Environmental Influences on the Epigenome: a Scoping Report

Project Leader:

Katherine Pelch, PhD

DNTP, Office of Health Assessment and Translation

INTRODUCTION

Background

The National Institutes of Health (NIH) defines epigenetics as "the study of changes in the regulation of gene activity and expression that are not dependent on gene sequence" (NIH 2009, Project 2010). There is great interest in understanding how genome-wide chemical modifications to DNA may regulate gene activity without altering the DNA sequence itself and what role these modifications may play in health and disease, including: cancer, autoimmune disease, mental disorders, and diabetes, among other illnesses (NIH 2014). NIH is a major sponsor of epigenetics research, spending over $700 million on epigenetics in 2012 (Burris and Baccarelli 2014).

The National Institute of Environmental Health Science (NIEHS) is interested in understanding the effects of the environment on the epigenetic regulation of biological and pathological processes (NIEHS 2012).[1] A wide array of environmental factors has been reported to cause disruption to the epigenome, including: diet and nutrition, stress, and chemical and pharmaceutical exposures. Epigenetic modifications are thought to be a key mechanism behind the growing literature demonstrating developmental programing and transgenerational inheritance of health effects related to environmental exposures (Aiken and Ozanne 2014).

In 2012, NIEHS spent 7.1% of its total budget on environmental epigenetics research (Burris and Baccarelli 2014). In order to assess the impact of this research commitment, methods are needed to systematically identify the relevant literature so that we can begin the process of translating findings from individual studies into knowledge based on synthesizing critical findings across epigenetic studies. A systematic review of the evidence for "environmental influences on the epigenome" would be a challenge given the size of the literature to date, estimated at greater than 100,000 records (Table 1). Part of the complexity of identifying the relevant literature arises from the term epigenetics, the definition of which is still debated (Ledford 2008). Depending on which definition is used, the term epigenetics can encompass many different molecular modifications/mechanisms, for example, alterations in DNA methylation, histone modifications, microRNA expression, or other newer molecular modifications that are still being described.

One way to begin to investigate such a large and complex literature is to use text-mining algorithms and machine learning approaches to identify and characterize the type of research being conducted. Text mining and machine learning add meaning to text that is retrieved, extracted, and mined in an automated fashion (Ananiadou *et al.* 2006). In systematic review, this type of analysis is usually published as a scoping report to either show trends in research or identify targeted questions to address in future systematic reviews (Levac et al. 2010, Colquhoun et al. 2014, Wilson 2014).

Objective and Specific Aims

The overall objective of this scoping report is to identify published findings relevant to understanding the extent of the evidence linking environmental exposures to health outcomes via genome-wide alterations in DNA methylation.

---

[1] Goal 1 of the NIEHS 2012-2017 Strategic Plan was to *"identify and understand fundamental shared mechanisms of common biological pathways, e.g., inflammation, epigenetic changes, oxidative stress, mutagenesis, etc., underlying a broad range of complex disease, in order to enable the development of applicable prevention and intervention strategies."* Specifically, goal 1b was to "*investigate the effects of the environment on the epigenetic regulation of biological and pathological processes."*

Specific aims

1) Search the published literature (PubMed) for research that has used genome-wide analyses of DNA methylation;

2) Use topic-modeling capabilities of SWIFT (Sciome Workbench for Interactive, Computer-Facilitated Text-mining) to identify relevant literature;

3) Use the machine-learning capabilities of SWIFT to rank the search results based on relevance to the objective question;

4) Use the text-mining and machine-learning capabilities of SWIFT to categorize records by type of exposure, health outcome, and evidence stream (human, animal, in vitro); and

5) Use SWIFT to visualize and summarize the results of the relevancy ranking and categorization of the studies.

The overall objective and specific aims were based on a series of problem formulation steps that included assembling an NIEHS cross-divisional evaluation team with expertise in epigenetics, environmental health, text-mining/machine-learning software, systematic review, and information science. This scoping report can be used to identify and prioritize topic areas for future research and serve as a proof-of-concept analysis to evaluate the feasibility and reliability of using text mining/machine learning to evaluate large and complex literature bases.

METHODS

Problem Formulation

Understanding the role of the environment on altering the epigenetic state was identified as a primary goal of the NIEHS in the 2012-2017 Strategic Plan. Specifically, goal 1b was to "*investigate the effects of the environment on the epigenetic regulation of biological and pathological processes*." A cross-divisional implementation planning committee was formed within NIEHS to discuss how to execute this goal. The committee asked the Office of Health Assessment and Translation (OHAT) to assist in identifying literature that provides improved clarity on the extent of the evidence linking exposures to health outcomes via epigenetic mechanisms and ultimately to identify research areas where epigenetics links are the strongest. An evaluation team was subsequently formed with expertise in epigenetics, environmental health, text-mining/machine-learning software, systematic review, and information science.

Preliminary search of the literature and prioritization of genome-wide studies

Initial literature search strategies were developed for PubMed in July 2013 to identify all types of epigenetic studies ("gene-by-gene" as well as genome-wide) using a wide number of terms and concepts related to epigenetics, including DNA methylation, histone modifications, and microRNA signaling. The epigenetics search strategy was crossed with a search strategy designed to capture a broad range of environmental exposures or one targeted on selected exposures (air pollution, endocrine disruptors, heavy metals, flame retardants, pesticides, and phthalates). The resulting size of the queries is outlined below in Table 1 and specific search terminology can be found in Supplemental Materials (Table S1). The number of search results was >100,000 records when the broadest search strategy was used (epigenetics + environmental exposures) and ~4,500 for a focused search on DNA methylation + selected exposures.

Table 1. Exploratory PubMed Searches		
	Search Strategies	**# Records Retrieved**
1	Epigenetics + Environmental Exposures	107,647
2	Epigenetics + Selected Exposures (air pollution, endocrine disruptors, heavy metals, flame retardants, pesticides, phthalates)	26,631
3	DNA methylation + Exposure	19,558
4	DNA methylation + Selected Exposures (air pollution, endocrine disruptors, heavy metals, flame retardants, pesticides, phthalates)	4,472

Given the size of the literature base, a decision was made by the evaluation team to prioritize the focus on DNA methylation because most research investments to date have been focused on understanding environmental effects on this specific type of epigenetic modification (Burris and Baccarelli 2014). The focus of the evaluation was further prioritized to studies of mammalian systems that used genome-wide analyses of DNA methylation, rather than gene-by-gene analyses, as this seems to be where the future of this research lies.

Identifying Relevant Studies

Literature search strategy

A literature search strategy (Table S2) was developed to query PubMed to focus specifically on records that used genome-wide DNA methylation analyses. Genome-wide techniques were identified based on input from scientists with expertise in DNA methylation techniques and by consulting a recent review of DNA methylation analysis techniques (Laird 2010). The search was designed to be broad in order to capture the most relevant records regardless of the terminology used. The search strings were written to capture the concepts of "global DNA methylation," "DNA methylation", "genome-wide," "genome-wide techniques," and "epigenetics". These concepts were then combined in a stepwise manner resulting in the following query: "(Epigenetics AND (genome-wide OR genome-wide techniques)) OR Global DNA methylation OR (DNA methylation AND genome wide) OR (DNA methylation AND genome-wide techniques)." No limitations were imposed on publication year, evidence stream (i.e., human, animal, *in vitro*), type of health outcome evaluated, or type of exposure (or lack thereof). This search was run in PubMed on February 25, 2015 and retrieved 35,536 records.

There were several challenges encountered when designing this literature search strategy. First, the terms "genome-wide", "whole-genome", and "global" have been used interchangeably and imprecisely over time (e.g. "global" has been used to refer to analyses of repetitive sequence elements as well as locus specific analyses in some instances). Second, these terms are also descriptive of other types of molecular biology analyses such as measuring RNA or micro-RNA transcription or histone modifications. Third, the PubMed database is unable to search for concepts or keywords that are separated by intervening words (e.g. "*DNA-methylation* immunoprecipitation followed by *microarray* analysis").

Text mining to prioritize search results

The flow of records through the SWIFT software tool indicating how records were processed and the number of records evaluated at each step is shown in Figure 1. The retrieved records (n=35,119)[2] were uploaded into SWIFT and a series of queries was constructed to further refine the focus on identifying original research papers that used genome-wide technologies in model systems most applicable for human health. First, a query utilizing key words and machine learning was performed to tag and remove reviews, commentaries, editorials and other types of non-research article records. These queries used machine learning to discern what features research articles have compared to non-research articles and from this a fingerprint for "non-research articles" was created. The records were then assessed for their degree of similarity to the "non-research article" fingerprint. Next, research articles published prior to 1999 were removed because genome-wide techniques of interest were not developed prior to that time. Finally, in order to eliminate non-animal organisms (plants, fungi, viruses, etc.), records were restricted to having an 'animal' MeSH organism tag.

After these refinements, the topic-modeling functionality of SWIFT was used to survey topics in order to identify pockets of relevant and irrelevant literature. Topic modeling uses a generative Latent Dirichlet Allocation (LDA) model to probabilistically assign records based on the words they contain to topics (Blei *et al.* 2003, Blei and Lafferty 2009). The resulting topics, which are summarized by their most frequently used words, often have intuitive meanings (Figure 2). In this case, the topic modeling was used to identify "seed records" for the next step, which is training the machine-learning algorithm for relevancy ranking where the records are ranked from most relevant to the objective question (top of list) to least relevant to the objective question (bottom of list).

Training sets (also referred to as "seed sets") of 60 "positive" and 67 "negative" records were created by skimming the titles and abstracts of records contained within the topic model clusters. Positive records were confirmed to have at least one genome-wide DNA methylation analysis. Negative records did not have genome-wide analyses of DNA methylation, but may have had genome-wide analyses of mRNA or miRNA. Other records that used "global" analyses of DNA methylation, including those for multiple DNA repetitive elements, such as Alu elements and long interspersed nucleotide elements (LINE) were marked as negative seed records since repetitive elements were not the focus of this review. The remaining literature set (21,221 records) was relevancy ranked based on these training sets.

Categorizing search results

The top 25% of relevancy ranked records (n=5,306) were then tagged by health outcome, evidence stream, and exposure. The tags for evidence stream and exposure were based on targeted batch queries based on MeSH headings and title and abstract key words. Health-outcome tagging was performed using machine learning within SWIFT. For each top-level MeSH disease and mental disorder code, 5,000 records were selected at random from PubMed, and used as a training set to develop a classifier for the corresponding codes. Based on the output of the resulting set of classifiers, a given record can be assigned to multiple codes, and within each code, matching records are ranked according to the strength of the association

[2] There can be small differences between the downloadable version of PubMed used by SWIFT and the records available online. The missing PMIDs were absent in the downloadable PubMed database at the time the records were imported into SWIFT. For example records that were "e-published" ahead of their true publication date may not have been available for import into SWIFT.

RESULTS

Identifying the most relevant records

The PubMed literature search for records that used genome-wide DNA methylation analyses retrieved 35,119 records available for export to SWIFT. Filters to remove non-research articles such as reviews and commentaries, records published before 1999, and studies in plants, fungi, and viruses resulted in 21,221 records available for relevancy ranking. A training set of 60 positive and 67 negative records was used to rank the 21,221 records and the top 25% (n=5,306) were selected as most likely relevant and used for subsequent analysis (Figure 1).

Ranking performance was assessed by evaluating the lowest and highest ranked records and determining if each was relevant to the objective question. None of the 50 lowest ranked records were relevant to the question of "What is the extent of the evidence linking environmental exposures to health outcomes via genome-wide alterations in DNA methylation?" therefore the ranking of these at the bottom of the list was appropriate. Of the 50 highest ranked records, 18 were directly relevant to the objective question since they evaluated environmental exposures leading to altered DNA methylation that was evaluated on a genome-wide scale. The remaining 32 records all pertained to DNA methylation analyses, but 26 of these were not genome-wide analyses, 3 were methods papers and therefore lacked the exposure component, and 3 evaluated DNA methylation in non-mammalian systems. An additional 100 records were chosen using a random number generator and evaluated to calculate the predicted ranking performance, which is shown in Figure 3. This analysis suggests that at least 80% of all relevant records are expected to occur within the top 25% of records (i.e. recall is predicted to be >80% when the top 25% of ranked records are evaluated) (Figure 3).

The results of the filters and relevancy ranking are presented in word clouds that show the most overrepresented terms in a specified set of records (Figure 4). The size of the words is proportional to the word frequency and inversely proportional to the number of records containing that word (Manning *et al.* 2008). The word cloud from the unfiltered and unranked set of 35,119 records (Figure 4A) prominently exhibits many "off-topic" words such as histones, chromatin, and small interfering RNA. This is not unexpected as the literature search strategy was concept-based and these concepts can be applied to many different biological processes other than just DNA methylation. In comparison, more specific words such as DNA, methyl, and CpG are prominent in the word cloud from the top 25% of ranked records (Figure 4B). The side-by-side comparison of these two word clouds provides one way to visually assess the enrichment for relevant records that was achieved using SWIFT.

Categorization of relevancy ranked results

SWIFT was used to categorize the top 25% of relevancy ranked records (n = 5,306) by type of health outcome, evidence stream, and exposure. All of the records could be categorized based on health outcome or evidence stream. Only 21% (n=1130) of the top 25% ranked records were associated with an exposure. It is important to note that records can map to more than one category for any given categorization scheme. For example, see below.

The neoplasms category was the largest health-outcome category containing 3,391 records (Figure 5). The second and third largest categories were digestive system diseases (n=1,524) and female urogenital diseases (n=787), respectively. Over half of the records were associated with a human evidence stream (Figure 6). Approximately 24% of the records were associated with *in vitro* and 15% associated with the animal evidence stream. The largest category of exposure (Figure 7) was "general environment" and contained general terms such as "environmental pollutant", "endocrine disruptor", "hazardous

substances", and "water pollutant". The second largest category was diet and nutrition. Several (207 of 742) of the records that mapped to "general environment" also mapped to the more specific exposure categories because the titles and/or abstracts from these records contained both the specific and non-specific exposure terms. The second largest category was "diet and nutrition" which contained 212 records.

A key feature of SWIFT is that it allows the user to explore how various categories overlap by using Boolean operators to intersect selected categories. As an example, Figure 8 shows how the health outcome categorization changes when exploring two different exposure categories. There were 26 records associated with pesticide exposure and these were primarily related to endocrine system diseases, female urogenital disease and neoplasm categories. In comparison, there were 41 records tagged with stress as an exposure. The largest health outcomes associated with stress were anxiety disorders, and nervous system diseases. An alternative way of displaying how the exposure and health outcome categories intersect is provided in the heat map in Figure 9. The heat map shows the intersection of the exposure categories on the horizontal axis and the health-outcome categories on the vertical axis and the number of records in each exposure/health-outcome intersection are indicated. The largest pockets of literature are colored red, small pockets are colored blue, and pockets with no literature are colored white.

DISCUSSION

The current text-mining analysis focuses on genome-wide studies of DNA methylation and can be considered a proof-of-concept analysis for ways to use advances in text mining and machine learning to prioritize and categorize large and complex literature bases. A potential future project is to apply these approaches to the much larger literature base of gene-by-gene studies as well as other types of epigenetic modifications in addition to DNA methylation.

In the current work, we used SWIFT to first remove "out-of-scope" records (non-research articles, those published prior to 1999, and non-animal records). We then used topic modeling to explore pockets of literature that were relevant or irrelevant to the objective question. While exploring the topics, a training set containing positive and negative seed records was created which was subsequently used to train the program to relevancy rank the remaining 21,221 records. The top 25% of ranked records were then categorized by health outcome, evidence stream, and exposure.

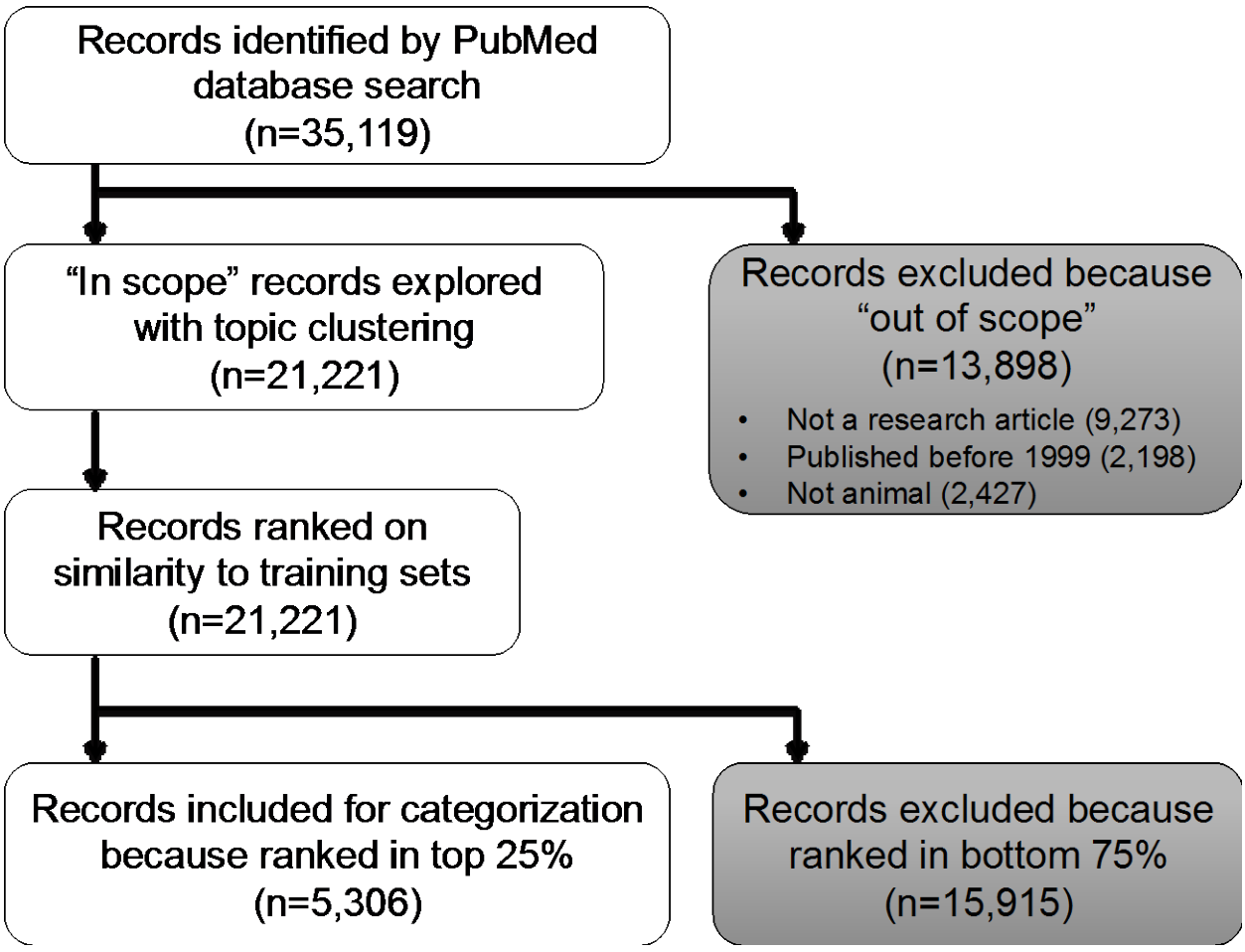
SWIFT was a useful tool in regards to both prioritization and categorization of the records based on health outcome, evidence stream and exposure. A current limitation and area of active methods development, however, is the lack of precision that is observed even in the top ranked records. That is, at this point not all of the records in the top 25% or ranked records are actually relevant to the objective question (the ranking procedure has good recall performance but the precision is unknown). One improvement could be to construct a more specific literature search; however this could reduce the recall of relevant records. Another approach is to improve the seed record selection and/or to use iterative training and limited screening in order to increase the precision while also maintaining high recall.

Many advances have been made in text-mining and machine-learning approaches to exploring large literature databases and these approaches may be the most efficient way to summarize the state of science for a research area as broad as epigenetics (Blei and Lafferty 2009, Thomas *et al.* 2011, Mimno 2012, Miwa *et al.* 2012, BioCreativeIV 2013, Miwa *et al.* 2014, Shemilt *et al.* 2014, O'Mara-Eves *et al.*

2015). The use of text mining tools to identify and assist researchers in identifying records for inclusion during the screening phase of systematic review is an area of active development and results have been promising so far (O'Mara-Eves *et al.* 2015). Text-mining tools, when used for prioritizing the order of record screening, seems to be a safe use of the technology and can potentially reduce the time burden associated with screening. A recent review on the use of text mining in systematic review found that most papers suggested a saving in workload between 30% and 70% might be possible although the saving in workload would be accompanied by some loss of relevant records, i.e., a 95% recall rate at meaningful workload reduction rates (O'Mara-Eves *et al.* 2015).

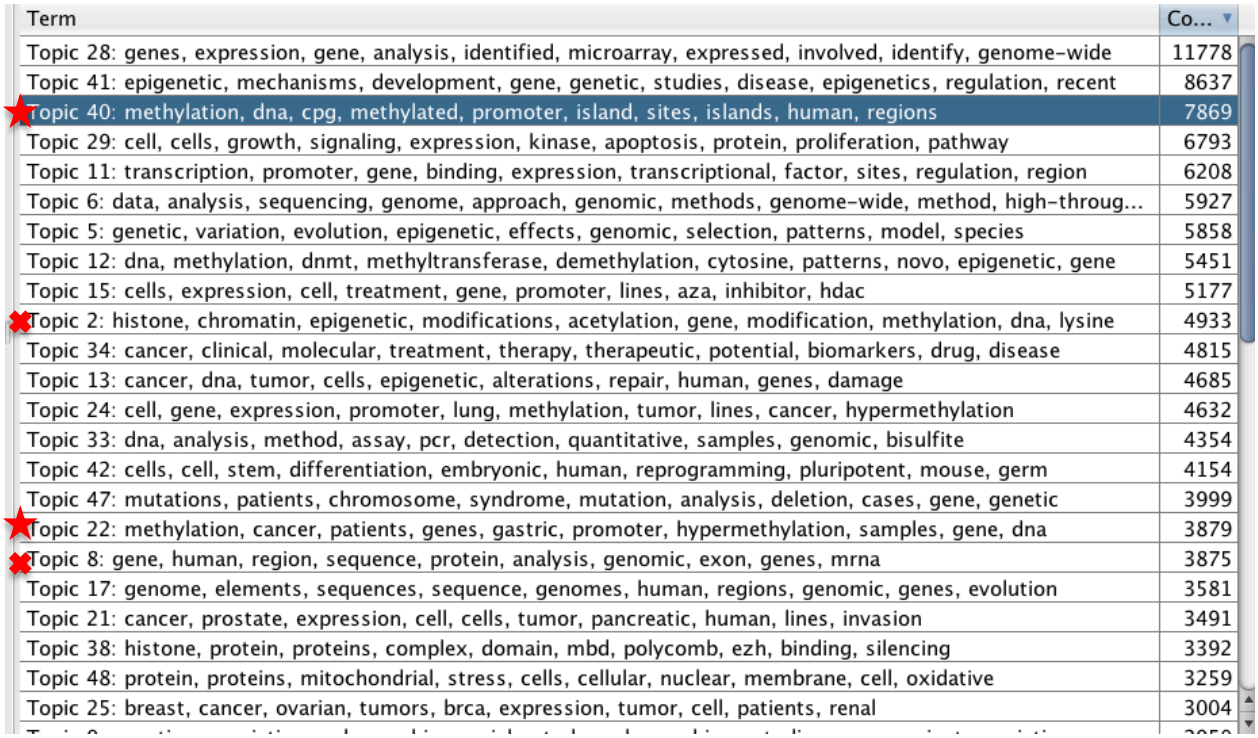
SWIFT incorporates state of the art information retrieval approaches including search, topic modeling and supervised classification into an interactive, integrated workbench. Although classification methods have previously been shown to be helpful in the context of systematic review (e.g. Abstrackr, GAPScreener, EPPI-Reviewer), SWIFT's interactive approach and exploratory nature may make it more useful than related tools for the purpose of preparing scoping reviews and during the problem formulation steps undertaken during full systematic reviews (Yu *et al.* 2008, Thomas *et al.* 2010, Wallace *et al.* 2012). At this time, we are testing the incorporation of a modified approach where text mining, machine learning and expert screening will be used in an iterative fashion to improve both precision and recall, while reducing the screening burden such that accurate and comprehensive scoping reports can be produced in the most efficient manner. Future developments also include full-text searching, extracting data/language from tables and graphs, collating gene and protein names within a record, and developing metrics to evaluate research trends over time.

Figure 1. Study flow diagram



Study flow diagram indicating how records were processed and the number of records evaluated at each step.

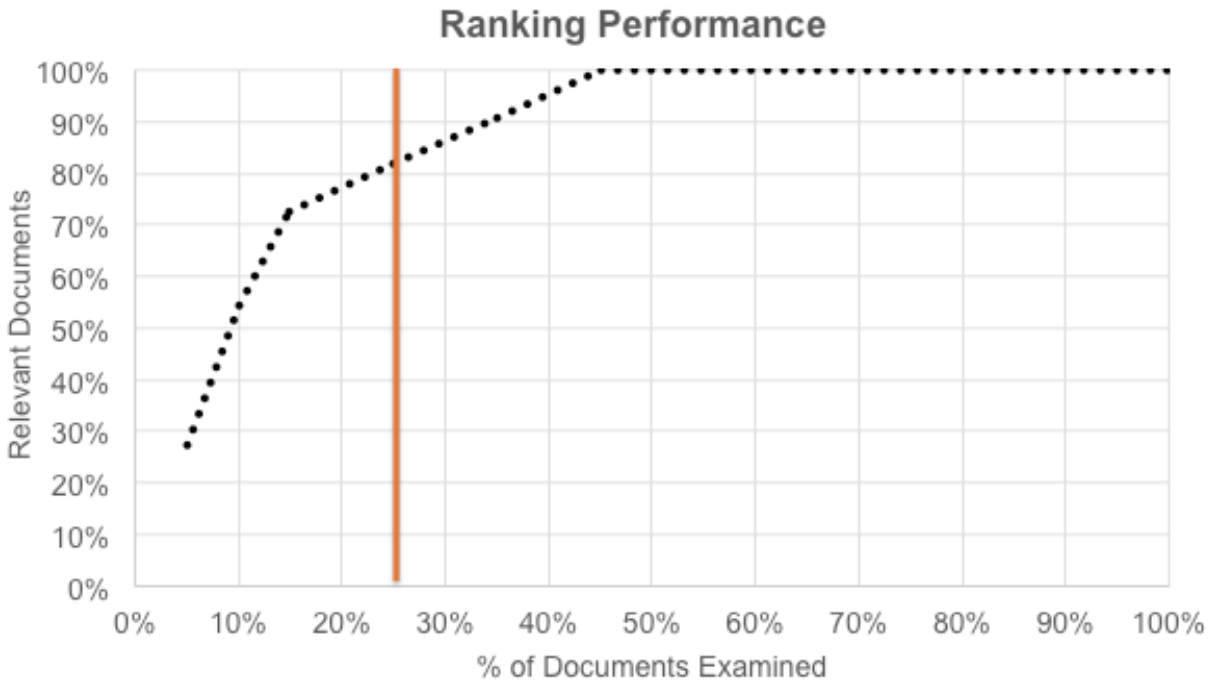
Figure 2. Topic modeling



Term	Co...
Topic 28: genes, expression, gene, analysis, identified, microarray, expressed, involved, identify, genome-wide	11778
Topic 41: epigenetic, mechanisms, development, gene, genetic, studies, disease, epigenetics, regulation, recent	8637
Topic 40: methylation, dna, cpg, methylated, promoter, island, sites, islands, human, regions	7869
Topic 29: cell, cells, growth, signaling, expression, kinase, apoptosis, protein, proliferation, pathway	6793
Topic 11: transcription, promoter, gene, binding, expression, transcriptional, factor, sites, regulation, region	6208
Topic 6: data, analysis, sequencing, genome, approach, genomic, methods, genome-wide, method, high-throug...	5927
Topic 5: genetic, variation, evolution, epigenetic, effects, genomic, selection, patterns, model, species	5858
Topic 12: dna, methylation, dnmt, methyltransferase, demethylation, cytosine, patterns, novo, epigenetic, gene	5451
Topic 15: cells, expression, cell, treatment, gene, promoter, lines, aza, inhibitor, hdac	5177
Topic 2: histone, chromatin, epigenetic, modifications, acetylation, gene, modification, methylation, dna, lysine	4933
Topic 34: cancer, clinical, molecular, treatment, therapy, therapeutic, potential, biomarkers, drug, disease	4815
Topic 13: cancer, dna, tumor, cells, epigenetic, alterations, repair, human, genes, damage	4685
Topic 24: cell, gene, expression, promoter, lung, methylation, tumor, lines, cancer, hypermethylation	4632
Topic 33: dna, analysis, method, assay, pcr, detection, quantitative, samples, genomic, bisulfite	4354
Topic 42: cells, cell, stem, differentiation, embryonic, human, reprogramming, pluripotent, mouse, germ	4154
Topic 47: mutations, patients, chromosome, syndrome, mutation, analysis, deletion, cases, gene, genetic	3999
Topic 22: methylation, cancer, patients, genes, gastric, promoter, hypermethylation, samples, gene, dna	3879
Topic 8: gene, human, region, sequence, protein, analysis, genomic, exon, genes, mrna	3875
Topic 17: genome, elements, sequences, sequence, genomes, human, regions, genomic, genes, evolution	3581
Topic 21: cancer, prostate, expression, cell, cells, tumor, pancreatic, human, lines, invasion	3491
Topic 38: histone, protein, proteins, complex, domain, mbd, polycomb, ezh, binding, silencing	3392
Topic 48: protein, proteins, mitochondrial, stress, cells, cellular, nuclear, membrane, cell, oxidative	3259
Topic 25: breast, cancer, ovarian, tumors, brca, expression, tumor, cell, patients, renal	3004

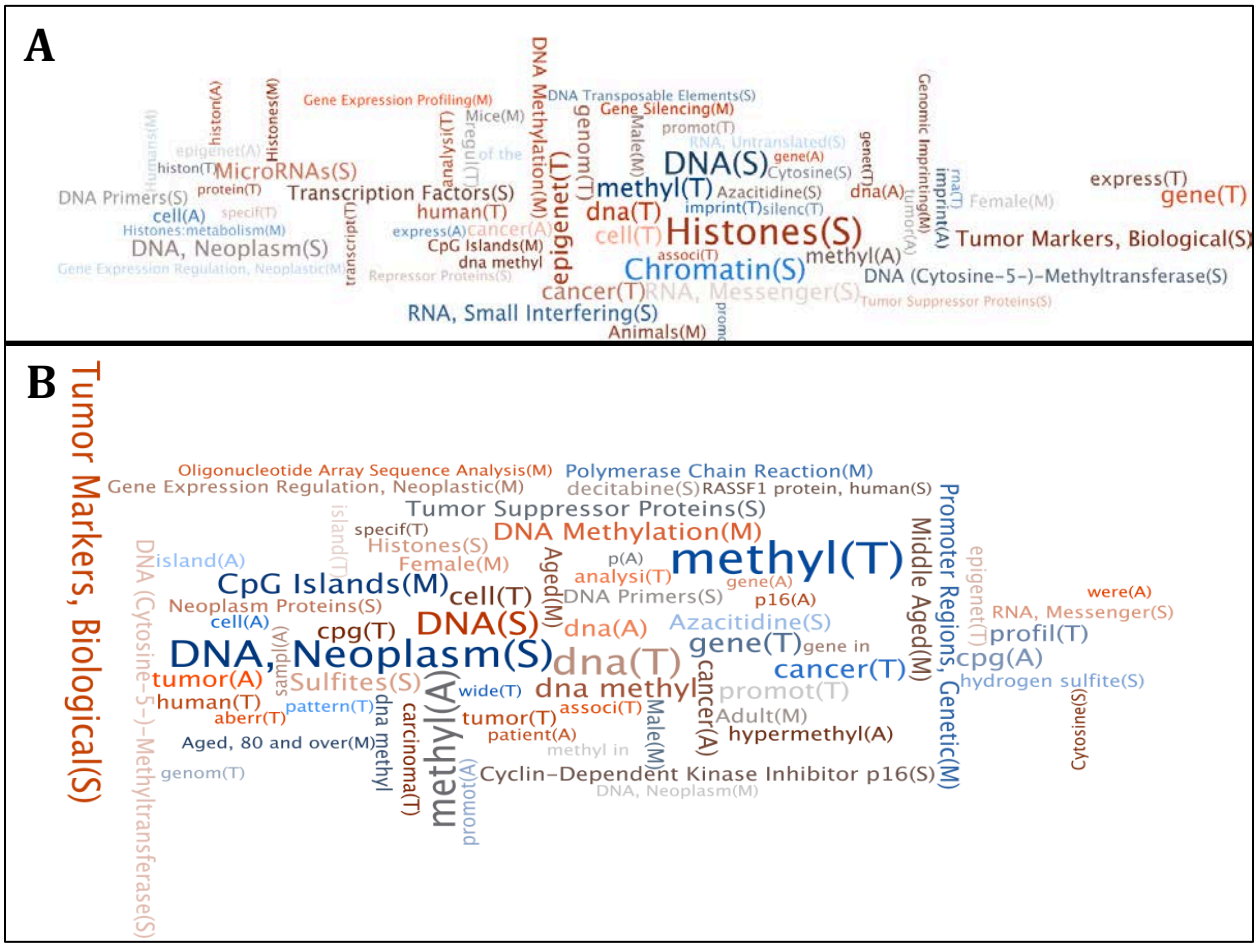
Topic modeling uses a generative Latent Dirichlet Allocation (LDA) model to probabilistically assign records based on the words they contain to topics. The machine-learning algorithms within SWIFT automatically populate 50 topics for a given project. The topics are summarized by their most frequently used words. The “Topic Number” has no relevance. In this figure topics are listed according to how many records are in each topic. Records can occur in more than one topic. Red stars and X’s indicate topics that more or less likely, respectively, to contain records relevant to the objective question. Various topics were explored to identify positive and negative seed records to be used for training the relevancy-ranking algorithm.

Figure 3. Predicted ranking performance



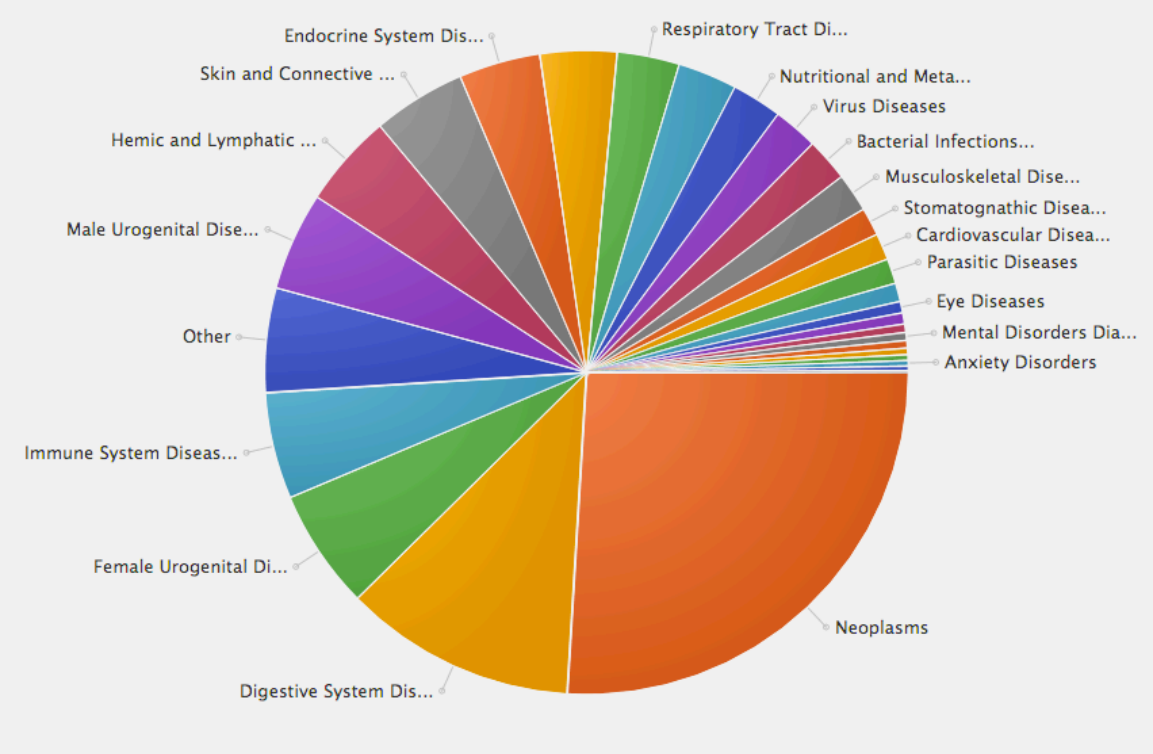
Ranking performance was predicted based on evaluating 200 records (of 21,221 possible). The black dotted line suggests that as more records are examined, then the larger the percentage of relevant records would be captured. The orange vertical line indicates that once 25% of records are evaluated, you are likely to have captured 80% of the relevant records.

Figure 4. Word clouds indicating overrepresented terms



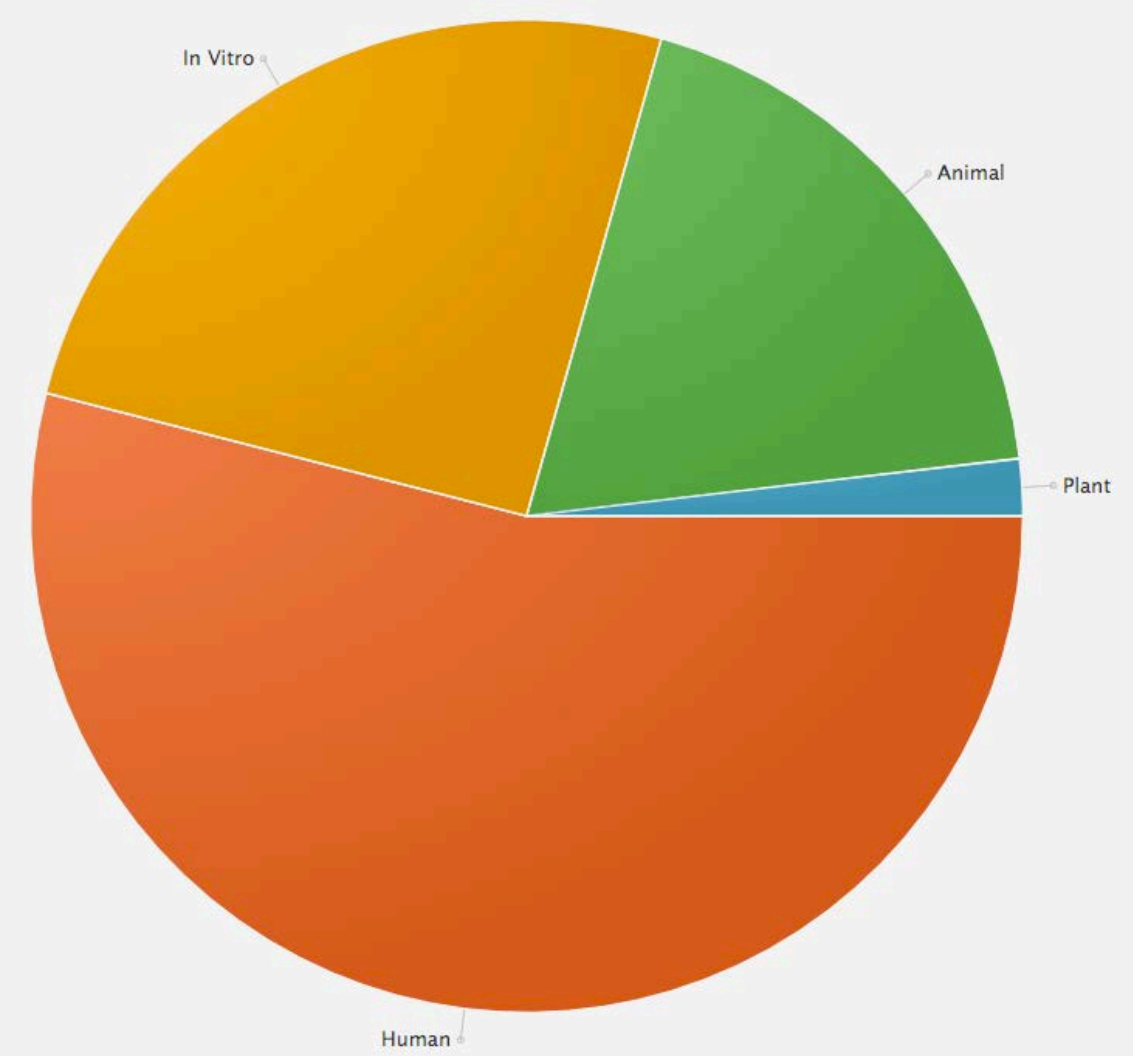
Word clouds indicate overrepresented words within a literature set. Word clouds from the (A) original, unfiltered search resulting in 35,119 records and (B) the top 25% ranked records are shown demonstrate an enrichment of relevant terms after filtering and ranking. Letters in parenthesis indicate where the word occurs: (T) title, (A) abstract, (M) MeSH (S) MeSH Supplementary Chemical. The size of the words is proportional to the word frequency and inversely proportional to the number of records containing that word.

Figure 5. Categorization by health outcome



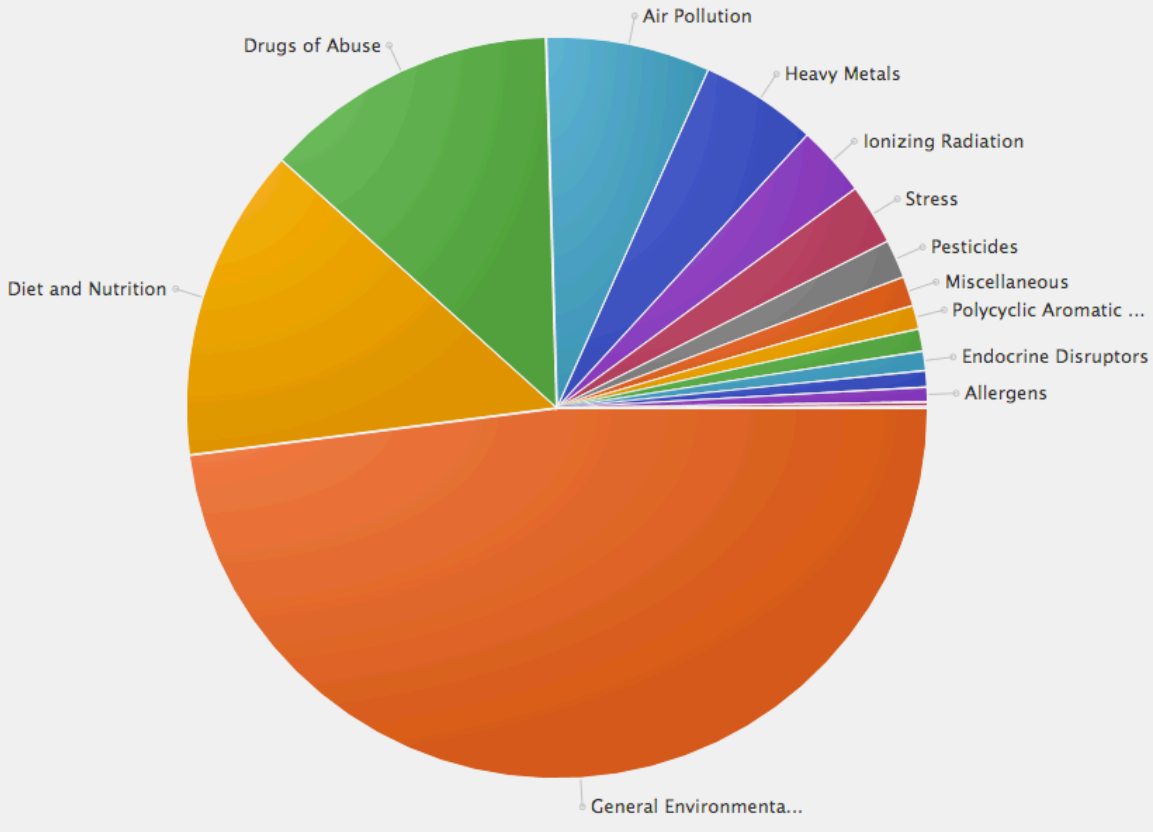
Categorization of the top 25% of ranked records based on health outcome. Health-outcome tagging was performed using machine learning within SWIFT where fingerprints for the top level MeSH disease and mental disorder codes were previously developed based on a random sampling of 5,000 records from each code. In order for a record to map to a given health outcome, it must match a pre-determined number of words in the health-outcome fingerprint.

Figure 6. Categorization by evidence stream



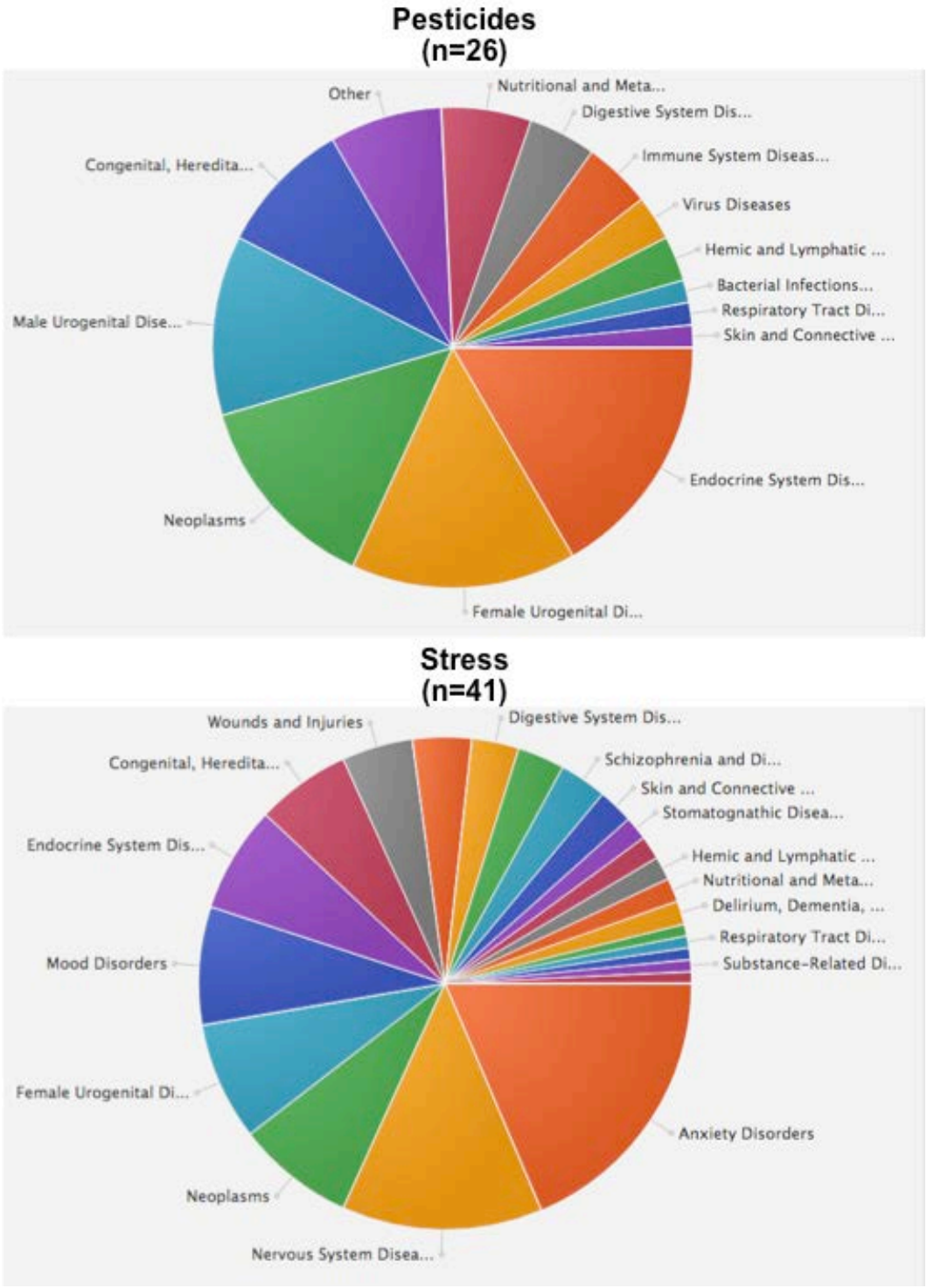
Categorization of the top 25% of ranked records based on evidence stream. Evidence stream was determined using a text-word search (see Supplemental Materials for more information).

Figure 7. Categorization of exposure



Categorization of the top 25% of ranked records based on exposure. Exposure was determined using a text-word search (see Supplemental Materials for more information). Only 1130 of 5306 records were associated with an exposure.

Figure 8. Intersection of exposure and health-outcome categories



Various exposure categories were selected and the numbers of records associated with each exposure are shown in parenthesis. Each exposure category was then intersected with the health-outcome categories.

Figure 9. Intersection of exposure and health-outcome categories

	Air Pollution	Allergens	Diet and Nutrition	Drugs of Abuse	Endocrine Disruptors	Flame Retardants	General Environmental Exposures	Heavy Metals	Ionizing Radiation	Miscellaneous	Occupational	Pesticides	Phthalates	Polycyclic Aromatic Hydrocarbons	Solvents	Stress
Bacterial Infections and Mycoses	22	19	106	162	2	0	315	33	23	12	0	28	0	10	6	41
Virus Diseases	19	13	101	220	2	0	239	15	20	3	3	14	1	4	1	26
Parasitic Diseases	13	10	56	122	0	0	148	6	13	3	0	13	0	2	0	24
Neoplasms	198	16	580	342	43	2	1913	187	223	32	47	51	16	59	7	92
Musculoskeletal Diseases	5	2	55	28	3	0	146	6	15	2	0	5	0	3	0	22
Digestive System Diseases	86	4	466	209	15	0	970	72	65	17	7	29	6	18	4	39
Stomatognathic Diseases	27	0	27	47	2	0	103	4	8	1	2	1	1	2	0	5
Respiratory Tract Diseases	153	52	59	177	2	0	325	46	15	12	21	10	0	17	3	14
Otorhinolaryngologic Diseases	18	54	12	20	0	0	67	3	2	2	5	0	0	0	0	7
Nervous System Diseases	16	4	158	111	10	4	465	43	31	6	1	17	3	1	0	175
Eye Diseases	0	0	11	3	0	0	44	2	3	0	0	0	0	0	0	5
Male Urogenital Diseases	38	4	135	62	28	1	393	81	19	7	16	36	16	11	3	28
Female Urogenital Diseases and Pregnancy Com	72	20	356	137	46	0	619	86	16	2	14	41	25	14	4	60
Cardiovascular Diseases	25	2	165	55	2	0	159	17	5	2	4	6	1	1	1	26
Hemic and Lymphatic Diseases	16	9	67	21	3	0	243	27	51	4	14	10	0	19	1	9
Congenital, Hereditary, and Neonatal Diseases a	39	15	298	79	39	1	457	36	24	3	2	25	19	7	2	53
Skin and Connective Tissue Diseases	15	5	84	28	17	1	317	17	46	5	6	11	10	9	1	19
Nutritional and Metabolic Diseases	40	15	895	122	27	4	500	50	17	10	5	24	13	7	4	59
Endocrine System Diseases	23	7	356	85	52	2	562	41	36	9	3	43	21	5	3	89
Immune System Diseases	48	89	217	79	7	1	479	27	47	7	16	15	5	14	3	45
Wounds and Injuries	8	2	29	37	4	0	79	10	30	1	2	4	0	2	0	42
Anxiety Disorders	4	0	25	23	1	1	102	14	6	1	0	3	0	1	0	356
Delirium, Dementia, Amnestic, Cognitive Disord	7	2	44	17	0	0	90	15	1	1	0	5	0	0	0	34
Dissociative Disorders	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	15
Eating Disorders	1	0	12	1	0	0	5	0	0	0	0	0	0	0	0	0
Mental Disorders Diagnosed in Childhood	1	0	22	25	0	2	88	1	2	1	0	3	0	0	0	22
Mood Disorders	2	0	12	23	1	0	72	3	2	0	0	0	0	1	0	79
Schizophrenia and Disorders with Psychotic Fea	3	0	17	27	0	0	79	1	1	0	0	1	0	1	0	34
Sexual and Gender Disorders	0	0	4	0	1	0	33	1	0	0	0	0	0	0	0	4
Sleep Disorders	0	0	4	0	1	0	6	0	0	0	0	0	0	0	0	3
Substance-Related Disorders	201	4	64	400	1	0	126	9	3	1	6	2	1	4	0	13
Other	6	0	108	125	10	2	512	66	72	32	4	40	5	10	7	40

The heat map shows the intersection of the exposure categories on the horizontal axis and the health-outcome categories on the vertical axis and the number of records in each exposure/health-outcome intersection are indicated. The largest pockets of literature are colored red, small pockets are colored blue, and pockets with no literature are colored white.

REFERENCES

Aiken CE, Ozanne SE. 2014. Transgenerational developmental programming. *Human reproduction update* 20(1): 63-75.

Ananiadou S, Kell DB, Tsujii J. 2006. Text mining and its potential applications in systems biology. *Trends in biotechnology* 24(12): 571-579.

BioCreativeIV. Arighi CN, Cohen KB, Hirschman L*et al., eds.* 2013. BioCreative IV, Biocreative IV, Bethesda, Maryland. Available: http://www.biocreative.org/resources/publications/biocreative-iv-proceedings/.

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blei DM, Lafferty JD. 2009. *Topic Models*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. . Available: https://http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.

Burris HH, Baccarelli AA. 2014. Environmental epigenetics: from novelty to scientific discipline. *J Appl Toxicol* 34(2): 113-116.

Colquhoun HL, Levac D, O'Brien KK, Straus S, Tricco AC, Perrier L, Kastner M, Moher D. 2014. Scoping reviews: time for clarity in definition, methods, and reporting. *Journal of clinical epidemiology* 67(12): 1291-1294.

Laird PW. 2010. Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics* 11(3): 191-203.

Ledford H. 2008. Language: Disputed definitions. *Nature* 455(7216): 1023-1028.

Levac D, Colquhoun H, O'Brien KK. 2010. Scoping studies: advancing the methodology. *Implementation science : IS* 5: 69.

Manning CD, Raghavan P, Schutze H. 2008. *Introduction to Information Retrieval*, New York: Cambridge University Press.

Mimno D. 2012. Computational historiography. *Journal on Computing and Cultural Heritage* 5(1): 1-19.

Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. 2012. Extracting semantically enriched events from biomedical literature. *BMC bioinformatics* 13: 108.

Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. 2014. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics* 51: 242-253.

NIEHS. 2012. *NIEHS Strategic Plan*. Available: http://www.niehs.nih.gov/about/strategicplan/index.cfm [accessed 1/14/15].

NIH. 2009. *NIH Epigenomics Roadmap Initiative*. National Institute of Health. Available: http://www.ncbi.nlm.nih.gov/bioproject?Cmd=DetailsSearch&Term=34535%5Buid%5D [accessed 2/25/15].

NIH. 2014. *Epigenomics*. Available: http://commonfund.nih.gov/epigenomics/index [accessed 3/12/2015].

O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 4: 5.

Project RE. 2010. *Overview*. Available: http://www.roadmapepigenomics.org/overview [accessed 2/25/15].

Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP, Thomas J. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Res Synth Methods* 5(1): 31-49.

Thomas J, J. B, Graziosi S. 2010. EPPI-Reviewer 4.0: Software for Research Synthesis. Software E-C. London, Social Science Research Unit, Institute of Education.

Thomas J, McNaught J, Ananiadou S. 2011. Applications of text mining within systematic reviews. *Res Synth Methods* 2: 1-14.

Wallace BC, Small K, Broadley CE, Lau J, A. TT. 2012. Deploying an interactive machine learning system in an evidence-based practice center, Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics - IHI '12: 819.

Wilson V. 2014. Research Methods: Scoping Studies. *Evidence Based Library & Information Practice* 9(4): 97-99.

Yu W, Clyne M, Dolan SM, Yesupriya A, Wulf A, Liu T, Khoury MJ, Gwinn M. 2008. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC bioinformatics* 9: 205.