Proof-of-Concept of Environment-Wide Association Studies (EWAS) on Common Disease

Chirag Patel
Division of Systems Medicine, Department of Pediatrics
Stanford University School of Medicine

National Academy of Sciences Exposome Workshop
12/8/2011

chirag.patel@stanford.edu
twitter: @chiragjp
Common disease is a function of genes and environment...

...and we’re exposed to many environmental factors...

...but, lack methods and data to comprehensively and systematically connect the environment with common disease.
... and the case is different with genetics (e.g., genomics)!
Common disease is function of genes and environment...

WHO prioritized diseases:
- 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

Characterize common variation

Measurement tools

Comprehensive, high-throughput analyses

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

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

e.g., Environment-wide Association Study (EWAS)
Proof of Concept of EWAS

1. Background and Methods
2. Examples: Type 2 Diabetes, Serum Lipid Levels
3. Checking Validity and An “LD” Map of the Environment?
4. Conclusions
5. Informatics for the Environment
Connecting Disease with Exposures: Drawbacks in Environmental Epidemiology

<table>
<thead>
<tr>
<th></th>
<th>E+</th>
<th>E-</th>
</tr>
</thead>
<tbody>
<tr>
<td>diseased</td>
<td>👨‍⚕️ 👨‍⚕️</td>
<td>👨‍⚕️</td>
</tr>
<tr>
<td>non-diseased</td>
<td>👨‍⚕️👨‍⚕️👨‍⚕️</td>
<td>👨‍⚕️👨‍⚕️👨‍⚕️👨‍⚕️</td>
</tr>
</tbody>
</table>

“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
Genome-Wide Association Studies (GWAS)

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

What genetic loci are associated to disease?
**Environment-Wide Association Studies (EWAS)**

- **β-carotene**
- **2-hydroxyfluorene**

What specific environmental “loci” are associated to disease?: ie, T2D, lipid levels, obesity, etc?
Aim 1: EWAS Methods

Why “EWAS”?

What environmental factors are associated to disease?

A Massive, Ongoing, and Significant Public Health Survey

http://www.cdc.gov/nchs/nhanes.htm
## NHANES

### Environmental Factor “E-Chip”

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>age, race/ethnicity, income, education</td>
<td>10,000</td>
</tr>
<tr>
<td>Questionnaires</td>
<td>Drug Use, Physical Activity, Occupation, Food Frequency</td>
<td>10,000</td>
</tr>
<tr>
<td>Physical Exam and Clinical Laboratory Measures</td>
<td>Blood Pressure, Height, Weight, Fasting Blood Glucose, Cholesterol</td>
<td>~3,000</td>
</tr>
<tr>
<td>Markers of Exposure (serum and urine)</td>
<td>Nutrients/Vitamins, Metals, Hydrocarbons, Phthalates, Phenols, Infectious Agents, Allergens</td>
<td>~3,000</td>
</tr>
</tbody>
</table>
## Aim 1: EWAS Methods

### Environmental Measures by Category and Cohort

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrylamide</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Allergen Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Bacterial</td>
<td>8</td>
<td>13</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Cotinine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Diakyl</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Dioxins</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Furans</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Heavy Metals</td>
<td>18</td>
<td>18</td>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td>Hydrocarbons</td>
<td>14</td>
<td>22</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>Latex</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Carotenoid Nutrients</td>
<td>0</td>
<td>6</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Mineral Nutrients</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Vitamin A</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Vitamin B</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Vitamin C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vitamin D</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vitamin E</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Polychlorinated Biphenyls</td>
<td>23</td>
<td>26</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>Perchlorate</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Pesticides, Atrazine</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Pesticides, Carbamate</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pesticides, Chlorophenol</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pesticides, Organochlorine</td>
<td>10</td>
<td>13</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Pesticides, Organophosphate</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Pesticides, Pyrethroid</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Phenols</td>
<td>15</td>
<td>11</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Phthalates</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Phytoestrogens</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Polybrominated Ethers</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Polyfluorochemicals</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Virus</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Volatile Compounds</td>
<td>29</td>
<td>14</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>169</strong></td>
<td><strong>182</strong></td>
<td><strong>258</strong></td>
<td><strong>96</strong></td>
</tr>
</tbody>
</table>
Aim 1: EWAS Methods

**EWAS Methodology**

4 individual cohorts

**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**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bisphenol A</td>
<td>.</td>
<td>.</td>
<td>0.002</td>
<td>0.01</td>
</tr>
<tr>
<td>PCB199</td>
<td>0.1</td>
<td>0.02</td>
<td>0.03</td>
<td>.</td>
</tr>
<tr>
<td>β-carotene</td>
<td>NA</td>
<td>0.0001</td>
<td>0.02</td>
<td>0.002</td>
</tr>
<tr>
<td>cotinine</td>
<td>0.03</td>
<td>0.01</td>
<td>0.9</td>
<td>.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### EWAS Methodology, cont’d

**False Discovery Rate Estimation**


**foreach:**

- bisphenol A
- PCB199
- β-carotene
- cotinine

... 

**foreach:** 1 to \(B\)

permute labels or residuals \(B\) times

compute FDR

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bisphenol A</td>
<td>.</td>
<td>.</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>PCB199</td>
<td>0.1</td>
<td>0.02</td>
<td>0.03</td>
<td>.</td>
</tr>
<tr>
<td>β-carotene</td>
<td>NA</td>
<td>0.01</td>
<td>0.1</td>
<td>0.05</td>
</tr>
<tr>
<td>cotinine</td>
<td>0.03</td>
<td>0.01</td>
<td>0.9</td>
<td>.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Tentative Validation**

\(FDR < \text{threshold}\) in 2 or greater cohorts?  
AND  
\(\text{sign(β}_{\text{factor}})\) equal for cohorts?
Proof of Concept of EWAS

1. Background and Methods
2. Examples: Type 2 Diabetes, Serum Lipid Levels
3. Checking Validity and An “LD” Map of the Environment?
4. Conclusion
5. Informatics for the Environment
What environmental factors are associated with Type 2 Diabetes?
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

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

<table>
<thead>
<tr>
<th>lipid type</th>
<th>risk for CHD (for 1% increase in lipids)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL-Cholesterol</td>
<td>1% increased risk</td>
</tr>
<tr>
<td>HDL-Cholesterol</td>
<td>2% decreased risk</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>(increased risk)</td>
</tr>
</tbody>
</table>

Aim 2: EWAS examples

EWAS on HDL-C

-\log_{10}(p\text{value})

0 1 2 3 4 5

log_{10}(HDL-C)

BMI, SES, ethnicity, age, age^2, sex
N=1000-3000

cohort markers
1999-2000 ●
2001-2002 ■
2003-2004 ●
2005-2006 ▲

Aim 2: EWAS examples

**Effect Sizes For Validated Factors:**

**HDL-C**

<table>
<thead>
<tr>
<th>cohort</th>
<th>N</th>
<th>pvalue</th>
<th>effect (mg/dl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cis–b–carotene</td>
<td></td>
<td></td>
<td>3e–04</td>
</tr>
<tr>
<td>2001–2002</td>
<td>7478</td>
<td>3e–04</td>
<td>2</td>
</tr>
<tr>
<td>2003–2004</td>
<td>6790</td>
<td>9e–04</td>
<td>3</td>
</tr>
<tr>
<td>2005–2006</td>
<td>6264</td>
<td>2e–04</td>
<td>3</td>
</tr>
<tr>
<td>combined</td>
<td>7151</td>
<td>3e–12</td>
<td>3</td>
</tr>
<tr>
<td>Iron, Frozen Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999–2000</td>
<td>6383</td>
<td>0.009</td>
<td>2</td>
</tr>
<tr>
<td>2001–2002</td>
<td>7457</td>
<td>0.003</td>
<td>2</td>
</tr>
<tr>
<td>2003–2004</td>
<td>2706</td>
<td>0.006</td>
<td>2</td>
</tr>
<tr>
<td>2005–2006</td>
<td>2524</td>
<td>0.002</td>
<td>2</td>
</tr>
<tr>
<td>combined</td>
<td>6764</td>
<td>6e–11</td>
<td>2</td>
</tr>
<tr>
<td>Retinyl stearate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001–2002</td>
<td>7251</td>
<td>0.002</td>
<td>-1</td>
</tr>
<tr>
<td>2003–2004</td>
<td>6790</td>
<td>0.003</td>
<td>-1</td>
</tr>
<tr>
<td>2005–2006</td>
<td>6337</td>
<td>0.002</td>
<td>-2</td>
</tr>
<tr>
<td>combined</td>
<td>8421</td>
<td>4e–05</td>
<td>-1</td>
</tr>
<tr>
<td>Folate, serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001–2002</td>
<td>7468</td>
<td>0.004</td>
<td>1</td>
</tr>
<tr>
<td>2003–2004</td>
<td>7267</td>
<td>0.02</td>
<td>1</td>
</tr>
<tr>
<td>combined</td>
<td>9559</td>
<td>2e–05</td>
<td>1</td>
</tr>
<tr>
<td>Vitamin C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003–2004</td>
<td>6799</td>
<td>0.006</td>
<td>2</td>
</tr>
<tr>
<td>2005–2006</td>
<td>6911</td>
<td>0.02</td>
<td>1</td>
</tr>
<tr>
<td>combined</td>
<td>4852</td>
<td>0.002</td>
<td>1</td>
</tr>
<tr>
<td>Vitamin D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001–2002</td>
<td>7056</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>2003–2004</td>
<td>7273</td>
<td>0.004</td>
<td>2</td>
</tr>
<tr>
<td>2005–2006</td>
<td>6966</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>combined</td>
<td>7401</td>
<td>1e–06</td>
<td>2</td>
</tr>
<tr>
<td>g–tocopherol</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001–2002</td>
<td>7428</td>
<td>0.001</td>
<td>-1</td>
</tr>
<tr>
<td>2003–2004</td>
<td>6790</td>
<td>0.01</td>
<td>-1</td>
</tr>
<tr>
<td>combined</td>
<td>9216</td>
<td>6e–06</td>
<td>-1</td>
</tr>
</tbody>
</table>

% change = Δ ± 1 SD in Exposure

18 validated factors

combined adjusted for:

BMI, SES, ethnicity, age, age², sex, waist circumference, diabetes (FBG > 125 mg/dL), blood pressure

comparable to genetic effect sizes¹!

% change


deity (FBG > 125 mg/dL), blood pressure

comparable to genetic effect sizes¹!
Proof of Concept of EWAS

1. Background and Methods
2. Examples: Type 2 Diabetes, Serum Lipid Levels
3. Checking Validity and An “LD” Map of the Environment?
4. Conclusion
5. Informatics for the Environment
Assessing Validity of Estimates

*example: HDL-C*

Could the disease “lead” to exposure? “Reverse causality”

Could there something *confounding* the association?

Are associations *independent*? How to untangle the web of exposure?
Longitudinal Study:
“Gold Standard” for Validation

- exposure changing through time
- reverse causality bias
- compute disease risk
- how will we use “exposome” longitudinally?

HDL-Cholesterol (mg/dL)

γ-tocopherol supplements for CHD individuals?

[low] [high]

γ-tocopherol

age/time
Addressing Confounding Bias with the Exposome

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

“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
Aim 3: EWAS and Validity

**Correlation Structure of the Exposome?**

- β-carotene
- hydrocarbons
- γ-tocopherol

Are associations *independent*? How to untangle the complex web of exposure?

**Analogy: “Linkage Disequilibrium”**

Correlation between occurrence of genetic loci

In GWAS, allows one to trace to the “causal” locus.

Sladek et al., *Nature Genetics*. 2007
Aim 3: EWAS and Validity

"LD" of Exposure Biomarkers in NHANES

for each pair:
- Partial $\rho$
  - age, BMI (serum)
  - age, creatinine (urine)

permuted data to produce
- "null $\rho$"

filtered those $\rho > 0.3$ or $< -0.2$
- (5th, 95th percentile)
- sought replication in $> 1$ cohort

Red: positive $\rho$
Blue: negative $\rho$
thickness: $|\rho|$

EWAS: Conclusion and Discussion

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

• novel associations for T2D and HDL-C

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

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

• Correlative web is dense: what is the “LD” of the exposome?
Proof of Concept of EWAS

1. Background and Methods
2. Examples: Type 2 Diabetes, Serum Lipid Levels
3. Checking Validity and An “LD” Map of the Environment?
4. Conclusion
5. Informatics for the Environment
Sequencing of the genome

Characterize common variation

Measurement tools

Nature, 2001

Science, 2001

International HapMap Project

HapMap project:

2001-2003 (ongoing)

“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*

Human Genome Project: Informatics Contributions

http://genome.ucsc.edu
Informatics for the Environment?  
A GEO equivalent?

Gene Expression Omnibus

required submission

> 600,000 individuals (!)

enabled new investigations

collaboration & transparency

driven standardization

What if we had the same for

environmental exposure measures?


accessed 12/1/2011
Acknowledgments

- Atul Butte
- Jay Bhattacharya
- Mark Cullen
- John Ioannidis
- Rob Tibshirani
- Biomedical Informatics at Stanford
- CDC / NCHS
- NHANES and participants!
- National Academy of Sciences
  - Steve Rappaport
- National Library of Medicine
- March of Dimes Preterm Birth Research Center
- Stanford Prevention Center / SPECTRUM Population Studies Innovation Grant

email: chirag.patel@stanford.edu
twitter: @chiragjp