Identifying OMICs markers related to inflammation as measured by targeted proteins

Emilie Lambourg, Louis Fisher, Yalu Su

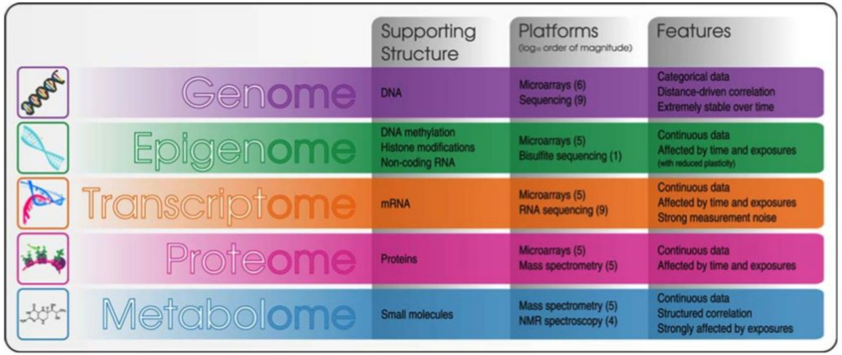
Overview

PEM data (Personal Monitoring Exposure):

- Part of the EXPOsOMICS project[1].
- Aims to explore the impact of high priority environmental pollutants.
- Air pollution -> chronic inflammation -> chronic diseases ?
- 150 healthy participants : measure of exposure during 24 hours (external exposome) + blood sample (internal exposome, which we focus on)
- Repeated measurement design : multiple sessions for one participant.

[1] The exposome in practice: Design of the EXPOsOMICS project. Vineis et al, 2016

The Dataset



	Supporting Structure	Platforms ($\log_{10}$ order of magnitude)	Features
Genome	DNA	Microarrays (6) Sequencing (9)	Categorical data Distance-driven correlation Extremely stable over time
Epigenome	DNA methylation Histone modifications Non-coding RNA	Microarrays (5) Bisulfite sequencing (1)	Continuous data Affected by time and exposures (with reduced plasticity)
Transcriptome	mRNA	Microarrays (5) RNA sequencing (9)	Continuous data Affected by time and exposures Strong measurement noise
Proteome	Proteins	Microarrays (5) Mass spectrometry (5)	Continuous data Affected by time and exposures
Metabolome	Small molecules	Mass spectrometry (5) NMR spectroscopy (4)	Continuous data Structured correlation Strongly affected by exposures

4 OMICs levels

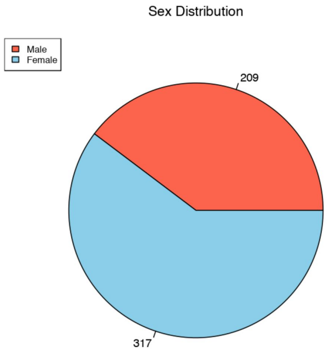
Size of Data

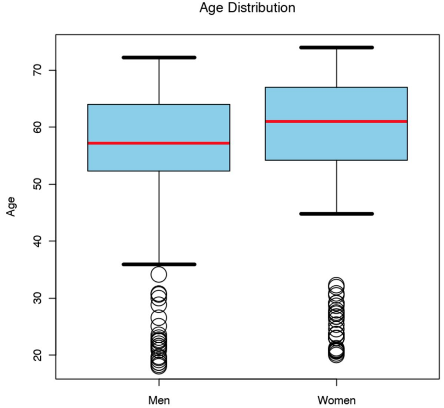
	Proteomics	Metabolomics	Transcriptomics	Epigenomics (Methylation)
Dimensions	336 X 13	400 X 11,217	227 X 23, 557	390 X 485, 512

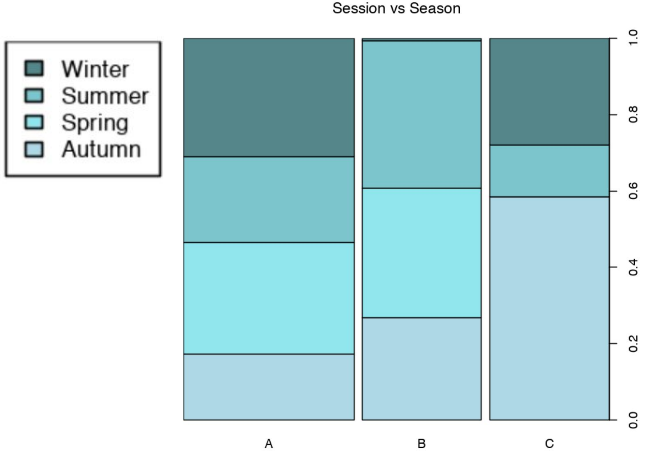
n<<p

- 19 covariates : technical (plate, chip…) and non-technical (age, gender, city…)
- Not everyone has every OMICs measurement.

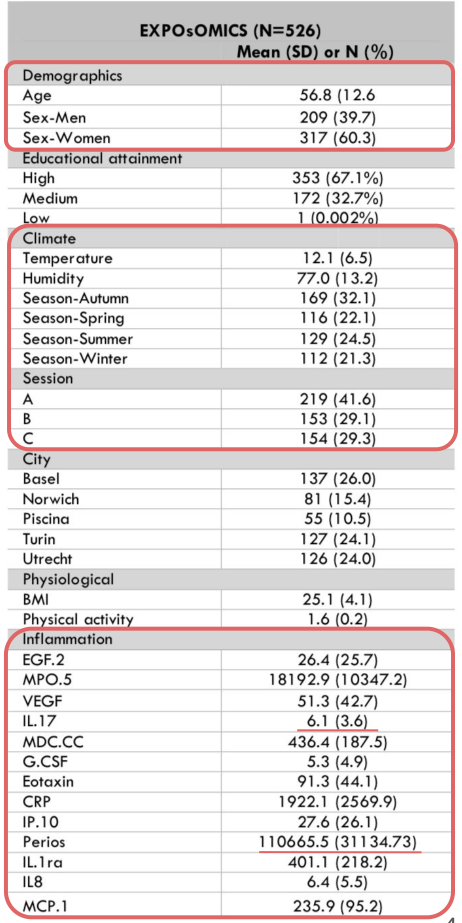
Exploratory Data Analysis







Wide range of protein intensities measured



	EXPOsOMICS (N=526) Mean (SD) or N (%)
Demographics	
Age	56.8 (12.6
Sex-Men	209 (39.7)
Sex-Women	317 (60.3)
Educational attainment	
High	353 (67.1%)
Medium	172 (32.7%)
Low	1 (0.002%)
Climate	
Temperature	12.1 (6.5)
Humidity	77.0 (13.2)
Season-Autumn	169 (32.1)
Season-Spring	116 (22.1)
Season-Summer	129 (24.5)
Season-Winter	112 (21.3)
Session	
A	219 (41.6)
B	153 (29.1)
C	154 (29.3)
City	
Basel	137 (26.0)
Norwich	81 (15.4)
Piscina	55 (10.5)
Turin	127 (24.1)
Utrecht	126 (24.0)
Physiological	
BMI	25.1 (4.1)
Physical activity	1.6 (0.2)
Inflammation	
EGF.2	26.4 (25.7)
MPO.5	18192.9 (10347.2)
VEGF	51.3 (42.7)
IL.17	6.1 (3.6)
MDC.CC	436.4 (187.5)
G.CSF	5.3 (4.9)
Eotaxin	91.3 (44.1)
CRP	1922.1 (2569.9)
IP.10	27.6 (26.1)
Perios	110665.5 (31134.73)
IL.1ra	401.1 (218.2)
IL8	6.4 (5.5)
MCP.1	235.9 (95.2)

Table 1. Descriptive statistics of the EXPOsOMICS dataset

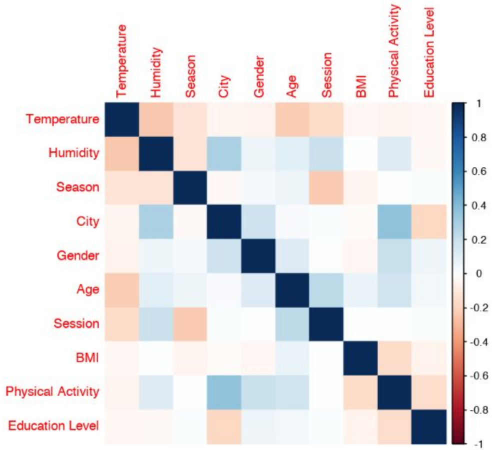
Pre-Processing

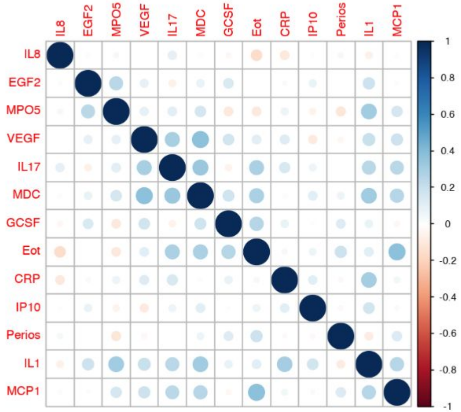
Covariates:

- Dropped the covariates with large proportions of missing values (temperature and humidity).
- Assess correlation within covariates.

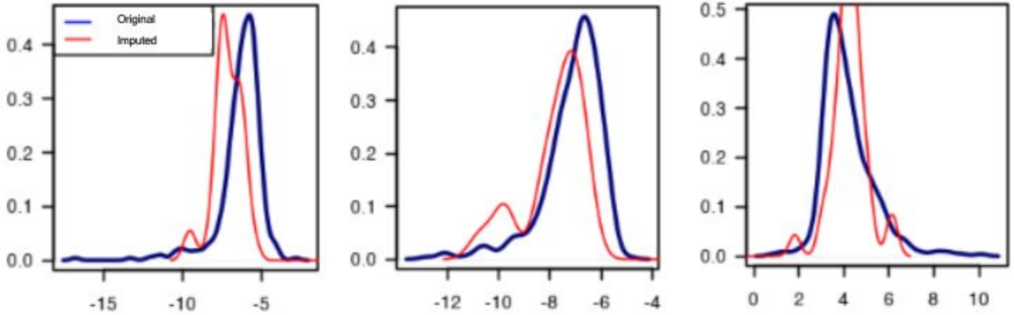
Proteins:

- Assess correlation between inflammatory proteins.
- No major correlation observed





Pre-Processing



Transcripts:

- No missing values.
- Already log-transformed.

Metabolites:

- Drop any metabolites with > 30% missing values.
- Imputed any remaining missing values using a quantile regression approach (favoured for left-censored MNAR data) [2].

Methylation:

- Dropped any methylation site with >10% missing values.
- Transformed beta values of methylation to M-values using logit-2 transformation (more statistically valid for differential methylation analysis)[3].
- Imputed any remaining missing values using k-nearest neighbours imputation.

[2] Missing Value Imputation Approach for Mass Spectrometry-Based Metabolomics Data, Wang et al 2018

[3] Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis, Du et al, 2010.

Pre-Processing

Data Denoising:

- Technical covariates exist for the measurement of each OMIC due to experimental variability:
  - Proteomics - Plate number.
  - Methylation - Chip number, chip position
  - Transcriptomics - Isolation date, labelling date and hybridisation date.
- Fit these as random effects in linear mixed models and carry out further analysis on residuals from these models.
- Formulation:$y = \alpha + X\beta + Zu + \varepsilon$
- Statistical model:

  proteins ~ (1 | plate) + (1 | id) + age + gender + bmi + season + city

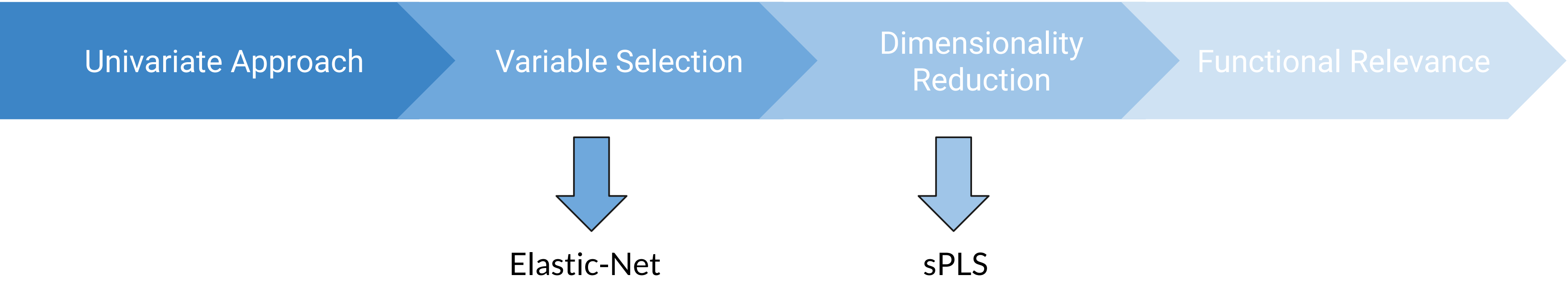
Aims

1. Explore the relationship between individual inflammatory proteins and individual transcriptomic, metabolomic and epigenomic features.

2. Identify a set of OMICs features that best predict inflammatory protein levels.

3. How do OMICs features jointly affect inflammatory protein levels?

4. Assess the functional relevance of any identified OMICs markers of inflammation.



Univariate Approach

Variable Selection

Dimensionality Reduction

Functional Relevance

Elastic-Net

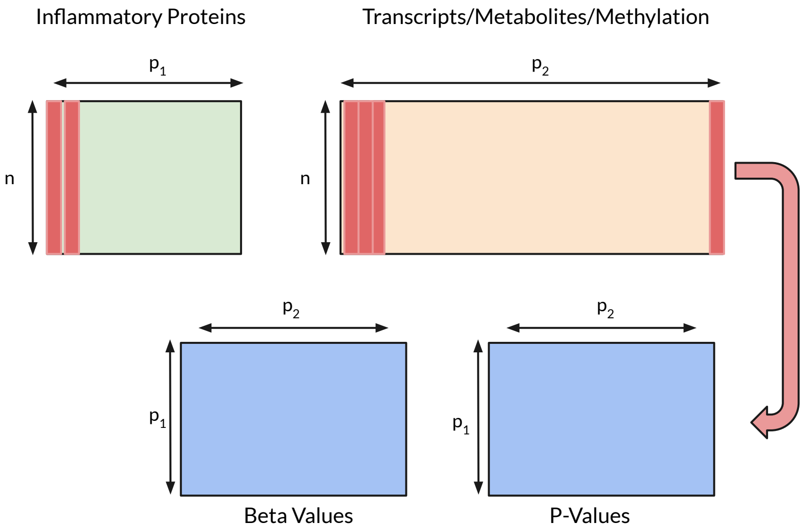
sPLS

Univariate Models

Aim: Explore relationship between individual protein and individual OMICs feature.

$$Y_{ij} = \alpha + \beta X_{ij} + \varepsilon_{ij}$$

Where:
Y_{ij} is the measurement levels of jth inflammatory protein
α is the intercept
β is the regression coefficient
X_{ij} is the observed value of jth alternative OMIC feature
ε_{ij} is the residual error measuring the random deviation from the linear relationship



Advantages:

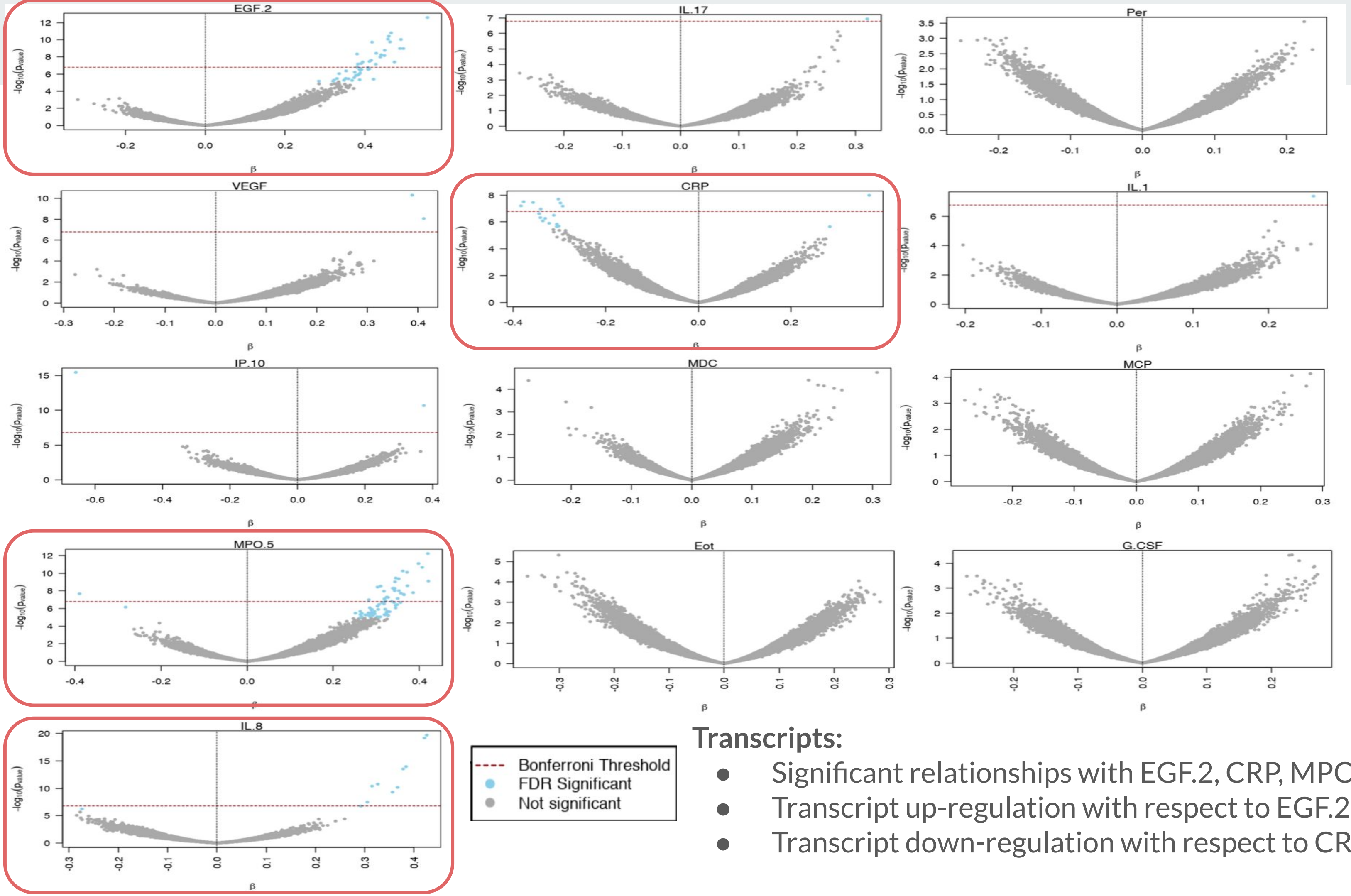
- Simple first exploration of relationships between inflammatory proteins and other OMICs.
- Efficient for exploring large p.
- Straightforward adjustment on confounders.

Disadvantages:

- Does not account for covariance structure within the data.
- Need to account for multiple testing during analysis.

Multiple Testing Correction:

- Run p_1 x p_2 tests.
- Large number of false positives.
- Account for using Bonferroni and Benjamini-Hochberg correction.



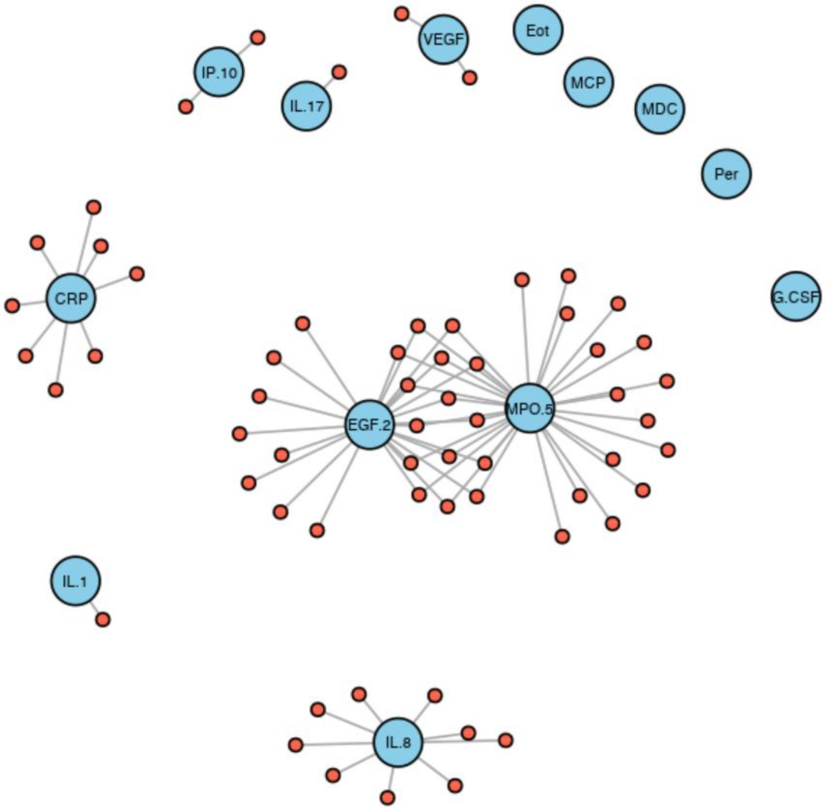
Transcripts:

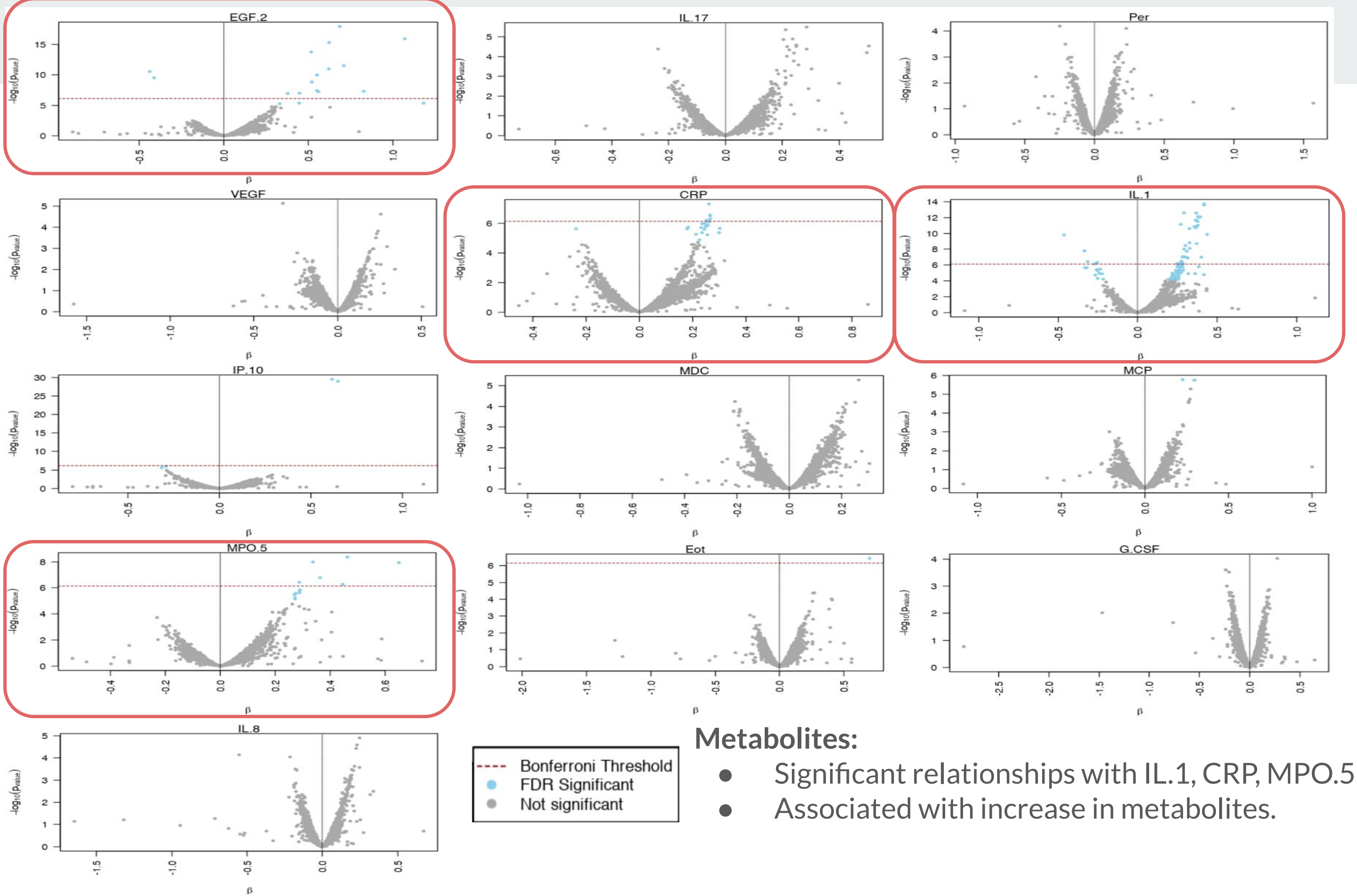
- Significant relationships with EGF.2, CRP, MPO.5 and IL.8.
- Transcript up-regulation with respect to EGF.2, MPO.5 and IL.8.
- Transcript down-regulation with respect to CRP.

Network Analysis

Transcripts:

- Significant relationships with 8/13 inflammatory proteins.
- EGF.2 and MPO.5 are the most significantly related.
- Overlap between EGF.2 and MPO.5.

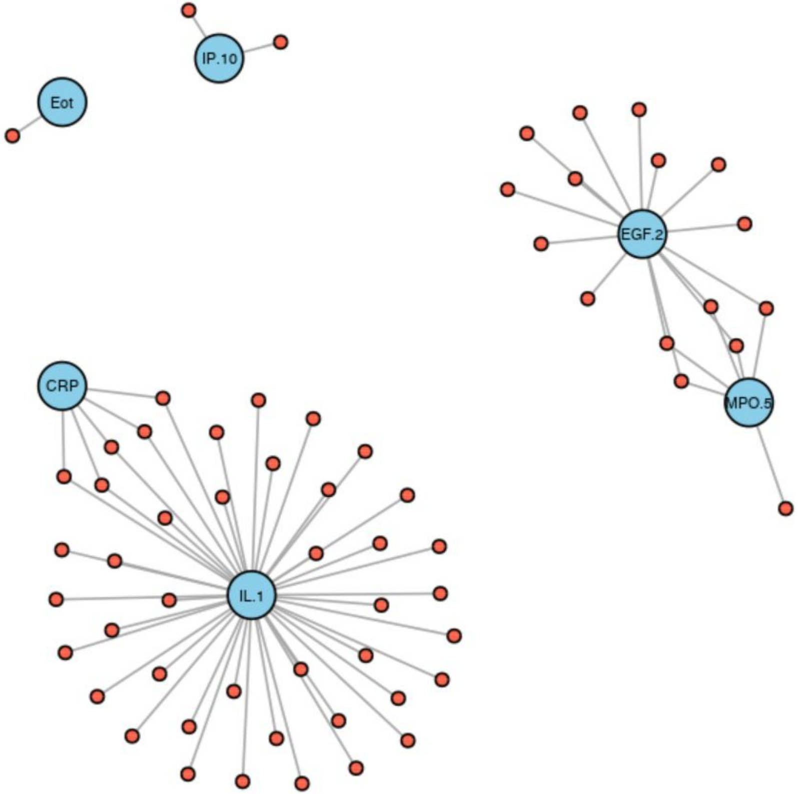




Metabolites:

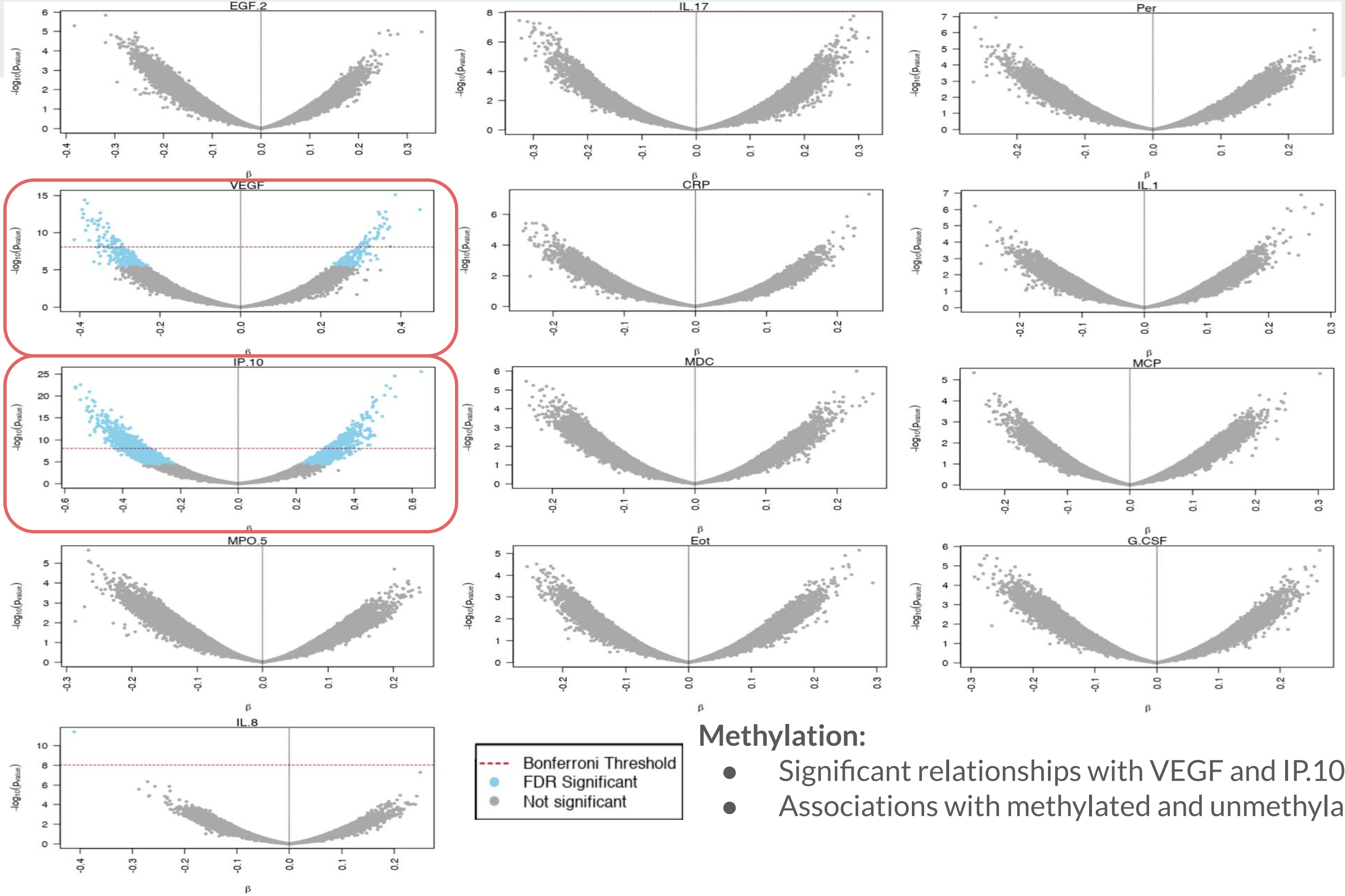
- Significant relationships with IL.1, CRP, MPO.5 and EGF.2.
- Associated with increase in metabolites.

Network Analysis



Metabolites:

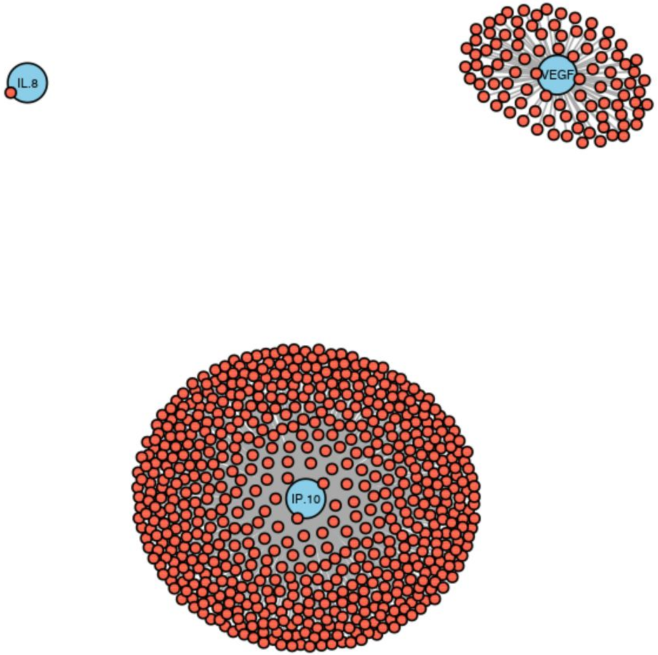
- Significant relationships with 6/13 inflammatory proteins.
- Overlap between IL.1 (many significant relationships) and CRP.
- Overlap between EGF.2 and MPO.5.



Methylation:

- Significant relationships with VEGF and IP.10.
- Associations with methylated and unmethylated sites.

Network Analysis

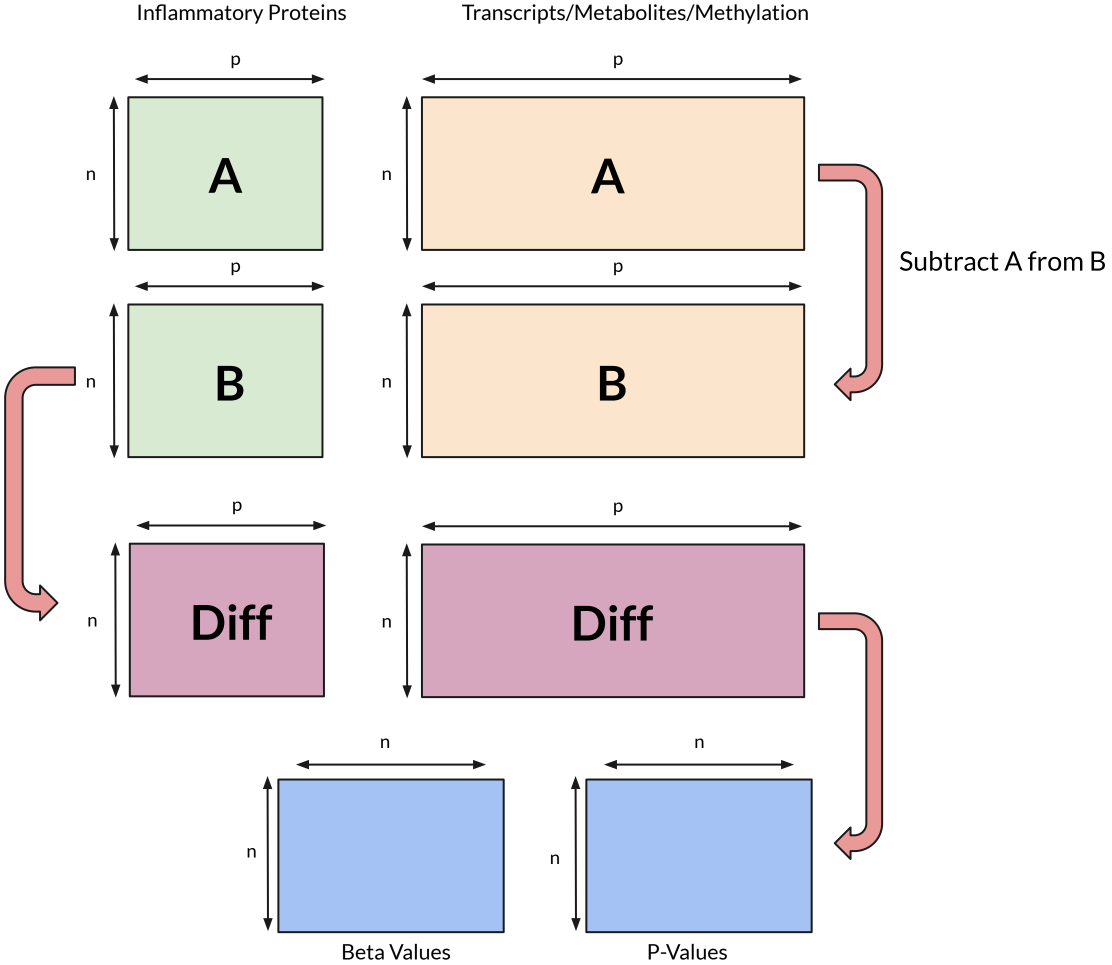


Methylation:

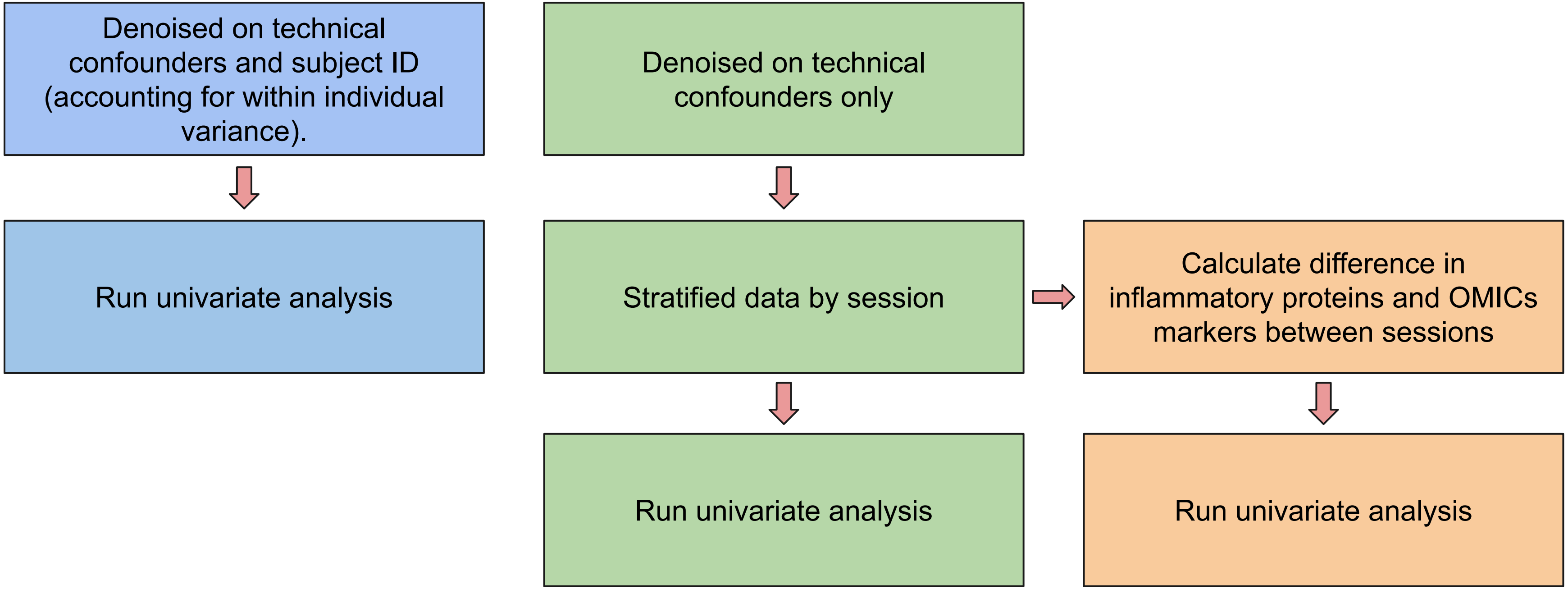
- Significant relationships with 3/13 inflammatory proteins.
- IP.10 has many significant relationships.
- No overlap.

Sensitivity Analysis

- Univariate models on data stratified by session.
- Regress difference in protein levels vs difference in individual OMICs features.
- Significant relationships strengthen previously identified relationships.
- Note: Lack of significant relationship does not negate the relationship.



Stratified Networks



Denoised on technical confounders and subject ID (accounting for within individual variance).

Run univariate analysis

Denoised on technical confounders only

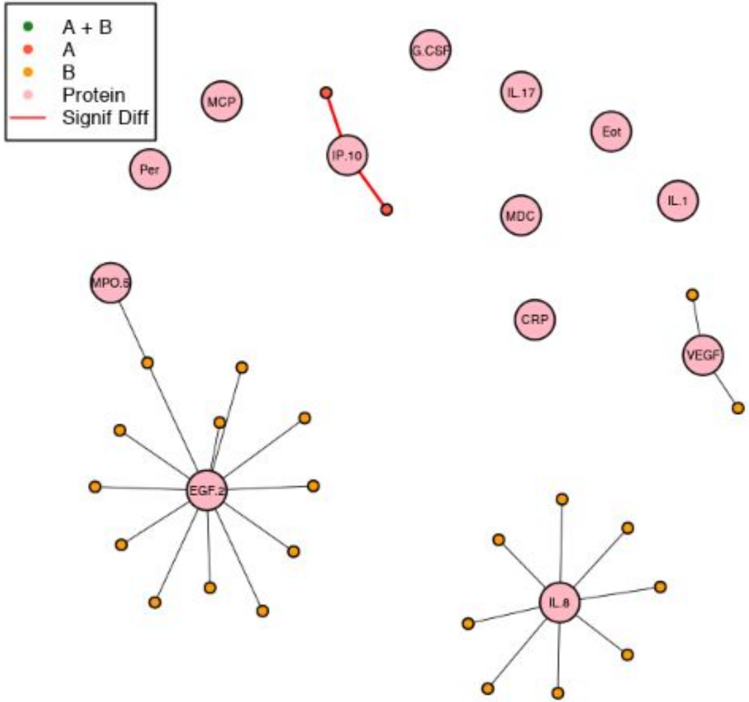
Stratified data by session

Calculate difference in inflammatory proteins and OMICs markers between sessions

Run univariate analysis

Run univariate analysis

Stratified Networks



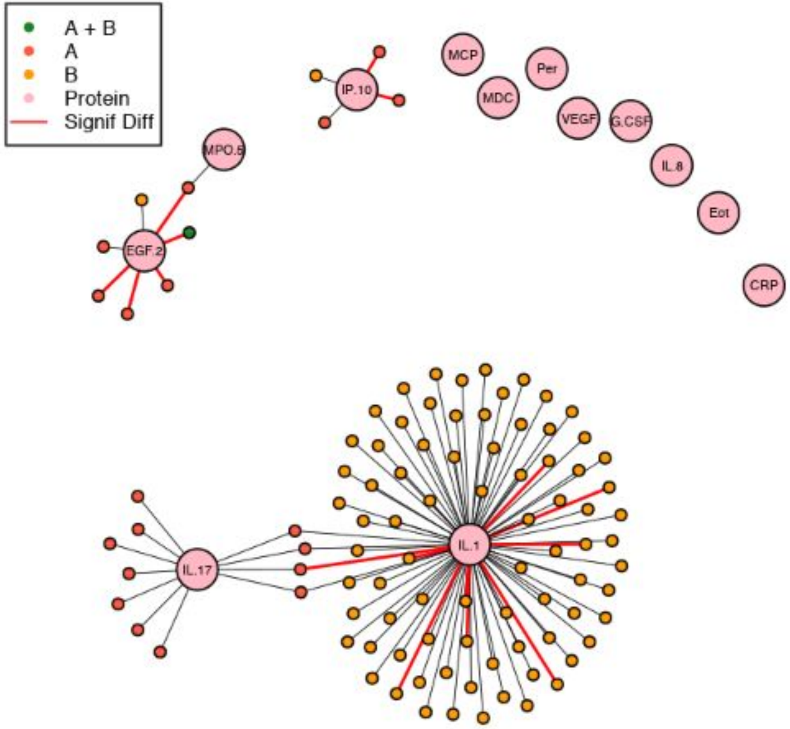
Transcripts:

- No associations found in both sessions.
- Small number of associations seen in session A.
- Associations with VEGF, IL8, EGF.2 and MPO.5 remain in session A.

Stratified Networks

Metabolites:

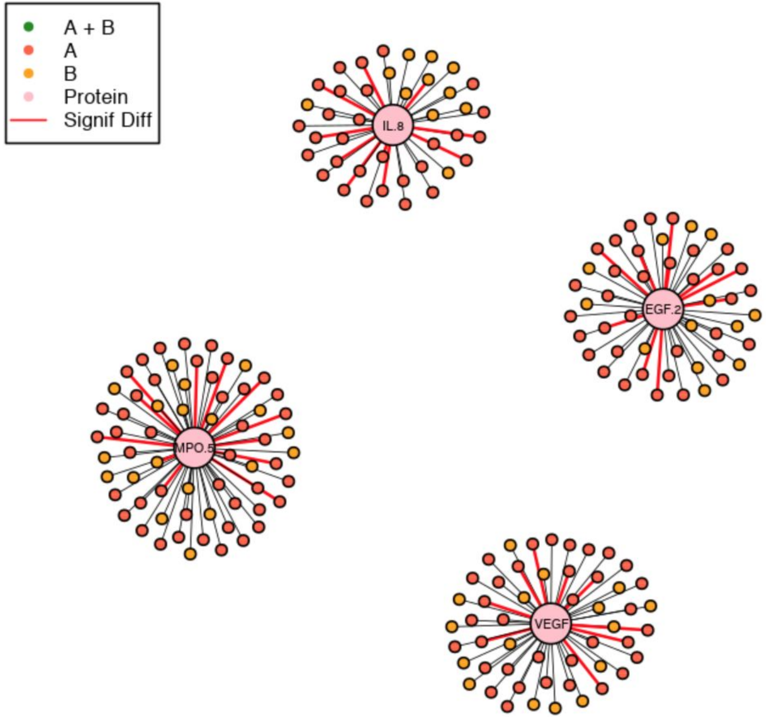
- 1 common association
- Large number of associations with IL.1 only found in session B.



Stratified Networks

Methylation:

- Associations found with 4 inflammatory proteins in both sessions.
- But, no common associations.



Penalised Regression

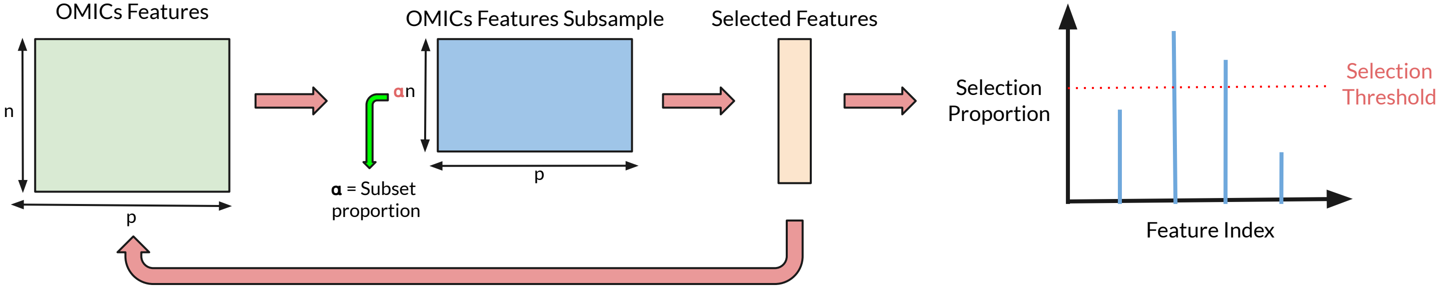
Aim: Identify a sparse set of OMICs features that best predict inflammatory protein levels

Advantages:

- Penalisation approaches impose sparsity on regression coefficients.
- Stable estimates of coefficients when p >n.
- Can use to select most informative predictors

Disadvantages:

- The max number of non-penalised variables is limited to the number of observations.
- Instability in variable selection - basis behind using stability selection approach[4].



e.g 500 iterations

[4] Stability Selection, Buhlmann P et al. 2010.

Elastic Net

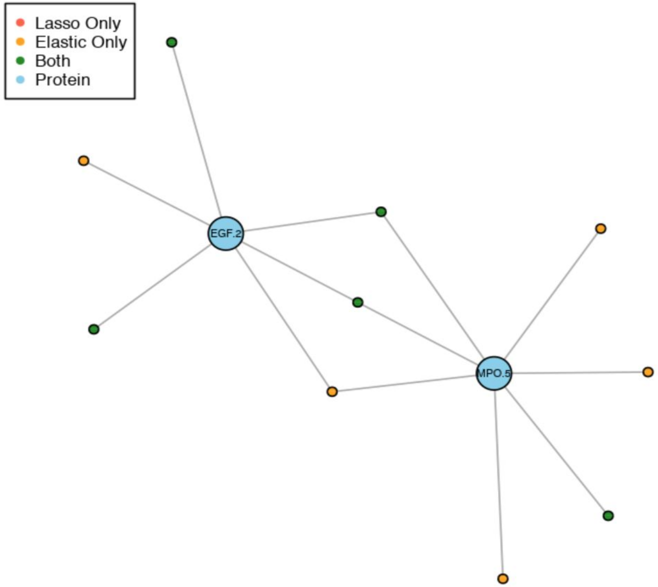
- Weighted sum of Lasso and Ridge.

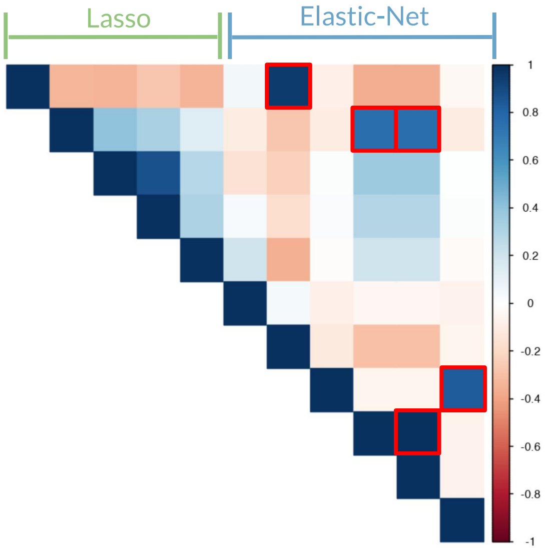
$$\lambda \sum_{j=1}^{p} (\lambda_0 \beta_j^2 + (1 - \lambda_0)|\beta_j|)$$

- Get numerical stability of Ridge and the sparsity of Lasso.
- When there is strong correlation between predictors (such as OMICs), Lasso may disregard significant predictors.

Motivating Example

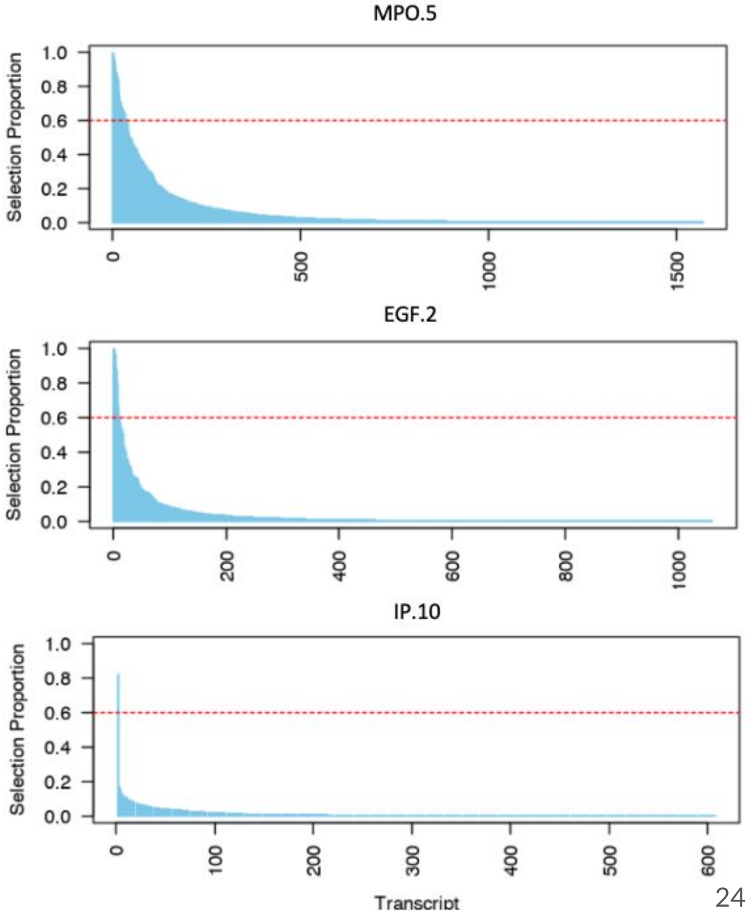
- Strong correlation structure exists in OMICs data.
- Lasso can disregard highly correlated predictors - can lead to loss of predictive power.



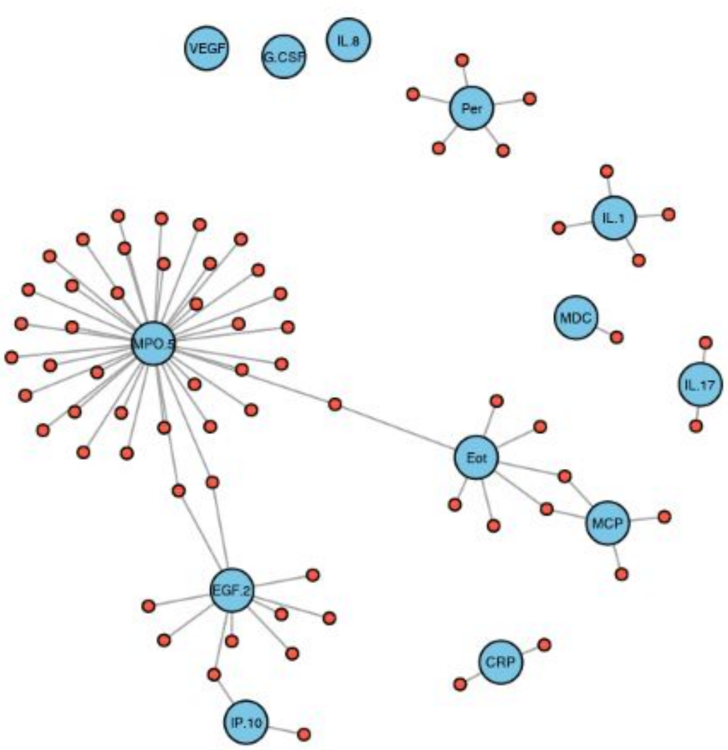


Metabolites

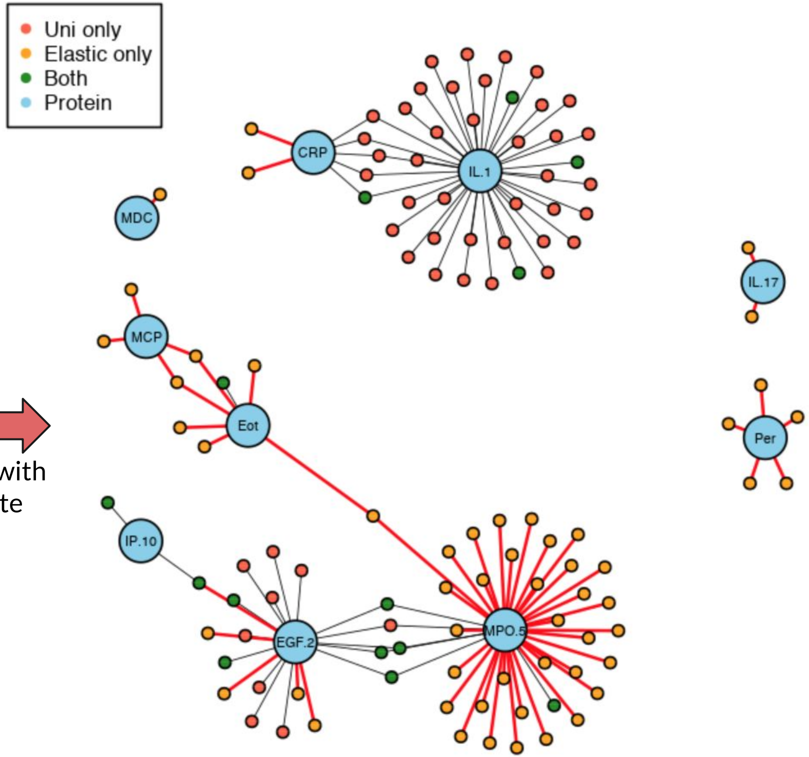
- Selection proportion threshold set to 60%.
- Significant metabolite associations found with 10/13 inflammatory proteins.
- Most associations seen with EGF2 and MPO5.



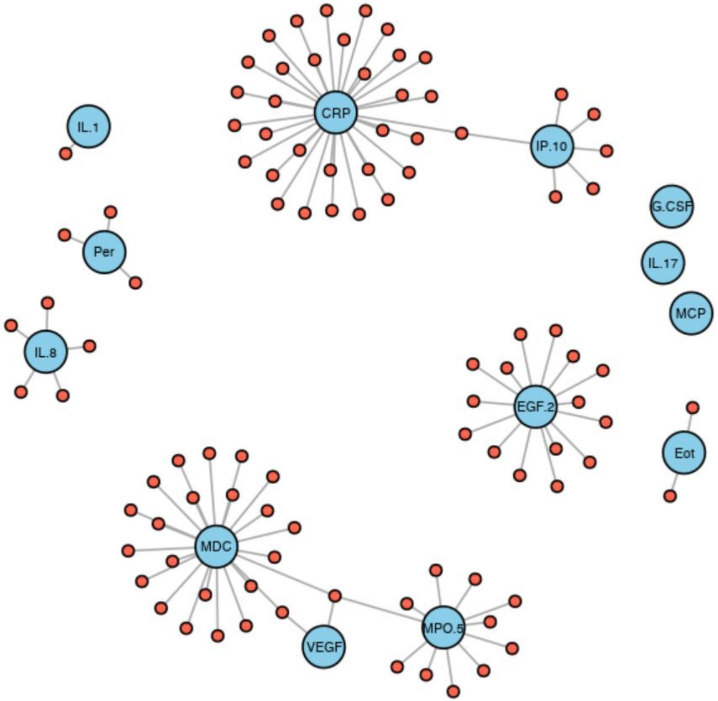
Metabolites



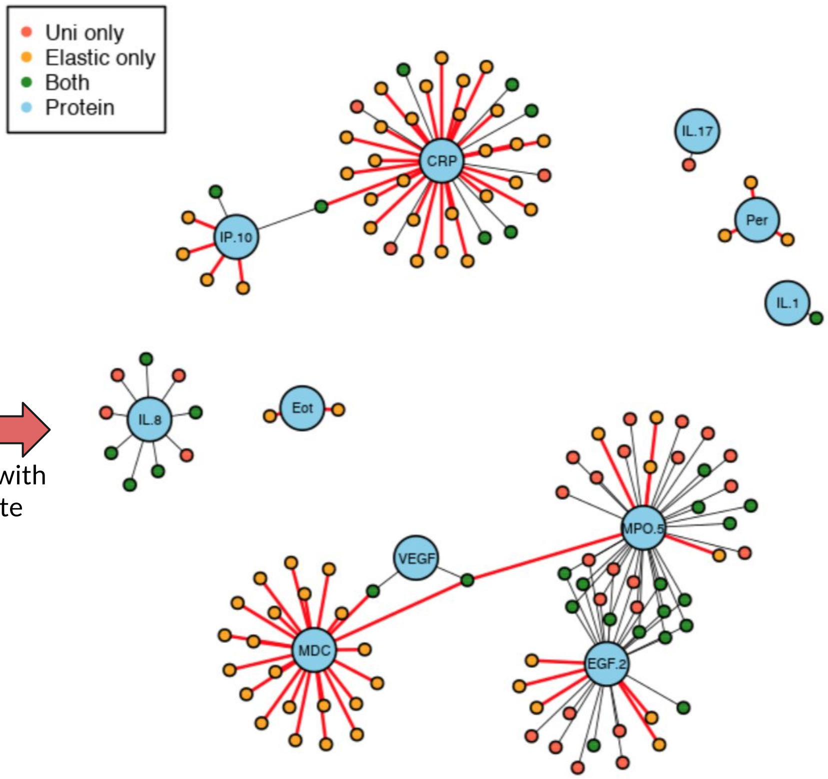
Integrate with univariate



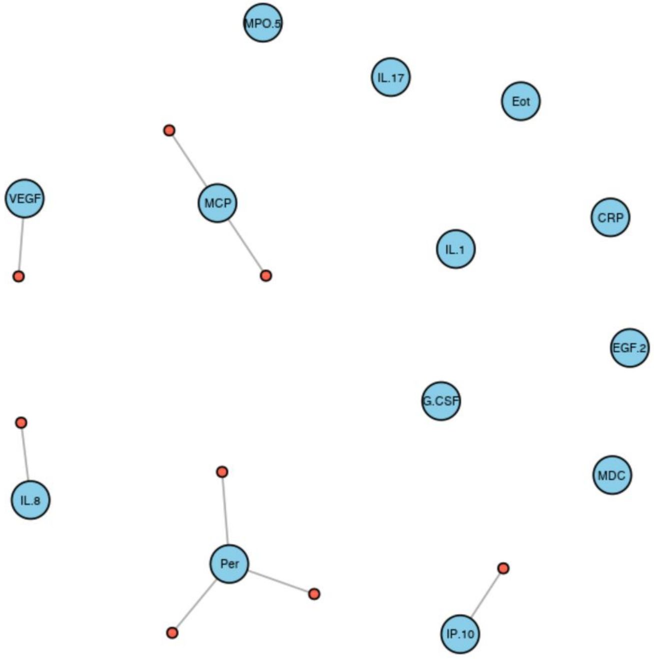
Transcripts



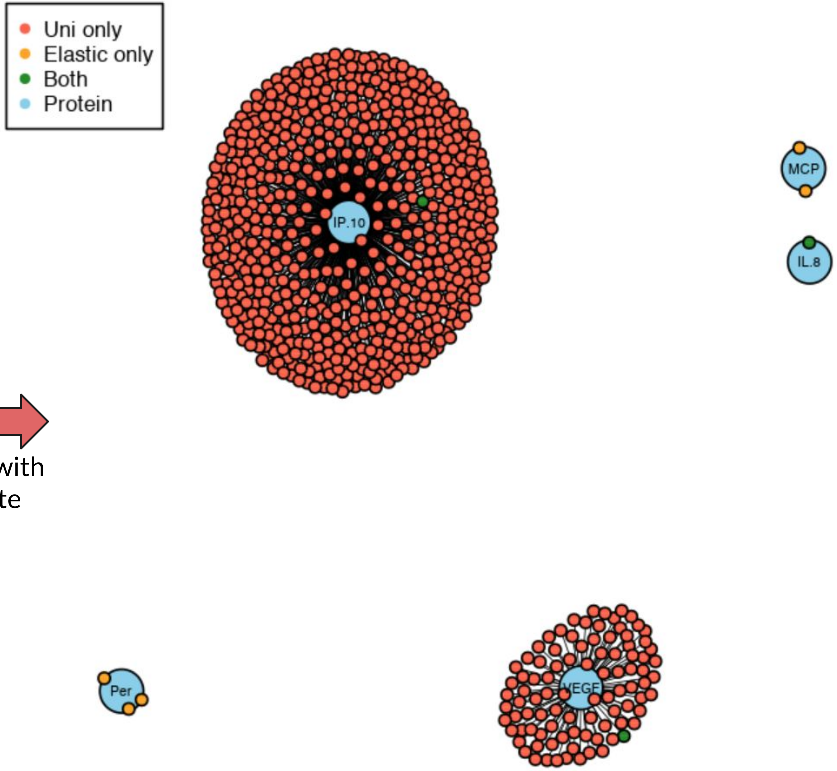
Integrate with univariate



Methylation



Integrate with univariate



Single sPLS

Aim: How do OMICs features jointly affect inflammatory protein levels?

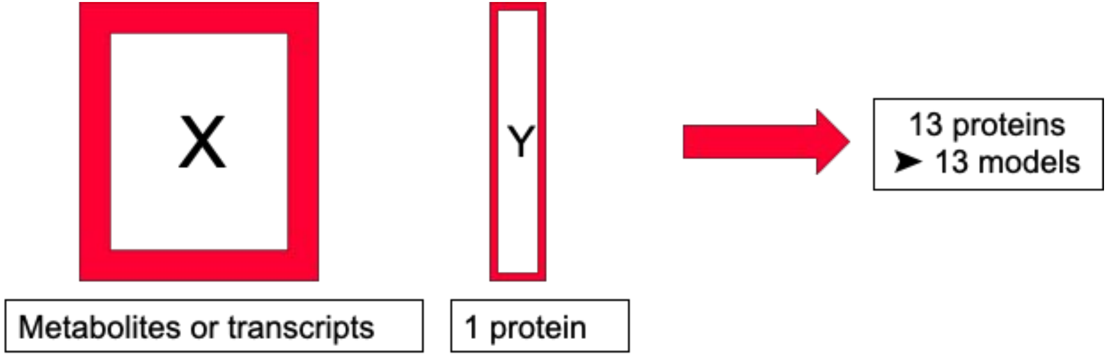
Why PLS ?

-to find inflammatory signatures -> need for a method that finds predictors **relevant to the outcome (inflammatory proteins)** and maximizes the variance in X AND Y

- can handle many noisy, collinear and missing variables

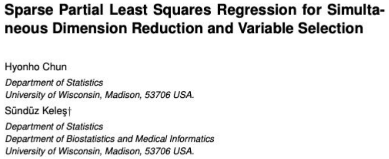
Why sparse PLS ?

- n<<p (PLS not suitable for very large p and small n (1)) , highly correlated -> need for sparsity
- increase interpretability



13 proteins
➤ 13 models

(1)



Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection

Hyonho Chun
Department of Statistics
University of Wisconsin, Madison, 53706 USA.
Sündüz Keleş†
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison, 53706 USA.

Compare sPLS with previous models

Not many links in common when comparing single spls with univariate linear models and elastic net

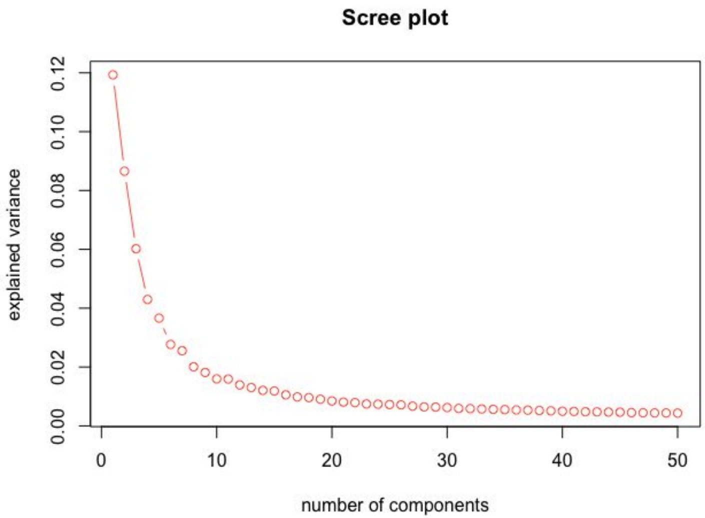
Hypothesis :

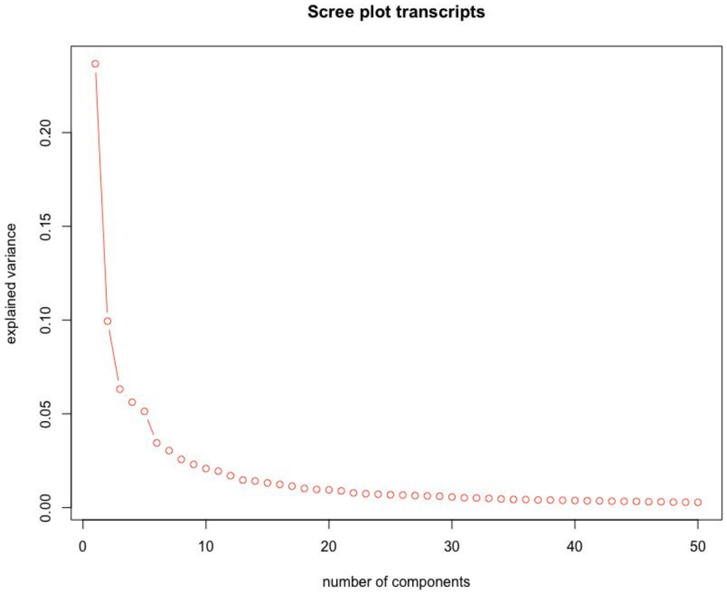
1.univariate linear models might not be good models as it misses the joint effects - so important in OMICs data.

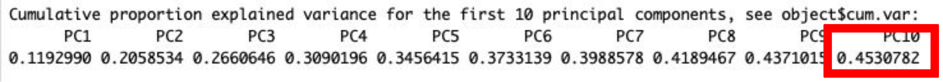
2.different denoising method

Multilevel PCA

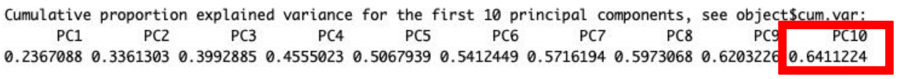
- Multilevel : before running the model, the “withinVariation” function decomposes the **within** from the **between** variance
- PCA applied on the within subject deviation matrix







```
Cumulative proportion explained variance for the first 10 principal components, see object$cum.var:
      PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8       PC9      PC10
0.1192990 0.2058534 0.2660646 0.3090196 0.3456415 0.3733139 0.3988578 0.4189467 0.4371015 0.4530782
```



```
Cumulative proportion explained variance for the first 10 principal components, see object$cum.var:
      PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8       PC9      PC10
0.2367088 0.3361303 0.3992885 0.4555023 0.5067939 0.5412449 0.5716194 0.5973068 0.6203226 0.6411224
```

10 components -> 45% of the variance explained

10 components -> 64% of variance explained

Multilevel sPLS - sparsity on X and Y

Use of *sPLS* function from *“mixOmics”* package

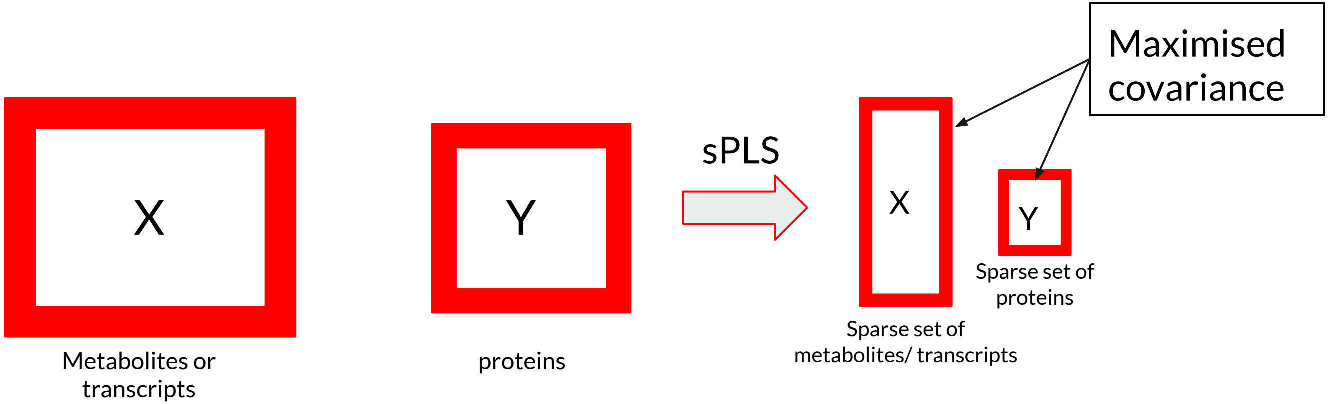
-Regression mode

- maximize the covariance between 2 matrices : metabolites or transcripts and the set of inflammatory proteins

-Variable selection through LASSO penalization on the pair of loading vectors

- respect of the repeated measurement design of the study

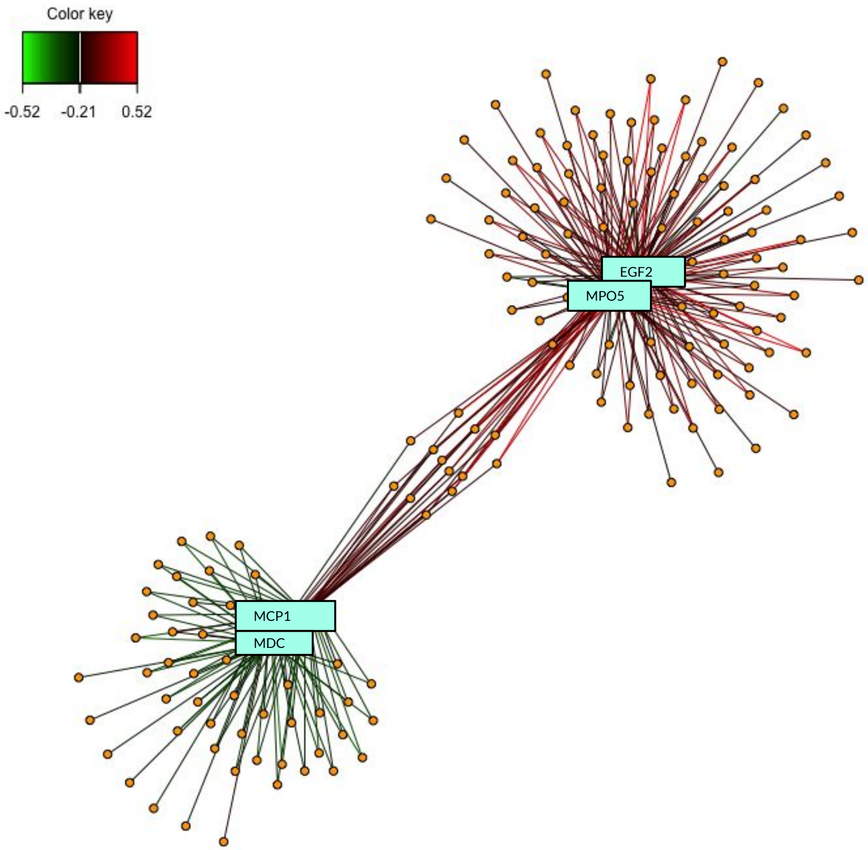
- attempt to predict the metabolites/transcripts selected with respect to the chosen set of inflammatory proteins


X
Metabolites or transcripts
Y
proteins
sPLS
X
Sparse set of metabolites/ transcripts
Y
Sparse set of proteins
Maximised covariance

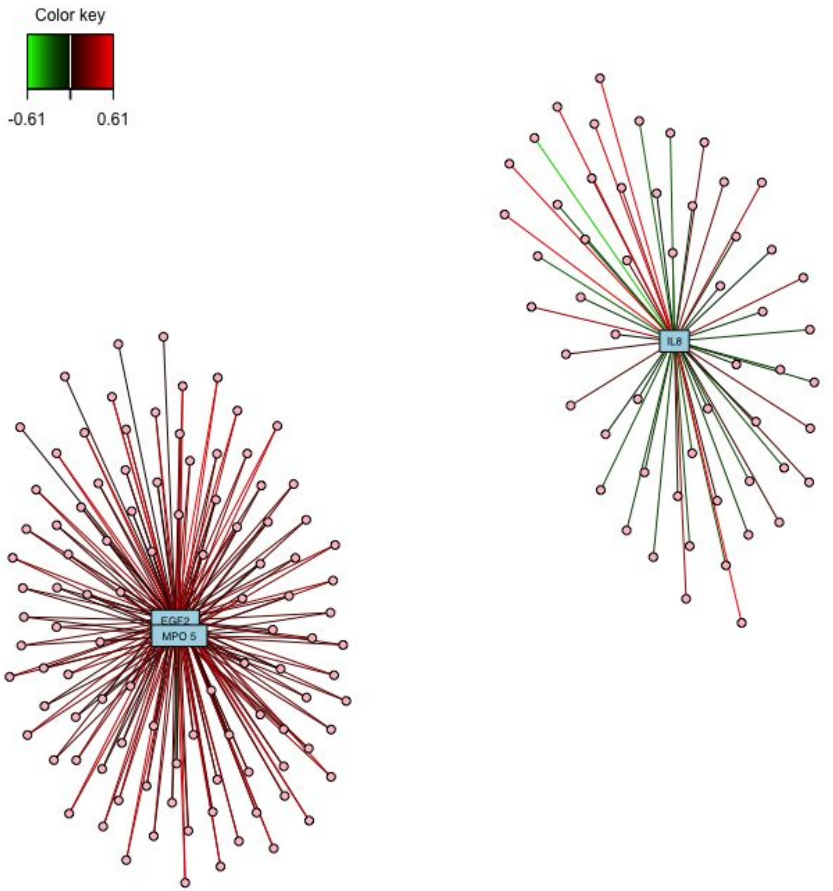
Component 1 : 2 proteins selected, 210 metabolites selected
Component 2 : 2 proteins selected, 103 metabolites selected

http://mixomics.org/methods/spls/

Multilevel spls metabolites / proteins



Multilevel spls transcripts / proteins



Component 1 : 131 transcripts and 3 proteins selected

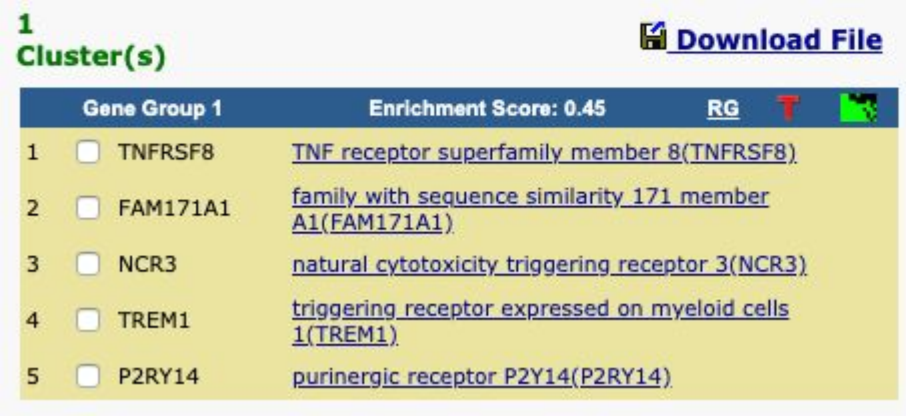
Component 2 : 71 transcripts and 3 proteins selected (same ones as component 1)

2 of the 3 proteins selected were also selected by spls metabolites/proteins : **EGF2 and MPO5**

Functional Interpretation

Which biological terms/functions are specifically enriched in the list of significant transcripts?

What are the major gene functional groups in the list of selected transcripts?

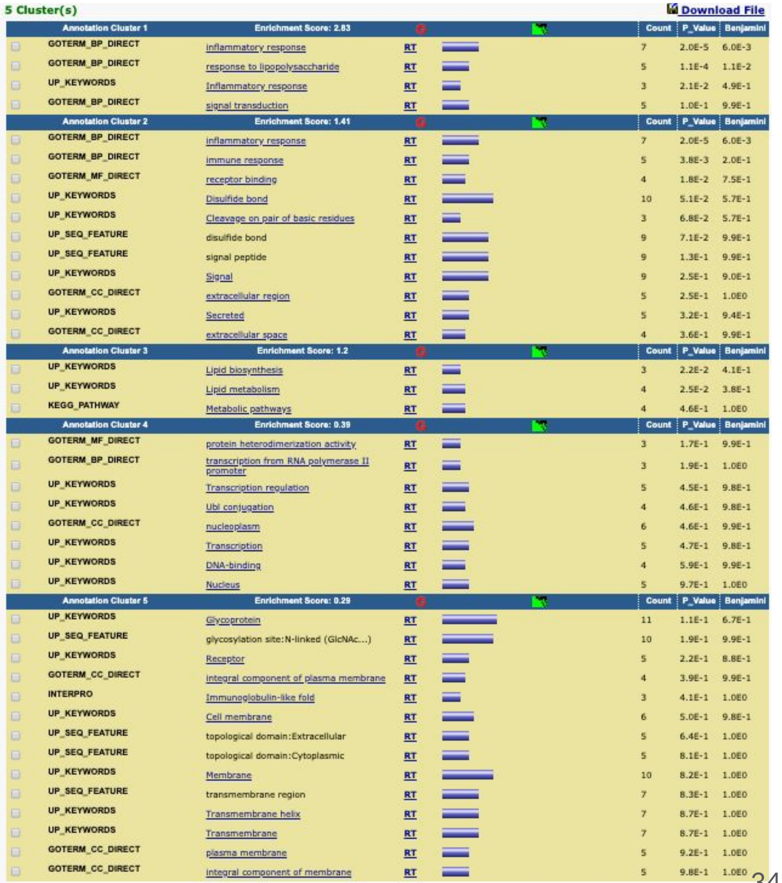


1 Cluster(s)

Download File

		Gene Group 1	Enrichment Score: 0.45
1	☐	TNFRSF8	TNF receptor superfamily member 8(TNFRSF8)
2	☐	FAM171A1	family with sequence similarity 171 member A1(FAM171A1)
3	☐	NCR3	natural cytotoxicity triggering receptor 3(NCR3)
4	☐	TREM1	triggering receptor expressed on myeloid cells 1(TREM1)
5	☐	P2RY14	purinergic receptor P2Y14(P2RY14)

RG



5 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 2.83		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	inflammatory response	RT	7	2.0E-5	6.0E-3
☐	GOTERM_BP_DIRECT	response to lipopolysaccharide	RT	5	1.1E-4	1.1E-2
☐	UP_KEYWORDS	Inflammatory response	RT	3	2.1E-2	4.9E-1
☐	GOTERM_BP_DIRECT	signal transduction	RT	5	1.0E-1	9.9E-1
	Annotation Cluster 2	Enrichment Score: 1.41		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	inflammatory response	RT	7	2.0E-5	6.0E-3
☐	GOTERM_BP_DIRECT	immune response	RT	5	3.8E-3	2.0E-1
☐	GOTERM_MF_DIRECT	receptor binding	RT	4	1.8E-2	7.5E-1
☐	UP_KEYWORDS	Disulfide bond	RT	10	5.1E-2	5.7E-1
☐	UP_KEYWORDS	Cleavage on pair of basic residues	RT	3	6.8E-2	5.7E-1
☐	UP_SEQ_FEATURE	disulfide bond	RT	9	7.1E-2	9.9E-1
☐	UP_SEQ_FEATURE	signal peptide	RT	9	1.3E-1	9.9E-1
☐	UP_KEYWORDS	Signal	RT	9	2.5E-1	9.0E-1
☐	GOTERM_CC_DIRECT	extracellular region	RT	5	2.5E-1	1.0E0
☐	UP_KEYWORDS	Secreted	RT	5	3.2E-1	9.4E-1
☐	GOTERM_CC_DIRECT	extracellular space	RT	4	3.6E-1	9.9E-1
	Annotation Cluster 3	Enrichment Score: 1.2		Count	P_Value	Benjamini
☐	UP_KEYWORDS	Lipid biosynthesis	RT	3	2.2E-2	4.1E-1
☐	UP_KEYWORDS	Lipid metabolism	RT	4	2.5E-2	3.8E-1
☐	KEGG_PATHWAY	Metabolic pathways	RT	4	4.6E-1	1.0E0
	Annotation Cluster 4	Enrichment Score: 0.39		Count	P_Value	Benjamini
☐	GOTERM_MF_DIRECT	protein heterodimerization activity	RT	3	1.7E-1	9.9E-1
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT	3	1.9E-1	1.0E0
☐	UP_KEYWORDS	Transcription regulation	RT	5	4.5E-1	9.8E-1
☐	UP_KEYWORDS	Ubl conjugation	RT	4	4.6E-1	9.8E-1
☐	GOTERM_CC_DIRECT	nucleoplasm	RT	6	4.6E-1	9.9E-1
☐	UP_KEYWORDS	Transcription	RT	5	4.7E-1	9.8E-1
☐	UP_KEYWORDS	DNA-binding	RT	4	5.9E-1	9.9E-1
☐	UP_KEYWORDS	Nucleus	RT	5	9.7E-1	1.0E0
	Annotation Cluster 5	Enrichment Score: 0.29		Count	P_Value	Benjamini
☐	UP_KEYWORDS	Glycoprotein	RT	11	1.1E-1	6.7E-1
☐	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT	10	1.9E-1	9.9E-1
☐	UP_KEYWORDS	Receptor	RT	5	2.2E-1	8.8E-1
☐	GOTERM_CC_DIRECT	integral component of plasma membrane	RT	4	3.9E-1	9.9E-1
☐	INTERPRO	Immunoglobulin-like fold	RT	3	4.1E-1	1.0E0
☐	UP_KEYWORDS	Cell membrane	RT	6	5.0E-1	9.8E-1
☐	UP_SEQ_FEATURE	topological domain:Extracellular	RT	5	6.4E-1	1.0E0
☐	UP_SEQ_FEATURE	topological domain:Cytoplasmic	RT	5	8.1E-1	1.0E0
☐	UP_KEYWORDS	Membrane	RT	10	8.2E-1	1.0E0
☐	UP_SEQ_FEATURE	transmembrane region	RT	7	8.3E-1	1.0E0
☐	UP_KEYWORDS	Transmembrane helix	RT	7	8.7E-1	1.0E0
☐	UP_KEYWORDS	Transmembrane	RT	7	8.7E-1	1.0E0
☐	GOTERM_CC_DIRECT	plasma membrane	RT	5	9.2E-1	1.0E0
☐	GOTERM_CC_DIRECT	integral component of membrane	RT	5	9.8E-1	1.0E0