PRECISION HEALTH
ANALYSIS BOOTCAMP

Unsupervised multiomics data integration

Amrit Singh, PhD

Department of Anesthesiology, Pharmacology and Therapeutics, UBC

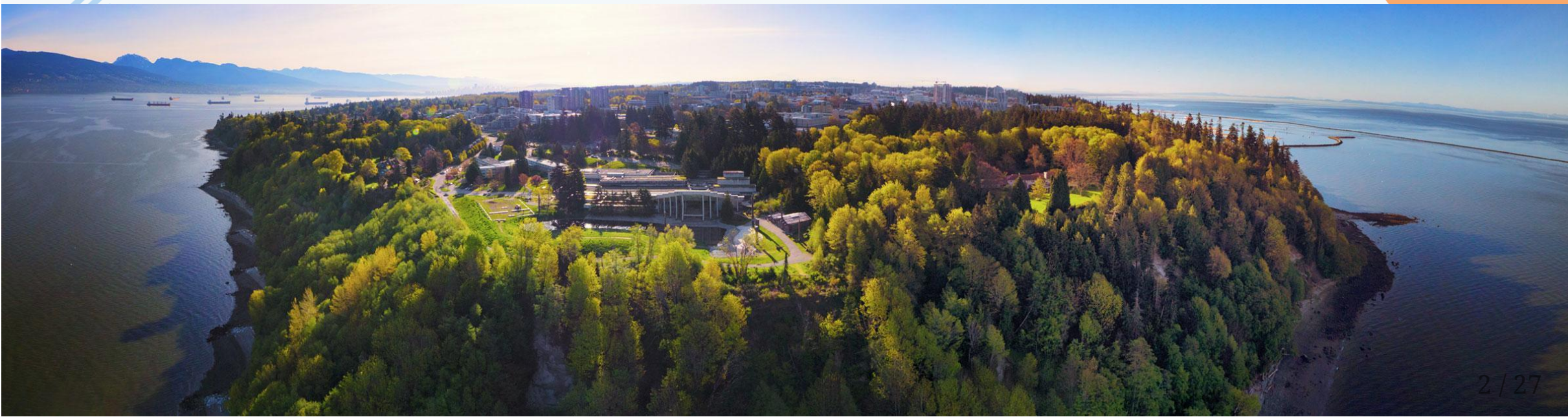
Centre for Heart Lung Innovation

August 08, 2022, 2022 | 09:00-11:00

lab
code
asingh_22g

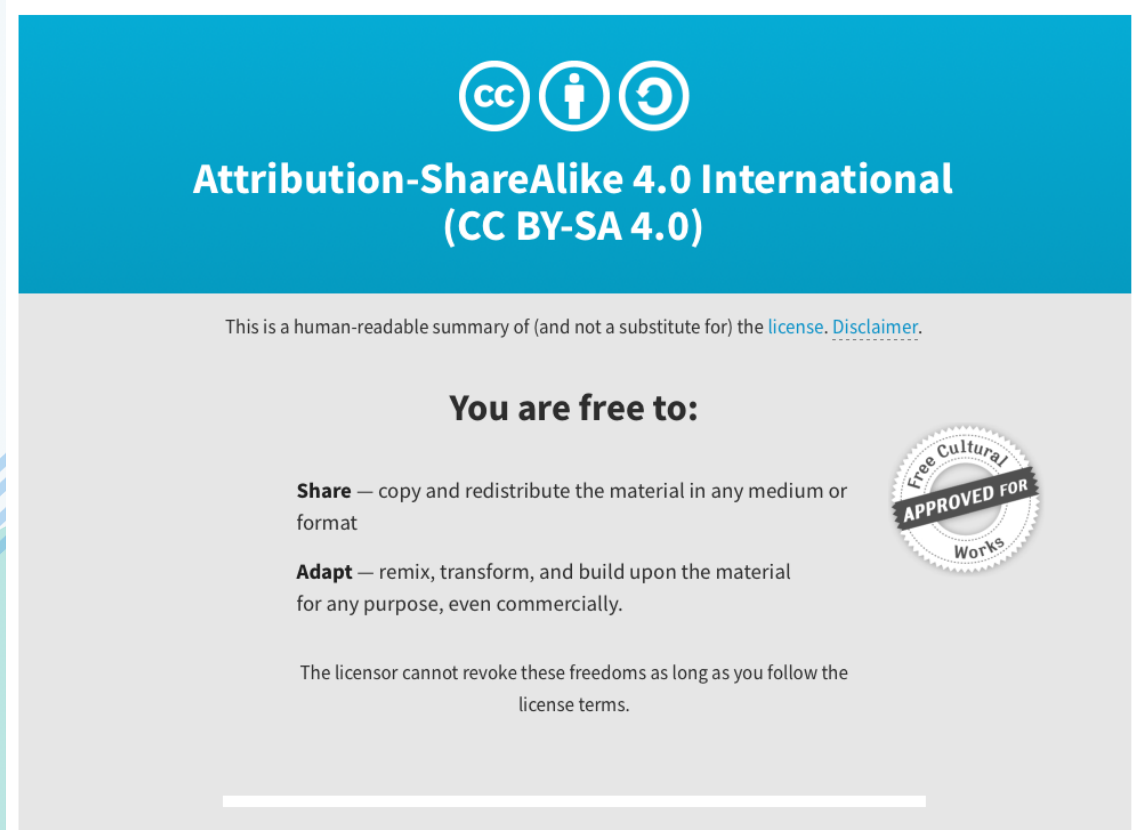
PRECISION HEALTH
ANALYSIS BOOTCAMP

We would like to begin by acknowledging that the land on which we gather is the traditional, ancestral, and unceded territory of the xwməθkwəy̓əm (Musqueam) People.



PRECISION HEALTH
ANALYSIS BOOTCAMP

Copyright Information



Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Free Cultural APPROVED FOR Works

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Read more here:
https://creativecommons.org/licenses/by-sa/4.0/

Learning outcomes

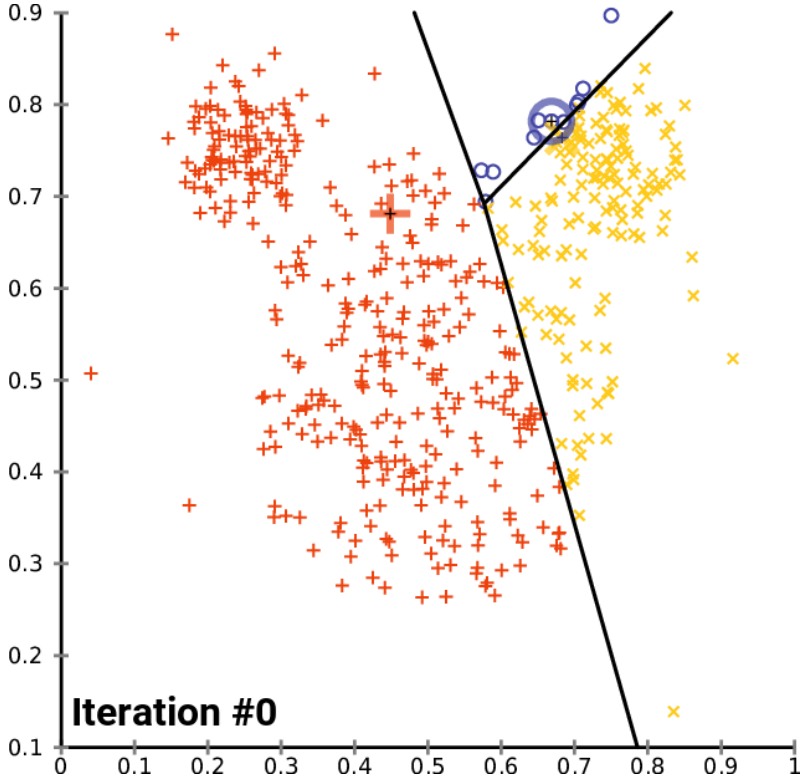
1. Contrast between unsupervised and supervised learning

2. Give examples of methods that integrate multiomics data and what each is trying to acheive

3. Apply a few unsupervised integrative methods to real world data

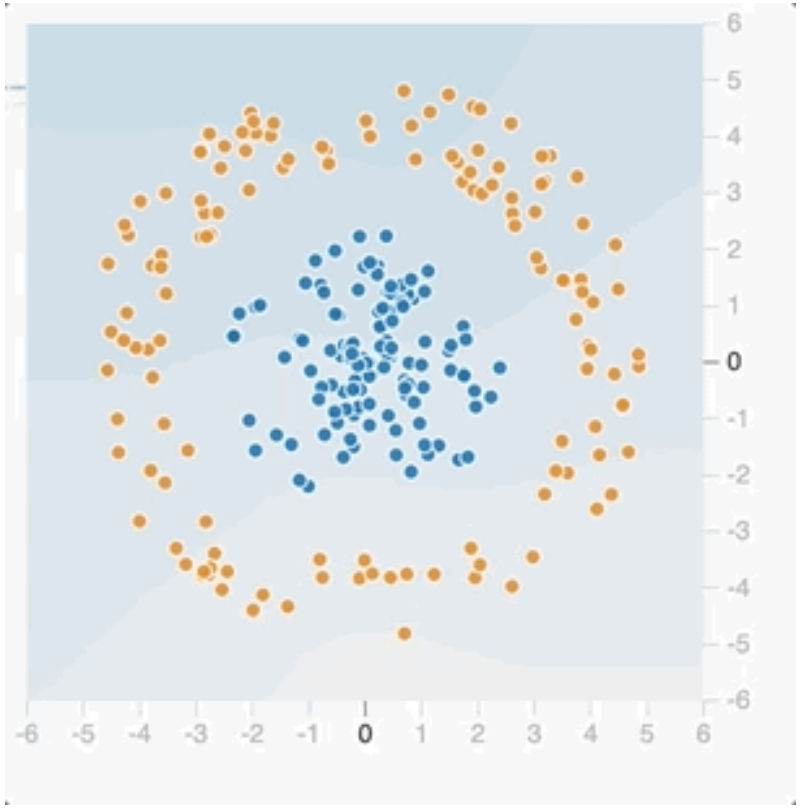
"In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures." Elements of statistical learning, 2008 page 3

Unsupervised (clustering)



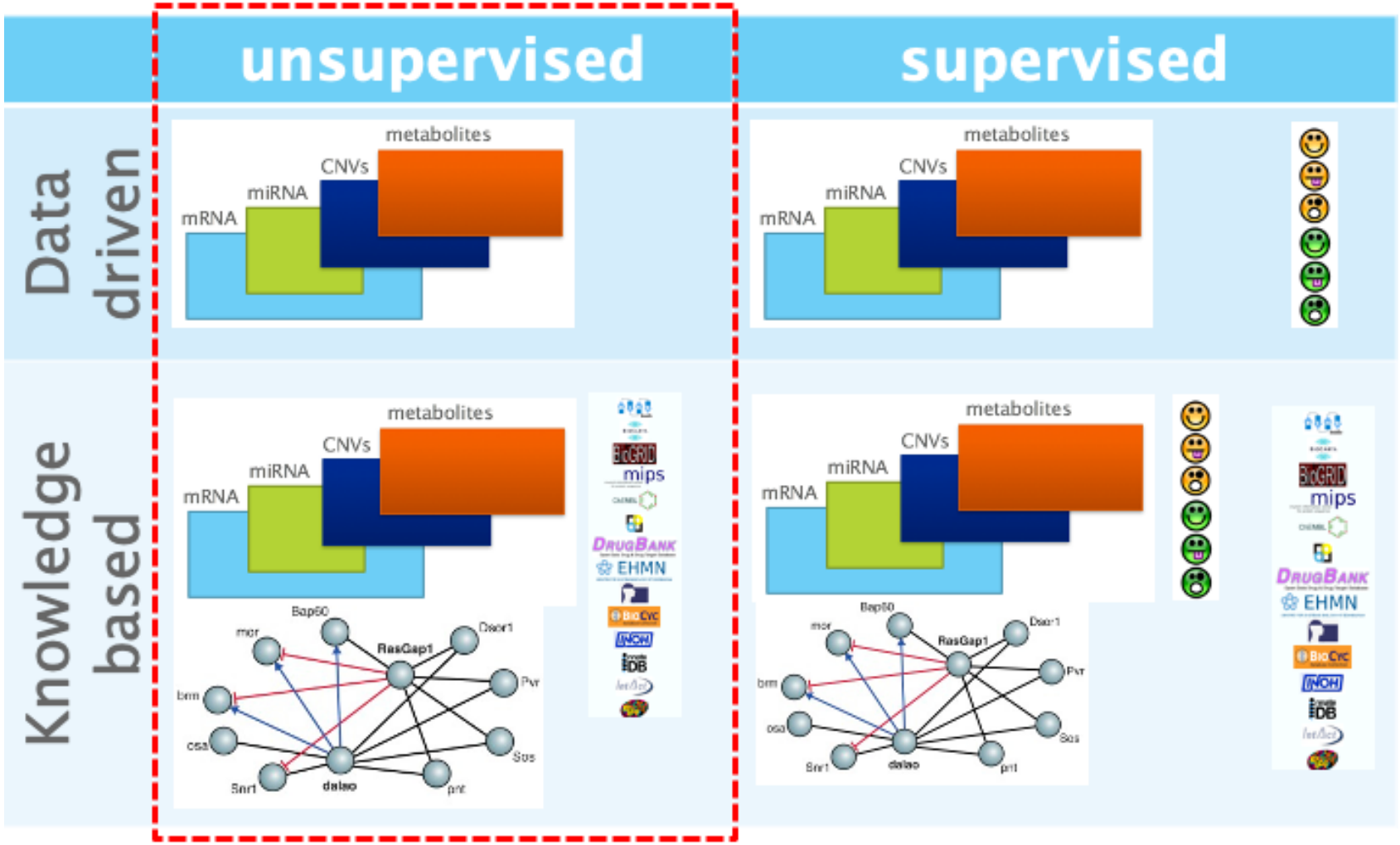
Chire 2017

Supervised (regression/classification)



Tensorflow playground

Unsupervised multiomics



Terminology review

Term	Synonym	Description
multiomics	multiview, multidomain, mulitmodal	confirmation of findings using the same samples, (e.g. measuring the same protein using different technologies in the same set of samples)
validation		confirmation of findings using samples from the original discovery cohort
dataset	block, table, data matrix	a 2D matrix comprises of observations (n) and variables (p)
latent variable	component, embedding, hidden variable	low-dimensional variable that contains information about the high dimensional dataset

REVIEW article

Front. Genet., 28 August 2019
Sec. Statistical Genetics and Methodology
https://doi.org/10.3389/fgene.2019.00627

This article is part of the Research Topic

Statistical and Computational Methods for Microbiome Multi-Omics Data

View all 11 Articles >

Multitable Methods for Microbiome Data Integration

Kris Sankaran[1*] and Susan P. Holmes[2]

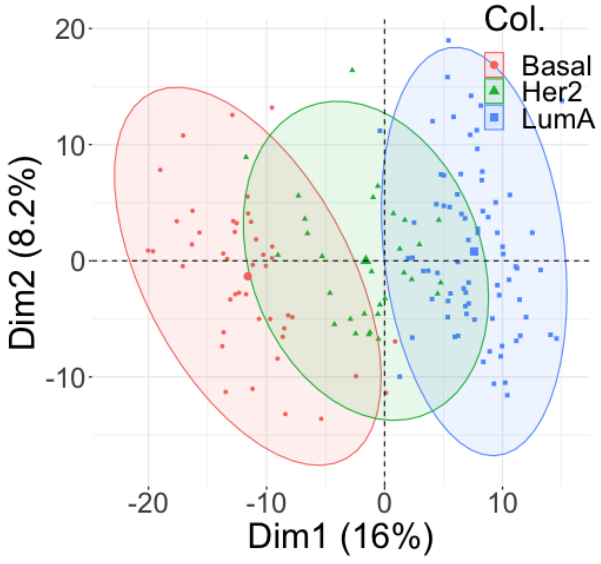
[1] Mila, Universite de Montréal, Montréal, QC, Canada
[2] Department of Statistics, Stanford University, Stanford, CA, United States

Sankaran K and Holmes SP. Front Genet. 2019 Aug 28;10:627

Concatenated PCA

$$X = [X^1 | \ldots | X^J]$$

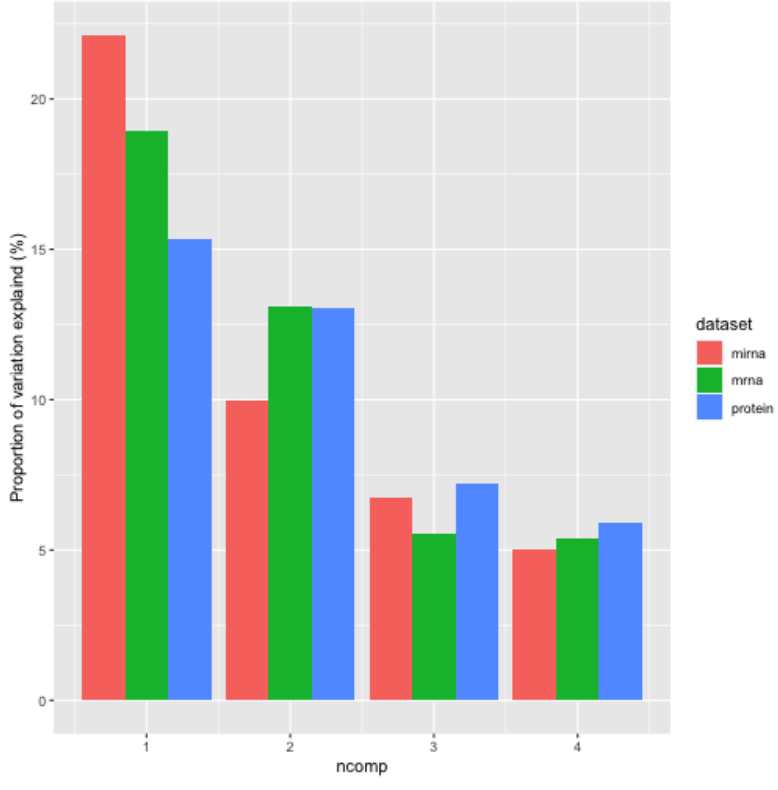
```
pca_res <- breast.TCGA$data.train[1:3] %>%
  do.call(cbind, .) %>%
  prcomp(x = ., rank. = 2, center = TRUE, scale. = TRUE)
```



Limitation of Concatenated PCA

- provides relationship between variables but not sets of variables
- datasets with more variables can dominate the resulting PCs; the mirna, mrna, protein datasets have 184, 200, 142 variables respectively.

```
p <- lapply(breast.TCGA$data.train[1:3], function(X){
  result <- prcomp(x = X, rank. = 4, center = TRUE, sc
  propvar <- 100*(result$sdev^2)/sum(result$sdev^2)
  propvar[1:4]
}) %>%
  do.call(cbind, .) %>%
  as.data.frame() %>%
  mutate(ncomp = 1:4) %>%
  gather(dataset, propvar, -ncomp) %>%
  ggplot(aes(x = ncomp, y = propvar, fill = dataset))
  geom_bar(stat="identity", position=position_dodge())
  ylab("Proportion of variation explaind (%)")
```



Weighted PCA (Multiple Factor Analysis, MFA)

$$X = \left[\frac{X^1}{\lambda_1(X^1)} | \dots | \frac{X^J}{\lambda_1(X^J)} \right]$$

```
res <- FactoMineR::MFA(as.data.frame(do.call(cbind, breast.TCGA$data.train[1:3])),
           group=sapply(breast.TCGA$data.train[1:3], ncol), type=rep("s",3),
    ncp=5, name.group=names(breast.TCGA$data.train[1:3]), graph=FALSE)
```

Guide to MFA interpretation

MFA

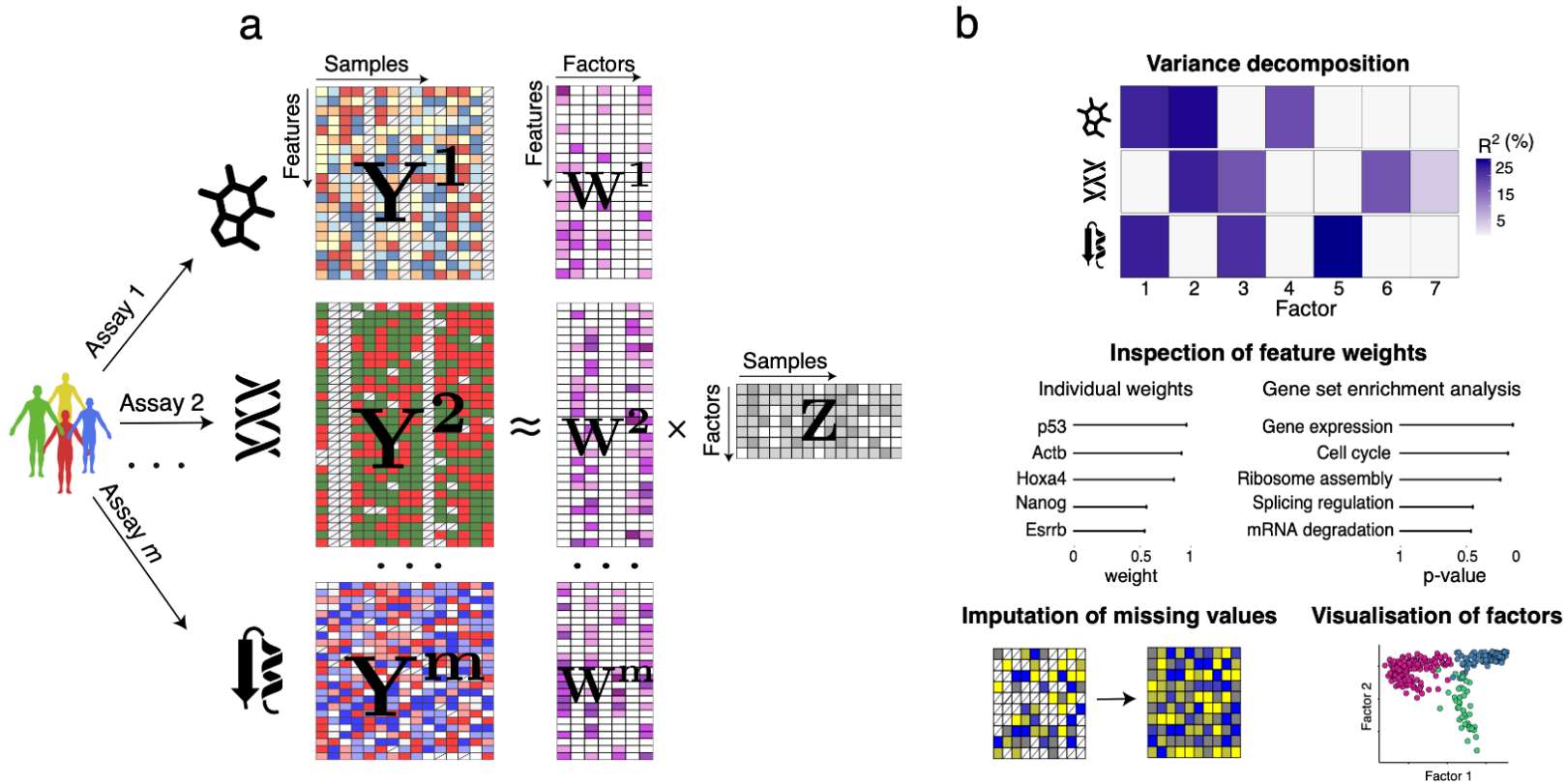
```
plot(res,choix="var", select = "cos2 0.6")
```

```
plot(res,choix="ind",partial="all")
```

```
## Warning: ggrepel: 120 unlabeled data points (too many overlap
## increasing max.overlaps
```

Multi-Omics Factor Analysis (MOFA)

- PCA for multiple datasets; feature selection for associated latent variables

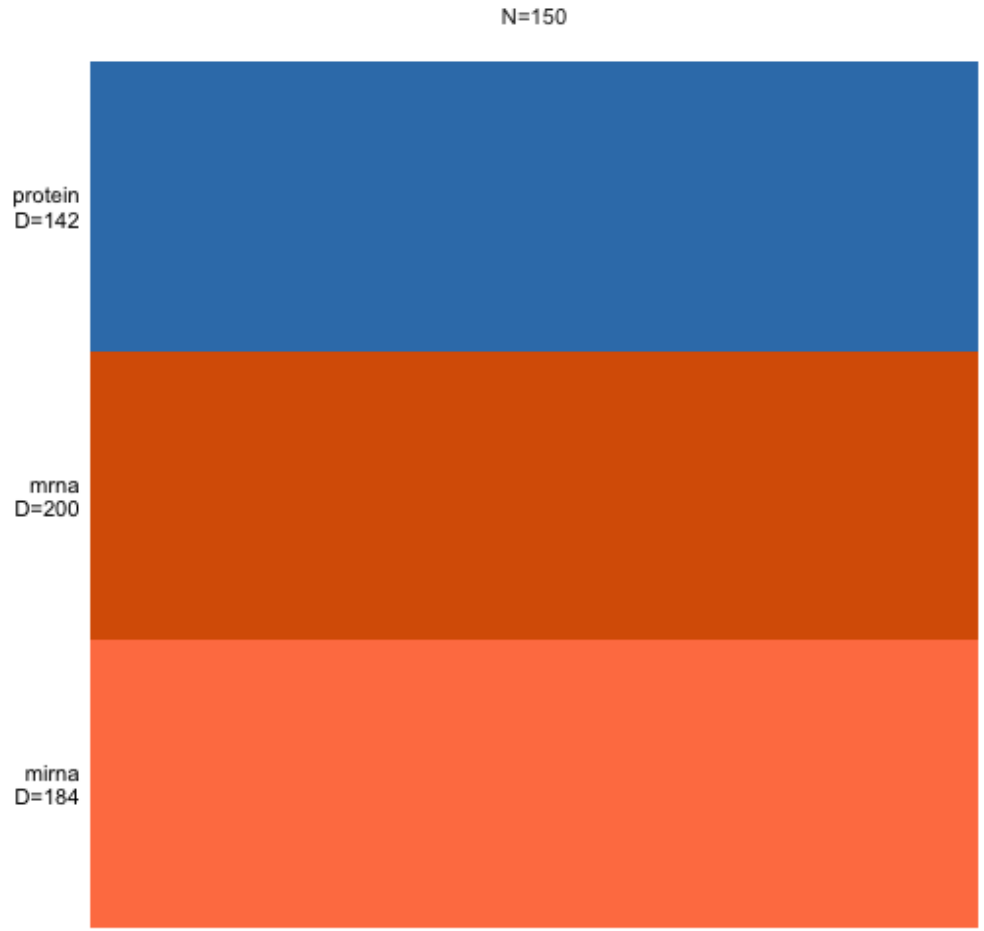


MOFA Genome Biology volume 21, Article number: 111 (2020)

MOFA

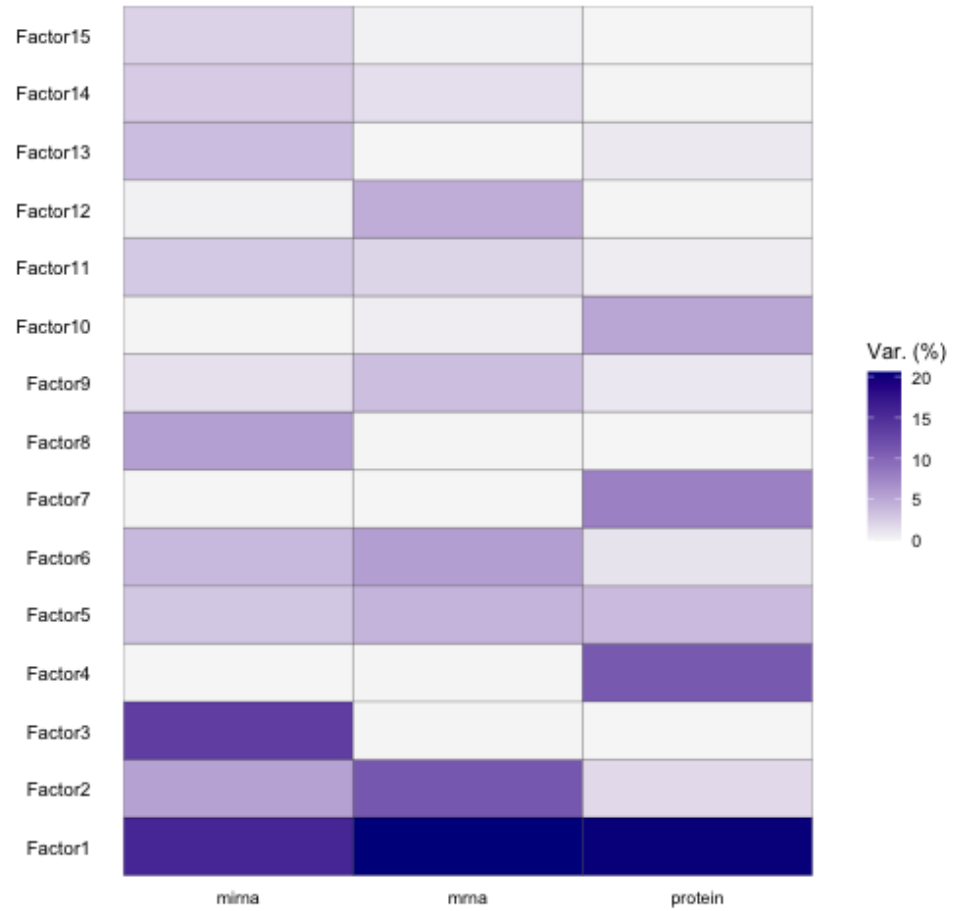
```
library(MOFA2)

MOFAobject <- create_mofa(lapply(breast.TCGA$data.trai

data_opts <- get_default_data_options(MOFAobject)
model_opts <- get_default_model_options(MOFAobject)
train_opts <- get_default_training_options(MOFAobject)

MOFAobject <- prepare_mofa(
  object = MOFAobject,
  data_options = data_opts,
  model_options = model_opts,
  training_options = train_opts
)
```

```
p <- plot_data_overview(MOFAobject)
```



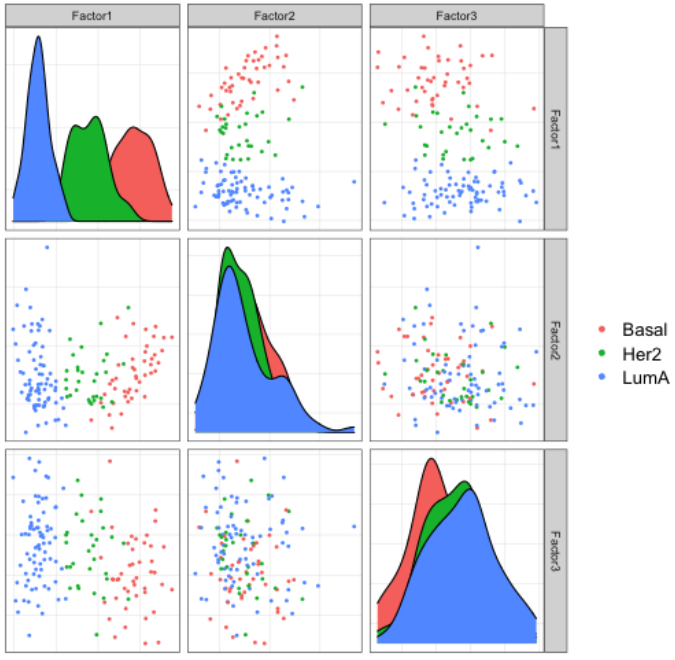
MOFA

```
outfile = file.path(getwd(),"model.hdf5")
model <- run_mofa(MOFAobject, outfile)
samples_metadata(model) <- data.frame(
  sample = samples_names(model)[[1]],
  subtype = breast.TCGA$data.train$subtype
)

p <- plot_variance_explained(model)
```

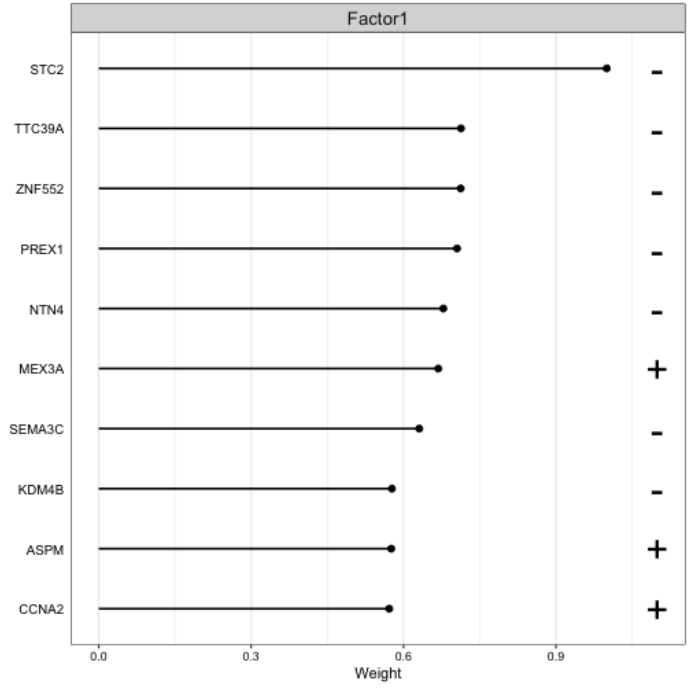


MOFA

```
plot_factors(model,
  factors = 1:3,
  color_by = "subtype"
)
```



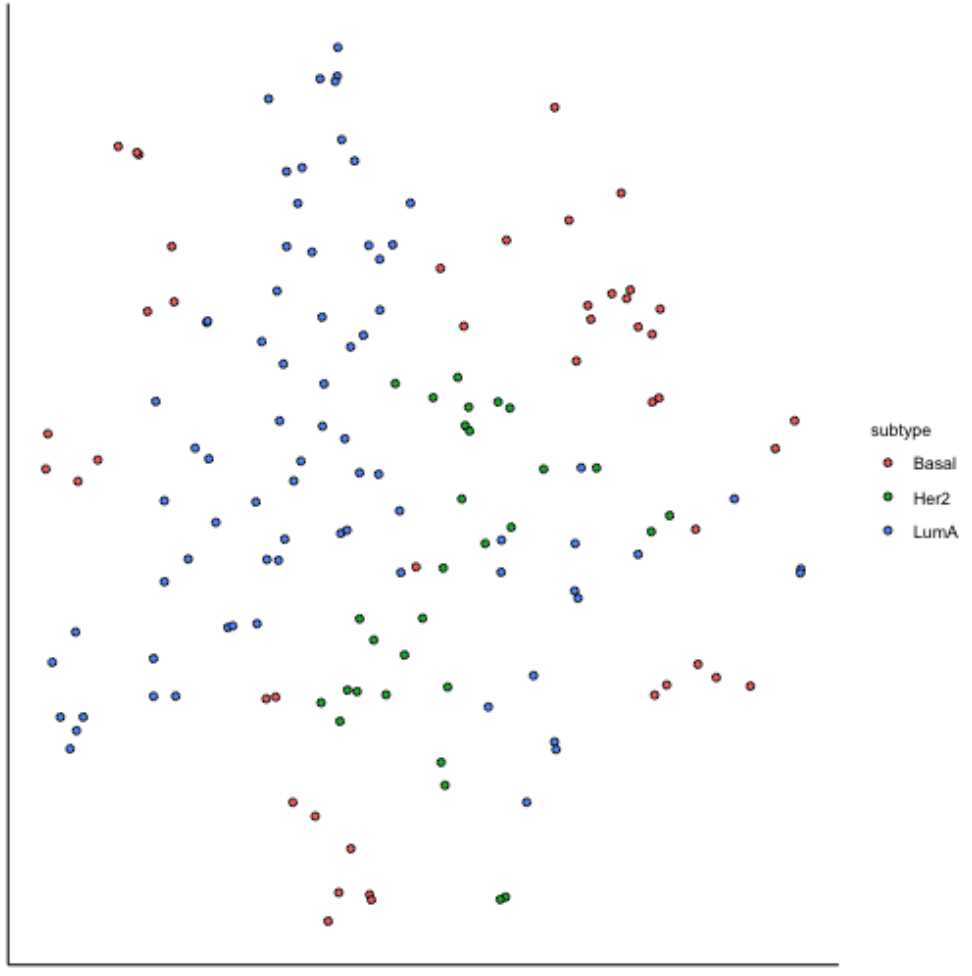

```
plot_top_weights(model,
  view = "mrna",
  factors = 1,
  nfeatures = 10
)
```



MOFA + tSNE

```
set.seed(42)
# model <- run_umap(model)
model <- run_tsne(model)

p <- plot_dimred(model,
  method = "TSNE",  # method can be either "TSNE" or "
  color_by = "subtype"
)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please us
## "none")` instead.
```



Partial Least Squares (PLS)

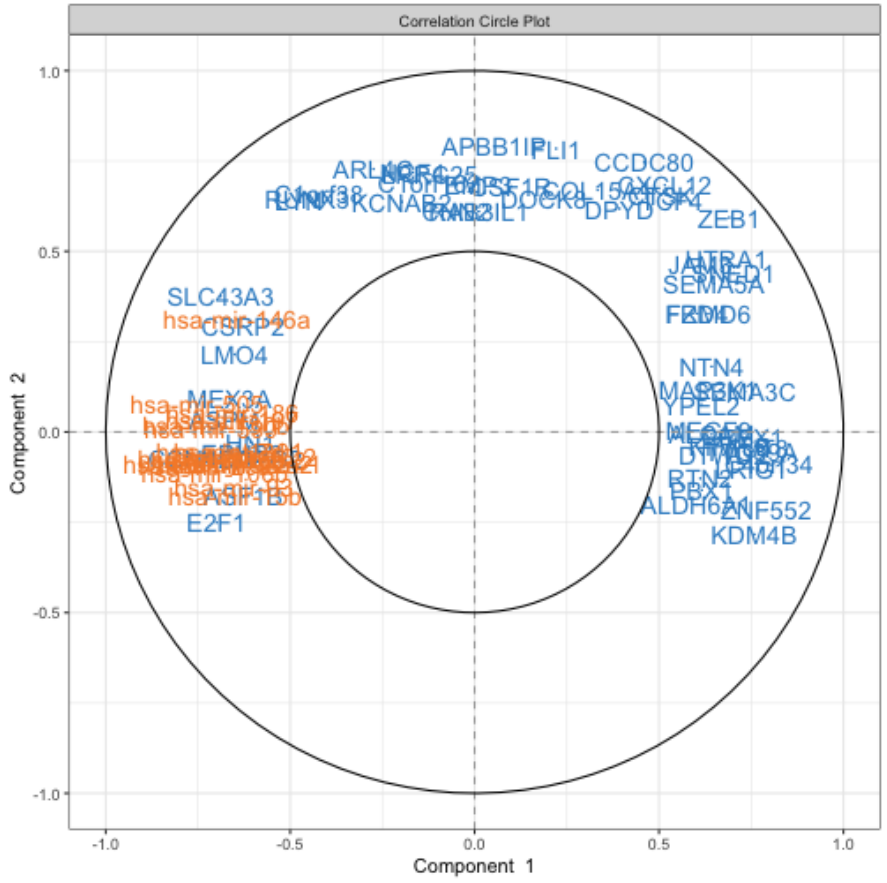
- unlike PCA which maximizes variance, PLS maximize covariance between two data matrices
- doesnt work with more than 2 data matrices

```
cim(cor(pls_res$variates$X, pls_res$variates$Y), margi
    xlab = "mRNA", ylab = "miRNA")
```

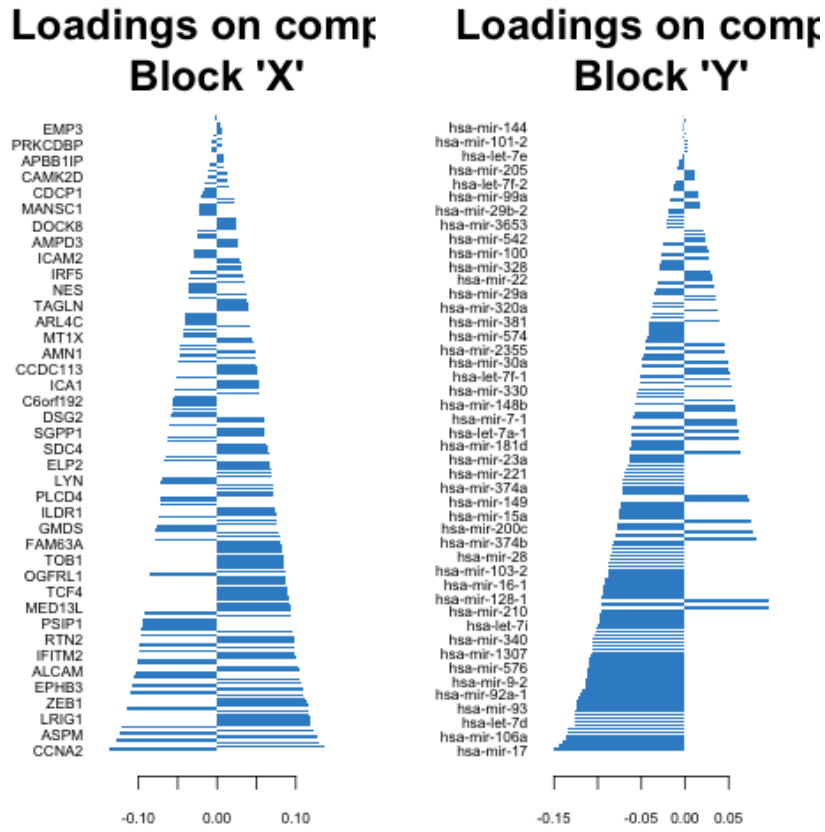
```
plotIndiv(pls_res, rep.space = "XY-variate", group = b
          ellipse = TRUE)
```

PLS

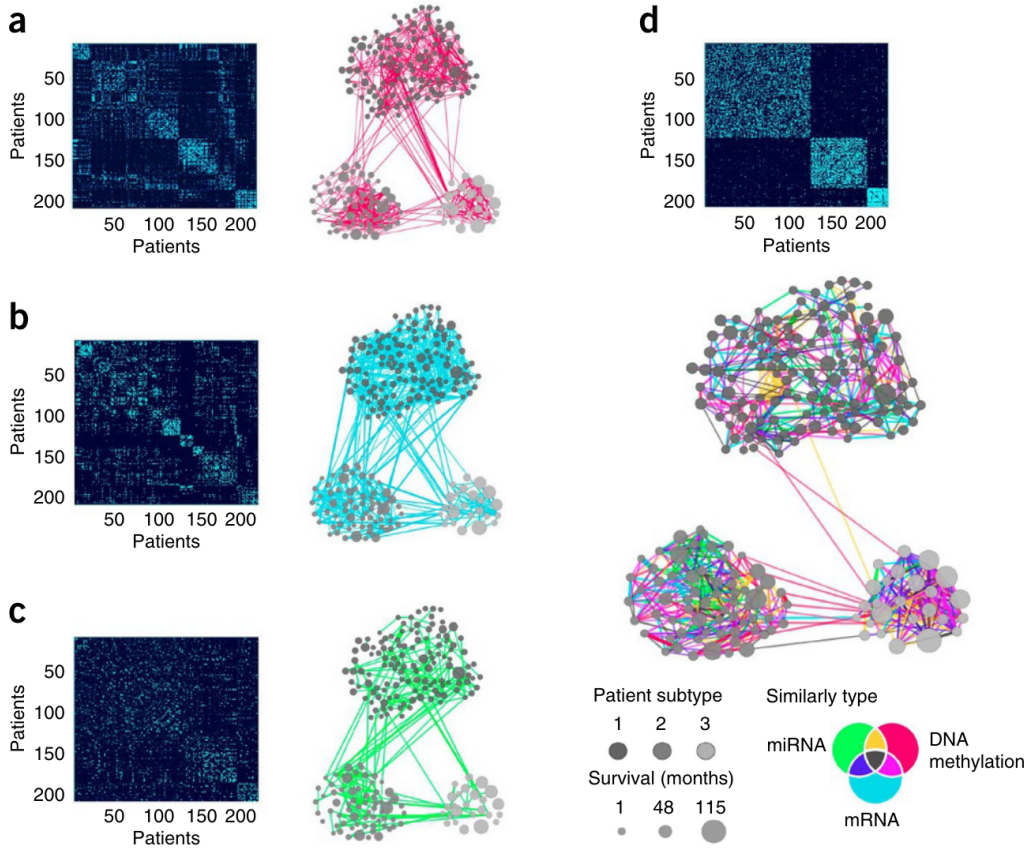
```
plotVar(pls_res, cutoff = 0.6)
```




```
plotLoadings(pls_res)
```



Similarity Network Fusion (SNF)

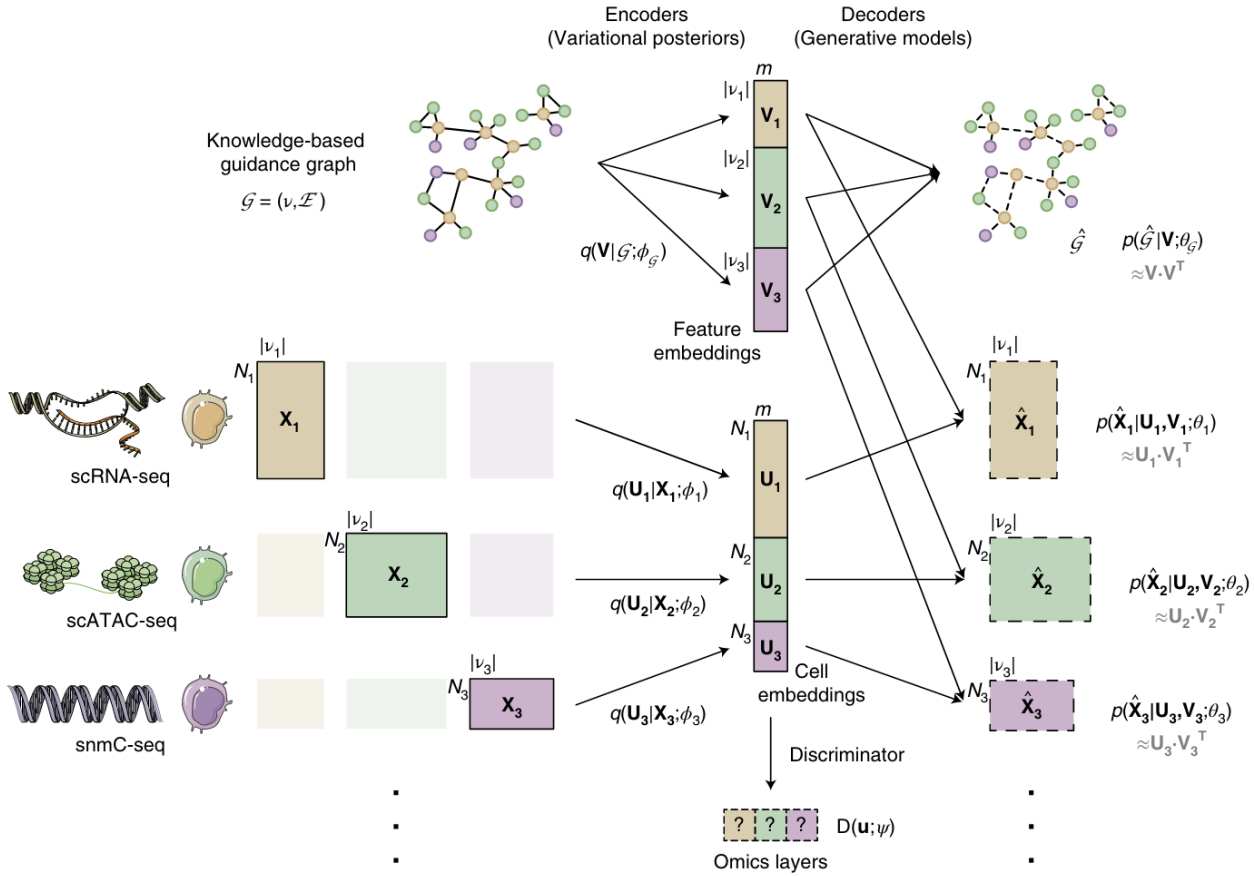


?SNF

```
library(SNFtool)

## First, set all the parameters:
K = 20;         # number of neighbors, usually (10~30)
alpha = 0.5;       # hyperparameter, usually (0.3~0.8)
T = 20;     # Number of Iterations, usually (10~20)

std_norm <- breast.TCGA$data.train[1:3]
std_norm$mrna <- standardNormalization(std_norm$mrna)
std_norm$mirna <- standardNormalization(std_norm$mirna

W <- lapply(std_norm, function(i){
  # distance matrix
  dist_matrix <- (dist2(as.matrix(i),as.matrix(i)))^(1
  # affinity matrix calculation
  affinityMatrix(dist_matrix, K, alpha)
})

W = SNF(W, K, T)


C = 3
labels = spectralClustering(W, C)
```

```
displayClusters(W, labels)
```

Unpaired multiomics data integration (GLUE)



Cao Z and Gao G. Nat Biotechnol. 2022 May 2. | GLUE code

timeOmics

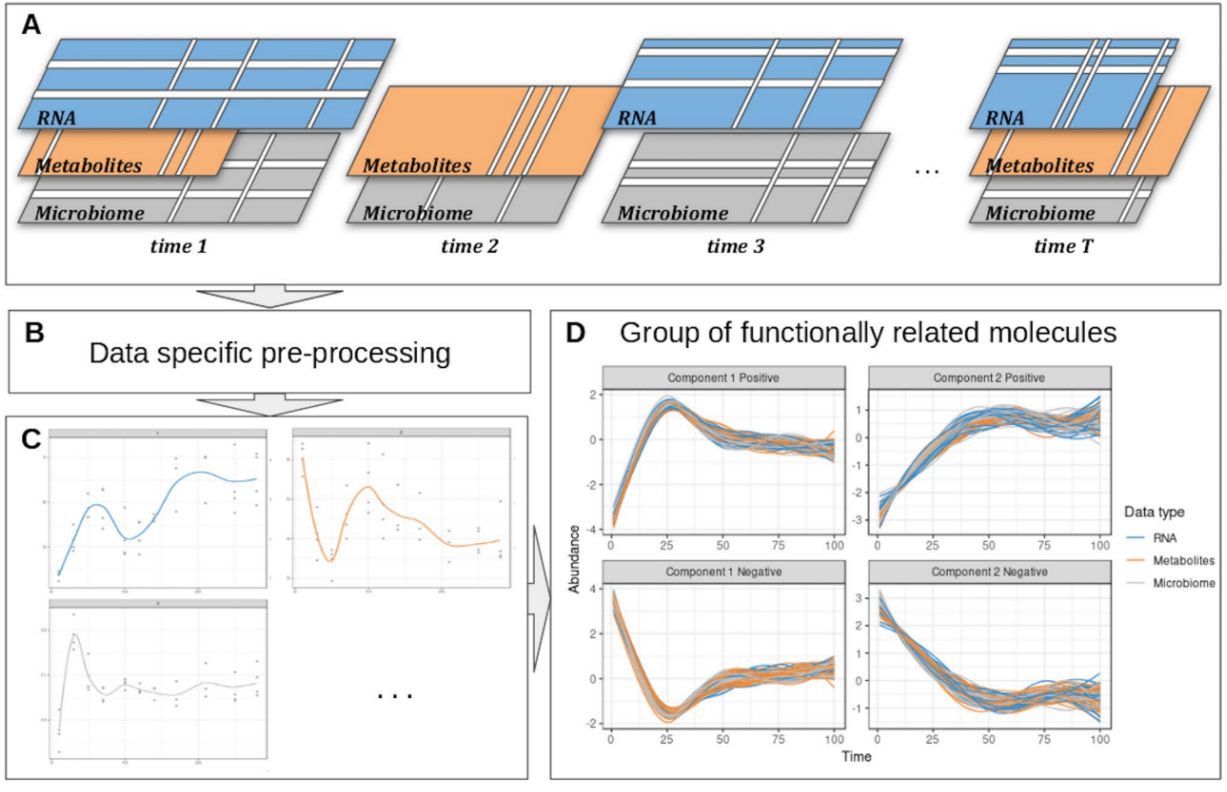
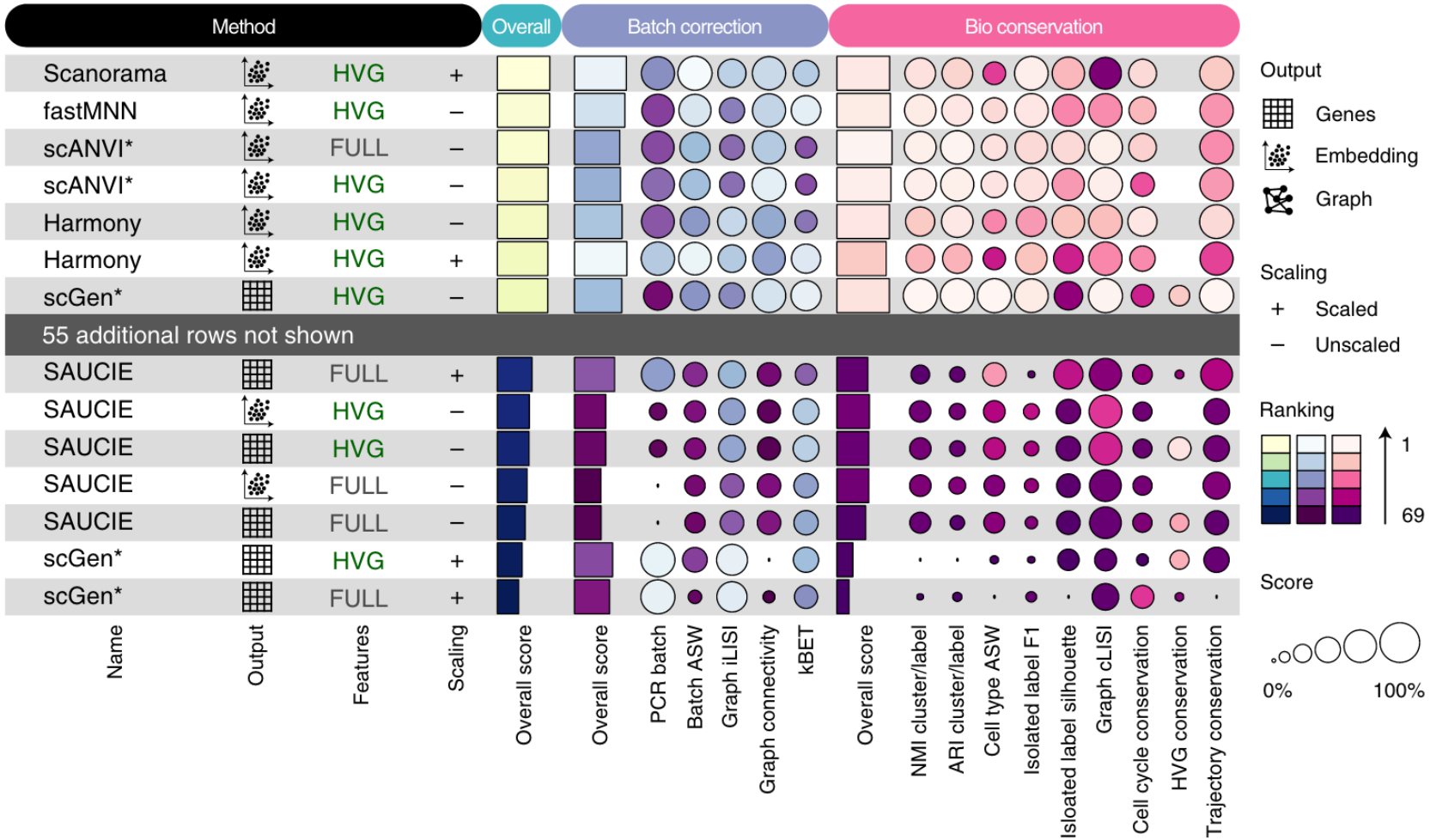


FIGURE 1 | Overview of the proposed approach. **(A)** Description of the experimental design: the same biological material is sampled at several time points across several omic layers indicated in different colors. Blank lines indicate potential missing values per time point or per feature in a given time point. **(B)** Specific pre-processing and normalization are applied according to the type of data. **(C)** Each molecule is modeled as a function of time by taking into account all the variabilities of the different biological replicates in a linear mixed model spline framework. **(D)** The modeled trajectories across all omics layers are clustered using a multivariate integrative method.

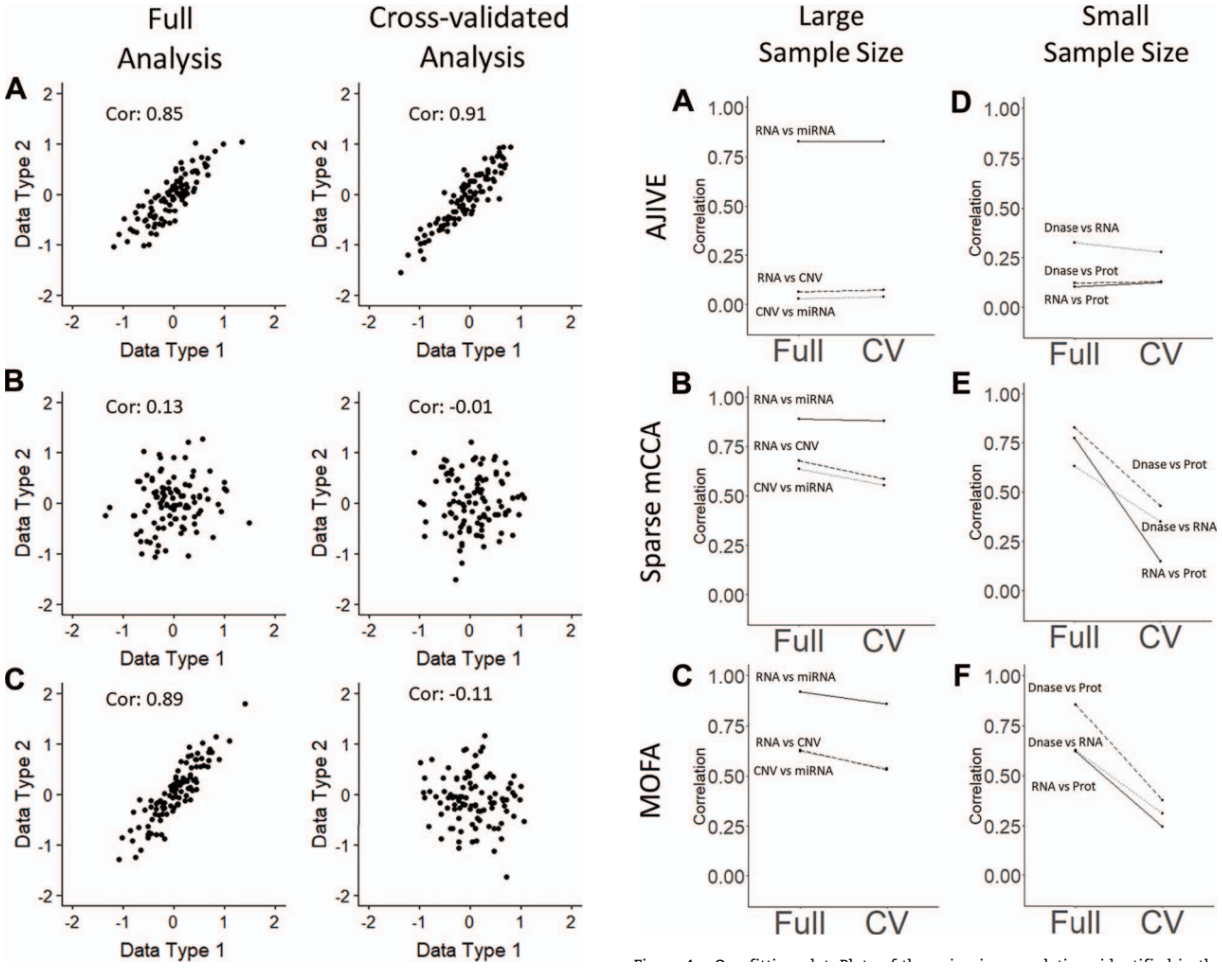
Bodein et al., Front Genet. 2019 Nov 7;10:963.

Single cell data integration



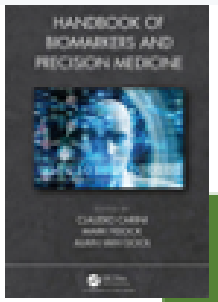
Nat Methods. 2022 Jan;19(1):41-50. | Explanation of metrics

Assessing overfitting and consistency



mikelove / awesome-multi-omics Public Watch 45 Fork 123 Star 502

seandavi / awesome-single-cell Public Watch 238 Fork 773 Star 2.2k



Unsupervised

Supervised

Data-driven

Knowledge-based

MCIA: omicade4
sMB-PLS[a]
SGCCA: RGCCA/mixOmics

Multiblock data analyses
Extensions of generalized CCA

SMSMA: msma
SGCCA+DA: mixOmics

SNF: SNFtool

Message-passing algorithms

PANDA: pandaR

Joint NMF[a]

Factorization methods

BCC: bayesCC

iBAG[b]

Bayesian Methods

Bayesian Networks[c]

Concatenation: caret
Ensemble: caretEnsemble

Classification and regression algorithms

SNMNMF[a]

Network-constrained

stSVM: netclass
GELnet: gelnet

Singh A et al., Handbook of Biomarkers and Precision Medicine CRC Press 2019:596

PRECISION HEALTH
ANALYSIS BOOTCAMP

THANK YOU!

August 08, 2022, 2022 | 09:00-11:00

lab
code
asingh_22g