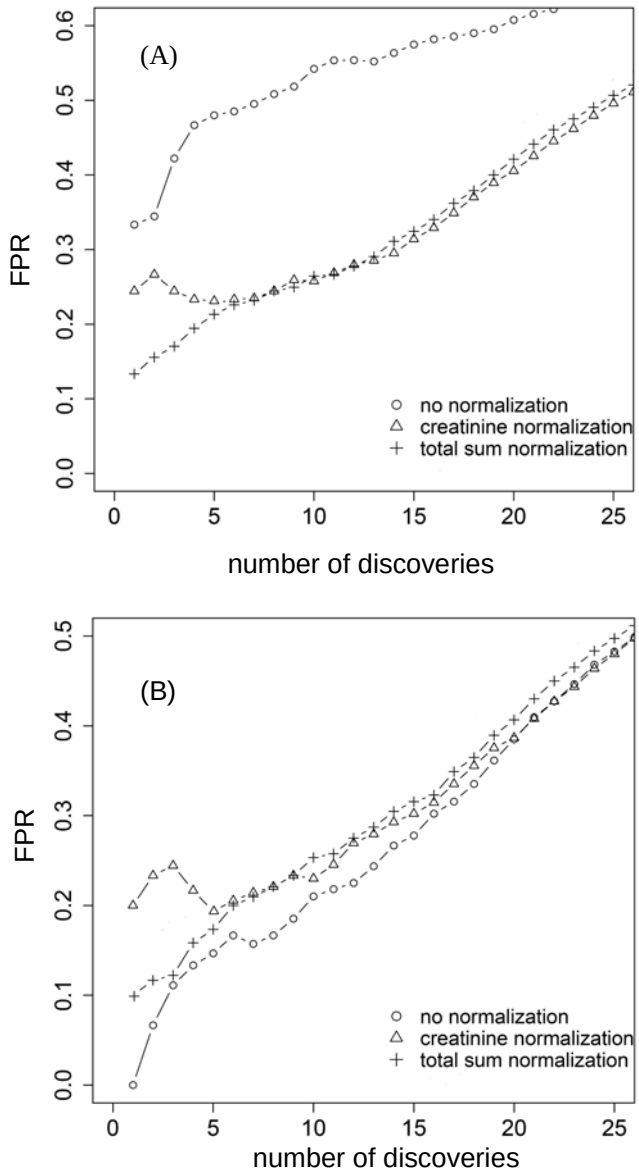

Figure S7. False positive rate (FPR) versus number of detected differentially abundant features in the spike-in experiment, after peak picking, but without log transform. X-axis: number of features discovered. Y-axis: false discovery rate calculated as dividing the number of false discoveries by the given number of discoveries; averaged FPR from all possible comparisons are used in this plot. In the legend, shapes and color indicate specific normalization methods and different aspects of testing, respectively. (A) The comparisons were made only for sample pairs with diverse dilution levels (90 comparisons) (B) The comparisons were made only for sample pairs with the same dilution levels (45 comparisons) Since VSN incorporates the log as part of its normalization procedure, there is no alternative to VSN on the raw scale.

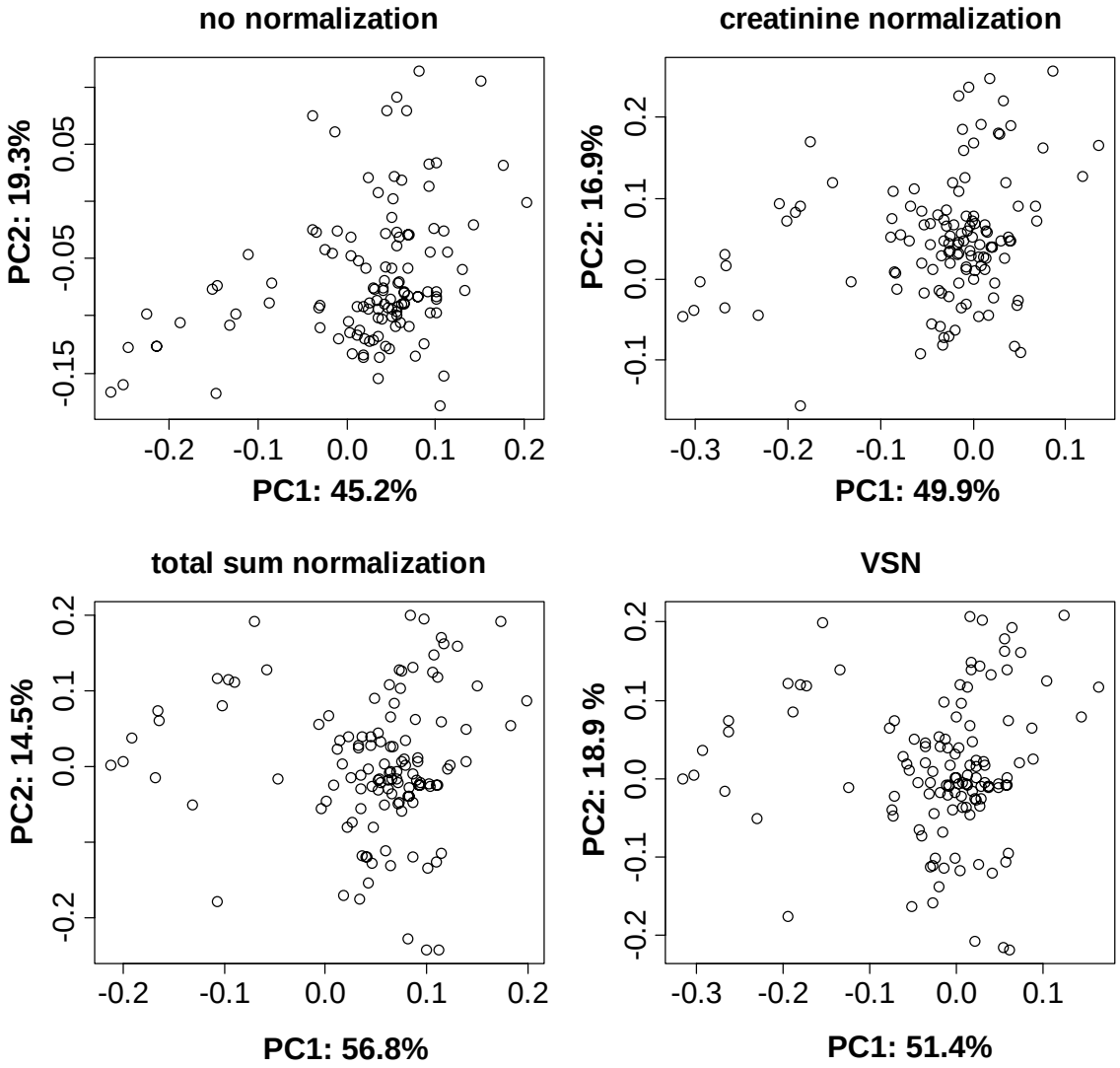

Figure S8. PCA loading plots for the diabetes dataset. X and Y axes are the first and the second principle components, respectively.

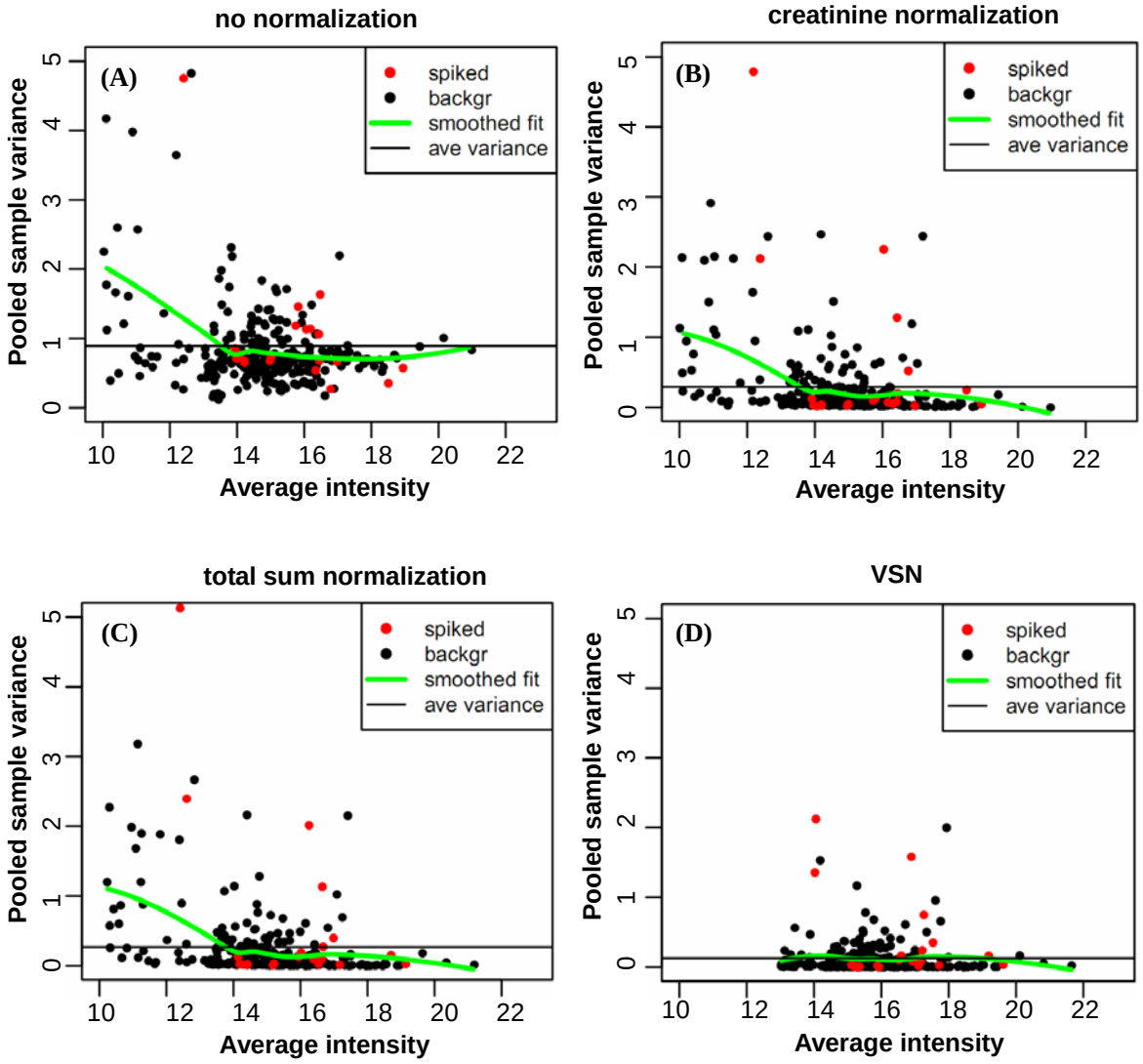

Figure S9. Relationship between average peak intensity and pooled sample variance after the (generalized-) log transform. All three normalization procedures reduce the average variance of the peak intensities. However VSN is the only procedure enforcing a roughly equal variance for all means.

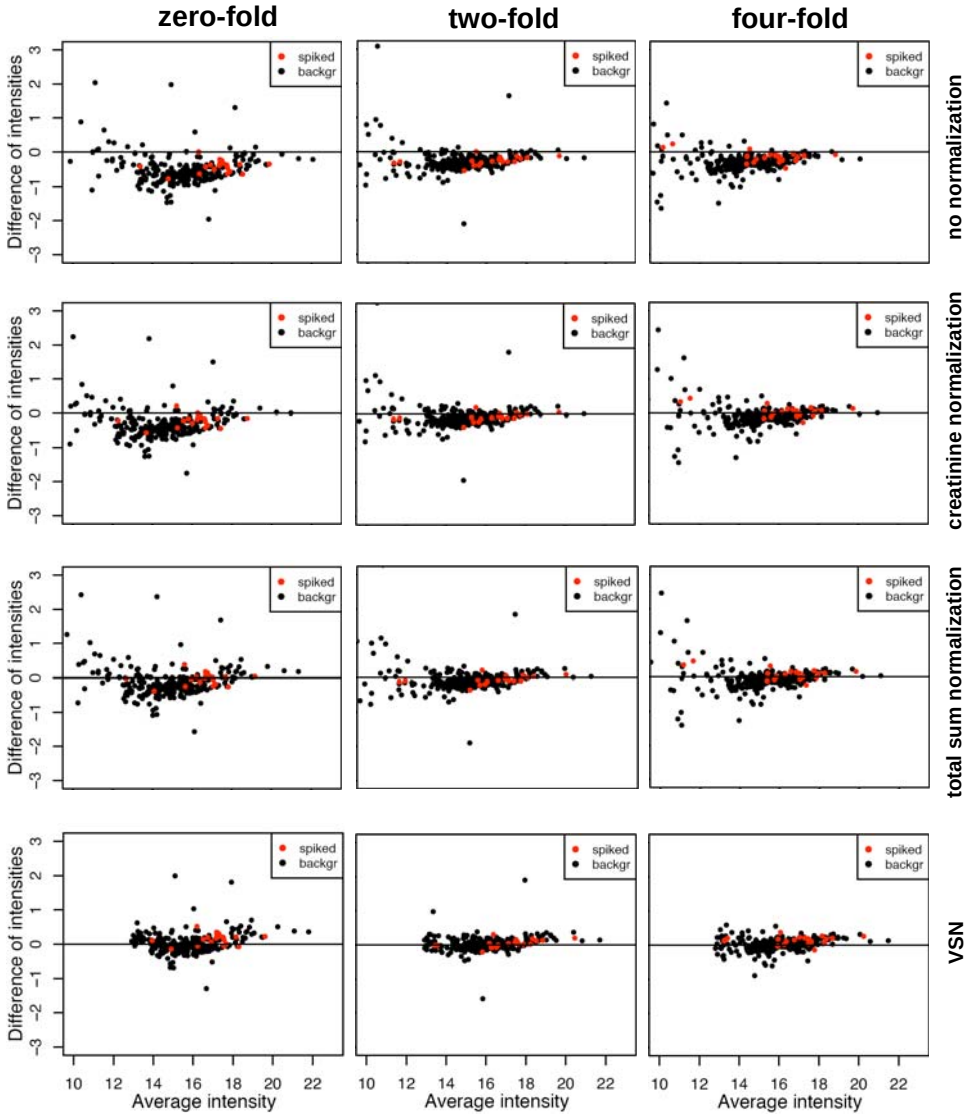

Figure S10. Relationship between the average and the difference of peak intensities of two replicates of a same mixture after the (generalized-) log transform. All the procedures except VSN show that the differences between replicates are larger for peaks with low intensities. The dependence of the variation on the mean is more pronounced for increased dilutions.

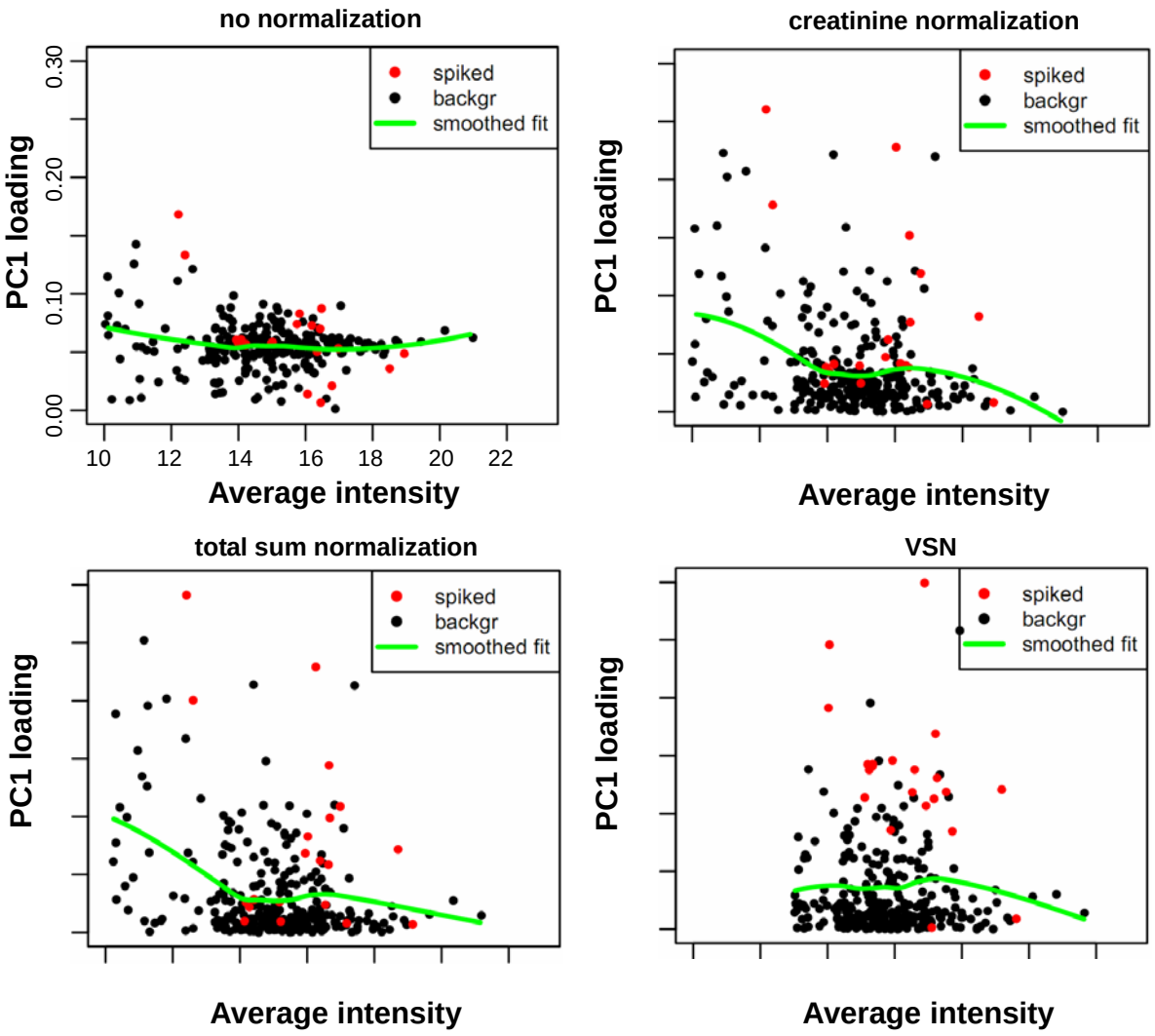

Figure S11. Relationship between the absolute values of the peak loadings obtained as part of the output of the PCA, and the average intensity. Loadings of low-abundant background peaks are unduly inflated by all procedures except VSN.

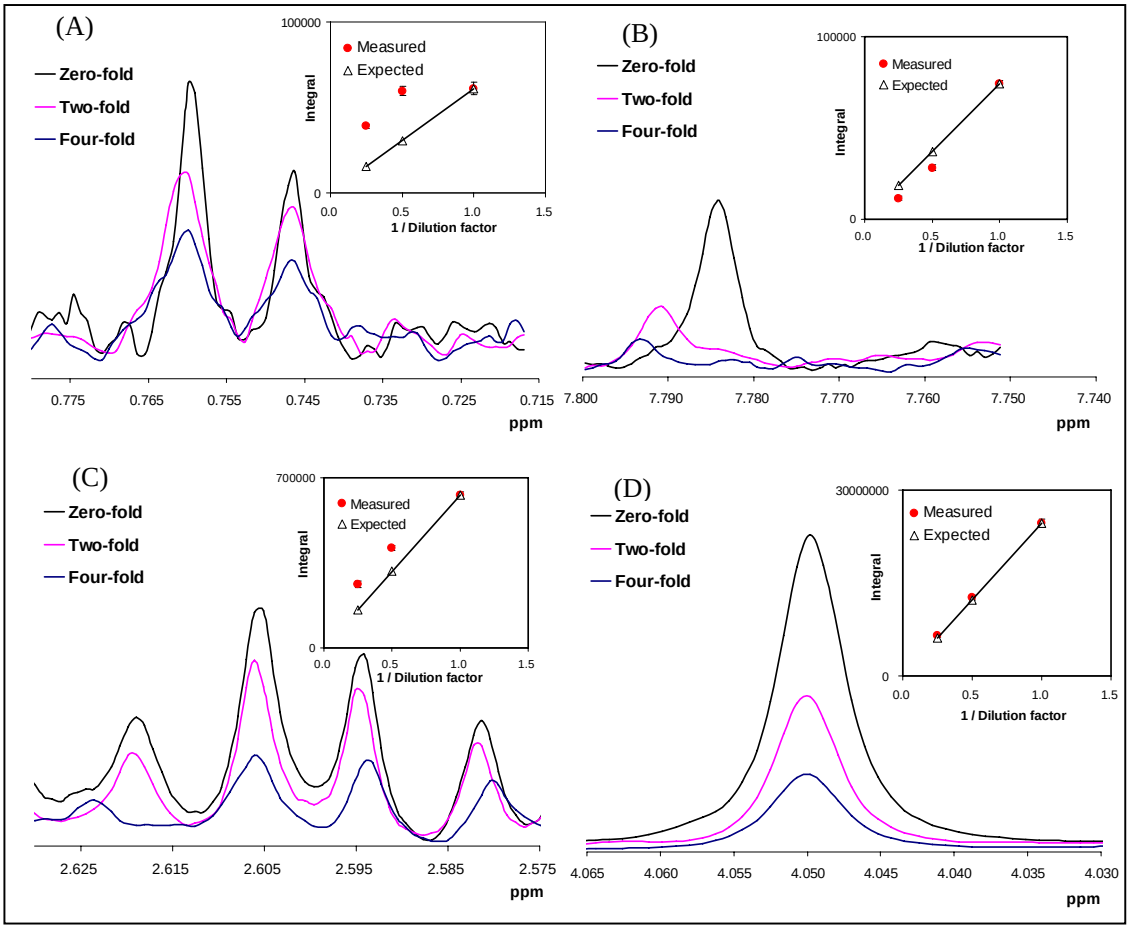

Figure S12. Raw baseline-corrected spectral plots showing three representative non-linearly behaving peak integrals (**A,B,C**) and the creatinine peak (**D**) for comparison using one mixture at three dilution levels. Insets are plots of the measured individual peak integral versus the inverse dilution factor and the expected integral plotted for reference. The expected values were calculated from the integral at the zero-fold dilution level and the dilution factor. Peak integrals shown in plots A, B and C for the three dilution conditions are clearly not strictly inversely proportional to dilution factors. These non-linearly behaved peaks are different from most highly concentrated metabolites, represented by the case shown in plot D, in which the peak integral does change in a direct inverse proportionality with the dilution factor.