Feature Selection and Supervised OMICs Integration

ISMB / ECCB 2021

Tutorial 4: A practical introduction to multi-omics integration and network analysis

Nikolay Oskolkov, NBIS SciLifeLab, 22.07.2021

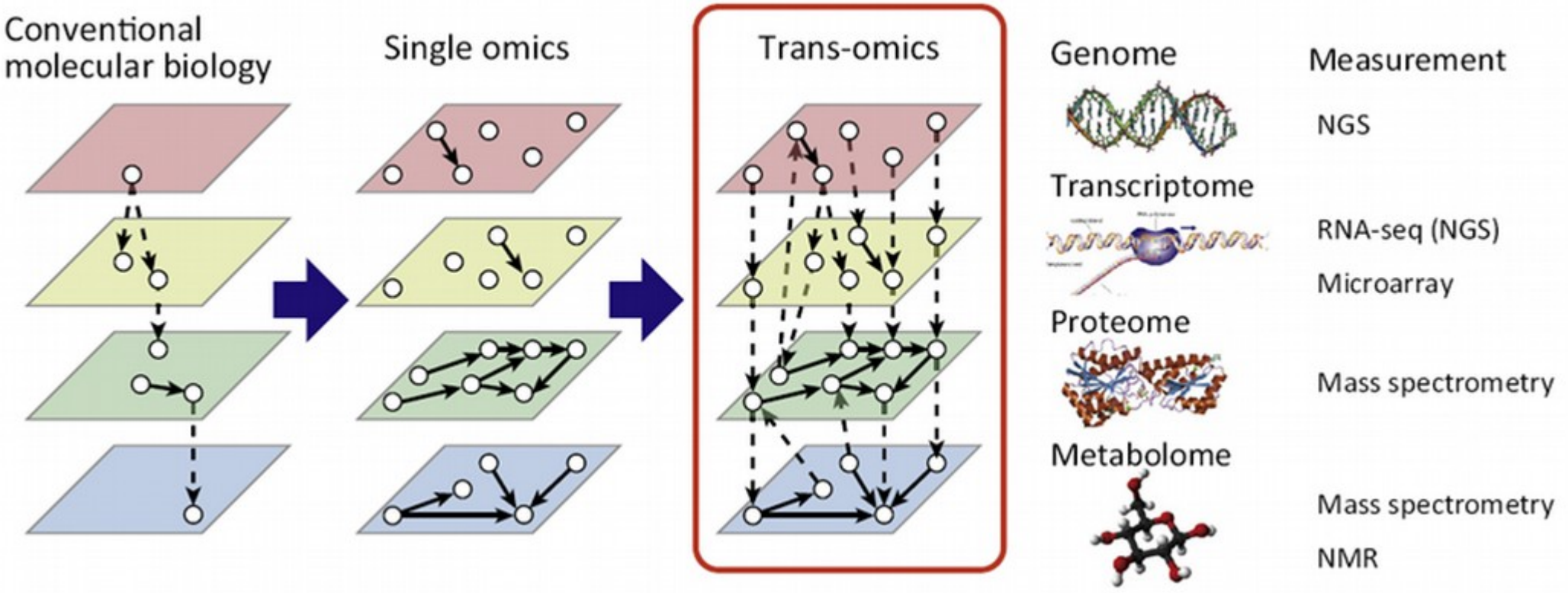


Image adapted from Yugi et al., Trends Biotechnol. 2016 Apr;34(4):276–290

Protocol of OMICs Integration

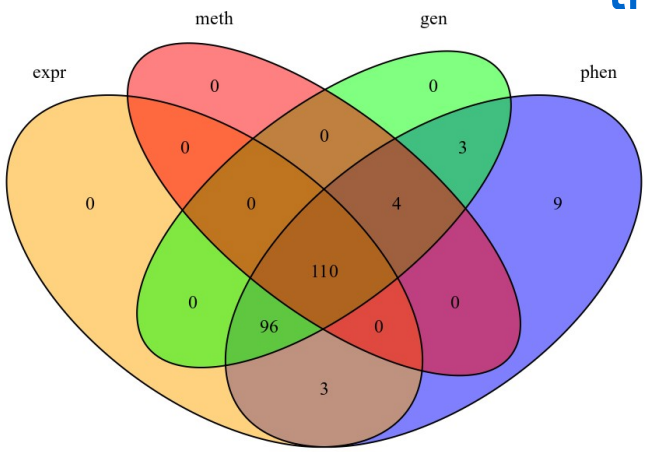
What Method of Integration to Select?

	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

Feature Pre-Selection to Overcome the Curse of Dimensionality



29% of Diabetics

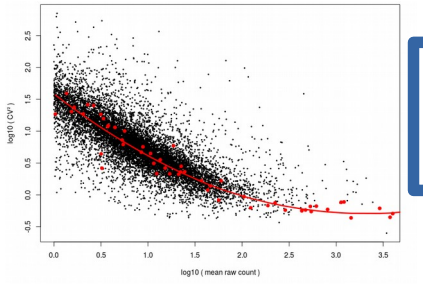
Data Set (4 OMICs)

~~WGS (~30 mln dims)~~
~~BSseq (~30 mln dims)~~

Train Set (n = 88)

Test Set (n = 22)

unsupervised



supervised

Feature Pre-Selection

Evaluation

OMICs Integration

Trained Model

Protocol of Integrative OMICS Analysis

1) Check that there is a relation between the OMICs (MOFA)

2) Choose integrative model based on amount of data and goal (linear, supervised)

3) Do feature pre-selection (supervised or unsupervised) on train data set

4) Integrate the OMICs using your favorite model chosen in 2) on train data set

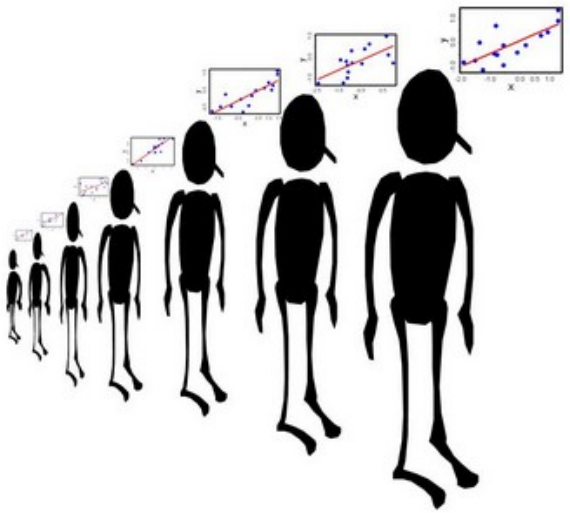
5) Compare prediction of integrative model with predictions from individual OMICs

Maximum Likelihood does not Stand Non-Independence in Data

```
##            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

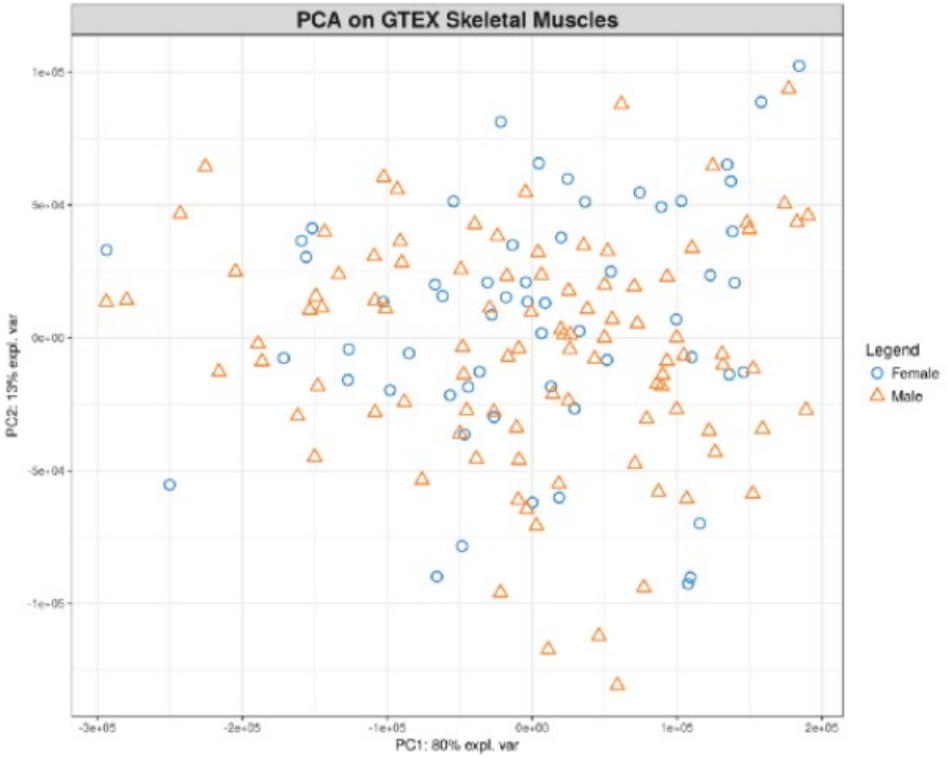
Image adapted from http://anythingbutrbitrary.blogspot.com/2012/06/random-regression-coefficients-using.html

Univariate and Multivariate Feature Selection

Univariate Feature Selection

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

ReadGTEX.R hosted with ❤ by GitHub view raw




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

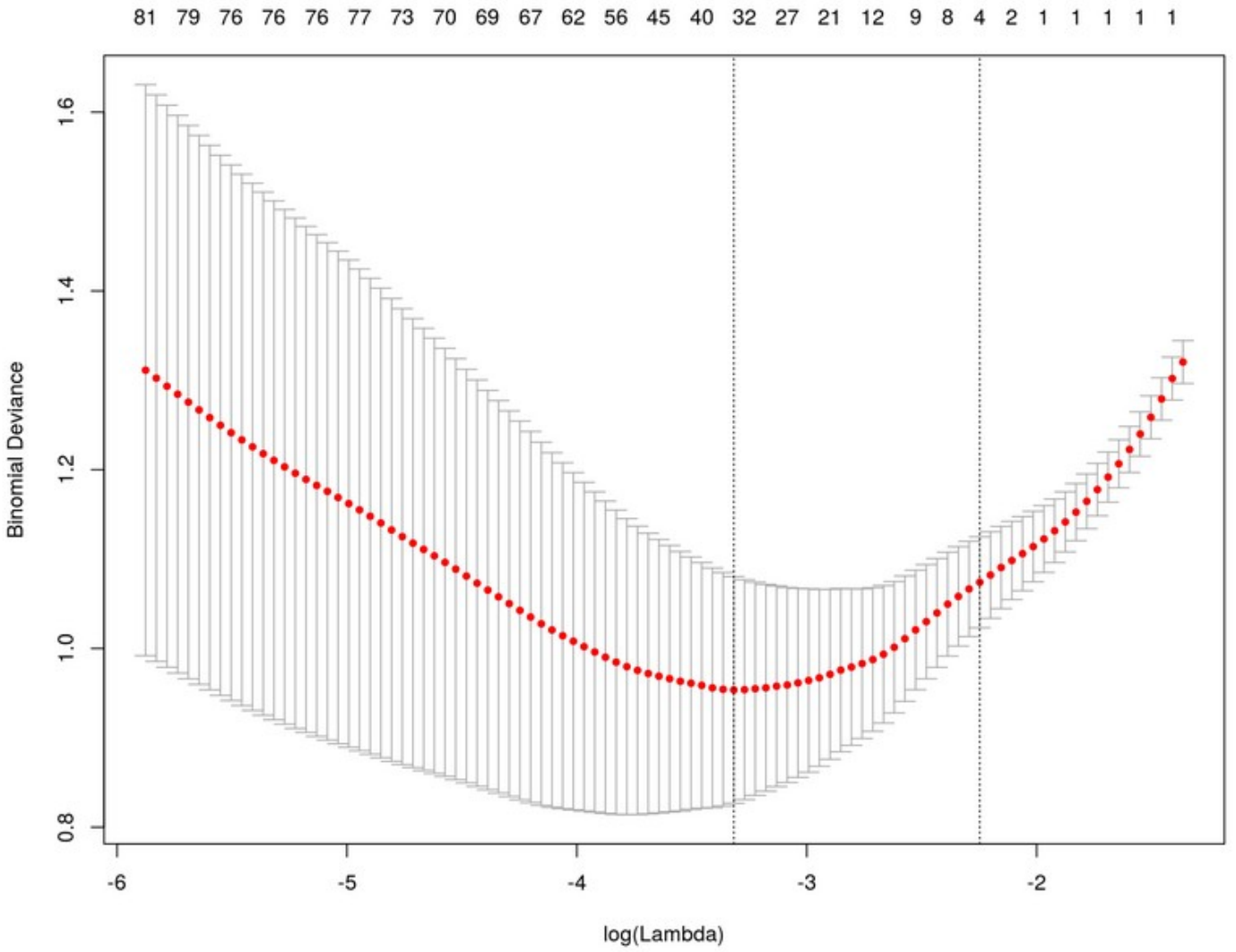
Generally acknowledged that univariate feature selection has a poor predictive capacity compared to multivariate feature selection

Multivariate Feature Selection: LASSO

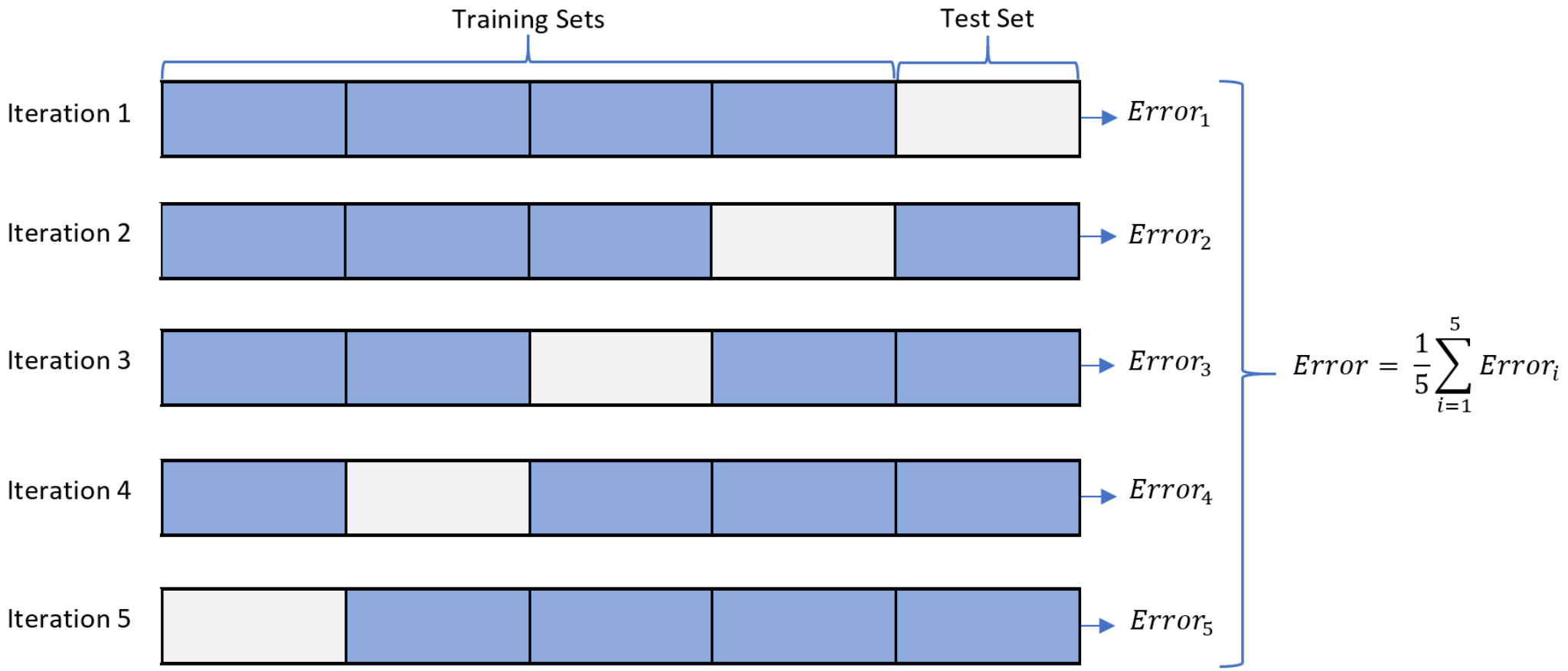
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

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

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



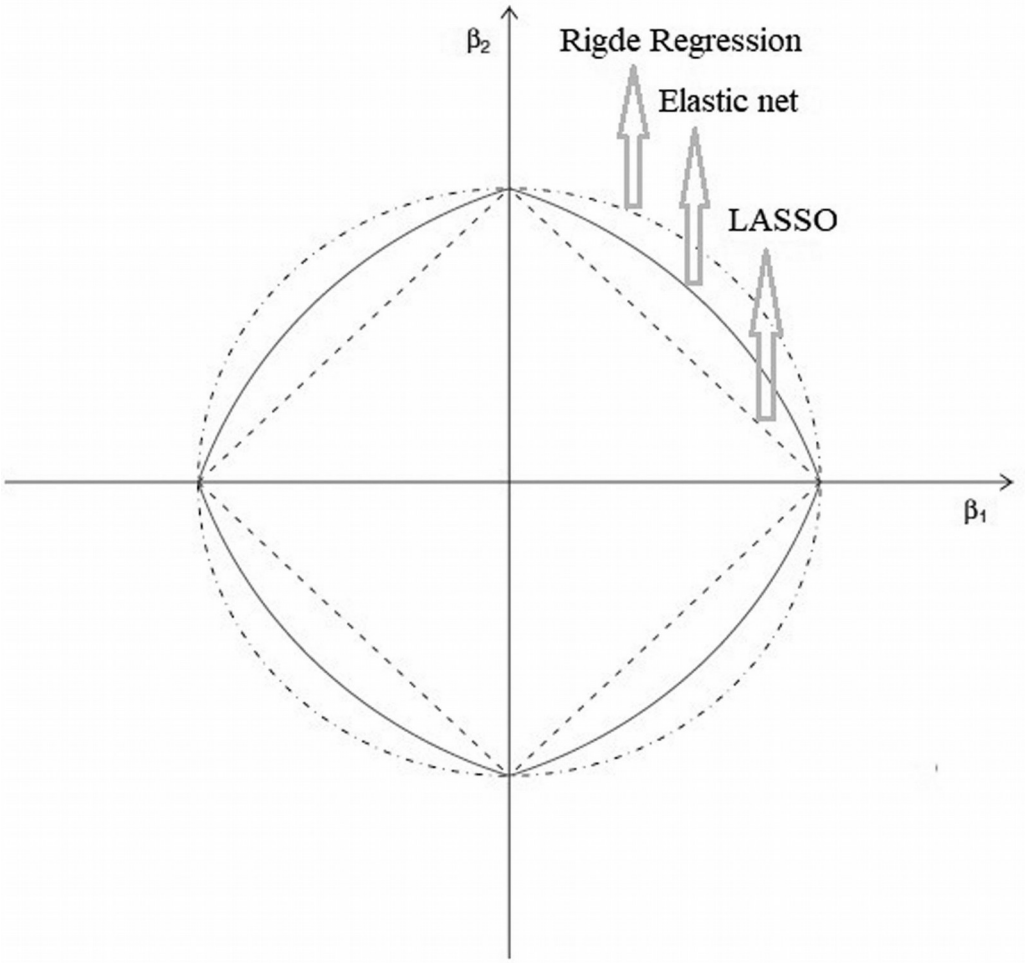
K-fold Cross Validation



Cross-validation is a standard way to tune model hyperparameters such as λ in LASSO

Image adapted from https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b

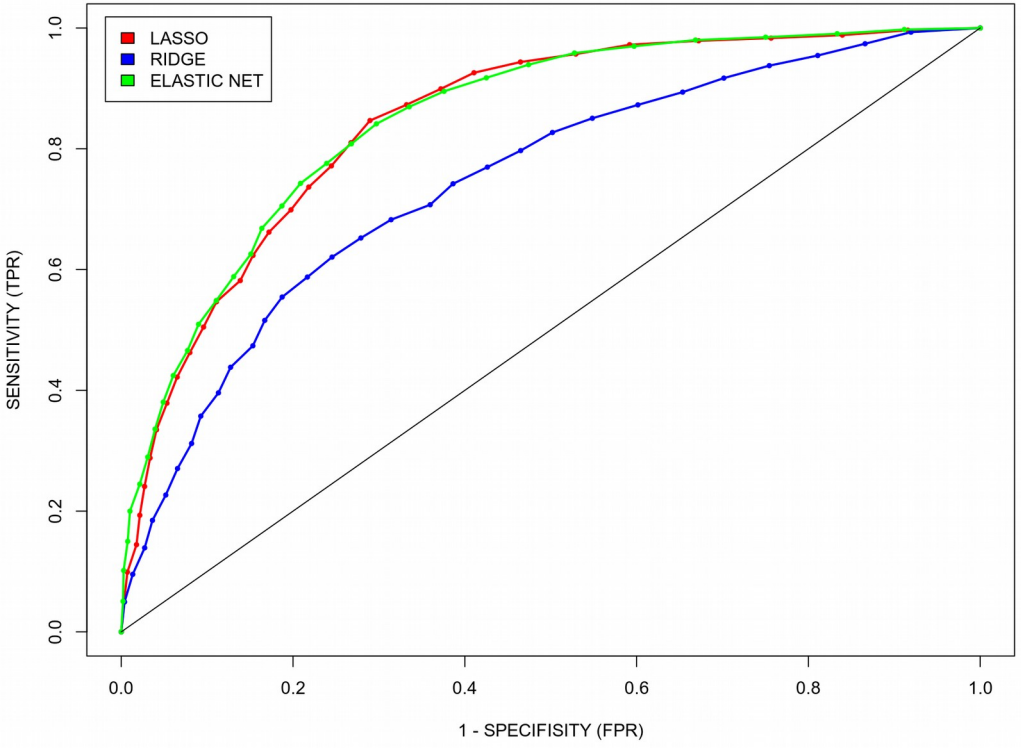
Difference between Lasso, Ridge, and Elastic Net



Zou and Hastie, J R Stat Soc Ser B, 2005

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

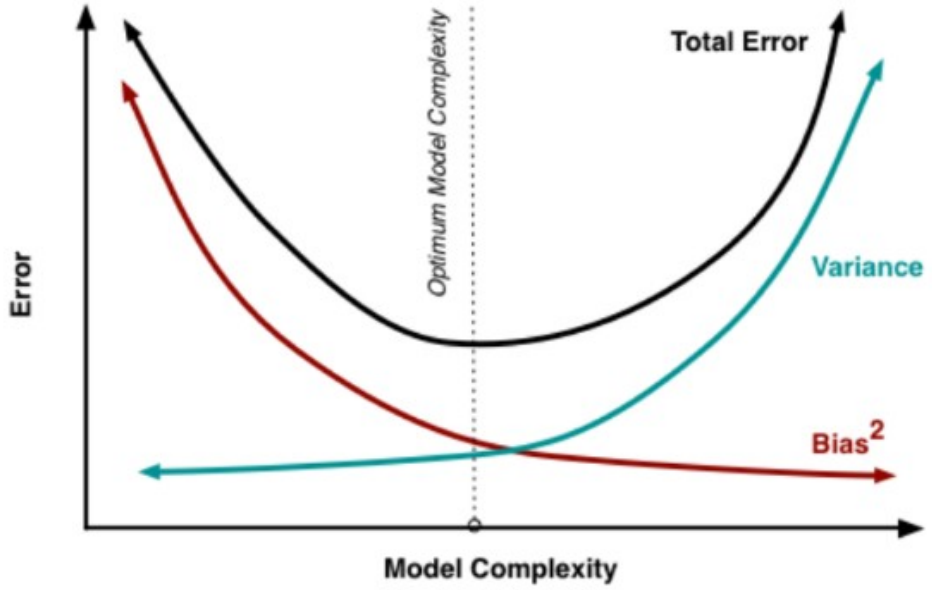
$$\text{Ridge} : \beta_1^2 + \beta_2^2 \leq \lambda$$

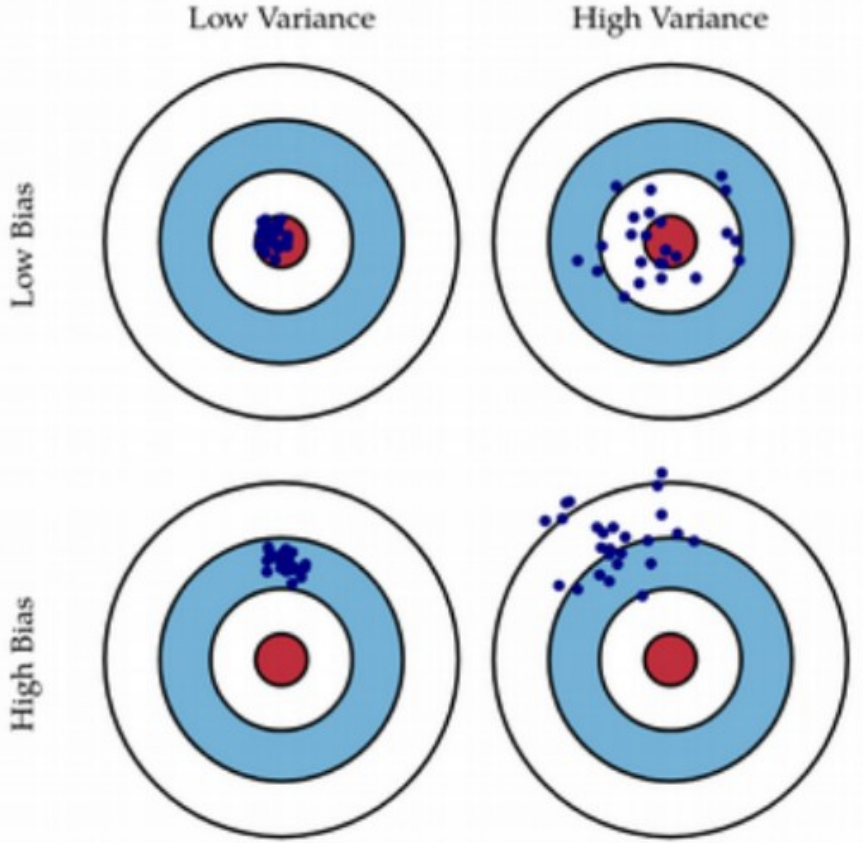


Lasso is more conservative

Ridge is more permissive

Penalized Regression Interpretation





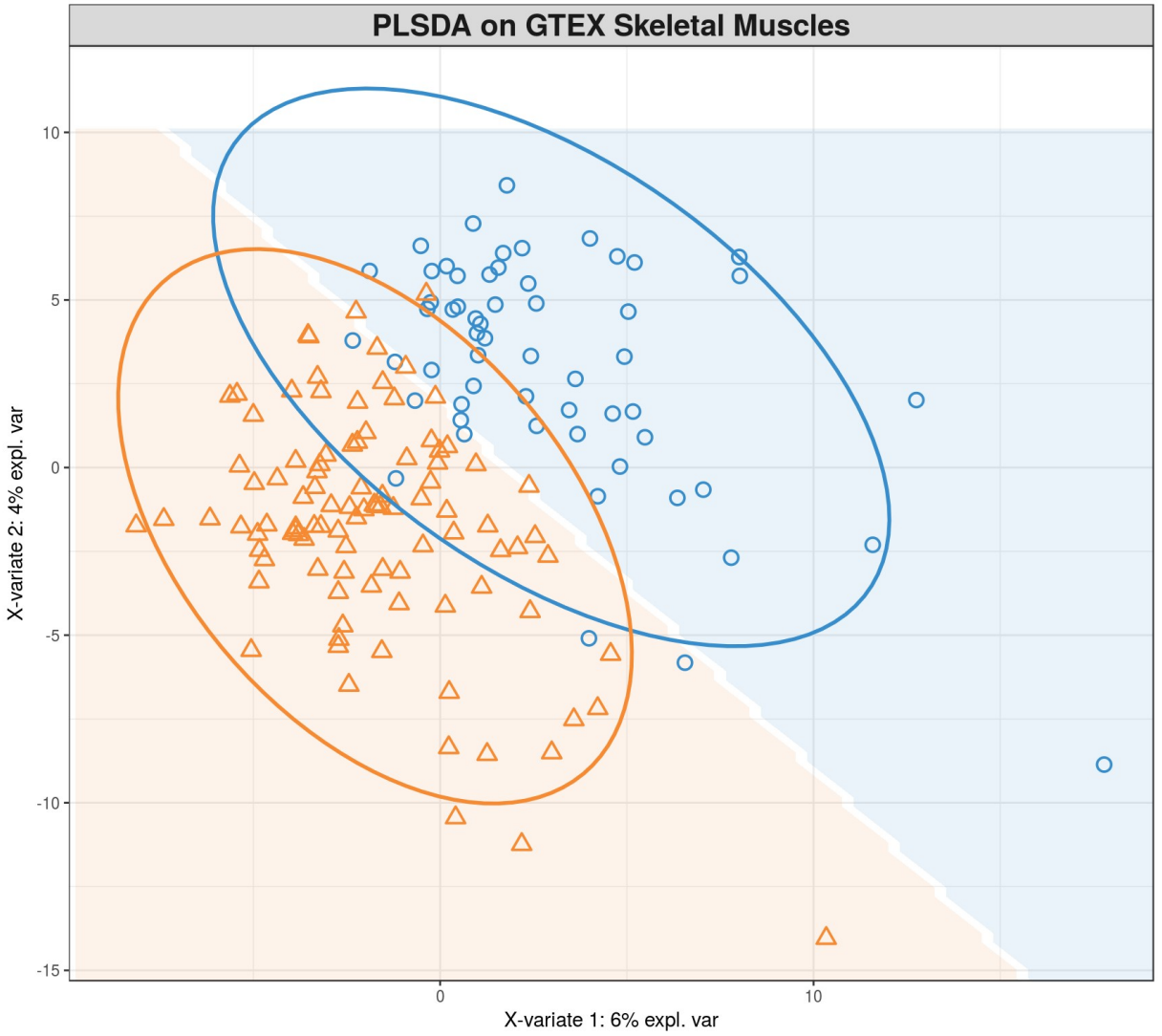
$$Y = f(X) \Longrightarrow \text{Reality}$$

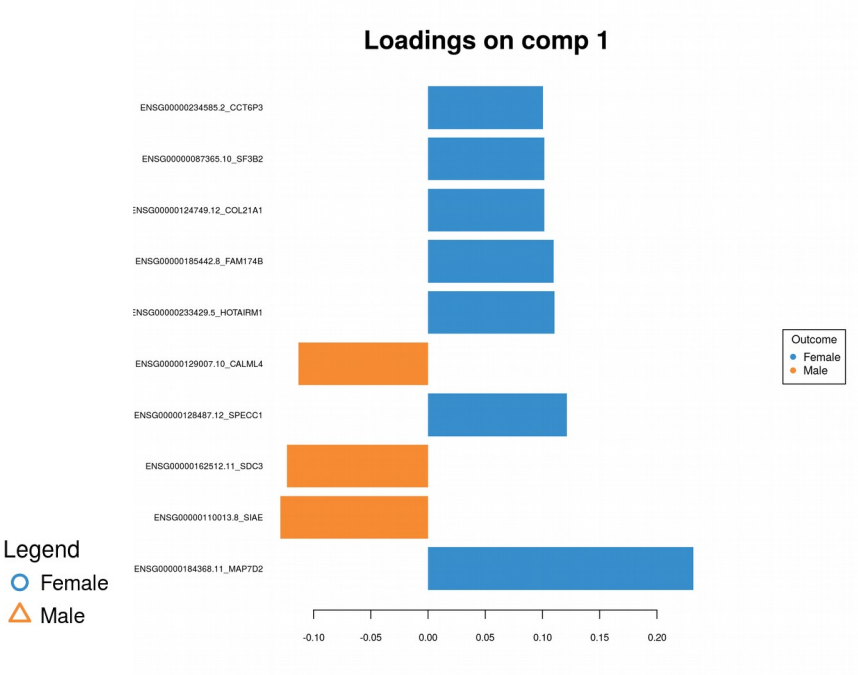
$$\mathrm{Y} = \hat{\mathrm{f}}\,(\mathrm{X}) + \text{Error} \Longrightarrow \text{Model}$$

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

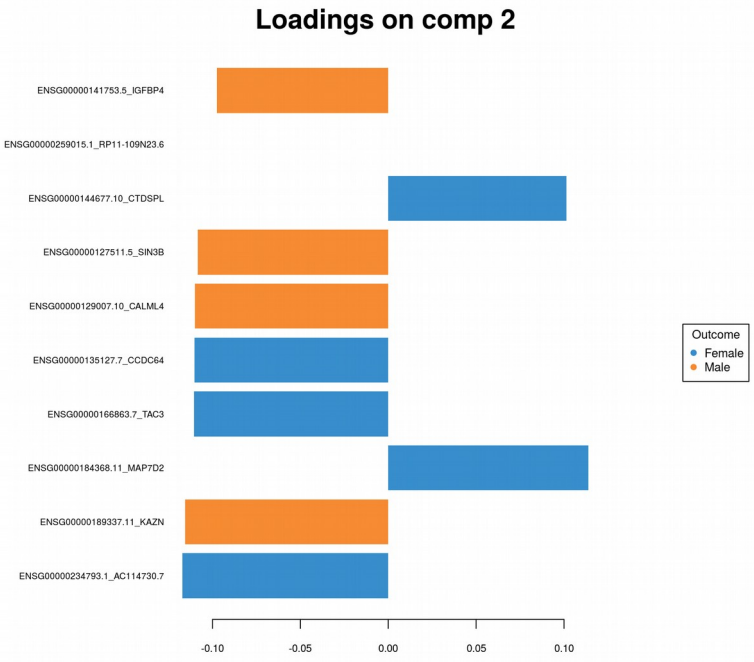
Images adapted from https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

Multivariate Feature Selection: PLS



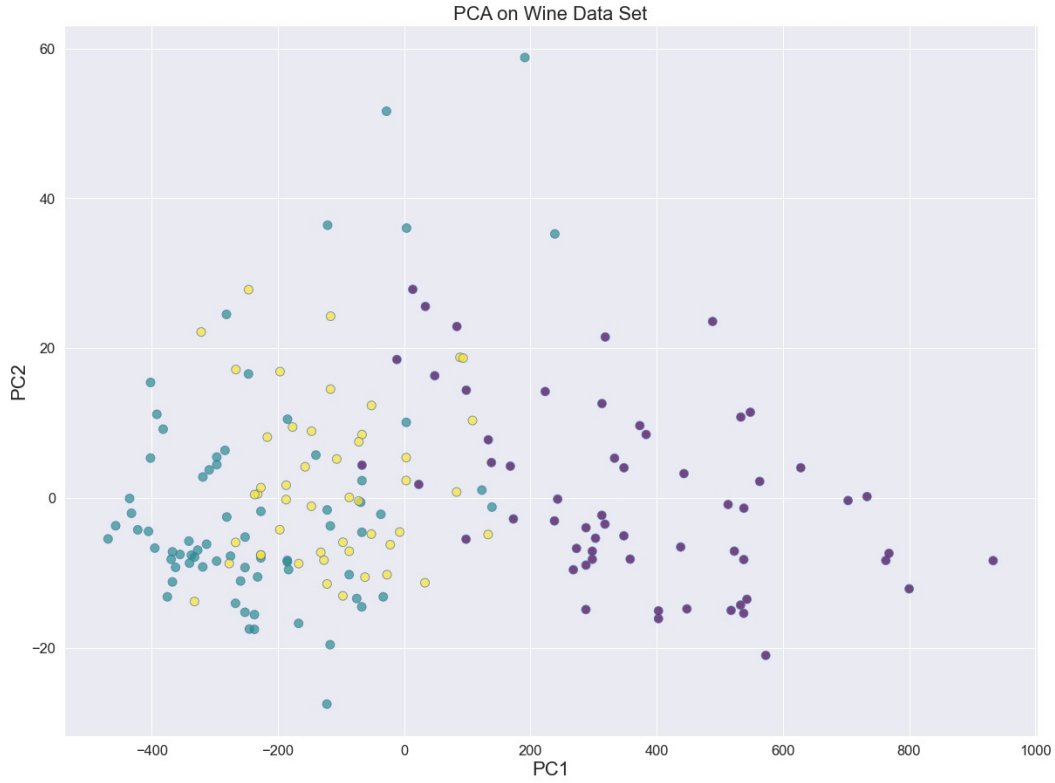


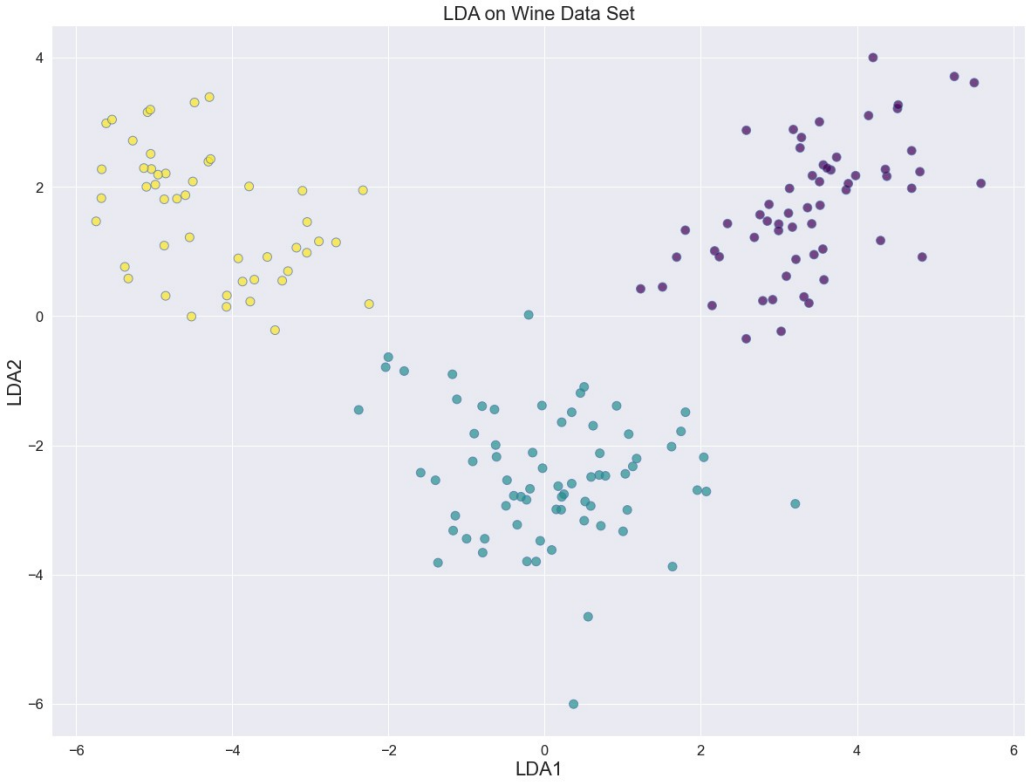
Legend
- Female
- Male



Select features that separate two groups of samples the most

Multivariate Feature Selection: Linear Discriminant Analysis (LDA)

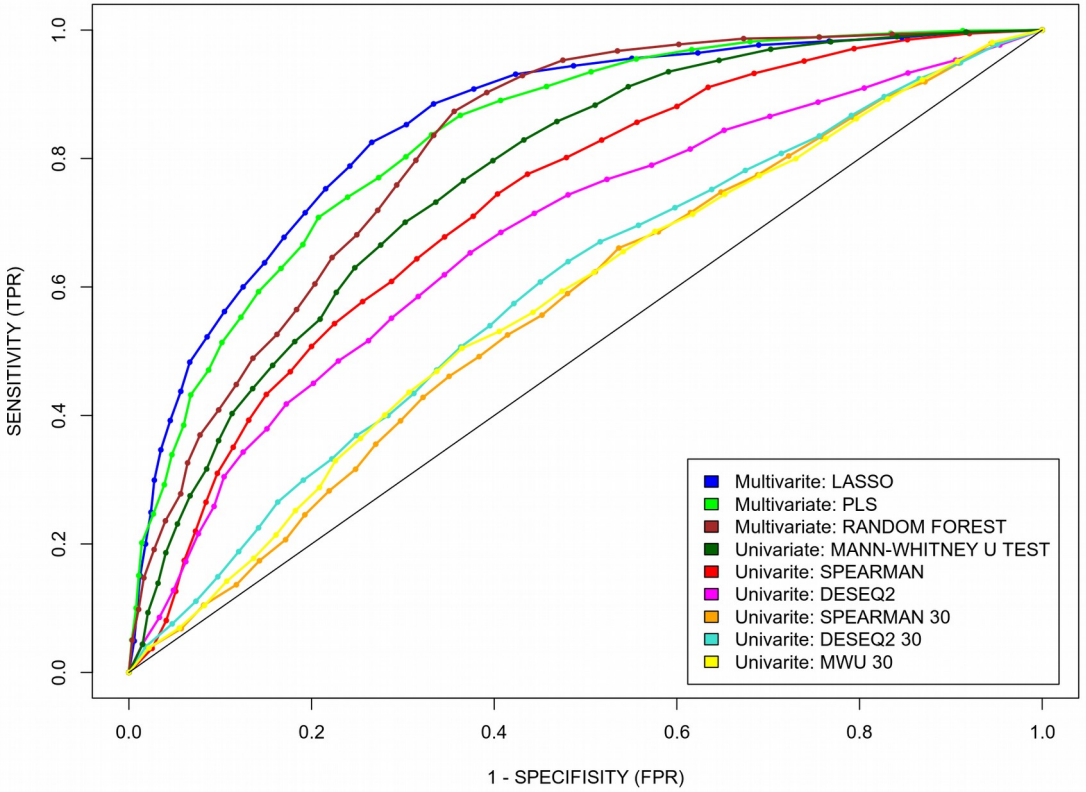




Minimize variance within clusters and maximize variance between clusters

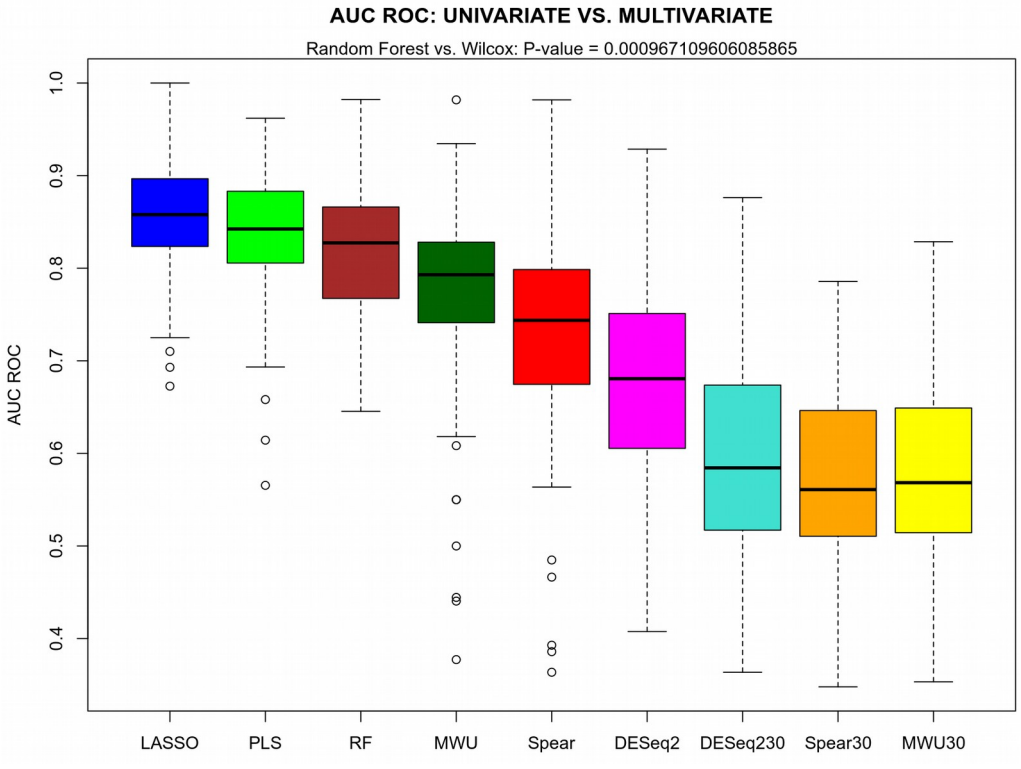
Similar to what ANOVA / t-test is doing, therefore LINEAR Discriminant Analysis (LDA)

Univariate vs. Multivariate Prediction



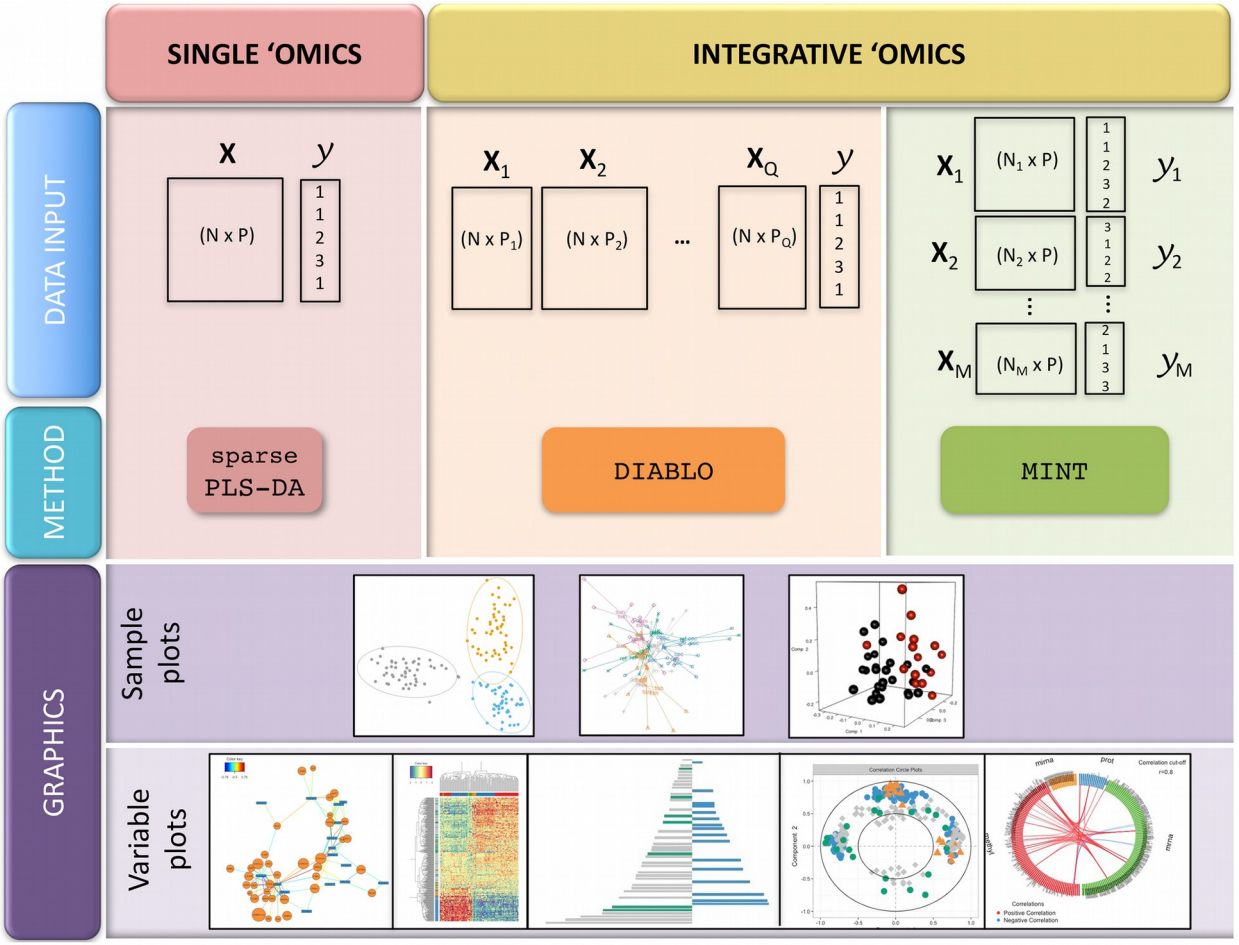
Multivariate methods (LASSO, PLS, RanFor) have significantly higher AUC ROC than univariate methods (Spear, MWU, DESeq2) on skeletal muscle gene expression data

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



DIABLO Omics Integration

DIABLO PLS-Based Algorithm



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

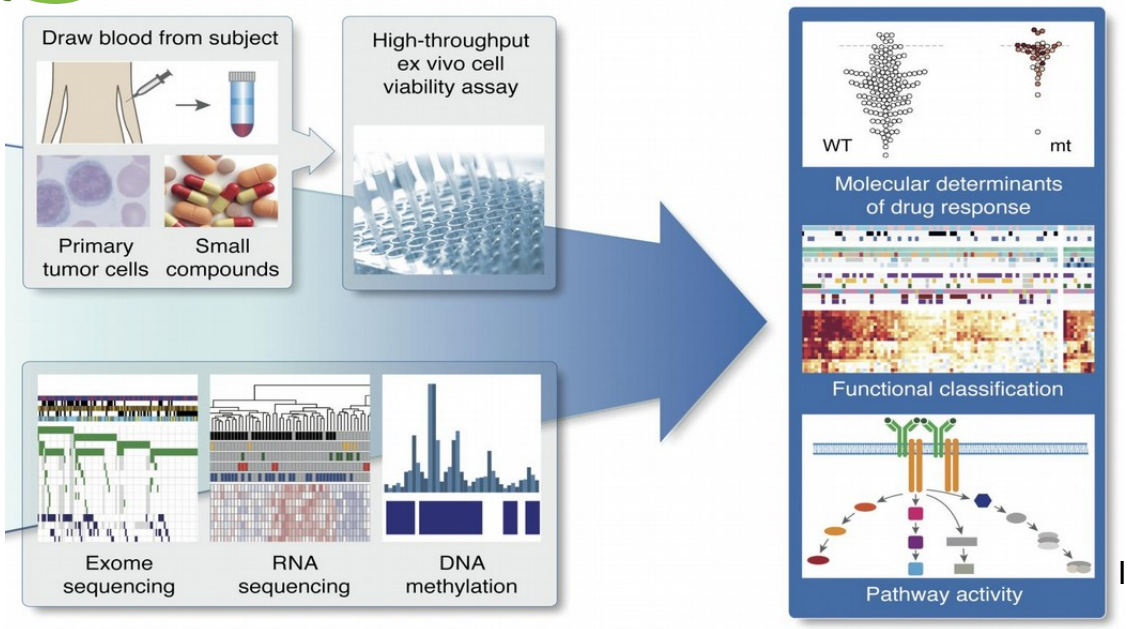
$$\max_{a_h^{(1)},\ldots,a_h^{(Q)}} \sum_{i,j=1, i\neq j}^{Q} c_{i,j}\ \mathrm{cov}(X_h^{(i)} a_h^{(i)}, X_h^{(j)} a_h^{(j)}), \quad (1)$$

$$\text{s.t. } ||a_h^{(q)}||_2 = 1 \text{ and } ||a_h^{(q)}||_1 \leq \lambda^{(q)} \text{ for all } 1 \leq q \leq Q$$

where $a_h^{(q)}$ is the variable coefficient or loading vector on dimension h associated to the residual matrix $X_h^{(q)}$ 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 $a_h^{(q)}$. Similar to the LASSO (Tibshirani, 1996) and other ℓ_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 $t_h^{(q)} = X_h^{(q)} a_h^{(q)}$. The result is the identification of variables that are highly correlated *between* and *within* omics datasets.

$$\max_{\beta} \mathrm{cov}(X, Y) \Longrightarrow \hat{\beta}$$

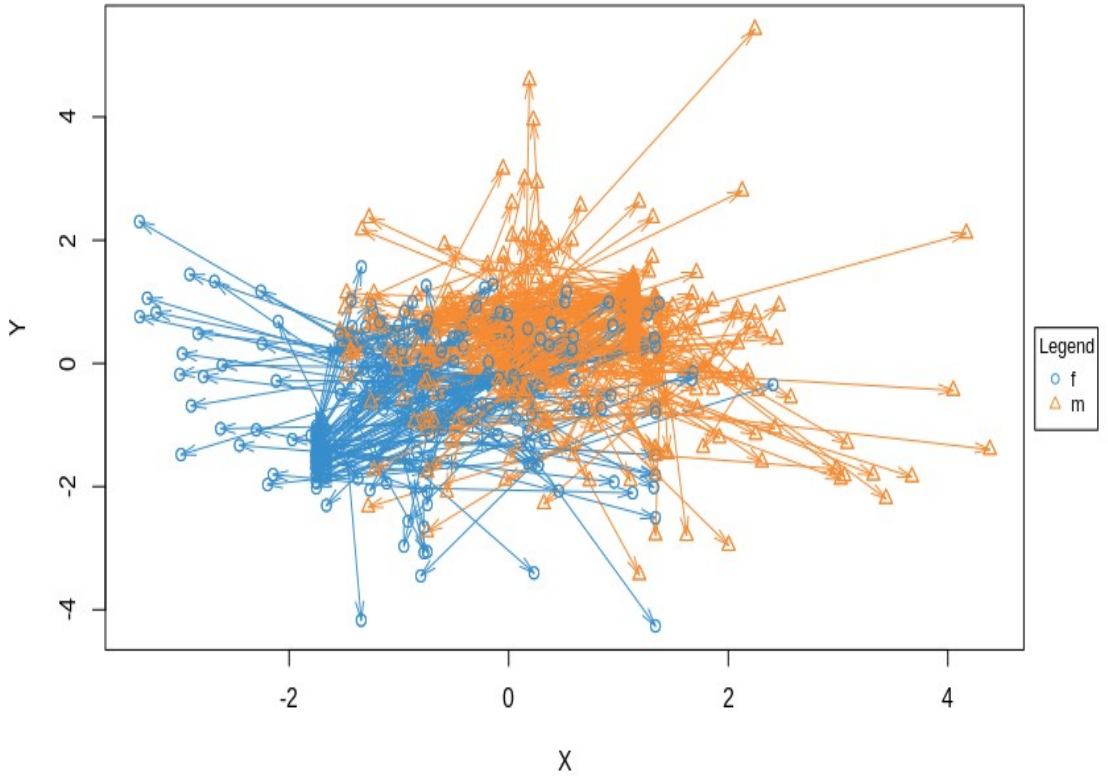
DIABLO Omics Integration

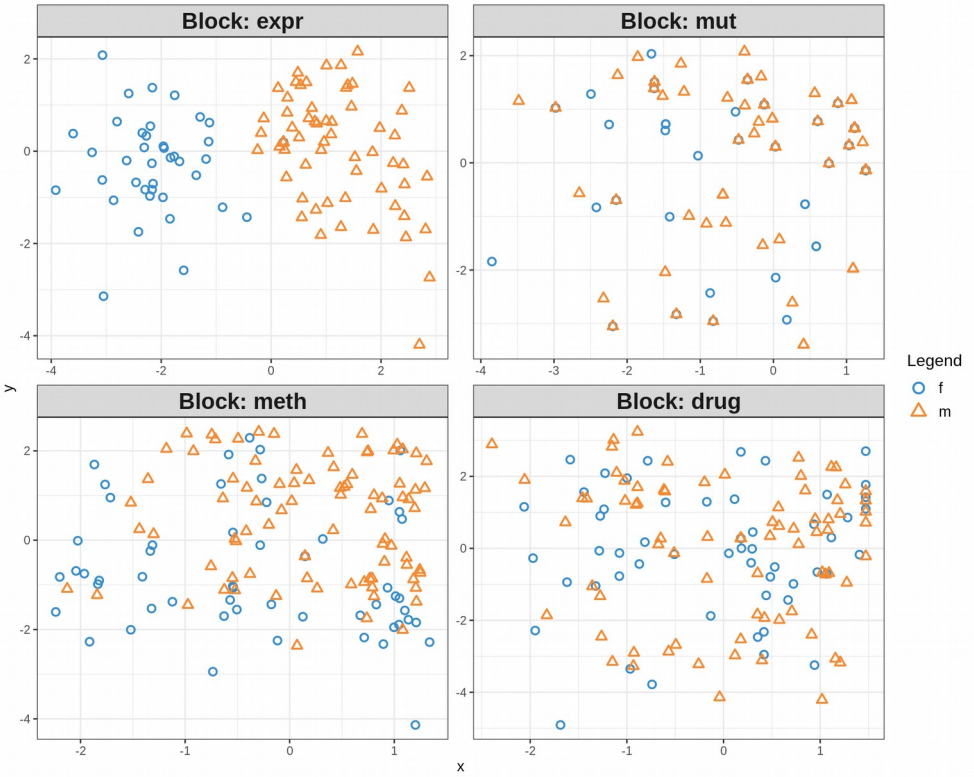


Chronic Lymphocytic Leukaemia (CLL):

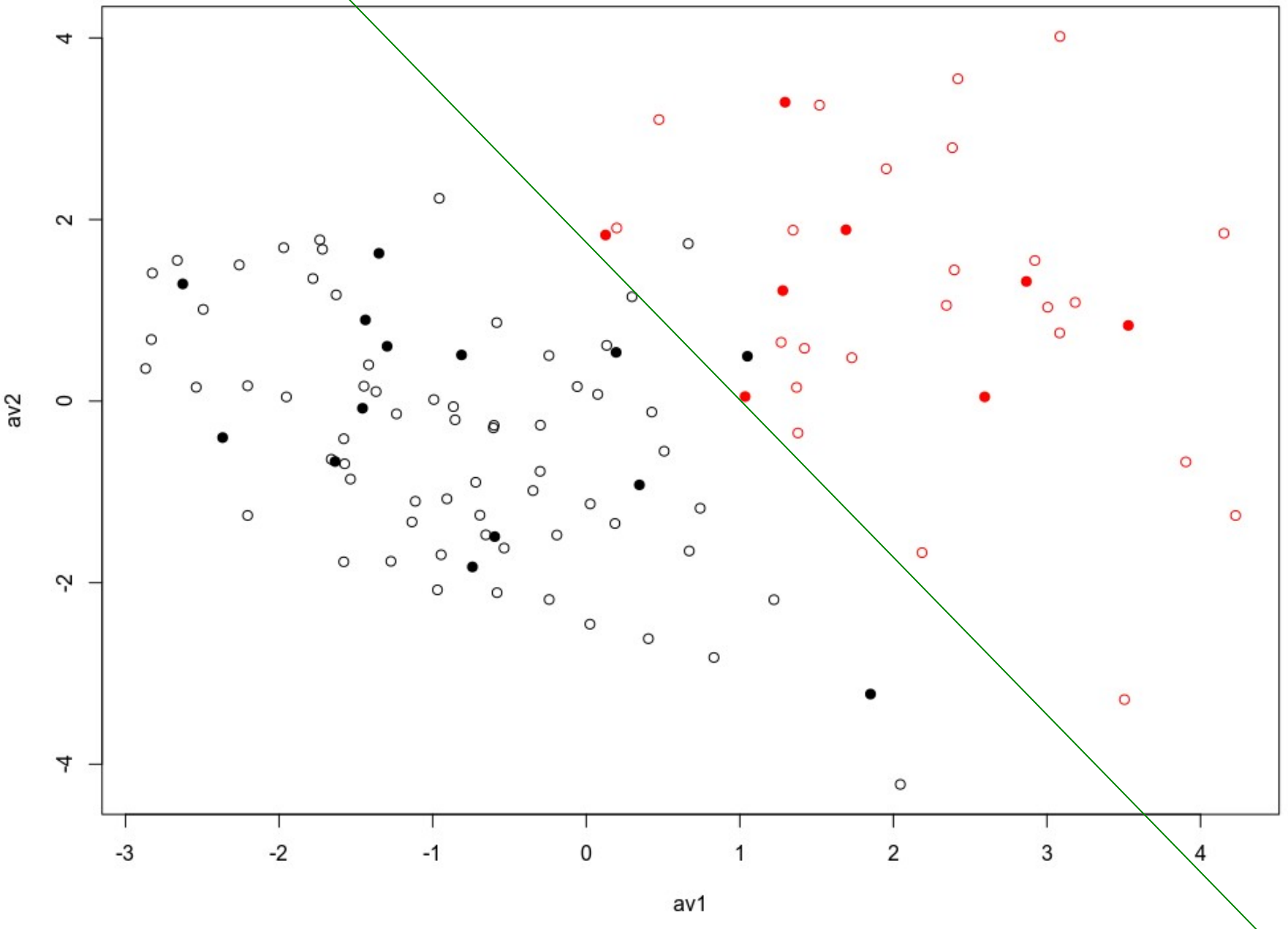
gene expression (RNAseq), mutations, methylation, drug response

Image adapted from Dietrich et al., J. Clin. Inv. 2017

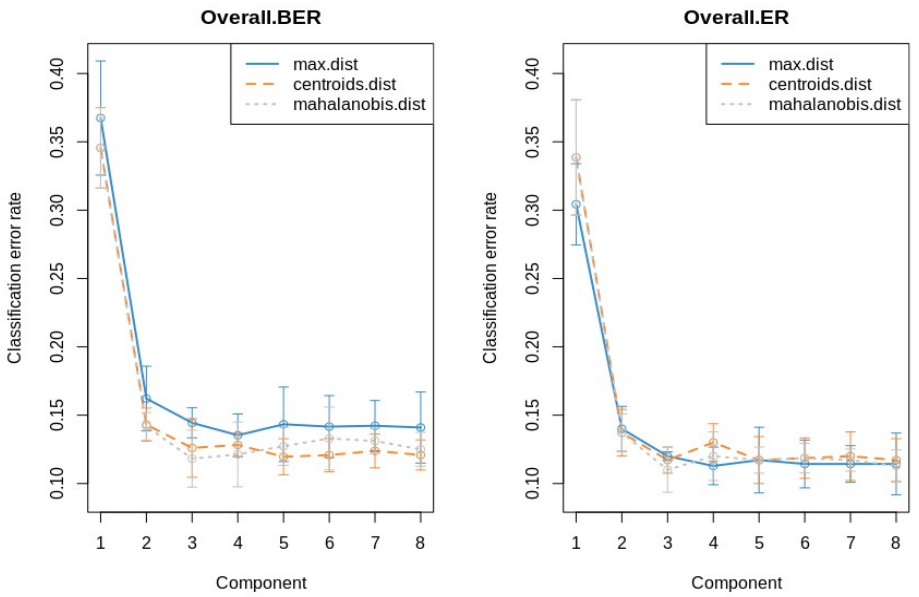


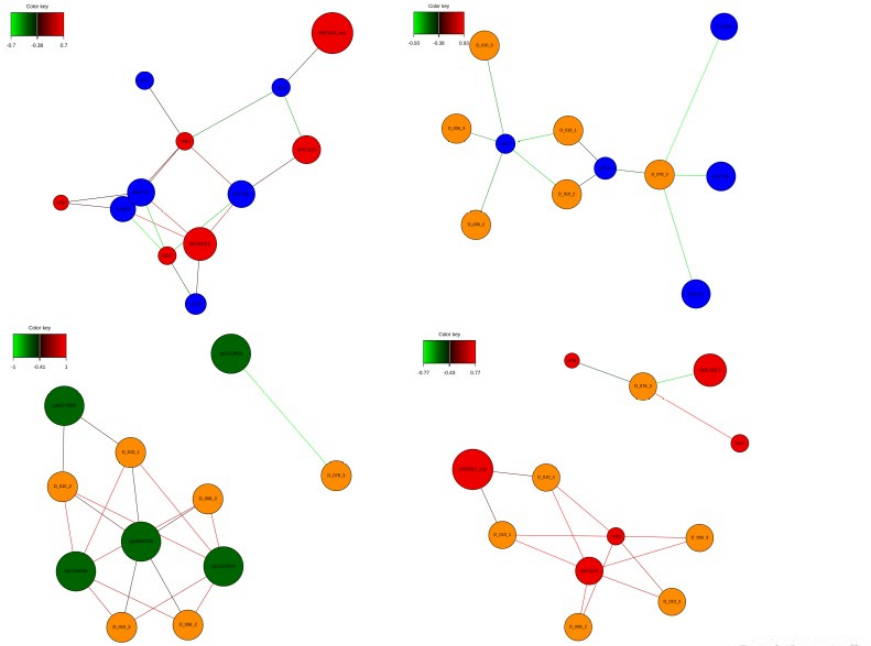


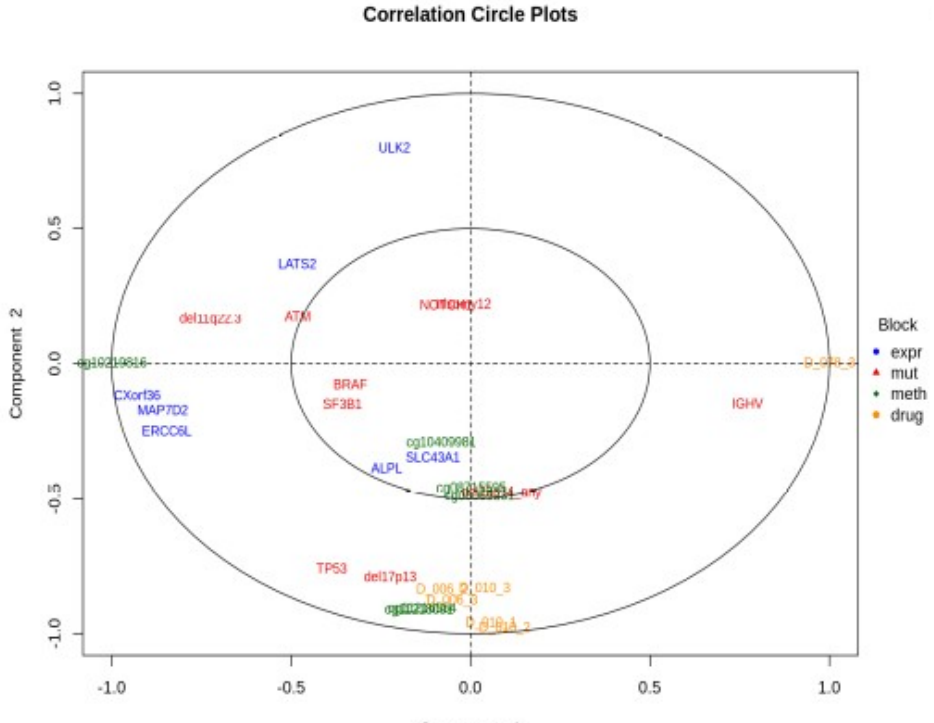
Prediction on Test Data Set

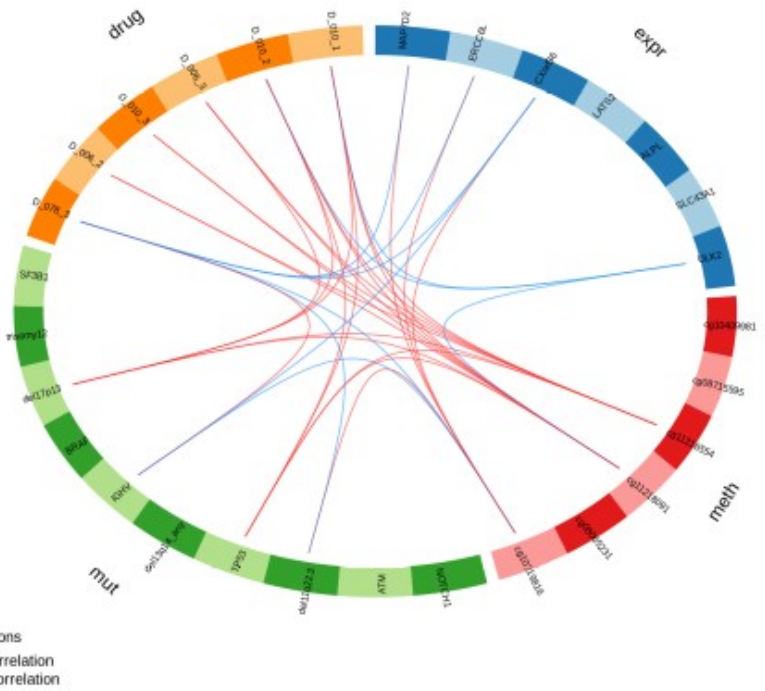


DIABLO Visualization

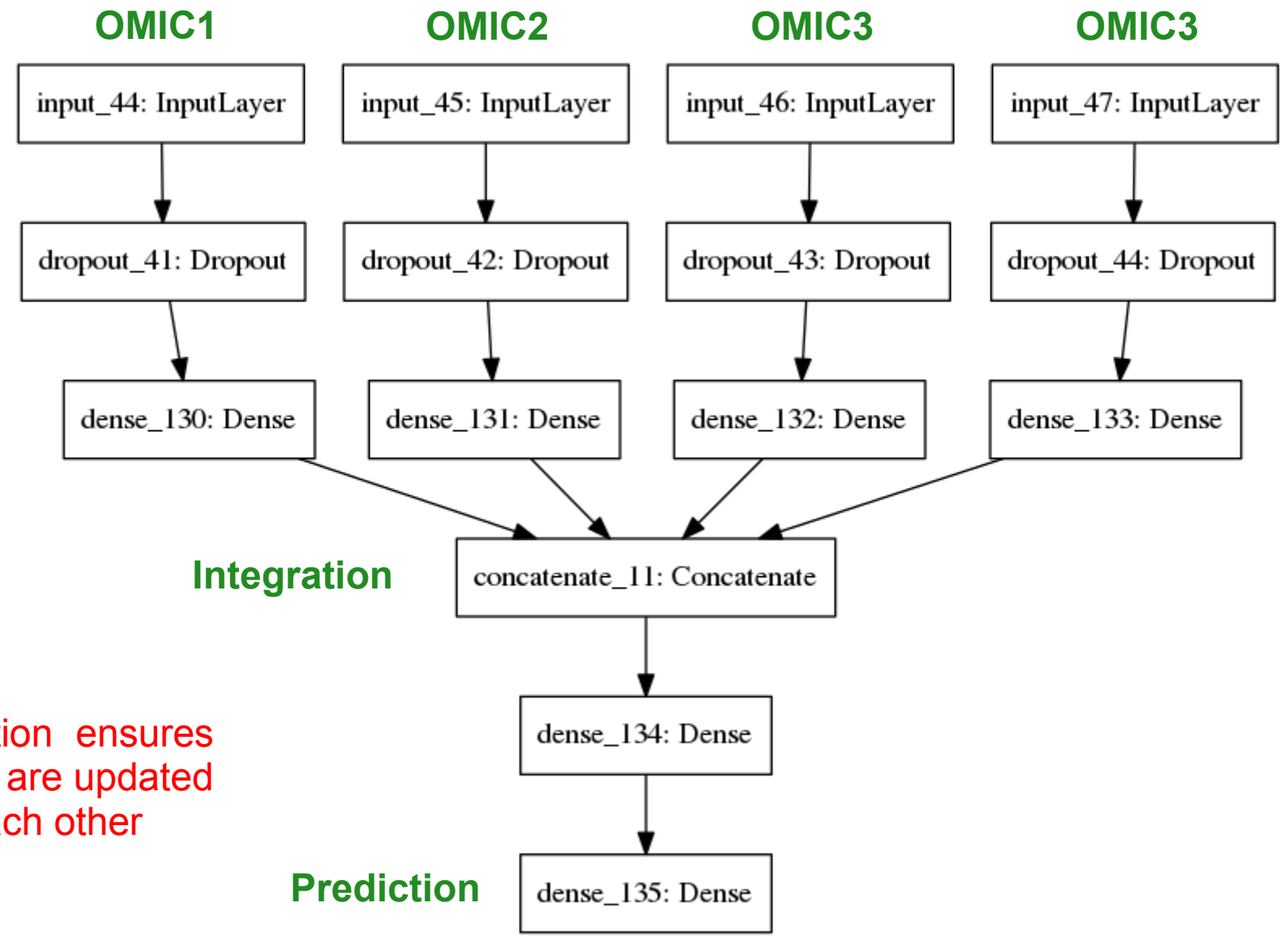








Deep Learning for Omics Integration



Back propagation ensures that all weights are updated in context of each other

Bayesian Networks for Omics Integration

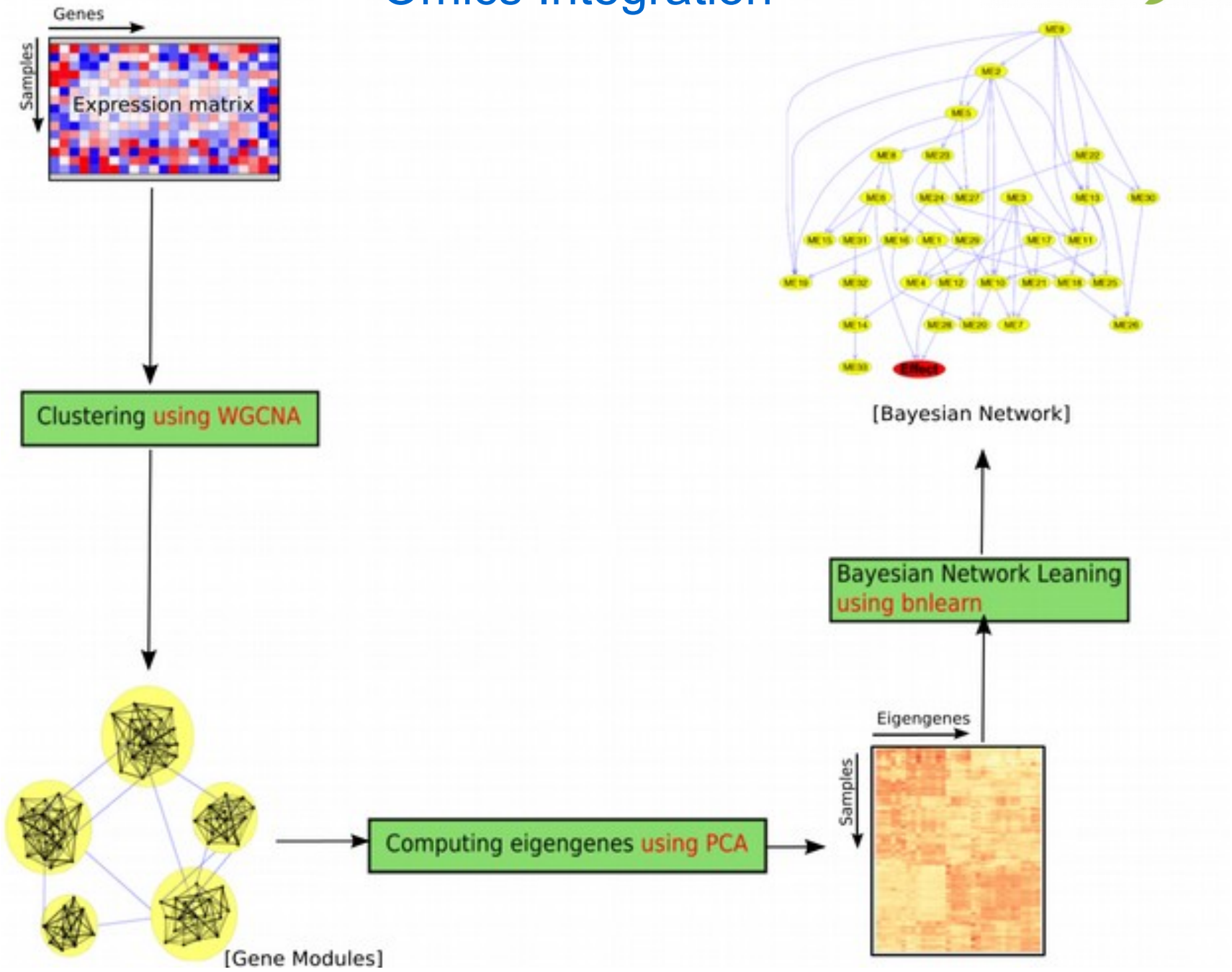


Image adapted from Agrahari, R., Foroushani, A., Docking, T.R. et al. Sci Rep 8, 6951 (2018)

Random Forest for Omics Integration

Acharjee *et al. BMC Bioinformatics* 2016, **17**(Suppl 5):180
DOI 10.1186/s12859-016-1043-4

BMC Bioinformatics

RESEARCH Open Access

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

Animesh Acharjee[1,3], Bjorn Kloosterman[1,2], Richard G. F. Visser[1] and Chris Maliepaard[1*]

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 selecting and integrating multiple ~ omics data using random

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET