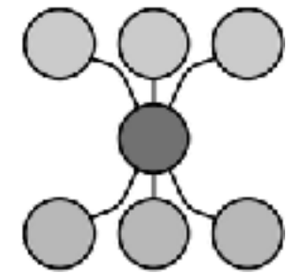


Data analytics to enable ***X-wide Association Studies (XWASs)***

Chirag J Patel
(with Nam Pho, Jake Chung, and Arjun Manrai)

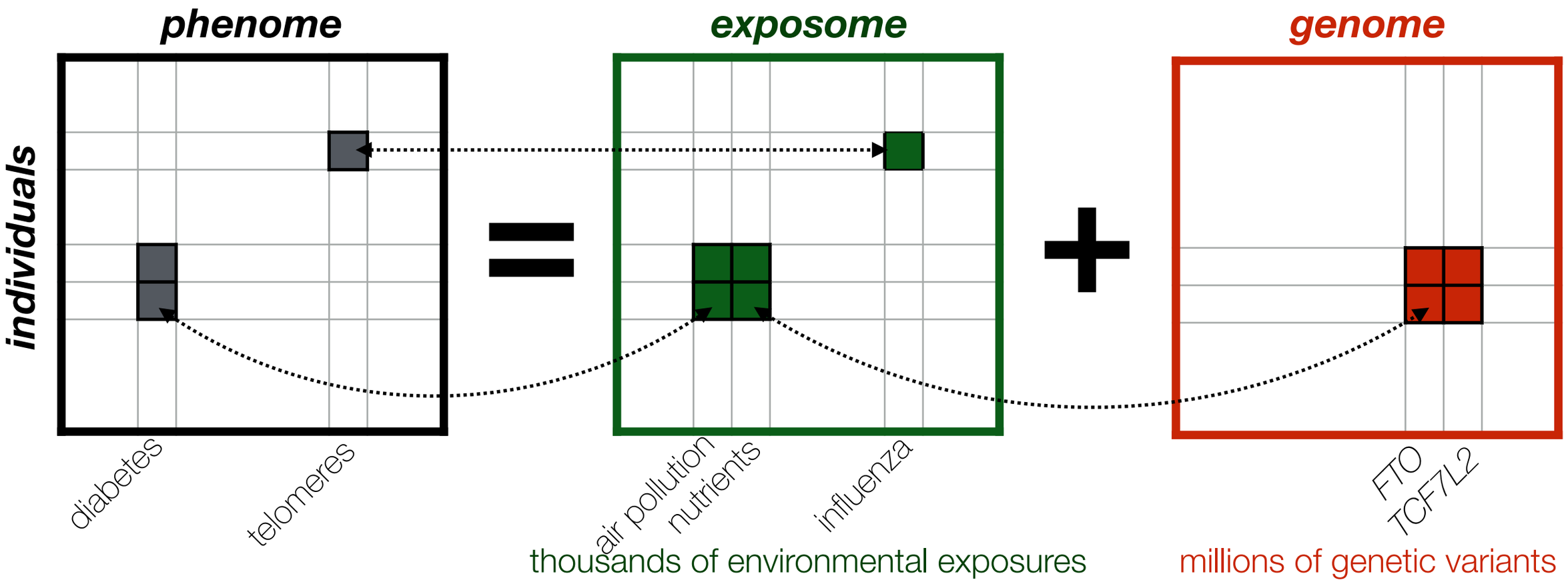
ISEE pre-conference tutorial, part 1
Ottawa, Canada
8/26/18

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu
@chiragjp
www.chiragjpgroup.org

Patel Group in Translational and Environmental Bioinformatics

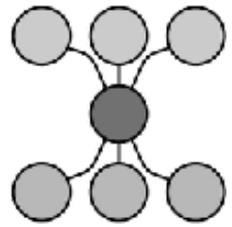
Studying the elusive environment in large scale with the ***phenome***, ***exposome***, and ***genome*** for translational discovery



ARPH, 2017
JAMA, 2014
PLoS ONE, 2010
IJE, 2012, 2013
Reprod Tox, 2014
Hum Genet 2013
JECH, 2014
AJE, 2015
Proc Symp Biocomp, 2015

http://www.chiragjpgroup.org

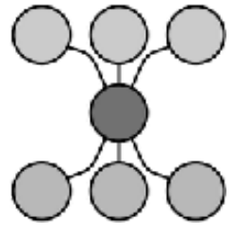
Patel Group in Translational and Environmental Bioinformatics



Studying the elusive environment in large scale with the ***phenome***, ***exposome***, and ***genome*** for translational discovery

ARPH, 2017
JAMA, 2014
PLoS ONE, 2010
IJE, 2012, 2013
Reprod Tox, 2014
Hum Genet 2013
JECH, 2014
AJE, 2015
Proc Symp Biocomp, 2015

http://www.chiragjpgroup.org



Studying the elusive environment in large scale with the ***phenome***, ***exposome***, and ***genome*** for translational discovery

Large sample sizes ***and*** number of variables!

ARPH, 2017
JAMA, 2014
PLoS ONE, 2010
IJE, 2012, 2013
Reprod Tox, 2014
Hum Genet 2013
JECH, 2014
AJE, 2015
Proc Symp Biocomp, 2015

http://www.chiragjpgroup.org

Real quick:

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

exposome

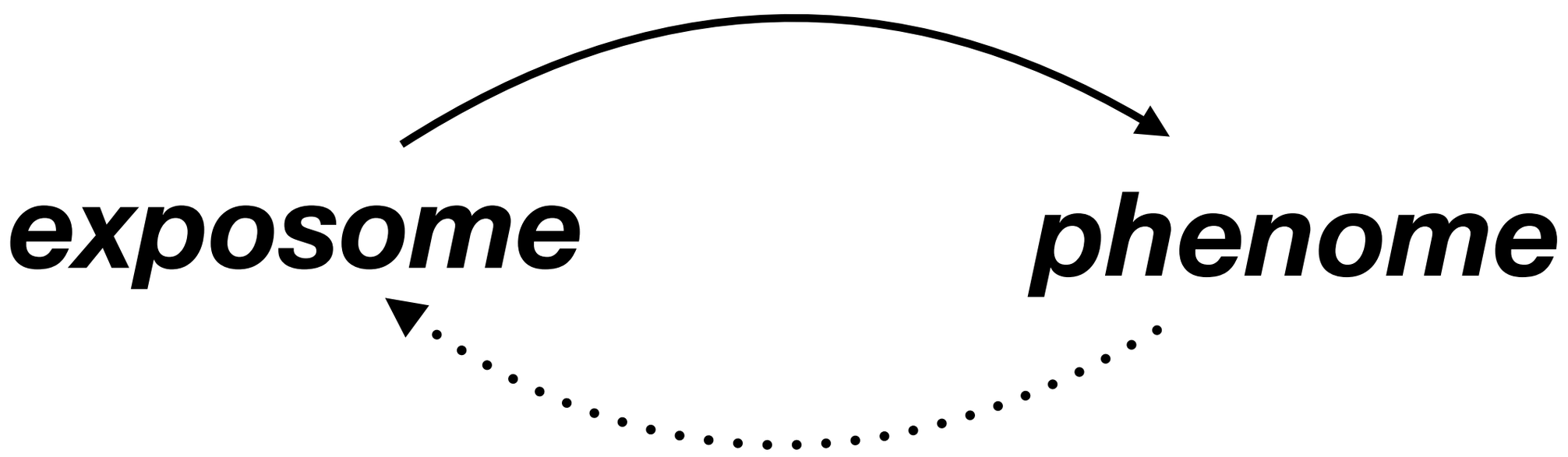
internal	external
lead (serum)	geography
nutrients (serum)	air pollution
infection (urine)	income
metabolome	

phenome

function	diseases
expression	diabetes
telomeres	cancer
metabolome	heart disease

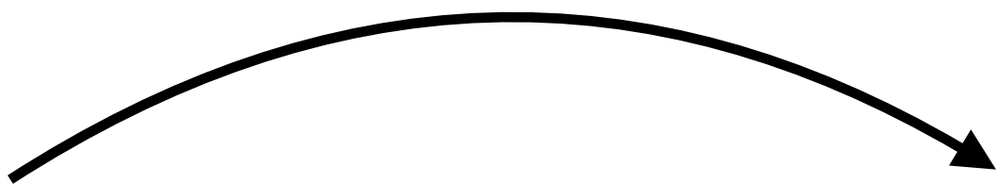
Exposome associated with the ***phenome***?

…and vice versa?



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

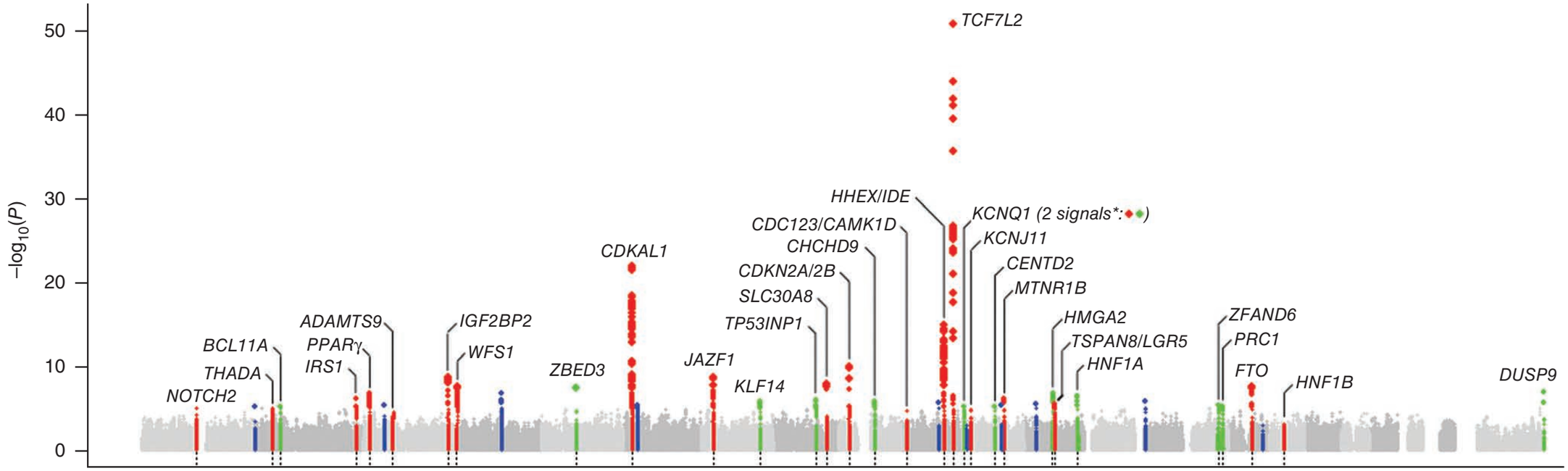
We can learn a thing or two from ***genomics*** investigation…



genome ***phenome***

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

Computational 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…

XWAS tutorial in R markdown

(R *code*)

Freely available *exposome* data for your research

(NHANES: ~40,000 individuals and 1,000 variables)

Materials for teaching and demonstration

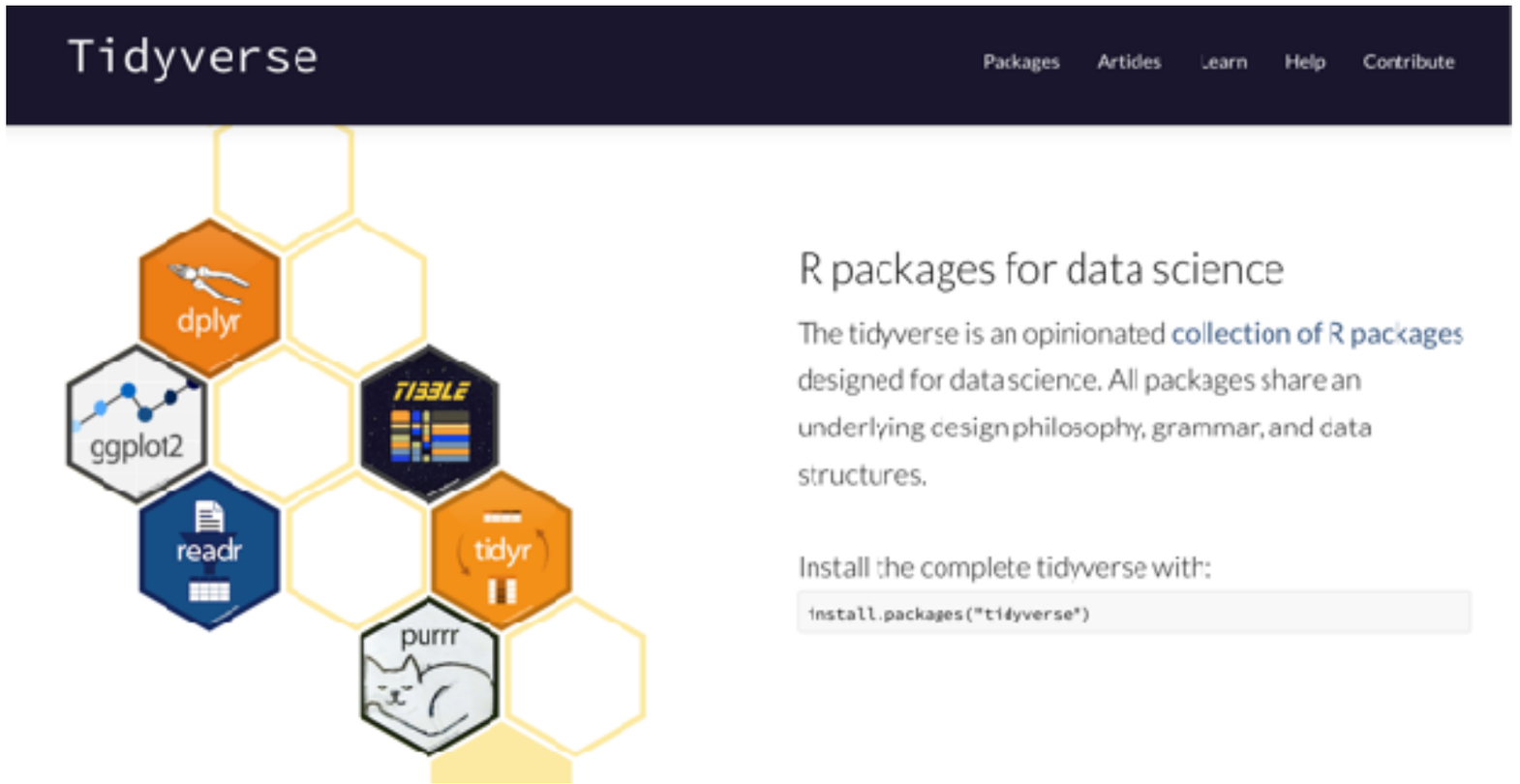
first…

learn ***RStudio,*** the ***tidyverse***, and ***github***

*…*practice ***regression methods***!

Intro to Data Science: *Integrating Genomes, Exposomes, and Phenomes*

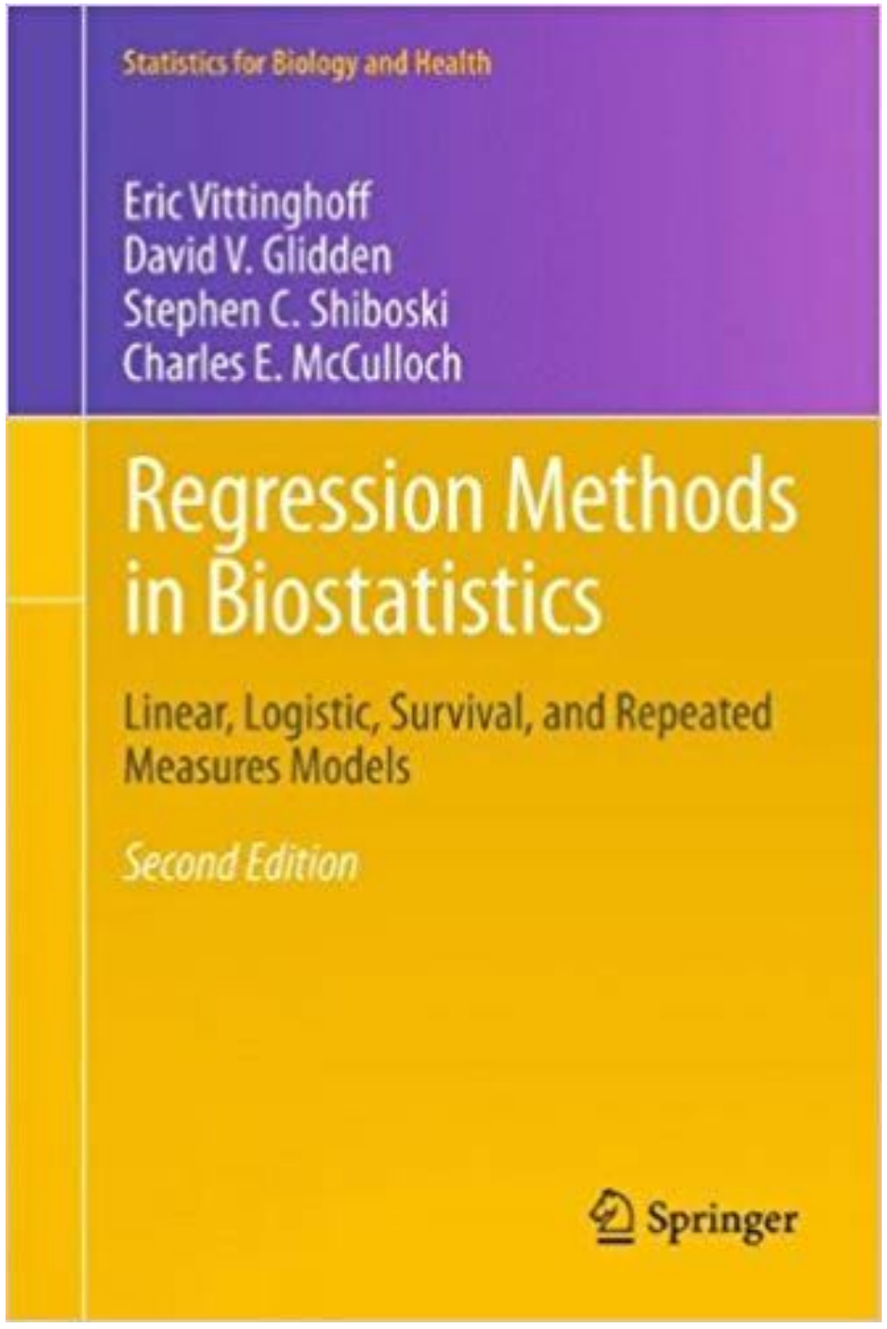
Syllabus, references and readings

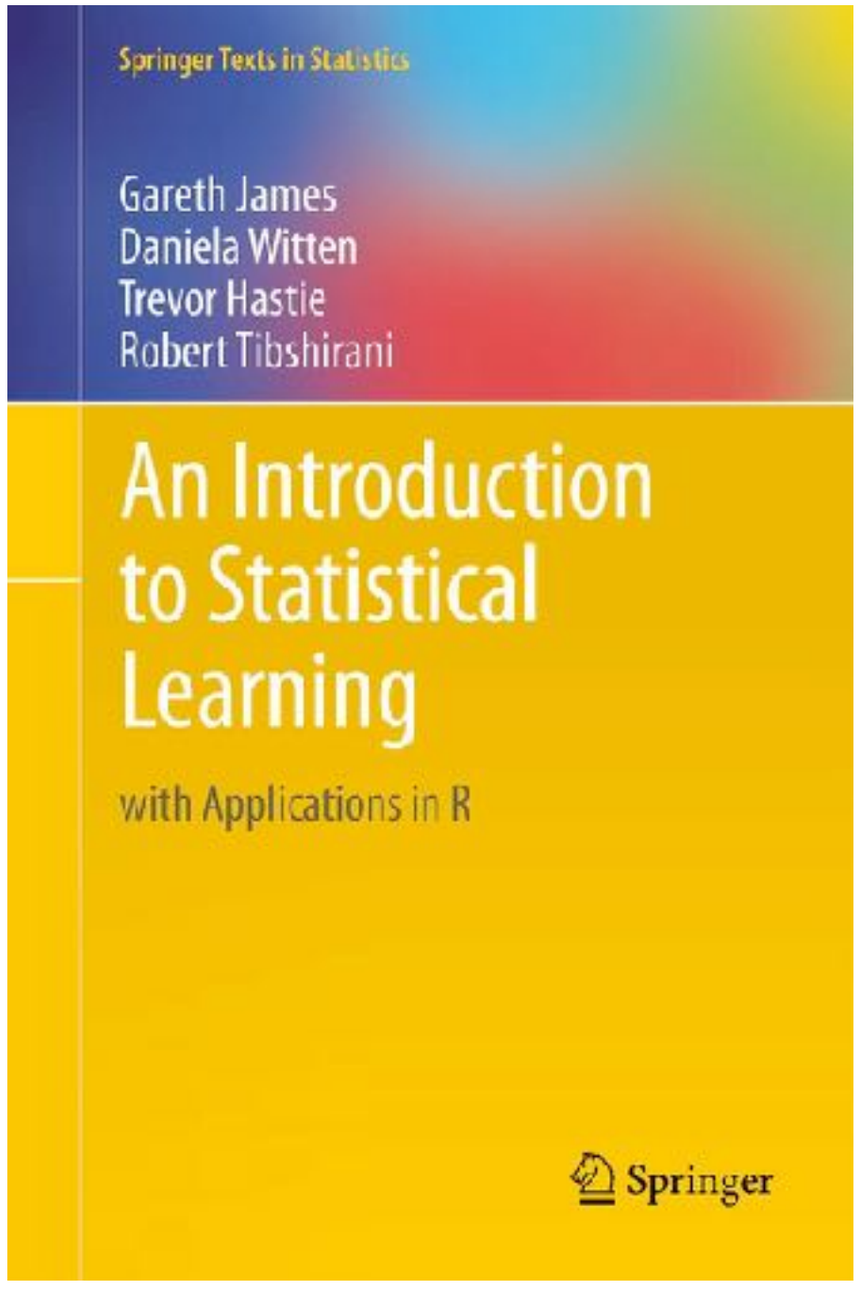


~90% (or more) of data science effort in ***processing*** — learn to do it efficiently, reproducibly, and quickly!

Intro to Data Science: *Integrating Genomes, Exposomes, and Phenomes*

Syllabus, references and readings





Resources Index
(for today's session)

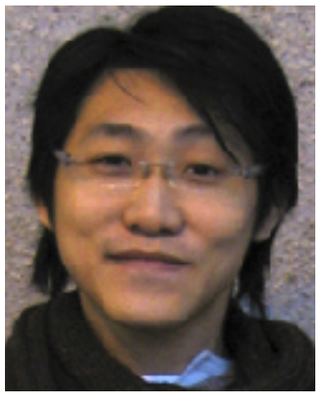
http://bit.ly/xwas_with_nhanes

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

Chirag

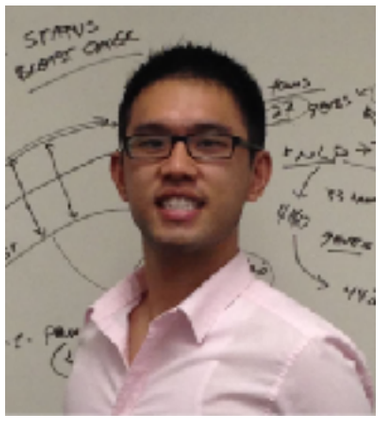
@chiragjp

Jake



@jakemkc

Nam



@nampho2

But genotypes are static and have little correlation —

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

(but we will get started today!)

JAMA 2014
Pac Symp Biocomp 2015

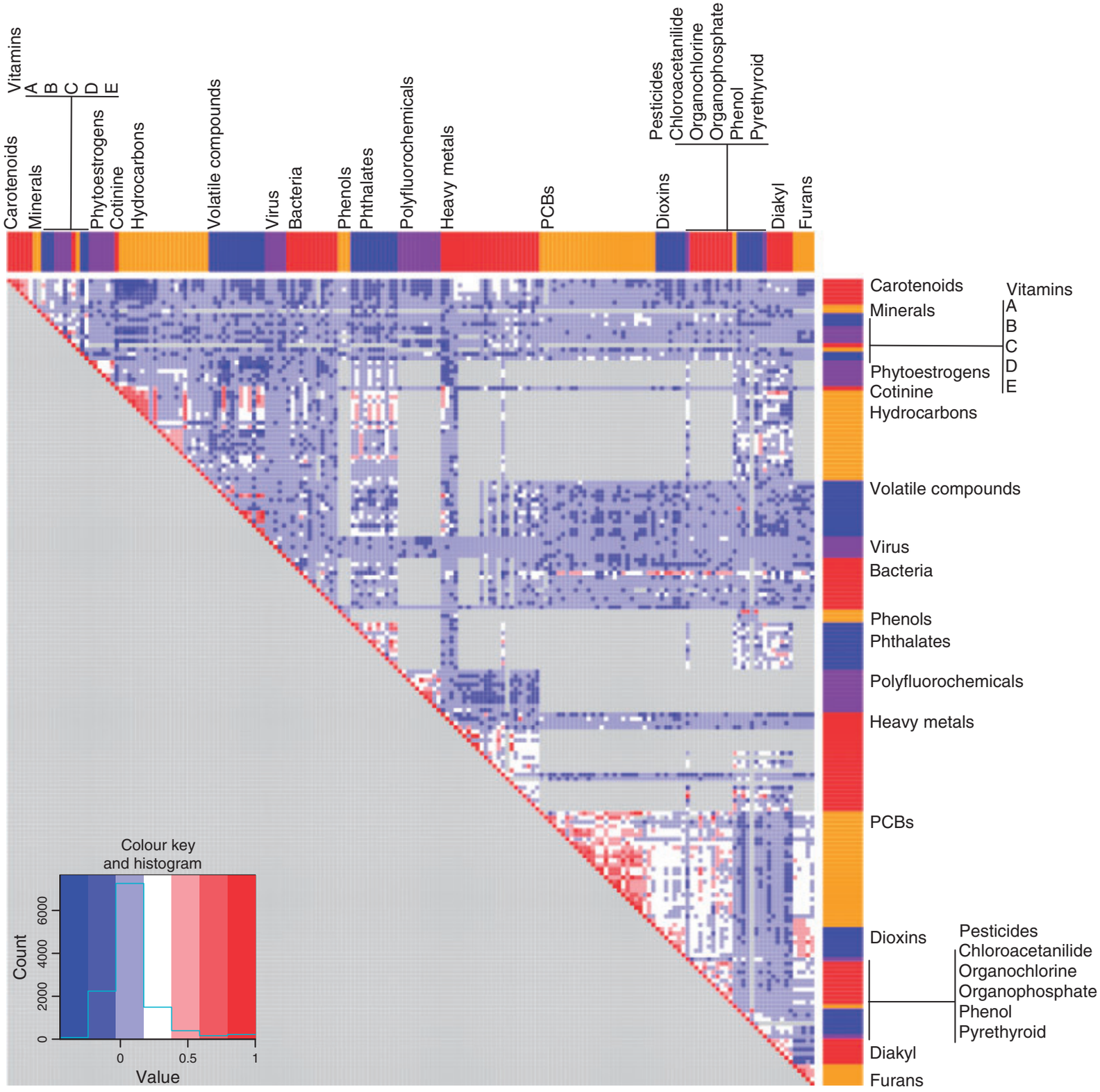
Dense correlational web of the *exposome* leads to complications in determining:

what causes what?

the influence of confounding bias?

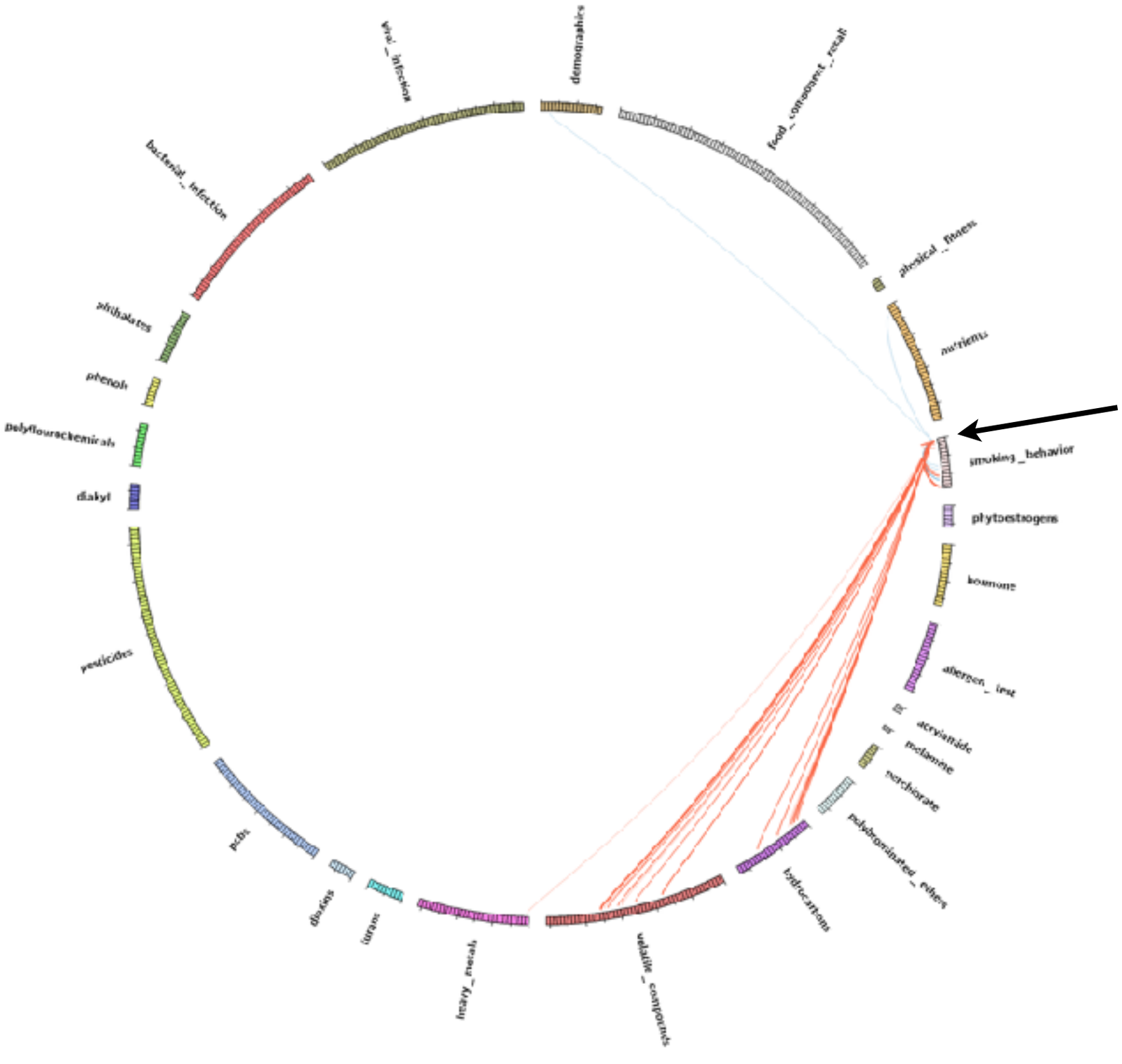
how <u>many</u> environmental factors are part of the exposome??? (10? 1000? 10000? Infinite?)

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



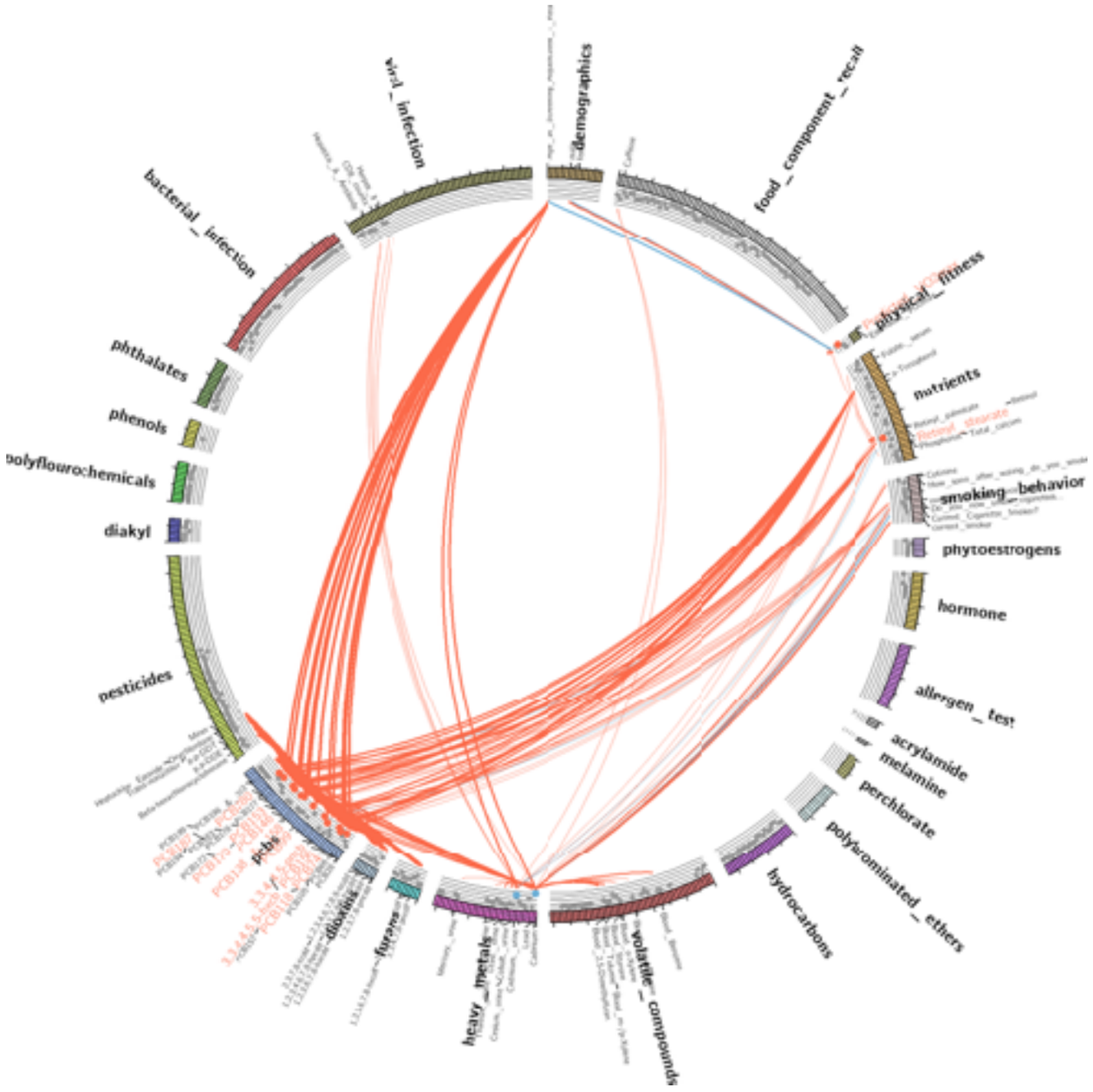
for each pair of ***E***:
Spearman ρ
(575 factors: 81,937 correlation

permuted data to produce
“null ρ”
sought replication in > 1
cohort

Red: positive ρ
Blue: negative ρ
thickness: |ρ|

Pac Symp Biocomput. 2015
JECH. 2015

Interdependencies of the ***exposome***: Telomeres vs. all-cause mortality

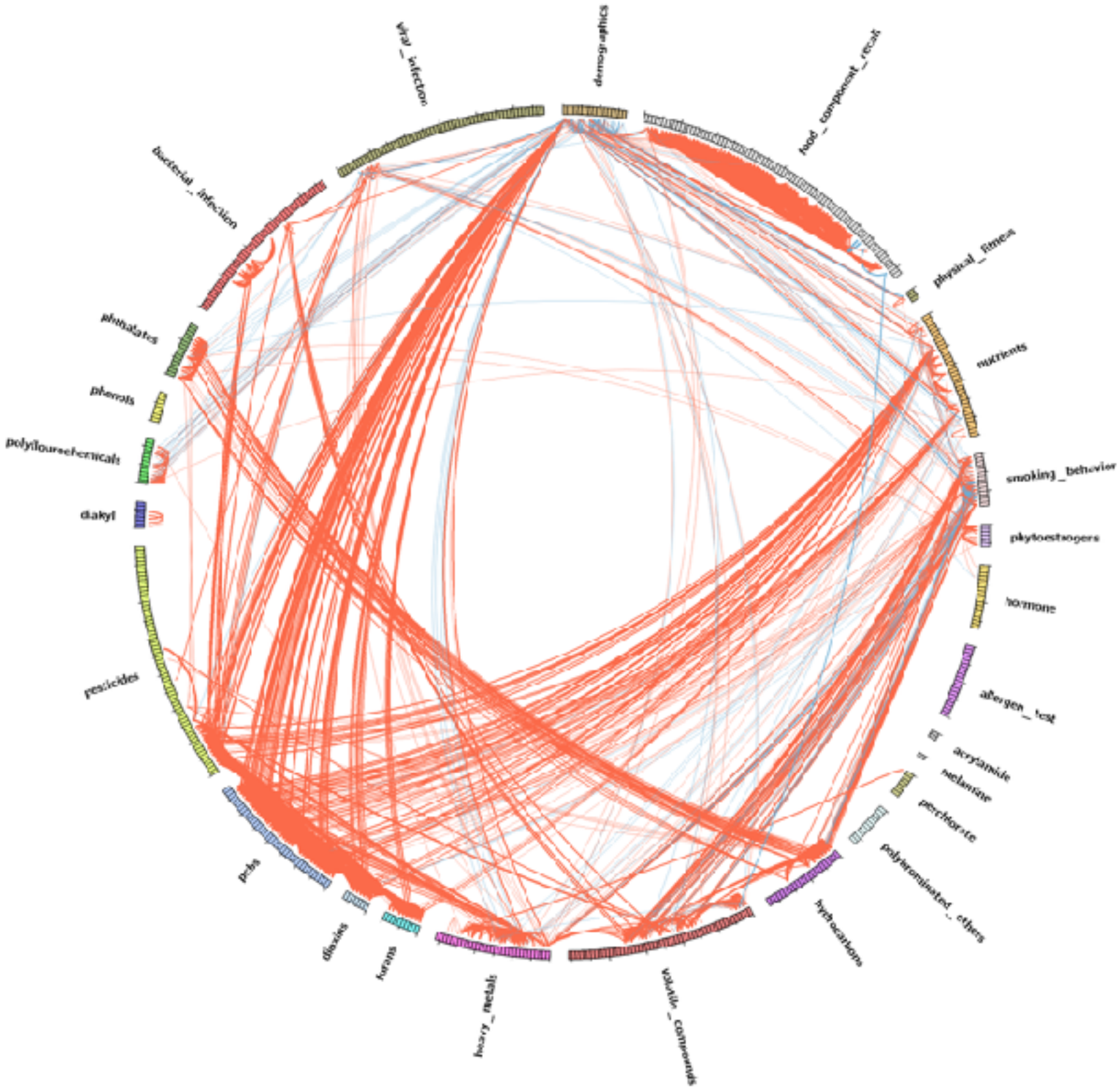


Telomere Length

All-cause mortality

http://bit.ly/globebrowse

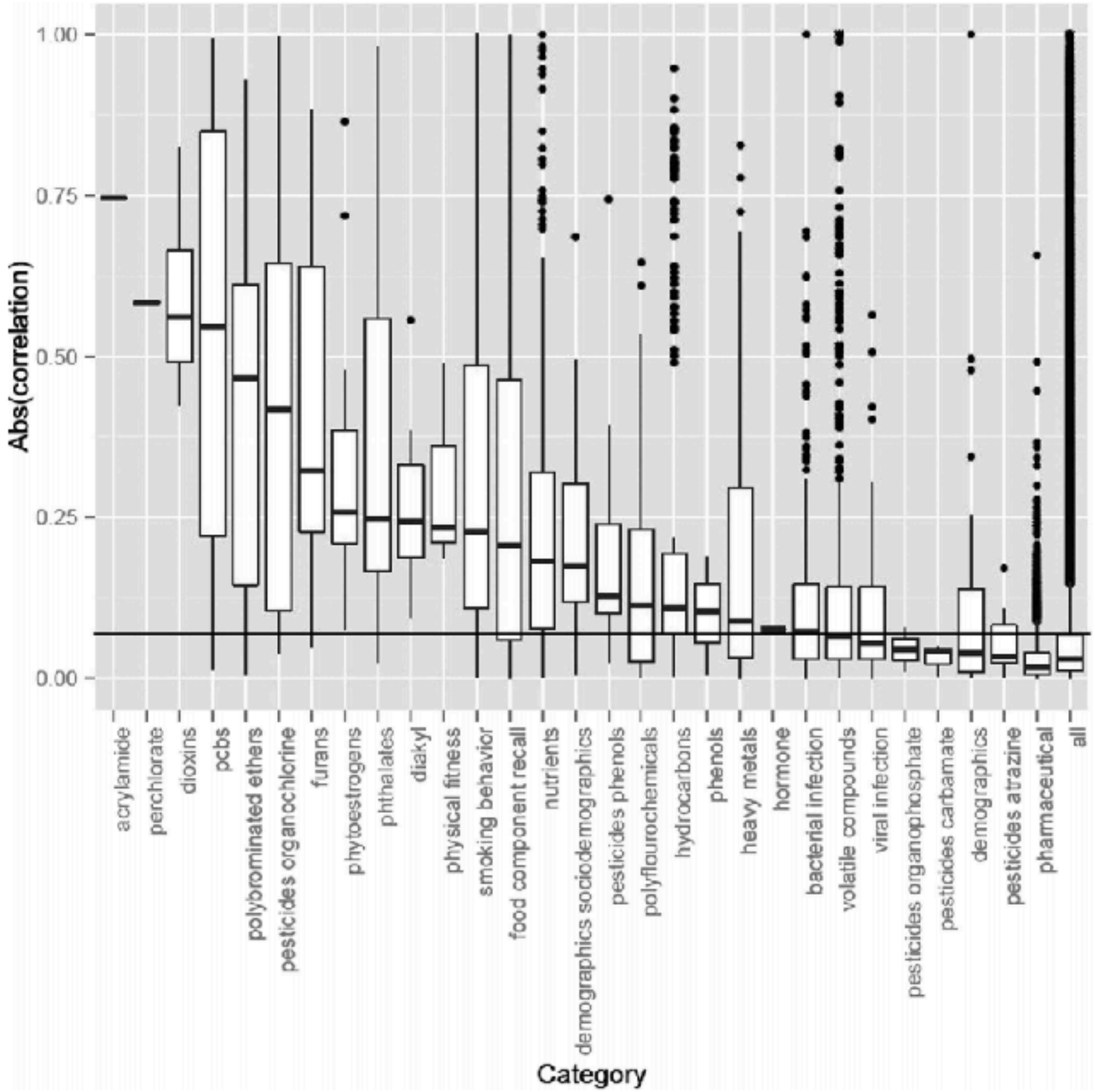
Interdependencies of the ***exposome***: Correlation globes paint a complex view of exposure



http://bit.ly/globebrowse

Pac Symp Biocomput. 2015
JECH. 2015

Interdependencies of the ***exposome***:
Modest correlation size (Spearman/biserial ρ <0.5)



JECH. 2015

Interdependencies of the ***exposome***: Modest correlations ~ number of ***effective variables (Meff)*** ~= the ***number of variables measured (M)***!

Table 2 Number of variables in each of the 32 exposure, demographics or behaviour categories

Category (measurement type)	M	Meff	Mdiff	p Value
Polychlorinated biphenyls (s)	38	23.79	14.21	0.002
Dioxins (s)	7	4.90	2.10	0.01
Acrylamide (s)	2	1.44	0.56	0.03
Organochlorine pesticides (s)	13	9.95	3.05	0.005
Polybrominated ethers (s)	11	8.47	2.53	0.006
Furans (s)	10	7.99	2.01	0.006
Perchlorate (u)	2	1.66	0.34	0.03
Smoking behaviour (s/sr)	14	11.65	2.35	0.004
Phthalates (u)	13	10.85	2.15	0.005
Food component recall (sr)	74	63.92	10.08	0.0008
Phytoestrogens (u)	6	5.22	0.78	0.01
Hydrocarbons (u)	21	18.47	2.53	0.003
Nutrients (s)	29	26.32	2.68	0.002
Volatile compounds (s)	38	34.60	3.40	0.001
Physical fitness (sr*)	3	2.78	0.22	0.02
Dialkyl metabolites (u)	6	5.59	0.41	0.009
Socioeconomics (SES) (sr)	9	8.49	0.51	0.006
Demographics (sr)	8	7.55	0.45	0.007
Polyfluorochemicals (s)	12	11.39	0.61	0.004
Phenol pesticides (u)	7	6.65	0.35	0.008
Heavy metals (s/u)	29	27.68	1.32	0.002
Bacterial infection (s/u)	33	32.19	0.81	0.002
Viral infection (s)	16	15.72	0.28	0.003
Phenols (u)	3	2.97	0.03	0.02
Atrazine-like pesticides (u)	6	5.98	0.02	0.008
Hormone (s)	2	1.99	0.01	0.03
Pharmaceutical (sr)	107	106.78	0.22	0.0005
Organophosphate pesticides (u)	4	4.00	0.00	0.01
Carbamate pesticides (u)	4	4.00	0.00	0.01
Melamine (u)	1	1.00	0.00	0.05
Chloroacetanilide pesticides (u)	1	1.00	0.00	0.05
Pyrethyroid pesticides (u)	1	1.00	0.00	0.05
Total	530	476.0	54.0	0.0001

M, number of variables in category; Meff, effective number of variables after taking into account correlation; Mdiff, difference between M and Meff; p value, indicative correlation-adjusted Bonferroni p value threshold (0.05/Meff).
*Denotes quantitative assessment of physical activity (VO2Max) was also measured.
s, serum; sr, self-report; u, urine.

$$M_{\text{eff}} = 1 + (M - 1)\left(1 - \frac{\text{Var}(\lambda_{\text{obs}})}{M}\right).$$

If everything was correlated:

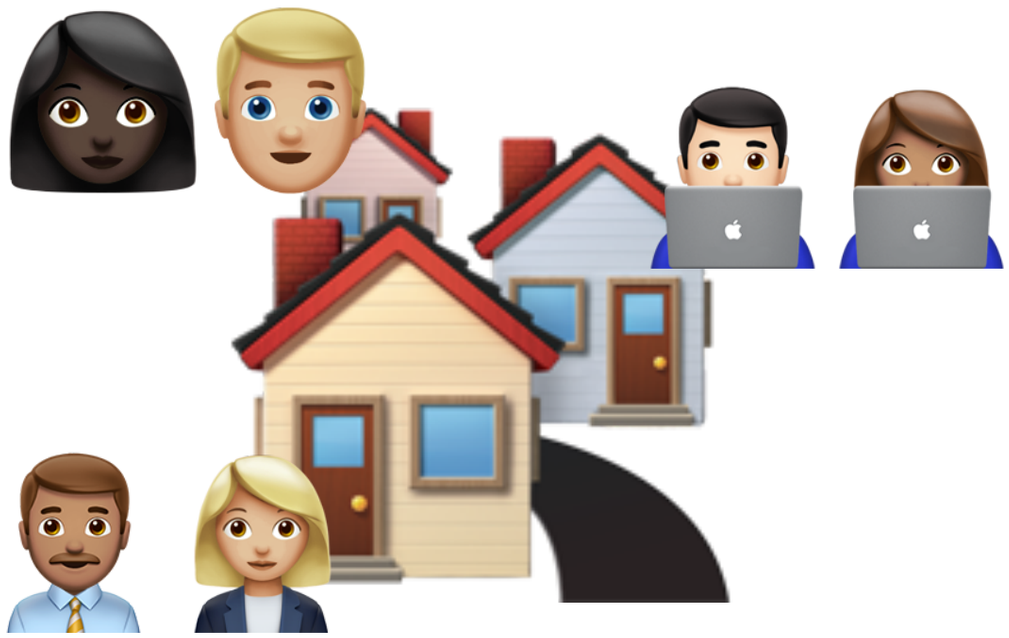
Meff = 1!

If everything was ***not*** correlated (independent):

Meff = M!

JECH 2015

How large of a role does ***shared environment*** play a significant role in co-exposure of the exposome?



Possible to capture ***household*** and ***gender*** influence on variation of ***E***?

Does living together mean higher correlation?
E of individuals in the same home ***predictive*** of others?

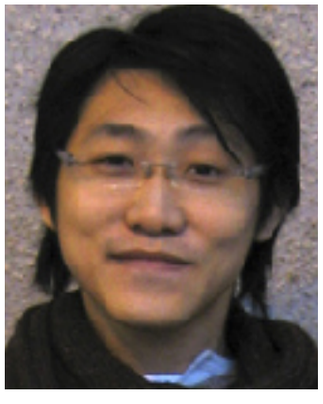
https://github.com/jakemkc/exposome_variability

Chung et al, ES&T 2018

Longitudinal Investigation of Fertility and the Environment (*LIFE*): a prospective study of **couples** desiring to become pregnant

- Reproductive age (**18-40** for females; **>18** for males)
- **N=501 couples** (Michigan and Texas) in 2005-2007 living in the same home
- Data collected in couples' home
  - urine, blood, semen (at baseline and at month 1)

Germaine Buck Louis (NICHD)



Jake Chung

https://github.com/jakemkc/exposome_variability

Buck Louis et al, 2013
Buck Louis et al, 2014
Chung et al, ES&T 2018

Longitudinal Investigation of Fertility and the Environment (*LIFE*): a prospective study of **couples** with many indicators of ***E!***

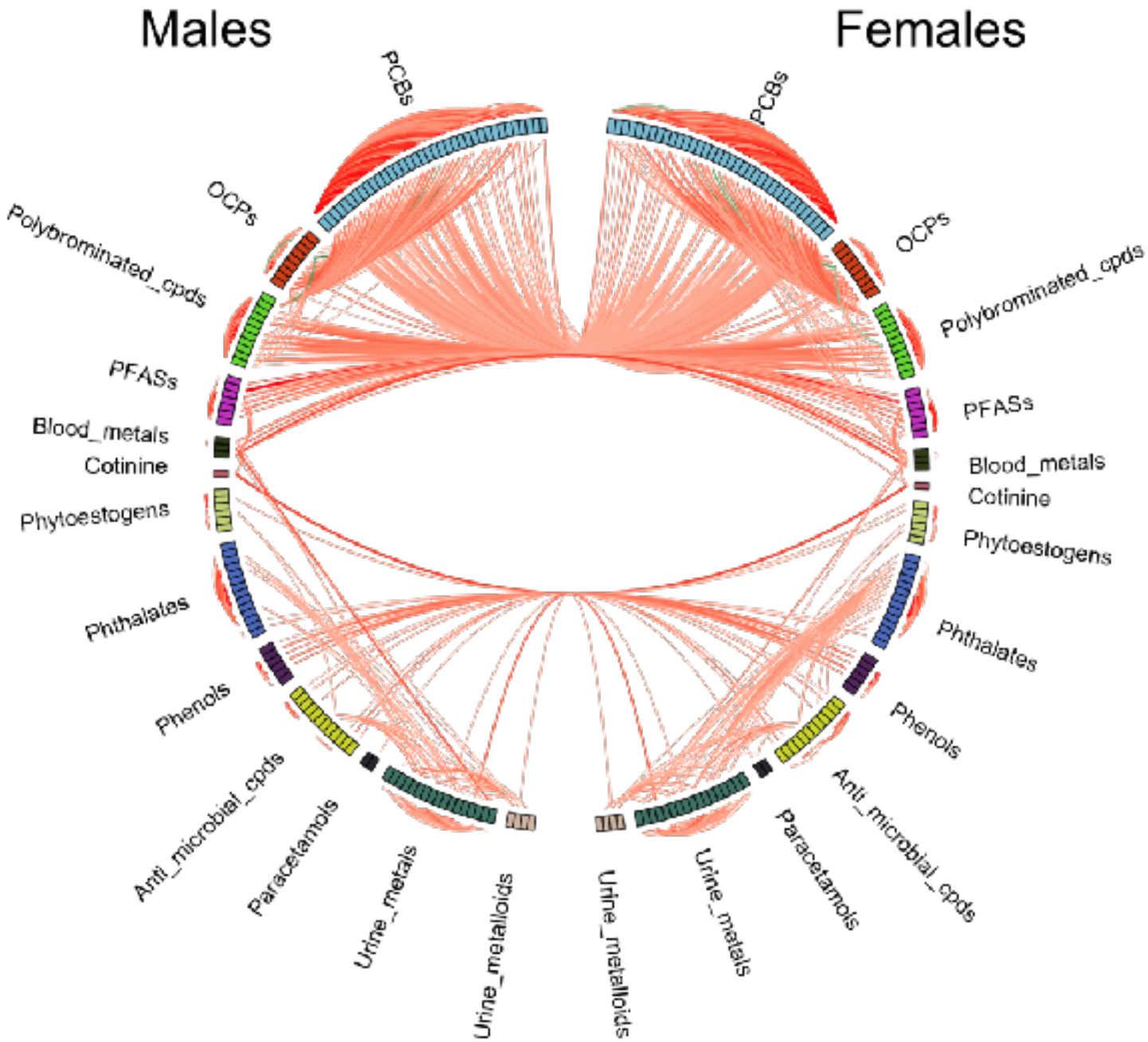
Chemical classes	No.	Chemicals
Serum persistent organic compounds		
Polychlorinated biphenyls (PCBs)	36	Congeners: 28, 44, 49, 52, 66, 74, 87, 99, 101, 105, 110, 114, 118, 128, 138, 146, 149, 151, 153, 156, 157, 167, 170, 172, 177, 178, 180, 183, 187, 189, 194, 195, 196, 201, 206, and 209
Organochlorine pesticides (OCPs)	9	Hexachlorobenzene (HCB), β-hexachlorocyclohexane (β-HCH), γ-hexachlorocyclohexane (γ-HCH), oxychlordane, *trans*-nonachlor, *p,p'*-DDT, *o,p'*-DDT, *p,p'*-DDE, and mirex
Polybrominated chemicals	11	Brominated biphenyl (BB 153); brominated diphenyl ethers (BDEs) congeners: 17, 28, 47, 66, 85, 99, 100, 153, 154, and 183
Polyfluoroalkyl substances (PFASs)	7	2-(*N*-ethyl-perfluorooctane sulfonamido) acetate (Et-PFOSA-AcOH), 2-(*N*-methyl-perfluorooctane sulfonamido) acetate (Me-PFOSA-AcOH), perfluorodecanoate (PFDeA), perfluorononanoate (PFNA), perfluorooctane sulfonamide (PFOSA), perfluorooctane sulfonate (PFOS), and perfluorooctanoate (PFOA)
Urinary non-persistent organic compounds		
Phytoestrogens	6	Genistein, daidzein, O-desmethylangolensin, equol, enterodiol, and enterolactone
Phthalate metabolites	14	Mono (3-carboxypropyl) phthalate (mCPP), monomethyl phthalate (mMP), monoethyl phthalate (mEP), mono (2-isobutyl phthalate) (miBP), mono-n-butyl phthalate (mBP), mono (2-ethyl-5-carboxyphentyl) phthalate (mECPP), mono-[(2-carboxymethyl) hexyl] phthalate (mCMHP), mono (2-ethyl-5-oxohexyl) phthalate (mEOHP), mono (2-ethyl-5-hydroxyhexyl) phthalate (mEHHP), monocyclohexyl phthalate (mCHP), monobenzyl phthalate (mBzP), mono (2-ethylhexyl) phthalate (mEHP), mono-isononyl phthalate (mNP), and monooctyl phthalate (mOP).
Phenols	6	Total bisphenol A (BPA); benzophenones (BPs): 4-hydroxybenzophenone (4-OH-BP), 2,4-dihydroxybenzophenone (2,4-OH-BP), 2,2',4,4'-tetrahydroxybenzophenone (2,2'4,4'-OH-BP), 2-hydroxy-4-methoxybenzophenone (2-OH-4-MeO-BP), and 2,2'-dihydroxy-4-methoxybenzophenone (2,2'-OH-4-MeO-BP)
Anti-microbial chemicals	12	Triclosan (TCS) and triclocarban (TCC); parabens: methyl paraben (MP), ethyl paraben (EP), propyl paraben (PP), butyl paraben (BP), benzyl paraben (BzP), heptyl paraben (HP), 4-hydroxy benzoic acid (4-HB), 3,4-dihydroxy benzoic (3,4-DHB), methyl-protocatechuic acid (OH-Me-P), and ethyl-protocatechuic acid (OH-Et-P)
Paracetamol & derivatives	2	Paracetamol and 4-aminophenol
Others		
Blood metals	3	Cadmium (Cd), lead (Pb), and mercury (Hg)
Serum cotinine	1	Cotinine
Urine metals	17	Manganese (Mn), chromium (Cr), beryllium (Be), cobalt (Co), molybdenum (Mo), cadmium (Cd), tin (Sn), caesium (Cs), barium (Ba), nickel (Ni), copper (Cu), zinc (Zn), tungsten (W), platinum (Pt), thallium (Tl), lead (Pb), and uranium (U)
Urine metalloids	4	Selenium (Se), arsenic (As), antimony (Sb), and tellurium (Te)

https://github.com/jakemkc/exposome_variability

Chung et al, ES&T 2018

Dense correlational web in ***LIFE*** comparable to ***NHANES:***

90% spearman correlations ranging from -0.3 to 0.3



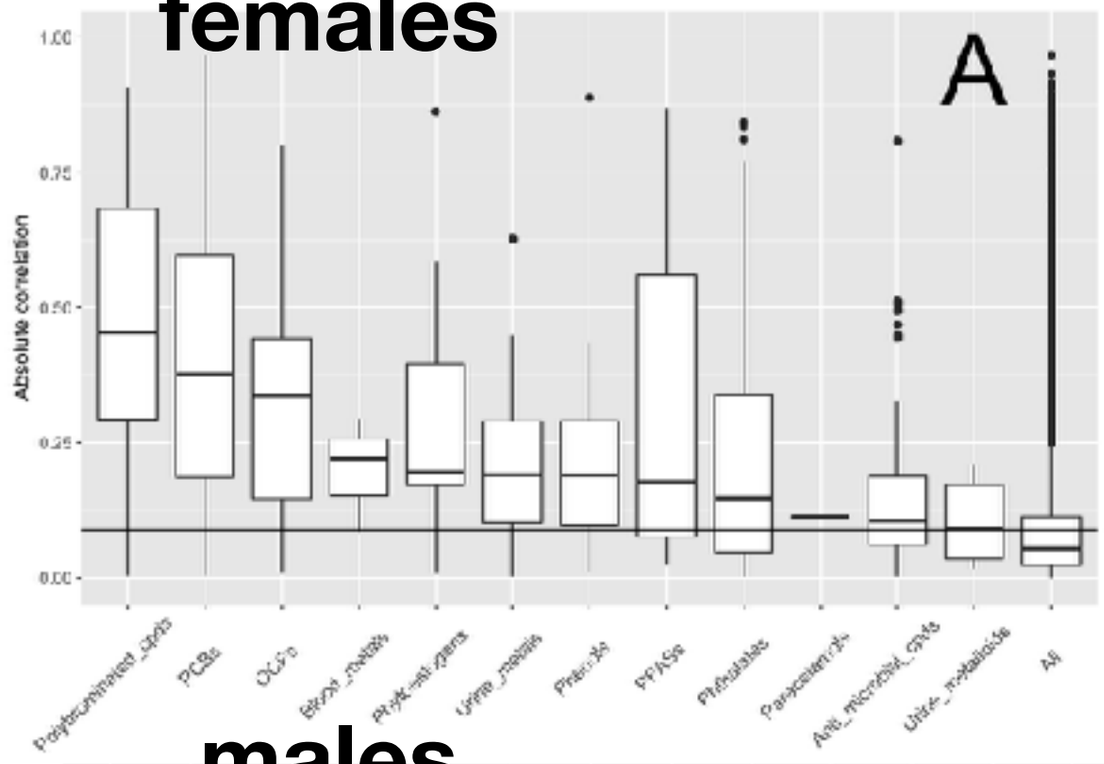
https://github.com/jakemkc/exposome_variability

Chung et al, ES&T 2018

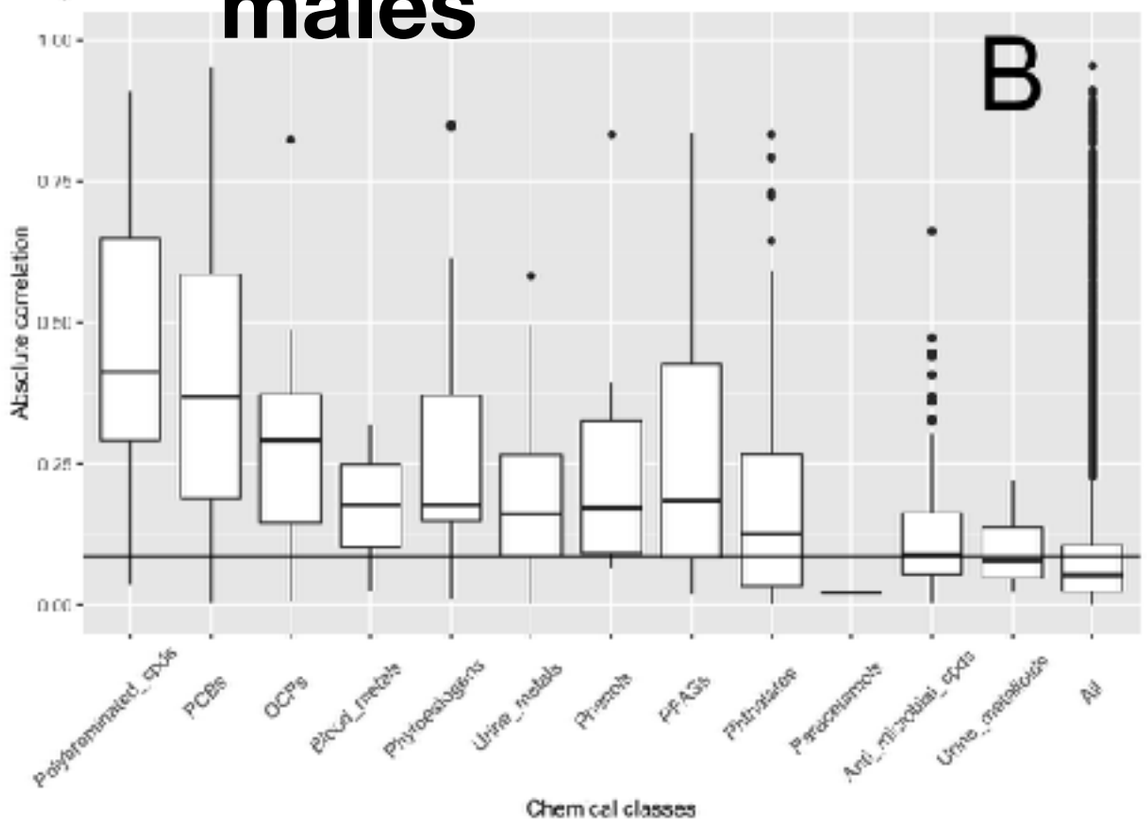
cor(persistent ***E***) **>** ***cor***(non-persistent ***E***)

(but this pattern **diminishes** between couples in the same household!)

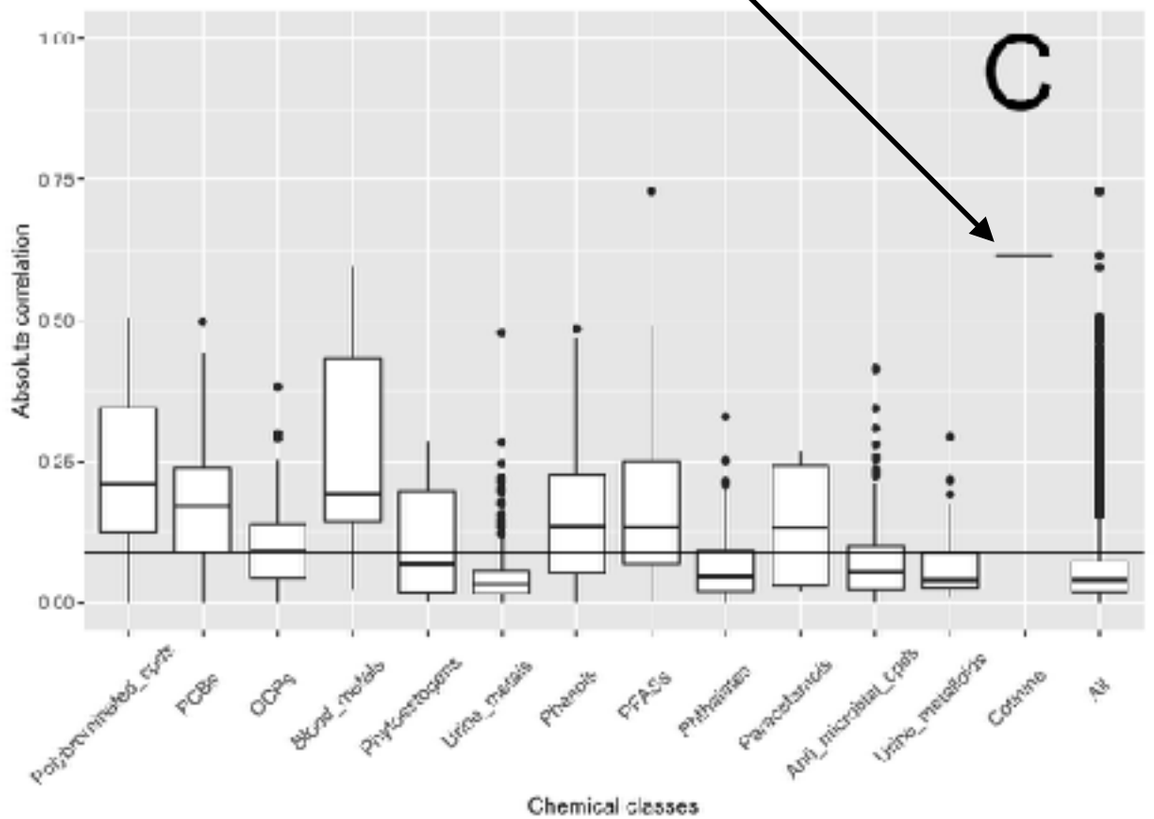
females



males



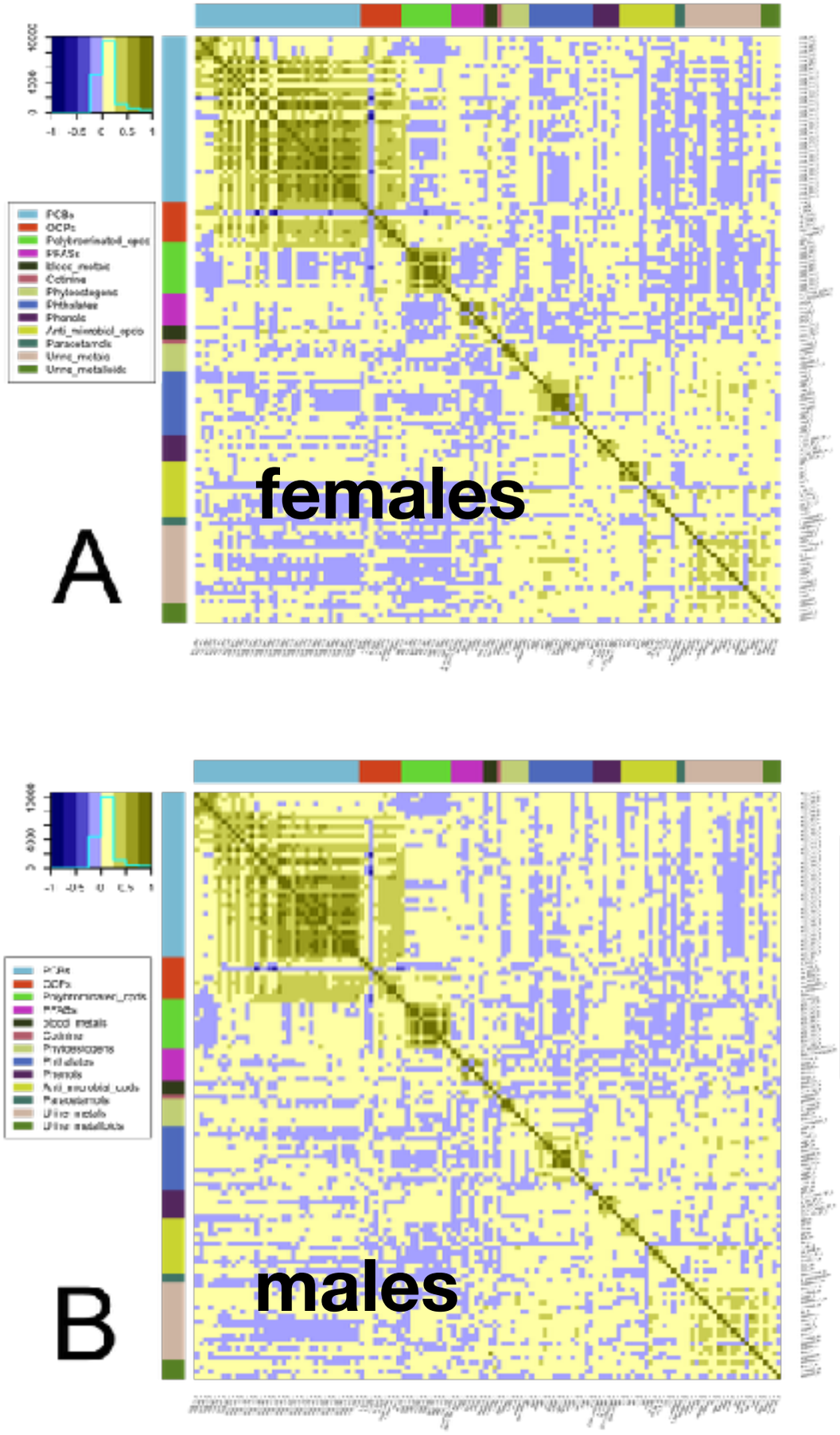
cotinine

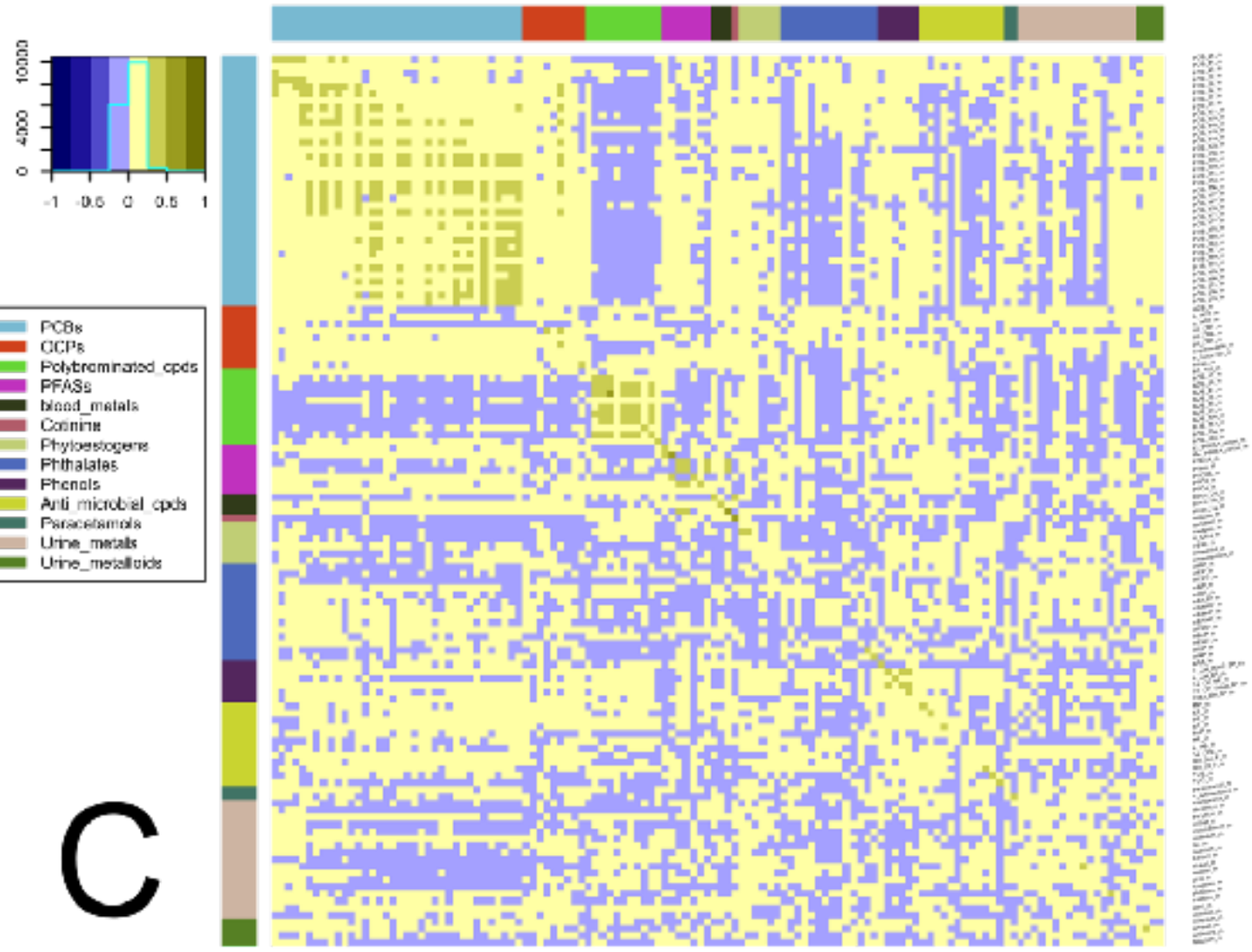


Chung et al, ES&T 2018

Co-***E*** patterns between
females and **males** are ***similar***

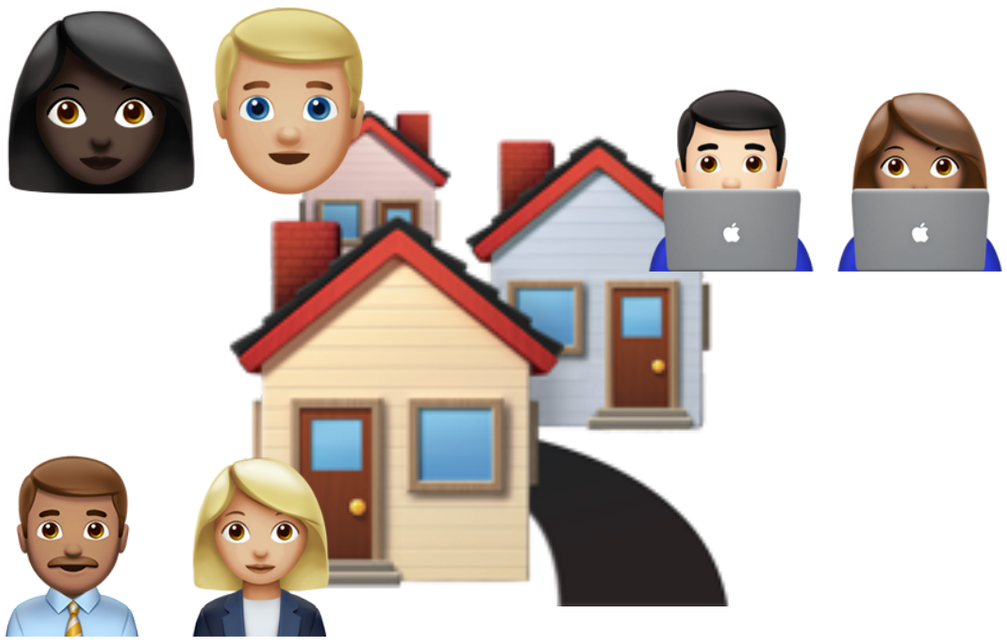
… ***however:***
males and **females *within*** household
weaker!





Chung et al, ES&T 2018

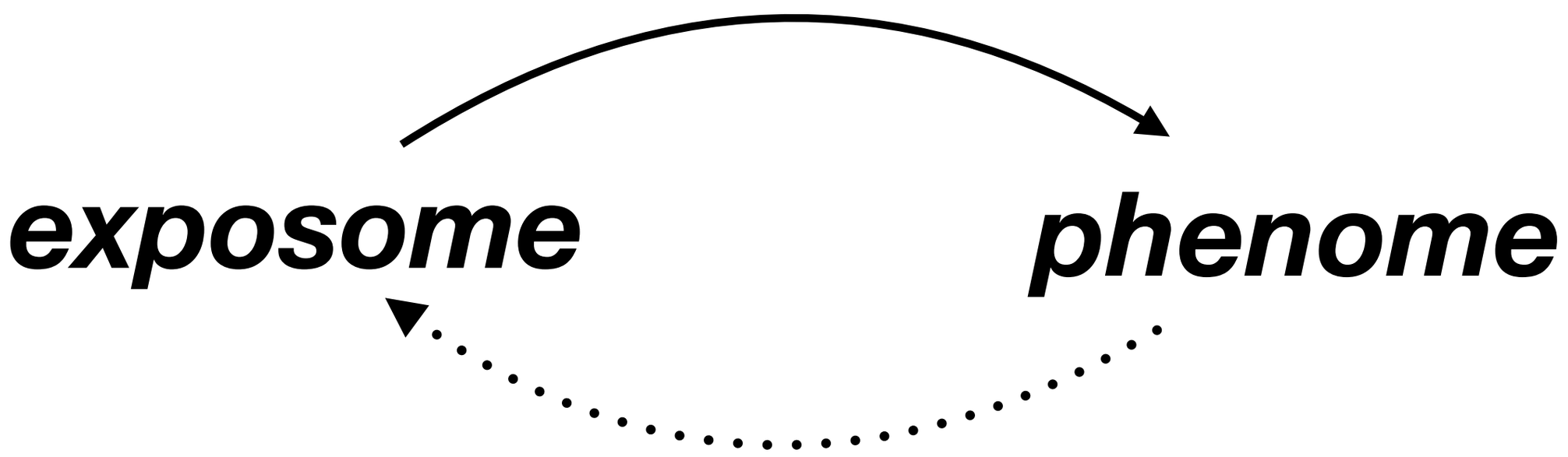
Household < ***specific*** **environment** in co-occuring exposures



implications for **sampling**, **routes** of exposure:
where do we sample? - *individuals!*

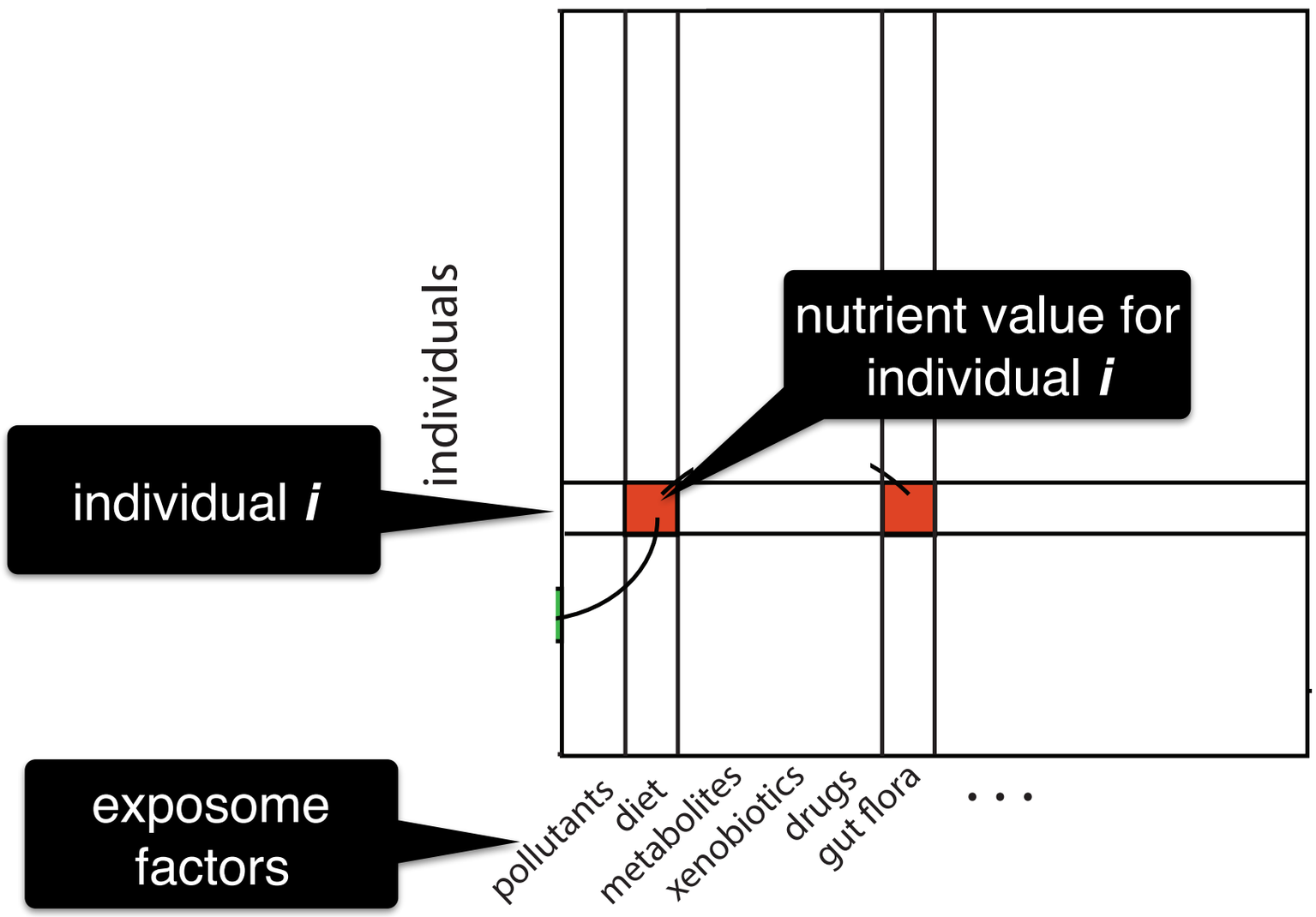
Chung et al, ES&T 2018

X-wide Association Investigations: Correlating the exposome with the phenome



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

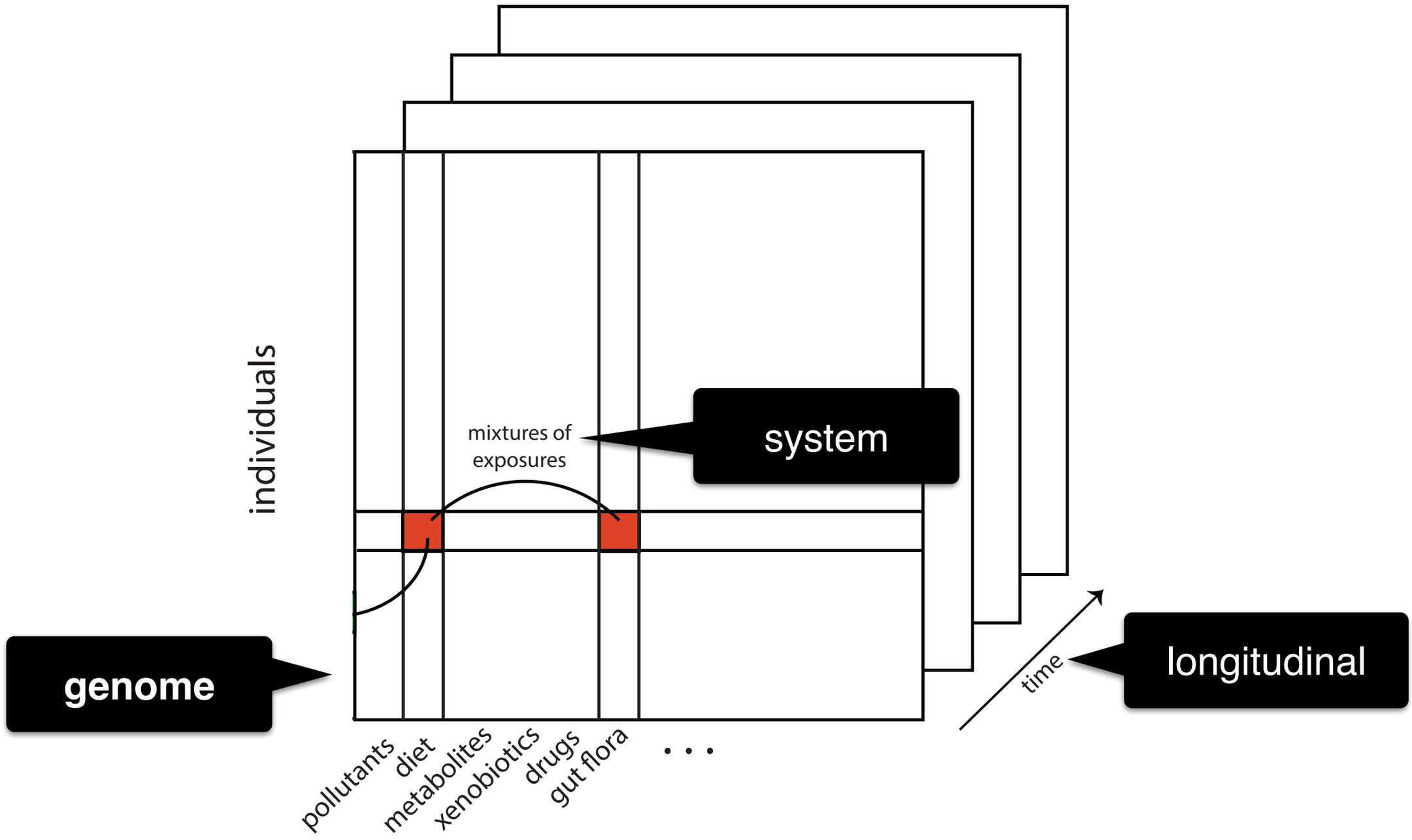
a ***high-dimensioned 3D*** matrix of
(1) ***exposure*** measurements
on (2) ***individuals*** as a function of (3) ***time***



Manrai et al ARPH 2017

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

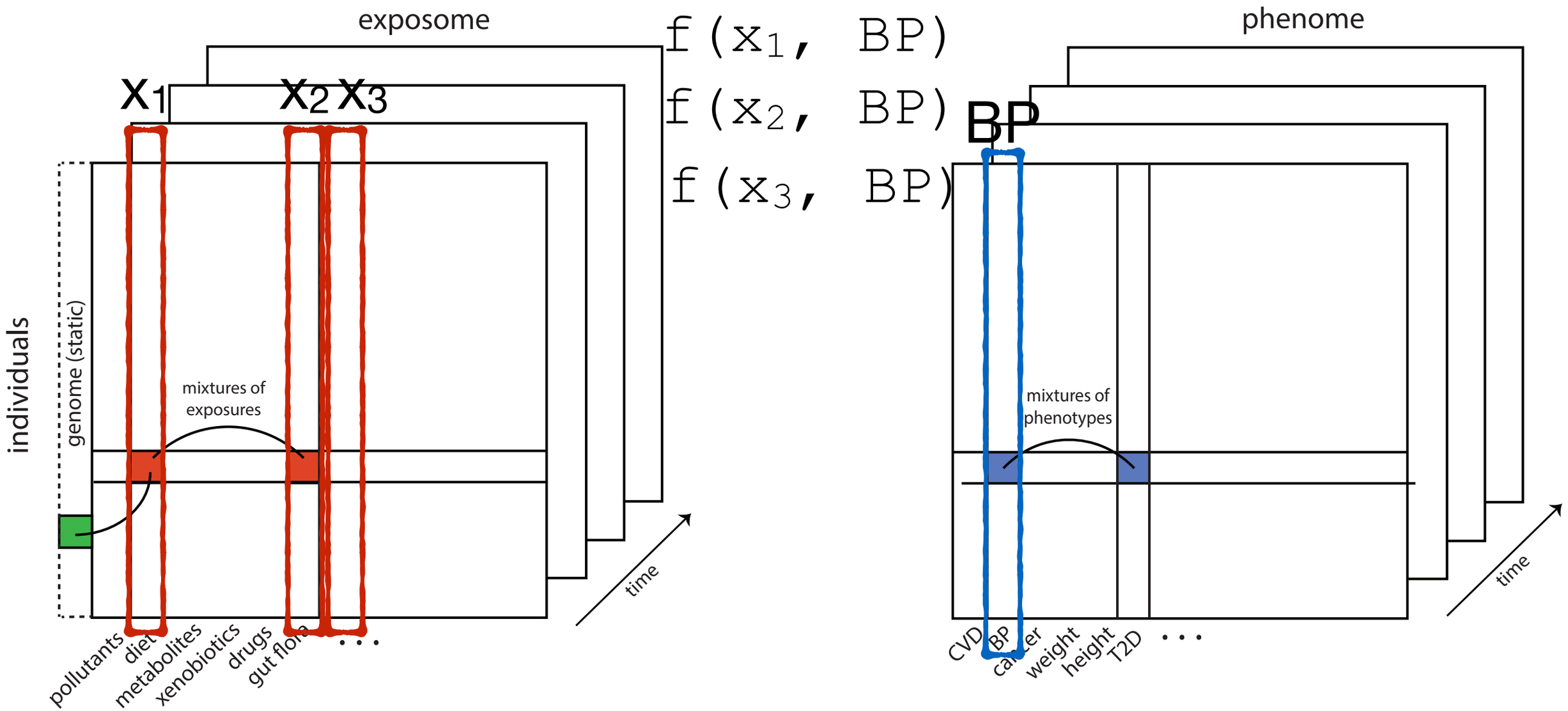
a ***high-dimensioned 3D*** matrix of
(1) ***exposure*** measurements
on (2) ***individuals*** as a function of (3) ***time***



Manrai et al ARPH 2017

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

Associating all of the 'red' with the 'blue'



Where `f` = association function (regression, correlation, etc)

Manrai et al ARPH 2017

‘pseudo-code’ for implementation of an *XWAS: how would you do it?*

```
y = [blood pressure values for cohort]

association_list = empty_list()

for each x in list of exposures:
   association_test=f(x,y)
   append(association_list, association_test)

multiplicity_correct(association_list)

volcano_plot(y, x, association_list)
```

What is stored in *y*?
What is stored in *association_list*?
What can *f* be?
What does *append* do?
What is the reason for *multiplicity_correct?*

Manrai et al ARPH 2017

Multiplicity: how to determine signal from noise?

type 1 error (spurious findings)

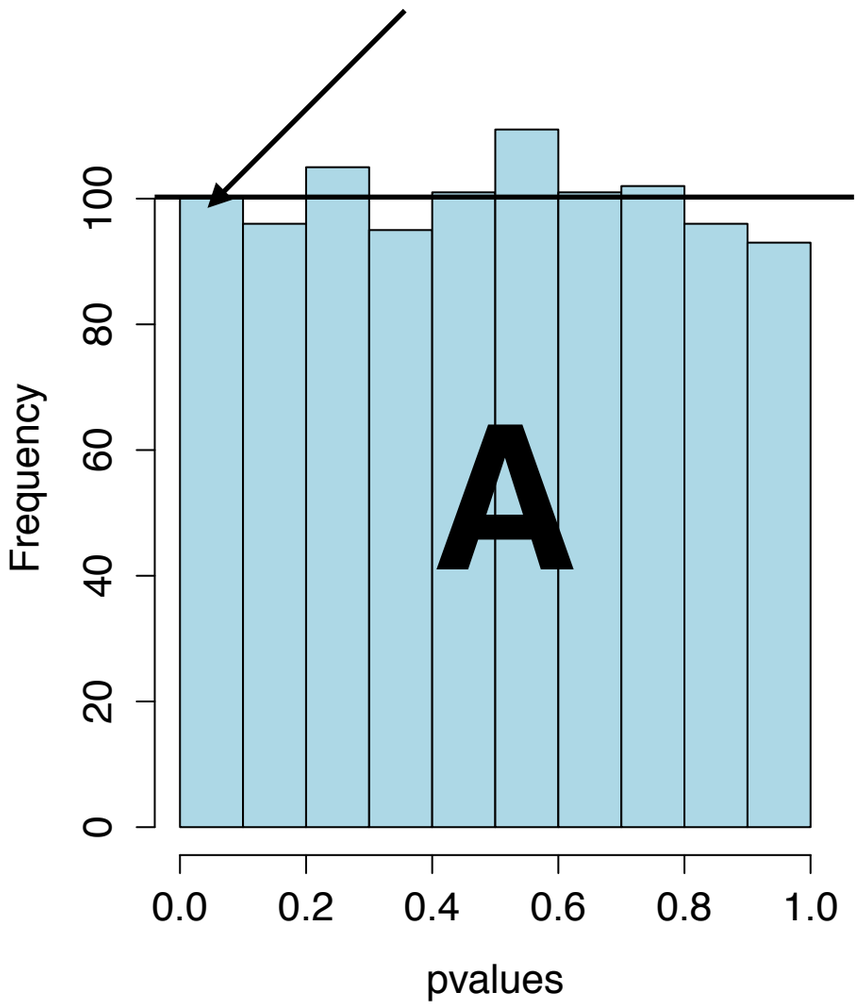
JAMA 2014

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

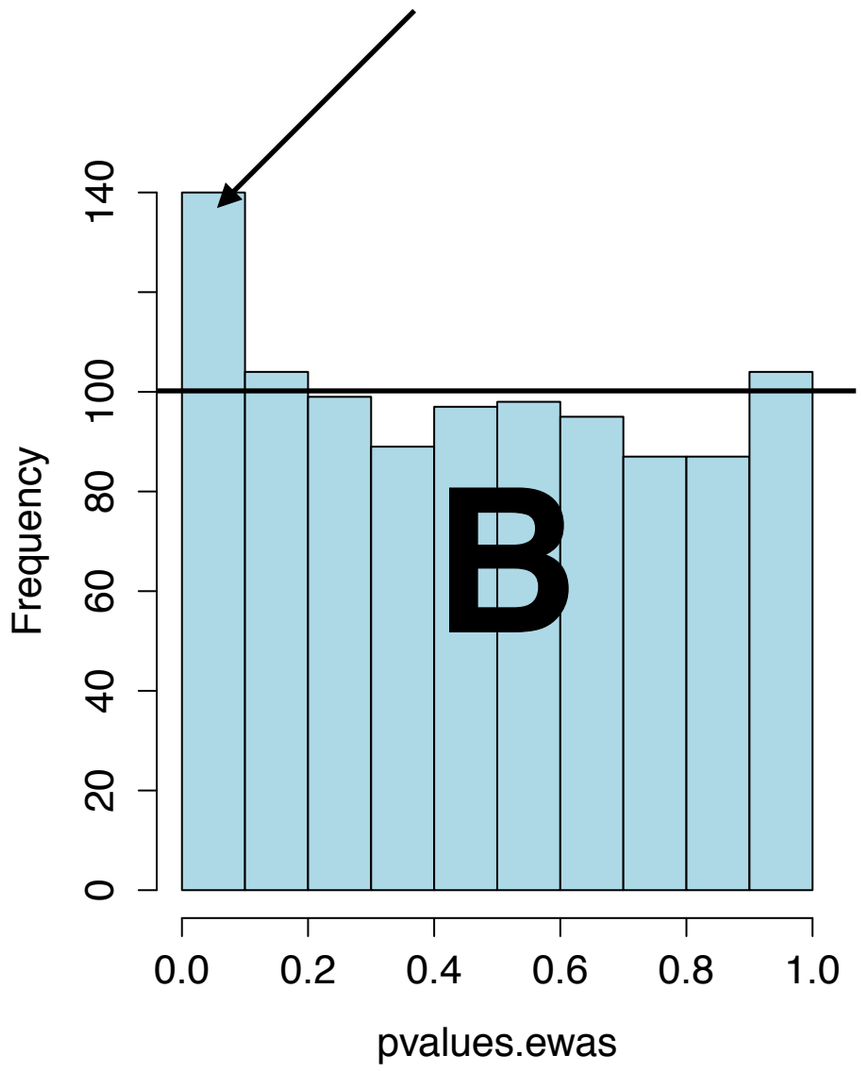
… and there were no difference between the cases and controls…

…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

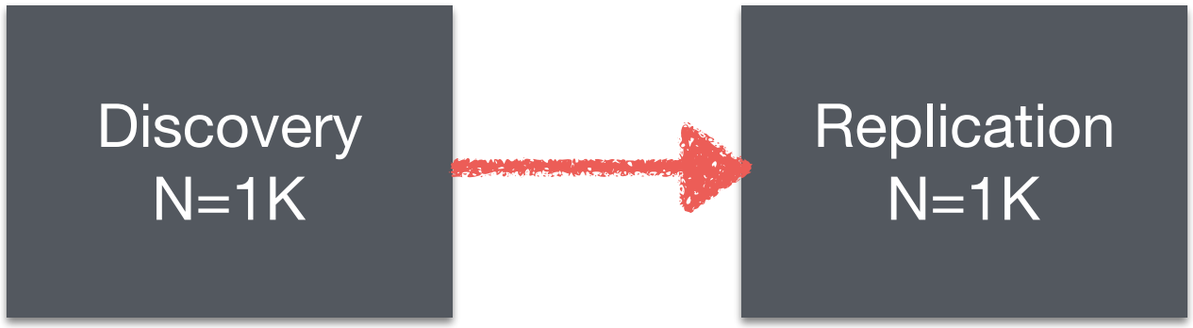


No difference
(no true associations)



5% exposures different
(5% true associations)

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

https://github.com/jakemkc/ewas_sample_size

Chung et al, in review

The ***false discovery rate***: A ***powerful*** approach for multiple hypothesis correction

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

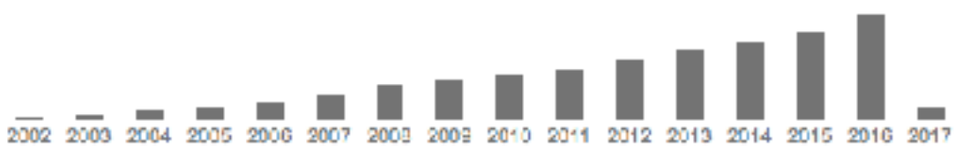
- ***"powerful"***: p-value threshold less stringent than Bonferroni
- an estimate of frequency of ***false discoveries*** at a given threshold!

The ***false discovery rate***: A ***powerful*** approach for multiple hypothesis correction

Yoav Benjamini

Controlling the false discovery rate: a practical and powerful approach to multiple testing

Authors	Yoav Benjamini, Yosef Hochberg
Publication date	1995/1/1
Journal	Journal of the royal statistical society. Series B (Methodological)
Pages	289-300
Publisher	Blackwell Publishers
Description	The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses-the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is ...
Total citations	Cited by 37060



Scholar articles Controlling the false discovery rate: a practical and powerful approach to multiple testing
Y Benjamini, Y Hochberg - Journal of the royal statistical society. Series B (..., 1995
Cited by 37060 - Related articles - All 47 versions

- ***"powerful"***: pvalue threshold less stringent than Bonferroni
- an estimate of frequency of ***false discoveries*** at a given threshold!

The ***false discovery rate***: A ***powerful*** approach for multiple hypothesis correction

TABLE 1

Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

FDR = **V / R**

- ***"powerful"***: pvalue threshold less stringent than Bonferroni
- an estimate of frequency of ***false discoveries*** at a given threshold!

How can I compute the ***False Discovery Rate?***

False Discovery Rate Estimation:
The *expected* rate of false positives

$$= \frac{\#\text{ false positives} \le \alpha}{\#\text{ findings} \le \alpha} \qquad \frac{50\text{ false positives} \le 0.05}{100\text{ findings} \le 0.05} = 0.5$$

?

$$\#\text{ false positives } (\alpha)$$

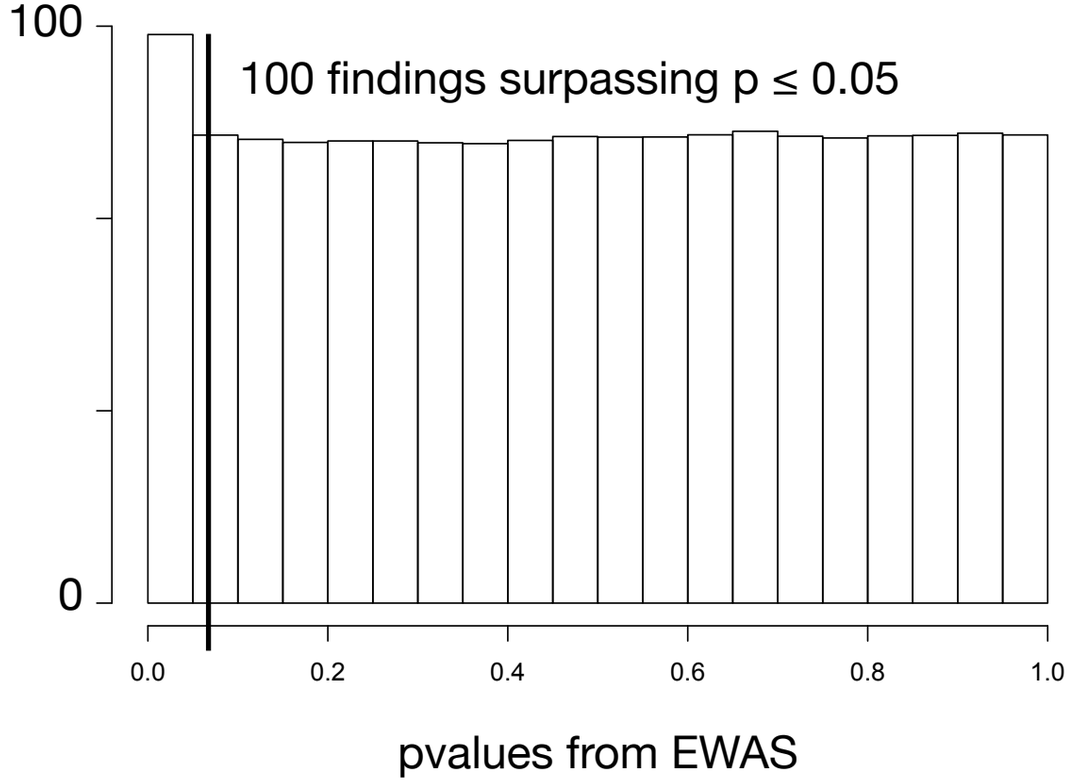


TABLE 1

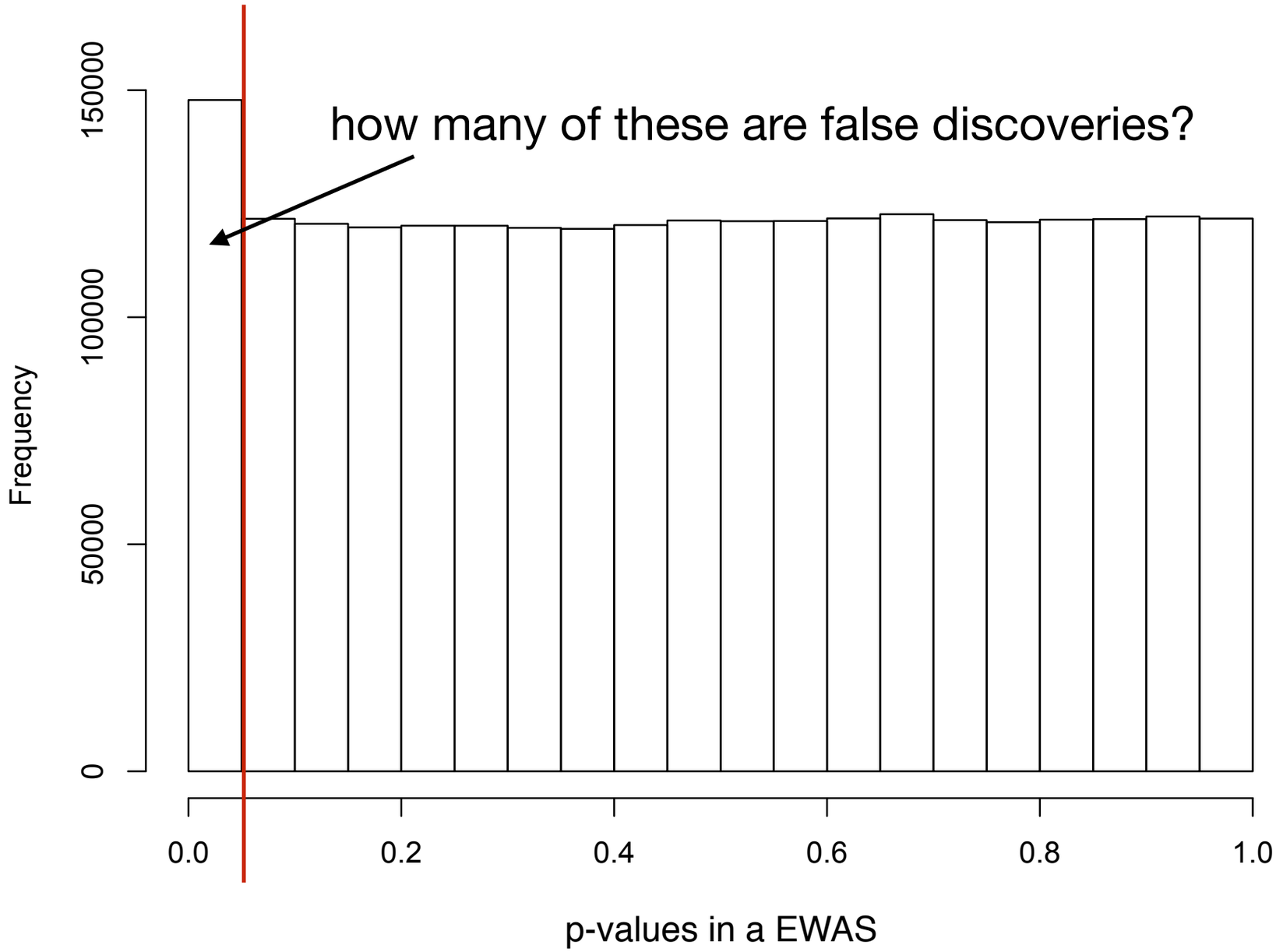
Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

We don’t know **V**!

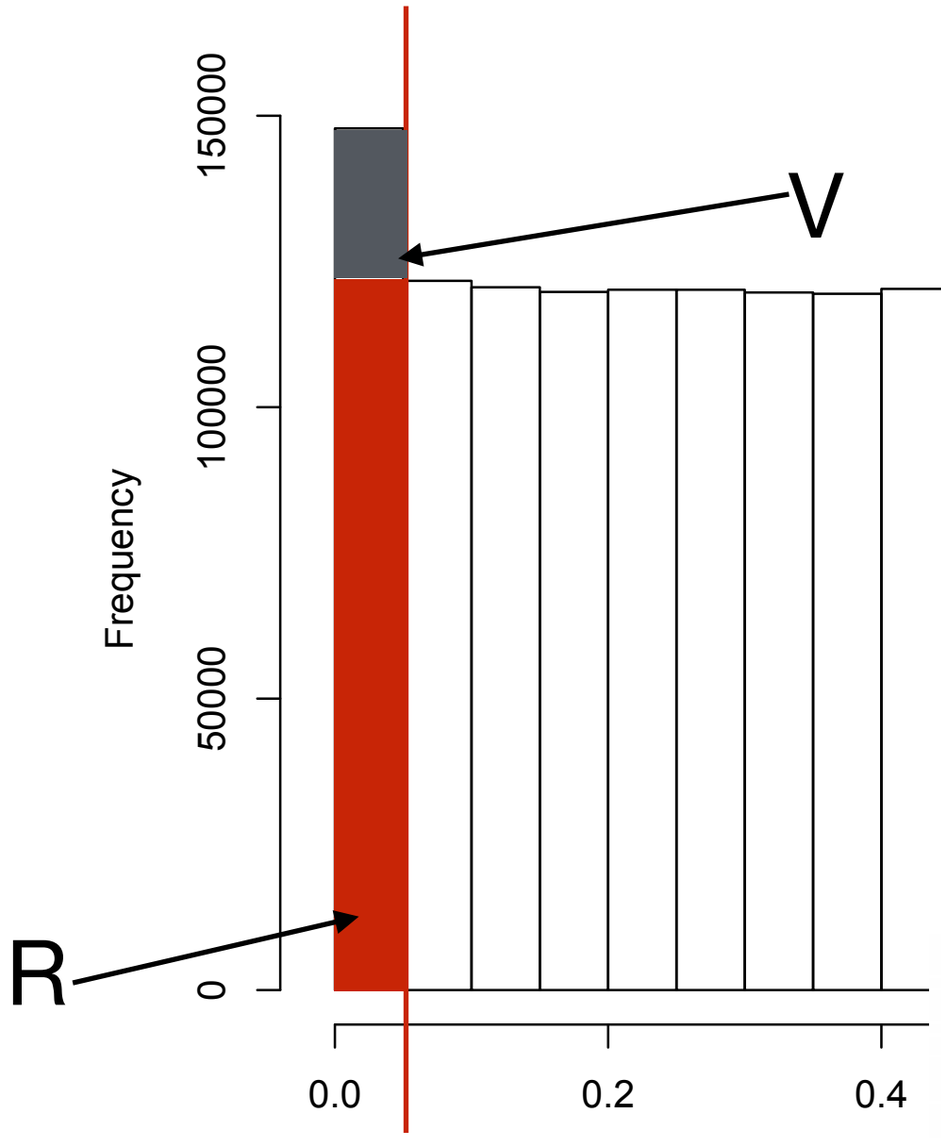
What is the ***False Discovery Rate?***

The expected number of false discoveries at a given significance threshold



What is the ***False Discovery Rate?***

The expected number of false discoveries at a given significance threshold



FDR = **V / (S+V)** = **V / R**

TABLE 1

Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Benjamini, Hochberg Roy Stat Soc Ser B1995
Noble Nat Biotech 2006

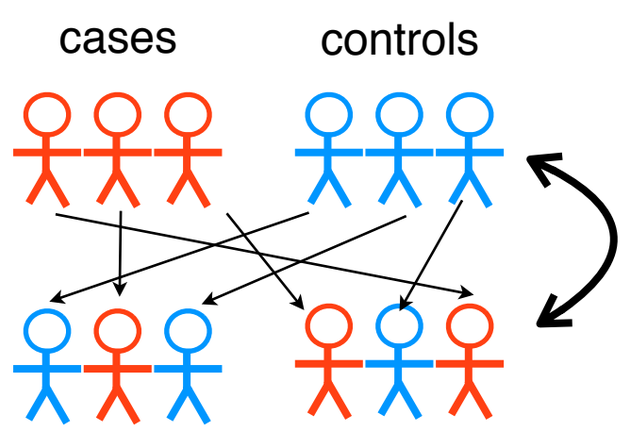
How can I compute the ***False Discovery Rate?***
empirically deriving the null distribution through permutation tests

False Discovery Rate Estimation:
The *expected* rate of false positives

$$= \frac{\text{\# false positives} \leq \alpha}{\text{\# findings} \leq \alpha} \qquad \frac{50 \text{ false positives} \leq 0.05}{100 \text{ findings} \leq 0.05} = 0.5$$

? # false positives (α)

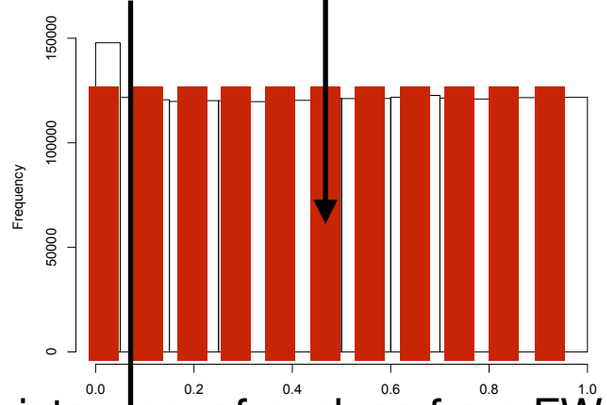
“Shuffle” (permute) disease and non-diseased participants



Repeat many times

Re-run EWAS

100 findings surpassing p ≤ 0.05



histogram of pvalues from EWAS

How can I compute the ***False Discovery Rate?***
Benjamini-Hochberg "step-up" method

first, choose a threshold, q

Benjamini and Hochberg (1995) showed that when the test statistics are independent the following procedure controls the FDR at level $q \cdot m_0/m \leq q$.

Next, sort p-values

THE BENJAMINI HOCHBERG PROCEDURE. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values. Define

find *k*

$$k = \max\left\{i: p_{(i)} \leq \frac{i}{m}q\right\}, \tag{1}$$

and reject $H^0_{(1)} \cdots H^0_{(k)}$. If no such i exists, reject no hypothesis.

```
p(i) * m/i <= q
```

Benjamini and Hochberg (1995) showed that when the test statistics are independent the following procedure controls the FDR at level $q \cdot m_0/m \leq q$.

THE BENJAMINI HOCHBERG PROCEDURE. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered observed p-values. Define

$$k = \max \left\{ i\colon p_{(i)} \leq \frac{i}{m} q \right\}, \tag{1}$$

and reject $H^0_{(1)} \cdots H^0_{(k)}$. If no such i exists, reject no hypothesis.

```
p(i) * m/i <= q
```

0.01, 0.2, 0.5, 0.45, 0.06, 0.04,0.004, 0.02, 0.1, 0.07

0.5, 0.45, 0.2, 0.1, 0.07, 0.06, 0.04, 0.02, 0.01, 0.004

0.004*10/1 <= 0.05?

0.01*10/2 <= 0.05?

0.02*10/3 <= 0.05?

Any reach significance at an FDR < 0.05?

How can I compute the ***False Discovery Rate?***
Benjamini-Hochberg "step-up" method

0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04

Any here reach significance at an FDR < 0.05?

How can I compute the ***False Discovery Rate?***
Benjamini-Hochberg “step-up” method

```
p.adjust(p, ‘fdr’)
```

What does this return?

Consideration of multiplicity of modeling scenarios.

Example: *Vibration of Effects*, the empirical distribution of effect sizes due to model choice

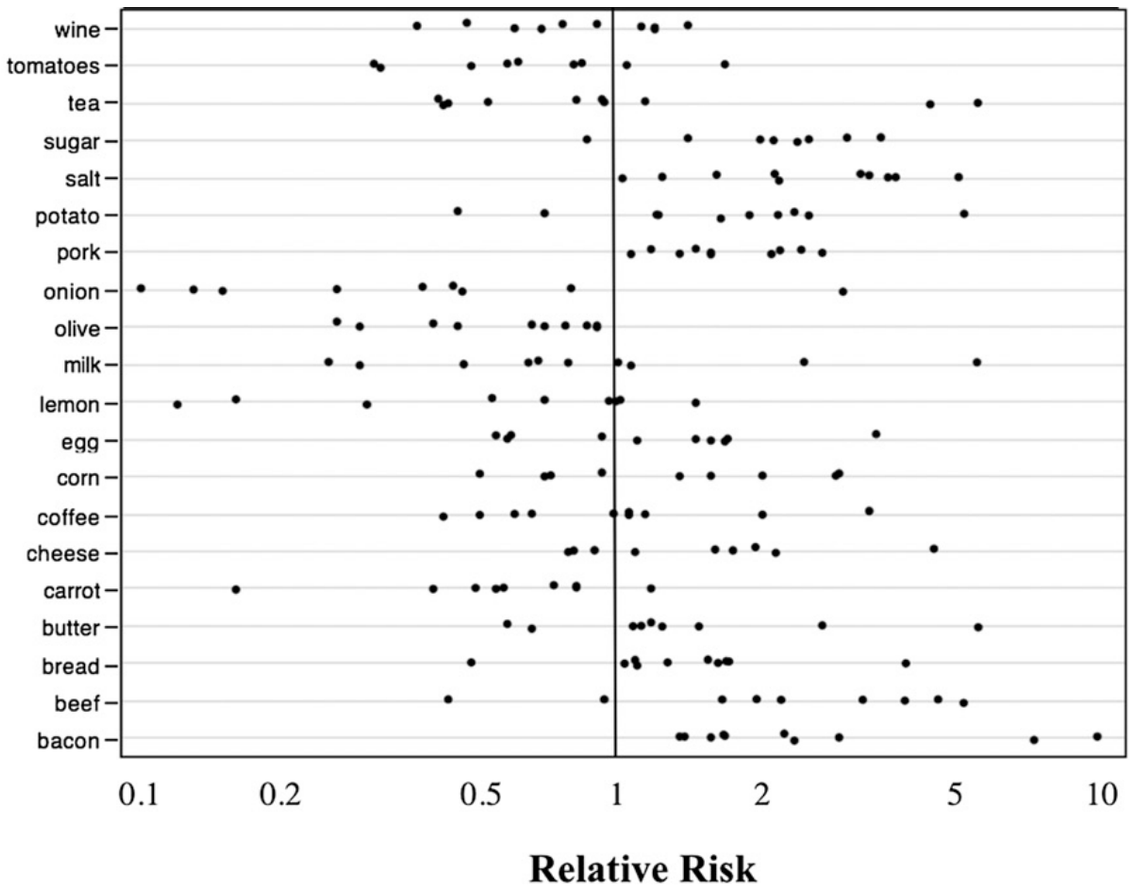
Example of *fragmentation:* Is everything we eat associated with cancer?

50 random ingredients from *Boston Cooking School Cookbook*

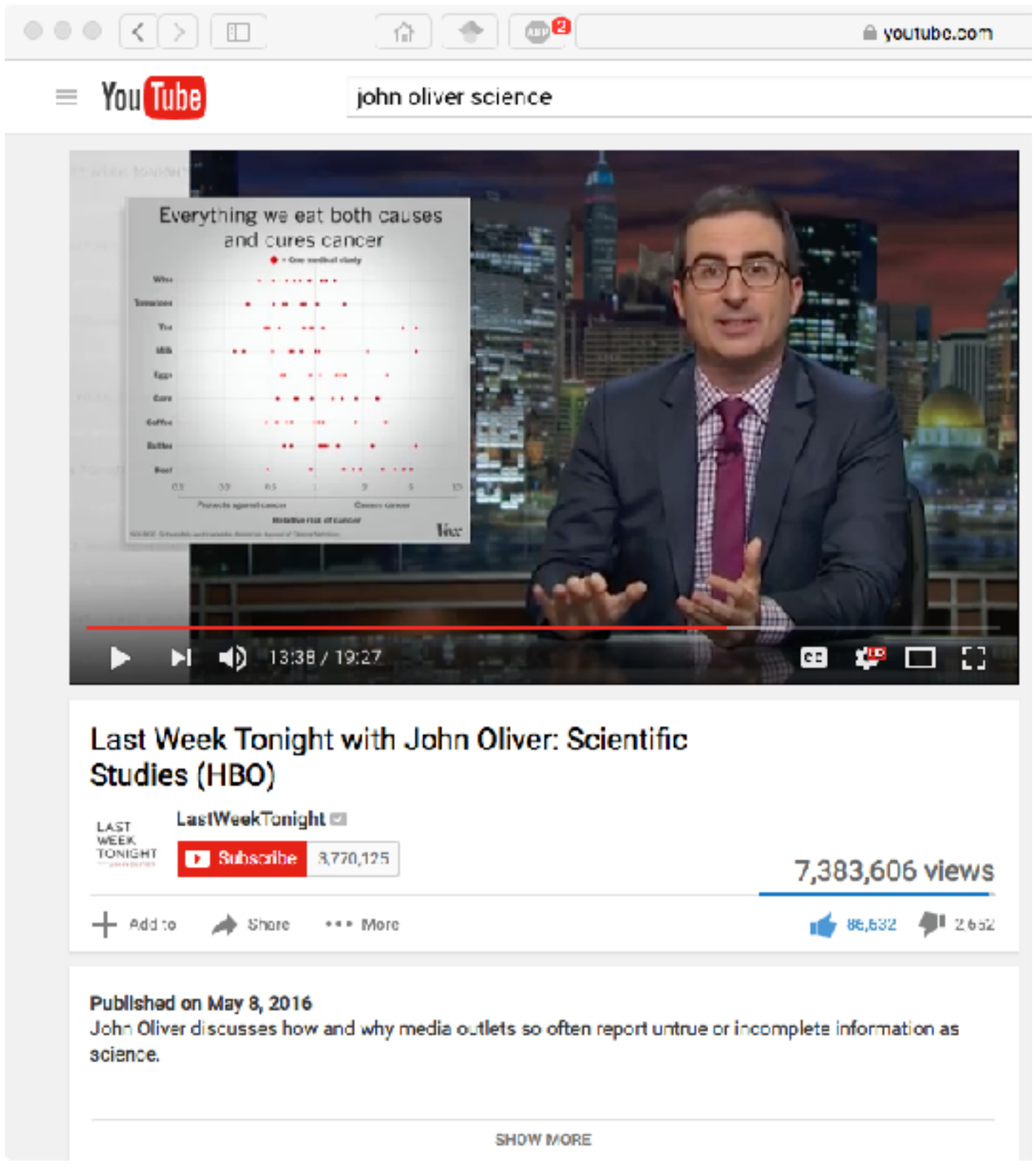
Any associated with cancer?

Of 50, 40 studied in cancer risk

Weak statistical evidence:
non-replicated
inconsistent effects
non-standardized



Schoenfeld and Ioannidis, *AJCN* 2012



https://www.youtube.com/watch?v=0Rnq1NpHdmw

A *maze* of associations is one way to a ***fragmented*** literature and ***Vibration of Effects***

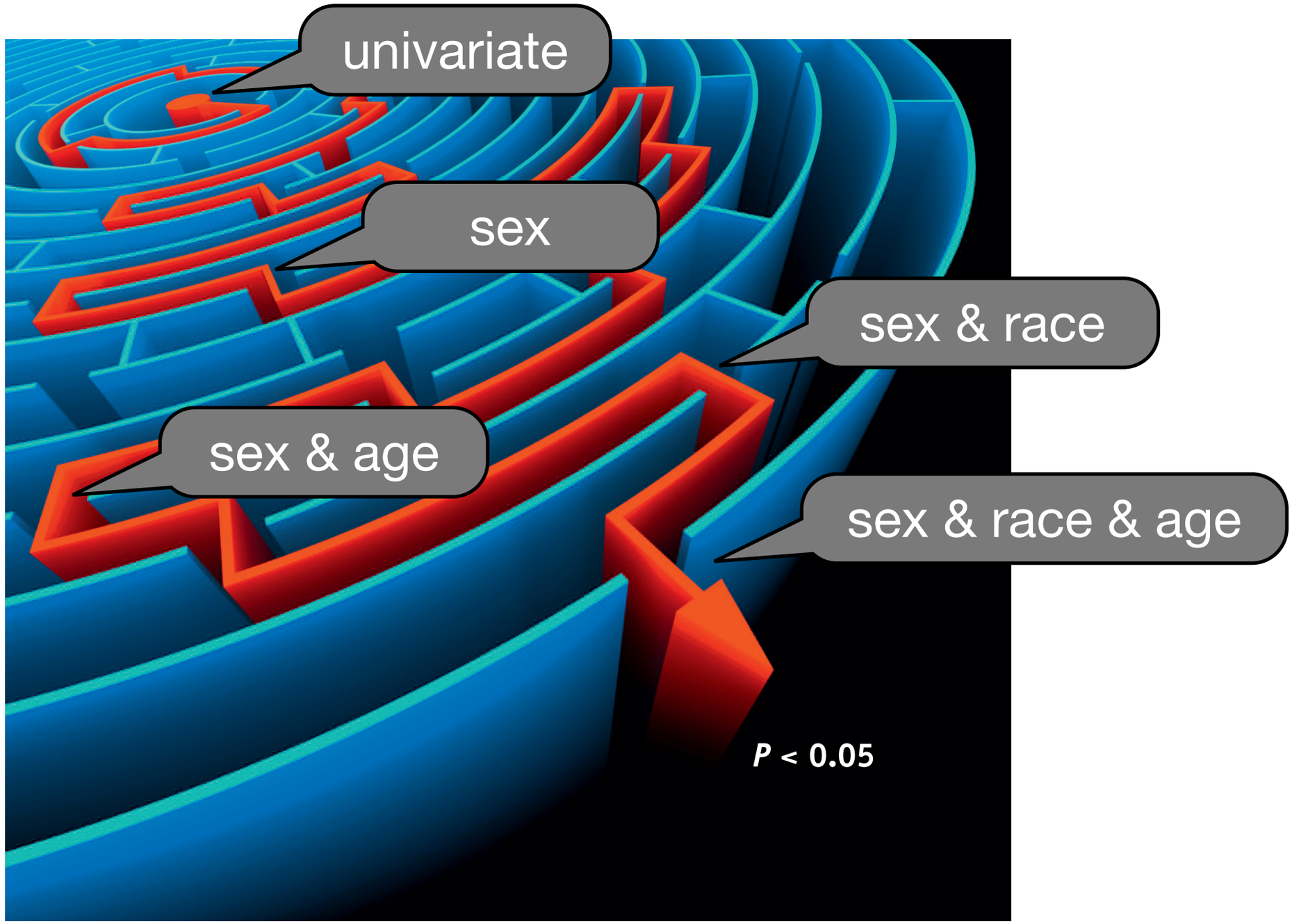


Figure 3. The path through a complex process can appear quite simple once the path is defined. Which terms are included in a multiple linear regression model? Each turn in a maze is analogous to including or not a specific term in the evolving linear model. By keeping an eye on the p-value on the term selected to be at issue, one can work towards a suitably small p-value. © ktsdesign – Fotolia

Young, 2011

JCE, 2015

Distribution of associations and p-values due to model choice: Estimating the ***Vibration of Effects (or Risk)***

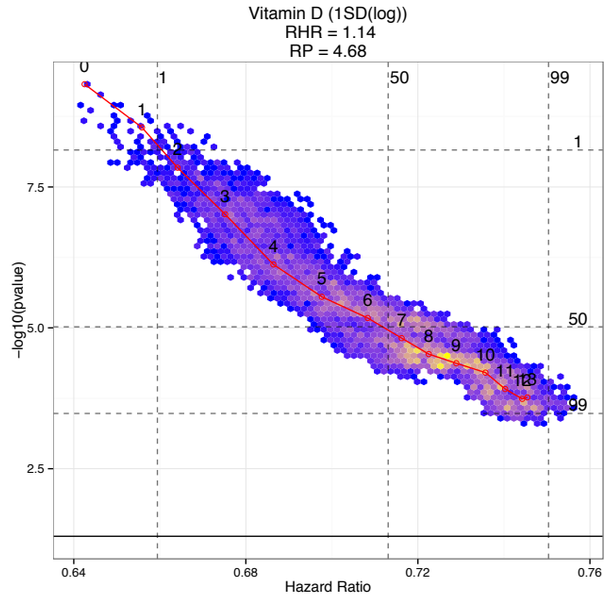
A **Data Source**
NHANES 1999-2004
417 variables of interest
time to death
N≧1000 (≧100 deaths)

B **Variable of Interest**
e.g., 1 SD of log(serum Vitamin D)

C **Adjusting Variable Set**
n=13

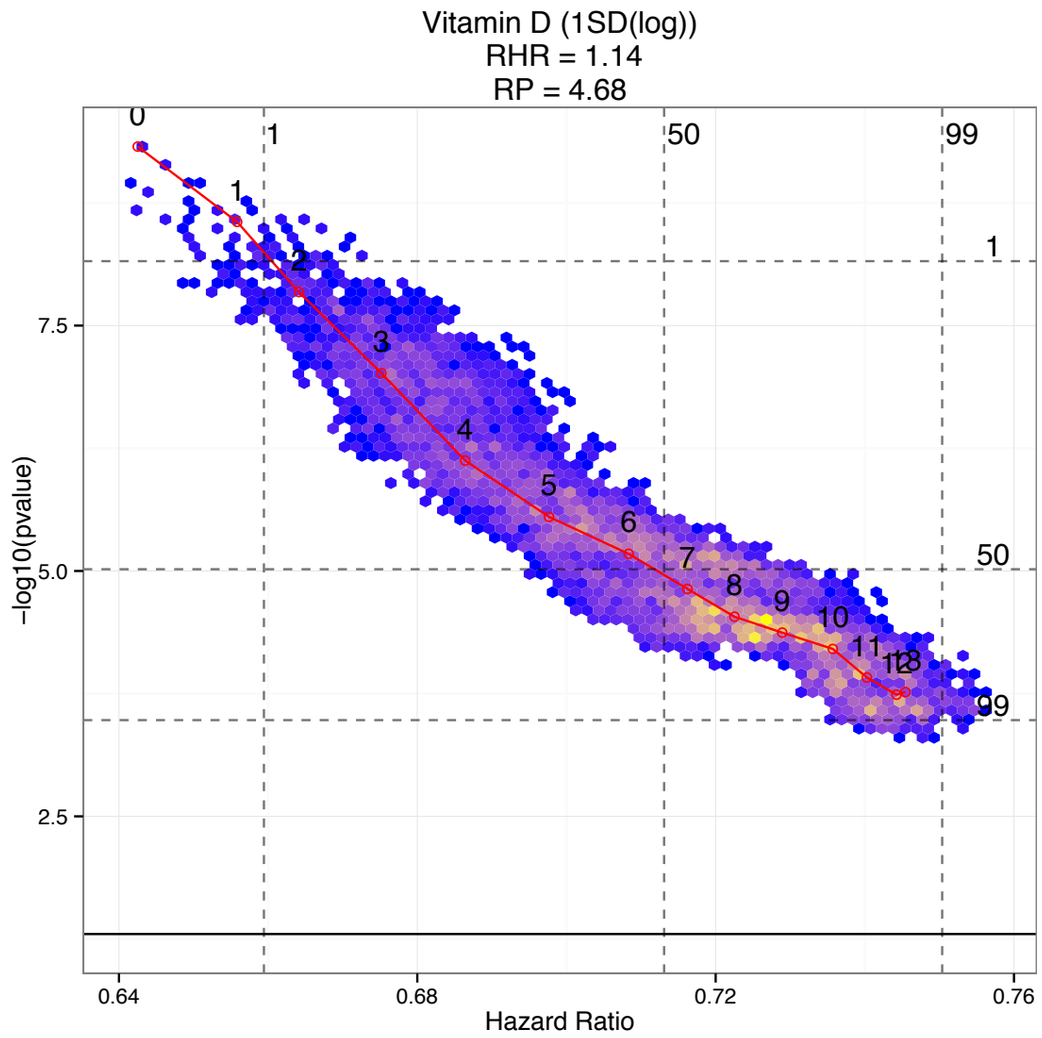
SES [3rd tertile]
education [>HS]
race [white]
body mass index [normal]
total cholesterol
any heart disease
family heart disease
any hypertension
any diabetes
any cancer
current/past smoker [no smoking]
drink 5/day
physical activity

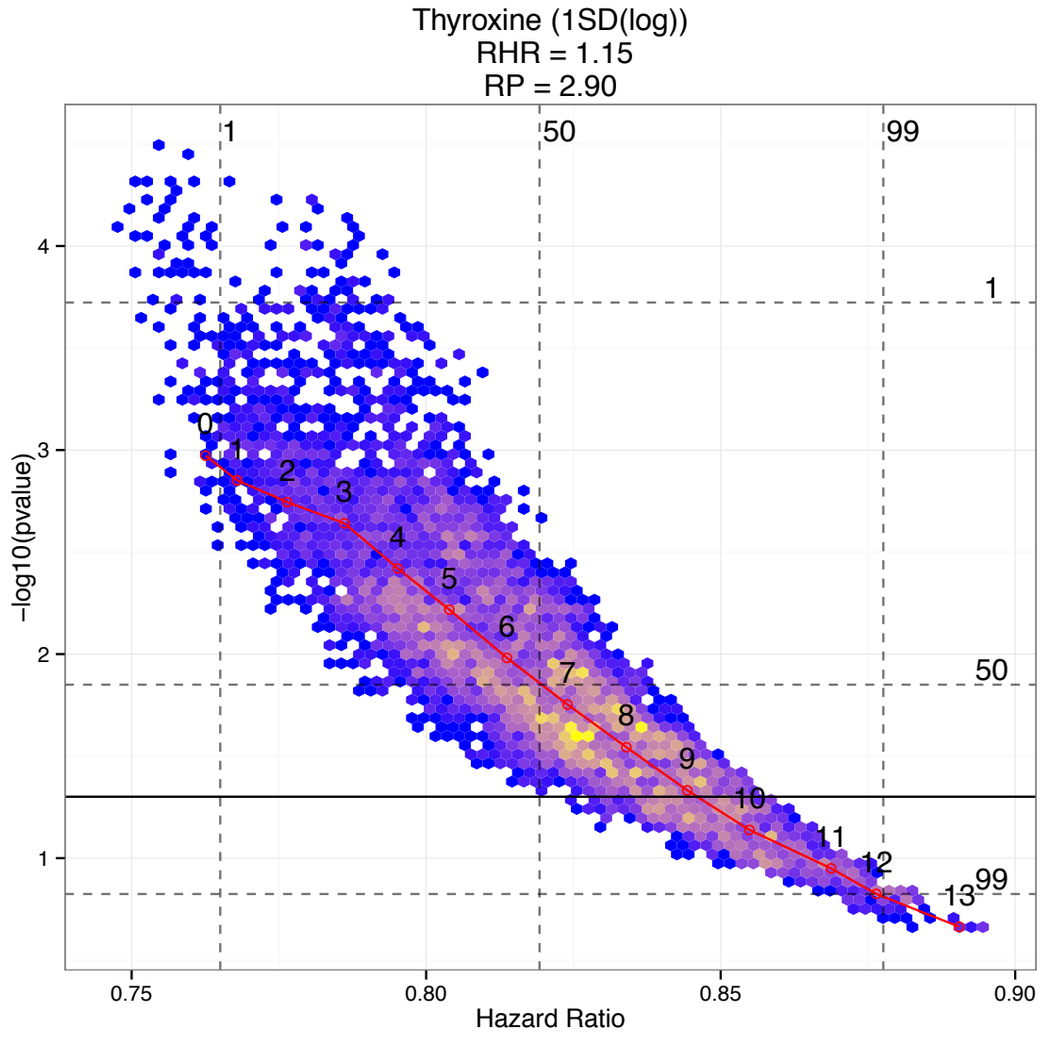
D **All-subsets Cox regression**
$2^{13}+1 = 8{,}193$ *models*



JCE, 2015

The ***Vibration of Effects:***
Vitamin D and Thyroxine and attenuated risk in mortality

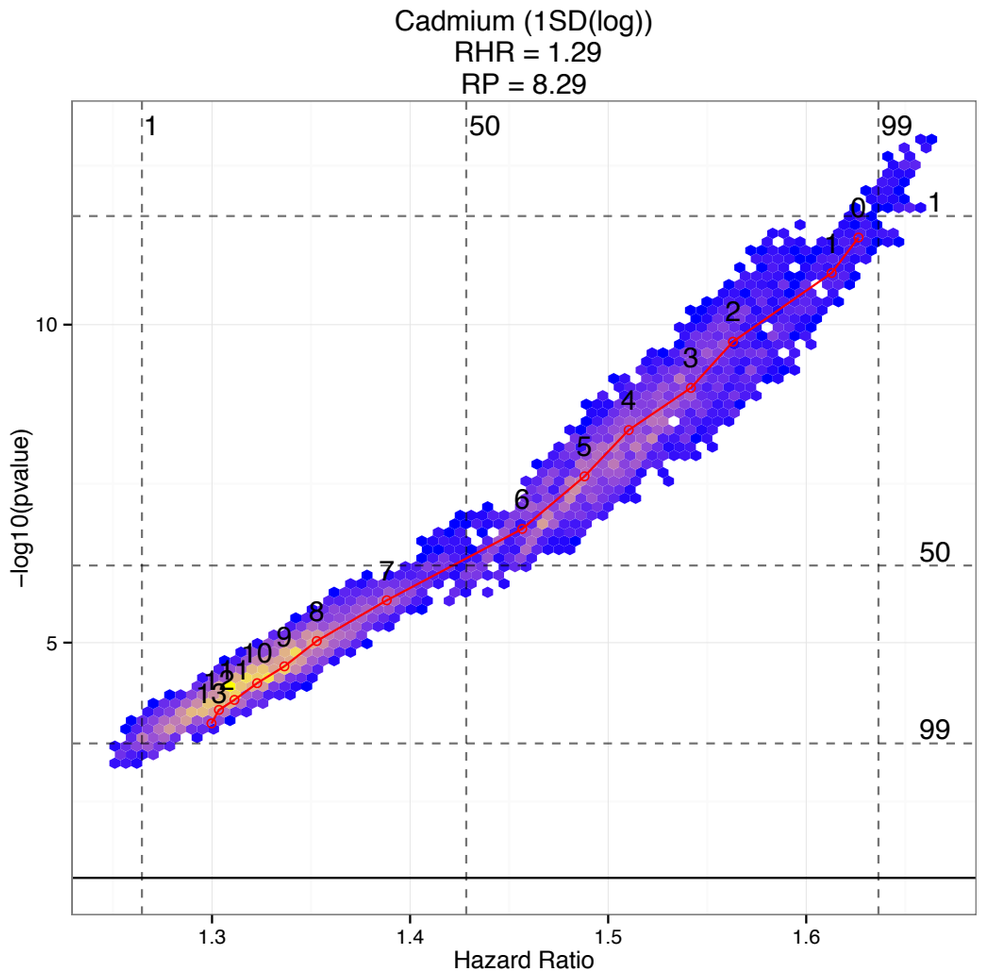


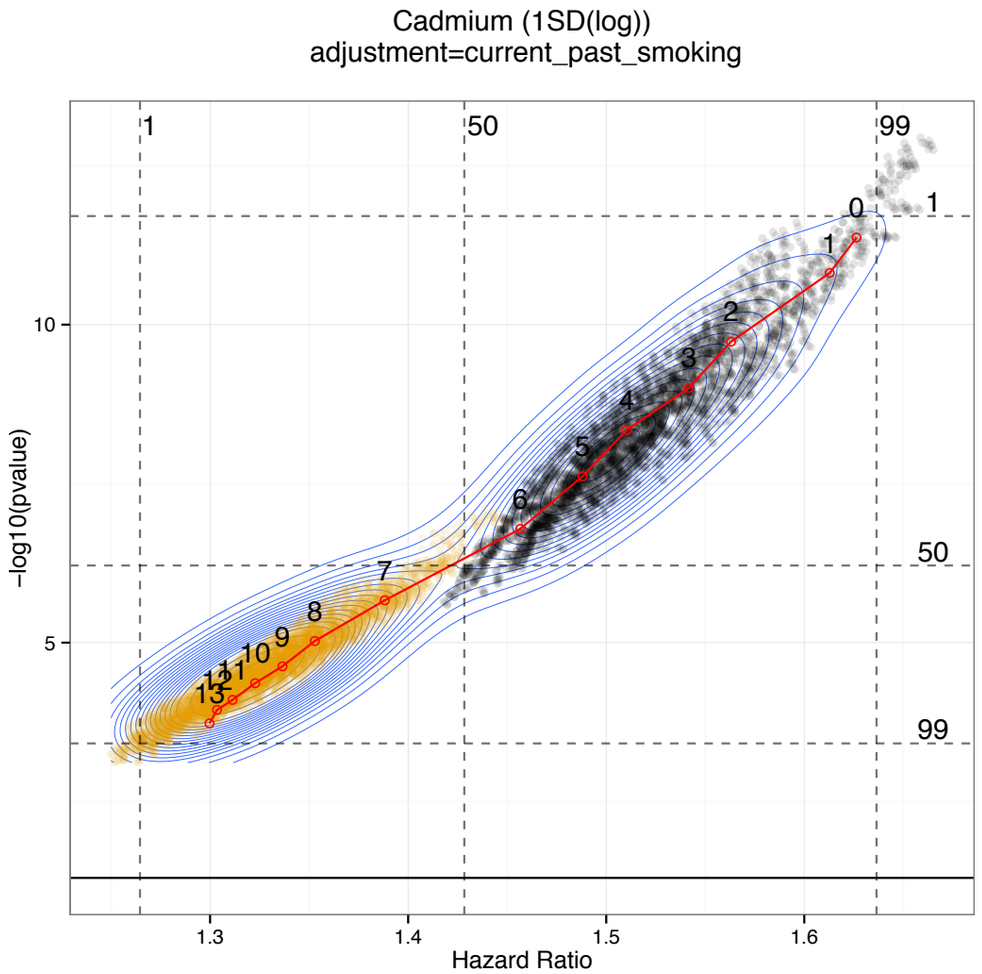


http://www.chiragjpgroup.org/voe/

JCE, 2015

The ***Vibration of Effects:*** shifts in the effect size distribution due to select adjustments (e.g., adjusting ***cadmium levels with smoking status***)



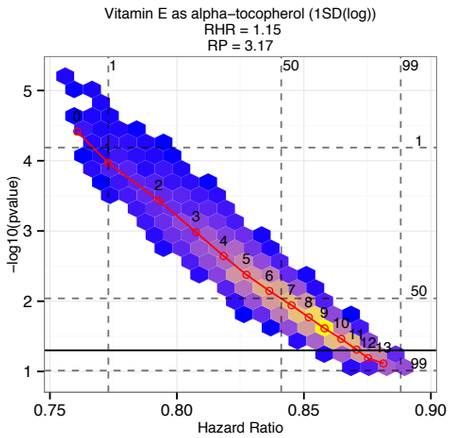


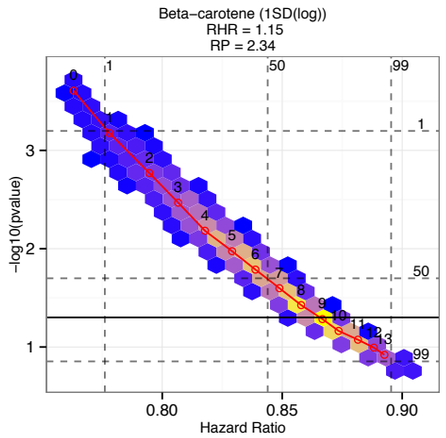
http://www.chiragjpgroup.org/voe/

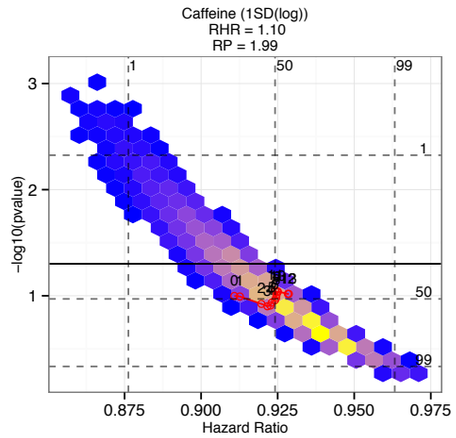
JCE, 2015

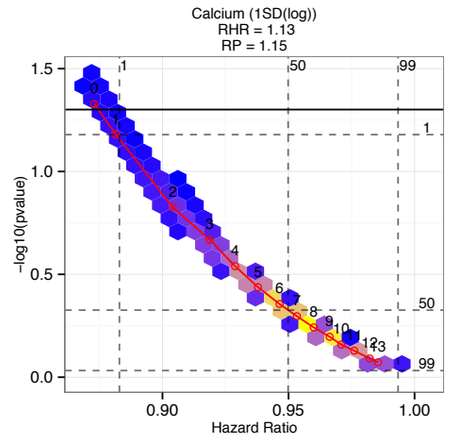
β-carotene

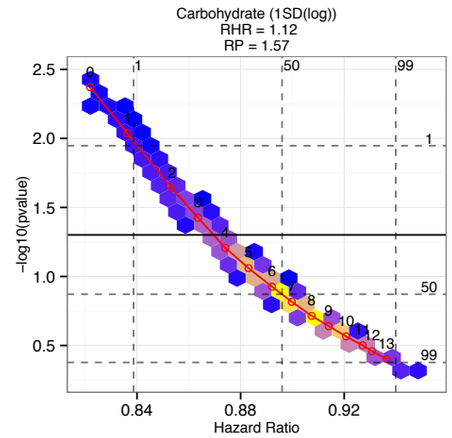
caffeine

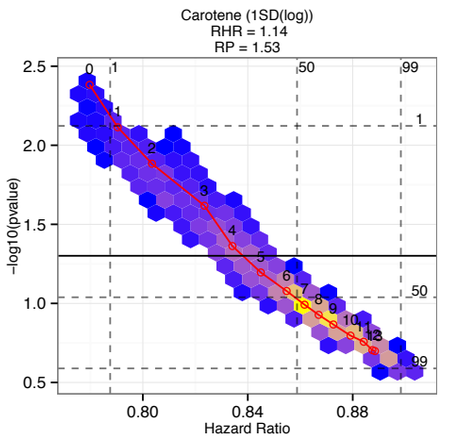




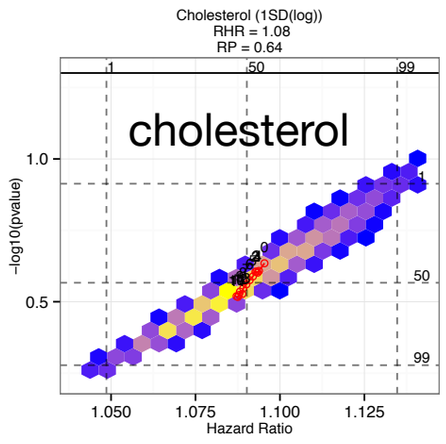


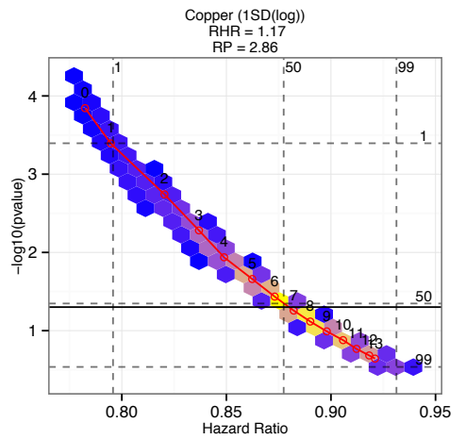


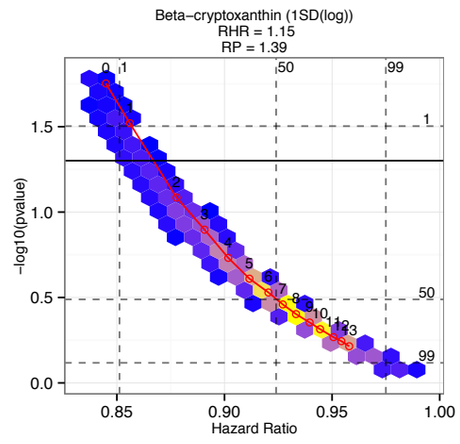


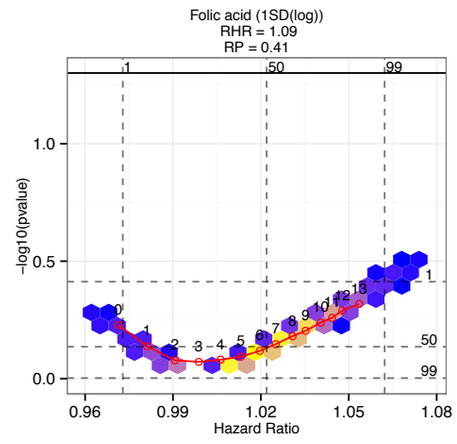


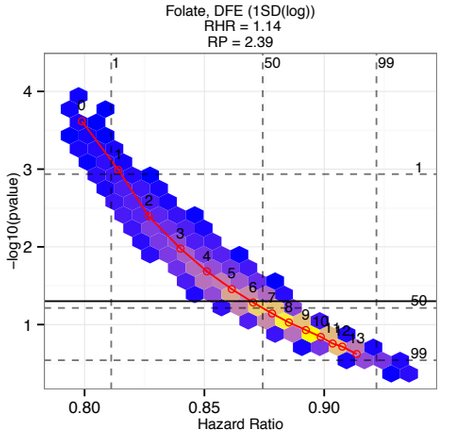
cholesterol



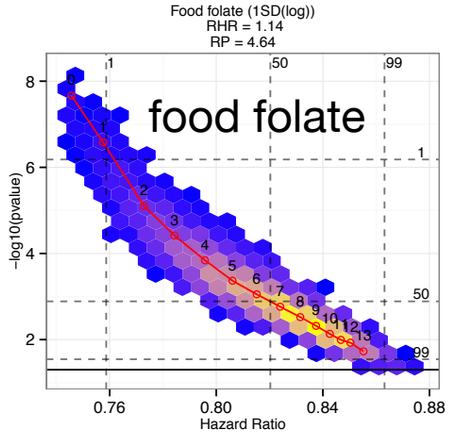


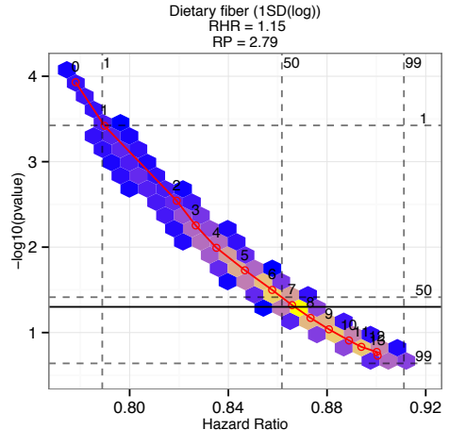


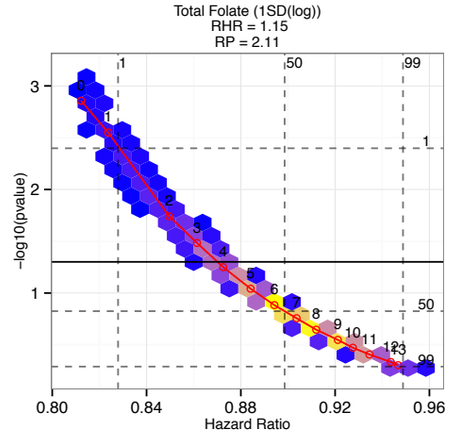


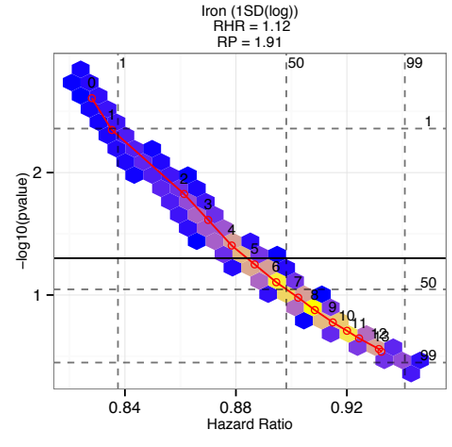


food folate

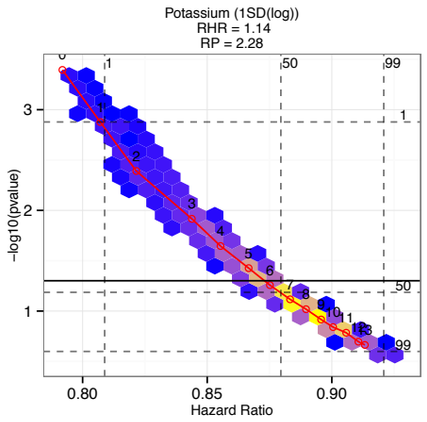


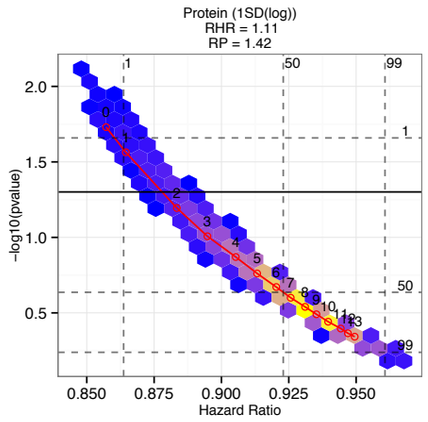


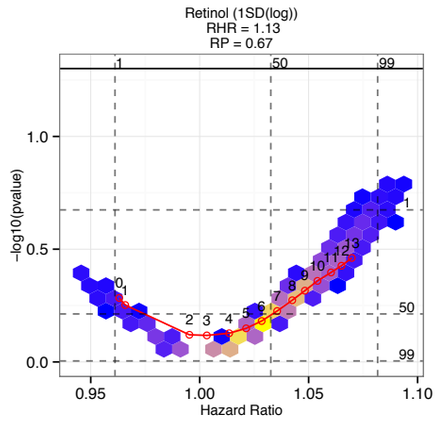


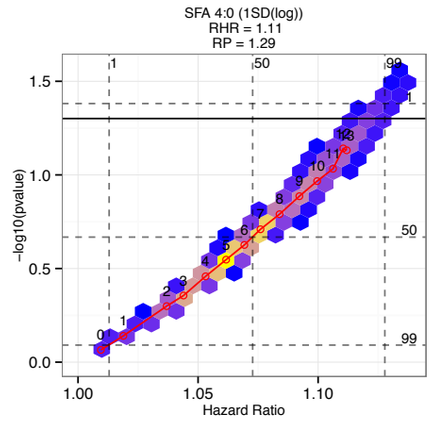


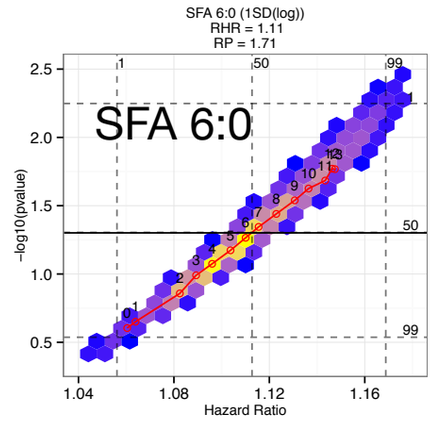
http://www.chiragjpgroup.org/voe/

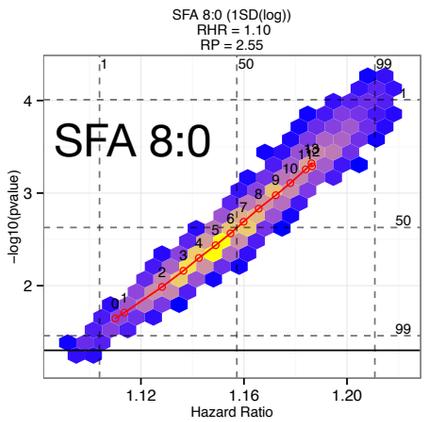


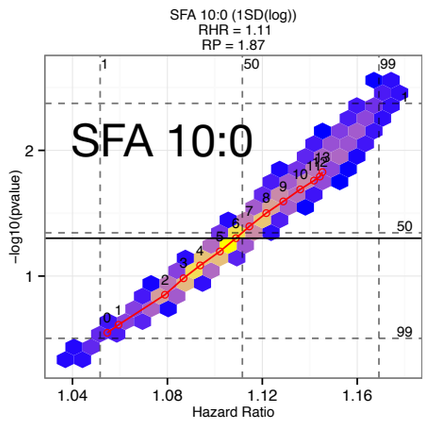


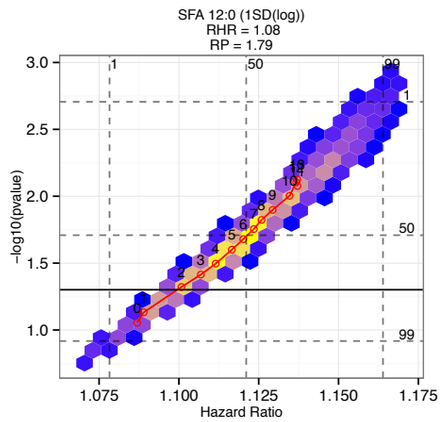


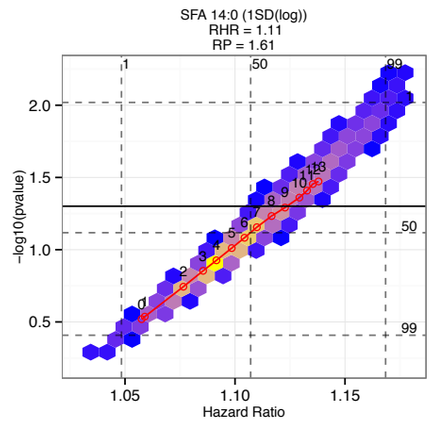


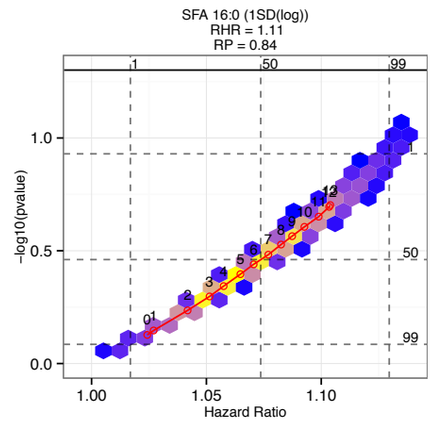


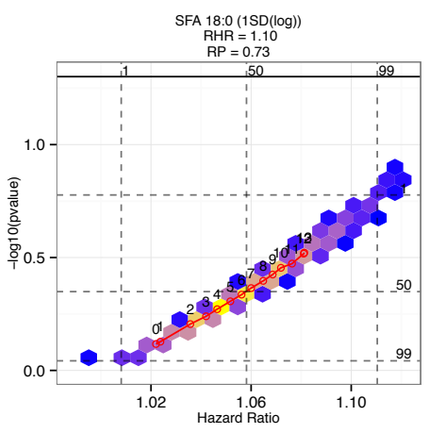


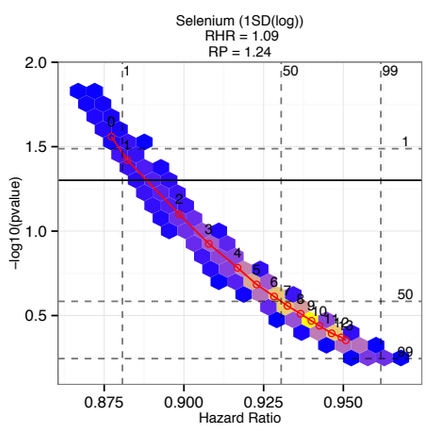


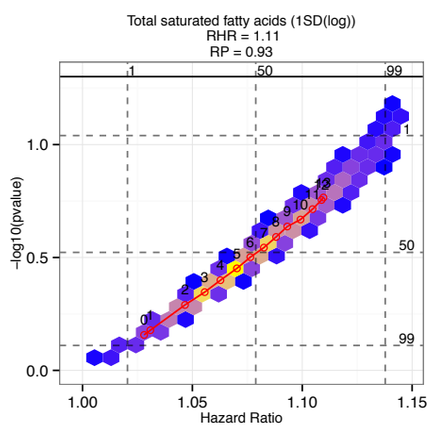


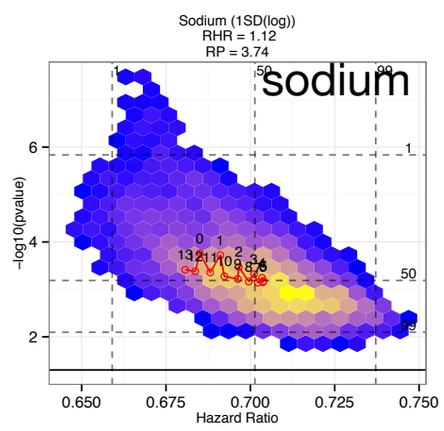


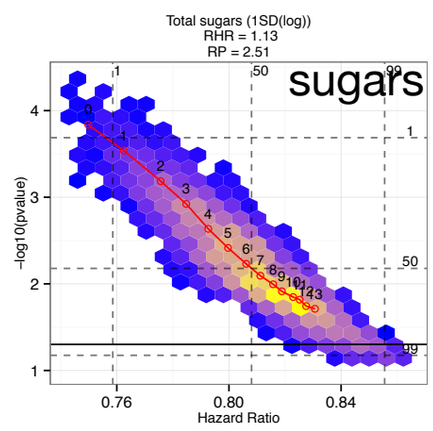


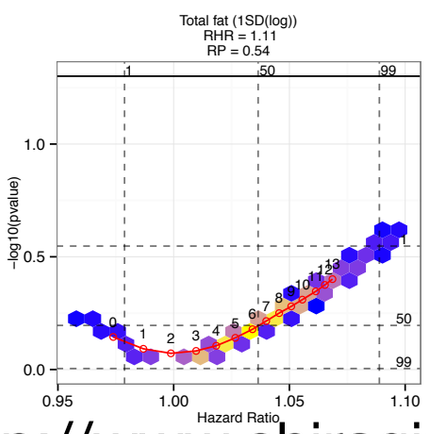


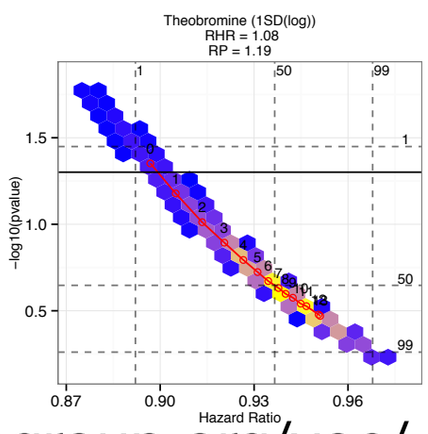


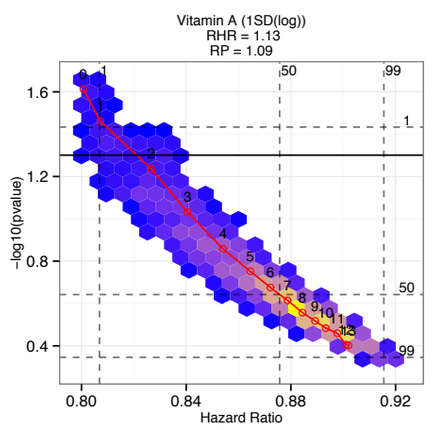


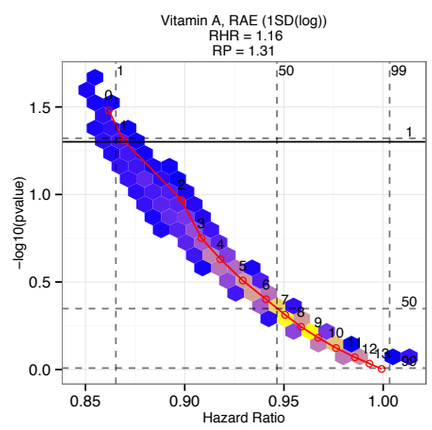


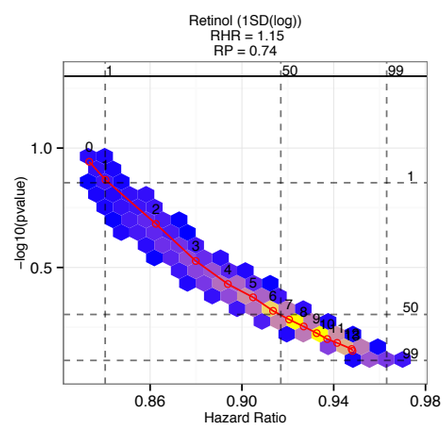




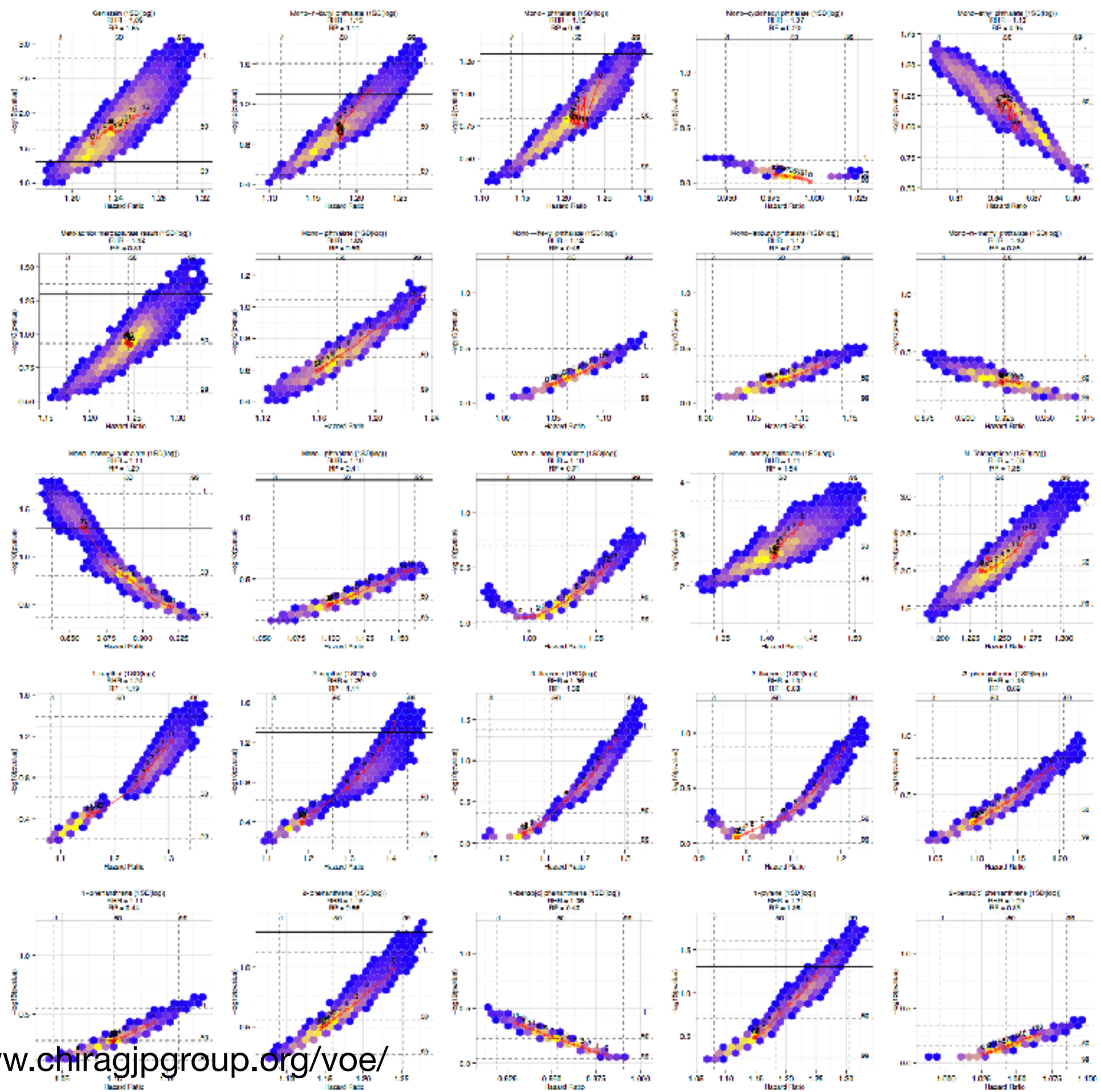




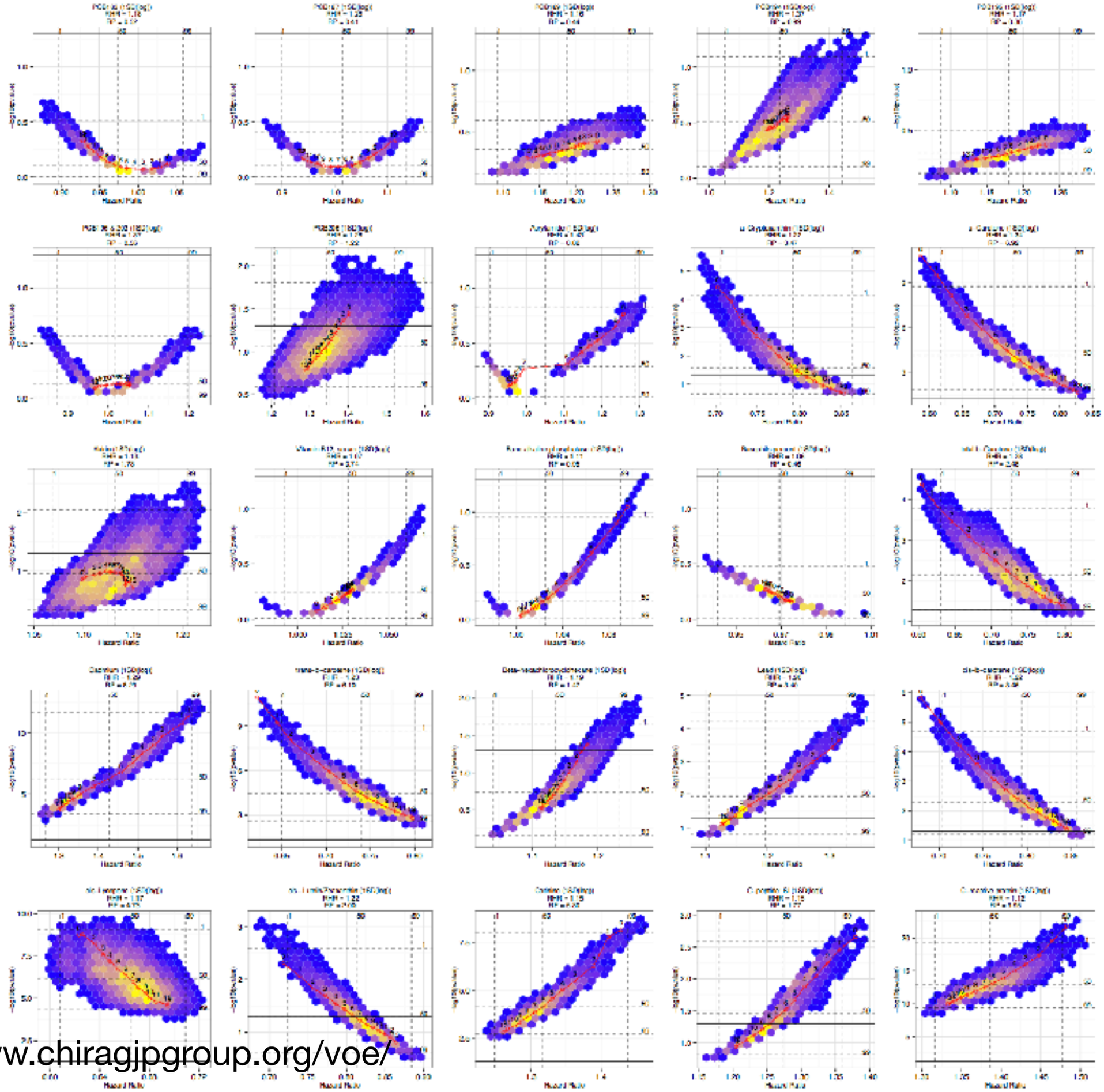




http://www.chiragjpgroup.org/voe/

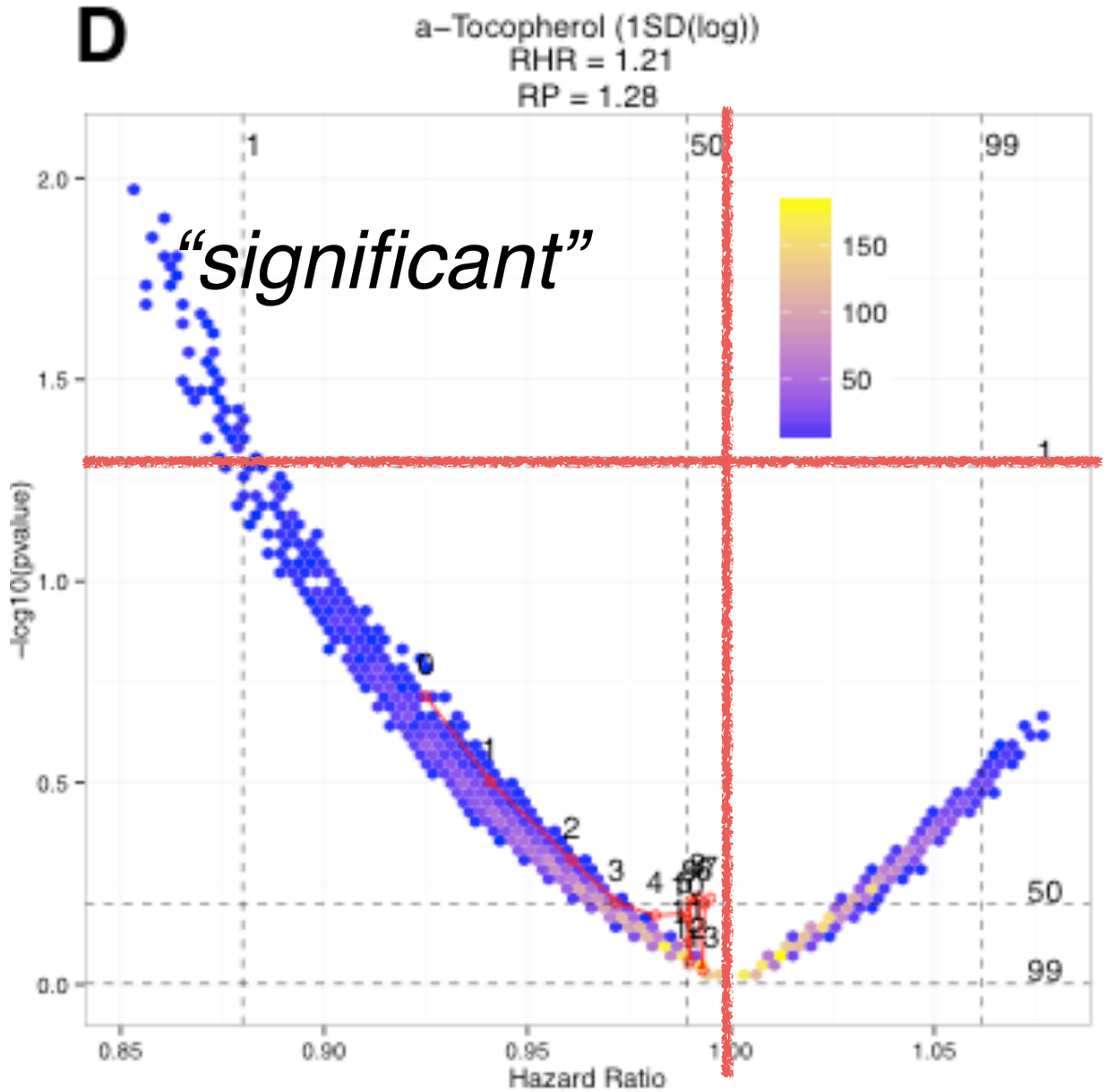


http://www.chiragjpgroup.org/voe/



http://www.chiragjpgroup.org/voe/

The ***Vibration of Effects:*** beware of the *Janus* effect (both ***risk*** *and* ***protection****?!)*



“significant”

“protection” *“risk”*

Brittanica.com

Janus (two-faced) risk profile

Risk and ***significance*** depends on modeling scenario!

http://www.chiragjpgroup.org/voe/

JCE, 2015

For the second session today…

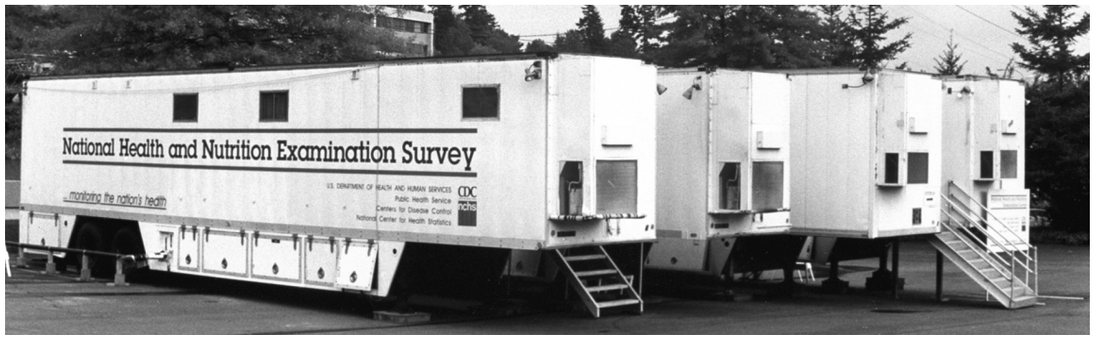
Extensible & open-source analytics software
(R *code*)

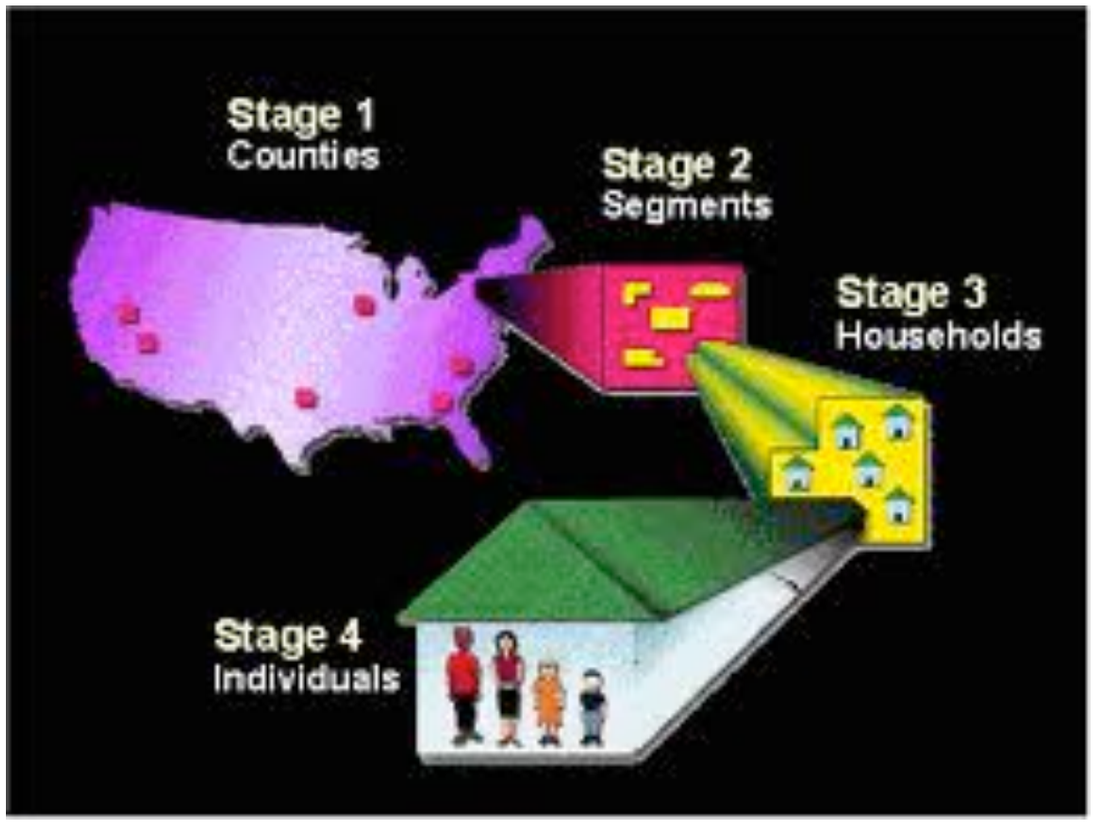
Freely available *exposome* data for your research
(NHANES: ~40,000 individuals and 1,000 variables)

Materials for teaching and demonstration

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)

>200 quantitative clinical traits (e.g., serum glucose, lipids, body mass index, telomeres)

Death index linkage (cause of death)

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

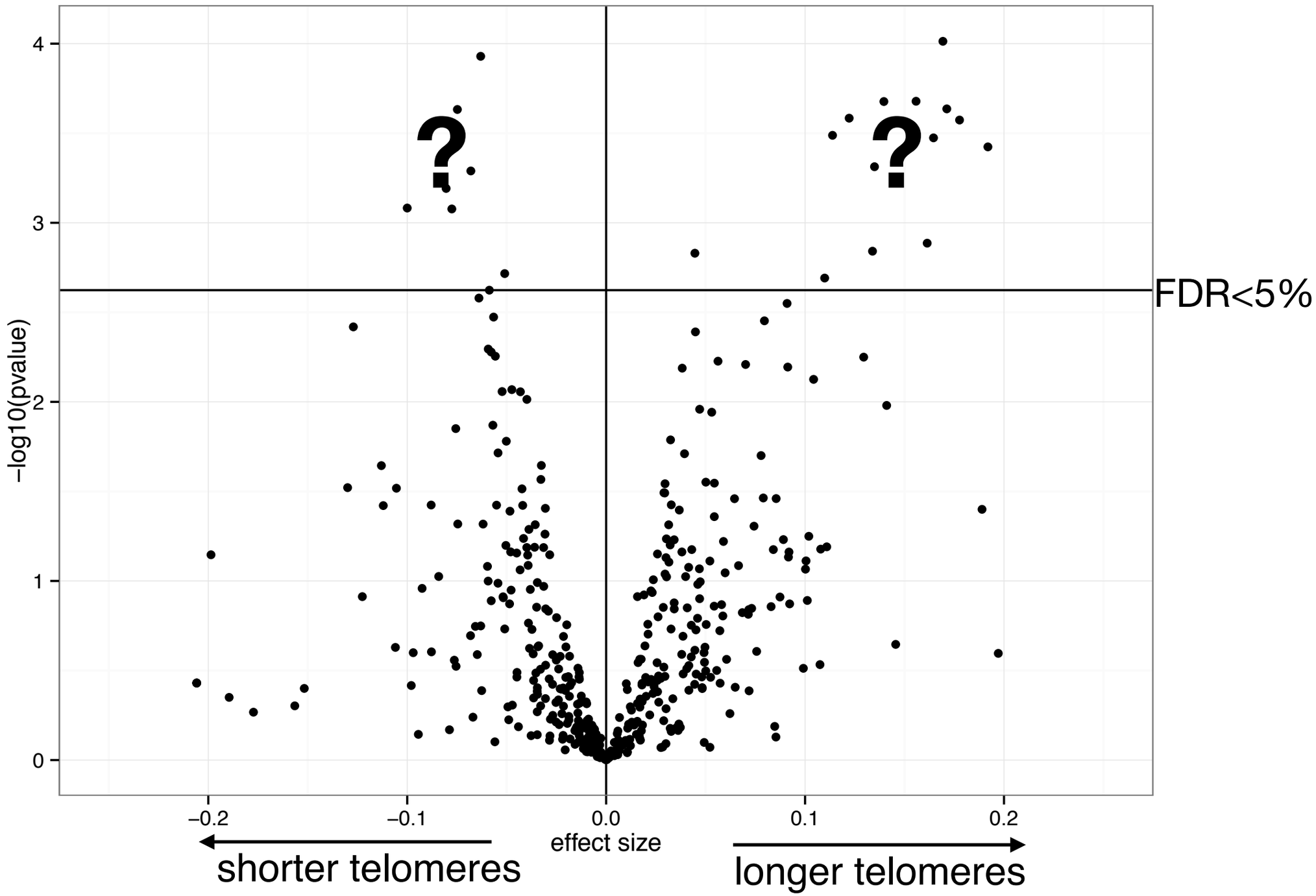
in review

13 X***WAS***-related manuscripts

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

http://correct

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



median N=3000; N range: 300-7000

IJE, 2016

Resources Index
(for today's session)

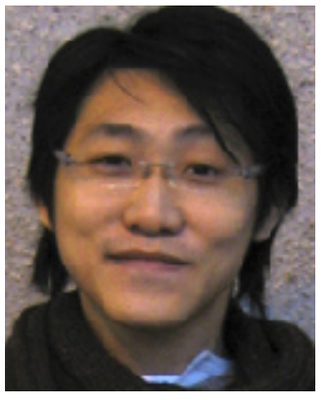
http://bit.ly/xwas_with_nhanes

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

Chirag

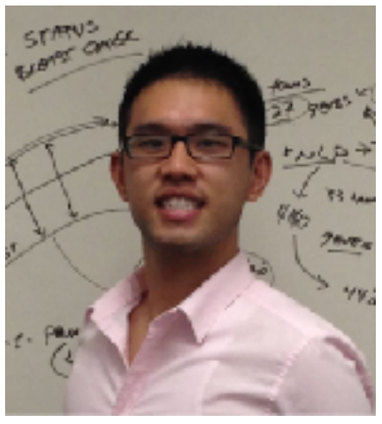
@chiragjp

Jake



@jakemkc

Nam



@nampho2

Resources Index
Papers and repositories

Papers:
https://paperpile.com/shared/PtvEae

Sample size for EWAS:
https://github.com/jakemkc/ewas_sample_size

Correlation Globes (stratified):
https://github.com/jakemkc/exposome_variability

Correlation Globes:
https://github.com/chiragjp/exposome_correlation

Vibration of Effects
https://github.com/chiragjp/voe

Exposome cohort data (NHANES 1999-2006)
https://github.com/chiragjp/nhanes_scidata

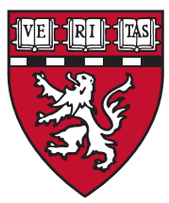
Acknowledgements

RagGroup
Nam Pho
Arjun Manrai
Jake Chung
Chirag Lakhani
Danielle Rasooly
Grace Mahoney

Harvard DBMI
Isaac Kohane

Stanford
John PA Ioannidis

NIEHS R00 ES023504
NIEHS R21 ES025052
NIAID R01 AI127250
NSF 1636870



HARVARD MEDICAL SCHOOL | DEPARTMENT OF Biomedical Informatics

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