A combined test for feature selection on sparse metaproteomics data - alternative to missing value imputation

Sandra Plancade[1,†,*], Magali Berland[2,†], Mélisande Blein-Nicolas[3], Olivier Langella[3], Ariane Bassignani[2,3], Catherine Juste[4].

(1) INRAE, UR875 MIAT, F-31326 Castanet-Tolosan, France - Univ. Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France
(2) Univ. Paris-Saclay, INRAE, MGP, 78350, Jouy-en-Josas, France
(3) Univ. Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE-Le Moulon, F-91190, Gif-sur-Yvette, France - PAPPSO, doi:10.15454/1.5572393176364355E12, GQE-Le Moulon, F-91190, Gif-sur-Yvette, France
(4) Univ. Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France
∗ To whom correspondence should be addressed: sandra.plancade@inrae.fr
† Shared co-first authorship

June 22, 2021

Abstract

One of the difficulties encountered in the statistical analysis of metaproteomics data is the high proportion of missing values, which are usually treated by imputation. Nevertheless, imputation methods are based on restrictive assumptions regarding missingness mechanisms, namely "at random" or "not at random". To circumvent these limitations in the context of feature selection in a multi-class comparison, we propose a univariate selection method that combines a test of association between missingness and classes, and a test for difference of observed intensities between classes. This approach implicitly handles both missingness mechanisms. We performed a quantitative and qualitative comparison of our procedure with imputation-based feature selection methods on two experimental data sets. Whereas we observed similar performances in terms of prediction, the feature ranking from various imputation-based methods was strongly divergent. We showed that the combined test reaches a compromise by correlating reasonably with other methods.

Keywords: Metaproteomics, feature selection, missing value imputation, combined test

Acronyms

FSM	Feature Selection Method
FDR	False Discovery Rate
KNN	k-Nearest Neighbors
SVD	Singular Value Decomposition
SVM	Support Vector Machine
RF	Random Forest
MAR	Missing At Random
MNAR	Missing Not At Random

1 Introduction

Metaproteomics refers to the study of all proteins present in an ecosystem (soil, water, gut...) at a given time. It allows for the qualitative and quantitative profiling of the tremendous diversity of proteins in complex biological samples. It is the method of choice to learn about which microorganisms are doing what in a microbial ecosystem. Therefore, metaproteomics moves beyond the genetic potential addressed by metagenomics, and it is generating rising interest and new international initiatives (www.metaproteomics.org). Yet metaproteomics has long lagged behind metagenomics due to the lack of appropriate tools, but impressive progress in LC-MS/MS technologies (Liquid Chromatography coupled with tandem Mass Spectrometry) makes it possible to decipher metaproteomes in a deep, broad and high throughput manner. However, processing of metaproteomics data is much less developed than for metagenomics and statistical approaches developed for proteomics of single organisms cannot necessarily be transposed to complex ecosystems. Indeed, metaproteomics data are characterized by a huge diversity and specificity within and between samples; this generates large and sparse matrices of protein abundances which require dedicated analytical methods. In particular, selecting metaproteomic features that are shared by homogeneous clinical groups could facilitate the diagnosis or prognosis of a disease.

Feature selection methods (FSMs) can be classified in two categories. Wrapper and embedded methods make use of a classifier to select a set of features based on their discrimination ability, either with a recursive selection (wrapper) or by including a filtering into the classifier (embedded) (Saeys *et al.*, 2007). While these methods enable the extraction of a reduced list of predictors, they are pointed out as potentially generating overfit (Saeys *et al.*, 2007), and lead to the elimination of correlated features which may be detrimental to biological intepretation. In univariate methods, features are examined separately. These methods do not account for potential interactions amongst variables, but they enable the inclusion of more complex designs (batch effects, multiple effects, censoring,...).

One of the difficulties encountered to implement FSMs on shotgun proteomics and metaproteomics is to handle the missing data. Indeed, LC-MS/MS technologies are known to generate a high rate of missing values and this phenomenon is enhanced in metaproteomics. Indeed, on the one hand, microbiota composition is largely specific to individuals, leading to a significant proportion of truly missing proteins. On the other hand, the high complexity of microbiota samples makes data acquisition and pre-processing particularly sensitive, and generates a higher technical variability than observed on proteomics data, leading to important measurement errors as well as missing values. The processes leading to missingness are diverse and may originate from any step of the pipeline, either biochemical, analytical or bioinformatics (Lazar *et al.*, 2016). These mechanisms can be analysed in the framework developped by Rubin (1976), who distinguishes Missing At Random (MAR) in which the probability for a feature to be missing is independent of its true abundance, and Missing Not At Random (MNAR) in which missingness depends on the abundance, including notably thresholding due to device detection limit. It is commonly recognized that both MAR and MNAR occur with LC-MS/MS technologies (O'Brien *et al.*, 2018; Lazar *et al.*, 2016), but neither the proportion of each mechanism on a data set nor the precise mechanism at the origin of a given missing value are known a priori.

Methods to address missing data in proteomics mostly rely on either missing value imputation (Wang *et al.*, 2020) or statistical modelling of censoring mechanisms (Karpievitch *et al.*, 2009; Luo *et al.*, 2009; O'Brien *et al.*, 2018), even if a few alternative have arisen. Borrowing from both above mentioned categories, Berg *et al.* (2019) have recently proposed a multiple imputation approach based on a MAR/MNAR model. Besides, Webb-Robertson *et al.* (2010) developed a filtering approach that circumvents missing values imputation by means of two successive filtering based on difference in terms of peptide occurrence, and difference of intensities among the non-missing observations. But to our best knowledge, in the metaproteomics contex, the treatment of missing values mostly relies on imputation (Tang *et al.*, 2020a). A large number of imputation methods for proteomics or metaproteomics have been proposed in literature (R package `NAguideR`, Jin *et al.* (2021)), and can be classified in three categories (i) single value imputation, where missing intensities are replaced by the same value for all samples; (ii) global structure methods, in which imputation is based on correlations between the whole set of observations; (iii) local similarity imputation, based only on the most similar samples.

In this paper, we propose an approach which circumvents the limitations of missing value imputation and implicitly handles both MAR and MNAR mechanisms. This univariate feature selection method combines a presence/absence test which detects if the frequence of missingness is different between classes, and a test of the difference in observed intensities between classes, embedded in a permutation test procedure. We compared our method with three imputation-based FSMs, on two metaproteomics data sets: the first one from human gut microbiota of a cohort of coronary artery disease patients, and the second one from gut microbiota samples of pigs repeatedly measured in a diet perturbation experiment. Moreover, we made use of a set of technical replicates to explore missingness mechanisms.

2 Material and methods

2.1 Experimental data sets

2.1.1 *ProteoCardis.*

We used a subset of the data set generated in the ProteoCardis project, an association study between the human intestinal metaproteome and cardiovascular diseases (Bassignani, 2019, Section 1.6). Two classes were considered: patients with acute cardiovascular disease (n=49) and healthy controls (n=50). For each of these 99 subjects, the extracted gut microbiota was fractionated into its cytosolic and envelope compartments, which were analysed separately for their metaproteome, giving a total of 198 metaproteomes. Details on metaproteomics analyses can be found in Bassignani (2019) (Section 4.1.2). The cytosolic and envelope data sets are denoted by *ProteoCardis-cyto* and *ProteoCardis-env.*

In addition, we used eight biological samples from the ProteoCardis cohort, which were each replicated seven times for both their cytosolic and envelope compartments analysis.

The peptides and the proteins they come from were identified using an original iterative method described in Bassignani *et al.* (2021). Indistinguishable proteins, i.e. those identified with a same set of peptides, were grouped into metaproteins (or protein subroups) using the parsimonious grouping algorithm X!TandemPipeline (Langella *et al.*, 2017). To simplify the writing, those protein assemblages are denoted "proteins" in the following (Bassignani *et al.*, 2021; Bassignani, 2019). Finally, intensities of proteins were calculated as the sum of the extracted ion currents of their specific peptides, using MassChroQ (Valot *et al.*, 2011). Data are available at https://doi.org/10.15454/ZSREJA.

2.1.2 *Pigs.*

The data set *Pigs* consists in fecal microbiota analyses on 12 pigs observed at six times points during a four-week diet. Samples from weeks one and two (one observation per week) were gathered into a "metabolic period" and samples from weeks three and four (two observations per week) into an "equilibrium period". These two periods represent our classes of interest for this analysis, similarly to Tang *et al.* (2020a). All details can be found in Tilocca *et al.* (2017), and the data are available at ProteomeXchange PXD006224.

2.2 Filtering of sparse features

FSMs were applied on log-transformed data after filtering out proteins with less than τ non-missing values, with τ equal to 40% of the size of the smallest class ($\tau = 20$ for *ProteoCardis* and 10 for *Pigs*). This value represents a compromise between a high threshold that may lead to the deletion of a large part of the features, and a low threshold where too little information would be available for some variables. Nevertheless, as the impact of the missing value treatment may depend on the proportion of missing values, complementary analyses were performed with higher threshold values (30, 40 and 50 for *ProteoCardis* data sets; 20 and 30 for *Pigs*).

2.3 Combined test

We propose an univariate combined test that accounts for both missing and non-missing data. Consider m features (here, proteins) observed in n samples belonging to two or more classes. For each feature, let p_1 be the p-value of the Fisher exact test for association between class and missingness, computed by transforming observed intensities into 1 (not missing) or 0 (missing). Let p_2 be the p-value of the t-test (for two classes) or the F-test (for more than two classes) computed on observed log-intensities. Let S bet the Fisher combined statistic defined as:

$$S = -\frac{1}{2}(\log p_1 + \log p_2).$$

S is large if at least one of the two p-values p_1 and p_2 is small; Moreover, if only one of the two p-values is small, S is weakly affected by the value of the largest one. If the statistics of the two tests were independent, S would be χ^2-distributed under the global null hypothesis, but this assumption may be violated, especially under MNAR assumption, since low protein abundances could simultaneously lead to low observed intensities and high probability to be missing. Therefore, the distribution under the global null hypothesis is obtained by repeated permutations (N^{perm}) of the classes. In order to reduce the computing time and to increase the number of distinct values that can be taken by S, the distribution under the null hypothesis is assumed to be identical for all variables with the same proportion of missing values.

2.3.1 Permutations framework.

In the case of a complex design, single shuffling may be inappropriate since data are not freely exchangeable under the null hypothesis. Thus for *Pigs*, permutations were implemented such that the number of observations on each animal over each period was preserved (two time points and four time points over the first and second period respectively), using the R package `permute`.

2.3.2 Number of permutations.

For the prediction accuracy and the replicability on independent subsets, we considered $N^{perm} = 10^4$, and for the implementation on the whole data set, we considered $N^{perm} = 10^5$. A larger number of permutations leads to a better precision of the p-values (of order $1/N^{perm}$) but at the cost of a larger computation time which increases linearly with N^{perm}. Note that the procedure can be parallellized very easily, since the distribution under the null hypothesis is computed separately for each possible number of missing values.

2.3.3 Sample size requirement.

The combined test requires a sufficient sample size, so that enough permutations can be realised to compute the statistic distribution under the null hypothesis with a good precision. Notably, let τ be the filtering level, i.e. features with less than τ non-missing values are removed, then the number of possible distinct values for the t-test statistic is 2^τ, thus the p-value precision for the combined test is limited to $1/2^\tau$.

2.4 FSMs based on NA imputation methods

The combined test was compared with feature selection procedures based on missing value imputation. We considered the three imputation methods proposed by Tang *et al.* (2020b) (package `metaFS`), namely (i) single value imputation, where all missing value are replaced by the smallest intensity observed in the data set; (ii) k-Nearest Neighbours (KNN); (iii) Singular Value Decomposition (SVD). Following Wang *et al.* (2020), KNN was implemented using the R function `SeqKNN` (package `SeqKnn`) with a number of neighbours $k = 10$ and SVD was implemented using the function `pca` (package `pcaMethods`) with two components. In all cases, imputation was followed by the t-test; This choice guarantees that the differences observed between the imputation-based FSMs and the combined test are exclusively due to the treatment of missing values. The imputation-based FSMs are denoted as follows.

- **single-tt:** log-transformation + single value imputation + t-test
- **KNN-tt:** log-transformation +KNN imputation + t-test
- **SVD-tt:** log-transformation + SVD imputation + t-test

2.5 Resampling-based procedure for control of False Discovery Rate (FDR)

To account for correlation between variables, the False Discovery Rate (FDR) was controled using the resampling-based procedure proposed by Reiner *et al.* (2003): p-values were reestimated by resampling (100 times) from the marginal distribution prior to p-value adjustment (Benjamini and Hochberg, 1995). Even if most FDR procedures only guarantee an upper-bound control and are subject to assumptions on dependence between variables, the number of selected variables for a given FDR threshold is an indicator of the FSM's power. Moreover, the resampling-based FDR procedure considered here enables the bias due to complex dependence structures to be circumvented.

2.6 Criteria for method comparison

2.6.1 Stability between independent data sets.

The set of samples was repeatedly (100 times) split into two independent subsets while preserving the proportion of each class. FSMs were applied on each subset, and the stability was quantified as the proportion of common variables among the top N features selected on each of the two subsets.

2.6.2 Classification accuracy.

Classification accuracy was computed on a k-fold cross-validation loop ($k = 10$ for *ProteoCardis* and $k = 6$ for *Pigs*), repeated 10 times in order to evaluate the standard deviation. The classification procedure consists in selecting the top N proteins, and then infer either random forest (RF) or support vector machine (SVM) based on these N features. For prediction, missing values were replaced by zero. Filtering was performed on the complete data set (as this step does not involve class labels). For each cross-validation split, the whole classification procedure including feature selection was performed on the training set only, then the labels of the samples in the validation set were predicted.

2.6.3 Agreement between FSMs.

The Pearson correlation between vectors of log-transformed p-values quantifies the overall similarity between FSMs on all proteins, while putting more weights on those with low p-values via the log-transformation. The proportion of common selected features among the top N directly targets agreement in terms of feature selection. Values of N were chosen for each data set according to the number of significant features. For *ProteoCardis* data sets which display a small number of significant values, we considered $N = 30, 100, 200$. For *Pigs* for which a large number of proteins are significantly different between the two classes, we considered larger values $N = 200, 500, 1000$. Moreover, for *Pigs*, sample splitting in cross-validation loop and stability computation was implemented such that all observations from the same animal remained in the same subset.

2.7 Analysis of replicates

Consider a technical replication of the analysis of a biological sample i. The probability that a feature j is missing in the technical replicate, given that the observed intensity $X_{i,j}$ is equal to x in the original analysis, was defined as follows:

$$p^0_{i,j}(x) = \mathbb{P}(X'_{i,j} = \mathrm{NA} | X_{i,j} = x)$$

with $X'_{i,j}$ the observed intensity in the replicate. The replicate data sets were used to infer the missingness probability function. First of all, we assumed that the probability was independent of the biological sample and of the feature $p^0_{i,j}(x) = p^0(x)$. Moreover, observed intensities were stratified in 5% quantiles: $(x_0, \ldots, x_{20})$ and p^0 was approximated by:

$$p\left(\frac{x_\ell + x_{\ell+1}}{2}\right) = \mathbb{P}(X'_{i,j} = \mathrm{NA} | X_{i,j} = [x_\ell, x_{\ell+1})) = \frac{\mathbb{P}(X'_{i,j} = \mathrm{NA}, X_{i,j} \in [x_\ell, x_{\ell+1}))}{\mathbb{P}(X_{i,j} \in [x_\ell, x_{\ell+1}))}$$

which was estimated by its empirical counterpart:

$$\frac{\sum_{j=1}^{J}\sum_{i=1}^{8}\sum_{r,r'=1,\dots,7,r\neq r'} \mathbb{1}_{X_{i,j}^{r}=x, X_{i,j}^{r'}=\mathrm{NA}}}{J\times 8\times 7\times 6} \times \frac{J\times 8\times 7}{\sum_{j=1}^{J}\sum_{i=1}^{8}\sum_{r=1,\dots,7}\mathbb{1}_{X_{i,j}^{r}=x}}$$

with $X_{i,j}^{r}$ the intensity of the protein j in the replicate r of the biological sample i.

3 Results

3.1 Statistical characteristics of the experimental data sets.

Even after filtering of sparse features, *ProteoCardis* data sets were still highly sparse, with most proteins having more than half missing values while *Pigs* displays a larger proportion of proteins with very few missing values (Figure S1, top). These differences of sparsity may originate from a higher similarity in terms of genetic background and diet among experimental animals. Moreover, many more proteins were significantly different between the two classes in *Pigs* than in *ProteoCardis* for all FSMs (Figure S1, bottom), and the prediction accuracy was higher for *Pigs* (cf. Section 3.5). Thus *Proteocardis* and *Pigs* display different statistical characteristics, which enhance the robustness of the FSM comparison carried out in this paper.

3.2 Missingness mechanisms: both MAR and MNAR

The replicate data set, including seven technical replicates for eight biological samples, allows for the assessment of the technical variability and the analysis of the MAR/MNAR hypotheses. Figures 1 and S2 (left) display the average observed intensity of a protein as a function of the number of times it is missing in the replicate samples. The observed intensity decreased as the number of missing values increased, which suggests that the probability to be missing is higher when the protein abundance is lower, so missingness mechanisms is at least partially MNAR. In particular, we observed a pronounced difference of intensity between proteins with no missing value and protein with one, or *a fortiori* more than one, missing values. Nevertheless, even when the protein was missing in a large proportion of replicates, the average observed intensity could still be high, indicating that missingness mechanisms are not exclusively MNAR. These observations were confirmed by the probability of being missing, that decreased when the observed intensity increased, but remained non-negligible even for consistent observed intensities (Figures 1 and S2, right). For example, for an intensity equal to the median of the observed values on the data set, the probability of being missing was 0.23-0.25, and even for an intensity equal to the 0.9 quantile, the probability was still 0.05.

3.3 Concordance between the combined test and the imputation-based FSMs.

We compared our combined test with the t-test implemented after three imputation methods: KNN and SVM, based respectively on local and global structure similarity, and single value imputation by the smallest observed intensity.

First of all, FSMs were analysed in terms of correlation between log-transformed p-values and proportion of common selected features (Figures 2 and S3). Regarding imputation-based FSMs, we observed a strong agreement between SVD-tt and KNN-tt, but both methods showed a poor concordance with single-tt. The combined test reaches a compromise as it displayed a strong agreement with single-tt and a moderate agreement with KNN-tt and SVD-tt. These observations were common to the *ProteoCardis* and *Pigs* data sets, but the concordance between methods was globally higher for *Pigs* due to a weaker proportion of missing values.

The scatterplots between log-transformed p-values of the combined test and each imputation-based FSM (first row of Figures 3, 4 and S4) confirms the compromise reached by the combined test. Indeed, features found highly significant by any of the imputation based FSMs were at least moderately significant with the combined test, while the opposite was not true. For *ProteoCardis* data sets, the proteins with very low p-values ($p < 5.10^{-4}$) with the imputation-based FSMs also had low p-values with the combined test

($p < 5.10^{-2}$ for comparison with KNN-tt, and $p < 5.10^{-3}$ for comparison with SVD-tt and single-tt), but some proteins which displayed very low p-values with the combined test were found non significant with some of the imputation-based FSMs. On *Pigs*, most of the proteins that were found significant with SVD-tt or KNN-tt and non-significant with the combined test were very sparse (Figure S5), and thus included a large proportion of imputed values, which made them weakly reliable. On the contrary, the variables found to be significant with the combined test but non-significant with SVD-tt and KNN-tt had a sparsity level that was either low or high. These observations are coherent with the objectives of the combined test which targets the two missingness mechanisms.

As a complement, the two tests involved in the combined test, Fisher test on missingness and t-test on observed intensities, were implemented separately. These additional analyses enabled us to unravel the structure of the features detected by the combined test but not by one of the imputation-based FSMs (Rows 2 and 3 of Figures 3, 4 and S4). On the one hand, we observed a very strong correlation between p-values of the Fisher test on missingness and of single-tt. Indeed, as the smallest intensity used for imputation is far from most of the observed intensities (Figure S9), the proportion of imputed values among a class strongly impacts the average intensity after single value imputation. Therefore, the comparisons in terms of average intensity via the t-test and in terms of presence/absence lead to coherent p-values. On the other hand, the p-values from KNN-tt and SVD-tt correlated well with the p-values of the t-test on observed intensities (0.72-0.75 for *ProteoCardis*, and 0.61-0.62 for *Pigs*, row 3 of Figures 3, 4 and S4). Moreover, all proteins from *ProteoCardis* data sets (rows 3 of Figures 3 and S4), and almost all proteins from *Pigs* (row 3 of Figure 4), which were found to be highly significant with the t-test on observed values, were also significant with KNN-tt and SVD-tt. This indicates that the low p-values obtained with the combined test but not with KNN-tt or SVD-tt actually corresponded to features displaying a difference of occurrence between classes.

3.4 Impact of the filtering threshold

The impact of the imputation method is expected to decrease with the proportion of missing values, which itself depends on the filtering threshold. Therefore, we repeated our analyses with higher filtering thresholds: 30, 40 and 50 for *ProteoCardis*, and 20 and 30 for *Pigs*, to examine to what extent the comparisons between FSMs were impacted. Figures S6, S7 and S8 display the concordance between FSMs for several filtering thresholds. The comparisons between methods still held when threshold varied: SVD-tt/KNN-tt on the one hand, and single-tt/combined test on the other hand were strongly concordant, while KNN-tt/SVD-tt were moderately concordant with the combined test, and poorly concordant with single-tt. As expected, the agreement between all pairs of FSMs globally increased with the threshold.

A harder thresholding also results in a higher number of discarded variables, potentially leading to a loss of biological information. Indeed, Figure S10 displays the distribution of the p-values as a function of the protein sparsity; we observed that some very sparse proteins display very small p-values, in particular for *ProteoCardis*. Moreoever, Table S1 indicates that the sets of most significant proteins for each FSM included 57-82% (average 70%) of features with more than half of missing values for *ProteoCardis*, and 10-57% (average 42%) for *Pigs*. Note that this proportion of sparse proteins among the selected ones was only slightly lower than in the entire data sets (79% for *ProteoCardis* and 43% for *Pigs*). This indicates that sparse proteins were selected almost as frequently as less sparse ones.

3.5 Comparison of quantitative performances

Tables 1 and S2 display the average prediction accuracy and its standard deviation for SVM and RF classifiers applied on selected proteins. Even though the four FSMs selected very different sets of features, their performances in terms of prediction were not significantly different.

Figures 5 and S11 display the concordance between variable selection performed on independent data sets. On *ProteoCardis-cyto*, reproducibility of feature selection was similar with the four FSMs, while on *ProteoCardis-env* and *Pigs* the combined test and single-tt outperformed imputation-based FSMs, with a slight benefit of combined test on single-tt and of KNN-tt on SVD-tt. Therefore variable selection based on the combined test is either equally or more reproducible than imputation-based FSMs.

Figure S1 displays the number of selected variables for various values of FDR. The methods ranking varied with the data set and the FDR threshold, but the combined test remained competitive in terms of number of selected variables.

4 Discussion

4.1 Missingness blends MAR and MNAR mechanisms and our method addresses both assumptions.

Metaproteomics by LC-MS/MS generates a large proportion of missing values, usually imputed prior to statistical analysis. Several categories of imputation methods are routinely considered. Methods based on local similarity (e.g. KNN) or global structure (e.g. SVD) implicitely assume that missingness occurs independently of the true feature concentration (MAR). But the analysis of the replicate data set clearly indicates that missingness is more likely to occur when the feature has a low abundance. On the other hand, the single imputation method relies on the assumption of a left censoring mechanism, but the distribution of observed intensities as well as the analysis of the replicate data set are not consistent with this assumption. Even if the co-existence of MAR and MNAR mechanisms is admitted in LC-MS/MS proteomics, exploration of their prevalence is often based on biological replicates (e.g. Karpievitch *et al.* (2009)), assuming similar protein abundances in all samples from a given class. This assumption is questionable, in particular in human gut metaproteomics characterised by a strong individual specificity. On the contrary, our analysis based on technical replicates guarantees that the true protein abundances are identical.

The combined test presented in this paper addresses the two types of missingness by combining a Fisher exact test for association between missingness and classes, which corresponds to a MNAR assumption, and a test of difference between distribution of observed intensities after removal of missing values, which amounts to assume MAR mechanisms.

4.2 The combined test reaches a compromise between imputation-based methods.

Beyond differences in underlying assumptions, distinct imputation-based FSMs lead to very different sets of selected variables. While selection with the two MAR imputation methods (SVD and KNN) are concordant, these two methods display almost no agreement with the FSM based on MNAR assumption (single value imputation) for the highly sparse data sets *ProteoCardis*, and a low agreement for the moderately sparse data set *Pigs*. Therefore, the choice of an imputation method can strongly impact the biological conclusions and our combined test displays a correct agreement with each imputation-based FSM. In greater details, we observed that the features detected using single value imputation were recovered by the Fisher test on missingness, while the features selected using KNN or SVD were recovered by the t-test on observed values, which is consistent with the assumption on missingness mechanisms. Note that pairs of FSMs can simultaneously present a good concordance between lists of top variables for a given number of selected features, and lead to very different lists of variables for a fixed FDR threshold, due to power differences between testing procedures. Indeed, the combined test relies on permutations while the t-test used in imputation-based FSMs is based on parametric assumptions.

Finally, despite strong differences in terms of selected features, all FSMs analysed in this paper performs similarly in terms of classification, while the combined test may outperfom global and local structure imputations in terms of replicability.

4.3 Limits of missing value imputation in metaproteomics

Missing value imputation is a flexible approach that enables to address any type of statistical questions (e.g. prediction, network inference...) using methods developed for non missing data. But the downside is the risk to "forget" which values were imputed and to treat them equally to observed values, regardless of implicit assumptions underlying imputation that can strongly impact biological findings when a large proportion of values are missing (O'Brien *et al.*, 2018; Karpievitch *et al.*, 2012; Lazar *et al.*, 2016). In particular, we observed that global and local structure imputations specifically led to selection of features with a large proportion of imputed and thus weakly reliable values. Therefore, despite its easiness of use, imputation has a cost in terms of reliability and should be limited to moderately sparse data sets. On sparse metaproteomics data, this condition would require to filter out a large part of the features, which may be harmful since we demonstrated that a large part of the potentially interesting proteins have more than half of missing values.

As an alternative to missing value imputation, censored statistical models developed for proteomics data can account simultaneously for MAR and MNAR mechanisms (Karpievitch *et al.*, 2009; Luo *et al.*, 2009; O'Brien *et al.*, 2018). Moreover, Berg *et al.* (2019) proposed a multiple imputation model which handles MAR and MNAR assumptions,but their method suffers from a methodological bias since imputation is performed within each class, which artificially increases significance of inter-classes differences by enhancing intra-class similarities. Even though these models may address the complexity of missingness mechanisms more acutely than imputation methods, they also heavily rely on assumptions regarding missingness mechanisms (e.g. hard thresholding) as well as signal distribution (e.g. additive effect, Gaussian distribution). Therefore, they can not be applied to metaproteomics data whose structure and characteristics strongly differ from proteomics.

4.4 Perspectives

In this paper, we focused on the missing data issue and we restricted our analysis to FSMs based on the t-test, but the combined test could further be compared with more diverse FSMs, including wrapped and embedded methods (Tang *et al.*, 2020a). Moreover, the qualitative comparison of concordant and discordant features could be more widely used in FSM comparisons.

On the biological side, the conclusion of this article could be reinforced either through validation by targeted proteomics measurements on a subset of variables, or using spike-in data sets where the concentration of a small number of features is controlled.

Finally, the combined test developed in this article is not restricted to metaproteomics data and may be implemented on other meta-omics data or on any data including a large part of missing values, whatever the missingness mechanisms. Moreover, the proposed approach could be generalised to univariate feature selection in other frameworks than multi-class comparison (e.g. time series) provided that a test of presence/absence is available (e.g. a rank test for time series)

5 Software

Software in the form of R code, together with the ProteoCardis data sets are available at https://doi.org/10.15454/ZSREJA.

6 Supplementary Material

Contribution statements

S. Plancade and M. Berland conceived the original idea, performed the numerical implementation, analysed the data and wrote the article. M. Blein Nicolas and C. Juste participated to discussions about the method and revised critically the draft. ProteoCardis data were produced by C. Juste and pre-processed by O. Langella and A. Bassignani from samples provided by the Metacardis European FP7 initiative, and prepared and analyzed for the ANR project Proteocardis.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR) as part of the ProteoCardis (ANR-15-CE14-0013) project. We thank Alfred AMEADAN (Univ. Paris-Saclay, INRAE, APPSO-Micalis, 78350, Jouy-en-Josas, France) for preparing, analysing, and generating the raw LC-MS/MS data sets of the ProteoCardis study. Sylvie HUET (senior research scientist at MaIAGE-INRAE, UniversitÃľ Paris-Saclay, 78350 JOUY-en-JOSAS), Christine CARAPITO (senior research scientist at IPHC-CNRS, 67037 STRASBOURG) and Magali ROMPAIS (research engineer at LSMBO, IPHC-CNRS, 67037 STRASBOURG) are gratefully acknowledged for their support and valuable contribution to scientific discussions. *Conflict of Interest*: None declared.

References

Bassignani, A. (2019). *Metaproteomics analysis to study functionalities of the gut microbiota in large cohorts*. Theses, Sorbonne Université.

Bassignani, A., Plancade, S., Berland, M., Blein-Nicolas, M., Guillot, A., Chevret, D., Moritz, C., Huet, S., Rizkalla, S., Clément, K., Doré, J., Langella, O., and Juste, C. (2021). Benefits of Iterative Searches of Large Databases to Interpret Large Human Gut Metaproteomic Data Sets. *Journal of Proteome Research*, **20**(3), 1522–1534.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Berg, P., McConnell, E. W., Hicks, L. M., Popescu, S. C., and Popescu, G. V. (2019). Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. *BMC Bioinformatics*, **20**(102), 1471–2105.

Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X., and Tian, Y. (2021). A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports*, **11**(1), 1760.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W.-J., Yoon, H., Smith, R. D., and Dabney, A. R. (2009). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, **25**(16), 2028–2034.

Karpievitch, Y. V., Dabney, A. R., and Smith, R. D. (2012). Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics*, **13**(16), S5.

Langella, O., Valot, B., Balliau, T., Blein-Nicolas, M., Bonhomme, L., and Zivy, M. (2017). X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research*, **16**(2), 494–503.

Lazar, C., Gatto, L., Ferro, M., Bruley, C., and Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, **15**(4), 1116–1125.

Luo, R., Colangelo, C. M., Sessa, W. C., and Zhao, H. (2009). Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Stat Bioscis*, **1**(2), 228–245.

O'Brien, J. J., Gunawardena, H. P., Paulo, J. A., Chen, X., Ibrahim, J. G., Gygi, S. P., and Qaqish, B. F. (2018). The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *The annals of applied statistics*, **12**(4), 2075–2095.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**(3), 368–375.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Saeys, Y., Inza, I., and LarraÃśaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**(19), 2507–2517.

Tang, J., Wang, Y., Fu, J., Zhou, Y., Luo, Y., Zhang, Y., Li, B., Yang, Q., Xue, W., Lou, Y., Qiu, Y., and Zhu, F. (2020a). A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies. *Briefings in bioinformatics*, **21**(4), 1378–1390.

Tang, J., Mou, M., Wang, Y., Luo, Y., and Zhu, F. (2020b). MetaFS: Performance assessment of biomarker discovery in metaproteomics. *Briefings in Bioinformatics*.

Tilocca, B., Burbach, K., Heyer, C. M. E., Hoelzle, L. E., Mosenthin, R., Stefanski, V., Camarinha-Silva, A., and Seifert, J. (2017). Dietary changes in nutritional studies shape the structural and functional composition of the pigs? fecal microbiome?from days to weeks. *Microbiome*, **5**, 2049–2618.

Valot, B., Langella, O., Nano, E., and Zivy, M. (2011). Masschroq: A versatile tool for mass spectrometry quantification. *PROTEOMICS*, **11**.

Wang, S., Li, W., Hu, L., Cheng, J., Yang, H., and Liu, Y. (2020). NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Research*, **48**(14), e83–e83.

Webb-Robertson, B.-J. M., McCue, L. A., Waters, K. M., Matzke, M. M., Jacobs, J. M., Metz, T. O., Varnum, S. M., and Pounds, J. G. (2010). Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from ms-based proteomics data. *Journal of Proteome Research*, **9**(11), 5748–5756.

Figures

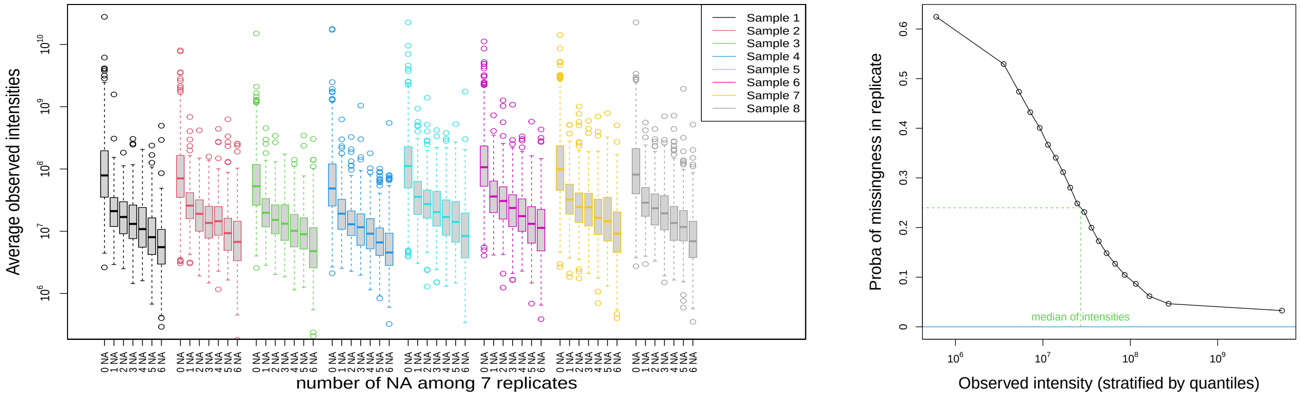
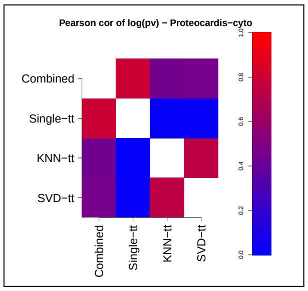


Figure 1: **Analysis of replicates - cytosolic fraction.** Left: log10-transformed average intensities of non-missing observations as a function of the number of missing values, for all proteins and for each biological sample. Right: Estimate of the probability that a protein is missing when the analysis is replicated, as a function of the average of its non-missing values.

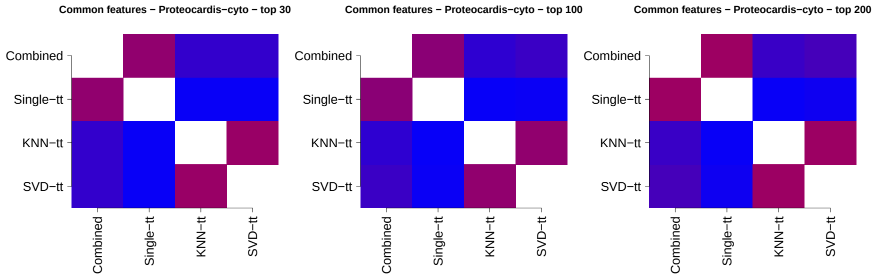
ProteoCardis-cyto

A - Pearson correlation



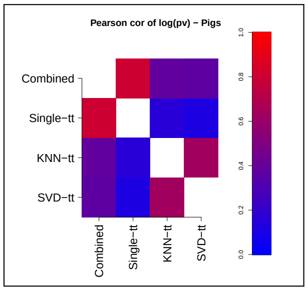
B - Proportion of common selected features

	Top 30			Top 100			Top 200		
	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt
combined	0.20	0.57	0.20	0.18	0.54	0.23	0.23	0.61	0.27
KNN-tt		0.03	0.60		0.03	0.57		0.04	0.61
single-tt			0.03			0.04			0.06



Pigs

C - Pearson correlation



D - Proportion of common selected features

	Top 200			Top 500			Top 1000		
	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt
combined	0.45	0.45	0.34	0.40	0.65	0.37	0.43	0.72	0.44
KNN-tt		0.15	0.45		0.21	0.53		0.30	0.61
single-tt			0.14			0.19			0.29

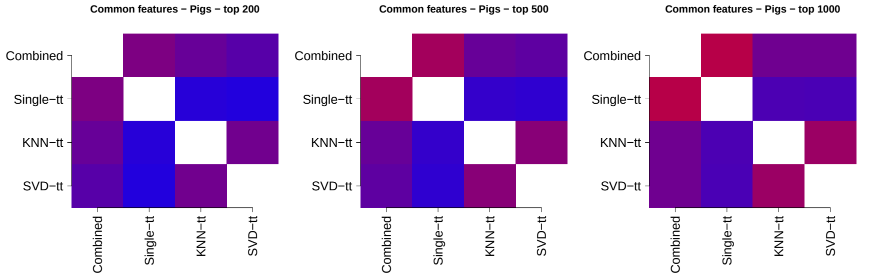


Figure 2: **Pairwise agreement between p-values of FSMs.** A,B: *ProteoCardis-cyto*; C,D: *Pigs*. A,C: Pearson correlation between log-transformed p-values. B,D: Proportion of common features among the top N for each pair of FSMs, as a table and a heatmap.

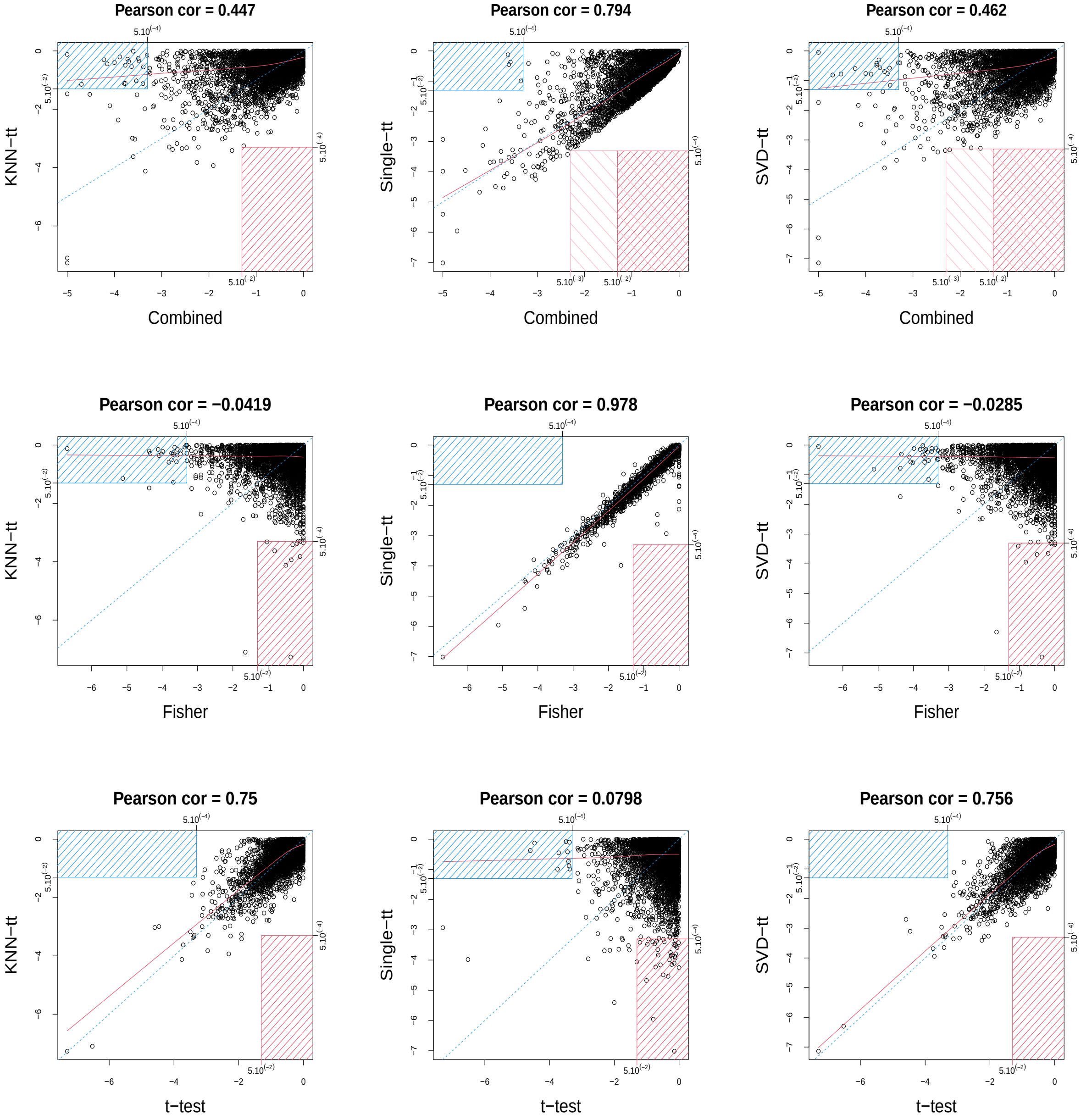


Figure 3: **Scatterplots between log10-transformed p-values of pairs of FSMs for** *ProteoCardis-cyto*. Row 1: combined test and imputation-based FSM. Row 2: Fisher test for missingness and imputation-based FSMs; proteins with less than 2 non-missing values are not displayed. Row 3: t-test on observed values and imputation-based FSMs. For each pair of testing procedure, the red rectangle corresponds to proteins with $p > 5.10^{-2}$ with the first procedure and with $p < 5.10^{-4}$ for the second procedure; conversely, the blue rectangle corresponds to proteins with $p < 5.10^{-4}$ with the first procedure and with $p > 5.10^{-2}$ for the second procedure.

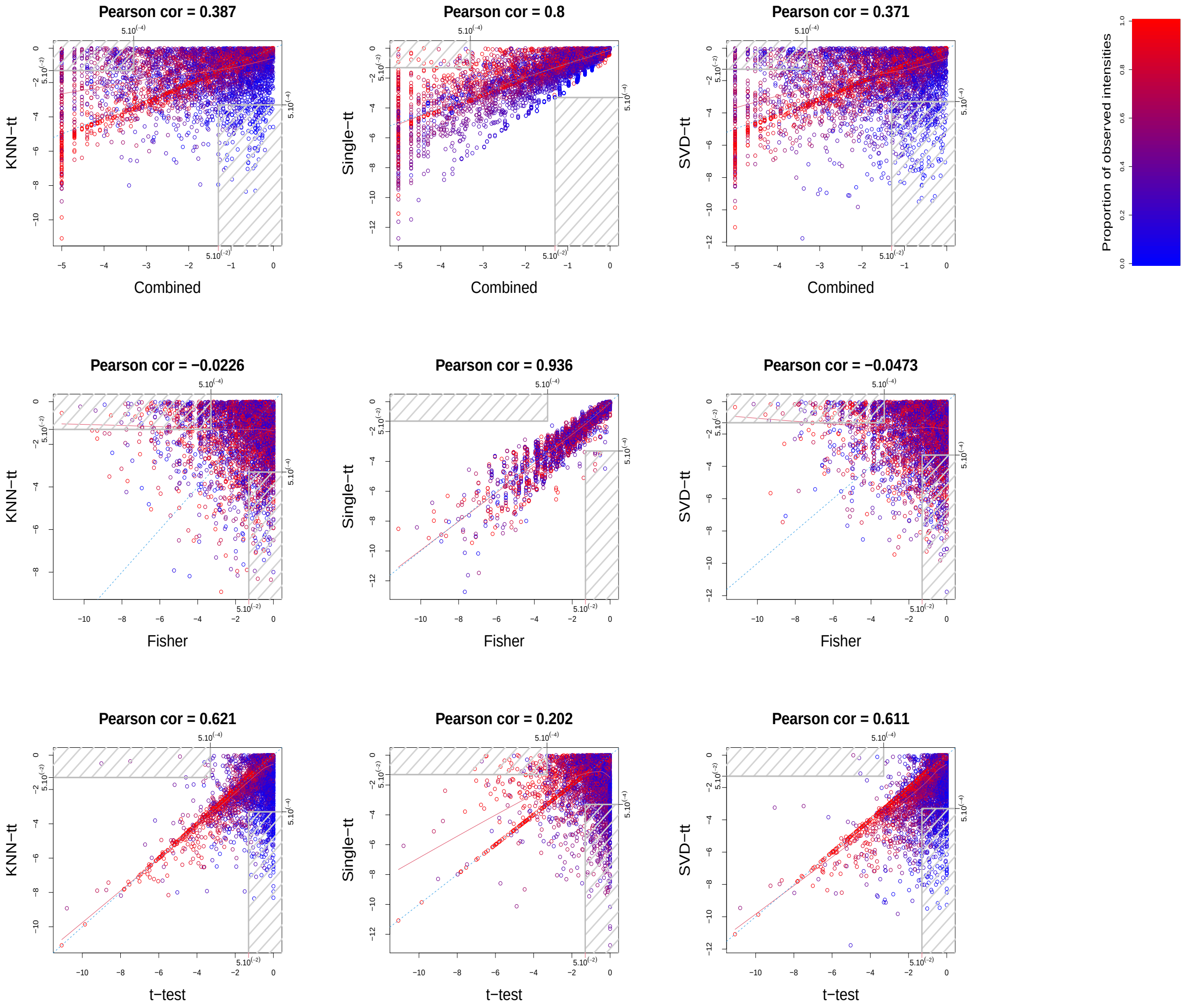


Figure 4: **Scatterplots between log10-transformed p-values of pairs of FSMs for** *Pigs.* Row 1: combined test and imputation-based FSMs. Row 2: Fisher test and imputation-based FSMs; proteins with less than 2 non-missing values are not displayed. Row 3: t-test on observed values and imputation-based FSM. Color gradient corresponds to the proportion of non-missing values for each protein. Gray rectangle correspond to features with $p < 5.10^{-4}$ with one FSM and $p > 5.10^{-2}$ with the other.

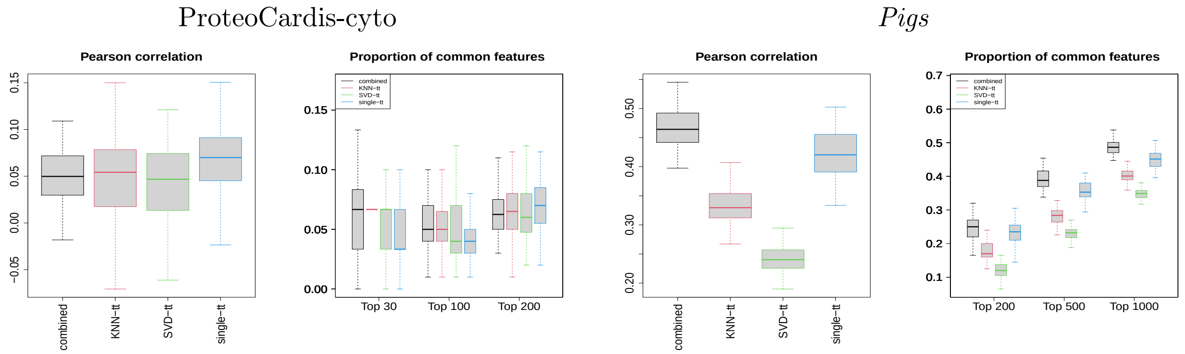


Figure 5: **Replicability of variable selection on independent subsets.** Pearson correlation between log10-transformed p-values and proportion of common selected features among the top N, for 100 splitting of samples into two subsets. Datasets: *ProteoCardis-cyto* and *Pigs*

Tables

ProteoCardis-cyto

		Combined	KNN-tt	SVD-tt	single-tt
Top 30	RF	**0.715**(0.029)	0.682(0.026)	0.706(0.022)	0.695(0.028)
	SVM	**0.708**(0.029)	0.612(0.043)	0.615(0.031)	0.68(0.032)
Top 100	RF	**0.715**(0.021)	0.681(0.027)	0.703(0.027)	0.713(0.028)
	SVM	**0.701**(0.025)	0.59(0.034)	0.656(0.029)	0.698(0.022)
Top 200	RF	0.711(0.017)	0.676(0.024)	0.69(0.038)	**0.712**(0.019)
	SVM	0.706(0.027)	0.608(0.028)	0.67(0.034)	**0.712**(0.014)

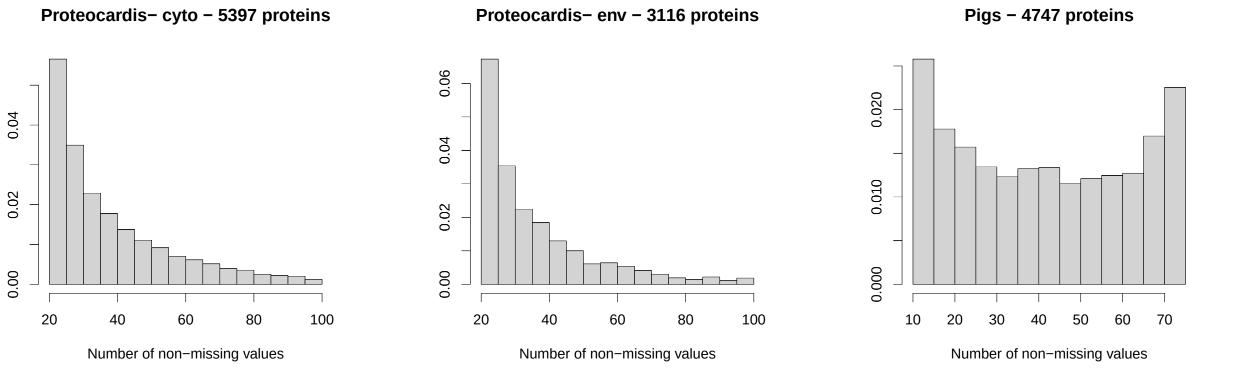
Pigs

		Combined	KNN-tt	SVD-tt	single-tt
Top 200	RF	**0.904**(0.01)	0.901(0.012)	**0.904**(0.0079)	**0.904**(0.0079)
	SVM	0.868(0.019)	0.862(0.021)	**0.906**(0.013)	0.896(0.024)
Top 500	RF	**0.904**(0.012)	0.903(0.0093)	0.901(0.012)	0.897(0.0097)
	SVM	0.867(0.018)	0.86(0.02)	**0.883**(0.0072)	0.844(0.016)
Top 1000	RF	0.901(0.01)	0.903(0.0093)	**0.904**(0.0079)	0.9(0.011)
	SVM	**0.897**(0.012)	0.85(0.016)	0.876(0.012)	0.824(0.0094)

Table 1: **Prediction accuracy** for two classification procedures on *ProteoCardis-cyto* and *Pigs*. The selection of the top N variables ($N = 30, 100, 200$) was followed by SVM or RF. Accuracy was computed in a 10-fold cross-validation loop, repeated 10 times. Each cell provides the average accuracy (standard deviation of accuracy) computed over the 10 repetitions of the cross-validation. Bold numbers correspond to the highest accuracy among the four FSMs.

.

Supplementary material



	ProteoCardis-cyto				*ProteoCardis-env*				*Pigs*			
	combined	KNN-tt	single-tt	SVD-tt	combined	KNN-tt	single-tt	SVD-tt	combined	KNN-tt	single-tt	SVD-tt
q < 0.01	0	2	3	2	0	4	2	1	962	828	789	1409
q < 0.05	11	2	33	2	31	7	11	2	1709	1547	1584	2096
q < 0.1	37	5	73	2	49	18	22	4	2140	1985	1993	2430
q < 0.2	151	14	230	14	138	105	69	4	2725	2557	2566	2878

Figure S1: **Statistical characteristics of the three data sets** *ProteoCardis-cyt*, *ProteoCardis-env*, *Pigs*. Top: frequencies of the number of non-missing values for all proteins after filtering (threshold 20 for *ProteoCardis*, and 10 for *Pigs*). Bottom: number of selected variables with the resampling FDR procedure with 100 resampling repetitions, with various values of the FDR threshold values.

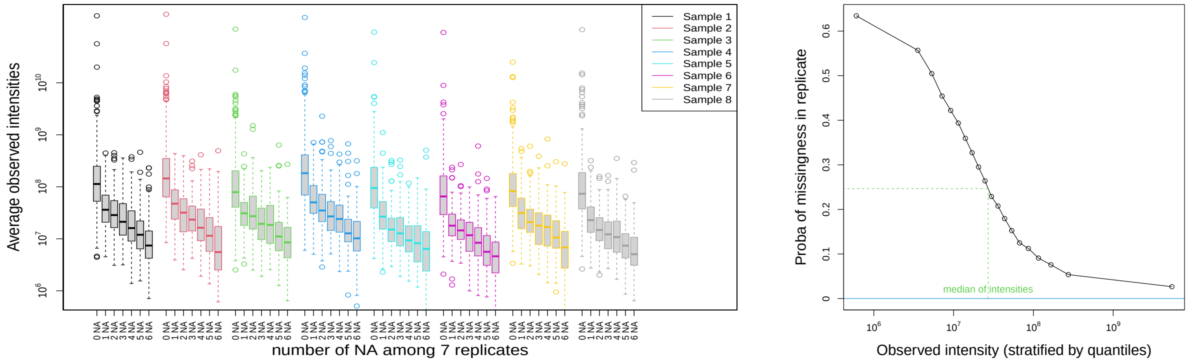
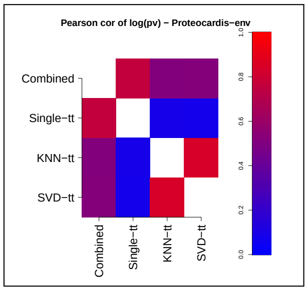


Figure S2: **Analysis of replicates - envelope fraction** Left: log10-transformed average intensities of non-missing observations, as a function of the number of missing values, for all proteins and for each biological sample. Right: Estimate of the probability that a protein is missing in a technical replicate as a function of the average of its non-missing values.

ProteoCardis-env

A - Pearson correlation



B - Proportion of common selected features

	Top 30			Top 100			Top 200		
	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt	KNN-tt	single-tt	SVD-tt
combined	0.27	0.40	0.30	0.39	0.52	0.38	0.39	0.58	0.41
KNN-tt		0.07	0.53		0.10	0.68		0.15	0.67
single-tt			0.07			0.10			0.15

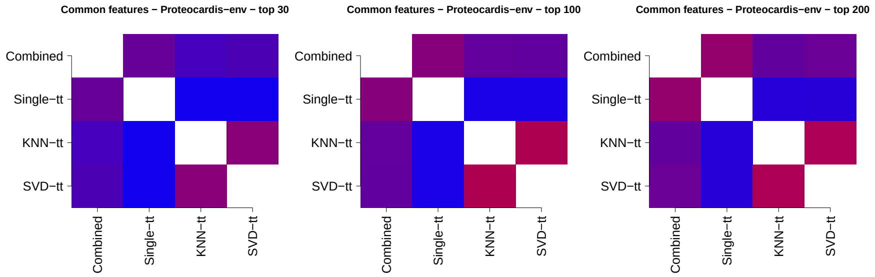


Figure S3: **Pairwise agreement between p-values of FSMs for** *ProteoCardis-env***.** A: Pearson correlation between log of p-values. B: Proportion of common features among the top N ($N = 30, 100, 200$) for each pair of FSMs, as a table and a heatmap.

Proportion of sparse features among selected

	ProteoCardis-cyto			*ProteoCardis-env*			*Pigs*		
	top30	top100	top200	top30	top100	top200	top200	top500	top1000
combined	0.60	0.66	0.69	0.73	0.73	0.76	0.12	0.21	0.29
KNN-tt	0.67	0.61	0.68	0.70	0.76	0.82	0.45	0.53	0.54
SVD-tt	0.70	0.66	0.69	0.67	0.75	0.80	0.37	0.43	0.47
single-tt	0.63	0.68	0.72	0.57	0.69	0.78	0.58	0.55	0.54

Table S1: Proportion of selected variables with less than half observed intensities, among the top N variables (between 20 and 50 non-missing values for *ProteoCardis* data sets, and between 10 and 36 for *Pigs*).

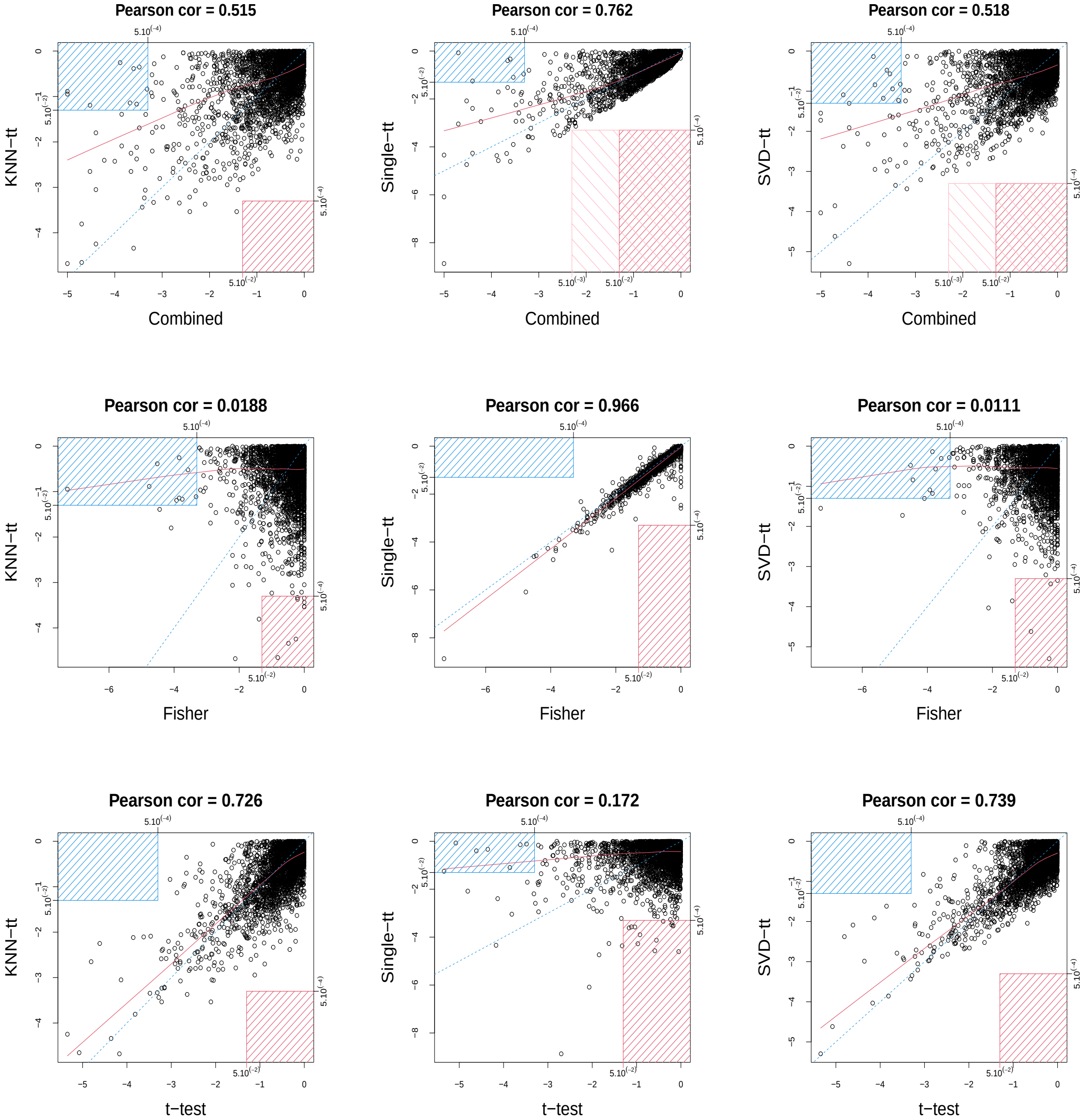


Figure S4: **Scatterplots between log10-transformed p-values of pairs of FSMs for** *ProteoCardis-env*. Row 1: combined test and imputation-based FSMs. Row 2: Fisher test for missingness and imputation-based FSMs; proteins with less than 2 non-missing values are not displayed. Row 3: t-test on observed values and imputation-based FSMs. For each pair of testing procedure, the red (resp. red and pink) rectangle corresponds to proteins with $p > 5.10^{-2}$ (resp.$p > 5.10^{-3}$) with the first procedure and with $p < 5.10^{-4}$ for the second procedure; conversely, the blue rectangle corresponds to proteins with $p < 5.10^{-4}$ with the first procedure and with $p > 5.10^{-2}$ for the second procedure.

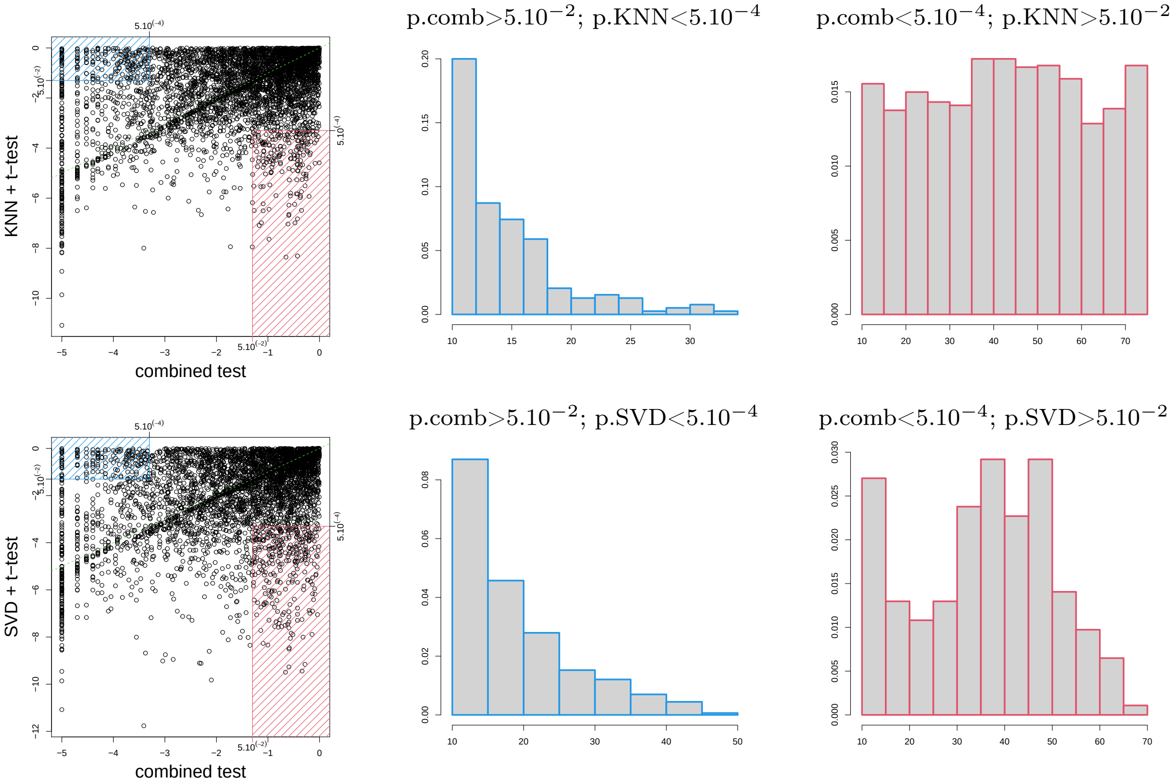


Figure S5: **Sparsity for proteins which are discordant** between the combined test and KNN-tt (first row) or SVD-tt (second row), on *Pigs*. Column 1: scatterplot of log10-transformed p-values of pairs of FSMs; the red rectangle corresponds to proteins with $p > 5.10^{-2}$ with the first procedure and with $p < 5.10^{-4}$ for the second procedure; conversely, the blue rectangle corresponds to proteins with $p < 5.10^{-4}$ with the first procedure and with $p > 5.10^{-2}$ for the second procedure. Column 2 (resp. 3): frequencies of proportion of observed values for proteins in the blue (resp. red) rectangle.

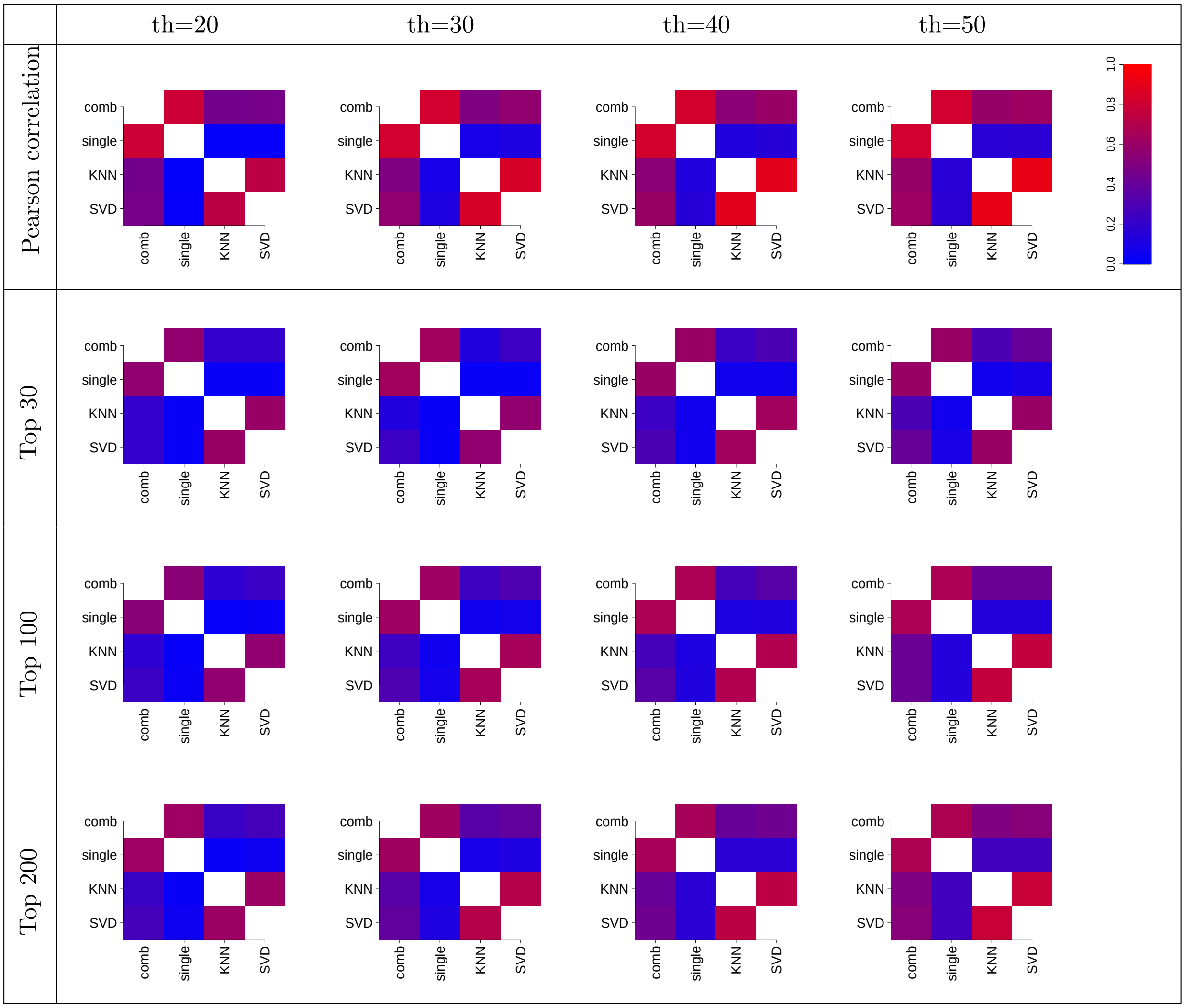


Figure S6: **Pairwise agreement between p-values from the four FSMs, for filtering threshold of 20, 30, 40 and 50 for** *ProteoCardis-cyto*. Each row correspond to a criterion; row 1: Pearson correlation between log-transformed p-values; rows 2 to 4: proportion of common variables among the top N variables with $N = 30, 100, 200$. Each column correspond to a threshold value.

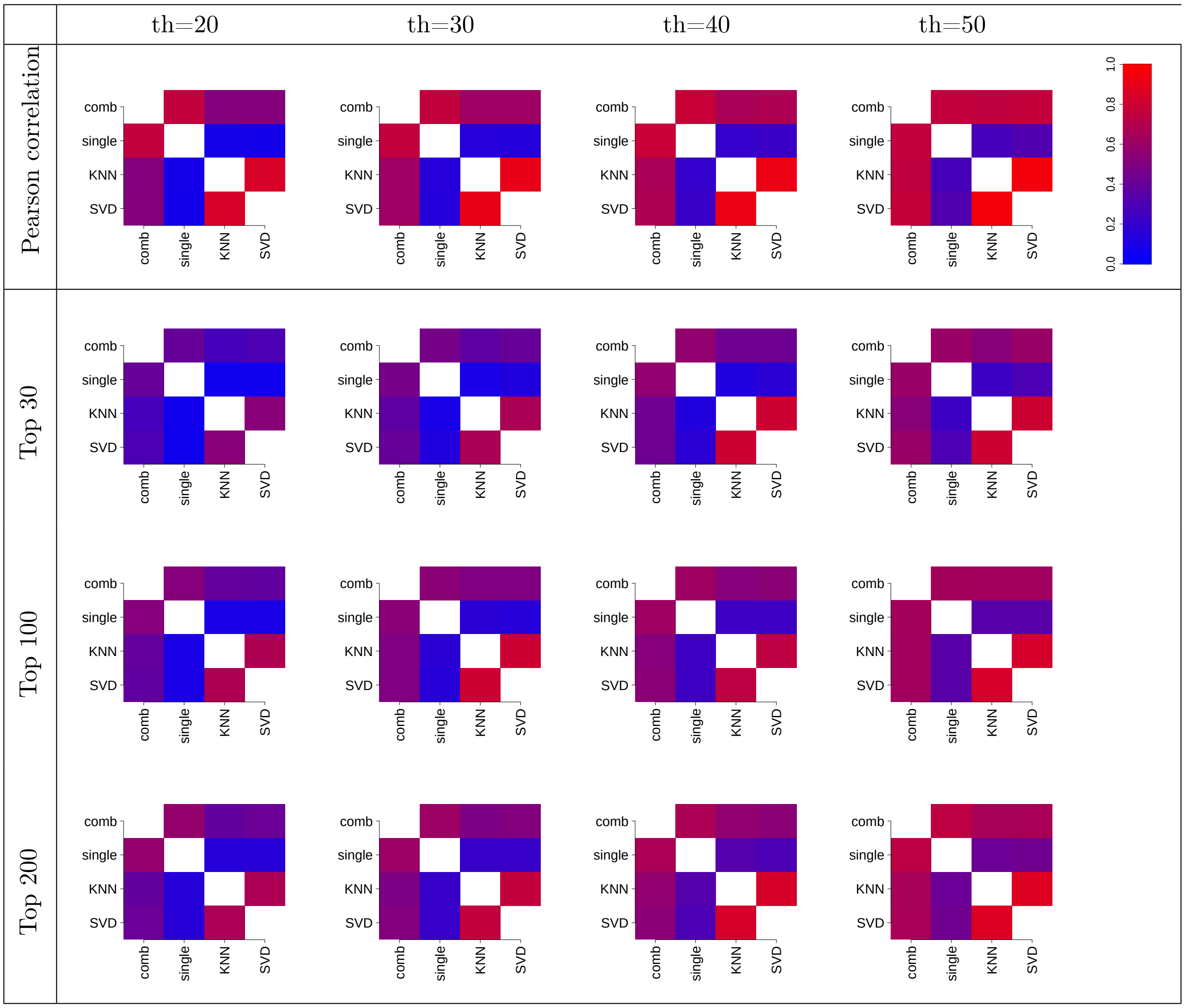


Figure S7: **Pairwise agreement between p-values from the four FSMs, for filtering threshold of 20,30,40 and 50 for** *ProteoCardis-env*. Each row correspond to a criterion; row 1: Pearson correlation between log-transformed p-values; rows 2 to 4: proportion of common variables among the top N variables with $N = 30, 100, 200$. Each column correspond to a threshold value.

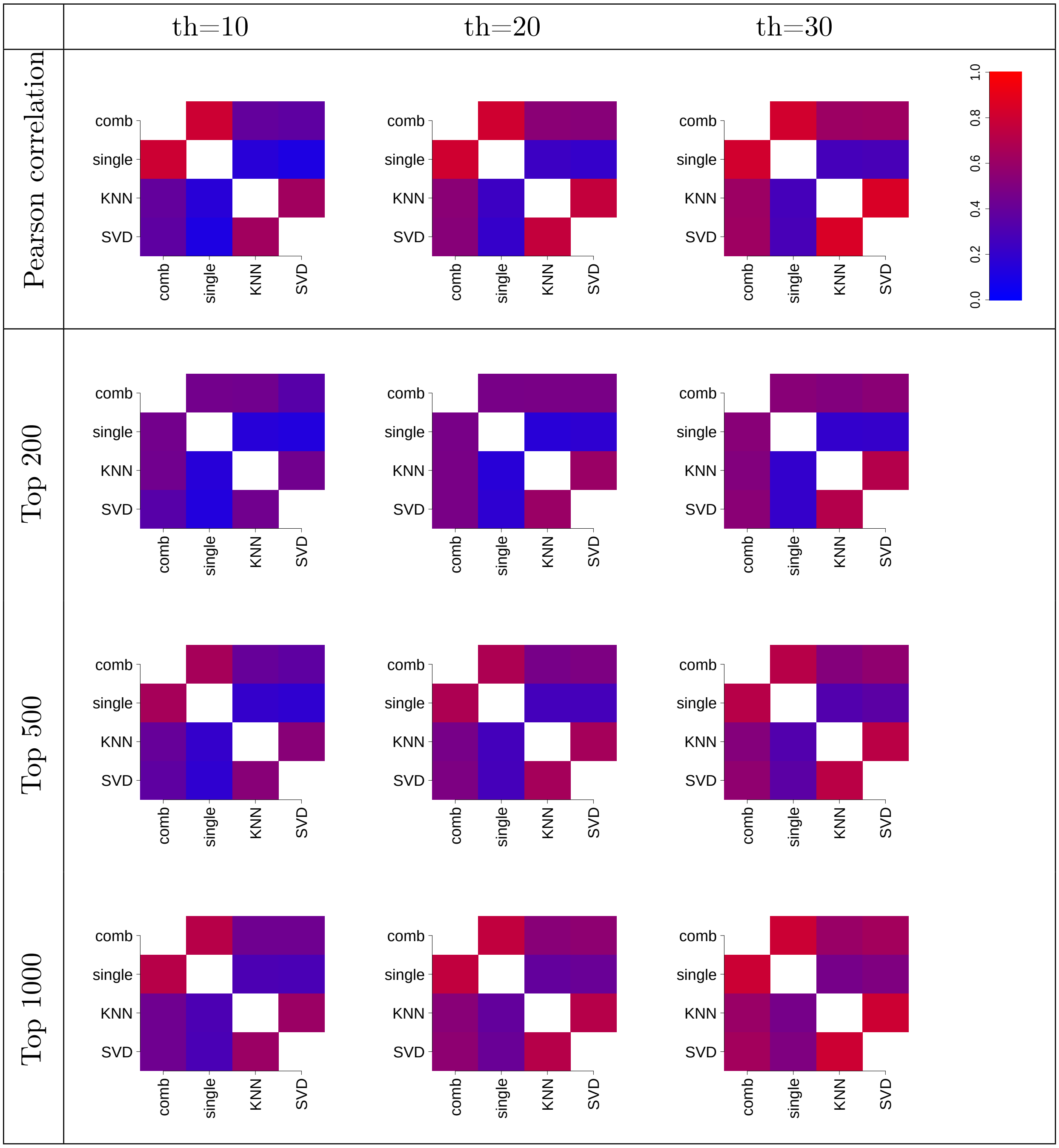


Figure S8: **Pairwise agreement between p-values from the four FSMs, for filtering threshold of 20 and 30 for** *Pigs*. Each row correspond to a criterion; row 1: Pearson correlation between log-transformed p-vlaues; rows 2 to 4: proportion of common variables among the top N variables with $N = 200, 500, 1000$. Each column correspond to a threshold value.

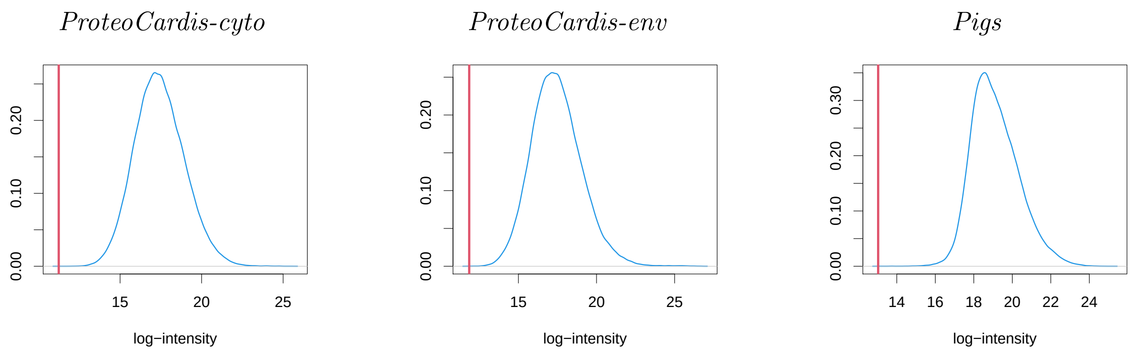


Figure S9: **Single value imputation.** Distribution of observed log-transformed intensities (blue) and imputed value (red) with single value imputation.

		FS combined test	FS KNN + t-test	FS SVD + t-test	FS single value +t-test
Top 30	RF	**0.775**(0.027)	0.727(0.022)	0.729(0.019)	0.763(0.02)
	SVM	**0.761**(0.03)	0.667(0.05)	0.678(0.038)	0.731(0.025)
Top 100	RF	0.755(0.018)	**0.766**(0.021)	0.745(0.028)	0.742(0.023)
	SVM	**0.768**(0.022)	0.697(0.027)	0.69(0.036)	0.719(0.019)
Top 200	RF	**0.743**(0.019)	0.735(0.029)	0.741(0.021)	0.737(0.022)
	SVM	**0.768**(0.021)	0.689(0.027)	0.693(0.037)	0.714(0.024)

Table S2: **Prediction accuracy** for two classification procedures on *ProteoCardis-env*. The selection of the top N variables ($N = 30, 100, 200$) was followed by SVM or RF. Accuracy was computed in a 10-fold cross validation loop, repeated 10 times. Each cell provides the average accuracy (standard deviation of accuracy) computed over the 10 repetitions of the cross-validation. Bold numbers correspond to the highest accuracy among the four FSMs

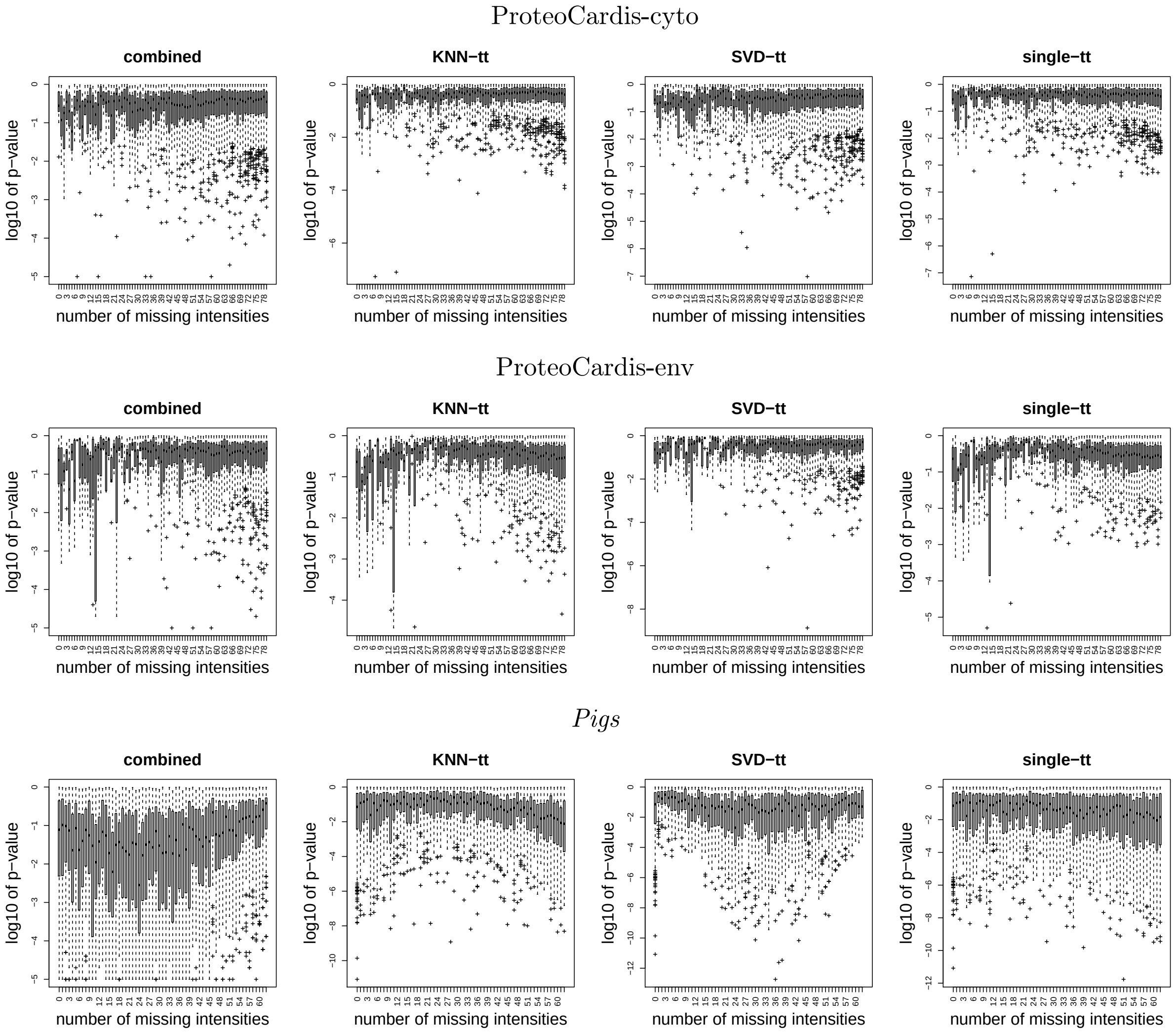


Figure S10: **Log10-transformed p-values as a function of sparsity**. The x-axis corresponds to the number of missing values among the 99 samples for *ProteoCardis* data sets, and among the 72 samples for *Pigs*.

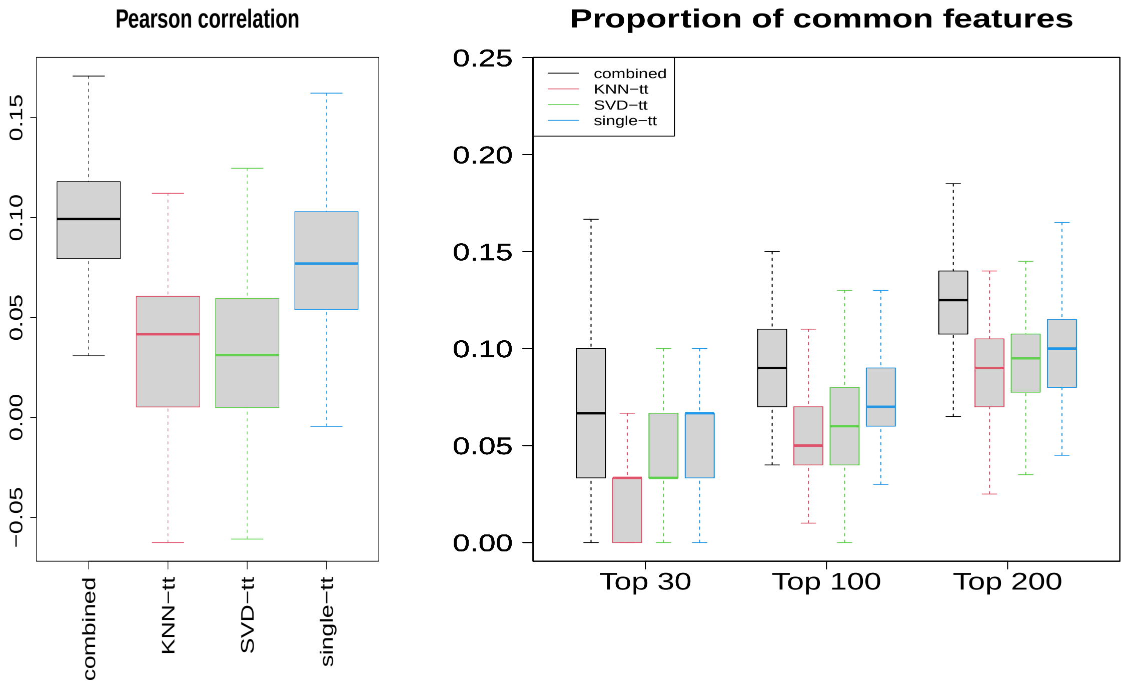


Figure S11: **Replicability of variable selection on independent subsets.** Pearson correlation between log-transformed p-values and proportion of common variables among the top N for 100 splitting of samples into two subsets. Dataset: *ProteoCardis-env*.