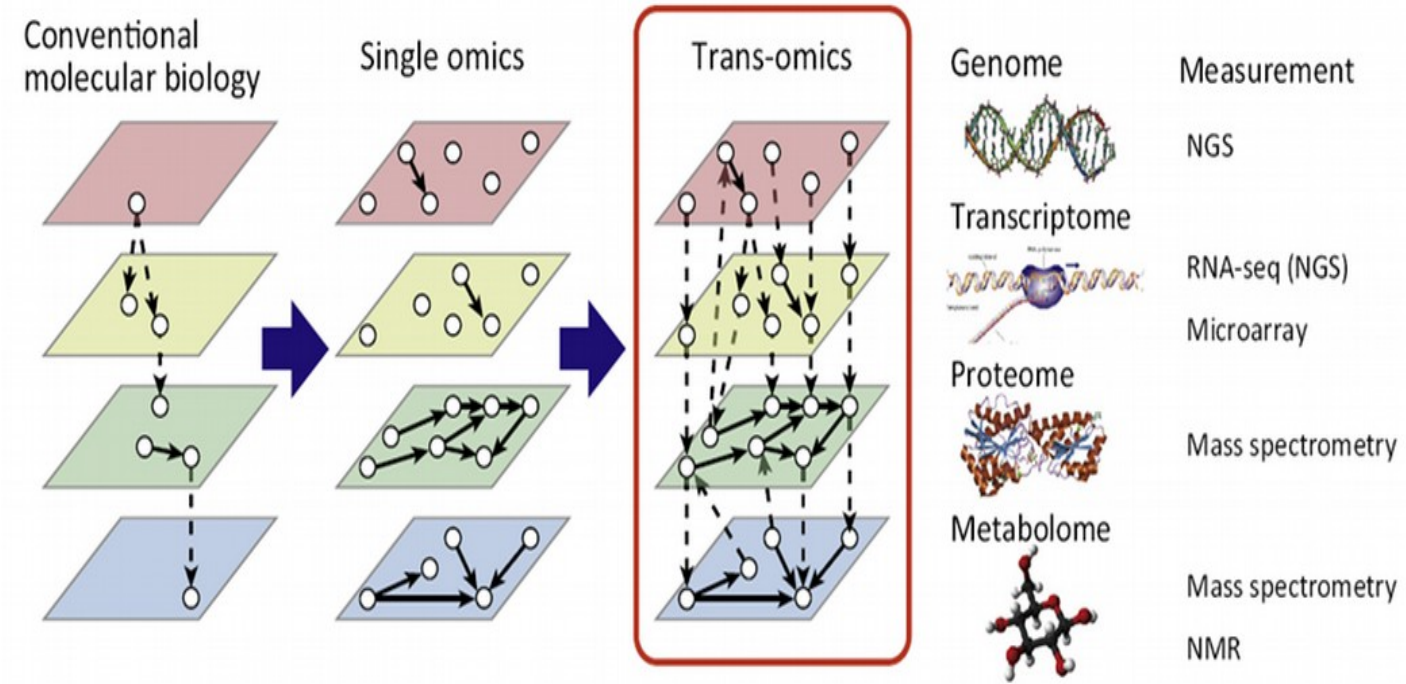
Supervised Machine Learning MultiOmics Integration

Physalia course, online via zoom

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

How does Machine Learning work?

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

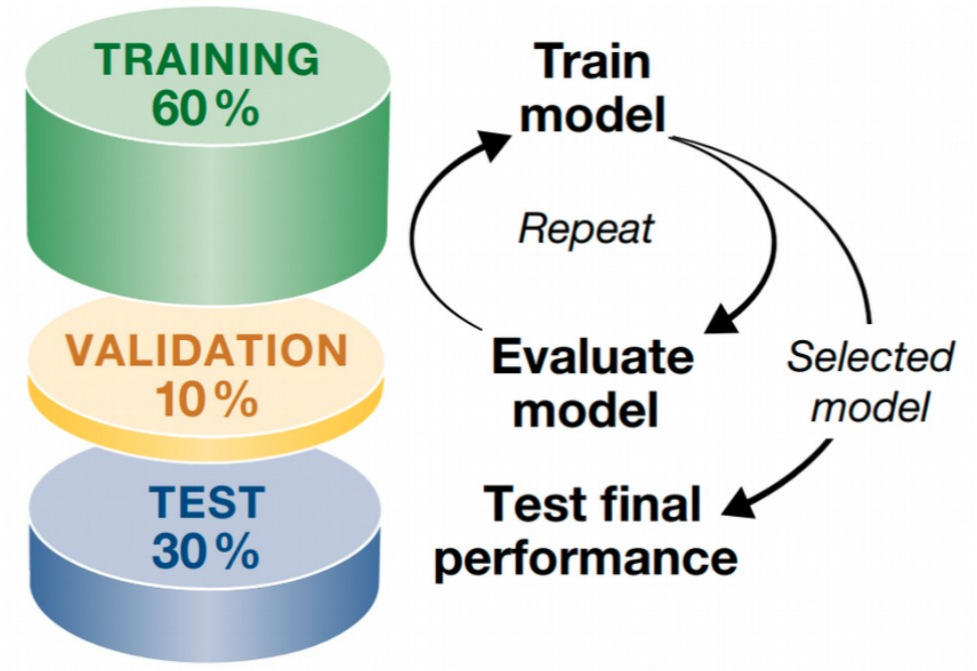


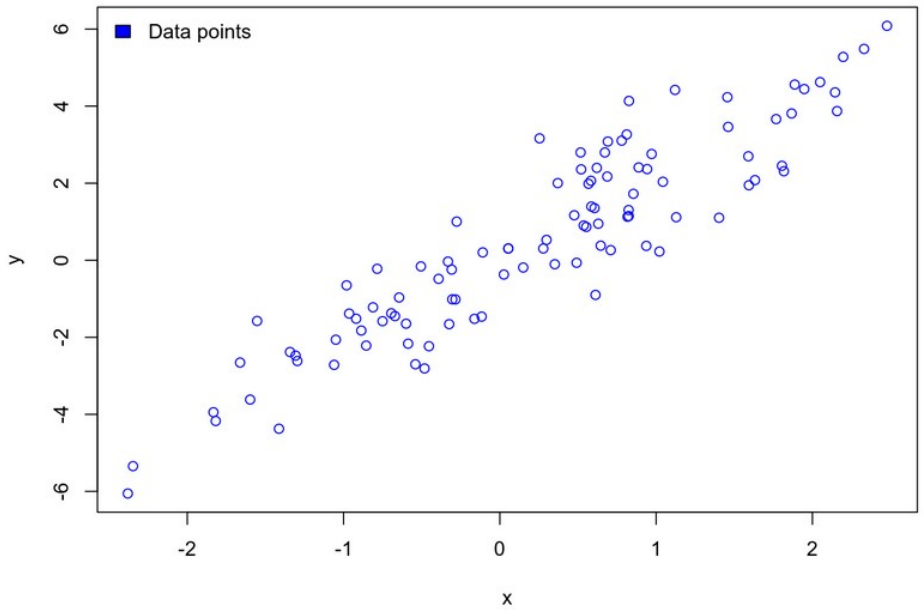
Image adapted from Angenmueller et al., Mol. Syst. Biol. 2016

Machine Learning typically involves five basic steps:

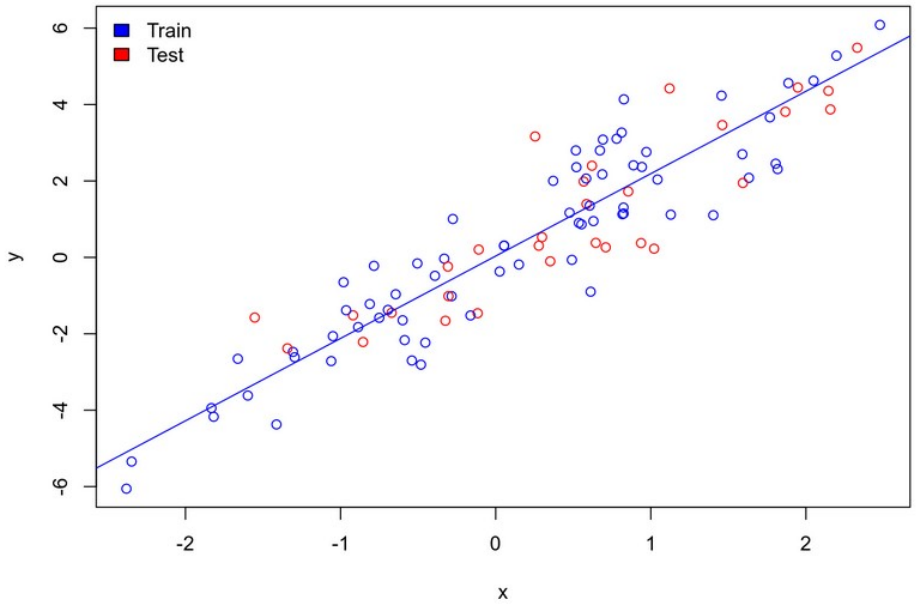
1. Split data set into **train**, **validation** and **test** subsets
2. Fit the model on the train subset
3. Validate your model on the validation subset
4. Repeat train - validation split many times and tune **hyperparameters**
5. Test the accuracy of the optimized model on the test subset.

Toy example of supervised machine learning

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

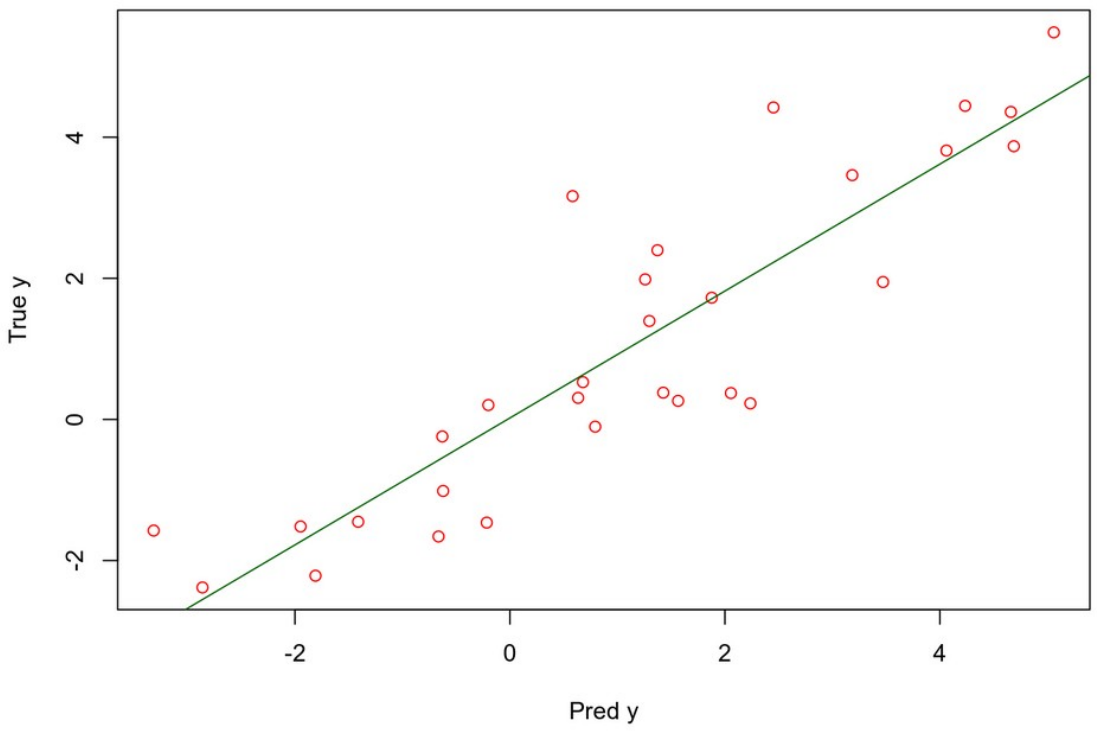



```r
train <- df[sample(1:dim(df)[1], 0.7 * dim(df)[1]), ]
test <- df[!rownames(df) %in% rownames(train), ]
df$col <- ifelse(rownames(df) %in% rownames(test), "red", "blue")
plot(y ~ x, data = df, col = df$col)
legend("topleft", c("Train","Test"), fill=c("blue","red"), bty="n")
abline(lm(y ~ x, data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



Machine learning model validation

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




```r
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

Multi-Omics from human pancreatic islets

www.nature.com/scientificreports

scientific reports

OPEN

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

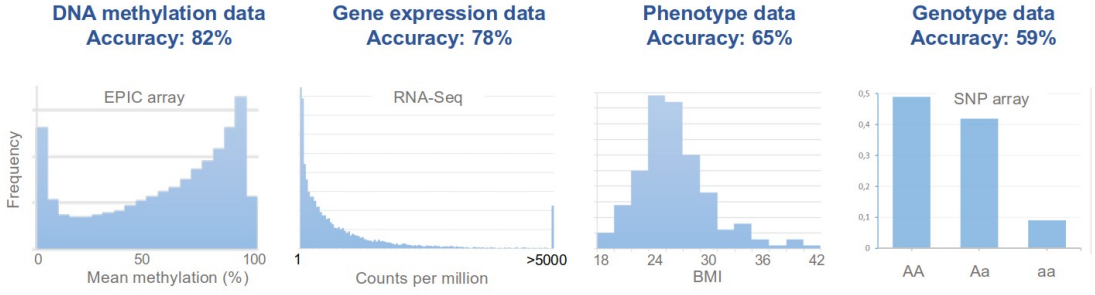
Tina Rönn[1], Alexander Perfilyev[1], Nikolay Oskolkov[2,3] & Charlotte Ling[1,3]✉

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-sequening, 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-sequencing, and SNPs were analyzed using HumanOmniExpress arrays. Supervised linear multiOmics integration via DIABLO based on Partial Least Squares (PLS) achieved an accuracy of 91 ± 15% of T2D prediction with an area under the curve of 0.96 ± 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 *ANO1*, 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-sequencing, 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[1]. 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[2]. 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-sequencing (RNA-Seq)[3–7] and DNA methylation analysis[8–11] in human islets from 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[10,12]. 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

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



LUND UNIVERSITY | LUDC DIABETES CENTRE

Search this site | Search | Listen

LUDC | Research groups | LUDC-IRC | Resources | Courses and training | For early career researchers

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 upon agreed consent instead dedicated to research and distributed to research centers

Related information

The nordic network for clinical islets transplantation

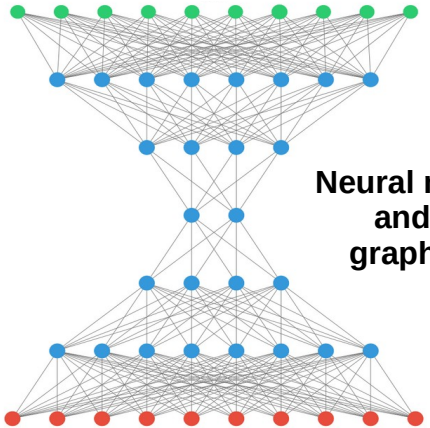
Publications

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

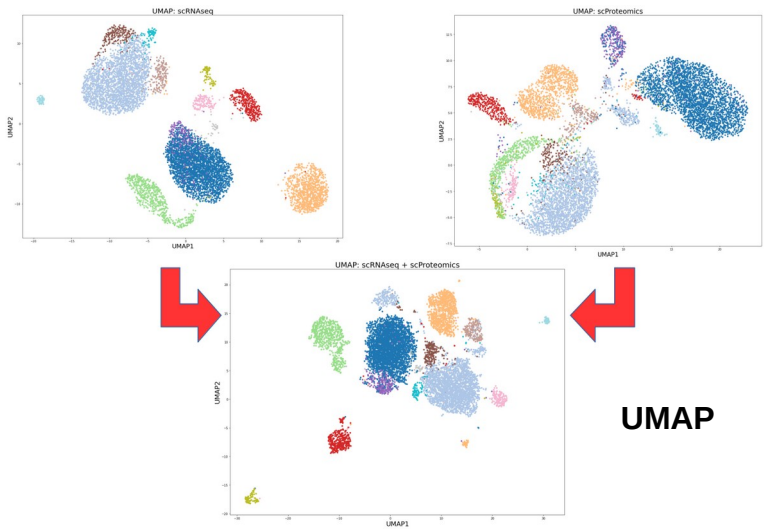
Publications

Overview of Integrative Methods

Convert to common space

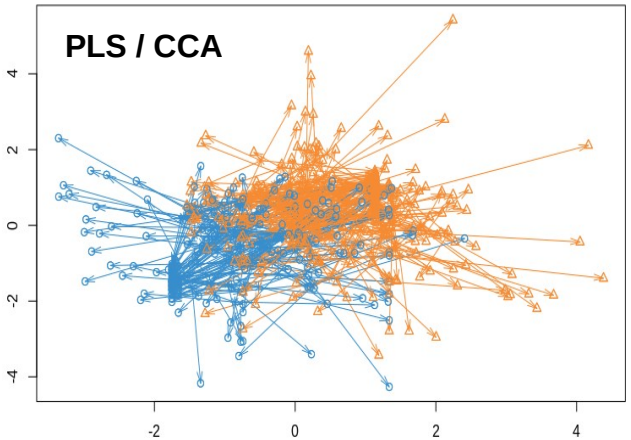


Neural nets and graphs

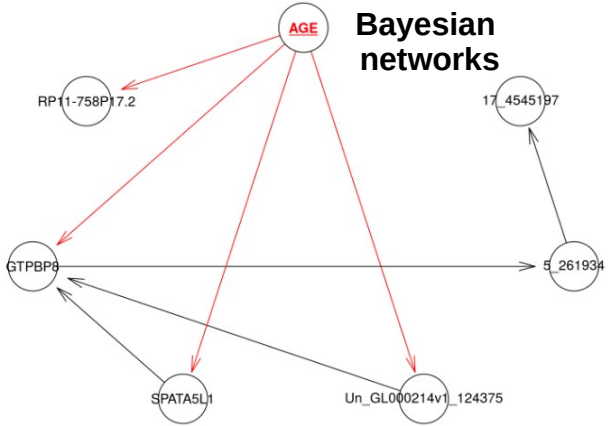


UMAP

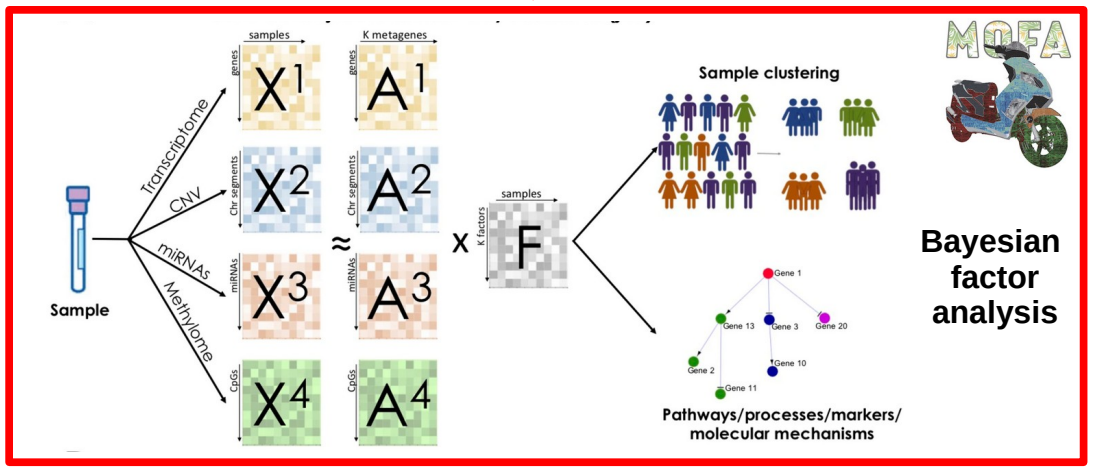
Extract common variation



Combine via Bayes rule



Bayesian networks



Bayesian factor analysis

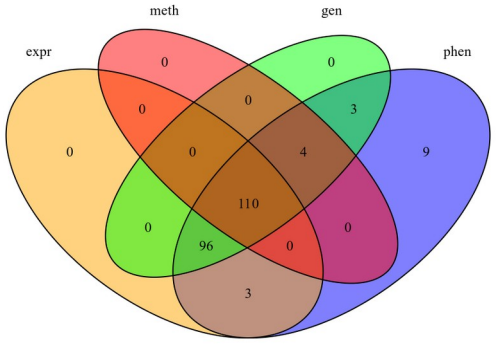
What Method of Integration to Select?

	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

For Example:

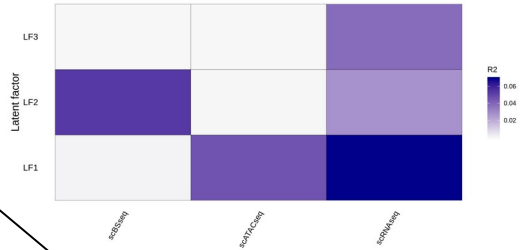
1) With ~100 samples it is a good idea to do **linear** Omics integration
2) T2D is a phenotype of interest, therefore **supervised** integration

Typical Workflow for Omics Integration



Data Set (4 Omics)
110 overlapping individuals

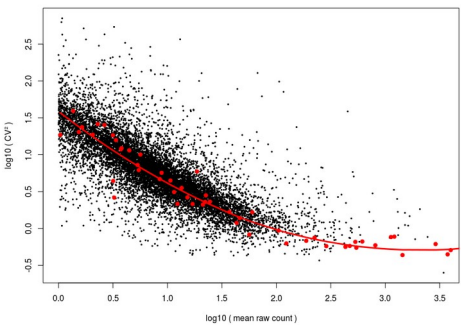
Check covariance



Train Set (n = 80)

Test Set (n = 30)

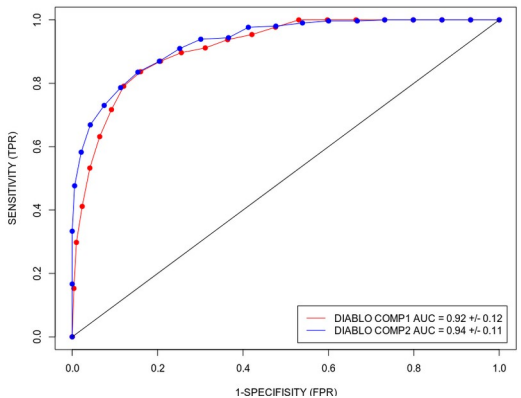
Unsupervised:
remove low-variance



Supervised:
LASSO

Feature Selection

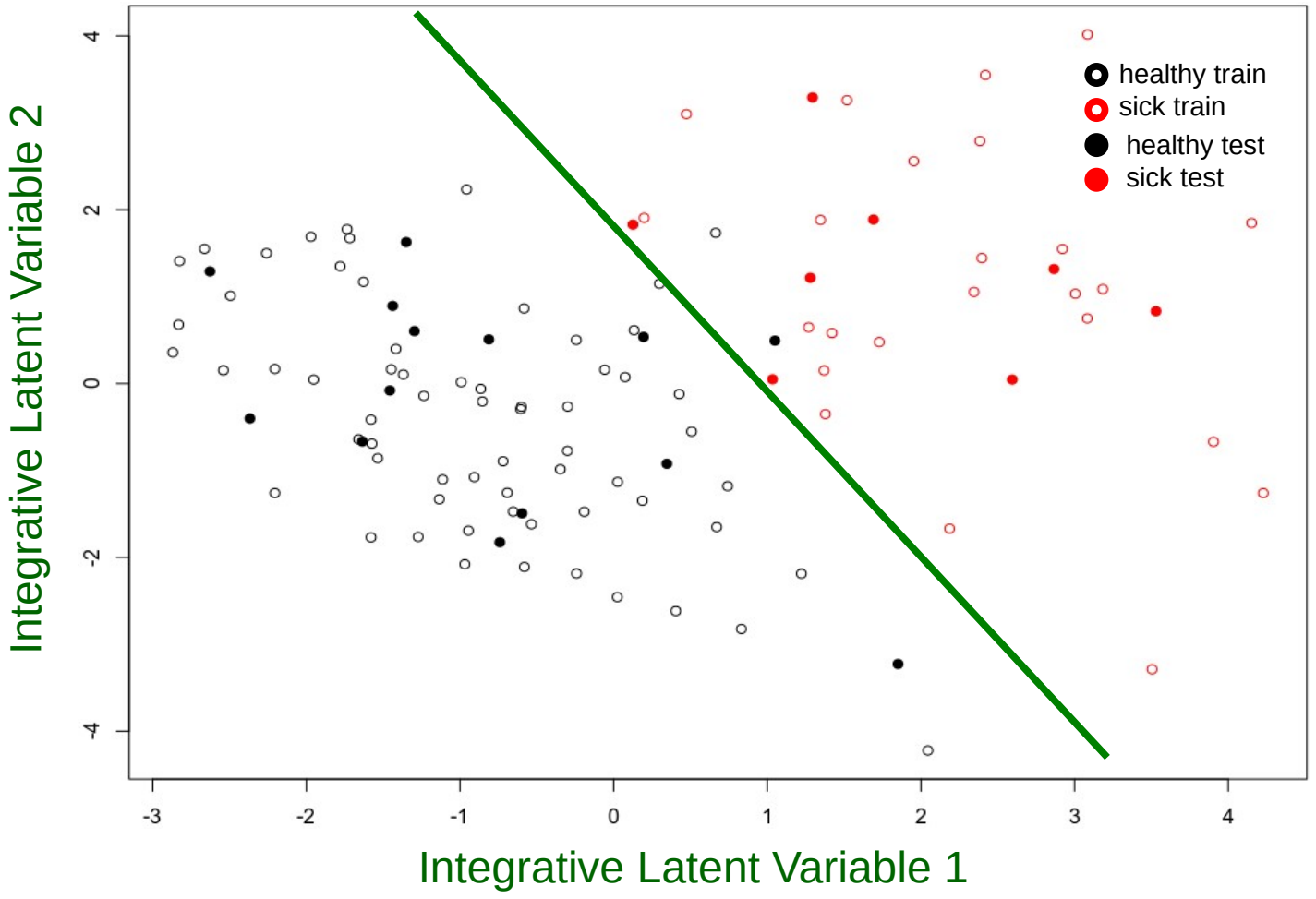
Evaluation

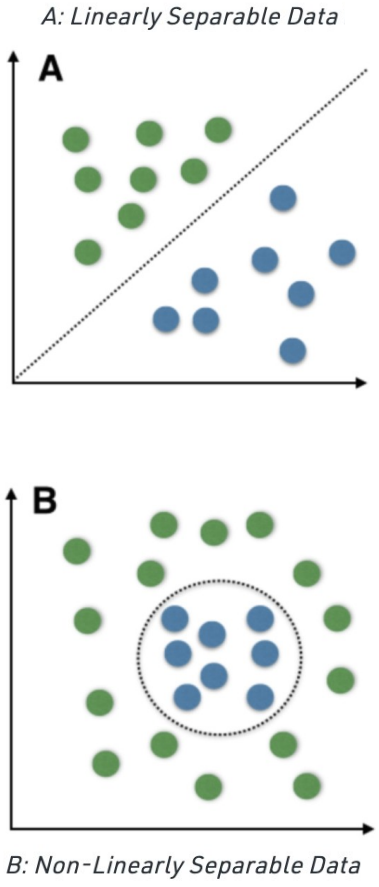


Omics Integration

Trained Model

Linear Separation: Decision Boundary



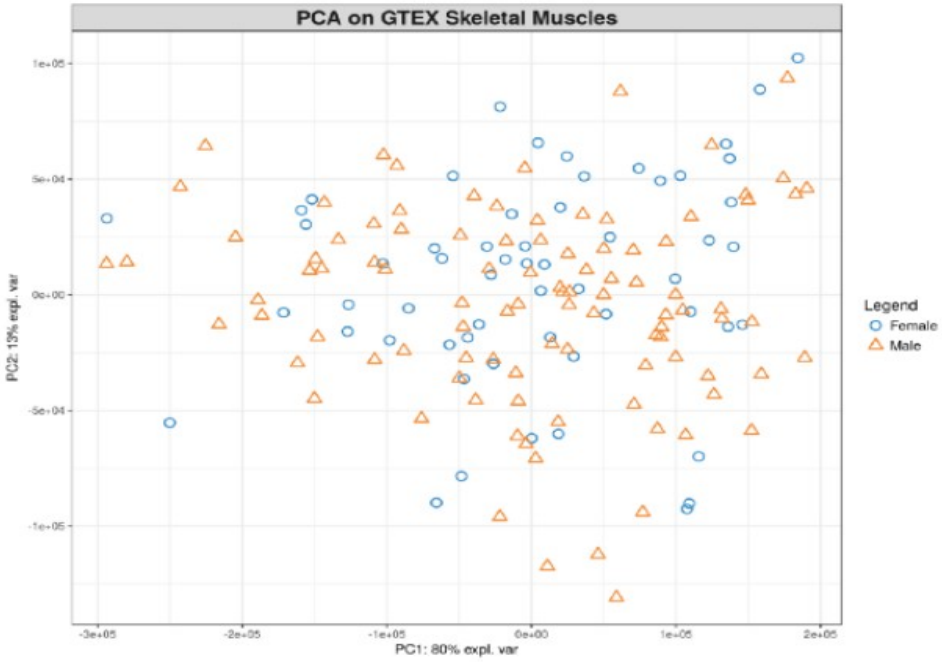


Univariate and Multivariate Feature Selection

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
```

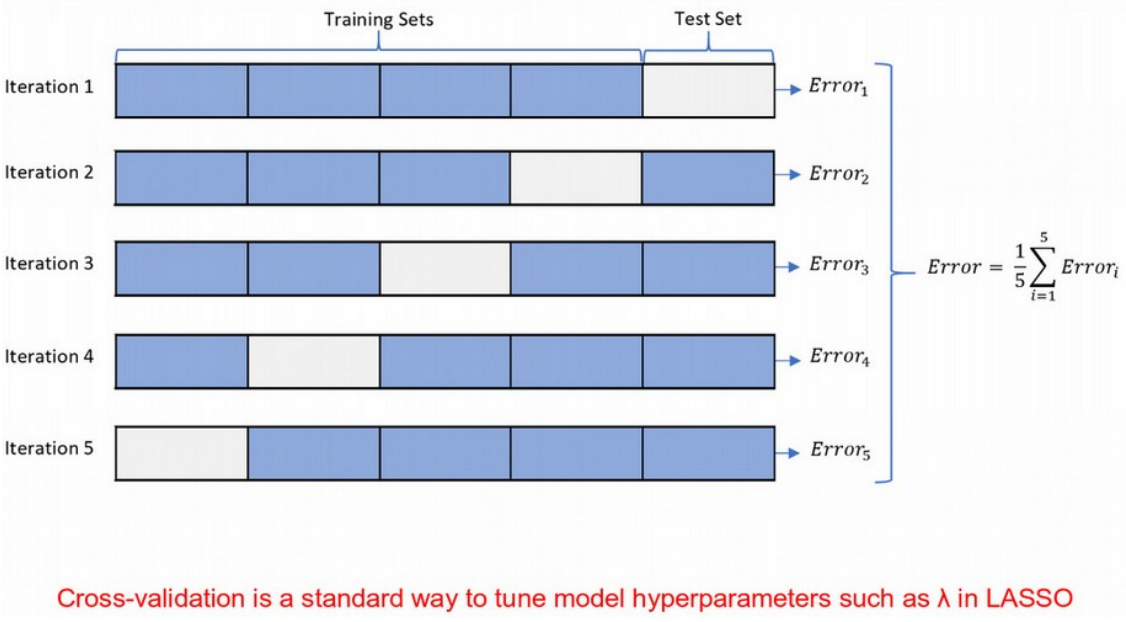
Generally acknowledged that univariate feature selection has poor predictive capacity compared to multivariate feature selection

Multivariate Feature Selection: LASSO

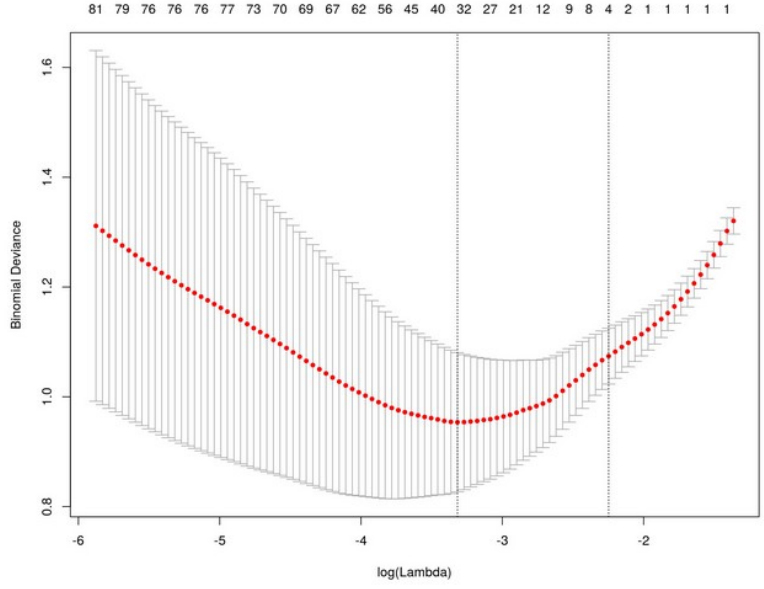
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

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

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



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



Regularizations are Priors in Bayesian stats

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

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

$$\mathrm{L}(\mathrm{Y} \mid \beta_1, \beta_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{(\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2}{2\sigma^2}}$$

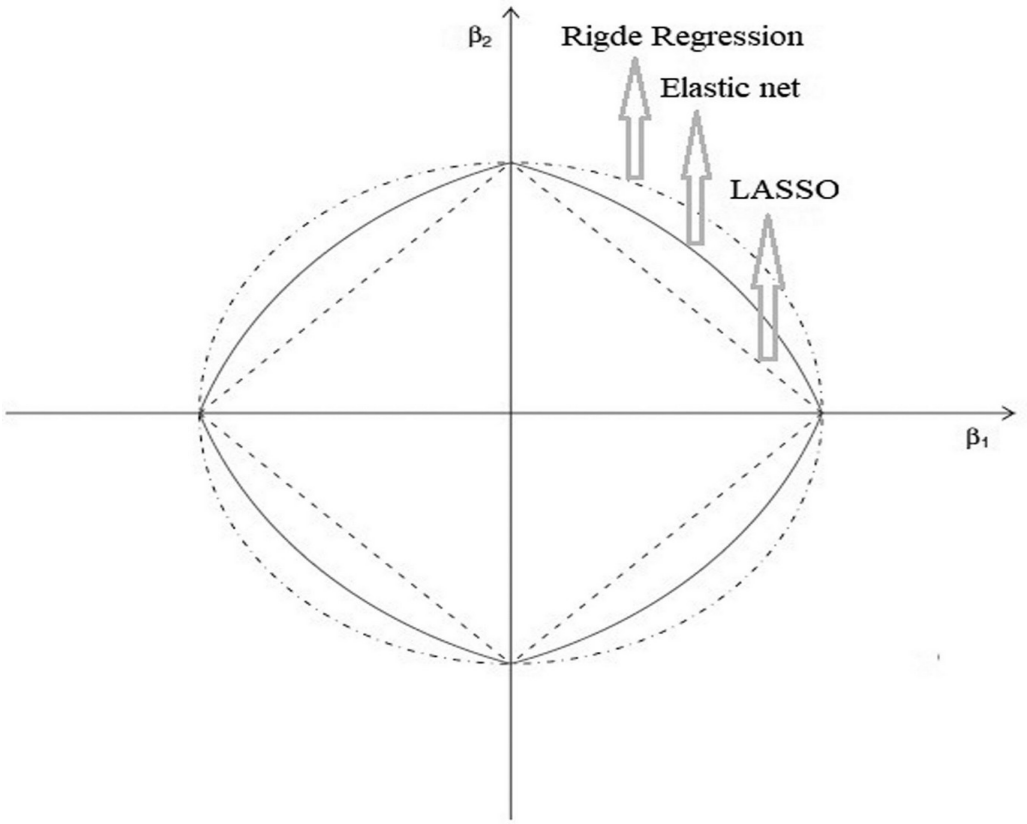
- **Bayes theorem**: maximize posterior probability of observing parameters given data:

$$\mathrm{Posterior}(\mathrm{params} \mid \mathrm{data}) = \frac{\mathrm{L}(\mathrm{data} \mid \mathrm{params}) * \mathrm{Prior}(\mathrm{params})}{\int \mathrm{L}(\mathrm{data} \mid \mathrm{params}) * \mathrm{Prior}(\mathrm{params})\, \mathrm{d}(\mathrm{params})}$$

$$\mathrm{Posterior}(\beta_1, \beta_2 \mid \mathrm{Y}) \sim \mathrm{L}(\mathrm{Y} \mid \beta_1, \beta_2) * \mathrm{Prior}(\beta_1, \beta_2) \sim \exp^{-\frac{(\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2}{2\sigma^2}} * \exp^{-\lambda(|\beta_1| + |\beta_2|)}$$

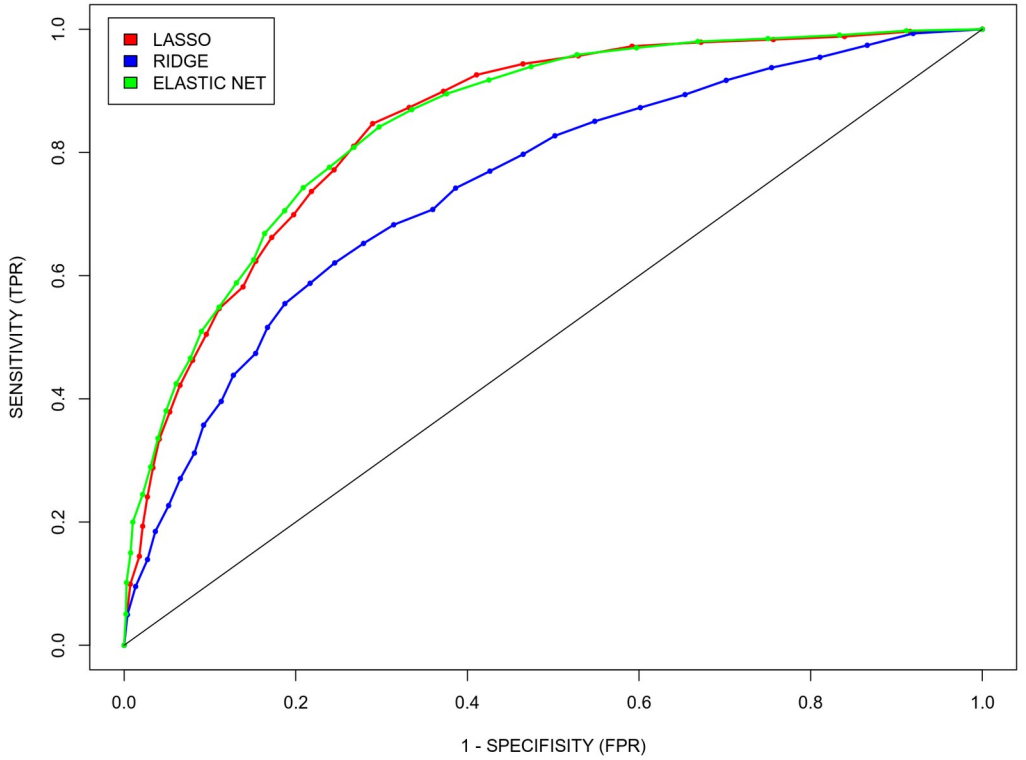
$$-\log[\mathrm{Posterior}(\beta_1, \beta_2 \mid \mathrm{Y})] \sim (\mathrm{Y} - \beta_1 \mathrm{X}_1 - \beta_2 \mathrm{X}_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$

Lasso vs. Ridge vs. Elastic Net



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

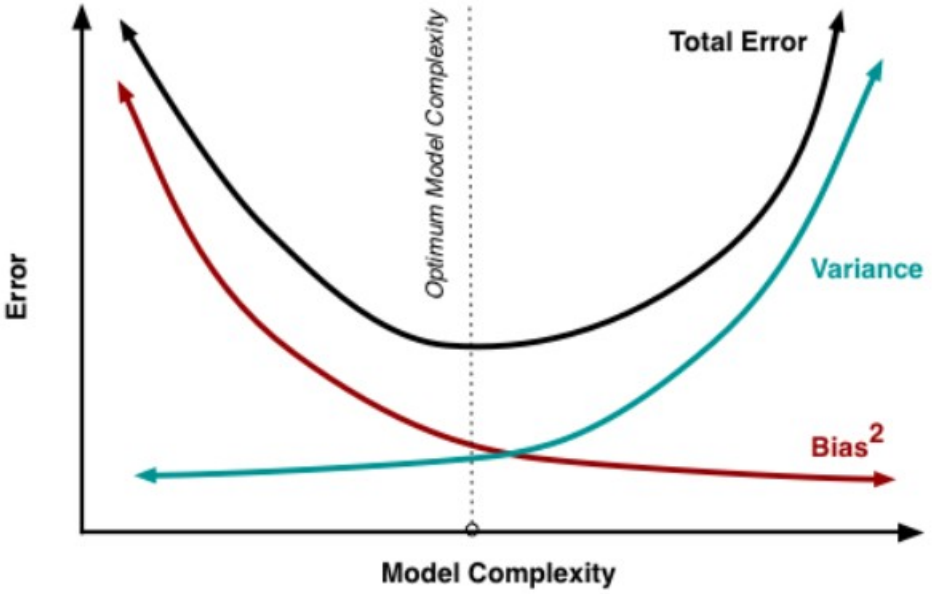
$$\text{Ridge} : \beta_1^2 + \beta_2^2 \leq \lambda$$

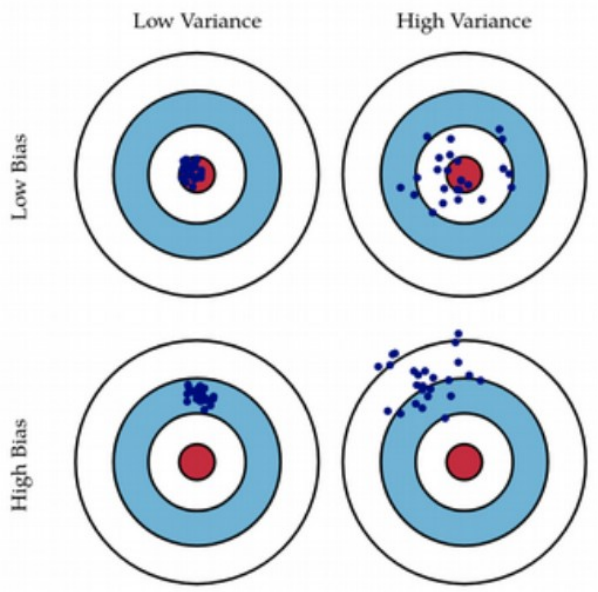


Lasso is more conservative

Ridge is more permissive

Penalized regression interpretation

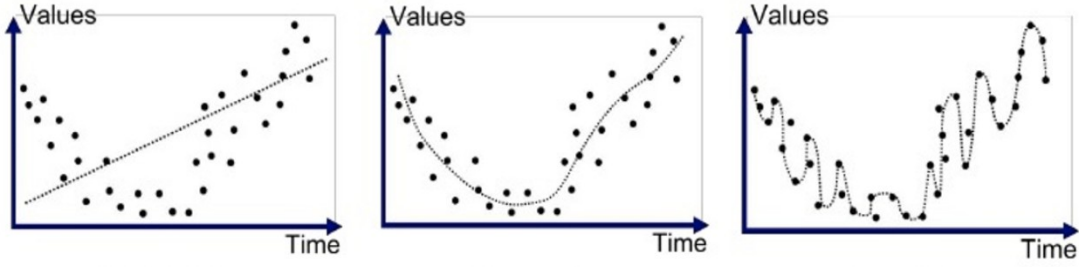




$$Y = f(X) \Longrightarrow \text{Reality}$$

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

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



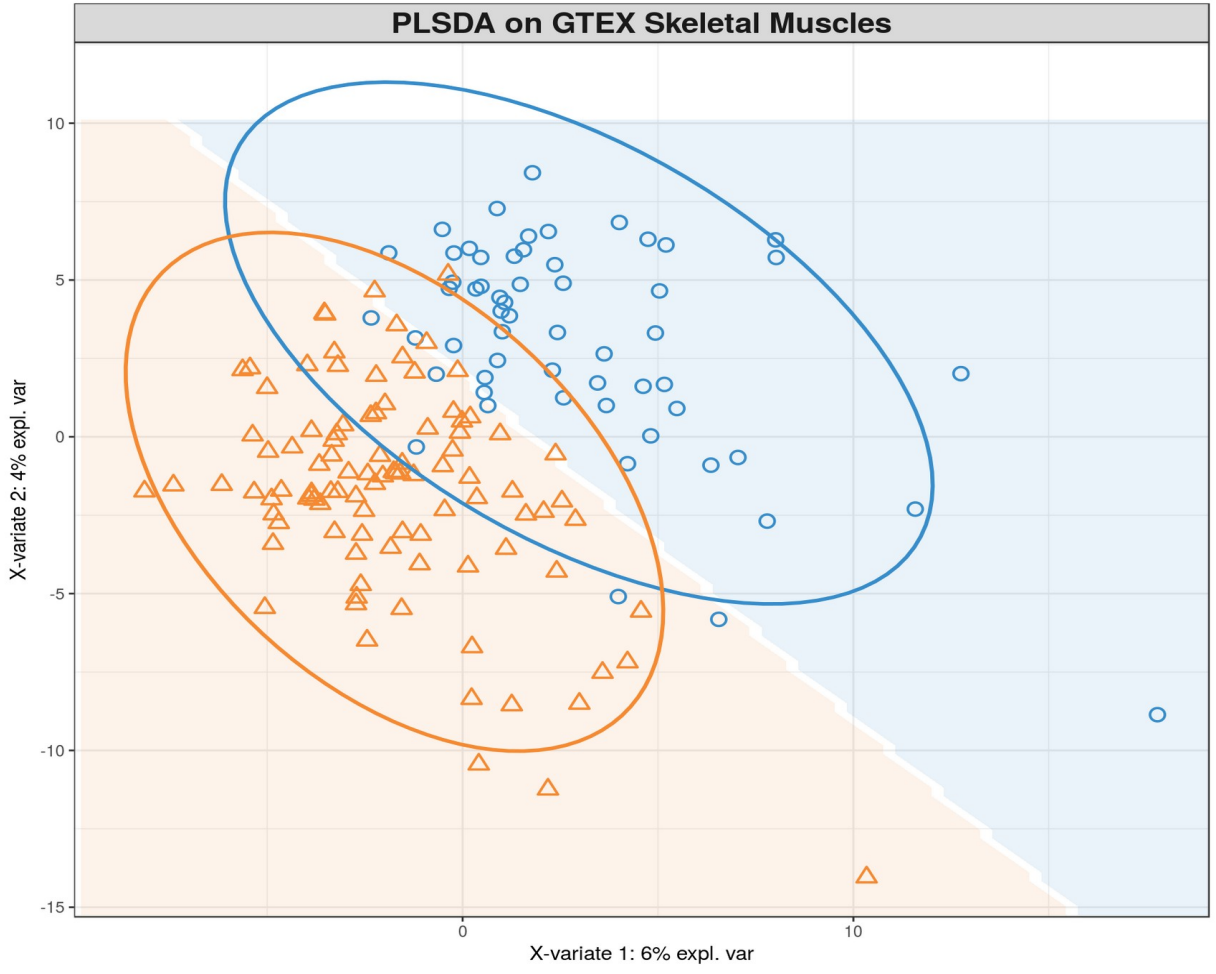
LASSO – high bias, low variance

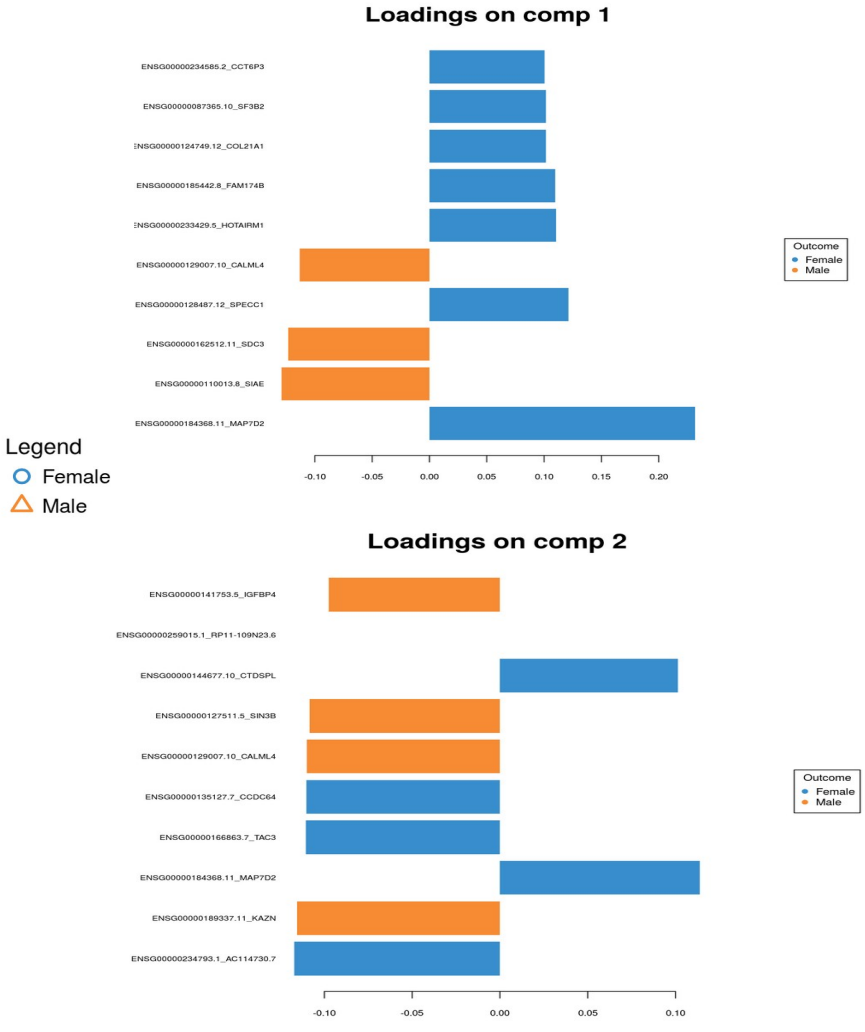
Underfitted

Good Fit/Robust

Overfitted

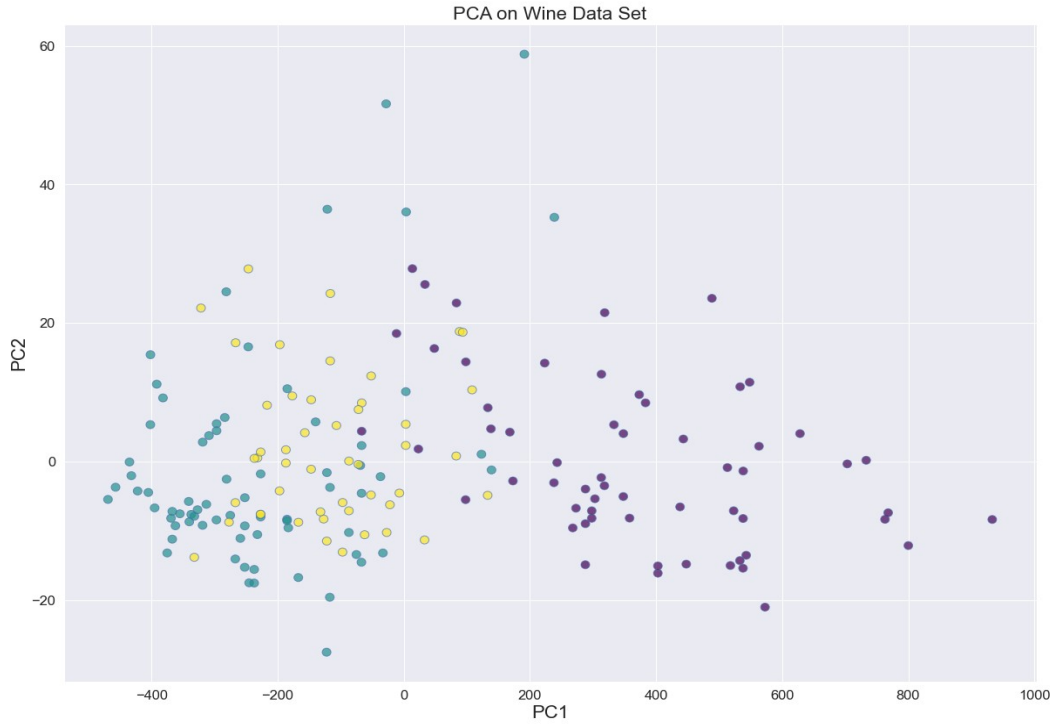
Multivariate Feature Selection: PLS

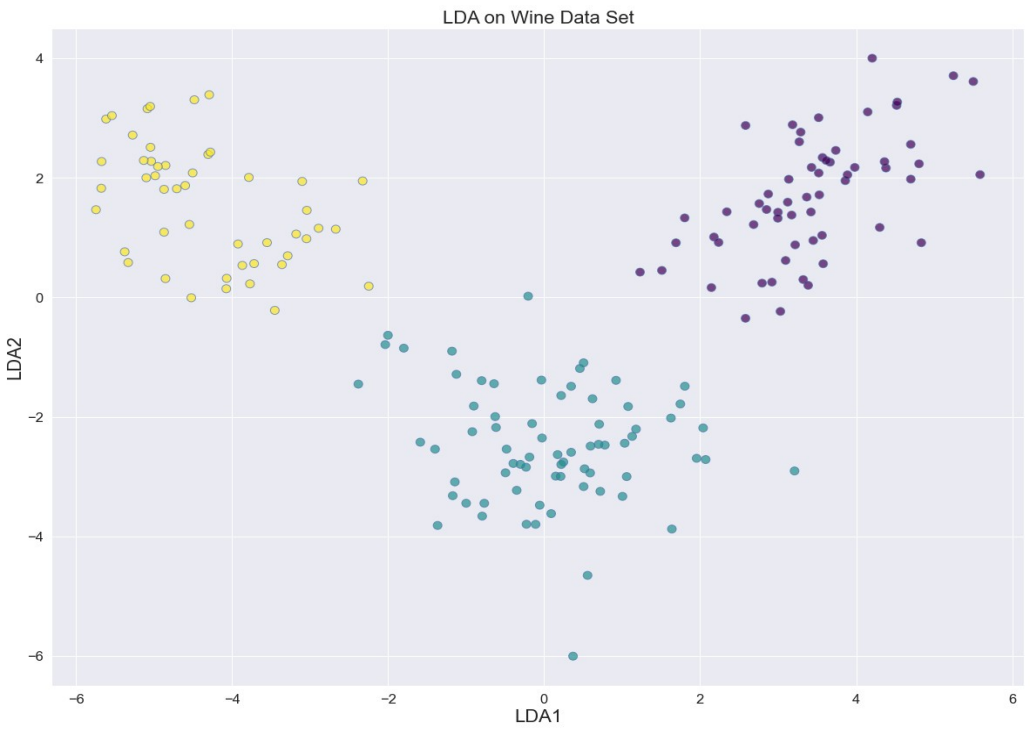




Select features that separate two groups of samples the most

Multivariate Feature Selection: Linear Discriminant Analysis (LDA)

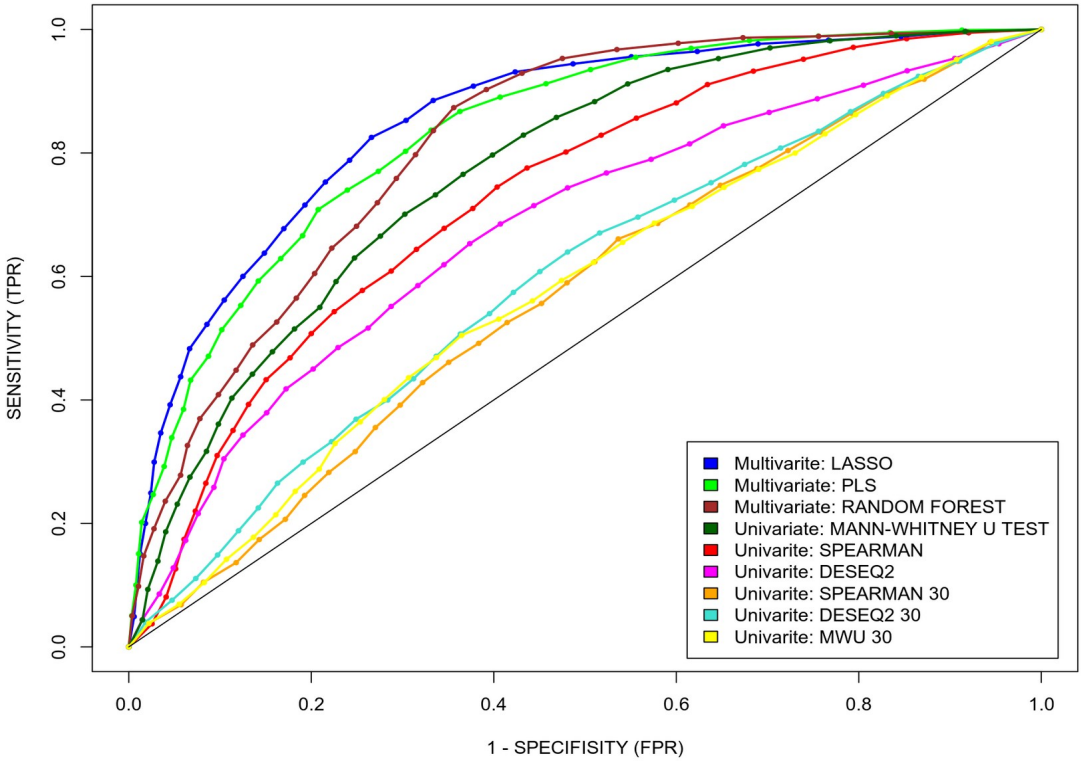




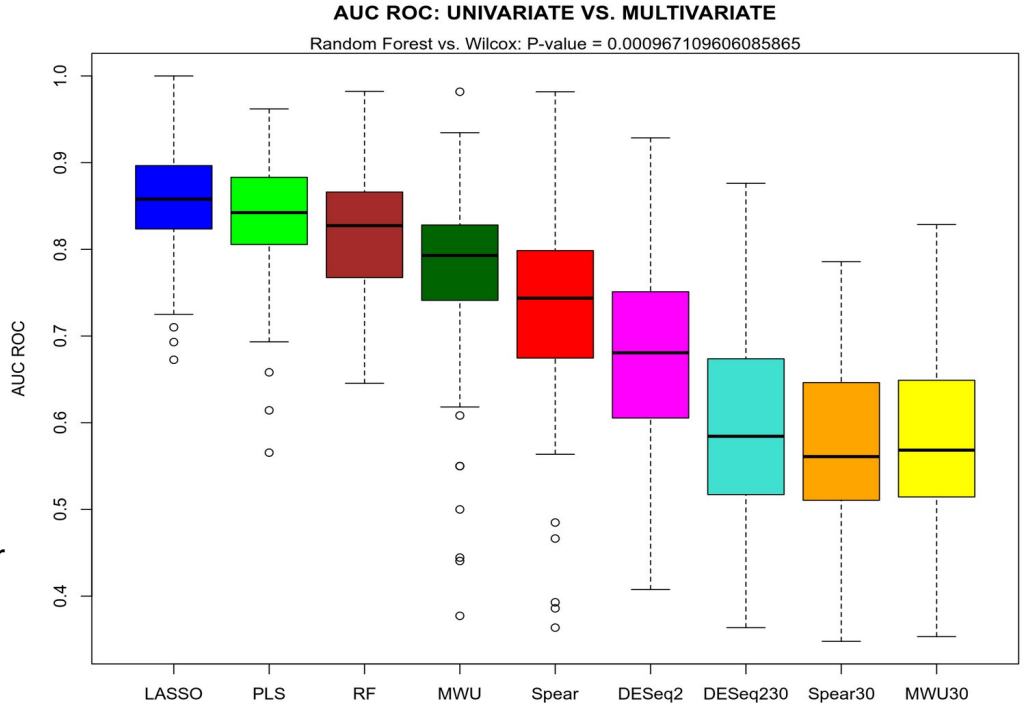
Minimize variance within clusters and maximize variance between clusters

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

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

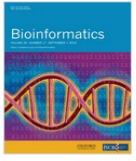
DIABLO Omics Integration

OXFORD ACADEMIC | Journals | Books

Bioinformatics

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Issues | Advance articles | Submit | Alerts | About | Bioinformatics | Search | Advanced Search



Volume 35, Issue 17
1 September 2019

Article Contents

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**

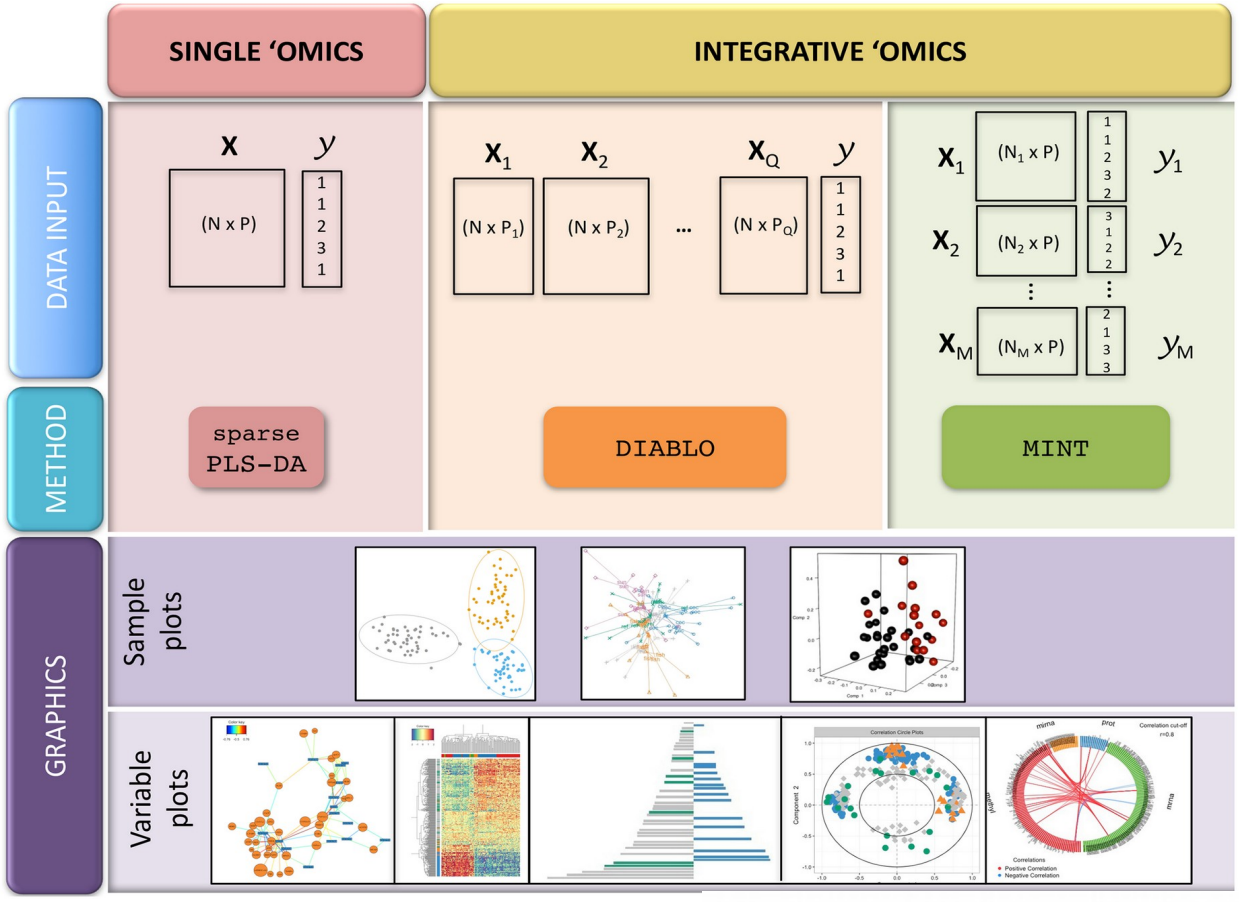
PDF | Split View | Cite | Permissions | Share

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 *D*ata *I*ntegration *A*nalysis for *B*iomarker discovery using *L*atent c*O*mponents (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.

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 $h = 1, \ldots, H$:

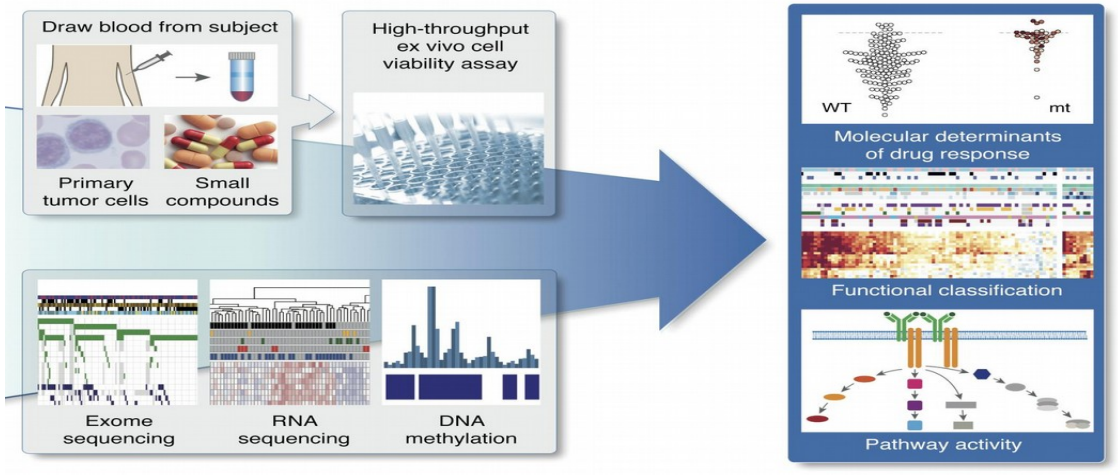
$$\max_{a_h^{(1)},\ldots,a_h^{(Q)}} \sum_{i,j=1, i\neq j}^{Q} c_{i,j}\ \mathrm{cov}(X_h^{(i)} a_h^{(i)}, X_h^{(j)} a_h^{(j)}), \quad (1)$$

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

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.

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

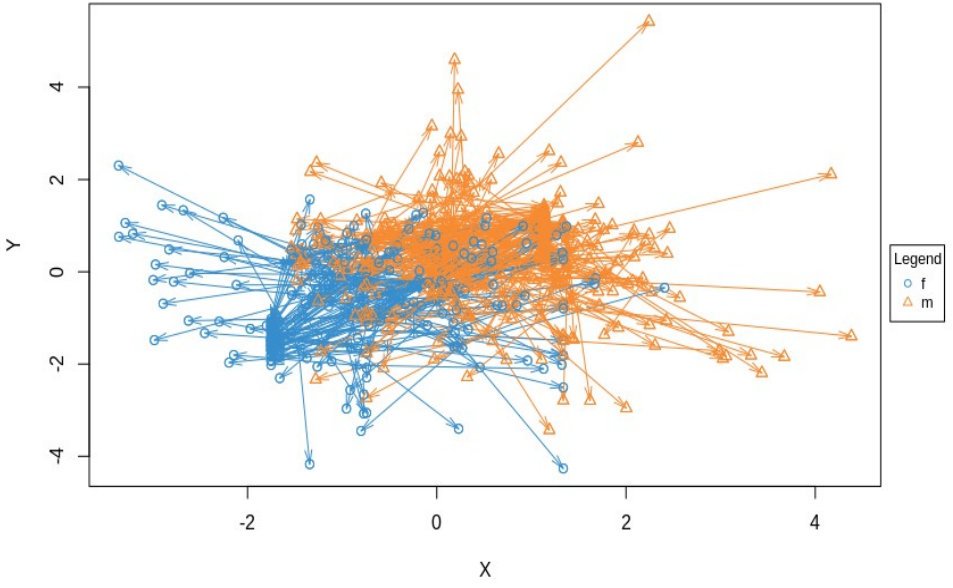
DIABLO Omics Integration

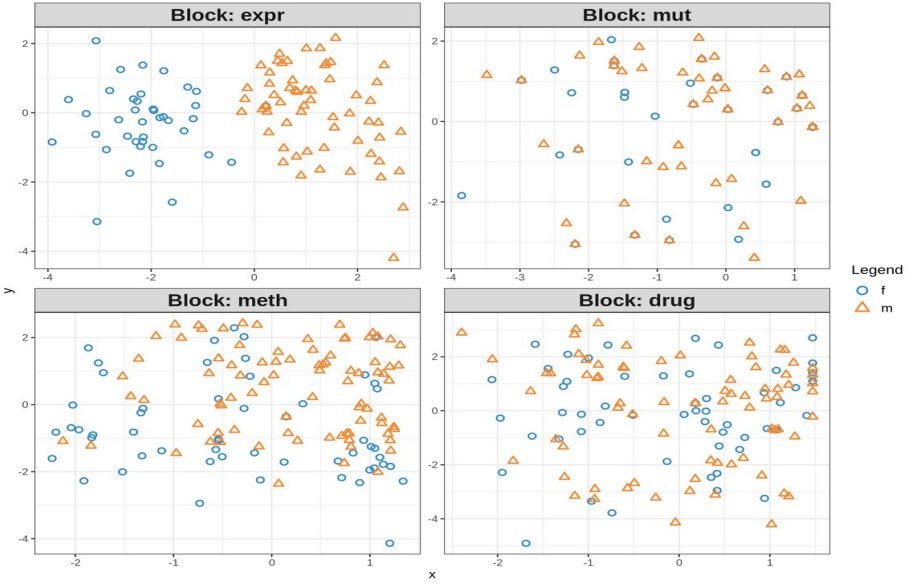


Chronic Lymphocytic Leukaemia (CLL):

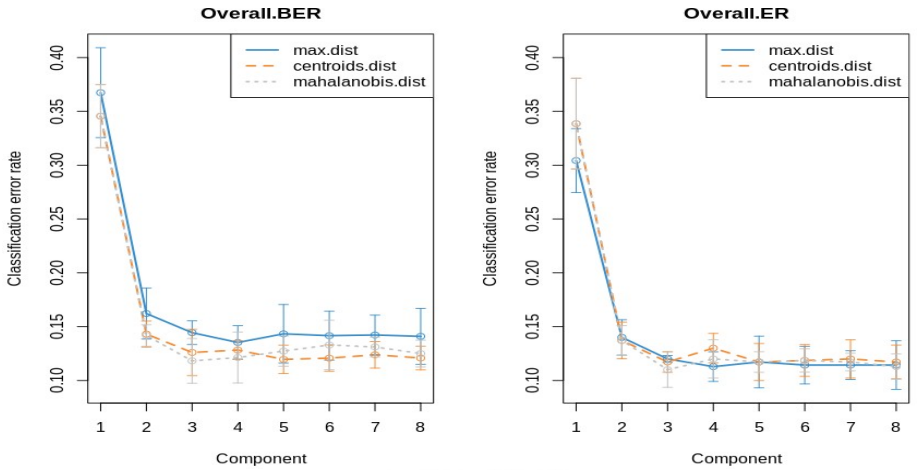
gene expression (RNAseq), mutations, methylation, drug response

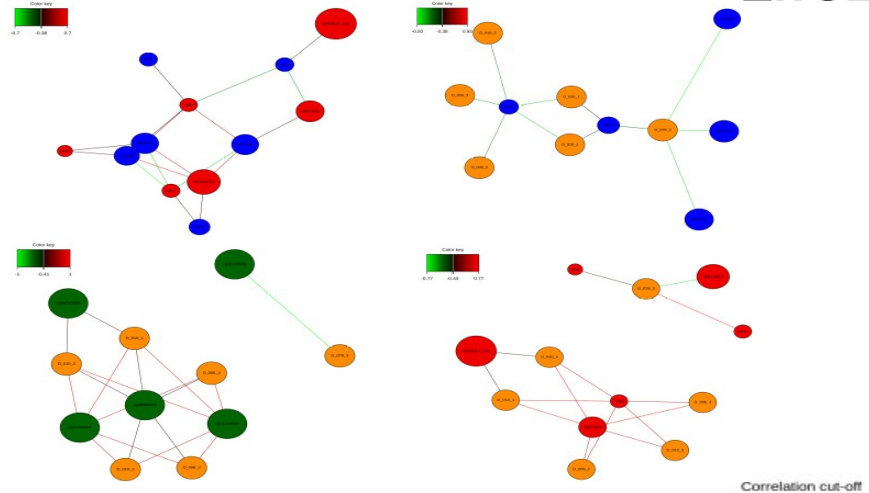
Dietrich et al., J Clin Invest. 2018

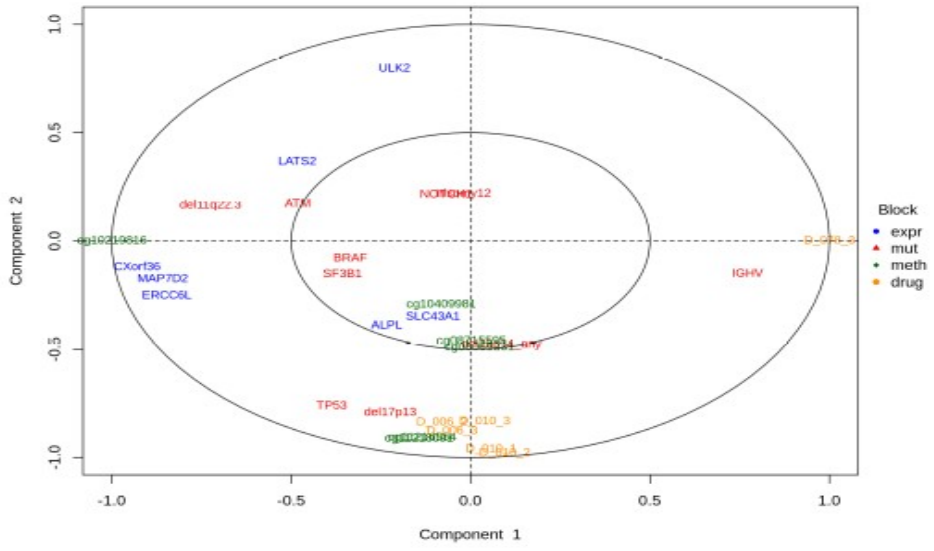


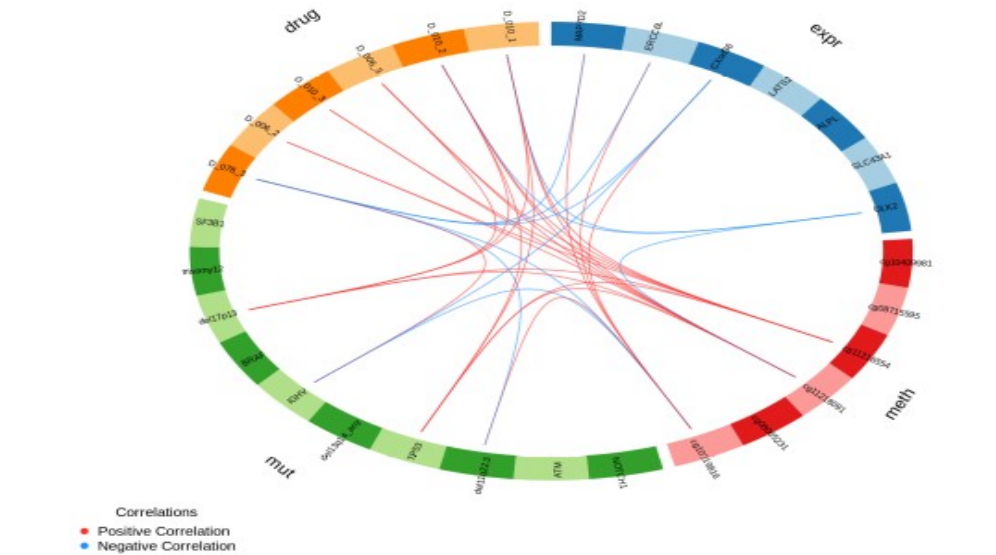


DIABLO visualization

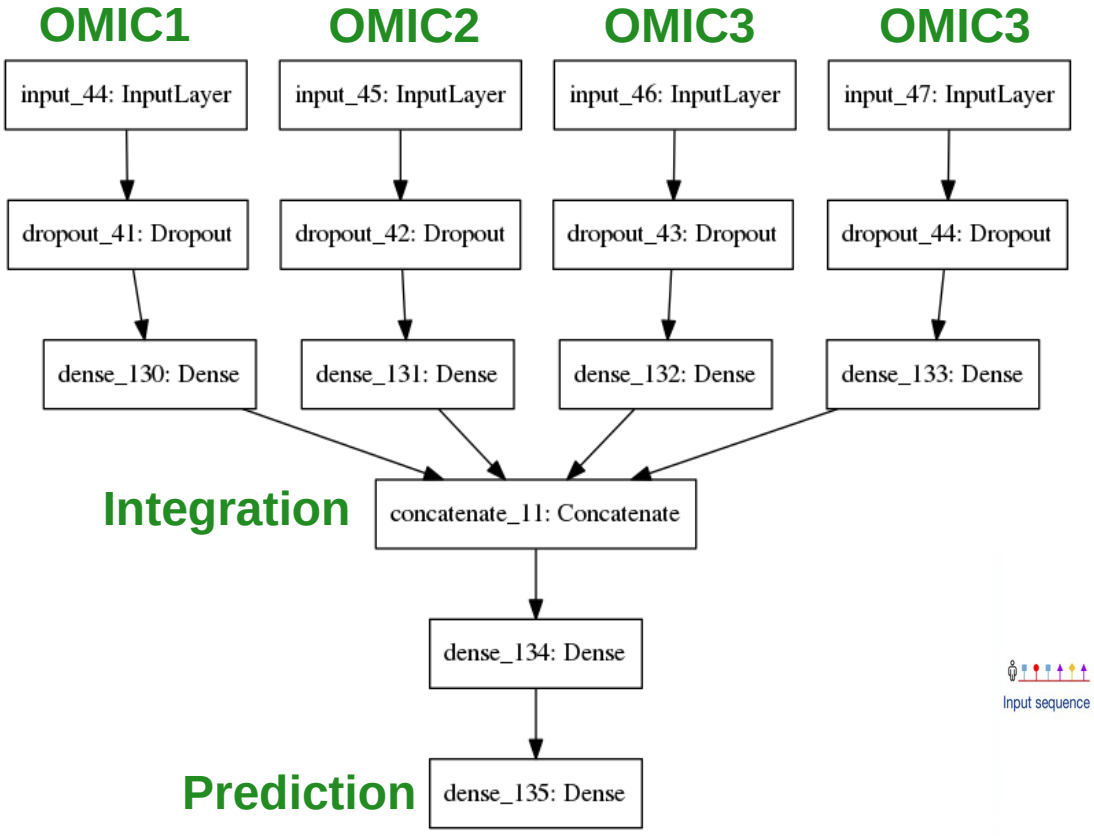








Supervised deep learning Omics integration



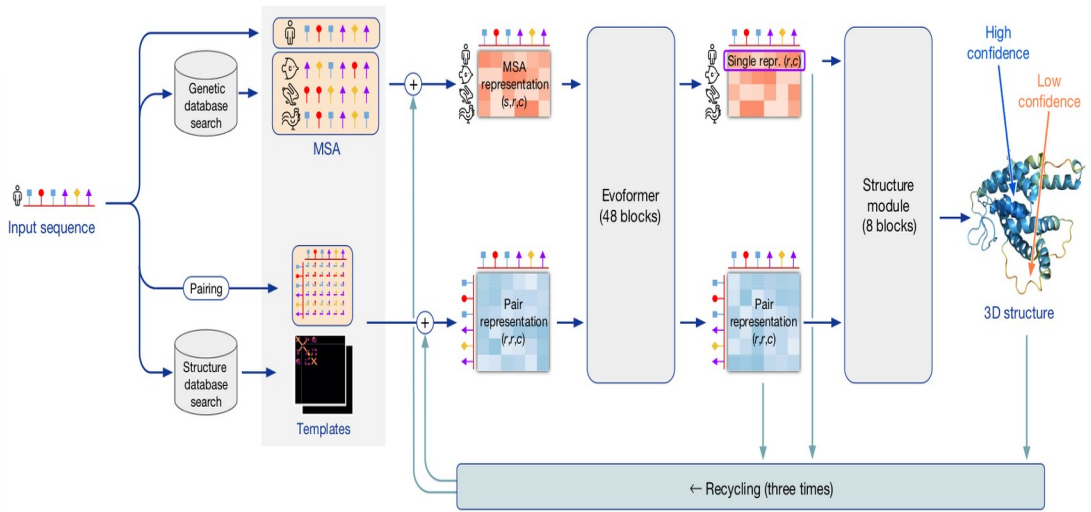
Backward propagation ensures that the parameters across Omics are optimized in context of each other

Article

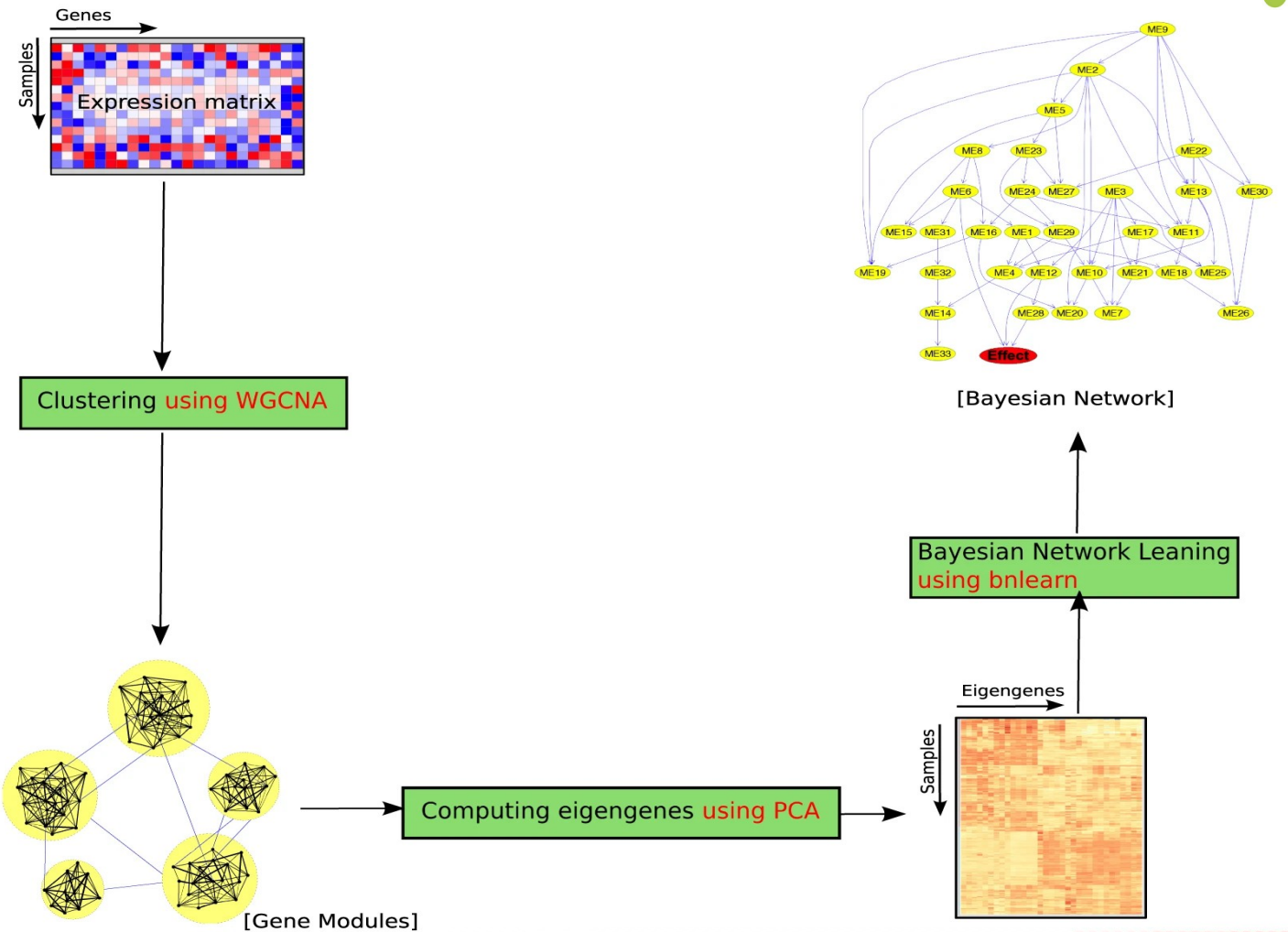
Highly accurate protein structure prediction with AlphaFold

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



Bayesian Networks for Omics Integration



Agrahari R. et al., Sci Rep. 2018 3;8(1):6951

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.

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet