Statistical glossary and perspectives: main approaches relevant for Exposomics

MRC-PHE Centre Investigator's Seminar – Exposomics Update – London

Marc Chadeau-Hyam

Overview: Exposomics Aims and Design

- Aims: develop am new methodological framework to:
  - Assessing the biological/molecular effect of high priority environmental exposures (internal exposome)
  - Identify mixture(s) of exposure driving future risks of health outcomes (external exposome)
  - Identify how the internal and external exposomes overlap and concur to future risk of chronic disease
  - Account to age-related differential effects & susceptibility function
- Three main types of effects investigated: different study designs

Type of effect	Timescale	Design	Exposures
Acute effect	<2 hours	Intervention study	Pre-post experiment meas.
Short-term effect	24 Hours	Personalised Exposure Measurement Campaigns (PEM)	Real-time monitoring (e.g., backpack)
Long-term effect	Years	Cohort Studies	Modelled exposure (LUR...)

Exposomics data

- Exposure data
  - Air pollution data
  - Water pollution data
- OMICs data
  - In all studies: adductomics, transcriptomics, metabolomics (MS)
  - In long term (cohort) studies: proteomics and epigenetics
- Age ranges:
  - Young children 0-4 years old
  - Children: 5-9 years old
  - Young adults/Adults: 18-70 years old
- Health outcomes:
  - Children: birth weight, neurodevelopment
  - Adults: CVD, CRC, Asthma

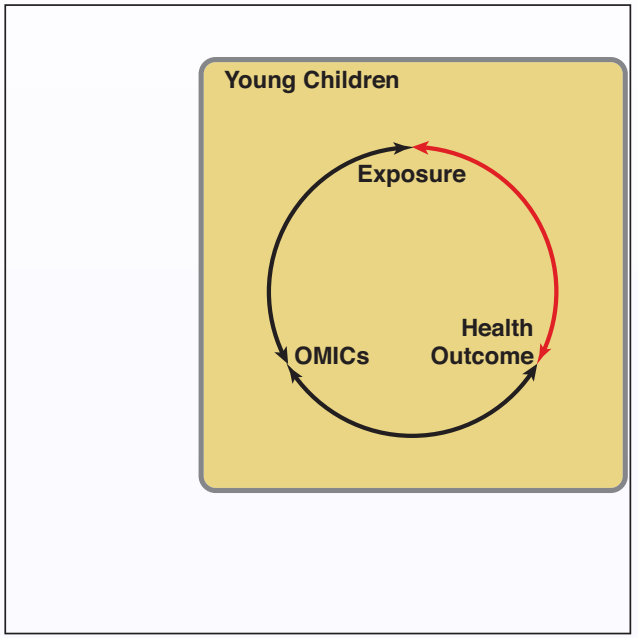
Exposomics data

	Study	Exp markers Source	Time scale	Epigenetics	Age
AIR	Oxford Street 2	PMNOxUFP$^{(1)}$	<2hr	×	50-70
	TAPAS 2	PMNOxUFP$^{(1)}$	<2hr; long-term	×	18-60
	PEM-adults	PMNOxUFP$^{(1)}$	24hr; long-term	✓	50-70
	PEM-kids INMA	PMNOxUFP$^{(1)}$	24hr	✓	7-9
	Piscina air	PMNOxUFP$^{(1)}$	24hr	×	18-40
	EPIC-NL	ESCAPE	Long-term	✓	50-70
	EPIC-Torino	ESCAPE	Long-term	✓	50-70
	East Anglia	ESCAPE-extension	Long-term	✓	50-70
	Sapaldia	ESCAPE country specific models	Long-term	✓	50-70
	ALSPAC	LUR	Long-term	✓	0-7
	RHEA	ESCAPE	Long-term	✓	0-4
	Piccoli+	ESCAPE	Long-term	✓	0-4
	EPIGENAIR	ESCAPE	Long-term	✓	35-70
WATER	Piscina	Water pollutants$^{(1)}$	<2hr (40 mins)	×	18-40
	MCC	Water pollutants	Long-term	✓	

Main analyses: general analytical plan

For a given health outcome

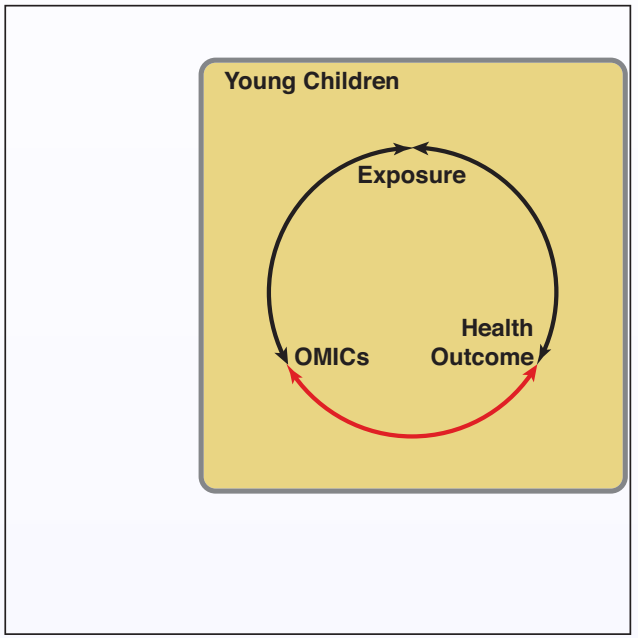
1. Exposure profiling
   - External exposome relating to health outcome
   - Aim: Identify (mixtures of) exposures that drive future risk of the health outcome
   - Specifics: several tens of highly correlated measures



Main analyses: general analytical plan

For a given health outcome

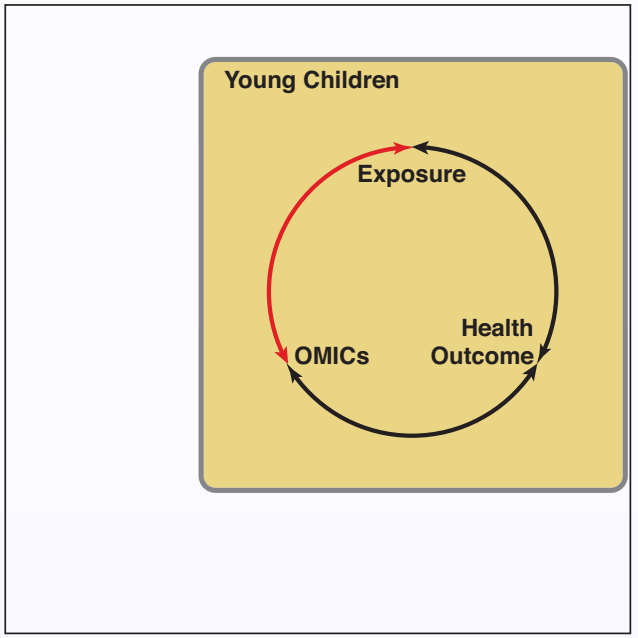
1. Exposure profiling
2. OMICs-health outcome profiling
   - Internal exposome $vs.$ health outcome
   - Aim: Identify sets of OMICs prospective and early disease markers
   - Specifics: several thousand of correlated measures
   - Investigate each platform separately
   - Integrate the different platforms (cross-omic analyses)



Main analyses: general analytical plan

For a given health outcome

1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
   - Internal $vs.$ external exposomes
   - Aim: biologically relevant markers of exposures
   - Specifics: multivariate X and Y
   - Investigate each platform separately
   - Integrate the different platforms (cross-omic analyses)
   - Possibility to match in experimental studies



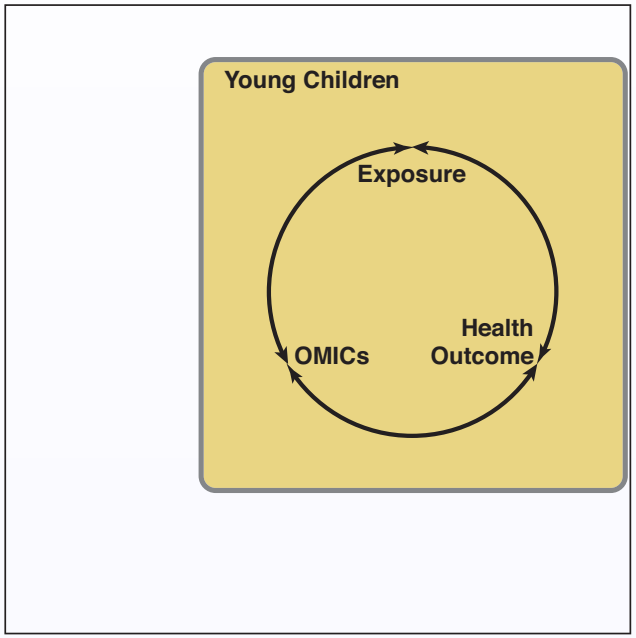
Main analyses: general analytical plan

For a given health outcome

1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3

$\Rightarrow$ investigation of the markers co-action

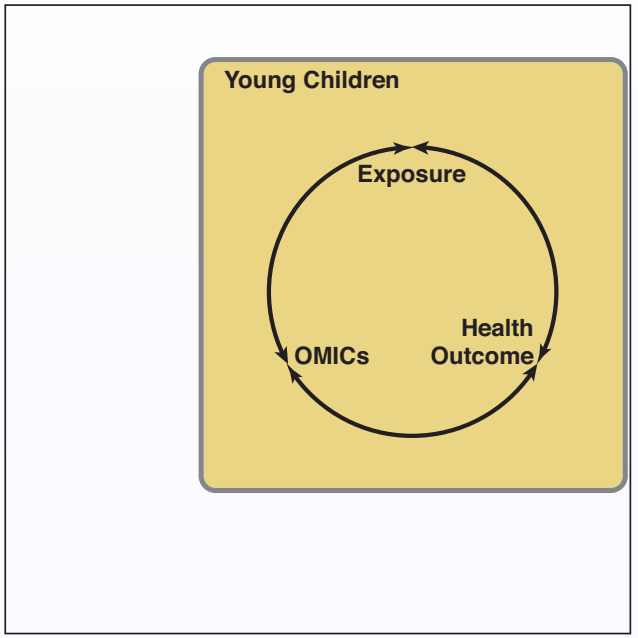
$\Rightarrow$ insights into biological mechanisms involved



Main analyses: general analytical plan

For a given health outcome

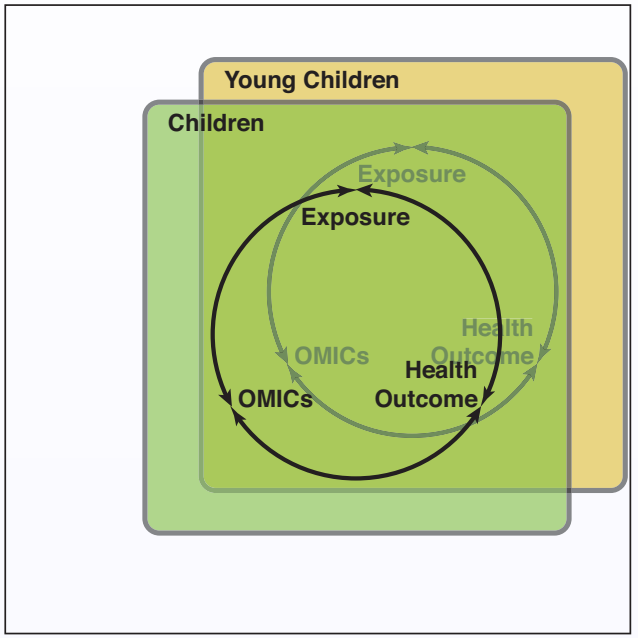
1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
   - young children (0-4)



Main analyses: general analytical plan

For a given health outcome

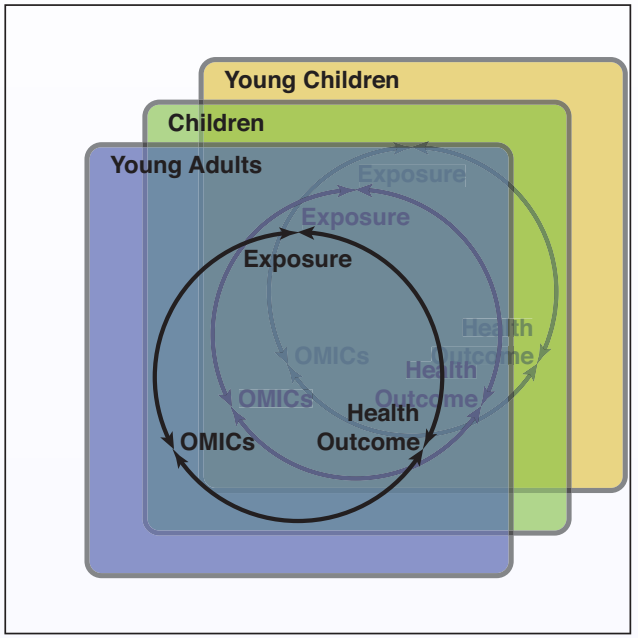
1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
   - young children (0-4); children (4-10)



Main analyses: general analytical plan

For a given health outcome

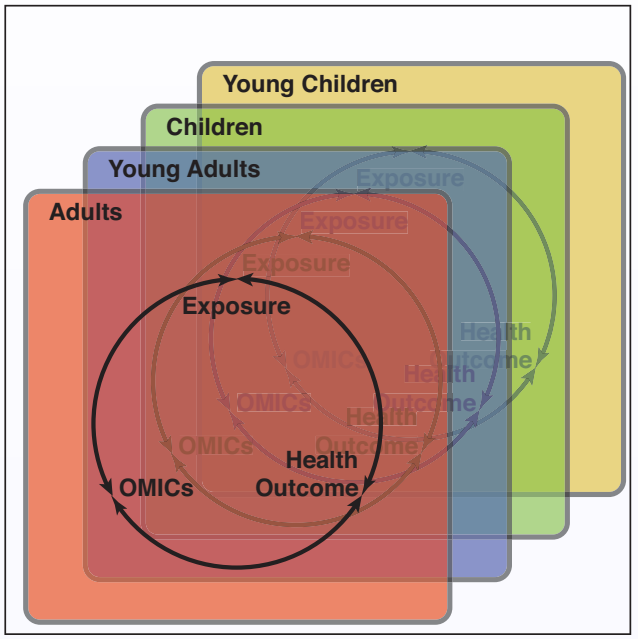
1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
   - young children (0-4); children (4-10); young adults (18-40)



Main analyses: general analytical plan

For a given health outcome

1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
   - young children (0-4); children (4-10); young adults (18-40); adults (>40)

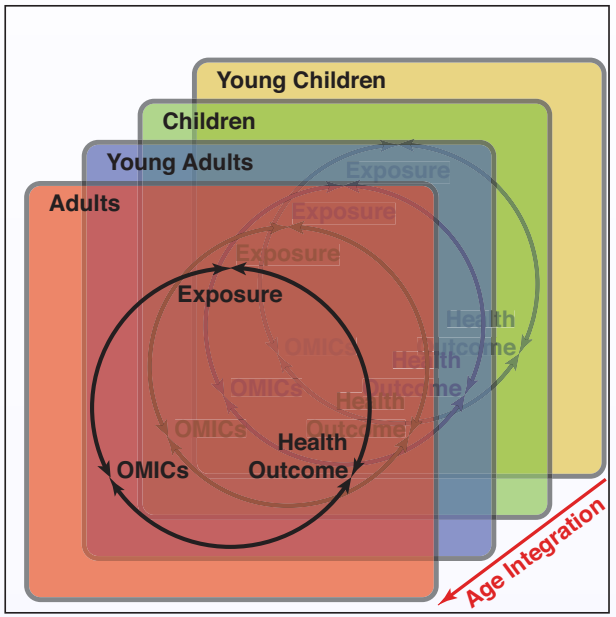


Main analyses: general analytical plan

For a given health outcome

1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
6. Integration across age classes:

$\Rightarrow$ investigate age-related effect modifications & susceptibility functions

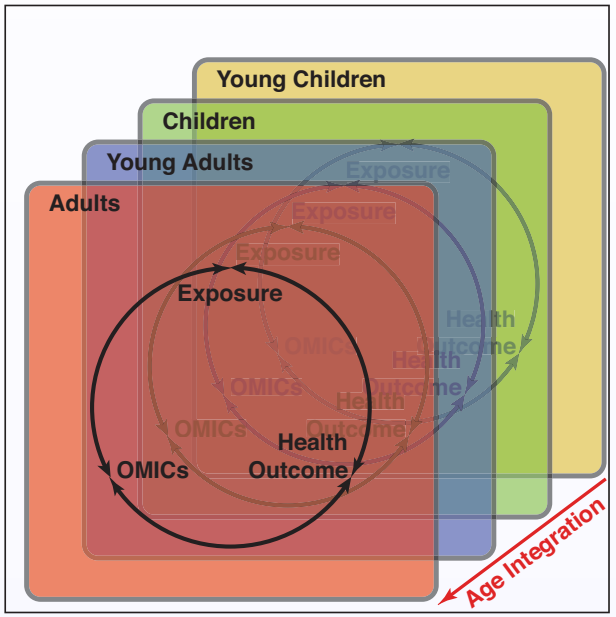


Main analyses: general analytical plan

For a given health outcome

1. Exposure profiling
2. OMICs-health outcome profiling
3. OMICs-exposure profiling
4. Integrate biomarkers identified in 1-3
5. Re-iterate steps 1-4 for other co-horts/age ranges:
6. Integration across age classes
7. Integration across outcomes:

$\Rightarrow$ investigate potential common pathological pathways



$\Rightarrow$ highly dimensional project!

OMICs/Exposure data: diverse and complex data

- Nature of the data
  - Categorical variables ($e.g.$ genotype data)
  - Continuous variables ($e.g.$ methylation, exposures $\ldots$)
- Dimension: wide range of scales
  - Tens of measurements (exposures)
  - Hundreds of measurements (proteins levels)
  - Tens of thousands of variables: (NMR-MS spectral data)
  - Hundreds of thousands of variables (epigenome scans)
- Correlated structure in the data:
  - Strength of the correlation varies
  - Correlation structure can either be 'distance-driven' ($e.g$ LD in genomics data) or more complex ($e.g.$ NMR spectral data).

$\Rightarrow$ need for computationally efficient and flexible models providing interpretable results

Exposomics: further challenges

- Effect of environmental exposures
  - Exposure are expected to have subtle effects
  - Mixtures of exposures are active (non-additive effects)

  $\Rightarrow$ need for powerful methods handling multivariate X and Y
- Complex effect: molecular signatures at different levels

  $\Rightarrow$ need to integrate the different OMICs data and explore molecular mechanisms
- Complex effect: the temporal component reflected in the study design
  - Exposures effects have different time scales: acute (experimental studies), mid-term (PEM), and long term (modelled exposures)
  - Potential age effect modification (age-related susceptibility to exposures, and disease)

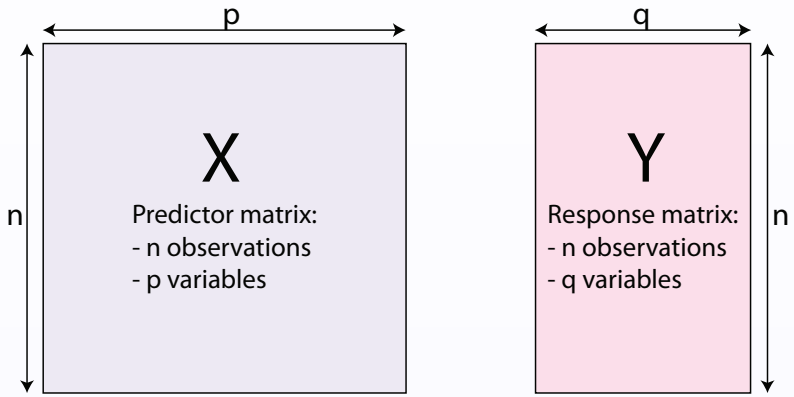
  $\Rightarrow$ incorporate a longitudinal component in the models

Exposomics: 3 main analytical streams

- Screening models: ‘OMICs & Exposure profiling’
  - Aim: identify within each OMIcs platforms & (sets of) exposures, relevant signatures of exposures
  - Status: established methods, benchmark for Exposomics
- Integrative models: ‘Cross-omic’ analyses
  - Aim: integrate data arising form several OMIC platforms and explore their interplay
  - Status: methods/strategies are developing
- Models including a temporal component
  - Aim: model the temporal component of the exposome
  - Status: experimental...

Profiling methods: *-*WAS*

Data definition:



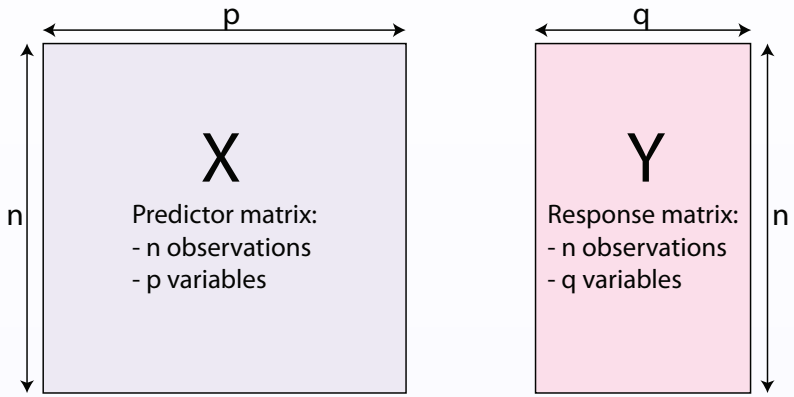
Aim: identify which of the p variables in X (OMICs/ exposure data) are associated with the outcome Y (disease status or (mixtures of) exposure(s))

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

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

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

Univariate approaches

- Principle: assess the association between each column of X and the outcome Y
- Model formulation: linear model for individual i and predictor j

$$Y_i = \alpha + \beta X_{ij} + \epsilon_{ij},$$

where:

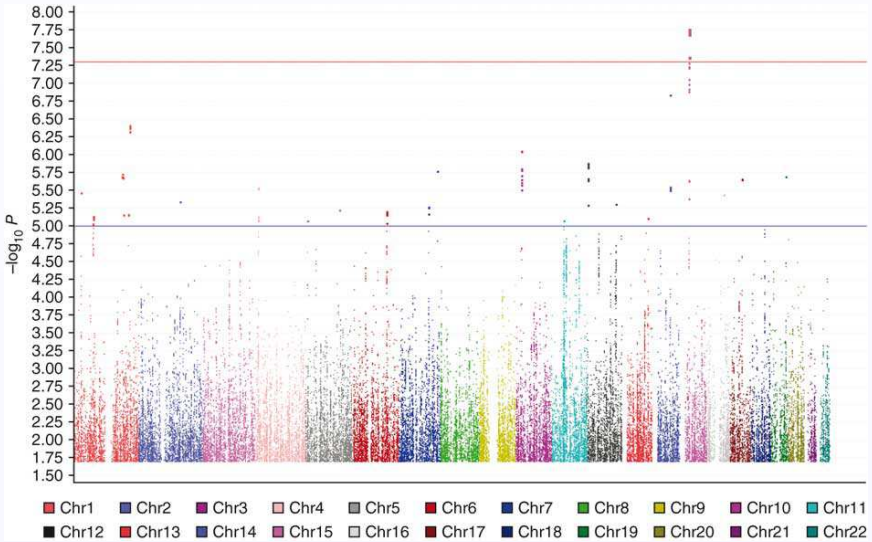
- Y_i is the measured outcome (possibly multivariate)
- X_{ij} is the observed value for j^{th} predictor
- α is the intercept
- β is the regression coefficient
- ϵ_{ij} is the residual error measuring the random deviation from the linear relationship

$\Rightarrow p$ models are estimated (one per predictor)

Univariate approaches

- Principle: assess the association between each column of X and the outcome Y
- Model formulation: linear model for individual i and predictor j

$$Y_i = \alpha + \beta X_{ij} + \epsilon_{ij},$$



$\Rightarrow$ how to draw a general conclusion over all p tests performed?

Multiple Testing correction Strategies

	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: define the per-test significance α' ensuring 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

    $\Rightarrow$ FDR control may be preferred in an exploratory context

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)
  - Main limitation: results may not guarantee easy interpretability

    $\Rightarrow$ need to ensure sparsity of the results
- 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

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

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

- Method: singular value decomposition (eigenvalues/eigenvectors)

  $\Rightarrow$ eigenvalues measures the proportion of variance explained

- Limitation: unsupervised method; the variation in the data may not be relevant to the outcome of interest.

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

  $\Rightarrow$ need for supervised methods

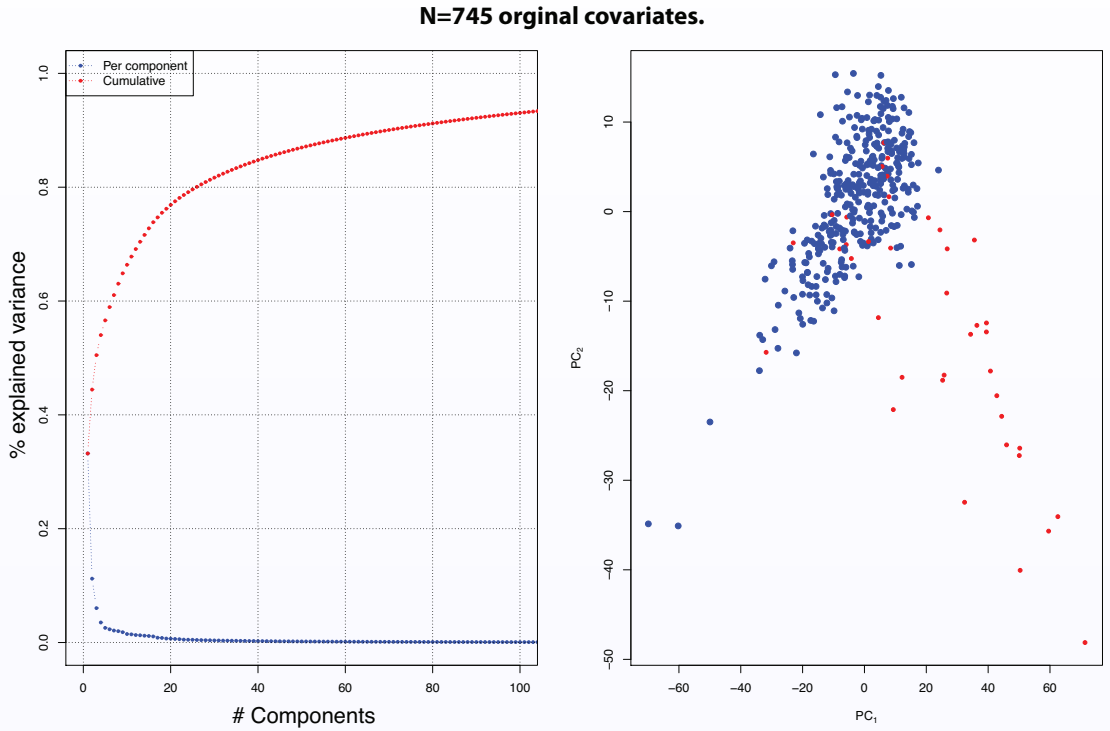
- Principle: PLS seeks for PCs that are the most correlated to the outcome: Definition of the objective function:

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

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

Dimension Reduction Techniques in practice

- Scree plot and Score plot:

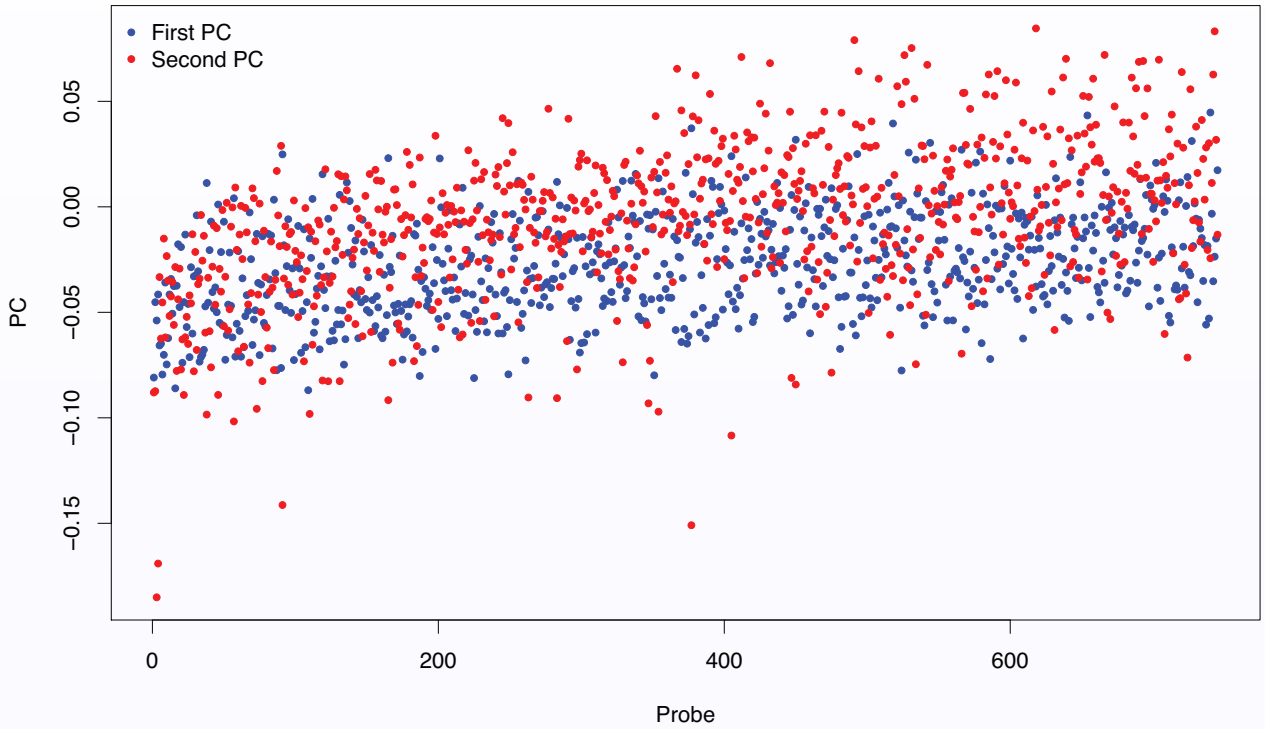


  - 90% variance explained for 80 PC's (≈10%)
  - Clear discrimination of cases and controls (PC_1 and PC_2 are strongly associated to Ca/Co)

$\Rightarrow$ 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 $n \leq p$
  - The number of predictors with $\beta \neq 0$ is upper bounded by p
- Main features of Lasso
  - No constraint on the number of retained markers ($i.e$ with $\beta \neq 0$)
  - Shrinkage: for large values of λ, regression coefficients are shrunk towards 0

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

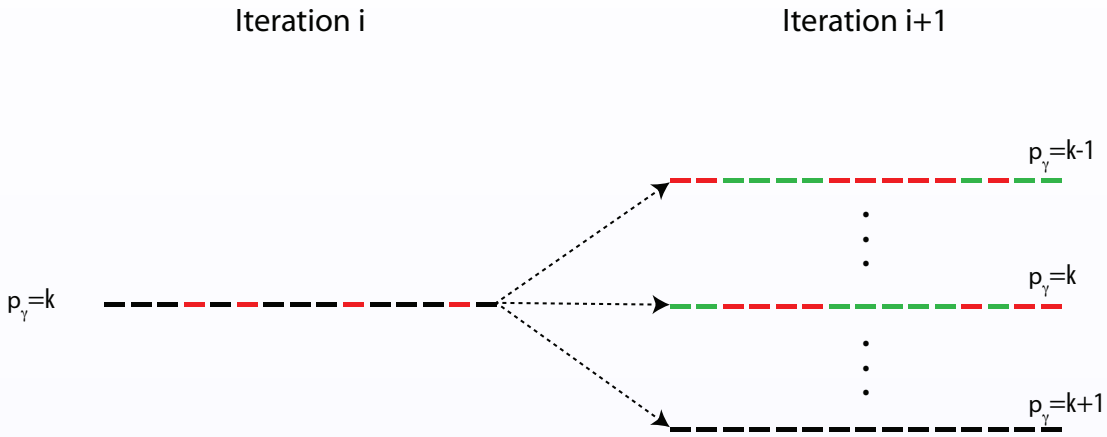
- Key issues:
  - Defining f and the subsequent S

    $\Rightarrow$ 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?

SSS, one intuitive search algorithm

- Shotgun Stochastic Search (SSS):



- Identification of the best model based on gains in S (quality of fit statistics)
- GUESS: a BVS for multiple outcomes
  - Optimised Search algorithm: EMC
  - Computational optimisation: enabling GPU capacity

$\Rightarrow$ GUESS is tailored for exposome investigation

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

Variable Selection approaches: strenghts and limitations

- Multivariate models accounting for combined effects of predictors

  $\Rightarrow$ implicit correction for multiple testing

  $\Rightarrow$improved power to detect multivariate/complex effects
- Main features of penalised regression:
  - Computationally efficient
  - Provides easily interpretable results
  - Accommodates all types of outcomes
  - Perfs. are hampered by calibration of the penalisation parameter
- Main features of BVS:
  - Provides easily interpretable results
  - Accommodates multiple outcomes and most outcomes
  - Easily incorporates adjustment on confounders
  - Integrative analyses (no expensive calibration)
  - Subtle parametrization (although mostly automated)

Wrap-up summary

- Univariate approaches and multiple testing correction
  - Computationally efficient and highly flexible models
  - Not accounting for potential combined effects of predictors

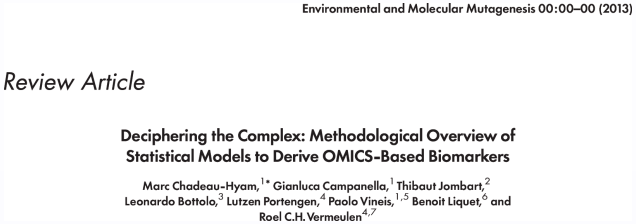
    $\Rightarrow$ towards multivariate approaches
- Dimension reduction techniques
  - Computationally efficient methods
  - Results may be difficult to interpret

    $\Rightarrow$need to impose sparsity
- Variable selection approaches
  - Joint modelling of predictors effects
  - Sparse results
  - Computationally intensive

    $\Rightarrow$ complementary methods to derive OMICs biomarkers

Achievements

- 2 publications



Environmental and Molecular Mutagenesis 00:00–00 (2013)

Review Article

Deciphering the Complex: Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers

Marc Chadeau-Hyam,[1]* Gianluca Campanella,[1] Thibaut Jombart,[2] Leonardo Bottolo,[3] Lutzen Portengen,[4] Paolo Vineis,[1,5] Benoit Liquet,[6] and Roel C.H. Vermeulen[4,7]

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

R2GUESS: a Graphics Processing Unit-based R package for Bayesian variable selection regression of multivariate responses

Benoît Liquet
The University of Queensland, Brisbane

Leonardo Bottolo
Imperial College London

Gianluca Campanella
Imperial College London

Sylvia Richardson
MRC Biostatistics Unit, Cambridge

Marc Chadeau-Hyam
Imperial College London

Achievements

- 2 publications
- Short course: Stat-XP, 1^{st} London edition. Dec 8-12, 2014

Please visit:

`http://www1.imperial.ac.uk/publichealth/education/shortcourses/stat_xp/`

Still a few seats available!!!!!

Achievements

- 2 publications
- Short course: Stat-XP, 1^{st} London edition, Dec 8-12
- Helix-Exposomics interactions
  - One simulation study investigating the applicability of aforementioned methods to exposures and comparing their performances
  - Identifying and FDR control issue when highly correlated exposures (coll. K Strimmer)

$\Rightarrow$ Expected outcome: 2 publications in 2015

'Cross-omic' analyses: some ways forward

- Step-wise procedure
  - Analyse each platform separately
  - Combine candidates from each platform in a single model (clustering approaches, network models)

    $\Rightarrow$ identify/visualise correlation patterns among candidates
- Integrative models: pooling OMICs data
  - Work in progress: testing BVS on a Transcriptomics Proteomic dataset (EGM data)

    $\Rightarrow$ restricted to few biologically relevant pools of proteins

    $\Rightarrow$ need to move towards an hypothesis-free framework
  - Methods: multivariate regression models, networks, canonical correlation algorithms
  - Challenges: dimensionality, interpretability, and correlation among omics data

    $\Rightarrow$ preliminary feature selection (filtering and clustering), or hierarchical framework

Further methodological challenges

- Mechanistic investigations
  - Seq$*$ – WAS: ordered lists of markers associated to exposure and to future disease risk
  - Network models within and across classes indicate how these co-act

    $\Rightarrow$ explore/visualise molecular mechanisms involved
- Investigation the role of age

  1- Use of mother-child cohorts
  - Aim: Identify differential OMICs/Exposure signal within pairs
  - Expected outcome: Candidate signals whose effect is modulated by age

Further methodological challenges

- Mechanistic investigations
  - Seq$*$ – WAS: ordered lists of markers associated to exposure and to future disease risk
  - Network models within and across classes indicate how these co-act

    $\Rightarrow$ explore/visualise molecular mechanisms involved
- Investigation the role of age

  2- Cross-studies investigations
  - Aim: Identify potential age-related effect modifications
  - Model: match participants wrt exposure levels (across studies) and investigate differential OMICs signals
  - Expected outcome: OMICs signals whose level is affected by age

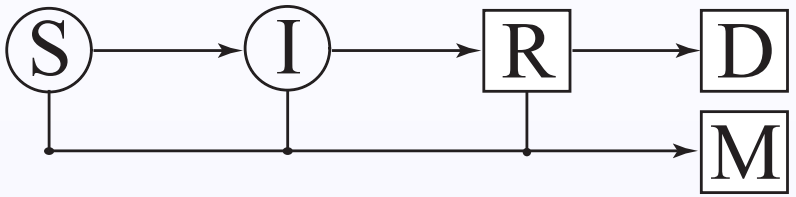
    $\Rightarrow$ towards the identification of age-related susceptibility functions

Further methodological challenges

- Mechanistic investigations
  - Seq$*$ – WAS: ordered lists of markers associated to exposure and to future disease risk
  - Network models within and across classes indicate how these co-act

    $\Rightarrow$ explore/visualise molecular mechanisms involved
- Investigation the role of age

  3- Explicit modelling of the exposure history
  - Methods: Compartmental (multi-state) models using exposure history



  - Parametrisation: age-related susceptibility functions explicitly defined, possible inclusion of OMICs markers

    $\Rightarrow$ quantification of the exposure effect on health outcomes