PLS-models in Practice: sparse and sparse group extensions

Lecture 1/3

MSc Health Data Analytics – Computational epidemiology – February 11, 2021

Marc Chadeau-Hyam

m.chadeau@imperial.ac.uk

Imperial College London

Implementing complex study designs: Experimental studies

- Experimental studies: controlling the environment and assess individual variation
    - Inter-individual variability: related to specific individual characteristics (e.g. genome, BMI, behaviours)
    - Intra-individual variability: related to changes between experimental conditions

⇒ need to decompose the sources of variability

⇒ what is the variability of interest?

Implementing complex study designs: Experimental studies

- Experimental studies: controlling the environment and assess individual variation
  - Inter-individual variability: related to specific individual characteristics (e.g. genome, BMI, behaviours)
  - Intra-individual variability: related to changes between experimental conditions

⇒ need to decompose the sources of variability

⇒ the intra-individual variability captures the response to the experimental changes

Implementing complex study designs: Experimental studies

- Experimental studies: controlling the environment and assess individual variation
  - Inter-individual variability: related to specific individual characteristics (e.g. genome, BMI, behaviours)
  - Intra-individual variability: related to changes between experimental conditions

⇒ need to decompose the sources of variability

⇒ the intra-individual variability captures the response to the experimental changes

- 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)
  - OMICs data: proteins (N=13), Metabolites ($N \sim 6{,}000$), Transcripts ($N \sim 30K$)

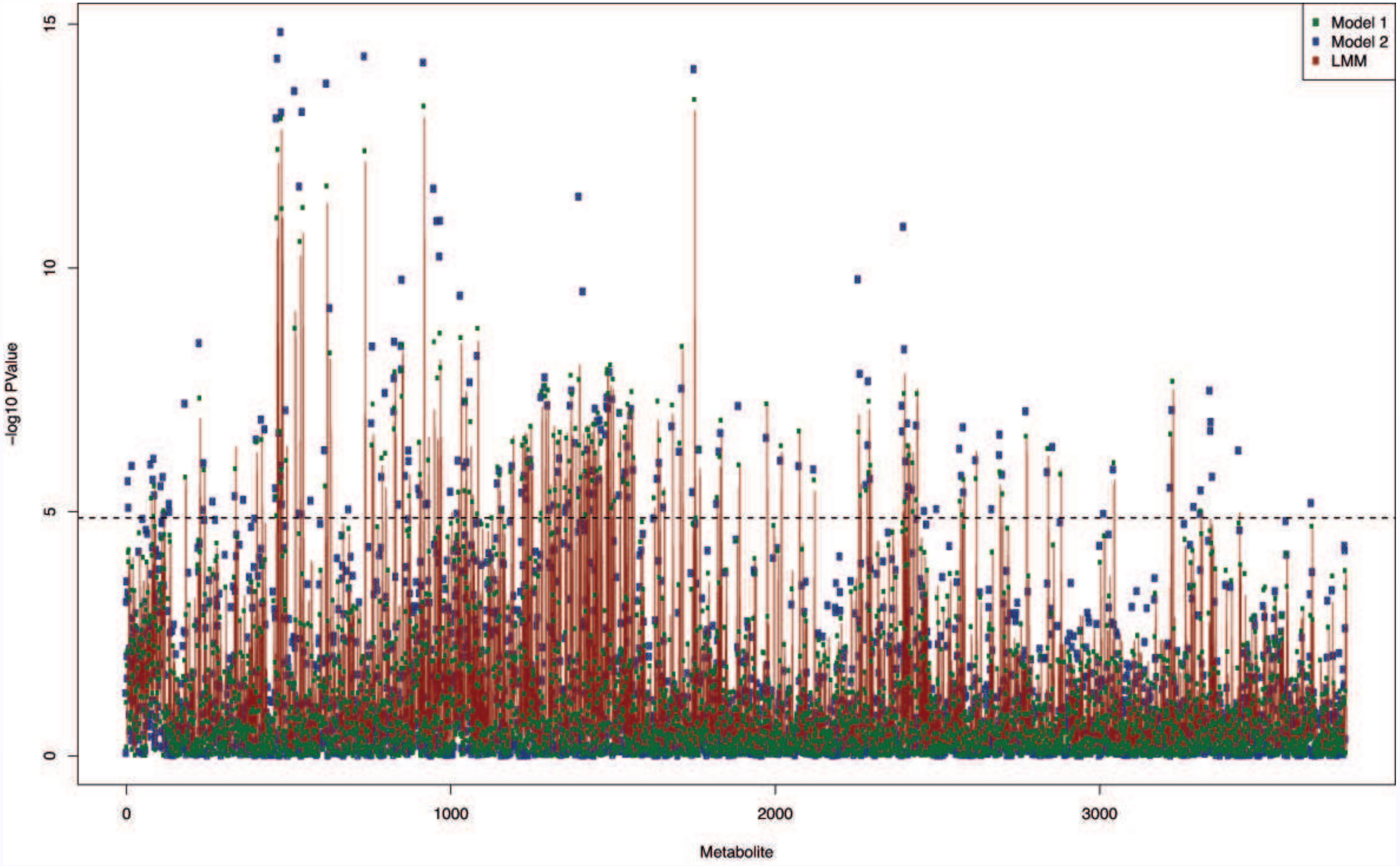
PISCINA study: LMM parametrisation

- Metabolite data: outcome, Y
  - N=6,471 peaks measured in the whole population
  - Data is standardised to unit variance (for comparability)
- Exposure data: predictor
  - Five DBP measured in exhaled breath
  - Log-transformed exposures
  - Exposures are centered on the average level across ’pre’-measurements
- Two measurements per participant: setting up a linear mixed model with an individual ID random intercept:

$$\texttt{Y} \sim \texttt{Expo + (1|ID)}$$

$$\Sigma = \begin{pmatrix} \sigma^2_{pre} & \delta \\ & \sigma^2_{post} \end{pmatrix}$$

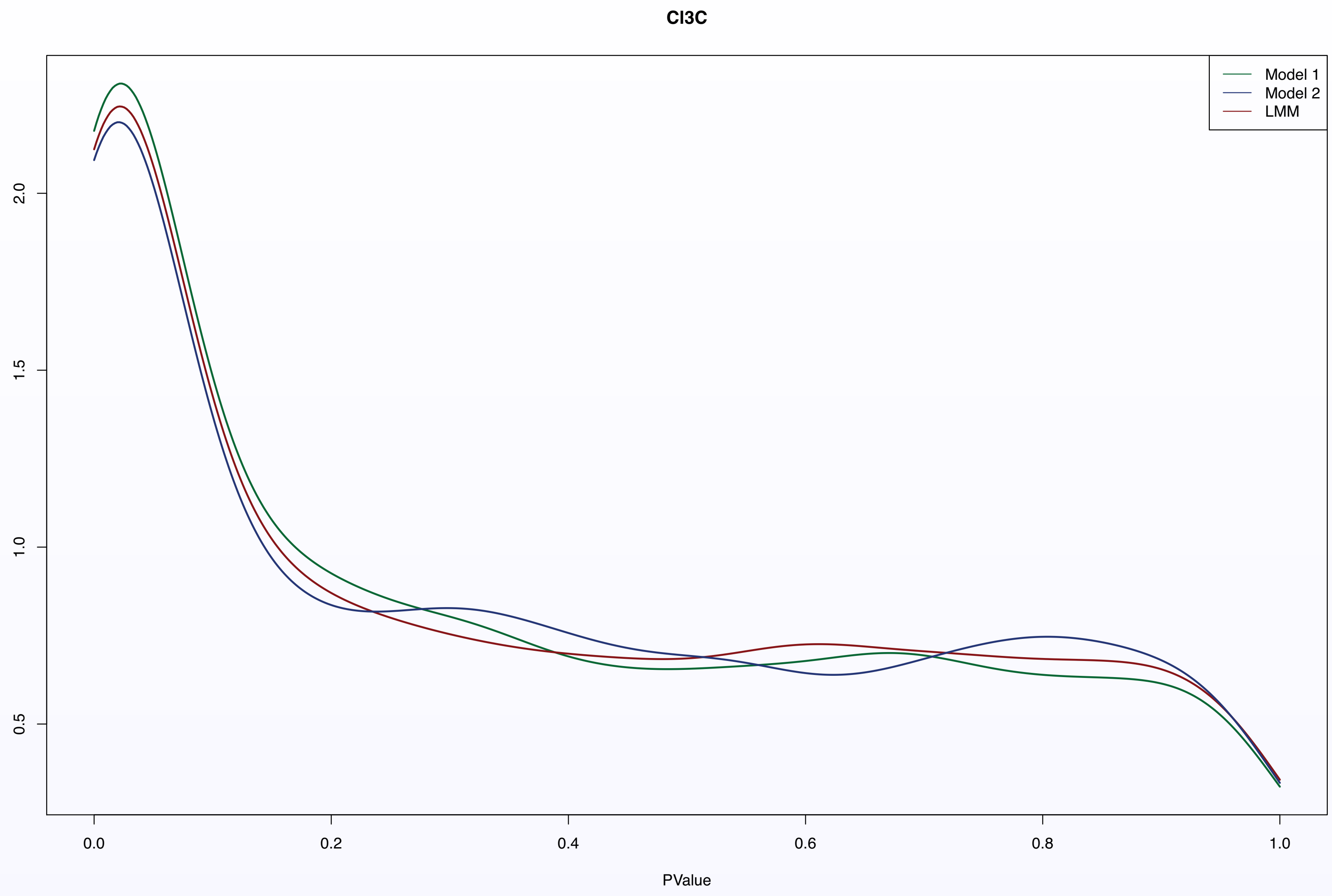
PISCINA study: model comparison



⇒ highly consistent results

⇒ Pre-post indicator may act as proxy for exposure

PISCINA study: model comparison

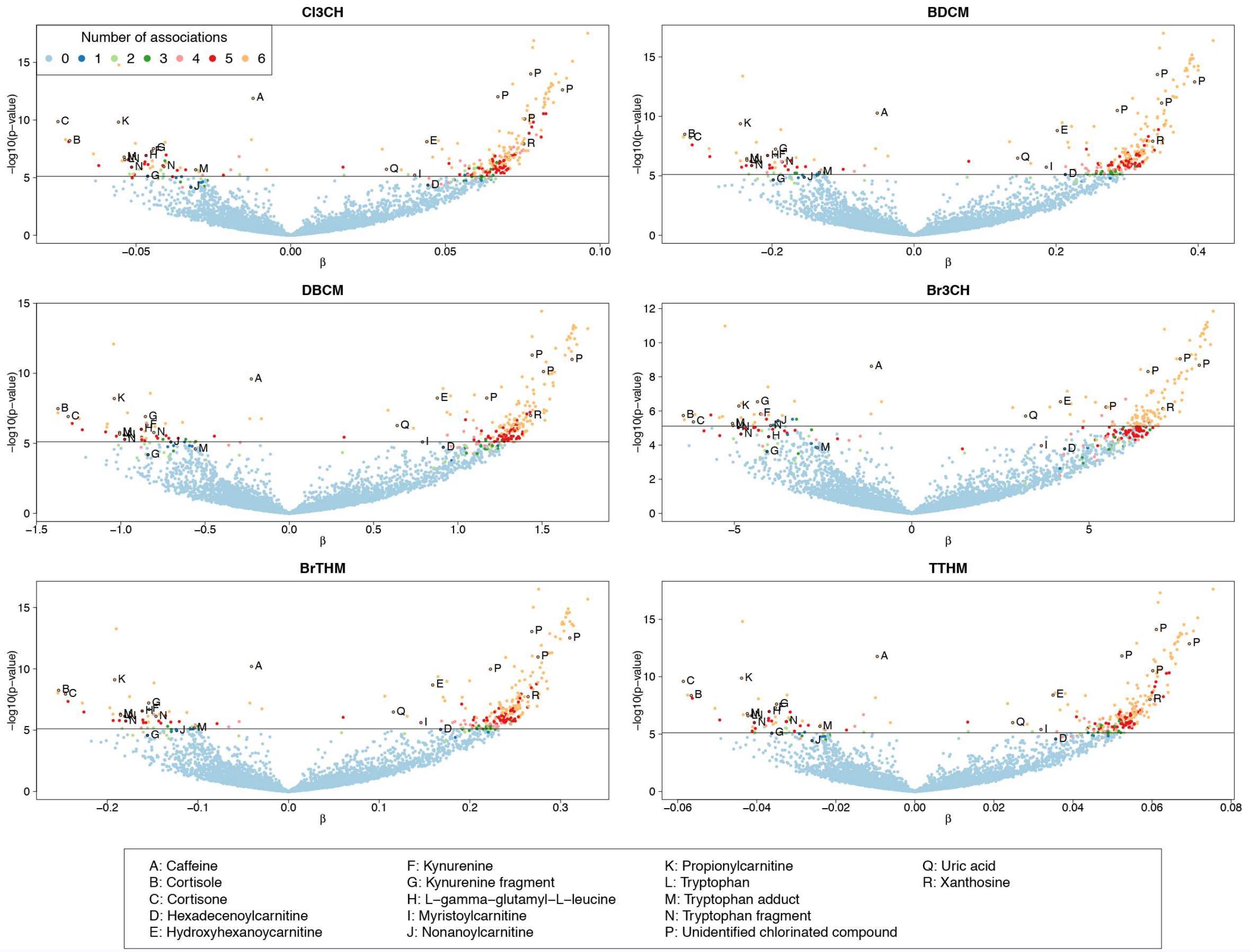


⇒ highly consistent p-value distributions

⇒ models are mostly equivalent

Results from PISCINA study: metabolomics

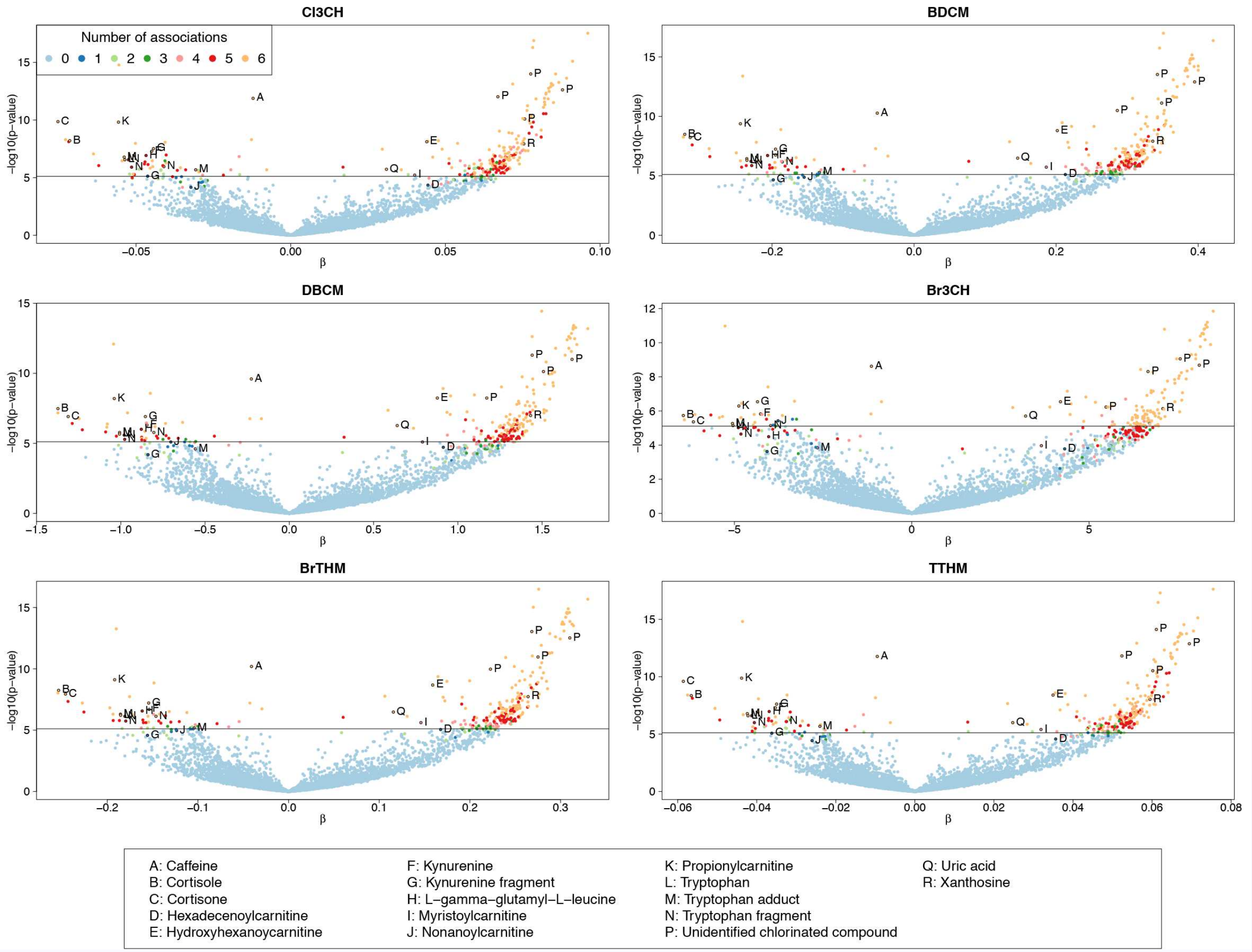
- Results overview (van Veldhoven *et al.*, *Env Int, 2018*):



⇒ 293 features associated to at least one exposure

Results from PISCINA study: metabolomics

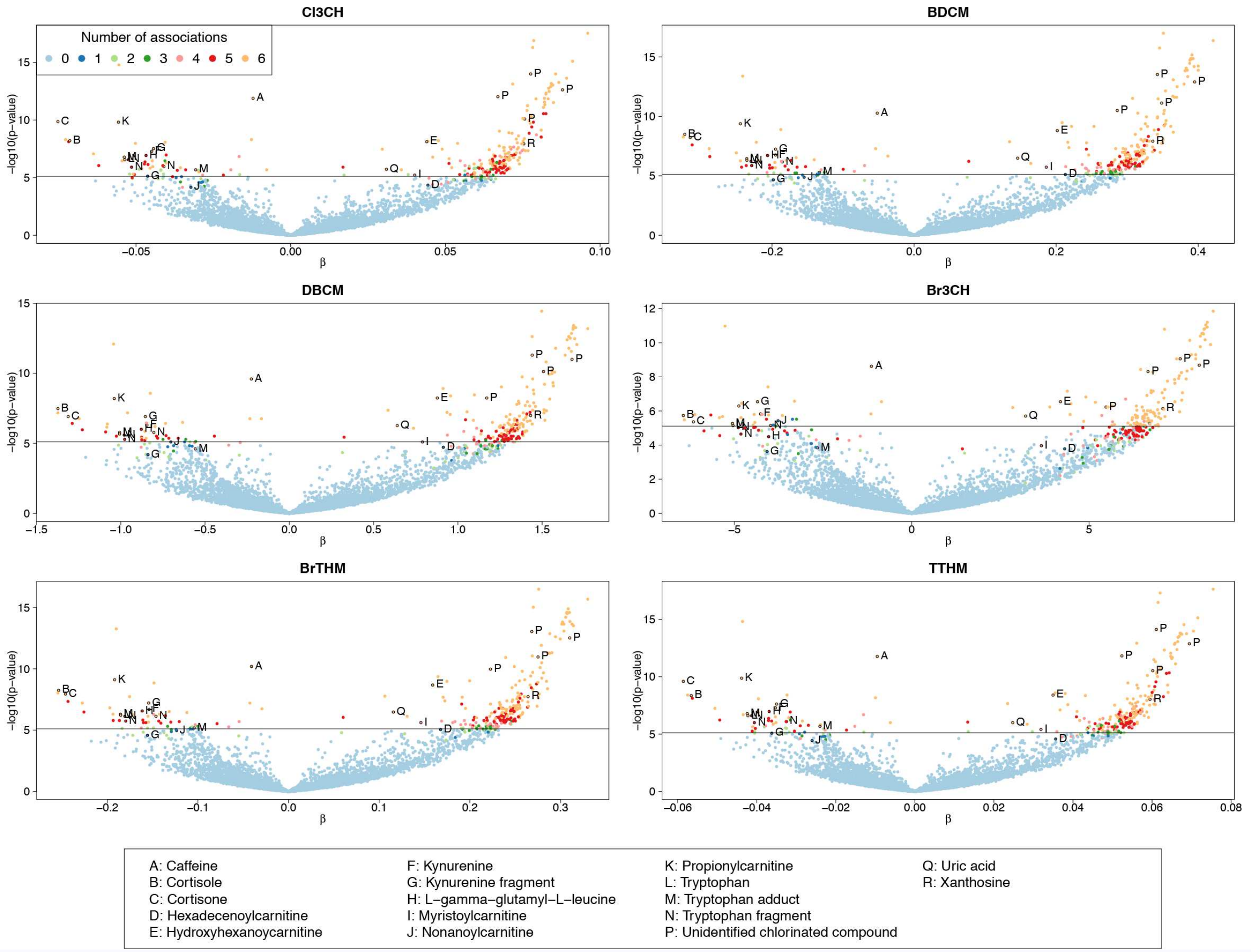
- Results overview (van Veldhoven *et al.*):



⇒ No association survives adjustment for Pre-post indicator

Results from PISCINA study: metabolomics

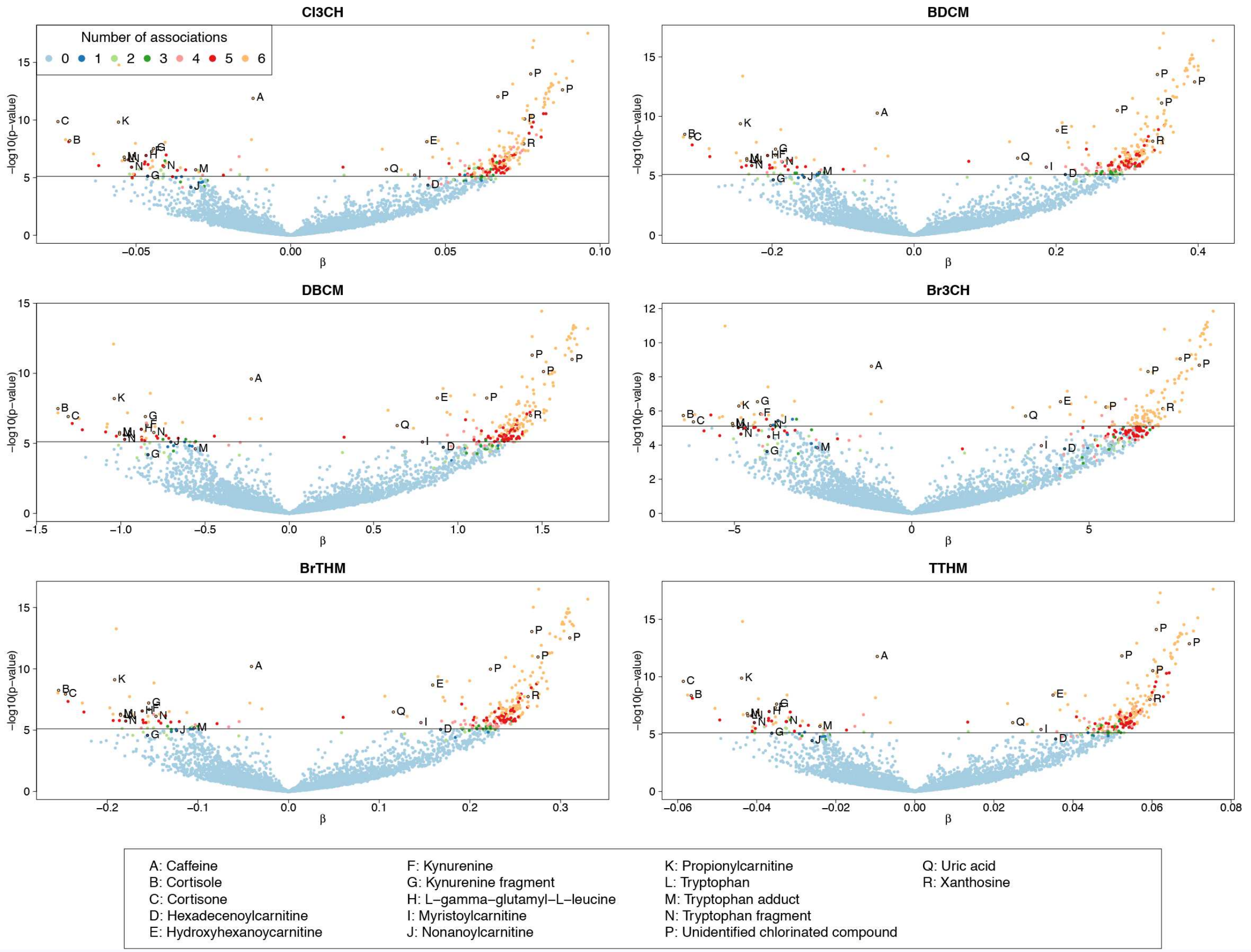
- Results overview (van Veldhoven *et al.*):



⇒ strong overlap across exposure-associated features (>60% associated to >3 exposures)

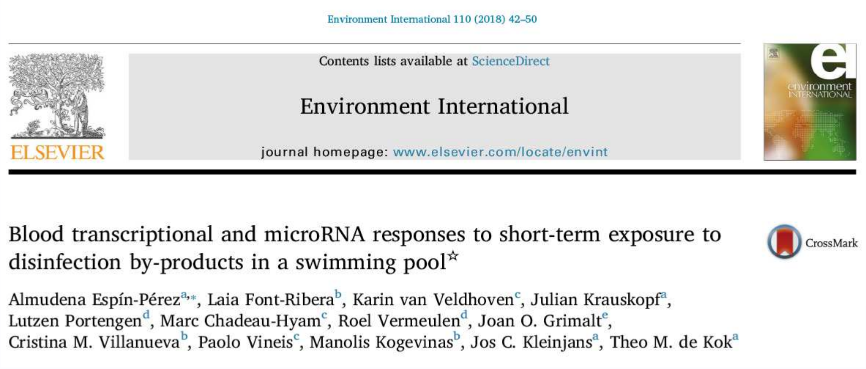
Results from PISCINA study: metabolomics

- Results overview (van Veldhoven *et al.*):



⇒ Confounding by the experiment (e.g. PA): feature annotation identified 13 chlorinated compounds)

Results from PISCINA study: transcriptomics (N$\sim$ 27,000)



Environment International 110 (2018) 42–50

Contents lists available at ScienceDirect

Environment International

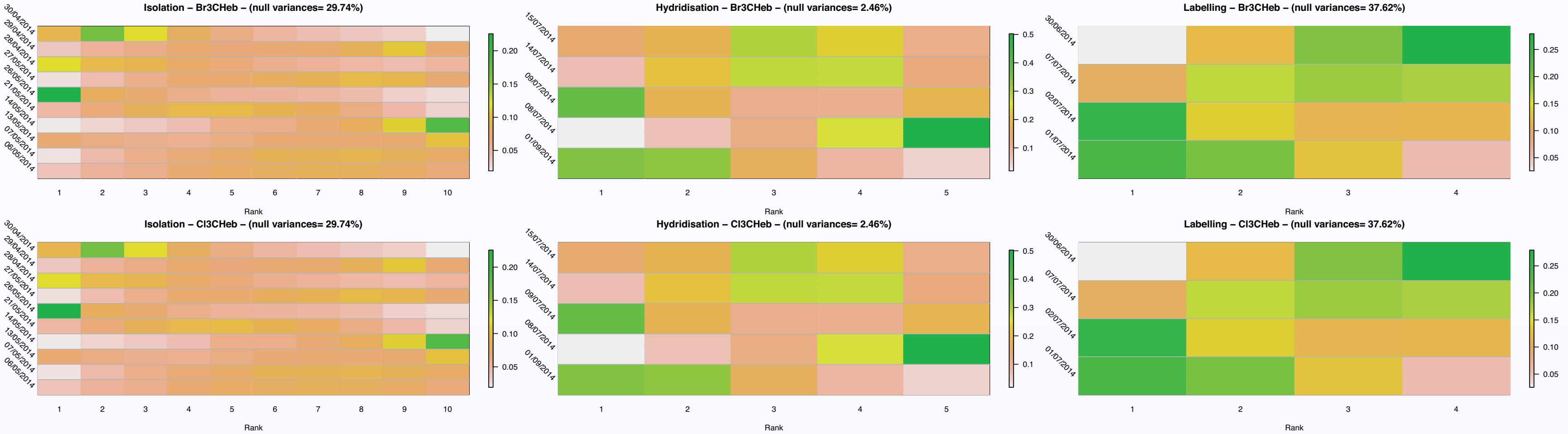
journal homepage: www.elsevier.com/locate/envint

ELSEVIER

Blood transcriptional and microRNA responses to short-term exposure to disinfection by-products in a swimming pool☆

Almudena Espín-Pérez[a,*], Laia Font-Ribera[b], Karin van Veldhoven[c], Julian Krauskopf[a], Lutzen Portengen[d], Marc Chadeau-Hyam[c], Roel Vermeulen[d], Joan O. Grimalt[e], Cristina M. Villanueva[b], Paolo Vineis[c], Manolis Kogevinas[b], Jos C. Kleinjans[a], Theo M. de Kok[a]

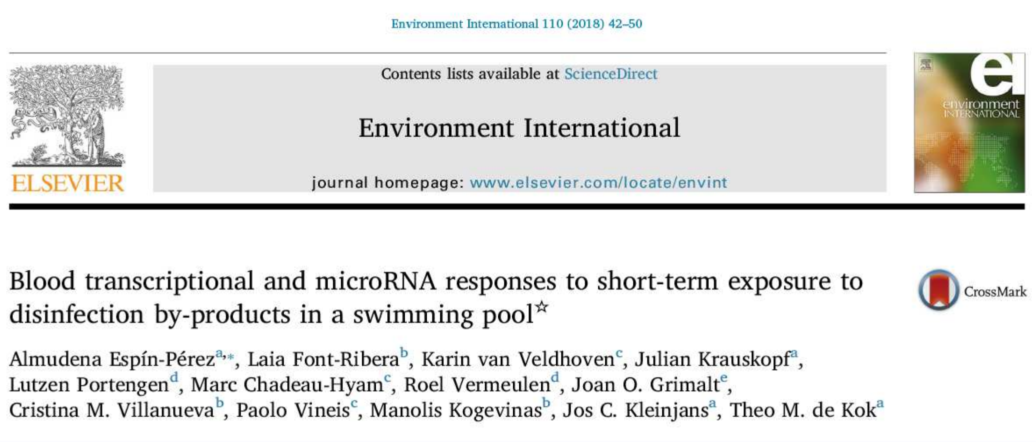
- Nuisance variation modelling: for $BrCH_3$ and $ClCH_3$



⇒ Hybridisation generated far more noise (<3% null variance estimates)

⇒ RE estimates are similar for all exposures

Results from PISCINA study: transcriptomics (N∼ 27,000)



Environment International 110 (2018) 42–50

Contents lists available at ScienceDirect

Environment International

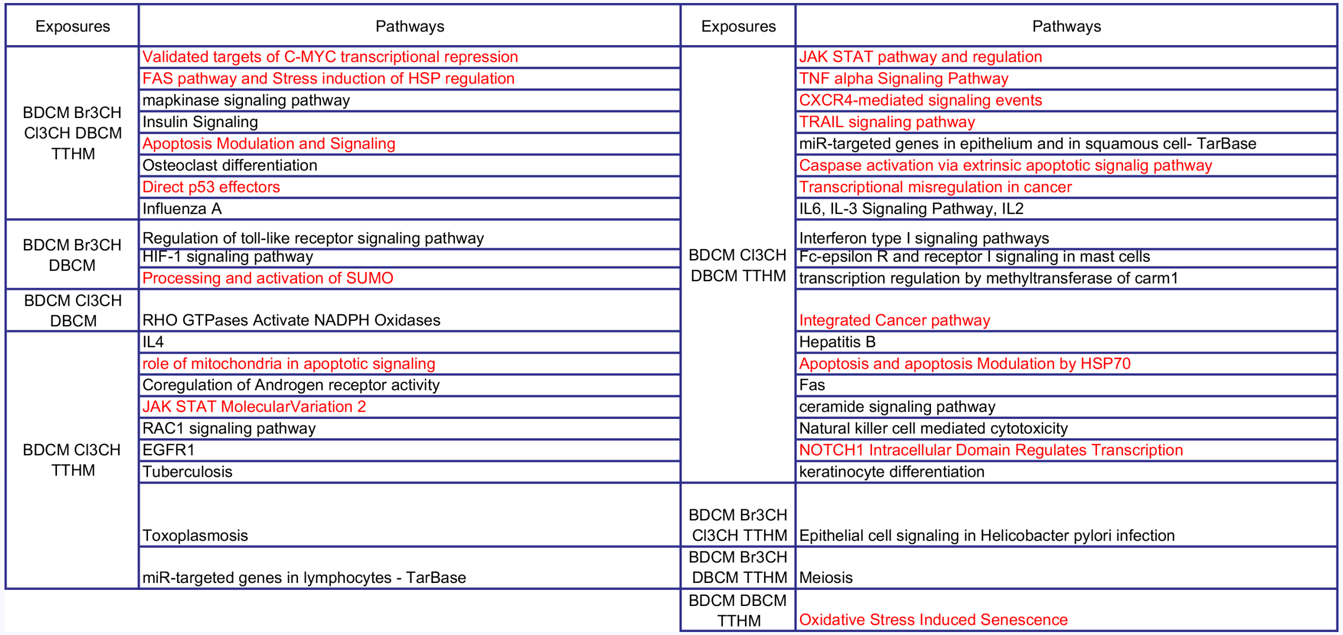
journal homepage: www.elsevier.com/locate/envint

ELSEVIER

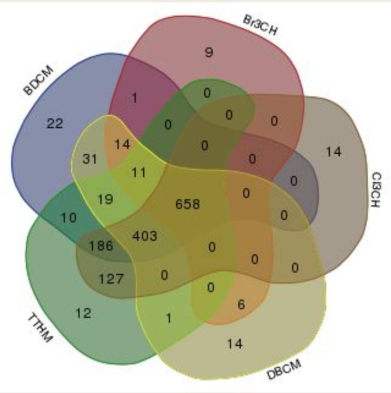
Blood transcriptional and microRNA responses to short-term exposure to disinfection by-products in a swimming pool☆

Almudena Espín-Pérez[a,*], Laia Font-Ribera[b], Karin van Veldhoven[c], Julian Krauskopf[a], Lutzen Portengen[d], Marc Chadeau-Hyam[c], Roel Vermeulen[d], Joan O. Grimalt[e], Cristina M. Villanueva[b], Paolo Vineis[c], Manolis Kogevinas[b], Jos C. Kleinjans[a], Theo M. de Kok[a]

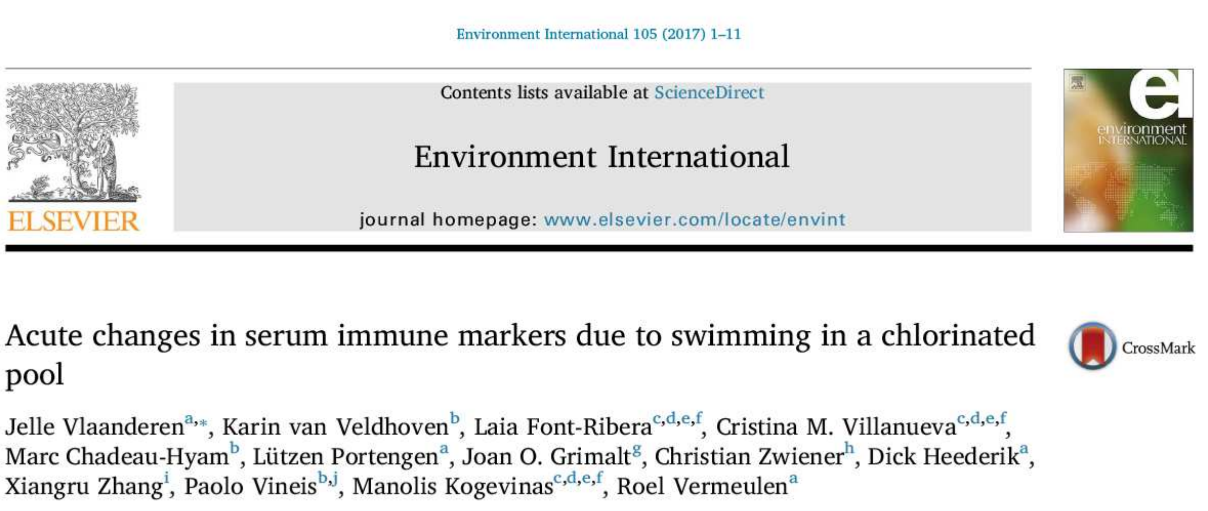
- Numerous associations: overlapping transcripts and enriched pathways



Exposures	Pathways	Exposures	Pathways
BDCM Br3CH Cl3CH DBCM TTHM	Validated targets of C-MYC transcriptional repression	BDCM Cl3CH DBCM TTHM	JAK STAT pathway and regulation
	FAS pathway and Stress induction of HSP regulation		TNF alpha Signaling Pathway
	mapkinase signaling pathway		CXCR4-mediated signaling events
	Insulin Signaling		TRAIL signaling pathway
	Apoptosis Modulation and Signaling		miR-targeted genes in epithelium and in squamous cell- TarBase
	Osteoclast differentiation		Caspase activation via extrinsic apoptotic signalig pathway
	Direct p53 effectors		Transcriptional misregulation in cancer
	Influenza A		IL6, IL-3 Signaling Pathway, IL2
BDCM Br3CH DBCM	Regulation of toll-like receptor signaling pathway		Interferon type I signaling pathways
	HIF-1 signaling pathway		Fc-epsilon R and receptor I signaling in mast cells
	Processing and activation of SUMO		transcription regulation by methyltransferase of carm1
BDCM Cl3CH DBCM	RHO GTPases Activate NADPH Oxidases		Integrated Cancer pathway
BDCM Cl3CH TTHM	IL4		Hepatitis B
	role of mitochondria in apoptotic signaling		Apoptosis and apoptosis Modulation by HSP70
	Coregulation of Androgen receptor activity		Fas
	JAK STAT MolecularVariation 2		ceramide signaling pathway
	RAC1 signaling pathway		Natural killer cell mediated cytotoxicity
	EGFR1		NOTCH1 Intracellular Domain Regulates Transcription
	Tuberculosis		keratinocyte differentiation
	Toxoplasmosis	BDCM Br3CH Cl3CH TTHM	Epithelial cell signaling in Helicobacter pylori infection
	miR-targeted genes in lymphocytes - TarBase	BDCM Br3CH DBCM TTHM	Meiosis
		BDCM DBCM TTHM	Oxidative Stress Induced Senescence



Results from PISCINA study: proteins (N=13)



Environment International 105 (2017) 1–11

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Acute changes in serum immune markers due to swimming in a chlorinated pool

Jelle Vlaanderen[a,*], Karin van Veldhoven[b], Laia Font-Ribera[c,d,e,f], Cristina M. Villanueva[c,d,e,f], Marc Chadeau-Hyam[b], Lützen Portengen[a], Joan O. Grimalt[g], Christian Zwiener[h], Dick Heederik[a], Xiangru Zhang[i], Paolo Vineis[b,j], Manolis Kogevinas[c,d,e,f], Roel Vermeulen[a]

- Consistent associations for all exposure, and TTHM (total)

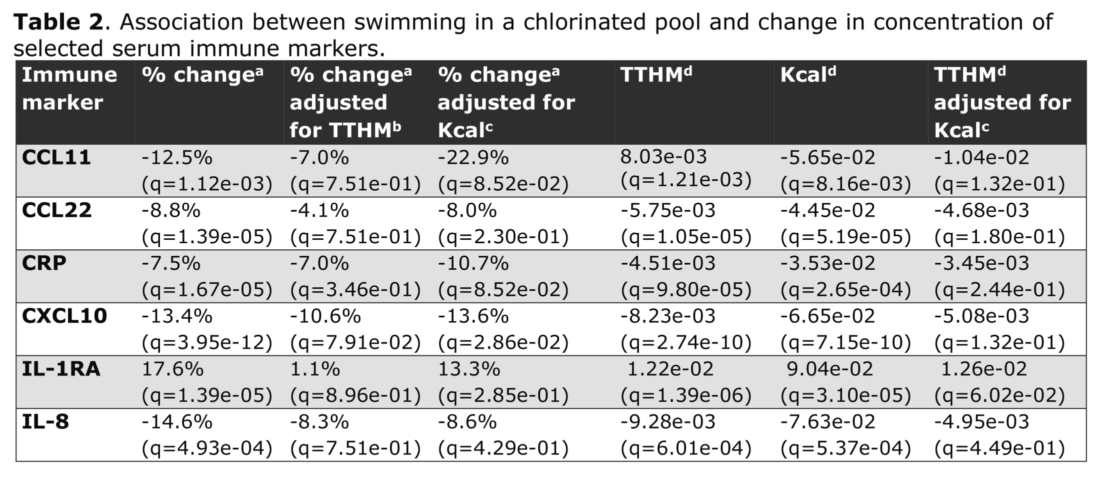
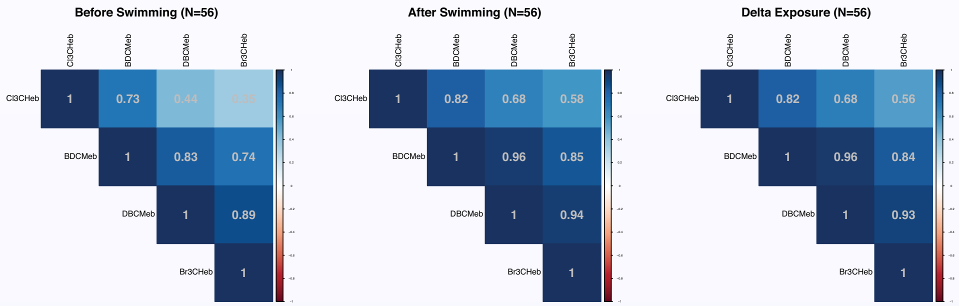


Table 2. Association between swimming in a chlorinated pool and change in concentration of selected serum immune markers.

Immune marker	% change^{a}	% change^{a} adjusted for TTHM^{b}	% change^{a} adjusted for Kcal^{c}	TTHM^{d}	Kcal^{d}	TTHM^{d} adjusted for Kcal^{c}
CCL11	-12.5% (q=1.12e-03)	-7.0% (q=7.51e-01)	-22.9% (q=8.52e-02)	8.03e-03 (q=1.21e-03)	-5.65e-02 (q=8.16e-03)	-1.04e-02 (q=1.32e-01)
CCL22	-8.8% (q=1.39e-05)	-4.1% (q=7.51e-01)	-8.0% (q=2.30e-01)	-5.75e-03 (q=1.05e-05)	-4.45e-02 (q=5.19e-05)	-4.68e-03 (q=1.80e-01)
CRP	-7.5% (q=1.67e-05)	-7.0% (q=3.46e-01)	-10.7% (q=8.52e-02)	-4.51e-03 (q=9.80e-05)	-3.53e-02 (q=2.65e-04)	-3.45e-03 (q=2.44e-01)
CXCL10	-13.4% (q=3.95e-12)	-10.6% (q=7.91e-02)	-13.6% (q=2.86e-02)	-8.23e-03 (q=2.74e-10)	-6.65e-02 (q=7.15e-10)	-5.08e-03 (q=1.32e-01)
IL-1RA	17.6% (q=1.39e-05)	1.1% (q=8.96e-01)	13.3% (q=2.85e-01)	1.22e-02 (q=1.39e-06)	9.04e-02 (q=3.10e-05)	1.26e-02 (q=6.02e-02)
IL-8	-14.6% (q=4.93e-04)	-8.3% (q=7.51e-01)	-8.6% (q=4.29e-01)	-9.28e-03 (q=6.01e-04)	-7.63e-02 (q=5.37e-04)	-4.95e-03 (q=4.49e-01)

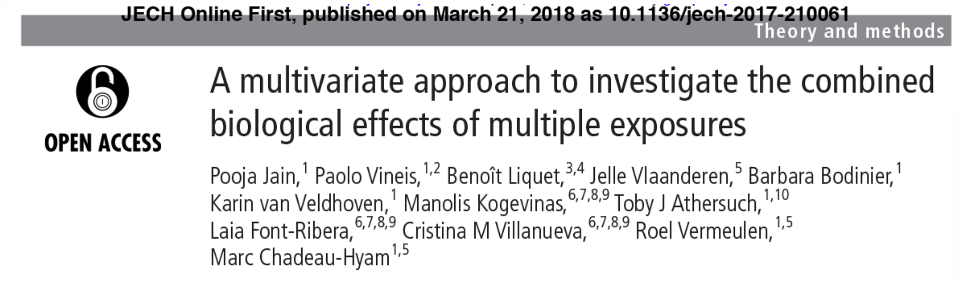
Results from PISCINA study: Conclusions

- The OMICS data sets investigated
  - Proteins (N=13 inflammatory markers)
  - Metabolomics (N$\sim$ 6,000 features)
  - Transcriptomics: (N$\sim$ 30,000 transcripts)
- Main conclusions:
  - Effects of the experiment was detected at all 3 molecular levels
  - Irrespective of the platform, strong overlap across markers of each exposure
- Exposure Correlations: strong co-occurence



$\Rightarrow$ is the strong overlap across exposure due to their correlation?

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?

  ⇒ is there a 'mixture effect'?

  ⇒ 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)

Refresher on Partial Least Square model

- Refresher on the PCA:
  - Unsupervised approach
  - For each principal component h, find loadings u_h such that:

$$\max_{||u_h||=1} \text{Var}(X_h u_h) \quad h \in \{1, \ldots, H\}$$

- Partial Least Square (PLS): supervised extension, $i.e.$ summarises the information in X that is relevant to a (multivariate) outcome Y
- Objective: estimate the loadings u_h and v_h summaring X and Y, respectively such that the variance covariance between the projections is maximal

$$\max_{||u_h||=1,\ ||v_h||=1} \text{Cov}(X_h u_h, Y_h v_h) \quad h \in \{1, \ldots, H\}$$

$$X_h = \begin{pmatrix} x_{h_{11}} & \ldots & x_{h_{1p}} \\ \ldots & \ldots & \ldots \\ x_{h_{n1}} & \ldots & x_{h_{np}} \end{pmatrix} u_h = \begin{pmatrix} u_h^1 \\ \ldots \\ u_h^p \end{pmatrix} Y_h = \begin{pmatrix} y_{h_{11}} & \ldots & y_{h_{1p}} \\ \ldots & \ldots & \ldots \\ y_{h_{n1}} & \ldots & y_{h_{np}} \end{pmatrix} v_h = \begin{pmatrix} v_h^1 \\ \ldots \\ v_h^p \end{pmatrix}$$

Partial Least Square: estimation procedure univariate case

- Initialisation: Find $\hat{u}_1$ such that

$$\hat{u}_1 = \underset{||u_1||=1}{\text{argmax}} \, \text{Cov}(Xu_1, Y) = \frac{X^T Y}{||X^T Y||}$$

$\Rightarrow$ Scores of the first component of X are computed from linear combination of X with loadings coefficients in u: $S_{X1} = Xu$ (rescaled coefficients from standardised linear regression)

- Iterative algorithm:
  1. Deflation step: the variance of X_{h-1} explained by component $(h-1)$ is removed in X_h to ensure orthogonality

     $$X_h = X_{h-1} - S_{h-1} c^T, \text{ where } c\text{: reg. coeff of } X_{h-1} \; S_{h-1}$$

     $\Rightarrow$ remove from X the information of X captured by comp. $h-1$

     $$Y_h = Y_{h-1} - dS_{h-1}, \text{ where } d\text{: reg. coeff of } Y_{h-1} \; S_{h-1}$$

     $\Rightarrow$ remove from Y the part explained by the X comp. $h-1$
  2. Find $\hat{u}_h$ such that: $\hat{u}_h = \underset{||u_h||=1}{\text{argmax}} \, \text{Cov}(X_h u_h, Y_h)$

PLS: estimation procedure multivariate case

- Parameter estimation for multivariate Y now we have loadings for Y (v)

$$\max_{||u_h||=1,\ ||v_h||=1} \mathrm{Cov}(X_h u_h, Y_h v_h) \quad h \in \{1, \ldots, H\}$$

- Initialisation: set $h = 1$, and $Y_1 = Y$
  1. Set w_h as the first column of Y_h
  2. Calculate X loadings: $u_h = \frac{X_h^T w_h}{w_h^T w_h}$ and scale u_h to 1
  3. Compute the scores of X_h: $S_h = X_h u_h$
  4. Derive the Y loadings (regressing Y_h on the X scores: $v_h = \frac{Y_h^T S_h}{S_h^T S_h}$
  5. Compute the scores of Y_h and set $w_h = Y_h v_h$ (Y scores)
  6. Repeat 2 to 5 until convergence (limited changes in v and u)
  7. Compute the regression coefficients c_h (or e_h) from the regression of X_h (or Y_h) onto S_h
  8. Deflation step: compute the residual matrices $X_{h+1} = X_h - S_h c_h^T$ and $Y_{h+1} = Y_h - S_h e_h^T$
  9. Increment h