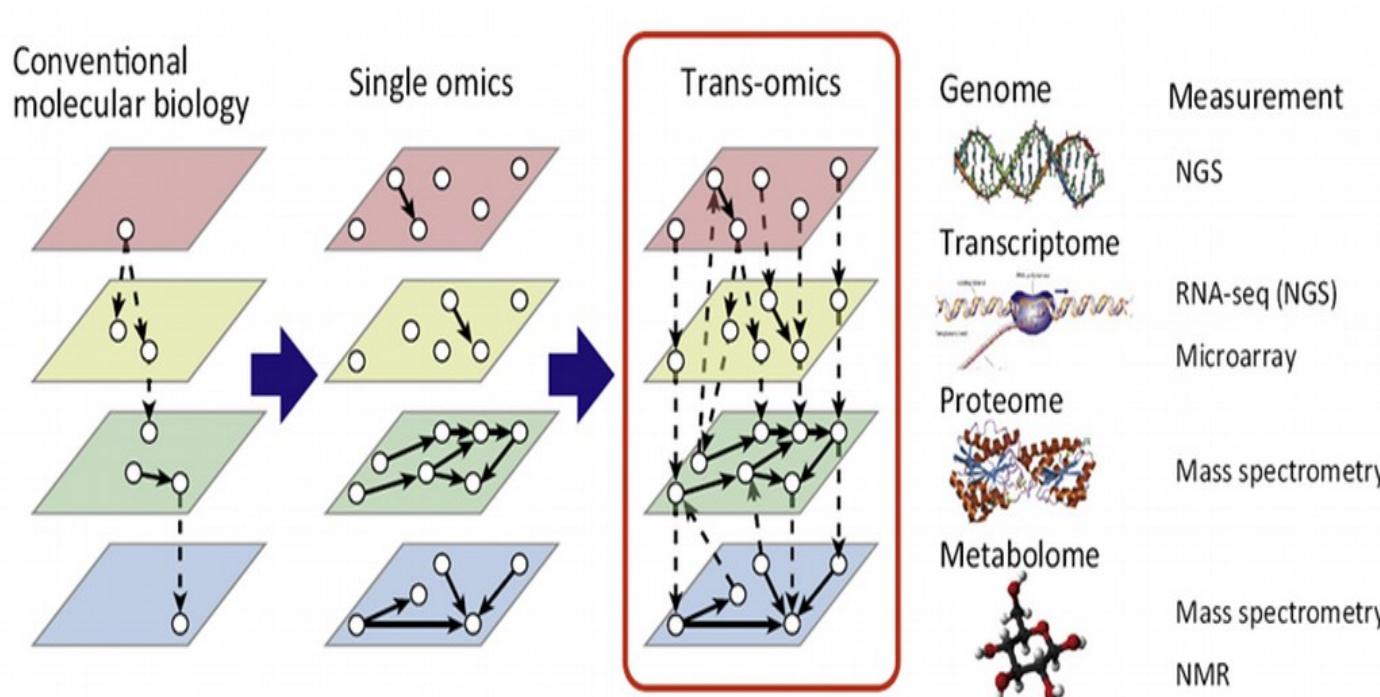


# Supervised Machine Learning MultiOmics Integration

Physalia course, 15-17 December 2025, University of Barcelona  
Nikolay Oskolkov, MRG Group Leader, LIOS, Riga, Latvia



@NikolayOskolkov



@oskolkov.bsky.social



Personal homepage:  
<https://nikolay-oskolkov.com>

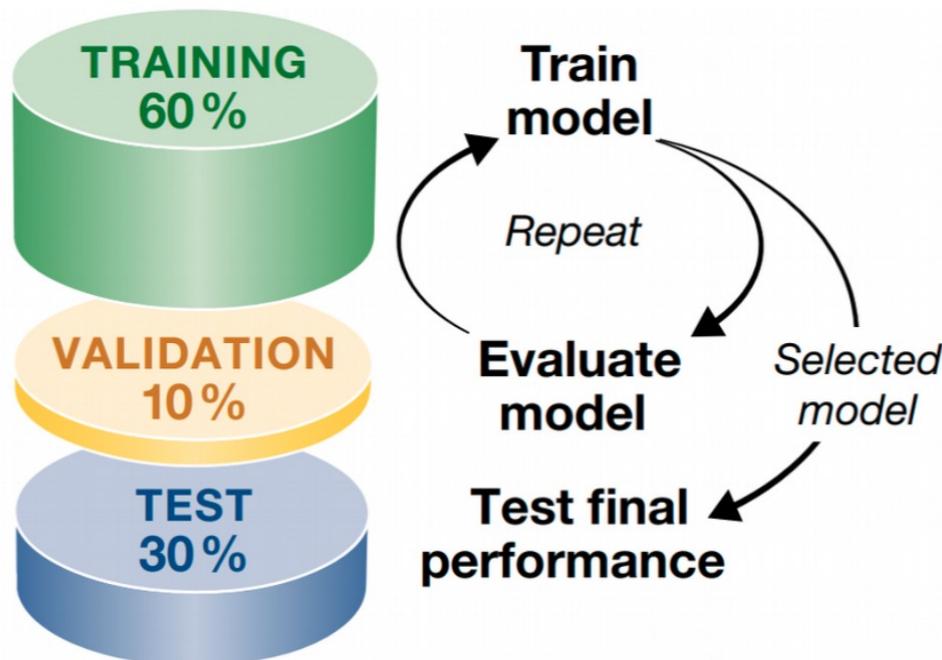
Image adapted from Yugi et al., Trends Biotechnol. 2016

# Brief Introduction to Supervised Machine Learning

**$Y = f(X)$ , where X is input (data) and Y is output (response)**

Y is present – supervised machine learning

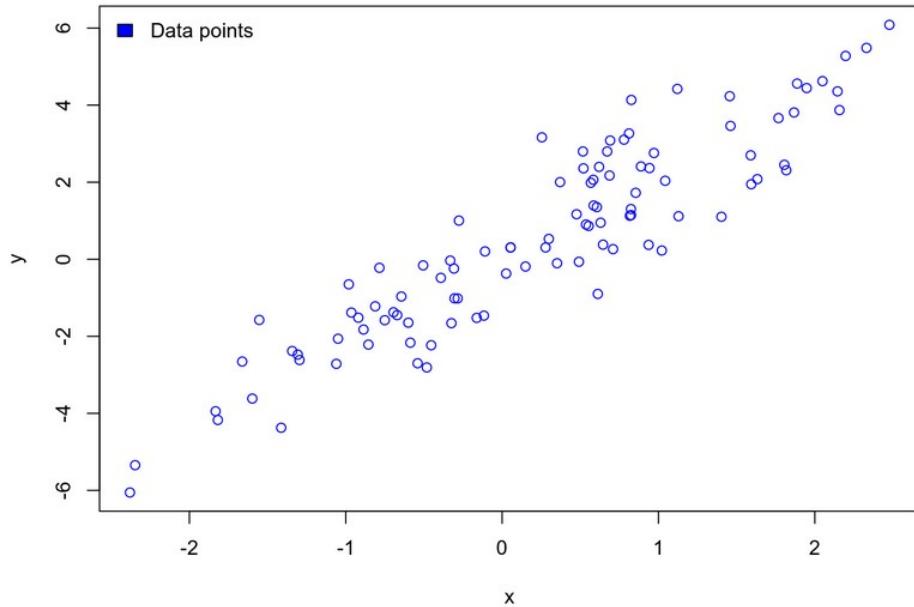
Y is absent – unsupervised machine learning



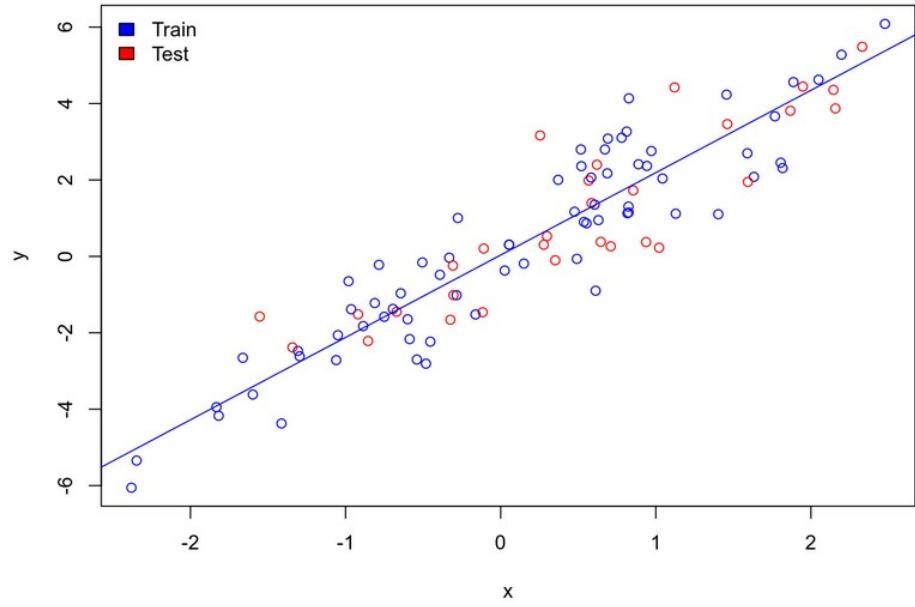
Machine Learning typically involves five basic steps:

1. Split data set into train, validation and test subsets
2. Fit the model on the train subset
3. Validate your model on the validation subset
4. Repeat train - validation split many times and tune hyperparameters
5. Test the accuracy of the optimized model on the test subset.

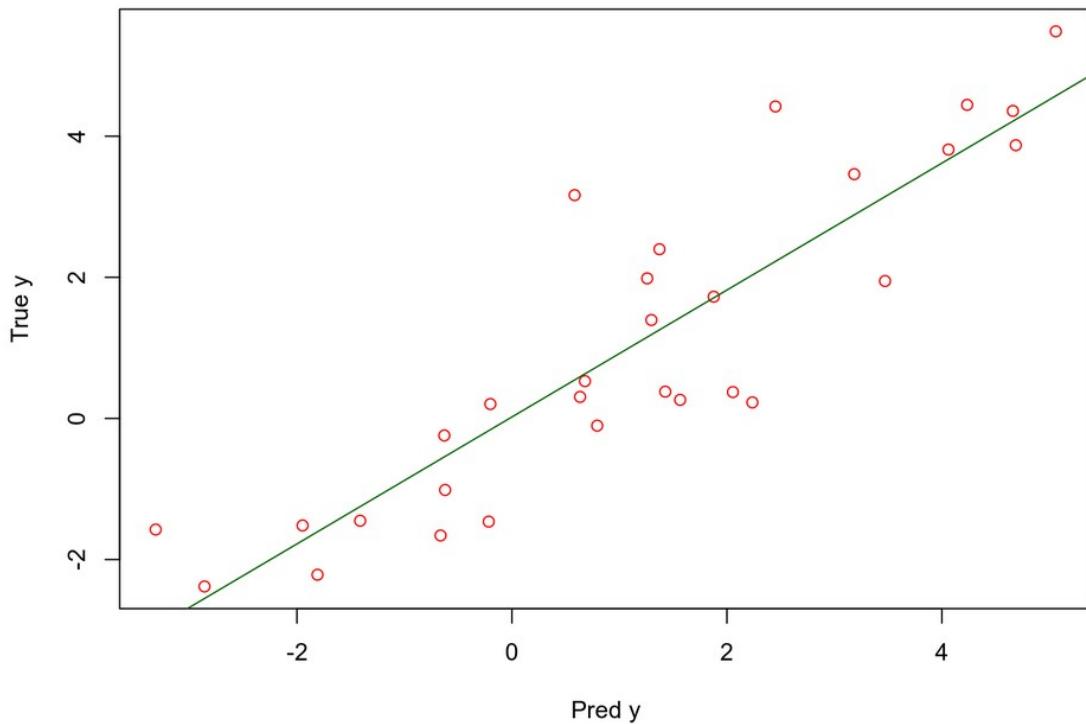
```
1 N <- 100
2 x <- rnorm(N)
3 y <- 2 * x + rnorm(N)
4 df <- data.frame(x, y)
5 plot(y ~ x, data = df, col = "blue")
6 legend("topleft", "Data points", fill = "blue", bty = "n")
```



```
1 train <- df[sample(1:dim(df)[1], 0.7 * dim(df)[1]), ]
2 test <- df[!rownames(df) %in% rownames(train), ]
3 df$col <- ifelse(rownames(df) %in% rownames(test), "red", "blue")
4 plot(y ~ x, data = df, col = df$col)
5 legend("topleft", c("Train", "Test"), fill=c("blue", "red"), bty="n")
6 abline(lm(y ~ x, data = train), col = "blue")
```



```
1 test_predicted <- as.numeric(predict(lm(y ~ x, data = train), newdata = test))
2 plot(test$y ~ test_predicted, ylab = "True y", xlab = "Pred y", col = "red")
3 abline(lm(test$y ~ test_predicted), col = "darkgreen")
```



```
1 summary(lm(test$y ~ test_predicted))
```

```
Call:
lm(formula = test$y ~ test_predicted)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.80597 -0.78005  0.07636  0.52330  2.61924 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.02058   0.21588   0.095   0.925    
test_predicted 0.89953   0.08678  10.366 4.33e-11 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 
' ' 1

Residual standard error: 1.053 on 28 degrees of freedom
Multiple R-squared:  0.7933,    Adjusted R-squared:
0.7859
F-statistic: 107.4 on 1 and 28 DF,  p-value: 4.329e-11
```

Thus the model explains 79% of variation on the test subset.

# Supervised Machine Learning applied to Omics Integration



OPEN

## Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets

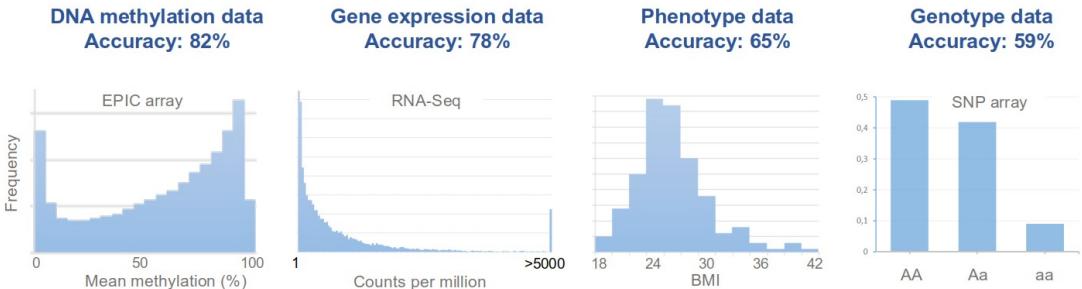
Tina Rönn<sup>1</sup>, Alexander Perfiljev<sup>1</sup>, Nikolay Oskolkov<sup>2,3</sup> & Charlotte Ling<sup>1,3,✉</sup>

Type 2 diabetes (T2D) is the fastest growing non-infectious disease worldwide. Impaired insulin secretion from pancreatic beta-cells is a hallmark of T2D, but the mechanisms behind this defect are insufficiently characterized. Integrating multiple layers of biomedical information, such as different Omics, may allow more accurate understanding of complex diseases such as T2D. Our aim was to explore and use Machine Learning to integrate multiple sources of biological/molecular information (multiOmics), in our case RNA-seqencing, DNA methylation, SNP and phenotypic data from islet donors with T2D and non-diabetic controls. We exploited Machine Learning to perform multiOmics integration of DNA methylation, expression, SNPs, and phenotypes from pancreatic islets of 110 individuals, with ~ 30% being T2D cases. DNA methylation was analyzed using Infinium MethylationEPIC array, expression was analyzed using RNA-seqencing, and SNPs were analyzed using HumanOmniExpress arrays. Supervised linear multiOmics integration via DIABLO based on Partial Least Squares (PLS) achieved an accuracy of  $91 \pm 15\%$  of T2D prediction with an area under the curve of  $0.96 \pm 0.08$  on the test dataset after cross-validation. Biomarkers identified by this multiOmics integration, including SACS and TXNIP/DNA methylation, OPRD1 and RHOT1 expression and a SNP annotated to ANO2, provide novel insights into the interplay between different biological mechanisms contributing to T2D. This Machine Learning approach of multiOmics cross-sectional data from human pancreatic islets achieved a promising accuracy of T2D prediction, which may potentially find broad applications in clinical diagnostics. In addition, it delivered novel candidate biomarkers for T2D and links between them across the different Omics.

**Keywords** DNA methylation, RNA-seqencing, Genetic variation, Metabolic disease, Omics integration, Machine learning, Epigenetics, MultiOmics analysis, Insulin secretion, Beta-cell, EWAS, GWAS

The complexity of the human genome necessitates integration of several Omics, i.e. different layers of information on top of the DNA sequence, to get further insights in the pathogenesis of type 2 diabetes (T2D). Next generation sequencing (NGS) technologies have revolutionized T2D research and provided unique information about the disease with unprecedented depth and scale<sup>1</sup>. Importantly, genome-wide association studies (GWAS) showed that a large proportion of risk single nucleotide polymorphisms (SNPs) for T2D are associated with impaired insulin secretion<sup>2</sup>. Indeed, pancreas is the key organ for understanding T2D pathogenesis since insulin and glucagon secretion from pancreatic beta and alpha cells, respectively, largely control blood glucose levels. Large efforts have therefore been made to dissect the molecular mechanisms that contribute to impaired insulin and glucagon secretion from pancreatic islets in patients with T2D. These include genome-wide RNA-seq (RNA-Seq)<sup>3</sup> and DNA methylation analysis<sup>4</sup> in human islet donors with T2D and non-diabetic controls, which identified candidate genes for the disease. DNA methylation, i.e., the attachment of a methyl group to the DNA, is an epigenetic mark indicative of gene activity. As DNA methylation mainly occurs on the nucleotide cytosine, it is also dependent on SNPs<sup>5,6</sup>. The genetic and epigenetic codes are eventually connected to RNA transcription in each cell, comprehensively quantified by RNA-Seq. A combined analysis of these complementary Omics data from different layers of cell organization in human pancreatic islets may provide synergistic effects for the

<sup>1</sup>Epigenetics and Diabetes Unit, Department of Clinical Sciences, Lund University Diabetes Centre, Scania University Hospital, Lund University, 205 02 Malmö, Sweden. <sup>2</sup>Science for Life Laboratory, Department of Biology, National Bioinformatics Infrastructure Sweden, Lund University, Sölvegatan 35, 223 62 Lund, Sweden. <sup>3</sup>These authors contributed equally: Nikolay Oskolkov and Charlotte Ling. <sup>✉</sup>email: charlotte.ling@med.lu.se


[LUDC](#) [Research groups](#) [LUDC-IRC](#) [Resources](#) [Courses and training](#) [For early career researchers](#)

Search this site

Search


[Start](#) > [Resources at LUDC](#) > [Human Tissue Laboratory](#)

- [LUDC Bioinformatics unit](#)
- [Flow Cytometry Core Facility at LUDC](#)
- [Innovation support](#)
- [Population studies](#)
- [Human Tissue Laboratory](#)
  - [Publications](#)
- [Confocal Platform](#)

## Human Tissue Laboratory



HTL receives and handles donor materials in the form of human insulin-producing cells, muscles, and liver and adipose tissue. Photo: Kennet Ruona

The human tissue laboratory (HTL) is collaboration between EXODIAB and the Nordic Network for Clinical Islet Transplantation. The main tissue handled is human pancreatic islets. The islets are isolated at the islet isolation facility in Uppsala, primarily for transplantation purposes. A fraction of the islets that can not be used for transplantation, are under agreed consent instead dedicated to research and distributed to research centers

## Related information

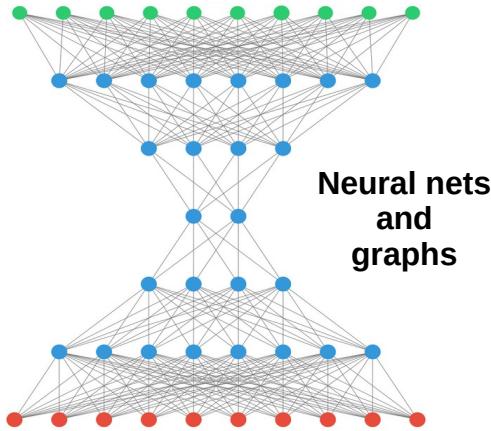
[The nordic network for clinical islets transplantation](#)

## Publications

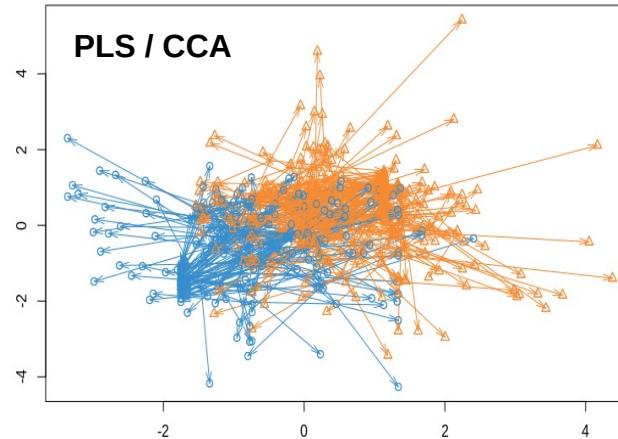
A growing number of publications are based on results from the Human Tissue Laboratory.

[Publications](#)

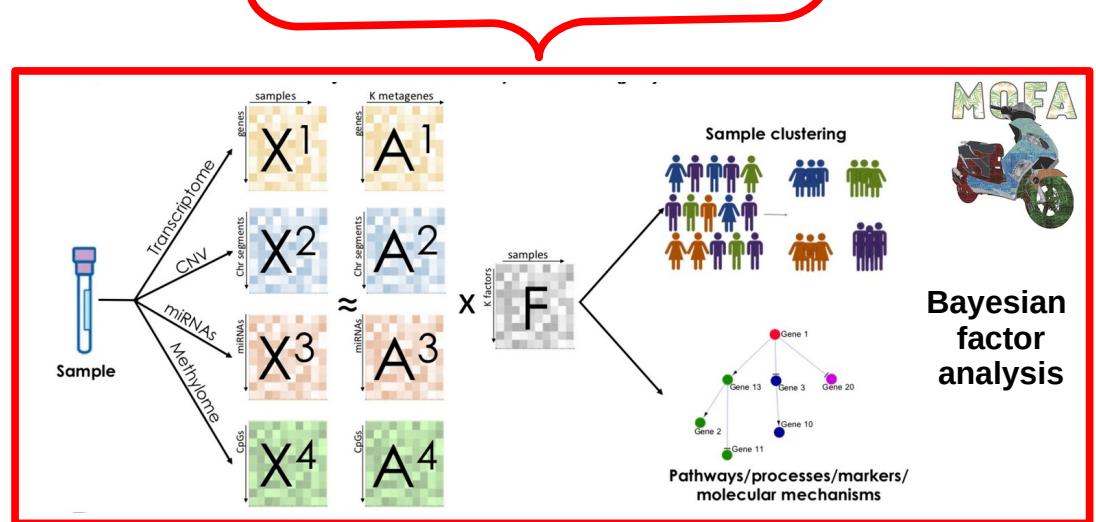
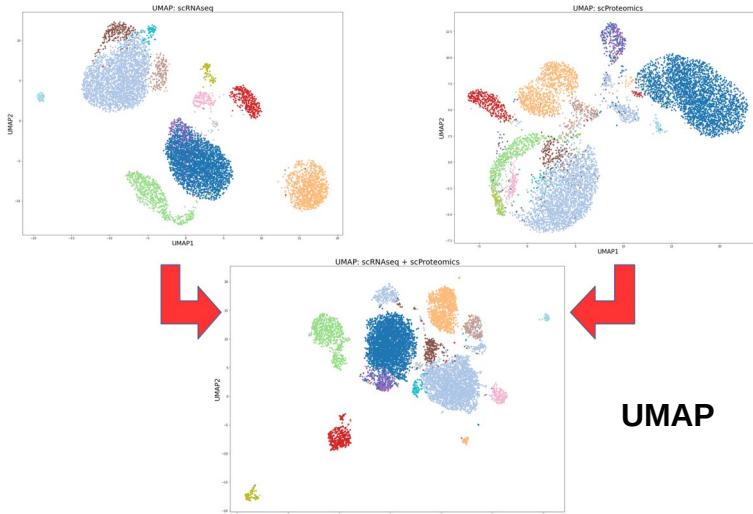
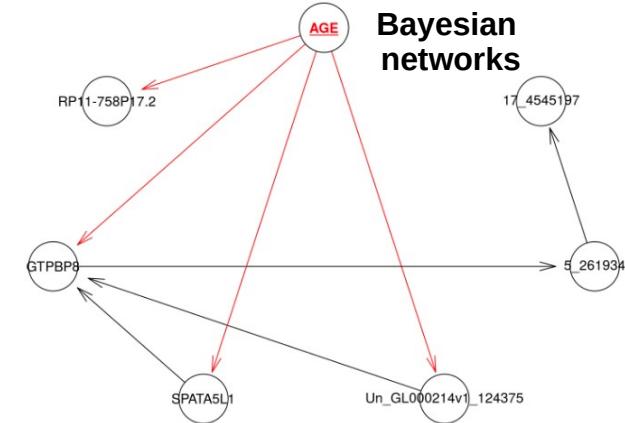
## Convert to common space



## Extract common variation



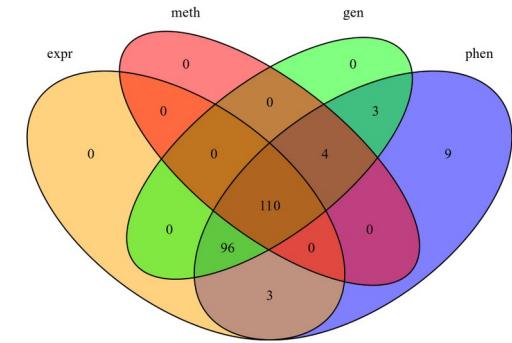
## Combine via Bayes rule



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

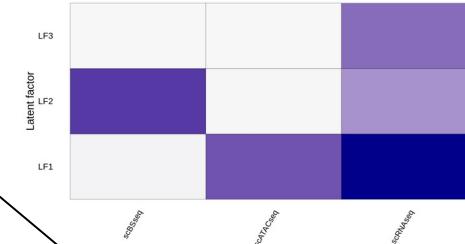
For Example:

- 1) With ~100 samples it is a good idea to do **linear** Omics integration
- 2) T2D is a phenotype of interest, therefore **supervised** integration



Data Set (4 Omics)  
110 overlapping individuals

Check covariance



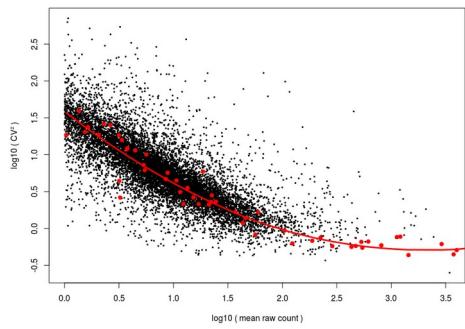
Train Set ( $n = 80$ )

Supervised:  
LASSO

Test Set ( $n = 30$ )

Evaluation

Unsupervised:  
remove low-variance

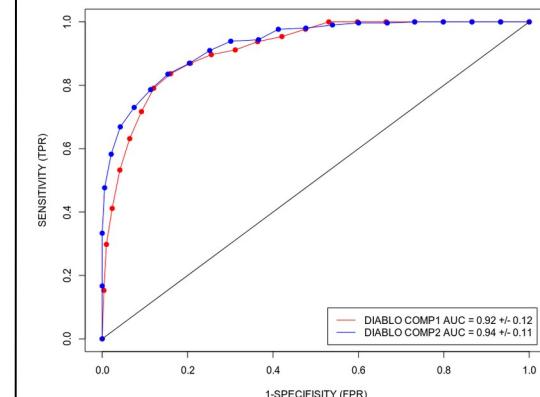


Feature Selection

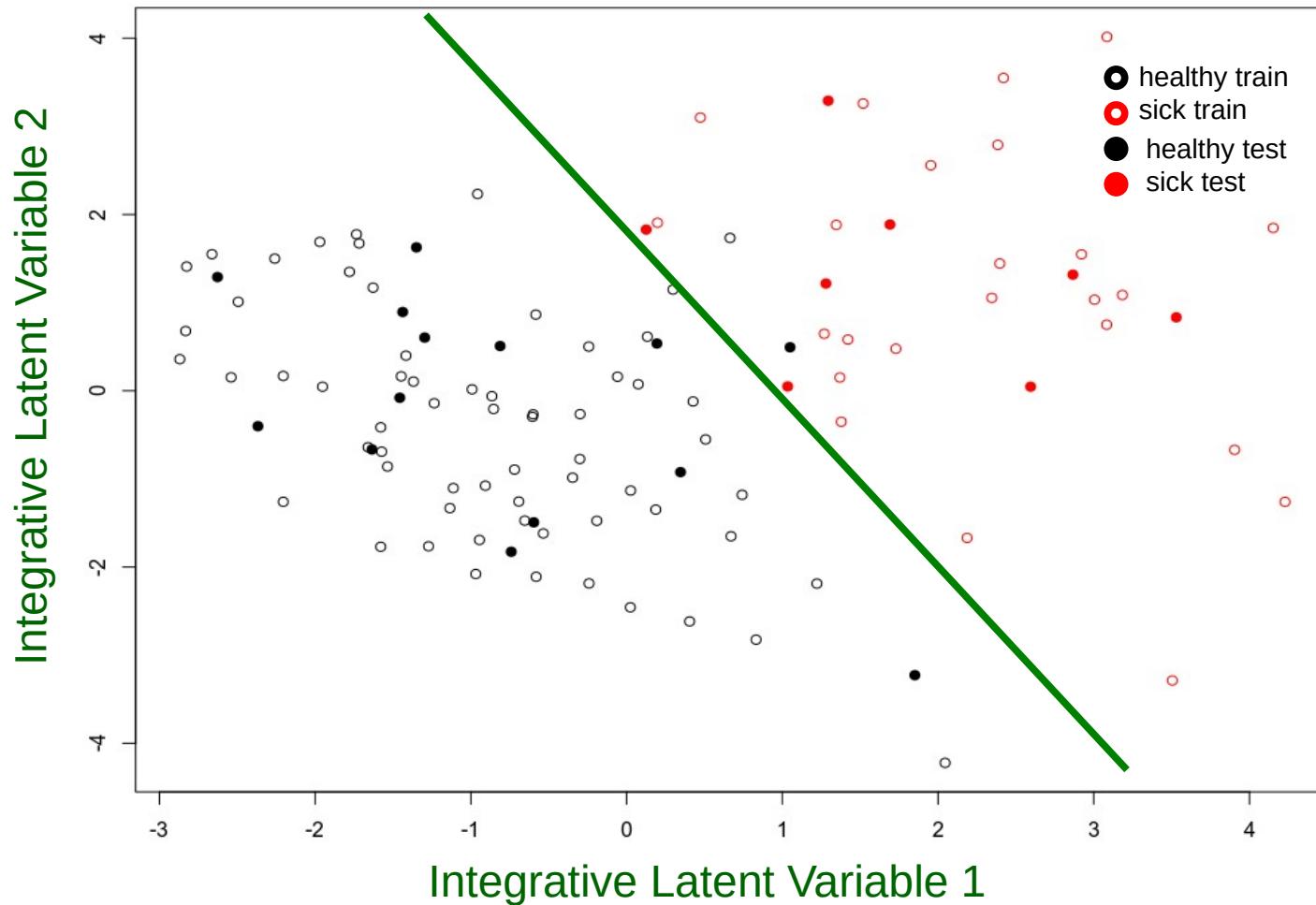


Omics Integration

Trained Model

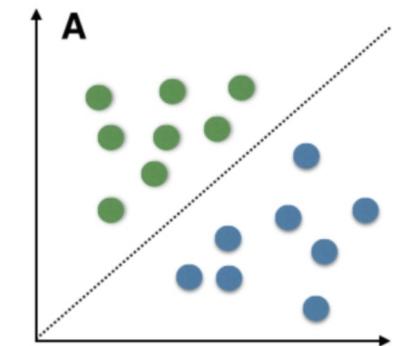


## Linear Separation: Decision Boundary

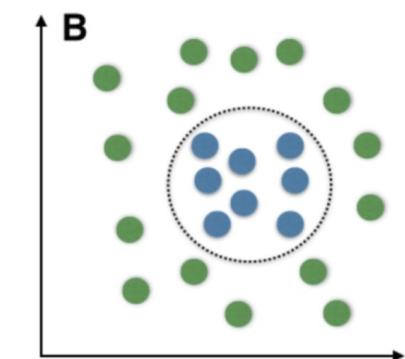


A: Linearly Separable Data

A



B



B: Non-Linearly Separable Data

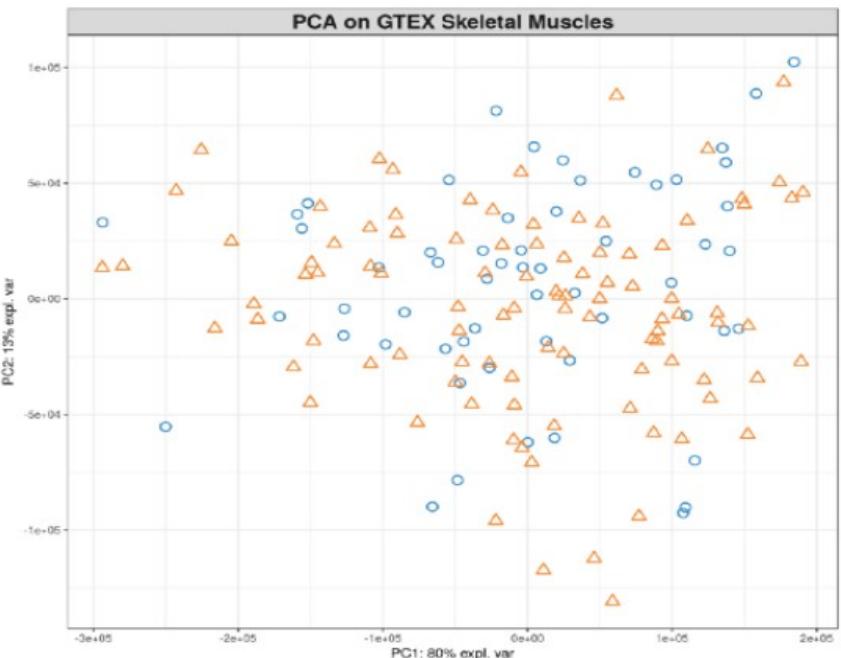
# Univariate and Multivariate Feature Selection

```

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](#)

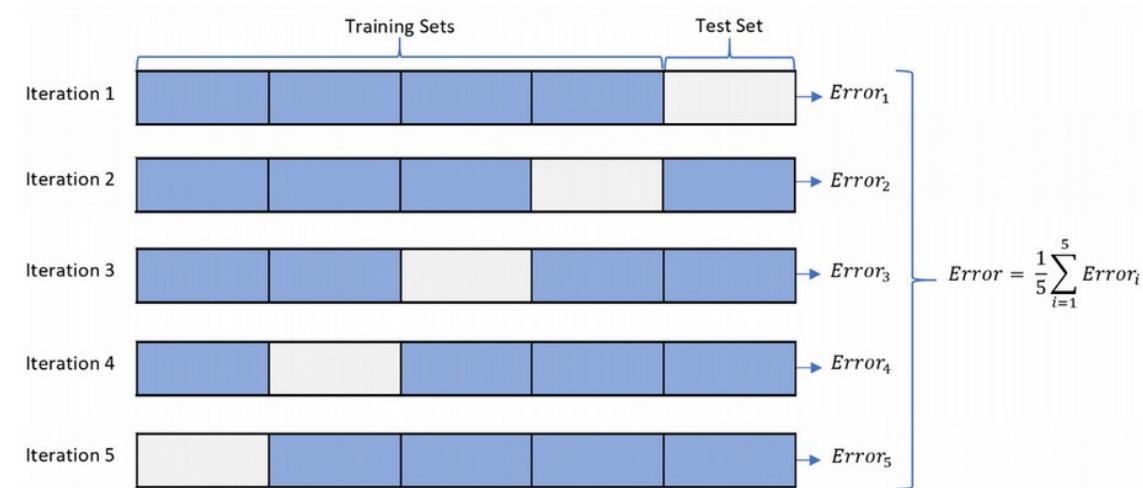
	GENE	SPEARMAN_RHO	PVALUE	FDR
## 256	ENSG00000184368.11_MAP7D2	-0.5730196	4.425151e-15	2.416132e-12
## 324	ENSG00000110013.8 SIAE	0.3403994	1.288217e-05	3.516833e-03
## 297	ENSG00000128487.12_SPECC1	-0.3003621	1.323259e-04	2.408332e-02
## 218	ENSG00000162512.11_SDC3	0.2945390	1.807649e-04	2.467441e-02
## 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 poor predictive capacity compared to multivariate feature selection

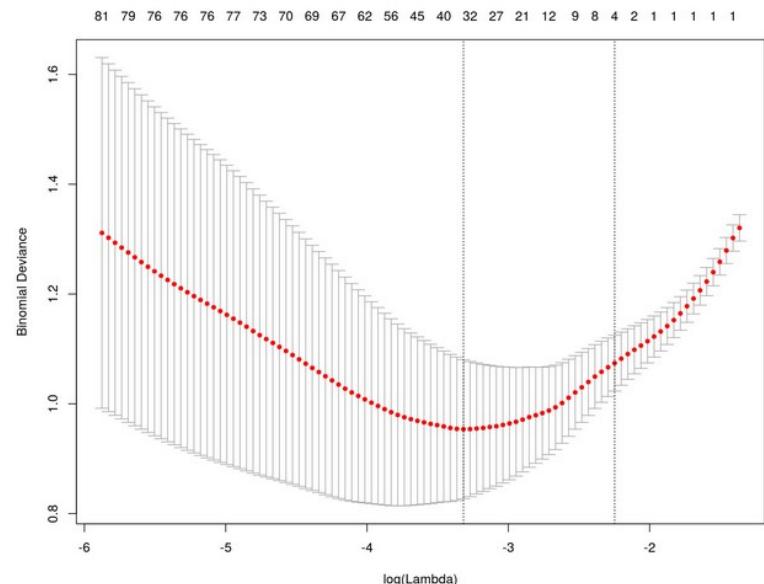
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

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

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



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



$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon; \quad Y \sim N(\beta_1 X_1 + \beta_2 X_2, \sigma^2) \equiv L(Y | \beta_1, \beta_2)$$

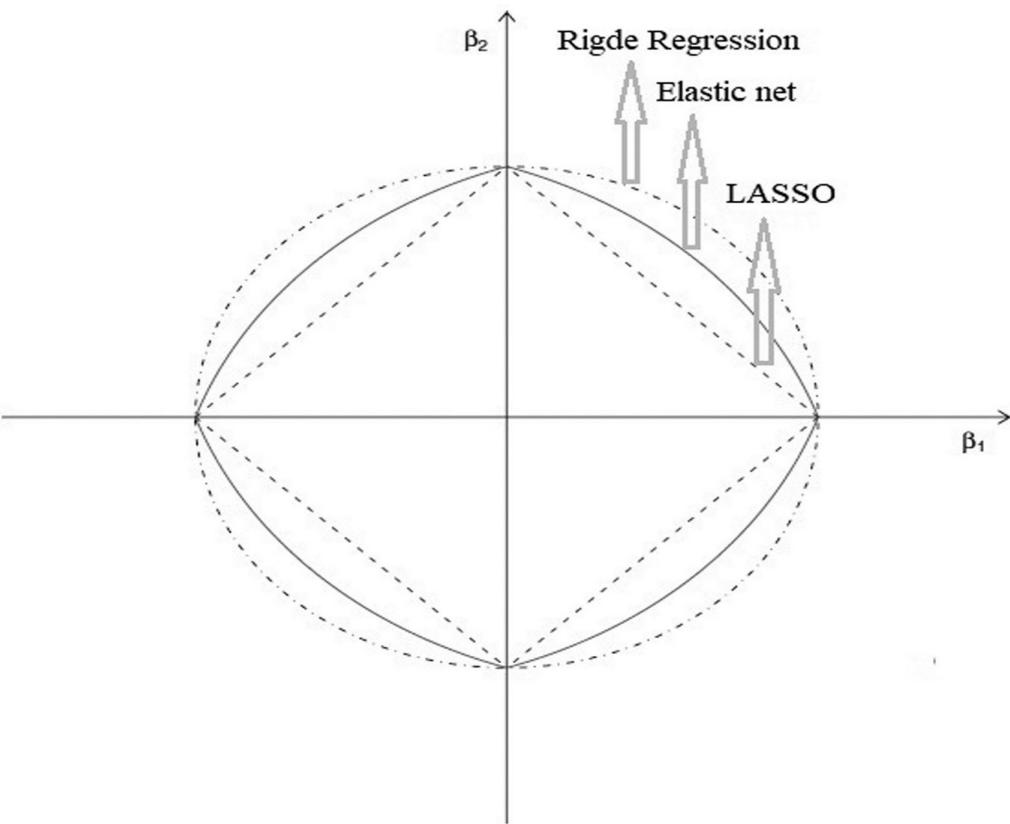
- Maximum Likelihood principle: maximize probability to observe data given parameters:

$$L(Y | \beta_1, \beta_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(Y - \beta_1 X_1 - \beta_2 X_2)^2}{2\sigma^2}}$$

- Bayes theorem: maximize posterior probability of observing parameters given data:

$$\text{Posterior}(\text{params} | \text{data}) = \frac{L(\text{data} | \text{params}) * \text{Prior}(\text{params})}{\int L(\text{data} | \text{params}) * \text{Prior}(\text{params}) d(\text{params})}$$

$$\begin{aligned} \text{Posterior}(\beta_1, \beta_2 | Y) &\sim L(Y | \beta_1, \beta_2) * \text{Prior}(\beta_1, \beta_2) \sim \exp^{-\frac{(Y - \beta_1 X_1 - \beta_2 X_2)^2}{2\sigma^2}} * \exp^{-\lambda(|\beta_1| + |\beta_2|)} \\ &- \log [\text{Posterior}(\beta_1, \beta_2 | Y)] \sim (Y - \beta_1 X_1 - \beta_2 X_2)^2 + \lambda(|\beta_1| + |\beta_2|) \end{aligned}$$

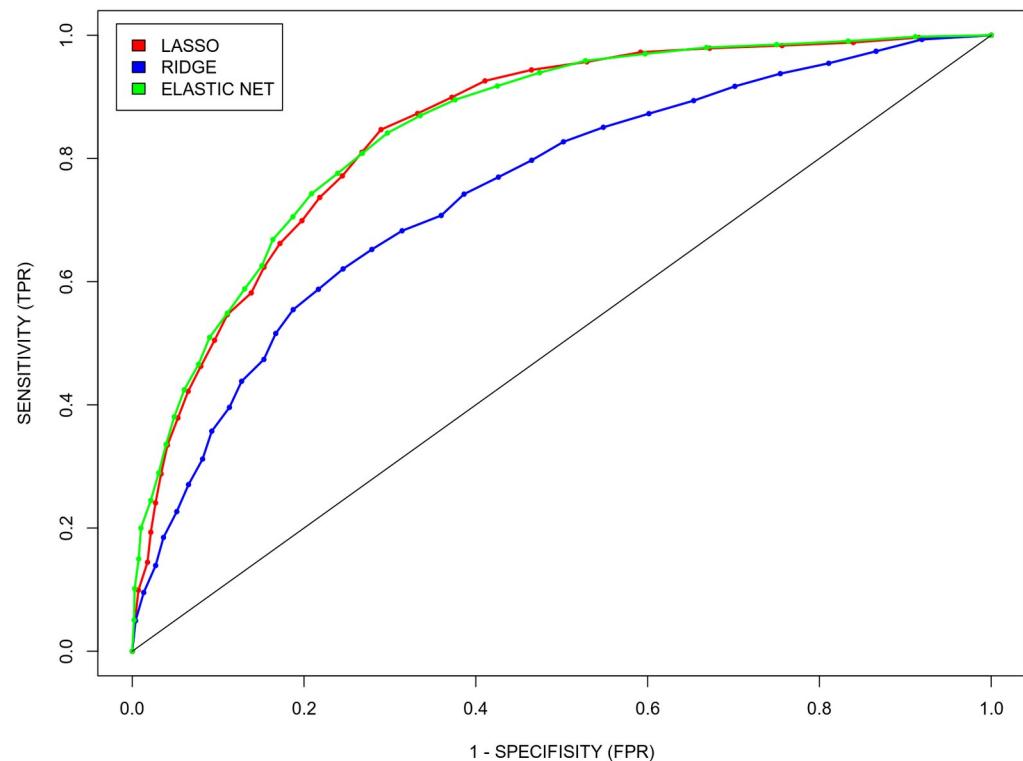


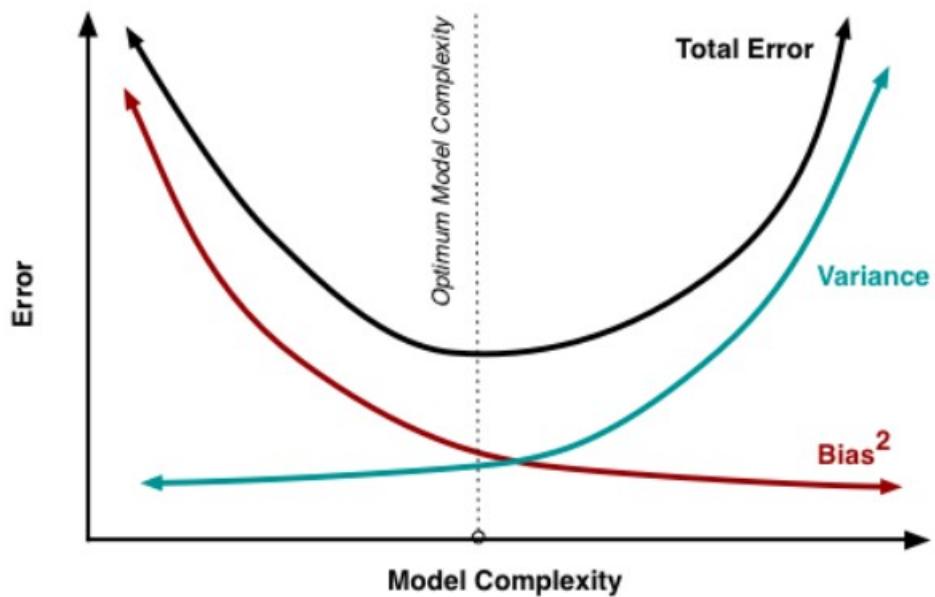
Lasso is more conservative

Ridge is more permissive

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

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



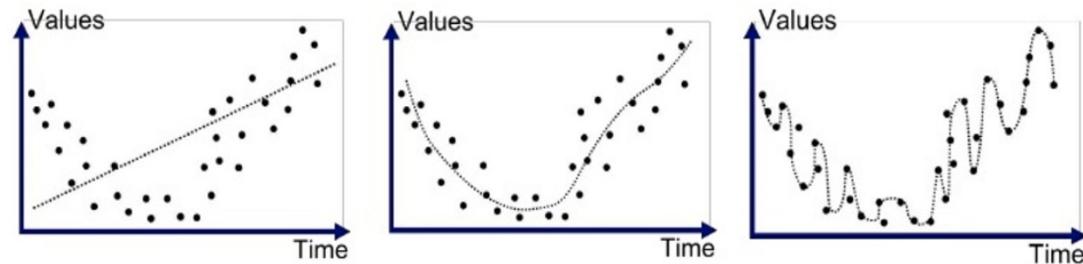
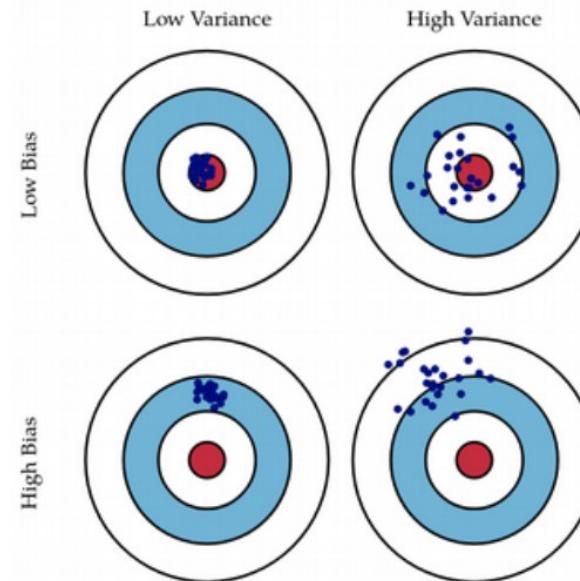


$$Y = f(X) \implies \text{Reality}$$

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

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

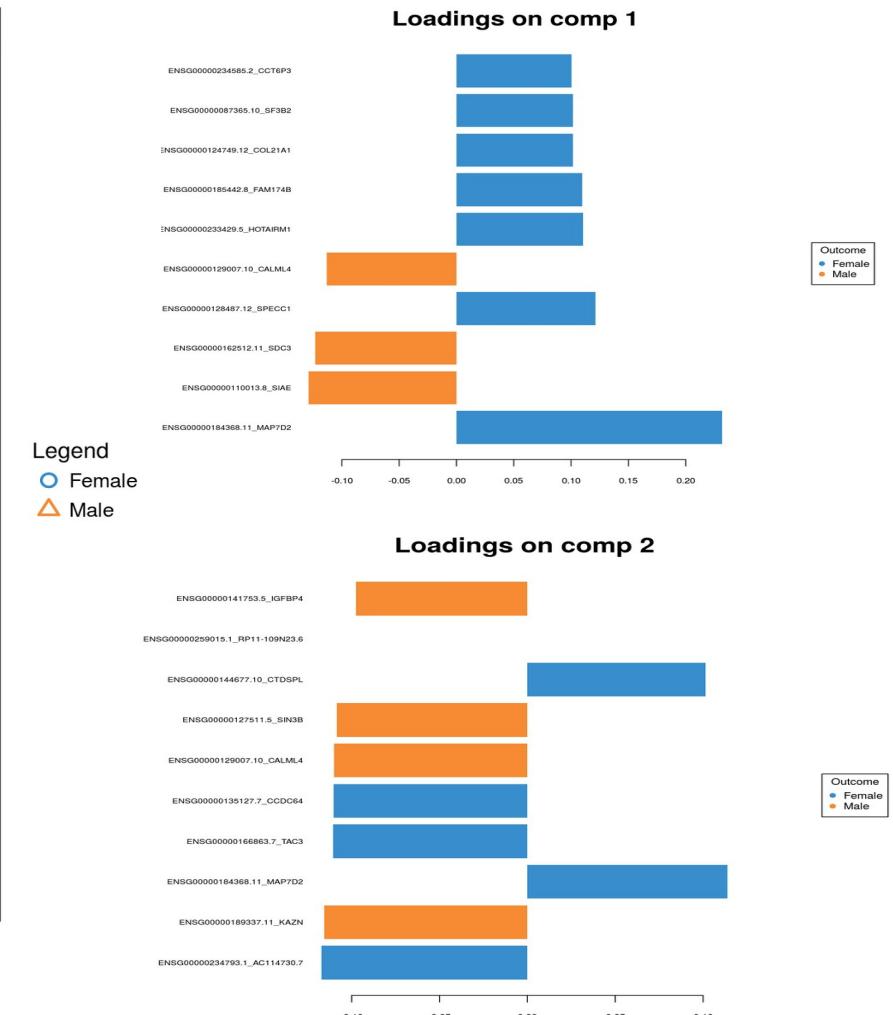
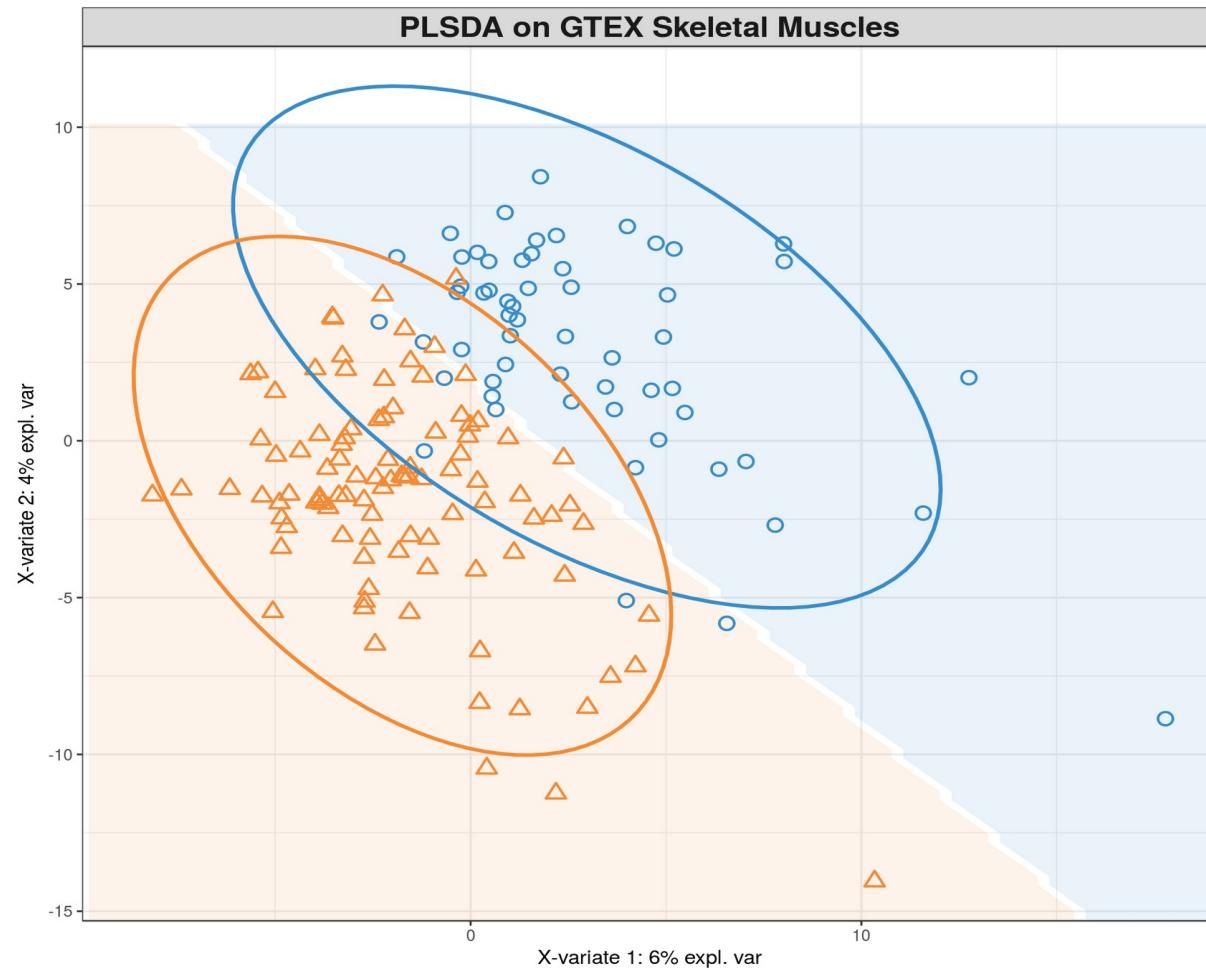
LASSO – high bias, low variance



Underfitted

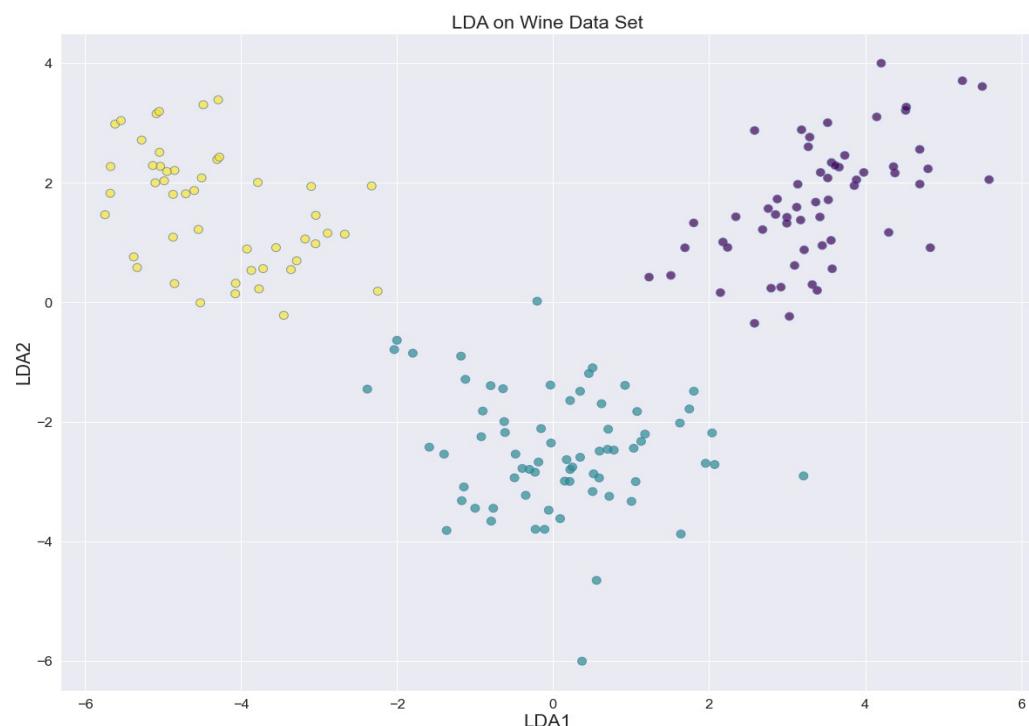
Good Fit/Robust

Overfitted



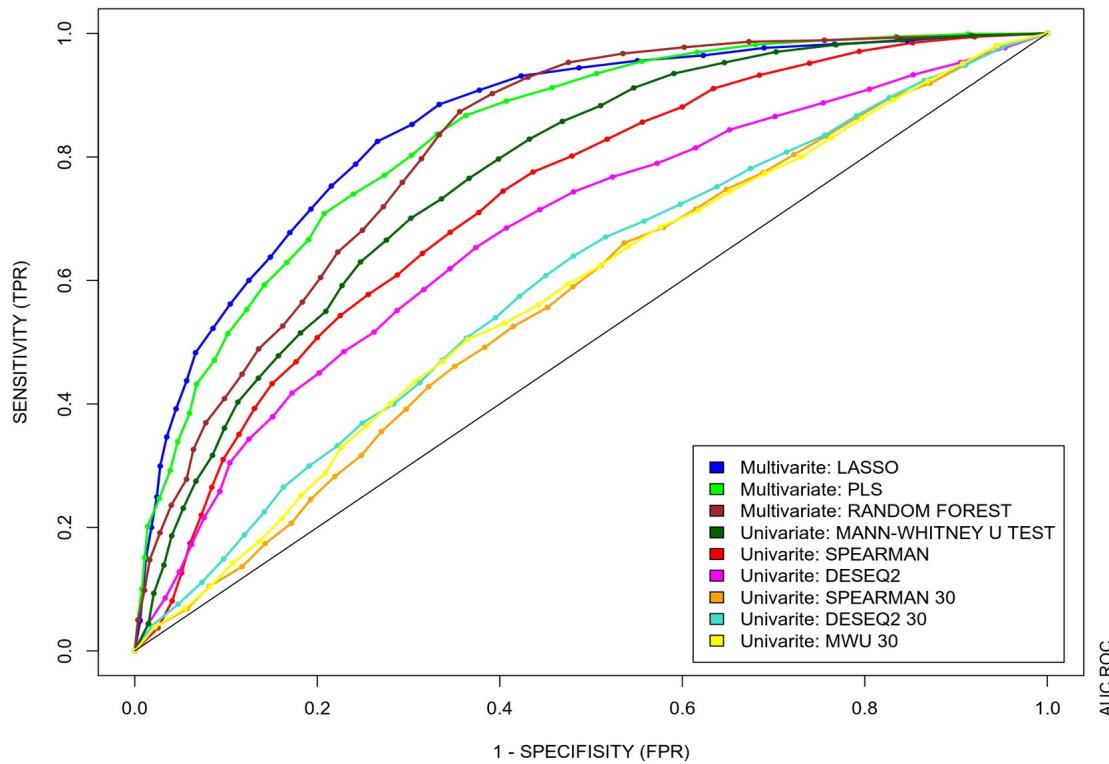
Select features that separate two groups of samples the most

# Multivariate Feature Selection: Linear Discriminant Analysis (LDA)



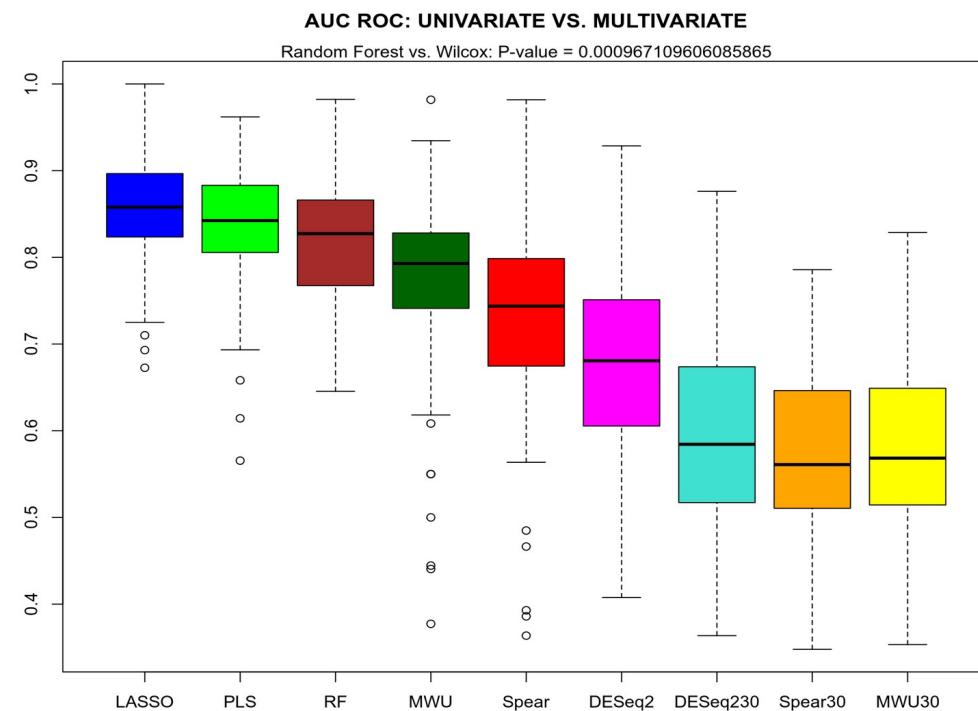
Minimize variance within clusters and maximize variance between clusters

Similar to what ANOVA is doing, therefore LINEAR Discriminant Analysis (LDA)



If you find a dataset where univariate feature selection has higher predictive capacity than multivariate one, please let me know

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



# DIABLO Omics Integration

OXFORD  
ACADEMIC

Journals

Books



## Bioinformatics

Issues

Advance articles

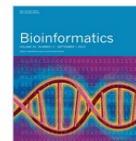
Submit ▾

Alerts

About ▾

Bioinformatics

▼ Search

Advanced  
SearchiSCBS  
INTERNATIONAL SOCIETY FOR  
COMPUTATIONAL BIOLOGYVolume 35, Issue 17  
1 September 2019

### Article Contents

Abstract

1 Introduction

2 Materials and methods

3 Results

4 Discussion

Acknowledgements

Funding

References

### JOURNAL ARTICLE

## DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays

Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, Kim-Anh Lê Cao

Bioinformatics, Volume 35, Issue 17, 1 September 2019, Pages 3055–3062, <https://doi.org/10.1093/bioinformatics/bty1054>

Published: 18 January 2019 Article history ▾

PDF Split View Cite Permissions Share ▾

### Abstract

### Motivation

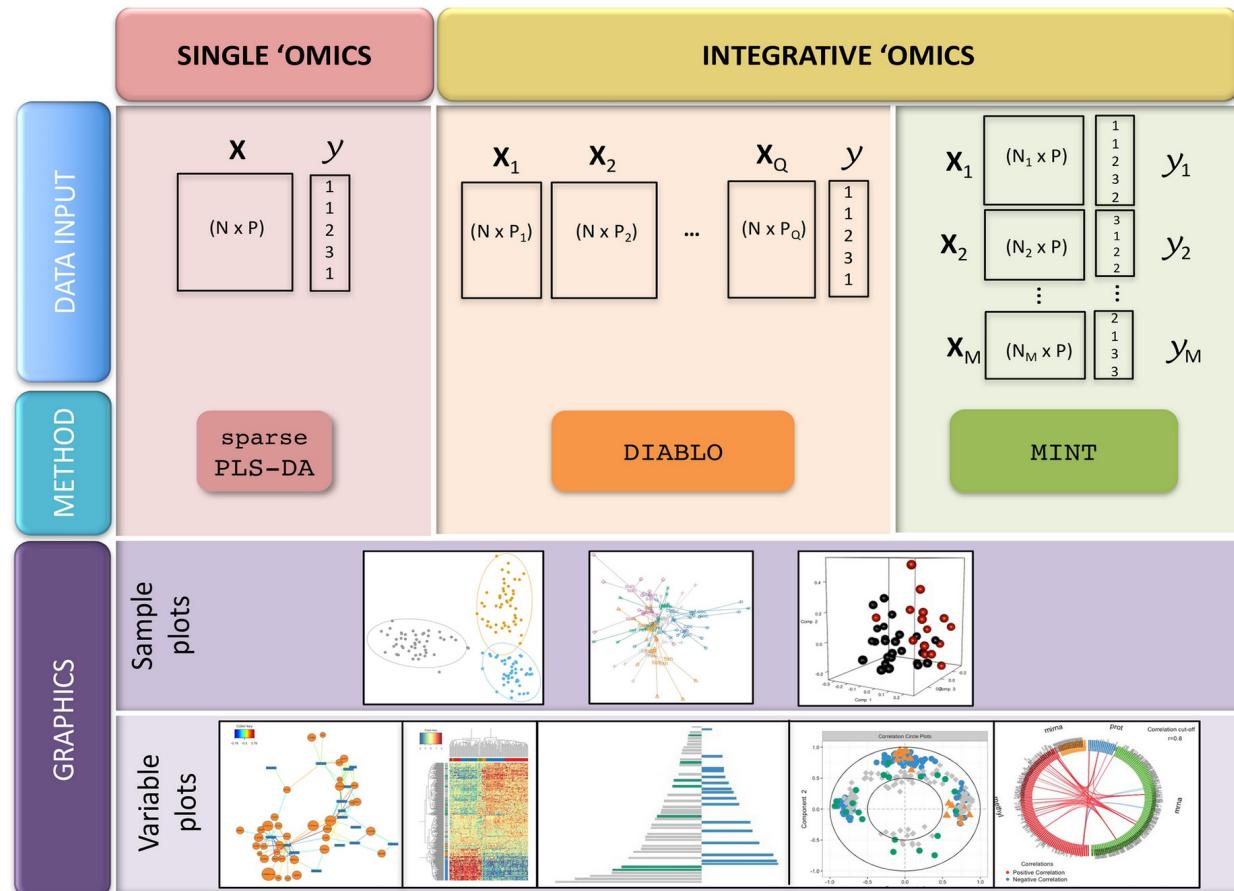
In the continuously expanding omics era, novel computational and statistical strategies are needed for data integration and identification of biomarkers and molecular signatures. We present Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO), a multi-omics integrative method that seeks for common information across different data types through the selection of a subset of molecular features, while discriminating between multiple phenotypic groups.



High-  
Impact  
Articles in

BIOINFORMATICS  
AND  
COMPUTATIONAL  
BIOLOGY

READ  
NOW



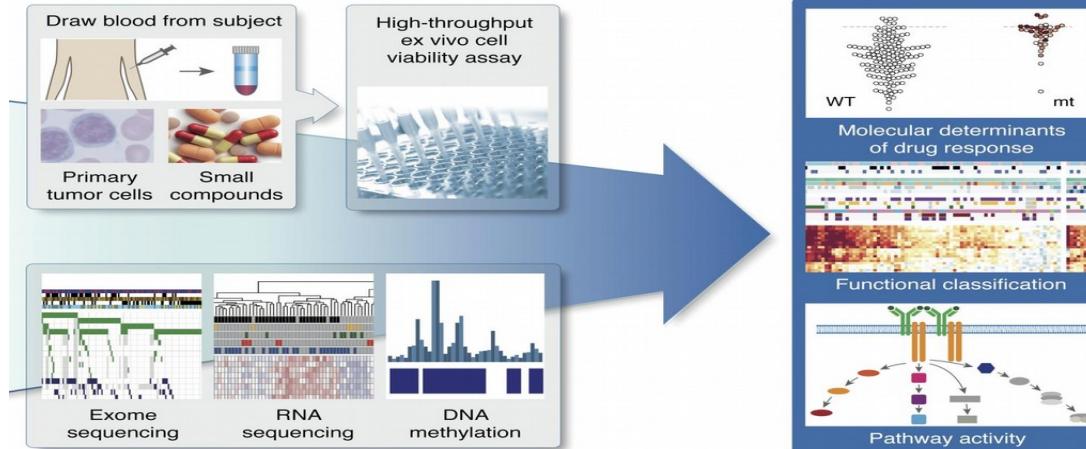
Denote  $Q$  normalized, centered and scaled datasets  $X^{(1)}$  ( $N \times P_1$ ),  $X^{(2)}$  ( $N \times P_2$ ), ...,  $X^{(Q)}$  ( $N \times P_Q$ ) measuring the expression levels of  $P_1, \dots, P_Q$  'omics variables on the same  $N$  samples'. sGCCA solves the optimization function for each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

$$\max_{a_b^{(1)}, \dots, a_b^{(Q)}} \sum_{i,j=1, i \neq j}^Q c_{i,j} \text{cov}(X_b^{(i)} a_b^{(i)}, X_b^{(j)} a_b^{(j)}), \quad (1)$$

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

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 (Lé 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.

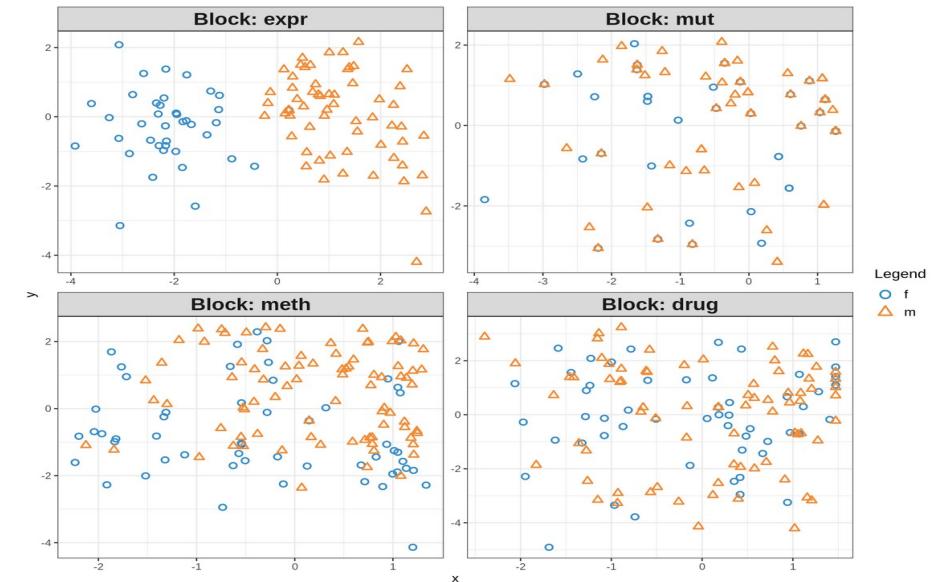
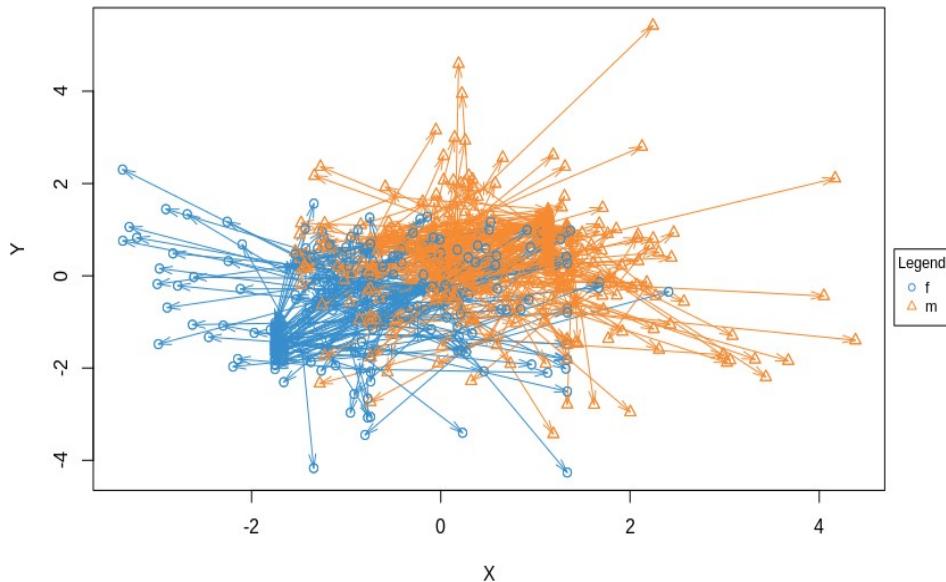
$$\max_{\beta} \text{cov}(X, Y) \implies \hat{\beta}$$

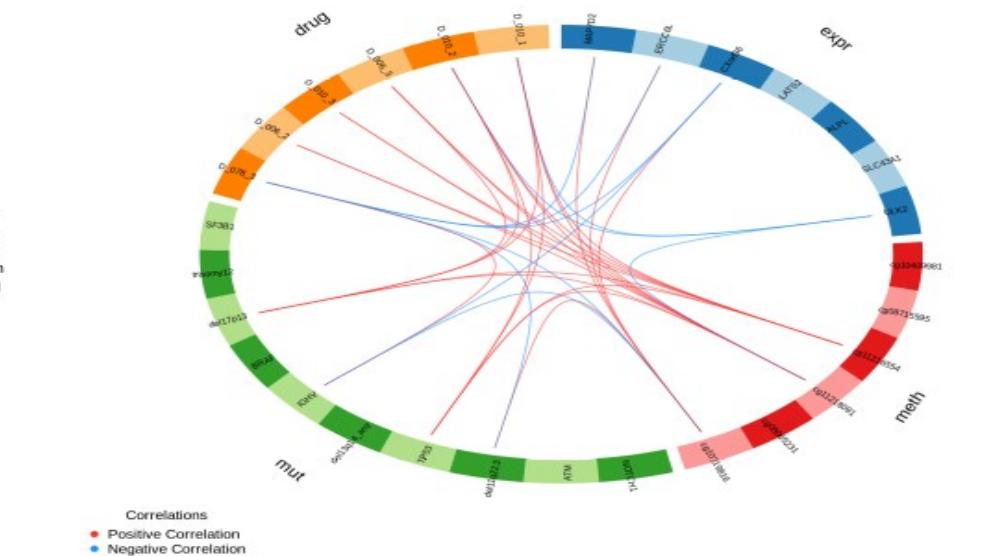
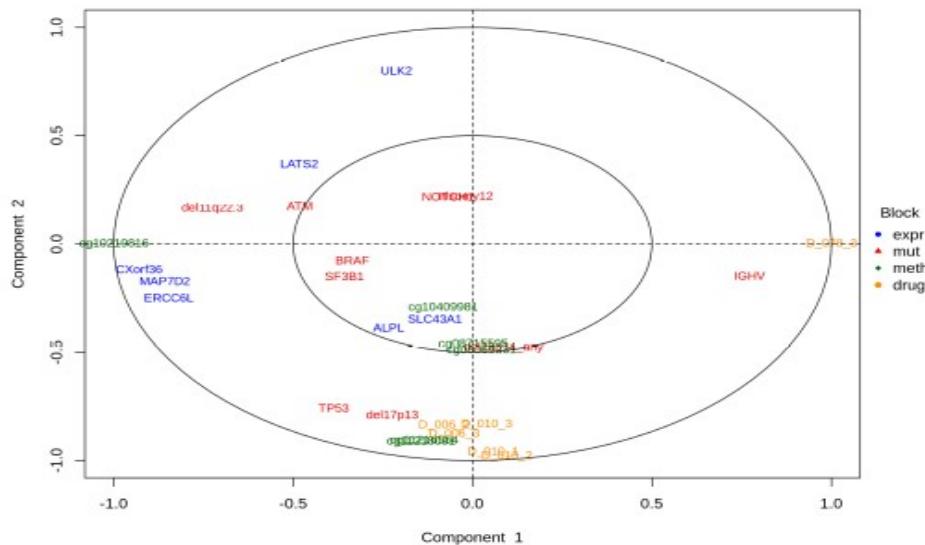
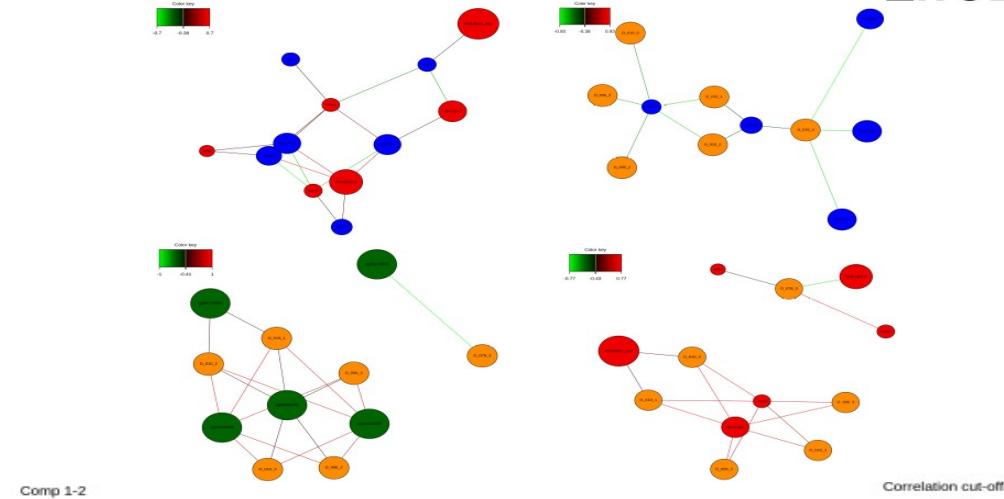
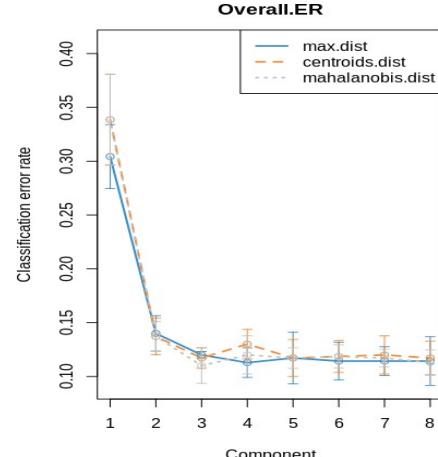
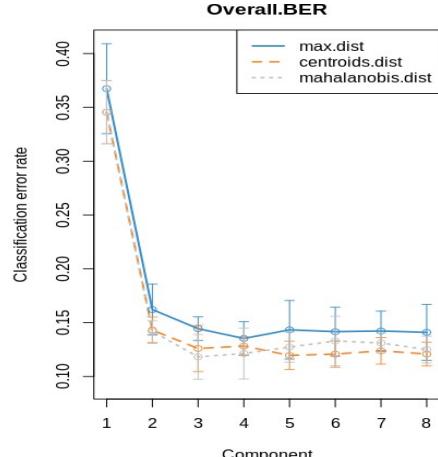


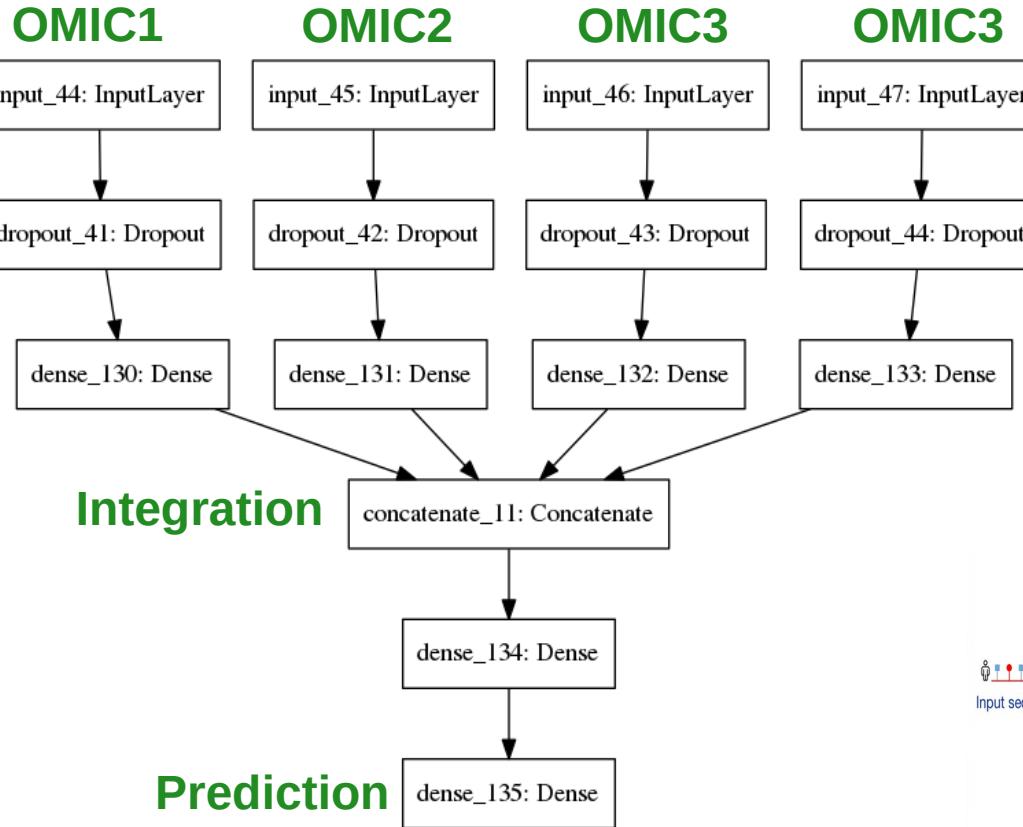
### Chronic Lymphocytic Leukaemia (CLL):

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

Dietrich et al., J Clin Invest. 2018







Backward propagation ensures that the parameters across Omics are optimized in context of each other

## Article

# Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

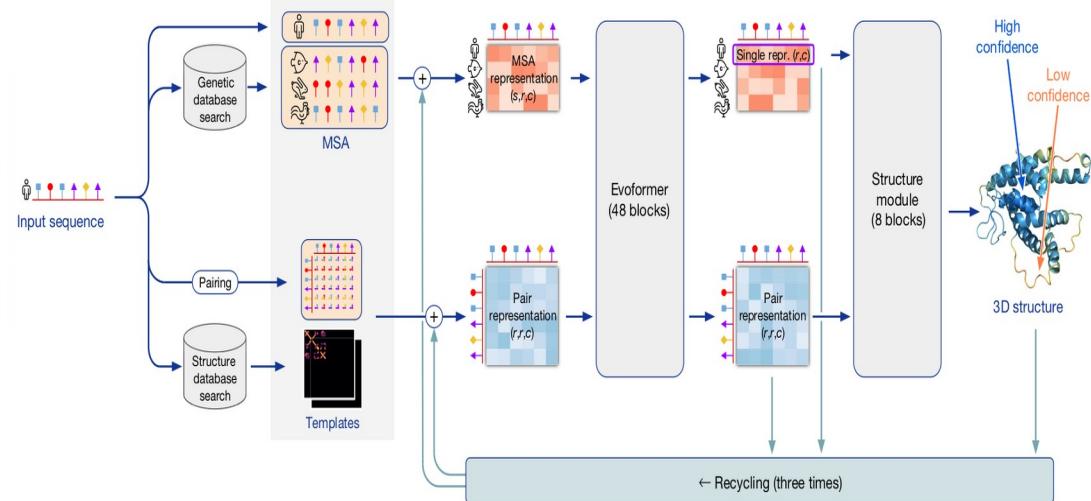
Accepted: 12 July 2021

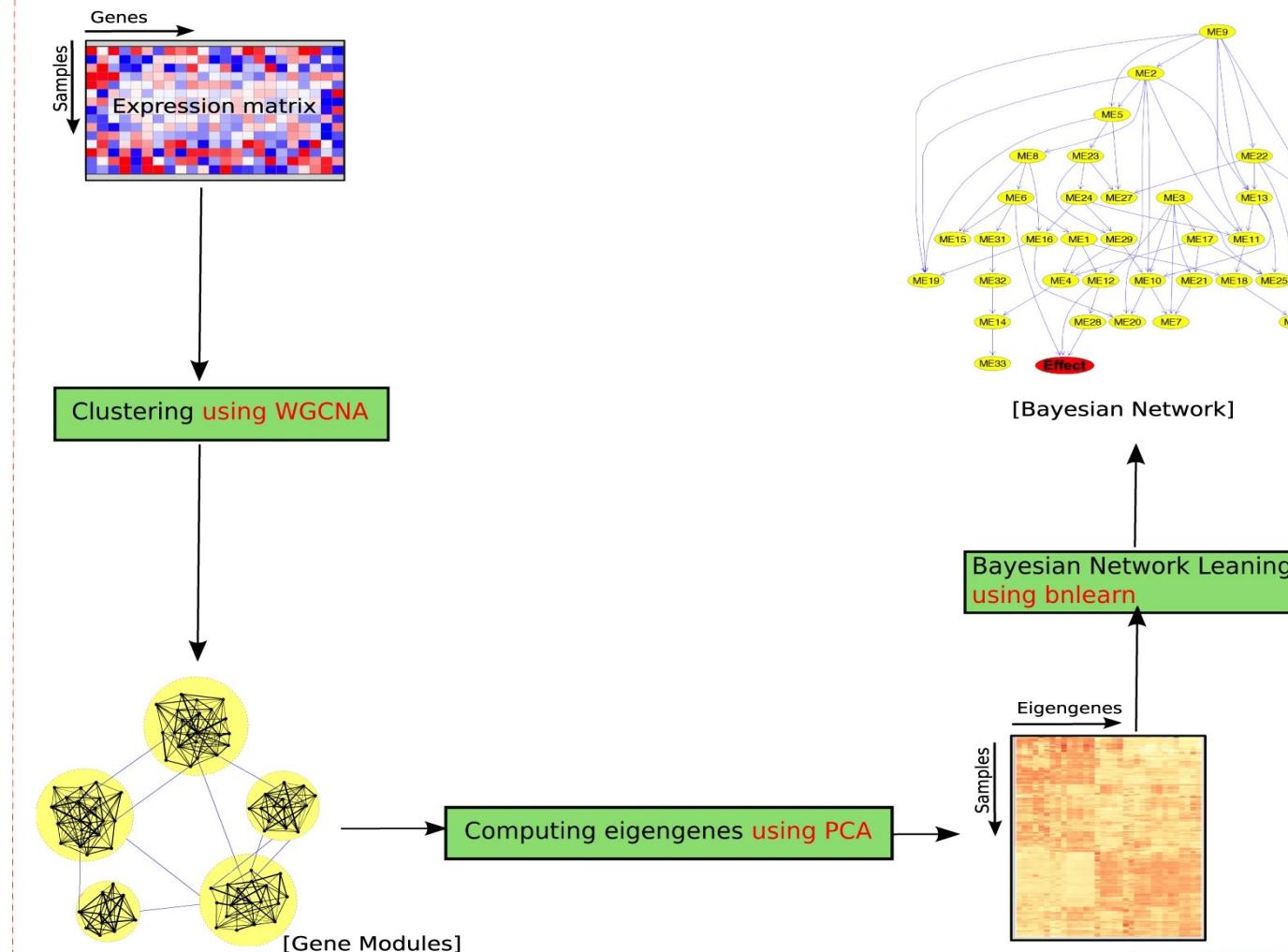
Published online: 15 July 2021

Open access

Check for updates

John Jumper<sup>1,4</sup>, Richard Evans<sup>1,4</sup>, Alexander Pritzel<sup>1,4</sup>, Tim Green<sup>1,4</sup>, Michael Figurnov<sup>1,4</sup>, Olaf Ronneberger<sup>1,4</sup>, Kathryn Tunyasuvunakool<sup>1,4</sup>, Russ Bates<sup>1,4</sup>, Augustin Žídek<sup>1,4</sup>, Anna Potapenko<sup>1,4</sup>, Alex Bridgland<sup>1,4</sup>, Clemens Meyer<sup>1,4</sup>, Simon A. A. Kohl<sup>1,4</sup>, Andrew J. Ballard<sup>1,4</sup>, Andrew Cowie<sup>1,4</sup>, Bernardino Romera-Paredes<sup>1,4</sup>, Stanislav Nikolov<sup>1,4</sup>, Rishabh Jain<sup>1,4</sup>, Jonas Adler<sup>1</sup>, Trevor Back<sup>1</sup>, Stig Petersen<sup>1</sup>, David Reiman<sup>1</sup>, Ellen Clancy<sup>1</sup>, Michał Zieliński<sup>1</sup>, Martin Steinegger<sup>2,3</sup>, Michałina Pacholska<sup>1</sup>, Tamas Berghammer<sup>1</sup>, Sebastian Bodenstein<sup>1</sup>, David Silver<sup>1</sup>, Oriol Vinyals<sup>1</sup>, Andrew W. Senior<sup>1</sup>, Koray Kavukcuoglu<sup>1</sup>, Pushmeet Kohli<sup>1</sup> & Demis Hassabis<sup>1,4</sup>





## 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.



*Knut och Alice  
Wallenbergs  
Stiftelse*



Vetenskapsrådet



LUNDS  
UNIVERSITET