

# Feature Selection and Supervised OMICs Integration

ISMB / ECCB 2021

Tutorial: a practical introduction to multi-omics integration and network analysis

Nikolay Oskolkov, NBIS SciLifeLab, 22.07.2021

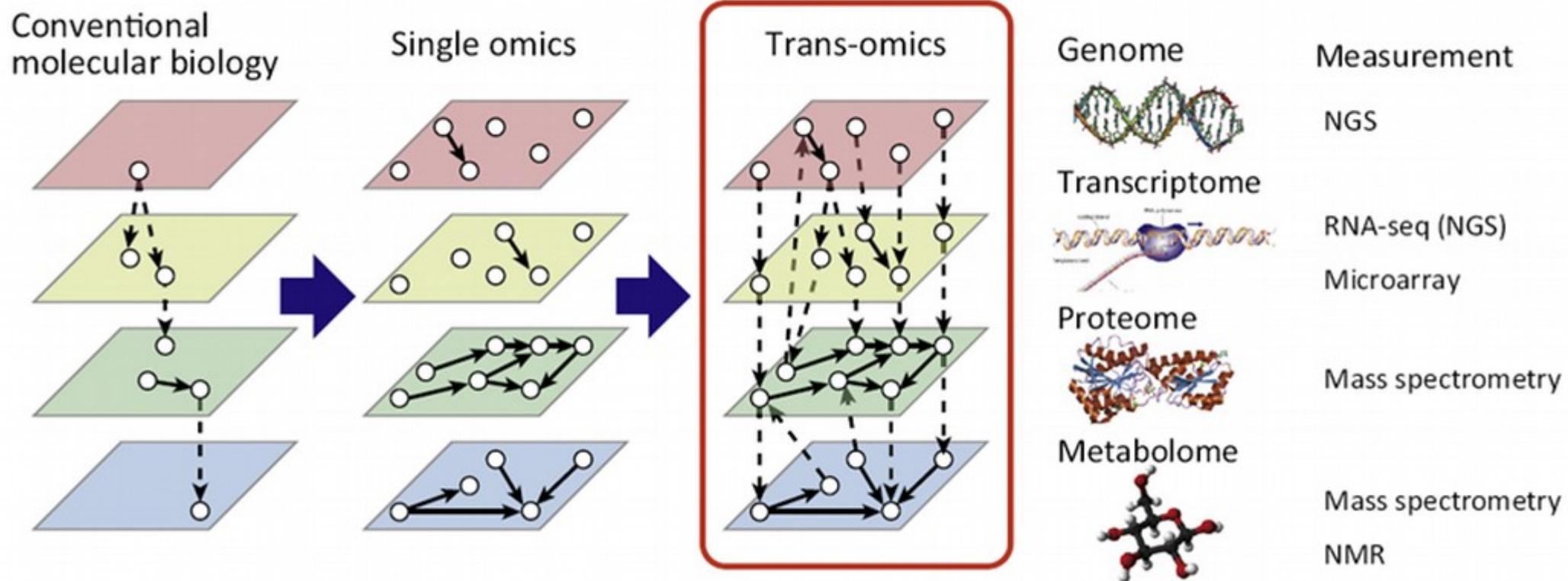
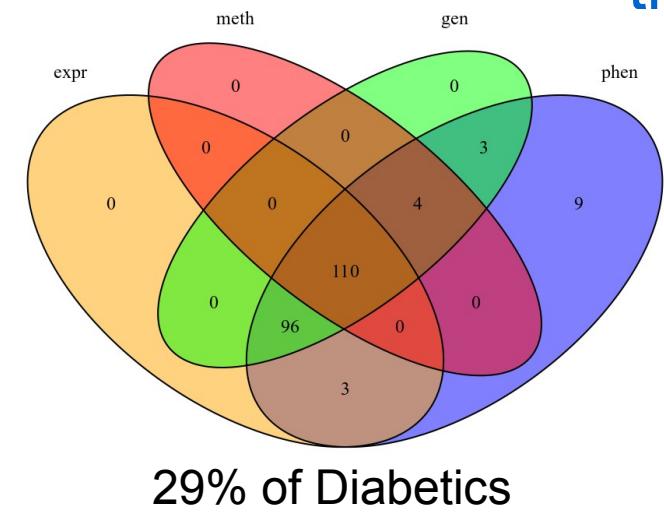


Image adapted from Yugi et al., Trends Biotechnol. 2016 Apr;34(4):276–290

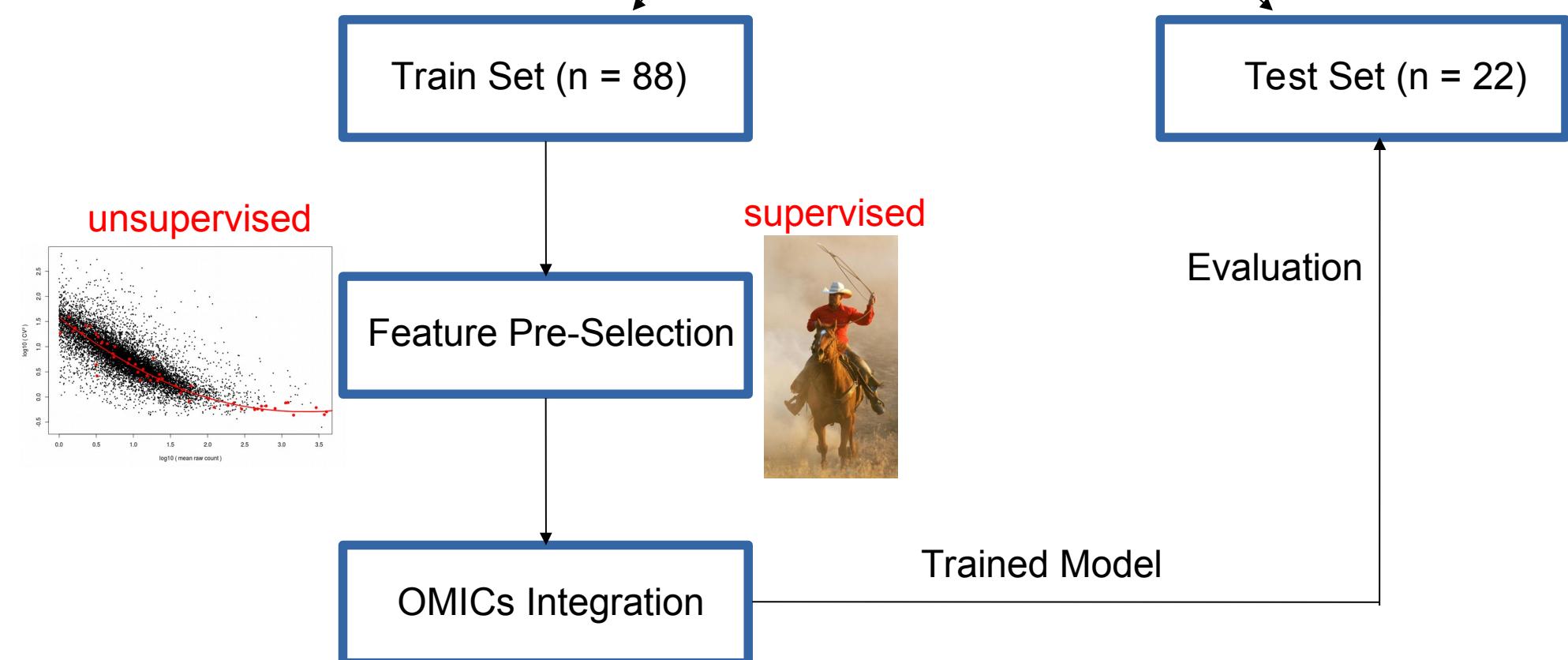
	Linear	Non-Linear
Supervised	PLS / OPLS / mixOmics, LASSO / Ridge / Elastic Net	Neural Networks, Random Forest, Bayesian Networks
Unsupervised	Factor Analysis / MOFA	Autoencoder, SNF, UMAP, Clustering of Clusters

## Example:

- 1) With ~110 samples it is a good idea to do **linear OMICs integration**
- 2) T2D is a phenotype of interest, therefore **supervised integration**



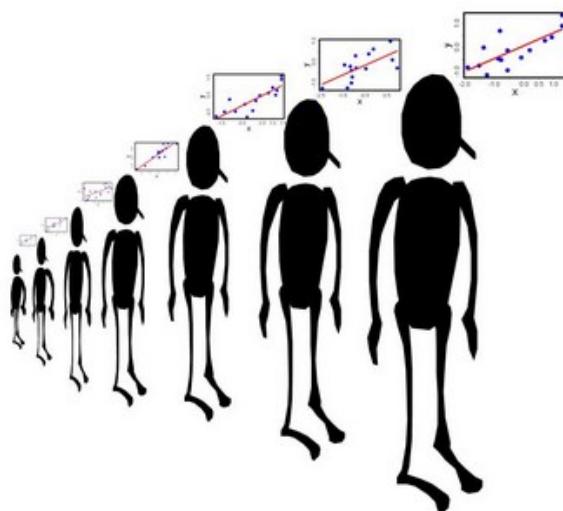
## 29% of Diabetics



```
##          n1          n2          n3          n4          n5
## p1 -0.6760258 -1.2307634 1.66039982 0.196033326 -0.2981471
## p2 -1.5834993  0.6494188 -0.01267663 -1.064763128 -0.1792141
## p3  0.3152418 -0.5791937 -1.79593465 -0.312303710  0.2671534
## p4 -0.9359010  0.1212546 -0.36279328 -0.553364109  1.0598898
## p5 -2.0411903  0.6899356 -1.03923098  0.008958754 -0.2249498
```

- Two types of non-independence in data
  - between samples
  - between features

## Random Effects



## Lasso



# Univariate and Multivariate Feature Selection

# Univariate Feature Selection

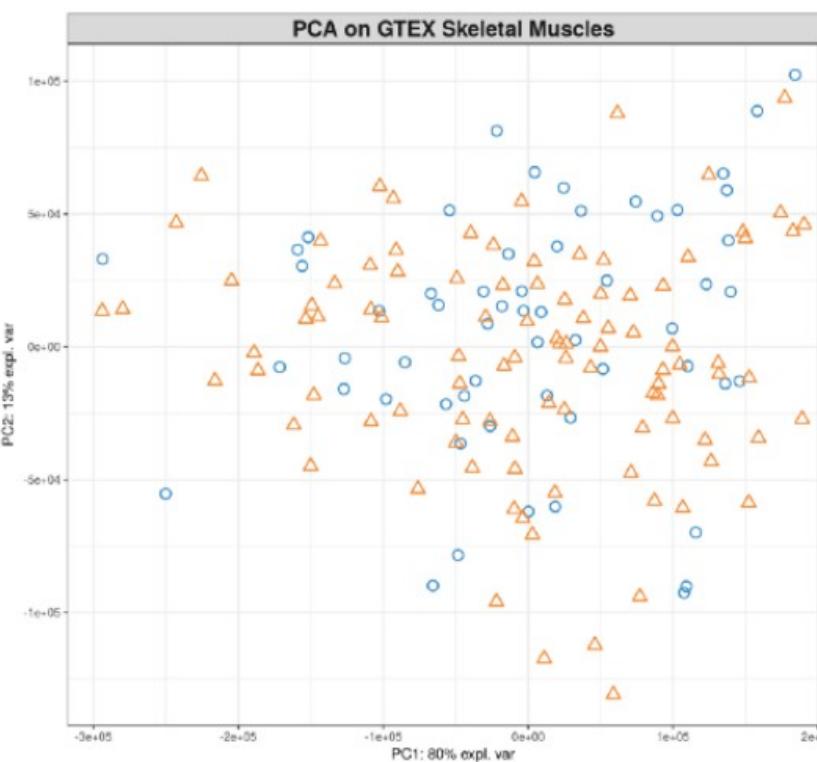
```

1 X <- read.table("GTEX_SkeletalMuscles_157Samples_1000Genes.txt",
2   header=TRUE, row.names=1, check.names=FALSE, sep="\t")
3 X <- X[, colMeans(X) >= 1]
4 Y <- read.table("GTEX_SkeletalMuscles_157Samples_Gender.txt",
5   header=TRUE, sep="\t")$GENDER
6 library("mixOmics")
7 pca.gtex <- pca(x, ncomp=10)
8 plot(pca.gtex)
9 plotIndiv(pca.gtex, group = Y, ind.names = FALSE, legend = TRUE,
10   title = 'PCA on GTEX Skeletal Muscles')

```

ReadGTEX.R hosted with ❤ by GitHub

[view raw](#)



```

1 rho <- vector()
2 p <- vector()
3 a <- seq(from=0, to=dim(X)[2], by=100)
4 for(i in 1:dim(X)[2])
5 {
6   corr_output <- cor.test(X[,i], as.numeric(Y), method="spearman")
7   rho <- append(rho, as.numeric(corr_output$estimate))
8   p <- append(p, as.numeric(corr_output$p.value))
9   if(isTRUE(i %in% a)==TRUE){print(paste("FINISHED ",i," FEATURES", sep=""))}
10 }
11 output <- data.frame(GENE=colnames(X), SPEARMAN_RHO=rho, PVALUE=p)
12 output$FDR <- p.adjust(output$PVALUE, method="fdr")
13 output <- output[order(output$FDR, output$PVALUE, -output$SPEARMAN_RHO), ]
14 head(output, 10)

```

UnivarFeatureSelect.R hosted with ❤ by GitHub

[view raw](#)

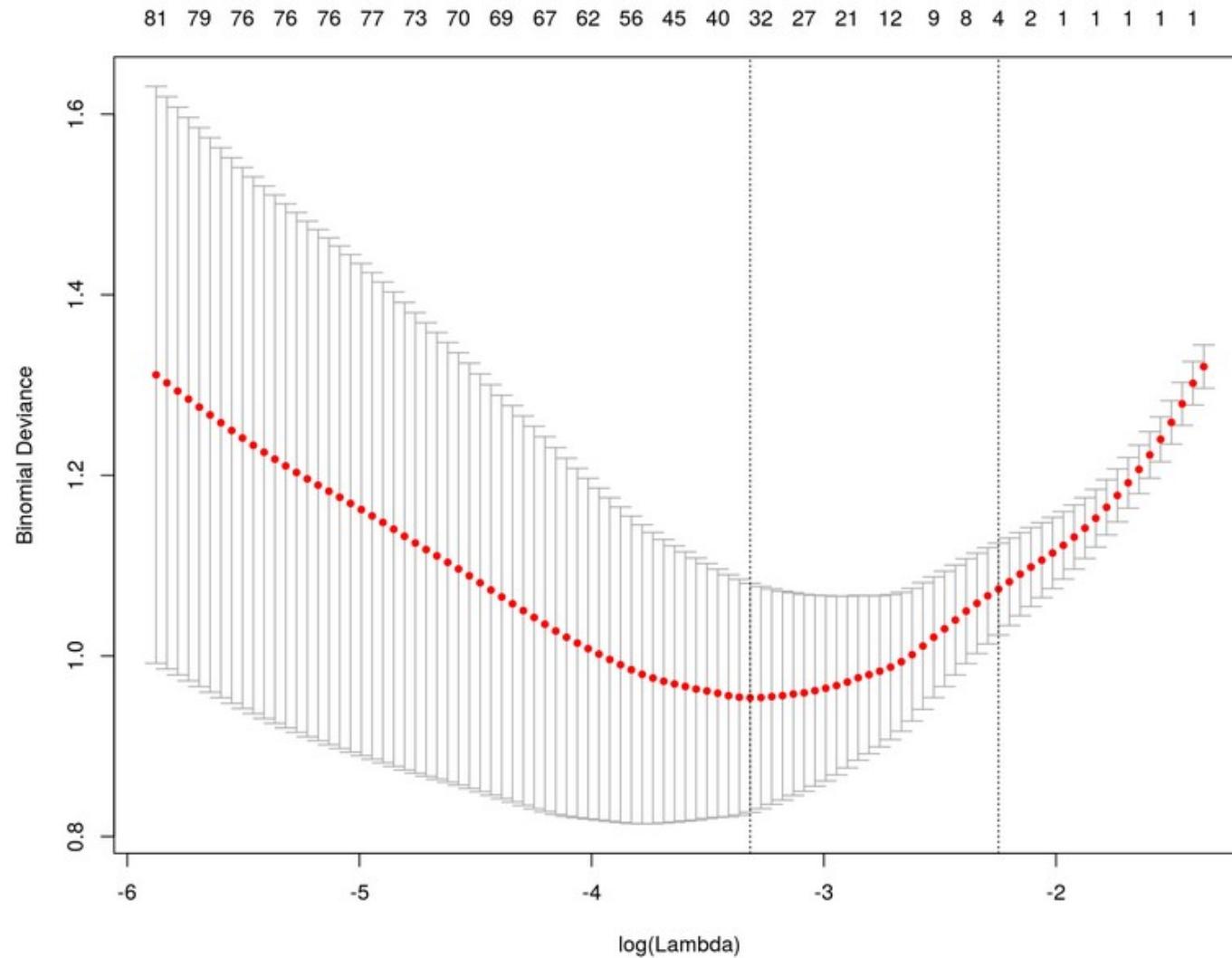
##	GENE	SPEARMAN_RHO	PVALUE	FDR
## 256	ENSG00000184368.11_MAP7D2	-0.5730196	4.425151e-15	2.416132e-12
## 324	ENSG00000110013.8_SIAE	0.3403994	1.288217e-05	3.516833e-03
## 297	ENSG00000128487.12_SPECC1	-0.3003621	1.323259e-04	2.408332e-02
## 218	ENSG00000162512.11_SDC3	0.2945390	1.807649e-04	2.467441e-02
## 38	ENSG00000129007.10_CALML4	0.2879754	2.549127e-04	2.783647e-02
## 107	ENSG00000233429.5_HOTAIRM1	-0.2768054	4.489930e-04	4.085836e-02
## 278	ENSG00000185442.8_FAM174B	-0.2376098	2.731100e-03	2.130258e-01
## 421	ENSG00000234585.2_CCT6P3	-0.2322268	3.426233e-03	2.338404e-01
## 371	ENSG00000113312.6_TTC1	0.2284351	4.007655e-03	2.431310e-01
## 269	ENSG00000226329.2_AC005682.6	-0.2226587	5.064766e-03	2.523944e-01

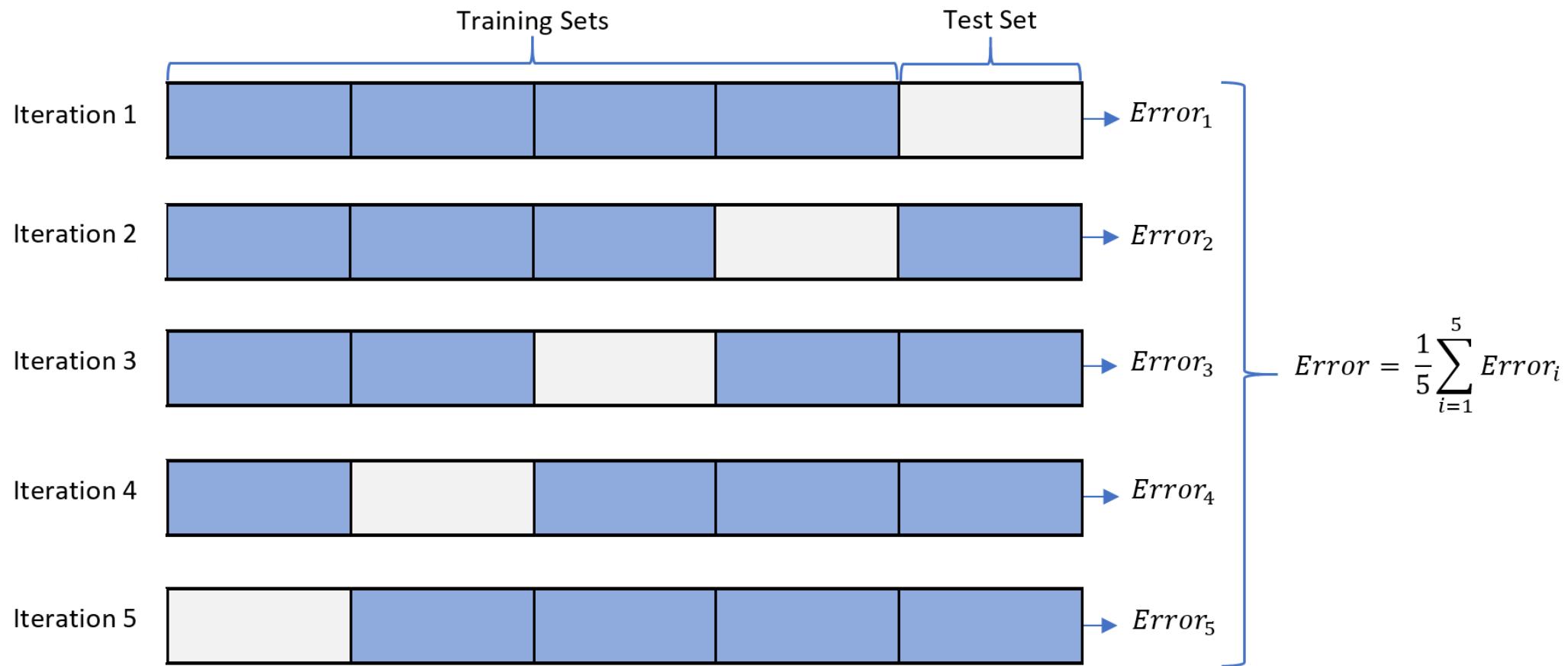
Generally acknowledged that univariate feature selection has a poor predictive capacity compared to multivariate feature selection

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

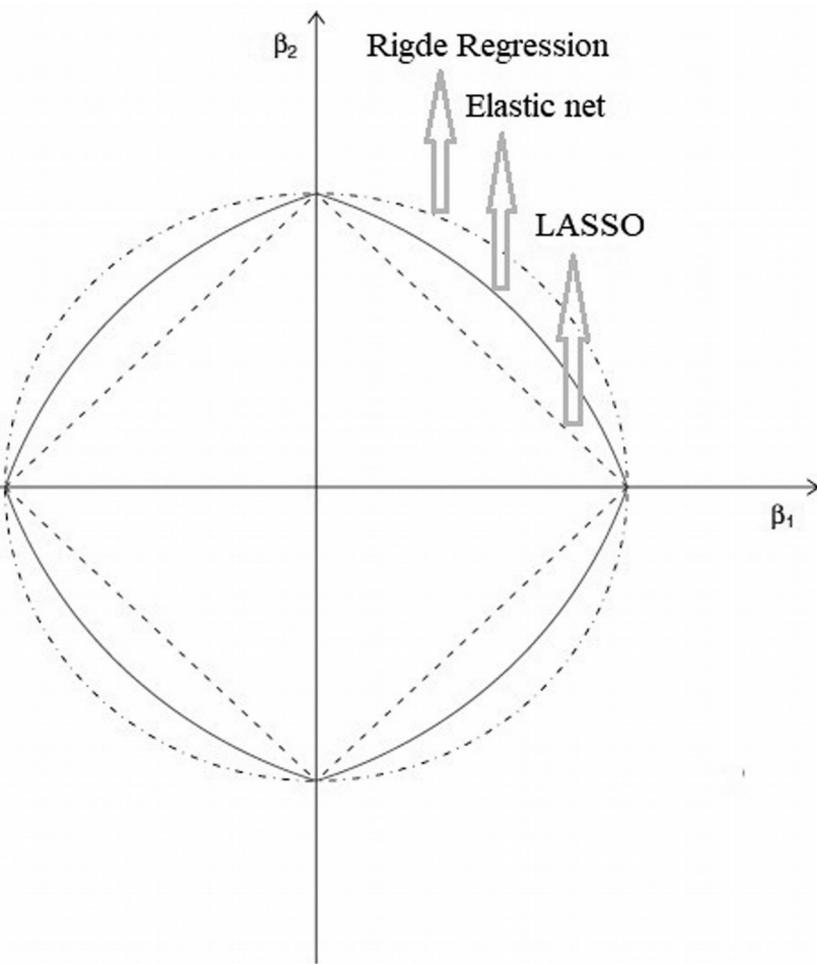
$$\text{OLS} = (y - \beta_1 X_1 - \beta_2 X_2)^2$$

$$\text{Penalized OLS} = (y - \beta_1 X_1 - \beta_2 X_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$





Cross-validation is a standard way to tune model hyperparameters such as  $\lambda$  in LASSO

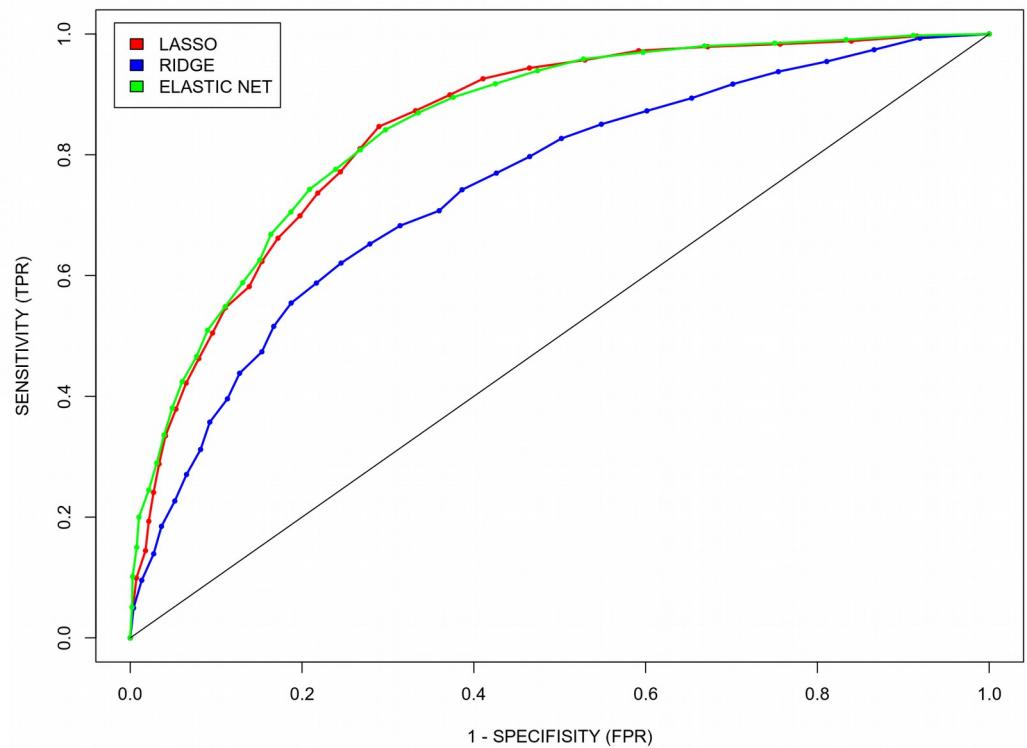
Difference between Lasso, Ridge,  
and Elastic Net

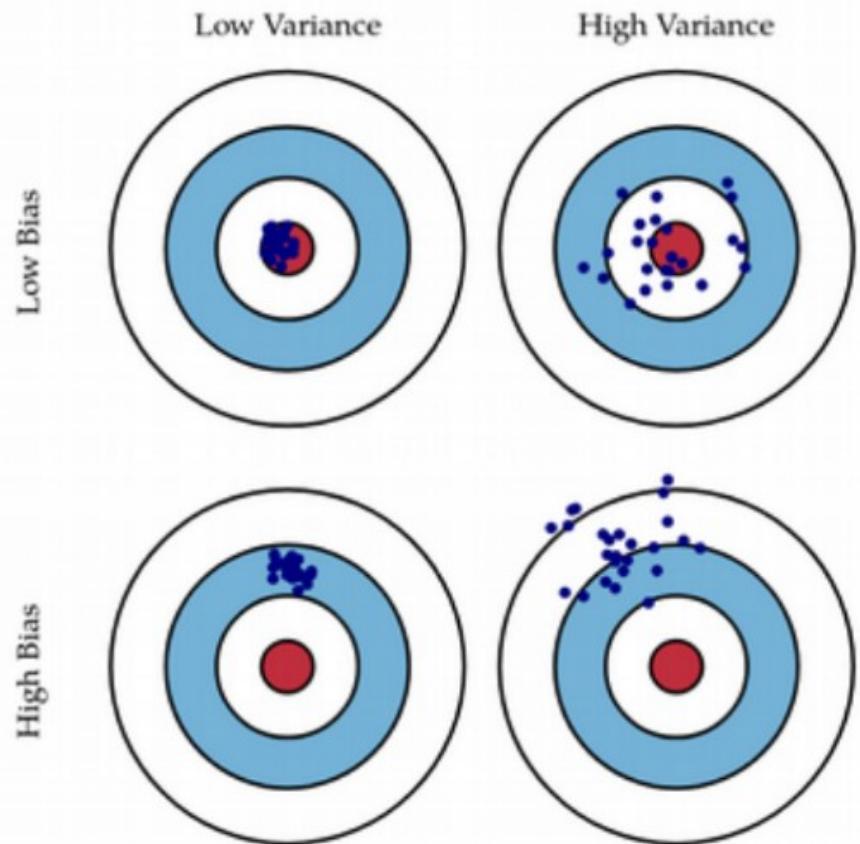
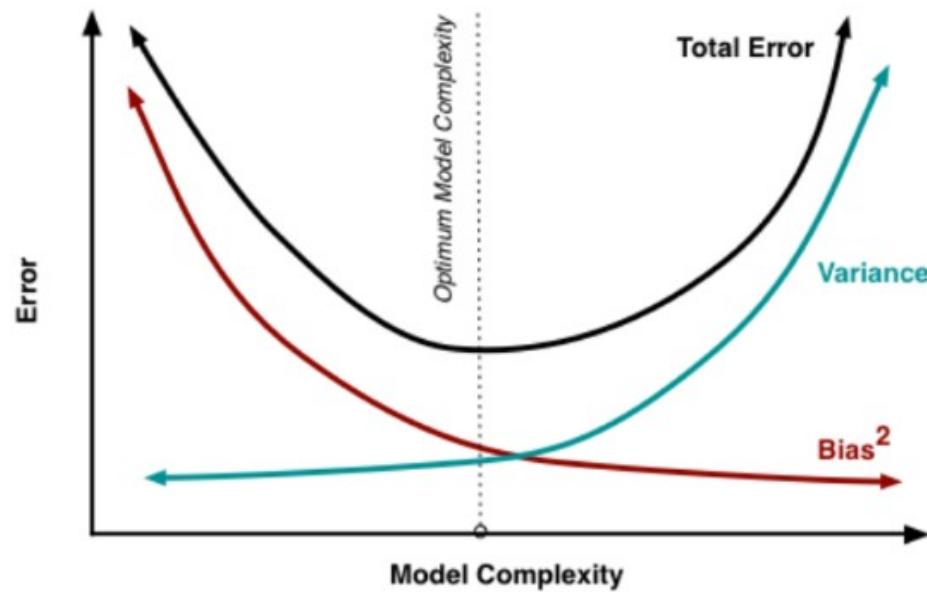
Lasso is more conservative

Ridge is more permissive

$$\text{Lasso} : |\beta_1| + |\beta_2| \leq \lambda$$

$$\text{Ridge} : \beta_1^2 + \beta_2^2 \leq \lambda$$

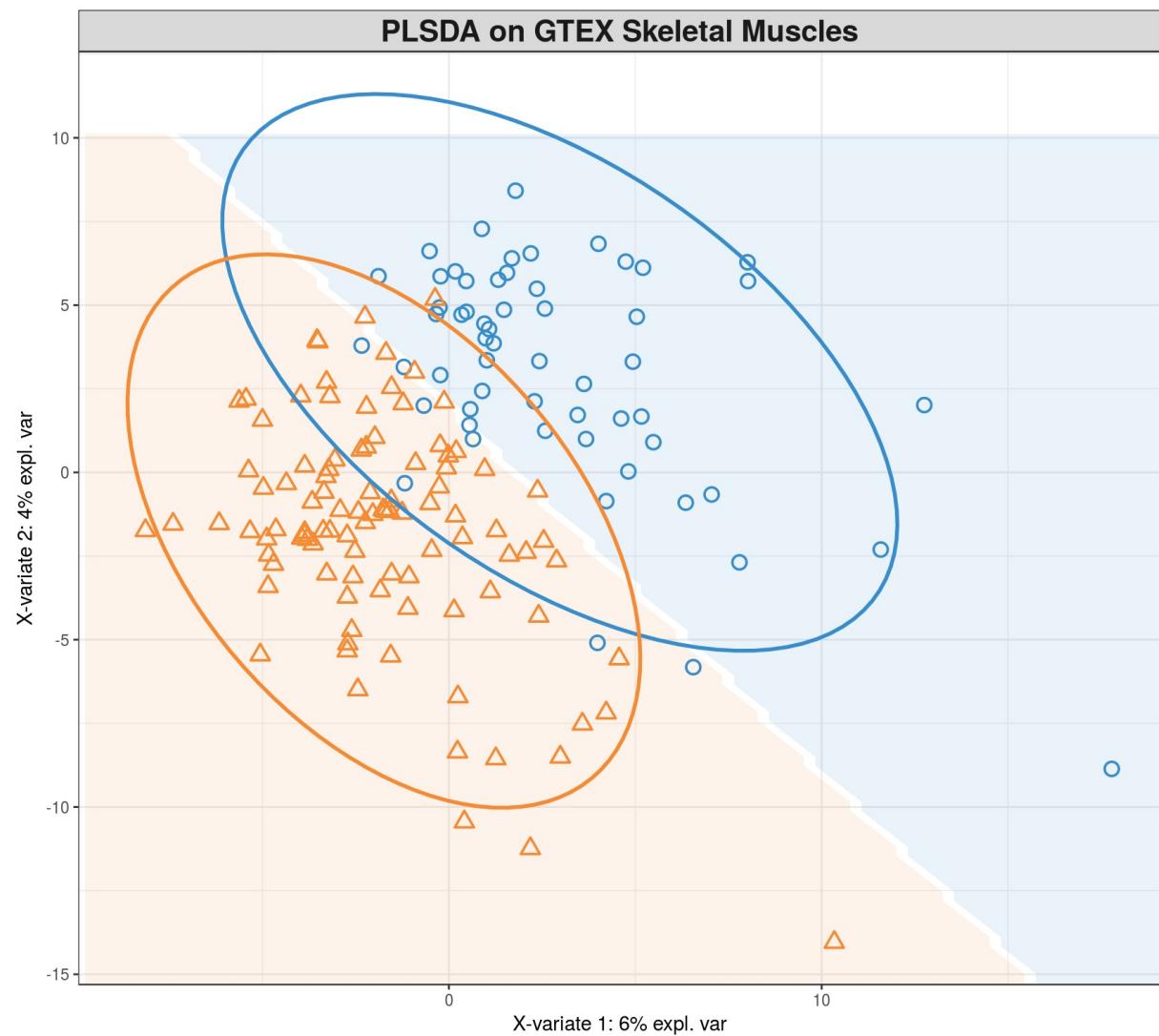




$$Y = f(X) \implies \text{Reality}$$

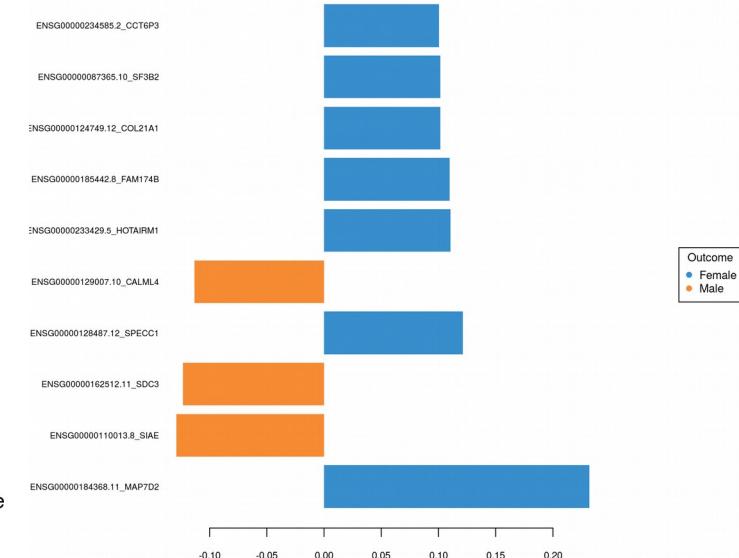
$$\hat{Y} = \hat{f}(X) + \text{Error} \implies \text{Model}$$

$$\text{Error}^2 = (Y - \hat{f}(X))^2 = \text{Bias}^2 + \text{Variance}$$

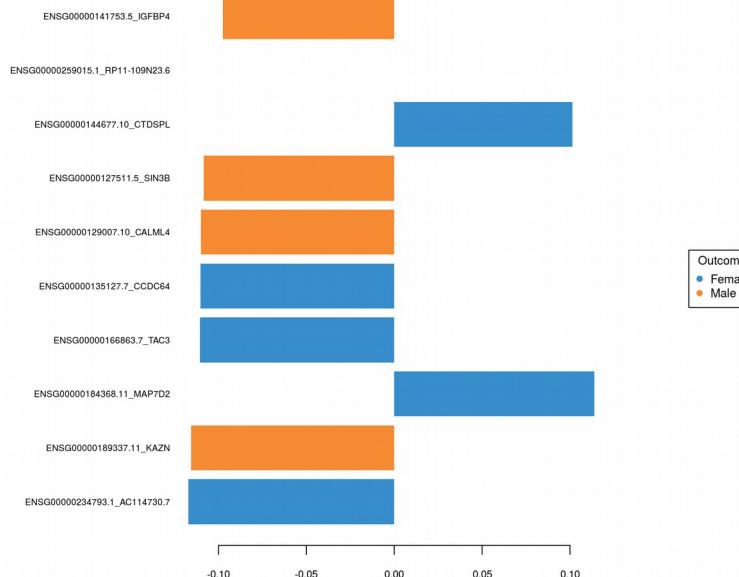


Select features that separate two groups of samples the most

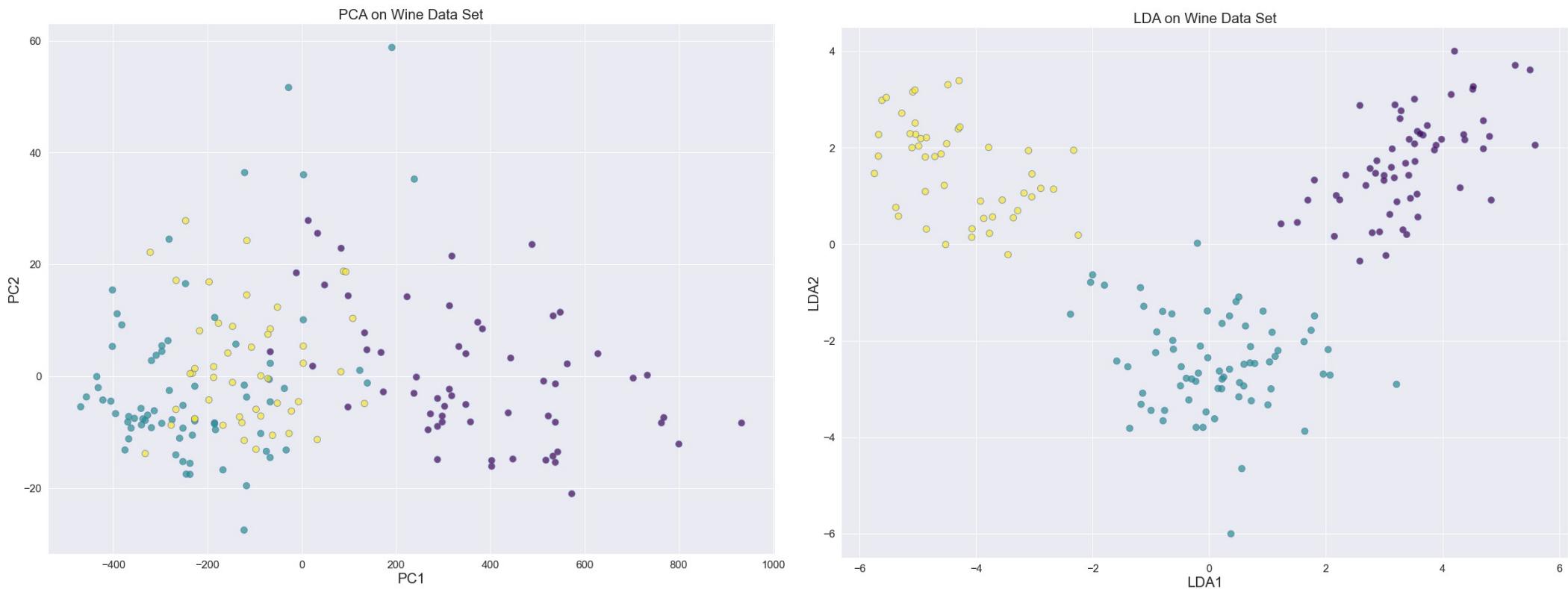
Loadings on comp 1



Loadings on comp 2

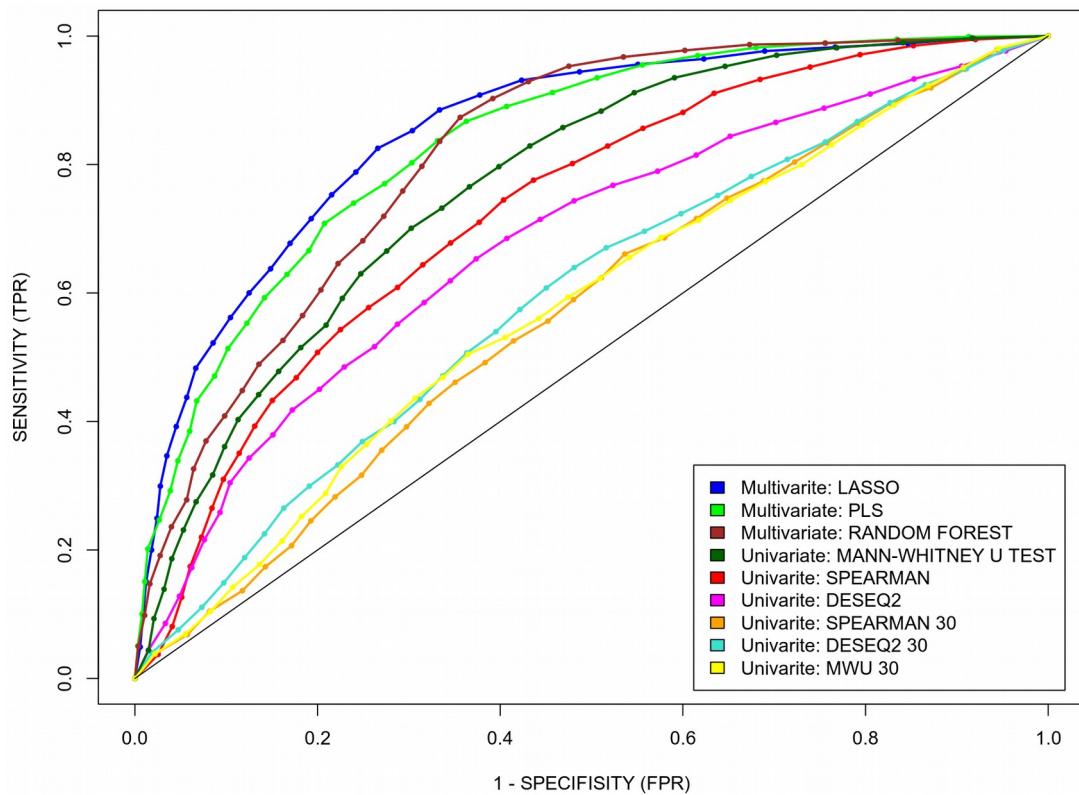


# Multivariate Feature Selection: Linear Discriminant Analysis (LDA)

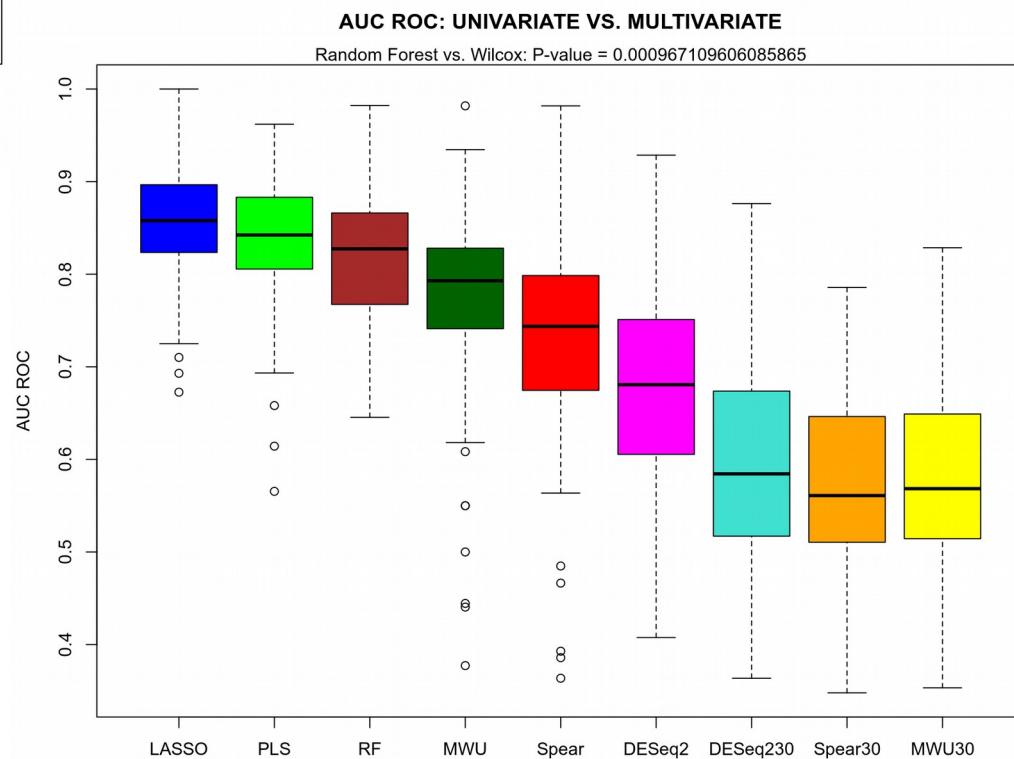


Minimize variance within clusters and maximize variance between clusters

Similar to what ANOVA / t-test is doing, therefore LINEAR Discriminant Analysis (LDA)

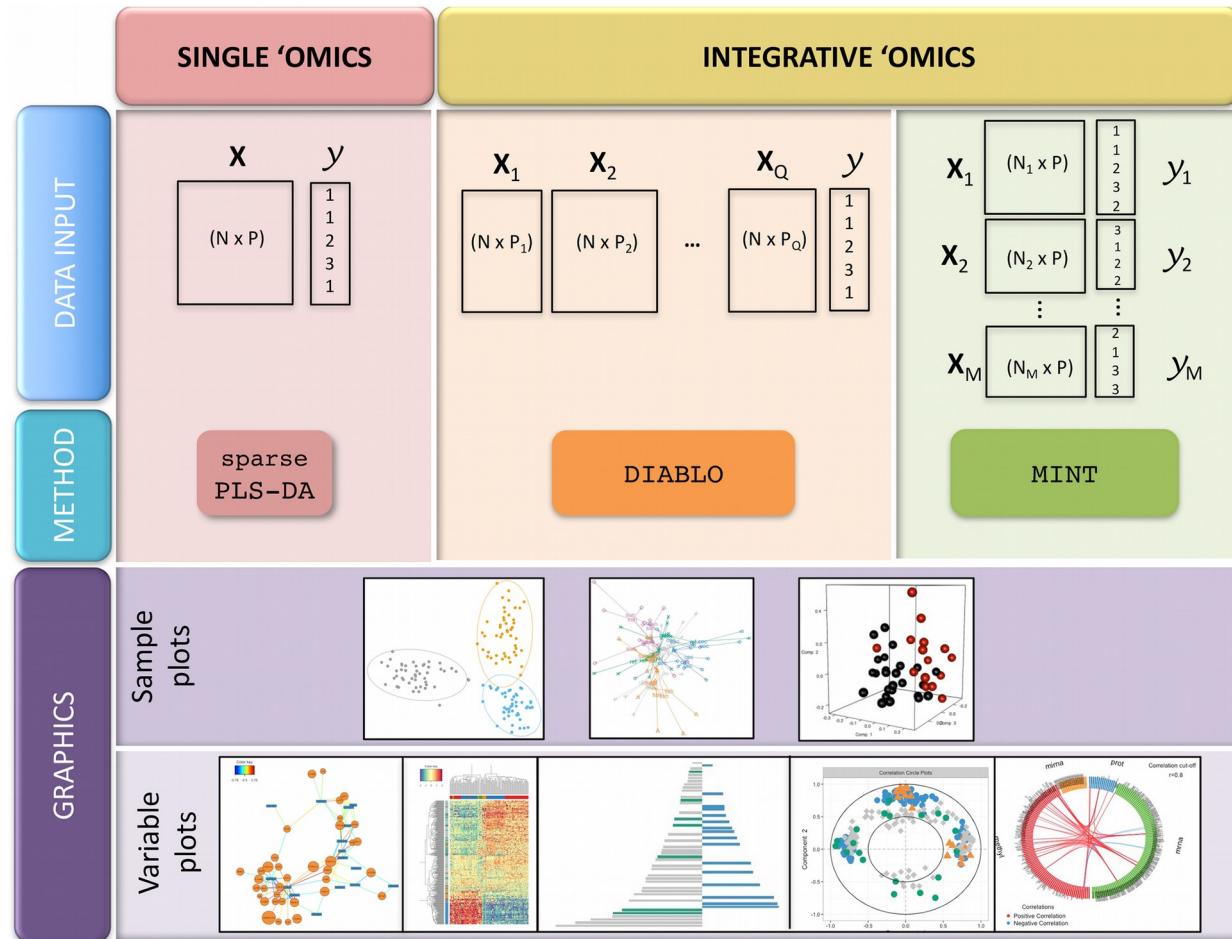


Multivariate methods (LASSO, PLS, RanFor) have significantly higher AUC ROC than univariate methods (Spear, MWU, DESeq2) on skeletal muscle gene expression data



If you find a data set where it is not true,  
please let me know

# DIABLO Omics Integration

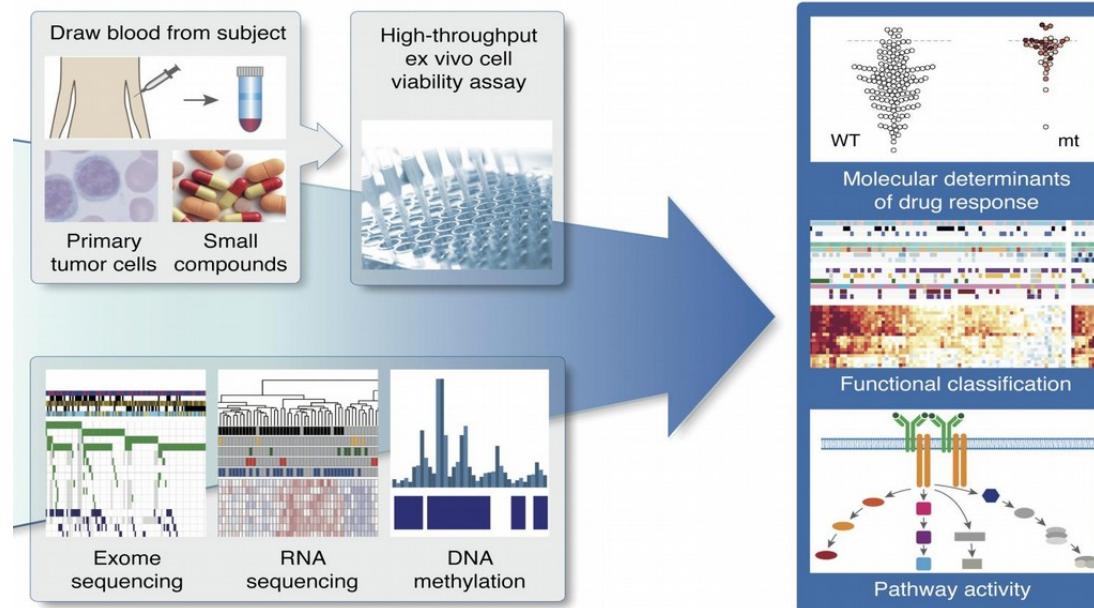


$$\max_{\beta} \text{cov}(X, Y) \implies \hat{\beta}$$

Denote  $Q$  normalized, centered and scaled datasets  $X^{(1)} (N \times P_1), X^{(2)} (N \times P_2), \dots, X^{(Q)} (N \times P_Q)$  measuring the expression levels of  $P_1, \dots, P_Q$  'omics variables on the same  $N$  samples'. sGCCA solves the optimization function for each dimension  $b = 1, \dots, H$ :

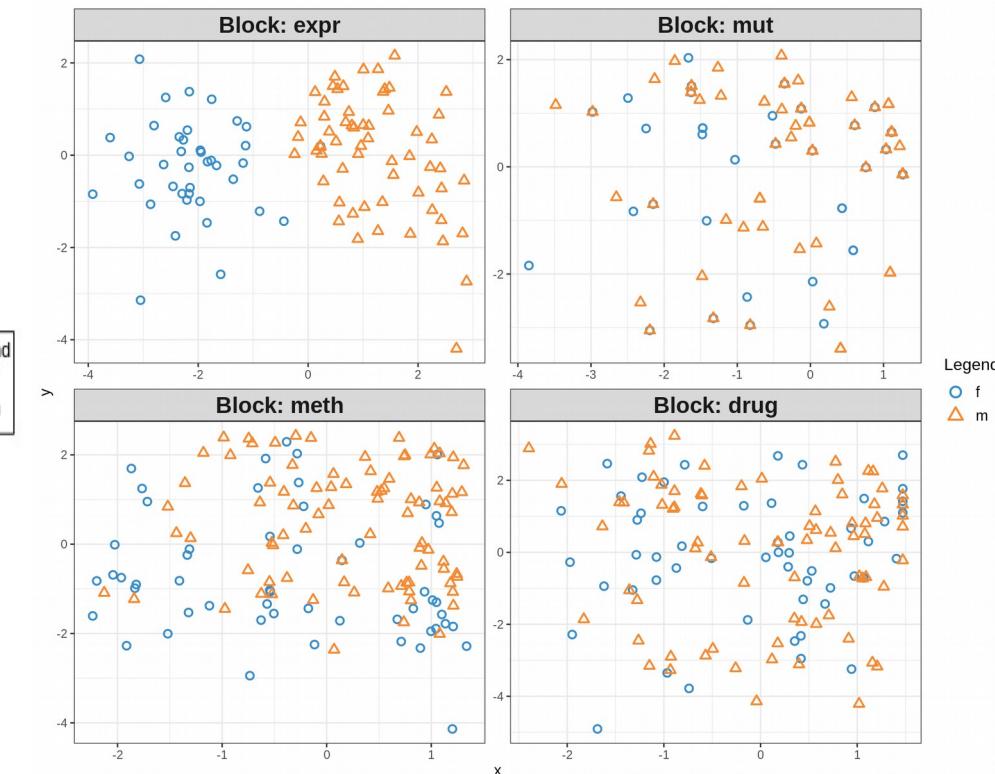
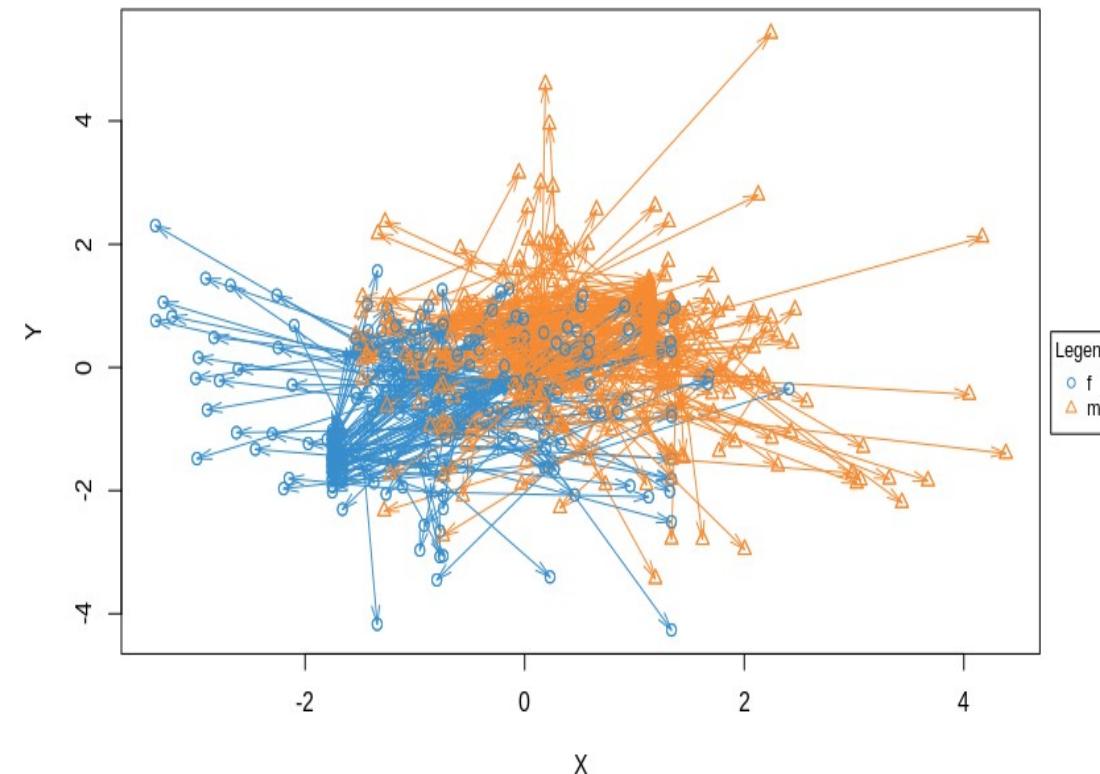
$$\begin{aligned} & \max_{a_b^{(1)}, \dots, a_b^{(Q)}} \sum_{i,j=1, i \neq j}^Q c_{i,j} \text{cov}(X_b^{(i)} a_b^{(i)}, X_b^{(j)} a_b^{(j)}), \\ & \text{s.t. } \|a_b^{(q)}\|_2 = 1 \text{ and } \|a_b^{(q)}\|_1 \leq \lambda^{(q)} \text{ for all } 1 \leq q \leq Q \end{aligned} \quad (1)$$

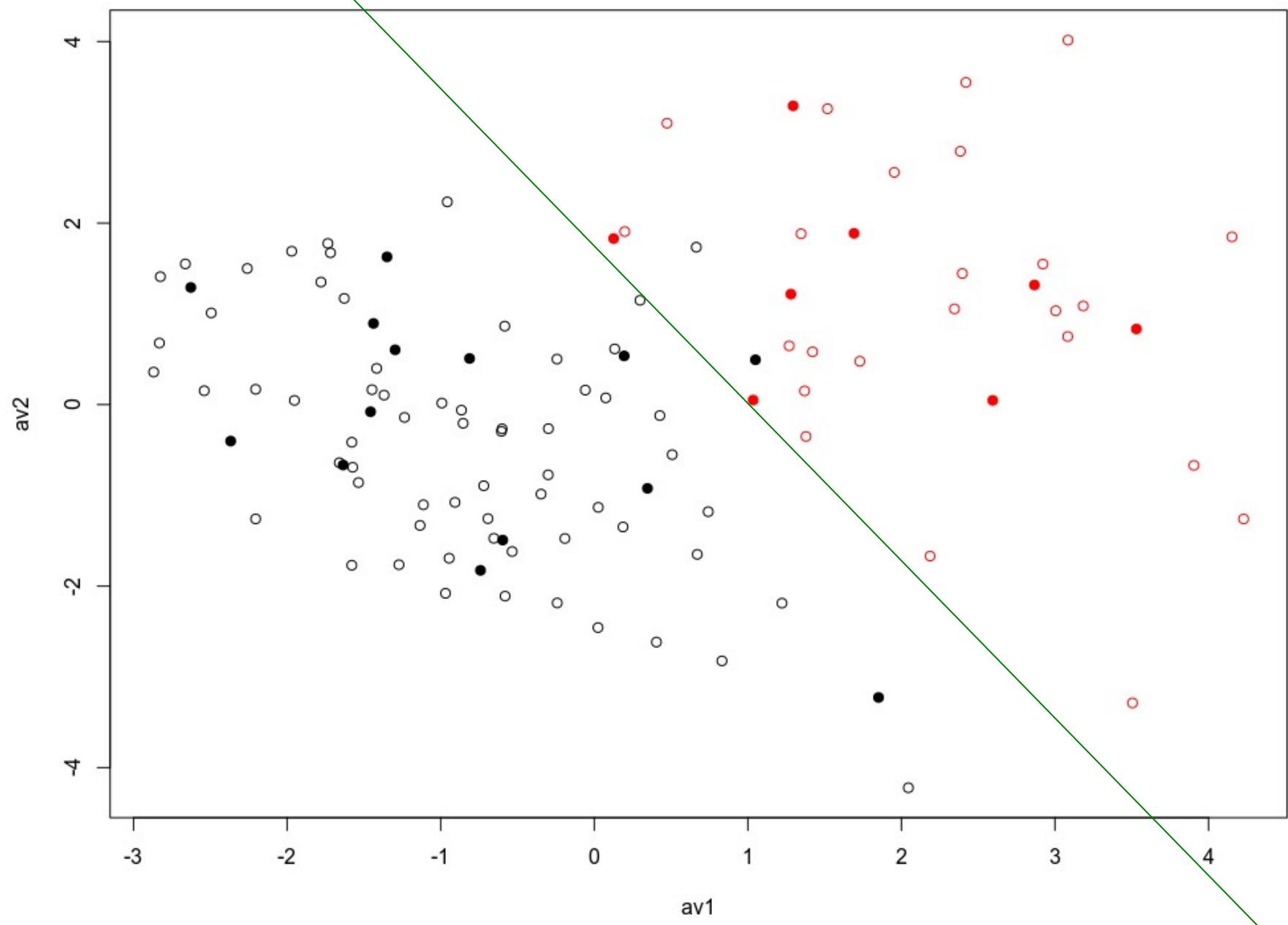
where  $a_b^{(q)}$  is the variable coefficient or loading vector on dimension  $b$  associated to the residual matrix  $X_b^{(q)}$  of the dataset  $X^{(q)}$ .  $C = \{c_{i,j}\}_{i,j}$  is a  $(Q \times Q)$  design matrix that specifies whether datasets should be connected. Elements in  $C$  can be set to zeros when datasets are not connected and ones where datasets are fully connected, as we further describe in Section 2.2. In addition in (1),  $\lambda^{(q)}$  is a non-negative parameter that controls the amount of shrinkage and thus the number of non-zero coefficients in  $a_b^{(q)}$ . Similar to the LASSO (Tibshirani, 1996) and other  $\ell_1$  penalized multivariate models developed for single omics analysis (Lê Cao *et al.*, 2011), the penalization enables the selection of a subset of variables with non-zero coefficients that define each component score  $t_b^{(q)} = X_b^{(q)} a_b^{(q)}$ . The result is the identification of variables that are highly correlated *between* and *within* omics datasets.

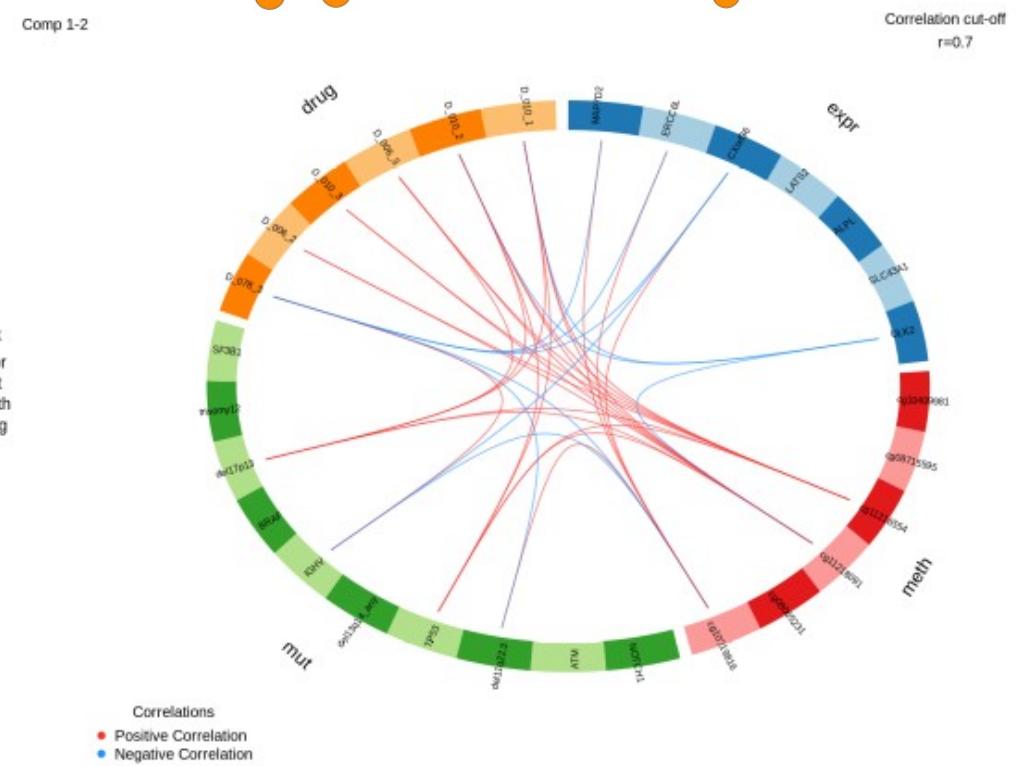
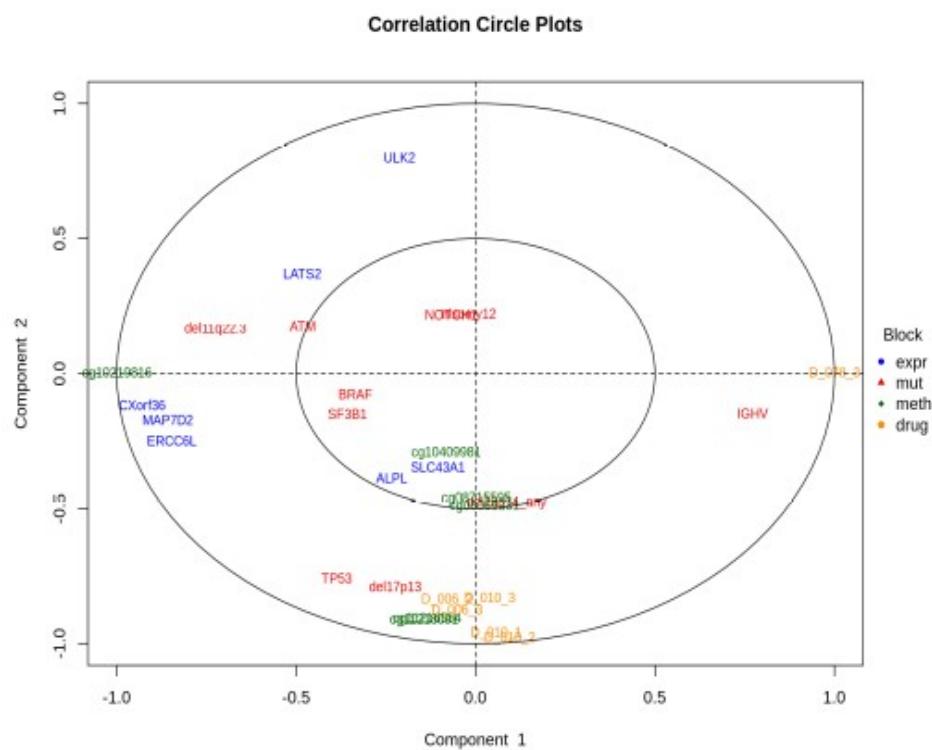
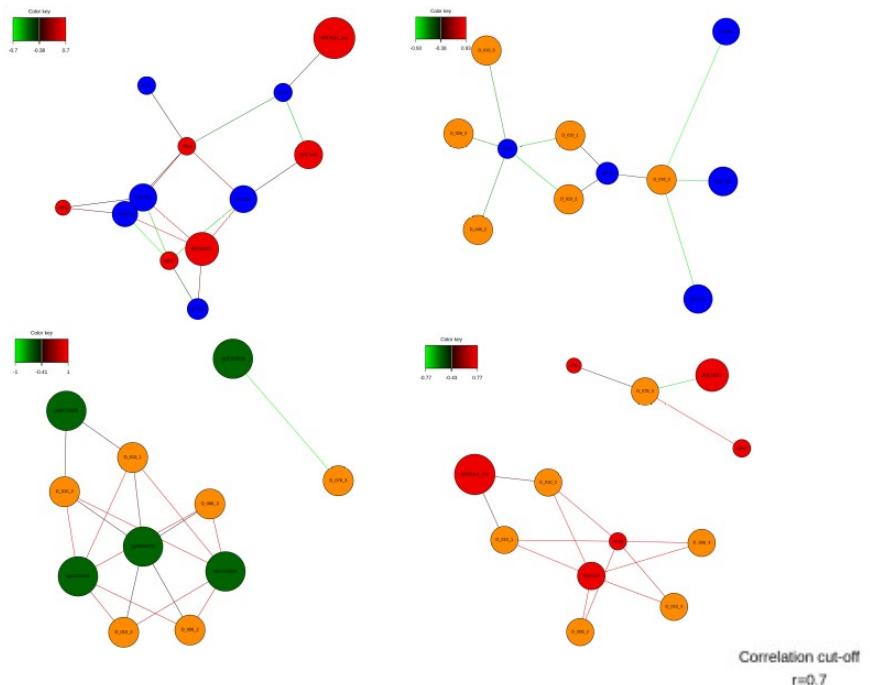
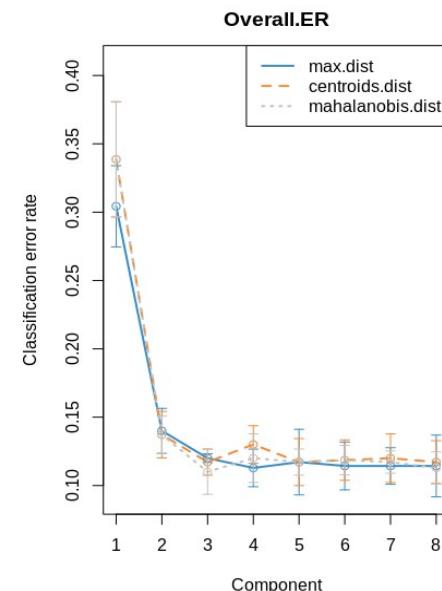
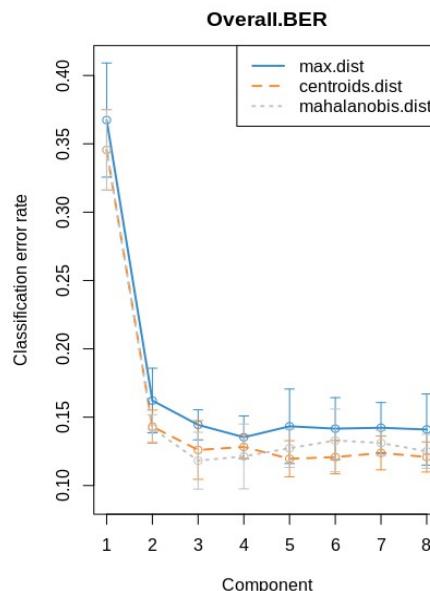


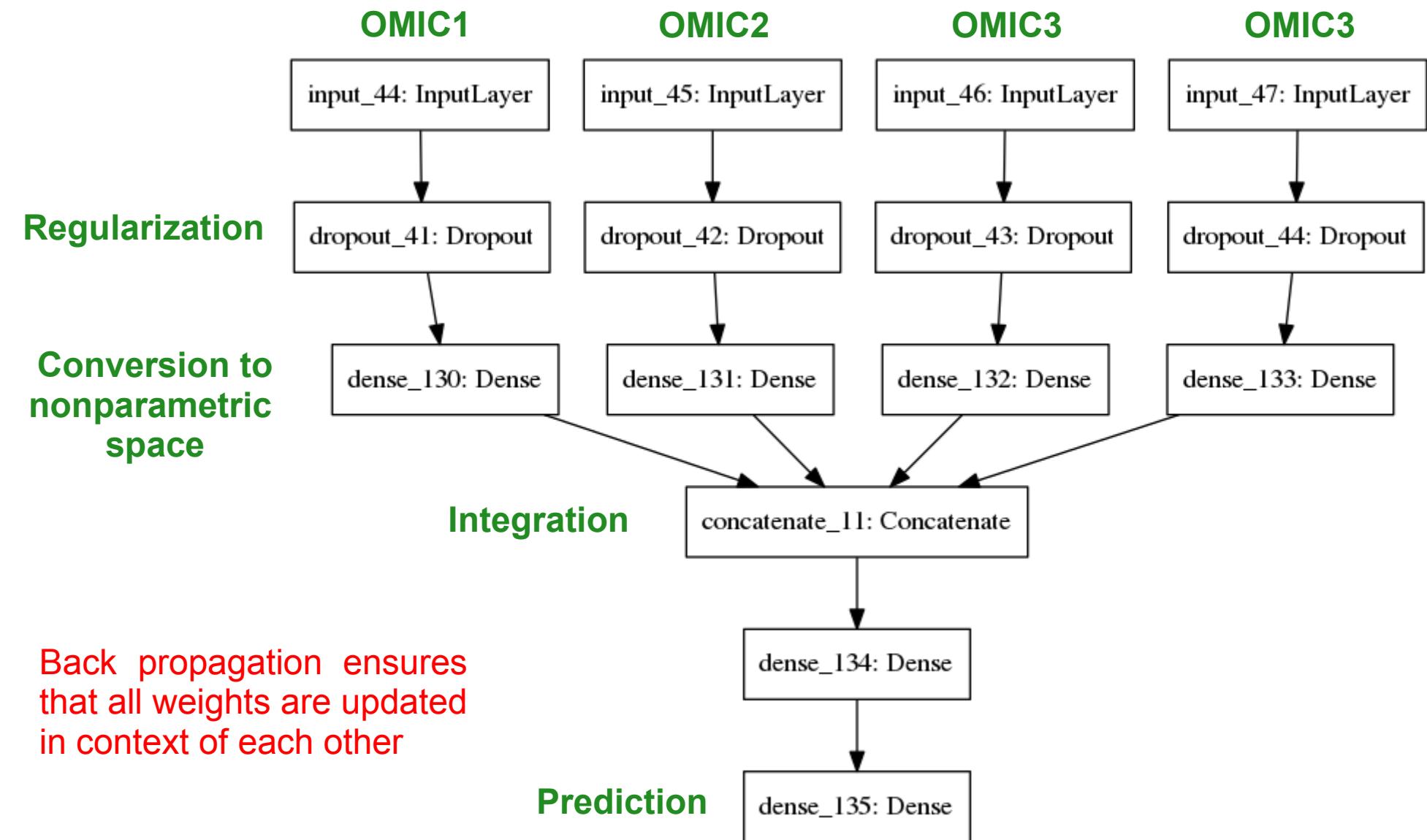
## Chronic Lymphocytic Leukaemia (CLL):

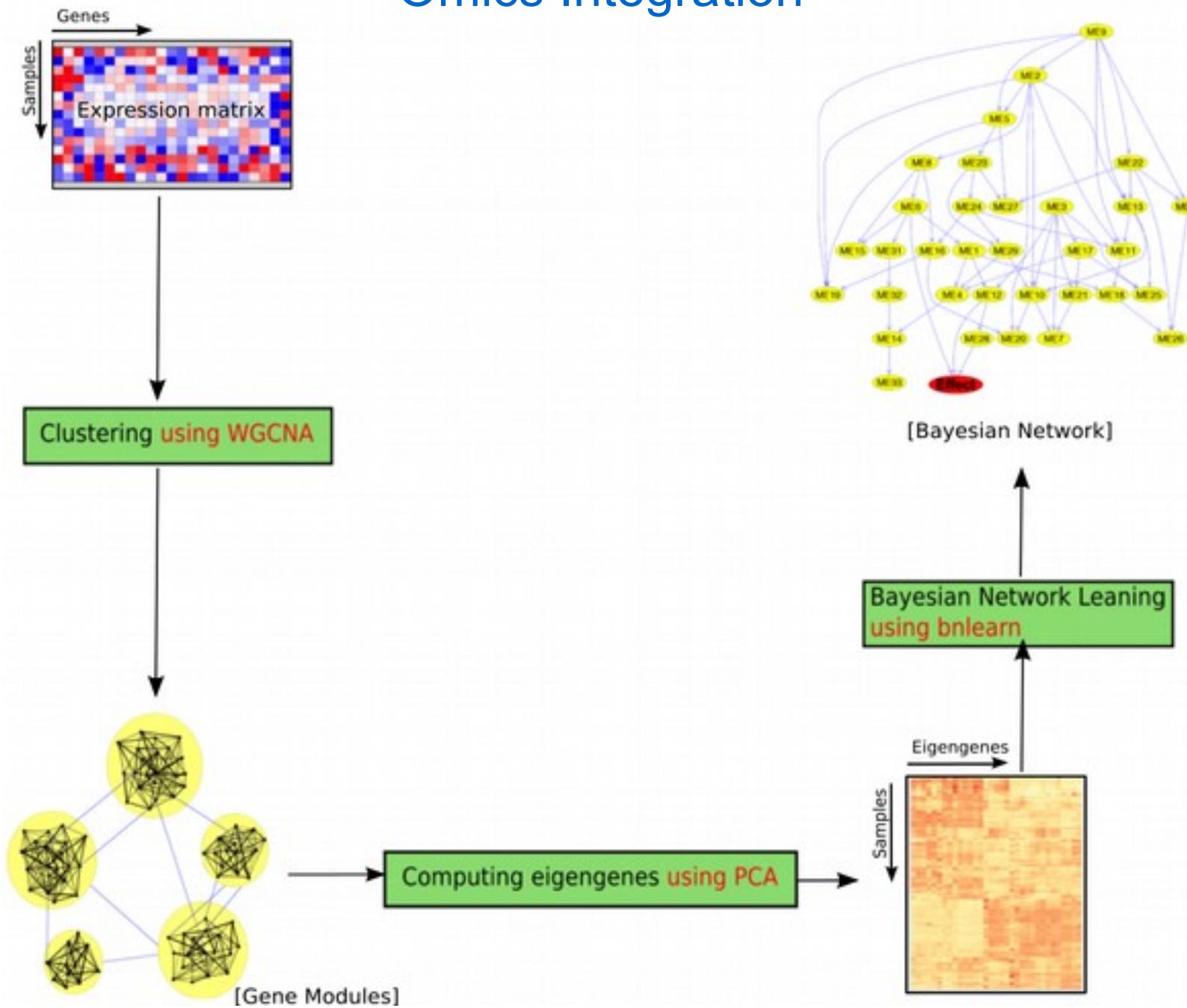
gene expression (RNAseq), mutations, methylation, drug response











## RESEARCH

## Open Access



CrossMark

# Integration of multi-omics data for prediction of phenotypic traits using random forest

Animesh Acharjee<sup>1,3</sup>, Bjorn Kloosterman<sup>1,2</sup>, Richard G. F. Visser<sup>1</sup> and Chris Maliepaard<sup>1\*</sup>

From Statistical Methods for Omics Data Integration and Analysis 2014  
Heraklion, Crete, Greece. 10-12 November 2014

## Abstract

**Background:** In order to find genetic and metabolic pathways related to phenotypic traits of interest, we analyzed gene expression data, metabolite data obtained with GC-MS and LC-MS, proteomics data and a selected set of tuber quality phenotypic data from a diploid segregating mapping population of potato. In this study we present an approach to integrate these ~ omics data sets for the purpose of predicting phenotypic traits. This gives us networks of relatively small sets of interrelated ~ omics variables that can predict, with higher accuracy, a quality trait of interest.

**Results:** We used Random Forest regression for integrating multiple ~ omics data for prediction of four quality traits of potato: tuber flesh colour, DSC onset, tuber shape and enzymatic discoloration. For tuber flesh colour beta-carotene hydroxylase and zeaxanthin epoxidase were ranked first and forty-fourth respectively both of which have previously been associated with flesh colour in potato tubers. Combining all the significant genes, LC-peaks, GC-peaks and proteins, the variation explained was 75 %, only slightly more than what gene expression or LC-MS data explain by themselves which indicates that there are correlations among the variables across data sets. For tuber shape regressed on the gene expression, LC-MS, GC-MS and proteomics data sets separately, only gene expression data was found to explain significant variation. For DSC onset, we found 12 significant gene expression, 5 metabolite levels (GC) and 2 proteins that are associated with the trait. Using those 19 significant variables, the variation explained was 45 %. Expression QTL (eQTL) analyses showed many associations with genomic regions in chromosome 2 with also the highest explained variation compared to other chromosomes. Transcriptomics and metabolomics analysis on enzymatic discoloration after 5 min resulted in 420 significant genes and 8 significant LC metabolites, among which two were putatively identified as caffeoylquinic acid methyl ester and tyrosine.

**Conclusions:** In this study, we made a strategy for collecting and integrating multiple ~ omics data using random



# National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab



*Knut och Alice  
Wallenbergs  
Stiftelse*



LUNDS  
UNIVERSITET