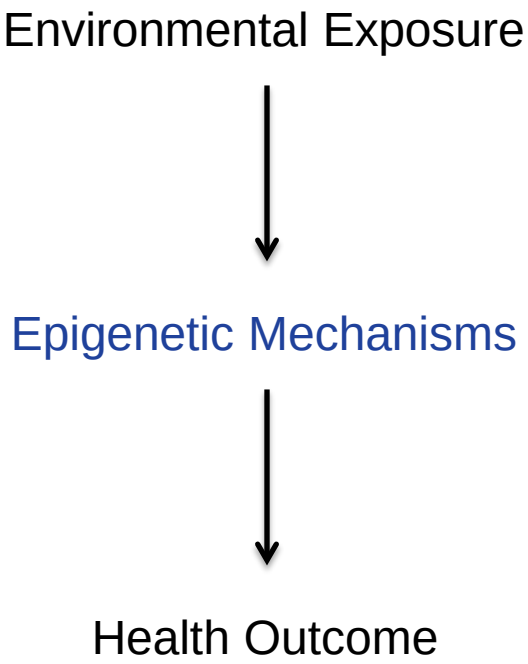
NTP
National Toxicology Program

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

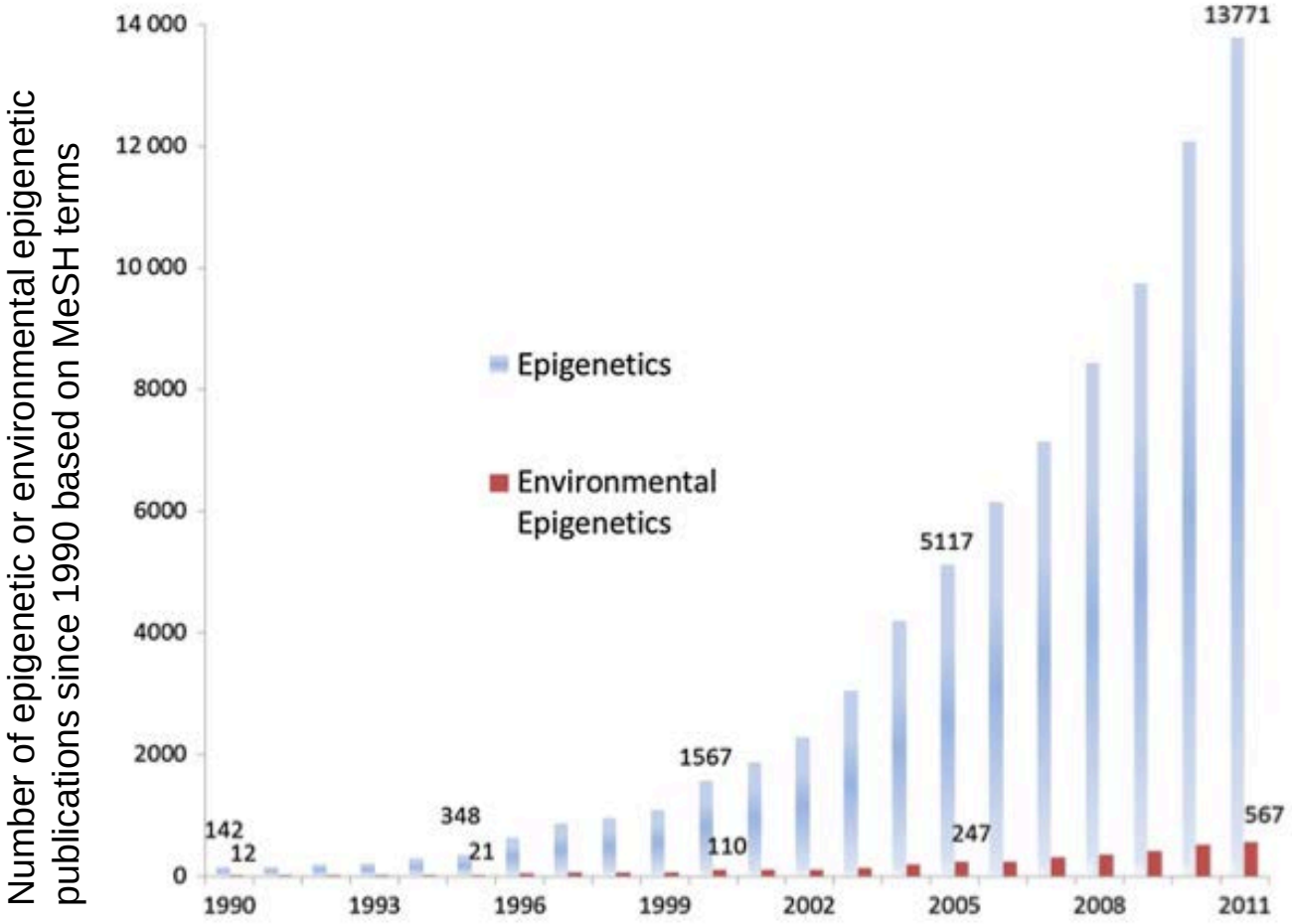
Katherine Pelch, PhD
Office of Health Assessment and Translation
National Institute of Environmental Health Sciences

NTP Board of Scientific Counselors Meeting
June 16, 2015

Where are these links the strongest?



There is a growing interest in epigenetics

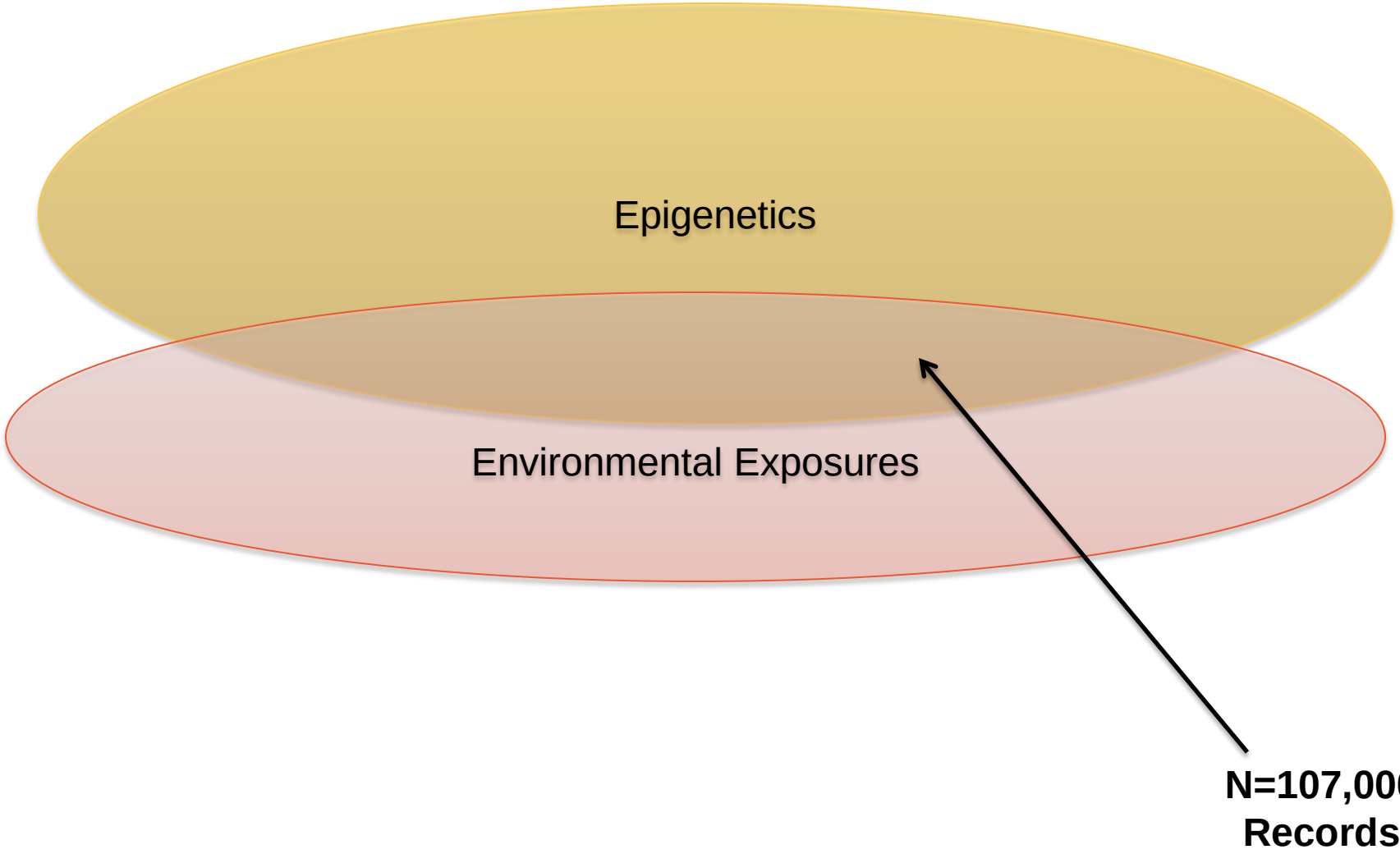


Burris HH, Baccarelli AA. 2014. Environmental epigenetics: From novelty to scientific discipline. J Appl Toxicol 34:113-116.

Addressing the need

- 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

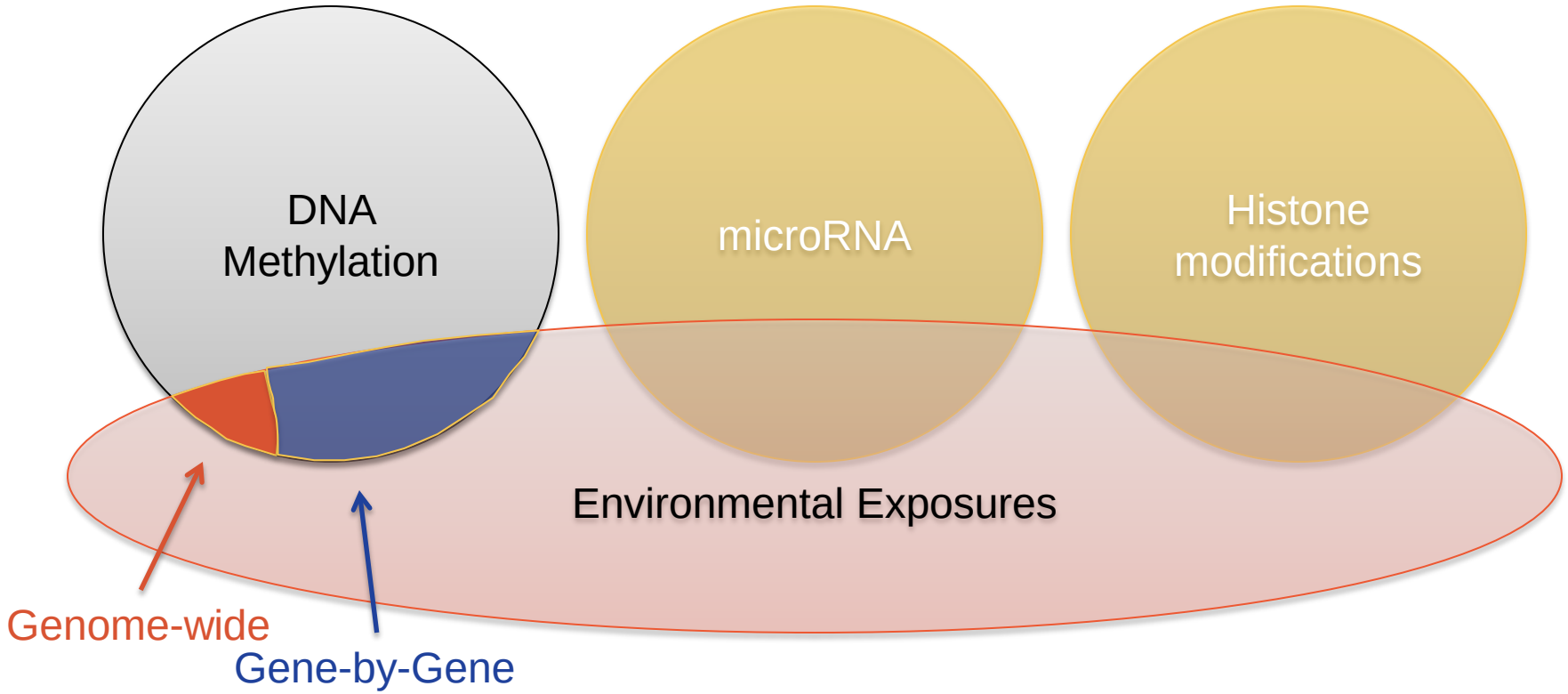
How was the research question refined?



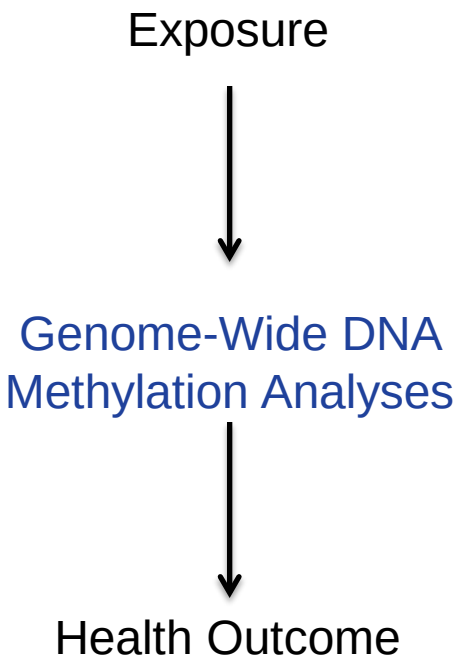
How was the research question refined?

- 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 m
- Genon
- Genon
- Epiger

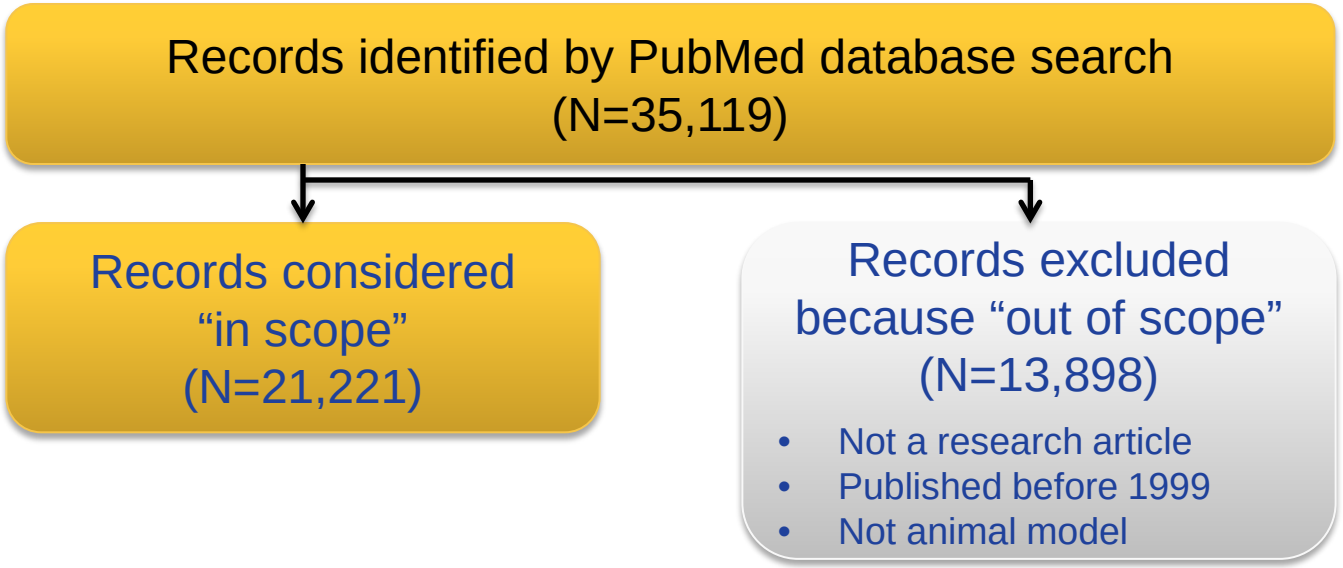
DNA METHYLATION - Search "DNA Methylation"[mh] OR ((DNA[mh] OR DNA[tiab] OR cytosine[mh] OR cystosine*[tiab] OR Alu[tiab] OR LINE-1[tiab] OR genes[tiab] OR genom*[tiab]) AND methylat*[tiab]) OR "CpG islands"[mh] OR ((CpG[tiab] OR CpGs[tiab] OR CPG's[tiab] OR non-CG[tiab] OR non-CPG) AND methylat*[tiab]) OR differential-methylat*[tiab] OR differentially-methylated[tiab] OR partially-methylated[tiab] OR "partial methylation"[tiab] OR hemimethylat*[tiab] OR hemi-methylat*[tiab] OR hypermethylat*[tiab] OR hyper-methylat*[tiab] OR hypomethylat*[tiab] OR hypo-methylat*[tiab] OR demethylat*[tiab] OR unmethylat*[tiab] OR nonmethylat*[tiab] OR DNA modification methylases[mh] OR methylas*[tiab] OR demethylas*[tiab] OR methyltransferase*[tiab] OR "DNA Mtase"[tiab] OR Dnmt*[tiab] OR 5-methylcytosine[mh] OR "5-methylcytosine"[tiab] OR 5-methyl-cytosine[tiab] OR 5mc[tiab] OR 5meC[tiab] OR Hydroxymethylat*[tiab] OR 5-hydroxymethylcytosine[tiab] OR 5-hydroxy-methylcytosine[tiab] OR 5-hydroxymethyl-cytosine[tiab] OR 5hmC[tiab] OR Methyl-binding-domain*[tiab] OR methyl-binding-protein*[tiab] OR MBD2*[tiab] OR MBD3*[tiab] OR MBD4*[tiab] OR "methyl-cpg-binding protein 2"[mh] OR methyl-CpG*[tiab] OR MeCP2[tiab]

Global DNA methylation OR

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

N=35,119 records retrieved

How were records processed?

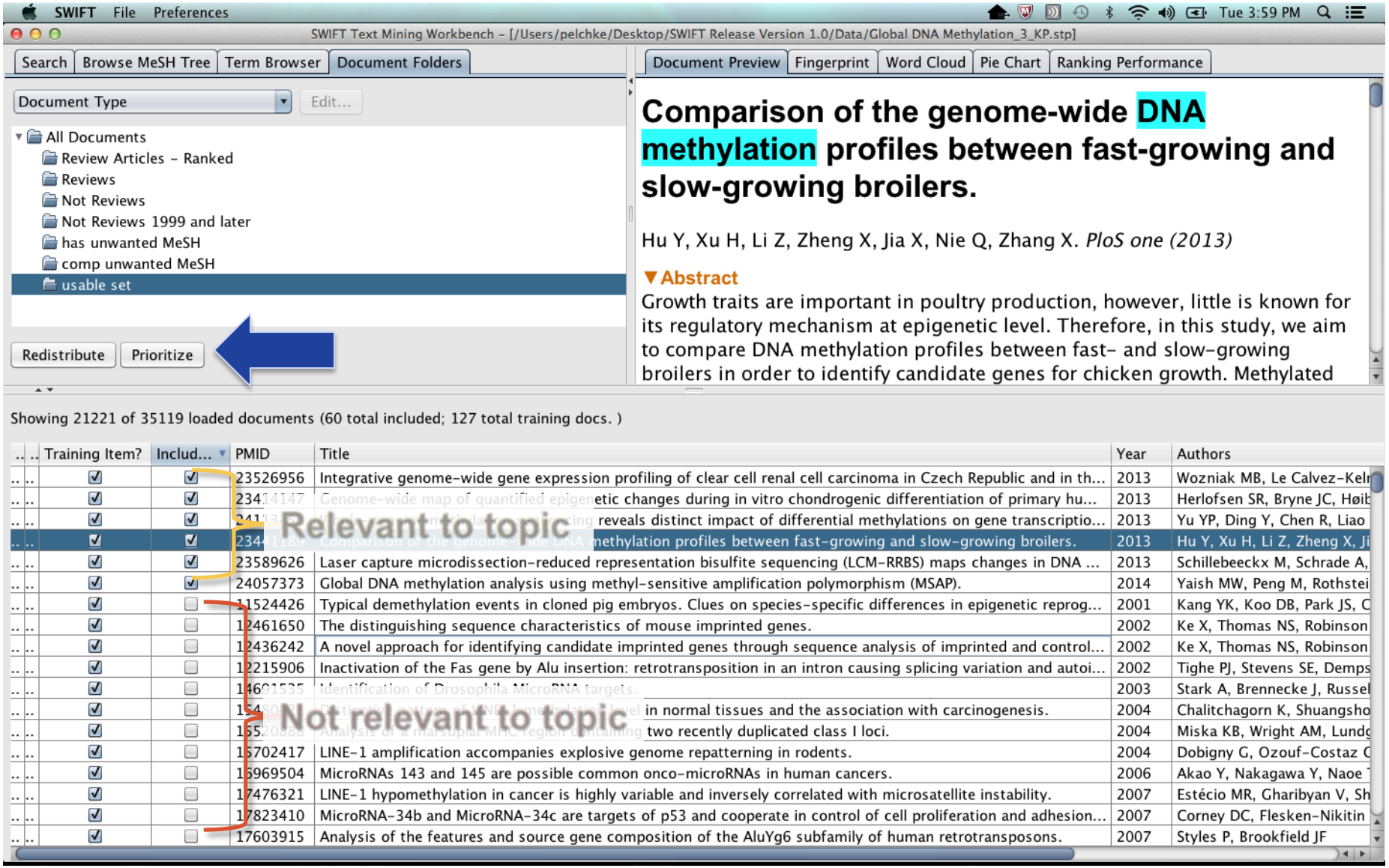


What topics are in this literature?

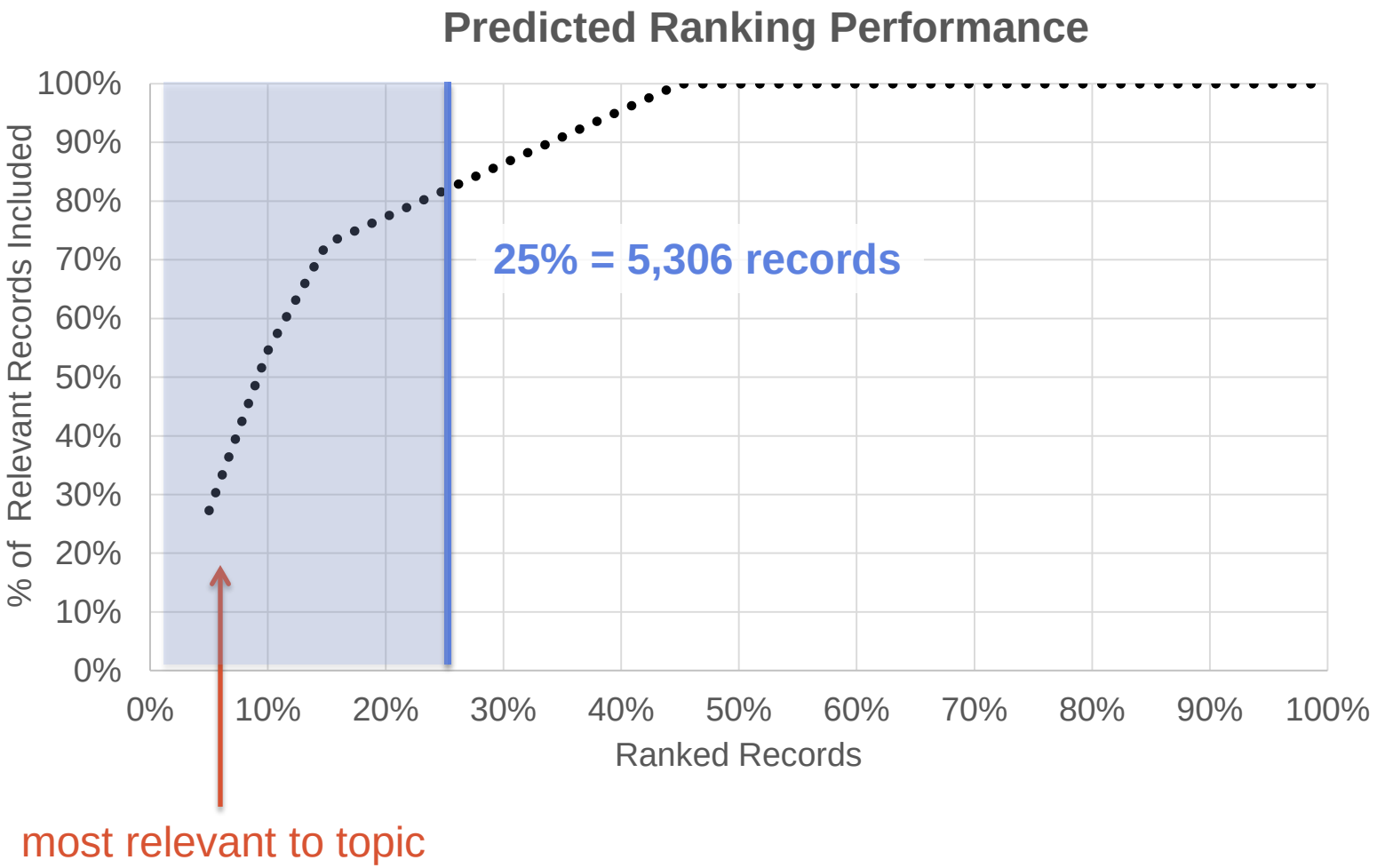
	# records
Topic 28: genes, expression, gene, analysis, identified, microarray, expressed, involved, identify, genome-wide	11778
Topic 41: epigenetic, mechanisms, development, gene, genetic, studies, disease, epigenetics, regulation, recent	8637
Topic 40: methylation, dna, cpg, methylated, promoter, island, 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 2: histone, chromatin, epigenetic, modifications, acetylation, gene, modification, methylation, dna, lysine	4933
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

- 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 ✖)

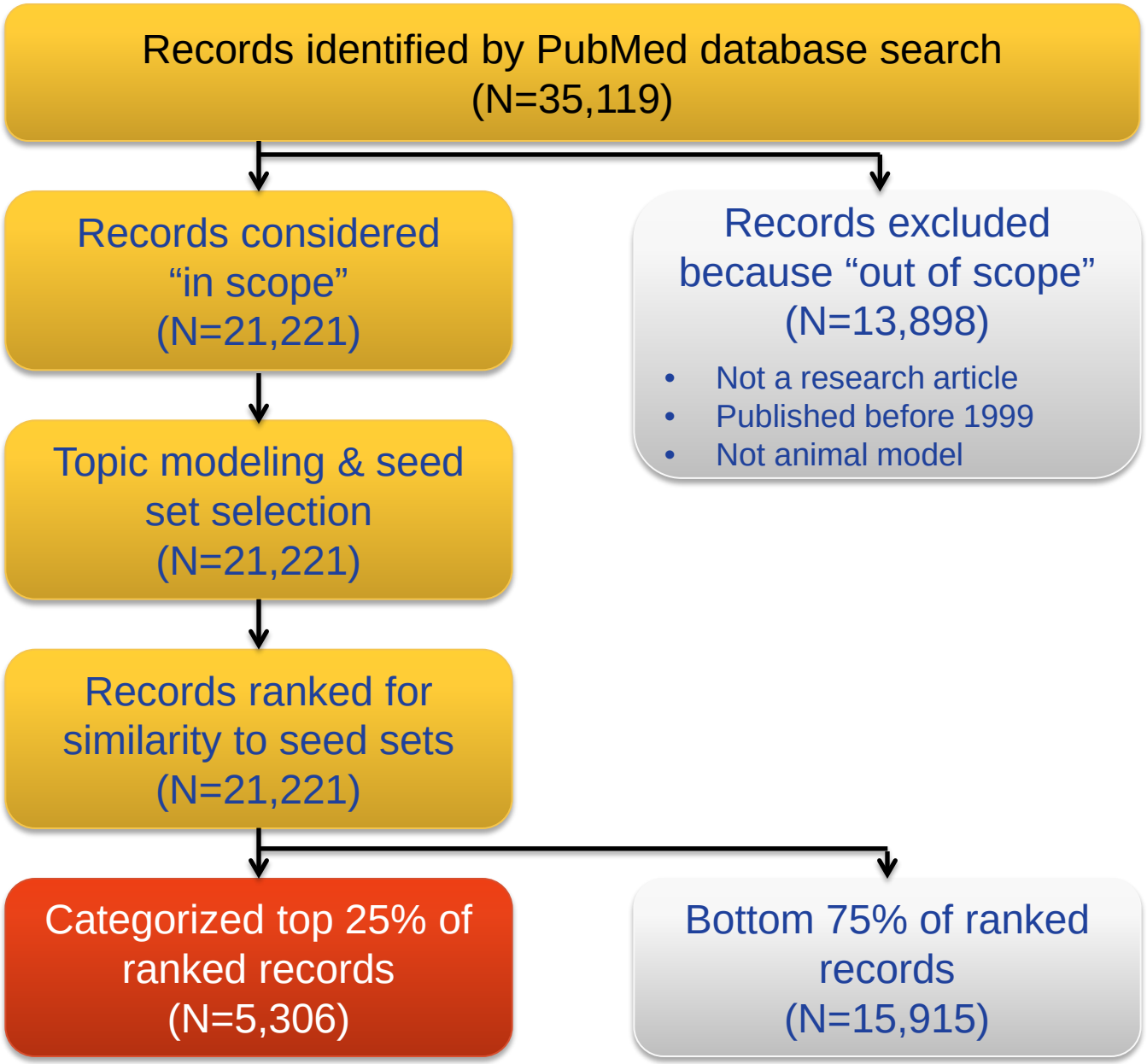
How do the records compare to seed sets?



Which were the most relevant records?



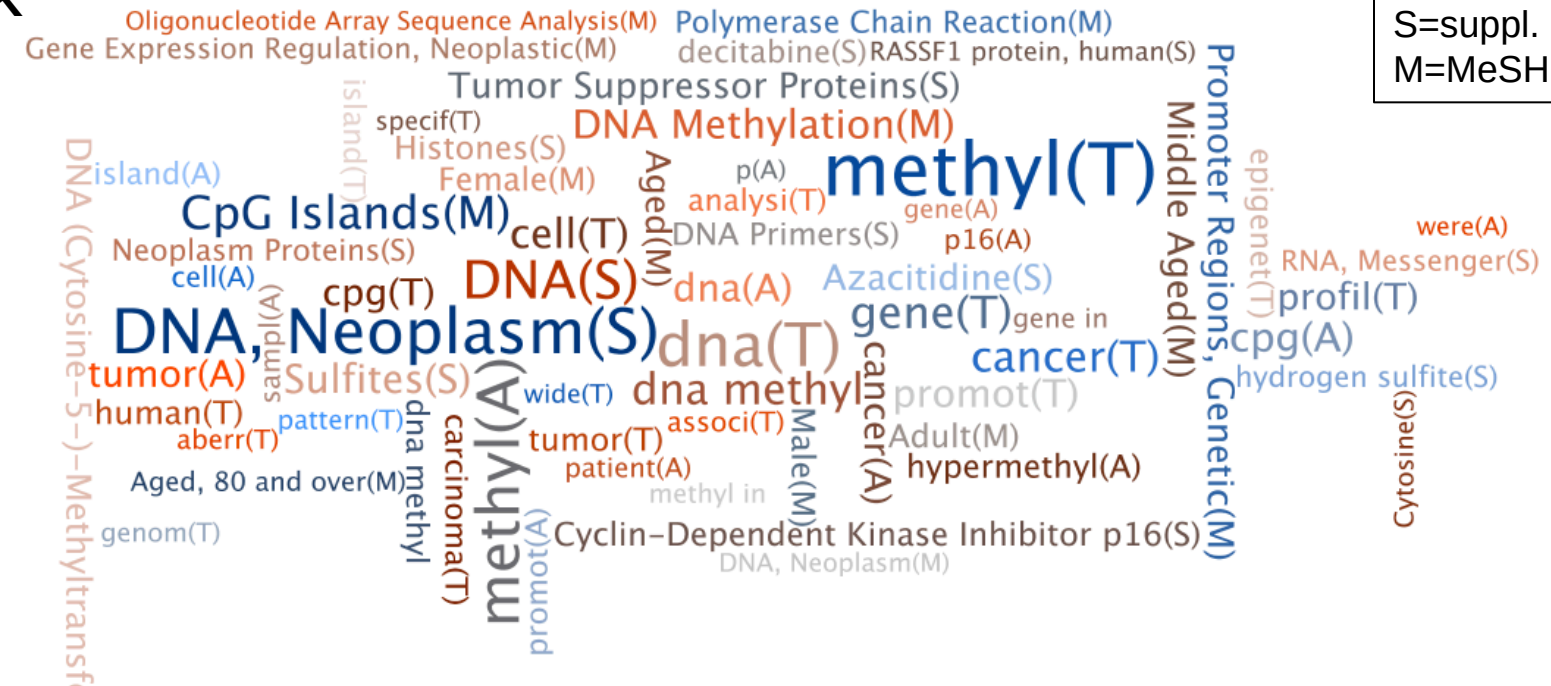
How were records processed?



How effective was the enrichment?

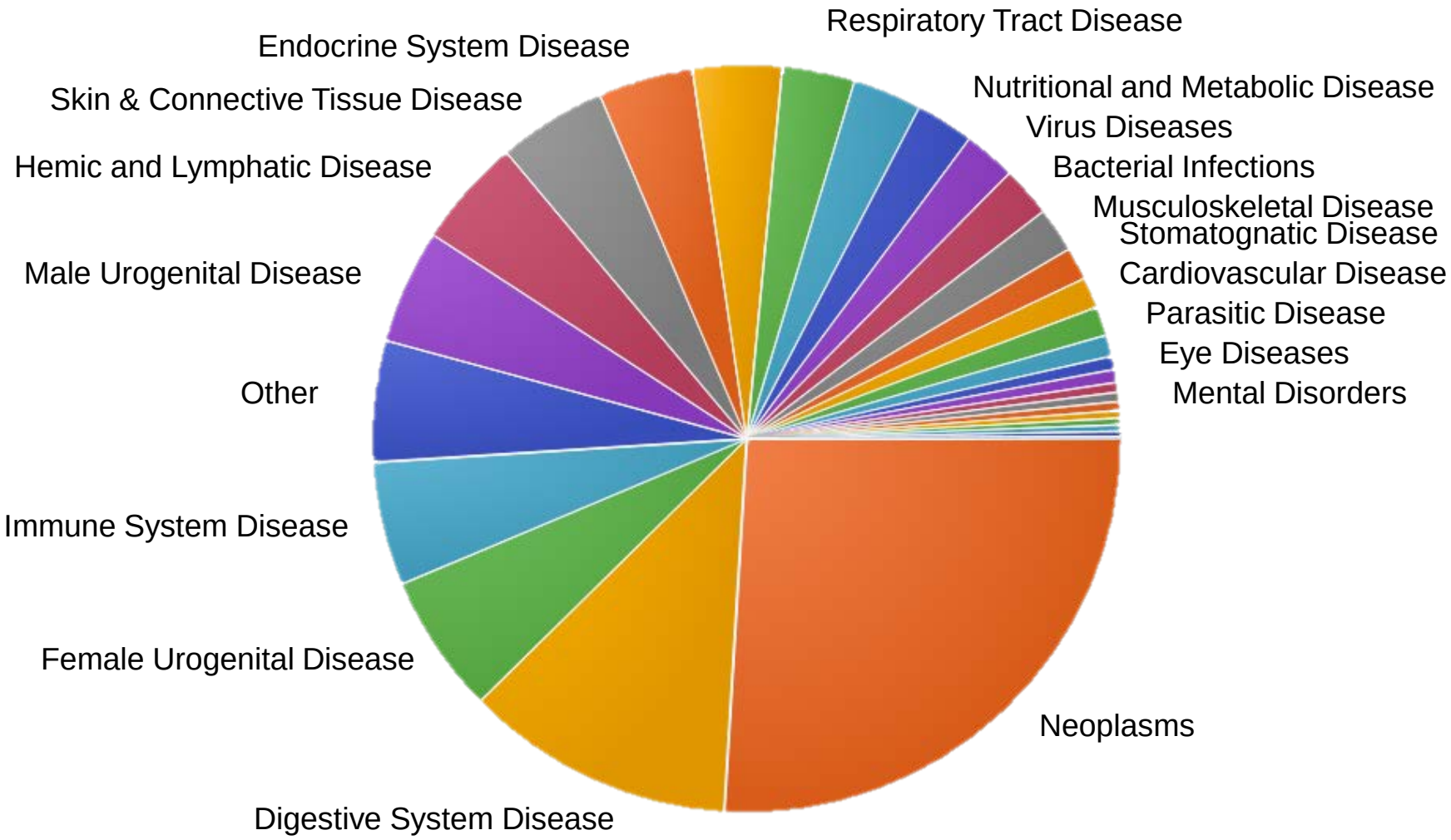
Original
N=35K

Top 25%
N=5K



T=title
A=abstract
S=suppl. info
M=MeSH

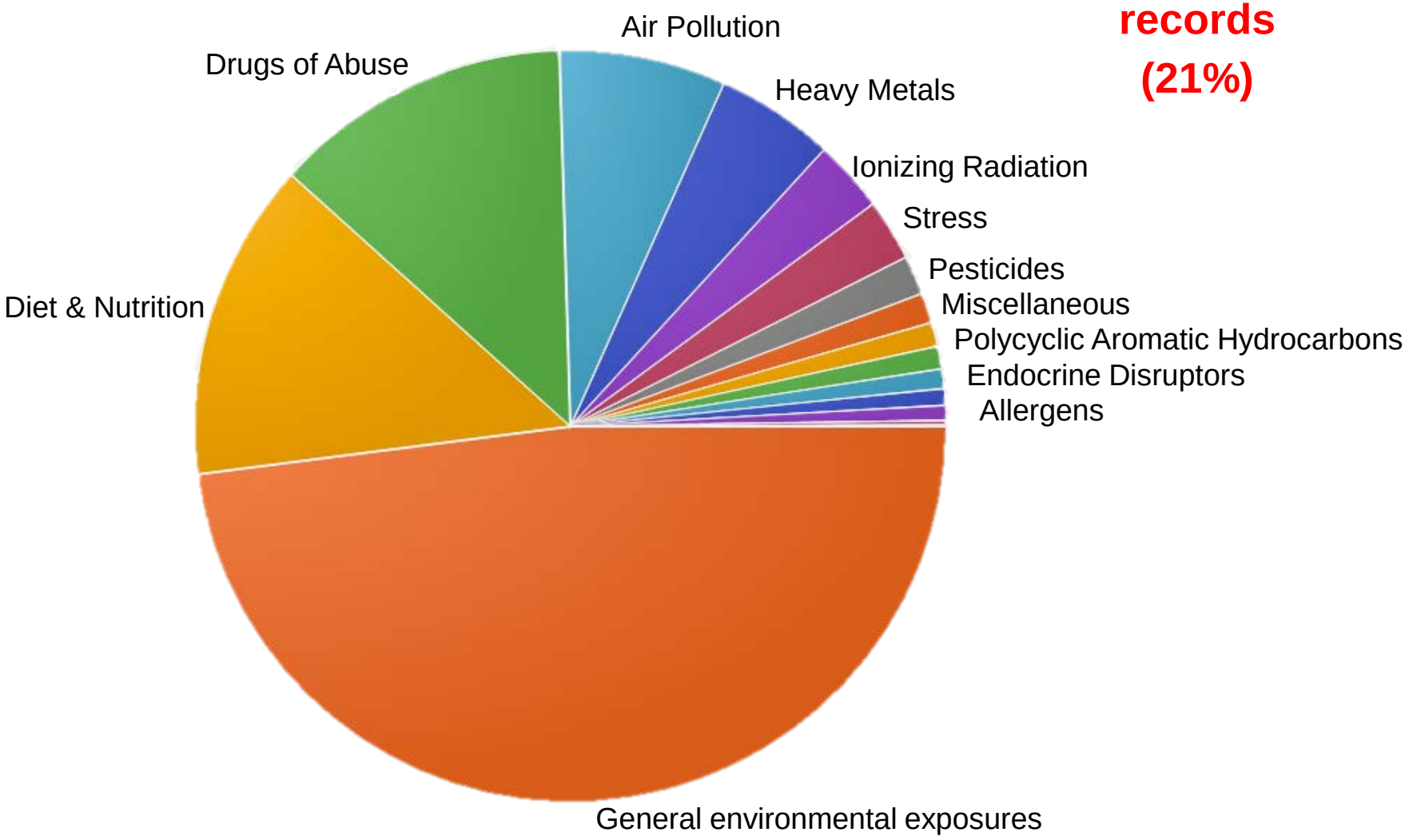
What health outcomes are investigated?



N=5,306 records

What types of exposures are investigated?

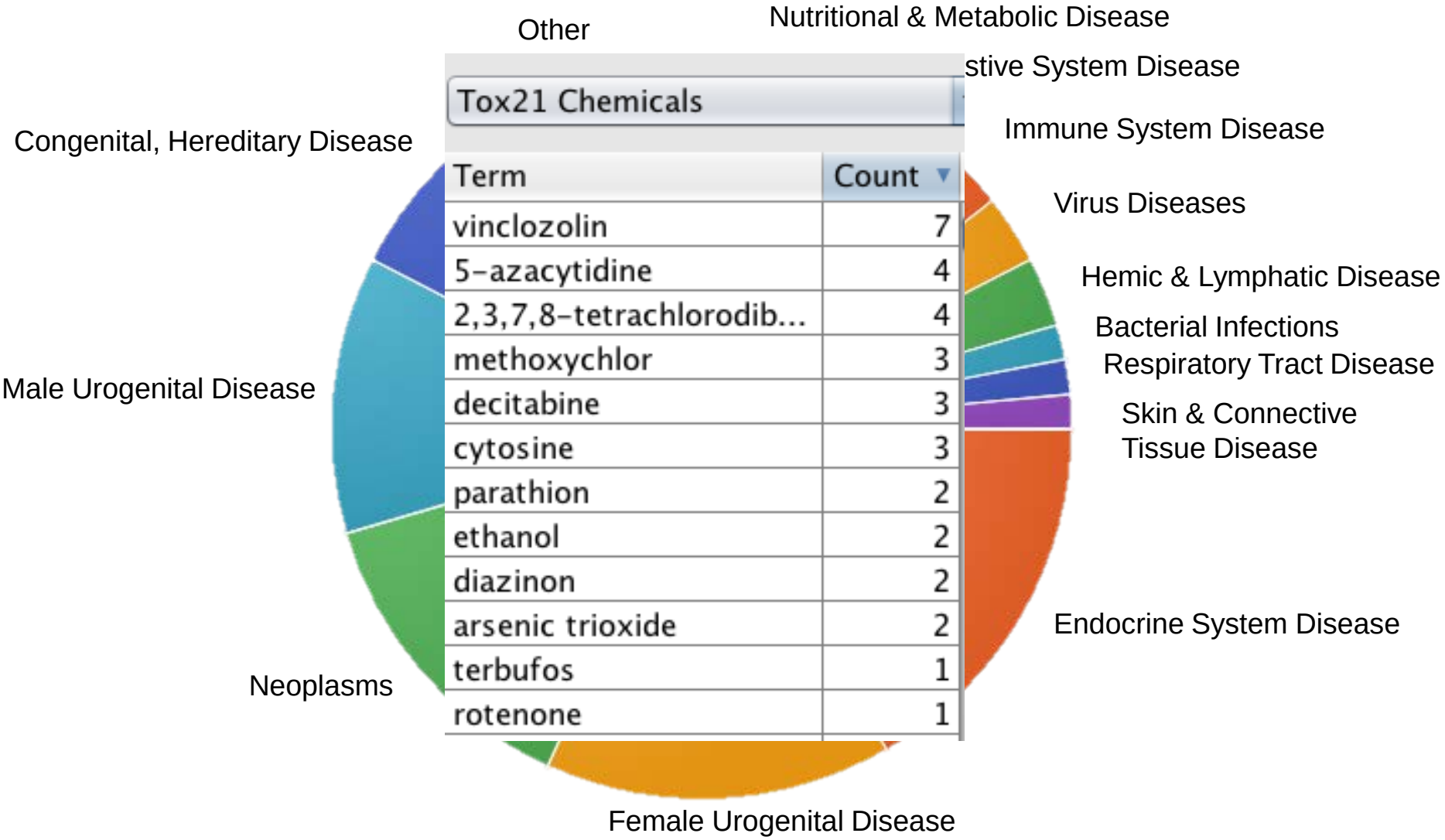
N=1,130 of 5,306 records (21%)



Evaluating exposure & health outcomes

Exposure = pesticide

N=26 records



Term	Count
vinclozolin	7
5-azacytidine	4
2,3,7,8-tetrachlorodib...	4
methoxychlor	3
decitabine	3
cytosine	3
parathion	2
ethanol	2
diazinon	2
arsenic trioxide	2
terbufos	1
rotenone	1

Summary

- 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