Data integration using Network and Partial Least Square methods

Jeanine Houwing-Duistermaat

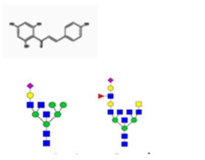
Department of Statistics, University of Leeds
Department of Biostatistics and Research Support, Julius Center, UMC Utrecht
Alan Turing Institute, London

WiN Workshop February 2019

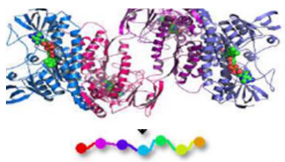
Omics Data

Small datasets
Noisy measurements

Metabolomics

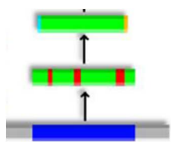


Glycomics



Proteomics

Methylation



Gene Expression

Huge datasets
Accurate measurements

DNA

Data sets represent same biological mechanism

Why omics research?

- My background: statistical genetics
- DNA markers appeared to have limited prediction ability
- My collaborators moved on to omics research (2010)
- Omics datasets: Transcriptomics, Proteomics, Glycomics, Metabolomics
- Two European Projects: MIMOmics and IMforFUTURE

Acknowledgements

- Said el Bouhaddani, Angga Fuady, and Renaud Tissier
- Partners MIMOmics consortium
- Partners IMforFUTURE Network

Properties of Omics Data

- Correlation within and between datasets
- Omic datasets differ in
  - size (number of variables and subjects),
  - scale (type of variable)
  - measurement error (depends on platform)
- High dimensional ($p > n$)
- Noisy

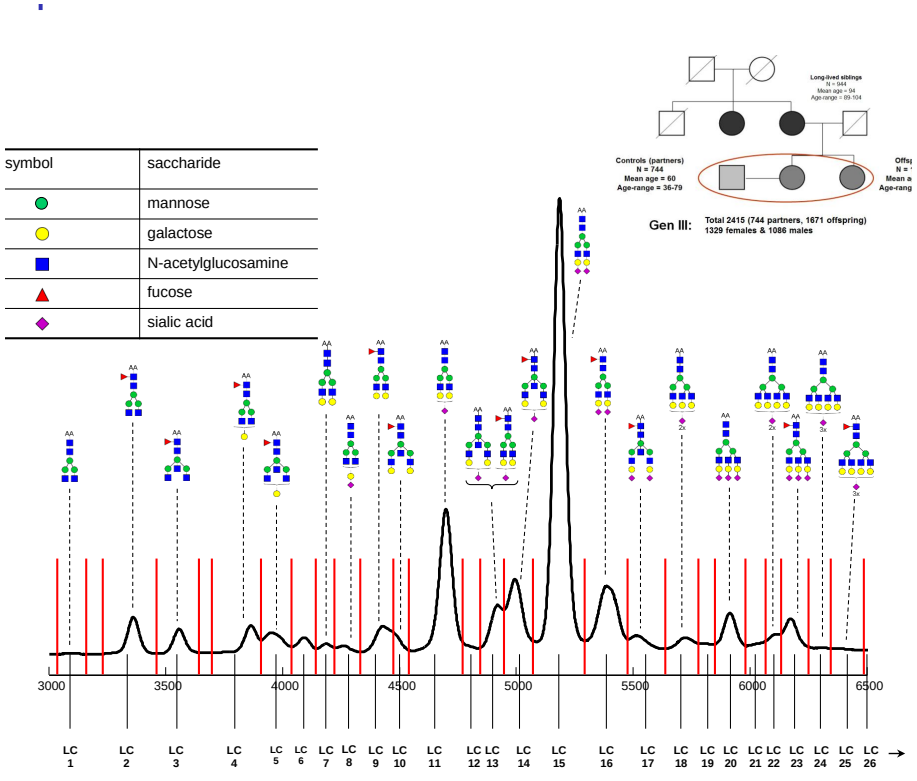
Noisy data

- Measurement process not automated and standardized
- Degeneration of samples
- Detection limit
- Non normal data
- Technique differences can be huge

Statistical Methodology for Omics data

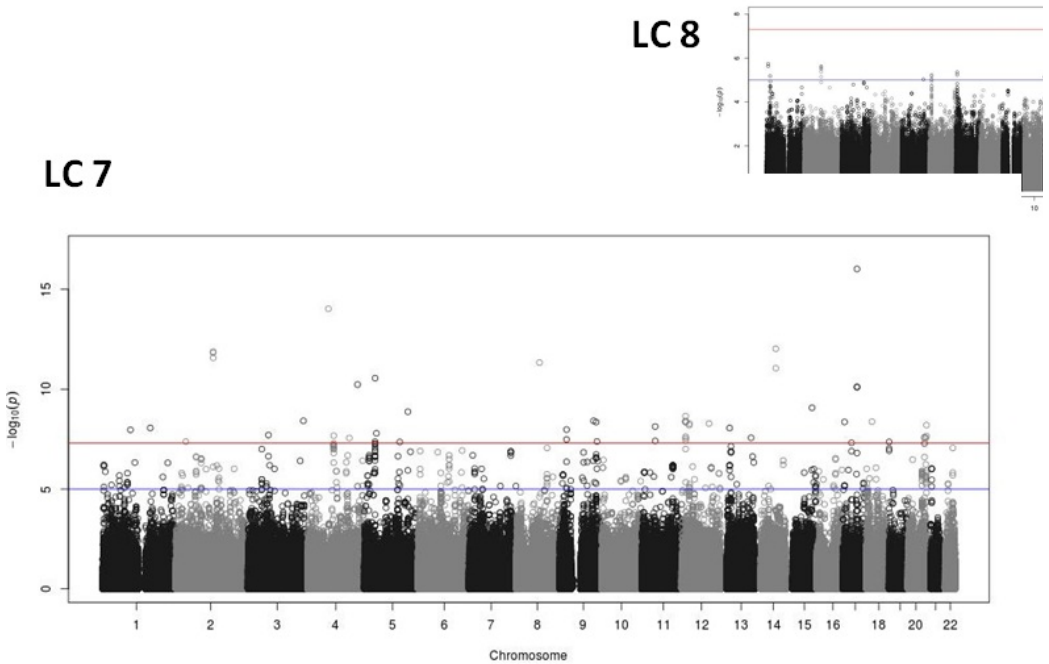
- Data Cleaning, Data Wrangling
- Data Reduction, Integration, Descriptives
- Statistical Inference: association and causality
- Prediction

Data cleaning



- Glycan LC7 was significantly associated with offspring-partner status
- Glycan LC7 was also significantly associated with Diabetes
- But when we performed a GWAS with the Glycomics variables as outcome..

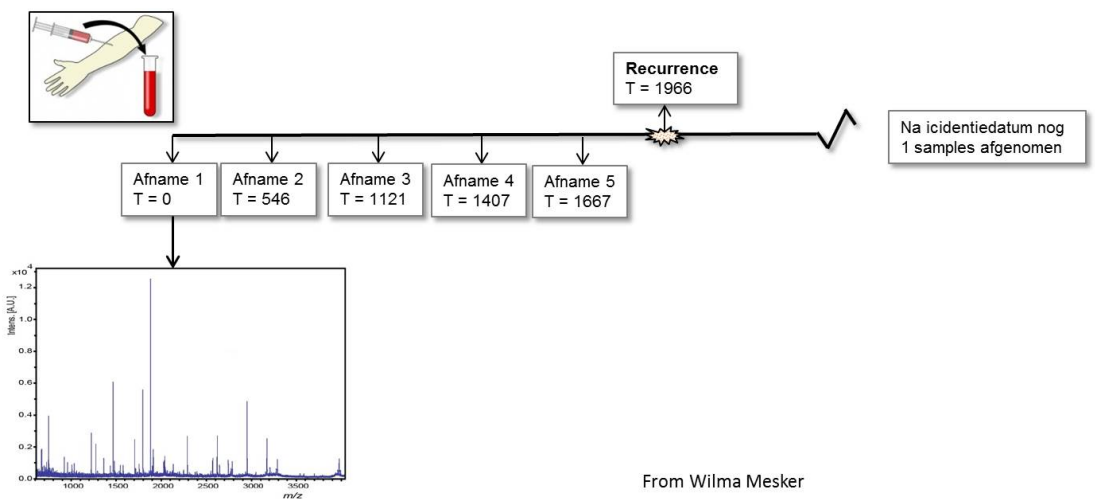
Data cleaning



- GWAS for LC7 yielded too many significant results
- Cause: One batch was not well measured

Classification with outcome: Colorectal Cancer

- **Aim:** to identify subgroups based on Proteomics (FTMS)
- Study Design I: Case-Control, for cases additional information on tumor stage and size is available
- Study Design II: Cases who had surgery and are cancer free (paired pre-post)
- Study Design III: Follow up study of patients who had surgery and are cancer free (screening)



From Wilma Mesker

Results

- Case-control cohort: a set of proteomic biomarkers with prediction accuracy, AUC 0.980 (cross validated)
  - Explanation: Case and control samples appeared not to be equally handled
- **Solution** Study Design II: we build predictor based on data from cases before and after surgery (pre post)
- Resulted in a set of 42 proteomic biomarkers
- AUC of this set was 0.769 in pre-post sample (cross validated) and when applied to case control set the AUC was 0.722

Gene Expression and Huntington Disease

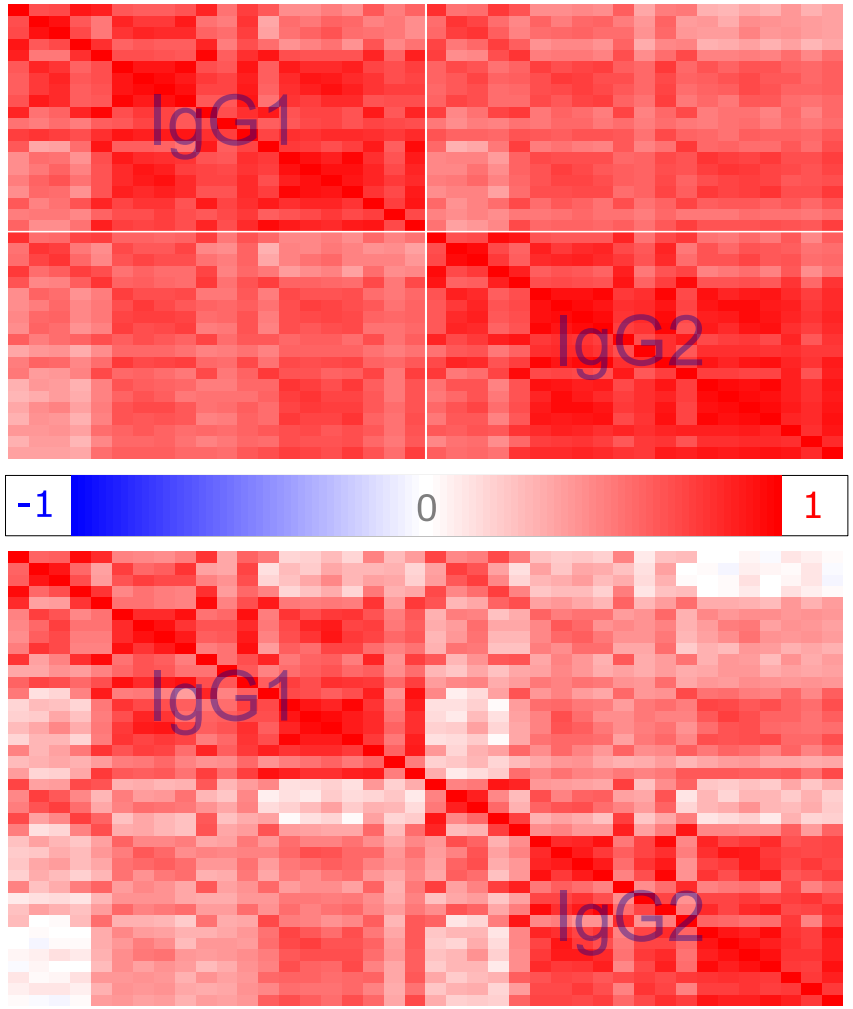
- Analysis of DeepSAGE measurements in Huntington cases and relatives yielded 167 associated genes
- After three years, new samples of same patients were measured with RNAseq and no significant associations were found
- We performed joint analysis of both datasets using measurement error methodology (a small calibration set is available)
- From 167 genes, 14 passed quality control, 9 were significant at time point 1, these were all significant in joint analysis
- Conclusion: Replication failed due to differences in techniques and the quality control was in the original study was not sufficient.

Glycomics datasets Descriptives

- IgG = Immunoglobulins (or antibodies) type G
- Bind to many kinds of pathogens
- Two cohorts: Korcula and Vis
- Two datasets: IgG1 and IgG2 glycans

⇒ Integrate IgG1 and IgG2 glycan data

Correlation heatmaps Korcula and Vis



Various technical platforms - Predicting

- Two measurement platforms for Glycomics
- Three types of studies
  - Cohort 1: UPLC, LCMS
  - Cohort 2: UPLC
  - Cohort 3: LCMS
- Goal: Performing Glycomics GWAS using all data
- Needed: mapping from LCMS to UPLC

Prediction based on two omics datasets

- CNV and Gene Expression to predict treatment response in cancer cell lines
- 45 cell lines, 637 CNVs, 5375 probes
- Problem two omics datasets are very different
- Stacking the datasets and applying standard prediction methods result in bad prediction accuracy:

Method	CNV	gene expression	prediction
Lasso	.934	.571	.576
Ridge	.454	.610	.614

Dimension Reduction and Integration

- Facilitating insights
- Predicting one dataset from another dataset
- Predicting an outcome from one or multiple omics datasets
  - One dataset: Correlation within a dataset gives different results in different Cross Validation steps
  - Multiple datasets: Stacking of datasets does not work.

Notation

- Y $n \times r$ matrix of r (response) variables for n individuals.
- X $n \times p$ matrix of p (explanatory) variables for n individuals.
- var$(X) = \Sigma_x$, var$(Y) = \Sigma_y$ and cov$(X, Y) = \Sigma_{xy}$.
- mean of $Y = \mu_y$ and $X = \mu_x$.
- Distribution of vector (Y_i, X_i) of observations of individual i of length $r + p$ follows multivariate normal distribution.

Model for Relationship between X and Y

- Multiple multivariate regression: $Y = \mu_y + (X - \mu_x)\beta + e$
- Here β is $p \times r$ matrix of regression coefficients
- Best linear predictor: $\hat{Y} = \hat{\mu}_y + (X - \hat{\mu}_x)\hat{\beta}$
- with $\hat{\beta} = \Sigma_x^{-1}\Sigma_{xy}$
- Estimation of β might be hampered by large dimension and or correlations between features of X.
- Idea is to use only relevant part of X (and of Y)

Decompose X in subspaces

Components g spanning a subspace of X are relevant if

- They are correlated with Y.
- Not correlated with the irrelevant part.
- Further, the irrelevant part is not correlated with Y.

PLS

- Reduction of original space of dimension p to $K < p$
- p-dimensional vectors g_k form basis subspace, $k = 0, \cdots, K$. Here $g_0 = 0$.
- Let G_k be $(g_0, \cdots, g_k)$.
- Vectors g_k can be obtained using the following algorithm: $g_{k+1} = \underset{g}{\operatorname{argmax}}\ g\Sigma_{xy}\Sigma_{xy}^T G_k$, under conditions $g_{k+1}^T \Sigma_x G_k = 0$ and $g^T g = 1$.
- Find compontents g_k with largest covariance with Y and orthogonal on previous identified components.

PLS pros and cons

- Pros
  - Fast
  - Robust, does not assume normality
- Cons
  - Algorithm and no model
  - No unique solution
  - There are many versions of the algorithm
- Recently O2PLS was developed. This method considers a third subspace: it splits the relevant part in a part correlated with Y and a X specific part

Probabilistic Partial Least Squares

Latent variable model: Probabilistic PLS (PPLS)[a]

[a] el Bouhaddani et al, 2018 (accepted), J. Multivar. Analysis

$$x = tW^{\mathrm{T}} + e$$
$$y = uC^{\mathrm{T}} + f$$
$$u = tB + h$$

Probabilistic model

- Normal distribution
- Correlated t and u, independent noise variables e and f
- Parameters of interest: W and C (weights)
- Number of JPCs: K

Identifiability

- Assumption 1: $W^{\mathrm{T}}W = C^{\mathrm{T}}C = I$
- Assumption 2: Independent $\{t_1, \ldots, t_K\}$ and $\{u_1, \ldots, u_K\}$

EM algorithm

- Latent variable model: EM algorithm
- Complete data likelihood:

$$f(x, y, t, u) = f(x|t)f(y|u)f(u|t)f(t)$$

- Distinct parts: e.g. maximise only over W
- Constrained optimisation over $W^{\mathrm{T}}W = I$

$$\log f(x|t) \propto ||x - tW^{\mathrm{T}}||^2 + \Lambda||W^{\mathrm{T}}W - I||^2 + \text{const.}$$

- E step: $\hat{t} = \mathbb{E}(t|x, y)$
- M step: $W^{\text{next}} = \text{orth}(x^{\mathrm{T}}\hat{t})$

Standard Errors

- Observed Fisher information matrix I
- First and second derivative simultaneously over all components
- $\sqrt{\text{diag}(-I_{\text{obs}}^{-1})}$ contain standard errors

R packages available:

- OmicsPLS[1] (for PLS *and more*), available on `cran`
- PPLS, see `github.com/selbouhaddani`

[1] el Bouhaddani et al, 2018 (minor rev.), BMC Bioinformatics

PPLS pros and cons

- Pros
  - A model
  - Likelihood based
  - Can deal with $p > n$
  - Decomposes X and Y
  - Unique solution
- Cons
  - Several constraints
  - Correlated latent spaces have same dimension
- Currently we are finishing the PO2PLS version

PO2PLS

- Extension of PPLS

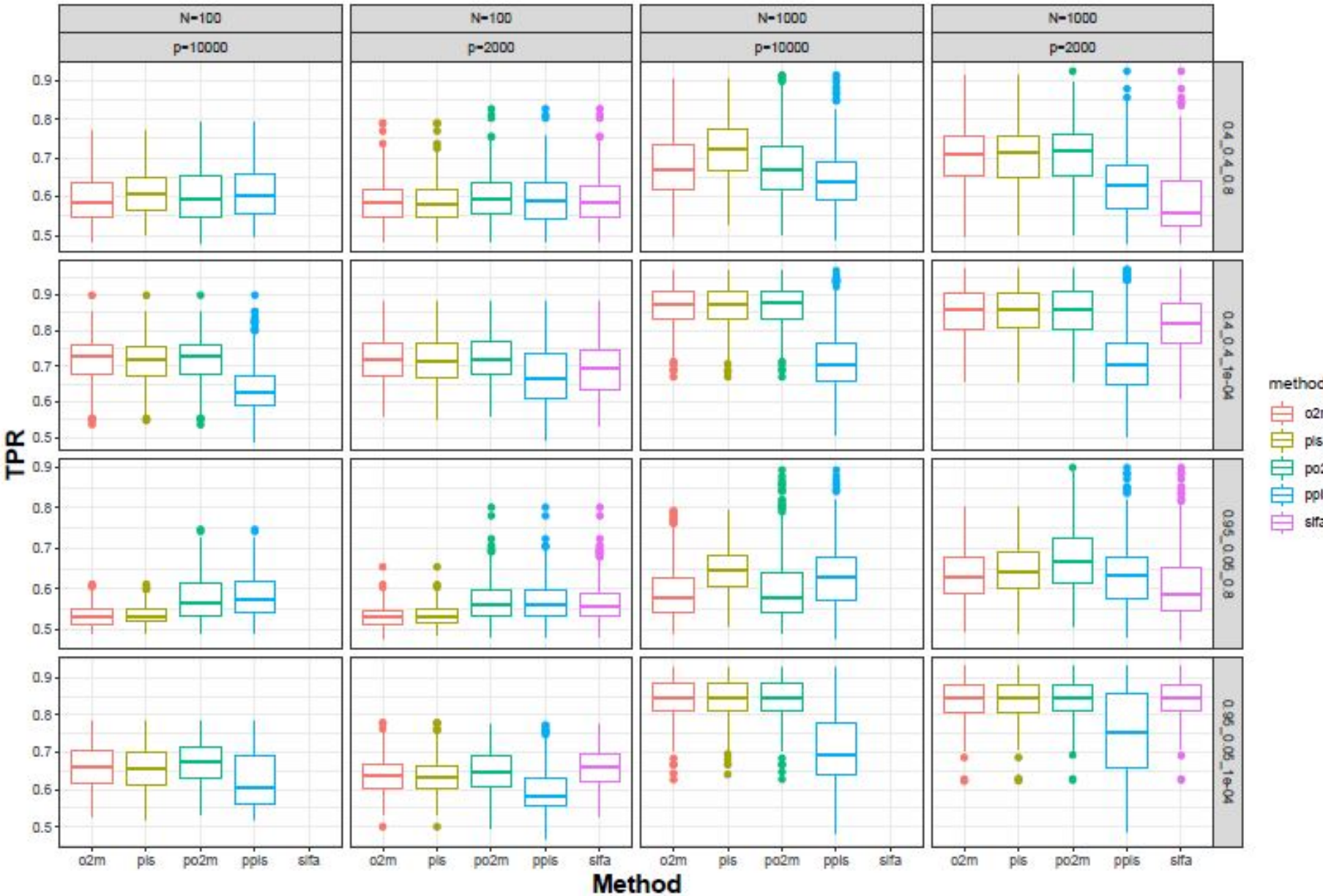
$$x = tW^{\mathrm{T}} + t_x W_x^{\mathrm{T}} + e$$
$$y = uC^{\mathrm{T}} + u_y C_y^{\mathrm{T}} + f$$
$$u = tB + h$$

- Several constraints
- EM algorithm for estimation
- Standard errors

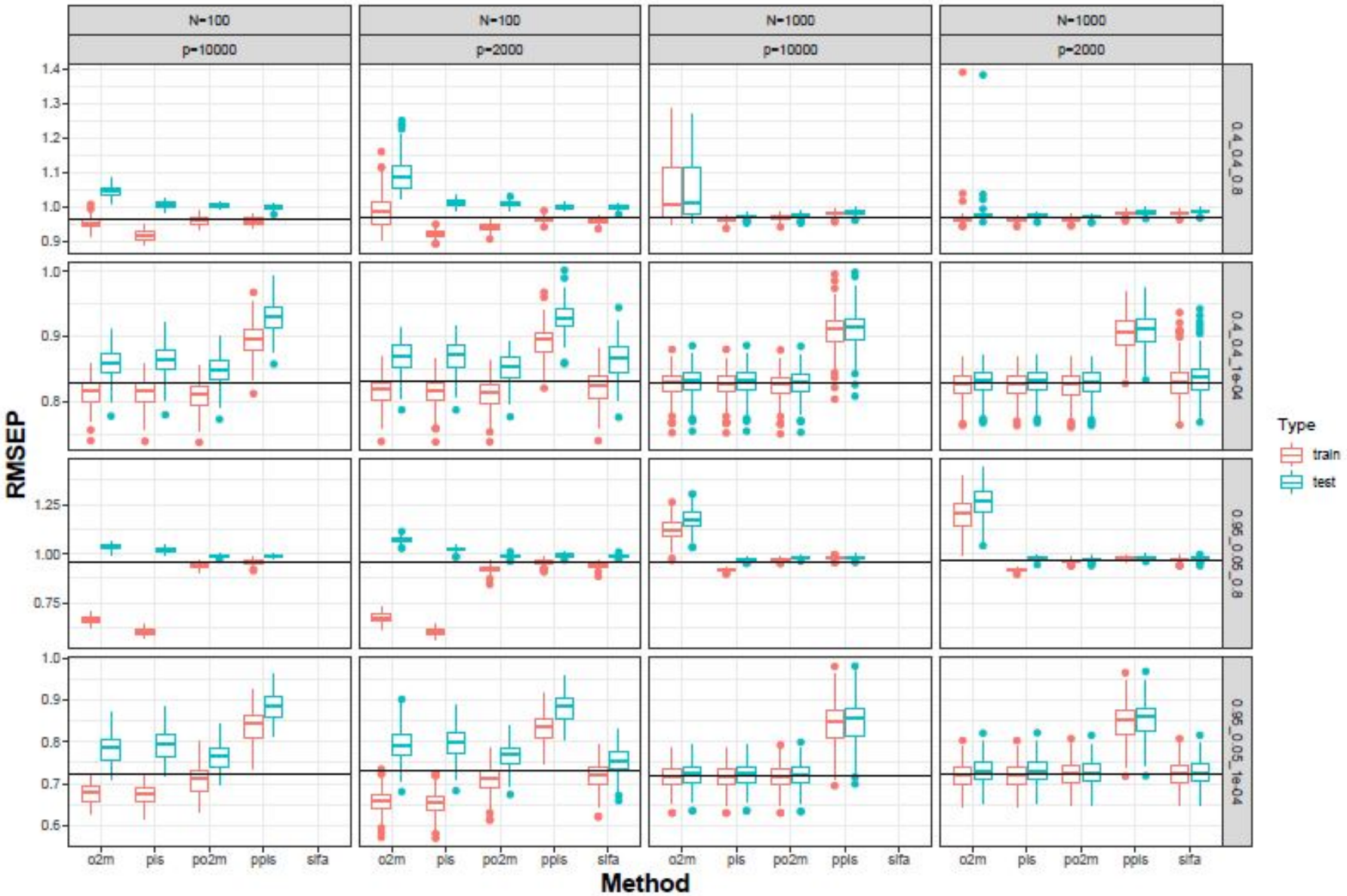
Simulation setting

- PLS, O2PLS, PPLS, PO2PLS
- Evaluation measures true positive rates and prediction error
- $N = 100, 1000$
- $p = 2000, 10000; r = 125, 25$
- Noise levels $(95\%, 5\%), (40\%, 40\%)$
- $q_x = q_y = 5$
- Data Specific part $0\%80\%$
- True Positive Rates in top 25%
- Prediction accuracy $\left\| Y - \hat{Y} \right\|_F$

Results TPR


N=100
p=10000
N=100
p=2000
N=1000
p=10000
N=1000
p=2000
0.4_0.4_0.8
0.4_0.4_1e-04
0.95_0.05_0.8
0.95_0.05_1e-04
TPR
0.9
0.8
0.7
0.6
0.5
o2m
pls
po2m
ppls
sifa
Method
method

Results Prediction



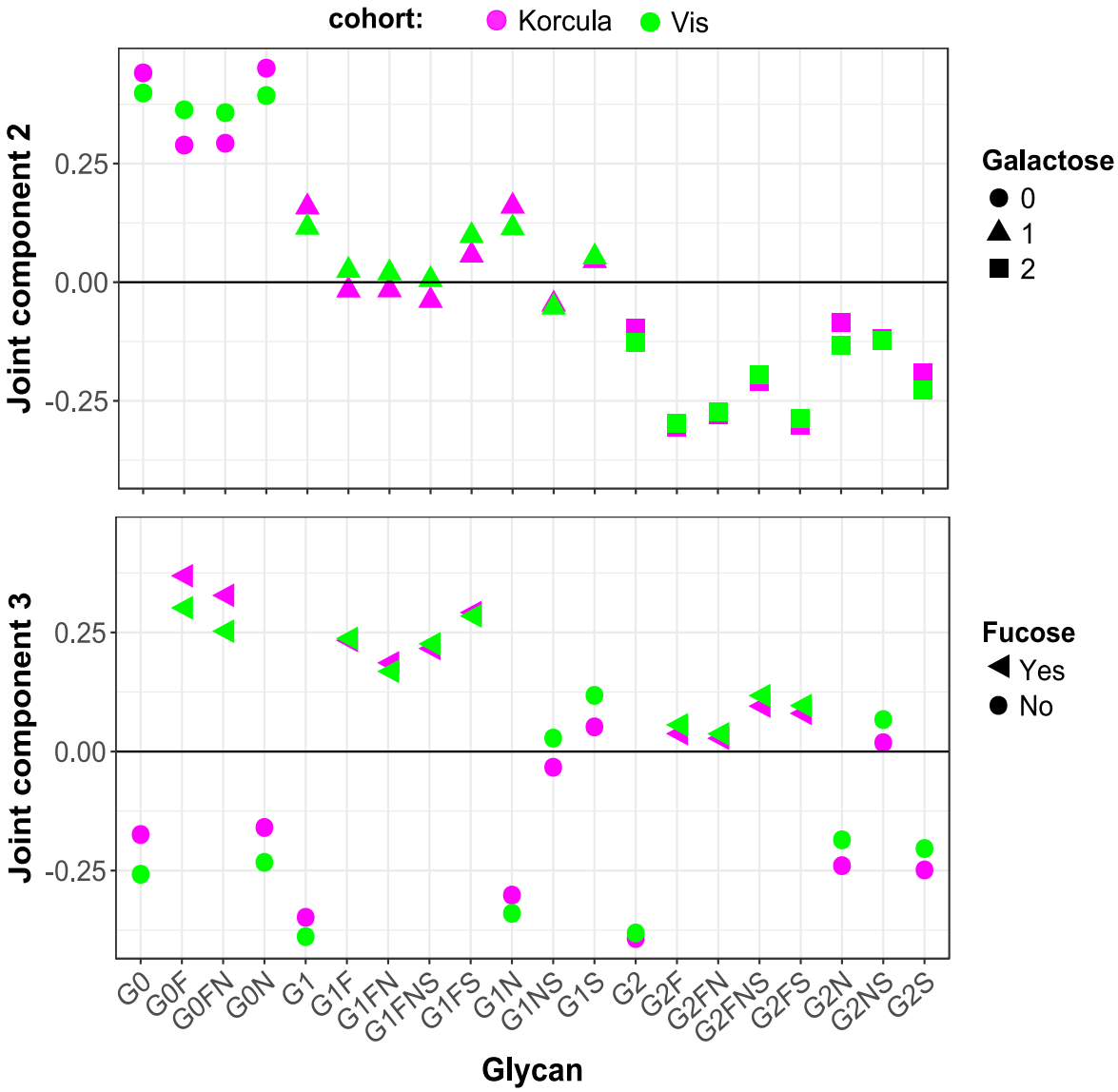
Conclusion Simulations

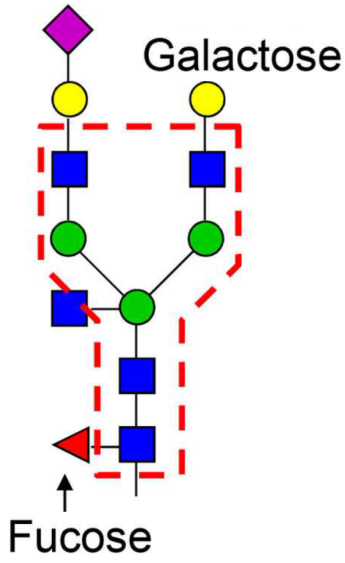
- TPR: PLS performs well in large datasets, PO2PLS performs well in small noisy datasets.
- In training set PLS and O2PLS overestimate

Application: IgG glycan datasets

- **Aim**: Integrate IgG1 and IgG2 glycan datasets
- Pre-processed IgG1 and IgG2 glycan abundances from two cohorts (Korcula and Vis)
- 20 variables in each dataset
- Sample sizes: 951 in Korcula and 756 in Vis
- 4 joint PPLS components retained: explain about 90% of data
- We present results for second and third component

Results: Loadings

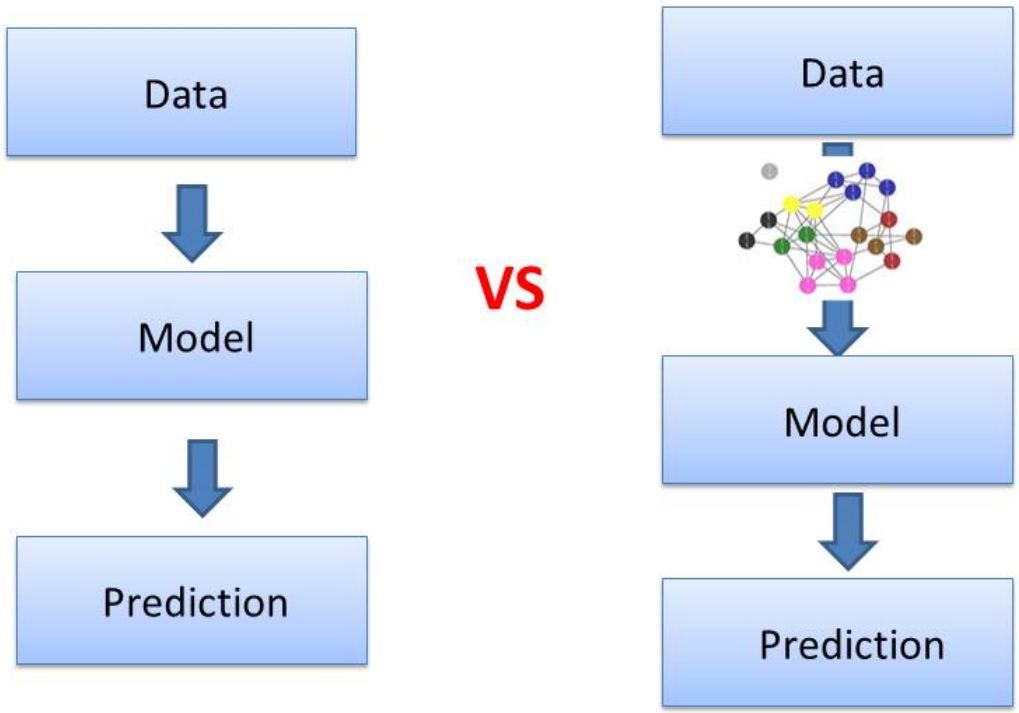




Galactose and Fucose:

- Reflect enzymatic reactions
- Associate with several disease statuses

Our proposal



Motivation

	WGCNA + Group Lasso			Lasso		
Variable	Average beta	Frequency	Cluster	Average beta	Frequency	Rank
GLOL	.064	10	1	.074	10	4
TYR	.060	10	1	.080	10	3
ALB	-.059	10	1	-.086	10	2
GLY	-.041	10	1	-.037	10	6
PHE	.038	10	1	.038	10	5
XSVLDLL	.038	10	2	.036	10	7
XLHDLL	-.038	10	3	-.089	9	1
HIS	-.036	10	1	-.024	9	8
SM	.034	10	2	.018	8	10
FAW6	.031	10	2	.017	7	12
GLC	.031	10	1	.018	10	11
SHDLL	.030	10	2	.003	3	20

Three steps, Tissier et al 2019 Plos One

- One Dataset (X)
- One Outcome (y)
- Network (X)
  - nodes: variables
  - edges: connection or association between variables
- Building networks: WGCNA based on correlation and penalizing edges by scale free topology: $w_{ij} = |cor(x_i, x_j)|^{\beta}$
- Obtaining L clusters via hierarchical clustering
- Prediction model(y) using Group Lasso

Group Lasso (Yuan, 2006)

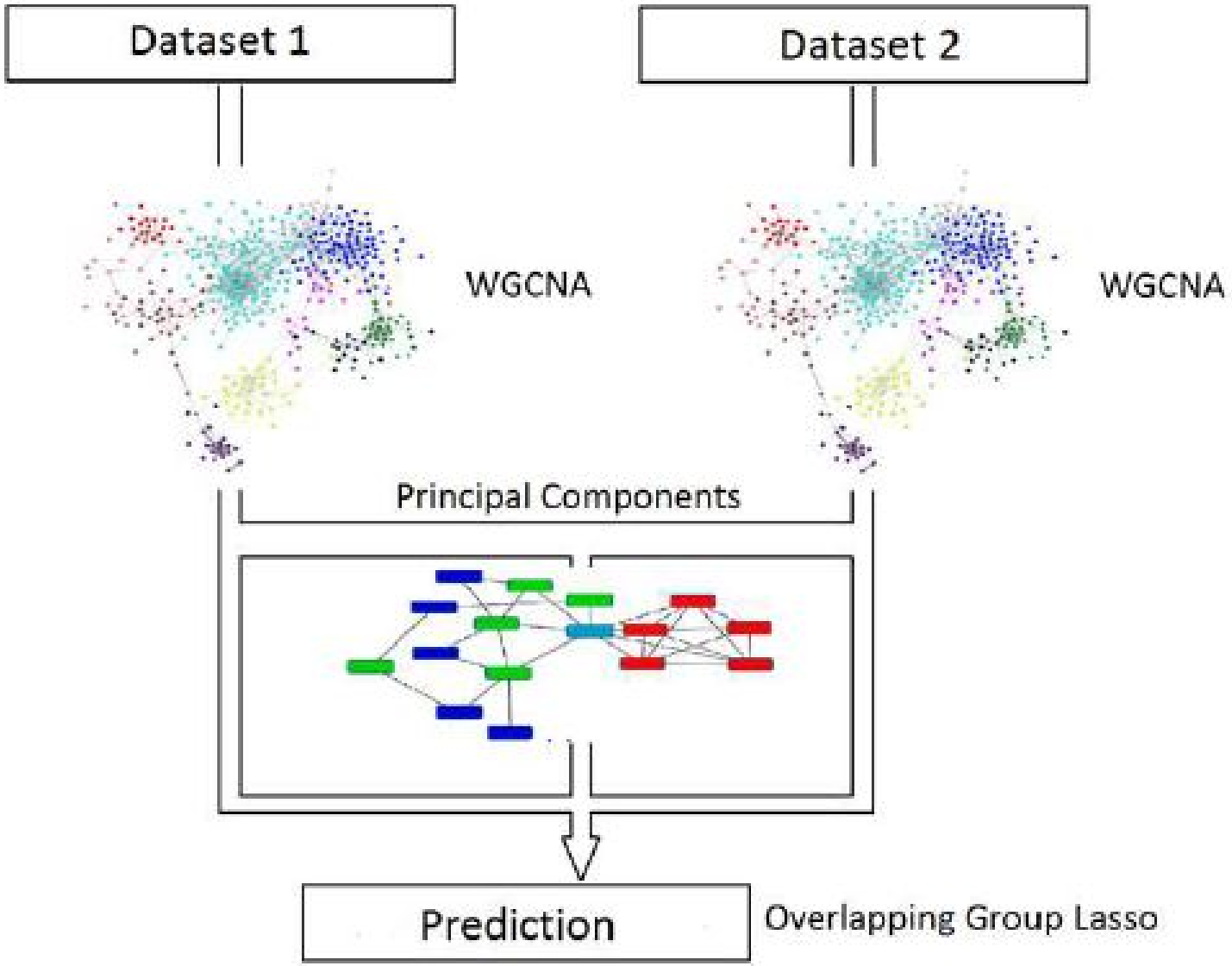
- Selects groups of variables
- All the coefficients belonging to the same pre-specified group are simultaneously shrunk towards zero .

$$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \left\| \beta_l \right\|_2 \right)$$

Results Simulation Study

- Simulated 4 clusters, vary number of predictors (200-1000), number of subjects (50-100), number of associated predictors (latent)
- Measure for prediction accuracy
  - Network methods: WGCNA obtains the correct number of clusters
  - Prediction models: Group Lasso performs well
  - Prediction performance similar (better) to standard approach Lasso
  - Group Lasso improves interpretation when variables are clustered

Two Omic datasets



Overlapping Lasso Jacob et al 2009

- X dimension r, Z dimension p two omics datasets
- Allows predictors to be part of several clusters
- $$\min_{\beta \in R^p} \left(\left\| y - \sum_{l=1}^{L} M\gamma_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \left\| \gamma_l \right\|_2 \right)$$

  with $\gamma_1, \cdots \gamma_L$ L L latent variables of dimension $p + r$ with $\gamma_{lm} = 0$ if variable m is not in group l
- Elements of correlated groups are included twice. Once in the combined group and once in the dataset specific group.

Simulation study

- X dimension $p = 100$ and Z dimension $r = 1000$
- X and Z correlated via a latent variable (second component of X is replaced by second component of Y)
- Outcome y was simulated from linear combination of elements of X and Y
- Noise was added to X and Y: Same noise structure, Different noise structure

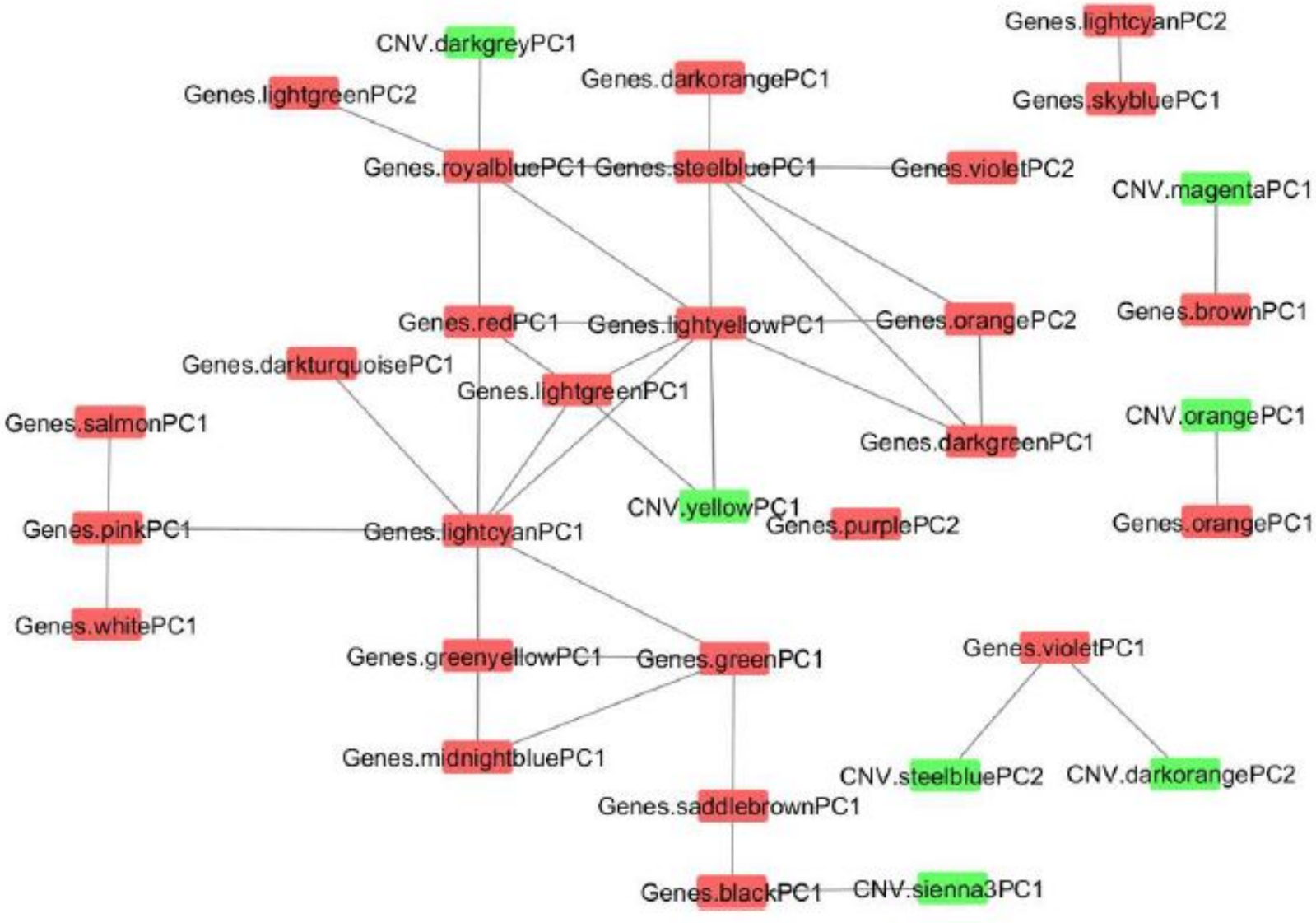
Results Simulation study

- Stacking is often no a good idea, especially when the noise structure is different
- Overlapping Lasso performs well if there is correlation.
- If the correlation is small, grouping the features in the two datasets separately and then consider all groups for predicting the outcome works well

Data application

Method	CNV	gene expression	prediction
Group Lasso	.476	.651	.504
OverlapLasso			.933
Lasso	.934	.571	.576
Ridge	.454	.610	.614

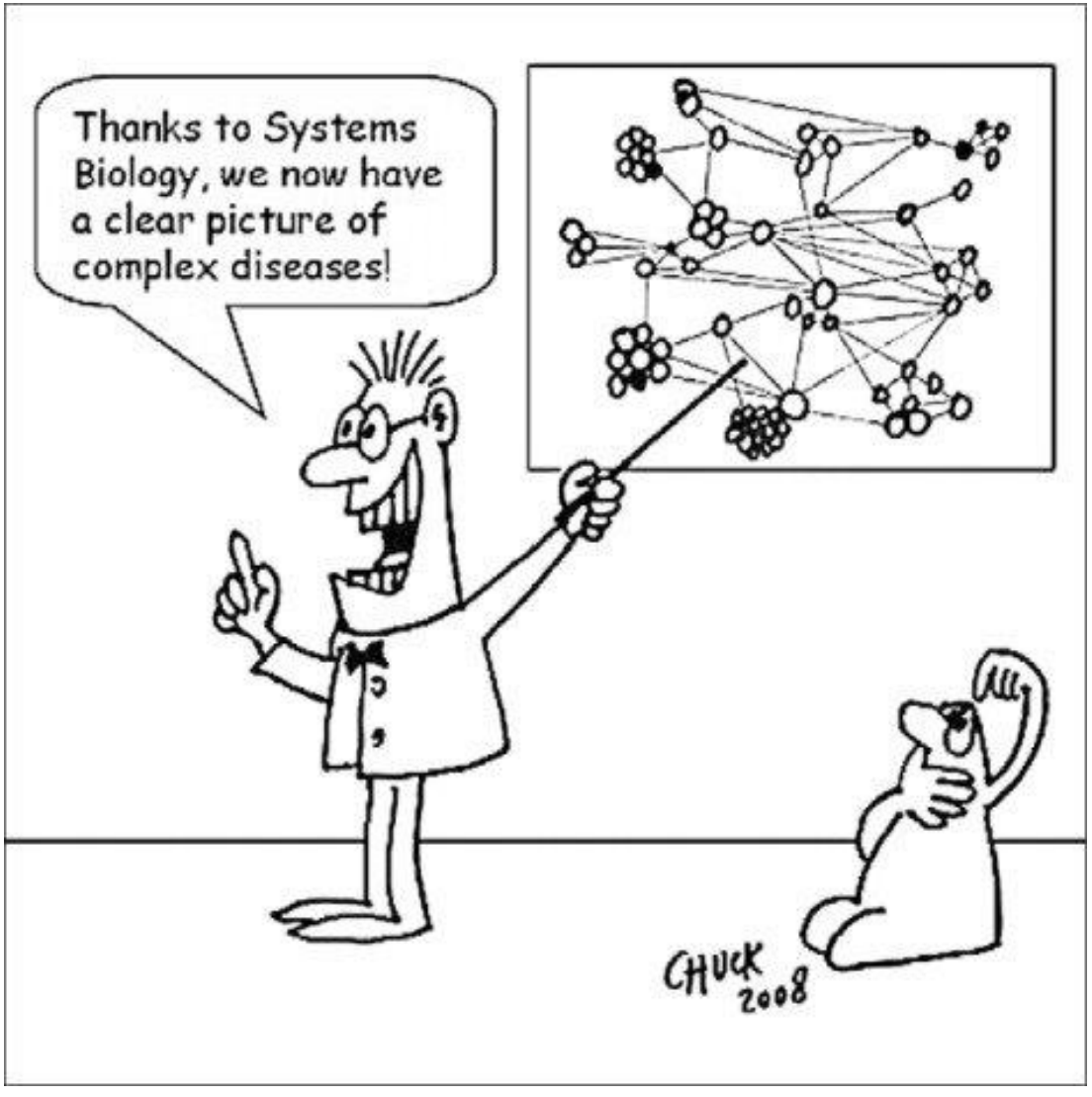
Results Prediction



Conclusion

- Data cleaning is an essential step
- PLS methods form useful set of tools for omic research
- Depending on type of data and research question, probabilistic versions are useful extensions
- Currently we are working on extensions to more than two datasets and to obtain sparse solutions
- Network methods useful to reduce dimensions
- More research needed in how to deal with multiple omics datasets

Making sense of data?



References

- Yuan M, Lin Y, 2006 Model selection and estimation in regression with grouped variables. Journal of Royal Stat Soc B. 68
- Jacob L et al, 2009, Group Lasso with overlap and Graph Lasso. ICML 09
- Tissier R et al, 2018, Improving stability of prediction models based on correlated omics data by using network approaches. Plos one 13.
- el Bouhaddani S et al, 2016, Evaluation of O2PLS in Omics data integration. BMC Bioinformatics 17(2)
- el Bouhaddani S et al, 2018 Probabilistic partial least squares model: Identifiability, estimation and application. Journal of Multivariate Analysis, 167
- Morris J and Baladandayuthapani V, 2017, Statistical contributions to bioinformatics: Design, modelling, structure learning and integration and discussion Houwing-Duistermaat JJ et al Statistical Modelling, 17 (4-5)
- Wold H, 1973, Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In: Multivar. Anal. III (Proc. Third Internat. Symp. Wright State Univ., Dayton, Ohio, 1972), Academic Press, New York