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

S.M. Rappaport
University of California, Berkeley

Research support from NIEHS
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)

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

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.”

Familial risks of cancer
(From Swedish Family-Cancer database)

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

Over 22 sites the median PAF = 1.4%

*PAF was doubled to reflect both parental lineages.
Environmental risks of cancer

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

- Urban air pollution
- Contamin. injections
- Indoor smoke
- Unsafe sex
- Alcohol
- Diet & exercise*

Not attributed (65%)

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
  o 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!


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

*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*
Exposure assessment for cancer (2010)

Table 1  Exposures considered, and theoretical optimum exposure level

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Optimum exposure level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tobacco smoke</td>
<td>Nil</td>
</tr>
<tr>
<td>Alcohol consumption</td>
<td>Nil</td>
</tr>
<tr>
<td>Diet</td>
<td></td>
</tr>
<tr>
<td>1 Deficit in intake of fruit and vegetables</td>
<td>≥5 servings (400 g) per day</td>
</tr>
<tr>
<td>2 Red and preserved meat</td>
<td>Nil</td>
</tr>
<tr>
<td>3 Deficit in intake of dietary fibre</td>
<td>≥23 g per day</td>
</tr>
<tr>
<td>4 Excess intake of salt</td>
<td>≤6 g per day</td>
</tr>
<tr>
<td>Overweight and obesity</td>
<td>BMI ≤25 kg m⁻²</td>
</tr>
<tr>
<td>Physical exercise</td>
<td>≥30 min 5 times per week</td>
</tr>
<tr>
<td>Exogenous hormones</td>
<td>Nil</td>
</tr>
<tr>
<td>Infections</td>
<td>Nil</td>
</tr>
<tr>
<td>Radiation – ionising</td>
<td>As in 1903 birth cohort</td>
</tr>
<tr>
<td>Radiation – solar (UV)</td>
<td>Nil</td>
</tr>
<tr>
<td>Occupational exposures</td>
<td>Minimum of 6 months</td>
</tr>
<tr>
<td>Reproduction: breast feeding</td>
<td></td>
</tr>
</tbody>
</table>

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.

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

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 (gene 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 exposures 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 categories are chronic diseases are collectively rather than individually. To develop a more comprehensive exposure, one that toxic effects

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, 8:30–9:00; Friday, 8:30–Noon • NAS Building, 2100 C Street, NW, Auditorium

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)
Capturing exogenous and endogenous exposures

The exposome includes all chemicals in the internal chemical environment.

External Environment (e.g.)
- Tobacco Smoke
- Infections
- Diet
- Food Toxins
- Drugs

Internal Environment (e.g.)
- Obesity
- Chronic Inflammation

Cancer Genome and Epigenome (e.g.)
- Genetic Susceptibility
- Driver Gene Targets

Cancer Biomarkers of Risk and Prognosis

Cancer Prevention and Screening Strategies

Cancer Therapy

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 of carcinogenesis, cancer therapy, and cancer prevention.

A. Schetter and C. Harris, PNAS, 2012, 109: 7955-6
Exposome-wide association studies (EWAS)

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

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

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

Causal pathway (c)

G → Rc → Pc → Disease traits

E

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

Disease traits → Rr → Pr → Secondary traits

Reactive pathway (r)

Mr

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

Genetic modifications (mutations)
Post-translational modifications
Epigenetic modifications

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9
If causal exposures operate primarily through small molecules ($M_c$) and proteins ($P_c$), then EWAS require metabolomics and/or proteomics.
Biospecimens for EWAS?

Causal biomarkers
(exposure)

Reactive biomarkers
(disease)

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

Microbiome products

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

DATA-DRIVEN DISCOVERY (EWAS)

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
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
- Genomics, epigenomics, transcriptomics, & experiments

Biomarkers of disease
- Disease stage and response to therapy

DATA-DRIVEN DISCOVERY (EWAS)

KNOCKLEGE-DRIVEN APPLICATIONS

- Molecular epidemiology
- Exposure biology
- Systems biology
- Drug development

Causality and prevention
- Diagnosis, prognosis and treatment

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9
Diseased vs. healthy (case-control studies)

Untargeted designs

Discriminating features

Chemical identification

Candidate biomarkers

Biomarkers of exposure

Biomarkers of disease

Targeted designs

Molecular epidemiology

Exposure biology

Systems biology

Drug development

Causality and prevention

Diagnosis, prognosis and treatment

Molecular systems biology

Exposure biology

Identify sources & measure exposures

Genomics, epigenomics, transcriptomics & experiments

Drug development

Disease stage and response to therapy

Dr. S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9
EWAS: proof of concept (Metabolomics via NMR & MS)

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).

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

S. Rappaport, *Biomarkers*, 2012, 17(6), 48: 3-9
An EWAS of colorectal cancer

Possible omic features:
900 Da x 500 features/Da ≈ 0.5M features
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
Follow up measurements of CRC-446

Biomarker highly associated with CRC
Uncorrelated with CRC stage
Does not return to normal after treatment
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