Understanding complex data

Pietro Ferrari, PhD

Head, Nutritional Methodology and Biostatistics Group
IARC, Lyon, France

Outline

- Metabolomics as an emerging resources
  - The PC-PR2 technique to investigate drivers of variability in large dimension data
  - Biomarker discovery strategy
- The meeting-in-the-middle (MITM) principle
  - Applications from the EPIC study

Metabolomics

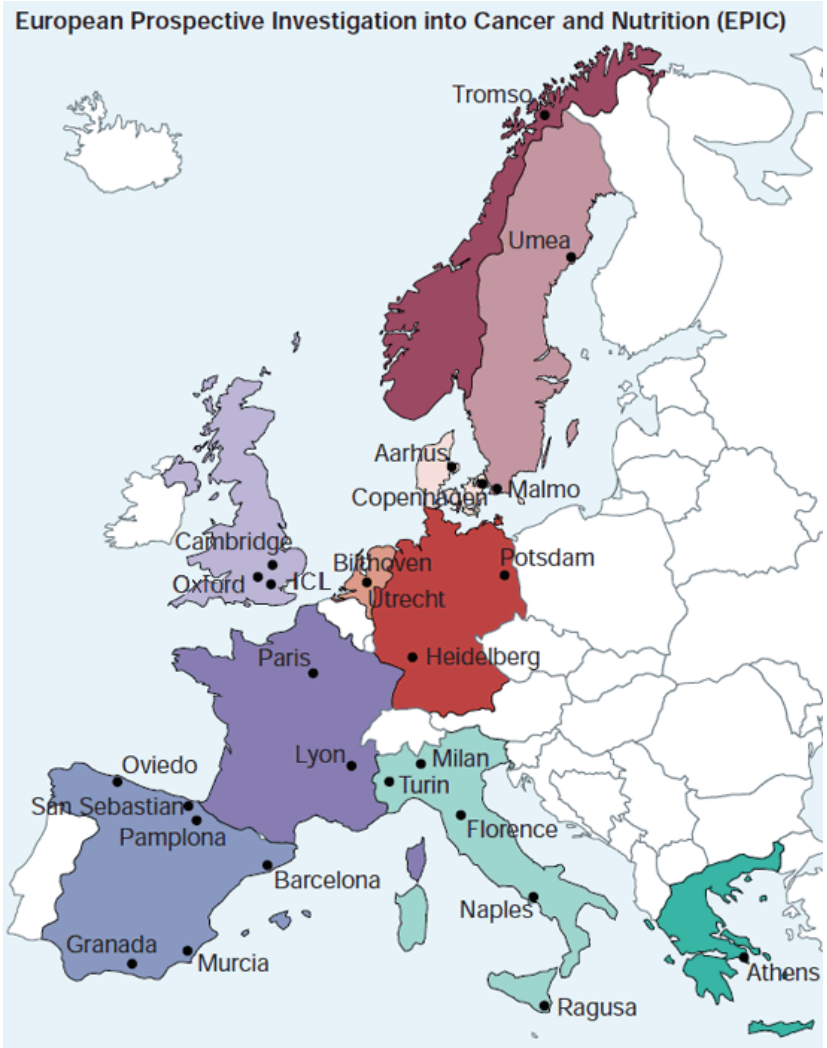
- Comprehensive and quantitative analysis of wide arrays of metabolites in biological samples
- Progressively available in prospective epidemiological investigations

Metabolomics

- Comprehensive and quantitative analysis of wide arrays of metabolites in biological samples
  - Progressively available in prospective epidemiological investigations
- Powerful tool for potential identification of causal pathways in disease development
  - Need for statistical methodologies

The EPIC Study

- Prospective cohort with 500,000 participants from 23 centres
  - Dietary and lifestyle exposures assessed at baseline
  - Biological samples collected at baseline from 80% disease-free participants



International Agency for Research on Cancer

World Health Organization

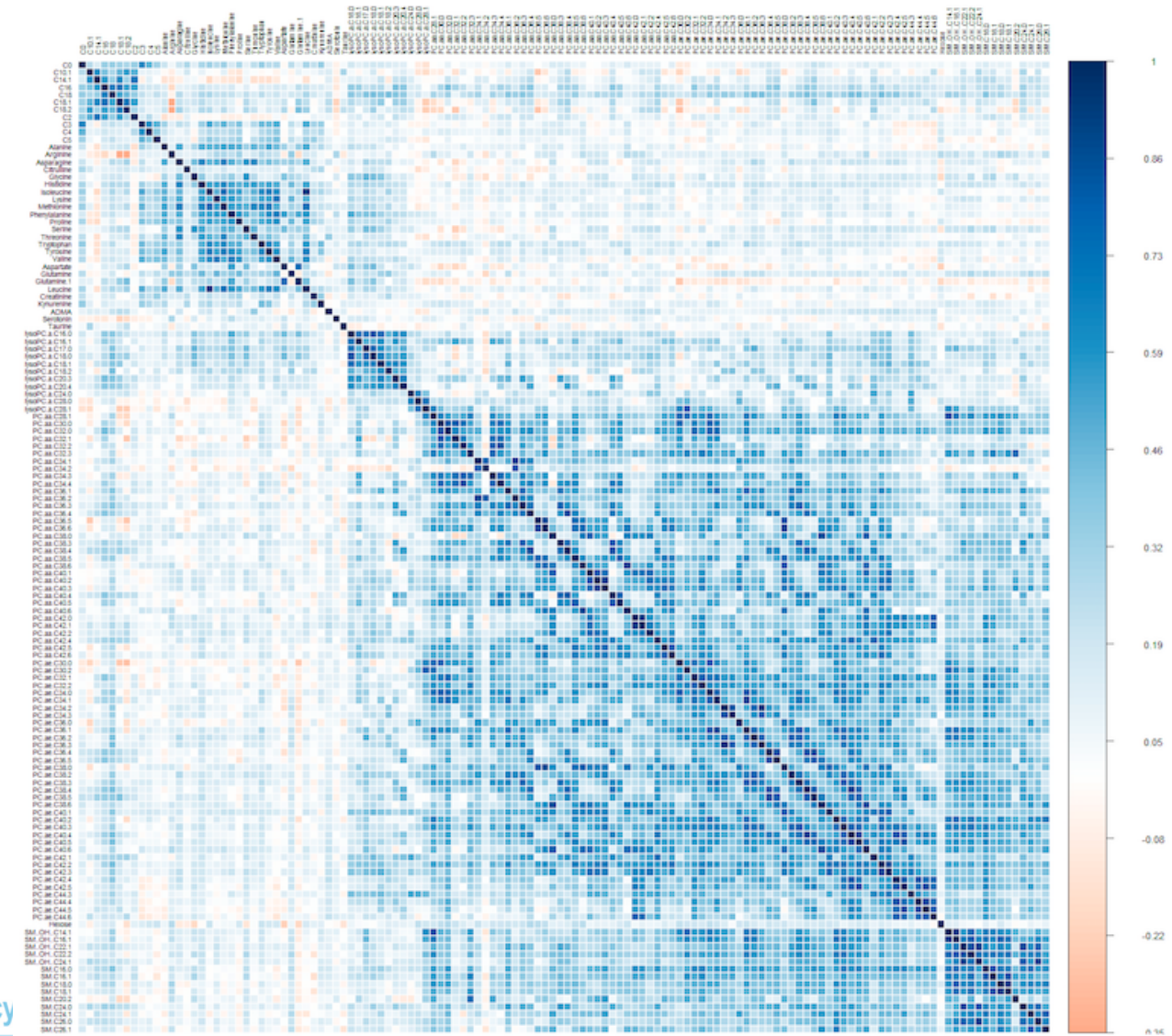
EPIC study on HCC

- Nested case-control study on hepatocellular carcinoma (HCC)
  - 147 cases and 147 matched controls
- 132 blood metabolites acquired in cancer-free individuals with Biocrates AbsoluteIDQ-p180 Kit using UPLC coupled to mass spectrometer

Explore variation in *-omics* data

- Before doing anything else, spend time exploring variability in metabolomics data

Biocrates metabolites



International Agency

World Health Organization

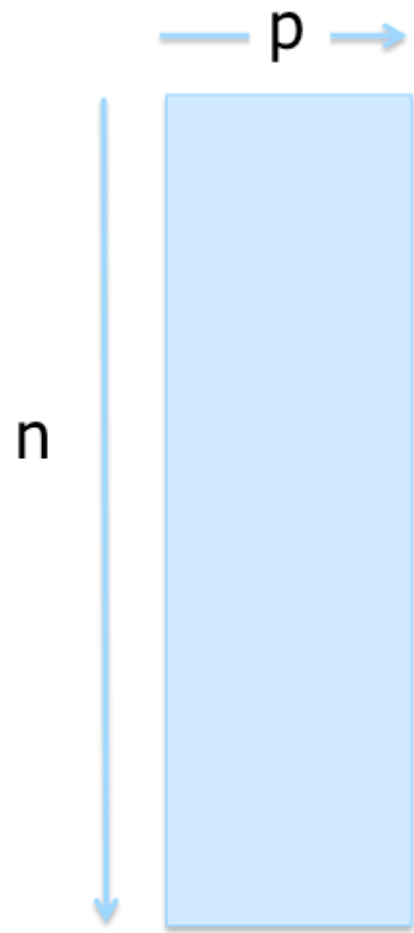
Explore variation in -omics data

- Before doing anything else, spend time exploring variability in metabolomics data
- What drives systematic variation in -omics?

Explore variation in -omics data

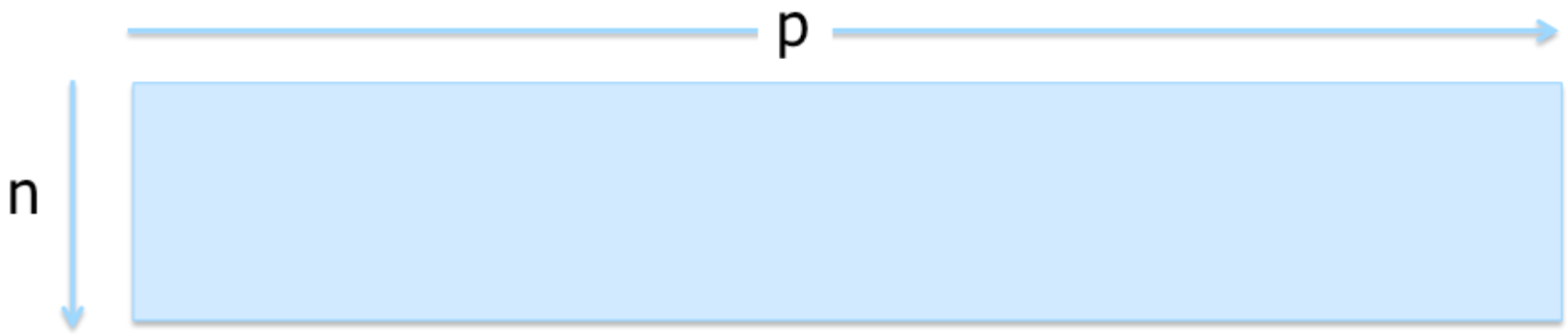
- Before doing anything else, spend time exploring variability in metabolomics data
- What drives systematic variation in -omics?
  - Subjects general characteristics (e.g. age, BMI, smoking status)
  - Variables describing samples' technical treatment (e.g., serum clot contact time, length of storage, fasting status)

The structure of *-omics* data



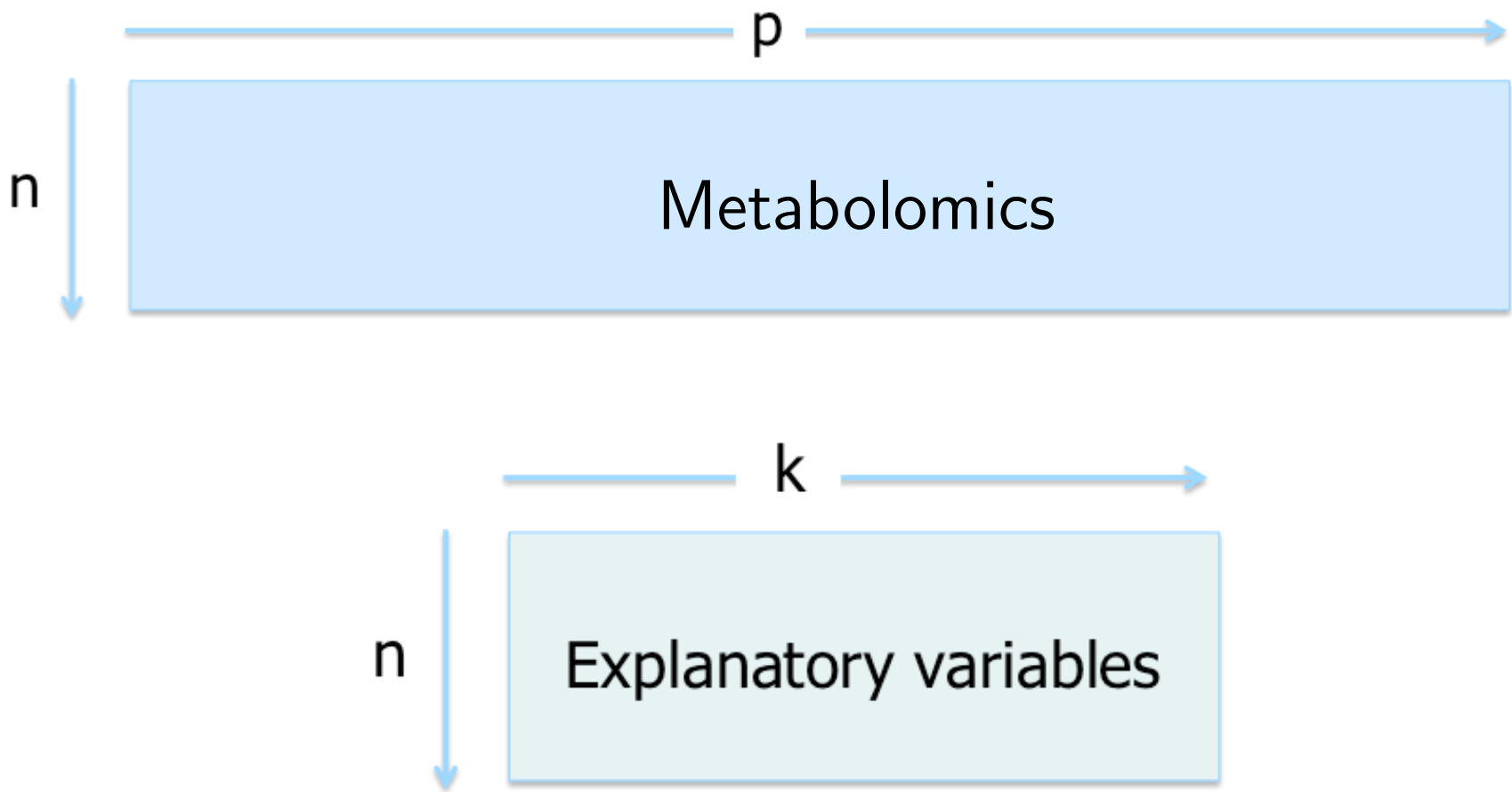
Standard scenario

The structure of *-omics* data (ii)

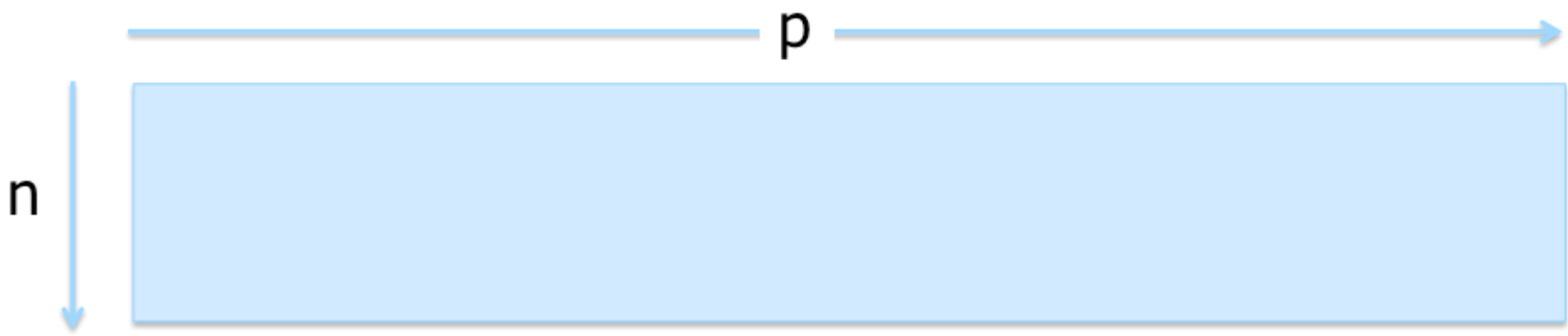


New generation scenario

The structure of *-omics* data (ii)

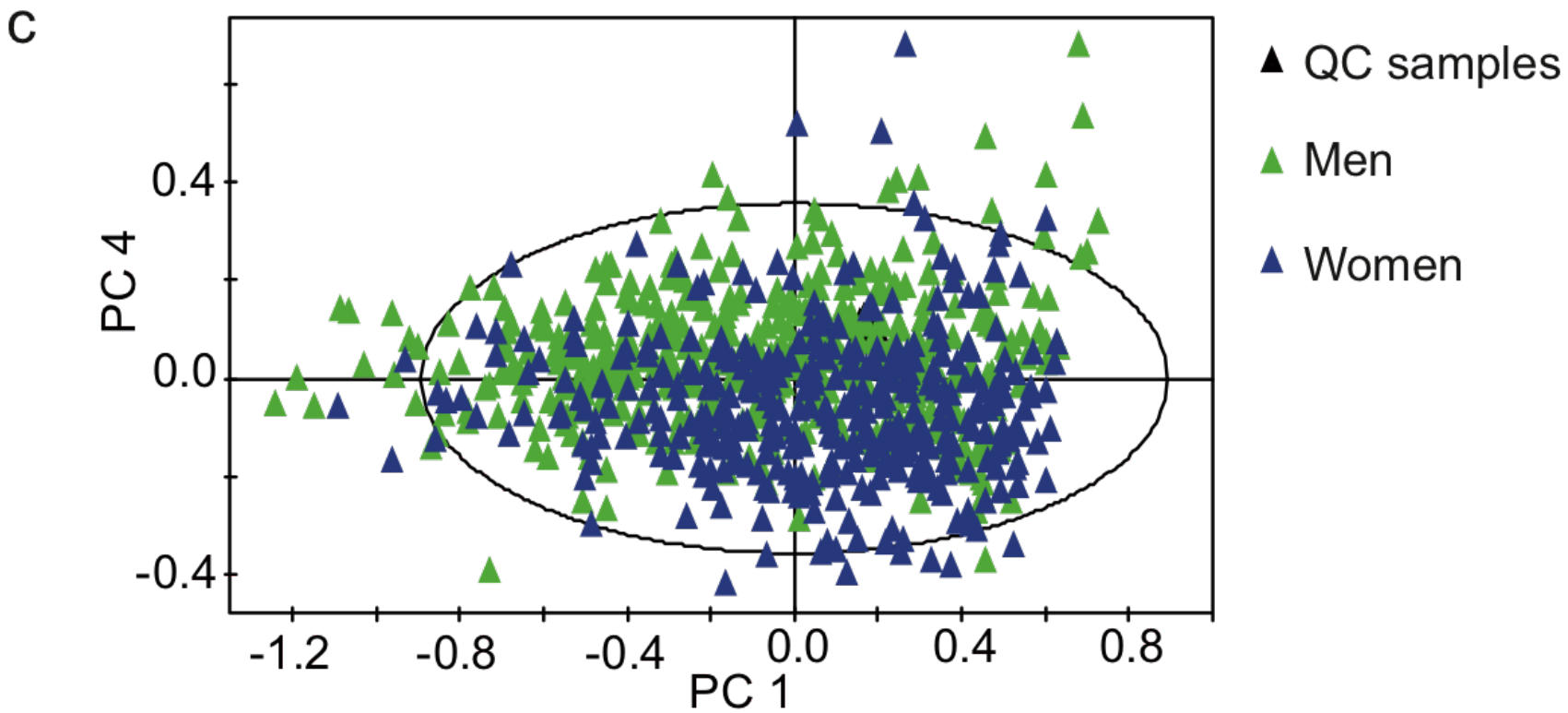


A practical approach



- Run principle component analysis (PCA)
- Retain the first q ($q \ll p$) components

Displaying PCA structure



A new method: PC-PR2

- Only two components at the time are displayed
- Only one explanatory factor can be visualised
- Important inter-correlations between explanatory variables NOT accounted for

A new method: PC-PR2

- Only two components at the time are displayed
- Only one explanatory factor can be visualised
- Important inter-correlations between explanatory variables NOT accounted for
- Principal Component Partial R-square (PC-PR2) method developed

Analytical steps of PC-PR2

(a) Perform PCA on the -omics data

Analytical steps of PC-PR2

(a) Perform PCA on the -omics data

(b) Retain **q** components, explaining -omics variability above a given threshold, ie. 80%

Analytical steps of PC-PR2

(a) Perform PCA on the -omics data

(b) Retain **q** components, explaining -omics variability above a given threshold, ie. 80%

(c) Fit **q** linear regression models, where each component (dependent variables) is explained in terms of **k** explanatory variables

Step (c) of PC-PR2

- For each component q with $(q = 1, \ldots, q_{tot})$

$$PC_{qi} = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_{qi}$$

with (i=1,. . . , n) and $\epsilon_{qi} \sim N(0, \sigma^2_{\epsilon q})$,

Analytical steps of PC-PR2 (ii)

(c) Fit **q** linear regression models, where each component (dependent variables) is explained in terms of **k** explanatory variables

(d) Determine the $R^2_{partial(q,k)}$ for each explanatory variable, in each model q

Step (d) of PC-PR2

- The partial R-square ($R^2_{partial}$) is a statistics quantifying the amount of variability of the dependent variable that each explanatory variable X_k contributes to explain, conditional on the other (**k-1**) variables in the model

$$R^2_{partial(\mathbf{X_1})q}$$

Step (d) of PC-PR2

- The partial R-square ($R^2_{partial}$) is a statistics quantifying the amount of variability of the dependent variable that each explanatory variable X_k contributes to explain, conditional on the other (**k-1**) variables in the model

$$R^2_{partial(\mathbf{X_1})q}, \; R^2_{partial(\mathbf{X_2})q}$$

Step (d) of PC-PR2

- The partial R-square ($R^2_{partial}$) is a statistics quantifying the amount of variability of the dependent variable that each explanatory variable X_k contributes to explain, conditional on the other ($\mathbf{k\text{-}1}$) variables in the model

$$R^2_{partial(\mathbf{X_1})q},\ R^2_{partial(\mathbf{X_2})q},\ \ldots,\ R^2_{partial(\mathbf{X_k})q}$$

Analytical steps of PC-PR2 (ii)

(c) Fit **q** linear regression models, where each component (dependent variables) is explained in terms of **k** explanatory variables

(d) Determine the $R^2_{partial(q,k)}$ for each explanatory variable, in each model q

(e) For each **k** variable, determine an overall $R^2_{partial(k)}$ as a weighted average, using eigenvalues as weights

Step (e) of PC-PR2

- For each variable **k**, over the **q** PCs:

Step (e) of PC-PR2

- For each variable **k**, over the **q** PCs:

$$R^2_{partial(\mathbf{X_1})} = \sum R^2_{partial(\mathbf{X_1})q} w_q$$

Step (e) of PC-PR2

- For each variable **k**, over the **q** PCs:

$$R^2_{partial(\mathbf{x_1})} = \sum R^2_{partial(\mathbf{x_1})q} w_q$$

$$R^2_{partial(\mathbf{x_2})} = \sum R^2_{partial(\mathbf{x_2})q} w_q$$

Step (e) of PC-PR2

- For each variable **k**, over the **q** PCs:

$$R^2_{partial(\mathbf{x_1})} = \sum R^2_{partial(\mathbf{x_1})q} w_q$$

$$R^2_{partial(\mathbf{x_2})} = \sum R^2_{partial(\mathbf{x_2})q} w_q$$

$$\ldots$$

$$R^2_{partial(\mathbf{x_K})} = \sum R^2_{partial(\mathbf{x_K})q} w_q$$

Step (e) of PC-PR2 (ii)

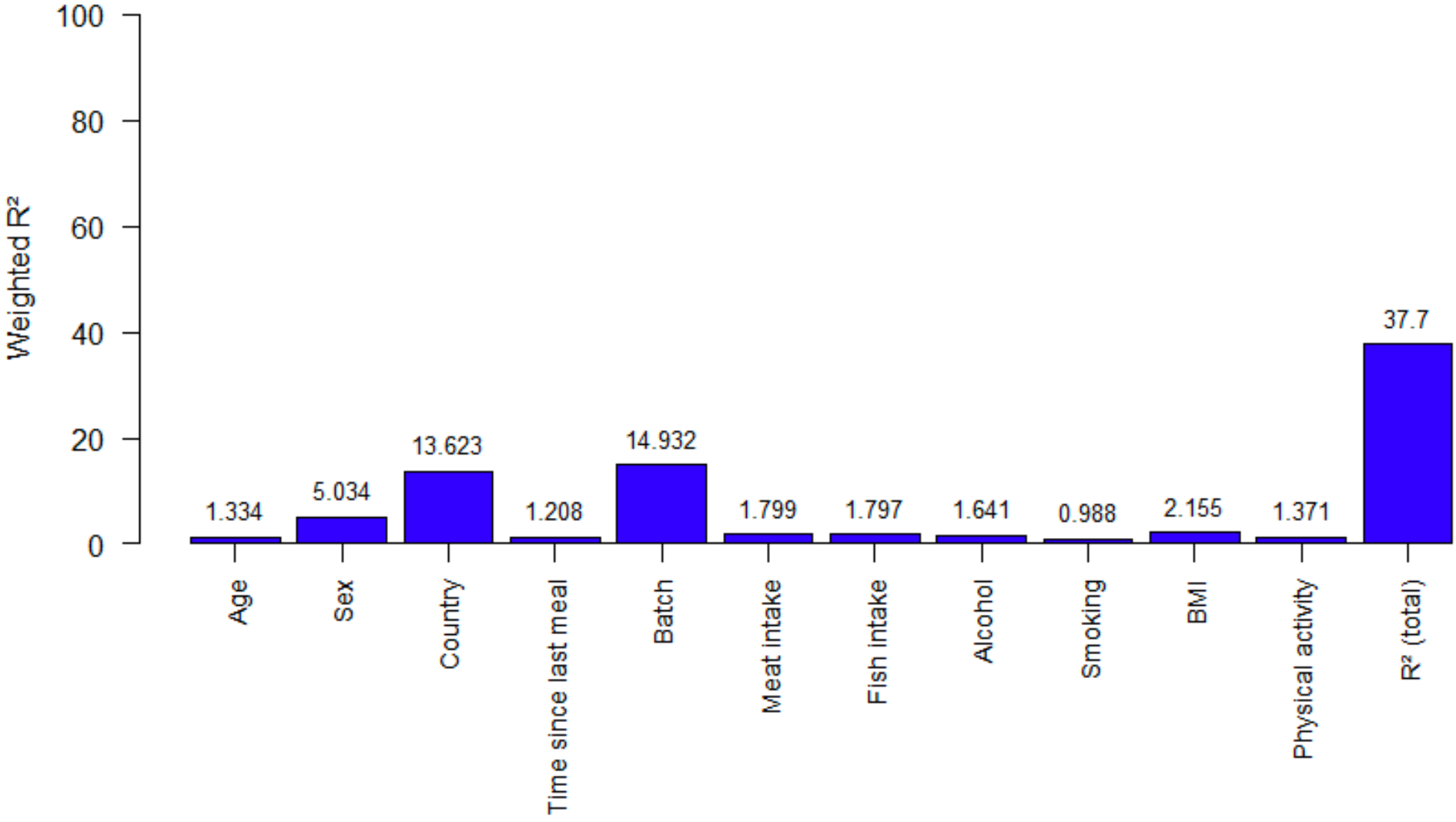
- For each variable **k**, over the **q** PCs:

$$R^2_{partial(\mathbf{X_K})} = \sum R^2_{partial(\mathbf{X_K})q} w_q$$

- where $w_q = \lambda_q / \sum \lambda_q$, for $(q = 1, \ldots, q_{tot})$

  with $\lambda_q =$ eigenvalue that expresses the amount of variability captured by PC_q

The PC-PR2 analysis



(Fages & Ferrari *et al.*, Metabolomics, 2014)

Biomarker discovery

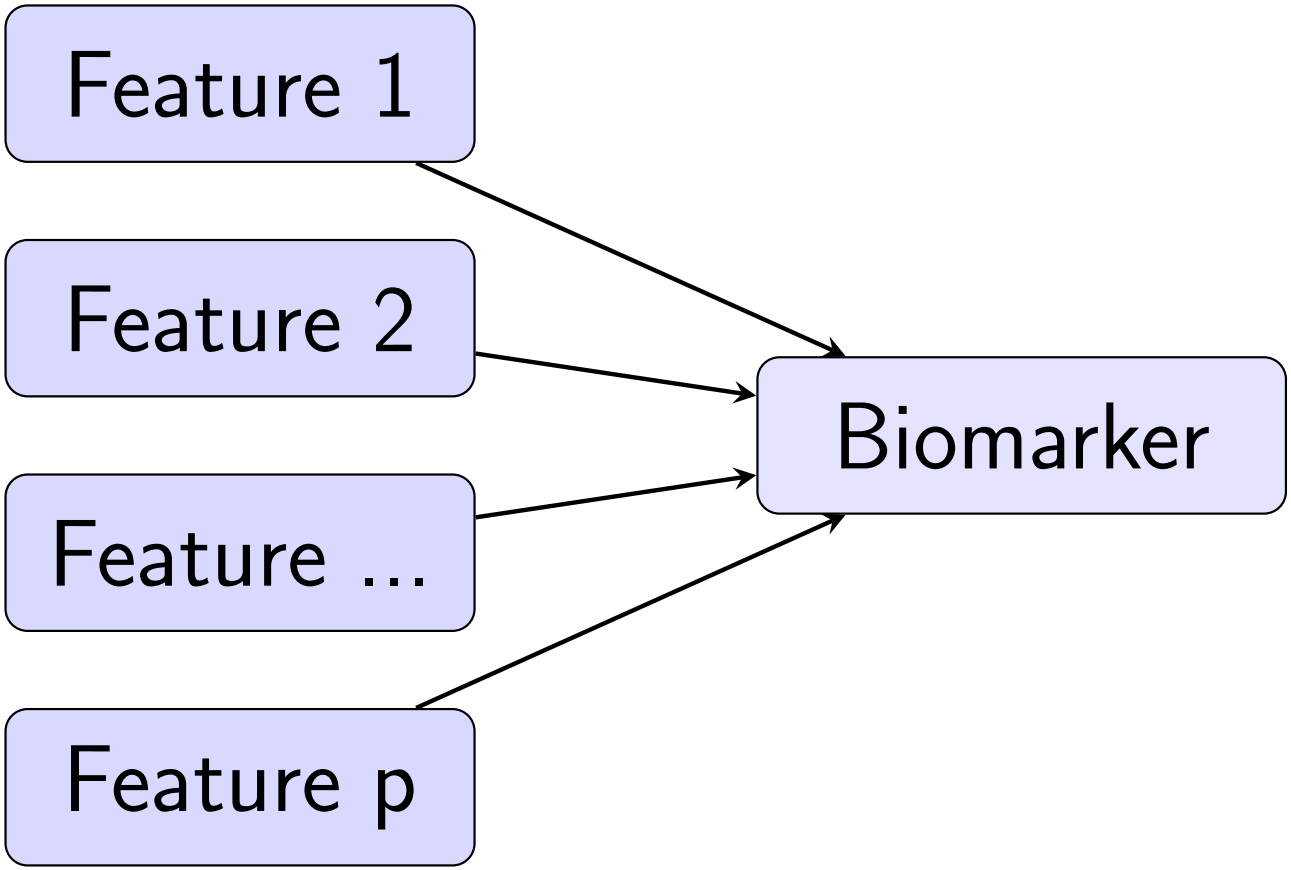
Feature 1

Feature 2

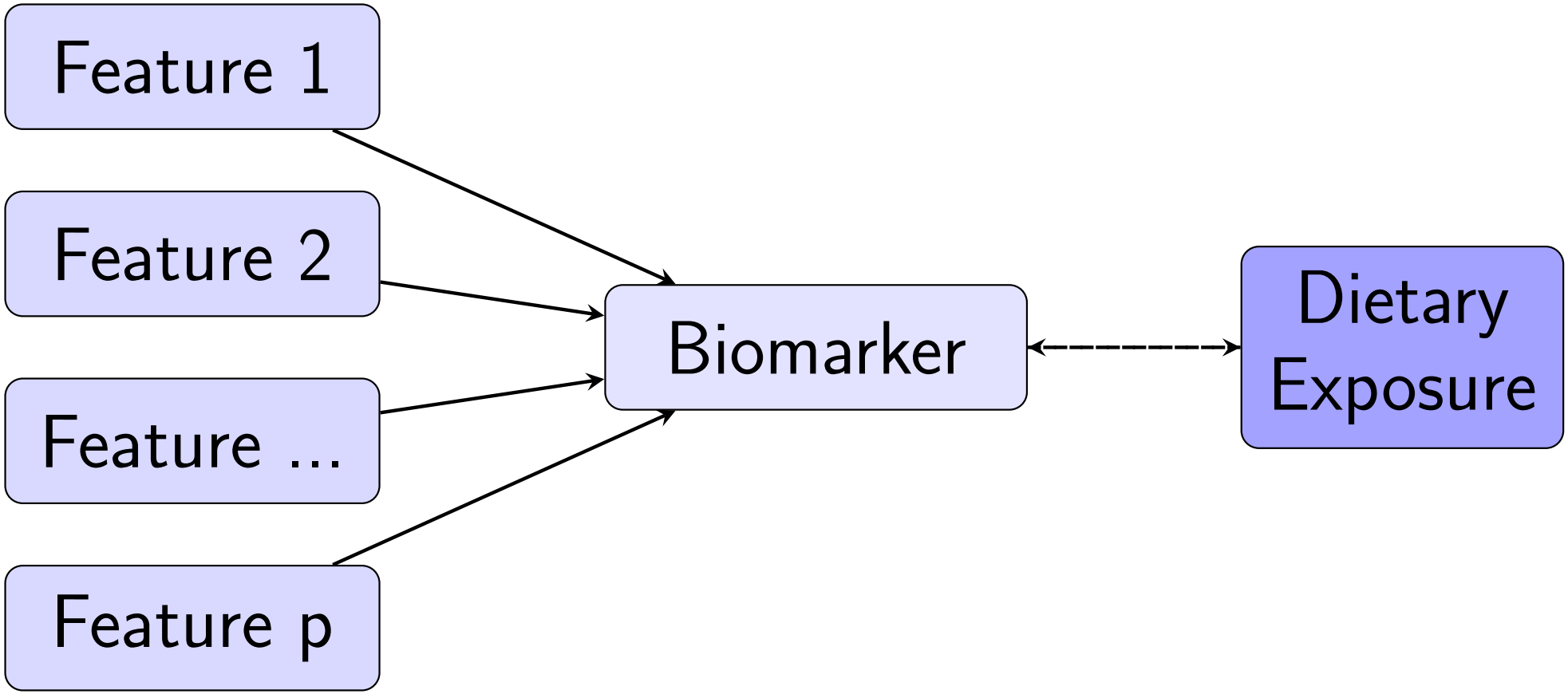
Feature ...

Feature p

Biomarker discovery



Biomarker discovery



Diet and polyphenols, n=475

Exposure[1]	Selected PP[2]	$\hat{r}_{adj}$	AUC[3] (95% CI)
Coffee	Single PP (caffeic acid)	0.42	86% (78%, 94%)
	PP by LASSO (p=11)	0.51	89% (83%, 95%)
Red wine	Single PP (Gallic acid EE)	0.66	89% (84%, 95%)
	PP by LASSO (p=2)	0.66	89% (84%, 95%)

[1] 24-hour dietary recall measurements; [2]Urinary poliphenols metabolites measured by UPLC-ESI-MS/MS in 24-hour urine; [3]Estimated by cross-validation in test and training sets.

(Noh *et al.*, submitted to JN)

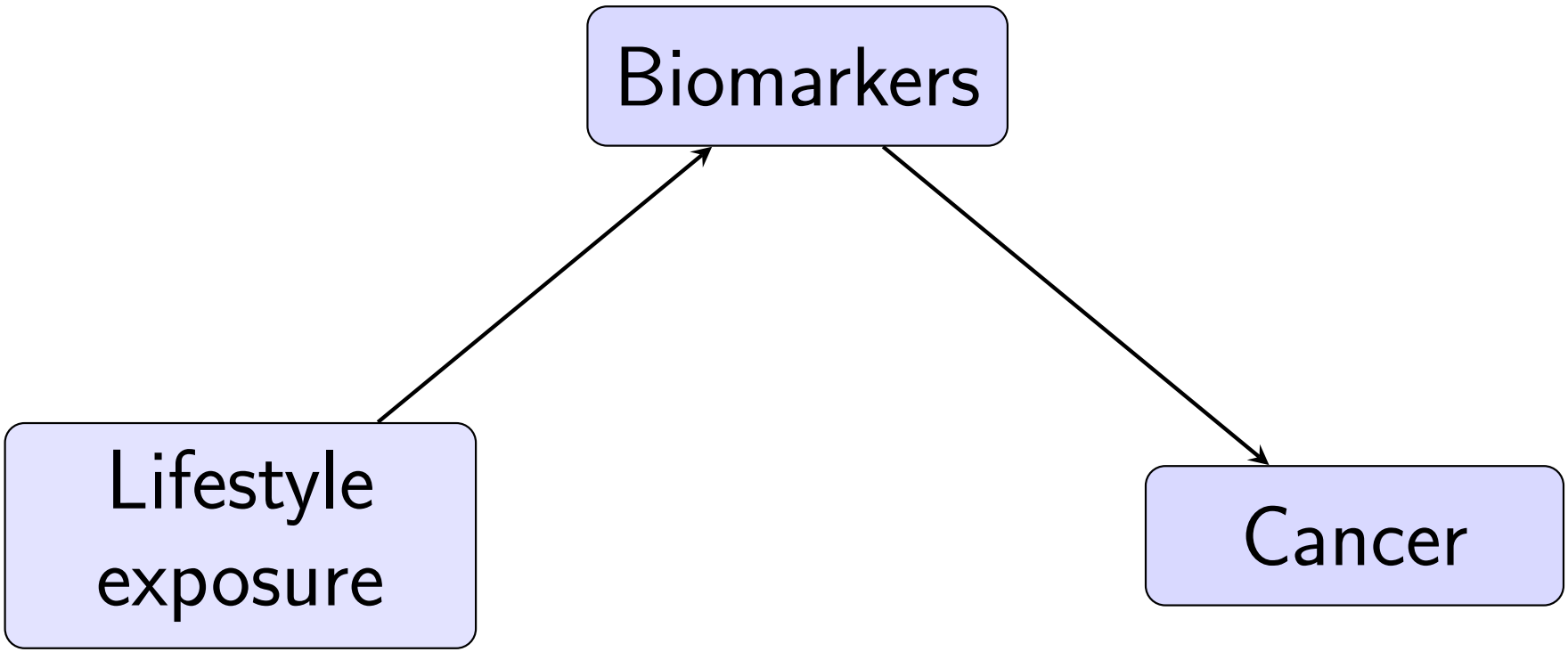
Biomarker discovery (ii)

Exposure[1]	Selected PP[2]	$\hat{r}_{adj}$	AUC[3] (95% CI)
Coffee	Single PP (caffeic acid)	0.42	86% (78%, 94%)
	PP by LASSO (p=11)	0.51	89% (83%, 95%)
Red wine	Single PP (Gallic acid EE)	0.66	89% (84%, 95%)
	PP by LASSO (p=2)	0.66	89% (84%, 95%)

[1] 24-hour dietary recall measurements; [2]Urinary poliphenols metabolites measured by UPLC-ESI-MS/MS in 24-hour urine; [3]Estimated by cross-validation in test and training sets.

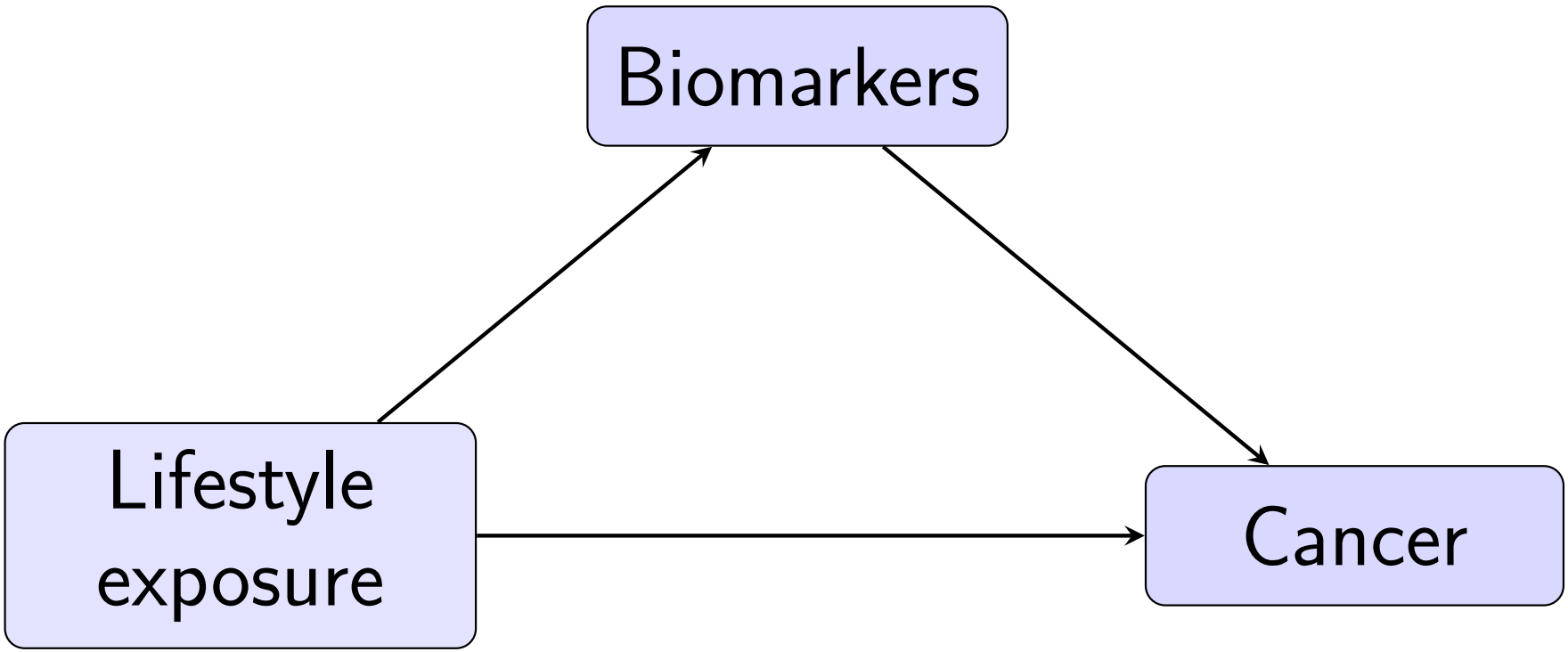
(Noh *et al.*, submitted to JN)

Meeting-in-the-middle (MITM)



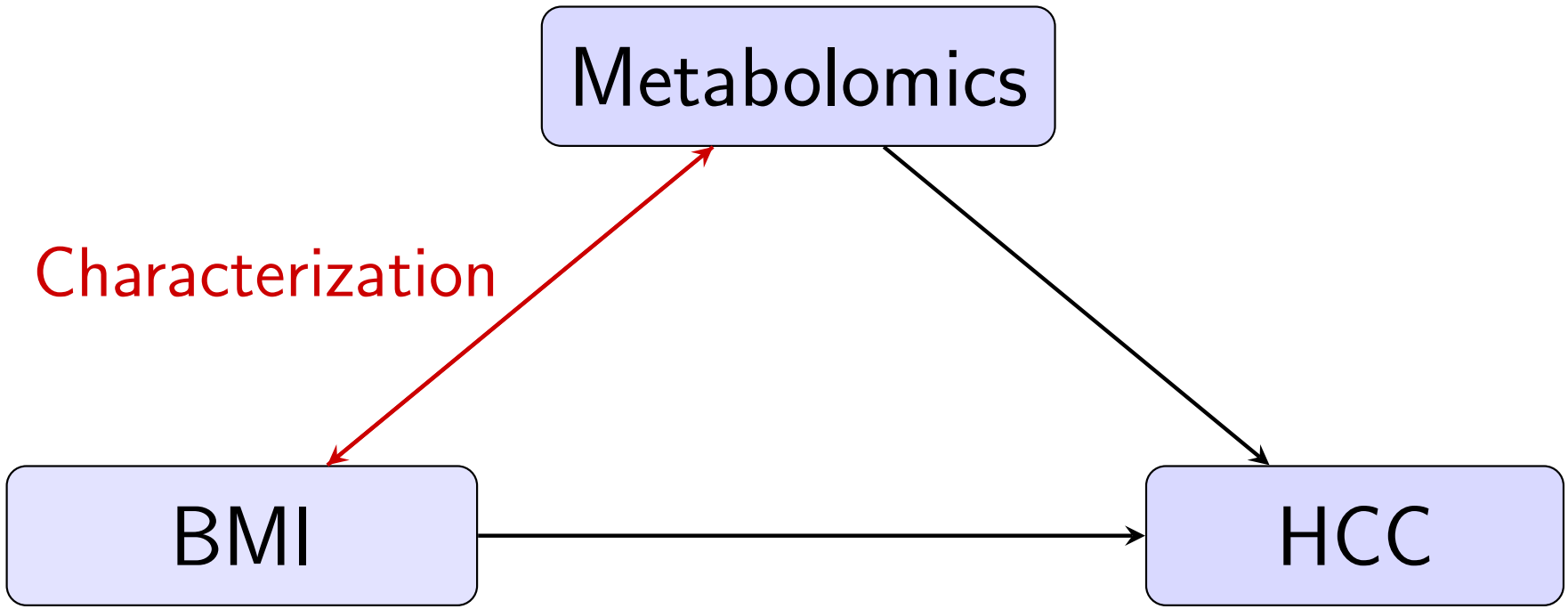
(Vineis & Perera, CEBP, 2007)

Meeting-in-the-middle (MITM)

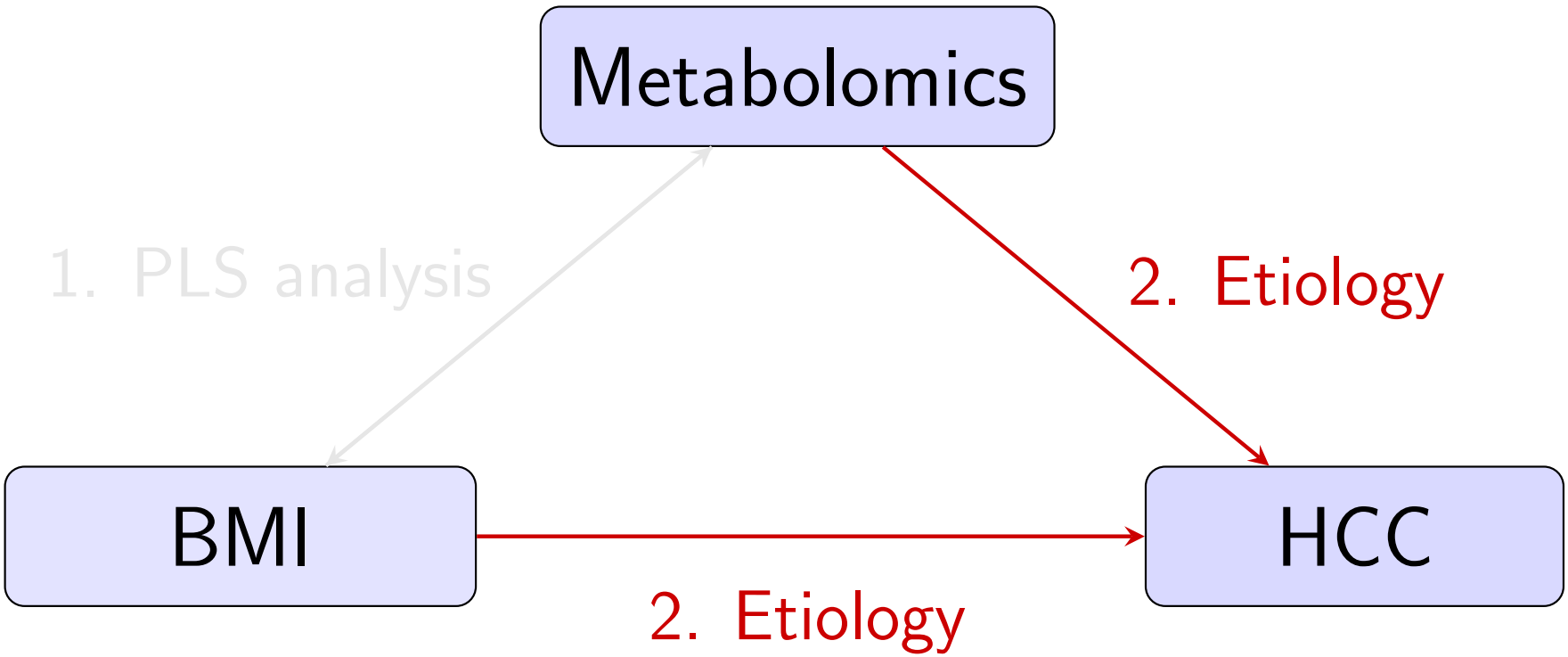


(Assi *et al.*, Mutagenesis, 2015)

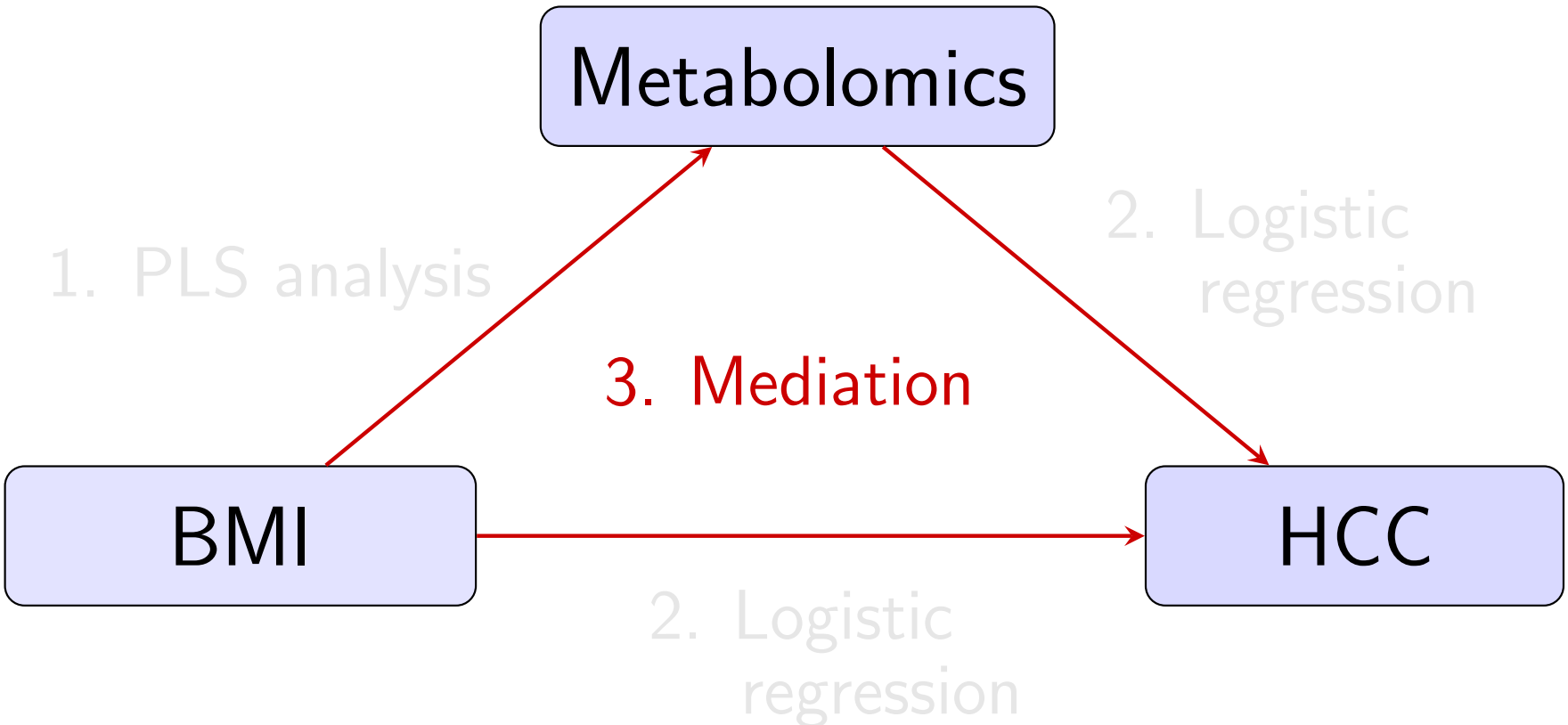
Analytical strategy of MITM



Analytical strategy of MITM



Analytical strategy of MITM



1. BMI-driven PLS factor

Metabolites	Loadings
Glutamine	-0.19
Glutamate	0.23
Tyrosine	0.24
Lyso PC a C17:0	-0.22
Lyso PC a C18:2	-0.23
PC ae C36:2	-0.20
Liver function score	0.19

(Assi *et al.*, submitted to PLOS Medicine)

2. BMI and HCC

Exposure	OR[1] (95%CI)	p_{value}
BMI	1.23 (0.93, 1.62)	0.149
Metabolites, PLS score	4.04 (2.22, 7.36)	4.8E-07
% Mediated[2]	100	

[1] Expressing HCC relative risk estimate of 1-SD increase in the PLS score; [2] Estimated as ln(NIE)/[ln(NIE)+ln(NDE)], with NIE=natural indirect effect, NDE=natural direct effect.

(Assi *et al.*, submitted to PLOS Medicine)

2. BMI and HCC

Exposure	OR[1] (95%CI)	p_{value}
BMI	1.23 (0.93, 1.62)	0.149
Metabolites, PLS score	4.04 (2.22, 7.36)	4.8E-07
% Mediated[2]	100	

[1] Expressing HCC relative risk estimate of 1-SD increase in the PLS score; [2] Estimated as ln(NIE)/[ln(NIE)+ln(NDE)], with NIE=natural indirect effect, NDE=natural direct effect.

(Assi *et al.*, submitted to PLOS Medicine)

Limitations

- Sample size, accuracy, reliability
- Mediation analysis offers a framework to link inter-correlated factors, but . . .

Limitations

- Sample size, accuracy, reliability
- Mediation analysis offers a framework to link inter-correlated factors, but ...
- It involves massive use of underlying assumptions, i.e. chronological sequence, confounding structure

Concluding remarks

- Strategies that tackle the complexity of dietary (and lifestyle) exposures are to be commended
  - Vastly unexplored potential of *-omics* data
- But again, clearly biology is way more complex than statistical modeling
  - Key to create multi-disciplinary settings, with a common language

Acknowledgments

- Anne Fages, Hwayoung Noh, Nada Assi (NMB Group), Vivian Viallon (Lyon I University)
- Pekka Keski-Rahkonen, Augustin Scalbert, Marc Gunter, Mazda Jenab, (NME Section)
- Paolo Vineis and all EPIC PIs