Integrative analysis of genetic and glycan data using O2PLS

Said el Bouhaddani, Hae-Won Uh, Geurt Jongbloed and Jeanine Houwing

June 7, 2017

BMTL 2017, Napoli

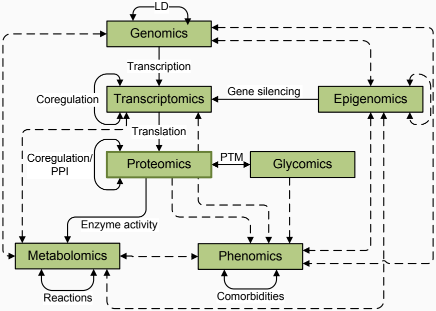
Introduction

Multiple Omics

- Recent advances in technology provided many types of omics
- Different levels of biological variation measured

⇒ How much is the overlap between these data?

⇒ Which (types of) molecules are responsible for this overlap?



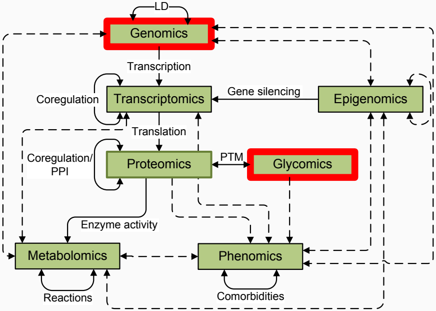
Zierer et al, 2015, Aging Cell

Multiple Omics

- Recent advances in technology provided many types of omics
- Different levels of biological variation measured

⇒ How much is the overlap between these data?

⇒ Which (types of) molecules are responsible for this overlap?



Zierer et al, 2015, Aging Cell

What are IgG glycans?

- IgG glycans are highly correlated
- Not clear how this correlation is built up
  - Genetics
  - Environment
  - Measurement error
- How much is genetic part?
- Which genes are correlated with which glycans?

⇒ Study relationship between genes and glycans

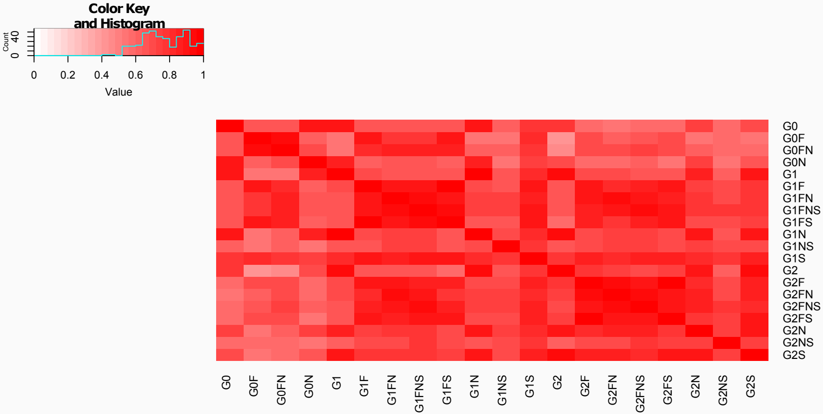
Lauc et al. Biochimica et biophysica acta (2016)

IgG glycan datasets

Data:

- 20 measurements from CROATIA_Korcula ($N = 885$)
- Corrected for age and sex

- Data matrix: Y (885×20)
- High correlations

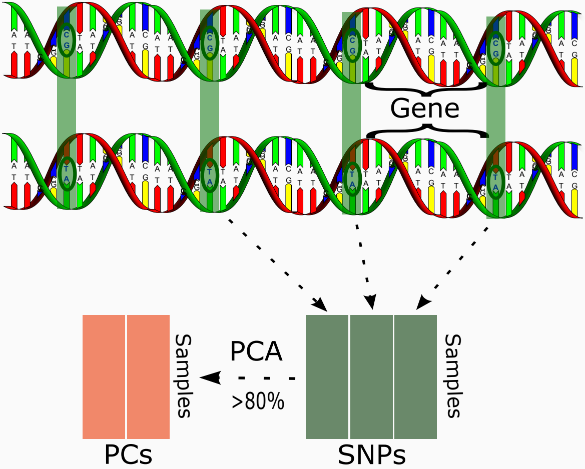


Genetic data

Data:

- 300.000 SNPs ($N = 885$)
- Structured correlation

- Data matrix: X (885×37.819)
- High dimensionality



Properties and aim

Properties:

- High correlations among glycans
- High dimensional genetic data

Challenge:

- Relationships between genes and glycans

Approach:

- Latent variables to model high correlations (what combination of glycans cause which part of correlation?)
- Dimension reduction to reduce data dimensionality

⇒ Partial Least Squares approach to include both

Methods

Principal components analysis

Population model associated with PCA

$$X = TW^{\mathrm{T}} + E$$

Properties

- T latent scores or PC's
- W loadings
- Components maximize $\mathrm{Var}(\mathrm{XW})$

Principal components analysis

Population model associated with PCA

$$X \quad = \quad TW^{\mathrm{T}} \quad + \quad E$$

Properties

- T latent scores or PC's
- W loadings
- Components maximize $\mathrm{Var}(\mathrm{XW})$

Interpretation

- Scores: Which *subjects* contribute most to components
- Loadings: Which *variables* contribute most to components

Partial Least Squares (PLS)

Population model associated with PLS

$$\begin{aligned} X &= TW^{\mathrm{T}} + E \\ \underbrace{Y}_{\textit{Data}} &= \underbrace{UC^{\mathrm{T}}}_{\textit{Joint}} + \underbrace{F}_{\textit{Noise}} \end{aligned}$$

Properties

- T and U joint scores
- W and C joint loadings
- Regression of U on T:

$$U = TB + H$$

- Joint Principal Components maximize $\mathrm{Cov}(\mathrm{XW}, \mathrm{YC})$

Partial Least Squares (PLS)

Population model associated with PLS

$$\begin{aligned} X &= TW^{\mathrm{T}} + E \\ \underbrace{Y}_{Data} &= \underbrace{UC^{\mathrm{T}}}_{Joint} + \underbrace{F}_{Noise} \end{aligned}$$

Properties

- T and U joint scores
- W and C joint loadings
- Regression of U on T:

$$U = TB + H$$

- Joint Principal Components maximize $\mathrm{Cov}(\mathrm{XW}, \mathrm{YC})$

Omic-specific variation

- Systematic differences between datasets
- One example: batch effects independently in both datasets
- We need extension of PLS to take this into account

Two-way Orthogonal PLS

Population model associated with O2PLS

$$\begin{array}{ccccccccc} X & = & TW^{\mathrm{T}} & + & T_{\perp}P_{Y\perp}^{\mathrm{T}} & + & E \\ \underbrace{Y}_{Data} & = & \underbrace{UC^{\mathrm{T}}}_{Joint} & + & \underbrace{U_{\perp}P_{X\perp}^{\mathrm{T}}}_{Specific} & + & \underbrace{F}_{Noise} \end{array}$$

- $T_{\perp}$ and $U_{\perp}$ omic-specific scores
- $P_{Y\perp}$ and $P_{X\perp}$ omic-specific loadings

$\Rightarrow$ T, U, W and C are corrected for omic-specific variation

Trygg & Wold, 2003, J. Chemometrics

Estimation

- Three-step estimation

$$\boxed{\begin{array}{ccccccc} X & = & TW^{\mathrm{T}} & + & T_{\perp}P_{Y\perp}^{\mathrm{T}} & + & E \\ \underbrace{Y}_{\textit{Data}} & = & \underbrace{UC^{\mathrm{T}}}_{\textit{Joint}} & + & \underbrace{U_{\perp}P_{X\perp}^{\mathrm{T}}}_{\textit{Specific}} & + & \underbrace{F}_{\textit{Noise}} \end{array}}$$

- Suppose we want r joint components, r_X X-specific components and r_Y Y-specific components.

1 Retain $r + \max(r_X, r_Y)$ components from PLS on X and Y

  - W and T contain both joint and specific part

2a Retain r_X components from PLS on E and T

2b Correct X: $X_c = X - T_{\perp}P_{Y\perp}^{\mathrm{T}}$

3 PLS on X_c and Y_c yield corrected joint components

We implemented O2PLS in the **OmicsPLS** package on CRAN

Data analysis

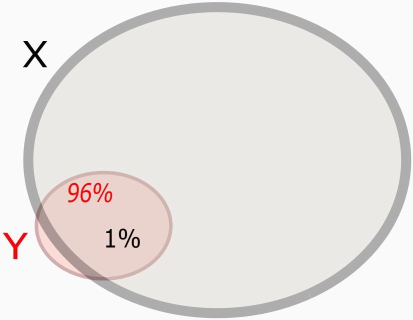
Two key questions

- How much is overlap between genes and glycans?
- Which genes/glycans are in this overlap?

Summary of variation

PLS

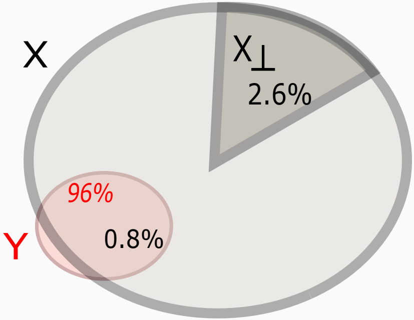
5 joint components



O2PLS

5 joint components

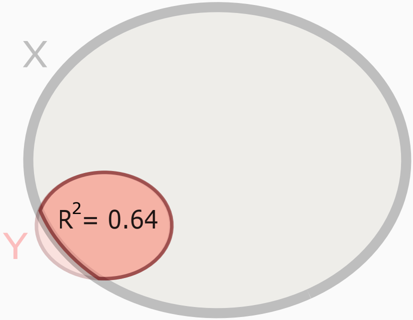
5 genetic-specific components



Summary of variation

PLS

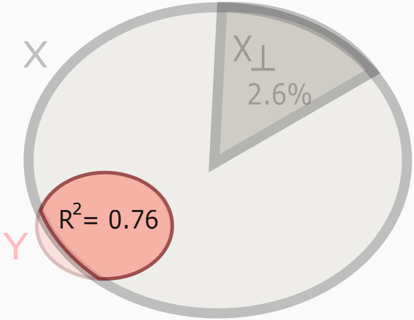
5 joint components



O2PLS

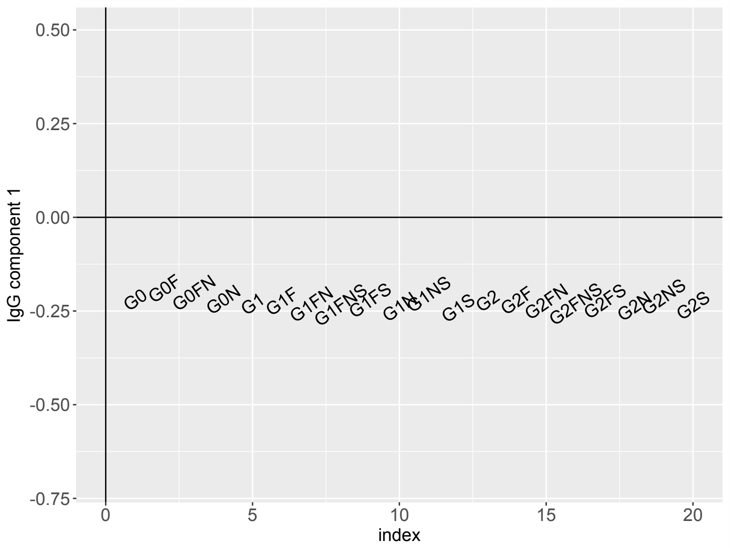
5 joint components

5 genetic-specific components



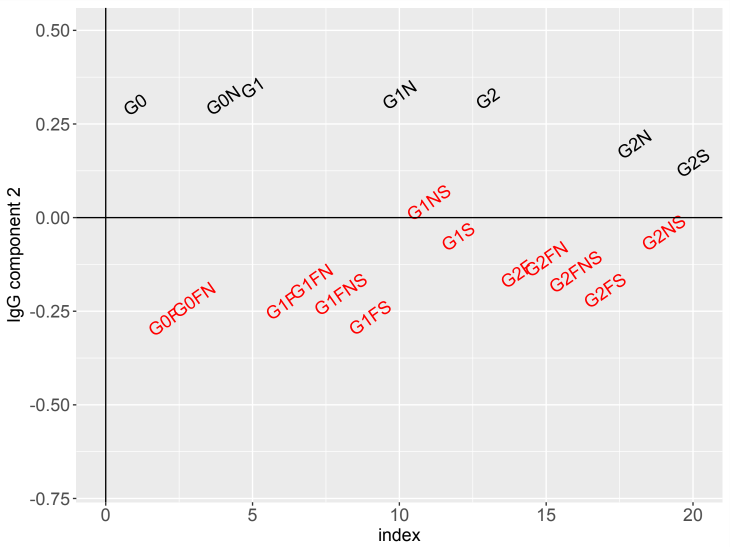
Top genes in component 1

Comp. 1
DNAJC10
ARID3B
ZNF502



Top genes in component 2

Comp. 2
FUT8
LGALS8
LDB3



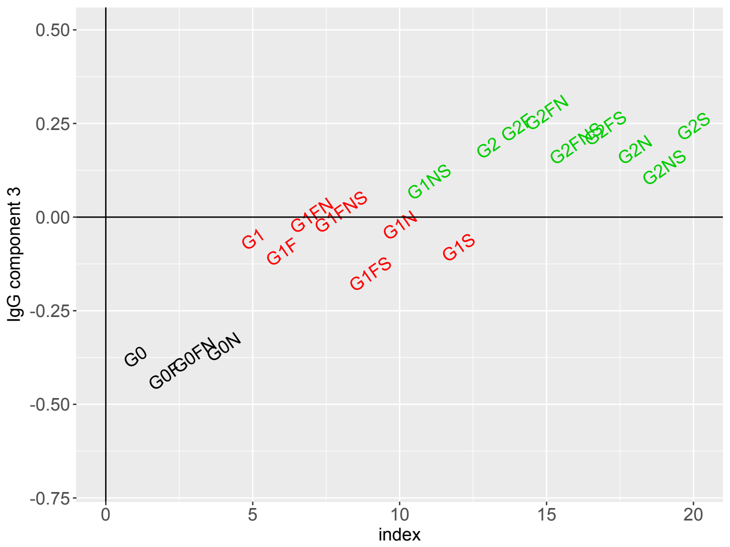
Top genes in component 3

Comp. 3

MTO1

AKAP9

MRPL33



Interpretation

Comp.	**Gene**	**Protein involved in**	**Pattern**
1	*DNAJC10*	recognizing and degrading misfolded glycoproteins	Average
2	*FUT8*	the transfer of fucose	F vs non-F
2	*LGALS8*	detecting and restricting proliferation of pathogens	F vs non-F
3	*AKAP9*	maintaining integrity of the Golgi apparatus	G0 vs G2

Do we understand these relationships?

Discussion

Summary

- O2PLS yields interpretable components
  - Glycan patterns reflect enzymatic reactions and disease pathways
- Estimation of genetic contributions
  - How much overlap: 96% with R^2 of 0.76
  - Which genes: Established and new findings

Future work

Data results

- Validate on Vis cohort
- Try other "summarizing" approaches for SNPs

Probabilistic O2PLS model and extensions

- Epidemiological studies:
  - Missing data
  - Several studies available
- Multiple imputation of missing blocks/datasets
- Meta analysis of several datasets across multiple cohorts
- Extend to complex models involving covariates and outcome

Acknowledgments

MIMOmics

$\hat{\sigma}^2$

MIMOmics and

LUMC

- Jeanine Houwing
- Hae-Won Uh
- Szymon Kielbasa
- Karli Reiding

TU Delft

- Geurt Jongbloed

Univ Edinburgh

- Lucija Klaric

LU MC Leids Universitair Medisch Centrum

UNIVERSITY OF LEEDS

TU Delft Delft University of Technology