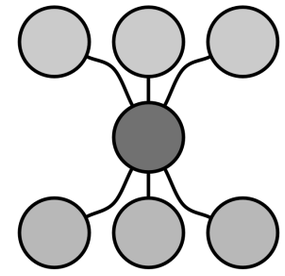


Data analytics approaches to enable ***EWAS***

Chirag J Patel and Nam Pho

Emory Exposome Workshop

06/16/16

HARVARD MEDICAL SCHOOL

DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu

@chiragjp

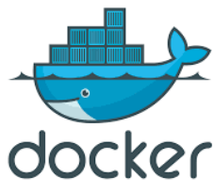
www.chiragjpgroup.org

Extensible & open-source analytics software library
(*XWAS R package*)

Freely available *exposome* data for your research
(NHANES: 40,000 individuals and 1,000 variables)

Computer "environment" to conduct *EWASs*
(Docker container in RStudio)

Materials for teaching and demonstration

XWAS **+**  **+**  **=** ***goodness!***

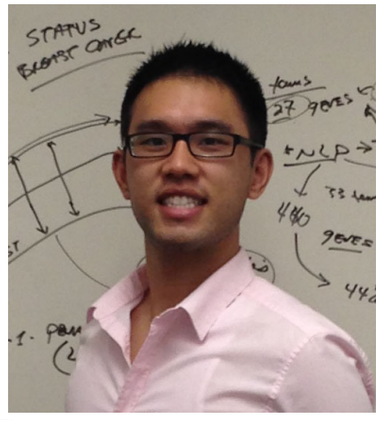
http://bit.ly/exposome-analytics-course

Please let us know if you are using the resources
(or provide feedback)!

Chirag

@chiragjp

Nam



@nampho2

Real quick:

What is the *exposome*? What is the *phenome*?

exposome

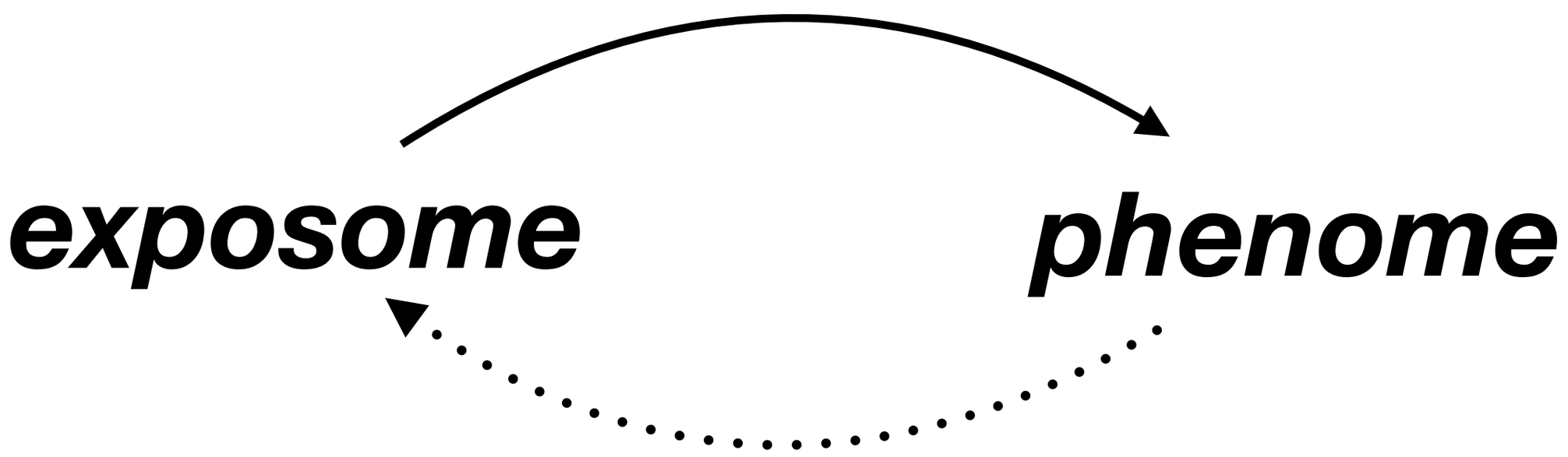
internal	external
lead (serum)	geography
nutrients (serum)	air pollution
infection (urine)	income
metabolome	

phenome

function	diseases
expression	diabetes
telomeres	cancer
metabolome	heart disease

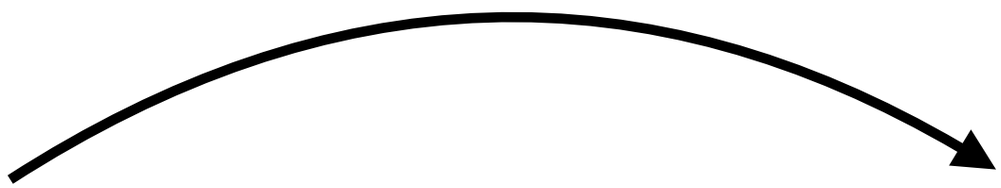
Exposome associated with the ***phenome***?

…and vice versa?



Analytic **tools** and big data **infrastructure** required to associate *exposome* with *phenome!*

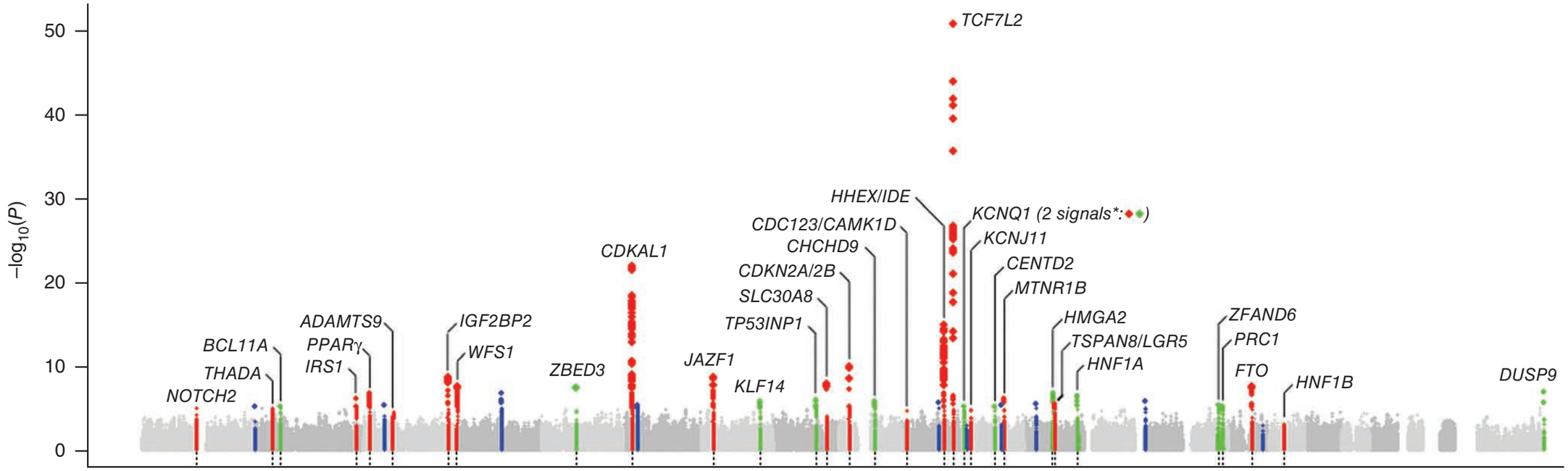
We can learn a thing or two from ***genomics*** investigation…



genome ***phenome***

e.g., ***GWAS***

Big data approaches fueled discovery of genetic variants in disease (example: genome-wide association [**GWAS**])



GWAS in Type 2 Diabetes

Voight et al, Nature Genetics 2012

N=8K T2D, 39K Controls

A search engine for robust, reproducible genotype-phenotype associations…

There are ***non-trivial*** data analytic challenges in searching for exposome-phenome associations!

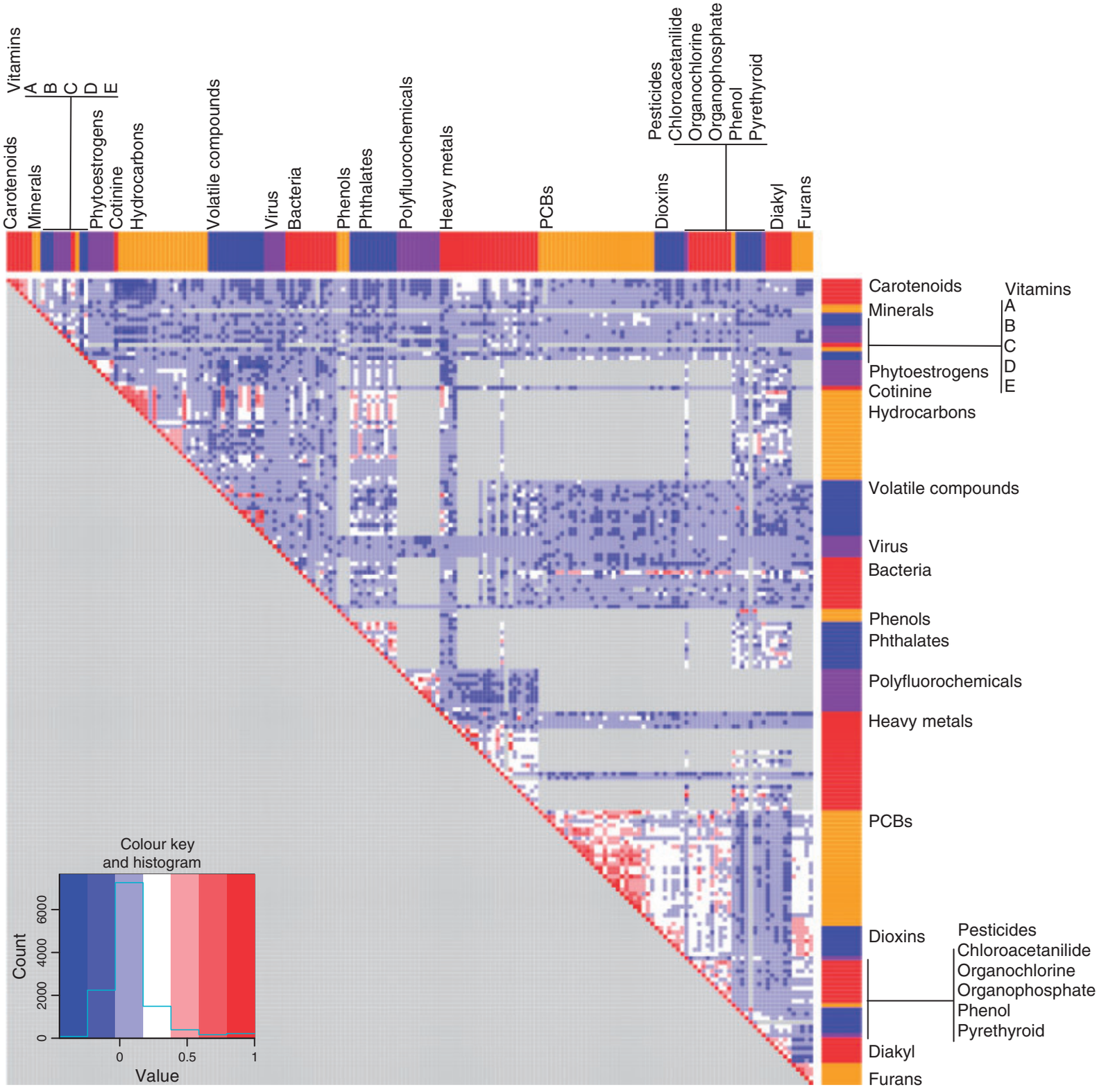
JAMA 2014
Pac Symp Biocomp 2015

Dense correlational web!

what causes what?

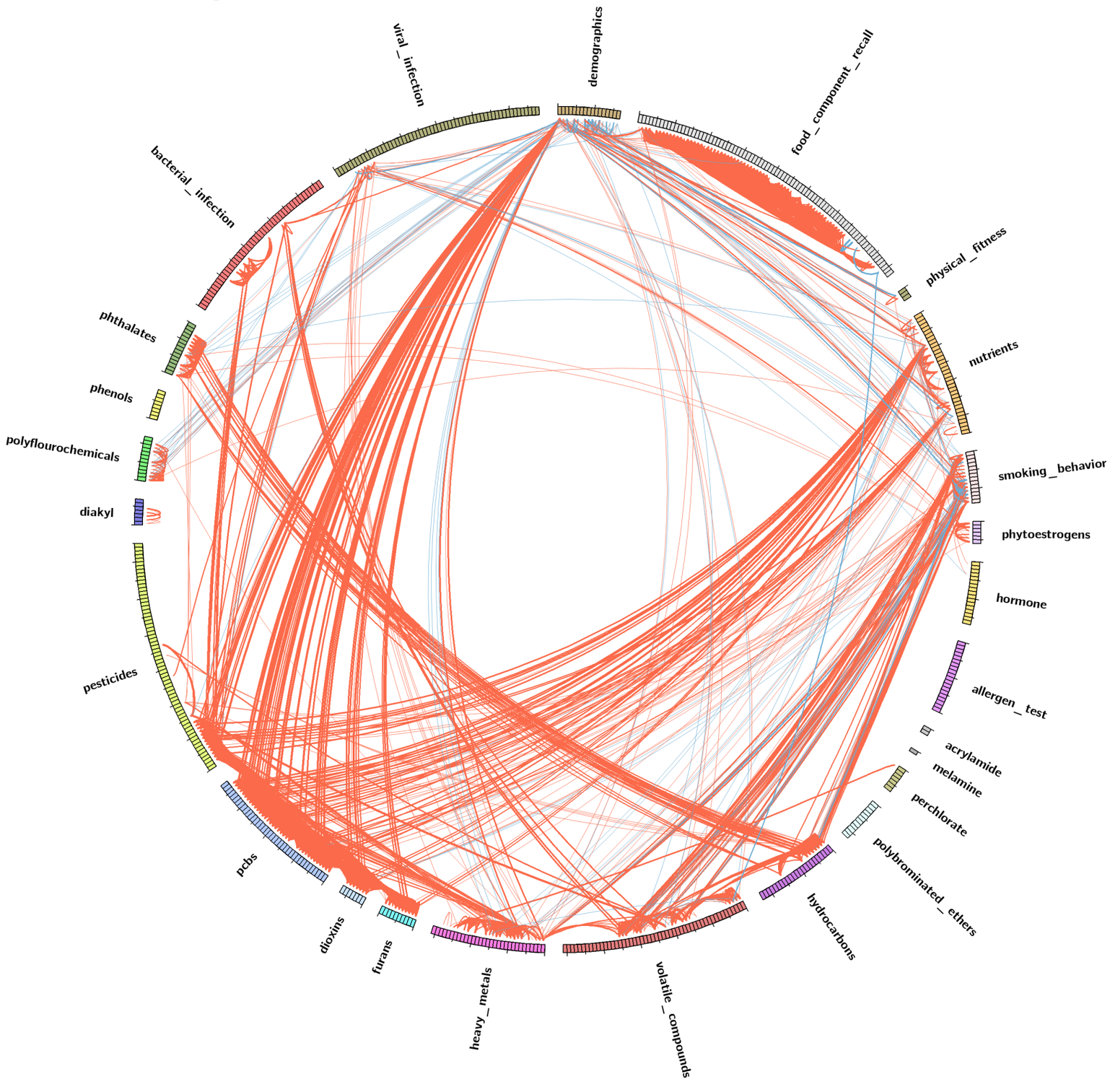
confounding bias?

JAMA 2014
Pac Symp Biocomp 2015


Vitamins
A
B
C
D
E
Carotenoids
Minerals
Phytoestrogens
Cotinine
Hydrocarbons
Volatile compounds
Virus
Bacteria
Phenols
Phthalates
Polyfluorochemicals
Heavy metals
PCBs
Dioxins
Pesticides
Chloroacetanilide
Organochlorine
Organophosphate
Phenol
Pyrethyroid
Diakyl
Furans
Colour key
and histogram
Count
0
2000
4000
6000
0
0.5
1
Value

IJE 2013

Interdependencies of the ***exposome***: Correlation globes paint a complex view of exposure



http://bit.ly/globebrowse

Pac Symp Biocomput. 2015
JECH. 2015

Multiplicity: how to determine signal from noise?

type 1 error (spurious findings)

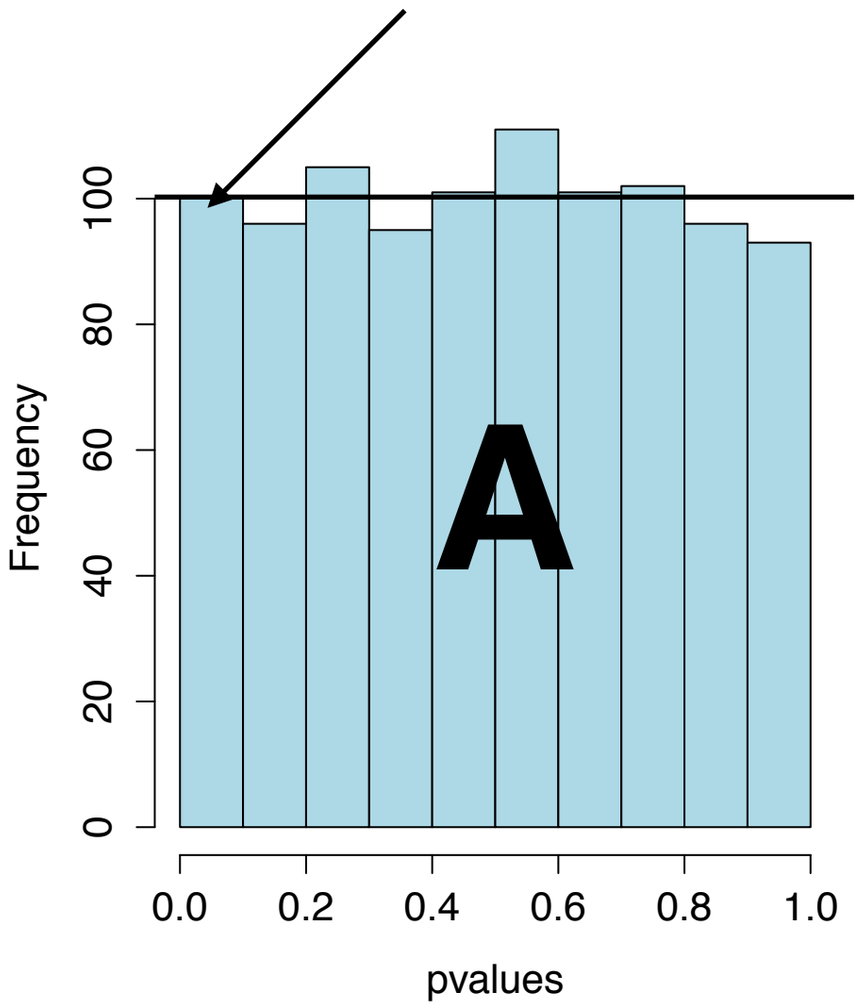
JAMA 2014

Suppose you are testing 1000 exposures in case-control study (disease vs. healthy)…

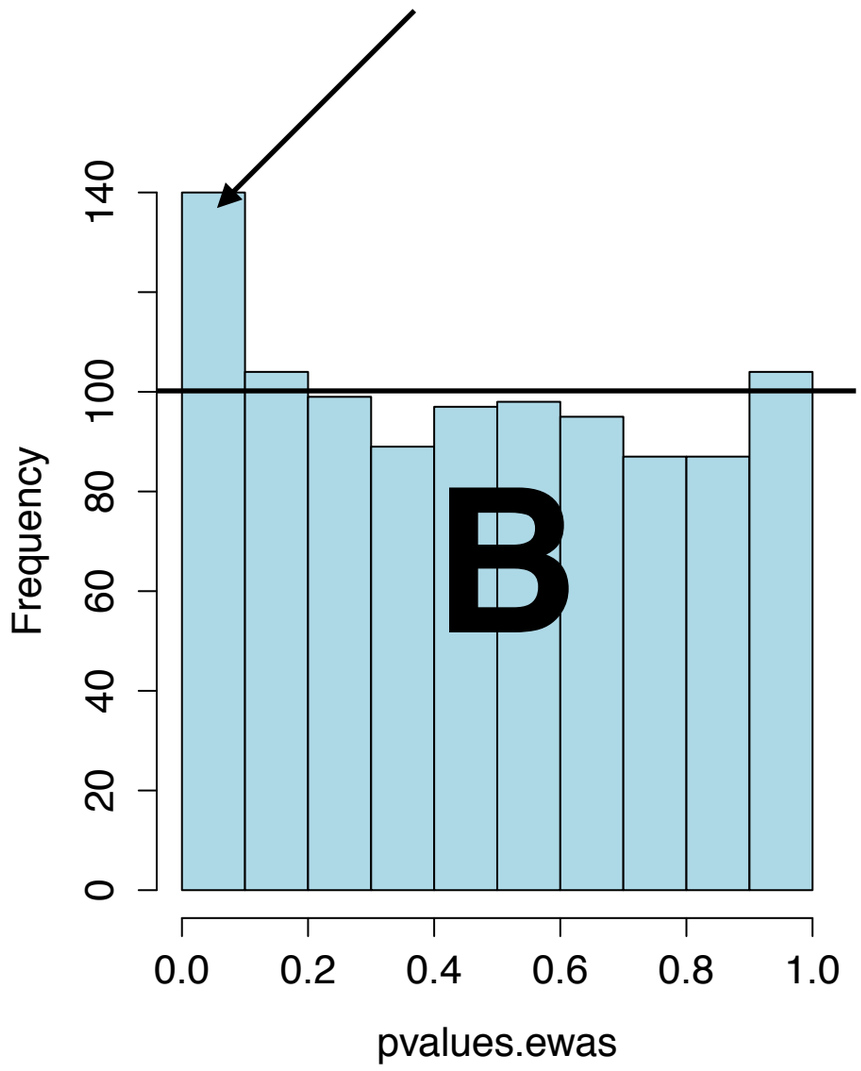
… and there were no difference between the cases and controls…

…how many findings would be “significant” at a p-value threshold of 0.05 (due to chance)?

Regime of multiple tests and "*signal to noise*":
Histogram of p-values in 2 scenarios: no difference and 5% different



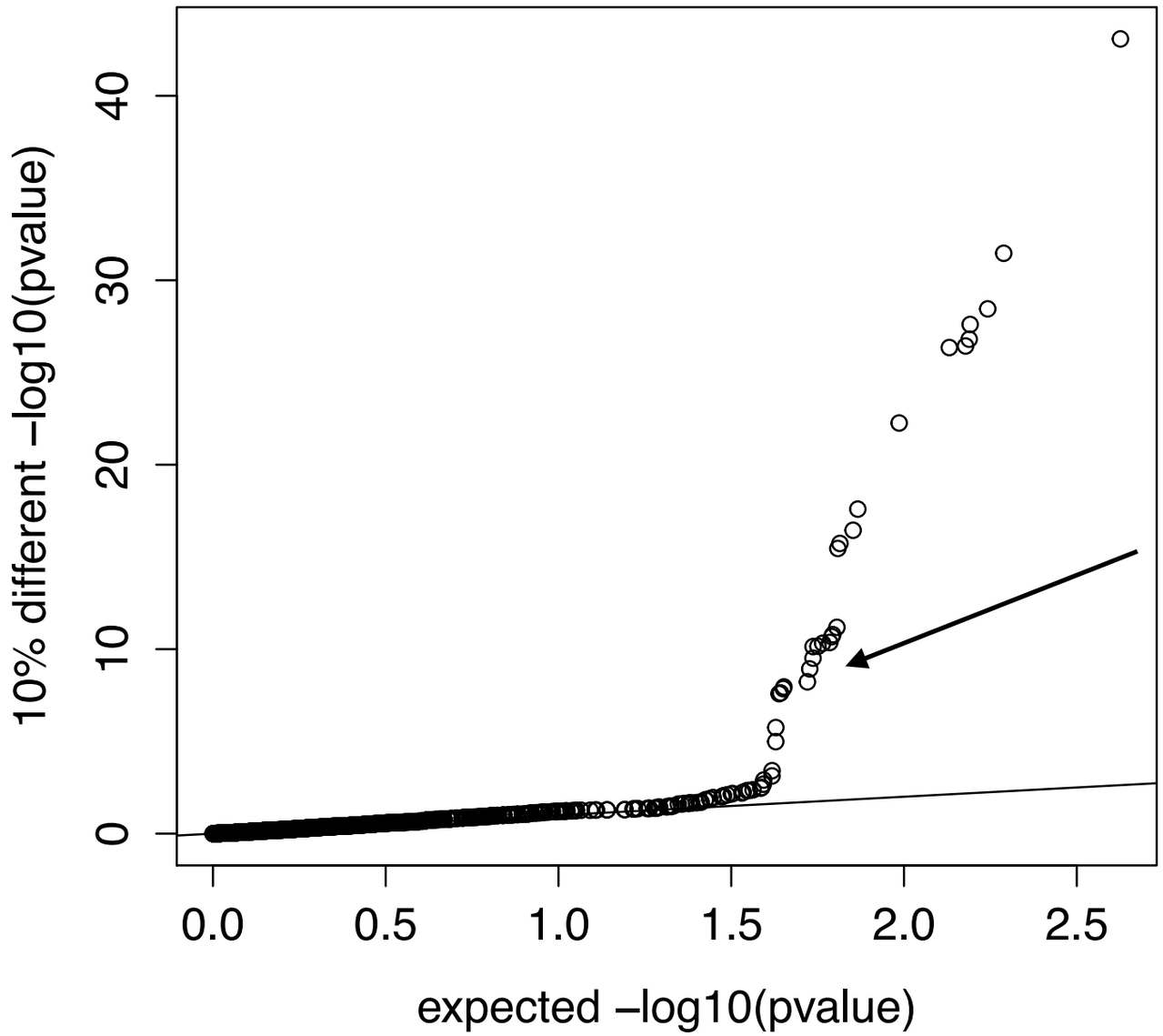
No difference
(no true associations)



5% exposures different
(5% true associations)

Estimating the deviation from null:
QQplot: -log10(pvalues) in the null and *EWAS* distributions

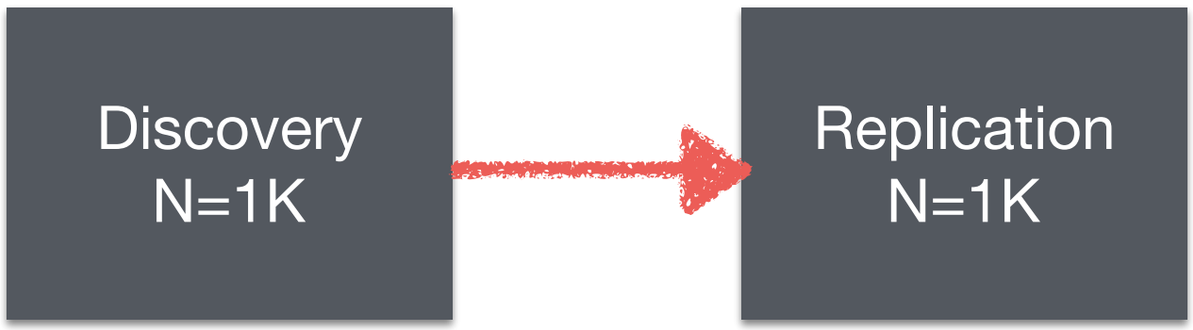
B



A

Bonferroni
False discovery rate

The tension between type 1 and type 2 errors: ***Power*** and ***replication*** for robust associations!

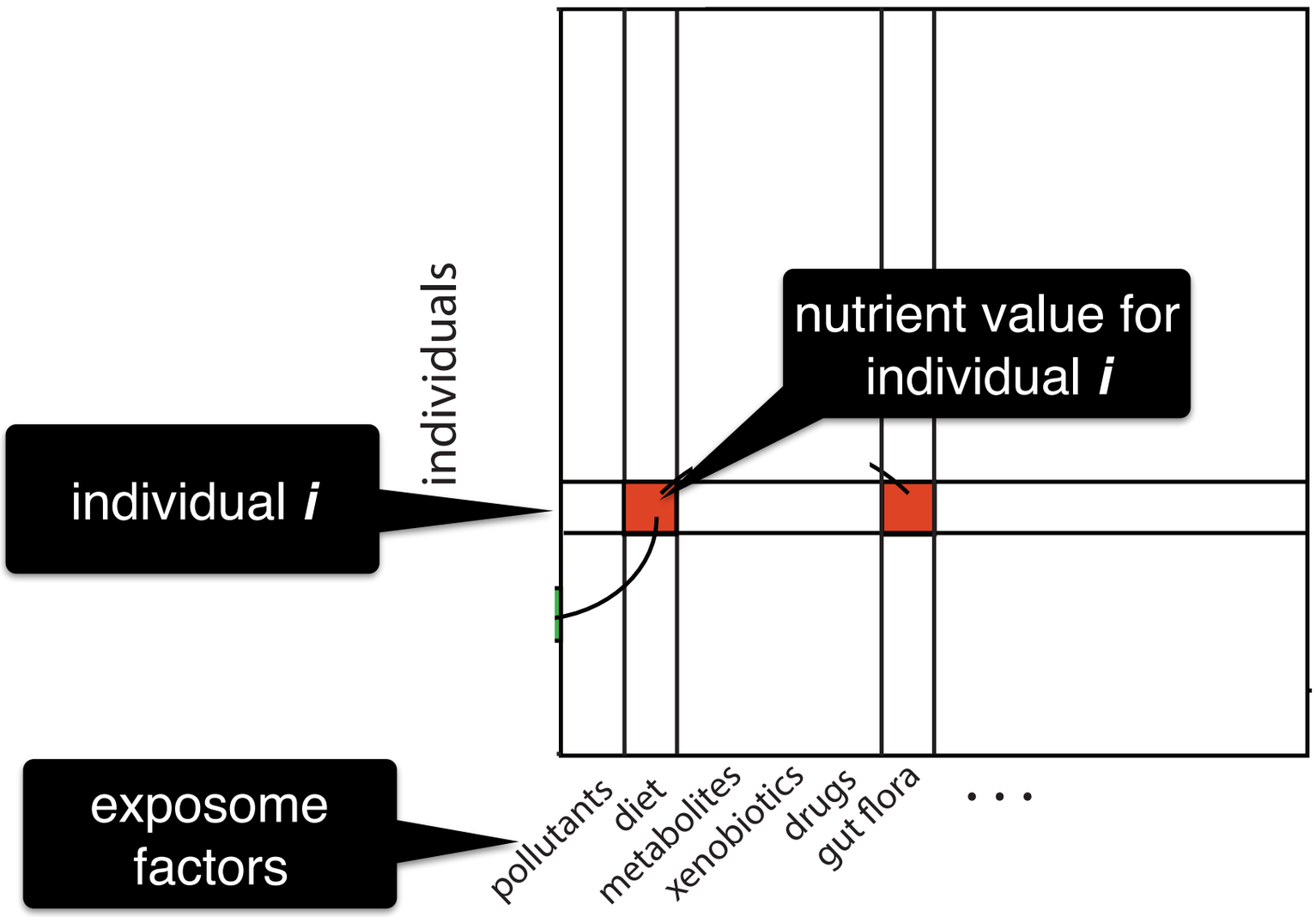


Discovery sample sizes must be large to overcome ***multiple testing*** and mitigate ***winner's curse***

Replication sample size must be large to detect association

What will the ***exposome*** data structure look like?:

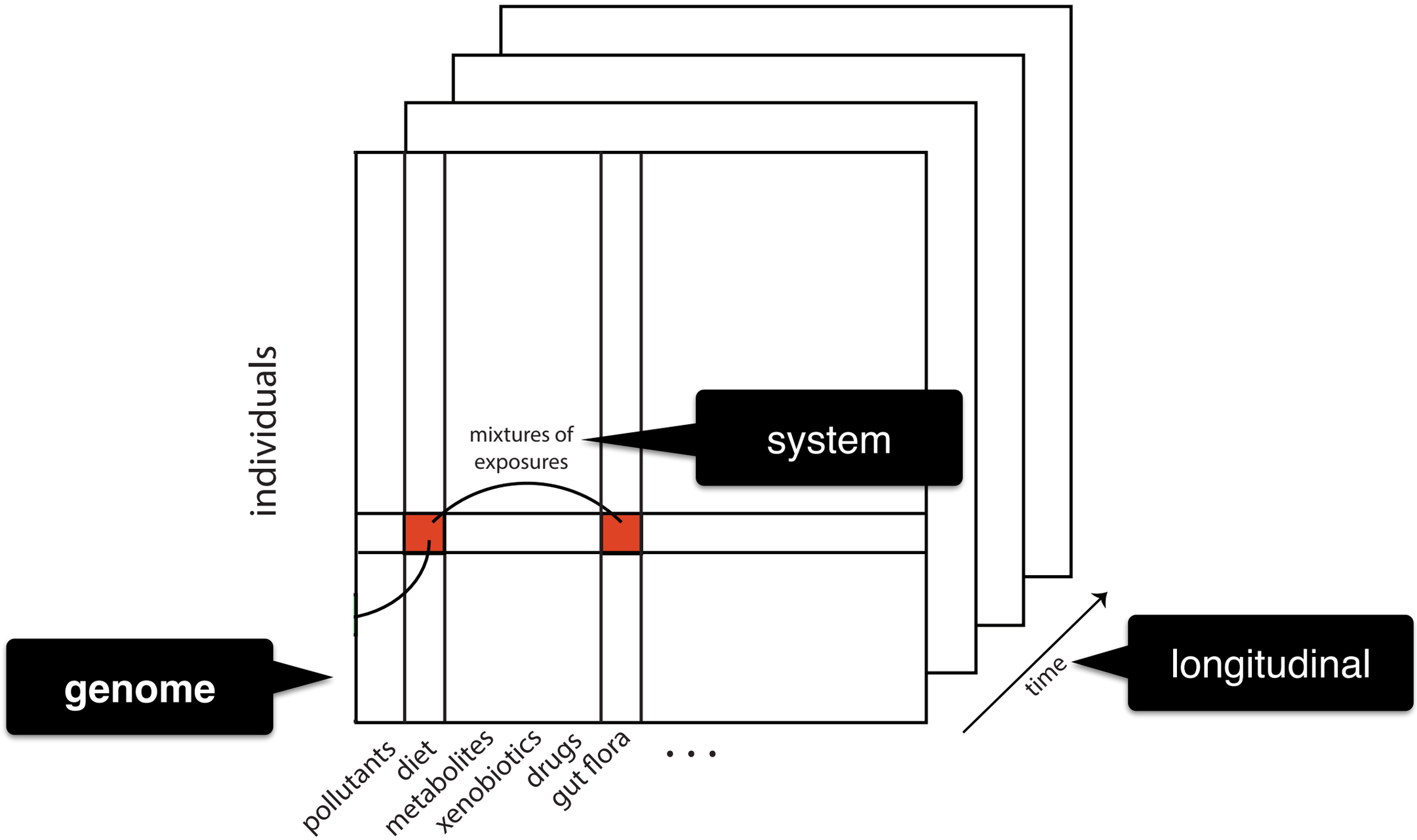
a ***high-dimensioned 3D*** matrix of
(1) ***exposure*** measurements
on (2) ***individuals*** as a function of (3) ***time***



in review

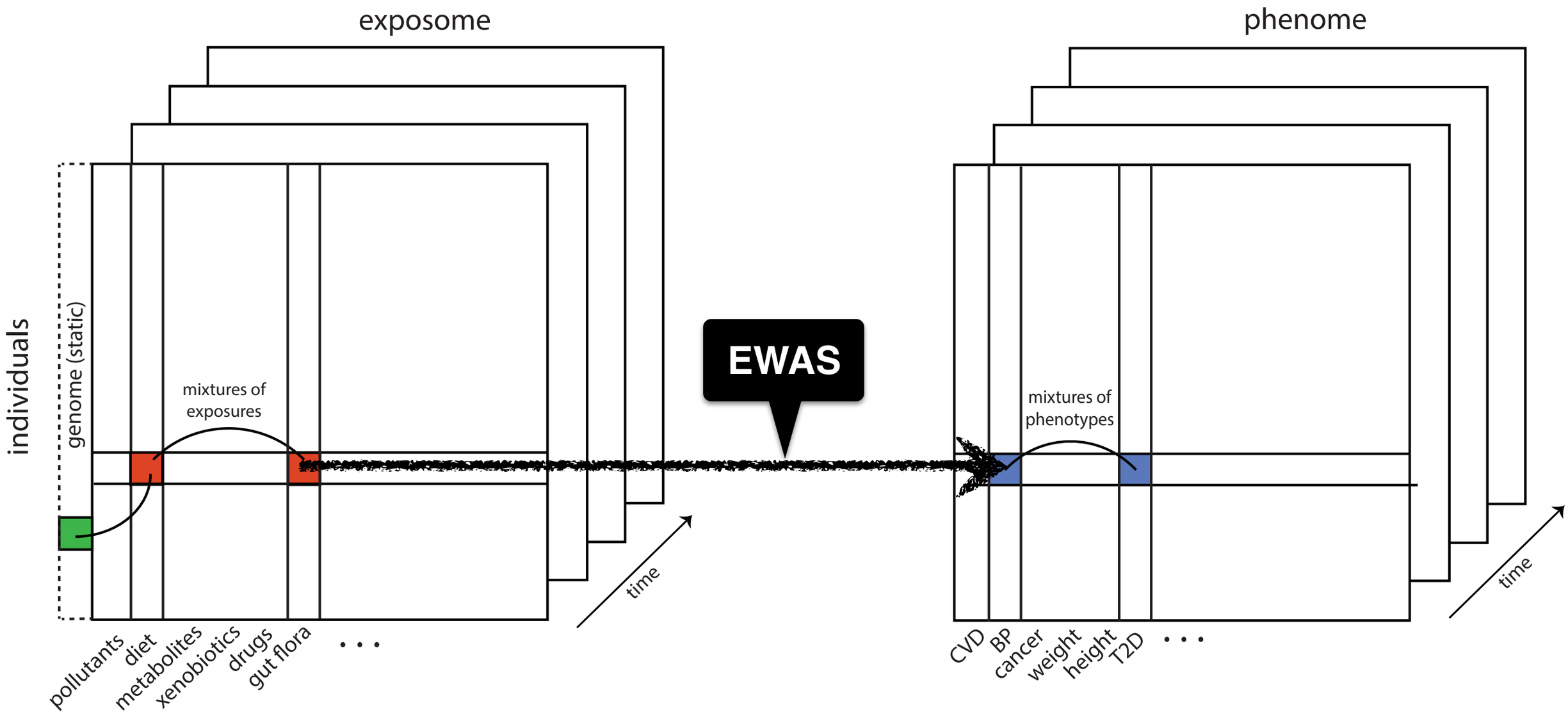
What will the ***exposome*** data structure look like?:

a ***high-dimensioned 3D*** matrix of
(1) ***exposure*** measurements
on (2) ***individuals*** as a function of (3) ***time***



in review

A schematic of a data-driven search for ***exposome-phenome*** associations…



in review

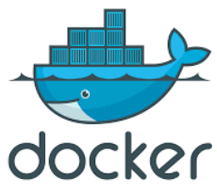
Time for you to give it a try!

Extensible & open-source analytics software library
(*XWAS R Package*)

Freely available *exposome* data for your research
(NHANES: 40,000 individuals and 1,000 variables)

Computer "environment" to conduct *EWASs*
(Docker container in RStudio)

Materials for teaching and demonstration

XWAS **+**  **+**  **=** ***goodness!***

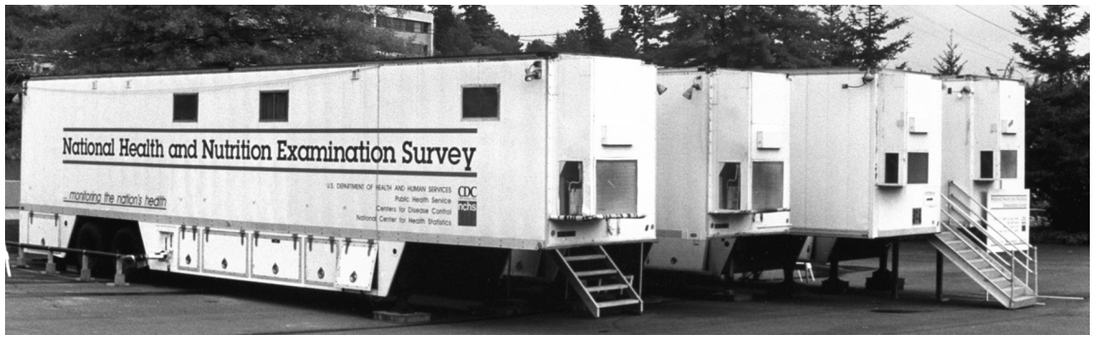
Fully merged dataset:
National Health and Nutrition Examination Survey

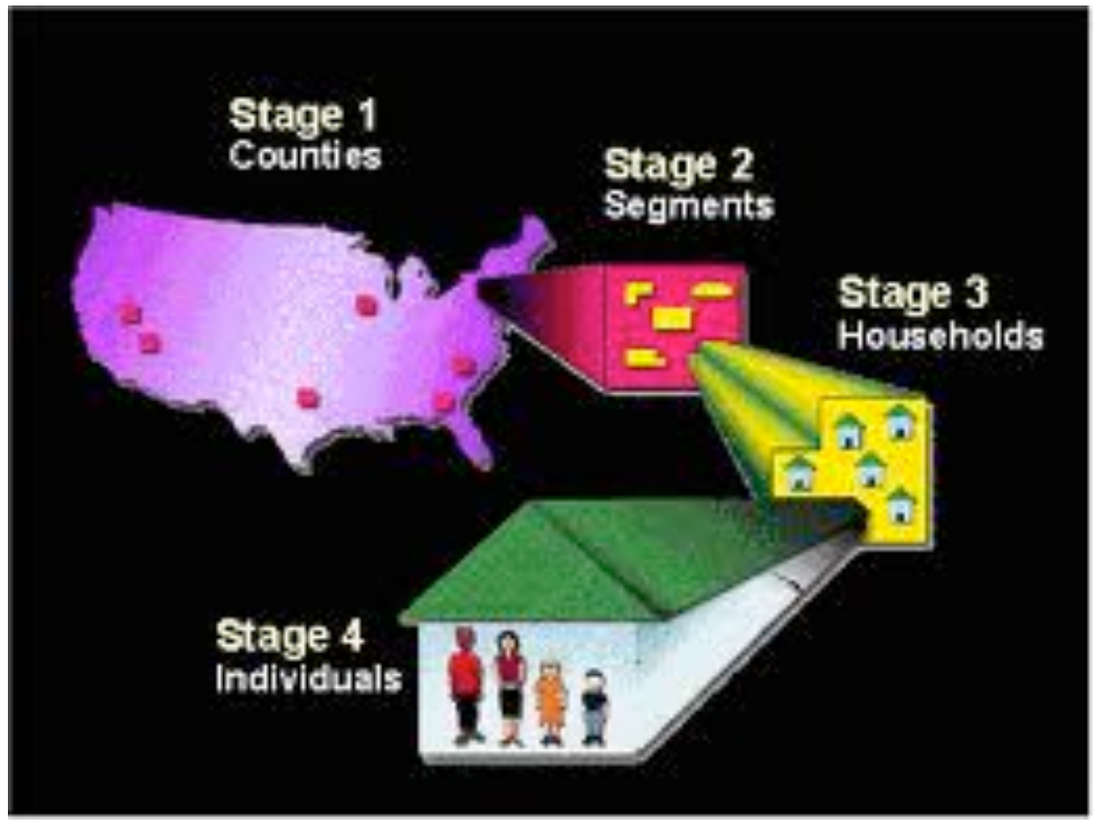
since the 1960s
now biannual: 1999 onwards
10,000 participants per survey

>250 exposures (serum + urine)

>85 quantitative clinical traits (e.g., serum glucose, lipids, body mass index)

Death index linkage (cause of death)





Ready to analyze! N=41K with >1000 variables
(let us know; we can give you a DOI)

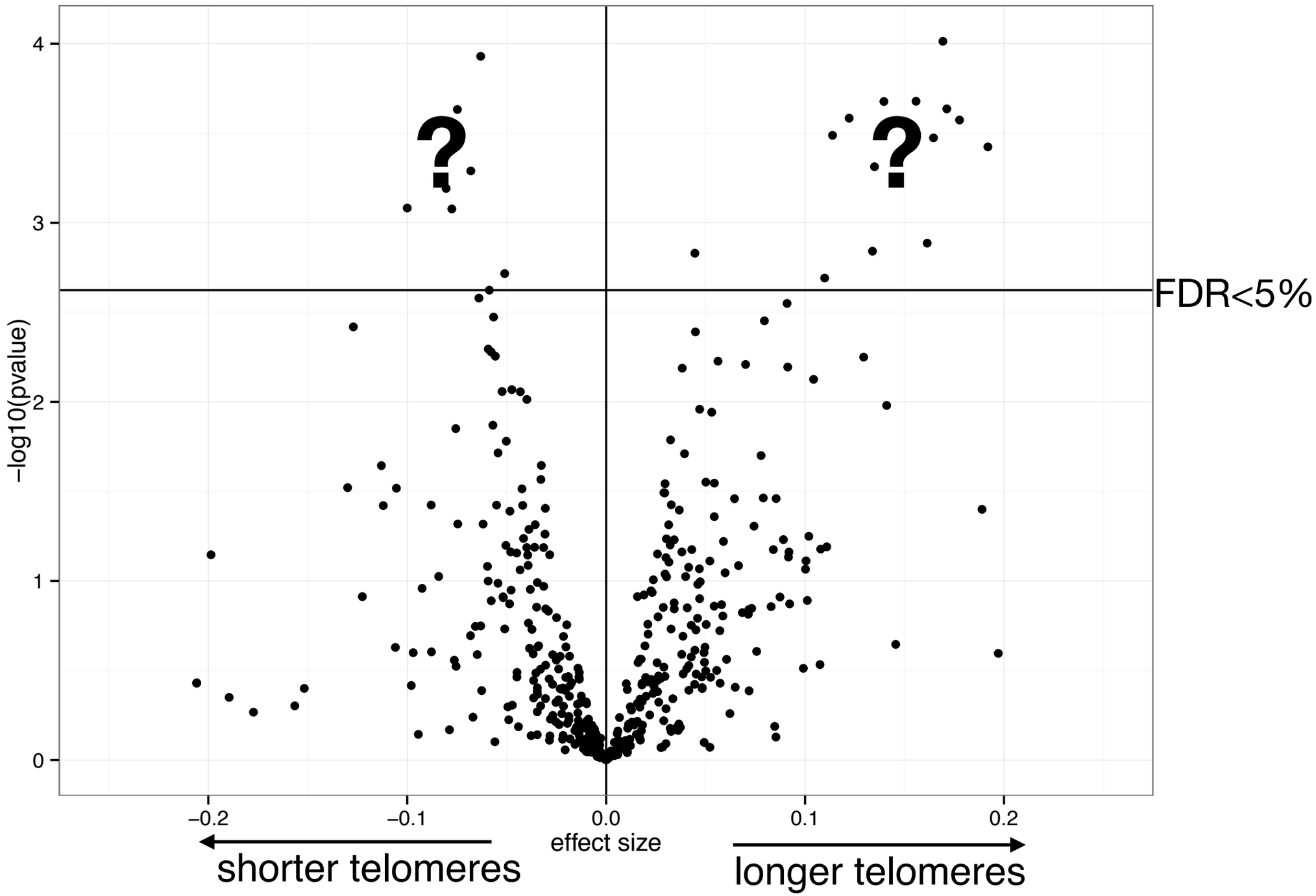
in review

13 ***EWAS***-related manuscripts

preterm birth
type 2 diabetes
type 2 diabetes genetics
lipids
blood pressure
income
mortality
telomere length
methodology (5)

http://bit.ly/ewas_nhanes

Associations in ***Telomere Length***:
Can you identify the associations in this graph?



median N=3000; N range: 300-7000

IJE, 2016

Associations in ***Telomere Length***:
Can you identify the associations in this graph?

Nam will show you how!

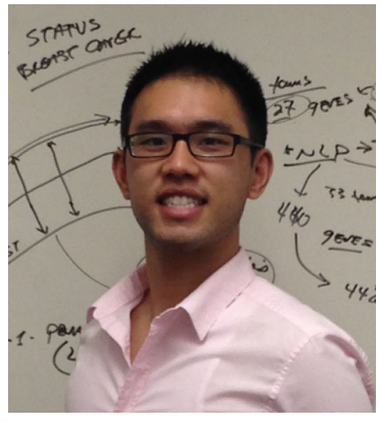
http://bit.ly/exposome-analytics-course

Please let us know if you are using the resources
(or provide feedback)!

Chirag

@chiragjp

Nam



@nampho2

Acknowledgements

RagGroup
Nam Pho
Chirag Lakhani
Adam Brown
Danielle Rasooly
Arjun Manrai
Grace Mahoney
Matthew Roy

Harvard DBMI
Isaac Kohane
Susanne Churchill
Stan Shaw
Jenn Grandfield
Michal Preminger

Stanford
John PA Ioannidis

Gary Miller
Kristine Dennis

NIEHS National Institute of Environmental Health Sciences

NIH Common Fund
Big Data to Knowledge

PhRMA RESEARCH • PROGRESS • HOPE

Agilent Technologies

HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

Chirag J Patel
chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org