OMICs analyses Introduction and Refresher: models and examples

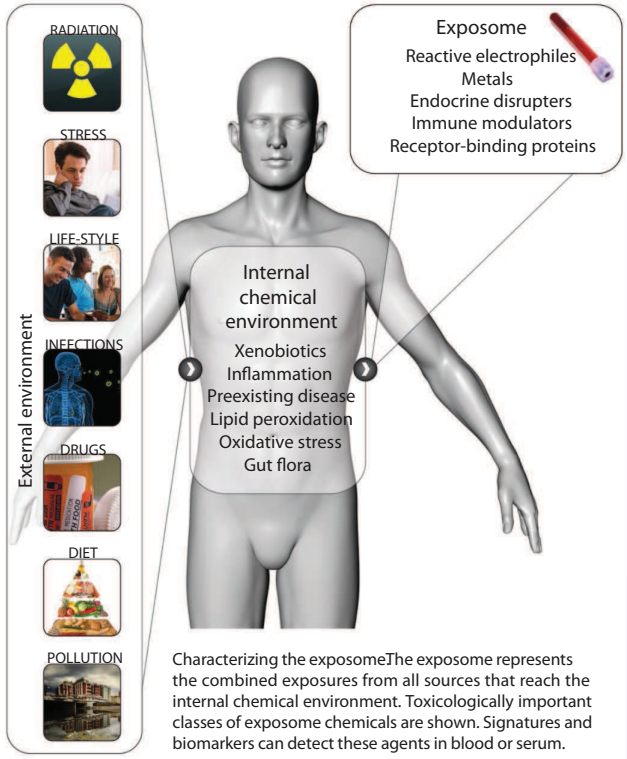
Surf64 – 25 June, 2018

Marc Chadeau-Hyam

m.chadeau@imperial.ac.uk

A practice-driven definition of the Exposome

Rappaport S, Smith M, Science (2010): 'The human exposome includes combined exposures reaching the internal chemical environment'



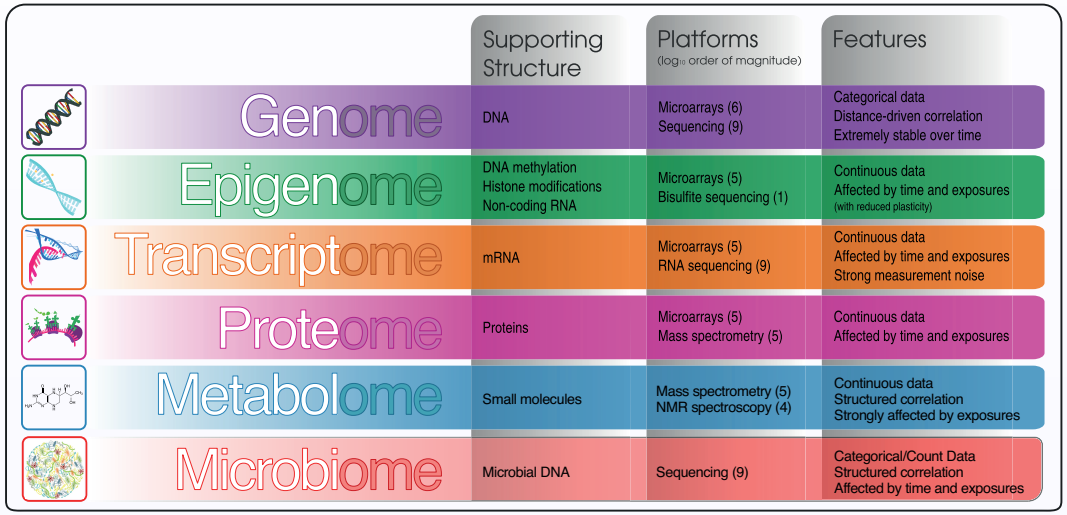
Characterizing the exposomeThe exposome represents the combined exposures from all sources that reach the internal chemical environment. Toxicologically important classes of exposome chemicals are shown. Signatures and biomarkers can detect these agents in blood or serum.

- Assumption: any effective exposure should be detected in the internal environment
- Exposome describes the human internal environment
- It captures exogenous interactions with the environment
- The internal environment comprises responses to external and internal stressors

$\Rightarrow$ how to characterise the exposome?

Deriving -OMICs biomarkers to characterise the exposome

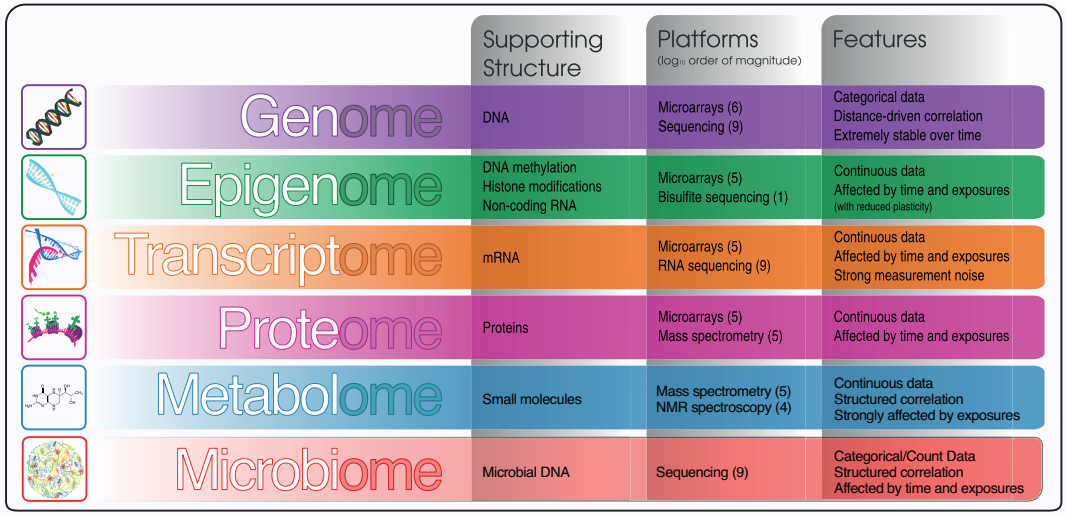
- OMICs data: high throughput biochemical measures of the abundance and/or structural features of molecules
- OMICs profiles: heterogeneous and complementary data



  - Ordered set of molecules along cellular pathways
  - Large range of molecules involved
  - Interacting molecules in the cellular activity and its regulation

Deriving -OMICs biomarkers to characterise the exposome

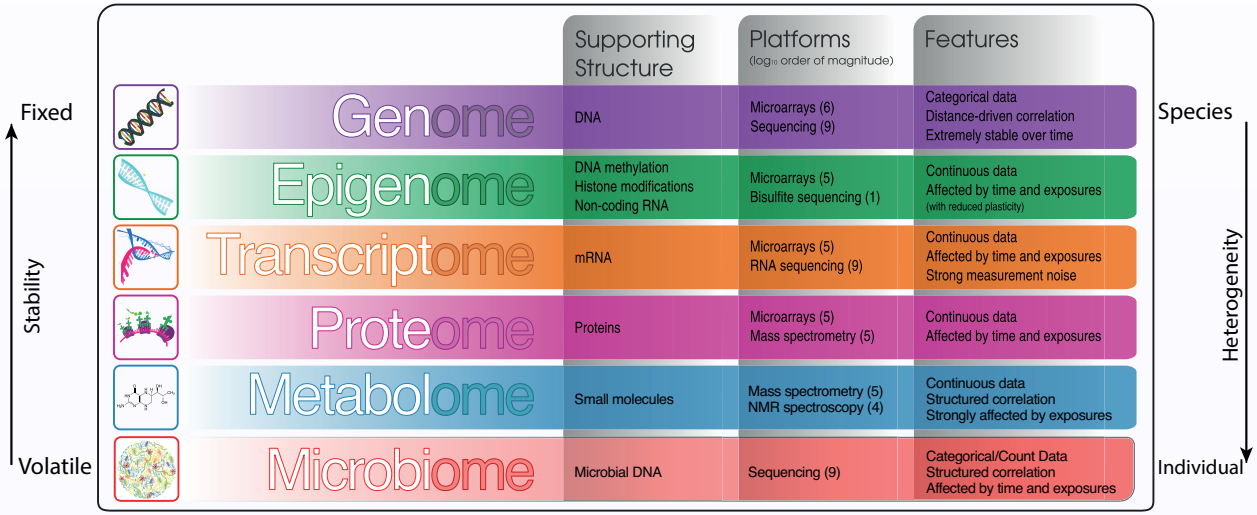
- OMICs data: high throughput biochemical measures of the abundance and/or structural features of molecules
- OMICs profiles: heterogeneous and complementary data



  - Dimension: ranging from hundreds to millions
  - Nature: continuous/binary/categorical/counts
  - Noise: more or less sensitive to experimental conditions
  - Correlation structure in the data: varies in strength and complexity

Deriving -OMICs biomarkers to characterise the exposome

- OMICs data: high throughput biochemical measures of the abundance and/or structural features of molecules
- OMICs profiles: heterogeneous and complementary data



  - Up pointing gradient of stability
  - Down pointing gradient of heterogeneity

  $\Rightarrow$ heterogeneity is driven by exposures and defines individual exposome and risk profiles

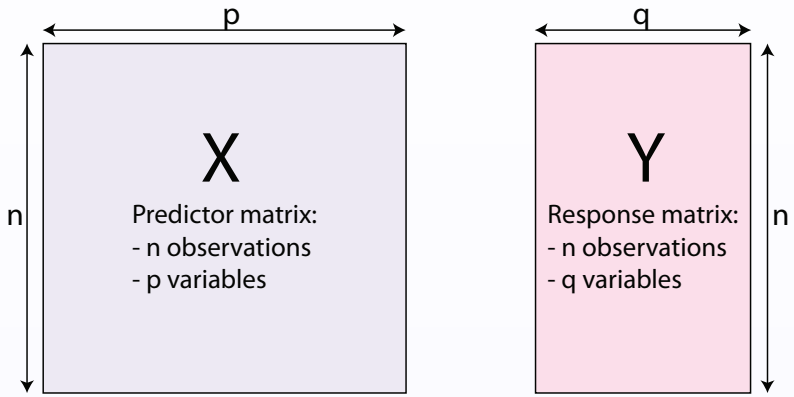
$\Rightarrow$ Potential to detect internal (effective) responses to external stresses

Statistical Challenges: beyond dimensionality

- Screening models‘OMICs & Exposure profiling’
  - Aim: identify within each OMIcs platforms & (sets of) exposures, relevant signatures of exposures health outcomes
  - Status: established methods
  - Challenge: model effects of mixtures & accommodate complex study designs
- Interpretation: functional and biological characterisation of identified OMICs biomarkers
  - Aim: integrate results arising form several OMIC platforms and explore their interplay
  - Status: methods/strategies are developing
  - Challenge: prioritize OMICs biomarkers

OMICs Profiling methods: **-WAS*

Data definition:



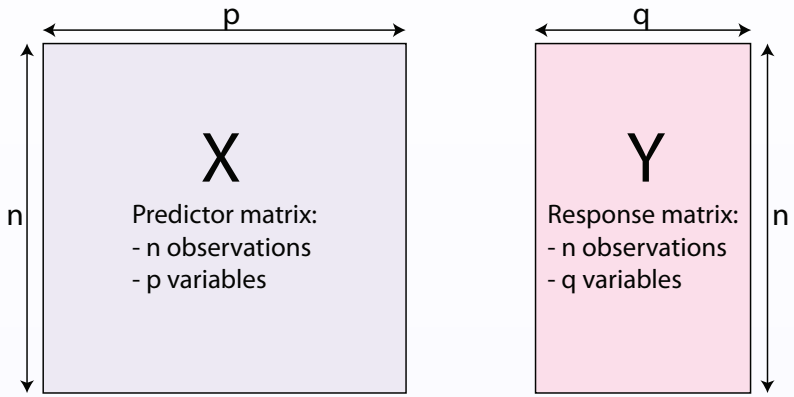
Aim: identify which of the p variables in X (OMICs/ exposure data) are associated with the outcome Y (disease status or (mixtures of) exposure(s))

- The $n < p$ situation:
  - More predictors than observations

    $\Rightarrow$ numerically intractable statistical inferences
  - Three main approaches have been proposed to get a situation where $n > p$

OMICs Profiling methods: **-WAS*

Data definition:



Aim: identify which of the p variables in X (OMICs/ exposure data) are associated with the outcome Y (disease status or (mixtures of) exposure(s))

- Univariate approaches: look at each predictor in X separately
- Dimension reduction techniques: summarize X into a lower dimension matrix
- Variable selection approach: define the best combination of variables in X to predict Y

Overarching model: Linear Framework

- Principle: assess the association between column(s) of X and the outcome Y
- Model formulation: generalised linear mixed model for individual i and predictor j

$$f(Y_i) = \alpha + \beta X_{ij} + u^{A^i} + \epsilon_{ij},$$

where:

  - Y_i is the measured outcome (OMICs or Health outcome)
  - X_{ij} is the observed value for j^{th} predictor (typically OMICs measurements, or exposures)
  - β is the effect size
  - u^{A^i}: random intercept modelling nuisance variation
  - ϵ_{ij} is the residual error measuring the random deviation from the linear relationship

$\Rightarrow$ linear models accommodates to most types of outcomes/predictors

Univariate Approaches: Multiple Testing correction

- Approach: run p linear models (one for each of the predictor separately)

	H_0 true	H_0 false	Total
H_0 rejected	V	S	R
H_0 accepted	U	T	p-R
Total	p_0	p-p_0	p

- FWER control:
  - $FWER$=α=$p(V \geq 1)$: the probability to have at least one FP
  - Aim: define the per-test significance α' ensuring $p(V = 0) \geq (1 - \alpha)$, where α is arbitrarily set.
- FDR control:
  - FDR=$E(V/R)$: the expected prop. of FP among positive calls
  - Aim: Identify FDR is upper bounded by the desired value
- FDR vs. FWER control: FDR is less stringent than FWER
  - FWER 5%: over 100 experiments <5 contain one (or more) FP
  - FDR control: over the 100 experiments the average #FP ≤ 5

$\Rightarrow$ need to accounts for the correlation among the tests performed

Univariate approaches: strengths and limitations

- Computational efficiency
  - Numerous numerically optimized implementations available
  - Possible parallelisation

    $\Rightarrow$ can accommodate $p > 10^6$
- Modelling flexibility
  - Linear models are restricted continuous covariates
  - Generalised linear models adapts to most types of outcomes (binary, categorical, count, survival)
  - No need to model the correlation within X in the model
  - Straightforward adjustment on potential confounders

    $\Rightarrow$ application to most OMICs data
- Limitations
  - Restricted to parametric marker-outcome relationship

    $\Rightarrow$ generalised additive models (computationally intensive)
  - Models do not account for potential combined effects of predictors

    $\Rightarrow$ need for multivariate approaches

Two main families of multivariate approaches

- Dimension Reduction techniques:
  - Aim: Identify summary covariates (components) constructed as linear combinations of original variables which accurately reconstruct in a lower dimension the structure of the original data
  - Main approaches: unsupervised ($e.g.$ PCA) and supervised ($e.g.$ PLS-based approaches)

    $\Rightarrow$ builds upon intrisic redundancies in the data
- Variable selection approaches
  - Aim: identify a sparse set of predictors that jointly predicts Y
  - Two main approaches: penalised regression ($e.g.$ lasso approaches), and Bayesian Variable Selection approaches (BVS)

    $\Rightarrow$ variable selection approaches implicitly correct for multiple testing

The principle of dimension reduction techniques

- Aim: Summarize the high dimensional X matrix in a lower dimension space.
- Definitions/Properties:
  - The original matrix X contains p predictors: $X_1, \ldots, X_p$
  - The i^{th} principal component PC_i is a linear combinations of the original variables such that:

$$PC_i = \alpha_{i1} X_1 + \cdots + \alpha_{ip} X_p$$

  - Any X can be decomposed in p orthogonal (non-redundant) PC's
    $\Rightarrow$ dimension reduction techniques seek for the linear combination coefficients to define each of the component.
  - Loadings (linear combination coefficients) measure the contribution of the original variables to each PC.
  - PC's can be ordered in terms of information restitution
    $\Rightarrow$ do not necessarily need all PC's for a accurate representation of the data

Main dimension reduction techniques

- PCA: sequentially estimate the loadings such that they maximize the variation in the X matrix

  $\Rightarrow$ this assumes that data X are characterized by their variance-covariance

- Method: singular value decomposition (eigenvalues/eigenvectors)

  $\Rightarrow$ eigenvalues measures the proportion of variance explained

  $\Rightarrow$ Limitation: unsupervised method, no guarantee that PC's are explanatory of the outcome ($e.g$ noise)

- Supervised alternative: Definition of the objective function:

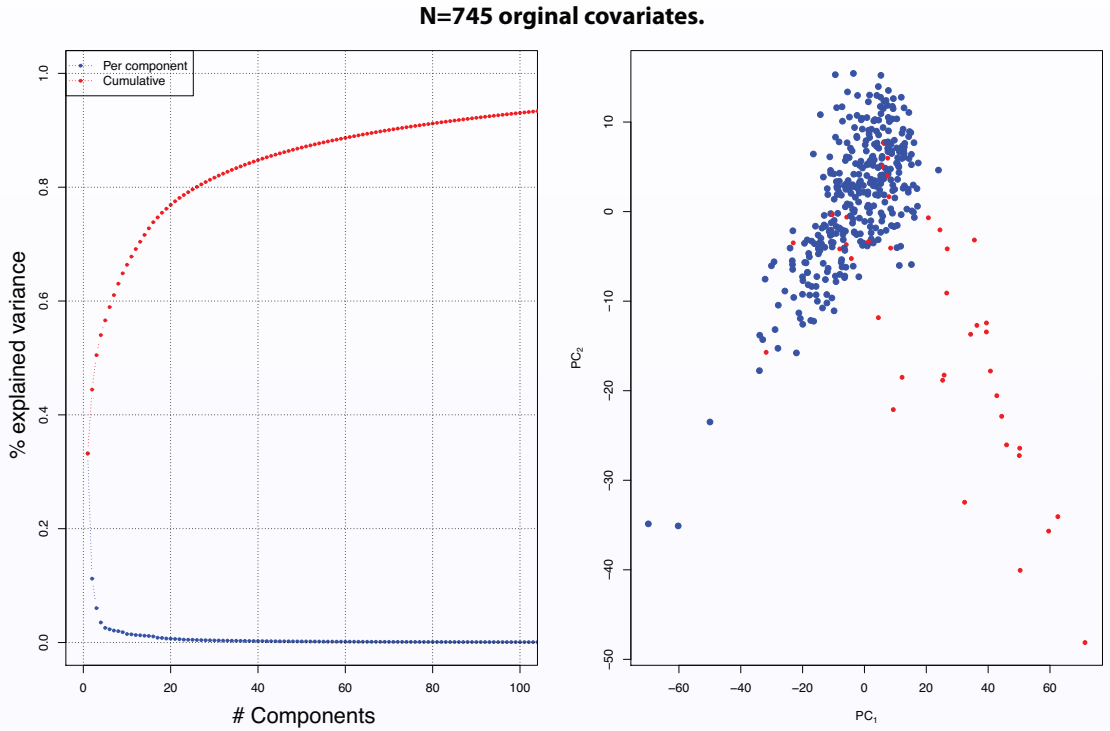
$$\max_{||\boldsymbol{u}_h||=1,||\boldsymbol{v}_h||=1} cov(X_h\boldsymbol{u}_h, Y_h\boldsymbol{v}_h) \qquad h = 1 \ldots H$$

$\Rightarrow$ PCs are defined to max. the covariance between X and Y

$\Rightarrow$ PLS identifies the variation in X that is related to Y

Dimension Reduction Techniques in practice

- Scree plot and Score plot:

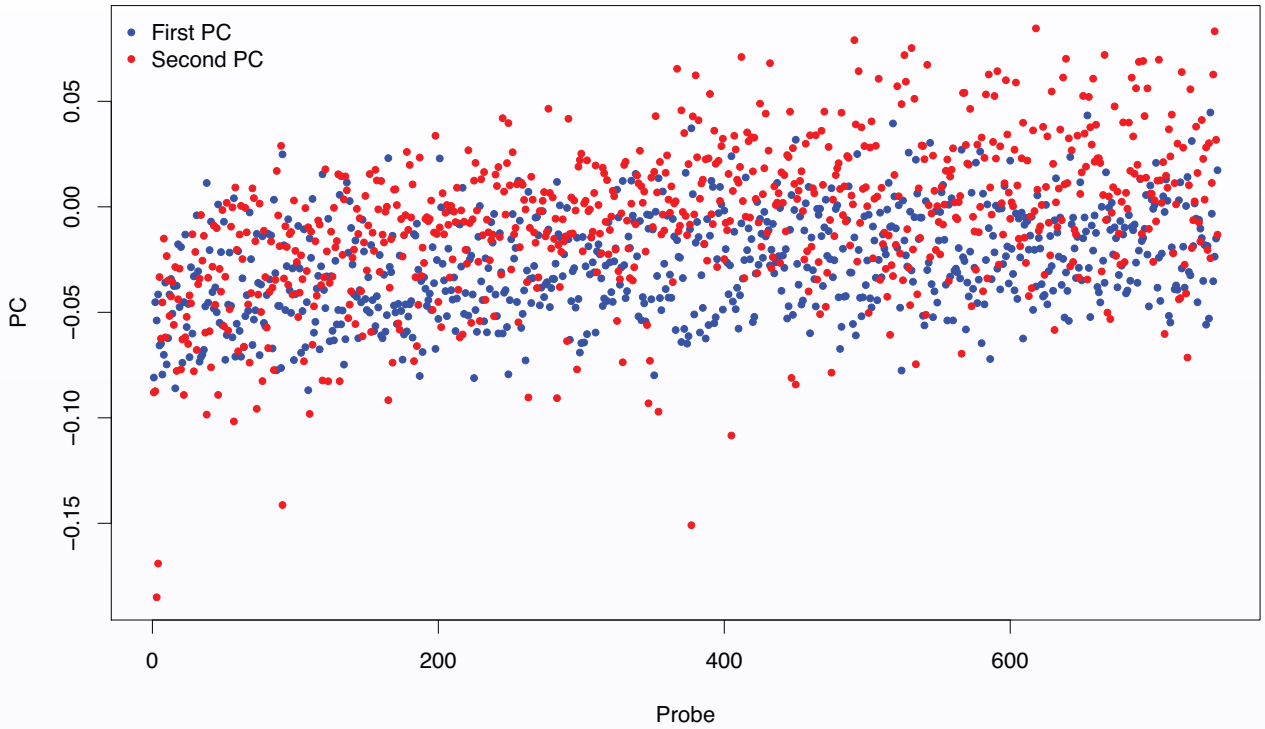


  - 90% variance explained for 80 PC's (≈10%)
  - Clear discrimination of cases and controls (PC_1 and PC_2 are strongly associated to Ca/Co)

$\Rightarrow$ efficient visualisation tool

Dimension Reduction Techniques in practice

- How to interpret results: Loadings plots



  - Loadings measure the contribution of original variables to the PC
  - No probes are clearly driving the PC's

    $\Rightarrow$ dimension reduction techniques may yield interpretation problems

    $\Rightarrow$ need to impose sparsity and/or to use supervised methods

Overview of penalised regression models

- Underlying model: linear model
- Principle: estimating the regression coefficients under a constraint
  - Ridge Regression: constraining the $L^2 = \sum_i \beta_j^2$ norm

    $\Rightarrow L^2$ constraint ensures numerical stability if $n \leq p$ and favours low β's
  - LASSO model: constraining the $L^1 = \sum_i |\beta_j|$ norm

    $\Rightarrow L^1$ constraint ensures sparsity of the results
- Penalised regression in practice
  - Set a calibration parameter λ
  - For a given value of λ the model will return β estimates satisfying the constraint (L^1 or L^2 =λ)

    $\Rightarrow$ How to determine λ?

    $\Rightarrow$ k-fold validation procedure: the optimal λ will minimise the prediction mean square error

Overview of penalised regression models

- Main features of Ridge regression
  - Numerical stability if $n \leq p$
- Main features of Lasso
  - The number of predictors with $\beta \neq 0$ is upper bounded by p

    $\Rightarrow$ LASSO ensures sparsity (and interpretability) of the results

    $\Rightarrow$ Elastic Net uses both penalties and combines both properties
- Main outcomes of penalized regression approaches: penalized regression coefficients
  - A vector of p regression coefficients
  - Due to the constraint most are estimated to be 0

    $\Rightarrow$ predictors with non-null regression coefficient are to be interpreted as jointly being associated to the outcome

    $\Rightarrow$ putative biomarkers are jointly identified and no measure of significance is provided

Bayesian Variable Selection Paradigm

Underlying Concept: given a certain function linking X and Y, among the p variables in X only a subset is informative regarding the response Y

- Definitions:
  - Let γ be a vector of 0's and 1's such that its i^{th} element:

$$\gamma_i = \begin{cases} 1 & \text{if the i}^{th} \text{ column of } X \text{ is in} \\ 0 & \text{otherwise} \end{cases}$$

  - Set p_γ as the number of variables of X that are in the model.
  - Let X_γ denote the design matrix of dimension $n \times p_\gamma$, collating all the columns of X for which $\gamma = 1$.
  - Formulation of one model: $Y - f(X_\gamma) = \epsilon$, where function f defines the relation between X and Y ($e.g.$ linear function)

$\Rightarrow$ *Aim: given f, estimate the vector γ that best predicts Y*

General Approach in Model Selection

- Comparing k models in that context relies on the following steps for each model $j, \in [1, k]$:
  - Set $\gamma = \gamma^j$ (e.g. null model contains only 0's)
  - Extract X_{γ^j} from X
  - Fit the model $Y - f(X_{\gamma^j}) = \epsilon$
  - Calculate a 'quality-of-fit' statistic S^j

  $\Rightarrow$ the best model (γ^{opt}) is the one providing the optimal value for S

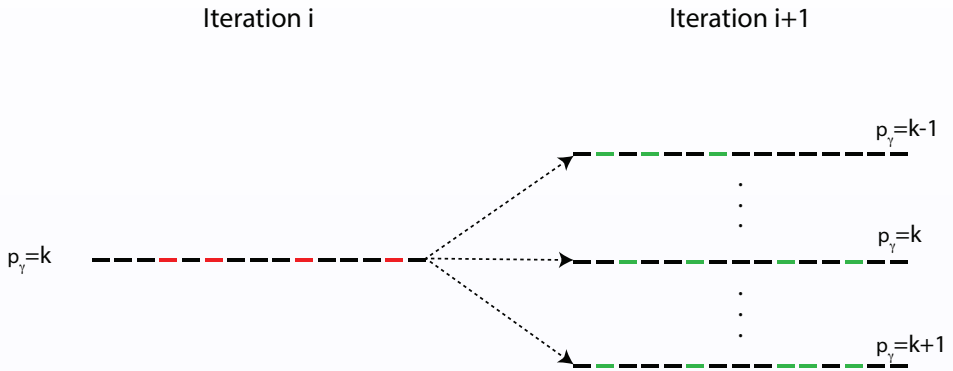
- Key issues:
  - Defining f and the subsequent S

    $\Rightarrow$ depends on nature of X and Y
  - Model space size: 2^p ($p = 50 \Rightarrow$1 million of billons of models)

    $\Rightarrow$ how to wander efficiently in that huge space?

Available implementations of BVS

- Shotgun Stochastic Search (SSS): intuitive search algorithm



- PiMASS: optimised search algorithm
- GUESS (R2GUESS): a BVS for multiple outcomes
  - Optimised Search algorithm: EMC
  - Computational optimisation: enabling GPU capacity
  - Scales to GWAS data and copes with 10 outcomes

$\Rightarrow$ GUESS is tailored for exposome investigation

$\Rightarrow$ BUT restricted to linear models (so far)

Outputs and interpretation

- Main outputs:
  - The history of all visited models and their conditional posterior

    $\Rightarrow$ each visited model is associated to a value of g, and a conditional posterior
  - The posterior distribution of g (the shrinkage factor)
- From these output, the following are computed:
  - The posterior of all model visited (and their rank)
  - The posterior model size
  - The posterior shrinkage factor (g)
  - The marginal probability of inclusion
  - And many more (R^2, β,...)

GUESS Example (from R2GUESS package)

Realistic simulation model: n=3,122 ind, p=273,675 SNPs, q=3 outcomes

$$\mathbf{Y} = \mathbf{X}_\alpha \mathbf{B} + \mathbf{E}, \quad \text{vec}\left(\mathbf{E}^{\mathrm{T}}\right) \sim \mathcal{N}_{n\times 3}(\mathbf{0}_{n\times 3},\, c \times \mathbf{I}_n \otimes \mathbf{\Sigma}),$$

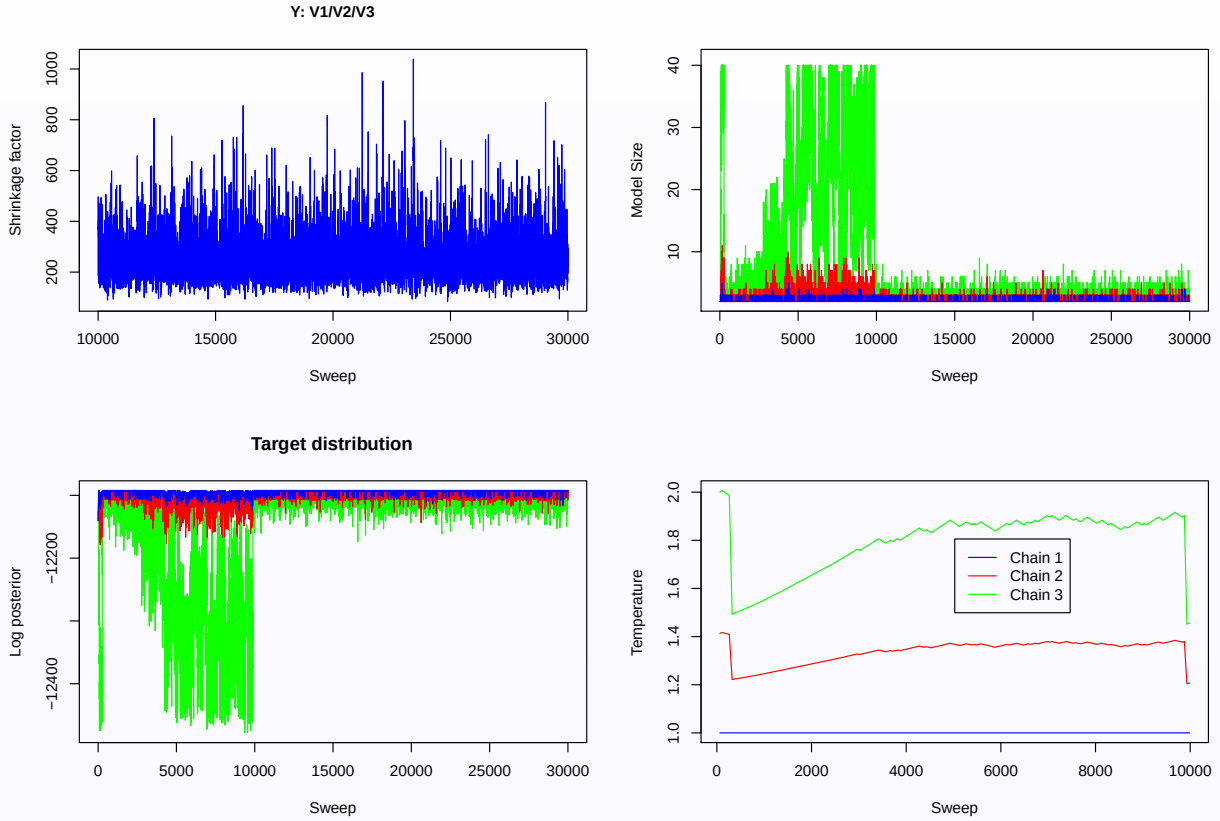
where B is the $r \times q$ matrix of regression coefficients, E is the $q \times q$ residual correlation matrix, and c is a scalar controlling the signal over noise ratio.

$$\mathbf{B} = \begin{pmatrix} 0.2 & 0.1 & 0.075 \\ 0.1 & 0.075 & 0.1 \end{pmatrix}, \text{ and } \mathbf{B} = \begin{pmatrix} 0.2 & 0.1 & 0.075 \\ 0.1 & 0.075 & 0.1 \\ 0.2 & 0.1 & 0.075 \\ 0.1 & 0.075 & 0.1 \\ 0.075 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.1 \\ 0.075 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.1 \end{pmatrix},$$

for r =2 and 8 'causal' SNPs, respectively. The value of c was calibrated such that the expected proportion of variance explained for each trait did not exceed 5%, to mimic small effects usually found in GWAS.

GUESS Example (from R2GUESS package)

GUESS run: 30K iterations, 3 chains, E_{p_γ}=5, σ_{p_γ}=5
Convergence assessment: monitor plot



- Assessing convergence and mixing

GUESS Example (from R2GUESS package)

GUESS run: 30K iterations, 3 chains, E_{p_γ}=5, σ_{p_γ}=5
Best Models: output

```
  Rank nVisits FirstVisit nEvalBefore1st ModeSize logCondPost postProb jeffrey
1    1   19641          1          284482        2   -12092.34 0.991000   89.60
2    2      51       6216          379173        3   -12098.56 0.001980   90.18
3    3      21        213          286608        3   -12099.13 0.001120   89.94
4    4      19        885          293542        3   -12099.87 0.000535   89.62
       modelName
1       1396 2009
2 1396 2009 2633
3 1396 2009 2538
4 1396 2009 2639
```

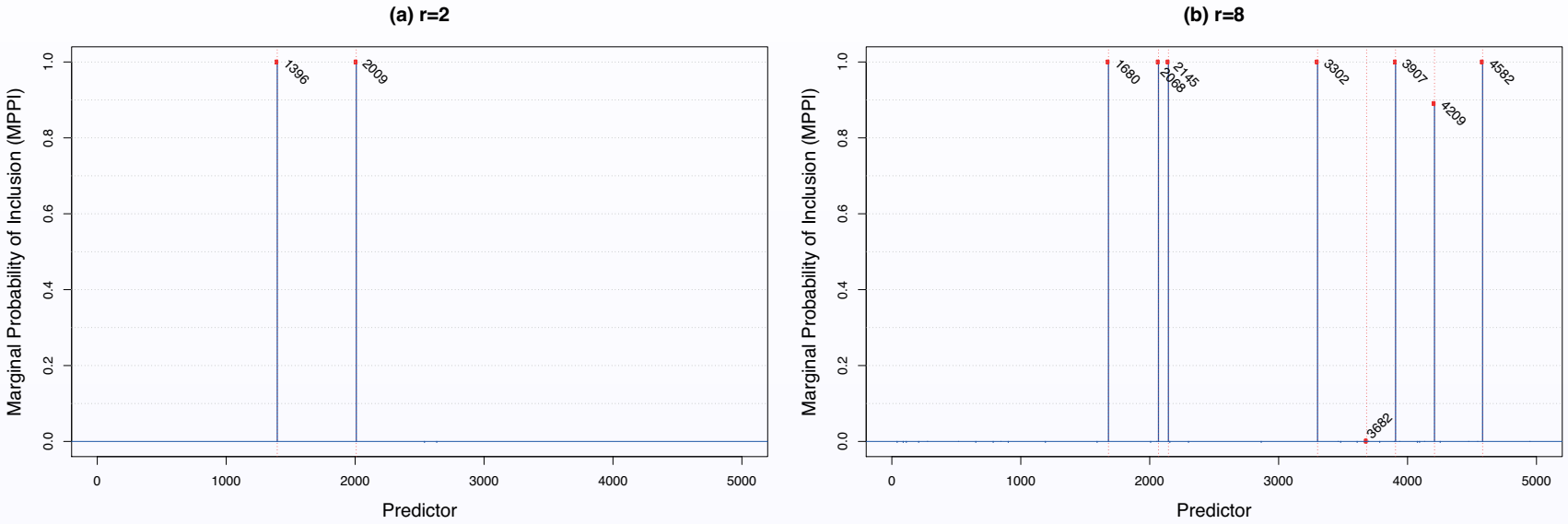
- Investigate # visits
- Model Posterior probability
- Look at model composition

$\Rightarrow$ can we infer the marginal contribution of the variable across all models?

GUESS Example (from R2GUESS package)

GUESS run: 30K iterations, 3 chains, E_{p_γ}=5, σ_{p_γ}=5

Marginal Probability of Inclusion (MPPI): posterior strength of association between a single predictor and the outcome

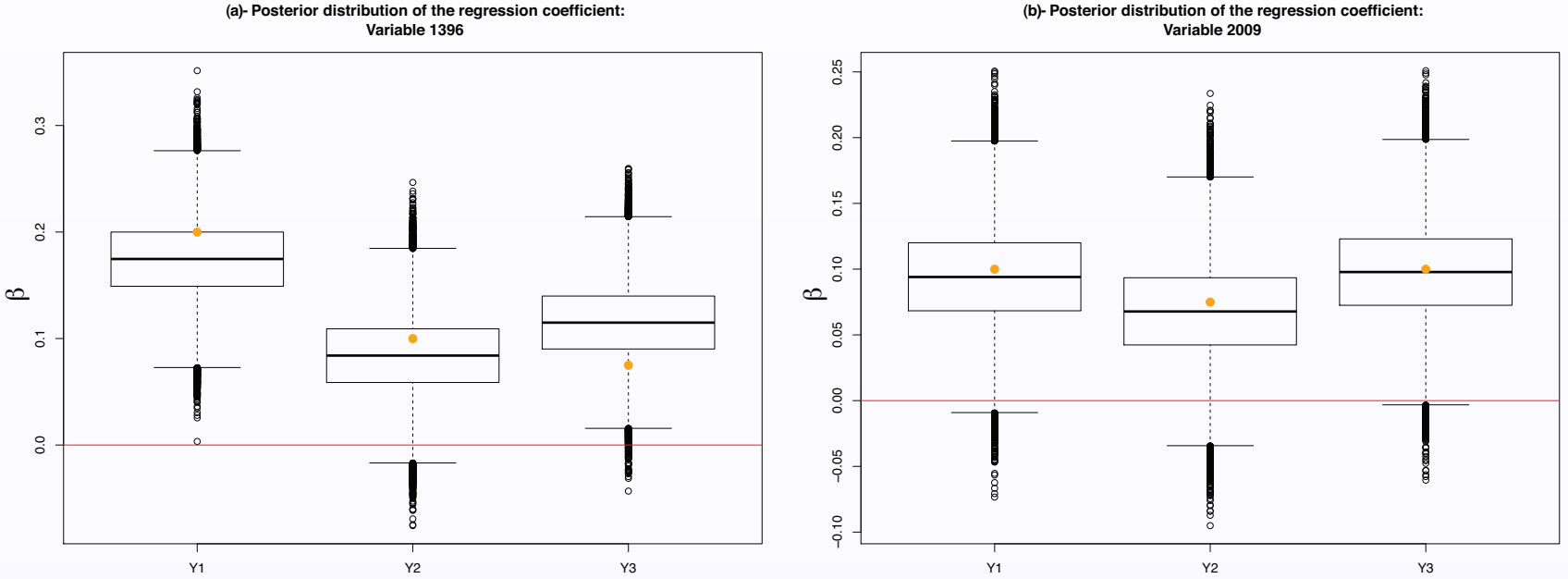


- Most 'causal' signals were recovered (2/2 for r=2, 7/8, for r=8)
- Missed signal was weaker, associated with greater variance and, for some (r=8 simulation), highly correlated to other causal signals

GUESS Example (from R2GUESS package)

GUESS run: 30K iterations, 3 chains, E_{p_γ}=5, σ_{p_γ}=5

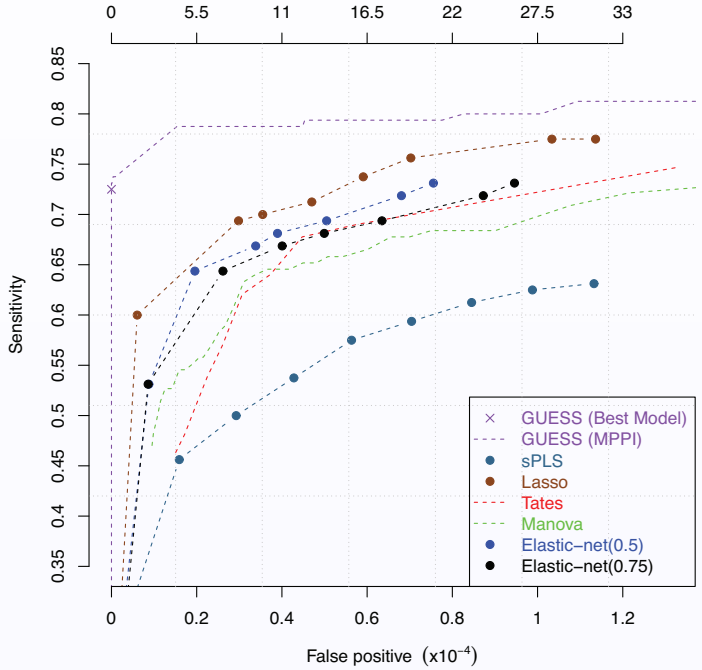
Effect size estimates: for a given covariate, the effect size can be simulated from posterior distributions



- Sharp posterior distribution of the regression coefficients
- Estimates are consistent with the values used for the simulation

GUESS Example (from R2GUESS package)

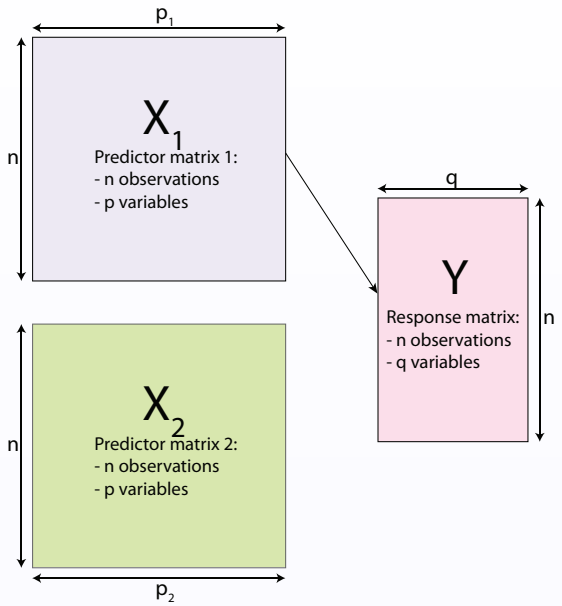
Performance assessment: comparison with alternative methods



- GUESS outperforms all alternative methods
- Same applies to $q = 1$ situation

$\Rightarrow$ overall good performances of GUESS

OMICs Integration: a Two-Step Strategy

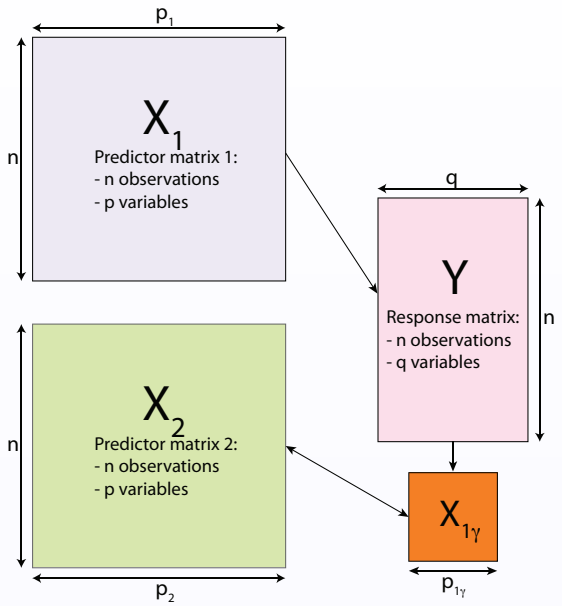


Full-resolution profiling of one OMIC dataset

- Using established profiling approaches

  $\Rightarrow$ Identify OMIC-specific biomarkers

OMICs Integration: a Two-Step Strategy



Identify correlated features in the second OMIC profile

- Using established profiling approaches

  $\Rightarrow$ Identify $p_{1\gamma}$ OMIC-specific biomarkers

- Assess the correlation between the $p_{1\gamma}$ biomarkers and the p_2 OMICs features

  $\Rightarrow$ perform $p_{1\gamma}$ OMIC-wide scans

OMICS integration: smoking and methylation example

Human Molecular Genetics, 2015, Vol. 24, No. 8 2349–2359
doi: 10.1093/hmg/ddu751
Advance Access Publication Date: 2 January 2015
Original Article

ORIGINAL ARTICLE

Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation

Florence Guida[1,†], Torkjel M. Sandanger[2,†], Raphaële Castagné[1,†], Gianluca Campanella[1], Silvia Polidoro[3], Domenico Palli[4], Vittorio Krogh[5], Rosario Tumino[6], Carlotta Sacerdote[3], Salvatore Panico[7], Gianluca Severi[3,8,9], Soterios A. Kyrtopoulos[10], Panagiotis Georgiadis[10], Roel C.H. Vermeulen[1,11,12], Eiliv Lund[2], Paolo Vineis[1,3,‡] and Marc Chadeau-Hyam[1,11,‡,*]

- Data: 745 prospective blood samples, DNA methylation profiles from Illumina Infinium HumanMethylation450
- Model: Epigenome-wide association study for smoking status (current-former $vs.$ never)
- Results: 751 differentially methylated CpG loci

  $\Rightarrow$ how to interpret the role of these methylation changes?

OMICS integration: smoking and methylation example

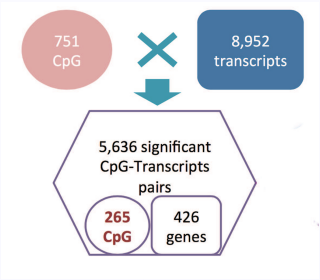
Human Molecular Genetics, 2015, Vol. 24, No. 8 2349–2359
doi: 10.1093/hmg/ddu751
Advance Access Publication Date: 2 January 2015
Original Article

ORIGINAL ARTICLE

Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation

Florence Guida[1,†], Torkjel M. Sandanger[2,†], Raphaële Castagné[1,†], Gianluca Campanella[1], Silvia Polidoro[3], Domenico Palli[4], Vittorio Krogh[5], Rosario Tumino[6], Carlotta Sacerdote[3], Salvatore Panico[7], Gianluca Severi[3,8,9], Soterios A. Kyrtopoulos[10], Panagiotis Georgiadis[10], Roel C.H. Vermeulen[1,11,12], Eiliv Lund[2], Paolo Vineis[1,3,‡] and Marc Chadeau-Hyam[1,11,‡,*]

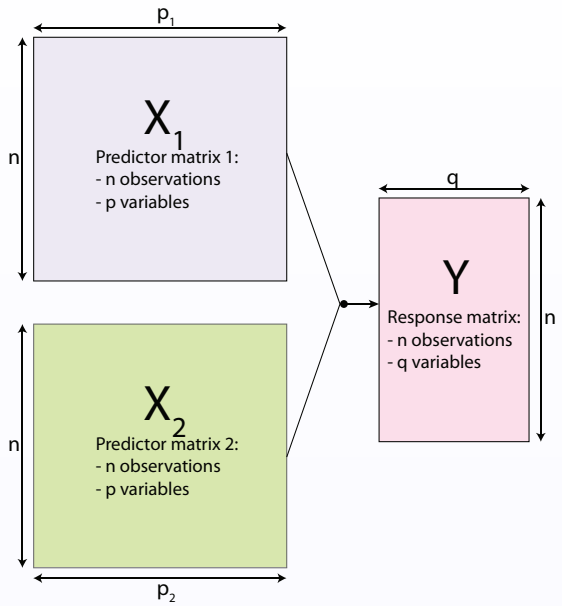
- Additional data (N=295 samples) with gene expression data
- We linearly regress candidate CpG sites vs all transcripts: 6.7×10^6 CpG-Transcript pairs



- 5,600 Bonferroni significant pairs (265 unique CpG, and 426 genes)
- Only 5 pairs involved CpG and transcripts on the same gene

  $\Rightarrow$ methylation - transcripts correlations respond to complex patterns

Multi-omic profiling



Aim: Agnostic investigation of 'multi-omic' markers of the outcome

- OMICs profiles are pooled and regressed against Y

  $\Rightarrow$ Profiling approaches handling (ultra-) high dimensional X matrix

  $\Rightarrow$ restricted to Multivariate approaches

Univariate approaches to OMICs integration

- Appropriate features:
  - Computational efficiency
  - Modelling flexibility

  $\Rightarrow$ to ensure feasibility and improve interpretability filtering may be required
- Scalability: feature pre-selection procedures
  - 2-step strategies
  - Combination with clustering approaches
  - Using external information (e.g. pathways)

  $\Rightarrow$ affects computational efficiency
- Limitation: no modelling of the (complex) correlation structures within and across OMIC profiles

  $\Rightarrow$ need to investigate multivariate approaches

Dimension reduction techniques and OMICs integration

- Methods of choice: PLS approaches
  - Canonical Correlation (symmetric)
  - PLS-R (asymmetric, adding a predictor/response structure)

    $\Rightarrow$ these handle blocks of multivariate matrices
- Interpretability improvements: grouping and sparsity
  - Sparsity achieved through penalisation
  - Grouping signals a priori (e.g. pathways, genes)

- sparse PLS components (sPLS)

$$C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$$

- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

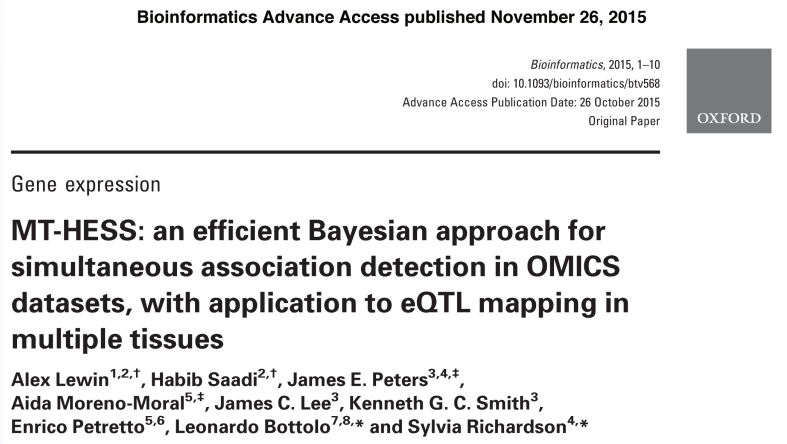
- sparse group PLS components (sgPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{=0} X_4 + \underbrace{u_5}_{=0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

Bayesian variable selection and OMICs integration

- GUESS/R2GUESS handles multivariate predictor and response
  - Hundreds of thousands of predictors
  - Few responses

    $\Rightarrow$ need to extends to higher dimensional Y's
- HT-HESS: handling high dimensional Y



Bioinformatics Advance Access published November 26, 2015

Bioinformatics, 2015, 1–10
doi: 10.1093/bioinformatics/btv568
Advance Access Publication Date: 26 October 2015
Original Paper

OXFORD

Gene expression

MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues

Alex Lewin[1,2,†], Habib Saadi[2,†], James E. Peters[3,4,‡], Aida Moreno-Moral[5,‡], James C. Lee[3], Kenneth G. C. Smith[3], Enrico Petretto[5,6], Leonardo Bottolo[7,8,*] and Sylvia Richardson[4,*]

$\Rightarrow$ includes a preliminary response filtering step in a hierarchical setting

- Limitation: assumes a global predictor-response between the two block of data

OMICs profiling in Exposome Research: challenges

- Modelling nuisance variation
- Accounting for repeated measurements
- Modelling multivariate response to multivariate exposures
- Accounting for a group structure in the predictor matrix

Accounting for technically-induced noise

- Definition: 'nuisance variation'
  - During profile acquisition bio-samples undergo numerous processing steps (extraction etc...)
  - During each of these steps physical processes and biochemical reaction are sensitive to experimental reaction ($e.g.$ temperature, pH, ...)

    $\Rightarrow$ while processing numerous samples, differences in experimental conditions may result in additional variation in the observations which is not related to the sample

    $\Rightarrow$ nuisance variation: additional variance diluting true signals
- Controlling/Accounting for 'nuisance variation'
  - During data acquisition (randomisation, control of experimental conditions)
  - At the analytical stage (normalisation techniques, including error in the model)

    $\Rightarrow$ how to model nuisance variation?

Mixed models and nuisance variation

- Simplifying assumption: technical covariates affects the efficacy of intermediate physical/chemical transformation during sample processing regardless of the nature of the sample

  $\Rightarrow$ technical variation will uniformly only shift the OMICs marker intensity measurement

  $\Rightarrow$ random intercept model

- Formulation, for individual i and predictor j:
  - Variable of interest: X_i (Ca/Co, or exposure levels)
  - Predictors: Y_i, observed value for the j^{th} OMICs marker
  - Fixed effects: FE_i; adjustment variables ($e.g.$ age)
  - Random Effect variables: u^{A_i}, where A_i are nuisance variables ($e.g.$ batch in which sample i was processed)
  - Full Model:

$$Y_{ij} \sim \alpha + \beta_1 X_i + \beta_2 FE_i + u^{A_i} + \epsilon_{ij}$$

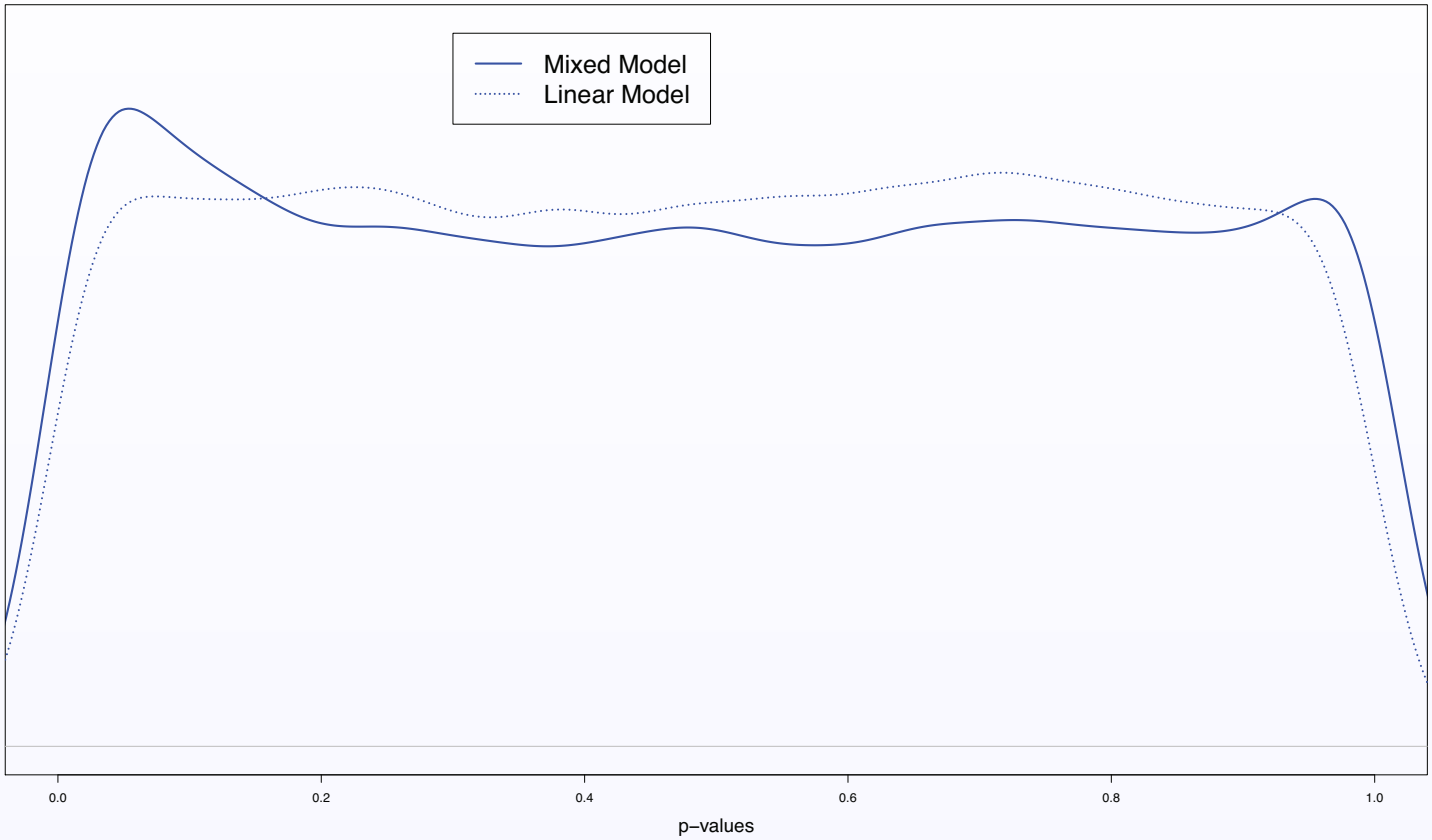
Mixed models and nuisance variation

- Statistical inference: likelihood ratio test
  - Run the model with variable of interest (X_{ij})
  - Run the model without variable of interest
  - Compare both models (likelihood ratio test)

    $\Rightarrow$ for each OMIC marker j we obtain a p-value testing the association between the $X_{.j}$ and the disease status/or exposure
- Main outcomes over the p models
  - A matrix of $p \times k$ regression coefficients (and corresponding p-values) measuring the effect of the k fixed effects on the p markers
  - A list of p regression coefficients (and corresponding p-values) measuring the association of the outcome of interest and the p markers
  - A set of p estimates of the random effects (for each marker l estimates are provided; one for each RE class)

Mixed models and nuisance variation

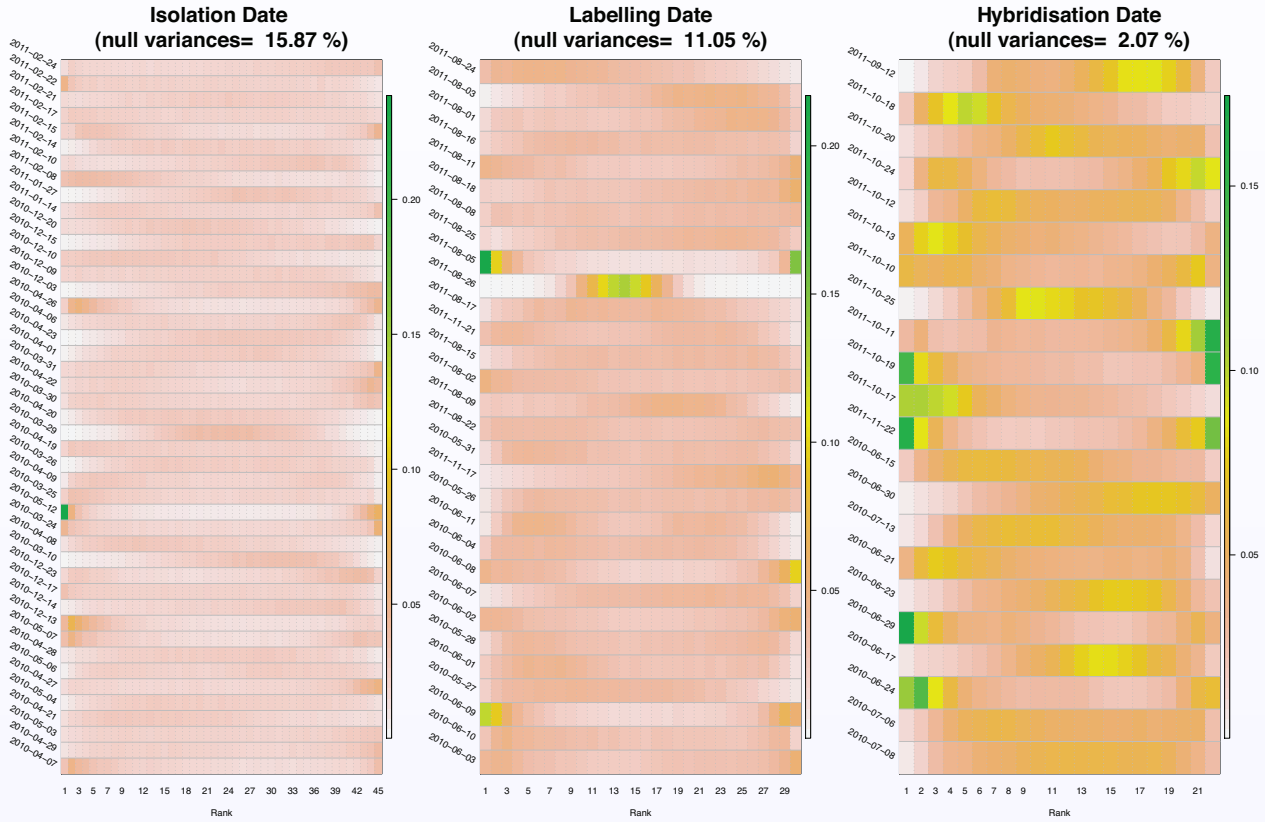
- Results interpretation: Assess the impact of nuisance variation - comparison with linear model



$\Rightarrow$ Mixed model efficiently corrects for nuisance variance

Mixed models and nuisance variation

- Results interpretation: Assess the impact of nuisance variation - investigating RE estimates
  We rank each date wrt to their variance estimate across the $\sim$ 30K values:



Mixed models and nuisance variation

Summary:

- Isolation date: 45 dates
  - 19.3% probes were associated with a null variance
  - 12/05/2010 generate higher variances
- Hybridisation date: 22 dates
  - 2.1% probes were associated with a null variance
  - 22/11/2011; 19/10/2011; 17/10/2011; 13/10/2011; 29/06/2010; 24/06/2010 are outliers
- Labelling Date: 30 dates
  - 11.1% probes were associated with a null variance
  - 05/08/2011; 09/06/2010 are outliers

$\Rightarrow$ higher contribution of hybridisation to the nuisance variation

Nuisance variation and multivariate models

- Accounting for nuisance variation:
  - Penalised regression: (generalised) linear mixed models are implemented
  - BVS: some implementations include random effects

    $\Rightarrow$ computationally intensive approaches

    $\Rightarrow$ what are the alternative approaches?
- Alternative to mixed models:
  1. Run a linear mixed model on the data:

$$Y_{ij} \sim \alpha + \beta_1 X_i + \beta_2 FE_i + u^{A_i} + \epsilon_{ij}$$

  2. From estimates get de-noised data:

$$\bar{Y}_{ij} = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 FE_i + \hat{\epsilon_{ij}}$$

  3. Run the variable selection approach on the de-noised data

Implementing complex study designs: Experimental studies

- PISCINA study: a pre-post intervention study
  - Design: 60 participants were enrolled to swim for 40 minutes in a chlorinated pool
  - Data: exposure (exhaled breath) and OMICs (blood) measured before and after swimming (N=2/participant)
- Oxford Street study: a cross-over design
  - 60 participants walking in Hyde Park and Oxford Street
  - 6 OMIC profiles/ participants (1 before and 2 after each walk)
  - 2 exposure measurements (one during each walk)

The multivariate normal model: formalism

- Aim: identify a unique model to analyse all these data
- Generation of general design:
  - 3 blood samples per participant and situations
  - 2 Exposure situations

$\Rightarrow$ Modelling approach: we consider all OMIC profiles at once as multivariate outcome Y

- Formalism:
  - Y_1, Y_2, Y_3,: samples in unexposed environment (t=0, 2, 24)
  - Y_4, Y_5, Y_6: samples in exposed environment (t=0, 2, 24)
- Underlying Model:
- $Y = \{Y_1, Y_2, Y_3, Y_4, Y_5, Y_6\}$: for one given individual a set of 6 OMICs levels

$$Y = \{Y_1, Y_2, Y_3, Y_4, Y_5, Y_6\} \sim \mathcal{N}(\mu, \Sigma), \text{ where}$$

  - $\mu(6, 1)$: expected (mean) OMIC intensity
  - $\Sigma(6, 6)$: the variance covarionce across all 6 observations

The MVN model for Oxford street study

Notations:

- $Y_{i,j,k}$: the metabolite level at location j (1 for HP, 2 for Ox), and time point k (1,2 or 3) for individual i
- $X_{i,j}$: the exposure level for individual i during the walk at location j

$$
\begin{aligned}
Y_{i,1,1} &= \beta_{0,1,1} + \beta_{1,1,1} * X_{i,1} + \epsilon_{i,1,1} \\
Y_{i,1,2} &= \beta_{0,1,2} + \beta_{1,1,2} * X_{i,1} + \epsilon_{i,1,2} \\
Y_{i,1,3} &= \beta_{0,1,3} + \beta_{1,1,3} * X_{i,1} + \epsilon_{i,1,3} \\
Y_{i,2,1} &= \beta_{0,2,1} + \beta_{1,2,1} * X_{i,1} + \epsilon_{i,2,1} \\
Y_{i,2,2} &= \beta_{0,2,2} + \beta_{1,2,2} * X_{i,1} + \epsilon_{i,2,2} \\
Y_{i,2,3} &= \beta_{0,2,3} + \beta_{1,2,3} * X_{i,1} + \epsilon_{i,2,3}, \text{ where}
\end{aligned}
$$

the residual error $\epsilon_{i,j,k}$ is such that $\epsilon_{i,j,k} \sim MVN_6(0_6, \Sigma_6)$

$\Rightarrow$ how to define Σ?

$\Rightarrow$ once the effect of predictors in accounted for, how do metabolic levels (co)-vary within each participant, across experimental conditions?

Devising the residual variance-covariance matrix

- Interpretation of the residual variance covariance matrix:
  - Residuals measure individual OMIC variability once the effect of exposure (and confounders) have been removed (calculated over the entire population).
  - How do these residuals should correlate within one individual?

$\Rightarrow$ residual variance: unexplained variability within participants

- Naive approach: consider repeated measures as independent
  Residual variance-covariance: for 6 measurements per individuals

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

$\Rightarrow$ ignores the natural correlation within individuals & experimental conditions

$\Rightarrow$ loss of power

Devising the residual variance-covariance matrix

- Interpretation of the residual variance covariance matrix:
  - Residuals measure individual OMIC variability once the effect of exposure (and confounders) have been removed (calculated over the entire population).
  - How do these residuals should correlate within one individual?

  $\Rightarrow$ residual variance: unexplained variability within participants

- Mixed models: accounts for correlation within individuals
  Residual variance-covariance: for 6 measurements per individuals

$$\begin{pmatrix} \sigma^2 & \delta & \delta & \delta & \delta & \delta \\ \delta & \sigma^2 & \delta & \delta & \delta & \delta \\ \delta & \delta & \sigma^2 & \delta & \delta & \delta \\ \delta & \delta & \delta & \sigma^2 & \delta & \delta \\ \delta & \delta & \delta & \delta & \sigma^2 & \delta \\ \delta & \delta & \delta & \delta & \delta & \sigma^2 \end{pmatrix}$$

$\Rightarrow$ the link between the 6 observations is the same and does not depend on experimental conditions

$\Rightarrow$ per-subjet random intercept

Devising the residual variance-covariance matrix

- Interpretation of the residual variance covariance matrix:
  - Residuals measure individual OMIC variability once the effect of exposure (and confounders) have been removed (calculated over the entire population).
  - How do these residuals should correlate within one individual?

  $\Rightarrow$ residual variance: unexplained variability within participants

- Multivariate Normal Models: fully flexible model
  Residual variance-covariance: for 6 measurements per individuals

$$\begin{pmatrix} \sigma^2_{11} & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} & \delta_{16} \\ \delta_{21} & \sigma^2_{22} & \delta_{23} & \delta_{24} & \delta_{25} & \delta_{26} \\ \delta_{31} & \delta_{32} & \sigma^2_{33} & \delta_{34} & \delta_{35} & \delta_{36} \\ \delta_{41} & \delta_{42} & \delta_{43} & \sigma^2_{44} & \delta_{45} & \delta_{46} \\ \delta_{51} & \delta_{52} & \delta_{53} & \delta_{54} & \sigma^2_{55} & \delta_{56} \\ \delta_{61} & \delta_{62} & \delta_{63} & \delta_{64} & \delta_{65} & \sigma^2_{66} \end{pmatrix}$$

$\Rightarrow$ covariance within individual depends on experimental conditions

Devising the residual variance-covariance matrix

- Interpretation of the residual variance covariance matrix:
  - Residuals measure individual OMIC variability once the effect of exposure (and confounders) have been removed (calculated over the entire population).
  - How do these residuals should correlate within one individual?

$\Rightarrow$ residual variance: unexplained variability within participants

- Multivariate Normal Models: fully flexible model
  Residual variance-covariance: for 6 measurements per individuals

$$\begin{pmatrix} \sigma^2_{11} & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} & \delta_{16} \\ & \sigma^2_{22} & \delta_{23} & \delta_{24} & \delta_{25} & \delta_{26} \\ & & \sigma^2_{33} & \delta_{34} & \delta_{35} & \delta_{36} \\ & & & \sigma^2_{44} & \delta_{45} & \delta_{46} \\ & & & & \sigma^2_{55} & \delta_{56} \\ & & & & & \sigma^2_{66} \end{pmatrix}$$

$\Rightarrow$ covariance within individual depends on experimental conditions

$\Rightarrow$ symmetric matrix: $\delta_{ij} = \delta_{ji}$

Oxford Street 2 Study: MVN modelling

- Exposure measurement:
  - One exposure measurements during each walk
  - 6 main exposures measured (NO_2, PM_{10}, $PM_{2.5}$, CBLK, Noise, Humidity)
- Model parametrisation: which exposure for which time point
  - t_0: average (modelled) exposure over 1-year before the walk

    $\Rightarrow$ background/long term exposure
  - t_1 and t_2: measured exposure

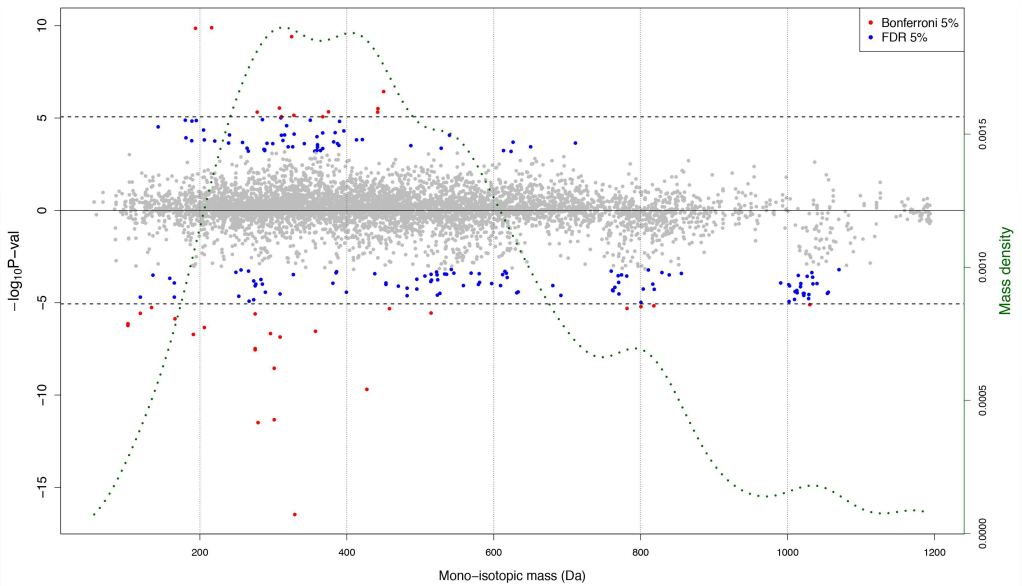
    $\Rightarrow$ exposure changes due to the experiment
- Model formulation: disentangling effect of background exposure and exposure change from the background levels induced by the walk

  $\Rightarrow$ investigate molecular changes induced by exposure stresses after the experiment

  $\Rightarrow$ setting t_1=t_2 accounts for a lagging effect
- To ensure estimates comparability, Y values are scaled.

MWAS from Oxford Street 2 study

- Results from MWAS (N 5,000 features) for NO_2:

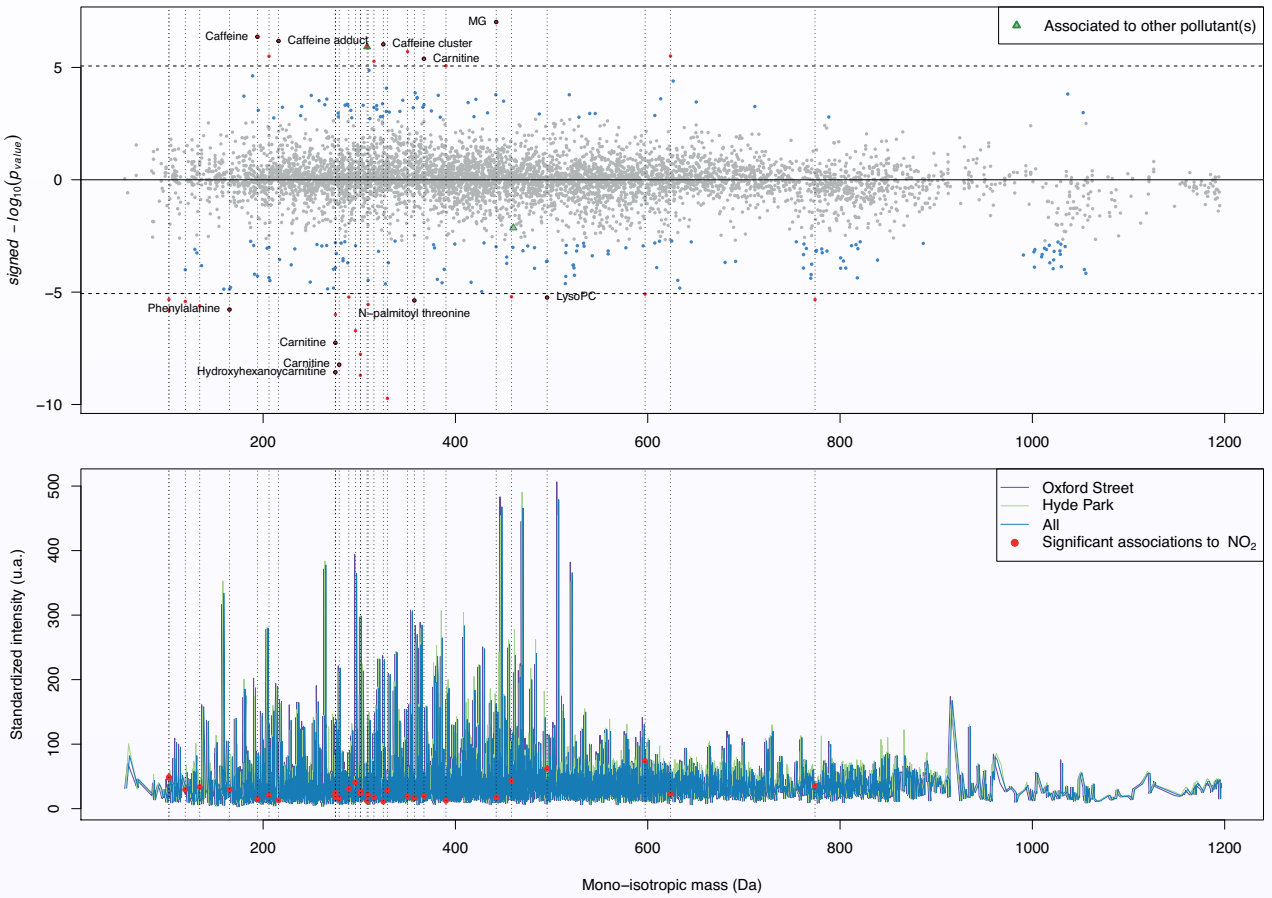


$\Rightarrow$ we identified several metabolome-wide associations linked to experimentally-induced changes in NO_2 levels: 51 Bonferroni and 464 FDR 5% hits

- Strong clustering and correlation across the metabolic features: 15 and 50 principal component explain >90% variance
- Associations are also present in low density mass regions

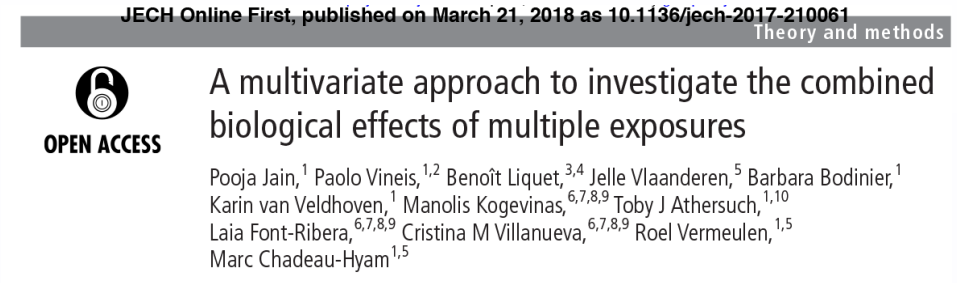
MWAS from Oxford Street 2 study: NO_2 example

- Results from MWAS (N 5,000 features) for NO_2:



$\Rightarrow$ we identified several metabolome-wide associations linked to experimentally-induced changes in NO_2 levels

Investigating effects of multivariate exposures (Jain *et al.*)


JECH Online First, published on March 21, 2018 as 10.1136/jech-2017-210061

Theory and methods

OPEN ACCESS

A multivariate approach to investigate the combined biological effects of multiple exposures

Pooja Jain,[1] Paolo Vineis,[1,2] Benoît Liquet,[3,4] Jelle Vlaanderen,[5] Barbara Bodinier,[1] Karin van Veldhoven,[1] Manolis Kogevinas,[6,7,8,9] Toby J Athersuch,[1,10] Laia Font-Ribera,[6,7,8,9] Cristina M Villanueva,[6,7,8,9] Roel Vermeulen,[1,5] Marc Chadeau-Hyam[1,5]

- Question: due to exposure co-occurence, are all exposures needed to explain the inflammatory response?

  $\Rightarrow$ is there a 'mixture effect'?

  $\Rightarrow$ use all exposures as predictor and assess the most relevant ones
- Need to account for the multidimensional nature of the response
- Method: (sparse) PLS model of (N=4) exposures $vs.$ (N=13) proteins
- Multi-level extension accounts for the repeated measure design
- Aim: identify molecular signatures of exposures:
  - which (sets of) exposure are affecting proteins level (X selection)
  - which (sets of) proteins are affected by exposures (Y selection)
  - what set of exposures most affect a subset of the proteins (X& Y selection)

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ Due to exposure correlation: C_{1X} explains $\sim 95\%$ of the variance in X

$\Rightarrow$ All exposures have negative loadings

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ Weaker correlations in proteins: C_{1Y} explains $\sim$ 20% of the variance in Y

$\Rightarrow$ only 4 negative loadings (including CXCL10)

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow C_{1X}$ explains $\sim$ 10% of the variance in Y

$\Rightarrow$ limited explanatory performances of the exposures

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow C_{2X}, \ldots, C_{4X}$ explain lesse than 5% of the variance in X

$\Rightarrow$ they explain less than 2% of the Y variance

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on X excludes $BrCH_3$

$\Rightarrow$ explained variance of X and Y by $C_{1X'}$ are not affected

Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on Y excludes 4 proteins

$\Rightarrow$ resulting improvements in the Y explained variance

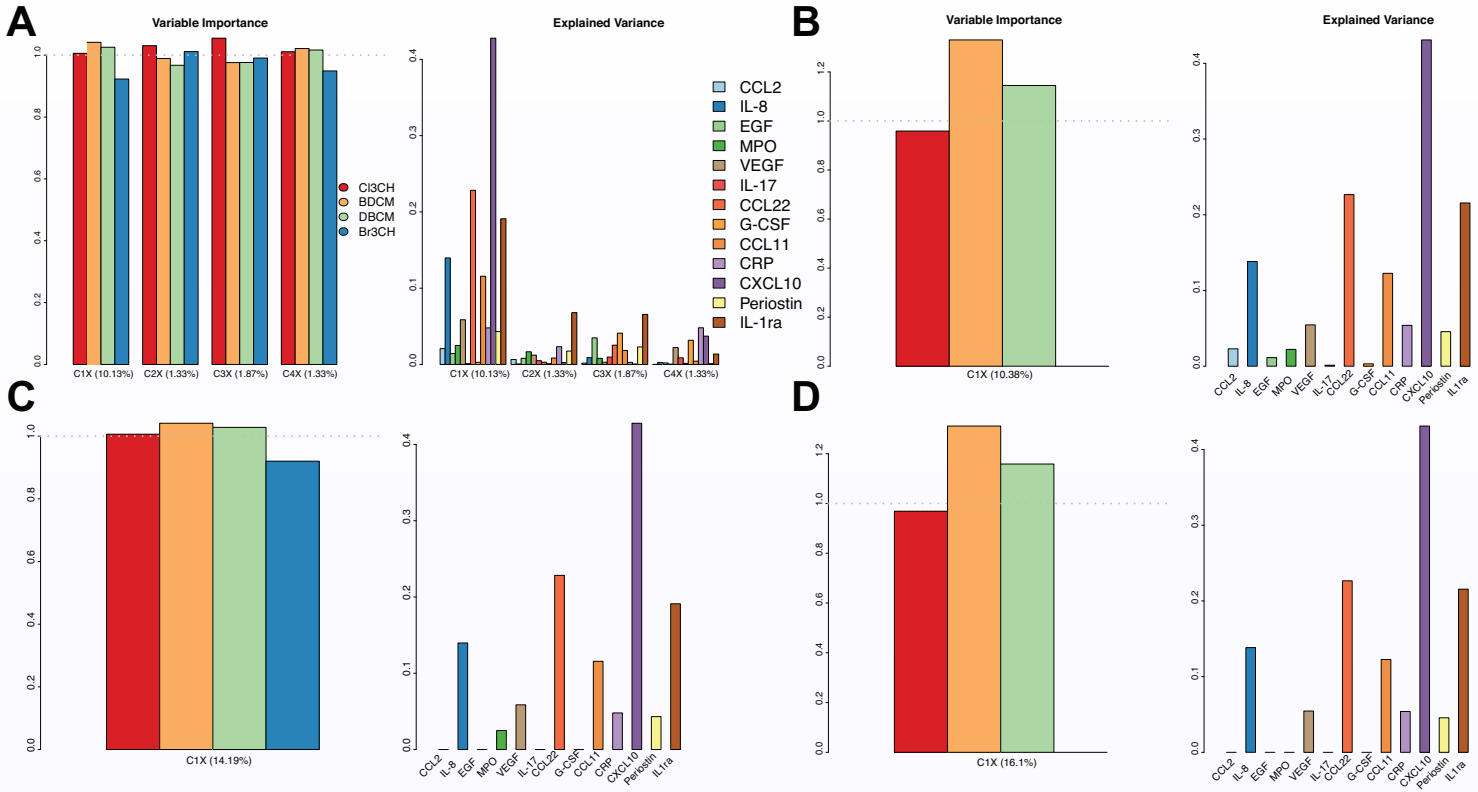
Parameter estimates overview:

	PLS				sPLS on X	sPLS on Y	sPLS on X and Y
Exposures (X matrix)	C_{1X}	C_{2X}	C_{3X}	C_{4X}	$C_{1X'}$	$C_{1X''}$	$C_{1X'''}$
Cl_3CH	-0.50	-0.60	-0.60	-0.17	-0.48	-0.50	-0.48
BDCM	-0.52	-0.21	0.45	0.70	-0.67	-0.52	-0.66
DBCM	-0.51	0.11	0.51	-0.68	-0.57	-0.51	-0.58
$BrCH_3$	-0.46	0.76	-0.42	0.15	0.00	-0.46	0.00
Explained Variance in X	94.8%	4.5%	0.6%	0.04%	94.0%	94.8%	94.0%
Explained Variance in Y	10.1%	1.3%	1.9%	1.3%	10.4%	14.2%	16.1%
Protein levels (Y matrix)	C_{1Y}	C_{2Y}	C_{3Y}	C_{4Y}	$C_{1Y'}$	$C_{1Y''}$	$C_{1Y'''}$
CCL2	0.12	0.195	-0.09	-0.02	0.13	0.00	0.00
IL-8	0.31	0.062	0.19	0.12	0.32	0.30	0.29
EGF	-0.10	0.216	-0.38	-0.11	-0.09	0.00	0.00
MPO	-0.14	0.310	0.18	0.05	-0.13	-0.02	0.00
VEGF	0.21	-0.266	-0.11	-0.36	0.20	0.13	0.11
IL-17	0.03	0.169	0.20	0.22	0.03	0.00	0.00
CCL22	0.42	-0.131	-0.32	-0.09	0.41	0.44	0.43
G-CSF	0.05	-0.079	-0.41	-0.43	0.05	0.00	0.00
CCL11	0.29	0.221	-0.27	-0.16	0.30	0.26	0.26
CRP	0.19	0.367	-0.11	-0.53	0.20	0.09	0.11
CXCL10	0.57	0.121	-0.05	0.46	0.57	0.68	0.67
Periostin	-0.18	-0.318	-0.31	-0.08	-0.18	-0.08	-0.08
IL-1ra	-0.38	-0.627	0.52	-0.28	-0.40	-0.39	-0.41
Explained Variance in Y	19.7%	6.9%	19.5%	23.3%	19.8%	17.7%	17.4%

$\Rightarrow$ variable selection on X & Y excludes 5 proteins and $BrCH_3$

$\Rightarrow$ yields maximal Y explained variance by C'''_{1X}

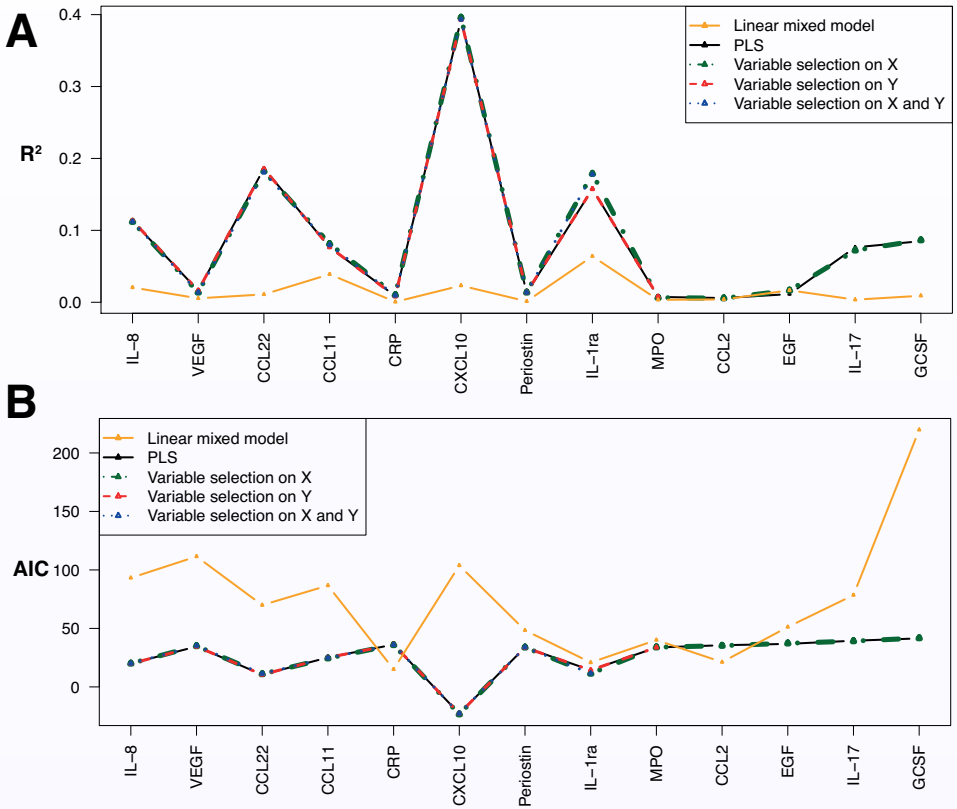
Multi level PLS analyses of PISCINA data



- Despite correlation, 3 exposures selected
  - Bromoform appears less influential (not selected in sPLS-X)
  - Variance is heterogeneously explain across proteins (e.g. IP10, MDCCC, and IL1ra) which are selected in sPLS (on Y)
  - Variable selection on X and Y selects 3 exposures and 8 proteins

Multi level PLS analyses of PISCINA data

- Comparison with mixed models (investigating protein separately):



$\Rightarrow$ accounting for the correlation across proteins (multivariate Y) is improving the model's fit

$\Rightarrow$ All PLS variants have similar performances

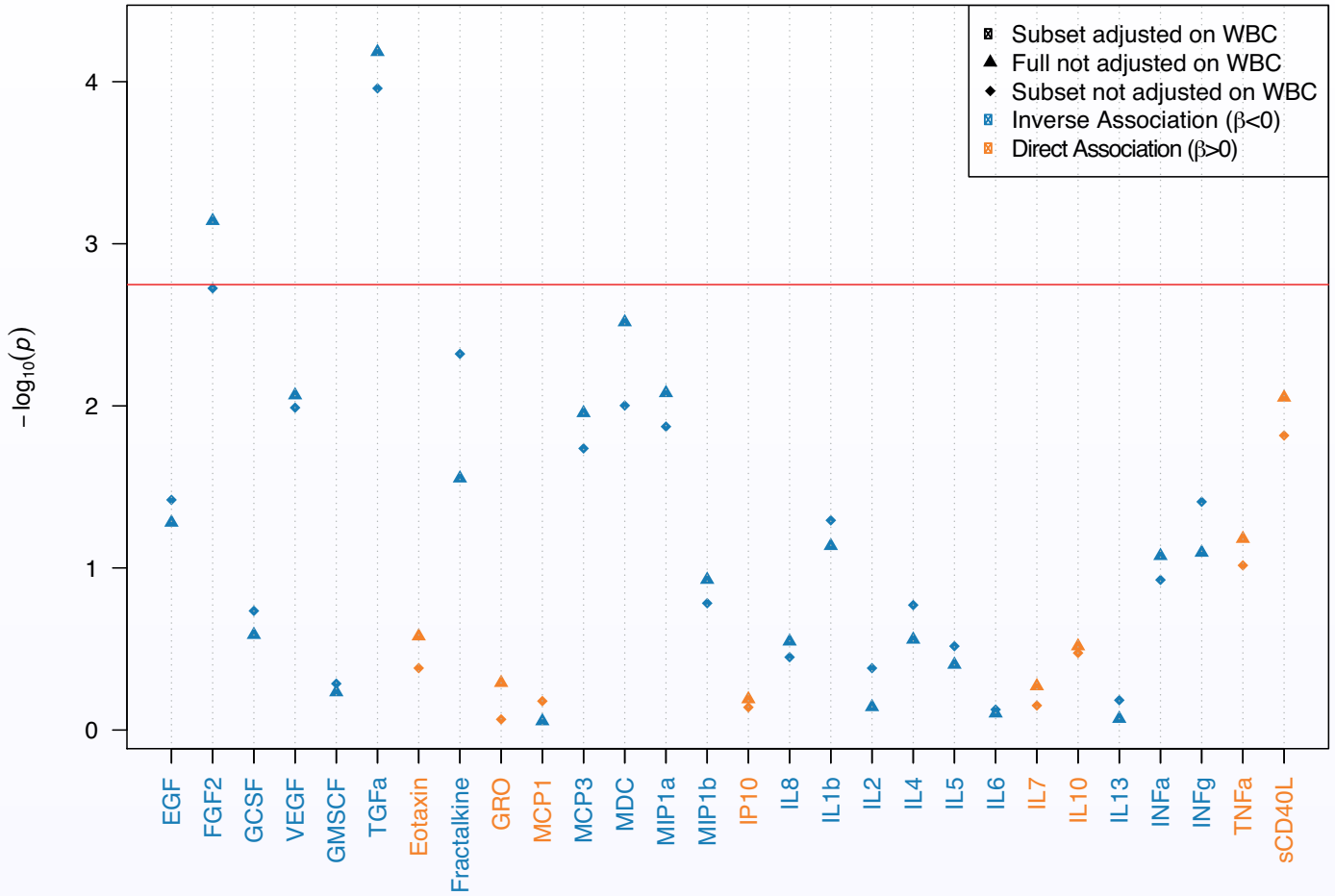
$\Rightarrow$ Efficient prioritisation tool

Lymphoma cases by subtypes and TtD

- EnviroGenoMarkers: a multi-OMIc study of NHL
  - Two contributing cohorts: EPIC Italy, and NHSDS
  - Transcriptomics, Proteomic (N=28) data available
- Four subtypes were identified:
  - B-cell Chronic Lymphatic Leukemia (BCLL): 14.8%
  - Diffuse Large B-cell Lymphoma (DLBCL): 15.6%
  - Follicular Lymphoma (FL): 14.4%
  - Multiple Myeloma (MM): 27.4%
- Study population:

Subtype	TtD<6	TtD>6	**Total**
BCLL	15	24	**39**
DLBCL	18	23	**41**
FL	18	20	**38**
MM	42	30	**72**
Others	41	32	**73**
Total	**93**	**97**	**263**

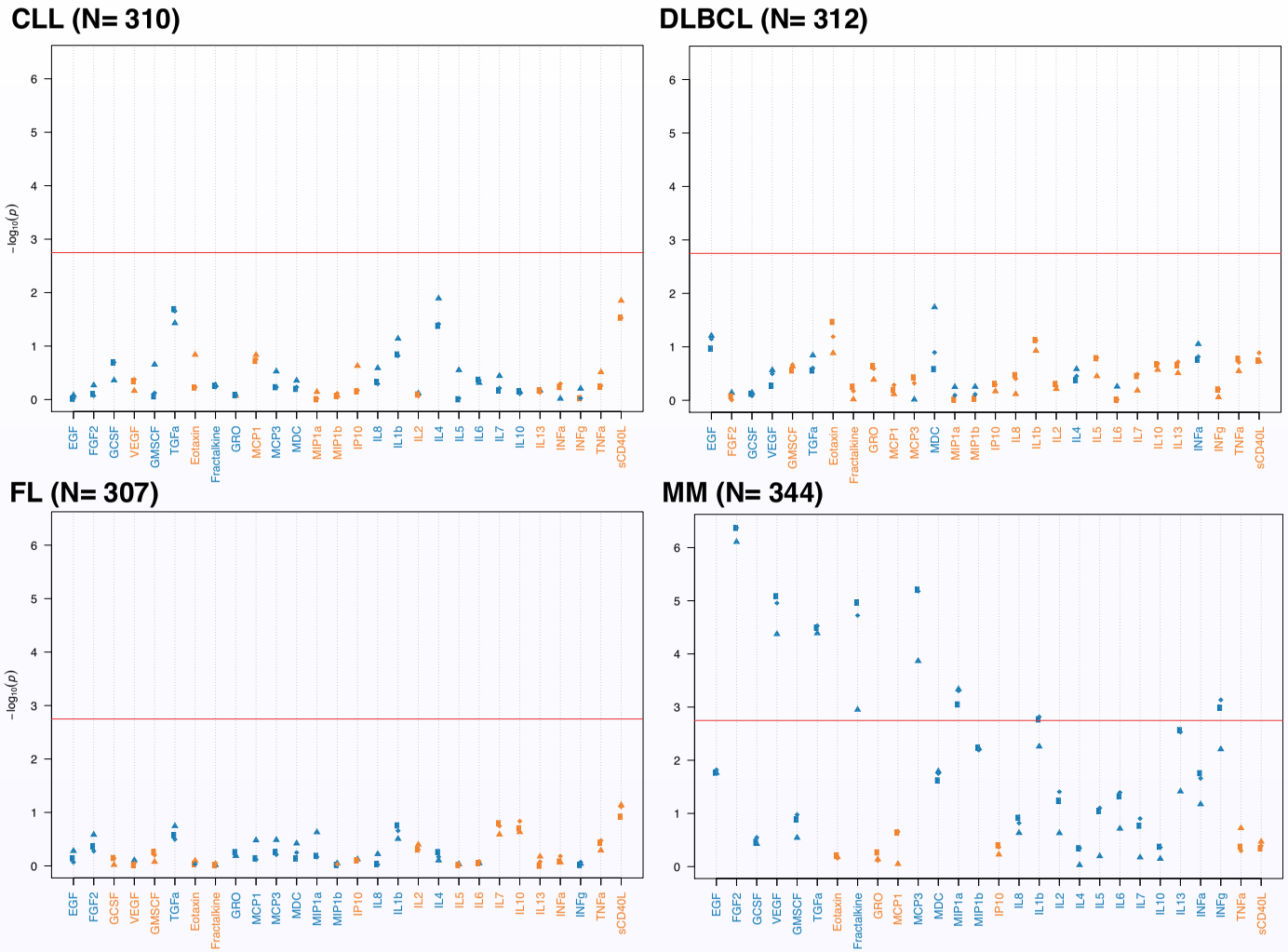
Analysis of all BCL cases



$\Rightarrow$ Two Bonferroni significant associations involving FGF2 & TGFα

$\Rightarrow$ weak effect of WBC adjustments

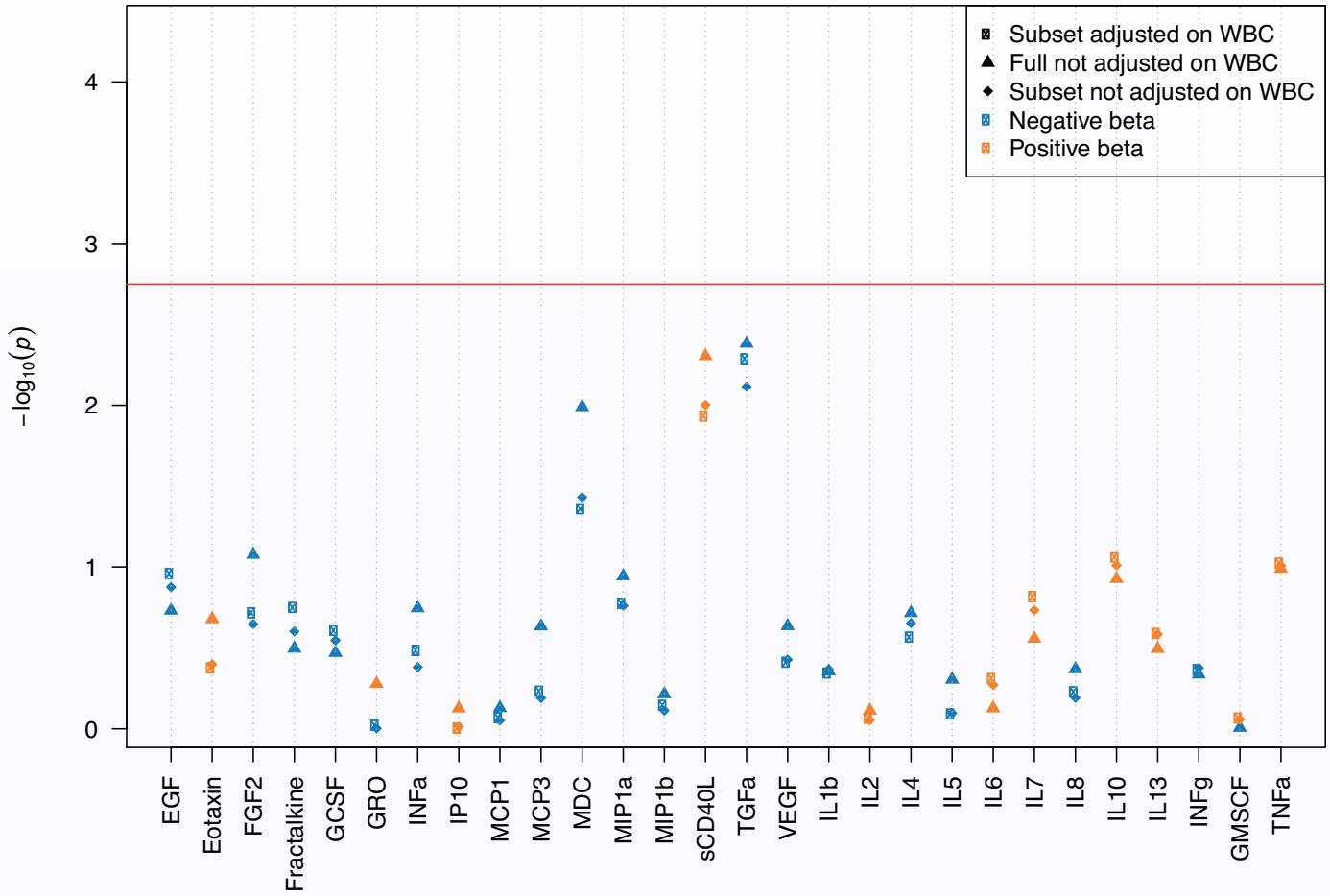
Histological subtype analyses



$\Rightarrow$ 8 (strong) and inverse associations for MM

$\Rightarrow$ no association for the other subtypes

All BCL excluding MM



$\Rightarrow$ Both BCL-related associations lose significance upon exclusion of MM cases

$\Rightarrow$ MM may have driven the BCL associations

PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset?
  - Is the functional grouping relevant to the disease?
  - Are there groups (and proteins within each group) more associated to disease?

One Million $ question: which models????

PLS analyses: Rationale and plan

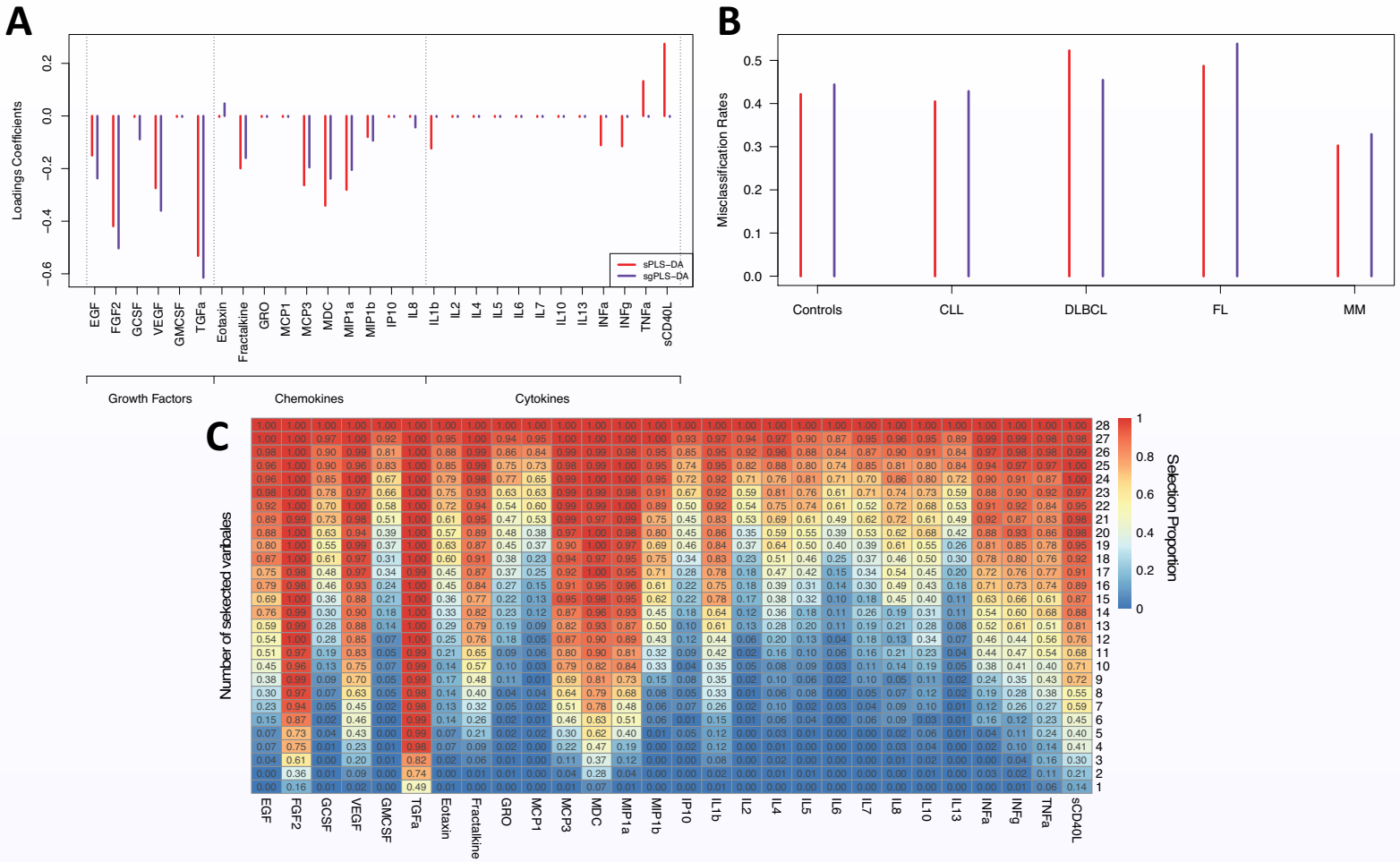
- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset? – (s)PLS
  - Is the functional grouping relevant to the disease? – gPLS
  - Are there groups (and proteins within each group) more associated to disease? – sgPLS

One Million $ response

PLS analyses: Rationale and plan

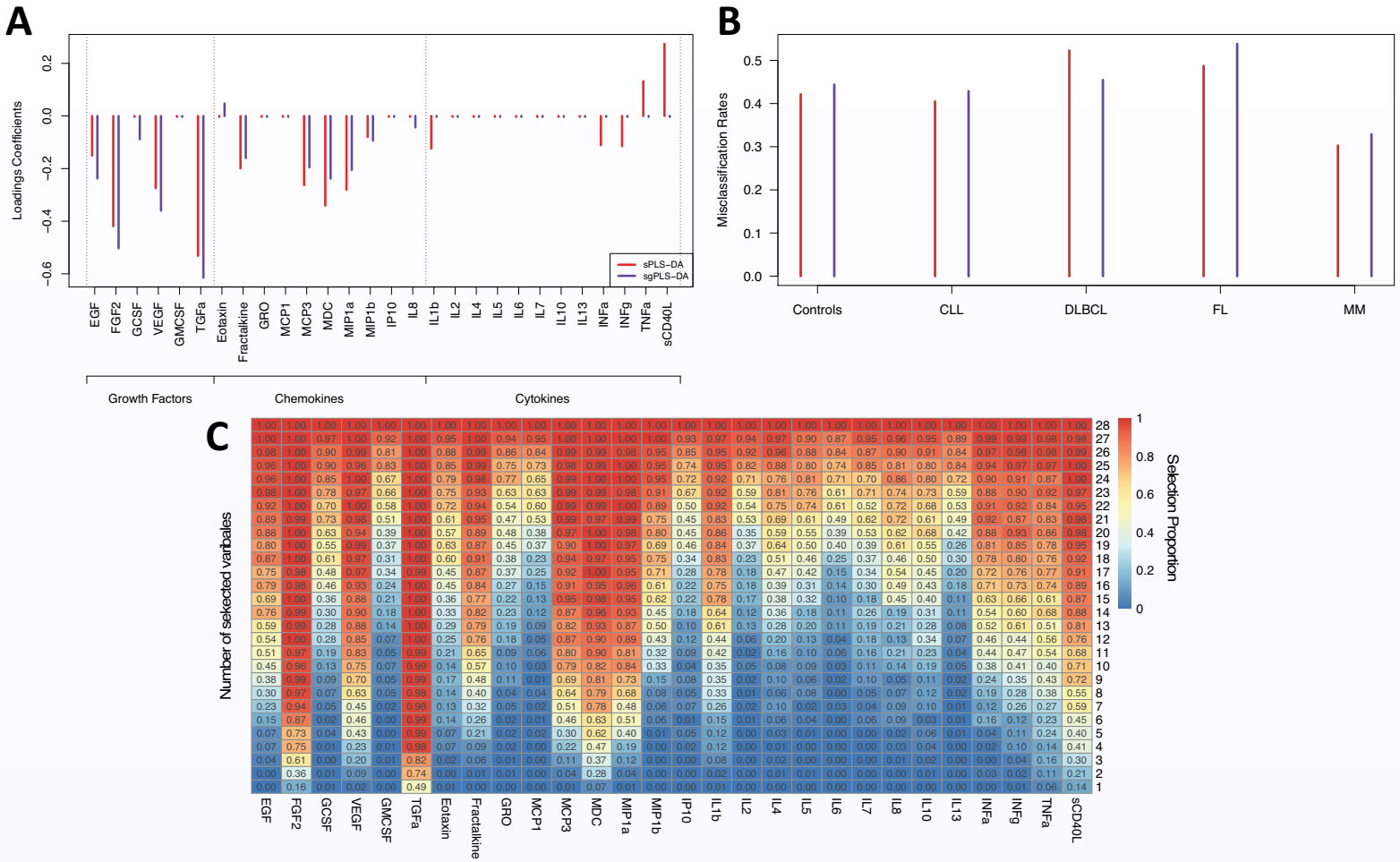
- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset? – (s)PLS
  - Is the functional grouping relevant to the disease? – gPLS
  - Are there groups (and proteins within each group) more associated to disease? – sgPLS
- Analytical Plan: all PLS variants to analyse
  - All BCL
  - Each subtype separately
  - In cases only: the time to diagnosis

PLS analyses: All BCL



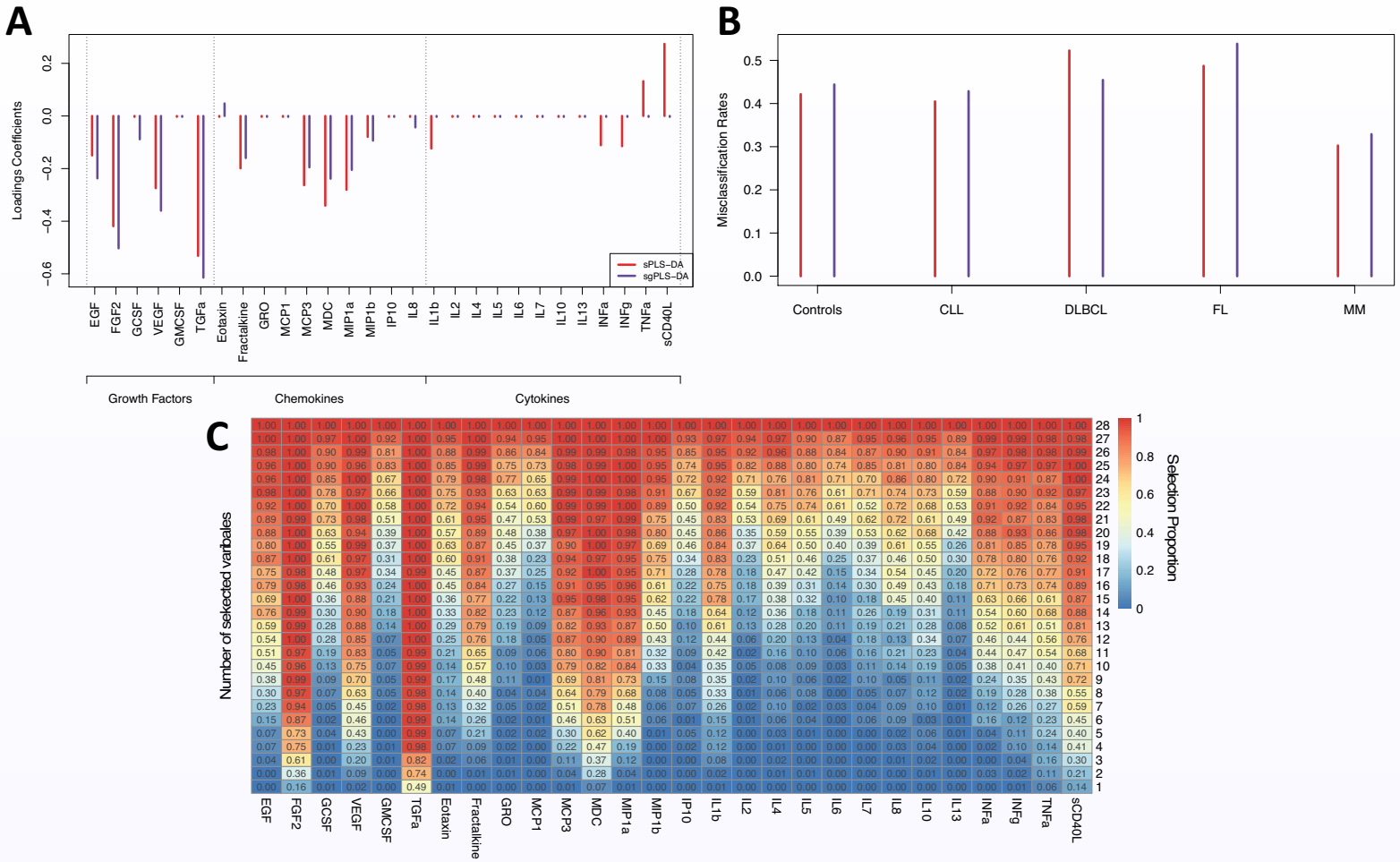
- sPLS mainly selects variables in GF an chemokines groups
- Two cytokines proteins selected with larger loadings (TNF-α, sCD40)
- sgPLS selects the two group with more non zero loadings

PLS analyses: All BCL



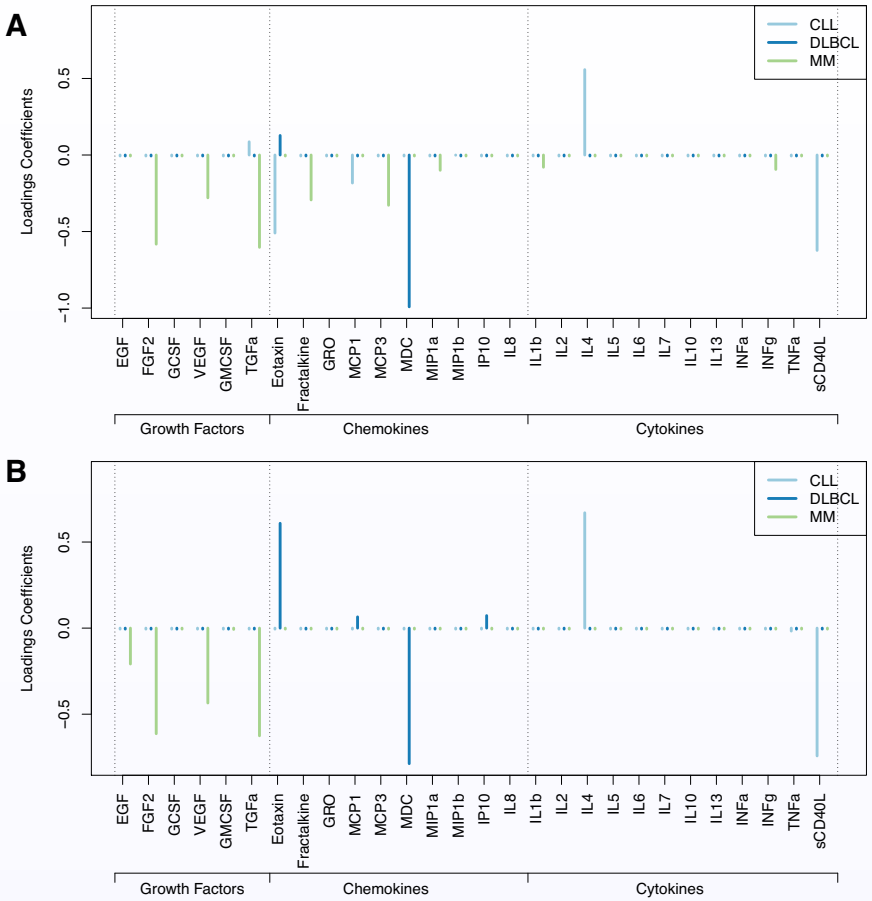
- sPLS and sgPLS yield comparable misclassification rates (unimportant exclusion of cytokines)
- Better misclassification rates for MM

PLS analyses: All BCL



- Assessing the sensitivity to calibration via stability analyses
- The largest loadings are the first and most frequently selected (sPLS)

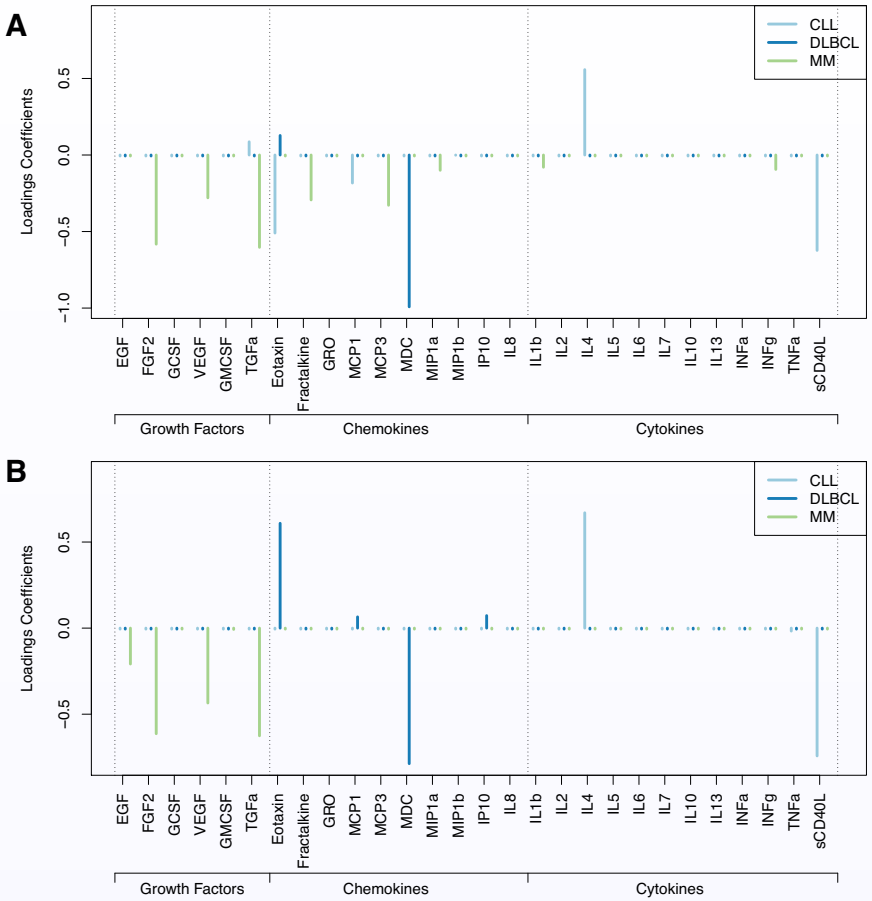
PLS analyses: subtype analyses



sPLS analyses select:

- **MM**: proteins mainly in chemokines and growth factors
- **CLL**: chemokines and cytokines (though only 2/12 proteins)
- **DLBCL**: 2 chemokines are selected

PLS analyses: subtype analyses

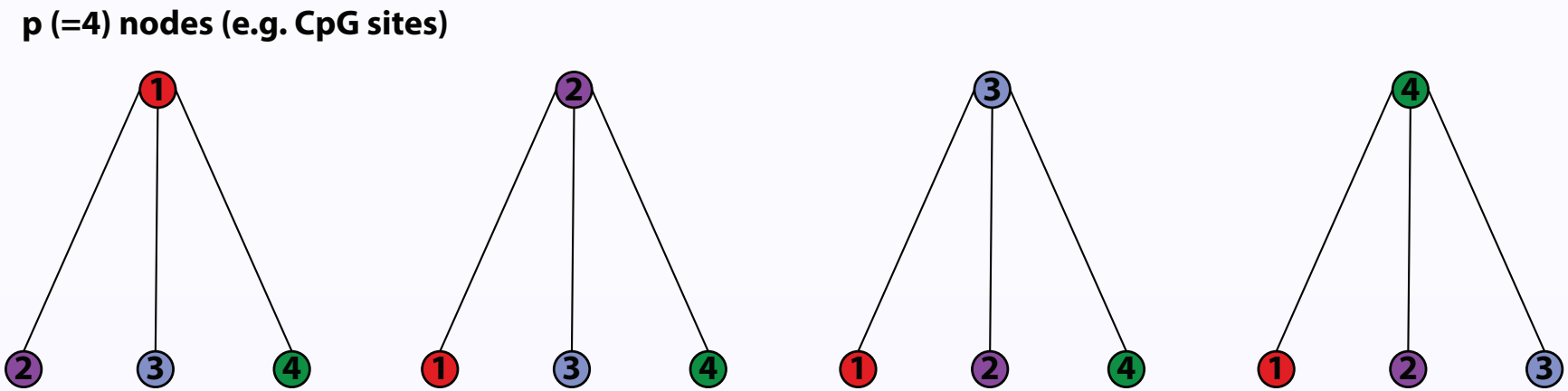


sgPLS analyses select:

- **MM**: growth factors and within the group the same variables as sPLS
- **CLL**: cytokines and both the sPLS proteins
- **DLBCL**: Chemokines are selected (including the the 2 sPLS proteins)

Network models: introduction and challenges

- Data: Variables of interest are the p nodes (e.g. selected CpG sites)
- Aim: Describe and summarise the relationships among the p nodes
- Main steps:
  1. Define a measure for the relationship
  2. Identify the most important ones
- Challenge: dimensionality

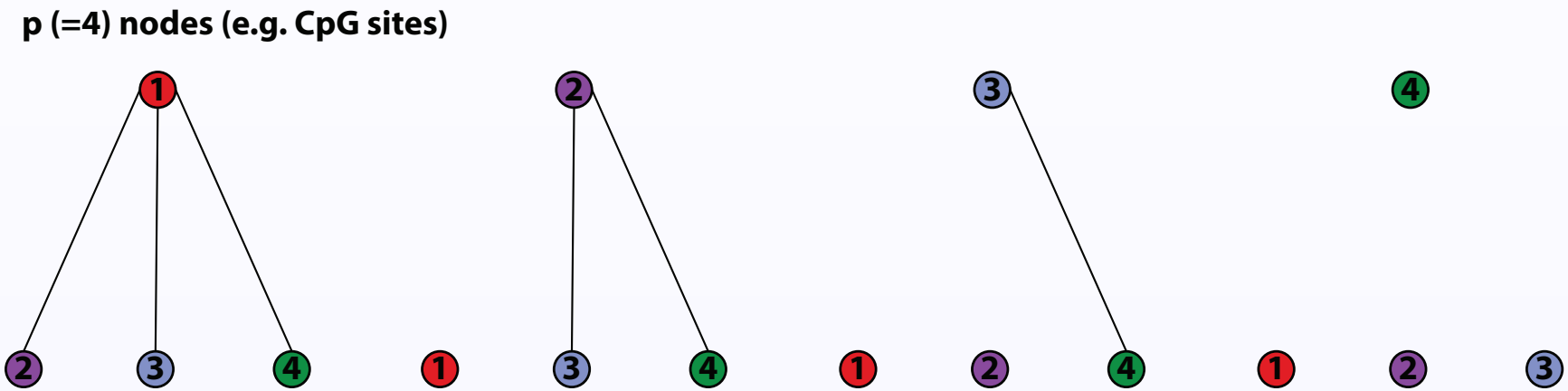


$\Rightarrow$ each of the p nodes is investigated against the $p-1$ others

$\Rightarrow p \times (p-1)$ relationships

Network models: introduction and challenges

- Data: Variables of interest are the p nodes (e.g. selected CpG sites)
- Aim: Describe and summarise the relationships among the p nodes
- Main steps:
  1. Define a measure for the relationship
  2. Identify the most important ones
- Challenge: dimensionality



$\Rightarrow$ accounting for symmetry, fewer links to investigate

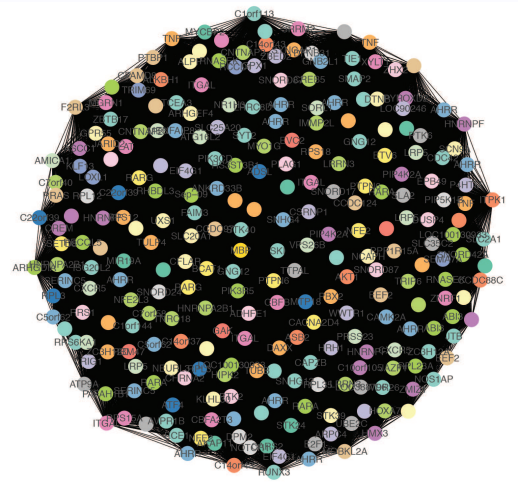
$$\Rightarrow (p-1) + (p-2) + \cdots + 1 = \frac{(p) \times (p-1)}{2} \text{ relationships}$$

Network models: introduction and challenges

- Data: Variables of interest are the p nodes (e.g. selected CpG sites)
- Aim: Describe and summarise the relationships among the p nodes
- Main steps:
  1. Define a measure for the relationship
  2. Identify the most important ones
- Challenge: dimensionality
- For 265 CpG sites: 35,425 edges

Network models: introduction and challenges

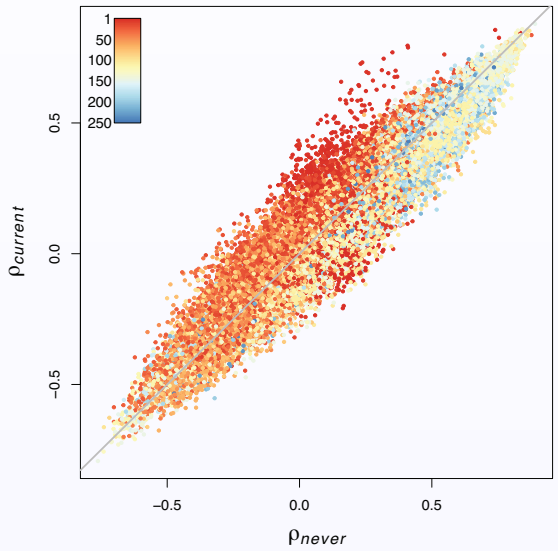
- Data: Variables of interest are the p nodes (e.g. selected CpG sites)
- Aim: Describe and summarise the relationships among the p nodes
- Main steps:
  1. Define a measure for the relationship
  2. Identify the most important ones
- Challenge: dimensionality
- For 265 CpG sites: 35,425 edges



$\Rightarrow$ need to reduce dimensionality and/or perform edge selection

Devising a differential network approach

- Aim 1: can we elucidate how the 265 CpG sites concur to the methylation response to smoking?
- Aim 2: what is the joined role of CpG and transcripts?
- Aim 3: are there signals associated to lung cancer?
- Exploring pairwise correlation (N$\sim$ 35,000 pairs)



$\Rightarrow$ pairs involving smoking-related CpG site have stronger correlations in current smoker

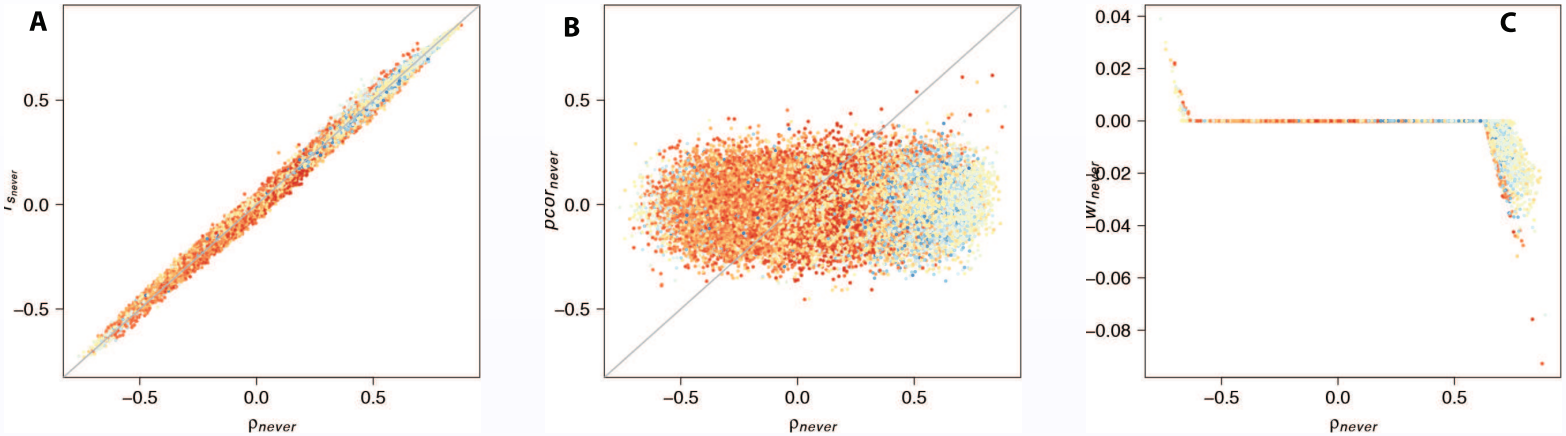
$\Rightarrow$ suggestive of a differential correlation patterns

Devising a differential network approach

- Overall approach in differential network inference:
  - Step 1: infer individual networks (i.e. in smokers and non smokers separately)
  - Step 2: define the metrics for the change in correlation between the two populations
  - Step 3: Identify significant changes in correlations
- Main statistical challenges:
  1. Define the correlation metrics
  2. Devise the calibration procedure to control the sparsity of the network
  3. How to achieve significance assessment

Defining the pairwise metric of interest

- Exploring different correlation metrics

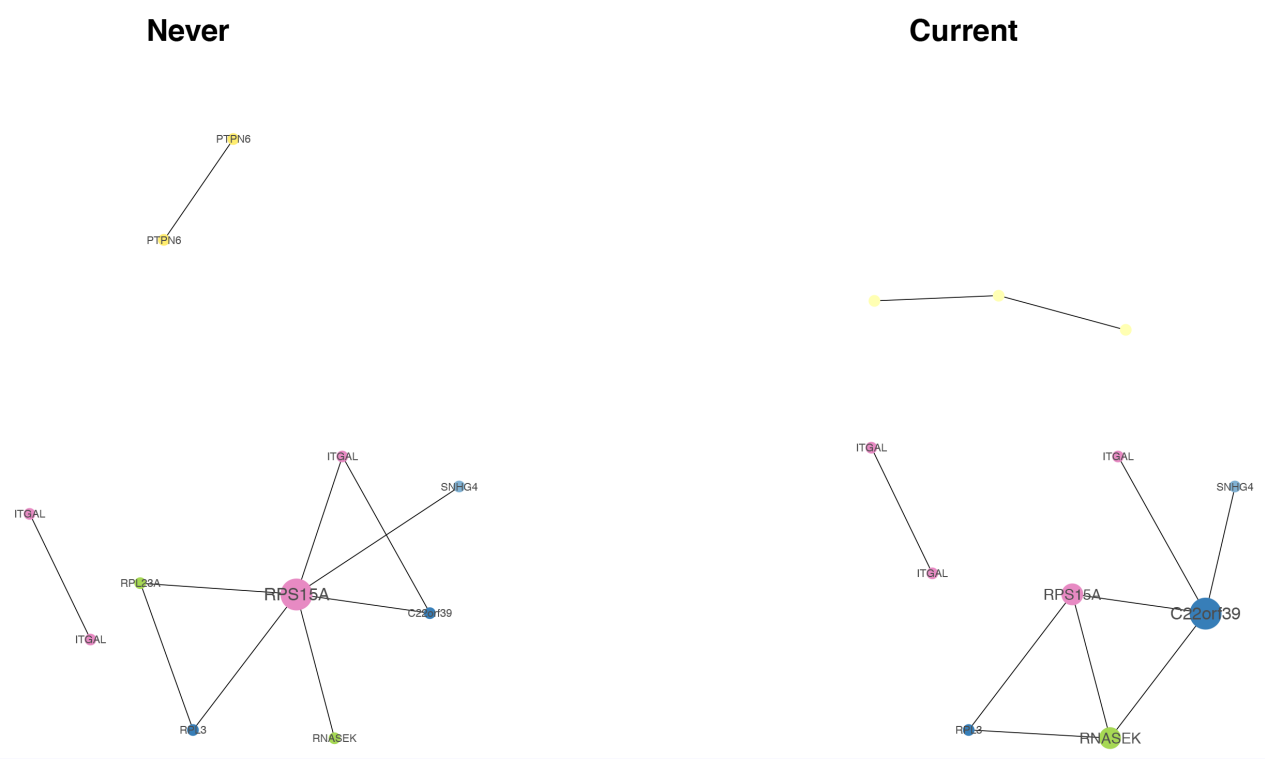


- Conclusions
  - Pearson and Spearman correlation are consistent (A)
  - Partial (shrinkage) correlation destroys the correlation pattern (owing to the preselection of CpG sites, that are by nature partially redundant) - (B)
  - Inverse variance from the GLASSO shrinks correlations <0.5 (C)

$\Rightarrow$ for targeted analyses, partial correlation cannot be used

Representation of individual networks

- Individual networks in never (left) and current (right) smokers
- The 10 strongest correlations are represented

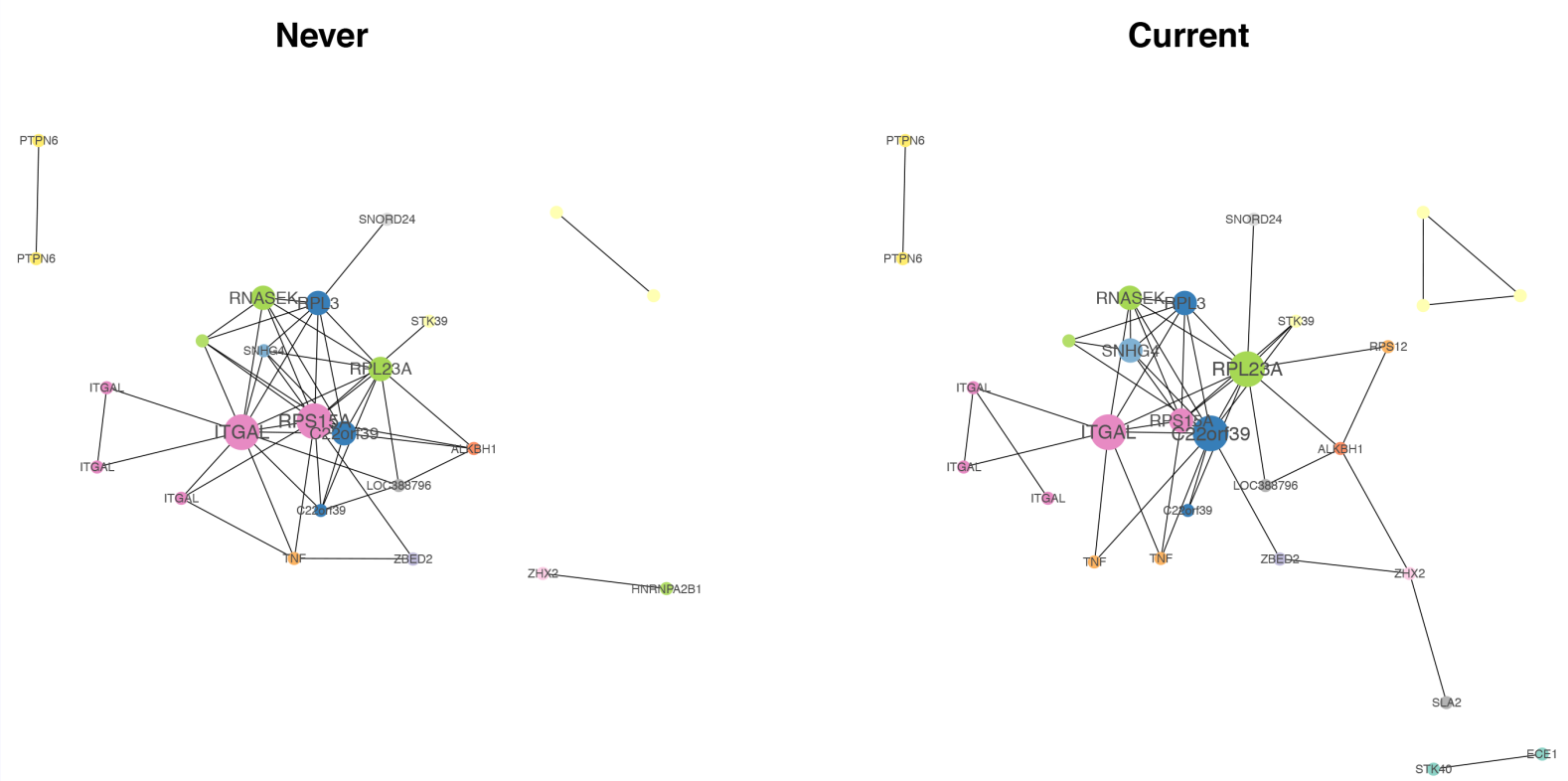


$\Rightarrow$ out of the 11 nodes involved, 7 are common in both populations, including those with high degree

$\Rightarrow$ apparent differences between sub-populations

Representation of individual networks

- Individual networks in never (left) and current (right) smokers
- The 50 strongest correlations are represented

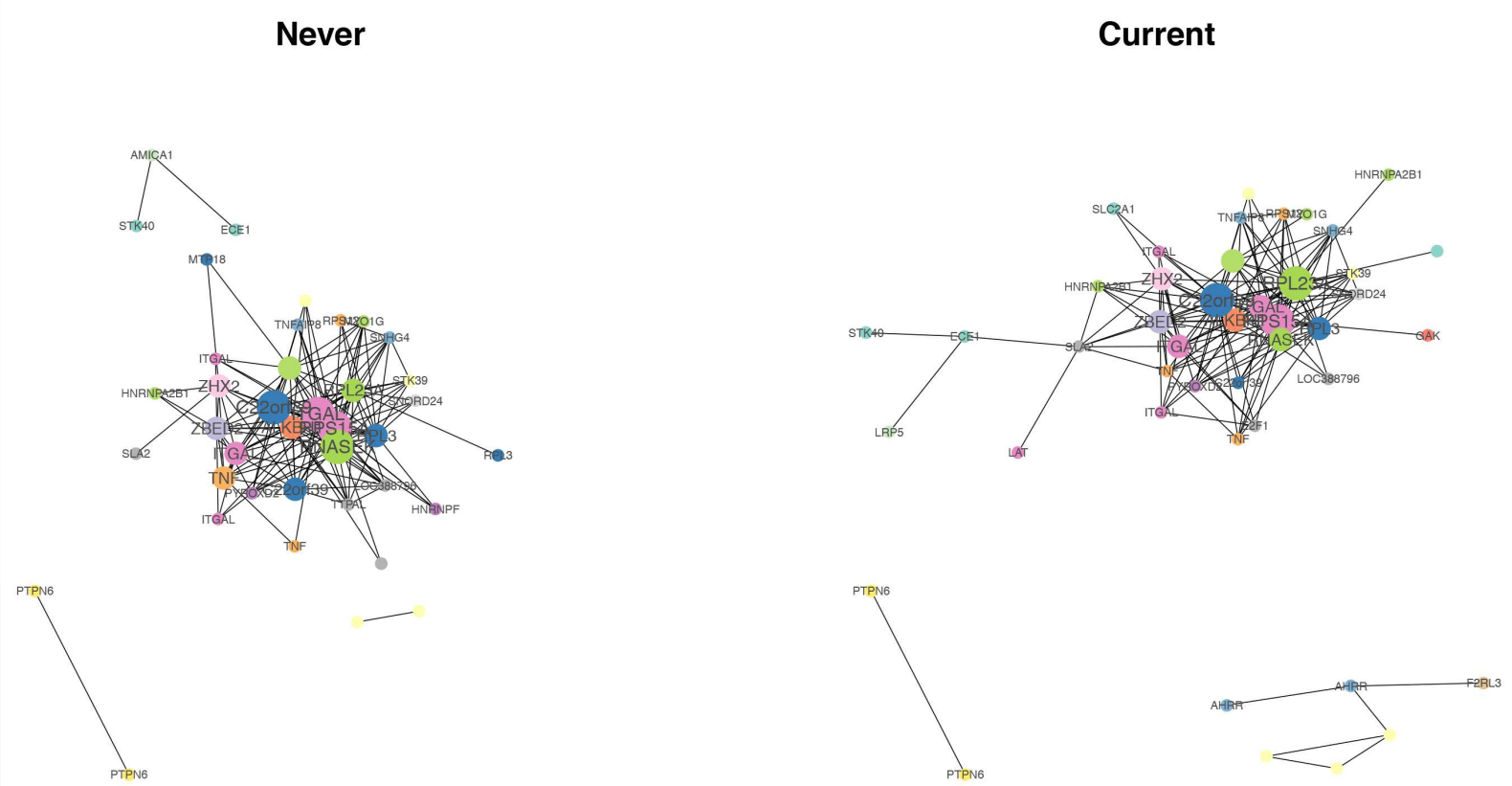


⇒ networks composition is consistent in the subpopulations

⇒ structural differences emerge in denser networks

Representation of individual networks

- Individual networks in never (left) and current (right) smokers
- The 150 strongest correlations are represented



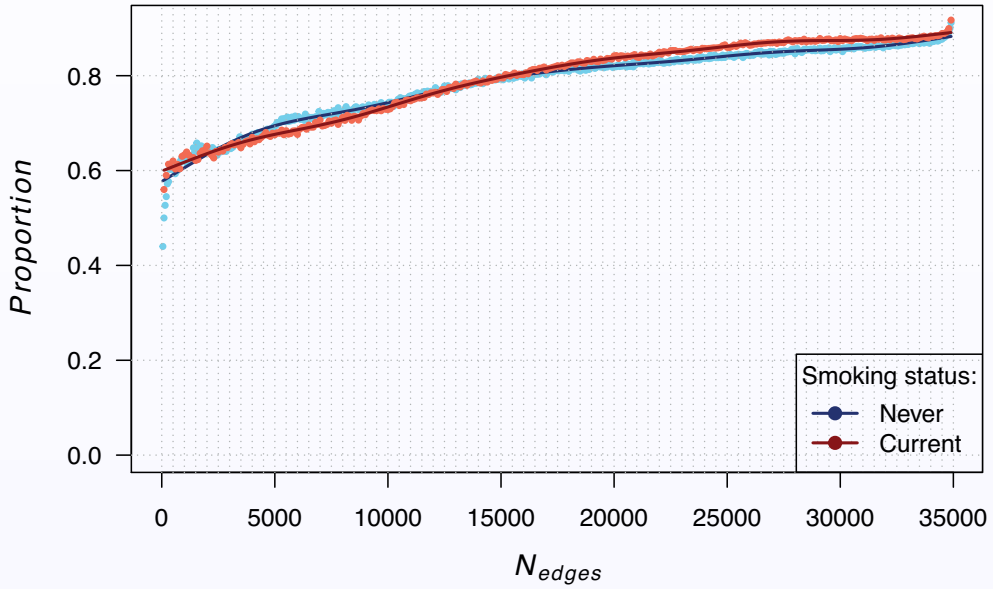
⇒ how many edges to retain?

⇒ define a calibration procedure

Network Calibration

- Defining the calibration procedure
  - Classical approaches rely on cross validation optimising prediction
  - Here we are interested in identify robust findings

$\Rightarrow$ Stability analyses (100 replicates of 80% subsamples): investigate the proportion of edges that are systematically selected



$\Rightarrow$ proportion ranging from 0 to 1

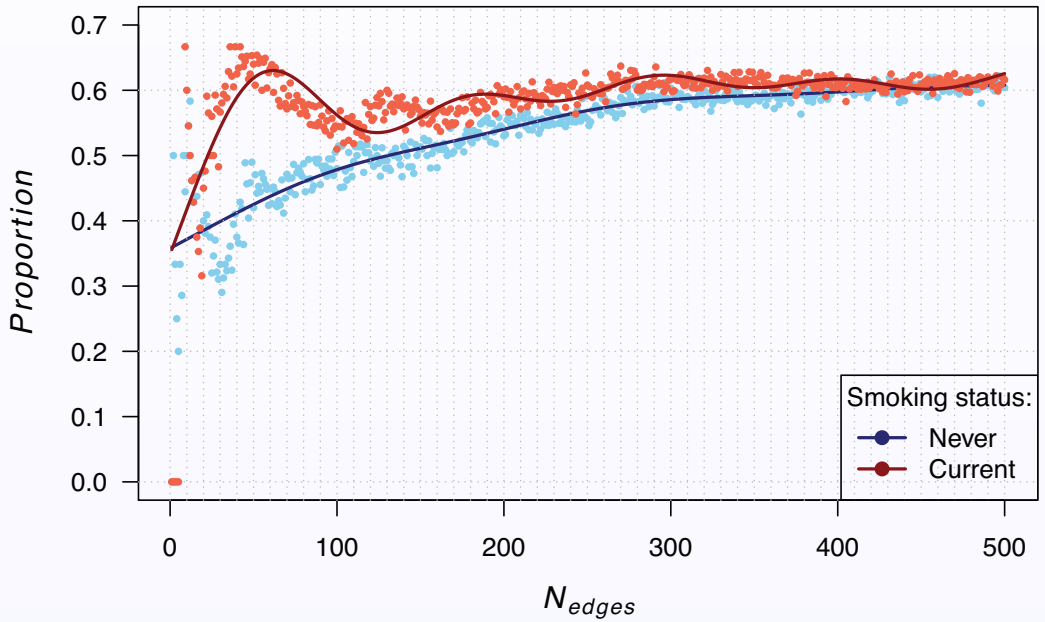
$\Rightarrow$ increasing function of # edges

$\Rightarrow$ identify the optimal #edges

Network Calibration

- Defining the calibration procedure
  - Classical approaches rely on cross validation optimising prediction
  - Here we are interested in identify robust findings

$\Rightarrow$ Stability analyses (100 replicates of 80% subsamples): investigate the proportion of edges that are systematically selected

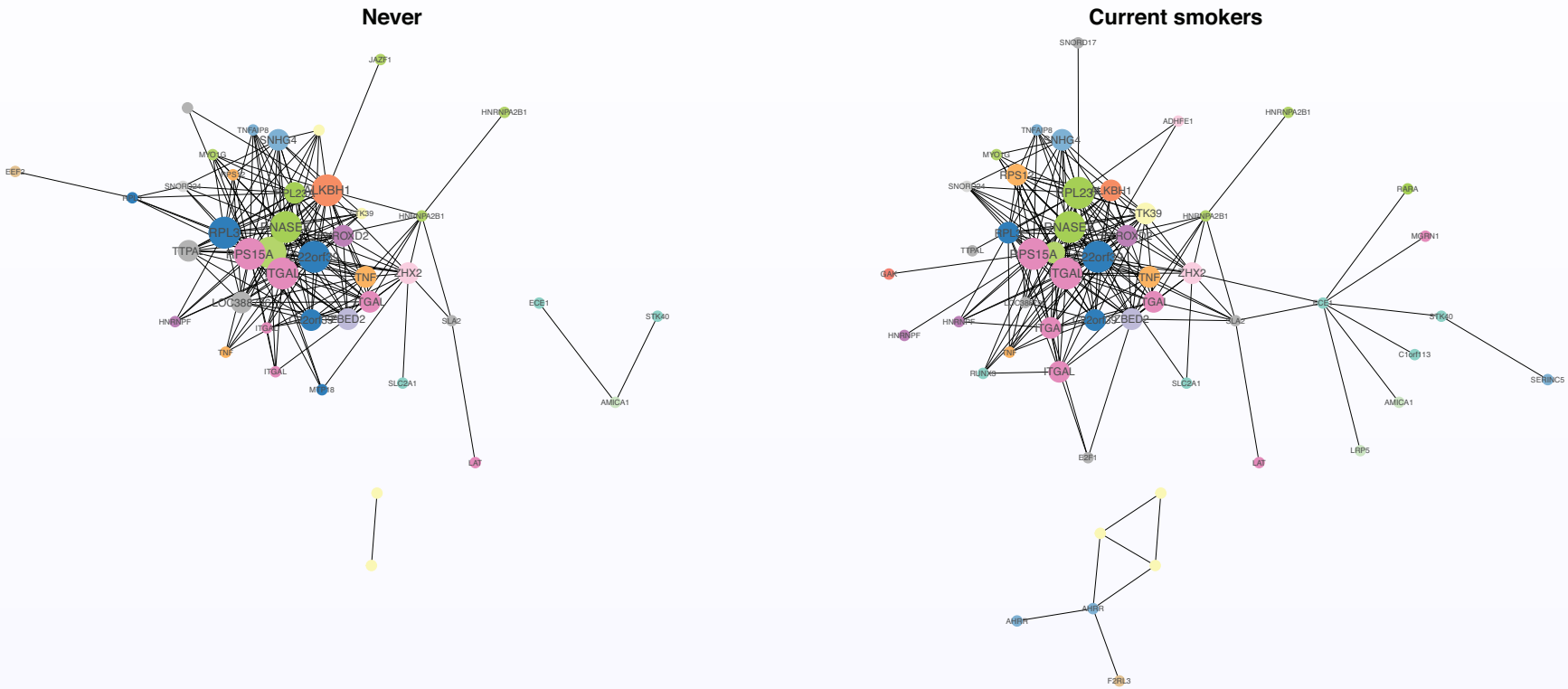


$\Rightarrow$ after 300 edges: moderate increases in the selection proportion

$\Rightarrow$# 300 edges seem to provide the good balance between sparsity and stability

Calibrated individual networks

- Resulting calibrated individual networks (N=300 edges)

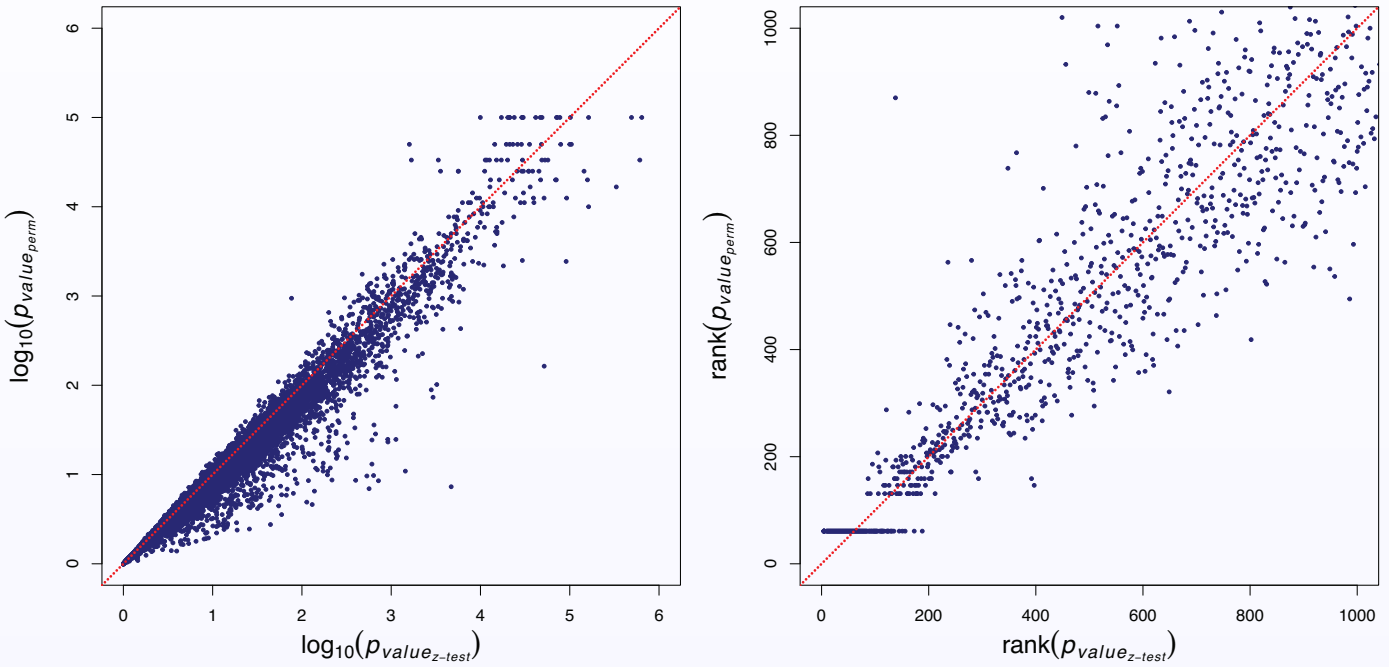


⇒ apparent differences between the two networks

⇒ which are the ones that are significant?

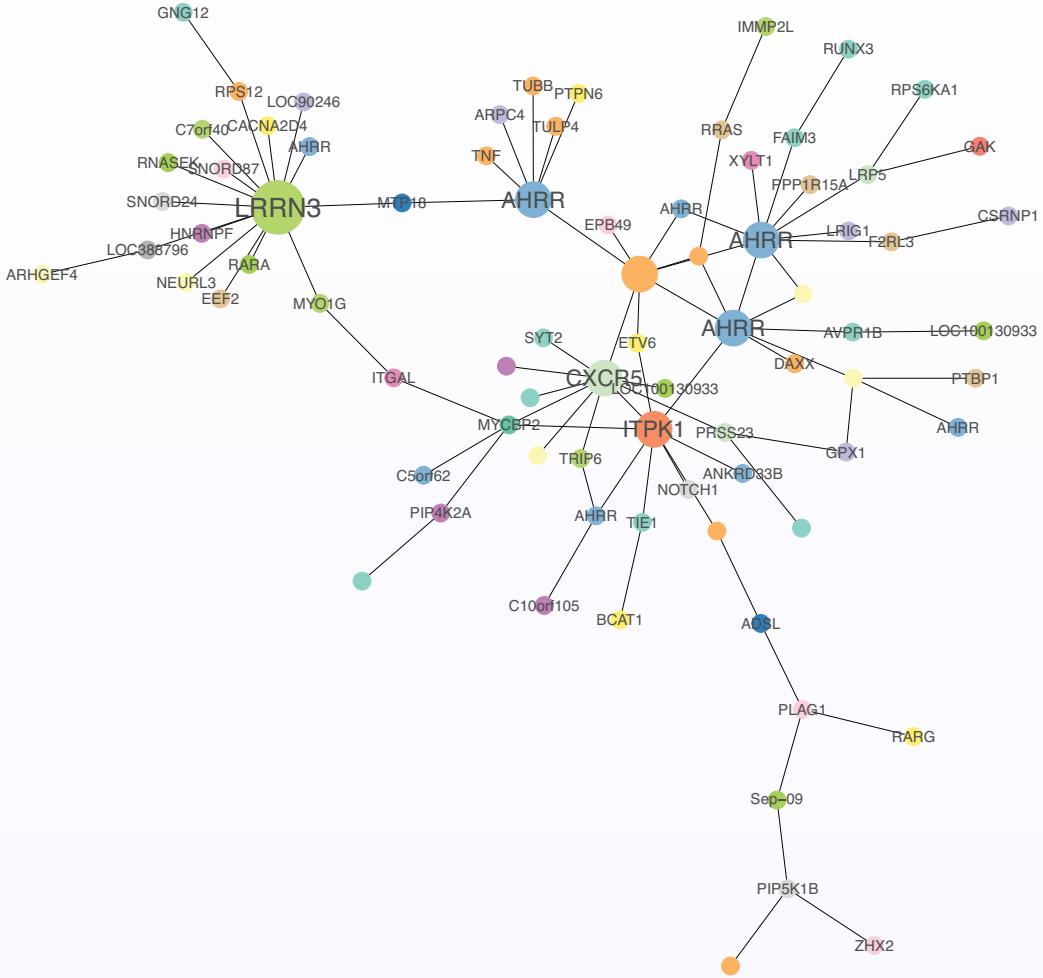
Significance assessment

- Testing the differences in correlation
  - Classical approaches rely on permutation tests
  - Here we have full correlation coefficients use of an exact test relying on Fisher transformation (Half normal distribution under the null).



$\Rightarrow$ exact test is in line with permutation test (without the resolution issue)

Representation of the Differential network (N=100 edges)



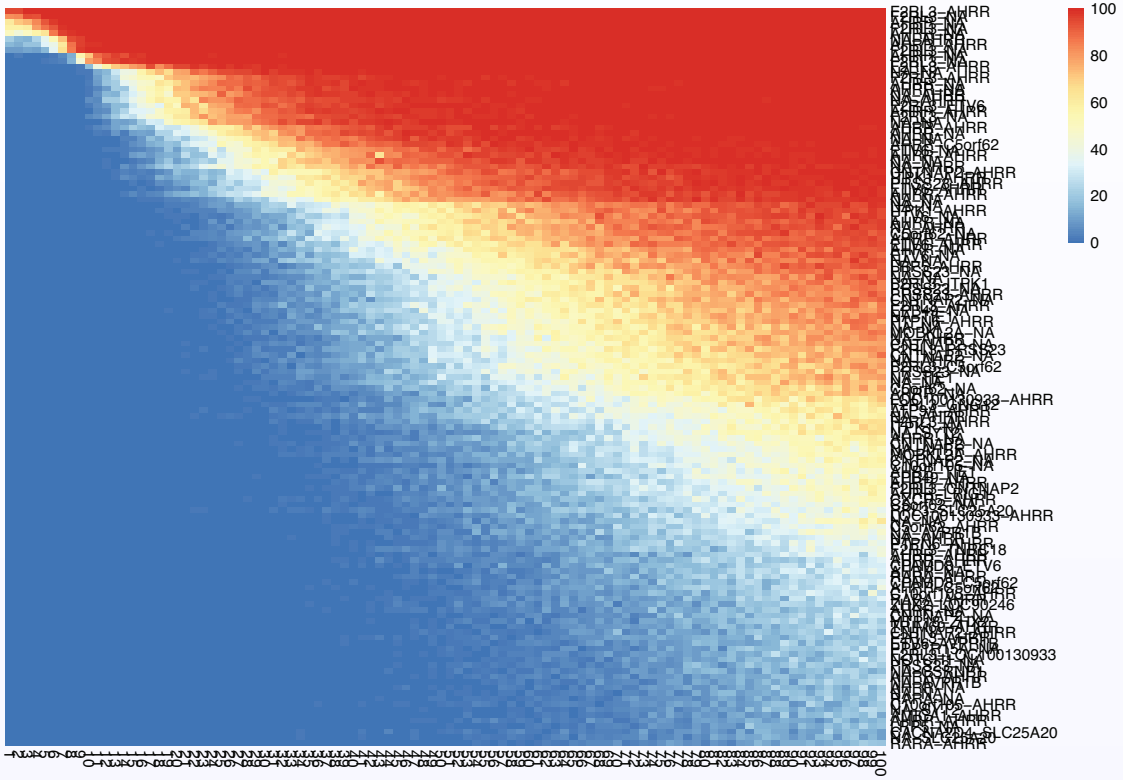
$\Rightarrow$ modular structure of the network

$\Rightarrow$ central role of LRRN3 and AHHR

Calibration of the Differential network

- Egdes can be selected on their p-values
- Here again, we want to prioritise generalisable findings

$\Rightarrow$ calibration through stability

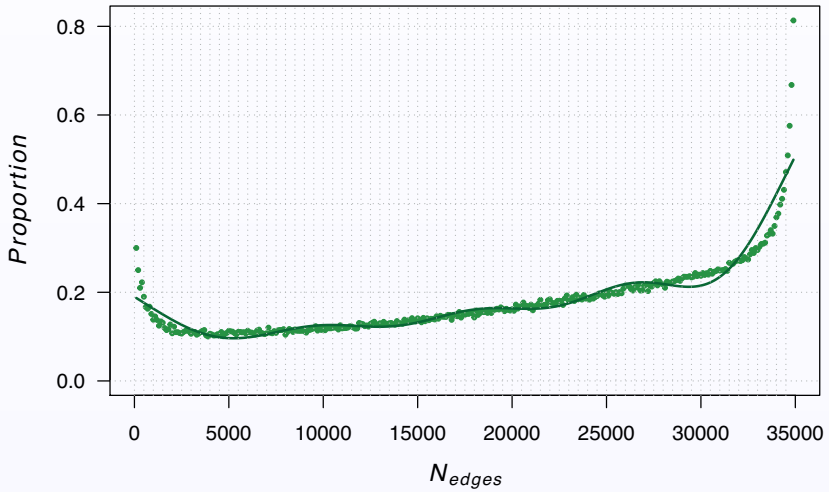


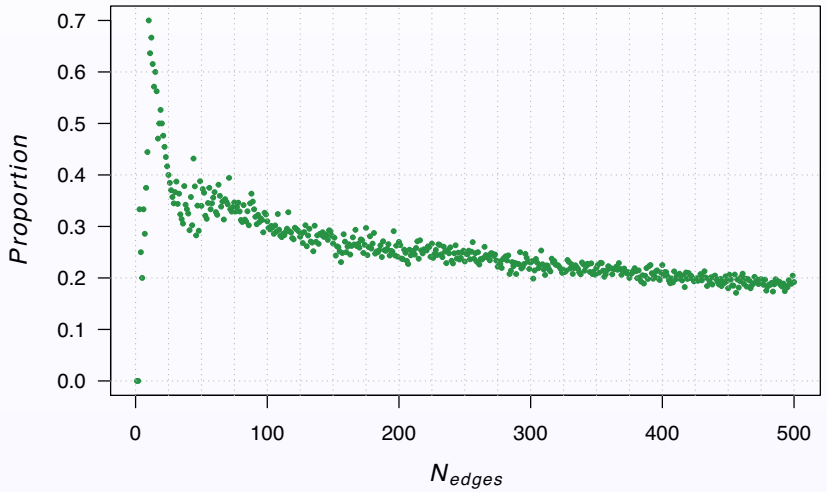
$\Rightarrow$ edges with high selection probability are the first ones to be selected

Calibration of the Differential network

- Egdes can be selected on their p-values
- Here again, we want to prioritise generalisable findings

$\Rightarrow$ calibration through stability





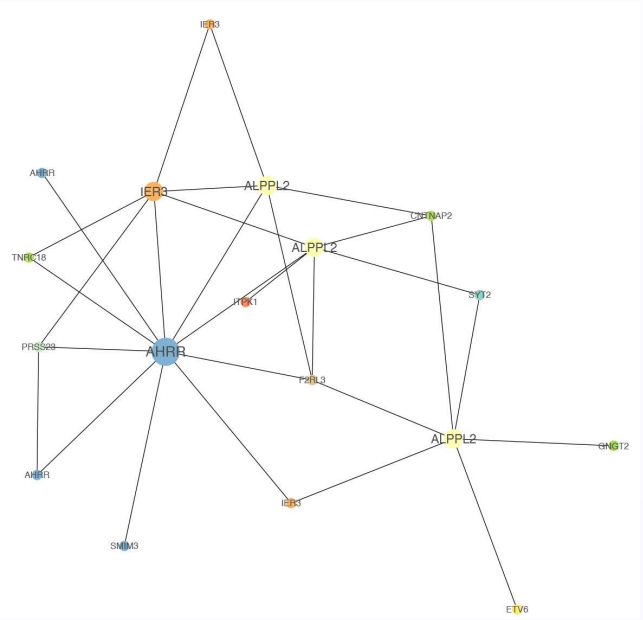
$\Rightarrow$ selection proportion peaks for N=10 edges

Calibrated Differential network

- Egdes can be selected on their p-values
- We want to prioritise generalisable findings

$\Rightarrow$ calibration through stability

EPIC



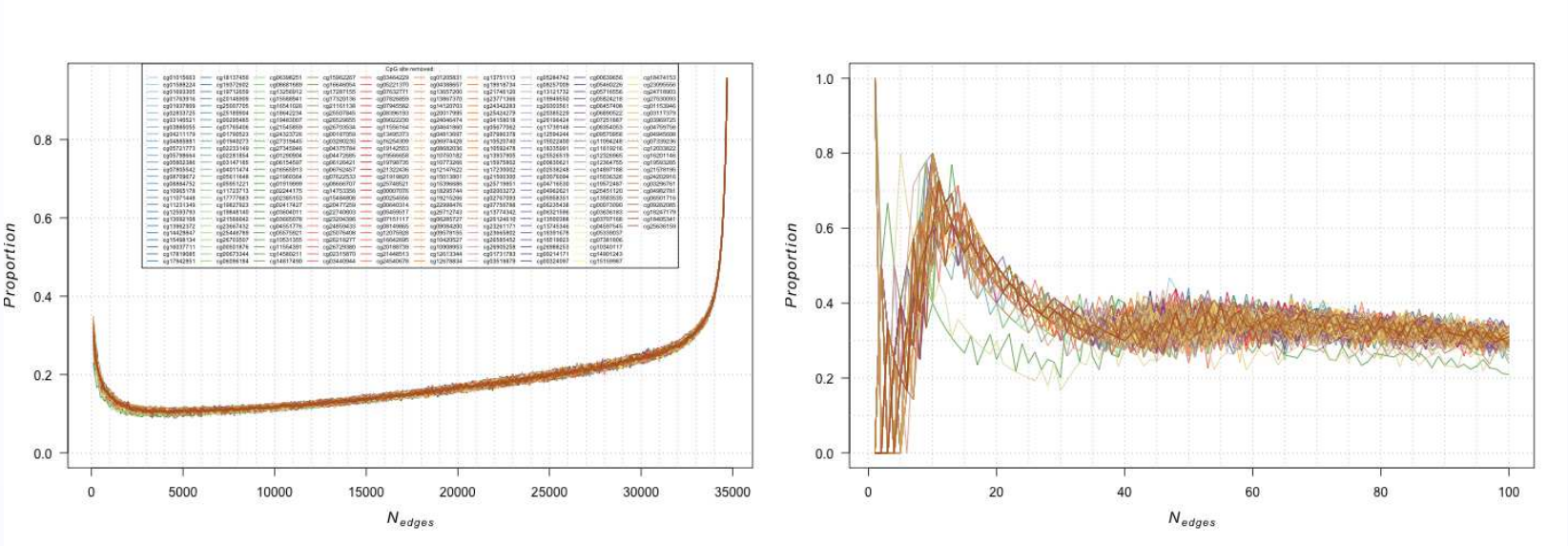
NOWAK

$\Rightarrow$ In both studies central role of AHHR and similar topology

$\Rightarrow$ In NOWAK study, LRRN3 also appears central (but separate from AHRR)

Perturbation analyses:response to node removal

Method: assess network stability for 265 differential networks (with one CpG removed)



$\Rightarrow$ Overall, little changes in stability across the 265 differential networks

$\Rightarrow$ 2 slightly less stable differential networks: those excluding AHRR_cg05575921 and F2RL3_cg03636183

$\Rightarrow$ these are the most central nodes

OMICs Integration: multi-omic network

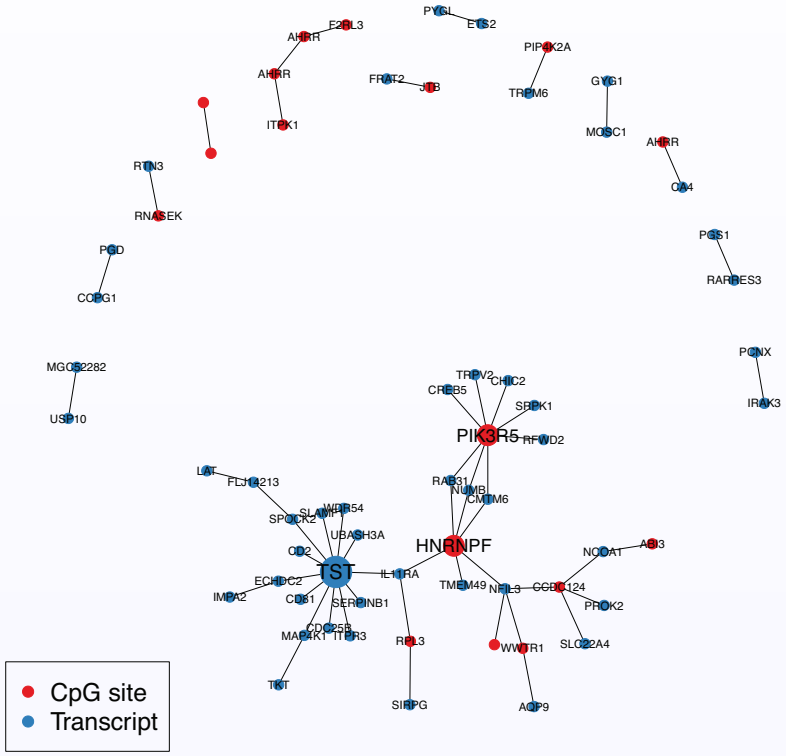
- Integration: including both smoking-related transcripts (N=426) and CpG sites (N=265) in the network

$\Rightarrow$ 239,086 edges to explore

OMICs Integration: multi-omic network

- Integration: including both smoking-related transcripts (N=426) and CpG sites (N=265) in the network

$\Rightarrow$ 239,086 edges to explore

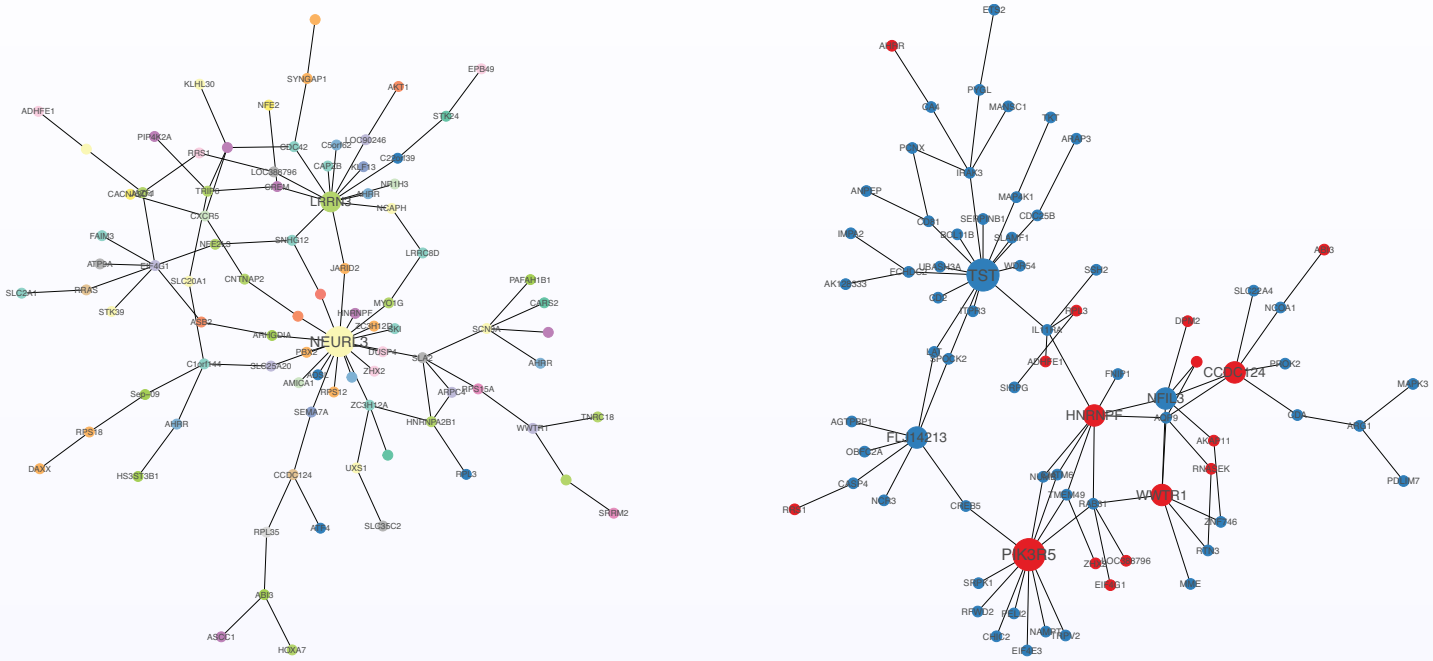


$\Rightarrow$ identification of multi-Omic connected components (module)

- Key statistic: Network stability & Transcript enrichment

Differential networks and lung cancer outcome

- Data: Lung Cancer case control study (NOWAC & EPIC)
- Approach: differential (left) & integrative differential network (right, NOWAC only)

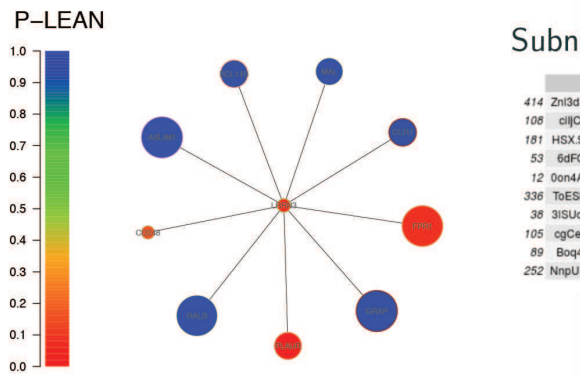


- Results:
  - NEURL3 appears central CpG site
  - When integrating gene expression: vast enrichment in transcripts vs. CpG loci (not identified in regression models)

Subnetwork analyses: improved interpretability

- Aim: Identify sub-networks enriched in outcome-relevant features
- Definition: sub-networks are star-shaped
- Approach: LEAN-R identified the enriched ’stars’

Look LRRN3 subnetwork : Strong p-value but weak PLEAN !



Subnetwork details of p-values

	NodeName	PLEAN	Gene_Symbol	pvalue
414	Znl3dSA4HuOuhEnO4o	0.08165946	LRRN3	4.726392e-20
108	cilJCU2Vf5d4g_nnv0	1.00000000	RALB	1.239907e-02
181	HSX.93nrVTNK66V8yU	0.09419762	FPR1	2.729603e-02
53	6dFQSN.UitTroIYwV4	1.00000000	MAL	8.205835e-02
12	0on4A3Ikg6WUu8SeQo	0.04150967	PLAUR	2.147870e-01
336	ToESkBW4.651_wdJXk	1.00000000	BCL11B	3.176783e-01
38	3ISUq13dBRyH6d0oHc	0.18535079	CD248	6.458755e-01
105	cgCe7bvV7gHISu7vpk	1.00000000	ABLIM1	7.694093e-01
89	Boq4S6WVCIRIfvf738	1.00000000	CCR7	8.567964e-01
252	NnpUIV8gV4uekDdHnU	1.00000000	GRAP	8.883546e-01

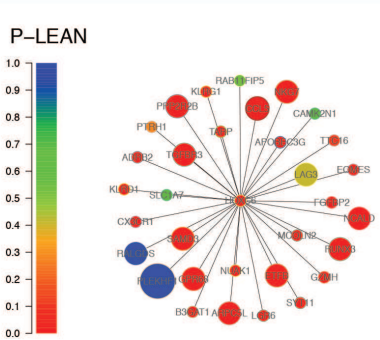
NodeName	Ptilde	mean_bg_p	k	m	pk	pstar	z.score	PLEAN	Gene_Symbol	pvalue
Znl3dSA4HuOuhEnO4o	-18.325470	-5.455686	1	10	4.726392e-20	0.0268	-4.18893042	0.0816594600	LRRN3	4.726392e-20

- Result:
  - Few smoking-associated nodes in the LRRN3 subnetwork

Subnetwork analyses: improved interpretability

- Aim: Identify sub-networks enriched in outcome-relevant features
- Definition: sub-networks are star-shaped
- Approach: LEAN-R identified the enriched 'stars'

Look HOXC6 subnetwork : Weak p-value but strong PLEAN



Subnetwork details of p-values

	NodeName	PLEAN	Gene_Symbol	pvalue
346	ud8S4okRZevuog6d5E	0.0007961468	KLRD1	1.369228e-06
13	0PIPXs9FbUhtvRHdE4	0.0007961468	CXCR1	2.122660e-06
216	KSe4E6HVUdd8.nRGo	0.0007961468	ADRB2	1.817189e-05
418	ZroYe7xJfp5nLJ5f04	0.0007961468	SAMD3	3.116227e-05
381	xAgIQRj3qNUoqgRTgg	0.0007961468	TARP	7.802702e-05
213	KlS.3eurUu.71yu3nk	0.0007961468	PPP2R2B	8.795381e-05
246	NeJemR7prVbXpzgF8	0.0007961468	NKG7	1.492241e-04
176	Hkie59ehuPV.6Ceheg	0.0007961468	APOBEC3G	2.141614e-04
390	xmX900n31rnK4CFXE	0.0007961468	FGFBP2	2.220655e-04
274	ood7Xu0Xd1K7Unt05c	1.0000000000	PLEKHF1	5.069431e-04
100	c5OkDXIrdS_vN4Mu0U	0.0007961468	TGFBR3	5.609440e-04
232	lO0dd56nvT7T57vSoU	0.0007961468	NCALD	6.680052e-04
117	cs6AcOeUd01W1QU3e0	0.0007961468	KLRG1	1.236487e-03
405	ZBTVinpIKOleroB5.k	0.0007961468	RUNX3	2.934214e-03
93	BtSj5dH4h_cpd57ueo	0.0007961468	EOMES	2.996968e-03
316	rgmXcEkpOlcF9d6l4U	0.0007961468	GZMH	4.291649e-03
145	EWW3QVdEonx6QE4ANo	0.0246805508	CCL5	5.849135e-03
91	BqhUlOt0kAeK_NKF0U	0.1155428081	TTC16	1.558687e-02
41	3NLo9Vu7e4eSXI6pIM	0.0400502323	ETFB	1.629153e-02
383	xCEuCq6KvjSpEuhKiE	0.0007961468	LGR6	2.054070e-02
78	9V4lF_Shv9VBdtRfXo4	0.0007961468	B3GAT1	5.979682e-02
182	HU4S4T5XRTjV3UC6_0	0.0007961468	SYT11	9.274594e-02
36	3JneuleWOqV67SB1UA	0.5996138704	SLC1A7	1.053335e-01
421	ZumFCGP95ltKj0V1k	0.3850079857	LAG3	1.144694e-01
279	oSqUQ6LW9eTyeXQdQ4	0.2736832462	PTRH1	1.167481e-01
411	ZqjR.0q4D0pI.5Upc	0.0007961468	NUAK1	1.579770e-01
115	cpREiCCJU5B0tSJ.Ho	0.0007961468	GPR68	2.973813e-01
106	cHkgeE.4a3FiJumnrvo	0.0007961468	ARPC5L	3.338156e-01
315	rCF0505LoT9_I7.Xuo	0.0007961468	MCOLN2	3.624062e-01
138	EJRH6KAvkrq_pKiSk4	0.5236872352	RAB11FIP5	3.860593e-01
157	fLiPRR5tep.GlU5CGo	1.0000000000	RALGDS	4.130992e-01
219	KzhYQpF167cHVUCEOI	0.5981401860	CAMK2N1	8.130708e-01
422	Zz5Cu463p_rg.SqA4o	0.0007961468	HOXC6	9.776866e-01

NodeName	Ptilde	mean_bg_p	k	m	pk	pstar	z.score	PLEAN	Gene_Symbol	pvalue
Zz5Cu463p_rg.SqA4o	-29.55826	-11.95706	12	33	0.0006680052	0	-4.822627	0.0007961468	HOXC6	0.9776866

- Result:
  - Weak smoking association, but many smoking associated node in the sub network