Cancer epidemiology in the "big-data" era

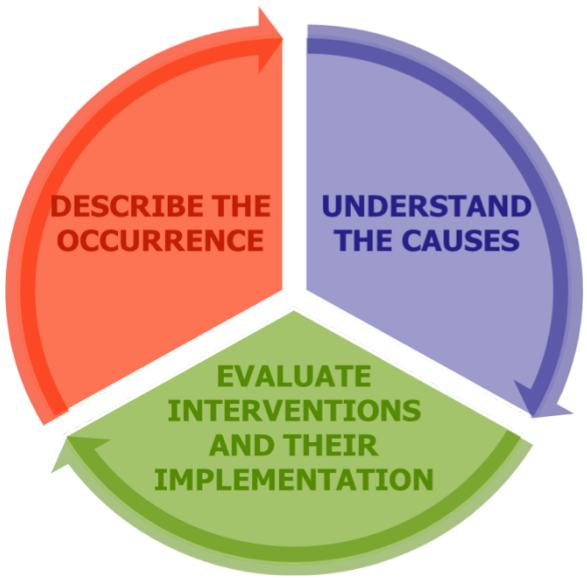
Vivian Viallon

Scientist (Statistician)
NMB Group, IARC

Big Data In Medicine: Challenges and Opportunities
CERN, June 7, 2019

International Agency for Research on Cancer
World Health Organization

International Agency for Research on Cancer (IARC)

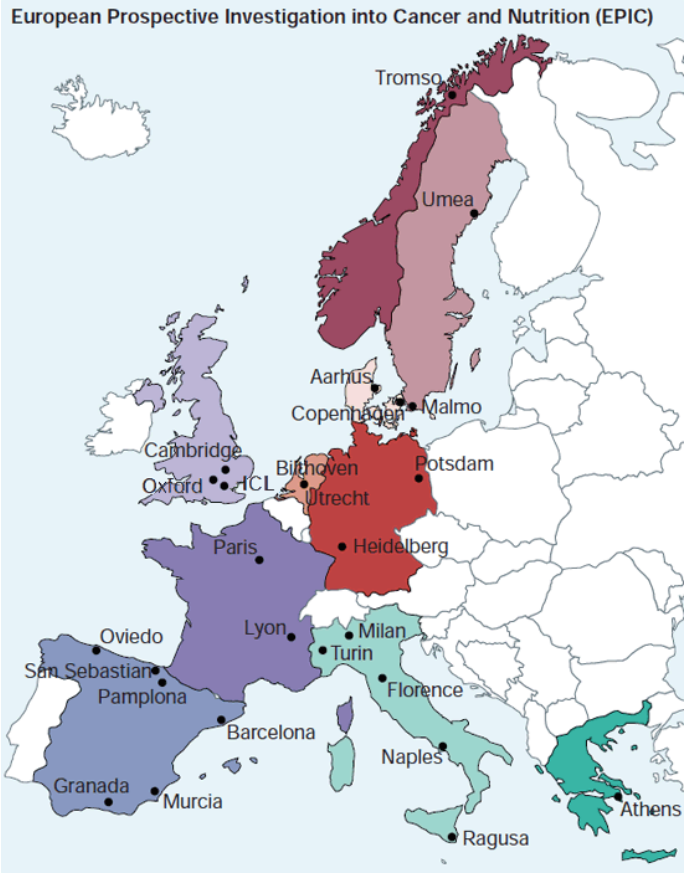


- Established in 1965 in Lyon (France)
- Autonomous agency of WHO
- To promote international collaboration in cancer research
- Epidemiology:
  - observational cohort studies
  - to study causes of cancer...

International Agency for Research on Cancer
World Health Organization

The EPIC cohort (IARC and Imperial College)

European Prospective Investigation into Cancer and Nutrition study



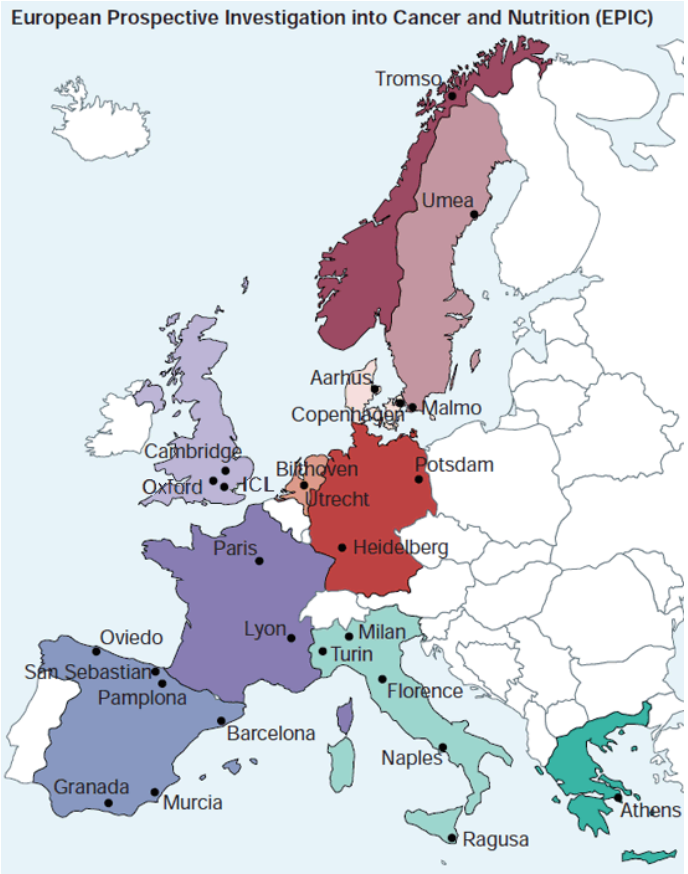
- $>$ 500,000 participants
- recruited over 1992-1999
- across 10 European countries

International Agency for Research on Cancer

World Health Organization

The EPIC cohort (IARC and Imperial College)

European Prospective Investigation into Cancer and Nutrition study

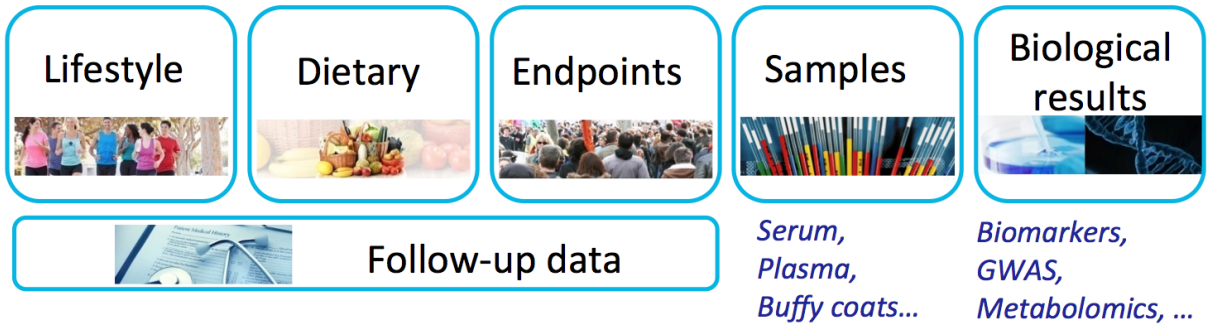


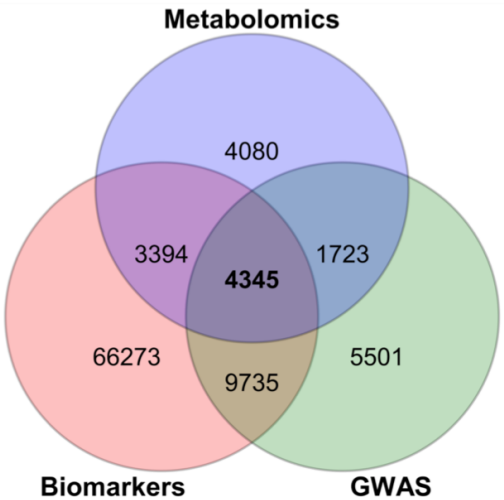
- $>$ 500,000 participants
- recruited over 1992-1999
- across 10 European countries
- to study the associations between
  - lifestyle and environmental factors
  - and the incidence of cancer and other chronic diseases (CVD, diabetes)
- $>$ 15 years of follow-up (still ongoing)
  - $\sim$ 50,000 incident cancers

International Agency for Research on Cancer

World Health Organization

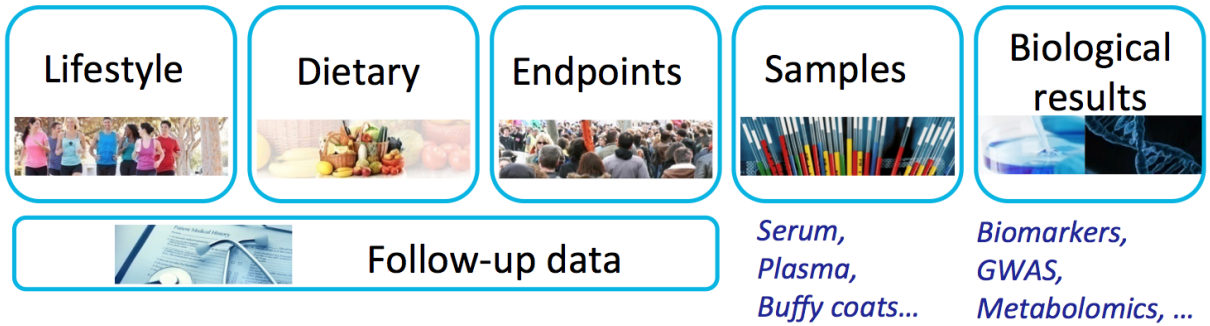
Data available in EPIC

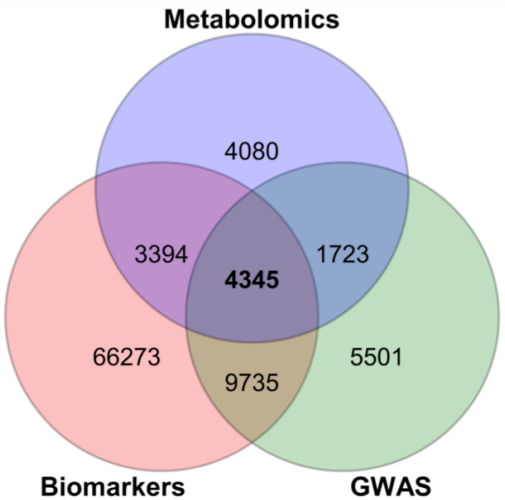




- Ex.: ~ 50,000 EPIC subjects with GWAS data

Data available in EPIC





- Ex.: $\sim$ 50,000 EPIC subjects with GWAS data
- for now, "large p"; in the sense $n \gg p$;
  - $n \sim 10^2 - 10^4$;
    $p \sim 10^2 - 10^6$

Bigger data: ~ To – Po

- "large n, large-huge p":
  - UK Biobank (500K ind., ~15K cancers):
    - GWAS for all participants
    - whole exome sequencing for 50,000 participants.
  - Mutographs: whole genome sequencing of 10,000 cancer patients

- "huge n": primary care data (SIDIAP – Catalonia)
  - (virtually) describes the complete medical history for (virtually) the entire population
  - blood test results, BMI, etc.
  - Rk: E3N + SNDS in France

New opportunities?

- omics data
  - to identify genetic determinants of diseases
  - to investigate biological pathways underlying the exposure-cancer relationship: [Assi et al., Mutagenesis, 2015].
  - to assess the causal interpretation of observed association through Mendelian Randomization.
  - to investigate new cancer subtypes (using multi-omics clustering of tumors; [Alcala et al., Nature Communications, to appear])

New opportunities?

- omics data
  - to identify genetic determinants of diseases
  - to investigate biological pathways underlying the exposure-cancer relationship: [Assi et al., Mutagenesis, 2015].
  - to assess the causal interpretation of observed association through Mendelian Randomization.
  - to investigate new cancer subtypes (using multi-omics clustering of tumors; [Alcala et al., Nature Communications, to appear])

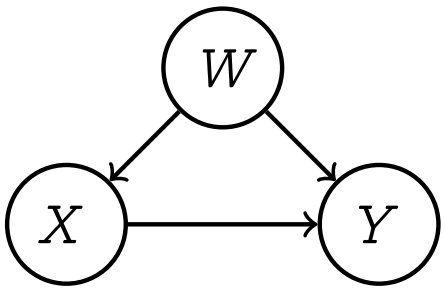
- primary care data
  - to study the effect of previous conditions, use of treatments, etc. on the occurence of cancer, etc.
  - to adjust for medical history and get more reliable causal effect estimates when studying one exposure of interest

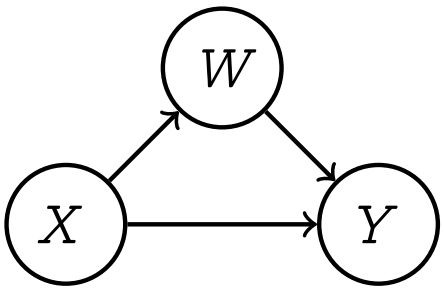
Challenges?

- Causal validity of statistical associations?
  - measurement error; e.g., -omics: unwanted variation due to batch effects, fasting status,...
  - confounding/collider bias: agnostic research

Challenges?

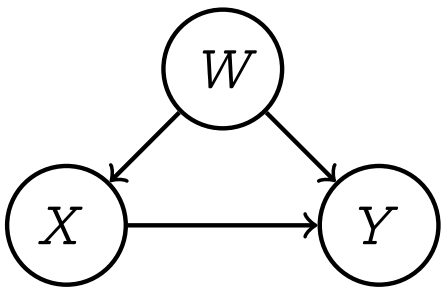
- Causal validity of statistical associations?
  - measurement error; e.g., -omics: unwanted variation due to batch effects, fasting status,...
  - confounding/collider bias: agnostic research

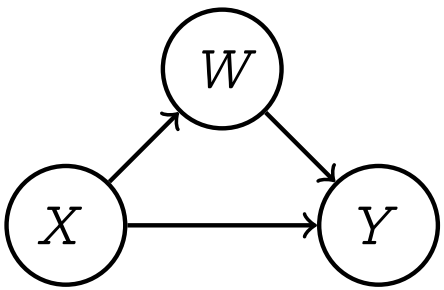




Challenges?

- Causal validity of statistical associations?
  - measurement error; e.g., -omics: unwanted variation due to batch effects, fasting status,...
  - confounding/collider bias: agnostic research

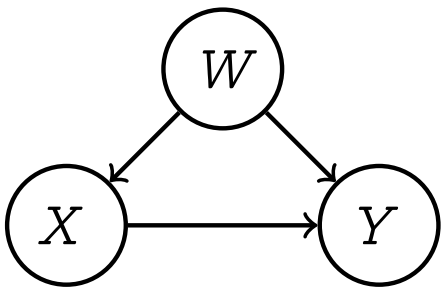


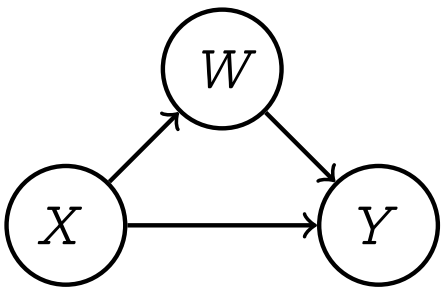


  - selection bias; e.g., primary care datasets
    - blood test results are only known for individuals who had blood tests.

Challenges?

- Causal validity of statistical associations?
  - measurement error; e.g., -omics: unwanted variation due to batch effects, fasting status,...
  - confounding/collider bias: agnostic research





  - selection bias; e.g., primary care datasets
    - blood test results are only known for individuals who had blood tests.
- (partial?) solutions exist...
  - PC-PR2 [Fages et al. Metabolomics, 2014], DAG structure estimation [Maathuis et al, Ann. Statist, 2009], ..., weighted approaches [Cole and Hernan, Am. J. Epid., 2008]

The curse of dimensionality: $n \gg p$

- Consider the simple case where
  - $Y \in \{0, 1\}$: outcome (e.g., cancer occurence)
  - $\mathbf{X} = (X_1, \ldots, X_p)$: pre-processed omics features
  - (we omit potential confounders here for simplicity).

The curse of dimensionality: $n \gg p$

- Consider the simple case where
  - $Y \in \{0, 1\}$: outcome (e.g., cancer occurence)
  - $\mathbf{X} = (X_1, \ldots, X_p)$: pre-processed omics features
  - (we omit potential confounders here for simplicity).

- omics signature discovery: $S = \mathbf{X}^T \boldsymbol{\beta}^*$
  - use, e.g., a logistic model: find $(\alpha^*, \boldsymbol{\beta}^*) \in \mathbb{R}^{p+1}$ such that

$$\Pr(Y = 1|\mathbf{X}) = g(\mu^* + \mathbf{X}^T \boldsymbol{\beta}^*),$$

where $g(x) = [1 + \exp(-x)]^{-1}$ is the sigmoid function.

Towards penalized approaches

- when $n \gg p$, standard estimation procedures are inefficient.
  - errors of standard estimates/predictions: $\mathcal{O}\big(\min(p/n, 1)\big)$
  - $\sim$ over-fitting

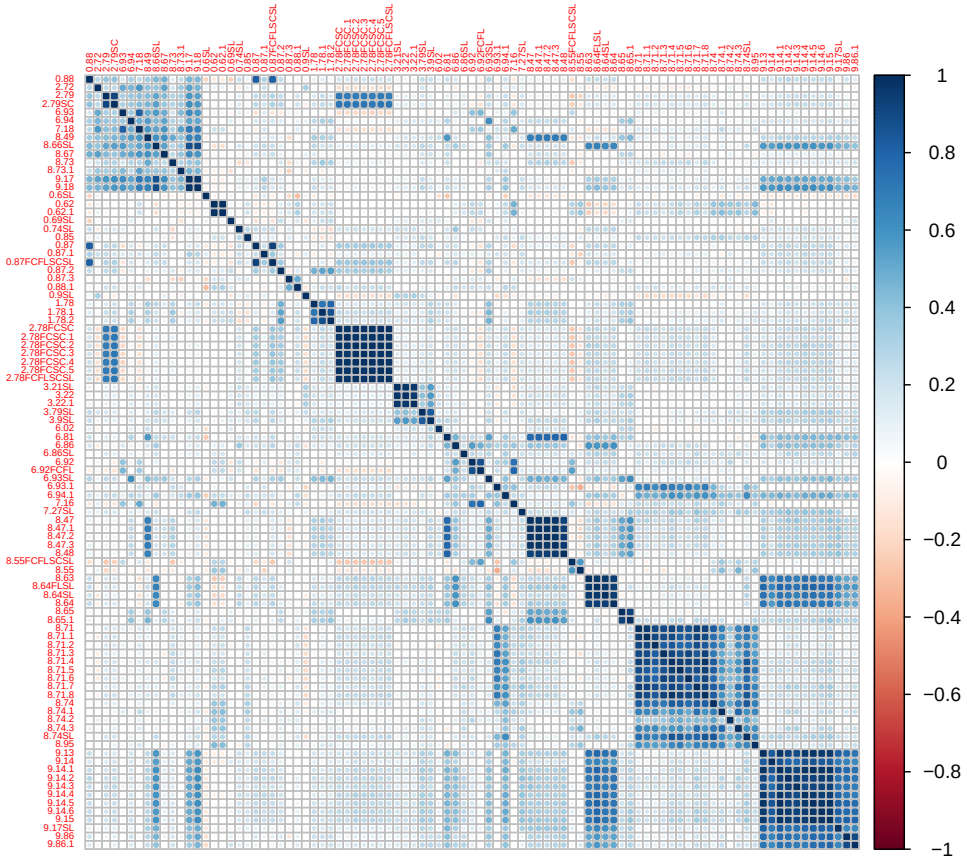
$\Rightarrow$ use prior knowledge/assumption to reduce model "complexity".

- Sparsity assumption: $\boldsymbol{\beta}^*$ is "sparse", i.e.
  - only a small fraction of its components are non-null
  - only a few omics features are related to cancer

  $\Rightarrow$ the Lasso, based on an L_1-penalty (Tibshirani, 1996, JRSS-B).

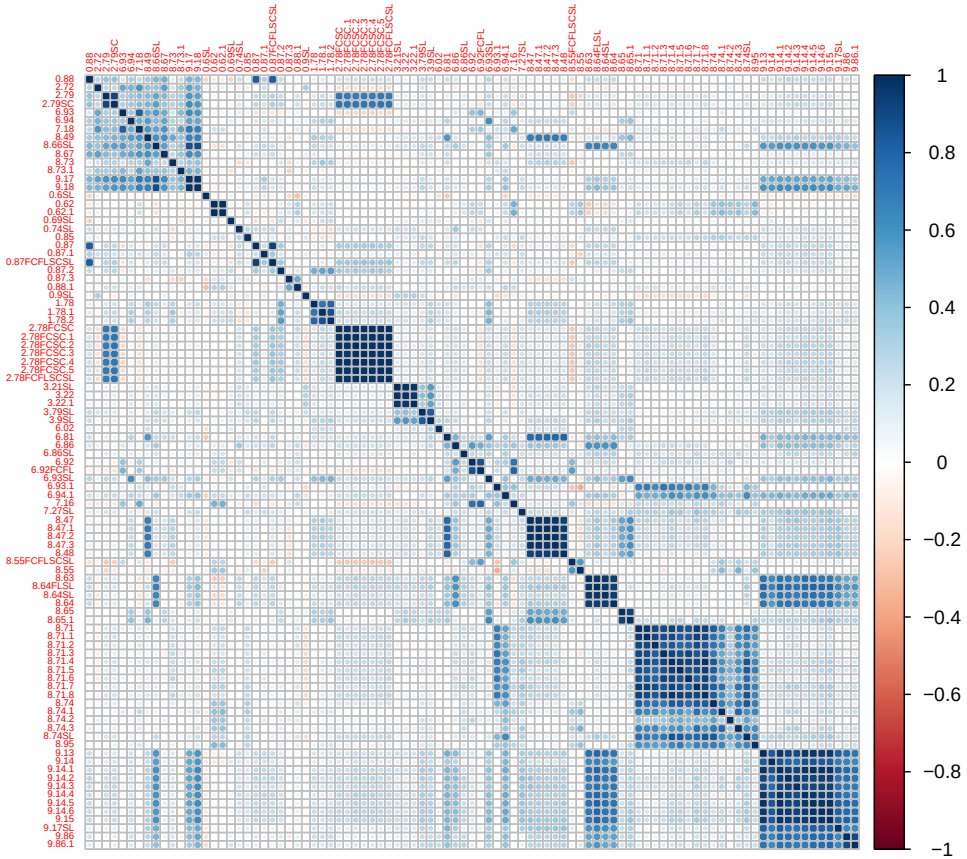
$$\hat{\boldsymbol{\beta}}(\lambda) \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmax}} \{\mathcal{L}(\boldsymbol{\beta}) - \lambda \|\boldsymbol{\beta}\|_1\},$$

Beyond sparsity: group structure



- Some groups appear:
  - several features from the same metabolite
  - several metabolites from the same nutrient, ...

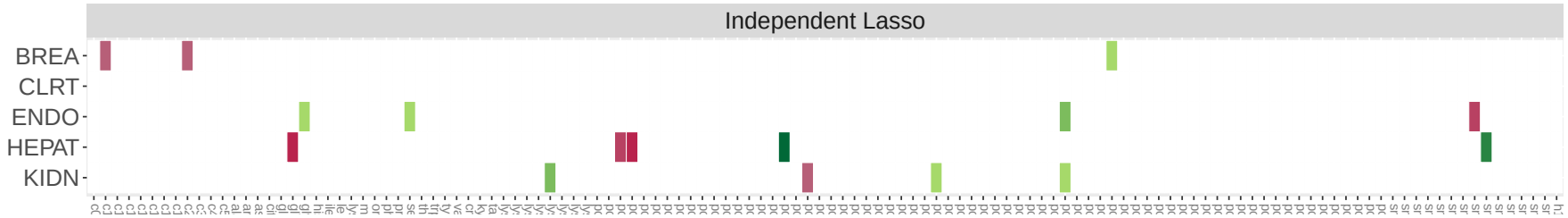
Beyond sparsity: group structure



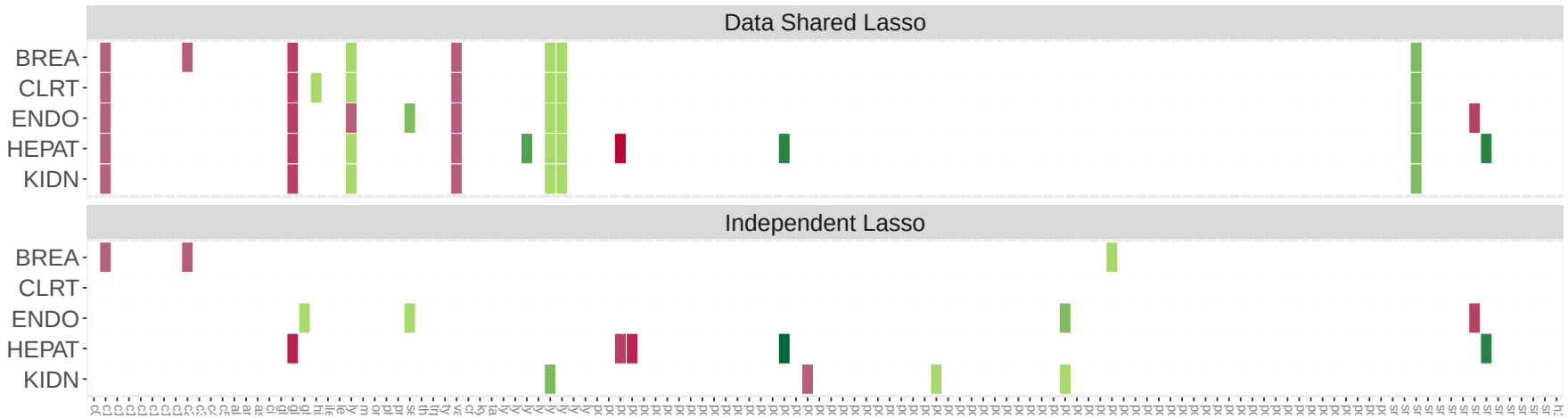
- Some groups appear:
  - several features from the same metabolite
  - several metabolites from the same nutrient, ...
- Similar correlation patterns for other omics

$\Rightarrow$ Selection at the group level (e.g., with the group lasso)

Beyond sparsity: commonalities across cancer sites



Beyond sparsity: commonalities across cancer sites



- here, "structure" is two-fold:
  - each vector $\boldsymbol{\beta}_k^*$ is sparse
  - vectors $\boldsymbol{\beta}_k^*$ look alike

Conclusion

- Increasingly large data available in cancer epidemiology
- offer great opportunities
- may raise data storage issues soon
- already raise computational issues
  - ML approaches are computationally intensive; e.g.,
    - one lasso for one value of λ is intensive
    - many λ values have to be tested (e.g., when several λ are involved)
    - sub-sampling for optimal λ value ; highly parallelizable though
- causal validity?
  - estimation of high-dimensional DAG structures
  - need for repeated measurements (time-varying omics data)