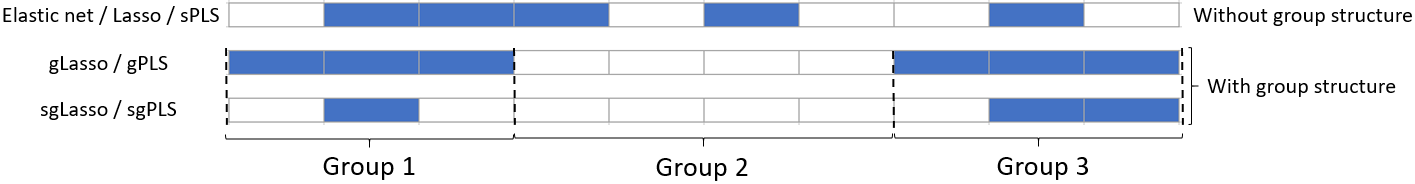
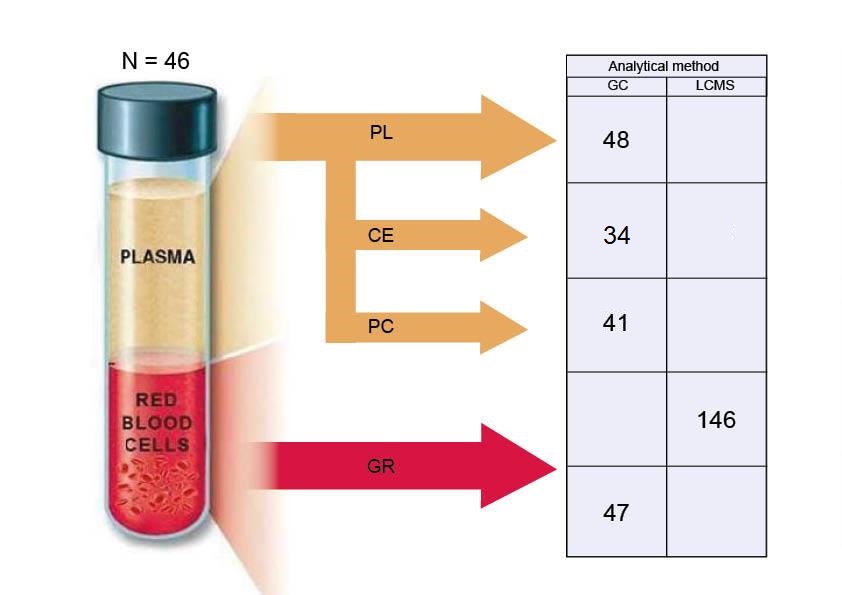
Supplementary material: “Benefits of dimension reduction in penalized regression methods for high dimensional grouped data: a case study in low sample size”

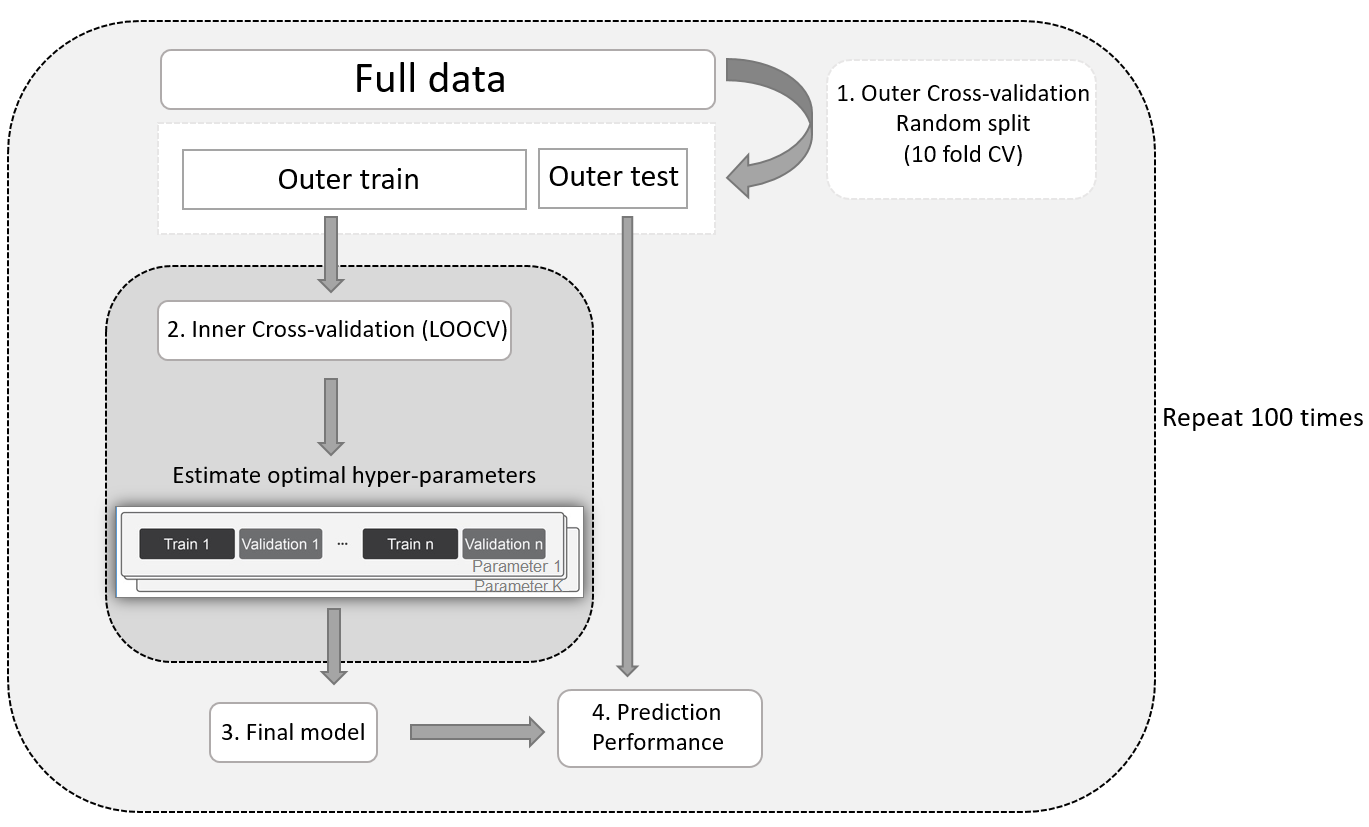


**Fig S1:** Comparison of penalized regression and dimension reduction methods in terms of variable selection while considering or not the group structure. Group 1, 2 and 3 are composed of 3, 4 and 3 predictors respectively. The selected (resp. discarded) predictors are displayed in blue (resp. in white).

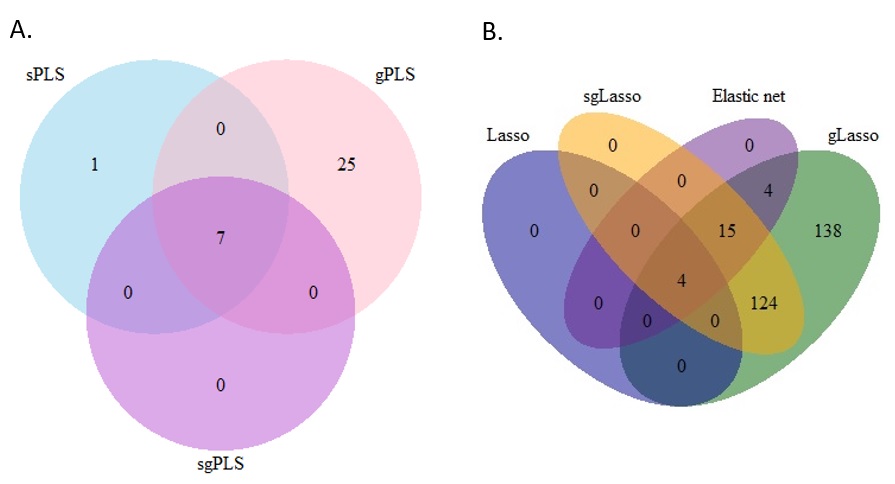
1. **BLISAR study:**



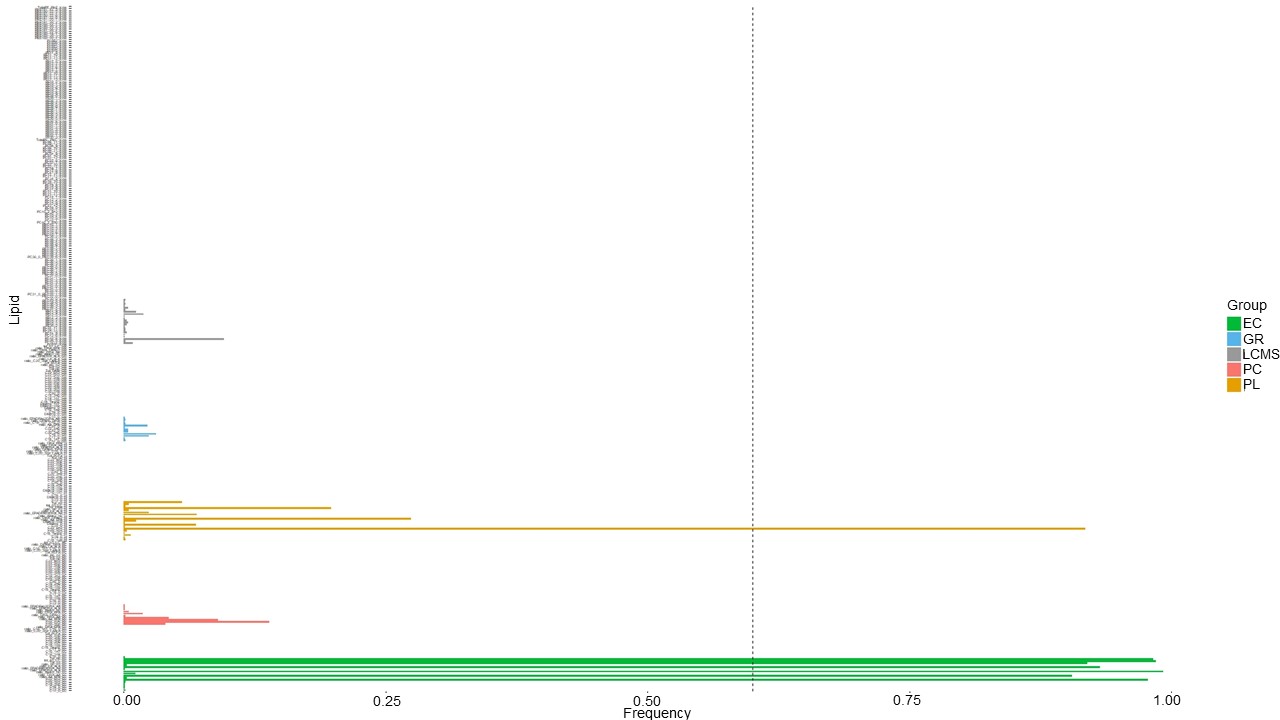
**Fig S2**. Scheme of the BLISAR post-mortem study (GC: gas chromatography; LCMS: liquid chromatography coupled to electroSpray ionization tandem mass spectrometry, CE: cholesteryl esters, PC: phosphatidylcholines, PL: total plasma, GR: red blood cells). The table represents the analytical methods used and the number of predictors measured in each group.



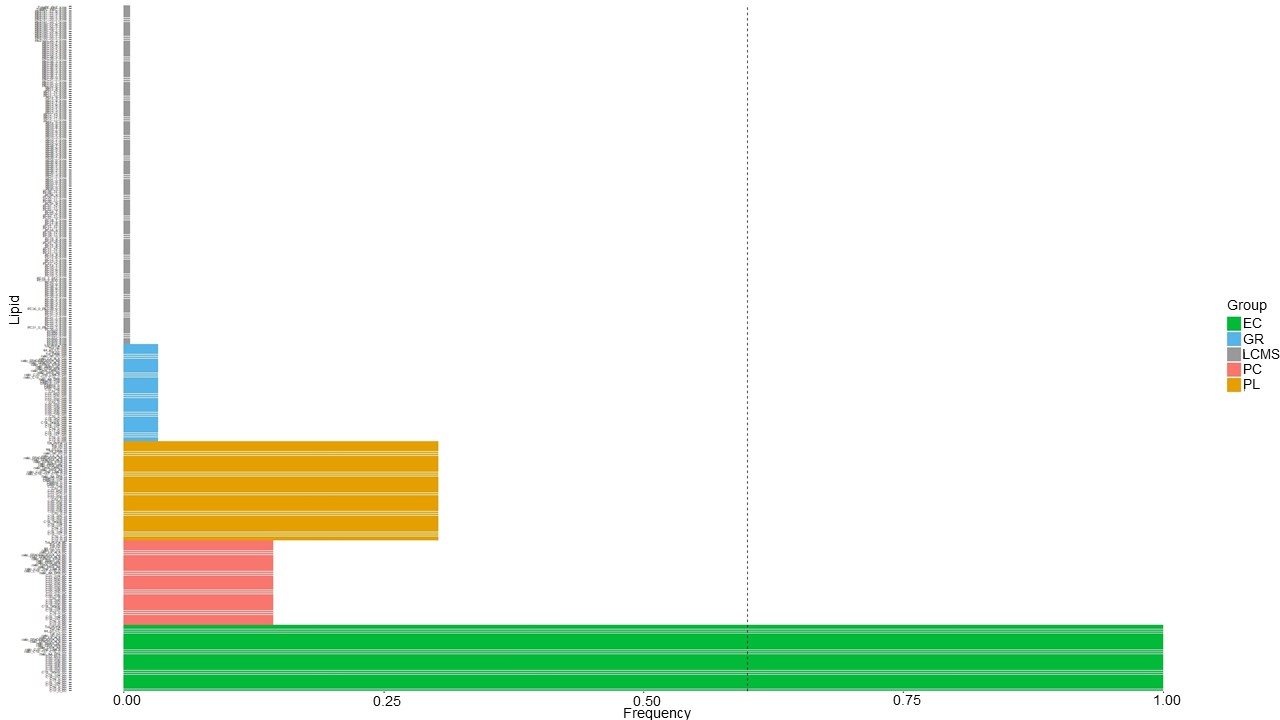
**Fig S3.** Repeated double cross validation scheme.

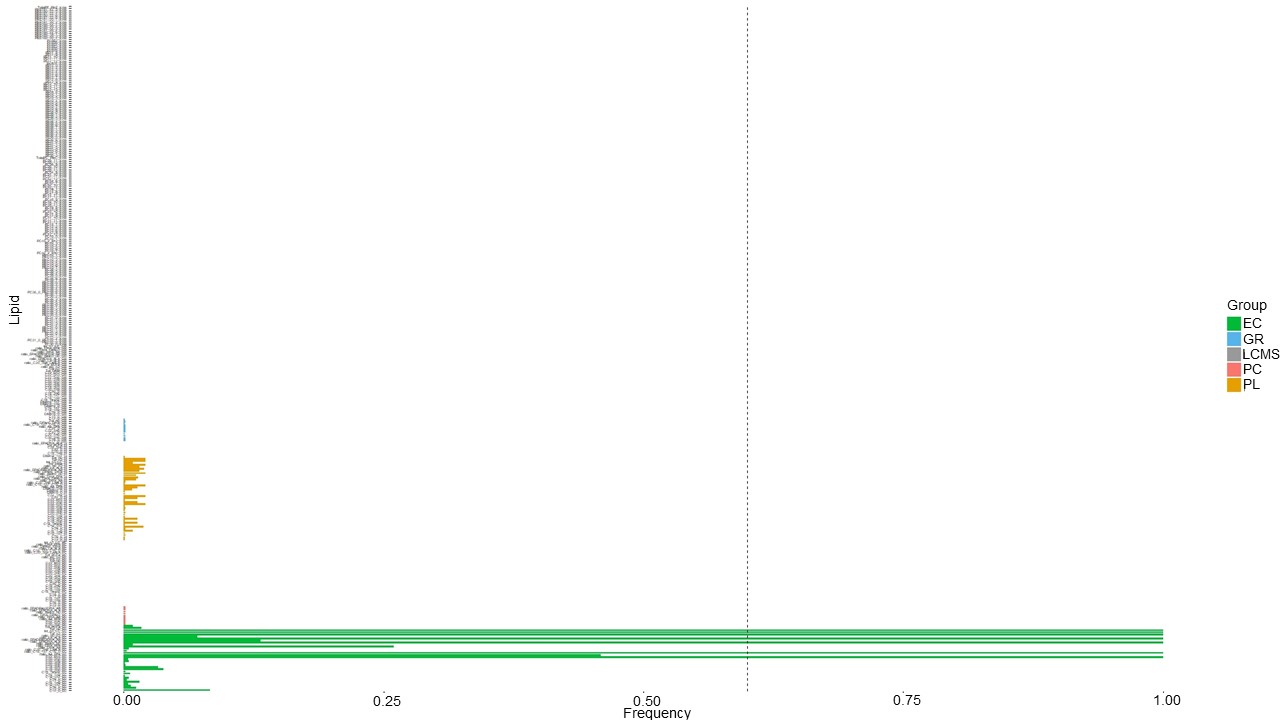
****

**Fig S4**. Venn diagrams of the predictors selected by dimension reduction methods (resp. penalized regression methods) on part A (resp. part B)

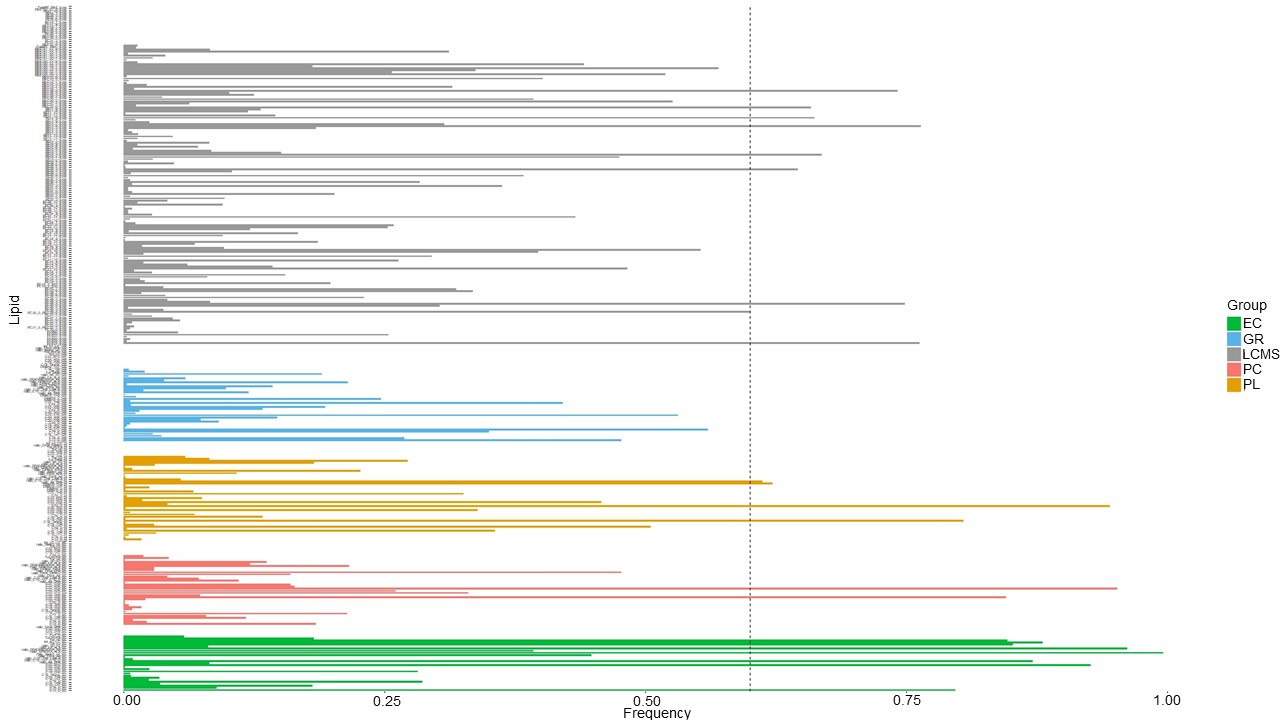
****

**Fig S5.** Lipids frequency selection with sPLS over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

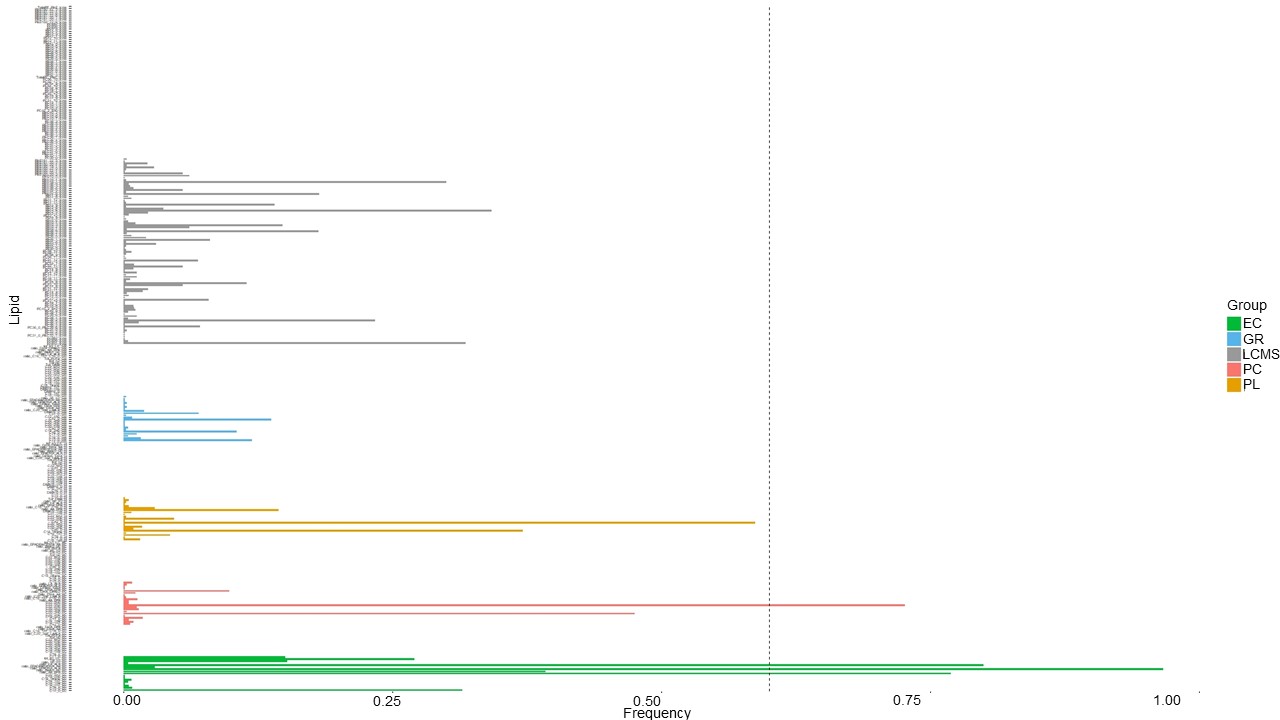
**Fig S6.** Lipids frequency selection with gPLS over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

****

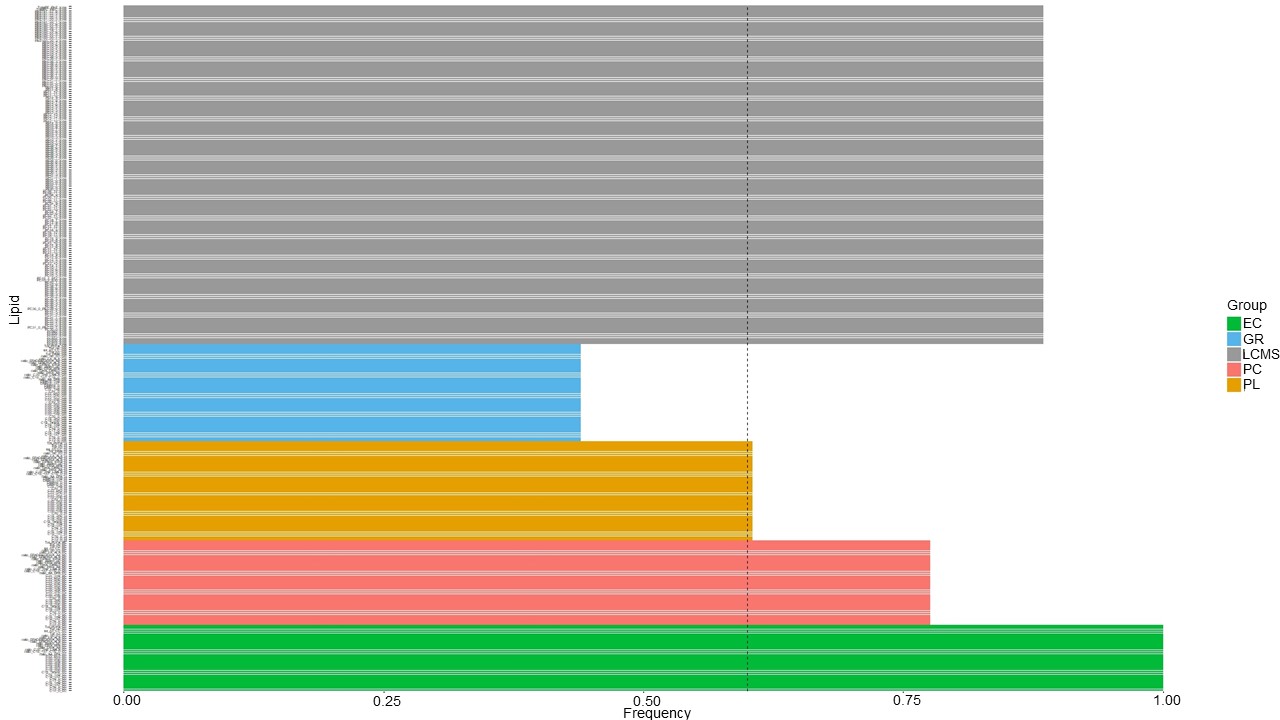
**Fig S7.** Lipids frequency selection with sgPLS over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

****

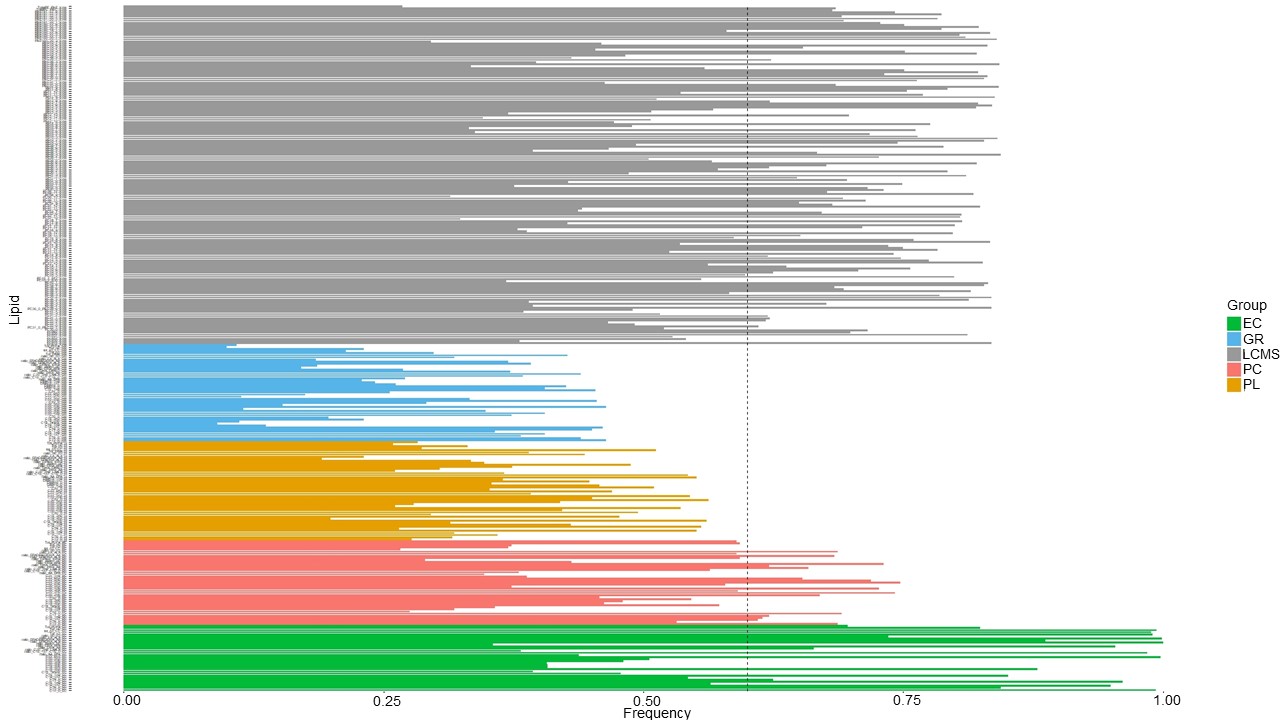
**Fig S8.** Lipids frequency selection with elastic net over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

****

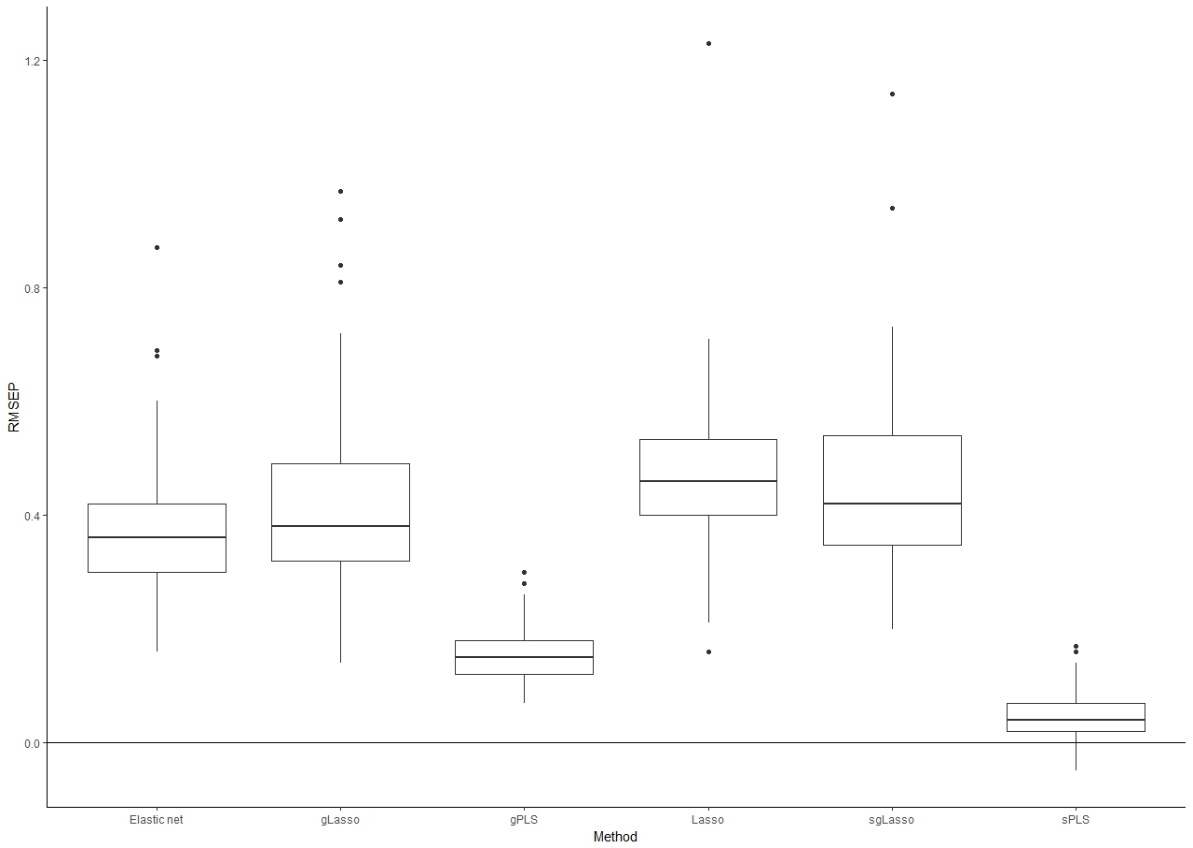
**Fig S9.** Lipids frequency selection with lasso over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

****

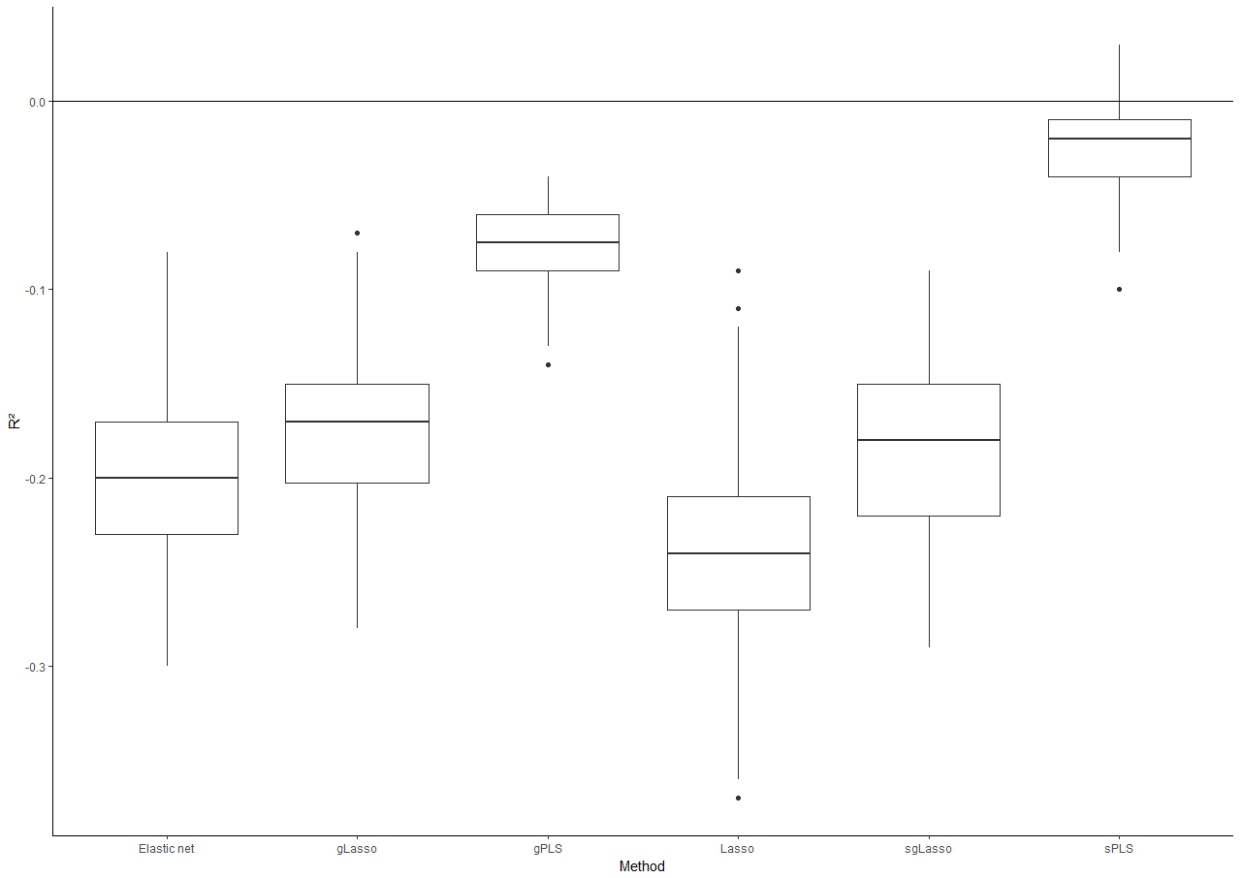
**Fig S10.** Lipids frequency selection with glasso over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.

****

**Fig S11.** Lipids frequency selection with sglasso over 100 runs. The vertical dashed line correspond to a frequency selection of 60%.



**Fig S12.** The boxplots of the difference in RMSEP between sgPLS (reference line) and the other six methods over the 100 runs



**Fig S13.** The boxplots of the difference in R² between sgPLS (reference line) and the other six methods over the 100 runs.

1. **DALIA trial:**
   1. **Description**

Dalia trial is an HIV vaccine trial evaluating the safety and the immunogenicity response of a dendritic-cell-based vaccine on HIV-infected patients. A detailed description of this study has been published elsewhere (Lévy *et al.*, 2014). The objective was to predict the immune response (CD4 polyfunctionnality) of 16 participants from the observation of the change of gene expressions during vaccination. The 5399 genes included in our dataset were grouped in 69 gene modules (groups of correlated genes) defined by Chaussabel et al. (Chaussabel *et al.*, 2008). It is noteworthy that each gene contributed to only one module (i.e., the defined modules were not overlapping) (Liquet *et al.*, 2016).

* 1. **Results:**

We compared the performances of the seven methods investigated in this work on this dataset in terms of prediction accuracy and variable selection (see Table S1). Among the methods taking into account the group structure, sgPLS reached the lowest prediction error (RMSEP=1.05) and selected only one group (M5.15) and 21 genes. Interestingly, gPLS obtained similar prediction performances (RMSEP=1.06) than sgPLS while selecting all the 24 genes of the same group (M5.15). In contrast, gLasso and sgLasso obtained lower prediction accuracies compared to gPLS and sgPLS while selecting more genes from different groups. Of note, among the four groups selected by gLasso (M5.15, M4.1, M4.11, M4.2), three were selected by sgLasso (M5.15, M4.1, M4.2). When the group structure was not considered, sPLS reached the best prediction performance (R²=0.37, RMSEP=1.38) compared to elastic net and lasso. However, sPLS and elastic net selected more groups (6) compared to lasso which selected only one group. Interestingly, the only group selected by lasso (M5.4) was consistently selected by elastic net and sPLS. It is noteworthy that all the compared methods except lasso have consistently selected the M5.15 group.

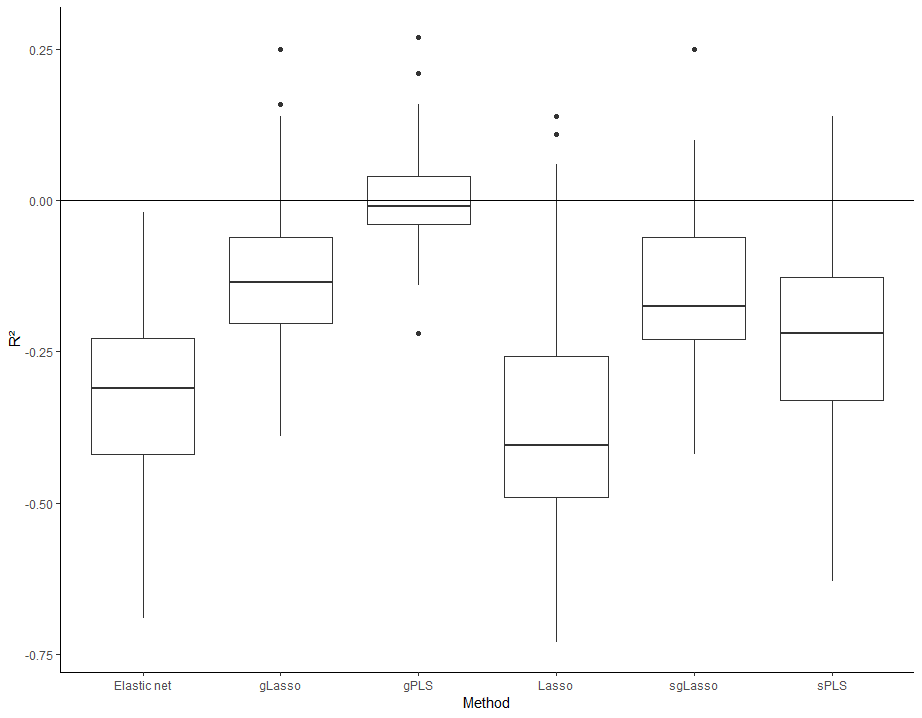
Supplementary figure S14 shows the difference in R² between sgPLS (considered as a benchmark) and the other methods over the 100 runs. It clearly shows that sgPLS outperformed the other methods in most of the runs (except gPLS which is close to sgPLS in this application). We observe similar results for the RMSEP criterion (Supplementary figure S15).

To summarize, even in a very low sample size (N=16) and a high number of features (p=5399) configuration, sgPLS reached the best predictive performances while selecting the most predictive genes from a single group.

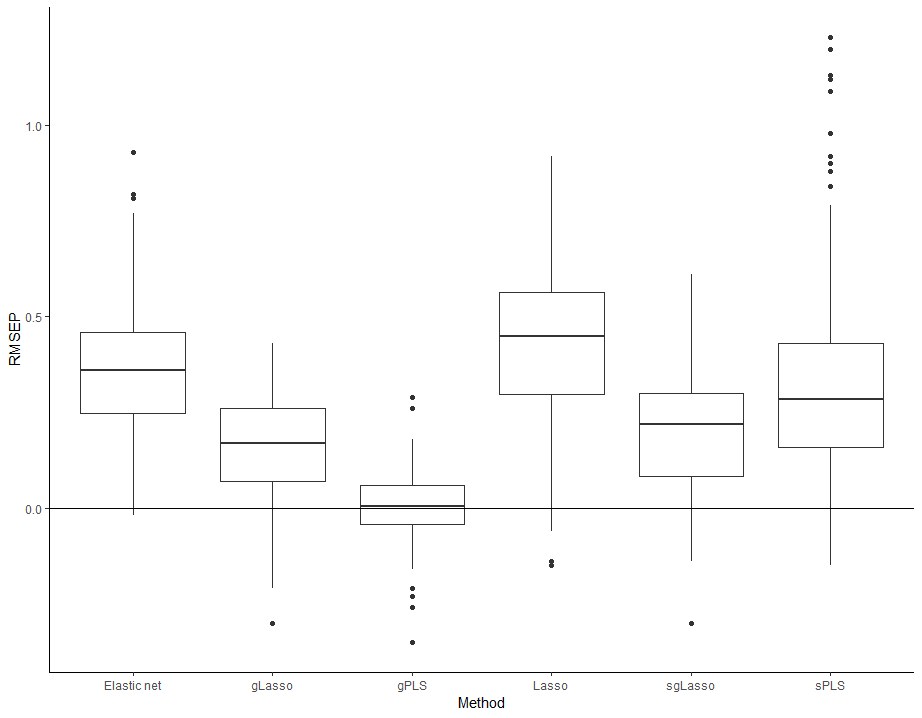
**Table S1.** Comparison of the multivariable regression methods for 10 random divisions with 100 runs (N=16, p=5399)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Test data R² (SD)** | **Test data RMSEP (SD)** | **Number of selected predictors\*** | **Selected groups\*** |
| Lasso | 0.21 (0.15) | 1.49 (0.19) | 1 | M5.4 |
| sgLasso | 0.43 (0.10) | 1.25 (0.13) | 45 | M5.15, M4.1, M4.2 |
| gLasso | 0.45 (0.09) | 1.22 (0.11) | 144 | M5.15, M4.1, M4.11, M4.2 |
| Elastic net | 0.26 (0.15) | 1.42 (0.16) | 14 | M5.15, M7.11, M5.4, M7.27, M4.2, M4.6 |
| sPLS | 0.37 (0.16) | 1.38 (0.31) | 19 | M5.15, M5.4, M4.2, M7.27, M3.1, M4.6 |
| gPLS | **0.58 (0.09)** | **1.06 (0.11)** | **24** | **M5.15** |
| **sgPLS** | **0.58 (0.11)** | **1.05 (0.14)** | **21** | **M5.15** |

**\*** In at least 60% of the samples



**Fig S14.** The boxplots of the difference in R² between sgPLS (reference line) and the other six methods over the 100 runs



**Fig S15.** The boxplots of the difference in RMSEP between sgPLS (reference line) and the other six methods over the 100 runs

1. **Bibliography:**

Chaussabel,D. *et al.* (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*, **29**, 150–164.

Lévy,Y. *et al.* (2014) Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load: Clinical immunology. *Eur. J. Immunol.*, **44**, 2802–2810.

Liquet,B. *et al.* (2016) Group and sparse group partial least square approaches applied in genomics context. *Bioinforma. Oxf. Engl.*, **32**, 35–42.