Epidemiological considerations to results interpretation in the age of « multi-omics science »

Dr. Cyrille Delpierre,

Director of EQUITY Team (Embodiment, social ineQualities, lifecoUrse epidemiology, cancer and chronIc diseases, intervenTions, methodology), UMR1027 – French Institute of Health and Medical Research (Inserm)*/ University Paul Sabatier*

Toulouse, France

Inserm
La science pour la santé
From science to health

UNIVERSITÉ TOULOUSE III PAUL SABATIER
Université de Toulouse

Genome wide association studies

- Main objective
  - Which DNA regions explain the phenotype
  - In reality to identify DNA regions associated with one specific phenotype
- Huge development…

Omics science: a developing field…

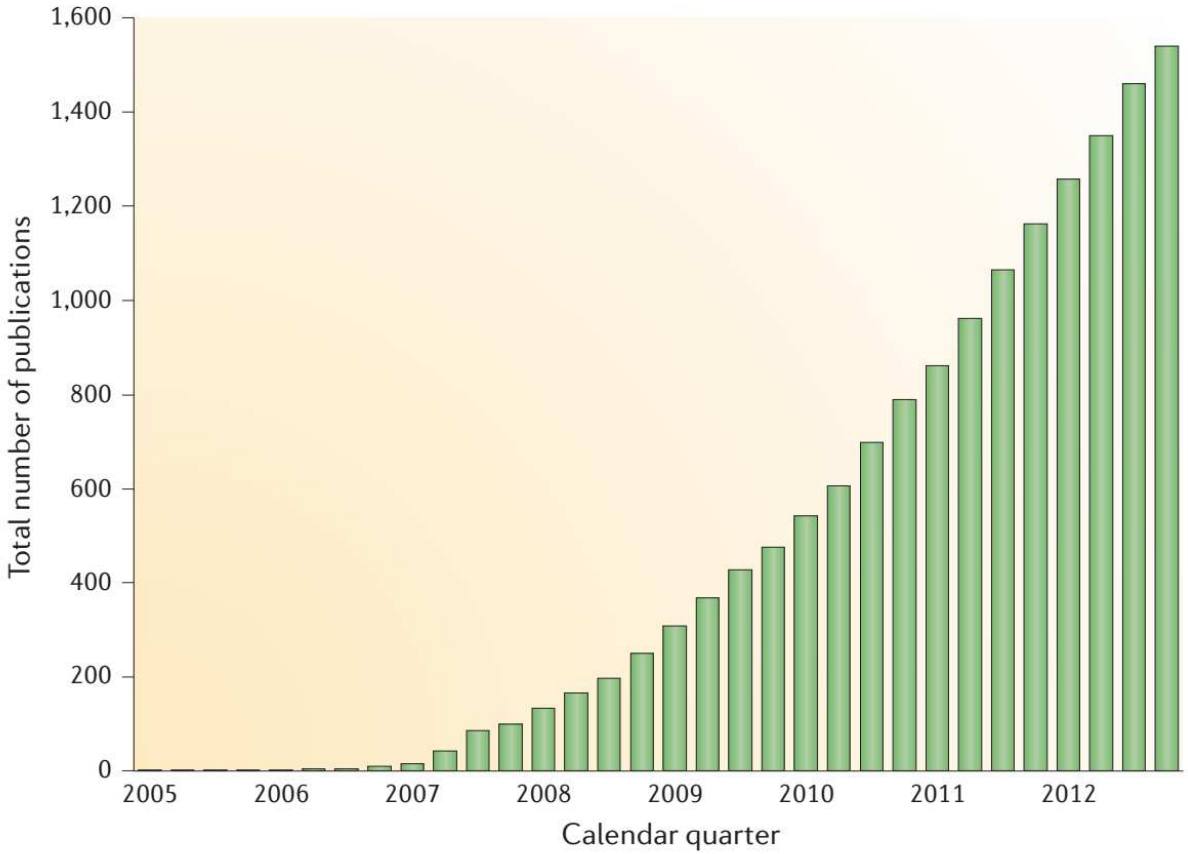
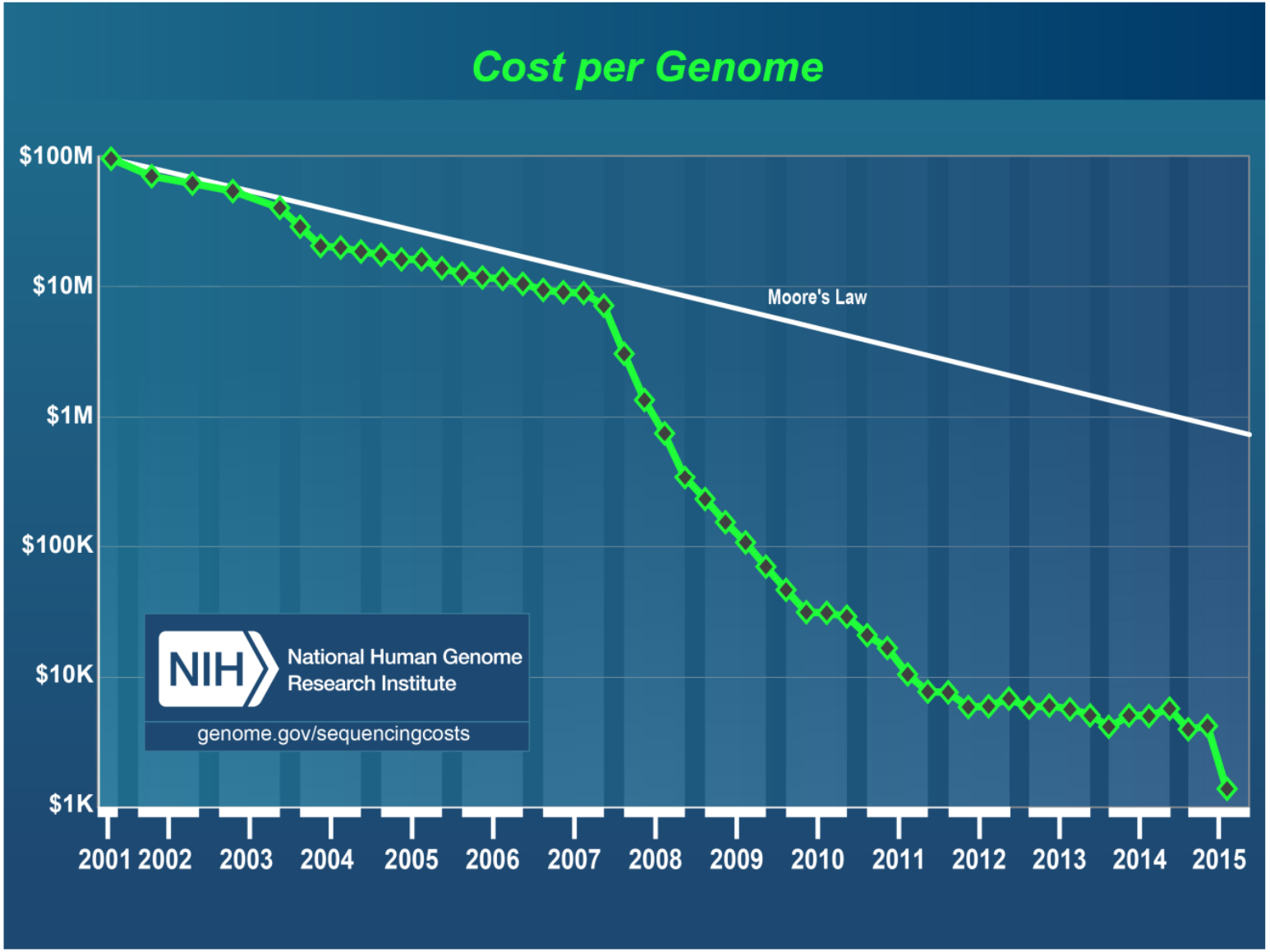


Figure 1 | **Pace of genome-wide association study publications since 2005.** The pace of genome-wide association study (GWAS) publications has increased dramatically over 7 years and shows no signs of slowing. The figure is based on data from the US National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies.

Manolio TA. Nature Review Genetics. 2013

Genome wide association studies

- Main objective
  - Which DNA regions explain the phenotype
  - In reality to identify DNA regions associated with one specific phenotype
- Huge development … and increasingly cheaper


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

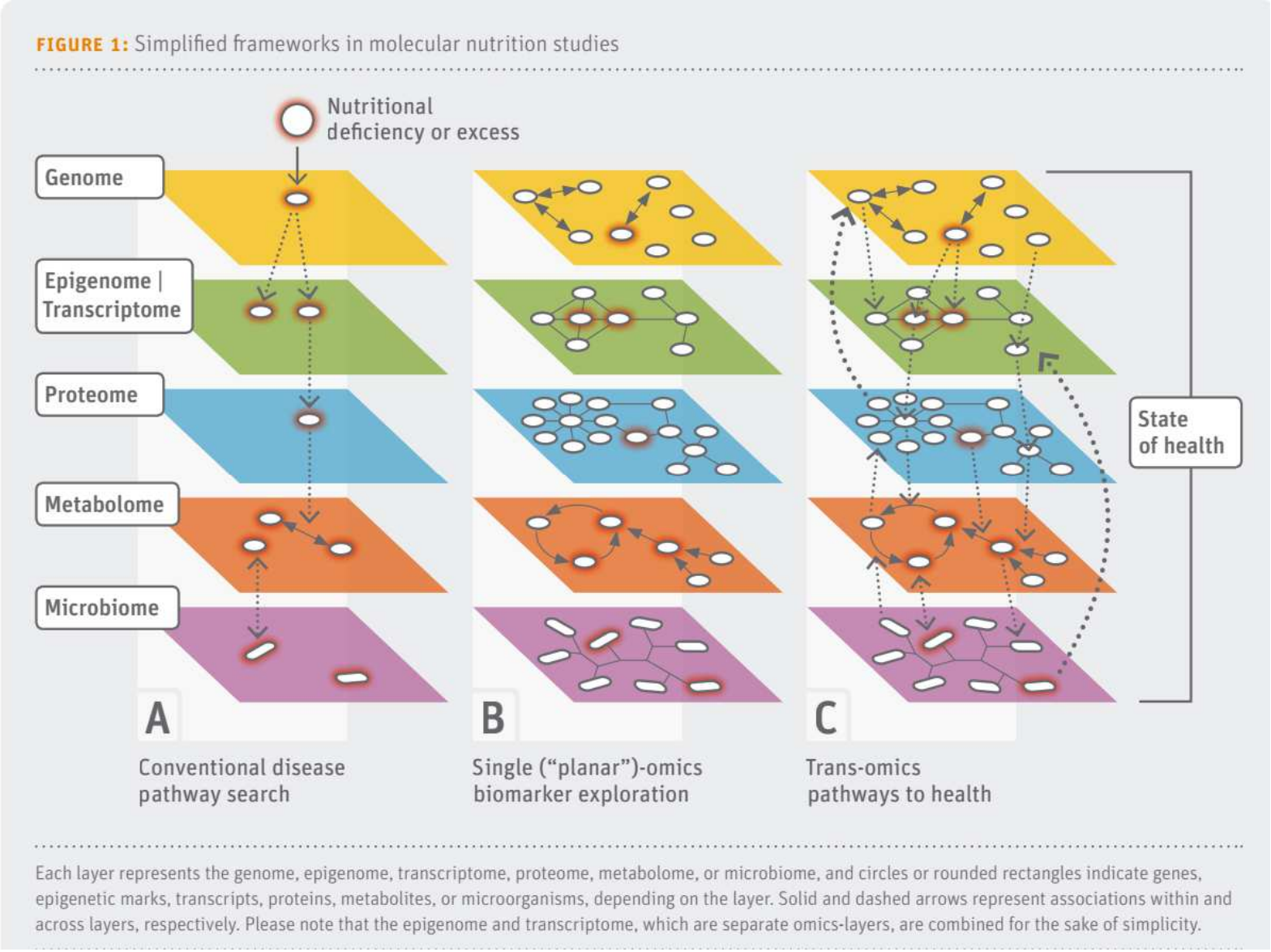
Genome wide association studies

- Main objective
  - Which DNA regions explain the phenotype
  - In reality to identify DNA regions associated with one specific phenotype
- Huge development … and increasingly cheaper
- With some sucesses
  - Mapping of thousands of genetic variants contributing to disease (Hasin et al. Genome Biology 2017)
- But with important weaknesses
  - Most variants discovered have relativelly small effect: OR rarely above 1.3 (Price et al. Proc R Soc B 2015)
  - Explain small part of heritability, often 10% or less (Manolio TA. Nature Reviews Genetics. 2013): missing heritability

⇨ Development of integrative analysis using multiple omics data

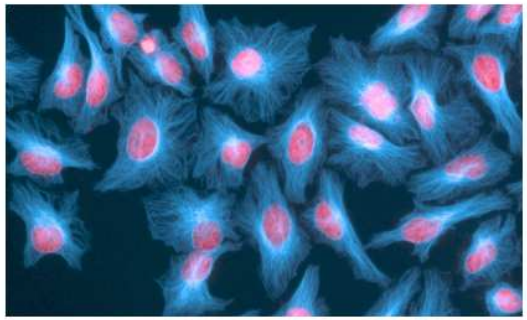
↳ The case of the majority of the projects in this summer school

Multi-omics approaches to disease

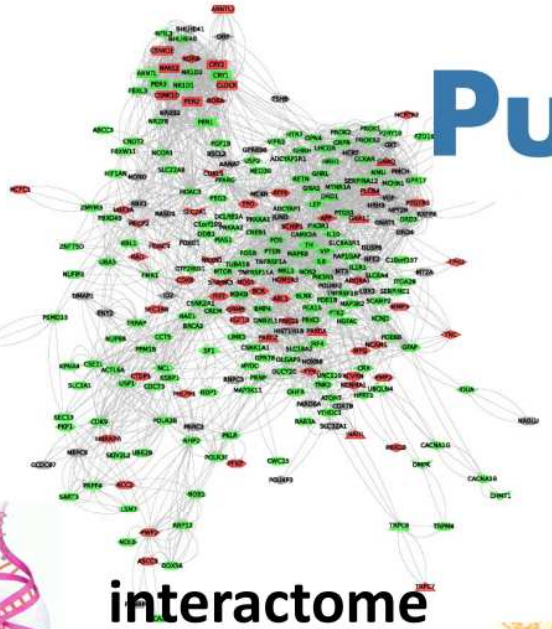


Lee SE. Sight and life. 2017

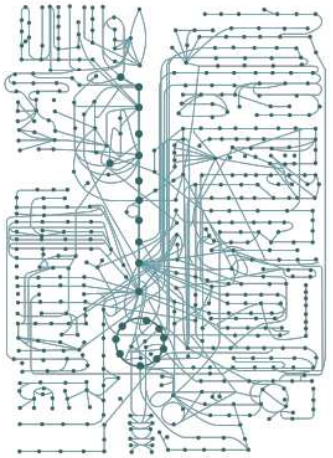
Multi-omics science as a big data science



phenome

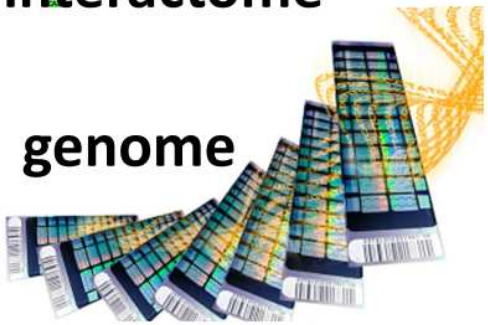


interactome



metabolome

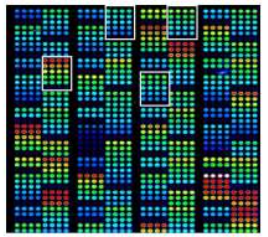
methylome



genome

publications

transcriptome



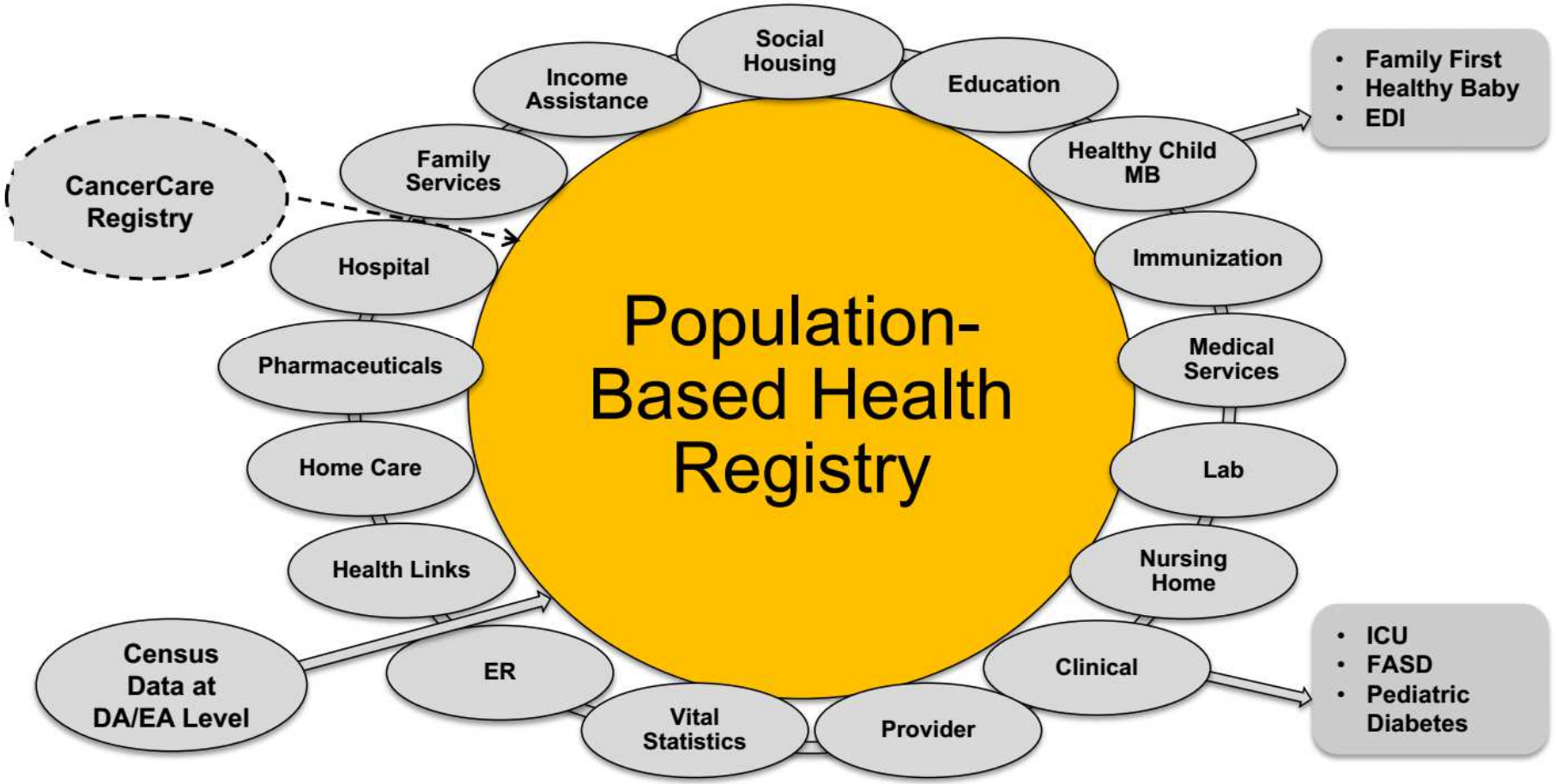
proteome

Image sources: ajc1@ flickr; Zlir'a@wikimedia

As classical epidemiology

Source	Method	Example
Individual persons	• Questionnaire • Survey	• British birth cohorts • CDC's National Health and Nutrition Examination Survey • EPIC
Biological measures	• Blood samples • Tissue samples (tumors)	• Inflammatory markers • DNA methylation
Environment	• Samples from the environment (river water, soil) • Sensors for environmental changes	• Collection of water from area streams — check for chemical pollutants • Air-quality ratings
Health care providers	• Notifications of health problems in medical records	• Report cases of cancer…
Nonhealth–related sources (financial, legal)	• Sales records • Court records • Administrative records	• Cigarette sales • Intoxicated driver arrests • Social characteristics

Manitoba Centre for Health Policy Data Repository



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

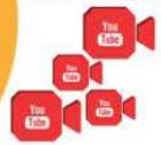
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

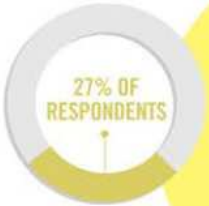
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Massive amount of large, heterogeneous, unstructured data not easy to apprehend

That need analysis pipelines

Multi-omics/big data challenges

- Modern technologies (genomic, transcriptomic, proteomic, metabolomic, epigenomic) facilitate the generation and collection of new types of data

- Main issues:
  - How to deal with this amount of data to propose useful information for health
  - Challenges in using current methods to process, analyze, and interpret the data systematically and efficiently or to find relevant signals in potential oceans of noise

Multi-omics/big data challenges

- Modern technologies (genomic, transcriptomic, proteomic, metabolomic, epigenomic) facilitate the generation and collection of new types of data
- Main issues:
  - How to deal with this amount of data to propose useful information for health
  - Challenges in using current methods to process, analyze, and interpret the data systematically and efficiently or to find relevant signals in potential oceans of noise
- Ask biostatistical problems:
  - What you are working here for
- Ask epidemiological/methodological problems
  - What I will try to expose

How epidemiological approaches can help for multi-omics science

- Epidemiology always been a discipline that uses large quantities of information with the goal of identifying risk factors that can be targeted in individuals or populations to improve health
  - Big data in epidemiological already a reality
  - Literature discussing big data issues in epidemiology
    - From « big epidemiology to « colossal epidemiology »
      When all eggs are in one basket
      Hernan MA, Savitz DA
    - Is size the next big thing in epidemiology.
      Toh S, Platt R

Epidemiology, Vol 24, N°3 2013

- Epidemiology as a tool to cope with some big data issues
  - « Epidemiologists are poised to play a central role in shaping the directions and investment in building infrastructures for the storage and robust analysis of massive and complex datasets. Given experience with multidisciplinary teams, epidemiologists are also equipped to direct the interpretation of the data in collaboration with experts in clinical and basic health sciences, biomedical informatics, computational biology, mathematics and biostatistics, and exposure sciences." Using 21st Century Science to Improve Risk-Related Evaluations. National Academies of Sciences, Engineering, and Medicine; 2017.

Classical epidemiology

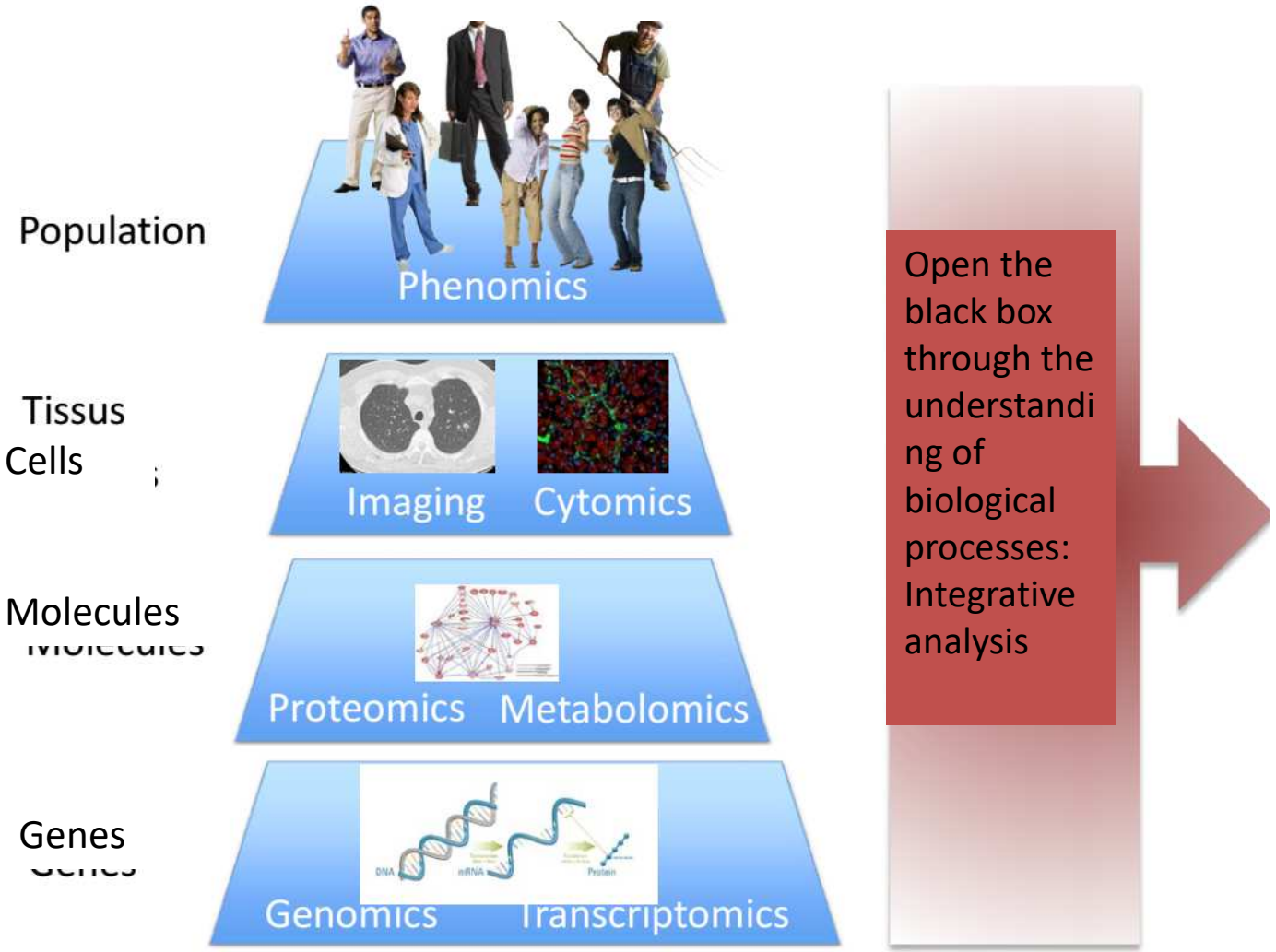
Population

Regression model

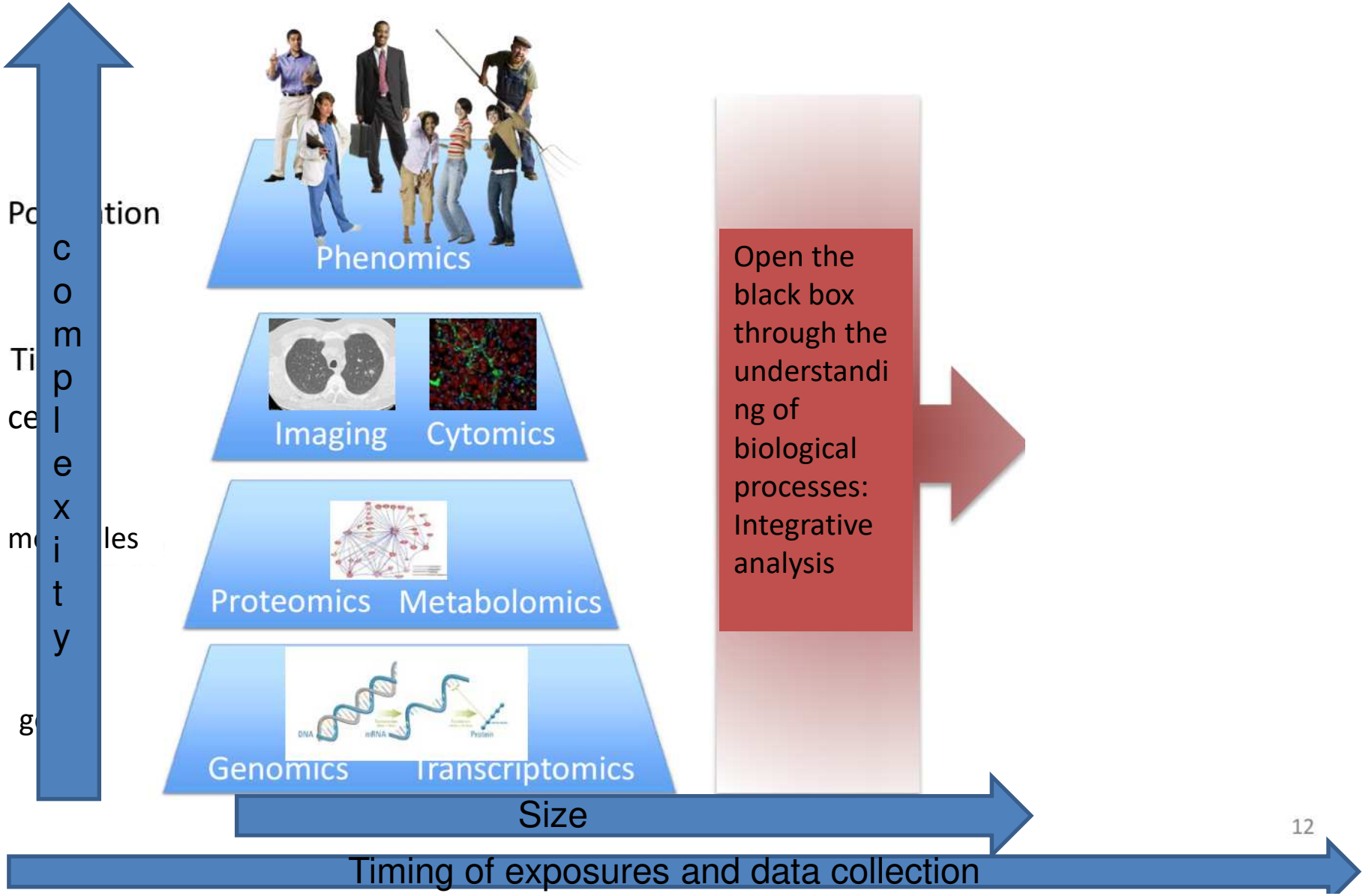
$$Logit\ P(Y=1) = X^T\beta$$

Associations of risk factors with disease: « Black box » epidemiology

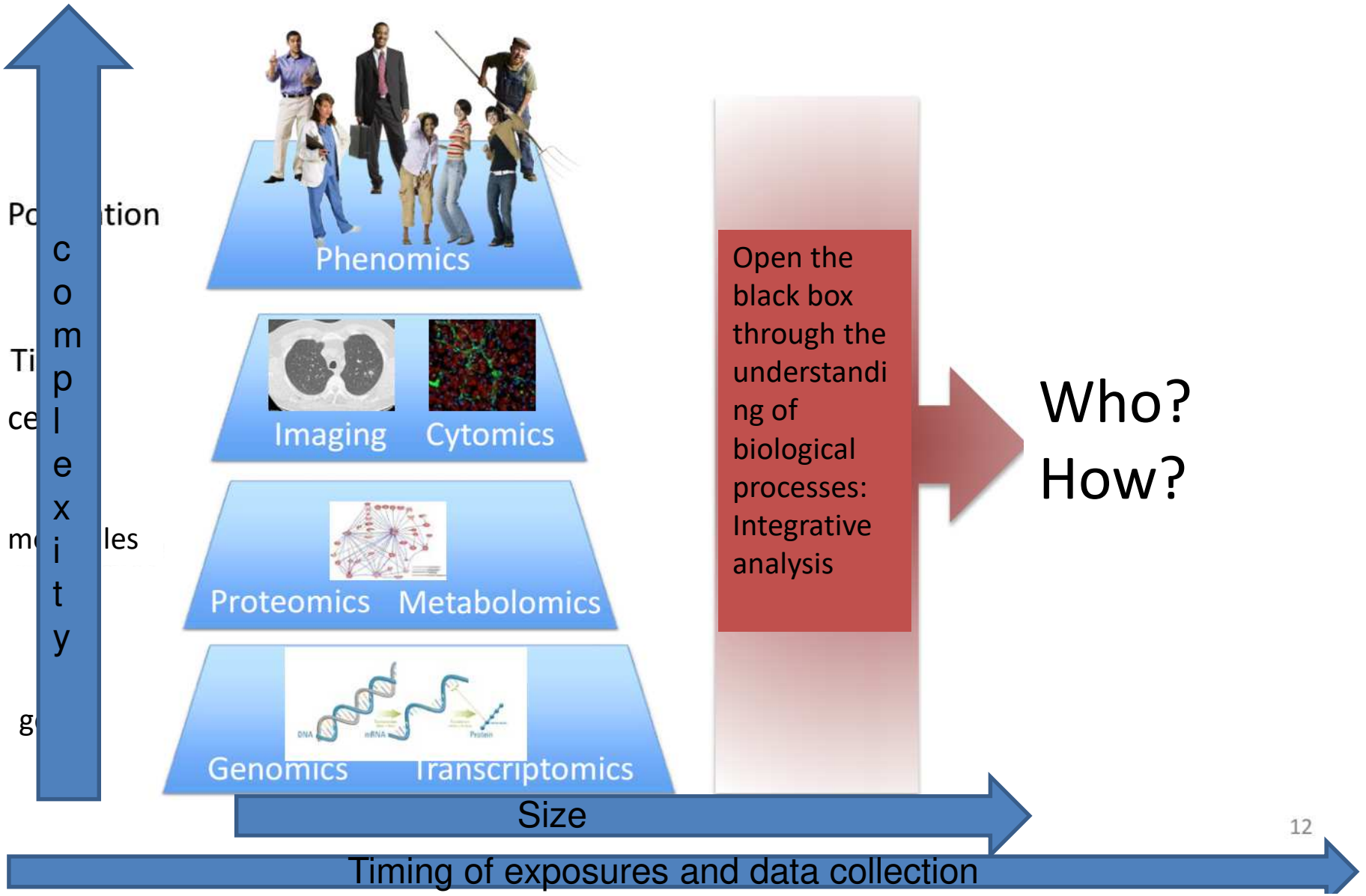
Multi-omics epidemiology



Multi omics epidemiology



Multi omics epidemiology



Common problems with multi-omics/big data analyses

- Unable to influence choice of measures/data collection/design studies
- Quality of measures
- Inadequate control groups
- Incomplete data
- Hierarchical data structure
- Correlated variables
- Everything if statistically significant
- Multiple comparisons
- Data management, linkage, security, identifiability

Translation in epidemiological biases

- Information bias
  - From error in measurements of exposures, outcomes and covariates
- Selection bias
  - From the ways in which participants are chosen to take part in epidemiological studies
- Confounding
  - From the mingled effects of exposures of interest and other exposures

⇒ External validity and the generalizability of findings

Causality

Causality

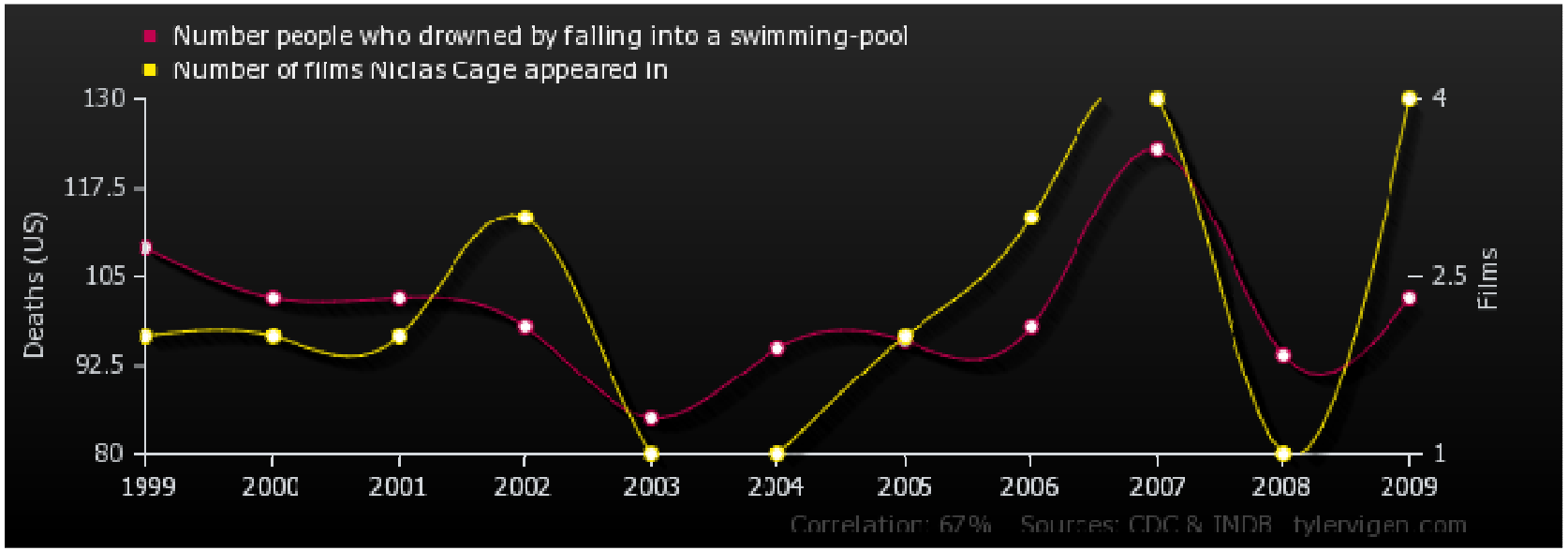
- Hypothesis driven vs. Data driven

Cartoon deriding chronic disease epidemiology, for randomly generating fears by investigating seemingly unrelated risk factors and diseases

This cartoon contains a grain of truth: observational research is at its methodological best in discovering unexpected adverse effects.

Spurious correlations

Number of People Who Drowned by Falling into a Swimming Pool Correlates with Number of Nicholas Cage Films



http://www.tylervigen.com/view_correlation?id=359

34 | NATURE | VOL 500 | 1 AUGUST 2013

From big data to mechanism

The gene variant *APOE4* is the most common genetic risk factor associated with late-onset Alzheimer's disease. A comprehensive, multilayer study reveals the molecular and cellular signatures of *APOE4* in humans. SEE ARTICLE P.45

doi:10.1038/nature12415

Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease

Herve Rhinn[1,2]*, Ryousuke Fujita[1,2]*, Liang Qiang[1,2], Rong Cheng[2], Joseph H. Lee[2,3] & Asa Abeliovich[1,2]

CORRECTIONS & AMENDMENTS

RETRACTION

doi:10.1038/nature14591

Retraction: Integrative genomics identifies APOE ε4 effectors in Alzheimer's disease

Herve Rhinn, Ryousuke Fujita, Liang Qiang, Rong Chen, Joseph H. Lee & Asa Abeliovich

Nature **500,** 45–50 (2013); doi:10.1038/nature12415

In this Article, we described integrative genomics analyses of Alzheimer's disease and associated risk factors. However, reanalysis of the data has showed that sample numbers, image panels and data points were inappropriately manipulated and inaccurate in the ELISA and subcellular localization studies presented in Figs 2d, e, 3b, g, h and 4c, as well as in corresponding Supplementary Figs 10–16. We are in the process of repeating these cell-based studies. We remain confident in the transcriptomics and human genetics analyses reported in the Article. However, given these issues, we wish to retract the Article in its entirety. We deeply regret this circumstance and apologize to the community.

Reporting associations without thinking about causality

- This stuff makes me crazy in epidemiology
- It undermines the credibility of research
- Can we do better?
  - Yes but
    - We have to think much more carefully about causation and bias
  - How
    - Use a matching technique, like propensity scores, to make treatments and controls more comparable
    - Use a quasi-experimental design (instrumental variables) to exploit natural randomness
    - Define conceptual model using direct acyclic graphs and management of confusion/mediation

Meet in the middle approach*

- Involves generally
  - A prospective search for intermediate biomarkers linked to the underlying disease
  - A retrospective search that links the intermediate biomarkers to past exposures of the environmental agent of concern
- Can be considered as three steps
  - An investigation into the association between exposure and disease
  - An assessment of the relationship between exposure and biomarkers of exposure and early effects
  - An assessment of the relationship between the disease outcome and intermediate biomarkers
- Inference of a causal relationship strengthened if associations are documented for each of the three key relationships
  - Not always the case: quid in your projects (link with health in particular) ???

Vineis et al. Environ Mol Mutagen 2013)

A LIFE COURSE VIEW OF HEALTHY AGEING

Theortical framework

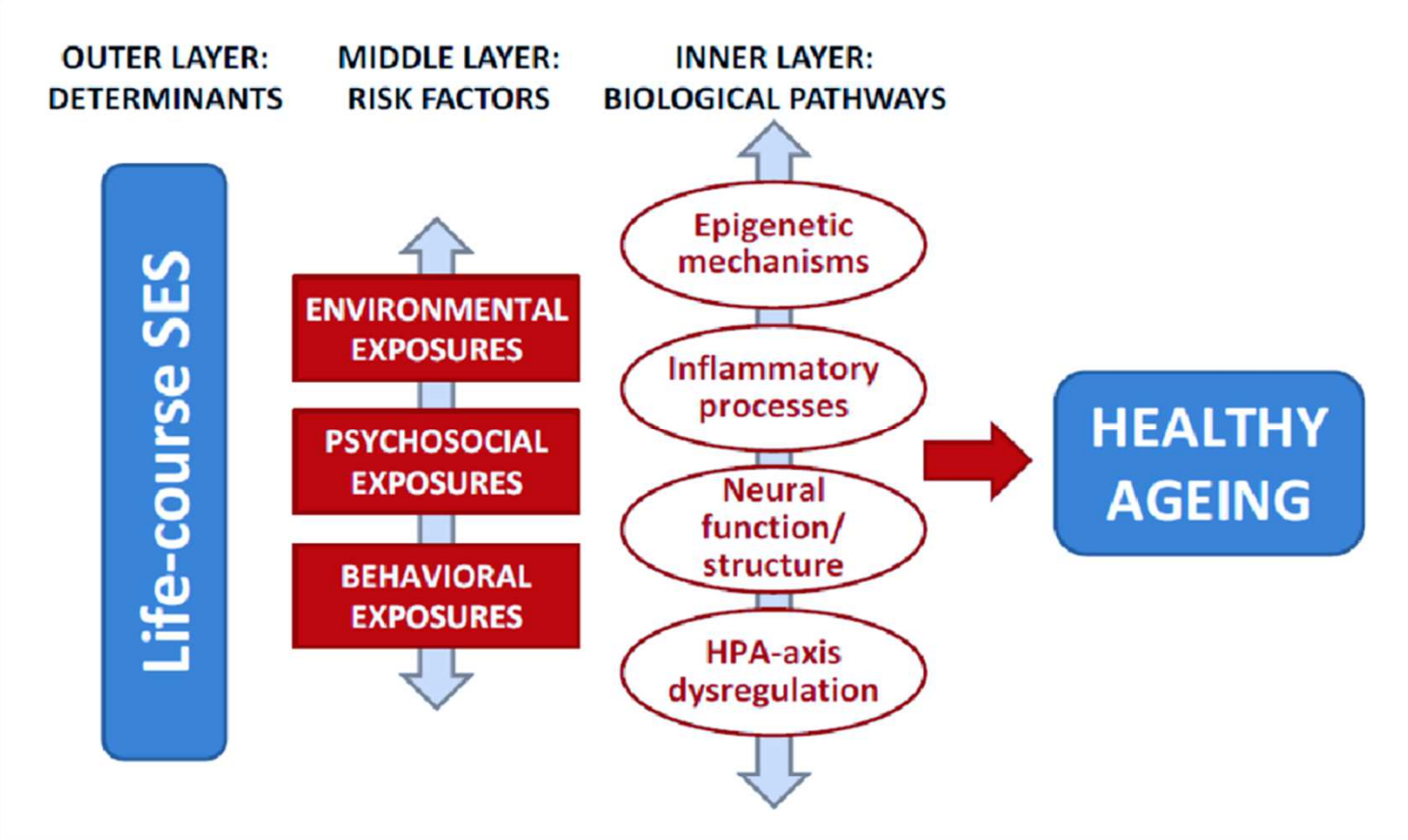


Figure: The graph represents the theoritical framework of the life-course pathways underlying social differences in healthy ageing.

→ In line with the exposome concept, there is a need to incorporate a temporal dimension in the model

The Bradford Hill criteria

- Nine viewpoints to help determine if observed associations are causal
  - **Strength of the association**: A small association does not mean that there is not a causal effect, though the larger the association, the more likely that it is causal
  - **Consistency** (**reproducibility**): Consistent findings observed by different persons in different places with different samples strengthens the likelihood of an effect
  - **Specificity**: The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship
  - **Temporality**: The effect has to occur after the cause. Perhaps the only criterion which epidemiologists universally agree is essential to causal inference or even "inarguable" (Rothman et al. Am J Public Health 2005)
  - **Biological gradient**: Greater exposure should generally lead to greater incidence of the effect
  - **Plausibility**: A plausible mechanism between cause and effect is helpful
  - **Coherence**: Coherence between epidemiological and laboratory findings increases the likelihood of an effect
  - **Experiment**: Occasionally it is possible to appeal to experimental evidence
  - **Analogy**: The effect of similar factors may be considered

The Bradford Hill criteria & big data*

- **Strength of the association**:
  - Statistically significant results presented not always biologically meaningful or methodologically appropriate for contributing to causal inference
  - Need an examination of underlying methods, comparison to the weight of evidence in the literature, and consideration of other contextual factors
- **Consistency** (**reproducibility**):
  - Particularly difficult due to the lack of control on data collection & study design (sample size, confounders): "Human studies are affected by a multitude of confounding factors that are difficult or impossible to control for. The ability of omics approaches to produce meaningful insight into human disease is very much dependent on available sample sizes, and in many settings an underpowered study may not only be a shot in the dark, missing true signals, but it is also more likely to produce false positive results." (Hasin et al. Genome Biology 2017)
- **Specificity**:
  - Most exposure and health concerns around complex chemical mixtures and low-dose environmental and occupational exposures complicated by a variety of risk factors
  - The same exposure can influence various biological systems at various omics scales
- **Temporality**:
  - Many exposure involve low levels of exposure over extended time frames, and low incidence, micro-scale outcomes that occur following long latency periods
  - Multi-omics studies mainly cross sectional with only one single biological measure

*Fedak et al. Emerg Themes Epidemiol 2015

The Bradford Hill criteria & big data*

- **Biological gradient**:
  - Most dose–response curves are non-linear and can even vary in shape from one study to the next depending on characteristics of the given population, exposure routes, and molecular endpoints assessed
  - New tools and technical capabilities allow researchers to characterize a variety of low-level molecular endpoints that may not lead to disease or observable adverse outcomes on a larger scale: responses at these low levels may not be indicative of disease, but rather adaptive responses

- **Plausibility**:
  - Opening the 'black box' through integrating molecular epidemiological advancements is not always similar to a better understanding of biological plausibility for suggested causal relationships

- **Coherence**:
  - Not always coherent results between the different omics dimensions

* Fedak et al. Emerg Themes Epidemiol 2015

An example from a pilot study including epidemiological considerations

Castagné R et al. Biological marks of early life socioeconomic experience is detected in the adult inflammatory transcriptome. Sci Rep. 2016 Dec 9;6:38705.

Castagné R, et al. A life course approach to explore the biological embedding of socioeconomic position and social mobility through circulating inflammatory markers. Sci Rep. 2016 Apr 27;6:25170

Pilot dataset: EPIC-Italy & EnviroGenoMarkers study

Biological measures

EpiGenomics

- Illumina Infinium Human Methylation 450 BeadChip
- 485, 512 Methylation sites
- 1, 716 samples

Transcriptomics

- Agilent 44k
- 29, 662 probes
- 268 samples

Proteins

- Luminex Multianalyte Profiling
- 28 inflammatory-related proteins
- 268 samples

Life course socioeconomic position (SEP)

Childhood SEP

- Father's occupation
- 2 classes: 'Manual' and 'Non-manual'

Young adulthood SEP

- Participant's education
- 2 classes: 'High' (above the minimum legal education level) and 'Low'

Adulthood SEP

- Highest household occupation
- 2 classes: 'Manual' and 'Non-manual'

OMICs PROFILING: OBJECTIVES & MODELS

- **Aim:** to identify biological imprints of soci-economic experiences at different molecular levels.
- **Approach:** full resolution screening using univariate approaches.
- **Model formulation:** linear mixed models to account for nuisance variation.
  - Variable of interest: X^i (SEP, 2 classes)
  - Predictors: Y^i, Proteins or Gene expression
  - Fixed effects: FE^i, age and gender, phase and centre, case-control status
  - Random effect variables: u^{A^i} where A^i are nuisance variables (i.e. sample position on the array, ...)
  - Full model

$$Y^i \sim \alpha + \beta_1 X^i + \beta_2 FE^i + u^{A^i} + \epsilon^i$$

$\rightarrow$ estimation using likelihood ratio test comparing the model with and without the variable of interest

Dimensionality reduction: Defining a score

Hypothesis: consistent positive direction of the association between biomarkers and SEP

Definition:

1. Get the denoised data to remove nuisance variation
2. Calculate quartiles for each biomarker
3. Assign 0 for quartile 1-3
4. Assign 1 for quartile 4
5. **Global score**: Sum across biomarkers

→ Additional assumption: dose response is set
→ Need for sensitivity analyses (on the coding)

Alternative measures:

- First PC from a principal component analyses based on 'de-noised' biomarker levels
- Cumulative rank-based approach

Life-course model using single observations

- Aim: identify critical lifes stage at which SEP can be imprinted.
- Adjusting on: Age, gender, case-control status, phase and center
- Principle: leverage available measures of time-ordered SEP-indicator.
- Approach: running sets of models sequentially adjusting for SEP:
  - **Model A:** Adjustment covariates and father job
    → early-life embedding
  - **Model B-1:** Model A + education
    → signals emerging and specific to young adulthood experience
  - **Model B-2:** Model A + highest household's occupation
    → signals emerging and specific to adulthood experience
  - **Model C:** Model B-1 + highest household's occupation
    → as above and independently of young adulthood experience
  - **Model D:** Model C + BMI + Smoking status + Alcohol
    → adjusting on behavioural factors

Life course model and inflammatory proteome score

Table: Life course multiple regression analyses for father's occupational position and inflammatory proteome score **(A)** and PC1 **(B)**. Estimates are based on 230 participants with full SEP and lifestyle information.

		Model A		Model B-1		Model B-2		Model C		Fully Adjusted Model	
(B) Inflammatory score											
Variables	**Levels**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**
Father's occupational position	Manual	1.96 (0.89)	0.029	2.88 (0.97)	0.003	2.64 (0.93)	0.005	3.08 (0.98)	0.002	2.93 (1.00)	0.004
Participant's education	Low			-2.22 (0.98)	0.024	-	-	-1.54 (1.08)	0.156	-1.5 (1.10)	0.174
Household's highest occupation	Manual					-2.22 (0.97)	0.023	-1.56 (1.07)	0.149	-1.49 (1.09)	0.174
BMI										-0.07 (0.13)	0.617
Smoking status	Former									-0.62 (1.16)	0.594
	Current									-0.57 (1.16)	0.621
Alcohol										-0.02 (0.03)	0.433
(C) Principal component 1											
Variables	**Levels**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**
Father's occupational position	Manual	-0.60 (0.45)	0.182	-1.05 (0.49)	0.031	-0.84 (0.47)	0.074	-1.10 (0.49)	0.026	-1.06 (0.50)	0.034
Participant's education	Low			1.10 (0.49)	0.025	-	-	0.93 (0.54)	0.088	0.95 (0.55)	0.086
Household's highest occupation	Manual					0.79 (0.49)	0.104	0.39 (0.54)	0.466	0.40 (0.55)	0.462
BMI										-0.01 (0.07)	0.856
Smoking status	Former									0.27 (0.58)	0.639
	Current									0.48 (0.58)	0.411
Alcohol										0.01 (0.01)	0.637

- The association with early life SEP is detected by the score
- For PC1 the association is significant upon adjustment on participant's education
- Results are robust to behavioural factors
- For both scores, association with participant's education in model B-1 only

Application to gene expression data: genome wide screen

- Expected result from genome wide scan: no significant associations

  → Need to reduce dimensionality

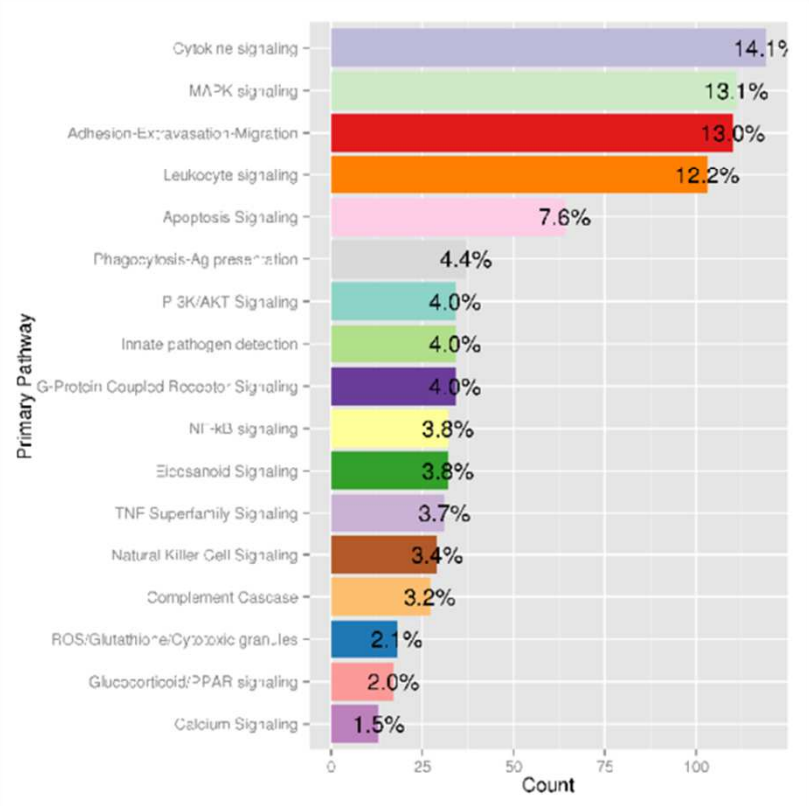
  → Towards a definition of the inflammatory transcriptome
- Genes were choosen based on the phases of the immune response
  - Hematopoiesis
  - Danger signal
  - Mobilisation of immune cells
  - Extravation
  - Migration to site of inflammation
  - Interactions between resident cells, immune cells and pathogens at site of inflammation
  - Activation of inflammatory cells
  - Effector function of inflammatory cells
  - Response of target cells
  - Resolution of immune response vs chronic inflammation

Loza *et al.*, 2007

Defining the inflammatory transcriptome (Loza et al., 2007)

Choosing genes from pathways

- Pathways were build using Ingenuity Pathways Analysis
- Genes chosen were assigned to one of the functionnal pathways:
- 1027 genes in the paper, 845 genes present in our dataset



Loza *et al.*, 2007

LIFECOURSE LINEAR REGRESSION AND INFLAMMATORY TRANSCRIPTOME SCORE

Results:

Table: Life course multiple regression analyses for father's occupational and the inflammatory transcriptome in the EPIC-Italy participants from EGM. Results are presented for the inflammatory transcriptome score (A).

		Model A		Model B		Model C		Model D		Fully Adjusted Model	
(A) Inflammatory transcriptome score											
Variables	Levels	β (SE)	P-value	β (SE)	P-value	β (SE)	P-value	β (SE)	P-value	β (SE)	P-value
Father's occupational position	Manual	21.81 (10.32)	0.036	26.25 (11.26)	0.021	25.41 (11.39)	0.027	25.37 (11.46)	0.028	23.59 (11.40)	0.040
Participant's education	Low			-11.21 (11.35)	0.324	-14.09 (12.61)	0.265	-9.81 (12.67)	0.440	-9.07 (12.57)	0.472
Household's highest occupation	Manual					6.59 (12.49)	0.599	9.14 (12.5)	0.465	10.24 (12.40)	0.410
BMI								-3.23 (1.53)	0.036	-3.30 (1.52)	0.031
Smoking status	Former							19.23 (13.44)	0.154	18.78 (13.32)	0.160
	Current							10.19 (13.33)	0.445	11.44 (13.30)	0.391
Alcohol								-0.06 (0.34)	0.868	0.06 (0.34)	0.855

- Inflammatory transcriptome global score is associated with father occupational position
- Association remains significant after adjusting behavioural factors
- No association with education after adjusting on early life SEP

Lifecourse linear regression and inflammatory transcriptome score

Sensitivity analyses results:

Table: Life course multiple regression analyses for father's occupational and the inflammatory transcriptome in the EPIC-Italy participants from EGM. Results are presented for the first PC (B) and the cumulative gene ranking-based score (C).

		Model A		Model B		Model C		Model D		Fully Adjusted Model	
(B) Principal components 1											
Variables	**Levels**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**
Father's occupational position	Manual	-4.03 (2.14)	0.061	-4.36 (2.34)	0.063	-4.12 (2.36)	0.083	-3.93 (2.39)	0.102	-2.52 (2.05)	0.219
Participant's education	Low			0.85 (2.35)	0.720	1.68 (2.62)	0.522	1.42 (2.64)	0.591	0.73 (2.26)	0.746
Household's highest occupation	Manual					-1.9 (2.59)	0.464	-1.98 (2.61)	0.449	-2.67 (2.23)	0.232
BMI								0.14 (0.32)	0.657	0.22 (0.27)	0.414
Smoking status	Former							-0.61 (2.8)	0.829	-0.44 (2.39)	0.854
	Current							4.51 (2.78)	0.106	3.10 (2.39)	0.196
Alcohol								0.04 (0.07)	0.535	-0.01 (0.06)	0.852
(C) Cumulative gene ranking-based score											
Variables	**Levels**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**	**β (SE)**	**P-value**
Father's occupational position	Manual	11.75 (5)	0.020	13.16 (5.46)	0.017	12.34 (5.51)	0.026	10.95 (5.55)	0.050	8.21 (4.96)	0.099
Participant's education	Low			-3.56 (5.5)	0.518	-6.36 (6.1)	0.298	-5.5 (6.14)	0.371	-4.21 (5.47)	0.442
Household's highest occupation	Manual					6.41 (6.04)	0.290	7.32 (6.06)	0.228	8.77 (5.40)	0.106
BMI								-0.71 (0.74)	0.344	-0.85 (0.66)	0.201
Smoking status	Former							1.86 (6.51)	0.775	1.43 (5.80)	0.805
	Current							-6.42 (6.46)	0.321	-3.89 (5.79)	0.805
Alcohol								-0.28 (0.16)	0.091	-0.15 (0.15)	0.502

- Results are weakened with the PC1 but robust with the cumulative gene ranking-based score

→ Score rely on strong assumptions → adopt pathway specific scores

Replication

Independent Dataset GSE15180 (Kobor MS *et al.*, 2009)

- **Overall design:** Samples from 30 adults with low early-life SES and 30 adults with high early-life SES
- **Summary:** This study conducted transcriptional profiling of PBMC in healthy adults who were low vs. high in early-life SES to explore the long-lasting genomic effects of early experience
- **Platform:** Illumina HumanRef-8 v3.0 expression beadchip

Table: Linear regression results for the inflammatory transcriptome and the early life SEP in participants from the GSE15180 dataset.

	β	β (se)	P-val
Global score	24.50	10.21	0.020
PC1	-2.80	2.86	0.332
Cumulative gene ranking based score	4.48	1.11	0.00002

→ The association between the inflammatory transcriptome score and early life SEP is replicated in the dataset GSE15180

Inflammatory cis-acting regulatory methylation sites (cis-eMS)

- → Multiple testing correction using a Bonferroni threshold (P$<$4.3E-06)
- → **141** Gene-CpG pairs
- → Extract strongest assocation by pair (minimum p-value)
- → **61** *cis*-eMS
- → 78.7% negatively associated with gene expression
- → 21.3% positively associated with gene expression
- → P-value ranging from $9.5x10^{-28}$ to $4.0x10^{-6}$
- → 24.6% of the identified *cis*-eMS already reported in the MESA database

---

Liu *et al.*, 2013

Inflammatory methylome score definition

→ **Global DNAm score**: average beta-values across CpG

→ **z-score**

→ **Alternative summary measure of the score:** First PC from a principal component analyses based on 'de-noised' DNA methylation levels

Linear regression model and inflammatory methylome scores

High SEP group used as reference

	Model 1			Model 1 + bmi			Model 1 + smoking			Model 1 + alcohol		
	β	SE	P-value	β	SE	P-value	β	SE	P-value	β	SE	P-value
Father job												
DNAm.global	-0.002	0.003	0.418	-0.002	0.003	0.423	-0.002	0.003	0.479	-0.002	0.003	0.510
z.score	-2.325	4.079	0.570	-2.267	4.045	0.576	-1.783	4.080	0.663	-1.464	4.179	0.727
PC1	-0.713	0.919	0.439	-0.702	0.915	0.444	-0.609	0.924	0.511	-0.568	0.947	0.550
Education												
DNAm.global	-0.008	0.003	0.007	-0.007	0.003	0.011	-0.007	0.003	0.016	-0.007	0.003	0.018
z.score	-9.722	3.978	0.016	-8.938	3.986	0.026	-8.320	4.057	0.042	-8.214	4.091	0.047
PC1	-1.626	0.906	0.075	-1.475	0.910	0.107	-1.346	0.926	0.148	-1.327	0.934	0.158
Highest household occupation												
DNAm.global	-0.008	0.003	0.007	-0.007	0.003	0.011	-0.007	0.003	0.013	-0.007	0.003	0.012
z.score	-11.419	4.026	0.005	-10.696	4.029	0.009	-10.482	4.045	0.011	-10.610	4.060	0.010
PC1	-2.051	0.917	0.027	-1.912	0.920	0.039	-1.868	0.925	0.045	-1.890	0.929	0.044

→ Similar, albeit slightly stronger, associations with highest household occupation compared to education

→ Association with education & highest household occupation remain significant after adjusting for behavioural factors

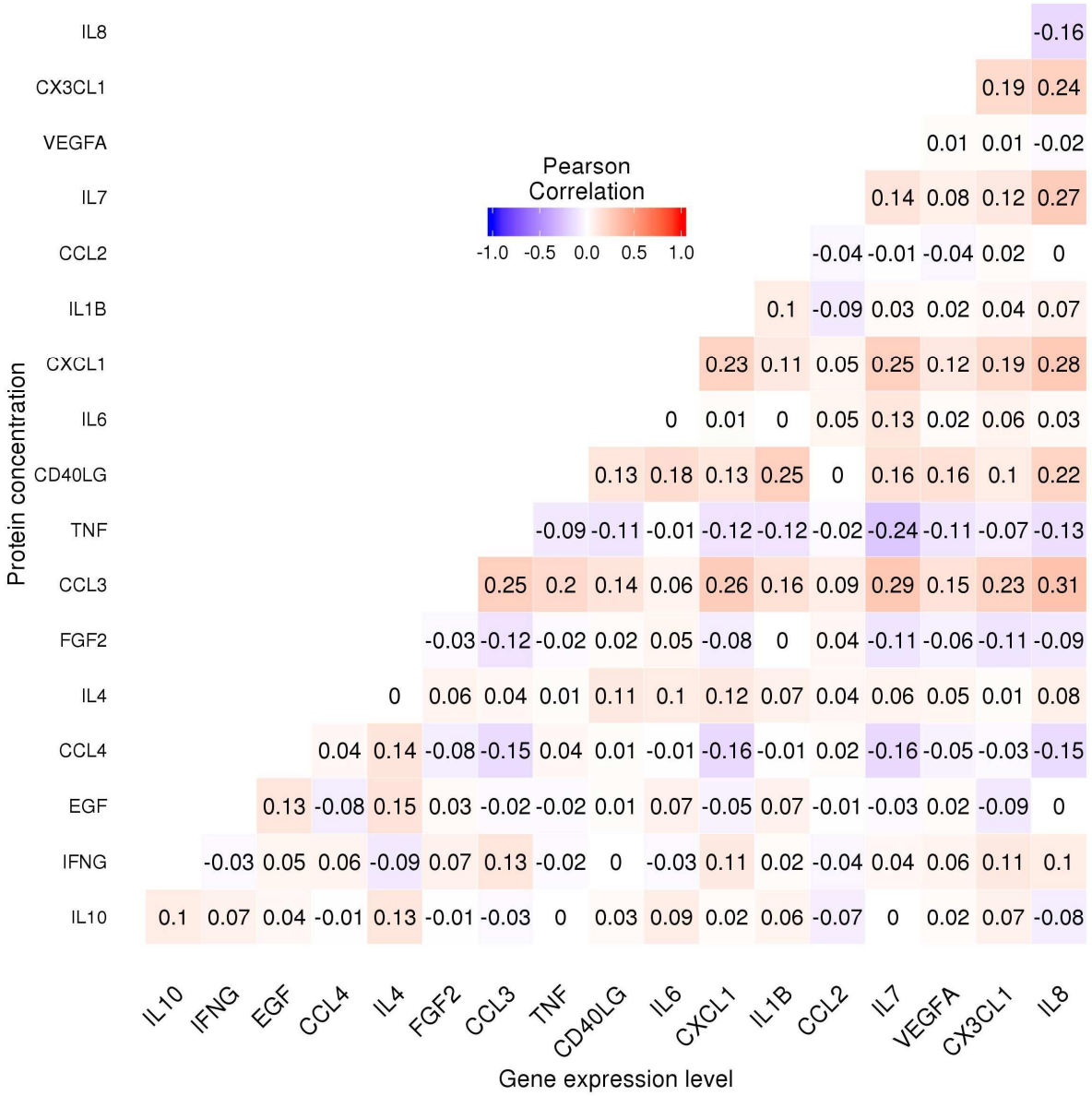
Conclusion

- Overall concordant results
  - Participants reporting a father with a 'manual' occupation had a higher proteome and transcription inflammatory score later in life
  - Participants reporting low SEP during adulthood (education and highest household occupation) showed global hypomethylation of the inflammatory methylome score
  - Associations were robust to the measure of the score, and robust to the 'functionnal' definition for methylation (cis-eMS & pMS)
  - Estimates and conclusions remained stable upon adjustments for confounders and mediators (later SEP and behavioural factors) following a a priori model
  - Replication of the inflammatory transcriptome association with early life SEP in an independent dataset (GSE15180, 60 adults)

  ⇒ Strength of associations, coherence, temporality +++

- Limitations
  - Limited sample size, no choice on data collection and study design
  - Representativeness of our study population which derives from a cancer case control study nested in a cohort with large proportion of breast cancer
    - Lack of consistency
  - Exposure = SEP
    - Lack of specificity
  - Weak correlation between omics dimensions
    - Plausibility?

Correlation between protein level and gene expression for 18 inflammatory proteins



Multi-omics science : key points

- Human data and biosamples potentially available for application of omics technologies may come from opportunistic studies based on data sources that may have been collected and stored for non research purposes
- Evidence from studies that use human tissue and medical data gained through convenience sampling from special populations may not be readily generalized
- No opportunity with these studies to address some biases via a well-thought out study design, data collection, and protocols for obtaining biospecimens
- Having a large sample size does not mitigate the potential for biases, and it increases the likelihood of statistically significant false positive findings

Data integration should represent an opportunity to expand our abilities as researchers to think about causation

Some suggestions from a classical epidemiologist

- Focus on quality of data rather than size of data
- Focus on clinical/health importance, not statistical significance
- Focus & prioritise the research questions
- Plan statistical analyses a priori
- Replicate/test results with different data
- Augment/calibrate results with higher quality « small » data
- Enchance interpretation using other methods: result triangulation (strategic use of multiple approaches to address one question)
- Develop interdisciplinarity: critical mass of expertise accross different disciplines

Thanks for your attention