Environmental Influences on the Epigenome: Using SWIFT Text Mining Tool to Explore the State of the Science

Katherine Pelch, PhD
Office of Health Assessment and Translation
National Institute of Environmental Health Sciences

NTP Board of Scientific Counselors Meeting
June 16, 2015
Where are these links the strongest?

Environmental Exposure

Epigenetic Mechanisms

Health Outcome
There is a growing interest in epigenetics.

• 2012-2017 NIEHS Strategic Plan
  - “the effects of the environment on the epigenetic regulation of biological and pathological processes”

• Cross-divisional implementation group
  - “Where are the links between environmental exposures and health outcomes via epigenetic mechanisms the strongest?”

• Evaluation design team
How was the research question refined?

Epigenetics

Environmental Exposures

N=107,000 Records
How was the research question refined?

- More focused research question
- Test drive text-mining and machine learning tools
- Develop a new report format ("scoping report")
How was the research question refined?
What is the extent of the evidence?

Specific Aims

- Search for genome-wide analyses of DNA methylation
- Identify literature for seed sets
- Relevancy rank the search results
- Categorize records by type of exposure, health outcome, and evidence stream (human, animal, *in vitro*)
- Visualize the results

SCOPING REPORT
How was the literature search crafted?

- Global DNA methylation
- DNA methylation
- Genome-wide
- Genome-wide techniques
- Epigenetics

Global DNA methylation OR
(Epigenetics OR DNA methylation) AND
(genome-wide OR genome-wide techniques)

N=35,119 records retrieved
How were records processed?

- Records identified by PubMed database search (N=35,119)
- Records considered “in scope” (N=21,221)
- Records excluded because “out of scope” (N=13,898)
  - Not a research article
  - Published before 1999
  - Not animal model
What topics are in this literature?

- 50 topics are automatically generated based on the most frequently used words

- Topics were surveyed to get a sense of the types of records that had been retrieved

- Topics were screened and ~120 records were identified to serve as a seed set (60 relevant★ 67 not relevant ✗)
How do the records compare to seed sets?

Comparison of the genome-wide DNA methylation profiles between fast-growing and slow-growing broilers.


**Abstract**

Growth traits are important in poultry production, however, little is known for its regulatory mechanism at epigenetic level. Therefore, in this study, we aim to compare DNA methylation profiles between fast- and slow-growing broilers in order to identify candidate genes for chicken growth. Methylated
Which were the most relevant records?

25% = 5,306 records

most relevant to topic
How were records processed?

- Records identified by PubMed database search (N=35,119)
- Records considered “in scope” (N=21,221)
  - Records excluded because “out of scope” (N=13,898)
    - Not a research article
    - Published before 1999
    - Not animal model
- Topic modeling & seed set selection (N=21,221)
- Records ranked for similarity to seed sets (N=21,221)
- Categorized top 25% of ranked records (N=5,306)
- Bottom 75% of ranked records (N=15,915)
How effective was the enrichment?

Original
N=35K

Top 25%
N=5K

T=title
A=abstract
S=suppl. info
M=MeSH
What health outcomes are investigated?

- Endocrine System Disease
- Respiratory Tract Disease
- Nutritional and Metabolic Disease
- Virus Diseases
- Bacterial Infections
- Musculoskeletal Disease
- Stomatognatic Disease
- Cardiovascular Disease
- Parasitic Disease
- Eye Diseases
- Mental Disorders
- Neoplasms
- Digestive System Disease
- Skin & Connective Tissue Disease
- Hemic and Lymphatic Disease
- Male Urogenital Disease
- Female Urogenital Disease
- Immune System Disease
- Other

N=5,306 records
What types of exposures are investigated?

N=1,130 of 5,306 records (21%)
Evaluating exposure & health outcomes

Exposure = pesticide

N=26 records

- Female Urogenital Disease
- Male Urogenital Disease
- Neoplasms
- Congenital, Hereditary Disease
- Nutritional & Metabolic Disease
- Immune System Disease
- Virus Diseases
- Hemic & Lymphatic Disease
- Bacterial Infections
- Respiratory Tract Disease
- Skin & Connective Tissue Disease
- Endocrine System Disease

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>vinclozolin</td>
<td>7</td>
</tr>
<tr>
<td>5-azacytidine</td>
<td>4</td>
</tr>
<tr>
<td>2,3,7,8-tetrachlorodibenzofuran</td>
<td>4</td>
</tr>
<tr>
<td>methoxychlor</td>
<td>3</td>
</tr>
<tr>
<td>decitabine</td>
<td>3</td>
</tr>
<tr>
<td>cytosine</td>
<td>3</td>
</tr>
<tr>
<td>parathion</td>
<td>2</td>
</tr>
<tr>
<td>ethanol</td>
<td>2</td>
</tr>
<tr>
<td>diazinon</td>
<td>2</td>
</tr>
<tr>
<td>arsenic trioxide</td>
<td>2</td>
</tr>
<tr>
<td>terbufos</td>
<td>1</td>
</tr>
<tr>
<td>rotenone</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

Tox21 Chemicals
• Scoping reports useful for large, complex or emerging literatures
  – Get a sense of the literature
  – Highlights the variety in the field
  – Line up future projects
  – Research tool for data mining

• Implemented a new text mining tool
  – Useful for visualizing
Acknowledgements

• Office of Health Assessment and Translation
  – Abee Boyles
  – Kembra Howdeshell
  – Katie Pelch
  – Andrew Rooney, Deputy Director
  – Kyla Taylor
  – Kristina Thayer, Director
  – Vickie Walker

• Office of Scientific Information Management
  – Stephanie Holmgren

• SCIOME, LLC
  – Brian Howard
  – Jason Phillips
  – Ruchir Shah

• Evaluation Design Team
  – Douglas Bell
  – Lisa Chadwick
  – David Fargo
  – Stephanie London
  – Alex Merrick
  – Jack Taylor
  – Christina Teng
  – Fred Tyson
  – Paul Wade
  – Lauren Wilson