ToxCast Data Expands Universe of Chemical-Gene Interactions

Sean Watford

Student Services Contractor

USEPA National Center for Computational Toxicology

Graduate Student, UNC-Chapel Hill

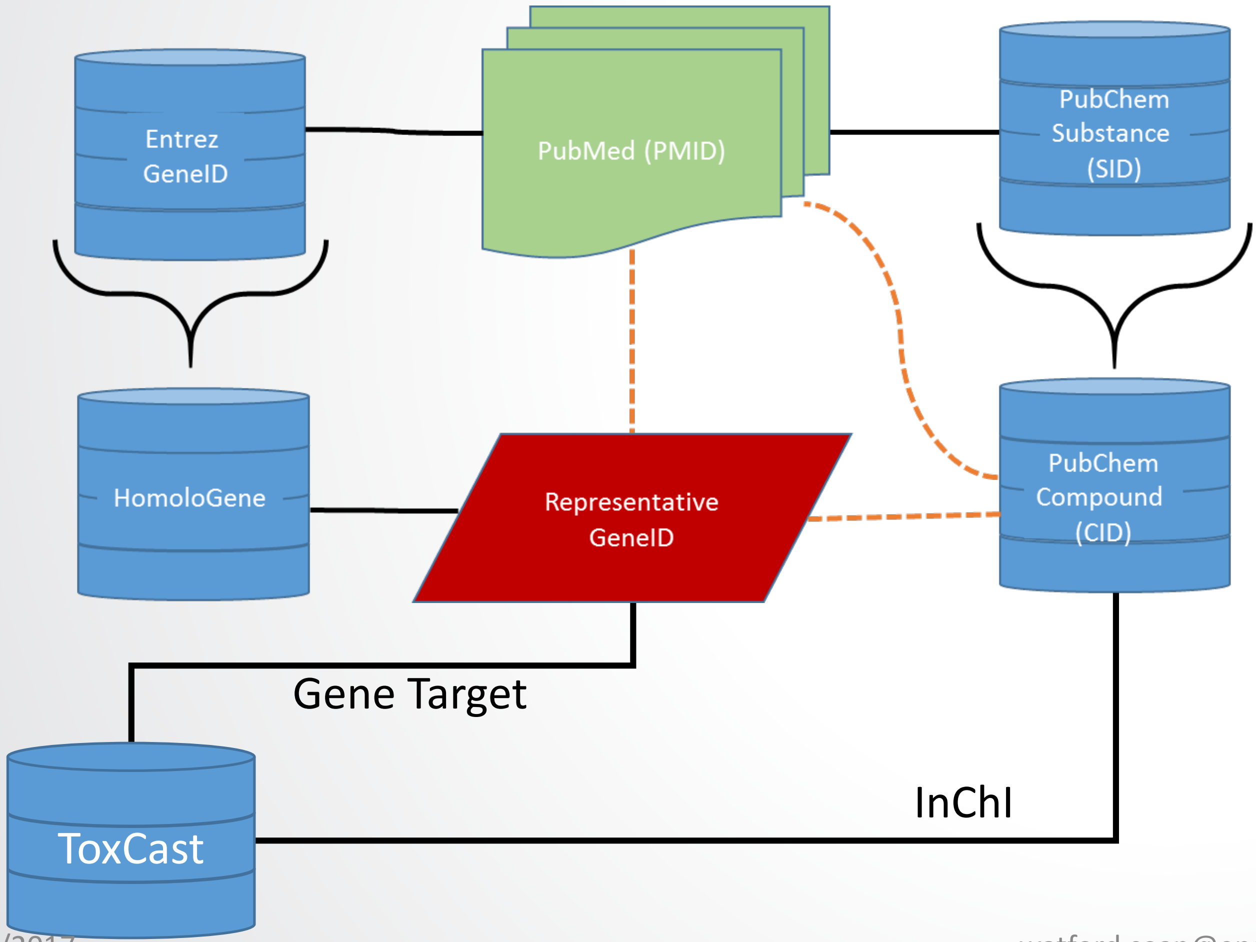
Watford.Sean@epa.gov

Disclosure

- I have no actual or potential conflict of interest in relation to this presentation
- This presentation does not necessarily reflect U.S. EPA policy

watford.sean@epa.gov

Curated Literature and ToxCast



Summary of ToxCast Chemical-Gene Activity	
CID-GeneID actives	47,423 (**165**)
CIDs	5,011
GeneIDs	321

Total of 165 Chemical-Gene activities with corresponding associations in curated literature, but **does it infer bioactivity**?

Used CTD as an external resource, but GeneID mappings from NCBI didn't overlap well with CTD when compared to Chemicals

Conclusion: **More gene-article mappings are needed**

watford.sean@epa.gov

Genes in Literature:
Manually Curated GeneID-Article Associations

Resource	gene2pubmed	generif	ctd	UniProt/ Swiss	Reactome	RGD	MGI	Total
gene2pubmed	(4625706, 998833, 8707898)	(78404, 611872, 969857)	(39111, 13369, 9639)	(216310, 755490, 1917108)	(10799, 11194, 21391)	(30094, 56918, 118972)	(41890, 174596, 601460)	
generif		(84929, 628432, 995232)	(19966, 9634, 7220)	(50463, 511544, 743138)	(9423, 5341, 5657)	(20075, 21229, 21639)	(12070, 77778, 76837)	
ctd			(40862, 55568, 763056)	(22129, 10828, 7548)	(9346, 177, 250)	([illegible], 1884, 829)	(3585, 2329, 189)	
UniProt/Swiss				(231051, 874281, 2812099)	(10371, 10115, 13554)	(22740, 38943, 62347)	(15653, 166141, 492417)	
Reactome					(10850, 15045, 198103)	(5240, 682, 399)	(810, 2227, 738)	
RGD						(30172, 58144, 187776)	(12086, 5828, 24347)	
MGI							(42125, 177324, 615203)	
Total								(4649012, 1178220, 10627745)

Overlap

Only Human Genes?

UniRef50 Cluster ESR1

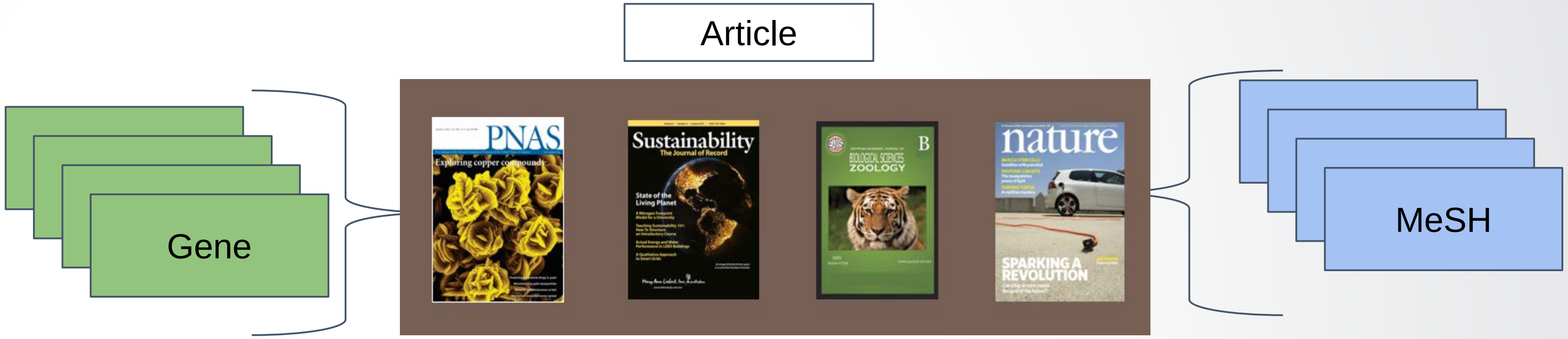
Entry name	Organism
P03372	**Homo sapiens**
P03372-3	Homo sapiens
P49884	Bos taurus
Q29040	Sus scrofa
Q53AD2	Felis catus
Q9TV98	Equus caballus
P06211	Rattus norvegicus
P19785	Mus musculus
Q9QZJ5	Mesocricetus auratus
P06212	Gallus gallus
Q91250	Taeniopygia guttata

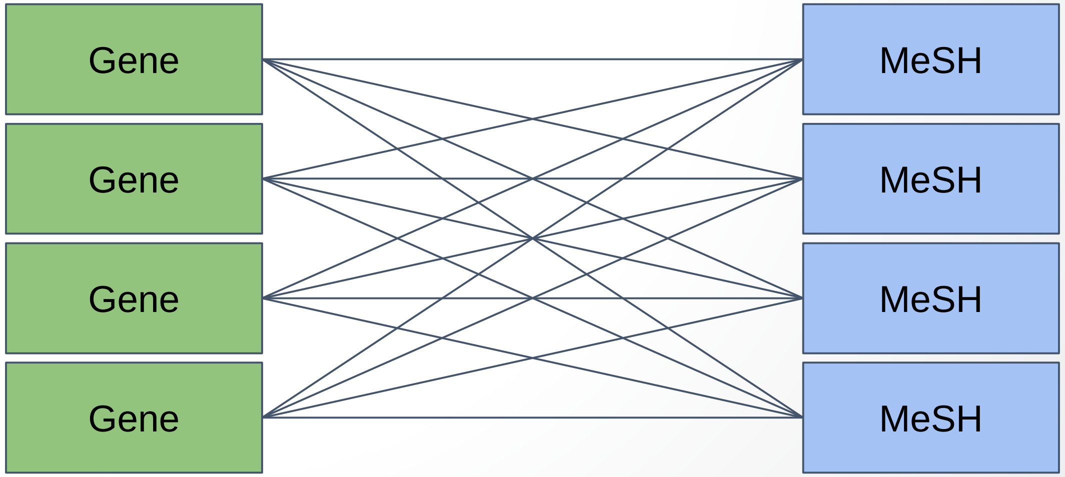
~700K Articles

(vs <500k for humans only)

Gene-MeSH Network

- National Library of Medicine: Medical Subject Headings (MeSH, MeSH terms)
  - Keywords used to categorize an article
  - Descriptors
    - Hierarchical tree
  - Qualifiers
    - Qualify a descriptor
  - Supplementary Concept Records
    - Not frequent enough to be a descriptor
    - Mapped to descriptor
  - Major Topic



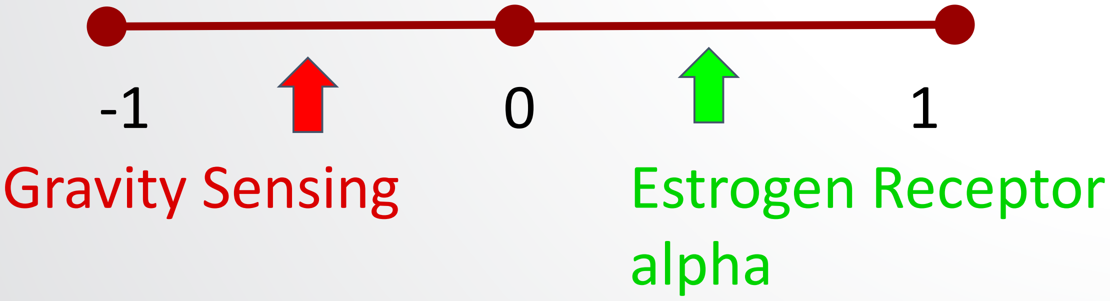


Ranking Gene-MeSH Associations: Separating Signal from Noise

- Normalized Pointwise Mutual Information (NPMI)
  - Rank measure
  - Amount of information two concepts share when compared to all other occurrences to another concept
  - Allows for comparison across datasets

Gene	MeSH Term	NPMI
Estrogen Receptor alpha (ESR1)	Estrogen Receptor alpha	0.4304
Estrogen Receptor alpha (ESR1)	Receptors, Estrogen	0.4163
Estrogen Receptor alpha (ESR1)	Estrogen Receptor beta	0.3809
Estrogen Receptor alpha (ESR1)	Receptors, Steroid	0.3496

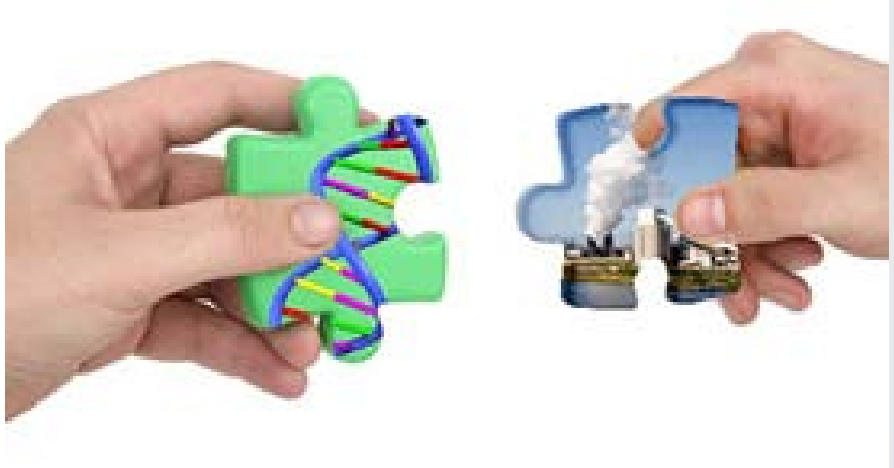
Search with Gene Estrogen Receptor alpha



watford.sean@epa.gov

Identifying Critical Genes in Breast Cancer

- Silent Spring Institute (SSI) has a focus on breast cancer prevention through research on the effects the environment plays in breast cancer risk
- SSI has compiled a list of nearly 300 genes through expert solicitation that are linked to breast cancer with experimental data
    - The list was created to cover genes involved with mammary tissue growth and development along with key cancer characteristics
    - Created to form a panel to further experimentally explore the function and how chemicals interact with the genes



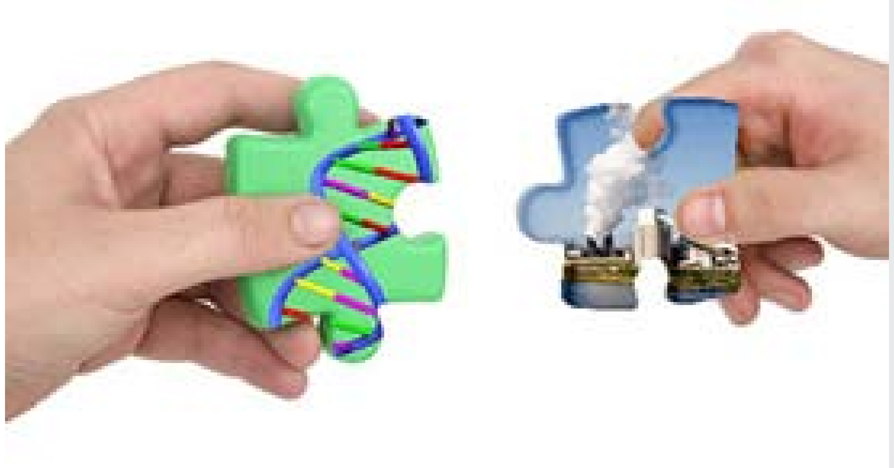
SilentSpring.org

SILENT SPRING INSTITUTE

watford.sean@epa.gov

Identifying Critical Genes in Breast Cancer

- Because Breast Cancer is a complex disorder influenced by both genetic and environmental factors, the biological scope is not entirely understood
- Questions remain about whether the reference list of 300 genes from SSI adequately represents a meaningful portion of breast cancer-related genes



Can we use Big Data approaches to evaluate the list of 300 genes and potentially identify missing critical genes?

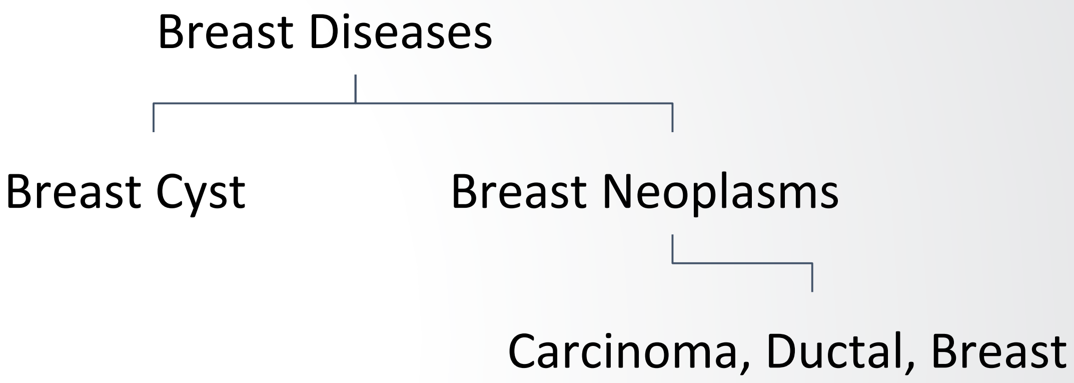
SilentSpring.org

SILENT SPRING INSTITUTE

watford.sean@epa.gov

Approach :
Finding Breast Cancer-Related MeSH Terms

SSI Identified Characteristic	MeSH terms
Angiogenesis	Angiogenesis, Pathologic; Angiogenesis, Physiologic
Evading apoptosis	Apoptosis
Cell cycle changes	Cell Cycle
Cell proliferation	Cell Proliferation
Epigenetics	Epigenomics
Genotoxicity	DNA Damage; DNA Repair
Altered peptide (growth) hormone activity	Growth Hormone
Receptor-mediated effects	Gonadal Steroid Hormones
Immortalization	Cell Survival
Immune modulation	Immune System
Inflammation	Inflammation
Mammary	Breast Diseases; Breast
Metabolic activation	Xenobiotics
Oxidative stress	Oxidative Stress



Breast Diseases

Breast Cyst

Breast Neoplasms

Carcinoma, Ductal, Breast

Assumption:

Children are more specific than parents in the hierarchical MeSH tree

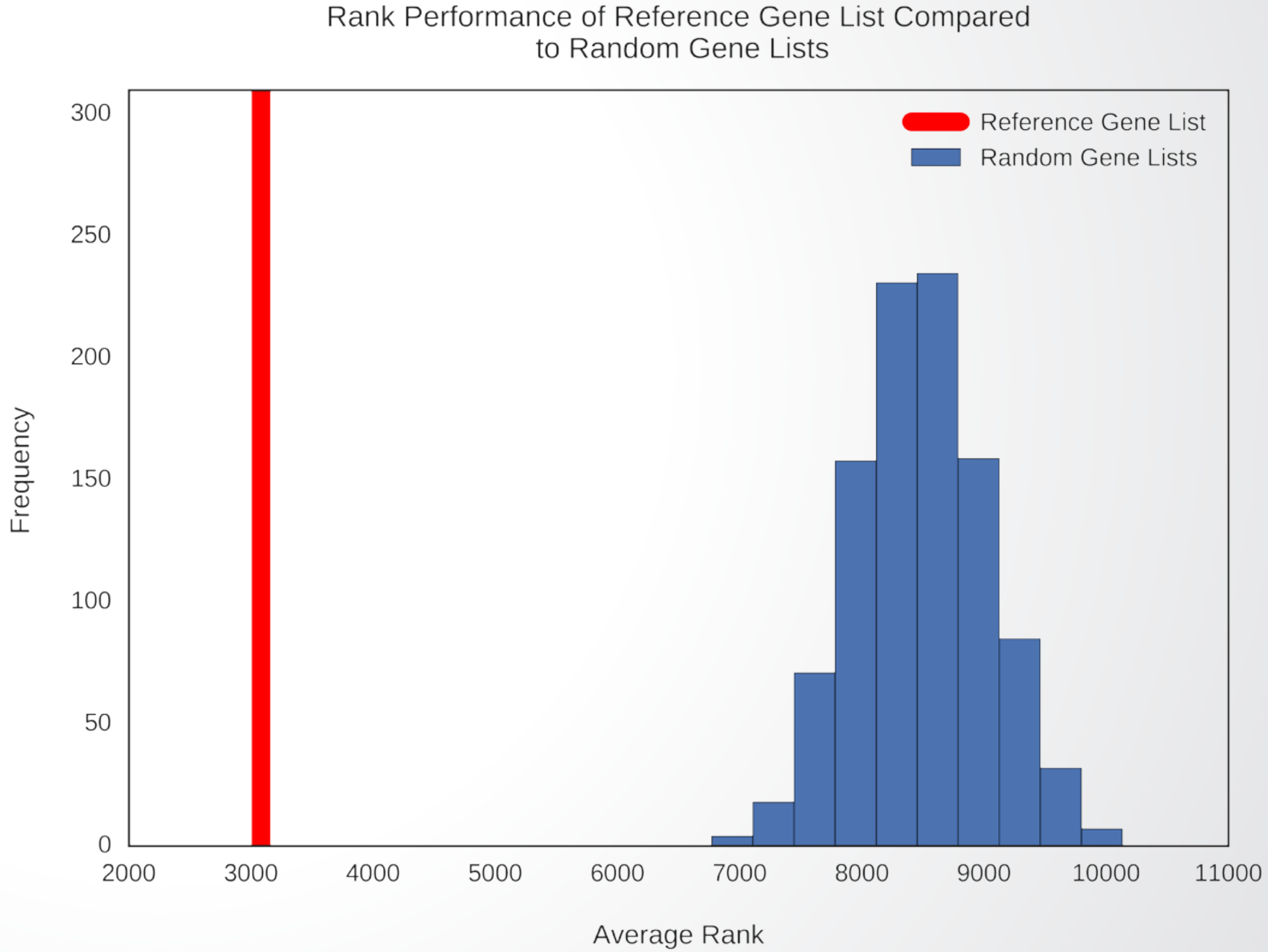
- Some articles are tagged with “Breast Diseases”, but others are only tagged with “Breast Neoplasms”
- **All are relevant**

Reference Gene Performance

How well do the 300 reference breast cancer genes perform against randomly generated gene lists?

- Maximize the ranks of the 300 reference breast cancer genes
    - Randomly generate gene lists of the same length and average the ranks using the search results from the integrated resource using the 17 selected MeSH terms

The 300 reference breast cancer genes outperformed randomly generated gene lists (p << 0.001)

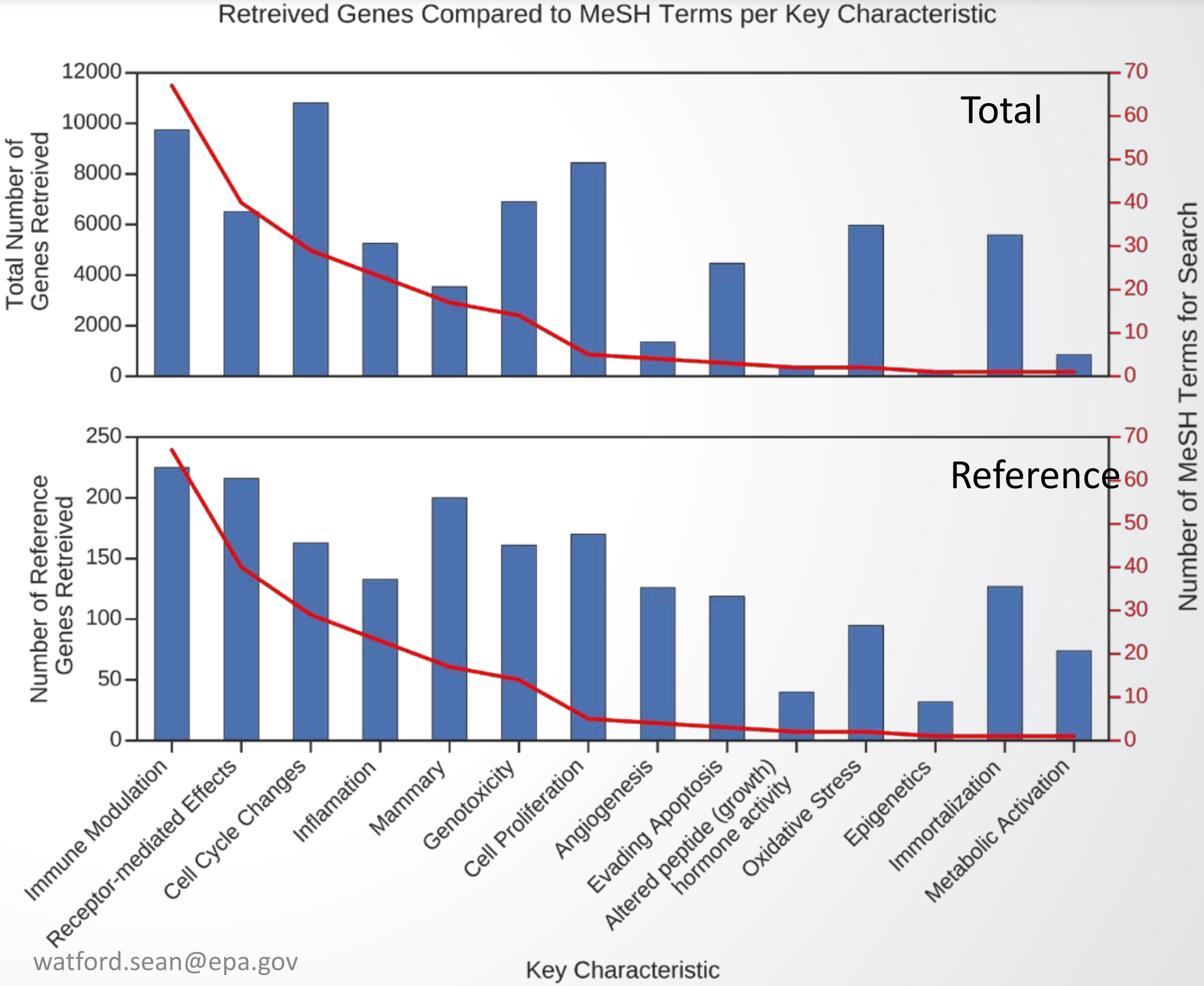


p << 0.001

watford.sean@epa.gov

Genes Returned Compared to Number of MeSH terms

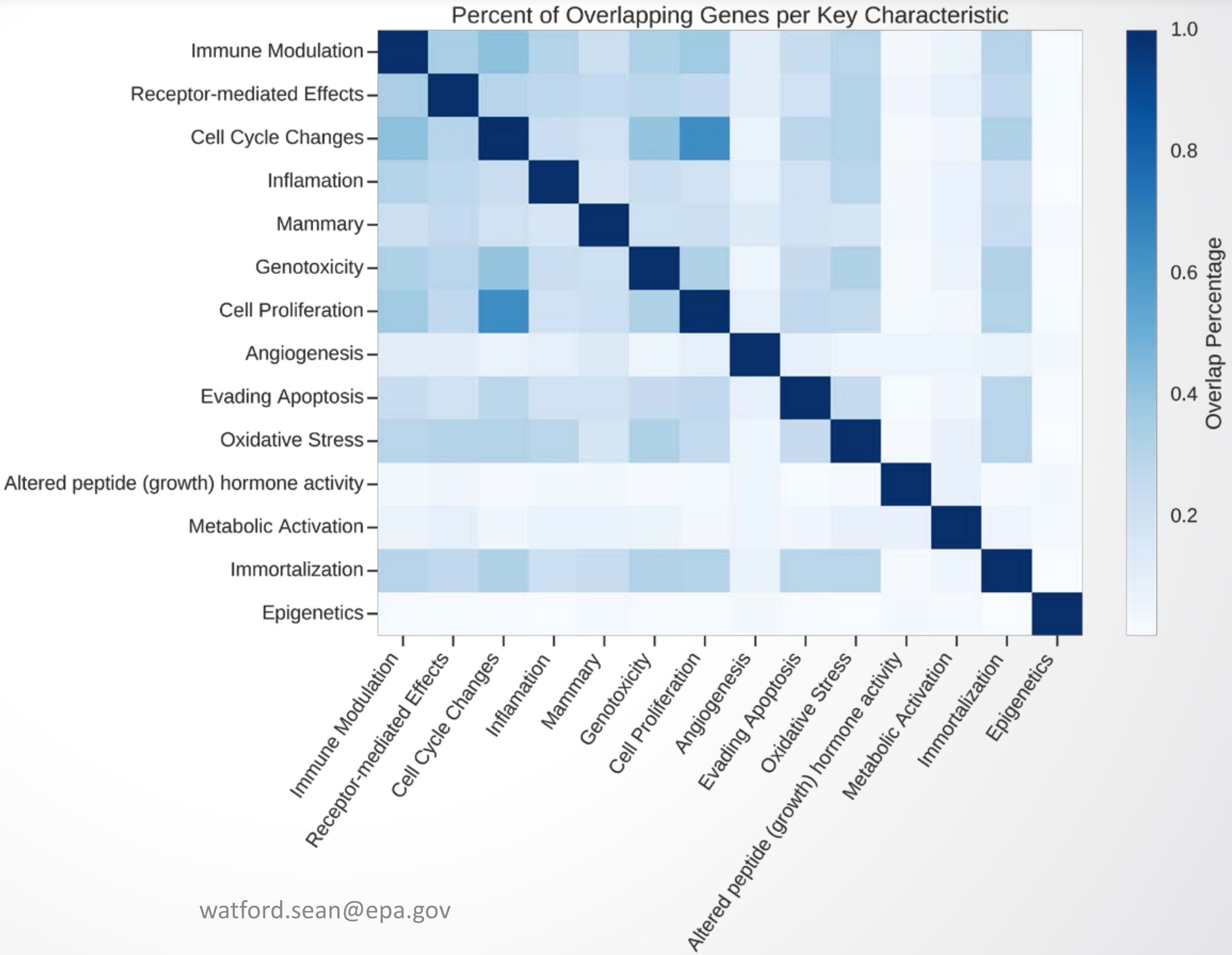
- Sorted descending by “Number of MeSH terms for Search”
- The number of genes retrieved does not directly correlate with number of MeSH terms, however, lower number of genes may be improved with better MeSH term selection
- Number of genes may also correlate with publication year
    - For example, Epigenetics is a newer term, so this may show gaps in knowledge surrounding a topic



watford.sean@epa.gov

Gene Overlap Between Key Characteristics

- Sorted descending by number of children
- Overlap Percentage $= \frac{(C_1 \cap C_2)}{(C_1 \cup C_2)}$
- The characteristics with more MeSH terms have more overlap, but, overall, overlap is low
- Indicates little redundancy across gene lists per characteristic



watford.sean@epa.gov

Conclusions

- Current state of curated biomedical literature is sparse, but adequate for many topics
  - Breast Cancer
  - Type II Diabetes
  - Other disorders with a lot of coverage
- Large gaps in knowledge exist in certain areas (e.g. Epigenetics)
- Need ongoing manual curation efforts, but systematic approaches could aid in identifying gaps and speed up curation process

watford.sean@epa.gov

Ongoing Work

- Currently, this process has been used to identify 500 breast cancer genes to experimentally explore in a pilot project
- Adapt to other complex disorders like Type 2 Diabetes
- Improve MeSH term selection
  - MeSH co-occurrence network to generate ranked MeSH-MeSH profiles for more targeted selection of MeSH terms for gene queries
- More gene-PMID associations
  - Systematic approaches to predict or infer gene-PMID associations
- Chemical-Gene interactions
  - Integrate chemical-PMID resources

watford.sean@epa.gov

Acknowledgments

National Center for Computational Toxicology

Matt Martin

Imran Shah

Silent Spring Institute

Rachel Grashow

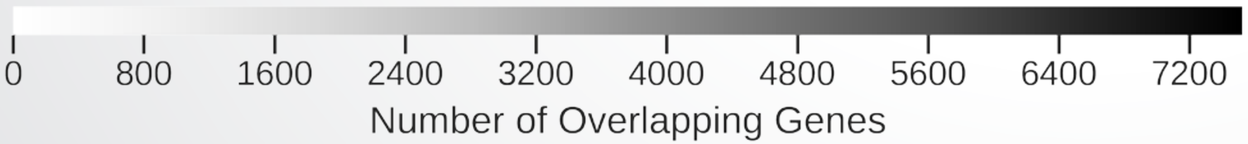
Vanessa De La Rosa

Ruthann Rudel

SilentSpring.org SILENT SPRING INSTITUTE

watford.sean@epa.gov

Key Characteristic Network


Number of Overlapping Genes