Environment-wide Association Studies (EWAS): Creating Data-Driven Hypotheses on a Population Scale

Chirag Patel
NIEHS Mixtures Workshop
9/27/2011
chirag.patel@stanford.edu

Division of Systems Medicine
Department of Pediatrics
Stanford University School of Medicine

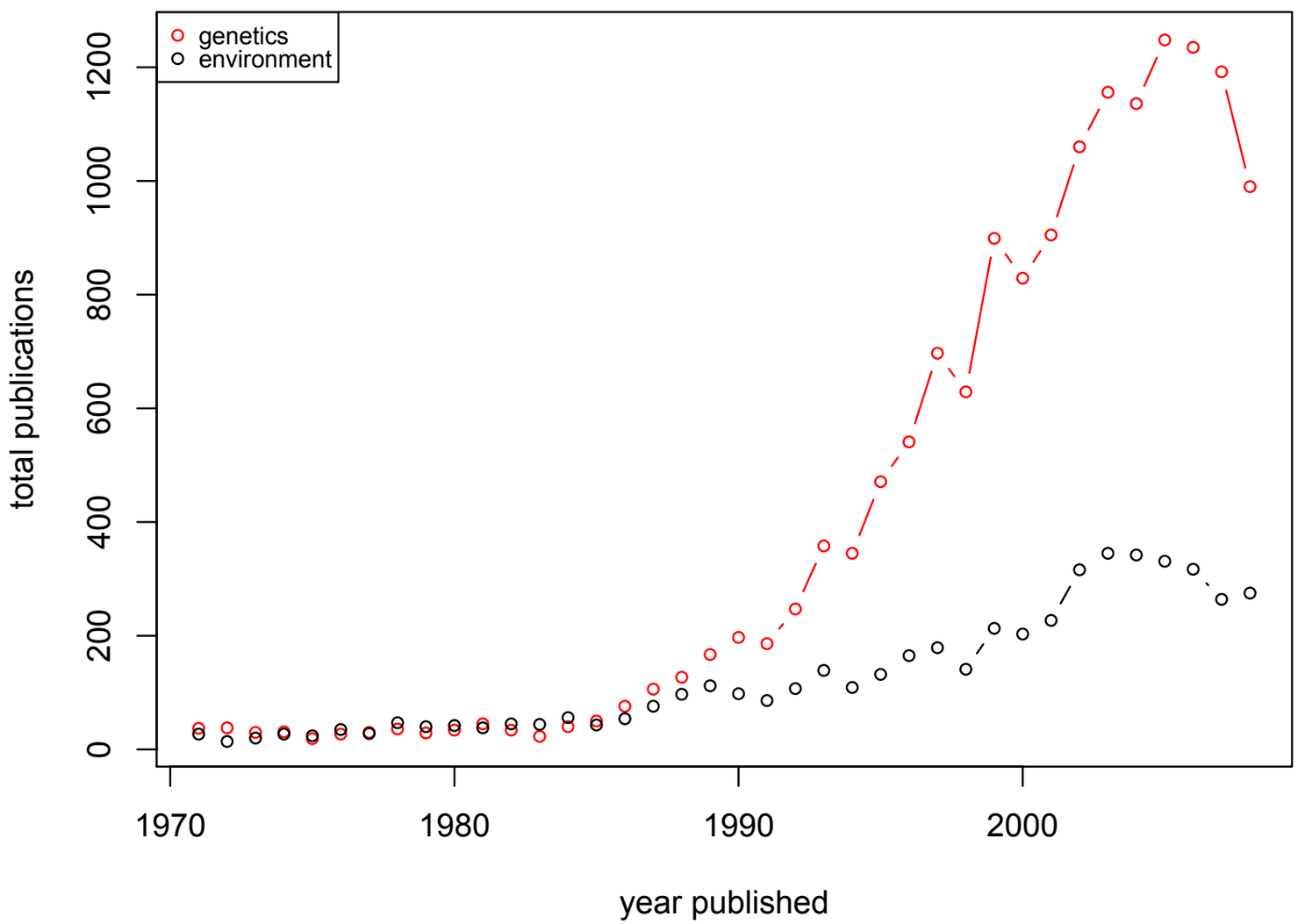
Common diseases are both a function of genes and environment, and the contribution of the environment is significant.

Heart Disease (30-50%)

Type 2 Diabetes (70%)

~2-13% of death worldwide
(World Health Organization)

Common disease is both a function of genes and environment...



WHO prioritized diseases: investigating either "genetics" or "environment" in MEDLINE

type 2 diabetes
cardiovascular disease
kidney disease
lung, colon, and prostate cancer
asthma
COPD
preterm births
Alzheimer Disease

... yet the target of investigation is biased toward genetics!

Human Genome Project to *GWAS*

Sequencing of the genome

2001

Characterize common variation

HapMap project:
http://hapmap.ncbi.nlm.nih.gov/

2001-2003 (ongoing)

Measurement tools

"Variant SNP chip"
~$400 for ~100,000 variants

~2003 (ongoing)

Comprehensive, high-throughput analyses

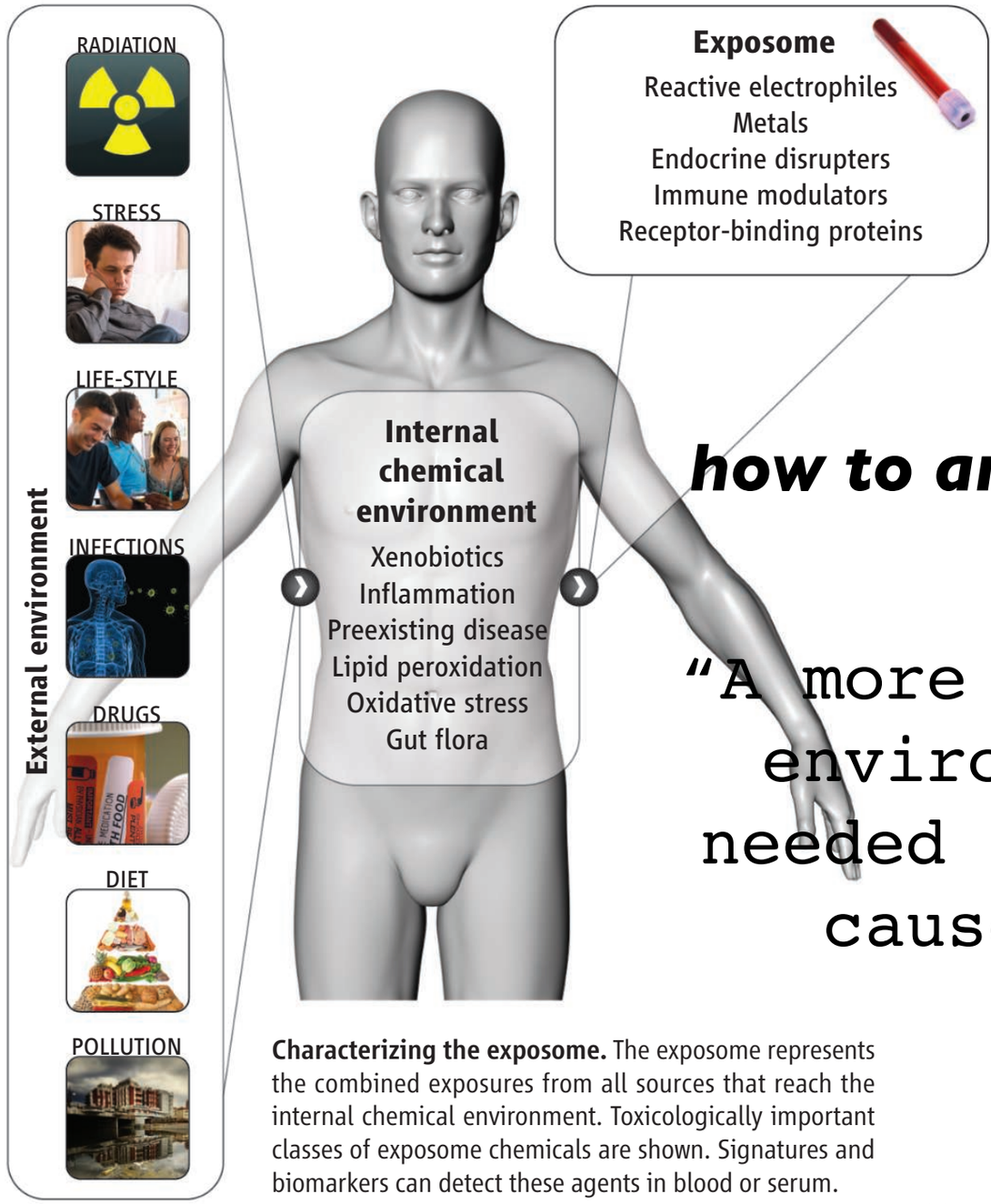
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

WTCCC, Nature, 2008.

Human *Exposome*[1] Project to "E"WAS?

what to measure?



Characterizing the exposome. The exposome represents the combined exposures from all sources that reach the internal chemical environment. Toxicologically important classes of exposome chemicals are shown. Signatures and biomarkers can detect these agents in blood or serum.

how to measure?

how to analyze in relation to health?

"A more comprehensive view of environmental exposure is needed ... to discover major causes of diseases..."

bl d

1. Rappaport S, Smith M. Environment and Disease Risks. Science (2010) vol. 330 (6003) pp. 460-461
2. Anthony, J.C., The promise of psychiatric enviromics. Br J Psychiatry Suppl. (2001) 40: p. s8-11.

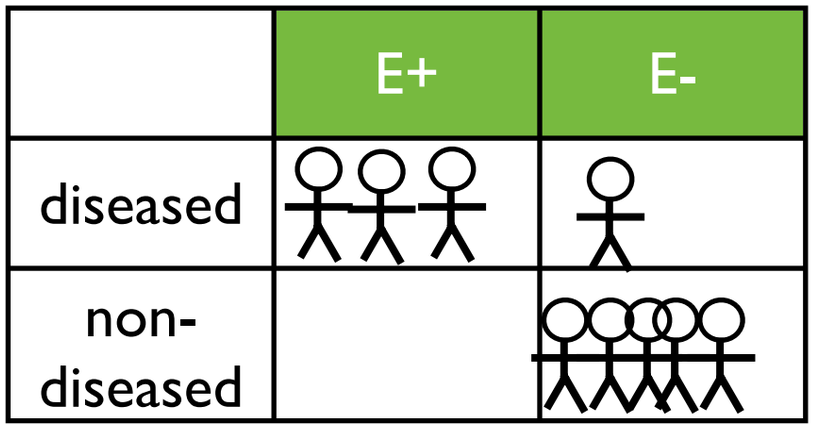
Hypothesis
Applying genome-based methods to the environment

We claim that comprehensive connection of environmental factors to disease is practicable using high-throughput analysis methods, now common in genome-based investigations.

Specific Aims:
EWAS for hypothesis generation

1. Background and Methods
2. Examples: Type 2 Diabetes, Serum Lipid Levels
3. Checking Validity
4. An "LD" Map of the *Exposome*?

Epidemiology Drawbacks



"candidate" E factors

multiple hypotheses often ignored

selective reporting

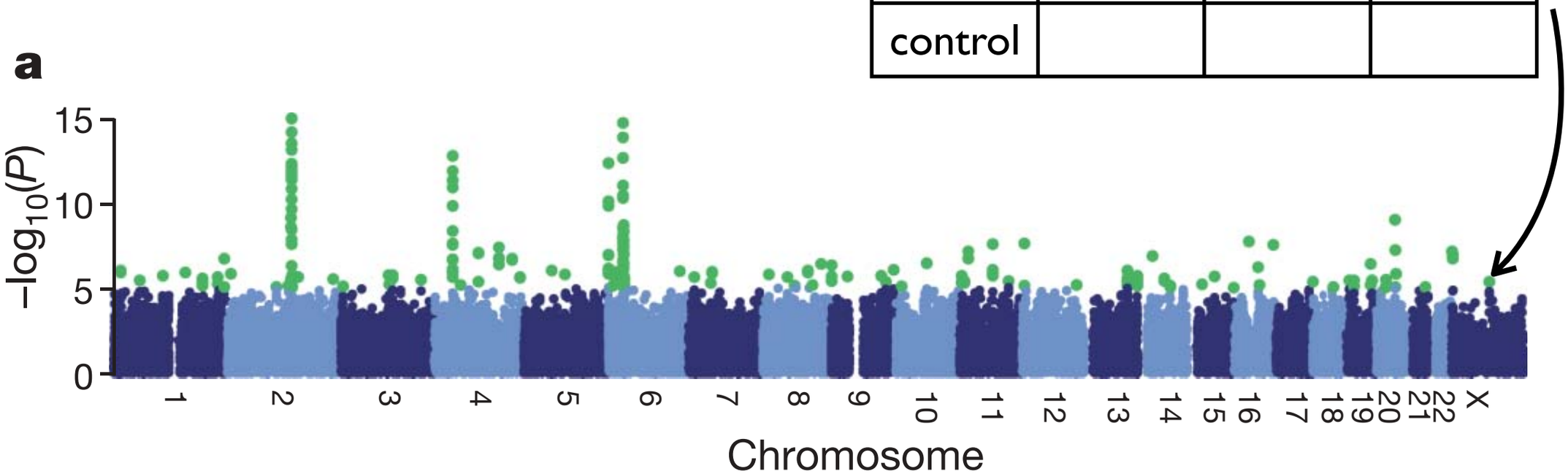
The lack of comprehension has led to a fragmented literature of environmental associations[1,2,3,4]

Genome-wide epidemiology has overcome some of these drawbacks[1]

1. Ioannidis et al. Science Translational Medicine, 2009. 1 (7) p. 6
2. Boffetta, P., et al.,. J Natl Cancer Inst, 2008. 100(14): p. 988-95.
3. Young, S.S., Int J Epidemiol, 2010. **39**(3): p. 934
4. Taubes, G.. Science, 1995. 269(5221): p. 164-9.

Aim 1: EWAS Methods

	AA	Aa	aa
case			
control			

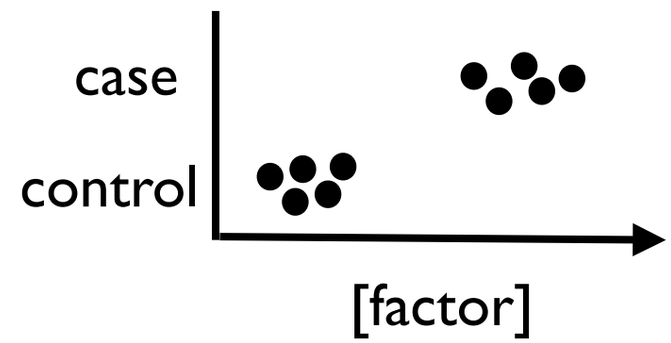


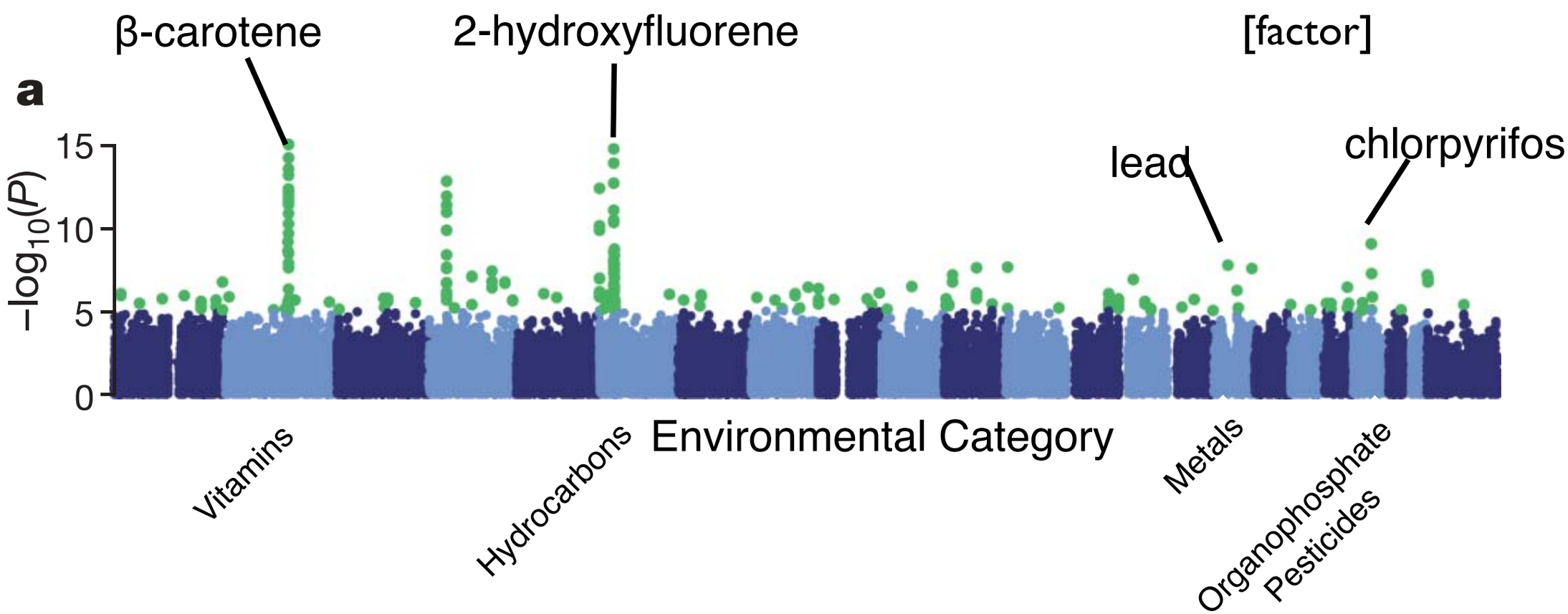
~100,000 - 1,000,000 association tests

What genetic loci are associated to disease?

WTCCC, 2007

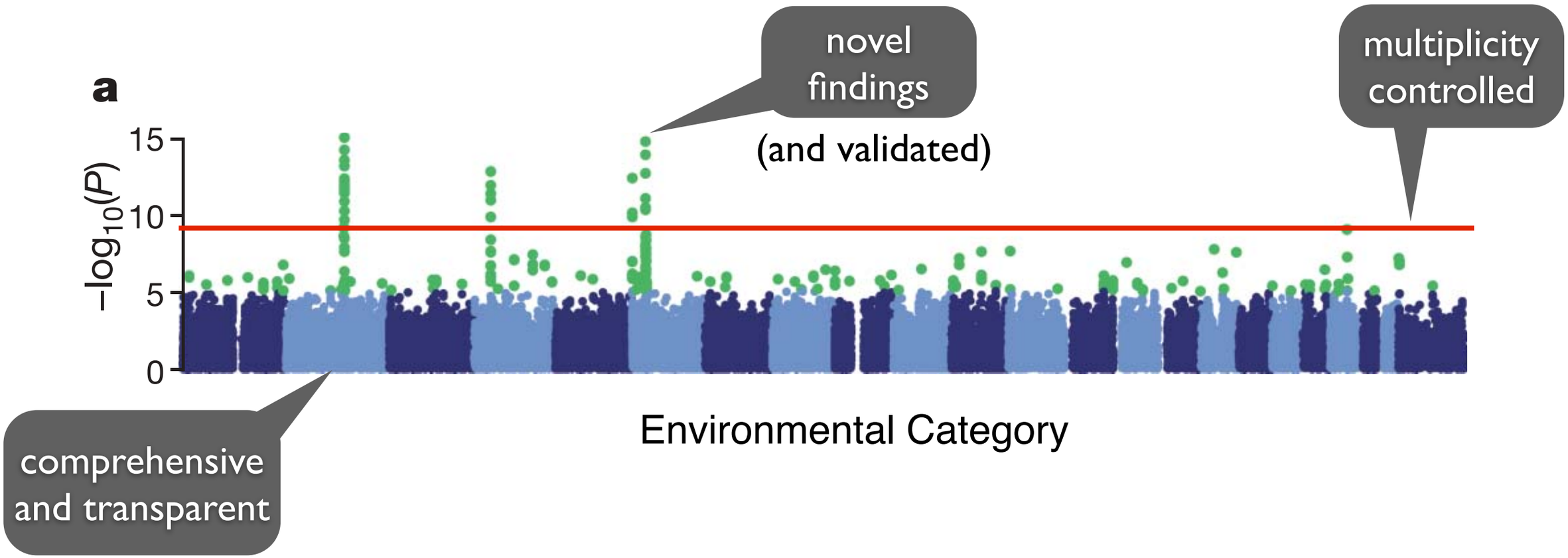
Environment-Wide Association Studies (EWAS)





What specific environmental "loci" are associated to disease?:
ie, T2D, lipid levels, obesity, etc?

Why "EWAS"?

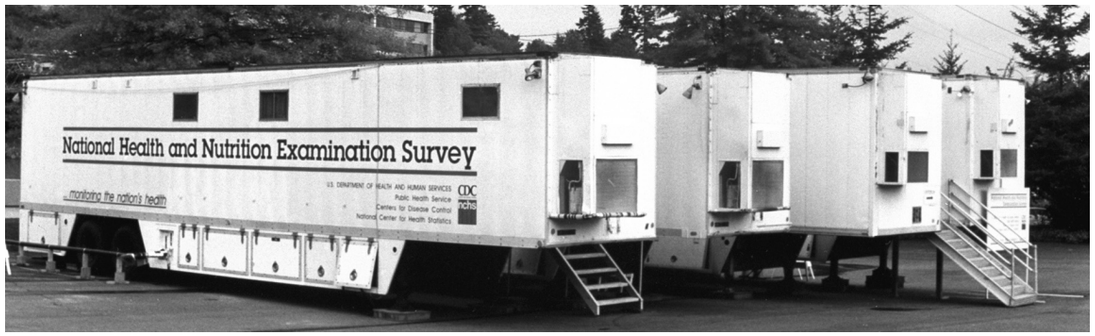


What environmental factors are associated to disease?

Ioannidis JPAI, Loy EY, et al, (2009) Researching genetic vs nongenetic determinants of diseases: a comparison and proposed unification. *Sci Trans Med* vol. 1(7)7ps8

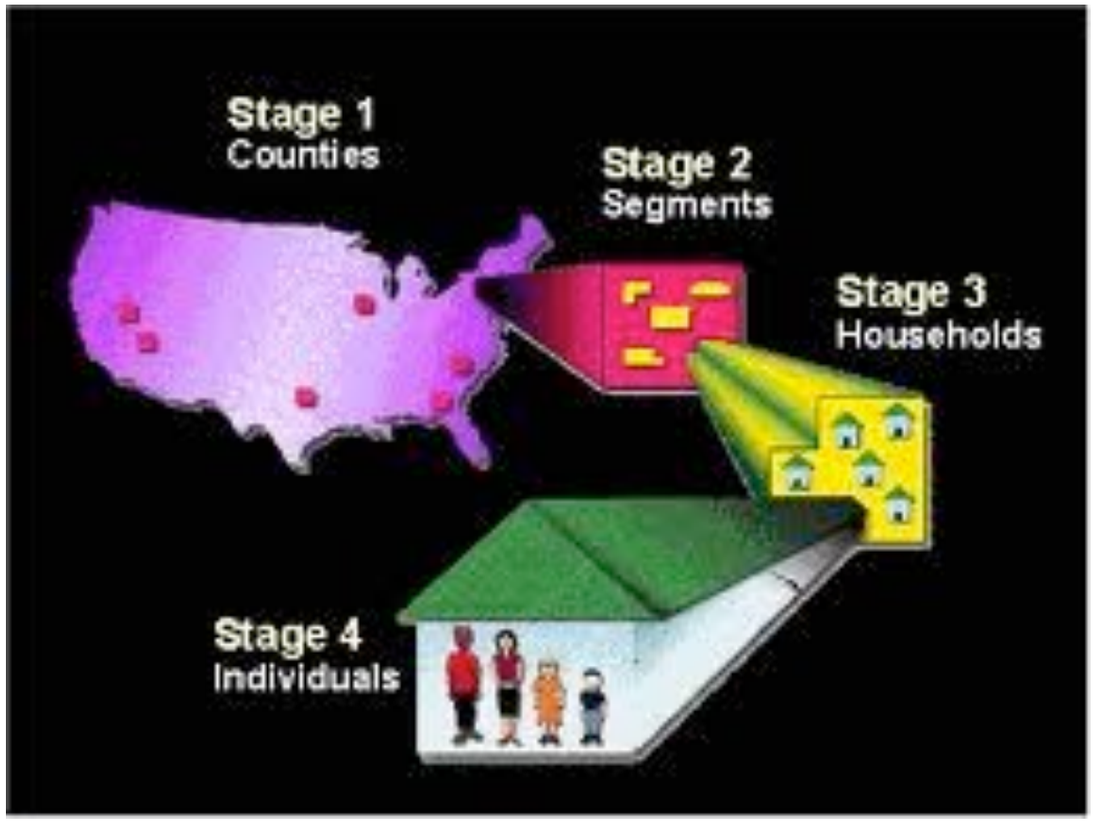
NHANES:
National Health and Nutrition Examination Survey[1]

since the 1960s: 50 years!
now biannual: 1999 onwards
10,000 participants per cohort



disease prevalence estimates:
T2D, obesity, cardiovascular disease

growth charts for development:
WHO standard



environment:
elimination of lead --
70% decline since '70s

A Massive, Ongoing, and Significant Public Health Survey

1 http://www.cdc.gov/nchs/nhanes.htm

NHANES
Environmental Factor "E-Chip"

Demographics (N ~=10,000)	Examination (N ~= 3000)
Age Sex Income Education Ethnicity	Blood Pressure Body Measurement Vision Oral Health

physician-led assessment

Laboratory (N ~= 3000)	Questionnaire (N ~=10,000)
Clinical Measures: Triglycerides, cholesterol, glucose **Exposure Markers**: Heavy metals, dioxins, PCBs, phenols, phthalates; Infectious Diseases Allergens	Disease & Health Status Drug use Physical Activity Health & Fitness History Occupation

biomarkers measured in blood and urine

self-report; verified where possible

One of a kind dataset that is representative of the US population

Environmental Measures by Category and Cohort

environmental categories

	1999-2000	2001-2002	2003-2004	2005-2006	
Acrylamide	0	0	2	0	
Allergen Test	0	0	0	20	IgE (cat, dog, milk ragweed...)
Bacterial	8	13	17	1	Staph. aureus, gonorrhea, chlamydia
Cotinine	1	1	1	1	
Diakyl	7	7	6	0	
Dioxins	5	7	7	0	
Furans	5	5	9	0	
Heavy Metals	18	18	23	25	lead, cadmium, arsenic
Hydrocarbons	14	22	21	0	
Latex	1	0	0	0	
Carotenoid Nutrients	0	6	15	7	carotenes, lutein/zeaxanthin
Mineral Nutrients	2	2	2	1	
Vitamin A	3	3	3	3	retinyl palmitate, retinol
Vitamin B	4	4	5	3	
Vitamin C	0	0	1	1	
Vitamin D	0	1	1	1	
Vitamin E	2	2	3	2	
Polychlorinated Biphenyls	23	26	38	0	
Perchlorate	0	0	2	0	
Pesticides, Atrazine	0	0	5	0	
Pesticides, Carbamate	0	0	1	0	
Pesticides, Chlorophenol	0	0	1	1	
Pesticides, Organochlorine	10	13	11	0	DDT, trans-nonachlor
Pesticides, Organophosphate	2	2	2	0	
Pesticides, Pyrethyroid	1	1	1	0	
Phenols	15	11	9	12	bisphenol A
Phthalates	7	12	12	0	
Phytoestrogens	6	6	6	0	
Polybrominated Ethers	0	0	12	0	
Polyflourochemicals	0	0	10	12	
Virus	6	6	10	6	HIV, measles, hepatitis A-D
Volatile Compounds	29	14	22	0	
total	**169**	**182**	**258**	**96**	

EWAS Methodology

4 individual cohorts

foreach: *{1999-2000, 2001-2002, 2003-2004, 2005-2006}*

foreach:

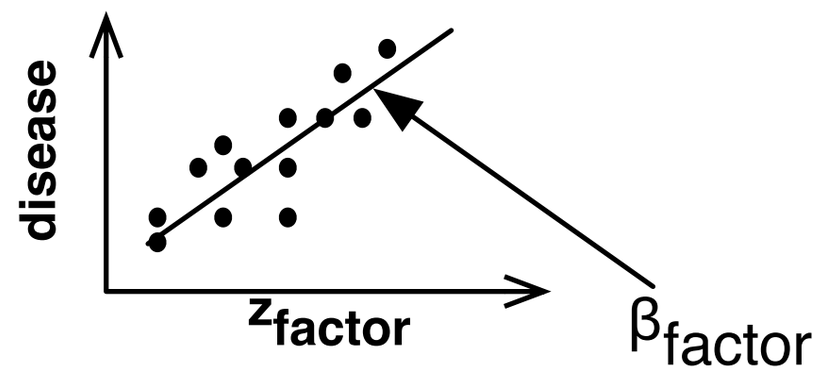
bisphenol A
PCB199
β-carotene
cotinine
...

M=96-258

Environmental factors:
log transformed & z-standardized
reference groups "negative"

Survey Regression (GEE):
adjusted for known confounding factors

age, sex, ethnicity, socioeconomic status, ...



Significance tests per cohort

p-value(β_{factor})

	1999-2000	2001-2002	2003-2004	2005-2006
bisphenol A	.	.	0.002	0.01
PCB199	0.1	0.02	0.03	.
β-carotene	NA	0.0001	0.02	0.002
cotinine	0.03	0.01	0.9	.
...	...	...	...	...

EWAS Methodology, cont'd

False Discovery Rate Estimation **foreach:** *{1999-2000, 2001-2002, 2003-2004, 2005-2006}*

foreach: { *bisphenol A*, *PCB199*, *β-carotene*, *cotinine* } M=96-258

...

permute labels or residuals *B* times **foreach:** *1 to B*

compute FDR

$$\frac{\#\ [\text{p-value}(\beta_{factor(permuted)}) < p] \times 1/B}{\#\ [\text{p-value}(\beta_{factor})]}$$

FDR(p-value)

	1999-2000	2001-2002	2003-2004	2005-2006
bisphenol A	.	.	0.1	0.2
PCB199	0.1	0.02	0.03	.
β-carotene	NA	0.01	0.1	0.05
cotinine	0.03	0.01	0.9	.
...	...	...	...	...

Tentative Validation

FDR < *threshold* in 2 or greater cohorts?
AND
sign(β_{factor}) equal for cohorts?

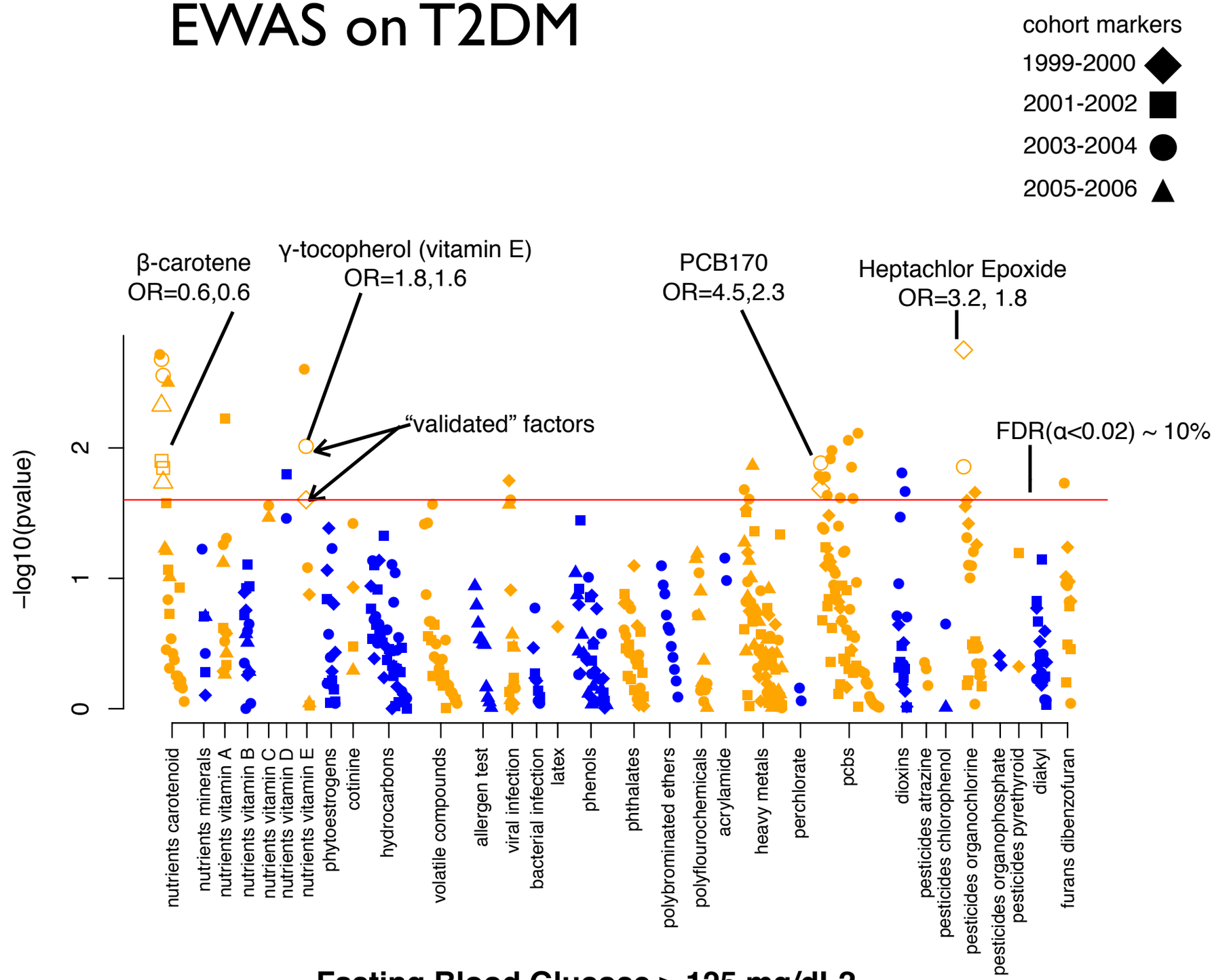
What environmental factors are associated with Type 2 Diabetes?

EWAS on T2DM

Novel Findings:
heptachlor epoxide
γ-tocopherol

Known Associations:
β-carotene
vitamin D
PCBs

Interesting Patterns:
pesticides, PCBs



Fasting Blood Glucose > 125 mg/dL?
BMI, SES, ethnicity, age, sex
OR: Δ 1SD of exposure
N=500-2000 per cohort

Patel CJ, Bhattacharya J, Butte AJ, (2010) An Environment-Wide Association Study (EWAS) on T2DM. *PLoS ONE* vol. 5(5)

What about other risk phenotypes? EWAS on Serum Lipid Levels

Risk factors for coronary heart disease (CHD)

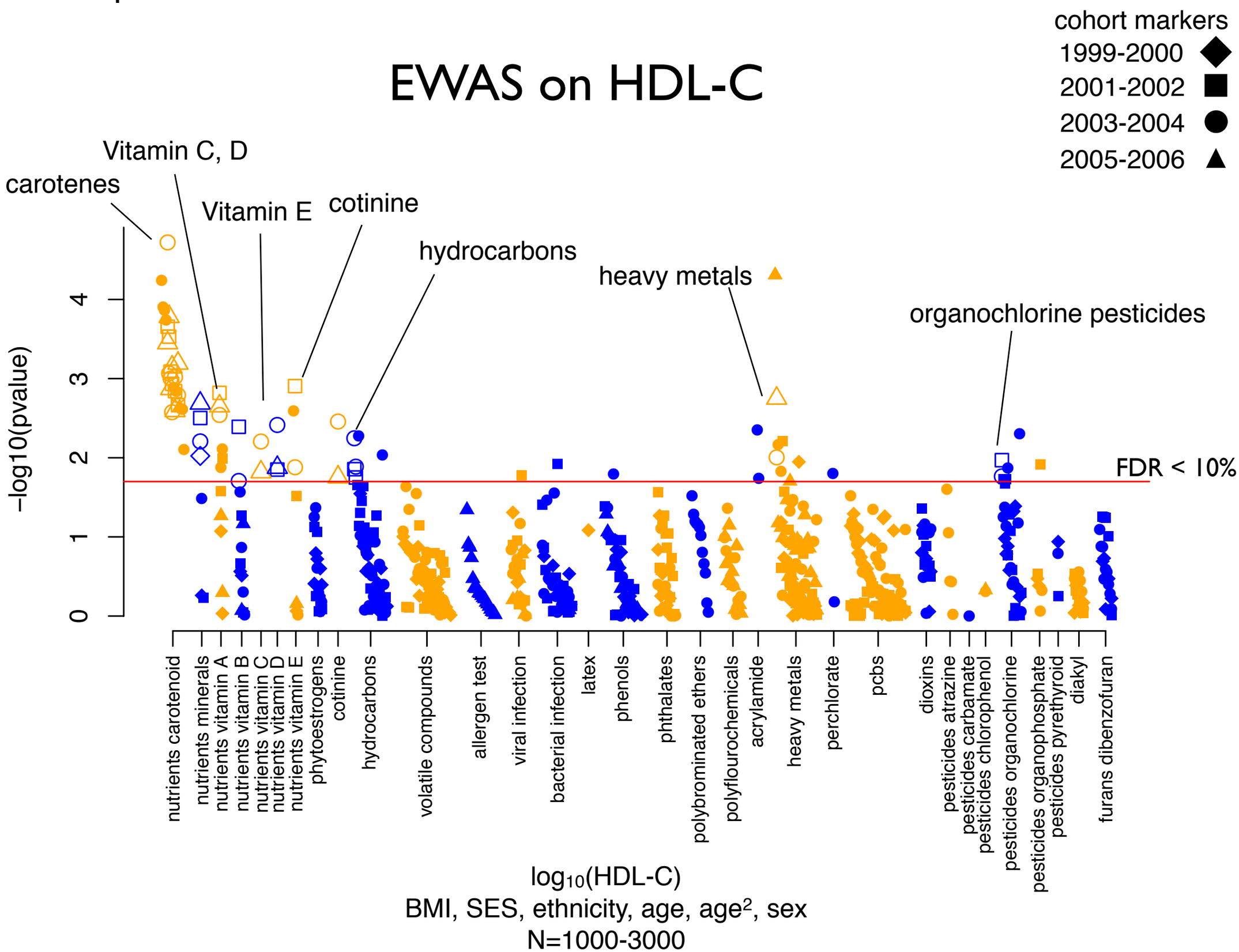
Targets for intervention (ie, statins)

Influenced by smoking, physical activity, diet, genetics[1]

lipid type	risk for CHD (for 1% increase in lipids)
LDL-Cholesterol	1% increased risk[2]
HDL-Cholesterol	2% decreased risk[3]
Triglycerides	(increased risk)

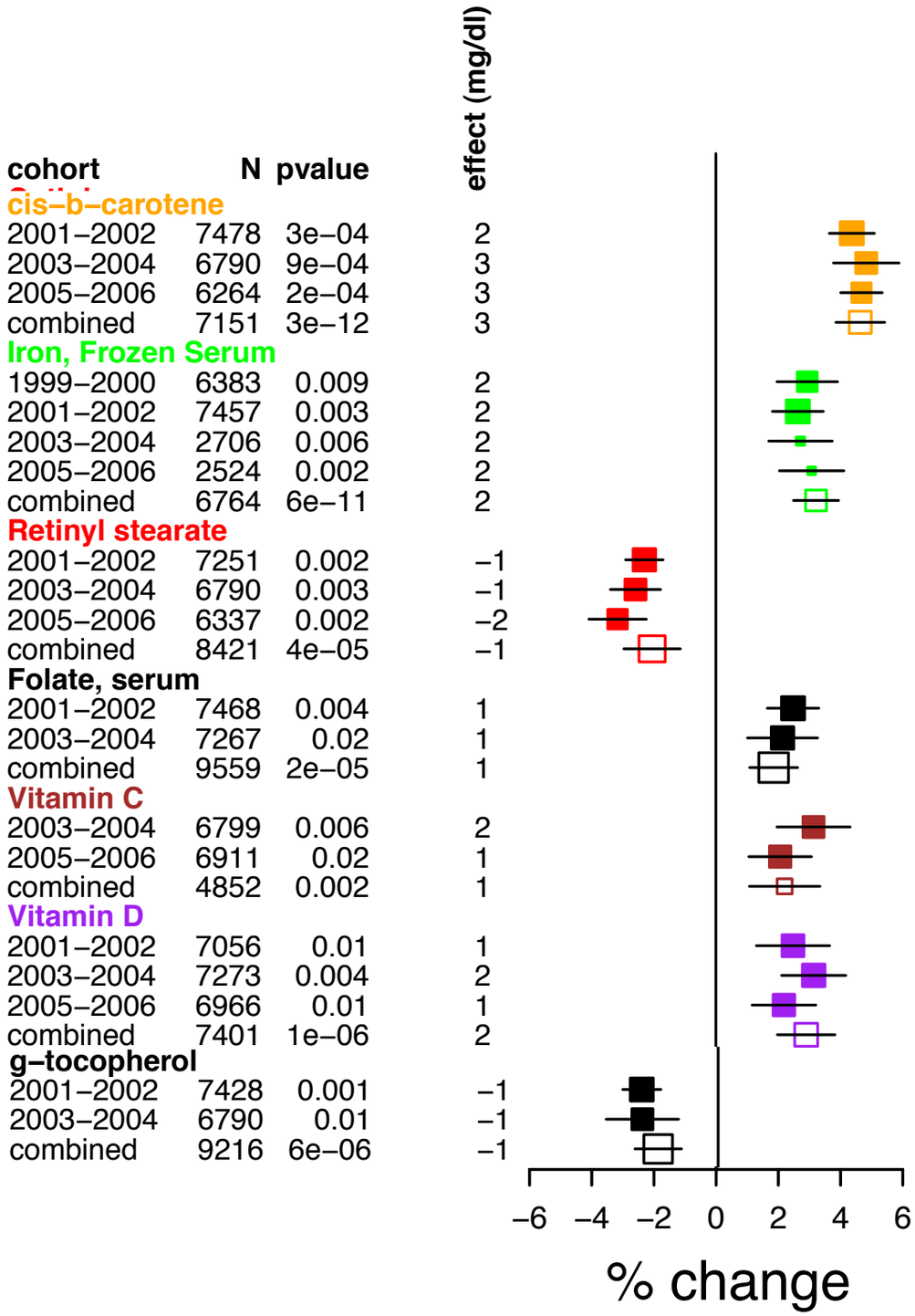
1. Tanya M. Teslovich et al. *Nature* (2010) vol. 466 (7307) pp. 707
2. Grundy et al. *Arteriosclerosis, Thrombosis, and Vascular Biology* (2004) vol. 24 (2) pp. e13
3. Gotto et al. *Journal of the American College of Cardiology* (2004) vol. 43 (5) pp. 717-24

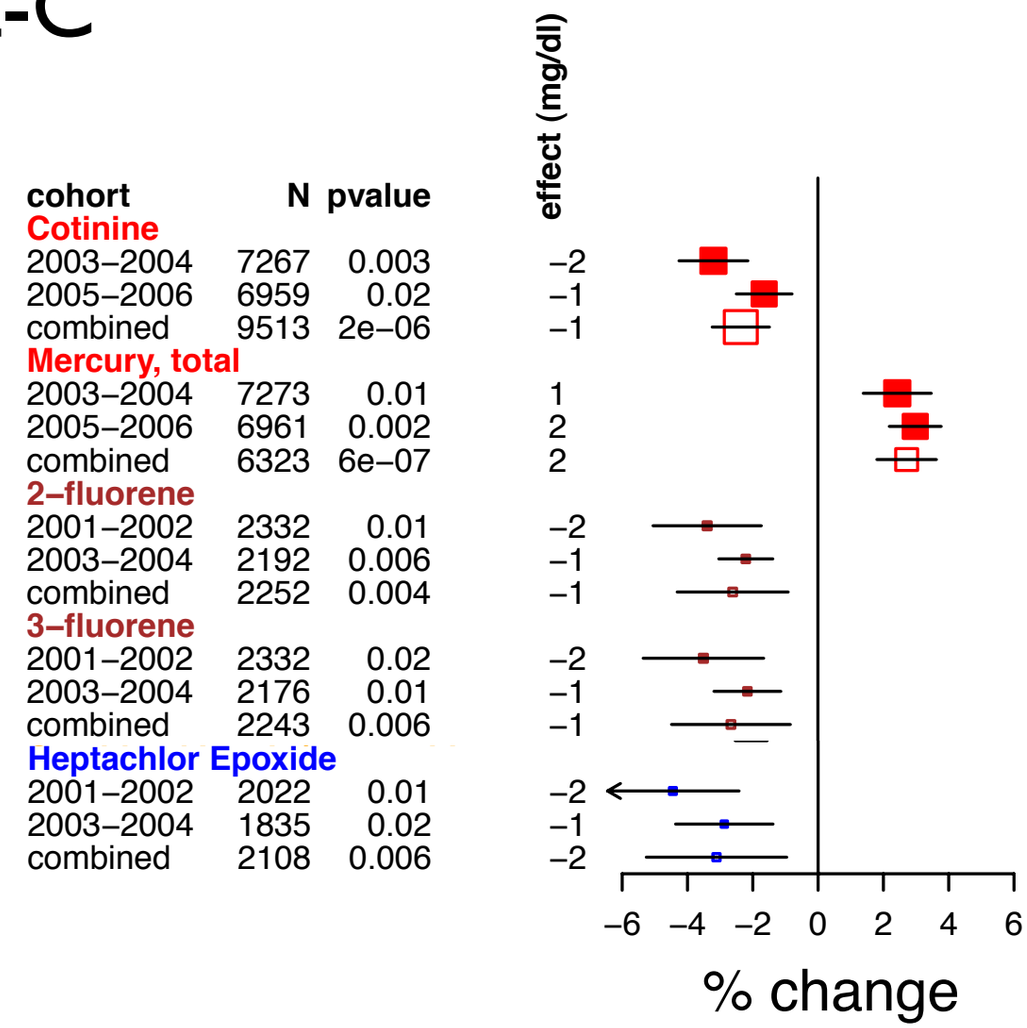
Aim 2: EWAS examples



Patel CJ, Cullen MR, Ioannidis JAP, Butte AJ, (2011). Non-genetic associations and correlation globes for determinants of Lipid Levels: an EWAS. In Review.

Effect Sizes For Validated Factors: HDL-C





% change = Δ 1 SD in Exposure

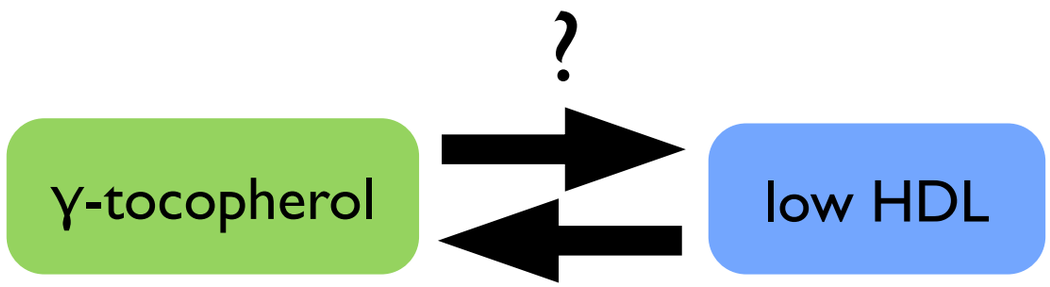
18 *validated* factors

combined adjusted for:
BMI, SES, ethnicity, age, age^2, sex, waist circumference, diabetes (FBG > 125 mg/dL), blood pressure

comparable to genetic effect sizes[1]!

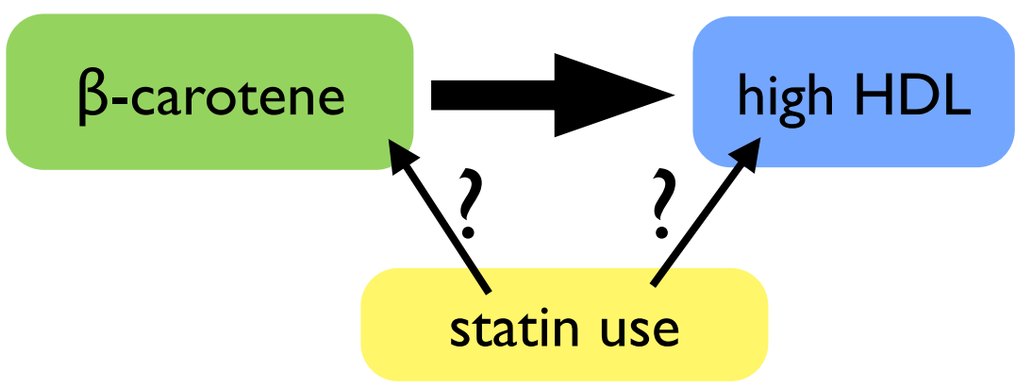
Checking for Validity

Assessing Validity of Estimates
example: HDL-C



tocopherol (vitamin e) supplements for CHD individuals?

Could the disease "lead" to exposure?
"Reverse causality"

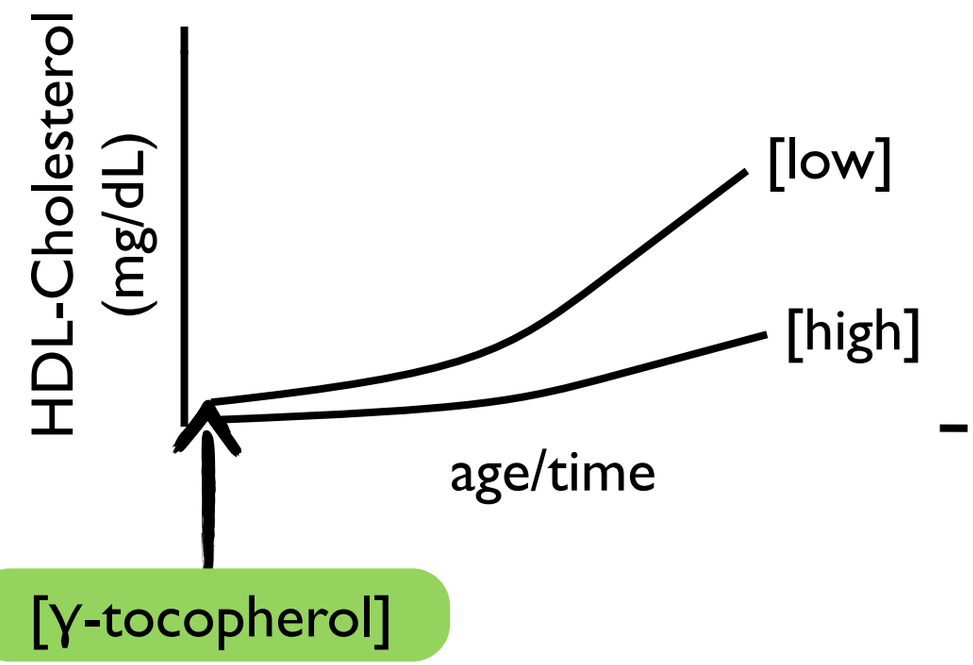


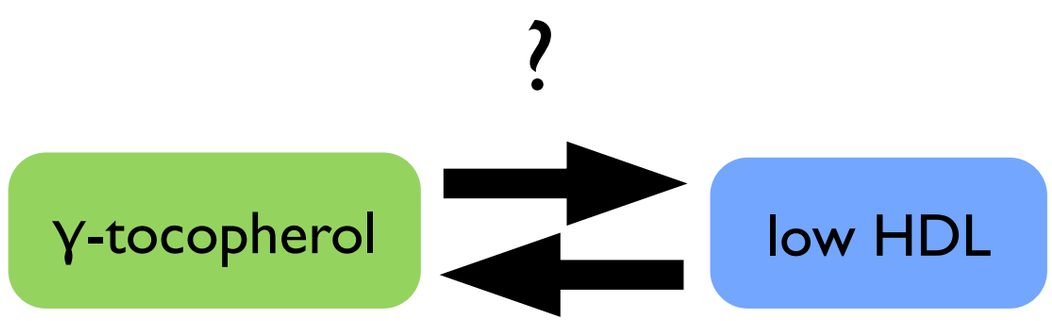
confounders

Could there something *confounding* the association?

Longitudinal Study: "Gold Standard" for Validation

- exposure changing through time
- reverse causality bias
- compute disease risk





tocopherol (vitamin e) supplements for CHD individuals?

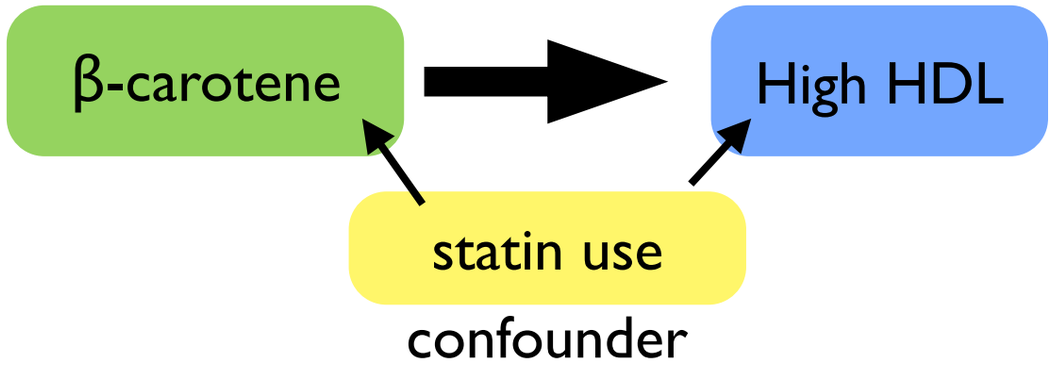
Addressing Confounding Bias with the Exposome

$E[\text{HDL-C}] = \alpha + \boldsymbol{\beta_{1,\,original}} * \text{carotene}$

$E[\text{HDL-C}] = \alpha + \boldsymbol{\beta_{1,extended}} * \text{carotene} + \boldsymbol{\beta_2} * \text{statin use}$

compare $\boldsymbol{\beta_{1,\,original}}$ and $\boldsymbol{\beta_{1,extended}}$

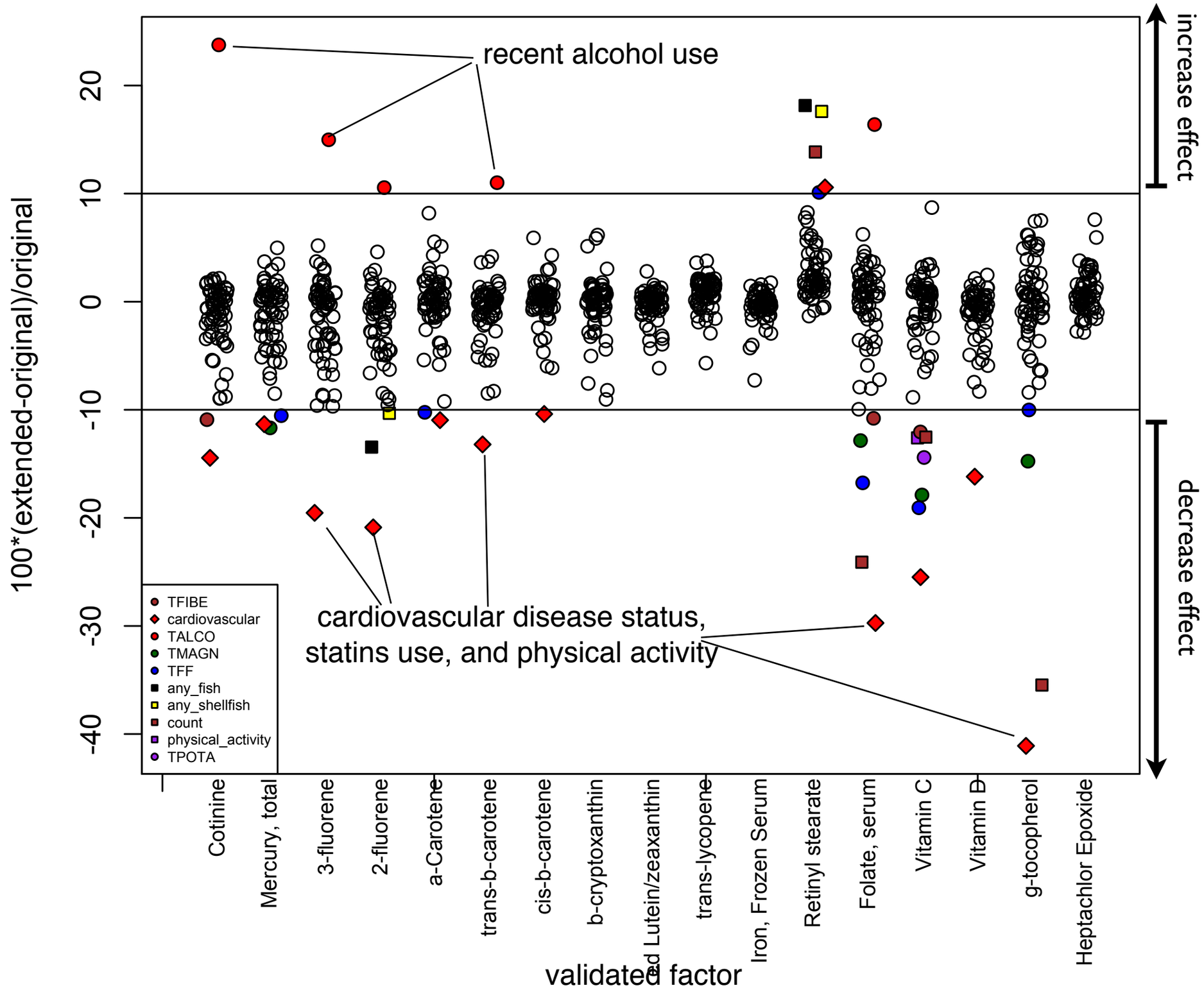
"account" for bias due to statin use



"source" of bias	example variables
disease status	diabetes, CHD, heart attack
drug use	metformin, statins use
supplement use	count of total supplements used
physical activity	daily estimated metabolic equivalents
recent food intake	total nutrients computed from frequency questionnaire

total: 62

Assessing Bias from Self-report data
ex.: HDL-C



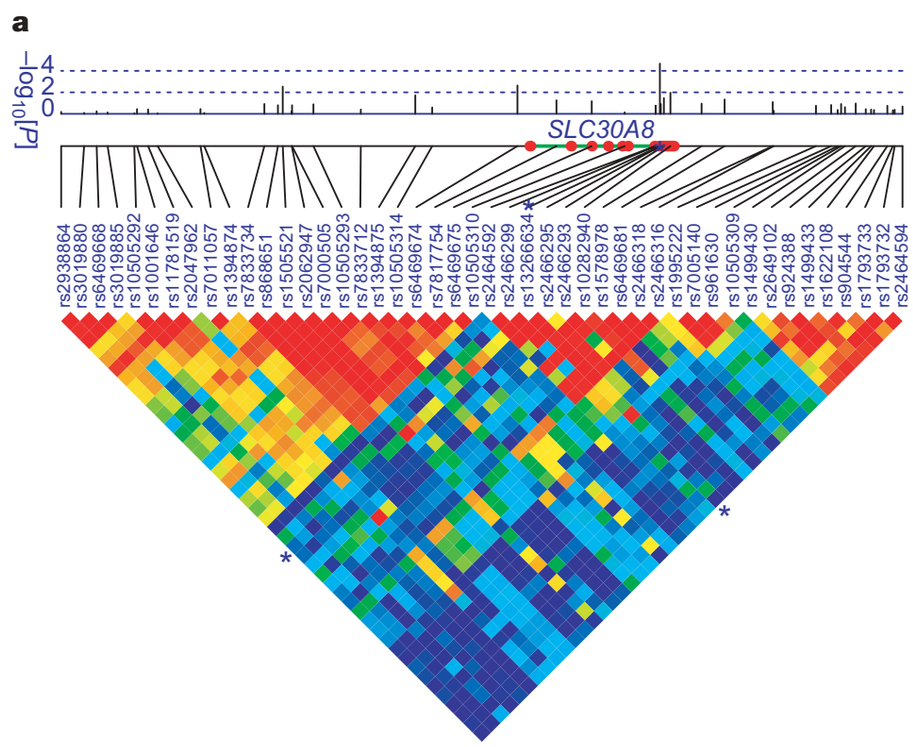
Can we use an "exposome" to elucidate combinations of exposures?

What is the *LD* of the exposome?

What is "Linkage Disequilibrium" of the *Exposome*?

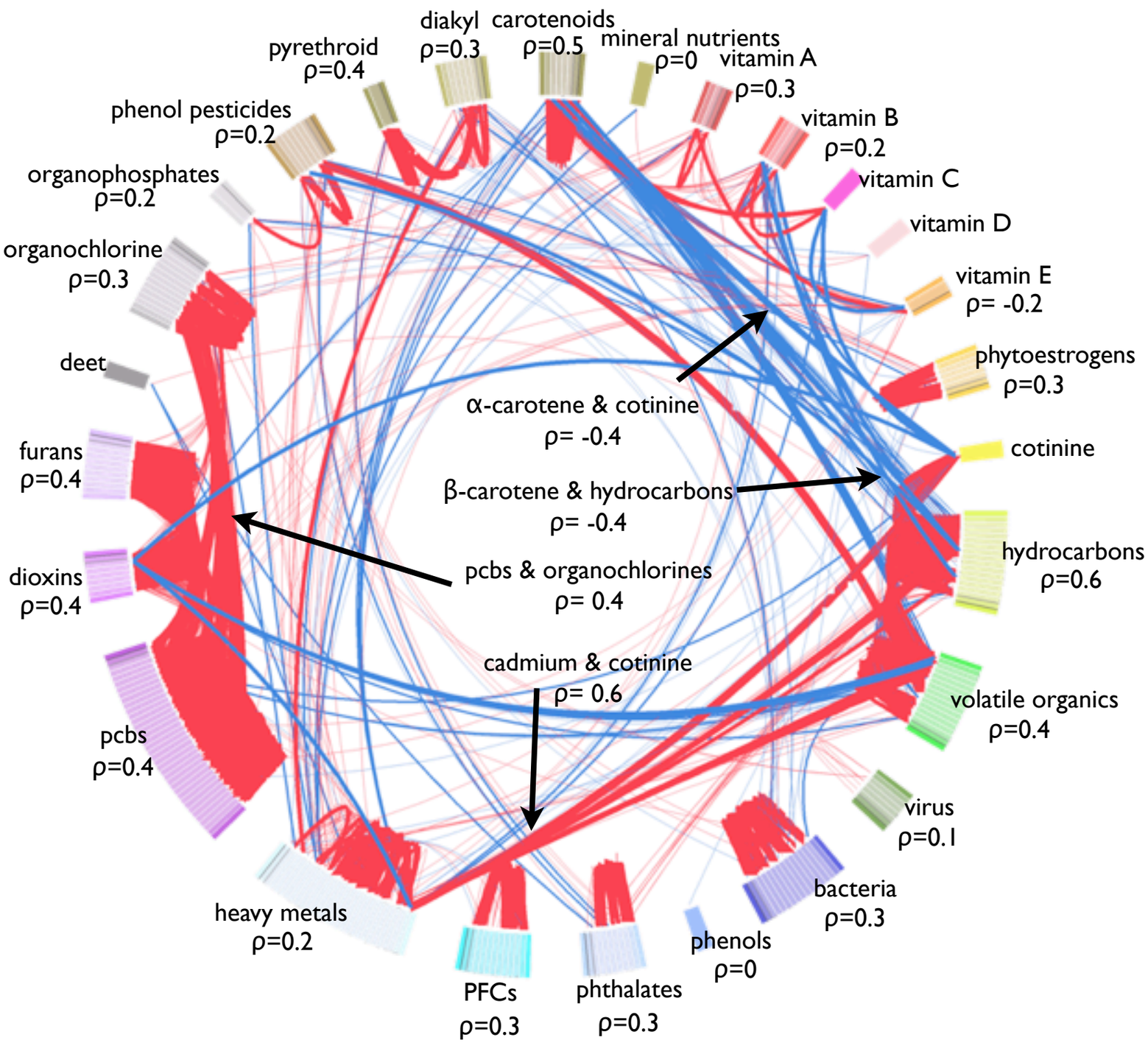
Given measurement of the exposome,
it is possible discern combinations of potential exposures.

Analogy: "Linkage Disequilibrium" and correlated loci



Sladek et al., *Nature Genetics*. 2007

Preliminary *LD* of NHANES Exposures

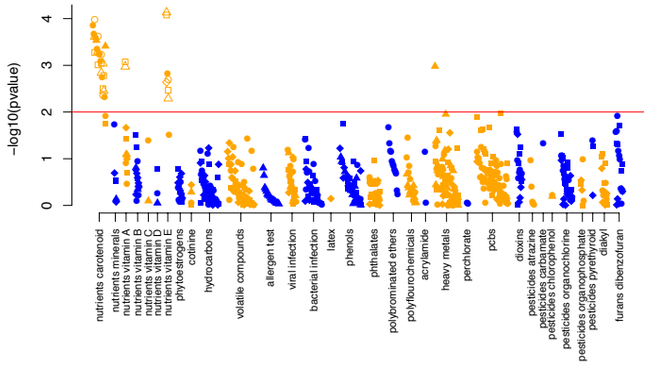


for each pair:
Partial ρ
age, BMI (serum)
age, creatinine (urine)

permuted data to produce
"null ρ"
filtered those ρ > .3 or < -0.2
(5th, 95th percentile)
sought replication in > 1 cohort

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

EWAS:
Conclusion and Discussion



- generalizable, comprehensive, transparent, and systematic study of environment

- novel associations for T2D and HDL-C

heptachlor epoxide
γ-tocopherol

- effects on disease are on par with genetics, calling for large-scale *exposomic* study

HDL-C: 1-10 mg/dL
T2D: ~2-3 OR

- *However*: confounding, reverse causal biases: need longitudinal and follow-up studies.

Acknowledgments

- Atul Butte
- Jay Bhattacharya
- Mark Cullen
- John Ioannidis
- Rob Tibshirani
- Biomedical Informatics at Stanford
- CDC / NCHS
- NHANES and participants!
- NIEHS
- NLM (T15 LM 007033)

Lucile Packard Children's Hospital

Stanford School of Medicine

email: chirag.patel@stanford.edu