

Multi-omic data analysis

Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology
Barcelona Institute for Global Health (ISGlobal)
e-mail: juanr.gonzalez@isglobal.org
<http://www.creal.cat/brge>

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets
- 4 Take home messages
- 5 Recommended lectures

Outline

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets
- 4 Take home messages
- 5 Recommended lectures

Introduction

- Understanding the genetic basis of complex traits is an open question for many researchers
- Recent advances in technology have made this problem even more complex by incorporating other pieces of information such as neuroimaging, daily measurements of environmental exposures among others
- The main challenge is to analyze the vast amount of data that is being generated, particularly how to integrate information from different tables

Introduction

- Understanding the genetic basis of complex traits is an open question for many researchers
- Recent advances in technology have made this problem even more complex by incorporating other pieces of information such as neuroimaging, daily measurements of environmental exposures among others
- The main challenge is to analyze the vast amount of data that is being generated, particularly how to integrate information from different tables

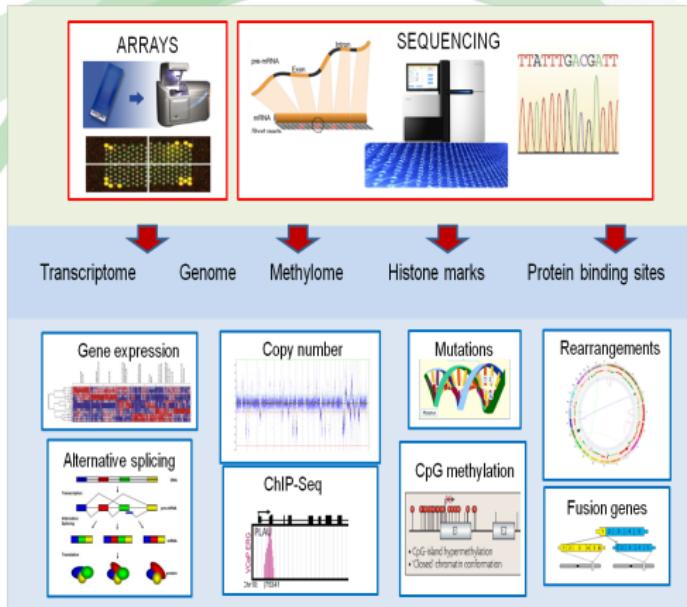
Introduction

- Understanding the genetic basis of complex traits is an open question for many researchers
- Recent advances in technology have made this problem even more complex by incorporating other pieces of information such as neuroimaging, daily measurements of environmental exposures among others
- The main challenge is to analyze the vast amount of data that is being generated, particularly how to integrate information from different tables

Introduction

One omic dataset: dimensionality reduction
Integrating two or more omic datasets
Take home messages
Recommended lectures

Introduction

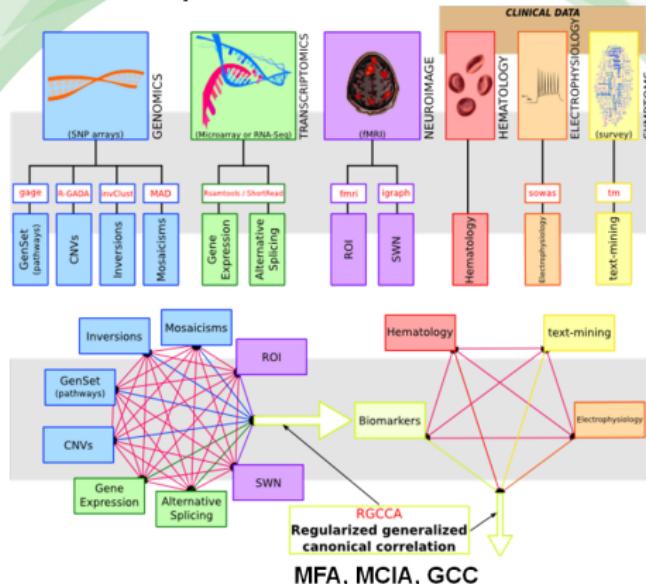


Introduction

One omic dataset: dimensionality reduction
Integrating two or more omic datasets
Take home messages
Recommended lectures

Introduction

Supervised



Introduction

- Data integration (Integrative bioinformatics, integrated analysis, crossomics, multi-dataset analysis, data fusion, ...) is being crucial in Bioinformatics/Biology/Epidemiology
- Data integration may refer to different aspects
 - Computational combination of data (sets)
 - Simultaneous analysis of different variables from different tables, different time points, different tissues, ...
 - Provide biological insights by using information from existing databases (ENCODE, GTEx, KEGG, ...)
- Here we mean the process by which different types of omic data are combined as predictor variables to allow more thorough and comprehensive modelling of complex traits or phenotypes

Introduction

- **Multi-staged analysis:** Involves integrating information using a stepwise or hierarchical analysis approach
- **Meta-dimensional analysis:** Refers to the concept of integrating multiple different data types to build a multivariate model associated with a given outcome

Multi-staged analysis

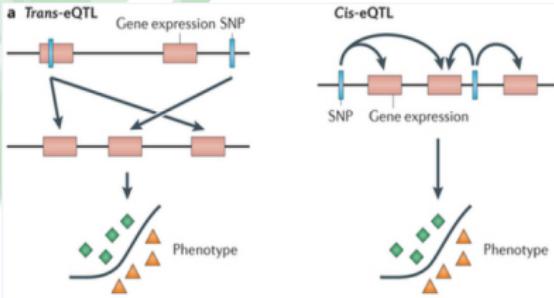
- **Genomic variation analysis approaches:** Triangle approach
 - ① SNPs are associated with the phenotype of interest and filtered based on p-values
 - ② Significant SNPs are tested for association with another level of omic data (i.e eQTLs)
 - ③ Omic data used in step 2 are then tested for correlation with phenotype of interest
- **Domain knowledge-guided approaches:** Integration of functional and pathway information
 - ① Significant results (regions of interest) are annotated in databases such as ENCODE, Roadmap Epigenomic Project, KEGG, ...
 - ② We can determine whether those regions are within pathways and/or overlapping with functional units, such as transcription factors binding, hyper- or hypo-methylated regions
 - ③ Overlapping regions are then correlated with the phenotype of interest
 - ④ The main limitation is that this method is biased by current knowledge

Multi-staged analysis

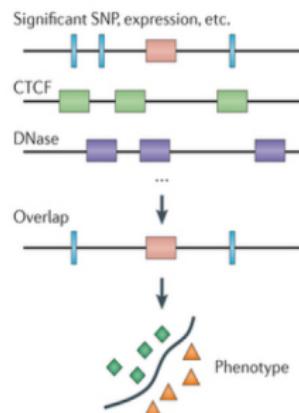
- **Genomic variation analysis approaches:** Triangle approach
 - ① SNPs are associated with the phenotype of interest and filtered based on p-values
 - ② Significant SNPs are tested for association with another level of omic data (i.e eQTLs)
 - ③ Omic data used in step 2 are then tested for correlation with phenotype of interest
- **Domain knowledge-guided approaches:** Integration of functional and pathway information
 - ① Significant results (regions of interest) are annotated in databases such as ENCODE, Roadmap Epigenomic Project, KEGG, ...
 - ② We determine whether those regions are within pathways and/or overlapping with functional units, such as transcription factors binding, hyper- or hypo-methylated regions
 - ③ Overlapping regions are then correlated with the phenotype of interest
 - ④ The main limitation is that this method is biased by current knowledge

Types of multi-staged analyses

(a) Genomic variation analysis



(b) Domain knowledge-guided



Introduction

- One omic dataset: dimensionality reduction
- Integrating two or more omic datasets
- Take home messages
- Recommended lectures

Genomic variation example

ARTICLES

nature
biotechnology

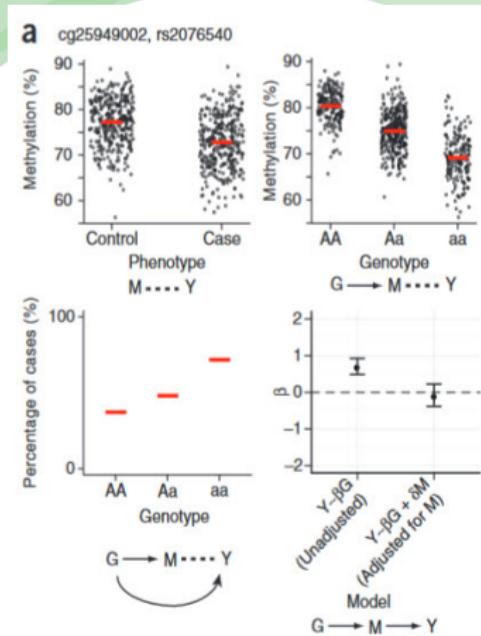
Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis

Yun Liu^{1,2,12}, Martin J Aryee^{1,3,12}, Leonid Padyukov^{4,5,12}, M Daniele Fallin^{1,6,7,12}, Espen Hesselberg^{4,5},
Arni Runarsson^{1,2}, Lovisa Reinarius⁸, Nathalie Acevedo⁹, Margaret Taub^{1,6}, Marcus Ronninger^{4,5},
Klementy Shchetyl'sky^{4,5}, Annika Scheynius⁹, Juha Kere⁸, Lars Alfredsson¹⁰, Lars Klareskog^{4,5},
Tomas J Ekström^{5,11} & Andrew P Feinberg^{1,2,6}

Introduction

- One omic dataset: dimensionality reduction
- Integrating two or more omic datasets
- Take home messages
- Recommended lectures

Genomic variation example



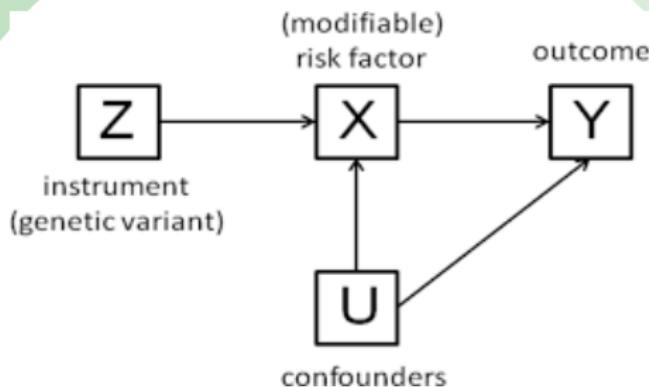
Genomic variation example

- Mendelian randomization
- Conditional Independent test (CIT)

MR: https://cran.r-project.org/web/packages/MendelianRandomization/vignettes/Vignette_MR.pdf

CIT: <https://cran.r-project.org/web/packages/cit/>

Mendelian randomization



Conditional Independent Test (CIT)

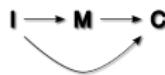
A) Independencia / Asociación



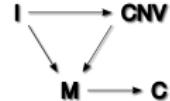
B) Causal



C) Causal / Asociación



D) Complejo

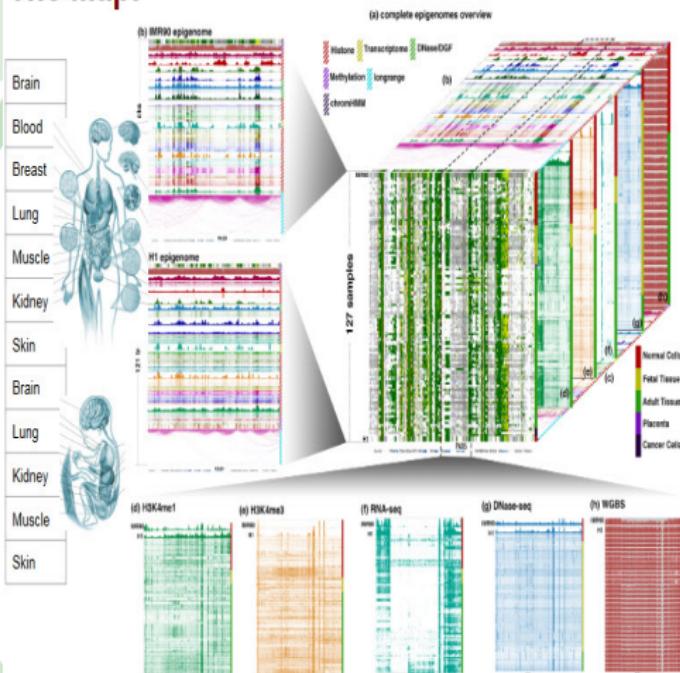


Introduction

One omic dataset: dimensionality reduction
Integrating two or more omic datasets
Take home messages
Recommended lectures

Domain knowledge-guided

The Map!



The RoadMap Epigenomics

- Marking of functional genomic elements
- Environmental exposures
- Understanding development and differentiation
- Regenerative medicine (stem and iPS cells)
- Human disease
- Interpreting GWAS
- Biomarkers, diagnostics and therapies
- Exploring cross-talk between epigenomic mechanisms

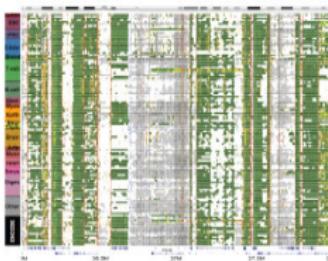
Domain knowledge-guided

Using Epigenomic Data to Interpret GWAS Data

Gene variants in
human disease



Epigenomic data for many
normal human cell/tissue types



- 77% of disease variants are in/near enhancer elements or promoters*
- Variants are in regulatory regions NOT protein coding regions
- Generate hypotheses about function

Introduction

One omic dataset: dimensionality reduction

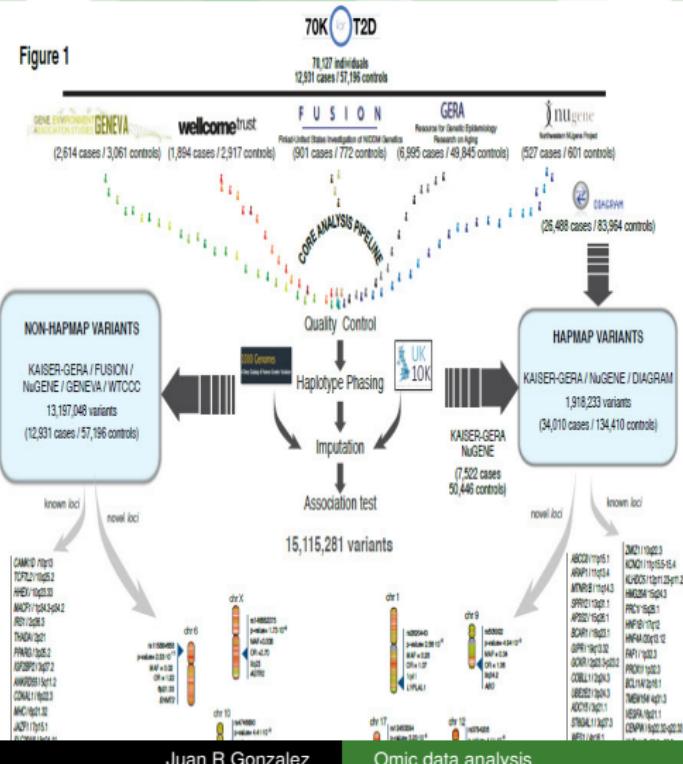
Integrating two or more omic datasets

Take home messages

Recommended lectures

Domain knowledge-guided

Figure 1



Introduction

One omic dataset: dimensionality reduction

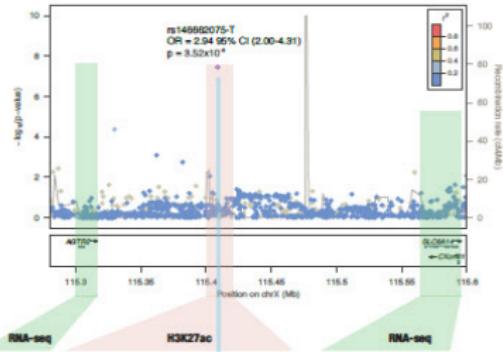
Integrating two or more omic datasets

Take home messages

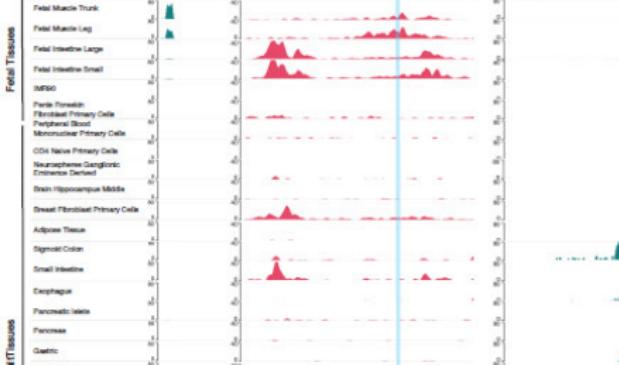
Recommended lectures

Domain knowledge-guided

(A)



Fetal Tissues

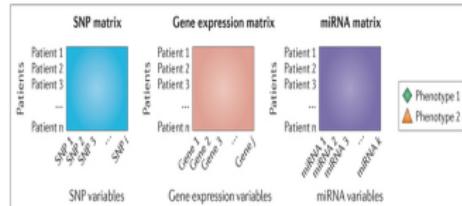
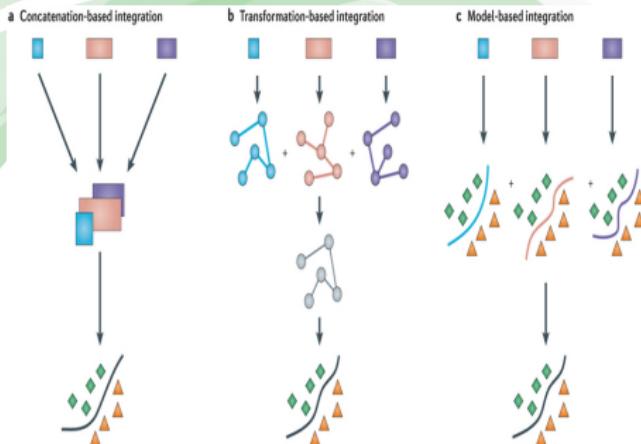


Adult Tissues

Introduction

One omic dataset: dimensionality reduction
Integrating two or more omic datasets
Take home messages
Recommended lectures

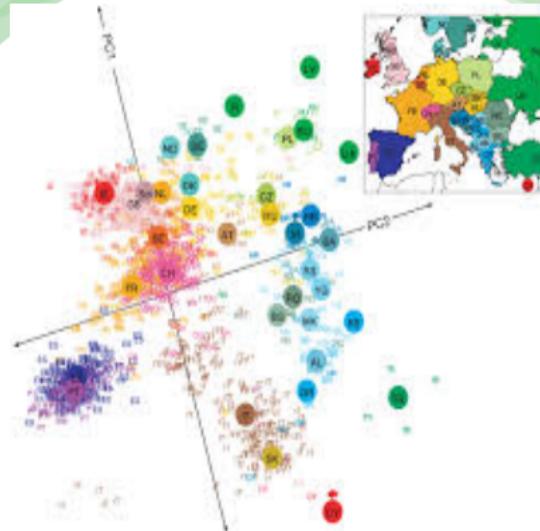
Types of meta-dimensional analyses



Outline

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets
- 4 Take home messages
- 5 Recommended lectures

One omic dataset: dimensionality reduction



One omic dataset: dimensionality reduction

- Used to study tissue, cell attributes or cancer signatures with regard to abundance of mRNAs, proteins and metabolites
- Given an omic data set X , which is a $n \times p$ matrix, of n individuals and p features

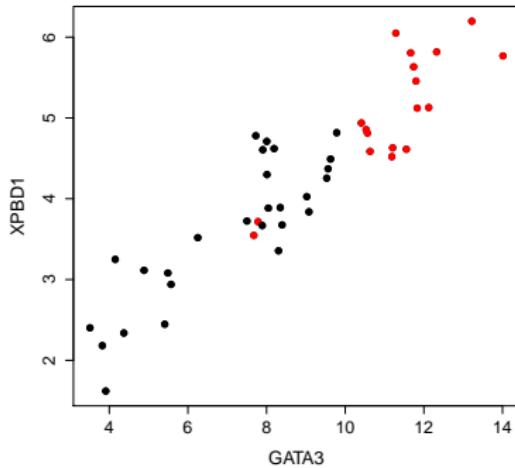
$$X = (x_1, x_2, \dots, x_p)$$

- we look for new variables that are linear combinations of the original variables $f = q_1 X_1 + q_2 X_2 + \dots + q_p X_p$ or $f = Xq$ where q are known as loadings.
- We introduce the restriction that for i th component, q 's should maximize the variance components of f 's

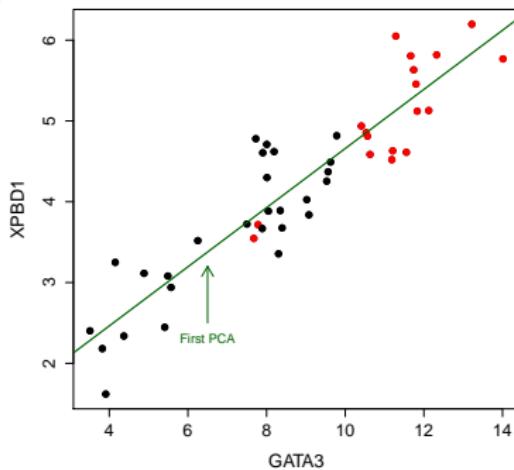
$$\arg \max_{q^i} \text{var}(Xq^i)$$

and q 's have to be orthogonal to each other.

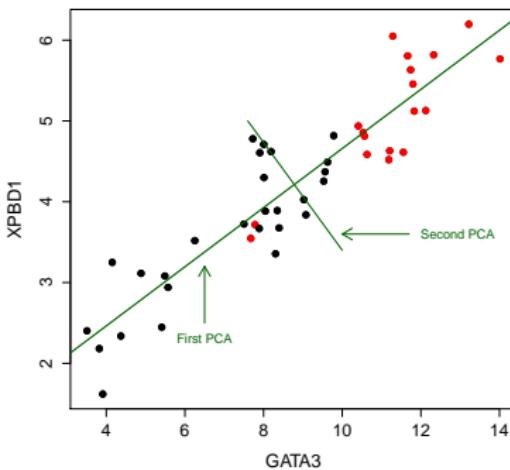
One omic dataset: dimensionality reduction



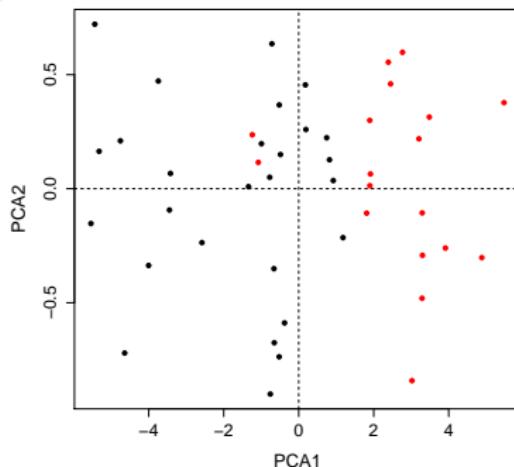
One omic dataset: dimensionality reduction



One omic dataset: dimensionality reduction

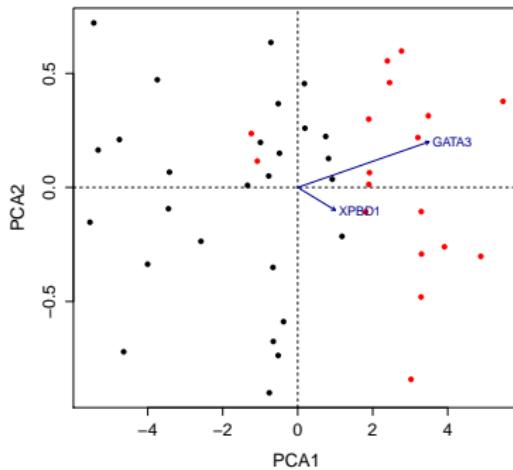


One omic dataset: dimensionality reduction



One omic dataset: dimensionality reduction

Data visualization: biplot



One omic dataset: dimensionality reduction

- There are other techniques such as: Principal co-ordinate analysis (PCoA), correspondence analysis (CA) and nonsymmetrical correspondence analysis (NSCA).
- They can be applied to other types of data (e.g. non-continuous). PCoA can also be applied to binary or count data.
- PCA is designed to analyze multi-normal distributed data. If data are skewed, contain extreme outliers, or display nonlinear trends, other methods such as Multidimensional Scaling (MDS)
- Nonnegative matrix factorization (NMF) and Independent Component Analysis (ICA) are applied when orthogonality or independence across components are not hold.

One omic dataset: dimensionality reduction

- There are other techniques such as: Principal co-ordinate analysis (PCoA), correspondence analysis (CA) and nonsymmetrical correspondence analysis (NSCA).
- They can be applied to other types of data (e.g. non-continuous). PCoA can also be applied to binary or count data.
- PCA is designed to analyze multi-normal distributed data. If data are skewed, contain extreme outliers, or display nonlinear trends, other methods such as Multidimensional Scaling (MDS)
- Nonnegative matrix factorization (NMF) and Independent Component Analysis (ICA) are applied when orthogonality or independence across components are not hold.

One omic dataset: dimensionality reduction

- There are other techniques such as: Principal co-ordinate analysis (PCoA), correspondence analysis (CA) and nonsymmetrical correspondence analysis (NSCA).
- They can be applied to other types of data (e.g. non-continuous). PCoA can also be applied to binary or count data.
- PCA is designed to analyze multi-normal distributed data. If data are skewed, contain extreme outliers, or display nonlinear trends, other methods such as Multidimensional Scaling (MDS)
- Nonnegative matrix factorization (NMF) and Independent Component Analysis (ICA) are applied when orthogonality or independence across components are not hold.

One omic dataset: dimensionality reduction

- There are other techniques such as: Principal co-ordinate analysis (PCoA), correspondence analysis (CA) and nonsymmetrical correspondence analysis (NSCA).
- They can be applied to other types of data (e.g. non-continuous). PCoA can also be applied to binary or count data.
- PCA is designed to analyze multi-normal distributed data. If data are skewed, contain extreme outliers, or display nonlinear trends, other methods such as Multidimensional Scaling (MDS)
- Nonnegative matrix factorization (NMF) and Independent Component Analysis (ICA) are applied when orthogonality or independence across components are not hold.

One omic dataset: dimensionality reduction

- Solving the problem for the i -th component

$$\arg \max_{q^i} \text{var}(Xq^i)$$

uses SVD decomposition and it requires an inversion step that can be problematic when $p \gg n$

- Several extensions based on regularization step or L-1 penalization (Least Absolute Shrinkage and Selection Operator, LASSO) can be applied
- Sparse, penalized and regularized extensions of PCA and related methods have been proposed in omic data analysis, recently (sPCA)

Multi-omic real data

- Data from the Cancer Genome Atlas (TCGA) will be analyzed.
- A subset of the TCGA breast cancer study from Nature 2012 publication have been selected.
- Data
<https://tcga-data.nci.nih.gov/docs/publications/brcal>
- Available data are: miRNA, miRNAPrecursor, RNAseq, Methylation, proteins from a RPPA array, and GISTIC SNP calls (CNA and LOH). Clinical data are also available.

Multi-omic real data

Data can be loaded into R by executing:

```
load("data/breast_TCGA_subset_multi_omic.RData")
summary(breast_multi)

##                                     Length Class      Mode
## mirNA                  79  data.frame  list
## miRNAPrecursor        79  data.frame  list
## RNAseq                 791580 -none-    numeric
## Methyl                  79  data.frame  list
## RPPA                   79  data.frame  list
## LOH                    996348 -none-    numeric
## CNA                   934649 -none-    numeric
## clin                   29  data.frame  list
```

PCA analysis

```
require(ade4)
dim(breast_multi$RNAseq)

## [1] 10020      79

breastPCA<-dudi.pca(breast_multi$RNAseq,
                      scannf=FALSE, nf=5)
```

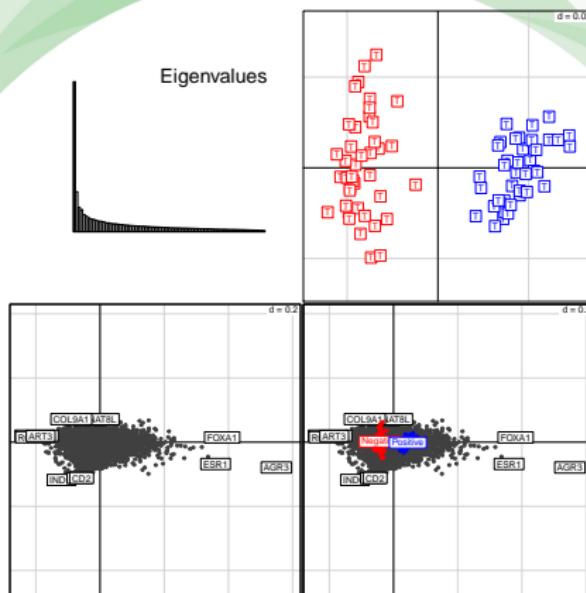
PCA analysis

There is a BioC package called made4 which is a wrapper around ade4 that utilizes Bioconductor data classes (such as Expression Set to help visualizing the results). The key function is ord. The panels shows plot of eigenvalues, projections of samples, projections of genes and biplot.

```
require(made4)
```

```
group<-droplevels(breast_multi$clin$ER.Status)
out <- ord(breast_multi$RNAseq, classvec=group)
plot(out, nlab=3, arraylabels=rep("T", 79))
```

PCA analysis

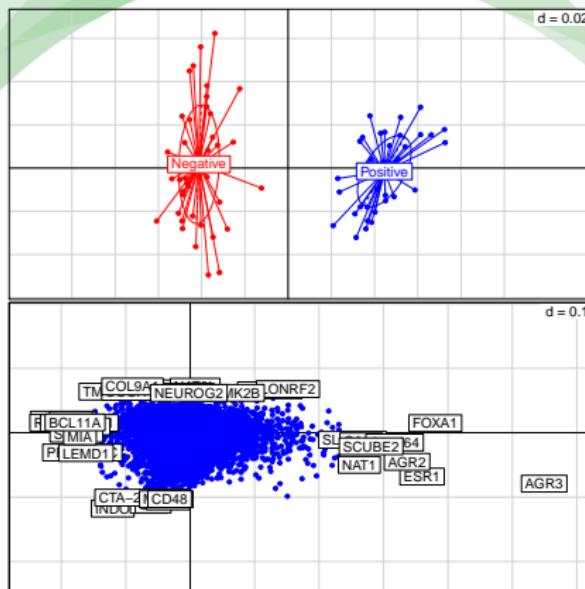


PCA analysis

Sample and genes projections can also be plotted using another wrapper around `ade4` package. By default the top-10 more important genes for each axis are highlighted.

```
par(mfrow=c(2, 1))
plotarrays(out$ord$co, classvec=group)
plotgenes(out, col="blue")
```

PCA analysis



PCA analysis

A list of variables with higher loadings on axes can be obtained using the following function

```
ax1 <- topgenes(out, axis=1, n=5)
ax2 <- topgenes(out, axis=2, n=5)
cbind(ax1, ax2)

##          ax1        ax2
## [1,] "ROPN1B"   "SH3GL2"
## [2,] "ROPN1"     "SLITRK6"
## [3,] "FOXA1"     "LONRF2"
## [4,] "SFRP1"     "IL7R"
## [5,] "ESR1"       "INDO"
## [6,] "AGR3"       "NAT8L"
## [7,] "AGR2"       "UBD"
## [8,] "Clorf64"   "CD2"
## [9,] "ART3"       "CD69"
## [10,] "BCL11A"    "COL9A1"
```

Sparse PCA analysis

Witten D, Tibshirani R, Hastie T (Biostatistics, 2009, 10(3):515-534) presented a penalized matrix decomposition for computing a rank-K approximation for a matrix based on L1-penalization. The method is also extended to Canonical Correlation Analysis (CCA).

```
require(PMA)
dd <- t(breast_multi$RNAseq)
sout <- SPC(dd, sumabsv=3,
            K=2, orth=TRUE)

## 12345678910
## 1234567891011121314151617181920
```

NOTE: `sumabsv` argument controls the degree of sparsity. It can be tuned by using `SPC.CV` function

Sparse PCA analysis

```
rownames(sout$u) <- rownames(dd)
rownames(sout$v) <- colnames(dd)
head(sout$u)

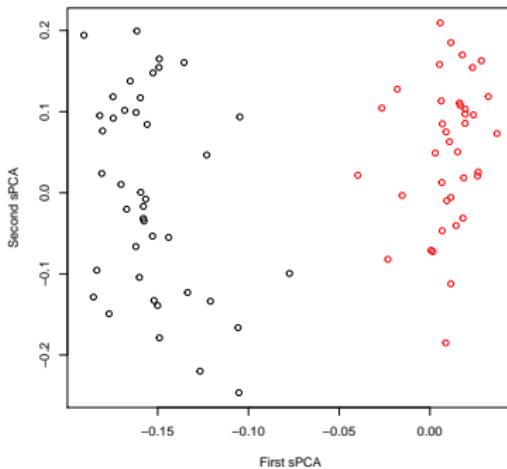
##           [,1]      [,2]
## TCGA-C8-A12V -0.1266630 -0.2198702
## TCGA-A2-A0ST -0.1052587 -0.2464092
## TCGA-E2-A159 -0.1057203 -0.1662398
## TCGA-BH-A0BW -0.1335251 -0.1229296
## TCGA-A2-A0SX -0.1520058 -0.1327320
## TCGA-AR-A1AI -0.1768082 -0.1491056

head(sout$v)

##           [,1]  [,2]
## CREB3L1      0   0
## PNMA1        0   0
## MMP2         0   0
## C10orf90     0   0
## GPR98        0   0
## APBB2        0   0
```

Sparse PCA analysis

```
plot(sout$u, type="n", xlab="First sPCA", ylab="Second sPCA")
points(sout$u, col=as.numeric(group))
```



Sparse PCA analysis

```
ss <- sout$v[,1]
ss[ss!=0]

##      FSIP1      ROPN1B      SLC44A4      ROPN1      FOXA1      SCGB2A2
## 0.12976786 -0.20047758  0.09854402 -0.22180706  0.35283734  0.02914304
##      CA12      HORMAD1      TFF3       ABCC11      ESR1      AGR3
## 0.08389962 -0.11978720  0.01092539  0.01843790  0.26810418  0.69491639
##      AGR2     Clorf64      SCUBE2      ART3
## 0.29917578  0.23368898  0.05023190 -0.18825577

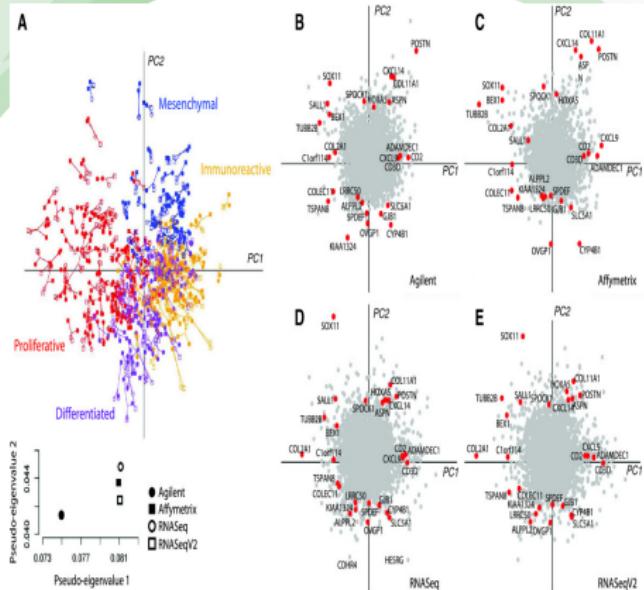
ax1

## [1] "ROPN1B"   "ROPN1"    "FOXA1"    "SFRP1"    "ESR1"    "AGR3"    "AGR2"
## [8] "Clorf64"  "ART3"     "BCL11A"
```

Outline

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets**
- 4 Take home messages
- 5 Recommended lectures

One omic dataset: dimensionality reduction



Integrating two or more omic datasets

- There are methods based on dimension reduction techniques: generalized SVD, Co-Inertia Analysis (CIA), sparse or penalized extensions of Partial Least Squares (PLS), Canonical Correlation (CCA), Multiple Factor Analysis (MFA), Generalized Canonical Correlation (GCA)
- There are methods that are based on path modelling: structural equation models (SEM)
- Regularized Generalized Canonical Correlation (RGCA) provides a unified framework for different approaches (Tenenhaus, M. and Tenenhaus, A. Psychometrika, 2011, 76(2):257-284). Other methods are special cases of RGCA.
- RGCA integrates a feature selection method, named sparse GCCA (SGCCA) (Tenenhaus, A et al. Biostatistics, 2014, 15(3):569-83).

Integrating two or more omic datasets

- There are methods based on dimension reduction techniques: generalized SVD, Co-Inertia Analysis (CIA), sparse or penalized extensions of Partial Least Squares (PLS), Canonical Correlation (CCA), Multiple Factor Analysis (MFA), Generalized Canonical Correlation (GCA)
- There are methods that are based on path modelling: structural equation models (SEM)
 - Regularized Generalized Canonical Correlation (RGCA) provides a unified framework for different approaches (Tenenhaus, M. and Tenenhaus, A. Psychometrika, 2011, 76(2):257-284). Other methods are special cases of RGCA.
 - RGCA integrates a feature selection method, named sparse GCCA (SGCCA) (Tenenhaus, A et al. Biostatistics, 2014, 15(3):569-83).

Integrating two or more omic datasets

- There are methods based on dimension reduction techniques: generalized SVD, Co-Inertia Analysis (CIA), sparse or penalized extensions of Partial Least Squares (PLS), Canonical Correlation (CCA), Multiple Factor Analysis (MFA), Generalized Canonical Correlation (GCA)
- There are methods that are based on path modelling: structural equation models (SEM)
- Regularized Generalized Canonical Correlation (RGCA) provides a unified framework for different approaches (Tenenhaus, M. and Tenenhaus, A. Psychometrika, 2011, 76(2):257-284). Other methods are special cases of RGCA.
- RGCA integrates a feature selection method, named sparse GCCA (SGCCA) (Tenenhaus, A et al. Biostatistics, 2014, 15(3):569-83).

Integrating two or more omic datasets

- There are methods based on dimension reduction techniques: generalized SVD, Co-Inertia Analysis (CIA), sparse or penalized extensions of Partial Least Squares (PLS), Canonical Correlation (CCA), Multiple Factor Analysis (MFA), Generalized Canonical Correlation (GCA)
- There are methods that are based on path modelling: structural equation models (SEM)
- Regularized Generalized Canonical Correlation (RGCA) provides a unified framework for different approaches (Tenenhaus, M. and Tenenhaus, A. Psychometrika, 2011, 76(2):257-284). Other methods are special cases of RGCA.
- RGCA integrates a feature selection method, named sparse GCCA (SGCCA) (Tenenhaus, A et al. Biostatistics, 2014, 15(3):569-83).

Integrating two or more omic datasets

- Canonical Correlation can be seen as an extension of PCA for more than two tables X and Y
- The two datasets can be decomposed as:

$$f = Xp$$

$$g = Yq$$

where p and q are the loading vectors

- CCA searches for association or correlations among X and Y by

$$\arg \max_{p^i q^i} \text{cor}(Xp^i Yq^i)$$

for the i -th component

- Xp^i and Yq^i are known as canonical variates and their correlations are the canonical correlations.

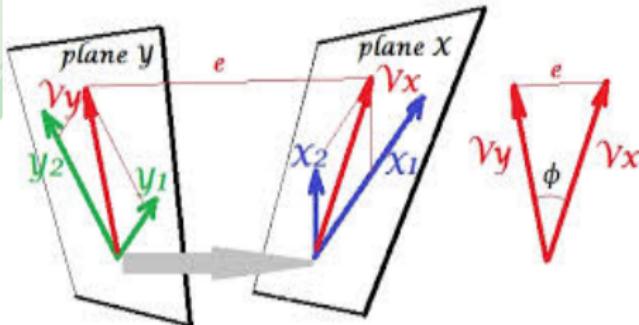
Integrating two or more omic datasets

- $(p >> n)$ is an issue. Additionally, there is often presence of multicollinearity within both sets of variables that requires a regularization step.
- This may be accomplished by adding a ridge penalty, that is, adding a multiple of the identity matrix to the correlation/covariance matrix.

$$\arg \max_{p^i q^i} \text{cor}(X p^i Y q^i) + \lambda I$$

- A sparse solution (filtering the number of variables) is the solution: pCCA, sCCA, CCA-I1, CCA-EN, CCA-group sparse have been used to integrate two omic data.

Integrating two or more omic datasets



V_y and V_x are selected to maximize:

- Correlation (CCA)
- Squared Covariance (CIA)

$$\arg \max_{p^i q^i} \text{cov}^2(X p^i Y q^i)$$

Canonical Correlation

- CCA has been used in omic data
- The main limitation is that the number of features generally greatly exceeds the number of observations
- Consequence 1: parameter estimation cannot be applied using standard methods
- Consequence 2: Most of markers are having no effect in the canonical axes (e.g. almost all components in a and b are 0)
- Penalized (sparse) CCA has been proposed
- Main advantages: More than two tables, easy interpretation (do not need computing p-values nor correcting for multiple comparisons)

Sparse canonical correlation

$$U_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$U_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

 \vdots

$$U_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q$$

$$V_2 = b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q$$

 \vdots

$$V_p = b_{p1}Y_1 + b_{p2}Y_2 + \cdots + b_{pq}Y_q$$

a and b for i-th canonical variable is obtained by maximizing

$$\rho_i^* = \frac{\text{cov}(U_i, V_i)}{\sqrt{\text{var}(U_i)\text{var}(V_i)}}$$

Subject to:

$$\|a\|^2 \leq 1 \text{ and } \|b\|^2 \leq 1$$

Co-Inertia

- CIA does not require an inversion step of the covariance matrix; thus, regularization or penalization implementation is not required
- CIA can deal with disperse variables
- CIA does consider quantitative or qualitative variables
- Can weight cases
- The method provides the RV coefficient. This is a measure of global similarity between the datasets, and is a number between 0 and 1. The closer it is to 1 the greater the global similarity between the two datasets.

Canonical Correlation: gene expression and proteins

```
require(CCA)
df1 <- t(breast_multi$RNAseq) [,1:1000]
df2 <- t(breast_multi$RPPA)
```

```
resCC <- cc(df1, df2)
```

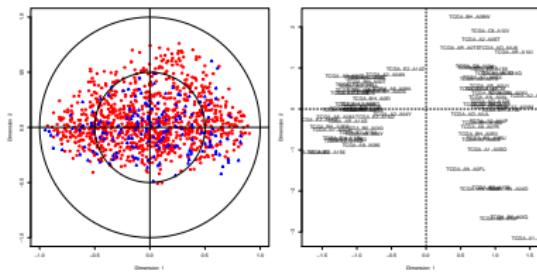
Error en chol.default(Bmat) :
la submatriz de orden 81 no es definida positiva

```
resRCC <- rcc(df1, df2, 0.2, 0.1)
```

```
regul <- estim.regul(df1, df2)
resRCC2 <- rcc(df1, df2, regul$lambda1, regul$lambda2)
```

```
plt.cc(resRCC)
```

Canonical Correlation: gene expression and proteins



Canonical Correlation: gene expression and proteins

```
require(PMA)
ddlist <- list(df1, df2)
perm.out <- MultiCCA.permute(ddlist,
                               type=c("standard", "standard"),
                               trace=FALSE)

resMultiCCA <- MultiCCA(ddlist,
                         penalty=perm.out$bestpenalties,
                         ws=perm.out$ws.init,
                         type=c("standard", "standard"),
                         ncomponents=1, trace=FALSE, standardize=TRUE)
```

Canonical Correlation: gene expression and proteins

```
rownames(resMultiCCA$ws[[1]]) <- colnames(df1)
rownames(resMultiCCA$ws[[2]]) <- colnames(df2)
head(resMultiCCA$ws[[1]])

## [,1]
## CREB3L1    0.03812575
## PNMA1     0.04078831
## MMP2      0.02402715
## C10orf90 -0.02896362
## GPR98     0.04723420
## APBB2     0.06279879

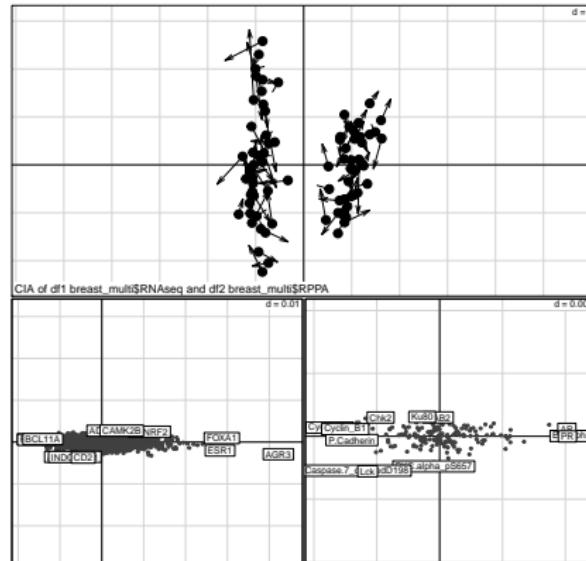
head(resMultiCCA$ws[[2]])

## [,1]
## c.Myc     -0.07099252
## HER3      0.02652238
## XBP1      0.00000000
## Fibronectin 0.01527691
## PAI.1     -0.02319958
## p21       0.02415011
```

Co-inertia: gene expression and proteins

```
resCIA <- cia(breast_multi$RNAseq, breast_multi$RPPA)
```

```
plot(resCIA, classvec=group, nlab=3, clab=0, cpoint=3 )
```



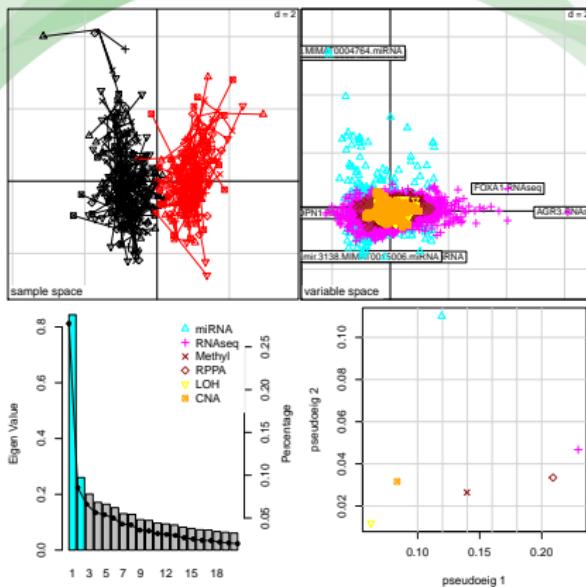
More than two tables

```
require(omicade4)
resMCIA <- mcia( breast_multi[ c(1, 3, 4, 5, 6, 7) ] )

plot(resMCIA, axes=1:2, sample.lab=FALSE, sample.legend=FALSE,
phenovec=group, gene.nlab=2,
df.color=c("cyan", "magenta", "red4", "brown", "yellow", "orange")
df.pch=2:7)

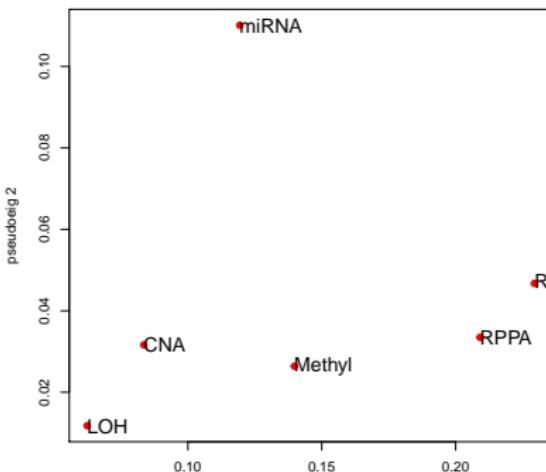
## Error in xy.coords(x, y, xlabel, ylabel, log): 'x' is a
list, but does not have components 'x' and 'y'
```

More than two tables



More than two tables

```
plot(resMCIA$mcoa$cov2, xlab = "pseudoeig 1",
      ylab = "pseudoeig 2", pch=19, col="red")
text(resMCIA$mcoa$cov2, labels=rownames(resMCIA$mcoa$cov2),
     cex=1.4, adj=0)
```



Outline

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets
- 4 Take home messages
- 5 Recommended lectures

Concluding remarks

Correlation $Y = X$

Multiple correlation $Y = X_1 X_2 X_3 \cdots X_k$

Canonical correlation $Y_1 Y_2 Y_3 \cdots Y_j = X_1 X_2 X_3 \cdots X_k$

- Co-inertia analysis (CIA) is similar to CCA but it optimizes the squared covariance between the eigenvectors while CC optimizes the correlation.
- CIA can be applied to datasets where the number of variables (genes) far exceeds the number of samples (arrays) such is the case in several omic data, while CCA requires a regularized version to be implemented.
- Sparse and regularized methods require a tuning parameter. This makes these methods computing demanding.

Take home messages

- Multivariate methods are purely descriptive methods that do not test a hypothesis to generate a p-value.
- They are not optimized for variable of biomarkers discovery, though the introduction of sparsity in variable loadings may help in the selection of variables for downstream analyses.
- Number of variables in omic data is a challenge to traditional visualization tools. New R packages including `ggord` are being developed to address this issue.
- Dynamic visualization is possible using `ggvis`, `ploty`, `explor` and other packages.
- Projection in the same space of variable annotation (GO or Reactome) may help to determine gene sets or pathways associated with our traits.

Outline

- 1 Introduction
- 2 One omic dataset: dimensionality reduction
- 3 Integrating two or more omic datasets
- 4 Take home messages
- 5 Recommended lectures

Recommended lectures

- Millstein et al. (2009). Disentangling molecular relationships with a causal inference test. *BMC Bioinf*, 10:23.
- Ebrahim and Smith (2008). Mendelian randomization: can genetic epidemiology help redress the failure of observational epidemiology?. *Hum Genet*, 123:15-33.
- Liu et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotech*, 31(2):142-148.
- Voight et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*, 380(9841): 572-580.
- Meng et al. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinf*, 15:162.
- Witten et al. (2009). A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis. *Biostatistics*, 10(3):515-34.
- Meng et al (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform*, 17(4): 628-641.