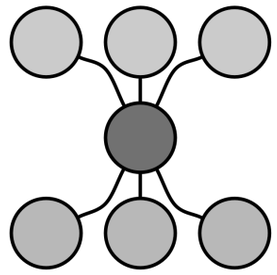


Informatics and **data analytics** to support ***exposome-based discovery***

Perspectives from a NIEHS workshop

Chirag J Patel

International Society of Exposure Science

Henderson, NV (by way of Boston, MA)

10/20/15

HARVARD MEDICAL SCHOOL

DEPARTMENT OF Biomedical Informatics

chirag@hms.harvard.edu

@chiragjp

www.chiragjpgroup.org

The ***workgroup*** discussed *informatics* capability for high-throughput ***exposome*** research
(late 2014 to early 2015)

Arjun Manrai (Harvard)*
Yuxia Cui (NIEHS)
Pierre Bushel (NIEHS)
Molly Hall (Penn State, now U Penn)*
Spyros Karakitsios (Aristotle U, Greece)
Carolyn Mattingly (NCSU)
Marylyn Ritchie (Geisinger Health/Penn State)
Charles Schmitt (NIEHS)
Denis Sarigiannis (Aristotle U, Greece)
Duncan Thomas (USC)
David Wishart (U Alberta, Canada)
David Balshaw (NIEHS)

We are now in the era of ***high-throughput*** **biology** and **biomedicine**.

(now possible to assay ***thousands*** to ***millions*** of datapoints today)

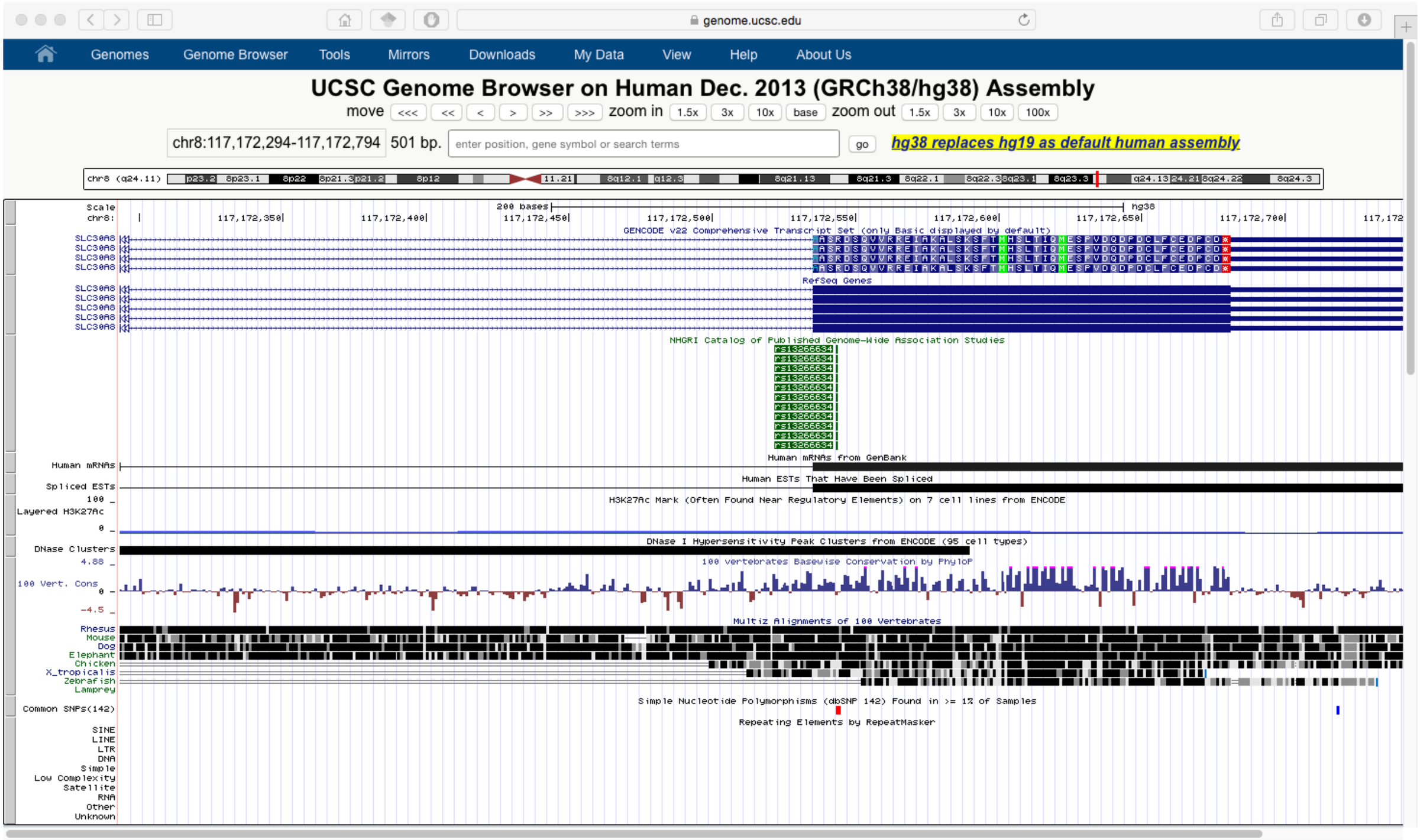
We are now in the era of ***high-throughput*** biology and biomedicine: examples of ***genomic*** advances

genetic arrays gene expression common genetic variants epigenome (methylation)	***3-4 x 10^4 genes*** ***10^6 to 10^7 variants***
whole genome sequencing (WGS) full genome sequencing mRNA-seq epigenome (3D, histone)	***3 x 10^9 nucleotide bases***

Informatics has enabled ***discovery*** in ***genomics*** investigations.

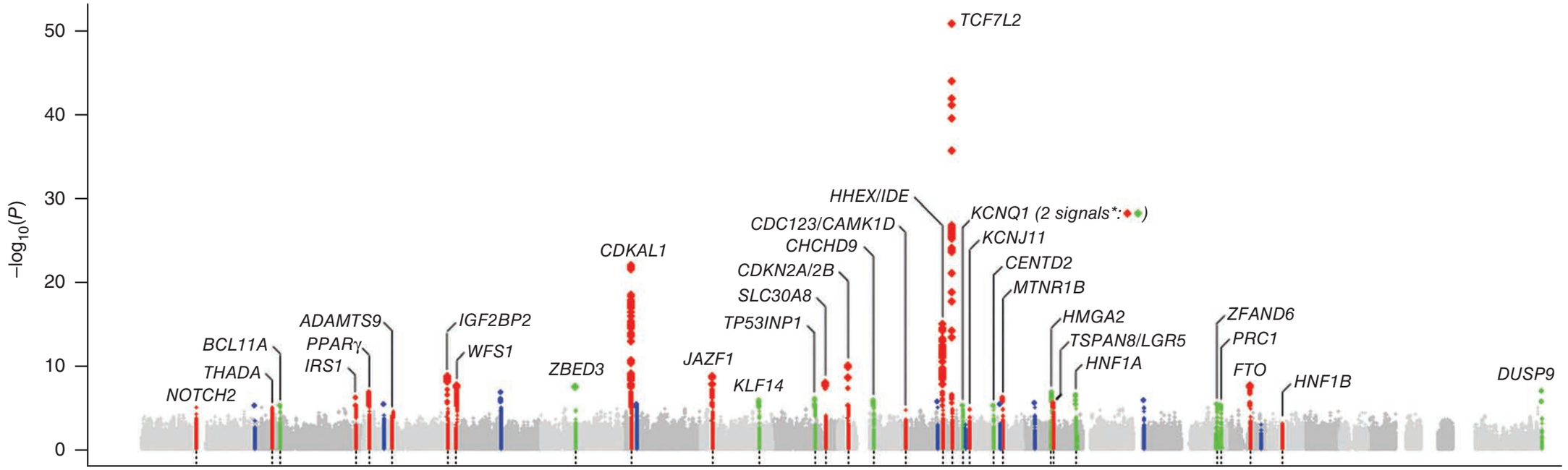
1. ***infrastructure/standards,***
2. ***analytics,***
3. ***databases***

Information infrastructure has enabled discovery in ***genomics*** (example: UCSC genome browser)



Analytic methods have enabled discovery in ***genomics***
(example: genome-wide association [**GWAS**])

GWAS in Type 2 Diabetes



Voight et al, Nature Genetics 2012

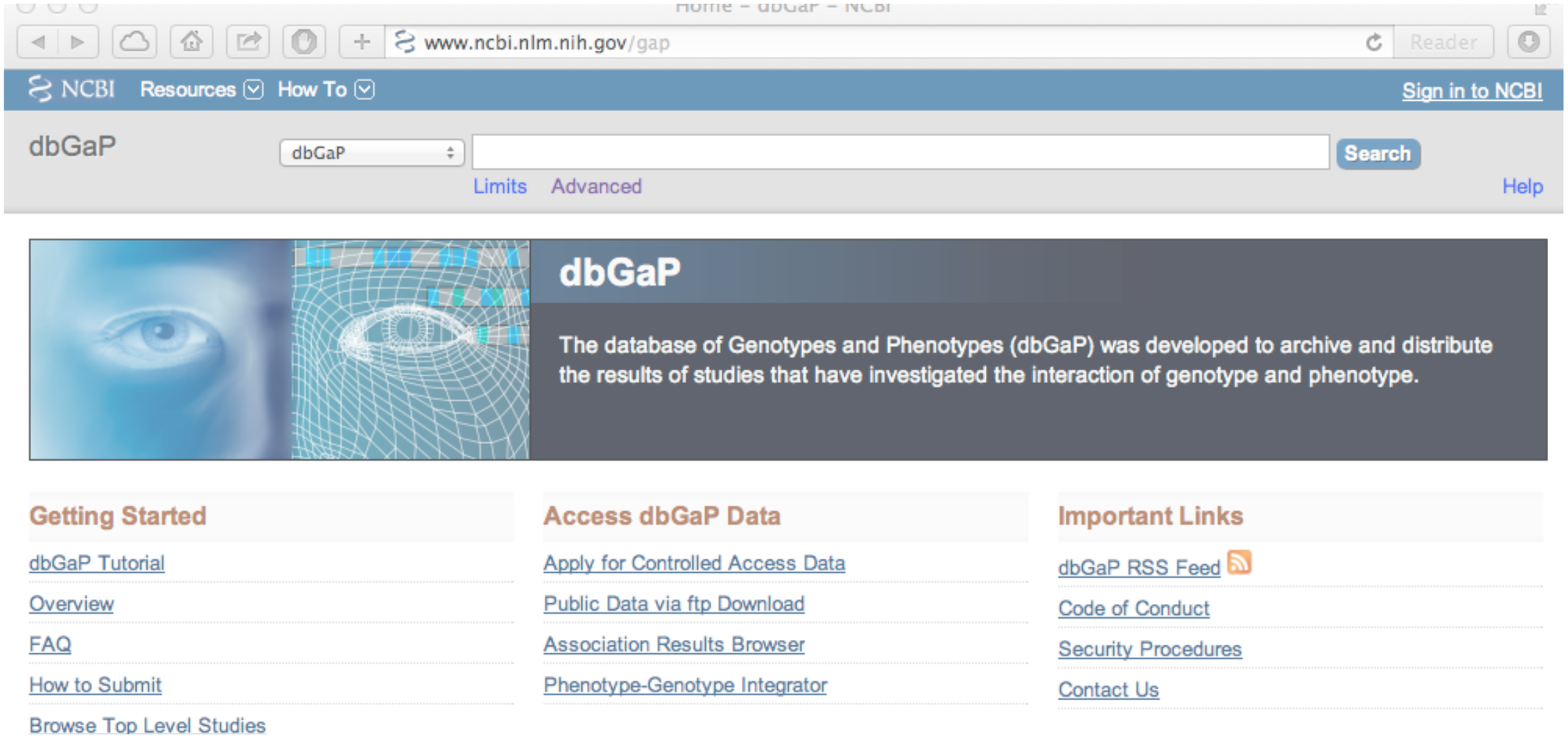
N=8K T2D, 39K Controls

A search engine for genetic influence in phenotypes

Genome-wide association studies (GWASs)

Accessible data repositories have enabled discovery in ***genomics*** investigation:

(ex: Databases of Genotypes and Phenotypes)



758,000 individuals

>400 studies

>>1B datapoints (genotypes and phenotypes)

>950 manuscripts (Paltoo et al., *Nature Genetics* 2014)

We claim that there is need for informatics **analytic methods**, ***databases***, and ***standards*** for the **exposome**-driven discovery.

EWAS *akin to* ***GWAS****?*

courtesy: colabria.com

Why?

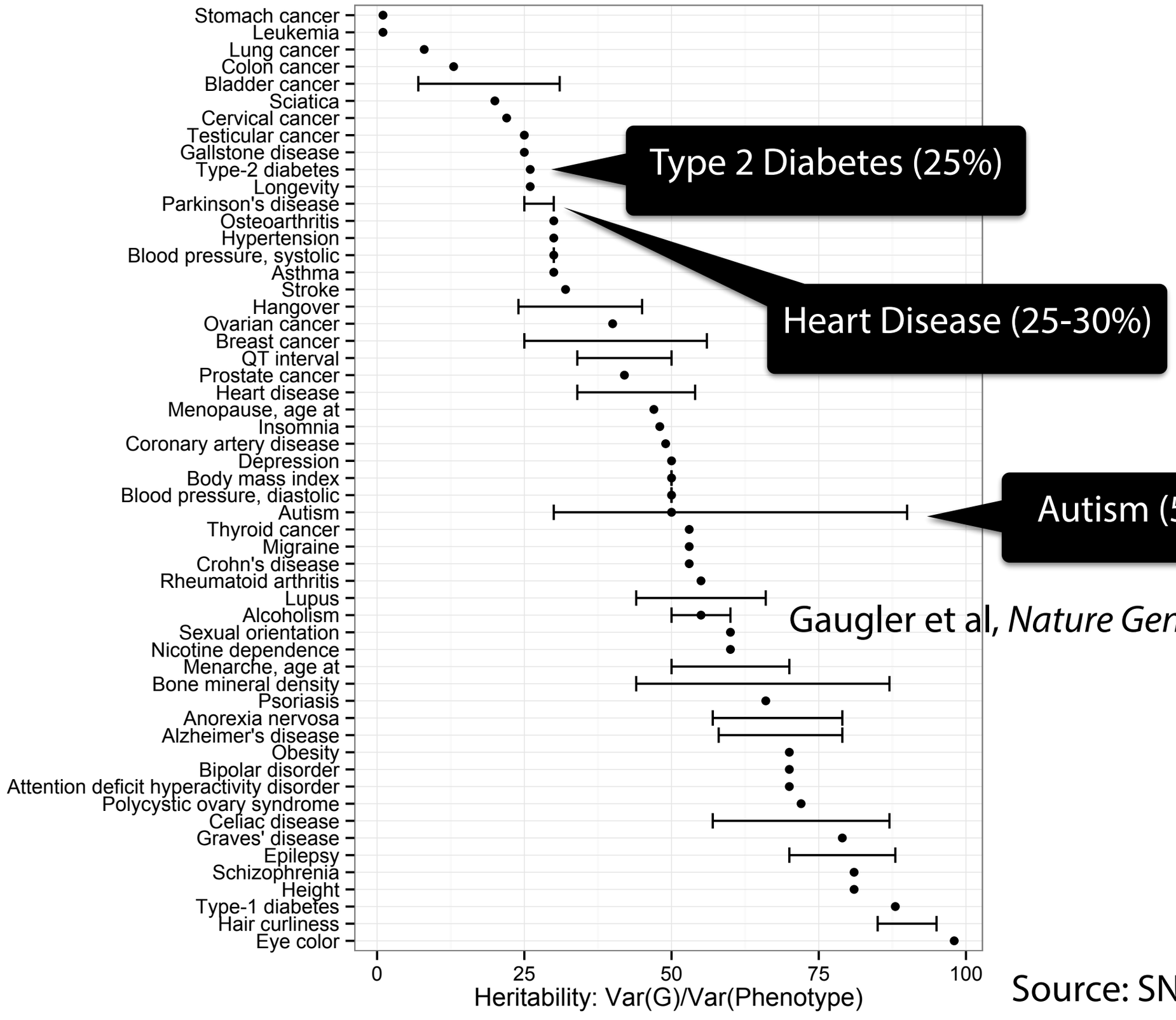
$$P = G + E$$

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

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

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

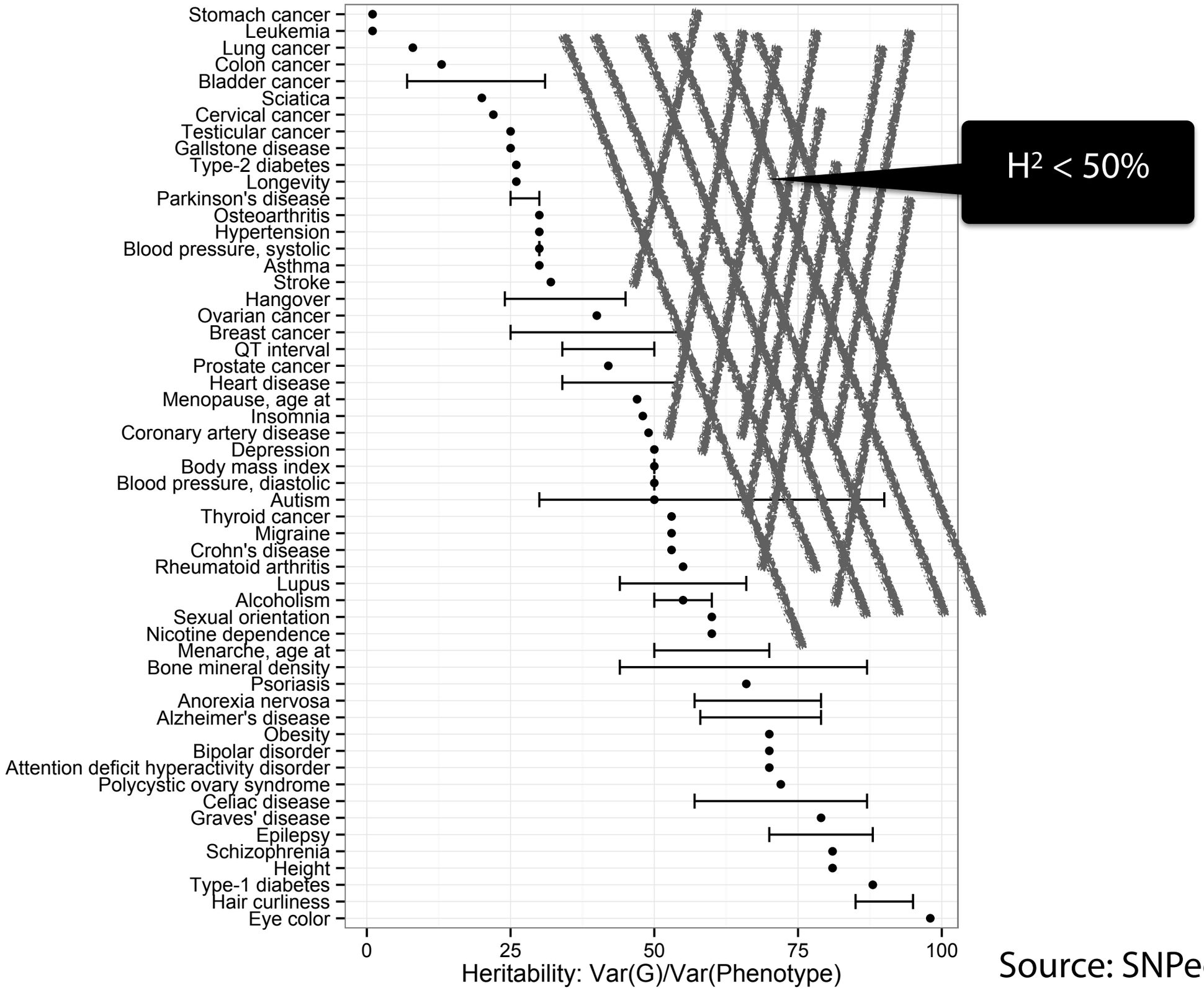
H^2 estimates for complex traits are **low and variable:** massive opportunity for *high-throughput **E*** research



Gaugler et al, *Nature Genetics* (2014)

Source: SNPedia.com

H^2 estimates for complex traits are **low and variable:** massive opportunity for *high-throughput **E*** research



Source: SNPedia.com

Meta-analysis of the heritability of human traits based on fifty years of twin studies

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

Nature Genetics, 2015

17,804 traits of the ***phenome***

2,748 publications

14,558,903 twin pairs

Average ***H² (genome)***: ***0.49***

Exposome plays an equal role.

What is the potential chemical (external and internal) space of the ***exposome?*:** perhaps on the order of ***thousands***.

toxicants drugs

3,600 + 1,634

Toxic Exposome Database
Wishart et al (2015)

>13,000

Davis et al
Comparative Tox DB (2015)

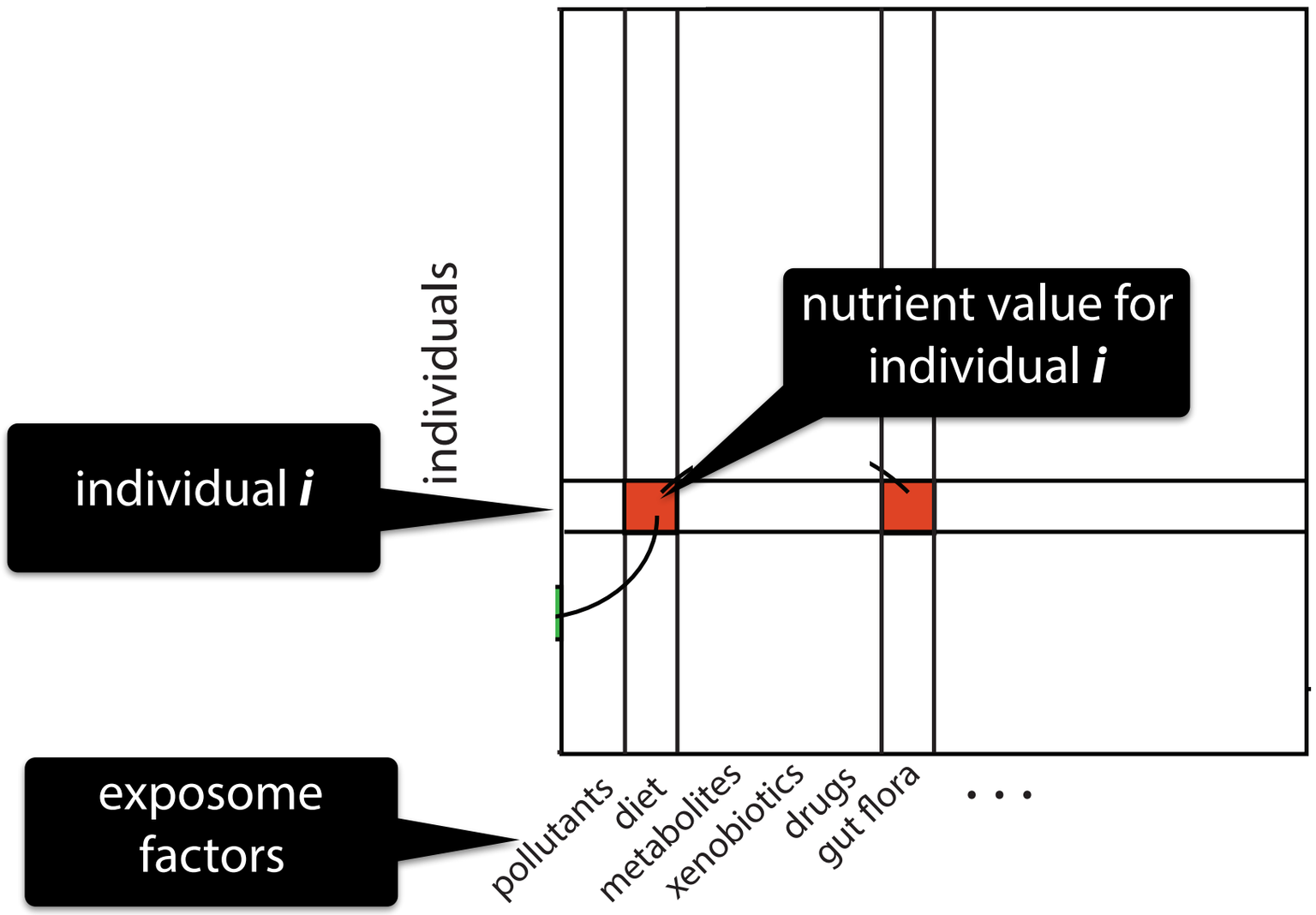
100-1,000?

uBiome

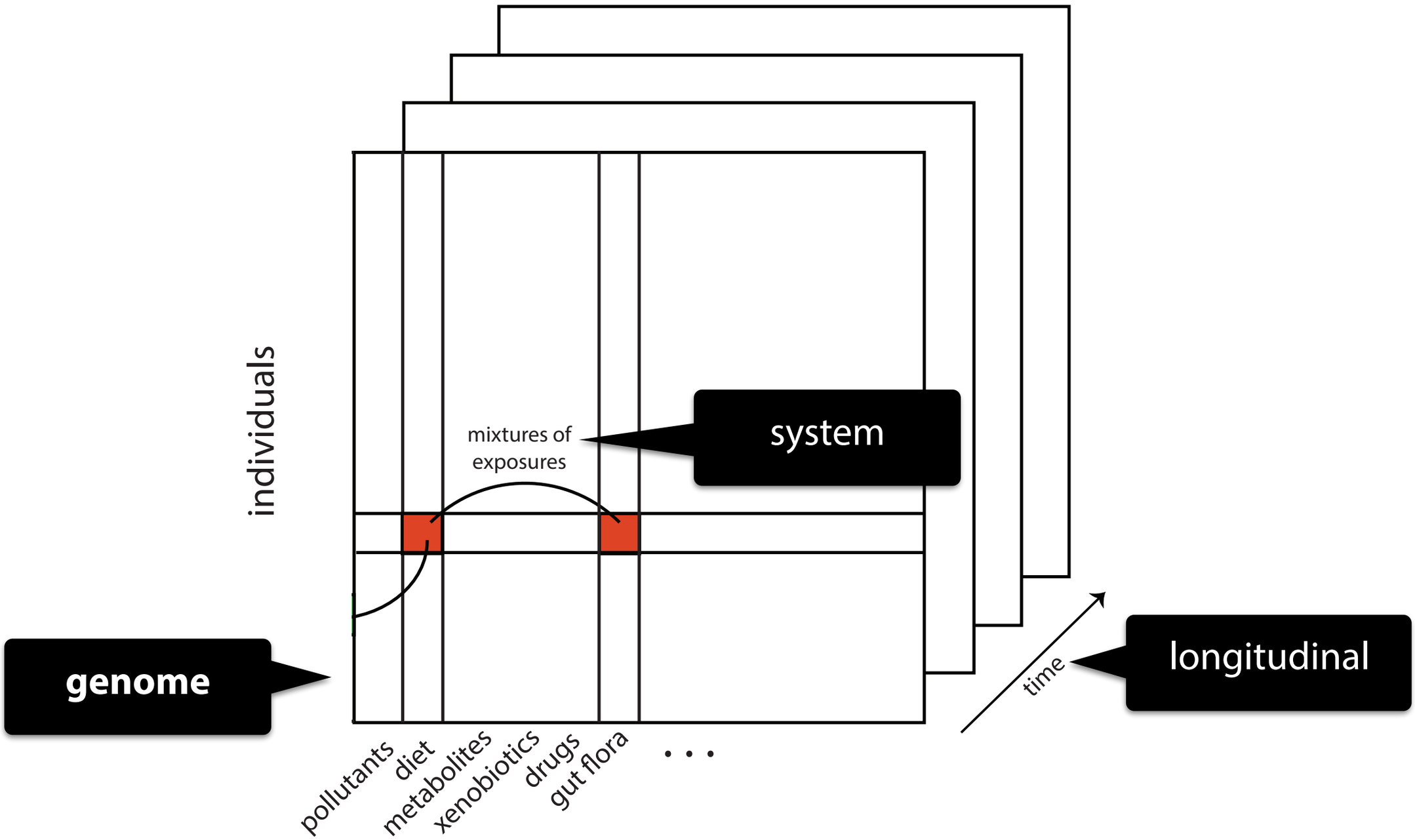
>84,000

TSCA and EPA Inventory
(2014)

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***

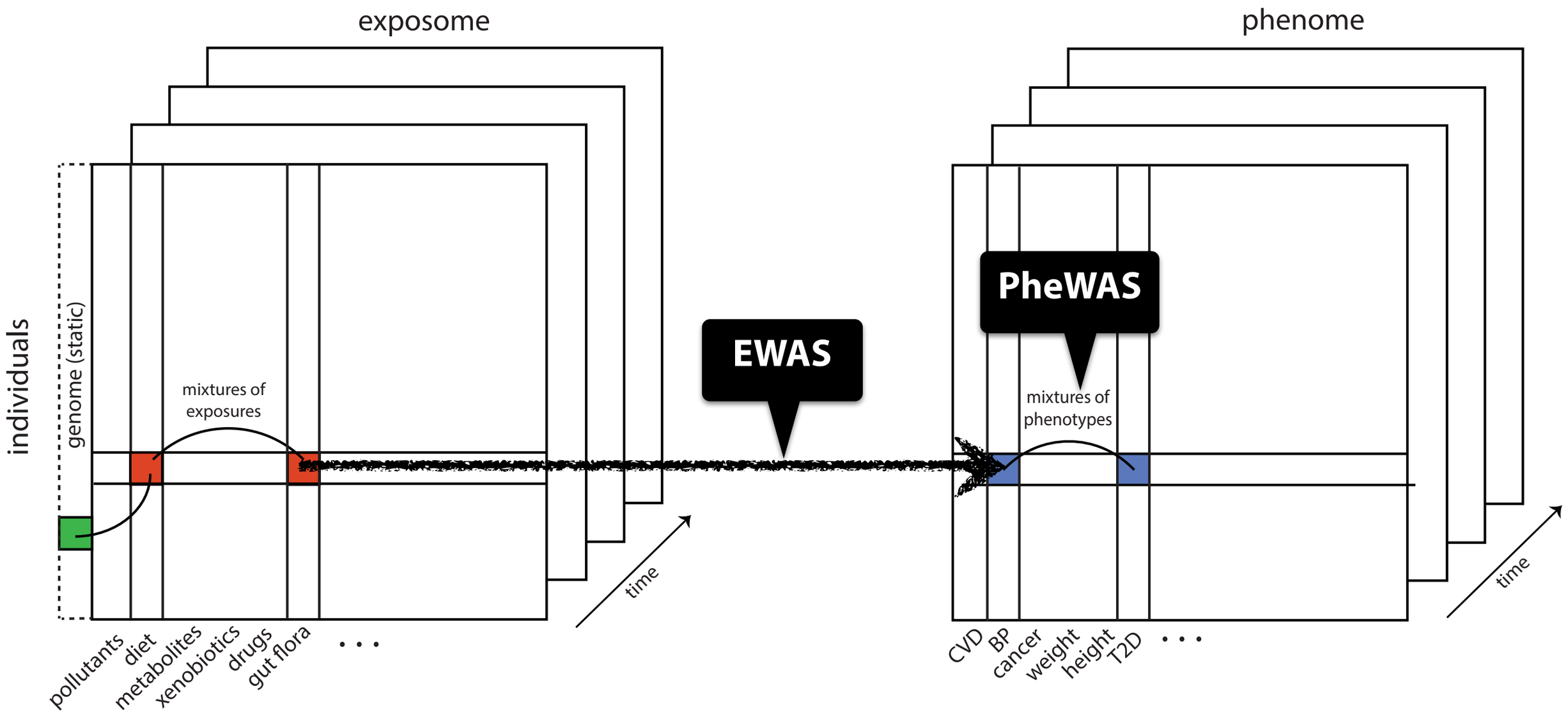


What will the ***exposome*** data structure look like?:
a ***high-dimensioned*** 3D matrix of (1) ***exposure*** measurements
on (2) ***individuals*** as a function of (3) ***time***



Data-driven investigation for novel ***exposome*** factors in the ***phenome:***

Exposome-wide, phenome-wide, and genome-exposome-wide discovery



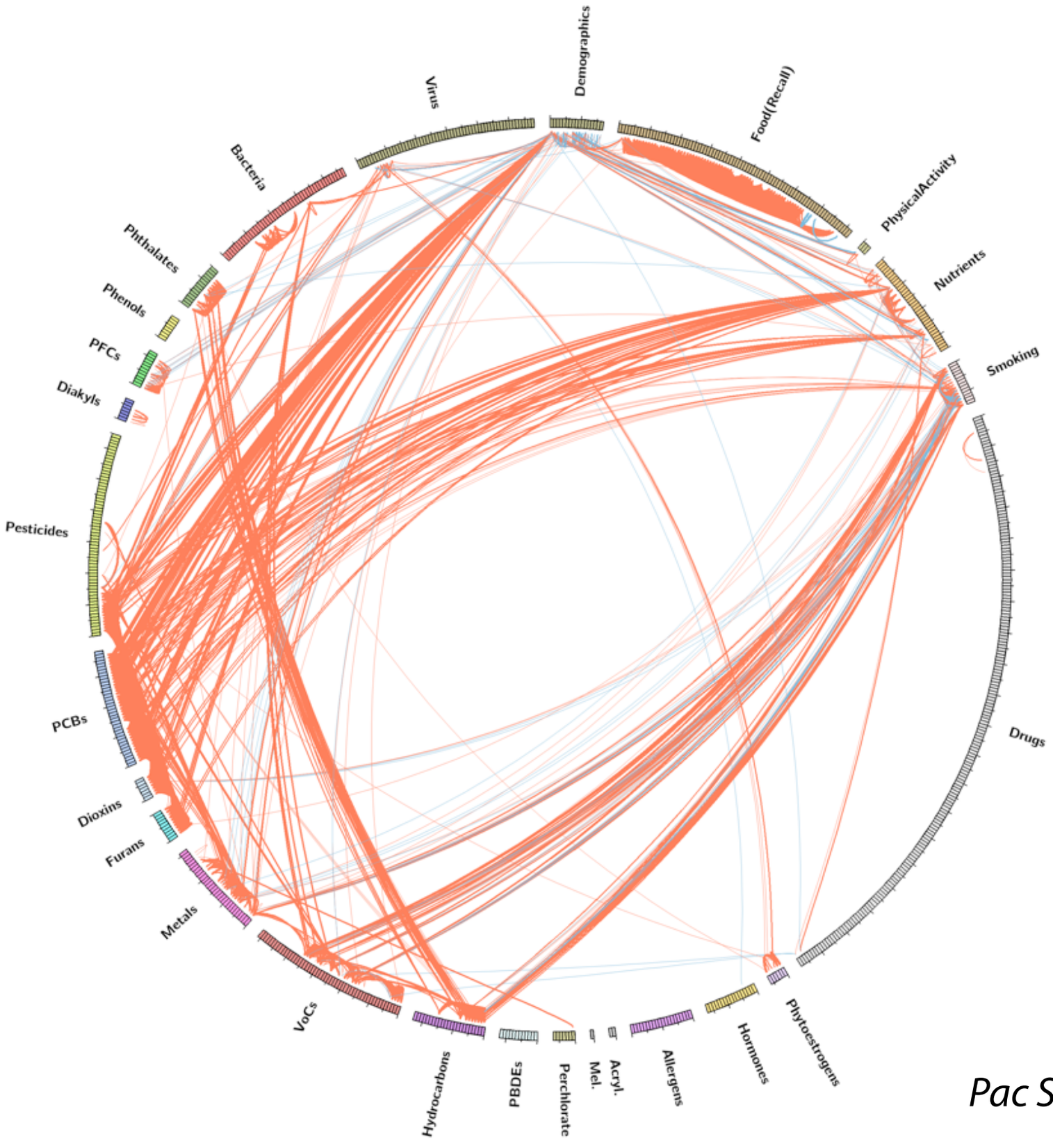
Informatics methods to ***integrate*** heterogeneous data (***E, G,*** and ***P***)
and to conduct ***EWAS, GxEWAS, and PheWAS***

Integration challenges in conducting
data-driven investigation for novel ***exposome*** factors in the ***phenome:***
The **exposome** is **heterogenous** and **G does not equal E.**

platform
scale
time-dependent
type
correlation

mass-spec: ***targeted*** **vs.** ***untargeted***
external **vs.** ***internal***
sampling and life trajectories
continuous vs. categorical
dense!

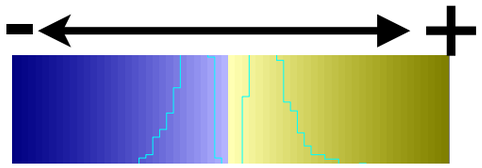
Interdependencies of the ***exposome***:
Correlation globes paint a dense and complex view of exposure



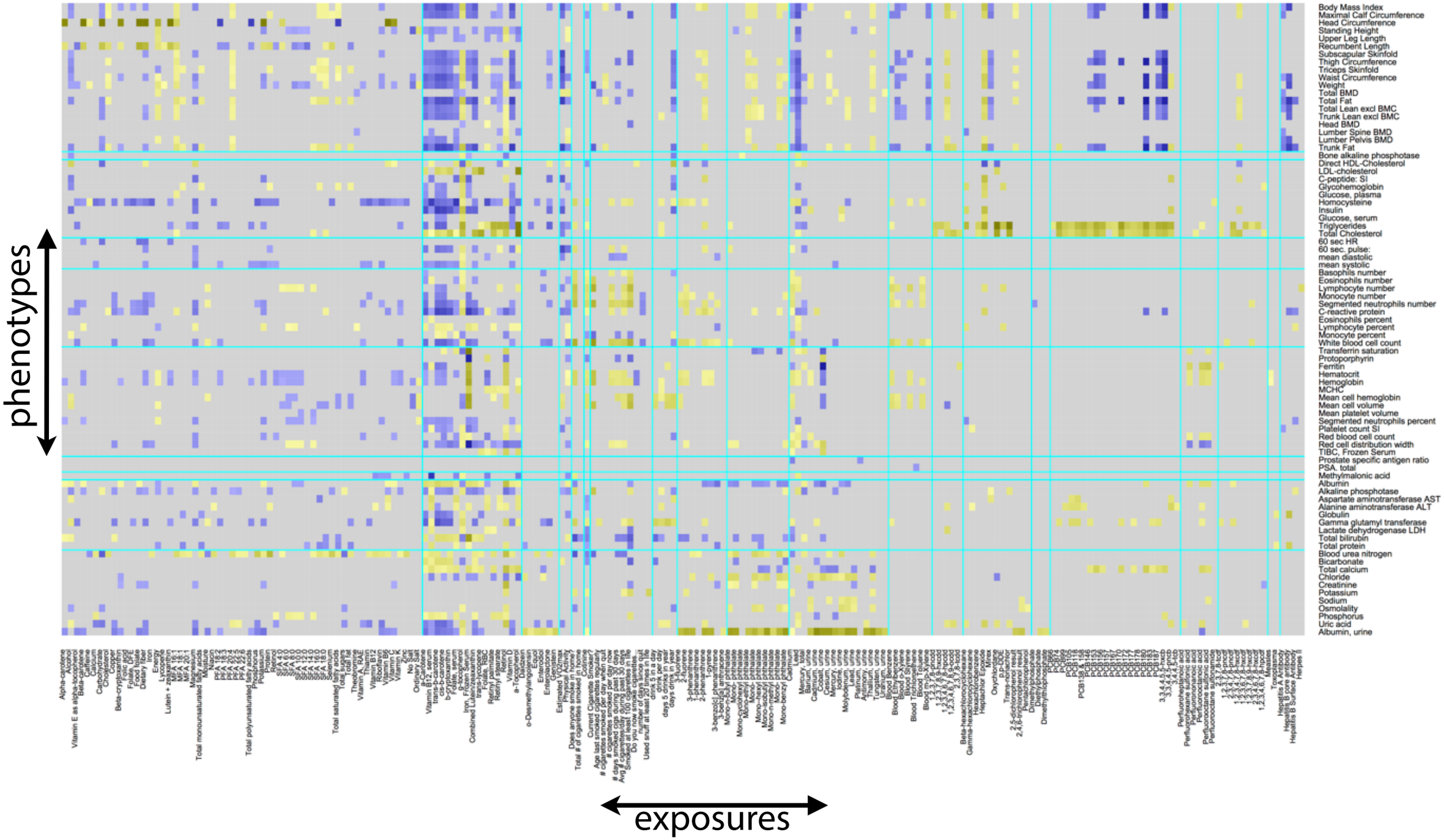
JAMA 2015
Pac Symp Biocomput. 2015

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

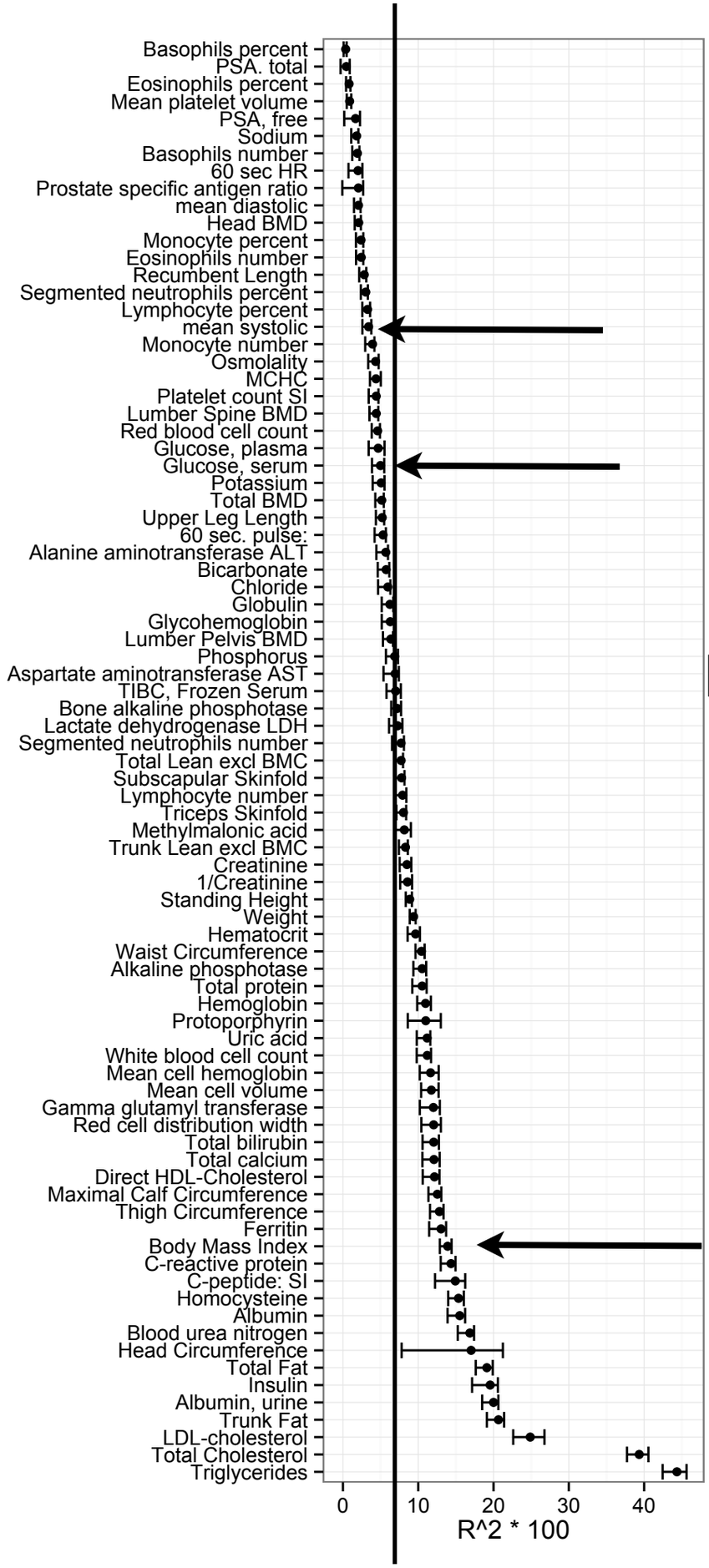
$$\sigma^2_E \; ???$$



EWAS-derived phenotype-exposure association *map*:
A 2-D view of 86 phenotype by 252 exposure associations



http://bit.ly/pemap



σ^2_E?

1 to 66 exposures identified for 81 phenotypes

Additive effect of ***E*** factors:
Describe less than 10% of variability in ***P***

(On average: 8%)

Recall: H^2 <= 50%

Exposome may enable realization of remainder of ***P*** (>40%)

Stan Shaw, Hugues Aschard, JP Ioannidis

What do we do now?

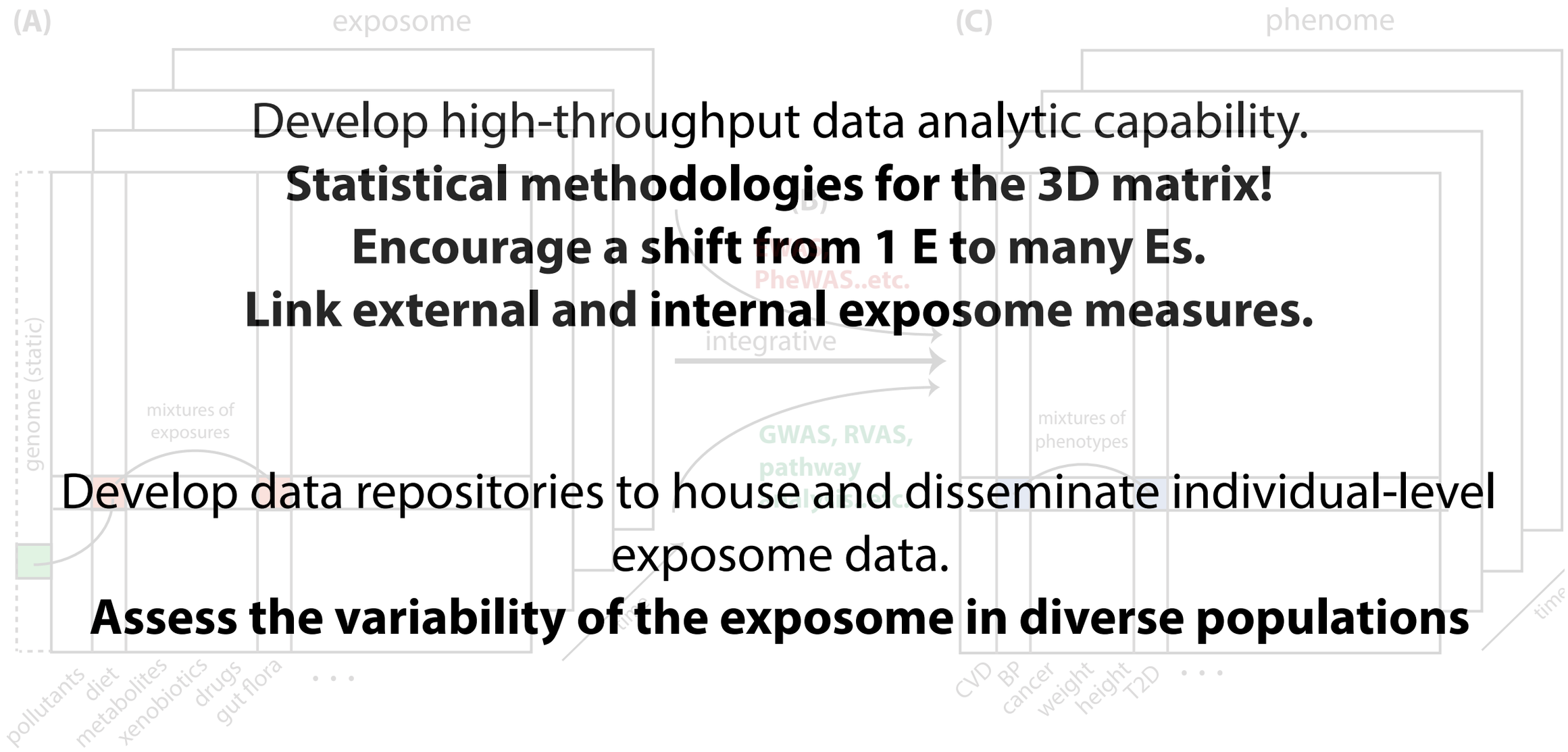
Recommendations from the workgroup

Data workgroup **recommendation** highlights

Comprehensive ***catalog*** of documented environmental associations (e.g., risk, variance explained) to strengthen case for ***exposome.***
Where is evidence robust (e.g., air pollution and CVD)?
Where do we see non-replication?
Where is *heritability* low and ripe for exposome?

Identify technologies that can measure the ***exposome.***
Targeted and untargeted metabolomics.

Data workgroup **recommendation** highlights



Develop high-throughput data analytic capability.

Statistical methodologies for the 3D matrix!

Encourage a shift from 1 E to many Es.

Link external and internal exposome measures.

Develop data repositories to house and disseminate individual-level exposome data.

Assess the variability of the exposome in diverse populations

Data workgroup **recommendation** highlights

Identify data standards for exposome research.
Develop data standards to enable the re-use of research to build large exposome-rich cohorts.

Identify analytics standards for reproducible research.
Software libraries and tools to share methods and findings.

Incentivize other parties (e.g., researchers, funders, and industry) to integrate the **exposome** in their existing programs.

Data workgroup **recommendation** highlights

Educate.

Identify example datasets (e.g., **NHANES**, **DEMOCOPHES**).
Hackathons and **challenges** to recruit data scientists.
Develop big **data training support** (e.g., K awards) directed at **exposome-related research**

niehs.nih.gov

National Institute of Environmental Health Sciences
Your Environment. Your Health.

GO

Health & Education | Research | Funding Opportunities | Careers & Training | News | About NIEHS

Research

- Funded by NIEHS Grants
  - About the Extramural Research and Training Division
    - Research Programs
      - Autism Research
      - Biphenol A (BPA) Research Program
      - Breast Cancer & the Environment Research Program
      - Centers for Children's Environmental Health & Disease Prevention Research
      - Centers for Neurodegeneration Science
      - **Children's Health Exposure Analysis Resource (CHEAR)**
        - Grantees
      - Children's Environmental Health Supplements

Children's Health Exposure Analysis Resource (CHEAR)

Children's health and wellbeing are influenced by interactions between environmental and genetic factors. NIEHS is establishing an infrastructure, the Children's Health Exposure Analysis Resource (CHEAR), to provide the extramural research community access to laboratory and data analyses to add or expand the inclusion of environmental exposures in their children's health research. The goal of CHEAR is to provide the tools for researchers to assess the full array of environmental exposures which may affect children's health. We anticipate CHEAR will be used by children's health researchers conducting epidemiological or clinical studies that currently have very limited consideration of environment, or those who have collected exposure data but seek more extensive analyses.

CHEAR is intended to:

- Expand the number of studies that include environmental exposure analysis in their studies,
- Implement the exposome concept in children's health studies,
- Create a public resource of children's exposures across the country, and
- Develop data and metadata standards for the environmental health sciences community.

NIH Director Francis S. Collins, M.D., Ph.D.

"Technology advances have become a powerful driver in studying and understanding the start and spread of disease," said NIH Director Francis S. Collins, M.D., Ph.D. "These projects will expand the toolbox available to researchers to improve our ability to characterize environmental exposures, understand how environmental exposures affect in utero development and function, and bolster the infrastructure for exposure research."

Share This Page: Page Options: Request Translation Services

google: “niehs chear”

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

Informatics will enable us to decipher the role of the emerging ***exposome*** in phenotypes to capture the missing σ^2_P

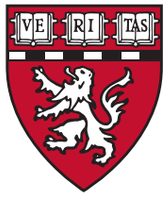
Thanks again to the group:

Arjun Manrai (Harvard)*
Yuxia Cui (NIEHS)
Pierre Bushel (NIEHS)
Molly Hall (Penn State, now Penn)*
Spyros Karakitsios (Aristotle U, Greece)
Carolyn Mattingly (NCSU)
Marylyn Ritchie (Geisinger/Penn State)
Charles Schmitt (NIEHS)
Denis Sarigiannis (Aristotle U, Greece)
Duncan Thomas (USC)
David Wishart (U Alberta, Canada)
David Balshaw (NIEHS)

Funded in part by the NIEHS.

Thank you.

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



HARVARD
MEDICAL SCHOOL

DEPARTMENT OF
Biomedical Informatics