Supplementary Material

SUPPLEMENTARY MATERIAL M1. Description of the multivariate statistical methods that were applied in the analyses of the kidney rejection data, including parameter tuning steps and splitting of the data into training and testing sets for the classification analysis.

a. Principal Component Analysis

PCA (Jolliffe, 2002) is a classical dimension reduction and feature extraction tool in exploratory analysis, and has been applied in a wide range of fields. PCA has often been used as a pre-processing step for subsequent analysis. PCA projects the data into a new space spanned by the *principal components* (pc), which are uncorrelated and orthogonal. The principal components are linear combinations of the original variables. Each *loading vector* associated to each pc, indicate the contribution of each gene or protein to each pc. The pcs and the loading vectors are defined so that the variance of each pc is maximised. One popular criterion to choose the number of pcs is based on the cumulative percentage of explained variance, as the pcs are ordered by decreasing explained variance. For the case of high dimensionality, many alternative ad hoc stopping rules have been proposed without, however, leading to a consensus (see (Cangelosi & Goriely, 2007) for a thorough review). In our case we applied the elbow criterion to the PCA scree plot.

b. Independent Principal Component Analysis

PCA proved unsuccessful to highlight a ‘natural’ separation between the AR and the NR samples. Another exploratory method, Independent Principal Component Analysis (IPCA,(Yao, Coquery, & Lê Cao, 2012)) was envisaged as an alternative to PCA. IPCA is a variant of Independent Component Analysis (Comon, 1994; Hyvarinen et al., 2002) which became a popular multivariate analysis tool in molecular biology. In contrast to PCA, which assumes orthogonal (i.e., uncorrelated) components, ICA requires the components to be statistically independent, which means that the values of one component provide no information about the values of other components. When using ICA, we make the assumption that the variables that we measure depend on some biological or environmental factors that are assumed to be statistically independent. These factors are the independent components we are searching for. When applying ICA, we assume that the observed data have been determined by some unknown fundamental factors, which are independent of each other. It has been shown that due to its independence condition, ICA might be more suitable to some metabolomic studies and outperform PCA (Scholz, Gatzek, Sterling, Fiehn, & Selbig, 2004; Wienkoop et al., 2008). IPCA was recently proposed to further denoise the independent components and remove as many irrelevant variables as possible (see (Yao et al., 2012) for more details).

c. Partial Least Squares Discriminant Analysis and its sparse variant

In this study, we were particularly interested in Partial Least Squares Discriminant Analysis (PLS-DA,(Barker & Rayens, 2003)), a special case of Partial Least Squares for a supervised framework. Similar to LDA, this approach seeks optimal components, which are linear combinations of variables (e.g. genes or proteins) that best discriminate the different conditions (e.g. rejection status).

In order to obtain more interpretable results regarding the role of the genes or proteins in PLS-DA, a sparse approach has recently been developed (sPLS-DA, (Lê Cao et al., 2011)) to select the best predictor variables within the model. The result is a parsimonious model built on a small subset of variables. In sPLS-DA two parameters need to be tuned: the total number of components (also called dimensions) onto which the data will be projected, and the number of variables to select on each component. It has been shown that often, G-1, where G is the number of classes or conditions, is the optimal number of components to use (Lê Cao et al., 2011). In the acute rejection case study with two classes (AR and NR) this suggests to focus on one component.

Parameter tuning. In this study, we divided the original data set into a training and a testing set from the initial cohort of 40 patients, 13 AR and their time-matched NR samples were randomly selected to constitute the training set. Stratified sampling was performed by time of rejection (early vs. late) to reduce the impact of this parameter as a potential confounder in light of the earlier exploratory results with PCA and IPCA. Two sPLS-DA classifiers were built using the genomic and proteomic training sets, and tested on the remaining 14 samples (7 AR and 7 time-matched NR).

The number of variables for a parsimonious model was selected for the first (smallest) model where the average classification error (CE) rate fell within one standard error of the minimum average CE rate based on 100 times 5-fold cross-validation runs on the training data. Figure S4 indicates that for the genomic data, 90 probe-sets were selected (a) while 21 protein groups were selected for the proteomic data (b).

d. sparse Partial Least Squares

Partial Least Square regression (S. Wold, Sjostrom, & Lennart, 2001) is a generalization to PLS-DA in the case where two data sets measure the expression or abundance of two different types of features (here gene expression and protein abundance on the same samples) on the same samples. The difference with PLS-DA is that the information from both data set is integrated in an unsupervised manner (no prior biological knowledge on the patients’ rejection status is included in the model). PLS relates and integrates the two data sets by a linear multivariate model while also modeling the data structure. PLS is particularly useful for analysing noisy, collinear, even incomplete highly dimensional data, see (Boulesteix & Strimmer, 2007) for a review.

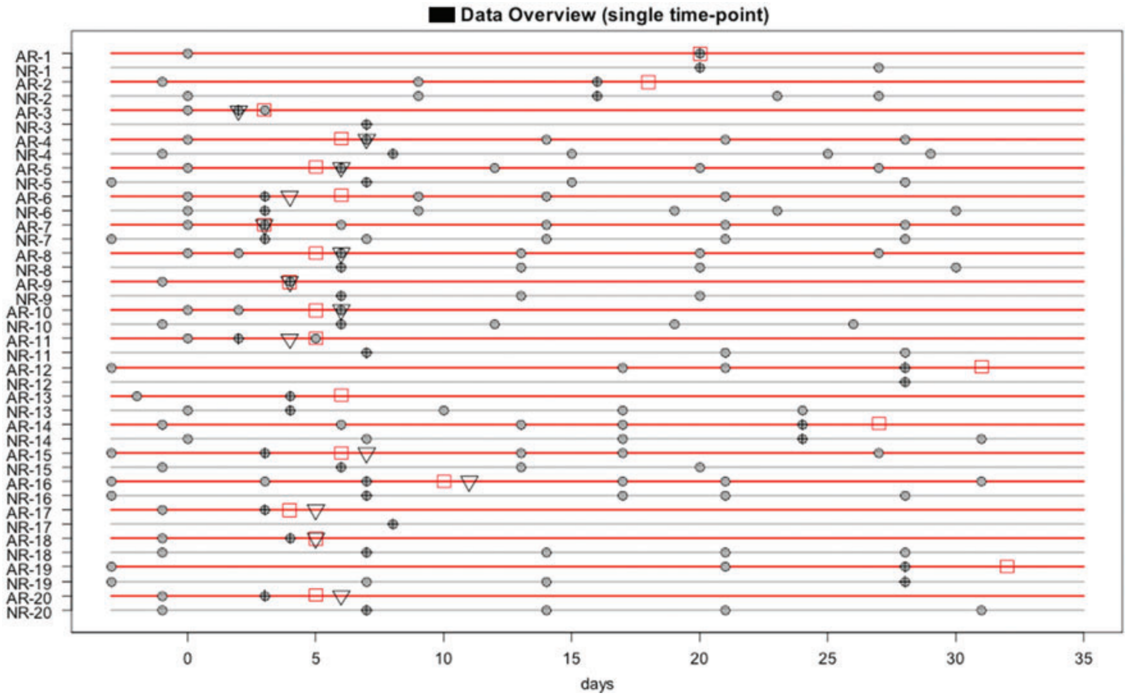
Similar to the multivariate approaches presented earlier, PLS performs successive decompositions of the two data sets into new variables, the PLS components, which should be fewer in number than the total number of measured features, orthogonal to each other within each data set, and estimated as linear combinations of the original variables from both data sets (the weight coefficient of each variable is indicated in the associated loading vectors). PLS relates both matrices by maximising the covariance between each pair of PLS components. PLS can be applied within a regression- or a canonical framework, the latter models a symmetric relationship between the two data sets, i.e. extracting the common information between the two platforms similar to the framework of another multivariate method called Canonical Correlation Analysis, CCA(Hotelling, 1936)). In this paper, we primarily focus on the canonical framework.

A 'sparse' approach was recently developed (sPLS, (Lê Cao, Martin, et al., 2009; Lê Cao et al., 2008)) to select relevant subsets of variables from both sets. When modeling a symmetric relationship between the two data sets, we are primarily interested in selecting the best subsets of correlated features, within each data set and across data sets (Lê Cao, Martin, et al., 2009). sPLS is an unsupervised approach; but, similar to supervised sPLS-DA, the number of components (dimensions) and the number features to select need to be tuned in each data set.

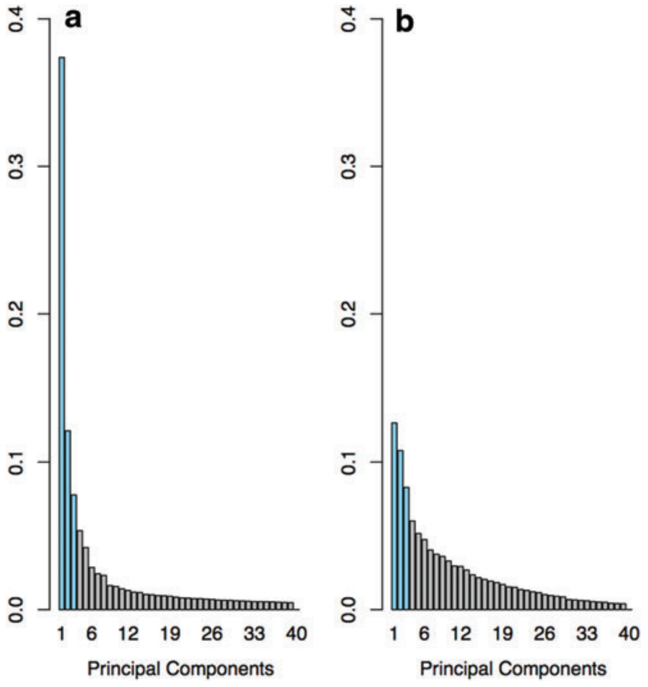
Parameter tuning. The sparse models were fitted with one component and a selection of 100 genes and 50 proteins on 10-folds 'training' data sets. We used a 'stability analysis', proposed by (Meinshausen & Bühlmann, 2010), which records the frequency of the same feature to be selected across several 'training' data sets where 1 fold out of 10 was removed to perturb the original data. This was performed 1,000 times to determine the most frequently selected genes and proteins. The final selections include the variables selected more than 70% of the time across the runs.

e. sparse Generalised Canonical Correlation Analysis

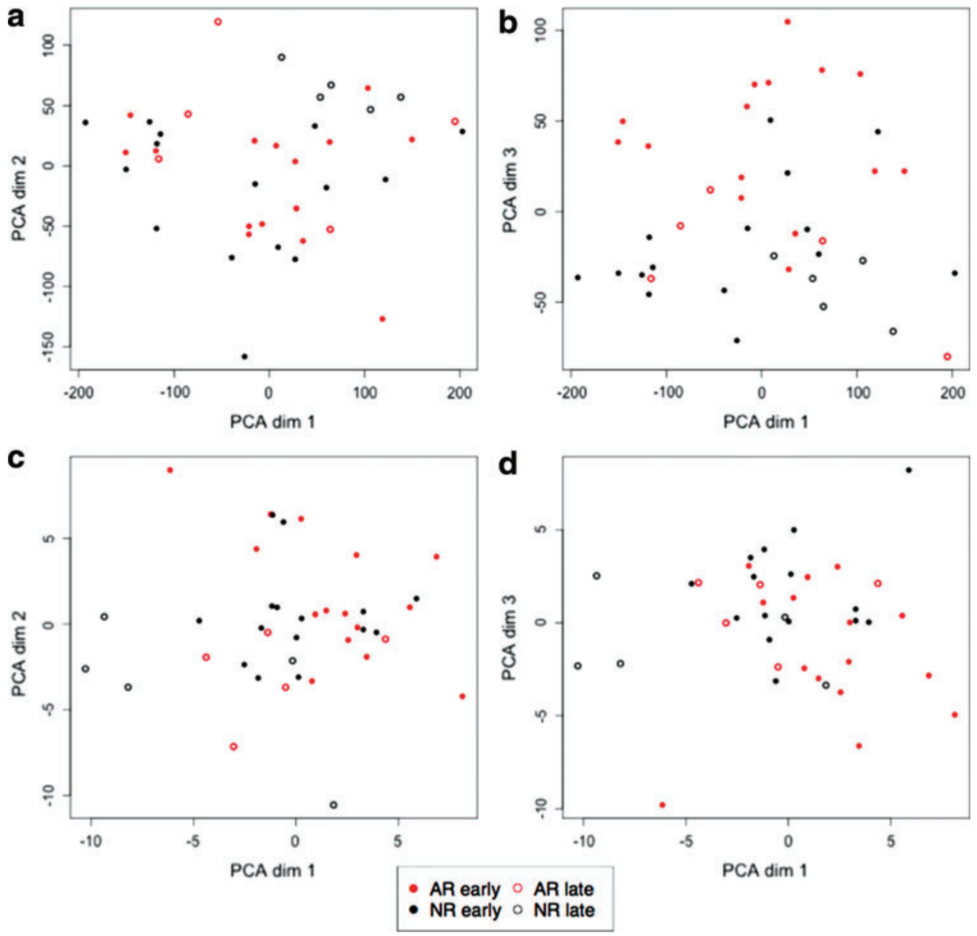
sGCCA is a sparse version of the methodology called Generalized Canonical Correlation Analysis (Tenenhaus et al., 2014), which provides a unifying framework for the integration of multiple data sets. Similar to sPLS but in the case of more than two data sets, the objective is to seek linear combinations of variables from each data set (components) so that each component explains its own data set, as well as the other data sets for which we assume there is a high correlation with its data set. The criterion to maximize is the covariance between pairwise components from two data sets (also called two 'blocks') at a time. The approach therefore requires an input design matrix to define the connection between each data set. Similarly to sPLS-DA, sGCCA can include an outcome factor, therefore allowing a supervised framework. Parameter tuning was performed as described above for the sPLS approach.



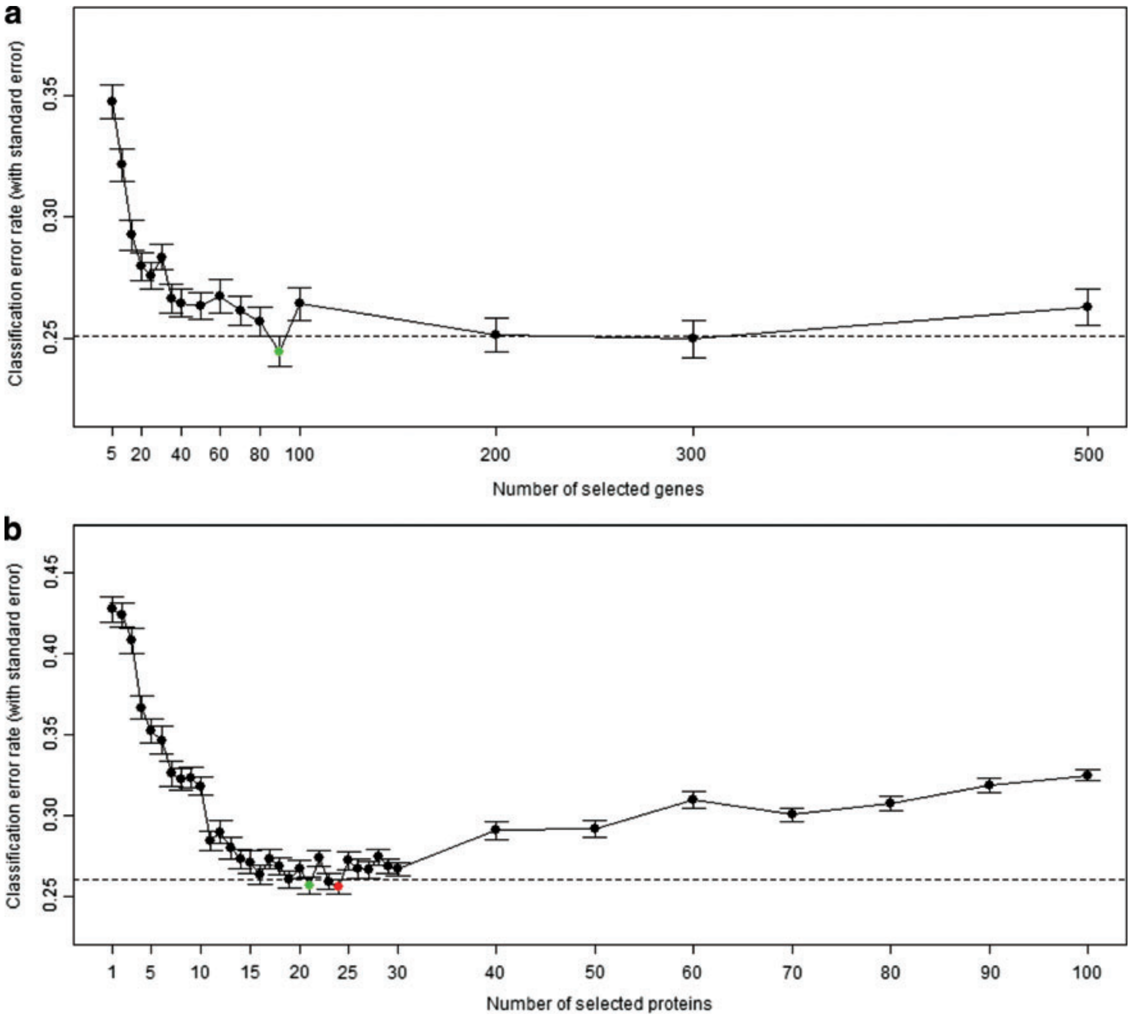
SUPPLEMENTARY FIG. S1. Overview of blood samples that were collected for the 40 patients in this study. Figure S1 displays the data available, the selected samples and time-point matching for all 40 patients, together with biopsy- and rejection treatment start-dates for the 20 AR patients. *Red lines* indicate Acute Rejection-patients (AR) and *gray lines* represent Non-rejection patients (NR). *Gray filled circles* show availability of genomics and proteomics sample data at the respective time post-transplant along the x-axis. *Circles with an inside plus-sign* represent the samples that were used in the study. *Red squares* indicate the time of biopsy for acute rejection patients, and *black inverted triangles* show when rejection treatment was initiated.



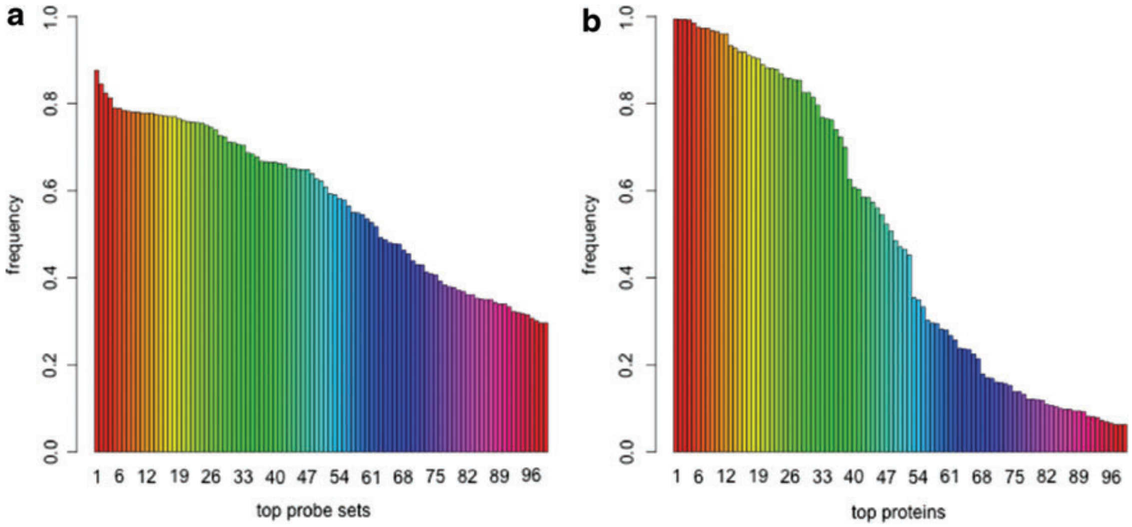
SUPPLEMENTARY FIG. S2. Scree plot of the Principal Component Analysis. **(a)** Genomics and **(b)** Proteomics data. *Light blue* indicate the number of chosen components (3).



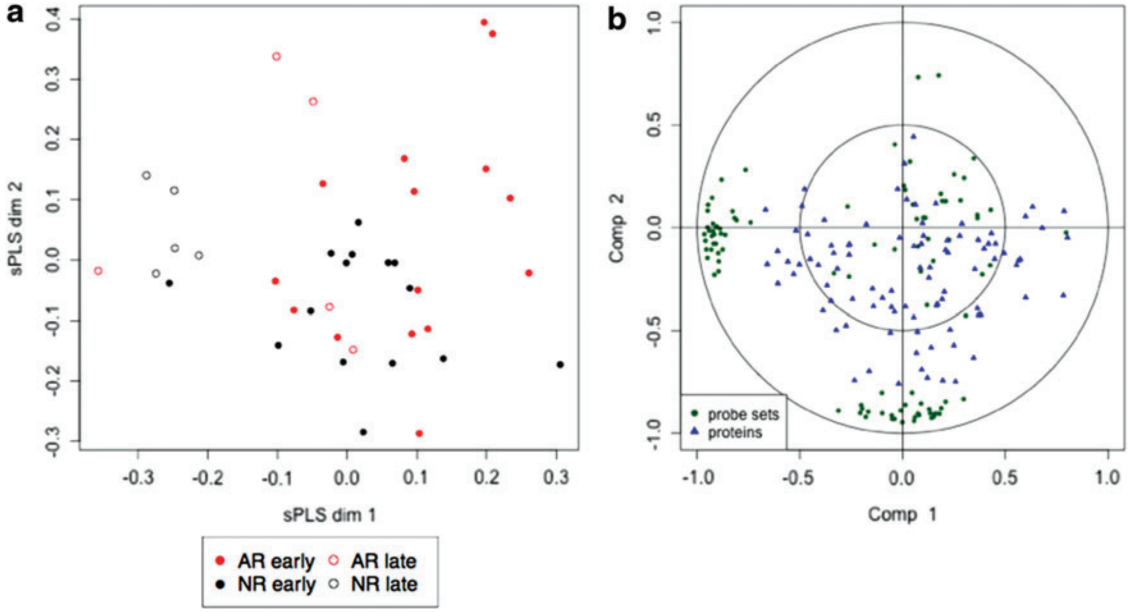
SUPPLEMENTARY FIG. S3. PCA sample representation. Samples were projected on the first two to three principal components for the genomics data **(a, b)** and the proteomics data **(c, d)**.



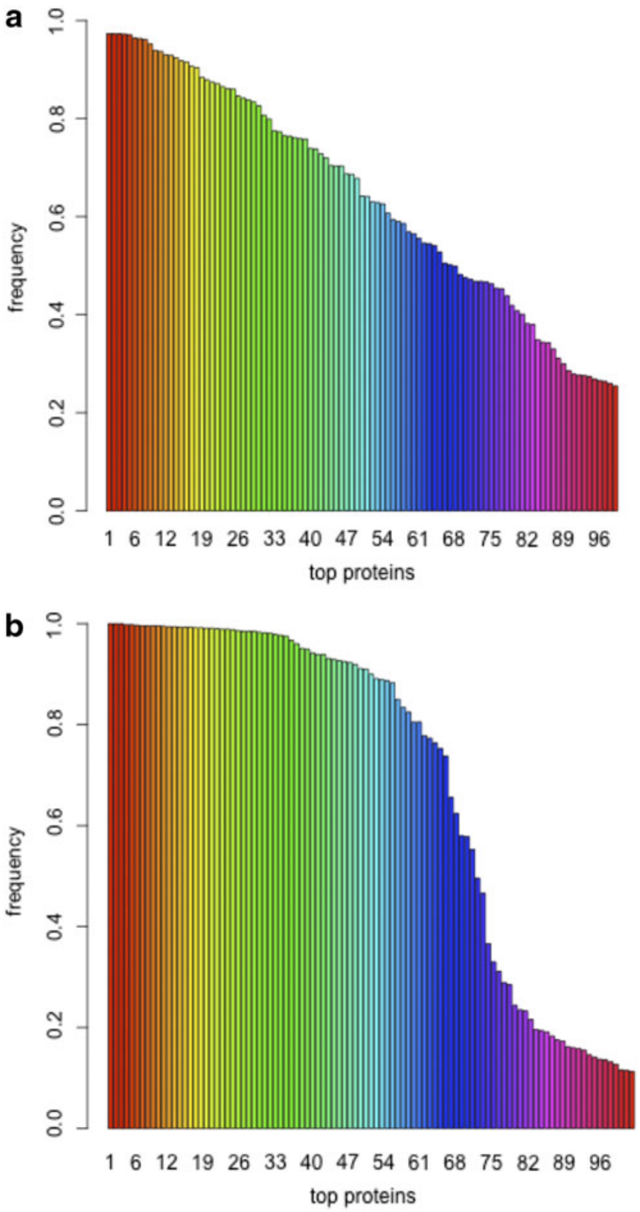
SUPPLEMENTARY FIG. S4. Average classification error rate with standard error with respect to the number of selected variables in a sPLS-DA model (1 component) for genomics **(a)** and proteomics **(b)** data. Classification errors are determined by averaging over 100×5-fold cross-validation runs for each of the number of selected features on the 26-sample training data. Shown in *red* is the location of the minimum classification error rate, while the number of selected features based on the one-standard-error-rule is indicated in *green*. For the genomics classifier, the two points were the same. The seven testing sample pairs that were used for testing in the supervised sPLS-DA analysis are: AR-2/NR-2, AR-5/NR-5, AR-8/NR-8, AR-10/NR-10, AR-13/NR-13, AR-14/NR-14 and AR-15/NR-15. The remaining 13 AR/NR sample pairs were used as the training set.



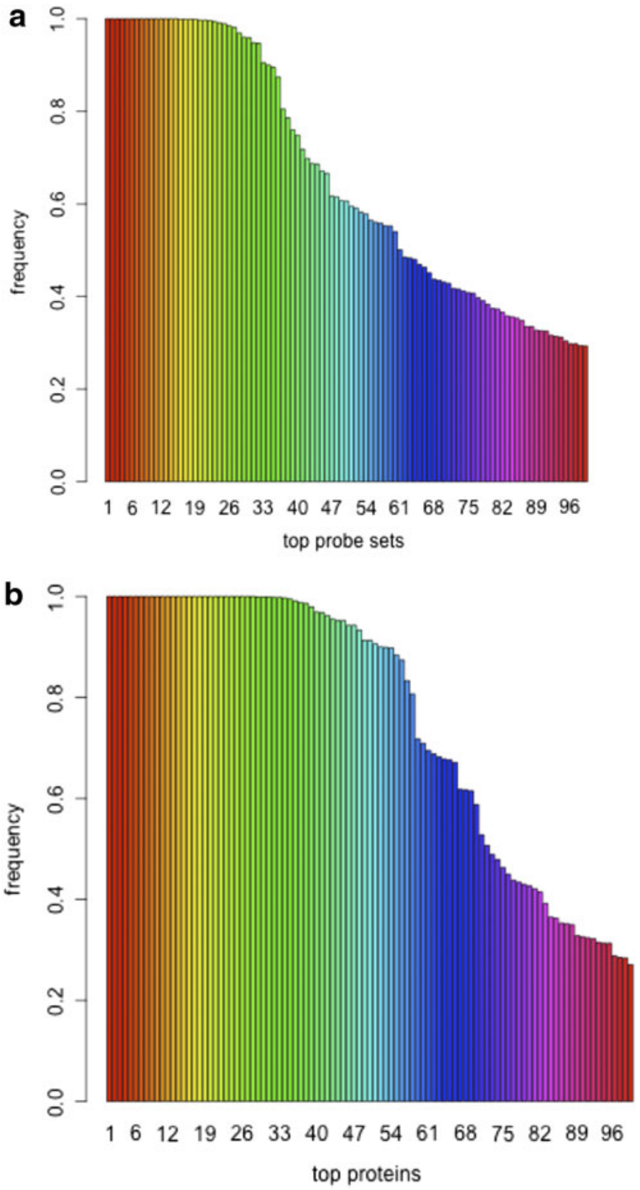
SUPPLEMENTARY FIG. S5. Stability analysis with sPLS. Using 10-fold cross validation over 1000 repetitions, the most frequently selected genes and proteins were determined with an arbitrary cutoff of 0.7.



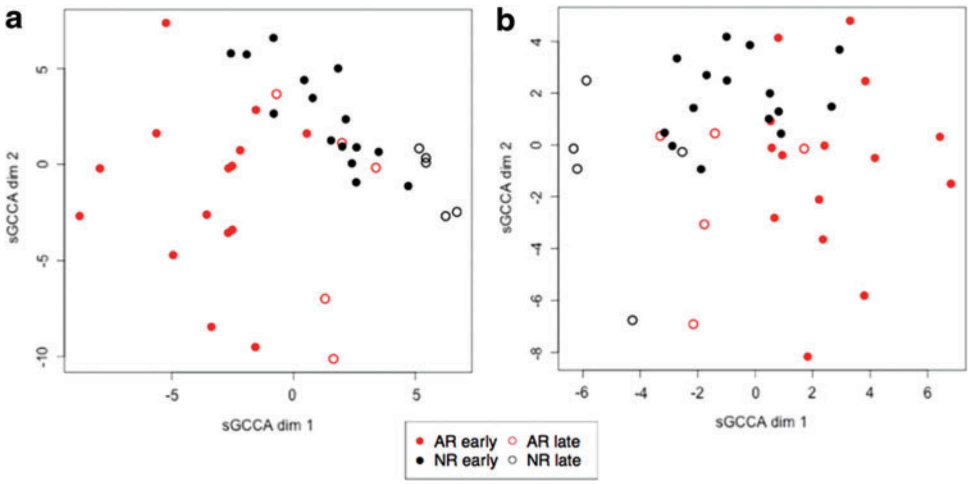
SUPPLEMENTARY FIG. S6. sPLS analysis. Sample representation **(a)** and variable representation **(b)** for the first two dimensions.



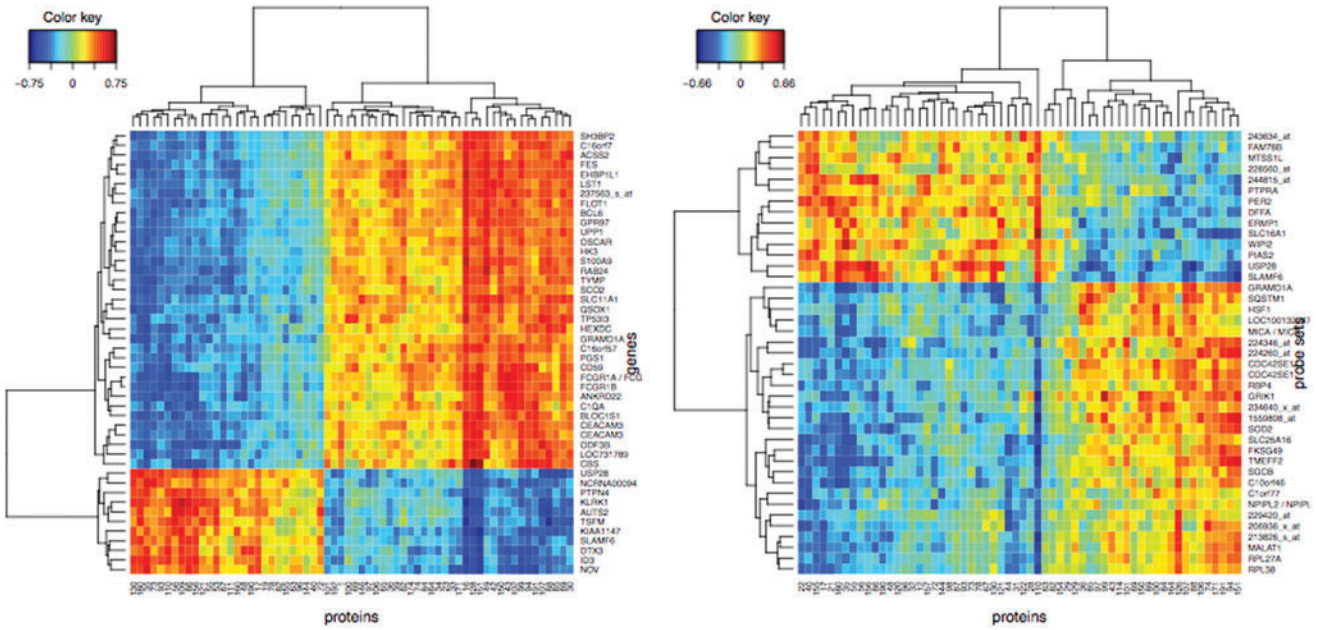
SUPPLEMENTARY FIG. S7. Stability analysis with sGCCA design 1. Using 10-fold cross validation over 1000 repetitions, the most frequently selected genes **(a)** and proteins **(b)** were determined with an arbitrarily chosen cutoff of 0.7.



SUPPLEMENTARY FIG. S8. Stability analysis with sGCCA design 2. Using 10-fold cross validation over 1000 repetitions, the most frequently selected genes **(a)** and proteins **(b)** were determined with an arbitrarily chosen cutoff of 0.7.



SUPPLEMENTARY FIG. S9. sGCCA analysis with design 1. Sample representation on the genomics space **(a)** and the proteomics space **(b)** for the first two dimensions for a selection of 46 genes and 64 proteins.



SUPPLEMENTARY FIG. S10. Clustered Image Maps (CIM) of 46 genes and 64 proteins selected with sGCCA-design 1 (*left*) and 41 genes and 60 proteins selected with sGCCA-design 2 (*right*). Genes are displayed in rows while proteins are shown in columns. The *blue* (*red*) color represents regions where genes and proteins are highly negatively (positively) correlated.