Feature selection in metaproteomics: how to deal with missing values? Toward a more qualitative analysis

Sandra Plancade[1,2], Magali Berland[3], Catherine Juste[4], Melisande Blein-Nicolas[5], Olivier Langella[5]

(1) Unité MaIAGE, INRAE, (2) Unité MIAT, INRAE, (3) Unité MGP, INRAE (4) Unité MICALIS, INRAE, (5) Unité GQE Le Moulon, INRAE

4 february 2021

Statistical analysis of untargeted omics data

Statistical analysis of untargeted omics data

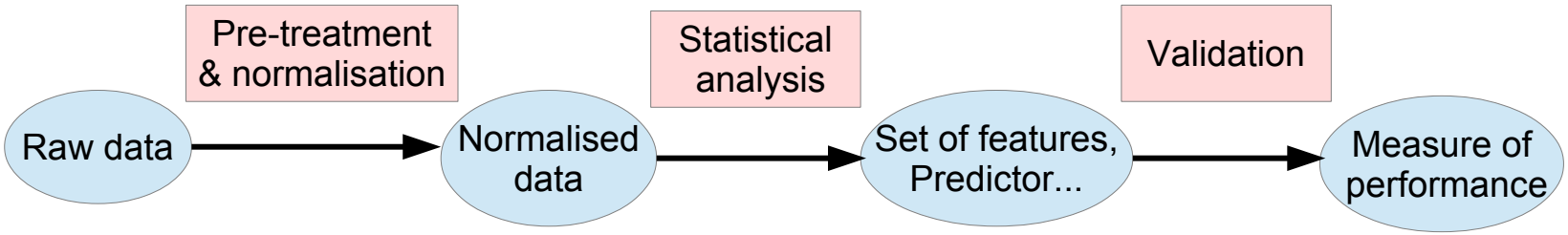
- Large dimension, noisy data

Statistical analysis of untargeted omics data

- Large dimension, noisy data
- Huge number of methods/variants to address the same question (feature selection, prediction...)
  - Results strongly impacted by the method
  - Issue : comparison of methods

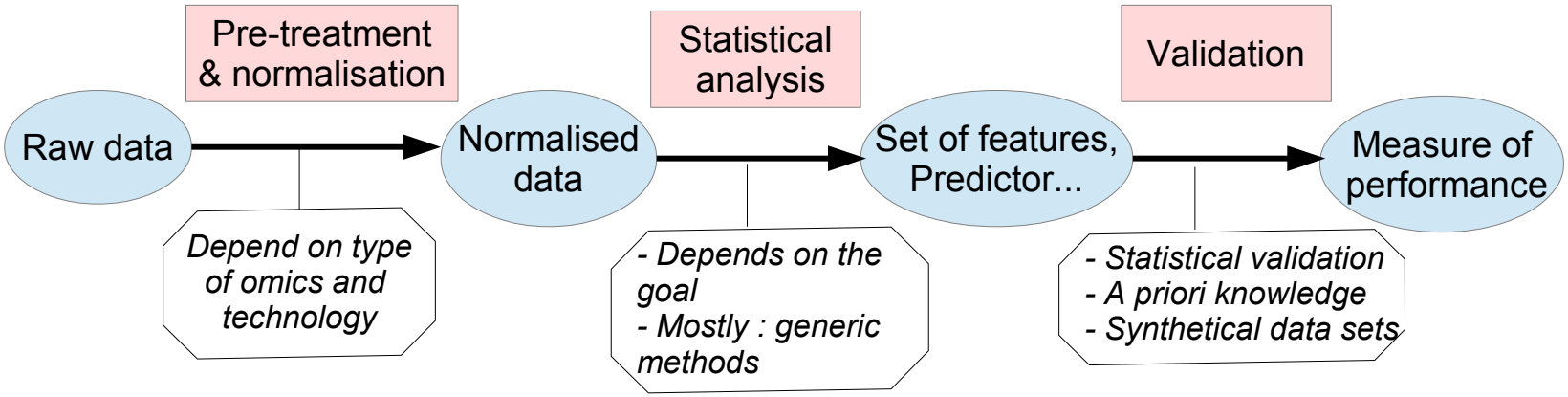
Statistical analysis of untargeted omics data

- Large dimension, noisy data
- Huge number of methods/variants to address the same question (feature selection, prediction...)
  - Results strongly impacted by the method
  - Issue : comparison of methods
  - Workflow



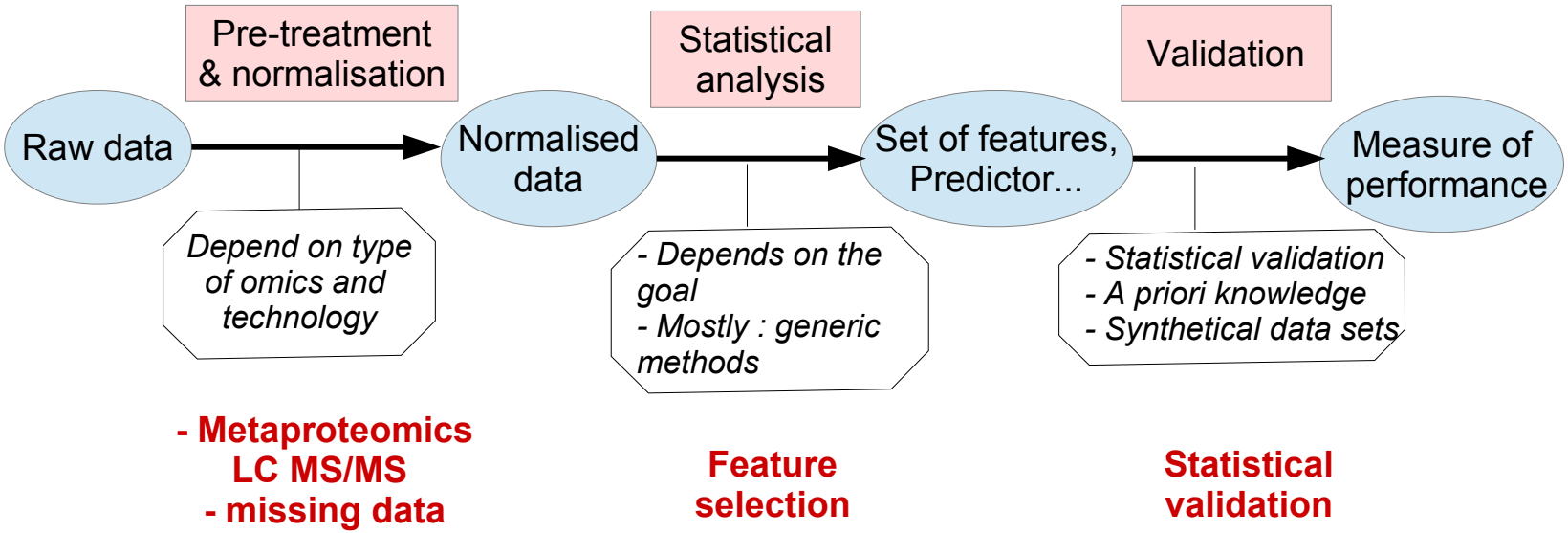
Statistical analysis of untargeted omics data

- Large dimension, noisy data
- Huge number of methods/variants to address the same question (feature selection, prediction...)
  - Results strongly impacted by the method
  - Issue : comparison of methods
  - Workflow



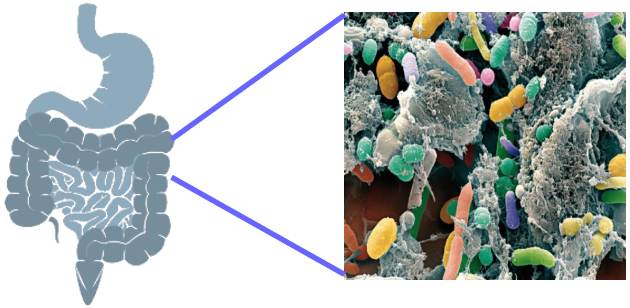
Statistical analysis of untargeted omics data

- Large dimension, noisy data
- Huge number of methods/variants to address the same question (feature selection, prediction...)
  - Results strongly impacted by the method
  - Issue : comparison of methods
  - Workflow



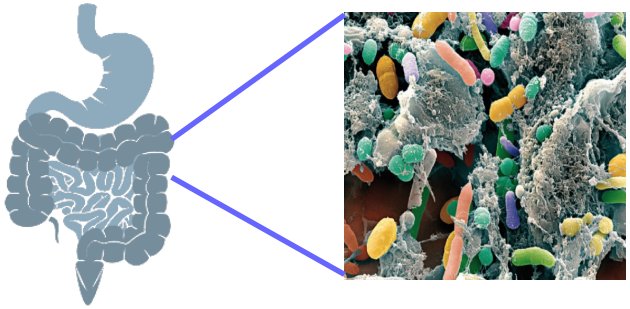
Metaproteomics

- Gut microbiote = complex bacterial ecosytem
  - Highly specific of the individual



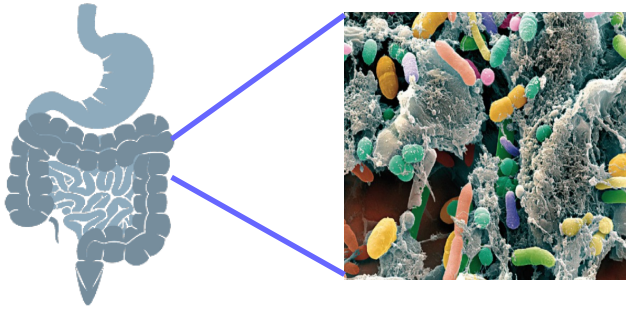
Metaproteomics

- Gut microbiote = complex bacterial ecosytem
  - Highly specific of the individual
- Metagenomics : measurements of all genes in a sample
  - Genetic potential

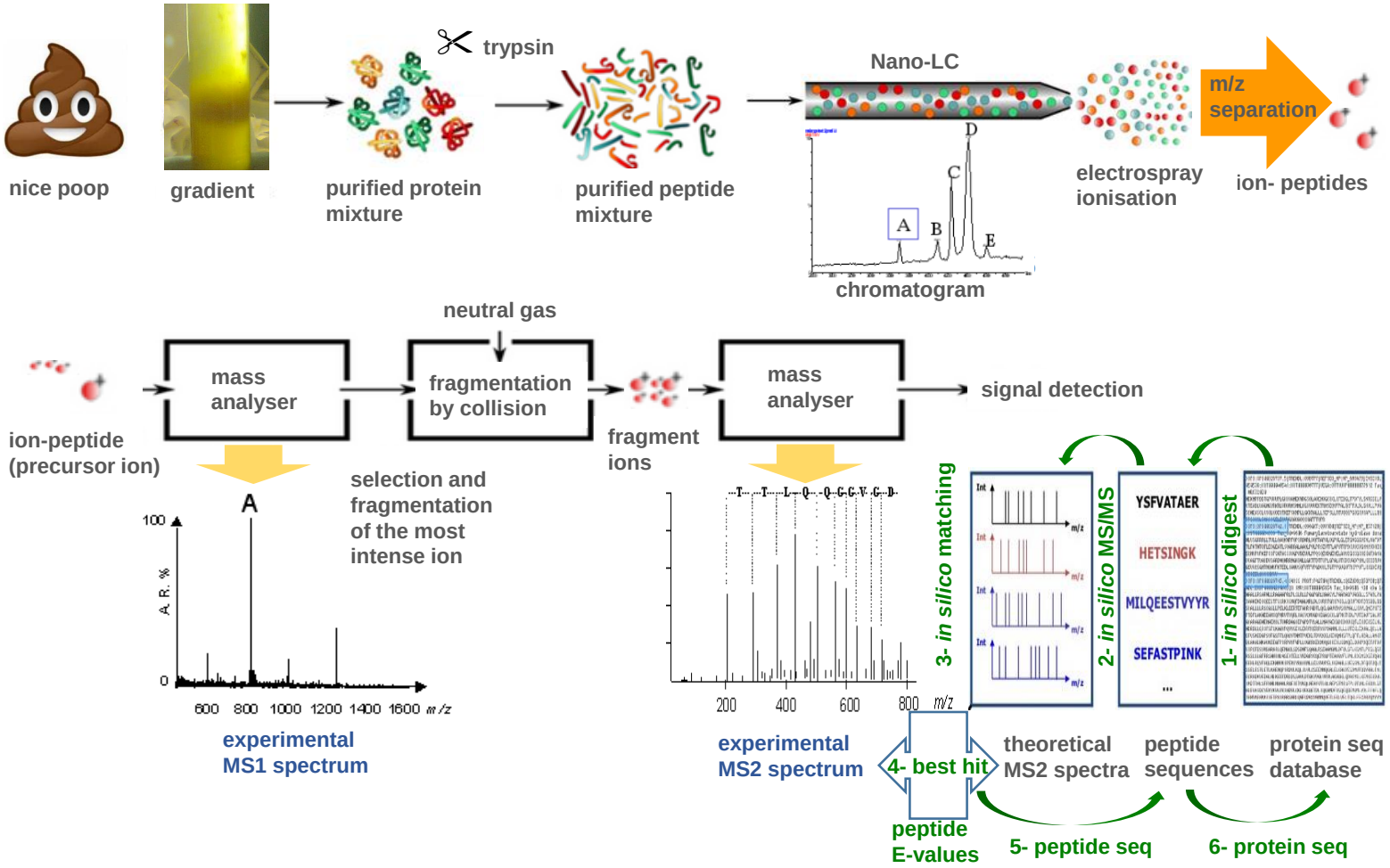


Metaproteomics

- Gut microbiote = complex bacterial ecosytem
  - Highly specific of the individual
- Metagenomics : measurements of all genes in a sample
  - Genetic potential
- Metaproteomics : measurements of all proteins in a sample
  - proteins actually expressed = functions realised in the gut



PROTEIN IDENTIFICATION USING MS/MS DATA



DATA ACQUISITION

DATA INTERPRETATION

A. Bassignani (2019)

Complex sources of technical variability

Factors that impact peptide measured intensity

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)
- Biochemical properties of the peptide
  - separation by retention time

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)
- Biochemical properties of the peptide
  - separation by retention time
- Abundance of other peptides with similar retention time
  - limit of resolution of the mass spectrometer

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)
- Biochemical properties of the peptide
  - separation by retention time
- Abundance of other peptides with similar retention time
  - limit of resolution of the mass spectrometer
- Global composition of the sample
  - Alignment of retention time based on peptides identified in other samples

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)
- Biochemical properties of the peptide
  - separation by retention time
- Abundance of other peptides with similar retention time
  - limit of resolution of the mass spectrometer
- Global composition of the sample
  - Alignment of retention time based on peptides identified in other samples
- Presence of similar peptides in the database

Factors that impact peptide measured intensity

- Peptide abundance (fortunately!)
- Biochemical properties of the peptide
  - separation by retention time
- Abundance of other peptides with similar retention time
  - limit of resolution of the mass spectrometer
- Global composition of the sample
  - Alignment of retention time based on peptides identified in other samples
- Presence of similar peptides in the database

These mechanisms lead to missing values

Classification of types of missingness

Missing values can be classified in two categories (Rubin, 1976)

Classification of types of missingness

Missing values can be classified in two categories (Rubin, 1976)

- Missing At Random (MAR) : missingness weakly/not related to the true feature concentration itself, but potentially related to the other feature concentration.

Classification of types of missingness

Missing values can be classified in two categories (Rubin, 1976)

- Missing At Random (MAR) : missingness weakly/not related to the true feature concentration itself, but potentially related to the other feature concentration.
- Missing Not At Random (MNAR) : missingness due to feature concentration close to the limit of detection of the device.

Classification of types of missingness

Missing values can be classified in two categories (Rubin, 1976)

- Missing At Random (MAR) : missingness weakly/not related to the true feature concentration itself, but potentially related to the other feature concentration.
- Missing Not At Random (MNAR) : missingness due to feature concentration close to the limit of detection of the device.

In practise : impossible to distinguish between these types of missingness

Handling missing data in (meta)-proteomics in literature

- One paper: model sources of technical variability in proteomics [O'Brien 2018]
  - (Gaussian) mixed model: not adapted to metaproteomics

Handling missing data in (meta)-proteomics in literature

- One paper: model sources of technical variability in proteomics [O'Brien 2018]
  - (Gaussian) mixed model: not adapted to metaproteomics
- Proteomics and metaproteomics: imputation of missing values
  - Replace NA by a single value (e.g. smallest observed intensity)
  - Local structure imputation (e.g. K Nearest Neighbors)
    - ★ Missing values infered based on the k most similar samples
  - Global structure imputation (e.g. Singular Value Decomposition)
    - ★ Based on structure of dependence between all peptides/proteins
    - ★ Not adapted to individual specificity of microbiote

Handling missing data in (meta)-proteomics in literature

- One paper: model sources of technical variability in proteomics [O'Brien 2018]
  - (Gaussian) mixed model: not adapted to metaproteomics
- Proteomics and metaproteomics: imputation of missing values
  - Replace NA by a single value (e.g. smallest observed intensity)
  - Local structure imputation (e.g. K Nearest Neighbors)
    - ★ Missing values infered based on the k most similar samples
  - Global structure imputation (e.g. Singular Value Decomposition)
    - ★ Based on structure of dependence between all peptides/proteins
    - ★ Not adapted to individual specificity of microbiote

Handling missing data in (meta)-proteomics in literature

- One paper: model sources of technical variability in proteomics [O'Brien 2018]
  - (Gaussian) mixed model: not adapted to metaproteomics
- Proteomics and metaproteomics: imputation of missing values
  - Replace NA by a single value (e.g. smallest observed intensity) [MNAR]
  - Local structure imputation (e.g. K Nearest Neighbors) [MAR]
    - ★ Missing values infered based on the k most similar samples
  - Global structure imputation (e.g. Singular Value Decomposition)
    - ★ Based on structure of dependence between all peptides/proteins
    - ★ Not adapted to individual specificity of microbiote

Handling missing data in (meta)-proteomics in literature

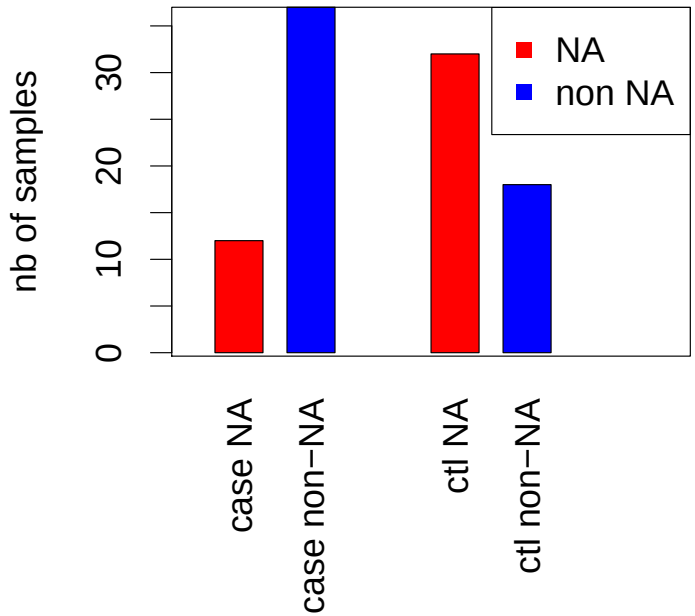
- One paper: model sources of technical variability in proteomics [O'Brien 2018]
  - (Gaussian) mixed model: not adapted to metaproteomics
- Proteomics and metaproteomics: imputation of missing values
  - Replace NA by a single value (e.g. smallest observed intensity) [MNAR]
  - Local structure imputation (e.g. K Nearest Neighbors) [MAR]
    - ⋆ Missing values infered based on the k most similar samples
  - Global structure imputation (e.g. Singular Value Decomposition)
    - ⋆ Based on structure of dependence between all peptides/proteins
    - ⋆ Not adapted to individual specificity of microbiote
- Statistical analyses : imputed and observed intensities treated equally
  - Problem with a large proportion of missing value

Alternative : combined test

- Statistical question: select features that differs between groups (e.g. case/ctl)

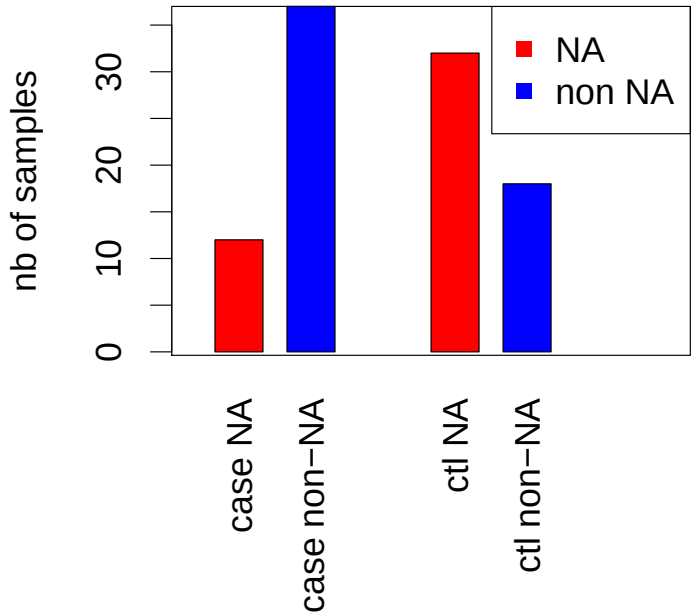
Alternative : combined test

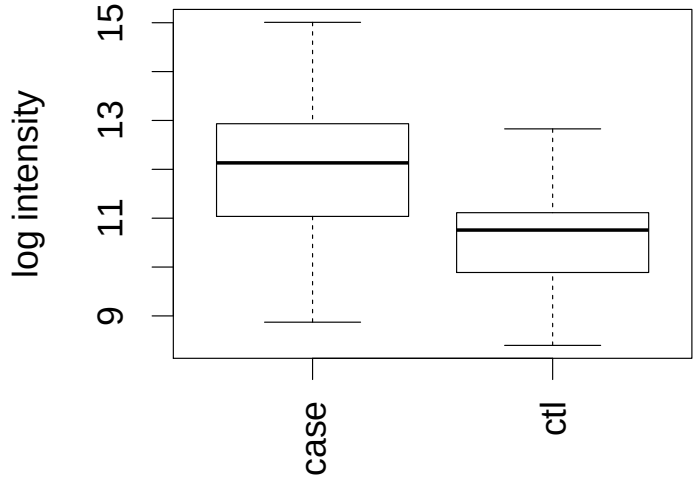
- Statistical question: select features that differs between groups (e.g. case/ctl)
- Test that targets two behaviours
  (a) Difference of missingness : protein is missing more frequently in one group than the other [MNAR]



Alternative : combined test

- Statistical question: select features that differs between groups (e.g. case/ctl)
- Test that targets two behaviours
  (a) Difference of missingness : protein is missing more frequently in one group than the other [MNAR]
  (b) Difference in intensities among non-missing value [MAR]





- Statistical procedure
  - Compute p-values p_a (Fisher exact test) and p_b (t-test)

- Statistical procedure
  - Compute p-values p_a (Fisher exact test) and p_b (t-test)
  - Combined test statistic $S = -(\log p_a + \log p_b)/2$
    - ★ Large if at least one p-value is small

- Statistical procedure
  - Compute p-values p_a (Fisher exact test) and p_b (t-test)
  - Combined test statistic $S = -(\log p_a + \log p_b)/2$
    - ★ Large if at least one p-value is small
  - Permutation test : repeated permutations of patient groups labels

- Statistical procedure
  - Compute p-values p_a (Fisher exact test) and p_b (t-test)
  - Combined test statistic $S = -(\log p_a + \log p_b)/2$
    - ★ Large if at least one p-value is small
  - Permutation test : repeated permutations of patient groups labels
    - ★ Technical detail : distribution under H_0 is assumed identical for all proteins with same proportion of NA

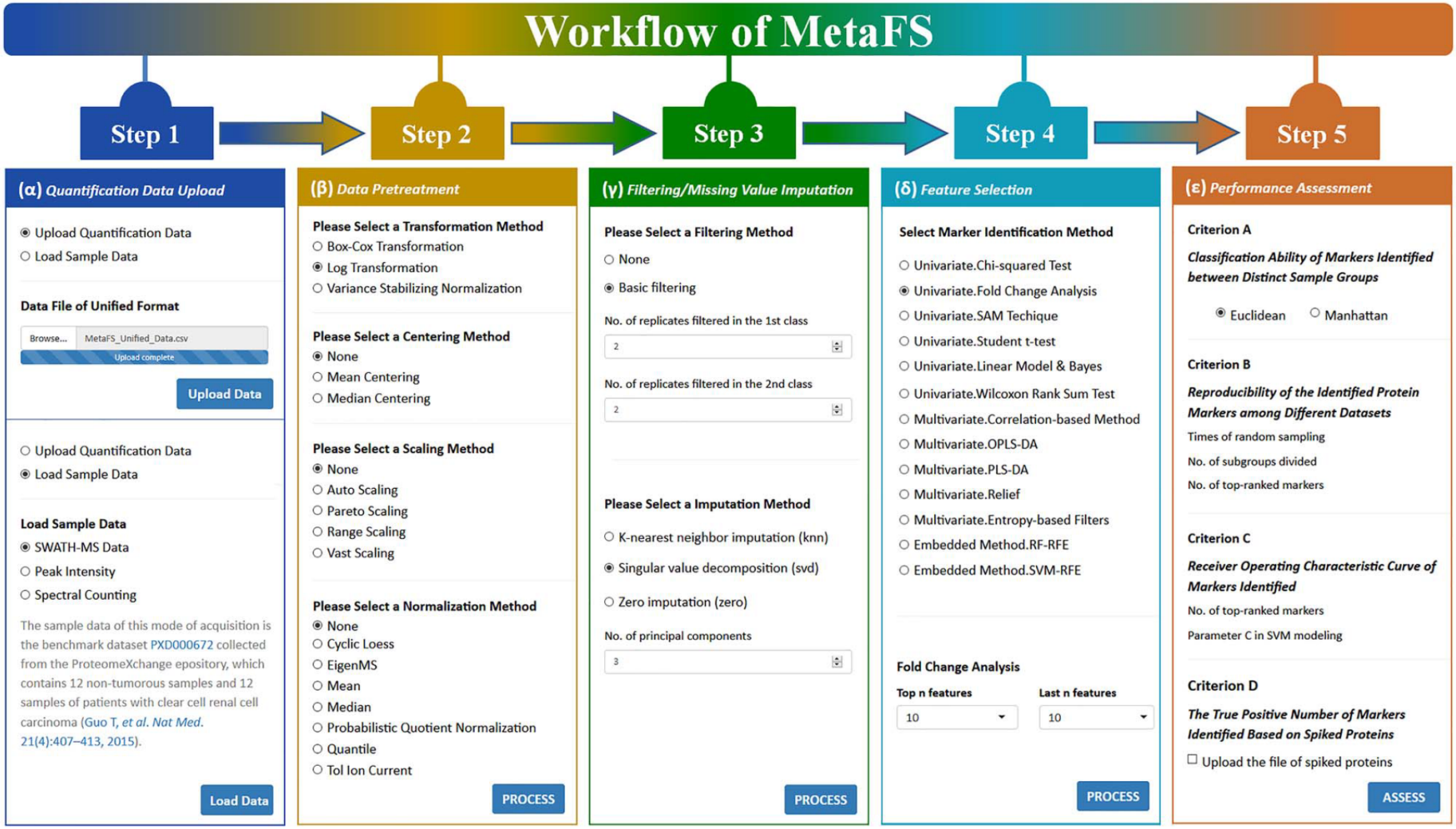
Comparison of methods to handle NA in feature selection

- Data sets from the ProteoCardis project
  - Metaproteome from 99 individuals in two groups (50 cases / 49 controls)
  - 8 biological samples with 7 technical replicates

Comparison of methods to handle NA in feature selection

- Data sets from the ProteoCardis project
  - Metaproteome from 99 individuals in two groups (50 cases / 49 controls)
  - 8 biological samples with 7 technical replicates
- Comparison of the following feature selection workflows
  (1) log-transformation + filter 20 non-NA + KNN + t-test [MAR]
  (2) log-transformation + filter 20 non-NA+ Single value imputation + t-test [MNAR]
  (3) log-transformation + filter 20 non-NA+ combined test [MAR + MNAR]

Comparison of feature selection methods in literature



Tang *et al* (2020), Briefings in Bioinformatics

Comments on criteria of performance assessment

Comments on criteria of performance assessment

- Ranking of methods depends on the criterion and the data set

Comments on criteria of performance assessment

- Ranking of methods depends on the criterion and the data set
- Criteria often relies on an arbitrary choice of statistical method

Comments on criteria of performance assessment

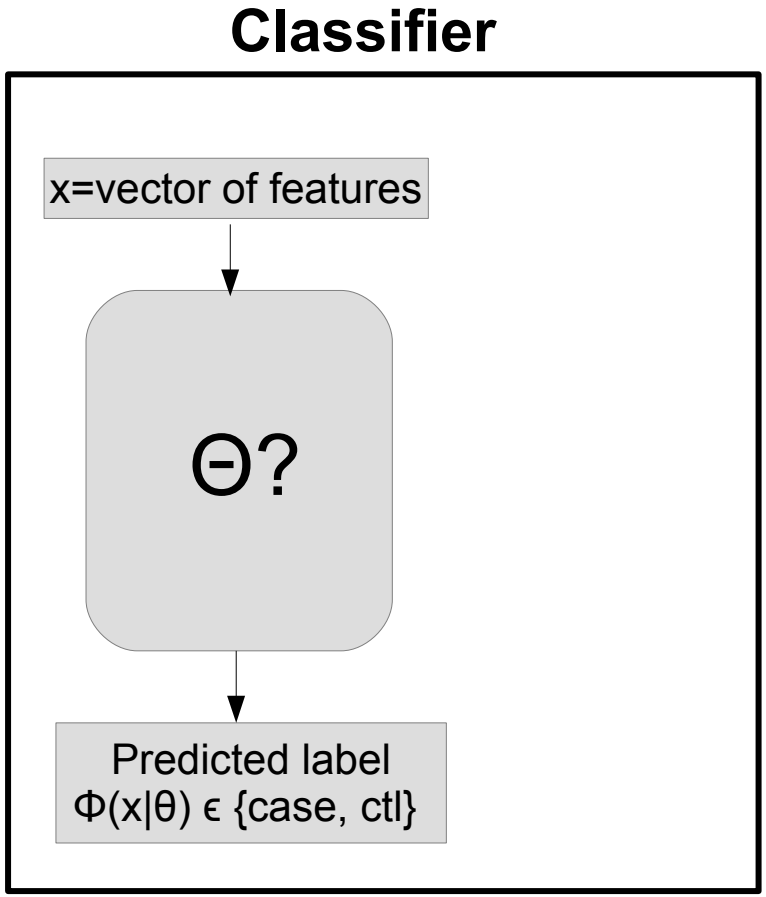
- Ranking of methods depends on the criterion and the data set
- Criteria often relies on an arbitrary choice of statistical method
- Caution : cross validation implementation is often biased

Comments on criteria of performance assessment

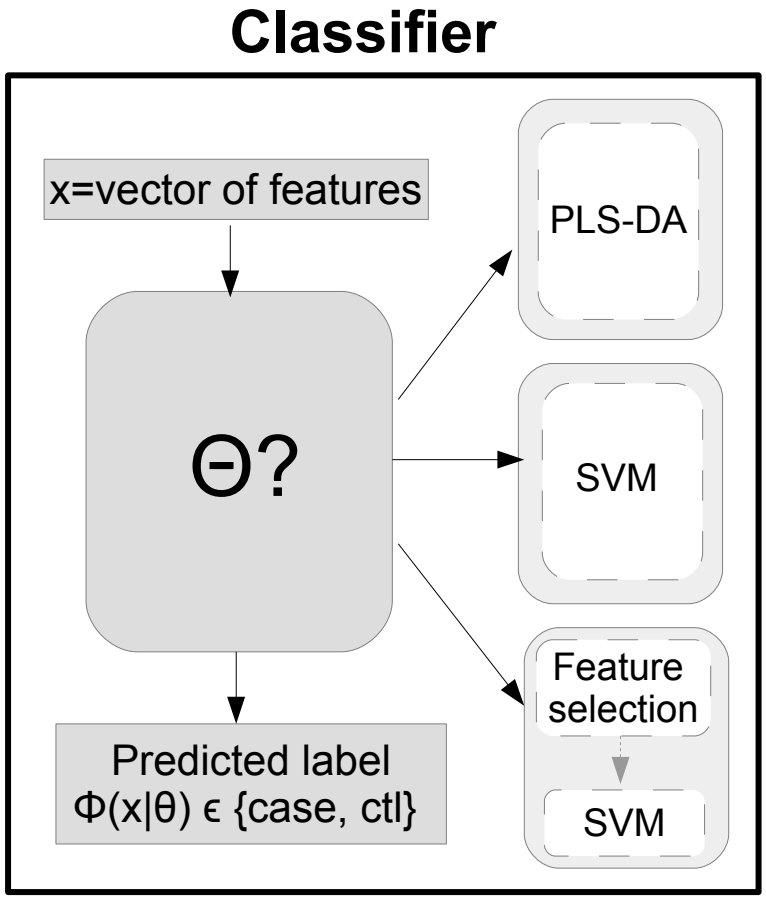
- Ranking of methods depends on the criterion and the data set
- Criteria often relies on an arbitrary choice of statistical method
- Caution : cross validation implementation is often biased

  $\hookrightarrow$ Not specific to this particular paper!

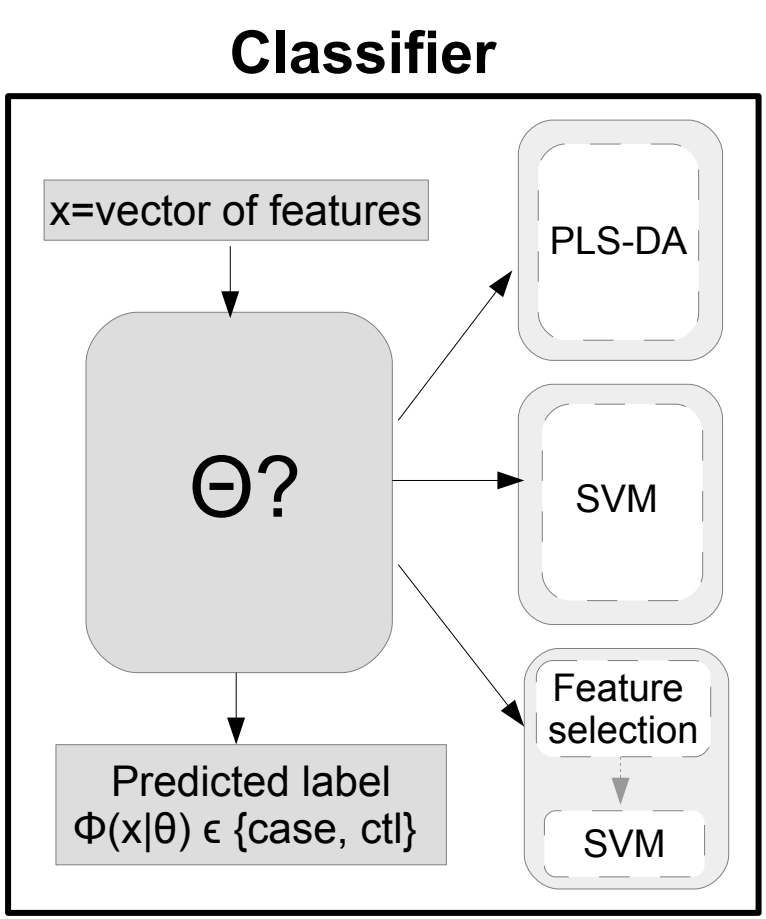
Focus on criterion of classification accuracy

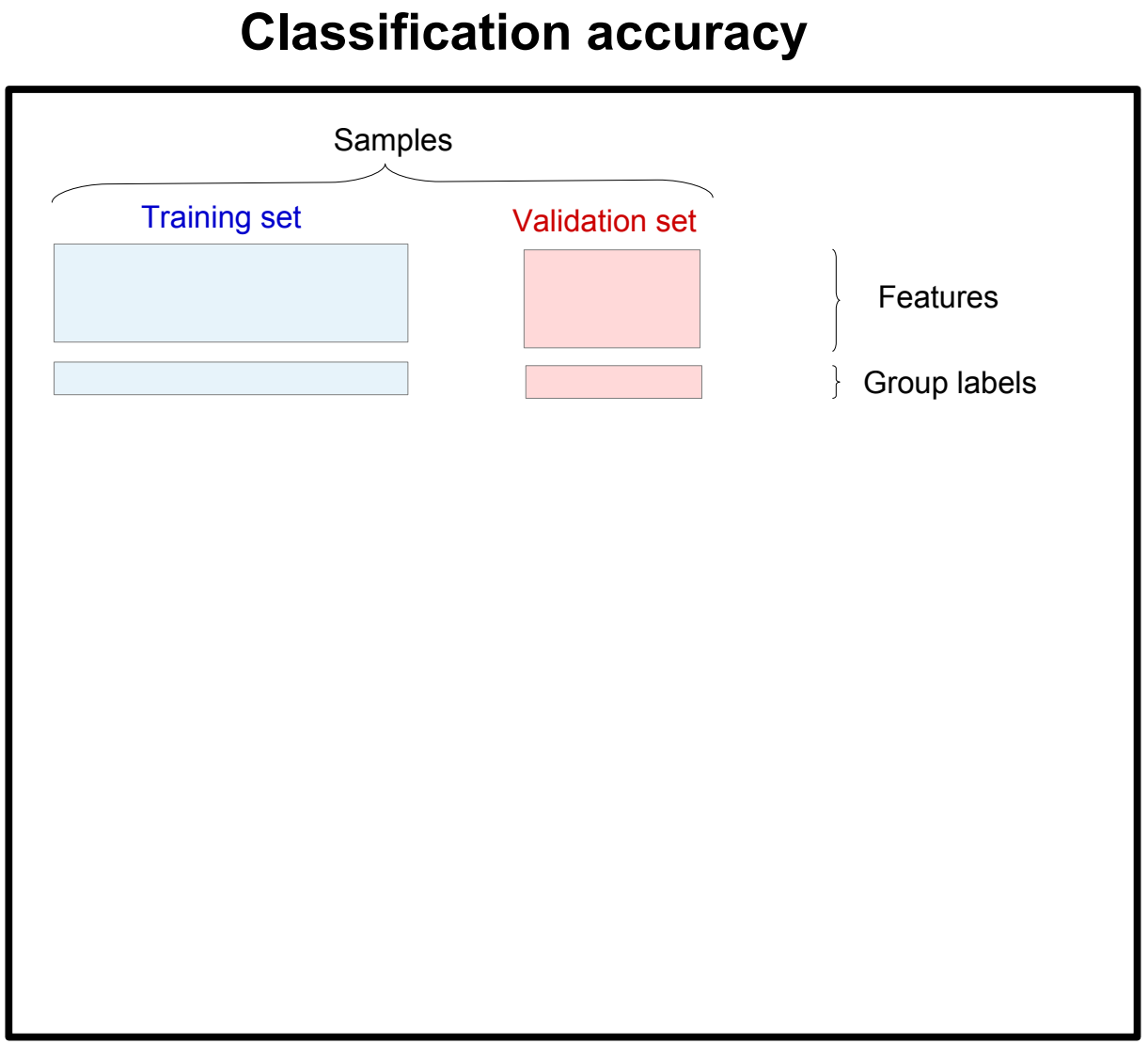


Focus on criterion of classification accuracy

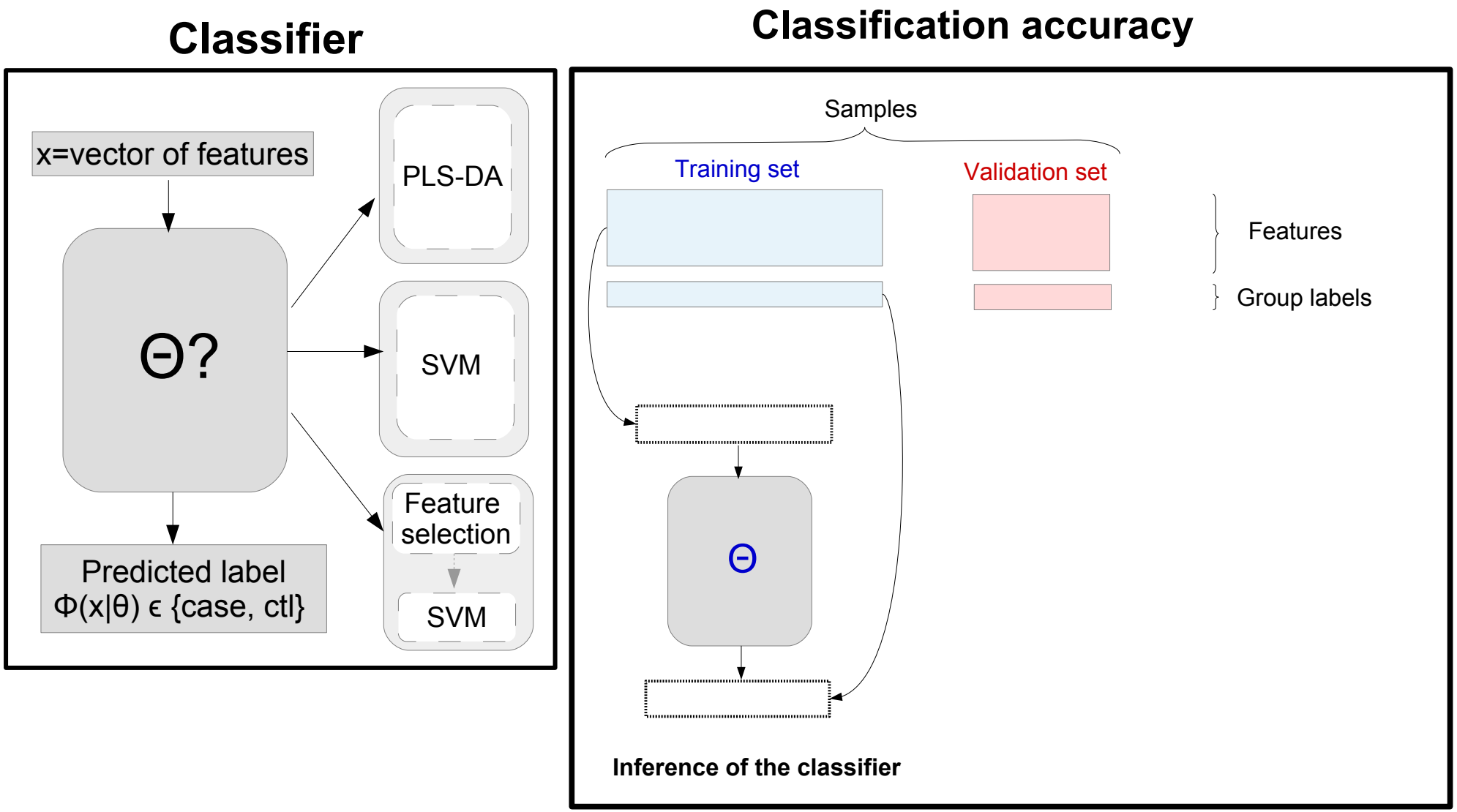


Focus on criterion of classification accuracy

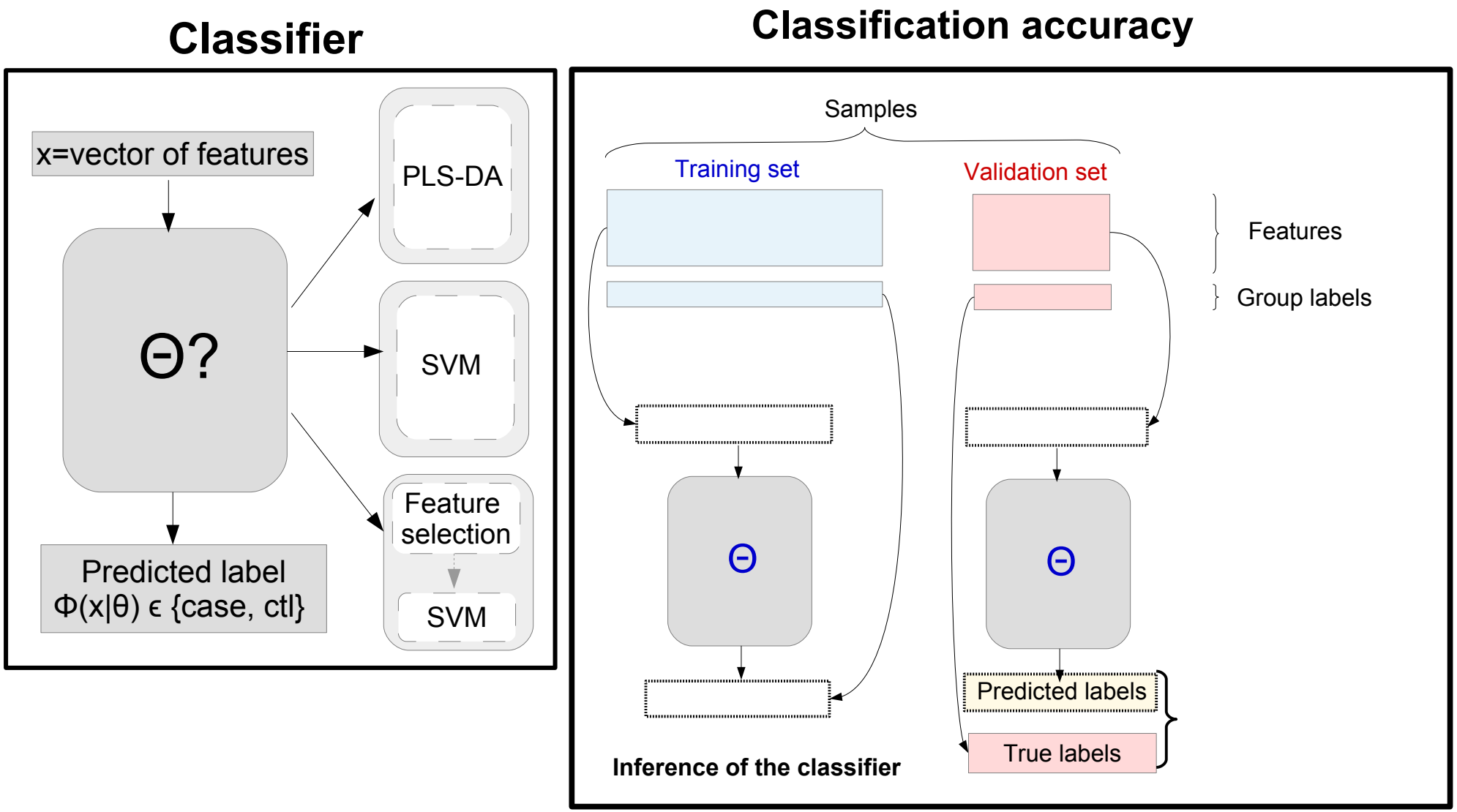




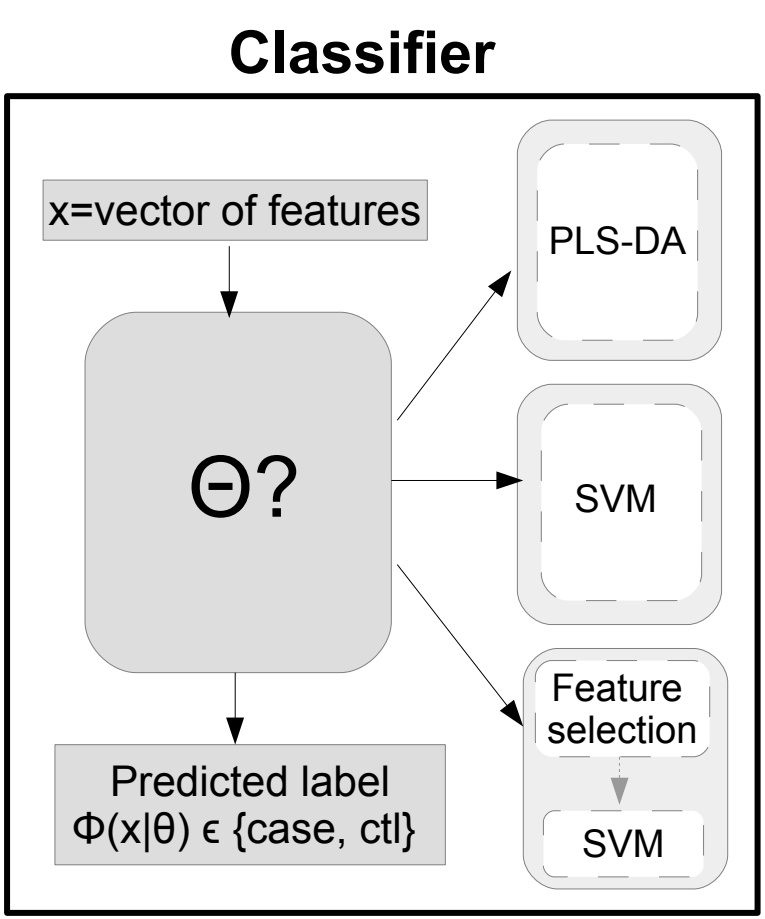
Focus on criterion of classification accuracy

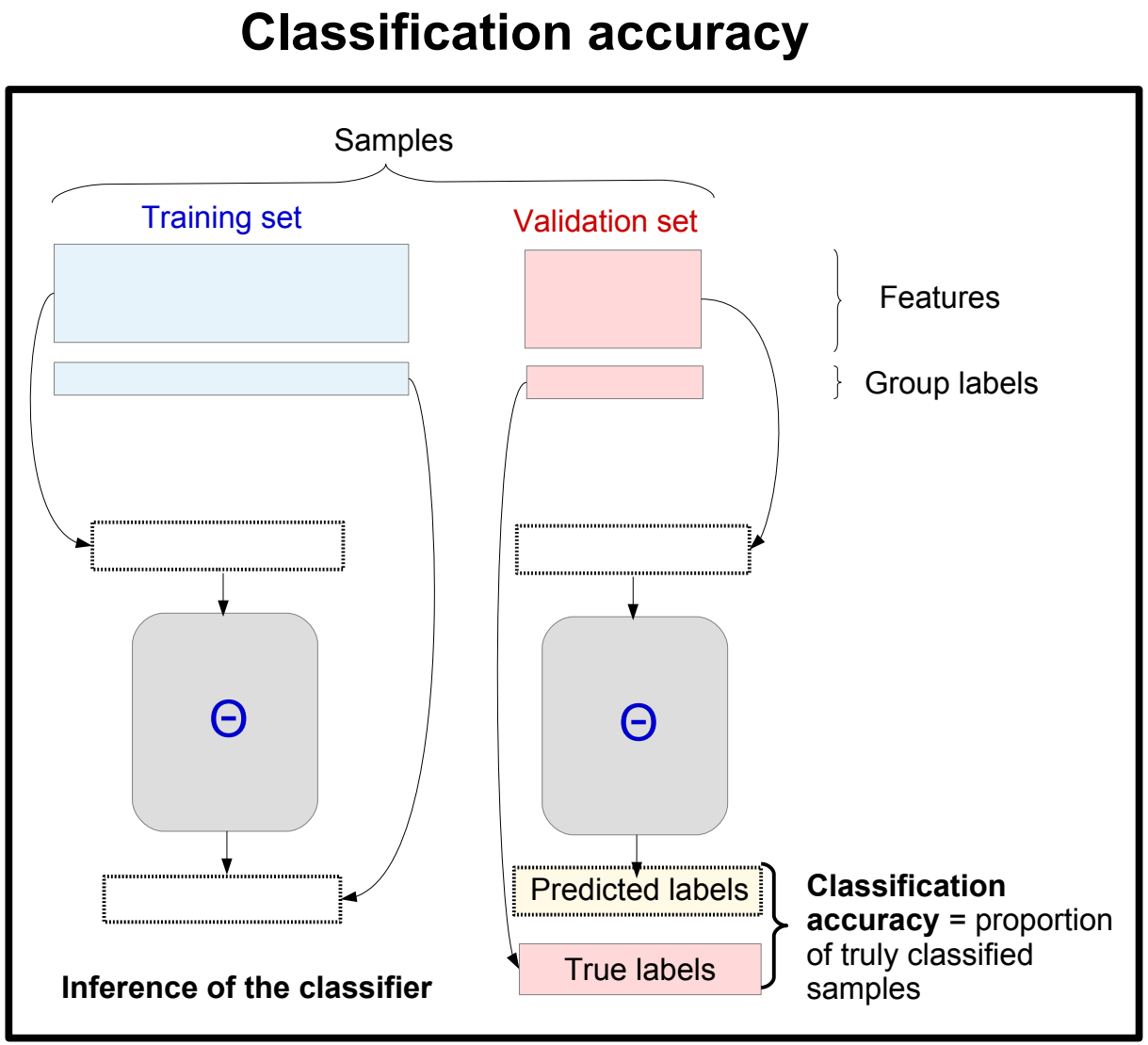


Focus on criterion of classification accuracy



Focus on criterion of classification accuracy





Illustration

14 FSMs (Feature Selection Methods) were assesed in this study, which included [...] (xii) support vector machine - recursive features eliminiation (SVM-FE). [...] Classification accuracy was used to judge the reliability of the selected biomarkers candidates. [...]. First, the discriminative proteins were identified and ranked using the FSMs. Then the top-ranked proteins (top 20, top 50, [...] top 450) were identified. Third SVM was applied to assess the performances of FSMs [...] using 5-fold cross validation.

Tang *et al* (2020), Briefings in Bioinformatics

- Bias in CV: both feature selection and inference of the classified should be performed on **the training data set.**

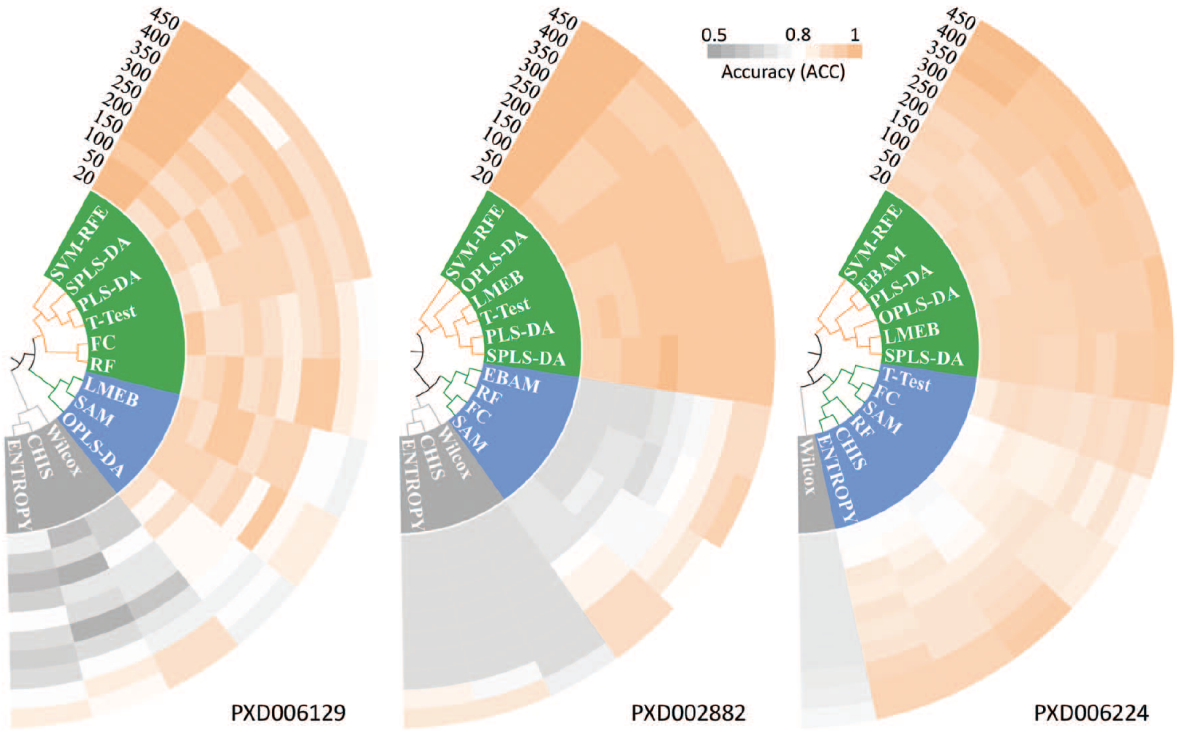
Illustration

14 FSMs (Feature Selection Methods) were assesed in this study, which included [...] (xii) support vector machine - recursive features eliminiation (SVM-RFE). [...] Classification accuracy was used to judge the reliability of the selected biomarkers candidates. [...]. First, the discriminative proteins were identified and ranked using the FSMs. Then the top-ranked proteins (top 20, top 50, [...] top 450) were identified. Third SVM was applied to assess the performances of FSMs [...] using 5-fold cross validation.

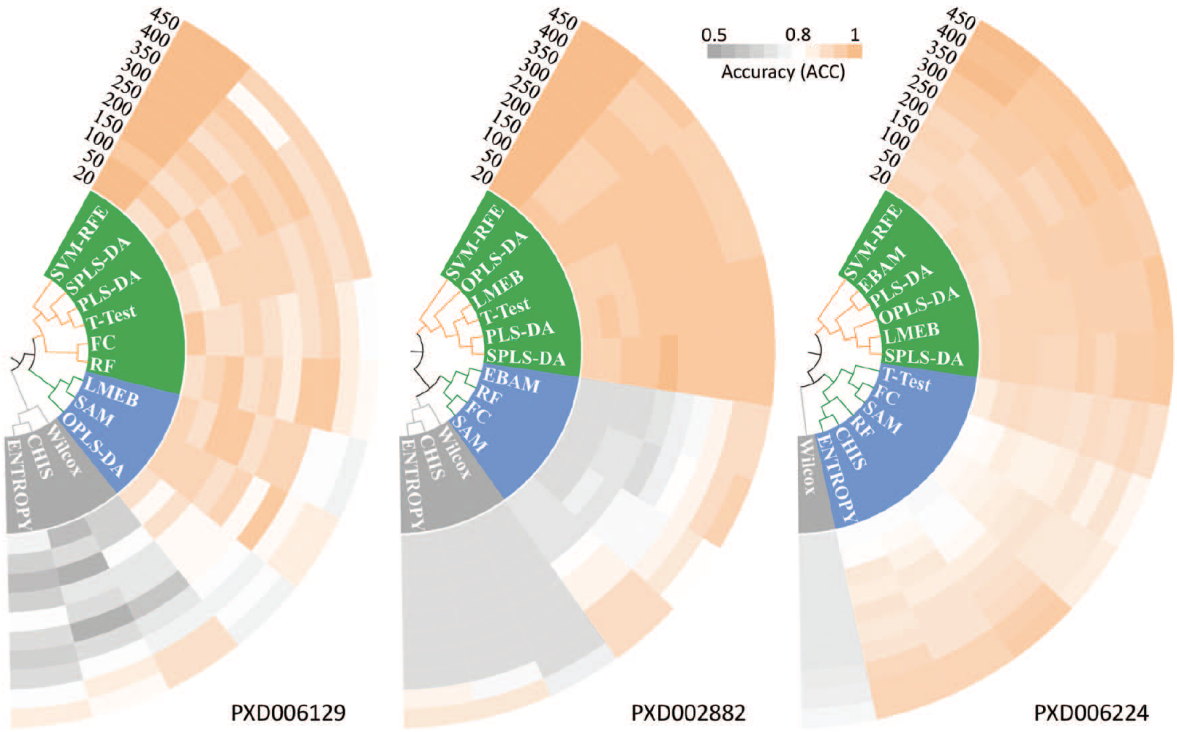
Tang *et al* (2020), Briefings in Bioinformatics

- Bias in CV: both feature selection and inference of the classified should be performed on **the training data set.**
- (Arbitrary) choice of classifier to compute prediction accuracy : SVM

Classification accuracy on 3 data sets

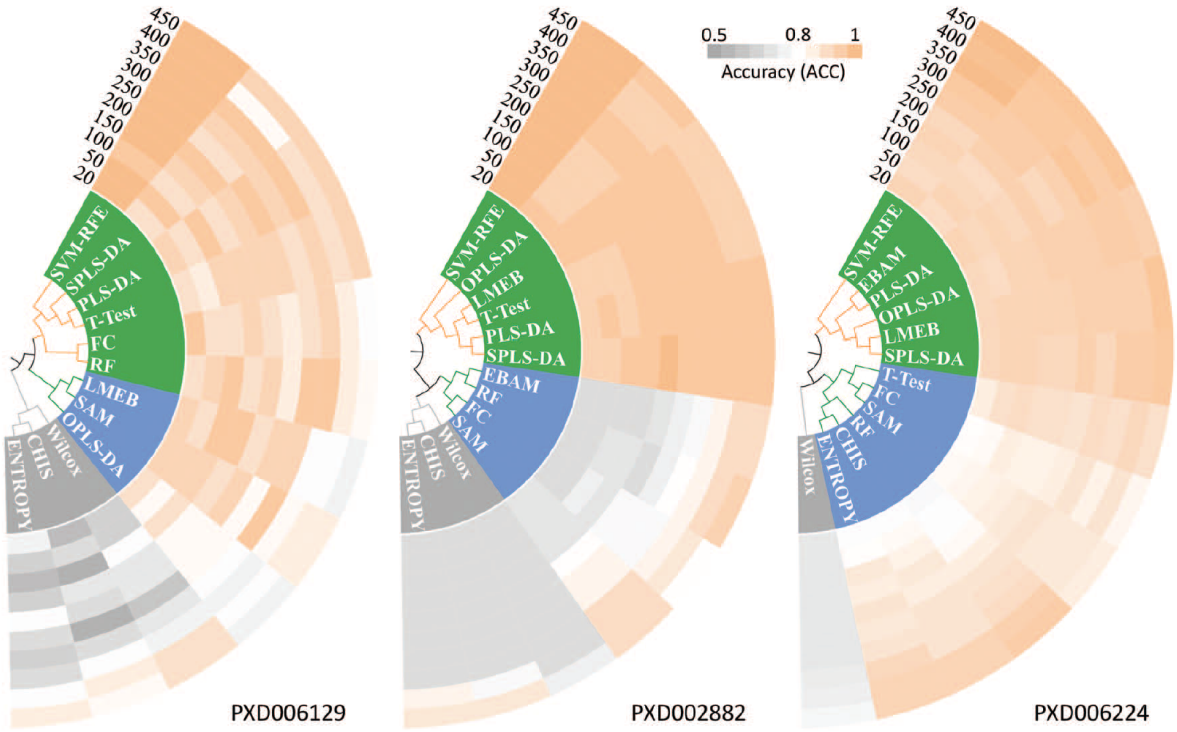


Classification accuracy on 3 data sets



- Best feature selection method : SVM-RFE
  - method based on the same classifier used for validation

Classification accuracy on 3 data sets



- Best feature selection method : SVM-RFE
  - method based on the same classifier used for validation
- Coincidence? I think not...

Selection of top 100 features + classification in a CV loop

(FSM1) log-transf. + filter 20 non-NA + KNN + t-test

(FSM2) log-transf. + filter 20 non-NA+ Single value imput. + t-test

(FSM3) log-transf. + filter 20 non-NA+ combined test

Selection of top 100 features + classification in a CV loop

(FSM1) log-transf. + filter 20 non-NA + KNN + t-test

(FSM2) log-transf. + filter 20 non-NA+ Single value imput. + t-test

(FSM3) log-transf. + filter 20 non-NA+ combined test

Classif with random forest

	Proteins		Specific peptides	
	NA = 0 for RF	KNN imput for RF	NA = 0 for RF	KNN imput for RF
FSM1	0.71	0.63	0.72	**0.72**
FSM2	0.67	0.58	**0.77**	0.64
FSM3	**0.72**	**0.68**	0.73	0.70

Classif with SVM

	Proteins		Specific peptides	
	NA = 0 for SVM	KNN imput for SVM	NA = 0 for SVM	KNN imput forSVM
FSM1	0.57	0.55	0.70	0.64
FSM2	0.55	**0.60**	0.70	**0.72**
FSM3	**0.59**	0.58	**0.72**	0.59

Selection of top 100 features + classification in a CV loop

(FSM1) log-transf. + filter 20 non-NA + KNN + t-test

(FSM2) log-transf. + filter 20 non-NA+ Single value imput. + t-test

(FSM3) log-transf. + filter 20 non-NA+ combined test

Classif with random forest

	Proteins		Specific peptides	
	NA = 0 for RF	KNN imput for RF	NA = 0 for RF	KNN imput for RF
FSM1	0.71	0.63	0.72	**0.72**
FSM2	0.67	0.58	**0.77**	0.64
FSM3	**0.72**	**0.68**	0.73	0.70

Classif with SVM

	Proteins		Specific peptides	
	NA = 0 for SVM	KNN imput for SVM	NA = 0 for SVM	KNN imput forSVM
FSM1	0.57	0.55	0.70	0.64
FSM2	0.55	**0.60**	0.70	**0.72**
FSM3	**0.59**	0.58	**0.72**	0.59

- Method ranking depends on data set and classifier
- Similar performances

MAR or MNAR? Evaluation on replicates

- 8 biological samples with 7 technical replicates.

MAR or MNAR? Evaluation on replicates

- 8 biological samples with 7 technical replicates.
- For each biological sample s and protein p : $x_{s,p}$= average intensity of non missing values and $n_{s,p}$=number of missing values.

NA	12	NA	NA	NA	16	10

$x_{s,p} = 12.67; \quad n_{s,p} = 3$

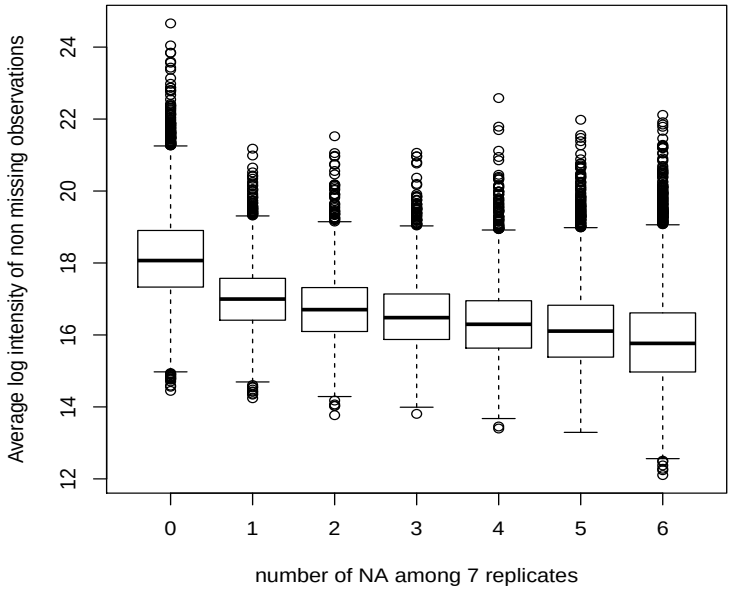
MAR or MNAR? Evaluation on replicates

- 8 biological samples with 7 technical replicates.
- For each biological sample s and protein p : $x_{s,p}$= average intensity of non missing values and $n_{s,p}$=number of missing values.

NA	12	NA	NA	NA	16	10

$x_{s,p} = 12.67; \quad n_{s,p} = 3$

- Boxplot of $\log(x_{s,p})$ as a function of $n_{s,p}$



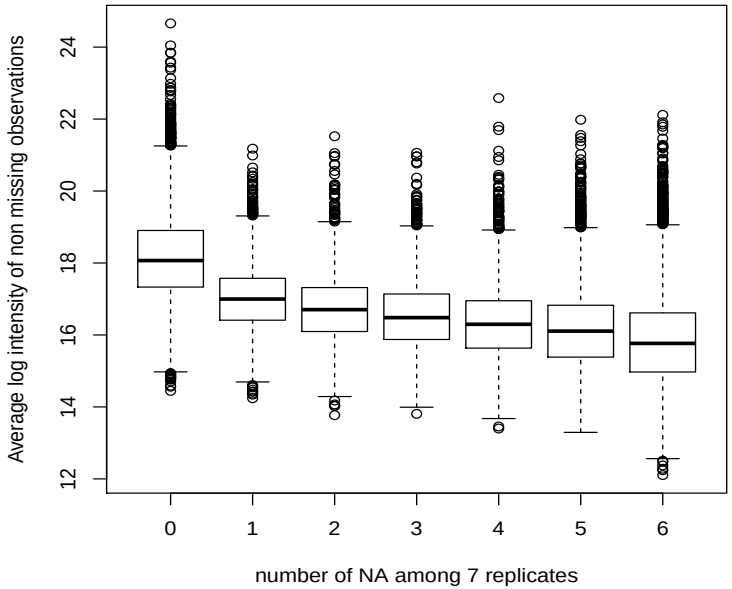
MAR or MNAR? Evaluation on replicates

- 8 biological samples with 7 technical replicates.
- For each biological sample s and protein p : $x_{s,p}$= average intensity of non missing values and $n_{s,p}$=number of missing values.

NA	12	NA	NA	NA	16	10

$x_{s,p} = 12.67; \quad n_{s,p} = 3$

- Boxplot of $\log(x_{s,p})$ as a function of $n_{s,p}$



- Observed intensity decreases when probability of missingness increases
- Even whith high proba of missingness, observed intensity can be high

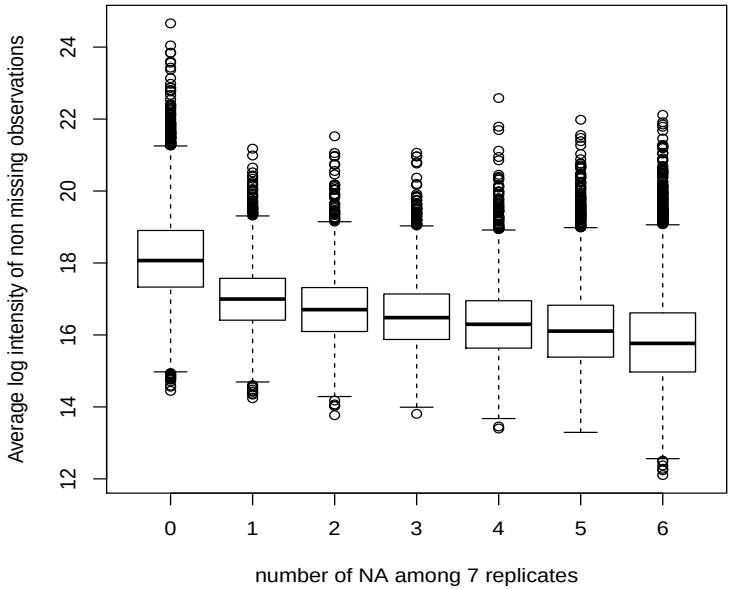
MAR or MNAR? Evaluation on replicates

- 8 biological samples with 7 technical replicates.
- For each biological sample s and protein p : $x_{s,p}$= average intensity of non missing values and $n_{s,p}$=number of missing values.

NA	12	NA	NA	NA	16	10

$x_{s,p} = 12.67; \quad n_{s,p} = 3$

- Boxplot of $\log(x_{s,p})$ as a function of $n_{s,p}$



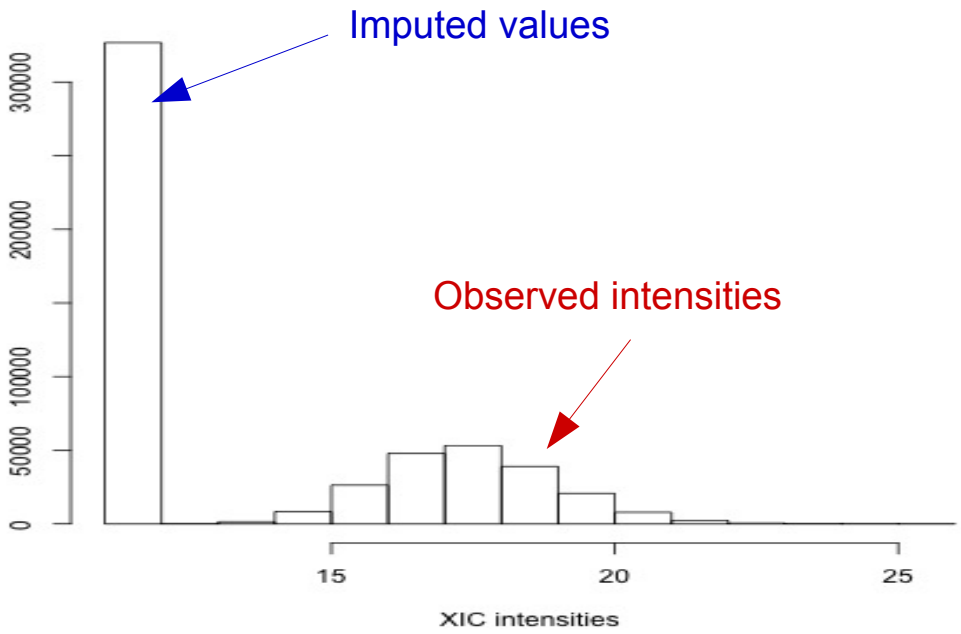
- Observed intensity decreases when probability of missingness increases
- Even whith high proba of missingness, observed intensity can be high

Both MAR and MNAR

Single value imputation

- Imputed values = smallest observed intensity

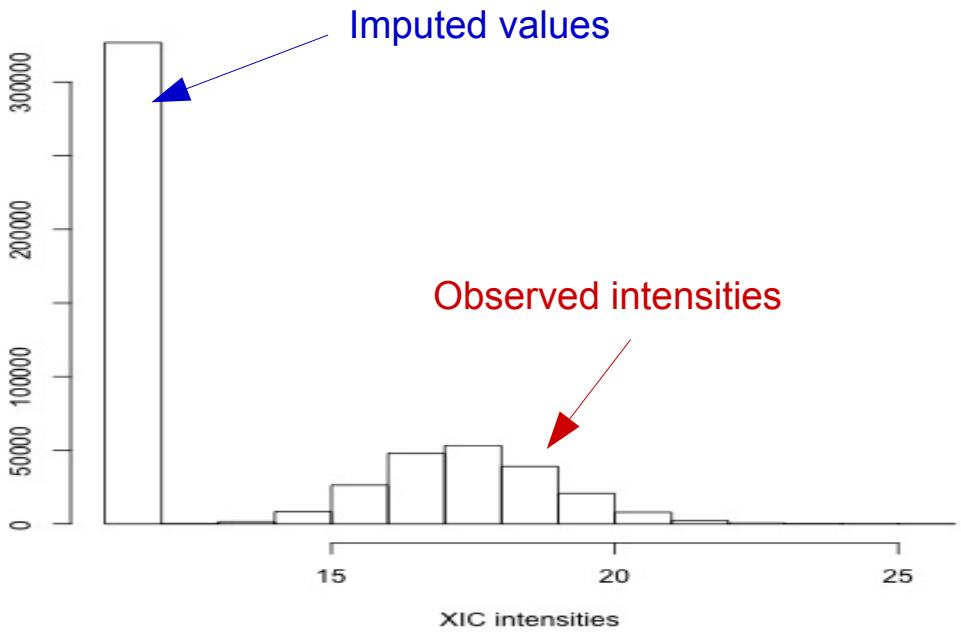
  Histogram of intensities after single value imputation



Single value imputation

- Imputed values = smallest observed intensity

  Histogram of intensities after single value imputation

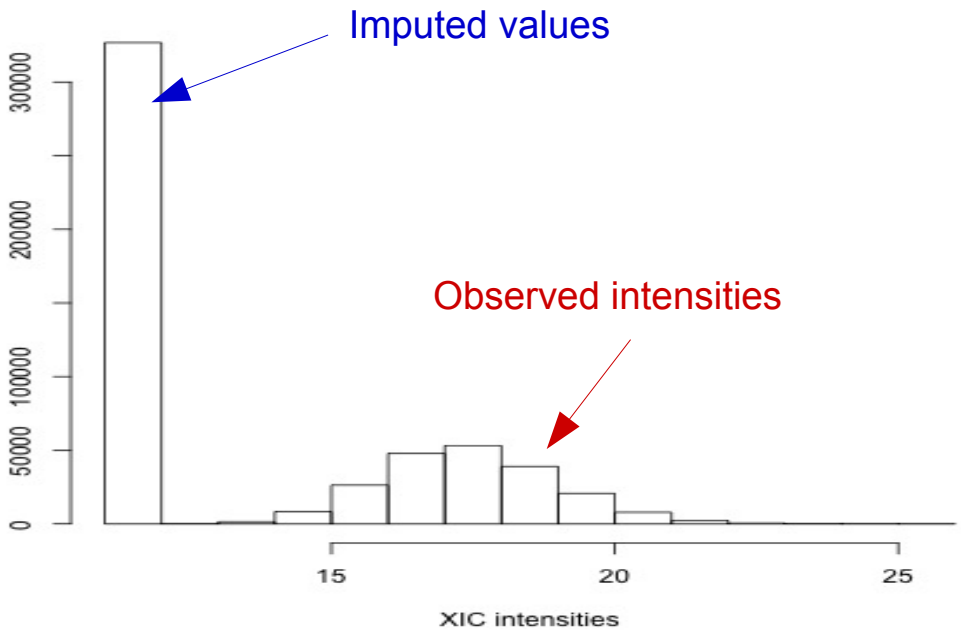


- Imputed value = too small

Single value imputation

- Imputed values = smallest observed intensity

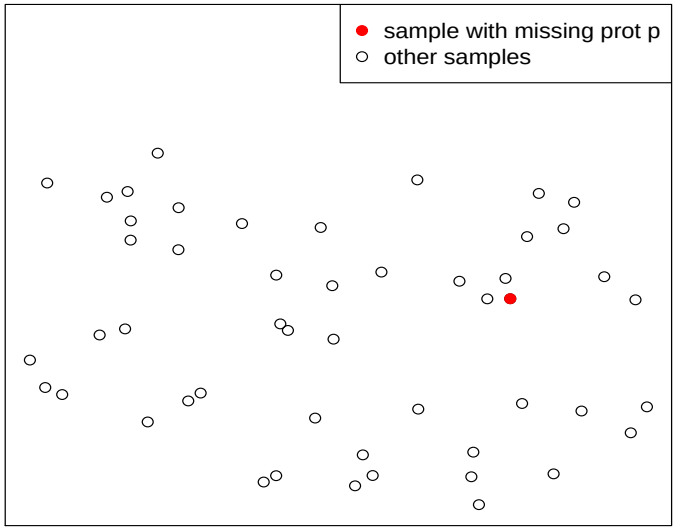
Histogram of intensities after single value imputation



- Imputed value = too small
- Rather a decreasing probability of detection than a threshold

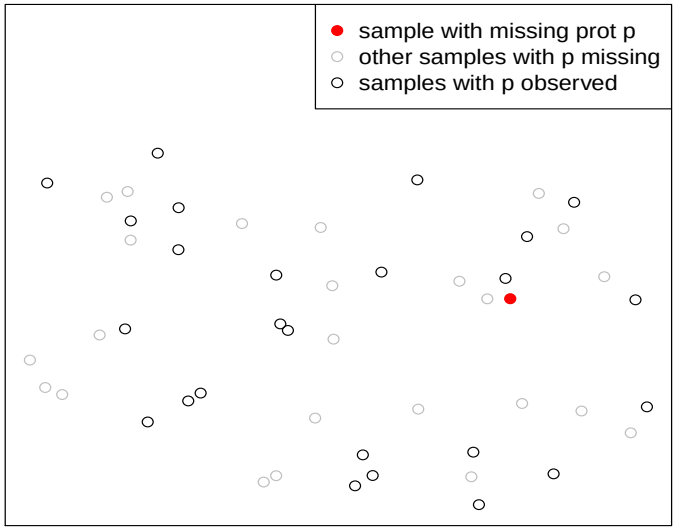
K Nearest Neighbors

- KNN imputation depends on
  - number of neighbors k
  - distance



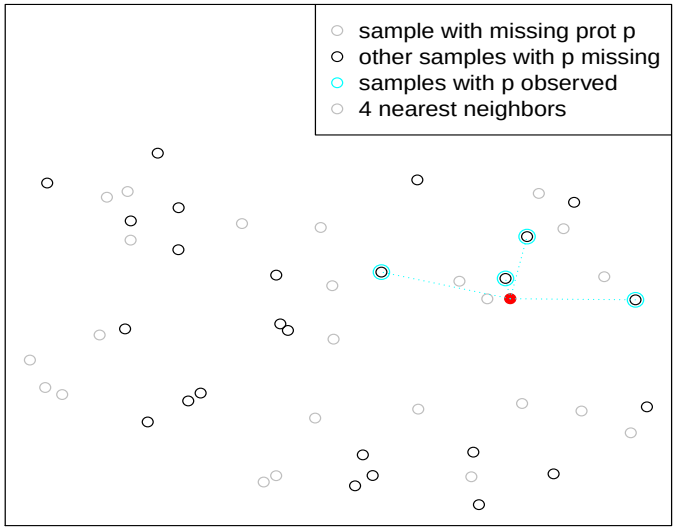
K Nearest Neighbors

- KNN imputation depends on
  - number of neighbors k
  - distance



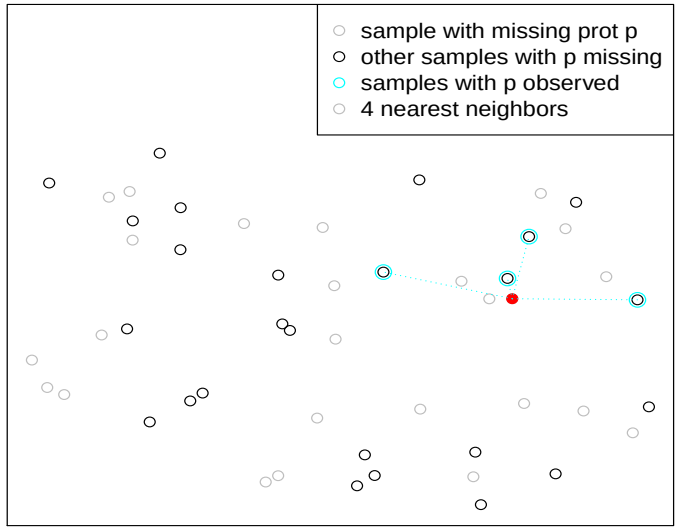
K Nearest Neighbors

- KNN imputation depends on
  - number of neighbors k
  - distance



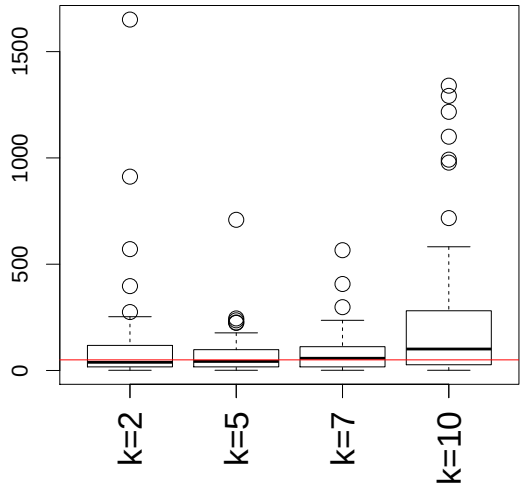
K Nearest Neighbors

- KNN imputation depends on
  - number of neighbors k
  - distance

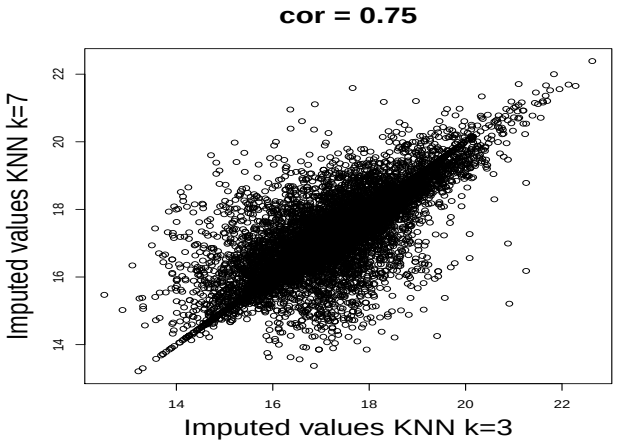


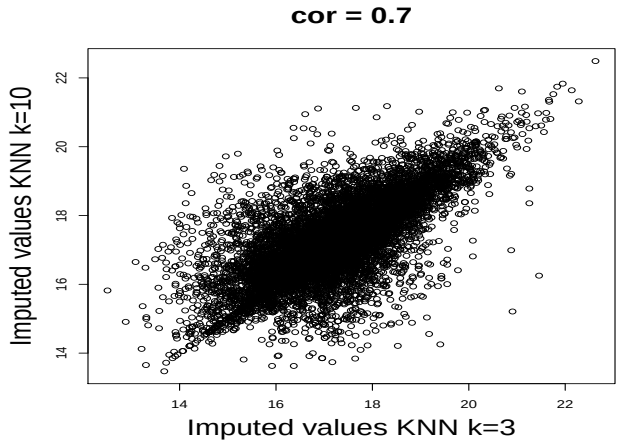
- Impact of k on feature selection

- KNN + t-test with $k_0 = 3$: rank variables
- KNN + t-test with $k \neq k_0$: extract top 50
- Look at order of the top 50 in the first ranking (with k_0)

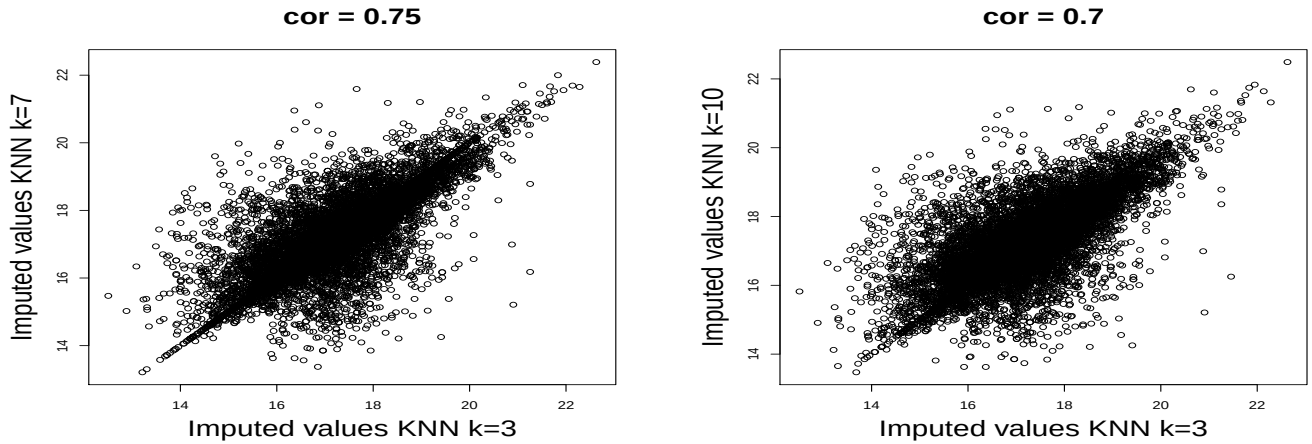


- Impact of k on imputed values





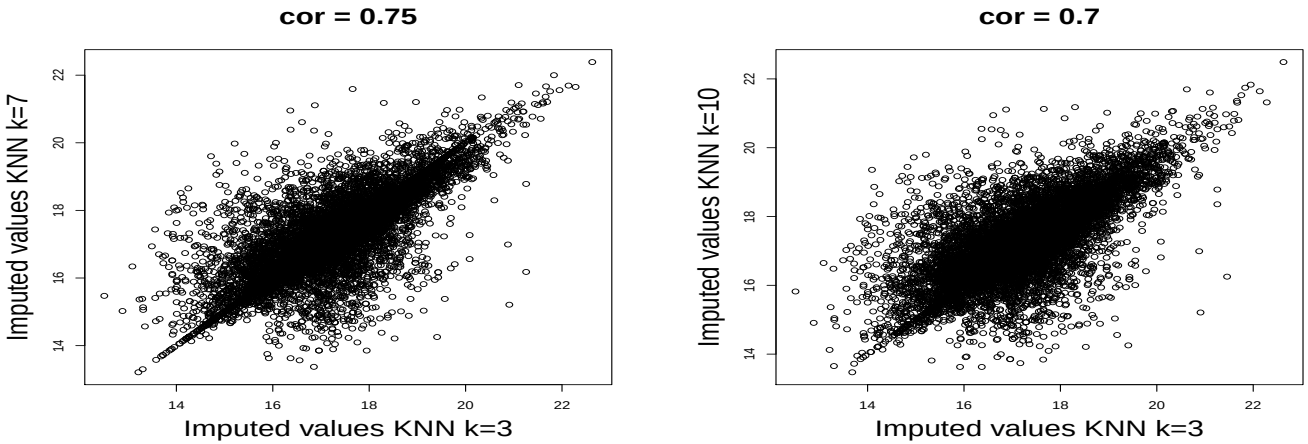
- Impact of k on imputed values



- Comparison with technical noise.
  - For each protein p and each replicated sample s,

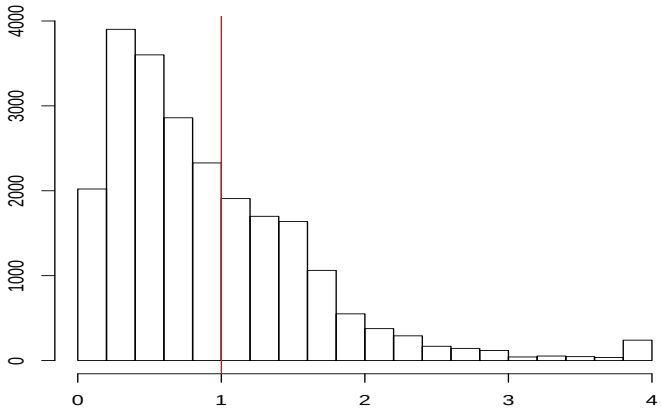
$$\text{mean of } |\log X_{srp} - \log X_{sr'p}| \begin{cases} \text{on pairs with both observed: } d_{sp}^{\text{obs}} \\ \text{on pairs with at least one imputed: } d_{sp}^{\text{imput}} \end{cases}$$

- Impact of k on imputed values



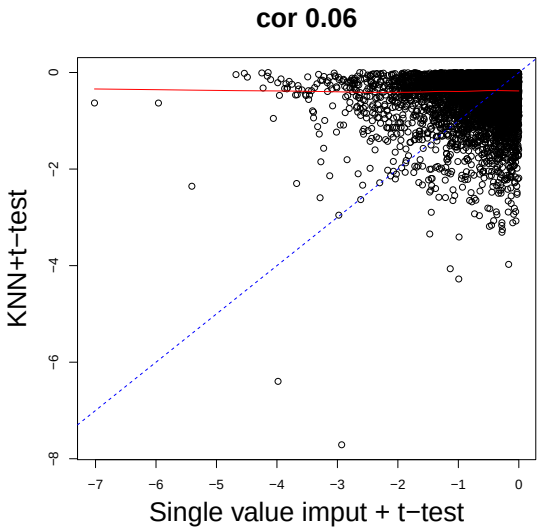
- Comparison with technical noise.
  - For each protein p and each replicated sample s,

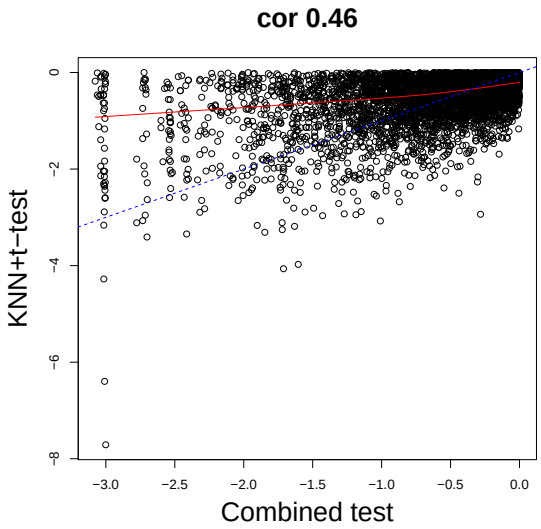
$$\text{mean of } |\log X_{srp} - \log X_{sr'p}| \begin{cases} \text{on pairs with both observed: } d_{sp}^{\text{obs}} \\ \text{on pairs with at least one imputed: } d_{sp}^{\text{imput}} \end{cases}$$

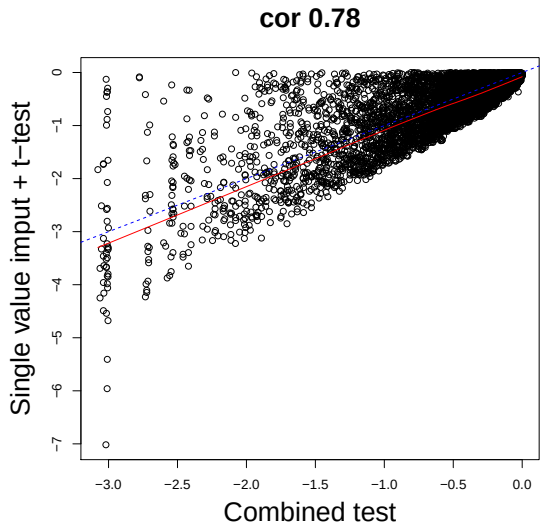


- Histogram of $d_{sp}^{\text{obs}} / d_{sp}^{\text{imput}}$
- Imputation increases moderately variability

log10-p-values for the 3 procedures

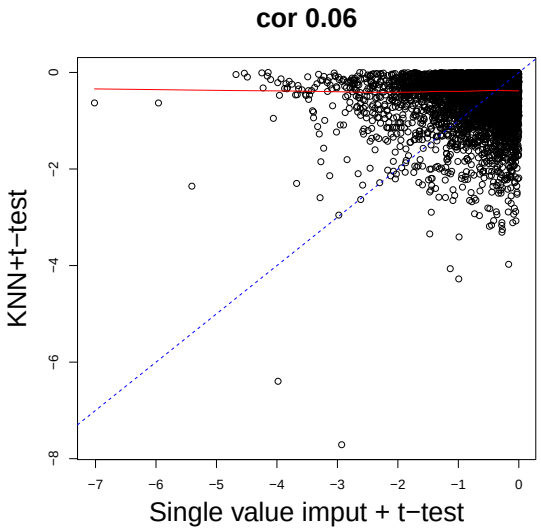


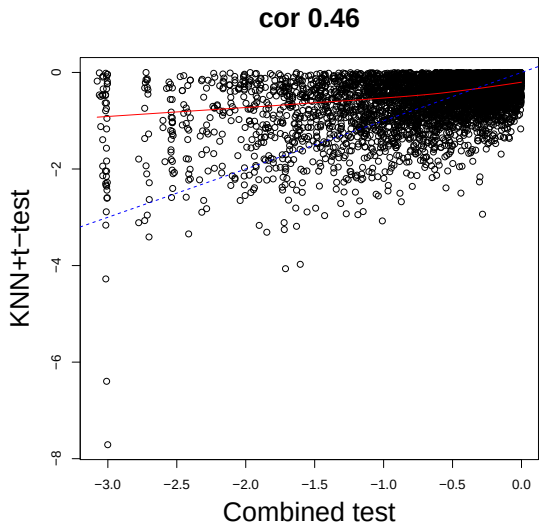


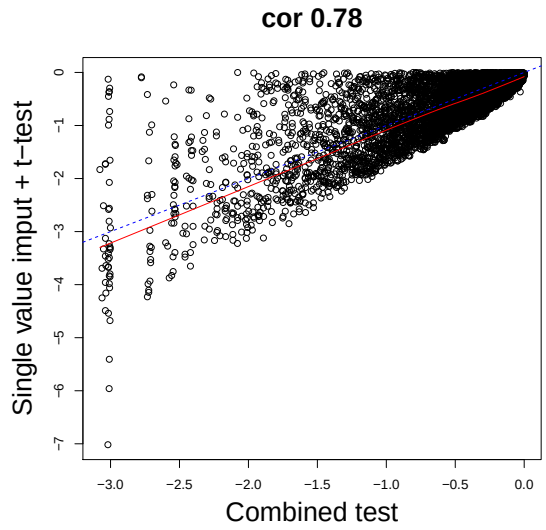


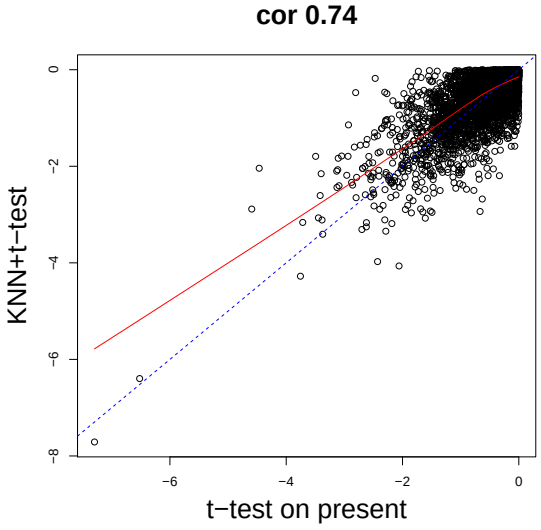
- KNN and single value: very different pv

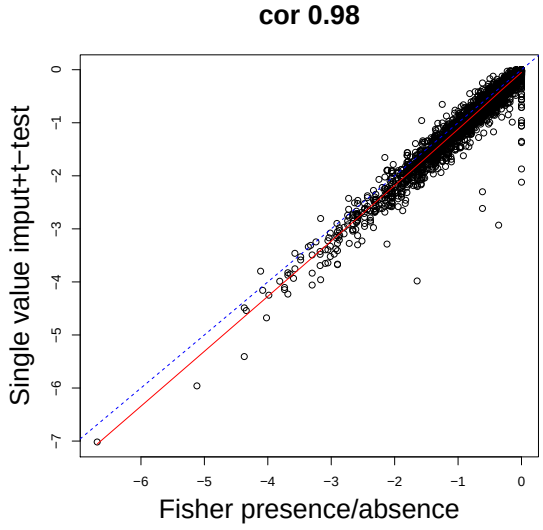
log10-p-values for the 3 procedures





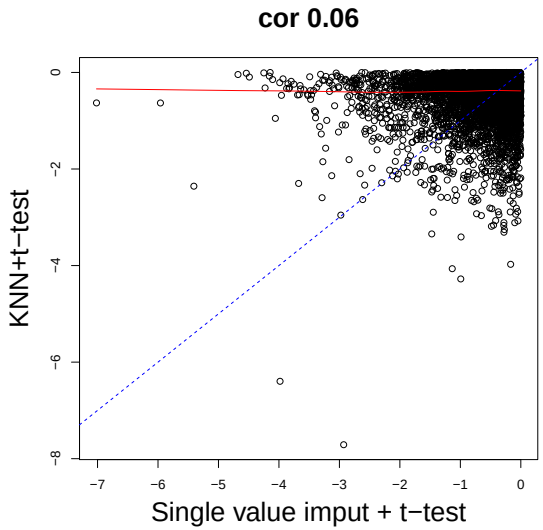


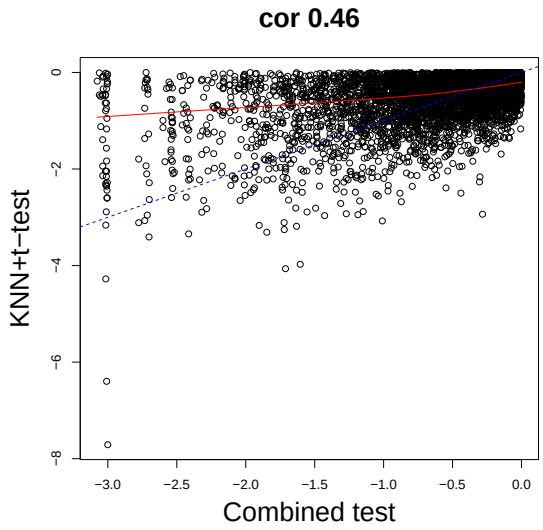


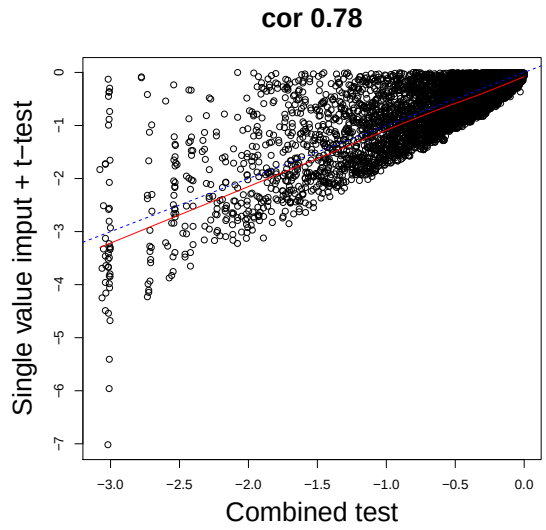


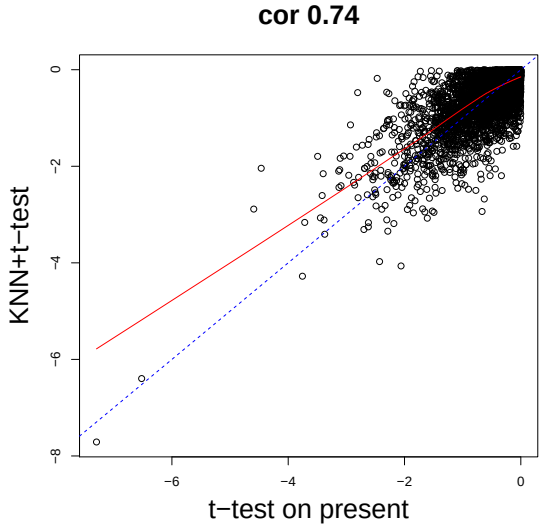
- KNN and single value: very different pv

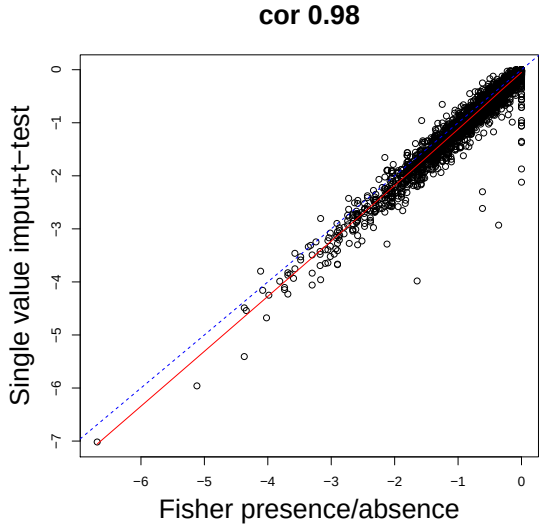
log10-p-values for the 3 procedures











- KNN and single value: very different pv
- Combined test ≈ recover results with KNN and single value

Summary of qualitative analysis

NA imputation

- ⊕ Flexible: enables any statistical analysis
- ⊖ The information of missingness is "lost"
- ⊖ Results highly dependent on methods and parameters
- ⊖ Only MAR or MNAR
- ⊕/⊖ KNN: makes use of correlation structure between variables

Combined test

- ⊖ Less flexible: only univariate statistical analysis
- ⊕ Preserve information of missingness
- ⊕ Both MAR and MNAR
  - Recover variables from KNN and single value imput.
- ⊕/⊖ Do not use correlation structure between variables
- ⊖ **Require sufficient sample size**

(Subjective) conclusion on comparison of statistical methods

- Quantitative performances: of interest but should be considered cautiously
  - Ranking varies with data set
  - Criteria may depend on (arbitrary) parametrisation
  - Cross-validation may be erroneous
- Complementary qualitative analysis
  - Examine underlying assumption/modelling
    - ⋆ Combine skills of biologist/biochemist and statistician/mathematician
  - If possible: produce technical replicates
- Implement various statistical strategies
  - More robust biological findings