Using exposome-wide association studies (EWAS) to discover causes of cancer

S.M. Rappaport

University of California, Berkeley

Research support from NIEHS

CEB Center For Exposure Biology

Some background

- About three fourths of all people die from chronic diseases, mainly CVD and cancer
- These diseases likely result from a combination of genetic (G) and environmental (E) factors
- *But how much of the risk can be attributed to G, E and GxE?*

Explained variance of cancer incidence

(From structural equation modeling of the Swedish Family-Cancer database of 10M individuals born after 1934)

Site	Genetic	Shared environmental	Childhood environmental	Non-shared environmental
Stomach	0.01	0.15	0.13	0.71
Colon	0.13	0.12	0.06	0.69
Rectum	0.12	0.09	0.03	0.75
Lung	0.08	0.09	0.04	0.79
Breast	0.25	0.09	0.06	0.60
Cervix (invasive)	0.22	0.00	0.03	0.75
Cervix (*in situ*)	0.13	0.00	0.13	0.74
Testis	0.25	0.00	0.17	0.58
Kidney	0.08	0.08	0.06	0.78
Bladder	0.07	0.12	0.04	0.77
Melanoma	0.21	0.02	0.08	0.69
Nervous system	0.13	0.05	0.02	0.80
Thyroid	0.53	0.01	0.10	0.36
Endocrine	0.28	0.03	0.11	0.58
Non-Hodgkin's lymphoma	0.10	0.06	0.02	0.83
Leukemia	0.01	0.08	0.04	0.88
Median	***0.13***	***0.07***	***0.06***	***0.75***

K Czene, P. Lichtenstein and K Hemminki, Int J Cancer 2002, 99: 260-6

Attributable risk

“The population attributable fraction (*PAF*) can be interpreted as *the proportion of disease cases over a specified time that would be prevented following elimination of the exposures*, assuming the exposures are causal.”

B Rockhill, B Newman and C Weinberg, AJPH, 1998, 88: 15-19.

Familial risks of cancer

(From Swedish Family-Cancer database)

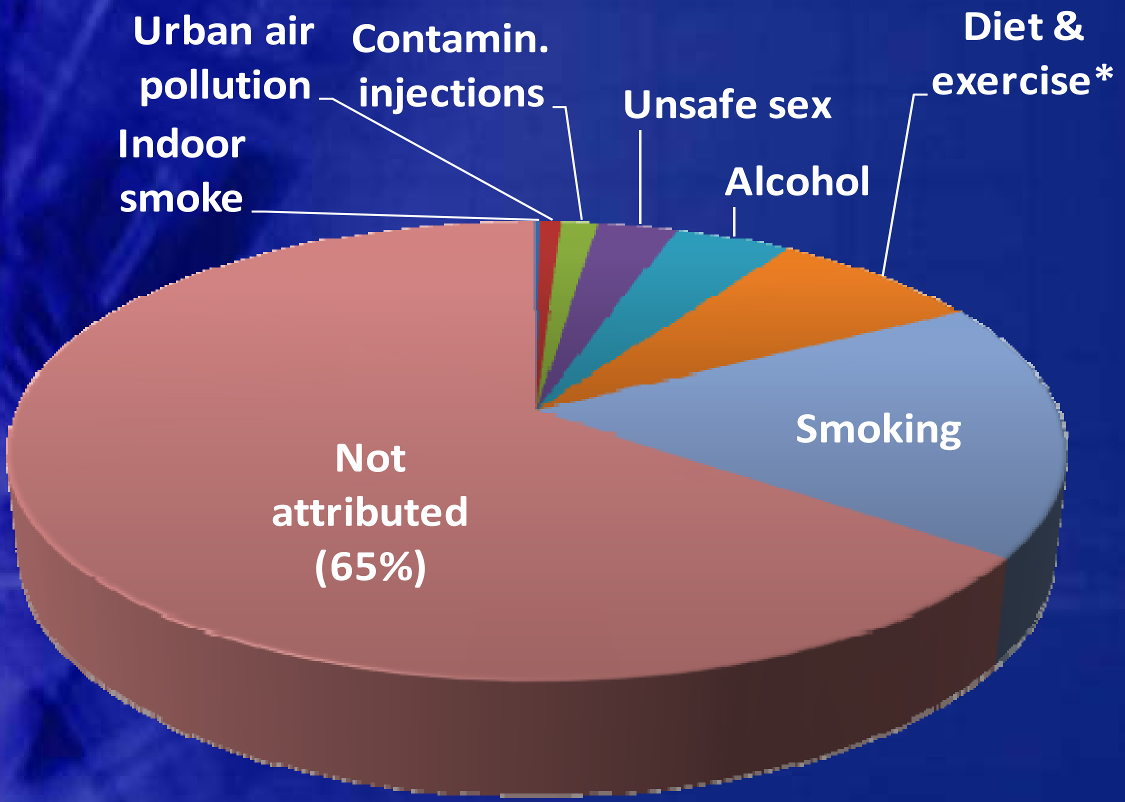
Site	Case pairs	Familial PAF (%)
Prostate	559	20.55*
Breast	2784	10.61*
Colorectum	771	6.87
Endometrium	119	5.31*
Ovary	155	4.90*
Lung	330	3.81
Thyroid	102	3.56
Melanoma	382	2.74
Testis	63	2.71*
Cervix	122	2.43
Skin	75	2.35
Bladder	146	2.03
All others		< 2.00

*** PAF was doubled to reflect both parental lineages.**

Over 22 sites the median PAF = 1.4 %

K Hemminki and K Czene, CEBP 2002, 11: 1638-44

Environmental risks of cancer



Attributable risks for cancer
(worldwide, all tumor types, joint PAF=35%)

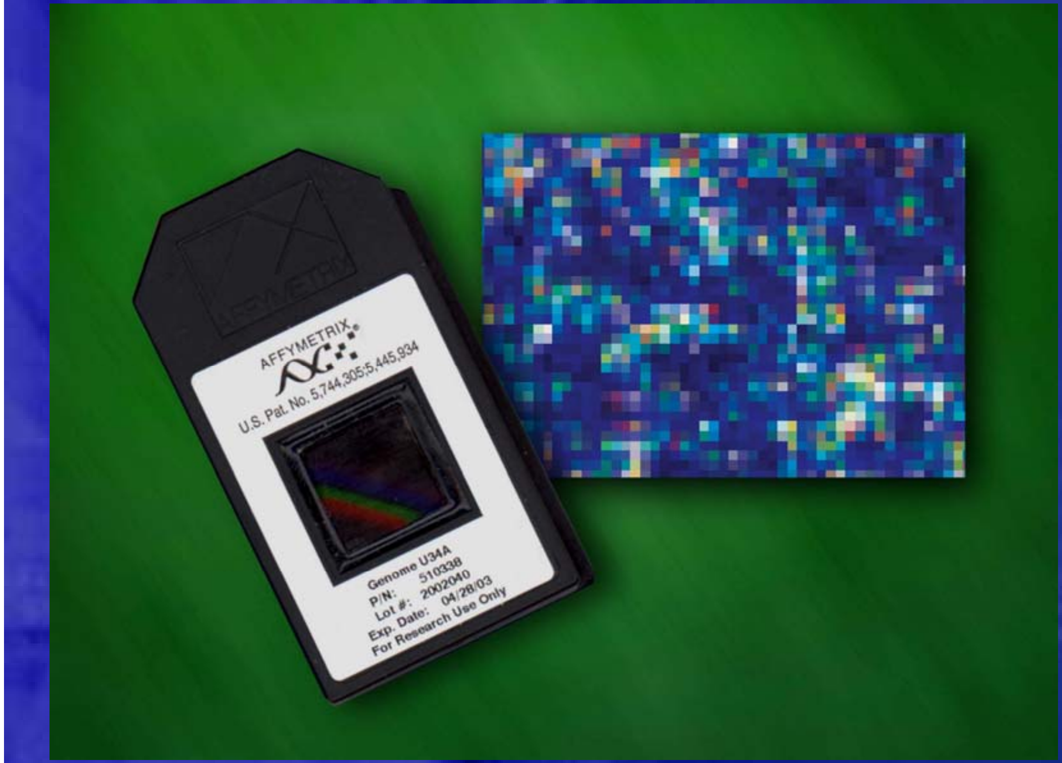
SM Rappaport

Data from Ezzati *et al.*, "Comparative Quantification of Mortality and Burden of Disease Attributable to Selected Risk Factors," *Global Burden of Disease and Risk Factors,* Chapter 4, WHO, 2006.

Discovering causes of cancer

- Cancer risks attributable to genetic factors (G) are typically small (1 – 2%)
- Most cancers must be caused by non-genetic factors (E) or GxE
  - However, two thirds of attributable E risks have not been identified
- *What tools are available for identifying G, E and GxE causes of cancer?*

Human genotyping: major technology advances



SNPs per assay

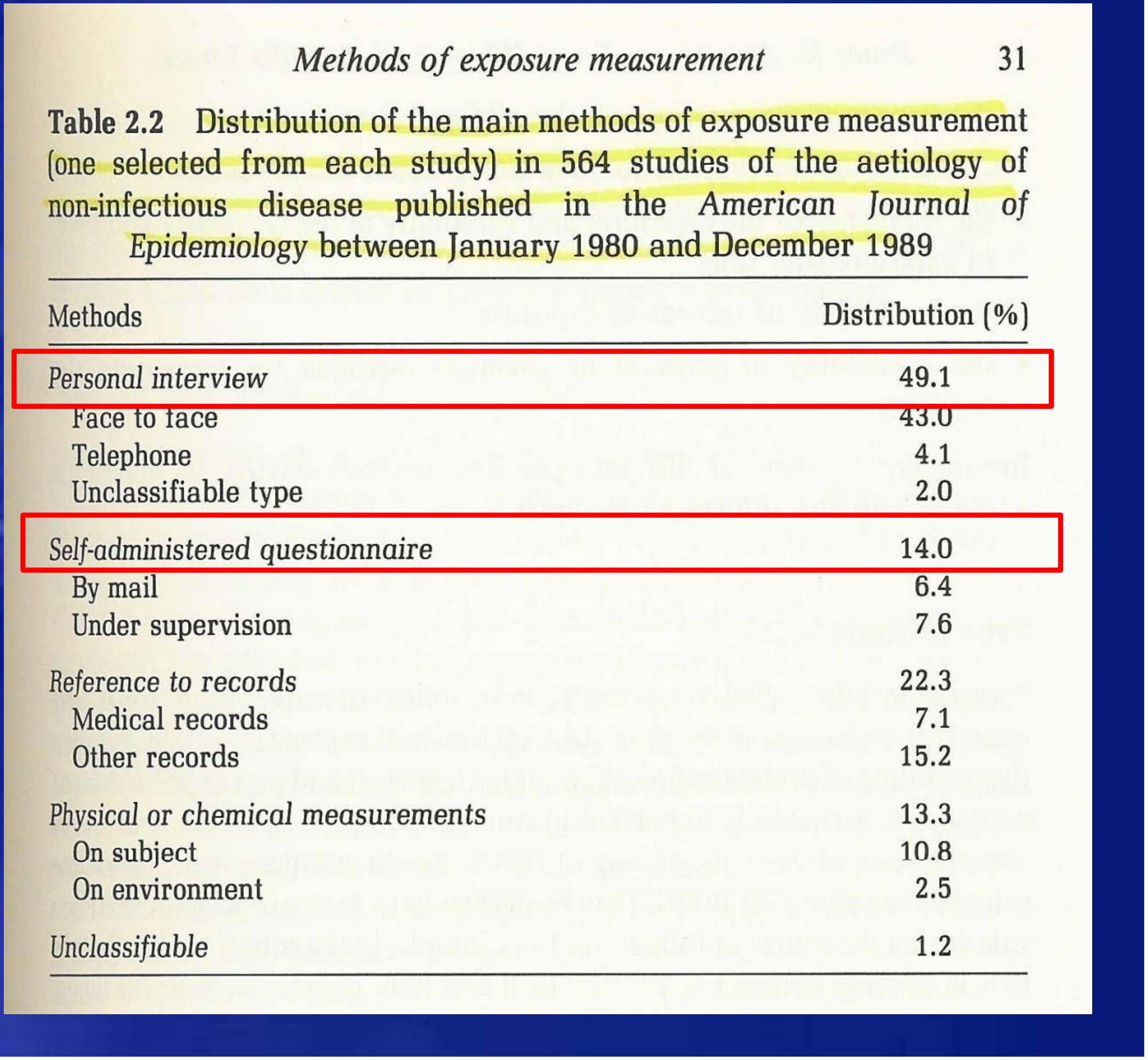
1997	1
2001	10
2002	1,000
2004	50,000
2006	500,000
2007	1,000,000
2010	>>1,000,000

Genome-Wide Association Studies (GWAS) **now possible with 2,000-20,000 samples (2 billion - 20 billion genotypes)**

Courtesy of E. Lander, MIT/Broad

Environmental factors in epidemiology

Two thirds of studies relied upon subjects to assess their own exposures!



Methods of exposure measurement 31

Table 2.2 Distribution of the main methods of exposure measurement (one selected from each study) in 564 studies of the aetiology of non-infectious disease published in the *American Journal of Epidemiology* between January 1980 and December 1989

Methods	Distribution (%)
Personal interview	49.1
Face to face	43.0
Telephone	4.1
Unclassifiable type	2.0
Self-administered questionnaire	14.0
By mail	6.4
Under supervision	7.6
Reference to records	22.3
Medical records	7.1
Other records	15.2
Physical or chemical measurements	13.3
On subject	10.8
On environment	2.5
Unclassifiable	1.2

B.K. Armstrong *et al. Principles of Exposure Measurement in Epidemiology,* **Oxford Med. Pubs., 1992**

Exposure assessment for cancer (2010)

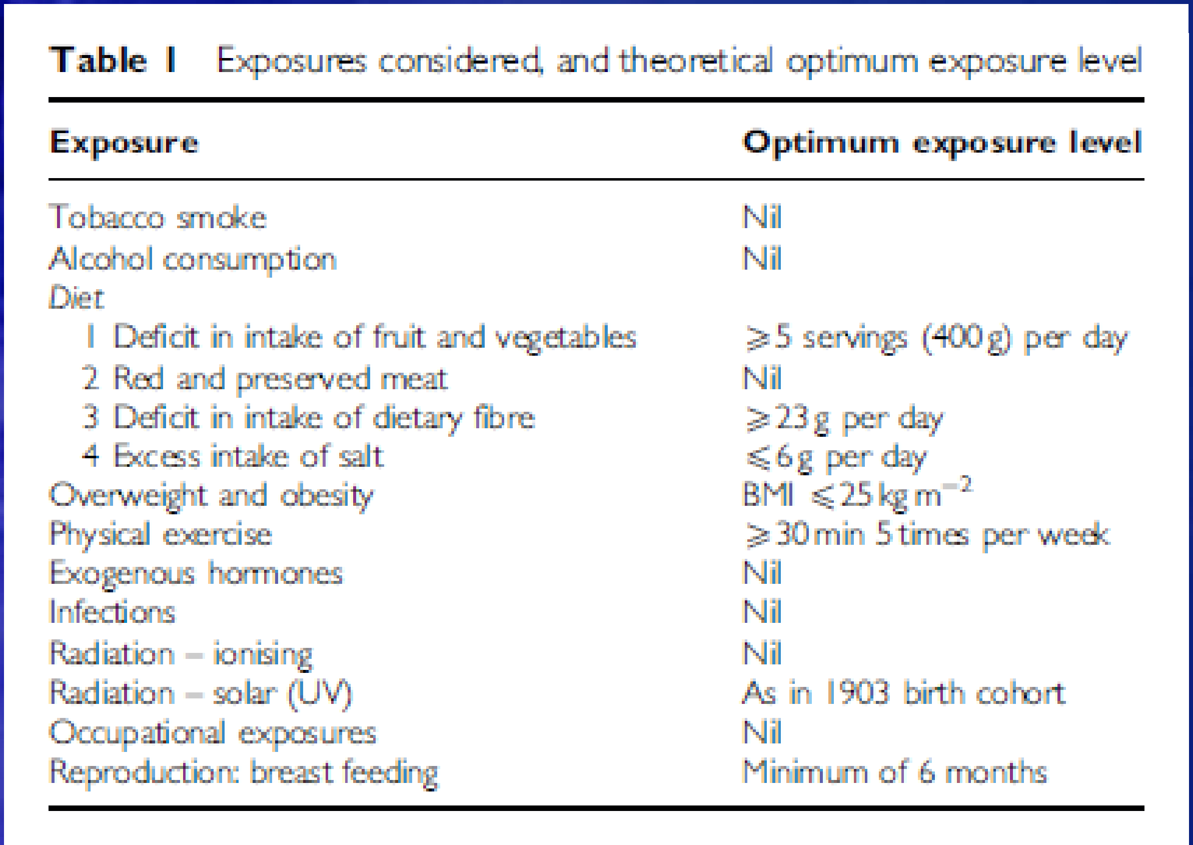


Table 1 Exposures considered, and theoretical optimum exposure level

Exposure	**Optimum exposure level**
Tobacco smoke	Nil
Alcohol consumption	Nil
Diet	
1 Deficit in intake of fruit and vegetables	⩾5 servings (400 g) per day
2 Red and preserved meat	Nil
3 Deficit in intake of dietary fibre	⩾23 g per day
4 Excess intake of salt	⩽6 g per day
Overweight and obesity	BMI $\leqslant 25\,\mathrm{kg\,m^{-2}}$
Physical exercise	⩾30 min 5 times per week
Exogenous hormones	Nil
Infections	Nil
Radiation – ionising	Nil
Radiation – solar (UV)	As in 1903 birth cohort
Occupational exposures	Nil
Reproduction: breast feeding	Minimum of 6 months

DM Parkin, The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010, *Brit J Cancer 105, S1-S5 (2011).*

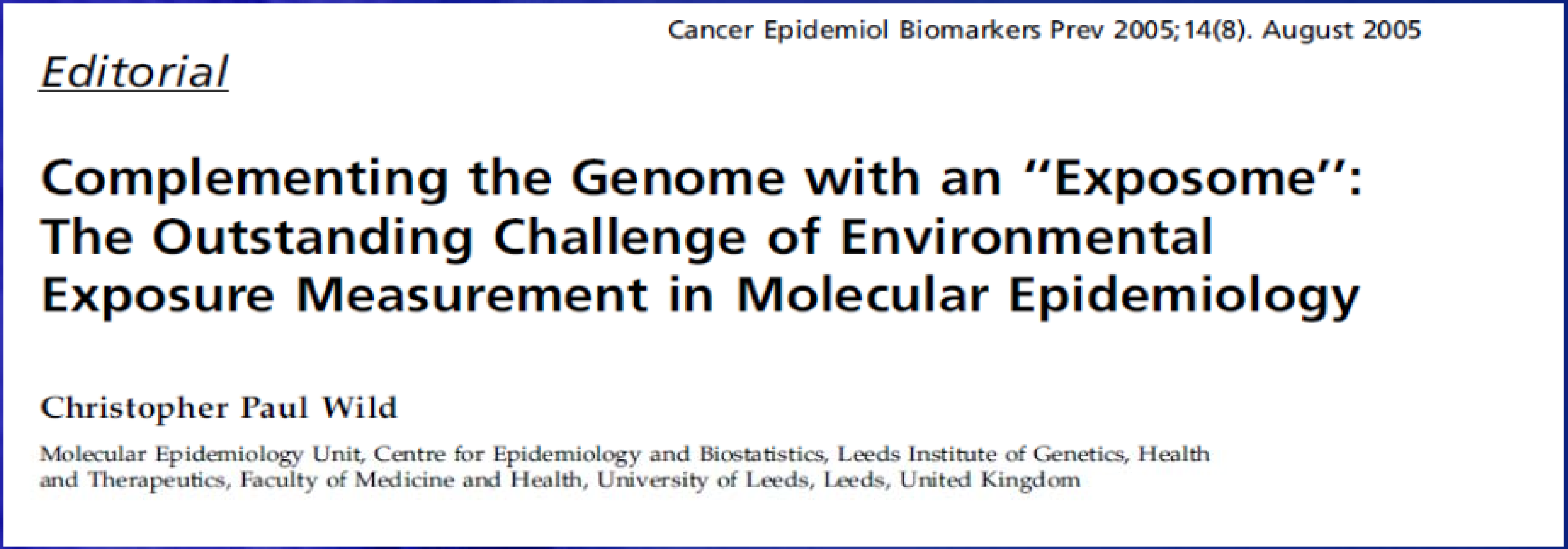
Finding unknown causes of cancer

- Elaboration of the G matrix with modern GWAS has been stunningly comprehensive
  - but has explained relatively little cancer risk
- Elaboration of the E matrix relies on questionnaires, geographic information and targeted measurements
  - much as it did a century ago

The exposome – **promoting discovery of environmental causes of disease**

Christopher Wild defined the ‘exposome’, representing *all* environmental exposures (including diet, lifestyle, and infections) from conception onwards, as a complement to the genome in studies of disease etiology.

Wild, C.P., *Cancer Epidemiol Biomarkers Prev 14 (8), 1847-1850 (2005).*



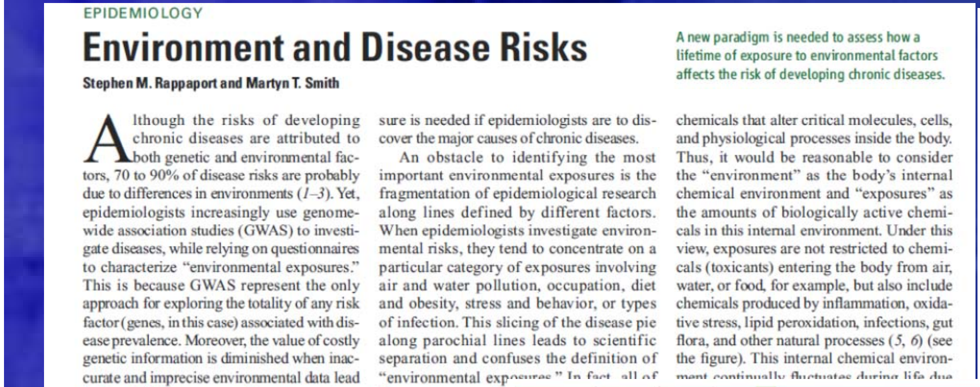
Cancer Epidemiol Biomarkers Prev 2005;14(8). August 2005

Editorial

Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology

Christopher Paul Wild

Molecular Epidemiology Unit, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom



EPIDEMIOLOGY

Environment and Disease Risks

Stephen M. Rappaport and Martyn T. Smith

A new paradigm is needed to assess how a lifetime of exposure to environmental factors affects the risk of developing chronic diseases.

Although the risks of developing chronic diseases are attributed to both genetic and environmental factors, 70 to 90% of disease risks are probably due to differences in environments (1–3). Yet, epidemiologists increasingly use genome-wide association studies (GWAS) to investigate diseases, while relying on questionnaires to characterize "environmental exposures." This is because GWAS represent the only approach for exploring the totality of any risk factor (genes, in this case) associated with disease prevalence. Moreover, the value of costly genetic information is diminished when inaccurate and imprecise environmental data lead to biased inferences regarding gene-environment interactions (4). A more comprehensive and quantitative view of environmental exposure is needed if epidemiologists are to discover the major causes of chronic diseases.

An obstacle to identifying the most important environmental exposures is the fragmentation of epidemiological research along lines defined by different factors. When epidemiologists investigate environmental risks, they tend to concentrate on a particular category of exposures involving air and water pollution, occupation, diet and obesity, stress and behavior, or types of infection. This slicing of the disease pie along parochial lines leads to scientific separation and confuses the definition of "environmental exposures." In fact, all of these exposure cate chronic diseases and collectively rather th

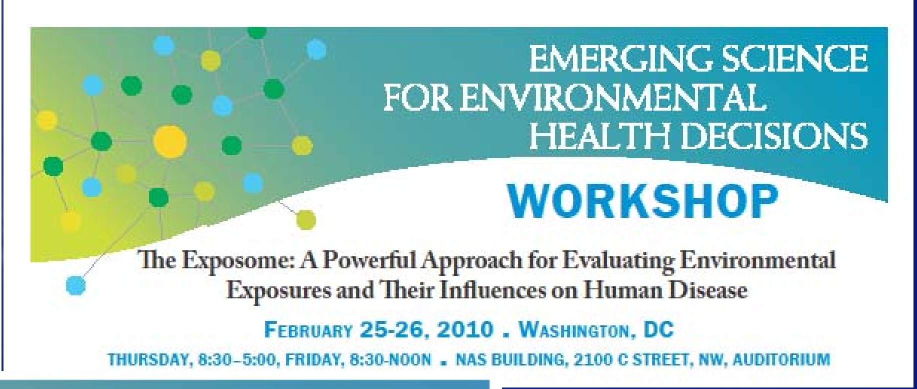
To develop a mor ronmental exposure, nize that toxic effect

chemicals that alter critical molecules, cells, and physiological processes inside the body. Thus, it would be reasonable to consider the "environment" as the body's internal chemical environment and "exposures" as the amounts of biologically active chemicals in this internal environment. Under this view, exposures are not restricted to chemicals (toxicants) entering the body from air, water, or food, for example, but also include chemicals produced by inflammation, oxidative stress, lipid peroxidation, infections, gut flora, and other natural processes (5, 6) (see the figure). This internal chemical environment continually fluctuates during life due

School of Public Health, University of California, Berkeley, CA 94720–7356, USA. E-mail: srappaport@berkeley.edu

460 22 OCTOBER 2010 VOL 330

Publishe



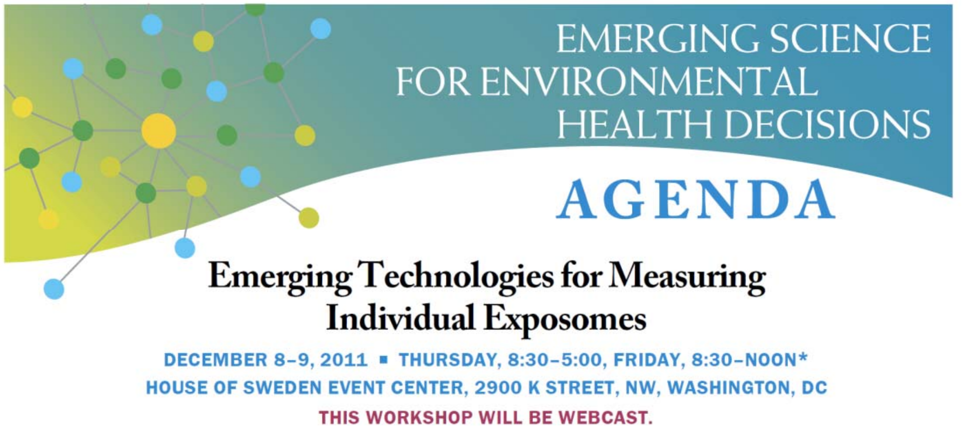
EMERGING SCIENCE
FOR ENVIRONMENTAL
HEALTH DECISIONS

WORKSHOP

The Exposome: A Powerful Approach for Evaluating Environmental Exposures and Their Influences on Human Disease

FEBRUARY 25-26, 2010 . WASHINGTON, DC

THURSDAY, 8:30–5:00, FRIDAY, 8:30-NOON . NAS BUILDING, 2100 C STREET, NW, AUDITORIUM



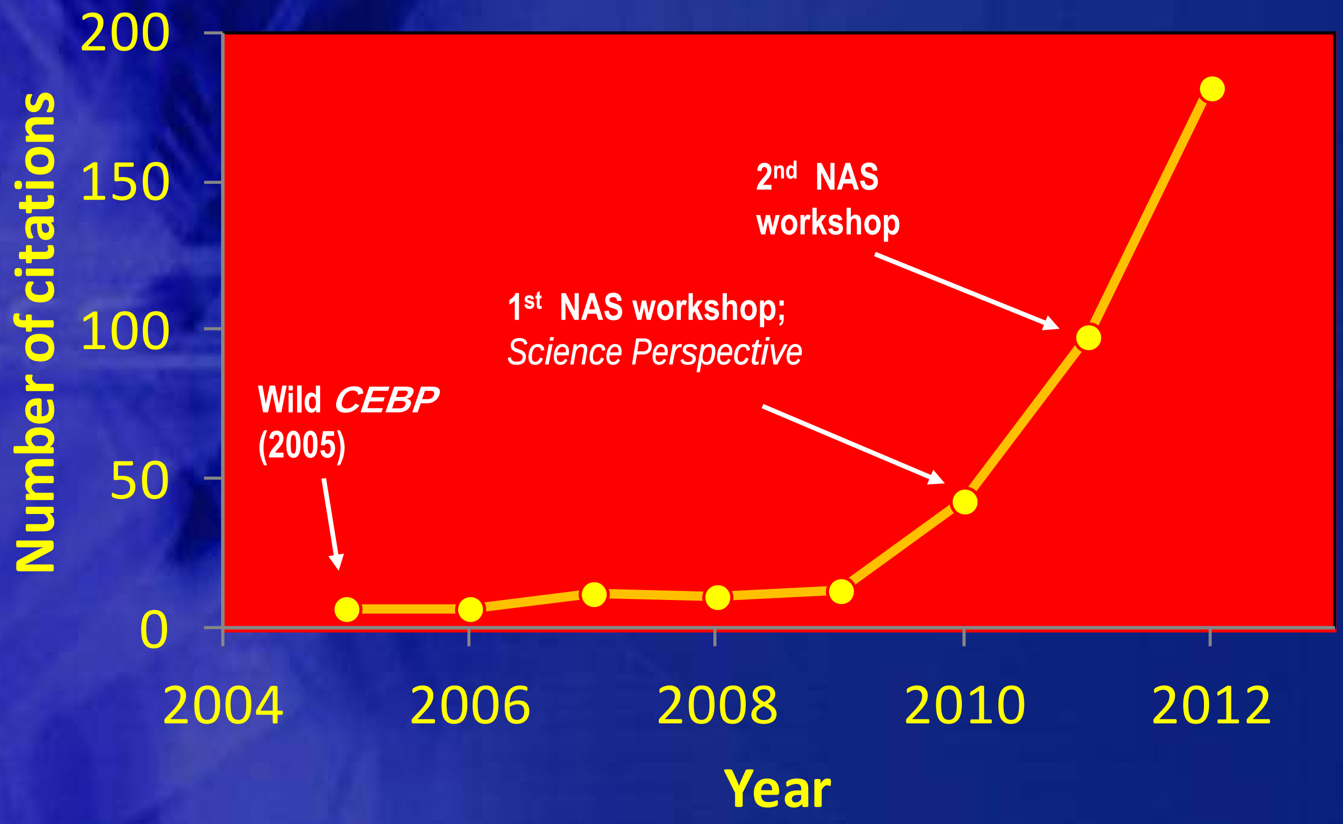
EMERGING SCIENCE
FOR ENVIRONMENTAL
HEALTH DECISIONS

AGENDA

Emerging Technologies for Measuring Individual Exposomes

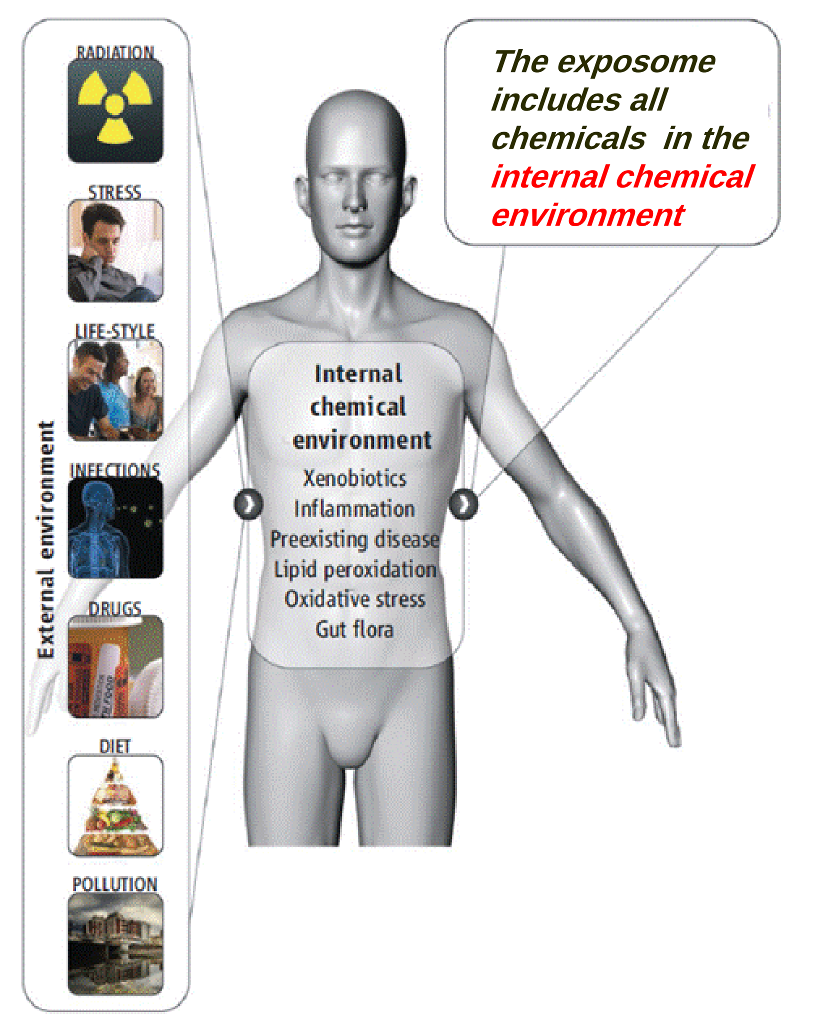
DECEMBER 8–9, 2011 ▪ THURSDAY, 8:30–5:00, FRIDAY, 8:30–NOON*
HOUSE OF SWEDEN EVENT CENTER, 2900 K STREET, NW, WASHINGTON, DC
THIS WORKSHOP WILL BE WEBCAST.

Scientific citations to ‘exposome’ (Google Scholar)



SM Rappaport

Capturing exogenous and endogenous exposures



The exposome includes all chemicals in the internal chemical environment

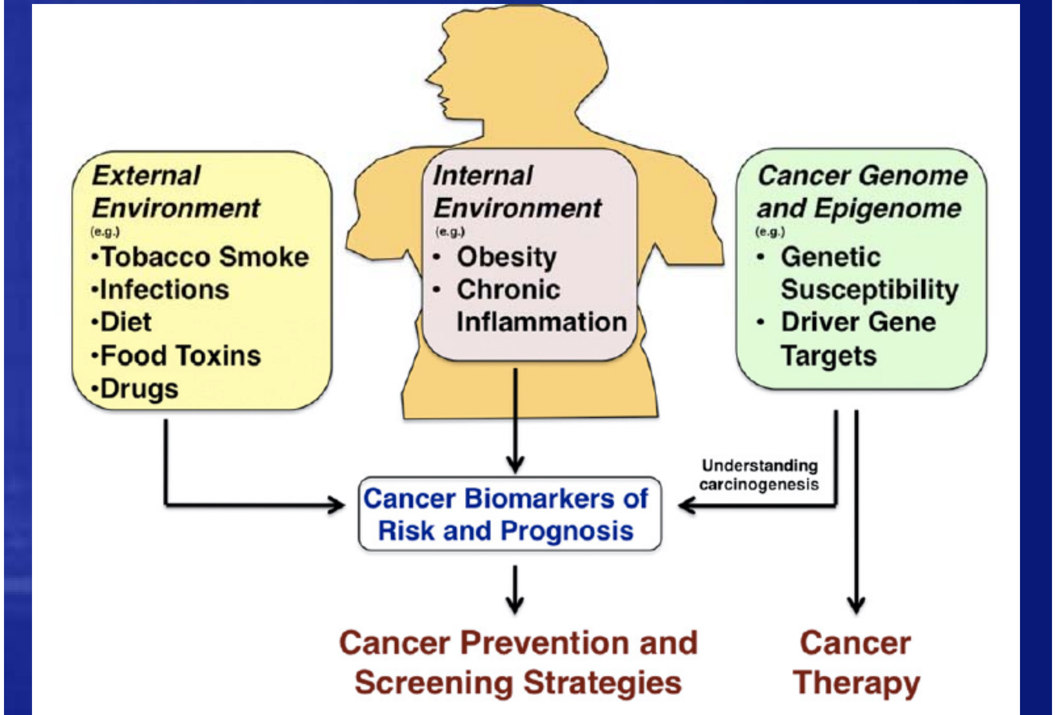


Fig. 1. Analysis of the exposome (external and internal environment) and the cancer genome (somatic and germ-line mutations and epigenetic changes, e.g., DNA methylation and noncoding RNAs) will improve the understanding carcinogenesis, cancer therapy, and cancer prevention.

A. Schetter and C. Harris, PNAS, 2012, 109: 7955-6

S.M. Rappaport and M.T. Smith, Science, 2010: 330, 460-461

Exposome-wide association studies (EWAS)

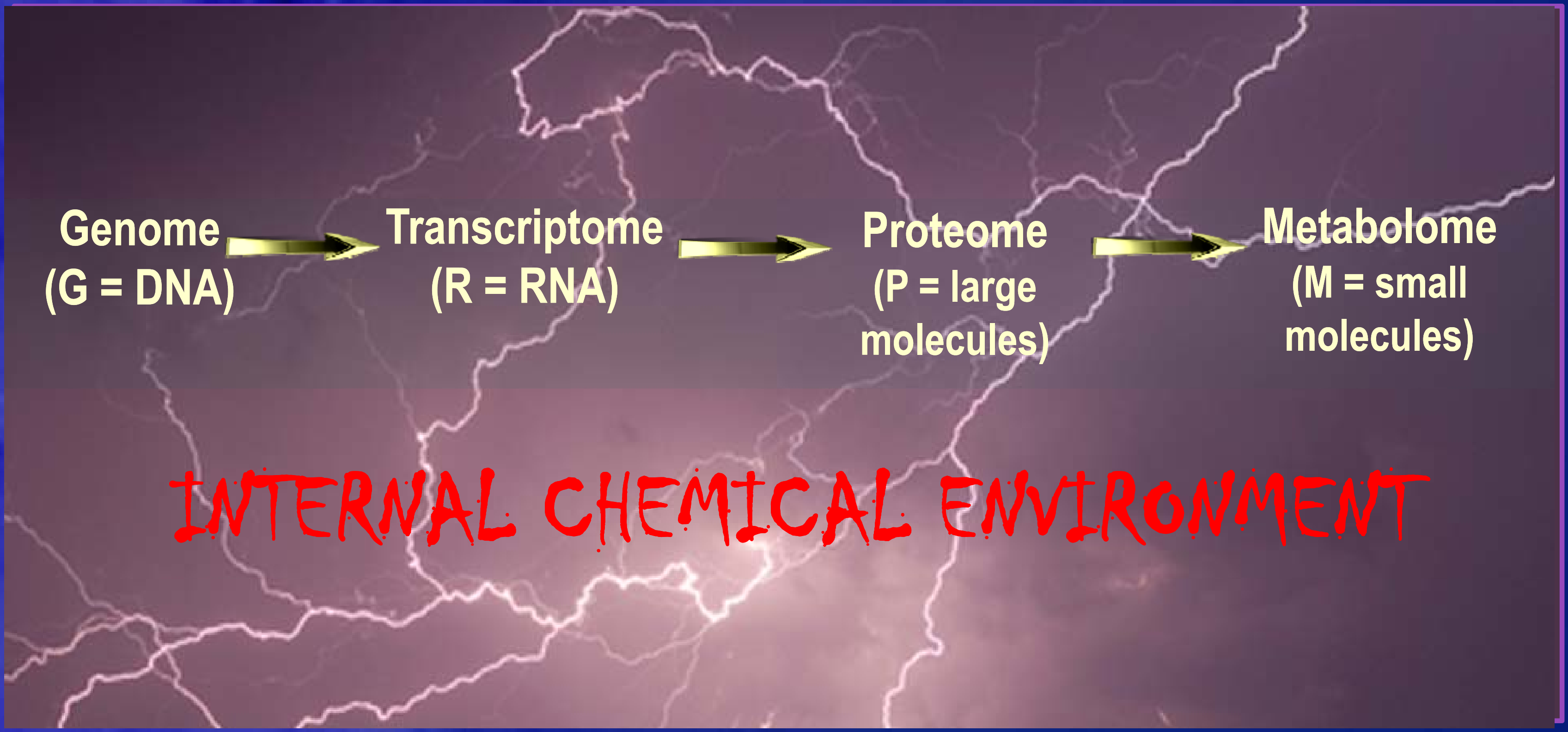
By applying EWAS to biospecimens from healthy and diseased subjects, we can discover causal environmental exposures.

http://www.flickr.com/photos/paulieparker/246707763/

But which ‘omes’ offer the most promise for EWAS and follow-up studies?

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

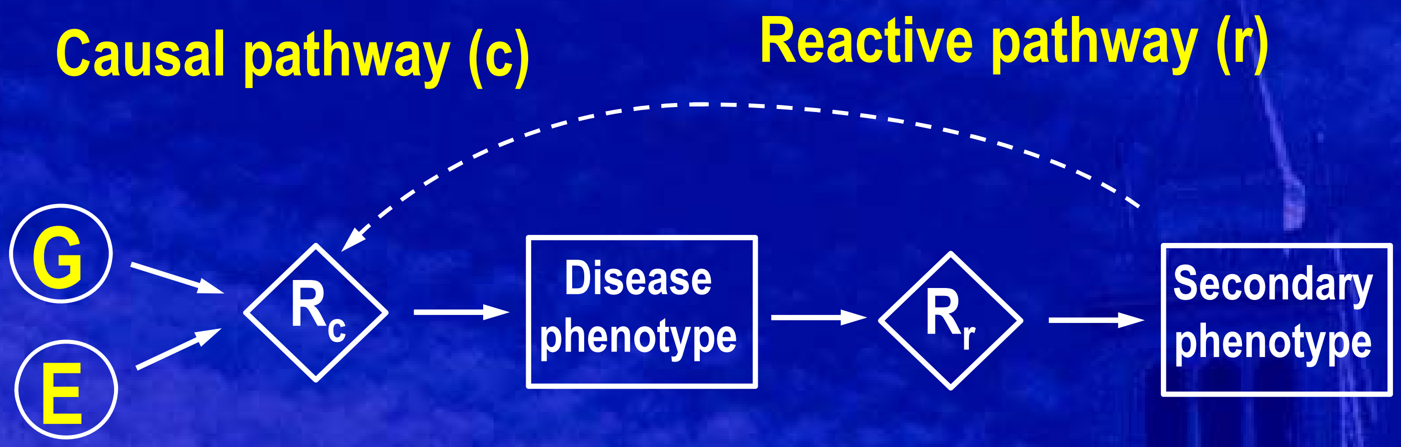
The molecular basis of life (and disease)



Genome (G = DNA) → Transcriptome (R = RNA) → Proteome (P = large molecules) → Metabolome (M = small molecules)

INTERNAL CHEMICAL ENVIRONMENT

Disease pathways



Causal pathway (c)

Reactive pathway (r)

G

E

R_c

Disease phenotype

R_r

Secondary phenotype

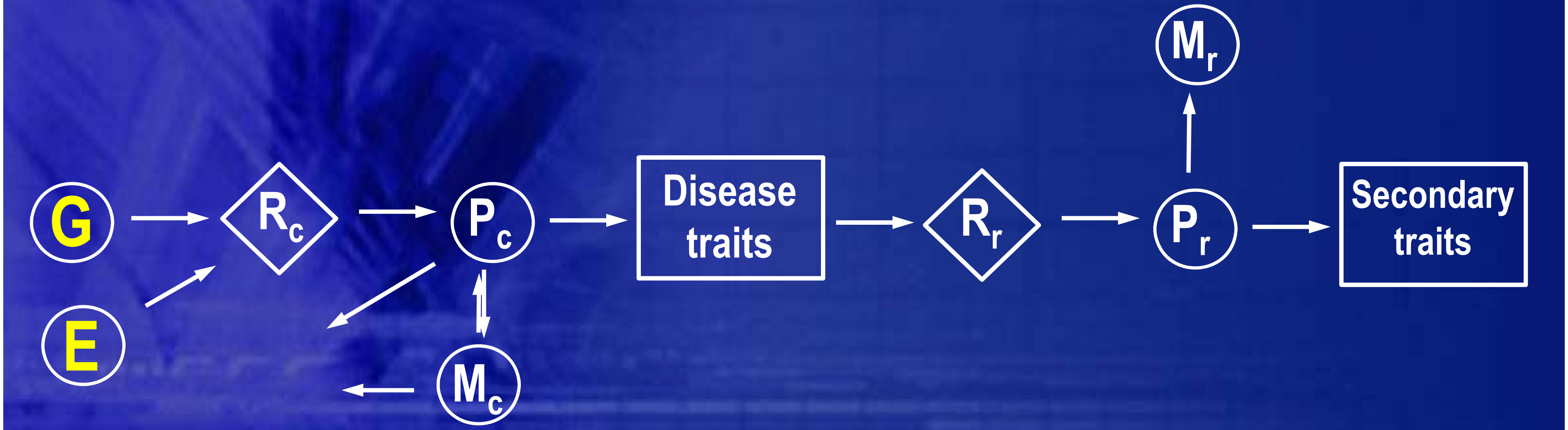
G = genome
E = environment
R = transcriptome (gene expression)

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9
Based on: E. Shadt *et al.*, *Nat Gen*, 2005, 37: 710-717

Adding omes

Causal pathway (c)

Reactive pathway (r)



G

E

R_c

P_c

M_c

Disease traits

R_r

P_r

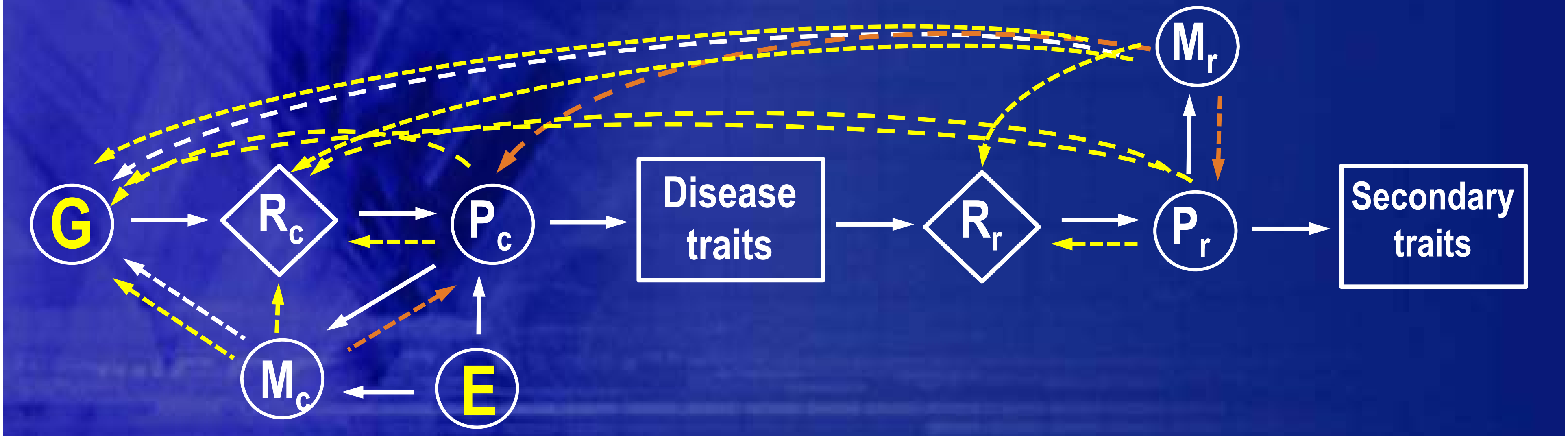
M_r

Secondary traits

G = genome
E = environment
R = transcriptome (gene expression)
P = proteome (protein expression)
M = metabolome (all small molecules and metals)

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

More omic connections



G

R_c

P_c

Disease traits

R_r

P_r

Secondary traits

M_r

M_c

E

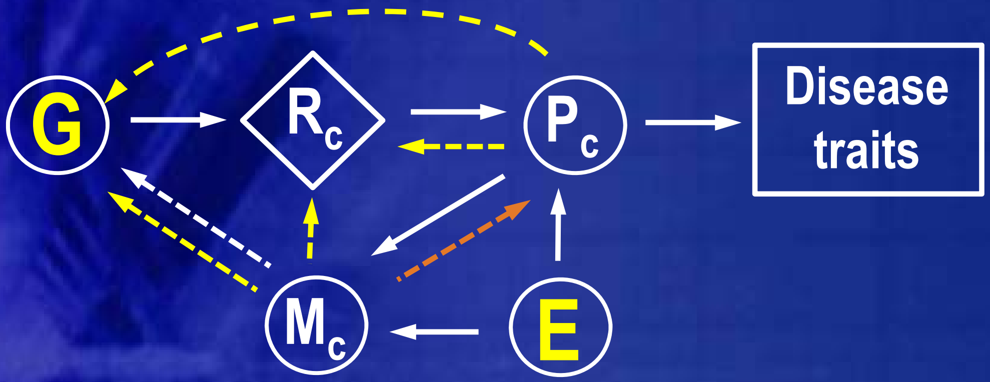
----- Genetic modifications (mutations)

----- Post-translational modifications

----- Epigenetic modifications

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

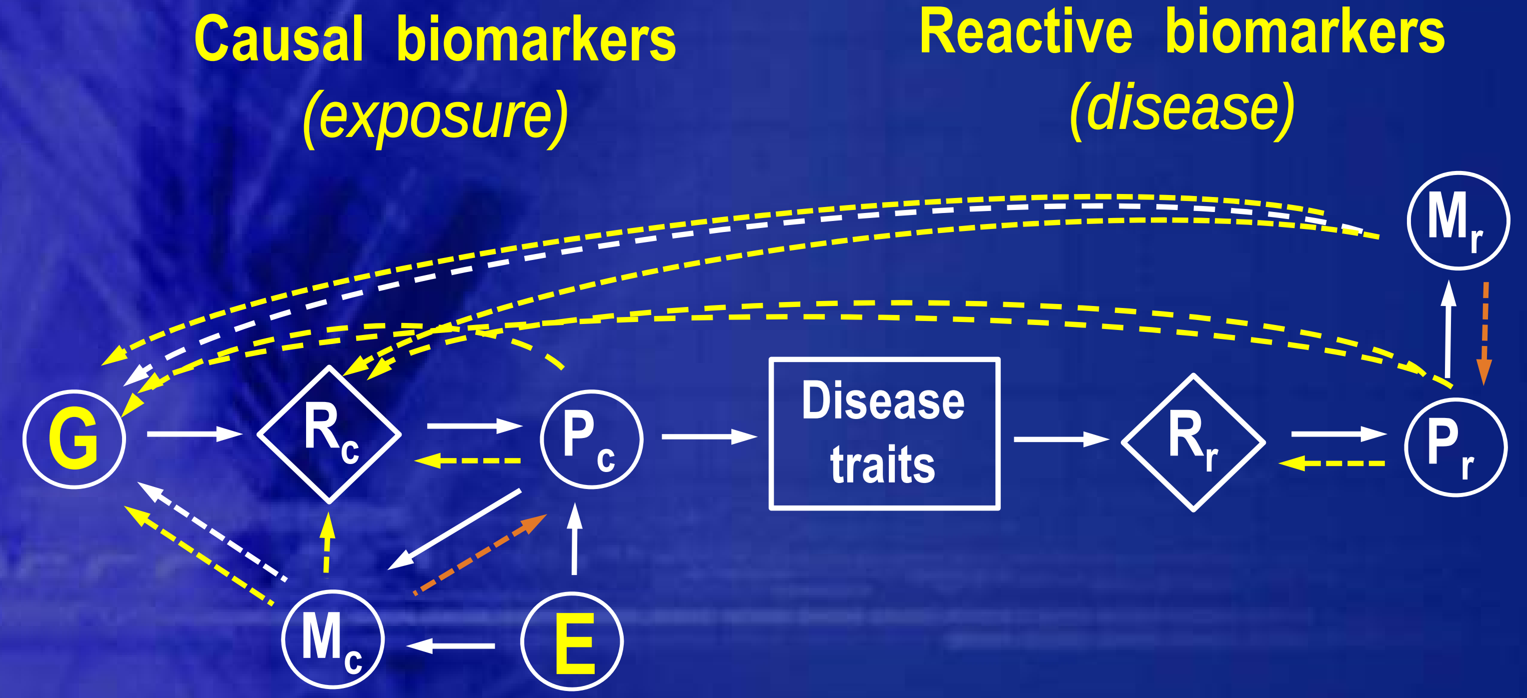
Which omes for EWAS?



If causal exposures operate primarily through small molecules (M_c) and proteins (P_c), then EWAS require metabolomics and/or proteomics.

SM Rappaport

Biospecimens for EWAS?



Reactive biomarkers obscure causal pathways.

For validation of exposure biomarkers, biospecimens should be obtained prior to disease (prospective cohorts)

Bioactive molecules

Reactive electrophiles:
Reactive O, N & Cl species
Aldehydes
Epoxides
Quinones

Metabolome:
Lipids
Sugars
Nucleotides
Amino acids
Metabolites
Xenobiotics

Inflammation markers:
Cytokines
Chemokines
Eicosanoids
Vasoactive amines
Growth factors

SERUM EXPOSOME

Micronutrients

Metals

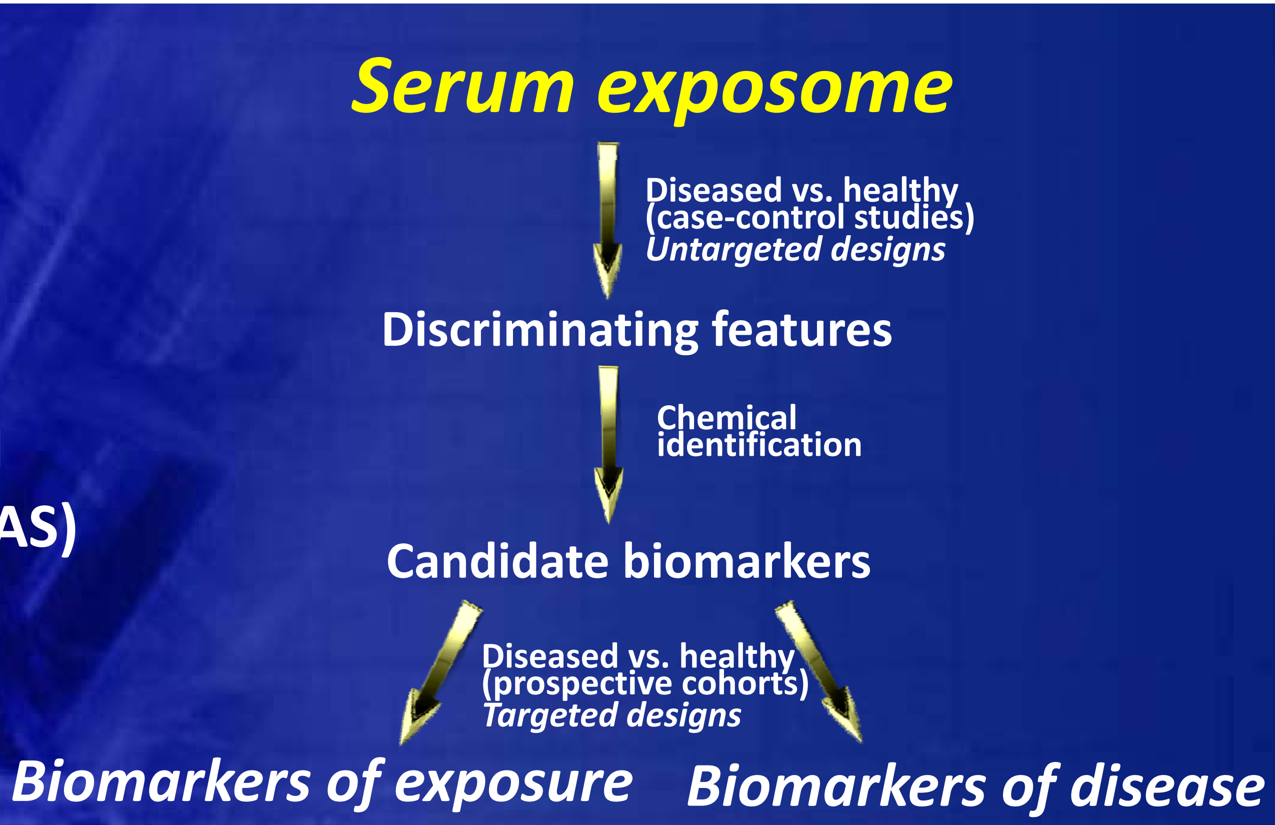
Microbiome products

Drugs

Receptor-binding agents:
Hormones
Xenoestrogens
Endocrine disruptors

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

DATA-DRIVEN DISCOVERY (EWAS)



Serum exposome

Diseased vs. healthy
(case-control studies)
Untargeted designs

Discriminating features

Chemical
identification

Candidate biomarkers

Diseased vs. healthy
(prospective cohorts)
Targeted designs

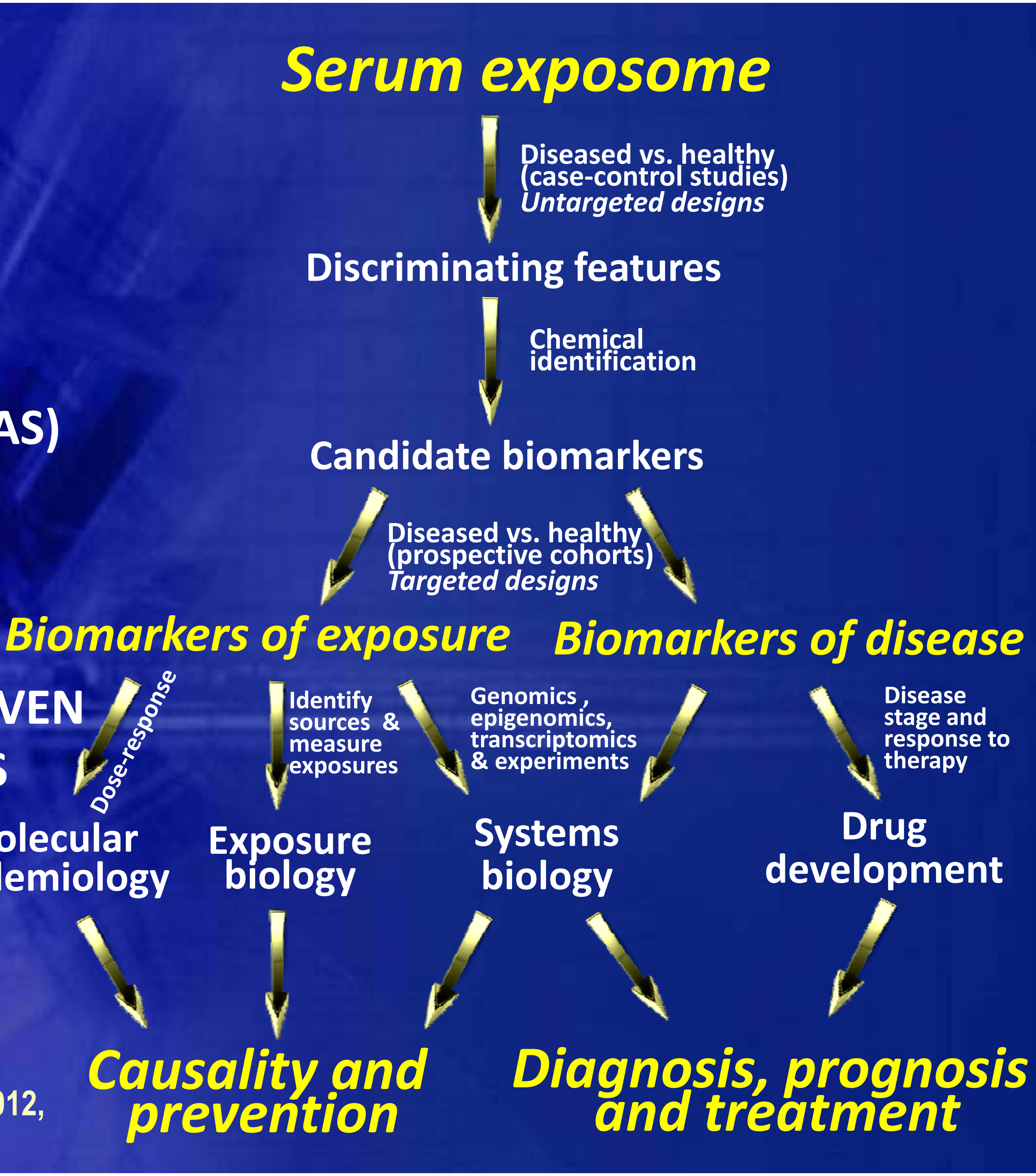
Biomarkers of exposure

Biomarkers of disease

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

DATA-DRIVEN DISCOVERY (EWAS)

KNOWLEDGE-DRIVEN APPLICATIONS



Serum exposome

Diseased vs. healthy
(case-control studies)
Untargeted designs

Discriminating features

Chemical identification

Candidate biomarkers

Diseased vs. healthy
(prospective cohorts)
Targeted designs

Biomarkers of exposure

Biomarkers of disease

Dose-response

Identify sources & measure exposures

Genomics, epigenomics, transcriptomics & experiments

Disease stage and response to therapy

Molecular epidemiology

Exposure biology

Systems biology

Drug development

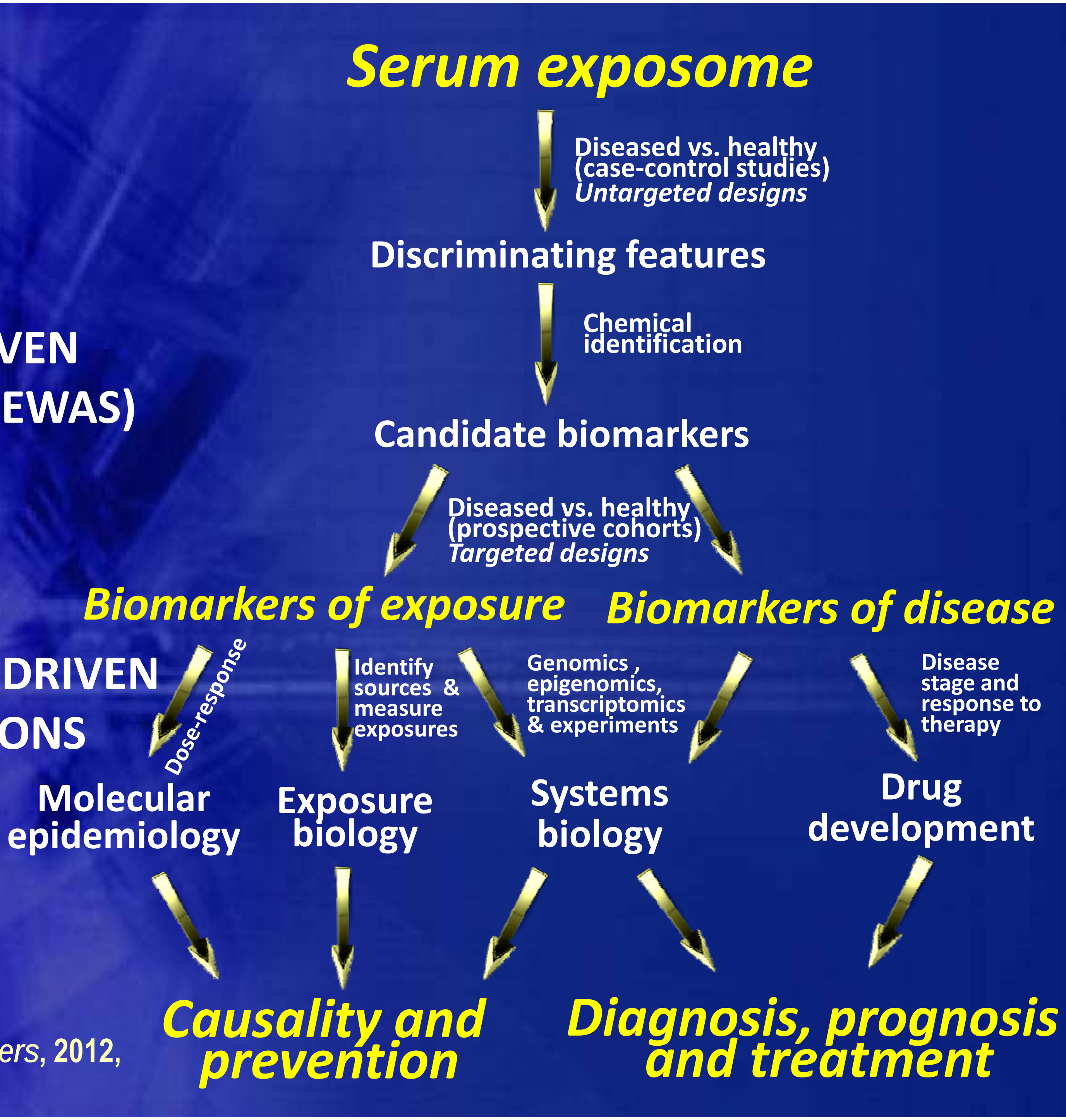
Causality and prevention

Diagnosis, prognosis and treatment

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

DATA-DRIVEN DISCOVERY (EWAS)

KNOWLEDGE-DRIVEN APPLICATIONS



Serum exposome

Diseased vs. healthy
(case-control studies)
Untargeted designs

Discriminating features

Chemical identification

Candidate biomarkers

Diseased vs. healthy
(prospective cohorts)
Targeted designs

Biomarkers of exposure

Biomarkers of disease

Dose-response

Identify sources & measure exposures

Genomics, epigenomics, transcriptomics & experiments

Disease stage and response to therapy

Molecular epidemiology

Exposure biology

Systems biology

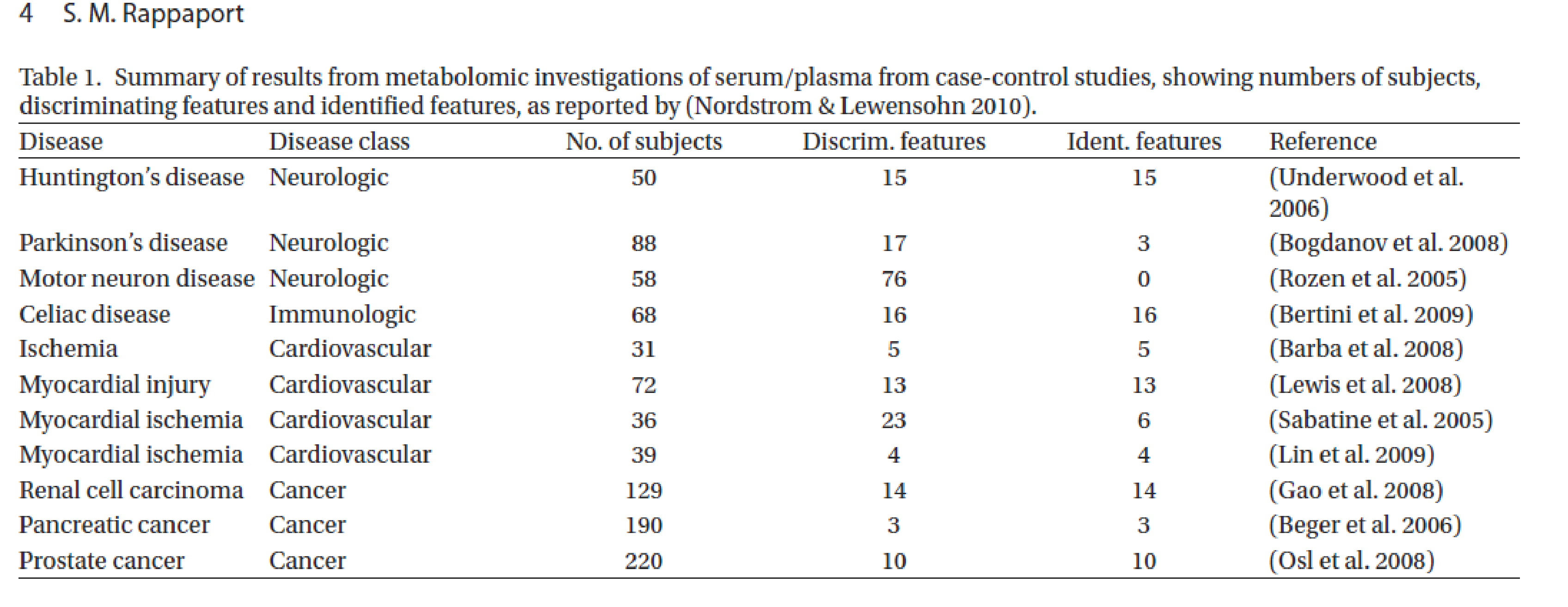
Drug development

Causality and prevention

Diagnosis, prognosis and treatment

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

EWAS: proof of concept (Metabolomics via NMR & MS)



4 S. M. Rappaport

Table 1. Summary of results from metabolomic investigations of serum/plasma from case-control studies, showing numbers of subjects, discriminating features and identified features, as reported by (Nordstrom & Lewensohn 2010).

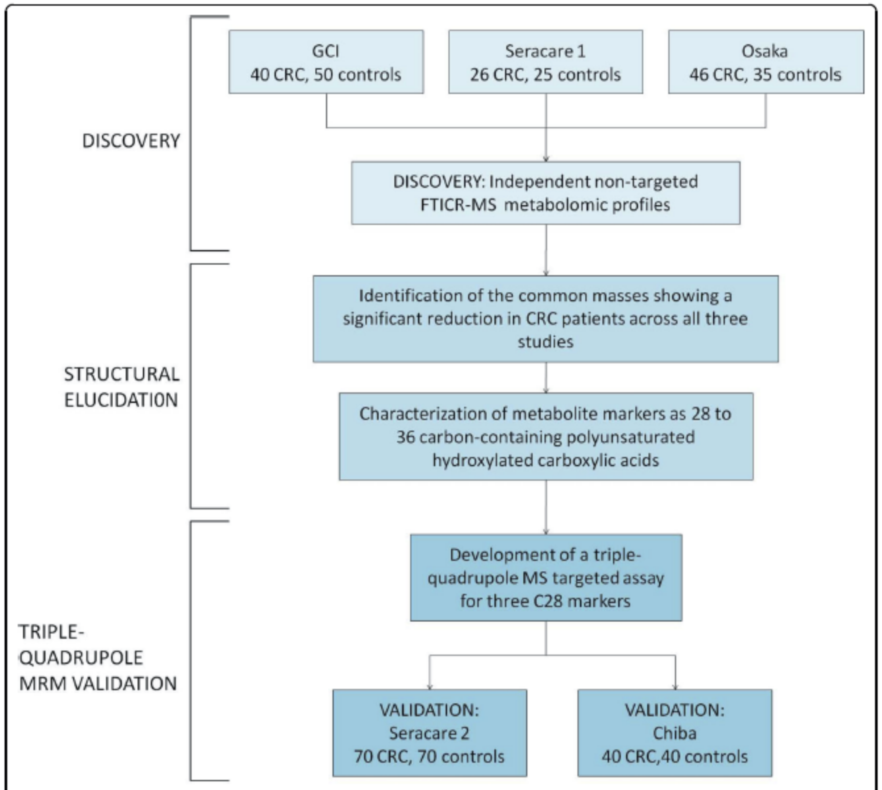
Disease	Disease class	No. of subjects	Discrim. features	Ident. features	Reference
Huntington's disease	Neurologic	50	15	15	(Underwood et al. 2006)
Parkinson's disease	Neurologic	88	17	3	(Bogdanov et al. 2008)
Motor neuron disease	Neurologic	58	76	0	(Rozen et al. 2005)
Celiac disease	Immunologic	68	16	16	(Bertini et al. 2009)
Ischemia	Cardiovascular	31	5	5	(Barba et al. 2008)
Myocardial injury	Cardiovascular	72	13	13	(Lewis et al. 2008)
Myocardial ischemia	Cardiovascular	36	23	6	(Sabatine et al. 2005)
Myocardial ischemia	Cardiovascular	39	4	4	(Lin et al. 2009)
Renal cell carcinoma	Cancer	129	14	14	(Gao et al. 2008)
Pancreatic cancer	Cancer	190	3	3	(Beger et al. 2006)
Prostate cancer	Cancer	220	10	10	(Osl et al. 2008)

Modest numbers of subjects

Candidate biomarkers

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9

An EWAS of colorectal cancer



DISCOVERY

GCI
40 CRC, 50 controls

Seracare 1
26 CRC, 25 controls

Osaka
46 CRC, 35 controls

DISCOVERY: Independent non-targeted FTICR-MS metabolomic profiles

STRUCTURAL ELUCIDATION

Identification of the common masses showing a significant reduction in CRC patients across all three studies

Characterization of metabolite markers as 28 to 36 carbon-containing polyunsaturated hydroxylated carboxylic acids

TRIPLE-QUADRUPOLE MRM VALIDATION

Development of a triple-quadrupole MS targeted assay for three C28 markers

VALIDATION:
Seracare 2
70 CRC, 70 controls

VALIDATION:
Chiba
40 CRC,40 controls

Figure 1 Study design. The study comprised three phases: Fourier transform ion cyclotron resonance mass spectrometry metabolomic discovery in three independent sample sets, structural investigation and determination of metabolic biomarkers as hydroxylated polyunsaturated ultra long-chain fatty acids and validation using a triple-quadrupole multiple reaction monitoring targeted assay.

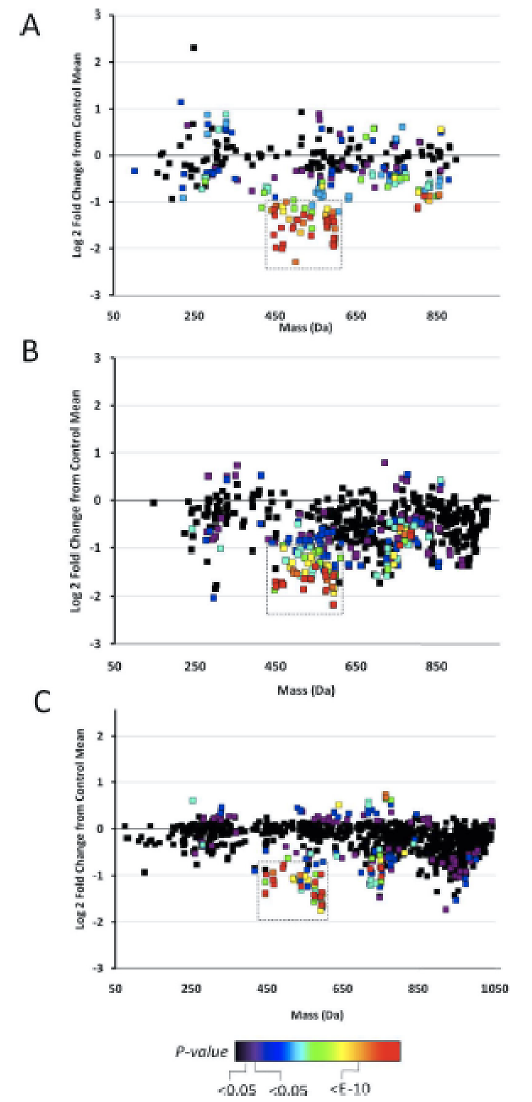


Figure 2 Scatter plots of average sample peak intensity fold change between colorectal cancer (CRC) and normal patient sera in three independent studies. Sample-specific peaks for all subjects were log2 normalized to the mean of the control population, and plotted according to mass (Da). Points are coloured according to significance based on an unpaired Studentst-test (see legend). (A) Genomics Collaborative Inc discovery population, (B) Seracare 1 discovery population, (C) Osaka discovery population. The region boxed in grey represents the cluster of masses between 440 and 600 Da consistently reduced in the CRC patient population compared to controls in all three cohorts.

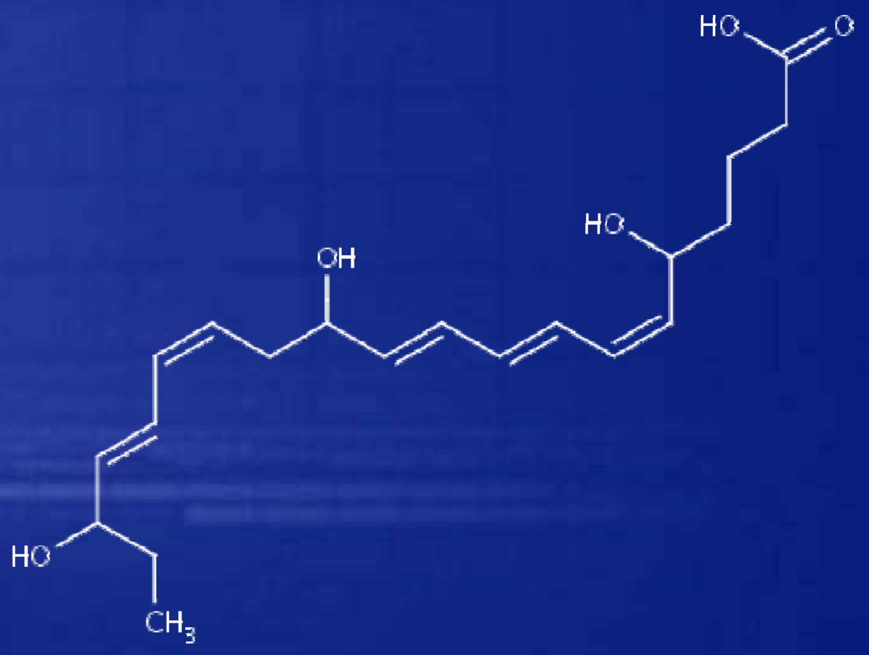
Possible omic features:
900 Da x 500 features/Da ≈ 0.5M features

SM Rappaport

Ritchie *et al.*, *BMC Medicine,* 2010, 8, 13

Biomarker identification

- Structures not confirmed
  - Hydroxylated ultra-long-chain fatty acids (C_{28} – C_{36})
  - Unique-mass spectra permit precise measurements
- Probably anti-inflammatory agents similar to resolvins, protectins and lipoxins (products of omega-3 fatty acids)



Resolvin E1

SM Rappaport

Follow up measurements of CRC-446

Biomarker highly associated with CRC

Uncorrelated with CRC stage

Does not return to normal after treatment

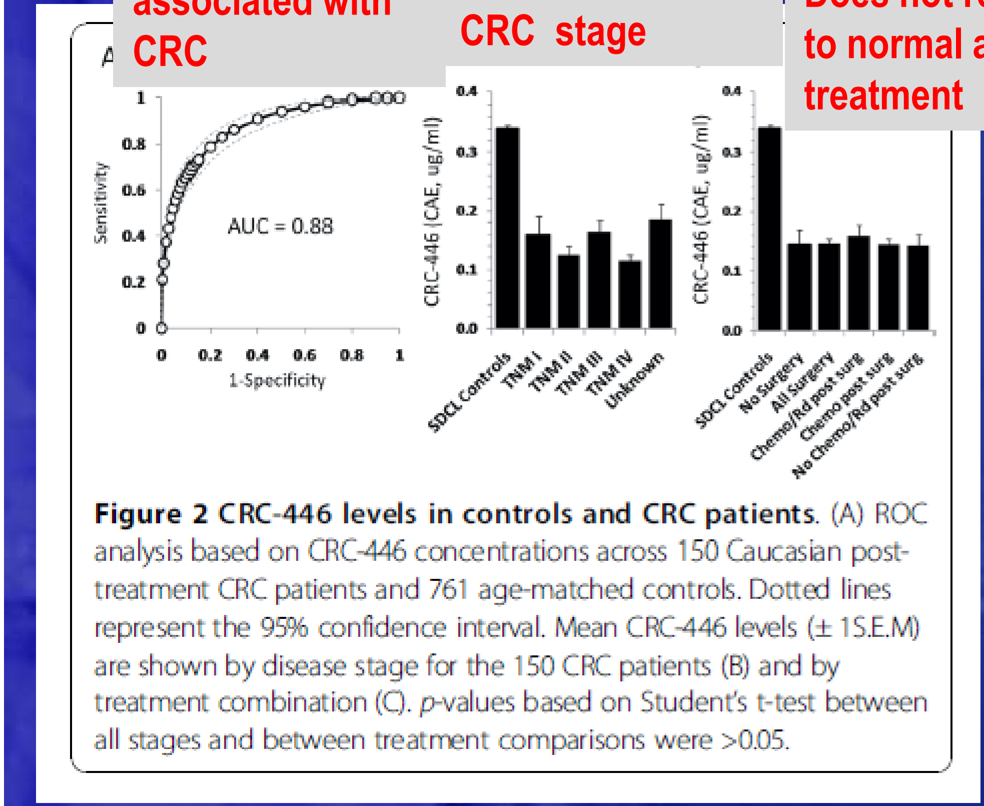
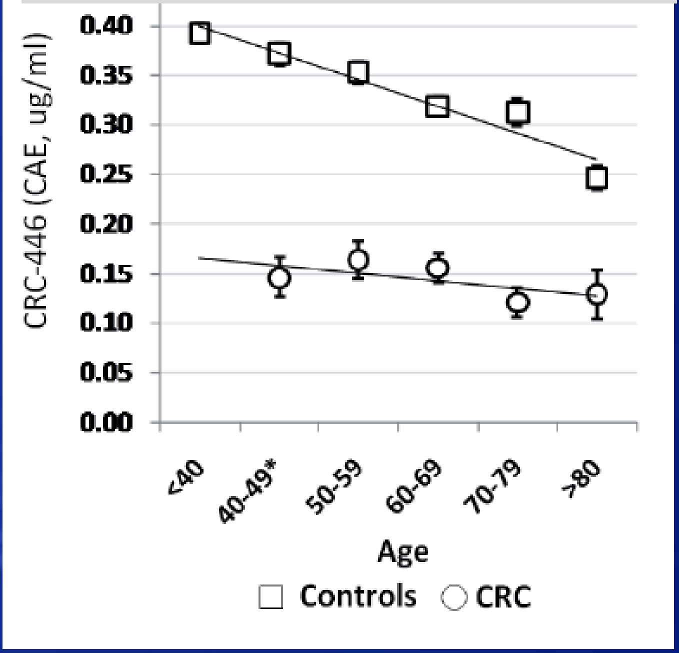


Figure 2 CRC-446 levels in controls and CRC patients. (A) ROC analysis based on CRC-446 concentrations across 150 Caucasian post-treatment CRC patients and 761 age-matched controls. Dotted lines represent the 95% confidence interval. Mean CRC-446 levels (± 1S.E.M) are shown by disease stage for the 150 CRC patients (B) and by treatment combination (C). *p*-values based on Student's t-test between all stages and between treatment comparisons were >0.05.

Biomarker also decreases with age



Results indicate that CRC-446 may be a causal biomarker of (protective) exposure!

SM Rappaport

Ritchie *et al.*, *BMC Gastroenterology,* 2010, 10, 140

Two biomarker-research agendas

EWAS

- For disease etiology
- Data-driven, untargeted designs
- Focus on small molecules and proteins
- To identify biomarkers
- Proof of concept has been established

Follow-up studies

- *For etiology, diagnosis and prognosis*
- *Knowledge-driven, targeted designs*
- For causative or suspicious factors
- Use biomarkers to confirm causality, etc.
- Provide feedback for public health and treatment

Needs for EWAS and follow-up

1. Interdisciplinary research teams (e.g. epidemiology, medicine, toxicology, analytical chemistry and statistics/bioinformatics)
2. Apply untargeted omics (metabolomics, proteomics and *adductomics*) to multiple case-control studies
   - State-of-the-art equipment (HR-MS/MS)
   - Method development/validation
   - Identify discriminating features (candidate biomarkers)
3. Follow up with biospecimens from prospective-cohort studies (targeted designs)
   - Add transcriptomics and systems biology
   - Advanced bioinformatics and statistics

Best wishes from Berkeley

Major support from NIEHS through grants U54ES016115 and P42ES04705

CEB Center For Exposure Biology

Genes & Environment Laboratory