Building a ***search engine*** to find ***environmental factors*** associated with ***disease and health***

Chirag J Patel
IEA-WCE 2017 Symposium
Saitama, Japan
8/20/17

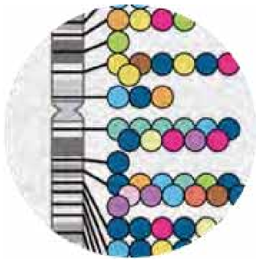
HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics

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

Phenotype *Genome* *Environment*

P = G + E

Phenotype	Genome	Environment
Type 2 Diabetes	Variants	Infectious agents
Cancer		Diet + Nutrients
Alzheimer's		Pollutants
Gene expression		Drugs



G

We are great at ***G*** investigation!

2,940 (as of 6/1/17)
36,066 *G-P* associations
Genome-wide Association Studies (GWAS)
https://www.ebi.ac.uk/gwas/

E: ???

Nothing comparable to elucidate ***E*** influence!

We lack high-throughput methods
and data to discover new ***E*** in ***P…***

A similar paradigm for discovery should exist for ***E***!

Why?

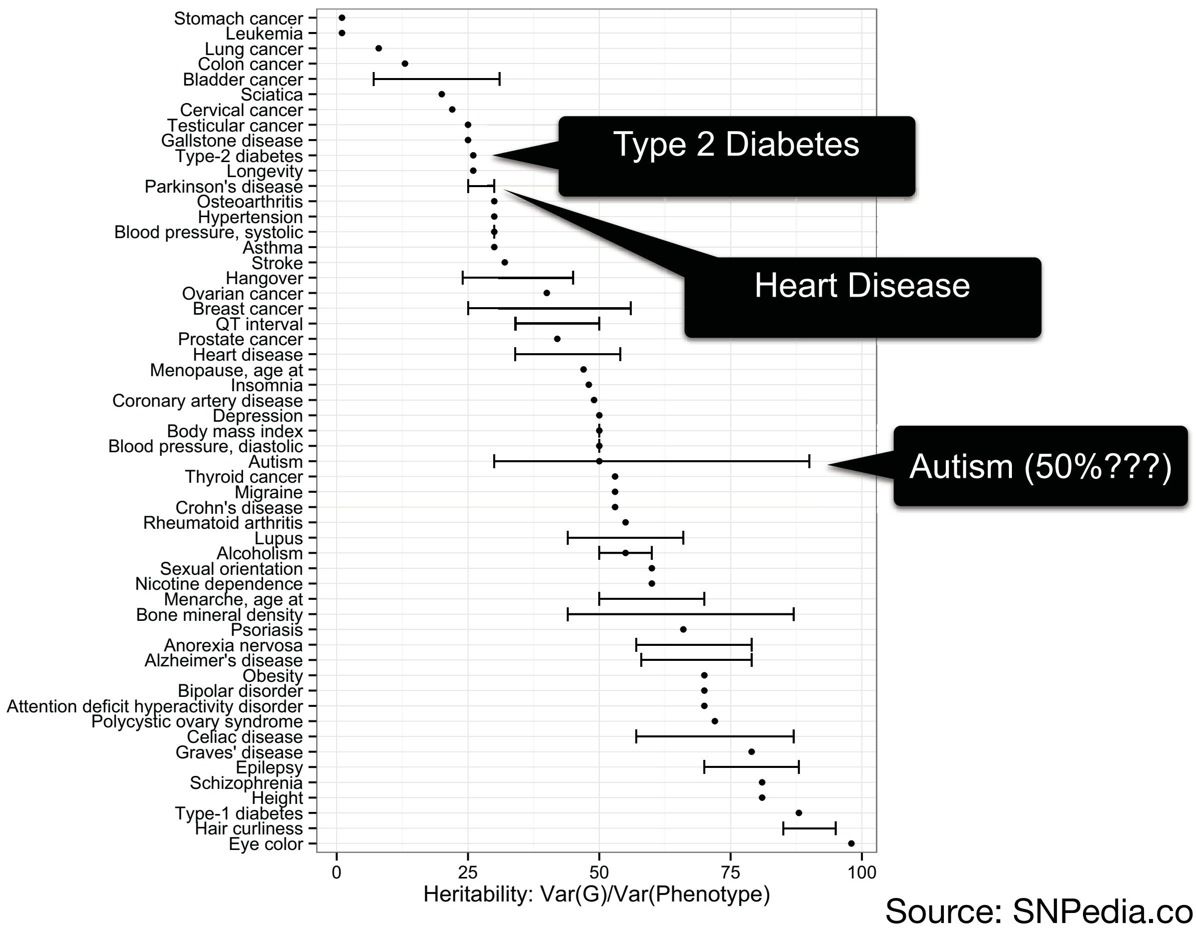
$$\sigma^2_P = \sigma^2_G + \sigma^2_E$$

Heritability (H^2) is the range of phenotypic variability attributed to genetic variability in a population

$$H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

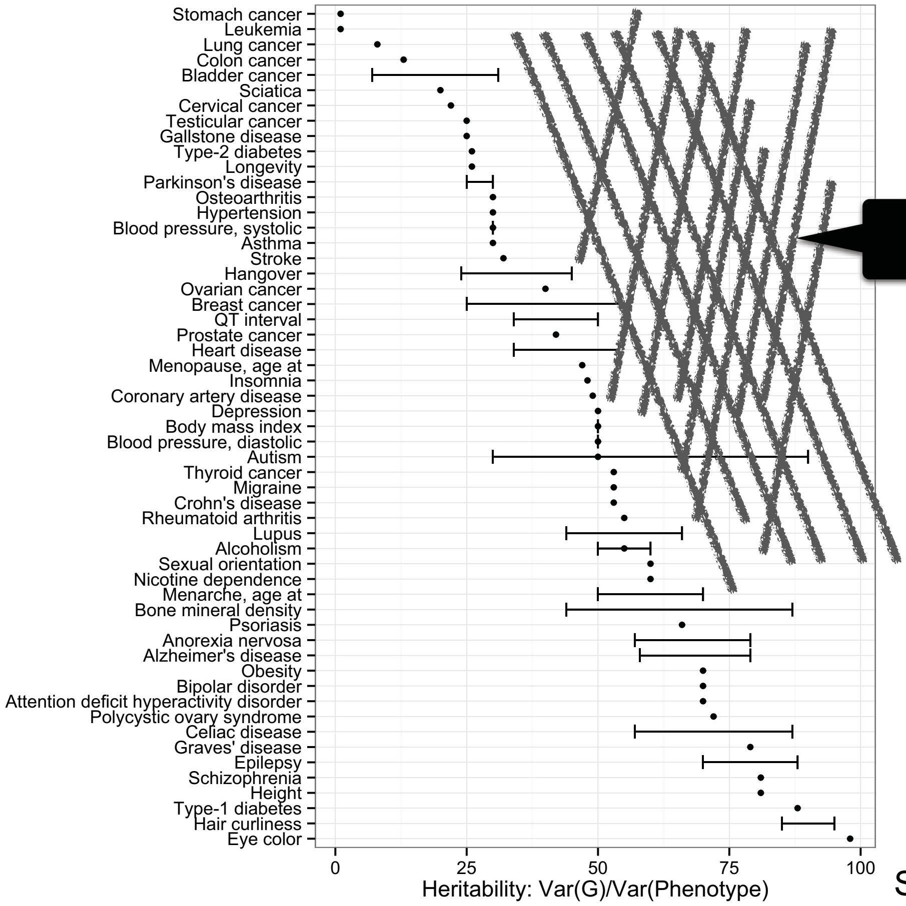
Indicator of the proportion of phenotypic differences attributed to **G**.

G estimates for burdensome diseases are **low and variable:** massive opportunity for ***high-throughput E*** *discovery*



Source: SNPedia.com

G estimates for complex traits are **low and variable:** massive opportunity for *high-throughput **E** discovery*



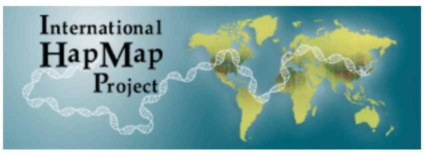
Source: SNPedia.com

It took a new paradigm of ***GWAS*** for discovery**:** Human Genome Project to ***GWAS***

Sequencing of the genome

2001

Characterize common variation



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

2001-current day

Measurement tools

High-throughput variant assay
< $99 for ~**1M** variants
~2003 (ongoing)

Comprehensive, high-throughput analyses

GWAS

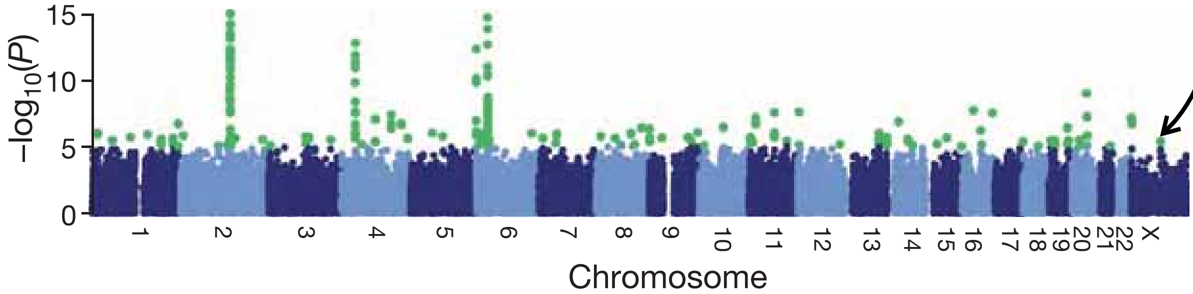
Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

WTCCC, Nature, 2008.

What is a Genome-Wide Association Study (GWAS)?: Data-driven search for ***G*** factors in ***P***

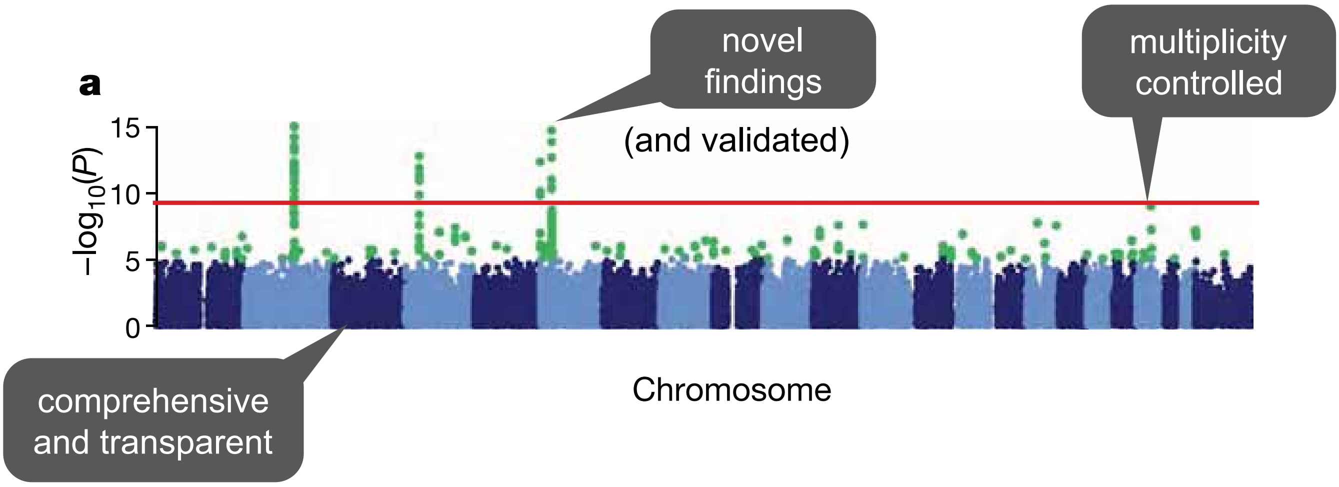
	AA	Aa	aa
case			
control			



Robust, transparent, and comprehensive search for ***G*** in ***P***

WTCCC *Nature,* 2007

Why carry out a Genome-Wide Association Study: Analytically robust, transparent, and comprehensive search for ***G*** in ***P***



JAMA 2014
JECH 2014

Promises and ***Challenges*** in creating a search engine for identifying ***E*** in ***P***

Studying the Elusive Environment in Large Scale

JAMA 2014

Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health

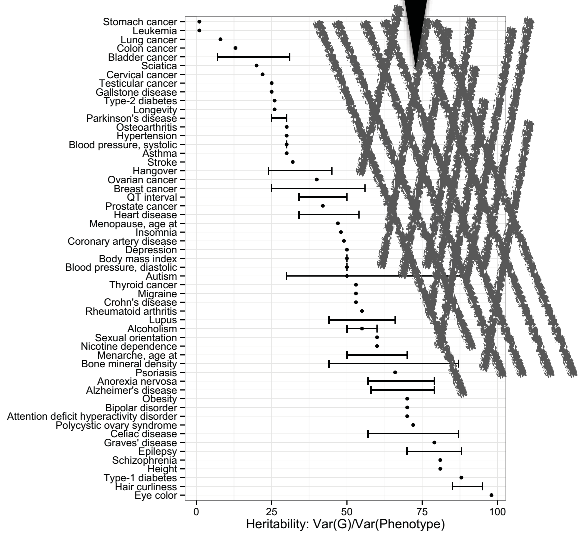
ARPH 2016

Placing epidemiological results in the context of multiplicity and typical correlations of exposures

JECH 2014

Promises and ***Challenges*** in creating a search engine for ***E*** in ***P***

σ^2_E : Exposome!



High-throughput E = **discovery**!
systematic; reproducible
multiple hypothesis control
prioritization

Arjun Manrai
(Yuxia Cui, David Balshaw)

ARPH 2016
JAMA 2014
JECH 2014

Examples of ***exposome-driven*** discovery machinery, or ***EWASs***

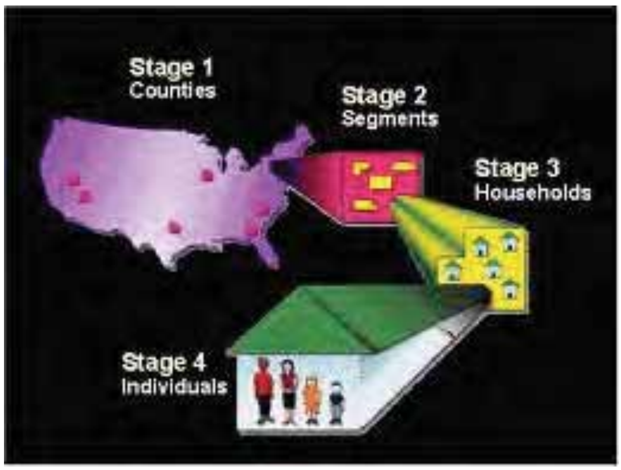
Gold standard for ***breadth*** of human exposure information:
National Health and Nutrition Examination Survey[1]

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

>250 exposures (serum + urine)
GWAS chip

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

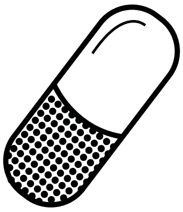
Death index linkage (cause of death)



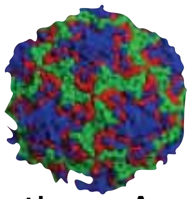
1 http://www.cdc.gov/nchs/nhanes.htm

Gold standard for ***breadth*** of exposure & behavior data:
National Health and Nutrition Examination Survey

Nutrients and Vitamins
vitamin D, carotenes



Drugs
statins; aspirin



Infectious Agents
hepatitis, HIV, Staph. aureus

Plastics and consumables
phthalates, bisphenol A

Pesticides and pollutants
atrazine; cadmium; hydrocarbons

Physical Activity
e.g., steps

What ***E*** are associated with ***aging***:
all-cause mortality and
telomere length?

Int J Epidem 2013
Int J Epidem 2016

How does it work?:
Searching *for* exposures and behaviors associated with **all-cause mortality.**

NHANES: 1999-2004
National Death Index linked mortality
246 behaviors and exposures (serum/urine/*self-report*)

NHANES: 1999-2001
N=330 to 6008 (26 to 655 deaths)
~5.5 years of followup

Cox proportional hazards
baseline exposure and time to death

False discovery rate < 5%

?

NHANES: 2003-2004
N=177 to 3258 (20-202 deaths)
~2.8 years of followup

$p < 0.05$

Variance explained (**R^2**):
Proportion of variance in death correlated with ***E***

Int J Epidem 2013

How does it work?:
Discriminating signal from noise using family-wise error rate with the ***False Discovery Rate***

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.

Benjamini and Hochberg, J R Stat Soc B 1993

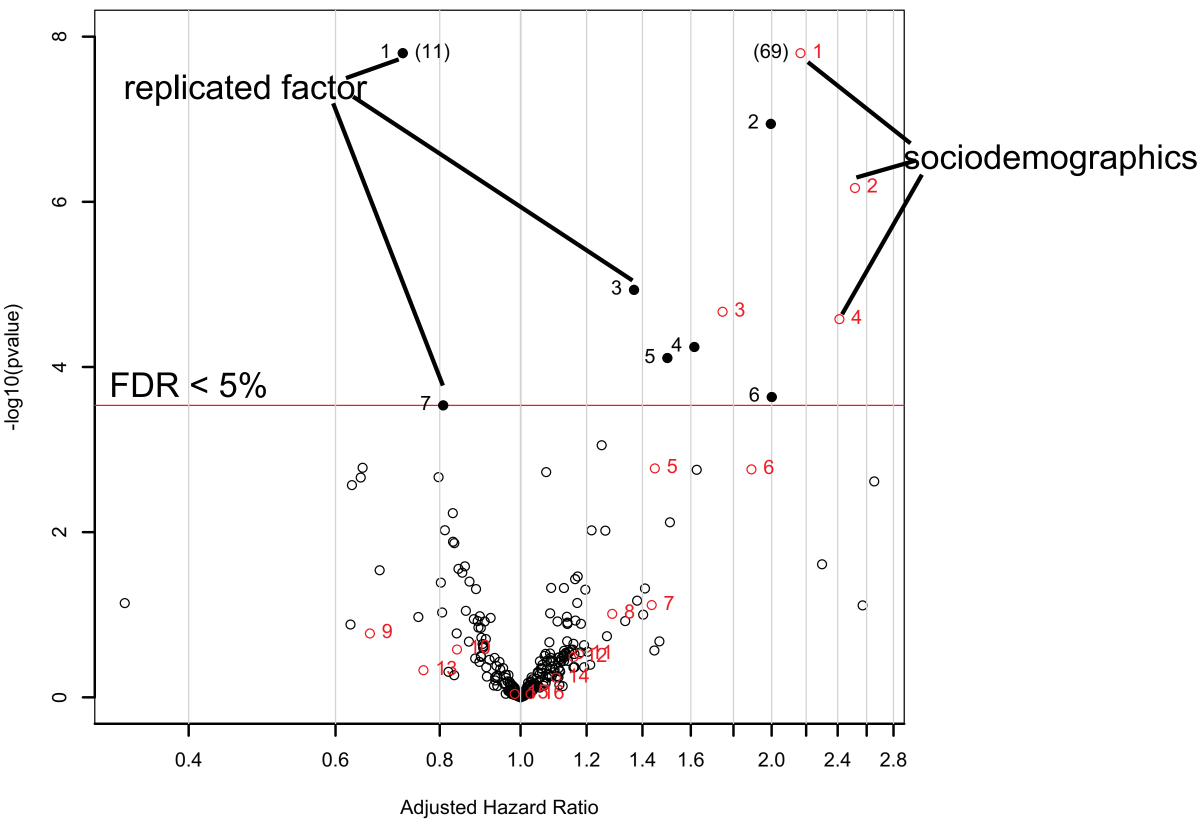
How does multiple testing correction work?

William S Noble

When prioritizing hits from a high-throughput experiment, it is important to correct for random events that falsely appear significant. How is this done and what methods should be used?

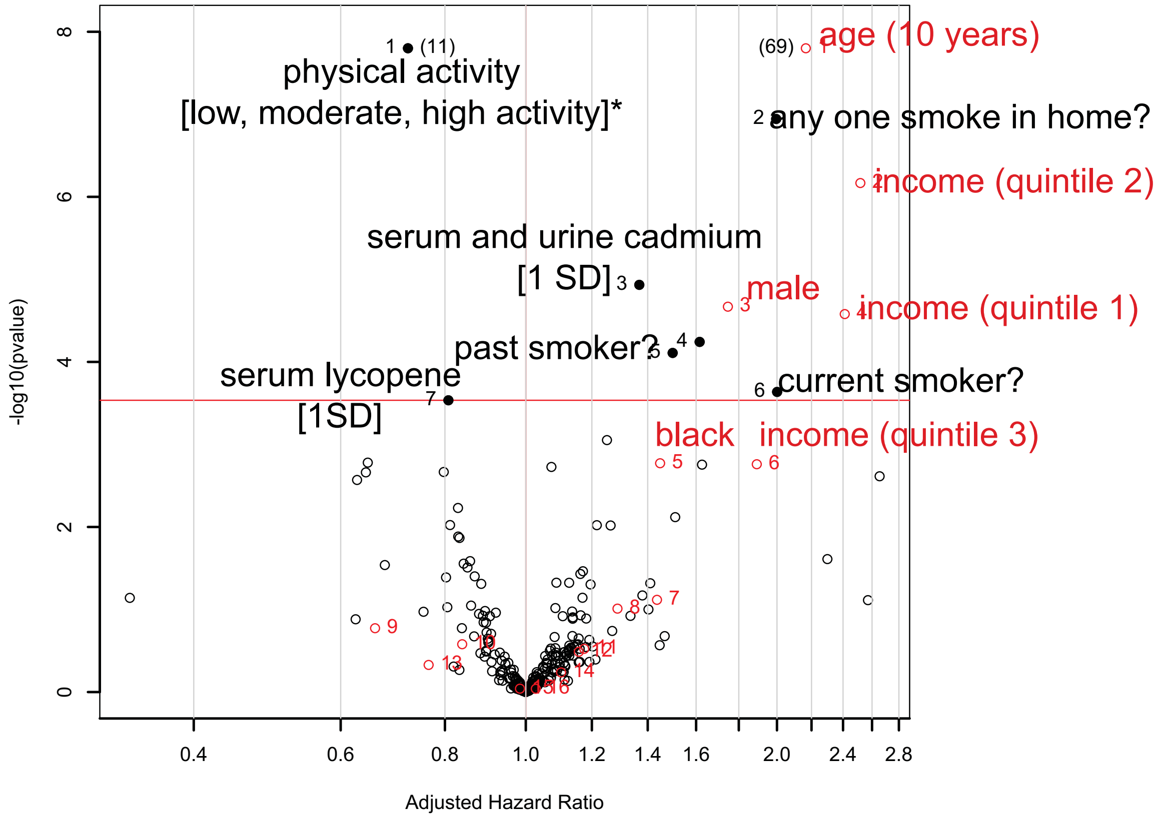
Noble, Nature Biotech 2009

EWAS in all-cause mortality: 253 exposure/behavior associations in survival



Int J Epidem 2013

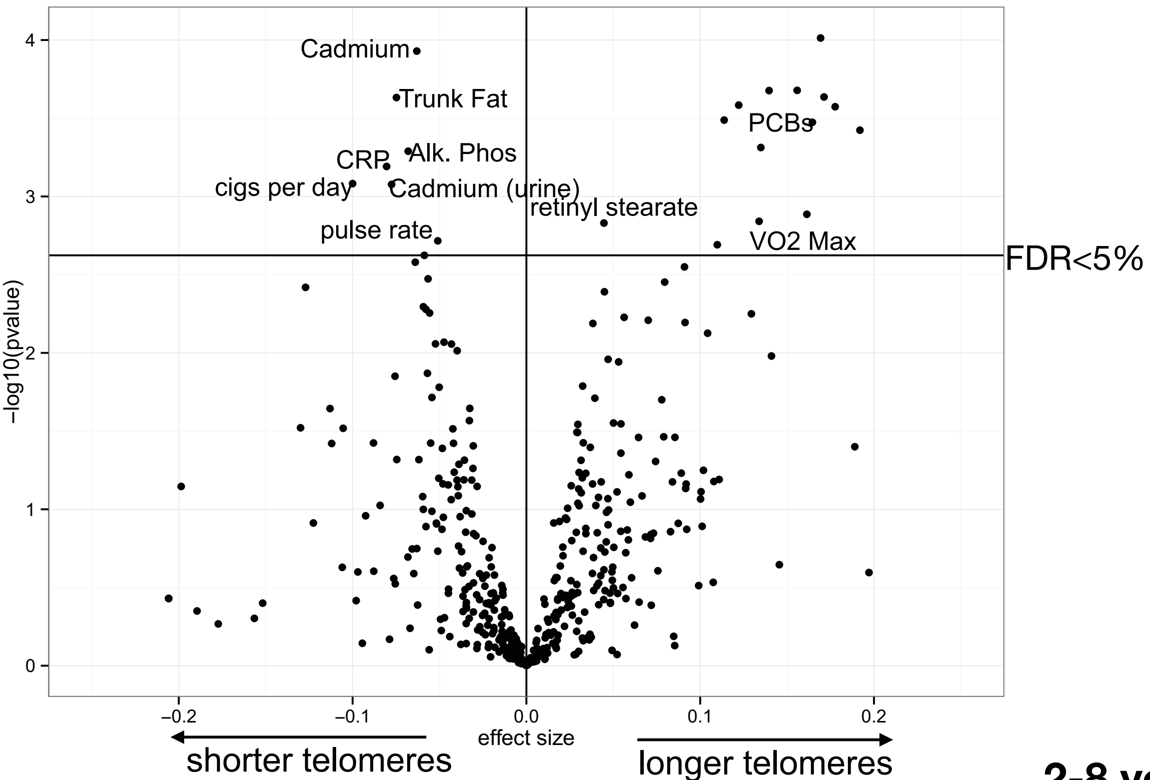
EWAS identifies factors associated with ***all-cause mortality***: Volcano plot of 200 associations



Multivariate cox (age, sex, income, education, race/ethnicity, occupation [in red])
*derived from METs per activity and categorized by Health.gov guidelines

R^2 ~ 2%

452 associations in ***Telomere Length***: Polychlorinated biphenyls associated with longer telomeres?!



adjusted by age, age^2, race, poverty, education, occupation
median N=3000; N range: 300-7000

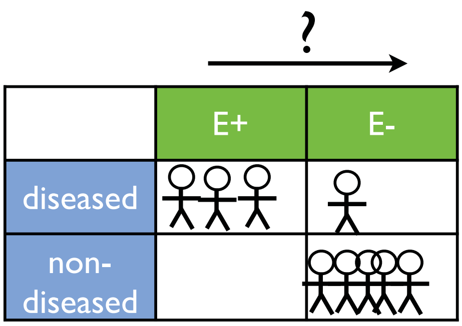
2-8 years
R^2 ~ 1%

Int J Epidem 2016

20 more examples:
https://paperpile.com/shared/PtvEae

diabetes
preterm birth
income
blood pressure
lipids
kidney disease
telomere length
mortality

It is possible to capture ***E*** in high-throughput to create biomedical hypotheses using tools such as ***EWAS***



Versus

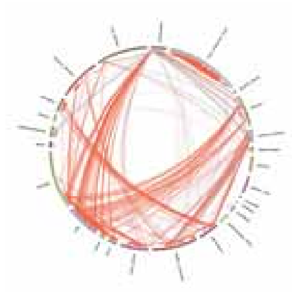
candidates

comprehensive

Promises and **Challenges** in creating a search engine for ***E*** in ***P***

High-throughput assays of E!
scalable and standard technologies

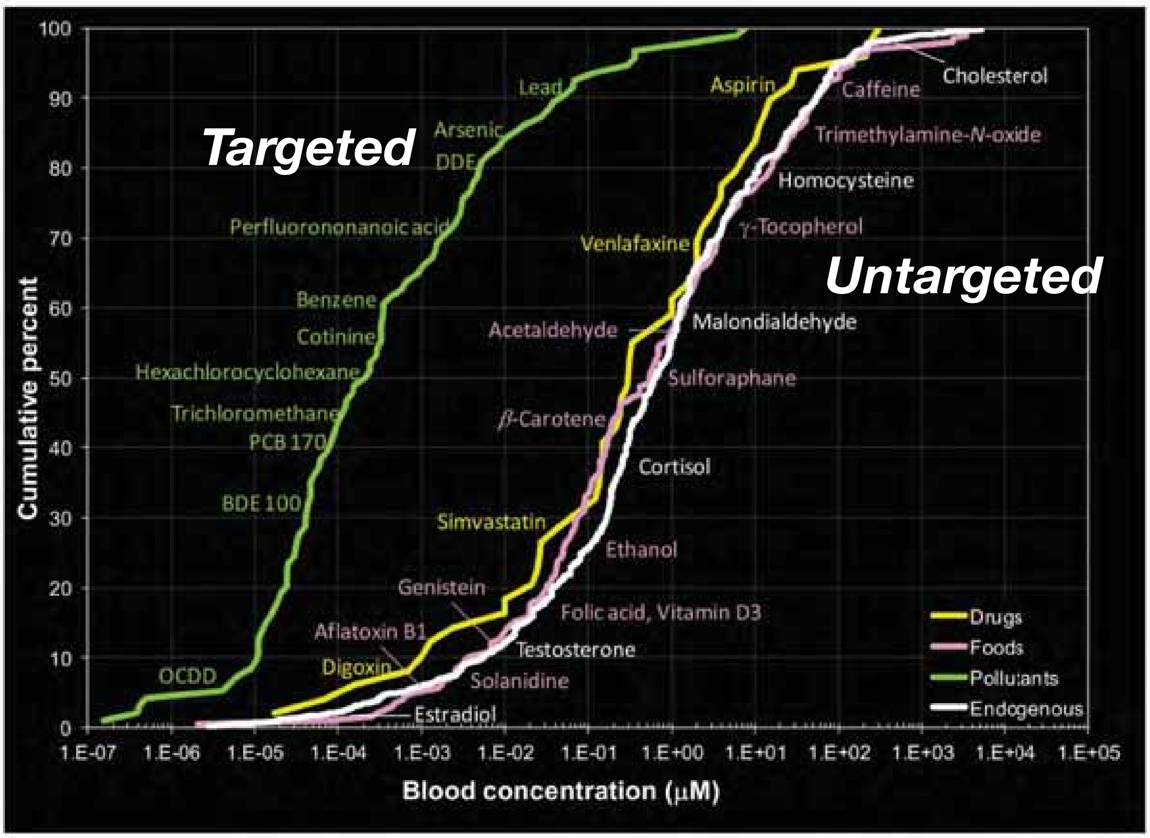
Big data = big bias!
Confounding; reverse causality
Dense correlational web of ***E*** and ***P***
Fragmented and small ***E***-***P*** associations
Influence of time and life-course



Arjun Manrai
(Yuxia Cui, David Balshaw)

ARPH 2016
JAMA 2014
JECH 2014

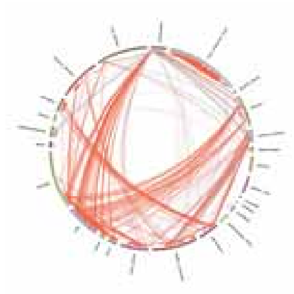
Challenge to scale ***absolute E*** due to heterogeneity and large dynamic range.



Rappaport et al, *EHP* 2015

Promises and **<u>Challenges</u>** in creating a search engine for ***E*** in ***P***

High-throughput assays of E!
scalable and standard technologies



Big data = big bias!
Confounding; reverse causality
Dense correlational web of ***E*** and ***P***
Fragmented and small ***E***-***P*** associations
Influence of time and life-course

Arjun Manrai
(Yuxia Cui, David Balshaw)

ARPH 2016
JAMA 2014
JECH 2014

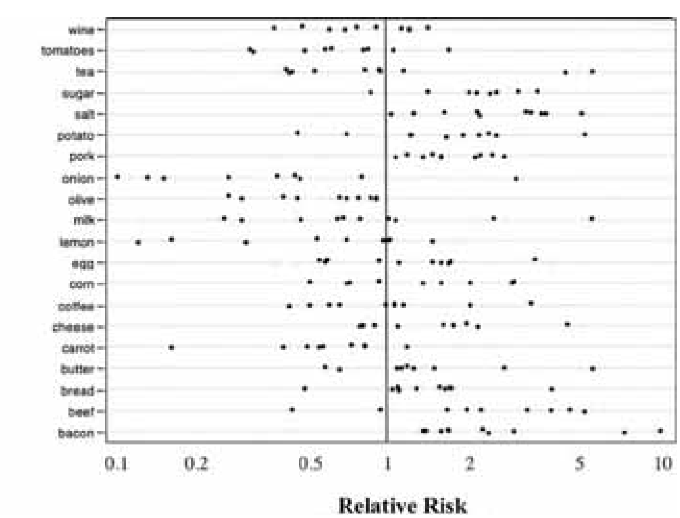
Example of ***fragmentation****:*
Is everything we eat associated with cancer?

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

Are all the ***drugs*** we take associated with ***cancer***?

Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study

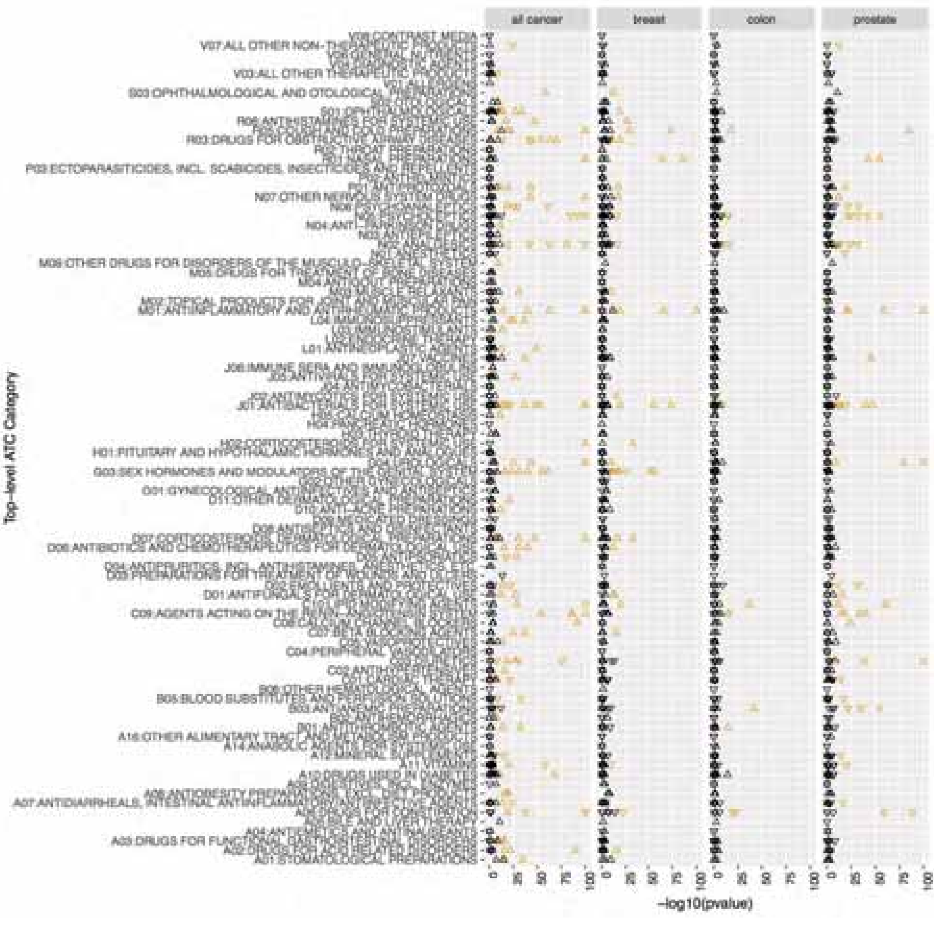
Chirag J. Patel[1], Jianguang Ji[2], Jan Sundquist[2], John P. A. Ioannidis[3] & Kristina Sundquist[2]

Associated all (~500) drugs prescribed in
entire population of Sweden
(N=9M) with time to cancer

Assessed 2 modeling techniques (**Cox** and **case-crossover**)

Sci Reports 2016

What drugs are associated with time to cancer?
Too **many** to be plausible (up to **26**%!)



any cancer:
141 (26%)
prostate:
56 (10%)
breast:
41 (7%)
colon:
14 (3%)

Modest *concordance between Cox and case-crossover:* ***12*** *out of* ***141****!*

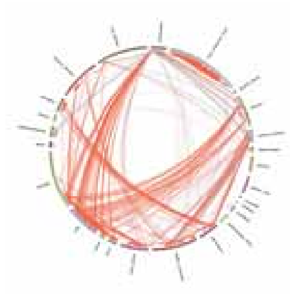
Most correlations small (HR < 1.1); residual confounding?

Sci Reports 2016

Promises and **<u>Challenges</u>** in creating a search engine for ***E*** in ***P***

High-throughput assays of E!
scalable and standard technologies

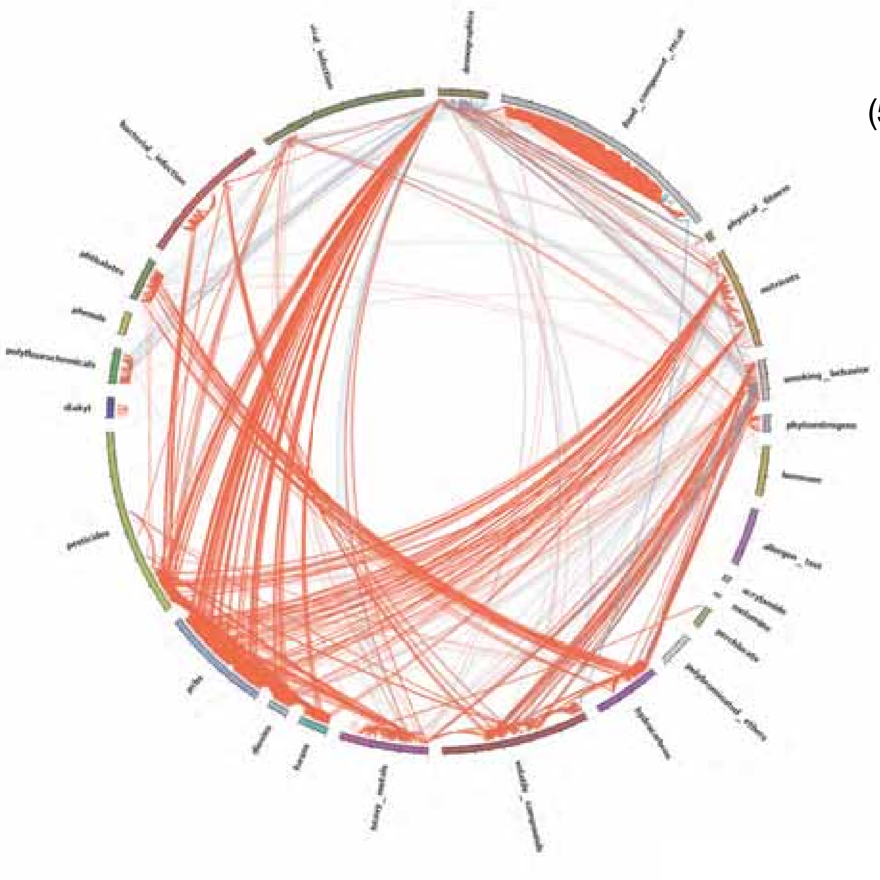
Big data = big bias!
Confounding; reverse causality
Dense correlational web of ***E*** and ***P***
Fragmented and small ***E***-***P*** associations
Influence of time and life-course



Arjun Manrai
(Yuxia Cui, David Balshaw)

ARPH 2016
JAMA 2014
JECH 2014

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



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

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

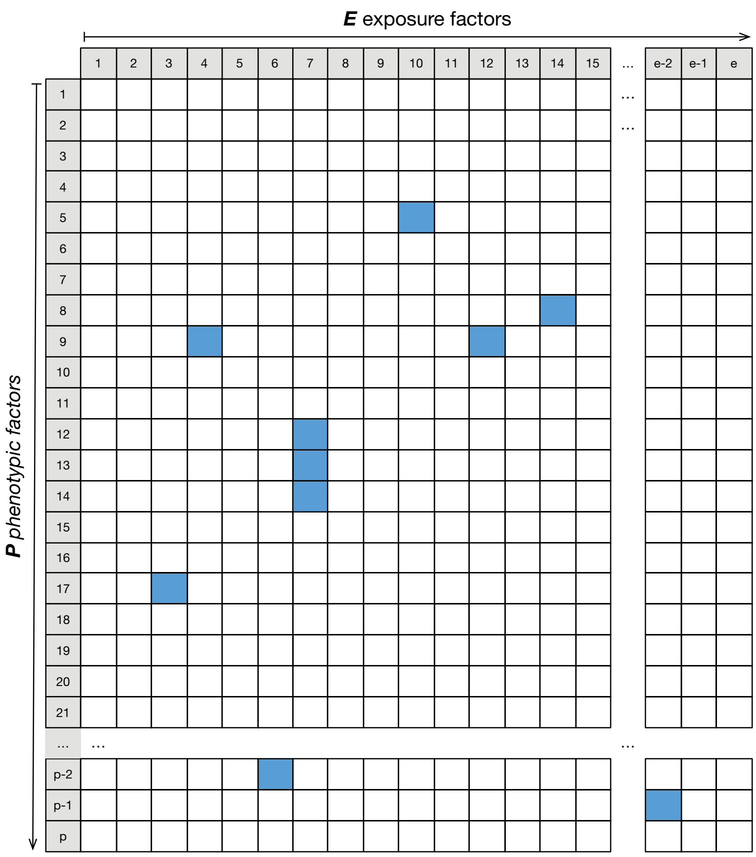
Red: positive ρ
Blue: negative ρ
thickness: |ρ|

Effective number of variables:
500 (10% decrease)

Pac Symp Biocomput. 2015
JECH. 2015

*Does my single association between **E** and **P** matter?*

Does my association between *E* and *P* matter in the entire possible space of associations?



which ones to test?
all?
the ones in ***blue***?

E times ***P*** possibilities!
how to detect signal from noise?

ARPH 2017
Hum Genet 2012
JECH 2014
Curr Epidemiol Rep 2017
Curr Env Health Rep 2016

Scaling up the search in multiple phenotypes:
does my single association between E and P matter?

P

<u>Body Measures</u>
Body Mass Index
Height

<u>Metabolic</u>
Glucose
LDL-Cholesterol
Triglycerides

<u>Blood pressure & fitness</u>
Systolic BP
Diastolic BP
Pulse rate
VO_2 Max

<u>Kidney function</u>
Creatinine
Sodium
Uric Acid

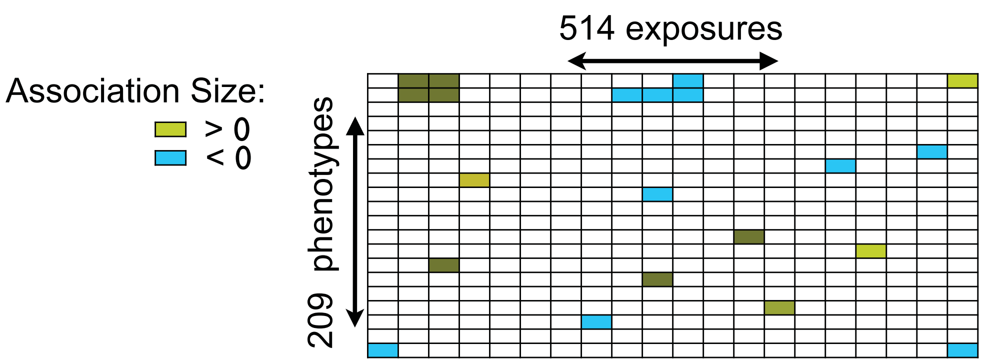
<u>Aging</u>
Telomere length
Time to death

<u>Inflammation</u>
C-reactive protein
white blood cell count

<u>Liver function</u>
Aspartate aminotransferase
Gamma glutamyltransferase

Raj Manrai, Hugues Aschard, JPA Ioannidis, Dennis Bier

Creation of a phenotype-exposure association ***map***:
A 2-D view of 209 phenotype by 514 exposure associations

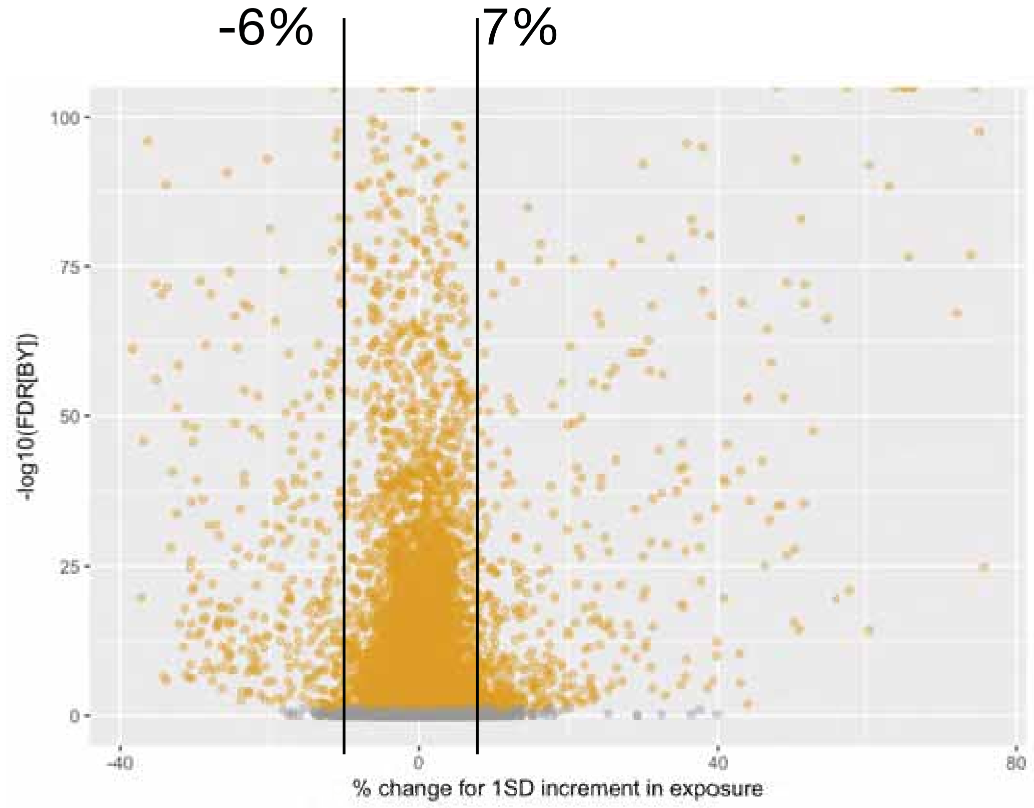


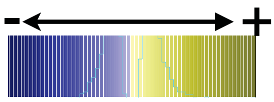
504 ***E*** exposure and diet indicators × 209 clinical trait phenotypes
NHANES 1999-2000, 2001-2002, 2005-2006, …, 2011-2012 (8)
Median N: 150-5000 per survey

~83,092 E-P associations!
significant associations (FDR < 5%)
adjusted by age, age^2, sex, race, income

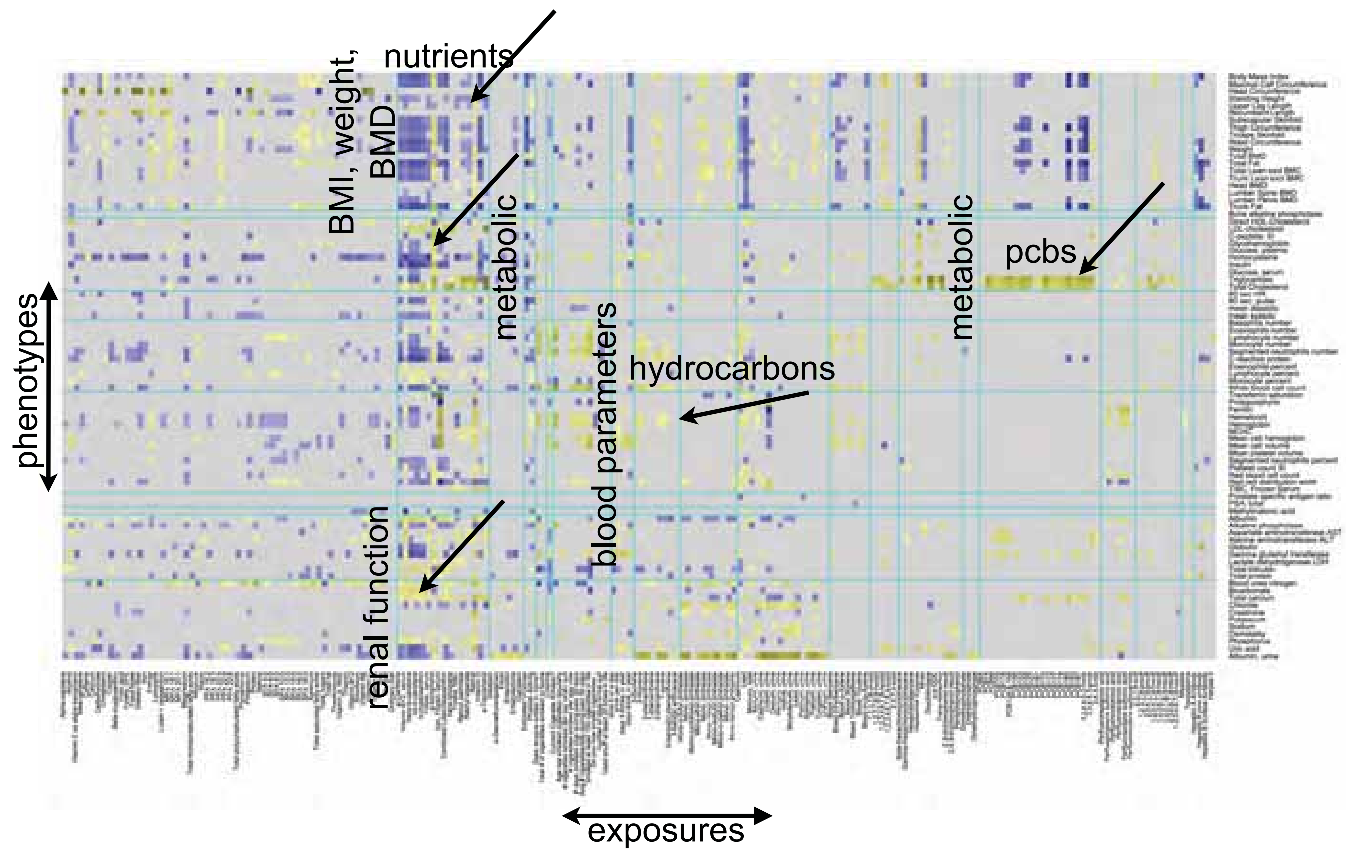
Raj Manrai, Hugues Aschard, JPA Ioannidis, Dennis Bier

83,092 total associations between ***E*** and ***P***
12,237 significant associations (**6%, in yellow**):
Average association size: <1% for 1SD change in ***E***

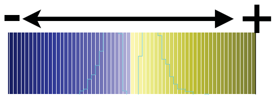




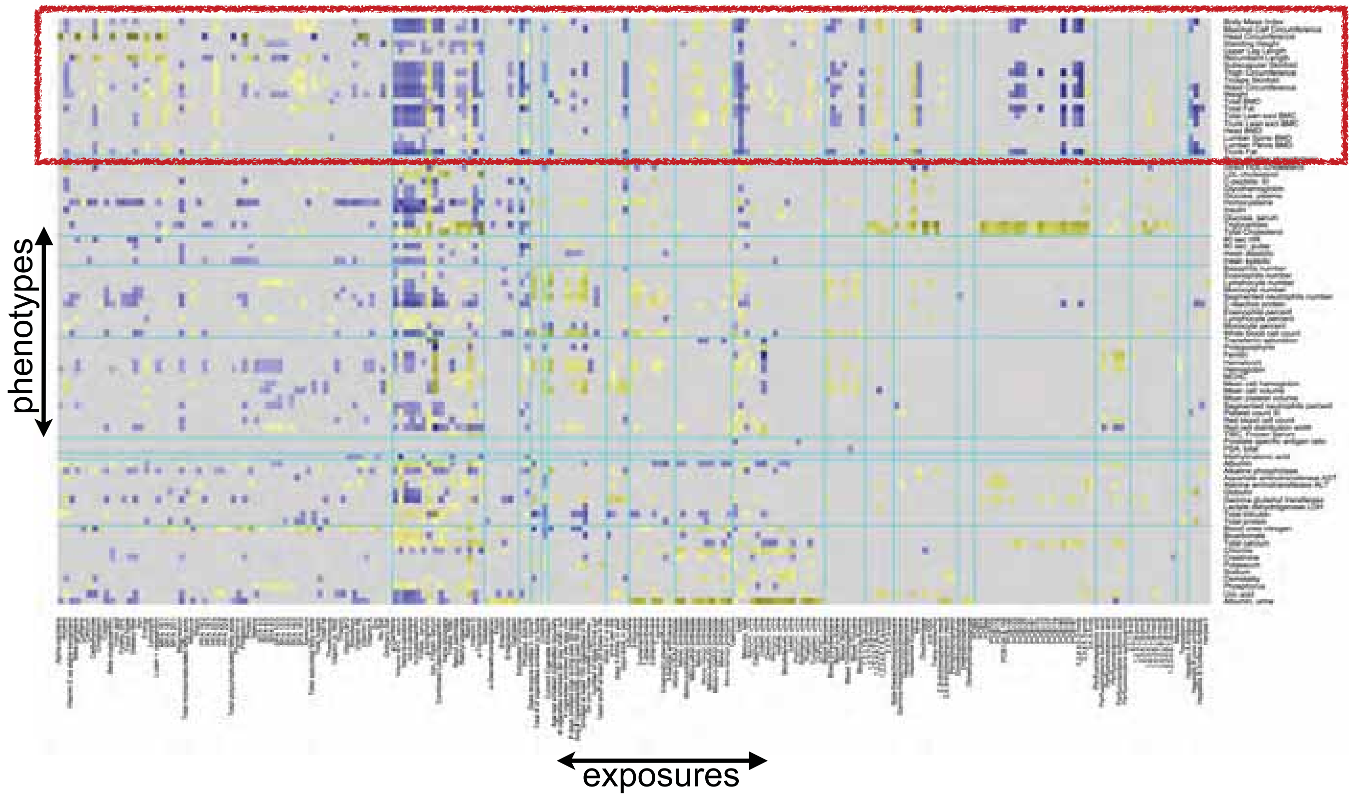
EWAS-derived phenotype-exposure association ***map***:
A 2-D view of connections between ***P and E***

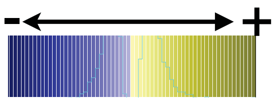


R^2: ~1-40% (average of 20%)

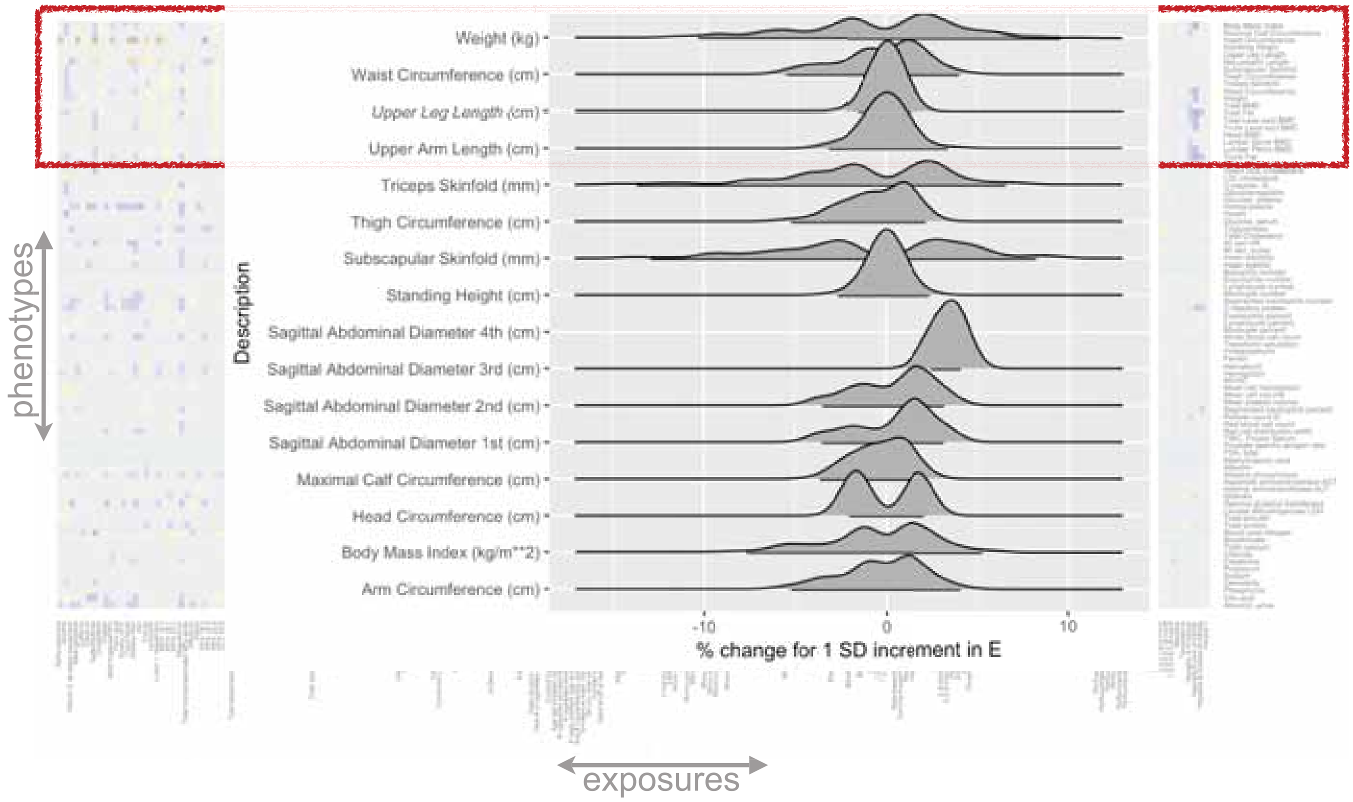


***EWAS*-derived phenotype-exposure association *map*:**
A 2-D view of connections between ***P* *and* *E*:**
does my correlation matter?

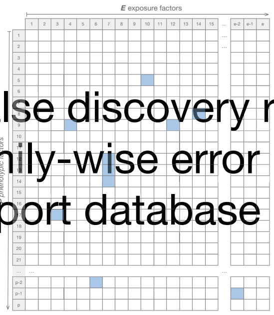




***EWAS*-derived phenotype-exposure association *map*:**
A 2-D view of connections between ***P* *and* *E*:**
does my correlation matter?

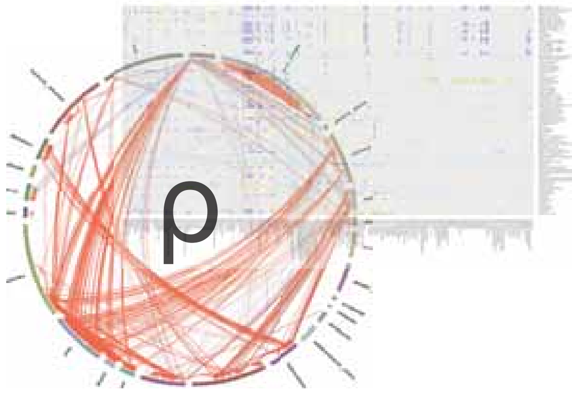


High-throughput data analytics to mitigate analytical challenges of exposome-based research:
Consider ***multiplicity of hypotheses*** and ***correlational web***!



Explicit in number of hypotheses tested

False discovery rate;
family-wise error rate;
Report database size!



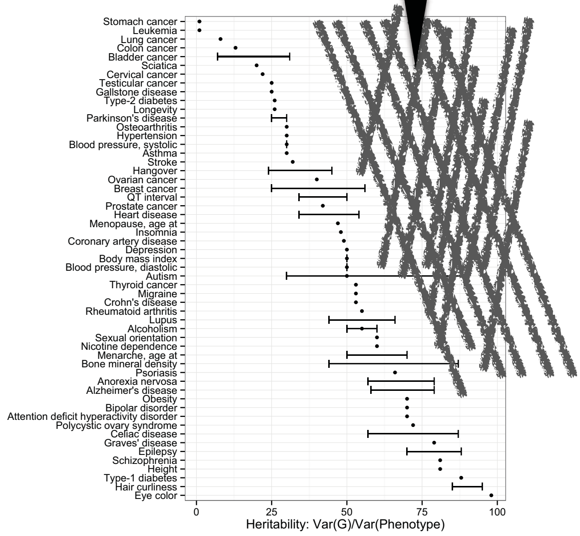
Does my correlation matter?
How does my new correlation compare to the family of correlations?
What is the total variance explained(σ^2_E)?

saturated fatty acids and BMI: 0.5%
does it matter? (i.e., 1.2% is average!)

ARPH 2016
JAMA 2014
JECH 2015

Bottom line: high-throughput ***E*** research will enable ***discovery to explain missing variation in P!***

σ^2_E : Exposome!



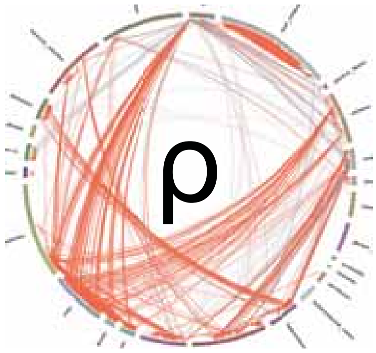
1.) Find elusive E in P and explain variation of disease risk

2.) Consideration of totality of evidence:
Does my correlation matter?

3.) Reproducible research and increase data literacy.

Bottom line: high-throughput ***E*** research will enable
discovery to explain missing variation in P!

1.) Find elusive E in P and explain variation of disease risk



2.) Consideration of totality of evidence: Does my correlation matter?

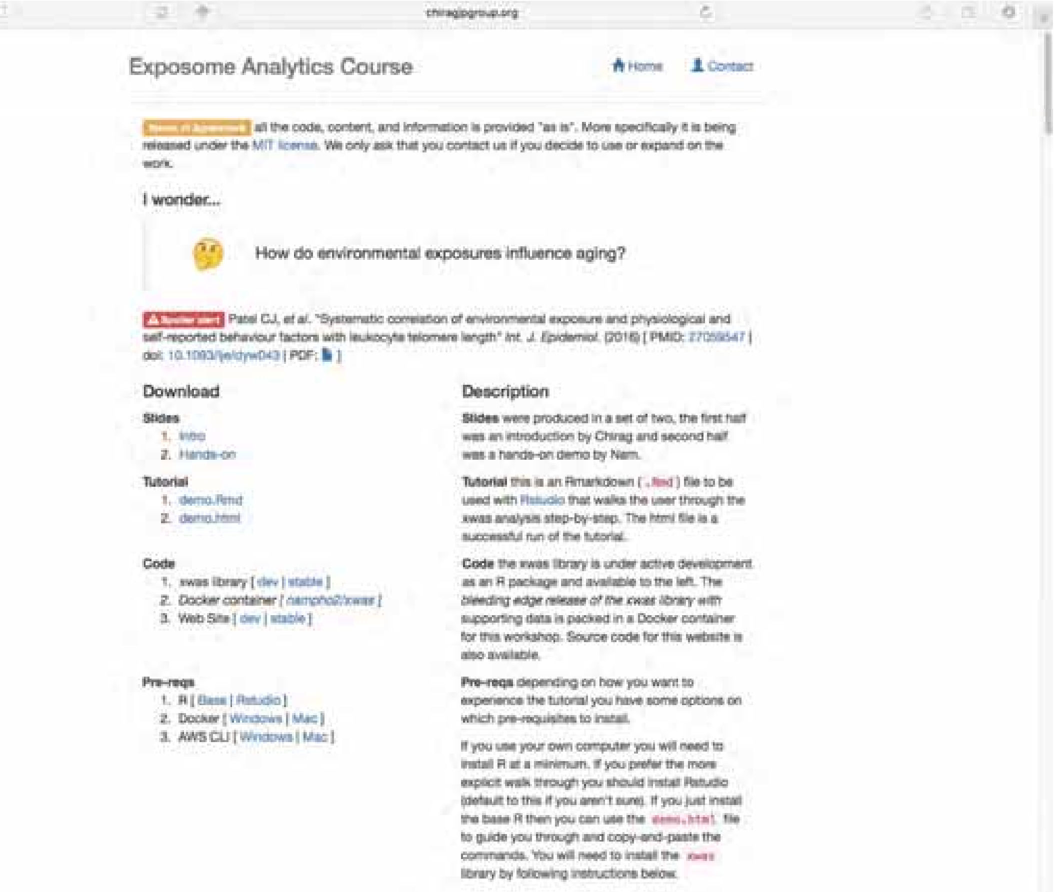
3.) Reproducible research and increase data literacy.

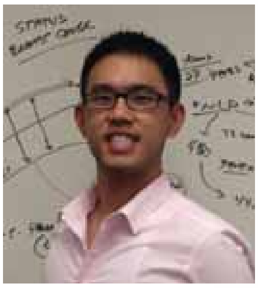
Bottom line: high-throughput ***E*** research will enable
discovery to explain missing variation in P!

1.) Find elusive E in P and explain variation of disease risk

2.) Consideration of totality of evidence: Does my correlation matter?

3.) Reproducible research and increase data literacy.





Nam Pho

Please contact me for help or project ideas!

http://chiragjpgroup.org/exposome-analytics-course

Designing a ***new*** children's study:

(1) Increase sample sizes and make data publicly available

(2) Measure ***G*** to discover role of ***E*** in ***P***

Designing a ***new*** children's study
(1) Increase sample sizes and make data ***publicly available***

biobank uk — Improving the health of future generations

N=500,000

China Kadoorie Biobank — 中国慢性病前瞻性研究

N=500,000

Generate wide interest and visibility
Enhance reproducibility (decrease false positives)

Designing a ***new*** children's study:
(2) Measure ***G*** to discover role of ***E*** in ***P***

The environmental contribution to gene expression profiles

Greg Gibson

Gibson, G. *Nature Reviews Genetics* 2008

biological function

Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies

Chirag J. Patel, Jacqueline Kerr, Duncan C. Thomas, Bhramar Mukherjee, Beate Ritz, Nilanjan Chatterjee, Marta M Jankowska, Juliette Madan, Margaret R. Karagas, Kimberly A McAllister, Leah E. Mechanic, M. Daniele Fallin, Christine Ladd-Acosta, Ian A Blair, Susan L Teitelbaum, and Christopher I. Amos

DOI: 10.1158/1055-9965.EPI-17-0459

Patel CJ et al, *CEBP* 2017

gene-by-environment interactions

VIEWPOINTS

G = E: What GWAS Can Tell Us about the Environment

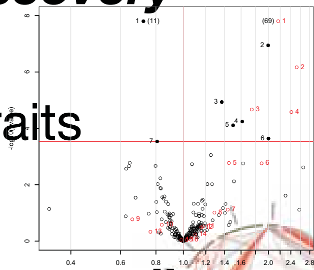
Suzanne H. Gage[1,2], George Davey Smith[1,3], Jennifer J. Ware[1,3], Jonathan Flint[4], Marcus R. Munafò[1,2]*

Gage S et al. *PLoS Genetics* 2016

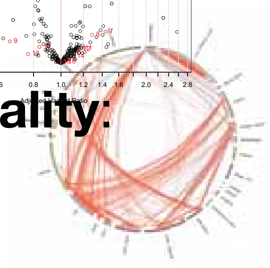
GWAS and mendelian randomization

In conclusion:

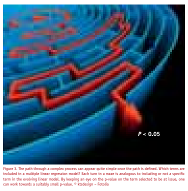
Data science inspired approaches to ascertain ***exposome*** *and* ***genome*** will enable biomedical ***discovery***



EWASs in aging: mortality and quantitative traits



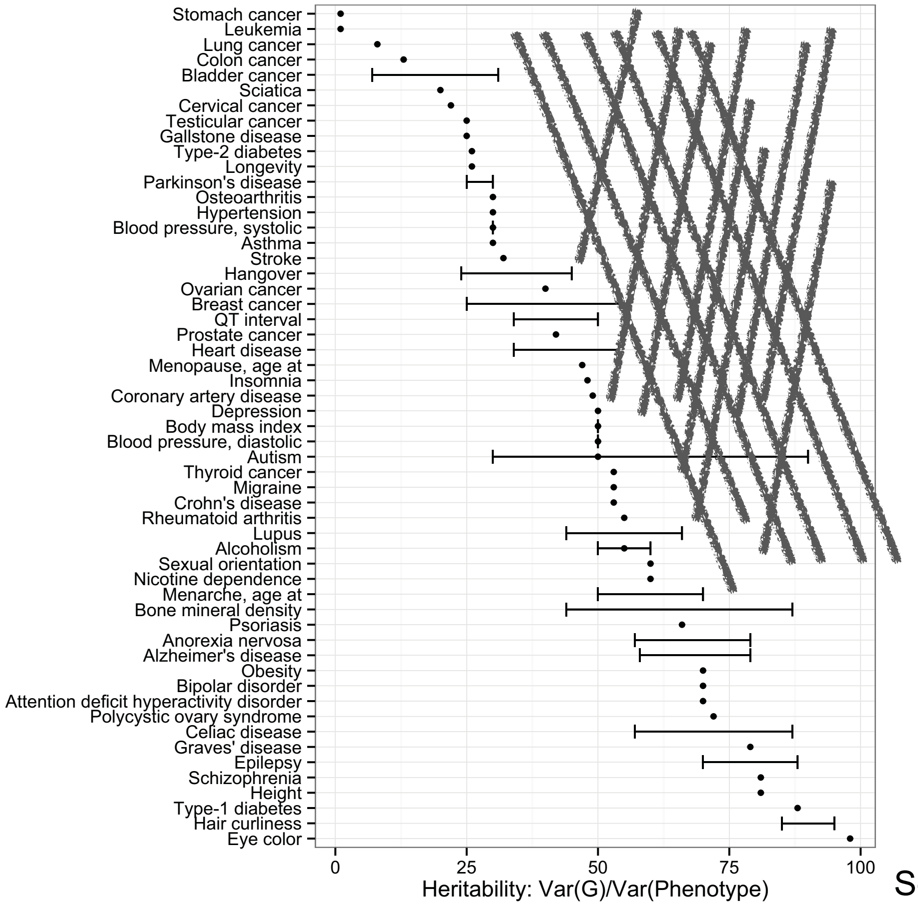
Dense **correlations, confounding, reverse causality**: how to assess at high dimension?



Mitigate **fragmented** literature of associations.

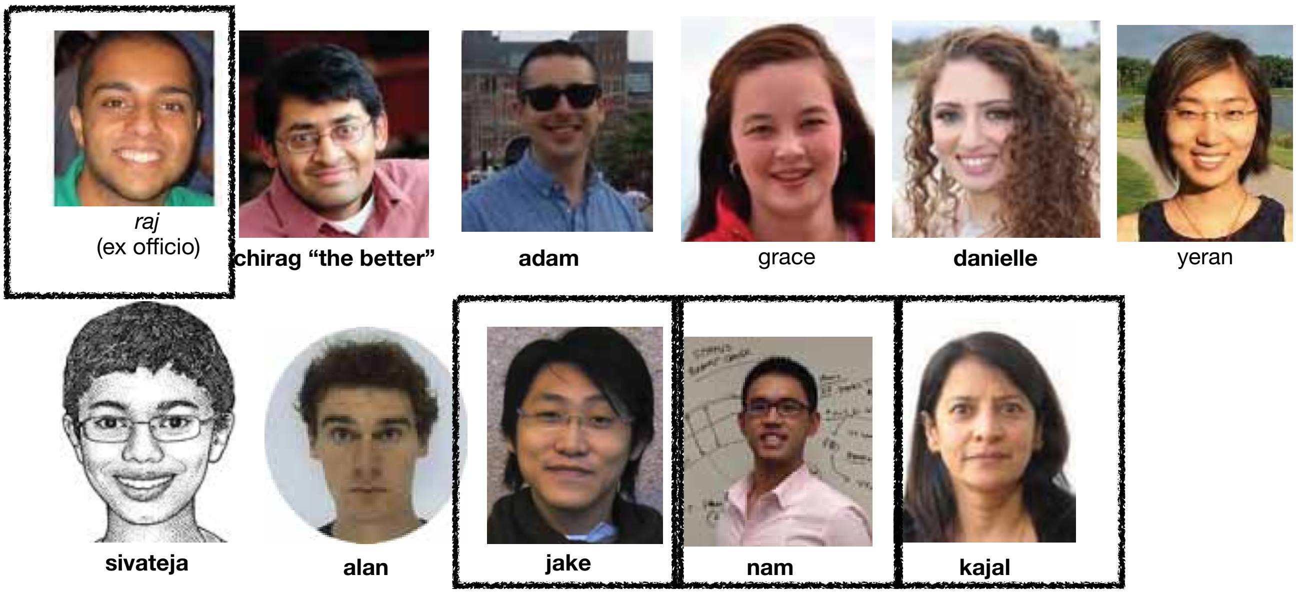
Understand interacting ***G*** and ***E*** for causation

Use high-throughput tools and data (e.g., exposome) will enhance discovery of the role of ***E*** (and ***G***) in ***P***.



Source: SNPedia.com

RagGroup Data Science Team:
2 post-docs, **3** PhD, **2** MS, **1** HS, **2** visiting



PhD: systems biology, integrative genomics
MS: statistics (HSPH)
Post-docs: biology, medicine, and mathematics

Acknowledgements

RagGroup
Nam Pho
Jake Chung
Kajal Claypool
Arjun Manrai
Chirag Lakhani
Adam Brown
Danielle Rasooly
Alan LeGoallec
Sivateja Tangirala

Harvard DBMI
Susanne Churchill
Nathan Palmer
Sophia Mamousette
Sunny Alvear
Michal Preminger

Mentioned Collaborators
Isaac Kohane
John Ioannidis
Dennis Bier
Hugo Aschard

IEA-WCE 2017 symposium
Shoji Nakayama
Junya Kasamatsu
Ministry of Environment (Japan)

NIEHS National Institute of Environmental Health Sciences
PhRMA
NSF
National Institute of Allergy and Infectious Diseases
Agilent Technologies

NIH Common Fund
Big Data to Knowledge

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

HARVARD MEDICAL SCHOOL
DEPARTMENT OF Biomedical Informatics